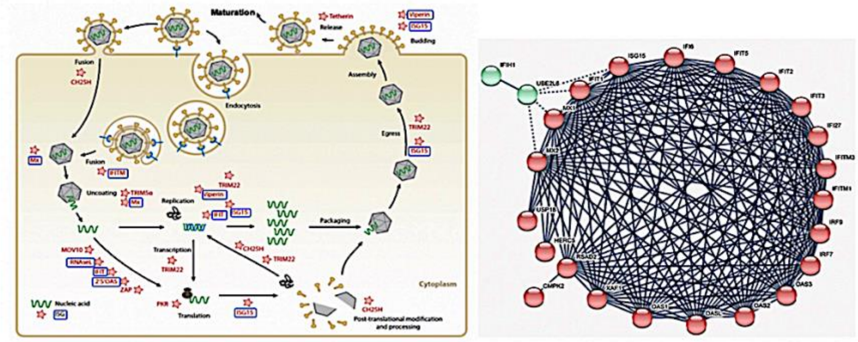
UNIVERSITY OF CRETE
SCHOOL OF MEDICINE

FORTH
FOUNDATION FOR RESEARCH AND TECHNOLOGY - HELLAS

Postgraduate MSc program in Bioinformatics



Exploring the molecular landscape of SARS-CoV-2 infection: Differential Expression, Pathway Analysis and Prognostic Modeling

POLYMNIA GKOUBLIA

MSc Thesis

Supervisor
Dr. George Potamias

Heraklion, Crete
December 2021

Abstract

Historical note. At the close of 2019, WHO China Country Office was informed of a pneumonia of unknown cause, detected in the city of Wuhan in Hubei province, China. On 10 January 2020 WHO declared the outbreak of 2019-nCoV (the first name assigned to the new disease), and on February 11 ,2020, it named the disease as COVID-19 (Coronavirus Disease 2019). WHO declared the outbreak a *Public Health Emergency of International Concern* on 30 January 2020, and a pandemic on 11 March 2020. Latest data from 'Our World in Data' report about 238 million SARS-CoV-2 confirmed infections and about 4.3 million deaths attributed to SARS-CoV-2 [access to 'Our World in Data COVID-19 data explorer' on 27 December 2021].

Background. This postgraduate thesis is concentrated on SARS-CoV-2 infection. In particular, it focuses on exploring and investigating the molecular landscape that characterizes and governs the genes/proteins interactions and biological the processes taking place during the different infection stages. One basic quest of the thesis concerns the **molecular patterns that differentiate SARS-CoV-2 infection stages**, and which may guide to the so-called 'cytokine storm' that underlie the infection's acute respiratory syndrome.

Methodology. The thesis is organized around four-(4) **biological questions** that we formulated and posted, and which are directly related to SARS-CoV-2 infection, namely: (i) *Does SARS-CoV-2 exhibits a two-stage infection profile*? (ii) *SARS-CoV-1 vs. SARS-CoV-2: Do they differ*? (iii) *Does and how SARS-CoV-2 differs from Influenza infection*? and (iv) *Does SCOV2 two-stage profile relates to Covid-19 severity*? We attempted to provide answers to these questions by analyzing and exploring the **gene-expression profiles** of (preserved cell-lines or human) samples infected with wild-type SARS-CoV-2 (or SARS-CoV, the first SARS!) strains. All relevant datasets come from the GEO (gene expression omnibus) database repository of high throughput gene expression. We followed a Bioinformatics approach for the analysis of gene-expression profiles focusing mainly on the identification of **differentially expressed genes** (**DEG**) and **enrichment/pathway (EP) analysis,** organized around a **multi-step analysis pipeline**. We back-up and validate our findings with **biological interpretation** via the most relevant bibliographic references. DEG and EP analysis was performed using state-of-the-art analytical methods and tools like DESeq2, limma, EdgeR and the iDEP (integrated differential expression and pathway analysis) server. Finally, and relying on a Machine Learning framework and relevant techniques, we attempted to devise **classification models** that could forecast the **severity** of COVID-19 cases based on **predictions** for the expected duration of infection symptoms.

Results. The fundamental finding of our research refers to the identification of a **two-stage** SARS-CoV-2 infection profile, an EARLY (or, EARLY-MID) and a LATE (or, MID-LATE). These stages are **clearly differentiated** by specific up-/down-regulated DEGs and engaged molecular pathways. Most of the differentiated DEGs and enriched molecular pathways play key-roles in **fundamental host immune** and **viral defense biological processes** and are found as **down-regulated at the early stages** of the infection. In addition, the performance of the devised classification/predictive models are quite encouraging, at-least for the prognosis of the duration of infection symptoms as a marker for the severity of the disease.

Conclusions. DEG and enrichment/pathway analysis present a valuable and effective methodology to explore the molecular fingerprints of SARS-CoV-2 infection. In addition, the devise of prognostic models for the progress and severity of the infection seems feasible. Other questions that could be tackled with the same methodology in a future work concern the exploration and identification of putative antiviral drugs and their molecular targets, and even, the molecular events underlying vaccination and triggering of host immune responses.

Περίληψη

Ιστορικό σημείωμα. Στο τέλος του 2019, το Γραφείο της ΠΟΥ στην Κίνα ενημερώθηκε για μια πνευμονία άγνωστης αιτίας, η οποία εντοπίστηκε στην πόλη Wuhan στην επαρχία Hubei της Κίνας. Στις 10 Ιανουαρίου 2020 ο ΠΟΥ ανακοίνωσε το ξέσπασμα του 2019-nCoV (το πρώτο όνομα που δόθηκε στη νέα ασθένεια) και στις 11 Φεβρουαρίου 2020 ονόμασε την ασθένεια ως COVID-19 (Coronavirus Disease 2019). Ο ΠΟΥ κήρυξε την ασθένεια *έκτακτης ανάγκης για τη δημόσια υγεία διεθνούς ανησυχίας* στις 30 Ιανουαρίου 2020 και πανδημία στις 11 Μαρτίου 2020. Τα τελευταία δεδομένα από το 'Our World in Data' αναφέρουν περίπου 238 εκατομμύρια επιβεβαιωμένες λοιμώξεις SARS-CoV-2 και περίπου 4,3 εκατομμύρια θανάτους αποδιδόμενων στη λοίμωξη με SARS-CoV-2 [πρόσβαση στον 'Εξερευνητή δεδομένων της ΨΟΩΙΔ-19 του Our World in Data' στις 27 Δεκεμβρίου 2021].

Υπόβαθρο. Η παρούσα μεταπτυχιακή εργασία επικεντρώνεται στη μόλυνση από τον ιό SARS-CoV-2. Ειδικότερα, εστιάζει στην εξερεύνηση και τη διερεύνηση του μοριακού τοπίου το οποίο χαρακτηρίζει και διέπει τις αλληλεπιδράσεις γονιδίων/πρωτεϊνών και των βιολογικών διεργασιών οι οποίες λαμβάνουν χώρα κατά τα διάφορα στάδια μόλυνσης. Μια βασική αναζήτηση της εργασίας αφορά τα μοριακά μοτίβα τα οποία διαφοροποιούν τα στάδια της μόλυνσης και τα οποία μπορεί να οδηγήσουν στη λεγόμενη 'καταιγίδα κυτοκινών' η οποία βρίσκεται στη βάση του οξέος αναπνευστικού συνδρόμου της λοίμωξης.

Μεθοδολογία. Η μεταπτυχιακή εργασία οργανώνεται γύρω από τέσσερα-(4) βιολογικά ερωτήματα τα οποία διατυπώσαμε και θέσαμε και τα οποία σχετίζονται άμεσα με τη μόλυνση από SARS-CoV-2, πιο συγκεκριμένα: (i) *Εμφανίζει η λοίμωξη από SARS-CoV-2 προφίλ δύο σταδίων*; (ii) *Διαφέρουν ο SARS-CoV-1 από τον SARS-CoV-2*; (iii) *Διαφέρει και πώς ο SARS-CoV-2 από τη μόλυνση από γρίπη*; και (iv) *Σχετίζεται το προφίλ δύο σταδίων της λοίμωξης SARS-COV-2 με τη σοβαρότητα της νόσου COVID-19*; Προσπαθήσαμε να δώσουμε απαντήσεις σε αυτά τα ερωτήματα αναλύοντας και διερευνώντας τα προφίλ γονιδιακής έκφρασης δειγμάτων (διατηρημένων κυτταρικών-σειρών ή ανθρώπινων) μολυσμένων με στελέχη αναφοράς αρχέγονου-τύπου (wild-type) SARS-CoV-2 (ή SARS-CoV-1, ο πρώτος SARS!). Όλα τα σχετικά σύνολα δεδομένων προέρχονται από το αποθετήριο δεδομένων γονιδιακής-έκφρασης υψηλής απόδοσης (throughput) GEO (gene expression omnibus). Ακολουθήσαμε μια προσέγγιση Βιοπληροφορικής για την ανάλυση των προφίλ γονιδιακής έκφρασης η οποία εστιάζει στον εντοπισμό διαφορικά εκφραζόμενων γονιδίων (differentially expressed genes / DEG) και στην ανάλυση εμπλουτισμού μοριακών διεργασιών και μονοπατιών (enrichment/pathway / EP analysis), οργανωμένη γύρω από έναν αγωγό ανάλυσης (pipeline) πολλαπλών βημάτων. Τα ευρήματα μας επικυρώνονται με τη κατάλληλη βιολογική ερμηνεία τους μέσω σχετικών βιβλιογραφικών αναφορών. Η ανάλυση DEG και EP πραγματοποιήθηκε χρησιμοποιώντας αναλυτικές μεθόδους και εργαλεία Βιοπληροφορικής τα οποία αποτελούν αιχμή στο σχετικό πεδίο έρευνας, όπως τα DESeq2, limma, EdgeR καθώς και ο διακομιστής iDEP (διαδικτυακή εφαρμογή για την ολοκληρωμένη διαφορική ανάλυση γονιδιακών εκφράσεων και ανάλυσης μοριακών μονοπατιών). Τέλος, και βασιζόμενοι σε προσεγγίσεις και τεχνικές Μηχανικής Μάθησης, προσπαθήσαμε να δημιουργήσουμε μοντέλα ταξινόμησης τα οποία θα μπορούσαν να προβλέψουν τη σοβαρότητα περιπτώσεων COVID-19 με βάση προβλέψεις για την αναμενόμενη διάρκεια των συμπτωμάτων μόλυνσης.

Αποτελέσματα. Το θεμελιώδες εύρημα της έρευνας μας αναφέρεται στον εντοπισμό ενός προφίλ μόλυνσης του SARS-CoV-2 το οποίο ακολουθεί δύο στάδια, ένα EARLY (ή, EARLY-MID) και ένα LATE (ή MID-LATE). Αυτά τα στάδια διαφοροποιούνται σαφώς από συγκεκριμένα υπό-/υπέρ ρυθμιζόμενων (up-/down-regulated) DEGs και σχετικών υπό-/υπέρ ρυθμιζόμενων βιολογικών διεργασιών και μονοπατιών. Τα περισσότερα από τα διαφοροποιούμενα DEGs και εμπλουτισμένα μοριακά μονοπάτια διαδραματίζουν βασικό ρόλο στις θεμελιώδεις βιολογικές διεργασίες του ανοσοποιητικού και της ιικής άμυνας του ξενιστή/οργανισμού και βρίσκονται υπό-ρυθμισμένα στα αρχικά στάδια της μόλυνσης. Επιπλέον, η απόδοση των μοντέλων ταξινόμησης/πρόβλεψης είναι αρκετά ενθαρρυντική, τουλάχιστον για την πρόγνωση της διάρκειας των συμπτωμάτων μόλυνσης ως δείκτη για τη σοβαρότητα της νόσου.

Συμπεράσματα. Οι τεχνικές DEG και ανάλυσης/εμπλουτισμού μοριακών μονοπατιών αποδεικνύονται και αποτελούν μια πολύτιμη και αποτελεσματική μεθοδολογία για τη διερεύνηση των μοριακών αποτυπωμάτων της λοίμωξης SARS-CoV-2. Επιπλέον, η δημιουργία προγνωστικών μοντέλων για την εξέλιξη και τη σοβαρότητα της λοίμωξης φαίνεται εφικτή. Άλλα ερωτήματα που θα μπορούσαν να αντιμετωπιστούν με την ίδια μεθοδολογία σε μελλοντική εργασία αφορούν την εξερεύνηση και τον εντοπισμό δυνητικών αντ-ιικών φαρμάκων και των μοριακών στόχων τους, όπως και η εξερεύνηση των υποκείμενων μοριακών αποτυπωμάτων και συμβάντων μετά από τον εμβολιασμό για τη COVID-19 και του εντοπισμού των ενεργοποιούμενων ανοσολογικών αποκρίσεων του ξενιστή.

Purpose and Objectives

1. Identification of key genes and molecular pathways that segregates SCOV2 early and late infection stages.
2. Inference of molecular fingerprints that differentiate SCOV2 from other common viral infections (e.g., influenza).
3. Identification of key molecular imprints that characterize different SCOV2 severity phenotypes (e.g., mild vs. severe)
4. Induction and assessment of Covid-19 prognostic models that could aid therapeutic decision-making.

- To meet our aims and targets, we state and post specific biological questions that relate to SCOV2 infection. Answers to these questions are approached and realized by a series of documented bioinformatics analysis methodologies such as differential gene expression and enrichment/ pathway analysis, as well as machine-learning/ML approaches for the induction of prognostic/predictive models.
- The aforementioned tasks are accomplished by utilizing a spectrum of public-domain gene expression datasets (RNAseq and microarrays) from respective well-documented studies.

Acknowledgments

This study was conducted in the framework of the Postgraduate Program in Bioinformatics at the School of Medicine of the University of Crete, in collaboration with the Computational BioMedicine Laboratory (CBML), FORTH-ICS.

Upon completion of this Master Thesis, I would like to thank very much, my supervising professor and Principal Researcher at the Computational BioMedicine Laboratory (CBML), FORTH-ICS, Dr. George Potamias. His valuable support, guidance, organization and advice from the first moment to the end and during all phases of the thesis, brought my work to a higher level. So, I feel very lucky to have worked with him and gained a lot of knowledge in the research of Bioinformatics. I would also thank the collaborating Researcher from the Computational BioMedicine Laboratory (CBML), FORTH-ICS, Alexandros Kanterakis for the valuable advice and our excellent cooperation during the preparation of my master thesis, as well as all the professors from the Bioinformatics postgraduate program for the knowledge they provided me through their courses during my postgraduate studies.

Finally, I would like to thank my family for the support and encouragement throughout my postgraduate studies.

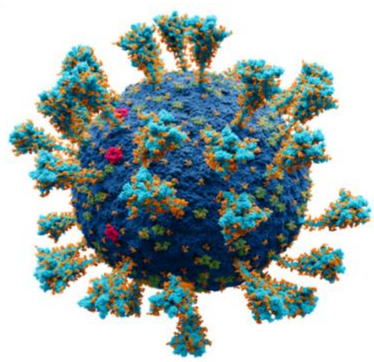
Table of Contents

1. Introduction

1.1 COVID-19: The pandemic background

Infectious diseases continue to make a huge impact on public health even though scientific research is evolving at a rapid pace. Some of these diseases are sporadic while others can cause pandemics such as bird flu and swine flu. In 2009 the World Health Organization (WHO) declared H1N1 flu as a pandemic. The H1N1 flu lasted about 1 year (March 2009 to August 2010). The first cases of the virus appeared in Mexico and then spread to the United States and around the world. Thus, 3 months after the outbreak of the virus, the number of confirmed cases reported by WHO was 94,512 in 110 countries with 429 reported deaths. After 9 months, the number of cases increased significantly, as well as the number of deaths. From the end of the H1N1 flu pandemic until today, the flu virus continues to exist and is transmitted from person to person, albeit with a lower frequency and a much lower number of confirmed deaths.

The story continues with SARS. At the close of 2019, WHO China Country Office was informed of a pneumonia of unknown cause, detected in the city of Wuhan in Hubei province, China. According to the authorities, some patients were operating dealers or vendors in the Huanan Seafood market. As of 3 January ,2020, a total of 44 patients with pneumonia of unknown etiology have been reported to WHO by the national authorities in China. Of the 44 cases reported, 11 were severely ill, while the remaining 33 patients were in stable condition. According to media reports, the concerned market in Wuhan was closed on 1 January 2020 for environmental sanitation and disinfection[1,2]. On 10 January 2020 WHO declared the outbreak of **2019-nCoV** (the first name assigned to the new disease), and on February 11 2020 it named the disease as **COVID-19** (COronaVIrus Disease 2019). WHO declared the outbreak a *Public Health Emergency of International Concern* on 30 January 2020, and a **pandemic on 11 March 2020**.



Cause, Symptomatology and Epidemics. A new type of **coronavirus** is considered as the cause of COVID-19, *severe acute respiratory syndrome coronavirus 2* (SARS-CoV-2, for now on, **SCOV2** for short). *Coronaviruses* (CoVs), named such because of the spikes on their surface when examined under a microscope, are a family of viruses that can cause illnesses such as the common cold, severe acute respiratory syndrome (SARS-CoV) and Middle East respiratory syndrome (MERS-CoV)[3]. Symptoms of COVID-19 are variable (including fever, cough, headache, fatigue, breathing difficulties, and loss of smell and taste). Most of COVID-19 infected individuals (~80%) develop mild to moderate symptoms or, they are asymptomatic; about 14% develop severe symptoms (dyspnea, hypoxia, or more than 50% lung involvement on imaging), and 5% critical symptoms (respiratory failure, shock, or multiorgan dysfunction), with older people to be at a higher risk. **It is estimated that more than a third of people who are infected do not develop noticeable symptoms, and this is one of the causes for the widespread of the disease** (Oran & Topol, 2020, 2021). Relevant studies postulate that most people affected by SCOV2 are adults; in general (>75% for ages over 18). Figure 1 shows the age distribution of COVID-19 cases in 105 countries and Greece.

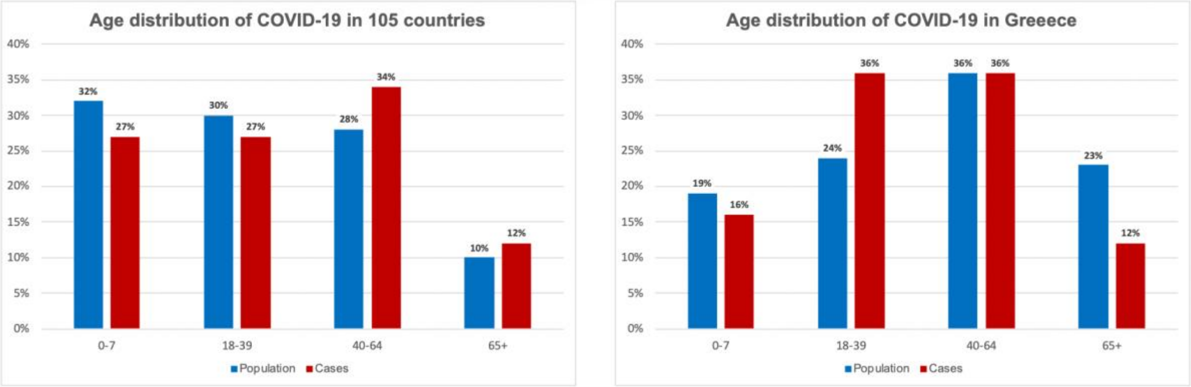
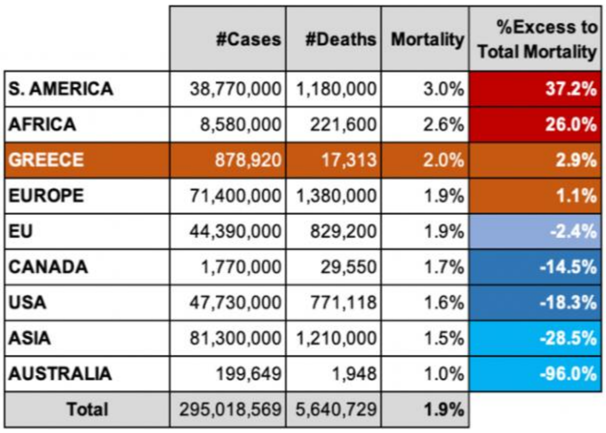
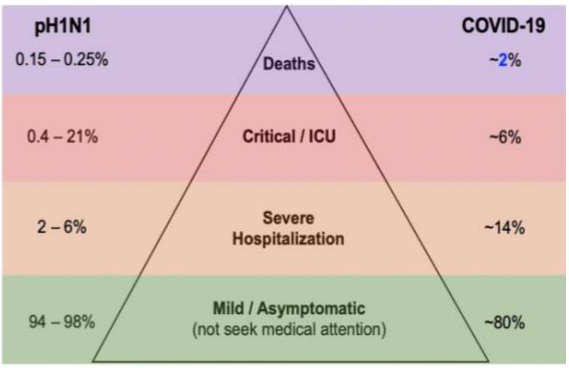


Figure 1. Age-distribution of COVID-19 cases over 105 countries (left); from data.unicef.org/resources/ covid-19-confirmed-cases-and-deaths-dashboard) and Greece (right);from EODY, 22/11/2021, eody. gov.gr/covid-gr-daily-report-20211122 and from www.statistics.gr/en/statistics/-/publication/SPO18/2020), with reference to the respective population percentages.

Mortality and Prevalence. As the COVID-19 pandemic is in progress, the figures about COVID-19 prevalence and mortality rates are still "under construction"(!). Nevertheless, searching over various sources and studies we were able to summarize some relative figures and estimates about COVID-19 incidents and deaths (see **Error! Reference source not found.**). Of interest is the comparison between Influenza (INFL) infection (all different types of the virus) and COVID-19. According to WHO, INFL results in 3-5 million serious cases worldwide every year, with about 300,000 - 650,000 deaths attributed to the disease. In addition, as the majority of INFL infected people do not seek for medical attention, it is estimated that the real INFL cases every year are about 100 times more!, i.e., ~4 billions (see the 'mild/asymptomatic' cases for the 2009 H1N1 pandemic at the right part of **Error! Reference source not found.**). Under this assumption, the INFL mortality rate is estimated at ~0.1% which, is less than the current estimates of COVID-19 mortality where, at present, about 80% of the cases (i.e., 'mild/asymptomatic') do not seek medical attention.



	#Cases	#Deaths	Mortality	%Excess to Total Mortality
S. AMERICA	38,770,000	1,180,000	3.0%	37.2%
AFRICA	8,580,000	221,600	2.6%	26.0%
GREECE	878,920	17,313	2.0%	2.9%
EUROPE	71,400,000	1,380,000	1.9%	1.1%
EU	44,390,000	829,200	1.9%	-2.4%
CANADA	1,770,000	29,550	1.7%	-14.5%
USA	47,730,000	771,118	1.6%	-18.3%
ASIA	81,300,000	1,210,000	1.5%	-28.5%
AUSTRALIA	199,649	1,948	1.0%	-96.0%
Total	295,018,569	5,640,729	1.9%	



pH1N1		COVID-19
0.15 – 0.25%	Deaths	~2%
0.4 – 21%	Critical / ICU	~6%
2 – 6%	Severe Hospitalization	~14%
94 – 98%	Mild / Asymptomatic (not seek medical attention)	~80%

Figure 2. (left) Number of reported cases and deaths attributed to COVID-19 –data are aggregated from 'OurWorldInData' excellent source of relative information (cases, deaths); (right) related estimates for deaths, critical/ICU, sever and mild cases that contrast between the H1N1 pandemic at 2009 (pH1N1) and COVID-19 respective figures, from (da Costa et al, 2020).

Here we have to make an important note regarding the COVID-19 mortality figures reported in **Error! Reference source not found.**. The *true severity* of a disease as a measure for its mortality can be described by the **Infection Fatality Ratio (IFR)**[1]:

[1] "Estimating mortality from COVID-19", WHO report, 4 August 2020.

$$\mathbf{IFR}(\%) = \frac{\#\text{Deaths\_from\_disease}}{\#\mathbf{infected\_individuals}}$$

But it is difficult to measure the 'true' number of "*infected individuals*", not only because the pandemic is still on, but mainly, because of the ***difficulty to assess the asymptomatic cases***, and as a consequence, the ***difficulty to assess the prevalence of the disease***. As a proxy, the **Case Fatality Rate** (**CFR**) can be utilized. It takes in consideration not, the difficult to assess, number of "infected individuals", but the number of "*confirmed cases*":

$$\mathbf{CFR}(\%) = \frac{\#\text{Deaths\_from\_disease}}{\#\mathbf{confirmed\_cases}}$$

Under the above observations, the mortality figures reported in **Error! Reference source not found.** reflect IFR estimates.

In any case, an exact estimate for COVID-19 prevalence is still pending, and results from various studies are already published (mainly on regional, country, city, hospital etc.) levels. Some recent prevalence studies (see Table 1) are reported and used for the induction of a COVID-19 prevalence model (Toulis, 2021); all the studies refer to 2020. As it can be observed the figures are quite diverging, leaving the question about the 'true' prevalence rates of COVID-19 an open problem.

Table 1. COVID-19 prevalence studies; midpoints (from the reference studies) are reported

Prevalence	Location	Month (2020)	Method	Notes (Publication)
6.14%	China	1	PCR	Used 80% of entire population in Vó, Italy (Lavezzo et al., 2020)
2.60%	Italy	2	PCR	Used 131 patients with ILI symptoms (Spellberg et al., 2020)
5.30%	USA	3	PCR	Sample of 215 pregnant women in NYC (Sutton et al., 2020)
13.70%	USA	3	PCR	(Yadlowsky et al., 2020)
9.40%	Spain	3	SER*	Sample of 578 healthcare workers (Garcia-Basteiro et al., 2020)
3%	Japan	3	SER	Random set of 1000 blood samples in Kobe Hospital (Doi et al., 2020)
36%	USA	4	PCR	Study in large homeless shelter in Boston (Baggett et al., 2020)
1.50%	USA	4	SER	Recruited 3330 people via Facebook (Bendavid et al., 2020)
9.10%	Switzerland	4	SER	Uses ILINet data; implies 96% unreported cases (Lu et al., 2020)
14%	Germany	4	SER	Sample of 1335 individuals in Geneva (Stringhini et al., 2020)
3.10%	Netherlands	4	AntG*	Self-reported 400 households in Gangelt. (Streeck et al., 2020)

*SER: Serological; AntG: Antigen

From a pandemic to an endemic disease? In a recent paper is *Science* (Lavine, Bjornstad, & Antia, 2021) some interesting results are reported regarding the immunological characteristics of COVID-19 and its) transition to an **endemic** disease (see Figure 3 at the next page). The authors postulate three rational assumptions that support their hypothesis and estimates: (i) faster *transmission* results in a quicker transition to the endemic state but more total deaths; (ii) *social distancing* saves lives, delays endemicity and allows crucial time for vaccine roll-out, and (iii) *vaccination* speeds up the transition to the endemic state and reduces the death toll. Their modeling framework provides forecasts about the progress of SCOV2 IFR figures in a time scale of 2.5, 5, 7.5 and 10 years with relation to different disease **reproduction numbers** (R_0), and contrasted with respective figures for **SCOV1** (the 1st SARS!) and MERS-CoV. Furthermore, the guide researcher of the publication (Jennie Lavine) provides a very interesting figure which, summarizes the authors' forecast. The forecast states that **COVID-19 will reach in about 2.5 to 3 years after its emergence into an IFR of 0.001 (the influenza rate!), and even less in the coming years.** Of course, this will happen with a **virus spread up to ~98% in the general population** (follow the blue dotted line)**,** either by early childhood infections (with no or low symptomatology) and/or mass vaccination programs[2].

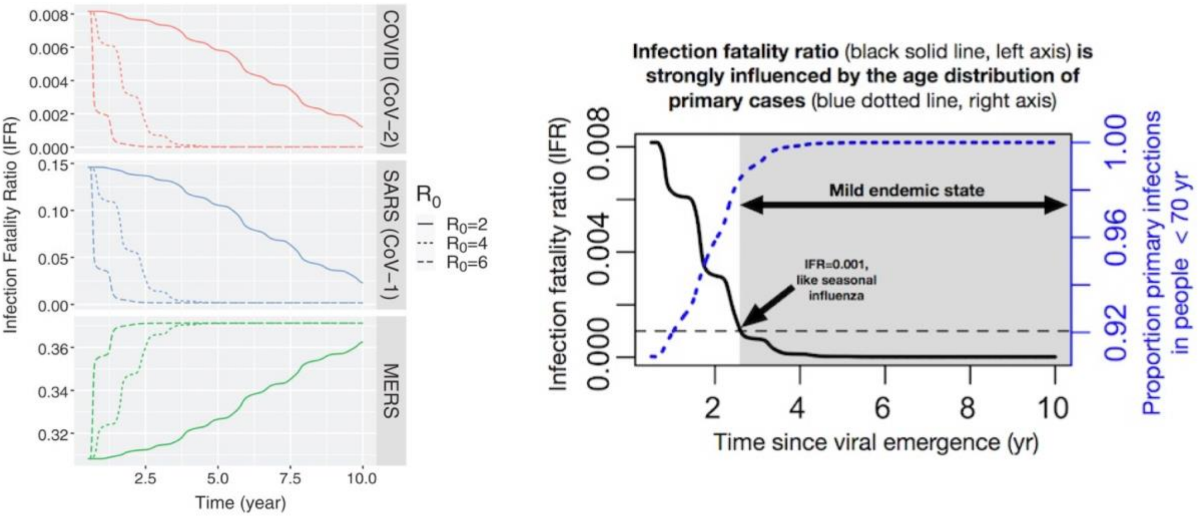


Figure 3. Forecasts for the progress of SCOV2 infection compared to SCOV1 and MERS-CoV IFR figures (left); endemicity of SCOV2 infection is strongly influenced by the virus spread (right).

