Characterizing host-virus interactions through host factor curation and network analysis

Sean Nesdoly

Department of Bioengineering

Graduate Program in Biological and Biomedical Engineering

McGill University

Montreal, Quebec, Canada

December 2020

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Engineering

Abstract

The biomolecular interactions that take place between a virus and its host are complex in nature and play a significant role in governing functional and evolutionary dynamics within a cell. Moreover, the mechanics by which viruses infect, replicate, and otherwise impact its host varies across individuals, species, and between generations. Therefore, determining what interactions occur, how they evolve, and their importance within the context of the set of all known interactions that take place between genes within a cell, known as the interactome, is of significant value and has yet to be fully established. This gap in knowledge makes it difficult to develop therapeutic interventions for active viral infections, as well as hinders the establishment of efficient cell-based vaccine manufacturing platforms, which rely on viruses to make their product.

To this end, a systems biology approach has been taken to elucidate a comprehensive set of human-influenza genetic and functional interactions. These are relationships that occur between and within viral and host genes, all the way from the level of the genome, up to the proteome, and beyond. To characterize these interactions, host genes that have evidence of interacting with the influenza virus, also known as host factors, were curated from the literature. These genes were then classified, with varying degrees of confidence based on the nature of the interaction and experimental source, as either being essential for or acting against viral propagation. Next, a candidate set of antiviral genes was derived computationally from the analysis of second-generation sequencing data obtained from a genome-wide CRISPR/Cas9-mediated knockout screen of HEK-293SF cells that were infected with the influenza virus. This powerful type of experiment, whereby the human genome is systematically perturbed and the resulting phenotypic effects are measured, can be used to probe the function of genes, and therefore the nature of their interactions, under a variety of conditions, and, more importantly, at the level of the cell. These annotated datasets, one with support from the literature and the other derived from experimental data, were then used as a framework to investigate the human interactome,

wherein various biological properties and functional relationships were explored. From this, a set of antiviral genes has been identified and characterized as potential targets for knocking out, or otherwise genetically engineering, as a way to increase viral titres in cell-based vaccine manufacturing platforms. This study also extends upon the knowledge base for proviral genes, which may be used to develop more effective or novel treatments for inhibiting active viral infections.

Résumé

Les interactions biomoléculaires qui ont lieu entre un virus et son hôte sont de nature complexe et jouent un rôle important dans la gouvernance de la dynamique fonctionnelle et évolutive au sein d'une cellule. De plus, les mécanismes par lesquels les virus infectent, se reproduisent et ont un impact sur leur hôte varient selon les individus, les espèces et les générations. Par conséquent, déterminer quelles interactions se produisent, comment elles évoluent et leur importance dans le contexte de l'ensemble de toutes les interactions identifiées qui ont lieu entre les gènes au sein d'une cellule, autrement connu sous le nom d'interactome, est d'une valeur significative et doit encore être pleinement établi. Cette lacune du point de vu des connaissances rend difficile la mise au point de protocoles d'interventions thérapeutiques contre les infections virales actives, et entrave la mise en place de plateformes cellulaires de fabrication de vaccins efficaces, qui dépendent des virus pour produire les composantes du dit vaccin.

À cette fin, une approche de biologie systémique a été adoptée pour élucider un ensemble complet d'interactions génétiques et fonctionnelles entre le virus de la grippe et le système immunitaire humain. Il s'agit de relations qui se produisent entre et au sein des gènes viraux et hôtes, depuis le niveau du génome jusqu'au protéome, et au-delà. Pour caractériser ces interactions, les gènes de l'hôte, collectivement appelés facteurs de l'hôte lorsqu'ils sont étudiés par rapport aux agents pathogènes, qui ont été identifiés comme interagissant avec le virus de la grippe, ont été tirés de la littérature. Ces gènes ont ensuite été classés, avec des degrés de confiance variables en fonction de la nature de l'interaction et de la source expérimentale, comme étant essentiels à la propagation du virus ou agissant contre celle-ci. Ensuite, un ensemble candidat de gènes antiviraux a été dérivé par calcul à partir de l'analyse des données de séquençage deuxième génération obtenues en effectuant un criblage pangénomique, via la technologie CRISPR/Cas9, de cellules HEK-293SF infectées par le virus de la grippe. Ce type d'expérience, qui consiste à perturber systématiquement le génome humain et à mesurer les effets phénotypiques qui en résultent, peut être utilisé pour son-

der la fonction des gènes, et donc la nature de leurs interactions, dans diverses conditions et, surtout, au niveau de la cellule. Ces ensembles de données annotées, l'un étayé par la littérature et l'autre dérivé de données expérimentales, ont ensuite été utilisés comme cadre d'étude de l'interactome humain, dans lequel diverses propriétés biologiques et relations fonctionnelles ont été explorées. À partir de là, des gènes antiviraux ont été identifiés et caractérisés comme cibles potentielles pour l'élimination par génie génétique, afin d'augmenter les titres viraux dans les plateformes cellulaires de fabrication de vaccins. Cette étude élargit également la base de connaissances sur les gènes proviraux, qui peuvent aussi être utilisés pour développer des traitements plus efficaces, plus innovants pour inhiber les infections virales actives.

Acknowledgements

I would like to express my sincere thanks to both of my supervisors, Professors Yu Xia and Amine Kamen, for their dedicated support, patience, and mentorship throughout my journey as a Master's student. Their perpetual passion and relentless curiosity for science and the pursuit of robust research was, is, and will continue to be a constant inspiration for me. This, combined with their commitment to quality supervision and teaching, made for a truly wonderful experience.

I am also very fortunate to have been surrounded by numerous creative, experienced, knowledgeable, and kind individuals within the research groups of Professors Yu Xia and Amine Kamen. I am particularly grateful to David Sharon, who generously took me under his arm as both a friend and fellow student; this thesis would not have been possible without him. Special thanks to Marie-Angélique Sene for her thoughtful advice and discussions, as well as for translations to French. Thanks to Yilun Han for his kind friendship, generosity, and willingness to share both knowledge and experiences. It was also a pleasure to interact and work with Léah Pollet, Mohamed Ghadie, Yangchun (Frank) Chen, Rodrigo Migueles Ramirez, Avital Sharir-Ivry, Jean-François Gélinas, Sascha Kiesslich, Pablo Diego Moço, Michelle Yen Tran, Pranav Joshi, José Pedro Losa, and other members of both research groups.

Many thanks to Prof. W. Robert J. Funnell, whose insightful feedback and suggestions during my committee meetings kept me on track, and to all of the individuals that support the administrative activities of the McGill Department of Bioengineering and the Biological & Biomedical Engineering program. I am also greatly appreciative for the financial support that was provided through the Natural Sciences and Engineering Research Council of Canada (NSERC) as part of the Alexander Graham Bell Canada Graduate Scholarships; moreover, this work would not have been possible without additional financial support from both of my supervisors, Professors Yu Xia and Amine Kamen.

Finally, I am beyond thankful for the support, love, and patience that my family and close friends have shown me; their willingness to believe in me is truly appreciated.

Contribution of Authors

Professors Yu Xia and Amine Kamen provided critical guidance and supervision towards the completion of each of the chapters within this thesis.

Chapter 3: Host Factor Curation

Dr. Jean-François Gélinas provided the initial idea for the host factor classification scheme illustrated in figure 1.

David Sharon designed all experiments, executed most of them, and prepared the manuscript for the pooled genome-wide knockout screening strategy described in the Sharon *et al.* (2020) study[89]; the computational analysis of the output from the screen applied to HEK-293SF cells infected with influenza, given in sections 3.1.3 and 3.2.2, was executed by Sean Nesdoly, with direct guidance from David Sharon and supervision from Professors Yu Xia and Amine Kamen.

Fold enrichment calculations given in table 4 were developed with guidance from Professor Yu Xia.

Chapter 4: Host-Virus Interactome Analysis

All figures and tables were planned, executed, and analyzed by Sean Nesdoly, with guidance from Professor Yu Xia.

Dr. Mohamed Ghadie provided critical analysis and discussion of the vertex degree distributions given in figures 7, 8, and 15.

Contents

List of Tables

List of Figures

Nomenclature

E	The set of edges for a graph G
G	A graph, as defined by its vertex set V and edge set E
$G[S]$	Induced subgraph of G formed by restricting vertices and edge endpoints to those in the set S
$N(v)$	Neighbourhood of a vertex v; that is, the subgraph of G formed by the vertices adjacent to v.
V	The set of vertices for a graph G
Cas9	CRISPR-associated protein 9
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
FDR	False Discovery Rate
GO	Gene Ontology
HA	Hemagglutinin
HEK-293SF	Human Embryonic Kidney 293SF cells, adapted to serum-free media and suspension culture
HI-union	Union of Human Interactomes
HPIDB3.0	Host-Pathogen Interaction Database, version 3.0
HuRI	Human Reference Interactome
NA	Neuraminidase
NAI	Neuraminidase Inhibitors
PCC	Pearson Correlation Coefficient
PPI	Protein-Protein Interaction
RNAi	RNA interference
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2

sgRNA	single guide RNA
SGS	Second-Generation Sequencing
siRNA	small interfering RNA
STRING	Search Tool for Retrieval of Interacting Genes/Proteins
VLP	Virus-like Particle

Chapter 1: Introduction

The ubiquitous nature of viruses as biological entities within our ecosystems has had, and continues to have, profound effects on living organisms. The extent of their impact on life is significant and far-reaching, especially when one considers their ability to scale rapidly by infecting a host, hijacking its molecular machinery, and subsequently replicating, packaging, and producing itself orders of magnitude over. This reliance of a virus on its host for survival and replication reinforces the notion that there exists an intimate relationship between them; viruses are, by their very nature, obligate parasites. The pursuit of understanding their relationship is complicated by the variability in their interactions across viral and host species, the stage of viral life cycle that they are in, their molecular environment, and numerous other factors. As such, this necessitates a careful and critical analysis to understand what is going on at a systems-wide level. Therefore, this thesis sets out to characterize some of the interactions that viruses—specifically, influenza—have with their living hosts by relying upon insights derived from large biological datasets. By doing so, a systems-level understanding of the set of all known molecular interactions that take place between host and viral genes, known as the host-virus interactome, may be developed. Importantly, this type of analysis compliments the knowledge derived from more traditional, reductionist-style experimental approaches. Whichever strategy is taken, the common goal remains to understand the extent and impact that viruses have had, and continue to have, on human health and disease, and, more broadly, on the evolutionary trajectory of life.

1.1 Motivation and Rationale of Research

The interactions that viruses have with their living hosts range from simple, direct physical interactions to more complex and nuanced indirect relationships; as such, the extent to which viruses perturb their hosts, the identity of the interacting components, and the methods by which they perform their function are all areas of study that are currently under intensive research[10,38,39,43,62,64,89,119]. This is especially the case in the year of 2020, with the advent of coronavirus disease 2019 (COVID-19) and the ensuing pandemic caused

by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The threat of infectious disease outbreaks also remains for other viruses; in particular, the influenza virus has a comparatively high burden of disease through its seasonal epidemics[54,91] and continues to have the potential to develop into much worse[70]. These factors have significantly increased the pressure to understand the intricacies of host-virus interactions; by doing so, more effective antiviral therapies may be developed, and, improvements may be made to cell-based vaccine production platforms.

Although the methods that different viral species use to hijack the cellular machinery of their host varies, similarities do exist. Therefore, knowledge of the mechanisms of infection and viral production—and, more generally, the targets of the host cell—from one species of virus can provide insight to the understanding of others. It is with this theme that this thesis sets out to explore and characterize the interactions that viruses—specifically, influenza—have with their hosts through the analysis of large biological datasets. Importantly, the work in this thesis aims to disambiguate the critical differences between anti- and pro-viral host genes, which have historically been studied independently of one another, given their distinct motivations. This unique aspect allows for the exploration of host factors that either improve viral production for use in cell-based vaccine development, or, restrict viral infection for use as an antiviral therapeutic.

1.2 Objectives

The main objective of this thesis is to use bioinformatic techniques to further characterize human-influenza genetic and functional interactions; by doing so, antiviral genes—those which may be inhibited to increase viral production for applications in cell-based vaccine development—and proviral genes—those which may be targeted for inhibition as a host-based antiviral therapy, or over-expressed to improve viral production capacity—will be identified. To achieve this, three specific aims have been defined:

1. Curate, annotate, and analyze host factors from the literature that have evidence of interacting with viruses, and specifically that of influenza.

2. Identify and rank antiviral host factors through the analysis of second-generation sequencing (SGS) reads derived from a genome-wide knockout screen of human embryonic kidney (HEK-293SF) cells infected with influenza. This bioinformatic analysis is included as part of the study conducted by Sharon *et al.* (2020)[89].

3. Build interactomes based on various interaction types and perform network analyses on them using both the literature-derived host factors from (1) and the computationally-derived putative antiviral host factors from (2) as a guide.

Through fulfillment of the three specific aims listed above, influenza-specific anti- and proviral host factors—along with a subset of their within-host and viral interactions—will be identified and further characterized within the context of several interactome networks. This information may be used to engineer and optimize cell-based vaccine production platforms, and, to extend upon the knowledge base for proviral genes, which may be used to develop more effective or novel treatments for inhibiting active viral infections.

1.3 Thesis Outline

This thesis is divided into six chapters, including this introductory one.

Chapter two provides some background information to set the stage for the thesis, as well as surveys the relevant literature pertaining to the topics of gene function, systems biology, and the applications of host-virus interactions in cell-based vaccine development and antiviral therapy.

Then, chapter three introduces a framework for classifying host factors (figure 1), provides several sets of literature-curated host factors, and walks through the computational analysis of SGS reads derived from a genome-wide knockout screen of HEK-293SF cells infected with influenza, as published in the Sharon *et al.* (2020) study[89]. It then concludes with a simple comparison of putative antiviral gene screen hits identified in a selection of influenza-specific host factor screening studies.

Chapter four introduces several different interactome datasets and explores their network properties (4.2.1 and 4.2.2). The host factors identified in chapter three are then used to

probe the various interactomes as an integrative approach for exploring their properties at the systems-level (4.2.3 and 4.2.4).

Finally, a general discussion is given in chapter five on: (5.1) methods for improving host factor curation efforts; (5.2) variations in high-throughput screening experiments that complicate their integration; and, (5.3) potential strategies for further characterization of host-virus interactions.

The thesis then concludes with chapter six, which provides a concise summary of the research findings with specific reference to the objectives.

Chapter 2: Literature Review

A literature review has been conducted that focuses on three main topics: (2.1) a selection of methods used to determine gene function; (2.2) computational techniques in systems biology that are used to probe large biological datasets; and, (2.3) the various applications that host-virus interactions have within the context of improving cell-based vaccine development, or, identifying targets for antiviral therapy.

2.1 Gene Function

This section introduces the gene as one of the basic units of function within a cell or virus. Then, a selection of methods used for determining the function of a gene of interest are discussed. Finally, some terminology is provided around the types of genes that are involved in host-virus interactions, and the ways in which they can be classified. By defining the function of a gene and exploring the methods that are used to determine it, we introduce a formal way of thinking about the nature of the interactions between a virus and its host.

2.1.1 Genes as Heritable Units of Function

At the most basic level, genes are sequences of DNA that are transcribed into messenger RNA (mRNA), translated into a sequence of amino acids, and then folded into a three-dimensional structure; this structure, the protein, is then capable of carrying out molecular function. At each level of this exchange in information between the different biopolymers, modifications may be made or regulation imparted.

Importantly, the complete sequence of DNA (genome), the set of transcribed RNA molecules (transcriptome), and the set of translated proteins (proteome) are all major biological components that contain the variation for which the mechanisms of evolution act upon. By extension, the interactions between an invading pathogen and its host also occur between these functional molecules, among others. As a consequence of these interactions, many pathogen and host genes are specifically designed, constrained, or otherwise altered

due to inter-species co-evolution; for the influenza virus, many of the mechanisms that drive this form of evolution and adaptation are unknown[96]. Furthermore, the impact that these processes have on host-virus interactions is significant, although the extent to which this occurs, and how it does so, remains unclear[85].

When considering the evolution of the influenza virus—where the scale of viral infection and replication is high within the host—there is more than sufficient opportunity for mutations to be introduced and subsequent selection to act upon. With recent advancements that have significantly reduced the cost per nucleotide of sequencing, it has made it possible to measure these genetic changes and therefore track the source and subsequent evolution of specific influenza virus species and their subtypes[115]. One such resource that has accomplished this is the 'Virus Pathogen Resource'[2] and its associated 'Influenza Research Database'[118], which actively collects and stores a compendium of data on viruses. However, the numerous other molecules that may act as the intermediary between viruses and their hosts—such as lipid, carbohydrate, and phosphate post-translational modifications on proteins—further complicates the understanding and tracking of heritable units of function; this makes it difficult to predict the functional consequences of host-virus interaction evolution.

2.1.2 Probing Gene Function

The majority of approaches aimed at determining the function of a gene—and thereby the nature of its interactions—involve perturbing the gene under study in some way and observing the phenotypic effect that it has on the organism, cell, or virus. However, the process of identifying a biological signal that corresponds to some aspect of the function of the gene under study, qualitatively or quantitatively measuring it, and finally distinguishing it from biological noise, is a formidable challenge. A selection of techniques used for measuring gene function—directly or indirectly—are discussed in the following sections.

Forward and Reverse Genetics

In many cases, the underlying genetic determinants of a given phenotype are unknown, especially with complex traits, where many genes act in combination to influence the final observable phenotype. As such, the process of forward genetics may be applied to discover the gene, or genetic elements, that contribute to a phenotype.

This process begins by identifying a specific qualitative or quantitative phenotype of interest. If sufficient phenotype and genotype data is available, patterns of heritability may be analyzed, which can roughly narrow down candidate genomic loci that are associated with the phenotype. Prior to the introduction of many of the essential tools of molecular genetics, this approach relied heavily on the presence of naturally occurring mutations within a population; this eventually evolved with the ability to synthetically induce targeted or random genetic mutations within an organism or cell at scale. This molecular genetic 'toolset'—which includes CRISPR/Cas9 and RNA interference (RNAi) systems, among others—then became essential for probing gene function and led to the feasibility of reverse genetics. Here, genetic modifications are purposely introduced into an organism, and any phenotypic alterations that result can be screened for and traced back to the corresponding genetic mutation. The principles of reverse genetics has paved the way towards numerous technological advancements in high-throughput genetic screening, making it possible to individually probe gene function at a genome-wide scale and thereby advance our fundamental understanding of cell biology[59,107].

High-Throughput Genetic Screening

High-throughput genetic screening is a systems-wide approach for probing gene function that is an incredibly powerful strategy for understanding basic biology. This screening strategy works by introducing a genetic perturbation to each gene—or other genetic element—within the target genome in turn and subsequently measuring the phenotypic effect that is produced by each of the perturbations. The general technique can be adapted in numerous ways to probe various aspects of cellular biology and gene function; examples

of this include: changing the method of gene perturbation, altering the experimental readout that is measured, or conducting the screen in a pooled versus arrayed format[83]. Each of these experimental alterations has their own positives and negatives. For example, the arrayed screening format has superior options for sensitivity in signal readout, but requires significant investment in time and equipment to efficiently carry out the screen and tends to introduce variation into the results; in contrast, the pooled format of a genetic screen—where a population of cells is perturbed within a single vessel and subsequently sorted based on a specific selection criteria—ensures that all cells are subjected to similar experimental conditions and can be conducted in a fraction of the cost, time, and effort[26].

As alluded to, one of the big limitations of high-throughput screening strategies is the comparatively low signal obtained from the readout of the experiments. As such, studies have to be carefully planned out in order to discover anything meaningful, let alone significant. Moreover, the methods used to distinguish signal from noise often rely on the application of statistical techniques that can deal with the high false discovery rates (FDR) that are inherent in biology. Given that such a large number of tests are being carried out simultaneously in high-throughput genetic screens, a number of different algorithms and software suites have been developed to deal with this challenge. A selection of commonly used software for the analysis of the output of perturbation-based genetic screens includes: `MAGeCK`[65,66,105], `HiTSelect`[25], `PinAPL-Py`[92], and `CRISPRBetaBinomial`[55]. Importantly, `MAGeCK` was selected to be used in the genome-wide knockout screen analysis described in chapter three of this thesis because of its ease of customization as a command-line tool, statistical rigour, supporting documentation, and publicly available code repository[44].

High-throughput genetic screening is increasingly being used to probe gene function under numerous experimental conditions, each of which provides significant value and potential for discovery in a wide variety of applications. Of particular interest is the use of these genetic screens for understanding and modelling host-virus interactions. By systematically perturbing host genes at a genome-wide scale, host factors may be identified that either positively or negatively alter viral production. One such study by Deans *et al.* (2016) used RNAi- and CRISPR/Cas9-based perturbations together in two separate screens to characterize the mechanism of action of a specific antiviral drug, for which improvements

were then made to an existing treatment regime against RNA viruses[24]. In another study by Orchard *et al.* (2016), a critical host factor that serves as the point of attachment for a norovirus was identified through screening; subsequent functional validation by analysis of its crystal structure identified the exact site of attachment[79]. Similarly, Park *et al.* (2017) utilized a genome-wide CRISPR/Cas9 screen to identify host factors that are essential for HIV infection and—importantly—not essential for the host cell, thereby making them prime host targets for therapy[81]. In conclusion, there are numerous studies that have used high-throughput genetic screening as a strategy for understanding host-virus interactions; further discussion on influenza-specific host factor screens is given in chapter three.

Observing Gene Function via Protein Structure

An alternative to indirectly measuring gene function through perturbation-based experiments is to observe the gene itself. Structural characteristics of the protein product of a gene can provide insight into its mechanism of action, potential binding partners, sites of post-translational modification, or location within the cell. In addition, the function of a given protein can be deduced if it has sequence or structural conservation with a homologous protein in another species for which its function is already known.

To obtain this type of insight into the function of a protein, its three-dimensional structure must first be determined. X-ray crystallography has been the primary method of doing this, whereby a 'snapshot' of the crystal structure of a macromolecule is taken by shooting X-ray's at it and subsequently analyzing the diffraction patterns made. This technology has made a significant impact on biomedical research since its discovery, and continues to do so; however, one limitation is in its requirement for the macromolecule under study to be in a crystal form. The process of protein crystallization can be quite difficult, and, more importantly, restricts molecular motion. Therefore, observing protein structure in its native environment of water—where molecular motion is not completely lost—would provide further mechanistic insight into protein function. In May of 2020, a breakthrough in cryo-electron microscopy—which images proteins in solution and thus addresses one of the limitations of X-ray crystallography—was made, where the individual atoms of a

protein were observed at an unprecedented resolution[72,117]; it is believed that this marks the turning point for the next-generation of protein structure determination, and, by extension, for probing gene function.

To conclude, the indirect measurement of gene function is often required as an alternative to directly observing gene function via its protein structure due to: the inherent complexities contained within cellular machinery that make it difficult to observe or deduce function from directly; the necessity for probing function under native conditions, rather than isolated; and, the significant amount of time and effort that is required to determine protein structure, especially under certain conditions of interest, such as during viral infection.

2.1.3 Host and Viral Factors

Interacting host and viral genes—collectively referred to as host and viral factors, respectively—are in a never-ending 'arms race' against one another in a battle to win the 'competition' of evolution; this also extends to the RNA- and protein-based functional counterparts of genes, along with any other functional genetic element. Mutations on one side that convey a functional advantage for itself or against its opponent are promptly responded to—in the timescale of evolution—with the positive or purifying selection of counter-mutations in the other[115]. It is this interplay between host and viral factors that has largely determined the severity of viral infections throughout history, with a significant turning point for the influenza virus being the 1918 flu pandemic, which introduced a highly pathogenic version of the virus (H1N1) to humans[70]. Furthermore, the small size of viral genomes—which is indicative of their role as obligate parasites—makes annotating the function of each of their genes a comparatively easier task than that of the human genome, which is significantly larger and more complex. This suggests that the host has a larger pool of available genetic components with which one may engineer to their advantage. Furthermore, the fact that viruses, by necessity, directly and indirectly interact with—and often perturb—a large proportion of the host genome, transcriptome, proteome, and other functional components, reinforces how important it is to have knowledge of the interactions that occur

between host and viral factors, along with the mechanisms by which they do so. To this end, specific definitions for the different types of host factors that are involved in the life cycle of a virus, and methods for their detection, are given in the following sections.

Antiviral Host Factors

A host factor whose function is antiviral in nature acts to restrict or inhibit invading viruses, thereby preventing or reducing viral infection. Collectively, antiviral host factors can be thought of as the set of immune system genes within a host. The function of each antiviral host factor varies considerably, with some directly interacting with viral components to restrict or inhibit infection, and others indirectly doing so, such as through the activation of other host factors. Moreover, some antiviral host factors completely abolish virus production, whereas others only reduce it; importantly, many viruses carry specific genes whose function is to inhibit antiviral host factors—with emphasis on those that completely restrict their production—so that they may successfully infect and replicate.

As a concrete example, the human innate immune response is triggered in part by pattern recognition receptors, such as the gene *DDX58* (`HGNC:19102`*) and toll-like receptors *TLR7* (`HGNC:15631`) and *TLR8* (`HGNC:15632`)[19]. These antiviral host factors specifically recognize pathogen-associated molecular patterns that are present on invading viruses or their products, and in response, trigger various signalling cascades that eventually culminate in the innate immune response.

Proviral Host Factors

In contrast to that of antiviral host factors, proviral host factors are necessary for or support the viral infection process. For example, membrane-bound sialic acid-containing host cell receptors act as the site of attachment for the influenza virus; without these, the virus cannot recognize, attach, and enter the host cell for infection[14].

Proviral factors may also be referred to as 'host dependency factors', such as in Li *et al.* (2020)[64], meaning that they assist, or are required for, viral function; that is, the virus

* 'HGNC' refers to the set of unique, official gene symbols & identifiers from the HUGO Gene Nomenclature Committee at the European Bioinformatics Institute[100].

depends on a specific set of host factors in order for it to complete its life cycle. Importantly, targeting proviral factors in host cells with drugs that inhibit their function may result in a reduction or complete loss of viral production by the host cell. As such, proviral factors are considered to be powerful therapeutic targets[109] that are able to mitigate viral infection without having to rely on the conservation of specific structural domain targets on viral proteins, which are comparatively fewer in number and under significant evolutionary pressure[115], thus making them susceptible to mutations that render their respective antivirals ineffective[52,78].

Viral Factors

The DNA-, RNA-, and protein-based functional components of a virus, or a subset of these depending on the type of virus, make up the set of viral factors that interact with anti- and pro-viral host factors. Viral factors tend to be highly specific to each family of virus, with significant variation also existing between subtypes of a single species; this is especially the case for the influenza virus, where numerous factors can alter its evolutionary trajectory (e.g. pandemic vs. seasonal subtypes of influenza)[115]. As viruses have compact DNA- or RNA-based genomes out of constraints in their size, each of their genes performs a critical function that is necessary for host infection. An example of a viral factor is the nonstructural protein 1 (NS1) of the influenza A virus; it functions as one of the major inhibitors of the host antiviral innate immune response, and has been shown to directly impair a number of antiviral host factors that are expressed upon detection of viral infection[95]. This reinforces that research conducted on viral factors is an important endeavour, as knowledge of the underlying genetic elements that determine virus pathogenicity, host range, and mechanisms of transmission is critical for informing infectious disease treatment and management[34,96].

Detecting Host-Virus Interactions by Probing Gene Function

Most of the techniques introduced in section 2.1.2 are applicable for the detection of interactions between host and viral factors. This is rationalized due to gene function

being partly defined by who a gene interacts with and how it does so, which is in turn defined by its sequence and associated protein structure. Therefore, knowledge of gene function—and having the ability to directly or indirectly probe it—is critical for detecting and characterizing host-virus interactions. A selection of experimental techniques specific to the detection of interactions between host and viral factors is herein discussed.

The first major technique for detecting host-virus interactions is through the measurement of viral titre upon perturbation of one or many target host factors, as introduced in the section on 'High-Throughput Genetic Screening'. Here, the viral titre of a cell with a specific perturbation is compared to that of a control; any observed differences can then be attributed to the perturbation, and further functional validation or rationalization can provide insight into the corresponding mechanism of action. Common genetic perturbations include: CRISPR/Cas9-mediated knockout, knockins, and transcriptional activation or inhibition; RNAi-based knockdown of gene expression; and gene inhibition, such as with a drug inhibitor. Importantly, many of these perturbations can be adapted to work within a screening strategy, whereby genes can be systematically probed at a genome-wide scale[69].

In contrast to perturbing host factors, viral factors may also be perturbed. Any changes in viral titre produced by the mutant virus may provide insight into the relative importance of the perturbed viral factor, as well as its mechanism of action. A study by Gack *et al.* (2009) used this strategy to identify host factor targets of the influenza A NS1 protein[35].

An alternative technique for probing the function of host and viral factors at a systems-wide level is the measurement of viral infection-induced host gene expression. By tracking the expression of host genes over various time-points of the viral life cycle, one can obtain molecular 'snapshots' of what is going on. The 'Virus Pathogen Resource' and its associated 'Influenza Research Database' is one such resource that has conducted these time-course transcriptomic experiments for numerous different strains of influenza, and other viruses[2,118]. However, characterizing the dynamic nature of host-virus interactions over the course of the life cycle of a virus is particularly challenging, especially when it requires the analysis of large datasets. Reproducibility of experimental results can be dif-

ficult, as technical artifacts—such as batch effects, or any number of other experimental variations—can easily introduce false positives or false negatives. Nonetheless, improvement in sequencing technology, which has led to the widespread use of RNA-sequencing over microarray-based technology, and software suites for rigorous bioinformatic analysis, has greatly improved the sensitivity of detection.

Lastly, host-virus interactions may be determined directly by querying for interactions between a specified 'bait' protein—such as the host or viral factor that is under study—and a set of possible interactors, also known colloquially as its 'prey'. This type of experiment, known as affinity purification followed by mass spectrometry (AP-MS), has been an instrumental technique for the detection and understanding of protein-protein interactions (PPIs)[71]. One such study by Thulasi Raman and Zhou (2016) used AP-MS to identify and analyze the set of host factors that interact with the influenza virus NS1 protein[97].

The recurrent theme of systems biology that underlies many of the techniques used to detect host-virus interactions, as presented in this section, is the core of what this thesis is based upon; as such, a more formal introduction to the field of systems biology is given next.

2.2 Systems Biology

In this section, basic terminology for the representation of large biological networks is introduced. Then, a selection of studies that use computational techniques to understand biological interactions and other cellular properties, at the systems-level, are described. Importantly, it should be kept in mind that each of the approaches used in systems biology share the common theme of collectively building towards a cellular-wide understanding, or model, of the cell.

2.2.1 Graph Theory and Network Analysis

The main data structure underpinning many of the computational techniques in systems biology is that of the *graph*. This data structure can be used to efficiently represent the set of all known biological interactions that take place between genes, or their products, within a cell; when a graph is used in this context, it is referred to as an *interactome*. The formal definition of a graph G is as follows:

$$G = \{V, E\}$$

Here, V is the set of vertices and E is the set of edges—or interactions—between vertices:

$$V = \{v_1, v_2, ..., v_n\}$$

$$E = \{(a, b) \mid a, b \in V\}$$

Importantly, edges can be directed, as in enzyme a catalyzes the reaction of substrate b, but not the other way around; or, undirected, where an interaction (a, b) is equivalent to (b, a), such as a binary physical interaction between two proteins a and b. The majority of the edges that will be considered in this thesis are *undirected*.

The function $N(v)$ denotes the *neighbourhood* of a vertex v; that is, the subgraph of G formed by the vertices adjacent to that of v and the edges that connect those neighbours,

if any. This can be thought of as all of the interacting partners of a protein v, along with any interactions that occur between them (excluding v). Importantly, the *degree* of a vertex v is the number of edges that connect to the vertex within the graph G (its 'neighbours'). The *degree sequence* of a graph G is the set of all vertex degree values, sorted in decreasing order, and is denoted by the variable k.

If we have a set S of vertices $S = \{s_1, s_2, ..., s_m\}$, we can compute the *induced subgraph* of G, denoted $G[S]$, by restricting vertices and edge endpoints to those in the set S. This notation is particularly useful when analyzing a set of genes of interest; for example, we can consider a set of antiviral host factors that are known to interact with the influenza virus and compute its induced subgraph within a target interactome G to get a sense of how they relate to one another.

Another important concept is that of a *path*; this is the sequence of edges that connects two vertices a and b (e.g. $\{(a, x), (x, y), (y, b)\}$). This can be used to query how 'close' two genes are in the 'space' of an interactome, which provides some measure of biological similarity.

Lastly, the operations performed on a graph may be thought of in terms of their computational complexity, which is often expressed using 'big O notation'. This gives an approximation as to how the run time or space requirements grow as the input to a given function grows; this has important implications for what can and cannot be computed in reasonable amounts of time or space. As interactome datasets inevitably grow in size, more and more consideration will have to be given to efficiently use storage space and optimize algorithm run times.

2.2.2 Host-Virus Interactions and their Network Properties

It has long been a goal of molecular biology to determine and understand the mechanisms of a cell at a systems-wide level[6]; by doing so, this would enable the ability to truly engineer and refine cellular systems, which has significant implications for the treatment of human disease. Moving beyond the 'single-gene' approach to disease and viewing the cell in its entirety—where combinations of genes act together to create a disease phenotype

or fight against invading pathogens—opens up the door to a new frontier of 'network-based medicine'[7]. Of particular importance for this thesis is the mapping and characterization of host-virus interactions within the context of various interactomes.

Current approaches for identifying host-virus interactions—as discussed in the section on gene function (2.1)—involves systematically perturbing individual host genes and analyzing the effect that it has on a phenotypic trait of interest (here, viral titre). The set of gene 'hits' and their phenotypic effects may then be further probed by constructing a biological network—or interactome—that allows for the identification of patterns and the ability to compare with other networks derived from other experiments. When host factors of interest are analyzed in combination with biological interactions through the lens of graph theory, emergent properties of the system may be discovered that perhaps would not have been identifiable when considered individually.

For instance, Ackerman *et al.* (2018)[1] used human PPI networks integrated with known virus-host interactions to predict—using a network algorithm that takes advantage of shortest paths between genes of interest—proviral host factors that are likely 'druggable' candidates for the inhibition of influenza virus replication. In another study by Watanabe *et al.* (2014)[111], a comprehensive co-immunoprecipitation screen for host-influenza interactions, followed by validation in a small interfering RNA (siRNA) screen and functional annotation within an interactome, identified even more candidate antiviral drug targets. With many host-virus interaction datasets available for computational analysis, such as the Host-Pathogen Interaction Database (`HPIDB3.0`)[3] and the 'Virus Pathogen Resource'[2] (with its associated 'Influenza Research Database'[118]), there remains significant opportunity for the discovery of systems-wide properties of cells, especially as they relate to host-virus interactions[38,94].

