Isabelle Franco moscardini

UNIVERSITÀ
DI SIENA
1240

University of Siena

Department of Medical Biotechnologies

PhD Programme in Medical Biotechnologies

XXXV Cycle

SYSTEMS BIOLOGY AND TRANSCRIPTOMIC ANALYSIS TO STUDY THE IMMUNE RESPONSE TO INFECTIONS AND VACCINES

Supervisor:
Prof. Francesco Santoro

Co-supervisor:
Dr. Alice Gerlini

Isabelle Franco moscardini

PhD Candidate:
Isabelle Franco Moscardini

Academic Year: 2022/2023

INDEX

ABSTRACT

Given the complexity of living systems, and the difficulty of measuring and interpreting data from these systems, biomedical science has been adopting a reductionist approach over the years. However, the rapid technological advances and the progress in molecular biology and computation are changing this establishment.

Transcriptomics is one of the technologies that has revolutionized the way we study the response of organisms to various situations, such as infections, vaccines and cancer. By measuring the changes in the gene expression, we can capture important information about the pathophysiology of diseases, mechanism of action of vaccines, among other biological processes. By integrating transcriptomic data with cytokines, for instance, we uncovered a systemic recall immune response to lung pneumococcal infection, describing the main factors driving this process.

By combining systems biology and Machine Learning algorithms, the biological signatures of three different cohorts studying the recombinant vesicular stomatitis virus vaccine against Ebola were compared. We showed that different methods can capture distinct signatures, especially when the molecular perturbation is less evident. The use of Feature Selection and Machine learning algorithms can help us to focus on a gene level characterization, which is an important feature in the precision medicine era.

Finally, in this work transcriptomics has also contributed to characterize the response to a mucosal immunization with a recombinant bacteria expressing the CTH522, a *Chlamydia trachomatis* antigen. We have shown that the intravaginal priming with the recombinant vector modulated the systemic response to the antigen, using a model of splenocytes *in vitro* stimulated after different immunization schedules.

Rather than focus on a specific vaccine or infection, the aim of this thesis was to explore the range of tools available for the analysis of transcriptomics data in a systems biology perspective. Using data from different studies, involving both experimental models and clinical studies, the thesis offered a great opportunity to approach different themes and leverage different tools to deal with the challenges of extracting meaningful biological information from large data sets.

ABBREVIATIONS

AUC: Area under the curve

BioFeatS: Biological Feature Selection tool

BTM: Blood Transcriptional Modules

CbpA: choline binding protein A

Chop: phosphorylcholine

CERNO: Coincident Extreme Ranks in Numerical Observations

CIM: Clustered Image Map

CPS: Capsular Polysaccharides

CWPS: Cell Wall Polysaccharides

DAMPs: Damage or Danger-associated Molecular Patterns

DE: Differential Expression

DEGs: Differentially Expressed Genes

DNA: Deoxyribonucleic acid

eIF2: eukaryotic translation initiation factor 2 complex

EVD: Ebola Virus Disease

FCS: Functional Class Scoring

FDR: False Discovery Rate

GSEA: Gene Set Enrichment Analysis

ICA: Independent Component Analysis

ICEBOV: Ivory Coast ebolavirus

IPCA: Independent Principal Component Analysis

IVAG: Intravaginal route

kNN: k-Nearest Neighbors

MAMPs: Microbe-associated Molecular Patterns

MDP: Molecular Degree of Perturbation

ML: Machine Learning

mRNA: messenger Ribonucleic acid

MOMP: Major outer membrane protein

MZ: marginal zone

ORA: Over Representation Analysis

RNA: Ribonucleic acid

OVA: Chicken egg albumin

PAFR: Platelet-activating Factor Receptors

PAMPs: Pathogen-associated Molecular Patterns

PBMC: Peripheral blood mononuclear cells

PCA: Principal component Analysis

PcpA: Pneumococcal choline-binding protein A

PCVs: pneumococcal conjugate vaccines

PLS: Partial Least Squares

PLS-DA: Partial Least Squares - Discriminant Analysis

PPI: Protein-Protein Interaction

PPV: Polysaccharide Pneumococcal Vaccines

PRRs: Pattern Recognition Receptors

PVCA: Principal Variance Component Analysis

REBOV: Reston ebolavirus

RFE: Recursive Feature Elimination

SEBOV: Sudan ebolavirus

sPLS: sparse Partial Least Squares

sPLS-DA: sparse Partial Least Squares - Discriminant Analysis

SVM: Support Vector Machine

URT: Upper Respiratory Tract

VCA: Variance Components Analysis

VSV: Vesicular Stomatitis Virus

WGCNA: Weighted correlation network analysis

WT: Wild-Type

ZEBOV: Zaire ebolavirus

CHAPTER 1

An introduction to Systems Biology: the origin, its applications and limitations

"Systems biology... is about putting together rather than taking apart, integration rather than reduction. It requires that we develop ways of thinking about integration that are as rigorous as our reductionist programmes, but different.... It means changing our philosophy, in the full sense of the term"
(Noble, 2006)

1.1 The origin of Systems Biology

The beginning

Despite the exponential growth of Systems Biology applications in recent years, the importance of the holistic view has been known for a long time (Erickson, 2007; Harte, 2002). However, because of the extreme complexity of living systems, the reductionist approach has been driving biological science over the years (Wellstead et al., 2008). Then, scientists started to realize that understanding the system's behavior and why the components interact together is as important as showing how they interact in a reductionist approach (Trewavas, 2006).

This process of change began in the late 1960s, with the application of Systems Theory to biology. At this point, Systems Theory was defined as "the theory of formal (mathematical) models of real-life (or conceptual) systems" (Mesarović, 1968). An example is the development of the Biochemical Systems Theory in the same period, given the growing demand for means to model the nonlinear behavior of biological systems (Savageau, 1969a, 1969b, 1970).

In the 1970s and 1980s, Robert Rosen studied the transfer of concepts from physics to biology and emphasized that biological systems are a special case of physical systems (Rosen, 1978, 1985). Some years later, Reilly and collaborators described System analyses as a method that could be applied to different levels of the biological hierarchy, from molecular and cellular components, to whole organisms, populations and ecosystems. However, at that

time, in 1994, the authors highlighted that the research oriented across system levels was not yet part of the standard disease research (Reilly et al., 1994).

The advances in technology and molecular biology culminated in the Human Genome Project during the 1990s and, later, enabled us to quantify mRNAs, proteins, and metabolites on a large scale. These new technological features allowed Systems Biology to be increasingly established as a new field (Ideker et al., 2001).

H. Kitano emphasized the importance and contributions of Molecular Biology to Systems Biology and described this field as a way to understand biological systems at a system level, through three main approaches: Bottom-Up, Top-Down, and Hybrid. The Bottom-up approach tries to establish gene regulatory networks from independent experiment data, while the Top-down seeks to find meaning in data from high-throughput measurements. The Hybrid, as the name suggests, is a mix of both approaches. Kitano also highlighted the difficulty of establishing network structures from Top-down approaches, as well the need to continuously develop new methods (Kitano, 2000).

By that time, the Systems Biology approach was already presented with huge expectations as a method that would support scientific discoveries in the most diverse areas. In the following years, we have watched systems biology arise and consolidate as a new biological field and not just a simple branch of physical or mathematical systems.

ISABELLE FRANCO MOSCARDINI
07/09/2023 13:53:04

Recent years

Isabelle Franco moscardini

The new century arrived bringing important technological and computational advancements, permitting the establishment of Systems Biology and its application to countless research areas, including infectious diseases, vaccination, chronic diseases and cancer.

Systems biology became a very dynamic area, with new methods and technologies constantly arising. In only a few years we jumped from a limited availability of less specific methods to a multitude of high-throughput technologies, capable of measuring changes at

different biological levels, including the expression of genes, proteins, microRNAs, and the production of cytokines and metabolites, generating what we today address as OMIC data. Moreover, the advances in biology could be wrapped up in huge databases, and new computational methods evolved to permit data analysis and integration.

1.2 Transcriptomics and its role in Systems Biology

Among the different biological layers studied within Systems Biology approach, transcriptomics has been one of the most used, thanks to its high coverage and relatively easy assessment (Hagan et al., 2015). Even though studies usually focus on the messenger RNAs (mRNAs), by definition, transcriptomics is the study of the whole range of the RNA transcripts that are produced by the genotype, and their quantities, for a specific stage or condition (Lowe et al., 2017; Milward et al., 2016). Thus, it includes not only mRNAs but also noncoding RNAs, such as long non-coding RNA and micro-RNAs.

The Central Dogma of Molecular Biology was introduced by Francis Crick, establishing the idea of information transfer from the genome (DNA) to messenger RNAs (mRNAs) through transcription, and then to amino-acid chains through translation (Crick, 1970). Half a century later, we now understand that the way between genotype and phenotype is a very complex road composed of many different steps and specific regulation points. Different mechanisms such as alternative splicing, epigenetics and regulation by non-coding RNAs arised as influencing factors of this process (Jafari et al., 2017).

The phenotypic diversity observed in our genetically identical cells is linked to the fact that each cell expresses a different set of genes, presenting different transcriptomes (Morozova et al., 2009). Both genetic and environmental factors contribute to the transcriptional activity, and assessing this information provides an overview of which cellular processes are taking place in a given situation (Chaussabel et al., 2010).

Despite the very complex scenario in which all these processes occur, the assessment of changes in gene expression at different conditions has been proved to be a valuable tool to

explore the molecular mechanisms ruling the host's responses to different conditions and perturbations. In the next topics, we will go through some applications of transcriptomics coupled with Systems Biology approach in studying immune responses to infection and vaccination.

1.3 Systems Biology and the study of the immune system

The response of organisms to vaccines, infections and other diseases involves many factors from different levels, including biological and environmental aspects. Systems biology offers frameworks and tools that enable us to deal with such complexity. This approach has possibilitated a better understanding of many aspects of the biomedical research, including host-pathogen interactions, disease progression, discovery of biomarkers and study of immune system development. Moreover, this approach allows a comprehensive look at the innate immunity and its bridging to adaptive immunity, which is a key point when studying the memory generated by vaccines and infections (Dhillon et al., 2020).

Application in the study of Infectious diseases

Systems biology approach has been extremely helpful in deeping our knowledge in the study of infectious diseases, especially for its ability to capture patterns in complex circumstances, where the usual targeted methods struggle in going deeper into the mechanisms. Previous studies have advanced our understanding on the host factors that could affect the susceptibility to developing disease in human challenge models, correlating gene expression at baseline, or very early after challenge, with clinical outcomes (Humphreys et al., 2007; Yang et al., 2016).

Thanks to the possibility of assessing a greater number of variables we are able to look at in a single experiment, transcriptomics has enabled us to gain insights on molecular mechanisms related to the immune responses and disease progression. Chikungunya (Soares-Schanoski et al., 2019), Dengue (Kwissa et al., 2014), Malaria (Tran et al., 2019) and

COVID-19 (Islam and Khan, 2020; Melms et al., 2021) are examples of diseases for which this approach has allowed a more comprehensive understanding of the pathogenesis.

Dual transcriptomics experiments (simultaneous profiling of host and pathogen transcriptomics) have been very useful in the study of infection biology and host-pathogen interactions. Successful examples include the study of the pathogenesis of *Streptococcus pneumoniae* (D'Mello et al., 2020; Minhas et al., 2020; Ritchie and Evans, 2019), *Chlamydia trachomatis* (Humphrys et al., 2013), *Haemophilus influenzae* (Baddal et al., 2015) and many others (Westermann et al., 2017).

In the same direction, transcriptomic studies have been used to expand our knowledge regarding the heterogeneity of symptoms observed in many diseases. In typhoid fever, for instance, inflammation and innate immunity pathways were found correlated with the severity of symptoms (Blohmke et al., 2016). For influenza infection, not only Interferon and pattern recognition pathways were enriched in patients with moderate/severe disease, but a signature of 6 interferon genes was able to distinguish the groups with 100% accuracy (Davenport et al., 2015). Recently, the Systems Biology approach was also applied to understand the signature of disease severity in COVID-19 (Arunachalam et al., 2020).

Insights empowered by computation can and should be further complemented and validated using appropriate methods. Aguiar et al. combined different OMIC data, histopathological results and the validation with immunohistochemistry to uncover important alterations in the brains of babies who developed Congenital Zika syndrome (Aguiar et al., 2020).

ISABELLE FRANCO MOSCARDINI
07/09/2023 14:49:55

Isabelle Franco moscardini

Application in the study of vaccines

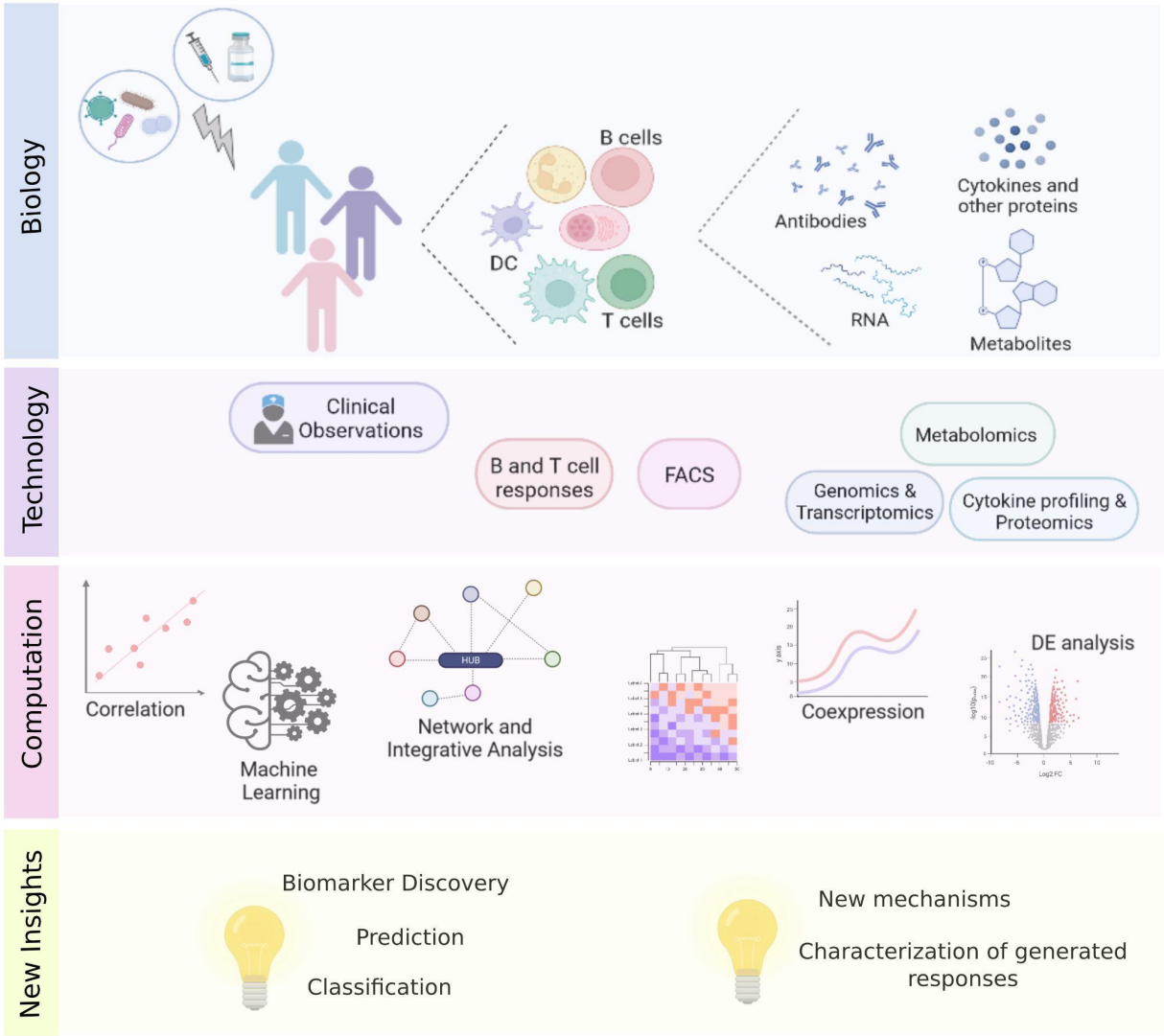
The use of Systems Biology in vaccinology became so widespread in the last few years, that a new term emerged to refer to this new area: Systems Vaccinology. The first application in this field has successfully identified gene signatures correlated with the antibody and $CD8^+$ T responses to Yellow Fever vaccine, suggesting that transcriptomics in

the early days after vaccination is a plausible method to predict the immunogenicity of vaccines (Querec et al., 2009). From then on, this approach allowed the characterization of the immune responses to different vaccines, including Influenza (Nakaya et al., 2011), Malaria (Vahey et al., 2010; Kazmin et al., 2017), HIV (Ehrenberg et al., 2019; Zak et al., 2012) and Ebola (Rechtien et al., 2017; Santoro et al., 2021), in many cases enabling the discovery of gene signatures linked to the antibody response or protection against disease.

Data integration and network analysis permitted the establishment of context-specific gene modules, reconstructed from 540 publicly available data series, originating the blood transcription modules (BTMs). Thanks to its comprehensive approach, BTMs were able to unveil new hypotheses on the mechanisms of vaccine induced immune responses (Li et al., 2014). Since then, this framework has been used along different statistical methods commonly used for pathway analysis to characterize immune responses and find correlations with immunological data.

This approach has also contributed to the molecular characterization of different vaccine adjuvants, providing insights on the mechanism of action of these substances and enabling a more reasoned direction for their use for the future vaccines (Olafsdottir et al., 2016).

Furthermore, systems vaccinology has also shown its usefulness in uncovering previously unknown mechanisms involved in vaccine responses by highlighting the importance of the microbiome during the generation of antibodies. One example is the correlation found between bacterias in the stool and skin prior vaccination to the levels of neutralizing antibodies after administration of a vaccine against HIV. A second important example is the discovery of how important the gut microbiota is for the immune response to the trivalent inactivated Influenza vaccine (Gonçalves et al., 2021; Oh et al., 2014).



ISABELLE FRANCO MOSCARDINI
07/09/2023 14:51:06

Figure 1. Systems Biology Overview. Three main components work together in the Systems Biology approach: Biology, technology and Computation. The biological aspect includes all the aspects of the system that are perturbed in a specific condition. It also includes all our knowledge in molecular biology and immunology, not only to interpret results but also to allow the improvements in the employed technologies. The technological aspect includes all the equipment needed to measure the changes occurred in the biological level. Finally, the computational aspect includes all the mathematical and statistical methods needed to assess and analyze information provided by the technological level, bringing meaningful information from complex data sets, leading to new insights and the formulation of hypotheses. (Adapted from Li et al., 2013)

Isabelle Franco moscardini

1.4 The limitations and challenges

Systems Biology is a recent and interdisciplinary field, combining three different perspectives: the biological, the technological and the computational, as illustrated in Figure 1. All of these three faces need to work together and all of them present challenges and limitations.

Biological variation, for example, is an important challenge to overcome, especially when analyzing human data. Host's factors like sex, age, presence of chronic diseases and life style (diet, exercise, smoking, etc) can have an huge impact in the biological system and influence how organisms respond to perturbations like infection and vaccination (Duraisingham et al., 2012; Kau et al., 2011; Orrù et al., 2013; Woods et al., 2009). Moreover, larger sample sizes could improve statistical power and reduce the impact of these variabilities, but due to the high cost and ethical and logistical constraints, the number of available samples is usually small (Hagan et al., 2015).

Limitations of transcriptomics include the assessment of a mix of different cell types, often being impossible to address the observed response to a specific cell type. In addition, enrichment analysis provides insights on the biological pathways activated in a specific condition, but it cannot provide cause-and-effect relationships, being extremely challenging to identify upstream processes and responses caused by technical artifacts (Barton et al., 2017). Thus, further investigation and validation of the findings obtained through computational analysis are of extreme importance.

From a technological point of view, the measurements made by high-throughput technologies can present varying degrees of noise. This affects especially the reliance on the detection of milder signals, that are not necessarily irrelevant. In fact, some low expressed transcripts or features present in very little amounts are known to have significant biological impact, such as transcription factors and long non-coding RNAs (Kornienko et al., 2016; Spitz and Furlong, 2012). Another very common technical challenge is the presence of batch effects, a source of bias that can arise due to different factors, such as the capacity of

machines, changes in the experimental conditions and the use of data obtained in different machines (Leek et al., 2010).

The computational methods need to walk together with technological development, providing ways to assess the information extracted and helping to attribute biological meaning to complex data. It is not an easy task and cases where the current methods were refuted or unintentionally misused have not been rare in the recent past (Li et al., 2022; Zhao et al., 2020).