1.2 A quick look at the molecular background of SARS-CoV-2 infection

The pathogenesis of SCOV2 infection (not excluding other viruses) is quite complex, and the exact reasons of its fatality are still under exploration. SCOV2 shares many clinical features with INFL, both in terms of the transmission routes, as both spread very easily between people through oral and nasal drops, as well as in terms of symptoms, because both affect the respiratory system. But, there is a major difference, **the severe and acute respiratory defects that comes as a syndrome to SCOV2 infection does not occur with INFL infection**. So, a key question relates to the mechanisms underlying, govern and guide the SCOV2 syndrome as a consequence of the so-called "***cytokine storm***" (Fajgenbaum & June, 2020).

The cytokine-storm. Many studies have made evident that the cytokine-storm plays a decisive role in the progress of SCOV2 infection and is an important factor for severe and fatal outcomes (Chen et al., 2021). The cytokine-storm encompasses an excessive immune response elicited between cytokines and immune system cells. Under infective events, a strong and healthy host immune system releases more than 150 (pro)-inflammatory cytokines. Complex regulations may cause **uncontrolled** pro-inflammatory and anti-inflammatory cytokines to be elevated in the blood serum, with lethal interactions. Cytokine-storm may occur in many viral infections (e.g., Ebola virus, Dengue virus, H1N1 flu virus, SCOV1 and MERS-CoV), in hematopoietic diseases as well as after the use of certain drugs. it can cause the

[2] Refer to "Just another common cold virus? Modeling SARS-CoV-2's future fade"

severe acute respiratory distress syndrome (ARDS) and multiple organ failure. It is the leading cause of death for many diseases. Studies have also shown that COVID-19 patients with severe symptoms exhibit much higher levels of white blood cells, neutrophils, procalcitonin and other inflammatory markers compared to patients with mild symptoms (Tang et al., 2020). The postulate states: **cytokine-storm presents a systemic inflammatory response to infections that leads to over-activation of immune cells and the uncontrolled production of pro-inflammatory cytokines**. (Costela-Ruiz, Illescas-Montes, Puerta-Puerta, Ruiz, & Melguizo-Rodríguez, 2020). See Figure 4 for a high-level illustration of events taeking place during the cytokine-storm situation.

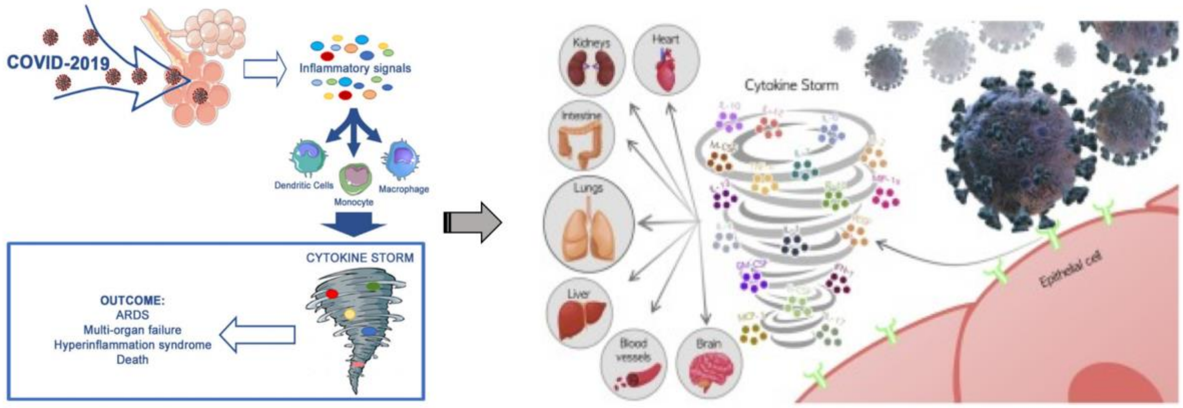


Figure 4. Immune host response during SCOV2 infection (left); adapted from (Castelli, Cimini, & Ferri, 2020); Cytokine-storm causes organ injury (right); adapted from (Costela-Ruiz et al., 2020)

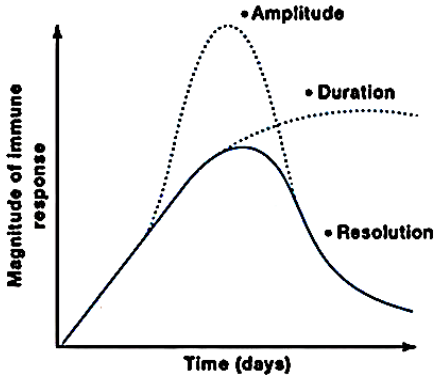
Inflammatory and pro-inflammatory cytokines at a glance. Cytokines include *interferons*, *tumor necrosis factors* (TNFs), *interleukins*, and *chemokines* (see Appendix I for a full annotated list.) Table 2, is adapted from (Rabaan et al., 2021) and shows the most significant inflammatory factors involved in the cytokine-storm and their functional roles.

Table 2. Significant inflammatory factors involved in the cytokine storm and their function.

Inflammatory Factors	Function
Interferons	Induces immunity and directs the expression of antiviral protein by expressing specific coding genes.
TNF Tumor Necrosis Factor	Released during severe infection and is an important indicator for chronic inflammatory and autoimmune diseases.
Interleukins	They help differentiate and activate cells and act as transporters of immune cells against infection. They also signal the production of secondary cytokines and are an indicator of the severity of the disease.
Chemokines	They help transport immune cells, regulate cell growth and differentiation, and control the immune response. Finally, they are an important indicator for controlling the severity of ARDS syndrome.

These cytokines regulate host defense responses and play an important role in mediating innate immune responses, mainly by regulating inflammatory reactions. Pro-inflammatory cytokines act to aggravate the disease, while anti-inflammatory cytokines act to heal inflammation. Their excessive chronic production contributes to inflammatory diseases such as atherosclerosis and cancer, while their deregulation has been linked to the occurrence of neurological diseases. So, a **balance between pro-inflammatory and anti-inflammatory cytokines to maintain a healthy state.**

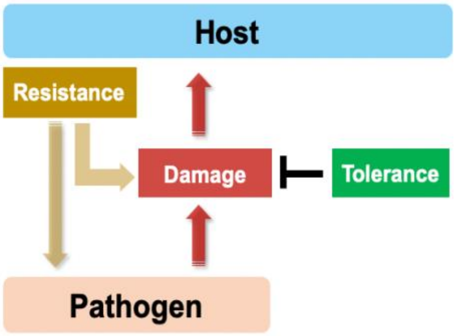
1.3 A closer look at SCOV2 infection progress



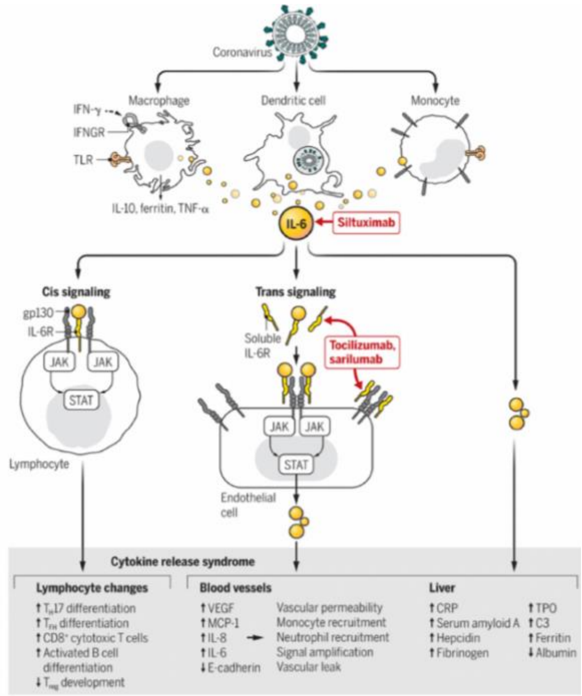
It is reported, and in a large-extend established, that progress of COVID-19 disease presents a **two-stage** profile, which is directly linked to **immune suppression** actions (Tian et al., 2020). Most of the cases present a *mild-to-moderate* phenotype at the early stages, typically around seven to nine days after the initial onset of symptoms, followed by a *severe* phenotype presented with worsening of respiratory function (Hadjadj et al., 2020). Various studies demonstrate that SCOV2 causes a **suppression of immune responses at the early stage** of infection, which results to an *uncontrolled replication of the virus*. The early immune suppression is attributed to ***defective type I IFN (IFN-I, a special type of interferons) immunity in the first hours and days of infection that leads to uncontrolled viral replication, with subsequent excessive leukocyte recruitment, underlying uncontrolled inflammation*** (Q. Zhang, Bastard, Bolze, et al., 2020). A sketch that provides a background to such an infection progress profile (i.e., from early to late stage) is provided in the above side figure that shows and contrasts between the *magnitude of normal and impaired immunity reactions to pathogen infections*. The solid line shows the *natural* course immune response to infection over a period of days to weeks that then leads to a *resolution* phase as the pathogen is controlled (the figure is adapted from (Mangalmurti & Hunter, 2020)). The largely ***uncontrolled replicative profile*** of SCOV2 conforms to such an infection progress profile, clearly indicated and stated in (Mangalmurti & Hunter, 2020): "… *for microorganisms with a high replicative potential*, *changes in the magnitude and duration of the immune response can result in systemic immune pathology associated with a cytokine storm through either an increased amplitude or a failure to enter the resolution phase*". It is essential to note that for severe COVID-19 cases with extended hospitalization, death events occur mostly after the virus is cleared, a fact that designates the ***continuation of host resistance mechanisms even if the threat is eliminated***! (Lee et al., 2009). In recent study that included patients died from severe pneumonia, a two-phase COVID-19 disease evolution is also suggested, and it is also linked with *viral RNA loads* (Desai et al., 2020).

In order to understand the COVID-19 disease we have to gain insight into the physiology and molecular mechanisms that putatively govern and guide the transition from an early to a late progress of the disease.

1.3.1 Host defense response mechanisms



A key to unlock and characterize the transition from suppressed immune responses at the early stage of SCOV2 infection to their over-activation at late stages, could be given in terms of two fundamental host defense strategies, namely: ***resistance*** and ***tolerance*** (D. S. Schneider & Ayres, 2008), with an interplay between them taking place (see side figure adapted from: (Carvalho et al., 2012)). Disease resistance engages various physiological and molecular host immunity processes (both innate and adaptive) and targets the *reduction of the pathogen's load*. Disease *tolerance* triggers host responses aiming to constrain the damages in affected tissue and support its function, and in that sense, it '*tolerates*' the pathogen's burden. In the upper respiratory tract (URT), the initial entrance and replication site for SCOV2 (and other viruses as well), the tolerance defense mechanisms aim to sustain the exchange of gas and blood oxygenation. In the case of URT infection, host resistance triggers a dynamic immune response in order to constrain the spread of the virus, with *type I and type III interferons* to play a crucial role (see next sections for a more detailed explanation of their roles). When resistance is weakened and the virus spreads to the lower respiratory tract (LRT), tolerant defense processes are triggered in order to preserve the alveolar structures which are crucial for the exchange of gas. It is indicative that in severe COVID-19 cases with extended hospitalization, death events occur mostly after the virus is cleared, a fact that designates the ***continuation of host resistance mechanisms even if the threat is eliminated***! (Lee et al., 2009). In a recent study that included succumbed patients with severe pneumonia a two-phase COVID-19 disease evolution is suggested which also depends on the presence of high viral RNA loads (Desai et al., 2020). As a first hint to the molecular background of defense and resistance mechanisms to SCOV2 infection have a look at **Error! Reference source not found.**. It illustrates the key-molecular events and gene/protein products (with putative drug targets) associated with the SCOV2 ***cytokine release syndrome*** during the progress of the infection (Moore & June, 2020).



C3: complement; CRP: C reactive protein; IFN-γ: interferon-γ (IFNG); IL: interleukin; IL-6R: IL-6 receptor; JAK: Janus kinase; MCP-1: monocyte chemoattractant protein; STAT3: signal transducer and activator of transcription 3; T_{FH}: T follicular helper cell; T_H17: T helper 17 cell; TNF-α: tumor necrosis factor-α; TLR: Toll-like receptor; TPO: thrombopoietin; T_{reg}: T regulatory cell; VEGF: vascular endothelial growth factor.

Figure 5. Key-molecular events and gene/protein products engaged with SCOV2 cytokine syndrome

1.3.2 The SCOV2 molecular framework

In the course of SCOV2 infection various host defense processes are triggered including direct antiviral molecules and inflammatory mediators, such as type I/III interferons, tumor necrosis factor (TNF), interleukin 6 (IL-6) and other chemokines, with pathogen-associated molecules to be recognized by dedicated recognition receptors (e.g., retinoic acid-inducible gene I, RIG-I) and Toll-like receptors (e.g., TLR3). As in most RNA viruses, SCOV2's RNA is recognized by various genes with the role to trigger a series of pathways that lead to the production of type I interferons (IFN-I), other **interferon stimulated genes** (**ISG**s), as well as various proinflammatory **cytokines** (Tan, Sun, Chen, & Chen, 2015)[3]. But SCOV2 has employed several different mechanisms to escape the IFN response, as for example, its non-structural protein 1 (NSP1) binds host's 40S ribosomal subunit to shutdown mRNA translation of IFNs and ISGs (Thoms et al., 2020). It is well-established that not only SCOV2, but other viral infections interfere with interferon signaling (Lei et al., 2020). But, in contrast to other viral infections (influenza for example; see section 0 for relative experiments and results), it is now clear that COVID-19 patients demonstrate **downregulation of interferon signaling pathways at the early stages** of the infection (Mudd et al., 2020). As a guide to the aforementioned observations, Figure 6 illustrates the key molecular processes and genes engaged and triggered in the course of SCOV2 infection (Y.-M. Kim & Shin, 2021):

[3] Cytokines include chemokines, interferons, interleukins, and tumor necrosis factors, see Appendix I for a full annotated list.

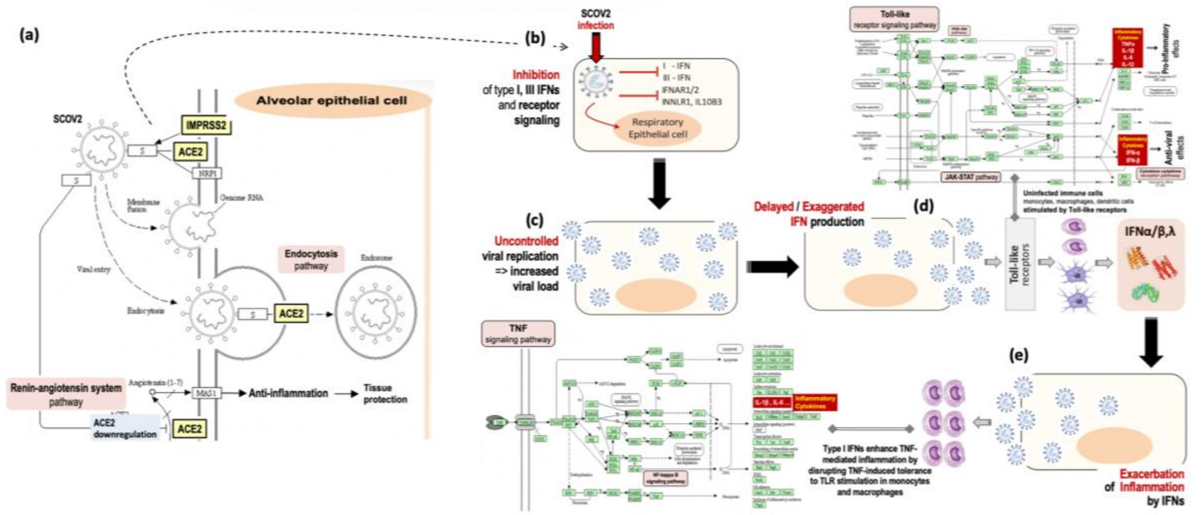

(a)
SCOV2
Alveolar epithelial cell
IMPRSS2
ACE2
S
NRP1
Membrane fusion
Genome RNA
Viral entry
Endocytosis
Endocytosis pathway
Endosome
Renin-angiotensin system pathway
Angiotensin (1-7)
MAS1
Anti-inflammation
Tissue protection
ACE2 downregulation
(b)
SCOV2 infection
Inhibition of type I, III IFNs and receptor signaling
I - IFN
III - IFN
IFNAR1/2
INNLR1, IL10B3
Respiratory Epithelial cell
(c)
Uncontrolled viral replication => increased viral load
Delayed / Exaggerated IFN production
(d)
Toll-like receptor signaling pathway
JAK-STAT pathway
Inflammatory Cytokines
Pro-inflammatory effects
Anti-viral effects
Uninfected immune cells monocytes, macrophages, dendritic cells stimulated by Toll-like receptors
Toll-like receptors
IFNα/β,λ
TNF signaling pathway
IL-1β, IL-6 ...
Inflammatory Cytokines
NF-kappa B signaling pathway
Type I IFNs enhance TNF-mediated inflammation by disrupting TNF-induced tolerance to TLR stimulation in monocytes and macrophages
(e)
Exacerbation of Inflammation by IFNs

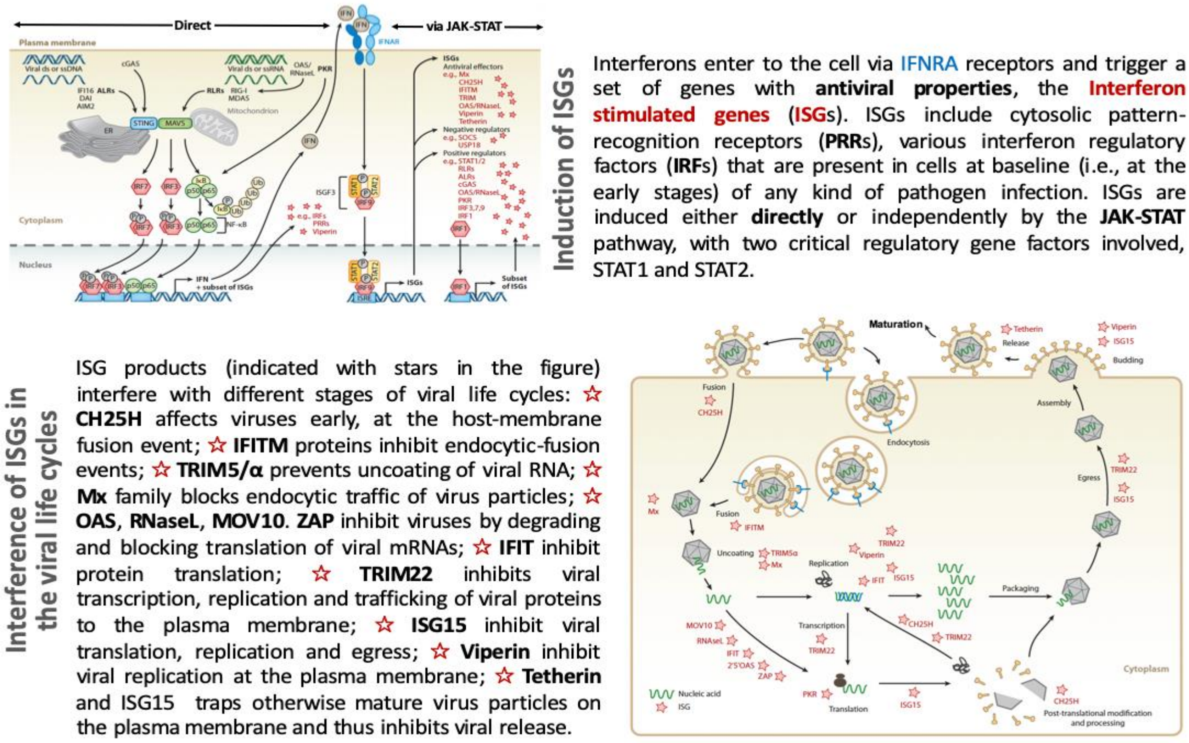
Figure 6. The molecular regulatory framework of SCOV2 infection progress; adapted from (Y.-M. Kim & Shin, 2021).

(a) The part of KEGG Coronavirus pathway that relates to the initial (alveolar) epithelial cells infection; interference and role of the key *ACE2* and *IMPRS22* genes is highlighted in various pathway processes; note the blockage of 'anti-inflammation' and 'tissue-repair processes' caused by the interference of *ACE2* => **(b**,**c)** Respiratory epithelial cells are infected, **type I/III IFN responses are blocked** (b), and viral load increases (c) => **(d)** Uninfected innate immune cells (*monocytes*, *macrophages*, and *dendritic cells*) are stimulated by viral components via Toll-like receptors (the KEGG Toll-like receptor signaling pathway is shown) and produce type I/III IFNs => **(e)** Type I/III IFNs further induce the accumulation and activation of monocytes and macrophages, leading to the production of large amounts of IFNs and proinflammatory cytokines; type I IFNs also enhance TNF-mediated inflammation by disrupting TNF-induced tolerance to TLR stimulation in monocytes and macrophages (the KEGG RIG-I-like receptor signaling pathway is shown); the figure is adapted from (Y.-M. Kim & Shin, 2021) and was accordingly enhanced and annotated.

1.3.3 Going deeper: Deficiency of early IFN response in SCOV2 infection

During viral infection the first task of host cells is to identify virus invasion via the recognition of particular viral molecular structures, the so-called pathogen-associated molecular patterns / *PAMPs*; a type of foreign viral RNAs introduced or produced in viral life cycle to trigger innate immune responses (Min et al., 2021). In this course type I interferons / **IFN-I**[4] are engaged with the role to block virus replication at many levels, via the triggering of **interferon simulated genes** / **ISGs**, a set of viral replication blocking genes (Katze, He, & Gale, 2002; McNab, Mayer-Barber, Sher, Wack, & O'Garra, 2015). But a series of studies indicate that IFN and especially IFN-I responses are weak during SCOV2 infection. As a consequence to such IFN-I defected response, productive viral replication take place at the early stage of SCOV2 infection, a situation that greatly contributes to COVID-19 pathology and severity (Hadjadj et al., 2020; Q. Zhang, Bastard, Liu, et al., 2020). Figure 7 **highlights the different induction forms of ISGs as well as their role in the course of viral life cycle, presenting a canvas on which we may study, assess and finally uncover the unique molecular characteristics of SCOV2 infection**.

[4] IFN-I is a pleiotropic cytokine composed by a family of IFN proteins which are encoded by at least 13 IFNα (IFNA) subtype genes (IFNA1, -A2, -A4, -A5, -A6, -A7, -A8, -A10, -A13, -A14 -A16, -A17 and -A21), and IFNB1, IFNE, IFNK and IFNW1 genes, all of which bind to IFNAR1 and IFNAR2 receptors (López de Padilla & Niewold, 2016).



Induction of ISGs

Interferons enter to the cell via IFNRA receptors and trigger a set of genes with **antiviral properties**, the **Interferon stimulated genes** (**ISG**s). ISGs include cytosolic pattern-recognition receptors (**PRR**s), various interferon regulatory factors (**IRF**s) that are present in cells at baseline (i.e., at the early stages) of any kind of pathogen infection. ISGs are induced either **directly** or independently by the **JAK-STAT** pathway, with two critical regulatory gene factors involved, STAT1 and STAT2.

Interference of ISGs in the viral life cycles

ISG products (indicated with stars in the figure) interfere with different stages of viral life cycles: ☆ **CH25H** affects viruses early, at the host-membrane fusion event; ☆ **IFITM** proteins inhibit endocytic-fusion events; ☆ **TRIM5/α** prevents uncoating of viral RNA; ☆ **Mx** family blocks endocytic traffic of virus particles; ☆ **OAS**, **RNaseL**, **MOV10**. **ZAP** inhibit viruses by degrading and blocking translation of viral mRNAs; ☆ **IFIT** inhibit protein translation; ☆ **TRIM22** inhibits viral transcription, replication and trafficking of viral proteins to the plasma membrane; ☆ **ISG15** inhibit viral translation, replication and egress; ☆ **Viperin** inhibit viral replication at the plasma membrane; ☆ **Tetherin** and ISG15 traps otherwise mature virus particles on the plasma membrane and thus inhibits viral release.

Figure 7. Induction of interferon stimulated genes/ISGs (upper-left) and their interference into the viral life cycle (down-right); figure and commentary captions accordingly adapted from (W. M. Schneider, Chevillotte, & Rice, 2014).

Based on the observations made so far, it should be clear that in order to **elucidate SCOV2 infection and uncover the underlying molecular regulatory background we have to focus on the evolution and progress stages of the disease**. With this as a guide we state the fundamental aims and targets of the thesis.

Aims and Targets of the Thesis

1. **Identification of key genes and molecular pathways that segregates SCOV2 early and late infection stages.**
2. **Inference of molecular fingerprints that differentiate SCOV2 from other common viral infections (e.g., influenza).**
3. **Identification of key molecular imprints that characterize different SCOV2 severity phenotypes (e.g., mild vs. severe)**
4. **Induction and assessment of Covid-19 prognostic models that could aid therapeutic decision-making.**

Methodology

- ❑ **To meet our aims and targets, we state and post specific biological questions that relate to SCOV2 infection. Answers to these questions are approached and realized by a series of documented bioinformatics analysis methodologies such as differential gene expression and enrichment/ pathway analysis, as well as machine-learning/ML approaches for the induction of prognostic/predictive models.**
- ❑ **The aforementioned tasks are accomplished by utilizing a spectrum of public-domain gene expression datasets (RNAseq and microarrays) from respective well-documented studies.**

2. Computational Framework and Bioinformatics Analysis Pipeline

For our experiments and analysis, we utilized the well-known and widely utilized **iDEP** server. iDEP is an R-Shiny web-based application equipped with state-of-the-art bioinformatics methodologies via the exploitation and smooth integration of respective R-packages (Ge, Son, & Yao, 2018): (a) **differential expression gene** (DEG) analysis, e.g., *DSEeq2*, *voom*, *EdgeR* (Ge et al., 2018); (b) **exploratory** analysis (e.g., *k-means* clustering, *PCA* and *MDS*); (c) **enrichment/pathway analysis** (e.g., *GSEA*); (d) advanced **visualization** capabilities (e.g., heatmaps, hierarchical clustering trees, enriched pathway maps). A unique and very useful function of iDEP is that it can *read and convert gene/transcript/probe annotations from various gene-expression platforms* (RNAseq or microarray). The full spectrum of iDEP analytical operations can be inspected at idepsite.wordpress.com. Figure 8 Illustrates the overall iDEP workflow with all relevant analytical operation utilized in the current thesis.

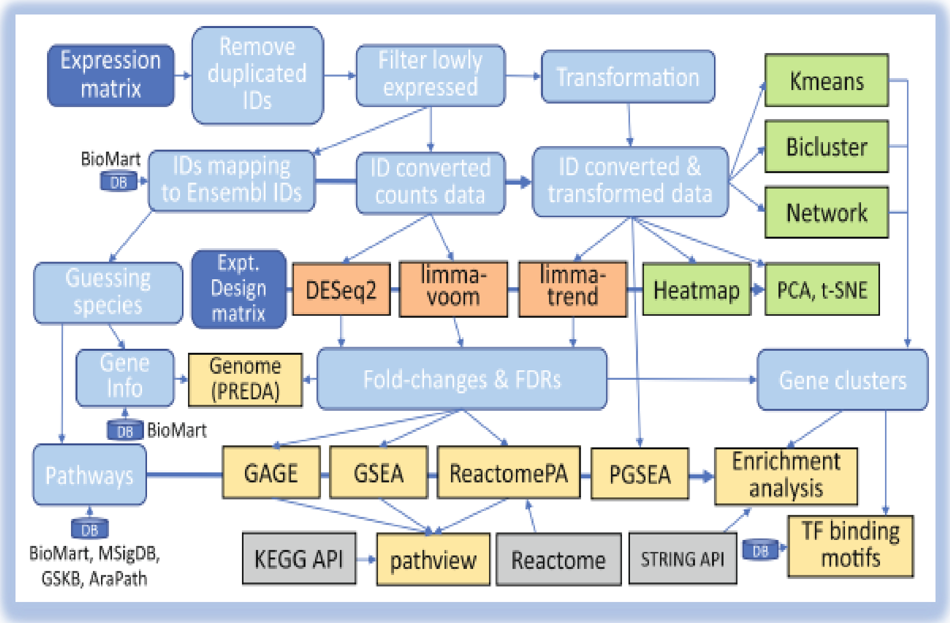


Figure 8. The overall iDEP analysis workflow (adapted from the iDEP web-server site) with analytical operation utilized in the current thesis.

To support our analyses we also utilized a series of other well-established genomics platforms including Ensembl/Biomart and g:profiler for the conversion/annotation of gene IDs between different gene-expression platforms, and STRING for functional annotation, clustering and visualization of associations between genes and proteins.

2.1 Analysis pipeline: Outline

❑ **Data Pre-Processing.** Gene-expression (microarray and/or RNA-seq) data are loaded to the iDEP duplicate gene/transcript IDs. **Gene/Transcripts Filtering.** The low expressed genes are filtered-out according to a (user-specified) minimum expression value; in most of our experiments we set a cutoff so that 30% of the low expressed gene/transcripts are discarded. For the genes/transcripts not recognized by iDEP we utilized the g:profiler and Ensembl/Biomart servers and services. **Normalization.** The reliability of gene-expression analysis results is heavily depended from the distribution of gene-expression values across samples iDEP offers various normalization methods, including DSEeq2's **vst** (variance stabilizing transformation)[5] and EdgeR's **log2(CPM+c)**[6] transformations, which according to the task at hand were appropriately utilized. **Visualization/Quality control.** The **box-** and **density-plots** of transformed data are visualized in order to evaluate the well-formed distribution of gene-expression values and get the respective mean and median values (these values are needed in further analysis in order to contrast between the gene-expression levels of differential expressed genes).