2.3 Host Factors in Vaccine Development and Antiviral Therapy

This section provides a review of the following: (2.3.1) the human immune system and the role of vaccines in priming it; (2.3.2) the influenza virus and some of its mechanisms of infection; (2.3.3) current developments in upstream influenza vaccine manufacturing techniques for cell-based and other host platforms; (2.3.4) and, the role of antivirals in the treatment of viral infections. These topics highlight the relevant applications of anti- and pro-viral host factors in the improvement of cell-based vaccine manufacturing platforms, and, for the identification of host-virus interactions that may be targeted for therapeutic intervention in viral infections, respectively. Importantly, these applications would not be possible without the initial identification and characterization of host factors, as already discussed in the gene function and systems biology sections of this chapter.

2.3.1 The Human Immune System and Vaccines

The human immune system has developed in a competitive evolutionary 'arms race' against invading pathogens for a significant period of time. Both humans and pathogens have developed strategic mechanisms of defence and infection, respectively, in an attempt to outcompete the other. As a first line of defence against invading pathogens, humans have the innate immune system, which broadly identifies and initiates a nonspecific attack against unfamiliar antigens. The subsequent priming and action of the adaptive immune system—through activation of helper T cells, cytotoxic T cells, and B cells—ensures a co-ordinated destruction of infected cells to prevent further spread of the invading pathogen. Importantly, the activation of B cells leads to the secretion of antibodies that bind directly to their target antigen on the invading pathogen, which flags it for destruction by other immune cells; it may also suppress its function directly. Immunity against the invading pathogen persists through the production of memory B and T cells, which ensures that subsequent infection by the same pathogen is met with an already-primed immune system.

Within this context, vaccines play an incredibly important role as an artificial, controlled

immune system stimulant. The introduction of an attenuated version of a pathogen, or a modified functional component of it, leads to an immune response and the subsequent persisting immunity that accompanies it. By taking advantage of the incredibly complex and effective human immune system, vaccines have become one of the most important and effective prophylactic measures for infectious disease intervention that is available in the world today. However, that being said, vaccine development does not come without its challenges, as has been observed in the current coronavirus disease 2019 pandemic[87], as well as others in the past. If anything, this supports further research into vaccines, host factors, and the mechanisms of viral infection that are able to subvert our immune system.

2.3.2 Influenza

The influenza virus is an enveloped, single-stranded, negative-sense RNA virus that belongs to the *Orthomyxoviridae* family. Influenza type A is the most clinically relevant influenza virus among that of genera—or types—A, B, and C; it can be further characterized by its subtype (e.g. H1N1), which denotes the type of hemagglutinin (HA) and neuraminidase (NA) proteins that are present on its surface[11]. Influenza A is highly contagious, with a recent study estimating that 10% of unvaccinated adults are infected in any given year[91]. Moreover, it has a comparatively high burden of disease through its seasonal epidemics, with yearly deaths estimated to be between 291,000 and 646,000[54]. Of serious concern is the potential for the introduction of pandemic strains of influenza, which can be highly pathogenic; for instance, the deadly 1918 flu pandemic resulted in approximately 50 million deaths, with descendants of this strain of virus continuing to persist within the population, threatening future pandemics[70]. Given this, having the capacity to rapidly manufacture influenza vaccines for pandemic situations is imperative; as such, a discussion on methods for influenza vaccine manufacturing, and considerations related to vaccine efficacy, is provided next.

2.3.3 Influenza Vaccine Manufacturing

Vaccine development and manufacturing is a complex process, with many different strategies available. As each pathogen is unique in its structure and function, so to are the vaccines that induce immunity against them. Each type of vaccine, as discussed here, differs in their method and ease of production, efficacy, and safety. Nonetheless, vaccines remain as a powerful prophylactic against a wide variety of infectious diseases.

Selection of Influenza Virus Subtypes for Seasonal Vaccination

Seasonal vaccination is recommended due to the ability of the influenza virus to rapidly adapt and subvert our immune system[11]. It has been shown that a significant reduction in influenza disease can be achieved through modest increases in influenza vaccine coverage, along with the accurate selection of vaccine formulations[51]. The process of vaccination works by eliciting an antibody response against the currently circulating HA influenza surface protein. A vaccine that is deemed effective will have induced a humoral immune response that is capable of recognizing—and successfully responding to—future infections by the same subtype of virus; however, the significant capacity for mutations to occur in circulating influenza A subtypes[115] makes it difficult to predict which vaccine to produce to have the greatest impact in any given season.

As such, the World Health Organization (WHO) monitors changes in circulating influenza A virus subtypes through their 'Global Influenza Programme'[120] and its associated initiatives, like that of the 'FluWatch' system in Canada. Through this, they report annually, or more frequently for some regions, the expected set of subtypes that will likely dominate the upcoming flu season for a specific region of the world. These predictions are made based on data gathered in the most recent flu season in the opposite hemisphere of the globe, as influenza virus infections typically follow a cyclical pattern within each individual hemisphere[73]. Based on this data, the HA and NA influenza surface proteins from the selected strains are formulated into a vaccine that is then administered to the population.

After a specific influenza vaccine is produced, it is essential to determine the extent to which it induces an immune response, as vaccine and circulating influenza strain mismatches frequently occur due to the necessary reliance on strain predictions. One method used to measure this is the hemagglutination inhibition assay; this tests for the ability of the generated antibody to inhibit the natural agglutination property of HA in a sample of red blood cells. Despite this and other sophisticated techniques and programs for influenza virus surveillance, such as the 'Influenza Research Database'[118] and the previously mentioned 'Global Influenza Programme'[120], there remains the issue of having the sufficient capacity to rapidly generate vaccines at scale in response to a pandemic strain of influenza. This challenge is discussed after a brief introduction to influenza vaccine types.

Influenza Vaccine Types

The most commonly administered influenza vaccine types are the inactivated and live-attenuated viral vaccines. The inactivated form is derived from live influenza viruses that have been subjected to various downstream processing steps, the goal of which is to restrict its ability to replicate while also retaining the structural integrity of the HA viral surface protein, which contains the majority of its immunogenic properties; in contrast, the live-attenuated version actively replicates within its host, but does so only in regions that are below body temperature—like that of the nose—as the virus has been adapted to efficiently replicate at colder temperatures[31]. The methods by which each of these vaccine types are produced is an area of active research, and is discussed next.

Methods for Accelerating Influenza Vaccine Manufacturing

The ability to rapidly generate and scale-up the production of a vaccine against pandemic strains of influenza—which may spread very quickly within a population—is of paramount importance. Moreover, production of the seasonal influenza vaccine, which differs from year to year, will benefit from further developments in technology that accelerate the vaccine manufacturing pipeline. Two general approaches that aim to improve the vaccine

manufacturing process are the intensification of bioprocess technologies and the genetic engineering of host factors for increasing viral yield per cell. An integrated discussion of each of these strategies is given below, starting with the selection of a suitable host for virus production.

Irrespective of the final influenza vaccine format, a host system must first be selected to initially produce the influenza virus. This is one of the areas where the work presented in this thesis may be applied, as knowledge of specific host factors that may be engineered to increase influenza virus production, as well as for other viral species, can lead to improvements in existing vaccine manufacturing platforms.

Currently, the majority of influenza vaccines that are approved for use are produced in embryonated chicken eggs. The importance of having alternative methods for influenza vaccine production is particularly relevant when considering the potential threat of pandemic strains of the influenza virus. Here, the virus tends to spread very rapidly and often has a greater potential to be lethal within its host[57], which, among other negative effects, can severely complicate production platforms that rely on generating the live virus. For example, when producing an influenza virus vaccine in embryonated chicken eggs, the wildtype strain of virus used—which, in pandemic strains, often have genetic components that are of avian origin[90] —should not be lethal to the host, or else viral production can be significantly impacted. Furthermore, it is critical to have a consistent and reliable source of embryonated chicken eggs; however, lengthy production times and susceptibility to infectious diseases, among other factors, limits the capacity for rapid scale-up[68,80]. Therefore, recent developments in alternative host platforms, such as mammalian cells and plants, are challenging the egg-based production of influenza vaccines.

One such strategy for egg-independent influenza vaccine manufacturing uses recombinant technology to insert the influenza HA protein into a viral vector, which then infects a host and leads to the expression of HA. This production strategy can rapidly generate recombinant HA, contingent on knowing its subtype-specific genetic sequence. The ‘FluBlok’ recombinant influenza vaccine—which relies on a baculovirus as the viral vector and insect

cells as the production platform—was shown to be safe and immunogenic for the prophylaxis of influenza[98]. More recently, the Quebec City-based company Medicago completed a phase III clinical trial for its candidate seasonal influenza vaccine that is produced in the plant *Nicotiana benthamiana* as a recombinant virus-like particle (VLP)[108]. The vaccine production process involves infection of the plant by the bacterial vector *Agrobacterium tumefaciens*; this vector carries the gene encoding the influenza HA protein, which is taken up by the plant and subsequently expressed on the surface of VLPs that it produces[22]. Importantly, both of these manufacturing platforms show that the shift towards egg-independent influenza vaccine production is possible, with significant potential for scale-up[68].

The second general strategy for optimizing influenza vaccine manufacturing is through the identification of host factors that may be genetically engineered for the improvement of cell-based viral yields. Importantly, one may either knockout or knockdown an antiviral factor, or over-express a proviral factor, as methods for theoretically improving viral production in host cells. Notably, a number of perturbation-based genetic screens have been conducted with the goal of identifying influenza-specific anti- or pro-viral host factors[43,47,56,64,88,89,99,111]; however, the majority of these have focused on identifying proviral factors as candidate drug targets for the restriction or inhibition of viral infection. Studies that have conducted perturbation-based screens followed by genetic engineering of cell lines—with the specific goal of improving viral yield—include those by van der Sanden *et al.* (2016)[101] and Hoeksema *et al.* (2018)[49] for the production of poliovirus, and by Wu *et al.* (2017)[114] for rotavirus. Each of these studies used the Vero cell line as their choice of host, and have had variable success. This suggests that more research is required to understand the intricacies of host-virus interactions. An in-depth analysis of the genome-wide perturbation screen by Sharon *et al.* (2020)[89], which was conducted to identify influenza-specific antiviral host factors, is given as part of this thesis.

In conclusion, the selection of a host that is robust, can be rapidly generated from a reliable source, has sufficient capacity to adapt to variations in influenza strains that may impact production, and is amenable to genetic engineering for optimization of viral output, is highly desirable. The ideal situation would have the technology sufficiently developed

and readily available for a variety of manufacturing platforms; this would enable flexibility in the choice of response to changes in seasonal influenza strains, or, to the introduction of a novel pandemic strain, for which a new vaccine must be generated quickly and at scale.

2.3.4 Antiviral Therapeutics: a Last Line of Defence

A number of pharmaceutical antimicrobial drugs exist for the treatment of a wide variety of infectious diseases. Each one can be classified based on their mechanism of action; this, in turn, depends upon the molecular characteristics of their target pathogen. For example, antiviral drugs are often designed to inhibit a particular stage of the virus life cycle, such as viral replication, viral subunit assembly, or release from its host; in contrast, many antibiotics work by directly destroying the target bacteria itself—rather than its ability to replicate—through disruption of an essential molecular function. In the case of the penicillin group of antibiotics, interference with the synthesis of the bacterial cell wall, which is a critical function necessary for the survival of many bacterial species, leads to lysis and eventual cell death[30]. An example of an antiviral is the well-established mechanism of neuraminidase inhibitors (NAIs) for interfering with the influenza virus; here, inhibition of the NA viral protein, which normally catalyzes the release of a budding virion from the host cell membrane, leads to blocking of viral release and a concomitant reduction in its spread within the host[42]. Extrapolating from these examples, there is a tendency for antiviral therapeutics to be designed with their target virus specifically in mind, which differs from that of antibiotics, where the more commonly administered types target a comparatively larger number of species. This minor distinction between antibiotics and antivirals, among others, has implications for the way in which each one is used in practice. Antibiotics tend to be administered proactively as a form of empirical therapy; whereas, antivirals are more often used in cases of active viral infections where the disease-causing virus is definitively known, or, where there is a potential for severe illness[32]. In either case, antimicrobial drugs have been and continue to be used extensively in modern medicine as a critical component of the never-ending fight against infectious diseases.

Limitations of Antivirals as a Strategy for Disease Intervention

In contrast to the extensive use of antibiotics for infectious disease treatment and prevention, antiviral therapeutics play a comparatively less significant role, albeit still of importance. This can be attributed to the use of vaccination as an effective method of disease intervention. For the vast majority of viral infections—and particularly for influenza—the use of an antiviral is typically reserved for cases where patients are critically ill and require hospitalization, or, are at higher risk of developing severe complications[32]. In these circumstances, having an immediate treatment option in the form of an antiviral drug for preventing the onset of severe symptoms—or, in some cases, for reducing the risk of death—is critical. However, for influenza, their ability to effectively treat disease declines rapidly with the amount of time elapsed since the initial infection[102]. More importantly, antivirals are only applicable to the subset of people that are infected with the virus; as such, their capacity to impart a long-lasting global impact is limited in comparison to that of vaccines, as they fail to prevent the spread of disease. These factors suggest that antiviral therapeutics be treated as one of the last lines of defence against influenza infection; by doing so, they may reduce the chance that serious complications develop in at-risk patients. Therefore, other avenues of intervention should be emphasized—like that of vaccine development and immunization campaigns—in order to effectively reduce the rates of infectious disease morbidity and mortality.

For viruses that spread quickly and that have a vaccine available with sufficient efficacy, like that of influenza, preventing infection through immunization is more cost-effective[63] and less burdensome[51] for both people and health care systems. This is especially the case for influenza, where those at high risk for infection-related complications have significant overall improvements when vaccinated[40,74]. The stockpiling of antivirals in preparation for potential viral outbreaks, although important, is more of a reactionary precaution that fails to address the higher priority factors that aim to limit the spread of disease. Additionally, a major point of concern with relying on antivirals is the potential for the selection of viral mutants that are resistant to the administered drug[41,52]. If such an event occurs, treatment options for infection become limited and the reliance is shifted

to that of the human immune system to do its job as best it can. In direct response to this ever-present threat, circulating strains of the influenza virus are constantly monitored by the 'Global Influenza Programme' and its associated initiatives, like the 'Global Influenza Surveillance and Response System', for mutations that make antivirals ineffective[120]. Moreover, there is an increasing demand for new antivirals with novel mechanisms of action. However, barriers to such discoveries remain; the enormous cost and time investment that is required for the initial research and development of such a drug is often significant. This further complicates matters, making it unclear as to the amount of effort that should be put into such endeavours. Nonetheless, this reinforces the major advantage that vaccines have over antiviral therapeutics: a novel virus, or strain of an existing one, can emerge, and, without in-depth knowledge of its molecular characteristics, a vaccine can be formulated that triggers the human immune system to recognize and attack it. As is often the case, utilizing such a system that has had millions of years of evolutionary pressure to hone its function is the superior choice over trying to invent a new antiviral that, if successfully made, only reduces the risk of developing severe complications from the infection.

Antivirals for the Treatment of Influenza

The earliest class of antiviral drug for the treatment of influenza—which includes that of amantadine and rimantadine—was discovered by Davies *et al.*[23] in 1964 as an inhibitor of the M2 ion channel protein of influenza A[45]. However, these antiviral drugs—despite being shown to be effective for the treatment of influenza A infection[28] —are particularly susceptible to the development of viral resistance[29], do not have activity against many other viral strains like that of influenza B[23], and can have negative side effects[28].

The next class of influenza antivirals that were discovered—introduced as NAIs at the beginning of this section—work by inhibiting the viral NA protein on budding virions to prevent release of its HA from host cell membrane receptors that contain sialic acid[42]. By doing so, new virions are prevented from escaping their host cell, thereby stopping the spread of infection. It is worth noting that NAIs were discovered through rational drug

design based on the crystal structure of NA in complex with sialic acid[15] —an impressive feat for structural biology at the time. Commonly administered NAIs, listed in order of their discovery, include zanamivir[104], oseltamivir[58], and peramivir[4]; these antivirals are active against both influenza types A and B. However, like that of the M2 ion channel antivirals, the threat of emerging viral strains that are increasingly resistant to NAIs is still present[52,86].

Finally, the most recent class of influenza antivirals works through a novel mechanism of action by targeting the viral polymerase complex for inhibition; of these, the viral endonuclease inhibitor baloxavir marboxil has shown significant efficacy in reducing the duration of infection in people without comorbidities[46]. A clinical trial (NCT02949011) has been completed to test baloxavir marboxil in at-risk patients, which showed that it has similar efficacy to that of oseltamivir[53]. Again, despite it targeting a new viral protein, strains of influenza have been identified that exhibit resistance[78]. Nonetheless, this discovery has provided a new option for the treatment of influenza in the case that other antiviral types prove ineffective.

Chapter 3: Host Factor Curation

To fulfill the first objective of this thesis, the literature was searched for experimental evidence of host factors interacting with viruses. This was initially conducted to understand the common techniques used for the detection of host factors, and to generate a 'gold-standard' list to which comparisons may be made to. Then, the priority shifted to the identification of host factors specifically involved in the mechanisms of influenza virus infection; to ensure sufficient numbers were obtained, evidence from perturbation-based screening experiments were included. Lastly, to address the second objective of this thesis, SGS reads—derived from a genome-wide knockout screen of HEK-293SF cells infected with influenza—were analyzed computationally to identify and rank antiviral host factors; this analysis is included as part of the study by Sharon *et al.* (2020)[89].

3.1 Methodology

3.1.1 Host Factor Classification Scheme

To introduce a formal way of thinking about the nature of the interactions between a virus and its host, a classification scheme for host factors has been defined, as depicted in figure 1. Four different categories exist, each providing an indication of the relationship that a given host factor has with an invading virus: restriction, inhibitory, auxiliary, and essential. A specific host factor can be classified into one of the categories based on the extent to which its perturbation within a host alters the viral production capacity of a target virus. Importantly, the perturbation considered here is that of gene knockouts, such as those made by the CRISPR/Cas9 system; however, the type and direction of perturbation can be freely altered, with the only requirement being a corresponding change in the predicted viral output upon the application of the specific perturbation. As host-virus interactions may be direct or indirect in their interference or support of the life cycle of a virus, this classification scheme attempts to remove this complexity and focus on the resulting output that is produced by a given perturbation. This makes this scheme particularly useful for applications in the engineering of mammalian cell lines for the

improvement of vaccine manufacturing platforms.

Shown in figure 1 is the predicted relative viral titre produced by a host cell upon perturbation of a particular type of host factor. Importantly, the host factors are assumed to be host genes, but the general framework may be extended to that of other cellular components, such as noncoding (regulatory) DNA, post-translational modifications, and metabolites. The specific function of each gene within a given host factor category likely varies substantially between one another; however, they relate to each other through their common ability to, upon perturbation, alter a host cell's viral production capacity in the same direction and relative magnitude. It should be noted that for antiviral factors, the predicted impact on the cell's viral titre is assumed to be in the absence of any viral regulatory or accessory proteins that alter the host cell state to make it more susceptible to infection.

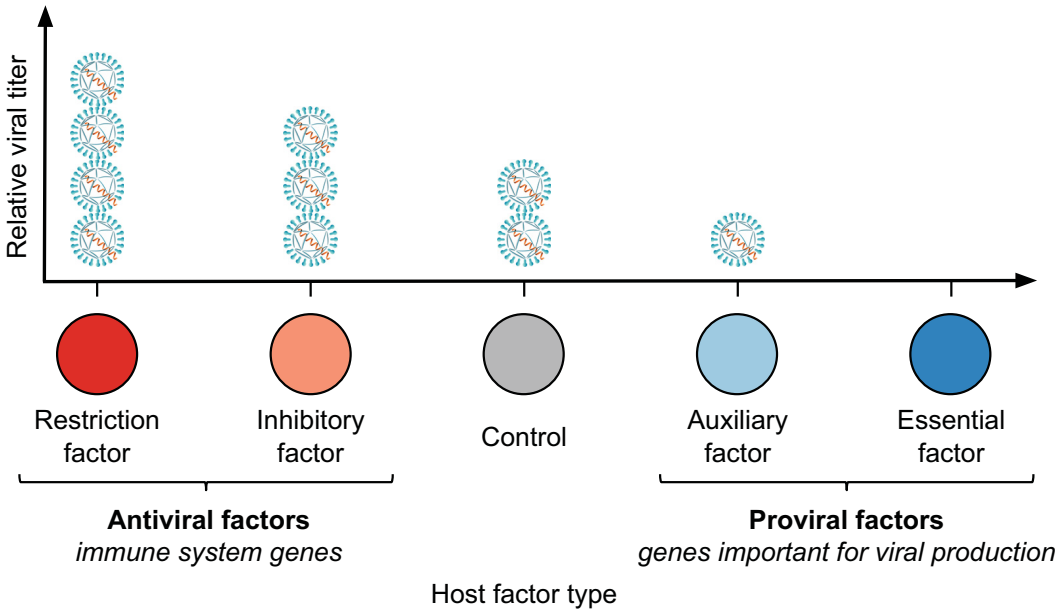

Figure 1: Host factor classification scheme as a tool to reason about the nature of host-virus interactions. The predicted impact of a gene-level genetic perturbation on the viral production capacity of a host cell is represented schematically in terms of arbitrary viral titre units (y-axis, diagrams of blue virus); here, titre is understood to be relative to the production capacity of a control cell (grey) that has no genetic perturbations. The classification of a given host factor into one of four types, as depicted along the x-axis, is therefore based on the nature of the interaction that occurs between the host factor and the virus. The interaction, and by extension, the host factor type, can be deduced indirectly by perturbing the host factor under study—through gene knockout, knockdown, or over-expression experiments—and comparing the viral titre produced to that of a control cell (grey) that lacks the genetic perturbation. Figure adapted from Gélinas *et al.* (2018)[37] with permission from the American Society for Microbiology (copyright 2018).

In the subsequent sections, specific definitions for each type of host factor are given, along with an example of each.

Restriction Factor

Under normal cell conditions and in the absence of any viral-encoded antagonists of host factors, a restriction factor *completely—or nearly—abolishes* viral production; upon knockout, the virus is free to infect and replicate without 'restriction'.

Examples of restriction factors include the interferon-inducible transmembrane (IFITM) family of proteins, specifically that of *IFITM2* (`HGNC:5413`) and *IFITM3* (`HGNC:5414`), which have been shown to restrict influenza A viral replication as part of the host innate immune system[12].

Inhibitory Factor

Under normal cell conditions and in the absence of any viral-encoded antagonists of host factors, an inhibitory factor *partially reduces* viral production; upon knockout, the virus has an increased capacity to replicate without 'inhibition'.

An example of an inhibitory factor is *DDX58* (`HGNC:19102`), which senses viral double-stranded RNA within the cytosol and initiates a non-specific immune response cascade against it. For a review on influenza A-specific host restriction and inhibitory factors, see the review paper by Villalón-Letelier *et al.* (2017)[103].

Auxiliary Factor

Under normal cell conditions, an auxiliary factor assists in the life cycle of a virus, but is not necessary for it to progress to completion; upon knockout, viral titre is *reduced*, but not entirely.

The host factor *UBR4* (`HGNC:30313`) has been shown to increase the efficiency at which influenza virions transit to the cell membrane when escaping; importantly, this factor is *not* essential for virus production[99].

Essential Factor

Under normal cell conditions, an essential factor is explicitly necessary for the completion of the life cycle of a virus; upon knockout, viral titre is *completely lost*.

The host factor *UVRAG* (`HGNC:12640`) has been shown to be an essential component of the mechanisms of host cell membrane-mediated entry for influenza A and vesicular stomatitis virus (VSV)[82].

Another example is the host cell membrane-bound sialic acid-containing glycan receptors that act as the site of attachment for the influenza virus by specifically binding to HA; without these glycans, the virus cannot recognize, attach, and enter the host cell through receptor-mediated endocytosis to initiate infection[14]. Although glycans themselves are not encoded by a gene within the host cell, it is indirectly encoded through the proper production of host glycosylation enzymes that decorate the cell membrane with cell-type specific glycans via post-translational modification of host cell proteins[48].

3.1.2 Curation of Host Factors from the Literature

Two rounds of literature-based host factor curation were conducted. The first identified host factors as part of an exploratory exercise, without restriction of the type of virus involved; however, a slight bias towards influenza is present. The second took a more targeted approach and identified host factors from screening experiments that specifically were involved in interactions with the influenza virus. A brief description of the methods for each are discussed in the following two sections.

Exploratory Curation of Host Factors from the Literature

A general set of host factors—without restriction on the type of virus involved—was curated from the literature as an exploratory exercise to become familiar with the experimental techniques used for their detection, and to generate a 'gold-standard' set, where each of the host factors have experimental evidence of interacting with a virus.

To do so, the following databases and search engines were queried: Google Scholar, PubMed, and UniProt. In addition, the section on PubMed that lists 'similar' publications to that of the currently selected was used to identify host factors that were 'close' in the PubMed search space. Finally, the bibliography of already-selected publications were searched to find similar experiments.

Importantly, for each publication that has evidence supporting a given host factor, manual extraction of the following data was done: a snippet of text that describes the mechanism by which viral activity is altered, the technique used to detect the host factor, and the extent to which viral production was perturbed. This supporting evidence was excluded from the results for brevity, but it assisted in classification of the host factors and with the understanding of the methods by which host-virus interactions are detected.

With regards to classification of the host factors, this was manually done based on the supporting publication and is thus a subjective measurement.

Curation of Influenza Host Factors Identified in Screening Experiments

To create a sufficiently large set of host factors specific to that of the influenza virus, studies that performed a perturbation-based screening strategy for the identification of genes that alter influenza viral production—either positively or negatively—were identified in the literature.

To ensure that the host factors identified in each study could be compared to one another, all gene hits were converted to the common format of NCBI gene identifiers using the DAVID gene conversion tool[50] and the Ensembl BioMart database[116]. In addition, study-specific details were noted, such as the method of gene perturbation—for example, knockdowns with siRNA vs. knockouts with CRISPR/Cas9-based systems—and the method of detection—for example, viral titre measurements vs. co-immunoprecipitation pull-down experiments.

3.1.3 Computational Analysis of a Genome-wide Knockout Screen to Identify Anti-influenza Host Factors

A significant component of the work in this thesis was dedicated to the analysis of SGS reads that were produced from a genome-wide CRISPR/Cas9-mediated knockout screen of HEK-293SF cells that were infected with the influenza virus. The methods and results of this computational analysis are published as part of the study by Sharon *et al.* (2020)[89]. This section presents an overview of the bioinformatic methods that were used in this study, with some of the text having been published as part of it.

Background on Screening Strategy used to Identify Putative Influenza-specific Antiviral Host Factors

To introduce the problem, it is necessary to give some background. This part of the experiment was designed and executed by David Sharon; for more details than what is described here, refer to the paper (**specifically, figure 1**)[89]. The general objective of the screening strategy is to identify host factors within the genome of the HEK-293SF cell line whose perturbation—in this case, a knockout mutation—increases the amount of influenza viral titre produced by the cell. It is assumed that the host factors identified by the screen will be classified as antiviral—as per the host factor classification scheme described in figure 1—but, this does not necessarily have to be the case, as long as the perturbation increases the relative viral titre in comparison to that of a control. The screen probes a heterogeneous mix of gene knockouts—where each cell has one knockout event—in a pooled screening format within a single vessel, which has the unique characteristic of exposing each cell to a similar environmental condition; moreover, it allows for the number of genes queried to be scaled easily to reach a genome-wide coverage, which is not trivial to achieve in an array-based screening format.

In order to query the effect that knocking out a gene has on the influenza viral production capacity of an HEK-293SF cell, a knockout mutation must first be introduced. To do this for every gene in the human genome, a pool of HEK-293SF cells were in-

fected with a library of single-guide RNAs (sgRNA) packaged within lentiviral vectors, with each one also carrying the sequence for Cas9 to complete the CRISPR/Cas9 system; this sgRNA library—known as the 'Brunello Human CRISPR/Cas9 Knockout Pooled Library'—contains 76,441 unique sgRNA sequences that target 19,114 genes with four-fold redundancy, and an additional 1,000 sgRNAs reserved as non-targeting controls[27]. The Brunello library of sgRNAs thus permits the generation of a heterogeneous pool of HEK-293SF cells—where each cell has one gene knocked out—with genome-wide coverage being achieved in the population of cells.

With the heterogeneous population of HEK-293SF knockout cells established, there remains the problem of detecting influenza viral titre at the level of each individual cell. To do so, an influenza virus with the HA gene swapped out for the gene encoding green fluorescent protein (GFP)—which permits measurement of its relative viral titre via the intensity of its fluorescence—is used to infect the heterogeneous pool of HEK-293SF cells. The cells are then subjected to a selection step that sorts them into two populations based on the relative viral titre produced by each individual cell; this selection is accomplished using fluorescence-activated cell sorting. The 'high yield' population is enriched for gene knockouts that improve influenza virus production capacity, and thus contains putative antiviral host factors. The other population—referred to as the 'control'—contains all of the other cells that did not meet the minimum relative viral titre during selection, as measured by GFP fluorescence intensity; gene knockouts that correspond to putative proviral host factors may be present, but the cell population is not specifically depleted for this.

Now that a 'high yield' population of HEK-293SF knockout cells has been enriched for via sorting by GFP fluorescence intensity—which is used as a proxy for influenza viral titre as it is expressed in direct proportion to that of wild-type HA—all that remains is the quantification and identification of the individual gene knockouts that are present within each population ('high yield' vs. 'control'). To do so, the sgRNA inserts used to induce gene knockouts are extracted from the two cell populations and sequenced on an Illumina HiSeq 4000 machine. The next task, as described in the following sections, quantifies the differences in sgRNA abundance between the 'high-yield' and 'control' cell populations in

order to identify genes that are enriched or depleted.

Sequencing and Knockout Screen Quality Control Metrics

First, `FastQC` (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) was used to assess the quality of raw sequencing data; given raw reads in `FASTQ` format, it detects any biases or errors that may have been introduced during sample preparation or the sequencing itself.

Next, the Brunello sgRNA library file[27] —which maps each sgRNA sequence to its corresponding target gene name and unique identifier—was modified from its original format to work in the subsequent analyses.

Then, the 'Model-based Analysis of Genome-wide CRISPR/Cas9 Knockout' (`MAGeCK`) software suite (version 0.5.9.2)[65,66,105] was used to assess a number of CRISPR screen quality control metrics that are calculated at the levels of: raw sequencing data, aligned sgRNA read counts, collapsed gene counts, and between samples; descriptions of these and their expected thresholds are neatly summarized in table 1 of the paper by Li *et al.* (2015)[66].

Quantification of sgRNA Abundance using MAGeCK count

Quantification of sgRNA abundance in the two HEK-293SF knockout cell populations ('high yield' and 'control' conditions, three biological replicates each) was carried out using the `MAGeCK` software suite (version 0.5.9.2)[65,66,105]. Specifically, the `count` function of `MAGeCK` was used in command-line mode to map the reads to the Brunello sgRNA library, thereby 'counting' up the number of cells with a given sgRNA sequence, for each sgRNA in the library. The sgRNA sequence contained within a given cell indicates the occurrence of a gene knockout event corresponding to the target gene of that sgRNA. Importantly, `MAGeCK count` only keeps primary alignments with no mismatches, which prevents the mapping of one read to two library sgRNAs, effectively counting it twice; additionally, reads that contain an 'N' nucleotide, indicating a low-confidence base call, are discarded.

Custom code was also written in the Julia programming language[9] to analyze the output of `MAGeCK count` for the purpose of exploration, verification, and adjustment of input parameters.

Comparing sgRNA Abundance Between High Yield and Control Cell Populations using MAGeCK test

To test for the significance of enrichment or depletion of sgRNA abundance between the 'high yield' and 'control' cell populations, the `MAGeCK` function `test` was used.

Briefly, `MAGeCK test` identifies significantly enriched or depleted genes by modelling the relationship between the mean and variance of sgRNA abundance in the control condition in order to estimate the parameters of a negative binomial distribution; this is then used to conduct a hypothesis test which robustly quantifies the significance of a change in average sgRNA abundance between conditions[65]. In this analysis, read counts were normalized using a set of 1,000 non-targeting control sgRNAs that were provided in the Brunello sgRNA library[27]. The initial set of 1,000 was reduced to 963 after the removal of outliers whose difference in normalized read counts between conditions were outside of the range:

$$[Q1 - 1.5 * IQR, Q3 + 1.5 * IQR] \tag{1}$$

where $Q1$, $Q3$, and IQR are the first quartile, third quartile, and interquartile range $(Q3 - Q1)$, respectively.

`MAGeCK test` was then used on the sgRNA read counts, as produced by `MAGeCK count`, with additional parameters `--remove-zero` and `--remove-zero-threshold` set to 'control' and '30', respectively. This removed 1,760 sgRNAs that have a median read count in the control condition that is less than 30, indicating insufficient representation. All other parameters were left at the default setting. Of note, sgRNA-level p-values were adjusted using the Benjamini-Hochberg procedure[8], which was set to control the FDR at level $\alpha = 0.25$. To obtain gene-level p-values from multiple sgRNAs targeting a single gene, version 0.5.9 of the modified Robust Rank Aggregation (RRA) algorithm[60] —named α-

RRA and included as part of `MAGeCK`—was used[65]. Similarly, the *log* fold change (LFC) of a gene is calculated as the median of the LFC values corresponding to the individual sgRNAs that target it. Custom analyses for exploration and verification of results were carried out using the programming language `Julia`[9]; plots were generated from the `MAGeCK` software suite[65,66,105] and the `R`[84] package `ggplot2`[112].

Replication of Results in PinAPL-Py

The 'Platform-independent Analysis of Pooled Screens using Python' (`PinAPL-Py`) web service[92] was used to replicate the results obtained from the `MAGeCK` software suite. The purpose of doing this was to verify the accuracy of results in another system that uses similar, but distinct normalization and statistical testing techniques. The results of this analysis were excluded from the results section for brevity, as they—for the most part—matched that of what was produced by `MAGeCK`.

3.1.4 Fold Enrichment Calculations for Gene Set Overlaps

Given two gene sets $A = \{a_1, a_2, ..., a_n\}$ and $B = \{b_1, b_2, ..., b_m\}$, which each correspond to gene hits derived from different perturbation-based genome-wide screens of cells infected with influenza, we can calculate a measure of enrichment for the overlapping genes by comparing the observed value to the number that are expected to overlap by chance.