Despite the limitations described above, the advancements in technology, computation and biology are emerging ever faster. New sophisticated technologies are constantly rising, the substitution of Microarray technology for RNA sequencing and now the possibility of performing single-cell RNA sequencing are examples of this process. Statistical methods are constantly being updated, focusing on overcoming biological and technical limitations, allowing, for instance, the identification and removal of batch effects (Leek et al., 2010; Papiez et al., 2019). Moreover, the accumulation of public data allows scientists to perform meta-analysis studies, increasing the strength of the hypothesis.

In a scientific context, understanding the challenges and limitations of available methods is of extreme importance to make the most of your data and decrease the risks of forming a wrong hypothesis. When consciously implemented, Systems Biology is a powerful tool that has been enabling a different look at biological processes, giving insights on molecular mechanisms, permitting the discovery of biomarkers and supporting the development of new treatments and vaccines.

ISABELLE FRANCO MOSCARDINI
07/09/2023 14:53:44

Isabelle Franco moscardini

1.5 The future

Two decades after the completion of the Human Genome Project, which has taken 13 years and around 3 billion dollars to sequence a single human genome, the scenario of sequencing technologies has completely changed. Today, technologies like Illumina can generate 200-6000 gigabase sequences per run, at an incomparably lower cost. A huge

progress in the detection of proteins, cytokines and metabolites was also observed (Veenstra, 2021).

The technological advances and the decreasing in the costs contributed to the exponential growth of systems biology applications observed in the past years. Despite the great progress observed in different fields, a broad knowledge on how cells and organisms react to internal and external events, which would enable the full prediction of how systems function, is yet to be achieved (Veenstra, 2021). However, it is highly probable that we keep seeing the exponential emergence of new methodologies, allowing more comprehensive and accurate measurements, taking us closer and closer to this goal.

1.6 Aim of the thesis

Among the range of new possibilities emerging along transcriptomics and Systems Biology, the aim of this work was to underline the use of these approaches to understand host's responses to infection and vaccination, with focus on the immune system. Through the analysis of different datasets, this work made use of a variety of available methods and frameworks to analyze and integrate transcriptomics data, extracting biological knowledge from complex datasets.

More specifically, this thesis aimed to use transcriptomics analysis and Systems Biology approach in three main contexts:

1. To study the systemic responses and the immune memory generated in a model of *Streptococcus pneumoniae* lung infection
2. To compare the transcriptomics profiles of the rVSV-ZEBOV vaccination in three independent cohorts, using two different approaches
3. To characterize the host's responses induced by the immunization with an engineered *S. gordonii* expressing the *Chlamydia trachomatis* CTH522 antigen

References

Aguiar, R.S., Pohl, F., Morais, G.L., Nogueira, F.C.S., Carvalho, J.B., Guida, L., Arge, L.W.P., Melo, A., Moreira, M.E.L., Cunha, D.P., Gomes, L., Portari, E.A., Velasquez, E., Melani, R.D., Pezzuto, P., de Castro, F.L., Geddes, V.E.V., Gerber, A.L., Azevedo, G.S., Schamber-Reis, B.L., Gonçalves, A.L., Junqueira-de-Azevedo, I., Nishiyama, M.Y., Ho, P.L., Schanoski, A.S., Schuch, V., Tanuri, A., Chimelli, L., Vasconcelos, Z.F.M., Domont, G.B., Vasconcelos, A.T.R., Nakaya, H.I., 2020. Molecular alterations in the extracellular matrix in the brains of newborns with congenital Zika syndrome. Sci. Signal. 13, eaay6736. https://doi.org/10.1126/scisignal.aay6736

Arunachalam, P.S., Wimmers, F., Mok, C.K.P., Perera, R.A.P.M., Scott, M., Hagan, T., Sigal, N., Feng, Y., Bristow, L., Tak-Yin Tsang, O., Wagh, D., Coller, J., Pellegrini, K.L., Kazmin, D., Alaaeddine, G., Leung, W.S., Chan, J.M.C., Chik, T.S.H., Choi, C.Y.C., Huerta, C., Paine McCullough, M., Lv, H., Anderson, E., Edupuganti, S., Upadhyay, A.A., Bosinger, S.E., Maecker, H.T., Khatri, P., Rouphael, N., Peiris, M., Pulendran, B., 2020. Systems biological assessment of immunity to mild versus severe COVID-19 infection in humans. Science 369, 1210–1220. https://doi.org/10.1126/science.abc6261

Baddal, B., Muzzi, A., Censini, S., Calogero, R.A., Torricelli, G., Guidotti, S., Taddei, A.R., Covacci, A., Pizza, M., Rappuoli, R., Soriani, M., Pezzicoli, A., 2015. Dual RNA-seq of Nontypeable Haemophilus influenzae and Host Cell Transcriptomes Reveals Novel Insights into Host-Pathogen Cross Talk. mBio 6, e01765-15. https://doi.org/10.1128/mBio.01765-15

Barton, A.J., Hill, J., Pollard, A.J., Blohmke, C.J., 2017. Transcriptomics in Human Challenge Models. Front. Immunol. 8, 1839. https://doi.org/10.3389/fimmu.2017.01839

Blohmke, C.J., Darton, T.C., Jones, C., Suarez, N.M., Waddington, C.S., Angus, B., Zhou, L., Hill, J., Clare, S., Kane, L., Mukhopadhyay, S., Schreiber, F., Duque-Correa, M.A., Wright, J.C., Roumeliotis, T.I., Yu, L., Choudhary, J.S., Mejias, A., Ramilo, O., Shanyinde, M., Sztein, M.B., Kingsley, R.A., Lockhart, S., Levine, M.M., Lynn, D.J., Dougan, G., Pollard, A.J., 2016. Interferon-driven alterations of the host's amino acid metabolism in the pathogenesis of typhoid fever. J. Exp. Med. 213, 1061–1077. https://doi.org/10.1084/jem.20151025

Chaussabel, D., Pascual, V., Banchereau, J., 2010. Assessing the human immune system through blood transcriptomics. BMC Biol. 8, 84. https://doi.org/10.1186/1741-7007-8-84

Crick, F., 1970. Central Dogma of Molecular Biology. Nature 227, 561–563. https://doi.org/10.1038/227561a0

Davenport, E.E., Antrobus, R.D., Lillie, P.J., Gilbert, S., Knight, J.C., 2015. Transcriptomic profiling facilitates classification of response to influenza challenge. J. Mol. Med. 93, 105–114. https://doi.org/10.1007/s00109-014-1212-8

Dhillon, B.K., Smith, M., Baghela, A., Lee, A.H.Y., Hancock, R.E.W., 2020. Systems Biology Approaches to Understanding the Human Immune System. Front. Immunol. 11, 1683. https://doi.org/10.3389/fimmu.2020.01683

D'Mello, A., Riegler, A.N., Martínez, E., Beno, S.M., Ricketts, T.D., Foxman, E.F., Orihuela, C.J., Tettelin, H., 2020. An in vivo atlas of host–pathogen transcriptomes during *Streptococcus pneumoniae* colonization and disease. Proc. Natl. Acad. Sci. 117, 33507–33518. https://doi.org/10.1073/pnas.2010428117

Duraisingham, S.S., Rouphael, N., Cavanagh, M.M., Nakaya, H.I., Goronzy, J.J., Pulendran, B., 2012. Systems Biology of Vaccination in the Elderly, in: Katze, M.G. (Ed.), Systems Biology, Current Topics in Microbiology and Immunology. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 117–142. https://doi.org/10.1007/82_2012_250

Ehrenberg, P.K., Shangguan, S., Issac, B., Alter, G., Geretz, A., Izumi, T., Bryant, C., Eller, M.A., Wegmann, F., Apps, R., Creegan, M., Bolton, D.L., Sekaly, R.P., Robb, M.L., Gramzinski, R.A., Pau, M.G., Schuitemaker, H., Barouch, D.H., Michael, N.L., Thomas, R., 2019. A vaccine-induced gene expression signature correlates with protection against SIV and HIV in

multiple trials. Sci. Transl. Med. 11, eaaw4236. https://doi.org/10.1126/scitranslmed.aaw4236
Erickson, H.L., 2007. Philosophy and Theory of Holism. Nurs. Clin. North Am. 42, 139–163. https://doi.org/10.1016/j.cnur.2007.03.001
Gonçalves, E., Guillén, Y., Lama, J.R., Sanchez, J., Brander, C., Paredes, R., Combadière, B., 2021. Host Transcriptome and Microbiota Signatures Prior to Immunization Profile Vaccine Humoral Responsiveness. Front. Immunol. 12, 657162. https://doi.org/10.3389/fimmu.2021.657162
Hagan, T., Nakaya, H.I., Subramaniam, S., Pulendran, B., 2015. Systems vaccinology: Enabling rational vaccine design with systems biological approaches. Vaccine 33, 5294–5301. https://doi.org/10.1016/j.vaccine.2015.03.072
Harte, V., 2002. Plato on Parts and Wholes. Oxford University Press. https://doi.org/10.1093/0198236751.001.0001
Humphreys, T.L., Li, L., Li, X., Janowicz, D.M., Fortney, K.R., Zhao, Q., Li, W., McClintick, J., Katz, B.P., Wilkes, D.S., Edenberg, H.J., Spinola, S.M., 2007. Dysregulated Immune Profiles for Skin and Dendritic Cells Are Associated with Increased Host Susceptibility to *Haemophilus ducreyi* Infection in Human Volunteers. Infect. Immun. 75, 5686–5697. https://doi.org/10.1128/IAI.00777-07
Humphrys, M.S., Creasy, T., Sun, Y., Shetty, A.C., Chibucos, M.C., Drabek, E.F., Fraser, C.M., Farooq, U., Sengamalay, N., Ott, S., Shou, H., Bavoil, P.M., Mahurkar, A., Myers, G.S.A., 2013. Simultaneous Transcriptional Profiling of Bacteria and Their Host Cells. PLoS ONE 8, e80597. https://doi.org/10.1371/journal.pone.0080597
Ideker, T., Galitski, T., Hood, L., 2001. A new approach to decoding life: systems biology. Annu. Rev. Genomics Hum. Genet. 2, 343–372. https://doi.org/10.1146/annurev.genom.2.1.343
Islam, A.B.M.Md.K., Khan, Md.A.-A.-K., 2020. Lung transcriptome of a COVID-19 patient and systems biology predictions suggest impaired surfactant production which may be druggable by surfactant therapy. Sci. Rep. 10, 19395. https://doi.org/10.1038/s41598-020-76404-8
Jafari, M., Ansari-Pour, N., Azimzadeh, S., Mirzaie, M., 2017. A logic-based dynamic modeling approach to explicate the evolution of the central dogma of molecular biology. PLOS ONE 12, e0189922. https://doi.org/10.1371/journal.pone.0189922
Kau, A.L., Ahern, P.P., Griffin, N.W., Goodman, A.L., Gordon, J.I., 2011. Human nutrition, the gut microbiome and the immune system. Nature 474, 327–336. https://doi.org/10.1038/nature10213
Kazmin, D., Nakaya, H.I., Lee, E.K., Johnson, M.J., van der Most, R., van den Berg, R.A., Ballou, W.R., Jongert, E., Wille-Reece, U., Ockenhouse, C., Aderem, A., Zak, D.E., Sadoff, J., Hendriks, J., Wrammert, J., Ahmed, R., Pulendran, B., 2017. Systems analysis of protective immune responses to RTS,S malaria vaccination in humans. Proc. Natl. Acad. Sci. 114, 2425–2430. https://doi.org/10.1073/pnas.1621489114
Kitano, H., 2000. Perspectives on systems biology. New Gener. Comput. 18, 199–216. https://doi.org/10.1007/BF03037529
Kornienko, A.E., Dotter, C.P., Guenzl, P.M., Gisslinger, H., Gisslinger, B., Cleary, C., Kralovics, R., Pauler, F.M., Barlow, D.P., 2016. Long non-coding RNAs display higher natural expression variation than protein-coding genes in healthy humans. Genome Biol. 17, 14. https://doi.org/10.1186/s13059-016-0873-8
Kwissa, M., Nakaya, H.I., Onlamoon, N., Wrammert, J., Villinger, F., Perng, G.C., Yoksan, S., Pattanapanyasat, K., Chokephaibulkit, K., Ahmed, R., Pulendran, B., 2014. Dengue Virus Infection Induces Expansion of a CD14+CD16+ Monocyte Population that Stimulates Plasmablast Differentiation. Cell Host Microbe 16, 115–127. https://doi.org/10.1016/j.chom.2014.06.001
Leek, J.T., Scharpf, R.B., Bravo, H.C., Simcha, D., Langmead, B., Johnson, W.E., Geman, D., Baggerly, K., Irizarry, R.A., 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat. Rev. Genet. 11, 733–739. https://doi.org/10.1038/nrg2825

Li, S., Nakaya, H.I., Kazmin, D.A., Oh, J.Z., Pulendran, B., 2013. Systems biological approaches to measure and understand vaccine immunity in humans. Semin. Immunol. 25, 209–218. https://doi.org/10.1016/j.smim.2013.05.003

Li, S., Rouphael, N., Duraisingham, S., Romero-Steiner, S., Presnell, S., Davis, C., Schmidt, D.S., Johnson, S.E., Milton, A., Rajam, G., Kasturi, S., Carlone, G.M., Quinn, C., Chaussabel, D., Palucka, A.K., Mulligan, M.J., Ahmed, R., Stephens, D.S., Nakaya, H.I., Pulendran, B., 2014. Molecular signatures of antibody responses derived from a systems biology study of five human vaccines. Nat. Immunol. 15, 195–204. https://doi.org/10.1038/ni.2789

Li, Y., Ge, X., Peng, F., Li, W., Li, J.J., 2022. Exaggerated false positives by popular differential expression methods when analyzing human population samples. Genome Biol. 23, 79. https://doi.org/10.1186/s13059-022-02648-4

Lowe, R., Shirley, N., Bleackley, M., Dolan, S., Shafee, T., 2017. Transcriptomics technologies. PLOS Comput. Biol. 13, e1005457. https://doi.org/10.1371/journal.pcbi.1005457

Melms, J.C., Biermann, J., Huang, H., Wang, Y., Nair, A., Tagore, S., Katsyv, I., Rendeiro, A.F., Amin, A.D., Schapiro, D., Frangieh, C.J., Luoma, A.M., Filliol, A., Fang, Y., Ravichandran, H., Clausi, M.G., Alba, G.A., Rogava, M., Chen, S.W., Ho, P., Montoro, D.T., Kornberg, A.E., Han, A.S., Bakhoum, M.F., Anandasabapathy, N., Suárez-Fariñas, M., Bakhoum, S.F., Bram, Y., Borczuk, A., Guo, X.V., Lefkowitch, J.H., Marboe, C., Lagana, S.M., Del Portillo, A., Tsai, E.J., Zorn, E., Markowitz, G.S., Schwabe, R.F., Schwartz, R.E., Elemento, O., Saqi, A., Hibshoosh, H., Que, J., Izar, B., 2021. A molecular single-cell lung atlas of lethal COVID-19. Nature 595, 114–119. https://doi.org/10.1038/s41586-021-03569-1

Mesarović, Mihajlo D., 1968. Systems Theory and Biology—View of a Theoretician, in: Mesarović, M. D. (Ed.), Systems Theory and Biology. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 59–87.

Milward, E.A., Shahandeh, A., Heidari, M., Johnstone, D.M., Daneshi, N., Hondermarck, H., 2016. Transcriptomics, in: Encyclopedia of Cell Biology. Elsevier, pp. 160–165. https://doi.org/10.1016/B978-0-12-394447-4.40029-5

Minhas, V., Aprianto, R., McAllister, L.J., Wang, H., David, S.C., McLean, K.T., Comerford, I., McColl, S.R., Paton, J.C., Veening, J.-W., Trappetti, C., 2020. In vivo dual RNA-seq reveals that neutrophil recruitment underlies differential tissue tropism of Streptococcus pneumoniae. Commun. Biol. 3, 293. https://doi.org/10.1038/s42003-020-1018-x

Morozova, O., Hirst, M., Marra, M.A., 2009. Applications of New Sequencing Technologies for Transcriptome Analysis. Annu. Rev. Genomics Hum. Genet. 10, 135–151. https://doi.org/10.1146/annurev-genom-082908-145957

Nakaya, H.I., Wrammert, J., Lee, E.K., Racioppi, L., Marie-Kunze, S., Haining, W.N., Means, A.R., Kasturi, S.P., Khan, N., Li, G.-M., McCausland, M., Kanchan, V., Kokko, K.E., Li, S., Elbein, R., Mehta, A.K., Aderem, A., Subbarao, K., Ahmed, R., Pulendran, B., 2011. Systems biology of vaccination for seasonal influenza in humans. Nat. Immunol. 12, 786–795. https://doi.org/10.1038/ni.2067

Noble, D., 2006. The music of life: biology beyond the genome. Oxford University Press, Oxford ; New York.

Oh, J.Z., Ravindran, R., Chassaing, B., Carvalho, F.A., Maddur, M.S., Bower, M., Hakimpour, P., Gill, K.P., Nakaya, H.I., Yarovinsky, F., Sartor, R.B., Gewirtz, A.T., Pulendran, B., 2014. TLR5-Mediated Sensing of Gut Microbiota Is Necessary for Antibody Responses to Seasonal Influenza Vaccination. Immunity 41, 478–492. https://doi.org/10.1016/j.immuni.2014.08.009

Olafsdottir, T.A., Lindqvist, M., Nookaew, I., Andersen, P., Maertzdorf, J., Persson, J., Christensen, D., Zhang, Y., Anderson, J., Khoomrung, S., Sen, P., Agger, E.M., Coler, R., Carter, D., Meinke, A., Rappuoli, R., Kaufmann, S.H.E., Reed, S.G., Harandi, A.M., 2016. Comparative Systems Analyses Reveal Molecular Signatures of Clinically tested Vaccine Adjuvants. Sci. Rep. 6, 39097. https://doi.org/10.1038/srep39097

Orrù, V., Steri, M., Sole, G., Sidore, C., Virdis, F., Dei, M., Lai, S., Zoledziewska, M., Busonero, F.,

Mulas, A., Floris, M., Mentzen, W.I., Urru, S.A.M., Olla, S., Marongiu, M., Piras, M.G., Lobina, M., Maschio, A., Pitzalis, M., Urru, M.F., Marcelli, M., Cusano, R., Deidda, F., Serra, V., Oppo, M., Pilu, R., Reinier, F., Berutti, R., Pireddu, L., Zara, I., Porcu, E., Kwong, A., Brennan, C., Tarrier, B., Lyons, R., Kang, H.M., Uzzau, S., Atzeni, R., Valentini, M., Firinu, D., Leoni, L., Rotta, G., Naitza, S., Angius, A., Congia, M., Whalen, M.B., Jones, C.M., Schlessinger, D., Abecasis, G.R., Fiorillo, E., Sanna, S., Cucca, F., 2013. Genetic Variants Regulating Immune Cell Levels in Health and Disease. Cell 155, 242–256. https://doi.org/10.1016/j.cell.2013.08.041

Papiez, A., Marczyk, M., Polanska, J., Polanski, A., 2019. BatchI: Batch effect Identification in high-throughput screening data using a dynamic programming algorithm. Bioinformatics 35, 1885–1892. https://doi.org/10.1093/bioinformatics/bty900

Querec, T.D., Akondy, R.S., Lee, E.K., Cao, W., Nakaya, H.I., Teuwen, D., Pirani, A., Gernert, K., Deng, J., Marzolf, B., Kennedy, K., Wu, H., Bennouna, S., Oluoch, H., Miller, J., Vencio, R.Z., Mulligan, M., Aderem, A., Ahmed, R., Pulendran, B., 2009. Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans. Nat. Immunol. 10, 116–125. https://doi.org/10.1038/ni.1688

Rechtien, A., Richert, L., Lorenzo, H., Martrus, G., Hejblum, B., Dahlke, C., Kasonta, R., Zinser, M., Stubbe, H., Matschl, U., Lohse, A., Krähling, V., Eickmann, M., Becker, S., Thiébaut, R., Altfeld, M., Addo, M., Agnandji, S.T., Krishna, S., Kremsner, P.G., Brosnahan, J.S., Bejon, P., Njuguna, P., Addo, M.M., Becker, S., Krähling, V., Siegrist, C.-A., Huttner, A., Kieny, M.-P., Moorthy, V., Fast, P., Savarese, B., Lapujade, O., 2017. Systems Vaccinology Identifies an Early Innate Immune Signature as a Correlate of Antibody Responses to the Ebola Vaccine rVSV-ZEBOV. Cell Rep. 20, 2251–2261. https://doi.org/10.1016/j.celrep.2017.08.023

Reilly, S.L., Sing, C.F., Savageau, M.A., Turner, S.T., 1994. Analysis of systems influencing renal hemodynamics and sodium excretion. I. Biochemical systems theory. Integr. Physiol. Behav. Sci. 29, 55–73. https://doi.org/10.1007/BF02691281

Ritchie, N.D., Evans, T.J., 2019. Dual RNA-seq in Streptococcus pneumoniae Infection Reveals Compartmentalized Neutrophil Responses in Lung and Pleural Space. mSystems 4, e00216-19. https://doi.org/10.1128/mSystems.00216-19

Rosen, R., 1985. Anticipatory Systems. Pergamon Press.