❑ **Exploratory analysis.** First the input genes are ranked according to their correlation/distance across all samples, and then, using the top genes (a user-specified value, in most of our experiments we used 1000 genes) the **hierarchical clustering tree of samples** is formed. This is done in order to get at a glance the contrasting profile between the different samples' classes. This is further supported by the respective samples' **PCA** (principal component analysis) and **MDS** (multi-dimensional scaling) plots. Hierarchical clustering, PCA and MDS are unsupervised methods for examining relationships between samples. These methods are suitable for exploratory analysis because class membership of samples is used to obtain the graphical representation. The purpose of the hierarchical clustering algorithm is to separate the data into homogeneous groups. In such a clustering the measure of similarity can have a very large effect on the result as the algorithm consists of the measure of similarity or dissimilarity between each pair of samples. In a hierarchical grouping tree, each leaf corresponds to a sample. The samples which are similar to each other are combined into branches. The height of the fusion on our vertical axis shows the distance of similarity or dissimilarity between the samples. MDS is a statistical method used in order to get insight into *hidden variables* that putatively explain (dis-)similarities between analyzed objects. MDS doesn't assume a specific relationship between objects and aims to optimize the fit of dissimilarities in the MDS (usually Euclidian) space. On the other hand, PCA assumes a linear relationship between the data objects which is represented by the induced *principal components* (PCs). As PCs do not maximize the separation between groups, PCA and MDS may provide different results. So, we utilize both plots in order to get insight into putative interdependencies in data. We also utilize the **Correlation Matrix** of samples in an effort to identify **coherently contrasting profiles between specific samples, and then use these samples for further analysis.**

❑ **Differential expression analysis.** The well-known and widely utilized R packages **DESeq2** (Anders, 2010) and/or **limma** (Smyth, Ritchie, & Thorne, 2011) were utilized for the identification of differentially expressed genes/**DEG**, setting a fixed minimum **FDR** level (0.05 was used in most cases), and different minimum **fold-change/FC** thresholds (in most cases the value of 2 was set). For datasets with big numbers of DEGs we used bigger FC thresholds in order to *keep the identified DEGs in manageable numbers so that their biological interpretation is eased*. The DEG's **heatmap** as well as the respective **Volcano**, **M-A** and **Scatter** plots were used in order to visually inspect the degree to which the DEGs separate the different samples (a nice introduction to these plots and their use in the interpretation of gene-expression analysis results is provided in (McDermaid, et al, 2019)).

❑ **Enrichment/Pathway analysis.** The Up- and Down-regulated DEGs are subjected to **Enrichment** and **Pathway Analysis**, founded on the utilization of *hypergeometric distribution* to compute the p-value of pathways in which the identified DEGs occur (Falcon

[5] Analyzing RNA-seq data with DESeq2

[6] edgeR: differential analysis of sequence read count data - User's Guide

& Gentleman, 2008). Pathway analysis was performed utilizing various methods available in iDEP, e.g., *fgsea* Bioconductor package.

- ❑ **Network analysis.** The STRING server (Franceschini, 2013) was used in order to find significant connections and relations between the identified DEGs and form the respective **genes' interaction network**. This network provides a kind of 'orchestration' of the molecular events underlying SCOV2 infection and proved a valuable tool for the biological interpretation of our results and findings.

- ❑ **SCOV2 prognosing modeling – The Machine Learning framework.** For the devise of **predictive models** that could be used for the prognosis of SCOV2 severity we relied on the **MLSeq** R/Bioconductor package – a Machine Learning interface to RNA-seq data analysis (Goksuluk et al., 2019). We used four MLseq classifiers to build our initial models: **SVM** (support vector machines), **Random Forests** (**RF**), **Voom Based Nearest Shrunken Centroids** (**VoomNSC**) (Zararsiz et al., 2017), and **Poisson Linear Discriminant Analysis** (**PLDA**) (Witten, 2011), following either a **train/test split** or a **k-fold cross-validation** procedure. In a second line of experiments, we utilized the **Weka** environment. *Weka* is an open-source software implemented in Java incorporating a collection of machine learning algorithms for data mining tasks, including tools for data preparation, classification, regression, clustering, attribute selection and other (Hall et al., 2009). For the devise of prognostic models, we utilized the SVM (named SMO in Weka), Random Forests / RF, **Decision Tree** / **DT** and **kNN** (named iBK in Weka) algorithmic approaches. All experiments were conducted following a standard cross-validation procedure and in particular, Leave-One-Out-Cross-Validation (LOOCV).

- ❑ **Gene-expression datasets.** The utilized datasets refer to samples infected with SCOV2 or SCOV1, and all come from the NCBI's **Gene Expression Omnibus** (**GEO**) database. The gene-expression profiles are acquired either by **microarray** of **NGS/RNA-seq** platforms and refer either to ***cell-line*** or ***human tissue***. In the current thesis we utilized four GEO gene-expression datasets that cover a wide range of tasks related to SCOV1/2 infection and COVID-19 disease, namely:

  - **GSE151513**. Contains human lung epithelial cells from specific cell-lines[7] infected with SCOV2 (uninfected control samples are also included). ***RNAseq*** gene-expression profiles were acquired via the 'Illumina HiSeq 2500' platform. Cell-line samples were collected at different *hours post-infection* (hpi), namely 0, 1, 2, 3, 6 and 12 in triplicates (i.e., a total of 3 x 6 = 18 samples). This dataset was used in order **to explore (and finally validate) the two-stage profile of SCOV2 infection**.

  - **GSE33267**. Contains a total of 99 human lung epithelial cell-line samples (mock-treated samples are also included) infected with the wild-type SCOV1 strain, with the respective gene-expression profiles acquired via a specific Agilent ***microarray*** platform. Cell-line samples were collected at different hpis, namely: 3, 7, 12, 24, 30, 36, 48, 54, 60 or 72 hours. This dataset was used in order **to explore (and finally validate) the resemblance between the molecular profiles underlying the early stages of the first SCOV1 and SCOV2.**

  - **GSE47960**. Contains a total of 163 HAE (human airway epithelial) cell cultures which were infected with the first SARS-CoV and H1N1 influenza strains, with samples collected at different hpis: 0, 12, 24, 36, 48, 60, 72, 84 and 96 for SCOV, and 0, 6, 12,

[7] *Human cell lines are commonly used for **research investigation**. For decades, cell lines have been the workhorse of programs to identify and interrogate mechanisms of action, discover and/or test drug/compounds/factors, and show relevance of findings to human disease*. "Cell lines", Elsevier/Science Direct

18, 24, 36 and 48 hrs for H1N1, as influenza progresses more quickly than SARS viruses! The gene-expression profiles were acquired via a specific Agilent ***microarray*** platform. This dataset was used in order **to explore and contrast the progress of infection between SCOV and Influenza/H1N1.**

- **GSE161731**. Contains a total of 77 samples from respective COVID-19 patients, with the ***RNA-seq*** gene-expression profiles acquired via the 'Illumina NovaSeq 6000' platform Whole blood samples were collected between 1-35 days post infection and divided based on disease severity and time from symptom onset. This dataset was used: (a) in order to explore and **contrast between the molecular backgrounds of COVID-19 patients that exhibit a short vs. medium vs. long periods of symptoms**; and (b) in order to devise predictive models that could **forecast the duration of symptoms for COVID-19 patients**, and from that to **estimate the putative severity** of the disease.

3. Uncovering SCOV2 infection: A bioinformatics approach

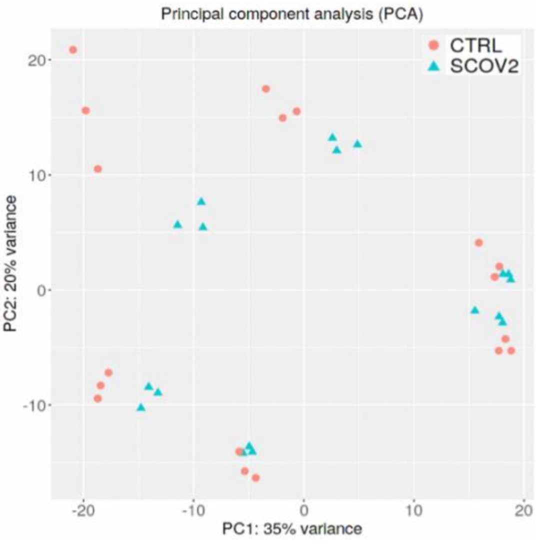
The sections that follow posts specific **biological questions** related to SCOV1/2 infection which are tackled with a carefully designed Bioinformatics methodology. The main task is to **uncover and identify the key-molecular events underlying SCOV2 infection** and provide respective answers to the posted biological questions.

3.1 Does SARS-CoV-2 exhibits a two-stage infection profile?

Dataset. The utilized dataset was download from the Gene Expression Omnibus (GEO) under the code GSE151513. In this study, human lung epithelial cells (Calu-3[8]) were infected with SCOV2 (18 samples) or kept uninfected (i.e., mock treated / CTRL, 18 samples), and their **RNAseq** gene-expression profiles were acquired via the 'Illumina HiSeq 2500' platform (GPL16791). Cell samples were collected at different six post-infection time-points/hours (hpi), namely 0, 1, 2, 3, 6 and 12 in triplicates (i.e., a total of 3 x 6 = 18 samples for SCOV2 and 18 for CTRL), with a *multiplicity of infection*[9] (MOI) of 2. The genes are presented as Ensembl transcripts. The results of this study are published in (Banerjee et al., 2021).

3.1.1 SCOV2 vs. CTRL: Could they be contrast?

We tried to induce differential expressed genes for SCOV2 vs. CTRL comparison. In other words, we were exploring if ***is there an indicative gene-expression profile, i.e., a molecular gene signature, that contrast SCOV2 infected from uninfected CTRL cases***? The data was loaded to iDEP using the "Normalized expression values" option, as the available GSE151513 study data were already normalized (with the DSEeq2 'median ratio method'). Following the key-suggestion given in the help-guide of the iDEP server we filtered the data according to a minimum expression value in at least 3 samples (the number of replicates for each time-point)_so that 30% of the low expressed transcripts are discarded. For the current dataset this cutoff value was set equal to 3.15, which left us with a total of 10,793 (from the original 15,761) gene/transcripts. The data was them log2 transformed in order to deal with too large expression values and avoid large kurtosis.



We performed a PCA analysis in order assess the separability between SCOV2 and CTRL samples. As it can be easily observed from the PCA plot (side Figure 9), the samples between the two classes are mixed. In addition, PCA results indicate that both PC1 and PC2 components are strongly correlated with the hpi variable, i.e., with the infection time-points (p=9.01e-20 and p=2.90e-16 for PC1 and PC2, respectively). MDS analysis plot gave similar results.

Figure 9. PCA plots of GSE151513 SCOV2 vs. CTRL data. (18+18 = 36 samples).

[8] Calu-3: a cell-line originated from lung adenocarcinoma human airway epithelial cells utilized for respiratory modeling and studying of the effects of gas concentrations, exposure time, biophysical stress, and biological agents on human airway epithelial cells (Y. Zhu, Chidekel, & Shaffer, 2010); calu-3 are used for studying SCOV2 infection as they are sensitive and express ACE2, the main entry receptor of the virus (Tseng et al., 2005).

[9] MOI: https://www.virology.ws/2011/01/13/multiplicity-of-infection/

The inability to separate SCOV2 from CTRL cases was further confirmed by the **inability to infer any differentiable gene/transcript** for the comparison SCOV2 vs. CTRL, even with relatively high FDR and relatively low fold-change cutoff values (0.1 and 1.5, respectively); the *limma* differential gene-expression (**DEG**) analysis method was utilized (Law, Chen, Shi, & Smyth, 2014; Smyth, 2005). So we may safely state that:

- ❑ Infected SCOV2 cases **could not be contrasted** to uninfected CTRL ones
- ❑ The infection gene expression profiles are strongly correlated with and influenced by the **progress** of the infection (i.e., infection time-points)
- ❑ The molecular background of SCOV2 infection needs to be explored according to the **infection progress stages**.

3.1.2 Defining the SCOV2 progress stages

Normalization. We focus just on the SCOV2 samples in order to search for putative molecular fingerprints that characterize the different infection stages. To this end, and in order to eliminate the 'normal' molecular events taking place in both SCOV2 and CTRL/mock-treated cell-lines, we *normalized* the gene/transcript expression values for each SCOV2 replicate case with reference to the respective CTRL/mock-treated replicate. For this we use following formula:

$$s'_{t,r} = \left(\frac{s_{t,r}}{m_{t,r}}\right) \times 100$$

where, $s'_{t,r}$ is the normalized value of replicate r for an SCOV2 sample s at hpi time-point t; $m_{t,r}$ the (original) value of the respective CTRL/mock-treated sample; to take care of gene expression values equal to zero we add 0.1 to both nominator and denominator. The normalized gene/transcript value does not reflect anymore an absolute gene expression value but the degree it deviates from its respective mock-treated value, with the deviation expressed as a percentage. In other words, we are now **seeking for molecular profiles that deviate (up- or down-regulated) from the 'normal' uninfected case, and contrast them between different SCOV2 infection stages**.

Designating EARLY / LATE infection stages. The normalized data were fed to iDEP using again the "Normalized expression values" option. The samples are labeled as, `SCOV2_<hpi>_<replicate>`, `hpi` = 0, 1, 2, 3, 6, 12, `replicate` = 1, 2, 3. Inspecting the heatmap of samples' expression values (side Figure 10), two contrasting profiles may be identified. The first includes samples for hpi time-points 1, 2 and 3, and one for hpi time-point 12. These samples are assigned to EARLY and LATE infection stages, respectively (a total of 12 samples). As we are interested to <u>intensely contrast</u> between different infection stage profiles the samples for time-points 0 (*the infection is not established*) and 6 (the '*intermediate*' stage) were left out.

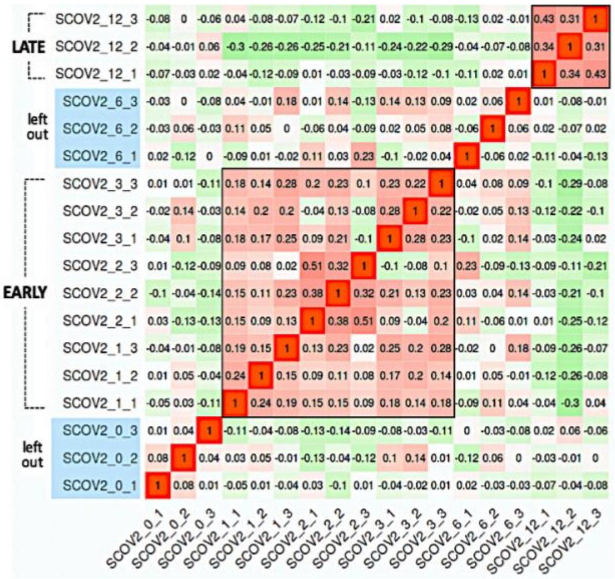


Figure 10. Heatmap of normalized SCOV2 gene-expression values.

Differentiating between EARLY and LATE SCOV2 infection stages. The gene-expression profiles of the retained 12 samples were uploaded to iDEP using again the “Normalized expression values” option. The experimental *design* includes now an extra variable with values EARLY and LATE assigned to the respective early and late staged samples. The data was also log2 transformed in order to deal with too large expression values and avoid large kurtosis. The upper-left part of Figure 11 shows the *correlation matrix*, and the down-left part the *hierarchical clustering tree* of samples, as formed by using the genes with maximum expression level at the top 75%. The hierarchical tree is induced using correlation as distance (i.e., 1 – correlation coefficient) and the average linkage method[10]. Note the separation of samples into EARLY (hpi time-points 1, 2, 3) and LATE (hpi time-point 12) into two groups, i.e., the **two SCOV2 infection stages**. This separation could be further justified by the respective *MDS* (upper-right of the figure) and *PCA* (down-right) plots.

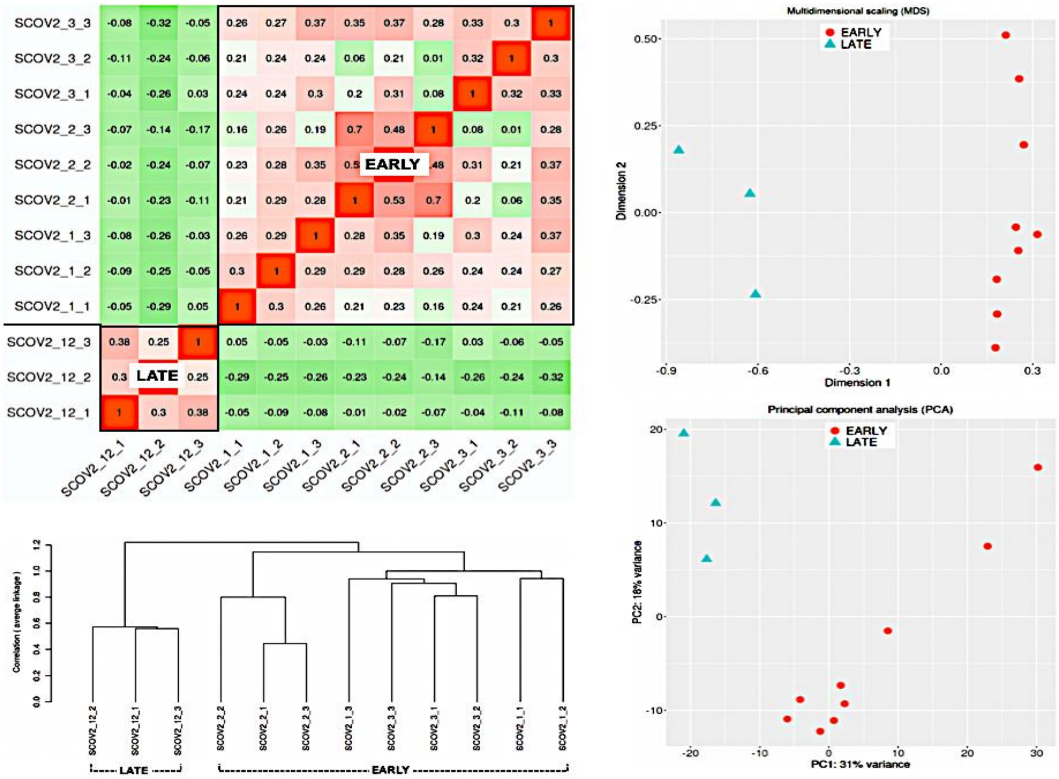


Figure 11. Correlation matrix (upper-left), hierarchical clustering tree (down-left), MDS (upper-right) and PCA (down-right) plots of the 12 SCOV2 samples.

3.1.3 EARLY vs. LATE SCOV2 infection stages: Differential expression and enrichment analysis

EARLY vs. LATE SCOV2 infection stages: Differential Expressed Genes (DEG analysis). The *limma* DEG method, as provided by iDEP for normalized gene-expression data., was applied. Setting FDR cutoff and minimum fold-change equal to 0.05 and 2, respectively, a total of **71** gene/transcripts were found as significantly differentiating between early and late SCOV2 infection stages; **41 Down**- and **30 Up-**regulated. The differentiation contrast is dual, that is: gene/transcripts down-regulated in the EARLY stage are up-regulated in the LATE stage. The respective heatmap of identified differentially expressed genes are shown in Figure 12.

[10] towardsdatascience.com/introduction-hierarchical-clustering-d3066c6b560e

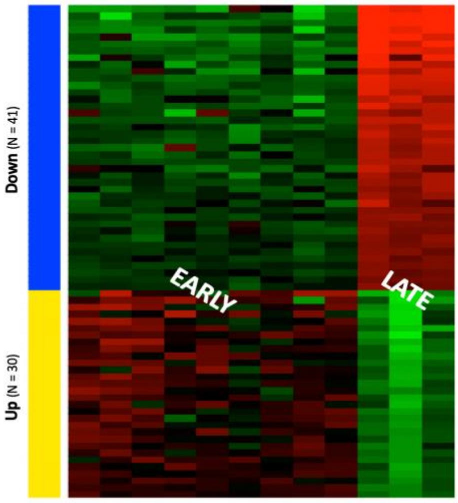
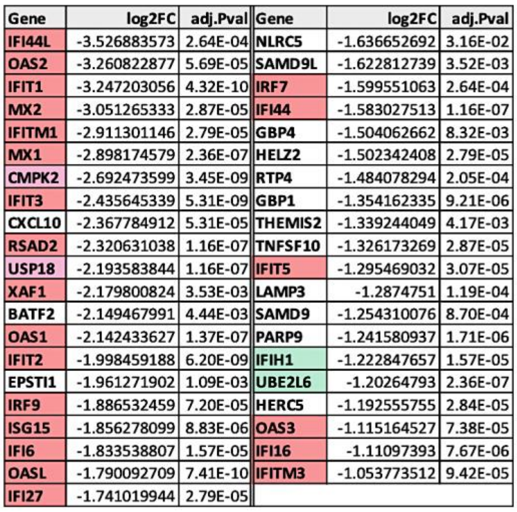


Figure 12. Heatmap (left), of differentially expressed genes – Down- and Up-regulated genes refer to SCOV2 EARLY and LATE infection stages, respectively.

We focus on the identified down-regulated genes in the early SCOV2 infection stage (left part of Figure 13, 41 genes). Note the negative log2FC sign that indicates down-regulation. In an effort to get insight into the interactions between these genes and uncover their functional roles we input them into the **STRING** database. STRING database collects, scores and integrates a spectrum of publicly available sources (e.g., ENSEMBL, GeneCards, KEGG, NextProt, RefSeq and UniProt) of physical and functional protein-protein interaction information. It complement these interactions with computational predictions, offering operations that ease the formation of **comprehensive protein networks** (Szklarczyk et al., 2019). In order to explore the main molecular functions of the identified down-regulated genes we conducted a **clustering** analysis using the *MCL* algorithm[11] (as provided by STRING). From the 41 down-regulated genes, 24 were **coherently clustered** and grouped into two main clusters (right part of **Error! Reference source not found.**). The coloring of genes (light-red and light-green) is made according to their cluster/group inclusion.



Gene	log2FC	adj.Pval	Gene	log2FC	adj.Pval
IFI44L	-3.526883573	2.64E-04	NLRC5	-1.636652692	3.16E-02
OAS2	-3.260822877	5.69E-05	SAMD9L	-1.622812739	3.52E-03
IFIT1	-3.247203056	4.32E-10	IRF7	-1.599551063	2.64E-04
MX2	-3.051265333	2.87E-05	IFI44	-1.583027513	1.16E-07
IFITM1	-2.911301146	2.79E-05	GBP4	-1.504062662	8.32E-03
MX1	-2.898174579	2.36E-07	HELZ2	-1.502342408	2.79E-05
CMPK2	-2.692473599	3.45E-09	RTP4	-1.484078294	2.05E-04
IFIT3	-2.435645339	5.31E-09	GBP1	-1.354162335	9.21E-06
CXCL10	-2.367784912	5.31E-05	THEMIS2	-1.339244049	4.17E-03
RSAD2	-2.320631038	1.16E-07	TNFSF10	-1.326173269	2.87E-05
USP18	-2.193583844	1.16E-07	IFIT5	-1.295469032	3.07E-05
XAF1	-2.179800824	3.53E-03	LAMP3	-1.2874751	1.19E-04
BATF2	-2.149467991	4.44E-03	SAMD9	-1.254310076	8.70E-04
OAS1	-2.142433627	1.37E-07	PARP9	-1.241580937	1.71E-06
IFIT2	-1.998459188	6.20E-09	IFIH1	-1.222847657	1.57E-05
EPSTI1	-1.961271902	1.09E-03	UBE2L6	-1.20264793	2.36E-07
IRF9	-1.886532459	7.20E-05	HERC5	-1.192555755	2.84E-05
ISG15	-1.856278099	8.83E-06	OAS3	-1.115164527	7.38E-05
IFI6	-1.833538807	1.57E-05	IFI16	-1.11097393	7.67E-06
OASL	-1.790092709	7.41E-10	IFITM3	-1.053773512	9.42E-05
IFI27	-1.741019944	2.79E-05			

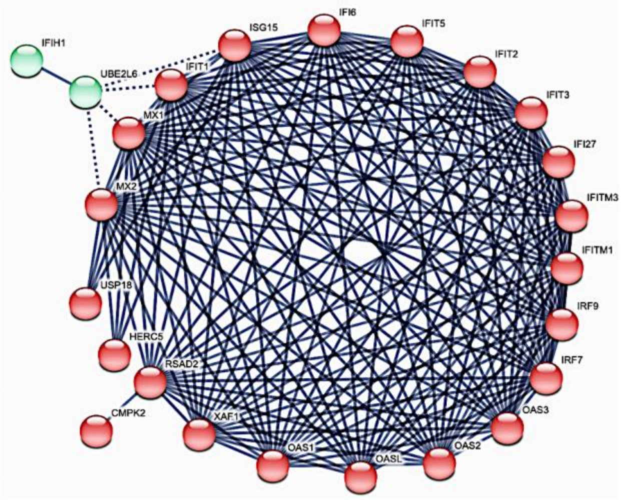


Figure 13. Down-regulated differential genes in the EARLY SCOV2 stage (left); (Right) Clustered network organization of highly confident correlated genes inferred with MCL clustering (right); minimum correlation/edge confidence score was set to 0.7

[11] MCL Markov Cluster algorithm is a fast and scalable unsupervised graph clustering algorithm (Enright, Van Dongen, & Ouzounis, 2002) which, is effectively used in protein association network analysis (Enright et al., 2002).

Table 3 summarizes the functional annotation of these down-regulated genes, with a **focus on their involvement in (anti-)viral activities and molecular regulations**.

Table 3. Functional annotation of down-regulated differential genes at the early SCOV2 infection stage

Interferon Stimulated Genes (ISGs)	Functional Annotation / Antiviral Activity
IFI16, IFI27, IFI44, IFI44L, IFI6, IFIH1, IFIT, IFIT2, IFIT3 IFIT5, IFITM1, IFITM3	– Inhibit the entry of viruses to the host cell cytoplasm, permitting endocytosis, but preventing subsequent viral fusion and release of viral contents into the cytosol – Active against multiple viruses (e.g., influenza A virus, SARS coronavirus/SARS-CoV, Marburg virus/MARV, Ebola virus/EBOV, Dengue virus/DNV, West Nile virus/WNV, human immunodeficiency virus type 1/HIV-1, hepatitis C virus/HCV)
IRF7, IRF9	– IRF7: plays a critical role in the innate immune response against DNA and RNA viruses – IRF9: Associate with the phosphorylated STAT1:STAT2 dimer to form a complex termed ISGF3 transcription factor that enters the nucleus and binds to the IFN stimulated response element (ISRE) to activate the transcription of interferon stimulated genes, which drive the cell in an antiviral state
MX1, MX2	– Sensors of viral single-stranded (ss)RNAs; Inhibit expression of viral mRNAs – Provide a molecular signature to distinguish between self and non-self mRNAs by the host during viral infection
OAS1, OAS2, OAS3, OASL	– Antiviral enzymes – Critical role in cellular innate antiviral response – Activation of **RNaseL** leading to degradation of viral RNA – OASL: antiviral activity against encephalomyocarditis virus (EMCV) and hepatitis C virus (HCV) via an alternative antiviral pathway independent of RNaseL
RSAD2 (Viperin)	– Plays major role in the cell antiviral state induced by IFN-I, II – Inhibit wide range of DNA/RNA viruses (cytomegalovirus/ HCMV, hepatitis C virus/HCV, west Nile virus/WNV, dengue virus, sindbis virus, influenza A virus, human immunodeficiency virus/HIV-1 and other)
CMPK2	– Adjacent to / Co-expressed with RSAD2 – Restrict human immunodeficiency virus (HIV) infection
ISG15	– Key role in the innate immune response to viral infection
UBEL26, USP18	– Critical role in ISG15 regulation
XAF1	– Apoptosis-related antiviral activity

Inspecting Table 3 it becomes evident that **most of the down-regulated genes are directly related to interferon regulation, especially to interferon stimulated** genes (light-red colored genes). The down-regulation profile of these genes confirms the hypothesis and supports the discussion made in previous sections regarding **impaired immune responses during early SCOV2 infection stage**. In particular, **all the INF-stimulated genes that interfere in the viral life-cycle genes are found as down-regulated**. Figure 14 highlights (in blue rectangles) these down-regulated ISGs in the course of the viral life-cycle (recall also Figure 7 at section 1.3.3).

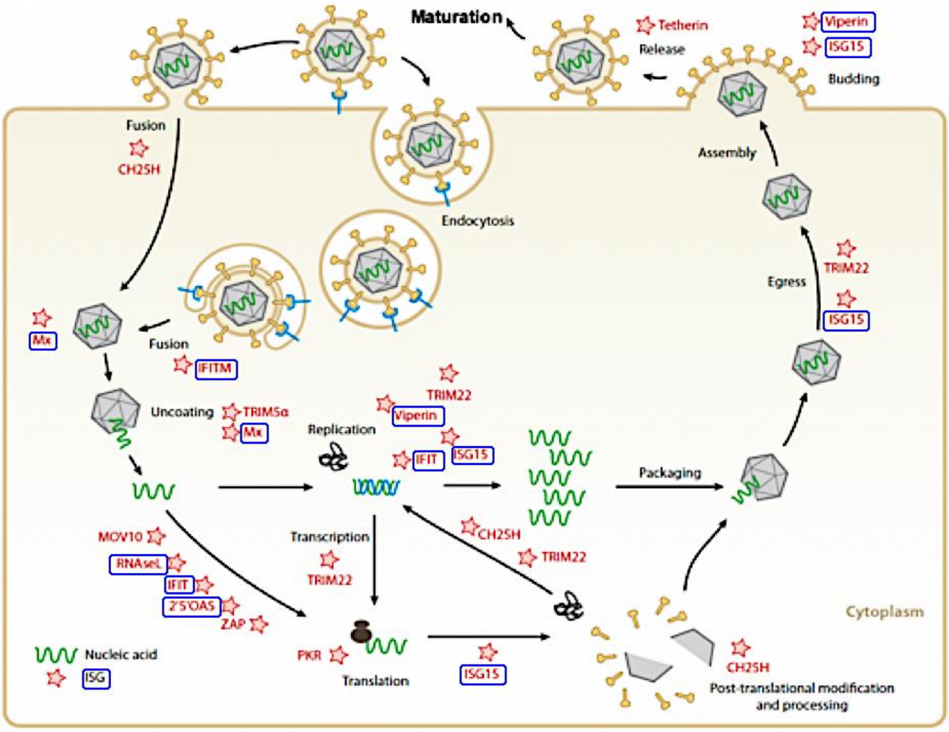


Figure 14. ISGs (interferon stimulated genes), highlighted with in blue rectangles, involved in the viral life-cycle and found down-regulated in the EARLY stage of SCOV2 infection.