We first calculate how many genes are expected to be overlapping between sets A and B:

$$\text{Expected Overlap}(A, B) = \frac{|A|}{n} \times \frac{|B|}{n} \times n \tag{2}$$

where n is the number of genes targeted for perturbation in each screen; for simplicity, we assume that $n = 20,000$ for both screens. The actual value of n may differ between screening experiments, but its effect on fold enrichment calculations is insignificant.

We then compare the observed and expected number of overlapping genes to obtain a fold enrichment value:

$$\text{Fold Enrichment}(A, B) = \frac{|A \cap B|}{\text{Expected Overlap(A,B)}} \tag{3}$$

The results of applying these enrichment calculations to a selection of the gene sets in table 3 are given in section 3.2.4.

3.2 Results

3.2.1 Host Factors from the Literature

To get a sense of the various types of host factors, the methods by which they are detected, and the nature of their interactions with invading viruses, the literature was searched for experimental evidence of host factors interacting with viruses.

In total, $n = 47$ host factors—each with experimental evidence of interacting with a virus—were curated from 45 publications and subsequently classified based on the scheme described in figure 1. The number of genes in each host factor type are as follows: 14 restriction, 21 inhibitory, 6 auxiliary, and 6 essential. Importantly, the classification of each host factor was manually annotated based on the supporting evidence given in its associated publication. The set of literature-curated host factors are given in table 1.

As previously stated, the exercise of curating host factors from the literature was initially conducted to understand the common techniques used for the detection of host factors, and to generate a 'gold-standard' list to which comparisons may be made to. Subsequent analyses that use this list of host factors may not use the complete set, depending on the data that is available; for example, gene *Fv1* (NCBI gene ID `14349`) is only found in *Mus musculus* (mice), and thus is not included in human-specific analyses.

Table 1: Host factors with literature-based evidence of interacting with a virus ($n = 47$).

HGNC Symbol*	NCBI Gene ID	Effect on Virus**	Virus†	PubMedID
ADAR	103	inhibitory	InfluenzaA; MeV	21159878
APOBEC3G	60489	inhibitory	HIV-1	12167863;12808465
ATP6V0D1	9114	essential	Influenza	18615016
BST2	684	restriction	HIV-1	18200009; 18342597
CCL5	6352	inhibitory	HIV-1	8525373
CDKN1B	1027	auxiliary	Influenza	20081832
CH25H	9023	inhibitory	Enveloped viruses	23273844
CLK1	1195	auxiliary	Influenza	20081832
COX6A1	1337	essential	Influenza	18615016
DDX58	23586	inhibitory	dsRNA viruses	15208624
EIF2AK2	5610	inhibitory	HIV-1	19229320
Fv1	14349	restriction	MLV	16474118
GNPTAB	79158	essential	Ebola virus	30655525
HDAC6	10013	restriction	InfluenzaA; HIV-1	30518648; 16148047; 25031336
HERC5	51191	inhibitory	InfluenzaA	20385878
IFI16	3428	inhibitory	KSHV	21575908
IFIH1	64135	inhibitory	HIV-1	14645903
IFIT1	3434	inhibitory	InfluenzaA; VSV	21642987
IFIT5	24138	inhibitory	NDV; SeV	23942572
IFITM1	8519	restriction	InfluenzaA	20064371
IFITM2	10581	restriction	InfluenzaA	20064371
IFITM3	10410	restriction	InfluenzaA	20064371
IFNB1	3456	inhibitory	RSV	19193793
ISG15	9636	restriction	InfluenzaA	20133869
MAVS	57506	inhibitory	SeV	16125763
METTL3	56339	essential	InfluenzaA	28910636
MOV10	4343	inhibitory	RNA viruses	27016603
MX1	4599	inhibitory	InfluenzaA	26202236
MX2	4600	restriction	HIV-1	24048477
NXF1	10482	essential	Influenza	18615016
OAS1	4938	inhibitory	Dengue virus	19923450
RNASEL	6041	inhibitory	SeV	17653195
RSAD2	91543	inhibitory	InfluenzaA	18005719
SAMHD1	25939	restriction	HIV-1	21720370; 21613998
SERINC3	10955	restriction	HIV-1	26416733
SERINC5	256987	restriction	HIV-1	26416733
SLFN11	91607	restriction	HIV-1	23000900
SON	6651	auxiliary	Influenza	20081832
STING1	340061	inhibitory	VSV; HSV-1	18724357
TP53	7157	auxiliary	InfluenzaA	29904383
TREX1	11277	auxiliary	HIV-1	20871604
TRIM25	7706	inhibitory	InfluenzaA	19454348
TRIM26	7726	inhibitory	VSV; NDV; SeV	26611359
TRIM28	10155	restriction	MLV	17923087
TRIM5	85363	restriction	HIV-1; MLV	16474118; 17156811
UVRAG	7405	essential	InfluenzaA; VSV	24550300
YTHDF2	51441	auxiliary	InfluenzaA; KSHV	28910636; 29659627

* 'HGNC Symbol' refers to the set of unique, official gene symbols from the HUGO Gene Nomenclature Committee at the European Bioinformatics Institute [100].

** The column 'Effect on Virus' denotes the manually annotated host factor classification, as supported by the evidence given in its associated publication (column 'PubMedID'); consult figure 1 for reference.

† Viruses are referred to by: type of virus; general class of virus, as defined by the Baltimore classification

system (e.g. 'dsRNA viruses'); or, a common characteristic shared by a group of viruses (e.g. 'Enveloped viruses').

Abbreviations: MeV, Measles morbillivirus; HIV-1, Human immunodeficiency virus type 1; MLV, Murine leukemia virus; KSHV, Kaposi's sarcoma-associated herpesvirus; VSV, Vesicular stomatitis virus; NDV, Newcastle disease virus; SeV, Sendai virus; RSV, Respiratory syncytial virus; HSV-1, Herpes simplex virus 1.

By going through the process of searching for and reading relevant publications that contain evidence of interactions occurring between viruses and their hosts, common experimental techniques used for identifying and validating host factors were determined. This provided the necessary background for the section in chapter 2 on probing gene function, where a selection of these experiments are described.

In the next section, results from the computational analysis of a genome-wide knockout screen of HEK-293SF cells—where the relative viral titre produced by each cell was used as a selection step—are presented. The results from this screen provide an experimental source of evidence for influenza-specific host factors, which supports the literature-curated set of nonspecific host factors provided here, in table 1.

3.2.2 Genome-wide Knockout Screen of HEK-293SF Cells Infected with Influenza for the Identification of Antiviral Host Factors

A genome-wide screening strategy for the identification of putative antiviral host factors, which may be used to genetically engineer cell-based vaccine cell lines to improve viral titres, is presented in the study by Sharon *et al.* (2020)[89]. Here, the results from the computational analysis of SGS reads derived from the application of the screening strategy to the HEK-293SF cell line—where a modified version of the influenza A virus was used for infection—is described in detail. For more details on the screening strategy, refer to the paper (**specifically, figure 1**) or the brief description already provided in the methods section of this chapter.

Knockout Screen Quality Control Metrics

To begin, a series of knockout screen quality control metrics were calculated using `FastQC` and `MAGeCK`[65,66,105]. This step is crucial for validating the screen output, as it detects the presence of any experimental biases and defines thresholds for each metric that should ideally be met to ensure proper interpretation of downstream functional analyses.

The first metric considered was the quality of sequencing data generated for each sample by the Illumina HiSeq 4000 machine. The figures associated with this analysis have been left out for brevity; refer to figure 4, subplots a and b, in the Sharon *et al.* (2020)[89] paper for a subset of these results.

Briefly, the reads generated for all six samples ('control' and 'high yield' conditions, each with three replicates) individually met the necessary quality control thresholds for the following metrics:

- Median per base sequence quality: Phred score >30 across all base pairs.
- Mean per sequence quality: single peak observed in distribution at Phred score of 39.
- Distribution of mean GC content over all sequences matches that of a theoretical normal distribution; here, GC content is defined as the percentage of G and C nucleotides in a given sequence.
- All reads are 50 base pairs in length, as defined by the sequencer.

Given the satisfactory quality of the sequencing data, the reads were mapped to the Brunello sgRNA library using the `count` function of `MAGeCK`; this 'counts' the number of cells with a given sgRNA sequence—each of which indicates a knockout corresponding to the target gene of that sgRNA—for each sgRNA in the Brunello library. A number of sgRNA-level quality control metrics associated with this mapping procedure are given in table 2.

Table 2: Knockout screen read mapping and sgRNA-level quality control metrics. sgRNA reads were aligned to the Brunello sgRNA library—which contains 77,441 targets—using the `MAGeCK` software suite[65,66,105]. Data corresponds to figure 4 (subplots c, d, and e) of the Sharon *et al.* (2020) study[89] and is under a Creative Commons license.

Sample*	Total Reads	Mapped Reads**	Zero Counts†	Gini index††
R1 high yield	36,119,880	24,370,285 (0.67)	209	0.09
R1 control	37,159,178	24,492,676 (0.66)	151	0.08
R2 high yield	37,239,926	25,136,938 (0.67)	187	0.09
R2 control	34,066,043	22,924,926 (0.67)	192	0.09
R3 high yield	38,394,069	25,419,311 (0.66)	248	0.09
R3 control	34,406,196	23,112,824 (0.67)	202	0.09

* R1, R2, and R3 denote biological replicates.

** The fraction of total reads successfully mapped to the Brunello sgRNA library is given in brackets.

† Zero counts are the number of unique sgRNAs in the Brunello library for which no read was mapped.

†† The Gini index is a measure for how even the distribution of sgRNA read counts is; values less than 0.1 indicate that the sgRNA representation in the population of cells is of high quality.

An important conclusion from table 2 is that the average number of reads per sgRNA in the Brunello library is ~ 300, indicating that there is sufficient representation of each gene knockout within the population of cells, for each sample[66].

The `count` function of `MAGeCK` also performs normalization on the sgRNA read counts; specifically, they are median-normalized—using the 'median ratio method'[65] —and then log_2 –transformed. Histograms of these transformed sgRNA read counts for all knockout screen samples are given in figure 2.

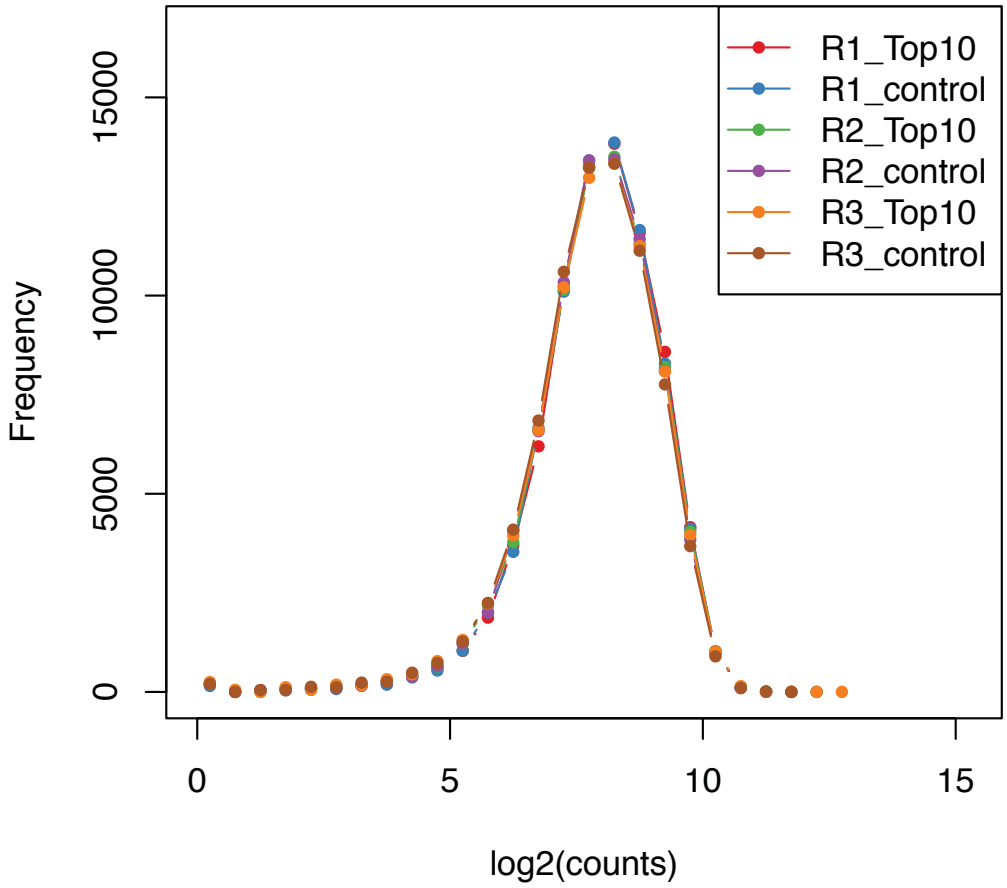

FIGURE 2: Frequency polygon (histogram) of median-normalized, log_2 –transformed sgRNA read counts for all knockout screen samples. Here, 'Top10' refers to the top 10% of GFP-expressing cells, which is synonymous with the 'high yield' cell population. Figure generated by the `MAGeCK` software suite[65,66,105].

Lastly, to assess the similarity of samples within an experimental condition (either the 'control' or 'high yield' cell populations), pairwise Pearson Correlation Coefficients (PCC) were calculated based on the sgRNA read count distributions of each sample; a heat map depicting the clustering of the knockout screen samples—which, importantly, cluster together within their respective conditions—is given in figure 3.

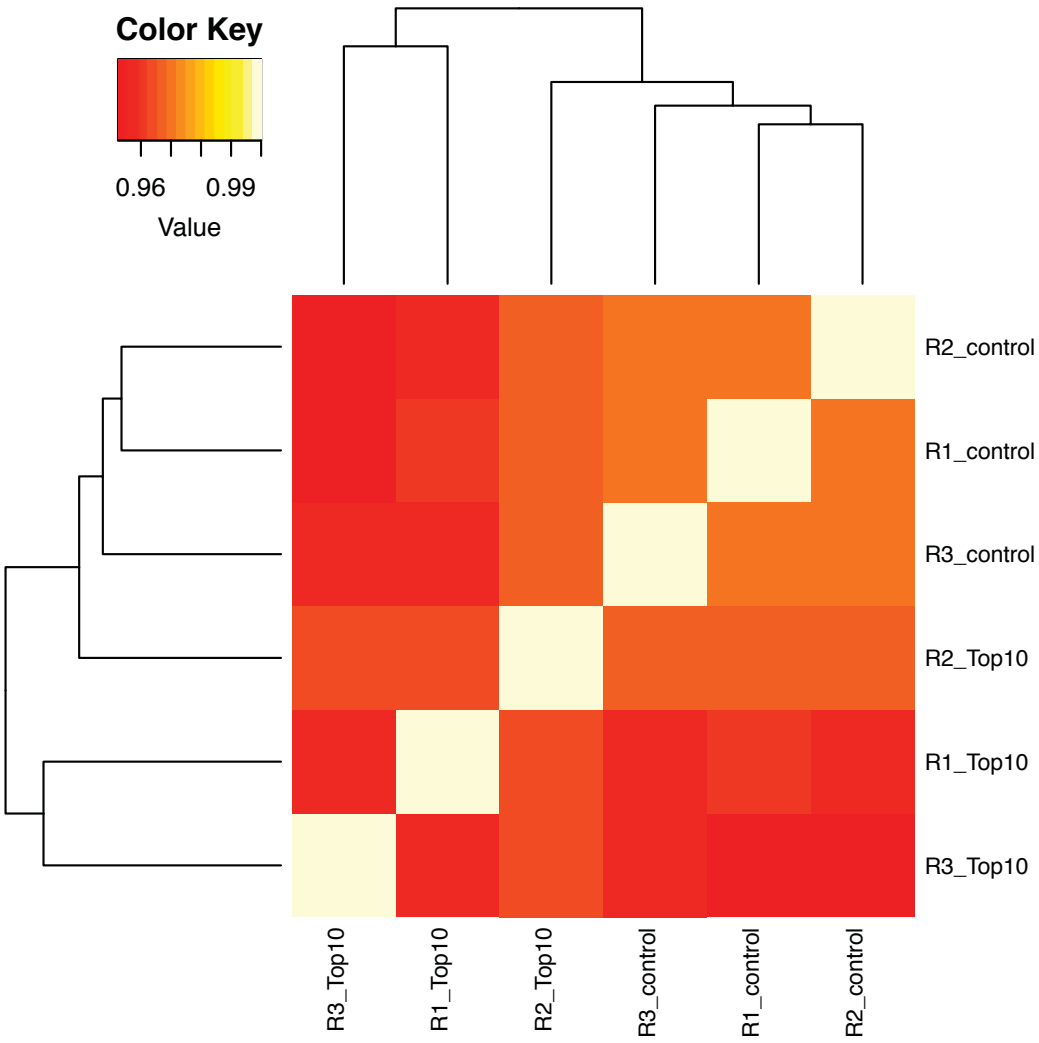

FIGURE 3: Heat map of knockout screen samples, clustered by pairwise PCCs of log_2 –transformed, median-normalized sgRNA read counts. As expected, the 'high yield' and 'control' cell populations each cluster together, indicating that their within-condition sgRNA read count distributions are similar. Here, 'Top10' refers to the top 10% of GFP-expressing cells, which is synonymous with the 'high yield' cell population. Figure generated by the `MAGeCK` software suite[65,66,105] and corresponds to figure 4f of the Sharon *et al.* (2020) study[89], which is under a Creative Commons license.

Comparison of sgRNA Abundance Between High Yield and Control Cell Populations Identifies Putative Antiviral Host Factors

After mapping reads to the Brunello sgRNA library and normalizing the resulting counts across samples, the abundance of each sgRNA is compared between 'high yield' and 'control' conditions using the `test` function of `MAGeCK`, as described in the methods section. As a brief reminder, the strategy used by `MAGeCK` to identify gene knockouts of biological importance is to identify those that are both enriched/depleted between conditions—as mea-

sured by the LFC in sgRNA abundance—and considered to be statistically significant—as tested using the negative binomial distribution[65]. To get an informative view of how the sgRNA abundances differ between conditions—where the median value is taken across replicates—individual log_2 –transformed sgRNA pseudocounts are plotted against each other, as depicted in figure 4.

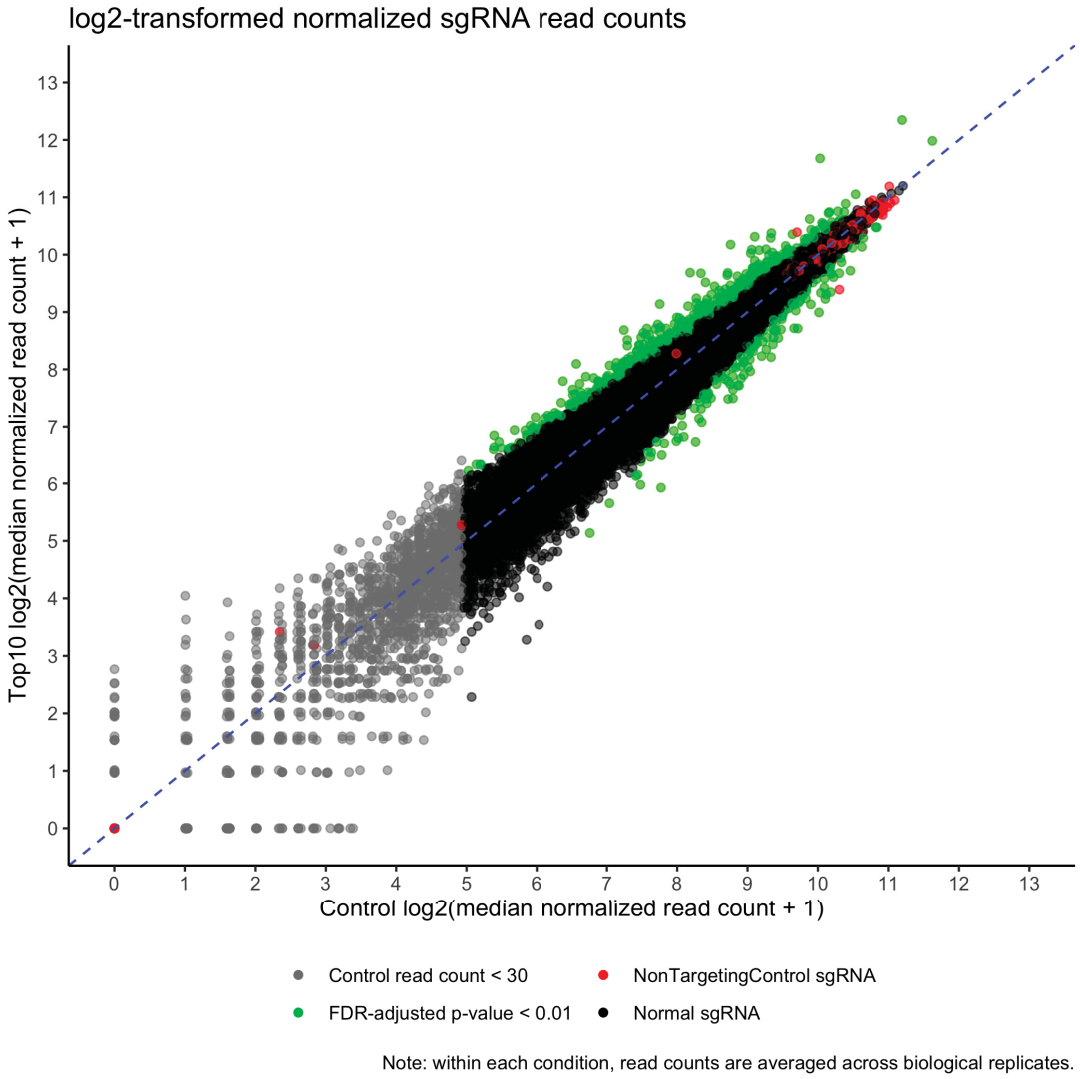

Figure 4: Scatter plot comparing sgRNA abundance between 'high yield' and 'control' cell populations. Abundance of normalized sgRNAs are depicted as the median value across replicates, with a log_2 transformation applied to values with a pseudocount of +1. Each sgRNA has been classified as: having a median read count in the control condition < 30 (n=1,760; grey); a non-targeting control (n=1,000; red); statistically significant with an FDR-adjusted p-value < 0.01 (n=754; green); or, none of the prior classifications (black). The blue dotted line denotes $y = x$, where points above or below this indicate enrichment or depletion of the corresponding sgRNA. Here, 'Top10' refers to the top 10% of GFP-expressing cells, which is synonymous with the 'high yield' cell population. Figure generated using `Julia`[9] and `ggplot2`[112], with data produced by the `MAGeCK` software suite[65,66,105]; it corresponds to figure 5a of the Sharon *et al.* (2020) study[89], which is under a Creative Commons license.

From this, the sgRNAs with median read counts of less than 30 in the control condition

(n=1,760; grey) were removed from subsequent downstream analyses, as they do not have sufficient representation in the population of cells.

Next, sgRNA-level FDR-adjusted p-values and LFCs were converted to the level of the gene using `MAGeCK`, as described in the methods section. To consider the LFC of gene abundance between conditions as a measure of biological significance, a volcano plot was produced—shown in figure 5—which compares individual gene FDR-adjusted p-values to their corresponding LFC. Putative gene hits were considered to be those with a $|LFC| > 0.4$ and an FDR-adjusted p-value < 0.01.

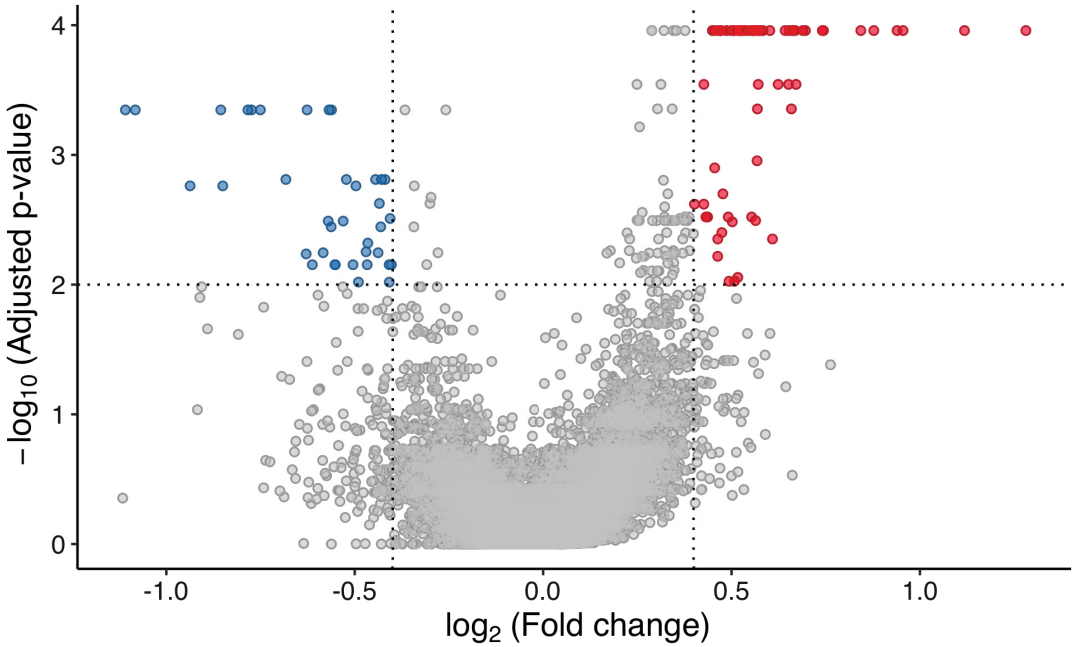

FIGURE 5: Volcano plot depicting significantly enriched and depleted knockout screen gene hits by comparing FDR-adjusted p-values to the LFC in gene abundance between conditions. Vertical dashed lines indicate arbitrarily chosen thresholds for putative gene hits as $|LFC| > 0.4$ and FDR-adjusted p-values < 0.01; genes coloured red ($n = 64$) and blue ($n = 37$) indicate putative anti- & pro-viral gene hits, respectively. Figure generated by the `MAGeCK` software suite[65,66,105] and corresponds to figure 6a of the Sharon *et al.* (2020) study[89], which is under a Creative Commons license.

In total, `MAGeCK test` identified 135 significant genes under the FDR-adjusted p-value threshold of 0.01. With the additional biological significance restriction of $|LFC| > 0.4$, 64 and 37 putative anti- and pro-viral influenza-specific host factors were identified, respectively. As noted by Sharon *et al.* (2020)[89], there is a lack of adequate controls within the screening strategy for the identification of proviral genes; thus, they should be interpreted with caution, especially when used as the basis for other functional analyses,

such as those presented in chapter four. The putative anti- and pro-viral influenza-specific host factors—corresponding to the red and blue gene 'hits' in figure 5—are included in section C of the appendix; there, the significance threshold was lowered to $\alpha = 0.05$ to increase numbers.

In conclusion, individual cells with specific gene knockouts that result in an abnormal level of viral replication—as measured by the expression of GFP-tagged influenza—have been identified computationally through the analysis of SGS reads derived from the Sharon *et al.* (2020)[89] genome-wide knockout screen of HEK-293SF cells. The candidate antiviral host factors identified will be used to further the understanding of host viral defence mechanisms, and, for guiding the development of optimized cell-based vaccine manufacturing platforms.

3.2.3 Summary of Influenza Host Factors Identified in Screening Experiments

To compliment the set of putative antiviral host factors identified in the screen by Sharon *et al.* (2020)[89], a selection of perturbation-based influenza-specific screening studies were found in the literature. Each study identified anti- and/or pro-viral host factors, with each one varying in their motivation for conducting the screen and in their method of perturbation. The curated screen 'hit lists' have all been converted to a common gene identifier for the purpose of comparison and are included in section C of the appendix.

Table 3: A selection of studies that have performed a perturbation-based screening strategy to identify host factors that are involved in the influenza virus life cycle, sorted by year of publication. Although each study was conducted in the form of a screen, it is worth noting that they differ from one another in experimental setup and statistical methods used and thus the comparison of their raw count values is not appropriate. See table for corresponding citations; the curated sets of host factors are provided in section C of the appendix.

First Author	Year	Journal	Perturbation Method	Number of Screen Hits	
				Antiviral Genes	Proviral Genes
Brass[12]	2009	*Cell*	siRNA	4	129
Carette[16]	2009	*Science*	Insertional Mutagenesis	0	2
Shapira[88]	2009	*Cell*	siRNA	176	221
Karlas[56]	2010	*Nature*	siRNA	0	168
Watanabe[111]	2014	*Cell Host & Microbe*	siRNA	34	358
Tripathi[99]	2015	*Cell Host & Microbe*	Meta-analysis	485	1445
Heaton[47]	2017	*Cell Reports*	CRISPR activation	1190	0
Sharon[89]	2020	*Scientific Reports*	CRISPR knockout	64 (89*)	37 (60*)
Total				1953	2360

*Number of genes identified in the screening study by Sharon *et al.* (2020)[89] under a significance level threshold of $\alpha = 0.05$; the values not in brackets correspond to $\alpha = 0.01$.

3.2.4 Comparison of Putative Antiviral Gene Screen Hits

To get a sense of the overlap of *antiviral* gene screen hits identified in perturbation-based genome-wide screening studies of cells infected with the influenza virus, a subset of studies from table 3 were compared by creating a Venn diagram, as shown in figure 6. This depicts their intersections at the level of the gene, where comparisons are made using NCBI gene identifiers; see section C of the appendix for the host factor sets.

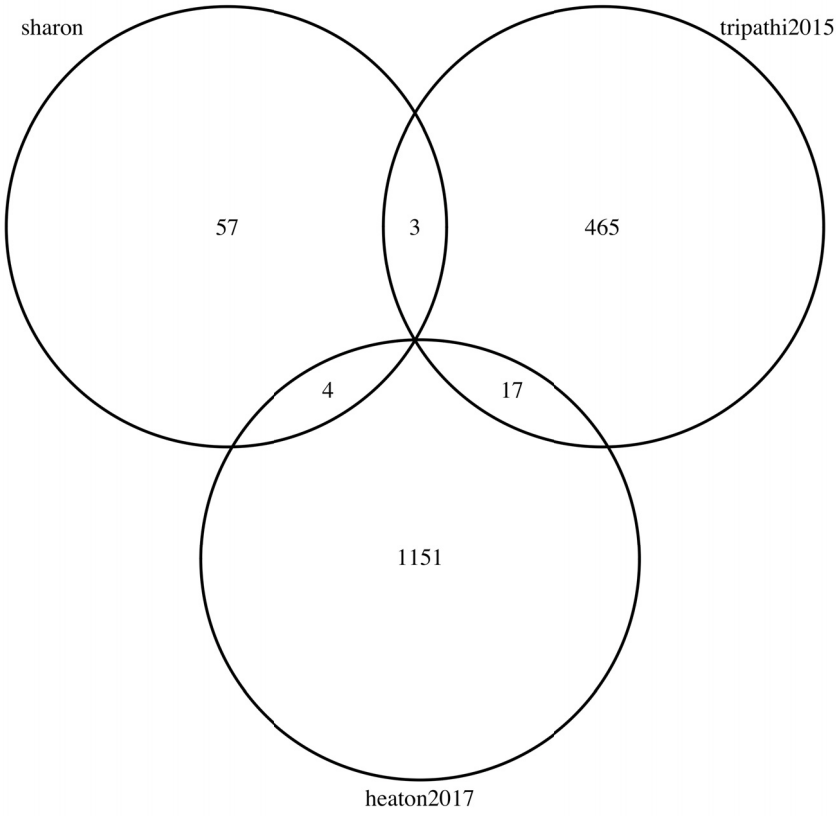

Figure 6: Venn diagram depicting the overlap of putative antiviral gene sets identified in two perturbation-based genome-wide screens of cells infected with influenza[47,89] and one meta-analysis study[99]. The three gene sets correspond to the following studies: Sharon *et al.* (2020)[89] as 'sharon' ($n = 64$; refers to those under a significance level threshold of $\alpha = 0.01$); Heaton *et al.* (2017)[47] as 'heaton2017' ($n = 1190$); Tripathi *et al.* (2015)[99] as 'tripathi2015' ($n = 485$). Refer to table 3 for further details on the studies. A gene-to-gene comparison of screen hits does not necessarily capture the full output of each study, as between-experiment variation exists, and thus the overlaps should be interpreted with caution. Figure created in `R`[84].

A caveat to this analysis is that gene-to-gene comparisons of screen hits do not necessarily capture the full output of each study, as between-experiment variation and other nuances exist; thus the gene set overlaps should be interpreted with caution. Importantly, the studies by Sharon *et al.* (2020)[89] ('sharon') and Heaton *et al.* (2017)[47] ('heaton2017') use CRISPR-mediated technologies to induce gene knockouts and activations, respectively; in contrast, Tripathi *et al.* (2015)[99] conducted a meta-analysis of previously published genome-wide RNA interference (RNAi) screens and integrated these results with protein interaction datasets—one of which was generated as part of their study. Although these differences exist, it does not preclude performing this trivial type of analysis, as it re-inforces the level of care that must be taken when comparing screen-based studies. A few reasons for the lack of congruity between systems-level technologies are given in the

discussion chapter of this thesis (5).

That being said, the main takeaways from the intersections of these putative antiviral gene screen hits can be summarized by calculating relative fold enrichments between each pair of sets (see 3.1.4 for details); this is given in table 4.

Table 4: Fold enrichments of overlapping putative antiviral gene sets identified in two perturbation-based genome-wide screens of cells infected with influenza[47,89] and one meta-analysis study[99]. See section 3.1.4 for calculation details. As noted previously, gene-by-gene comparison of screen hits is only one measure for comparing screens against each other and thus should be interpreted with caution. These calculations are contained in supplemental `S1` of the Sharon *et al.* (2020) study[89] and is under a Creative Commons license.

Gene Set A	Gene Set B	Fold Enrichment(A, B)
Sharon *et al.* (2020)[89]	Heaton *et al.* (2017)[47]	1.1
Sharon *et al.* (2020)[89]	Tripathi *et al.* (2015)[99]	1.9
Heaton *et al.* (2017)[47]	Tripathi *et al.* (2015)[99]	0.6

In conclusion, the putative antiviral genes derived from the selected perturbation-based genome-wide screens of cells infected with influenza—when compared at the level of the gene—have weak, if any, enrichment between each other. This supports further computational or experimental validation of the screen hits; therefore, the next chapter carries out a series of network-based computational analyses on various interactome datasets—using the host factors identified in this chapter as a guide—with the aim of further characterizing host-virus interactions.

