Multi-level proteomics profiling of virus-restrictive and -permissive cellular environments to discover novel regulators of virus infection



Honglin Chen

St Hugh's College

Department of Biochemistry

University of Oxford

A thesis presented for the degree of Doctor of Philosophy

Trinity 2022

Abstract

Viruses rely on host cell resources to replicate and propagate. To mitigate, human cells can express a diverse set of antiviral proteins to combat virus infection. Understanding the host proteins that aid or repress virus infection can be instrumental for the development of novel antiviral strategies. This recent decade has witnessed rapid developments in discovering host proteins that regulate virus infection. However, our current knowledge of the landscape of host proteins that drive or restrict virus infection remains largely incomplete. Here I employed a multi-level proteomics analysis to identify novel regulators of virus infection in a model system that comprises cells highly permissive to viruses (HEK293T), cells at steady-state (HEK293), and cells restrictive to viruses (HEK293 stimulated with interferon alpha (IFN-α)). This analysis examines changes in protein and RNA abundance, phosphorylation, and RNA binding across different cellular states. Whole-cell proteomic analysis revealed that HEK293T cells are depleted of well-established antiviral factors and enriched in proteins that interact with viral protein and RNA, which could explain its high permissiveness to virus infection. Proteome-wide comparative analysis also highlights that a number of proteins exhibit distinct expression patterns across permissive, steady-state, and restrictive cellular states, including genes that have not been previously associated with IFN-α response. Further analysis comparing cellular proteome and transcriptome discloses substantial discordances between protein- and RNA-level responses after IFN-α stimulation. The analysis of RNA-bound proteome (RBPome) shows, for the first time, that IFN-a stimulation alters the protein-RNA interaction landscape in human cells. An in-depth phosphoproteomic analysis of IFN-α stimulated cells provides evidences that IFN-α could regulate the cellular RBPome through phosphorylation of RNA binding proteins. By integrating evidences from multiple levels, I have selected and generated stable expression cell lines for 19 candidate genes to test their potential regulatory roles in viral fitness. Collectively, this study provides an information-rich dataset that examines host determinants of virus-permissive or restrictive cellular states from multiple levels. This could be a valuable source for further analysis and validation studies to discover regulators of virus infection.

Acknowledgements

Firstly, I would like to thank my supervisors Alfredo Castello and Shabaz Mohammed for their unconditional helps throughout my DPhil project, and their patience in supervising and guiding me through the journey of DPhil study. Thanks Alfredo Castello for his remarkable enthusiasm and optimism in science. Every conversation with Alfredo fuels me with the confidence on this project and insights into the biology behind complexed data. Thanks Shabaz Mohammed for his admirable attitude in pursuing high quality data in researches. His supervision always reminds me that a good scientific practice and progressions based on solid data are the prerequisites of science researches. Thank you both for allowing me to explore varies possibilities in this project, and helping me shape it to its current stage.

I would like to thank all the members from Castello lab for their trainings and supports in molecular biology, cell biology, and virology part of my DPhil study. Thanks Marko Noerenberg for his training on tissue culture, handling of SILAC cells, and HIV-1 related experiments. Thanks Aino Jarvelin for her trainings on various types of bioinformatics analyses. Thanks Wael Kamel for his training on RNA interactome capture, and for the insightful discussions and ever-welcoming chats. Thanks Manuel Garcia-Moreno and Vincenzo Ruscica for their advices in tissue cultures.

I would like to thank all the members from Shabaz lab for their trainings and supports in proteomics part of my DPhil study. Thanks Sabrina Liberatori and Svenja Hester from Shabaz lab for trainings on proteomics sample preparation, mass spectrometer operation, and regular maintenance of LC instruments. Thanks Philip Charles for all the discussions regarding biostatistics in proteomic data analysis, his intellectual contributions and collaborations in multi-variable regression and in missing-value imputation have irreplaceable roles in this project.

I would also like to thank people that provided trainings and collaborations in my DPhil. Thanks Richard Parton from Ilan Davis lab for training on the API Delta-Vision Elite microscope operation. Thanks Quan Gu from CVR for his collaboration works in analysis of RNA-sequencing data. Thanks Sam Wilson, David Robertson, and Joseph Hughes from CVR for discussions in this project.

I would like to give my special thanks to my friends and my family. Thanks my new friends in Oxford and old friends stayed in touch for all the chats and happy hours together. Thanks my parents for their selfless supports in finance and in my life. A huge thank to my dear wife Yue Li, for all her supports in our lives. I would not be able to go so far without her supports and pushes.

Thanks to all, sincere,

Honglin Chen

Table of contents

Chapter 1 ..... **7**

Overview **8**

1.1 Viruses hijack cellular resources 8

1.1.1 The alphavirus life cycle and its reliance on host cellular proteins. 8

1.1.2 The lentivirus life cycle and its reliance on host cellular proteins. 11

1.2 Host cells antagonise virus infection through the innate immune response. 14

1.2.1 Pathogen-associated molecular patterns (PAMP) recognition pathways. 15

1.2.2 Interferon (IFN) signalling pathway drives expressions of antiviral factors. 17

1.2.3 IFN-stimulated genes (ISGs) are antiviral factors in host cells 19

1.3 Proteomics analysis based on mass spectrometry 21

1.4 Proteomics expands our understandings of host factors involved in virus infection 25

1.5 Aims 30

Chapter 2 ..... **31**

Introduction **32**

Results **35**

2.1 HEK293T cells are more permissive to virus infection than HEK293 cells 35

2.2 'Single-shot' proteomic analysis tentatively suggests differential enrichment in innate immunity pathway 40

2.3 Deep proteomic profiling unveiled extensive differences in proteome landscapes between HEK293 and HEK293T cells 44

2.4 Deep proteomic analysis revealed enrichment of innate immunity factors in HEK293 cell, and viral protein/RNA interactors in HEK293T cells 47

2.5 IFN-α triggers antiviral responses in HEK293 and HEK293T with different strength 51

Discussion **54**

Acknowledgements **57**

Methods & Materials **57**

Chapter 3 ..... **65**

Introduction **66**

Results **69**

3.1 Serial titration defined time- and dose-dependent kinetics in IFN-α induced antiviral response 69

3.2 Application of deep proteomic profiling with SILAC labelling improved detection of IFN-α stimulated proteome landscape. 72

3.3 Quantitative proteomic analysis reveals proteome level changes at both early- and late-stage of IFN-α stimulation. 81

3.4 Comparative analysis of IFN-α regulated proteome and HEK293/HEK293T proteome highlighted common regulators of virus infection across different cellular states. 87

Discussion 89

Acknowledgements 94

Methods and Materials 94

Chapter 4 101

Introduction 102

Results 105

4.1 Widespread discordance between IFN-α responses in RNA- and protein-levels were revealed by integrative RNA-seq and proteomics. 105

4.2 Exploring regulatory signatures underlying discordance in RNA and protein levels after IFN-α stimulation 111

4.3 Exploring IFN-α induced transcriptome regulations in alternative splicing and expression of long non-coding RNA 117

Discussion 121

Acknowledgements 126

Methods and materials 126

Chapter 5 130

Introduction 131

Results 134

5.1 Deep phosphoproteome profiling revealed wide impact of IFN-α in cellular phosphorylation landscape. 134

5.2 Characterisation of cellular RNA-binding proteome in IFN-α stimulated cells. 140

5.3 Development of novel analysis method that partially impute data and utilise empirical Bayesian estimation to improve robustness of statistical test against missing values. 142

5.4 Comparative analysis with IFN-α proteome showed RNA-binding activity changes independent of virus infection. 147

5.5 Mechanism of IFN-α mediated regulations in RNA-binding activities could involve domain-specific regulation and phosphorylation. 151

Discussion 155

Acknowledgements 158

Methods and Materials 158

Chapter 6 168

Introduction 169

Result 173

6.1 Selection of candidates for follow-up studies using evidence-based scoring 173

6.2 Generation of inducible-expression stable lines for virology assays. 176

Discussion 179

Acknowledgements 182

Method and material 182

Chapter 7 185

Overview 186

7.1 High quality whole-cell proteome data from cells of distinct permissiveness to viruses 186

7.2 Differential proteome landscapes determine cell permissiveness to viruses 187

7.3 Profiling of phosphoproteome and RNA-bound proteome landscapes during IFN-α response 189

7.4 Development of novel imputation method suitable for label-free proteomic analysis . 190

7.5 Integrated multi-level proteomics facilitates selection of proteins that potentially have roles in infection 191

Appendices 195

References 196

Chapter 1

General introduction

Overview

Viruses need to complete their life cycles inside host organisms in order to replicate and propagate. However, due to the restricted size of viral genomes, viruses have to hijack host factors to complete their life cycles and escape host immune response[1-4]. When infected by viruses, human cells contain a diversified set of antiviral factors to combat virus infection[5-7]. Understanding how host factors promote or restrict virus infection is important to identify novel antiviral strategies. Recent advances in mass spectrometry-based proteomics have allowed omics-scale analysis of protein abundance, post-translational modifications, and protein-protein and protein-RNA interactions[8-10]. This chapter will briefly introduce how host factors are hijacked by viruses to facilitate their replication, how antiviral factors interfere with virus infection, and how mass spectrometry technique is applied in proteomic analyses to discover new host-virus interactions in host cells.

1.1 Viruses hijack cellular resources

A common feature shared across viruses is their reliance on host resources to replicate and spread. Genome length and protein coding capacity varies across virus species. Large viruses such as vaccinia virus has 191 kb long double-stranded DNA (dsDNA) genome and encodes over 250 viral proteins[11]. By contrast, smaller viruses like Dengue virus has a 10 kb long positive-sense single-stranded RNA (ss(+)RNA) genome that encode a large polyprotein, which is cleaved into 10 mature viral proteins[12]. Despite the differences in genome size, viruses cannot undergo their life cycle without a host cell. Host proteins participate virtually in all steps of the viral life cycle, from entry, genome replication and protein production, to viral particle assembly and release[1-3]. In this section, I will focus on the host-virus interactions mediated by two classes of viruses – alphavirus and lentivirus. They both belong to RNA virus, but have hugely different life cycles. Alphaviruses are cytoplasmic viruses and they encode their own RNA polymerase for replication; by contrast, lentiviruses replicate in host nucleus and rely on host RNA polymerase. These two classes of viruses are the main research focus in our lab and are model viruses in this thesis, because they are clinically important, and their life cycles involve a plethora of interactions between host and viral RNA/protein.

1.1.1 The alphavirus life cycle and its reliance on host cellular proteins.

The *Alphavirus* is a genus of enveloped ss(+)RNA viruses. Many members of alphavirus can be transmitted to human via mosquito, including Chikungunya virus (CHIKV), Semliki Forest virus (SFV), Sindbis virus (SINV), Ross River virus (RRV) and others[13]. They cause a range of human diseases including rash, fever, arthritis, and encephalitis[14]. The virion is formed by a spherical capsid associated to the 11kb genomic (g)RNA, surrounded by a lipid bilayer membrane that is coated with viral glycoprotein spikes[15] (Fig.1.1.1 A, left). The gRNA contains cistrons regulated by two different promoters. The first promoter produces gRNA which encodes the non-structural proteins (NSPs) necessary for viral RNA replication. The second promoter produces subgenomic (sg)RNA which encodes structural proteins that form capsid and glycoproteins[15] (Fig.1.1.1 A, right). The life cycle of alphavirus begins with receptor recognition by the viral glycoproteins, for example SINV glycoprotein spike recognises laminin receptor on human cells[16]. This binding induces endocytosis of the viral particle[17]. The low pH environment in mature endosome triggers fusion between viral and endosomal membrane and entry of viral nucleocapsid into the cellular cytosol[18]. The viral genomic RNA is translated in the cytosol to produce the non-structural proteins (NSPs), which assemble into the replication complex on the plasma membrane forming the replication spherules. Through endocytosis, multiple replication spherules are assembled into a cytopathic vacuoles type I (CPV-I), a viral RNA replication factory that mass produces gRNA and sgRNAs[19]. The sgRNA is translated into structural proteins, including the Capsid and glycoproteins. The gRNA and Capsid proteins are assembled into new nucleocapsids, which attach to secretory vesicle that contain viral glycoproteins, forming cytopathic vacuoles type II (CPV-II)[20]. Upon delivery to plasma membrane, the CPV-II fuses with membrane, and new viral particles are released through budding[21] (Fig.1.1.1 B).

Host cellular proteins are widely involved in the alphavirus life cycle. This can be exemplified by the formations of viral CPV-I and CPV-II. CPV-I is composed of virus replication spherules, which are semi-enclosed environments on plasma membrane. The viral RNA template is enveloped within the spherule, which is gated by a ring-like structure formed by viral protein NSP1[22]. The RNA-dependent RNA polymerase complex formed by NSP2/3/4 catalyses replication of viral RNA. The YXXM motif in NSP3 is reported to bind p85, a regulatory subunit of phosphatidylinositol 3-kinase (PI3K)[23]. PI3K is then activated by releasing its catalytic subunit p110 from p85, leading to phosphorylation of phosphatidylinositol 4,5-bisphosphate (PIP2), and subsequent activation of Akt kinase[24,25]. Through local recruitment and activation of PI3K-Akt pathway, the alphavirus utilises this pathway to internalise the replication spherules on plasma membrane, forming CPV-I (Fig.1.1.1 C). Another example is the formation and trafficking of CPV-II. Matured viral glycoproteins on Golgi apparatus need to be packed into vesicle to reach the plasma membrane. A siRNA-knockdown screening study shows that Arf1 is a key regulator of this process[26]. Arf1 is a small GTPase that

regulates formation of coat protein complex I (COP I) vesicles in secretory pathway[27]. Therefore alphavirus could exploit Arf1-regulated secretory pathway to form CPV-II (Fig.1.1.1 D). The knockdown screening also showed that actin-remodelling factors Arp3, Rac1, and PIP5K1-α are essential for CPV-II trafficking[26]. Therefore, it is hypothesised that transportation of CPV-II to plasma membrane could be dependent on actin filament elongation[28].

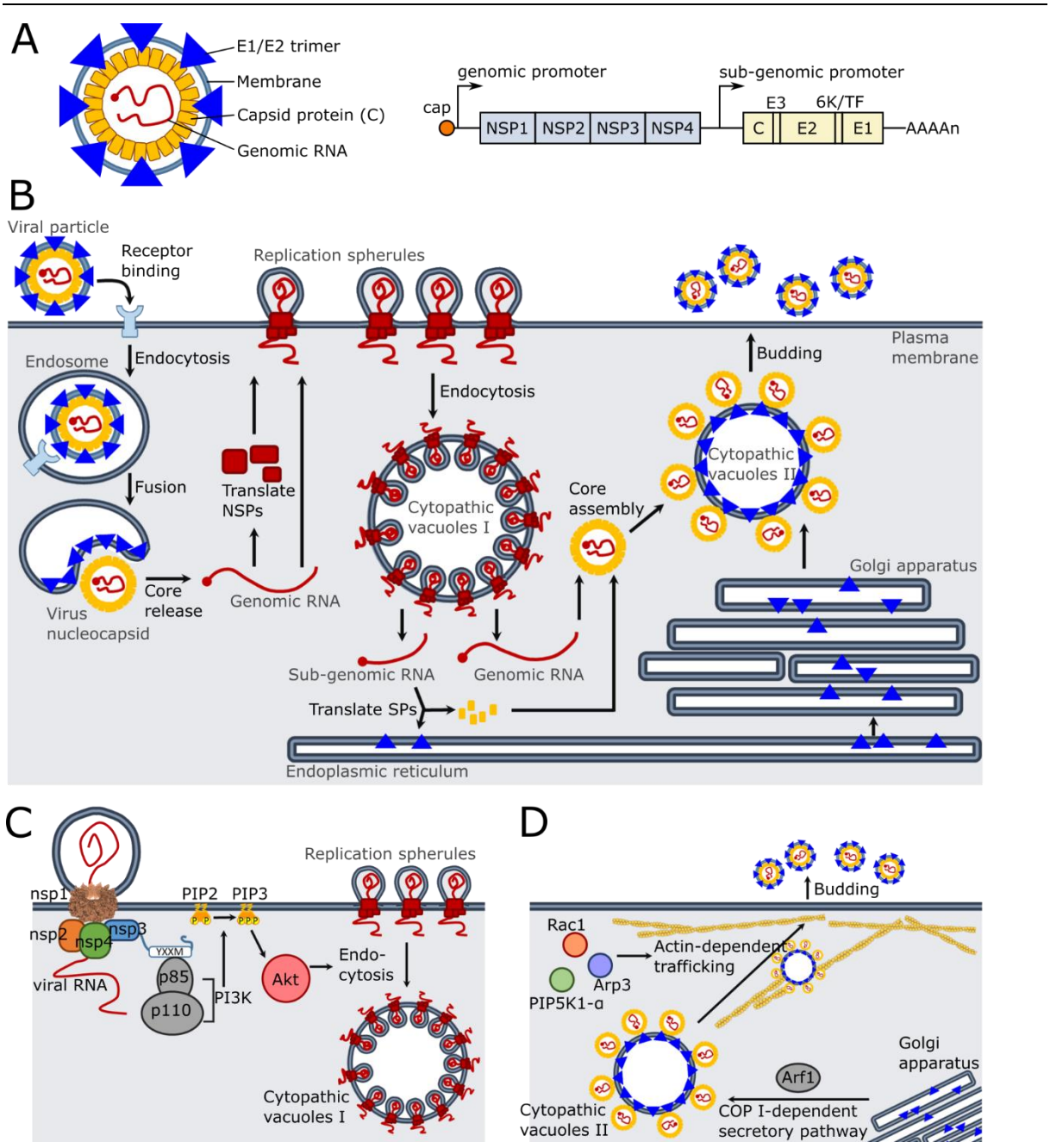


Figure.1.1.1 Alphavirus utilise host resources during life cycle. (**A**), schematics of alphavirus mature viral particle and genome. (**B**), schematic life cycle of alphavirus. (**C**), host factors involved in formation of CPV-I. The nsp3 in viral replication complex activates PI3K, and induce endocytosis

through PI3K-AKT pathway[23,25]. CPV-I is formed by endocytosis of multiple replication spherules. (**D**), host factors involved in trafficking of CPV-II. Arf1 is a regulator of coat protein complex I (COP I) vesicle formation, and has a central role in packing viral glycoprotein into vesicles from Golgi apparatus[27]. This vesicle and viral nucleocapsid form CPV-II, which is trafficked to plasma membrane in actin-dependent process[28].

1.1.2 The lentivirus life cycle and its reliance on host cellular proteins.

The *Lentivirus* genus belongs to *Retroviridae* family, which is characterised by the reverse transcription of the ss(+)RNA genome into double-stranded DNA (dsDNA) during its life cycle[29]. One typical example of Lentivirus is the human immunodeficiency virus type 1 (HIV-1), the causal agent of acquired immune deficiency syndrome (AIDS). It is affecting over 38 million people, and is responsible for death of 650,000 people in year 2021[30]. The mature HIV-1 particle contains two gRNA molecules within its cone-shaped capsid[31], which is surrounded by a spherical shell formed by viral protein Matrix and a lipid bilayer membrane that contain viral glycoproteins[31] (Fig.1.1.2 A). The HIV-1 genome is a capped and poly(A)-tailed ss(+)RNA. Its coding region is flanked by long terminal repeat (LTR) sequences that induce transcription initiation and termination at 5'- and 3'-end, and it contains multiple overlapping open reading frames. HIV-1 viral RNA has diverse splicing isoforms, each is able to express different viral protein (Fig.1.1.2 B). The HIV-1 life cycle begins with membrane fusion induced by the interaction between HIV-1 glycoprotein spike and the receptor CD4-CXCR4/ CCR5[32] (Fig.1.1.2 C). Viral capsid is released to host cytoplasm after fusion, and transported toward nuclear pore, disassembling in the nucleus[33]. The reverse transcribed viral dsDNA for integration into the host chromosome by the viral integrase. After integration, the viral DNA is packed with nucleosomes which blocks the viral promoter[34]. To initiate gRNA translation, HIV-1 early viral protein Tat hijacks the host chromatin remodelling complex SWI/SNF to rearrange nucleosome positions, therefore increase accessibility of transcription factors to viral promoter region[35] (Fig.1.1.2 D). After nucleosome relocating, the HIV-1 viral promoter still cannot efficiently transcribe viral RNA on its own. Most of its products are short RNA strands of the first 50 nucleotides, which contain a transactivation-responsive region (TAR) that is rich in secondary structure [36]. Tat can bind to TAR of nascent viral transcripts and recruit transcription elongation factor complex P-TEFb. This complex hyperphosphorylates the C-terminal domain of RNA polymerase II, which drives transcription elongation[37,38] (Fig.1.1.2 D).

The transcribed HIV-1 viral RNA has multiple splicing isoforms[1,39]. Genomic viral RNA and partially-spliced viral RNA contain introns and are trapped in the nucleus. Viral protein Rev can

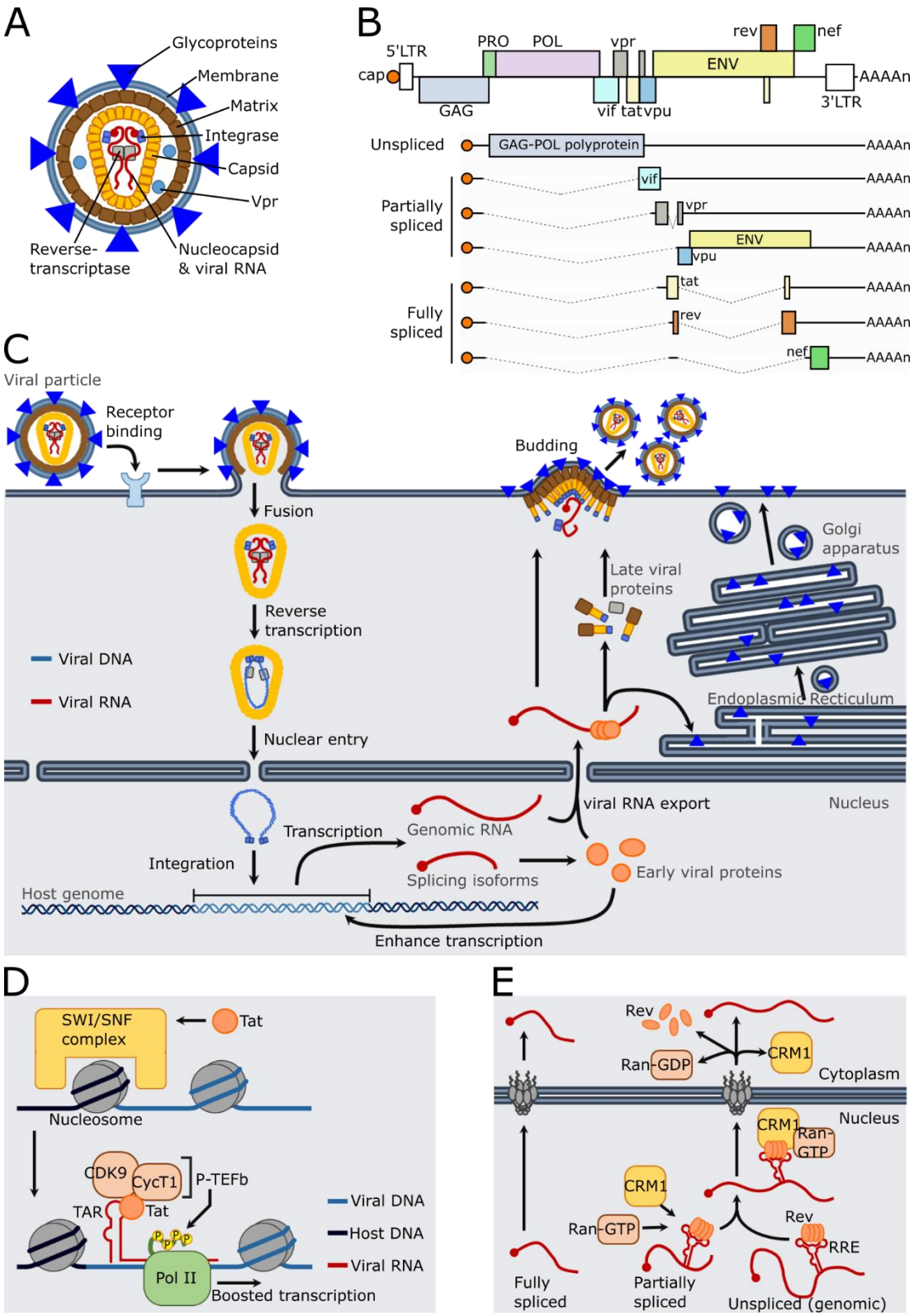


Figure.1.1.2 Retrovirus utilise host resources during life cycle. (**A**), schematics of mature HIV-1 viral particle. (**B**), schematics of HIV-1 genome and splicing isoforms of its RNA. (**C**), schematic life cycle of HIV-1. (**D**). host factors involved in transcription of HIV-1 viral RNA. Early viral protein Tat recruit

chromatin-remodelling complex SWI/SNP to expose transcription start site of integrated HIV-1 genome. Tat also binds to transactivation-responsive region (TAR) of nascent RNA, recruits P-TEFb complex to boost transcription initiation via hyperphosphorylation of RNA pol II C-terminal domain. (**E**), host factors involved in nuclear export of HIV-1 viral RNA. Partially-spliced and unspliced viral RNA contain Rev-responsive element (RRE) sequence. Rev binds to RRE, and recruits nuclear export complex members CRM1 and Ran-GTP to transport RRE-containing viral RNA to cytoplasm.

---

facilitate nuclear export of these viral RNA splicing isoforms exploiting the host RNA export mechanism[40]. Genomic and partially-spliced viral RNAs contain a secondary structure-rich sequence known as Rev-responsive element (RRE). In nucleus, the Rev protein binds and oligomerises on RRE[41]. The RRE-bound Rev then recruits exportin-1 (CRM1) and GTPase Ran-GTP to form the nuclear export complex, therefore exploit this host nuclear export pathway to translocate RRE-containing viral RNA[41,42]. Once transported into cytoplasm, Ran-GTP hydrolyse its bound GTP into GDP. This process destabilises the complex formed by RRE, Rev, CRM1 and Ran GTPase, leading to release of viral RNA into cytoplasm[43] (Fig.1.1.2 E). Partially-spliced and unspliced viral RNA produce the viral structural proteins which include the glycoproteins (gp120 and gp41) and the polyprotein Gag (comprises Matrix, capsid and nucleocapsid), and viral enzymes Pol (comprises protease, integrase and reverse transcriptase). Viral structural proteins recruits the unspliced gRNA to form new viral particles at the plasma membrane[44] (Fig.1.1.2 C).

Apart from hijacking host machineries for progression of virus life cycle, interactions between host and HIV-1 viral proteins are also involved in its immune evasion[45]. APOBEC family is a protein family of RNA editing enzymes, among which APOBEC3G and APOBEC3F can induce lethal hypermutations in HIV-1 genome[46,47]. Viral protein Vif can interfere this antiviral process through inducing poly-ubiquitination, and subsequently degradation of APOBEC3G[48]. Tetherin (also BST2) is a transmembrane protein that contains a single-pass transmembrane domain on N-terminus, a coiled-coil domain in the middle, and a glycophosphatidylinositol anchor on C-terminus. It can restrict viral particle production by anchoring both membranes of nascent virion and parent cells, therefore preventing virus budding[49]. To antagonise this process, viral protein Vpu binds to tetherin transmembrane domain and directs it to proteasome-dependent ERAD (ER-associated degradation) pathway or lysosome[50,51]. Another viral protein Nef can also antagonise tetherin function, it also down-modulate antigen presenting pathway by routing CD4 and MHC-I to endocytic pathway and lysosome for degradation[52].

1.2 Host cells antagonise virus infection through the innate immune response.

Almost all human cells have an antiviral innate immunity system which can detects virus-derived molecules and activates the expressions of an arsenal of proteins that antagonise virus infection[5-7] (Fig.1.2.1). Many virus-derived molecules contain pathogen-associated molecular patterns (PAMPs) that can be detected by sensors in the host cell[5]. Sensing of virus molecules initiates a series of signalling cascade and activates transcription factors that drive the expression of interferon (IFN)[5]. IFN is a family of cytokines that is secreted and binds IFN receptors on the producer cell as well as neighbouring cells. The binding of IFN to its receptor activates another signalling cascade that drives the expression of interferon-stimulated genes (ISGs, Fig.1.2.1)[6,7]. A large number of ISGs are known to be antiviral factors that directly antagonise virus infection[7,53]. The innate immune response is present in almost all somatic cells as our first line of defence against viruses. In this section, I will expand on the antiviral innate immune response in human cells, including PAMP recognition pathway, IFN signalling pathway, and antiviral effects exerted by ISGs.

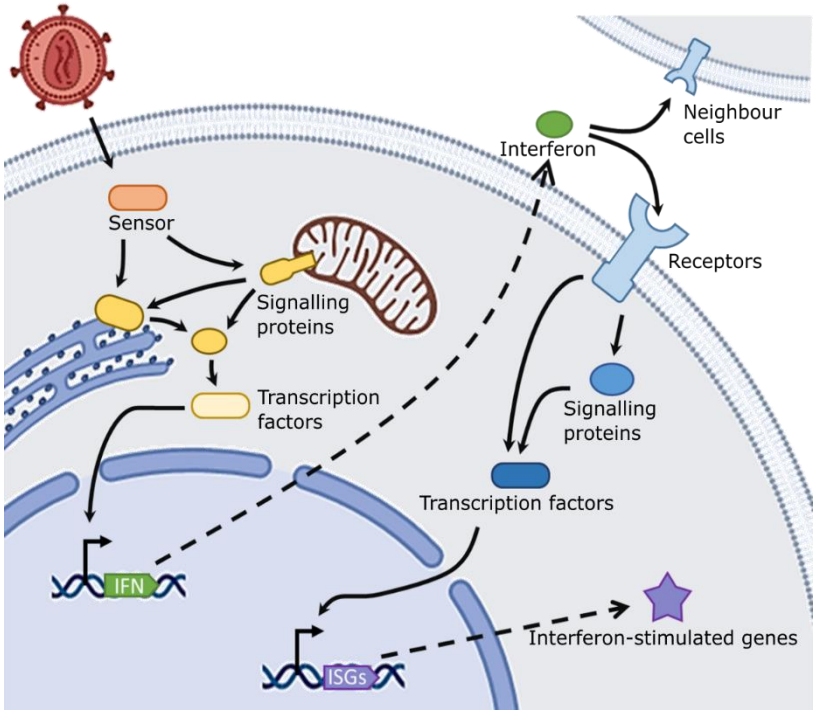


Figure.1.2.1 Schematic of the overall view of innate immune response against virus.

1.2.1 Pathogen-associated molecular patterns (PAMP) recognition pathways.

PAMP recognition pathway alarms the cell when it is infected by a pathogen. A wide range of pathogen-derived molecules contain PAMPs that can be recognised by the host, including modified lipids, double stranded (ds)RNA, 5'-triphosphylated single stranded RNA, foreign DNA, and other unusual molecules[54]. Different types of PAMPs are recognised by specialised sensors in the cell. Toll-like receptor (TLR) is a large family of transmembrane protein that recognise various types of PAMPs through their extracellular domains, and activate downstream signalling events via their intercellular domain (Fig.1.2.2)[55]. TLR1/2/4/6 are located on cell surface membrane, they can sense pathogen-specific lipids such as lipopolysaccharide and lipopeptide[56,57]. TLR3/7/9 are located on endosomal membrane, recognising viral nucleic acids that are released during endocytosis[58-60]. Ligand-bound TLRs activates TIR-domain-containing signalling factors such as MyD88 and TRIF[61]. Their downstream signalling cascades lead to expression of inflammation cytokines through transcription factor NF-κB, and expression of IFN through transcription factors IRF3/7[62].

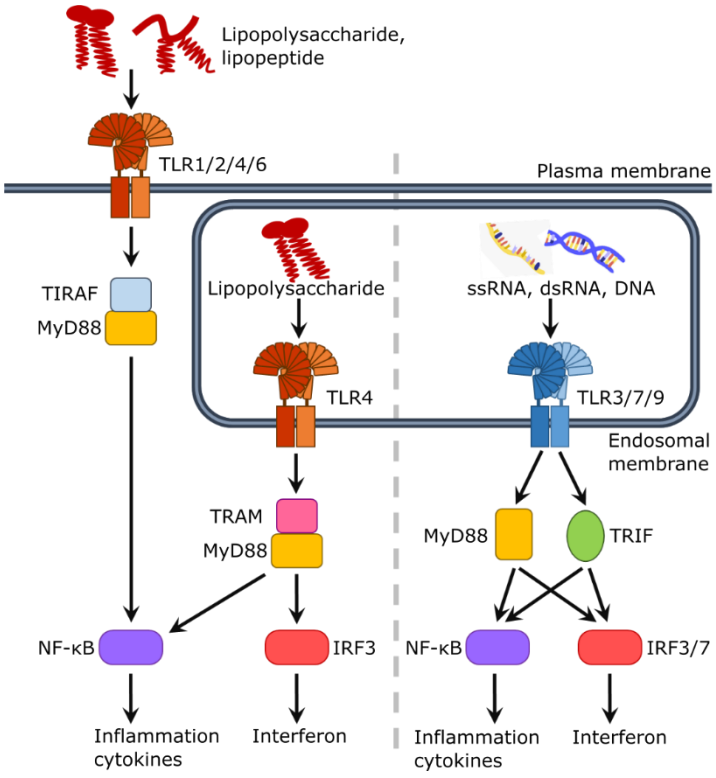


Figure.1.2.2 Toll-like receptor (TLR) based PAMP recognition signalling pathways. Lipid-based PAMPs are recognised by TLR1/2/4/6 mainly on plasma membrane, leading to activation of NF-κB and inflammation response. Pathogenic nucleic acids are recognised by TLR3/7/9 in endosome, leading to production of interferon and inflammation cytokines.

Virus-derived nucleic acids in the cytoplasm are recognised by a variety of PAMP-sensing factors. MDA5 (also IFIH1) and RIG-I (also DDX58) are two most well understood cytoplasmic sensors of dsRNA. They share high sequence similarity and a common DExD/H-box helicase domain, but they have different substrate preference[63]. MDA5 preferentially binds to internal regions of long dsRNA (>1000 bp), while RIG-I prefers short dsRNA (<300 bp) with blunt ends and 5'-triphosphate ends[64,65]. Both RNA sensors activate the signalling protein MAVS, promoting its oligomerisation on mitochondrial membrane[66]. This induces a TRAF-family protein-dependent signalling cascade, and results in production of proinflammatory cytokines and IFNs through activation of the transcription factors NF-κB and IRF3/7 (Fig.1.2.3 A)[67,68]. Sensing of cytosolic viral DNA by cGAS-STING signalling pathway has been well-characterised[69]. Upon recognition of viral DNA in cytoplasm, cGAS catalyses production of the second messenger cyclic GMP-AMP (2'3' or 3'5'-cGAMP)[70,71]. These molecules then bind to signalling protein STING on the endoplasmic reticulum (ER) membrane, and subsequently activate transcription factors IRF3 and NF-κB for production of IFNs and inflammation cytokines (Fig.1.2.3 B)[72-74].

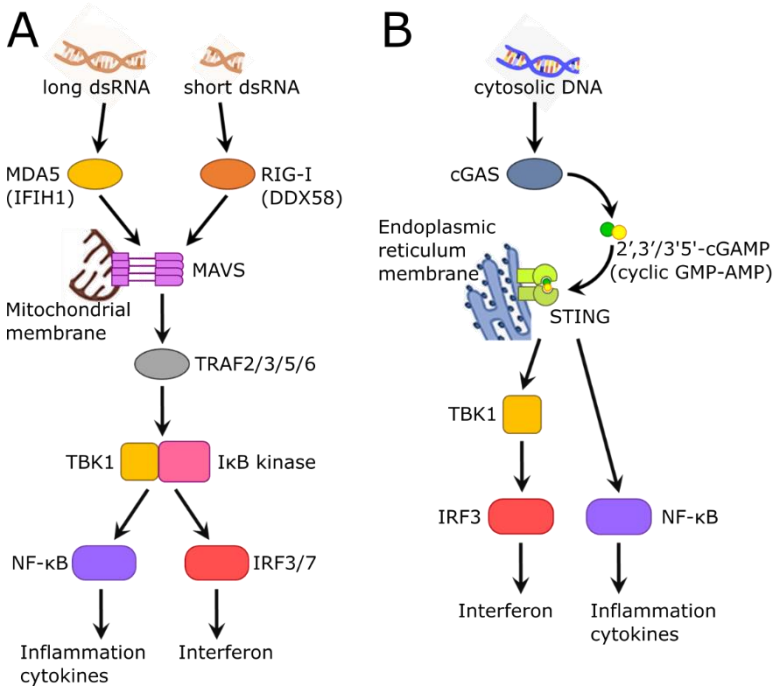


Figure.1.2.3 PAMP recognition signalling pathways for cytoplasmic nucleic acid. (**A**), cytoplasmic RNA sensing by RIG-I and MDA5 initiates MAVS-dependent signalling pathway that induce expression of interferon and other inflammation cytokines. (**B**), cytoplasmic DNA sensing by cGAS-STING signalling pathway that induce expression of interferon and other inflammation cytokines.

Apart from TLRs, MDA5, RIG-I, and c-GAS, virus-derived molecules are also recognised by many other cellular sensors[54,75]. In recent years, an increasing number of DExD/H box helicase (DDX) family members are identified to be sensor of viral RNA. For example, DDX3 can directly interact with viral RNA and activate IRF3 for production of IFN[76,77]; DDX60 also binds to viral RNA and is essential for RIG-I-mediated signal transduction[78]. ZBP1 (also DAI or DLM1) is characterised as a sensor of viral DNA. It can recognise left-handed DNA helix (Z-DNA) in cytoplasm and induce IFN production via kinase TBK1[79]. IFI16 is a multi-functional nucleic acid sensor that can recognise viral DNA in the nucleus and induce innate immune response through cGAS-STING pathway[80,81], it can also recognise viral RNA in the cytoplasm and upregulate RIG-I expression[82]. For these host sensors, a lot remain unknown regarding details of their mechanisms in virus recognition and signalling transduction. In addition, whether there are more currently unknown sensors that can detect virus-derived molecules remains an open question for further studies.

1.2.2 Interferon (IFN) signalling pathway drives expressions of antiviral factors.

Detection of virus infection generally results in production and secretion of IFNs. This is a class of cytokine that induces cellular response to virus. There are three types of IFN in human, type I to III. Each type of IFN is recognised by a different tyrosine kinase receptor complex (Fig.1.2.4 A). Type I IFN binds to IFNAR1/2 heterodimer. Almost all somatic cells can produce and recognise type I IFN when facing virus infection, while notably, most of the type I IFN *in vivo* are produced by dendritic cells[83]. Type II IFN is produced only by immune system cells, but can be recognised by almost all cell types through heterodimeric receptor IFNGR1/2[84]. Type III IFN is recognised by IL-10Rβ and IFNLR1 heterodimer receptor. It is newly discovered in the 21st century, initially found to act on epithelial cells, now known to have antiviral effect for a wide range of cells[85,86]. Upon receptor binding, all three types of IFN transduce signals mainly through the JAK-STAT signalling pathway, which is also known as the classical IFN signalling pathway (Fig.1.2.4 A). In type I IFN induced signalling, IFN-bound receptors activates kinases TYK2 and JAK1, which subsequently mediate tyrosine phosphorylation of STAT family proteins[87]. STAT family proteins are transcription factors that form homo/hetero-dimer upon activation[88]. The activated STAT dimer recruits IRF9 and bind to IFN stimulated response element (ISRE), a DNA sequence in IFN-responsive promoter that drives the expression of ISGs[6]. Type II IFN, when binds to its receptor, activates kinases JAK1/2 and induce tyrosine phosphorylation of STAT1. Activated STAT1 forms homodimer and drives ISG expressions via promoters that contain IFN-gamma activated site (GAS)[89]. Type III IFN share a similar signalling

cascade as type I IFN, and is currently believed to act as a front-line local defence that minimises detrimental inflammation response[86].

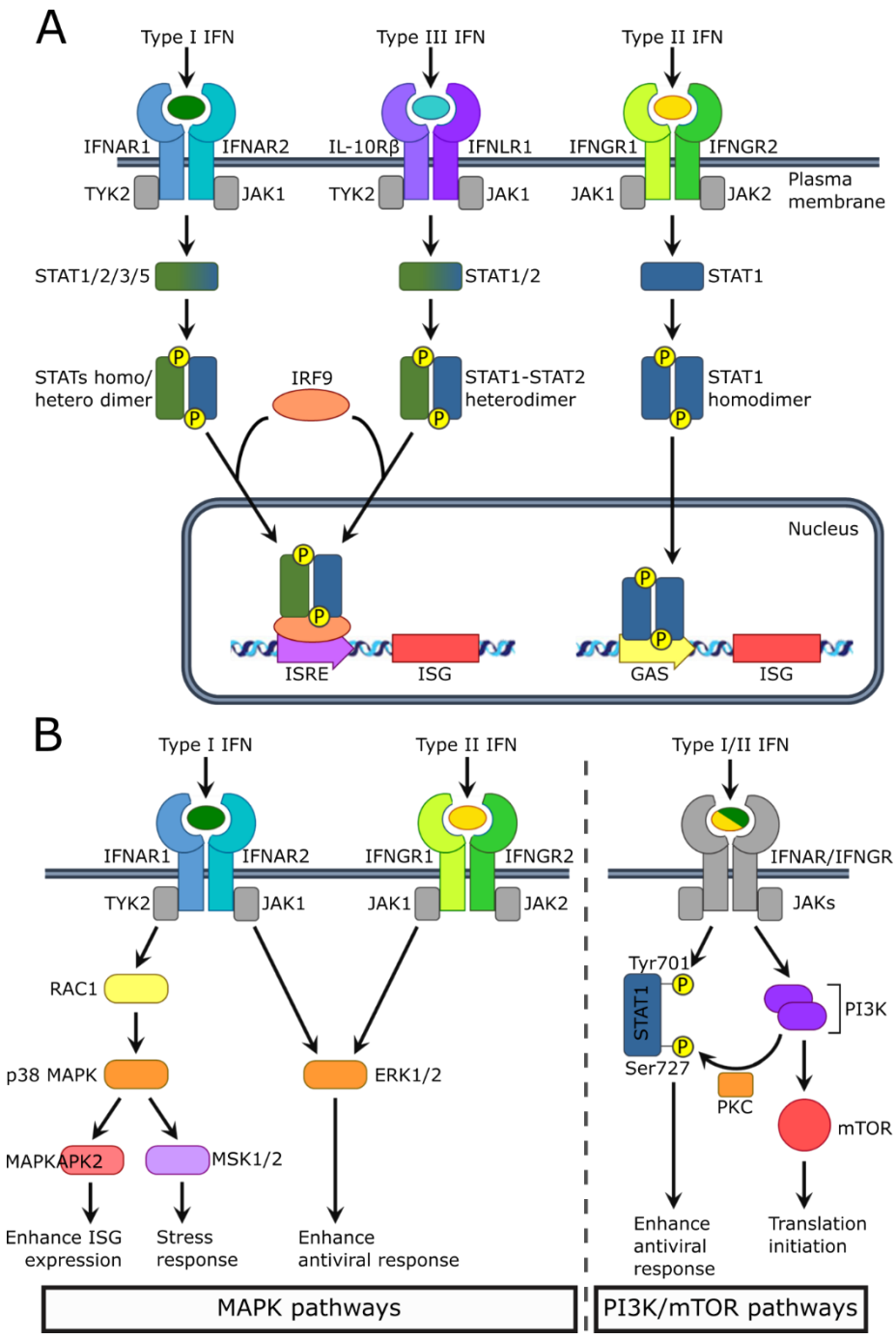


Figure.1.2.4 IFN activated signalling pathways. (**A**), overview of classical IFN signalling pathways. Receptor recognition of IFNs activate transcription factor family STATs and drive expression of interferon-stimulated genes (ISGs) through activation of promoters that contain IFN stimulated response element (ISRE) and IFN-gamma activated site (GAS). (**B**), overview of non-classical IFN

signalling pathways. Type I IFN activates MAPK pathway to induce stress response. Both type I and II IFN activate MAPK and PI3K/Akt/mTOR pathways to enhance antiviral response.

IFN can also activate a number of non-classical IFN signalling pathways (Fig.1.2.4 B)[6,90]. Type I IFN can activate Mitogen-activated protein kinases (MAPKs). There are 3 main groups of MAPKs: p38 family, extracellular-signal-regulated kinase (ERK) family, and JUN amino-terminal kinase (JNK) family[91]. Activation of p38 is important for ISRE promoter expression in type I IFN signalling[92]. This is mediated by small GTPase RAC1, and its downstream effects include enhancement of ISG expressions via MAPKAPK2, and stress response through stress-activated kinases MSK1/2[93,94]. ERK1/2 also have important role in enhancing ISG expressions in type I IFN signalling, but the details in this pathway remain unclear[6]. JNK has not been related to type I IFN. Type II IFN is also linked to MAPKs, but only through ERK1/2[95]. Both type I and type II IFN are reported to regulate the PI3K/Akt/mTOR signalling pathway. Activated PI3K induces phosphorylation of serine 727 on STAT1 via protein kinase C (PKC)[96,97], which is crucial for effective antiviral response[98]. Type I and II IFN can also activate the mammalian target of rapamycin (mTOR) downstream of PI3K, stimulating translation initiation[99,100].

1.2.3 IFN-stimulated genes (ISGs) are antiviral factors in host cells

IFN signalling pathway activates transcription of hundreds of ISGs. These ISGs are mainly comprised of PAMP-recognising factors and IFN signalling factors that enhance innate immune response, as well as other type of antiviral factors that directly restrict virus infection[53,101]. The dsRNA-dependent kinase PKR (also EIF2AK2) is one of the well characterised antiviral factors[102] (Fig.1.2.5 A). PKR is inactivated by its RNA-binding domain RRMs in steady state. Upon binding to viral RNA, the auto-inhibition by RRM domains is released, and PKR is activated through auto-phosphorylation and dimerisation[103]. The activated PKR phosphorylate translation initiation factor EIF2A, preventing the translation of viral proteins[104]. The 2'-5'-oligoadenylate synthase (OAS) family proteins are also strongly induced by IFN and confer antiviral activity against a wide range of viruses[105]. OASs are inactive until they binds to viral RNA, and can synthesize 2'-5'-oligoadenylate as second messenger[106]. The endoribonuclease RNaseL is activated by 2'-5'-oligoadenylate, and exert antiviral activity through cleavage of viral RNA[107,108] (Fig.1.2.5 B). The ubiquitin-like protein ISG15 is one of the most highly expressed ISGs after IFN stimulation[109]. Similar to ubiquitin, ISG15 can be covalently linked to substrate proteins, known as ISGylation, through ISG15-specific E1/E2/E3

conjugation enzymes[109,110] (Fig.1.2.5 A). Currently known E1 activating enzyme for ISGylation is UBE1L, which passes the activated ISG15 to E2 enzyme UBE2L6 and subsequently E3 ligases HERC5 and TRIM25, and ISGylation can be hydrolysed by USP18[111-115]. A diverse range of proteins can be ISGylated[109]. Unlike ubiquitin, conjugation of ISG15 does not cause protein degradation. The exact antiviral mechanism of ISG15 differs depending on substrate proteins. Signalling factors in PAMP-sensing and IFN signalling pathways are found to be ISGylated, including RIG-I and STAT1. These ISGylation events can either suppress innate immune response by inducing phagosome-dependent degradation[116], or enhance innate immunity by competing with ubiquitin-dependent degradation[117]. E3 ligase NEDD4 is essential for budding of enveloped viruses. ISGylation of NEDD4 inhibits its function, therefore inhibits virus egress[118]. Viral protein NS1 from Influenza A virus inhibits innate immune response, and its ISGylation abolishes the inhibitory effect[119]. Despite having these reports, antiviral mechanisms for the majority of ISGylated proteins are still largely unknown. For example, HERC5 is associated with polyribosomes and mediates ISGylation of newly synthesised viral proteins, but little is known regarding its downstream antiviral mechanism[120].

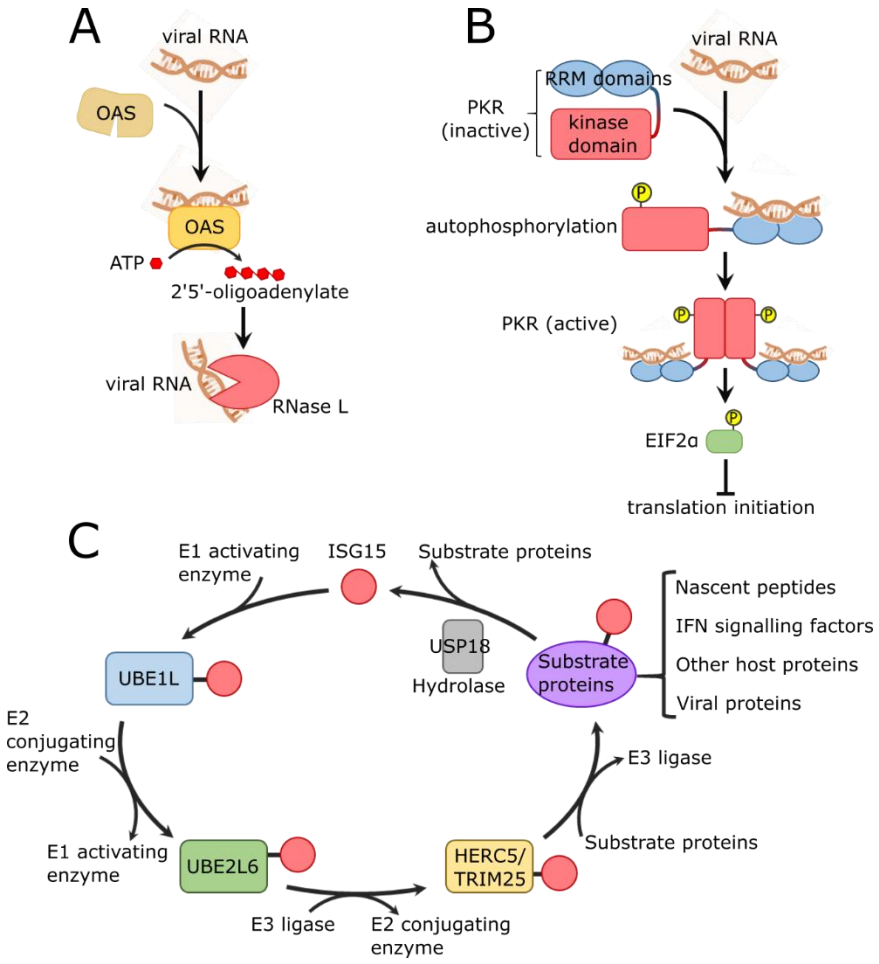


Figure.1.2.5 Examples of antiviral processes mediated by ISGs. (**A**), OAS-mediated antiviral process.

OAS catalyse ATP into 2'-5'-oligoadenylate upon binding to viral RNA. Oligoadenylate activates RNase L for viral RNA cleavage. (**B**), PKR-mediated antiviral process. Viral RNA is recognised by and activates PKR, which autophosphorylate and dimerise. The activated PKR dimer phosphorylates EIF2α and inhibits translation initiation of viral proteins. (**C**) ISG15-mediated antiviral process. Ubiquitin-like protein ISG15 is covalently linked to various types of substrate proteins, involving newly synthesised proteins, IFN signalling factors, host proteins in other processes, and viral proteins. ISG15 modification can be hydrolysed by host hydrolase.

Various other ISGs also act as antiviral effectors during innate immune response. GTPase MX family proteins are broad-spectra antiviral factors[121]. They are reported to form oligomers near Golgi apparatus and endoplasmic reticulum, and probably antagonise virus at early stage of infection through trapping viral nucleocapsid[122-124]. However their exact antiviral mechanism remain unclear. TRIM family proteins have multifaceted roles in human cells and many are responsive to IFN stimulation[125]. Members of TRIM family have diverse roles in antiviral response. For example, TRIM25 is an E3 ligase for ISG15[112], TRIM5α binds and disassemble HIV-1 nucleocapsid[126]. RNA binding proteins ADAR and APOBEC family proteins introduce hyper-mutations in viral genomes, leading to generation of replication-defective virus offsprings[127,128]. Besides ISGs introduced above, many ISGs also contribute to the host's antiviral response[101]. Whether there are more antiviral ISGs to be discovered, and the mechanisms how ISGs restrict virus infection are all valuable questions for further researches.

1.3 Proteomics analysis based on mass spectrometry

Proteomics based on mass spectrometry (MS) is a powerful technology that allows simultaneous identification of multiple proteins in the studied biological system. A key technique in MS-based proteomics is protein identification in millisecond time scale by tandem MS (MS/MS)[129,130]. In MS/MS analysis, proteins are first digested into peptides with protease, typically trypsin protease. Identities of target proteins can be determined by their tryptic peptides because of the high complexity of amino acid compositions. Peptide products are transformed into ions in gas phase by the ionisation process, commonly through electrospray ionisation (ESI) or matrix-assisted laser desorption/ionisation (MALDI)[131,132]. Peptide ions then enter the mass spectrometer which uses electromagnetic fields to distinguish ion species based on their mass-to-charge (m/z) ratios. This process generates an m/z spectrum of all the peptides (also known as the MS1 spectrum) (Fig.1.3.1

A). After obtaining the MS1 spectrum, peptide ions of particular m/z ratio (also precursor ion) are isolated and fragmented, which is typically achieved by collision with neutral inert gas molecules after acceleration (collision induced disassociation (CID))[133,134]. Fragmentation generally takes place on the amide bond between amino acid residues, resulting in ions that retain information about the amino acid sequence (b-ions for N-terminal fragments, and y-ions for C-termial fragments). The mass analyser then generate an m/z spectrum of all fragment ions for this precursor ion, which is referred to as MS2 spectrum. Amino acid sequence of the precursor peptide can be inferred by m/z ratios of its fragment ions[130]. Composition of the peptide mixture can be determined by sampling through peptide ions identified in the MS1 for fragmentation and MS2 spectrum generation (Fig.1.3.1 A). These peptide sequences are subsequently matched to proteome databases to determine the identity of the analysed proteins[135,136].

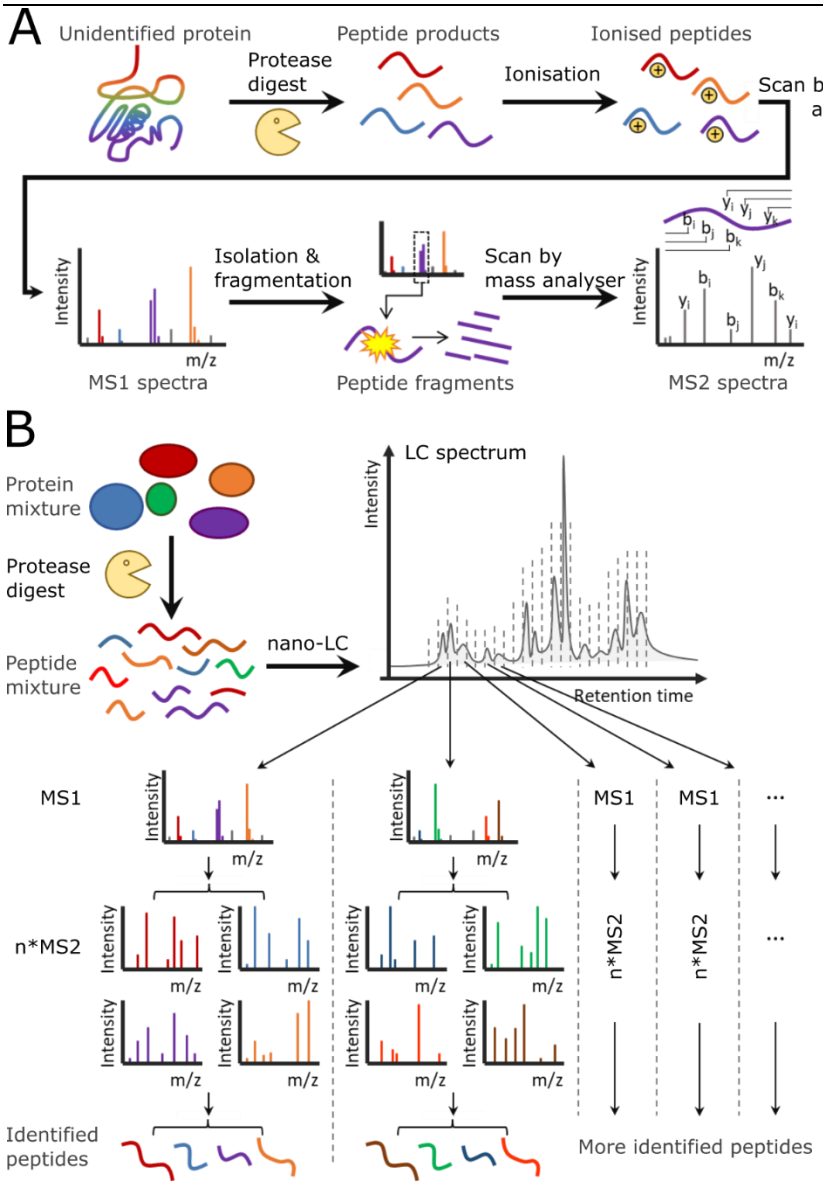


Figure.1.3.1 Schematics of principles in mass spectrometry (MS)-based proteomic analysis. (**A**)

protein identification through tandem MS (MS/MS). Proteins are digested into peptides, which are then charged by ionisation. Peptide ions are analysed by mass analyser to generate MS1 spectrum. Peptides ions identified in MS1 spectra are isolated and fragmented for sequencing in MS2. The $y_{i,j,k}$ and $b_{i,j,k}$ represent ions from fragmented peptide. (**B**) analysis of complex protein mixture by LC-MS/MS. Peptides from digested protein mixture are separated by liquid chromatogram (LC). Peptide species in LC-separated samples are identified by MS/MS.

Proteomic analysis of biological samples, such as pulldown eluates or lysates, faces three major challenges: sample complexity, dynamic range, and detection limit[8]. The sample complexity issue is evidenced by the large amount of protein species in the sample – a pulldown eluate typically contains hundreds of proteins, which yield over 10,000 unique peptides after trypsin protease digestion. Therefore, a comprehensive protein identification requires a large number of MS2 spectra to sequence these peptides. The dynamic range issue arises from the fact that protein abundance range in a biological sample generally spans several orders of magnitude – the abundance of a highly expressed housekeeping protein can be seven orders of magnitudes higher than low-abundant proteins[137]. Therefore, peptides from low-abundant proteins have much lower chance to be detected. The detection limit issue stems from how mass spectrometers control the trajectory of ions. For mass analysers used in this study (linear ion trap and Orbitrap), peptide ions are controlled by electromagnetic fields, which distinguish ions based on their m/z ratios. However, peptide ions can repulse each other, resulting in interfered ion trajectories and incorrect m/z measurement. As a consequence, only a limited amount of peptide ions can be analysed by mass spectrometer in each scan, therefore complexity and dynamic range issues cannot be resolved by simply increasing the amount of input material.

One effective solution in proteomic analysis of complex sample is to perform MS/MS on peptide mixtures that are separated by liquid chromatography (also referred to as LC-MS/MS, Fig.1.3.1 B)[129]. The most common type of separation is the reverse-phase chromatography which separate peptides based on hydrophobicity[138]. In reverse-phase chromatography, C18-coated particles are packed into the column and form an organic stationary phase. The peptide mixture are loaded on the column in aqueous solution, where peptides are retained in the organic stationary phase. Peptides are then gradually eluted with solutions of increasing concentration of organic solvent. The resolution of peptide separation is mainly determined by C18 particle size, flow rate, and column length. By using column packed with C18-coated 3-µm porous particles in nanolitre flow rate, Shen et al. have described an reverse phase chromatography can separate more than 1,000

peptide peaks in one retention window[139]. The ultra-high performance liquid chromatography (UHPLC) developed with similar configurations permits extensive peptide separation prior to MS/MS analysis. Eluates of UHPLC are typically connected to an electrospray ioniser where a high voltage is applied to evaporate each droplet of eluate, ionise peptides within the droplet, and inject peptide ions into the mass spectrometer[132]. This technique massively reduces the complexity of the peptide mixture that mass spectrometer analyse. By repeating MS/MS cycles that each comprises one MS1 followed by 10~20 MS2 events throughout the UHPLC elution, LC-MS/MS can achieve comprehensive identification of peptide species in a complex sample (Fig.1.3.1 B).

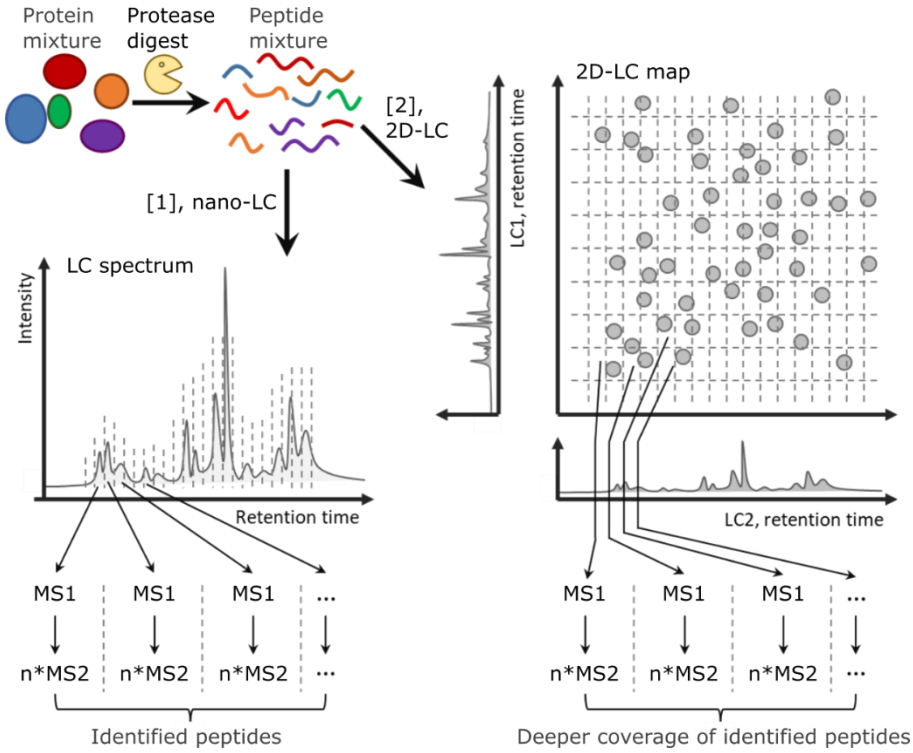


Figure.1.3.2 Schematics of principles in deep proteomic analysis. [**1**], in regular LC-MS/MS analysis, the digested peptide mixture is separated by nano-LC, and separated sample is analysed by MS/MS. [**2**], in deep proteomic analysis, peptide mixture is separated by two LC systems that separate peptides in distinct manner. The more thoroughly separated sample is analysed by MS/MS, allowing deeper coverage in peptide identification.

LC-MS/MS analysis with a single UHPLC gradient has proven a successful technique for proteomic analysis of samples like pulldown eluates, which typically contain a few hundred proteins. However, it remains insufficient for a comprehensive analysis of the whole cell proteome of human

cells. The Human genome contains ~20,000 genes, each could produce multiple protein isoforms and have varies post-translational modifications (PTMs)[140]. Therefore, a whole-cell lysate could yield millions of unique peptides after trypsin digestion. Moreover, a substantial proportion of proteins have expression levels that are several orders of magnitudes lower than housekeeping proteins, rendering them difficult to be sampled in MS/MS cycles. Increasing LC gradient length can improve the depth of detection, however, the LC gradient cannot be extended indefinitely because prolonged LC analysis time risks peak dispersion in the column. Combining "orthogonal" chromatography separation methods prior to MS/MS analysis has proven a powerful method to further reduce sample complexity in proteomic analysis (Fig.1.3.2)[141,142]. These chromatography separations exploit different physicochemical properties of the peptide, for example reverse phase chromatography separates peptides based on hydrophobicity, ion exchange chromatography separates based on charge state, size exclusion chromatography separates by molecular weight, etc.[142]. An appropriate choice of LC systems that has high separation orthogonality is important for improved separation. The most common separation mode in LC-MS/MS is the reverse phase chromatography in acidic solution because of its high resolution and highly compatibility with MS/MS analysis[143]. Reverse phase chromatography running in high pH solutions can yield a near-orthogonal separation compared with running in low pH solutions[144]. This combination exploits the difference in isoelectric points between peptides, as peptides have distinct charge states in solutions of different pH values, and has been proven one of the most powerful 2D-LC separation method[141]. The application of 2D-LC followed by MS/MS analysis allows researchers to identify up to 10,000 proteins in one cell lysate sample, and enables in-depth proteomic analysis of highly complexed samples.

1.4 Proteomics expands our understandings of host factors involved in virus infection

Virus infection triggers a massive overhaul of the host cellular environment. Its impacts includes, but not are limited in host gene expressions, signal transductions, protein-protein interactions, and protein-RNA interactions[145-148]. Proteomic analysis based on mass spectrometry, due to its unique ability to simultaneously identify and quantify multiple proteins in the biological sample, has become an important tool to study virus infection and host response to infection[146]. When coupled with different sample preparation techniques, proteomics allows systematic study of biological samples at protein abundance, post-translational modification, protein-protein interaction, and protein-RNA interaction levels[9,10,129,149]. In this section, I will briefly introduce how proteomics empowers studies of virus infection and host antiviral response at multiple levels.

Protein abundance

Virus infection can induce drastic changes in the whole cell proteome of host cells. This includes processes such as synthesis of viral protein, interference of host translation by viral proteins, expression of host antiviral proteins, and downregulation of host proteins. Quantitative proteomic analysis of whole cell lysate can reveal global changes in protein abundance in the cell. When applied to virus infected cells, whole-cell proteome (WCP) analysis enables systematic analysis of host proteins and pathways that are altered by infection. Different viruses are found to induce diverse host responses at proteome level. For example, WCP analysis of hepatitis B virus infected cells highlights overexpressions of innate immunity proteins in host cells[150]. WCP of Sindbis virus infected cells is found to have little changes compared with uninfected cells[151]. SARS-CoV-2 infection, however, can reshape host proteome by altering abundance of proteins in series of host pathways, including translation, splicing, carbon and nucleic acid metabolism[152]. Application of WCP profiling to virus restrictive cells, on the other hand, allows researchers to study proteome-wide changes during host antiviral response. WCP analyses of IFN stimulated cells have helped us understand how cellular proteome landscapes are affected during innate immune response, how different types of IFN induce diverse response in host cells, and how these differential innate immune responses are related to suppression of virus [153-156].

The major limiting factors of the WCP analysis are the sample complexity issue and spatial-temporal resolution. As introduced in previous section, whole cell proteome is an incredibly complexed system that yields over one million unique peptides spanning a huge dynamic range. Most of the LC-MS/MS studies performed with one LC gradient can identify 2,000 ~ 4,000 proteins in one cell lysate sample. This depth of detection is insufficient to observe regulations of low-abundance proteins. WCP analysis with 2D-LC can greatly improve the depth of detection, thus provide insightful data regarding how low-abundance host proteins respond during infection and antiviral response[141,142]. However, these studies are scarce due to technical difficulties. The host proteome undergoes dynamic regulations during infection and antiviral response, however, WCP analysis can only depict the global proteome composition at the timepoint when the cells are harvested. The information of protein localisation is lost in cell lysate, and the recovery of this information require organelle enrichment prior to quantitative proteomic analysis[157]. The temporal changes in proteome cannot be resolved in one WCP analysis. Recovery of temporal regulations requires comparative analysis of WCP datasets from multiple timepoints.

Post-translational modification

Post-translational modifications (PTM) on proteins have major roles in virus infection and host response to virus[146]. MS-based proteomics can be applied to identify and quantify changes of PTMs when coupled with PTM-specific affinity enrichment methods[9]. Both viral proteins and host proteins can have a wide variety of PTMs, such as phosphorylation, ubiquitination, ADP-ribosylation, ISGylation (by ISG15-conjugation as introduced in previous sections) and more others. Studying PTM regulations at the whole-cell level permits identification of host proteins that participate in regulation of virus but have no change in abundance. This can be exemplified by a phosphoproteomic study that analyse total phosphopeptides enriched from HIV-1 infected cells[148]. The analysis revealed how HIV-1 infection impact host phosphorylation landscape, and identified transcription factor ETS and splicing factor SRRM2 as regulators of HIV-1 replication[148]. The major limiting factors of PTM study is also the sample complexity issue. In quantitative proteomic analysis, data from multiple peptides can contribute to quantification of protein abundance. By contrast, PTM analysis can only rely on data of the modified peptides. The difficulty of quantifying PTM-containing peptides can be reduced by performing highly specific PTM enrichment prior to LC-MS/MS analysis, for example ion affinity enrichment for phosphorylation and antibody-based enrichment for ubiquitination[158,159]. However, comprehensive analysis of PTM landscapes during infection and antiviral response remain challenging tasks.

Protein-protein interaction

Progression of virus life cycles heavily rely on host proteins. This hijack of host resources is commonly mediated by viral proteins through (direct or indirect) interactions with host proteins. Characterisation of the composition of viral protein interactors (also viral protein interactome) can be achieved by proteomic analysis of eluates from pulldown studies that use viral proteins as "baits"[149,160]. In a large scale study that coupled proteomic analysis with pulldown experiments of 70 viral proteins from 30 virus species, 579 host proteins are found to have high-confidence interactions with viral proteins[147]. These include host proteins with known roles in innate immunity, such as STAT1/2, as well as proteins with important roles in progressions of virus life cycle, such as splicing factors and translation factors. Similar proteomics-based interactome studies have also been performed in various virus strains, including HIV-1, Influenza viruses, hepatitis virus, flavivirus, coronavirus, etc[161-165]. Each of these works has led to discovery of connections between host and viral proteins, as well as characterisation of novel regulator of virus infection that can either facilitate or antagonise virus life cycle. Protein-protein interactions are also important for host proteins to

exert their functions. Large scale interactome analyses using ISGs as "baits" have suggested that hundreds of human proteins can form protein complexes with ISGs of known antiviral roles[166]. These host proteins are also found to be enriched in viral protein interactors, and could be involved in regulation of viruses despite showing no abundance changes during antiviral response.

Despite having yield fruitful results, protein interactome analyses face a major limiting factor that protein interactors characterised in experimental condition may not reflect the real interactors during infection or antiviral response in physiological condition. Many interactome studies employs cells that overexpress "bait" proteins as research material to study the *in vivo* protein interactome. However, cellular environments are drastically changed upon virus infection or activation of antiviral response, therefore could affect the composition of interactome of the target protein. Interactome analysis of virus infected samples using more carefully designed tagging and pulldown methods is an attractive solution and have started to update us with information of protein interactome in more physiological conditions[167].

Protein-RNA interaction

RNA-binding proteins (RBPs) are important players in host response to virus. Many RBPs are key antiviral factors that can detect and dissect viral RNA[101,168]. A large number of RBPs are also hijacked by viruses to facilitate their life cycles. Quantitative analysis of RNA-bound proteome (RBPome) can be achieved by proteomics with multiple sample preparation methods. The shared principle of these methods are crosslinking of proteins to RNAs that they bound to, then purify RNAs of interests and analyse the RNA-crosslinked proteins with LC-MS/MS[169]. These RBPome analysis methods can identifying host proteins that bind to viral RNA when applied to analyse viral RNA samples[169]. RBPome analysis can also provide insightful data regarding how the host RBPome shifts during virus infection when applied to analyse whole cell RNA samples[151,170]. In a recent study in our lab, we profiled RBPome changes during SARS-CoV-2 infection with two methods: one examines changes in whole cell RBPome by extracting total mRNA from SARS-CoV-2 infected cells and analyse RBPs that are crosslinked to total mRNA by 254-nm wavelength *uv* radiation; the other examines RBPome of SARS-CoV-2 viral RNA by selectively tagging viral RNA with 4-thio-uridine and crosslinking RBPs bound to viral RNA by 365-nm wavelength *uv* light which can only induce crosslinking of 4-thio-uridine[170]. Both methods can provide unique perspectives in observing regulations of RNA bindings, and have revealed series of RBPs that have regulatory roles during SARS-CoV-2 replication. Similar studies have also been performed with other virus types, and have proven successful in unveiling virus-host relations in the RNA binding levels[151,169].

One major limiting factor of RBPome analysis in studying viral RNA interactome is whether interactions between viral RNA its bound RBPs can be detected with high specificity. Each existing RBPome analysis method has its own issue, for example formaldehyde crosslinking risks unspecific crosslinking, antisense oligo nucleotide-based viral RNA extraction risks unspecific binding[169]. Another limiting factor is interpretation of biological roles of the observed RBPome changes. RBPome analyses only provide observations about changes in RNA binding, but not why these RBPs show changed RNA binding. These altered RNA-binding events could be recruited by viruses as part of their life cycle, or could be part of the antiviral response that targets viral RNA for destruction. So far, RBPome studies are focused on virus infection, but little is known regarding how the host antiviral response influence the cellular RBPome.

1.5 Aims

Cellular proteins widely participate in the virus life cycle, by either promoting or restricting virus infection. However, our current understandings regarding regulators of virus infection in host cells remain incomplete. MS-based proteomics is a powerful and versatile tool to investigate host factors in multiple levels. In this project, my aim was to identify (novel) host proteins that can interfere with viral infection using proteomic-based approaches.

- In **Chapter 2** I will start with a comparative proteomic analysis of two cell lines with different permissiveness to virus infection, HEK293 and HEK293T cells. These are closely-related cell lines, but with HEK293T being more permissive and HEK293 more restrictive to multiple viruses. My hypothesis is that this proteome comparison can highlight cellular proteins or/and pathways that determine cell permissiveness to virus infection.
- In **Chapter 3** I will expand this comparative proteomic analysis to include cells in a virus-restrictive state by treating HEK293 with IFN-α. I expect that addition of this analysis will provide us a complete 'picture' of cellular environments that have distinct responses to infection, involving cells at highly permissive state, steady-state, and restrictive state.
- In **Chapter 4** I will investigate the relationship between transcriptome and proteome in IFN-stimulated HEK293 cells. I expect that this analysis will reveal the correlation and discordance between IFN-α responses at the RNA and protein levels. This could help highlight protein expression regulations that are not associated with known IFN-α induced changes in the transcriptome.
- In **Chapter 5** I will investigate the impact of IFN-α stimulation on the cellular environment that is not dependent on changes in gene expression. My studies in this chapter will focus on phosphorylation level and RNA-binding level, and highlight host factors that participate in antiviral response at PTM and protein activity levels.
- In **Chapter 6** I will integrate multiple levels of proteomics datasets from previous chapter, and select 25 putative regulators of virus infection to generate inducible-overexpression cell lines for follow-up studies.

Chapter 2

Proteome-wide comparison between HEK293 and HEK293T cells uncovered host factors and pathways regulating virus replication

Introduction

Immortalised human cell lines are important cellular models in virology and biomedical research. Using these cell lines, researchers can generate homogenous cell samples, manipulate gene expressions, or directly edit DNA sequences of the cell in a robust and highly feasible manner. One of the most commonly used cell lines in virology studies is Human Embryonic Kidney 293 (HEK293) cell. This cell line was derived from human primary kidney tissue of an aborted embryo, and immortalised after exposure to DNA fragments of adenovirus type 5 (AD5) in 1973[171]. The transformed cell line bears an insertion of 4 kbp AD5 DNA fragment in its chromosome 19, expressing adenovirus early proteins E1A and E1B[172]. The viral protein E1A can interfere with cell cycle checkpoint E2F to drive cells from non-dividing state into S-phase[173], and viral protein E1B can also promote the cell cycle through upregulation of several cell cycle control factors and interaction with p53[174]. Affected by these viral proteins, HEK293 cells exhibit tumorigenicity upon injection in mice[171]. The advantages that HEK293 cells offer include ease of handling, and high and reproducible yield of viral particles[175]. Due to these properties, HEK293 cells are widely used in production of recombinant viral proteins, viral vectors, or infective viral particles for both research and biopharmaceutical purposes. The applications include production of adenovirus, retrovirus, lentivirus, rhabdovirus, and influenza[176-178].

HEK293T cell is also a commonly used cell line in virology study. It was derived from HEK293 cells by stable transfection of a temperature sensitive allele of Simian Virus 40 (SV40) large tumour antigen (T antigen) gene. SV40 is an oncogenic DNA virus that can cause cancer in animal models[179]; and large T antigen is the master regulator of SV40 DNA replication[180]. Therefore, the expression of SV40 large T antigen (SV40 LT) enables replication of vectors that contain SV40 *ori* sequence[179]. Besides acting as viral DNA replication initiator, SV40 LT is an oncogenic viral protein that modulates cellular environment through interactions with multiple key nodes in the signalling network[181,182]. This includes several master tumour suppressors, such as p53[183], and all members of retinoblastoma protein (pRB) family[184]. Despite inactivation of p53, which is known to be crucial in maintaining genome stability[185], the genome stability of HEK293T cells was not seem to be affected. This was demonstrated by a recent genome sequencing study that compared HEK293T cells stocks of different sources. These stocks were separated and passaged independently over a long period of time, and all tested stocks showed high similarities in both single-nucleotide polymorphism (SNP) and copy-number variation (CNV)[186].

HEK293T is a preferred choice in production of HIV-1-based lentivirus due to its higher yield compared with HEK293 cells[187]. Recent studies revealed that HEK293T cells also had higher production of various types of viruses than HEK293 cells, including other RNA viruses (i.e. murine

stem cell virus), and DNA viruses (i.e. adenovirus)[188,189]. These examples indicate that the cellular environment in HEK293T cells could be more permissive to multiple types of viruses than HEK293 cells; however, the underlying mechanism(s) of this difference is largely unknown. One may suspect that SV40 LT could directly enhance replication of viruses or dampen host defence response, because it is uniquely expressed in HEK293T cells and not in HEK293 cells. Series of studies have, indeed, reported links between SV40 LT and the cellular innate immunity. However, reported effects in different cellular contexts are highly contradictory: overexpression of SV40 LT has been shown to upregulate a wide range of interferon-stimulated genes (ISGs) in immortalised mouse embryonic fibroblasts (MEF)[190] and human fibroblast[191]; however, it has also been shown to have no effect on virus sensing in primary MEF[192], and even inhibit the PKR-mediated antiviral response in monkey kidney cell line[193], and inhibit cGAS-STING pathway-dependent interferon production in HEK293 cells[194]. In addition to these contradictory results, studies focusing on lentivirus production showed that expression of SV40 LT have little contribution to the different lentivirus yield in HEK293 and HEK293T cells — neither overexpression of SV40 LT in HEK293 cells, nor shRNA-mediated knock-down of SV40 LT in HEK293T cells affect lentivirus titre in respective cell line[195]; CRISPR-mediated SV40 LT knockout also have little effect on lentivirus or adenovirus production in HEK293T cells[189]. Collectively, these studies show that expression of SV40 LT alone cannot explain why HEK293 and HEK293T cells exhibit different permissiveness to viruses.