The findings can be further justified by inspecting Figure 15, where the bar-plots for the expression levels of the key down-regulated genes in the EARLY SCOV2 infection stage are plotted. Note that for all these genes their **expression level is nearly reaching the mean/median expression level of all input gene/transcripts; the inverse holds for the LATE SCOV2 infection stage**.

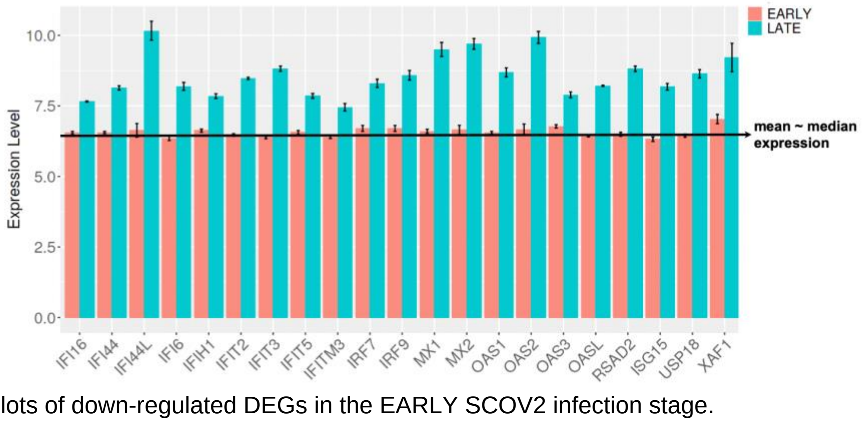


Figure 15. Bar-plots of down-regulated DEGs in the EARLY SCOV2 infection stage.

EARLY vs. LATE SCOV2 infection stages: Enrichment/Pathway analysis. In order to gain insight into the molecular events that take place during the progress of SCOV2 infection we proceed to the identification of the **enriched biological processes and pathways that contrast between early and late SCOV2 infection stages**. The focus is on the processes and pathways which are down-regulated and in a way are 'blocked' at the EARLY stage of the infection. Based on the identified DEGs, and utilizing the *fgsea* Bioconductor package for enrichment analysis (Korotkevich et al., 2021), a series of **GeneOntology**/GO-biological processes as well as a series of **Reactome** and **KEGG** pathways were found as significantly enriched and down-regulated at the early ECOV2 stage. Table 4 summarizes the findings; biological processes and pathways are sorted according to their adjusted p-value.

Table 4. Enriched down-regulated biological processes and pathways in the EARLY SCOV2 stage.

	GO-Biological Processes	adj.Pval	#genes
DOWN regulated at the **EARLY** SCOV2 stage	Defense response to virus	9.60E-37	27
	Type I interferon signaling pathway	7.60E-36	21
	Cellular response to type I interferon	7.60E-36	21
	Response to virus	2.00E-35	28
	Response to type I interferon	3.60E-35	21
	Defense response to other organism	1.40E-29	31
	Response to external biotic stimulus	9.80E-29	33
	Innate immune response	5.10E-27	28
	Defense response	3.20E-26	32
	Immune response	5.70E-24	33
	Response to cytokine	2.80E-22	28
	Reactome Pathways	**adj.Pval**	**#genes**
DOWN regulated at the **EARLY** SCOV2 stage	Interferon alpha/beta signaling	8.80E-35	19
	Interferon Signaling	2.80E-32	23
	Cytokine Signaling in Immune system	1.30E-20	24
	Immune System	2.80E-15	27
	Antiviral mechanism by IFN-stimulated genes	1.70E-14	11
	Interferon gamma signaling	1.10E-10	8
	ISG15 antiviral mechanism	2.50E-08	7
	OAS antiviral response	3.80E-08	4
	DDX58/IFIH1-mediated induction of interferon-alpha/beta	1.60E-07	6
	Negative regulators of DDX58/IFIH1 signaling	2.40E-07	5
	TRAF3-dependent IRF activation pathway	2.00E-03	2
	KEGG Pathways	**adj.Pval**	**#genes**
DOWN regulated at the **EARLY** SCOV2 stage	Influenza A	1.90E-09	9
	Hepatitis C	3.80E-08	8
	Coronavirus disease	2.40E-07	8
	Measles	2.20E-06	6
	NOD-like receptor signaling pathway	4.80E-06	6
	Epstein-Barr virus infection	2.50E-04	5
	RIG-I-like receptor signaling pathway	7.50E-04	3
	Viral protein interaction with cytokine and cytokine receptor	4.80E-03	2

Furthermore, and in order to reveal the interplay between the identified enriched down-regulated GO-biological processes, Figure 16 illustrates their network organization.

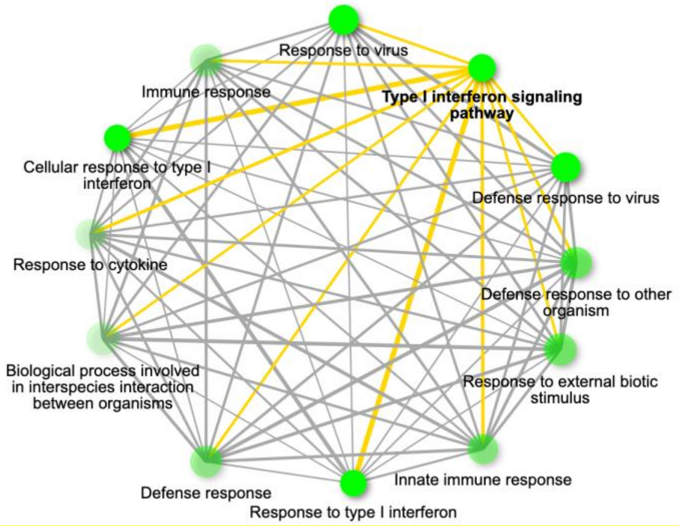

Response to virus
Immune response
Type I interferon signaling pathway
Cellular response to type I interferon
Defense response to virus
Response to cytokine
Defense response to other organism
Biological process involved in interspecies interaction between organisms
Response to external biotic stimulus
Innate immune response
Defense response
Response to type I interferon

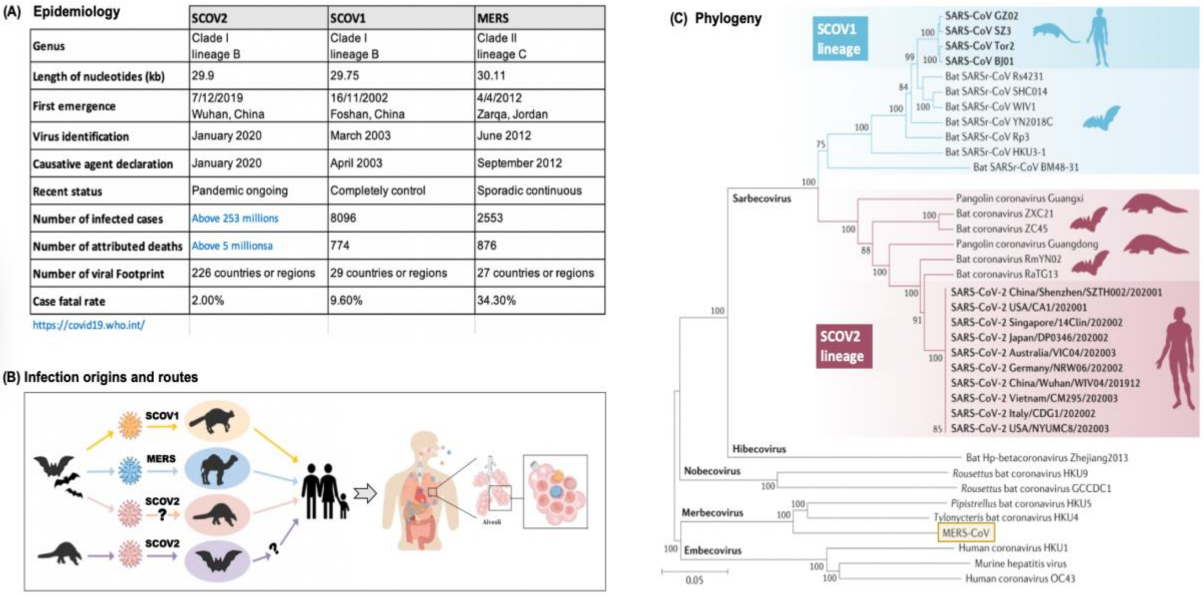
Figure 16. Network organization of enriched down-regulated biological processes (edge thickness indicates the number of shared genes among the connected processes; a filter of 30% was set.

- Not unexpectedly, based on the discussion made in section **Error! Reference source not found.** and previous sub-sections, the results signify the fact that at the early SCOV2 infection stage **key biological processes and pathways engaged in first-line innate immune response, such as IFN and cytokine signaling, are down-regulated**. **The key gene/transcripts engaged in these processes/pathways, e.g., ISGs (interferon stimulated genes), are also down-regulated.**
- Down-regulation of key immune processes at the early course of SCOV2 infection may present the **background for the later occurrence of acute inflammatory responses and severe disease outcomes**.

3.2 SARS-CoV-1 vs. SARS-CoV-2: Do they differ?

3.2.1 Background to SARS-CoVs infections

1st, 2nd, 3rd … SARS-CoVs: **Epidemiology, Origins and Phylogeny.** Despite the fact that SCOV2 became a pandemic, it is actually the third serious outbreak caused by Coronaviruses in the last 20 years. The '*first SARS*' (SCOV1) outbreak happened at Hong Gong back to 2002–2003 (Peiris et al., 2003), and MERS (maybe the *'2nd SARS'* !) at Saudi Arabia and Jordan in 2012 (Assiri et al., 2013; Memish, Perlman, Van Kerkhove, & Zumla, 2020). Despite the still on-going research and argumentation, the **zoonotic origin** of all three infections presents the most justified theory so-far with two scenarios proposed for their evolution and transfer to humans: (a) natural selection in an animal host before zoonotic transfer; and (b) natural selection in humans following zoonotic transfer (Andersen, Rambaut, Lipkin, Holmes, & Garry, 2020). In any case, various studies evidence a **strong similarity between the three infections in terms of their clinical manifestations** (Bi et al., 2020; Z. Zhu et al., 2020). In addition, it is well established that SCOV2 shares 79% of its genome with SCOV1. **This allows comparative analyses between SCOV1 and SCOV2 towards the discovery of commonalities in their molecular fingerprints**. The background and context of the three infections is shown in Figure 17; notice the close phylogeny of SCOV1 and SCOV2 (right-part C of the figure).



(A) Epidemiology

	SCOV2	SCOV1	MERS
Genus	Clade I lineage B	Clade I lineage B	Clade II lineage C
Length of nucleotides (kb)	29.9	29.75	30.11
First emergence	7/12/2019 Wuhan, China	16/11/2002 Foshan, China	4/4/2012 Zarqa, Jordan
Virus identification	January 2020	March 2003	June 2012
Causative agent declaration	January 2020	April 2003	September 2012
Recent status	Pandemic ongoing	Completely control	Sporadic continuous
Number of infected cases	Above 253 millions	8096	2553
Number of attributed deaths	Above 5 millionsa	774	876
Number of viral Footprint	226 countries or regions	29 countries or regions	27 countries or regions
Case fatal rate	2.00%	9.60%	34.30%

https://covid19.who.int/

(B) Infection origins and routes

(C) Phylogeny

SCOV1
MERS
SCOV2
?
SCOV2
Alveoli
SCOV1 lineage
SCOV2 lineage
SARS-CoV GZ02
SARS-CoV SZ3
SARS-CoV Tor2
SARS-CoV BJ01
Bat SARSr-CoV Rs4231
Bat SARSr-CoV SHC014
Bat SARSr-CoV WIV1
Bat SARSr-CoV YN2018C
Bat SARSr-CoV Rp3
Bat SARSr-CoV HKU3-1
Bat SARSr-CoV BM48-31
Sarbecovirus
Pangolin coronavirus Guangxi
Bat coronavirus ZXC21
Bat coronavirus ZC45
Pangolin coronavirus Guangdong
Bat coronavirus RmYN02
Bat coronavirus RaTG13
SARS-CoV-2 China/Shenzhen/SZTH002/202001
SARS-CoV-2 USA/CA1/202001
SARS-CoV-2 Singapore/14Clin/202002
SARS-CoV-2 Japan/DP0346/202002
SARS-CoV-2 Australia/VIC04/202003
SARS-CoV-2 Germany/NRW06/202002
SARS-CoV-2 China/Wuhan/WIV04/201912
SARS-CoV-2 Vietnam/CM295/202003
SARS-CoV-2 Italy/CDG1/202002
SARS-CoV-2 USA/NYUMC8/202003
Hibecovirus
Bat Hp-betacoronavirus Zhejiang2013
Nobecovirus
Rousettus bat coronavirus HKU9
Rousettus bat coronavirus GCCDC1
Merbecovirus
Pipistrellus bat coronavirus HKU5
Tylonycteris bat coronavirus HKU4
MERS-CoV
Embecovirus
Human coronavirus HKU1
Murine hepatitis virus
Human coronavirus OC43
100
99
84
75
88
91
85
0.05

Figure 17. The background and context of the three SARS-CoVs (SCOV1, SCOV2 and MERS): **(A)** Epidemiology, **(B)** Infection origins and routes (adapted from (Z. Zhu et al., 2020)), **(C)** Phylogeny of the three SARS-CoVs and other CoVs (from (Hu, Guo, Zhou, & Shi, 2021).

Delayed immune responses: the common SCOV1/2 fingerprint. There is supporting evidence that SCOV2 infection portrayed strong suppression of innate immune response and inhibition of IFN-I responses (Vabret et al., 2020), and our results and findings in the previous section justifies this. Furthermore, studies with animal models show that **SCOV1 and MERS infections also cause failures of IFN-I responses** and link this dysregulation with severe disease outcomes. In particular, it is evidenced that **timing in the induction of IFNs is the key for the pathogenicity of both SCOV1 and SCOV2 profiles, with early induction to be protective and later to cause pathologic situations** (Channappanavar et al., 2016, 2019). Moreover, recent *in vitro* studies provide evidence that SCOV2 is sensitive to IFN-I pretreatment, even to a higher level than SCOV1 (Lokugamage et al., 2020; Mantlo, Bukreyeva, Maruyama, Paessler, & Huang, 2020). It is also likely that IFN-induced (transmembrane) IFITM proteins family prevent SCOV2 cell entrance, a fact already demonstrated for SCOV1 (Huang et al., 2011), but not clearly evidenced for other CoVs (Zhao et al., 2014, 2018). **ACE2**, the key receptor that mediates SCOVs cell entrance, is also implicated with this phenomenon. In a recent study it is demonstrated that a **putative cause for the inhibition of IFN-I in the early stage of infection is linked with the role of ACE2 as an IFN-I induced/stimulated gene** (Ziegler et al., 2020).

The aforementioned observations validate and necessitate studies that contrast between SCOV1 and SCOV2 infections in an effort to discover and reveal a putative common molecular background for the two infections.

3.2.2 The SCOV1 two-stage profile

Dataset. To tackle this task we utilized the **GSE33267** GEO dataset. In this study, Calu-3 cells were infected with SCOV1 (33 samples) or kept uninfected (mock/CTRL, 33 samples), and their **microarray** gene-expression profiles were acquired via the 'Agilent-014850 Whole Human Genome Microarray 4x44K G4112F' platform (GPL4133). Cell samples were collected at eleven different post-infection time-points/hours (hpi), namely 0, 3, 7, 12, 24, 30, 36, 48, 54, 60 and 72 in triplicates (i.e., a total of 3 x 11 = 33 for each class), with a *multiplicity of infection* (MOI) of 5. The results of this study are published in (Sims et al., 2013).

Filtering and contrasting the time-course of SCOV1 infection. Following an analogous to the previous experiment filtering process (section 3.1) we managed to identify critical hpi time-points that could be used as a reference to contrast and differentiate between EARLY and LATE SCOV1 infection stages. To do so, we utilized the correlation matrix of samples in order to identify clearly contrasting samples' gene-expression profiles. In particular, hpi time-points 0,3,7 and 12 compose the EARLY, and 48,54,60 and 72 the LATE infection stages, respectively for the retained samples (see Figure 18).

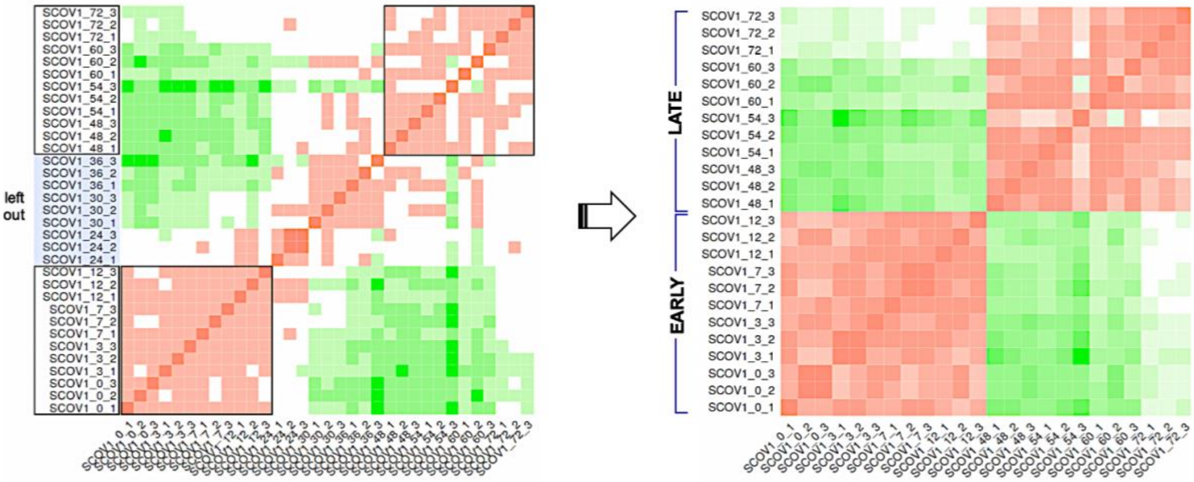


Figure 18. Filtering-out samples for hpi time-points 24,30 and 36 (left); The correlation matrix of retained samples that strongly contrast EARLY vs. LATE SCOV1 infection stages (right).

Figure 19 shows the PCA plot (left) and the hierarchical clustering tree of the retained for further analysis samples (hierarchical tree is induced as in the previous experiment, refer to Figure 11). The clear-cut contrasting profiles between EARLY and LATE SCOV1 infection stages is witnessed.

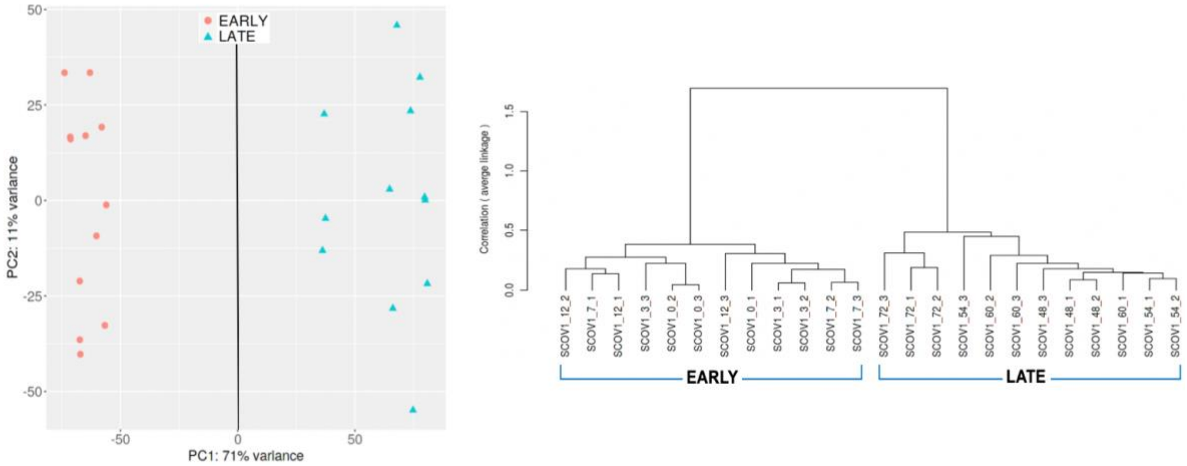


Figure 19. PCA plot (left) and hierarchical clustering of EARLY (defined for samples at hpi time-points 0, 3, 7 and 12 and LATE (defined for hpi time-points 48, 54, 60 and 72) SCOV1 infection stages (right).

Differential expressed genes between early and late SCOV1 stages. Analogous to the previous experiment we filtered-out gene probes with minimum expression values lower than a specified threshold in an adequate number of samples. After inspection of the data the cutoff value was set to 7 in at-least 3 (the number of replicates) samples, leaving 22,286 gene probes for further analysis (the median and mean of the transformed gene expression values of these genes is near 10). The *limma* DEG method was then applied with FDR cutoff equal to 0.05, and minimum fold-change equal to 5, as we are looking for strongly differentiating genes (smaller values gave too many DEGs). A total of **573** genes were found as significantly differentiating between early and late SCOV1 infection stages; **494 Down**- and **79 Up-**regulated. Again, we note that down-regulated genes refer to SCOV1 EARLY infection stage and up-regulated to the LATE stage (i.e., the duality of contrast between the two classes).

Contrasting SCOV1 / SCOV2 early infection stages. We focus on the common, if any, down-regulated DEGs between the two infections. From the top DEGs of the two infections there are 13 genes in common namely: IFI44L, IFI6, MX1, MX2, OAS1, OAS2, RSAD2, **all of which related to IFN signaling and belong to the ISGs family**; and MT1E, MT1G, MT1H, MT1X and MT2A all of which are metallothioneins (MTs); refer to the end of this section for a special discussion on metallothioneins and their role in SCOV2 and other viral infections. The down-regulated XAF1 gene is also shared between the two infections.

So, **all the differentially down-regulated expressed in SCOV2 early infection stage are also down-regulated in SCOV1 early infection stage, and all of them are ISGs that relate to IFN and cytokine signaling. In other words, both infection types share a common molecular background during their early progress stage.**

To further justify the findings, and following the same enrichment methodology as in the previous experiment, we were able to identify significantly enriched biological processes and pathways for the identified down-regulated DEGs (see Table 5). Comparing with the enriched processes/pathways during the early stage of SCOV2 infection (please refer to Table 4) it can be easily observed that **both viral infections share most of their down-regulated molecular processes during early infection stages, and all of them relate to IFN and cytokine signaling**. For the Reactome pathways the observation is more profound.

Table 5. Enriched down-regulated GO-biological processes and Reactome pathways in EARLY vs. LATE SCOV1 infection stages.

	GO-Biological Processes	**adj.Pval**	**#Genes**
DOWN regulated at the **EARLY** SCOV1 stage	Response to cytokine	5.80E-28	79
	Response to biotic stimulus	7.90E-28	84
	Defense response	7.90E-27	88
	Response to external biotic stimulus	7.90E-27	81
	Response to external stimulus	2.40E-25	112
	Response to virus	9.40E-25	42
	Defense response to virus	1.90E-24	37
	Cytokine-mediated signaling pathway	2.60E-24	60
	Cellular response to cytokine stimulus	3.20E-24	70
	Defense response to other organism	4.40E-23	66
	Immune system process	1.40E-22	113
	Innate immune response	1.40E-21	58
	Immune response	4.00E-21	89
	Reactome pathways	**Adj.Pv**	**#Genes**
	Cytokine Signaling in Immune system	4.10E-22	58
	Interferon Signaling	3.70E-19	28
	Interferon alpha/beta signaling	4.70E-18	18
	Immune System	5.80E-12	71
	Antiviral mechanism by IFN-stimulated genes	1.20E-09	14
	Interferon gamma signaling	2.60E-06	10
	OAS antiviral response	2.60E-06	5
	ISG15 antiviral mechanism	5.60E-06	10
	Signaling by Interleukins	8.90E-06	27
	DDX58/IFIH1-mediated induction of interferon-	2.30E-05	9
	Interleukin-10 signaling	5.60E-05	7
	Metallothioneins bind metals	5.60E-05	4
	Negative regulators of DDX58/IFIH1 signaling	2.30E-04	6
	Chemokine receptors bind chemokines	5.80E-04	6

Pathway analysis. Using the pathway analysis module of iDEP, and utilizing the PAGE/**PGSEA** enrichment analysis method (S.-Y. Kim & Volsky, 2005), we managed to identify a number of significantly down-regulated KEGG pathways during the early SCOV1 infection stage. Figure 20 (left) shows these pathways alongside their down/up regulated profiles in the respective samples; note the down-regulation of all these pathways in the early stage of SCOV1 infection (hpi time-points 0, 3, 7 and 12) in contrast to their up-regulation status in the late infection stage (48, 54, 60 and 72).

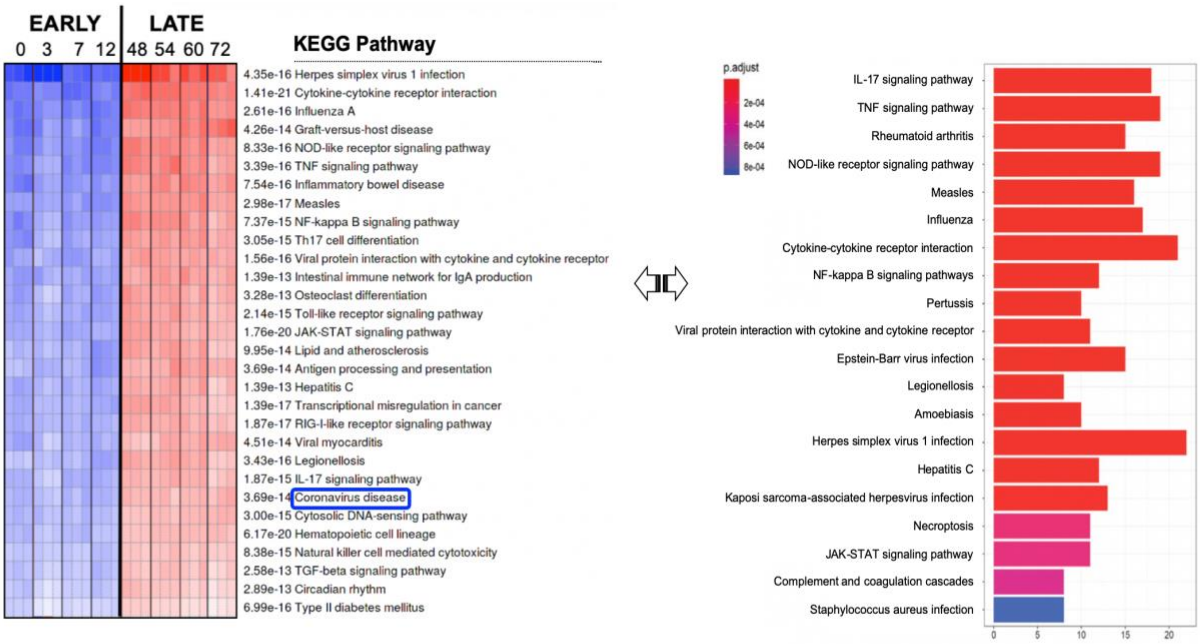


Figure 20. (Left): Differential down-regulated KEGG pathways during SCOV1 early infection stage; (Right): differential KEGG pathways reported in (H. Zhang et al., 2021).

Of considerable interest is the **down-regulation of 'Coronavirus disease' pathway**, as well as other, disease or biological molecular pathways which, relate to immune responses, e.g., 'Cytokine-cytokine receptor interaction', NOD-like receptor signaling, 'TNF signaling', 'NF-kappa B signaling', 'Viral protein interaction with cytokine and cytokine receptor', 'Toll-line receptor signaling', 'JAK-STAT signaling', 'Antigen processing and presentation', 'RIG-I-like receptor signaling', IL-17 signaling', etc. To further validate our findings, the right part of Figure 20 shows the differential KEGG pathways reported in (H. Zhang et al., 2021) for the comparison between SCOV2- vs. Mock-infected (i.e., CTRL) cell lines. Most of KEGG pathways are shared among our and reference study's results.