Rosen, R., 1978. Fundamentals of measurement and representation of natural systems. Elsevier Sci. Ltd. Vol 1.

Santoro, F., Donato, A., Lucchesi, S., Sorgi, S., Gerlini, A., Haks, M., Ottenhoff, T., Gonzalez-Dias, P., Consortium, Vsv-Ebovac, Consortium, Vsv-Eboplus, Nakaya, H., Huttner, A., Siegrist, C.-A., Medaglini, D., Pozzi, G., 2021. Human Transcriptomic Response to the VSV-Vectored Ebola Vaccine. Vaccines 9, 67. https://doi.org/10.3390/vaccines9020067

Savageau, M.A., 1970. Biochemical systems analysis. J. Theor. Biol. 26, 215–226. https://doi.org/10.1016/S0022-5193(70)80013-3

Savageau, M.A., 1969a. Biochemical systems analysis. J. Theor. Biol. 25, 365–369. https://doi.org/10.1016/S0022-5193(69)80026-3

Savageau, M.A., 1969b. Biochemical systems analysis. J. Theor. Biol. 25, 370–379. https://doi.org/10.1016/S0022-5193(69)80027-5

Soares-Schanoski, A., Baptista Cruz, N., de Castro-Jorge, L.A., de Carvalho, R.V.H., Santos, C.A. dos, Rós, N. da, Oliveira, Ú., Costa, D.D., Santos, C.L.S. dos, Cunha, M. dos P., Oliveira, M.L.S., Alves, J.C., Océa, R.A. de L.C., Ribeiro, D.R., Gonçalves, A.N.A., Gonzalez-Dias, P., Suhrbier, A., Zanotto, P.M. de A., Azevedo, I.J. de, Zamboni, D.S., Almeida, R.P., Ho, P.L., Kalil, J., Nishiyama, M.Y., Nakaya, H.I., 2019. Systems analysis of subjects acutely infected with the Chikungunya virus. PLOS Pathog. 15, e1007880. https://doi.org/10.1371/journal.ppat.1007880

Spitz, F., Furlong, E.E.M., 2012. Transcription factors: from enhancer binding to developmental control. Nat. Rev. Genet. 13, 613–626. https://doi.org/10.1038/nrg3207

Tran, T.M., Guha, R., Portugal, S., Skinner, J., Ongoiba, A., Bhardwaj, J., Jones, M., Moebius, J., Venepally, P., Doumbo, S., DeRiso, E.A., Li, S., Vijayan, K., Anzick, S.L., Hart, G.T., O'Connell, E.M., Doumbo, O.K., Kaushansky, A., Alter, G., Felgner, P.L., Lorenzi, H., Kayentao, K., Traore, B., Kirkness, E.F., Crompton, P.D., 2019. A Molecular Signature in Blood Reveals a Role for p53 in Regulating Malaria-Induced Inflammation. Immunity 51, 750-765.e10. https://doi.org/10.1016/j.immuni.2019.08.009

Trewavas, A., 2006. A Brief History of Systems Biology: "Every object that biology studies is a system of systems." Francois Jacob (1974). Plant Cell 18, 2420–2430. https://doi.org/10.1105/tpc.106.042267

Vahey, M.T., Wang, Z., Kester, K.E., Cummings, J., Heppner, Jr, D.G., Nau, M.E., Ofori‐Anyinam, O., Cohen, J., Coche, T., Ballou, W.R., Ockenhouse, C.F., 2010. Expression of Genes Associated with Immunoproteasome Processing of Major Histocompatibility Complex Peptides Is Indicative of Protection with Adjuvanted RTS,S Malaria Vaccine. J. Infect. Dis. 201, 580–589. https://doi.org/10.1086/650310

Veenstra, T.D., 2021. Omics in Systems Biology: Current Progress and Future Outlook. PROTEOMICS 21, 2000235. https://doi.org/10.1002/pmic.202000235

Wellstead, P., Bullinger, E., Kalamatianos, D., Mason, O., Verwoerd, M., 2008. The rôle of control and system theory in systems biology. Annu. Rev. Control 32, 33–47. https://doi.org/10.1016/j.arcontrol.2008.02.001

Westermann, A.J., Barquist, L., Vogel, J., 2017. Resolving host–pathogen interactions by dual RNA-seq. PLOS Pathog. 13, e1006033. https://doi.org/10.1371/journal.ppat.1006033

Woods, J.A., Vieira, V.J., Keylock, K.T., 2009. Exercise, Inflammation, and Innate Immunity. Immunol. Allergy Clin. North Am. 29, 381–393. https://doi.org/10.1016/j.iac.2009.02.011

Yang, W.E., Suchindran, S., Nicholson, B.P., McClain, M.T., Burke, T., Ginsburg, G.S., Harro, C.D., Chakraborty, S., Sack, D.A., Woods, C.W., Tsalik, E.L., 2016. Transcriptomic Analysis of the Host Response and Innate Resilience to Enterotoxigenic *Escherichia coli* Infection in Humans. J. Infect. Dis. 213, 1495–1504. https://doi.org/10.1093/infdis/jiv593

Zak, D.E., Andersen-Nissen, E., Peterson, E.R., Sato, A., Hamilton, M.K., Borgerding, J., Krishnamurty, A.T., Chang, J.T., Adams, D.J., Hensley, T.R., Salter, A.I., Morgan, C.A., Duerr, A.C., De Rosa, S.C., Aderem, A., McElrath, M.J., 2012. Merck Ad5/HIV induces broad innate immune activation that predicts CD8 $^{+}$ T-cell responses but is attenuated by preexisting Ad5 immunity. Proc. Natl. Acad. Sci. 109. https://doi.org/10.1073/pnas.1208972109

Zhao, S., Ye, Z., Stanton, R., 2020. Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. RNA 26, 903–909. https://doi.org/10.1261/rna.074922.120

CHAPTER 2

Methodologies in transcriptomics and Systems Biology

"The trick to being a scientist is to be open to using a wide variety of tools."
(Breiman, 2001)

To successfully achieve the objectives established in the first chapter of this thesis, a broad range of packages, tools and databases were used. This chapter elaborates on the available methods for the main topics on bioinformatics downstream analysis for transcriptomics, with a special focus on the resources used during the course of the PhD.

2.1 Exploratory Data Analysis and Quality control

2.1.1 Exploration of the descriptive data with DataExplorer R package

Before starting to analyze a dataset it is crucial to have a thorough understanding of the information and variables related to samples, especially when performing data integration and biomarker analysis. However, this step can be really time-consuming depending on the number of variables available.

In this context, packages and tools that summarize this type of information can be very useful. The DataExplorer R package is a great example, providing different functions to perform data exploration. Using different types of plots, DataExplorer examines the structure of the data, the type of each variable, the distribution of variables and the distribution of missing values.

In addition, this package also performs correlation analysis, Principal Component Analysis and allows the slicing of data in different ways, with boxplots and scatterplots for visualization. In a very useful way, the DataExplorer package performs most of the functionalities in only one personalized function, creating an html report with all the requested results (Boxuan Cui, n.d.).

2.1.2 Exploration of data variability

High-throughput technologies allow us to monitor thousands of variables in a single experiment. However, this comes with the challenge of dealing with high-dimensional data. In addition, the variability among samples can be attributed to several factors besides biological information. In this context, unsupervised techniques focused on dimensionality reduction are very useful. In particular, Principal component Analysis (PCA), Independent Principal Component Analysis (IPCA) and Principal Variance Component Analysis (PVCA) are examples of techniques used to explore the presence of experimental bias such as the presence of outliers, batch effects, and other sources of variability in the data. Moreover, these analyses also allow the identification of trends and patterns in the data.

Principal Component Analysis (PCA)

In a PCA the principal components are artificial variables, built as a linear combination of the initial variables. This results in the attribution of a weight value for each of the initial variables, which reflects the importance of the variable to the component. These weight values are stored in the loading vectors associated with each PC. Importantly, the components are uncorrelated and each of them aims to explain a maximal amount of variance, starting by the first one, which will capture the largest source of variation between samples. This allows an efficient dimension reduction while preserving the majority of the variability in the data (Hotelling, 1933). PCA can be performed in R by the *prcomp* function.

Independent Principal Component Analysis (IPCA)

However, sometimes PCA is not informative to the biological question for two main reasons: (i) the PCA assumes a multivariate normal distribution in the data, which is not always the case and (ii) the PCA decomposes data by maximizing its variance, but sometimes the main source of variance in the data is not related to the biological question. In these cases, the IPCA can be helpful.

The IPCA uses Independent Component Analysis (ICA) as a denoising process for PCA since the ICA is a good technique to reduce the effects of noise or artifacts of the signal. The assumption of the IPCA is that biological meaning can be highlighted by the components when most of the noise is removed from the loading vectors (Yao et al., 2012). IPCA is implemented by *mixOmics* R package.

Principal Variance Component Analysis (PVCA)

On the other hand, PVCA leverages the advantages of both PCA and Variance Components Analysis (VCA). The VCA fits a mixed linear model using factors of interest as random effects, aiming to understand the importance of each factor to the total variability (Boedigheimer et al., 2008). This is especially useful to estimate the importance of technical parameters and batch effects. PVCA is implemented in R by the *PVCA* package.

PCAtools: everything Principal Component Analysis

PCAtools (Kevin Blighe, n.d.) is an R package that provides exploratory functionalities for highly-dimensional data through PCA techniques, including biplots, loading plots and correlation of components with clinical data or even technical factors like known batch effects. Its main advantages comprise automated and high-quality report generation, with publication-ready figures, while providing optimization options for the user.

ISABELLE FRANCO MOSCARDINI
08/09/2023 13:55:17

2.1.3 Data exploration with MDP (Molecular Degree of Perturbation)

Isabelle Franco moscardini

MDP (Molecular Degree of Perturbation), described by Gonçalves et al., is a tool based on the Molecular Distance to Health (Pankla et al., 2009) algorithm which tries to quantify how perturbed samples from a given phenotype are compared to their controls, enabling us to measure the heterogeneity of a given dataset as well which genes are responsible for such perturbation.

This tool has been applied not only to compare the perturbation caused by different diseases but also to understand underlying differences among populations and disease severity (Oliveira-de-Souza et al., 2019). Moreover, this framework can be used to detect possible outliers, improving differential expression analysis results (Gonçalves et al., 2019).

2.1.4 Dealing with Batch effect

For logistical and practical reasons, genomic data is often produced in batches. The variation among different batches can result in discrepancies in the statistical distributions, which can directly impact the downstream analysis and the biological interpretation. While some packages for differential expression analysis, such as edgeR, have developed methods to deal with the differences across batches, for other downstream analyses this process needs to be done in advance.

Combat-Seq

ComBat is one of the most popular tools for adjusting known batch effects for downstream analysis when its sources are known (Johnson et al., 2007; Leek et al., 2010). ComBat-seq is an extension of the ComBat model that addresses several issues specific to bulk RNA-Seq data, such as using a negative binomial regression to model batch effects instead of a Gaussian distribution assumed by ComBat. Combat-seq allows the adjustment of the batches while preserving the integer nature of counts, which is compatible with most downstream analyses. However, the adjustment of batch effects should only be done when they are present and result in an unfavorable impact on downstream analysis (Zhang et al., 2020).

2.2 Differential Expression analyses

Differential Expression analysis is probably the most used approach to extract biological information from transcriptomic data. Through the application of statistical

procedures, the quantitative changes in the gene expression of individuals in two different conditions are evaluated. This process attributes a value concerning the magnitude of the increase or decrease in the expression (Fold-Change) and the associated p-value. Due to the extremely high number of statistical tests performed during this process, comparing the expression of thousands of transcripts, many false positive genes can be detected. Therefore, adjusting the p-values is a crucial step, and it is usually integrated into the DE analysis.

The R packages edgeR and DESeq2 are the two main used workflows for differential expression analysis in RNA sequencing data. Both are based on the negative binomial distribution, but they have their particularities. For instance, in estimating the dispersion, DESeq2 assumes a similar dispersion in genes with similar average expression strength over all samples, detecting and correcting dispersion estimates that are too low. On the other hand, edgeR will moderate the estimate of dispersion for each gene toward a common estimate across all genes, or toward a local estimate for genes with similar expression strength.

The defaults present differences as well. DESeq2 automatically finds the optimal value for filtering low expressed genes, spots the genes with outlier values, and excludes genes with very high within-group variance of the dispersion estimates. The edgeR package offers similar functionality in the estimateGLMRobustDisp function.

2.3 Enrichment and functional analysis

2.3.1 Gene sets

Enrichment analysis is a very used method to gain insights from transcriptomic data. The aim is to identify the processes that are activated or depleted, based on the expression of a set of genes, previously defined according to specific criteria. For instance, genes that are co-expressed, or genes that participate in a given biological process can assemble a gene set (Maleki et al., 2020). Gene sets can be organized in databases such as MSigDB, the Molecular Signatures Database (Subramanian et al., 2005).

Many gene sets are currently available. Among the most used ones, there are Reactome (Jassal et al., 2019) and KEGG (Kanehisa et al., 2016), providing information on the biological pathways and cellular processes. Gene Ontology (GO), describes the relation of genes with biological processes under three aspects: Molecular Function, Cellular Component and Biological Process (The Gene Ontology Consortium et al., 2021). Finally, the Blood Transcription Modules (BTM) combine gene ontology, cell type specific gene expression, interactome and bibliome to build sets of genes coexpressed under a particular biological condition (Li et al., 2014a).

2.3.2 Gene set enrichment analysis

Over Representation Analysis (ORA)

ORA is a commonly used method to perform enriched analysis and for this reason, it is implemented in a wide range of R packages. Based on a list of genes of interest, such as differentially expressed genes, the intersection of this list with a specific pathway in a gene set will be considered significant if this overlap could not be due to chance. For this, ORA uses a reference or background, which usually is the list of all the genes under study. The null hypothesis is that there is no association between differential expression and the specific gene set, assuming that the gene set is the result of a random sampling of genes from the background, and therefore, the probability of having differentially expressed genes in the gene set list can be calculated using the hypergeometric or binomial distribution and the significance of this association can be assessed by the Fisher's exact test (Drăghici et al., 2003; Maleki et al., 2020).

The tmod R package implements this approach through the tmodHGtest function (Weiner 3rd and Domaszewska, 2016). Many other packages implement ORA-based approaches for enrichment, with slight differences among them (Huang et al., 2009; Maleki et al., 2020).

Functional Class Scoring (FCS)

Despite the wide use of ORA approaches, the method has its limitations. Firstly, the need of setting thresholds to select a specific set of variables that are considered of interest, which might discard genes that do not fit the threshold, but could contribute to the understanding of the biological signatures in a data set. Secondly, ORA assumes that genes are independent, but actually, it is known that genes act in concert, especially when belonging to the same biological process.

Functional class scoring (FCS) deals with these factors by applying a statistical test based on the expression matrix, using the information provided by all the variables. There are two classes of FCS: univariate and multivariate methods. In univariate methods, there is the calculation of a score for each gene using its specific row of the expression matrix and then this score is used to calculate a gene set score and the significance of each gene set score is assessed. In multivariate methods, the gene set scores are calculated directly from the expression matrix, without the generation of gene scores.

The Gene Set Enrichment Analysis (GSEA), implemented through the fgsea R package, is among the most used FCS approaches. It is a univariate method based on the ranking of genes by the correlation between their expression and the distinct classes studied. (Subramanian et al., 2005). Following the same approach, the Single-sample GSEA (ssGSEA) is an extension of GSEA that calculates the enrichment scores separately for each pairing of a sample and gene set, representing the degree to which the genes of a given gene set are coordinately up- or down-regulated within a sample (Barbie et al., 2009)

As an example of multivariate FCS, the tmod package provides the CERNO test. The Coincident Extreme Ranks in Numerical Observations (CERNO) approach is based on Fisher's method of combining probabilities, a widely used approach to combine p-values in order to evaluate the presence of a global effect (Croze et al., 2013; Fisher, 1992). The CERNO test uses all the expression measurements, emphasizing the extreme changes within a dataset, and providing a measure of the strength of effect for each gene set. This method has

improvements in performance compared to the GSEA approach, especially with small sample sizes, a very common context in biomedical studies (Weiner 3rd and Domaszewska, 2016).

2.4 Data Integration

Data integration consists of the process of combining data that reside in different sources, to provide users with a unified view of such data. There are two main approaches: data-driven and knowledge-based. In the first one, statistical analysis, supervised and unsupervised methods are conducted without the need for strong prior knowledge. This includes correlations, clustering, dimension reduction and discriminant analysis by machine learning algorithms.

On the other hand, the knowledge-based approach requires strong literature, prior knowledge such as metabolic networks or mechanisms of enzyme kinetics are commonly used as a first step, focusing the analysis in a specific context. Then, statistical or machine learning methods are used to validate the models.

With the increase in data generation, *Multi-Omics* and *Omics and non-Omics* data integration have been gaining attention in the last few years. While the *Multi-Omics* is focused on integrating data from high-throughput experiments like transcriptomics, metabolomics and proteomics, the *Omics and non-omics* integrate these datasets with other types of data such as clinical, epidemiological or imaging. Despite the increase in the availability of computational methods to perform such analyses, the complexity of the methodologies and the presence of different limitations make these processes an ongoing challenge (Picard et al., 2021).

The Multi-Omics data integration is focused on analyzing numeric variables, such as two different Omic data sets measured in the same individuals. The main idea behind this process is that by the integration of different biological layers, the similarities between the data sets can be highlighted.

Partial Least Squares (PLS) and sparse Partial Least Squares (sPLS)

The PLS is a widely used supervised approach to integrating two numerical variables (matrices). In this linear multivariate visualization technique, the covariance between the components (linear combination of variables) of the two datasets is maximized and allows the modeling of the shared information that underlies the variables. It is a flexible method, able to deal with missing values and with the presence of correlations among independent variables, unlike traditional multiple regression models (Wold, Herman, 1966).

The mixOmics package implements not only the PLS but also its sparse version, the sPLS, with the aim of improving the interpretability of the method by including the Lasso penalization in each pair of loading vectors (Cao et al., 2008). The package also proposes a tuning parameter for the sPLS, performing feature selection by pinpointing the best number of features to be chosen in a specific data set.

DIABLO

The mixOmics package also proposes DIABLO - Data Integration Analysis for Biomarker discovery using Latent variable approaches for Omics studies. This approach leverages the generalization of the PLS method for multiple matching datasets. Therefore, DIABLO demands the same individuals in all datasets analyzed. The main objective of the tool is to identify, among the heterogeneous data, co-expressed variables. In a supervised way, this workflow can be applied to an outcome of interest, to uncover biological signatures.

DIABLO, like most of the mixOmics workflows, provides many graphical outputs to assist data interpretation, including networks of correlated features, circos plot and Clustered Image Map (CIM). Moreover, there is the possibility of tuning parameters and predicting an external test set (Singh et al., 2016).

Reactome GSA

One of the recently developed tools for multi-omics analysis is the ReactomeGSA R package and web tool (Griss et al., 2020). ReactomeGSA is a gene set analysis system that allows the comparative analysis of multiple independent datasets that are submitted together in a single pathway identification analysis, using the Reactome database. This method is suitable for both bulk and Single-cell level analysis and allows the direct integration with public data from ExpressionAtlas and Single Cell ExpressionAtlas.

2.5 Biomarker Discovery

In the new era of precision medicine, science has been fighting to identify biomarkers that could be useful in many fields, from diagnostics to the prediction of responses to a vaccine or a medication. In this aspect, the assessment of the expression of more than 20 thousand genes by transcriptomics offers a great opportunity to identify possible markers. This context has spurred the application of different methods and the creation of new frameworks to address this need.

DaMiR-seq

The DaMiR-seq package, described by Chiesa et al. offers a structured pipeline to perform data mining in R. The package receives the count data from RNA sequencing experiments, and besides the normalization of data, DaMiR-seq checks for the presence of outliers and unwanted sources of variation. Then, to improve classification performance and limit overfitting, redundant and irrelevant genes are excluded by feature selection approaches.

