UPPSALA
UNIVERSITET

Integrative bioinformatic analysis of SARs-CoV-2 data

Mirela Balan

Degree project in bioinformatics, 2021
Examensarbete i bioinformatik 30 hp till masterexamen, 2021
Biology Education Centre
Supervisors: Prof. Dr. Jan Komorowski and Dr. Fredrik Barrenäs

Abstract

During the current pandemic, a global effort has been deployed to discover the mechanisms through which SARs-CoV-2 virus infects humans, and how to use them against the virus in medical therapy. In this thesis the attention has been focused on analyzing the immune response of non-human primates (NHP) to receiving a novel SARs-CoV-2 vaccine and then being challenged with the virus.

The way in which classical biology works is by finding what one molecule does by taking apart the system that it belongs to. However, a biological system is very complex, and studying just one piece in isolation might not give you an understanding of the system. This is the case when looking at how an organism responds to vaccination and to viral infections. For this reason, in this work I have used approaches from systems biology to be able to integrate all the information, see at a glance which areas are affected and to analyze the changes on various levels.

The tool used in this thesis (DINA) is created by extracting useful gene-gene interaction from data repositories such as GEO and then merging the information into an interconnected network. This organized framework is made up of connections that are preserved in a variety of diseases and conditions (in this case for the same tissue). This ‘universality’ but in a restrained, focused sense, is what increases the probability that these gene-gene interactions are actually meaningful from a biological perspective. The advantage of this approach is that it allows scientists to derive meaningful associations even if the dataset they have obtained from an experiment is too small (in a classic sense) to extract meaningful observations.

The experiments consisted of 12 macaques that have been vaccinated either with a SARs-CoV-2 vaccine or a zika vaccine as control, both formulated with the same delivery backbone. Every week blood was sampled to find the unique signature that identifies the monkeys protected from SARs-CoV-2. After five weeks, all the monkeys were infected with SARs-CoV-2. Blood was drawn every two days to follow the response of the organism to infection and to find the gene signature that characterized monkeys protected from SARs-CoVs2.

During the investigation of the novel vaccination and viral challenge dataset, I have found unusually high gene expression signatures for both the time points before and after the challenge. The broad signatures for the whole time series showed that overall, the NHP immune systems responded in similar ways to the new SARs-CoV-2 vaccine and to the zika vaccine that was used as control. This is due most likely to the fact that the vaccine used in delivering the genetic information to the cells has a sustained expression of viral mRNA for long periods of time to drive the antiviral inflammatory responses, and thus acting as its own boost.

This increased priming of the organism led to an equally strong response to the viral challenge, so strong in fact that the samples containing them cluster separately from the rest of the samples during unsupervised clustering.

Using DINA I could detect both the similarities in the gene expression for the two experimental conditions (upregulated biosynthetic processes, genes modulating entry into the host, leukocyte and

dendritic cell differentiation) but also the differences that set them apart (gene responsive to vitamin A in SARs-CoV-2, reduction in angiogenesis and neuronal differentiation in zika).

At the end, I have extracted the gene expression signature that characterize NHP protected from SARs-CoV-2: a cluster of gene with functions in antiviral innate immunity, most of them placed in the interferon response pathway. With their concerted action, IFI44L, IFIT2, IFIT3, PARP9, PARP12, DHX58 , DDX60L and FFAR2 orchestrate the innate immune response together and help the organism fight the infection.

Popular Science Summary

We need to understand more about COVID-19 to create better therapies

The topic of this thesis is the subject of intense debate in the public forums due to the severe disruption that it has caused both to the world economy and to everyone's lives. The breakneck speed at which the vaccines were created to protect us from infection with SARs-CoV-2 is something new for the scientists, more used to going through the rigorous and slow pace of basic research, where you go from breakthrough to vaccine within a decade.

However, even if we have created these vaccines, they can be seen at best as 'emergency solutions'. There are many things that have not been understood in the short time the scientists had since December 2019. We do know how the virus infects the host. We are very familiar with its manifestations in patients, ranging from asymptomatic, to mild flu-like features, and all the way to extreme shortness of breath due to tissue damage, comma and death. But how exactly does the body respond to this threat? And what is the difference between people that have no or mild symptoms and those who are permanently affected by it?

There is a wealth of information regarding the response of the human immune system to viral infection, so why is this so difficult to investigate? The immune system is a very complex structure, with many genes participating in its regulation by interacting with each other in many, mostly unknown ways. In the context of this multi-dimensional spider-like web of interactions, it is difficult to see which of these interactions are important enough to affect the outcome of an immune response to a virus. Ontop of that, the classical way to do research is to focus on one gene at a time and discover what it does, then other scientists expanding on that knowledge. But how can you quickly find the area of interest in this web of interactions? How do you get enough dynamic information to characterize changes brought by infection?

This is where genetic material sequencing and computational biology come together in a powerful way to characterize complex systems. The aim of this project is to study the immune viral response in vaccinated monkeys exposed to SARs-CoV-2, in the hope of finding those molecular levers that can specifically allow the development of better drugs. To this end I have turned to methods used in systems biology, where you understand the bigger picture by putting together all the known pieces of that system. So in this case, the first step is to map the immune system response upon SARs-CoV-2 infection, both for protected and non-protected animals. Afterwards, we are zooming out to see the whole web and identify key areas that differ between the two cases. Then we focus into those promising areas to find which genes or network of genes make a difference in this case.

Abbreviations

ACE2 - angiotensin-converting enzyme 2

CENs - co-expression networks

COVID-19 - coronavirus disease 2019

DE - differentially expressed (genes)

DINA - dynamic integrative network analysis

DPC – days post-challenge

GO - gene ontology

ID – identity

IFN - interferon

logFC – log-fold-changes

NHP - nonhuman primate

PPIs - protein-protein interaction networks

RM – rhesus macaque

RNA – ribonucleic acid

S - spike protein of the SARs-CoV-2 virus

SARS-CoV2 - severe acute respiratory syndrome corona-virus 2

WPV – weeks post-vaccination

Table of Contents

1. Introduction

Viral contagious diseases are a continuous threat for human existence. Usually contained to their animal host where they multiply sometimes without obvious signs of infection, sometimes they acquire mutations that allow them to switch hosts, and even adapt to living in humans. Often we are lucky that such animal-borne viruses die out quickly because they do not have the opportunity to infect many people at once. Either they appear in an isolated location, or they are not easily transmissible or they represent such a deadly variant that they kill the host too fast without spreading. Every now and then, a pathogen arises that manages to balance all these requirements, and on top of that takes advantage of global transportation to spread quickly to all the corners of the world.

This is the case currently with the severe acute respiratory syndrome corona-virus 2 (SARs-CoV-2) that has started the coronavirus disease 2019 (COVID-19) pandemic. The whole scientific world has started a concerted campaign to find the specific mechanisms through which this virus is transmitted, infects, multiplies, and then leads to the outcome of the disease. We have managed to create emergency vaccines to stop the spread of SARs-CoV-2 but there is space for improvement.

In the context of the current pandemic, my master project is an exploratory analysis, aimed primarily at studying the immune response of non-human primates (NHP) to a novel vaccine against SARs-CoV-2, using both traditional bioinformatic tools but also methods from systems biology.

1.1. The SARs-CoV-2 virus and COVID-19 epidemic

The culprit behind the current pandemic, SARS-CoV-2 is a virus that belongs to the Coronaviridae family, known to produce a wide range of respiratory conditions, from the mild common cold to the potentially deadly COVID-19 and Middle East Respiratory Syndrome. SARS-CoV-2, just like its predecessor SARS-CoV, is a single-stranded ribonucleic acid (RNA) virus enveloped in a membrane studded with spike proteins (Fig. 1) (Lai & Cavanagh 1997).

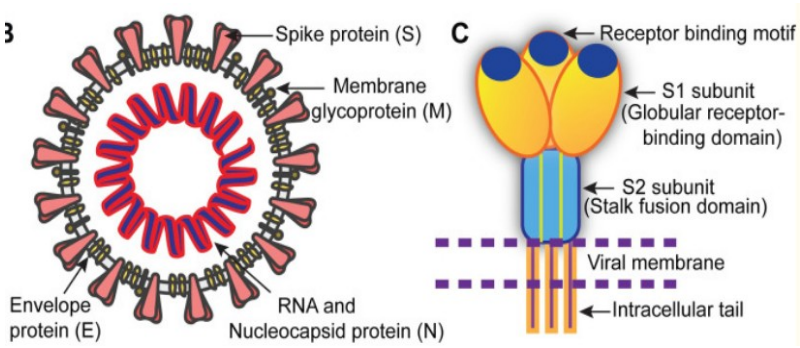

Figure 1: The internal and external structure of the SARs-CoV-2 virus, with a close-up of the structure of the S spike surface protein. Adapted from Mittal et al.

The S membrane protein is especially important because it is the main factor that mediates entry into the host cell. For SARs-CoV-2, the receptor is angiotensin-converting enzyme 2 (ACE2), (Ziegler *et al.* 2020). The distribution of the receptors within the body determines which tissues are

affected by the virus (tissue tropism). ACE is a receptor found on alveolar cells in the lung, but also in a variety of other tissues such as kidney and heart (Wan *et al.* 2020), but also the mucosa, one of the primary sites of infection (Xu *et al.* 2020). Also, depending on the eventual mutations of the S protein, the virus can become more contagious and/ or deadly, or it can completely change the type of host it can access for survival (host tropism). This structural specificity makes the S protein the perfect candidate for pattern recognition by the immune system. For this exact reason, the S protein has been used to create vaccines and as a bait to find small molecules in drug discovery against SARs-CoV-2.

Besides this dangerous possibilities, COVID-19 brings one more layer of complexity from the fact that it triggers a wide range of responses in the affected people, ranging from asymptomatic (but still contagious), to mild symptoms (flu-like symptoms such as fever, cough, difficulty breathing), and in some severe cases it can trigger uncontrollable immune reactions (cytokine storm), coma and death (Azkur *et al.* 2020). An unusual symptom of COVID-19 is the loss of smell or taste. Since no correlation has been so far found between severity of symptoms and disease development, and because there are so many people affected, there is no way to prioritize the people that would need intervention the most.

1.2. The immune response against viral infections

There are three types of responses against invading pathogens, depending on the type of invader the body has to fight. Type 1 immunity is specifically antiviral and it depends on the expression of interferon genes. The innate antiviral immune system is the first to respond to the threat, trying to reduce and to fight the infection until the adaptive immune response can ramp up and join in the battle.

A cell infected with a virus will produce interferon as a warning signal to attract the attention of the T cells. The destruction of the infected cell and the removal of any leftover viral particles is done via effector cells such as natural killer (NK) cells, cytotoxic T lymphocytes, T1 helper cells. These cells also release cytokines such as inteferon-g and tumour necrosis factor-a to increase the apoptotic effect.

Specifically to the immune response against SARs-CoV-2, the virus targets the cells in the lung and this creates an inflammatory environment that attracts lymphocytes such as macrophages and monocytes, which secrete cytokines to prime the body to fight. More immune cells are recruited as a response to the cytokine release, and the inflammatory environment is maintained, mediated by IL-1b, IL-6, TNF (Azkur *et al.* 2020). From this point, the body can resolve the infection and bring the body to homeostasis.

However, in certain cases, the safety measures of the immune system fail (such as the ineffective synthesis of anti-inflammatory molecules: IL-10, tumor growth factor beta (Rouse & Sehrawat 2010)), and the intense immune response triggers a strong inflammatory response, mediated among others by IL-2, IL-7 and tumour necrosis factor (TNF) (Tay *et al.* 2020). The systemic release of cytokines also known as ‘cytokine storm’, leads to progressive severe tissue damage and in the end to the death of the patient. The mechanisms that trigger this disregulation are not yet known for

SARs-CoV-2 but it is assumed that the virus can dampen the normal interferon response to viral infections (Narayanan *et al.* 2008).

Around week 1 after the onset of symptoms, the adaptive immune response starts to mount an attack against the virus. But since the experimental part of the thesis follows the immune response of the NHP for up to a week, the adaptive response does not fall within the scope of this thesis.

1.3. Vaccine design and delivery

Vaccination is the act of administering a compound that prepares the body's adaptive immune system to fight against a pathogen, and thus conferring protection to the subject. Over the years, this approach has eradicated several severe diseases, such as smallpox and polio.

There are several types of vaccines depending on how they are created to deliver the information to the immune system and for COVID-19 several strategies were employed (Fig. 2). The traditional method is to deliver live, weakened, inactivated pathogens or just some immunogenic parts of pathogens (such as surface proteins). Besides whole virus vaccines, other categories are represented by viral vector vaccines (non-replicating viral vector and replicating viral vector), nucleic acid-based vaccines (mRNA vaccine and DNA vaccine) and nanoparticle and virus-like particles vaccines (Flanagan *et al.* 2020).