The findings add additional support to the postulate**:**

SCOV1 and SCOV2 infections exhibit similar molecular profiles during the early stage of infection

A special note on metallothioneins. Besides ISGs, another group of genes are also identified as down-regulated. This group is composed by genes solely belonging to the **Matallothioneins (MT**), namely: *MT1E*, *MT1F*, *MT1G*, MT1H, *MT1X*, *MT2A*. MTs are a family of small, highly conserved, cysteine-rich metal-binding proteins[12]. It regulates **zinc (Zn)** (Ruttkay-Nedecky et al., 2013) which exhibits a beneficial role in various physiological and molecular host defense mechanisms during various pathogen infections including SCOV2. (i) **Anosmia/Taste.** It is known that Zn deficiencies relate directly to anosmia and taste dysfunctions (ageusia), an established and common symptomatology in SCOV2 infected cases. (Propper, 2021)**,** especially when decreased levels occurs in the nasopharyngeal tract (Equils, Lekaj, Fattani, Wu, & Liu, 2020). There is evidence that acute viral infection of the nasopharyngeal mucosa lead to a decrease in local Zn levels as part of normal defense against respiratory pathogens (Wessels, Maywald, & Rink, 2017). (ii) **Host defense and molecular machinery during infection.** It is known than Zn contributes to host defense

[12] www.sciencedirect.com/topics/neuroscience/metallothionein

responses by maintaining the membrane barrier structure and function (Finamore, Massimi, Conti Devirgiliis, & Mengheri, 2008) via the modulation of cytokine-induced epithelial cell barrier absorptiveness (Bao & Knoell, 2006). In addition, there is evident that Zn helps to enhance IFN-I response during SCOV2 infection, it shows an inhibition capacity of SCOV2 RNA polymerase, and its deficiency is associated with severe infection (Mayor-Ibarguren, Busca-Arenzana, & Robles-Marhuenda, 2020). (iii) **MTs and Zn.** In-vitro experiments on mice revealed direct and strong increase in the mRNA levels of MTs during acute influenza (A) infection of the upper respiratory tract (Ghoshal et al., 2001). The physiology underlying this increase is attributed to the beneficial antioxidant role, of MTs as they are triggered in order to effectively 'clean-up' the reactive oxygen species (ROS) generated by the host defense phagocytes during infection. A diverse of molecular signaling mechanisms are involved in the induction of MT in response to virus infection, including cytokines, glucocorticoids, and zinc. At the molecular level, recent studies demonstrate that Zn is required for interferon-, especially IFNL-mediated expression of MTs (Read et al., 2017).

3.3 Does and how SARS-CoV-2 differs from Influenza infection?

3.3.1 SARS-CoV-2 vs. Influenza: Background

Influenza (**INFL**) is the most common and a long-standing viral infection worldwide with a well-established epidemiological profile and an extensive scientific literature devoted to its physiological and molecular background[13]. As the knowledge about the SCOV2 infection and its disease manifestations in humans still accumulates, **highlighting the similarities and differences between SCOV2 and INFL is a rational approach to follow towards uncovering and understanding the key physiological and molecular mechanisms guiding and governing the two viral infections**.

INFL vs. SCOV2: Epidemics and Phenotypes. As a starting-point, Table 6 summarizes and contrast between the basic epidemic figures that contrast between the two viral infections (Table 4a), as well as the pathogenesis profiles and phenotypes that characterize them (Table 4b); adapted from (Flerlage, Boyd, Meliopoulos, Thomas, & Schultz-Cherry, 2021).

Table 6. SCOV2 vs. INFL: Epidemics, Pathogenesis and Phenotypes

(a) Epidemics and risk-factors	**SCOV2**	**INFL**
Yearly infections	~ 200 million	3-5 million
Yearly deaths	> 3 million	190.000 - 650.000
Main risk-factors for severe cases	○ Male sex ○ Obesity ○ Genetics (blood type; genes relate to interferon I, III) ○ Comorbidities [diabetes; chronic kidney disease; heart disease; hypertension]	○ Smoking ○ Obesity ○ Genetics (genes related to viral recognition and interferon signaling) ○ Comorbidities (heart diseases; COPD) ○ Age (bias towards type-2 immunity (e.g., Th2 vs. Th1 pathway activation); lack of prior immunity) ○ Pregnancy / Sex (tolerated immunological state; sex steroids influence on immune response)

(b) Pathogenesis and Phenotypes

[13] 'Influenza', *Nature Outlook*, www.nature.com/collections/jicdgbcgda

Receptor usage	ACE2	Sialic acid
Cellular tropism	Respiratory epithelial cells: type II alveolar epithelial cells, ciliated cells and secretory cells; sustentacular and horizontal basal cells of the olfactory epithelium Intestinal epithelial cells; endothelial cells; renal parenchymal cells	Respiratory epithelial cells: types I and II alveolar epithelial cells; ciliated cells
Tissues affected and pathology	Upper respiratory tract; lower respiratory tract; intestinal tract; cardiovascular or endothelial system; kidneys; nervous system	Upper respiratory tract; lower respiratory tract (severe cases)
Site of viral replication	Cytoplasmic	Nuclear
Extrapulmonary complications	Extensive; olfactory: anosmia; endothelial: thrombosis; neurological: stroke, encephalitis, neuropsychiatric; gastrointestinal: nausea, vomiting, diarrhea	Limited; cardiac: myocarditis (rare); neurological: encephalitis (rare)
Prior immunity	No specific SARS-CoV-2 immunity prior to late 2019–2020; protective immunity from other human coronaviruses unclear; vaccination started December 2020	Previous infection; vaccination (w. subtype specificity)

INFL vs. SCOV2: Infection phenotypes and transmissibility:

- **INFL.** Various studies have demonstrated that in most of the cases IFLN virus is detected on the first day after infection with viral titers coming at a peak in the second to third day after and falling to undetectable levels in the next six to seven days. Symptoms occur already from the first day after infection, coming to a peak on the second to third day, and diminish after five to six days (Carrat et al., 2008). Severe disease states are more frequent to individuals at high-risk (i.e., older people with comorbidities or even younger individuals not exposed to various INFL infections) and include, hospitalization, pneumonia, acute respiratory distress syndrome (ARDS), even death.
- **SCOV2.** Symptoms start after an incubation period of about five days (after exposure to the virus) with the majority of cases to display symptoms for about two weeks after (Bi et al., 2020). From estimations, highest transmissibility occurs for a period of about four days –two days before and one day after symptoms occur; mainly from pre-symptomatic individuals (He et al., 2020). In a recent meta-study, comprising 35 studies and a total of 3,385 participants (Yan et al., 2021), about SCOV2 symptomatology, a crucial factor that greatly differentiated SCOV2 from INFL infections, the following interesting figures are reported: the mean viral shedding time (VST) pooled mean is estimated to about 17 days (~17–20d), being significantly longer in symptomatic cases (19.7d) than in asymptomatic ones (10.9). It was also significantly longer in adults (23.2d) compared to children (9.9d). Furthermore, it was significantly longer for individuals with chronic diseases (24.2d) than in those without chronic diseases (11.5d). The aforementioned observations rationalize and evidence the postulate that the **longer incubation and manifestation periods of SCOV2 as well as the longer shedding rates result into more pre- or asymptomatic cases into the population, making SCOV2 enough more transmissible than INFL**.

The aforementioned observations guide us to the formation and adoption of a rational hypothesis that designates the targets of our exploration:

- **Uncovering the key molecular and regulatory mechanisms of SCOV2 that, in contrast to INFL infection, drive to delayed and uncontrolled immune responses at early stages of infection is of fundamental importance for a deeper understanding of SCOV2 progress.**
- Moreover, the similar molecular profiles of SCOV2 and SCOV1 during the early infection stages, as showcased and highlighted in the previous sections, **allows us to contrast SCOV1 with INFL infection and make inferences that also holds for the SCOV2** case, at least for the basic molecular processes.

3.3.2 Early vs. Late immune responses in SCOV1 and INFL

Dataset. We utilized again a relevant dataset from GEO under the code **GSE47960**. In this study HAE (human airway epithelial) cells[14] were infected with **SCOV1** (and other SCOV1 strains) and compared to A/CA/04/2009 **H1N1** influenza-infected cultures based on their gene-expression profiles. The microarray "Agilent-014850 Whole Human Genome Microarray 4x44K G4112F (Probe Name version)" platform (GPL6480) is used, with a total of **32067 gene/probes**. Cell samples were collected at various hours of post-infection (hpi): 0, 12, 24, 36, 48, 60, 72, 84 hpi and 96 for SCOV1, and 0, 6, 12, 18, 24, 36 and 48 hpi for H1N1 (in triplicates or quadruplicates), at a multiplicity of infection (MOI) of 2. The results of this study are published in (Mitchell et al., 2013).

Contrasting the SCOV1 and H1N1 infection time-course. In order to identify gene-expression profiles that contrast between (sequential) hpi time-points we followed the same methodology as in the previous experiments. The original data, separately for H1N1 and SCOV1, were analyzed with the iDEP platform using the normalized gene-expression option. Filtering, using a cutoff value 9 for at-least 3 samples, left us with 15,262 gene/probes, from the original 32,067, for further analysis. The correlation matrices for both H1N1 and SCOV1 were produced and visualized. With a careful inspection of these matrices we were able to identify contrasted hpi time-points that designate the early_to_medium and late stages for both infections. Figure 21 shows the respective correlation matrices, and the contrasted sample profiles for **EARLY_MID** and **LATE** infection stages are indicated by surrounding boxes. At the lower part of Figure 21 the respective PCA plots are shown from which it can easily observed the separated profiles between the two stages for both infections. The time-points 6, 12 designate the EARLY_MID H1N1 infection stage, and the time-points 12, 24, 36 and 48 the EARLY_MID stage for the SCOV1 infection. **The bigger number and the extend of EARLY_MID time-points for SCOV1 is to be expected because the onset of SCOV1 infection is generally delayed, mainly due to delayed immune responses**, as it was showcased in the previous sections. This is to be justified in the next section where the EARLY_MID profiles between the two infections are contrasted.

[14] The pulmonary epithelium is divided into three regions; upper (nasal and oral cavities), lower (trachea and primary bronchi), and distal small airway epithelia (alveolar). **Human airway epithelial** (**HAE**) cell cultures effectively mimic the human bronchial environment, allowing the cultivation of a wide variety of human respiratory viral pathogens (Pickles, 2013; S Banach et al., 2009)

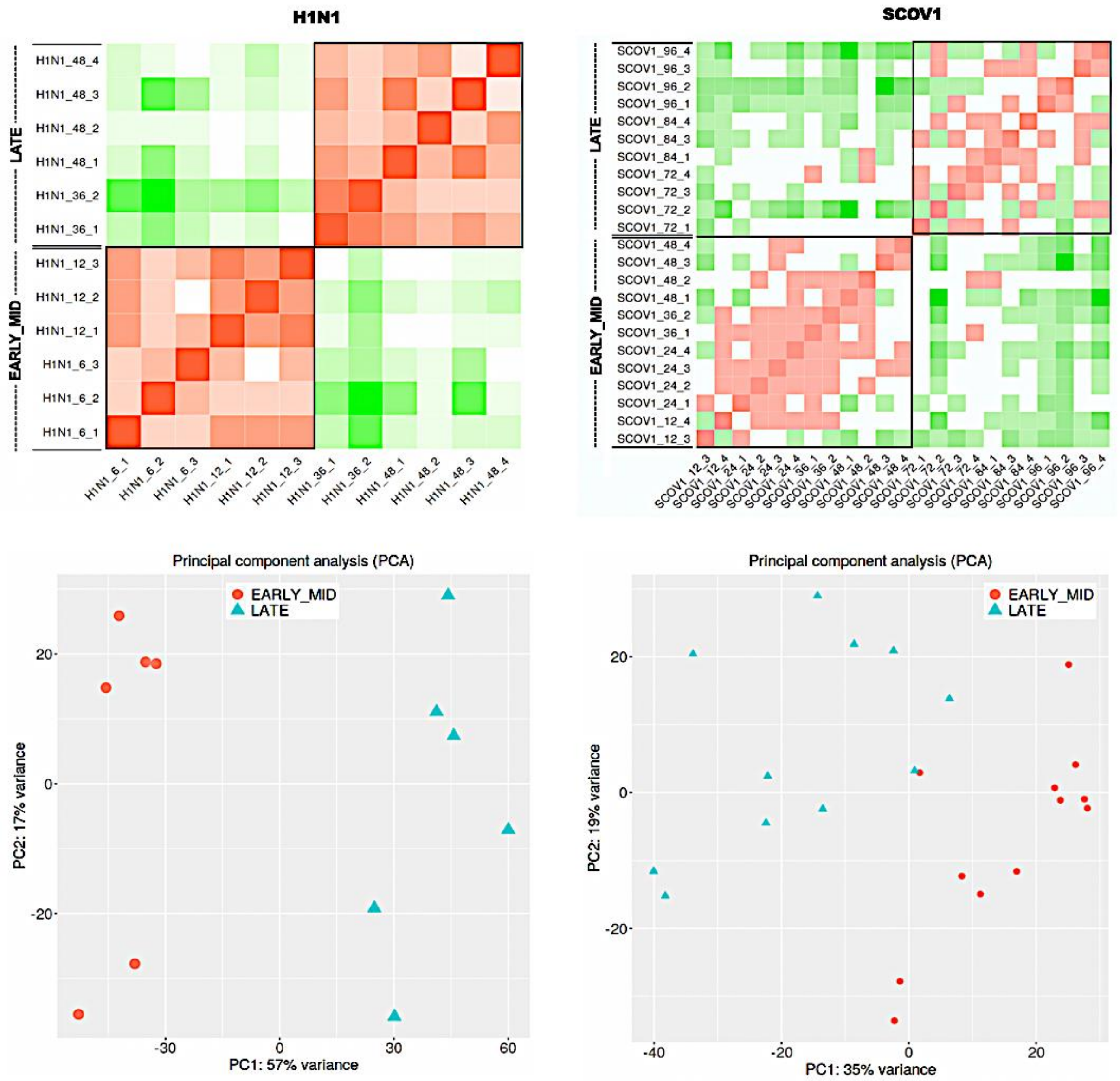


Figure 21. Correlation matrix for H1N1 (upper-left) and SCOV1 (upper-right) samples; boxes indicate the identified EARLY_MID and LATE stages of infection. The lower part of the figure shows the respective PCA plots for the two infections.

Differential expression analysis. Using the *limma* DEG method, with an FDR cutoff value equal to 0.05 and a minimum fold-change equal to 2, a total of **430** genes was found as differentially expressed, **364 down**- and **66 up**-regulated for the comparison SCOV1 vs. H1N1 EARLY_MID infection stages; hpi variable were set as a factor in order to take care of experimental batch effects (Luo et al., 2010). The left part of Figure 22 shows the hierarchical tree organization of samples, and the upper-right part the respective PCA plot; note the perfect separation of H1N1/SCOV1 samples into their respective EARLY_MID infection stage. At the down-right of the figure, the heatmap of identified DEGs is illustrated. Again, we note that down-regulated genes refer to SCOV1 EARLY_MID infection stage, and these genes are (dually) up-regulated the H1N1EARLY_MID infection stage.

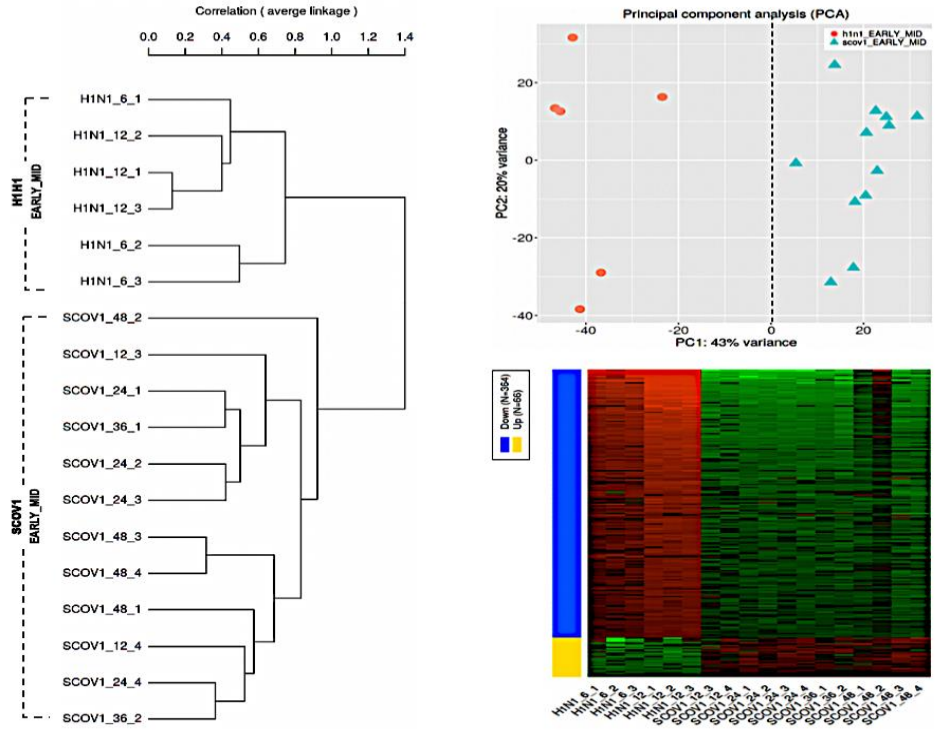


Figure 22. SCOV1 vs. H1N1 EARLY_MID contrast: Hierarchical tree (left), PCA plot (upper-right) of samples, and heatmap of differentially expressed gene/probes (down-right).

As in the case of SCOV2 experiment (see section 3.1.3) we concentrated on the identified down-regulated genes at the EARLY_MID SCOV1 infection stage. In order to get a better insight into the role of these genes, we focus on the gene/probes that: (i) exhibit fold-change values over 3, and (ii) are clustered into coherent groups; the MCL clustering algorithm was used (provided by the STRING server), as in the previous experiment. The result of these screening operations was a set of 118 genes (Table 7). The coloring of genes (light-red, green and yellow) is made according to their cluster/group inclusion.

Table 7. SCOV1 vs. H1N1 EARLY_MID contrast: Coherently grouped 118 down-regulated differentially expressed genes at the EARLY_MID SCOV1 infection stage.

Gene	log2FC	adj.Pval
CXCL10	-7.302	7.79E-06
IFNL3	-7.161	3.73E-08
CXCL11	-6.843	7.58E-06
IFNL1	-6.128	6.07E-09
OASL	-5.920	5.83E-09
IFNL2	-5.783	8.81E-09
CXCL9	-5.661	7.17E-08
SSTR2	-5.444	1.26E-07
RSAD2	-5.320	1.49E-08
IFIT2	-5.303	4.03E-07
CCL5	-5.143	1.53E-09
AIM2	-4.976	6.29E-08
IFNB1	-4.918	2.55E-06
IFIT3	-4.893	4.53E-09
IFIT1	-4.775	1.68E-08
TNFSF13B	-4.696	1.66E-08
TNLG7A	-4.696	1.66E-08
MX2	-4.536	2.14E-09
EPSTI1	-4.488	6.67E-10
IFI44L	-4.272	1.98E-08
IFIT1P1	-4.258	4.17E-04
ISG15	-4.256	1.14E-08
COMMD9	-4.168	3.57E-10
DDX58	-4.030	6.18E-10
APOBEC3A	-4.027	6.67E-06
FAM247A	-3.849	5.57E-10
OAS3	-3.763	5.20E-09
LAMP3	-3.745	2.41E-09
FAM247B	-3.713	5.49E-10
OAS2	-3.659	4.26E-09
USP30-AS1	-2.370	7.88E-08
IFIT5	-2.350	2.30E-10
TAP1	-2.278	2.30E-10
DHX58	-2.275	6.76E-10
LGP2	-2.275	6.76E-10
CD38	-2.268	1.22E-05
CRISPLD2	-2.174	1.70E-08
ZC3HDC1	-2.173	1.61E-09
NUPR1	-2.164	1.37E-09
SLC25A28	-2.163	5.28E-12
WARS	-2.160	2.77E-07
WARS1	-2.160	2.77E-07
MLKL	-2.156	3.73E-08
APOBEC3G	-2.127	2.58E-08
IFI27	-2.114	5.54E-05
BCL2L14	-2.113	2.42E-09
CHROMR	-2.101	4.53E-09
IFITM4P	-2.081	2.79E-08
SUSD3	-2.074	2.25E-06
TMEM106A	-2.065	3.97E-09
CD68	-2.062	2.25E-09
TRIM69	-2.057	5.07E-11
STARD5	-2.057	2.30E-10
TOR1B	-2.056	2.37E-11
SERPINB9P1	-2.006	8.22E-06
TDRD7	-1.973	2.37E-11
ZC3HAV1	-1.958	1.37E-10
IFITM3	-1.946	2.85E-08
TRIM21	-1.865	4.94E-10
TRANK1	-1.864	1.19E-09
BATF2	-3.653	8.46E-11
DDX60L	-3.598	2.63E-10
MX1	-3.485	1.63E-09
THEMIS2	-3.436	6.57E-10
OAS1	-3.425	9.24E-10
DDX60	-3.227	2.36E-10
STAT1	-3.154	2.10E-09
SAMD9L	-3.134	1.53E-09
LINC02574	-3.113	1.28E-07
CD274	-3.105	3.76E-09
IFIH1	-3.007	5.03E-10
IFI44	-2.992	2.97E-09
ETV7	-2.870	3.77E-11
HSH2D	-2.860	3.57E-10
GBP1	-2.790	2.16E-10
IDO1	-2.748	1.19E-05
TMEM140	-2.741	8.52E-10
IRF7	-2.739	2.55E-09
SLC15A3	-2.706	7.23E-09
GBP4	-2.697	1.13E-09
ISG20	-2.688	1.87E-09
SAMD9	-2.687	7.90E-09
IFI35	-2.645	1.69E-10
GBP5	-2.643	1.91E-07
BST2	-2.623	2.19E-07
GBP1P1	-2.608	4.52E-10
PPM1K	-2.592	1.19E-09
RTP4	-2.430	3.82E-10
XAF1	-2.408	1.84E-06
PSAT1	-1.860	1.87E-06
GNB4	-1.860	1.28E-06
RNF213	-1.846	7.70E-10
TGM2	-1.846	7.90E-09
ATP10A	-1.809	5.49E-10
ACKR4P1	-1.796	1.46E-08
IFITM3P7	-1.790	1.47E-07
APOL2	-1.786	2.12E-08
PSMB9	-1.785	1.65E-09
RBM3	-1.777	1.48E-08
TRIM14	-1.766	2.88E-10
APOL6	-1.746	4.22E-12
TREX1	-1.745	5.72E-10
SCAMP1-AS1	-1.744	5.12E-09
RP4-560B9.4	-1.724	1.30E-07
LAP3	-1.698	3.85E-07
CBSL	-1.684	5.03E-05
AHNAK	-1.668	2.86E-06
IFITM2	-1.658	7.56E-08
APOL4	-1.637	2.51E-05
HSPA6	-1.633	2.83E-03
DUSP5	-1.632	2.08E-07
IFITM8P	-1.624	1.64E-07
ANO7L1	-1.618	2.02E-09
ANGPTL4	-1.617	8.12E-05
Lnc-GRAP-3	-1.613	1.47E-03
IFI16	-1.604	6.20E-09
SHFL	-1.595	5.03E-10
IFITM3P1	-1.592	6.16E-08

The network organization of these coherently clustered genes is illustrated in Figure 23.

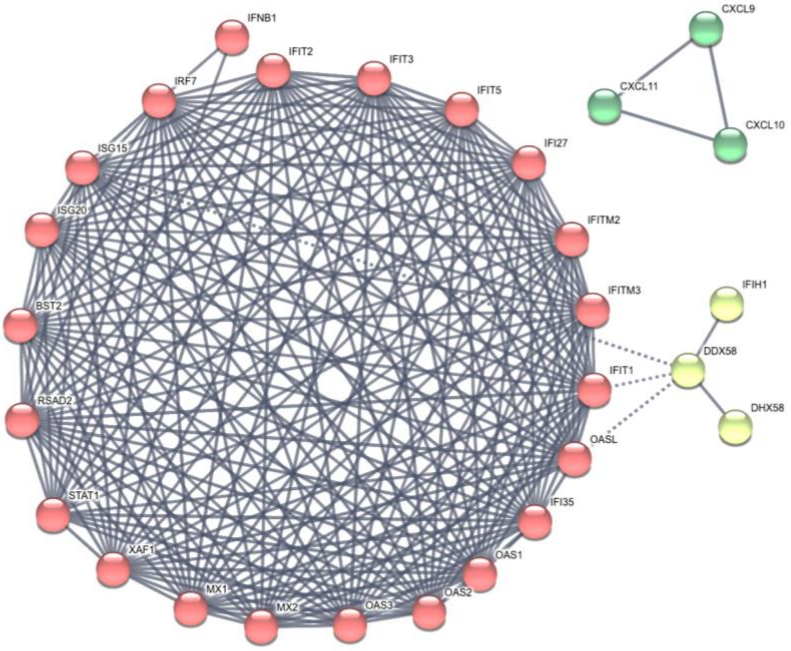


Figure 23. SCOV1 vs. H1N1 EARLY_MID contrast: Network organization of the 118 down-regulated differentially expressed genes at the EARLY_MID SCOV1 infection stage.

The results could be further justified by inspecting the expression levels of these genes. Figure 24 shows the bar-plots of those down-regulated genes that exhibit clearly contrasted expression levels; lower for SCOV1 and higher for H1N1 at the EARLY_MID respective infections' stages; also, most of these genes exhibit higher/lower expression levels for SCOV1/H1N1 compared to the average expression value of all input gene/probes (it is about 11).

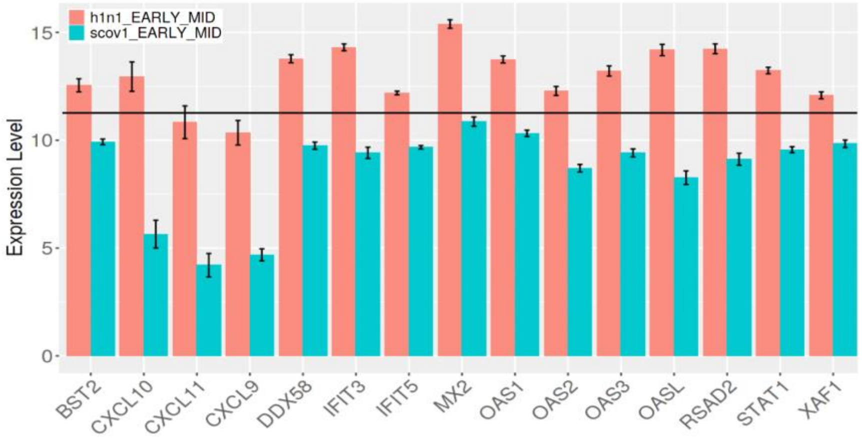
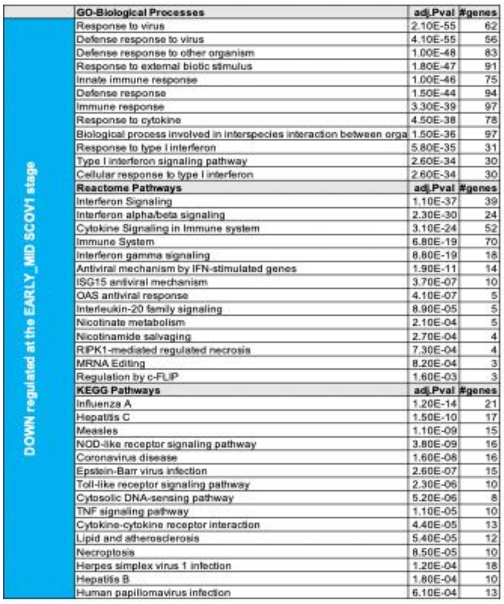


Figure 24. SCOV1 vs. H1N1 EARLY_MID contrast: Subset of the 118 down-regulated differentially expressed genes at the EARLY_MID SCOV1 infection stage with expression values over and down the average for H1N1 and SCOV1 samples, respectively.

Enrichment/Pathway analysis. As in the previous experiments, we conducted an enrichment/pathway analysis for all of the identified DEGs regarding GO-Biological Processes, Reactome and KEGG pathways. The left part of Figure 25 shows these processes and pathways, the left part of the figure the network organization of the GO-biological processes. It may be easily observed that most of the **down-regulated, i.e., impaired biological processes and pathways during the EARLY_MID SCOV1 infection stage are engaged to interferon/cytokine signaling and to defense/immune responses.**



	GO-Biological Processes	adj.Pval	#genes
DOWN regulated at the EARLY_MID SCOV1 stage	Response to virus	2.10E-55	62
	Defense response to virus	4.10E-55	56
	Defense response to other organism	1.00E-48	83
	Response to external biotic stimulus	1.80E-47	91
	Innate immune response	1.00E-46	75
	Defense response	1.50E-44	94
	Immune response	3.30E-39	97
	Response to cytokine	4.50E-38	78
	Biological process involved in interspecies interaction between orga	1.50E-36	97
	Response to type I interferon	5.80E-35	31
	Type I interferon signaling pathway	2.60E-34	30
	Cellular response to type I interferon	2.60E-34	30
	Reactome Pathways	**adj.Pval**	**#genes**
	Interferon Signaling	1.10E-37	39
	Interferon alpha/beta signaling	2.30E-30	24
	Cytokine Signaling in Immune system	3.10E-24	52
	Immune System	6.80E-19	70
	Interferon gamma signaling	8.80E-19	18
	Antiviral mechanism by IFN-stimulated genes	1.90E-11	14
	ISG15 antiviral mechanism	3.70E-07	10
	OAS antiviral response	4.10E-07	5
	Interleukin-20 family signaling	8.90E-05	5
	Nicotinate metabolism	2.10E-04	5
	Nicotinamide salvaging	2.70E-04	4
	RIPK1-mediated regulated necrosis	7.30E-04	4
	MRNA Editing	8.20E-04	3
	Regulation by c-FLIP	1.60E-03	3
	KEGG Pathways	**adj.Pval**	**#genes**
	Influenza A	1.20E-14	21
	Hepatitis C	1.50E-10	17
	Measles	1.10E-09	15
	NOD-like receptor signaling pathway	3.80E-09	16
	Coronavirus disease	1.60E-08	16
	Epstein-Barr virus infection	2.60E-07	15
	Toll-like receptor signaling pathway	2.30E-06	10
	Cytosolic DNA-sensing pathway	5.20E-06	8
	TNF signaling pathway	1.10E-05	10
	Cytokine-cytokine receptor interaction	4.40E-05	13
	Lipid and atherosclerosis	5.40E-05	12
	Necroptosis	8.50E-05	10
	Herpes simplex virus 1 infection	1.20E-04	18
	Hepatitis B	1.80E-04	10
	Human papillomavirus infection	6.10E-04	13

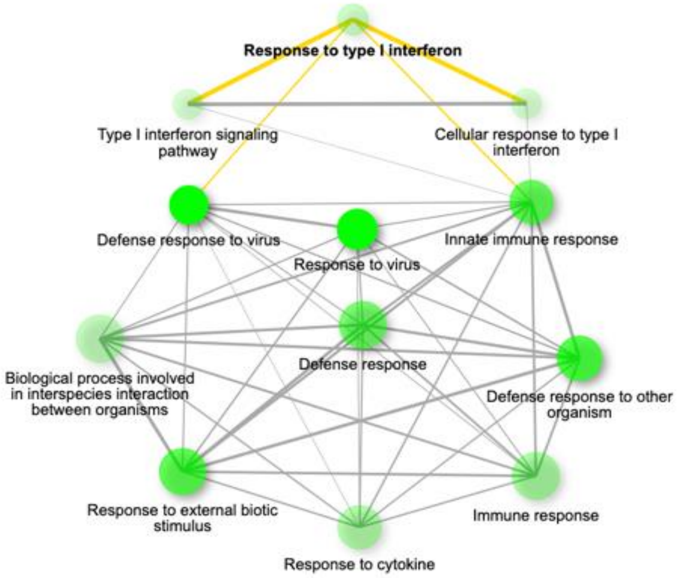

Response to type I interferon
Type I interferon signaling pathway
Cellular response to type I interferon
Defense response to virus
Response to virus
Innate immune response
Defense response
Biological process involved in interspecies interaction between organisms
Defense response to other organism
Response to external biotic stimulus
Immune response
Response to cytokine

Figure 25. SCOV1 vs. H1N1 EARLY_MID contrast: Enriched down-regulated biological processes and pathways (left); network organization of the GO-biological processes (right).