To identify the most informative variables capable of classifying samples according to the provided classes, the package uses an ensemble of different Machine Learning algorithms (Random Forest, Naïve Bayes, 3-Nearest Neighbours, Logistic Regression, Linear Discriminant Analysis, Support Vectors Machines, Neural Networks and Partial Least Squares). Since different datasets can fit differently into a model, weighting the prediction of

distinct classifiers and combining them may reduce the risk of classification errors (Polikar, 2006).

Moreover, the package offers many graphical outputs at each step of the process, allowing one to visually follow the pipeline and provide figures for the interpretation of the results.

PLS-DA and sPLS-DA

We have previously discussed the Partial Least Squares methodology in this chapter. Although PLS was developed to deal with numerical matrices, the flexibility of this method allows its application for discriminant analysis. In this case, a numeric matrix is integrated with a qualitative response (outcome), which will be treated as a continuous matrix. Despite being an adaptation, many studies have shown the method working in practice (Barker and Rayens, 2003; Boulesteix and Strimmer, 2006; Chung and Keles, 2010; Nguyen and Rocke, 2002).

The PLS-DA is implemented by the mixOmics R package, together with its sparse version (sPLS-DA), which also applies Lasso penalization in the loading vectors associated with the numerical matrix, performing feature selection.

2.6 Coexpression and Network analysis

Although different types of network analysis have been employed in Systems Biology, protein-protein interactions are the most extensively used. PPI are specific and intentional physical contacts between pairs of proteins that occur in a particular biological context, playing a key role in predicting the function of a target protein and the drug ability of molecules. De Las Rivas and Fontanillo have reviewed important concepts in PPI and network analysis.

Network Analyst

NetworkAnalyst is a powerful web-based bioinformatics tool released in 2014 to support systems-level data understanding. Its user-friendly platform offers visual analytics with statistical meta-analysis resources for interpreting gene expression data within the context of protein-protein interaction (PPI) networks. With the recent updates, it is now possible not only to visually compare multiple gene lists through interactive visualizations, but also generate gene co-expression networks, regulatory networks and cell-type specific PPI networks (Zhou et al., 2019).

CEMiTool

CEMiTool is an R package and a web tool that allows to perform co-expression analysis in an automated pipeline (Russo et al., 2018). The package is based on the Weighted correlation network analysis (WGCNA), but also provides a structured pipeline that includes an unsupervised gene filtering method and an improved automatic selection of the β value, being more reproducible and requiring less bioinformatic skills. Moreover, CEMiTool outperformed WGCNA using even less computational resources (Cheng et al., 2020).

The tool allows the identification of modules of genes that are co-expressed among samples, using a 'soft' thresholding to assign a connection-weight to each gene pair, instead of classifying connections in a binary manner (connected or not connected). Genes are divided in modules using an hierarchical clustering approach. With CEMiTool it is also possible to include information about the class of the samples and perform functional analysis based on a provided gene set. CEMiTool can also integrate the results with interactome data, providing co-expression networks. Results are available in a HTML report, with high-quality plots and interactive tables. To date, CEMiTool has provided insights in the most different topics, including osteoarthritis (Zheng et al., 2021), neuropsychiatric disorders (Hemmings et al., 2022) and breast cancer (Cedro-Tanda et al., 2020).

2.7 Data visualization

Data visualization is a very important aspect of research, especially in bioinformatics. In this field, it is very common to deal with many variables from different biological layers and involve distinct biological pathways. A good figure is not only aesthetically important, but in many cases, it is a crucial feature to extract information.

The R Graph Gallery and Ggplot2 R package

The R Graph Gallery is an organized portfolio of hundreds of graphs and their respective scripts to reproduce them in R. The Gallery is also linked to *From Data to viz* website, which displays useful information about the different types of plots, guiding the user to the most appropriate representations to use.

Many of these graphs are built using the R package *ggplot2*. The package allows the representation of data in high-quality figures, providing a range of tools to personalize the aesthetics and customize your figure, including the colors, shapes, sizes, legends, backgrounds and many more. Moreover, ggplot2 contains dozens of extensions, further increasing the range of graphical outputs, including the creation of networks, diagrams, animated representations, statistical tests, among others.

DiVenn web tool

DiVenn is a web tool that provides visualization of overlapping features from different experiments, in an integrated force-directed graph. The final result can be personalized in many ways, such as subsetting data, including feature names, and choosing colors. Moreover, there is the possibility of visualizing Gene Ontology annotations of all or some of the genes. DiVenn is a great alternative to the usual Venn Diagram, providing new resources and a better overview (Sun et al., 2019).

References

Barbie, D.A., Tamayo, P., Boehm, J.S., Kim, S.Y., Moody, S.E., Dunn, I.F., Schinzel, A.C., Sandy, P., Meylan, E., Scholl, C., Fröhling, S., Chan, E.M., Sos, M.L., Michel, K., Mermel, C., Silver, S.J., Weir, B.A., Reiling, J.H., Sheng, Q., Gupta, P.B., Wadlow, R.C., Le, H., Hoersch, S., Wittner, B.S., Ramaswamy, S., Livingston, D.M., Sabatini, D.M., Meyerson, M., Thomas, R.K., Lander, E.S., Mesirov, J.P., Root, D.E., Gilliland, D.G., Jacks, T., Hahn, W.C., 2009. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature 462, 108–112. https://doi.org/10.1038/nature08460

Barker, M., Rayens, W., 2003. Partial least squares for discrimination. J. Chemom. 17, 166–173. https://doi.org/10.1002/cem.785

Boedigheimer, M.J., Wolfinger, R.D., Bass, M.B., Bushel, P.R., Chou, J.W., Cooper, M., Corton, J.C., Fostel, J., Hester, S., Lee, J.S., Liu, F., Liu, J., Qian, H.-R., Quackenbush, J., Pettit, S., Thompson, K.L., 2008. Sources of variation in baseline gene expression levels from toxicogenomics study control animals across multiple laboratories. BMC Genomics 9, 285. https://doi.org/10.1186/1471-2164-9-285

Boulesteix, A.-L., Strimmer, K., 2006. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. Brief. Bioinform. 8, 32–44. https://doi.org/10.1093/bib/bbl016

Boxuan Cui, n.d. DataExplorer.

Breiman, L., 2001. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). Stat. Sci. 16. https://doi.org/10.1214/ss/1009213726

Cao, K.-A.L., Rossouw, D., Robert-Granié, C., Besse, P., 2008. A Sparse PLS for Variable Selection when Integrating Omics Data. Stat. Appl. Genet. Mol. Biol. 7. https://doi.org/10.2202/1544-6115.1390

Cedro-Tanda, A., Ríos-Romero, M., Romero-Córdoba, S., Cisneros-Villanueva, M., Rebollar-Vega, R.G., Alfaro-Ruiz, L.A., Jiménez-Morales, S., Domínguez-Reyes, C., Villegas-Carlos, F., Tenorio-Torres, A., Bautista-Piña, V., Beltrán-Anaya, F.O., Hidalgo-Miranda, A., 2020. A lncRNA landscape in breast cancer reveals a potential role for AC009283.1 in proliferation and apoptosis in HER2-enriched subtype. Sci. Rep. 10, 13146. https://doi.org/10.1038/s41598-020-69905-z

Cheng, C.W., Beech, D.J., Wheatcroft, S.B., 2020. Advantages of CEMiTool for gene co-expression analysis of RNA-seq data. Comput. Biol. Med. 125, 103975. https://doi.org/10.1016/j.compbiomed.2020.103975

Chiesa, M., Colombo, G.I., Piacentini, L., 2018. DaMiRseq—an R/Bioconductor package for data mining of RNA-Seq data: normalization, feature selection and classification. Bioinformatics 34, 1416–1418. https://doi.org/10.1093/bioinformatics/btx795

Chung, D., Keles, S., 2010. Sparse Partial Least Squares Classification for High Dimensional Data. Stat. Appl. Genet. Mol. Biol. 9. https://doi.org/10.2202/1544-6115.1492

Croze, E., Yamaguchi, K.D., Knappertz, V., Reder, A.T., Salamon, H., 2013. Interferon-beta-1b-induced short- and long-term signatures of treatment activity in multiple sclerosis. Pharmacogenomics J. 13, 443–451. https://doi.org/10.1038/tpj.2012.27

De Las Rivas, J., Fontanillo, C., 2010. Protein–Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks. PLoS Comput. Biol. 6, e1000807. https://doi.org/10.1371/journal.pcbi.1000807

Drăghici, S., Khatri, P., Martins, R.P., Ostermeier, G.C., Krawetz, S.A., 2003. Global functional profiling of gene expression☆☆This work was funded in part by a Sun Microsystems grant awarded to S.D., NIH Grant HD36512 to S.A.K., a Wayne State University SOM Dean's Post-Doctoral Fellowship, and an NICHD Contraception and Infertility Loan to G.C.O. Support from the WSU MCBI mode is gratefully appreciated. Genomics 81, 98–104. https://doi.org/10.1016/S0888-7543(02)00021-6

Fisher, R.A., 1992. Statistical Methods for Research Workers, in: Kotz, S., Johnson, N.L. (Eds.),

Breakthroughs in Statistics, Springer Series in Statistics. Springer New York, New York, NY, pp. 66–70. https://doi.org/10.1007/978-1-4612-4380-9_6

Gonçalves, A.N.A., Lever, M., Russo, P.S.T., Gomes-Correia, B., Urbanski, A.H., Pollara, G., Noursadeghi, M., Maracaja-Coutinho, V., Nakaya, H.I., 2019. Assessing the Impact of Sample Heterogeneity on Transcriptome Analysis of Human Diseases Using MDP Webtool. Front. Genet. 10, 971. https://doi.org/10.3389/fgene.2019.00971

Griss, J., Viteri, G., Sidiropoulos, K., Nguyen, V., Fabregat, A., Hermjakob, H., 2020. ReactomeGSA - Efficient Multi-Omics Comparative Pathway Analysis. Mol. Cell. Proteomics 19, 2115–2125. https://doi.org/10.1074/mcp.TIR120.002155

Hemmings, S.M.J., Swart, P., Womersely, J.S., Ovenden, E.S., van den Heuvel, L.L., McGregor, N.W., Meier, S., Bardien, S., Abrahams, S., Tromp, G., Emsley, R., Carr, J., Seedat, S., 2022. RNA-seq analysis of gene expression profiles in posttraumatic stress disorder, Parkinson's disease and schizophrenia identifies roles for common and distinct biological pathways. Discov. Ment. Health 2, 6. https://doi.org/10.1007/s44192-022-00009-y

Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. J. Educ. Psychol. 24, 417–441. https://doi.org/10.1037/h0071325

Huang, D.W., Sherman, B.T., Lempicki, R.A., 2009. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 37, 1–13. https://doi.org/10.1093/nar/gkn923

Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R., Loney, F., May, B., Milacic, M., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Weiser, J., Wu, G., Stein, L., Hermjakob, H., D'Eustachio, P., 2019. The reactome pathway knowledgebase. Nucleic Acids Res. gkz1031. https://doi.org/10.1093/nar/gkz1031

Johnson, W.E., Li, C., Rabinovic, A., 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8, 118–127. https://doi.org/10.1093/biostatistics/kxj037

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., Tanabe, M., 2016. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 44, D457–D462. https://doi.org/10.1093/nar/gkv1070

Kevin Blighe, A.L., n.d. PCAtools: everything Principal Components Analysis.

Leek, J.T., Scharpf, R.B., Bravo, H.C., Simcha, D., Langmead, B., Johnson, W.E., Geman, D., Baggerly, K., Irizarry, R.A., 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat. Rev. Genet. 11, 733–739. https://doi.org/10.1038/nrg2825

Li, S., Rouphael, N., Duraisingham, S., Romero-Steiner, S., Presnell, S., Davis, C., Schmidt, D.S., Johnson, S.E., Milton, A., Rajam, G., Kasturi, S., Carlone, G.M., Quinn, C., Chaussabel, D., Palucka, A.K., Mulligan, M.J., Ahmed, R., Stephens, D.S., Nakaya, H.I., Pulendran, B., 2014. Molecular signatures of antibody responses derived from a systems biology study of five human vaccines. Nat. Immunol. 15, 195–204. https://doi.org/10.1038/ni.2789

Maleki, F., Ovens, K., Hogan, D.J., Kusalik, A.J., 2020. Gene Set Analysis: Challenges, Opportunities, and Future Research. Front. Genet. 11, 654. https://doi.org/10.3389/fgene.2020.00654

Nguyen, D.V., Rocke, D.M., 2002. Tumor classification by partial least squares using microarray gene expression data. Bioinformatics 18, 39–50. https://doi.org/10.1093/bioinformatics/18.1.39

Oliveira-de-Souza, D., Vinhaes, C.L., Arriaga, M.B., Kumar, N.P., Cubillos-Angulo, J.M., Shi, R., Wei, W., Yuan, X., Zhang, G., Cai, Y., Barry, C.E., Via, L.E., Sher, A., Babu, S., Mayer-Barber, K.D., Nakaya, H.I., Fukutani, K.F., Andrade, B.B., 2019. Molecular degree of perturbation of plasma inflammatory markers associated with tuberculosis reveals distinct disease profiles between Indian and Chinese populations. Sci. Rep. 9, 8002. https://doi.org/10.1038/s41598-019-44513-8

Pankla, R., Buddhisa, S., Berry, M., Blankenship, D.M., Bancroft, G.J., Banchereau, J.,

Lertmemongkolchai, G., Chaussabel, D., 2009. Genomic transcriptional profiling identifies a candidate blood biomarker signature for the diagnosis of septicemic melioidosis. Genome Biol. 10, R127. https://doi.org/10.1186/gb-2009-10-11-r127

Picard, M., Scott-Boyer, M.-P., Bodein, A., Périn, O., Droit, A., 2021. Integration strategies of multi-omics data for machine learning analysis. Comput. Struct. Biotechnol. J. 19, 3735–3746. https://doi.org/10.1016/j.csbj.2021.06.030

Polikar, R., 2006. Ensemble based systems in decision making. IEEE Circuits Syst. Mag. 6, 21–45. https://doi.org/10.1109/MCAS.2006.1688199

Russo, P.S.T., Ferreira, G.R., Cardozo, L.E., Bürger, M.C., Arias-Carrasco, R., Maruyama, S.R., Hirata, T.D.C., Lima, D.S., Passos, F.M., Fukutani, K.F., Lever, M., Silva, J.S., Maracaja-Coutinho, V., Nakaya, H.I., 2018. CEMiTool: a Bioconductor package for performing comprehensive modular co-expression analyses. BMC Bioinformatics 19, 56. https://doi.org/10.1186/s12859-018-2053-1

Singh, A., Shannon, C.P., Gautier, B., Rohart, F., Vacher, M., Tebbutt, S.J., Lê Cao, K.-A., 2016. DIABLO: from multi-omics assays to biomarker discovery, an integrative approach (preprint). Bioinformatics. https://doi.org/10.1101/067611

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P., 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. 102, 15545–15550. https://doi.org/10.1073/pnas.0506580102

Sun, L., Dong, S., Ge, Y., Fonseca, J.P., Robinson, Z.T., Mysore, K.S., Mehta, P., 2019. DiVenn: An Interactive and Integrated Web-Based Visualization Tool for Comparing Gene Lists. Front. Genet. 10, 421. https://doi.org/10.3389/fgene.2019.00421

The Gene Ontology Consortium, Carbon, S., Douglass, E., Good, B.M., Unni, D.R., Harris, N.L., Mungall, C.J., Basu, S., Chisholm, R.L., Dodson, R.J., Hartline, E., Fey, P., Thomas, P.D., Albou, L.-P., Ebert, D., Kesling, M.J., Mi, H., Muruganujan, A., Huang, X., Mushayahama, T., LaBonte, S.A., Siegele, D.A., Antonazzo, G., Attrill, H., Brown, N.H., Garapati, P., Marygold, S.J., Trovisco, V., dos Santos, G., Falls, K., Tabone, C., Zhou, P., Goodman, J.L., Strelets, V.B., Thurmond, J., Garmiri, P., Ishtiaq, R., Rodríguez-López, M., Acencio, M.L., Kuiper, M., Lægreid, A., Logie, C., Lovering, R.C., Kramarz, B., Saverimuttu, S.C.C., Pinheiro, S.M., Gunn, H., Su, R., Thurlow, K.E., Chibucos, M., Giglio, M., Nadendla, S., Munro, J., Jackson, R., Duesbury, M.J., Del-Toro, N., Meldal, B.H.M., Paneerselvam, K., Perfetto, L., Porras, P., Orchard, S., Shrivastava, A., Chang, H.-Y., Finn, R.D., Mitchell, A.L., Rawlings, N.D., Richardson, L., Sangrador-Vegas, A., Blake, J.A., Christie, K.R., Dolan, M.E., Drabkin, H.J., Hill, D.P., Ni, L., Sitnikov, D.M., Harris, M.A., Oliver, S.G., Rutherford, K., Wood, V., Hayles, J., Bähler, J., Bolton, E.R., De Pons, J.L., Dwinell, M.R., Hayman, G.T., Kaldunski, M.L., Kwitek, A.E., Laulederkind, S.J.F., Plasterer, C., Tutaj, M.A., Vedi, M., Wang, S.-J., D'Eustachio, P., Matthews, L., Balhoff, J.P., Aleksander, S.A., Alexander, M.J., Cherry, J.M., Engel, S.R., Gondwe, F., Karra, K., Miyasato, S.R., Nash, R.S., Simison, M., Skrzypek, M.S., Weng, S., Wong, E.D., Feuermann, M., Gaudet, P., Morgat, A., Bakker, E., Berardini, T.Z., Reiser, L., Subramaniam, S., Huala, E., Arighi, C.N., Auchincloss, A., Axelsen, K., Argoud-Puy, G., Bateman, A., Blatter, M.-C., Boutet, E., Bowler, E., Breuza, L., Bridge, A., Britto, R., Bye-A-Jee, H., Casas, C.C., Coudert, E., Denny, P., Estreicher, A., Famiglietti, M.L., Georghiou, G., Gos, A., Gruaz-Gumowski, N., Hatton-Ellis, E., Hulo, C., Ignatchenko, A., Jungo, F., Laiho, K., Le Mercier, P., Lieberherr, D., Lock, A., Lussi, Y., MacDougall, A., Magrane, M., Martin, M.J., Masson, P., Natale, D.A., Hyka-Nouspikel, N., Orchard, S., Pedruzzi, I., Pourcel, L., Poux, S., Pundir, S., Rivoire, C., Speretta, E., Sundaram, S., Tyagi, N., Warner, K., Zaru, R., Wu, C.H., Diehl, A.D., Chan, J.N., Grove, C., Lee, R.Y.N., Muller, H.-M., Raciti, D., Van Auken, K., Sternberg, P.W., Berriman, M., Paulini, M., Howe, K., Gao, S., Wright, A., Stein, L., Howe, D.G., Toro, S., Westerfield, M., Jaiswal, P., Cooper, L., Elser, J., 2021. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res.

49, D325–D334. https://doi.org/10.1093/nar/gkaa1113

Weiner 3rd, J., Domaszewska, T., 2016. tmod: an R package for general and multivariate enrichment analysis (preprint). PeerJ Preprints. https://doi.org/10.7287/peerj.preprints.2420v1

Wold, Herman, 1966. Estimation of principal components and related models by iterative least squares. Multivar. Anal. 391–420.

Yao, F., Coquery, J., Lê Cao, K.-A., 2012. Independent Principal Component Analysis for biologically meaningful dimension reduction of large biological data sets. BMC Bioinformatics 13, 24. https://doi.org/10.1186/1471-2105-13-24

Zhang, Y., Parmigiani, G., Johnson, W.E., 2020. ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genomics Bioinforma. 2, lqaa078. https://doi.org/10.1093/nargab/lqaa078

Zheng, L., Chen, W., Xian, G., Pan, B., Ye, Y., Gu, M., Ma, Y., Zhang, Z., Sheng, P., 2021. Identification of abnormally methylated–differentially expressed genes and pathways in osteoarthritis: a comprehensive bioinformatic study. Clin. Rheumatol. 40, 3247–3256. https://doi.org/10.1007/s10067-020-05539-w

Zhou, G., Soufan, O., Ewald, J., Hancock, R.E.W., Basu, N., Xia, J., 2019. NetworkAnalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis. Nucleic Acids Res. 47, W234–W241. https://doi.org/10.1093/nar/gkz240

CHAPTER 3

A Systems biology approach to study the host's responses to pneumococcal lung infection

3.1 *Streptococcus pneumoniae*

Streptococcus pneumoniae, also known as pneumococcus, is a lancet-shaped gram-positive bacteria, typically observed in pairs or short chains. These bacteria present a thick cell wall, composed mainly of peptidoglycan and teichoic acids, known as cell wall polysaccharides (CWPS). Surrounding the cell wall, there is a polysaccharide capsule composed of repeating units of monosaccharides, which is a major pneumococcal virulence factor and dictates the ability of the bacteria to cause invasive diseases (Brown et al., 2015; Mitchell and Mitchell, 2010).