Apart from expression of SV40 LT, differences in other aspects could also lead to cellular environments of distinct virus permissiveness between HEK293 and HEK293T cells. Genome sequencing showed that these two cell lines have similar SNP distribution. Compared with reference human genome in RefSeq, hundreds of SNPs were commonly held by both HEK lineages, while only 11 SNPs are specific to HEK293 or HEK293T and located in protein coding region[186]. By contrast, more CNVs and structural variations (SVs) were found in HEK293T cells[186]. Microarray-based transcriptome comparison between these two cell lines showed that 490 genes were differentially expressed, possibly caused by the observed variations in genome structure and gene copy number. Interestingly, it has been shown that interferon stimulation was not able to restrict lentivirus production in HEK293T cells, whereas virus titre in monocytic cells was significantly suppressed by interferon[195]. This suggests that innate immune response might be compromised in HEK293T. However, none of the existing studies has systematically compared the innate immune response between HEK293 and HEK293T cells. Based on the current knowledge, transcriptome comparison of these two lines revealed no enrichment in innate immunity-related pathway among differentially expressed genes[186]. No previous study has compared their expression profile differences at the protein level, and our knowledge remain largely blank regarding whether differential expression of

innate immunity genes, or genes involved in other processes, are related to the distinct permissiveness to viruses between HEK293 and HEK293T cells.

Current evidence shows that virus production is more efficient in HEK293T than in HEK293 cells. This finding has been extended to several viruses from different families, indicating that these two cell lines have different permissiveness to virus infection. However, the observed differences cannot be explained by the expression of SV40 LT in HEK293T cells, and the underlying cellular factors remain unclear. Due to the similar genomic contexts and distinct response to virus infection, I believe that these two closely-related cell lines could serve as a system for exploration of virus-restrictive or virus-facilitating factors in the host. In this chapter, I aimed to further assess the difference in virus production capacity between HEK293 and HEK293T cells using two distinct types of viruses – Sindbis virus (SINV) and HIV-1, and to explore cellular factors that are differentially expressed between HEK293 and HEK293T cells using proteomic profiling. Although HEK cells are not natural host of SINV or HIV-1, they permit full life cycle for both viruses and are established model cell lines for virology studies. Through the comparison of proteome landscapes in HEK293 and HEK293T cells, I intended to highlight the underlying factors that could regulate virus production.

Results

2.1 HEK293T cells are more permissive to virus infection than HEK293 cells

HEK293T cells have been reported to show higher virus productions, including HIV-1 and HIV-1-based lentivirus, than HEK293 cells[188,189,195]. Before investigating the factors behind this phenomenon, I first repeated the HIV-1 viral vector transfection experiment using HEK293 and HEK293T cells as described by Li, et al.[188]. The HIV-1 viral vector used in this study were derived from pNL4-3.Luc.R-E- (strain NL4-3), which contains two frameshifts in viral protein Vpr and Env open reading frames[196]. This allows safe use in a biosafety category II laboratory. The viral vector was engineered by removing luciferase tag and insert a fluorescence protein, mCherry, in-frame into Gag polyprotein (HIV-$1_{R\text{-}E\text{--}Gag\text{-}mCherry}$) at the C-terminus of viral protein matrix (p17) linked with a Gly-Ser linker. Insertion into this site had been shown to have minimal impact on production of infective HIV-1 particles[197]. Gag is known to be a late-phase viral protein whose expression relies on production of non-spliced HIV-1 genomic RNA, therefore, fluorescent signal from Gag-inserted mCherry could indicate expression levels of late-phase viral proteins.

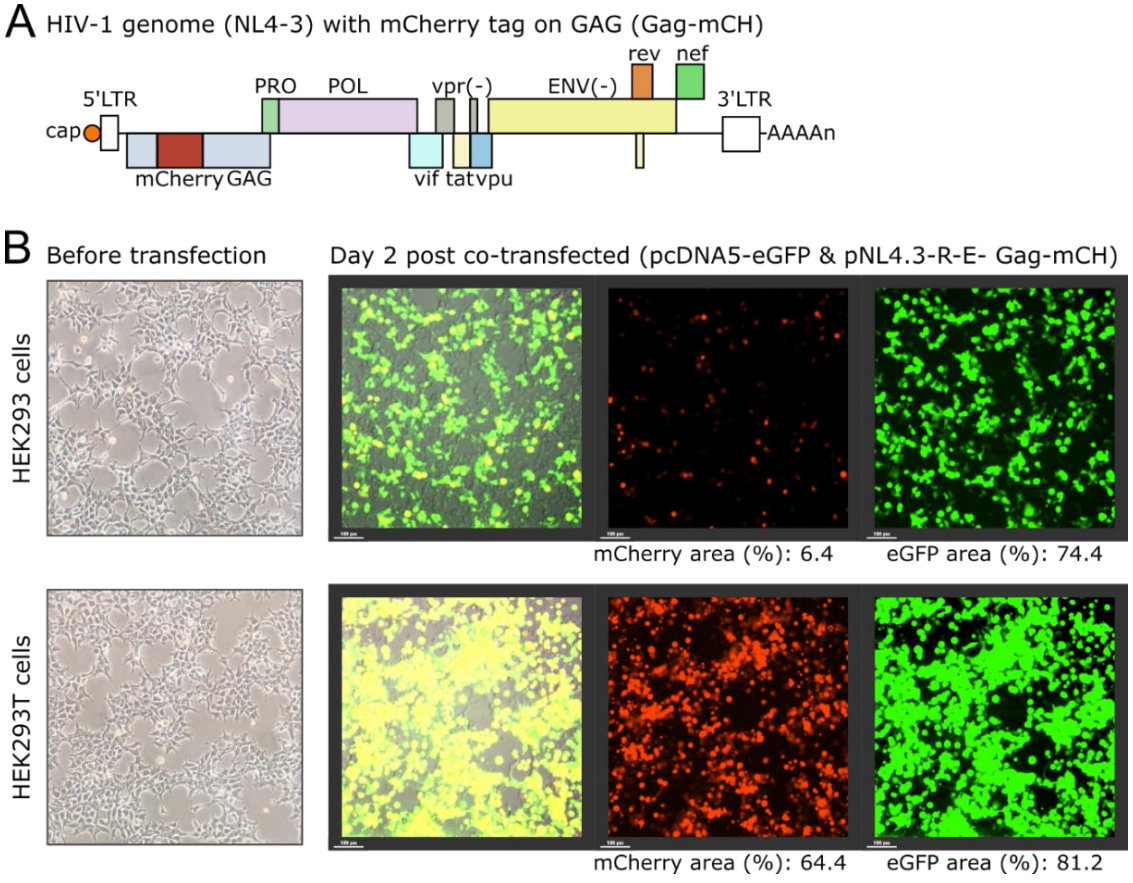


Figure.2.1.1 Higher viral protein expression levels in HEK293T cells transfected with an HIV-1 encoding plasmid than in HEK293 cells. (**A**), diagram of vpr & env double- frameshifted HIV-1 genome engineered from pNL.4-3.Luc.R-E- by removal of luciferase and in-frame insertion of mCherry into Gag polyprotein (HIV-$1_{R\text{-}E\text{--}Gag\text{-}mCherry}$)[196]. (**B**), Image of HEK293 and HEK293T cells before and 2-days after co-transfection of HIV-$1_{R\text{-}E\text{-}Gag\text{-}mCherry}$ viral vector and eGFP empty vector. Microscope acquisition parameters for HEK293 were identical to HEK293T cells. Areas covered by fluorescent signals were calculated for each image with ImageJ using ImageJ Macro "Color Thresholder" with in-house editions (see methods).

HEK293 and HEK293T cells were cultured in 6-well plates to the same confluency prior to transfection (Fig.2.1.1 B, left). A mixture of HIV-$1_{R\text{-}E\text{-}Gag\text{-}mCherry}$ and eGFP-expressing empty vector was co-transfected to each well of cells. Cells were imaged by fluorescence microscopy at 2 days post co-transfection, and the area covered by fluorescent signal was calculated using FIJI Macron (see methods). HEK293T cells exhibit appreciably higher level of mCherry signal than HEK293, with 10-fold higher incidence of mCherry-positive cells (Fig.2.1.1 B, right). However, HEK293T cells also showed a stronger signal for eGFP, which served as control for plasmid intake of the cells. Under the acquisition parameter that were applicable for HEK293 cells, the eGFP signal from HEK293T cells reached saturating levels (Fig.2.1.1 B, right). These results suggest that more copies of plasmid DNA were successfully transfected in HEK293T cells, leading to higher expression. Therefore, it is possible that the distinct transfection efficiencies between HEK293 and HEK293T cells lead to different initial copy numbers of viral vector and subsequent differential levels of viral proteins.

To rule out the bias from different transfection efficiencies and determine if HIV-1 gene expressions differ between HEK293 and HEK293T cells, I infected the cells with single round HIV-1 particles pseudotyped with vesicular stomatitis virus G protein (VSV-G)[198] (Fig.2.1.2 A). VSV-G-mediated cell entry relies on recognition of a highly ubiquitous receptor family – LDLR[199]. LDLR family are conserved and constitutively expressed across cell types and tissues, allowing VSV to infect a wide range of hosts, including insect and vertebrate cells from various species and sources[200]. Therefore, VSV-G pseudotyped HIV-1 particles likely have negligible entry bias between HEK293 and HEK293T cells. I used two HIV-1 constructs with mCherry tag at the matrix protein (MA) and being thus expressed as part of Gag polyprotein (HIV-$1_{R\text{-}E\text{-}Gag\text{-}mCherry}$, Fig.2.1.1 A), or in-frame with the viral protein Nef separated by the linker peptide T2A (HIV-$1_{R\text{-}E\text{-}Nef\text{-}mCherry}$, Fig.2.1.2 B). This T2A linker is a self-cleavage peptide sequence, so the tagged mCherry is separated from Nef after translation and have little impact on virus proliferation[201]. Viral protein nef is an early-phase protein expressed

through a fully spliced splicing isoform of HIV-1 RNA, therefore, the mCherry signal from HIV-$1_{R\text{-}E\text{-}Nef\text{-}mCherry}$ reflects early-phase viral protein production.

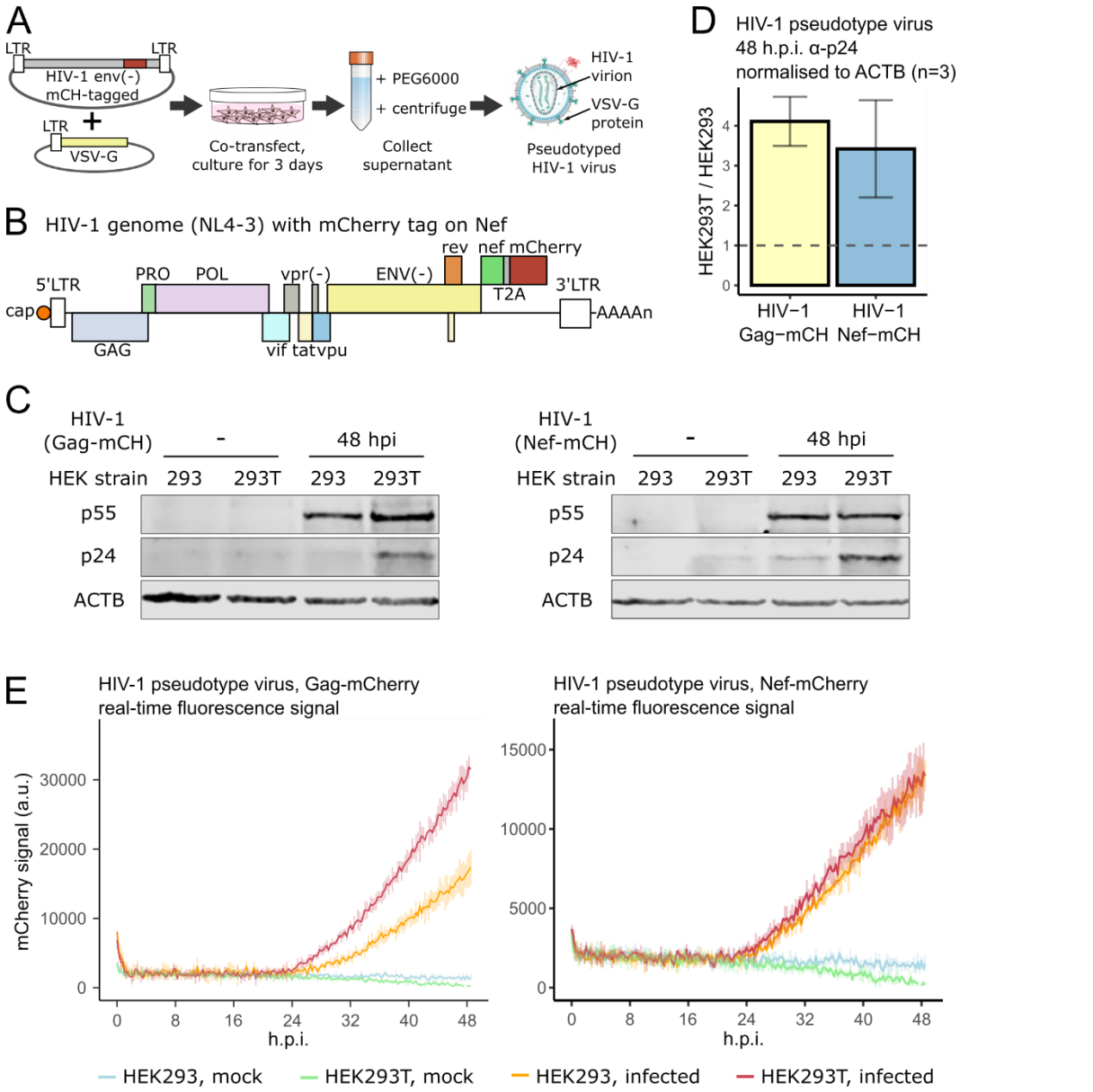


Figure.2.1.2 Infection of pseudotyped HIV-1 showed higher late-phase viral protein level in HEK293T than HEK293. (**A**), schematic workflow of pseudotyping HIV-1 viral particles with vesicular stomatitis virus G protein (VSV-G). (**B**), diagram of vpr & env double-frameshifted HIV-1 genome engineered from pNL.4-3.Luc.R-E- by replacement of luciferase with mCherry (HIV-$1_{R\text{-}E\text{--}Nef\text{-}mCherry}$)[196]. (**C**), immunoblotting against HIV-1 viral protein capsid (p24) for HEK293 and HEK293T cells were challenged with pseudotyped HIV-$1_{R\text{-}E\text{-}Gag\text{-}mCherry}$ or HIV-$1_{R\text{-}E\text{-}Gag\text{-}mCherry}$ at multiplicity of infection (MOI) of 1. The p24-containing polyprotein p55 was also blotted. (**D**), virus challenge and immunoblotting were performed in triplicate, signal from p24 was normalised to ACTB, and ratios of signals between

HEK293T and HEK293 cells were summarised into a barplot. (**E**), fluorescent signal from virus-challenged (MOI=1) HEK293 and HEK293T cells were measured every 15 min in culture condition using a plate reader. The experiment was performed in triplicate.

Cultured HEK293 and HEK293T cells were challenged with pseudotyped HIV-$1_{R\text{-}E\text{-}Gag\text{-}mCherry}$ or HIV-$1_{R\text{-}E\text{-}Gag\text{-}mCherry}$ particles at multiplicity of infection (MOI) of 1. Lysates at 48 hours post infection (hpi) were immunoblotted against late-phase viral protein capsid (p24). Blotting results of signals from p24 (Capsid) and p24-containing polyprotein p55 confirmed higher expression level of viral proteins in HEK293T than HEK293 cells (Fig.2.1.2 C). Immunoblotting was performed using secondary antibody modified with a near-infrared fluorophore, which makes the correlation between signal and abundance more linear, resulting in a more accurate quantification of blotting result. Quantitated results from triplicate experiments showed that HEK293T produces 3~4 fold more Gag and p24 than HEK293 cells (Fig.2.1.2 D).

To distinguish expression levels of early- and late-phase HIV-1 viral proteins, and to further determine the differences in viral protein expression kinetics between HEK293 and HEK293T cells, experiments were setup to monitor mCherry signal from infected cells in real time. Briefly, cells challenged with HIV-$1_{R\text{-}E\text{-}Gag\text{-}mCherry}$ or HIV-$1_{R\text{-}E\text{-}Nef\text{-}mCherry}$ were analysed in a plate reader, where cells were maintained in 37 °C isothermal and 5 % CO_2 condition. Fluorescent signal from mCherry was measured every 15min, allowing near real-time measurement of viral protein expression levels using mCherry as a proxy. Within the first 30 min after infection, both HEK293 and HEK293T cells showed a spike in mCherry signal that correspond to the fluorescent viruses (Fig.2.1.2 E). The similar levels of early mCherry detection suggests that the virus intake was comparable in both lines. After 30 min, mCherry signal reached the basal line and both cell lines showed detectable mCherry signals again at 24 hours post infection. The mCherry signal increased exponentially afterwards as virus life cycle progressed. Throughout the experiment, HIV-$1_{R\text{-}E\text{-}Gag\text{-}mCherry}$-infected HEK293T exhibited substantially higher mCherry signal than HEK293 cells, confirming higher expression level of late HIV-1 viral proteins (Fig.2.1.2 E). Interestingly, no significant difference was observed using HIV-$1_{R\text{-}E\text{-}Nef\text{-}mCherry}$, with the mCherry signal remaining almost identical between the two lines throughout the infection (Fig.2.1.2 E). Together, these data showed that HIV-1 infected HEK293 and HEK293T cells expressed similar levels of early viral proteins, and different levels of late viral proteins.

Several types of viruses also display differential infection capacity in HEK293 and HEK293T beyond HIV-1[188,189]. I hypothesize that this phenomenon could be applicable to other types of viruses that have not been tested in previous studies. To test this, further viral fitness assays were carried on with Sindbis virus (SINV). This is a positive sense, single-stranded RNA virus from the alphavirus

genus that replicates in cytoplasm. Wild-type SINV was used for Western blots, whereas a SINV-mCherry, which expresses the fluorescent protein from a duplicated subgenomic promoter, was used for real-time viral fitness assays. Expression of SINV subgenomic promoter is activated at late times post infection (5h) upon maturation of the non-structural proteins (NSPs) that are expressed at early times post infection (Fig.2.1.3 A).

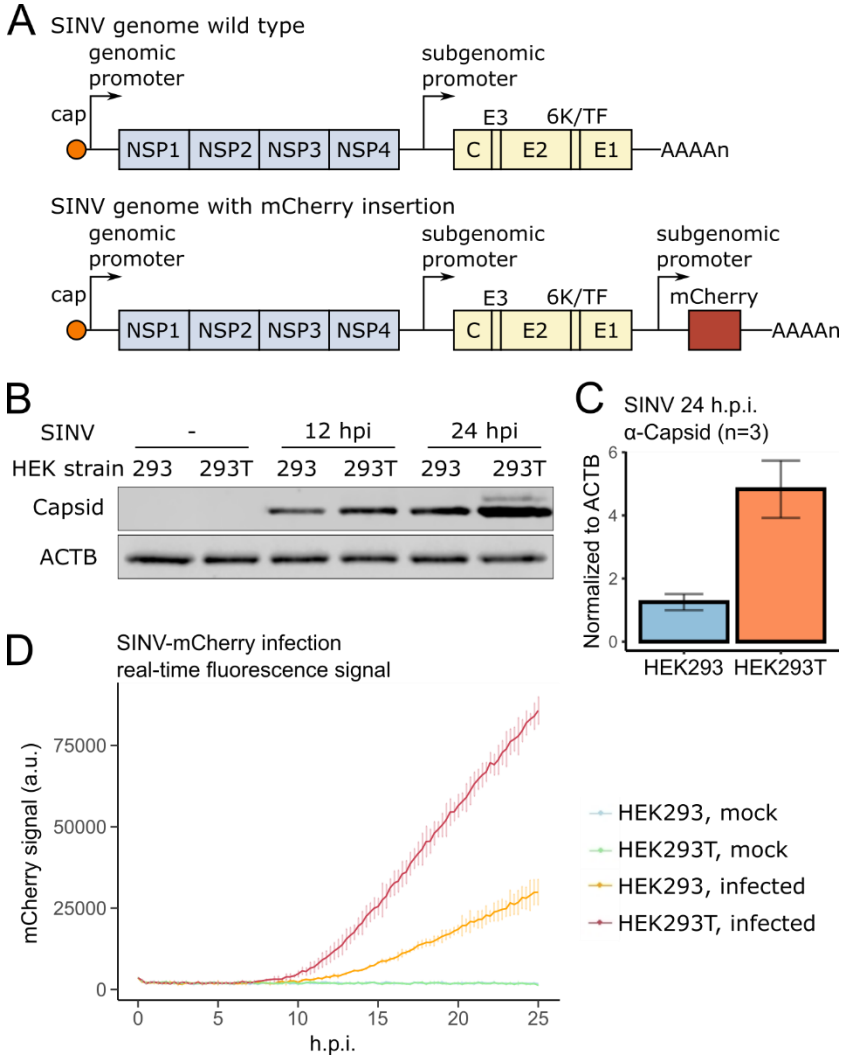


Figure.2.1.3 Infection of Sindbis virus (SINV) showed higher viral protein expression in HEK293T than HEK293 cells. (**A**), diagram of wild-type SINV viral genome, and the mCherry-inserted genome. (**B**), immunoblotting against SINV viral protein Capsid for SINV-challenged HEK293 and HEK293T cells. (**C**), virus-challenge and immunoblotting were performed in triplicate, Capsid signals from HEK293 or HEK293T lysates were normalised to ACTB, and summarised in barplot. (**D**), fluorescent signal from virus-challenged (MOI=1) HEK293 and HEK293T cells were measured every 15min under culture condition using plate reader. The experiment were performed in triplicate.

SINV-infected HEK293 and HEK293T cells were collected at 12 and 24 hpi for Western blot against viral protein Capsid. As with HIV-1, HEK293T also exhibited approx. 4 fold higher expression of SINV capsid than HEK293 cells (Fig.2.1.3 B, C). Viral replication kinetics were acquired with near real-time plate reader assay using mCherry signal as a proxy for gene expression. Fluorescent signals from infected HEK293T cells were substantially higher than background noise from uninfected HEK293T cells at 8 hpi. On the other hand, mCherry signal from infected HEK293 cells was only detectable at 10 hpi, suggesting a delay in the viral life cycle. Throughout the experiment, SINV-infected HEK293T cells consistently displayed higher mCherry signal than HEK293 cells (Fig.2.1.3 D). These results strengthened the conclusion that HEK293T cells are more permissive to viruses than HEK293 cells.

2.2 'Single-shot' proteomic analysis tentatively suggests differential enrichment in innate immunity pathway

After confirming that HEK293T cells are more permissive to viruses than HEK293 cells, I was interested in understanding the host proteins behind this phenomenon. A whole-cell proteomic profiling was performed to explore proteins differentially expressed between HEK293 and HEK293T cells. These two cell lines were raised with stable isotope labelling in tissue culture (SILAC) for high-accuracy quantification by mass spectrometry (MS). Briefly, carbon and nitrogen atoms of amino acids Arg and Lys in tissue culture media were replaced by heavy stable isotopes C_{13} and N_{15}. These isotopologues can be distinguished by mass spectrometer base on shifts in molecular weights, but are biochemically indistinguishable from the natural amino acids[202]. Due to this property, light- and heavy-isotope incorporated cells exhibit the same phenotype, and when mixed prior to sample processing, the sample mixture will experience similar losses and handling errors throughout the processing pipeline. This allows more accurate quantification when comparing signals from light and heavy channels within the same sample mixture. HEK293 and HEK293T cells from the same replicate were differentially labelled and combined prior to processing, and cells from different replicates were labelled with "label swap" strategy, where both cell types were allocated light and heavy isotopes to avoid bias in labelling step (Table.2.1). A small fraction of lysates in each sample were trypsin digested individually to assess isotope incorporation rate using a short MS gradient. For SILAC-based quantitative proteomic analysis, the remaining samples were analysed using longer MS gradient for SILAC-based quantitative proteomic analysis (Fig.2.2.1 A).

Table.2.1 Label swap and sample combine strategy in SILAC experiment

	Sample (isotope label)
Rep.1	HEK293 (L) + HEK293T(H)
Rep.2	HEK293 (H) + HEK293T(L)
Rep.3	HEK293 (L) + HEK293T(H)

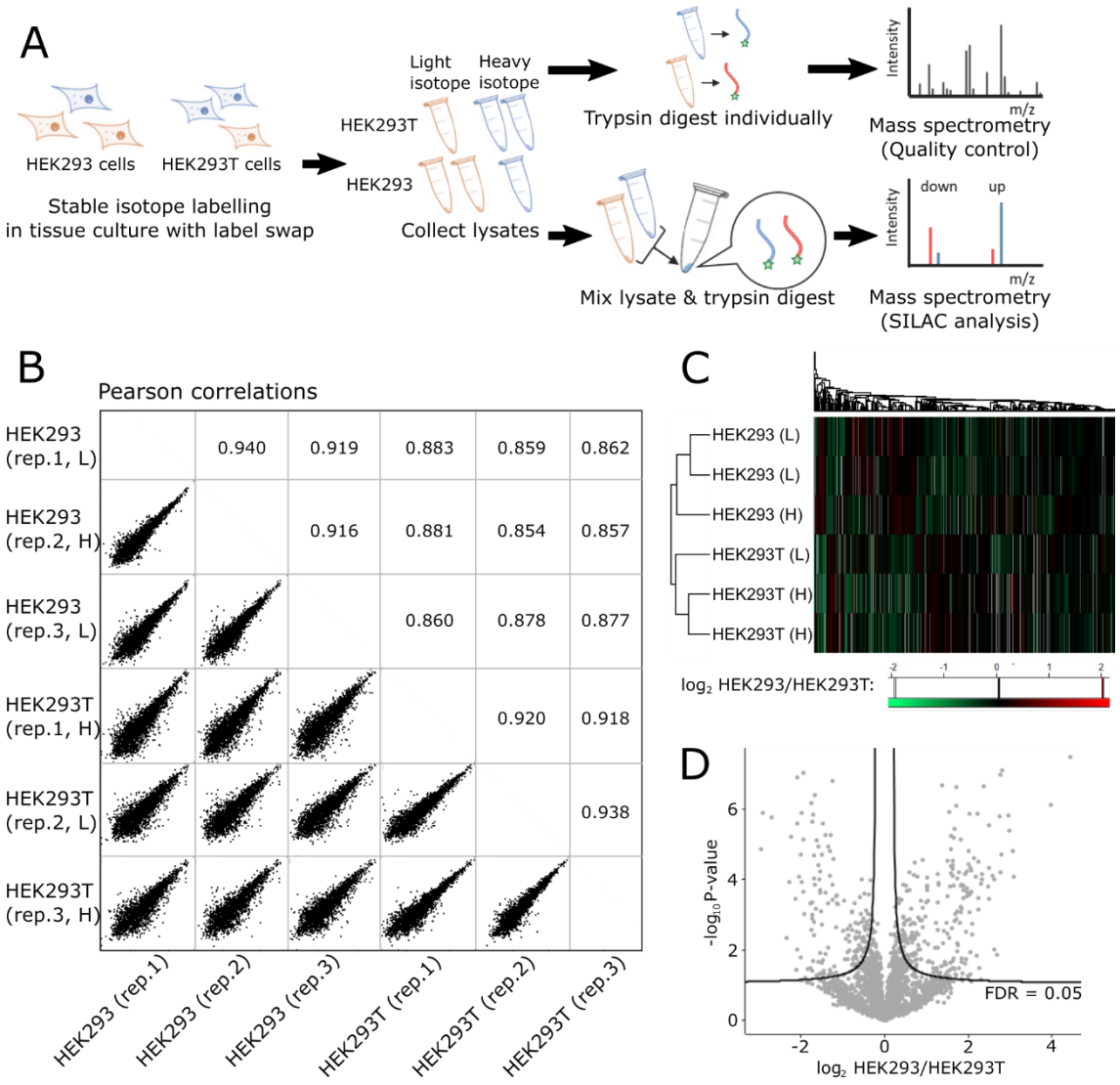


Figure.2.2.1 'Single-shot' whole-cell proteome analysis revealed hundreds of proteins differentially expressed between HEK293 and HEK293T cells. (**A**), schematic workflow of single-gradient profiling of HEK293 and HEK293T whole cell proteome with SILAC. Red and blue colour indicates samples incorporated with light and heavy isotopes. (**B**), scatter plots of log-transformed intensities, Pearson correlation coefficients for each contrast pair were shown in corresponding blocks. (**C**), protein intensities were medium-normalised and plotted into heatmap to visualise differential expressions of each protein among 6 samples. Relative protein expressions were labelled in colour gradient,

missing values were indicated by grey colour. Hierarchical clustering were performed on rows (samples) and columns (proteins). (**D**), volcano plot of proteome-wide comparison test between HEK293 and HEK293T cells. Log_2 fold changes were plotted on x-axis, $-log_{10}$ P-values were plotted on y-axis. Cut-off line for FDR = 0.05 were drawn in solid curves.

Mass spectrometry raw data were searched against human proteome using software MaxQuant. Based on search results from quality control experiments, the isotope incorporation rates were above 95% incorporation rates for all heavy-labelled samples, demonstrating high quality in isotope labelling (Table.2.2). Meanwhile, peptide/protein identification and quantification results were obtained from search results of the SILAC-based MS analysis. On average ~3,000 proteins were identified in each sample, and 4243 proteins were identified in total (Table.2.2).

Table.2.2 Quality control for single-gradient HEK293 & HEK293T proteome profiling.

	HEK293 rep1, L	**HEK293T rep1, L**	**HEK293 rep2, L**	**HEK293T rep2, H**	**HEK293 rep3, H**	**HEK293T rep3, H**
Protein identified	3320	3261	3062	2922	2804	2721
Peptide identified	18107	17881	16793	14654	13924	13781
Isotope incorporation (%)	99.528	99.536	99.625	96.722	96.740	96.363

Protein intensities from search results were log-transformed and proceeded to downstream analysis. Reproducibility of measurements were assessed by Pearson correlation comparing intensity distributions of each sample. As a result, correlation coefficients of samples of the same cell type exceeded 0.9, demonstrating high reproducibility between replicates (Fig.2.2.1 B). On the contrary, correlations comparing samples across cell types were consistently lower than 0.9, suggesting wide range of differences in protein expressions between HEK293 and HEK293T cells. Hierarchical clustering was performed to cluster both samples and proteins using medium-normalised protein intensities. For clustering at sample level, data derived from HEK293 and HEK293T cells clustered into two corresponding groups, this confirmed the Pearson correlation results showing that replicates were highly reproducible, and that the two cell lines have distinct proteome landscapes. For clustering at protein level, numerous proteins displayed distinct expression patterns between two cell lines. To further determine the significance of these changes, statistical analysis was performed using software Perseus. Proteins that were detected and quantified in all triplicates of either HEK293 or HEK293T samples were selected and defined as proteins with reliable identification. As many as 2,765 proteins met this criteria and were taken forward to a student T-

test. Proteins with significant changes in intensities were determined by false discovery rate (FDR) <0.05 (Fig.2.2.1 D). In total, 163 and 111 proteins were identified to be significantly enriched in HEK293 and HEK293T cells respectively.

Gene Ontology (GO) enrichment analysis was performed to gain insight into the biological processes behind the proteins that show distinct expressions between HEK293 and HEK293T cells (Fig.2.2.2 A). In HEK293 cells, proteins involved in organisation of cytoskeleton and cell junction were enriched, these include key players in adhesion junctions such as integrin subunits ITGA6, ITGB4, and actin-anchoring and branching proteins such as ACTN1/4, FLNA. By contrast, in HEK293T cells, proteins related to macromolecule metabolism were enriched, involving ribosome biosynthesis, RNA metabolism, and amino acid biosynthesis.

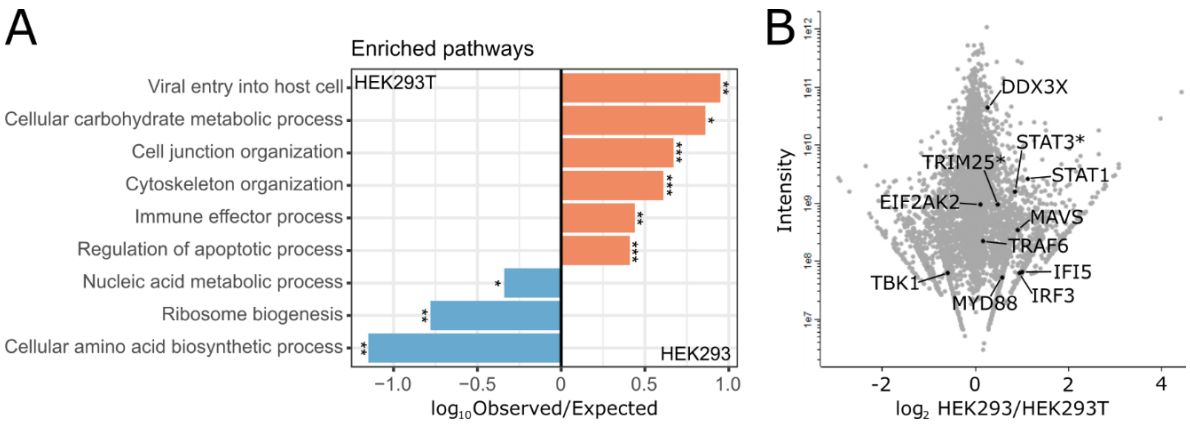


Figure.2.2.2 Enrichment analysis revealed differential expression of innate immune effectors between HEK293 and HEK293T cells. (**A**), enrichment in GO-BP terms were analysed through string.org, and significantly enriched terms ($p<0.05$) were shown in barplot. Strength of enrichment were measured by ratios between observed and expected counts (x-axis), p-values of enrichment were * labelled (*: $p<0.05$, **: $p<0.01$, ***: $p<0.001$). (**B**), scatter plot with fold changes plotted against intensities. Known factors involved in innate immune response were highlighted in black dots and text labelled, proteins shown significant change in T-test were * labelled.

Interestingly, without presence of exogenous immunomodulatory cytokine (i.e. interferon), HEK293 cells were enriched in immune effector proteins over HEK293T cells. These include members of antigen-presentation complex HLA and B2M, and signalling proteins in innate immune response STAT3 and TRIM25. Following this observation, I further assessed key regulatory proteins involved in innate immune signalling. Proteomic profiling result was visualised in scatter plot that draw fold change against intensity, and these proteins were highlighted with text labels (Fig.2.2.2 B). Surprisingly, most of the identified proteins involved in innate immune signalling showed higher

expression levels in HEK293 cells. Although only STAT3 and TRIM25 were determined as significant hits by T-test, proteins such as IFI5, IRF3, and MAVS showed substantially higher expression in HEK293 cells, but were detected with low peptide intensities. It is known that signal-to-noise level has major impact on detection error in mass spectrometry, therefore proteins whose abundance are near detection limit generally bear higher measurement errors, leading to dampened T statistics. Beside the issue of low confidence in statistics, only limited numbers of innate immunity factors were detected, and it remained unknown whether the other factors are differentially expressed between these two cell lines. Taken together, observations from single-shot proteomic data suggest that differential expressions innate immunity factors could potentially explain why HEK293 and HEK293T showed distinct permissiveness to viruses. However, current data generated with single MS gradient remained insufficient in demonstrating this hypothesis.

2.3 Deep proteomic profiling unveiled extensive differences in proteome landscapes between HEK293 and HEK293T cells

In order to effectively quantify innate immunity factors that were expressed in low abundance, I aimed to perform a deep proteomics profiling with improved detection limit. Sample fractionation uses two distinct liquid chromatogram (LC) systems to divide the sample into multiple fractions prior to MSMS analysis[141,142]. It can greatly improve the depth of detection by dividing a high-complexity sample into multiple fractions that have reduced complexity of peptide composition[144]. Each combined SILAC sample was separated into 20 fractions by the first LC, then analysed by LC-MSMS for deep proteomic profiling of HEK293 and HEK293T cells (Fig.2.3.1 A). Chromatograms for fractionation were measured by absorbance at 214 nm, and the profiles of triplicate samples showed high reproducibility in peaks separation (Fig.2.3.1 B). The peptide search was checked for MS/MS identification rate (i.e. identifiable peptide molecules in each fraction) and number of peptide sequences identified (i.e. peptide complexity in each fraction). All fractions showed high homogeneity in peptide abundance and complexity, and all 3 replicates showed high reproducibility in identification (Fig.2.3.1 C). The combined search of all fractions dramatically increased the level of protein identification – over 8,000 proteins were reliably identified and eligible for statistical test in deep proteome data, increased almost 2 times than that in single-shot proteome data (Fig.2.3.1 D).

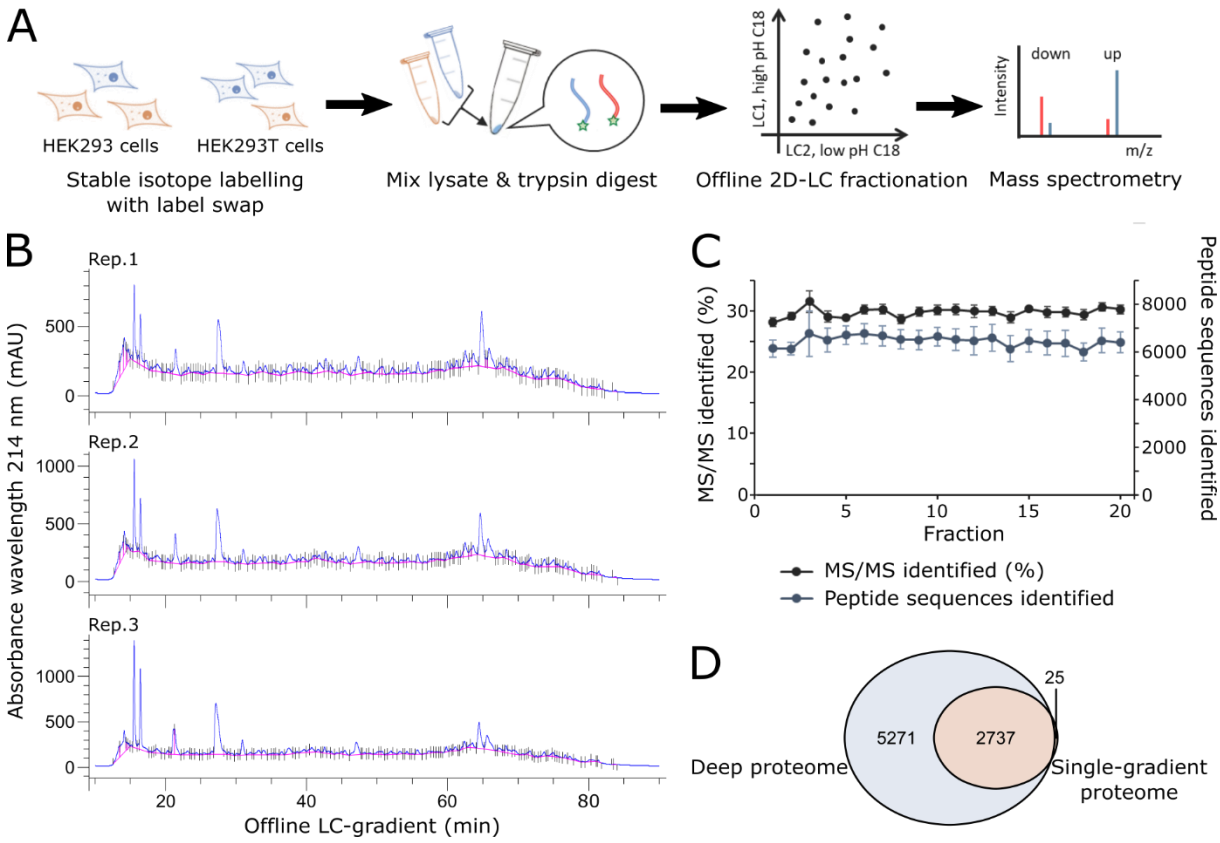


Figure.2.3.1 Deep profiling of whole-cell proteome increased depth of detection by thousands of proteins. (**A**), schematic workflow of deep profiling of HEK293 and HEK293T whole-cell proteome with SILAC. (**B**), chromatogram of offline-LC gradient with reverse-phase separation column in high pH. Eluates from 10 to 90 min gradient were collected and divided into 20 fractions. (**C**), MS/MS identification rates and numbers of peptide sequences identified in each fraction were drawn in line plot for quality control. Error bar represent standard deviation from three replicates. (**D**), venn diagram of number of proteins that are reliably identified in single-gradient proteome and deep proteome.

Apart from the improvement in proteomics methodology, I also focused on improving the statistical test pipeline. Missing values in mass spec detection has always been a difficult and important issue to address. They can be caused by: i) genuine absence of a protein in a given sample, ii) near-or-below detection limit in abundance, iii) or random losses in MSMS precursor selection. Ratio-based SILAC analysis is not able to quantify proteins that show high proportions of missing values in detection. However, these proteins could be important factors that show cell line-specific expression. Therefore, I took protein intensities quantified from the search result, filtered out known contaminants and proteins that had too many missing values, then proceeded to statistical analysis

(Fig.2.3.2 A). In our original pipeline, paired Student's T-test was called to pair-up samples from the same replicate for statistical test (Fig2.3.2 A, left branch). In this way, intensities of differentially labelled samples within the same SILAC mixture can be compared in pairwise manner. Under the default testing setups, paired T-test results revealed astonishingly large differences in protein expressions between HEK293 and HEK293T cells. Over 1,000 proteins and close to 900 proteins significantly enriched in HEK293 and HEK293T cells, respectively (Fig2.3.2 C, left). However, when carefully examined, I found that among the proteins that showed substantial changes but remained as non-significant in test result. A large number of these were well quantified in one cell line and had high proportion of missing values in the other cell line due to low abundance. Imputation of these missing values with signal at background-noise level inevitably introduced artificial variations, which dampened the T-statistics of these proteins. An alternative approach is thus required to fully analyse this data.

One attractive alternative to conventional Student's T-test is a method proposed by G.Smyth that use empirical Bayes model to moderate variances in T-test, termed moderated T-test[203]. It "borrows" information across all genes in the whole dataset to moderate the variance of each gene. Various works have proved its effectiveness in processing data that has large observation number and small sample size. This property rendered it an appropriate tool for this dataset, however, the fact that this dataset was generated with SILAC needed to be also considered. Because cells of the same replicate were differentially labelled and combined prior to sample processing steps, these cells experienced similar losses and handling errors; by contrast, cells of different replicates were processed in parallel and thus had errors from all sample processing steps accumulated. Principal component analysis (PCA) showed clear separations between HEK293 and HEK293T samples in the first dimension; but at the same time, separation by replicate was also evident in the second dimension, which accounts for 23.5% of variations in the dataset (Fig.2.3.2 B, left). This showed that the processing errors accumulated between SILAC replicates be effectively regarded as a layer of the batch effect. Following this rationale, moderated T-test was performed with SILAC replicate effect accounted as batch effect. Briefly, the "replicate" was assigned as a co-variable and incorporated into a linear model for T-statistics, and variances in this direction were subsequently partitioned off from the rest of variances (Fig.2.3.1 A). PCA of the batch effect corrected data showed more than 77 % of variances explained by the first dimension, which represented cell types; meanwhile, separation by replicate was eliminated, showing that accumulated processing errors between SILAC replicates were successfully addressed (Fig.2.3.2 B, right). This improved pipeline revealed 1,486 and 1,441 proteins significantly enriched in HEK293 and HEK293T, respectively, with FDR<0.1 cut-off. These results highlighted a strikingly large difference between these two closely related cell lines.

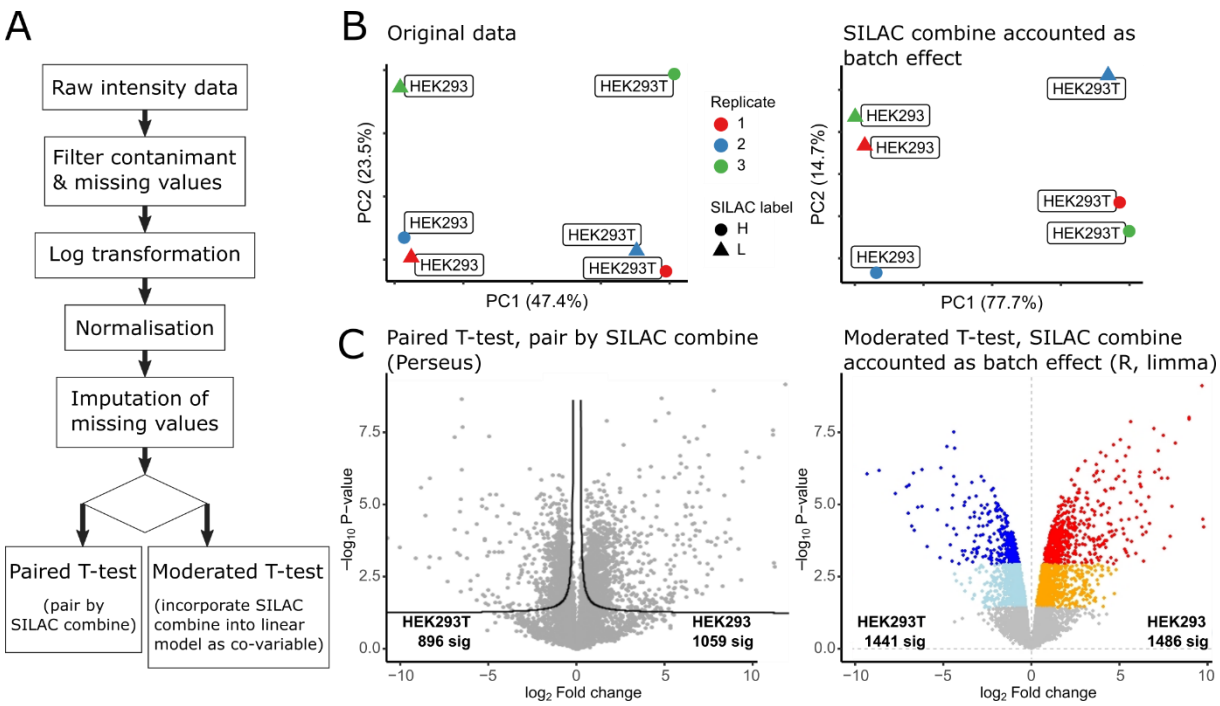


Figure.2.3.2 Improved statistics with batch effect correction. (**A**), flow chart of analysis pipeline. (**B**), principle component analysis (PCA) of HEK293-vs-HEK293T deep proteome data, plotting the first two principle components with percentage of variance explained in respective directions. (**C**), volcano plots showing deep proteome data processed by Student's paired T-test paired by SILAC replicate using Perseus (left), or moderated T-test accounted for batch effect using *limma* package in R (right). Significant hits in paired T-test were determined by FDR<0.1 after permutation-based multiple test correction (curve line). Significant hits in moderated T-test were determined by FDR<0.1 after multiple test correction with Benjamini-Hochberg method (orange/light-blue coloured for FDR<0.1, red/blue coloured for FDR<0.01).

2.4 Deep proteomic analysis revealed enrichment of innate immunity factors in HEK293 cell, and viral protein/RNA interactors in HEK293T cells

To globally assess pathway enrichment in HEK293 and HEK293T cells, deep proteomic profiling results were subjected to a Gene-set enrichment analysis (GSEA) through the website-based analysis tool *string*. In conventional enrichment analyses, all differentially expressed genes are assigned equal weight in the pathway enrichment analysis. Conversely, the GSEA algorithm also

considers the amplitude of the change for each gene, and estimate enrichment scores based on not only on the number of genes enriched but also their fold changes[204]. Pathway enrichments from GSEA of deep proteome were overall similar to the enrichment results the single shot proteome. HEK293 cells were enriched in cytoskeleton organisation-related processes, and HEK293T cells were enriched in processes related to macromolecule biosynthesis and cell proliferation (Fig.2.4.1 A). One of the major differences was that GSEA of deep proteomic data highlighted strong enrichments of cytokine signalling in HEK293 over HEK293T cells, including interferon and interleukin signalling, while single-gradient proteome only showed enrichment of a broad term “immune effector process”.

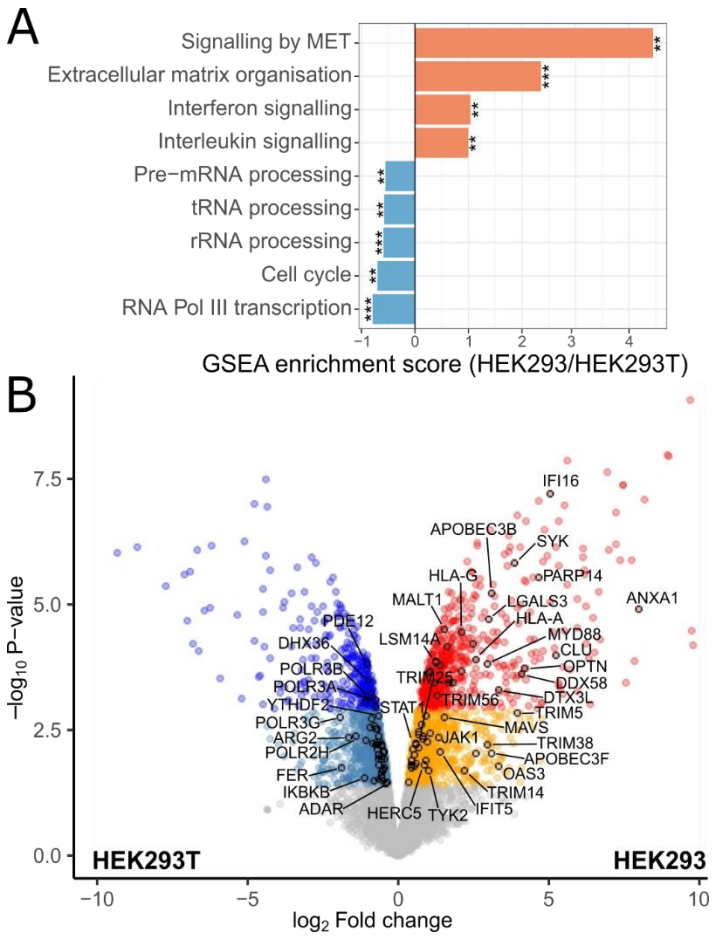


Figure.2.4.1 Gene-set enrichment analysis (GSEA) showed enrichment of interferon signalling in HEK293 over HEK293T cells without presence of exogenous cytokine. (**A**), representative Reactome pathways in GSEA were shown in barplot. P-values of enrichments were * labelled (**: p<0.01, ***: p<0.001). (**B**), volcano plot of HEK293-vs-HEK293T deep proteome analysis, with significance level colour labelled (red/blue: FDR<0.01, orange/lightblue: FDR<0.1). Black circle highlighted proteins

annotated with GO-BP terms associated with innate immunity or interferon signalling, proteins with large fold changes were text labelled.

Deep proteomic analysis allowed detection of numerous innate immunity factors that was not achievable with single-gradient proteomic analysis. Notably, without presence of exogenous immunomodulatory cytokine, key regulators in virtually all steps of innate immune signalling were found to be enriched HEK293 over HEK293T cells. Various proteins that have well-characterised functions in innate immune response were amongst them, including the RNA sensor RIG-I (DDX58)[64], DNA sensor IFI16[81], key node in PAMP signalling MAVS[68], interferon receptor IFNGR1, core member in JAK/STAT signalling JAK1 and STAT1[205], and well characterised antiviral factors OAS3, PKR, and APOBEC family members[101] (Table.2.3).

Table.2.3 HEK293/HEK293T-enriched proteins that have well-characterised role in innate immune response signalling. Proteins involved in multiple processes were not repeated in this table.

	HEK293-enriched	**HEK293T-enriched**
Virus sensing	DDX58, IFI16, TRIM5	DDX41, ADAR, DHX36
PAMP-signalling	MAVS, TRIM25, MYD88, TRIM21, TRAF2, TRIM56, TRIM38, DTX3L, OPTN	YTHDF2
IFN receptor	IFNGR1	
JAK/STAT signalling	STAT1, JAK1, TYK2	
Antiviral effectors	HERC5, APOBEC3B, APOBEC3F, OAS3, RNASEL, PKR	

To widen our views and not limit to well-studied regulators in innate immunity signalling pathway, I expanded the analysis to include proteins that were related to innate immunity by their GO annotations. Briefly, all GO-BP annotation terms that contain keywords “innate immunity”, or “interferon” were extracted, lists of proteins annotated under these terms were compared against this deep proteomic data, and the overlapping proteins were defined as “GO-annotated ISGs”. Similar to pathway-level analysis, the GO-annotated ISGs also showed enrichment in HEK293 cells. When visualised in volcano plot of HEK293-vs-HEK293T deep proteome data, these GO-annotated

ISGs had comparable levels of enrichment among low-fold-change proteins but were overrepresented in HEK293 cells among proteins of high fold-changes or low p-values (Fig.2.4.1 B).

Host proteins are involved in virus life cycle in multiple ways. Apart from restricting virus infection through innate immune response, many of them are also hijacked by viruses to facilitate virus replication. Therefore, we asked whether virus-associated proteins are differentially expressed between HEK293 and HEK293T cells. Lists of viral protein interactors and viral RNA interactors were consolidated based on literature reviews. Due to the large variety of methods focused on this field, it is difficult to define interactors under universal criteria. Therefore, only a few proteomics-based studies with standardized enrichment methods were included in this study – viral protein interactors were retrieved from anti-GFP beads-based affinity-purification mass spectrometry (AP-MS) analyses[147,161,164,206] viral RNA interactors were obtained from mass spectrometry analyses of protein crosslinked to viral RNA[169] (Fig.2.4.2 A). Fisher's exact test was applied to calculate enrichments of these virus interactors among proteins that showed significantly different expressions in HEK293 and HEK293T cells. Surprisingly, both viral protein interactors and viral RNA interactors were enriched in HEK293T over HEK293 cells (Fig.2.4.2 B). Further analysis of these viral protein interactors showed

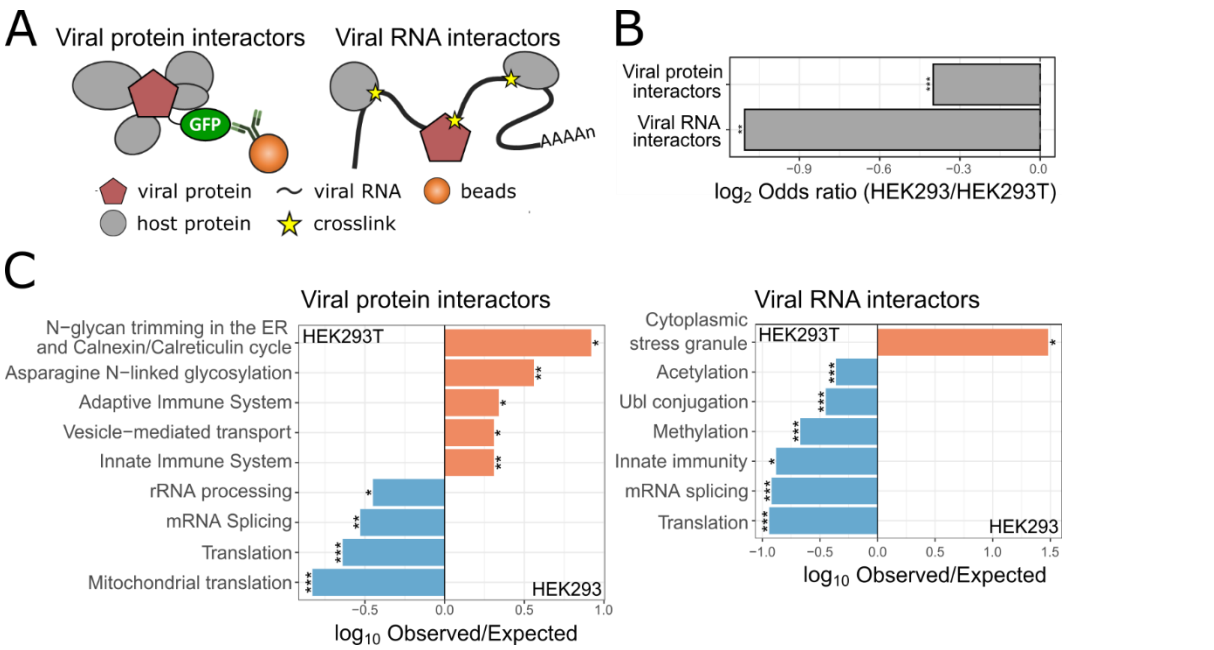


Figure.2.4.2 Known viral protein and viral RNA interactors are enriched in HEK293T cells. (**A**), diagrams depicting criteria for host proteins to be determined as viral protein or viral RNA interactors. Experimental evidences were obtained from literatures. (**B**), enrichment of viral protein or viral RNA interactors between HEK293 and HEK293T cells, p-value of enrichment were * labelled

(**: p<0.01, ***: p<0.001). (**C**), GO-BP enrichments of viral protein or viral RNA interactors were drawn in barplot, with p-value represented by * label (*: p<0.05, **: p<0.01, ***: p<0.001).

strong enrichments of protein linked to mRNA translation in HEK293T cells (Fig.2.4.2 C, left). These included members of translation initiation and elongation families EIFs and EEFs, members of SEC61 family that direct ER-localisation of ribosome, and mitochondrial ribosomal proteins MRPLs and MRPSs. In the HEK293 cells, viral protein interactors were enriched in innate immunity and ER-associated N-glycosylation processes, the later included crucial regulator of protein-folding quality control UGGT1 (Fig.2.4.2 C, left). Viral RNA interactors enriched in HEK293 were enriched in stress granule components, including the antiviral proteins TRIM25 and LSM14A. Unexpectedly, few RBPs involved in innate immunity were enriched in HEK293T cells. This included the RNA editing enzyme ADAR, the antiviral effector ZC3HAV1, as well as proteins that can inhibit innate immune response, for example C1QBP that functions through PI3K-AKT pathway, and YTHDF2 which acts through RNA methylation-dependent degradation.

2.5 IFN-α triggers antiviral responses in HEK293 and HEK293T with different strength.

Analysis of this deep proteomics data revealed the enrichment of innate immunity factors in HEK293 over HEK293T cells. To evaluate how the differential proteome regulate cell's responses, I treated HEK293 and HEK293T with different doses of interferon alpha (IFN-α) and analysed the IFN-response by Western blot, with antibodies against the IFN-α-activated transcription factor STAT1, and phosphorylation level at STAT1 Tyr701 (pSTAT1) (Fig.2.5.1 A). Blotting results showed that HEK293 and HEK293T cells were able to increase expression of STAT1 and induce its phosphorylation in respond to IFN-α in all tested concentrations. When normalising pSTAT1 signal to either ACTB or STAT1, both HEK293 and HEK293T cells exhibit similar kinetic properties in response to different concentrations of IFN-α in triplicate experiments (Fig.2.5.1 B). Immunoblotting was also performed at different times post IFN-α treatment, and both HEK293 and HEK293T cells showed similar increases in STAT1 expression overtime after IFN-α treatment (Fig.2.5.1 C). Notably, the STAT1 overexpression in HEK293 cells appeared to be initiated at an earlier timepoint than HEK293T cells (Fig.2.5.1 D). Interestingly, both cell lines showed consistent drop in STAT1 shortly after IFN-α treatment followed by a recovery and overexpression.

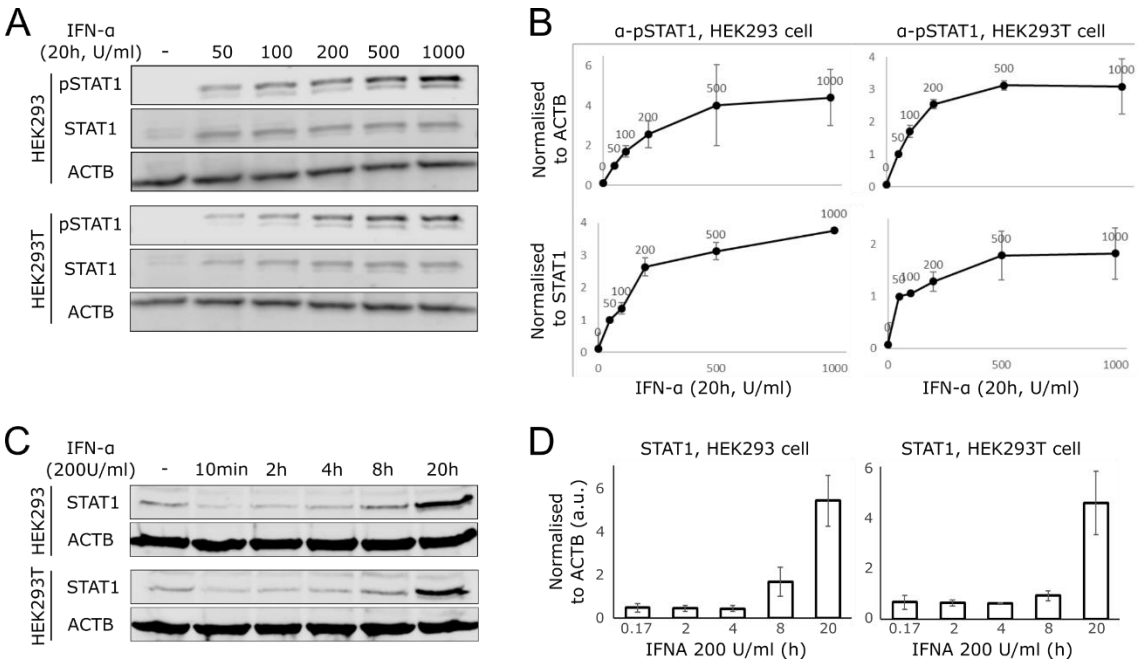


Fig.2.5.1 HEK293 and HEK293T cells showed similar responses to IFN-α stimulation in both time- and dosage-dependent manners. (**A**), HEK293 and HEK293T cells were treated with different dosages of IFN-α for 20 hours, lysates were blotted against STAT1, STAT1-pTyr701 (labelled as pSTAT1). (**B**), STAT1-pTyr701 blotting signals from triplicate experiments were normalised to STAT1 or ACTB, and summarised into line plots. (**C**), HEK293 and HEK293T were treated with 200 U/ml of IFN-α for different periods of time, and blotted against STAT1. (**D**), STAT1 signal from triplicate experiments were normalised to ACTB, and summarised into barplots.

After discovering that HEK293 and HEK293T cells response to IFN-α in a similar manner in virus-free conditions, I tested whether IFN-α also inhibits replication of viruses in both HEK293 and HEK293T cells to a similar extent. Based on IFN-α-response kinetics measured in previous experiments (Fig.2.5.1), the conditions for IFN-α stimulation were set to be 200 U/ml in dosage, and 4 h and 20 h of treatment, representing early- and late-timepoints post IFN-α stimulation. IFN-α treated cells were challenged with SINV or pseudotyped HIV-$1_{\text{R-E-Gag-mCherry}}$, lysates at 24 hpi or 48 hpi were collected for immunoblotting against pSTAT1 and corresponding viral proteins. In both types of viruses, IFN-α-treated HEK293 cells showed substantially lower levels of viral proteins than in IFN-α-treated HEK293T cells (Fig.2.5.2 A). This observation was consistent across triplicate experiments, indicating that IFN-α could efficiently drive HEK293 cells into virus-restricting state, but it was less effective in HEK293T cells (Fig.2.5.2 B). Interestingly, pSTAT1, which was shown to be induced by IFN-α in both HEK293 and HEK293T cells in virus-free conditions, showed very low signal in virus-

infected HEK293T cells (Fig.2.5.2 A). This phenomenon was consistent in SINV and HIV-$1_{R\text{-}E\text{-}Gag\text{-}mCherry}$ infected cells.

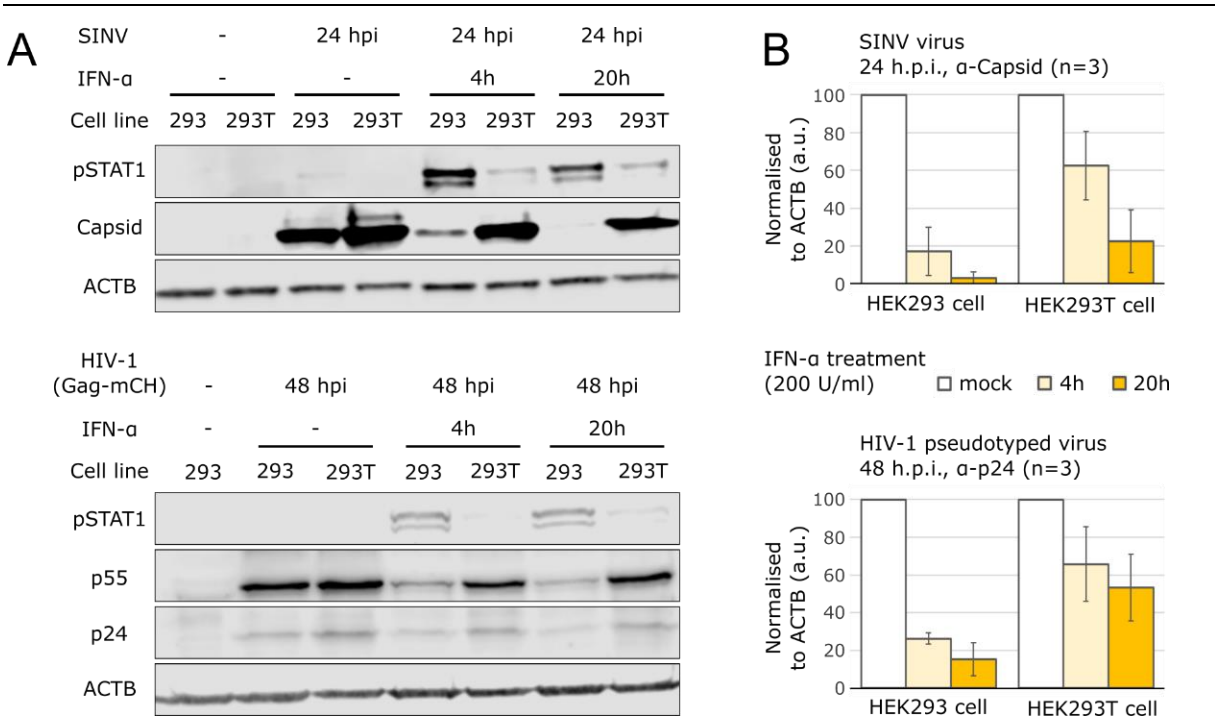


Figure.2.5.2 IFN-α treatment drives HEK293 cells into a virus-restrictive state, but is less effective in HEK293T cells. (**A**), HEK293 and HEK293T cells were stimulated with 200 U/ml of IFN-α for 4h or 20h prior to virus challenge (SINV or pseudotyped HIV-$1_{R\text{-}E\text{-}Gag\text{-}mCherry}$, MOI = 1). Lysates were immunoblotting against respective viral proteins. (**B**), Western blot triplicates of IFN-α-stimulated and virus-challenged HEK293 and HEK293T cells were summarised into barplots. Viral protein signals were normalised to ACTB, and transformed into percentages of mock-treated conditions.

Discussion

Both HEK293 and HEK293T cells are widely used in virology research, and often employed for similar purposes. However, some earlier evidence suggested that they are not analogous in their way to deal with virus infection[187,188,189,195]. The results in this chapter showed that HEK293T cells are more permissive to SINV and HIV-1 when compared with HEK293 cells. The viral fitness assays were conducted with infective viral particles instead of transfection of viral vectors, therefore, bypassing the bias in transfection efficiency between these two cell lines. In addition, two independent methods were used to evaluate virus progression, including immunoblotting and a near real-time fluorescence-based, viral fitness assay. One unexpected observation is that the expression of the early HIV-1 protein Nef was not affected by the cell type. Translation of viral protein Nef depends on expression of fully-spliced HIV-1 RNA splicing variant that to all respects behave as a normal cellular RNA as they are transcribed and processed by the host RNA polymerase II and the spliceosome[1]. By contrast, late HIV-1 proteins are translated from partially-spliced or non-spliced HIV-1 viral RNA, and these transcripts bypass normal host regulation through the viral protein Rev[1]. This unconventional behaviour could make them more vulnerable to virus sensing and antiviral mechanisms, for example, unspliced viral RNAs will not only contain an intron, being perceived by the cell as immature, but they lack host factors deposited during the normal processing of the RNA, such as the exon junction complex[207]. The presence of pathogen-associated features in viral RNA transcripts could explain why expression of late HIV-1 viral protein Gag was significantly lower in HEK293 cells than in HEK293T, while Nef is unaffected.

A whole cell proteomic profiling was performed to explore the proteome landscapes of HEK293 and HEK293T cells and use them to identify host factors that might be accountable for the observed difference in permissiveness to infection. Both single-gradient and deep proteomic profiling revealed strong enrichment of cytoskeleton- and cell junction-related proteins in HEK293 cells, while RNA/protein synthesis and cell cycle related pathways were enriched in HEK293T cells. These conclusions are in-line with their phenotypes in tissue culture – HEK293 cells attach to culture surface more firmly than HEK293T cells, while HEK293T cells have higher proliferation rate. The expression of SV40 LT in HEK293T cells could be the main driving force behind its high proliferation rate and enrichment of cell cycle regulators, because SV40 LT could bind to and inactivate pRB family member, which are known to be key regulator of cell cycle checkpoint G1/S[184,208]. Deep proteomic analysis added to this picture that a wide range of innate immunity factors are higher expressed in HEK293 cells than in HEK293T cells. Due to limited ability to detect highly complexed peptide mixture from whole-cell proteome, single-shot proteomics only identified a few core members in

innate immunity signalling, and suggested a trend of enrichment in HEK293 cells, which was low in confidence. This is not surprising because innate immunity factors are expressed to a low level in virus-free environment. Detection of antiviral proteins and low-abundance proteins was massively improved by offline fractionation-based deep proteomic analysis, which uncovered numerous high-confidence differentially expressed innate immunity related factors. It was striking to find innate immunity factors enriched in HEK293 cells at such a whole-pathway scale, without the presence of exogenous immunomodulatory cytokine. Previous studies of SV40 LT have shown that it could either induce or repress innate immunity-related genes depending on cellular systems studied. However, modulation of innate immunity in whole-pathway level has not been reported. It is possible that perturbations of protein expressions by SV40 LT were more wide-stretching than previously thought, and the proteomic data presented in this study provides new insights into immune-modulating functions of SV40 large T antigen. Alternatively, the observed difference in permissiveness to virus and expressions of innate-immunity protein could be caused by factors other than SV40 LT. It is worth noting that apart from harbouring SV40 large T antigen, differences at the genomic level exist between HEK293 and HEK293T cells despite their close lineage. As mentioned in the introduction, these two lines have similar SNPs, but HEK293T cells harbour more CNVs and SVs[186]. Investigation of whether it is the SV40 LT or the genome structural variation modulates cellular innate immune response in a pathway-scales is an exciting topic for future research.

Apart from the enrichment of innate immunity factors in HEK293, the enrichment of viral RNA and protein interactors in HEK293T cells was also surprising. It is worth noting that many known virus-restricting RBPs, like RIG-I and APOBEC family members, were not classified as “viral RNA interactors” by criteria introduced in this chapter. One possible reason is that the viral RNAs bound by these strongly virus-restricting RBPs are efficiently removed by the host, and therefore, unable to be captured by viral RNA pulldown assay and subsequent proteomic studies. Considering this potential sampling bias, it is possible that a large proportion of proteins classified as viral RNA interactors could facilitate virus infection *in vivo*. For example, protein PURA was classified as “viral RNA interactors”, and it has been shown to facilitate HIV-1 replication through interaction with HIV-1 viral protein Tat[209]. Therefore, it is possible that the permissive cellular environment in HEK293T cells is collectively shaped by both the relative depletion of innate immunity factors, and the enrichment of virus-facilitating factors. The caveat in this enrichment analysis, however, is that limited species of viruses were covered. In this chapter, only proteins from several viral protein AP-MS studies were enlisted into viral protein interactors, and proteins from recent crosslinking-based viral RNA pulldown studies were included in viral RNA interactors. These criteria were introduced for quality control across studies at costs of coverages in both host proteins and virus types. The

conclusion might be subject to changes if the analysis was further expanded by incorporating more viral protein/RNA interactome studies in a way that is manually curated or quality controlled. Given the current results that viral RNA/protein interactors were enriched in HEK293T cells, it is interesting to investigate whether or how these proteins modulate the cellular environment into a more virus permissive state.

Acknowledgements

I want to thank all members of Castello lab for the supports in cell cultures and virology assays, thank all members of Shabaz lab for the supports in the proteomics experiments and data analysis. Thanks Alfredo Castello-Palomares & Shabaz Mohammed for conceptualising the research in this chapter and their supervisions. Thanks Marko Noerenberg for kindly providing cell stocks of SILAC cell lines, engineered viral vectors of HIV-$1_{R\text{-}E\text{--}Gag\text{-}mCherry}$ and HIV-$1_{R\text{-}E\text{--}Nef\text{-}mCherry}$, and training on HIV-1 pseudotyping. Thanks Manuel Garcia-Moreno for kindly providing SINV stocks for virus infection test. Thanks Philip Charles for discussions regarding methods for proteomics data analysis.

Methods & Materials

Methods

Cell culture

HEK293 cells (ECACC) were obtained from commercial products, HEK293T were a kind gift from Dr. Jan Rehwinkel (Radcliffe Department of Medicine, University of Oxford). HEK293 and HEK293T cells were cultured in DMEM with 10 % FBS and 1x penicillin/streptomycin at 37°C with 5% CO_2. Cells at 80~90% confluency were passaged regularly by PBS wash and brief incubation with trypsin EDTA solution at 37°C, followed by resuspension and dilution in the culture media.

Culture with IFN-α stimulation was performed by supplying IFN-α stocks into culture media to 200 U/ml for desired period of time prior to harvest. IFN-α stocks were obtained by dissolving commercial IFN-α powder (PBL assay science) with 0.1% BSA in water and stored at -80°C.

Virus plasmid transfection and fluorescence imaging

HEK293 and HEK293T cells at similar confluency (~80%) were co-transfected with pNL4-3 HIV-$1_{Gag\text{-}mCherry}$ and pcDNA5-eGFP plasmid using XtremeGene 9 transfection reagent and Opti-MEM reduced serum medium with a DNA/reagent/medium ratio of 5/15/500 (µg/µl/µl), the ratio between pNL4-3

HIV-$1_{Gag\text{-}mCherry}$ and pcDNA5-eGFP was 3/2 (w/w). Cells were imaged 2 days after transfection on an API Delta-Vision Elite widefield fluorescence microscope using a 10X objective.

Areas covered by fluorescent signal were calculated using ImageJ Macro "Color Thresholder" (2.0.0-rc-65) with in-house editions. Briefly, a 512 by 512 squared region was selected on the centre of each image, a brightness threshold at 255 was set to determine fluorescence signal covered area. The percentage of fluorescence signal covered area was calculated using area passing threshold divided by selected area.

Generation of viruses

SINV with mCherry-tag was kindly generated by Dr Manuel Garcia-Moreno as communal lab stocks. The pT7-SVmCherry plasmid was used for generation of mCherry-tagged SINV. Viral particles were generated by Dr M. Garcia with the following process: plasmids linearized by XhoI were used as a template for T7 RNA polymerase transcription kit (New England Biolabs). Transcribed viral gRNA was transfected into BHK-21 cells. Viruses were collected after 24 h from the supernatant and purified by centrifugation at 2000 rpm for 3 min and filtration with 0.45 µm PVDF syringe filter, then stored at -80°C.

Pseudotyped HIV-$1_{Gag\text{-}mCherry}$ and HIV-$1_{Nef\text{-}mCherry}$ were generated following the protocol by Kutner et al.[210]. Viral vectors were engineered by Dr Marko Noerenberg based on pNL.4-3.R-E- (Centre For AIDS Reagents) by replacing luciferase with mCherry (HIV-$1_{Nef\text{-}mCherry}$), or removing luciferase and inserting mCherry into Gag polyprotein open reading frame (HIV-$1_{Gag\text{-}mCherry}$). Briefly, HEK293T cells at 30~40% confluency were treated with 25 µM chloroquine and co-transfected with a DNA/$CaCl_2$ mixture of 250 mM $CaCl_2$, pHEF-VSVG and pNL4-3 HIV-1 plasmid (HIV-$1_{Gag\text{-}mCherry}$ or HIV-$1_{Nef\text{-}mCherry}$) at 1:3 (w/w) ratio. Media was exchanged after 1 day. The supernatant was collected after 2 days, cell debris was removed by 500 g centrifugation for 10 min and 0.45 µm filter. Viral particles were precipitated with 50% PEG6000, 4 M NaCl and PBS at 4°C for 1.5 h. Precipitated virus were collected by 4,500 g centrifugation for 30 min and resuspended in 50 mM Tris-HCl pH 7.4, then stored at -80°C.

Western blotting

Cells were washed with PBS and lysed with 4°C RIPA buffer (50 Mm Tris-HCl pH 7.5, 150 mM NaCl, 1% Triton-X100, 0.1% SDS, 0.5% sodium deoxycholate) supplied with 2000x protease inhibitor AEBSF

and 1000x nuclease Benzonase. Lysates were centrifuged and boiled at 95oC for 5 min before loaded on house-made 12% polyacrylamide gels for SDS-PAGE analysis. Gels were transferred to nitrocellulose membranes using the Trans-Blot turbo transfer system (Bio-Rad). Membranes were blocked with 5% milk in PBS at room temperature for 1 h, then washed with 0.1% Tween20 in PBS (PBST) for 5 min before incubation with primary antibody in PBST with 5% milk at 4^{o}C for overnight. Membranes were washed twice with PBST for 5 min before incubation with secondary antibody in PBST with 5% milk at room temperature for 1h. Membranes were washed twice with PBST for 5 min before fluorescence imaging on Odyssey Fc imaging system (LI-COR Bioscience).

Plate reader assay

HEK293 and HEK293T cells were seeded on 96-well plates at 0.25 million cell per well, and infected with SINV-mCherry or pseudotyped HIV-$1_{Gag\text{-}mCherry/}$ or HIV-$1_{Nef\text{-}mCherry}$ at the desired multiplicity of infection (MOI) at 4 h after cell seeding. After infection, 96-well plates were transferred to the CLARIOstar microplate reader (BMG labtech). Fluorescent signals (mCherry gain 2900) were recorded every 15 min under 37^{o}C and 5% CO_2 over a 50 h period. Plate reader results of infected cells were normalised to uninfected cells by subtraction. SINV-mCherry and pseudotyped HIV-1 signals from first 25 h and 50 h were normalised to signal at 24 h and 48 h for plotting, respectively.