Chapter 4: Host-Virus Interactome Analysis

This chapter addresses the third objective of this thesis by using a systems biology approach to further characterize host-virus interactions. This is accomplished by applying a number of network-based computational analyses on several interactome datasets—each of which has distinct properties due to the variation in their underlying interaction types—using the host factors identified in chapter three as a guide.

4.1 Methodology

4.1.1 Interactome Datasets

The interactome datasets used in the network analyses of this chapter are described in the following sections. Each interactome is represented as a graph $G = \{V, E\}$, where V is the set of vertices (genes) and E is the set of edges (interactions) that occur between them. For all datasets used in this thesis, an edge connecting two genes is undirected, meaning that $(a, b) = (b, a)$. The existence of a weight on an edge, which denotes the strength or confidence of an interaction, is variable within the datasets used and as such will be specifically noted.

The Human Reference Interactome (`HuRI` and `HI-union`)

The 'Human Reference Interactome' (`HuRI`) and the 'union of Human Interactomes' (`HI-union`) are primary binary PPI datasets that can be downloaded from the Human Reference Protein Interactome Mapping Project website as part of the study by Luck *et al.* (2020)[67]. These datasets are part of the larger effort to map all human PPIs, as led by the Center for Cancer Systems Biology (CCSB) at the Dana-Farber Cancer Institute. Binary interactions in these datasets are identified using the yeast two-hybrid screening experiment; here, an interaction is detected by proxy through the transcription of a reporter gene, which only occurs if the bait and prey query proteins physically associate. As with all biological experiments, false positives and false negatives do occur; however,

significant effort has been made to reduce the extent of these errors through the use of technical and biological replicates. The `HI-union` dataset is an aggregate of all PPIs that have been identified by protein interaction mapping efforts at the CCSB; as such, `HI-union` is a superset of `HuRI` in both its vertex (V) and edge (E) sets. As this database is based on primary evidence, it is considered as the 'gold-standard' for binary PPIs.

`STRING`: A Database of Known and Predicted Protein-Protein Interactions

The `STRING` database ('Search Tool for Retrieval of Interacting Genes/Proteins') contains a collection of both direct and indirect PPIs that are curated from public sources, as well as computationally predicted[93]. This resource, available online at https://string-db.org/, supports large query sets, produces interactive web-based graphs with various edge weights ('scores') as a measure of confidence for exploration purposes, and provides numerous downstream functional analysis tools. Importantly, the interactions used for the analyses in this thesis were taken from `STRING` version 11.0 and were restricted to *physical* associations, thereby excluding those that relate proteins by function, as is defined by `STRING`.

`COXPRESdb`: An Aggregated Source for Ranked Gene Coexpression Data

`COXPRESdb` is a database of aggregated, ranked gene coexpression data, available online at https://coxpresdb.jp/[77]. Data used in this thesis corresponds to `COXPRESdb` version 7.2 (correlation tables `Hsa-r.c4-0` & `Hsa-u.c2-0`, released on 2019-02-25). Instead of storing physical interactions, it calculates the functional, indirect relationship between two genes based on the correlation of their expression across a large number of transcriptomic experiments; gene expression, in turn, is measured by counting the number of messenger RNA transcripts expressed by a gene under a given environmental condition. As such, the coexpression of two genes is used as a proxy for their indirect interaction; the nature and extent of this relationship is left up to interpretation, and often requires further experimental validation.

4.1.2 Basic Network Properties

Representing Interactome Data as an Adjacency List

Interactome datasets are typically stored in files as a list of interactions in the form (a, b), separated by some delimiter. When working with interactome data in practice, however, it is often more efficient to transform the list of edges into an adjacency list or matrix. By doing so, the computational complexity of common graph operations is significantly reduced. As such, each interactome dataset was read from file as a list of edges and subsequently converted into an adjacency list; this data structure is represented as an array of vertices, with each array element consisting of an array of all of its neighbours within the graph.

Calculating General Network Properties for an Interactome

Calculating general network properties for a graph that is stored as an adjacency list (or matrix) is trivial, and, for many of the common graph operations, computationally efficient. For a brief overview of the graph theory concepts that are used in this chapter, refer to section 2.2.1.

For example, to calculate the degree of a vertex v, one only has to lookup the array of neighbours corresponding to vertex v within the adjacency list and retrieve its length, which is a constant time operation ($\in O(1)$). To calculate the *degree sequence* of a graph G, the degree operation is simply repeated for each vertex in the graph, and the resulting values are sorted (operation is $\in O(n \log n)$).

One property of importance is that of the 'self-loop': it is defined as an edge (a, a) that connects the vertex a to itself. Within the context of an interactome network, this might refer to a protein that physically interacts with itself. Calculating the number of self-loops within a graph G involves iterating over all edges within the graph and checking for equivalence between the two vertices in each edge (operation is $\in O(|E|)$).

Power-law Distribution of Interactome Networks

The probability that a given gene (vertex) participates in k interactions (its degree) within an interactome has been shown to follow a power-law distribution[5]; this means that many empirical interactome datasets approximate a scale-free network, wherein a large number of genes have very few interactions and only a few genes—known as *hubs*—have a large number of interactions[6]. The power-law probability distribution, typically reserved for explaining large values of k, or the tail end of the distribution, is given below:

$$P(k) \approx k^{-\alpha} \tag{4}$$

where $\alpha > 1$. When the *log* function is applied to both sides of this equation, we can plot the following relationship:

$$log(P(k)) \approx log(k^{-\alpha}) \tag{5}$$

If the above relationship is plotted and a straight-line relationship is observed, this indicates that the degree distribution of the interactome under analysis approximates that of a power-law.

One property that arises from the nature of the power-law relation is that it is invariant under scaling of its parameter k; that is, when multiplying the variable k by a constant factor c, the resulting function is a multiple, or scaling, of that factor:

$$P(ck) = (ck)^{-\alpha} = (c^{-\alpha}k^{-\alpha}) = c^{-\alpha}P(k) \tag{6}$$

This property leads to the definition of 'scale-free networks', as already described.

4.1.3 Candidate Gene Set Network Analysis

The methods contained within this section all share the common theme of analyzing a set of query genes—or 'candidate' host factors—within the space of a target interactome

G. By conducting these gene set analyses within the context of an interactome network, systems-level properties may be determined.

Comparing Network Connectivity Between Gene Sets S_1 and S_2

To begin, the problem of quantifying the network connectivity for a single set of genes S within an interactome G is defined as follows:

> Given an undirected graph $G = \{V, E\}$ and a set of vertices S, such that $S \subseteq V$, compute the fraction of vertices in S that have at least one direct neighbour (edge) with a *different* vertex in the same set S; this can be interpreted as testing $N(v)$, the neighbourhood of a vertex v where $v \in S$, for set membership in S. From a biological standpoint, this problem aims to determine how well connected a set of genes is within a particular interaction network. If many of the genes within the query set S have neighbours that are also within S, this may suggest that they are functionally related. In any case, this simple analysis can efficiently test whether or not it is worth further investigating a particular set of genes.

To calculate the observed number of intra-set interactions that occur between genes in the set S within the induced subgraph $G[S]$, as described above, we can use the following equation:

$$\sum_{v \in S} \sum_{x \in N(v)} \begin{cases} 1 \text{ if } (v, x) \in E \\ 0 \text{ otherwise} \end{cases}, \text{ such that } x \in S \tag{7}$$

where $N(v)$ is the neighbourhood of vertex v; that is, the subgraph of G formed by the vertices adjacent to v. To normalize this value, we calculate the maximum number of edges possible for a graph with the same number of vertices. A graph with this property is known as a *complete* graph, where every pair of unique vertices is connected by an edge; the number of edges in a complete graph with $|V| = n$ vertices is given by the equation:

$$\binom{n}{2} = \frac{n(n-1)}{2} \tag{8}$$

At this point, we are now able to compare the relative connectivity of two gene sets S_1 and S_2 within their corresponding induced subgraphs, $G[S_1]$ and $G[S_2]$. To do so, we normalize the observed number of intra-set interactions, calculated in equation 7, by dividing by the maximum number of edges possible for a complete graph with the same number of vertices ($|V| = |S_1|$ or $|S_2|$, respectively), as calculated by equation 8.

Simulated Network Connectivity Enrichment Analysis

The 'simulated network connectivity enrichment analysis' attempts to answer the following question: do a set of genes tend to be enriched for the number of interactions that they have within an interactome?

More formally, this is testing if the induced subgraph $G[S]$, formed by an input set of vertices S and their 1-degree neighbours, has a distribution of vertex degree values that is, on average, larger than the degree distributions of randomly sampled, equivalent-sized sets of vertices from the same graph G.

To test this, two scoring schemes were used to compare between two vertex degree distributions:

1. **Scoring scheme 1**: Given a set S of vertices $\{v_1, v_2, ..., v_n\}$, where $v_i \in V$ and V is the vertex set of graph G, calculate the **arithmetic** mean of vertex degree values:

$$s1(S, G) = \frac{\sum_{i=1}^{n} \deg(v_i)}{n} \tag{9}$$

2. **Scoring scheme 2**: Given a set S of vertices $\{v_1, v_2, ..., v_n\}$, where $v_i \in V$ and V is the vertex set of graph G, calculate the **geometric** mean of vertex degree values:

$$s2(S, G) = \left(\prod_{i=1}^{n} \deg(v_i) \right)^{\frac{1}{n}} \tag{10}$$

Scoring scheme 2 ($s2$) was selected as a way to put less emphasis on values that may be

very large in comparison to the rest of the set. As the vertex degree distribution of HuRI was shown to approximate the power law—where few vertices have large degree values and many are small—this is applicable. The geometric mean can be expressed alternatively in terms of logarithms, where multiplication becomes addition, as shown below:

$$\left(\prod_{i=1}^{n} \deg(v_i)\right)^{\frac{1}{n}} = e^{\left[\frac{1}{n}\sum_{i=1}^{n} ln(\deg(v_i))\right]} \tag{11}$$

This alternative formulation of the geometric mean—which is essentially taking the *arithmetic* mean of the *log* –transformed degrees—puts less weight on extremely large values. This effectively removes any bias introduced by the presence of a few genes that have very large degree values, which, in the context of the biological world, are known as *hub* genes.

Induced Subgraph of Host Factors within the STRING Physical Interactome

For this analysis, the putative anti- and pro-viral host factors identified in the genome-wide knockout screen by Sharon *et al.* (2020)[89] ($n = 89$ and 60, respectively) were both used as input to the STRING database[93]. The STRING network was restricted to physical associations by selecting the 'physical' network type setting.

4.1.4 Gene Coexpression Network Analysis

Genes that have similar expression profiles over a number of different conditions and experiments are defined as being *coexpressed* and thus can be considered to be functionally related. However, in order to compare between experiments and different detection modalities, a framework of normalization, aggregation, and comparison must be defined. This has been done by the COXPRESdb database[77], whose methods and contents are briefly described in the following section. Use of this valuable dataset for the analysis of host factors is also described in detail.

The 'Mutual Rank' Score as a Measure of Gene Coexpression

To begin, it is useful to understand what gene expression data looks like; an example is given below in table 5.

Table 5: Structure of gene expression data. The number of transcripts expressed by a given gene g_i is measured across n samples ($S = \{s_1, s_2, ..., s_n\}$); here, samples may be partitioned into two sets that correspond to two different experimental conditions being tested. Comparison of counts between these two conditions across all genes may lead to the identification of those that are differentially expressed, and thus of interest for further analysis.

	s_1	s_2	...	s_n
g_1				
g_2				
...				
g_m				

`COXPRESdb` calculates a score, referred to as a 'Mutual Rank', for every pair of genes in the human genome; that is, $\binom{n}{2} = \binom{22897}{2} = 262,124,856$ gene pairs (the n here is based on `COXPRESdb`). A single score is derived from two corresponding sets of transcriptomic experiments for the two genes being analyzed, each of which looks like table 5. Importantly, potential variations in environmental conditions and detection modalities exist both within and between gene expression datasets. As such, the transcriptomic datasets for each gene undergo a series of transformations that permit data integration, prior to calculating the PCCs[†] between them; then, the set of correlations for a given gene A against all other genes ($PCC(A, X)$ for $X \in$ All Genes) is ranked, producing a list of rank-based scores (PCC_{rank}). Taking the rank of a correlation coefficient normalizes for variability in sample conditions, choice of method for normalizing gene expression data, and the relative strength of expression required for a gene to perform its function with its partners[76]. The final step forces the scoring system to be symmetric by taking the geometric average of the two ranks for a gene pair (A, B)[76]:

$$\text{MutualRank}(A, B) = \sqrt{PCC_{\text{rank}}(A, B) * PCC_{\text{rank}}(B, A)} \tag{12}$$

Importantly, the *smaller* the 'Mutual Rank' value is, the *greater* the coexpression between two genes.

[†] PCC = Pearson Correlation Coefficient

For all figures created in the corresponding results section, data wrangling and analysis was performed in `Julia`[9] and plots were produced using `ggplot2`[112] in `R`[84].

4.2 Results

4.2.1 Basic Network Properties of the Human Reference Interactome

To get an idea of what the PPI networks `HuRI` and `HI-union` look like, a number of general network properties were calculated; the results are summarized in table 6 and were used to inform subsequent analyses.

Table 6: General network properties of two human interactome datasets. Data wrangling and analysis was performed in `Julia`[9].

Network Property	Interactome Dataset	
	`HuRI`	`HI-union`
Number of vertices ($\|V\|$)	8272	9094
Number of edges ($\|E\|$)	52548	64006
Number of self-loops*	480	764
Mean vertex degree ($\bar{k}$)	12.65	13.99
Geometric mean of vertex degrees	4.65	4.98
Median vertex degree ($\tilde{k}$)	4.0	4.0
$Q1$ of vertex degree values	1	2
$Q3$ of vertex degree values	12	13
s(vertex degree)**	25.53	29.67
Range of vertex degree values	$[1, 499]$	$[1, 641]$

Here, V and E denote the vertex and edge sets of the specified interactome dataset, respectively. The variable k refers to the degree sequence of the vertices in a graph, which is the set of all vertex degree values, sorted in descending order. $Q1$ and $Q3$ denote the first and third quartiles, respectively.

* Self-loops within a graph represent interactions that occur between a given protein and itself, often at the quaternary-structure level; a case example is the heterotetramer hemoglobin, which consists of two α and two β protein subunits that assemble together to form the final protein product ($(\alpha\beta)_2$). When a self-interacting protein is used as both the bait and prey within a yeast two-hybrid screening experiment—such as that used by Luck *et al.* (2020)[67] to generate `HuRI`—the reporter gene will be transcribed, denoting that the protein interacts with itself.

**Function $s()$ is the corrected sample standard deviation.

Based on the relative similarity in network properties between `HuRI` and `HI-union`, `HuRI` was selected as the interactome for which subsequent analyses will be performed on. Apart from being a rather arbitrary decision, this was partially motivated by the fact

that `HI-union` is an aggregation of data from multiple experiments, some of which differ in the specifics of their methodology.

4.2.2 Degree Distribution of `HuRI` Approximates a Power-law Distribution

To further characterize the network properties of `HuRI`—prior to conducting further downstream analyses that are specific to host factors—its vertex degree distribution was explored with various transformations applied in succession. This analysis is important for calculating other network properties, as the expected results, and therefore biological interpretations, can change based on the type of degree distribution exhibited by a given graph.

To begin, a histogram of the vertex degree values for `HuRI` is plotted in figure 7. To give an idea of the density of vertices with small values of k, 98.5% ($\frac{8146}{8272}$) of vertices have degree values in the range $[1, 100]$; as such, the x-axis of figure 7 is restricted to only show values in this interval.

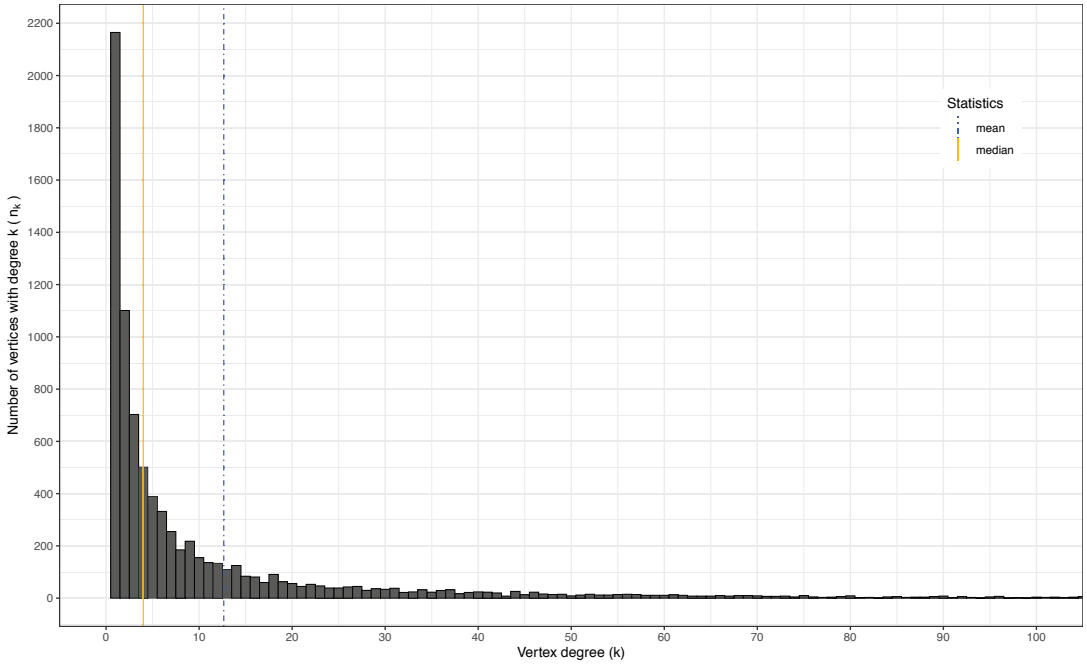

FIGURE 7: Degree histogram of the Human Reference Interactome (`HuRI`). This figure shows the distribution of vertex degree values by plotting the number of vertices (n_k) for each vertex degree value (k). Each bin has a width of 1, meaning that each bin in the histogram corresponds to n_k for *one* value of k. As n_k quickly decreases as k increases, the x-axis has been restricted to only show values in the interval $[1, 100]$ for clarity; in the full set of degree values, k goes all the way up to 499. The interactome dataset `HuRI` was obtained from Luck *et al.* (2020)[67]. Data wrangling and analysis was performed in `Julia`[9]; plots were produced using `ggplot2`[112] in `R`[84].

It is immediately clear by the characteristic heavy upper-tail region of figure 7 that the distribution of degrees within `HuRI` is highly skewed to the right. This observation supports further investigation into this network property: what is the expected number of vertices with degree k within a graph G? That is, how does the probability $P(k)$ relate to the degree k of a vertex? To answer this question, we can plot $P(k)$ against k and observe the effect that *log* –transforming each axis in turn has on their relationship. The motivation behind the choice of the *log* function is to spread out the density of values along the axis in order to aid in visualization. This exploratory analysis of the degree distribution of `HuRI` is given in figure 8; for the same analysis on `HI-union`, see figure 15 in section A of the appendix.

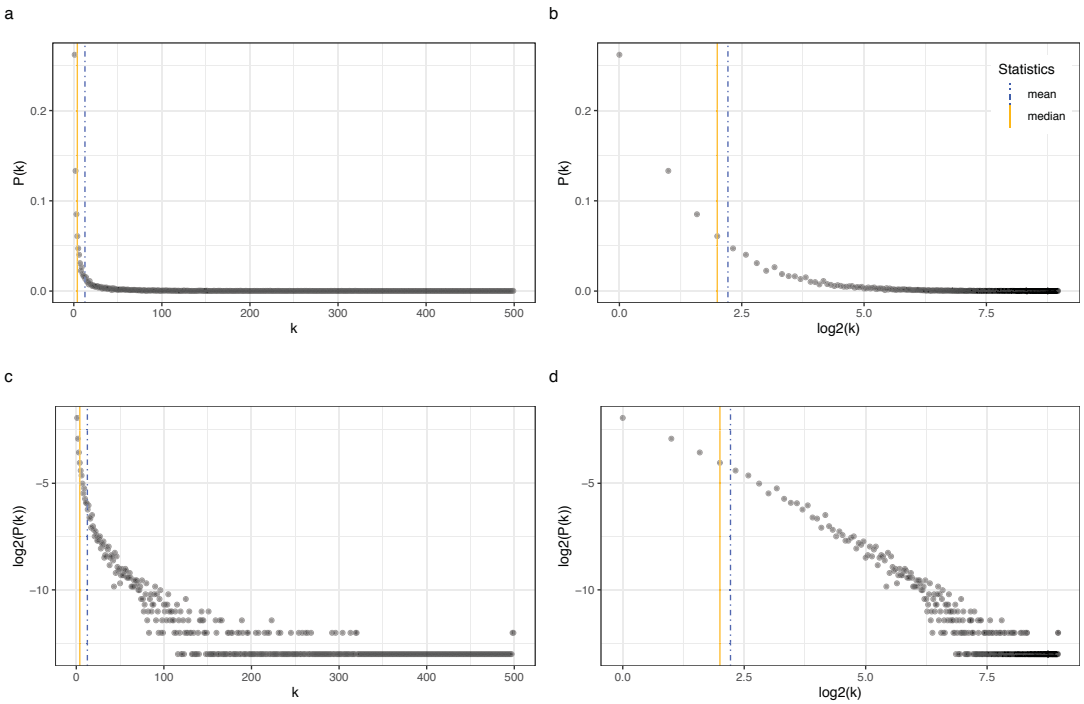

FIGURE 8: Exploration of the degree distribution of the Human Reference Interactome (`HuRI`). Each vertex within the graph, or interactome, represents a human protein-coding gene. The degree of a vertex, denoted as k, is the number of edges, or interactions, that it has with other vertices within the interactome. Each subplot depicts the relationship between $P(k) = \frac{n_k}{|V|}$, the number of vertices with degree k (n_k) divided by the total number of vertices in the graph ($|V|$), and k, the vertex degree value. To ensure that log_2 transformations may be applied to $P(k)$, we make $P(k) > 0, \forall k$ by adding a pseudocount of 1 to the numerator, giving $P(k) = \frac{n_k+1}{|V|}$. To better show the relationship between $P(k)$ and k, scatter plots (**a**, **b**, **c**, and **d**), each with 499 data points based on $|V| = 8272$ degree values, were created with log_2 –transformations applied successively to each axis in turn. The log_2 transformation effectively spreads out the distribution of values along a given axis, making it easier to visualize. Subplot (**d**) has the log_2 transform applied to both axes; here, the depiction of a linear correlation between $log_2(P(k))$ and $log_2(k)$ suggests that `HuRI`, at its current stage of completeness, approximates a power-law relationship[5]. That is, many genes have few interactions, and few genes have many. The interactome dataset `HuRI` was obtained from Luck *et al.* (2020)[67]. Data wrangling and analysis was performed in `Julia`[9]; plots were produced using `ggplot2`[112] in `R`[84].

This exploratory analysis of the degree distribution of `HuRI` suggests that it approximates a power-law distribution. Of interest, the genes with the largest values of k are *MEOX2* (`HGNC:7014`; $k = 498$) and *CYSRT1* (`HGNC:30529`; $k = 499$).

Although the accuracy of estimating the exponent α from the slope of a $log - log$ plot of $P(k)$ vs. k has been debated—as it depends partly on the completeness of the interactome[13,20] —this analysis still shows that the majority of genes within `HuRI` have few connections, and only a few genes have many. This, by itself, is interesting enough to warrant further investigation into `HuRI`; this will be accomplished by probing it with the host factors identified in chapter three. To conclude, the relationship discerned between

$P(k)$ and k is a guiding principle that must be kept in mind when making biological interpretations based on the `HuRI` dataset.

4.2.3 A Candidate Gene Set Network Analysis Reveals Emergent Systems-level Properties

The aim of this section is to use the candidate anti- and pro-influenza host factors identified in the genome-wide knockout screen by Sharon *et al.* (2020)[89] to probe the interactomes `HuRI` and `STRING`; as such, this is referred to as a 'candidate gene set network analysis', as the genes under study require further characterization. Specific focus will be kept on the putative anti-influenza host factors, as the screening strategy used was carried out with the primary motivation of improving cell-based vaccine manufacturing platforms.

Host Factor Interactome Network Properties

To begin, an in-depth analysis of the candidate anti- and pro-viral host factors identified in the Sharon *et al.* (2020) study[89] has been conducted within the target binary PPI interactomes `HuRI` and `HI-union`. A comprehensive set of the network properties resulting from this analysis are given in table 7. For reference, a summary of the interactome datasets that these network properties are based on—both of which were obtained from Luck *et al.* (2020)[67] —can be found in table 6.

Table 7: Interactome network properties for the candidate anti- & pro-viral host factors identified in the genome-wide influenza screen by Sharon *et al.* (2020)[89]. Values are given based on the query gene set used (`S`) and the version of the human binary interactome that was probed (`G`). Data wrangling and analysis was performed in `Julia`[9]. See table footnotes for specific details of network properties.

Query Gene Set (`S`)		Network Property	Interactome Dataset (G)	
			`HuRI`	`HI-union`
Antiviral genes (n=89)	(1)	Fraction of genes in `S` present in interactome	$\frac{41}{89}$ (46.1%)	$\frac{48}{89}$ (53.9%)
	(2.1)	Fraction of intra-set interactions present[†]	$\frac{11}{820}$ (1.34%)	$\frac{13}{1128}$ (1.15%)
	(2.2)	Fraction of intra-set interactions that are self-loops[*]	$\frac{7}{11}$ (63.6%)	$\frac{8}{13}$ (61.5%)
	(2.3)	Fraction of genes in `S` that participate in intra-set interactions, including self-loops[*]	$\frac{14}{41}$ (34.1%)	$\frac{17}{48}$ (35.4%)
	(2.4)	Fraction of genes in `S` that participate in intra-set interactions, **not** including self-loops[*]	$\frac{8}{41}$ (19.5%)	$\frac{10}{48}$ (20.8%)
	(3.1)	Fraction of simulated trials with scores $>$ `score(S, G)`, using scoring scheme 1 (9) and target interactome `G`	$\frac{136,236}{1\times10^6}$ (13.6%)	$\frac{262,632}{1\times10^6}$ (26.3%)
	(3.2)	Fraction of simulated trials with scores $>$ `score(S, G)`, using scoring scheme 2 (10) and target interactome `G`	$\frac{9,832}{1\times10^6}$ (0.98%)	$\frac{56,516}{1\times10^6}$ (5.7%)
Proviral genes (n=60)	(1)	Fraction of genes in `S` present in interactome	$\frac{30}{60}$ (50%)	$\frac{36}{60}$ (60%)
	(2.1)	Fraction of intra-set interactions present[†]	$\frac{4}{435}$ (0.92%)	$\frac{5}{630}$ (0.79%)
	(2.2)	Fraction of intra-set interactions that are self-loops[*]	$\frac{3}{4}$ (75%)	$\frac{4}{5}$ (80%)
	(2.3)	Fraction of genes in `S` that participate in intra-set interactions, including self-loops[*]	$\frac{4}{30}$ (13.3%)	$\frac{5}{36}$ (13.9%)
	(2.4)	Fraction of genes in `S` that participate in intra-set interactions, **not** including self-loops[*]	$\frac{2}{30}$ (6.7%)	$\frac{2}{36}$ (5.6%)
	(3.1)	Fraction of simulated trials with scores $>$ `score(S, G)`, using scoring scheme 1 (9) and target interactome `G`	$\frac{258,324}{1\times10^6}$ (25.8%)	$\frac{372,124}{1\times10^6}$ (37.2%)
	(3.2)	Fraction of simulated trials with scores $>$ `score(S, G)`, using scoring scheme 2 (10) and target interactome `G`	$\frac{166,790}{1\times10^6}$ (16.7%)	$\frac{289,571}{1\times10^6}$ (29.0%)

Network properties have been categorized into three general groups: (1) trivial; (2) network connectivity of the induced subgraph $G[S]$, as described in 'Comparing Network Connectivity Between Gene Sets S_1 and S_2' within section 4.1.3; (3) *simulated* network connectivity of the induced subgraph $G[S]$, as described in 'Simulated Network Connectivity Enrichment Analysis' within section 4.1.3.

*Self-loops within a graph represent interactions that occur between a given protein and itself; see footnotes of table 6 for further details. Where a network property states that it does not include self-loops, the property is calculated on the set of edges with the self-interactions removed ($(a, a) \notin E$).

†The denominator here is the maximum number of edges possible for the induced subgraph $G[S]$; that is, the number of edges in a *complete* graph—where every pair of unique vertices is connected by an edge—of size $|V| = 41$ or 30 for the anti- & pro-viral query gene sets, respectively. For a complete graph with n vertices, there are $\frac{n(n-1)}{2}$ edges. See 'Comparing Network Connectivity Between Gene Sets S_1 and S_2' in section 4.1.3 for specific details of this analysis.

As noted by Sharon *et al.* (2020)[89], any insight gleaned from the analysis of the *proviral* candidate gene set must be interpreted with caution, as the screening strategy lacked adequate controls for their proper identification. Nonetheless, it is still valuable to calculate their network properties, as they can be considered as a 'background' gene set for which comparisons to the putative antiviral host factors may be made.

Through analysis of the candidate host factors within the interactome `HuRI`, as presented in table 7, the following basic properties are observed (organized by network property category):

- `2.1`: Antiviral genes have slightly more intra-set interactions than proviral genes (1.34% vs. 0.92%).
- `2.2`: Both have high proportions of intra-set interactions that are self-loops (63.6% and 75%).
- `2.3` and `2.4`: Significantly more antiviral genes participate in intra-set interactions compared to proviral genes (34.1% vs. 13.3%), including when self-loops are omitted (19.5% vs. 6.7%).

To get a grasp of the values from network property (`2.1`)—the fraction of intra-set interactions that are present—we can consider the following logic:

> What is the probability of choosing two different genes a and b ($a \neq b$) from a set V, where $|V| = n$, such that $a, b \in S$, $S \subset V$ with $|S| << n$, and $(a, b) \in E$? Said in simpler terms, what is the probability that two antiviral

genes are neighbours within an interactome? This turns out to be quite difficult to calculate if the constraints of $|V|$ and $|E|$ are included; so, instead, we keep it simple by asking what the probability is of selecting genes a and b from V without replacement, such that $a, b \in S$:

$$P(a \cap b \mid a, b \in S) = \frac{|S|}{|V|} \times \frac{|S| - 1}{|V| - 1} \tag{13}$$

For the candidate anti- and pro-viral gene sets from Sharon *et al.* (2020)[89], with `HuRI` as the target interactome, the probability of this occurring is, respectively:

$$P(a \cap b \mid a, b \in \text{Antiviral Gene Set}) = \frac{41}{8272} \times \frac{40}{8271} = 0.0024\%$$

$$P(a \cap b \mid a, b \in \text{Proviral Gene Set}) = \frac{30}{8272} \times \frac{29}{8271} = 0.0013\%$$

These probabilities give some perspective on the values calculated for network property `2.1`, where $\frac{11}{820}$ (1.34%) and $\frac{4}{435}$ (0.92%) edges have endpoints that are **both** anti- or proviral genes, respectively.

With regards to network properties `3.1` and `3.2`—calculated as part of the 'simulated network connectivity enrichment analysis'—when genes were sampled randomly without replacement from `HuRI` one million times, 13.6% and 0.98% of these trials were larger than that of the candidate *antiviral* gene set based on the arithmetic (equation 9) and geometric (equation 10) mean scoring schemes, respectively. This can be interpreted as the candidate *antiviral* gene set being enriched for large vertex degree values, or interactions, within the interactome `HuRI` compared to that of randomly sampled sets of genes of the same size. Thus, the candidate *antiviral* genes identified in the genome-wide knockout screen by Sharon *et al.* (2020)[89] tend to have higher degree values than that expected by chance alone.

Furthermore, there is a comparatively larger reduction in percentages between scoring schemes $s1$ (9) and $s2$ (10) for antiviral genes in comparison to that of proviral genes (13.6% $\rightarrow$ 0.98%, difference of 12.6% vs. 25.8% $\rightarrow$ 16.7%, difference of 9.1%). This

indicates that the candidate *antiviral* gene set has a greater proportion of hub genes (those with large degrees), as scoring scheme $s2$ (10) places less weight on degree outliers through the use of the geometric mean function (see methods section for details).

Lastly, all of the observations just described regarding the network properties given in table 7 also hold for the interactome `HI-union`.

Next, figure 9 was created to present the results of the candidate gene set network analysis in a more visual, non-tabular format. This analysis stratifies the distributions of vertex degree values for comparison of the candidate anti- and pro-viral host factors identified in the genome-wide knockout screen by Sharon *et al.* (2020)[89].

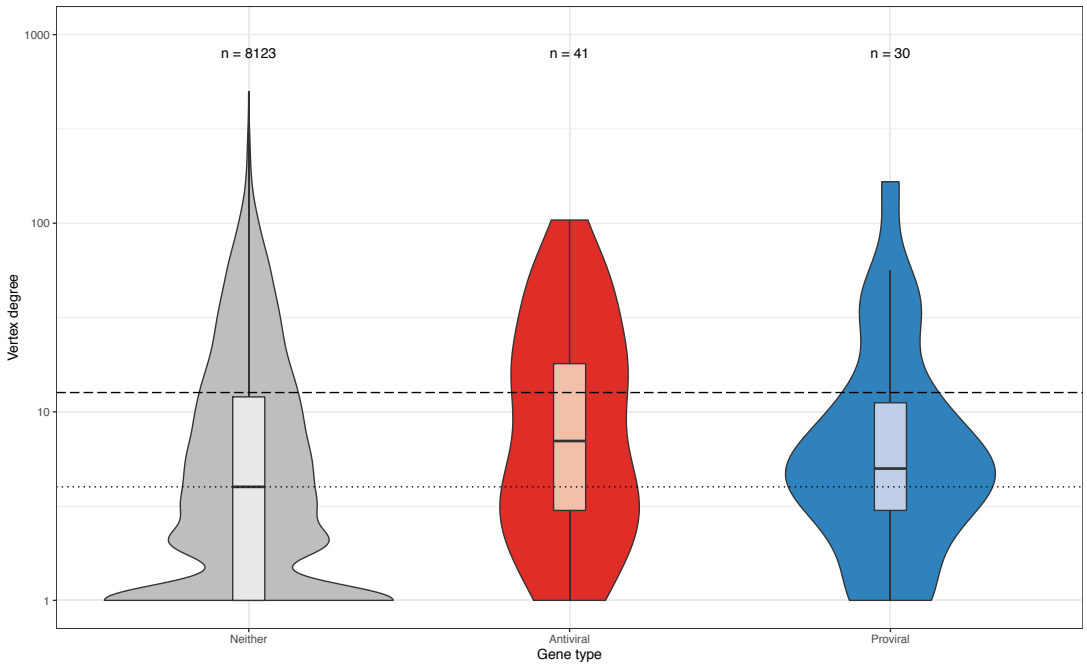

FIGURE 9: `HuRI` log_{10} –transformed vertex degree distributions, stratified by gene type. The anti- & pro-viral host factors depicted (red and blue distributions, respectively) are the putative host factors identified in the genome-wide knockout screen by Sharon *et al.* (2020)[89], restricted to those that are present in `HuRI`. The bottom dotted and upper dashed lines indicate the median and mean vertex degree for all vertices within `HuRI`, respectively. Data wrangling and analysis was performed in `Julia`[9]; plots were produced using `ggplot2`[112] in `R`[84].