While the CWPS presents a similar structure among different strains, the capsular polysaccharide is present in distinct structures, grouping *S. pneumoniae* strains sharing a unique capsular structure into serotypes (AlonsoDeVelasco et al., 1995). Today, more than 100 different serotypes were identified (Ganaie et al., 2020). Serotypes that are antigenically similar are grouped into serogroups.

Besides the CWPS and the capsule, *S. pneumoniae* present many proteins involved in various stages of its pathogenesis (Mitchell and Mitchell, 2010). Different surface proteins contributes to the interaction with the host's tissues and evasion of the immune system, including three important adhesins: pneumococcal surface antigen A (PsaA) and pneumococcal surface protein A and C (PspA, PspC) (Berry and Paton, 1996; Brock et al., 2002; Ogunniyi et al., 2007). Moreover, the extracellular glycosidase neuraminidase A (NanA) is also known to be involved in adhesion to host's cells (Tong et al., 2000).

Other two important virulence factors are the Pneumolysin, a pore-forming toxin that binds to membrane cholesterol and has a major role in inflammation (Weiser et al., 2018) and the sIgA1 protease, which cleaves the hinge of the human immunoglobulin IgA1 (Weiser et

al., 2003). The role of these virulence factors will be further discussed across this chapter, along the host's responses to this bacteria.

3.2 Pneumococcal disease epidemiology

S. pneumoniae colonizes the upper respiratory tract (URT) in an age-dependent manner, being higher in young children and decreasing in adult age and the prevalence is also higher in low-middle countries, compared to developed countries. The colonization is usually asymptomatic but is the leading source of transmission between individuals and is a prerequisite for pneumococcal infection (Le Polain de Waroux et al., 2014).

Infection occurs if the bacteria gains access to other tissues. When the spread is local, it causes noninvasive diseases like sinusitis, otitis, and non bacteremic pneumonia. When the bacteria reach the bloodstream it leads to invasive manifestations such as bacteremia, meningitis, and bacteremic pneumonia, presenting high mortality rates (Bogaert et al., 2004; Kadioglu et al., 2008).

Although pneumococcal diseases attain people at every age, infection and invasive diseases are more common in children and in the elderly , along with immunocompromised people (Melegaro et al., 2006). According to the World Health Organization, in 2019 pneumonia was responsible for 22% of all deaths in children aged 1 to 5, with *S. pneumoniae* being the leading cause of bacterial pneumonia ("Pneumonia," n.d.).

Vaccines are the most important strategy in preventing pneumococcal disease. The currently available vaccines target the polysaccharide capsule and are classified in two categories: the polysaccharide pneumococcal vaccines (PPV) and the pneumococcal conjugate vaccines (PCVs). The first group consists of the purified capsular polysaccharides (CPS) of 23 serotypes and elicits a T cell independent response. Despite effectively preventing pneumococcal bacteraemia and meningitis in adults and elderly people (Moberley et al., 2013), this vaccine does not induce affinity maturation and immunological memory (Rose et al., 2005), aside from being poorly immunogenic in children (Pollard et al., 2009).

In the pneumococcal conjugate vaccines (PCVs) a carrier protein is covalently linked to the purified CPS of 10 or 13 different serotypes, eliciting T cell dependent responses and leading to class switching, affinity maturation and antigen specific memory B cells, which are important for the long-term protection against pneumococcal diseases (Papadatou et al., 2019). The introduction of these vaccines resulted in an important decrease in invasive manifestations of pneumococcal infection. However, there is a growing concern regarding the replacement of the common serotypes by the serotypes not covered by the available vaccines and, given the impossibility of covering all serotypes in one product, scientists have been seeking serotype-independent strategies (Briles et al., 2019; Kaplan et al., 2013; Pichichero, 2017).

3.3 Host-pathogen interactions

The balance between the host's specific and nonspecific defenses against *S. pneumoniae* and the ability of the bacteria to counteract these defenses will dictate the colonization and the spread to other tissues (Weiser et al., 2018). Moreover, once the pneumococcal infection is established, the host's inflammatory responses play a major role, influencing the symptoms and disease outcome (Musher D. M, 1992; Dockrell, D. H., 2012).

Therefore, to better understand the systemic responses to a lung infection, it is important to have an overview of how the innate and adaptive branches of the immune system fight *S. pneumoniae* and how the bacteria escape from these responses.

3.3.1. Colonization, a prerequisite for infection

Physical defenses protect the respiratory tract, including the presence of antimicrobials, neutralizing immunoglobulins and an epithelial layer that provides a mucociliary escalator through the rapid beating cilia. In addition, resident leucocytes help to maintain the airway integrity. To succeed in infecting humans, the pathogen needs to overcome these different barriers (LeMessurier et al., 2020).

In humans, *S. pneumoniae* asymptomatically colonizes the nasopharynx, a process known as carriage. The colonization is a process mediated by different pneumococcal proteins, including the phosphorylcholine (Chop), a component of the cell wall that binds epithelial cells through platelet-activating factor receptors (PAFR), serving as an anchor for different choline-binding proteins (Cundell et al., 1995). Particularly, choline binding protein A (CbpA) and the pneumococcal choline-binding protein A (PcpA) contribute to the adhesion to epithelial cells (Khan et al., 2012; Zhang et al., 2000). On the other hand, the pneumococcal protease sIgA1 cleaves human IgA1 antibodies, the most abundant immunoglobulin in mucosal sites, and the remaining Fab fragments in the surface of the bacteria are associated with the increase in adhesion as well (Weiser et al., 2003).

In healthy individuals, the mucociliary escalator usually eliminates *S. pneumoniae* before the arrival to the lower respiratory tract and the bacteria is cleared within a couple of weeks. The process of overtaking the mucus escalator is facilitated by *S. pneumoniae*'s exoglycosidases, such as the neuraminidase NanA, that plays its role by deglycosylation of the mucus (Loughran A, 2019). Moreover, ciliary beating is inhibited by the cytotoxin pneumolysin (Fliegauf M, 2013) and the capsule, that is negatively charged, promotes evasion of mucus via electrostatic repulsion since the mucus is also negatively charged (Dockrell and Brown, 2015).

Depending on factors from both the host and the bacterial strain, *S. pneumoniae* can escape the clearance and spread to other tissues. The likelihood of penetration to other tissues, including microaspiration to the lungs, rises with the increase of bacterial loads in the nasopharynx. This can occur due to the induction of pro-inflammatory chemokines and cytokines, upregulation of target receptors and viral infections that cause damage to the respiratory epithelium (Loughran et al., 2019; Weiser et al., 2018).

The respiratory epithelium also presents high amounts of antimicrobial peptides and immunoglobulins, which is overcomed by the autolytic enzyme LytA that promotes the shedding of the capsule, leading to a considerable resistance to these peptides and permitting

a closer interaction with the host's cells, facilitating the invasion of epithelial cells (Kietzman C, 2016).

3.3.2. Pneumonia development and Inflammation

If the bacteria manage to arrive in the lungs, one of the first stages of infection involves the action of the pneumococcal esterase, EstA, that contributes to the cleavage of the sialic acid by NanA, exposing the ligands in the host's lung cells. With the progression of the disease, the bronchioles and alveoli have their extracellular matrix exposed and free to interact with different pneumococcal surface proteins, promoting the attachment of the bacteria (Loughran et al., 2019).

The first cells to recognize *S. pneumoniae* in the lungs are the alveolar macrophages. They subsequently stimulate the alveolar epithelium to amplify the response, recruiting additional immune cells (Koppe et al., 2012). Our innate immune system recognizes *S. pneumoniae* through pattern recognition receptors (PRRs), such as Toll-Like receptors 2, 4 and 9, NOD-like receptors, and cytosolic DNA sensors. PRRs can recognize pathogen-associated molecular patterns (PAMPs), microbe-associated molecular patterns (MAMPs) and damage or danger-associated molecular patterns (DAMPs) (Koppe et al., 2012).

During pneumococcal infection, PRRs are activated by different *S. pneumoniae* components, including the cell wall, pneumolysin and bacterial DNA. As a consequence, they trigger the transcription of NF-kB, resulting in the production of inflammatory molecules, including the cytokines TNFα, IL-6, IFNα/β, KC, MCP-1 and pro-IL-1β. PRRs also induce the formation of inflammasomes, like the NLRP3, which is activated by pneumolysin and controls IL-1 production at a post-translational level. All these mediators will promote an acute response to the bacteria, recruiting additional neutrophils and macrophages (Opitz et al., 2010).

At this stage, inflammation becomes very intense, driven especially by the bacterial cell wall, pneumolysin and hydrogen peroxide. The epithelial cells and capillaries become inflamed, the leukocytes enter the lesion and bacteria is engulfed by macrophages. Importantly, Pneumolysin inhibits the oxidative burst of neutrophils and macrophages and disrupt tight junctions, including the alveoli-capillary barrier, contributing to leakage and consequently allowing serous exudates to enter the lungs and the bacteria to cross into the bloodstream (Loughran et al., 2019). Indeed, Pneumolysin seems to be required for both development of severe pneumonia and bacterial survival in the blood (Loughran et al., 2019; Orihuela et al., 2004).

The development of invasive pneumococcal diseases directly from asymptomatic colonization is not common, but can take place in situations of innate immunity disruption. In fact, the access to the bloodstream occurs generally through the lungs. The intense inflammation at alveolar level, with presence of edema fluid, accumulation of erythrocytes, lymphatic dilatation and overabundance of bacteria is a propitious environment for bloodstream access (Loughran et al., 2019; Orihuela et al., 2004).

Although bacterial pneumonia generally leads to a compartmentalized inflammation, lung infection often results in a systemic response, even without bacteremia (Deng and Standiford, 2005) .

3.3.3. The development of an adaptive response to *S. pneumoniae*

In humans, different studies have shown that colonization of the nasopharynx is a immunizing process and leads to the generation of an adaptive response involving acquisition of anti-capsular and anti-protein antibodies (Goldblatt et al., 2005; Weinberger et al., 2008; Zhang et al., 2006), and CD4+ cellular responses (Lebon et al., 2011; Mureithi et al., 2009). The development of antibodies against capsular and non capsular components of pneumococcus seem to contribute to the gradual resistance to colonization, but cellular

responses also play a role in this process, especially through CD4+ IL-17A producing T cells (Malley, 2010).

Antibodies have their importance in pneumococcal diseases demonstrated through the success of vaccines developed to target the humoral response. In the context of an infection, the humoral response can be triggered against different surface proteins and the polysaccharide capsule, the later one inducing thymus-independent responses mediated by marginal zone B cells. Although this gives rise to plasma cells that secrete low affinity antibodies without generating memory B cells, thymus-independent responses develop faster than T cell dependent responses, being crucial to decrease pathogenic burden early in infection (Coutinho and Möller, 1973; Martin et al., 2001).

Since pneumococcus can cleave IgA, the most common antibody in mucosal surfaces, nasal and lung mucosal protection likely requires the action of other immunoglobulin classes, like IgG (Janoff et al., 2014).

Within the cellular responses, it is known that CD4+ T cells provide protection against *S. pneumoniae* in an antibody-dependent manner (Malley et al., 2005). These cells act through the release of cytokines and are activated through co-stimulatory molecules and antigen presentation cells, differentiating in Th1 and Th2 cells. Th1 cells can stimulate the cellular responses by releasing cytokines such Interferon-gamma, activating and recruiting other immune cells, like macrophages. In an *in vitro* study with human monocytes, live pneumococcus triggered a Th1-biased response, while killed pneumococcus triggered a Th17 response (Olliver et al., 2011). On the other hand, Th2 cells support humoral responses through the release of IL-4 cytokine and interaction with B cells (Romagnani, 1999).

T-helper 17 and regulatory T cells are very important in the context of a pneumococcal infection. Peripheral blood mononuclear cells (PBMC) from healthy adults living in regions with high incidence of pneumococcal carriage and infection respond to pneumococcal antigens by the production of IFN and IL-17, pointing out the development of a memory mediated by T cells (Mureithi et al., 2009). The release of IL-17 cytokine by Th17

cells is a pro-inflammatory process and recruits leukocytes to the site of infection, promoting clearance of the bacteria. Regulatory T cells act to regulate Th17 responses, avoiding autoimmunity caused by an over-inflammation process (Hoe et al., 2017).

3.4 The role of the spleen in pneumococcal infection

The structure of the spleen comprises the red pulp and the white pulp, the latter being surrounded by an interface region, the marginal zone (MZ). Afferent arterial blood ends in sinusoids in the MZ and flows through sinusoids spaces and red pulp into venous sinuses (Bronte and Pittet, 2013).

The MZ contains specific subsets of macrophages and B cells, acting as a bridge between innate and adaptive immune systems (Mebius RE, 2005). In a mouse model, blood-borne pathogens in circulation can enter the splenic MZ easily through a fenestrated marginal sinus, or have their antigens captured by dendritic cells and granulocytes that will then transport them to the MZ. (Mebius RE, 2005; Balázs M, 2002). The white pulp contains T cell and B cell zones, being structurally similar to a lymph node and permitting the generation of antigen-specific responses (Bronte and Pittet, 2013).

Thanks to its structure and composition, presenting a high perfusion and a multitude of cell types, the spleen plays an important role in the response to blood-borne pathogens. In fact, the spleen is the main responsible for the clearance of *S. pneumoniae* from the blood (Shinefield HR, 1066; Theilacker Cm 2016).

In brief, besides clearing pathogens from the circulation, the spleen is capable of triggering innate and adaptive responses against these pathogens. Therefore, when establishing a model to study pneumococcal pneumonia in mice, spleens are an important organ to investigate.

3.5 The aim of this chapter

In the present work, mice were intranasally infected with 10^7 CFU of TIGR4, causing a lung infection without indication of systemic spread. Spleens were collected at different time points after infection, stimulated with the same strain of inactivated bacteria and changes in gene expression and cytokines' concentration were assessed. From this perspective, the aim of this chapter is to understand the systemic response and memory generated in a pneumococcal lung infection model using a systems biology approach.

ISABELLE FRANCO MOSCARDINI
12/09/2023 16:09:31

Isabelle Franco moscardini

Immune Memory After Respiratory Infection With *Streptococcus pneumoniae* Is Revealed by *in vitro* Stimulation of Murine Splenocytes With Inactivated Pneumococcal Whole Cells: Evidence of Early Recall Responses by Transcriptomic Analysis

Isabelle Franco Moscardini[1], Francesco Santoro[2], Monica Carraro[2], Alice Gerlini[1], Fabio Fiorino[2], Chiara Germoni[2], Samaneh Gholami[2], Elena Pettini[2], Donata Medaglini[2], Francesco Iannelli[2] and Gianni Pozzi[2]*

[1] Microbiotec srl, Siena, Italy, [2] Laboratory of Molecular Microbiology and Biotechnology (LAMMB), Department of Medical Biotechnologies, University of Siena, Siena, Italy

The *in vitro* stimulation of immune system cells with live or killed bacteria is essential for understanding the host response to pathogens. In the present study, we propose a model combining transcriptomic and cytokine assays on murine splenocytes to describe the immune recall in the days following pneumococcal lung infection. Mice were sacrificed at days 1, 2, 4, and 7 after *Streptococcus pneumoniae* (TIGR4 serotype 4) intranasal infection and splenocytes were cultured in the presence or absence of the same inactivated bacterial strain to access the transcriptomic and cytokine profiles. The stimulation of splenocytes from infected mice led to a higher number of differentially expressed genes than the infection or stimulation alone, resulting in the enrichment of 40 unique blood transcription modules, including many pathways related to adaptive immunity and cytokines. Together with transcriptomic data, cytokines levels suggested the presence of a recall immune response promoting both innate and adaptive immunity, stronger from the fourth day after infection. Dimensionality reduction and feature selection identified key variables of this recall response and the genes associated with the increase in cytokine concentrations. This model could study the immune responses involved in pneumococcal infection and possibly monitor vaccine immune response and experimental therapies efficacy in future studies.

Keywords: *Streptococcus pneumoniae*, *in vitro* stimulation, Transcriptomic Analysis, Recall immune responses, lung infection, cytokines/chemokines

Edited by:
Elsa Bou Ghanem,
University at Buffalo, United States

Reviewed by:
Sarah Clark,
University of Colorado, United States
Olanrewaju B. Morenikeji,
University of Pittsburgh at Bradford, United States

***Correspondence:**
Francesco Santoro
santorof@unisi.it

Specialty section:
This article was submitted to Molecular Bacterial Pathogenesis, a section of the journal Frontiers in Cellular and Infection Microbiology

Received: *05 February 2022*
Accepted: *21 April 2022*
Published: *20 June 2022*

Citation:
Moscardini IF, Santoro F, Carraro M, Gerlini A, Fiorino F, Germoni C, Gholami S, Pettini E, Medaglini D, Iannelli F and Pozzi G (2022) Immune Memory After Respiratory Infection With Streptococcus pneumoniae Is Revealed by in vitro Stimulation of Murine Splenocytes With Inactivated Pneumococcal Whole Cells: Evidence of Early Recall Responses by Transcriptomic Analysis. Front. Cell. Infect. Microbiol. 12:869763. doi: 10.3389/fcimb.2022.869763

1 INTRODUCTION

Streptococcus pneumoniae is a major human pathogen responsible for various diseases, including life-threatening conditions such as pneumonia, sepsis, and meningitis (Weiser et al., 2018). Current vaccines have been very efficient in reducing the death toll caused by this pathogen. However, strains not covered by the available vaccines represent a growing concern, demanding new serotype-independent strategies (Kaplan et al., 2013; Pichichero, 2017; Briles et al., 2019). Models to assess the response to new vaccine candidates would be of great use.

In vitro stimulation with live or killed bacteria has been used for decades for understanding the host's response to different pathogens, including *S. pneumoniae* (Zhan and Cheers, 1995; Schultz et al., 1998; Wu et al., 2011; de Stoppelaar et al., 2016). This technique has also been applied in vaccine studies, characterizing the immune response after a second stimulus (Paranavitana et al., 2010; Moffitt et al., 2011; Shao et al., 2015). Changes in gene expression and in cytokine concentration were metrics assessed by some of these works to study the immune profile of pneumococcal infection. In the present work, we propose the combination of transcriptomic and cytokine assays from murine splenocytes to assess the immune memory built in the days following pneumococcal infection.

The spleen plays a vital role in host defenses against encapsulated blood-borne pathogens due to its elevated perfusion and efficient immune surveillance of the circulatory system (Cerutti et al., 2013). In a pneumococcal bacteremia model, bacteria present a tropism to the spleen, and macrophages present in the splenic Red Pulp (RP) are responsible for an initial binding and subsequent clearance of *S. pneumoniae* mediated by mature neutrophils present on the RP (Deniset et al., 2017; Ercoli et al., 2018). Moreover, the splenic Marginal Zone (MZ) is a crucial area of antigen presentation to MZ B cells that are capable of rapidly differentiating into plasmablasts, secreting low-affinity IgM and IgG (Cerutti et al., 2013).

RNA sequencing technologies permit us to gain insights into the host's response due to the possibility of analyzing the changes in gene expression in different conditions. The transcriptomic information can be integrated with other biological layers or clinical data, permitting a more comprehensive understanding of biological processes in response to perturbations such as infection and vaccination.

In the current work, we propose the study of the host systemic responses to a pneumococcal lung infection by assessing gene expression and cytokine profiles of splenocytes, identifying biological pathways and the key features involved in this process.

2 METHODS

2.1 Animals and Animal Infection

Seven-week female C57BL/6 mice (Charles River Italia, Italy) were treated according to national guidelines (Decreto Legislativo 26/2014) utilizing the three R's principles. Animals were maintained under specific pathogen-free conditions in the animal facility of the Laboratory of Molecular Microbiology and Biotechnology (LA.M.M.B.), Department of Medical Biotechnologies at University of Siena, at 20-24°C, with 55 ± 10% of humidity, with food and water *ad libitum*. The study was approved by the Italian Ministry of Health with authorization n° 304/2018-PR. As previously described (Kadioglu et al., 2011) male and female mice respond differently to pneumococcal infection and, therefore the use of only female mice can be considered a limitation of this study.