A relatively new vaccine design is to deliver the necessary genetic information that uniquely corresponds to the virus (epitope) using a virus-derived replicon RNA system: the vector contains the open reading frame of a viral RNA polymerase complex that drives an antigen-encoding gene from the pathogen. Once in the body, the cellular machinery starts to actively synthesize the exogenous RNA (Vogel *et al.* 2018). This process mimics an ongoing viral infection, leading to the activation of the innate inflammatory pathways and thus to the production of interferon and other inflammatory molecules. This concerted reaction of the body to the vaccine acts in effect similarly to administering a booster dose (a second dose of vaccination to enhance the immune response). Thus the sustained RNA transcription means only one dose would be enough to get protection and this makes it ideal in a global pandemic.

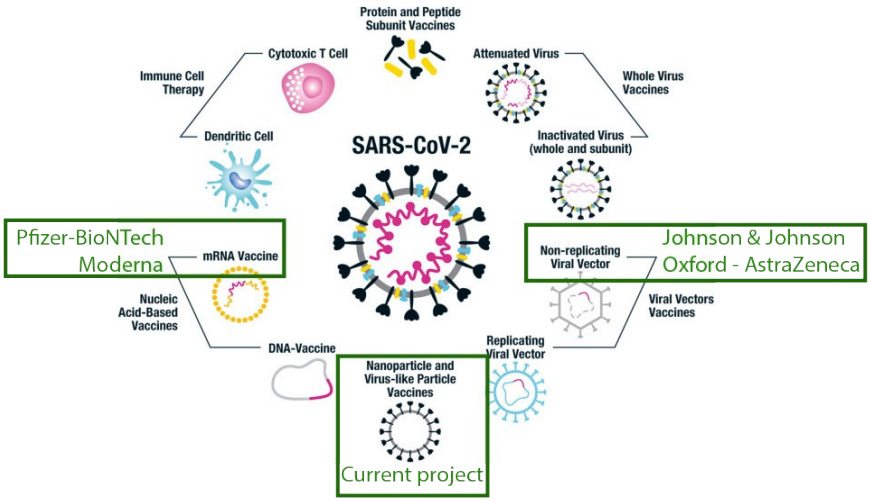

Figure 2: Types of platforms used to design vaccines for COVID-19. The four main vaccines that were approved for emergency use world wide are grouped in their respective category. Adapted from Flanagan et al, 2020

The vaccine for this project was constructed in the laboratory of Prof. Deborah Fuller at the University of Washington. The viral RNA polymerase complex came from the Alphavirus genus and it drives either the full-length spike protein of SARs-CoV-2 virus (Wuhan isolate, GenBank: MN908947.3) (Erasmus *et al.* 2020) or the pre-membrane and envelope coding sequence (prME) of the zika virus (Erasmus *et al.* 2018) for the mock vaccine.

1.4. Systems biology approach to characterize SARS-CoV-2 responses

1.4.1. Protein-protein interactions and co-expression networks

In order to accomplish their functions, proteins act together with other molecules, arranged in either transient or stable supramolecular structures. While one of the molecules might perform the main function of the molecular complex, the rest of them are just as important. The subcomponents either regulate the function or behave as docking sites for various substrates that the molecular complex acts upon (Rao *et al.* 2014). Protein-protein interactions (PPIs) refer to mapped physical protein interaction that can be determined experimentally (Fig. 3).

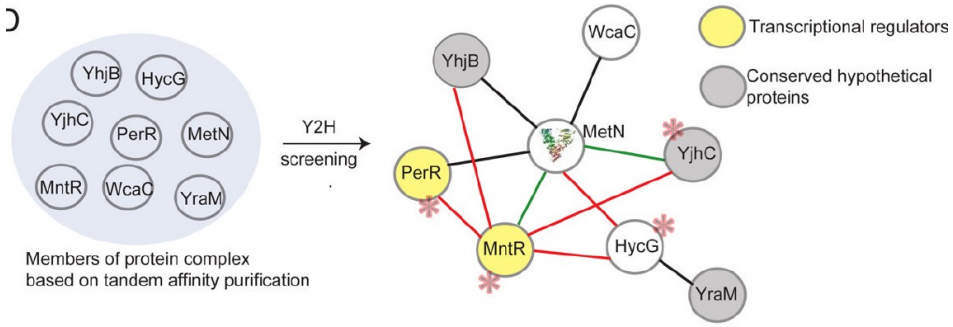

Figure 3: Example of creating a protein-protein interaction network from purified proteic complexes. Adapted from Rajagagopala et. al.

This method has several disadvantages, one of them starting from the very techniques used to obtain the information. The yeast two-hybrid experimental disadvantage is that the system can't perform all the functions of a mammalian system (such as post-translational modifications that can only occur in mammalian cells) and leads to high false negative and false positive rates (Rajagopala *et al.* 2012). Mass spectrometry experiments give better results because it incorporates an affinity purification step and because these cells express the protein of interest at physiological levels. In yeast two-hybrid, the protein can be over-expressed and sometimes the modification of the molecule itself can change its behavior. The biggest disadvantage is that PPIs can be made only from known information.

On the other hand, co-expression networks (CENs) can be implemented on systems/ organisms that lack information about direct protein-protein interaction (Vella *et al.* 2017). The methodology is based on statistical inference of the dependency between several variables, such as gene/ protein expression. It can be assumed that the observed deviations during system perturbation are a direct consequence of the intervention and may indicate active regulation of internal processes.

In the current thesis, I am studying the differences between two different vaccination interventions (COVID-19 and zika). It is plausible to assume that the differences I observe in gene expression might point to the genes that play a specific role in triggering protection from SARs-CoV-2. The genes that are statistically linked and placed together within a sub-network can be assumed to be influenced by at least one of the other genes, usually the central gene (the hub - it is the gene with the highest number of connections)

In this setup, I have used zika vaccine as a control to the SARs-CoV-2 vaccine. It was chosen because it is a vaccine designed for another viral disease and it has the exact same formulation, but it could have been replaced by any other vaccine that fulfilled these requirements.

This method too has several disadvantages such as a high number of false positive connections, especially in dense networks. And since it relies on data gathered from experiments, it is limited by the quality and availability of such data.

1.4.2. Dynamic Integrative Network Analysis

As it is obvious from the disadvantages discussed in the previous section, we would need a method that can be used in novel systems where pre-existing knowledge is scarce or in well characterized systems, where there is always space to discover novel and important interactions. The dynamic integrative network analysis (DINA) uses big data approaches to overcome the limitations faced by PPIs and CENs, and it was introduced by Dr. Fredrik Barrenäs from Uppsala University
(Barrenas *et al.* 2019) . It utilizes the wealth of publicly available transcriptomic data to identify gene modules with very strong statistical support, and uses gene expression signatures in the public datasets to assign biological functions to each module.

The main advantage is that it provides stronger statistical power to any transcriptomic or similar study, especially if those experiments are limited, such as in the case of a rapidly expanding outbreak like the COVID-19 pandemic.

The interactions between the genes are represented as a tree-and-leaf network for a much clearer overview of interactions. Each branch corresponds to specific biological processes (such as protein modification, or lung development), while the neighboring modules on the same main branch represent related modules.

The network represented in figure 4 was created using relevant and freely available datasets, and represent the most commonly found interactions in blood samples collected from various diseases. More details about its construction are in section 2.5.

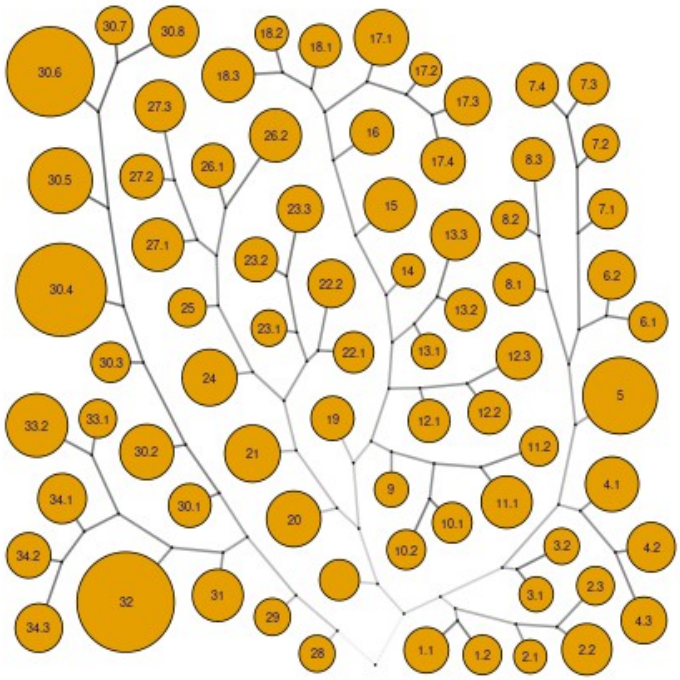

Figure 4: Tree-and-leaf network represents gene-gene interactions as modules, with similar modules clustered together based on the functions served by those interactions. The numbers of the modules are assigned in the order in which the module was calculated and added to the tree.

DINA has been used in a previous version to work on pathogenic mechanisms and vaccine development for human immunodeficiency virus/ simian immunodeficiency virus (Barrenas *et al.* 2019), so testing its analytical powers in the case of SARs-CoV-2 is a novel aspect of the thesis.

1.5. Using Rhesus macaque as experimental model

Finding a representative animal model for a biological experiment is a balancing act between several, sometimes dissenting requirements. First, you need to choose an animal model that most closely recapitulates the phenomenon in humans so that the findings are relevant. In terms of evolutionary proximity, the closest animals to a human are the great apes, and it is exactly this kinship that either strictly forbids or heavily restricts their use in research. Second, from an economic and practical point of view, you need a system that thrives in relatively inexpensive environments, reproduces easily and reaches appropriate experimental age quite quickly. That is why mice, flies, fish and worms are the most common models used in experiments. Third, from an ethical and humanistic point of view, you want to use the simplest animal model so as to not produce useless suffering, according to what we understand about how these animals experience pain and experimental constraints.

In the case of SARS-CoV-2, we need an animal that has a comparable immune system with ours, that displays the clinical symptoms given by the infection and if possible, to show variation in the response to the viral infection, just like in humans. Various animal models have been used in research, each with advantages and disadvantages. Mice infected with SARs-CoV-2 have no symptoms even though the virus is detected and it replicates (Haagmans & Osterhaus 2006). This

could be due to the fact that the virus does not efficiently bind to mouse ACE2 receptor (Bao *et al.* 2020). However they display clinical symptoms to various degrees, when the mice are genetically modified to express human hACE receptor either by creating classical transgenic mice or by expressing it using adenoviral or clustered regularly interspaced short palindromic repeats (CRISPR) systems (Johansen *et al.* 2020).

Ferrets are naturally susceptible to SARs-CoV-2 infection but they have mild symptoms and fast disease resolution. This is believed to be caused by a difference in tissue tropism of the virus between humans and ferrets. In the latter, the virus replicates mainly in the upper respiratory tract, unlike in humans. Therefore this model is mostly used to study viral transmission. Golden hamsters are more promising, but they don't develop severe symptoms. Tests have been carried also on cats, dogs, chickens and pigs, but they were found to not be suitable for a variety of reasons (Johansen *et al.* 2020).

Overall, these impediments are further compounded by the lack of 'realism' of standard experimental models. Most small animals used in research are inbred in order to minimize the variability between subjects and to be able to see the results easier. This inherently means they have a very low genetic diversity. And it is this difference in the composition of the genetic material that determines the differences in how the body reacts. Also, it is difficult to model the complex comorbidities associated with increased infection susceptibility and mortality in humans, such as chronic diseases of the lung (e.g. chronic obstructive pulmonary disease, asthma, lung fibrosis), cardiovascular diseases, diabetes or obesity. Imperfect models compound problems or respond in unexpected or incomplete ways when models of various diseases are bred together.
Considering all these complications, the model that has been found to be most appropriate to study immunological responses to SARs-CoV-2 is non-human primate (NHP), among them rhesus macaque (RM) showing a great degree of similarity to humans regarding symptoms, tissue damage and disease evolution (Munster *et al.* 2020). However the severe outcomes, such as the cytokine storm, are not present (Harrison *et al.* 2020).

2. Methods

The following chapter details the methods and procedures used to assess, process and analyze the data, the software and packages required to repeat the process.

2.1. Study animals and ethical approval

The RM were housed and treated according to the animal welfare regulations of the Washington National Primate Research Center and all animal experiments have received ethical approval from the institute where they took place. The actual sequencing was done at the Rocky Mountain Labs under the supervision of Dr. David Hawman and Dr. Heinz Feldmann from the National Institute of Health.

2.2. Study design

The data used in this project has not been published before and it was obtained through a collaboration between the laboratory of Prof. Michael Gale from University of Washington and Prof. Jan Komorowski from Uppsala University. Twelve male Rhesus Macaque (Macaca mulatta) NHPs have been randomly split into two groups and vaccinated either against SARs-CoV-2 (6 animals, numbers #308 - # 313) or against zika virus (6 animals, numbers #314 - # 319). For the next five weeks, blood was extracted weekly for transcriptomics (Fig. 5). At the end of this period, both groups of monkeys were challenged with SARs-CoV-2 virus and blood was extracted every two days to monitor the change in gene expression. One week after challenge, the animals were sacrificed and tissue samples from lungs were used for further analyses that don't make it into the scope of this thesis.