Importantly, the candidate antiviral gene distribution (red) is shifted more towards the higher degree values; this supports the observations previously made from the analysis of table 7, which suggests that there is an enrichment in the candidate antiviral set for

genes that interact with each other more (so-called 'hubs'). In conclusion, this supports the hypothesis that antiviral genes tend to work together to—in the form of an immune response—to fight against an invading virus; in contrast, proviral genes tend to be more specific in their functions, with a lower proportion of 'hubs' and fewer intra-set interactions. Viruses that target proviral genes are therefore likely doing so to inhibit a specific function, thereby permitting continuation of their life cycle and the process of infection.

Induced Subgraphs of Host Factors within the STRING Physical Interactome

The STRING networks, created by querying the STRING database with either candidate anti- or pro-viral genes, was restricted to physical associations by selecting the 'physical' network type setting, thereby reducing the possible evidence types for interactions to: curated databases (blue), experimentally determined (pink), textmining (green), and protein homology (purple). The induced subgraph of STRING formed by the candidate antiviral gene set from the Sharon *et al.* (2020) study[89] is given in figure 10.

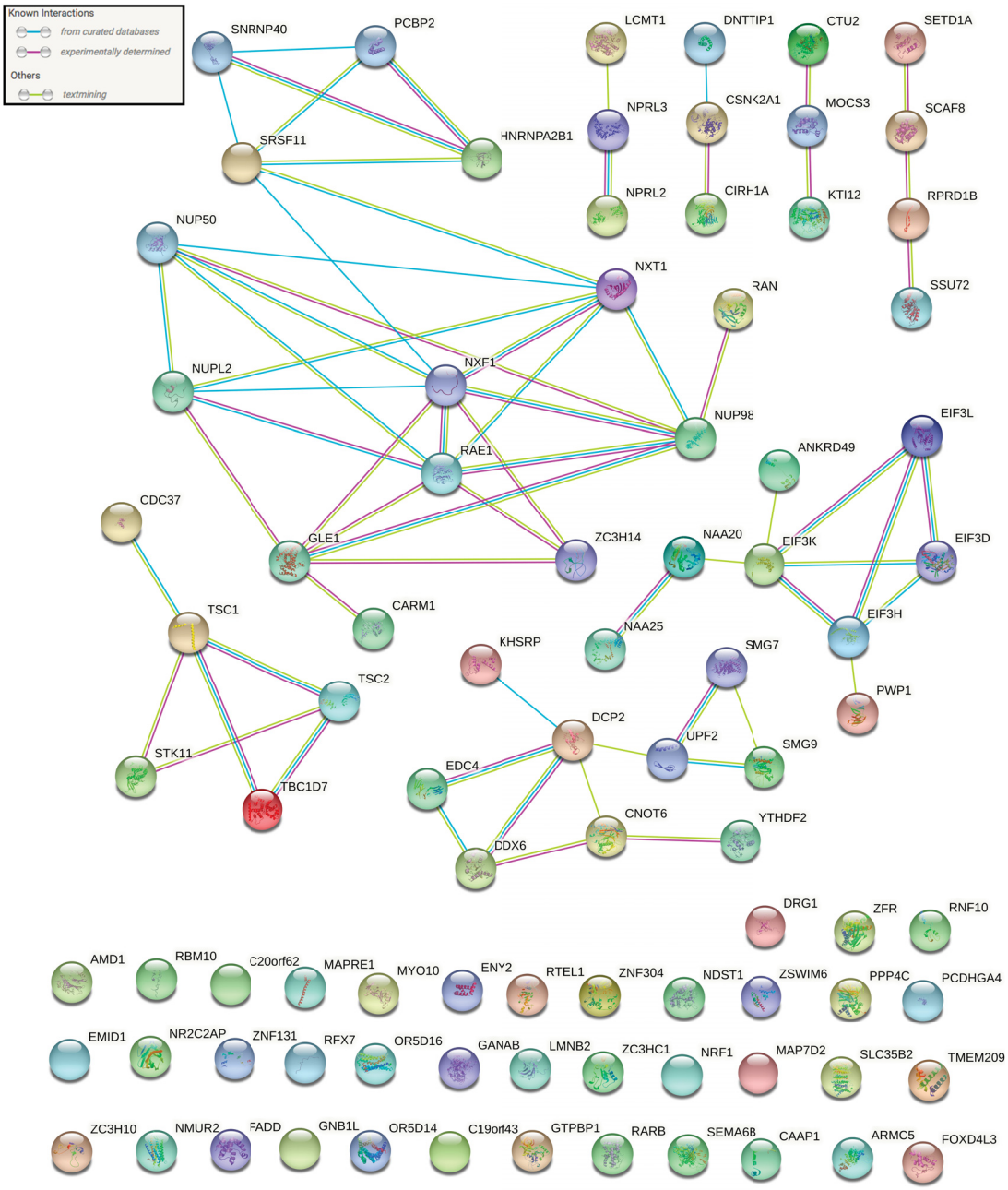

Figure 10: Physical interaction network of the 89 candidate **antiviral** genes identified in Sharon *et al.*'s (2020)[89] genome-wide influenza screen, as computed by `STRING`[93]. 67 physical interactions, where the type of evidence supporting a given interaction is denoted by the colour of the edge (see legend), are displayed between 88 vertices, with an average vertex degree of 1.52; these numbers are based on the inclusion of only physical—and not functional—associations ('physical' network type selected in `STRING` settings, minimum interaction score of 0.4). As calculated by `STRING`, the query antiviral gene set has an enrichment for interactions between themselves, as only 19 edges are expected to occur by chance from a random network of the same size with genes drawn from the entire human genome. The missing gene *DPH3P1* (`HGNC:16136`) was not recognized by `STRING`, as evidence for its existence as an expressed protein in common databases is lacking. Network produced using `STRING` version 11.0[93]; permanent link to network: https://version-11-0b.string-db.org/cgi/network?networkId=bwurWAxHj381.

The molecular function Gene Ontology (GO) term 'N6-methyladenosine-containing RNA binding' (GO:1990247) is enriched within this network—with a strength of 1.8 at a

FDR of 0.0255 —by the presence of genes *HNRNPA2B1* (`HGNC:5033`) and *YTHDF2* (`HGNC:31675`); here, strength is computed as $log_{10}(\frac{observed}{expected})$, where the 'expected' term is defined as the number of genes with the corresponding GO term that one would expect to find within a random network of the same size[93].

This enrichment can be partially explained, as the influenza A virus has been shown to express RNAs with methylated adenosines on the N^6 position (denoted m^{6A}), and removal of this methylation mark significantly decreases viral gene expression and replication; in the opposite case, over-expression of the m^6-A 'reader' *YTHDF2* results in an increase in viral gene expression and replication[21]. Furthermore, Winkler *et al.* (2019) found that deletion of *YTHDF2* after viral infection resulted in an increase in expression of interferon-stimulated genes, which collectively act to inhibit viral replication[113]. As this is the opposite effect of what one would expect from a putative antiviral gene, this is likely a false positive hit in the Sharon *et al.* (2020)[89] screen.

Similar to that of figure 10, the induced subgraph of `STRING` formed by the candidate proviral gene set from the Sharon *et al.* (2020) study[89] is given in figure 11.

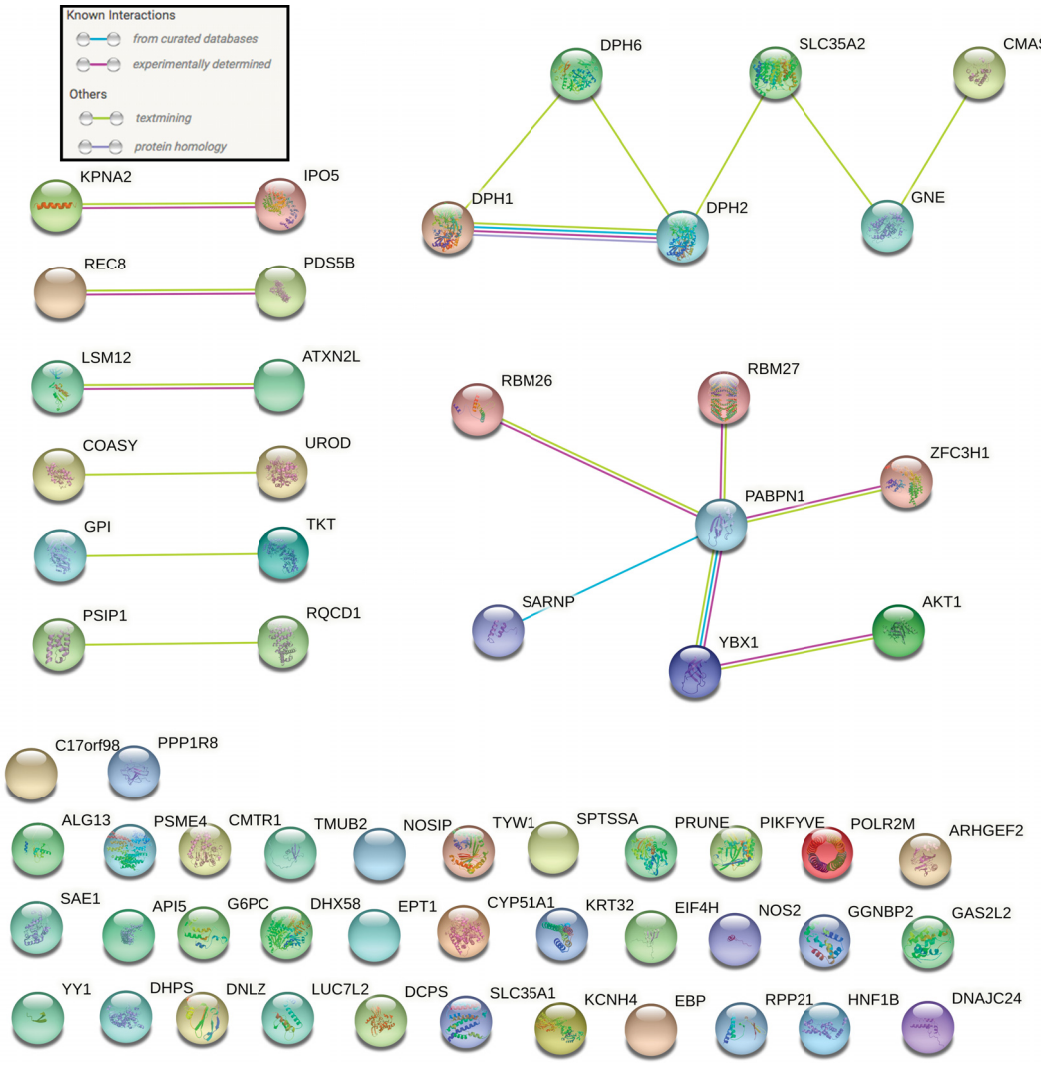

Figure 11: Physical interaction network of the 60 candidate **proviral** genes identified in Sharon *et al.*'s (2020)[89] genome-wide influenza screen, as computed by `STRING`[93]. 18 physical interactions, where the type of evidence supporting a given interaction is denoted by the colour of the edge (see legend), are displayed between 60 vertices, with an average vertex degree of 0.6; these numbers are based on the inclusion of only physical—and not functional—associations ('physical' network type selected in `STRING` settings, minimum interaction score of 0.4). As calculated by `STRING`, the query proviral gene set has an enrichment for interactions between themselves, as only 5 edges are expected to occur by chance from a random network of the same size with genes drawn from the entire human genome. Network produced using `STRING` version 11.0[93]; permanent link to network: https://version-11-0b.string-db.org/cgi/network?networkId=beMuD57KN5Ei.

Both of the presented induced subgraphs of the `STRING` database provide a helpful graphical representation of host factors and how they relate to one another. This permits critical analysis of their interactions, and, as such, may guide in the identification of 'druggable' proviral targets that are capable of mitigating active influenza infections, or, for the proposal of a set of candidate antiviral genes that may be perturbed for optimization of vaccine cell lines.

4.2.4 Host Factors in Gene Coexpression Space

As an alternative to exploring host factors in the space of binary PPIs, like that of HuRI, gene coexpression values from the COXPRESdb[77] database were considered. Gene coexpression is a form of interaction between genes that is considered to be the most permissive type, with binary PPIs being the most restrictive. As such, the analysis of host factors within the gene coexpression space is both unique and interesting.

In all of the figures presented in this section, the query set of host factors used are those that were curated from the literature, as provided in table 1. The advantage of using this set of host factors is that each one has literature-based evidence of interacting with a virus and thus has been classified according to the 'host factor classification scheme' presented in figure 1. However, the disadvantage of their use is that the host factors are not specific to that of the influenza virus; nonetheless, this provides valuable insight into the general nature of host factors, irrespective of their associated virus.

To begin, the distribution of 'Mutual Rank' scores corresponding to all pairs of genes ($n = 1035$) from the complete set of literature-curated host factors ($N = 46$, which excludes *Fv1*) was plotted alongside a randomly sampled (without replacement) set of genes ($N = 46$, with $n = 1035$ gene pairs). Then, to remove any doubt, the background set of genes was randomly sampled 100 times from the set of all genes in COXPRESdb, and for each iteration, the mean gene coexpression 'Mutual Rank' value—corresponding to all gene pairs—was calculated. This analysis is given in figure 12. A violin plot of this same figure, without the 100 randomly sampled sets of background genes, is given in figure 16 within section B of the appendix.

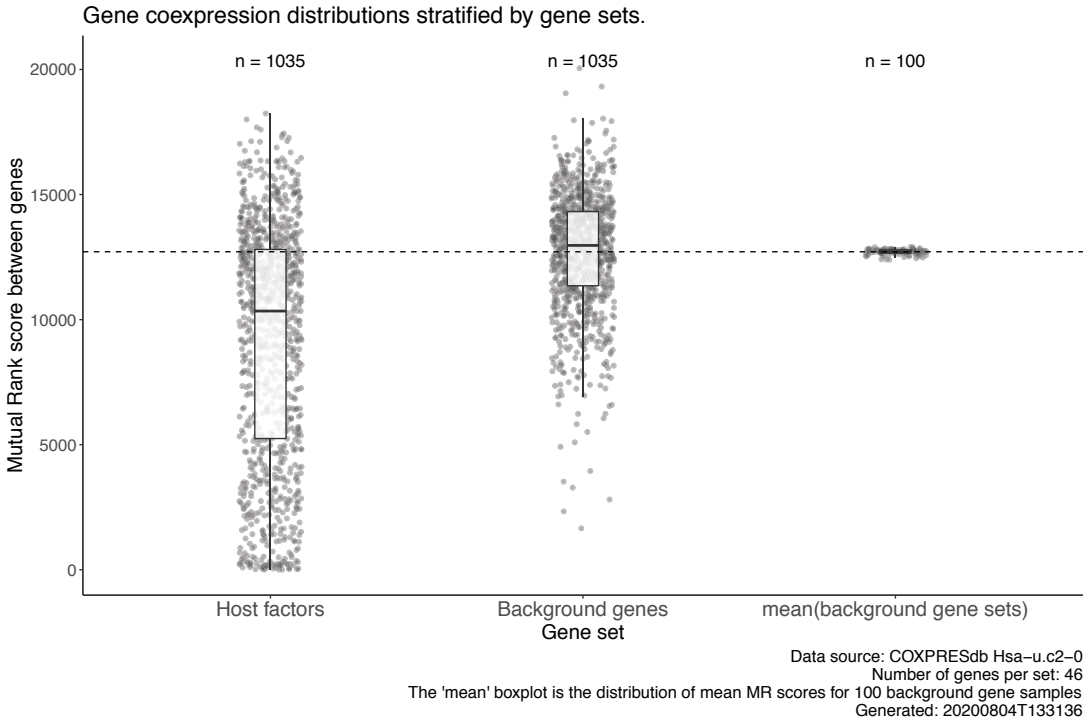

FIGURE 12: Distribution of `COXPRESdb` 'Mutual Rank' scores corresponding to all pairs of genes ($n = 1035$) from the complete set of literature-curated host factors ($N = 46$), as contained in table 1, is plotted alongside a randomly sampled (without replacement) set of genes ($N = 46$, with $n = 1035$ gene pairs). The additional 'mean' distribution contains the mean 'Mutual Rank' value from 100 randomly sampled sets of background genes. The black dotted line indicates the mean 'Mutual Rank' score for all values in `COXPRESdb`.

This indicates that the literature-curated host factors have an enrichment for coexpression between each other—as indicated by the lower 'Mutual Rank' distribution—in comparison to that of randomly sampled background gene sets.

Next, the same type of analysis was conducted, except that the host factors were stratified as either 'antiviral' or 'proviral' based on their manually annotated classification; this is given in figure 13.

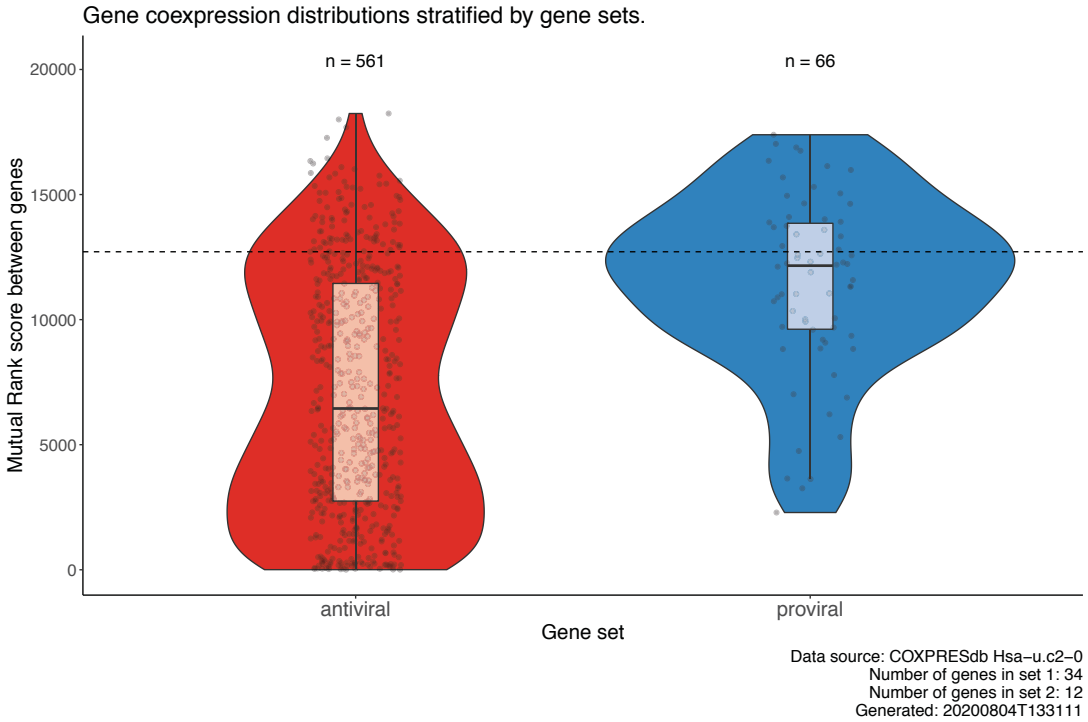

FIGURE 13: Distribution of `COXPRESdb` 'Mutual Rank' scores, stratified by antiviral ($N = 34$, with $n = 561$ gene pairs) and proviral ($N = 12$ with $n = 66$ gene pairs) host factor types. Annotations were done manually based on evidence from the literature, as given in table 1. The black dotted line indicates the mean 'Mutual Rank' score for all values in `COXPRESdb`.

From this, it can be posited that there is a subset of antiviral genes—perhaps those related to the innate immune system—that are significantly enriched for coexpression between each other; this is indicated by the smaller median value and the peak in the antiviral distribution (red) at the lower end of the 'Mutual Rank' scale. However, further curation of a wider selection of proviral genes may have to be done for a more accurate comparison to be made.

Lastly, the distribution of 'Mutual Rank' scores was further stratified by considering the complete set of host factor types, as defined by the classification scheme given in figure 1; the result of this analysis is provided in figure 14.

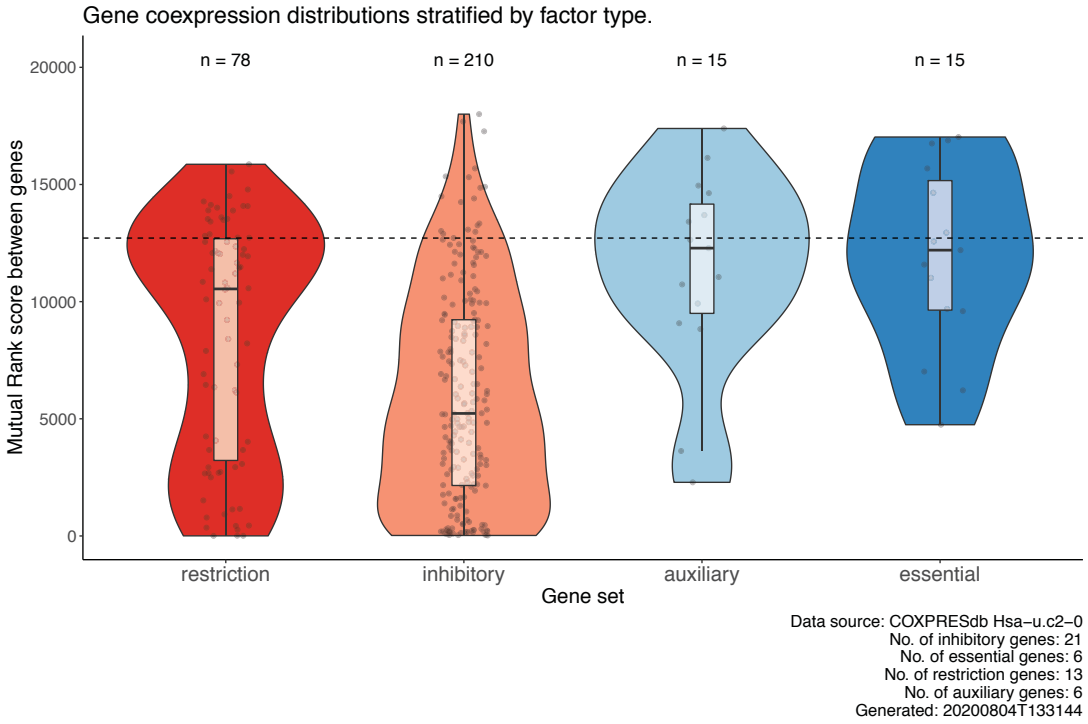

FIGURE 14: Distribution of `COXPRESdb` 'Mutual Rank' scores, stratified by restriction ($N = 13$, with $n = 78$ gene pairs), inhibitory ($N = 21$, with $n = 210$ gene pairs), auxiliary ($N = 6$, with $n = 15$ gene pairs), and essential ($N = 6$ with $n = 15$ gene pairs) host factor types. Annotations were done manually based on evidence from the literature, as given in table 1. The black dotted line indicates the mean 'Mutual Rank' score for all values in `COXPRESdb`. Note that the low numbers corresponding to auxiliary and essential host factor types may skew the graphical depiction of the distribution.

By stratifying the distribution of 'Mutual Rank' scores by the complete set of host factor types, the relative utility of the classification scheme given in figure 1 was tested. From this, the antiviral host factors (restriction and inhibitory) have coexpression distributions that are lower than that of proviral host factors (auxiliary and essential); this was also supported by figure 13. Of interest, the inhibitory host factors seem to have a significant number of highly-coexpressed genes, as indicated by the skewed distribution; these may be genes that are part of the innate immune system, which act together to initiate the immune response cascade and thus may not fully inhibit viral infection on their own. In contrast, the set of restriction host factors were annotated as such because of their ability to completely abolish virus production; therefore, the function of these antiviral genes may be more specific, in that they directly interact with viral components to restrict infection. However, it may also be the case that some of the host factors were annotated incorrectly within their anti- or pro-viral superset.

Chapter 5: Discussion

5.1 Improving and Extending Host Factor Curation Efforts

The large-scale curation of genes along with their associated functions has been going on for many decades, with the publishing of the draft sequence of the human genome in 2001 marking a significant turning point. This opened the door to a myriad of functional genomic experiment types—such as the yeast two-hybrid method and perturbation-based screening—along with a concomitant increase in the generation and storage of high-throughput biological data. As such, a major goal of contemporary biology is to integrate and make sense of all of this data in order to move towards a systems-level understanding of the cell. This was a large motivation for chapter three of this thesis, in which a classification scheme for host factors—presented in figure 1—was proposed as a general framework to reason about host-virus interactions. The importance of having specific definitions for host factors has been made clear, as there is currently ambiguity in the terminology used in the literature, which makes manual curation a particularly laborious task. Moreover, as numerous experiments are able to detect host-virus interactions in many different ways, it very quickly becomes overwhelming to try to parse through and make sense of heterogeneous sources of evidence. As such, many biological databases—such as `UniProt`, `BioGRID`, `IntAct`, and `STRING`—are moving increasingly towards the annotation of gene function using automated pipelines that mine the literature for specific key words, often with a manual verification step in place. Therefore, by using curation techniques that have already been well established, along with the appropriate implementation of database and software tools, more virus-specific host factors may be derived from the literature. Ultimately, the ideal goal of host factor curation efforts should be to move towards the identification and complete characterization of *all* host factors and their interactions, irrespective of the virus type.

A secondary goal of host factor curation efforts is to know the stage of the virus life cycle

that a given host or viral factor functions in. This would enable the ability to integrate this valuable information into the bioprocess of cell-based vaccine platforms, thereby unlocking a new avenue for engineering in which host factors may be perturbed at very specific time points as a way of optimizing their effect. Moreover, knowing when specific proviral host factors function during the course of a viral infection would permit optimization of potential drugs that inhibit them, resulting in improved therapeutic outcomes. The 'Virus Pathogen Resource'[2] and its associated 'Influenza Research Database'[118] are two prominent resources that have attempted to do this by carrying out numerous viral infection-induced time-course transcriptomics experiments. By characterizing host gene expression patterns over time upon infection with specific viruses, one may up- or down-regulate genes at specific points in cell culture platforms to attempt to optimize viral output.

Moreover, the definition of a host factor may be extended to that of other functional components of the cell, such as noncoding DNA, post-translational modifications, and metabolites. Of specific interest is that of the so-called 'dark matter' of the human genome, which has since been shown to encode many functional molecules and regulatory elements that have diverse roles within the cell. This has the potential to open up the 'flood gates' for the number of possible modifications that can be made to a cell; however, as the vast majority of the noncoding region of the genome remains a mystery in terms of its function, its potential for use is currently limited. As current genome-wide screening techniques predominantly target protein-coding regions of the genome, expanding the available repertoire of genomic perturbations would have a significant impact on the potential for discovery.

5.2 Variations in the Output of High-Throughput Screening Experiments

As explored in section 3.2.4 with figure 6 and table 4, there seems to be an apparent lack of congruity in the output of perturbation-based screening experiments. This has also

been observed elsewhere, especially in RNAi-based screens, where a significant amount of variation in screen output has been observed in the detection of influenza[18] and HIV[81] host factors. Although the comparison between screens presented in this thesis was trivial and only considered the matching of gene identifiers as a measure of screen overlap, there are a few underlying reasons for why this occurs—not only in perturbation-based screens, but also in other high-throughput technologies. First and foremost is the presence of false positives and false negatives; with any biological experiment, these are almost unavoidable. Furthermore, Tripathi *et al.* (2015) proposed that a general lack of congruity in the output of systems-level technologies can be attributed to: perturbation methods having a tendency to have some level of off-target activity; the lack of complete silencing within RNAi experiments; and, issues with limits of detection and the presence of non-specific binding in AP-MS experiments[99]. Therefore, further technology improvements and methods of integration must be developed to overcome this, as the current variation observed in the output of genome-wide screening strategies complicates the pursuit of conclusive results.

5.3 Further Characterization of Host-Virus Interactions using Techniques from Systems Biology

Despite the comprehensive analysis of host factors within the context of various interactome datasets presented in chapter four, further characterizing can—and should—be done. Firstly, to get a better sense of the extent of perturbation that the influenza virus has on the human interactome, within-host interaction data can be integrated with virus-host and virus-virus interaction data; these types of datasets can be found in the `HPIDB3.0` database[3], among others. By doing so, host-virus interactions may be further characterized at a whole new level of resolution, whereby individual host and viral factors can be specifically engineered through targeted perturbation for use in numerous applications. To extend this even further, one may consider integrating in protein structural information to host-virus interactomes, thereby improving the resolution even further; efforts have previously been made to do this, each of which has produced interesting results[17,33,36].

Furthermore, techniques for the prediction of host-virus interactions are also being developed[61,75], which may improve the response efforts to novel viruses that may arise in the future. A recent example of this is with SARS-CoV-2 and its associated COVID-19, for which comprehensive interactome datasets already exist[39] and efforts have been made to repurpose antiviral drugs through integrative network analyses[119].

A critical limitation of extracting meaningful insight from host-virus interactomes, or any big dataset, is where to begin to look. Expert knowledge of a domain can greatly assist in this, but gaps may still remain, especially when tackling systems-scale biological problems. More and more, the limitation on novel discoveries is not on the generation of data, but rather on the question that is being asked and the choice of methods for data analysis. As such, the comparison of host-influenza interactomes, followed by functional validation of interesting hits, is another powerful technique for the identification of novel interactions[106] from large biological datasets. An advantage of this type of analysis is that it can be conducted on datasets that already exist to produce new insights. By varying the influenza virus strain used, one can identify common mechanisms or interactions of interest that may be taken advantage of for therapeutic intervention, or, engineered to increase viral production in cell-based vaccine manufacturing platforms. In either case, it shows the importance of systematically comparing high-throughput datasets to discover commonalities or disparities that arise from variation inherent in the biological world. Although determining which interactions are specific to a given strain of influenza and which are not—or even which are of importance—is a difficult task given the scale of data under analysis, new approaches that are both efficient and accurate are being developed to accomplish this[61,119].

When considering the integration of interaction datasets for which the types of interactions differ, careful consideration must be placed on the biological interpretations that are made. For example, one notable caveat in the analysis of gene coexpression data in section 4.2.4 is that it only reflects the regulation of genes at the level of the transcriptome; any other form of regulation or control is lost. This can be addressed to some extent by incorporating other types of interactions to bolster the available sources of evidence, much like that of the `STRING` database[93]; however, as already mentioned, caution in the interpretation of

results from this type of network is required. As such, one may consider using high-quality interaction data that are deemed to be 'gold standard'—such as that of the `HuRI` dataset from Luck *et al.* (2020)[67] —as a basis for the interpretation of other types of interactions. These ideas thus support the extension of network analyses to integrate in other datasets—or combinations of other datasets—to obtain a broader understanding of the cell.

Goodacre *et al.* (2020) describes the perturbation of host factors by viruses as 'shells' of viral interactions that 'reach' into the human interactome[38]. This perspective is enlightening, as it emphasizes that both viral and host factors contribute to the pathogenicity of a given virus; this is especially the case for pandemic strains of influenza, which tend to be particularly pathogenic and deadly to its host by virtue of certain mutations within its HA and NA surface proteins[34]. Therefore, knowledge of which factors do what, and how certain host or influenza mutations confer increased virulence, is incredibly important for reducing its burden of disease. As is a recurrent theme of this thesis, the use of systems-level approaches for understanding the mechanisms of viral infection and host-virus interactions is a natural choice. A series of studies by Watanabe *et al.* (2010, 2014) has experimentally identified and critically analyzed many of the host factors and networks that the influenza virus utilizes for the completion of its life cycle[110,111]. A comprehensive understanding of the host-virus interactome for a given species of virus permits the application of targeted modulations of subcellular networks for either mitigating active viral infections, or, for improving viral titre for use in cell-based vaccine production platforms. Importantly, candidate sets of host factor targets may be reduced in size—in order to minimize disruption of cellular systems while also achieving the maximum impact—only if there is a thorough understanding of host-virus interactions to begin with. Therefore, there is sufficient motivation to conduct more high-throughput screening experiments, such as those performed by Sharon *et al.* (2020)[89] and others (see table 3), to further characterize host factors.

Chapter 6: Conclusion

To conclude, this thesis has presented a general framework for the detection, understanding, classification, and characterization of host factors; the subsequent integration of this information with techniques from systems biology and graph theory provided a unique approach to further characterize these host factors, especially with regards to how they function within host-virus interactions.

This thesis was introduced by emphasizing the ubiquitous influence of viruses throughout biological history; this fact alone supports the ever-increasing need to understand the nature of the interactions that viruses have with their hosts. In doing so, the increased knowledge of host factors permits their potential use in mitigating active viral infections, or, in improving upon existing cell-based vaccine manufacturing platforms. Specific aims were given for how this thesis approached this problem, including: (1) the curation, annotation, and analysis of host factors derived from the literature, as introduced in chapter two and carried out in parts of chapter three; (2) the computational analysis of the output from a genome-wide knockout screen of HEK-293SF cells infected with influenza for the identification of antiviral host factors, also included in chapter three; and, (3) the integrated analysis of these host factors within the space of various interactome datasets to further characterize the nature of their host-virus interactions, as presented in chapter four.

Bibliography

1. E. E. Ackerman, E. Kawakami, M. Katoh, T. Watanabe, S. Watanabe, Y. Tomita, T. J. Lopes, Y. Matsuoka, H. Kitano, J. E. Shoemaker, and Y. Kawaoka. Network-Guided Discovery of Influenza Virus Replication Host Factors. *mBio*, 9(6), 2018. ISSN 21507511. doi: 10.1128/mBio.02002-18.