Mouse-passaged TIGR4 strain of *S. pneumoniae* (Gerlini et al., 2014) was inoculated 1:50 in TSB (Tryptic-Soy Broth, Becton Dickinson, USA) supplemented with 0.1% of glucose (PanReac, Applichem, Italy), 1% of yeast extract (Oxoid, UK) and 0.016 M K_2HPO_4 (Sigma-Aldrich, USA) (TSB-GYP). The bacterial culture in mid-exponential phase (≈OD_{590} = 0.6), was centrifuged at 2,000 x *g* for 10 minutes and resuspended in an appropriate volume of saline. Before the centrifugation, the culture was Gram-stained and bacterial vital counts were performed using the multilayer plating method (Iannelli et al., 2021). Each mouse was anesthetized by intraperitoneal administration of 15 mg/kg tiletamine hydrochloride/zolazepam and 4 mg/kg xylazine and intranasally infected by instillation of 10^7 CFU of TIGR4, prepared as described above, in the volume of 25 µl/nostril in TSB. Mice were euthanized at different time points with overdose of anesthesia and cervical dislocation, as shown in **Figure 1**. Non-infected mice composed the baseline group. Each group included 6 animals.

2.2 Sample Collection

After aseptic removal at different time points (days 0, 1, 2, 4, and 7 after infection), spleens were meshed onto 70 µm nylon screens (Sefar Italia, Italy) using a scalpel and scraper. Cells were washed two times in RPMI (Sigma-Aldrich) supplemented with 10% of fetal bovine serum (FBS, Gibco, USA) and 1% of penicillin-streptomycin (Sigma-Aldrich)(cRPMI), treated with red blood cells lysis buffer according to manufacturer instruction (eBioscience, USA), and resuspended in cRPMI for cell counting by an automatic cell counter (Bio-Rad, USA). Lungs were aseptically removed, meshed onto 40 µm nylon screens (Sefar Italia, Italy), suspended in 1 ml of TSB containing glycerol at a final concentration of 10% and frozen at -70°C.

2.3 Bacterial Cells Counts in Lungs

Bacterial colony forming units (CFUs) were determined by plating appropriate dilutions of frozen lungs using a multilayer plating procedure (Iannelli et al., 2021). The lower limit of detection was 10 CFU/ml. CFUs were counted at 1, 2, 4, and 7 days after intranasal infection and in the mock infected control group.

2.4 Preparation of Inactivated Whole Cells of *S. pneumoniae*

TIGR4 was inoculated 1:1000 (v:v) in 1 L of pre-heated (37°C) TSB-GYP in a GLS80 1 liter bottle (Duran, USA). Temperature was maintained constant at 37°C, the pH was continuously measured with a probe (InPro3030, Mettler Toledo) and kept

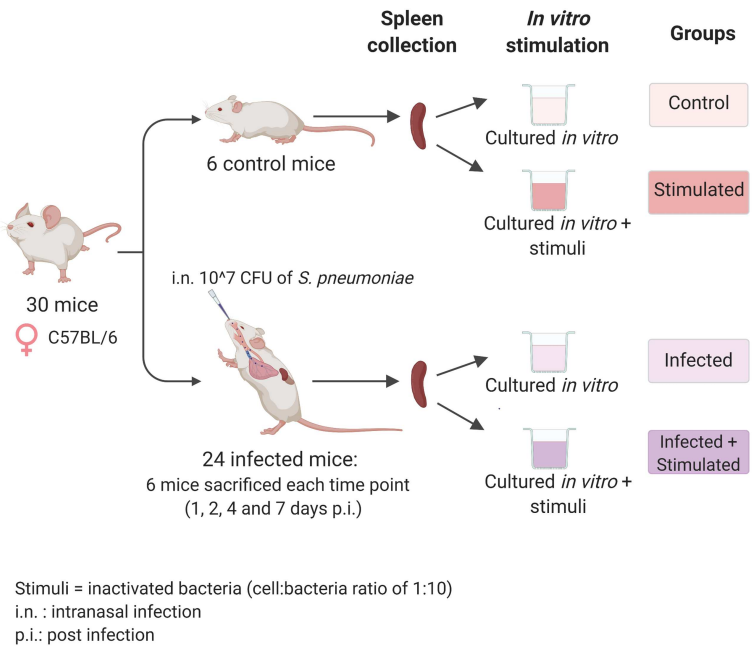


FIGURE 1 | Experimental design. Seven-weeks old C57BL/6 female mice were infected intranasally with a dose of 10^7 CFU/mouse of pneumococcal serotype 4, TIGR4 strain. After 1, 2, 4, and 7 days post-infection each group of mice (n=6) was euthanized and their spleens were collected for the isolation of splenocytes. Splenocytes from each single mouse were stimulated with formalin-inactivated TIGR4 (cell:bacteria 1:10), or maintained in cRPMI medium alone. The incubation was performed at 37°C in 5% CO_2 in a period of 6 hours for samples used in the RNA-sequencing experiment or 72 hours for samples used in the Bio-Plex Multiplex immunoassay. Non-infected mice were euthanized at day 0, splenocytes were collected and maintained in the same *in vitro* conditions of the unstimulated group and used as controls. Created with biorender.com.

at 6.9 by peristaltic pump controlled addition of 3M NaOH. Agitation was set at 100 rpm. Growth was monitored by aseptically drawing aliquots and measuring their OD_{590} in a Spectronic 200 spectrophotometer (ThermoFisher). At the peak OD_{590} (about 2.5, corresponding to 10^9 CFU/ml) bacteria were harvested by centrifugation, resuspended in PBS/10% glycerol and frozen at -70°C. TIGR4 bacteria were then thawed and inactivated by treatment with 1.5% formalin for 2 hours on a roller mixer at room temperature, then washed twice and resuspended in water.

2.5 Splenocyte Stimulation and Cytokine Secretion Assay

Splenocytes were cultured in a U-bottom 96-well plate in triplicate for 72 hours at 37°C with 5% CO_2 in cRPMI.

Splenocyte cultures were incubated in the presence or not of inactivated TIGR4, at a cell:bacteria ratio of 1:10. Unstimulated control splenocytes were cultured in cRPMI alone, and positive control splenocytes were stimulated with 50 ng/ml of phorbol 12-myristate 13-acetate (PMA) and 1 µM of Ionomycin (both from Sigma-Aldrich). After stimulation, cells were harvested and centrifuged at 500 x *g* for 15 minutes at 4°C. The supernatant was recovered and frozen at -70°C for subsequent Luminex immunoassay.

A broad screening panel consisting of a biologically-relevant collection of adaptive immunity cytokines, pro-inflammatory cytokines, and anti-inflammatory cytokines was used. In particular, IL-1α, IL-1β, IL-2, IL-3, IL-4, IL-5, IL-6, IL-9, IL-10, IL-12p40, IL-12p70, IL-13, IL-17, G-CSF, GM-CSF, IFNg, and TNF-α, and of the chemokines Eotaxin, KC, MCP-1 (MCAF), MIP-1α, MIP-1β, and RANTES production by *in vitro* stimulated splenocytes was assessed with the BioPlex pro mouse cytokine group 1 - panel 23-plex immunoassay (Bio-Rad, USA) according to manufacturer guidelines, and analyzed by Bio-Plex Magpix Multiplex reader (Bio-Rad). Cytokine and chemokine concentration was expressed as picograms per milliliter (pg/ml) and were calculated using Bio-Plex Manager 6.1.

2.6 Splenocyte Stimulation for RNA-Sequencing

In a U-bottom plate, 1 x 10^6 splenocytes/well were seeded in quintuplicate and cultured for 6 hours at 37°C with 5% CO_2 in cRPMI in the presence of inactivated TIGR4 at a cell:bacteria ratio of 1:10. Unstimulated control splenocytes were cultured in cRPMI alone. Upon stimulation, cell replicates were centrifuged at 500 x *g* for 10 minutes at 4°C. The supernatant was discarded, the pellet resuspended in 50 µl of lysis buffer RA1 (Macherey-Nagel, Germany), flash-frozen in liquid nitrogen, and stored at -70°C for subsequent RNA extraction.

2.7 RNA Extraction, Library Preparation, and Sequencing

The RNA purification was performed with the NucleoSpin® RNA kit (Macherey-Nagel) following manufacturer's

instructions, and, before DNAse treatment, the extracted RNA was quantified by the Qubit® 2.0 Fluorometer (Invitrogen by Thermo Fisher Scientific, USA), using the Qubit RNA BR (Broad-Range) Assay Kit.

Contaminating DNA was removed from the extracted RNA by adding 10X TURBO DNase Buffer (TURBO DNase, Ambion by Thermo Fisher Scientific) and 1 µl of TURBO DNase (Ambion), and samples were incubated at 37°C for 30 minutes. After purification by the RNA Clean & Concentrator Kit (Zymo Research, USA), the obtained RNA was quantified using the Qubit® RNA BR Assay Kit.

Library preparation was performed as described in a previous publication (Santoro et al., 2021), using the Ion AmpliSeq™ Transcriptome Mouse Gene Expression Kit from AmpliSeq (Thermo Fisher Scientific), allowing the amplification of 23,930 target genes. Libraries were diluted to 50 pM and pooled in equal volumes (7 µl), with eight individual samples per pool and loaded onto Ion PI™ Chip v3 using the Ion Chef™ Instrument. Sequencing was performed using Ion Proton™ Sequencer.

All described steps were performed according to the manufacturer's instructions.

2.8 RNA-Sequencing Data Analysis

R software in version 3.6.3 was used for transcriptomic data analysis. The DESeq2 package (Love et al., 2014), performs differential expression analysis and multiple test correction, returning values of LogFC, and adjusted p values. Genes with an adjusted p-value smaller than 0.05 and an absolute value of log2 Fold Change greater than 0.5 were classified as differentially expressed and then used in the enrichment analysis performed by the hypergeometric test from the *tmod* package (Weiner 3rd and Domaszewska, 2016) using the Blood Transcription Modules (BTM) database (Li et al., 2014).

2.9 Cytokine Data Analysis

R software in version 3.6.3 and the software GraphPad Prism 8.0 were used to perform the statistical analysis. The cytokine concentrations between stimulated and unstimulated samples were compared using the Wilcoxon signed-rank test, a non-parametric test used to compare two related samples. Samples from different time points were analyzed using the Mann-Whitney test, a non-parametric test for non-matched samples. A p-value ≤0.05 was considered statistically significant.

2.10 Biomarker Analysis

The DaMiRseq (Chiesa et al., 2018) package was used to find possible biomarkers of the host response to the second stimulus, the inactivated bacteria. Stimulated samples were selected and divided into three groups: baseline, infected samples at days 1 and 2 (early time points), and infected samples at days 4 and 7 (late time points). Following a pipeline that permits normalization, data adjustment, and feature selection, the DaMiRseq package ranked the most important features to distinguish the three classes. The number of selected genes was chosen based on the importance established by the package; genes with a scaled importance score higher than 0.5 were chosen (**Supplementary Image 1**).

2.11 Data Integration

To integrate RNA-sequencing results and the Cytokines Bioplex, we selected the concurrent samples from both experiments, and we performed the sparse version of Partial Least Squares (sPLS), provided by the MixOmics package. The PLS is a multivariate method to integrate two high dimensional matrices, maximizing the covariance between components from two data sets, in our case, the transcriptomic and cytokines assay data. The sparse version, sPLS, applies LASSO penalization in each pair of loading vectors from PLS, performing feature selection and providing the correlation values between the main features in each data set (Lê Cao et al., 2008).

According to the Q^2 criterion, two components would be sufficient to run the model (Q^2 of 0.33931900 and 0.08639437). As suggested by the *tune.spls* function, the optimal variable number was chosen in each component resulting in 16 genes for component 1 and 25 from component 2.

3 RESULTS

Groups of six C57BL/6 mice were intranasally inoculated with 10^7 CFUs of TIGR4 *S. pneumoniae* to generate lung infection. To study the systemic response induced at early time points afterinfection, we sacrificed animals after 1, 2, 4, and 7 days andisolated their splenocytes. We then stimulated splenocytesfrom each single mouse with formalin-inactivated whole pneumococcal cells and investigated the host responses by transcriptomic analysis and assessment of cytokine production (**Figure 1**).

3.1 Evidence of Pneumococcal Lung Infection in Mice

The weight loss of animals after infection is a critical clinical parameter of disease in the mouse model of infection, evaluated in different challenge murine models with different pathogens (Trammell and Toth, 2011; Pettini et al., 2015; Fiorino et al., 2021). Mouse body weight was measured every 24 hours for a period of seven days. Uninfected mice increased their body weight over time, which reflected their health status. Compared to naive mice, infected mice experienced a significant decrease in body weight soon after infection, and the average difference in the weight between the classes increased over time, being 1.07 grams at day 2 after infection and 1.52 grams at day 7 (**Figure 2A**).

The significance of these findings was assessed using the Mann-Whitney test, which showed significant differences in the weight from day 2 to day 7 after infection, indicating a long-lasting effect of the infection.

Pneumococcal cells were counted in the lungs of infected mice. Cell counts had a peak at day 4 after infection (**Figure 2B**). It is worth to note that, for each time point assayed, 2-5 mice had no detectable pneumococcal cells, suggesting that mice are able

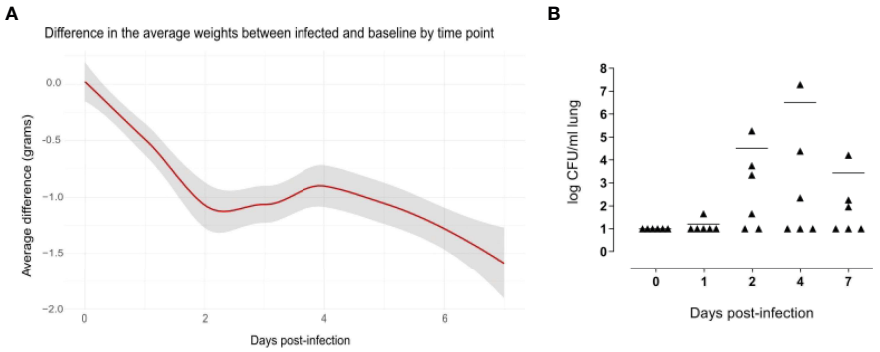


FIGURE 2 | Evidence of pneumococcal lung infection. Comparison of body weight variation of infected versus baseline **(A)**. Infected (n=36) and uninfected (n=5) mice were weighed every 24 hours for a period of 7 days. Values were obtained subtracting the pre-inoculum body weight from the body weight at each time-point, and then subtracting the mean in the control group from mean of the infected group in each time point. The average differences are expressed in grams and the gray shade represents the 0.95 confidence interval. From days 2 to 7 there were significant differences between the infected and baseline groups ($p < 0.05$, calculated using Mann-Whitney test. Pneumococcal cell counts in the lungs **(B)**. The lungs of mice sacrificed at 1, 2, 4 and 7 days after infection (n=6 per group) were collected, homogenized in a final volume of 1 ml and plated using a multilayer plating procedure. Pneumococcal cells were counted after 24 hours and 48 hours of incubation. Lungs of uninfected mice (0 days post-infection) were plated as a negative control. Data are expressed as CFUs/ml lung. The lower limit of detection was 10 CFU/ml lung. Average cell counts had a peak at day 4. For each time point, there were at least 2 mice without detectable pneumococci.

to spontaneously clear pneumococcal lung infection at an infectious dose of 1x10^7 CFUs of *S. pneumoniae* TIGR4. When setting up the mouse model, we also counted pneumococcal cells in the blood of six mice at 6 and 12 hours after intranasal infection, and in the blood of 12 mice at 24 and 96 hours after infection. Of those, only one animal had detectable pneumococcal cells in the blood (2.4x10^3 CFU/ml) at 24 hours after infection, suggesting that the infection is essentially limited to the mouse lungs without significant systemic spreading.

3.2 *In vitro* Splenocyte Stimulation With Pneumococcal Strain TIGR4 Activates Several Genes Related to Both Branches of the Immune System

Transcriptomic data from spleens of infected and uninfected mice with or without homologous *in vitro* stimulation were analyzed. We performed an Independent Principal Component Analysis (IPCA) and its sparse version, sIPCA, both proposed by MixOmics package (Lê Cao et al., 2011), (i) to observe the distribution of our data, (ii) to understand how stimulation at different time points affects the clustering of samples, (iii) to identify the genes responsible for the main variance among samples, and (iv) to find possible outlier samples.

The IPCA approach (**Supplementary Image 2**) yielded a better clusterization among experimental groups and time points compared to PCA (data not shown). An outlier control sample was detected and removed. The sparse version of the IPCA (sIPCA, **Figure 3**) applies soft-thresholding in the independent loading vectors in IPCA, performing feature selection. The graph shows the presence of two well-defined groups in the sIPC1: the stimulated and unstimulated samples.

To better understand the genes that drive the formation of these clusters, the normalized expression values of the 50 genes selected by the first component of the sIPCA were divided into two heatmaps (**Figure 3**). Genes positively correlated with the first component (driving the unstimulated cluster) included positive and negative regulators of the immune response, and they presented a decreased expression after stimulation. The genes negatively regulated with the first component (clustering the stimulated samples) were related to cytokines, chemokines, and inflammation, all of them presenting an increased expression compared to unstimulated samples.

3.3 Stimulation of Splenocytes From Infected Mice Highlights Biological Pathways of Pneumococcus Infection

We then proceeded with the differential expression analysis using the *DESeq2* package. To understand the biological alterations caused by the infection and the subsequent *in vitro* stimulation with inactivated pneumococcus, enrichment analysis was performed using three different comparisons. (i) Spleens from infected mice, (ii) stimulated spleens from non-infected mice, and (iii) stimulated spleens from infected mice, at different time points after infection, were all compared with control spleens.

The number of differentially expressed genes (DEGs) for each condition at each time point is presented in **Figure 4**. As expected, the stimulation of infected samples led to a higher number of DEGs compared to only infected samples, including specific genes that were not differentially expressed in the infection or stimulation alone. Days 4 and 7 presented the highest values of specific DEGs, in particular, new immune related genes and microRNAs were found, such as Il2, Foxp3, Il16a, Ccr1 and Mir155hg.

The enrichment analysis was performed using the *Blood Transcription Modules (BTM)* database and the *tmod* package, the complete results of the different groups are reported in the **Supplementary Data Sheet 1**. **Figure 5** shows a summary of the

FIGURE 3 | Distribution of gene expression data and changes after stimulation. The sIPCA proposed by MixOmics revealed two main clusters, one formed by stimulated samples (shaded in purple) and other formed by unstimulated samples. The IPC1 (x-axis) represents the variance between stimulated and unstimulated samples of infected and non-infected mice. IPC2 (y-axis) shows the variance between baseline samples with respect to the other groups in different study days (1, 2, 4, 7). After infection, samples from days 1 and 2 (early time points) form a smaller cluster closer to baseline samples (lighter green), while samples from days 4 and 7 (late time points) present a higher dispersion (darker green). The normalized expression values for the 50 most important genes selected by the sIPCA in the first component were represented in heatmaps, with lower values represented in blue and higher values in red and time points represented in the bottom part, following the same green scale of the sIPCA. The heatmap on the right, shows the 27 genes positively correlated with the first component, which include positive and negative regulators of the immune system and drive the formation of the cluster of unstimulated samples. The heatmap on the left shows the 23 genes negatively correlated with the first component, driving the stimulated samples to form a separate cluster. Most of these genes are related to cytokines and chemokines (Csf2, Ccl4, Il1a, Il1b, Il1rn, Il2ra, Cxcl1, Cxcl3, Cxcl2, Ccl3, Slc7a5) while others are related to immunity and inflammation (Cd83, Acod1, Slc7a11, Gadd45g, Nlrp3, Ptgs2 and Igfbp4).

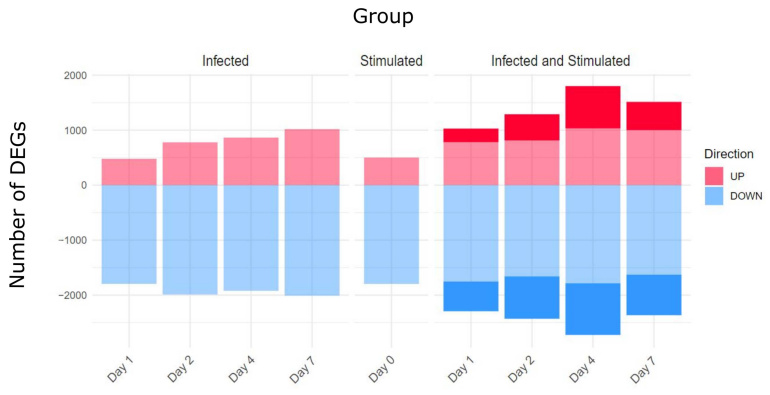


FIGURE 4 | Differentially Expressed genes (DEGs) for comparison and Time point. Three main comparisons against baseline controls were defined: infected samples, stimulated samples and infected and stimulated samples. Differentially Expressed Genes (DEGs) were obtained using the DESeq2 package and establishing thresholds of FDR < 0.05 and absolute logFC > 0.5. The stimulation of spleens from infected mice led to a higher number of DEGs compared to only infected or only stimulated spleens (number of specific genes highlighted in the figure).