The immunization was done using a virus-derived replicon RNA vaccine (see Section 1.3. in Introduction), constructed in the laboratory of Prof. Deborah Fuller at the University of Washington. The vector contained either the spike protein (S) from the SARs-CoV-2 WA-1 isolate, or the zika pre-membrane and envelope coding sequence (prME).

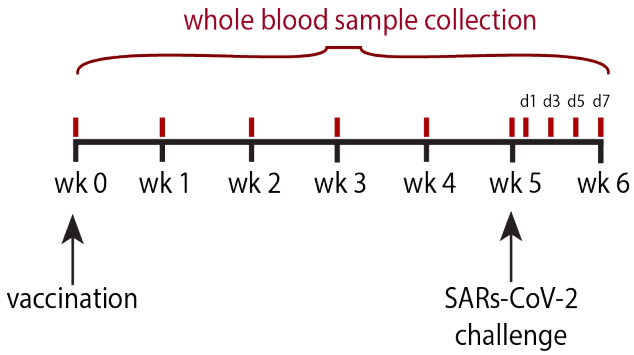

Figure 5: Study design. The monkeys were vaccinated after the first blood collection at week 0. Blood was collected every week, until week 5 when the monkeys were challenged with SARs-CoV-2 virus. Sampling was continued every other day, until day 7.

2.3. Extracting differentially expressed genes

All the assessments and statistical analyses were executed using Rstudio (version 1.4.1106 on R version 4.0.5). For extracting the lists of differentially expressed genes for each time point, I used the R packages ‘limma’ (version 3.46.0) and ‘egdeR’ (version 3.32.1). In short, I have created a DGEList target dataframe, an object that contains the counts for each gene and each time point, the groups to which these files belonged to (group = vaccine type x time point) and the gene ID. After plotting the distribution of the data to visually check for any extreme outliers, I have pre-processed the data to only retain the relevant genes. After trying several filtering methods, in the end I have settled on a manual filter that removed any gene with a count of less than 100 in more than 25% of the time points. After normalizing the gene expression distributions, I did a unsupervised clustering of the samples to see if they cluster based on a certain parameter, such as time point, animal ID or vaccination. These sources of variation need to be taken into consideration when creating the design matrix. At the end, the differentially expressed (DE) genes (genes that are significantly up- or downregulated as compared to the baseline) are extracted by fitting a linear model for the comparisons that we are interested in. In this case, I have defined DE genes as genes with an adjusted p-value < 0.05 and an absolute log fold change value > 1.5.

2.3.1. Library normalization

To get an accurate estimation of the DE genes, the initial libraries have to be resized to comparable levels. But just scaling them is too simple and it ignores several biological phenomena that can appear. For example, if in one particular time point we have a large number of uniquely or highly expressed genes, then the number of reads that could map to the other genes is reduced, leading to a sampling artifact. This skews the data leading to more false positive rates and decreases detection power for DE genes.

One way to overcome this issue is to use the ‘trimmed mean of M-values method’ (TMM), where a trimmed mean is the average after removing the upper and lower percentage of the data (Robinson & Oshlack 2010). This method presumes that the majority of genes in the experiment are not DE and this assumption is tested with a likelihood ratio testing. Library normalization has been performed using the ‘*calcNormFactors*’ function from the ‘edgeR’ package.

2.3.2. Unsupervised clustering

In unsupervised clustering, data is sorted into categories (‘clusters’) in the absence of a label that would assign the samples. It shows how similar or dissimilar samples are to each other. The plots thus give a good indication if there are DE genes to be discovered in the experimental datasets.

For my samples I have done unsupervised clustering using the ‘*MDSplot*’ function of the ‘limma’ package. This plot displays leading log-FC changes between each pair of RNA samples as distances.

2.3.3 Hierarchical clustering

Hierarchical clustering is a method of grouping objects or variables according to how close they are to each other, based on a certain attribute. In this case, the genes are clustered based on their expression. Each iteration of the algorithm places together the two most similar samples or clusters. Regarding the distance between the clusters, in this thesis I have used the biweight mid-correlation (Wilcox 2012, page 399), because it is more robust to influence from outliers compared to other methods such as Pearson correlation (Langfelder & Horvath 2012).

2.4. Functional enrichment analysis: GOsim

When performing a functional enrichment analysis what we are looking for is to determine if a class of objects is over-represented compared to the other classes of objects in a group. The objects can be genes, proteins, functions or cellular locations. This analysis was performed with 'GOSim' package (version 1.28.0) using the DE gene lists, after I have translated them from Ensembl macaque IDs (ENSMMUG*) to human Ensemble IDs (ENSG*).

2.5. Creating the DINA network

The DINA network was created by downloading freely available datasets done on whole or peripheral blood monocytes. The diseases analyzed ranged from smoking/ chronic obstructive pulmonary disease, fetal lung development, asthma/ pulmonary hypertension, to infectious diseases such as SARs-CoV-2, MERS-CoV and H1N1, H5N1, H3N2 and H7N9 influenza. All these datasets are representative of the various diseases that can affect the upper respiratory pathways just like SARs-CoV-2.

The first part of the pipeline takes the gene nomenclature and translates it to human Ensembl Gene ID, using biomaRt package. The array is then annotated to connect the microarray probes to the human Ensembl Gene ID. After the expression data passes quality control, it goes through rank normalization. If the data is RNAseq, then it is transformed to counts per million (cpm). The end processed data is the average of the gene expression. The gene-gene interaction is calculated using the biweight mid-correlation algorithm described in section 2.3.3. Together with the metadata, the gene-gene interactions are placed in biological modules and displayed as a tree-leaf-network (Fig. 4). Basically, modules contain genes with similar function.

2.6. Visualizing networks

After applying the data from the NHP experiment ontop of the DINA network, I can extract gene-gene interactions that I am interested in as subnetworks and then display them graphically using the 'igraph' R package (version 1.2.6).

3. Results

3.1. Extracting the differentially expressed genes

3.1.1. Exploring various threshold for filtering out potentially irrelevant genes

The first step in obtaining DE genes is to filter out the genes with counts under a certain threshold. Biologically, it is improbable that genes with very low levels of expression have an important role in the physiological process that we are studying and their low count will interfere with the statistical approximations that are used in the subsequent steps. Also their presence in the data set will increase the number of comparisons because they will be considered for the multiple testing correction when estimating the false discovery rates. This will lead to a decrease in the power of detection, meaning that we can loose some of the DE genes.

For this purpose I have tried various thresholds, both manual and automatic (*filterByExpr* function from 'edgeR' package, Fig. 6.D). The manual thresholds were chosen to be: A) count > 100, B) count > 30, and C) count > 10. As the value of the threshold decreases, genes with low counts appear on the left side of each density plot (Fig. 6, black arrows). The threshold for the automatic filtering could not be displayed in fig 6.A. because the expression that calculates it takes in account both the min nr of reads and library size.

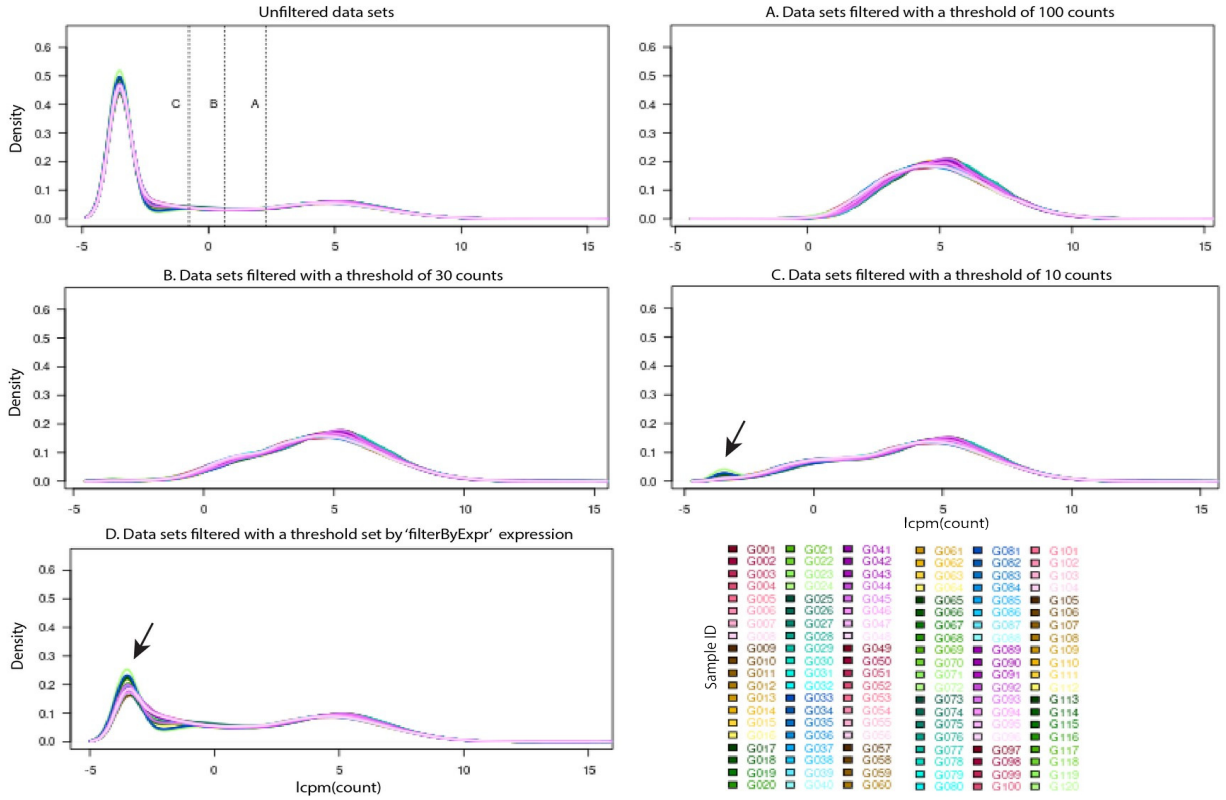

Figure 6: Filtering log transformed counts using various thresholds. First panel displays unfiltered data sets and the thresholds used for the subsequent panels. Threshold D could not be plotted because it is not a single value, it is an expression. Panels A-D show the density plot for each of the respective filter. The black arrows point to the low expressing genes that start to appear in the data sets when the threshold decreases.

I chose to continue processing the datasets using the most stringent threshold (> 100 counts) because it is commonly used in processing vaccination data from NHP, and it is likely to remove from analysis genes that are not biologically meaningful.

3.1.2. Unsupervised clustering of samples

First, I calculated the normalization factors using the TMM method (see Section 2.3.1). For each of the filtered DGEList from before, I have applied the factors and plotted the results (Fig. 7). The plots show that the normalization was successful and we obtained the effective sample size.

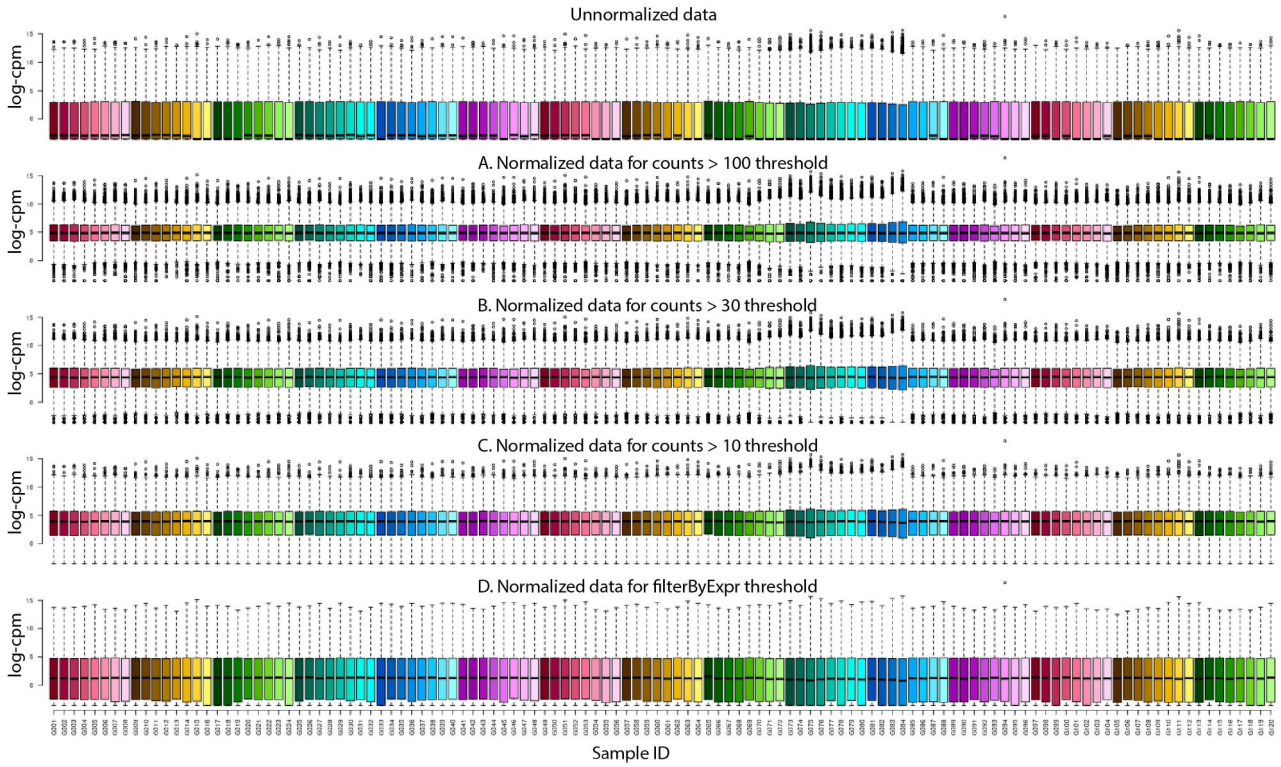

Figure 7: Normalization of the samples using the TMM method for data that has been filtered with the different thresholds.