Cell culture with SILAC

Culture media for SILAC were arginine and lysine depleted DMEM (Thermo Scientific) with 10% dialysed FBS, 1x penicillin/streptomycin, and supplied with heavy-isotope labeled arginine and lysine (L-Arginine 13C,15N; L-Lysine 13C,15N; Silantes GmbH). SILAC cell stocks were kindly provided by Dr Marko Noerenberg, and these cells were cultured in corresponding SILAC media for at least 5 more passages before experiments. Cell passage process was the same as unlabelled cell.

Proteomics sample preparation

HEK293 and HEK293T cells with SILAC for proteomic analysis were cultured on 6-well plate to ~80% confluency. Cells were gently washed with 4^{o}C PBS, and lysed on plate with 200 µl urea lysis buffer (8 M urea, 100 mM ammonium bicarbonate (AmBic), 2 µl protease inhibitor cocktail (Sigma) and 0.5 µl phosphatase inhibitor (Sigma)). An aliquot of each lysate was reserved for protein concentration measurement using Pierce 660-nm Protein Assay (Thermo Scientific) and isotope incorporation rate

measurement by independently performing sample preparation and 60-min LC-MS/MS gradient with Q Exactive (Thermo Fischer Scientific). The Pierce 660-nm protein assay was performed with NanoDrop 2000 spectrophotometer (Thermo Scientific), and the standard curve of protein concentration was obtained using BSA standard samples (Thermo Scientific). Reducing reagent Tris(2-carboxyethyl)phosphine hydrochloride (TCEP) and alkylating reagent 2-Chloroacetamide (CAA) were added to lysate to 10 mM and 50 mM, respectively, then incubated in dark for 30 min for alkylation. Alkylated samples were diluted to 6 M urea with 25 mM AmBic for LysC protease (Wako) digestion with protein/protease ratio at 40:1 (w/w) at 37^{o}C for 4 h with 250 rpm shaking, then diluted to 1 M urea with 25 mM AmBic for trypsin protease (Promega) digestion with protein/protease ratio at 40:1 (w/w) at 37^{o}C for overnight with shake. Digestion was quenched with 0.5% (v/v) formic acid (FA). Equal weights of SILAC samples were combined according to the label-swap strategy shown in Table.2.1, each combined sample contains 200 µg worth of samples.

Tryptic peptides for deep proteomic analysis were fractionated with a HPLC prior to LC-MS/MS analysis. The peptides were loaded with solvent A (10 mM AmBic 2% acetonitrile (CAN) in water, pH 8.3), and separated by a Zorbax 300 Extended-C18 column (2.1 x 150 mm, 3.5 µm) using a linear gradient (length: 100 min, 8% to 60% solvent B (80% ACN in water), flow rate: 200 µl/min). Fractions were collected every 1-min interval between 12 - 92 min of the gradient. The 80 collected fractions were combined at an even interval (1st, 21st, 41st, 61st fractions were combined, and so on), resulting in 20 fractions.

Mass spectrometry

Each fraction of tryptic peptides mixed with loading buffer (5% DMSO, 5% FA in water) in 10:1 (v/v). Loading mixture were analysed on a nanoUHPLC (Thermo) connected to a Q Exactive mass spectrometer (Thermo Fischer Scientific) through an EASY-Spray nano-electrospray ion source (Thermo Fischer Scientific). The peptides were trapped on a C18 PepMap100 pre-column (300 µm i.d. x 5 mm, 100 Å, Thermo Fisher) using solvent A (0.1% formic acid in water). The peptides were separated on an EASY-spray Acclaim PepMap analytical column (75 µm i.d. × 500 mm, RSLC C18, 2 µm, 100 Å) using a linear gradient (length: 120 min, 8 % to 28 % solvent B (0.1% formic acid, 5 % DMSO in acetonitrile), flow rate: 200 nL/min). The separated peptides were electrosprayed directly into the mass spectrometer operating in a data-dependent mode. Full scan MS spectra were acquired in the Orbitrap (scan range 350-1500 m/z, resolution 70000, AGC target 3e6, maximum injection time 100 ms). After the MS scans, the 20 most intense peaks were selected for HCD

fragmentation at 30 % of normalised collision energy. HCD spectra were also acquired in the Orbitrap (resolution 17500, AGC target 5e4, maximum injection time 120 ms).

Protein identification and quantification were performed using Andromeda search engine implemented in MaxQuant (1.5.0.35)[211]. Peptides were searched Human Uniprot database (version, January 2016). False discovery rate (FDR) was set at 1% for both peptide and protein identification. For deep proteomic analysis, data files from the same sample were assigned as near-by fractions, and match between run was turned on to match peak information from near-by fractions. Other options were set as defaults.

Proteomics quantitative analysis

Standard Student's T statistics was performed with software Perseus[212]. ProteinGroup files from MaxQuant search results were imported into Persues, proteins that have at least 3 valid values in at least one condition were subjected to log transformation, median normalisation, and imputation of missing values by background-level signal, then proceeded to paired T-test. All options for paied T-test were set as defaults.

Moderated T-statistics was performed with R programming. ProteinGroup files from MaxQuant search results were imported into R, proteins that have at least 3 valid values in at least one condition were subjected to log transformation, normalisation, and imputation of missing values. Normalisation was performed with vsn method using package *vsn*[213], imputation was performed by background-level signal using package *ImputeLCMD*[214]. Processed data was subjected to package *limma* for moderated T-test, with block variable set to replicate number of corresponding samples and correlation consensus calculated following the *limma* user guide. P-values were corrected for multiple testing by controlling the FDR with the Benjamini-Hochberg method.

Principal component analysis (PCA) was performed with R programming using base R functions. Data processed by log-transformation and normalisation was selected, and only proteins that have no missing values were proceeded to PCA.

Bioinformatic analysis

Pathway enrichment was performed with online platform STRING (string-db.org) using the "Multiple proteins" search option[215]. Gene set enrichment analysis (GSEA) was performed with STRING

platform using the "Proteins with Values/Ranks" option[215]. Results of *Reactome* pathway enrichment results were downloaded and plotted in R programming.

Enrichment in viral protein and viral RNA interactors were calculated by first matching significant proteins to a compilation of viral protein/RNA interactors. Viral protein interactors were compiled based on series of published large-scale GFP-pulldown experiments[147,161,164,206]. Viral RNA interactors were obtained from a published review article[169]. Next, enrichments were calculated using Fisher's exact test.

Materials

Chemicals and reagents

REAGENT or RESOURCE	SOURCE	IDENTIFIER
100 bp ladder	NEB	B7025
2-Chloroacetamide	Sigma Aldrich	C0267
6x purple loading dye	NEB	B7024S
Acetonitrile	Sigma Aldrich	#34851
Acrylamide/Bis-acrylamide 30%	Thistle Scientific	#20-2100-05
AEBSF hydrochloride	BioChemica	A14210100
Ammonium Bicarbonate	Sigma Aldrich	#9830
Bensonase	EMD Millipore	#70746
Calcium chloride	Sigma Aldrich	#499609
Chloroquine	ACROS Organics	#50-63-5
dialysed FBS	Silantes GmbH	#281000900
Dimethyl sulfoxide, DMSO	Thermo Scientific	T6508
DMEM	Gibco	#11995-065
FBS	Sigma Aldrich	F9665
Formic acid	Thermo Fisher	A117-50
Interferon alpha 2a	Pbl assay science	#11100-1
L-Arginine 13C,15N labeled	Silantes GmbH	#201604102
L-Lysine 13C,15N labeled	Silantes GmbH	#211604102
LysC protease	Wako	#129-02541
Milk powder	Sigma Aldrich	#70166
Opti-MEM reduced serum medium	Thermo Fisher	#31985062
PBS 1x, pH 7.4	Gibco	#10010023
PEG6000	Sigma Aldrich	#25322-68-3
penicillin/streptomycin	Sigma Aldrich	P4458
Phosphatase inhibitor	Sigma Aldrich	P0044
Pierce 660nm Protein Assay Reagent	Thermo Scientific	#22660
Pierce Bovine Serum Albumin Standard Pre-Diluted Set	Thermo Scientific	#23208
Protease inhibitor cocktail	Sigma Aldrich	P8340

SILAC DMEM media	Thermo Scientific	#10107883
Sodium Deoxycholate Detergent	Thermo Scientific	#89904
Sodium dodecyl sulfate	Sigma Aldrich	#151-21-3
Tris(2-carboxyethyl)phosphine hydrochloride	Thermo Scientific	#77720
Triton-X100	Sigma Aldrich	#9036-19-5
Trypsin EDTA	Sigma Aldrich	T3924
Trypsin, sequencing grade	Promega	V5111
Tween-20	Sigma Aldrich	P1379
Urea	Sigma Aldrich	U1250
X-tremeGENE 9 DNA transfection reagent	Sigma Aldrich	#6365787001

Instruments

INSTRUMENTS	SUPPLIER/INSTITUTE
API DeltaVision Elite widefield fluorescence microscope	Oxford Micron
Countess II FL Automated Cell Counter	Thermo Fisher
Odyssey Fc imaging system	LI-COR Bioscience
Nanodrop 2000 Spectrophotometer	Thermo Fisher
Trans-Blot Turbo Transfer system	Bio-Rad
CLARIOstar microplate reader	BMG labtech
LTQ XL Linear Ion Trap Mass Spectrometer	Thermo Scientific
UltiMate 3000 RSLCnano System	Thermo Scientific
Q Exactive mass spectrometer	Thermo Scientific

Antibodies

PRIMARY ANTIBODY	SUPPLIER	IDENTIFIER
β-actin	Sigma Aldrich	A2228
p24	Center for AIDS Reagents	ARP3279
anti-SINV C (304 and 306)	Laboratory of L. Carrasco	N/A
Stat1 (9H2) Mouse mAb	Cell Signaling technology	#9176
Phospho-Stat1 (Tyr701) (58D6) Rabbit mAb	Cell Signaling technology	#9167
SECONDARY ANTIBODY	**SUPPLIER**	**IDENTIFIER**
IRDye® 680LT Goat anti-Mouse IgG	LI-COR bioscience	#926-68020
IRDye® 680LT Goat anti-Rabbit IgG	LI-COR bioscience	#926-68021
IRDye 800CW Goat anti-Mouse IgG	LI-COR bioscience	#926-32210
IRDye® 800CW Goat anti-Rabbit IgG	LI-COR bioscience	#926-32211
IRDye® 800CW Goat anti-Human	LI-COR bioscience	#926-32232

Plasmids

PLASMIDS	SOURCE
pNL4-3.R-E- Nef-mCherry	Castello lab
pNL4-3.R-E- Gag-mCherry	Castello lab
pHEF-VSV-G	Sattentau lab

pcDNA5 eGFP	Castello lab

Cell lines

CELL LINE	SOURCE	IDENTIFIER
HEK293	ECACC	# 85120602
HEK293T	Rehwinkel lab	N/A

Chapter 3

Deep proteomic profiling of IFN-α stimulated cells reveals host factors that regulate cell defences against viruses

Introduction

Interferons (IFN) are a family of cytokines that activate cell innate immunity against viral infection in many animal species. Evidenced by genome sequencing and evolution studies, IFN and its receptors are likely functional in all jawed vertebrates[216]. In the human genome, there are 3 classes of IFN (type I, II, and III IFN) and each class is recognised by a specific receptor. Type I IFN consists of IFN-α, IFN-β, IFN-ε, IFN-κ, IFN-ω and interact with the receptor IFNAR1/2 heterodimer. Type II IFN is solely comprised of IFN-γ and interacts with IFNGR. Type III IFN was discovered more recently, it includes IFN-λ1, IFN- λ2, IFN-λ3 that are recognised by receptor IFNLR[217,218]. Almost all types of human cells are capable of producing type I and III IFNs in response to virus invasion[219]. By contrast, type II IFN is only produced by cells in the immune system, although it can also induce antiviral response in almost all somatic cells due to the ubiquitous expression of its receptor[220]. The strength of IFN response in somatic cells depend heavily on the binding affinity of each IFN with their receptors. IFN-β has the highest receptor-binding affinity (<0.1 nM) among all type I IFN, and could form stable complex with receptors during immunoprecipitation[221]. IFN-α comprises of 13 subtypes that are encoded by genes in a cluster in chromosome 9[222], and the binding affinities of these subtypes span 3 orders of magnitudes ranging from 0.4 to 200 nM[223]. IFN-β also has a strongest antiviral activity, which closely correlates with receptor-binding affinities for almost all type I IFN[224]. In another study, an engineered IFN-α with increased receptor-binding affinity successfully reproduced the high antiviral activity of IFN-β in a human amnion cell line[225].

Since the discovery of IFN in the 1950s as antiviral factor[226], researchers have been trying to understand the players that contribute to this regulatory process. In the past few decades, each time a new technology is introduced to this research field, we can gain new insights into IFN-mediated regulation. The earliest studies were carried out by comparing extracts of IFN-treated and not-treated cells through low-throughput biochemical assays, which leads to the identification of the first few known interferon-stimulated genes (ISGs)[227]. The inventions of oligonucleotide array (known as microarray)[228] and next generation sequencing (NGS)[229] provided powerful tools for simultaneous detection and quantification of numerous mRNA from cell extracts, thus opened a new path to study biological system at transcriptomic-level. When applied to IFN-related studies, researchers are able to understand how IFN induce changes in gene expression with high throughput and accuracy. This lead to identification of a plethora of ISGs that were not revealed in Northern-blot-based studies[230]. In online database *Interferome* that consolidates IFN-related transcriptome studies from various cell types and treatment conditions, thousands of genes are archived to be responsive to IFN at RNA level[231,232].

In contrast to the fruitful results in RNA-level studies, much less is understood regarding IFN-induced changes at protein level. Changes in mRNA abundance do not necessarily result in changes in protein abundance, as multi-omics studies have highlighted the limited coordination between regulations at transcriptome and proteome level[233,234]. This apparent discordance is likely the result of a wide variety of post-transcriptional and translational regulatory mechanisms[234]. Since a large proportion of genes involved in innate immune response carry out functions by their protein products, it is crucial to profile IFN-induced proteomic changes in order to understand the cellular environment that viruses encounter during infection. Advances in liquid chromatography-coupled tandem mass spectrometry (LC-MS/MS) based proteomic analysis have allowed profiling of proteome landscape of the cell to an increasingly higher depth of detection and accuracy of quantification. Early studies of IFN proteome have identified less than 1,400 proteins and determined dozens of IFN regulated proteins based on either measured fold change in one replicate[155], or T-test result in triplicate samples[235]; the number of protein identification has then be improved to 3,000 ~ 4,000 proteins in single-gradient MS analysis[150,156,236-238]. Nevertheless, each study provided insightful results even though the levels of detection depth in these studies are several folds shallower than studies in transcriptome, which typically profile over 10,000 RNA species in either microarray- or NGS-based studies. The main issue underlying this sharp contrast is the insufficient ability to detect changes in proteins of low abundance. This could be an issue of particular importance for the IFN field because proteins involved in IFN response are generally expressed at low levels in steady-state conditions, and their abundance, while increased in response to infection, would still be marginal when compared to highly abundant housekeeping cellular proteins.

In recent years, studies using orthogonal liquid chromatography to fractionate samples have allowed reduction of sample complexity prior to LC-MS/MS analysis, therefore enabled profiling of proteome landscape in great depth[142,239]. When applied to the field of IFN study, these methods have reached detection of over 6,000 proteins[240,241] to over 8,000 proteins[242], providing valuable insights into IFN responses at protein level in greater detail. However, the complete picture of IFN response at protein level remain far from being thoroughly elucidated. Studies in different cell systems have shown that IFN response is a complex system that could vary depending on IFN type and cellular context[243,244] For example, all 7 members of STAT transcription factor family could be activated by type I IFN, but STAT6 is only activated in B-cells[245], STAT1/3/5 are activated in promonocyte cells[246], STAT1-5 are activated in endothelial cells[247], and these differential responses result in regulation of different ISGs in cell-type dependent manner[244]. Therefore, whether a gene responses to IFN at protein level still need to be determined in the context of the cellular

environment of interest. Moreover, only part of the IFN-regulated proteins in these proteomics studies have known roles in innate immune or antiviral response[27-29], while functions and regulatory mechanisms of other proteins remain to be explored.

In previous chapter, deep proteomic profiling of HEK293 and HEK293T cells revealed thousands of proteins that are differentially expressed between these two cell lines and could contribute to their different permissiveness to virus infection. Proteins enriched in HEK293T cells were overrepresented in interactors of viral proteins and RNAs, while known innate immunity factors were more prevalent in HEK293 cells. These distinct expression patterns suggest that the cell permissiveness to virus is regulated by two separate populations of pro- and anti-viral factors. To further investigate how these factors regulate cellular environment, it was also necessary to assess their expression patterns in a virus-hostile environment under similar genomic background. IFN is a strong stimulus that drives cells into an antiviral state, therefore, we reasoned that analysis of the IFN-stimulated HEK293 cells would complement the existent analysis of steady-state HEK293 cells and virus-permissive HEK293T cells. In this chapter, I aimed to profile the proteomic landscape of IFN-stimulated HEK293 cells with a similar depth of detection as in HEK293/ HEK293T proteome in previous chapter. The deep proteomic data would itself be valuable to examine the protein-level responses upon IFN stimulation, and the comparison with HEK293/ HEK293T proteome would allow me to study the behaviours of virus-regulatory factors across different cellular states under similar genomic background.

Results

3.1 Serial titration defined time- and dose-dependent kinetics in IFN-α induced antiviral response.

Before profiling proteome landscape of virus-restrictive cellular environment, appropriate conditions for IFN stimulation needed to be determined. Interferon subtype IFN-α was chosen in this study for two reasons: 1, it belongs to type I IFN, which activates innate immunity for almost all somatic cells, whereas type II IFN is only produced and primarily used by immune effector cells in physiological condition; 2, it causes a much milder anti-proliferative effect than IFN-β in working concentration[225]. HEK293 cells were treated with IFN-α in different concentrations for overnight (20h) prior to infection. The same mCherry tagged SINV and HIV-$1_{R\text{-}E\text{-}}$ were used in this virus infection experiment as in previous chapter (Chapter 2, Section 2.1). Series of IFN-α dosages were tested for both SINV and VSV-G pseudotyped HIV-$1_{R\text{-}E\text{-}}$. Fluorescent signals from infected cells were measured in real-time using an isothermal culture condition. While high levels of fluorescence were observed in untreated cells infected with SINV, a close-to-background signal was detected for all the doses of IFN-α (Fig.3.1.1 A, left), showing a strong inhibition of SINV replication after 20h of IFN-α stimulation.

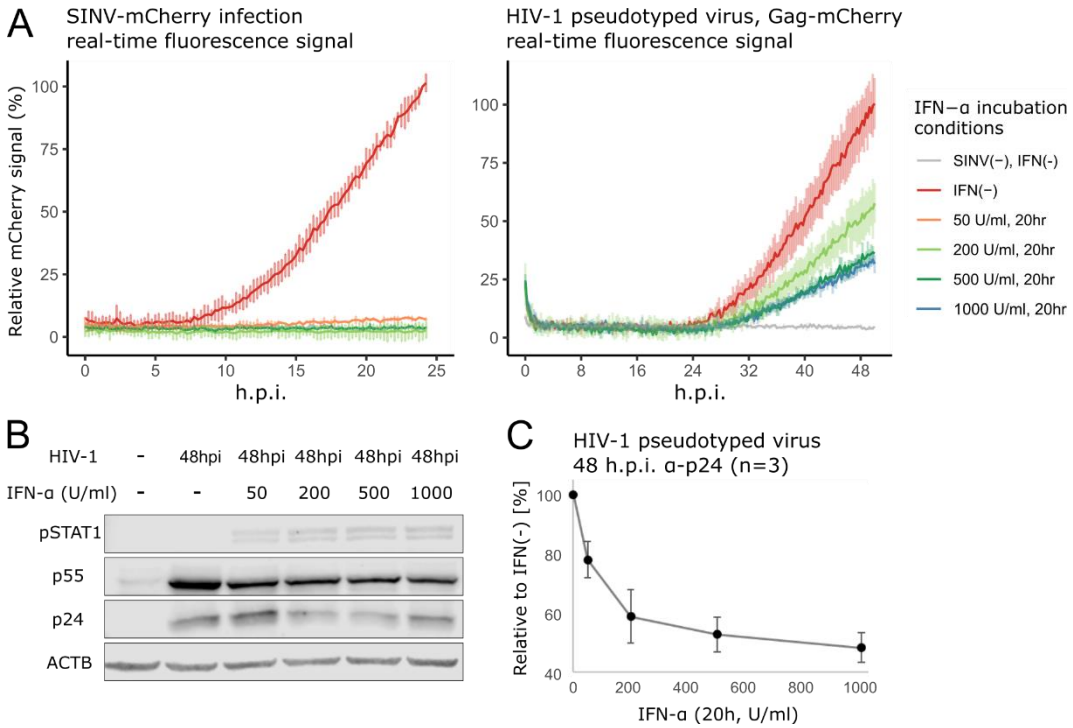


Figure.3.1.1 IFN-α induced antiviral response in HEK293 cells in dosage-dependent manner. (**A**), fluorescent signals from HEK293 cells challenged with mCherry-tagged SINV (MOI=1, left) or VSV-G

pseudotyped HIV-$1_{R\text{-}E\text{--}Gag\text{-}mCherry}$ (MOI=1, right) were measured every 15 min in culture condition in real-time using plate reader. IFN-α treatment conditions prior to infection were shown in colour labels. Signal readouts were normalised to signals at 24 or 48 h.p.i. (**B**), immunoblotting from HEK293 cells infected with VSV-G pseudotyped HIV-$1_{R\text{-}E\text{--}Gag\text{-}mCherry}$, probed against pSTAT1 and viral protein capsid (p24). HIV-1 p55 were probed also, because it is the precursor of p24. Cells were treated with IFN-α for 20h at indicated concentrations prior to infection. (**C**), summarised immunoblotting results from triplicate experiments, p55 signals were normalised to HIV-1 infected and IFN-α untreated samples.

By contrast, IFN-α treatment did not completely supress mCherry signals in HIV-$1_{R\text{-}E\text{--}Gag\text{-}mCherry}$ infected HEK293 cells, and the inhibition at 500 U/ml of IFN-α was similar in magnitude to that with 1000 U/ml, indicating a close-to-saturation level of antiviral activity (Fig.3.1.1 A, right). These results were verified by Western blot (WB) with antibodies against HIV-1 capsid (p24) and the IFN response marker pSTAT1 (Fig.3.1.1 B). Further quantification of WB result using normalised p55 signal revealed an exponential decrease in capsid abundance as IFN-α concentration increased (Fig.3.1.1 C). Interestingly, this dosage dependency of antiviral activity is similar to the kinetics of IFN-α induced STAT1 phosphorylation (Chapter 2, section 2.5), both results approached saturation at 1000 U/ml, and reached an "elbow" point in dosage-to-activity curve at 200 U/ml. Apart from activating antiviral response, IFN stimulation in high dosage and long duration is also known to induce an anti-proliferative effect[248]. Because the main goal in this chapter was to identify regulators of virus infection, 200 U/ml of IFN-α was chosen for downstream analysis to minimise IFN-induced anti-proliferative response.

The amount of time that cells were exposed to IFN-α prior to infection is also an important factor that governs cell defence against virus. To determine the time dependency of IFN-α stimulation, HEK293 cells were infected with mCherry-tagged SINV after exposed to IFN-α for different time frames, and subjected to plate reader fluorescence recording. The tested conditions included treatment times ranging from 10 min to 20h prior to infection, and concentrations from 50 to 500 U/ml. Fluorescence profiles from each condition was plotted separately after being normalised to mCherry signal at 24h.p.i, in IFN-α untreated cells (Fig.3.1.2 A). Fluorescent signals at 24h.p.i. were normalised to non-treated cells to estimate time-dependent kinetics (Fig.3.1.2 B). Interestingly, the resulting kinetic curves were highly similar to the dosage-dependent kinetic curve in a way that they both displayed a curve of exponential decrease. Compared to non-treated cells, cells with a short exposure to IFN-α were able to develop a significant antiviral activity. Notably, overnight exposure to IFN-α almost completely prevented SINV infection in all tested

concentrations. The observed time-dependent antiviral-activity kinetic was verified by WB using both SINV and HIV-$1_{R\text{-}E\text{-}}$, as blotting results showed a decrease in viral protein expression as IFN-α incubation time increased (Fig.3.1.2 C, D). To profile the changes in the whole-proteome after IFN-α induction, two time points were selected for downstream analysis: 4h IFN-α treatment was chosen as "early time point" because it reached "elbow" in time-to-activity curve, and 20h IFN-α treatment was the "late time point" as cells exhibited close-to-saturation antiviral activity.

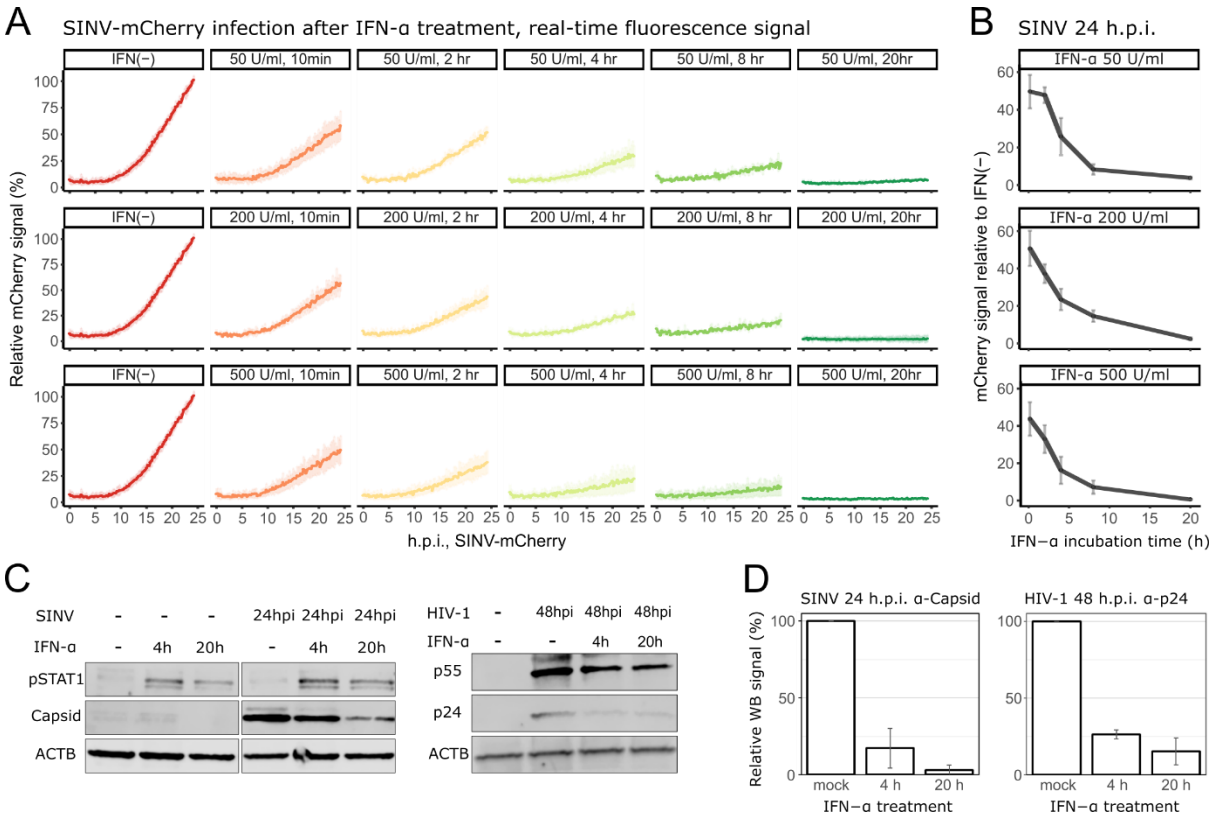


Figure.3.1.2 IFN-α induced antiviral response in HEK293 cells in time-dependent manner. (**A**), real-time fluorescent signals from HEK293 cells challenged with mCherry-tagged SINV (MOI=1). IFN-α treatment conditions are labelled above corresponding plots. Experiments were performed in triplicates, readouts were all normalised to signal at 24h.p.i. in IFN-α untreated cells. (**B**), relative mCherry signals at 24h.p.i. from fluorescent profiles in **A** were grouped by IFN-α dosage, and plotted against IFN-α treatment times. (**C**), western blots against viral proteins (Capsid for SINV, left; p24 for HIV-1, right) and IFN signalling marker pSTAT1 in virus-challenged HEK293 cells. Cells were treated with IFN-α at 200 U/ml for indicated amount of time prior to infection. (**D**), viral protein signals from western blot triplicate experiments were normalised to signals in IFN-α untreated conditions, and summarised into barplots. IFN-α treatments were all conducted at 200 U/ml for indicated amount of times.

3.2 Application of deep proteomic profiling with SILAC labelling improved detection of IFN-α stimulated proteome landscape.

After determining the conditions of IFN-α treatment, deep proteomic analysis was performed to depict IFN-α induced alterations at the whole-cell proteome level. HEK293 cells were cultured with stable isotope labelling in tissue culture (SILAC) for high-accuracy quantification in mass spectrometry (MS). Cells were labelled with light (L), medium (M) or heavy (H) isotope labels, and lysates were combined with label swap to avoid bias in isotope labelling (Table.3.1).

Table.3.1 label swap strategy in each replicate for SILAC-based IFN-α whole-cell proteome analysis

	Sample (isotope label)
Rep.1	mock (L) + 4h (M) + 20h (H)
Rep.2	mock (M) + 4h (H) + 20h (L)
Rep.3	mock (H) + 4h (L) + 20h (M)

A small fraction was collected from each sample prior to mixing, and subjected to MS analyses in short gradient to check isotope incorporation rates. The search result confirmed above 99% isotope incorporation rates in all samples (Table.3.2). Deep proteomic analysis was then achieved through off-line 2D-LC fractionation using reverse-phase separation (Fi.g3.2.1 A). Peptide separation profiles for off-line fractionation were visualised by absorbance at 214 nm. Chromatograms from triplicate samples were nearly identical, showing high reproducibility in separations of fractions (Fig.3.2.1 B). For search results of fractionated data, the MS/MS identification rate and the number of identified peptides were summarised by each fraction, and drawn in line plot with error bar representing standard deviations from triplicate data (Fig.3.2.1 C). MS/MS identification rate indicates percentage of identifiable peptide molecules in sample, and the observed high consistency proved high homogeneity of peptide abundance across fractions and replicates; number of peptide identification is a measure of peptide complexity in sample, and as shown line-dot plot, all fractions displayed high complexities in reproducible manner despite slightly lower number of identifications in first three fractions. Number of protein and peptide identifications from fractionated data were summarised by sample, and both metrics evidenced MS identification in high depth and consistency (Table.3.2). Apart from assessing reproducibility in

identification, the reproducibility in quantification was also checked by comparing log-transformed intensity profiles between replicates. As a result, Pearson correlation coefficients of all comparisons exceeded 0.95, evidencing a high reproducibility in quantification (Fig.3.2.1 D).

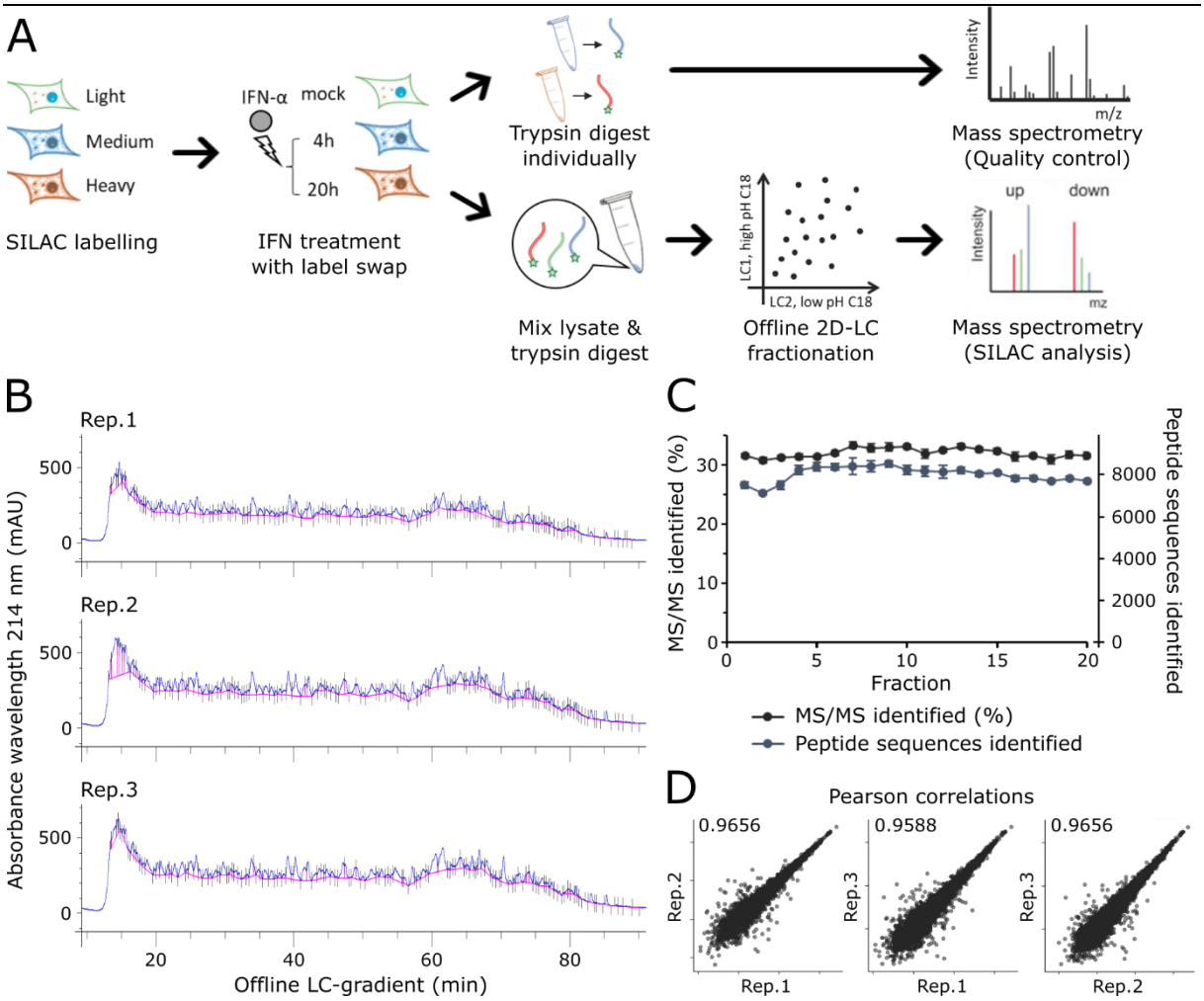


Figure.3.2.1 Deep proteomic profiling results of IFN-α treated cells were generated with high reproducibility. (**A**), schematic workflow of in-depth whole-cell proteome profiling that aimed to investigate IFN-α induced changes through SILAC analysis, and quality control experiment of each sample. (**B**), chromatogram of offline-LC gradient with reverse-phase separation column in high pH. Eluates from 10 to 90 min gradient were collected and divided into 20 fractions. (**C**), line plot of MS/MS identification rates and numbers of identified peptides in each fraction were drawn for quality control. Error bar represent standard deviation from three replicates. (**D**), combined intensities from L/M/H isotope channels were log-transformed and draw in scatter plot, Pearson correlation coefficients for each contrast pair were shown in corresponding plot.

Table.3.2 Quality control for IFN-α deep proteome profiling.

	mock, rep1, L	**4h, rep1, M**	**20h, rep1, H**	**mock, rep2, M**	**4h, rep2, H**	**20h, rep2, L**	**mock, rep3, H**	**4h, rep3, L**	**20h, rep3, M**
Protein identified	10587	10535	10542	10542	10539	10577	10537	10576	10538
Peptide identified	104764	104328	104459	104362	104443	104637	104346	104548	104240
Isotope incorporation (%)	99.953	99.775	99.499	99.772	99.580	99.953	99.606	99.906	99.835

After confirming the high reproducibility in MS identification and quantification, statistical test was performed to determine proteins of significant change in response to IFN-α stimulation. At first, the quantification result was subjected to Student's paired T-test, in which samples of different isotope labels within the same replicate were paired for direct comparison. Unexpectedly, the test result showed that only 3 proteins were determined "significant" by cut-off line of FDR<0.1, despite many proteins possessing large fold changes, low p values, and were appreciably outlying the non-changing hits (Fig.3.2.2 A-[1]). Upon closer examination, most of these non-significant hits were found to be known ISGs with well characterised functions in cellular antiviral response, including virus sensors RIG-I (DDX58), transcription factors STAT1 and IRF7, antiviral effectors PKR (EIF2AK2) and protein families OAS and MX proteins. This suggested that IFN-α induced proteomic changes were genuine but suffered of high incidence of false negative under Student's paired T-test.

Moderated T-test by *R* programing package *limma* was then applied as an alternative to Student's T-test as it was developed for large datasets that have small sample size[203,249]. Processed intensities were either subjected to the moderated T-test directly, or first applying "replicate" as a source of batch effect for correction (the same method as in HEK293/HEK293T test in Chapter.2, Section.2). The number of significant changes increased to 22 with moderated T-test by default settings, and to 36 when applied moderated T-test with batch effect accounted. When visualised in volcano plot, results from both methods showed improvements as more proteins with large fold changes and low p-values were assigned as "significant" hits (Fig.3.2.2 A-[2,3]). Dozens of functionally well-characterised ISGs were included among these improved hits. However, in both methods, a substantial amount of proteins that show low p-values and high fold changes remained not significant. These proteins included well characterised ISGs, such as ADAR and IRF7. Therefore, this pipeline was still affected by false negatives despite considerable improvement compared with Student's paired T-test.

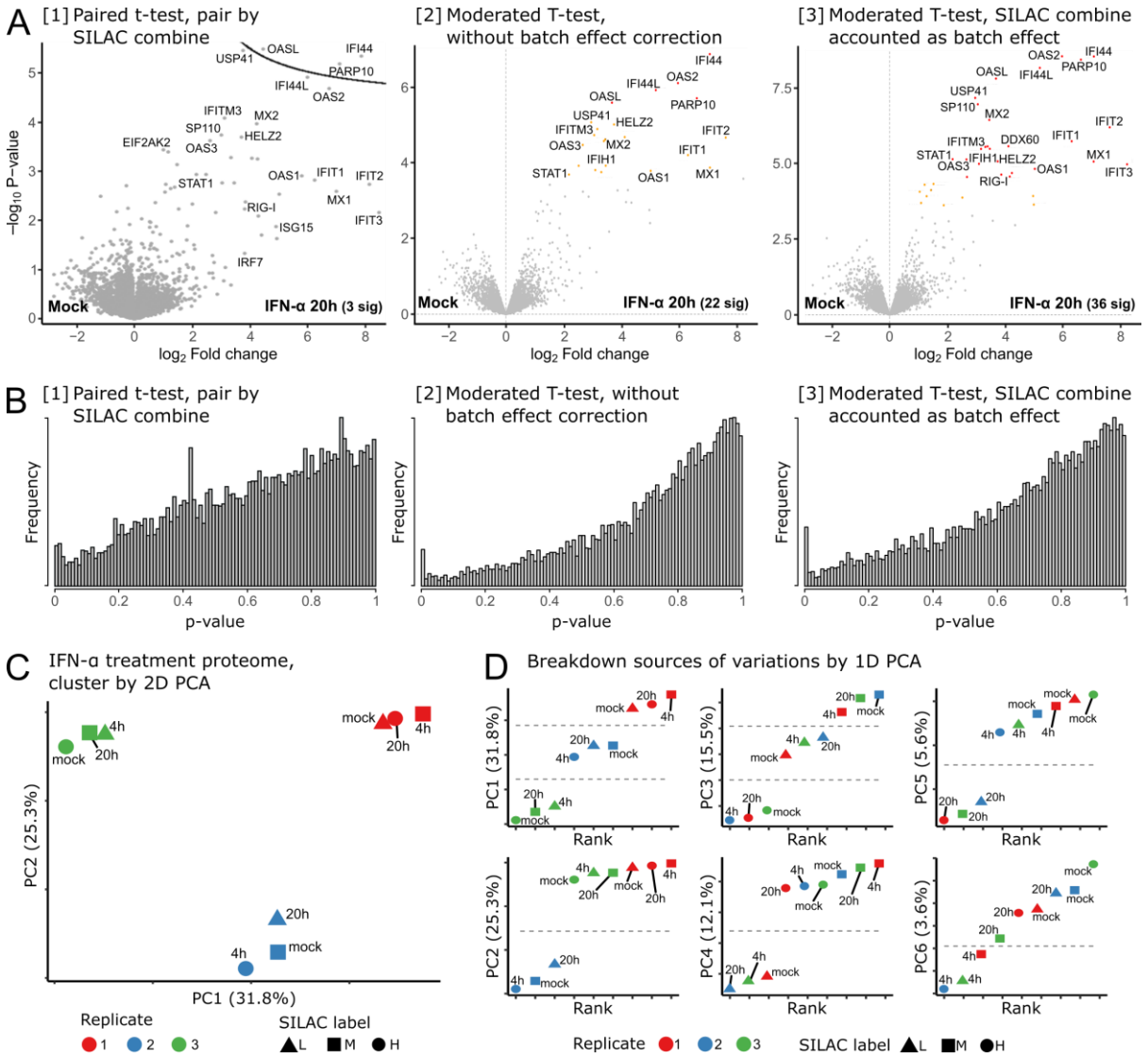


Figure.3.2.2 Conventional statistical tests faced issues from multiple layers of batch effects. (**A**), volcano plots showing deep proteome data processed through different pipelines: [1] Student's paired T-test paired by replicate using Perseus, cut-off for FDR<0.1 shown in curve; [2] *limma* moderated T-test without additional processing; [3] moderated T-test modelling replicate as batch effect, cut-off for significant hits shown in colour label (red for FDR<0.01, orange for FDR<0.1). Well characterised interferon-stimulated genes (ISGs) were text labelled. (**B**), histograms of p-values from corresponding test results. (**C**), principle component analysis (PCA) of IFN-α whole-cell proteome, plotted in 2D scatter plot showing first two principle components with percentage of variance explained in respective directions. (**D**), contributions of the first six principle components plotted separately, with y-axis showing the percentage of variance explained in the respective direction, and x-axis ranking 9 samples by their rotated value in the respective PC. Grey dash lines were drawn arbitrarily to highlight separations of samples in each 1D-PCA plot.

Distribution of p-values was checked to diagnose these test results. In a test where the "null" hypothesis is true (no genuine difference exists between tested groups), the p-value distribution would appear a uniformed flat line; when the test subjects contain genuine differences in the tested direction, p-values distribution would display a peak close to the origin as more hits have low p-values. In Student's paired T-test results, the p-value histogram did show a small spike close to 0, but the distribution were generally inclined toward high-p-value region instead of forming a flat line (Fig.3.2.2 B-[1]). P-value histograms from the other two methods showed similar trend, with the main difference that the peak close to the origin was sharper (Fig.3.2.2 B-[2,3]). The observed trend that p-value histogram inclined toward p-equal-1 indicated that the major source of non-random variances in the data was not in-line with the tested direction (mock versus IFN-α treated). This type of confounding variations was also generally referred to as batch effect. Given that the test with "replicate" accounted as batch effect still produced p-values that inclined toward p-equal-1, there could be another layer of batch effect that affected test results.

To assess sources of variations, principal component analysis (PCA) was performed on the proteomic data. When plotting first two principal components into 2D scatter plot, all samples tightly clustered into three groups by replicate, and there was no clear separation by IFN-α treatment condition (Fig.3.2.2 C). This was expected, because the principle of SILAC labelling experiment is to purposely group experimental errors within replicates into errors between replicates to partition-off this part of errors from other random errors. Contributions of top 6 principal components were further checked in scatter plot. In the first two principal components (PC1 & PC2), samples were all separated into three groups by replicate as expected (Fig.3.3.2 D, left). Surprisingly, the third and fourth principal components (PC3 & PC4) showed that samples were separated by their SILAC isotope labels (Fig.3.3.2 D, middle), whereas in PC5 & PC6, samples were separated by their IFN-α treatment conditions (Fig.3.2.2 D, right). This breakdown view of PCA result confirmed that the IFN-α treatment induced alterations in proteome were genuine, but confounded by two layers of batch effects – "replicate" and "SILAC label".

Several attempts had been made to resolve this issue. First, I focused on data normalisation. The method I used was variance stabilisation normalisation (VSN), which was built based on analytic continuation of log-transformation in the entire real number field[213]. It yields a result similar to medium normalisation in high-signal region, and performs better in reducing variations in low-signal region. Originally developed for microarray analysis, this method has been demonstrated applicable and powerful in transcriptomic and proteomic data analyses. However, I suspected that the VSN method might be inappropriate in analysing SILAC data. L/M/H isotope labelled samples were combined at an early step of sample processing to account for between-replicate variance, but

normalisation by VSN method would take all samples as input in an undistinguished manner. This might risk blending the between-replicate variance with the within-replicate variance, leading to underestimation of between-replicate variance and exaggeration of within-replicate variance.

To address this potential issue, I developed a novel method that normalise between-replicate variance and within-replicate variance separately. Briefly, L channels from three replicates were first aligned by medium value to normalise for between-replicate variance, then M/H channels were normalised to L channel in the respective replicate to normalise for within-replicate variance (Fig.3.2.3 A). This is named Light-channel aligned normalisation (LAN) in following discussions. When examined with boxplots, intensity distributions of samples processed by LAN was almost identical to that by VSN (Fig.3.2.3 B). To assess whether LAN method showed improvement in reducing variance of the data, intensities from IFN-α 20h triplicate samples were normalised using different methods, and standard deviation (sd) of each protein were calculated and plotted against the corresponding intensity. As a result, data without normalisation showed a more scattered sd-to-intensity plot and the average sd for all proteins were 0.565; in comparison, VSN processed data resulted in a lower average sd of 0.491, and LAN processed data produced a similar average sd of 0.489 (Fig.3.2.3 C). Subsequently, I performed the moderated T-test and the result of LAN-processed data was very similar to that of VSN processed data as shown in volcano plot (Fig.3.2.3 D). Under a cut-off line of FDR<0.1, LAN method determined 24 significant changes, which is fewer than the VSN method. Distribution of p-values of LAN-processed data was also almost identical to that of the VSN method. To summarise, the performance of LAN was comparable to VSN in all assessed aspects, and was unable to resolve the identified issue when performing the statistical test.

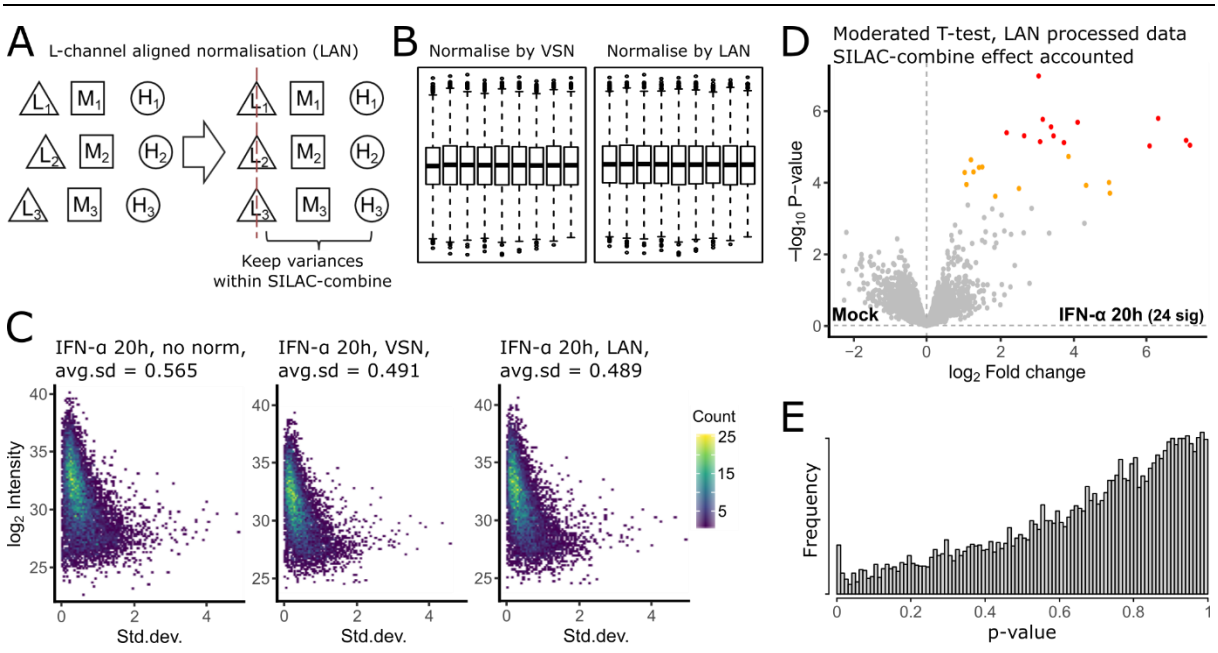


Figure.3.2.3 Novel normalisation method did not improve test result. (**A**), schematic workflow of normalisation by L-channel. (**B**), boxplot comparing intensity distributions of all samples under different normalisation methods. (**C**), scatter plot of each protein's standard deviation versus intensity, only results from IFN-α 20h triplicate were included, data were either non-normalised (left), normalised by VSN method (middle), or by L-channel aligned normalisation (LAN) introduced in panel **A**. (**D**), volcano plot of moderated T-test using data normalised by LAN method, with colour label on significant hits (red for FDR<0.01, orange for FDR<0.1). Replicate was accounted as batch effect. (**E**), histogram showing p-value distribution of the test result.

Next I sought to find a test method that allows incorporation of multiple layers of batch effects, so that the effects from replicate and isotope-label could be accounted for simultaneously. Student T-test and moderated T-test were both based on linear modelling, where intensity and experimental design were input as matrix Y and X, error term ε was calculated based on intensity matrix assuming that variations follow normal distribution, and fold change was estimated by coefficient β (Formula.3.1). Linear model allows incorporation of multiple batch effects as co-variables in the design matrix X, and these effects share the same error term with the main variable.

Formula.3.1 Linear model

$$Y = X \cdot \beta + \varepsilon\,,$$
where $X = design\,matrix,$
$\beta = coefficient\ (logFC),$
$\varepsilon \sim N(0, R)$

However, when there are multiple layers of batch effects, variations from each batch effects might not have the same distribution as that from the main variable. Therefore it could be inappropriate for this SILAC data. Linear mixed model (LMM) is a method that separately models main variable and batch effects – it models batch effects in matrix Z and calculates its error term μ separately (Formula.3.2)[250].

Formula.3.2 Linear mixed model

$$Y = X \cdot \beta + Z \cdot \mu + \varepsilon\,,$$
where $X = design\,matrix,$
$\beta = coefficient\ (logFC),$
$Z = batch\ effect\ design\,matrix,$
$\mu \sim N(0, R_1),\ \varepsilon \sim N(0, R_2)$

The difference in test rationales was also visualised in schematic diagrams (Fig.3.2.4 A) – for batch

effects-confounded data where linear modelling failed to detect significant changes, LMM allows separation of batch effects and testing in the correct direction.

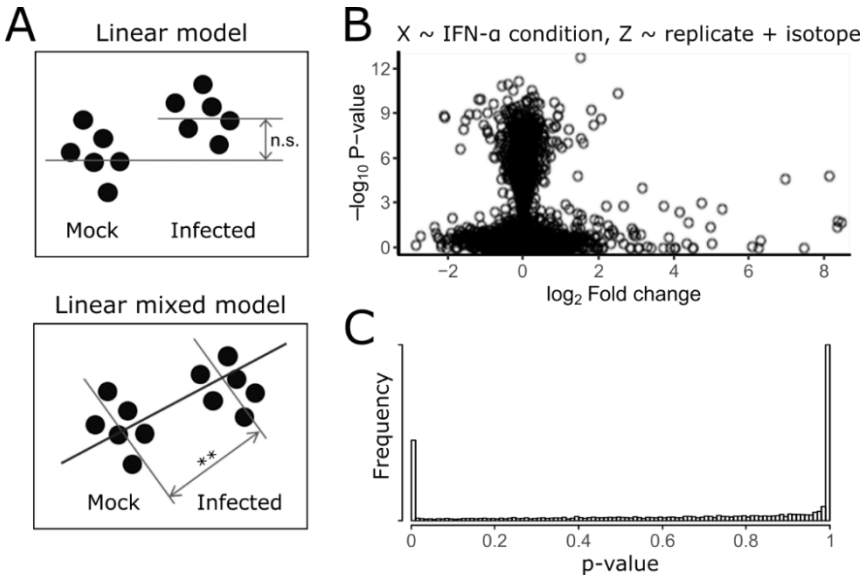


Figure.3.2.4 Linear mixed model (LMM) did not correctly accounted for multiple batch effects. (**A**), diagram showing principles of linear model and linear mixed model in T-test. (**B**), volcano plot of LMM test result, design matrixes for tested variable (X) and batch-effect variables (Z) were labelled. (**C**), histogram showing distribution of p-values from the test result.

Normalised and imputed intensity data were subjected to LMM where the experimental condition was modelled as main variable, and two layers of batch effects were modelled as co-variables. The test result, however, clearly had a different problem when visualised in volcano plot – significance for proteins of low fold changes were falsely boosted up, resulting a fountain-like shape (Fig.3.2.4 B). These false positives for low-fold-change hits resembled overfitting. When checking p-value distribution, it was evident that the batch effects were not correctly accounted as a sharp peak close to p-equal-1 emerged (Fig.3.2.4 C). Therefore, LMM method was also unable to address the statistical issue.

Given that Bayesian-based variance moderation had been shown to be powerful in proteomics data analysis, I next explored a method that allows variance moderation while modelling multiple batch effects. The limma package provides a powerful tool to achieve variance moderation, but it was initially built for microarray (MA) data analysis, therefore many of its functions were designed for MA-specific issues. The “block” argument in *limma* moderated T-test is designed to account for batch effect, because a major source of batch effect in MA experiment is the position on

which oligonucleotide probes are printed on silicon plate, and the signal variance between probes on the same "block" of silicon is typically smaller than variance between probes printed on different blocks. This "block" argument models the designated batch as co-variable in linear model, and has been shown to have outstanding performance251. However, it only allows modelling of one batch effect. To model the second layer of batch effect (SILAC label) for moderated T-test, I incorporated it into a design matrix together with the IFN-α treatment condition (Fig.3.2.5 A). This measure also

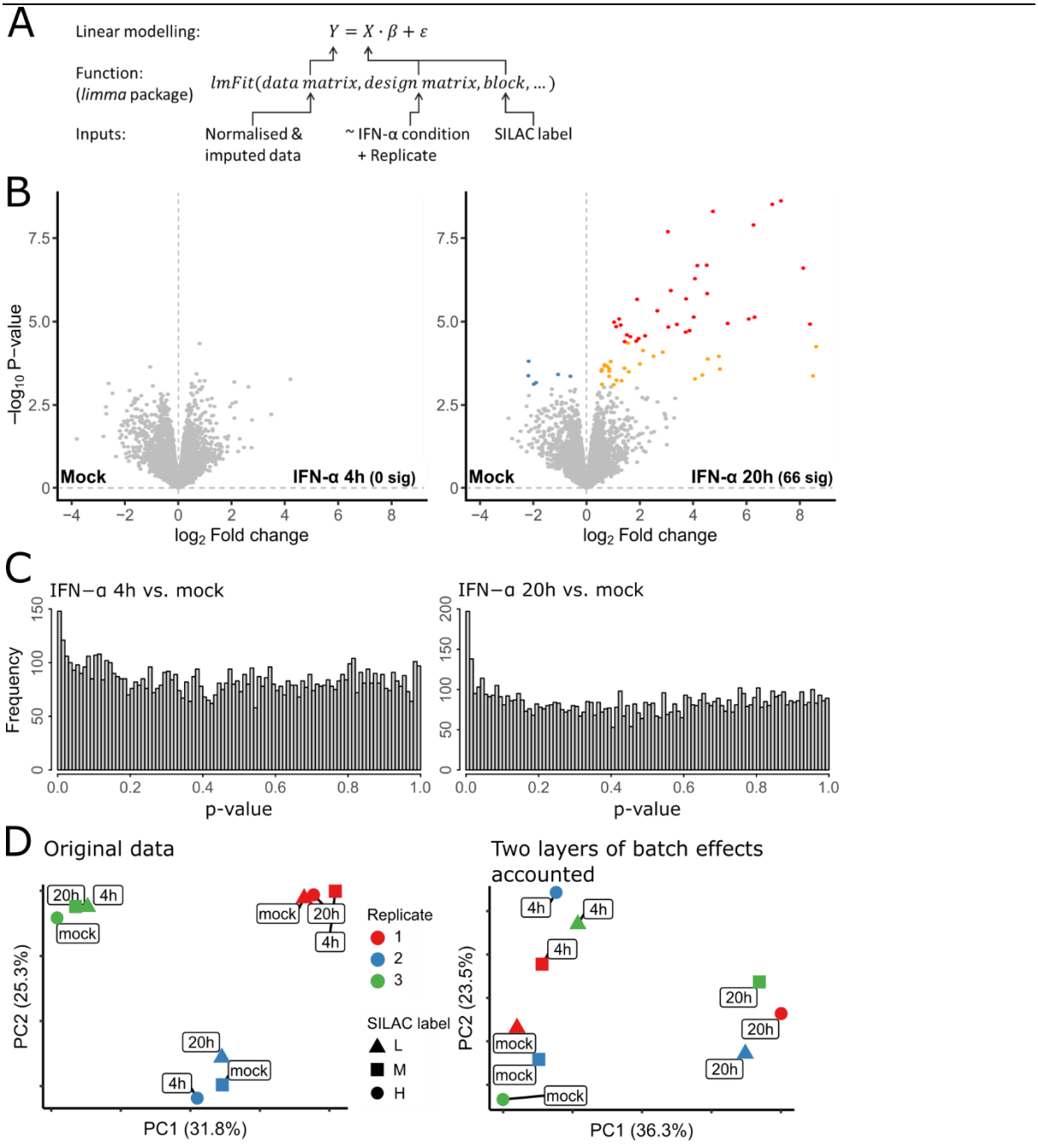


Figure.3.2.5 Modelling two co-variables in moderated T-test successfully accounted for two batch effects and revealed more IFN-α regulated proteins. (**A**), schematic workflow of moderated T-test

modelling both layers of batch effects. (**B**), volcano plots showing test results of IFN-α proteome using pipeline introduced in **A**. (**C**), histogram showing p-value distribution of corresponding test results. (**D**), PCA plot showing composition of variations before and after two layers of batch effects accounted.

---

allowed batch effect to be modelled as co-variable, but unlike default method that use "block" argument, the variance from this co-variable would not be moderated. Results of the test using this method, as visualised in volcano plot, were substantially improved compared with all previous methods (Fig.3.2.5 B). By FDR<0.1 cut-off , 66 proteins were determined to show significant changes after 20h of IFN-α treatment compared with mock, whereas 4h of IFN-α treatment showed no significant protein-level changes.

Distribution of p-values demonstrated the successful modelling of all batch effects – p-value frequency formed a peak close to origin, and maintained a flat line in other regions (Fig.3.2.5 C). PCA plot also showed that when both layers of batch effects were properly modelled, and as a result, the IFN-α treatment condition became the primary driver of variances of the data (Fig.3.2.5 D). After exploring multiple methods, the statistical issue in IFN-α deep proteomic profiling was solved using linear modelling with Bayesian-based variance moderation and two layers of identified batch effects incorporated as co-variable.

3.3 Quantitative proteomic analysis reveals proteome level changes at both early- and late-stage of IFN-α stimulation.

The optimisation of the data analysis pipeline afforded a considerable improvement in number of significantly altered proteins. From the volcano plot (Fig.3.3.1 A), it can be seen that a large number of functionally well-characterised ISGs were present among these significant hits. These included viral RNA sensors RIG-I (DDX58) and MDA5 (IFIH1), IFN signalling factors STATs and IRFs, Ubiquitin-like protein ISG15 and its E3 conjugators HERC proteins, and various antiviral effectors such as viral RNA editing enzyme ADAR, viral protein translation suppressor PKR (EIF2AK2), viral DNA synthesis suppressor SAMHD1, and antiviral protein families IFIs, MXs, OASs (Fig.3.3.1 A). To distinguish known ISGs from proteins that have no known functions in IFN response, these genes were mapped to the MA-based ISG database *interferome* and gene ontology (GO) annotation database[252]. Genes shown to be regulated by IFN-α in MA database were classified as "ISG by MA

database”; genes bearing GO terms related to innate immunity or IFN-α were classified as “ISG by GO annotation”. As visualised in the volcano plot (Fig.3.3.1 A, right), most of the IFN-α upregulated proteins had known relation with IFN-α by either MA database (19 proteins), annotation in GO terms (1 protein), or both evidences (36 proteins), whereas 4 proteins were not linked to IFN-α according

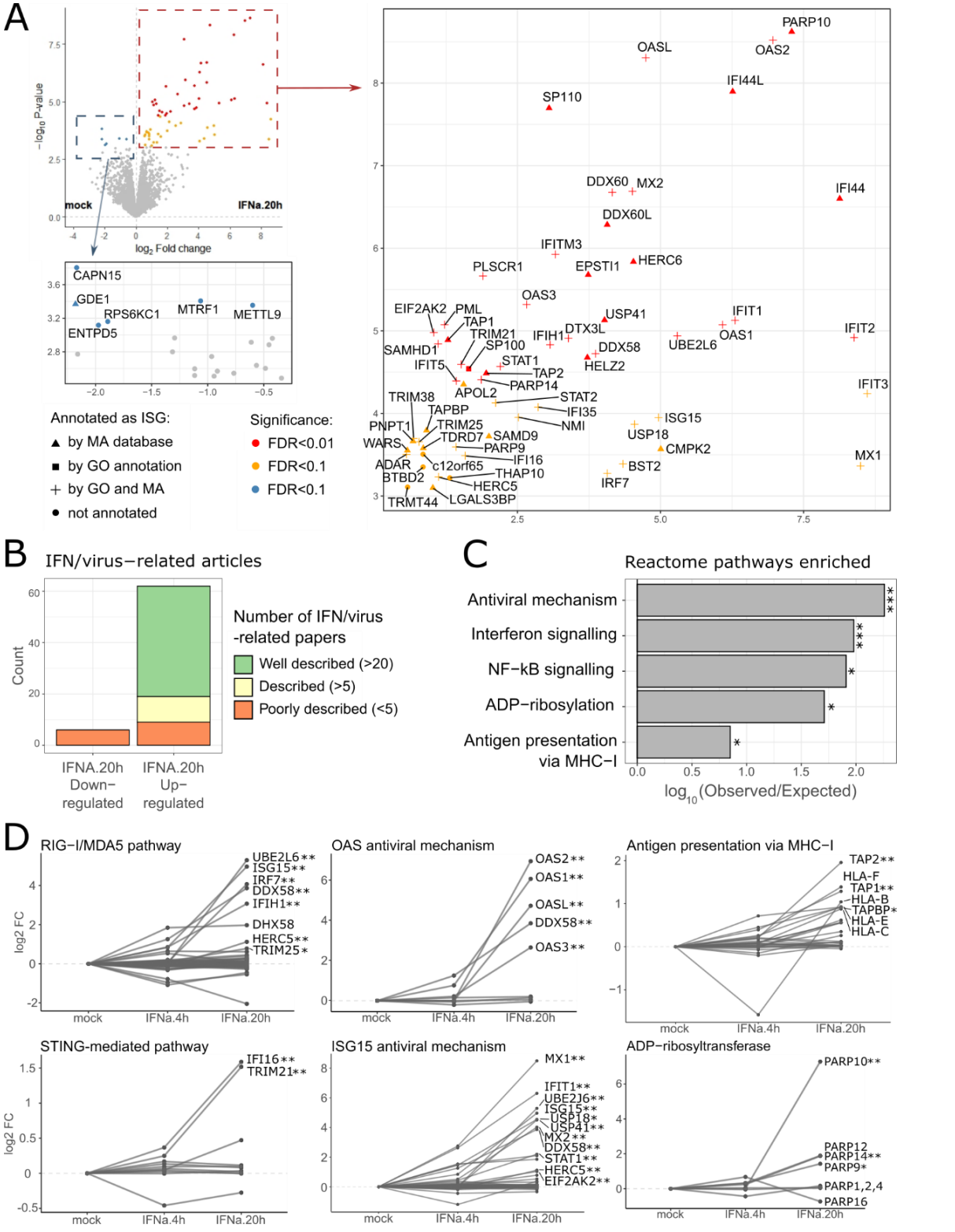


Figure.3.3.1 Enrichment analysis and text-mining revealed both known and novel proteins regulated by IFN-α. (**A**), volcano plot with text label showing proteins of significant changes after 20h IFN-α treatment. Proteins known to be related to IFN-α were drawn in different shapes – triangle: regulated by IFN-α in microarray (MA) studies; square: response to IFN-α in GO annotation; cross: related to IFN-α in both MA studies and GO annotation; circle: not known to be related to IFN-α. (**B**), boxplot showing count of IFN-α- or virus-related research articles in PubMed for proteins of significant changes in the proteomics data. (**C**), barplot showing enrichment in *Reactome* pathways, with strength of enrichment represented by ratio of observed count versus expected count by random selection, and p-value of enrichment labelled by * (*: p<0.05, **: p<0.01, ***: p<0.001). (**D**), dynamics of expression levels shown in parallel coordinate plots for genes in pathways of interests. Significance of changes in proteomic data were labelled by * (*: FDR<0.1, *: FDR<0.01).

---

to any of these data sources. Interestingly, most of the downregulated proteins had no known relation with IFN-α in either RNA level (MA database) or curated functional studies (GO annotation), implying that these protein downregulations were poorly understood in their relations with IFN-α (Fig.3.3.1 A, left).

Not all proteins that have role in cellular response to virus are known to be IFN-related. To further explore IFN-α regulated proteins that have known role during virus infection, a text mining was performed to search the PubMed library. Briefly, a research article of a given gene would be determined “IFN/virus-related” if it contains interferon- or virus-related keywords (excluding irrelevant words such as “lentiviral transduction”) in the abstract. Based on number of IFN/virus-related articles, genes were divided into three groups (“Well described”, “Described”, and “Poorly described”). Boxplot visualisation showed that a large proportion of IFN-α upregulated proteins had well described roles in antiviral response, whereas dozens of proteins were less well-described based on article counts (Fig.3.3.1 B). These less well understood genes includes part of IFN-α upregulated proteins and all of IFN-α down-regulated proteins.

Functional investigation of IFN-α regulated proteins were also performed at the pathway level. Enrichment analysis of *Reactome* pathways revealed that, as expected, pathways known to exert IFN signalling and antiviral activity were highly enriched (Fig.3.3.1 C). Members in NF-κB signalling were also enriched, which is consistent with current understanding that it is widely involved in cell stress and immune responses[253]. Apart from conventional innate immunity pathways, this analysis also highlighted IFN-α regulated proteins that have no well-defined role in innate immunity and showed interesting expression patterns. The entire MHC-I antigen peptide transporting complex (TAP1/2 and protein chaperon TAPBP) were found upregulated at protein

level, and multiple HLA proteins (subunits of MHC-I complex) also had large fold changes at late timepoint. Previous studies have described links between IFN response and adaptive immunity in cells of the immune system, for example type I IFN induces lymphocyte differentiation[254], IFN-α activates natural killer cells through dendritic cells[255], and formation of immunoproteasome which facilitates MHC-II antigen presentation[256]. In-line with what we observed here in HEK293 cells, IFN-α is also reported to upregulate MHC-I complex in other somatic and tumour cells[257,258], but less is known regarding the role of IFN-α in regulating antigen presentation and adaptive immunity in non-immune system cells. PARP family members are ADP-ribosyl transferases that catalyse post-translational modification of one or multiple copies of [N^6-(ADP-D-ribosyl)-] group on lysine residue. Most members in PARP family, which catalyse ADP-ribosylation process, are recently shown to be involved in regulation of innate immunity and inflammation[259]. Different PARP proteins appear to have a multifaceted role in innate immunity, for example PARP7 suppress IFN production by inactivating TBK1[260], PARP9 forms complex with ISG DTX3L to promotes STAT1 phosphorylation[261], and PARP14 enhances IFN promoter by promoting histone acetylation[262]. In this dataset, 4 out of 8 detected PARPs showed significant or close-to-significant (FDR~0.11) changes after IFN-a stimulation, and notably, all of these upregulated PARPs (PARP9/10/12/14) are reported to have antiviral activities[261-263].

Besides pathway-centred analysis, dynamics of protein-level responses was also explored at the whole-proteome level. IFN-α treatment response was divided into initial response and progressive response, defined by fold change from mock to early time point (4h post-treatment) and fold change from early to late time point (20h post-treatment). Proteins were divided into 8 clusters according to their patterns in initial and progressive responses to IFN-α: 1), inhibition at early time point followed by negative feedback; 2), induction after delayed response; 3), early and continuous upregulation; 4), early inhibition and reached plateau; 5), early induction and reached plateau; 6), early and continuous downregulation; 7), downregulation after delayed response; 8), early induction followed by negative feedback (Fig.3.3.2 A). Enrichment in *Reactome* pathways were then calculated for each dynamic group and visualised in barplots (Fig.3.3.2 B). Expectedly, proteins involved in IFN signalling and antiviral mechanisms were enriched in group 2 & 3. However, it was surprising that group 4 & 5, which largely comprised proteins that were regulated by IFN-α at early time point and maintained a plateau at late time point, was not enriched in any known pathways. Other groups also showed no pathway enrichment, indicating lack of understandings in processes involved in IFN-α-induced protein-level downregulation and temporal regulation.

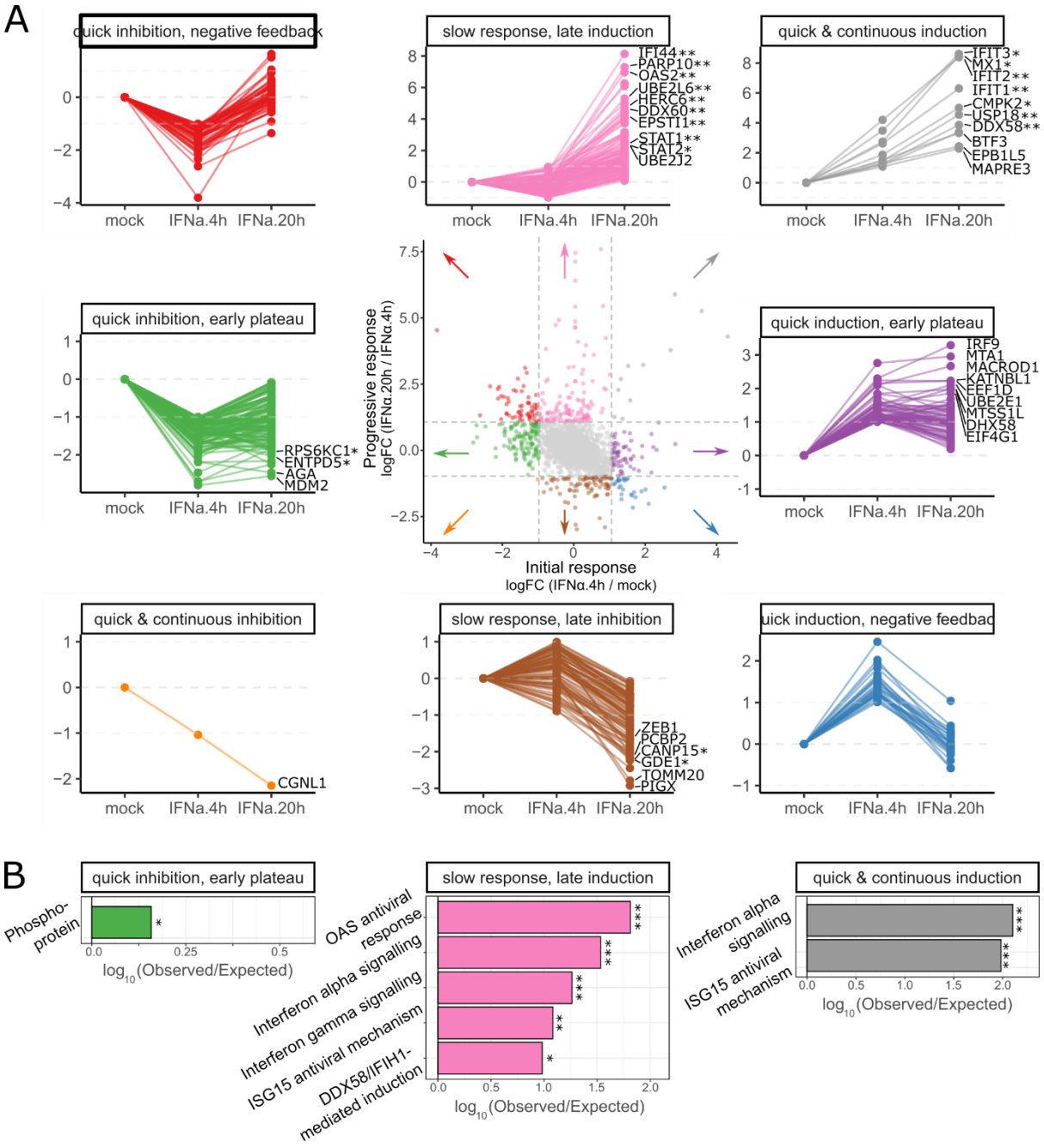


Figure.3.3.2 IFN-α induced proteome-level changes displayed distinct groups in dynamics. (**A**), scatter plots drawing IFN-α induced initial responses against progressive responses (middle). Dynamic changes were divided into 8 groups based on cut-off line at log-2 fold change of 1, and were displayed in parallel coordinate plots. (**B**), barplots showing enrichment of *Reactome* pathways in different dynamic groups, with p-values of enrichment labelled in * (*: p<0.05, **: p<0.01, ***: p<0.001). Dynamic groups not plotted here contained no significant enrichments.

Generally, functionally-associated genes are likely associated in expression levels in varies regulatory processes. Base on this trend, proteins that showed similar expression patterns with

known ISGs were also likely to be functionally associated with IFN-α. To explore proteins that had similar expression pattern with known ISGs, hierarchical clustering was performed on proteins that had larger than 2-fold change in proteomics result (Fig.3.3.3 B). Fold changes were scaled to unit variant prior to clustering, and cut-off for cluster number was set at 6 based on "elbow" method[264] (Fig.3.3.3 A). Expression patterns in each cluster were further visualised in parallel coordinate plots where scaled intensities were plotted against corresponding IFN-α treatment conditions (Fig.3.3.3 C). Similar to dynamic group analysis, hierarchical clustering result also distinguished proteins that were down-regulated (cluster 1 & 2), up-regulated (cluster 3 & 4), or having negative feedback

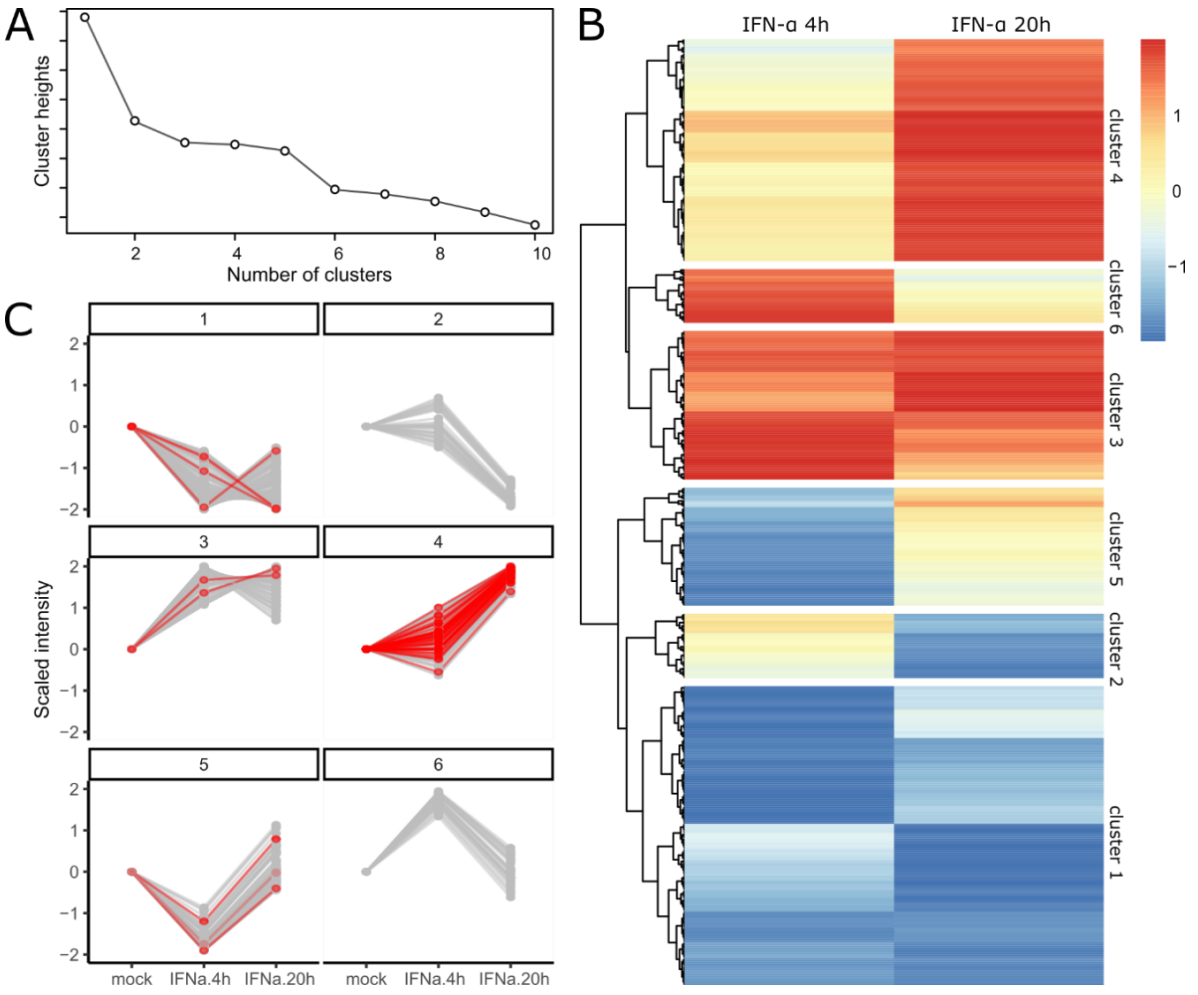


Figure.3.3.3 Clustering analysis highlighted proteins that share similar regulation pattern with known ISGs. (**A**), scatter plot determining optimal number of clusters by "elbow" method. Number of clusters was plotted against corresponding heights in Euclidean distance. (**B**), hierarchical clustering result displayed in heatmap. (**C**), clustering results displayed in parallel coordinate plot, with genes annotated as ISG by GO terms highlighted in red lines.