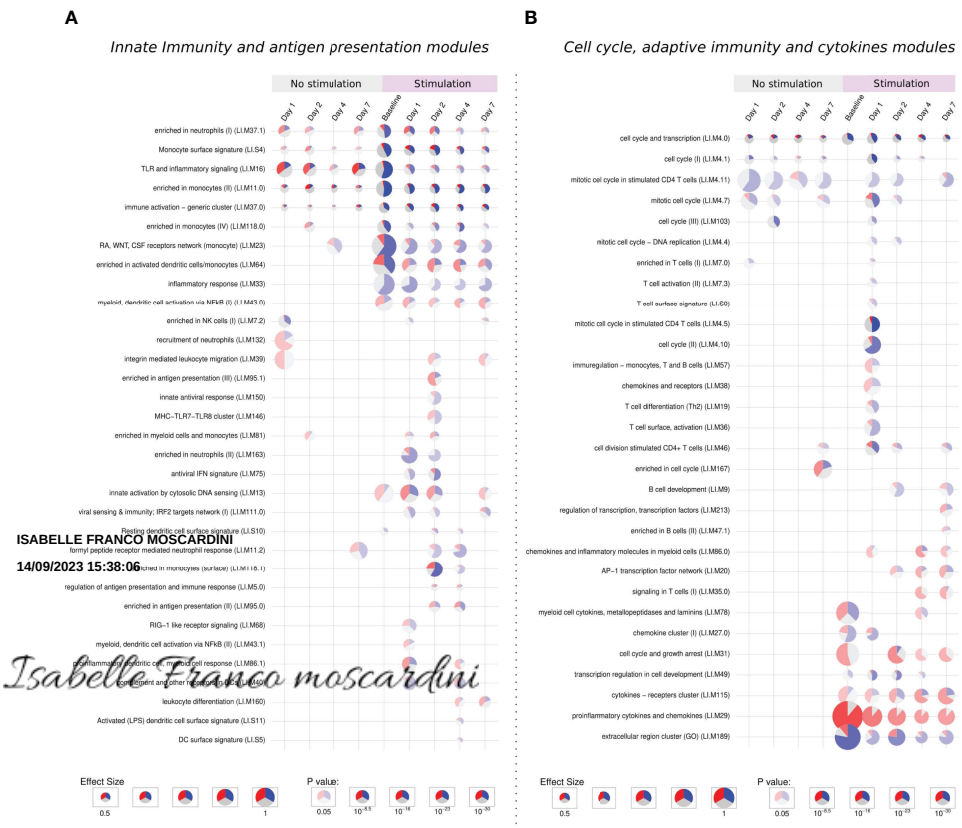


FIGURE 5 | Tmod enrichment analysis. From the list of genes provided by the DESeq2 package, genes were selected by their FDR value (< 0.05) and absolute log2 Fold Change (> 0.5). The enriched blood transcription modules were obtained by the Hypergeometric test. **(A)** Modules related to innate immunity and antigen presentation, **(B)** modules related to cell cycle, adaptive immunity and cytokines modules. The effect size is proportional to the size of the pie, while the adjusted p-value is proportional to colour intensity. Within each pie, the proportion of significantly upregulated and downregulated genes is shown in red and blue, respectively. The gray portion of the pie represents genes that are not significantly differentially regulated.

main immune system modules activated for each comparison and time point. In total, 87 modules were significantly enriched, only 3 of them being specifically activated in infected samples, while 40 modules were only activated after stimulation of previously infected samples.

3.3.1 Activation of Extracellular Matrix, Cell Adhesion, and Innate Immune Response Modules

Five modules were consistently activated at almost all time points in both unstimulated and *in vitro* stimulated groups. Related to monocytes, immune activation, TLR signaling, and cell cycle, these modules showed a different pattern after stimulation, presenting more down-regulated genes. Following the same direction, modules related to the extracellular region, monocytes, and cell cycle are especially enriched in down-regulated genes after stimulation (**Figures 5A, B**).

The "extracellular region cluster" module shows the downregulation of genes involved in the interaction with extracellular components, growth control, and the vascular endothelium/angiogenesis (HSPG2, GH1, ENG). Moreover, the monocyte chemoattractant CCL2 is also down-regulated, while CCL18, important for the recruitment of T lymphocytes but not monocytes, is up-regulated.

The stimulation down-regulates genes responsible for the proliferation and differentiation of monocytes and macrophages like CSF2RA, CSF1R, and CSF3R, the latter one also important for adhesion. In monocytes modules, other genes linked to the extracellular matrix and cell adhesion followed the same behavior.

The downregulation of extracellular matrix genes could be due to the process of *in vitro* stimulation, decreasing the cell adhesion to the plate surface.

3.3.2 Activation of Cell Cycle, Cytokines, and Adaptive Immune Response Modules

The stimulation of infected samples led to the enrichment of many biological pathways not activated in the previous comparisons, including modules related to antiviral response, antigen presentation, T cells, B cells, and chemokines (**Figures 5A, B**).

On the first day, unstimulated samples presented the enrichment of T cell and cell cycle modules. After stimulation, these same modules are activated, together with many others related to T cells and cell cycle, in both cases enriched mainly by down-regulated genes.

In T cell modules we find down-regulated cell-cycle genes and genes linked to cell adhesion, like VCAM1 and SIR3PG, while ITGA4, another adhesion-related gene, was up-regulated in unstimulated samples but presented no change after stimulation. Negative regulators of the T cell activity (LILRB4, LILRB3, SIT1) were also downregulated, while the few up-regulated genes were mainly related to T cell activation (CD3E, GRAP2, CDCA7, and LAT).

On the other hand, the specific modules in late time points were mostly activated by up-regulated genes. We observe a stronger activation of the "cytokines – receptors cluster" module and the specific enrichment of pathways like leukocyte differentiation, signaling in T cells, enriched in B cells, among other modules.

Cytokine modules are activated from the stimulation of baseline samples and looking inside these modules, indeed we see many up-regulated genes independently of the time point, especially those from the CCL family, IL1A, IL1RN, and TNF. Other genes such as CSF2, IL2RA, IL6, IL10 and IL1B present a modest increase in stimulation of baseline samples, but a major up-regulation at late time points.

Despite the high number of activated modules, the response to the stimuli after a previous infection does not show general up-regulation of the immune and inflammatory response. This second contact with the pathogen through the *in vitro* stimulation permitted us to appreciate biological processes which could not be detected in the primary infection, especially those related to antigen presentation, adaptive immunity, and cytokines. These processes are possibly related to a recall immune response starting within the first days after infection.

3.4 Cytokines Assay Suggests the Promotion of Innate and Adaptive Immune Responses From Day 4 After Infection

Regarding the concentration of cytokines in splenocyte culture supernatants, the infection without subsequent stimulation did not result in significant increases in the concentration of cytokines, with exception of IL-17a at day 7 after infection (data not shown).

The stimulation process induced a significant increase in KC and MIP-1a, compared to baseline samples (median of differences of 42.46 and 184.4 pg/ml, respectively). Despite cytokine changes between stimulated and control samples being noticeable in early stages, they increased considerably upon *in vitro* stimulation, at days 4 and 7 after the infection (**Figure 6**).

The comparison of stimulated samples from days 4 and 7 after infection with only infected samples showed a significant increase in all cytokine concentrations, with exception of MCP-1 and IL12p40 (**Figure 6** and **Supplementary Image 3**), suggesting the involvement of both innate and adaptive branches of the immune system. This increase was more accentuated on day 7, in which the difference in the median between the stimulated and unstimulated groups was 2597 pg/ml for IL-17A, 769 pg/ml for GM-CSF and 374 pg/ml for IFN-gamma.

3.5 Gene Expression and Cytokines Data Integration Indicate Specific Patterns of Recall Immune Response After Stimulation

To identify the genes correlated with the increase in the concentration of cytokines, especially at late time points, data integration was performed using the sparse version of Partial Least Squares (sPLS), from MixOmics package. PLS can integrate two types of data measured on the same sample by maximizing the covariance between the components of each data set. The sparse version applies LASSO $\ell 1$ penalizations in PLS analysis to perform feature selection.

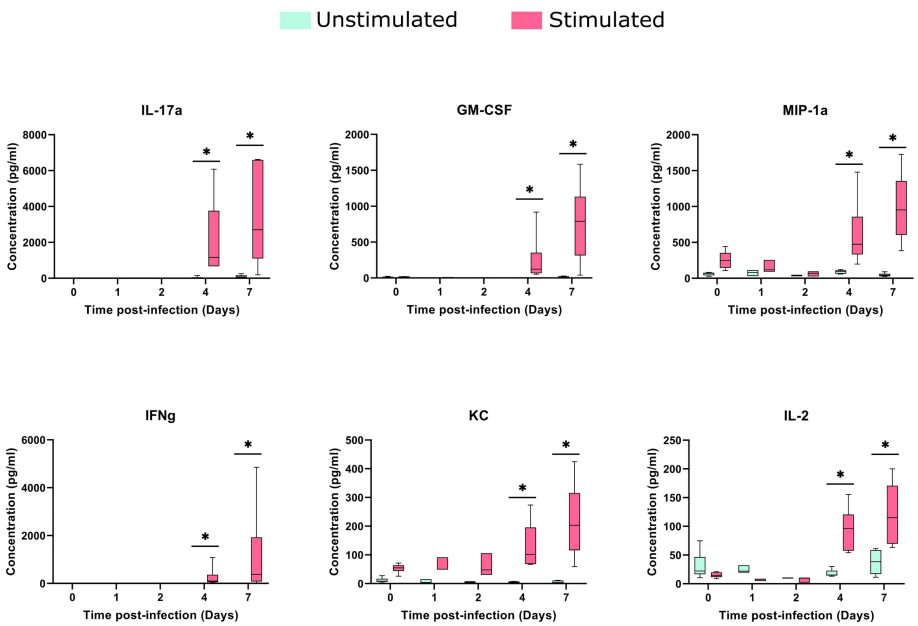


FIGURE 6 | Cytokines concentrations in the spleens after TIGR4 pneumococcal infection. The spleens were collected from infected mice at different time points and splenocytes were cultured for 72 hours in the presence or not of a stimulus (formalin-inactivated TIGR4 pneumococcal strain). The supernatants were collected and the concentration of 23 cytokines were assessed by Luminex immunoassay. Data was analyzed using GraphPad Prism software. To compare Stimulated to Unstimulated samples we used the Wilcoxon paired test. Concentrations from six of the cytokines are presented in the figure (for all the 23 cytokines see **Supplementary Image 3**). When compared to only infected samples, all the six cytokines presented a significant (*= $P<0.05$) increase when stimulated in days 4 and 7 after infection.

As expected, the *in vitro* stimulated samples formed a different cluster compared to the non-stimulated samples, although there is a different behavior regarding time points in each cluster (**Figure 7A**). In the non-stimulated cluster there is a perturbation caused by infection, but some samples from day 7 cluster together with control samples from day 0. On the other hand, *in vitro* stimulated samples presented a different pattern, samples from days 4 and 7 form a new cluster, driven by the increase in cytokine concentration and the expression of certain genes (**Figure 7B**).

By performing data integration and feature selection, sPLS identifies the genes whose expression is strongly associated with the concentrations of cytokines, providing the correlation value for each variable. The genes with the highest values of correlation with the 23 cytokines were Cd69, Csf2, Il2ra, and Il2 (**Figure 7C**). Other genes related to the immune system (Foxp3, Tnfrsf4, Tnfrsf9, Il10, and Il6) were also found positively correlated with the cytokines.

3.6 Possible Biomarkers Elicited by *In Vitro* Stimulation

We aimed to understand if feature selection could summarize the impact of a previous infection on stimulated samples, indicating possible biomarkers of this infection. We applied the DaMiR-seq package, which provides data normalization, feature selection, and classification, based on different machine learning techniques. Three groups were established based on the transcriptomics and cytokine data distribution, focusing on the stimulation of uninfected samples, samples from early time points after infection (days 1 and 2), and samples from late time points (days 4 and 7).

Eleven genes were chosen by applying a threshold of 0.5 to the scaled importance score identified by the DaMiR-seq package (**Supplementary Image 1**). These 11 genes allowed a clear clusterization of the three groups (**Figure 8**).

When compared to baseline stimulated samples, Fpr1, Nlrp3, and Slpi presented an increased expression in infected stimulated samples, independently of the time point. The stimulation of samples from early time points after the infection led to the increase in the expression of other inflammatory genes like Serpinb2 and Chil1 (Chitinase-3-like protein 1), and these values started to decrease in the subsequent days. At late time points, three other important genes, related to cytokine activity, had their expression increased when compared to the other groups: Ccr4, Csf2, and Il2.

Despite the small number of samples being a limitation for this type of analysis, the feature selection summarizes the new immunological processes that arise after the stimulation of infected samples and suggests the use of the *in vitro* stimulation model to detect the presence of a previous pneumococcal infection by measuring the expression of a few genes.

4 DISCUSSION

To characterize the host response to *S. pneumoniae* we proposed a murine model of intranasal infection followed by an *in vitro*

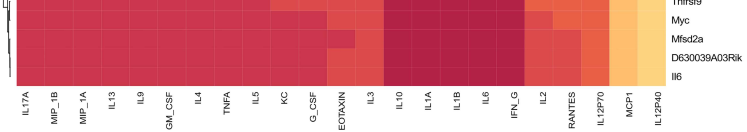


FIGURE 7 | Data integration using sPLS (Sparse Partial Least Squares). Data integration suggests a different pattern in stimulated samples from days 4 and 7 after infection, driven by the expression of a few genes and the concentration of most cytokines. **(A)** sPLS: Partial Least Squares regression applying the tune.spls function (MixOmics), 16 genes were selected for component 1 and 25 genes for component 2. Colors represent samples from different groups: infected or non-infected and with or without *in vitro* stimulation. Numbers indicate the study days. **(B)** Correlation Circle Plot: the selected genes (blue) and cytokines (red) are represented in a correlation circle plot. Variables positively associated are projected in the same direction from the origin, variables negatively correlated are projected in opposite directions. Variables displayed in a perpendicular angle are not correlated and the greater the distance from the origin the stronger the association. For example, the cytokine IL1B is positively correlated with the IL2 gene, but is negatively correlated with the gene Pdzd2. In general, cytokines and several genes related to cytokines are accumulated on the left side of the graph, indicating a high correlation among them. **(C)** Clustered Image Map (CIM): correlation between genes (mRNA), reported in vertical, and Cytokine production, in horizontal. The map highlights the correlation values for these variables for the genes selected on the first component of the sPLS. High positive correlations are represented by dark red, while high negative correlations by dark blue. Genes related to cytokines such as Il2ra, Csf2, Il10, Il6 and Il2 are positively correlated with most cytokines concentration. The select genes are strongly correlated with almost all cytokines, with exception of MCP1 and IL12p40.

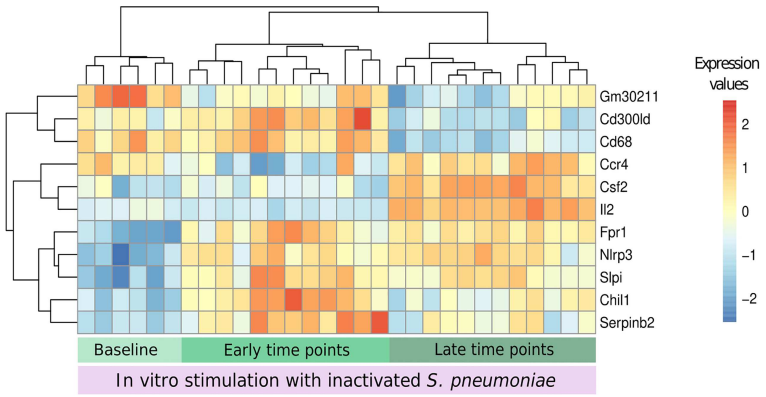


FIGURE 8 | Biomarkers of recall immune response. The DaMiR-seq package combines different classification methods to select possible biomarkers. After selecting the stimulated samples, this package was used to find genes involved in the differences between infected and uninfected samples after stimulation. Groups were established based on the results of data distribution (IPCA and sIPCA) and cytokines, where we observed a shared pattern between early time points (days 1 and 2) and another pattern at late time points (4 and 7). Eleven genes were selected based on the drop of the feature importance (**Supplementary Image 1**). The normalized expression values of the selected genes permitted to form three clear clusters.

stimulation of splenocytes with inactivated bacterial whole cells, at different time points after infection. Using a transcriptomic-based approach, our study has highlighted genes and biological pathways associated with the stimulation of baseline and previously infected samples, as well as the cytokines involved in the same processes.

In accordance with the genes selected by the sIPCA, the enrichment analysis has shown that the simple presence of inactivated bacteria leads to the activation of cytokines genes and different immune system pathways, mainly related to innate immunity. However, when this stimulation occurs in previously infected samples, there is a higher number of DEGs, revealing 40 new biological modules distributed across time points.

Stimulated samples presented downregulation of monocytes modules, together with the upregulation of antigen presentation and cytokine modules, which were also reported in gene expression data from alveolar macrophages (AM) in a pneumococcal colonization study, when comparing volunteers that developed carriage or not after experimental human pneumococcal challenge (Mitsi et al., 2020). In fact, the study did not suggest an increase in monocytes, but a higher monocyte-AM differentiation in people that developed carriage. On the other hand, monocytes seem to be recruited at the nose after the establishment of carriage (Jochems et al., 2018).

Recent vaccine studies have emphasized that innate immunity modules, including antigen presentation and dendritic cell activation, demonstrate stronger activation after a second contact with the antigen. Using an *in vivo* boost with the antigen alone following the priming with a chimeric vaccine against *Mycobacterium tuberculosis*, Santoro et al. have observed a faster and more robust response of dendritic cells and antigen presentation (Santoro et al., 2018). Similar results were recently observed in a different context, with an mRNA vaccine against SARS-CoV-2, in which the second dose activated new antigen presentation modules (Arunachalam et al., 2021).

Transcriptomics results have shown the presence of a particular response after a second contact with the pathogen and the concentrations of the cytokines suggested that this response is marked by different patterns of activation, with the stimulation allowing a better classification between early time points (1 and 2 days) and late time points (4 and 7) after infection. Moreover, the activated modules and the concentration of the soluble modulators suggested that both innate and adaptive branches of the immune system are promoted by the stimulation, suggesting cooperation between them.

In our model, cytokines secreted by macrophages, like MIP-1a (CCL3), KC (CXCL1), IL-1a, IL-1b, and IL-6 had a small, although significant, increase after stimulation of baseline samples. After stimulation of early time points, a small increase in the concentration of MIP-1b, RANTES, and TNF-a was observed, although not statistically significant. When the *in vitro* stimulus occurs at late time points after infection, the concentration of all these cytokines significantly increases, especially at day 7. In fact, innate immune responses to pneumococcus are known for polarization towards Th1 and Th17 responses through the release of cytokines (Bogaert et al., 2009).

Following this reasoning, it was expected that after the stimulation of baseline samples the cytokines linked to the adaptive immunity activation such as IL-17, IFNg, and IL-2 did not present an increased concentration in comparison to unstimulated samples. G-CSF and GM-CSF presented a small increase, although not significant. Again, at early time points no important changes are seen, but at late time points, a significant increase in the concentration was observed for all these cytokines, suggesting the activation of a Th1 and Th17 response starting around day 4. A strong Th1 response characterized by high levels of IFNg was also demonstrated in a murine model of bacterial meningitis by type 4 *S. pneumoniae*, already 48 hours after infection. (Pettini et al., 2015).

A previous study has highlighted the action of CD8+ T cells in helping AM to develop high MHC II expression after adenovirus

infection, a process that started coincidently with the entry of T cells in the alveolar tissue, around 5 days after the infection (Yao et al., 2018). The activation of T cells in the spleen could follow a similar behavior in supporting macrophage activity and consequently increasing cytokine release.

Biomarker analysis and sPLS integration were employed to find genes that characterize the recall immune response and correlate with the increase in the concentration of the cytokines. Among the genes found positively correlated by the sPLS method, many were linked to the immune response. The TNF receptors Tnfrsf9 and Tnfrsf4, are important for Th1 promotion and CD4 responses and, together with Il2, Il2ra, Foxp3 and IL10 participate in the "NF-kappaB signaling" biological pathway (Cho et al., 2021). Moreover, these genes are also associated with regulatory T cells, along with Cd69 and Il6, two other features found correlated with cytokines in the same analysis (Maloy and Powrie, 2005; Kimura and Kishimoto, 2010; Chaudhry et al., 2011; Yu et al., 2018; Hinterbrandner et al., 2021).