The main source of variance in a transcriptomics experiment is the animal ID, meaning that samples originating in the same animal are more similar for various reasons, compared to samples picked at random from various animals. However this was not the case in the current experiment. The main source of variation seemed to be the time point, meaning that all the samples collected in the first day after challenge (DPC1) are obviously different and thus they group separately in the plots (Fig. 8, A and B). When I removed DPC1 from the data set, the clustering of the remaining samples behaved as expected, clustering by animal ID (Fig. 8, C and D).

Taking this unexpected information into consideration, I decided to take a closer look at the samples and I have plotted the counts in various ways (Fig. 9). By plotting the total number of raw counts for each sample grouped by time point (Fig. 9.A.), I wanted to see if the samples belonging to the DPC1 time point have abnormally large or small sizes. However they were within normal range,

especially comparing with the outlier samples G065 from WPV5 and G089 from DPC3 (Fig. 9.A., pink rectangle).

Even if the number of total counts is ‘within range’, it is still possible that the distribution of the counts is different. I have plotted again the samples grouped by time point, but this time I only focused on the genes that have a count of zero (Fig. 9.B.). While one of the DPC1 samples can be called an outlier (G074), having the highest value, the rest of the DPC1 samples are within the range of the group. This means that overall there are not more non-expressed genes in this time point compared to the others.

Still, considering the threshold that I have chosen as ‘biologically meaningful’, I have decided to look at one more parameter: all the genes that are expressed above that threshold of 100 counts. Figure 9.C clearly shows that the samples in the DPC1 time point are depleted of genes with high counts. Overall, this means that the gene expression pattern for DPC1 is enriched in genes with low expression, that would not make it into the analysis. And this pattern makes the DPC1 samples group together and stand out in the unsupervised clustering plot (Fig. 8.B).

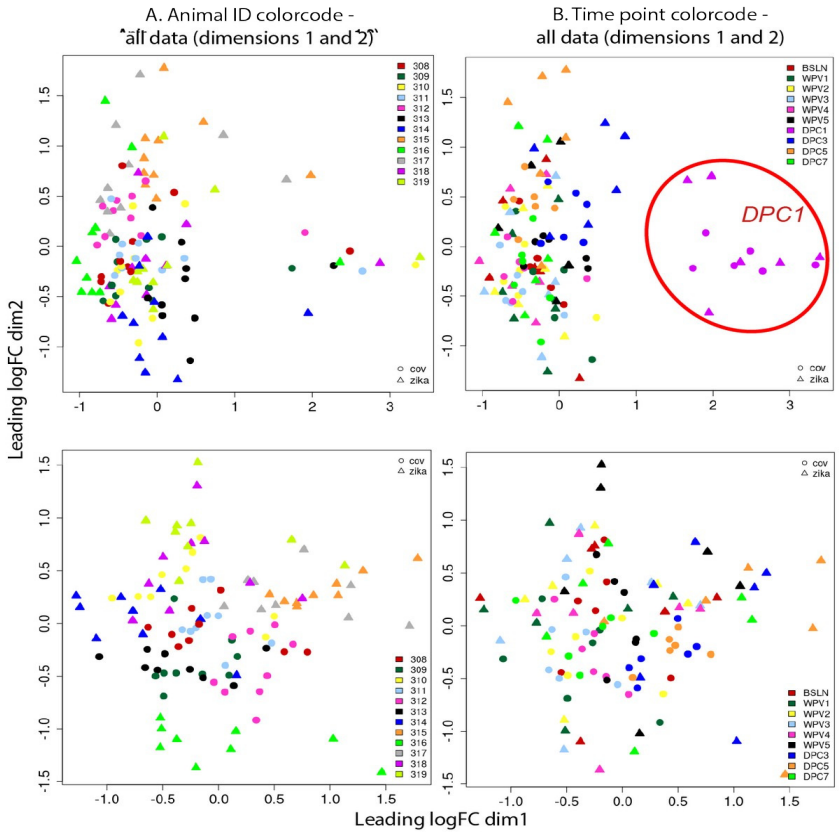

Figure 8: Unsupervised clustering of samples from the experiment. (A) samples are color coded on animal ID. (B) samples are color coded based on the time point when they have been collected. The first row of plots contain the outlier time point, while the lower row does not. Circles describe SARs-CoV-2 samples, which triangles are zika samples.

After extracting the number of DE genes from each time point, in relation to the initial baseline (tp 0), I did not get any DE genes for most of the post-vaccination time points for zika (Table 2., in Appendix). The possible reasons are presented in the Discussion section.

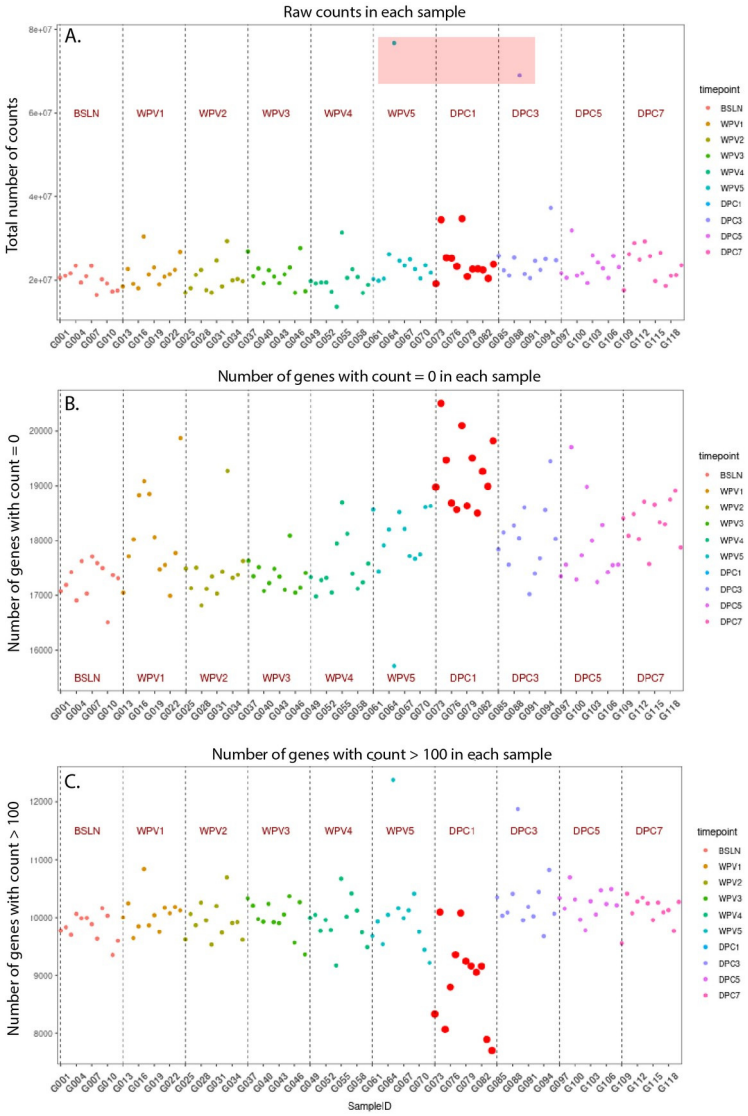

Figure 9: Exploring the unusual pattern observed for the DPC1 time point. (A) plotting the total raw counts for each samples to find outliers. (B) plotting the number of genes that have a count of zero in each sample. (C) plotting the number of genes that have a count above the chosen threshold (> 100 counts)

When plotting the total number of DE genes post-challenge versus the number of DE genes except in DPC1, I get a roughly 50% reduction in the number of DE genes, both for SARs-CoV-2 vaccine and for the zika vaccine (Fig. 10). This means that the genes expressed in DPC1 are mostly uniquely regulated genes, compared to the other time points. When comparing the numbers of DE genes from the two conditions (SARs-CoV-2 vs zika), the numbers are rather similar for each phase of the experiment: post-vaccination and post-challenge (Fig. 10).

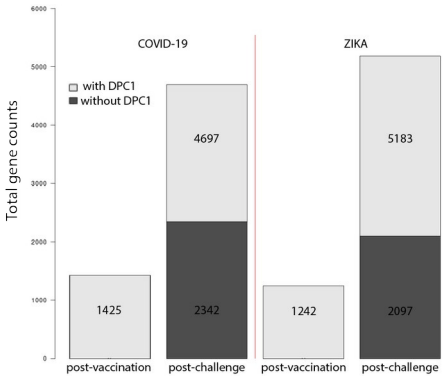

Figure 10: Exploring the total number of DE genes in each of the experimental conditions (SARs-CoV-2 and zika), with or without the presence of the DE genes from the DPC1 time point.

3.2. Gene clustering

From this point forward, I have looked at the data both in the absence and in the presence of the DPC1 time point. I then clustered together the genes for all time points and both experimental conditions to see if I can detect a differential signature for SARs-CoV-2 protection. When I clustered all the DE genes, I observed that the genes clustered within either 9 clusters if DPC1 was present (Fig. 11) or 5 clusters if DPC1 was absent (Fig. 12). It is normal that the presence of more genes that are unique to DPC1 contributes to the increased number of clusters in which the data segregates.

The unusual gene expression of the time point DPC1 is clearly visible (wk5 d1) for both vaccination conditions, but another pattern also became apparent for WPV5 (wk5) time point. What is equally surprising is how similar the post-challenge distribution of the genes in SARs-CoV-2 looks to the one for zika.

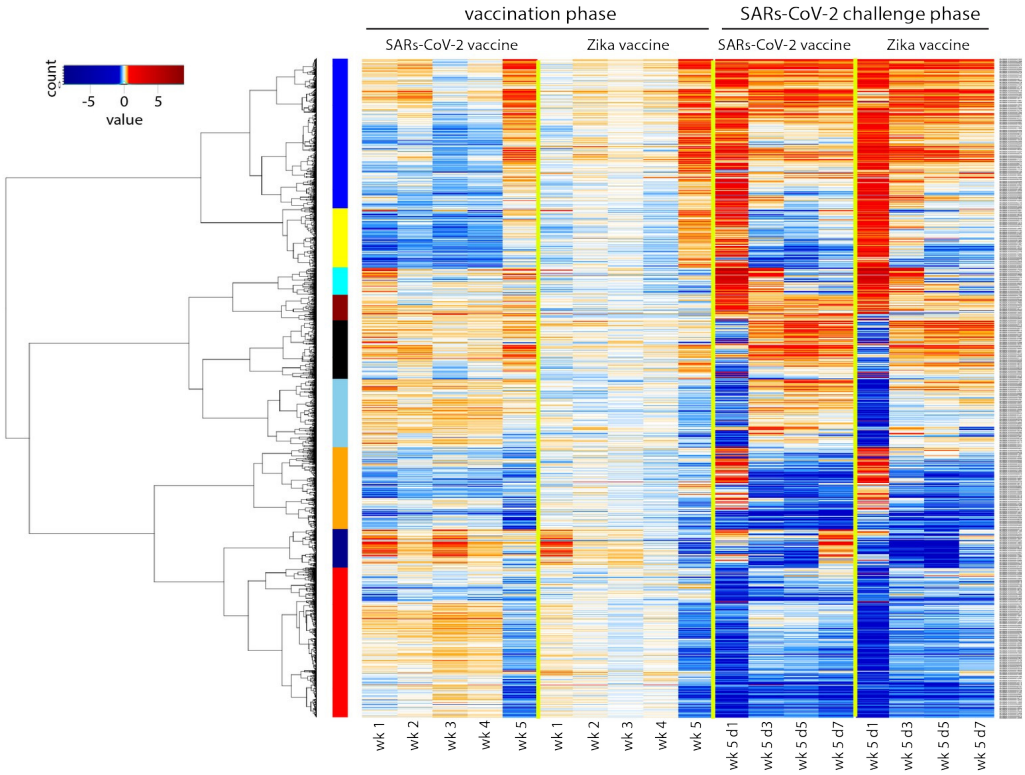

Figure 11: Comparison of differentially expressed genes between NHP given a specific vaccine designed to protect against SARs-CoV-2, and NHP that were given another vaccine, in this case against zika. The data displayed in the heatmap represents the union of all DE genes expressed at each time point and along both vaccinations.