(cluster 5 & 6). Known ISGs were determined based on their GO annotations, and were highlighted red. Interestingly, most of the known ISGs were assigned into cluster 4, where proteins were up-regulated by IFN-α at late time point after an initial delay. Interestingly, this cluster also consisted 86 proteins that had no GO term linked to innate immunity. These included proteins highlighted in previous pathway analysis, such as TAP1/TAP2 in antigen presentation and PARP10/PARP12 in ADP-ribosylation, and genes of no known role in innate immunity. Surprisingly, only two proteins (IRF9 and DHX58) were classified as "known ISG" by GO annotation in cluster 3, where proteins were induced rapidly at early time point and remained at plateau, and other proteins in this cluster were not enriched in any GO biological processes. This observation was consistent with that in dynamic group analysis – most of the proteins in innate immunity pathway only took over at a later time point post IFN-α treatment, and processes rapidly responsive to IFN-α were less understood.

3.4 Comparative analysis of IFN-α regulated proteome and HEK293/HEK293T proteome highlighted common regulators of virus infection across different cellular states.

Deep proteomic analysis of IFN-α stimulated HEK293 cells profiled the proteome landscape of a virus-restrictive cellular environment. Together with proteomics of HEK293 and HEK293T cells presented in previous chapter (Chapter.2, Section.3), these data revealed proteome landscapes from a continuum of cellular states – ranging from virus-permissive (HEK293T), to less virus-permissive (HEK293), then to virus-restrictive (IFN-α stimulated HEK293). Direct comparison was performed on test result of IFN-α treated versus mock-treated HEK293 cells, and result of HEK293 versus HEK293T cells. Due to difference in depth of MS identification (8,815 proteins in HEK293/HEK293T proteome, 10,155 proteins in IFN-α proteome), only proteins identified in both datasets were compared. The intersection of proteins enriched in HEK293 over HEK293T cells and proteins enriched in IFN-α stimulated cells was highlighted in the zoom-in view of the scatter plot (Fig.3.4.1 A). These intersecting proteins were mainly composed of known players of innate immunity pathway, included viral RNA sensor RIG-I (DDX58), viral DNA sensor IFI16, E3 ubiquitin ligase family proteins HERC5 and TRIM21/25/38, ISG-driving transcription factor STAT1, and antiviral effectors OAS3, PKR (EIF2AK2), IFIT5. This falls in-line with the observation in previous chapter that HEK293T cells are depleted of factors in every step of innate immune response. Apart from these ISGs, proteins involved in antigen presentation (TAP1, TAP2 and TAPBP) and ADP-ribosylation process (PARP14) also had higher expressions in HEK293 cells than in virus-permissive HEK293T cells, and higher expressions in IFN-α stimulated than in steady-state HEK293 cells. This expression pattern implied that, apart from

participating in IFN response, members in these two processes could also confer resistance to virus on cells without exogenous immunomodulatory cytokines.

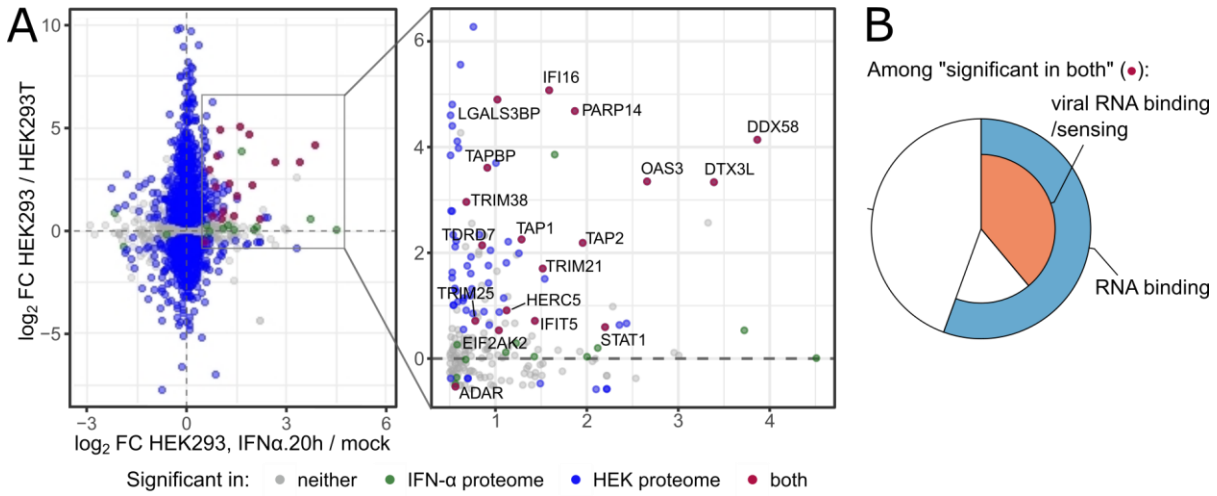


Figure.3.4.1 Comparative analysis between IFN-α regulated proteome, HEK293 proteome, and HEK293T proteome highlighted RNA binding proteins (RBPs) as mutual virus-restrictive factors. (**A**), scatter plot comparing proteins identified in proteomics data of IFN-α stimulated HEK293 cells, steady state HEK293 cells, and HEK293T cells. Significance of test results were colour labelled, and proteins significant in both datasets were text labelled. (**B**), pie chart highlighting viral RNA interactors and RBPs among proteins significant in both IFN-α and HEK293/HEK293T proteome.

Galectin-binding protein LGALS3BP and RNA binding protein TDRD7 were also found to have increased abundance in more virus-restrictive environment. LGALS3BP is known to function in integrin-mediated cell adhesion and have been reported to be involved in the host defence against virus[265]; TDRD7, however, has no known relation with innate immunity or cell response to virus. Enrichment analysis of these commonly regulated hits also revealed an interesting result – RNA binding proteins (RBPs) and, in particular, viral RNA interactors were strongly enriched among proteins that were significant in both datasets (Fig.3.4.1 B). This highlighted the important role of RBPs in antiviral response, which was not unexpected because various steps in antiviral response require direct binding to viral RNA, such as viral RNA sensing and viral translation inhibition. Toward the other direction of the cellular states continuum, however, no protein was found to be both enriched in the virus-permissive HEK293T cells and down-regulated by IFN-α. This suggested that proteins suppressed by IFN-α were not linked to the high virus permissiveness in HEK293T cells, but it remains unknown whether these proteins could affect virus replication in HEK293 or other human cells.

Discussion

Interferon induced antiviral effect has been intensively studied at the transcriptome level, but most studies at proteome level have limited depth of detection. This chapter provides a deep proteomic analysis of IFN-α stimulated HEK293 cells, in which more than 10,000 proteins are identified, and more than 9,000 proteins quantified. In this deep proteome dataset, one of the major challenge I faced was to solve the statistical issue where conventional test method failed to reveal IFN-α induced biological effects. This issue was caused by the two layers of batch effects that occurred in parallel with the biological effect. Separation by the effect "replicate" resembled a successful variance partitioning, where all technical error after SILAC sample combine was isolated and distinguished from random error. This allows improvement of measurement accuracy by accounting for this effect. The other effect from "SILAC label", however, was not beneficial because it represents extra error introduced by isotope labelling itself. This effect could have a mixed origin. Non-labelled contaminants from the environment could be the main source of this effect because the total intensity of "Light" channel, which has the same molecular weight as non-labelled contaminants, is consistently higher than the two samples labelled with heavy isotope. Factors such as impurity in isotope-labelled amino acids, and misidentification of isotope-labelled peptides in MS data analysis could all contribute to this "SILAC label" effect. In other SILAC experiments, incomplete isotope incorporation is known as a common source of error[266], but it should have limited effect on this dataset because I had experimentally confirmed an over 99.5% incorporation rate in all tested samples. It is also possible that the "SILAC label" effect arose from cell response to heavy isotope, but it is likely not the main source of error because no phenotypic difference was observed in this experiment, and a large number of SILAC studies have also confirmed that the labelling itself has little effect on cellular processes[267].

Simultaneous handling of these two batch effects is necessary for this work. In genomic studies, linear mixed model (LMM) is an effective approach in accounting for complicated confounding factors in data analysis[268,269]. However, a better solution I found for this dataset was to incorporate both effects as co-variables in linear model for moderated T-statistics. The application of Bayesian method-based variance moderation could be the main reason why it outperforms LMM method. Interestingly, recent study has proposed a new method that focuses on combining advantages of Bayesian-based variance moderation and LMM[270]. This new method was not applicable for my dataset because it only allows one effects being incorporated, however, development of more powerful computational method that allow variance moderation with multiple batch effects could further improve test result for proteome data that have complicated

experimental design. A different labelling strategy that can isolate technical error while minimising variance introduced by labelling itself would also benefit quantitative proteomics analysis.

Test results of this deep proteome data revealed no statistically significant changes at 4h post IFN-α treatment, and 66 high-confidence changes at 20h post-treatment. The lack of protein-level response at 4h IFN-α treated samples was puzzling, because an antiviral effect was observed at 4h post IFN-α treatment against both SINV and HIV-1, indicating that a certain level of antiviral factors should be active at early stage of IFN response. In agreement with this early-timepoint virus-inhibitory effects, RNA-centric studies also show that stimulation with type I IFN for 2-4 hours is able to induce gene upregulations in other human cells, and most of these early changes are persistent in later timepoints[154,271]. In fact, a number of proteins (mainly composed of well-established ISGs) did show substantial fold changes in this proteome data at early timepoint, but these were not determined "significant" in the T-test. For example, antiviral protein MX1 showed log-2 fold change of 2.75, but with FDR equal 0.72 (supplementary table 2). Therefore, this absence of significant changes could stem from low signal-to-noise ratio in the data. Regarding signal levels, not all cells in the sample response to IFN-α at the same pace, and early response could be more sensitive to cell-to-cell variations, resulting in a higher p-value in the T-test. As for the noise, as discussed above, the SILAC labelling process could introduce extra variance in the data. Although this effect was modelled and addressed, the variance-correction process reduced degree of freedom in the T-test, which globally increased p-values in test result. However, the absence of significant changes could also originate from genuine absence of protein-level response at early timepoint. System biological studies have shown that there is a time delay between transcription and translation[272], as synthesis and accumulation of protein products take much longer time than RNA[273]. Therefore, HEK293 cells could be readily responsive to IFN-α at RNA level and already started translation of antiviral proteins, whose accumulation to statistically significant level require longer period of time. At early stage of many acute responses (including IFN response), changes in proteins' post-translational modifications (PTMs) are frequently found to be main driver of activity changes[274,275], therefore it is possible that PTM-level regulations are better indicators of antiviral responses at early timepoint post IFN treatment. It is also worth noting that IFN response varies between cell types[244], and HEK293 cells could be genuinely slow-responsive to IFN stimulation at either RNA, protein, or PTM level compared with other cell types. A transcriptome-wide analysis of RNA-level responses at early timepoint could help unveil why protein-level early response in HEK293 cells was restricted. High-throughput studies into PTM changes at early timepoint post IFN-α stimulation could also help us understand factors underlying antiviral activities at early stage of IFN response.

At the late timepoint post IFN-α stimulation, under a detection depth of over 10,000 proteins, 66 were determined to be statistically significant. This scope of changes in proteome is more limited than we initially expected, given that 1/3 of genes identified in this proteome data have recorded IFN response in ISG database *Interferome*[231,232]. ISGs are generally of low abundance compared with highly expressed house-keeping genes. Quantitative proteomic analysis of these low-abundance proteins require higher detection depth and measurement accuracy, therefore, I expected that my deep proteomic profiling with SILAC will increase the chance of detecting differential expressed ISGs. However, the number of significant changes in my data does not show a substantial increase compared with other proteomics studies that also focus on IFN response. In addition, the number of changes determined in each study vary dramatically: Megger et al. reported 57 IFN-α regulated proteins in fibroblast from a proteome dataset that identified ~3,000 proteins[153], Tarasova et al. reported 70 out of ~4,000 identified proteins to be IFN-α regulated in glioblastoma cells[153], Colli et al. reported 267 IFN-α regulated proteins out of ~8,000 identified in pancreatic cells[154], and Zhao et al. reported 8 IFN-α regulated proteins out of ~7,000 identified in intestinal epithelial cells[276]. This large variation in detected IFN response could be partially explained by differences in proteomic profiling methods. The SILAC labelling method in this study can reduce technical error in quantification, but the number of replicate is restrained by limited isotope channels that cells can be labelled; by contrast, label-free MS analysis cannot reduce measurement error, but researchers could perform more replicates without limit. Megger et al. used 8 replicates per condition in their single-gradient proteomics analysis, Tarasova et al. used 10 replicates, and these high replicate numbers can improve confidence in statistical test. However, performing deep proteomic profiling with large replicate number in label-free is impractical because of the long instrumentation time. Sample labelling by tandem mass tag (TMT) allows more than 10 different channels[277], enabling both reduction of technical error and experiment with more replicates, but it tend to underestimate fold changes due to ratio compression issue[278]. Notably, both Colli et al. and Zhao et al. used proteomic profiling with TMT-labelling followed by off-line fractionation, but one reported 267 significant changes in pancreatic cells while the other recorded 8 changes in intestinal epithelial cells[154,276]. This contrast highlighted that differences between cell systems also have critical impact on differential IFN responses. Therefore, the scope of protein-level IFN-α response profiled in this study could be more representative for cells that are closely related to HEK cells, while detailed pictures of protein-level IFN-α response in other cell types still require case-specific analysis.

Among the 66 proteins regulated by IFN-α at late timepoint, 60 proteins were upregulated and 6 had decreased abundance. Most of the upregulated proteins have been recorded to respond to type I IFN in RNA-centric studies, however, many of them do not have well defined role in innate

immunity. More than one third of these upregulated proteins (23 out of 60) do not contain functional annotation related to innate immune response according to GO database. Part of these un-annotated proteins are revealed to have antiviral activities in very recent studies, for example RNA binding protein HELZ2 in suppression of Dengue virus and Zika virus[279,280], and mitochondrial nucleoside metabolism enzyme CMPK2 in restriction of Dengue virus and HIV[281,282]. The un-annotated proteins EPSTI1 and HERC6 have homologs that are involved in inflammation and innate immune response[283,284], but their role in human cells remain uncharacterised. Other un-annotated proteins, including APOL2, PNPT1, TDRD7, and USP41, have no known role in IFN response despite having been shown to respond to type I IFN in other transcriptome-level studies[232]. Notably, several IFN-α stimulated proteins have no record in RNA-centric ISG database. These include BTBD2, c12orf65, THAP10, and TRMT44, and most of them are scarcely covered by literature and have no known function in the cell. Their absence of record in RNA-centric studies suggest that they might be regulated by IFN-α at protein-level and not at RNA-level. However, this need to be confirmed by RNA-level data that is also obtained from HEK293 cells under the same experimental condition, because IFN response could vary between cell types. All of these proteins could be interesting candidates for follow-up studies to further elucidate whether they harbour antiviral activity, and how they participate in cellular response to IFN-α.

Apart from ISGs, hundreds of genes have been reported as IFN repressed genes (IRGs) in several transcriptomic and proteomic studies[153,232,285,286]. In this study, 6 proteins were found to be significantly downregulated by IFN-α. Interestingly, these repressions have limited overlap with recorded IRGs. GDE1, a glycerophosphoinositol metabolism enzyme, is the only protein that is downregulated in this study and has been reported to be repressed by type I IFN[287], and its role in IFN response remains unknown. Other IFN-α mediated downregulations identified in this dataset have no record in RNA-centric database *Interferome*. This lack of overlap with previous record is not unique to my IFN-α proteome dataset, a large-scale transcriptomic study analysing IFN responses across varies types of cells also shows that IFN-α induced suppressions are not conserved[286]. Another reason why these IFN-α downregulated proteins have no previous record could be that the IFN-α mediated repression only occur at protein level. Regulations of translation and protein stability can affect protein abundance without changing RNA abundance [288,289]. Notably, type I IFN stimulation can regulate translation process through mTOR pathway and protein degradation through ubiquitin-proteasome pathway[6,90,275]. More evidences at the RNA-level are required to elucidate whether protein downregulations observed in my data are caused by changes in RNA abundance, and this will be covered through a transcriptomic analysis in the following chapter.

When combining the IFN-α proteome dataset in this chapter with the proteomics analysis of HEK293/HEK293T cells in previous chapter, we are able to depict proteome landscapes among cells that showed a continuous spectrum of cellular states, ranging from permissive (HEK293T) to steady-state (HEK293), and to virus-restrictive (IFN-α stimulated HEK293). This series of proteome datasets were derived from cells of similar genomic background, conducted using the same experimental procedure, and acquired in a comparable depth of detection. These setups minimise biological and technical variations, therefore allow us to focus on proteins that shape these cellular states base on their expression patterns. Previous chapter showed that innate immunity factors were enriched in HEK293 cells over the more virus-permissive HEK293T cells. In the IFN-a proteome in this chapter, 17 proteins follow this trend and are enriched in IFN-a stimulated cells over the more permissive steady-state cells. These proteins include well-established antiviral factors such as MDA5, STAT1, and PKR, as well as proteins that are less-understood in regulating virus infection, such as LGALS3BP and TDRD7. These proteins are interesting candidates for follow-up functional studies, because they share similar expression pattern with known antiviral factors in my proteome datasets. Toward the other direction of this serial proteome datasets, HEK293T cells were found to be over-represented by viral proteins and viral RNA interactors in previous chapter, but none of these proteins were overexpressed in steady-state cells over the more virus-restrictive IFN-a stimulated cells. This indicates that cells are not actively downregulating these viral protein/RNA interactors to establish a virus-restrictive cellular state. However, we cannot conclude that these viral protein/RNA interactors are not regulated in restrictive cells, because change in abundance is not the only way to regulate protein function. Post translational modifications and even protein conformation changes could hugely affect protein activities and shape a restrictive cellular environment. For example, ADP-ribosylation can inhibit IFN response by targeting signalling protein TBK1[260], it can also enhance IFN response by promoting STAT1 phosphorylation[290], and conformation change from monomer to polymer form is critical for activation of IFN signalling factor MAVS [66]. These regulations cannot be captured by this proteome dataset which focuses on quantifying abundance changes. Further high-throughput analysis regarding protein modifications and activities could provide valuable insights into whether and how viral protein/RNA interactors are regulated by IFN-a without changes in abundance, therefore facilitate identification of proteins that regulate virus infection.

Acknowledgements

I want to thank all members of Castello lab for the supports in cell cultures, virology assays, and bioinformatics analysis, thank all members of Shabaz lab for the supports in the proteomics experiments and data analysis. Thanks Alfredo Castello-Palomares & Shabaz Mohammed for conceptualising the research in this chapter and their supervisions. Thanks Marko Noerenberg for kindly providing cell stocks of SILAC cell lines and training on HIV-1 pseudotyping. Thanks Manuel Garcia-Moreno for kindly providing SINV stocks for virus infection test. Thanks Philip Charles for discussions regarding methods for proteomics data analysis. Thanks the Department of Statistics, University of Oxford for the consulting service on multi-variable regression analysis. Thanks Aino Jarvelin for providing trainings and codes for GO term matching and text-mining.

Methods and Materials

Methods

Cell culture

Culture and maintenance of HEK293 cells were the same as Chapter 2. Culture with IFN-α stimulation was performed by supplying IFN-α stocks into culture media to desired concentrations and time periods prior to harvest. IFN-α stocks were obtained by dissolving commercial IFN-α powder (PBL assay science) with 0.1% BSA in water and stored at -80^{o}C.

Generation of viruses

The same as Chapter 2.

Western blotting

The same as Chapter 2.

Plate reader assay

HEK293 cells were seeded on 96-well plates at 0.25 million cell per well. Four hours after seeding, cells were and treated with IFN-α for desired concentrations and time periods prior to infection of SINV-mCherry or pseudotyped HIV-$1_{Gag\text{-}mCherry/}$ or HIV-$1_{Nef\text{-}mCherry}$ at multiplicity of infection (MOI) of 1. After infection, 96-well plates were transferred to the CLARIOstar microplate reader (BMG labtech). Fluorescent signal recording and analysis were the same as Chapter 2.

Cell culture with SILAC

Culture media for SILAC were arginine and lysine depleted DMEM (Thermo Scientific) with 10% dialysed FBS, 1x penicillin/streptomycin, and supplied with heavy-isotope labeled arginine and lysine (L-Arginine 13C,15N; L-Arginine 13C; L-Lysine 13C,15N; 4.4.5.5.-D4-L-Lysine; Silantes GmbH). SILAC cell stocks were kindly provided by Dr Marko Noerenberg, and these cells were cultured in corresponding SILAC media for at least 5 more passages before experiments. Cell passage process was the same as unlabelled cell.

Proteomics sample preparation

HEK293 cells with SILAC were cultured on twelve 150 cm^2 dishes to ~60% confluency. IFN-α were added to dishes at 20 h, 4 h, 10 min or not added prior to harvest. Cells were gently washed with 4^oC PBS and lysed on plate with cell scrapers using 1.4 ml urea lysis buffer (8M urea, 100 mM ammonium bicarbonate (AmBic), 20 µl protease inhibitor cocktail (Sigma) and 5 µl phosphatase inhibitor (Sigma)). Harvested samples were snap-frozen in liquid nitrogen. Three biological replicates were cultured and harvested on three separate dates.

The mock, IFN-α 4 h, and IFN-α 20 h samples were defrosted for deep proteomic analysis. An aliquot of each lysate was reserved for protein concentration measurement with Pierce 660-nm Protein Assay (Thermo Scientific), and isotope incorporation rate measurement by independently performing sample preparation and 60-min LC-MS/MS gradient with Q-Exactive (Thermo Fischer Scientific). The Pierce 660-nm protein assay was performed with NanoDrop 2000 spectrophotometer (Thermo Scientific), and the standard curve of protein concentration was obtained using BSA standard samples (Thermo Scientific). Equal weights of SILAC lysates were combined according to the label-swap strategy shown in Table.3.1, each combined sample contains 200 µg worth of

proteins. Reducing reagent Tris(2-carboxyethyl)phosphine hydrochloride (TCEP) and alkylating reagent 2-Chloroacetamide (CAA) were added to combined lysates to 10 mM and 50 mM, respectively, then incubated in dark for 30 min. Alkylated samples were purified with methanol-chloroform precipitation by: supplying each sample with 1.6 ml methanol, 400 µl chloroform, and 1.2 ml water sequentially, vortex after each reagent addition (final volume 3.6 ml), centrifuging in 3,000 g for 10 min to remove upper layer, supplying 1.2 ml methanol and vortex and centrifuge in 3,000 g for 10 min, removing all liquid and air dry, finally resuspending samples with 250 µl urea buffer. Purified alkylated proteins were diluted to 6 M urea with 25 mM AmBic for LysC protease (Wako) digestion with protein/protease ratio at 50:1 (w/w) at 37^{o}C for 4 h with 250 rpm shaking, then diluted to 1 M urea with 25 mM AmBic for trypsin protease (Promega) digestion with protein/protease ratio at 50:1 (w/w) at 37^{o}C for overnight with shake. Digestion was quenched with 0.5% (v/v) formic acid (FA).

Tryptic peptides were fractionated with a HPLC prior to LC-MS/MS analysis. The peptides were loaded with solvent A (10 mM AmBic 2% acetonitrile (ACN) in water, pH 8.3), and separated by a Zorbax 300 Extended-C18 column (2.1 x 150 mm, 3.5 µm) using a linear gradient (length: 100 min, 8% to 60% solvent B (80% ACN in water), flow rate: 200 µl/min). Fractions were collected every 1-min interval between 12 - 92 min of the gradient. The 80 collected fractions were combined at an even interval (1st, 21st, 41st, 61st fractions were combined, and so on), resulting in 20 fractions. These fractions were dried with SpeedVac and stored in -80^{o}C for mass spectrometry analysis.

Mass spectrometry

Each fraction of dried tryptic peptides were reconstituted with 50 µl of loading buffer (5% DMSO, 5% FA in water), 20 µl of reconstituted peptides mixture were analysed on a nanoUHPLC (Thermo) connected to a Q Exactive mass spectrometer (Thermo Fischer Scientific) through an EASY-Spray nano-electrospray ion source (Thermo Fischer Scientific). The peptides were trapped on a C18 PepMap100 pre-column (300 µm i.d. x 5 mm, 100 Å, Thermo Fisher) using solvent A (0.1% formic acid in water). The peptides were separated on an EASY-spray Acclaim PepMap analytical column (75 µm i.d. × 500 mm, RSLC C18, 2 µm, 100 Å) using a linear gradient (length: 120 min, 8 % to 28 % solvent B (0.1% formic acid, 5 % DMSO in acetonitrile), flow rate: 200 nL/min). The separated peptides were electrosprayed directly into the mass spectrometer operating in a data-dependent mode. Full scan MS spectra were acquired in the Orbitrap (scan range 350-1500 m/z, resolution 70000, AGC target 3e6, maximum injection time 100 ms). After the MS scans, the 20 most intense peaks were selected for HCD fragmentation at 30 % of normalised collision energy. HCD spectra

were also acquired in the Orbitrap (resolution 17500, AGC target 5e4, maximum injection time 120 ms).

Protein identification and quantification were performed using Andromeda search engine implemented in MaxQuant (1.6.3.4)[211]. Peptides were searched Human Uniprot database (version, January 2016). False discovery rate (FDR) was set at 1% for both peptide and protein identification. Data files from the same sample were assigned as near-by fractions, and match between run was turned on to match peak information from near-by fractions. Other options were set as defaults.

Proteomics quantitative analysis

Standard Student's T statistics was performed with software Perseus[212]. ProteinGroup files from MaxQuant search results were imported into Persues, proteins that have at least 3 valid values in at least one condition were subjected to log transformation, median normalisation, and imputation of missing values by background-level signal, then proceeded to paired T-test. All options for paied T-test were set as defaults.

Moderated T-statistics and T-statistics with linear mixed model was performed with R programming. ProteinGroup files from MaxQuant search results were imported into R, proteins that have at least 3 valid values in at least one condition were subjected to log transformation, normalisation, and imputation of missing values. Normalisation based on the vsn method was performed using package *vsn*[213], normalisation based of L-channel aligned method was performed as described in the main text. Imputation of missing values was performed by background-level signal using package *ImputeLCMD*[214].

For moderated T-test considering one layer of batch effect (replicate), processed data was subjected to package *limma* using moderated T-test function[249], with block variable set to replicate number of corresponding samples and correlation consensus calculated following the *limma* user guide. P-values were corrected for multiple testing by controlling the FDR with the Benjamini-Hochberg method. For T-statistics with linear mixed model, processed data was subjected to package *lme4* with IFN-α treatment condition set as main variable, while replicate number and isotope type set as co-variable according to its user guide[250]. For moderated T-test considering two layers of batch effects (replicate and isotope), processed data was subjected to package *limma* with IFN-α treatment condition and isotope input as two variables in design matrix, and replicate input as block variable. The correlation consensus for this block variable was calculated following user guide.

Principal component analysis (PCA) was performed with R programming using base R functions. Data processed by log-transformation and normalisation was selected, and only proteins that have no missing values were proceeded to PCA.

Bioinformatic analysis

Pathway enrichment was performed with online platform STRING (string-db.org) using the "Multiple proteins" search option[215]. Results of *Reactome* pathway enrichment results were downloaded and plotted in R programming.

Matching of IFN-related GO annotation was performed with R programming. The compilation of GO annotation was obtained from *GO.db* package[291]. GO terms containing keywords "innate immun/type I interferon/interferon-alpha" were considered IFN-related GO terms. Genes annotated with these GO terms were extracted with *biomaRt* package, then matched to the proteomic analysis result[292].

PubMed-wide text mining was performed with R programming. For each studied gene, the PubMed article searching was performed with package *RISmed* using gene name as input for searches[293], the abstracts of matched research articles were downloaded. Next, these abstracts were converted into one-gram with package *textstem*[294]. Searches of IFN/virus-related keywords were performed on one-gram tables. IFN-related keywords were "IFN/interferon/Interferon/innate immu/Innate immu", virus-related keywords were "virus/Virus/viral/Viral". Research articles whose abstract contain at least one IFN-related keyword or at least two virus-related keyword were counted as IFN/virus-related papers.

Hierarchical clustering was performed with R programming using base R functions. Protein log2 fold changes from each contrast were consolidated into one table, and normalised to unit variance prior to clustering. The number of clusters were determined based on Euclidean distances by plotting the cluster number against the "height" output of hierarchical clustering, and finding the "elbow" points of the curve. Data were divided into 6 clusters using cutree method.

Materials

Chemicals and reagents

REAGENT or RESOURCE	SOURCE	IDENTIFIER
100 bp ladder	NEB	B7025
2-Chloroacetamide	Sigma Aldrich	C0267
4.4.5.5.-D4-L-Lysine	Silantes GmbH	#211104113
6x purple loading dye	NEB	B7024S
Acetonitrile	Sigma Aldrich	#34851
Acrylamide/Bis-acrylamide 30%	Thistle Scientific	#20-2100-05
AEBSF hydrochloride	BioChemica	A14210100
Ammonium Bicarbonate	Sigma Aldrich	#9830
Bensonase	EMD Millipore	#70746
Calcium chloride	Sigma Aldrich	#499609
Chloroform	Sigma Aldrich	#67-66-3
Chloroquine	ACROS Organics	#50-63-5
dialysed FBS	Silantes GmbH	#281000900
Dimethyl sulfoxide, DMSO	Thermo Scientific	T6508
DMEM	Gibco	#11995-065
FBS	Sigma Aldrich	F9665
Formic acid	Thermo Fisher	A117-50
Interferon alpha 2a	Pbl assay science	#11100-1
L-Arginine 13C labeled	Silantes GmbH	#201204102
L-Arginine 13C,15N labeled	Silantes GmbH	#201604102
L-Lysine 13C,15N labeled	Silantes GmbH	#211604102
LysC protease	Wako	#129-02541
Methanol	Sigma Aldrich	#34860
Milk powder	Sigma Aldrich	#70166
Opti-MEM reduced serum medium	Thermo Fisher	#31985062
PBS 1x, pH 7.4	Gibco	#10010023
PEG6000	Sigma Aldrich	#25322-68-3
penicillin/streptomycin	Sigma Aldrich	P4458
Phosphatase inhibitor	Sigma Aldrich	P0044
Pierce 660nm Protein Assay Reagent	Thermo Scientific	#22660
Pierce Bovine Serum Albumin Standard Pre-Diluted Set	Thermo Scientific	#23208
Protease inhibitor cocktail	Sigma Aldrich	P8340
SILAC DMEM media	Thermo Scientific	#10107883
Sodium Deoxycholate Detergent	Thermo Scientific	#89904
Sodium dodecyl sulfate	Sigma Aldrich	#151-21-3
Tris(2-carboxyethyl)phosphine hydrochloride	Thermo Scientific	#77720
Triton-X100	Sigma Aldrich	#9036-19-5
Trypsin EDTA	Sigma Aldrich	T3924
Trypsin, sequencing grade	Promega	V5111
Tween-20	Sigma Aldrich	P1379

Urea	Sigma Aldrich	U1250
X-tremeGENE 9 DNA transfection reagent	Sigma Aldrich	#6365787001

Instruments

INSTRUMENTS	SUPPLIER/INSTITUTE
Countess II FL Automated Cell Counter	Thermo Fisher
Odyssey Fc imaging system	LI-COR Bioscience
Nanodrop 2000 Spectrophotometer	Thermo Fisher
Trans-Blot Turbo Transfer system	Bio-Rad
CLARIOstar microplate reader	BMG labtech
LTQ XL Linear Ion Trap Mass Spectrometer	Thermo Scientific
UltiMate 3000 RSLCnano System	Thermo Scientific
Q Exactive mass spectrometer	Thermo Scientific

Antibodies

The same as Chapter 2

Plasmids

PLASMIDS	**SOURCE**
pNL4-3.R-E- Nef-mCherry	Castello lab
pNL4-3.R-E- Gag-mCherry	Castello lab
pHEF-VSV-G	Sattentau lab

Cell lines

CELL LINE	SOURCE	IDENTIFIER
HEK293	ECACC	# 85120602

Chapter 4

Transcriptome profiling of IFN-α stimulated cells revealed correlations and divergences between IFN-α responses at RNA and protein levels

Introduction

The central dogma of molecular biology describes gene expression as a process that DNA is transcribed into mRNA and then translated by the ribosome to synthesize the final protein product. It is conventionally believed that regulation at transcription level by transcription factors is the main driving force of gene expression regulation, but this view has been challenged in the last decade[233,272,295]. Recent technical advancements in RNA sequencing and mass spectrometry-based proteomics have allowed quantitative analysis of RNA and protein abundance at omics level. Based on these techniques, series of works have devoted to investigate correlations between absolute abundance of mRNA and that of protein at whole cell level. In a study that quantified absolute mRNA and protein abundance in cultured cell line, it was found that around 40% of variations in protein abundance could be explained by mRNA abundance, while the remaining ~60% are correlated with other regulatory mechanisms[233,295]. This view has been further complemented by large-scale studies that compare protein and mRNA abundance in multiple tissues, and it was found that the mRNA-protein correlation vary dramatically depending on which type of correlation is examined. For correlations that compare protein-to-RNA ratios between different genes within the same cellular context (between-gene correlations), mRNA levels correlate reasonable well with protein levels, and the correlation could exceed 0.8[296,297]. However the high between-gene correlation is a result of the huge dynamic range in protein abundance. For example, highly expressed housekeeping genes such as ribosomal protein RPS14 has substantially higher mRNA and protein abundance than transcription factor AHCTF1 regardless of tissue types, therefore resulting in high mRNA-protein correlation[298]. When we focus on the contribution of post-transcriptional and post-translational regulations in protein expression, it is useful to compare protein-to-RNA ratios of the same gene between different cellular environments (within-gene correlation). Contradict to the high correlations in between-gene comparisons, with-gene correlations are much lower in general (median correlation 0.21)[299]. This imply that for a given gene, its mRNA level generally has low power in predicting its protein level in different cellular context. Supporting this view, comparative omics studies have highlighted that mRNA-protein correlations for each gene is highly dependent on the cellular context, as translation efficiencies and protein half-lives vary dramatically across different tissues[300,301]. Based on these findings, it is now widely accepted that both transcriptomic and proteomic data can provide valuable and irreplaceable information to comprehensively understand the landscape of gene expression regulations.

A wide variety of regulatory mechanisms can modulate protein abundance without affecting mRNA abundance. These generally include translation regulation, which determines whether an mRNA is translated and its translation rate[288,302], and protein stability regulation, which controls

protein abundance through degradation[289,303]. MicroRNA (miRNA) and RNA-binding protein (RBP) are two major classes of factors that participate translation regulations. MiRNAs are small non-coding RNA genes that typically have 20-25 nt and exert their biological functions through base-pairing with complementary sequences normally placed within the 3' untranslated region (3'UTR) of mRNAs[304]. The miRNA is best known for its ability to recruit deadenylation complex to induce mRNA degradation[305], meanwhile, it can also downregulate gene expression through translational inhibition[306,307]. RBPs participate in almost all steps in regulations of translation rate. For example, translation initiation factor eIF3 is a multi-subunit RBP complex that can be recruited to 5'UTR region to initiate translation. In other examples, the RBP HuR can compete with the miRNA miR-21 in mRNA binding, therefore antagonise the translation suppressive effect[308]; RBPs Nanos and Pumilio can mediate translation repression during posterior patterning in Drosophila embryo[309]. Apart from regulatory processes mediated by miRNA and RBP, elements and features on the mRNA sequence also affect translation process. RNA secondary structure on the mRNA can regulate translation. Secondary structures such as internal ribosome entry sites (IRESs) on the 5'UTR of mRNA can recruit ribosomes for translation initiation[310], while stable hairpin structures near 5′ methyl-G cap can block translation[311]. Codon composition of the mRNA could also affect translation rate[312]. Recent studies showed that the usage of frequent codons that relate to high-abundant tRNAs would result in faster translation elongation[313,314] than mRNAs with suboptimal codons linked to low-abundant tRNAs.

The mechanisms that regulate protein stability are also highly diverse. In most of the known protein degradation processes, proteins subject to degradation are covalently modified with poly-ubiquitin chain on the lysine residue, and directed to the proteasome for digestion[289]. Recent studies have also highlighted the roles of ubiquitin-independent proteasomal proteolysis and proteasome-independent proteolysis in protein degradation[303,315]. For example, degradation of p53 by cytosolic thiol-protease CAPN3 is reported to be unaffected by proteasome inhibition[315]. Regulations in these protein degradation processes can uncouple the direct relation between mRNA and protein abundance. The uncoupling of mRNA-protein correlation is examined in a greater detail in another comparative omics study. By comparing mRNA abundance, ribosome occupancy, and protein abundance in a system-wide level, researchers show that mRNA levels correlates better with ribosome occupancy than with protein abundance, and similarly, protein abundance also correlated better with ribosome occupancy than with mRNA abundance[316]. These results are in-line with our current understanding of gene expression regulation, where various regulatory processes participate at different steps of the central dogma, and collectively shape cellular proteome.

Using transcriptome profiling techniques such as microarray and RNA-seq, researchers have documented thousands of IFN-α induced RNA-level regulations in different samples and treatment

conditions[154,232,286]. In the proteomic analysis of IFN-α response in Chapter 3, the comparative analysis showed that these documented RNA-level regulations were strongly overrepresented among IFN-α upregulated proteins, and were depleted among IFN-α downregulated proteins. However, these studies were conducted in different cellular environments and IFN-α treatment conditions, which could affect IFN-α response of the cell. Therefore, a direct comparison between proteomic and transcriptomic data from the same cell line and treatment condition is necessary to comprehensively investigate the correlation and regulatory network between IFN-α induced RNA and protein level responses. In this chapter, I aimed to profile transcriptome landscape of IFN-stimulated cells using the same cell line and treatment conditions as in proteomic profiling in previous chapter. The resulting IFN-α transcriptome data would be compared with IFN-α proteome data to assess correlation between IFN-α responses at mRNA and protein levels, and to investigate features that could explain the observed mRNA-protein correlation or discordance.

Results

4.1 Widespread discordance between IFN-α responses in RNA- and protein-levels were revealed by integrative RNA-seq and proteomics.

To obtain transcriptomic data of IFN-α stimulated cells to complement the proteomic data obtained in the previous chapter, RNA-seq was performed using HEK293 cells and the same IFN-α treatment conditions (Chapter.3, Section.2). Total RNA was extracted and processed for cDNA library construction and next-generation sequencing after ribosomal (r)RNA depletion (Fig.4.1.1 A). Sequencing was performed to a depth of approximately 50 million reads per replicate, resulting in over 33,000 Ensembl gene IDs identified. Quantitative analysis was performed with T-statistics and a cut-off of FDR<0.05, revealing 126 and 419 genes as significantly differentially expressed genes (DEGs) at 4h and 20h post IFN-α treatment, respectively (Fig.4.1.1 B). Interestingly, most of genes that were regulated at 4h-post IFN-α treatment remained significant at 20h post treatment (Fig.4.1.1 C), showing that IFN-α induced regulations in RNA expressions had a highly consistent temporal profile.

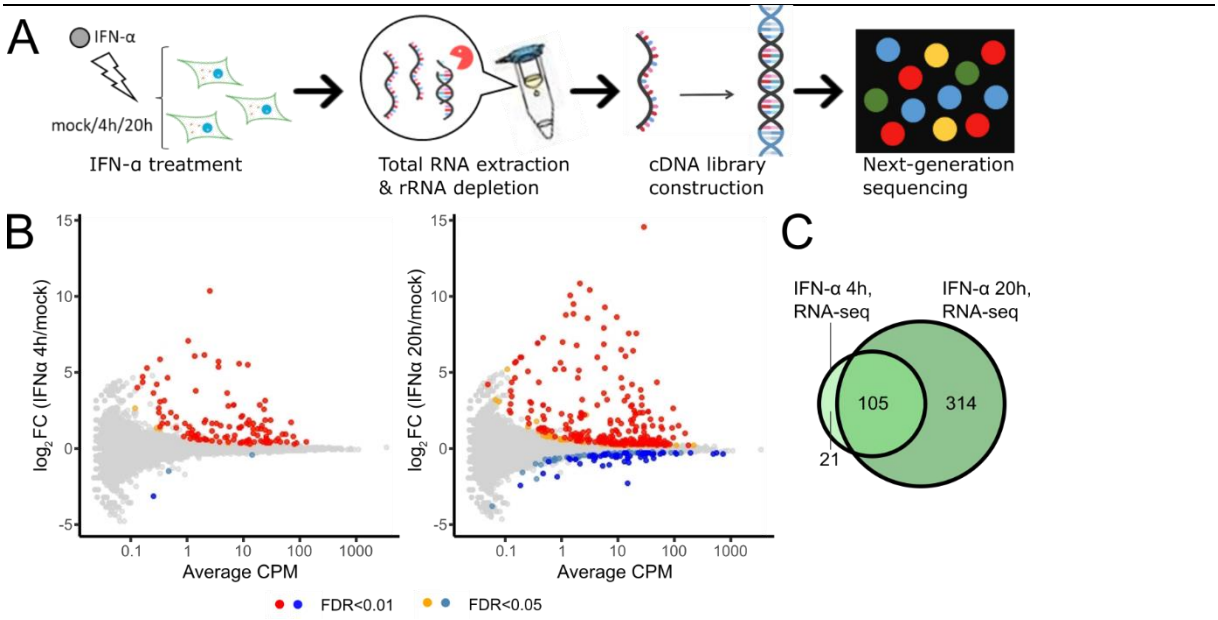


Figure.4.1.1 RNA-seq depicted IFN-α induced gene expression regulations at transcriptome level in HEK293 cells. (**A**), schematic workflow of RNA-seq of IFN-α stimulated HEK293 cells. (**B**), MA plot showing changes in transcriptome landscape after IFN-α treatment for 4h (left) or 20h (right), genes of "significant" changes were determined by FDR and colour labelled. (**C**), venn diagram comparing numbers of "significantly" regulated genes in each IFN-α treatment condition.

The IFN-α induced gene-expression regulations could vary between cell types. To confirm that IFN-α induced RNA-level response in HEK293 cells was comparable with that in other human cell types, my sequencing results were cross-referenced to published RNA-seq datasets of IFN-α stimulated human cells. Transcriptomic data of IFN-α treated primary human fibroblasts[286] and human pancreatic beta cells (EndoC-βH1 cell line)[154] were selected for comparative analysis. It's worth noting that the IFN-α used in these two compared studies were 1,000 U/ml, while IFN-α concentration in this study was set to 200 U/ml to mimic physiological condition and reduce anti-proliferative effects. Under the same cut-off for significant changes (FDR<0.05), the number of genes that were significantly regulated by IFN-α varied substantially between cell types: in total 305, 2045, and 872 genes were upregulated in HEK293, human fibroblast, and EndoC-βH1 cells, with only 172 genes upregulated in common (Fig.4.1.2 A). This contrast between the numbers of significant hits could arise from the lower IFN-α dosage in this study. However, a direct comparison of fold changes

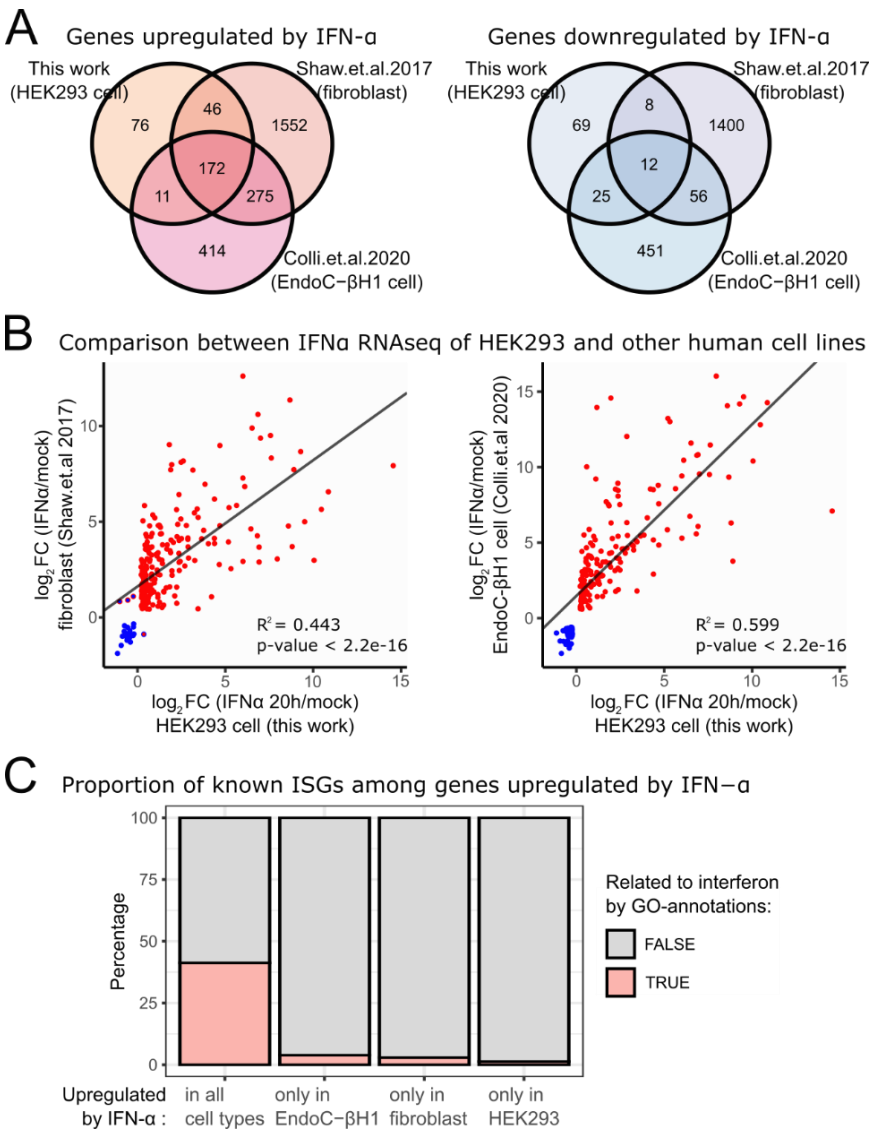


Figure.4.1.2 Transcriptome regulations after IFN-α stimulation in HEK293 cells were consistent with regulations in other cell types. (**A**), venn diagram comparing numbers of genes upregulated (left) or downregulated (right) by IFN-α in each cell line. Cut-offs for significant changes in three compared studies were all set to FDR<0.05 after corrected for multiple test effect. (**B**), scatter plots comparing RNA-seq results of genes that were "significantly" regulated in both IFN-α stimulated HEK293 cell and human fibroblast (left) or human pancreatic beta cell line (EndoC-βH1 cell, right). The up- and down-regulated genes in HEK293 cells were filled with red and blue, and genes up-/down-regulated in other cell types were outlined with red/blue. (**C**), proportions of known ISGs among genes upregulated by IFN-α in all three cell types, or specific to each cell type. Known ISGs were defined as genes containing GO annotation related to interferon.

---

revealed that that IFN-α responses in HEK293 cell were consistent with that in both human fibroblast and EndoC-βH1 cell, with correlation coefficients of 0.44 and 0.6 respectively, suggesting a common regulatory signature (Fig.4.1.2 B). Importantly, the majority of the genes that were up- or down-regulated in HEK293 cells were also regulated in the same direction in fibroblast and EndoC- βH1 cell, and only 4 genes were outliers. Further examines using gene ontology (GO) annotations also showed that IFN-α responses were consistent across cell types. Among DEGs in all three datasets, over 41% were linked to interferon by GO annotations. By contrast, less than 4% of DEGs in a single cell type were related to IFN-α response (Fig.4.1.2 C). These observations support that different cell types, including HEK293 cells employed in this study, share a set of functionally well-described ISGs that are consistently regulated upon IFN-α stimulation.

Regarding IFN-α repressed genes (IRGs), 114, 1476, and 544 genes were downregulated in HEK293, human fibroblast, and EndoC-βH1 cells, respectively. Notably, only 12 genes were commonly downregulated in all three cell types (Fig.4.1.2 A). A similar lack of consistency in IRGs between cells of different species were also reported in a published RNA-seq study, where researchers proposed that IRGs were evolutionally less conserved[286]. The limited overlap of IRGs among cross-referenced human cells indicate that IFN-α mediated downregulations could also be strongly dependent on the cellular context.

Having confirmed that the alterations in RNA levels in this study were representative of IFN-α response in human cells, I cross-referenced my RNA-seq dataset with the IFN-α proteomic results from Chapter 3. At early timepoint (4h post treatment), none of the genes that show transcriptome changes have corresponding significant changes at protein level (Fig.4.1.3 A, upper-left, green dots). However, many of the core ISGs do show substantial increase in protein abundance despite that they were not determined "statistically significant". At late timepoint (20h post treatment), RNA

upregulation positively correlated with protein upregulation (coefficient = 0.79, Fig.4.1.3 A, lower-right). Notably, despite having a high correlation, only 53 genes were significantly upregulated at both RNA and protein levels. By contrast, 366 genes were only regulated at RNA level, and 13 genes were only regulated at protein level (Fig.4.1.3 C). With regard to IFN-α mediated downregulations, none of the RNA downregulation overlapped with protein downregulation, leading to a close-to-zero correlation (coefficient = 0.05, Fig.4.1.3 A, lower-right). These discordances showed that a large part of the IFN-α responses at RNA level were not transferred to changes in protein abundance, and a number of IFN-α responses are independent of changes in RNA abundance.

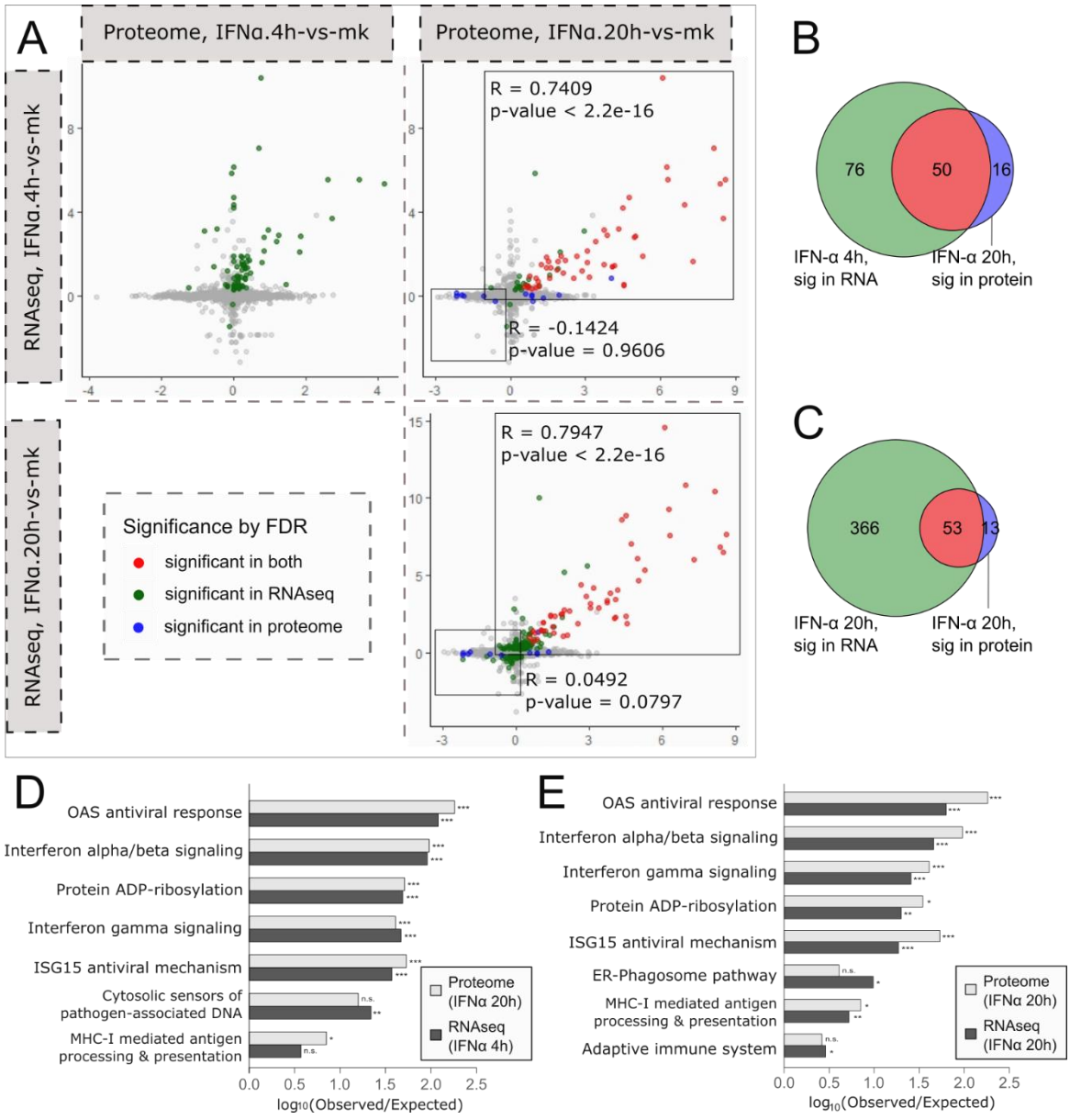


Figure.4.1.3 Comparative analysis revealed correlations and divergences in IFN-α induced changes in the transcriptome and proteome. (**A**), scatter plots comparing log-2 fold changes in RNA-seq and proteomics results of IFN-α stimulated HEK293 cells. Only genes detected in both datasets were

plotted, and genes were colour labelled based on their responses at RNA and protein levels. (**B, C**), venn diagram comparing number of "significantly" regulated genes in respective datasets. Genes only detected in one dataset were also included. (**D, E**) enrichment of *Reactome* pathways among genes of "significant" changes at either RNA-seq or proteomics results.

To assess whether IFN-α induced transcriptome and proteome regulations were also correlated at the pathway level, enrichment analysis was performed on genes significantly regulated in RNA-seq and proteome datasets. Expectedly, classical innate immunity pathways were highly enriched in DEGs at the RNA level, and at late-timepoint protein level (Fig.4.1.3 D,E). These included OAS antiviral response, interferon signalling, and ISG15-dependent antiviral mechanism. Regarding other enriched pathways, the ADP-ribosylation was enriched at all examined levels of responses, highlighting its important role in antiviral response. MHC-I-dependent antigen presentation was enriched at late-timepoints in both RNA and protein levels. Jointly, IFN-α responses at RNA and protein levels showed high similarity at the pathway level (Fig.4.1.3 D,E). It is worth noticing that a large number of DEGs showed discordance between RNA- and protein-levels responses, however these discordance did not result in dissimilarity at the pathway level.

A positive correlation was also found when comparing the RNA-seq data at early timepoint with the proteomics data at late timepoint (coefficient = 0.74, Fig.4.1.3 A, upper-right). Interestingly, this comparison showed a larger proportion of overlap between DEGs at RNA and protein levels: in total 50 genes were regulated at both levels, while 76 and 16 genes were only regulated at RNA and protein levels respectively (Fig.4.1.3 B). However, this overlap is also limited on IFN-α mediated upregulations, whereas the downregulated DEGs showed no correlation between RNA and protein level responses. This finding suggested that the increased RNA abundance in early IFN-α response is a reflection of the proteome upregulations at later time points. This time delay can be explained by the time that upregulated RNAs require to translate and accumulate to detectable changes in protein abundance.

To investigate if this interpretation is correct, I analysed the correlation between features observed by RNA-seq at early timepoint and proteomic result at the late timepoint. Briefly, IFN-α upregulated DEGs were categorised into two groups: 1) "common DEGs" for genes that had significant changes at both RNA and protein levels (Fig.4.1.4 A, red dots), and 2) "RNA-only DEGs" for genes that only had significant RNA-level changes (Fig.4.1.4 A, green dots). Different metrics in RNA-seq analysis result were extracted as features, and correlation tests were performed to compare these features between "common DEGs" and "RNA-only DEGs". As shown by the correlation test result, fold change in RNA-seq had a poor correlation with response at the protein level (coefficient =

0.237), indicating that absolute value of changes in mRNA abundance had limited power in explaining delayed protein-level changes (Fig.4.1.4 B, [1]). Reads count (CPM, count per million reads) and FDR in RNA-seq result were features that showed moderate correlations with protein-level response at late timepoint, with coefficients equal 0.417 and 0.401 respectively (Fig.4.1.4, [2,3]). Gene's position on the MA plot of RNA-seq data was also correlated with its protein level response – genes that located further away from non-significant hits in the MA plot was more likely to have protein level response at late timepoint (Fig.4.1.4 A). Notably, genes with low abundance tend to have higher noise, therefore have higher fold change than high-abundance genes in general. To take this effect into account, I used the principle of "significance B" statistics in proteomics data analysis[211] to calculate "local Z-score" as a feature that measures a gene's relative distance to other non-significant hits. Briefly, for each gene, its nearby 1,500 genes were considered to be a "local distribution" that have a similar measurement error, and its local Z-statistics score was calculated based on its fold change and its local distribution. Interestingly, compared with other features, this "local Z-score" metric showed the strongest correlation with protein-level responses. However, this correlation coefficient did not exceed 0.5 (Fig.4.1.4 B, [4]). Collectively, none of the analysed features in RNA-level early response showed a higher-than-0.5 correlation with protein-level late response. This indicate that the time-delay effect have limited power in explaining protein-level response, and IFN-α induced changes in proteome can only be partially explained by changes in transcriptome, even considering the time-delay effect.

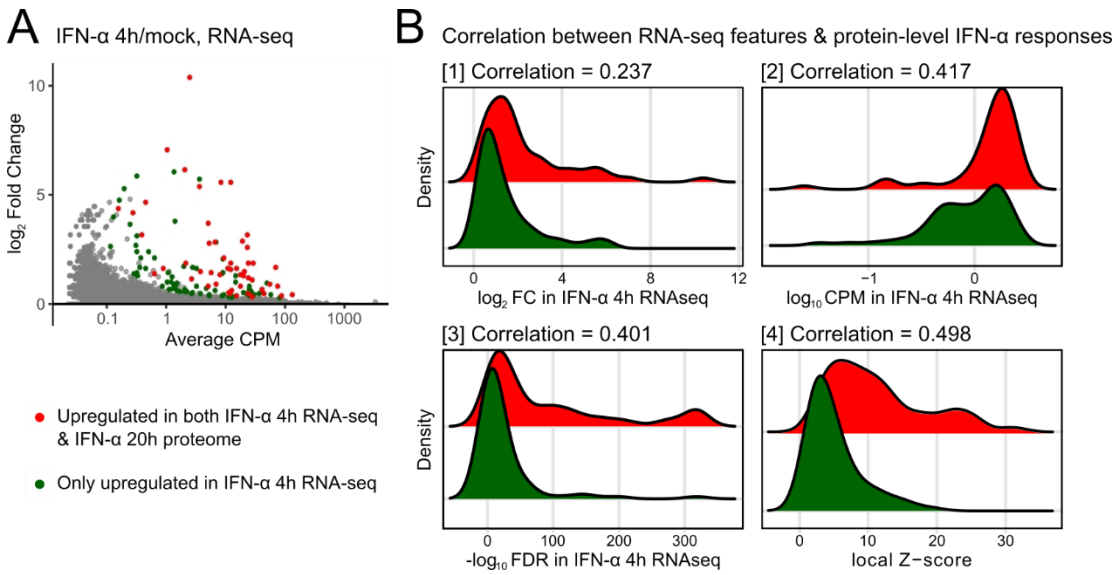


Figure.4.1.4 Early timepoint IFN-α response at RNA level showed limited correlation with late timepoint protein-level response. (**A**), MA plot of IFN-α 4h versus mock RNA-seq result. Only genes with positive fold changes are shown, and genes were colour labelled based on their responses at RNA and protein levels. (**B**), density plots comparing RNA-seq features among genes regulated at

both RNA and protein levels (red) and that among genes only regulated at RNA level (green). Four features were compared: fold change, reads count, FDR, and Z-statistic score within nearby 1,500 genes. Point-biserial correlations were calculated to correlate RNA-seq feature (x-axis, continuous variable) and protein-level response (y-axis, categorical variable). Correlation for each feature were shown above corresponding density plot.

4.2 Exploring regulatory signatures underlying discordance in RNA and protein levels after IFN-α stimulation

Given the discordance between RNA and protein was observed for many genes, I wondered what roles these genes play (if any) in innate immunity. Genes regulated by IFN-α were divided into three groups based on their RNA- and protein-level responses at late timepoint: 1) "protein-only DEGs" for genes that only had significant protein-level changes, 2) "RNA-only DEGs" and 3) "common DEGs" as described above. When searched against GO annotation database, expectedly, 2/3 of "common DEGs" contained GO terms linked to interferon (Fig.4.2.1 A, left). This agreed with enrichment analysis result at pathway level (Fig.4.1.3 E). Among "RNA-only DEGs", 11.2% (46 genes) were linked to interferon by GO annotations (Fig.4.2.1 A, middle). A closer examine showed that these "RNA-only DEGs" include well-characterised antiviral factors in TRIM and APOBEC protein family (Fig.4.2.1 B). TRIM family members TRIM5, 14, 22 and 56 are known to widely participate in host antiviral signalling and in the life cycles of many RNA and DNA viruses[125]; APOBEC family members APOBEC3G and APOBEC3F are DNA base-editing enzymes that restrict various types of virus by introducing detrimental mutations into viral genomes[127]. Their lack of response at protein level was puzzling, because these genes are known antiviral factors and had significant changes at RNA level. This discordance could be the result of post-transcriptional regulations, such as translational control[288].

For "protein-only DEGs", GO analysis revealed no known connection to interferon pathway (Fig.4.2.1 A, right). Functions of these genes were highly diverse. Tryptophan-tRNA ligase WARS was reported to be an ISG in human fibroblast[317], and its secreted form could act as ligand of Toll-like receptor complex to enhance innate immune response[318]. TAP1 and TAPBP are members of MHC-I antigen presenting complex[319], and they are both upregulated only at protein level (Fig.4.2.1 C). However, another member of MHC-I antigen presenting complex – TAP2 was upregulated at both RNA and protein levels. These changes indicated that innate immune response are linked to adaptive immunity, possibly by a complexed mechanism that involves both RNA-dependent and -independent regulations. MTRF1 and c12orf65 are both mitochondrial translation release factors that are

downregulated and upregulated respectively (Fig.4.2.1 C). This hints a potential shift in usage of mitochondria translation releasing factors in response to IFN-α. Other regulated genes included metabolism enzymes ENTPD5[320] and GDE1[321], methyltransferase METTL9[322], and several poorly characterised genes. Little was known regarding the roles of these genes in interferon response and the mechanisms underlying their protein-level responses.

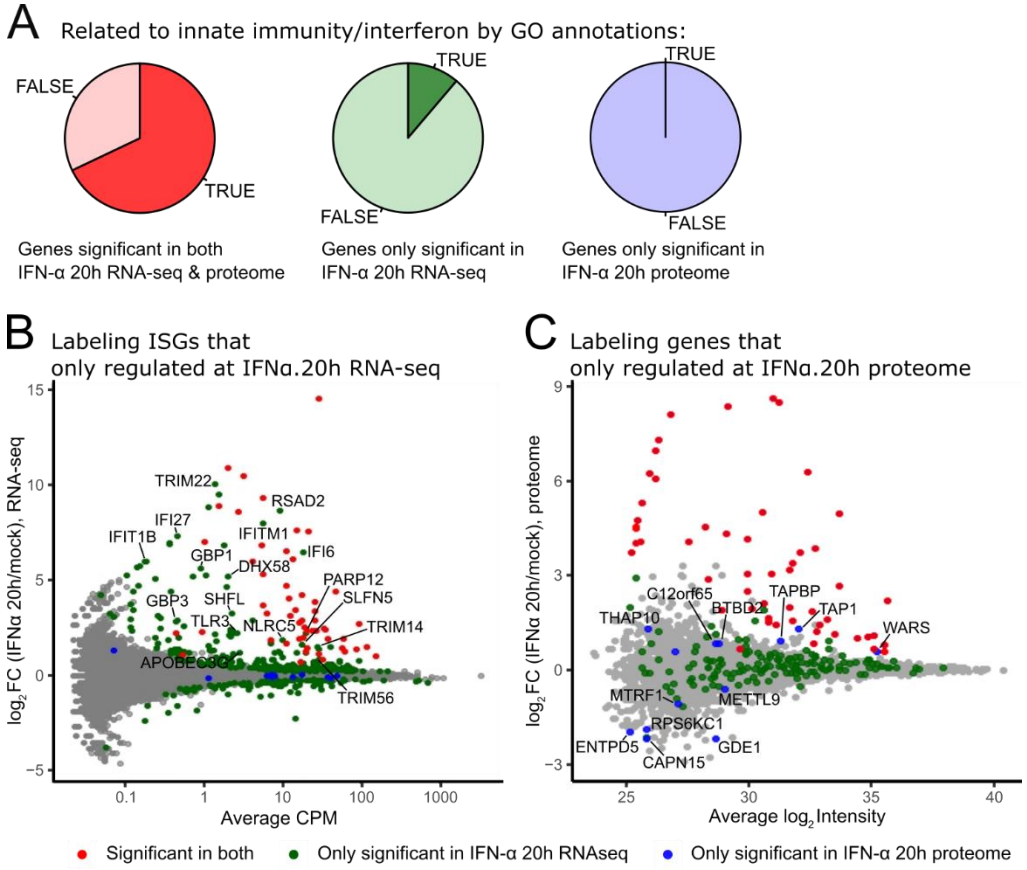


Figure.4.2.1 Genes that showed diverging IFN-α responses at RNA- and protein-levels contained both known ISGs and genes not related to interferon. (**A**), proportions of genes that were related to innate immunity or interferon by GO annotations. (**B**), MA plots of late-timepoint IFN-α responses in transcriptome highlighting known ISGs. (**C**), MA plot of late-timepoint IFN-α responses in proteome highlighting genes only regulated at protein level.

The observed discordance between RNA and protein levels in response to IFN-α indicates participation of post-transcriptional or post-translational regulatory mechanisms. Various factors could contribute to these regulations, RNA binding proteins (RBPs) and microRNAs (miRNAs) are two major classes of these factors[304,323]. To investigate whether RBPs and miRNAs are correlated with discordance in RNA/protein responses, IFN-α regulated genes were divided into three groups based on their RNA and protein responses, then correlation tests were performed to compare RBP- and

miRNA-related features between different groups. Because of the difficulty to thoroughly summarise binding sites of different types of RBPs across all IFN-α regulated genes, I analysed the length of mRNA's 3'UTR assuming a positive relationship between length and frequency of RBP-RNA interactions. As visualised in the density plot, little or no difference in 3'UTR length was observed between "common DEGs" and "RNA-only DEGs" (Fig.4.2.2 A, left). Surprisingly, "protein-only DEGs" appeared to have shorter 3'UTR in general, but it remained obscure whether this trend has any biological implication as the correlation was very low (coefficient = 0.15). The miRNA binding sites for each gene were obtained from online database *miRDB*, which consisted of both functionally identified and predicted miRNA target sites[324]. The test result showed that "protein-only DEGs" and "common DEGs" had similar miRNA binding sites. Interestingly, "RNA-only DEGs" appeared to have marginally more miRNA binding sites in general (Fig.4.2.2 A, right), indicating that these genes could be subjected to stronger miRNA-dependent translational regulation. However, this correlation was also weak (coefficient = 0.096), so no major conclusion could be drawn based on this result alone.

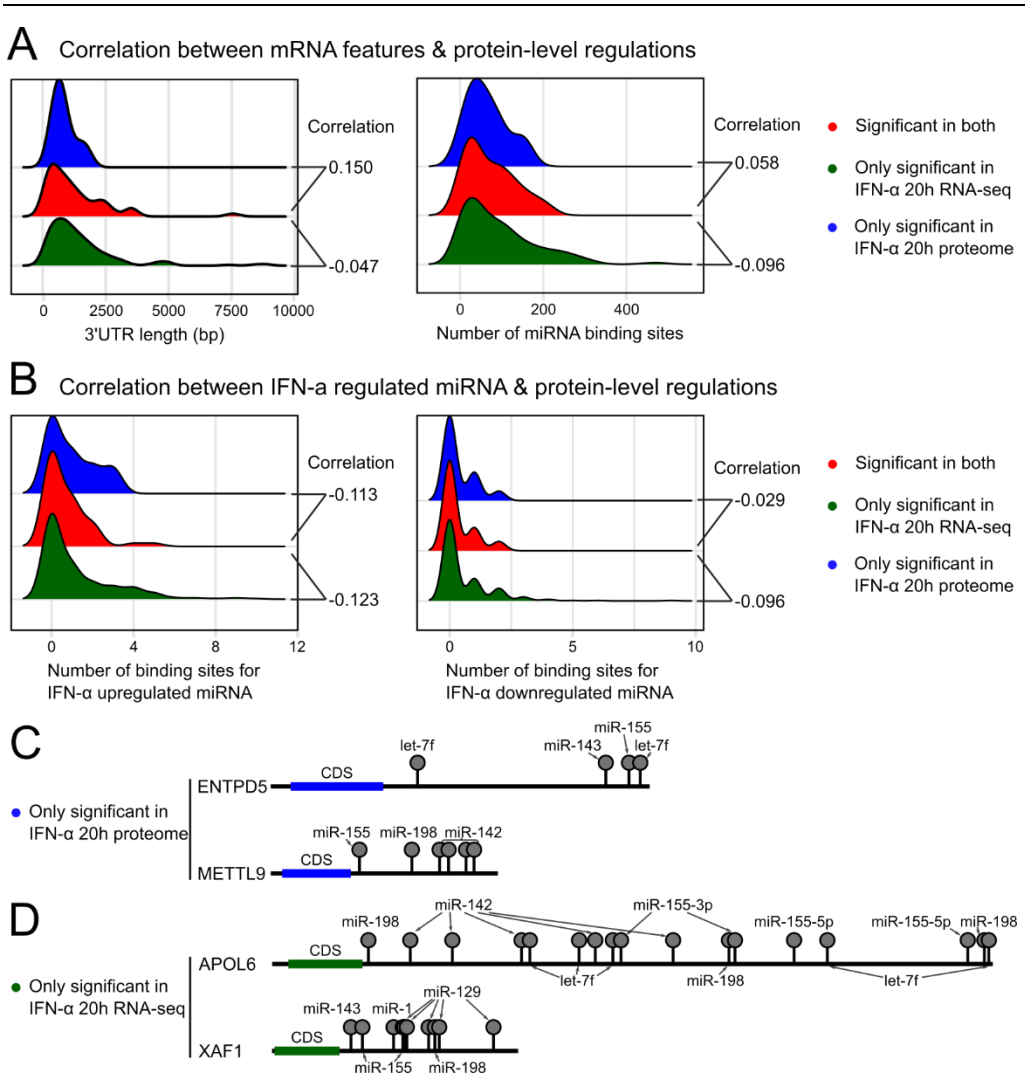


Figure.4.2.2 Diverging RNA- and protein-level responses in IFN-α regulated genes showed little correlation with 3'UTR length and number of miRNA binding site. (**A**), density plots comparing 3'UTR length and number of miRNA binding sites between genes that showed different RNA- and protein-level responses at late-time-point IFN-α stimulation. Positive correlation coefficients indicated higher occurrence in "common-DEGs". (**B**), density plots comparing number of binding sites for IFN-α upregulated and downregulated miRNA. (**C, D**), transcripts highlighting binding sites of IFN-α upregulated miRNA.

Series of studies have reported that type-I IFN could regulate miRNA abundance, and these miRNAs were participated in regulation of interferon response[325]. Correlation tests were performed to investigate whether these IFN-regulated miRNAs could have an impact on the RNA/protein discordance. The resulting density plot showed no correlation between RNA/protein response and the number of binding sites of IFN-α downregulated miRNA (Fig.4.2.2 B, right). By contrast, IFN-α upregulated miRNA appear to be correlated with RNA/protein discordance as "common DEGs" generally showed fewer binding sites (Fig.4.2.2 B, left). Genes with the highest number of miRNA binding sites were further examined. Phosphotase ENTPD5 and methyltransferase METTL9 were downregulated by IFN-α only at protein level, and their cognate mRNAs contained multiple binding sites of IFN-α upregulated miRNAs (Fig.4.2.2 C). The lipid binding protein APOL6 and the pro-apoptotic factor XAF1 were "RNA-only DEGs", and dozens of binding sites of IFN-α upregulated miRNAs were mapped to their cognate mRNA transcripts (Fig.4.2.2 D), indicating that these genes could be intensively regulated by miRNA during IFN-α response.

Bias in codon usage in the mRNA could also affect protein synthesis as usage of optimal or suboptimal codons change the elongation speed, which can affect protein abundance without changing RNA abundance. To investigate whether genes that had discordant RNA- and protein-level IFN-α responses showed bias in codon usage, frequency of each codon was calculated for each DEG. The result table, which contains the frequencies of all 64 codons, was subjected to principal component analysis (PCA). This unsupervised machine learning method could evaluate codon usage pattern of each gene based on their codon frequencies, and assess their differences in codon usage patterns. PCA result was visualised in scatter plot where each dot represent a gene, and the distance between two dots shows the difference in codon usage between two genes. All DEGs distributed homogenously on the first principal component (PC1, Fig.4.2.3, left), showing that they have highly diverse codon usage patterns. However, the scatter plot of PC1 & PC2 did not separate "common DEGs" from "RNA-only DEGs" or "protein-only DEGs", indicating that different groups of DEGs have similar codon usage. When examining more PCs, surprisingly, I found that DEGs were divided into

four distinct clusters in PC2 & PC3 (Fig.4.2.3 A). Clustering in PCA result indicates that certain feature(s) in codon usage was shared among genes within the cluster. To explore what is the codon usage feature that separates DEGs, I checked variable correlation plots that show the importance of each codon in each PC. As a result, it was found that the main differentiating variables in PC2 & PC3 were three stop codons (Fig.4.2.3 B). Because gene clusters on PCA plot showed no clear correlation with their IFN-α responses at RNA or protein level, this differential usage of stop codon was likely not correlated with IFN-α responses. After excluded stop codons from the PCA, the result showed that DEGs were widely distributed in all assessed PCs without forming obvious clusters (Fig.4.2.3 C). Dozens of codons were found to have high contribution in each PC (Fig.4.2.3 D). However, because different groups of DEGs were almost uniformly distributed in each principle component, it remained difficult to highlight the role of codon usage in IFN-α response.

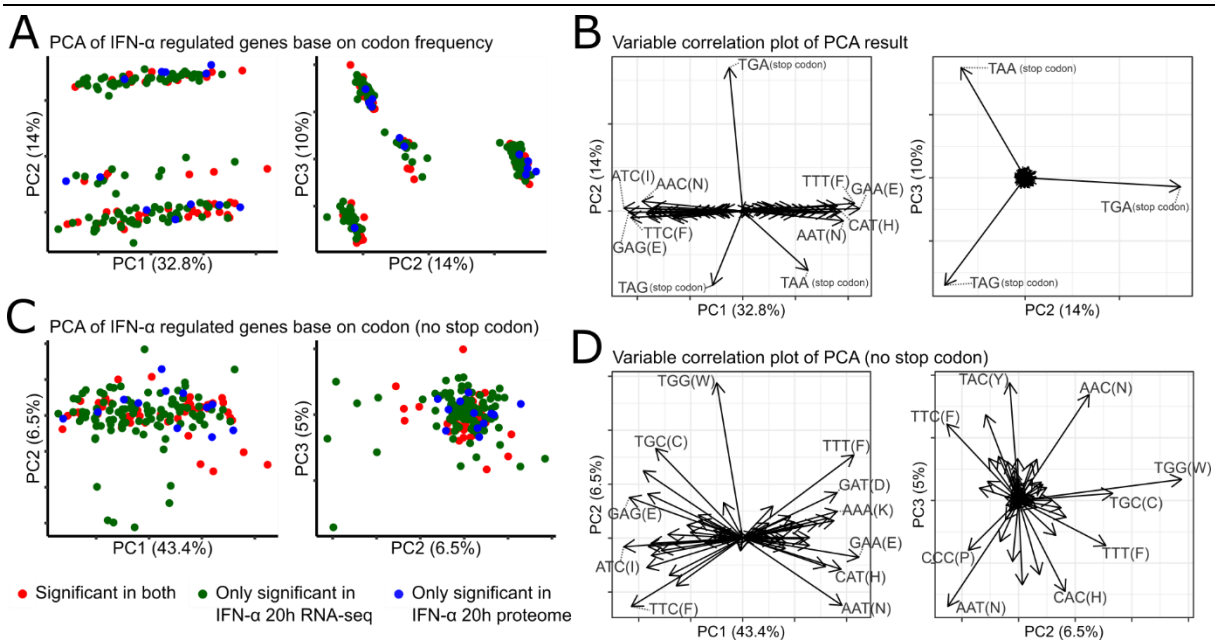


Figure.4.2.3 IFN-α regulated genes exhibited distinct groups based on bias in codon usage patterns. (**A**), principle component analysis (PCA) of codon frequencies in IFN-α regulated genes, all codons were included as variables. Genes that showed significant changes at RNA, protein, or both levels were colour labelled in green, blue, or red. (**B**), variable correlation plots showing contribution of each codon in the PCA result in panel **A**. (**C**), PCA of codon frequencies in IFN-α regulated genes, three stop codons were excluded. (**D**), variable correlation plots showing contributions of each codon in PCA in panel **C**.

Apart from analysis with the unsupervised machine learning (such as PCA), I also explored the correlation between codon usage bias and IFN-α response using supervised machine learning method. Unlike PCA which assess gene-to-gene difference in codon usage without prior grouping,

the supervised machine learning divides genes into pre-defined groups, and assess the feature(s) that maximally differentiate these groups. Briefly, DEGs were divided into three groups as described above, all 64 codons were taken as features, and the contribution of each codon in differentiating different groups of DEGs was estimated using random forest (RF) based feature selection method. When comparing "common DEGs" and "protein-only DEGs", codon CGA (encodes Arg) and GAC (encodes Asp) were shown to have high importance in differentiating these two groups (Fig.4.2.4 A, top). According to average codon frequencies in human RefSeq dataset[326], CGA is a rare codon that only covers 11.5% of arginine-encoding codons in human genome. "Common DEGs" also have low frequency of CGA codon, however "protein-only DEGs" showed increased usage of this codon, which could result in slower translation rate (Fig.4.2.4 B, left). The two codons that encode Asp (GAC and GAT) have almost identical frequency in human genome (50.2% and 49.8% respectively). "Common DEGs" showed similar pattern, but interestingly, "protein-only DEGs" had lower usage of GAC than GAT. It remains unclear though how this differential codon usage affects translation rates for "protein-only DEGs" (Fig.4.2.4 B, right).