2. B. D. Aevermann, B. E. Pickett, S. Kumar, E. B. Klem, S. Agnihothram, P. S. Askovich, A. Bankhead, M. Bolles, V. Carter, J. Chang, T. R. W. Clauss, P. Dash, A. H. Diercks, A. J. Eisfeld, A. Ellis, S. Fan, M. T. Ferris, L. E. Gralinski, R. R. Green, M. A. Gritsenko, M. Hatta, R. A. Heegel, J. M. Jacobs, S. Jeng, L. Josset, S. M. Kaiser, S. Kelly, G. L. Law, C. Li, J. Li, C. Long, M. L. Luna, M. Matzke, J. McDermott, V. Menachery, T. O. Metz, H. Mitchell, M. E. Monroe, G. Navarro, G. Neumann, R. L. Podyminogin, S. O. Purvine, C. M. Rosenberger, C. J. Sanders, A. A. Schepmoes, A. K. Shukla, A. Sims, P. Sova, V. C. Tam, N. Tchitchek, P. G. Thomas, S. C. Tilton, A. Totura, J. Wang, B.-J. Webb-Robertson, J. Wen, J. M. Weiss, F. Yang, B. Yount, Q. Zhang, S. McWeeney, R. D. Smith, K. M. Waters, Y. Kawaoka, R. Baric, A. Aderem, M. G. Katze, and R. H. Scheuermann. A comprehensive collection of systems biology data characterizing the host response to viral infection. *Scientific Data*, 1(1):140033, 2014. ISSN 2052-4463. doi: 10.1038/sdata.2014.33.

3. M. G. Ammari, C. R. Gresham, F. M. McCarthy, and B. Nanduri. HPIDB 2.0: a curated database for host–pathogen interactions. *Database (Oxford)*, 2016:baw103, 2016. ISSN 1758-0463. doi: 10.1093/database/baw103.

4. Y. S. Babu, P. Chand, S. Bantia, P. Kotian, A. Dehghani, Y. El-Kattan, T.-H. Lin, T. L. Hutchison, A. J. Elliott, C. D. Parker, S. L. Ananth, L. L. Horn, G. W. Laver, and J. A. Montgomery. BCX-1812 (RWJ-270201): Discovery of a Novel, Highly Potent, Orally Active, and Selective Influenza Neuraminidase Inhibitor through Structure-Based Drug Design. *Journal of Medicinal Chemistry*, 43(19):3482–3486, 2000. ISSN 0022-2623. doi: 10.1021/jm0002679.

5. A.-L. Barabási and R. Albert. Emergence of Scaling in Random Networks. *Science*,

286(5439):509 – 512, 1999. doi: 10.1126/science.286.5439.509.

6. A.-L. Barabási and Z. N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004. ISSN 1471-0064. doi: 10.1038/nrg1272.

7. A. L. Barabási, N. Gulbahce, and J. Loscalzo. Network medicine: A network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68, 2011. ISSN 14710056. doi: 10.1038/nrg2918.

8. Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. ISSN 00359246. doi: 10.1111/j.2517-6161.1995.tb02031.x.

9. J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. Julia: A Fresh Approach to Numerical Computing. *SIAM Review*, 59(1):65–98, 2017. ISSN 0036-1445. doi: 10.1137/141000671.

10. D. Bojkova, K. Klann, B. Koch, M. Widera, D. Krause, S. Ciesek, J. Cinatl, and C. Münch. Proteomics of SARS-CoV-2-infected host cells reveals therapy targets. *Nature*, 583(7816):469–472, 2020. ISSN 14764687. doi: 10.1038/s41586-020-2332-7.

11. N. M. Bouvier and P. Palese. The biology of influenza viruses. *Vaccine*, 26:D49–D53, 2008. ISSN 0264-410X. doi: 10.1016/j.vaccine.2008.07.039.

12. A. L. Brass, I.-C. Huang, Y. Benita, S. P. John, M. N. Krishnan, E. M. Feeley, B. J. Ryan, J. L. Weyer, L. van der Weyden, E. Fikrig, D. J. Adams, R. J. Xavier, M. Farzan, and S. J. Elledge. The IFITM Proteins Mediate Cellular Resistance to Influenza A H1N1 Virus, West Nile Virus, and Dengue Virus. *Cell*, 139(7):1243–1254, 2009. ISSN 00928674. doi: 10.1016/j.cell.2009.12.017.

13. A. D. Broido and A. Clauset. Scale-free networks are rare. *Nature Communications*, 10(1):1017, 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-08746-5.

14. P. J. Bruce-Staskal, R. M. Woods, O. V. Borisov, M. J. Massare, and T. J. Hahn. Hemagglutinin from multiple divergent influenza A and B viruses bind to a distinct

branched, sialylated poly-LacNAc glycan by surface plasmon resonance. *Vaccine*, 38 (43):6757–6765, 2020. ISSN 18732518. doi: 10.1016/j.vaccine.2020.08.037.

15. W. P. Burmeister, R. W. Ruigrok, and S. Cusack. The 2.2 A resolution crystal structure of influenza B neuraminidase and its complex with sialic acid. *The EMBO Journal*, 11(1):49–56, 1992. ISSN 0261-4189. doi: 10.1002/j.1460-2075.1992.tb05026.x.

16. J. E. Carette, C. P. Guimaraes, M. Varadarajan, A. S. Park, I. Wuethrich, A. Godarova, M. Kotecki, B. H. Cochran, E. Spooner, H. L. Ploegh, and T. R. Brummelkamp. Haploid Genetic Screens in Human Cells Identify Host Factors Used by Pathogens. *Science*, 326(5957):1231–1235, 2009. ISSN 0036-8075. doi: 10.1126/science.1178955.

17. Y. F. Chen and Y. Xia. Convergent perturbation of the human domain-resolved interactome by viruses and mutations inducing similar disease phenotypes. *PLoS Computational Biology*, 15(2):e1006762, 2019. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1006762.

18. Y.-C. Chou, M. Lai, Y.-C. Wu, N.-C. Hsu, K.-S. Jeng, and W.-C. Su. Variations in genome-wide RNAi screens: lessons from influenza research. *Journal of Clinical Bioinformatics*, 5(1):2, 2015. ISSN 2043-9113. doi: 10.1186/s13336-015-0017-5.

19. K. T. Chow, M. Gale, and Y.-M. Loo. RIG-I and Other RNA Sensors in Antiviral Immunity. *Annual Review of Immunology*, 36(1):667–694, 2018. ISSN 0732-0582. doi: 10.1146/annurev-immunol-042617-053309.

20. A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-Law Distributions in Empirical Data. *SIAM Review*, 51(4):661–703, 2009. ISSN 0036-1445. doi: 10.1137/070710111.

21. D. G. Courtney, E. M. Kennedy, R. E. Dumm, H. P. Bogerd, K. Tsai, N. S. Heaton, and B. R. Cullen. Epitranscriptomic Enhancement of Influenza A Virus Gene Expression and Replication. *Cell Host & Microbe*, 22(3):377–386.e5, 2017. ISSN 1934-6069. doi: 10.1016/j.chom.2017.08.004.

22. M. A. D'Aoust, P. O. Lavoie, M. M. Couture, S. Trépanier, J. M. Guay, M. Dargis, S. Mongrand, N. Landry, B. J. Ward, and L. P. Vézina. Influenza virus-like particles produced by transient expression in Nicotiana benthamiana induce a protective immune response against a lethal viral challenge in mice. *Plant Biotechnology Journal*, 6(9):930–940, 2008. ISSN 14677652. doi: 10.1111/j.1467-7652.2008.00384.x.

23. W. L. Davies, R. R. Grunert, R. F. Haff, J. W. McGahen, E. M. Neumayer, M. Paulshock, J. C. Watts, T. R. Wood, E. C. Hermann, and C. E. Hoffmann. Antiviral Activity of 1-Adamantanamine (Amantadine). *Science*, 144(3620):862 LP – 863, 1964. doi: 10.1126/science.144.3620.862.

24. R. M. Deans, D. W. Morgens, A. Ökesli, S. Pillay, M. A. Horlbeck, M. Kampmann, L. A. Gilbert, A. Li, R. Mateo, M. Smith, J. S. Glenn, J. E. Carette, C. Khosla, and M. C. Bassik. Parallel shRNA and CRISPR-Cas9 screens enable antiviral drug target identification. *Nature Chemical Biology*, 12(5):361–366, 2016. ISSN 1552-4450. doi: 10.1038/nchembio.2050.

25. A. A. Diaz, H. Qin, M. Ramalho-Santos, and J. S. Song. HiTSelect: a comprehensive tool for high-complexity-pooled screen analysis. *Nucleic Acids Research*, 43(3):e16–e16, 2015. ISSN 0305-1048. doi: 10.1093/nar/gku1197.

26. J. G. Doench. Am I ready for CRISPR? A user's guide to genetic screens. *Nature Reviews Genetics*, 19(2):67–80, 2018. ISSN 1471-0056. doi: 10.1038/nrg.2017.97.

27. J. G. Doench, N. Fusi, M. Sullender, M. Hegde, E. W. Vaimberg, K. F. Donovan, I. Smith, Z. Tothova, C. Wilen, R. Orchard, H. W. Virgin, J. Listgarten, and D. E. Root. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nature Biotechnology*, 34(2):184–191, 2016. ISSN 15461696. doi: 10.1038/nbt.3437.

28. R. Dolin, R. C. Reichman, H. P. Madore, R. Maynard, P. N. Linton, and J. Webber-Jones. A Controlled Trial of Amantadine and Rimantadine in the Prophylaxis of Influenza a Infection. *New England Journal of Medicine*, 307(10):580–584, 1982. ISSN 0028-4793. doi: 10.1056/NEJM198209023071002.

29. G. Dong, C. Peng, J. Luo, C. Wang, L. Han, B. Wu, G. Ji, and H. He. Adamantane-Resistant Influenza A Viruses in the World (1902–2013): Frequency and Distribution of M2 Gene Mutations. *PLoS ONE*, 10(3):e0119115, 2015. doi: 10.1371/journal.pone.0119115.

30. G. R. Donowitz and G. L. Mandell. Beta-Lactam Antibiotics. *New England Journal of Medicine*, 318(7):419–426, 1988. ISSN 0028-4793. doi: 10.1056/NEJM198802183180706.

31. J. Dunning, R. S. Thwaites, and P. J. Openshaw. Seasonal and pandemic influenza: 100 years of progress, still much to learn. *Mucosal Immunology*, 13(4):566–573, 2020. ISSN 19353456. doi: 10.1038/s41385-020-0287-5.

32. A. E. Fiore, A. Fry, D. Shay, L. Gubareva, J. S. Bresee, T. M. Uyeki, and Centers for Disease Control and Prevention. Antiviral agents for the treatment and chemoprophylaxis of influenza—recommendations of the Advisory Committee on Immunization Practices (ACIP). *MMWR Recomm Rep*, 60(1):1–24, 2011.

33. E. A. Franzosa and Y. Xia. Structural principles within the human-virus protein-protein interaction network. *Proceedings of the National Academy of Sciences of the United States of America*, 108(26):10538–10543, 2011. ISSN 00278424. doi: 10.1073/pnas.1101440108.

34. S. Fukuyama and Y. Kawaoka. The pathogenesis of influenza virus infections: the contributions of virus and host factors. *Current Opinion in Immunology*, 23(4):481–486, 2011. ISSN 09527915. doi: 10.1016/j.coi.2011.07.016.

35. M. U. Gack, R. A. Albrecht, T. Urano, K.-S. Inn, I.-C. Huang, E. Carnero, M. Farzan, S. Inoue, J. U. Jung, and A. García-Sastre. Influenza A Virus NS1 Targets the Ubiquitin Ligase TRIM25 to Evade Recognition by the Host Viral RNA Sensor RIG-I. *Cell Host & Microbe*, 5(5):439–449, 2009. ISSN 1931-3128. doi: 10.1016/j.chom.2009.04.006.

36. S. Garamszegi, E. A. Franzosa, and Y. Xia. Signatures of Pleiotropy, Economy and Convergent Evolution in a Domain-Resolved Map of Human-Virus Protein-Protein

Interaction Networks. *PLoS Pathogens*, 9(12):1–9, 2013. ISSN 15537374. doi: 10.1371/journal.ppat.1003778.

37. J.-F. Gélinas, D. R. Gill, and S. C. Hyde. Multiple Inhibitory Factors Act in the Late Phase of HIV-1 Replication: a Systematic Review of the Literature. *Microbiology and Molecular Biology Reviews*, 82(1):e00051–17, 2018. ISSN 1098-5557. doi: 10.1128/MMBR.00051-17.

38. N. Goodacre, P. Devkota, E. Bae, S. Wuchty, and P. Uetz. Protein-protein interactions of human viruses. *Seminars in Cell & Developmental Biology*, 99:31–39, 2020. ISSN 10849521. doi: 10.1016/j.semcdb.2018.07.018.

39. D. E. Gordon, G. M. Jang, M. Bouhaddou, J. Xu, K. Obernier, K. M. White, M. J. O'meara, V. V. Rezelj, J. Z. Guo, D. L. Swaney, T. A. Tummino, R. Huettenhain, R. M. Kaake, A. L. Richards, B. Tutuncuoglu, H. Foussard, J. Batra, K. Haas, M. Modak, M. Kim, P. Haas, B. J. Polacco, H. Braberg, J. M. Fabius, M. Eckhardt, M. Soucheray, M. J. Bennett, M. Cakir, M. J. Mcgregor, Q. Li, B. Meyer, F. Roesch, T. Vallet, A. M. Kain, L. Miorin, E. Moreno, Z. Zar, C. Naing, Y. Zhou, S. Peng, Y. Shi, Z. Zhang, W. Shen, I. T. Kirby, J. E. Melnyk, J. S. Chorba, K. Lou, S. A. Dai, I. Barrio-Hernandez, D. Memon, C. Hernandez-Armenta, J. Lyu, C. J. P. Mathy, T. Perica, K. B. Pilla, S. J. Ganesan, D. J. Saltzberg, R. Rakesh, X. Liu, S. B. Rosenthal, L. Calviello, S. Venkataramanan, J. Liboy-Lugo, Y. Lin, X.-P. Huang, Y. Liu, S. A. Wankowicz, M. Bohn, M. Safari, F. S. Ugur, C. Koh, N. Sadat Savar, Q. D. Tran, D. Shengjuler, S. J. Fletcher, M. C. O'neal, Y. Cai, J. C. J. Chang, D. J. Broadhurst, S. Klippsten, P. P. Sharp, N. A. Wenzell, D. Kuzuoglu, H.-Y. Wang, R. Trenker, J. M. Young, D. A. Cavero, J. Hiatt, T. L. Roth, U. Rathore, A. Subramanian, J. Noack, M. Hubert, R. M. Stroud, A. D. Frankel, O. S. Rosenberg, K. A. Verba, D. A. Agard, M. Ott, M. Emerman, N. Jura, M. Von Zastrow, E. Verdin, A. Ashworth, O. Schwartz, C. D'enfert, S. Mukherjee, M. Jacobson, H. S. Malik, D. G. Fujimori, T. Ideker, C. S. Craik, S. N. Floor, J. S. Fraser, J. D. Gross, A. Sali, B. L. Roth, D. Ruggero, J. Taunton, T. Kortemme, P. Beltrao, M. Vignuzzi, A. García-Sastre, K. M. Shokat, B. K. Shoichet, and N. J. Krogan. A SARS-CoV-2

protein interaction map reveals targets for drug repurposing. *Nature*, pages 1–13, 2020. doi: 10.1038/s41586-020-2286-9.

40. T. M. E. Govaert, C. Thijs, N. Masurel, M. J. W. Sprenger, G. J. Dinant, and J. A. Knottnerus. The Efficacy of Influenza Vaccination in Elderly Individuals: A Randomized Double-blind Placebo-Controlled Trial. *JAMA*, 272(21):1661–1665, 1994. ISSN 0098-7484. doi: 10.1001/jama.1994.03520210045030.

41. E. A. Govorkova, T. Baranovich, P. Seiler, J. Armstrong, A. Burnham, Y. Guan, M. Peiris, R. J. Webby, and R. G. Webster. Antiviral resistance among highly pathogenic influenza A (H5N1) viruses isolated worldwide in 2002–2012 shows need for continued monitoring. *Antiviral Research*, 98(2):297–304, 2013. ISSN 0166-3542. doi: 10.1016/j.antiviral.2013.02.013.

42. L. V. Gubareva, L. Kaiser, and F. G. Hayden. Influenza virus neuraminidase inhibitors. *The Lancet*, 355(9206):827–835, 2000. ISSN 01406736. doi: 10.1016/S0140-6736(99)11433-8.

43. J. Han, J. T. Perez, C. Chen, Y. Li, A. Benitez, M. Kandasamy, Y. Lee, J. Andrade, B. TenOever, and B. Manicassamy. Genome-wide CRISPR/Cas9 Screen Identifies Host Factors Essential for Influenza Virus Replication. *Cell Reports*, 23(2):596–607, 2018. ISSN 2211-1247. doi: 10.1016/J.CELREP.2018.03.045.

44. R. E. Hanna and J. G. Doench. Design and analysis of CRISPR–Cas experiments. *Nature Biotechnology*, pages 1–11, 2020. ISSN 1087-0156. doi: 10.1038/s41587-020-0490-7.

45. A. J. Hay, A. J. Wolstenholme, J. J. Skehel, and M. H. Smith. The molecular basis of the specific anti-influenza action of amantadine. *The EMBO Journal*, 4(11):3021–3024, 1985. ISSN 0261-4189. doi: 10.1002/j.1460-2075.1985.tb04038.x.

46. F. G. Hayden, N. Sugaya, N. Hirotsu, N. Lee, M. D. de Jong, A. C. Hurt, T. Ishida, H. Sekino, K. Yamada, S. Portsmouth, K. Kawaguchi, T. Shishido, M. Arai, K. Tsuchiya, T. Uehara, and A. Watanabe. Baloxavir Marboxil for Uncomplicated

Influenza in Adults and Adolescents. *New England Journal of Medicine*, 379(10): 913–923, 2018. ISSN 0028-4793. doi: 10.1056/NEJMoa1716197.

47. B. E. Heaton, E. M. Kennedy, R. E. Dumm, A. T. Harding, M. T. Sacco, D. Sachs, and N. S. Heaton. A CRISPR Activation Screen Identifies a Pan-avian Influenza Virus Inhibitory Host Factor. *Cell Reports*, 20(7):1503–1512, 2017. ISSN 22111247. doi: 10.1016/j.celrep.2017.07.060.

48. B. Henrissat, A. Surolia, and P. Stanley. A Genomic View of Glycobiology. In A. Varki, R. D. Cummings, J. D. Esko, P. Stanley, G. W. Hart, M. Aebi, A. G. Darvill, T. Kinoshita, N. H. Packer, J. H. Prestegard, R. L. Schnaar, and P. H. Seeberger, editors, *Essentials of Glycobiology [Internet]*, chapter 8. Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY), 3rd edition, 2017. doi: 10.1101/GLYCOBIOLOGY.3E.008.

49. F. Hoeksema, J. Karpilow, A. Luitjens, F. Lagerwerf, M. Havenga, M. Groothuizen, G. Gillissen, A. A. Lemckert, B. Jiang, R. A. Tripp, and C. Yallop. Enhancing viral vaccine production using engineered knockout vero cell lines – A second look. *Vaccine*, 36(16):2093–2103, 2018. ISSN 18732518. doi: 10.1016/j.vaccine.2018.03.010.

50. D. W. Huang, B. T. Sherman, and R. A. Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1):44–57, 2009. ISSN 17542189. doi: 10.1038/nprot.2008.211.

51. M. M. Hughes, C. Reed, B. Flannery, S. Garg, J. A. Singleton, A. M. Fry, and M. A. Rolfes. Projected Population Benefit of Increased Effectiveness and Coverage of Influenza Vaccination on Influenza Burden in the United States. *Clinical Infectious Diseases*, 70(12):2496–2502, 2020. ISSN 1058-4838. doi: 10.1093/cid/ciz676.

52. A. C. Hurt, T. Chotpitayasunondh, N. J. Cox, R. Daniels, A. M. Fry, L. V. Gubareva, F. G. Hayden, D. S. Hui, O. Hungnes, A. Lackenby, W. Lim, A. Meijer, C. Penn, M. Tashiro, T. M. Uyeki, and M. Zambon. Antiviral resistance during the 2009 influenza A H1N1 pandemic: public health, laboratory, and clinical perspectives.

The Lancet Infectious Diseases, 12(3):240–248, 2012. ISSN 1473-3099. doi: 10.1016/S1473-3099(11)70318-8.

53. M. G. Ison, S. Portsmouth, Y. Yoshida, T. Shishido, M. Mitchener, K. Tsuchiya, T. Uehara, and F. G. Hayden. Early treatment with baloxavir marboxil in high-risk adolescent and adult outpatients with uncomplicated influenza (CAPSTONE-2): a randomised, placebo-controlled, phase 3 trial. *The Lancet Infectious Diseases*, 20 (10):1204–1214, 2020. ISSN 1473-3099. doi: 10.1016/S1473-3099(20)30004-9.

54. A. D. Iuliano, K. M. Roguski, H. H. Chang, D. J. Muscatello, R. Palekar, S. Tempia, C. Cohen, J. M. Gran, D. Schanzer, B. J. Cowling, P. Wu, J. Kyncl, L. W. Ang, M. Park, M. Redlberger-Fritz, H. Yu, L. Espenhain, A. Krishnan, G. Emukule, L. van Asten, S. Pereira da Silva, S. Aungkulanon, U. Buchholz, M. A. Widdowson, J. S. Bresee, E. Azziz-Baumgartner, P. Y. Cheng, F. Dawood, I. Foppa, S. Olsen, M. Haber, C. Jeffers, C. R. MacIntyre, A. T. Newall, J. G. Wood, M. Kundi, T. Popow-Kraupp, M. Ahmed, M. Rahman, F. Marinho, C. V. Sotomayor Proschle, N. Vergara Mallegas, F. Luzhao, L. Sa, J. Barbosa-Ramírez, D. M. Sanchez, L. A. Gomez, X. B. Vargas, a. B. Acosta Herrera, M. J. Llanés, T. K. Fischer, T. G. Krause, K. Mølbak, J. Nielsen, R. Trebbien, A. Bruno, J. Ojeda, H. Ramos, M. an der Heiden, L. del Carmen Castillo Signor, C. E. Serrano, R. Bhardwaj, M. Chadha, V. Narayan, S. Kosen, M. Bromberg, A. Glatman-Freedman, Z. Kaufman, Y. Arima, K. Oishi, S. Chaves, B. Nyawanda, R. A. Al-Jarallah, P. A. Kuri-Morales, C. R. Matus, M. E. J. Corona, A. Burmaa, O. Darmaa, M. Obtel, I. Cherkaoui, C. C. van den Wijngaard, W. van der Hoek, M. Baker, D. Bandaranayake, A. Bissielo, S. Huang, L. Lopez, C. Newbern, E. Flem, G. M. Grøneng, S. Hauge, F. G. de Cosío, Y. de Moltó, L. M. Castillo, M. A. Cabello, M. von Horoch, J. Medina Osis, A. Machado, B. Nunes, A. P. Rodrigues, E. Rodrigues, C. Calomfirescu, E. Lupulescu, R. Popescu, O. Popovici, D. Bogdanovic, M. Kostic, K. Lazarevic, Z. Milosevic, B. Tiodorovic, M. Chen, J. Cutter, V. Lee, R. Lin, S. Ma, A. L. Cohen, F. Treurnicht, W. J. Kim, C. Delgado-Sanz, S. de mateo Ontañón, A. Larrauri, I. L. León, F. Vallejo, R. Born, C. Junker, D. Koch, J. H. Chuang, W. T. Huang, H. W. Kuo, Y. C. Tsai, K. Bundhamcharoen, M. Chittaganpitch, H. K. Green, R. Pebody,

N. Goñi, H. Chiparelli, L. Brammer, and D. Mustaquim. Estimates of global seasonal influenza-associated respiratory mortality: a modelling study. *The Lancet*, 391 (10127):1285–1300, 2018. ISSN 1474547X. doi: 10.1016/S0140-6736(17)33293-2.

55. H.-H. Jeong, S. Y. Kim, M. W. C. Rousseaux, H. Y. Zoghbi, and Z. Liu. Beta-binomial modeling of CRISPR pooled screen data identifies target genes with greater sensitivity and fewer false negatives. *Genome Research*, 29(6):999–1008, 2019. ISSN 1549-5469. doi: 10.1101/gr.245571.118.

56. A. Karlas, N. Machuy, Y. Shin, K.-P. Pleissner, A. Artarini, D. Heuer, D. Becker, H. Khalil, L. A. Ogilvie, S. Hess, A. P. Mäurer, E. Müller, T. Wolff, T. Rudel, and T. F. Meyer. Genome-wide RNAi screen identifies human host factors crucial for influenza virus replication. *Nature*, 463, 2010. doi: 10.1038/nature08760.

57. J. C. Kash, T. M. Tumpey, S. C. Proll, V. Carter, O. Perwitasari, M. J. Thomas, C. F. Basler, P. Palese, J. K. Taubenberger, A. García-Sastre, D. E. Swayne, and M. G. Katze. Genomic analysis of increased host immune and cell death responses induced by 1918 influenza virus. *Nature*, 443(7111):578–581, 2006. ISSN 14764687. doi: 10.1038/nature05181.

58. C. U. Kim, W. Lew, M. A. Williams, H. Liu, L. Zhang, S. Swaminathan, N. Bischofberger, M. S. Chen, D. B. Mendel, C. Y. Tai, W. G. Laver, and R. C. Stevens. Influenza Neuraminidase Inhibitors Possessing a Novel Hydrophobic Interaction in the Enzyme Active Site: Design, Synthesis, and Structural Analysis of Carbocyclic Sialic Acid Analogues with Potent Anti-Influenza Activity. *Journal of the American Chemical Society*, 119(4):681–690, 1997. ISSN 0002-7863. doi: 10.1021/ja963036t.

59. H. Koike-Yusa, Y. Li, E.-P. Tan, M. D. C. Velasco-Herrera, and K. Yusa. Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nature Biotechnology*, 32(3):267–273, 2014. ISSN 1087-0156. doi: 10.1038/nbt.2800.

60. R. Kolde, S. Laur, P. Adler, and J. Vilo. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics*, 28(4):573–80, 2012. ISSN 1367-4811. doi: 10.1093/bioinformatics/btr709.

61. I. A. Kovács, K. Luck, K. Spirohn, Y. Wang, C. Pollis, S. Schlabach, W. Bian, D. K. Kim, N. Kishore, T. Hao, M. A. Calderwood, M. Vidal, and A. L. Barabási. Network-based prediction of protein interactions. *Nature Communications*, 10(1), 2019. ISSN 20411723. doi: 10.1038/s41467-019-09177-y.

62. M. Kuroda, P. J. Halfmann, L. Hill-Batorski, M. Ozawa, T. J. Lopes, G. Neumann, J. W. Schoggins, C. M. Rice, and Y. Kawaoka. Identification of interferon-stimulated genes that attenuate Ebola virus infection. *Nature Communications*, 11(1):1–14, 2020. ISSN 20411723. doi: 10.1038/s41467-020-16768-7.

63. A. J. Leidner, N. Murthy, H. W. Chesson, M. Biggerstaff, C. Stoecker, A. M. Harris, A. Acosta, K. Dooling, and C. B. Bridges. Cost-effectiveness of adult vaccinations: A systematic review. *Vaccine*, 37(2):226–234, 2019. ISSN 0264-410X. doi: 10.1016/j.vaccine.2018.11.056.

64. B. Li, S. M. Clohisey, B. S. Chia, B. Wang, A. Cui, T. Eisenhaure, L. D. Schweitzer, P. Hoover, N. J. Parkinson, A. Nachshon, N. Smith, T. Regan, D. Farr, M. U. Gutmann, S. I. Bukhari, A. Law, M. Sangesland, I. Gat-Viks, P. Digard, S. Vasudevan, D. Lingwood, D. H. Dockrell, J. G. Doench, J. K. Baillie, and N. Hacohen. Genome-wide CRISPR screen identifies host dependency factors for influenza A virus infection. *Nature Communications*, 11(1):164, 2020. ISSN 2041-1723. doi: 10.1038/s41467-019-13965-x.

65. W. Li, H. Xu, T. Xiao, L. Cong, M. I. Love, F. Zhang, R. A. Irizarry, J. S. Liu, M. Brown, and X. S. Liu. MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biology*, 15(12):554, 2014. ISSN 1474-760X. doi: 10.1186/s13059-014-0554-4.

66. W. Li, J. Köster, H. Xu, C.-H. Chen, T. Xiao, J. S. Liu, M. Brown, and X. S. Liu. Quality control, modeling, and visualization of CRISPR screens with MAGeCK-VISPR. *Genome Biology*, 16(1):281, 2015. ISSN 1474-760X. doi: 10.1186/s13059-015-0843-6.

67. K. Luck, D.-K. Kim, L. Lambourne, K. Spirohn, B. E. Begg, W. Bian, R. Brignall, T. Cafarelli, F. J. Campos-Laborie, B. Charloteaux, D. Choi, A. G. Coté, M. Da-

ley, S. Deimling, A. Desbuleux, A. Dricot, M. Gebbia, M. F. Hardy, N. Kishore, J. J. Knapp, I. A. Kovács, I. Lemmens, M. W. Mee, J. C. Mellor, C. Pollis, C. Pons, A. D. Richardson, S. Schlabach, B. Teeking, A. Yadav, M. Babor, D. Balcha, O. Basha, C. Bowman-Colin, S.-F. Chin, S. G. Choi, C. Colabella, G. Coppin, C. D'Amata, D. De Ridder, S. De Rouck, M. Duran-Frigola, H. Ennajdaoui, F. Goebels, L. Goehring, A. Gopal, G. Haddad, E. Hatchi, M. Helmy, Y. Jacob, Y. Kassa, S. Landini, R. Li, N. van Lieshout, A. MacWilliams, D. Markey, J. N. Paulson, S. Rangarajan, J. Rasla, A. Rayhan, T. Rolland, A. San-Miguel, Y. Shen, D. Sheykhkarimli, G. M. Sheynkman, E. Simonovsky, M. Taşan, A. Tejeda, V. Tropepe, J.-C. Twizere, Y. Wang, R. J. Weatheritt, J. Weile, Y. Xia, X. Yang, E. Yeger-Lotem, Q. Zhong, P. Aloy, G. D. Bader, J. De Las Rivas, S. Gaudet, T. Hao, J. Rak, J. Tavernier, D. E. Hill, M. Vidal, F. P. Roth, and M. A. Calderwood. A reference map of the human binary protein interactome. *Nature*, pages 1–7, 2020. ISSN 0028-0836. doi: 10.1038/s41586-020-2188-x.

68. I. Manini, C. Trombetta, G. Lazzeri, T. Pozzi, S. Rossi, and E. Montomoli. Egg-Independent Influenza Vaccines and Vaccine Candidates. *Vaccines*, 5(3):18, 2017. ISSN 2076-393X. doi: 10.3390/vaccines5030018.

69. W. M. McDougall, J. M. Perreira, E. C. Reynolds, and A. L. Brass. CRISPR genetic screens to discover host–virus interactions. *Current Opinion in Virology*, 29:87–100, 2018. ISSN 18796265. doi: 10.1016/j.coviro.2018.03.007.

70. D. M. Morens, J. K. Taubenberger, and A. S. Fauci. The Persistent Legacy of the 1918 Influenza Virus. *New England Journal of Medicine*, 361(3):225–229, 2009. ISSN 0028-4793. doi: 10.1056/nejmp0904819.

71. J. H. Morris, G. M. Knudsen, E. Verschueren, J. R. Johnson, P. Cimermancic, A. L. Greninger, and A. R. Pico. Affinity purification–mass spectrometry and network analysis to understand protein-protein interactions. *Nature Protocols*, 9(11):2539–2554, 2014. ISSN 1750-2799. doi: 10.1038/nprot.2014.164.

72. T. Nakane, A. Kotecha, A. Sente, G. McMullan, S. Masiulis, P. M. Brown, I. T. Grigoras, L. Malinauskaite, T. Malinauskas, J. Miehling, L. Yu, D. Karia, E. V.

Pechnikova, E. de Jong, J. Keizer, M. Bischoff, J. McCormack, P. Tiemeijer, S. W. Hardwick, D. Y. Chirgadze, G. Murshudov, A. R. Aricescu, and S. H. Scheres. Single-particle cryo-EM at atomic resolution. *bioRxiv*, page 2020.05.22.110189, 2020. doi: 10.1101/2020.05.22.110189.

73. L. P. Newman, N. Bhat, J. A. Fleming, and K. M. Neuzil. Global influenza seasonality to inform country-level vaccine programs: An analysis of WHO FluNet influenza surveillance data between 2011 and 2016. *PLoS ONE*, 13(2):e0193263, 2018. ISSN 1932-6203. doi: 10.1371/journal.pone.0193263.

74. K. L. Nichol, A. Lind, K. L. Margolis, M. Murdoch, R. McFadden, M. Hauge, S. Magnan, and M. Drake. The Effectiveness of Vaccination against Influenza in Healthy, Working Adults. *New England Journal of Medicine*, 333(14):889–893, 1995. ISSN 0028-4793. doi: 10.1056/NEJM199510053331401.

75. E. Nourani, F. Khunjush, and S. Durmuş. Computational approaches for prediction of pathogen-host protein-protein interactions. *Frontiers in Microbiology*, 6:94, 2015. doi: 10.3389/fmicb.2015.00094.

76. T. Obayashi and K. Kinoshita. Rank of Correlation Coefficient as a Comparable Measure for Biological Significance of Gene Coexpression. *DNA Research*, 16(5): 249–260, 2009. ISSN 1340-2838. doi: 10.1093/dnares/dsp016.

77. T. Obayashi, Y. Kagaya, Y. Aoki, S. Tadaka, and K. Kinoshita. COXPRESdb v7: a gene coexpression database for 11 animal species supported by 23 coexpression platforms for technical evaluation and evolutionary inference. *Nucleic Acids Research*, 47(D1):D55–D62, 2019. ISSN 0305-1048. doi: 10.1093/nar/gky1155.

78. S. Omoto, V. Speranzini, T. Hashimoto, T. Noshi, H. Yamaguchi, M. Kawai, K. Kawaguchi, T. Uehara, T. Shishido, A. Naito, and S. Cusack. Characterization of influenza virus variants induced by treatment with the endonuclease inhibitor baloxavir marboxil. *Scientific Reports*, 8(1):9633, 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-27890-4.