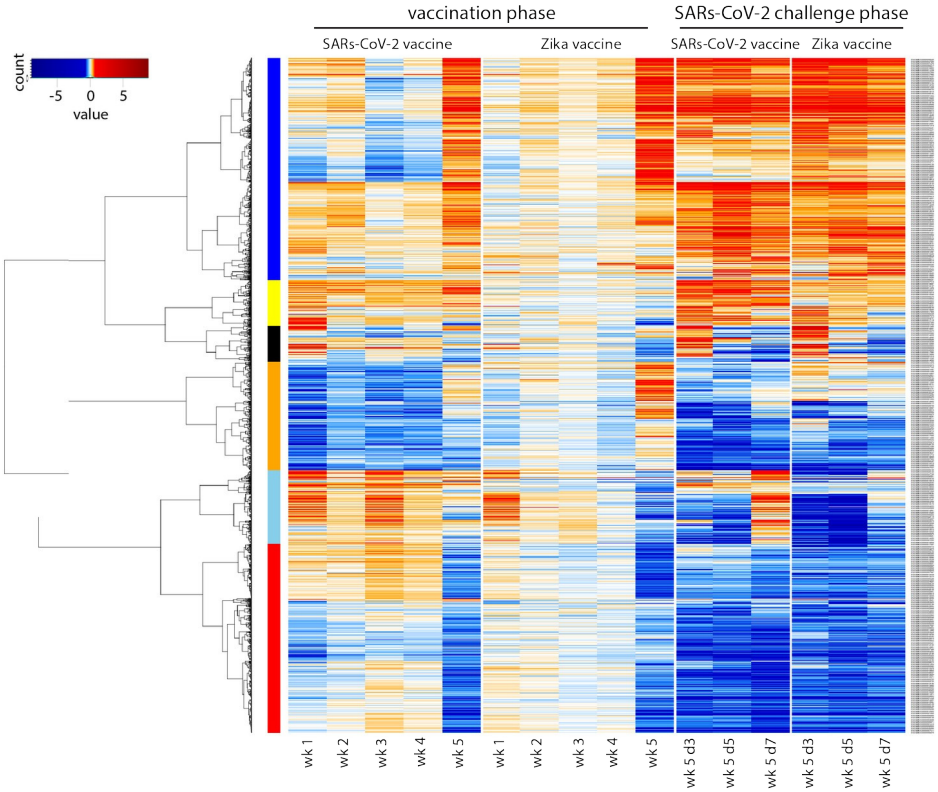

Figure 12: Comparison of differentially expressed genes between NHP given a specific vaccine designed to protect against SARs-CoV-2, and NHP that were given a vaccine against zika. The data displayed in the heatmap represents the union of all DE genes expressed at each time point and along both vaccinations. Missing from the union are the DE genes for the DPC1 time point to emphasize the simplification of the gene clustering.

3.3 Using DINA to display the results

After seeing some promising and interesting patterns in the heatmaps, all the data has been loaded onto the DINA network to extract further clarifying information about the how the genes react in these separate experimental conditions (Fig. 13).

The modules have been ranked and then the top upregulated and top downregulated modules have been extracted from each network. Overall, some modules are very similar between the two conditions and they display a trend in a similar direction, such as 13.1, 3.2 and 8.3. However, modules like 30.7 is among the top regulated modules, but it shows opposite regulation.

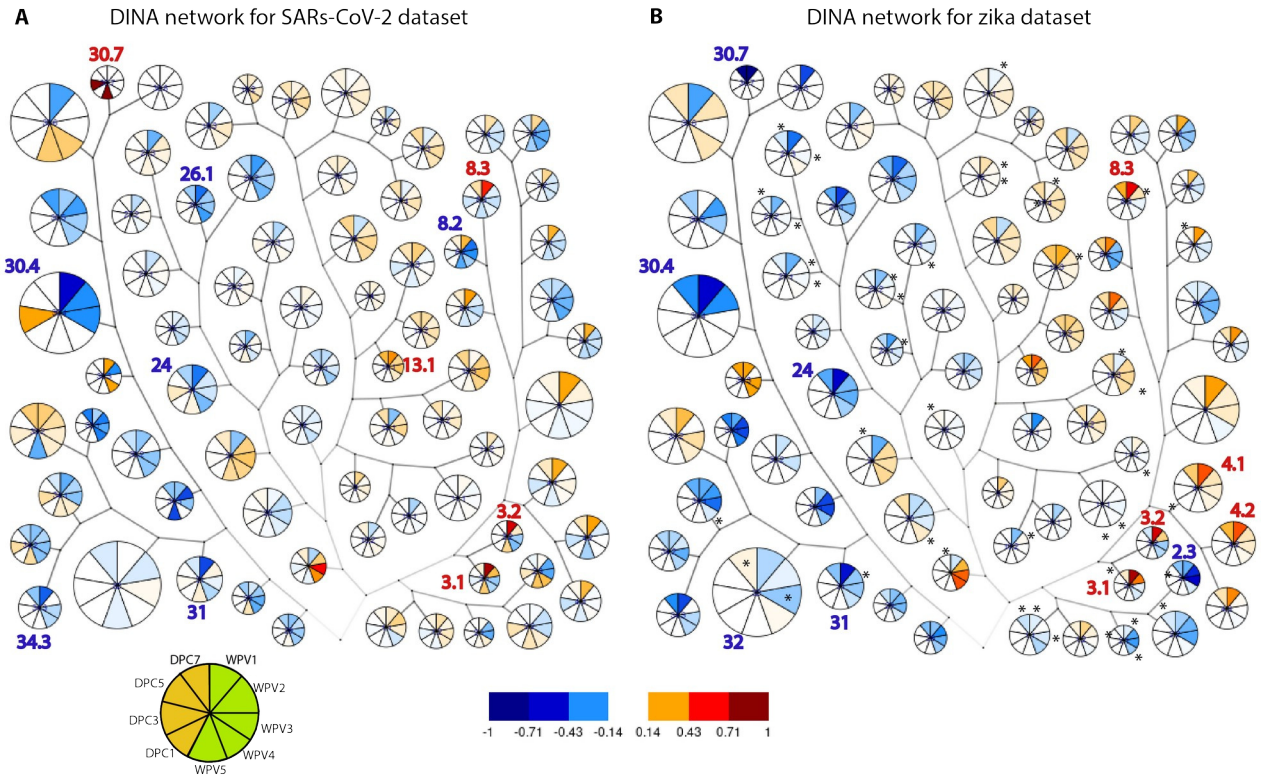

Figure 13: DINA network loaded with the data from the whole times series, both for covid (A) and for zika (B). The modules have been scored and ranked. For each dataset, the top up-regulated (numbers in red color) and top down-regulated modules (numbers in blue color) have been marked. The asterisks in the DINA for zika represent the modules that displayed opposing regulatory pattern between modules for zika and those for SARs-CoV-2, for each time point.

When looking at the gene ontology (GO) terms under which the gene-gene interactions are grouped (see Table 1 in the Appendix), the most representative belong to broad biological mechanisms, such as tissue differentiation, but there are also specific functions such as T cell differentiation and activation.

No GO terms could be retrieved for the module of interest 30.7. This could mean that there were not enough interactions that could be assigned to that module to give a significant identification.

Table 1: Classification of the GO terms extracted from the upregulated and downregulated modules in DINA. For a more detailed list, see Table 2 in Appendix.

Upregulated modules		Downregulated modules	
SARs-CoV-2	**zika**	**SARs-CoV-2**	**zika**
3.1 and 3.2 - cellular biosynthetic processes: macromolecule modification, lipoproteic metabolism	**3.1 and 3.2** - cellular biosynthetic processes: macromolecule modification, lipoproteic metabolism	**31** - regulation of leukocyte and dendritic cell differentiation	**32** - lysosomal transport, protein transport, export from the cell
8.3 - modulation by symbiont of entry into host, carbohydrate metabolic process	**8.3** - modulation by symbiont of entry into host, carbohydrate metabolic process	**34.3** - bone, skin and skeletal muscle development, cell differentiation - cellular hormone metabolic, lipid catabolic process, negative regulation of cells, response to nutrient/ extracellular stimulus (all vitamine A related) - T cell differentiation and activation - fluid homeostasis (water loss via skin) - cellular responses to chemicals	**2.3** - connective tissue, epithelium development - reduction in angiogenesis, endothelial cell differentiation and reduction in proliferation response to growth factors
13.1 - RNA metabolic processing	**4.1** - neuron differentiation	**26.1** - eye, heart, epithelium development - cellular response to stress, cell aging - tissue remodeling - positive regulation of cell , cell signaling	**24** - macromolecule biosynthetic processes
		8.2 - movement in host environment - carbohydrate metabolic process	**31** - regulation of leukocyte and dendritic cell differentiation

3.4. Network analysis for the gene of interest

Going back to the clustered genes displayed in figures 11 and 12, I have decided to extract groups of genes and analyze them based on the aims of the project.

3.4.1. The unique signature for SARs-CoV-2 vaccine

The first aim was to find out if I can pinpoint the genes or networks of genes that differentiate between the ‘ready state’ of the immune system after vaccination for SARs-CoV-2. For this objective I can’t compare the whole overall behaviour of the immune system during the post-

vaccination period, because there were no DE genes identified for the time points WPV1 – WPV4 for zika. I could still compare the WPV5 end time point between the two vaccines.

The first approach was to find all the DE genes from time point WPV5 that were significantly upregulated in one case, and significantly downregulated in the other. Unfortunately, none of the genes that showed marked difference between SARs-CoV-2 and zika at WPV5 belonged to a large subnetwork, they were mostly single genes.

Another approach was to extract the genes that are significantly upregulated or downregulated in SARs-CoV-2, and having the opposite behavior in zika, but in this case I have expanded the search to non-significant genes. In this case I got too many genes to make a useful conclusion from the experiment (Fig. 18 in the Appendix).

Next I extracted the DE genes that are present only in the SARs-CoV-2 data at time point WPV5 but not in the zika data. After extracting the 262 uniquely SARs-CoV-2 genes, I have displayed them in a co-expression network using the package ‘igraph’ in R (Fig. 14). The largest subnetwork contained only 12 genes, and they are presented in the Discussion section.

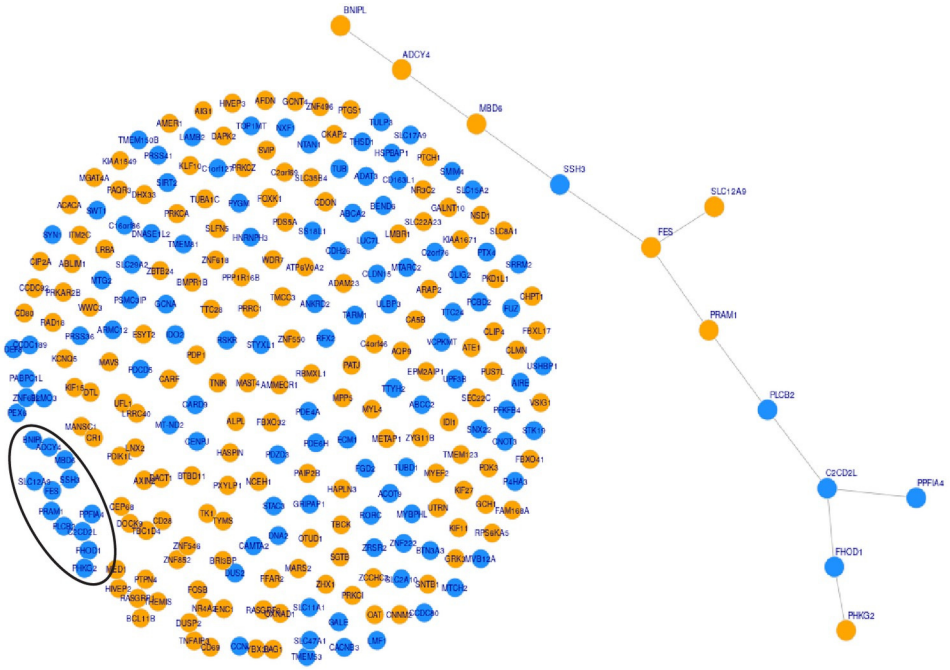

Figure 14: The complete network and the largest subnetwork showing all the unique DE genes expressed only in the WPV5 time point of the experiment. All the upregulated DE genes are displayed in orange, while the downregulated DE genes are displayed in blue. The black area points to the location of the extracted subnetwork.

To get a bigger picture regarding the role of these DE unique genes, I have displayed them on top of a co-expression network made of all the DE genes that I have found in this experiment. From the total of 6035 DE genes, I have obtained several large subnetworks, the largest being presented in Fig. 15.

The DE genes unique to SARs-CoV-2 cluster nicely in separate areas of this subnetwork. To know more about the network, I have extracted the top three hub-genes (genes with the largest number of connections to other genes). *CEP350* (red node, Fig. 15) has 15 connections, followed by *TAOK1*

(pink node, Fig. 15) with 9 vertices, and *ZNF692* with 8 vertices (asterisk, Fig. 15). The last gene is also one of the uniquely upregulated genes that I have found for time point WPV5. In short, Fig. 15 contains the signature for the SARs-CoV-2 vaccine.