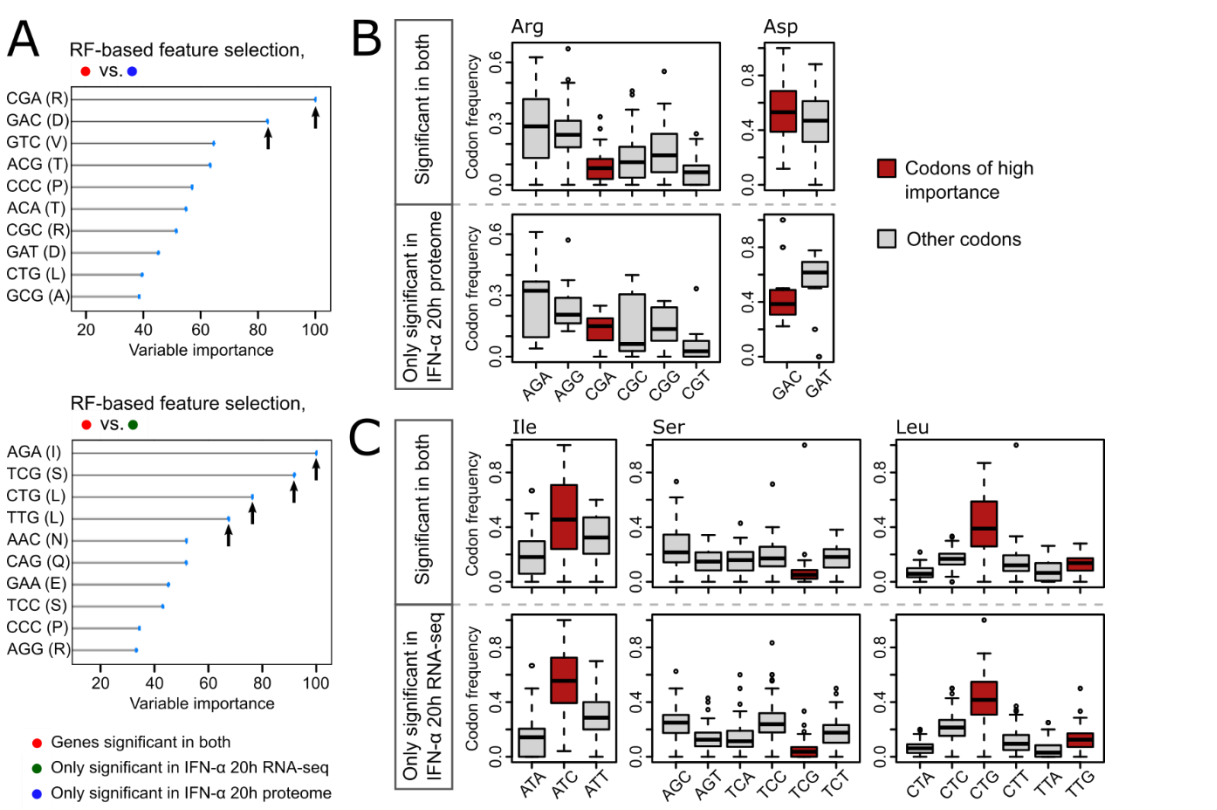


Figure.4.2.4 Supervised machine learning revealed codons that maximally differentiated genes of different IFN-α responses. (**A**), scaled variable importance for each codon in differentiating the designated groups. Only top-10 most important codons were shown in each contrast. Codons of

relatively high importance were arrow labelled. (**B, C**), boxplot showing frequencies of all codons that encode arginine or aspartic acid. Codons with arrow labels in panel **A** were filled red.

When comparing "common DEGs" and "RNA-only DEGs", codons AGA, TCG, CTG, and TTG showed high importance in differentiating these two groups (Fig.4.2.4 A, bottom). These codons encode for isoleucine, serine, and leucine. However, the frequencies of these codons in two groups of DEGs were found to be very similar, and are also similar to average frequencies in human genome (Fig.4.2.4 C). This observation showed that, even for the codons of highest differentiating power, "common DEGs" and "RNA-only DEGs" had little to no difference in codon usage.

4.3 Exploring IFN-α induced transcriptome regulations in alternative splicing and expression of long non-coding RNA

Apart from quantifying changes in mRNA abundance for protein coding genes, the high sequencing depth in this RNA-seq study (50 million reads per replicate) also allowed analyses of IFN-α induced transcriptome-wide regulations in different perspectives. The identified alternative splicing (AS) events were quantified based on event count, and subjected to differential analysis to compare AS events in IFN-α treated cells and mock cells using generalised linear modelling (GLM) method[327]. Under a cut-off line of FDR<0.05, in total 174 AS events showed statistically "significant" differential event occurrence. These consisted of 5 types of AS events: 53% of differential AS events were intron retention; one third of these events were single exon skip; alternative 3'/5'-splice site and multiple exon skip events were less frequent, comprising 6%, 4%, and 4% of differential AS events (Fig.4.3.1 A). Surprisingly, when cross-referenced to gene differential expression test results,

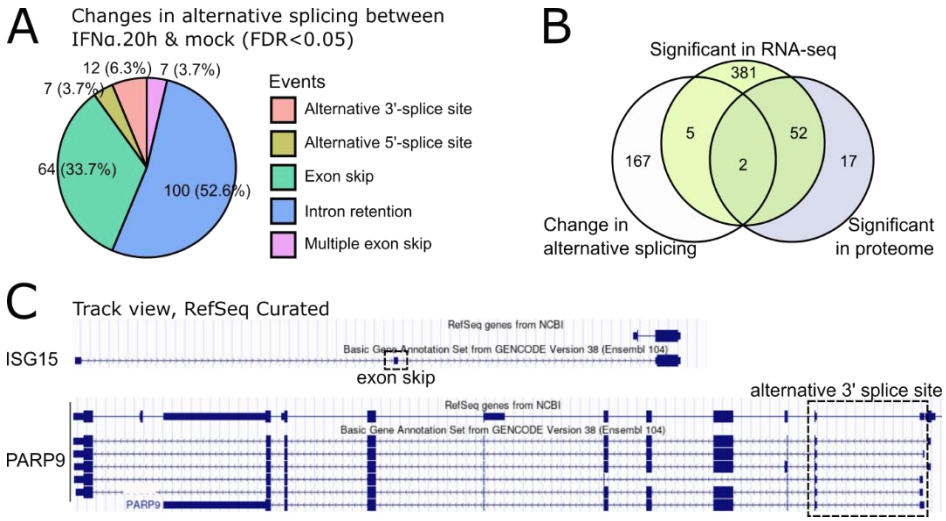


Figure.4.3.1 IFN-α regulated alternative splicing events had limited effect on protein-level IFN-α responses. (**A**), pie chart showing different types of alternative splicing events that were significantly regulated by IFN-α at 20h post-stimulation (FDR<0.05). (**B**), venn diagram comparing alternative splicing, RNA abundance, and protein abundance regulations induced by 20h of IFN-α stimulation. (**C**), IFN-α regulated alternative splicing events on ISG15 and PARP9. Intron were thin-lined, and open reading frame were extra-thick lines.

---

it was found that most of genes that had differential AS events were not regulated in abundance at either RNA or protein level (Fig.4.3.1 B). Only 7 differentially expressed genes had IFN-α regulated AS events, and among these genes, only ISG15 and PARP9 showed differential expression at protein level. In further examine, it was found that the IFN-α regulated AS event in ISG15 was a single exon skip at 5'-untranslated region (5'-UTR), and the AS event in PARP9 was an alternative 3'-splice site at 3'UTR (Fig.4.3.1 C). Both of these IFN-α regulated AS events had no effect on expression of protein isoforms.

The high coverage of this IFN-α RNA-seq dataset also allowed to study changes in abundance of long non-coding RNA (lncRNA). This RNA species lack canonical open reading frames, but have been revealed to participate in various regulatory processes[328]. Using a cut-off line of FDR<0.05, 55 lncRNAs were differentially regulated after 20h of IFN-α stimulation (Fig.4.3.2 A). These lncRNAs comprised 13% of "significant" RNA-level regulations at IFN-α 20h-post treatment, whereas protein-coding genes comprised 87% (364 out of 419) of RNA-level IFN-α regulations (Fig.4.3.2 B). Various types of lncRNA were found to be regulated by IFN-α, including 19 antisense RNA, 13 pseudogenes, 4 divergent transcripts (share the same promoter with coding gene and transcribe in the antisense direction), and 19 lncRNA that were poorly annotated (Fig.4.3.2 C). Interestingly, some of the IFN-α regulated lncRNA were found to be related to known ISGs – both mRNA and antisense RNA of IRF1 and IFITM3 were significantly upregulated; pseudogene and mRNA of PNPT1 were significantly upregulated with similar fold change; IRF2 and its divergent transcript were also upregulated by IFN-α (Fig.4.3.2 D, Table.4.3). Regarding genes that had no known role in interferon response, vesicular transporting protein NAPA and NAD metabolism enzyme NAMPT showed increased expression after IFN-α treatment in this study, and lncRNA related to these genes were also found to be regulated by IFN-α (Fig.4.3.2 D). These observed co-regulations of mRNA and related lncRNA indicated that these regulatory events could be functionally connected.

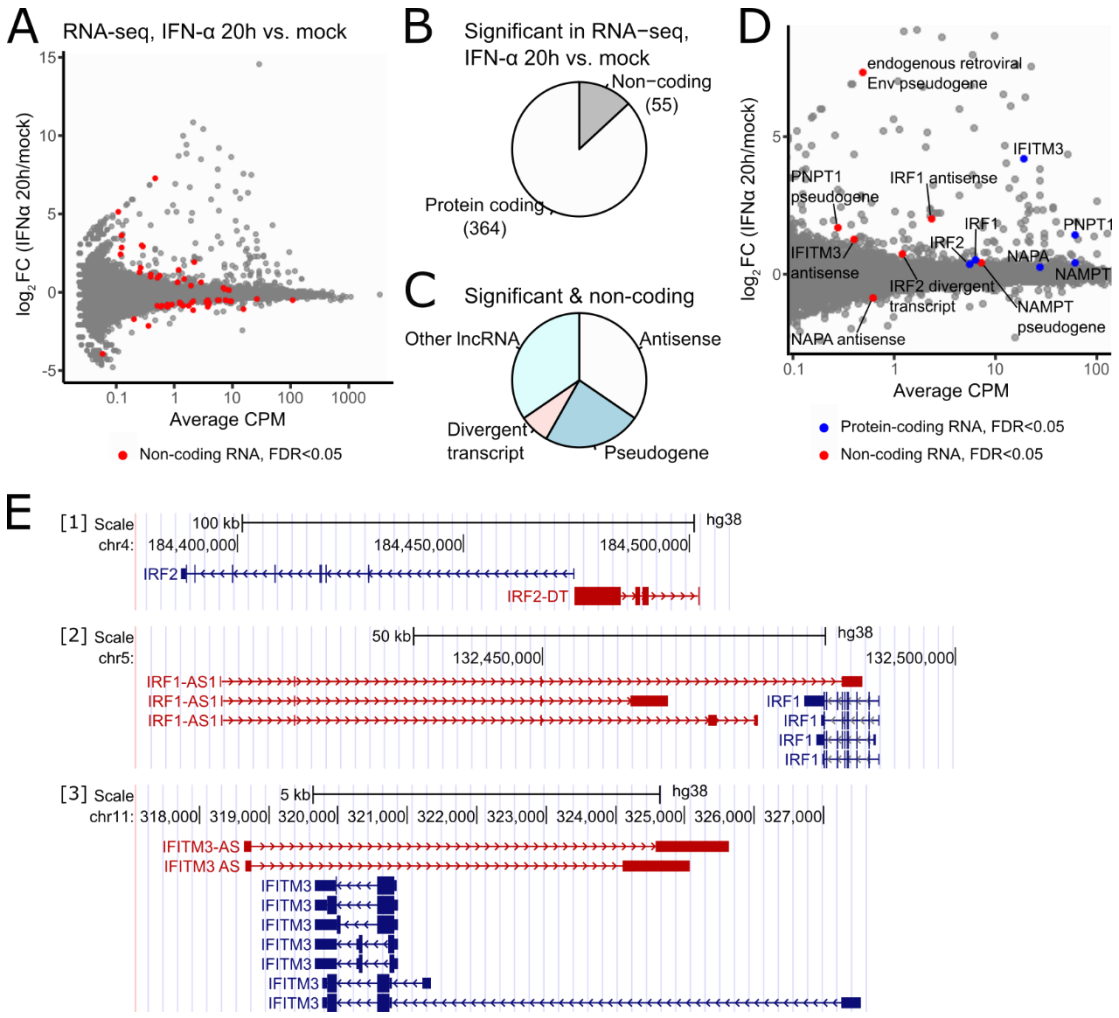


Figure.4.3.2 IFN-α regulated various types of long non-coding RNA (lncRNA). (**A**), MA plot of IFN-α 20h RNA-seq result, red highlight significantly regulated lncRNA. (**B**), pie chart showing proportion of protein coding and non-coding RNA that were significantly regulated by IFN-α at 20h post-treatment (FDR<0.05). (**C**), pie chart showing different types of lncRNA that were significantly regulated by IFN-α (FDR<0.05). (**D**), zoom-in view of MA plot of IFN-α 20h RNA-seq result, significantly regulated protein coding genes and their related lncRNA were highlighted in blue and red respectively. (**E**), genome architecture of known ISGs (blue) and IFN-α regulated lncRNA that were related to these ISGs (red).

Another interesting IFN-α regulated lncRNA is the pseudogene of endogenous retroviral gene (endogenous retroviral family W, Env(7), member 2). This gene showed the largest fold change among IFN-α regulated lncRNAs (Fig.4.3.2 D). This type of endogenous retroviral element was acquired from an ancient retrovirus and loss protein-coding ability during evolution[329]. Recent studies have shown that some endogenous retroviral (ERV) elements could participate in the regulation of innate immune response[330]. Overexpression of some ERV transcripts (i.e. Syncytin-1,

env-FC2) were found to activate innate immunity upon binding with cellular virus sensors MDA5[331]; while other ERVs such as ERVK-6 could inhibit innate immunity[332]. The strong overexpression of this endogenous retroviral lncRNA upon IFN-α stimulation indicated that it could be functionally involved in IFN-α response.

Table.4.3 IFN-α regulated lncRNAs that were linked to IFN-α regulated protein-coding genes

Gene ID	Relation to significant coding gene	Response to IFN-α
ENSG00000251661	Antisense to IFITM3	Upregulated
ENSG00000197536	Antisense to IRF1	Upregulated
ENSG00000268061	Antisense to NAPA	Downregulated
ENSG00000271646	IRF2 divergent transcript	Upregulated
ENSG00000229644	NAMPT pseudogene	Upregulated
ENSG00000229241	PNPT1 pseudogene	Upregulated

Although the functions for most of the lncRNA remained unknown, studies have revealed that some lncRNA could exert their regulatory functions co-transcriptionally to genes that had proximal localisation on the chromatin[333]. Interestingly, the mRNA encoding IFITM3, IRF1 and IRF2 were co-localised with IFITM3 antisense RNA, IRF1 antisense RNA, and IRF2 divergent transcript respectively (Fig.4.3.2 E). These mRNAs and lncRNAs were all significantly upregulated by IFN-α in this study, indicating that localisation and co-transcriptional regulatory mechanisms might also operate during IFN-α response in HEK293 cells.

Discussion

In this chapter, I provided an RNA-seq dataset that profiled the transcriptome landscape of IFN-α stimulated cells in the same cell type and treatment condition as proteome dataset in Chapter 3. In total, 126 and 419 genes showed statistically significant changes in this RNA-seq dataset at early (4h) and late (20h) timepoint post IFN-α stimulation, respectively. Current functional studies of ISGs focus mostly on genes that have increased RNA level in response to IFN-α[334,335]. However, the comparison of my IFN-α RNA-seq and proteome datasets revealed that a large population of genes exhibit discordance between mRNA- and protein-level changes after IFN-α stimulation. Genes that showed mRNA/protein discordance have a wide range of functions, include both known ISGs as well as genes that have no known role in IFN response. A number of genes were found respond to IFN-α only at protein level, including TAPBP, GDE1, and another 10 genes. Most of these genes have no previously known role in IFN response, marking them as interesting candidates for follow-up studies. Analysis of these genes in more cell types will help reveal whether their protein-only IFN-α response is specific to HEK293 cells or is a general phenomenon. Biochemical assays in virus-infected cells will help uncover their roles in innate immunity response. Notably, all of the downregulated proteins showed discordance in their mRNA changes, and *vice versa*. Various factors can lead to this strong discordance in IFN-α induced downregulations. From technical point of view, quantifications of low-abundance genes have higher measurement error, which can partially cause discordance. From biological aspect, proteins generally have longer half-life than mRNA, therefore temporal downregulations in mRNA will not result in changes at protein level[233]. Regulation of protein degradation rate is more effective in downregulating these long half-live proteins, and protein degradation processes generally function independent of mRNA abundance.

Although a large proportion of IFN-α regulated genes showed discordance in mRNA and protein level responses, we still found a >0.79 correlation between upregulation observed in the transcriptome and proteome at 20h post IFN-α stimulation. This is particularly evident among IFN-α upregulated proteins, as 90% (54 out of 60) of these proteins have corresponding upregulations at RNA level. Several studies have inferred that transcriptional regulation is the main driver of protein abundance changes, leading to high mRNA/protein correlations[296,336,337]. In a study that models dynamics of gene expression during cell's response to lipopolysaccharide, researchers combined RNA-seq and pulse-chase labelled proteomic data, and concluded that mRNA upregulations are driving protein-level changes for genes involved in gain of function[336]. This could explain the high mRNA/protein correlation observed in this study, because IFN-α induced antiviral response is a gain of function. In another study that investigated cell response to rapamycin, it was found that changes

in mRNA are strongly correlated with changes in protein abundance at later timepoints[337]. Similar result was present in this study. A >0.74 correlation was found between upregulations at early timepoint transcriptome and that at late timepoint proteome, suggesting that the increase in mRNA abundance with time-delay effect is the main driving force of protein upregulations.

However, these still cannot resolve why we observe high RNA-protein correlation values when a large number of genes showed no correlated RNA- and protein-level changes. Contradict to the explanation that change in mRNA level is the driving force in IFN-α response, correlation analysis in this chapter pointed that none of the examined features in RNAseq data has >0.5 correlation with IFN-α response at protein level (Fig.4.1.4). The most likely explanation behind these two seemingly opposite results is that, when comparing mRNA and protein upregulations, the type of correlation I assessed is between-gene correlation, which does not fully reflect the RNA/protein discordance for each gene. As mentioned in the Introduction, between-gene correlation compares RNA and protein abundance for all genes in a given system[298].This metric is not sensitive to gene expression changes because of the huge dynamic range of RNA/protein copy number[299]. For example, the mRNA and protein level of a highly expressed gene can be orders of magnitude higher than low-expression genes such as antiviral protein APOBEC3G. When APOBEC3G show discordant response, both its mRNA and protein level will remain lower than highly expressed gene, resulting in almost unchanged mRNA-protein correlation value. In this study, between-gene correlations were calculated based on RNA and protein fold changes, instead of absolute abundance. This measure can partially alleviate the dynamic range issue, and better reflect how changes in mRNA correlate with changes in protein level. However, dynamic range still has impact on these correlations because highly-expressed genes do not necessarily have the same magnitude of change as low-expression genes. The type of correlation that can better reflect gene's discordant response is the within-gene correlation, which compares RNA and protein abundance for the same gene in different systems[299]. In this study, however, the within-gene correlation cannot be estimated because of the lack of data points. This study only quantified RNA and protein abundance in HEK293 cells in 3 conditions (mock, IFN-α 4h, IFN-α 20h), while reliable estimation of within-gene correlation require more RNA and protein data points.

Another interesting observation from comparing RNA-seq and proteomic data is that the number of genes with protein-level responses was far fewer than that with RNA-level responses. Compared with transcriptome response where 419 genes showed significant changes at 20h post IFN-α stimulation, only 66 genes showed significant changes at protein level. This phenomenon is likely not restricted to HEK293 cells alone. As discussed in previous chapter, the number of significant hits differ among current proteomic studies in IFN-α response[153,154,276,338]. Significant hits

range from dozens to a few hundreds, and are substantially fewer than IFN-α responses observed by transcriptomics[232]. The fact that we observe much fewer protein-level changes than RNA-level changes could be a result of the "buffering" effect. Almost all forms of life live in a constantly changing environment, but a substantial part of environmental changes could be stochastic noise. Therefore mechanisms to buffer noise and variation is crucial for exerting accurate and robust biological processes[339]. Recent studies have highlighted that buffering mechanism at protein level plays a major role in maintaining cellular environment. In both prokaryotic and eukaryotic cells, studies have found that protein synthesis rate is more precisely and robustly correlated with a protein's stoichiometry in multi-protein complex than mRNA abundance[340,341]. Another evidence of buffering at protein level shows that cell can maintain its protein abundance, but not its mRNA abundance, when it has aberrant gene copy number[342]. Mechanisms leading to protein-level buffering are highly diverse. These include features examined in this chapter: codon composition, miRNA mediated regulation, RBP mediated regulation. Factors or processes such as 5'UTR structure, mRNA localisation, and protein degradation could all contribute to this buffering effect. In addition, system biology modelling has suggested that the limited energy and ribosome availability in the cell can also play a role[343]. The total amount of ribosome in fast-growing cells is limited[344], and this physical constrain can force overexpressed mRNA to queue for translation machineries, therefore result in a buffering effect without the need for active regulations[343]. However, not all RNA-protein discordance can be explained by this buffering effect. Dozens of known antiviral factors, such as IFITM1 and TRIM22[345-347], are highly expressed at mRNA level after IFN-α stimulation, but have no increase in their cognate proteins. These genes could have tightly controlled translation process, and require stronger or longer IFN-α stimulation to get expressed. Whether this hypothesis stand true require further experiments, for example evidence from ribosome occupancy data.

This chapter also provided an exploratory analysis of the potential regulatory processes underlying the mRNA/protein discordance using correlation test. Although none of the tested feature (3'UTR length, miRNA binding site, and IFN regulated miRNA) is highly correlated with mRNA/protein discordance, the analyses results indicated that IFN upregulated miRNA could be associated with the discordant response. Binding site of IFN regulated miRNA miR-155 is frequently found among genes that show discordant response. MiR-155 is a multifunctional miRNA, participate in blood cell differentiation, inflammation, and T-cell dependent antigen response[348]. It can be induced by lipopolysaccharide and type I IFN stimulation[349,350], infection of Epstein-Barr virus also lead to its upregulation[351]. The strong presence of its binding sites among discordant genes indicate that miR-155 could be more closely related to innate immune response than previously known. My analyses also showed that codon bias could contribute to the mRNA/protein discordance. "Protein-

only DEGs" tend to use rare codons of Arg and Asp more frequently, therefore could have a slower translation rate. However, whether this stand true require more supporting evidences, such as translation rate analysis via ribosome profiling data[352]. It is also worth noticing that this work only compared codon frequency of DEGs against the average codon frequency across human RefSeq database, therefore it cannot be ruled out that IFN-α stimulation could alter tRNA pool, leading to a different picture of codon bias. These correlation analyses also have several limitations. First, I only examined 3'UTR length, miRNA binding, and codon bias, while other regulatory processes or elements could also contribute to mRNA/protein discordance. For example, 5'UTR secondary structures are known to affect translation initiation[288]. In addition, RNA sequence motifs have closer correlation with RBP or antisense RNA binding sites than 3'UTR length. More correlation analyses using appropriate features or metrics are needed to assess whether these elements contribute to mRNA/protein discordance. Second, in correlation tests, I used the number of miRNA binding sites for all annotated miRNA sequences. This could "dilute" the contribution of miRNAs that cause discordant response. Solution to this issue is to perform the analysis at a larger scale, and assess the correlation for each miRNA individually. Third, different regulatory machineries could have synergy effect, for example crosstalk between miRNA and RBP[353]. This crosstalk could be co-operative, for example the RBP hnRNPD can facilitate substrate targeting of miRNA let-7[354]. It could also be antagonistic, exemplified by competitive regulation between the RBP HuR and miRNA miR-494[355]. These synergy effects are not detectable using my analysis approach, and identification of these effects require high-quality annotations and more advanced correlation analysis method.

Protein degradation is also a major factor in regulating protein abundance. However, the contribution from this aspect cannot be estimated base on my transcriptome and proteome datasets. Investigation of correlation between protein degradation and mRNA/protein discordance require further experiments that are designed for measurement of protein turnover rate. Metabolic labelling-based pulse-chase experiment coupled with mass spectrometry analysis is one of the most powerful tool to directly assess protein turnover rate in omic-scale[356]. This method has been implemented in a number of publications[336,357]. Apart from direct measurement of protein turnover rates, analysis of ubiquitination across whole proteome can also be insightful for understanding protein stability control. Ubiquitin-proteome pathway is responsible for degradation of a large proportion of cellular proteins, and recent technical advances have allowed identification of cellular ubiquitinated proteome in high sensitivity[159,358]. IFN-α regulation in alternative splicing and lncRNA expression were also assessed in this chapter. However, due to the scarcity of these events, their effects in determining mRNA/protein discordance cannot be estimated using statistical method. Emerging evidences highlighted that lncRNAs could either act in *cis* to recruit transcriptional

regulators to affect expression of ISGs co-transcriptionally, or act in *trans* as scaffolds for assembly of ribonucleoprotein complex that regulate expression of ISGs[359,360]. Other lncRNA, such as retroviral pseudogene, could also participate in modulation of cellular antiviral response[330,331]. These differentially expressed lncRNAs are also valuable targets for follow-up studies, because we only have limited understanding about their roles and mechanisms in cellular response to virus.

Acknowledgements

I want to thank all members of Castello lab for the supports in cell cultures, and thank Wilson lab and Robertson lab in Centre for Virus Research (CVR), University of Glasgow for collaboration works in RNA-sequencing and data analysis. Thanks Quan Gu for his collaboration works in differential expression analysis of RNA-seq data and supervised machine learning analysis of codon bias. Thanks Joseph Hughes for his discussions with me and Quan Gu regarding codon bias analysis.

Methods and materials

Methods

Cell culture

Culture and maintenance of HEK293 cells were the same as Chapter 2. Culture with IFN-α stimulation was the same as Chapter 3.

RNA sequencing

HEK293 cells were cultured on 6-well plates at ~60% confluency, IFN-α were added to dishes at 20 h, 4 h or not added prior to harvest. Total RNA were extracted from cells using Monarch Total RNA Miniprep Kit (NEB) following the manufacturer's protocol. Briefly, cells were gently washed with 4^{o}C PBS and lysed with 600 µl RNA Lysis Buffer provided in the RNA miniprep kit with gentle pipetting. Genomic DNA (gDNA) were removed using gDNA Removal Column with centrifuge, flow-throughs were supplied with ethanol and loaded on RNA Extract Columns with centrifuge. RNA on column was washed with Washing Buffer, then eluted with 100 µl nuclease-free water. Ribosomal RNAs were removed by rRNA Depletion Kit (NEB) following the manufacturer's protocol. Briefly, total RNA was hybridised with rRNA probes in rRNA Depletion Solution provided in the kit, then digested by RNase H at 37^{o}C for 30 min and by DNase I at 37^{o}C for 30 min. Extracted RNAs were supplied with RiboLock RNase inhibitor (Thermo Scientific) and subjected to SuperScript III reverse transcriptase (Invitrogen)

for cDNA library generation using random hexamer primers. RNase H was added to reverse transcription products to obtain cDNA library for RNA-seq.

RNA sequencing and differential expression analysis of RNA-seq results were performed by collaborator Dr Quan Gu in CVR, University of Glasgow. Libraries were pooled in equimolar concentrations and sequenced using a NextSeq 500 sequencer (Illumina). RNA-Seq reads quality was assessed using FastQC software (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). 89.1% of the reads generated presented a Q score of ⩾30. RNA-Seq reads were aligned to the Homo sapiens genome (GRCh38) downloaded via ENSEMBL using Hisat2[361]. After the alignment, software FeatureCount was used to count reads mapping to genes annotation files[362]. The *edgeR* package was used to calculate the gene expression level and to analyze differentially expressed genes between sample groups[363]. Apart from differential expressed genes analysis, software SplAdder was used for the quantification and differential testing of alternative splicing events[327].

Bioinformatic analysis

Pathway enrichments and matching of IFN-related GO annotation were the same as Chapter 3.

Correlation analyses between RNA-seq features and RNA/protein discordance were performed with R programming. Genes significantly upregulated in RNA-seq of 4h-post IFN-α stimulation were divided into two groups based on whether they have significant protein-level responses at 20 h-post IFN-a treatment in Chapter 3. Four features were selected and analysed independently: 1), log-2 Fold change; 2), log-10 count per million reads (CPM); 3), -log-10 FDR; 4), local Z-score. Local Z-scores were calculated by first ranking all genes in the IFN-α 4h RNA-seq result by their average CPM; next for each gene, take fold changes of its nearby 1,500 genes (ranked ±750 around it); then calculate this gene's Z-score based on its fold change against the average standard deviation of fold changes of its nearby genes; finally irritate this process for all analysed genes. Test of correlation was performed using base R function "cor()" by inputting the RNA-seq feature as continuous variable and RNA/protein discordance as binary variable.

Correlation analysis between of mRNA feature and RNA/protein discordance were performed with R programming. Genes significantly regulated in RNA-seq or proteomics results of 20h-post IFN-α stimulation were divided into three groups based on whether they were only significant in RNA-seq, only significant in proteomics, or significant in both datasets. Four mRNA features were selected and analysed independently: 1), 3'UTR length; 2), miRNA binding sites; 3), binding sites of IFN-α upregulated miRNA; 4), binding sites of IFN-α downregulated miRNA. Human genome annotations

from ENSEMBL was downloaded as GTF file, and imported into R by *rtracklayer* package[364]. For each gene, 3'UTR lengths of all isoforms were extracted from the genome annotation, and averaged to calculate its 3'UTR length. MiRNA binding sites were obtained by first importing miRNA-gene annotation file from online database miRDB_v6.0[324]; then matching analysed genes with this annotation table. IFN-α upregulated and downregulated miRNA was obtained from review article by Forster et al.[324]. The correlation between genes "Significant in both" and "Only significant in RNA-seq", and that between "Significant in both" and "Only significant in proteome" were calculated independently. Test of correlation was the same as mentioned above.

Analysis of codon bias was performed in collaboration with Dr Quan Gu from CVR, University of Glasgow. The usage of the existing 64 codons, used as features, was calculated in each nucleotide sequence across the human genome, and available from published study[365]. In unsupervised codon usage analysis, codon frequencies were input as features and genes as objectives into PCA using base R function. PCA results were visualised by drawing each gene's coordinates in interested PCs into scatter plots. Variable correlations were visualised by drawing each codon's rotations in interested PCs into arrow plots. In supervised codon usage analysis, we calculated the feature importance for the classification of different groups (mock vs IFN-α) based on the gradient boosting machine (GBM) model from the feature selection package *Caret*[366]. Feature importance was normalised into percentiles of the maximum importance value prior to plotting.

Materials

Chemicals and reagents

REAGENT or RESOURCE	SOURCE	IDENTIFIER
DMEM	Gibco	#11995-065
FBS	Sigma Aldrich	F9665
PBS 1x, pH 7.4	Gibco	#10010023
penicillin/streptomycin	Sigma Aldrich	P4458
Trypsin EDTA	Sigma Aldrich	T3924
Interferon alpha 2a	Pbl assay science	#11100-1
Monarch Total RNA Miniprep Kit	NEB	T2010S
NEBNext rRNA Depletion Kit	NEB	E6310
RiboLock RNase Inhibitor	Thermo Scientific	EO0381
SuperScript III Reverse Transcriptase	Invitrogen	#18080093

Instruments

INSTRUMENTS	SUPPLIER/INSTITUTE
T100 Thermal Cycler	Bio-Rad
NextSeq 500 sequencer	Illumina

Cell lines

The same as Chapter 3

Chapter 5

Phosphoproteome and RNA-bound proteome profiling of IFN-α stimulated cells revealed protein abundance-independent remodelling of the cellular proteome

Introduction

Protein phosphorylation is a class of post-translational modification (PTM) that is widely involved in various signalling cascades and regulations of protein activity, interactions, localisation, and stability[274]. Interferon signalling response is heavily dependent on phosphorylation. In canonical interferon signalling, the IFN-bound receptor activates tyrosine kinase JAK1/TYK2 which phosphorylate tyrosine 701 in STAT1 and tyrosine 690 in STAT2. The activated STAT proteins then form transcription initiation complex that drives expression of interferon stimulated genes (ISG)[6,205]. Beside this classical model, several studies have reported multiple non-canonical phosphorylation events in interferon response. These phosphorylation sites distribute in both canonical and non-canonical interferon signalling pathways. In the canonical interferon signalling mediated by JAK-STAT pathway, phosphorylation of serine 727 in STAT1 has been reported upon stimulation of either type I or type II IFN to promote innate immune response[97,98,367]. Phosphoserines in receptors IFNAR1 (serine 535) has also been reported to be induced by interferon[368]. The non-canonical interferon signalling pathways are initiated by activating tyrosine kinases JAK1/TYK2, but then diverge from canonical pathway as STAT1 phosphorylation is not involved in the non-canonical pathways[369]. So far, the main players we know in non-canonical interferon signalling are the MAP kinase (MAPK) pathway and the mammalian target of rapamycin (mTOR) pathway[369,370]. MAPK pathway is reported to mediate growth inhibitory effects in IFN stimulated cells through phosphorylation of threonine 180 and tyrosine 182 in p38 MAPK[371-373], and threonine 209/214 in MNK1[374]. MAPK pathway is also reported to negatively control canonical interferon signalling through modification of SPRY2 phosphoserine 112/121[375]. The known role of mTOR pathway in interferon signalling includes regulation of translation initiation and apoptosis[100,376,377]. Phosphorylation sites associate with translation initiation include threonine 421 and serine 424 on p70 S6 kinase[100], while sites mediating apoptosis remain poorly understood. It has been proposed that cooperation between these canonical and non-canonical pathways are important for effective activation of innate immune response[6], however, analysis of phosphorylation in IFN-stimulated cells has not been performed systematically.

Profiling of phosphorylation on amino acid residues within protein has been challenging for decades. In early studies, identification of phosphorylation sites heavily relies on biochemical assays such as ^{32}P isotope radiolabelling, antibody probing, or chromatography analysis of hydrolysed amino acid pool[378,379]. These methods generally suffer low throughput, sometimes only analyse one protein and one phosphosite at a time, or/and might have low sensitivity issues. In the last two decades, developments in mass spectrometry (MS)-based phosphopeptide profiling have allowed

the identification of phosphosites in a proteome-wide scale with high sensitivity[379-382]. This type of profiling require enrichment of phosphopeptides, which is commonly performed with affinity purification exploiting the high affinity interaction between phospho-groups and multivalent cations, such as Fe^{3+}, Ga^{3+}, Ti^{4+}, etc[383-385]. The phosphorylation of peptides can be detected through a mass shift of 79.983 Dalton in the mass spectrometer scans. Series of MS-based phosphoproteome analyses have been performed at cell lines[382,386] and tissue level[387,388]. These studies have dramatically expanded the realm of known phosphorylation sites, which would not be otherwise detected in biochemical assay-based studies. However, the landscape of phosphorylation events in IFN-treated cells remains largely unexplored.

RNA-binding proteins (RBP) are a heterogeneous group of cellular proteins that play a prominent role in the innate immune response against virus. Virus sensing machineries are largely composed of RBPs in order to detect viral RNA genomes or other types of virus-derived RNA[5]. For example, the well-characterised virus sensors RIG-I and MDA5 are both RBPs. MDA5 is known to recognise long-dsRNA (>1kb)[389], and RIG-I senses 5'-triphosphorylated ssRNA and short-dsRNA (<1kb)[390]. Other examples of virus-sensing RBPs are TLR3/7/8 that detect extracellular viral RNA[55], and IFIT1-5 that bind to 5'-tri-phosphorylated viral RNA[391]. A large proportion of proteins that exert antiviral activities are also RBPs, for example OAS family members OAS1-3 and OASL that trigger the cleavage of viral RNA by RNase L[392], RNA editing protein ADAR that impairs viral genome by introducing A-to-I mutations[393], and PKR which halt viral protein synthesis through inhibition of translation initiation factors[394]. Consistent with this knowledge, RBPs were found to be highly enriched among IFN-α regulated proteins identified in whole-cell proteome (WCP) analysis (Chapter.3, Section.3). Apart from antiviral factors, RBPs also have important roles in virtually all steps of virus life cycle, including viral RNA synthesis, viral protein translation, nucleus-cytoplasm trafficking, viral factory assembly, and viral particle packaging[168,323,395]. Recent studies have reported that the cellular RNA-binding proteome (RBPome) undergoes global alterations after virus infection – as much as 247 RBPs showed altered binding activities in SINV infected cells[151], and 335 RBPs were regulated in SARS-CoV-2 infected cells[170]. These RBPs cover a broad range of processes, including mRNA splicing, tRNA splicing, rRNA processing, translation initiation, and RNA decay. However, it is unclear which part of the RBPome changes were driven by the virus or due to an active host response to virus invasion.

Conventionally, identification and functional investigation of RBPs relied on biochemical assays, such as electrophoretic mobility shift assay (EMSA), which focus on a limited number of proteins at a time. Recent advances in methodology employs mass spectrometry (MS)-based analysis and allows the profiling of hundreds of RBPs in one experiment[10,396-398]. RNA interactome

capture (RIC) pioneered comprehensive RBPome analyses with quantitative proteomics followed by combining ultra-violet (UV) protein-RNA crosslinking and RNA isolation with oligo(dT) magnetic beads[399]. It offers several advantages over other MS-based RBP profiling methods: 1), the crosslinking event is initiated by UV light which is non-invasive and has high penetrance, therefore it can be applied in living cells or tissues to crosslink RNA-protein interactions *in vivo*[400-403]; 2), the crosslinking bond is formed by nucleobase-driven attacks to zero distance amino acids, therefore it does not promote protein-protein crosslinks; 3), different from methods that uses immobilised RNA probes, the UV-crosslinking and pulldown in RIC globally captures RNA in its native conformation and with its physiologic partners; 4), the crosslinking is mediated by covalent bond which is stable under stringent washing conditions, therefore contaminant and non-crosslinked proteins can be efficiently removed by washes using buffers containing ionic detergent and high salt. These advantages make RIC a versatile approach for compositional analysis of the RBPome in different biological contexts, such as Drosophila development[402] and in virus infection[151,170].

Previous chapters have investigated IFN-α induced regulation in gene expression at RNA and protein levels. Beyond these, IFN-α could also regulate cellular processes in a gene expression independent manner, for example, through posttranslational modifications. In this chapter, I focused on investigating the effect of IFN-α on protein phosphorylation and RNA-binding activity. Using mass spectrometry-based phosphoproteomic analysis, phosphorylation landscape of IFN-α stimulated cells could be profiled to study interferon signalling responses in full picture, and IFN-α dependent regulations of phosphorylation in other cellular proteins. Using RIC, I also profiled the effects of IFN-α on the responses of the cellular RBPome. These analyses results provided new insights into the protein abundance independent regulation of cell functions induced by IFN-α, and highlighted putative regulator of host-virus interactions.

Results

5.1 Deep phosphoproteome profiling revealed a wide impact of IFN-α on the cellular phosphorylation landscape.

Time-dependency of the IFN-α induced antiviral response was measured in previous chapter (Chapter.3, Section.1). IFN-α treatment for 4h showed emerging antiviral activity, and 20h treatment almost completely prevented SINV infection in HEK293 cells. These two timepoints were selected for proteomic analysis. However, western blot analysis showed that IFN-α induced response in phosphorylation could happen at a much earlier timepoint, followed by a gradual decay of the signal overtime (Fig.5.1.1). This was not surprising as signalling cascades mediated by phosphorylation are fast-responding processes, therefore, timepoint shortly after IFN-α treatment could provide valuable insights into the progression of signalling cascade. Considering this, the early and late timepoints for phosphoproteome profiling of IFN-α stimulated cells were set at 10min and 4h post stimulation. The mock and 4h post-treatment samples in phosphoproteomic analysis were collected from the same cell culture dish as the IFN-α WCP analysis in order to keep consistency between experiments (Fig.2.1.2 A). Similar to WCP, SILAC labelling with label permutation was introduced to account for bias in labelling (Table.5.1).

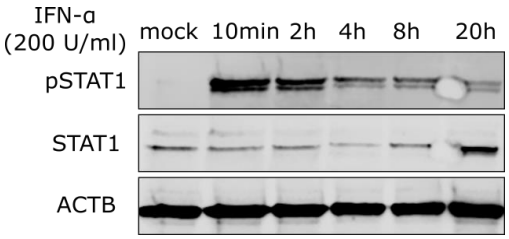


Figure.5.1.1 IFN-α induce phosphorylation of cellular protein in HEK293 cells. The pSTAT1 bands were probed against phosphotyrosine 701 of STAT1.

Table.5.1 Label swap and sample combine strategy in IFN-α phosphoproteomic experiment.

	Sample (isotope label)
Rep.1	mock (L) + 10min (H) + 4h (M)
Rep.2	mock (M) + 10min (L) + 4h (H)
Rep.3	mock (H) + 10min (M) + 4h (L)

Phosphoproteomic analysis requires enrichment of phosphopeptide before MS analysis. Effective enrichment can be achieved based on highly selective interaction between phospho-group and multivalent titanium ion (Ti^{4+}), but the different choice of ion carrier could affect enrichment results[158]. Ti-IMAC is a phospho-group enriching reagent where Ti^{4+} ions were fixed on beads with chelating agent[25]. In a pilot experiment using Ti-IMAC for phosphopeptide enrichment in HeLa lysate, 3241 unique phosphopeptides could be identified in a 60min LC-MSMS scan. The ratio between numbers of phosphopeptides versus non-phosphorylated peptides reached 51.4, showing a high selectivity in enriching phosphopeptide.

Deep phosphoproteomic analysis was achieved using offline fractionation before Ti-IMAC phosphopeptide enrichment, similar to the procedure applied in deep proteomic analysis in previous chapters (Chapter.2, Section.3 & Chapter.3, Section.2; Fig.5.1.2 A). Absorbance profiles at 214nm were highly similar between replicates, confirming high reproducibility during sample preparation and fractionation (Fig.5.1.2 B). Samples were divided into 10 fractions, which were analysed with two mass spectrometers. The two instruments used in this this study were Q-Exactive and Orbitrap-

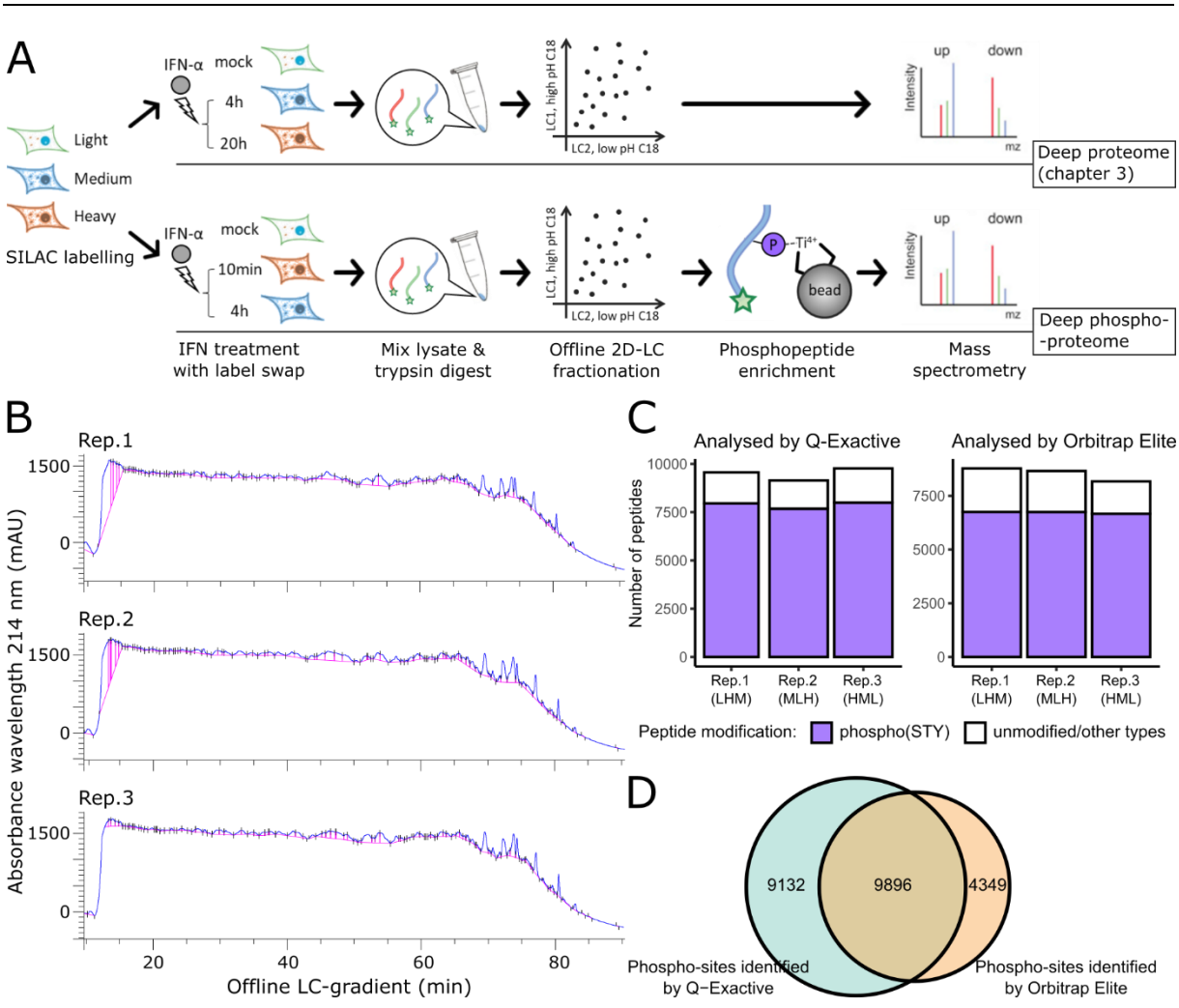


Figure.5.1.2 Phosphoproteomic profiling of IFN-α stimulated cells were generated from two mass spectrometer instruments with high depth of detection. (**A**), schematic workflow of deep phosphoproteome profiling. (**B**), chromatogram of offline-LC gradient with reverse-phase separation column in high pH. Eluates from 10 to 90 min gradient were collected and divided into 10 fractions. (**C**), barplots showing numbers of phosphorylated and non-phosphorylated peptides identified in each SILAC sample from two mass spectrometers, Q-Exactive and Orbitrap Elite. (**D**), venn diagram showing numbers of phosphorylation sites identified from two mass spectrometers.

-Elite, they both use the orbitrap as mass analyser for MS1, but have distinct approaches for peptide fragmentation, varying in both activation technique and the choice of mass analyser for detection. These differences endow them with complementary capabilities for identifying phosphopeptides. When examining search results, both instruments showed consistent phosphopeptide identifications across three replicates (Fig.5.1.2 C). In total 9,896 phosphosites were commonly identified by both instruments, and interestingly, 9,132 and 4,349 phosphosites were only detected in Q-Exactive and Orbitrap Elite respectively (Fig.5.1.2 D).

Analysing phosphoproteome with two instruments improved the coverage of phospho-peptide identification, but quantitative analysis using these two datasets required caution. Similar to quantitative proteomic analysis in previous chapters, phosphosites that had high proportion of missing values were excluded, and only phosphosites that were identified in all 3 replicates in at least one condition (mock or IFN-α treated) were considered for the quantitative analysis. In total 8,219 phosphosites remained after applying this stringent filter (Fig.5.1.3 A). This was not surprising because unlike proteomic analysis where multiple peptides could map to one protein, identification of phosphosites often relies on single detection events and are more affected by stochastic fluctuations. In total 3,496 phosphosites were quantified by both Q-Exactive and Elite, and 1,874 and 2,849 sites were uniquely quantified in respective instrument, confirming that these two datasets were highly complementary (Fig.5.1.3 B). For phosphosites that were not determined “quantified”, those that were detected in multiple replicates and only detected in one condition (IFN-treated or mock) were also retained as “semi-quantitative” changes, because these sites could be turned “ON” or “OFF” by IFN-α. In total 94 phosphosites were included in this category (Fig.5.1.3 A). Notably, data generated by the Q-Exactive and Elite instruments possess distinct intensity distributions (Fig.5.1.3 C,D), therefore requiring each instrumental dataset to treated independently.

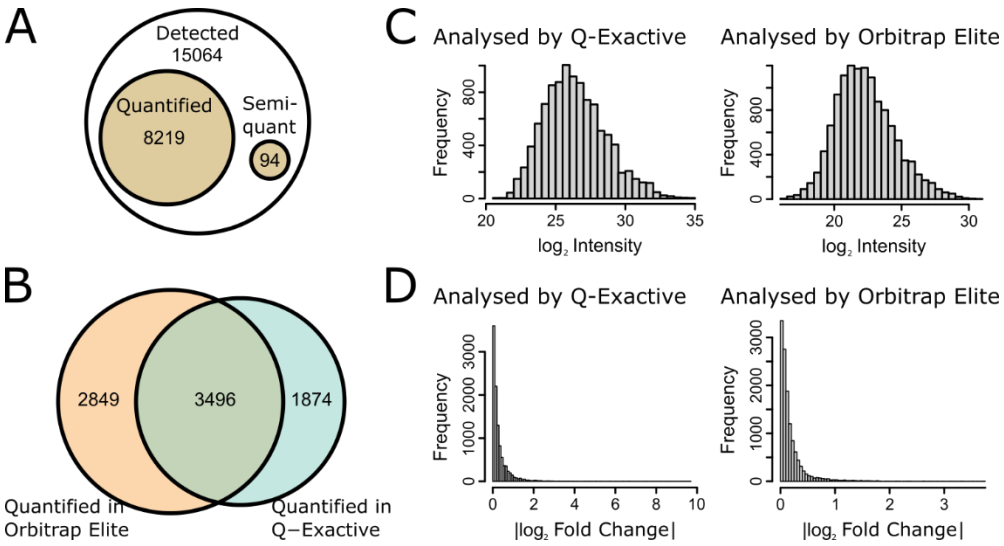


Figure.5.1.3 Phosphoproteomic data from two instruments showed distinct distributions and required separated analysis. (**A**), venn diagram showing number of phosphosites that were identified or quantified in this experiment, including data from both instruments. Phosphosites identified in all 3 replicates in at least one condition were considered "quantified". Phosphosites identified in 2 replicates and only identified in one condition were considered "semi-quantitative". (**B**), venn diagram showing number of phosphopeptides that were quantified in data generated from each instrument. (**C**), distributions of intensities for phosphosites quantified by each instrument. (**D**), distribution of absolute value of log-2 fold change for phosphosites quantified by each instrument.

Search results from Q-Exactive and Orbitrap Elite datasets were tested separately using moderated T-statistics, and the test results were combined *a posteriori*. The test pipeline for each dataset was analogous to that applied for IFN-α WCP analysis (Chapter.3, Section.2). As expected, when visualised by volcano plots, the test results from Q-Exactive data and Elite data differed considerably in both the scale of fold change and p-value (Fig.5.1.4 A). After individual test, these two test results were combined in following steps: 1), if a phosphosite was quantified only in one dataset, it was directly accepted into the combined result; 2), if quantified in both datasets, the result that had substantially higher phospho-group localization probability was accepted; 3), if both results had similar phospho-group localization probability, the result with higher phosphopeptide identification score was accepted; 4), if both results showed high identification scores, I accepted the result that had higher signal-to-noise level measured by the percentile rank of intensity (Fig.5.1.4 B).

In summary, the combined dataset contained 8312 phosphosites, where 624 and 618 sites showed p-value lower than 0.05 after 10min and 4h of IFN-α treatment, respectively. Among these phosphosites, 37 sites were located in known ISGs, including IFN-α receptor IFNARs, interferon

signalling transmitter MAVS, and ISG-driving transcription factor STAT1 (Fig.5.1.4 C). Phosphoserine at STAT1 position 727 was reported to be critical in response to IFN-α response[97], and was also upregulated at 4h-post IFN-α stimulation. Notably, when using statistical pipeline built for quantitative proteomic analysis, only 14 sites showed changes post-treatment with FDR lower than 0.1. This low confidence in detected changes could be explained by the fact that quantification of phosphopeptides often rely on very small number of detection events, which is vulnerable to stochastic fluctuations. As a result, quantitative phospho-proteomic analysis is inherently noisier than WCP analysis, therefore the T-statistics-based pipeline for proteomic data analysis is less effective in processing phosphoproteomic data. One widely-used alternative to T-statistic in SILAC-based proteomic data analysis is Z-statistic, which performs tests based on average ratio between tested conditions (IFN-α treated versus mock)[211,266]. Based on this method, 108 and 133 phosphosites showed changes in IFN-α 10min and IFN-α 4h treatments with FDR lower than 0.05.

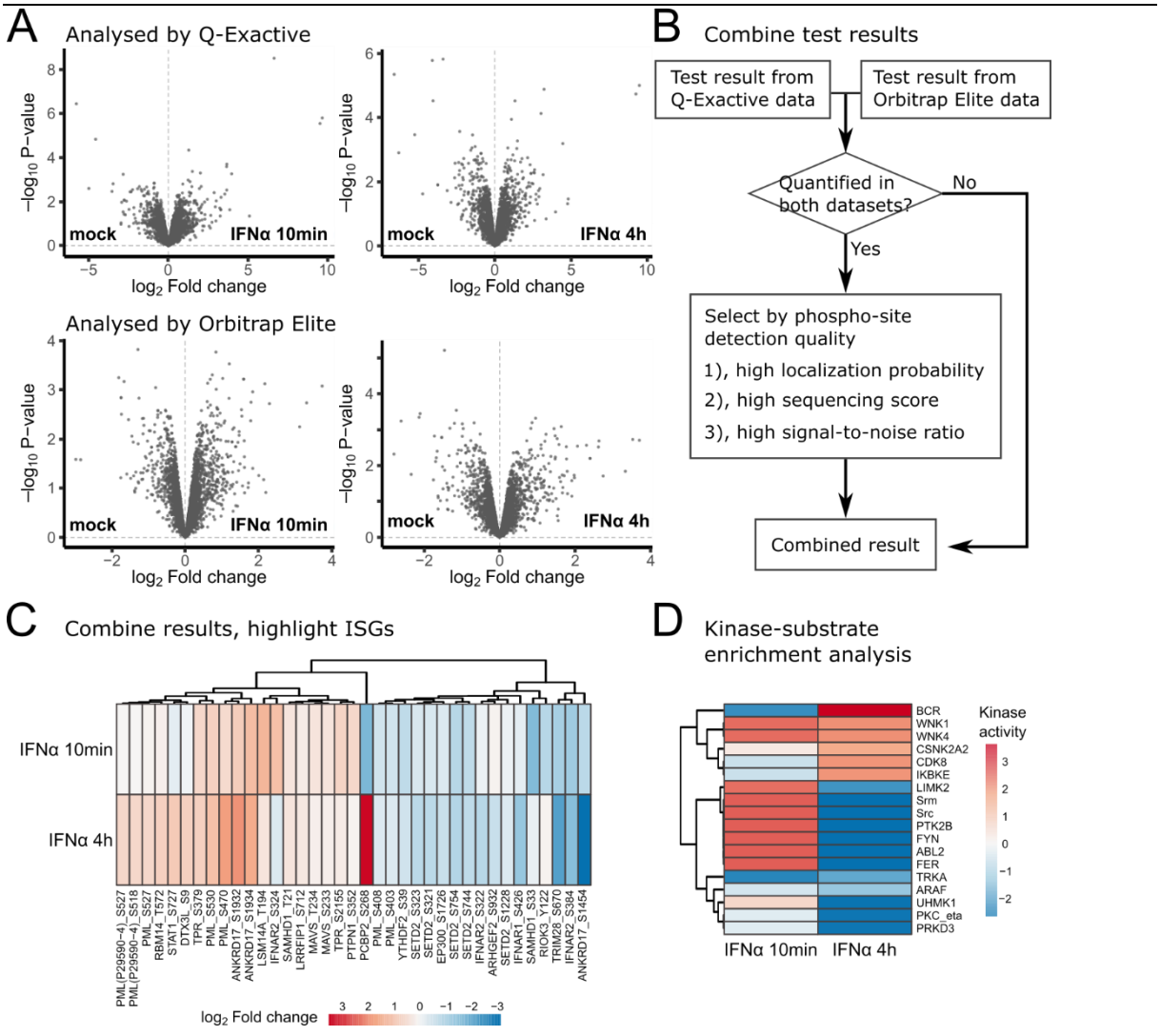


Figure.5.1.4 Combined quantitative analysis result revealed temporal landscape of IFN-α regulated

phosphorylation. (**A**), separated test result from each dataset. (**B**), flow chart of combining test results. (**C**), heatmap showing phosphosites with p-value <0.05 in known ISGs. (**D**), kinase-substrate enrichment analysis (KSEA) of quantified phosphoproteome in IFN-α stimulated cells.

To further investigate the processes involved in IFN-α regulation, enrichment analysis was performed on proteins that contain IFN-α regulated phosphosites. When using all phosphoproteins (irrespective of fold change) identified in this study as background, no enrichment in known signalling pathway was found. In another attempt to investigate pathways involved in IFN-α regulated phosphorylation, all quantified phosphosites were subjected to kinase-substrate enrichment analysis (KSEA). This method estimates activity of a kinase base on the average fold change of all phosphosites that it could regulate, and calculates the significance of changes in activity level of annotated kinases via Z-statistics. A combination of multiple kinase-substrate annotation databases was used to assign phosphosites quantified in this study to known kinases[404-410]. In total 11,035 kinase-substrate correlations were found that mapped 390 kinases to 4,698 sites in this phosphoproetome data. Through KSEA, 18 kinases showed "significantly" altered activities after IFN-α stimulation by a cut-off of FDR<0.05 (Fig.5.1.4 D). These kinases displayed no enrichment in specific pathway, however, several of these kinases were shown to be linked to immune response. These included IKBKE which could activate IRF3 and STAT proteins[76], and WNK kinases which could be involved in multiple immunity processes and antiviral response[4,411].

Not all phosphorylation have biological function in signalling pathways. When examining all proteins that contain phosphosites that could be up- or down-regulated by IFN-α at either 10min or 4h post treatment, it was found that these phosphoproteins were enriched in both viral protein interactors and viral RNA interactors (Fig.5.1.5 A). The compilation of viral protein interactors and viral RNA interactors were consolidated from publications by the same criteria as in previous chapter (Chapter.2, Section.4). Within the viral RNA interactors that contain IFN-α regulated phosphosites, multiple proteins had known link to innate immunity or role in virus infection (Fig.5.1.5 B), for example PCBP2 which could negatively regulate antiviral response[412] and initiate viral RNA replication in Poliovirus infection[413], and LSM14A which participates in initiation of antiviral response[414]. Known regulators of virus infection were also overrepresented in viral protein interactors that contain IFN-α regulated phosphosite (Fig.5.1.5 C), including ANKRD17[415], PCBP2[413], TRIM28[416], etc. These discoveries indicated that apart from known interferon signalling events, IFN-α regulated phosphorylation could be involved in host-virus interactions on a much wider range.

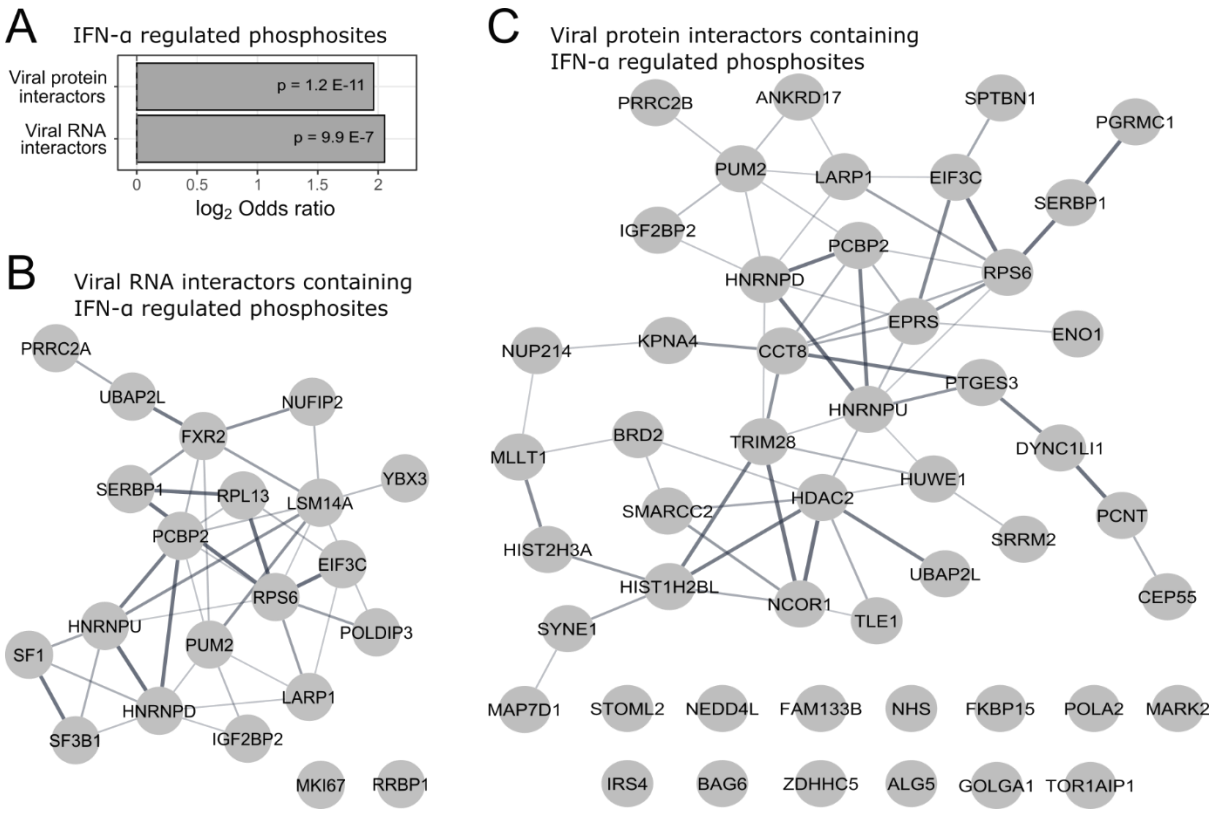


Figure.5.1.5 IFN-α regulated phosphorylation sites exhibited enrichment in viral protein and viral RNA interactors. (**A**), enrichment of viral protein or viral RNA interactors among proteins that contain IFN-α regulated phosphosites, which were determined by FDR < 0.05 (Z-statistics) at any timepoint, including both up- and down-regulations. (**B**), network map showing viral RNA interactors that contained IFN-α regulated phosphosites. (**C**), network map showing viral protein interactors that contained IFN-α regulated phosphosites.

5.2 Characterisation of cellular RNA-binding proteome in IFN-α stimulated cells.

Cellular RBPs play a central role in the regulation of virus infection. To investigate whether the host could actively regulate RNA-binding activities of its cellular RBPs upon IFN-α stimulation, cultured cells were either mock-treated or stimulated with IFN-α for 20h, and subjected to RNA interactome capture[399]. Briefly, IFN-α treated or mock treated cells were crosslinked by UV light to induce covalent bonds between cellular RNAs and proteins that directly bind to them. Crosslinked cells were lysed in a denaturing condition to reduce non-specific RNA-protein interactions while maintaining the stable RNA-protein adducts produced by UV crosslinking. The repertoire of poly(A)

RNAs were captured with oligo(dT) beads along with their bound proteins. These RBPs were then released by RNase digestion, and subjected to protease digestion and mass spectrometry analysis (Fig.5.2.1 A). Silver staining revealed a visible protein patterns in eluates from UV crosslinked samples, resembling poly(A)-RNA-bound proteins. By contrast, eluates in absence of UV crosslinking eluates were largely depleted of proteins, evidencing successful removal of proteins not covalently crosslinked to RNA (Fig.5.2.1 B). After MS data acquisition, quantitative proteomic analysis of crosslinked versus non-crosslinked samples was performed using moderated T-test after log-transformation and imputation of missing values with background-level signal. As visualised in volcano plot, the test results of crosslinked versus non-crosslinked data largely coincided with silver staining result: the non-crosslinked samples was almost depleted of proteins, while the UV-crosslinked samples showed clear enrichment of hundreds of RBPs (Fig.5.2.1C).

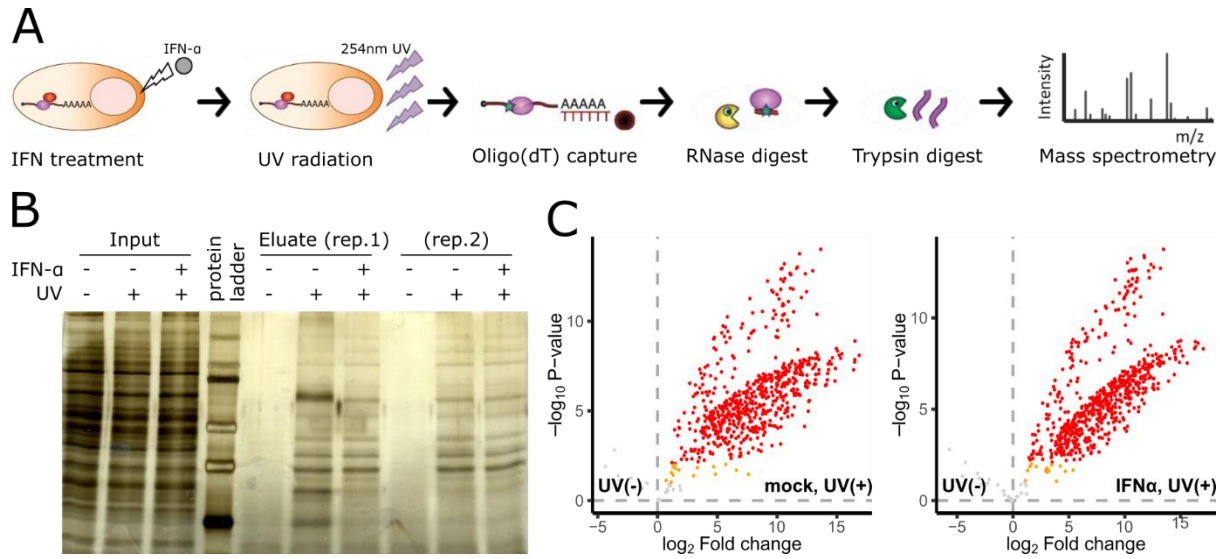


Figure.5.2.1 Cellular poly(A) RNA-bound proteome (RBPome) after IFN-α stimulation revealed by RNA interactome capture (RIC) experiment. (**A**), schematic of RIC pipeline. After IFN-α treatment for 20h, cells were crosslinked by UV light to induce RNA-protein adducts. The repertoire of poly(A) RNA-bound proteins were pulled down with oligo(dT) beads, then analysed by mass spectrometry after RNase and protease digestion. (**B**), silver staining of input and eluate of oligo(dT) capture samples. (**C**), volcano plots visualising poly(A) RBPome in each condition, RBPs were colour labelled by false discovery rate (red: FDR<0.01, orange: FDR<0.1).

RBPs in each condition were determined by a cut-off line of false discovery rate (FDR) <0.1 in the corresponding UV(-) versus UV(+) test. Under this criteria, 638 and 665 proteins were determined as RBPs in mock and IFN-α stimulated cells respectively. These two groups of proteins

were collectively defined as the RBPome identified in this experiment. When searched against Gene Ontology (GO) database, over 85% of RBPs identified in each condition were annotated to have RNA-binding activities (Fig.5.2.2 A). These high proportions of known RBPs indicated successful capture of high confidence RBPome in this RIC experiment. When comparing RBPs identified in IFN-α treated and mock conditions, most of RBPs in these two conditions overlapped, whereas 49 and 22 RBPs were uniquely identified in IFN-α stimulated and mock conditions (Fig.5.2.2 B). This indicated that IFN-α treatment could result in changes in poly(A) RNA-bound proteome. However, it needed to be further analysed using appropriate test method to compare RBPs identified in IFN-α treated and mock conditions.

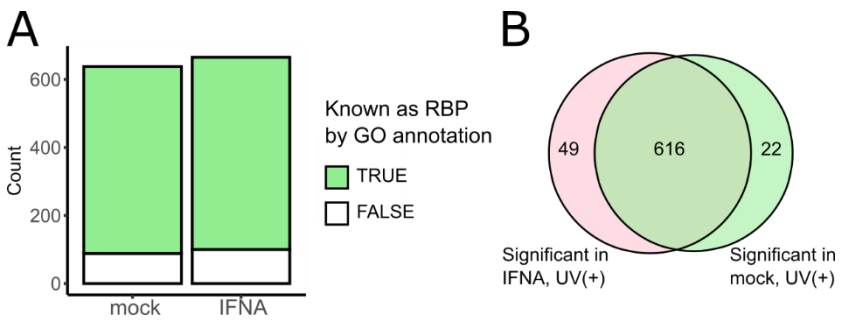


Figure.5.2.2 RIC experiment identified hundreds of known RBPs. (**A**), barplot showing number of proteins known to interact with RNA based on GO annotations. (**B**), venn diagram comparing RBPs in each condition.

5.3 Development of novel analysis method that partially impute data and utilise empirical Bayesian estimation to improve robustness of statistical test against missing values.

In proteomic data analysis, one of the most widely used methods to handle missing values is to impute them by a minimal deterministic value (MinDet) that resembles a background-noise level signal[214]. This imputation method was used in WCP and phosphoproteome analysis. It has proven useful in various proteomics works, because missing values in MS analysis are common due to peptides often falling below the instrument detection limit. However, the occurrence of missing values in MS data cannot be solely explained by low abundance. Various random factors could also result in missing values in high-abundance proteins, such as stochastic fluctuations in sampling and misidentification of peptides from MS2 spectra[417]. When missing values derived from these random effects occur, imputation by MinDet method could be detrimental to quantification of those

proteins, because the imputed value could largely deviate from observed values in other replicates. This issue will increase with size of dataset, because missing value in any of the replicate would impair test result. This issue could be more prominent in this RIC dataset as it contained more replicates than IFN-α WCP or phosphoproteome data I generated.

Interestingly, the moderated T-statistics itself could be helpful in bypassing the imputation issues. In T-statistics, the possibility that an observation deviates from the null hypothesis is estimated based on its fold change, standard deviation, and sample size. Moderated T-statistics follows the same principle, with an improvement that replaces standard deviation term by standard deviation moderated with Bayesian estimation method[249] (Formula.5.3.1).

Formula.5.3.1 modulated T-statistics

$$t = \frac{\beta}{\tilde{s}/\sqrt{n}},$$

where $\beta = coefficient\ (logFC)$,
$\tilde{s} = moderated\ standard\ deviation$,
$n = number\ of\ replicate$

The moderated standard deviation is calculated as a weighted average of global standard deviation, which derives from mean standard deviation of all proteins in the entire dataset, and the local standard deviation, which is the standard deviation of the tested protein. The weighing factor is the degree of freedom, which correspond to number of valid values (Formula.5.3.2).

Formula.5.3.2 modulated standard deviation

$$\tilde{s} = \frac{{s_g}^2 d_g + {s_0}^2 d_0}{d_g + d_0},$$

where $s_g = global\ standard\ deviation$,
$s_0 = local\ standard\ deviation$,
$d_g = global\ degree\ of\ freedom$,
$d_0 = local\ degree\ of\ freedom$

This algorithm “borrows” information from the entire dataset to help improve the measurement accuracy for each of the tested protein – when the tested protein has lower number of valid values, the algorithm gives higher weight to global standard deviation and *vice versa*. This feature endows

moderated T-test a higher tolerance to missing values and yield more robust results compared with conventional Student's T-test. The only scenario where moderated T-statistics fail to work is when a protein is only present in one condition and has no observation in the other condition. Based on these features, I hypothesized that missing values in data could be handled by moderated T-test algorithm without imputing all missing values in data. In this novel analytic method, the missing values in data are "partially" imputed based on their expression patterns – proteins that are only detected in one condition ("ON-OFF" pattern) are imputed on the "OFF" side, and other types of missing values are kept intact prior to moderated T-statistics (Fig.5.3.1 A).

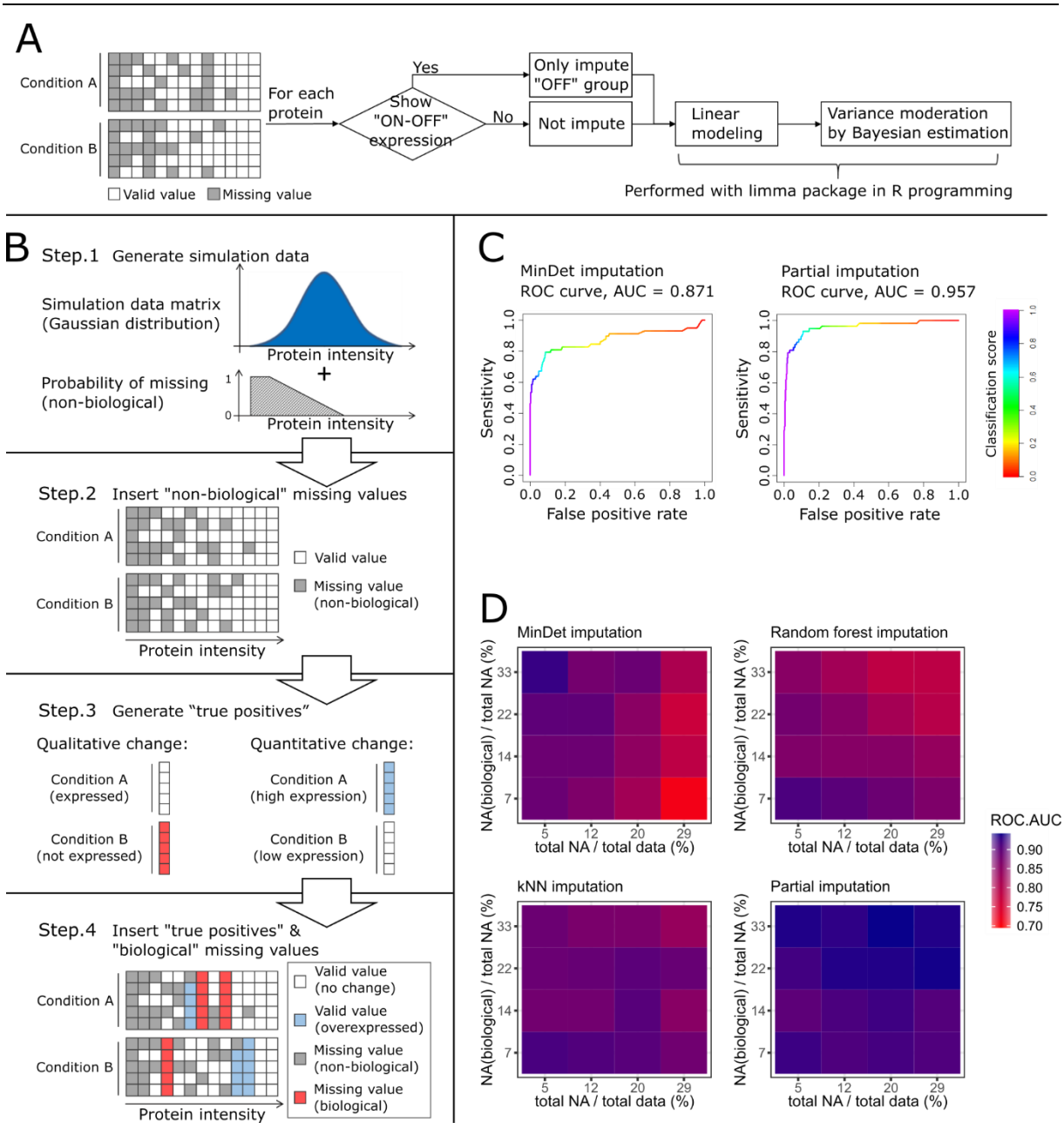


Figure.5.3.1 Novel data processing method (partial imputation) improved results of moderated T-test in simulation data. (**A**), workflow of novel imputation method (partial imputation) that check whether a protein is only detected in one condition ("ON-OFF" pattern), and only impute missing values for these "ON-OFF" proteins. (**B**), schematic of strategy used for generating simulation data. Missing values from two different sources were simulated. (**C**), Receiver Operating Characteristic (ROC) curves of test results using different imputation methods. The simulation data used contained 5 replicates, 30% missing values, and 20% of missing values being driven by biological factor. (**D**), heatmaps of Area Under Curve (AUC) values for ROC plots generated from different simulation data and imputation methods. Percentage of missing values (NA) in each simulation data were labelled in x- and y-axises.

Various imputation methods have been developed for proteomic data analysis[417,418]. Despite built on different rationales and models, the principle of these methods is to find more accurate approximates for all missing values in the data. It was unknown whether the performance of "partial" imputation can surpass the existing methods in proteomic data analysis. Comparison between different imputation methods were first conducted using simulation data, which had known "true positive" hits and missing values that had similar distribution as real-world proteomic data. Briefly, simulation data without missing values was first generated based on Gaussian distribution, and missing values were inserted in two steps: missing values caused by non-biological effects, such as low abundance and stochastic fluctuation, were randomly inserted following the assumption that the probability of missing value negatively correlate with protein intensity (Fig.5.3.1 B, Step.1); missing values resulted from "biological" effect were simulated by randomly select a group of proteins and replace all numbers in one condition by missing values, resembling proteins that are only present in one biological condition (Fig.5.3.1 B, Step.3). Compared with real-world proteomic data generated in our lab, the simulated data had similar distributions of intensity, missing value, and standard deviation.

The first test I conducted was to compare performances of MinDet and "partial" imputation. After filtering out proteins that had too few valid values to be quantified, the simulation data was subjected to these imputation methods before moderated T-test. The simulated data contained 5 replicates, 30% missing values, and 20% of missing values derived from "biological effect". Receiver Operating Characteristic (ROC) plot was drawn based on test results to evaluate the ability to correctly identify true positives in data (Fig.5.3.1 C). For an ideal classifier that correctly identify all true positive hits and make no false positive, its ROC curve would reach point (0, 1) and maintain a plateau; by contrast, a random classifier such as coin flipping would appear as a diagonal line in ROC

curve. Area under curve (AUC) is an important metric of the ROC curve, AUC of an ideal classifier would equal 1 and AUC of coin flipping would be 0.5. Importantly, partial-imputation pipeline showed higher performance than MinDet pipeline in both ROC curve shapes and AUC values.