Most of the identified down-regulated biological processes and pathways at the EARLY_MID SCOV1 stage are in common with the down-regulated processes/pathways at the early SCOV2 infection stage, as showcased in the previous sections. This is in contrast to H1N1 infection where all these processes and pathways are up-regulated and functional at the early stage of the infection.

Figure 26 illustrates the regulation of biological processes and KEGG pathways across the SCOV1 and H1N1 samples at the EARLY_MID stage of the two infections. It is noticeable that **even the KEGG 'Coronavirus' as well as the 'Influenza' KEGG pathways are down-regulated at the EARLY_MID SCOV1 stage!**

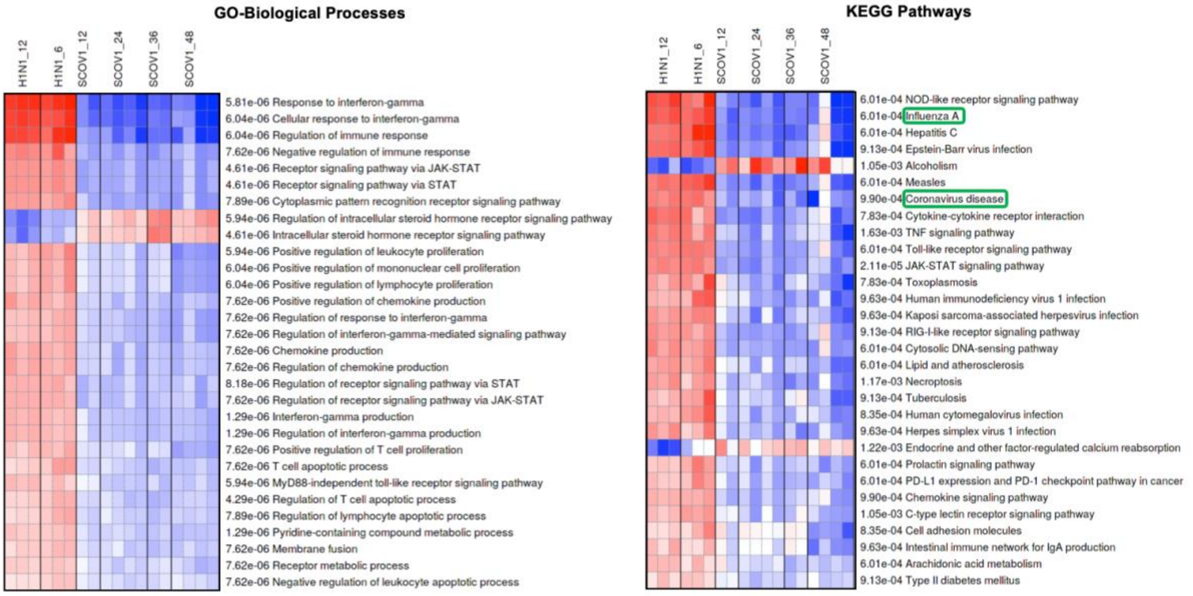


Figure 26. SCOV1 vs. H1N1 EARLY_MID contrast: Regulation of GO-biological processes (left) and KEGG pathways (right) across the SCOV1 and H1N1 EARLY_MID stages; blue color indicates down-regulation/dysfunctional and red color indicates up-regulation/functional process/pathway.

The result are in line with the findings in the previous sections about (a) **delayed immune/defense responses, mainly due to the down-regulation of interferon/cytokine signaling and activation**; (b) relevant observations and discussion made in section 1.3 , and (c) in accordance with recent studies related to SCOV2 infection and Covid-19 disease (Vastrad, Vastrad, & Tengli, 2020).

3.4 Does SCOV2 two-stage profile relates to Covid-19 severity?

In this section we move our attention on a more clinical task, trying to explore the **putative relation between Covid-19 disease severity and the two-stage SCOV2 infection profile**. The aim is to explore the **potential of SCOV2 delayed immune/defense responses to be linked with disease severity**. In particular, we focus on different Covid-19 phenotypes as defined by respective patients' clinical profiles, and especially by the **duration of infection symptoms**.

Dataset. The utilized dataset comes again form GEO, under the code **GSE161731**. The relevant study includes the RNA-seq profiles of 77 Covid-19 patients, acquired via the 'Illumina NovaSeq 6000' platform (**GPL24676**) comprising a total of 60,675 ENSEMBL gene transcripts. Whole blood samples were collected between 1-35 days post infection and divided based on disease severity and time from symptom onset (based on patients' self-reporting and subsequent follow-up): **SHORT** with ≤10 days duration of symptoms, **MEDIUM** with 11-21 days, and **LONG** with >21 days. Peripheral blood sample, for NGS analysis and production of the respective RNA-seq data, was received from patients at the time of enrolment, and duration of symptoms was based on patients' self-reporting and subsequent follow-up. The study and the available dataset include a total of 198 samples from patients with Covid-19 and other infections namely, 'other CoVs'/CoV, INFL, bacterial, as well as healthy controls (77 SCOV2, 61 CoV, 17 INFL, 24 bacterial, 19 healthy). The results of this study are published in (McClain et al., 2021).

Data filtering and phenotype re-assignment. The original unformalized RNA-seq counts data for the 77 SCOV2 infected individuals were utilized (normalized version of data is also available from GEO) in which, the distribution of the three SCOV2 infection **phenotypes** to have as follows: 19 SHORT, 36 MEDIUM and 22 LONG. But **self-reporting of the onset of symptoms is a bit 'subjective' and may not be so accurate**. So, we followed a careful **sample filtering** process followed by **re-assignment of samples to newly invented phenotypes**. To this end we produced the heatmap of samples (see upper-left part of Figure 27). Inspecting the heatmap it can be easily observed that the MEDIUM phenotypes are distributed and mixed with both SHORT and LONG samples (at the left and at the right part of samples' dendrogram, respectively) but **two tangibly contrasted groups** may be clearly identified. Some samples that belong to either SHORT or LONG phenotypes and interfere within these groups are deleted from the dataset, i.e., LONG samples from the left part of the dendrogram, and some SHORT samples from the right part (this process was repeated two times until two clearly contrasted groups are formed). The result is a newly formed dataset with the samples re-assigned to two new phenotypes with a *natural interpretation*: **SHORT_MEDIUM** with 25 samples, and **MEDIUM_LONG** with 43 samples; see upper-right part of Figure 27, and note the perfect separation of samples into the newly formed SCOV2 phenotypes. This may be also observed by inspecting the PCA plot of samples (see down part of Figure 27).

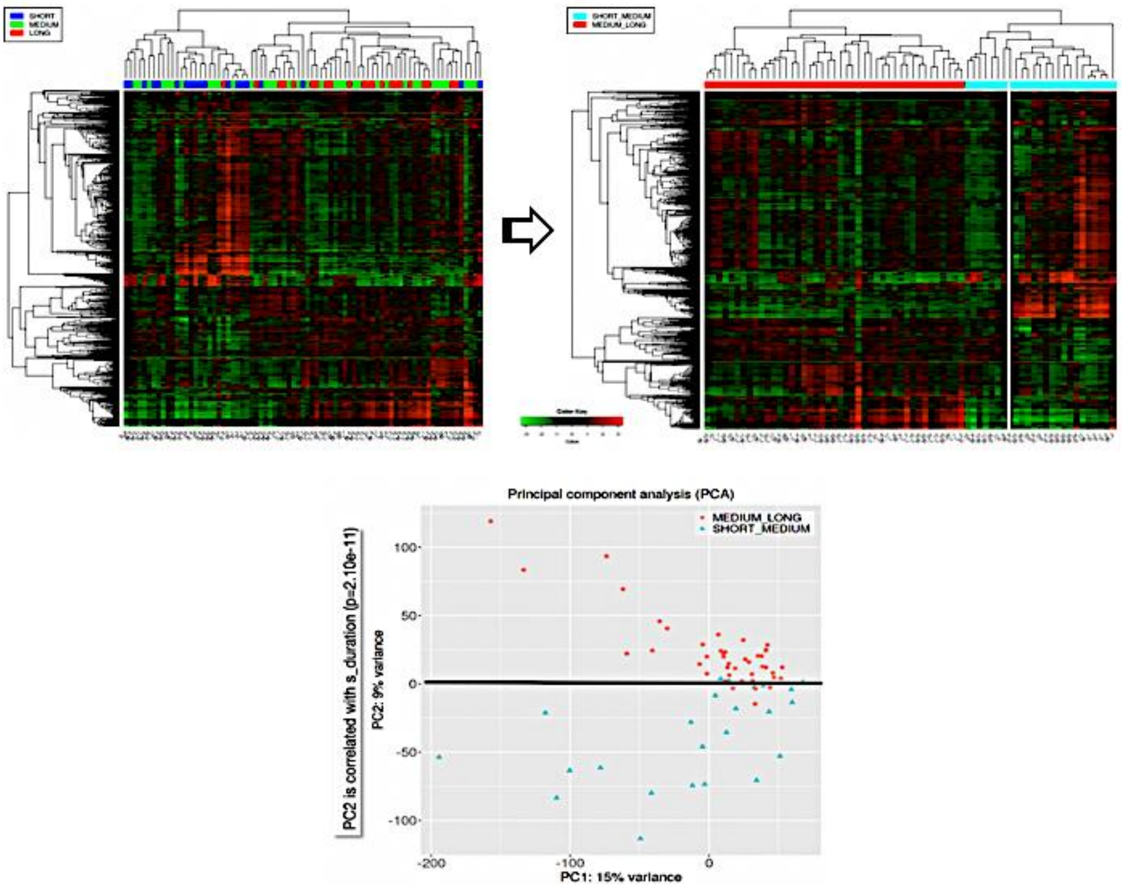


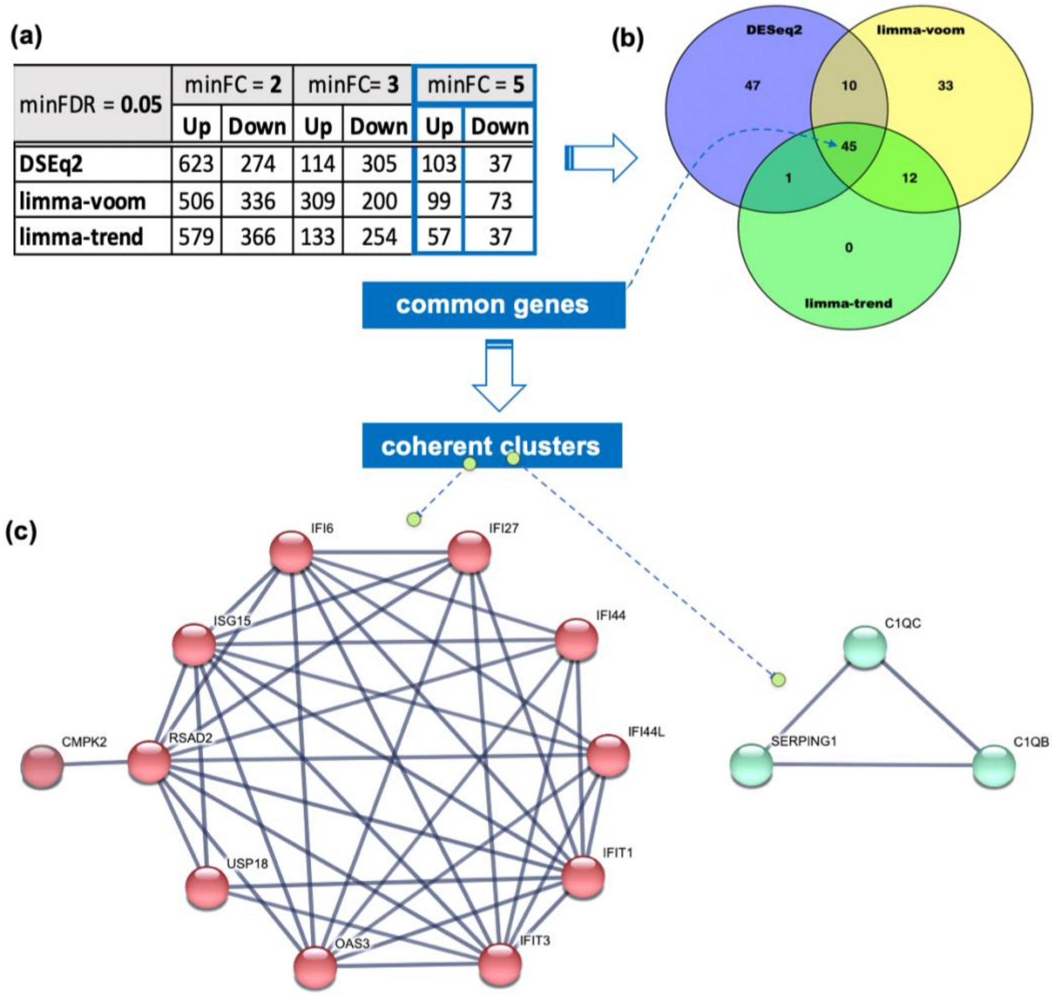
Figure 27. Heatmap of samples (in their original phenotypes, SHORT, MEDIUM, LONG) (upper-eft); Heatmap of filtered samples re-assigned to their new phenotypes, i.e., SHORT_MEDIUM, M~EDIUM_LONG (upper-right); PCA plot of samples (down); the heatmaps were produced using the 1000 most variable gene transcripts

The duration of symptoms for Covid-19 patients is linked to the severity of the disease. Contrasting between SHORT_MEDIUM and MEDIUM_LONG SCOV2 phenotypes is of great importance as it may reveal putative biomarkers that could support clinical decision making.

3.4.1 Contrasting SHORT_MEDIUM with MEDIUM_LONG phenotypes

Data filtering and pre-processing. The newly formed dataset, with the new phenotypes assigned to samples, was downloaded to iDEP. Following the same, as in the previous experiments filtering process (min CPM = 0.5 in at-least 25 samples; the number of samples for the phenotype with less samples, i.e., SHORT_MEDIUM), we were left with 15,998 gene-transcripts (from the original 60,675). The data were also normalized using the EdgeR method (Robinson, McCarthy, & Smyth, 2010) based on the *log2(CPM+c)* transformation of CPM count values (*c* was set equal to 4); average and median of gene-expression values for all input samples was near to 9.

Differential expression (DEG) analysis. The transformed data were analyzed using all the iDEP available DEG methods, i.e., *DSEeq2* (Love, Huber, & Anders, 2014), *limma-voom* (Law et al., 2014) and *limma-trend* (Phipson, Lee, Majewski, Alexander, & Smyth, 2016) from the limma R package, with minimum FDR cutoff set to 0.05. Different minimum fold-change cutoffs (2, 3 and 5) were set and used in order to test the stability of results, each one giving different numbers of up- and down-regulated gene/transcripts (see Figure 28a).



minFDR = **0.05**	minFC = **2**		minFC= **3**		minFC = **5**	
	Up	Down	Up	Down	Up	Down
DSEq2	623	274	114	305	103	37
limma-voom	506	336	309	200	99	73
limma-trend	579	366	133	254	57	37

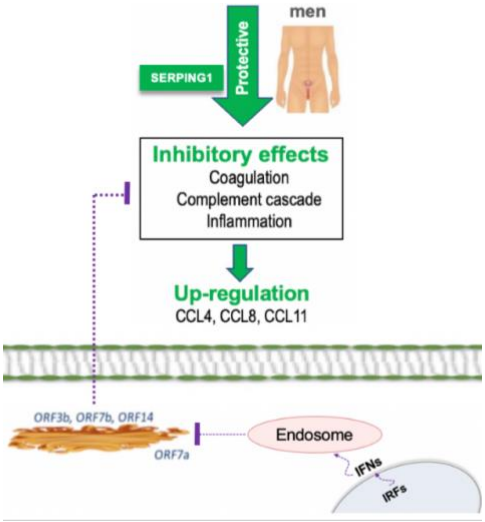
Figure 28. (a) Differentially expressed genes from the three DEG methods (DSEeq2, limma-voom, limma-trend) for the comparison SHORT_MEDIUM vs. MEDIUM_LONG (duration of SCOV2 infected individuals); **(b)** Common up-regulated genes (n = 45) shared among the three DEG methods; (c) Coherent clusters of the 46 common up-regulated genes.

As we are mainly concerned to provide an interpretable biological background for the results we decided to go with minFC = 5 so that to induce a **manageable number of highly contrasted DEGs**. In an effort to produce robust results we found the **common differentially expressed and up-regulated gene/transcripts among all DEG methods**. The focus is set on the up-regulated gene/transcripts as we observed that host immune/defense-related gene/transcripts are included in these. A total of **45 up**-regulated DEG gene/transcripts were found as common among the three methods. **Error! Reference source not found.**b illustrates the Venn diagram of these DEGs. We **focus on the up-regulated genes** (103, 99 and 57 for DSEeq2, limma-voom and limma-trend, respectively) **after observing that most of the DEGs found by all three DEG methods are mainly related to host immune/defense operations**. Uploading the 45 up-regulated genes to the STING server and following the same methodology as in the previous experiments, we were able to induce two robust and coherent clusters (**Error! Reference source not found.**c).

The red-colored genes (11 genes, left part of **Error! Reference source not found.**c) are all ISGs (interferon stimulated genes), and all of them are shared with the DEGs found as down-regulated at the early SCOV2/1 infection stage (as it was showcased in the previous experiments). A natural hypothesis amenable to this finding may be stated:

- ➢ **For Covid-19 patients with relatively short duration of symptoms the delayed immune/defense response profile, mainly manifested by the down-regulation of ISGs, and which characterizes the early SCOV1/2 stage seems to be ‘canceled’.**
- ➢ **In other words, the early induction of ISGs seems to be beneficial for the disease prognosis.**

Of special interest are the genes in the second coherent cluster (green-colored, right part of **Error! Reference source not found.**c). Some references from the gene annotations provided by STRING and other resources (from Wikipedia to published relevant papers) highlight their role in the pathophysiological features underlying SCOV2 infection.



■ **SERPING1.** A plasma protease of the (complement) C1 inhibitor; may play a potentially crucial role in regulating important physiological pathways including **blood coagulation**, fibrinolysis and the generation of kinins. ❍*fibrinolysis* ***prevents blood clots from growing and becoming problematic***. ❍***kinins mediate inflammatory responses by triggering the immune system***; regulate cardiovascular and renal function through ***mediating the effects of ACE inhibitors***. In a very recent paper that explores the links between gender and SCOV2 infection (Russo et al., 2021), it is demonstrated that, for men individuals, “… *the presence of SERPING1 in the testis* ***could prevent thrombotic risk*** *as SCOV2 may block SEPING1 function* ***increasing inflammatory processes****, and deteriorated SERPING1 expression caused by SCOV2 interacting proteins could activate the intrinsic coagulation pathway, inducing a pro-coagulant state*.” The side figure (an edited simplified version of Figure 3 in the aforementioned paper) illustrates and highlights the protective role of SRPING1 in the course of SCOV2 infection. Note the role of interferons IRF/IFNs in the inhibition of specific SCOV2 virus domains that inhibit the beneficial inhibition effect of SERPING1 to coagulation, complement and inflammation cascades. ■ **C1QC/B.** Both belong to the superfamily of SERPIN proteins. It is reported that C1Q complement proteins interacts with 7 different SCOV1 proteins and polypeptides, encoded by ORF3b, ORF7b, ORF14, nsp2ab, nsp13ab, nsp14ab and nsp8ab, with these SCOV1 proteins to be comparable to their SCOV2 homologous (Thomson, Toscano-Guerra, Casis, & Paciucci, 2020). It is also well established that the complement system rapidly and with high specificity detects, traces, targets and eradicates pathogens, by binding to antigen-antibody complexes during an adaptive immune response (Fodil & Annane, 2021).

Enrichment/Pathway analysis. We performed enrichment/pathways analysis (using the respective iDEP operations) on GO-biological Processes and Reactome, KEGG pathways for the differentially expressed genes found by the three DEG methods (DSEeq2, limma-voom and limma-trend). As each method found different processes/pathways as enriched we looked for the common ones between the three methods. A set of six GO-processes, three Reactome and three KEGG pathways were found to be shared among the three DEG methods (Table 8). As, in general, each method computes a different (adjusted) p-value for each process/pathway, we calculated the respective **combined p-values**. We utilized the metap R package for the Fisher and Edgington p-value combination methods (Heard & Rubin-Delanchy, 2018).

Table 8. Common up-regulated biological processes and pathways between three DEG methods (DSEeq2, limma-voom, limma-trend) for Covid-19 patients of the EARLY_MEDIUM phenotype, and their combined (adjusted) p-values.

	Combined adj.Pval	
	Fisher	**Edgington**
GO-Biological Processes		
Type I interferon signaling pathway	5.13E-29	1.10E-20
Cellular response to type I interferon	5.13E-29	1.10E-20
Response to type I interferon	9.71E-29	1.46E-20
Response to external biotic stimulus	8.51E-25	2.93E-17
Response to virus	8.22E-23	1.37E-16
Innate immune response	2.36E-21	7.18E-16
Reactome pathways		
Interferon alpha/beta signaling	2.46E-35	2.04E-24
Immune System	8.16E-13	3.01E-09
Classical antibody-mediated complement activation	1.46E-07	6.61E-09
KEGG pathways		
Coronavirus disease	1.98E-08	1.51E-07
Staphylococcus aureus infection	1.29E-06	1.09E-07
Complement and coagulation cascades	6.75E-06	4.29E-07

As can be easily observed, all the GO-biological processes and Reactome pathways concern host immune/defense responses, with interferon signaling ones at the top. The enriched up-regulated Reactome and KEGG complement-related pathways relates to the discussion made above about the SERPING1 and C1QB/C complement-related SCOV2 phenotype differentiating genes, which were also found as up-regulated.

- **It seems that the molecular profile underlying some Covid-19 patients (with SHORT_MEDIUM duration of symptoms) is characterized by early triggering of defense-immune responses, with early activation of specific IFN signaling pathways that allows the protective effects of various intrinsic host viral defense factors such as serpins (SERPING1) and coagulation (ICQ) inhibitors.**
- **These factors need to be further explored as putative Covid-19 prognostic biomarkers.**

3.4.2 Last minute update –A special note on gene IFI27

In the list of up-regulated gene in the SHORT_MEDIUM vs. MEDIUM_LONG comparison, gene **IFI27** is found as with the **highest differentiation** between the two SCOV2 phenotypes (fold-change = 29.4). Figure 29 shows the (normalized) expression level of IFI27 in the two phenotypes (left), and the network of genes immediate connecting with it. The direct link of IFI27 with most of the identified, either as down-regulated during the early SCOV2 infection stage or as up-regulated in Covid-19 patients with SHORT_MEDIUM symptoms phenotype, indicate its central role in SCOV2 infection and Covid-19 disease development.

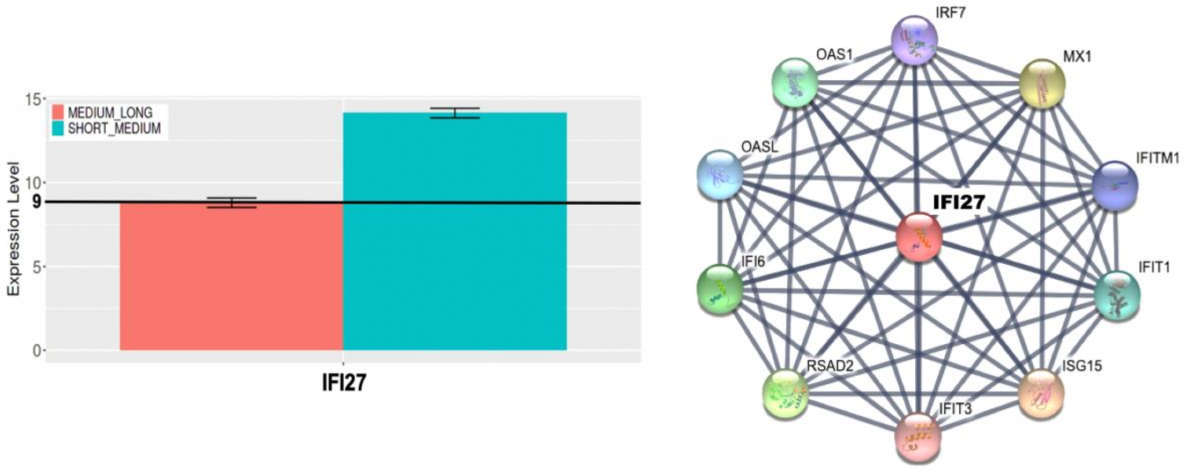


Figure 29. Bar-plot of IFI27 expression level in SHORT_MEDIUM and MEDIUM_LONG SCOV2 (duration of symptoms) phenotypes; 9 is the average expression value over all input gene/transcripts, (left); The network of genes which are immediate connected with IFI27 (produced by the STRING server).

❑ **IF27 as a Covid-19 infection detector.** In a recent publication in Lancet (Gupta et al., 2021) that reports results from a large-scale nested, case-control diagnostic accuracy study, it is clearly noted and stated: "… *blood transcriptomic biomarkers for viral infection, including* ***IFI27****, reflect IFN-I responses and* ***detect early SARS-CoV-2 infection with high accuracy***". In addition, it is suggested that these biomarkers should be included in scalable point-of-care tests for SCOV2 in a way to **facilitate early case detection and contact investigation**.

❑ **New Covid-19 vaccines? Targeting, not the invasion, but the virus replication**

- ***Cover story.*** We quote verbatim from a recent cover article in Nature[15]. "*In the study, published on 10 November in Nature, the authors examined blood samples collected in the first weeks of the pandemic from nearly 60 UK health-care workers. All worked in hospitals, putting them at high risk of contracting COVID-19, but* ***never tested positive or produced any antibodies to the virus for four months after enrolling in the study****. The researchers noticed that in 20 of these 'seronegative' participants,* ***T cells had multiplied*** *— a sign that the* ***immune system might be gearing up to fight an infection****. Nineteen of these individuals also had increased levels of an immune-system protein called* ***IFI27****, which the authors say might be an early marker of SARS-CoV-2 infection. The authors say that these data are evidence for 'abortive infections', meaning that the virus made an incursion into the body but failed to take hold. The authors hypothesized that* ***T cells halt SARS-CoV-2 by disabling a cluster of viral proteins called the replication transcription complex, which helps the virus to reproduce.****"*

- The study that the cover article refers is titled "***Pre-existing polymerase-specific T cells expand in abortive seronegative SARS-CoV-2***" (Swadling et al., 2021), it is published as an '*accelerated article*', a fact that highlights its importance, and includes intensively monitored health care workers (HCWs) as the trial subjects. Key findings, reports and suggestions from the article are:
  – *some of the SCOV2 infected HCWs do not show PCR- or antibody-positivity (i.e., sero-negative), and this is an indication for a sub-clinical (before seroconversion takes place)* ***rapid clearance*** *of the virus;*

[15] How do people resist COVID infections? Hospital workers offer a hint, Nature News, 11 November 2021.

- *this may be attributed to the **pre-existence of memory T-cell responses as the result of previous infections from other coronaviruses**, and results in **early cross-recognition of SCOV2 infection**;*
- *candidates for the source of these pre-existing T-cells are the **closely related human endemic common cold coronaviruses** (HCoVs);*
- *sero-negative HCWs (SN-HCWs) had: stronger, more frequently directed immune response against the replication and transcription complex (RTC)[16] of the virus, and an **increase in IFI27, a robust early innate signature of SCOV2** (Gupta et al., 2021), suggesting **abortive infection**;*
- *the results highlight RTC-specific T-cells as **targets for vaccines** against endemic and emerging infections from various coronavirus (in the family of Coronaviridae[17] where SCOV2 belongs).*

[16] SCOV2 ORF1a and ORF1b domains encode 15–16 non-structural proteins (nsp), of which 15 compose the viral RTC (V'kovski, Kratzel, Steiner, Stalder, & Thiel, 2021).