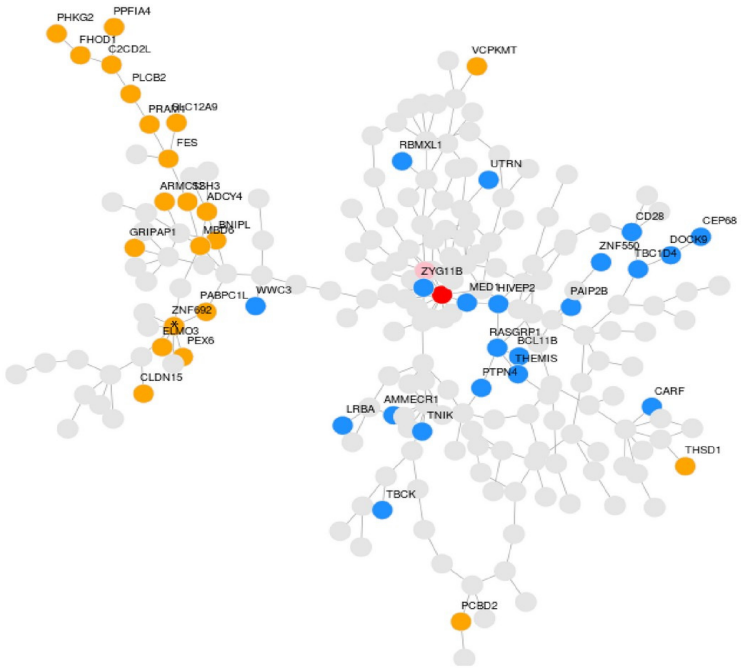

Figure 15: Largest subnetwork displaying the interconnectedness of DE genes expressed throughout the experiment, for both SARs-CoV-2 and for zika. All the upregulated DE genes are displayed in orange, while the downregulated DE genes are displayed in blue. The red dot represents the hub with the highest number of connections in this sub-network (CEP350), the pink dots represents the second largest hub (TAOK1), and the asterisk is the third hub (ZNF692).

3.4.2 Analyzing the response of the primed immune system to viral infection

The first step in this approach, was to look at the differences that appear in gene expression at every time point following infection (DPC1 - DPC7). I have proceeded similarly to the previous section, where I have extracted the DE genes that have the opposite response to challenge in each of the experimental groups. Unfortunately, all the largest subnetworks were extremely dense and I could not draw any conclusions from this analysis (Fig. 19, in the Appendix).

As before, I have decided to extract the DE genes that are expressed in the last time point of the challenge (DPC7) for SARs-CoV-2 but not for zika. The largest subnetwork that they are a part of is formed of nine genes (Fig. 16), the main hub-genes being IFI44L and IFIT3.

As a last analysis, I wanted to see how do these genes look like when displayed on the network formed of all the DE genes from this experiment. The largest subnetwork is made of 11 genes (Fig. 17), and several of them are common to the genes from only the uniquely DE genes at DPC7.

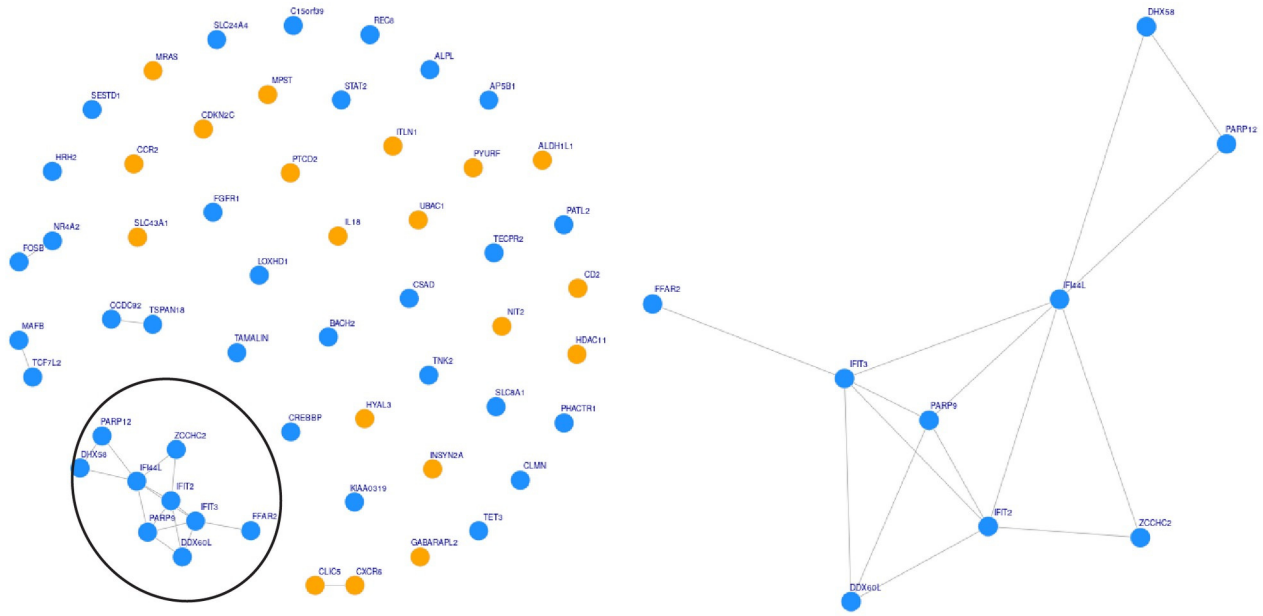

Figure 16: The complete network and the largest subnetwork showing all the unique DE genes expressed only in the WPV5 time point of the experiment. All the upregulated DE genes are displayed in orange, while the downregulated DE genes are displayed in blue. The black area points to the location of the extracted subnetwork.

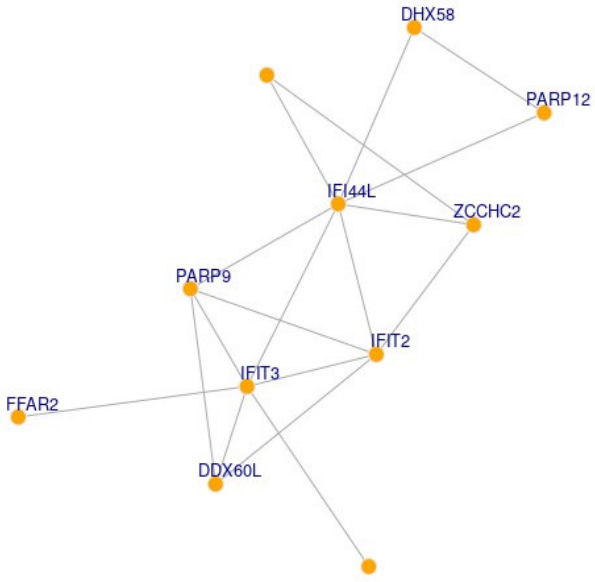

Figure 17: Largest subnetwork displaying the interconnectedness of DE genes expressed throughout the experiment, for both SARs-CoV-2 and for zika. All the DE genes are upregulated (orange). The hub with the highest number of connections in this sub-network is IFI44L, followed by IFIT3.

4. Discussion

The main purpose of this thesis is to determine how does the immune system get primed following immunization with a new SARs-CoV-2 vaccine and to explore the immuological mechanisms that confer resistance to the disease. The experiment was done in NHP in order to have the best chance of obtaining relevant findings. The ethics of this choice was discussed in section 1.5. All the animals used in this experiment had the approval from the ethical committee of the experimental institution. After preparing and normalizing the data, following the unsupervised clustering of the samples, I could detect an unusual grouping based on time point, instead of the usual animal ID (Fig. 8). The outlier group was made out of the samples collected 1 day after challenge (DPC1) with SARs-CoV-2 virus for both vaccination groups. Upon a closer inspection, it is probable that this unusual reaction is provoked by how the new vaccine was designed to function. Virus-derived replicon RNA vaccine are designed to mimic the behaviour of a viral particle, being able to continuously synthesize the viral RNA (Berglund *et al.* 1998). As such it induces strong inflammatory and innate immune reactions in the organism. Erasmus *et al.* (2020) have shown that a single dose of this vaccine (250 ug) in NHP triggers the same strong immune activation if you would use a fifth of this amount followed by a boost 28 days later. All measurements for interferon-gamma, T cells and various cytokines (such as IL-2, IL-17A and TNF) by this time point showed the immune system reached its peak activation, a state that was maintained until the end of the experiment at day 70. This state is confirmed by pattern observed in the heatmaps for the last time point before challenge, 5 weeks after vaccination (WPV5, Fig. 11 and Fig. 12). In these conditions, it is possible that the fully activated immune system of the NHP engaged the virus and reacted strongly to the challenge.

What makes the DPC1 time point stand out during clustering is the fact that the are more genes with low-to-intermediary gene expression (< 100 counts) when compared to the other samples (Fig. 9). Additionally, a big proportion of the genes expressed at time point DPC1 are unique, when compared to the DE genes extracted for the whole experiment. When removing DPC1 DE genes the number of DE genes for the whole experiment drops to 49,86% for SARs-CoV-2 series and to 40,46% for zika series (Fig. 10). This is a confirmation of the strong response from the body against the viral infection.

When studying the gene expression clustering, at the first glance it looks as if the response for both vaccines is rather similar (Fig. 11 and Fig. 12). Same can be said about the response to the challenge with SARs-CoV-2 virus. However this is likely because of the interference from the strong immune response to the vaccine. While gene clustering helps classify the genes into several clusters enriched in certain pathways or biological functions, still it is difficult to have a clear overview of what is happening overall with the gene expression.

To solve this issue, I took advantage of methods from systems biology that can boost the experimental observation and increase the confidence in the observed results. This is particularly helpful in the case of experiments where there the samples are precious (either the samples are prohibitively expensive or there are not many patients that have a certain disease). DINA is a framework made from freely available transcriptomic datasets, that are compatible or similar to the disease or the biological process that is being studied. Once the correlation matrices are extracted from these datasets, they are collapsed into a single matrix. Basically each gene pair is evaluated

over each of the constituent dataset (Barrenas *et al.* 2019), meaning that gene-gene interactions that are more frequently observed are more likely to be biologically relevant, therefore they are retained.

Besides grouping gene-gene interactions into modules, DINA can also visually display at a glance and color code the behaviour of each module at each time point in the experiment. In general, the top upregulated and downregulated modules coincided for both vaccines. The main upregulated functions for both vaccines were cellular biosynthetic processes (Modules 3.1 and 3.2) which are necessary when preparing to expand the immune system in order to fight a pathogen. Also pathways related to entry into a host (Module 8.3). The main common downregulated pathways were related to regulation of leukocyte and dendritic cell differentiation (Module 31), two very important components of the innate immune response.

The differences in the modules also made sense. For zika vaccine, the upregulated Module 4.1 contains gene-gene interactions that have to do with neuronal differentiation. Zika is a virus that through not yes completely understood mechanisms, affects the neuronal development in foetuses by depleting neural progenitor cells (Dang *et al.* 2016) and thus leading to babies suffering from microcephaly (Tetro 2016). RNA metabolic processing is a common biological function underlying a multitude of biological functions (Module 13.1).

Module 34.3 in SARs-CoV-2 series has a particular set of functions related to vitamin A utilization: T cell differentiation and activation, but also nutrient response. Supplementation with vitamin A has been associated with maintenance of the innate immunity, with minimizing inflammation in the lungs, and repair of the respiratory epithelium (Stephensen & Lietz 2021).

Besides ranking the modules in the order of their downregulation or upregulation, DINA also facilitates comparison, making it very easy to find the modules that react in opposite manner. For example, module 30.7 is overall upregulated in SARs-CoV-2 but overall downregulated in zika. However, I believe that much more interesting it is to study the modules that show this opposite response at the same time point. In figure 13.B, asterisks point to such behaviour. However this detailed investigation was not possible due to time constraints.

In a searching for how the specific immune response develops against SARs-CoV-2, I wanted to compare the DE genes obtained at each time point for each vaccination series (e.g. WPV1_cov vs WPV1_zika). However, when extracting the DE genes, I did not obtaine any DE genes for the intermediary time points for SARs-CoV-2, therefore this approach was not possible. The reason I did not find any DE genes at those time points can be due to having chosen a rather stringent threshold for keeping gene reads that are considered ‘biologically meaningful’ (> 100 counts). As can be seen from figure 6, another threshold that does a good job of eliminating the low counts is also > 30 counts. However, an argument for keeping the higher threshold is that, when extracting the DE genes and performing correction for multiple testing, the more comparisons we have to make, the lower the discovery power of DE genes will be.

To search for the signature for immune priming against SARs-CoV-2, I have used a few approaches. First I wanted to extract those DE genes that are significantly upregulated in SARs-CoV-2 and significantly downregulated in zika overall and the other way around (post-vaccination_cov vs post-vaccination_zika), and then plot them as gene-gene interaction networks. However these genes mapped exclusively outside the largest sub-network, mostly being disconnected. Then I switched tactics and decided to display genes that are significantly upregulated in SARs-CoV-2 and non-significantly downregulated in zika (and the other way around as well). In this case I got too many

genes that fit this description (Fig. 18 in the Appendix). However, this is why DINA is such a powerful approach, because it allows the user to see at a glance which modules are up/down regulated at each time point (Fig. 13, the asterisks at the opposing modules that are inversely regulated at the same time point).