79. R. C. Orchard, C. B. Wilen, J. G. Doench, M. T. Baldridge, B. T. McCune, Y. C. J.

Lee, S. Lee, S. M. Pruett-Miller, C. A. Nelson, D. H. Fremont, and H. W. Virgin. Discovery of a proteinaceous cellular receptor for a norovirus. *Science*, 353(6302): 933–936, 2016. ISSN 10959203. doi: 10.1126/science.aaf1220.

80. A. Pandey, N. Singh, S. Sambhara, and S. K. Mittal. Egg-independent vaccine strategies for highly pathogenic H5N1 influenza viruses. *Human Vaccines*, 6(2):178–188, 2010. ISSN 15548600. doi: 10.4161/hv.6.2.9899.

81. R. J. Park, T. Wang, D. Koundakjian, J. F. Hultquist, P. Lamothe-Molina, B. Monel, K. Schumann, H. Yu, K. M. Krupzcak, W. Garcia-Beltran, A. Piechocka-Trocha, N. J. Krogan, A. Marson, D. M. Sabatini, E. S. Lander, N. Hacohen, and B. D. Walker. A genome-wide CRISPR screen identifies a restricted set of HIV host dependency factors. *Nature Genetics*, 49(2):193–203, 2017. ISSN 15461718. doi: 10.1038/ng.3741.

82. S. D. Pirooz, S. He, T. Zhang, X. Zhang, Z. Zhao, S. Oh, D. O'Connell, P. Khalilzadeh, S. Amini-Bavil-Olyaee, M. Farzan, and C. Liang. UVRAG is required for virus entry through combinatorial interaction with the class C-Vps complex and SNAREs. *Proceedings of the National Academy of Sciences of the United States of America*, 111(7):2716–2721, 2014. ISSN 10916490. doi: 10.1073/pnas.1320629111.

83. A. S. Puschnik, K. Majzoub, Y. S. Ooi, and J. E. Carette. A CRISPR toolbox to study virus-host interactions. *Nature Reviews Microbiology*, 15(6):351–364, 2017. ISSN 17401534. doi: 10.1038/nrmicro.2017.29.

84. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL https://www.r-project.org.

85. G. Rodrigo, J. A. Daròs, and S. F. Elena. Virus-host interactome: putting the accent on how it changes. *Journal of Proteomics*, 156:1–4, 2017. ISSN 18767737. doi: 10.1016/j.jprot.2016.12.007.

86. R. Roosenhoff, V. Reed, A. Kenwright, M. Schutten, C. A. Boucher, A. Monto, B. Clinch, D. Kumar, R. Whitley, J. S. Nguyen-Van-Tam, A. D. M. E. Osterhaus,

R. A. M. Fouchier, and P. L. A. Fraaij. Viral Kinetics and Resistance Development in Children Treated with Neuraminidase Inhibitors: The Influenza Resistance Information Study (IRIS). *Clinical Infectious Diseases*, 71(5):1186–1194, 2020. ISSN 1058-4838. doi: 10.1093/cid/ciz939.

87. C. M. Saad-Roy, C. E. Wagner, R. E. Baker, S. E. Morris, J. Farrar, A. L. Graham, S. A. Levin, M. J. Mina, C. J. E. Metcalf, and B. T. Grenfell. Immune life history, vaccination, and the dynamics of SARS-CoV-2 over the next 5 years. *Science*, 2020. doi: 10.1126/science.abd7343.

88. S. D. Shapira, I. Gat-Viks, B. O. Shum, A. Dricot, M. M. de Grace, L. Wu, P. B. Gupta, T. Hao, S. J. Silver, D. E. Root, D. E. Hill, A. Regev, and N. Hacohen. A Physical and Regulatory Map of Host-Influenza Interactions Reveals Pathways in H1N1 Infection. *Cell*, 139(7):1255–1267, 2009. ISSN 00928674. doi: 10.1016/j.cell.2009.12.018.

89. D. M. Sharon, S. Nesdoly, H. J. Yang, J. F. Gélinas, Y. Xia, S. Ansorge, and A. A. Kamen. A pooled genome-wide screening strategy to identify and rank influenza host restriction factors in cell-based vaccine production platforms. *Scientific Reports*, 10 (1):12166, 2020. ISSN 20452322. doi: 10.1038/s41598-020-68934-y.

90. G. J. Smith, J. Bahl, D. Vijaykrishna, J. Zhang, L. L. Poon, H. Chen, R. G. Webster, J. S. Peiris, and Y. Guan. Dating the emergence of pandemic influenza viruses. *Proceedings of the National Academy of Sciences of the United States of America*, 106(28):11709–11712, 2009. ISSN 00278424. doi: 10.1073/pnas.0904991106.

91. M. P. Somes, R. M. Turner, L. J. Dwyer, and A. T. Newall. Estimating the annual attack rate of seasonal influenza among unvaccinated individuals: A systematic review and meta-analysis. *Vaccine*, 36(23):3199–3207, 2018. ISSN 0264-410X. doi: https://doi.org/10.1016/j.vaccine.2018.04.063.

92. P. N. Spahn, T. Bath, R. J. Weiss, J. Kim, J. D. Esko, N. E. Lewis, and O. Harismendy. PinAPL-Py: A comprehensive web-application for the analysis of CRISPR/Cas9 screens. *Scientific Reports*, 7(1):15854, 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-16193-9.

93. D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, N. T. Doncheva, J. H. Morris, P. Bork, L. J. Jensen, and C. Mering. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, 47(D1):D607–D613, 2019. ISSN 0305-1048. doi: 10.1093/nar/gky1131.

94. S.-L. Tan, G. Ganji, B. Paeper, S. Proll, and M. G. Katze. Systems biology and the host response to viral infection. *Nature Biotechnology*, 25(12):1383–1389, 2007. ISSN 1546-1696. doi: 10.1038/nbt1207-1383.

95. Y. Tang, G. Zhong, L. Zhu, X. Liu, Y. Shan, H. Feng, Z. Bu, H. Chen, and C. Wang. Herc5 attenuates influenza A virus by catalyzing ISGylation of viral NS1 protein. *The Journal of Immunology*, 184(10):5777–90, 2010. ISSN 1550-6606. doi: 10.4049/jimmunol.0903588.

96. J. K. Taubenberger and J. C. Kash. Influenza Virus Evolution, Host Adaptation, and Pandemic Formation. *Cell Host & Microbe*, 7(6):440–451, 2010. ISSN 19313128. doi: 10.1016/j.chom.2010.05.009.

97. S. N. Thulasi Raman and Y. Zhou. Networks of Host Factors that Interact with NS1 Protein of Influenza A Virus. *Frontiers in Microbiology*, 7:654, 2016. ISSN 1664-302X. doi: 10.3389/fmicb.2016.00654.

98. J. J. Treanor, H. E. Sahly, J. King, I. Graham, R. Izikson, R. Kohberger, P. Patriarca, and M. Cox. Protective efficacy of a trivalent recombinant hemagglutinin protein vaccine (FluBlok®) against influenza in healthy adults: A randomized, placebo-controlled trial. *Vaccine*, 29(44):7733–7739, 2011. ISSN 0264-410X. doi: 10.1016/j.vaccine.2011.07.128.

99. S. Tripathi, M. O. Pohl, Y. Zhou, A. Rodriguez-Frandsen, G. Wang, D. A. Stein, H. M. Moulton, P. DeJesus, J. Che, L. C. Mulder, E. Yángüez, D. Andenmatten, L. Pache, B. Manicassamy, R. A. Albrecht, M. G. Gonzalez, Q. Nguyen, A. Brass, S. Elledge, M. White, S. Shapira, N. Hacohen, A. Karlas, T. F. Meyer, M. Shales, A. Gatorano, J. R. Johnson, G. Jang, T. Johnson, E. Verschueren, D. Sanders, N. Krogan, M. Shaw, R. König, S. Stertz, A. García-Sastre, and S. K. Chanda.

Meta- and Orthogonal Integration of Influenza "OMICs" Data Defines a Role for UBR4 in Virus Budding. *Cell Host & Microbe*, 18(6):723–735, 2015. ISSN 19313128. doi: 10.1016/j.chom.2015.11.002.

100. S. Tweedie, B. Braschi, K. Gray, T. E. M. Jones, R. Seal, B. Yates, and E. A. Bruford. Genenames.org: the HGNC and VGNC resources in 2021. *Nucleic Acids Research*, 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa980.

101. S. M. G. van der Sanden, W. Wu, N. Dybdahl-Sissoko, W. C. Weldon, P. Brooks, J. O'Donnell, L. P. Jones, C. Brown, S. M. Tompkins, M. S. Oberste, J. Karpilow, and R. A. Tripp. Engineering Enhanced Vaccine Cell Lines To Eradicate Vaccine-Preventable Diseases: the Polio End Game. *Journal of Virology*, 90(4):1694–1704, 2016. ISSN 0022-538X. doi: 10.1128/jvi.01464-15.

102. S. Venkatesan, P. R. Myles, K. J. Bolton, S. G. Muthuri, T. Al Khuwaitir, A. P. Anovadiya, E. Azziz-Baumgartner, T. Bajjou, M. Bassetti, B. Beovic, B. Bertisch, I. Bonmarin, R. Booy, V. H. Borja-Aburto, H. Burgmann, B. Cao, J. Carratala, T. Chinbayar, C. Cilloniz, J. T. Denholm, S. R. Dominguez, P. A. D. Duarte, G. Dubnov-Raz, S. Fanella, Z. Gao, P. Gérardin, M. Giannella, S. Gubbels, J. Herberg, A. L. Higuera Iglesias, P. H. Hoeger, X. Y. Hu, Q. T. Islam, M. F. Jiménez, G. Keijzers, H. Khalili, G. Kusznierz, I. Kuzman, E. Langenegger, K. B. Lankarani, Y.-S. Leo, R. P. Libster, R. Linko, F. Madanat, E. Maltezos, A. Mamun, T. Manabe, G. Metan, A. Mickiene, D. Mikić, K. G. I. Mohn, M. E. Oliva, M. Ozkan, D. Parekh, M. Paul, B. A. Rath, S. Refaey, A. H. Rodríguez, B. Sertogullarindan, J. Skręt-Magierło, A. Somer, E. Talarek, J. W. Tang, K. To, D. Tran, T. M. Uyeki, W. Vaudry, T. Vidmar, P. Zarogoulidis, and J. S. Nguyen-Van-Tam. Neuraminidase Inhibitors and Hospital Length of Stay: A Meta-analysis of Individual Participant Data to Determine Treatment Effectiveness Among Patients Hospitalized With Non-fatal 2009 Pandemic Influenza A(H1N1) Virus Infection. *The Journal of Infectious Diseases*, 221(3):356–366, 2020. ISSN 0022-1899. doi: 10.1093/infdis/jiz152.

103. F. Villalón-Letelier, A. Brooks, P. Saunders, S. Londrigan, and P. Reading. Host Cell Restriction Factors that Limit Influenza A Infection. *Viruses*, 9(12):376, 2017.

ISSN 1999-4915. doi: 10.3390/v9120376.

104. M. von Itzstein, W.-Y. Wu, G. B. Kok, M. S. Pegg, J. C. Dyason, B. Jin, T. Van Phan, M. L. Smythe, H. F. White, S. W. Oliver, P. M. Colman, J. N. Varghese, D. M. Ryan, J. M. Woods, R. C. Bethell, V. J. Hotham, J. M. Cameron, and C. R. Penn. Rational design of potent sialidase-based inhibitors of influenza virus replication. *Nature*, 363(6428):418–423, 1993. ISSN 1476-4687. doi: 10.1038/363418a0.

105. B. Wang, M. Wang, W. Zhang, T. Xiao, C.-H. Chen, A. Wu, F. Wu, N. Traugh, X. Wang, Z. Li, S. Mei, Y. Cui, S. Shi, J. J. Lipp, M. Hinterndorfer, J. Zuber, M. Brown, W. Li, and X. S. Liu. Integrative analysis of pooled CRISPR genetic screens using MAGeCKFlute. *Nature Protocols*, 14(3):756–780, 2019. ISSN 1754-2189. doi: 10.1038/s41596-018-0113-7.

106. L. Wang, B. Fu, W. Li, G. Patil, L. Liu, M. E. Dorf, and S. Li. Comparative influenza protein interactomes identify the role of plakophilin 2 in virus restriction. *Nature Communications*, 8(1):13876, 2017. ISSN 2041-1723. doi: 10.1038/ncomms13876.

107. T. Wang, J. J. Wei, D. M. Sabatini, and E. S. Lander. Genetic screens in human cells using the CRISPR-Cas9 system. *Science*, 343(6166):80–4, 2014. ISSN 1095-9203. doi: 10.1126/science.1246981.

108. B. J. Ward, A. Makarkov, A. Séguin, S. Pillet, S. Trépanier, J. Dhaliwall, M. D. Libman, T. Vesikari, and N. Landry. Efficacy, immunogenicity, and safety of a plant-derived, quadrivalent, virus-like particle influenza vaccine in adults (18–64 years) and older adults (65 years): two multicentre, randomised phase 3 trials. *The Lancet*, 2020. ISSN 01406736. doi: 10.1016/s0140-6736(20)32014-6.

109. K. L. Warfield, K. R. Schaaf, L. E. DeWald, K. B. Spurgers, W. Wang, E. Stavale, M. Mendenhall, M. H. Shilts, T. B. Stockwell, D. L. Barnard, U. Ramstedt, and S. R. Das. Lack of selective resistance of influenza A virus in presence of host-targeted antiviral, UV-4B. *Scientific Reports*, 9(1):7484, 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-43030-y.

110. T. Watanabe, S. Watanabe, and Y. Kawaoka. Cellular networks involved in the

influenza virus life cycle. *Cell Host & Microbe*, 7(6):427–439, 2010. ISSN 19313128. doi: 10.1016/j.chom.2010.05.008.

111. T. Watanabe, E. Kawakami, J. E. Shoemaker, T. J. Lopes, Y. Matsuoka, Y. Tomita, H. Kozuka-Hata, T. Gorai, T. Kuwahara, E. Takeda, A. Nagata, R. Takano, M. Kiso, M. Yamashita, Y. Sakai-Tagawa, H. Katsura, N. Nonaka, H. Fujii, K. Fujii, Y. Sugita, T. Noda, H. Goto, S. Fukuyama, S. Watanabe, G. Neumann, M. Oyama, H. Kitano, and Y. Kawaoka. Influenza virus-host interactome screen as a platform for antiviral drug development. *Cell Host & Microbe*, 16(6):795–805, 2014. ISSN 19346069. doi: 10.1016/j.chom.2014.11.002.

112. H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. doi: 10.1007/978-0-387-98141-3. URL https://ggplot2.tidyverse.org.

113. R. Winkler, E. Gillis, L. Lasman, M. Safra, S. Geula, C. Soyris, A. Nachshon, J. Tai-Schmiedel, N. Friedman, V. T. K. Le-Trilling, M. Trilling, M. Mandelboim, J. H. Hanna, S. Schwartz, and N. Stern-Ginossar. m6A modification controls the innate immune response to infection by targeting type I interferons. *Nature Immunology*, 20(2):173–182, 2019. ISSN 1529-2916. doi: 10.1038/s41590-018-0275-z.

114. W. Wu, N. Orr-Burks, J. Karpilow, and R. A. Tripp. Development of improved vaccine cell lines against rotavirus. *Scientific Data*, 4(1):1–12, 2017. ISSN 20524463. doi: 10.1038/sdata.2017.21.

115. K. S. Xue, L. H. Moncla, T. Bedford, and J. D. Bloom. Within-Host Evolution of Human Influenza Virus. *Trends in Microbiology*, 26(9):781–793, 2018. ISSN 1878-4380. doi: 10.1016/j.tim.2018.02.007.

116. A. D. Yates, P. Achuthan, W. Akanni, J. Allen, J. Allen, J. Alvarez-Jarreta, M. R. Amode, I. M. Armean, A. G. Azov, R. Bennett, J. Bhai, K. Billis, S. Boddu, J. C. Marugán, C. Cummins, C. Davidson, K. Dodiya, R. Fatima, A. Gall, C. G. Giron, L. Gil, T. Grego, L. Haggerty, E. Haskell, T. Hourlier, O. G. Izuogu, S. H. Janacek, T. Juettemann, M. Kay, I. Lavidas, T. Le, D. Lemos, J. G. Martinez, T. Maurel, M. McDowall, A. McMahon, S. Mohanan, B. Moore, M. Nuhn, D. N. Oheh,

A. Parker, A. Parton, M. Patricio, M. P. Sakthivel, A. I. Abdul Salam, B. M. Schmitt, H. Schuilenburg, D. Sheppard, M. Sycheva, M. Szuba, K. Taylor, A. Thormann, G. Threadgold, A. Vullo, B. Walts, A. Winterbottom, A. Zadissa, M. Chakiachvili, B. Flint, A. Frankish, S. E. Hunt, G. IIsley, M. Kostadima, N. Langridge, J. E. Loveland, F. J. Martin, J. Morales, J. M. Mudge, M. Muffato, E. Perry, M. Ruffier, S. J. Trevanion, F. Cunningham, K. L. Howe, D. R. Zerbino, and P. Flicek. Ensembl 2020. *Nucleic Acids Research*, 48(D1):D682–D688, 2020. ISSN 0305-1048. doi: 10.1093/nar/gkz966.

117. K. M. Yip, N. Fischer, E. Paknia, A. Chari, and H. Stark. Atomic-resolution protein structure determination by cryo-EM. *Nature*, pages 1–5, 2020. ISSN 0028-0836. doi: 10.1038/s41586-020-2833-4.

118. Y. Zhang, B. D. Aevermann, T. K. Anderson, D. F. Burke, G. Dauphin, Z. Gu, S. He, S. Kumar, C. N. Larsen, A. J. Lee, X. Li, C. Macken, C. Mahaffey, B. E. Pickett, B. Reardon, T. Smith, L. Stewart, C. Suloway, G. Sun, L. Tong, A. L. Vincent, B. Walters, S. Zaremba, H. Zhao, L. Zhou, C. Zmasek, E. B. Klem, and R. H. Scheuermann. Influenza Research Database: An integrated bioinformatics resource for influenza virus research. *Nucleic Acids Research*, 45(D1):D466–D474, 2017. ISSN 0305-1048. doi: 10.1093/nar/gkw857.

119. Y. Zhou, Y. Hou, J. Shen, Y. Huang, W. Martin, and F. Cheng. Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. *Cell Discovery*, 6 (1):1–18, 2020. ISSN 20565968. doi: 10.1038/s41421-020-0153-3.

120. T. Ziegler, A. Mamahit, and N. J. Cox. 65 years of influenza surveillance by a World Health Organization-coordinated global network. *Influenza and Other Respiratory Viruses*, 12(5):558–565, 2018. ISSN 1750-2659. doi: 10.1111/irv.12570.

Appendix

A. Degree Distribution of the 'union of Human Interactomes' (HI-union)

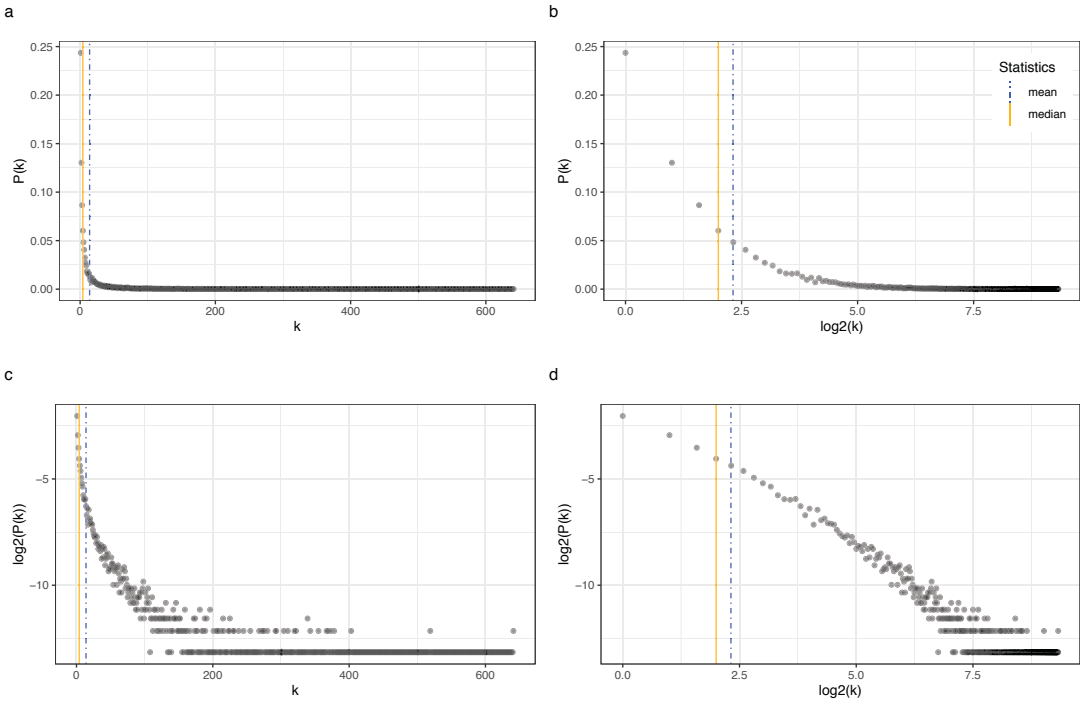

Figure 15: Exploration of the degree distribution of the union of Human Interactomes (`HI-union`). Each vertex within the graph, or interactome, represents a human protein-coding gene. The degree of a vertex, denoted as k, is the number of edges, or interactions, that it has with other vertices within the interactome. Each subplot depicts the relationship between $P(k) = \frac{n_k}{|V|}$, the number of vertices with degree k (n_k) divided by the total number of vertices in the graph ($|V|$), and k, the vertex degree value. To ensure that log_2 transformations may be applied to $P(k)$, we make $P(k) > 0, \forall k$ by adding a pseudocount of 1 to the numerator, giving $P(k) = \frac{n_k+1}{|V|}$. To better show the relationship between $P(k)$ and k, scatter plots (**a**, **b**, **c**, and **d**), each with 641 data points based on $|V| = 9094$ degree values, were created with log_2 –transformations applied successively to each axis in turn. The log_2 transformation effectively spreads out the distribution of values along a given axis, making it easier to visualize. Subplot (**d**) has the log_2 transform applied to both axes; here, the depiction of a linear correlation between $log_2(P(k))$ and $log_2(k)$ suggests that `HI-union`, at its current stage of completeness, approximates a power-law relationship[5]. That is, many genes have few interactions, and few genes have many. The interactome dataset `HI-union` was obtained from Luck *et al.* (2020)[67]. Data wrangling and analysis was performed in `Julia`[9]; plots were produced using `ggplot2`[112] in `R`[84]. These plots are the same as those in figure 8, except with the target interactome being `HI-union` instead of `HuRI`.

B. Literature-curated Host Factors in Gene Coexpression Space

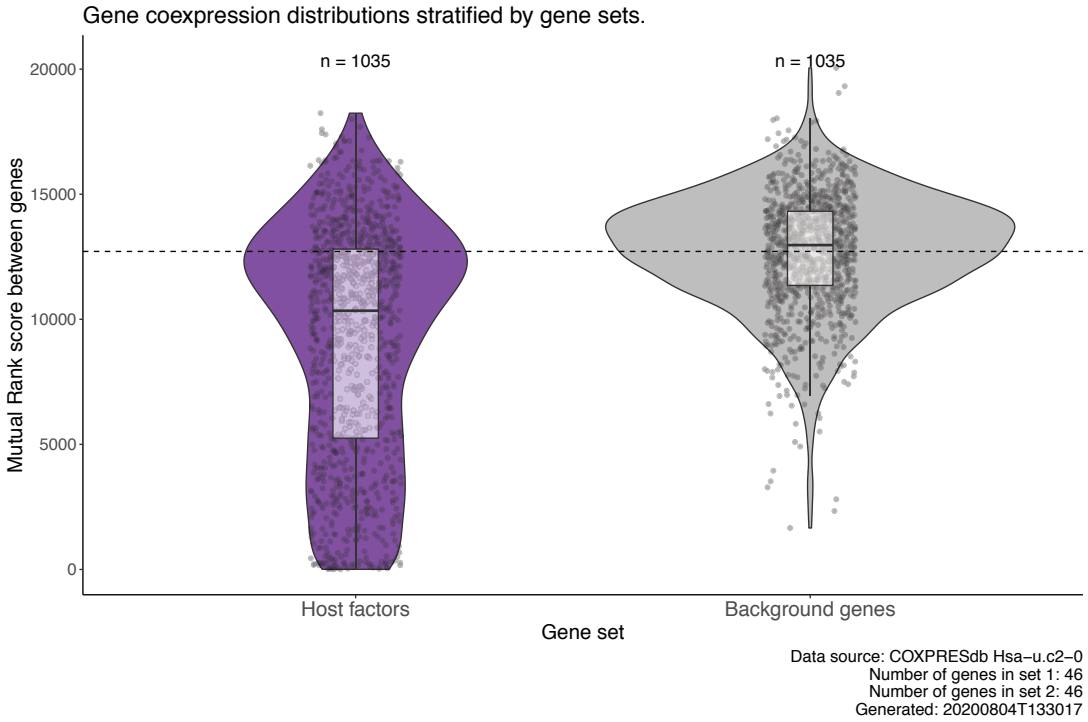

Figure 16: Distribution of `COXPRESdb` 'Mutual Rank' scores corresponding to all pairs of genes ($n = 1035$) from the complete set of literature-curated host factors ($N = 46$), as contained in table 1, is plotted alongside a randomly sampled (without replacement) set of genes ($N = 46$, with $n = 1035$ gene pairs). Data wrangling and analysis was performed in `Julia`[9] and plots were produced using `ggplot2`[112] in `R`[84]. The black dotted line indicates the mean 'Mutual Rank' score for all values in `COXPRESdb`.

C. Curated Sets of Influenza Host Factors Identified in Genome-wide Perturbation-based Screening Experiments

The putative influenza host factor sets are organized based on the study that they were identified in and their classification as either 'antiviral' or 'proviral'; refer to table 3 for a summary of the screening studies. Importantly, each screening study has variations in their experimental setup, and thus care must be taken when comparing between them at the level of the gene. Genes are given as NCBI gene identifiers for ease of comparison and the guarantee of unique values. This information is included in the Sharon *et al.* (2020) study[89] as 'Supplemental S1' and is thus under a Creative Commons license.

Putative Antiviral Gene Hits by Paper

- Brass *et al.* (2009)[12] ($n = 4$)
  - 10410, 126789, 8460, 55339
- Shapira *et al.* (2009)[88] ($n = 176$)
  - 154, 60489, 9048, 57130, 567, 637, 29760, 657, 659, 694, 54930, 56913, 23523, 8913, 786, 59284, 23705, 57658, 810, 29775, 836, 10344, 1019, 55602, 1031, 1050, 1052, 79643, 1153, 23529, 57396, 122011, 1487, 10217, 1491, 56259, 2919, 91966, 54205, 23500, 1616, 64421, 1677, 1718, 23312, 4189, 1871, 55840, 1960, 5610, 2048, 2113, 2131, 2149, 81558, 22868, 26234, 55030, 23768, 8061, 2299, 8456, 7855, 7107, 94239, 3052, 8334, 3099, 3106, 3159, 3228, 79803, 26353, 10581, 10410, 51447, 3551, 9641, 8517, 22806, 3556, 246778, 80895, 3620, 80789, 3661, 3705, 3709, 3714, 23030, 9665, 57148, 22920, 10365, 8609, 9903, 55958, 3976, 3988, 9516, 114569, 5609, 10746, 4293, 11184, 5598, 145282, 8569, 2872, 10884, 4627, 9612, 4814, 28511, 338321, 64127, 7025, 8829, 5029, 11235, 5195, 5298, 5277, 5287, 30849, 8399, 5322, 5333, 5362, 23654, 5371, 56342, 80148, 166336, 5566, 8575, 3275, 5699, 5997, 22838, 6102, 6195, 6197, 23076, 23168, 23429, 56681, 27111, 5272, 5054, 6432, 4091, 81848, 6772, 6788, 23075, 11346, 6867, 6934, 23424, 7029, 375346, 55281, 8718, 8600, 10766, 7157, 84676, 54970, 56995, 7342, 79465, 7390, 83878, 7484, 9589
- Watanabe *et al.* (2014)[111] ($n = 34$)
  - 71, 87, 128272, 908, 64708, 8727, 7818, 8663, 1973, 1975, 11160, 2664, 2771, 9328, 8570, 3875, 10226, 23787, 4673, 5315, 5518, 5631, 5644, 22931, 6223, 6201, 6281, 6432, 154091, 7922, 6636, 10972, 10959, 10382
- Tripathi *et al.* (2015)[99] ($n = 485$)
  - 148, 242, 725, 758, 1559, 2259, 3182, 3622, 3660, 4025, 5082, 5149, 5738, 6041, 6249, 6275, 6650, 6904, 7148, 8406, 8876, 9567, 9997, 10036, 10406, 22980, 26093, 26278, 27285, 29128, 51388, 51517, 54707, 56673, 56850, 57165, 57480, 58476, 79717, 79868, 80790, 81629, 83446, 83937, 84261, 85462, 90416, 92565, 93010, 117196, 121274, 136371, 148808, 197196, 285368, 400935, 6150, 6155, 6183, 6187, 6900, 6314, 8295, 9913, 10629, 1, 9, 20, 29, 87, 140, 176, 309, 384, 567, 611, 694, 786, 810, 908, 1031, 1050, 1052, 1102, 1130, 1153, 1305, 1404, 1420, 1491, 1667, 1668, 1947, 1960, 1982, 2041, 2048, 2299, 2328, 2533, 2664, 2872, 2880, 2919, 3052, 3099, 3106, 3159, 3228, 3275, 3502, 3661, 3705, 3875, 4293, 4617, 4627, 4814, 4951, 5029, 5099, 5195, 5272, 5277, 5298, 5303, 5362, 5426, 5598, 5699, 5731, 5780, 5997, 6102, 6195, 6251, 6281, 6303, 6449, 6451, 6523, 6607, 6730, 6788, 7025, 7054, 7107, 7328, 7342, 7347, 7371, 7484, 7772, 7855, 8243, 8260, 8334, 8427, 8456, 8460, 8539, 8569, 8570, 8575, 8600, 8609, 8674, 8682, 8718, 8727, 8879, 8880, 8904, 9048, 9125, 9205, 9211, 9217, 9276, 9328, 9516, 9612, 9641, 9665, 9739, 9829, 9903, 9929, 9953, 10217, 10344, 10365, 10425, 10581, 10746, 10766, 10873, 10959, 11160, 11184, 11235, 11274, 11346, 22838, 22856, 22868, 22920, 22931, 23075, 23076, 23312, 23357, 23410, 23424, 23429, 23500, 23523, 23527, 23641, 23654, 23768, 25875, 26234, 26353, 27111, 27173, 28511, 29079, 29775, 30811, 30820, 30849, 50804, 51032, 51057, 51275, 51480, 53342, 53944, 54345, 54407, 54441, 54467, 54555, 54577, 54930, 54970, 54981, 55030, 55164, 55281, 55339, 55602, 55894, 55958, 56259, 56342, 56547, 56913, 56916, 56924, 56979, 57127, 57130, 57148, 57182, 57188, 57396, 57512, 57531, 57541, 57544, 57578, 57580, 57658, 58489, 59284, 60489, 60558, 64063, 64130, 64421, 64708, 64711, 64776, 65264, 66000, 79025, 79074, 79096, 79176, 79465, 79643, 79682, 79741, 79799, 79803, 79876, 79902, 79922, 79933, 80148, 80174, 80198, 80339, 80789, 81848, 83551, 83638, 83878, 83935, 84056, 84191, 84553, 84676, 84978, 85235, 85301, 85508, 90233, 91181, 91966, 92579, 93650, 94239, 96626, 114928, 114984, 116138, 117156, 117195, 122011, 124411, 125206, 126248, 126789, 140597, 140686, 144383, 145282, 146850, 150275, 152206, 154091, 158800, 163071, 164832, 166336, 168448, 201305, 203111, 203611, 220906, 246778, 252839, 254048, 283554, 283742, 284001, 319101, 338321, 338799, 339977, 340596, 344807, 345757, 353345, 374819, 375346, 387590, 387680, 388815, 389217, 389630, 389692, 391533, 391560, 392255, 393078, 403274, 404552, 414062, 440097, 494115, 497048, 729264, 6432, 10410, 60, 2280, 3586, 23729, 55824, 83860, 125, 238, 335, 762, 1371, 2147, 2712, 2782, 2785, 3239, 3263, 3726, 3763, 3785, 4040, 5127, 5128, 5617, 5796, 6261, 6657, 6934, 7088, 7094, 8471, 8833, 9942, 10045, 10577, 10606, 23542, 51129, 51270, 51720, 54764, 84941, 117245, 170575, 1504, 3752, 7157, 8794, 23168, 51447, 57082, 93973, 3551, 4673, 6147, 6201, 6223, 7818, 8517, 10884,

71, 154, 301, 637, 657, 836, 983, 1019, 1677, 1718, 1871, 2113, 2149, 2771, 3620, 3976, 3988, 4091, 5054, 5322, 5333, 5371, 5518, 5609, 5631, 6772, 7029, 7390, 8399, 8829, 10226, 23529, 23787, 29760, 54205, 56681, 64127