[17] Family Coronaviridae, Viruses, 2017

3.5 COVID-19 prognosis: Predicting the duration of infection symptoms

SCOV2 infection is challenging the health care systems worldwide as the percentages of infected patients needing **hospitalization** care range in high levels. Taking the latest figures from the US CDC (Center for Disease Control and Prevention)[18], for the whole period of SCOV2 pandemic, about 7.5 million hospitalizations are estimated for a total of (symptomatic and asymptomatic) 146 million cases, a rate of about **5%**. Projecting this estimate to Greece, with a total of about 750,000 reported cases in 2021, the estimate is about 38,000 hospitalizations. If we divide this number by 3, the three COVID-19 waves during 2021, over 12,000 hospitalizations are taking place at a period of 2-3 months, that is, about 5,000 each month. These estimates, but also the real-life events took place, showcase the burden of COVID-19 over the national health care systems worldwide. So, **reliable predictive models for COVID-19 patients needing hospitalized care raises as an imperative need**.

Here we have to note that relevant predictive and prognostic models that rely solely or mainly on patients' clinico-physiological measurements are already in place (Wynants et al., 2020). In this section and based on the results reported in the previous section, we tried to **devise classifiers that predict the duration of symptoms (i.e., SHORT, MEDIUM, LONG) for SCOV2 infected individuals based solely on their gene-expression profiles**. With a reliable estimate for the duration of symptom, informed clinical decision making may take place. For example, if a COVID-19 patient is predicted (by his/her gene-expression profile) as an individual that will display a SHORT or even SHORT-MEDIUM disease progression profile (according to the duration of his/her symptoms) then, this patient may not need hospitalization care or even, will not end-up to ICU care. Of course, the clinico-physiological profile of patients should be also considered. In addition, ***our study aims to demonstrate the feasibility of the approach and by no a ready-to-use tool in the clinical practice***. After all, in a recent MIT Technology Review it is reported that, actually it is titled: "*Hundreds of AI tools have been built to catch covid. None of them helped*"[19]. The report exemplifies the situation and highlight the '*prons'* and *'cons'* of relevant approaches.

3.6 Initial line of experiments

In our initial experiments for the devise of classification models we used the **MLSeq** R/Bioconductor package – a Machine Learning interface to RNA-seq data analysis (Goksuluk et al., 2019). We relied on four classifiers to build our models: **SVM** (support vector machines), **Random Forests** (**RF**), **Voom Based Nearest Shrunken Centroids** (**VoomNSC**) (Zararsiz et al., 2017), and **Poisson Linear Discriminant Analysis** (**PLDA**) (Witten, 2011). Data were split into ***training*** and ***test***. Training set was used to build classification models and test set was used to assess the performance of each model. The ratio of splitting data was 80% for training and 20% for testing. The ***k*-fold (*k* = 5) *inner cross-validation*/CV** training process was followed in order to tune and optimize the model parameters. The tuned trained model was then applied on the left-out test set to get the final *testing* accuracy performance figures; performance figures (accuracy, sensitivity and specificity) from the 5-foldCV are also assessed and reported.

- ❑ **Multi-class model** (SHORT vs. MEDIUM vs. LONG**).** For this experiment the multi-class labeled dataset was used, i.e., the samples are assigned into three classes, SHORT, MEDIUM and LONG (regarding duration of SCOV2 infection). Furthermore, we devised models using either all input gene/transcripts or just the identified differentially expressed

[18] *Estimated COVID-19 Burden*, US CDC

[19] www.technologyreview.com/2021/07/30/1030329/machine-learning-ai-failed-covid-hospital-diagnosis-pandemic/

genes (DEGs), as identified in the previous section. The results are summarized in Table 9.

Table 9. Performance results of the classification model for multi-class labeled data (i.e., samples are assigned to three classes, SHORT, MEDIUM or LONG). Bold figures in blue indicate superior performance for the dataset with ALL genes/transcripts, and bold figures in red indicate superior performance for the dataset with just the identified DEGs.

Method	**Genes**	**ACC** (5-foldCV)	**SE**	**SP**	**ACC** (test)
SVM	ALL	51.6%	40.0%	**88.1**%	50.3%
	DEGs	66.7%	70.0%	**82.5**%	**62.5**%
RF*	ALL	45.9%	35.0%	80.5%	**68.8**%
	DEGs	61.3%	65.0%	78.6%	56.3%
VoomNSC	ALL	**80.3**%	**95.0**%	80.5%	43.8%
	DEGs	**67.2**%	**95.0**%	70.7%	56.3%
PLDA	ALL	59.0%	85.0%	65.9%	50.0%
	DEGs	62.3%	85.0%	70.7%	**62.5**%

*pre-specified number of random trees generated it was set equal to 500

The VoomSNC model achieves the best k-foldCV 'ACC'uracy and 'SE'ensitivity performance (80.3% and 95.0%, respectively, when ALL gene/transcripts are used; and 67.2%, 95.0%, respectively, when just the identified DEGs are used). The best 'SP'ecificity performance figures are achieved by SVM (88.1% for ALL, and 82.5% for DEGs). For the final accuracy on the test set, the best 'ACC'uracy for ALL is achieved by RF (68.8%), and the best 'ACC'uracy for DEGs is achieved by SVM and PLDA (62.5%). A general notice concerns the **inability of the classifier models to increase their performance when a more 'informed' set of features (i.e., the DEGs)** is used; in all metrics the performance figures are better when ALL gene/transcripts are used. This should be attributed to the fact that **SHORT, MEDIUM and LONG phenotypes and their respective gene-expression profiles are not well separated and contrasted** (as already demonstrated by the results in the previous section). So, we attempted to devise the same classifier models using just the samples assigned to the SHORT and LONG phenotypes, as these phenotypes are more well separated and contrasted.

❑ **Two-class model** (SHORT vs. LONG)**.** For this experiment the two-class labeled dataset was used, i.e., only the samples assigned to classes SHORT and LONG are kept. Again, we devised models using either all input gene/transcripts or just the identified differentially expressed genes (DEGs). The results are summarized in Table 10.

Table 10. Performance results of the classification model for two-class labeled data (i.e., only the samples assigned to classes, SHORT and LONG are kept). Bold figures in blue indicate superior performance for the dataset with ALL genes/transcripts, and bold figures in red indicate superior performance for the dataset with just the identified DEGs.

Method	**Genes**	**ACC** (5-foldCV)	**SE**	**SP**	**ACC** (test)
SVM	ALL	87.5%	83.8%	81.3%	66.7%
	DEGs	**100.0**%	**100.0**%	**100.0**%	**77.8**%
RF*	ALL	93.8%	87.5%	**100.0**%	66.7%
	DEGs	**100.0**%	**100.0**%	**100.0**%	**77.8**%
VoomNSC	ALL	**100.0**%	**100.0**%	**100.0**%	**77.8**%

	DEGs	90.6%	81.3%	**100.0**%	66.7%
PLDA	ALL	93.8%	87.5%	**100.0**%	66.7%
	DEGs	93.8%	87.5%	**100.0**%	55.7%

*pre-specified number of random trees generated it was set equal to 500

The VoomNSC model achieves the best k-foldCV 'ACC'uracy, 'SE'ensitivity, 'SP'ecificity (together with RF and PLDA, 100.0%), and test 'ACC'uracy (77.8%) when ALL gene/transcripts are used. When just the DEGs are used the best performing models across all metrics are SVM and RF (100.0% for 5-foldCV 'ACC'uracy, 'SE'ensitivity, 'SP'ecificity and 77.8% for test accuracy). In contrast to the previous results, **all the models exhibit better performance figures when just the DEGs are used**. The finding provides **further evidence for the adequacy of the identified DEGs and their ability to differentiate between the two contrasting SCOV2 phenotypes, i.e., SHORT vs. LONG** duration of symptoms.

3.7 Second line of experiments: A LOOCV assessment approach

Here we present results on the same (as in the previous experiment) dataset following a *Leave-One-Out-Cross-Validation* (**LOOCV**) procedure. In biomedical research, LOOCV is a common process followed in biomedical research for assessing the performance of prognostic models (Li, Wang, Chen, & Wang, 2020; Mistry, Davies, & Di Veroli, 2015; Qu, Zhao, & Yin, 2019). With this process, and for a dataset with *k* samples, *k-1* are used for training and the one left-out sample for testing; the process is repeated *k* times and the final performance figures are computed as the averages of the respective figures at each fold. It is demonstrated that when the number of instances in a data set is small or the number of instances in the classes is unbalanced, k-foldCV suffers from the *independence* and *randomness* assumption when splitting the data. With LOOCV both criteria are met, and the *point estimate of accuracy for a given data set is constant* (Wong, 2015). In our case, the number of samples is not big and the number of samples are unequal in terms of their class assignment. So, LOOCV seems a rational strategy to follow in order to assess the performance of our models.

In our experiments we utilized the processed data from the experiments performed in section 3.4 (i.e., the EdgeR log2 transformed counts). We utilized the **Weka** machine-learning framework for our experiments (Hall et al., 2009), and we devised models using the Random Forests (**RF**), **SVM** (called SMO in Weka), Decision Tree (**DT**, called J48 in Weka), and **kNN** (called iBK in Weka) methods.

❑ **Multi-class model** (SHORT vs. MEDIUM vs. LONG)**.** For this experiment the multi-class labeled dataset was used, i.e., the samples are assigned into three classes, SHORT, MEDIUM and LONG (regarding duration of SCOV2 infection). Again, we devised models using either ALL input gene/transcripts or just the identified differentially expressed (DEG) gene/transcripts. The results are summarized in Table 11.

Table 11. LOOCV performance results of the classification model for multi-class labeled data (i.e., samples are assigned to three classes, SHORT, MEDIUM or LONG). Bold figures in blue indicate superior performance for the dataset with ALL genes/transcripts, and bold figures in red indicate superior performance for the dataset with just the identified DEGs.

Method	Genes	ACC	SE	SP	AUC
RF*	ALL	57.4%	57.4%	66.9%	0.722
	DEGs	**61.8%**	**61.8%**	73.5%	**0.771**
SVM (SMO)	ALL	**63.2%**	**63.2%**	**72.7**%	**0.742**

	DEGs	57.4%	57.4%	**73.6%**	**0.711**
DT (J48)	ALL	58.8%	58.8%	71.5%	0.733
	DEGs	52.9%	52.9%	66.2%	0.650
kNN (iBK)	ALL	50.0%	50.0%	62.4%	0.562
	DEGs	48.5%	48.5%	65.7%	0.571

*pre-specified number of random trees generated it was set equal to 100

The SVM model achieves the best LOOCV performance across of metrics (‘ACC’uracy 63.2%, “SE’ensitivity 63.2%, ‘SP’ecificity 72.7% and Area Under the Curve/AUC 0.742) when ALL gene/transcripts are used. For the cased of DEGs, RF exhibits the best performance (ACC’uracy 61.8%, “SE’ensitivity 61.8%, and Area Under the Curve/AUC 0.771), except for ‘SP’ecificity where SVM is slightly better (73.5% vs. 73.6%). Most of the models could not achieve better results when just the DEGs are used. **This does not hold for RF which, manages to achieve significantly better results when just the DEGs are used**.

- **Two-class model** (SHORT vs. LONG)**.** For this experiment the two-class labeled dataset was used, i.e., only the samples assigned to classes SHORT and LONG are kept. Again, we devised models using either all input gene/transcripts or just the identified differentially expressed genes (DEGs). The results are summarized in Table 12.

Table 12. LOOCV performance results of the classification model for two-class labeled data (i.e., only the samples assigned to classes, SHORT and LONG are kept). Bold figures in blue indicate superior performance for the dataset with ALL genes/transcripts, and bold figures in red indicate superior performance for the dataset with just the identified DEGs.

Method	Genes	ACC	SE	SP	AUC
RF*	ALL	**94.3%**	**94.3%**	**91.4%**	**0.991**
	DEGs	**100.0%**	**100.0%**	**100.0%**	**1.000**
SVM (SMO)	ALL	**94.3%**	**94.3%**	**91.4%**	0.929
	DEGs	**100.0%**	**100.0%**	**100.0%**	**1.000**
DT (J48)	ALL	91.4%	91.4%	87.1%	0.893
	DEGs	97.1%	97.1%	95.7%	0.964
kNN (IBK)	ALL	80.0%	80.0%	79.5%	0.798
	DEGs	**100.0%**	**100.0%**	**100.0%**	**1.000**

*pre-specified number of random trees generated it was set equal to 100

The results from this experiment is quite encouraging. All classifier models exhibit highly performing figures, with RF and SVM achieving the best results (for both All and DEGs cases). It is noticeable that even kNN achieves perfect results for the case of DEGs. This is an additional indication for the suitability of the selected DEGs, and their potential to act as biomarkers that differentiate between SHORT (less severe / mild) and LONG (more severe) SCOV2 infection phenotypes, at-least with reference to the expected duration of symptoms.

- **Hybrid-class model** (SHORT-MEDIUM vs. MEDIUM-LONG)**.** For this experiment the hybrid-class labeled dataset was used, i.e., all the samples are used but assigned to the combined classes SHORT-MEDIUM and MEDIUM-LONG, as formed from the analysis in previous section. Again, we devised models using either all input gene/transcripts or just the identified differentially expressed genes (DEGs). The results are summarized in
- Table ***13***.

Table 13. LOOCV performance results of the classification model for hybrid-class labeled data (i.e., all samples are used and assigned to the hybrid-classes SHORT-MEDIUM and MEDIUM-LONG). Bold figures in blue indicate superior performance for the dataset with ALL genes/transcripts, and bold figures in red indicate superior performance for the dataset with just the identified DEGs.

Method	Genes	ACC	SE	SP	F1	AUC
RF	ALL	92.6%	92.6%	89.0%	0.925	**0.990**
	DEGs	97.1%	97.1%	96.6%	0.971	0.996
SVM (SMO)	ALL	**94.1%**	**94.1%**	**91.6%**	**0.941**	0.928
	DEGs	**100.0%**	**100.0%**	**100.0%**	**1.000**	**1.000**
DT (J48)	ALL	82.4%	82.4%	79.7%	0.824	0.654
	DEGs	89.7%	89.7%	89.0%	0.897	0.800
kNN (IBK)	ALL	91.2%	91.2%	88.2%	0.911	0.897
	DEGs	**100.0%**	**100.0%**	**100.0%**	**1.000**	**1.000**

Again, the results are very good, especially for SVM and kNN, which achieve perfect performance when just the differentially expressed gene/transcripts are used. The results are indicative for:

- **The well-formed hybrid SCOV2 phenotype classes (SHORT-MEDIUM / MEDIUM-LONG duration of symptoms) as they, except from their natural meaning, are also separable in terms of their respective patient's differential gene-expression profiles**, and
- **The identified DEGs are well suited for characterizing the SCOV2 phenotypes and may present putative biomarkers for the prognosis of SCOV2 infection progress and severity (at-least in terms of the expected duration of symptoms period).**

4. Conclusions & Future work

COVID-19 is one of the largest and deadliest pandemics on the planet that still continues to plague humanity with thousands of deaths worldwide every day. Different studies show that the severity of the disease and the mortality rate are directly related to the excessive secretion of **pro-inflammatory cytokines**.

Interferons/ISGs …

Hyperinflammation in severe COVID-19 infected patients and overexpression of cytokines, such as **interferons (**IFNs), interleukins, and TNF-α can lead to the so-called 'cytokine storm' and finally to severe pneumonia, lung failure, and multiple organ damage, with potentially fatal outcomes. Recent studies link COVID-19 severity with viral-load and its relation to a two-stage (early vs. late) infection profile (Walsh et al., 2020). Furthermore, it is established that IFNs, especial **type I IFNs** (**IFN-I**), and the induced **interferon stimulated genes** (ISGs) encode a variety of antiviral effects throughout the whole viral life-cycle (entry, uncoating, genome replication, particle assembly and egress; refer to Figures Figure ***7*** and Figure ***14***). Well-defined studies demonstrate and suggest that, targeting early, post-entry life cycle events is a common mode of ISG action (Schoggins et al., 2011). In particular, *IFITM*s, a specific ISG family, inhibit the life-cycle of various viruses (including Influenza A and H1N1, filoviruses that cause hemorrhagic fever such as Embola, and SCOV1) at their **early steps**, by blocking entry or viral particle trafficking (Brass et al., 2009; Huang et al., 2011; Lu et al., 2011; Mantlo et al., 2020). In addition, several ISGs (including, *IFI6*, *IFI27*, IRFs (IRF1 and *IRF9*), MX1, OAS1 and RSAD2/Viperin) reduce HCV (hepatitis C virus) replicon[20] activity (Itsui et al., 2006).

… **their inhibition by SCOV1**&**2**

As it is showed and reported in (Y.-M. Kim & Shin, 2021), in contrast to human common-cold coronaviruses that induce high expression levels of IFN-I SCOV1, SCOV2, and MERS-CoV induce reduced IFN-I responses. After the first SCOV1 outbreak, several studies showcased that SCOV1 and MERS-CoV use various mechanisms to avoid IFN-I-mediated immune responses (Totura & Baric, 2012), (Sa Ribero, Jouvenet, Dreux, & Nisole, 2020). As SCOV2 genome has 82% nucleotide identity with the SCOV1 genome, and most of SCOV2 proteins have high amino acid sequence homology with the corresponding SCOV1 proteins, many SCOV2 proteins have **inhibitory effects on IFN-I responses** similar to those of SCOV1 proteins. In the same study of Kim & Shin, a rational hypothesis is stated which demonstrates that **initially (at early infection stages) delayed but then (at the late infection stage) exaggerated IFN-I responses are involved in hyperinflammation and contribute to the severe progression of COVID-19** (see section 1.3.2 The SCOV2 molecular framework and Figure 6).

… **demonstrated by thesis results**

In our study, we performed **differential expression** and **enrichment / pathway analysis** utilizing public-domain gene expression datasets (RNA-seq and microarrays) from respective well-documented studies, which associate with infectious diseases including SCOV1, SCOV2

[20] ***Replicons*** are self-amplifying recombinant RNA molecules expressing proteins sufficient for their own replication but which do not produce infectious virions; they resemble virus-like particles that enter a target cell, undergo limited transcription and translation to synthesize encoded proteins, but will not produce infectious progeny (Morrison & Plotkin, 2016); in the context of **vaccine production**, administration of replicon RNA vectors has resulted in strong immune responses and generation of neutralizing antibodies in various animal models (Lundstrom, 2016).

and Influenza. We posted some basic **biological questions and tasks challenging to uncover the molecular landscape of SCOV2 infection**. We attempt to provide answers to these questions and tackle the respective tasks following a **multi-step Bioinformatics pipeline** realized by the utilization of state-of-the-art gene-expression and pathway analysis methodologies, services and tools. We were able to identify **key genes** and **molecular pathways** that: (a) **segregate SCOV2 early and late infection stages**; (b) **differentiate SCOV2 from other common viral infections** (such as influenza), (c) **characterize different SCOV2 severity phenotypes** according to the infection's duration of symptoms, and (d) **induce, based on Machine Learning methodology, classification models that could predict the progression and potential severity of the infection** (in terms of the infection duration of symptoms). The overwhelming majority of identified differential expressed genes and biological processes/pathways: (i) relate to **IFN/ISG genes and host immune/defense mechanisms**, and (ii) these genes and pathways were found as **down-regulated at the early stages** of SCOV2 infection. It is evident that our findings are in accordance to the aforementioned discussion and observations and demonstrate that: **the identification of key cytokines/IFNs/ISGs and molecular pathways that differentiate between: [i] early and late SCOV2 infection stages, and [ii] SCOV2 from other common viral infections, is crucial for the prognosis of COVID-19 disease and could aid therapeutic decision-making**.

What's next …

A. The findings reported in this thesis should be **coupled, validated and strengthen with additional experiments** on similar datasets from other studies; ➔ *the target is the **formation of a reliable and strongly predictive gene-signature as a biomarker for the progression and staging of SCOV2 infection***.

B. As the scientific research in infection diseases continues to evolve, a variety of therapeutic ***drug candidates*** have shown potential to combat the disease severity and balance the hypersecretion of pro-inflammatory cytokines; ➔ *utilizing **sets of genes known to be associated with drugs** (from relevant and well-established drug-gene association resources) **we may follow an enrichment analysis process to prioritize putative SCOV2 treatment drug candidates***.

C. As COVID-19 continues to cause an ongoing pandemic, the scientific community is working rapidly to collect and process COVID-19 data. Thus, a challenging task is the creation of a meta-analysis framework with gene-expression data from the mass-vaccination programs worldwide (relevant studies and respective data are slowly raising!). We have already highlighted the role of particular IFN/ISG genes as putative COVID-19 diagnostic biomarkers as well as putative vaccine targets (refer to section 3.4.2 **Error! Reference source not found.** about gene IFI27); ➔ *the target is the **exploration of the molecular profiles of vaccinated individuals in an effort to identify the molecular fingerprints underlying COVID-19 vaccination***.

References

Anders, S. (2010). HTSeq: Analysing high-throughput sequencing data with Python. *European Molecular Biology Laboratory*, 1–28.

Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C., & Garry, R. F. (2020). The proximal origin of SARS-CoV-2. *Nature Medicine*, *26*(4), 450–452. https://doi.org/10.1038/s41591-020-0820-9

Assiri, A., McGeer, A., Perl, T. M., Price, C. S., Al Rabeeah, A. A., Cummings, D. A. T., … Team, K. S. A. M.-C. I. (2013). Hospital outbreak of Middle East respiratory syndrome coronavirus. *The New England Journal of Medicine*, *369*(5), 407–416. https://doi.org/10.1056/NEJMoa1306742

Banerjee, A., El-Sayes, N., Budylowski, P., Jacob, R. A., Richard, D., Maan, H., … Mossman, K. (2021). Experimental and natural evidence of SARS-CoV-2-infection-induced activation of type I interferon responses. *IScience*, *24*(5), 102477. https://doi.org/10.1016/j.isci.2021.102477

Bao, S., & Knoell, D. L. (2006). Zinc modulates cytokine-induced lung epithelial cell barrier permeability. *American Journal of Physiology-Lung Cellular and Molecular Physiology*, *291*(6), L1132–L1141. https://doi.org/10.1152/ajplung.00207.2006

Bi, Q., Wu, Y., Mei, S., Ye, C., Zou, X., Zhang, Z., … Feng, T. (2020). Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: a retrospective cohort study. *The Lancet. Infectious Diseases*, *20*(8), 911–919. https://doi.org/10.1016/S1473-3099(20)30287-5

Brass, A. L., Huang, I.-C., Benita, Y., John, S. P., Krishnan, M. N., Feeley, E. M., … Elledge, S. J. (2009). The IFITM proteins mediate cellular resistance to influenza A H1N1 virus, West Nile virus, and dengue virus. *Cell*, *139*(7), 1243–1254. https://doi.org/10.1016/j.cell.2009.12.017

Carrat, F., Vergu, E., Ferguson, N. M., Lemaitre, M., Cauchemez, S., Leach, S., & Valleron, A.-J. (2008). Time Lines of Infection and Disease in Human Influenza: A Review of Volunteer Challenge Studies. *American Journal of Epidemiology*, *167*(7), 775–785. https://doi.org/10.1093/aje/kwm375

Carvalho, A., Cunha, C., Bozza, S., Moretti, S., Massi-Benedetti, C., Bistoni, F., … Romani, L. (2012). Immunity and Tolerance to Fungi in Hematopoietic Transplantation: Principles and Perspectives . *Frontiers in Immunology* , Vol. 3, p. 156. Retrieved from https://www.frontiersin.org/article/10.3389/fimmu.2012.00156

Castelli, V., Cimini, A., & Ferri, C. (2020). Cytokine Storm in COVID-19: “When You Come Out of the Storm, You Won’t Be the Same Person Who Walked in.” *Frontiers in Immunology*, *11*, 2132. https://doi.org/10.3389/fimmu.2020.02132

Channappanavar, R., Fehr, A. R., Vijay, R., Mack, M., Zhao, J., Meyerholz, D. K., & Perlman, S. (2016). Dysregulated Type I Interferon and Inflammatory Monocyte-Macrophage Responses Cause Lethal Pneumonia in SARS-CoV-Infected Mice. *Cell Host & Microbe*, *19*(2), 181–193. https://doi.org/10.1016/j.chom.2016.01.007

Channappanavar, R., Fehr, A. R., Zheng, J., Wohlford-Lenane, C., Abrahante, J. E., Mack, M., … Perlman, S. (2019). IFN-I response timing relative to virus replication determines MERS coronavirus infection outcomes. *The Journal of Clinical Investigation*, *129*(9), 3625–3639. https://doi.org/10.1172/JCI126363

Chen, R., Lan, Z., Ye, J., Pang, L., Liu, Y., Wu, W., … Zhang, P. (2021). Cytokine Storm: The Primary Determinant for the Pathophysiological Evolution of COVID-19 Deterioration . *Frontiers in Immunology* , Vol. 12, p. 1409. Retrieved from https://www.frontiersin.org/article/10.3389/fimmu.2021.589095

Costela-Ruiz, V. J., Illescas-Montes, R., Puerta-Puerta, J. M., Ruiz, C., & Melguizo-Rodríguez, L. (2020). SARS-CoV-2 infection: The role of cytokines in COVID-19 disease. *Cytokine and Growth Factor Reviews*, *54*(June), 62–75. https://doi.org/10.1016/j.cytogfr.2020.06.001

da Costa, V. G., Saivish, M. V., Santos, D. E. R., de Lima Silva, R. F., & Moreli, M. L. (2020). Comparative epidemiology between the 2009 H1N1 influenza and COVID-19 pandemics. *Journal of Infection and Public Health*, *13*(12), 1797–1804. https://doi.org/10.1016/j.jiph.2020.09.023

Desai, N., Neyaz, A., Szabolcs, A., Shih, A. R., Chen, J. H., Thapar, V., … Deshpande, V. (2020). Temporal and spatial heterogeneity of host response to SARS-CoV-2 pulmonary infection. *Nature Communications*, *11*(1), 6319. https://doi.org/10.1038/s41467-020-20139-7

Enright, A. J., Van Dongen, S., & Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, *30*(7), 1575–1584. https://doi.org/10.1093/nar/30.7.1575

Equils, O., Lekaj, K., Fattani, S., Wu, A., & Liu, G. (2020). Proposed mechanism for anosmia during COVID-19: the role of local zinc distribution. *J Transl Sci*, *7*, 1–2.

Fajgenbaum, D. C., & June, C. H. (2020). Cytokine Storm. *New England Journal of Medicine*, *383*(23), 2255–2273. https://doi.org/10.1056/NEJMra2026131

Falcon, S., & Gentleman, R. (2008). Hypergeometric Testing Used for Gene Set Enrichment Analysis. In *Bioconductor Case Studies. Use R!* https://doi.org/https://doi.org/10.1007/978-0-387-77240-0_14

Finamore, A., Massimi, M., Conti Devirgiliis, L., & Mengheri, E. (2008). Zinc Deficiency Induces Membrane Barrier Damage and Increases Neutrophil Transmigration in Caco-2 Cells. *The Journal of Nutrition*, *138*(9), 1664–1670. https://doi.org/10.1093/jn/138.9.1664

Flerlage, T., Boyd, D. F., Meliopoulos, V., Thomas, P. G., & Schultz-Cherry, S. (2021). Influenza virus and SARS-CoV-2: pathogenesis and host responses in the respiratory tract. *Nature Reviews Microbiology*, *19*(7), 425–441. https://doi.org/10.1038/s41579-021-00542-7

Fodil, S., & Annane, D. (2021). Complement Inhibition and COVID-19: The Story so Far. *ImmunoTargets and Therapy*, *10*, 273–284. https://doi.org/10.2147/ITT.S284830

Franceschini, A. (2013). *STRINGdb Package Vignette*. (July), 1–13.