Another approach was to extract the DE genes uniquely expressed only for SARs-CoV-2 at WPV5 (WPV5_cov - WPV5_zika) and map them to see which gene-gene interactions I can find (Fig. 14). The genes appearing in the largest subnetwork are involved in a variety of generic biological processes: cell signaling and motility, lipid transport and energy generation. The more promising was PRAM1, which is involved in T cell receptor mediated signaling, while SLC12A9 and BNIPL have been found in lung cancer and asthma patients. All were upregulated in the last time point after vaccination for SARs-CoV-2.

Since the genes in this network were not very specific for immune response, one more option was to look at all the DE genes for WPV5 time point, overlayed on the gene-gene co-expression network made from all the DE genes of the experiment (Fig. 15). The subnetwork was substantially bigger that the previous one (212 DE genes). In this case I searched for the most interconnected gene, because gene-hubs have a big influence in networks like this. CEP350 is a microtubule anchoring protein essential for the integrity of the microtubule network. TAOK1 has been involved in cancer mutation in lung adenocarcinoma, and it activates the MAP2K3/6 kinases. It is disregulated in inflammatory and immune disorders (Fang *et al.* 2020). On the other hand, ZNF692 is a zinc finger protein that has been found in lung adenocarcinoma where it promotes cell proliferation and motility (Zhang Q *et al.* 2017).

Similar to searching for the signature for specific immune protection against SARs-CoV-2, I decided to first compare each time point for each series (e.g. DPC1_cov vs DPC1_zika) and to look at the gene-gene interactions. However, the biggest subnetwork that I extracted for each time point had too many components and thus any biological meaning is hidden (Fig. 19 in Appendix).

I then went directly to mapping DE genes uniquely expressed at the last experimental time point (DPC7_cov – DPC7_zika) and extracting the largest network (Fig. 16). This time the network was more interconnected and it contained only down-regulated genes. Surprisingly, I obtained most of the same genes when plotting the overall DE genes from this time point (DPC7_cov vs DPC7_zika) and overlayed them on the network obtained from the overall DE genes extracted from the experiment (Fig. 17).

Unlike the previously obtained networks, these were composed in majority of genes know to be involved in immune responses. IFI44L, the gene with the greatest number of connections, has been shown to be involved in the defense response of the body to viruses and is a type I INF-related gene (Zhang J *et al. 2015)*. IFIT2 and IFIT3 are paralogs, being involved in INF-gamma signaling and are active in innate immune responses. IFIT2 is also connected to mitochondrial changes that appear when the apoptotic program is unfolding, having a positive regulatory effect in programmed cell death (Hsu *et al. 2013, Mears & Sweeney 2018)*. PARP9 is involved in INF-mediated antiviral response *(Iwata et al. 2016),* while PARP12 is involved in inflammatory responses via association with NF-kB signaling (Welsby *et al. 2014)*. DHX58 is active in the innate immune system, helping to mount an attack against the viruses through interacting directly with RIG-I (Saito *et al. 2007, Satoh et al. 2010)*. DDX60L is another INF stimulated gene that can bind to viral particles in the

cytosol, alerting the cell to the presence of an intruder (Gr*ünvogel et al. 2015)*. Lastly, FFAR2 in involved in mucosal immunity *(Chun et al. 2019).*

As a result, these networks can be considered the unique molecular signature that confers the NHP protection from SARs-CoV-2 virus.

Another approach could be to see if instead of using the samples collected at the initial time point as baseline for extracting the DE genes for the samples after the challenge (i.e. use week 0, before the vaccination as baseline for post-challenge: DEtp - DEbaseline), we define the DE genes in comparison to the signature we have at the last time point before the challenge (i.e. use week 5 WPV5 as baseline for post-challenge: DEtp - DE_WPV5). This could make sense in this experimental setting, seeing that we get such a specific gene expression signature due to the vaccine design. However I decided against it because my reason was, that the NHP were stable at baseline for a long period of time before the experiment began, therefore the expression should also be stable. However, each organism response to therapeutic intervention varies. And since this experiment was rather small (only 6 biological replicates per condition), I was not sure that this will compensate for the variation in response to the vaccine (meaning each NHP's immune system could take a varied amount of time before arriving at the same 'ready state').

5. Conclusion

The aim of this project is to explore the genetic signature of NHP that received either a new vaccine against SARs-CoV-2 or zika vaccine of the same design as a perfect baseline. The complex response of the immune system have been mapped onto a recently developed bioinformatic tool that incorporates data-driven biological annotation of gene modules and improves the inference of gene-gene interactions. The modules can then be ranked and the functions extracted to see which pathways or functions are mainly affected during the experiment. Using various approaches I have extracted the gene expression signature that provides the NHP with protection against SARs-CoV-2 virus. However, the gene networks that I have obtained for the activated state of the immune system after vaccination is less clear, since most of the genes involved in it can participate in a broad range of biological functions.

6. Future work

The dataset would benefit from reworking using other thresholds for filtering out the ‘biologically irrelevant’ expression levels considered for genes. Also going more into detail regarding the oppositely regulated DINA modules per each time point. The findings from this thesis should be reproduced using more data obtained with the same type of vaccine as in this work, because it seems that it triggers a peculiarly strong response.

Another point is for the future experiments to be made either on larger cohorts. In this case, most of the NHP were protected from the more severe effects of SARs-CoV-2, however one individual showed lung scarring similar to non-protected NHP, even if a more mild form. By increasing the number of replicates we can then exclude participants that have an intermediary response and then likely get a cleaner cell expression signature.

7. Acknowledgment

I would like to thank my supervisor, Prof. Dr. Jan Komorowski for accepting me into his group of hard working people where I have seen what a difference a steady focused attention can make. Also for his humor, constant encouragement in front of a new topic and a new area for me.

I would like to thank also my co-supervisor, Dr. Fredrik Barrenäs, for sharing his skills in bioinformatics, for his help in deciphering the secrets of systems biology and for helping me understand where my work is placed in the broader context. Without the tool he created (DINA), this project would have been a lot poorer in insights into immune gene expression.

This project could not have been possible without the contribution of the team of Prof. Dr. Michael Gale from the University of Washington, which has done the actual gathering of experimental data and have provided the useful feedback to interpret the bioinformatics analysis. Also, a big thank you to Prof. Dr. Deborah Fuller from the University of Washington which provided both the SARs-CoV-2 and the zika vaccines used in this project.

To my subject reader, Dr. Adam Ameur, for the suggestions and improvements, on how I can process the data and see it in a new light. Also thank you for the encouragement and confidence in my skills.

A great continuous support I got from Lena Henriksson, always willing to help and to find solutions when time was short.

Despite the fact that COVID-19 took a lot of freedom away from us, it did not dampen the spirits of the students. I am grateful to my fellow students for soldering on under these conditions and for making me going back to masters a wonderful second experience.

To all my other friends, thank you for your constant encouragement, cakes and zoom-therapy.

And not in the last, I want to thank my family for supporting and cheering me on every opportunity to figure out my path in life.

8. References

Azkur AK, Akdis M, Azkur D, Sokolowska M, Veen W, Brüggen M, O'Mahony L, Gao Y, Nadeau K, Akdis CA. 2020. Immune response to SARS-CoV-2 and mechanisms of immunopathological changes in COVID-19. Allergy 75: 1564–1581.

Bao L, Deng W, Huang B, Gao H, Liu J, Ren L, Wei Q, Yu P, Xu Y, Qi F, Qu Y, Li F, Lv Q, Wang W, Xue J, Gong S, Liu M, Wang G, Wang S, Song Z, Zhao L, Liu P, Zhao L, Ye F, Wang H, Zhou W, Zhu N, Zhen W, Yu H, Zhang X, Guo L, Chen L, Wang C, Wang Y, Wang X, Xiao Y, Sun Q, Liu H, Zhu F, Ma C, Yan L, Yang M, Han J, Xu W, Tan W, Peng X, Jin Q, Wu G, Qin C. 2020. The pathogenicity of SARS-CoV-2 in hACE2 transgenic mice. Nature 583: 830–833.

Barrenas F, Raehtz K, Xu C, Law L, Green RR, Silvestri G, Bosinger SE, Nishida A, Li Q, Lu W, Zhang J, Thomas MJ, Chang J, Smith E, Weiss JM, Dawoud RA, Richter GH, Trichel A, Ma D, Peng X, Komorowski J, Apetrei C, Pandrea I, Gale M. 2019. Macrophage-associated wound healing contributes to African green monkey SIV pathogenesis control. Nature Communications 10: 5101.

Berglund P, Smerdou C, Fleeton MN, Tubulekas loannis, Liljeström P. 1998. Enhancing immune responses using suicidal DNA vaccines. Nature Biotechnology 16: 562–565.

Chun E, Lavoie S, Fonseca-Pereira D, Bae S, Michaud M, Hoveyda HR, Fraser GL, Gallini Comeau CA, Glickman JN, Fuller MH, Layden BT, Garrett WS. 2019. Metabolite-Sensing Receptor Ffar2 Regulates Colonic Group 3 Innate Lymphoid Cells and Gut Immunity. Immunity 51: 871-884.e6.

Dang J, Tiwari SK, Lichinchi G, Qin Y, Patil VS, Eroshkin AM, Rana TM. 2016. Zika Virus Depletes Neural Progenitors in Human Cerebral Organoids through Activation of the Innate Immune Receptor TLR3. Cell Stem Cell 19: 258–265.

Erasmus JH, Khandhar AP, Guderian J, Granger B, Archer J, Archer M, Gage E, Fuerte-Stone J, Larson E, Lin S, Kramer R, Coler RN, Fox CB, Stinchcomb DT, Reed SG, Van Hoeven N. 2018. A Nanostructured Lipid Carrier for Delivery of a Replicating Viral RNA Provides Single, Low-Dose Protection against Zika. Molecular Therapy 26: 2507–2522.

Erasmus JH, Khandhar AP, O'Connor MA, Walls AC, Hemann EA, Murapa P, Archer J, Leventhal S, Fuller JT, Lewis TB, Draves KE, Randall S, Guerriero KA, Duthie MS, Carter D, Reed SG, Hawman DW, Feldmann H, Gale M, Veesler D, Berglund P, Fuller DH. 2020. An *Alphavirus* -derived replicon RNA vaccine induces SARS-CoV-2 neutralizing antibody and T cell responses in mice and nonhuman primates. Science Translational Medicine 12: eabc9396.

Fang C-Y, Lai T-C, Hsiao M, Chang Y-C. 2020. The Diverse Roles of TAO Kinases in Health and Diseases. International Journal of Molecular Sciences 21: 7463.

Flanagan KL, Best E, Crawford NW, Giles M, Koirala A, Macartney K, Russell F, Teh BW, Wen SC. 2020. Progress and Pitfalls in the Quest for Effective SARS-CoV-2 (COVID-19) Vaccines. Frontiers in Immunology 11: 579250.

Grünvogel O, Esser-Nobis K, Reustle A, Schult P, Müller B, Metz P, Trippler M, Windisch MP, Frese M, Binder M, Fackler O, Bartenschlager R, Ruggieri A, Lohmann V. 2015. DDX60L

Is an Interferon-Stimulated Gene Product Restricting Hepatitis C Virus Replication in Cell Culture. Journal of Virology 89: 10548–10568.

Haagmans BL, Osterhaus ADME. 2006. Nonhuman Primate Models for SARS. PLoS Medicine 3: e194.

Harrison AG, Lin T, Wang P. 2020. Mechanisms of SARS-CoV-2 Transmission and Pathogenesis. Trends in Immunology 17.

Hsu Y-L, Shi S-F, Wu W-L, Ho L-J, Lai J-H. 2013. Protective Roles of Interferon-Induced Protein with Tetratricopeptide Repeats 3 (IFIT3) in Dengue Virus Infection of Human Lung Epithelial Cells. PLoS ONE 8: e79518.

Iwata H, Goettsch C, Sharma A, Ricchiuto P, Goh WWB, Halu A, Yamada I, Yoshida H, Hara T, Wei M, Inoue N, Fukuda D, Mojcher A, Mattson PC, Barabási A-L, Boothby M, Aikawa E, Singh SA, Aikawa M. 2016. PARP9 and PARP14 cross-regulate macrophage activation via STAT1 ADP-ribosylation. Nature Communications 7: 12849.

Johansen MD, Irving A, Montagutelli X, Tate MD, Rudloff I, Nold MF, Hansbro NG, Kim RY, Donovan C, Liu G, Faiz A, Short KR, Lyons JG, McCaughan GW, Gorrell MD, Cole A, Moreno C, Couteur D, Hesselson D, Triccas J, Neely GG, Gamble JR, Simpson SJ, Saunders BM, Oliver BG, Britton WJ, Wark PA, Nold-Petry CA, Hansbro PM. 2020. Animal and translational models of SARS-CoV-2 infection and COVID-19. Mucosal Immunology 13: 877–891.

Lai MMC, Cavanagh D. 1997. The Molecular Biology of Coronaviruses. Advances in Virus Research, pp. 1–100. Elsevier,

Langfelder P, Horvath S. 2012. Fast R Functions for Robust Correlations and Hierarchical Clustering. Journal of Statistical Software 46:

Mears HV, Sweeney TR. 2018. Better together: the role of IFIT protein–protein interactions in the antiviral response. Journal of General Virology 99: 1463–1477.