Next, I conducted a more systematic comparison of imputation methods using multiple simulation datasets that had more diverse setups. Several characteristic imputation methods were chosen to represent different classes of missing value modelling algorithm: 1), , MinDet method was selected to represent left-censored imputation methods that model missing values as signals close to detection limit[214]; 2), k-Nearest Neighbour (kNN) method takes local similarity to impute missing value of target protein, and was selected to represent similarity-based imputation methods[214]; 3), decision tree-based imputation methods employs non-parametric machine learning in iterative manner to estimate missing values, algorithm built based on random forest (RF) was selected to represent this class of method[419]. Multiple simulation datasets were generated in 5 replicates with different setups: the percentage of missing values ranged from 5% to 30%, and the percentage of "ON-OFF" expression-driven missing values ranged from 7% to 33%. After processed with either MinDet, kNN, RF, or partial imputation methods, AUC values of ROC curves from each test result were calculated and summarised into heatmaps (Fig.5.3.1 D). For MinDet imputation, as expected, it showed better performance when a higher proportion of missing values were driven by "ON-OFF" expressions and had the lowest AUC values when the data had a large proportion of missing values that were originated from other factors. The kNN and RF imputation yielded similar AUC values, ranging from 0.8 to 0.9. Remarkably, the partial imputation method showed high and consistent AUC values even in dataset where 30% of points were missing. This observation implied that partial imputation could be a high-performance and robust method for T-statistics-based quantitative proteomics analysis.

After validated the feasibility of partial imputation by simulation data, it was further compared with existing imputation methods using real-world proteomic data from publications. These datasets covered various types of experiments and multiple research topics, including affinity purification-based pulldown of ubiquitin binding proteins[420], immunoprecipitation-based pulldown of ISG interactors[166], PTMscan-based pulldown of cellular ubiquitome[421], UV crosslinking-based RBPome profiling[151,170], and antisense RNA-based pulldown of viral RNAs[422]. Performance of the imputation methods were determined by moderated T-test results, the number of significant hits under several cut-offs are summarised in Table 5.3. Compared with other methods, partial imputation provided the best test result for various types of proteomic data, with only one exception in analysing PTMscan data. These test results confirmed the high versatility of partial imputation method in processing various types of proteomic data.

Table.5.3 Comparison between analysis methods using published proteomic data.

Publication	Experiment	MinDet imputation		kNN imputation		RF imputation		Partial imputation	
		FDR<0.1	FDR<0.01	FDR<0.1	FDR<0.01	FDR<0.1	FDR<0.01	FDR<0.1	FDR<0.01
X.Zhang, et.al., 2018	Pulldown with Di-Ub coated beads in HeLa lysate	1397	696	1228	309	1047	396	1602	1073
X.Zhang, et.al., 2018	Pulldown with poly-Ub coated beads in HeLa lysate	660	294	131	59	198	58	917	582
P.Hubel, et.al., 2019	Pulldown with anti-GFP beads in IFI44L-GFP stable line	132	21	122	18	128	28	175	93
P.Hubel, et.al., 2019	GFP beads-based IFI44L pulldown in mock and RVFV-infected cell	555	237	587	196	706	287	843	459
F.Hansen, et.al., 2021	PTMscan-based Ub-site detection in mock and TNF treated cell	732	205	1166	266	2232	878	1629	635
M.Garcia, et.al., 2019	RNA interactome capture in mock and SINV-infected cell	223	39	231	22	226	45	251	60
W.Kamel, et.al., 2021	RNA interactome capture in mock and SARS-CoV-2 infected cell	306	58	303	56	324	59	339	110
S.Lee, et.al., 2021	Antisense RNA-based SARS-CoV-2 viral RBP pulldown	271	215	244	138	276	223	324	287

5.4 Comparative analysis with IFN-α proteome showed RNA-binding activity changes independent of virus infection.

When applied to my IFN-α RIC data analysis, partial imputation introduced a much smaller scale of artificial intervention compared with MinDet imputation – close to 400 data points were imputed in MinDet, whereas ~170 points were imputed by partial imputation pipeline (Fig.5.3.2 A). The test result of data processed with partial imputation highlighted 83 proteins as significant with an FDR<0.1, and 47 proteins with FDR<0.01 (Fig.5.3.2 B). Among these hits, 51 were upregulated, and 32 downregulated (Fig.5.3.2 C). In comparison, the test result from MinDet-processed data also had 83 significant hits by a cut-off line of FDR<0.1, but only 30 proteins had FDR<0.01 (Fig.5.3.2 B). The results according to partial imputation were taken forward and proteins with significant change were defined as IFN-α regulated RBPs.

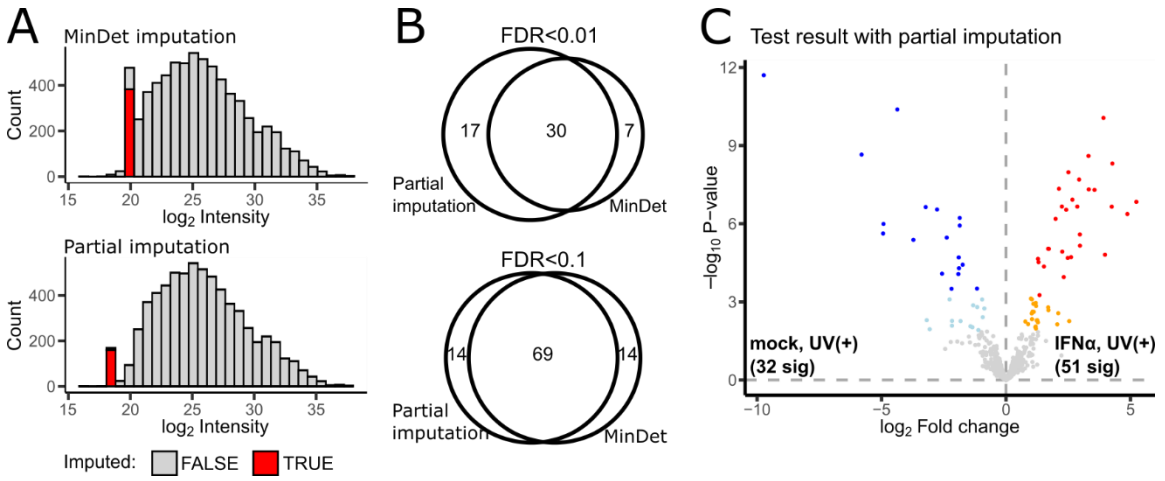


Figure.5.3.2 Novel imputation method improved test results of IFN-α RIC data. (**A**), histograms of protein intensities after processing pipelines, imputed values were highlighted red. (**B**), venn diagrams comparing number of proteins with FDR<0.1 or FDR<0.01 under different data processing method. (**C**), volcano plot of RBPs in IFN-α stimulated cells versus mock cells, processed with partial imputation method. Cut-off for significant hits were colour labelled (red and blue for FDR<0.01, orange and light-blue for FDR<0.1).

To investigate whether the observed changes in poly(A) RBPome after IFN-α treatment were resulted from changes in protein abundance, IFN-α RIC data was compared against the IFN-α WCP data generated in the previous chapter (Chapter.3, section.3). These two datasets were highly comparable because they were obtained from the same cell line under the same IFN-α treatment condition. In the comparative analysis, the majority of RBPs identified in IFN-α RIC were also identified in the WCP. Interestingly, fold changes in RIC were largely independent of changes in WCP (Fig.5.4.1 A). This indicated that the IFN-α induced changes in poly(A) RNA-bound proteome was likely a result of regulation in RNA-binding activity, instead of protein abundance. Applying a FDR<0.1, only 3 proteins were upregulated in both protein abundance and RNA-bound proteome: helicase HELZ2, ISG15-conjugator HERC5, and RNA deaminase ADAR. Several proteins also showed positive fold changes in both RIC and WCP, including ubiquitin-conjugator TRIM25, protein kinase PKR (EIF2AK2), and mono-ADP-ribosyltransferase PARP12. These proteins have well-described roles in innate immune response against virus[113,280,393,394,423,424]. By contrast, no enrichment in innate immunity-related GO term was found among proteins that showed significant changes in RIC and no changes in WCP, indicating that the lack of current understanding in IFN-α mediated RBP activity.

Series of RIC-based studies have shown that virus infection could cause massive changes in cellular RBPome[151,170]. However, the changes induced by the virus are difficult to separate from

those driven by host antiviral response. To distinguish the effect of these two components, the IFN-α RIC results in this study were compared against published cellular poly(A) RBPomes of cells infected with either Sindbis virus (SINV)[151] or SARS-CoV-2 (SARS-2)[170]. These two studies share the same RBP-ome identification method as in my IFN-α RIC study. The SINV cRIC data were also obtained in HEK293 cells, while the SARS2 cRIC data were obtained in lung cancer cell line (Calu-3). After applying the same cut-off (FDR<0.1) to all three datasets, IFN-α regulated RBPs showed limited overlap with RBPs regulated upon infection. Only 17 RBPs overlapped with SARS-CoV-2 infected cells, and 12 RBPs overlapped with SINV-infected cells (Fig.5.4.1 B). Surprisingly, when examining the fold change of these overlapping RBPs, it was found that IFN-α induced RBP activity changes were negatively correlated with virus-induced changes in general (Fig.5.4.1 C). For example, RNA deaminase ADAR, and rRNA processing proteins DDX27, RRP8, NOP14 showed increased RNA binding activities after IFN-α stimulation and decreased activities after virus infection; on the contrary, P-body associated protein PALT1, splicing factors HNRNPD, DDX42, and translation initiation factor EIF4A2, EIF5B showed the opposite pattern.

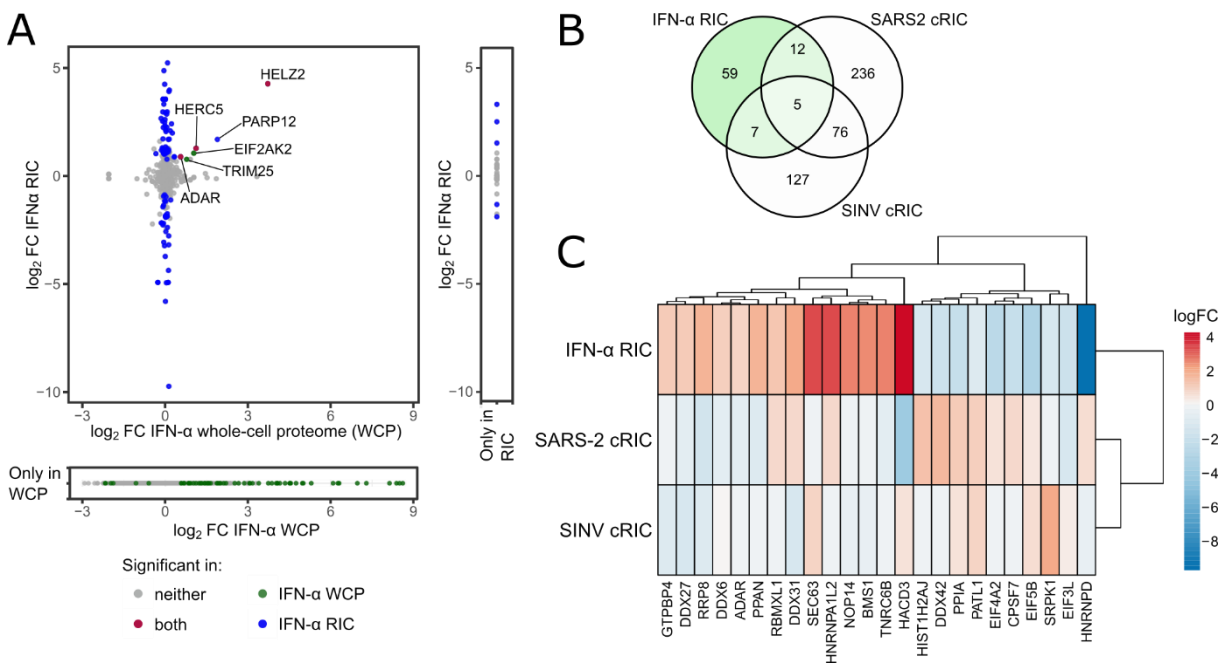


Figure.5.4.1 IFN-α regulated RBP activities independent of protein abundance and virus infection. (**A**), Proteins detected in both RIC and WCP datasets were plotted in the scatter plot in the middle, proteins that were only detected in WCP or RIC were plotted on bottom or right. Significant hits in each data were colour labelled (cut-off line FDR<0.1). (**B**), venn diagram comparing changes in the cellular poly(A) RBPome in after IFN-α stimulation, SINV infection (HEK293 cells), and SARS-CoV-2 infection (Calu-3 cells). The method for RBPome identification were the same across these three

compared datasets. (**C**), heatmap comparing magnitude of change in RBP activities regulated by IFN-α and virus infection, "logFC" represented log_2 fold change of each condition (IFN-α stimulation / SARS-2 infection / SINV infection) versus mock.

Recently, Hubel et al. used a large-scale mass spectrometry to reveal the compilation of cellular proteins that interact with known ISGs, which greatly expanded the realm of proteins that may participate in the interferon response[166]. IFN-α regulated RBPs could also be connected with known ISGs through protein-protein interactions. To explore this aspect, Fisher's exact test was applied to examine whether ISG interactors were enriched in IFN-α regulated RBPs. Interestingly, when using whole cell proteome as background for enrichment analysis, IFN-α regulated RBPs were found to be highly enriched in ISG interactors (p = 0.0001); and when considering only known RBPs as background, ISG interactors were also overrepresented among IFN-α regulated RBPs with a considerably low p-value (p = 0.0768, Fig.5.4.2 A). Network analysis showed that antiviral proteins

---

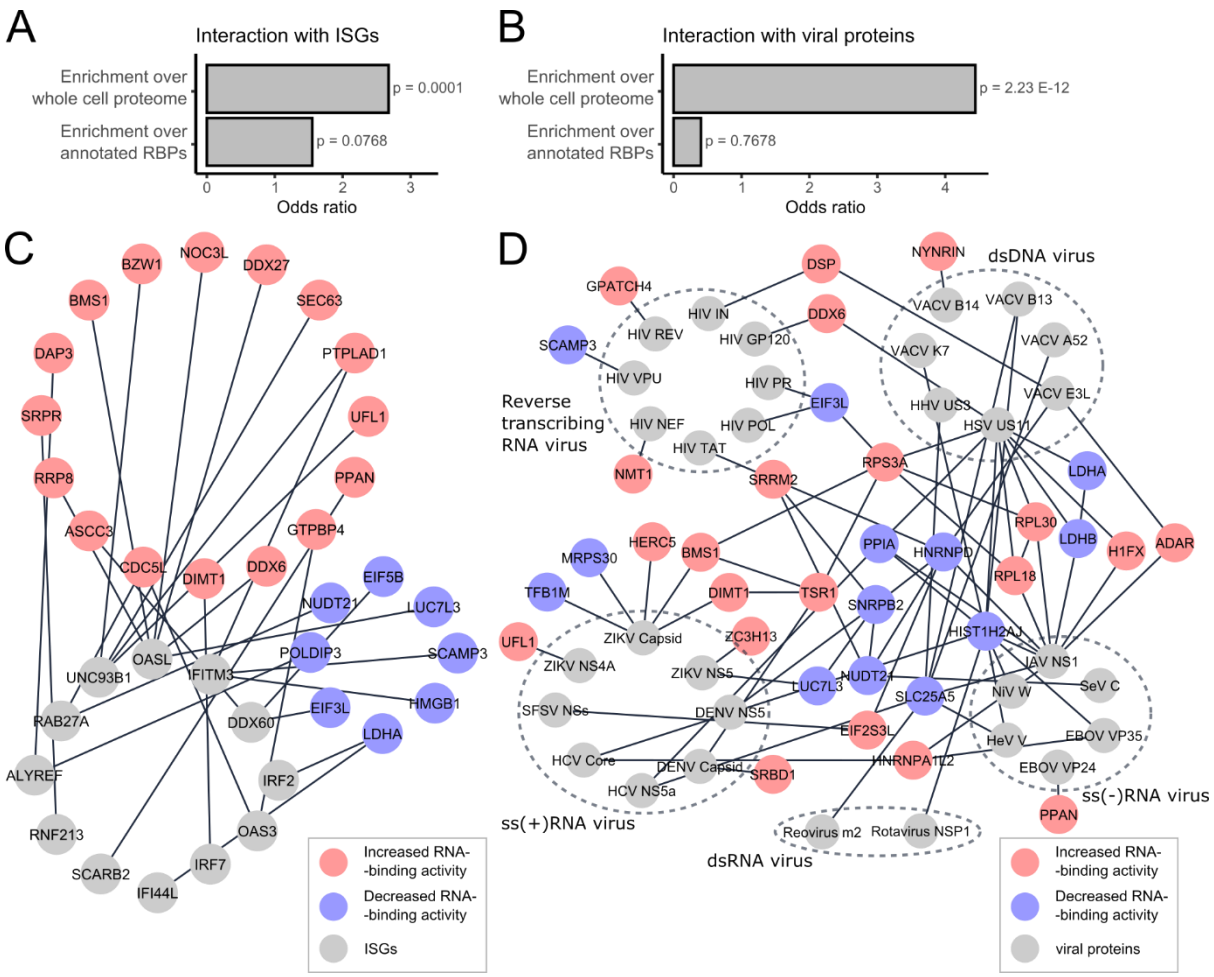


Figure.5.4.2 Interaction networks between IFN-α regulated RBPs and ISGs or viral proteins. (**A**),

enrichment of ISG interactors among IFN-α regulated RBPs over whole cell proteome or known RBPs. (**B**), enrichment of viral protein interactors among IFN-α regulated RBPs over whole cell proteome or known RBPs. (**C**), network of interactions between ISGs and IFN-α regulated RBPs. (**D**), network of interactions between viral proteins and IFN-α regulated RBPs.

OASL, IFITM3, and Toll-like receptor regulator UNC93B1 were connected with RBPs that showed increased RNA-binding activity after IFN-α stimulation (Fig.5.4.2 C). These interactions could localise ISGs to RNA that bound to IFN-α regulated RBPs. Meanwhile, several ISGs such as DDX60 and OAS3 could interact with RBPs that are either up-regulated or down-regulated by IFN-α, indicating a complexed response to IFN-α.

IFN-α regulated RBPs could also connected to viral proteins via protein-protein interactions. Similar analysis was also performed to assess whether IFN-α regulated RBPs were enriched in viral protein interactors. The repertoire of viral protein interactors was consolidated based on literature reviews as described in previous chapter (Chapter.2, section.4). As a result, IFN-α regulated RBPs were found to be significantly enriched in viral protein interactors over whole cell proteome, but no enrichment was found when using known RBPs as background for enrichment analysis (Fig.5.4.2 B). This indicated that viral protein interactors could be enriched in RBPs in general, and not in IFN-α regulated RBPs in particular. Formation of these ribonucleoprotein complexes could still be regulated by IFN-α, and may play important roles in host antiviral responses. Several IFN-α regulated RBPs were interestingly connected with multiple viral proteins, such as AU-rich element binding protein HNRNPD and ADP/ATP translocase SLC25A5 (Fig.5.4.2 D). Studies have shown that HNRNPD has multifaceted roles during infection, it could act as a virus facilitating factor for HIV-1[425], Epstein-Barr virus[426], West Nile virus[427], Hepatitis C virus[428], and it could inhibit Human Rhinovirus[429], and Nipah virus[430]. The IFN-α mediated regulation of HNRNPD could be part of the host defence response, and similar principle could also apply to other IFN-α regulated RBPs.

5.5 Mechanism of IFN-α mediated regulations in RNA-binding activities could involve domain-specific regulation and phosphorylation.

RNA binding domains (RBDs) mediate the interaction of proteins with RNA[431]. They can be classified as classical RBD if having well characterised interactions with RNA, and non-classical if not[323]. To test which domain types were prone to IFN-α regulation, domain annotations for proteins

characterised as RBPs and proteins with significant changes in my IFN-α RIC data were retrieved from Pfam database[432]. Only 27% and 22% of RBPs stimulated or inhibited by IFN-α were found to contain these classical RBDs, respectively (Fig.5.5.1 A). In comparison, the proportion of proteins harbouring classical RBDs was 35% for the total RBPome reported in IFN-α RIC, and 53% for RBPs identified in HeLa[433]. The most frequently occurred RBDs include RNA recognition motif (RRM), DEAD box helicase (DEAD) domain, zinc finger domain (zf-CCCH), and hnRNPK homology (KH) domain (Fig.5.5.1 B, right). Interestingly, the frequencies of these RBDs differ between IFN-α regulated RBPs and total RBPs. KH domain is the second most frequent classical RBDs in total RBPome, and it was not present among IFN-α regulated RBPs (Fig.5.5.1 B, left). Differences in domain occurrence were also observed for non-classical RBPs: Helicase_C domain and Ldh domains (Ldh_1_N, Ldh_1_C) showed higher occurrence among IFN-α regulated RBPs than the whole RBPome, while various other domains were less frequent among IFN-α regulated RBPs (Fig.5.5.1 C).

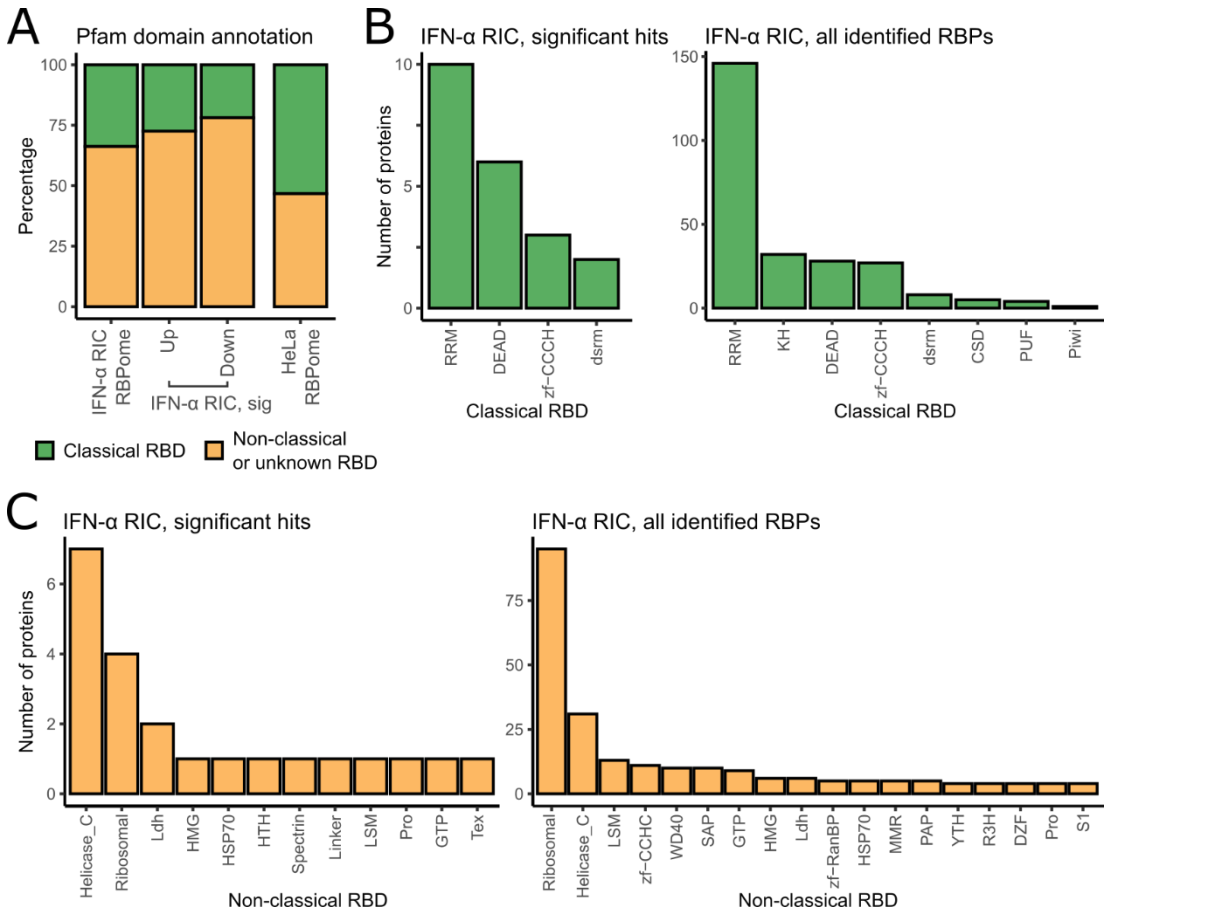


Figure.5.5.1 Domain analysis revealed changes in domain occurrence in RBPs regulated by IFN-α. (**A**), proportion of RBPs that harbour classical RNA binding domain (RBD) according to Pfam annotation. HeLa RBPome was acquired from published RIC study using HeLa cells[433]. (**B**), occurrence

of classical RBDs among “significantly” regulated RBPs and all identified RBPs in IFN-α RIC experiment. (**C**), occurrence of non-classical RBDs. Domains of less than 3 counts were not drawn in the barplot of all identified RBPs.

A series of studies have shown that for a wide range of RBPs, regulation of their RNA-binding activities is achieved through post-translational modifications (PTMs)[433-435]. Phosphorylation is one of the most common types of PTM that regulates protein activity. To explore whether IFN-α regulated phosphosites were related to RBP activity regulation, the IFN-α phosphoproteome analysis result was cross-referenced to this RIC result. Phosphosites that showed high Z-score and sites that showed “ON/OFF” changes in semi-quantitative analysis were included. As a result, several IFN-α regulated phosphosites were found to be present in both RBPs that showed up-regulated and down-regulated RNA-binding activities after IFN-α treatment (Table.5.5).

Table.5.5 Phosphoproteomic analysis results of IFN-α regulated sites in IFN-α regulated RBPs.

	IFNa.10min_vs_mock phosphoproteome				IFNa.4h_vs_mock phosphoproteome				RNA binding activity
	log2FC	P.Value	Z.score	identified	log2FC	P.Value	Z.score	identified	
SRRM2_S398	0.485	0.606	1.525	both sides	3.336	0.014	10.019	both sides	Up-regulated in IFNα-RIC
SRRM2_S420	NA	NA	NA	only in IFNα(+)	NA	NA	NA	only in IFNα(+)	Up-regulated in IFNα-RIC
SRRM2_S424	1.960	0.074	3.603	both sides	1.860	0.086	3.318	both sides	Up-regulated in IFNα-RIC
SRRM2_T384	0.332	0.580	1.045	both sides	1.981	0.019	5.949	both sides	Up-regulated in IFNα-RIC
CKAP5_S830	-0.880	0.040	-2.768	both sides	-1.650	0.004	-4.956	both sides	Up-regulated in IFNα-RIC
DDX42_S744	NA	NA	NA	only in IFNα(+)	NA	NA	NA	both sides	Up-regulated in IFNα-RIC
DDX42_S83	NA	NA	NA	only in IFNα(+)	NA	NA	NA	both sides	Up-regulated in IFNα-RIC
DDX42_S84	NA	NA	NA	only in IFNα(+)	NA	NA	NA	both sides	Up-regulated in IFNα-RIC
HNRNPD_T193	NA	NA	NA	only in IFNα(+)	NA	NA	NA	only in IFNα(+)	Down-regulated in IFNα-RIC
POLDIP3_T138	NA	NA	NA	only in IFNα(+)	NA	NA	NA	only in IFNα(+)	Down-regulated in IFNα-RIC
PATL1_T178	-0.267	0.540	-0.839	both sides	1.348	0.015	4.049	both sides	Down-regulated in IFNα-RIC

Among RBPs that showed down-regulated RNA-binding activity after IFN-α treatment, the IFN-α regulated phosphosites included Thr-193 in HNRNPD, Thr-138 in POLDIP3, and Thr-178 in PATL1. Interestingly, all of these phosphosites were found to be up-regulated by IFN-α (Fig.5.5.2 A, left). Notably, the IFN-α regulated phosphothreonine identified in phosphoproteome analysis locates in the second RRM domain of HNRNPD. According to the known 3D structure of HNRNPD, this threonine residue establishes hydrogen bond with nearby serine residue and hydrophobic interaction with isoleucine residue[436] (Fig.5.5.2 B). These interactions could be disrupted by IFN-α mediated phosphorylation, and potentially cause conformation change of this RRM domain that may

reduce RNA binding activity. Among RBPs that had up-regulated RNA-binding activity by IFN-α, spliceosome component SRRM2, dsRNA helicase DDX42, and microtube-associated protein CKAP5 were found to harbour IFN-α regulated phosphosites. These phosphosites were located at positions with no evidence for mediating RNA binding (Fig.5.5.2 A, right). Although not placed at known RNA-binding interface, it cannot be ruled out that these phosphosites could also affect RNA binding.

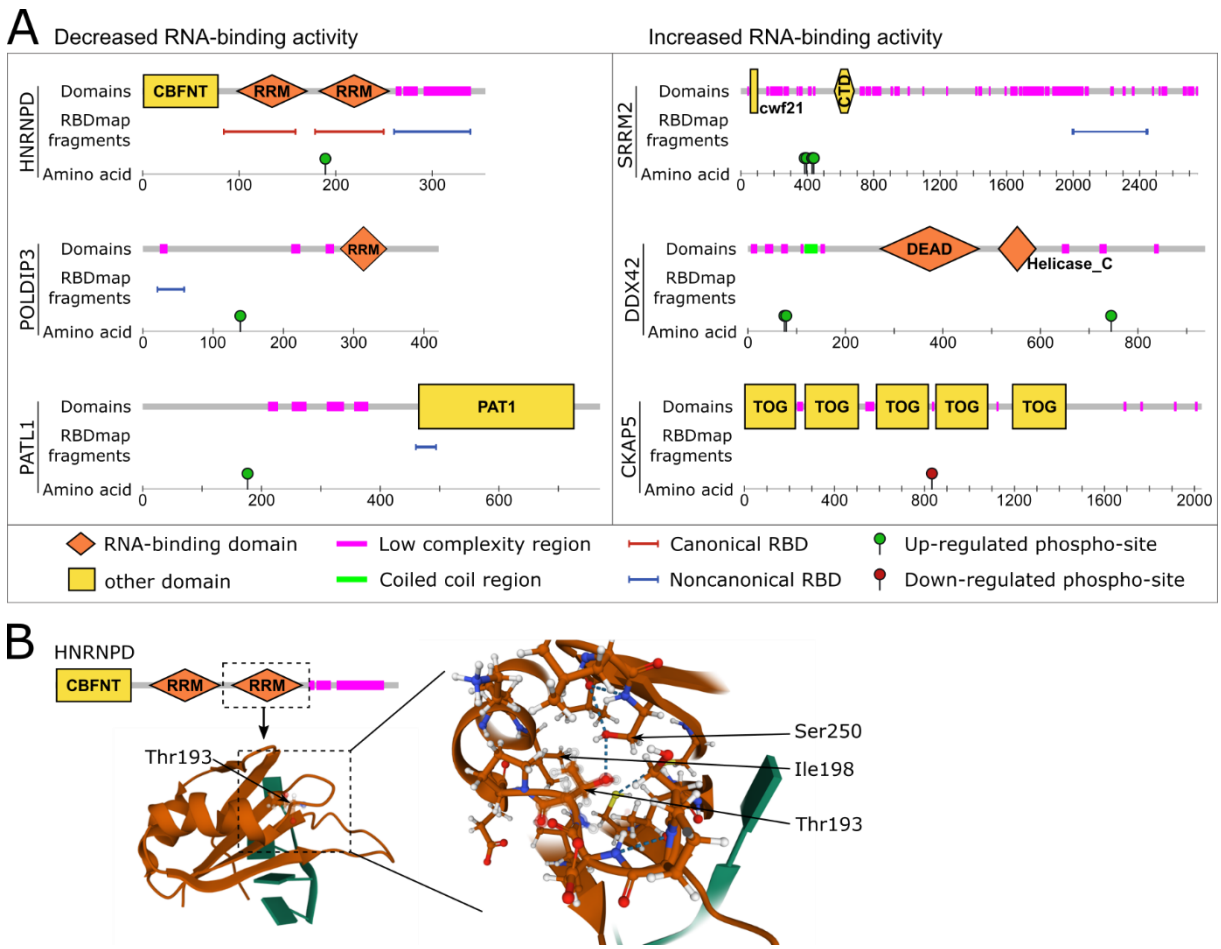


Figure.5.5.2 IFN-α dependent phosphorylation could be involved in regulation of RNA-binding activity. (**A**), schematics of IFN-α regulated RBPs that contains IFN-α regulated phosphosites. "Domains" showed annotations from SMART and Pfam; "RBDmap fragment" showed RNA-interacting peptide fragments identified in RIC-based RDB profiling study[433]. (**B**), structure of second RRM of HNRNPD (PDB-ID: 1WTB)[436,437].

Discussion

In-depth profiling of protein phosphorylation landscape in cells has been a challenging task. In this chapter, I profiled phosphoproteome of IFN-α stimulated HEK293 cells using off-line fractionation and employing two LC-MS/MS systems. Once combined, the resulting dataset reached a detection depth of 23,377 unique phosphosites across 3942 phosphoproteins. The two MS instruments use different methods for peptide sequencing, one uses collision induced disassociation (CID) and the other is higher energy collision disassociation (HCD). Jedrychowski et.al has evaluated performance of these two methods, and concluded that both methods possess unique advantages in phosphopeptide identification[438]. In-line with that report, the two MS instruments in this study generated highly complementary results when analysing the same sample set. SILAC labelling enabled quantification of phosphopeptide with improved accuracy because the sample combining occurred at an early step, allowing technical errors to be isolated and removed. However, measures that account for non-biological effects (such as replicate) could distort the test result if these effects are not distributed evenly between tested groups[439]. For phosphoproteome data, unbalanced distribution is a severe issue because of the high proportion of missing value. The solution in this study was to apply a stringent filter which only accept phosphosites that have valid values across all replicates in at least one condition. This filter ensures balanced distribution of all effects in at least one condition, however with a cost of discarding ~15,000 sites in quantification. The high occurrence of missing value has been a major challenge in phosphoproteome analysis since the emergence of this field. This is driven by: 1, the overwhelmingly rich diversity of protein phosphorylation compared with limited LC separation power and MSMS scan number; 2, phosphosite identification often rely on single detection event; and 3, the low abundance of phosphopeptide compared with total sequence-able peptide in sample. Methods that reduce the number of missing value could hugely expand the quantifiable phosphoproteome. To achieve this, one tempting direction in future phosphoproteome work is to apply tandem mass-tag (TMT) labelling. This method labels different samples with tags that are only distinguishable during peptide sequencing, mixes labelled samples prior to MS analysis, and perform quantification using ratios between tags[440]. Because TMT method quantify phosphopeptides during sequencing step, all samples in a sample mixture can share one sequencing result, therefore reduce missing values caused by missed identification.

Phosphoproteomic analysis in this work revealed dozens of IFN-α regulated phosphorylation sites on known ISGs. The detected increase in phosphorylation of STAT1 Ser727 has also been reported in previous studies, and is shown to augment interferon signalling[97,367]. However, changes in phosphorylation on other ISGs that participate in innate immunity, such as PML[441] and

ANKRD17[415], have not been functionally described. A decrease in phosphorylation on IFNAR2 Ser384 was observed in this data. This site locates in the region responsible for IFNAR2-STAT2 binding during interferon signalling[442], and the role of its phosphorylation is poorly understood. On the contrary, a number of functionally characterised phosphosites on ISGs were not observed in the present study. These unobserved sites are mostly phosphotyrosines on interferon signalling factors, including IFNAR1, IFNAR2, STAT family proteins, and MAVS[6,443]. The incomplete characterisation of phosphosites that participate in interferon signalling is expected because only a part of the total phosphoproteome could be observed. A complete portrayal of cellular phosphorylation sites remain hindered by high complexity of phosphorylation sites in phosphoproteome, high stochastic noise in phosphopeptide identification, and low stoichiometry (particularly for phosphotyrosines)[444,445].

The majority of observed IFN-α regulated phosphosites are not located on proteins known to participate in other signalling pathways. This is not unexpected because over 2/3 of proteins in the human proteome are estimated to be phosphorylated[445,446], and the majority of phosphorylation modifications are not necessarily involving signalling proteins[274]. Surprisingly, these IFN-α regulated phosphosites were found to be enriched in interactors of viral proteins and viral RNAs. These virus-interacting host proteins have highly diverse functions during virus infection, either acting as antiviral effectors, or supporting viral replication. As the most common type of PTM that regulates protein activity, the biological role of phosphorylation can be very diverse depending on function of the modified protein. The overrepresentation of viral protein and RNA interactors indicates that IFN-α dependent phosphorylation could play a major regulatory role in host-virus battlefront, whereas the exact role of each phosphosite may require case-by-case studies.

RBPs are also known to be active players in host-virus engagement. Recent studies have revealed large-scale changes in RNA binding landscape in virus infected cells[151,170]. In this chapter, the characterisation of the cellular poly(A) RBPome in IFN-α stimulated cells revealed that in the absence of virus, cells can still actively modulate its RBPome during innate immune response. Most of these changes in RNA binding were shown to be independent of changes in protein abundance, indicating the existence of a previously unknown layer of host antiviral response at RBPome level. Together with the studies focusing on RBPs that interact with viral RNA or total RNA during infection[151,170,447], my study on the RBPome responses induced by IFN-α provides a complementary perspective in understanding the RNA-protein interactions established during virus infection. Many of the known antiviral proteins were RBPs, such as viral RNA sensor RIG-I and MDA5, however most of these proteins are absent among IFN-α regulated RBPs in this dataset. This can be explained by the absence of viral RNA as the experiment was performed in uninfected condition. Therefore, RBPs that bind to viral RNA with high specificity would be omitted in this RBPome dataset. Only 4 known

ISGs have increased RNA binding in response to IFN-α without the presence pf viral RNA, include ADAR, HELZ2, HERC5, and PARP12. ADAR is an A-to-I RNA editing enzyme that is known to bind to viral RNA genome where it introduces detrimental mutations[393]. In this dataset, its increased RNA binding in the absence of viral RNA could be unspecific binding. However, we cannot rule out that interaction between ADAR and host RNA have uncharacterised role in IFN response. Accurate mapping of ADAR binding sites among host RNA is required to fully understand why ADAR showed increased RNA binding by IFN-α without a viral RNA target. HERC5 is known to act as ISG15 conjugator during IFN response[113], however, little is known regarding the function of its RNA binding activity. HELZ2 and PARP12 are participated in multiple processes besides IFN response. HELZ2 is a transcription coactivator that regulates lipid metabolism and adipocyte differentiation[448], and PARP12 can regulate stress response by localising to stress granule[449]. Their increased RNA binding by IFN-α could associate these processes with IFN-α response, but further studies are required to answer whether their RNA binding activities are essential for these non-immunity processes, and whether regulations in these processes have effect on IFN response.

For the rest of the IFN-α regulated RBPs identified in this study, whether their changes in RNA binding play a role in the innate immune response or in infection require further investigation. Several of the IFN-α regulated RBPs have been linked to virus replication. For example, AU-rich element interactor HNRNPD is reported to facilitate replication of HIV-1, EBV, WNV and HCV[425-428], and splicing factor SRRM2 was shown to regulate splicing of HIV-1 viral RNA[148]. My data shows that HNRNPD and SRRM2 also harbour phosphosites that could be regulated by IFN-α, and interestingly, the phosphothreonine site in HNRNPD is located in its RNA recognition motif (RRM), which mediates interaction with RNA. These observations imply a potential crosstalk between IFN-α regulation in phosphorylation and RNA binding. Another interesting discovery for IFN-α regulated RBPs is their enrichment in known viral protein interactors and ISG interactors. This lead to an interesting idea that IFN-α regulated RBPs could impact viral replication or host immune response through protein-protein interactions with viral protein or ISGs. Further biochemical and in vivo studies are required to investigate whether these IFN-α regulated RBPs have a role in IFN response, whether their roles in IFN response are related to their RNA binding activities, how they participate in regulation of virus infection or host immune response, and how IFN-α regulates their RNA binding activities.

Acknowledgements

I want to thank all members of Castello lab for the supports in cell cultures and RIC experiments, thank all members of Shabaz lab for the supports in the proteomics experiments and data analysis. Thanks Alfredo Castello-Palomares & Shabaz Mohammed for conceptualising the research in this chapter and their supervisions. Thanks Sabrina Liberatori for her pilot works in optimising the phosphopeptide enrichment. Thanks Louise Iseline for her pilot works in RIC analysis of IFN-α stimulated cells, and her collaboration works in generating cells for RIC analysis in this chapter. Thanks Philip Charles for discussions regarding methods for phosphoproteomic data analysis, his collaboration works in development of the partial imputation method, and discussions regarding *on silico* validations of the partial imputation method using simulation and real-world proteomics data.

Methods and Materials

Methods

Cell culture

Culture and maintenance of HEK293 cells were the same as Chapter 2. Culture with IFN-α stimulation was the same as Chapter 3.

Western blotting

Blocking of the nitrocellulose membranes was performed using The same as Chapter 2.

Cell culture with SILAC

The same as Chapter 3.

Sample preparation for phosphoproteomic analysis

HEK293 cells with SILAC for phosphoproteomic analysis were the same batch of cells generated in Chapter 3. The mock, IFN-α 10 min, and IFN-α 4 h samples were defrosted. An aliquot of each lysate was reserved for protein concentration and isotope incorporation measurements through the same processes as Chapter 3. Equal weights of SILAC lysates were combined according to the label-swap strategy shown in Table.5.1, each combined sample contains 1.6 mg worth of proteins. Reducing reagent Tris(2-carboxyethyl)phosphine hydrochloride (TCEP) and alkylating reagent 2-Chloroacetamide (CAA) were added to combined lysates to 10 mM and 50 mM, respectively, then incubated in dark for 30 min. Alkylated samples were purified with methanol-chloroform precipitation by: supplying each sample with 12.8 ml methanol, 3.2 ml chloroform, and 9.6 ml water sequentially, vortex after each reagent addition (final volume 28.8 ml), centrifuging in 3,000 g for 10 min to remove upper layer, supplying 9.6 ml methanol and vortex followed by centrifuge in 3,000 g for 10 min, removing all liquid and air dry, finally resuspending samples with 1.6 ml 8 Murea buffer. Purified alkylated proteins were diluted to 6 M urea with 25 mM AmBic for LysC protease (Wako) digestion with protein/protease ratio at 60:1 (w/w) at 37^{o}C for 4 h with 250 rpm shaking, then diluted to 1 M urea with 25 mM AmBic for trypsin protease (Promega) digestion with protein/protease ratio at 50:1 (w/w) at 37^{o}C for overnight with shake. Digestion was quenched with 0.5% (v/v) formic acid (FA).

Tryptic peptides were fractionated with a HPLC prior to LC-MS/MS analysis following the same process as Chapter 3. The resulting 20 fractions were dried with SpeedVac and resuspended in 125 µl loading buffer (5% DMSO, 5% FA in water), then combined into 10 fractions at an even interval (1st with 11th, 2nd with 12th, etc.). These 10 fractions were subjected to phosphopeptide enrichment by Ti-IMAC following procedures in published study[450]. Specifically, 350 µl Ti-IMAC beads slurry (Resyn Bioscience) was washed with 800 µl ethanol trice, 700 µl of 1% ammonium hydroxide once, and 1ml Ti-IMAC loading buffer (80% ACN, 1 M Glycolic acid, 5% trifluoroacetic acid (TFA) in water) twice, each with 1350 rpm shake, and equilibrated with 350 µl Ti-IMAC loading buffer. A 14 µl aliquot of equilibrated beads slurry was added to each fraction and incubated in 24^{o}C with 1350 rpm shake for 25 min to allow binding of phosphopeptides with Ti-IMAC beads. Supernatants were removed and beads were washed with 400 µl Ti-IMAC loading buffer once, 400 µl Ti-IMAC washing buffer 1 (80% ACN, 1% TFA in water) once, and 400 µl Ti-IMAC washing buffer 2 (10% ACN, 0.2% TFA) twice. Phosphopeptides were eluted from beads by incubation with 80 µl ammonium hydroxide at 24^{o}C with 1350 rpm shake for 25 min three times, three eluates were combined and acidified with 60 µl of 10% formic acid. Enriched phosphopeptides were dried with SpeedVac and stored in -80^{o}C for mass spectrometry analysis.

Mass spectrometry for phosphoproteomics

Each fraction of dried tryptic peptides were reconstituted with 50 µl of loading buffer (5% DMSO, 5% FA in water). Mass spectrometry analysis was performed with two independent LC-MS/MS systems. In one system, 20 µl of reconstituted peptides mixture were analysed on a nanoUHPLC (Thermo) connected to a Q Exactive mass spectrometer (Thermo Fischer Scientific) through an EASY-Spray nano-electrospray ion source (Thermo Fischer Scientific). The peptides were trapped on a C18 PepMap100 pre-column (300 µm i.d. x 5 mm, 100 Å, Thermo Fisher) using solvent A (0.1% formic acid in water). The peptides were separated on an EASY-spray Acclaim PepMap analytical column (75 µm i.d. × 500 mm, RSLC C18, 2 µm, 100 Å) using a linear gradient (length: 120 min, 8 % to 28 % solvent B (0.1% formic acid, 5 % DMSO in acetonitrile), flow rate: 200 nL/min). The separated peptides were electrosprayed directly into the mass spectrometer operating in a data-dependent mode. Full scan MS spectra were acquired in the Orbitrap (scan range 350-1500 m/z, resolution 70000, AGC target 3e6, maximum injection time 50 ms). After the MS scans, the 20 most intense peaks were selected for HCD fragmentation at 30 % of normalised collision energy. HCD spectra were also acquired in the Orbitrap (resolution 17500, AGC target 5e4, maximum injection time 120 ms).

In the other LC-MS/MS system, 20 µl of reconstituted peptides mixture were analysed on an EASY-nLC 1000 System (Thermo) connected to an Orbitrap Elite Hybrid mass spectrometer (Thermo Fischer Scientific) through an EASY-Spray nano-electrospray ion source (Thermo Fischer Scientific). The peptides were trapped on a C18 PepMap100 pre-column (300 µm i.d. x 5 mm, 100 Å, Thermo Fisher) using solvent A (0.1% formic acid in water). The peptides were separated on an EASY-spray Acclaim PepMap analytical column (75 µm i.d. × 500 mm, RSLC C18, 2 µm, 100 Å) using a linear gradient (length: 60 min, 8 % to 28 % solvent B (0.1% formic acid, 5 % DMSO in acetonitrile), flow rate: 200 nL/min). The separated peptides were electrosprayed directly into the mass spectrometer operating in a data-dependent mode. Full scan MS spectra were acquired in the Orbitrap (scan range 350-1500 m/z, resolution 70000, AGC target 1e6, maximum injection time 100 ms). After the MS scans, the 20 most intense peaks were selected for CID fragmentation at 35% normalised collision energy. CID spectra were acquired in the LTQ mass analyser (resolution 7500, AGC target 5e3, maximum injection time 100 ms).

Protein identification and quantification were performed using Andromeda search engine implemented in MaxQuant (1.6.3.4)[211]. Peptides were searched Human Uniprot database (version, January 2016). Phospho(STY) was set as a variable modification. False discovery rate (FDR) was set at

1% for both peptide and protein identification. Data files from the same sample were assigned as near-by fractions, and match between run was turned on to match peak information from near-by fractions. Other options were set as defaults.

Quantitative analysis for phosphoproteomics

Moderated T-statistics and T-statistics with linear mixed model was performed with R programming. Phospho(STY) files from MaxQuant search results were imported into R, proteins that have at least 3 valid values in at least one condition were subjected to log transformation, normalisation with vsn method using package *vsn*[213], and imputation of missing values by background-level signal using package *ImputeLCMD*[214]. Moderated T-test was performed using package *limma*[249]. Two layers of batch effects (replicate and isotope) were considered by incorporating isotope type as co-variable in design matrix, and inputting replicate number as block variable. Correlation consensus for this block variable was calculated following user guide. P-values were corrected for multiple testing by controlling the FDR with the Benjamini-Hochberg method.

Z-statistics of phosphosite quantifications were performed by first calculating the standard deviation (sd) of fold changes of all phosphosites, then divide fold change of each phosphosite by this sd value to obtain Z-scores, finally convert Z-scores to p-values based on normal distribution. P-values were corrected for multiple testing by controlling the FDR with the Benjamini-Hochberg method.

Statistical tests were performed separately to phosphoproteome data acquired from Q-Exactive and Orbitrap Elite mass spectrometers. Test results were combined in two methods. The first method retains all information from both datasets. The other method only retains one test result for duplicated measurements following the procedure visualised in Fig.5.1.4 B. Criteria for selecting phosphosites by detection quality were: 1), if the difference between localisation probabilities in two datasets was larger than 0.05, select the result with higher localisation probability, otherwise proceed to next criterion; 2), if the overall score of phosphosite identifications in two datasets were less than 90, select the result with higher identification score, otherwise proceed to next criterion; 3), calculate the quantile of intensity of the phosphosite against all intensities of all phosphosites in the dataset, select the result with higher quantile.

Bioinformatic analysis for phosphoproteomics

Heatmaps of IFN-α regulated phosphosites were plotted with R programming package *pheatmap* using hierarchical clustering[451]. Enrichment in viral protein and viral RNA interactors were calculated by first obtaining proteins that contain IFN-α regulated phosphosites (by Z-statistics) from quantitative analysis results, then match to a compilation of viral protein/RNA interactors. Viral protein/RNA interactors were compiled following the same process as Chapter 2. Enrichments were calculated using Fisher's exact test.

Kinase-substrate enrichment analysis (KSEA) was performed following the procedure described in published studies[452,453]. Specifically, a compilation of kinase-substrate correlation annotations were first downloaded from series of online databases (PhosphoSitePlus[404], PhosphoELM[405], SIGNOR[406], DEPOD[407], iPTMnet[408], NetworKIN and Netphorest[409]). Phosphorylation positions of phosphosites in IFN-α phosphoproteome data were converted to positions at the first isoform before matching to the kinase-substrate correlation compilation. For each kinase, its activity change was calculated by the average of fold changes of phosphosites that it can regulated. Enrichment of kinases was performed with Z-statistics as described in the phosphoproteomics quantitative analysis section.

RNA interactome capture

Comparative RNA interactome capture (RIC) was performed following the protocol of published study[10]. Culture of IFN-α treated HEK293 cells and uv crosslinking of cell lysates were performed by collaborator Louise Iseline. Cells were cultured on ten 10-cm dishes and treated with IFN-α or mock treated for 20h following the same procedure as described in Chapter 4. Cells were washed with PBS and irradiated with 150 mJ/cm^2 UV light at 254 nm, then harvested with cell scrapers in 1 ml RIC lysis buffer (20 mM Tris-HCl pH 7.5, 500 mM LiCl, 0.5% LiDS w/v, 1 mM EDTA, 0.1% IGEPAL (NP-40) and 5 mM DTT), and froze in -80^{o}C.

After defrost, lysates were homogenised by repetitively passing the lysate at high speed through a 2 mL syringe with a 27G needle (BD) until lysates were fully homogenised. 150 µl worth of oligo(dT) magnetic beads (NEB) were added to each lysate after three washes with 1 ml RIC lysis buffer and reconstitute in RIC lysis buffer. Lysates were incubated with beads at 4^{o}C for 2 h with gentle rogations to allow binding of poly(A) RNA with oligo(dT) beads. Supernatants were removed after binding, followed by one wash with 1.8 ml RIC lysis buffer, two washes with 1.8 ml RIC washing buffer 1 (20 mM Tris-HCl pH 7.5, 500 mM LiCl, 0.1% LiDS wt/vol, 1 mM EDTA, 0.1% IGEPAL and 5 mM DTT), and two washes with RIC washing buffer 2 (20 mM Tris-HCl pH 7.5, 500 mM LiCl, 1 mM EDTA, 0.01% IGEPAL and 5 mM DTT), each wash was performed on ice for 5 min inverting tubes ten times

every minute. Two additional washes was performed at room temperature with 1.8 ml washing buffer 3 (20 mM Tris-HCl pH 7.5, 200 mM LiCl, 1 mM EDTA and 5 mM DTT). Poly(A) RNA bound to beads were eluted with 100 µl elution buffer at 55^{o}C for 10 min with agitation (200 rpm), and supernatants were collected after magnetic rack separation. RNA-bound proteins were released by digestion with 1µl of 1:100 diluted Rnase A and Rnase T (Sigma).

Silver staining

RBPs obtained from RIC were analysed with silver staining using SilverQuest silver staining kit (Invitrogen) following the producer manual. Samples were centrifuged and boiled at 95oC for 5 min before loaded on house-made house-made 12% polyacrylamide gels for SDS-PAGE analysis. Gels were fixed with Fixing solution (40% ethanol, 10% acetic acid in water) at 4^{o}C overnight, and washed with 30% ethanol for 10 min with gentle shake. Fixed gels were treated with 60 ml Sensitising solution (30% ethanol, 10% sensitizer provided by the kit in water) for 10 min, washed with 30% ethanol for 10 min, and washed with water for 10 min. Gels were stained by incubation with 60 ml Staining buffer (1% stainer in water) for 15 min, followed by 1 min wash with water, and 4~8 min of incubation with Developing buffer (10 ml developer, 1 drop of developer enhancer in 100 ml water), and quenched with 10 ml stopper when protein patterns were clearly visible.

Proteomics sample preparation

RBPs obtained from RIC were cleared and digested with single-pot solid-phase-enhanced sample preparation (SP3) protocol as described in published study[151]. Samples were reduced and alkylated by adding TCEP to 10 mM and CAA to 50 mM before 30 min of in-dark incubation at room temperature. 40 µl worth of carboxylated magnetic beads slurry (Sigma) were equilibrated with with 160 µl water, washed with 200 µl water for three times, then reconstituted in 200 µl water. A 10µl aliquot was added to each sample, then supplied with ACN to final concentration of 77% ACN. The mixtures were incubated at room temperature for 30 min, and placed on magnetic rack to remove supernatant. Beads were washed with 10 washing cycles, each comprised two washes with 1 ml 70% ethanol and one wash with 1 ml 100% ACN. After washes, beads were reconstituted in 50 µl of 25 mM AmBic solution, proteins were digested on beads with trypsin protease (Promega) at protein/protease ratio of 50:1 (w/w) at 37^{o}C for overnight with shake. Eluates were collected after separation by magnetic rack. Beads were eluted again by reconstitution with 25 µl of 2% DMSO in

water, and supernatants were combined with the previous eluates (final volume 75 µl) followed by acidify with 1 µl of neat formic acid.

Mass spectrometry

SP3 cleared tryptic peptides were supplied with 5 µl of loading buffer (5% DMSO, 5% FA in water), 40 µl of peptides mixture was analysed on a nanoUHPLC (Thermo) connected to a Q Exactive mass spectrometer (Thermo Fischer Scientific) through an EASY-Spray nano-electrospray ion source (Thermo Fischer Scientific). The peptides were trapped on a C18 PepMap100 pre-column (300 µm i.d. x 5 mm, 100 Å, Thermo Fisher) using solvent A (0.1% formic acid in water). The peptides were separated on an EASY-spray Acclaim PepMap analytical column (75 µm i.d. × 500 mm, RSLC C18, 2 µm, 100 Å) using a linear gradient (length: 120 min, 8 % to 28 % solvent B (0.1% formic acid, 5 % DMSO in acetonitrile), flow rate: 200 nL/min). The separated peptides were electrosprayed directly into the mass spectrometer operating in a data-dependent mode. Full scan MS spectra were acquired in the Orbitrap (scan range 350-1500 m/z, resolution 70000, AGC target 3e6, maximum injection time 100 ms). After the MS scans, the 20 most intense peaks were selected for HCD fragmentation at 30 % of normalised collision energy. HCD spectra were also acquired in the Orbitrap (resolution 17500, AGC target 5e4, maximum injection time 120 ms).

Protein identification and quantification were performed using Andromeda search engine implemented in MaxQuant (1.6.3.4)[211]. Peptides were searched Human Uniprot database (version, January 2016). False discovery rate (FDR) was set at 1% for both peptide and protein identification. Data files from the same treatment condition (IFN-α or mock) were assigned as near-by fractions, and match between run was turned on to match peak information from near-by fractions. Other options were set as defaults.

Proteomics quantitative analysis

Moderated T-statistics was performed with R programming. ProteinGroup files from MaxQuant search results were imported into R, proteins that have at least 3 valid values in at least one condition were subjected to log transformation, normalisation, and imputation of missing values. Normalisation was performed with vsn method using package *vsn*[213]. Imputation by conventional method used package *ImputeLCMD* to impute missing values by background-level signal[214]. Imputation by "partial imputation" method developed in this chapter is described in Section 5.3, only proteins that have no detection in all 5 replicates in one condition were imputed by

background-level noise with package *ImputeLCMD*, remaining missing values were left unchanged. Processed data was subjected to package *limma* for moderated T-test, with block variable set to replicate number to account for the batch effect derived from processing dates (5 replicates were processed with RIC at different dates), and correlation consensus calculated following the *limma* user guide. P-values were corrected for multiple testing by controlling the FDR with the Benjamini-Hochberg method.

Evaluation of partial imputation result was performed with ROC curves using package PRROC[454]. Simulation data containing 5 replicates were generated as described in Section 5.3, then applied filter of at least 3 valid values in at least one condition, log transformation, normalisation by vsn method. Processed simulation data was imputed with partial imputation or other imputation methods prior to moderated T-test with package *limma*, FDR were calculated based on the Benjamini-Hochberg method. True positives and negatives in simulation data were input as binary variable, and FDR were input as continuous variable to draw ROC curve and calculate area under curve (AUC).

Bioinformatic analysis for RIC

Matching of RNA binding-related GO annotation was performed by first obtaining the compilation of GO annotation was obtained from R package *GO.db*[291]. Then GO terms containing the keyword "RNA binding" were considered RNA binding-related GO terms. Genes annotated with these GO terms were extracted with *biomaRt* package, then matched to the proteomic analysis result[292].

Enrichment of viral protein interactor was the same as described in Chapter 2. Enrichment of ISG interactors were performed by matching IFN-α regulated RBPs with a compilation of ISG interactors characterised by Hubel et al. using GFP-pulldown coupled with proteomic analysis[166]. Degree of enrichment was calculated with Fisher's exact test.

For RNA binding domain analysis, the compilations of Pfam domains that are annotated as classical or non-classical RNA binding domains were obtained from published study[170]. Pfam domains in all proteins identified in IFN-α RIC dataset were extracted after mapping proteins to the Human genome annotation database obtained from R package *org.Hs.eg.db*[455].

Materials

Chemicals and reagents

REAGENT or RESOURCE	SOURCE	IDENTIFIER
100 bp ladder	NEB	B7025
2-Chloroacetamide	Sigma Aldrich	C0267
6x purple loading dye	NEB	B7024S
Acetonitrile	Sigma Aldrich	#34851
Acrylamide/Bis-acrylamide 30%	Thistle Scientific	#20-2100-05
AEBSF hydrochloride	BioChemica	A14210100
Ammonium Bicarbonate	Sigma Aldrich	#9830
Bensonase	EMD Millipore	#70746
Bovine Serum Albumin (BSA) powder	Fisher Scientific	BP9700100
Carboxylate-modified magnetic beads	Merk	GE45152105050250
Chloroform	Sigma Aldrich	#67-66-3
Dimethyl sulfoxide, DMSO	Thermo Scientific	T6508
DL-dithiothreitol (DTT)	Sigma Aldrich	D1532
DMEM	Gibco	#11995-065
Ethanol	VWR chemicals	#20821.321
Ethylenediaminetetraacetic acid (EDTA)	VWR chemicals	100935V
FBS	Sigma Aldrich	F9665
Formic acid	Thermo Fisher	A117-50
Glycolic acid	Sigma Aldrich	#124737
IGEPAL CA-630	Sigma Aldrich	I3021
Interferon alpha 2a	Pbl assay science	#11100-1
Lithium chloride	Sigma Aldrich	#62476
Lithium-dodecyl sulfate	Sigma Aldrich	L9781
LysC protease	Wako	#129-02541
MagReSyn Ti-IMAC magnetic beads	Resyn Bioscience	MR-TIM002
Methanol	Sigma Aldrich	#34860
Oligo(dT)25 magnetic beads	NEB	S1419
PBS 1x, pH 7.4	Gibco	#10010023
penicillin/streptomycin	Sigma Aldrich	P4458
Phosphatase inhibitor	Sigma Aldrich	P0044
Pierce 660nm Protein Assay Reagent	Thermo Scientific	#22660
Pierce Bovine Serum Albumin Standard Pre-Diluted Set	Thermo Scientific	#23208
Protease inhibitor cocktail	Sigma Aldrich	P8340
Rnase A	Sigma Aldrich	R4642
RNAse T1	Sigma Aldrich	R1003
SilverQuest silver staining kit	Invitrogen	LC6070
Sodium Deoxycholate Detergent	Thermo Scientific	#89904
Sodium dodecyl sulfate	Sigma Aldrich	#151-21-3
Trifluoroacetic acid	Sigma Aldrich	T6508
Tris(2-carboxyethyl)phosphine hydrochloride	Thermo Scientific	#77720
Triton-X100	Sigma Aldrich	#9036-19-5
Trypsin EDTA	Sigma Aldrich	T3924
Trypsin, sequencing grade	Promega	V5111

Tween-20	Sigma Aldrich	P1379
Urea	Sigma Aldrich	U1250

Instruments

INSTRUMENTS	SUPPLIER/INSTITUTE
Countess II FL Automated Cell Counter	Thermo Fisher
Odyssey Fc imaging system	LI-COR Bioscience
Nanodrop 2000 Spectrophotometer	Thermo Fisher
Trans-Blot Turbo Transfer system	Bio-Rad
LTQ XL Linear Ion Trap Mass Spectrometer	Thermo Scientific
UltiMate 3000 RSLCnano System	Thermo Scientific
Q Exactive mass spectrometer	Thermo Scientific
Easy-nLC 1000 system	Thermo Scientific
Orbitrap Elite hybrid mass spectrometer	Thermo Scientific

Antibodies

The same as Chapter 2.

Cell lines

The same as Chapter 3.

Chapter 6

Integrated omics analysis highlighted putative virus regulators for the generation of stable expression cell lines and functional validation

Introduction

Technical advances in high-throughput methods have hugely expanded the total amount of data on various branches of biological studies. As researchers are able to access much more data than before, mining out useful information from massive datasets has become a major challenge. To address this challenge, series methods have been developed to integrate information from multiple datasets and use it to help highlight proteins of interest (POIs). A common strategy in pinpointing POIs is to analyse their places in cellular pathways. Databases of manually curated pathways, such as KEGG[456] and Reactome[457,458], integrate numerous individual interaction studies for assembly of pathways. These are valuable data sources to understand the known roles of a protein. However, they are less informative for proteins that are not thoroughly studied, because curated pathways are constructed based on known interactions. Integration of multiple high-throughput proteomic datasets can help expand pathway annotations to include less-studied proteins. In a study that integrate transcriptome, proteome, and protein interactome datasets, the mathematical principle of path optimisation was used to directly assemble pathways based on the integrated dataset[459]. In another study, researchers integrated multiple protein-protein interaction (PPI) datasets, and developed machine learning based method to train this integrated datasets using curated pathway data[460]. In both studies, pathways generated from integrated datasets can effectively reproduce curated pathways, and expand these pathways to include other processes and less-studied proteins.

Apart from pathway-centred analyses, studies that integrate multiple interactome datasets also try to highlight POIs through a network-centred view. In the cellular environment, proteins are interconnected to form a complex network. This network is composed of numerous basic elements known as motifs, such as feedback loop[461], and algorithms in detecting network motifs could help deduce the regulatory function of related POIs[462,463]. Another approach to identify POIs in network is to find highly-connected nodes, known as network's hubs[464]. Network hubs generally have high importance in cellular processes. They tend to result in higher lethality when mutated[465], and are more often targeted by pathogens to control the cellular environment[466]. The cellular PPI network contains multiple sub-networks (or communities) that are densely connected within each community, and have limited connections with other communities[467]. Many network hubs are highly connected within their local communities, these include key enzymes and master signalling proteins that have central roles in their respective biological processes[468,469]. Network hubs can also locate between multiple communities to bridge them. These hubs are also referred to as "bottlenecks", and tend to be essential for cell survival[470]. Computational methods in analysing network structures

can help identify network hubs, therefore be helpful in highlighting POIs of important functions for mechanistic studies[471-473].

Pathway- and network-centred analyses can provide valuable information for highlighting POIs when integrating multiple datasets. However, these methods strongly rely on evidence of direct PPI, therefore tend to omit proteins that exert their functions independent of direct PPI or by transient PPIs. These proteins can be picked up by analysis methods that focus on correlations between unknown proteins and proteins of known function. In a large-scale bioinformatics analysis that integrate over 5,200 human proteome raw files, researchers developed a co-regulation analysis that infers protein function based on other proteins that share similar expression patterns across different human cell lines[474]. In another bioinformatics study that integrated 112 phosphoproteome datasets from different species, a machine learning-based method was developed to highlight phosphosites of interest using functional score, which is estimated based on various features such as sequencing quality of phosphopeptide, protein structure, evolutionary conservation, etc.[475]. A common strategy shared by these analyses is that information from the integrated datasets were combined into a score to evaluate functional importance of the POI. This scoring-based approach bypass the need for PPI dataset, and can utilise information from different types of data for prediction of functional importance.

Overexpressing POI in the studied system is a regular route to investigate its function. The simplest method is to deliver plasmids that contain POI-coding sequence into cells via transient transfection. Transfected cells will have a surge in POI expression, which will be lost after several cell cycles because the transfected plasmid do not replicate in cultured cells. The transient transfection bear a lot of issues. After transfection, the copy number of POI sequence massively exceed that in genome, resulting in overexpression to a non-physiological level and artefact in phenotype. The transfected POI-encoding plasmids are cytoplasmic foreign DNA strands, which contain pathogen-associated molecular pattern and could trigger innate immune response[476]. This effect could mask the real biological function of the overexpressed protein. Some engineered plasmids can replicate in human cells, for example plasmid containing SV40 *ori* sequence can replicate in HEK293T cells[477], but their usages are restricted to limited cell types. Stable expression of the POI in a more physiological copy number can be achieved through direct integration of POI-coding sequence into the host genome. This is commonly achieved through viral vector-based stable transduction, transposon vector-based stable transfection, and DNA recombination-based stable transfection. Virus vector-based stable transduction uses part of the viral genome as carrier of POI sequence. Lentivirus-based system is a common option for stable transduction, it expresses viral integrase to transfer POI sequence into host genome[210]. Another frequent option is adenovirus-based system, which deliver

POI-containing viral vector into the nucleus as episomes, and integrated into the host genome during the DNA repair process[478]. The advantages of these stable transduction systems include high efficiency and compatibility to a wide range of host cell type. The restrictions of these systems are, for example the insertion site is unspecific so POI would not be expressed if the sequence is inserted into a silenced chromatin region, and would affect host gene expression if inserted inside another gene. Viral vector-based systems depend on expression of viral proteins, which could perturb the cellular environment therefore creating artefacts. The length of inserted sequence is also limited by the capacity of virus carrier in these stable transduction systems. Transposon-based stable transfection systems employ transposon-derived flanking sequence and transposase, allowing integration of large DNA fragments into genome[479]. Its main constrains is the limited efficiency and unspecific insertion. The copy number of sequence inserted by virus- and transposon-based system are variable, which risk causing artefact. Systems based on DNA recombination can insert DNA fragment in site-specific manner. Some methods flank POI sequence by homolog arms, which guide the insertion site in host genome, and uses host DNA recombinase to catalyse the integration event. Some use engineered site-specific DNA recombinase (i.e. CRE from bacteriophage, and FLP from yeast), and place POI near the enzyme recognition sequence for DNA recombination (loxP sequence for CRE recombinase, and FRT sequence for FLP)[480-482]. Both methods can achieve site-directed DNA recombination for generation of stable over-expression cell line (stable line), the methods using homolog arm allow almost arbitrary choice of insertion site but suffer lower efficiency, while methods using site-specific DNA recombinase are more efficient but require pre-existing recombinase-recognition sequence in host genome.

In previous chapters, I quantitatively compared the cellular proteome landscapes of cells that are highly permissive to virus (HEK293T), at steady-state (HEK293), and restrictive to viruses (post IFN-α stimulation). This is further supplemented with quantitative comparisons between restrictive and steady-state cellular environments at the transcriptome, phosphoproteome, and RBPome levels. Collectively, these datasets depict, with information from multiple levels of biological processes and great depth of detection, how a restrictive cellular environment differs from other cellular states. Utilising this information, we can highlight the factors that modulate cell permissiveness and subsequently characterise them as novel regulators of virus infection. It is impractical to achieve this by screening through all factors that differentiate cellular states with all types of viruses. Therefore it is crucial to select candidates that are more relevant to the host response to virus, and set up appropriate screening assay using representative viruses to test how candidate proteins affect virus infection. It's also worth noticing that many antiviral proteins are known to have redundancies where one protein's loss of function could be compensated by several

others[483,484]. Therefore, overexpression assays could be more effective in testing POIs' function than knock-down. In this chapter, I aimed to integrate omics datasets I generated in previous chapters to select a set of candidate genes that shape cell permissiveness to viruses. Because this multi-level omics dataset is rich in information, it allows investigations in various different perspectives. Here I intended to narrow down my focus on uncovering regulators of virus infection that are currently less understood. For these selected candidates, I also aimed to generate stable over-expression cell lines for follow-up studies.

Result

6.1 Selection of candidates for follow-up studies using evidence-based scoring

To facilitate the selection of candidates for follow-up studies, I first integrated datasets generated in previous chapters, and endowed them with scores that estimate how each line of evidence contributes to identification of regulator of infection that are less-understood at current stage. When integrating results from multiple datasets, a common challenge is that the selected datasets do not share the same dynamic range, therefore cannot be directly compared. The multi-level omics datasets I have were generated with different techniques using different types of samples, therefore have huge difference in their dynamic ranges. One simple method to overcome this issue is to count the number of times that a gene is determined "significant" in all datasets[474]. However, this method losses information regarding quantitative changes, and is sensitive to the choice of cut-off lines for statistical significance. By contrast, Z-statistics and T-statistics results are comparable across datasets, so using them for scoring can bypass the arbitrary selection of cut-off lines. Notably, the Z-statistics focuses solely on the average fold change and cannot reflect the reproducibility of measurement. Albeit having this drawback, Z-statistics results are still included in scoring because, in proteomic data of IFN-α stimulated cells, several known interferon-stimulated genes (ISGs) were found to have large fold change but relatively high p-value under T-statistics. These included transcription factor IRF9[485] and RNA helicase DHX58[486]. They were likely regarded not significant due to high measure error rather than genuine absence of response to IFN-α, and they can be picked up by Z-statistics.

Using both Z- and T-statistics results, I consolidated my multi-level omics data of IFN-α response as the first scoring factor in my candidate selection process (Fig.6.1 A & B-[1]). This includes both Z- and T-statistics of whole-cell proteome, and T-statistics of RNA-seq and RBPome of changes between IFN-α 20h-treated and mock-treated cells. IFN-α response has been intensively studied at the RNA level and most of the known ISGs found based on their RNA-level responses[334,335,476,487,488]. By contrast, protein-level responses were less well studied in previous publications, therefore I gave bonus to genes that show changes in the proteomics data based on both their Z-statistics and T-statistics, with an attempt to highlight both genes that have reproducible but small changes, and genes that have large changes but less stable measurement. For transcriptomic data, I gave penalties to genes that showed significant response in T-statistics for two reasons. Firstly, well-characterised ISGs were identified based on RNA-centric studies, therefore applying bonus to genes that show changes in transcriptome would put too much weight on well-studied ISG. Secondly, with a deep

proteomics data, I can potentially identify genes that show protein-level responses independent of changes at the RNA level. Applying penalty to RNA-level changes would help highlight these genes. Bonus was also given to genes that showed differential RNA binding after IFN-α treatment, because this level of IFN-α response has not been studied previously. Expression levels in HEK293/HEK293T proteome were also considered in the candidate selection. For genes that had significant differences in protein abundance between HEK293 and HEK293T cells, they received bonus scores because these genes could participate in shaping HEK293T's permissive cellular environment or in restricting virus infection in HEK293 cells.

Apart from my datasets, evidence from other sources were also included in the scoring process. They were incorporated using a count-based method, because some data were count-based by nature (i.e. number of matched publications in text-mining), and other quantitative data were difficult to compare directly since they were derived with different techniques and quality control criteria. Several factors were incorporated to show a gene's relation with known ISGs or innate immune response. These include GO annotation, clustering analyses results, and PubMed-wide text mining result (Fig.6.1 A & B-[2],[5]). A well-studied ISG would have clear GO annotation and surplus amount of literature describing its function in innate immunity. To avoid duplicating these works, penalties were applied to genes that contain GO annotations related to innate immunity, and genes that have more than 20 articles describing their functions in innate immunity (text-mining based on abstracts in Chapter.3, Section.3). To highlight potential virus regulating factors that we have limited knowledge about, bonuses were applied to genes that show similar expression patterns as well-characterised ISGs across mock, IFN-α 4h, and IFN-α 20h conditions (clustering result in Chapter.3, Section.3). Bonuses were also applied to genes that have little or no article that relate this gene with innate immune or antiviral response.

Several other factors were included in my scoring process to show whether a gene has known relations with viruses. Genes in my dataset were cross-referenced to our published cellular RBPome analyses in Sindbis virus (SINV) and SARS-CoV-2 infected cells[151,170], as well as unpublished works such as analyses of RBPome in HIV-1 and influenza infected cells, RBPome in mature SINV viral particle, and interactome of HIV-1 viral protein *Rev* (Fig.6.1 A & B-[3]). These genes were prioritised in the candidate selection by assigning high bonuses. I also included viral protein interactors from other publications in the scoring, the rationale is that genes interact with viral proteins are likely to affect the course of infection through protein-protein interactions (Fig.6.1 A & B-[4]). Bonuses were given to genes that interact with viral proteins, and the compilation of viral protein interactors is the same as described in Chapter 2. In this part of the scoring, although many known ISGs were also

interactors of viral proteins, ISGs were not specifically excluded because they had already received strong penalties in other scoring factors.

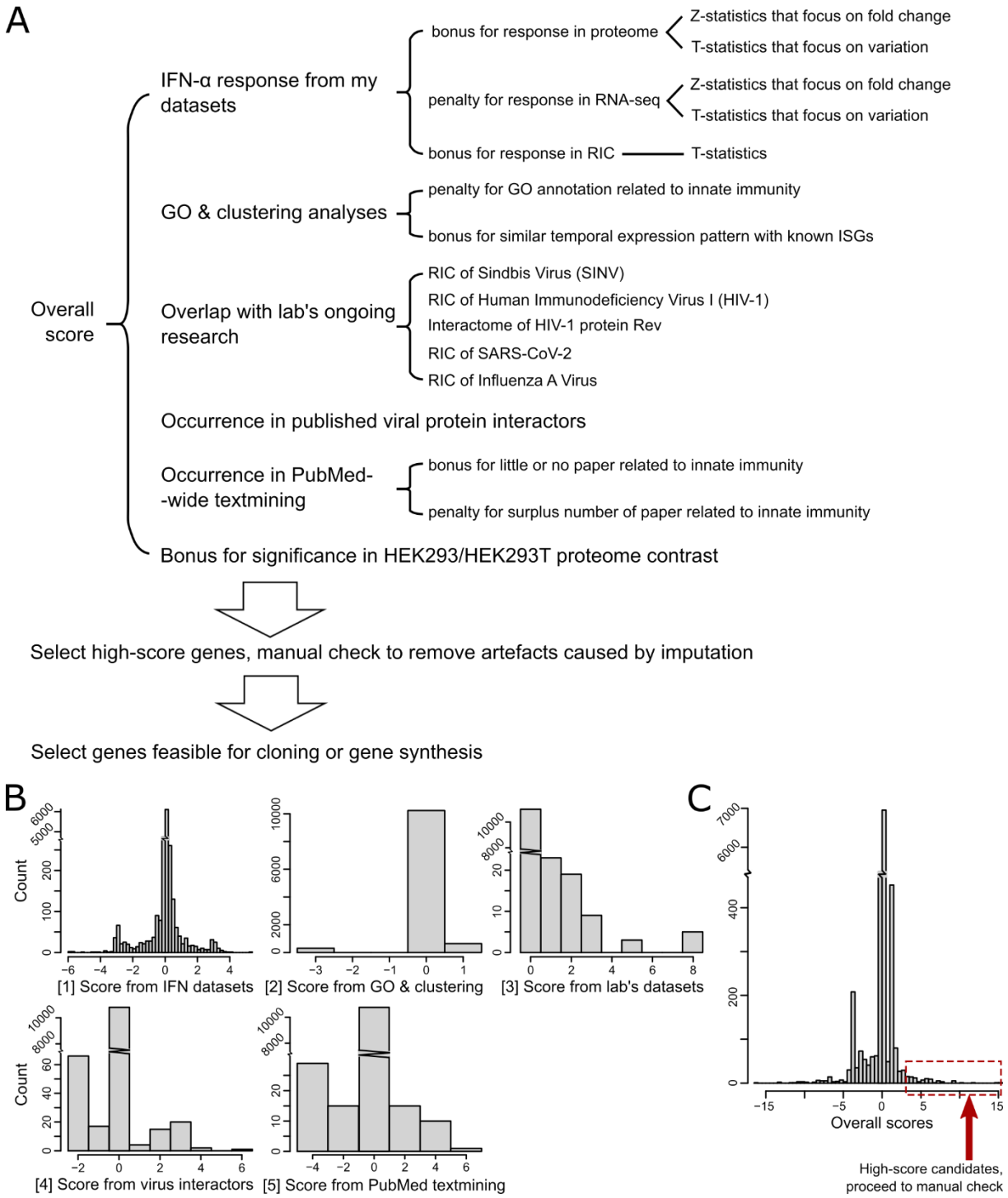


Figure.6.1 Selecting of candidates for follow-up studies using evidence-based scoring. (**A**), flow chart of evidence-based scoring process for selection of follow-up study candidates. Only genes that have FDR<0.2 in either Z-statistics or T-statistics in IFN-α proteome data are scored in other 5 major

scoring factors. (**B**), histograms showing scores of each aspect considered in the scoring procedure. (**C**), histogram of overall scores for all assessed genes. Red dashed box highlights high-score genes for manual check.