Ge, S. X., Son, E. W., & Yao, R. (2018). iDEP: an integrated web application for differential expression and pathway analysis of RNA-Seq data. *BMC Bioinformatics*, *19*(1), 534. https://doi.org/10.1186/s12859-018-2486-6

Ghoshal, K., Majumder, S., Zhu, Q., Hunzeker, J., Datta, J., Shah, M., … Jacob, S. T. (2001). Influenza virus infection induces metallothionein gene expression in the mouse liver and lung by overlapping but distinct molecular mechanisms. *Molecular and Cellular Biology*, *21*(24), 8301–8317. https://doi.org/10.1128/MCB.21.24.8301-8317.2001

Goksuluk, D., Zararsiz, G., Korkmaz, S., Eldem, V., Zararsiz, G. E., Ozcetin, E., … Karaagaoglu, A. E. (2019). MLSeq: Machine learning interface for RNA-sequencing data. *Computer Methods and Programs in Biomedicine*, *175*, 223–231. https://doi.org/https://doi.org/10.1016/j.cmpb.2019.04.007

Gupta, R. K., Rosenheim, J., Bell, L. C., Chandran, A., Guerra-Assuncao, J. A., Pollara, G., … Zahedi, D. (2021). Blood transcriptional biomarkers of acute viral infection for detection of pre-symptomatic SARS-CoV-2 infection: a nested, case-control diagnostic accuracy study. *The Lancet Microbe*, *2*(10), e508–e517. https://doi.org/10.1016/S2666-5247(21)00146-4

Hadjadj, J., Yatim, N., Barnabei, L., Corneau, A., Boussier, J., Smith, N., … Terrier, B. (2020). Impaired type I interferon activity and inflammatory responses in severe COVID-19 patients. *Science (New York, N.Y.)*, *369*(6504), 718–724. https://doi.org/10.1126/science.abc6027

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *SIGKDD Explor. Newsl.*, *11*(1), 10–18. https://doi.org/10.1145/1656274.1656278

He, X., Lau, E. H. Y., Wu, P., Deng, X., Wang, J., Hao, X., … Leung, G. M. (2020). Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nature Medicine*, *26*(5), 672–675. https://doi.org/10.1038/s41591-020-0869-5

Heard, N. A., & Rubin-Delanchy, P. (2018). Choosing between methods of combining p-values. *Biometrika*, *105*(1), 239–246. https://doi.org/10.1093/biomet/asx076

Hu, B., Guo, H., Zhou, P., & Shi, Z.-L. (2021). Characteristics of SARS-CoV-2 and COVID-19. *Nature Reviews Microbiology*, *19*(3), 141–154. https://doi.org/10.1038/s41579-020-00459-7

Huang, I.-C., Bailey, C. C., Weyer, J. L., Radoshitzky, S. R., Becker, M. M., Chiang, J. J., … Farzan, M. (2011). Distinct patterns of IFITM-mediated restriction of filoviruses, SARS coronavirus, and influenza A virus. *PLoS Pathogens*, *7*(1), e1001258–e1001258. https://doi.org/10.1371/journal.ppat.1001258

Itsui, Y., Sakamoto, N., Kurosaki, M., Kanazawa, N., Tanabe, Y., Koyama, T., … Watanabe, M. (2006). Expressional screening of interferon-stimulated genes for antiviral activity against hepatitis C virus replication. *Journal of Viral Hepatitis*, *13*(10), 690–700. https://doi.org/https://doi.org/10.1111/j.1365-2893.2006.00732.x

Katze, M. G., He, Y., & Gale, M. (2002). Viruses and interferon: a fight for supremacy. *Nature Reviews Immunology*, *2*(9), 675–687. https://doi.org/10.1038/nri888

Kim, S.-Y., & Volsky, D. J. (2005). PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics*, *6*, 144. https://doi.org/10.1186/1471-2105-6-144

Kim, Y.-M., & Shin, E.-C. (2021). Type I and III interferon responses in SARS-CoV-2 infection. *Experimental & Molecular Medicine*, *53*(5), 750–760. https://doi.org/10.1038/s12276-021-00592-0

Korotkevich, G., Sukhov, V., Budin, N., Shpak, B., Artyomov, M. N., & Sergushichev, A. (2021). Fast gene set enrichment analysis. *BioRxiv*, 60012. https://doi.org/10.1101/060012

Lavine, J. S., Bjornstad, O. N., & Antia, R. (2021). Immunological characteristics govern the transition of COVID-19 to endemicity. *Science*, *371*(6530), 741–745. https://doi.org/10.1126/science.abe6522

Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, *15*(2), R29. https://doi.org/10.1186/gb-2014-15-2-r29

Lee, N., Chan, P. K. S., Hui, D. S. C., Rainer, T. H., Wong, E., Choi, K.-W., … Sung, J. J. Y. (2009). Viral loads and duration of viral shedding in adult patients hospitalized with influenza. *The Journal of Infectious Diseases*, *200*(4), 492–500. https://doi.org/10.1086/600383

Lei, X., Dong, X., Ma, R., Wang, W., Xiao, X., Tian, Z., … Wang, J. (2020). Activation and evasion of type I interferon responses by SARS-CoV-2. *Nature Communications*, *11*(1), 3810. https://doi.org/10.1038/s41467-020-17665-9

Li, J., Wang, S., Chen, Z., & Wang, Y. (2020). A Bipartite Network Module-Based Project to Predict Pathogen–Host Association . *Frontiers in Genetics* , Vol. 10, p. 1357. Retrieved from https://www.frontiersin.org/article/10.3389/fgene.2019.01357

Lokugamage, K. G., Hage, A., de Vries, M., Valero-Jimenez, A. M., Schindewolf, C., Dittmann, M., … Menachery, V. D. (2020). Type I interferon susceptibility distinguishes SARS-CoV-2 from SARS-CoV. *BioRxiv*, 2020.03.07.982264. https://doi.org/10.1101/2020.03.07.982264

López de Padilla, C. M., & Niewold, T. B. (2016). The type I interferons: Basic concepts and clinical relevance in immune-mediated inflammatory diseases. *Gene*, *576*(1 Pt 1), 14–21. https://doi.org/10.1016/j.gene.2015.09.058

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Lu, J., Pan, Q., Rong, L., He, W., Liu, S.-L., & Liang, C. (2011). The IFITM proteins inhibit HIV-1 infection. *Journal of Virology*, *85*(5), 2126–2137. https://doi.org/10.1128/JVI.01531-10

Lundstrom, K. (2016). Replicon RNA Viral Vectors as Vaccines. *Vaccines*, *4*(4), 39. https://doi.org/10.3390/vaccines4040039

Luo, J., Schumacher, M., Scherer, A., Sanoudou, D., Megherbi, D., Davison, T., … Zhang, J. (2010). A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. *The Pharmacogenomics Journal*, *10*(4), 278–291. https://doi.org/10.1038/tpj.2010.57

Mangalmurti, N., & Hunter, C. A. (2020). Cytokine Storms: Understanding COVID-19. *Immunity*, *53*(1), 19–25. https://doi.org/10.1016/j.immuni.2020.06.017

Mantlo, E., Bukreyeva, N., Maruyama, J., Paessler, S., & Huang, C. (2020). Antiviral activities of type I interferons to SARS-CoV-2 infection. *Antiviral Research*, *179*, 104811. https://doi.org/10.1016/j.antiviral.2020.104811

Mayor-Ibarguren, A., Busca-Arenzana, C., & Robles-Marhuenda, Á. (2020). A Hypothesis for the Possible Role of Zinc in the Immunological Pathways Related to COVID-19 Infection. *Frontiers in Immunology*, *11*, 1736. https://doi.org/10.3389/fimmu.2020.01736

McClain, M. T., Constantine, F. J., Henao, R., Liu, Y., Tsalik, E. L., Burke, T. W., … Woods, C. W. (2021). Dysregulated transcriptional responses to SARS-CoV-2 in the periphery. *Nature Communications*, *12*(1), 1079. https://doi.org/10.1038/s41467-021-21289-y

McDermaid, A., Monier, B., Zhao, J., Liu, B., & Ma, Q. (2019). Interpretation of differential gene expression results of RNA-seq data: review and integration. *Briefings in Bioinformatics*, *20*(6), 2044–2054. https://doi.org/10.1093/bib/bby067

McNab, F., Mayer-Barber, K., Sher, A., Wack, A., & O'Garra, A. (2015). Type I interferons in infectious disease. *Nature Reviews Immunology*, *15*(2), 87–103. https://doi.org/10.1038/nri3787

Memish, Z. A., Perlman, S., Van Kerkhove, M. D., & Zumla, A. (2020). Middle East respiratory syndrome. *Lancet (London, England)*, *395*(10229), 1063–1077. https://doi.org/10.1016/S0140-6736(19)33221-0

Min, Y.-Q., Huang, M., Sun, X., Deng, F., Wang, H., & Ning, Y.-J. (2021). Immune evasion of SARS-CoV-2 from interferon antiviral system. *Computational and Structural Biotechnology Journal*, *19*, 4217–4225. https://doi.org/10.1016/j.csbj.2021.07.023

Mistry, H. B., Davies, M. R., & Di Veroli, G. Y. (2015). A new classifier-based strategy for in-silico ion-channel cardiac drug safety assessment . *Frontiers in Pharmacology* , Vol. 6, p. 59. Retrieved from https://www.frontiersin.org/article/10.3389/fphar.2015.00059

Mitchell, H. D., Eisfeld, A. J., Sims, A. C., McDermott, J. E., Matzke, M. M., Webb-Robertson, B.-J. M., … Waters, K. M. (2013). A network integration approach to predict conserved regulators related to pathogenicity of influenza and SARS-CoV respiratory viruses. *PloS One*, *8*(7), e69374–e69374. https://doi.org/10.1371/journal.pone.0069374

Moore, J., & June, C. (2020). Cytokine release syndrome in severe COVID-19. *Science*, *368*(6490), 473–474. https://doi.org/10.1126/science.abb8925

Morrison, J., & Plotkin, S. (2016). *Chapter 19 - Viral Vaccines: Fighting Viruses with Vaccines* (M. G. Katze, M. J. Korth, G. L. Law, & N. B. T.-V. P. (Third E. Nathanson, Eds.). https://doi.org/https://doi.org/10.1016/B978-0-12-800964-2.00019-7

Mudd, P. A., Crawford, J. C., Turner, J. S., Souquette, A., Reynolds, D., Bender, D., … Ellebedy, A. H. (2020). Distinct inflammatory profiles distinguish COVID-19 from influenza with limited contributions from cytokine storm. *Science Advances*, *6*(50), eabe3024. https://doi.org/10.1126/sciadv.abe3024

Oran, D. P., & Topol, E. J. (2020). Prevalence of Asymptomatic SARS-CoV-2 Infection. *Annals of Internal Medicine*, *173*(5), 362–367. https://doi.org/10.7326/M20-3012

Oran, D. P., & Topol, E. J. (2021). The Proportion of SARS-CoV-2 Infections That Are Asymptomatic : A Systematic Review. *Annals of Internal Medicine*, *174*(5), 655–662. https://doi.org/10.7326/M20-6976

Peiris, J. S. M., Lai, S. T., Poon, L. L. M., Guan, Y., Yam, L. Y. C., Lim, W., … group, S. study. (2003). Coronavirus as a possible cause of severe acute respiratory syndrome. *Lancet (London, England)*, *361*(9366), 1319–1325. https://doi.org/10.1016/s0140-6736(03)13077-2

Phipson, B., Lee, S., Majewski, I. J., Alexander, W. S., & Smyth, G. K. (2016). ROBUST HYPERPARAMETER ESTIMATION PROTECTS AGAINST HYPERVARIABLE GENES AND IMPROVES POWER TO DETECT DIFFERENTIAL EXPRESSION. *The Annals of Applied Statistics*, *10*(2), 946–963. https://doi.org/10.1214/16-AOAS920

Pickles, R. J. (2013). Human airway epithelial cell cultures for modeling respiratory syncytial virus infection. *Current Topics in Microbiology and Immunology*, *372*, 371–387. https://doi.org/10.1007/978-3-642-38919-1_19

Propper, R. E. (2021). Smell/Taste alteration in COVID-19 may reflect zinc deficiency. *Journal of Clinical Biochemistry and Nutrition*, *68*(1), 3. https://doi.org/10.3164/jcbn.20-177

Qu, J., Zhao, Y., & Yin, J. (2019). Identification and Analysis of Human Microbe-Disease Associations by Matrix Decomposition and Label Propagation . *Frontiers in Microbiology* , Vol. 10, p. 291. Retrieved from https://www.frontiersin.org/article/10.3389/fmicb.2019.00291

Rabaan, A. A., Al-Ahmed, S. H., Muhammad, J., Khan, A., Sule, A. A., Tirupathi, R., … Dhama, K. (2021). Role of inflammatory cytokines in covid-19 patients: A review on molecular mechanisms, immune functions, immunopathology and immunomodulatory drugs to counter cytokine storm. *Vaccines*, *9*(5). https://doi.org/10.3390/vaccines9050436

Read, S. A., O'Connor, K. S., Suppiah, V., Ahlenstiel, C. L. E., Obeid, S., Cook, K. M., … Ahlenstiel, G. (2017). Zinc is a potent and specific inhibitor of IFN-λ3 signalling. *Nature Communications*, *8*, 15245. https://doi.org/10.1038/ncomms15245

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, *26*(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

Russo, C., Morello, G., Malaguarnera, R., Piro, S., Furno, D. Lo, & Malaguarnera, L. (2021). Candidate genes of SARS-CoV-2 gender susceptibility. *Scientific Reports*, *11*(1), 21968. https://doi.org/10.1038/s41598-021-01131-7

Ruttkay-Nedecky, B., Nejdl, L., Gumulec, J., Zitka, O., Masarik, M., Eckschlager, T., … Kizek, R. (2013). The role of metallothionein in oxidative stress. *International Journal of Molecular Sciences*, *14*(3), 6044–6066. https://doi.org/10.3390/ijms14036044

S Banach, B., Orenstein, J. M., Fox, L. M., Randell, S. H., Rowley, A. H., & Baker, S. C. (2009). Human airway epithelial cell culture to identify new respiratory viruses: coronavirus NL63 as a model. *Journal of Virological Methods*, *156*(1–2), 19–26. https://doi.org/10.1016/j.jviromet.2008.10.022

Sa Ribero, M., Jouvenet, N., Dreux, M., & Nisole, S. (2020). Interplay between SARS-CoV-2 and the type I interferon response. *PLOS Pathogens*, *16*(7), e1008737. Retrieved from https://doi.org/10.1371/journal.ppat.1008737

Schneider, D. S., & Ayres, J. S. (2008). Two ways to survive infection: what resistance and tolerance can teach us about treating infectious diseases. *Nature Reviews Immunology*, *8*(11), 889–895. https://doi.org/10.1038/nri2432

Schneider, W. M., Chevillotte, M. D., & Rice, C. M. (2014). Interferon-stimulated genes: a complex web of host defenses. *Annual Review of Immunology*, *32*, 513–545. https://doi.org/10.1146/annurev-immunol-032713-120231

Schoggins, J. W., Wilson, S. J., Panis, M., Murphy, M. Y., Jones, C. T., Bieniasz, P., & Rice, C. M. (2011). A diverse range of gene products are effectors of the type I interferon antiviral response. *Nature*, *472*(7344), 481–485. https://doi.org/10.1038/nature09907

Sims, A. C., Tilton, S. C., Menachery, V. D., Gralinski, L. E., Schäfer, A., Matzke, M. M., … Baric, R. S. (2013). Release of severe acute respiratory syndrome coronavirus nuclear import block enhances host transcription in human lung cells. *Journal of Virology*, *87*(7), 3885–3902. https://doi.org/10.1128/JVI.02520-12

Smyth, G. K. (2005). Limma: linear models for microarray data. In Bioinformatics and computational biology solutions using R and Bioconductor. *Edited by Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W*, New York: Springer, 397-420.

Smyth, G. K., Ritchie, M., & Thorne, N. (2011). Linear Models for Microarray Data User ' s Guide. *Bioinformatics*, *20*(May), 3705–3706.

Swadling, L., Diniz, M. O., Schmidt, N. M., Amin, O. E., Chandran, A., Shaw, E., … investigators, Covid. (2021). Pre-existing polymerase-specific T cells expand in abortive seronegative SARS-CoV-2. *Nature*. https://doi.org/10.1038/s41586-021-04186-8

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., … Mering, C. von. (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, *47*(D1), D607–D613. https://doi.org/10.1093/nar/gky1131

Tan, X., Sun, L., Chen, J., & Chen, Z. J. (2015). Detection of Microbial Infections Through Innate Immune Sensing of Nucleic Acids. *Annual Review of Microbiology*, *72*(1), 447–478. https://doi.org/10.1146/annurev-micro-102215-095605

Tang, Y., Liu, J., Zhang, D., Xu, Z., Ji, J., & Wen, C. (2020). Cytokine Storm in COVID-19: The Current Evidence and Treatment Strategies. *Frontiers in Immunology*, *11*(July), 1–13. https://doi.org/10.3389/fimmu.2020.01708

Thoms, M., Buschauer, R., Ameismeier, M., Koepke, L., Denk, T., Hirschenberger, M., … Beckmann, R. (2020). Structural basis for translational shutdown and immune evasion by the Nsp1 protein of SARS-CoV-2. *Science (New York, N.Y.)*, *369*(6508), 1249–1255. https://doi.org/10.1126/science.abc8665

Thomson, T. M., Toscano-Guerra, E., Casis, E., & Paciucci, R. (2020). C1 esterase inhibitor and the contact system in COVID-19. *British Journal of Haematology*, *190*(4), 520–524. https://doi.org/10.1111/bjh.16938

Tian, W., Zhang, N., Jin, R., Feng, Y., Wang, S., Gao, S., … Wong, C. C. L. (2020). Immune suppression in the early stage of COVID-19 disease. *Nature Communications*, *11*(1), 5859. https://doi.org/10.1038/s41467-020-19706-9

Totura, A. L., & Baric, R. S. (2012). SARS coronavirus pathogenesis: host innate immune responses and viral antagonism of interferon. *Current Opinion in Virology*, *2*(3), 264–275. https://doi.org/10.1016/j.coviro.2012.04.004

Toulis, P. (2021). Estimation of Covid-19 prevalence from serology tests: A partial identification approach. *Journal of Econometrics*, *220*(1), 193–213. https://doi.org/10.1016/j.jeconom.2020.10.005

Tseng, C.-T. K., Tseng, J., Perrone, L., Worthy, M., Popov, V., & Peters, C. J. (2005). Apical entry and release of severe acute respiratory syndrome-associated coronavirus in polarized Calu-3 lung epithelial cells. *Journal of Virology*, *79*(15), 9470–9479. https://doi.org/10.1128/JVI.79.15.9470-9479.2005

V'kovski, P., Kratzel, A., Steiner, S., Stalder, H., & Thiel, V. (2021). Coronavirus biology and replication: implications for SARS-CoV-2. *Nature Reviews Microbiology*, *19*(3), 155–170. https://doi.org/10.1038/s41579-020-00468-6

Vabret, N., Britton, G. J., Gruber, C., Hegde, S., Kim, J., Kuksin, M., … Project, S. I. R. (2020). Immunology of COVID-19: Current State of the Science. *Immunity*, *52*(6), 910–941. https://doi.org/10.1016/j.immuni.2020.05.002

Vastrad, B., Vastrad, C., & Tengli, A. (2020). Bioinformatics analyses of significant genes, related pathways, and candidate diagnostic biomarkers and molecular targets in SARS-CoV-2/COVID-19. *Gene Reports*, *21*, 100956. https://doi.org/10.1016/j.genrep.2020.100956

Wessels, I., Maywald, M., & Rink, L. (2017). Zinc as a Gatekeeper of Immune Function. *Nutrients*, *9*(12), 1286. https://doi.org/10.3390/nu9121286

Witten, D. M. (2011). Classification and clustering of sequencing data using a Poisson model. *The Annals of Applied Statistics*, *5*(4), 2493–2518. https://doi.org/10.1214/11-AOAS493

Wong, T.-T. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition*, *48*(9), 2839–2846. https://doi.org/https://doi.org/10.1016/j.patcog.2015.03.009

Wynants, L., Van Calster, B., Collins, G. S., Riley, R. D., Heinze, G., Schuit, E., … van Smeden, M. (2020). Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ (Clinical Research Ed.)*, *369*, m1328–m1328. https://doi.org/10.1136/bmj.m1328

Yan, D., Zhang, X., Chen, C., Jiang, D., Liu, X., Zhou, Y., … Yang, S. (2021). Characteristics of Viral Shedding Time in SARS-CoV-2 Infections: A Systematic Review and Meta-Analysis. *Frontiers in Public Health*, *9*, 652842. https://doi.org/10.3389/fpubh.2021.652842

Zararsiz, G., Goksuluk, D., Klaus, B., Korkmaz, S., Eldem, V., Karabulut, E., & Ozturk, A. (2017). voomDDA: Discovery of diagnostic biomarkers and classification of RNA-seq data. *PeerJ*, *2017*(10), 1–27. https://doi.org/10.7717/peerj.3890

Zhang, H., Chen, H., Zhang, J., Chen, X., Guo, B., Zhi, P., … Lu, X. (2021). Bioinformatics analysis of SARS-CoV-2 infection-associated immune injury and therapeutic prediction for COVID-19. *Emergency and Critical Care Medicine*, *1*(1). Retrieved from https://journals.lww.com/eccm/Fulltext/2021/09000/Bioinformatics_analysis_of_SARS_CoV_2.6.aspx

Zhang, Q., Bastard, P., Bolze, A., Jouanguy, E., Zhang, S.-Y., Effort, C. H. G., … Casanova, J.-L. (2020). Life-Threatening COVID-19: Defective Interferons Unleash Excessive Inflammation. *Med (New York, N.Y.)*, *1*(1), 14–20. https://doi.org/10.1016/j.medj.2020.12.001

Zhang, Q., Bastard, P., Liu, Z., Le Pen, J., Moncada-Velez, M., Chen, J., … Casanova, J.-L. (2020). Inborn errors of type I IFN immunity in patients with life-threatening COVID-19. *Science (New York, N.Y.)*, *370*(6515), eabd4570. https://doi.org/10.1126/science.abd4570

Zhao, X., Guo, F., Liu, F., Cuconati, A., Chang, J., Block, T. M., & Guo, J.-T. (2014). Interferon induction of IFITM proteins promotes infection by human coronavirus OC43. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(18), 6756–6761. https://doi.org/10.1073/pnas.1320856111

Zhao, X., Sehgal, M., Hou, Z., Cheng, J., Shu, S., Wu, S., … Guo, J.-T. (2018). Identification of Residues Controlling Restriction versus Enhancing Activities of IFITM Proteins on Entry of Human Coronaviruses. *Journal of Virology*, *92*(6), e01535-17. https://doi.org/10.1128/JVI.01535-17

Zhu, Y., Chidekel, A., & Shaffer, T. H. (2010). Cultured Human Airway Epithelial Cells (Calu-3): A Model of Human Respiratory Function, Structure, and Inflammatory Responses. *Critical Care Research and Practice*, *2010*, 394578. https://doi.org/10.1155/2010/394578

Zhu, Z., Lian, X., Su, X., Wu, W., Marraro, G. A., & Zeng, Y. (2020). From SARS and MERS to COVID-19: a brief summary and comparison of severe acute respiratory infections caused by three highly pathogenic human coronaviruses. *Respiratory Research*, *21*(1), 224. https://doi.org/10.1186/s12931-020-01479-w

Ziegler, C. G. K., Allon, S. J., Nyquist, S. K., Mbano, I. M., Miao, V. N., Tzouanas, C. N., … Network, H. C. A. L. B. (2020). SARS-CoV-2 Receptor ACE2 Is an Interferon-Stimulated Gene in Human Airway Epithelial Cells and Is Detected in Specific Cell Subsets across Tissues. *Cell*, *181*(5), 1016-1035.e19. https://doi.org/10.1016/j.cell.2020.04.035

Appendix I – Common human cytokines

Interferons		
Reference Name	**Genes**	**Name**
IFN-I (type **I** Interferons)		
IFNα	IFNA1, IFNA2, IFNA4, IFNA5, IFNA6, IFNA7, IFNA8, IFNA10, IFNA13, IFNA14, IFNA16, IFNA17, IFNA21	interferon a
IFNβ	IFNB1	interferon beta 1
IFNω	IFNW1	interferon omega 1
IFNε	IFNE, IFNE1, IFNT1	interferon epsilon
IFNκ	IFNK, IFNK1, IFNT1, IFNE1	interferon kappa
IFN-II (type **II** Interferons)		
IFNγ	IFNG, IMD69, IFI	interferon g
IFN-III (type **III** Interferons)		
IFNλ	IFNL1, IFNL2, IFNL3, IFNL4	

Interleukins		
Reference Name	**Synonyms**	**Name**
IL-1α	**IL1A**, IL-1_alpha, IL-1A, IL1, IL1-ALPHA, IL1F1	interleukin 1 alpha
IL-1β	**IL1B**, IL-1, IL1-BETA, IL1F2, IL1beta	
IL-1RA	**IL1R1**, CD121A, D2S1473, IL-1R-alpha, IL1R, IL1RA, P80	
IL-18	**IL18**, IGIF, IL-18, IL-1g, IL1F4	
Common g chain (CD132)		
IL-2	**IL2**, TCGF, lymphokine, T cell growth factor	
IL-4	**IL4**, BCGF-1, BCGF1, BSF-1, BSF1	interleukin 4
IL-7	**IL7**	interleukin 7
IL-9	**IL9**, HP40, IL-9, P40	interleukin 9
IL-13	**IL13**, P600	interleukin 13
Il-15	**IL15**, AI503618	interleukin 15
Common b chain (CD131)		
IL-3	**IL3**, MCGF, MULTI-CSF	interleukin 3
IL-5	**IL5**, EDF, TRF	interleukin 5
Related		
GM-CSF	**CSF2**, CSF, GMCSF	colony stimulating factor 2
IL-6-like		
IL-6	**IL6**, BSF-2, BSF2, CDF, HGF, HSF, IFN-beta-2, IFNB2	interleukin 6
IL-11	**IL11**, AGIF	
G-CSF	**CSF3**, C17orf33, CSF3OS, GCSF	colony stimulating factor 3
IL-12A	**IL12A**, CLMF, NFSK, NKSF1, P35, NK cell stimulatory factor	interleukin 12A
IL-12B	**IL12B**, CLMF, CLMF2, IMD28, IMD29, NKSF, NKSF2	interleukin 12B
LIF	**LIF**, CDF, DIA, HILDA, MLPLI, leukemia inhibitory factor	LIF IL6 family cytokine
OSM		oncostatin M
IL-10	**IL10**, CSIF, GVHDS, IL10A, TGIF	interleukin 10
IL-20	**IL20**, IL10D, ZCYTO10	interleukin 20

Others		
IL-14	**IL14**, TXLNA, TXLN	taxin alpha
IL-16	**IL16**, LCF, FLJ42735, FLJ16806	interleukin 16
IL-17A	**IL17A**, CTLA8	interleukin 17A
IL-17B	**IL17B**, IL-20, NIRF, ZCYTO7	interleukin 17B

TNF		
CD154	**CD40LG**, CD40L, HIGM1, IGM, IMD3, T-BAM, TNFSF5, TRAP, gp39, hCD40L	CD40 ligand
LT-β	**LTB**, TNFC, TNFSF3, TNLG1C, p33	lymphotoxin beta
TNF-α	**TNF**, DIF, TNF-alpha, TNFA, TNFSF2, TNLG1F, cachectin	TNF
TNF-β	**LTA**, LT, TNFB, TNFSF1, TNLG1E	lymphotoxin alpha
4-1BBL	**TNFSF9**, CD137L, TNLG5A	TNF superfamily member 9
APRIL	**TNFSF13**, CD256, TALL-2, TALL2, TNLG7B, TRDL-1, UNQ383/PRO715, ZTNF2	TNF superfamily member 13
CD70	CD27-L, CD27L, CD27LG, LPFS3, TNFSF7, TNLG8A	TNF ligand superfamily member 7
CD153	**TNFSF8**, CD30L, CD30LG, TNLG3A	TNF ligand superfamily member 8
CD178	**FASLG**, ALPS1B, APT1LG1, APTL, CD95-L, CD95L, FASL, TNFSF6, TNLG1A	TNF ligand superfamily member 6
GITRL	**TNFSF18**, AITRL, TL6, TNLG2A, hGITRL	TNF ligand superfamily member 18
LIGHT	**TNFSF14**, CD258, HVEML, LTg	TNF ligand superfamily member 14
OX40L	**TNFSF4**, CD134L, CD252, GP34, OX-40L, TNLG2B, TXGP1	TNF ligand superfamily member 4

Chemokines		
Reference Name	**Synonyms**	**Receptor**
CCL1	I-309, P500, SCYA1, SISe, TCA3	CCR8
CCL2	GDCF-2, HC11, HSMCR30, MCAF, MCP-1, MCP1, SCYA2, SMC-CF	CCR2
CCL3	G0S19-1, LD78ALPHA, MIP-1-alpha, MIP1A, SCYA3	CCR1
CCL4	ACT2, AT744.1, G-26, HC21, LAG-1, LAG1, MIP-1-beta, MIP1B, MIP1B1, SCYA2, SCYA4	CCR1, CCR5
CCL5	RANTES, D17S136E, SCYA5, SIS-delta, SISd, TCP228, eoCP	CCR5
CCL7	FIC, MARC, MCP-3, MCP3, NC28, SCYA6, SCYA7	CCR2
CCL8	HC14, MCP-2, MCP2, SCYA10, SCYA8	CCR1, CCR2, CCR5
CCL11	SCYA11, Eotaxin	CCR2, CCR3, CCR5
CCL13	CKb10, MCP-4, NCC-1, NCC1, SCYA13, SCYL1	CCR2, CCR3, CCR5
CCL14	HCC-1, MCIF, Ckβ1, NCC-2, CCL	CCR1
CCL15	Leukotactin-1, HCC-2, HMRP-2B, LKN-1, LKN1, MIP-1_delta, MIP-1D, MIP-5, MRP-2B, NCC-3, NCC3, SCYA15, SCYL3, SY15	CCR1, CCR3
CCL16	CKb12, HCC-4, ILINCK, LCC-1, LEC, LMC, Mtn-1, NCC-4, NCC4, SCYA16, SCYL4	CCR1, CCR2, CCR5, CCR8
CCL17	A-152E5.3, ABCD-2, SCYA17, TARC , dendrokine	CCR4
CCL18	AMAC-1, AMAC1, CKb7, DC-CK1, DCCK1, MIP-4, PARC, SCYA18	
CCL19	ELC, Ckβ11, Exodus-3	CCR7

CCL20	CKb4, LARC, MIP-3-alpha, MIP-3a, MIP3A, SCYA20, ST38, Exodus-1	CCR6
CCL21	6Ckine, CKb9, ECL, SCYA21, SLC, TCA4, Exodus-2	CCR7
CCL22	A-152E5.1, ABCD-1, DC/B-CK, MDC, SCYA22, STCP-1	CCR4
CCL23	CK-BETA-8, CKb8, Ckb-8, Ckb-8-1, MIP-3, MIP3, MPIF-1, SCYA23, hmrp-2a	CCR1
CCL24	Ckb-6, MPIF-2, MPIF2, SCYA24, Eotaxin-2	CCR3
CCL25	Ck_beta-15, Ckb15, SCYA25, TECK	CCR9
CCL26	IMAC, MIP-4a, MIP-4alpha, SCYA26, TSC-1, Eotaxin-3	CCR3
CCL27	ALP, CTACK, CTAK, ESKINE, ILC, PESKY, SCYA27, Eskine, skinkine	CCR10
CCL28	CCK1, MEC, SCYA28	CCR3, CCR10

TGFB (transforming growth factor)		
TGF-β1	**TGFB1**, CED, DPD1, IBDIMDE, LAP, TGF-beta1, TGFB, TGFbeta	transforming growth factor beta 1
TGF-β2	**TGFB2**, G-TSF, LDS4, TGF-beta2	transforming growth factor beta 2
TGF-β3	**TGFB3**, ARVD, ARVD1, LDS5, RNHF, TGF-beta3	transforming growth factor beta 3