- Heaton *et al.* (2017)[47] ($n = 1190$)
  - 124872, 200150, 5884, 101059938, 5521, 92104, 63970, 54453, 9843, 252995, 83882, 10553, 400823, 283403, 5005, 79097, 100287171, 9375, 122664, 653720, 346517, 158326, 645425, 55638, 4000, 10979, 11067, 55731, 4717, 5510, 1031, 3006, 253635, 8091, 23384, 401124, 6046, 84541, 23025, 55975, 57761, 2543, 323, 79888, 3167, 7752, 10980, 57523, 90233, 374768, 3554, 161436, 9203, 60561, 5157, 94030, 11184, 91074, 56204, 10775, 6236, 6480, 100507055, 351, 9628, 9908, 7021, 9722, 10795, 727837, 333926, 10170, 283635, 5239, 56916, 9725, 253012, 5425, 6857, 4589, 59067, 1129, 51333, 441308, 5146, 100505841, 10189, 9229, 94, 5553, 64598, 6228, 140893, 2651, 64078, 57094, 202243, 162968, 5781, 10870, 337959, 2217, 100131094, 60680, 84335, 64320, 11190, 5641, 10044, 144347, 353143, 22871, 56896, 51326, 57096, 7084, 7306, 54900, 53829, 168391, 7067, 6492, 144125, 64772, 1017, 9537, 285242, 339559, 311, 4345, 3418, 8390, 80099, 9382, 81341, 55810, 27111, 9069, 143503, 23729, 9625, 132660, 124783, 5454, 55885, 1675, 10396, 130340, 112885, 6543, 2597, 9454, 970, 27244, 57447, 4281, 10044, 54970, 2258, 27297, 23526, 2348, 1846, 10621, 8715, 64062, 58538, 6907, 9185, 9201, 4543, 4654, 10076, 26100, 5284, 4995, 23779, 116238, 56899, 27018, 8874, 271, 149840, 7707, 4940, 57209, 2074, 23600, 25902, 339366, 1395, 93474, 940, 58499, 64506, 1829, 1734, 80221, 4062, 126695, 203447, 115572, 80823, 147929, 9788, 26223, 163479, 51473, 255743, 344658, 143678, 128674, 5787, 80700, 51399, 79939, 11036, 1496, 51002, 9813, 53826, 51300, 7763, 389792, 55139, 644150, 1435, 5144, 390916, 10465, 140578, 142891, 388341, 254428, 23671, 2306, 10612, 8218, 255252, 64840, 7374, 463, 5339, 10114, 138199, 3029, 163589, 9337, 84826, 6541, 2495, 9703, 23063, 712, 25911, 1278, 79676, 3656, 352999, 653125, 51441, 250, 51204, 10720, 627, 9610, 392862, 13, 112939, 56751, 54823, 57705, 51542, 57150, 130367, 5409, 51555, 55735, 84219, 2932, 317719, 2180, 145748, 3755, 144124, 11200, 84223, 401427, 23366, 55755, 6871, 7164, 8881, 157848, 9315, 521, 23399, 475, 1368, 823, 11170, 57633, 11252, 51134, 4481, 85416, 10749, 631, 55101, 285550, 3842, 90141, 285180, 90557, 129401, 55585, 191, 166979, 64168, 51729, 4831, 84707, 23360, 60468, 80254, 4151, 27010, 8189, 341947, 23142, 80255, 3707, 26095, 7453, 23648, 6651, 10232, 440093, 10399, 25925, 11174, 85465, 10935, 4838, 28999, 220323, 145258, 5723, 10678, 23366, 54940, 10390, 144715, 23185, 7140, 7884, 9955, 22882, 1307, 2832, 4239, 55240, 9633, 11259, 65250, 3422, 137682, 10585, 728279, 4255, 221302, 100288801, 9663, 221322, 6206, 5916, 8398, 6560, 124222, 164395, 7579, 10788, 494118, 92521, 5295, 9750, 29780, 2537, 283899, 50650, 23369, 124925, 283710, 55230, 3776, 5340, 4288, 9746, 120406, 55854, 80168, 353135, 85015, 64837, 23126, 2903, 29090, 1108, 7494, 169270, 91942, 94032, 4308, 168417, 257236, 390261, 60437, 80059, 5325, 6559, 56975, 388561, 5339, 55898, 8540, 3949, 9130, 25843, 200844, 80312, 89953, 3158, 2837, 2559, 51705, 10103, 388963, 27128, 9140, 79567, 5696, 100130988, 3028, 8505, 6156, 4594, 51035, 23710, 165100, 10409, 7419, 2132, 3670, 23081, 11186, 1836, 493860, 25970, 56243, 1102, 201625, 200185, 4665, 121130, 51751, 9470, 1731, 6520, 22894, 283870, 3416, 5568, 92196, 10468, 57621, 222642, 1048, 26999, 1769, 10089, 3185, 10138, 119391, 257364, 10371, 55088, 93408, 10590, 55565, 134265, 1456, 8139, 167838, 245928, 51435, 55651, 51088, 79858, 51206, 7490, 163126, 6522, 80739, 388523, 55341, 221477, 83636, 10717, 285601, 9881, 6582, 7365, 7700, 9473, 64756, 54466, 10333, 57717, 87178, 5775, 55862, 9644, 5704, 26270, 81892, 948, 81624, 10093, 26164, 58531, 55041, 1201, 54780, 23197, 7712, 9075, 9611, 56142, 79918, 11277, 146664, 2122, 10534, 80142, 7111, 266747, 6397, 5982, 254187, 10342, 140873, 23012, 10114, 389874, 196415, 23383, 51715, 49, 497190, 8975, 140890, 862, 4282, 5308, 55973, 9891, 9217, 643418, 51208, 5218, 57582, 51554, 6141, 145501, 5167, 58511, 25915, 5007, 10020, 100505591, 79712, 90060, 126402, 4284, 1981, 169270, 6455, 57658, 64754, 2099, 9472, 55778, 26166, 285242, 29128, 286514, 4070, 3615, 54494, 8631, 404281, 51118, 150082, 11126, 150221, 100038246, 3915, 256356, 79949, 11186, 4791, 6738, 50636, 51380, 341019, 5446, 1029, 27304, 10288, 9046, 51649, 23125, 84311, 255626, 56964, 30000, 201973, 10087, 79755, 58487, 5021, 8554, 57054, 10763, 54360, 57508, 9643, 80324, 64121, 70, 282770, 218, 199, 5464, 5509, 142685, 255738, 64926, 81539, 147700, 56834, 339287, 221786, 168451, 145482, 58485, 388931, 245911, 57111, 56915, 4975, 9931, 54520, 10249, 56341, 4277, 64761, 50832, 342898, 90627, 9133, 9517, 9718, 9087, 22848, 10848, 85415, 90427, 158431, 1837, 348262, 55635, 84264, 131920, 51363, 140690, 85417, 54462, 2954, 4174, 1434, 820, 3155, 256643, 678, 56243, 5291, 54989, 157869, 4208, 6138, 27095, 9672, 30014, 4148, 64419, 1959, 83853, 135886, 8819, 401541, 83872, 353134, 79973, 26074, 65018, 29071, 91584, 54906, 9607, 2057, 5236, 117157, 10533, 149465, 10509, 2982, 84292, 199, 84696, 79838, 181, 757, 84226, 3326, 84146, 137392, 80157, 91768, 151393, 90293, 399939, 390037, 10880, 64344, 10276, 7323, 51246, 387640, 27091, 8858, 55596, 83548, 7341, 51082, 90273, 51512, 2554, 83259, 2203, 2768, 338657, 30815, 64400, 6160, 984, 80209, 219557, 56907, 79772, 78986, 28988, 57097, 2982, 2286, 416, 653567, 10855, 1605, 26333, 55254, 6943, 27284, 30820, 55229, 3120, 8895, 391356, 25788, 8482, 25865, 4210, 283897, 643853, 100170229, 5190, 8386, 285311, 56133, 147183, 7138, 222584, 1780, 919, 10084, 9289, 1582, 5288, 6358, 27201, 83759, 390648, 286256, 219293, 26168, 9938, 7881, 8514, 390445, 4329, 7369, 353513, 84418, 22800, 55793, 64326, 26995, 25980, 7179, 142680, 55344, 55657, 92610, 4332, 8453, 26996, 58485, 9122, 100130274, 79661, 124790, 10991, 154075, 9892, 2208, 5635, 546, 54496, 283652, 51114, 79770, 9201, 2998, 9528, 90843, 81671, 10747, 3091, 120400, 3170, 5899, 7652, 27033, 120, 269, 126129, 160492, 56673, 4157, 80323, 100463482, 84305, 128338, 5745, 7401, 337967, 114134, 345079, 11154, 558, 100129924, 81627, 9320, 399687, 18, 140690, 100526739, 29946, 5803, 2115, 128153, 5136, 5158, 58493, 51506, 91227, 84933, 9860, 339965, 1996, 115677, 51530, 2208, 353333, 163688, 8706, 7634, 55532, 4009, 26150, 10229, 627, 26205, 838, 147841, 5744, 54914, 6829, 23327, 130026, 50856, 2852, 22994, 143425, 390152, 6002, 6504, 126205, 442361, 51324, 902, 54878, 5049, 93436, 148867, 121129, 55802, 64648, 5866, 80183, 51365, 4800, 221443, 55143, 5431, 23533, 283171, 100131980, 51021, 54587, 8796, 80110, 100131378, 219432, 388199, 79839, 8406, 283209, 23609, 6101, 8436, 83988, 7474, 7421, 9330, 84220, 399, 54785, 5524, 153657, 55322, 441150, 55130, 4108, 55997, 22999, 1436, 219479, 2052, 27085, 54212, 84539, 80346, 3007, 81788, 9815, 26137, 2180, 31, 10201, 350383, 6939, 64782, 7769, 27067, 219927, 245932, 9191, 4659, 54873, 23209, 154865, 83900, 93589, 346653, 55135, 55061, 84154, 653145, 2813, 57003, 221491, 50833, 59338, 5408, 337879, 377630, 3557, 10144, 1016, 83986, 57187, 7089, 316, 118812, 7767, 116113, 54107, 3663, 445, 55607, 10299, 1460, 5067, 3823, 6103, 392376, 10908, 3698, 79414, 29093, 100132476, 283777, 1359, 79981, 25833, 51117, 93978, 116441, 23746, 643, 84318, 54108, 55353, 2773, 149371, 10699, 90381, 9568, 4956, 27430, 51661, 27044, 6331, 221092, 1015, 4640, 29887, 90423, 9805, 81470, 54112, 11073, 2160, 2358, 79831, 10349, 653550, 27094, 8600, 55174, 5999, 6130, 93134, 4898, 137886, 1390, 84532, 10785, 57158, 23395, 55216, 55213, 10152, 29100, 139135, 11026, 79174, 55636, 338699, 1062, 10169, 10916, 8292, 4681, 400120, 10875, 64852, 83449, 390439, 152503, 254783, 135932, 57555, 9802, 3357, 9447, 126133, 341799, 79893, 22902, 282616, 84276, 51142, 5580, 11160, 80352, 55293, 79696, 2796, 3685, 56113, 2908, 272, 4681, 9028, 10859, 9666, 64981, 282991, 2263, 56606, 55277, 94234, 114926, 26231, 219348, 375190, 7352, 84520, 23644, 64859, 148545, 4898, 4795, 284618, 51147, 10514, 25961, 80036, 5494, 90342, 9913, 222235, 7760, 57711, 7139, 403273, 149233, 81602, 203547, 57569, 388394, 6902, 7167, 7091, 1807, 55246, 79728, 29999, 203100, 57708, 85452, 1292, 3694, 151516, 79713, 57786, 8618, 427
- Sharon *et al.* (2020)[89] ($n = 89$ at significance threshold $\alpha = 0.05$)
  - 9921, 23644, 51663, 9410, 3340, 5915, 8570, 23193, 51530, 10762, 9739, 51386, 9567, 8664, 79882, 8480, 1457, 22919, 58490, 116092, 8131, 7249, 11140, 56006, 4899, 1656, 4928, 80018, 57472, 10641, 9887, 9295, 6794, 2733, 79886, 56943, 3181, 79002, 262, 27304, 57688, 57343, 5901, 29107, 56923, 84872, 11137, 11097, 79798, 84916, 10498, 4651, 51256, 84928, 8667, 5531, 26019, 347734, 29101, 10482, 7248, 10501, 8772, 140834, 54851, 7690, 167227, 51126, 348180, 84823, 27335, 64864, 8241, 126382, 256714, 4733, 54584, 219436, 129080, 286380, 5094, 51441, 112970, 390144, 51451, 22828, 51750, 56111, 100132911

Putative Proviral Gene Hits by Paper

- Brass *et al.* (2009)[12] ($n = 129$)
  - 215, 191, 154810, 27329, 9716, 372, 515, 537, 533, 155066, 9550, 8704, 23621, 331, 65990, 58509, 253559, 10241, 90557, 91409, 25978, 1198, 1203, 53942, 1314, 1315, 9276, 22820, 10980, 22818, 10815, 8738, 51340, 286464, 51167, 64858, 1659, 84222, 1756, 1825, 1780, 9343, 1965, 2058, 90952, 132884, 9780, 2212, 220042, 2521, 81544, 51659, 2700, 2931, 3192, 8739, 11185, 9922, 57703, 163233, 51149, 53353, 80740, 4236, 84730, 255758, 197259, 4350, 9019, 9650, 93649, 4714, 4809, 4846, 115677, 57122, 4928, 9818, 10204, 10482, 56114, 84108, 27295, 5226, 5304, 9651, 56342, 26121, 10594, 80863, 5652, 5868, 8934, 10432, 25813, 9092, 57147, 85465, 23451, 10992, 23450, 118980, 6507, 6514, 10559, 54946, 10569, 6628, 6633, 6634, 6729, 246329, 80765, 27233, 96764, 7069, 284355, 10155, 7319, 59286, 57418, 91746, 7691, 7710, 7564, 7767, 9668, 25888, 79818
- Carette *et al.* (2009)[16] ($n = 2$)
  - 55907, 7355
- Shapira *et al.* (2009)[88] ($n = 221$)
  - 53, 2182, 88, 101, 9510, 113, 113146, 29929, 269, 116519, 64860, 1386, 468, 8553, 9184, 26175, 8209, 773, 10241, 818, 6348, 57018, 25819, 8837, 79094, 1154, 9976, 25932, 1191, 1277, 1314, 51226, 10898, 1437, 1452, 51076, 4283, 1540, 1573, 23002, 8527, 1746, 1847, 9451, 8894, 2002, 2033, 2043, 9415, 54537, 2194, 26190, 2263, 2289, 10691, 2802, 2820, 9289, 2885, 55127, 79654, 3105, 3135, 8091, 3383, 3433, 24138, 3460, 3489, 3601, 84639, 11009, 133396, 90865, 388324, 3621, 54556, 3654, 3659, 3660, 3664, 3394, 3707, 10625, 3726, 55709, 9711, 9675, 23379, 57707, 11004, 1316, 3845, 3911, 3953, 3959, 55957, 4000, 23175, 51599, 84445, 4110, 5606, 10454, 5600, 5603, 4236, 9645, 10797, 9788, 51594, 4683, 4758, 4763, 4779, 4783, 4790, 4792, 4794, 338322, 338323, 3164, 10482, 4938, 8638, 9180, 51585, 5106, 51449, 5154, 57162, 5226, 51230, 23469, 5286, 5293, 5296, 8503, 64600, 113026, 10714, 55703, 11230, 5578, 5616, 9266, 5698, 5781, 10076, 10966, 9693, 10235, 5930, 10616, 23180, 8780, 6038, 6041, 7844, 285830, 23521, 6160, 8986, 9252, 65117, 64108, 6258, 60485, 59342, 6385, 6464, 51763, 6507, 4092, 10073, 6653, 10252, 80765, 11329, 6830, 6840, 6850, 8867, 51347, 26000, 6904, 54103, 9338, 6942, 7020, 7022, 7046, 9874, 81793, 25880, 25816, 8793, 8771, 79931, 10346, 7706, 7726, 5987, 7316, 80329, 7372, 64854, 8936, 10810, 23001, 22884, 7474, 7494, 55596, 23503, 7545, 7586, 79698, 57178, 7764, 57473
- Karlas *et al.* (2010)[56] ($n = 168$)
  - 10768, 64400, 1173, 54518, 8260, 9048, 85300, 477, 27032, 537, 10159, 527, 9114, 523, 526, 567, 8938, 343472, 9256, 115708, 56911, 132001, 85417, 962, 965, 975, 1027, 1056, 23563, 25932, 1195, 9746, 26507, 1314, 1315, 9276, 22820, 57585, 1409, 10663, 1586, 113612, 1629, 166614, 10202, 1740, 55929, 196403, 92235, 1915, 8661, 8663, 8666, 9775, 64772, 2051, 2197, 10517, 115548, 83706, 54508, 2729, 2797, 115330, 2905, 9776, 8341, 3248, 3299, 3329, 23765, 3552, 3660, 9636, 55600, 3725, 10300, 3768, 23277, 284058, 3832, 3837, 23367, 8022, 84894, 9361, 9890, 4125, 4148, 10001, 84292, 4609, 4654, 284086, 91754, 4913, 23165, 4928, 10482, 611, 54510, 5253, 23533, 5300, 5328, 5338, 1263, 5437, 5441, 54866, 10594, 5631, 83886, 55851, 5682, 10213, 5708, 5798, 51560, 29127, 79171, 56729, 117584, 57484, 11224, 6204, 6208, 6217, 6233, 6193, 861, 6294, 6404, 10291, 23451, 51639, 6439, 10280, 9356, 55234, 6625, 6636, 22938, 58533, 51429, 6651, 23524, 23166, 55959, 6830, 10607, 6929, 7030, 8784, 10188, 55809, 9830, 6737, 166655, 83983, 10907, 337867, 8875, 7483, 56949, 7511, 7514, 57621
- Watanabe *et al.* (2014)[111] ($n = 358$)
  - 58, 158, 55750, 79026, 226, 10541, 161, 1175, 372, 10564, 55160, 128272, 93436, 51593, 23020, 55210, 10632, 539, 8455, 79888, 9531, 10409, 7919, 57448, 10902, 9184, 705, 64423, 29105, 55421, 374864, 79002, 92747, 790, 790, 55832, 826, 4076, 829, 832, 22794, 11335, 84229, 146849, 28958, 440193, 10694, 8099, 10523, 1153, 10970, 1207, 1209, 1213, 29097, 23019, 25920, 118881, 1314, 1315, 1315, 50813, 1465, 1487, 1499, 1509, 1537, 1537, 3491, 1615, 9188, 1654, 9775, 57696, 1665, 1665, 22907, 1659, 84061, 51726, 55735, 23234, 10059, 1786, 54344, 144455, 144455, 51366, 1917, 1917, 1937, 1938, 9343, 8661, 8668, 8668, 8664, 8662, 1975, 8672, 84895, 10160, 54888, 2521, 9513, 8729, 10985, 2665, 50628, 10399, 114928, 2873, 2923, 64151, 10614, 8339, 3106, 3151, 3190, 3181, 4670, 11319, 11319, 3326, 3304, 3306, 3309, 3313, 3315, 3326, 3329, 3376, 23463, 10644, 3609, 3609, 10989, 3613, 57508, 3675, 3675, 3692, 9445, 3716, 7965, 3735, 79734, 54442, 9675, 23277, 3861, 3852, 3854, 319101, 114294, 3927, 3958, 3959, 10128, 55379, 55692, 4085, 4131, 81631, 5606, 4141, 4174, 10627, 6182, 740, 28973, 51649, 23107, 28957, 10240, 50804, 4628, 4637, 4666, 23310, 3071, 56926, 57727, 4705, 4702, 4728, 4774, 83696, 4831, 79050, 4841, 23279, 23165, 23165, 9688, 9688, 9180, 56124, 5824, 5245, 5245, 11331, 84844, 23646, 8985, 56655, 5464, 5479, 5495, 5496, 5537, 9055, 10935, 5566, 5573, 55660, 10594, 5635, 167681, 5682, 5685, 5685, 5686, 5686, 5690, 5690, 5694, 5700, 5702, 5704, 5706, 5706, 5707, 5717, 5717, 5718, 5718, 5719, 10213, 5708, 5709, 5709, 5710, 9861, 9861, 5713, 5714, 5714, 5720, 5721, 51495, 284119, 54517, 5817, 5901, 5901, 3843, 8241, 5954, 64283, 6135, 6138, 6147, 6154, 6157, 6122, 6161, 6167, 6181, 6184, 6204, 6206, 6208, 6209, 6218, 6218, 6271, 6275, 9667, 25939, 9147, 23753, 51150, 51150, 871, 7536, 8175, 23451, 10992, 23450, 23450, 10262, 10262, 83443, 6434, 10929, 8683, 6449, 57619, 22827, 80143, 22938, 51763, 6566, 292, 6513, 27044, 6625, 6626, 6627, 6627, 6628, 6631, 6631, 6634, 6634, 6637, 8878, 8878, 6733, 6749, 11329, 10492, 8148, 6894, 64786, 54997, 57187, 3842, 10452, 10452, 7167, 7171, 10155, 7248, 7273, 7277, 79861, 203068, 7280, 81567, 11338, 6675, 29796, 7374, 9958, 8239, 7415, 7425, 7431, 79084, 7514, 11260, 11260, 22803, 7531, 7534, 55854
- Tripathi *et al.* (2015)[99] ($n = 1445$)
  - 390, 553, 1145, 1339, 1912, 2827, 4919, 8859, 9086, 9409, 23098, 23474, 25758, 54913, 55200, 55627, 57093, 57492, 79864, 84950, 89958, 91978, 92312, 114036, 144132, 154075, 219793, 284131, 338611, 389332, 648000, 100271374, 3550, 4832, 5935, 6406, 8899, 23148, 125476, 701, 5430, 6132, 6156, 6164, 6229, 6234, 6235, 7311, 7465, 8665, 25873, 983, 2889, 6134, 6142, 6165, 6188, 6203, 6891, 9088, 9611, 10569, 25942, 88, 101, 140, 158, 177, 215, 361, 558, 685, 705, 722, 826, 827, 829, 832, 871, 962, 965, 1056, 1130, 1191, 1198, 1209, 1305, 1316, 1329, 1348, 1394, 1465, 1511, 1521, 1537, 1540, 1573, 1586, 1613, 1717, 1737, 1740, 1746, 1756, 1780, 1787, 1810, 1825, 1832, 1847, 2012, 2072, 2108, 2162, 2182, 2212, 2322, 2327, 2328, 2334, 2346, 2356, 2357, 2488, 2550, 2580, 2620, 2638, 2662, 2665, 2700, 2703, 2820, 2829, 2880, 2905, 3029, 3071, 3151, 3164, 3248, 3299, 3363, 3460, 3489, 3543, 3547, 3601, 3613, 3621, 3631, 3652, 3707, 3752, 3778, 3854, 3911, 3927, 3958, 4025, 4046, 4058, 4076, 4110, 4125, 4148, 4174, 4185, 4283, 4582, 4637, 4643, 4666, 4752, 4774, 4783, 4829, 4849, 4886, 4920, 4923, 4938, 5063, 5067, 5069, 5096, 5106, 5165, 5245, 5265, 5304, 5310, 5338, 5495, 5537, 5616, 5619, 5635, 5697, 5724, 5732, 5783, 5795, 5805, 5824, 5954, 5961, 6013, 6038, 6182, 6258, 6271, 6275, 6328, 6334, 6357, 6385, 6439, 6442, 6449, 6478, 6489, 6494, 6514, 6566, 6574, 6604, 6605, 6624, 6645, 6653, 6675, 6715, 6729, 6809, 6875, 6894, 6900, 6929, 6932, 6942, 7005, 7022, 7030, 7084, 7163, 7167, 7171, 7178, 7277, 7280, 7366, 7374, 7431, 7474, 7483, 7490, 7511, 7545, 7564, 7586, 7691, 7706, 7710, 7726, 7764, 7767, 7844, 8022, 8091, 8099, 8209, 8216, 8320, 8323, 8335, 8339, 8341, 8455, 8519, 8527, 8539, 8553, 8558, 8672, 8677, 8726, 8739, 8784, 8793, 8831, 8834, 8875, 8934, 8936, 8938, 8985, 9019, 9048, 9055, 9101, 9201, 9217, 9230, 9231, 9289, 9338, 9356, 9415, 9445, 9464, 9497, 9509, 9510, 9531, 9625, 9645, 9650, 9651, 9667, 9672, 9683, 9685, 9693, 9711, 9736, 9746, 9752, 9753, 9776, 9780, 9788, 9791, 9806, 9830, 9874, 9890, 9943, 9958, 9976, 10036, 10052, 10076, 10105, 10152, 10160, 10202, 10206, 10240, 10252, 10296, 10300, 10346, 10409, 10410, 10432, 10452, 10461, 10462, 10517, 10541, 10559, 10595, 10600, 10607, 10627, 10663, 10691, 10725, 10783, 10786, 10797, 10810, 10815, 10898, 10900, 10902, 10929, 10935, 10962, 10965, 10966, 10970, 11004, 11102, 11185, 11214, 11230,

11237, 11319, 11322, 11331, 22794, 22803, 22838, 22856, 22884, 22907, 23001, 23002, 23076, 23102, 23166, 23173, 23174, 23175, 23180, 23234, 23242, 23324, 23326, 23379, 23386, 23410, 23456, 23464, 23469, 23503, 23531, 23538, 23552, 23563, 23620, 23637, 23646, 23753, 23770, 24138, 25813, 25816, 25819, 25831, 25880, 25888, 25897, 25939, 25978, 26000, 26043, 26058, 26175, 26190, 26507, 26692, 27044, 27087, 27233, 27243, 27295, 27347, 28957, 28958, 28978, 29035, 29063, 29097, 29115, 29929, 29959, 29999, 30815, 49860, 50488, 50804, 50833, 50839, 51046, 51061, 51076, 51150, 51167, 51172, 51257, 51275, 51294, 51307, 51365, 51390, 51393, 51429, 51449, 51454, 51495, 51526, 51534, 51594, 51599, 51637, 51649, 51726, 53342, 53826, 53942, 54103, 54344, 54442, 54494, 54507, 54508, 54510, 54517, 54518, 54537, 54556, 54776, 54888, 54946, 54964, 54980, 54991, 54997, 55031, 55160, 55229, 55323, 55511, 55534, 55577, 55588, 55600, 55652, 55692, 55703, 55709, 55735, 55793, 55809, 55850, 55854, 55872, 55957, 55959, 56114, 56124, 56164, 56300, 56311, 56342, 56660, 56729, 56911, 56924, 56926, 56997, 57018, 57038, 57085, 57143, 57150, 57448, 57473, 57475, 57484, 57502, 57508, 57531, 57534, 57544, 57551, 57561, 57579, 57585, 57619, 57621, 57695, 57696, 57707, 57709, 57727, 58491, 58526, 58528, 58533, 59342, 60370, 60437, 60485, 60489, 60526, 64093, 64108, 64151, 64283, 64284, 64400, 64421, 64423, 64600, 64601, 64772, 64854, 64858, 64860, 65117, 65220, 65268, 65990, 79026, 79050, 79094, 79171, 79363, 79574, 79616, 79641, 79654, 79698, 79705, 79734, 79818, 79872, 79888, 79931, 80005, 80055, 80108, 80143, 80148, 80231, 80345, 80818, 80863, 81567, 83550, 83638, 83696, 83706, 83886, 83903, 83983, 84061, 84108, 84171, 84187, 84197, 84229, 84307, 84330, 84445, 84458, 84639, 84894, 84895, 84899, 85300, 85438, 89781, 89891, 90060, 90557, 90736, 90865, 90952, 90987, 90990, 91409, 91754, 92292, 92610, 92747, 93436, 93611, 93953, 93973, 94234, 96626, 113026, 113146, 113540, 113612, 113878, 114299, 114788, 114880, 114928, 114971, 115330, 115548, 115677, 115701, 115708, 116519, 116535, 117584, 118442, 118980, 119772, 122258, 122525, 124583, 124602, 126541, 127495, 127733, 132001, 132884, 133396, 134391, 135154, 140701, 140886, 144455, 145788, 146849, 147166, 147409, 149420, 150248, 152992, 153571, 154810, 155006, 155435, 157983, 160492, 161823, 163882, 166655, 169792, 170712, 196403, 197259, 200185, 200424, 201305, 201456, 203611, 205327, 220906, 221545, 246329, 246777, 253143, 254065, 255239, 256051, 256126, 256892, 259286, 282969, 283377, 283455, 284029, 284040, 284086, 284230, 284269, 284355, 284366, 284393, 284573, 285440, 285830, 286464, 286827, 337867, 338322, 338323, 339512, 339766, 340260, 343472, 344807, 347240, 348327, 353376, 374864, 377841, 387590, 387867, 387911, 388324, 388403, 388428, 388795, 389941, 390538, 391194, 399706, 400658, 401007, 401565, 401665, 404281, 407977, 431707, 439931, 440396, 440400, 440738, 441239, 441670, 643358, 643641, 653712, 729324, 730974, 100129028, 100129482, 100129842, 100287898, 100505621, 102724699, 372, 523, 526, 527, 533, 537, 790, 975, 1195, 1263, 1314, 1315, 1659, 1956, 2186, 2263, 2521, 3265, 3329, 3433, 3660, 3675, 3725, 3837, 3959, 4193, 4236, 4350, 4809, 4928, 5226, 5253, 5566, 5603, 5606, 5682, 5707, 5708, 5798, 5868, 6204, 6208, 6217, 6404, 6507, 6625, 6627, 6628, 6634, 6737, 6830, 7273, 7514, 8021, 8661, 8780, 8837, 9114, 9180, 9184, 9256, 9276, 9343, 9575, 9675, 9775, 9818, 10001, 10155, 10159, 10188, 10213, 10241, 10280, 10291, 10482, 10594, 10616, 10625, 10714, 10733, 10992, 11329, 22820, 22938, 23165, 23216, 23277, 23450, 23451, 23604, 23765, 25932, 29127, 54866, 55421, 55851, 57418, 80765, 166614, 203068, 3764, 5045, 5347, 5599, 6240, 10606, 27302, 1069, 1128, 2515, 3263, 3673, 4093, 4139, 4509, 4881, 5465, 5579, 5588, 5617, 6512, 6790, 6910, 7328, 8626, 8833, 54820, 57082, 94235, 292, 335, 515, 611, 773, 1452, 2197, 2797, 3832, 5328, 5652, 6041, 6122, 6135, 6230, 6233, 6421, 6651, 6792, 6850, 8260, 8438, 8664, 8668, 8896, 9092, 9159, 9550, 10381, 10768, 11224, 22818, 30811, 30849, 51639, 51659, 53353, 55660, 55929, 59286, 64805, 80740, 81544, 85452, 92579, 96764, 153201, 167681, 284119, 340024, 207, 466, 539, 567, 658, 672, 699, 1019, 1026, 1213, 1385, 1386, 1437, 1499, 1615, 1654, 1665, 1915, 1937, 1938, 1965, 1974, 1975, 2011, 2033, 2058, 2885, 2932, 3105, 3106, 3135, 3181, 3190, 3304, 3320, 3326, 3376, 3394, 3516, 3609, 3659, 3716, 3717, 3735, 3842, 3843, 4085, 4141, 4609, 4670, 4790, 4792, 4794, 4914, 5156, 5296, 5437, 5441, 5573, 5594, 5601, 5607, 5685, 5686, 5690, 5694, 5698, 5700, 5702, 5704, 5706, 5709, 5710, 5713, 5714, 5717, 5718, 5719, 5720, 5721, 5781, 6015, 6138, 6139, 6154, 6157, 6160, 6181, 6187, 6191, 6193, 6206, 6209, 6218, 6224, 6427, 6434, 6464, 6626, 6631, 6633, 6636, 6637, 6749, 7046, 7248, 7531, 7533, 7534, 7536, 7919, 7965, 8175, 8662, 8663, 8666, 8683, 8878, 8894, 9135, 9521, 9688, 9861, 10181, 10262, 10454, 10492, 10523, 10642, 10907, 11335, 11338, 22827, 23020, 23279, 23521, 23524, 26121, 51226, 51340, 51366, 51585, 51593, 51763, 55596, 57122, 57187, 64105, 64946, 79228, 79902, 83443, 84292, 84844, 348995, 740, 1153, 1173, 1859, 2802, 2869, 2870, 2873, 2923, 4702, 4705, 4709, 4714, 4728, 4831, 5300, 5422, 5427, 5496, 5757, 5797, 5930, 5957, 5987, 6294, 6733, 7415, 8239, 8570, 8653, 9188, 9448, 9513, 9716, 10128, 10399, 10626, 10849, 10980, 10989, 23019, 23352, 23367, 23387, 25920, 28973, 28996, 29105, 29110, 50813, 51230, 51497, 55234, 56655, 56949, 57120, 57703, 58509, 79002, 83737, 91746, 151871, 284058, 53, 70, 91, 92, 113, 147, 157, 161, 191, 226, 269, 290, 331, 351, 369, 468, 477, 535, 572, 602, 816, 818, 861, 1027, 1054, 1072, 1111, 1154, 1175, 1203, 1207, 1277, 1280, 1409, 1434, 1455, 1487, 1509, 1629, 1733, 1786, 1845, 1917, 1967, 2002, 2022, 2043, 2045, 2048, 2050, 2051, 2194, 2260, 2264, 2289, 2324, 2342, 2444, 2475, 2539, 2597, 2729, 2931, 2936, 2956, 3009, 3084, 3306, 3312, 3313, 3315, 3356, 3383, 3491, 3552, 3568, 3581, 3586, 3605, 3654, 3664, 3674, 3684, 3687, 3692, 3710, 3726, 3760, 3767, 3768, 3773, 3845, 3861, 3953, 3984, 4000, 4092, 4131, 4221, 4296, 4599, 4600, 4628, 4654, 4683, 4758, 4763, 4779, 4841, 4846, 4913, 4915, 5062, 5154, 5286, 5289, 5293, 5464, 5479, 5578, 5580, 5584, 5600, 5602, 5605, 5610, 5631, 5886, 5901, 6093, 6184, 6196, 6340, 6348, 6446, 6545, 6610, 6613, 6672, 6811, 6840, 6904, 6934, 7020, 7069, 7294, 7316, 7319, 7341, 7372, 7423, 7494, 7786, 7879, 7980, 8148, 8241, 8290, 8476, 8503, 8536, 8638, 8704, 8729, 8738, 8741, 8744, 8771, 8805, 8867, 8986, 9149, 9252, 9266, 9361, 9451, 9474, 9578, 9636, 9641, 9922, 9972, 10055, 10059, 10073, 10114, 10204, 10235, 10297, 10564, 10614, 10644, 10694, 10892, 10985, 11009, 11035, 11113, 11213, 11260, 23049, 23309, 23310, 23396, 23463, 23533, 23534, 23621, 23677, 26269, 27032, 27092, 27327, 27329, 29843, 29882, 50616, 50628, 51347, 51422, 51560, 54856, 55127, 55379, 55750, 55832, 56000, 56893, 57147, 57162, 57178, 80329, 81793, 85417, 92235, 93649, 116447, 118881, 155066, 204851, 246721, 338599, 387082, 440193

- Sharon *et al.* (2020)[89] ($n = 60$ at significance threshold $\alpha = 0.05$)
  - 1595, 4843, 3843, 4904, 80347, 23198, 23047, 23415, 54439, 1725, 7528, 8106, 9985, 79868, 7355, 2821, 7458, 3882, 79132, 1801, 28960, 55907, 200576, 9181, 5511, 7389, 2538, 1802, 196441, 89978, 23070, 85465, 64062, 207, 10055, 51070, 58497, 9125, 51631, 10682, 10020, 124801, 7086, 10559, 11168, 171546, 8539, 11273, 79089, 120526, 3838, 55253, 84324, 728489, 79897, 81488, 246176, 6928, 388381, 79893