After determined each scoring factor, they were combined to yield an overall score, which has a distribution that most genes were rated close to zero, while a few genes were rated with high scores (Fig.6.1 C). High-score genes were selected and manually checked to remove artefacts, for example scores boosted by imputed values. Genes that are relatively difficult to clone or synthesize were also excluded in manual check, for example the highly-scored RNA-binding protein (RBP) HELZ2 because it has more than 2,600 amino acids in length. The remaining genes were considered candidates for follow-up studies, and in total 25 genes were constructed into overexpression vectors via either molecular cloning or gene synthesis.

6.2 Generation of inducible-expression stable lines for virology assays.

Many known antiviral proteins have non-negligible cytotoxic effects when overexpressed[248]. In innate immune response, cells have evolved mechanisms to timely suppress the antiviral ISGs after initial IFN stimulation, therefore keeping cytotoxic effect under control[489]. However constitutive overexpression of these antiviral proteins could be detrimental in stable transfection cell lines. To study virus-regulating activities of POIs in a close-to-physiological condition, I selected FlpIn-TREx system for stable transfection and drug-inducible overexpression of POIs (Fig.6.2 A). Briefly, POIs were inserted downstream a “Tet-ON” promoter, which is activated after exposed to Tetracycline or Doxycycline. A fluorescent protein eGFP was linked to POI with a Gly/Ser-rich linker, allowing detection of POI abundance through fluorescence signal. The overexpression vector contains a FRT sequence that can be recognised by DNA recombinase Flipase[480]. When co-transfected with Flipase-expressing vector, POIs were integrated through DNA recombination into another FRT sequence, which is engineered into the genome of HEK293 FlpIn-TREx cell line (Fig.6.2 A). This recombination event will shift the cell’s antibiotic resistance from Zeocin-resisting to Hygromycin-resisting, so cells with successful stable transfection can be isolated by antibiotic selection. These cell lines will be infected with different types of viruses to investigate whether corresponding POIs could affect virus infection. Virus strains prepared for this study include Chikugunya virus, Coxsackie virus, HIV-1, Semliki forest virus, Sindbis virus, and Zika virus. Another fluorescent protein mCherry is inserted into viral genomes either in-frame with viral protein or drove

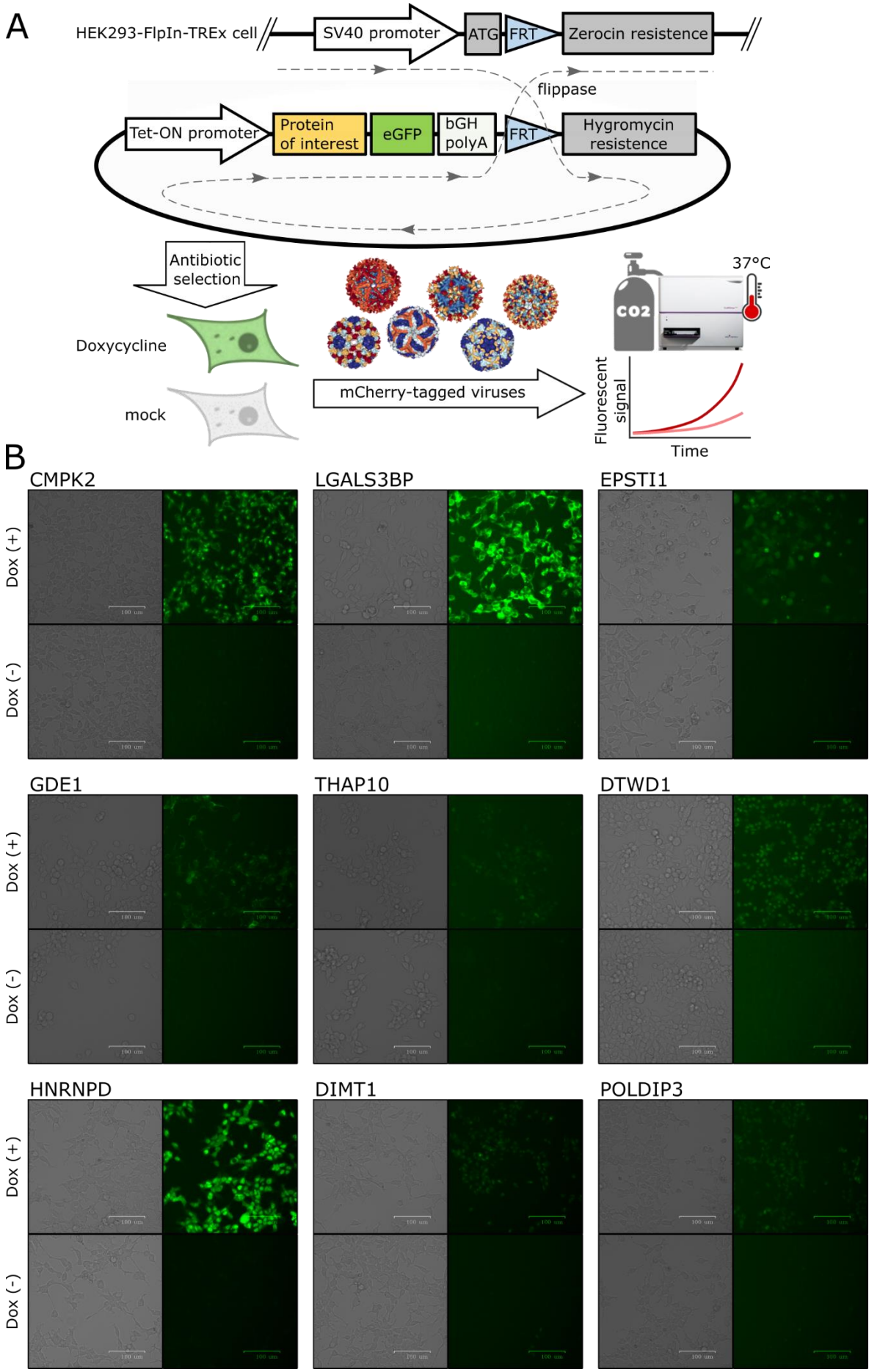

A
HEK293-FlpIn-TREx cell
SV40 promoter
ATG
FRT
Zerocin resistence
flippase
Tet-ON promoter
Protein of interest
eGFP
bGH polyA
FRT
Hygromycin resistence
Antibiotic selection
Doxycycline
mock
mCherry-tagged viruses
37°C
CO2
Fluorescent signal
Time
B
CMPK2
LGALS3BP
EPSTI1
GDE1
THAP10
DTWD1
HNRNPD
DIMT1
POLDIP3
Dox (+)
Dox (-)

Figure.6.2 Experiment plan and current stage for assessing virus-regulating activity of POIs using inducible expression stable lines. (**A**), schematic diagram of generation of inducible expression cell line and phenotype screen using viruses. (**B**), fluorescence imaging confirmed successfully generated cell lines for doxycycline-inducible expression. Only 9 out of 19 stable lines are presented in this panel. Dox (+) labels overnight doxycycline treatment.

---

by viral promoter. Therefore, the infected cells can be analysed by plate reader to record fluorescent signals in both eGFP and mCherry channels in real-time, and these signals can be used as proximate of POI expression and virus replication (Fig.6.2 A). If the analysed POI could affect virus replication, Doxycycline-treated infected cells would show a significantly altered mCherry profile compared with non-treated cells.

In my current work progress, I have successfully generated stable lines for 19 POIs. As confirmed in fluorescence imaging, these cell lines show little or no eGFP at steady-state, and have elevated eGFP signals after overnight Doxycycline treatment (Fig.6.2 B). Cellular localisation of eGFP in these cell lines followed the localisation of corresponding POI. For example, LGALS3BP is localised in both cytoplasm and extracellular spaces[490], and eGFP signal in the corresponding cell line was depleted in nucleus; DTWD1 and the eGFP signal in corresponding cell line are co-localised in the nucleus[491] (Fig.6.2 B). This evidence supported that eGFP is stably tagged on most POIs, therefore eGFP signal can infer the expression of POIs. Notably, fluorescent signals from different POI-expressing cell lines vary substantially. This could reflect the difference in expression rate, protein stability, or cytotoxicity between POIs. However, these cell lines are nevertheless competent for studying whether overexpression of each POI can affect cellular response to virus, because the fluorescent signal from mCherry-tagged virus in Doxycycline-treated cells will be compared with that from mock-treated cells, instead of other Doxycycline-treated cells. These stable lines are ready for virus infection and plate reader assays, and also for future mechanistic studies.

Discussion

In this chapter, I integrated the datasets I generated in previous chapters, and selected 25 POIs from over 9,800 proteins in the combined dataset, and successfully generated doxycycline-inducible stable expression cell lines for 19 POIs. This selection process uses my datasets as the main source of evidence, and combines it with other bioinformatic analysis results to estimate each protein's functional relevance to cellular response to virus, represented by functional score. POIs were selected from highly scored proteins after manual check. A similar approach has also been used in published studies that integrate multiple datasets to analyse functional importance of proteins or phosphosites[474,475]. These studies used sophisticated machine-learning based methods for scoring, whereas my POI selection is not. Machine learning algorithm was not introduced in this study because firstly, my datasets have major roles in POI selection while other evidences, such as text mining results, were used as reference. However, machine learning does not distinguish data of major role and other datasets. Secondly, machine learning uses datasets that have similar format and are acquired from different samples, for example whole cell proteome datasets from different cell lines. By contrast, the datasets integrated in my study were acquired with different methods and have different biological implications, for example transcriptome and RBPome datasets were focused on different levels of IFN-α response. The third reason is the lack of suitable training set for machine learning. POIs that I intend to find are putative virus regulators that are less-studied by current knowledge, but machine learning require training based on well-characterised proteins. Considering these factors, proteins in my integrated datasets were scored manually based on our interpretation of the evidences. Notably, the IFN-α phosphoproteome datasets generated in Chapter 5 were not directly integrated in scoring. This is because quantification of phosphosites bear high variations, and it is challenging to compare scores between singly and multiply phosphorylated proteins. Instead of being integrated directly, my phosphoproteome datasets were used as reference when manually checking POIs. They could provide further insights in future mechanistic studies, for example IFN-α regulated RNA binding of HNRNPD and POLDIP3 could be related to changes in their phosphorylation states (Chapter.5, Section.5). It is also worth noticing that POIs selected in this study are prioritised on those that show IFN-α responses at protein level. Proteins that had responses at RNA-binding level were also included in this study, but had a lower priority. This does not mean IFN-α regulated RNA-binding is unimportant; on the contrary, changes at RNA-binding level highlighted a new layer of IFN-α response that has not been analysed in previous works. These IFN-α regulated RBPs will be focused in future studies.

For each candidate selected in this study, I generated an inducible overexpression stable line using FRT/FLP recombination system. Overexpression of POIs could help bypass the functional redundancy between regulators of virus infection[483,484], however, it also have limitations. One POI is expressed in each stable line, therefore it could be difficult to identify synergy effects between virus-regulating factors. Many ISGs exert their functions in innate immune response through interaction with other ISGs. For example, transcription of many ISGs in IFN signalling require STAT1, STAT2 and IRF9 to form a transcription factor complex[205,485]; ubiquitin ligase DTX3L and ADP-ribosyltransferase PARP9 form a heterodimer complex to enhance ISG expressions[492]. Although these ISGs could have antiviral activities when solely overexpressed, their full antiviral capacities can only be revealed when innate immune response is activated. One possible solution to find this synergy effect is to knock-down or knock-out related proteins in IFN stimulated cells, but this would also require more accurate measurement of changes in virus replication rate, because loss of synergy effect in one protein could be partially compensated by antiviral activities of other proteins. Another limitation of this overexpression system is that the POI is artificially supplied in larger amount than physiological condition. Protein functions and interactions observed in this overexpression condition may not reflect their true functions under physiological condition. The site-specific insertion method can control that in every cell, only one copy of POI sequence is inserted in a transcriptionally active genome locus, therefore partially alleviate this issue. However, expression of the POI is controlled by Tet-ON promoter, which is engineered from virus-derived CMV promoter to respond to Doxycycline. This type of virus-derived promoters have higher transcription rate than most promoters in human genome, and could still cause artefact[233]. However, this level of artefact is acceptable because the aim of this study is not to thoroughly study each protein's physiological functions; instead, I aimed to use overexpression cell lines to screen through dozens of high-interest proteins and find proteins that could affect virus replication rate.

At current stage, I have not performed virus infection assay on cell lines I generated. The next experiment I planned to do is to infect these cell lines with a set of mCherry-tagged viruses, then monitor eGFP and mCherry signals in real time using plate reader (Fig.6.2 A). These virus strains cover a wide range of RNA viruses that are being studied our lab. Chikungunya virus, Semliki forest virus, and Sindbis virus belong to Alphavirus, which have single-strand (positive) RNA genome and are mostly transmitted by mosquitos to cause human diseases[14]. The mCherry protein is inserted into their genomes and driven by their sub-genomic promoter. Zika virus and Coxsakie virus are positive-strand RNA viruses that encode a single genomic polyprotein[493,494]. The mCherry protein is inserted in-frame with viral protein in their genomes. HIV-1 belongs to reverse transcribing RNA virus that its life cycle requires reverse transcription from RNA genome to DNA[1]. The mCherry is also

tagged in-frame with its viral protein. All these viruses have an RNA genome, therefore POIs that recognise viral RNA could have higher chance to be identified.

When analysing fluorescent signal profiles, an appropriate interpretation is crucial to identifying (novel) regulators of infection. If overexpression of a POI reduces viral protein signal, cell viability would also need to be checked because reduce in signal could be the result of an effective antiviral response or a strong anti-proliferative effect. The level of changes in viral protein signals could vary depending on the POI and the type of virus used in the assay. As shown in Fig.6.2 B, these POIs could have hugely different expression levels after 24h of doxycycline induction. Therefore, it could be misleading if we judge POIs' roles in infection by directly comparing changes in viral protein signals between different POIs. Whether a POI can modulate virus infection should be determined by comparing the fluorescence profiles between doxycycline-induced and un-induced cells within the same cell line.

When a POI only showed a moderated but significant impact on the viral protein signal (i.e. less than 50% increase/decrease), it could still be an important regulator of infection, because firstly, the time window of the fluorescence assay could be suboptimal for a POI to exert its antiviral effect. Some host proteins antagonise viruses by affecting their ability to deliver viable offsprings, for example tetherin (also BST2) can limit formation of HIV-1 particle by "tethering" the nascent viral particles onto the surface of parent cells[49], and APOBEC3G can be incorporated inside HIV-1 capsid to induce hypermutations on HIV-1 genome[48]. The antiviral effect of this type of POIs will be amplified as virus passage number increases, therefore cannot be captured by plate reader assays in a short time window. Measures such as performing multiple rounds of infection with supernatants could uncover this effect. Secondly, many antiviral proteins rely on other proteins or complexes to exert their functions, for example OASL require expression of RIG-I to exert its antiviral effect[495]. Solely overexpress one POI could have limited effect. Thirdly, antiviral effect of host proteins could be specific to certain type of virus. One example is TRIM79α, which direct degradation of NS5 of tick-borne encephalitis virus (TBEV) but have little impact on West Nile virus (WNV), despite that both viruses belong to flavivirus[496]. Therefore, the level of signal changes after POI overexpression are restricted to the viruses to be used in this assay. Similarly, when the overexpression assay shows no impact on viral protein signal, we cannot rule out the possibility that the analysed POI can regulate other types of viruses in a different cellular environment. It is also possible that the tested cell line cannot produce functional POI despite having GFP expression. This could be caused by factors such as misfolding and lack of post-translational modification.

Acknowledgements

I want to thank all members of Castello lab for the supports in cell cultures. Thanks Alfredo Castello-Palomares & Shabaz Mohammed for discussions regarding the candidate selection for generation of stable lines.

Method and material

Methods

Molecular cloning

HEK293 cells cultured in a 6-well plate at 80% confluency were harvested for cDNA library construction after treatment of 200 U/ml IFN-α for 20 h. Total RNA extraction was performed with Monarch Total RNA kit following the manufacturer's protocol. Library was generated by reverse transcription of extracted RNA with SuperScript III reverse transcriptase (Invitrogen) using oligo(dT) primer after supplied with RiboLock RNase inhibitor (Thermo Scientific). The coding sequence of CTDNEP1 and DTX3L were cloned from this cDNA library with PCR using primers designed to amplify the respective CDSs. DTX3L sequence was flanked by BamHI and NotI, and CTDNEP1 sequence was flanked by HindIII and NotI restriction endonucleases sites. They were cloned into the multiple cloning site (MCS) of pcDNA5 FRT/TO linker-eGFP vector, which is modified from pcDNA5 FRT/TO vector by adding a Gly-Ser linker and eGFP after the MCS.

Gene synthesis

Coding sequences of BTF3, CMPK2, DIMT1, DTWD1, ENTPD1, EPSTI1, FRMD5, GDE1, HERC6, HNRNPD, LGALS3BP, MTRF1, NCOA4, POLDIP3, THAP10, UBE2J2, and UBE2L6 were generated by gene synthesis through Bio Basics Inc. Synthesised coding sequences were flanked with BamHI/HindIII/KpnI at 5' and NotI at 3' ends restriction sites and cloned into the pcDNA5 FRT/TO linker eGFP vector.

Cell culture

HEK293 Flp-In TREx cells (commercially available from Invitrogen) were cultured in DMEM with 10 % FBS, 1x penicillin/streptomycin, 15 µg/ml Blasticidin S, and 100 µg/ml Zeocin at 37°C with 5% CO_2. Generation of POI-eGFP expressing cell lines, HEK293 Flp-In TREx cells were cultured on 6-well plates to 50% confluency then co-transfected with pOG44 and corresponding pcDNA5 FRT/TP POI-linker-eGFP vectors at 1:1 ratio (w/w) using XtremeGene 9 DNA transfection reagent and OptiMEM reduced serum in a DNA/reagent/serum ratio of 2µg/6µl/200µl. After 48 h, transfected cells were transferred to T25 culture flasks using selection media (DMEM with 10 % FBS, 1x pen/strep, 15 µg/ml Blasticidin S, and 150 µg/ml Hygromycin B), with replacement of fresh media every 3~4 days. Surviving colonies were transferred to 10-cm culture dish for amplification. Protein expression induction was performed by supplying the culture media with 1 µg/ml Doxycycline. Cells surviving Hygromycin(+) media and showing eGFP fluorescence after induction were considered positive cell lines.

Mateirals

Chemicals and reagents

REAGENT or RESOURCE	SOURCE	IDENTIFIER
2 log DNA ladder	NEB	N3200S
Agarose	MP Biomedicals	AGAP0100
BamHI-HF	NEB	R3136
Blasticidin	TOKU-E	B001
DMEM	Gibco	#11995-065
Doxycycline	Sigma Aldrich	D9891
FBS	Sigma Aldrich	F9665
HindIII-HF	NEB	R3104
Hygromycin B	TOKU-E	H011
Library Efficiency DH5α competent cells	Invitrogen	#18263012
Monarch Total RNA Miniprep Kit	NEB	T2010S
NEBNext rRNA Depletion Kit	NEB	E6310
NotI-HF	NEB	R3189
Opti-MEM reduced serum medium	Thermo Fisher	#31985062
PBS 1x, pH 7.4	Gibco	#10010023
penicillin/streptomycin	Sigma Aldrich	P4458
Phusion-HF DNA polymerase	NEB	M0530
QIAquick Gel Extraction kit	QIAgen	#28706X4
QIAquick PCR Purification kit	QIAgen	#28104
RiboLock RNase Inhibitor	Thermo Scientific	EO0381

SuperScript III Reverse Transcriptase	Invitrogen	#18080093
T4 DNA ligase	NEB	M0202
Trypsin EDTA	Sigma Aldrich	T3924
X-tremeGENE 9 DNA transfection reagent	Sigma Aldrich	#6365787001
Zeocin	Invitrogen	R25001

Instruments

INSTRUMENTS	SUPPLIER/INSTITUTE
T100 Thermal Cycler	Bio-Rad
Nanodrop 2000 Spectrophotometer	Thermo Fisher

Plasmids

PLASMIDS	SOURCE
pOG44 Flp-Recombinase vector	Invitrogen
pcDNA5-FRT/TO	Invitrogen
pcDNA5-FRT/TO eGFP-linker-MCS	Castello lab
pcDNA5-FRT/TO MCS-linker-eGFP	Castello lab

Oligonucleotides

NAME	SEQUENCE
CTDNEP1_F	CCCAAGCTTGTCATCGGGATGATGCGGACGCAGTGTCTG
CTDNEP1_R	TTTGCGGCCGCCTGTCACCAGAGCCGATGTTGG
DTX3L_F	CTCGGATCCATGGCCTCCCACCTGCG
DTX3L_R	TTTGCGGCCGCCATCTTCCAGCAGTTGTCTTACTCAATTCC

Cell line

CELL LINE	SOURCE	IDENTIFIER
HEK293 Flp-In TREx	Invitrogen	R78007

Chapter 7

General discussion

Overview

Host proteins are widely involved in virus life cycle and antiviral activities. Recent years have witnessed discoveries of various host factors that positively or negatively regulate virus infection[101,166,206]. Mass spectrometry (MS) based proteomic analysis has become increasingly important for these studies[146]. To comprehensively explore regulators of virus infection in the host, I used cells in three distinct states: 1), HEK293T cells that are highly permissive to multiple virus types; 2), HEK293 cells as cells in steady-state with competent interferon pathway; and 3), IFN-α stimulated HEK293 cell that is restrictive to virus infection. I employed deep proteomic analysis to study changes in protein abundance in these cellular states. In addition, I exploited the high versatility of proteomic analysis to study how the cellular states affect phosphorylations and RNA-binding activities of cellular proteins. This integrated multi-level proteomic analysis provide a rich dataset that covers different aspects of host responses to infection. From this dataset, I ranked and highlighted 25 proteins of interests (POIs) for construction of inducible stable expression cell lines, which lays a foundation stone for follow-up virology and mechanistic studies.

7.1 High quality whole-cell proteome data from cells of distinct permissiveness to viruses

As mentioned above, I studied three cell systems that consist of cells at highly-permissive state (HEK293T), steady-state (HEK293), and virus-restrictive state (IFN-α stimulated HEK293). I validated their different ability to sustain life cycles of alphavirus Sindbis virus (SINV) and lentivirus human immunodeficiency virus type 1 (HIV-1) in Chapter 2 & 3. All three cell states in this model system are all originated from HEK293 cells. They share highly similar genetic background and cell morphology[497], therefore comparative proteomics in this system can maximally highlight proteins that cause different permissiveness to viruses. Three cell states in my model system are all uninfected cells. Several studies have explored regulators of virus infection through proteomic analysis of virus-infected cells[165,498]. Their approaches can faithfully record the battlefront in host-virus confrontation. However, many viruses can suppress host responses, for example SINV can induce protein synthesis shutoff in host cells to prevent host antiviral response at the protein level[499,151]. By contrast, I used uninfected cells for proteomic analysis in this study, because my interest lay on understanding host factors that determine the cellular environment which viruses encounter during infection. I have validated the differences in viral fitness using two taxonomically

distant viruses (SINV & HIV-1)[13,29], indicating that the differences in virus permissiveness are robust and widespread.

Whole-cell proteome (WCP) profiling of these three cell states were performed with stable isotope labelling (SILAC) and off-line fractionation prior to LC-MS/MS[142,500]. This experimental design allows high accuracy in quantification and great depth in detection. The overall result achieved over 10,000 protein identification with more than 9,800 proteins are quantified in at least two cell states. In addition, the application of SILAC in deep proteomic profiling improved measurement accuracy, as well as reduced the total MS instrument running time, as differentially labelled samples are combined and analysed simultaneously. Running a large amount of samples through the LC-MS/MS device for a long period of time inevitably reduces its performance overtime. SILAC method alleviated this issue, therefore maintained a consistent instrument performance throughout the deep proteome profiling. In data analysis, I applied a linear regression method that is suitable for SILAC data. This method couples Bayesian estimation-based moderated T-test with appropriate regression of batch effects derived from the SILAC experiment[501,502]. In comparison, conventional ratio-based analysis is vulnerable to missing values therefore risks substantial loss of data points. Standard T-statistics cannot reflect the benefit of SILAC that processing errors downstream of sample combination are equally distributed to differentially-labelled samples. Quality control plots in Chapter 3 (p-value histograms & PCA plots) confirmed the superior performance of my data analysis approach over other methods in my SILAC datasets. This method, which combines moderated T-statistics with regression of batch effects, could be generalizable to other SILAC-based proteomics studies.

7.2 Differential proteome landscapes determine cell permissiveness to viruses

Deep proteome profiling revealed widespread differences in the cellular proteome between the highly-permissive HEK293T cells and steady-state HEK293 cells. With a FDR<0.1 cut-off, 1,486 and 1,441 proteins were found to be enriched in HEK293 and HEK293T respectively (Chapter 2). Among the proteins and processes that are differentially enriched in each cell line, there are two notable differences – the enrichment of almost the entire innate immune response pathway in HEK293 cells, and the enrichment of various viral protein/RNA interactors in HEK293T cells. This discovery strongly supports the notion that the highly permissive cellular environment in HEK293T cells is shaped by both depletion of antiviral factors and surplus of host factors that contribute to

virus life cycle. Conversely, the comparative analysis of steady-state and IFN-α stimulated HEK293 cells showed a moderate difference in proteome. No significant difference was observed at 4h-post IFN-α stimulation, and 66 proteins were differentially expressed at 20h-post IFN-α stimulation (Chapter 3). Despite the smaller scale of changes, these protein-abundance changes are sufficient to turn the cellular environment into a virus-restrictive state. These proteins comprise both well-studied ISGs and proteins that do not have well-established role in IFN-α response. Notably, none of the IFN-α down-regulated proteins have known links with IFN-α, with the exception of GDE1 which is downregulated by IFN-β in a microarray study[287]. IFN-induced downregulations of cellular mRNAs have been described in numerous IFN studies, hundreds of genes have been reported to be repressed by IFN[153,231,286,287]. However, our knowledge regarding their roles in IFN response remain largely blank. How these IFN-α down-regulated proteins and other less-understood IFN-α upregulated proteins participate in shaping virus-restrictive cell states remain to be determined.

When I collectively compared proteomes of these three cell states, I found that proteins of significant changes in both proteome-wide comparisons are enriched in in IFN-α stimulated cells over steady-state HEK293 cells, and enriched in steady-state HEK293 cells over HEK293T cells (Chapter 3). A substantial proportion of these common changes are well-studied IFN-stimulated genes (ISGs), such as viral RNA sensor RIG-I (also DDX58) and antiviral effector OAS3. By contrast, no common change was found in the opposite direction (enrichment in highly permissive over steady-state cells, and in steady-state over virus-restrictive cells). Collectively, these observations highlight that the effectiveness in innate immune response could be the main driver in forging a virus-permissive or virus-hostile cellular environment. However, we cannot rule out that overexpression of virus-facilitating factors could also make cells more permissive to viruses. The cellular proteins that are known to interact with viral protein or viral RNA are enriched in HEK293T cells. These proteins are highly involved in gene expression regulations, such as splicing and translation (Chapter 2). Some of these proteins have been shown to be relevant to virus replication, for example HTSF1 (also HTATSF1) is recruited by HIV-1 viral protein Tat to enhance elongation during the transcription of viral RNA[503]. I suspect there are more regulators of virus infection to be found within this comparative proteomics dataset, calling for follow-up screening and validation studies.

I also employed RNA-seq to analyse transcriptome changes induced by IFN-α in HEK293 cells. Combined with the comparative proteomic analysis, this dataset provides an in-depth insight into the IFN-α response at both transcriptome and proteome levels. A substantial number of RNAs were regulated by IFN-α at 4h and 20h (126 and 419 genes respectively, Chapter 4); however, only 0 and 66 proteins displayed statistically significant changes at corresponding timepoints. This antagonistic result highlights a strong discordance between IFN-α responses at the RNA and protein levels.

Through series of correlation analyses, I showed that IFN-α induced upregulations at protein level do not have high correlation with features in the RNA-seq data. In addition, none of the IFN-α induced downregulations at protein level have corresponding changes at RNA level, and *vice versa*. When I expanded the correlation analysis to assess other features, such as 3'-UTR length and miRNA binding sites, none of these features could account for the discordance between RNA and protein level changes. Discordance between RNA and protein abundance in the cell have also been reported and discussed in a series of studies[273,296,298,299]. The RNA-protein discordance in IFN-α responses in this study agree with the current understanding that RNA and protein abundance are correlated, but not strictly deterministic[298]. Both transcriptome and proteome analysis provides valuable and complementing insights in understanding IFN-α response. The observed RNA-protein discordance also highlights the necessity to directly measure protein abundance when we want to access changes in protein abundance in a particular biological system.

7.3 Profiling of the phosphoproteome and RNA-bound proteome landscapes during IFN-α response

This study assessed phosphorylation landscapes in steady-state and IFN-α stimulated HEK293 cells. This phosphoproteome analysis was also performed with SILAC. The result that included data from two instruments has identified over 22,000 phosphorylation sites in 3,942 proteins, and more than 8,000 phosphosites were reliably quantified (Chapter 5). Although this phosphoproteome dataset also identified phosphorylation events in IFN response pathways, such as phosphorylation at serine 727 in STAT1[96], the majority of IFN-α regulated phosphosites are not located on known IFN-related proteins. IFN-α regulated phosphosites in this dataset showed no enrichment in signalling pathway, and functions of many phosphosites are unknown. As a matter of fact, only less than 3% of phosphosites in the human phosphoproteome have reported functions, and a large part of them the known functions are related to signalling events[274]. Interestingly, IFN-α regulated phosphosites are enriched in viral protein and viral RNA interactors (Chapter 5). Some of the highly connected viral protein/RNA interactors, such as PCBP2, are known to participate in host response to virus[412,413]. This observation suggests that IFN-α regulated phosphorylation of non-signalling proteins could play a role in modulating protein-protein or protein-RNA interactions between the host and the virus. Whether this hypothesis stands true remain an interesting direction for follow-up studies.

Another level of the IFN-α response investigated in this study is the respond of RNA-binding proteins (RBPs). Host RBPs are heavily involved in virus life cycle and host antiviral response[168,323]. Series of studies have revealed pervasive changes of the cellular RNA-bound proteome (RBPome) after virus infection[151,170]. This study employed RNA-interactome capture (RIC) technique[10], and showed that the host can also actively modulate its RBPome during antiviral response, without the presence of viruses (Chapter 5). Interestingly, RBPome changes driven by IFN-α appear to have an opposite trend compared with those driven by infection of SINV and SARS-CoV-2[151,170]. This interesting result provides a complementary perspective to existing studies about host responses to virus in RBPome. However, many outstanding questions remain to be investigated regarding IFN-α induced RBPome changes. Firstly, ISGs that are known to interact with viral RNA are mostly absent in this dataset. This could due to the lack of "baits" (i.e. viral RNA) for RBPs. Ideally, this "bait" should contain pathogen-associated molecular patterns (PAMPs), be stable in IFN-α stimulated host cells, and be compatible with the RIC experiment. Without the presence of "baits", the RBPome landscape in antiviral response would still be incomplete. Secondly, IFN-α stimulation alters host transcriptome, therefore also the pool of available RBP binding sites. It is possible that the observed RBPome changes is a by-product of this shift in RBP binding sites. Methods to study the protein binding maps of RNAs, such as iCLIP, can provide system-wide information on the binding preferences of IFN-α regulated RBPs with single-nucleotide resolution[504]. Further research considering PAMP-containing "baits" and information about RBP binding sites is required to deconvolute how IFN-α induced RBPome changes are affected by passive change in available binding sites, and active change in RNA-binding activities.

7.4 Development of novel imputation method suitable for label-free proteomic analysis

The development of a partial imputation method is another major innovation provided by this study (Chapter 5). Missing values in MS detection is a major issue in quantitative proteomic analysis. Conventionally, missing values in proteomic data are imputed by numbers that mimic the signal of background noise, or proteins that have a similar expression pattern[417,418]. Each imputation method inserts artefactual numbers into the original data based on different models, however, these models inevitably face constraints when applied to complexed proteomics data. The method proposed in this study does not impute all missing values in the dataset, and instead utilises moderated T-statistics to estimate the effect of missing values[502]; meanwhile, imputation by background noise is only applied to proteins that showed "ON-OFF" changes to ensure that all

proteins in the dataset can be tested by moderated T-statistics. The empirical Bayesian estimation method in moderated T-statistics allows estimation of variance for proteins that do not have complete data, therefore bypass the need to artificially impute missing data points. Compared with existing full imputation methods, this partial imputation introduces minimal level of artefactual numbers into the original data. This minimises the risk that the model for imputation does not fit the real-world situation. In addition, leaving missing data points un-imputed can provide moderated T-test a genuine degree of freedom, which preserves the information that certain data points in this protein MS result was missing. By contrast, full imputation methods do not inform the T-test which protein contained missing value in the MS result, therefore falsely boost the confidence of the test result.

In Chapter 5, this study compared the effectiveness of partial imputation with popular imputation methods using both simulation datasets and real-world datasets of label-free proteomics[214,419]. The superior performance of partial imputation in almost all tests highlights the high reliability and versatility of this method in label-free proteomic analysis. However, it is worth noticing that the partial imputation method also has constraints in its application. It is not compatible with SILAC proteomics. This is because analysis of SILAC data require linear regression of multiple co-variables, and accurate multi-variable regression require complete data matrix[439]. It is also not compatible with feature-extraction analysis, such as PCA, because most of these analysis methods also require complete data matrix. The partial imputation method has provided substantial improvement in analysis of IFN-α RIC dataset in this study. It has also been successfully applied in different types of proteomic data in our lab's recent publication and ongoing projects, and could be a powerful data analysis tool for a variety of proteomics studies[170].

7.5 Integrated multi-level proteomics facilitates selection of proteins that potentially have roles in infection

Integrated analysis of multi-level datasets in this study has facilitated the identification of putative regulators of virus infection with high confidence. In this study, I prioritised the selection of proteins of interests (POIs) on genes that show protein abundance changes across different cell states, in particular those that show RNA-protein discordance. This approach focuses on studying virus-regulating functions that are dependent on protein abundance. As a result, I have selected 25 POIs and generated Doxycycline-inducible stable-expression cell lines for 19 of them. I aim to

perform phenotype screen using multiple types of viruses with these stable lines. This strategy can largely relief the issues of redundancy between antiviral factors, artefacts induced by plasmid transfection, and artefacts induced by overexpression at non-physiological level. It is worth noticing that the phenotype screen with viruses is a crude search to highlight POIs that have significantly impact on virus replication. For POIs that show phenotype in virus-infection screens, strategies for mechanistic study will be tailored for each of them. For example, if the POI was only regulated by IFN-α at protein level, it could subject to translational regulation or protein stability regulation mechanisms. Translation regulations can be assessed through ribosome footprinting experiments such as ribosome profiling[352]. Many translation regulations are mediated by 5'/3'-UTR sequence[288]. GFP reporter system containing the identical UTR sequences could help study the mechanism of translation regulation. Protein stabilities can be assessed by stable isotope labelling-based pulse-chase experiments[357,505]. Ubiquitin-proteasome pathway is the main protein degradation process in the cell[289], and factors regulating POI ubiquitination could be studied by point mutation of ubiquitination site and interactome analysis.

The selection of POIs in this study primarily focused on those that could regulate virus infection by altered protein abundance. However, the rich content in this integrated multi-level proteomic dataset allows POI selection from multiple aspects. Here are some of the possible directions to explore, and there could be many other angles awaits to be exploited:

IFN-α mediated regulations of RNA-binding

Follow-up studies can prioritise on proteins that only show changes in RBPome after IFN-α stimulation. It is important to understand whether these changes are caused by altered RNA-binding activities or shifts in the number of available RBP binding sites. Binding sites of many RBPs have been characterised by RNA-seq based methods such as iCLIP, and consolidated into databases[504,506]. Information of RBP binding sites can be mapped to the IFN-α transcriptome dataset in this study to determine whether an IFN-α regulated RBP has altered binding site availability. It is interesting to study the downstream effects of IFN-α regulated RNA-binding. The role of loss-of-function in RNA-binding can be studied by point mutation or truncation that diminish the RNA binding activity. The role of gain-of-function can be studied by fusion with engineered RNA-binding domain (RBD). As an example, Pumilio domain consists of four types of repeating units that each specifically recognises one type of base in the RNA[507]. Engineering Pumilio domain by rearranging the order of repeating units can adjust its RNA specificity, and fusion with POIs can create a constantly RNA-bound chimera protein for functional tests[508]. An IFN-α regulated RBP could be a member of a protein complex, its

loss- or gain-of-function could determine whether the protein complex is localised on the target RNA. This can be assessed by proteomics method like interactome analysis, biochemical methods like Western blots and EMSA[509], and cell biology methods like single-molecule fluorescence microscopy[510]. It is also interesting to study the upstream mechanism of IFN-α regulated RNA-binding. Several of these RBPs contain IFN-α regulated phosphosites when mapped to the phosphoproteome dataset in this study, such as HNRNPD, SRRM2, etc. (Chapter 5). Some phosphosites are located near or within RBDs, therefore could be involved in regulation of RNA-binding activities. It is worth noticing that, because of the huge sample complexity and the stochastic nature of phosphopeptide quantification, IFN-α regulated phosphosites in a RBP could be missed or quantified with low confidence in my phosphoproteome dataset. Eluates from pulldown of RBP or its bound RNA have greatly reduced sample complexity, and phosphoproteome analysis of these samples could reveal IFN-α regulated phosphosites that are missed or poorly quantified in my phosphoproteome data.

Virus-regulating roles of retroviral and transposon elements

Retroviral and transposable elements are conventionally thought as parasites in the human genome. However, researchers start to realise that these seemingly "redundant" elements could participate in cellular immune response[330,511]. Notably, these elements are also found in different levels of omics data in this study. IFN-α transcriptome data in this study showed strong overexpression of a pseudogene of endogenous retroviral *Env* gene[329]. Proteome data of virus-permissive HEK293T cells showed that protein ORF1 from LINE-1 transposable element is strongly overexpressed[512]. Although these two elements are overexpressed in two cell states that have opposite permissiveness to viruses, their presence suggest that these elements could be involved in shaping cell permissiveness to viruses. A recent study showed that overexpression of endogenous retroviral element HERVK endows the host with enhanced antiviral activity by activating innate immune response[513]. The overexpression of retroviral *Env* pseudogene could have similar role, and could form a positive feedback loop to strengthen innate immune response in host cells. The role of transposon-derived product ORF1 in the cell remain largely unknown. Overexpression of ORF1 could be the cause of virus-permissive cellular environment. Recent studies have reported that protein product of retroviral element could repress innate immunity through protein-protein interactions, and ORF1 could have similar function[332]. Overexpression of ORF1 could also be the consequence of virus-permissive cellular environment. This study has revealed that HEK293T has significantly lower expression of the entire innate immune response pathway. Transcript of ORF1 could escape cell

surveillance in this cellular environment. However, recent researches have pointed that piRNA (a class of microRNA[514]) pathway is the main suppressor of transposons in the cell[515,516]. It is therefore interesting to investigate why transposon-derived ORF1 is not suppressed by piRNA in HEK293T. It is possible that piRNA-mediated transposon silencing pathway is inactive in HEK293T compared with stead-state cells, and piRNA pathway could be involved in shaping virus-permissive/restrictive environments.

Regulatory roles of viral protein and viral RNA interactors

Differential enrichment of viral protein/RNA interactors are found in multiple levels of proteomics datasets in this study. Highly permissive HEK293T cells are enriched in viral protein/RNA RNA interactors over steady-state cells; proteins that contain IFN-α regulated phosphosites are also enriched in viral protein/RNA interactors over cellular phosphorylate-able proteome; IFN-α regulated RBPs are enriched in viral protein interactors over cellular proteome. Follow-up studies in this direction can prioritise on viral protein/RNA interactors that show significant changes in multiple omics data. Some of these virus interactors are well-established antiviral proteins. For example, HERC5 is an E3 ligase for ISG15 in ISG15-mediated antiviral process[113], it is depleted in HEK293T cells, and has increased RNA-binding after IFN-α stimulation. There are many other virus interactors that do not have well-studied roles in virus infection. LARP1 is a translation regulator downstream of mTOR complex[517]. It can interact with Reovirus viral protein $\sigma 3^4$, SARS-CoV-2 viral protein N[206], and viral RNA from multiple viruses[447]. Evidences showed that LARP1 is associated with replication of DENV and virion formation of HCV, however little is known about the role of LARP1 in cell response to virus[518,519]. HNRNP proteins are a large class of nucleus proteins participate in various steps of RNA processing[520]. Several HNRNPs, including HNRNPD and HNRNPU, are differentially expressed in multiple proteomic results in this study, and can interact with viral RNA from a broad range of virus species[447]. The rich content in this multi-level proteomics dataset can provide extra evidences on host-virus relation for viral protein/RNA interactors, therefore help highlight high-confidence POIs for downstream studies.

Appendices

Supplementary table 1 – Comparative deep whole-cell proteomic analysis result between HEK293 and HEK293T cells with SILAC.

Supplementary table 2 – Comparative deep whole-cell proteomic analysis result between mock, 4h IFN-α, and 20h IFN-α stimulated HEK293 cells with SILAC.

Supplementary table 3 – Clustering analyses results of IFN-α proteomics data. Included dynamic group analysis and hierarchical clustering analysis in Chapter 3.

Supplementary table 4 – Comparative RNA-sequencing analysis result between mock, 4h IFN-α, and 20h IFN-α stimulated HEK293 cells.

Supplementary table 5 – Comparative deep phosphoproteomic analysis result between mock, 10min IFN-α, and 4h IFN-α stimulated HEK293 cells with SILAC. Included a result table that separates phosphosite data acquired from Q-Exactive and Orbitrap-Elite, and a result table that combines data from two instruments following the pipeline introduced in Chapter 5.

Supplementary table 6 – Kinase-substrate enrichment analysis (KSEA) result of IFN-α phospho-proteome data. Included the sources of kinase-substrate annotations used in this study, the KSEA result that only uses curated kinase-substrate annotations, the KSEA result that uses curated and text-mining based annotations, and the KSEA result that uses curated, text-mining, and motif prediction based annotations.

Supplementary table 7 – Comparative RNA-bound proteome analysis result between mock and 20h IFN-α stimulated HEK293 cells.

Supplementary table 8 – Lists of Gene Ontology (GO) annotations and Pfam motif annotations used in enrichment analyses in this study. Included the list of GO terms related to innate immunity, the list of GO terms related to type I IFN, the list of GO terms related to virus, and the list of Pfam domains related to RNA binding.

References

1 Karn, J. & Stoltzfus, C. M. Transcriptional and posttranscriptional regulation of HIV-1 gene expression. *Cold Spring Harb Perspect Med* **2**, a006916, doi:10.1101/cshperspect.a006916 (2012).

2 Lloyd, R. E. Nuclear proteins hijacked by mammalian cytoplasmic plus strand RNA viruses. *Virology* **479-480**, 457-474, doi:10.1016/j.virol.2015.03.001 (2015).

3 Bushell, M. & Sarnow, P. Hijacking the translation apparatus by RNA viruses. *J Cell Biol* **158**, 395-399, doi:10.1083/jcb.200205044 (2002).

4 Pichlmair, A. *et al.* Viral immune modulators perturb the human molecular network by common and unique strategies. *Nature* **487**, 486-490, doi:10.1038/nature11289 (2012).

5 Barbalat, R., Ewald, S. E., Mouchess, M. L. & Barton, G. M. Nucleic acid recognition by the innate immune system. *Annu Rev Immunol* **29**, 185-214, doi:10.1146/annurev-immunol-031210-101340 (2011).

6 Platanias, L. C. Mechanisms of type-I- and type-II-interferon-mediated signalling. *Nat Rev Immunol* **5**, 375-386, doi:10.1038/nri1604 (2005).

7 Schoggins, J. W. Interferon-stimulated genes: roles in viral pathogenesis. *Curr Opin Virol* **6**, 40-46, doi:10.1016/j.coviro.2014.03.006 (2014).

8 Domon, B. & Aebersold, R. Options and considerations when selecting a quantitative proteomics strategy. *Nat Biotechnol* **28**, 710-721, doi:10.1038/nbt.1661 (2010).

9 Mann, M. & Jensen, O. N. Proteomic analysis of post-translational modifications. *Nat Biotechnol* **21**, 255-261, doi:10.1038/nbt0303-255 (2003).

10 Castello, A. *et al.* Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell* **149**, 1393-1406, doi:10.1016/j.cell.2012.04.031 (2012).

11 Johnson, G. P., Goebel, S. J. & Paoletti, E. An update on the vaccinia virus genome. *Virology* **196**, 381-401, doi:10.1006/viro.1993.1494 (1993).

12 Henchal, E. A. & Putnak, J. R. The dengue viruses. *Clin Microbiol Rev* **3**, 376-396, doi:10.1128/CMR.3.4.376 (1990).

13 Chen, R. *et al.* ICTV Virus Taxonomy Profile: Togaviridae. *J Gen Virol* **99**, 761-762, doi:10.1099/jgv.0.001072 (2018).

14 Ryman, K. D. & Klimstra, W. B. Host responses to alphavirus infection. *Immunol Rev* **225**, 27-45, doi:10.1111/j.1600-065X.2008.00670.x (2008).

15 Strauss, J. H. & Strauss, E. G. The alphaviruses: gene expression, replication, and evolution. *Microbiol Rev* **58**, 491-562, doi:10.1128/mr.58.3.491-562.1994 (1994).

16 Wang, K. S., Kuhn, R. J., Strauss, E. G., Ou, S. & Strauss, J. H. High-affinity laminin receptor is a receptor for Sindbis virus in mammalian cells. *J Virol* **66**, 4992-5001, doi:10.1128/JVI.66.8.4992-5001.1992 (1992).

17 Zhang, R. *et al.* Mxra8 is a receptor for multiple arthritogenic alphaviruses. *Nature* **557**, 570-574, doi:10.1038/s41586-018-0121-3 (2018).

18 Li, L., Jose, J., Xiang, Y., Kuhn, R. J. & Rossmann, M. G. Structural changes of envelope proteins during alphavirus fusion. *Nature* **468**, 705-708, doi:10.1038/nature09546 (2010).

19 Grimley, P. M., Berezesky, I. K. & Friedman, R. M. Cytoplasmic structures associated with an arbovirus infection: loci of viral ribonucleic acid synthesis. *J Virol* **2**, 1326-1338, doi:10.1128/JVI.2.11.1326-1338.1968 (1968).

20 Soonsawad, P. *et al.* Structural evidence of glycoprotein assembly in cellular membrane compartments prior to Alphavirus budding. *J Virol* **84**, 11145-11151, doi:10.1128/JVI.00036-10 (2010).

21 Elmasri, Z., Nasal, B. L. & Jose, J. Alphavirus-Induced Membrane Rearrangements during Replication, Assembly, and Budding. *Pathogens* **10**, doi:10.3390/pathogens10080984 (2021).

22 Jones, R., Bragagnolo, G., Arranz, R. & Reguera, J. Capping pores of alphavirus nsP1 gate membranous viral replication factories. *Nature* **589**, 615-619, doi:10.1038/s41586-020-3036-8 (2021).

23 Mazzon, M. *et al.* Alphavirus-induced hyperactivation of PI3K/AKT directs pro-viral metabolic changes. *Plos Pathog* **14**, e1006835, doi:10.1371/journal.ppat.1006835 (2018).
24 Spuul, P., Balistreri, G., Kaariainen, L. & Ahola, T. Phosphatidylinositol 3-kinase-, actin-, and microtubule-dependent transport of Semliki Forest Virus replication complexes from the plasma membrane to modified lysosomes. *J Virol* **84**, 7543-7557, doi:10.1128/JVI.00477-10 (2010).
25 Thaa, B. *et al.* Differential Phosphatidylinositol-3-Kinase-Akt-mTOR Activation by Semliki Forest and Chikungunya Viruses Is Dependent on nsP3 and Connected to Replication Complex Internalization. *J Virol* **89**, 11420-11437, doi:10.1128/JVI.01579-15 (2015).
26 Radoshitzky, S. R. *et al.* siRNA Screen Identifies Trafficking Host Factors that Modulate Alphavirus Infection. *Plos Pathog* **12**, e1005466, doi:10.1371/journal.ppat.1005466 (2016).
27 Dodonova, S. O. *et al.* 9A structure of the COPI coat reveals that the Arf1 GTPase occupies two contrasting molecular environments. *Elife* **6**, doi:10.7554/eLife.26691 (2017).
28 Brown, R. S., Wan, J. J. & Kielian, M. The Alphavirus Exit Pathway: What We Know and What We Wish We Knew. *Viruses* **10**, doi:10.3390/v10020089 (2018).
29 Coffin, J. *et al.* ICTV Virus Taxonomy Profile: Retroviridae 2021. *J Gen Virol* **102**, doi:10.1099/jgv.0.001712 (2021).
30 HIV/AIDS, J. U. N. P. o. *Global HIV & AIDS Statistics — 2022 Fact Sheet*, <https://www.unaids.org/en/resources/fact-sheet> (2022).
31 Briggs, J. A. & Krausslich, H. G. The molecular architecture of HIV. *J Mol Biol* **410**, 491-500, doi:10.1016/j.jmb.2011.04.021 (2011).
32 Chen, B. Molecular Mechanism of HIV-1 Entry. *Trends Microbiol* **27**, 878-891, doi:10.1016/j.tim.2019.06.002 (2019).
33 Francis, A. C. & Melikyan, G. B. Single HIV-1 Imaging Reveals Progression of Infection through CA-Dependent Steps of Docking at the Nuclear Pore, Uncoating, and Nuclear Transport. *Cell Host Microbe* **23**, 536-548 e536, doi:10.1016/j.chom.2018.03.009 (2018).
34 el Kharroubi, A. & Martin, M. A. cis-acting sequences located downstream of the human immunodeficiency virus type 1 promoter affect its chromatin structure and transcriptional activity. *Mol Cell Biol* **16**, 2958-2966, doi:10.1128/MCB.16.6.2958 (1996).
35 Gatignol, A. Transcription of HIV: Tat and cellular chromatin. *Adv Pharmacol* **55**, 137-159, doi:10.1016/S1054-3589(07)55004-0 (2007).
36 Kao, S. Y., Calman, A. F., Luciw, P. A. & Peterlin, B. M. Anti-termination of transcription within the long terminal repeat of HIV-1 by tat gene product. *Nature* **330**, 489-493, doi:10.1038/330489a0 (1987).
37 Kim, Y. K., Bourgeois, C. F., Isel, C., Churcher, M. J. & Karn, J. Phosphorylation of the RNA polymerase II carboxyl-terminal domain by CDK9 is directly responsible for human immunodeficiency virus type 1 Tat-activated transcriptional elongation. *Mol Cell Biol* **22**, 4622-4637, doi:10.1128/MCB.22.13.4622-4637.2002 (2002).
38 Buratowski, S. Progression through the RNA polymerase II CTD cycle. *Mol Cell* **36**, 541-546, doi:10.1016/j.molcel.2009.10.019 (2009).
39 Stoltzfus, C. M. Chapter 1. Regulation of HIV-1 alternative RNA splicing and its role in virus replication. *Adv Virus Res* **74**, 1-40, doi:10.1016/S0065-3527(09)74001-1 (2009).
40 Truman, C. T., Jarvelin, A., Davis, I. & Castello, A. HIV Rev-isited. *Open Biol* **10**, 200320, doi:10.1098/rsob.200320 (2020).
41 Rausch, J. W. & Le Grice, S. F. HIV Rev Assembly on the Rev Response Element (RRE): A Structural Perspective. *Viruses* **7**, 3053-3075, doi:10.3390/v7062760 (2015).
42 Matsuura, Y. & Stewart, M. Structural basis for the assembly of a nuclear export complex. *Nature* **432**, 872-877, doi:10.1038/nature03144 (2004).
43 Fischer, U., Huber, J., Boelens, W. C., Mattaj, I. W. & Luhrmann, R. The HIV-1 Rev activation domain is a nuclear export signal that accesses an export pathway used by specific cellular RNAs. *Cell* **82**, 475-483, doi:10.1016/0092-8674(95)90436-0 (1995).

44 Sundquist, W. I. & Krausslich, H. G. HIV-1 assembly, budding, and maturation. *Cold Spring Harb Perspect Med* **2**, a006924, doi:10.1101/cshperspect.a006924 (2012).
45 Kirchhoff, F. Immune evasion and counteraction of restriction factors by HIV-1 and other primate lentiviruses. *Cell Host Microbe* **8**, 55-67, doi:10.1016/j.chom.2010.06.004 (2010).
46 Wiegand, H. L., Doehle, B. P., Bogerd, H. P. & Cullen, B. R. A second human antiretroviral factor, APOBEC3F, is suppressed by the HIV-1 and HIV-2 Vif proteins. *EMBO J* **23**, 2451-2458, doi:10.1038/sj.emboj.7600246 (2004).
47 Sheehy, A. M., Gaddis, N. C., Choi, J. D. & Malim, M. H. Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. *Nature* **418**, 646-650, doi:10.1038/nature00939 (2002).
48 Yu, X. *et al.* Induction of APOBEC3G ubiquitination and degradation by an HIV-1 Vif-Cul5-SCF complex. *Science* **302**, 1056-1060, doi:10.1126/science.1089591 (2003).
49 Hinz, A. *et al.* Structural basis of HIV-1 tethering to membranes by the BST-2/tetherin ectodomain. *Cell Host Microbe* **7**, 314-323, doi:10.1016/j.chom.2010.03.005 (2010).
50 Iwabu, Y. *et al.* HIV-1 accessory protein Vpu internalizes cell-surface BST-2/tetherin through transmembrane interactions leading to lysosomes. *J Biol Chem* **284**, 35060-35072, doi:10.1074/jbc.M109.058305 (2009).
51 Mangeat, B. *et al.* HIV-1 Vpu neutralizes the antiviral factor Tetherin/BST-2 by binding it and directing its beta-TrCP2-dependent degradation. *Plos Pathog* **5**, e1000574, doi:10.1371/journal.ppat.1000574 (2009).
52 Roeth, J. F. & Collins, K. L. Human immunodeficiency virus type 1 Nef: adapting to intracellular trafficking pathways. *Microbiol Mol Biol Rev* **70**, 548-563, doi:10.1128/MMBR.00042-05 (2006).
53 Sadler, A. J. & Williams, B. R. Interferon-inducible antiviral effectors. *Nat Rev Immunol* **8**, 559-568, doi:10.1038/nri2314 (2008).
54 Kumar, H., Kawai, T. & Akira, S. Pathogen recognition by the innate immune system. *Int Rev Immunol* **30**, 16-34, doi:10.3109/08830185.2010.529976 (2011).
55 Kawai, T. & Akira, S. The role of pattern-recognition receptors in innate immunity: update on Toll-like receptors. *Nat Immunol* **11**, 373-384, doi:10.1038/ni.1863 (2010).
56 Jin, M. S. *et al.* Crystal structure of the TLR1-TLR2 heterodimer induced by binding of a tri-acylated lipopeptide. *Cell* **130**, 1071-1082, doi:10.1016/j.cell.2007.09.008 (2007).
57 Park, B. S. *et al.* The structural basis of lipopolysaccharide recognition by the TLR4-MD-2 complex. *Nature* **458**, 1191-1195, doi:10.1038/nature07830 (2009).
58 Hornung, V. *et al.* Sequence-specific potent induction of IFN-alpha by short interfering RNA in plasmacytoid dendritic cells through TLR7. *Nat Med* **11**, 263-270, doi:10.1038/nm1191 (2005).
59 Haas, T. *et al.* The DNA sugar backbone 2' deoxyribose determines toll-like receptor 9 activation. *Immunity* **28**, 315-323, doi:10.1016/j.immuni.2008.01.013 (2008).
60 Bell, J. K., Askins, J., Hall, P. R., Davies, D. R. & Segal, D. M. The dsRNA binding site of human Toll-like receptor 3. *Proc Natl Acad Sci U S A* **103**, 8792-8797, doi:10.1073/pnas.0603245103 (2006).
61 Dunne, A., Ejdeback, M., Ludidi, P. L., O'Neill, L. A. & Gay, N. J. Structural complementarity of Toll/interleukin-1 receptor domains in Toll-like receptors and the adaptors Mal and MyD88. *J Biol Chem* **278**, 41443-41451, doi:10.1074/jbc.M301742200 (2003).
62 Akira, S. & Takeda, K. Toll-like receptor signalling. *Nat Rev Immunol* **4**, 499-511, doi:10.1038/nri1391 (2004).
63 Reikine, S., Nguyen, J. B. & Modis, Y. Pattern Recognition and Signaling Mechanisms of RIG-I and MDA5. *Front Immunol* **5**, 342, doi:10.3389/fimmu.2014.00342 (2014).
64 Pichlmair, A. *et al.* RIG-I-mediated antiviral responses to single-stranded RNA bearing 5'-phosphates. *Science* **314**, 997-1001, doi:10.1126/science.1132998 (2006).

65 Kato, H. *et al.* Differential roles of MDA5 and RIG-I helicases in the recognition of RNA viruses. *Nature* **441**, 101-105, doi:10.1038/nature04734 (2006).
66 Hou, F. *et al.* MAVS forms functional prion-like aggregates to activate and propagate antiviral innate immune response. *Cell* **146**, 448-461, doi:10.1016/j.cell.2011.06.041 (2011).
67 Liu, S. *et al.* MAVS recruits multiple ubiquitin E3 ligases to activate antiviral signaling cascades. *Elife* **2**, e00785, doi:10.7554/eLife.00785 (2013).
68 Seth, R. B., Sun, L., Ea, C. K. & Chen, Z. J. Identification and characterization of MAVS, a mitochondrial antiviral signaling protein that activates NF-kappaB and IRF 3. *Cell* **122**, 669-682, doi:10.1016/j.cell.2005.08.012 (2005).
69 Hopfner, K. P. & Hornung, V. Molecular mechanisms and cellular functions of cGAS-STING signalling. *Nat Rev Mol Cell Biol* **21**, 501-521, doi:10.1038/s41580-020-0244-x (2020).
70 Zhang, X. *et al.* Cyclic GMP-AMP containing mixed phosphodiester linkages is an endogenous high-affinity ligand for STING. *Mol Cell* **51**, 226-235, doi:10.1016/j.molcel.2013.05.022 (2013).
71 Sun, L., Wu, J., Du, F., Chen, X. & Chen, Z. J. Cyclic GMP-AMP synthase is a cytosolic DNA sensor that activates the type I interferon pathway. *Science* **339**, 786-791, doi:10.1126/science.1232458 (2013).
72 Liu, S. *et al.* Phosphorylation of innate immune adaptor proteins MAVS, STING, and TRIF induces IRF3 activation. *Science* **347**, aaa2630, doi:10.1126/science.aaa2630 (2015).
73 Abe, T. & Barber, G. N. Cytosolic-DNA-mediated, STING-dependent proinflammatory gene induction necessitates canonical NF-kappaB activation through TBK1. *J Virol* **88**, 5328-5341, doi:10.1128/JVI.00037-14 (2014).
74 Burdette, D. L. *et al.* STING is a direct innate immune sensor of cyclic di-GMP. *Nature* **478**, 515-518, doi:10.1038/nature10429 (2011).
75 Chow, K. T., Gale, M., Jr. & Loo, Y. M. RIG-I and Other RNA Sensors in Antiviral Immunity. *Annu Rev Immunol* **36**, 667-694, doi:10.1146/annurev-immunol-042617-053309 (2018).
76 Gu, L., Fullam, A., Brennan, R. & Schroder, M. Human DEAD box helicase 3 couples IkappaB kinase epsilon to interferon regulatory factor 3 activation. *Mol Cell Biol* **33**, 2004-2015, doi:10.1128/MCB.01603-12 (2013).
77 Oshiumi, H., Sakai, K., Matsumoto, M. & Seya, T. DEAD/H BOX 3 (DDX3) helicase binds the RIG-I adaptor IPS-1 to up-regulate IFN-beta-inducing potential. *Eur J Immunol* **40**, 940-948, doi:10.1002/eji.200940203 (2010).
78 Miyashita, M., Oshiumi, H., Matsumoto, M. & Seya, T. DDX60, a DEXD/H box helicase, is a novel antiviral factor promoting RIG-I-like receptor-mediated signaling. *Mol Cell Biol* **31**, 3802-3819, doi:10.1128/MCB.01368-10 (2011).
79 Takaoka, A. *et al.* DAI (DLM-1/ZBP1) is a cytosolic DNA sensor and an activator of innate immune response. *Nature* **448**, 501-505, doi:10.1038/nature06013 (2007).
80 Jonsson, K. L. *et al.* IFI16 is required for DNA sensing in human macrophages by promoting production and function of cGAMP. *Nat Commun* **8**, 14391, doi:10.1038/ncomms14391 (2017).
81 Unterholzner, L. *et al.* IFI16 is an innate immune sensor for intracellular DNA. *Nat Immunol* **11**, 997-1004, doi:10.1038/ni.1932 (2010).
82 Jiang, Z. *et al.* IFI16 directly senses viral RNA and enhances RIG-I transcription and activation to restrict influenza virus infection. *Nat Microbiol* **6**, 932-945, doi:10.1038/s41564-021-00907-x (2021).
83 Pestka, S., Krause, C. D. & Walter, M. R. Interferons, interferon-like cytokines, and their receptors. *Immunol Rev* **202**, 8-32, doi:10.1111/j.0105-2896.2004.00204.x (2004).
84 Valente, G. *et al.* Distribution of interferon-gamma receptor in human tissues. *Eur J Immunol* **22**, 2403-2412, doi:10.1002/eji.1830220933 (1992).

85 Sommereyns, C., Paul, S., Staeheli, P. & Michiels, T. IFN-lambda (IFN-lambda) is expressed in a tissue-dependent fashion and primarily acts on epithelial cells in vivo. *Plos Pathog* **4**, e1000017, doi:10.1371/journal.ppat.1000017 (2008).
86 Lazear, H. M., Schoggins, J. W. & Diamond, M. S. Shared and Distinct Functions of Type I and Type III Interferons. *Immunity* **50**, 907-923, doi:10.1016/j.immuni.2019.03.025 (2019).
87 Silvennoinen, O., Ihle, J. N., Schlessinger, J. & Levy, D. E. Interferon-induced nuclear signalling by Jak protein tyrosine kinases. *Nature* **366**, 583-585, doi:10.1038/366583a0 (1993).
88 Darnell, J. E., Jr. STATs and gene regulation. *Science* **277**, 1630-1635, doi:10.1126/science.277.5332.1630 (1997).
89 Decker, T., Lew, D. J., Mirkovitch, J. & Darnell, J. E., Jr. Cytoplasmic activation of GAF, an IFN-gamma-regulated DNA-binding factor. *EMBO J* **10**, 927-932, doi:10.1002/j.1460-2075.1991.tb08026.x (1991).
90 Chen, K., Liu, J. & Cao, X. Regulation of type I interferon signaling in immunity and inflammation: A comprehensive review. *J Autoimmun* **83**, 1-11, doi:10.1016/j.jaut.2017.03.008 (2017).
91 Gaestel, M. & Kracht, M. Peptides as signaling inhibitors for mammalian MAP kinase cascades. *Curr Pharm Des* **15**, 2471-2480, doi:10.2174/138161209788682299 (2009).
92 Goh, K. C., Haque, S. J. & Williams, B. R. p38 MAP kinase is required for STAT1 serine phosphorylation and transcriptional activation induced by interferons. *EMBO J* **18**, 5601-5608, doi:10.1093/emboj/18.20.5601 (1999).
93 Li, Y. *et al.* Role of p38alpha Map kinase in Type I interferon signaling. *J Biol Chem* **279**, 970-979, doi:10.1074/jbc.M309927200 (2004).
94 Uddin, S. *et al.* The Rac1/p38 mitogen-activated protein kinase pathway is required for interferon alpha-dependent transcriptional activation but not serine phosphorylation of Stat proteins. *J Biol Chem* **275**, 27634-27640, doi:10.1074/jbc.M003170200 (2000).
95 Hu, J. *et al.* ERK1 and ERK2 activate CCAAAT/enhancer-binding protein-beta-dependent gene transcription in response to interferon-gamma. *J Biol Chem* **276**, 287-297, doi:10.1074/jbc.M004885200 (2001).
96 Nguyen, H., Ramana, C. V., Bayes, J. & Stark, G. R. Roles of phosphatidylinositol 3-kinase in interferon-gamma-dependent phosphorylation of STAT1 on serine 727 and activation of gene expression. *J Biol Chem* **276**, 33361-33368, doi:10.1074/jbc.M105070200 (2001).
97 Uddin, S. *et al.* Protein kinase C-delta (PKC-delta) is activated by type I interferons and mediates phosphorylation of Stat1 on serine 727. *J Biol Chem* **277**, 14408-14416, doi:10.1074/jbc.M109671200 (2002).
98 Wen, Z., Zhong, Z. & Darnell, J. E., Jr. Maximal activation of transcription by Stat1 and Stat3 requires both tyrosine and serine phosphorylation. *Cell* **82**, 241-250, doi:10.1016/0092-8674(95)90311-9 (1995).
99 Lekmine, F. *et al.* Interferon-gamma engages the p70 S6 kinase to regulate phosphorylation of the 40S S6 ribosomal protein. *Exp Cell Res* **295**, 173-182, doi:10.1016/j.yexcr.2003.12.021 (2004).
100 Lekmine, F. *et al.* Activation of the p70 S6 kinase and phosphorylation of the 4E-BP1 repressor of mRNA translation by type I interferons. *J Biol Chem* **278**, 27772-27780, doi:10.1074/jbc.M301364200 (2003).
101 Schneider, W. M., Chevillotte, M. D. & Rice, C. M. Interferon-stimulated genes: a complex web of host defenses. *Annu Rev Immunol* **32**, 513-545, doi:10.1146/annurev-immunol-032713-120231 (2014).
102 Meurs, E. *et al.* Molecular cloning and characterization of the human double-stranded RNA-activated protein kinase induced by interferon. *Cell* **62**, 379-390, doi:10.1016/0092-8674(90)90374-n (1990).

103 Nanduri, S., Carpick, B. W., Yang, Y., Williams, B. R. & Qin, J. Structure of the double-stranded RNA-binding domain of the protein kinase PKR reveals the molecular basis of its dsRNA-mediated activation. *EMBO J* **17**, 5458-5465, doi:10.1093/emboj/17.18.5458 (1998).
104 Dar, A. C., Dever, T. E. & Sicheri, F. Higher-order substrate recognition of eIF2alpha by the RNA-dependent protein kinase PKR. *Cell* **122**, 887-900, doi:10.1016/j.cell.2005.06.044 (2005).
105 Hornung, V., Hartmann, R., Ablasser, A. & Hopfner, K. P. OAS proteins and cGAS: unifying concepts in sensing and responding to cytosolic nucleic acids. *Nat Rev Immunol* **14**, 521-528, doi:10.1038/nri3719 (2014).
106 Donovan, J., Dufner, M. & Korennykh, A. Structural basis for cytosolic double-stranded RNA surveillance by human oligoadenylate synthetase 1. *Proc Natl Acad Sci U S A* **110**, 1652-1657, doi:10.1073/pnas.1218528110 (2013).
107 Silverman, R. H. Viral encounters with 2',5'-oligoadenylate synthetase and RNase L during the interferon antiviral response. *J Virol* **81**, 12720-12729, doi:10.1128/JVI.01471-07 (2007).
108 Floyd-Smith, G., Slattery, E. & Lengyel, P. Interferon action: RNA cleavage pattern of a (2'-5')oligoadenylate--dependent endonuclease. *Science* **212**, 1030-1032, doi:10.1126/science.6165080 (1981).
109 Perng, Y. C. & Lenschow, D. J. ISG15 in antiviral immunity and beyond. *Nat Rev Microbiol* **16**, 423-439, doi:10.1038/s41579-018-0020-5 (2018).
110 Potter, J. L., Narasimhan, J., Mende-Mueller, L. & Haas, A. L. Precursor processing of pro-ISG15/UCRP, an interferon-beta-induced ubiquitin-like protein. *J Biol Chem* **274**, 25061-25068, doi:10.1074/jbc.274.35.25061 (1999).
111 Basters, A. *et al.* Structural basis of the specificity of USP18 toward ISG15. *Nat Struct Mol Biol* **24**, 270-278, doi:10.1038/nsmb.3371 (2017).
112 Zou, W. & Zhang, D. E. The interferon-inducible ubiquitin-protein isopeptide ligase (E3) EFP also functions as an ISG15 E3 ligase. *J Biol Chem* **281**, 3989-3994, doi:10.1074/jbc.M510787200 (2006).
113 Wong, J. J., Pung, Y. F., Sze, N. S. & Chin, K. C. HERC5 is an IFN-induced HECT-type E3 protein ligase that mediates type I IFN-induced ISGylation of protein targets. *Proc Natl Acad Sci U S A* **103**, 10735-10740, doi:10.1073/pnas.0600397103 (2006).
114 Zhao, C. *et al.* The UbcH8 ubiquitin E2 enzyme is also the E2 enzyme for ISG15, an IFN-alpha/beta-induced ubiquitin-like protein. *Proc Natl Acad Sci U S A* **101**, 7578-7582, doi:10.1073/pnas.0402528101 (2004).
115 Yuan, W. & Krug, R. M. Influenza B virus NS1 protein inhibits conjugation of the interferon (IFN)-induced ubiquitin-like ISG15 protein. *EMBO J* **20**, 362-371, doi:10.1093/emboj/20.3.362 (2001).
116 Du, Y. *et al.* LRRC25 inhibits type I IFN signaling by targeting ISG15-associated RIG-I for autophagic degradation. *EMBO J* **37**, 351-366, doi:10.15252/embj.201796781 (2018).
117 Desai, S. D. *et al.* Elevated expression of ISG15 in tumor cells interferes with the ubiquitin/26S proteasome pathway. *Cancer Res* **66**, 921-928, doi:10.1158/0008-5472.CAN-05-1123 (2006).
118 Okumura, A., Pitha, P. M. & Harty, R. N. ISG15 inhibits Ebola VP40 VLP budding in an L-domain-dependent manner by blocking Nedd4 ligase activity. *Proc Natl Acad Sci U S A* **105**, 3974-3979, doi:10.1073/pnas.0710629105 (2008).
119 Tang, Y. *et al.* Herc5 attenuates influenza A virus by catalyzing ISGylation of viral NS1 protein. *J Immunol* **184**, 5777-5790, doi:10.4049/jimmunol.0903588 (2010).
120 Durfee, L. A., Lyon, N., Seo, K. & Huibregtse, J. M. The ISG15 conjugation system broadly targets newly synthesized proteins: implications for the antiviral function of ISG15. *Mol Cell* **38**, 722-732, doi:10.1016/j.molcel.2010.05.002 (2010).

121 Haller, O., Staeheli, P., Schwemmle, M. & Kochs, G. Mx GTPases: dynamin-like antiviral machines of innate immunity. *Trends Microbiol* **23**, 154-163, doi:10.1016/j.tim.2014.12.003 (2015).
122 Kochs, G., Haener, M., Aebi, U. & Haller, O. Self-assembly of human MxA GTPase into highly ordered dynamin-like oligomers. *J Biol Chem* **277**, 14172-14176, doi:10.1074/jbc.M200244200 (2002).
123 Kochs, G. & Haller, O. Interferon-induced human MxA GTPase blocks nuclear import of Thogoto virus nucleocapsids. *Proc Natl Acad Sci U S A* **96**, 2082-2086, doi:10.1073/pnas.96.5.2082 (1999).
124 Staeheli, P. & Pavlovic, J. Inhibition of vesicular stomatitis virus mRNA synthesis by human MxA protein. *J Virol* **65**, 4498-4501, doi:10.1128/JVI.65.8.4498-4501.1991 (1991).
125 Ozato, K., Shin, D. M., Chang, T. H. & Morse, H. C., 3rd. TRIM family proteins and their emerging roles in innate immunity. *Nat Rev Immunol* **8**, 849-860, doi:10.1038/nri2413 (2008).
126 Stremlau, M. *et al.* The cytoplasmic body component TRIM5alpha restricts HIV-1 infection in Old World monkeys. *Nature* **427**, 848-853, doi:10.1038/nature02343 (2004).
127 Harris, R. S. & Dudley, J. P. APOBECs and virus restriction. *Virology* **479-480**, 131-145, doi:10.1016/j.virol.2015.03.012 (2015).
128 Taylor, D. R., Puig, M., Darnell, M. E., Mihalik, K. & Feinstone, S. M. New antiviral pathway that mediates hepatitis C virus replicon interferon sensitivity through ADAR1. *J Virol* **79**, 6291-6298, doi:10.1128/JVI.79.10.6291-6298.2005 (2005).
129 Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198-207, doi:10.1038/nature01511 (2003).
130 Steen, H. & Mann, M. The ABC's (and XYZ's) of peptide sequencing. *Nat Rev Mol Cell Biol* **5**, 699-711, doi:10.1038/nrm1468 (2004).
131 Hillenkamp, F., Karas, M., Beavis, R. C. & Chait, B. T. Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers. *Anal Chem* **63**, 1193A-1203A, doi:10.1021/ac00024a002 (1991).
132 Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F. & Whitehouse, C. M. Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246**, 64-71, doi:10.1126/science.2675315 (1989).
133 Olsen, J. V. *et al.* Higher-energy C-trap dissociation for peptide modification analysis. *Nat Methods* **4**, 709-712, doi:10.1038/nmeth1060 (2007).
134 Wells, J. M. & McLuckey, S. A. Collision-induced dissociation (CID) of peptides and proteins. *Methods Enzymol* **402**, 148-185, doi:10.1016/S0076-6879(05)02005-7 (2005).
135 Zhang, Y., Fonslow, B. R., Shan, B., Baek, M. C. & Yates, J. R., 3rd. Protein analysis by shotgun/bottom-up proteomics. *Chem Rev* **113**, 2343-2394, doi:10.1021/cr3003533 (2013).
136 Cox, J. *et al.* Andromeda: a peptide search engine integrated into the MaxQuant environment. *J Proteome Res* **10**, 1794-1805, doi:10.1021/pr101065j (2011).
137 Zubarev, R. A. The challenge of the proteome dynamic range and its implications for in-depth proteomics. *Proteomics* **13**, 723-726, doi:10.1002/pmic.201200451 (2013).
138 Mitulovic, G. & Mechtler, K. HPLC techniques for proteomics analysis--a short overview of latest developments. *Brief Funct Genomic Proteomic* **5**, 249-260, doi:10.1093/bfgp/ell034 (2006).
139 Shen, Y. *et al.* High-efficiency nanoscale liquid chromatography coupled on-line with mass spectrometry using nanoelectrospray ionization for proteomics. *Anal Chem* **74**, 4235-4249, doi:10.1021/ac0202280 (2002).
140 Ponomarenko, E. A. *et al.* The Size of the Human Proteome: The Width and Depth. *Int J Anal Chem* **2016**, 7436849, doi:10.1155/2016/7436849 (2016).

141 Gilar, M., Olivova, P., Daly, A. E. & Gebler, J. C. Orthogonality of separation in two-dimensional liquid chromatography. *Anal Chem* **77**, 6426-6434, doi:10.1021/ac050923i (2005).
142 Di Palma, S., Hennrich, M. L., Heck, A. J. & Mohammed, S. Recent advances in peptide separation by multidimensional liquid chromatography for proteome analysis. *J Proteomics* **75**, 3791-3813, doi:10.1016/j.jprot.2012.04.033 (2012).
143 Premstaller, A. *et al.* High-performance liquid chromatography-electrospray ionization mass spectrometry using monolithic capillary columns for proteomic studies. *Anal Chem* **73**, 2390-2396, doi:10.1021/ac010046q (2001).
144 Wang, Y. X. *et al.* Reversed-phase chromatography with multiple fraction concatenation strategy for proteome profiling of human MCF10A cells. *Proteomics* **11**, 2019-2026, doi:10.1002/pmic.201000722 (2011).
145 Nagy, P. D. & Pogany, J. The dependence of viral RNA replication on co-opted host factors. *Nat Rev Microbiol* **10**, 137-149, doi:10.1038/nrmicro2692 (2011).
146 Greco, T. M., Diner, B. A. & Cristea, I. M. The Impact of Mass Spectrometry-Based Proteomics on Fundamental Discoveries in Virology. *Annu Rev Virol* **1**, 581-604, doi:10.1146/annurev-virology-031413-085527 (2014).
147 Pichlmair, A. *et al.* Viral immune modulators perturb the human molecular network by common and unique strategies. *Nature* **487**, 486-U101, doi:10.1038/nature11289 (2012).
148 Wojcechowskyj, J. A. *et al.* Quantitative phosphoproteomics reveals extensive cellular reprogramming during HIV-1 entry. *Cell Host Microbe* **13**, 613-623, doi:10.1016/j.chom.2013.04.011 (2013).
149 Miteva, Y. V., Budayeva, H. G. & Cristea, I. M. Proteomics-based methods for discovery, quantification, and validation of protein-protein interactions. *Anal Chem* **85**, 749-768, doi:10.1021/ac3033257 (2013).
150 Makjaroen, J. *et al.* Comprehensive Proteomics Identification of IFN-lambda3-regulated Antiviral Proteins in HBV-transfected Cells. *Mol Cell Proteomics* **17**, 2197-2215, doi:10.1074/mcp.RA118.000735 (2018).
151 Garcia-Moreno, M. *et al.* System-wide Profiling of RNA-Binding Proteins Uncovers Key Regulators of Virus Infection. *Mol Cell* **74**, 196-211 e111, doi:10.1016/j.molcel.2019.01.017 (2019).
152 Bojkova, D. *et al.* Proteomics of SARS-CoV-2-infected host cells reveals therapy targets. *Nature* **583**, 469-472, doi:10.1038/s41586-020-2332-7 (2020).
153 Megger, D. A., Philipp, J., Le-Trilling, V. T. K., Sitek, B. & Trilling, M. Deciphering of the Human Interferon-Regulated Proteome by Mass Spectrometry-Based Quantitative Analysis Reveals Extent and Dynamics of Protein Induction and Repression. *Front Immunol* **8**, 1139, doi:10.3389/fimmu.2017.01139 (2017).
154 Colli, M. L. *et al.* An integrated multi-omics approach identifies the landscape of interferon-alpha-mediated responses of human pancreatic beta cells. *Nat Commun* **11**, 2584, doi:10.1038/s41467-020-16327-0 (2020).
155 Yan, W. *et al.* System-based proteomic analysis of the interferon response in human liver cells. *Genome Biol* **5**, doi:ARTN R54

DOI 10.1186/gb-2004-5-8-r54 (2004).
156 Schuhenn, J. *et al.* Differential interferon-alpha subtype induced immune signatures are associated with suppression of SARS-CoV-2 infection. *Proc Natl Acad Sci U S A* **119**, doi:10.1073/pnas.2111600119 (2022).
157 Trost, M. *et al.* The phagosomal proteome in interferon-gamma-activated macrophages. *Immunity* **30**, 143-154, doi:10.1016/j.immuni.2008.11.006 (2009).
158 Fila, J. & Honys, D. Enrichment techniques employed in phosphoproteomics. *Amino Acids* **43**, 1025-1047, doi:10.1007/s00726-011-1111-z (2012).