Munster VJ, Feldmann F, Williamson BN, van Doremalen N, Pérez-Pérez L, Schulz J, Meade-White K, Okumura A, Callison J, Brumbaugh B, Avanzato VA, Rosenke R, Hanley PW, Saturday G, Scott D, Fischer ER, de Wit E. 2020. Respiratory disease in rhesus macaques inoculated with SARS-CoV-2. Nature 585: 268–272.

Narayanan K, Huang C, Lokugamage K, Kamitani W, Ikegami T, Tseng C-TK, Makino S. 2008. Severe Acute Respiratory Syndrome Coronavirus nsp1 Suppresses Host Gene Expression, Including That of Type I Interferon, in Infected Cells. Journal of Virology 82: 4471–4479.

Rajagopala SV, Sikorski P, Caufield JH, Tovchigrechko A, Uetz P. 2012. Studying protein complexes by the yeast two-hybrid system. Methods 58: 392–399.

Rao VS, Srinivas K, Sujini GN, Kumar GNS. 2014. Protein-Protein Interaction Detection: Methods and Analysis. International Journal of Proteomics 2014: 1–12.

Robinson MD, Oshlack A. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology 11: R25.

Rouse BT, Sehrawat S. 2010. Immunity and immunopathology to viruses: what decides the outcome? Nature Reviews Immunology 10: 514–526.

Saito T, Hirai R, Loo Y-M, Owen D, Johnson CL, Sinha SC, Akira S, Fujita T, Gale M. 2007. Regulation of innate antiviral defenses through a shared repressor domain in RIG-I and LGP2. Proceedings of the National Academy of Sciences 104: 582–587.

Satoh T, Kato H, Kumagai Y, Yoneyama M, Sato S, Matsushita K, Tsujimura T, Fujita T, Akira S, Takeuchi O. 2010. LGP2 is a positive regulator of RIG-I- and MDA5-mediated antiviral responses. Proceedings of the National Academy of Sciences 107: 1512–1517.

Stephensen CB, Lietz G. 2021. Vitamin A in resistance to and recovery from infection: relevance to SARS-CoV2. British Journal of Nutrition 1–10.

Tay MZ, Poh CM, Rénia L, MacAry PA, Ng LFP. 2020. The trinity of COVID-19: immunity, inflammation and intervention. Nature Reviews Immunology 20: 363–374.

Tetro JA. 2016. Zika and microcephaly: causation, correlation, or coincidence? Microbes and Infection 18: 167–168.

Vella D, Zoppis I, Mauri G, Mauri P, Di Silvestre D. 2017. From protein-protein interactions to protein co-expression networks: a new perspective to evaluate large-scale proteomic data. EURASIP Journal on Bioinformatics and Systems Biology 2017: 6.

Vogel AB, Lambert L, Kinnear E, Busse D, Erbar S, Reuter KC, Wicke L, Perkovic M, Beissert T, Haas H, Reece ST, Sahin U, Tregoning JS. 2018. Self-Amplifying RNA Vaccines Give Equivalent Protection against Influenza to mRNA Vaccines but at Much Lower Doses. Molecular Therapy 26: 446–455.

Wan Y, Shang J, Graham R, Baric RS, Li F. 2020. Receptor Recognition by the Novel Coronavirus from Wuhan: an Analysis Based on Decade-Long Structural Studies of SARS Coronavirus. Journal of Virology, doi 10.1128/JVI.00127-20.

Welsby I, Hutin D, Gueydan C, Kruys V, Rongvaux A, Leo O. 2014. PARP12, an Interferon-stimulated Gene Involved in the Control of Protein Translation and Inflammation. Journal of Biological Chemistry 289: 26642–26657.

Wilcox RR. 2012. Introduction to robust estimation and hypothesis testing, 3rd ed. Academic Press, Amsterdam

Peterson CW, Yu A, Zheng HB, Gideon HP, Winchell CG, Lin PL, Bingle CD, Snapper SB, Kropski JA, Theis FJ, Schiller HB, Zaragosi L-E, Barbry P, Leslie A, Kiem H-P, Flynn JL, Fortune SM, Berger B, Finberg RW, Kean LS, Garber M, Schmidt AG, Lingwood D, Shalek AK, Ordovas-Montanes J, Banovich N, Barbry P, Brazma A, Desai T, Duong TE, Eickelberg O, Falk C, Farzan M, Glass I, Haniffa M, Horvath P, Hung D, Kaminski N, Krasnow M, Kropski JA, Kuhnemund M, Lafyatis R, Lee H, Leroy S, Linnarson S, Lundeberg J, Meyer K, Misharin A, Nawijn M, Nikolic MZ, Ordovas-Montanes J, Pe’er D, Powell J, Quake S, Rajagopal J, Tata PR, Rawlins EL, Regev A, Reyfman PA, Rojas M, Rosen O, Saeb-Parsy K, Samakovlis C, Schiller H, Schultze JL, Seibold MA, Shalek AK, Shepherd D, Spence J, Spira A, Sun X, Teichmann S, Theis F, Tsankov A, van den Berge M, von Papen M, Whitsett J, Xavier R, Xu Y, Zaragosi L-E, Zhang K. 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: 1016-1035.e19.

9. Appendix

Table 2: The total number of DE genes extracted for each time point and for each vaccination. The genes were extracted with reference to the sample extracted before vaccination (tp0)

Name dataset	Nr of extracted DE genes
DE_cov_WPV1	505
DE_cov_WPV2	6
DE_cov_WPV3	205
DE_cov_WPV4	3
DE_cov_WPV5	891
DE_zika_WPV1	0
DE_zika_WPV2	0
DE_zika_WPV3	0
DE_zika_WPV4	0
DE_zika_WPV5	1242
DE_cov_DPC1	3479
DE_cov_DPC3	1389
DE_cov_DPC5	1577
DE_cov_DPC7	1336
DE_zika_DPC1	4248
DE_zika_DPC3	1410
DE_zika_DPC5	1566
DE_zika_DPC7	956

Table 3: The top upregulated and downregulated modules extracted from DINA networks for both SARs-CoV-2 and zika. The GO terms that belong to those modules are elaborated for each module.

Upregulated modules		Downregulated modules	
SARs-CoV-2	**zika**	**SARs-CoV-2**	**zika**
3.1 and 3.2 GO:0019348 "dolichol metabolic process" GO:0035269 "protein O-linked mannosylation" GO:0018279 "protein N-linked glycosylation via asparagine" GO:0006506 "GPI anchor biosynthetic process"	**3.1 and 3.2** GO:0019348 "dolichol metabolic process" GO:0035269 "protein O-linked mannosylation" GO:0018279 "protein N-linked glycosylation via asparagine" GO:0006506 "GPI anchor biosynthetic process"	**31** GO:2001199 "negative regulation of dendritic cell differentiation"	**32** GO:0033299 "secretion of lysosomal enzymes" GO:0006622 "protein targeting to lysosome" GO:0008333 "endosome to lysosome transport"
8.3 GO:2000535 "regulation of entry of bacterium into host cell" GO:0016139 "glycoside catabolic process" GO:0006004 "fucose metabolic process"	**8.3** GO:2000535 "regulation of entry of bacterium into host cell" GO:0016139 "glycoside catabolic process" GO:0006004 "fucose metabolic process"	**34.3** GO:2001037 "positive regulation of tongue muscle cell differentiation" GO:0034653 "retinoic acid catabolic process" GO:0048387 "negative regulation of retinoic acid receptor signaling pathway" GO:0043587 "tongue morphogenesis" GO:0001768 "establishment of T cell polarity" GO:0033189 "response to vitamin A" GO:0061436 "establishment of skin barrier" GO:0009954 "proximal/distal pattern formation" GO:0001709 "cell fate determination" GO:0007140 "male meiotic nuclear division" GO:0071300 "cellular response to retinoic acid" GO:0060349 "bone morphogenesis" GO:0070268 "cornification" GO:0006805 "xenobiotic metabolic process" GO:0030326 "embryonic limb morphogenesis" GO:0045580 "regulation of T cell differentiation" GO:0016125 "sterol metabolic process"	**2.3** GO:0035990 "tendon cell differentiation" GO:0030948 "negative regulation of vascular endothelial growth factor receptor signaling pathway" GO:0001886 "endothelial cell morphogenesis" GO:0001937 "negative regulation of endothelial cell proliferation" GO:0016525 "negative regulation of angiogenesis" GO:0071773 "cellular response to BMP stimulus"

- cont. on the next page -

Upregulated modules		Downregulated modules	
SARs-CoV-2	**zika**	**SARs-CoV-2**	**zika**
13.1 GO:0000470 "maturation of LSU-rRNA" GO:0000460 "maturation of 5.8S rRNA"	**4.1** GO:0021522 "spinal cord motor neuron differentiation"	**26.1** GO:0003408 "optic cup formation involved in camera-type eye development" GO:0090403 "oxidative stress-induced premature senescence" GO:0090399 "replicative senescence" GO:0060317 "cardiac epithelial to mesenchymal transition" GO:0043616 "keratinocyte proliferation" GO:0014068 "positive regulation of phosphatidylinositol 3-kinase signaling" GO:0046849 "bone remodeling" GO:0046330 "positive regulation of JNK cascade"	**24** GO:0006024 "glycosaminoglycan biosynthetic process"
		8.2 GO:2000535 "regulation of entry of bacterium into host cell" GO:0016139 "glycoside catabolic process" GO:0006004 "fucose metabolic process"	**31** GO:2001199 "negative regulation of dendritic cell differentiation"

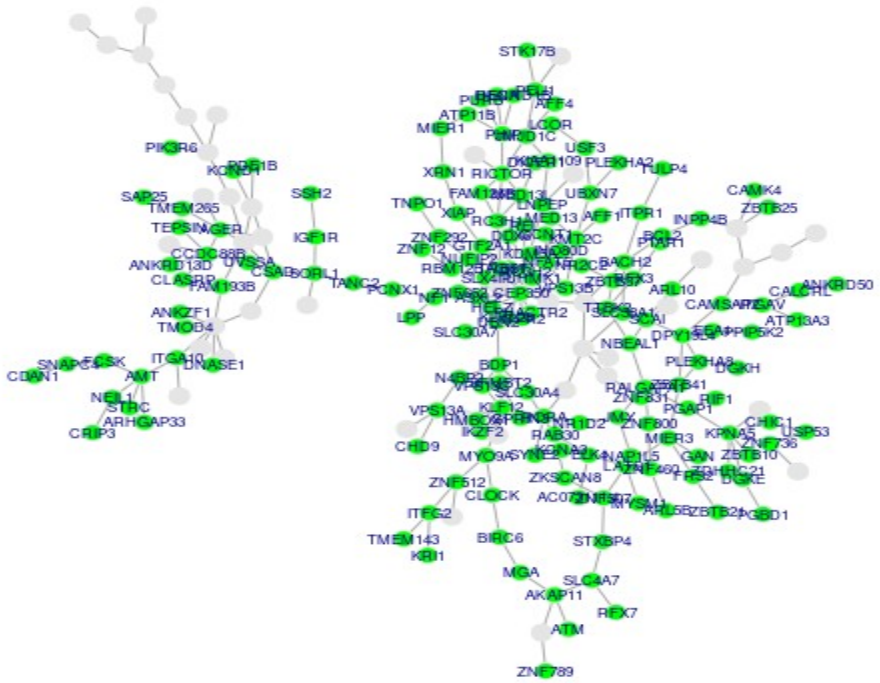

Figure 18: The largest subnetwork of genes from the DE genes in the experiment, both for SARs-CoV-2 and for zika. Marked in green are all the genes that are significantly upregulated or downregulated in SARs-CoV-2, but are non-significantly regulated in zika vaccine.

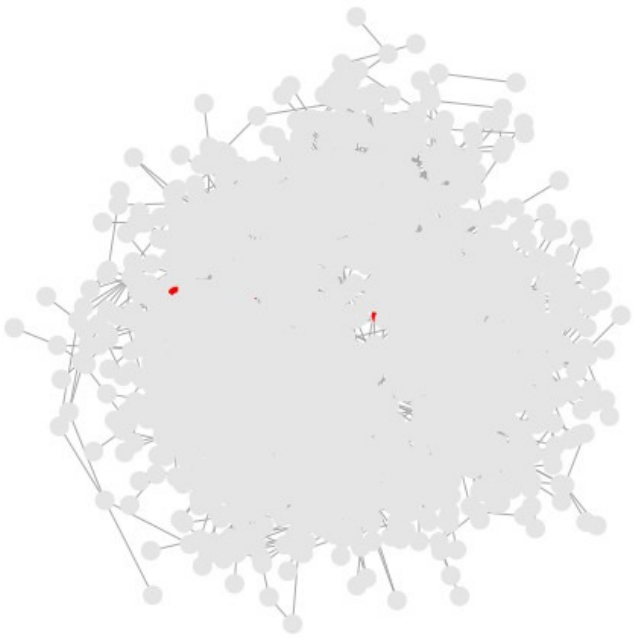

Figure 19: A dense subnetwork obtained when plotting DE genes for the time point DPC1. All the names of the genes have been removed to show just the network. The red nodes represent upregulated genes.