159 Fulzele, A. & Bennett, E. J. Ubiquitin diGLY Proteomics as an Approach to Identify and Quantify the Ubiquitin-Modified Proteome. *Methods Mol Biol* **1844**, 363-384, doi:10.1007/978-1-4939-8706-1_23 (2018).
160 Varjosalo, M. *et al.* Interlaboratory reproducibility of large-scale human protein-complex analysis by standardized AP-MS. *Nat Methods* **10**, 307-314, doi:10.1038/nmeth.2400 (2013).
161 Jager, S. *et al.* Global landscape of HIV-human protein complexes. *Nature* **481**, 365-370, doi:10.1038/nature10719 (2012).
162 Wang, L. *et al.* Comparative influenza protein interactomes identify the role of plakophilin 2 in virus restriction. *Nat Commun* **8**, 13876, doi:10.1038/ncomms13876 (2017).
163 Kang, S. M., Shin, M. J., Kim, J. H. & Oh, J. W. Proteomic profiling of cellular proteins interacting with the hepatitis C virus core protein. *Proteomics* **5**, 2227-2237, doi:10.1002/pmic.200401093 (2005).
164 Shah, P. S. *et al.* Comparative Flavivirus-Host Protein Interaction Mapping Reveals Mechanisms of Dengue and Zika Virus Pathogenesis. *Cell* **175**, 1931-+, doi:10.1016/j.cell.2018.11.028 (2018).
165 Stukalov, A. *et al.* Multilevel proteomics reveals host perturbations by SARS-CoV-2 and SARS-CoV. *Nature* **594**, 246-252, doi:10.1038/s41586-021-03493-4 (2021).
166 Hubel, P. *et al.* A protein-interaction network of interferon-stimulated genes extends the innate immune system landscape. *Nat Immunol* **20**, 493-502, doi:10.1038/s41590-019-0323-3 (2019).
167 Luo, Y. *et al.* HIV-host interactome revealed directly from infected cells. *Nat Microbiol* **1**, 16068, doi:10.1038/nmicrobiol.2016.68 (2016).
168 Garcia-Moreno, M., Jarvelin, A. I. & Castello, A. Unconventional RNA-binding proteins step into the virus-host battlefront. *Wiley Interdiscip Rev RNA* **9**, e1498, doi:10.1002/wrna.1498 (2018).
169 Iselin, L. *et al.* Uncovering viral RNA-host cell interactions on a proteome-wide scale. *Trends Biochem Sci* **47**, 23-38, doi:10.1016/j.tibs.2021.08.002 (2022).
170 Kamel, W. *et al.* Global analysis of protein-RNA interactions in SARS-CoV-2-infected cells reveals key regulators of infection. *Mol Cell* **81**, 2851-2867 e2857, doi:10.1016/j.molcel.2021.05.023 (2021).
171 Graham, F. L., Smiley, J., Russell, W. C. & Nairn, R. Characteristics of a Human Cell Line Transformed by DNA from Human Adenovirus Type-5. *J Gen Virol* **36**, 59-72, doi:Doi 10.1099/0022-1317-36-1-59 (1977).
172 Louis, N., Evelegh, C. & Graham, F. L. Cloning and sequencing of the cellular-viral junctions from the human adenovirus type 5 transformed 293 cell line. *Virology* **233**, 423-429, doi:DOI 10.1006/viro.1997.8597 (1997).
173 Sha, J. F., Ghosh, M. K., Zhang, K. M. & Harter, M. L. E1A Interacts with Two Opposing Transcriptional Pathways To Induce Quiescent Cells into S Phase. *J Virol* **84**, 4050-4059, doi:10.1128/Jvi.02131-09 (2010).
174 Hidalgo, P., Ip, W. H., Dobner, T. & Gonzalez, R. A. The biology of the adenovirus E1B 55K protein. *Febs Lett* **593**, 3504-3517, doi:10.1002/1873-3468.13694 (2019).
175 Petiot, E., Cuperlovic-Culf, M., Shen, C. F. & Kamen, A. Influence of HEK293 metabolism on the production of viral vectors and vaccine. *Vaccine* **33**, 5974-5981, doi:10.1016/j.vaccine.2015.05.097 (2015).
176 Nadeau, I. & Kamen, A. Production of adenovirus vector for gene therapy. *Biotechnol Adv* **20**, 475-489, doi:Pii S0734-9750(02)00030-7

Doi 10.1016/S0734-9750(02)00030-7 (2003).
177 Fontana, D., Kratje, R., Etcheverrigaray, M. & Prieto, C. Rabies virus-like particles expressed in HEK293 cells. *Vaccine* **32**, 2799-2804, doi:10.1016/j.vaccine.2014.02.031 (2014).
178 Emma, P. & Kamen, A. Real-time monitoring of influenza virus production kinetics in HEK293 cell cultures. *Biotechnol Progr* **29**, 275-284, doi:10.1002/btpr.1601 (2013).

179 Eddy, B. E., Borman, G. S., Grubbs, G. E. & Young, R. D. Identification of the oncogenic substance in rhesus monkey kidney cell cultures as simian virus 40. *Virology* **17**, 65-75 (1962).
180 An, P., Robles, M. T. S. & Pipas, J. M. Large T Antigens of Polyomaviruses: Amazing Molecular Machines. *Annu Rev Microbiol* **66**, 213-236, doi:10.1146/annurev-micro-092611-150154 (2012).
181 Ahuja, D., Saenz-Robles, M. T. & Pipas, J. M. SV40 large T antigen targets multiple cellular pathways to elicit cellular transformation. *Oncogene* **24**, 7729-7745, doi:DOI 10.1038/sj.onc.1209046 (2005).
182 Fine, D. A. *et al.* Identification of FAM111A as an SV40 Host Range Restriction and Adenovirus Helper Factor. *Plos Pathog* **8**, doi:ARTN e1002949

10.1371/journal.ppat.1002949 (2012).
183 Kierstead, T. D. & Tevethia, M. J. Association of P53 Binding and Immortalization of Primary C57bl/6 Mouse Embryo Fibroblasts by Using Simian-Virus 40 T-Antigen Mutants Bearing Internal Overlapping Deletion Mutations. *J Virol* **67**, 1817-1829, doi:Doi 10.1128/Jvi.67.4.1817-1829.1993 (1993).
184 Dyson, N. *et al.* Large T-Antigens of Many Polyomaviruses Are Able to Form Complexes with the Retinoblastoma Protein. *J Virol* **64**, 1353-1356, doi:Doi 10.1128/Jvi.64.3.1353-1356.1990 (1990).
185 Eischen, C. M. Genome Stability Requires p53. *Csh Perspect Med* **6**, doi:ARTN a026096

10.1101/cshperspect.a026096 (2016).
186 Lin, Y. C. *et al.* Genome dynamics of the human embryonic kidney 293 lineage in response to cell biology manipulations. *Nat Commun* **5**, doi:ARTN 4767

10.1038/ncomms5767 (2014).
187 Merten, O. W. *et al.* Large-Scale Manufacture and Characterization of a Lentiviral Vector Produced for Clinical Ex Vivo Gene Therapy Application. *Hum Gene Ther* **22**, 343-356, doi:10.1089/hum.2010.060 (2011).
188 Li, M. Q. *et al.* Codon-usage-based inhibition of HIV protein synthesis by human schlafen 11. *Nature* **491**, 125-U145, doi:10.1038/nature11433 (2012).
189 Bae, D. H. *et al.* Design and Testing of Vector-Producing HEK293T Cells Bearing a Genomic Deletion of the SV40 T Antigen Coding Region. *Mol Ther-Meth Clin D* **18**, 631-638, doi:10.1016/j.omtm.2020.07.006 (2020).
190 Rathi, A. V., Cantalupo, P. G., Sarkar, S. N. & Pipas, J. M. Induction of interferon-stimulated genes by Simian virus 40 T antigens. *Virology* **406**, 202-211, doi:10.1016/j.virol.2010.07.018 (2010).
191 Forero, A. *et al.* Simian Virus 40 Large T Antigen Induces IFN-Stimulated Genes through ATR Kinase. *J Immunol* **192**, 5933-5942, doi:10.4049/jimmunol.1303470 (2014).
192 Lau, L., Gray, E. E., Brunette, R. L. & Stetson, D. B. DNA tumor virus oncogenes antagonize the cGAS-STING DNA-sensing pathway. *Science* **350**, 568-571, doi:10.1126/science.aab3291 (2015).
193 Swaminathan, S., Rajan, P., Savinova, O., Jagus, R. & Thimmapaya, B. Simian virus 40 large-T bypasses the translational block imposed by the phosphorylation of eIF-2 alpha. *Virology* **219**, 321-323, doi:DOI 10.1006/viro.1996.0255 (1996).
194 Reus, J. B., Trivino-Soto, G. S., Wu, L. I., Kokott, K. & Lim, E. S. SV40 large T antigen is not responsible for the loss of STING in 293T cells but can inhibit cGAS-STING interferon induction. *Viruses* **12**, 137 (2020).
195 Ferreira, C. B. *et al.* Lentiviral vector production titer is not limited in HEK293T by induced intracellular innate immunity. *Mol Ther-Meth Clin D* **17**, 209-219 (2020).

196 He, J. *et al.* Human immunodeficiency virus type 1 viral protein R (Vpr) arrests cells in the G2 phase of the cell cycle by inhibiting p34cdc2 activity. *J Virol* **69**, 6705-6711, doi:10.1128/JVI.69.11.6705-6711.1995 (1995).
197 Muller, B. *et al.* Construction and characterization of a fluorescently labeled infectious human immunodeficiency virus type 1 derivative. *J Virol* **78**, 10803-10813, doi:10.1128/Jvi.78.19.10803-10813.2004 (2004).
198 Bartz, S. R. & Vodicka, M. A. Production of high-titer human immunodeficiency virus type 1 pseudotyped with vesicular stomatitis virus glycoprotein. *Methods* **12**, 337-342, doi:DOI 10.1006/meth.1997.0487 (1997).
199 Finkelshtein, D., Werman, A., Novick, D., Barak, S. & Rubinstein, M. LDL receptor and its family members serve as the cellular receptors for vesicular stomatitis virus. *P Natl Acad Sci USA* **110**, 7306-7311, doi:10.1073/pnas.1214441110 (2013).
200 Belot, L., Albertini, A. & Gaudin, Y. Structural and cellular biology of rhabdovirus entry. *Adv Virus Res* **104**, 147-183, doi:10.1016/bs.aivir.2019.05.003 (2019).
201 Edmonds, T. G. *et al.* Replication competent molecular clones of HIV-1 expressing Renilla luciferase facilitate the analysis of antibody inhibition in PBMC. *Virology* **408**, 1-13, doi:10.1016/j.virol.2010.08.028 (2010).
202 Ong, S. E. *et al.* Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics* **1**, 376-386, doi:10.1074/mcp.M200025-MCP200 (2002).
203 Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, doi:ARTN e47

10.1093/nar/gkv007 (2015).
204 Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *P Natl Acad Sci USA* **102**, 15545-15550, doi:10.1073/pnas.0506580102 (2005).
205 Stark, G. R. & Darnell, J. E., Jr. The JAK-STAT pathway at twenty. *Immunity* **36**, 503-514, doi:10.1016/j.immuni.2012.03.013 (2012).
206 Gordon, D. E. *et al.* A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* **583**, 459-468, doi:10.1038/s41586-020-2286-9 (2020).
207 Balistreri, G. *et al.* The host nonsense-mediated mRNA decay pathway restricts Mammalian RNA virus replication. *Cell Host Microbe* **16**, 403-411, doi:10.1016/j.chom.2014.08.007 (2014).
208 Harbour, J. W., Luo, R. X., Dei Santi, A., Postigo, A. A. & Dean, D. C. Cdk phosphorylation triggers sequential intramolecular interactions that progressively block Rb functions as cells move through G1. *Cell* **98**, 859-869, doi:10.1016/s0092-8674(00)81519-6 (1999).
209 Chepenik, L. G., Tretiakova, A. P., Krachmarov, C. P., Johnson, E. M. & Khalili, K. The single-stranded DNA binding protein, Pur-alpha, binds HIV-1 TAR RNA and activates HIV-1 transcription. *Gene* **210**, 37-44, doi:10.1016/s0378-1119(98)00033-x (1998).
210 Kutner, R. H., Zhang, X. Y. & Reiser, J. Production, concentration and titration of pseudotyped HIV-1-based lentiviral vectors. *Nat Protoc* **4**, 495-505, doi:10.1038/nprot.2009.22 (2009).
211 Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **26**, 1367-1372, doi:10.1038/nbt.1511 (2008).
212 Tyanova, S. *et al.* The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat Methods* **13**, 731-740, doi:10.1038/nmeth.3901 (2016).
213 Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A. & Vingron, M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18 Suppl 1**, S96-104, doi:10.1093/bioinformatics/18.suppl_1.s96 (2002).

214 Lazar, C. imputeLCMD: a collection of methods for left-censored missing data imputation. *R package, version* **2** (2015).
215 Szklarczyk, D. *et al.* STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* **47**, D607-D613, doi:10.1093/nar/gky1131 (2019).
216 Secombes, C. J. & Zou, J. Evolution of Interferons and Interferon Receptors. *Front Immunol* **8**, 209, doi:10.3389/fimmu.2017.00209 (2017).
217 Pestka, S., Krause, C. D. & Walter, M. R. Interferons, interferon-like cytokines, and their receptors. *Immunol Rev* **202**, 8-32, doi:DOI 10.1111/j.0105-2896.2004.00204.x (2004).
218 Kotenko, S. V. *et al.* IFN-lambda s mediate antiviral protection through a distinct class II cytokine receptor complex. *Nat Immunol* **4**, 69-77, doi:10.1038/ni875 (2003).
219 Iversen, M. B. & Paludan, S. R. Mechanisms of Type III Interferon Expression. *J Interf Cytok Res* **30**, 573-578, doi:10.1089/jir.2010.0063 (2010).
220 Valente, G. *et al.* Distribution of Interferon-Gamma Receptor in Human Tissues. *Eur J Immunol* **22**, 2403-2412, doi:DOI 10.1002/eji.1830220933 (1992).
221 Russell-Harde, D., Wagner, T. C., Perez, H. D. & Croze, E. Formation of a uniquely stable type I interferon receptor complex by interferon beta is dependent upon particular interactions between interferon beta and its receptor and independent of tyrosine phosphorylation. *Biochem Bioph Res Co* **255**, 539-544, doi:DOI 10.1006/bbrc.1998.0105 (1999).
222 Diaz, M. O. *et al.* Structure of the Human Type-I Interferon Gene-Cluster Determined from a Yac Clone Contig. *Genomics* **22**, 540-552, doi:DOI 10.1006/geno.1994.1427 (1994).
223 Lavoie, T. B. *et al.* Binding and activity of all human alpha interferon subtypes. *Cytokine* **56**, 282-289, doi:10.1016/j.cyto.2011.07.019 (2011).
224 Schreiber, G. The molecular basis for differential type I interferon signaling. *J Biol Chem* **292**, 7285-7294, doi:10.1074/jbc.R116.774562 (2017).
225 Jaitin, D. A. *et al.* Inquiring into the differential action of interferons (IFNs): an IFN-alpha 2 mutant with enhanced affinity to IFNAR1 is functionally similar to IFN-beta. *Mol Cell Biol* **26**, 1888-1897, doi:10.1128/Mcb.26.5.1888-1897.2006 (2006).
226 Isaacs, A. & Lindenmann, J. Virus interference. I. The interferon. *Proceedings of the Royal Society of London. Series B-Biological Sciences* **147**, 258-267 (1957).
227 Lengyel, P. Biochemistry of Interferons and Their Actions. *Annu Rev Biochem* **51**, 251-282, doi:DOI 10.1146/annurev.bi.51.070182.001343 (1982).
228 Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative Monitoring of Gene-Expression Patterns with a Complementary-DNA Microarray. *Science* **270**, 467-470, doi:DOI 10.1126/science.270.5235.467 (1995).
229 Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 621-628, doi:10.1038/nmeth.1226 (2008).
230 Der, S. D., Zhou, A. M., Williams, B. R. G. & Silverman, R. H. Identification of genes differentially regulated by interferon alpha, beta, or gamma using oligonucleotide arrays. *P Natl Acad Sci USA* **95**, 15623-15628, doi:DOI 10.1073/pnas.95.26.15623 (1998).
231 Samarajiwa, S. A., Forster, S., Auchettl, K. & Hertzog, P. J. INTERFEROME: the database of interferon regulated genes. *Nucleic Acids Res* **37**, D852-D857, doi:10.1093/nar/gkn732 (2009).
232 Rusinova, I. *et al.* INTERFEROME v2.0: an updated database of annotated interferon-regulated genes. *Nucleic Acids Res* **41**, D1040-D1046, doi:10.1093/nar/gks1215 (2013).
233 Schwanhausser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337-342, doi:10.1038/nature10098 (2011).
234 Buccitelli, C. & Selbach, M. mRNAs, proteins and the emerging principles of gene expression control. *Nat Rev Genet* **21**, 630-644, doi:10.1038/s41576-020-0258-4 (2020).

235 Ternette, N., Wright, C., Kramer, H. B., Altun, M. & Kessler, B. M. Label-free quantitative proteomics reveals regulation of interferon-induced protein with tetratricopeptide repeats 3 (IFIT3) and 5 '-3 '-exoribonuclease 2 (XRN2) during respiratory syncytial virus infection. *Virol J* **8**, doi:Artn 442

10.1186/1743-422x-8-442 (2011).
236 Megger, D. A., Philipp, J., Khanh, L. T. V. T., Sitek, B. & Trilling, M. Deciphering of the human interferon-regulated Proteome by Mass spectrometry-Based Quantitative analysis reveals extent and Dynamics of Protein induction and repression. *Frontiers in Immunology* **8**, doi:ARTN 1139

10.3389/fimmu.2017.01139 (2017).
237 Tarasova, I. A. *et al.* Comparative proteomics as a tool for identifying specific alterations within interferon response pathways in human glioblastoma multiforme cells. *Oncotarget* **9**, 1785-1802, doi:10.18632/oncotarget.22751 (2018).
238 Wobma, H. M. *et al.* The influence of hypoxia and IFN-gamma on the proteome and metabolome of therapeutic mesenchymal stem cells. *Biomaterials* **167**, 226-234, doi:10.1016/j.biomaterials.2018.03.027 (2018).
239 Wang, Y. *et al.* Reversed-phase chromatography with multiple fraction concatenation strategy for proteome profiling of human MCF10A cells. *Proteomics* **11**, 2019-2026, doi:10.1002/pmic.201000722 (2011).
240 Hodge, K., Makjaroen, J., Robinson, J., Khoomrung, S. & Pisitkun, T. Deep Proteomic Deconvolution of Interferons and HBV Transfection Effects on a Hepatoblastoma Cell Line. *ACS Omega* **5**, 16796-16810, doi:10.1021/acsomega.0c01865 (2020).
241 Zhao, M. Z. *et al.* Comparative Transcriptomic and Proteomic Analyses Prove that IFN-lambda 1 is a More Potent Inducer of ISGs than IFN-alpha against Porcine Epidemic Diarrhea Virus in Porcine Intestinal Epithelial Cells. *J Proteome Res* **19**, 3697-3707, doi:10.1021/acs.jproteome.0c00164 (2020).
242 Colli, M. L. *et al.* An integrated multi-omics approach identifies the landscape of interferon-alpha-mediated responses of human pancreatic beta cells. *Nat Commun* **11**, doi:ARTN 2584

10.1038/s41467-020-16327-0 (2020).
243 Gibbert, K., Schlaak, J. F., Yang, D. & Dittmer, U. IFN-alpha subtypes: distinct biological activities in anti-viral therapy. *Br J Pharmacol* **168**, 1048-1058, doi:10.1111/bph.12010 (2013).
244 van Boxel-Dezaire, A. H., Rani, M. R. & Stark, G. R. Complex modulation of cell type-specific signaling in response to type I interferons. *Immunity* **25**, 361-372, doi:10.1016/j.immuni.2006.08.014 (2006).
245 Gupta, S., Jiang, M. & Pernis, A. B. IFN-alpha activates Stat6 and leads to the formation of Stat2:Stat6 complexes in B cells. *J Immunol* **163**, 3834-3841 (1999).
246 Ho, H. H. & Ivashkiv, L. B. Role of STAT3 in type I interferon responses. Negative regulation of STAT1-dependent inflammatory gene activation. *J Biol Chem* **281**, 14111-14118, doi:10.1074/jbc.M511797200 (2006).
247 Gomez, D. & Reich, N. C. Stimulation of primary human endothelial cell proliferation by IFN. *J Immunol* **170**, 5373-5381, doi:10.4049/jimmunol.170.11.5373 (2003).
248 Bekisz, J., Baron, S., Balinsky, C., Morrow, A. & Zoon, K. C. Antiproliferative Properties of Type I and Type II Interferon. *Pharmaceuticals (Basel)* **3**, 994-1015, doi:10.3390/ph3040994 (2010).
249 Smyth, G. K. in *Bioinformatics and computational biology solutions using R and Bioconductor* 397-420 (Springer, 2005).
250 Bates, D., Machler, M., Bolker, B. M. & Walker, S. C. Fitting Linear Mixed-Effects Models Using lme4. *J Stat Softw* **67**, 1-48, doi:DOI 10.18637/jss.v067.i01 (2015).

251 Smyth, G. K., Michaud, J. & Scott, H. S. Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics* **21**, 2067-2075, doi:10.1093/bioinformatics/bti270 (2005).
252 The Gene Ontology, C. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res* **47**, D330-D338, doi:10.1093/nar/gky1055 (2019).
253 Brasier, A. R. The NF-kappaB regulatory network. *Cardiovasc Toxicol* **6**, 111-130, doi:10.1385/ct:6:2:111 (2006).
254 Iwasaki, A. & Medzhitov, R. Control of adaptive immunity by the innate immune system. *Nat Immunol* **16**, 343-353, doi:10.1038/ni.3123 (2015).
255 Jinushi, M. *et al.* Critical role of MHC class I-related chain A and B expression on IFN-alpha-stimulated dendritic cells in NK cell activation: impairment in chronic hepatitis C virus infection. *J Immunol* **170**, 1249-1256, doi:10.4049/jimmunol.170.3.1249 (2003).
256 Aki, M. *et al.* Interferon-gamma induces different subunit organizations and functional diversity of proteasomes. *J Biochem* **115**, 257-269, doi:10.1093/oxfordjournals.jbchem.a124327 (1994).
257 Oberg, K. The action of interferon alpha on human carcinoid tumours. *Semin Cancer Biol* **3**, 35-41 (1992).
258 Coomans de Brachene, A. *et al.* IFN-alpha induces a preferential long-lasting expression of MHC class I in human pancreatic beta cells. *Diabetologia* **61**, 636-640, doi:10.1007/s00125-017-4536-4 (2018).
259 Fehr, A. R. *et al.* The impact of PARPs and ADP-ribosylation on inflammation and host-pathogen interactions. *Gene Dev* **34**, 341-359, doi:10.1101/gad.334425.119 (2020).
260 Yamada, T. *et al.* Constitutive aryl hydrocarbon receptor signaling constrains type I interferon-mediated antiviral innate defense. *Nat Immunol* **17**, 687-694, doi:10.1038/ni.3422 (2016).
261 Juszczynski, P. *et al.* BAL1 and BBAP are regulated by a gamma interferon-responsive bidirectional promoter and are overexpressed in diffuse large B-cell lymphomas with a prominent inflammatory infiltrate. *Mol Cell Biol* **26**, 5348-5359, doi:10.1128/MCB.02351-05 (2006).
262 Caprara, G. *et al.* PARP14 Controls the Nuclear Accumulation of a Subset of Type I IFN-Inducible Proteins. *J Immunol* **200**, 2439-2454, doi:10.4049/jimmunol.1701117 (2018).
263 Atasheva, S., Frolova, E. I. & Frolov, I. Interferon-stimulated poly(ADP-Ribose) polymerases are potent inhibitors of cellular translation and virus replication. *J Virol* **88**, 2116-2130, doi:10.1128/JVI.03443-13 (2014).
264 Kassambara, A. & Mundt, F. Factoextra: extract and visualize the results of multivariate data analyses. *R package version* **1**, 337-354 (2017).
265 Xu, G. *et al.* Inducible LGALS3BP/90K activates antiviral innate immune responses by targeting TRAF6 and TRAF3 complex. *Plos Pathog* **15**, e1008002, doi:10.1371/journal.ppat.1008002 (2019).
266 Park, S. S. *et al.* Effective correction of experimental errors in quantitative proteomics using stable isotope labeling by amino acids in cell culture (SILAC). *J Proteomics* **75**, 3720-3732, doi:10.1016/j.jprot.2012.04.035 (2012).
267 Mann, M. Functional and quantitative proteomics using SILAC. *Nat Rev Mol Cell Biol* **7**, 952-958, doi:10.1038/nrm2067 (2006).
268 Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* **42**, 348-354, doi:10.1038/ng.548 (2010).
269 Listgarten, J., Kadie, C., Schadt, E. E. & Heckerman, D. Correction for hidden confounders in the genetic analysis of gene expression. *Proc Natl Acad Sci U S A* **107**, 16465-16470, doi:10.1073/pnas.1002425107 (2010).

270 Yu, L. B., Zhang, J. Y., Brock, G. N. & Fernandez, S. Fully moderated t-statistic in linear modeling of mixed effects for differential expression analysis. *Bmc Bioinformatics* **20**, doi:ARTN 675

10.1186/s12859-019-3248-9 (2019).
271 Shapira, S. D. *et al.* A physical and regulatory map of host-influenza interactions reveals pathways in H1N1 infection. *Cell* **139**, 1255-1267, doi:10.1016/j.cell.2009.12.018 (2009).
272 Liu, Y., Beyer, A. & Aebersold, R. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* **165**, 535-550, doi:10.1016/j.cell.2016.03.014 (2016).
273 Schwanhausser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337-342, doi:10.1038/nature10098 (2011).
274 Needham, E. J., Parker, B. L., Burykin, T., James, D. E. & Humphrey, S. J. Illuminating the dark phosphoproteome. *Sci Signal* **12**, doi:10.1126/scisignal.aau8645 (2019).
275 Davis, M. E. & Gack, M. U. Ubiquitination in the antiviral immune response. *Virology* **479-480**, 52-65, doi:10.1016/j.virol.2015.02.033 (2015).
276 Zhao, M. *et al.* Comparative Transcriptomic and Proteomic Analyses Prove that IFN-lambda1 is a More Potent Inducer of ISGs than IFN-alpha against Porcine Epidemic Diarrhea Virus in Porcine Intestinal Epithelial Cells. *J Proteome Res* **19**, 3697-3707, doi:10.1021/acs.jproteome.0c00164 (2020).
277 Li, J. *et al.* TMTpro reagents: a set of isobaric labeling mass tags enables simultaneous proteome-wide measurements across 16 samples. *Nat Methods* **17**, 399-404, doi:10.1038/s41592-020-0781-4 (2020).
278 Altelaar, A. F. *et al.* Benchmarking stable isotope labeling based quantitative proteomics. *J Proteomics* **88**, 14-26, doi:10.1016/j.jprot.2012.10.009 (2013).
279 Dukhovny, A. *et al.* A CRISPR Activation Screen Identifies Genes That Protect against Zika Virus Infection. *J Virol* **93**, doi:10.1128/JVI.00211-19 (2019).
280 Fusco, D. N. *et al.* HELZ2 Is an IFN Effector Mediating Suppression of Dengue Virus. *Front Microbiol* **8**, 240, doi:10.3389/fmicb.2017.00240 (2017).
281 El-Diwany, R. *et al.* CMPK2 and BCL-G are associated with type 1 interferon-induced HIV restriction in humans. *Sci Adv* **4**, eaat0843, doi:10.1126/sciadv.aat0843 (2018).
282 Lai, J. H. *et al.* Mitochondrial CMPK2 mediates immunomodulatory and antiviral activities through IFN-dependent and IFN-independent pathways. *iScience* **24**, 102498, doi:10.1016/j.isci.2021.102498 (2021).
283 Oudshoorn, D. *et al.* HERC6 is the main E3 ligase for global ISG15 conjugation in mouse cells. *PLoS One* **7**, e29870, doi:10.1371/journal.pone.0029870 (2012).
284 Kim, Y. H., Lee, J. R. & Hahn, M. J. Regulation of inflammatory gene expression in macrophages by epithelial-stromal interaction 1 (Epsti1). *Biochem Biophys Res Commun* **496**, 778-783, doi:10.1016/j.bbrc.2017.12.014 (2018).
285 Trilling, M. *et al.* Deciphering the modulation of gene expression by type I and II interferons combining 4sU-tagging, translational arrest and in silico promoter analysis. *Nucleic Acids Res* **41**, 8107-8125, doi:10.1093/nar/gkt589 (2013).
286 Shaw, A. E. *et al.* Fundamental properties of the mammalian innate immune system revealed by multispecies comparison of type I interferon responses. *PLoS Biol* **15**, e2004086, doi:10.1371/journal.pbio.2004086 (2017).
287 Henig, N. *et al.* Interferon-beta induces distinct gene expression response patterns in human monocytes versus T cells. *PLoS One* **8**, e62366, doi:10.1371/journal.pone.0062366 (2013).
288 Gebauer, F. & Hentze, M. W. Molecular mechanisms of translational control. *Nat Rev Mol Cell Biol* **5**, 827-835, doi:10.1038/nrm1488 (2004).
289 Hochstrasser, M. Ubiquitin-dependent protein degradation. *Annu Rev Genet* **30**, 405-439, doi:10.1146/annurev.genet.30.1.405 (1996).
290 Iwata, H. *et al.* PARP9 and PARP14 cross-regulate macrophage activation via STAT1 ADP-ribosylation. *Nat Commun* **7**, 12849, doi:10.1038/ncomms12849 (2016).

291 GO.db: A set of annotation maps describing the entire Gene Ontology (2020).
292 Steffen Durinck, P. T. S., Ewan Birney and Wolfgang Huber. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols* **4**, 7 (2009).
293 RISmed: Download Content from NCBI Databases (2021).
294 textstem: Tools for stemming and lemmatizing text (2018).
295 Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet* **13**, 227-232, doi:10.1038/nrg3185 (2012).
296 Wilhelm, M. *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* **509**, 582-587, doi:10.1038/nature13319 (2014).
297 Edfors, F. *et al.* Gene-specific correlation of RNA and protein levels in human cells and tissues. *Mol Syst Biol* **12**, 883, doi:10.15252/msb.20167144 (2016).
298 Buccitelli, C. & Selbach, M. mRNAs, proteins and the emerging principles of gene expression control. *Nat Rev Genet* **21**, 630-644, doi:10.1038/s41576-020-0258-4 (2020).
299 Fortelny, N., Overall, C. M., Pavlidis, P. & Freue, G. V. C. Can we predict protein from mRNA levels? *Nature* **547**, E19-E20, doi:10.1038/nature22293 (2017).
300 Franks, A., Airoldi, E. & Slavov, N. Post-transcriptional regulation across human tissues. *PLoS Comput Biol* **13**, e1005535, doi:10.1371/journal.pcbi.1005535 (2017).
301 Eraslan, B. *et al.* Quantification and discovery of sequence determinants of protein-per-mRNA amount in 29 human tissues. *Mol Syst Biol* **15**, e8513, doi:10.15252/msb.20188513 (2019).
302 Jackson, R. J., Hellen, C. U. & Pestova, T. V. The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat Rev Mol Cell Biol* **11**, 113-127, doi:10.1038/nrm2838 (2010).
303 Jariel-Encontre, I., Bossis, G. & Piechaczyk, M. Ubiquitin-independent degradation of proteins by the proteasome. *Biochim Biophys Acta* **1786**, 153-177, doi:10.1016/j.bbcan.2008.05.004 (2008).
304 Bartel, D. P. MicroRNAs: target recognition and regulatory functions. *Cell* **136**, 215-233, doi:10.1016/j.cell.2009.01.002 (2009).
305 Guo, H., Ingolia, N. T., Weissman, J. S. & Bartel, D. P. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* **466**, 835-840, doi:10.1038/nature09267 (2010).
306 Djuranovic, S., Nahvi, A. & Green, R. miRNA-mediated gene silencing by translational repression followed by mRNA deadenylation and decay. *Science* **336**, 237-240, doi:10.1126/science.1215691 (2012).
307 Wilczynska, A. & Bushell, M. The complexity of miRNA-mediated repression. *Cell Death Differ* **22**, 22-33, doi:10.1038/cdd.2014.112 (2015).
308 Kundu, P., Fabian, M. R., Sonenberg, N., Bhattacharyya, S. N. & Filipowicz, W. HuR protein attenuates miRNA-mediated repression by promoting miRISC dissociation from the target RNA. *Nucleic Acids Res* **40**, 5088-5100, doi:10.1093/nar/gks148 (2012).
309 Wharton, R. P., Sonoda, J., Lee, T., Patterson, M. & Murata, Y. The Pumilio RNA-binding domain is also a translational regulator. *Mol Cell* **1**, 863-872, doi:10.1016/s1097-2765(00)80085-4 (1998).
310 Macejak, D. G. & Sarnow, P. Internal initiation of translation mediated by the 5' leader of a cellular mRNA. *Nature* **353**, 90-94, doi:10.1038/353090a0 (1991).
311 Babendure, J. R., Babendure, J. L., Ding, J. H. & Tsien, R. Y. Control of mammalian translation by mRNA structure near caps. *RNA* **12**, 851-861, doi:10.1261/rna.2309906 (2006).
312 Quax, T. E., Claassens, N. J., Soll, D. & van der Oost, J. Codon Bias as a Means to Fine-Tune Gene Expression. *Mol Cell* **59**, 149-161, doi:10.1016/j.molcel.2015.05.035 (2015).
313 Berg, O. G. & Kurland, C. G. Growth rate-optimised tRNA abundance and codon usage. *J Mol Biol* **270**, 544-550, doi:10.1006/jmbi.1997.1142 (1997).

314 Gustafsson, C. *et al.* Engineering genes for predictable protein expression. *Protein Expr Purif* **83**, 37-46, doi:10.1016/j.pep.2012.02.013 (2012).
315 Tao, T. *et al.* Def defines a conserved nucleolar pathway that leads p53 to proteasome-independent degradation. *Cell Res* **23**, 620-634, doi:10.1038/cr.2013.16 (2013).
316 Battle, A. *et al.* Genomic variation. Impact of regulatory variation from RNA to protein. *Science* **347**, 664-667, doi:10.1126/science.1260793 (2015).
317 Bange, F. C., Flohr, T., Buwitt, U. & Bottger, E. C. An interferon-induced protein with release factor activity is a tryptophanyl-tRNA synthetase. *Febs Lett* **300**, 162-166, doi:10.1016/0014-5793(92)80187-l (1992).
318 Ahn, Y. H. *et al.* Secreted tryptophanyl-tRNA synthetase as a primary defence system against infection. *Nat Microbiol* **2**, 16191, doi:10.1038/nmicrobiol.2016.191 (2016).
319 Blees, A. *et al.* Structure of the human MHC-I peptide-loading complex. *Nature* **551**, 525-528, doi:10.1038/nature24627 (2017).
320 Huitema, L. F. *et al.* Entpd5 is essential for skeletal mineralization and regulates phosphate homeostasis in zebrafish. *Proc Natl Acad Sci U S A* **109**, 21372-21377, doi:10.1073/pnas.1214231110 (2012).
321 Zheng, B., Berrie, C. P., Corda, D. & Farquhar, M. G. GDE1/MIR16 is a glycerophosphoinositol phosphodiesterase regulated by stimulation of G protein-coupled receptors. *Proc Natl Acad Sci U S A* **100**, 1745-1750, doi:10.1073/pnas.0337605100 (2003).
322 Davydova, E. *et al.* The methyltransferase METTL9 mediates pervasive 1-methylhistidine modification in mammalian proteomes. *Nat Commun* **12**, 891, doi:10.1038/s41467-020-20670-7 (2021).
323 Hentze, M. W., Castello, A., Schwarzl, T. & Preiss, T. A brave new world of RNA-binding proteins. *Nat Rev Mol Cell Biol* **19**, 327-341, doi:10.1038/nrm.2017.130 (2018).
324 Chen, Y. & Wang, X. miRDB: an online database for prediction of functional microRNA targets. *Nucleic Acids Res* **48**, D127-D131, doi:10.1093/nar/gkz757 (2020).
325 Forster, S. C., Tate, M. D. & Hertzog, P. J. MicroRNA as Type I Interferon-Regulated Transcripts and Modulators of the Innate Immune Response. *Front Immunol* **6**, 334, doi:10.3389/fimmu.2015.00334 (2015).
326 Athey, J. *et al.* A new and updated resource for codon usage tables. *Bmc Bioinformatics* **18**, 391, doi:10.1186/s12859-017-1793-7 (2017).
327 Kahles, A., Ong, C. S., Zhong, Y. & Ratsch, G. SplAdder: identification, quantification and testing of alternative splicing events from RNA-Seq data. *Bioinformatics* **32**, 1840-1847, doi:10.1093/bioinformatics/btw076 (2016).
328 Mercer, T. R., Dinger, M. E. & Mattick, J. S. Long non-coding RNAs: insights into functions. *Nat Rev Genet* **10**, 155-159, doi:10.1038/nrg2521 (2009).
329 Grandi, N. & Tramontano, E. Type W Human Endogenous Retrovirus (HERV-W) Integrations and Their Mobilization by L1 Machinery: Contribution to the Human Transcriptome and Impact on the Host Physiopathology. *Viruses* **9**, doi:10.3390/v9070162 (2017).
330 Grandi, N. & Tramontano, E. Human Endogenous Retroviruses Are Ancient Acquired Elements Still Shaping Innate Immune Responses. *Front Immunol* **9**, 2039, doi:10.3389/fimmu.2018.02039 (2018).
331 Hurst, T. P. & Magiorkinis, G. Activation of the innate immune response by endogenous retroviruses. *J Gen Virol* **96**, 1207-1218, doi:10.1099/jgv.0.000017 (2015).
332 Morozov, V. A., Dao Thi, V. L. & Denner, J. The transmembrane protein of the human endogenous retrovirus--K (HERV-K) modulates cytokine release and gene expression. *PLoS One* **8**, e70399, doi:10.1371/journal.pone.0070399 (2013).
333 Faghihi, M. A. & Wahlestedt, C. Regulatory roles of natural antisense transcripts. *Nat Rev Mol Cell Biol* **10**, 637-643, doi:10.1038/nrm2738 (2009).
334 Schoggins, J. W. *et al.* A diverse range of gene products are effectors of the type I interferon antiviral response. *Nature* **472**, 481-485, doi:10.1038/nature09907 (2011).

335 Schoggins, J. W. *et al.* Pan-viral specificity of IFN-induced genes reveals new roles for cGAS in innate immunity. *Nature* **505**, 691-695, doi:10.1038/nature12862 (2014).
336 Jovanovic, M. *et al.* Immunogenetics. Dynamic profiling of the protein life cycle in response to pathogens. *Science* **347**, 1259038, doi:10.1126/science.1259038 (2015).
337 Fournier, M. L. *et al.* Delayed correlation of mRNA and protein expression in rapamycin-treated cells and a role for Ggc1 in cellular sensitivity to rapamycin. *Mol Cell Proteomics* **9**, 271-284, doi:10.1074/mcp.M900415-MCP200 (2010).
338 Tarasova, I. A. *et al.* Comparative proteomics as a tool for identifying specific alterations within interferon response pathways in human glioblastoma multiforme cells. *Oncotarget* **9**, 1785-1802, doi:10.18632/oncotarget.22751 (2018).
339 Felix, M. A. & Barkoulas, M. Pervasive robustness in biological systems. *Nat Rev Genet* **16**, 483-496, doi:10.1038/nrg3949 (2015).
340 Li, G. W., Burkhardt, D., Gross, C. & Weissman, J. S. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell* **157**, 624-635, doi:10.1016/j.cell.2014.02.033 (2014).
341 Taggart, J. C. & Li, G. W. Production of Protein-Complex Components Is Stoichiometric and Lacks General Feedback Regulation in Eukaryotes. *Cell Syst* **7**, 580-589 e584, doi:10.1016/j.cels.2018.11.003 (2018).
342 Dephoure, N. *et al.* Quantitative proteomic analysis reveals posttranslational responses to aneuploidy in yeast. *Elife* **3**, e03023, doi:10.7554/eLife.03023 (2014).
343 Mather, W. H., Hasty, J., Tsimring, L. S. & Williams, R. J. Translational cross talk in gene networks. *Biophys J* **104**, 2564-2572, doi:10.1016/j.bpj.2013.04.049 (2013).
344 Warner, J. R. The economics of ribosome biosynthesis in yeast. *Trends Biochem Sci* **24**, 437-440, doi:10.1016/s0968-0004(99)01460-7 (1999).
345 Barr, S. D., Smiley, J. R. & Bushman, F. D. The interferon response inhibits HIV particle production by induction of TRIM22. *Plos Pathog* **4**, e1000007, doi:10.1371/journal.ppat.1000007 (2008).
346 Di Pietro, A. *et al.* TRIM22 inhibits influenza A virus infection by targeting the viral nucleoprotein for degradation. *J Virol* **87**, 4523-4533, doi:10.1128/JVI.02548-12 (2013).
347 Wilkins, C. *et al.* IFITM1 is a tight junction protein that inhibits hepatitis C virus entry. *Hepatology* **57**, 461-469, doi:10.1002/hep.26066 (2013).
348 Faraoni, I., Antonetti, F. R., Cardone, J. & Bonmassar, E. miR-155 gene: a typical multifunctional microRNA. *Biochim Biophys Acta* **1792**, 497-505, doi:10.1016/j.bbadis.2009.02.013 (2009).
349 Taganov, K. D., Boldin, M. P., Chang, K. J. & Baltimore, D. NF-kappaB-dependent induction of microRNA miR-146, an inhibitor targeted to signaling proteins of innate immune responses. *Proc Natl Acad Sci U S A* **103**, 12481-12486, doi:10.1073/pnas.0605298103 (2006).
350 O'Connell, R. M., Taganov, K. D., Boldin, M. P., Cheng, G. & Baltimore, D. MicroRNA-155 is induced during the macrophage inflammatory response. *Proc Natl Acad Sci U S A* **104**, 1604-1609, doi:10.1073/pnas.0610731104 (2007).
351 Yin, Q. *et al.* MicroRNA-155 is an Epstein-Barr virus-induced gene that modulates Epstein-Barr virus-regulated gene expression pathways. *J Virol* **82**, 5295-5306, doi:10.1128/JVI.02380-07 (2008).
352 Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218-223, doi:10.1126/science.1168978 (2009).
353 Kakumani, P. K. AGO-RBP crosstalk on target mRNAs: Implications in miRNA-guided gene silencing and cancer. *Transl Oncol* **21**, 101434, doi:10.1016/j.tranon.2022.101434 (2022).
354 Min, K. W. *et al.* AUF1 facilitates microRNA-mediated gene silencing. *Nucleic Acids Res* **45**, 6064-6073, doi:10.1093/nar/gkx149 (2017).

355 Tominaga, K. *et al.* Competitive regulation of nucleolin expression by HuR and miR-494. *Mol Cell Biol* **31**, 4219-4231, doi:10.1128/MCB.05955-11 (2011).
356 Doherty, M. K., Hammond, D. E., Clague, M. J., Gaskell, S. J. & Beynon, R. J. Turnover of the human proteome: determination of protein intracellular stability by dynamic SILAC. *J Proteome Res* **8**, 104-112, doi:10.1021/pr800641v (2009).
357 Mathieson, T. *et al.* Systematic analysis of protein turnover in primary cells. *Nat Commun* **9**, 689, doi:10.1038/s41467-018-03106-1 (2018).
358 Kim, W. *et al.* Systematic and quantitative assessment of the ubiquitin-modified proteome. *Mol Cell* **44**, 325-340, doi:10.1016/j.molcel.2011.08.025 (2011).
359 Valadkhan, S. & Gunawardane, L. S. lncRNA-mediated regulation of the interferon response. *Virus Res* **212**, 127-136, doi:10.1016/j.virusres.2015.09.023 (2016).
360 Bassett, A. R. *et al.* Considerations when investigating lncRNA function in vivo. *Elife* **3**, e03058, doi:10.7554/eLife.03058 (2014).
361 Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**, 357-360, doi:10.1038/nmeth.3317 (2015).
362 Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930, doi:10.1093/bioinformatics/btt656 (2014).
363 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140, doi:10.1093/bioinformatics/btp616 (2010).
364 Lawrence, M., Gentleman, R. & Carey, V. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* **25**, 1841-1842, doi:10.1093/bioinformatics/btp328 (2009).
365 Chai, H., Gu, Q., Hughes, J. & Robertson, D. L. In silico prediction of HIV-1-host molecular interactions and their directionality. *PLoS Comput Biol* **18**, e1009720, doi:10.1371/journal.pcbi.1009720 (2022).
366 Kuhn, M. Building predictive models in R using the caret package. *J Stat Softw* **28**, 1-26 (2008).
367 Vanhatupa, S., Ungureanu, D., Paakkunainen, M. & Silvennoinen, O. MAPK-induced Ser727 phosphorylation promotes SUMOylation of STAT1. *Biochem J* **409**, 179-185, doi:10.1042/BJ20070620 (2008).
368 Kumar, K. G., Krolewski, J. J. & Fuchs, S. Y. Phosphorylation and specific ubiquitin acceptor sites are required for ubiquitination and degradation of the IFNAR1 subunit of type I interferon receptor. *J Biol Chem* **279**, 46614-46620, doi:10.1074/jbc.M407082200 (2004).
369 Mazewski, C., Perez, R. E., Fish, E. N. & Platanias, L. C. Type I Interferon (IFN)-Regulated Activation of Canonical and Non-Canonical Signaling Pathways. *Front Immunol* **11**, 606456, doi:10.3389/fimmu.2020.606456 (2020).
370 Saleiro, D. & Platanias, L. C. Interferon signaling in cancer. Non-canonical pathways and control of intracellular immune checkpoints. *Semin Immunol* **43**, 101299, doi:10.1016/j.smim.2019.101299 (2019).
371 Uddin, S. *et al.* Activation of the p38 mitogen-activated protein kinase by type I interferons. *J Biol Chem* **274**, 30127-30131, doi:10.1074/jbc.274.42.30127 (1999).
372 Mayer, I. A. *et al.* The p38 MAPK pathway mediates the growth inhibitory effects of interferon-alpha in BCR-ABL-expressing cells. *J Biol Chem* **276**, 28570-28577, doi:10.1074/jbc.M011685200 (2001).
373 Verma, A. *et al.* Activation of the p38 mitogen-activated protein kinase mediates the suppressive effects of type I interferons and transforming growth factor-beta on normal hematopoiesis. *J Biol Chem* **277**, 7726-7735, doi:10.1074/jbc.M106640200 (2002).
374 Joshi, S. *et al.* Type I interferon (IFN)-dependent activation of Mnk1 and its role in the generation of growth inhibitory responses. *Proc Natl Acad Sci U S A* **106**, 12097-12102, doi:10.1073/pnas.0900562106 (2009).

375 Sharma, B. *et al.* Sprouty proteins are negative regulators of interferon (IFN) signaling and IFN-inducible biological responses. *J Biol Chem* **287**, 42352-42360, doi:10.1074/jbc.M112.400721 (2012).
376 Thyrell, L. *et al.* Interferon alpha-induced apoptosis in tumor cells is mediated through the phosphoinositide 3-kinase/mammalian target of rapamycin signaling pathway. *J Biol Chem* **279**, 24152-24162, doi:10.1074/jbc.M312219200 (2004).
377 Kaur, S. *et al.* Role of the Akt pathway in mRNA translation of interferon-stimulated genes. *Proc Natl Acad Sci U S A* **105**, 4808-4813, doi:10.1073/pnas.0710907105 (2008).
378 Yan, J. X., Packer, N. H., Gooley, A. A. & Williams, K. L. Protein phosphorylation: technologies for the identification of phosphoamino acids. *J Chromatogr A* **808**, 23-41, doi:10.1016/s0021-9673(98)00115-0 (1998).
379 Salih, E. Phosphoproteomics by mass spectrometry and classical protein chemistry approaches. *Mass Spectrom Rev* **24**, 828-846, doi:10.1002/mas.20042 (2005).
380 Palumbo, A. M. *et al.* Tandem mass spectrometry strategies for phosphoproteome analysis. *Mass Spectrom Rev* **30**, 600-625, doi:10.1002/mas.20310 (2011).
381 Boersema, P. J., Mohammed, S. & Heck, A. J. Phosphopeptide fragmentation and analysis by mass spectrometry. *J Mass Spectrom* **44**, 861-878, doi:10.1002/jms.1599 (2009).
382 Zhou, H. *et al.* Toward a comprehensive characterization of a human cancer cell phosphoproteome. *J Proteome Res* **12**, 260-271, doi:10.1021/pr300630k (2013).
383 Andersson, L. & Porath, J. Isolation of phosphoproteins by immobilized metal (Fe3+) affinity chromatography. *Anal Biochem* **154**, 250-254, doi:10.1016/0003-2697(86)90523-3 (1986).
384 Posewitz, M. C. & Tempst, P. Immobilized gallium(III) affinity chromatography of phosphopeptides. *Anal Chem* **71**, 2883-2892, doi:10.1021/ac981409y (1999).
385 Zhou, H. *et al.* Specific phosphopeptide enrichment with immobilized titanium ion affinity chromatography adsorbent for phosphoproteome analysis. *J Proteome Res* **7**, 3957-3967, doi:10.1021/pr800223m (2008).
386 Swaney, D. L., Wenger, C. D., Thomson, J. A. & Coon, J. J. Human embryonic stem cell phosphoproteome revealed by electron transfer dissociation tandem mass spectrometry. *Proc Natl Acad Sci U S A* **106**, 995-1000, doi:10.1073/pnas.0811964106 (2009).
387 Han, G. *et al.* Phosphoproteome analysis of human liver tissue by long-gradient nanoflow LC coupled with multiple stage MS analysis. *Electrophoresis* **31**, 1080-1089, doi:10.1002/elps.200900493 (2010).
388 Han, G. *et al.* Large-scale phosphoproteome analysis of human liver tissue by enrichment and fractionation of phosphopeptides with strong anion exchange chromatography. *Proteomics* **8**, 1346-1361, doi:10.1002/pmic.200700884 (2008).
389 Wu, B. *et al.* Structural basis for dsRNA recognition, filament formation, and antiviral signal activation by MDA5. *Cell* **152**, 276-289, doi:10.1016/j.cell.2012.11.048 (2013).
390 Rehwinkel, J. & Reis e Sousa, C. RIGorous detection: exposing virus through RNA sensing. *Science* **327**, 284-286, doi:10.1126/science.1185068 (2010).
391 Fensterl, V. & Sen, G. C. Interferon-induced Ifit proteins: their role in viral pathogenesis. *J Virol* **89**, 2462-2468, doi:10.1128/JVI.02744-14 (2015).
392 Eskildsen, S., Justesen, J., Schierup, M. H. & Hartmann, R. Characterization of the 2'-5'-oligoadenylate synthetase ubiquitin-like family. *Nucleic Acids Res* **31**, 3166-3173, doi:10.1093/nar/gkg427 (2003).
393 Nishikura, K. Functions and regulation of RNA editing by ADAR deaminases. *Annu Rev Biochem* **79**, 321-349, doi:10.1146/annurev-biochem-060208-105251 (2010).
394 Garcia, M. A., Meurs, E. F. & Esteban, M. The dsRNA protein kinase PKR: virus and cell control. *Biochimie* **89**, 799-811, doi:10.1016/j.biochi.2007.03.001 (2007).
395 Li, Z. & Nagy, P. D. Diverse roles of host RNA binding proteins in RNA virus replication. *RNA Biol* **8**, 305-315, doi:10.4161/rna.8.2.15391 (2011).

396 Butter, F., Scheibe, M., Morl, M. & Mann, M. Unbiased RNA-protein interaction screen by quantitative proteomics. *Proc Natl Acad Sci U S A* **106**, 10626-10631, doi:10.1073/pnas.0812099106 (2009).
397 Schmidt, C., Kramer, K. & Urlaub, H. Investigation of protein-RNA interactions by mass spectrometry--Techniques and applications. *J Proteomics* **75**, 3478-3494, doi:10.1016/j.jprot.2012.04.030 (2012).
398 Kramer, K. *et al.* Photo-cross-linking and high-resolution mass spectrometry for assignment of RNA-binding sites in RNA-binding proteins. *Nat Methods* **11**, 1064-1070, doi:10.1038/nmeth.3092 (2014).
399 Castello, A. *et al.* System-wide identification of RNA-binding proteins by interactome capture. *Nat Protoc* **8**, 491-500, doi:10.1038/nprot.2013.020 (2013).
400 Beckmann, B. M. *et al.* The RNA-binding proteomes from yeast to man harbour conserved enigmRBPs. *Nat Commun* **6**, 10127, doi:10.1038/ncomms10127 (2015).
401 Matia-Gonzalez, A. M., Laing, E. E. & Gerber, A. P. Conserved mRNA-binding proteomes in eukaryotic organisms. *Nat Struct Mol Biol* **22**, 1027-1033, doi:10.1038/nsmb.3128 (2015).
402 Sysoev, V. O. *et al.* Global changes of the RNA-bound proteome during the maternal-to-zygotic transition in Drosophila. *Nat Commun* **7**, 12128, doi:10.1038/ncomms12128 (2016).
403 Bach-Pages, M., Castello, A. & Preston, G. M. Plant RNA Interactome Capture: Revealing the Plant RBPome. *Trends Plant Sci* **22**, 449-451, doi:10.1016/j.tplants.2017.04.006 (2017).
404 Hornbeck, P. V. *et al.* PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res* **40**, D261-270, doi:10.1093/nar/gkr1122 (2012).
405 Dinkel, H. *et al.* Phospho.ELM: a database of phosphorylation sites--update 2011. *Nucleic Acids Res* **39**, D261-267, doi:10.1093/nar/gkq1104 (2011).
406 Licata, L. *et al.* SIGNOR 2.0, the SIGnaling Network Open Resource 2.0: 2019 update. *Nucleic Acids Res* **48**, D504-D510, doi:10.1093/nar/gkz949 (2020).
407 Damle, N. P. & Kohn, M. The human DEPhOsphorylation Database DEPOD: 2019 update. *Database (Oxford)* **2019**, doi:10.1093/database/baz133 (2019).
408 Huang, H. *et al.* iPTMnet: an integrated resource for protein post-translational modification network discovery. *Nucleic Acids Res* **46**, D542-D550, doi:10.1093/nar/gkx1104 (2018).
409 Linding, R. *et al.* NetworKIN: a resource for exploring cellular phosphorylation networks. *Nucleic Acids Res* **36**, D695-699, doi:10.1093/nar/gkm902 (2008).
410 Horn, H. *et al.* KinomeXplorer: an integrated platform for kinome biology studies. *Nat Methods* **11**, 603-604, doi:10.1038/nmeth.2968 (2014).
411 Siew, K. & O'Shaughnessy, K. M. Extrarenal roles of the with-no-lysine[K] kinases (WNKs). *Clin Exp Pharmacol Physiol* **40**, 885-894, doi:10.1111/1440-1681.12108 (2013).
412 You, F. *et al.* PCBP2 mediates degradation of the adaptor MAVS via the HECT ubiquitin ligase AIP4. *Nat Immunol* **10**, 1300-1308, doi:10.1038/ni.1815 (2009).
413 Walter, B. L., Parsley, T. B., Ehrenfeld, E. & Semler, B. L. Distinct poly(rC) binding protein KH domain determinants for poliovirus translation initiation and viral RNA replication. *J Virol* **76**, 12008-12022, doi:10.1128/jvi.76.23.12008-12022.2002 (2002).
414 Li, Y. *et al.* LSm14A is a processing body-associated sensor of viral nucleic acids that initiates cellular antiviral response in the early phase of viral infection. *Proc Natl Acad Sci U S A* **109**, 11770-11775, doi:10.1073/pnas.1203405109 (2012).
415 Wang, Y. *et al.* Ankrd17 positively regulates RIG-I-like receptor (RLR)-mediated immune signaling. *Eur J Immunol* **42**, 1304-1315, doi:10.1002/eji.201142125 (2012).
416 Krischuns, T. *et al.* Phosphorylation of TRIM28 Enhances the Expression of IFN-beta and Proinflammatory Cytokines During HPAIV Infection of Human Lung Epithelial Cells. *Front Immunol* **9**, 2229, doi:10.3389/fimmu.2018.02229 (2018).

417 Webb-Robertson, B. J. *et al.* Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. *J Proteome Res* **14**, 1993-2001, doi:10.1021/pr501138h (2015).
418 Lazar, C., Gatto, L., Ferro, M., Bruley, C. & Burger, T. Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies. *J Proteome Res* **15**, 1116-1125, doi:10.1021/acs.jproteome.5b00981 (2016).
419 Stekhoven, D. J. & Buhlmann, P. MissForest--non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28**, 112-118, doi:10.1093/bioinformatics/btr597 (2012).
420 Zhang, X. *et al.* Proteome-wide identification of ubiquitin interactions using UbIA-MS. *Nat Protoc* **13**, 530-550, doi:10.1038/nprot.2017.147 (2018).
421 Wagner, S. A. *et al.* A proteome-wide, quantitative survey of in vivo ubiquitylation sites reveals widespread regulatory roles. *Mol Cell Proteomics* **10**, M111 013284, doi:10.1074/mcp.M111.013284 (2011).
422 Lee, S. *et al.* The SARS-CoV-2 RNA interactome. *Mol Cell* **81**, 2838-2850 e2836, doi:10.1016/j.molcel.2021.04.022 (2021).
423 Martin-Vicente, M., Medrano, L. M., Resino, S., Garcia-Sastre, A. & Martinez, I. TRIM25 in the Regulation of the Antiviral Innate Immunity. *Front Immunol* **8**, 1187, doi:10.3389/fimmu.2017.01187 (2017).
424 Welsby, I. *et al.* PARP12, an interferon-stimulated gene involved in the control of protein translation and inflammation. *J Biol Chem* **289**, 26642-26657, doi:10.1074/jbc.M114.589515 (2014).
425 Lund, N. *et al.* Differential effects of hnRNP D/AUF1 isoforms on HIV-1 gene expression. *Nucleic Acids Res* **40**, 3663-3675, doi:10.1093/nar/gkr1238 (2012).
426 Lee, N., Pimienta, G. & Steitz, J. A. AUF1/hnRNP D is a novel protein partner of the EBER1 noncoding RNA of Epstein-Barr virus. *RNA* **18**, 2073-2082, doi:10.1261/rna.034900.112 (2012).
427 Friedrich, S. *et al.* AUF1 p45 promotes West Nile virus replication by an RNA chaperone activity that supports cyclization of the viral genome. *J Virol* **88**, 11586-11599, doi:10.1128/JVI.01283-14 (2014).
428 Paek, K. Y., Kim, C. S., Park, S. M., Kim, J. H. & Jang, S. K. RNA-binding protein hnRNP D modulates internal ribosome entry site-dependent translation of hepatitis C virus RNA. *J Virol* **82**, 12082-12093, doi:10.1128/JVI.01405-08 (2008).
429 Cathcart, A. L., Rozovics, J. M. & Semler, B. L. Cellular mRNA decay protein AUF1 negatively regulates enterovirus and human rhinovirus infections. *J Virol* **87**, 10423-10434, doi:10.1128/JVI.01049-13 (2013).
430 Hino, K. *et al.* Downregulation of Nipah virus N mRNA occurs through interaction between its 3' untranslated region and hnRNP D. *J Virol* **87**, 6582-6588, doi:10.1128/JVI.02495-12 (2013).
431 Lunde, B. M., Moore, C. & Varani, G. RNA-binding proteins: modular design for efficient function. *Nat Rev Mol Cell Biol* **8**, 479-490, doi:10.1038/nrm2178 (2007).
432 El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res* **47**, D427-D432, doi:10.1093/nar/gky995 (2019).
433 Castello, A. *et al.* Comprehensive Identification of RNA-Binding Domains in Human Cells. *Mol Cell* **63**, 696-710, doi:10.1016/j.molcel.2016.06.029 (2016).
434 Lovci, M. T., Bengtson, M. H. & Massirer, K. B. Post-Translational Modifications and RNA-Binding Proteins. *Adv Exp Med Biol* **907**, 297-317, doi:10.1007/978-3-319-29073-7_12 (2016).
435 England, W. E. *et al.* An atlas of posttranslational modifications on RNA binding proteins. *Nucleic Acids Res*, doi:10.1093/nar/gkac243 (2022).
436 Enokizono, Y. *et al.* Structure of hnRNP D complexed with single-stranded telomere DNA and unfolding of the quadruplex by heterogeneous nuclear ribonucleoprotein D. *J Biol Chem* **280**, 18862-18870, doi:10.1074/jbc.M411822200 (2005).

437 Humphrey, S. J., James, D. E. & Mann, M. Protein Phosphorylation: A Major Switch Mechanism for Metabolic Regulation. *Trends Endocrinol Metab* **26**, 676-687, doi:10.1016/j.tem.2015.09.013 (2015).
438 Jedrychowski, M. P. *et al.* Evaluation of HCD- and CID-type fragmentation within their respective detection platforms for murine phosphoproteomics. *Mol Cell Proteomics* **10**, M111 009910, doi:10.1074/mcp.M111.009910 (2011).
439 Nygaard, V., Rodland, E. A. & Hovig, E. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics* **17**, 29-39, doi:10.1093/biostatistics/kxv027 (2016).
440 Dayon, L. & Sanchez, J. C. Relative protein quantification by MS/MS using the tandem mass tag technology. *Methods Mol Biol* **893**, 115-127, doi:10.1007/978-1-61779-885-6_9 (2012).
441 El Bougrini, J., Dianoux, L. & Chelbi-Alix, M. K. PML positively regulates interferon gamma signaling. *Biochimie* **93**, 389-398, doi:10.1016/j.biochi.2010.11.005 (2011).
442 Shemesh, M., Lochte, S., Piehler, J. & Schreiber, G. IFNAR1 and IFNAR2 play distinct roles in initiating type I interferon-induced JAK-STAT signaling and activating STATs. *Sci Signal* **14**, eabe4627, doi:10.1126/scisignal.abe4627 (2021).
443 Liu, B. & Gao, C. Regulation of MAVS activation through post-translational modifications. *Curr Opin Immunol* **50**, 75-81, doi:10.1016/j.coi.2017.12.002 (2018).
444 Johnson, H. & White, F. M. Toward quantitative phosphotyrosine profiling in vivo. *Semin Cell Dev Biol* **23**, 854-862, doi:10.1016/j.semcdb.2012.05.008 (2012).
445 Sharma, K. *et al.* Ultradeep human phosphoproteome reveals a distinct regulatory nature of Tyr and Ser/Thr-based signaling. *Cell Rep* **8**, 1583-1594, doi:10.1016/j.celrep.2014.07.036 (2014).
446 Kruger, R., Kubler, D., Pallisse, R., Burkovski, A. & Lehmann, W. D. Protein and proteome phosphorylation stoichiometry analysis by element mass spectrometry. *Anal Chem* **78**, 1987-1994, doi:10.1021/ac051896z (2006).
447 Iselin, L. *et al.* Uncovering viral RNA-host cell interactions on a proteome-wide scale. *Trends Biochem Sci* **47**, 23-38, doi:10.1016/j.tibs.2021.08.002 (2022).
448 Katano-Toki, A. *et al.* THRAP3 interacts with HELZ2 and plays a novel role in adipocyte differentiation. *Mol Endocrinol* **27**, 769-780, doi:10.1210/me.2012-1332 (2013).
449 Leung, A. K. *et al.* Poly(ADP-ribose) regulates stress responses and microRNA activity in the cytoplasm. *Mol Cell* **42**, 489-499, doi:10.1016/j.molcel.2011.04.015 (2011).
450 Batth, T. S. *et al.* Protein Aggregation Capture on Microparticles Enables Multipurpose Proteomics Sample Preparation. *Mol Cell Proteomics* **18**, 1027-1035, doi:10.1074/mcp.TIR118.001270 (2019).
451 pheatmap: Pretty Heatmaps (2019).
452 Lachmann, A. & Ma'ayan, A. KEA: kinase enrichment analysis. *Bioinformatics* **25**, 684-686, doi:10.1093/bioinformatics/btp026 (2009).
453 Bouhaddou, M. *et al.* The Global Phosphorylation Landscape of SARS-CoV-2 Infection. *Cell* **182**, 685-712 e619, doi:10.1016/j.cell.2020.06.034 (2020).
454 Grau, J., Grosse, I. & Keilwagen, J. PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics* **31**, 2595-2597, doi:10.1093/bioinformatics/btv153 (2015).
455 org.Hs.eg.db: Genome wide annotation for Human (2020).
456 Aoki-Kinoshita, K. F. & Kanehisa, M. Gene annotation and pathway mapping in KEGG. *Methods Mol Biol* **396**, 71-91, doi:10.1007/978-1-59745-515-2_6 (2007).
457 Joshi-Tope, G. *et al.* Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res* **33**, D428-432, doi:10.1093/nar/gki072 (2005).
458 D'Eustachio, P. Reactome knowledgebase of human biological pathways and processes. *Methods Mol Biol* **694**, 49-61, doi:10.1007/978-1-60761-977-2_4 (2011).

459 Huang, S. S. & Fraenkel, E. Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks. *Sci Signal* **2**, ra40, doi:10.1126/scisignal.2000350 (2009).
460 Wu, G., Feng, X. & Stein, L. A human functional protein interaction network and its application to cancer data analysis. *Genome Biol* **11**, R53, doi:10.1186/gb-2010-11-5-r53 (2010).
461 Alon, U. Network motifs: theory and experimental approaches. *Nat Rev Genet* **8**, 450-461, doi:10.1038/nrg2102 (2007).
462 Wong, E., Baur, B., Quader, S. & Huang, C. H. Biological network motif detection: principles and practice. *Brief Bioinform* **13**, 202-215, doi:10.1093/bib/bbr033 (2012).
463 Albert, I. & Albert, R. Conserved network motifs allow protein-protein interaction prediction. *Bioinformatics* **20**, 3346-3352, doi:10.1093/bioinformatics/bth402 (2004).
464 Barabasi, A. L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nat Rev Genet* **5**, 101-113, doi:10.1038/nrg1272 (2004).
465 Jeong, H., Mason, S. P., Barabasi, A. L. & Oltvai, Z. N. Lethality and centrality in protein networks. *Nature* **411**, 41-42, doi:10.1038/35075138 (2001).
466 Brodsky, I. E. & Medzhitov, R. Targeting of immune signalling networks by bacterial pathogens. *Nat Cell Biol* **11**, 521-526, doi:10.1038/ncb0509-521 (2009).
467 Newman, M. E. Modularity and community structure in networks. *Proc Natl Acad Sci U S A* **103**, 8577-8582, doi:10.1073/pnas.0601602103 (2006).
468 Kim, J. & Guan, K. L. mTOR as a central hub of nutrient signalling and cell growth. *Nat Cell Biol* **21**, 63-71, doi:10.1038/s41556-018-0205-1 (2019).
469 Borneman, A. R. *et al.* Target hub proteins serve as master regulators of development in yeast. *Genes Dev* **20**, 435-448, doi:10.1101/gad.1389306 (2006).
470 Yu, H., Kim, P. M., Sprecher, E., Trifonov, V. & Gerstein, M. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput Biol* **3**, e59, doi:10.1371/journal.pcbi.0030059 (2007).
471 Zhang, Q. *et al.* ACE2 inhibits breast cancer angiogenesis via suppressing the VEGFa/VEGFR2/ERK pathway. *J Exp Clin Cancer Res* **38**, 173, doi:10.1186/s13046-019-1156-5 (2019).
472 Missiuro, P. V. *et al.* Information flow analysis of interactome networks. *PLoS Comput Biol* **5**, e1000350, doi:10.1371/journal.pcbi.1000350 (2009).
473 Lin, C. Y. *et al.* Hubba: hub objects analyzer--a framework of interactome hubs identification for network biology. *Nucleic Acids Res* **36**, W438-443, doi:10.1093/nar/gkn257 (2008).
474 Kustatscher, G. *et al.* Co-regulation map of the human proteome enables identification of protein functions. *Nat Biotechnol* **37**, 1361-1371, doi:10.1038/s41587-019-0298-5 (2019).
475 Ochoa, D. *et al.* The functional landscape of the human phosphoproteome. *Nat Biotechnol* **38**, 365-373, doi:10.1038/s41587-019-0344-3 (2020).
476 Chen, Q., Sun, L. & Chen, Z. J. Regulation and function of the cGAS-STING pathway of cytosolic DNA sensing. *Nat Immunol* **17**, 1142-1149, doi:10.1038/ni.3558 (2016).
477 Rio, D. C., Clark, S. G. & Tjian, R. A mammalian host-vector system that regulates expression and amplification of transfected genes by temperature induction. *Science* **227**, 23-28, doi:10.1126/science.2981116 (1985).
478 Schultz, B. R. & Chamberlain, J. S. Recombinant adeno-associated virus transduction and integration. *Mol Ther* **16**, 1189-1199, doi:10.1038/mt.2008.103 (2008).
479 Kowarz, E., Loscher, D. & Marschalek, R. Optimized Sleeping Beauty transposons rapidly generate stable transgenic cell lines. *Biotechnol J* **10**, 647-653, doi:10.1002/biot.201400821 (2015).
480 Zhou, H., Liu, Z. G., Sun, Z. W., Huang, Y. & Yu, W. Y. Generation of stable cell lines by site-specific integration of transgenes into engineered Chinese hamster ovary strains using an FLP-FRT system. *J Biotechnol* **147**, 122-129, doi:10.1016/j.jbiotec.2010.03.020 (2010).

481 Kobayashi, N. *et al.* Efficient Cre/loxP site-specific recombination in a HepG2 human liver cell line. *Cell Transplant* **9**, 737-742, doi:10.1177/096368970000900525 (2000).
482 Grindley, N. D., Whiteson, K. L. & Rice, P. A. Mechanisms of site-specific recombination. *Annu Rev Biochem* **75**, 567-605, doi:10.1146/annurev.biochem.73.011303.073908 (2006).
483 Helbig, K. J. & Beard, M. R. The role of viperin in the innate antiviral response. *J Mol Biol* **426**, 1210-1219, doi:10.1016/j.jmb.2013.10.019 (2014).
484 Schmid, S., Mordstein, M., Kochs, G., Garcia-Sastre, A. & Tenoever, B. R. Transcription factor redundancy ensures induction of the antiviral state. *J Biol Chem* **285**, 42013-42022, doi:10.1074/jbc.M110.165936 (2010).
485 Qureshi, S. A., Salditt-Georgieff, M. & Darnell, J. E., Jr. Tyrosine-phosphorylated Stat1 and Stat2 plus a 48-kDa protein all contact DNA in forming interferon-stimulated-gene factor 3. *Proc Natl Acad Sci U S A* **92**, 3829-3833, doi:10.1073/pnas.92.9.3829 (1995).
486 Yoneyama, M. *et al.* Shared and unique functions of the DExD/H-box helicases RIG-I, MDA5, and LGP2 in antiviral innate immunity. *J Immunol* **175**, 2851-2858, doi:10.4049/jimmunol.175.5.2851 (2005).
487 Gitlin, L. *et al.* Essential role of mda-5 in type I IFN responses to polyriboinosinic:polyribocytidylic acid and encephalomyocarditis picornavirus. *Proc Natl Acad Sci U S A* **103**, 8459-8464, doi:10.1073/pnas.0603082103 (2006).
488 Yoneyama, M. *et al.* The RNA helicase RIG-I has an essential function in double-stranded RNA-induced innate antiviral responses. *Nat Immunol* **5**, 730-737, doi:10.1038/ni1087 (2004).
489 Kim, M. J., Hwang, S. Y., Imaizumi, T. & Yoo, J. Y. Negative feedback regulation of RIG-I-mediated antiviral signaling by interferon-induced ISG15 conjugation. *J Virol* **82**, 1474-1483, doi:10.1128/JVI.01650-07 (2008).
490 Lodermeyer, V. *et al.* 90K, an interferon-stimulated gene product, reduces the infectivity of HIV-1. *Retrovirology* **10**, 111, doi:10.1186/1742-4690-10-111 (2013).
491 Takakura, M., Ishiguro, K., Akichika, S., Miyauchi, K. & Suzuki, T. Biogenesis and functions of aminocarboxypropyluridine in tRNA. *Nat Commun* **10**, 5542, doi:10.1038/s41467-019-13525-3 (2019).
492 Zhang, Y. *et al.* PARP9-DTX3L ubiquitin ligase targets host histone H2BJ and viral 3C protease to enhance interferon signaling and control viral infection. *Nat Immunol* **16**, 1215-1227, doi:10.1038/ni.3279 (2015).
493 Bedard, K. M. & Semler, B. L. Regulation of picornavirus gene expression. *Microbes Infect* **6**, 702-713, doi:10.1016/j.micinf.2004.03.001 (2004).
494 Yun, S. I. & Lee, Y. M. Zika virus: An emerging flavivirus. *J Microbiol* **55**, 204-219, doi:10.1007/s12275-017-7063-6 (2017).
495 Zhu, J. *et al.* Antiviral activity of human OASL protein is mediated by enhancing signaling of the RIG-I RNA sensor. *Immunity* **40**, 936-948, doi:10.1016/j.immuni.2014.05.007 (2014).
496 Taylor, R. T. *et al.* TRIM79alpha, an interferon-stimulated gene product, restricts tick-borne encephalitis virus replication by degrading the viral RNA polymerase. *Cell Host Microbe* **10**, 185-196, doi:10.1016/j.chom.2011.08.004 (2011).
497 Lin, Y. C. *et al.* Genome dynamics of the human embryonic kidney 293 lineage in response to cell biology manipulations. *Nat Commun* **5**, 4767, doi:10.1038/ncomms5767 (2014).
498 Feng, X., Zhang, J., Chen, W. N. & Ching, C. B. Proteome profiling of Epstein-Barr virus infected nasopharyngeal carcinoma cell line: identification of potential biomarkers by comparative iTRAQ-coupled 2D LC/MS-MS analysis. *J Proteomics* **74**, 567-576, doi:10.1016/j.jprot.2011.01.017 (2011).
499 Gorchakov, R., Frolova, E. & Frolov, I. Inhibition of transcription and translation in Sindbis virus-infected cells. *J Virol* **79**, 9397-9409, doi:10.1128/JVI.79.15.9397-9409.2005 (2005).

500 Ong, S. E. *et al.* Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics* **1**, 376-386, doi:10.1074/mcp.m200025-mcp200 (2002).
501 Smyth, G. K., Michaud, J. & Scott, H. S. Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics* **21**, 2067-2075, doi:10.1093/bioinformatics/bti270 (2005).
502 Smyth, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **3**, Article3, doi:10.2202/1544-6115.1027 (2004).
503 Zhou, Q. & Sharp, P. A. Tat-SF1: cofactor for stimulation of transcriptional elongation by HIV-1 Tat. *Science* **274**, 605-610, doi:10.1126/science.274.5287.605 (1996).
504 Konig, J. *et al.* iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol* **17**, 909-915, doi:10.1038/nsmb.1838 (2010).
505 Pratt, J. M. *et al.* Dynamics of protein turnover, a missing dimension in proteomics. *Mol Cell Proteomics* **1**, 579-591, doi:10.1074/mcp.m200046-mcp200 (2002).
506 Yang, Y. C. *et al.* CLIPdb: a CLIP-seq database for protein-RNA interactions. *BMC Genomics* **16**, 51, doi:10.1186/s12864-015-1273-2 (2015).
507 Wang, X., McLachlan, J., Zamore, P. D. & Hall, T. M. Modular recognition of RNA by a human pumilio-homology domain. *Cell* **110**, 501-512, doi:10.1016/s0092-8674(02)00873-5 (2002).
508 Chen, Y. & Varani, G. Engineering RNA-binding proteins for biology. *FEBS J* **280**, 3734-3754, doi:10.1111/febs.12375 (2013).
509 Hellman, L. M. & Fried, M. G. Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. *Nat Protoc* **2**, 1849-1861, doi:10.1038/nprot.2007.249 (2007).
510 Yang, L. *et al.* Single molecule fluorescence in situ hybridisation for quantitating post-transcriptional regulation in Drosophila brains. *Methods* **126**, 166-176, doi:10.1016/j.ymeth.2017.06.025 (2017).
511 Broecker, F. & Moelling, K. Evolution of Immune Systems From Viruses and Transposable Elements. *Front Microbiol* **10**, 51, doi:10.3389/fmicb.2019.00051 (2019).
512 Wei, W. *et al.* Human L1 retrotransposition: cis preference versus trans complementation. *Mol Cell Biol* **21**, 1429-1439, doi:10.1128/MCB.21.4.1429-1439.2001 (2001).
513 Grow, E. J. *et al.* Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature* **522**, 221-225, doi:10.1038/nature14308 (2015).
514 Lau, N. C. *et al.* Characterization of the piRNA complex from rat testes. *Science* **313**, 363-367, doi:10.1126/science.1130164 (2006).
515 Czech, B. *et al.* piRNA-Guided Genome Defense: From Biogenesis to Silencing. *Annu Rev Genet* **52**, 131-157, doi:10.1146/annurev-genet-120417-031441 (2018).
516 Aravin, A. A., Hannon, G. J. & Brennecke, J. The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science* **318**, 761-764, doi:10.1126/science.1146484 (2007).
517 Tcherkezian, J. *et al.* Proteomic analysis of cap-dependent translation identifies LARP1 as a key regulator of 5'TOP mRNA translation. *Genes Dev* **28**, 357-371, doi:10.1101/gad.231407.113 (2014).
518 Suzuki, Y. *et al.* Characterization of RyDEN (C19orf66) as an Interferon-Stimulated Cellular Inhibitor against Dengue Virus Replication. *Plos Pathog* **12**, e1005357, doi:10.1371/journal.ppat.1005357 (2016).
519 Plissonnier, M. L. *et al.* LARP1 binding to hepatitis C virus particles is correlated with intracellular retention of viral infectivity. *Virus Res* **271**, 197679, doi:10.1016/j.virusres.2019.197679 (2019).
520 Dreyfuss, G., Matunis, M. J., Pinol-Roma, S. & Burd, C. G. hnRNP proteins and the biogenesis of mRNA. *Annu Rev Biochem* **62**, 289-321, doi:10.1146/annurev.bi.62.070193.001445 (1993).