The eleven genes selected as possible biomarkers are capable of correctly clustering the stimulated samples in the studied groups (baseline, early, and late time points, **Figure 8**). These genes could be cross validated in future studies using the same model of pneumococcal lung infection to study vaccine strategies and antimicrobial therapies. A link with pneumococcal infection, colonization, or vaccination was established in the literature for most of the selected genes. The Il2 and Csf2 genes were not only the first and third most important genes for the classification of samples regarding the presence of a previous infection but they were also among the genes with the highest correlation with the concentration of different cytokines, together with the Il2ra gene. This highlights the importance of the IL-2 signaling pathway to the described recall response. Indeed, different vaccine studies reported an increase in IL-2 cytokine after restimulation with pneumococcal proteins or peptides from these proteins (Kataoka et al., 2011; Singh et al., 2014; Elhaik Goldman et al., 2016; Converso et al., 2017).

Csf2 gene encodes for Granulocyte/Macrophage colony-stimulating factor (GM-CSF), a cytokine that presented one of the highest concentrations after stimulation of infected samples from late time points. Previous studies have described the increase in Csf2 expression and GM-CSF concentration in the lungs from mice infected intranasally with *S. pneumoniae* (Steinwede et al., 2011). *In vitro* stimulation of PBMCs with *S. pneumoniae* has also increased the concentration of this cytokine. Furthermore, a protective role of GM-CSF in pneumococcal infection was described with intra-alveolar administration of this cytokine (Schmeck et al., 2004; Steinwede et al., 2011) and the resistance to lung infection attributed to the microbiota was found to be through GM-CSF signaling (Brown et al., 2017).

The lack of Fpr1 and Chil1 led to a higher mortality rate in murine models of pneumococcal meningitis and pneumonia, respectively (Dela Cruz et al., 2012; Oldekamp et al., 2014).

Slpi is involved in the innate immune response to bacterial infections, regulating the NF-kappa-B activation and inflammatory responses. This gene was up-regulated in the lungs of mice infected with pneumococcus, but the same was not observed in the spleen, suggesting that its expression is modulated at the site of inflammation in the presence of inflammatory stimuli (Abe et al., 1997). Our data suggested a similar result, since Slpi expression did not change in the spleen of infected samples, but only increased after the *in vitro* stimulation.

Cd300ld presents no clear link with pneumococcal infection, but its encoded protein, an activating receptor in myeloid and mast cells, was downregulated in the blood of mice infected with *Streptococcus suis* (Dai et al., 2018).

The link of most of the selected genes with the physiopathology of pneumococcal infection supports the use of feature selection and machine learning techniques to unveil gene signatures, potentially finding new features and/or assigning new roles to genes involved in a process, such as recall responses. The changes in cytokines concentration and gene expression are two important ways to assess immunological information after infection or vaccination. Our findings suggest that *in vitro* stimulation is an important step to understanding the systemic response to pneumococcal lung infection and the immunological memory generated by this bacteria. The analysis of transcriptomic and cytokine data revealed a clustering of the samples based on the stage of infection (early vs late), with more intense signals at late time points. Integrative analysis identified few genes, related to the immune system, which could categorize the samples based on the infection stage and which may be useful in future studies to monitor vaccine immune response and experimental therapies efficacy.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the GEO database under accession number: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?&acc=GSE199605. The bioinformatic analysis can be accessed at https://github.com/IsaMoscardini/Spleen_stimulation.

ETHICS STATEMENT

The animal study was reviewed and approved by the Italian Ministry of Health with authorization n° 304/2018-PR.

AUTHOR CONTRIBUTIONS

GP, FI and FS conceived and designed the experiments. FS and MC prepared the bacteria for animal infection and for *in vitro* stimulation. MC, FF, SG and EP performed animal experiments. MC and CG performed transcriptomic analysis. MC and FF analysed cytokines. DM secured funding. IM analysed data and drafted the paper with contributions from FS, AG and GP. All the authors reviewed, edited and approved the final version of the manuscript.

FUNDING

This study was carried out with financial support from the Commission of the European Communities, Seventh

Framework Programme, Innovative Medicines Initiative Joint Undertaking "Biomarkers for Enhanced Vaccine Safety" project BioVacSafe (IMI JU Grant No. 115308).

IM received a PhD fellowship under the Marie Sklodowska-Curie actions (MSCA) – Innovative Training Networks (ITN), Project VacPath (Novel vaccine vectors to resist pathogen challenge) grant agreement No 812915 funded by the European Union's Horizon 2020 research and innovation programme.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcimb.2022.869763/full#supplementary-material

Supplementary Image 1 | Feature Importance. The DaMiRseq package ranks the selected features using RReliefF, a multivariate filter technique that assesses the relevance of the features. The graph shows the importance of each feature, indicating that Il2 gene is the best classificator.

Supplementary Image 2 | Independent Principal Component Analysis (IPCA). The IPCA analysis from mixOmics package displays the distribution of samples by their gene expression, pointing out two main clusters that split samples according to their stimulation status. Numbers represent the time point of each sample, leading to the formation of three groups: baseline (light green), early time points - days 1 and 2 (green), and late time points - days 4 and 7 (dark green). The purple shading highlights the clusters composed by stimulated samples.

Supplementary Image 3 | Cytokines boxplots (complete panel). Boxplots comparing the concentration of 23 different cytokines in stimulated and unstimulated samples at baseline and at different time points after infection.

Supplementary Data Sheet 1 | Modules enriched in each comparison and time point.

REFERENCES

Abe, T., Tominaga, Y., Kikuchi, T., Watanabe, A., Satoh, K., Watanabe, Y., et al. (1997). Bacterial Pneumonia Causes Augmented Expression of the Secretory Leukoprotease Inhibitor Gene in the Murine Lung. *Am. J. Respir. Crit. Care Med.* 156, 1235–1240. doi: 10.1164/ajrccm.156.4.9701075

Arunachalam, P. S., Scott, M. K. D., Hagan, T., Li, C., Feng, Y., Wimmers, F., et al. (2021). Systems Vaccinology of the BNT162b2 mRNA Vaccine in Humans. *Nature* 596, 410–416. doi: 10.1038/s41586-021-03791-x

Bogaert, D., Weinberger, D., Thompson, C., Lipsitch, M., and Malley, R. (2009). Impaired Innate and Adaptive Immunity to Streptococcus Pneumoniae and its Effect on Colonization in an Infant Mouse Model. *Infect. Immun.* 77, 1613–1622. doi: 10.1128/IAI.00871-08

Briles, D. E., Paton, J. C., Mukerji, R., Swiatlo, E., and Crain, M. J. (2019). Pneumococcal Vaccines. *Microbiol. Spectr.* 7(6). doi: 10.1128/microbiolspec.GPP3-0028-2018

Brown, R. L., Sequeira, R. P., and Clarke, T. B. (2017). The Microbiota Protects Against Respiratory Infection *via* GM-CSF Signaling. *Nat. Commun.* 8, 1512. doi: 10.1038/s41467-017-01803-x

Cerutti, A., Cols, M., and Puga, I. (2013). Marginal Zone B Cells: Virtues of Innate-Like Antibody-Producing Lymphocytes. *Nat. Rev. Immunol.* 13, 118–132. doi: 10.1038/nri3383

Chaudhry, A., Samstein, R. M., Treuting, P., Liang, Y., Pils, M. C., Heinrich, J.-M., et al. (2011). Interleukin-10 Signaling in Regulatory T Cells is Required for Suppression of Th17 Cell-Mediated Inflammation. *Immunity* 34, 566–578. doi: 10.1016/j.immuni.2011.03.018

Chiesa, M., Colombo, G. I., and Piacentini, L. (2018). DaMiRseq—an R/Bioconductor Package for Data Mining of RNA-Seq Data: Normalization, Feature Selection and Classification. *Bioinformatics* 34, 1416–1418. doi: 10.1093/bioinformatics/btx795

Cho, J.-W., Son, J., Ha, S.-J., and Lee, I. (2021). Systems Biology Analysis Identifies TNFRSF9 as a Functional Marker of Tumor-Infiltrating Regulatory T-Cell Enabling Clinical Outcome Prediction in Lung Cancer. *Comput. Struct. Biotechnol. J.* 19, 860–868. doi: 10.1016/j.csbj.2021.01.025

Converso, T. R., Goulart, C., Rodriguez, D., Darrieux, M., and Leite, L. C. C. (2017). Systemic Immunization With Recombinant Streptococcal Pneumoniae Nasopharyngeal Colonization in Mice. *Vaccine* 35, 149–155. doi: 10.1016/j.vaccine.2016.11.027

Dai, J., Lai, L., Tang, H., Wang, W., Wang, S., Lu, C., et al. (2018). Streptococcus Suis Synthesizes Deoxyadenosine and Adenosine by 5'-Nucleotidase to Dampen Host Immune Responses. *Virulence* 9, 1509–1520. doi: 10.1080/21505594.2018.1520544

Dela Cruz, C. S., Liu, W., He, C. H., Jacoby, A., Gornitzky, A., Ma, B., et al. (2012). Chitinase 3-Like-1 Promotes Streptococcus Pneumoniae Killing and Augments Host Tolerance to Lung Antibacterial Responses. *Cell Host Microbe* 12, 34–46. doi: 10.1016/j.chom.2012.05.017

Deniset, J. F., Surewaard, B. G., Lee, W.-Y., and Kubes, P. (2017). Splenic Ly6Ghigh Mature and Ly6Gint Immature Neutrophils Contribute to Eradication of S. Pneumoniae. *J. Exp. Med.* 214, 1333–1350. doi: 10.1084/jem.20161621

de Stoppelaar, S. F., Claushuis, T. A. M., Schaap, M. C. L., Hou, B., van der Poll, T., Nieuwland, R., et al. (2016). Toll-Like Receptor Signalling Is Not Involved in Platelet Response to Streptococcus Pneumoniae *In Vitro* or *In Vivo*. *PloS One* 11, e0156977. doi: 10.1371/journal.pone.0156977

Elhaik Goldman, S., Dotan, S., Talias, A., Lilo, A., Azriel, S., Malka, I., et al. (2016). Streptococcus Pneumoniae Fructose-1,6-Bisphosphate Aldolase, a Protein Vaccine Candidate, Elicits Th1/Th2/Th17-Type Cytokine Responses in Mice. *Int. J. Mol. Med.* 37, 1127–1138. doi: 10.3892/ijmm.2016.2512

Ercoli, G., Fernandes, V. E., Chung, W. Y., Wanford, J. J., Thomson, S., Bayliss, C. D., et al. (2018). Intracellular Replication of Streptococcus Pneumoniae Inside Splenic Macrophages Serves as a Reservoir for Septicaemia. *Nat. Microbiol.* 3, 600–610. doi: 10.1038/s41564-018-0147-1

Fiorino, F., Pettini, E., Koeberling, O., Ciabattini, A., Pozzi, G., Martin, L. B., et al. (2021). Long-Term Anti-Bacterial Immunity Against Systemic Infection by Salmonella Enterica Serovar Typhimurium Elicited by a GMMA-Based Vaccine. *Vaccines (Basel)* 9, 495. doi: 10.3390/vaccines9050495

Gerlini, A., Colomba, L., Furi, L., Braccini, T., Manso, A. S., Pammolli, A., et al. (2014). The Role of Host and Microbial Factors in the Pathogenesis of Pneumococcal Bacteraemia Arising From a Single Bacterial Cell Bottleneck. *PloS Pathog.* 10, e1004026. doi: 10.1371/journal.ppat.1004026

Hinterbrandner, M., Rubino, V., Stoll, C., Forster, S., Schnüriger, N., Radpour, R., et al. (2021). Tnfrsf4-Expressing Regulatory T Cells Promote Immune Escape of Chronic Myeloid Leukemia Stem Cells. *JCI Insight* 6, e151797. doi: 10.1172/jci.insight.151797

Iannelli, F., Santoro, F., Fox, V., and Pozzi, G. (2021). A Mating Procedure for Genetic Transfer of Integrative and Conjugative Elements (ICEs) of Streptococci and Enterococci. *Methods Protoc.* 4, 59. doi: 10.3390/mps4030059

Jochems, S. P., Marcon, F., Carniel, B. F., Holloway, M., Mitsi, E., Smith, E., et al. (2018). Inflammation Induced by Influenza Virus Impairs Human Innate Immune Control of Pneumococcus. *Nat. Immunol.* 19, 1299–1308. doi: 10.1038/s41590-018-0231-y

Kadioglu, A., Cuppone, A. M., Trappetti, C., List, T., Spreafico, A., Pozzi, G., et al. (2011). Sex-Based Differences in Susceptibility to Respiratory and Systemic Pneumococcal Disease in Mice. *J. Infect. Dis.* 204, 1971–1979. doi: 10.1093/infdis/jir657

Kaplan, S. L., Barson, W. J., Lin, P. L., Romero, J. R., Bradley, J. S., Tan, T. Q., et al. (2013). Early Trends for Invasive Pneumococcal Infections in Children After the Introduction of the 13-Valent Pneumococcal Conjugate Vaccine. *Pediatr. Infect. Dis. J.* 32, 203–207. doi: 10.1097/INF.0b013e318275614b

Kataoka, K., Fujihashi, K., Oma, K., Fukuyama, Y., Hollingshead, S. K., Sekine, S., et al. (2011). The Nasal Dendritic Cell-Targeting Flt3 Ligand as a Safe Adjuvant Elicits Effective Protection Against Fatal Pneumococcal Pneumonia. *Infect. Immun.* 79, 2819–2828. doi: 10.1128/IAI.01360-10

Kimura, A., and Kishimoto, T. (2010). IL-6: Regulator of Treg/Th17 Balance. *Eur. J. Immunol.* 40, 1830–1835. doi: 10.1002/eji.201040391

Lê Cao, K.-A., Rossouw, D., Robert-Granié, C., and Besse, P. (2008). A Sparse PLS for Variable Selection When Integrating Omics Data. *Stat. Appl. Genet. Mol. Biol.* 7. doi: 10.2202/1544-6115.1390

Li, S., Rouphael, N., Duraisingham, S., Romero-Steiner, S., Presnell, S., Davis, C., et al. (2014). Molecular Signatures of Antibody Responses Derived From a Systems Biology Study of Five Human Vaccines. *Nat. Immunol.* 15, 195–204. doi: 10.1038/ni.2789

Love, M. I., Huber, W., and Anders, S. (2014). Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data With Deseq2. *Genome Biol.* 15, 550. doi: 10.1186/s13059-014-0550-8

Maloy, K. J., and Powrie, F. (2005). Fueling Regulation: IL-2 Keeps CD4+ Treg Cells Fit. *Nat. Immunol.* 6, 1071–1072. doi: 10.1038/ni1105-1071

Mitsi, E., Carniel, B., Reiné, J., Rylance, J., Zaidi, S., Soares-Schanoski, A., et al. (2020). Nasal Pneumococcal Density Is Associated With Microaspiration and Heightened Human Alveolar Macrophage Responsiveness to Bacterial Pathogens. *Am. J. Respir. Crit. Care Med.* 201, 335–347. doi: 10.1164/rccm.201903-0607OC

Moffitt, K. L., Gierahn, T. M., Lu, Y., Gouveia, P., Alderson, M., Flechtner, J. B., et al. (2011). T(H)17-Based Vaccine Design for Prevention of Streptococcus Pneumoniae Colonization. *Cell Host Microbe* 9, 158–165. doi: 10.1016/j.chom.2011.01.007

Oldekamp, S., Pscheidl, S., Kress, E., Soehnlein, O., Jansen, S., Pufe, T., et al. (2014). Lack of Formyl Peptide Receptor 1 and 2 Leads to More Severe Inflammation and Higher Mortality in Mice With of Pneumococcal Meningitis. *Immunology* 143, 447–461. doi: 10.1111/imm.12324

Paranavitana, C., Zelazowska, E., DaSilva, L., Pittman, P. R., and Nikolich, M. (2010). Th17 Cytokines in Recall Responses Against Francisella Tularensis in Humans. *J. Interferon Cytokine Res.* 30, 471–476. doi: 10.1089/jir.2009.0108

Pettini, E., Fiorino, F., Cuppone, A. M., Iannelli, F., Medaglini, D., and Pozzi, G. (2015). Interferon-γ From Brain Leukocytes Enhances Meningitis by Type 4 Streptococcus Pneumoniae. *Front. Microbiol.* 6. doi: 10.3389/fmicb.2015.01340

Pichichero, M. E. (2017). Pneumococcal Whole-Cell and Protein-Based Vaccines: Changing the Paradigm. *Expert Rev. Vaccines* 16, 1181–1190. doi: 10.1080/14760584.2017.1393335

Santoro, F., Donato, A., Lucchesi, S., Sorgi, S., Gerlini, A., Haks, M. C., et al. (2021). Human Transcriptomic Response to the VSV-Vectored Ebola Vaccine. *Vaccines (Basel)* 9, 67. doi: 10.3390/vaccines9020067

Santoro, F., Pettini, E., Kazmin, D., Ciabattini, A., Fiorino, F., Gilfillan, G. D., et al. (2018). Transcriptomics of the Vaccine Immune Response: Priming With Adjuvant Modulates Recall Innate Responses After Boosting. *Front. Immunol.* 9. doi: 10.3389/fimmu.2018.01248

Schmeck, B., Zahlten, J., Moog, K., van Laak, V., Huber, S., Hocke, A. C., et al. (2004). Streptococcus Pneumoniae-Induced P38 MAPK-Dependent Phosphorylation of RelA at the Interleukin-8 Promotor. *J. Biol. Chem.* 279, 53241–53247. doi: 10.1074/jbc.M410376200

Schultz, M. J., Speelman, P., Zaat, S., van Deventer, S. J., and van der Poll, T. (1998). Erythromycin Inhibits Tumor Necrosis Factor Alpha and Interleukin 6 Production Induced by Heat-Killed Streptococcus Pneumoniae in Whole Blood. *Antimicrob. Agents Chemother.* 42, 1605–1609. doi: 10.1128/AAC.42.7.1605

Shao, J., Zhang, J., Wu, X., Mao, Q., Chen, P., Zhu, F., et al. (2015). Comparing the Primary and Recall Immune Response Induced by a New EV71 Vaccine Using Systems Biology Approaches. *PloS One* 10, e0140515. doi: 10.1371/journal.pone.0140515

Singh, R., Gupta, P., Sharma, P. K., Ades, E. W., Hollingshead, S. K., Singh, S., et al. (2014). Prediction and Characterization of Helper T-Cell Epitopes From Pneumococcal Surface Adhesin A. *Immunology* 141, 514–530. doi: 10.1111/imm.12194

Steinwede, K., Tempelhof, O., Bolte, K., Maus, R., Bohling, J., Ueberberg, B., et al. (2011). Local Delivery of GM-CSF Protects Mice From Lethal Pneumococcal Pneumonia. *J. Immunol.* 187, 5346–5356. doi: 10.4049/jimmunol.1101413

Trammell, R. A., and Toth, L. A. (2011). Markers for Predicting Death as an Outcome for Mice Used in Infectious Disease Research. *Comp. Med.* 61, 492–498.

Weiner, 3. J., and Domaszewska, T. (2016). Tmod: An R Package for General and Multivariate Enrichment Analysis. *Peer J Inc*. Preprints 4:e2420v1. doi: 10.7287/peerj.preprints.2420v1

Weiser, J. N., Ferreira, D. M., and Paton, J. C. (2018). Streptococcus Pneumoniae: Transmission, Colonization and Invasion. *Nat. Rev. Microbiol.* 16, 355–367. doi: 10.1038/s41579-018-0001-8

Wu, Y., Mao, H., Ling, M.-T., Chow, K.-H., Ho, P.-L., Tu, W., et al. (2011). Successive Influenza Virus Infection and Streptococcus Pneumoniae Stimulation Alter Human Dendritic Cell Function. *BMC Infect. Dis.* 11, 201. doi: 10.1186/1471-2334-11-201

Yao, F., Coquery, J., and Lê Cao, K.-A. (2012). Independent Principal Component Analysis for Biologically Meaningful Dimension Reduction of Large Biological Data Sets. *BMC Bioinf.* 13, 24. doi: 10.1186/1471-2105-13-24

Yao, Y., Jeyanathan, M., Haddadi, S., Barra, N. G., Vaseghi-Shanjani, M., Damjanovic, D., et al. (2018). Induction of Autonomous Memory Alveolar Macrophages Requires T Cell Help and Is Critical to Trained Immunity. *Cell* 175, 1634–1650.e17. doi: 10.1016/j.cell.2018.09.042

Yu, L., Yang, F., Zhang, F., Guo, D., Li, L., Wang, X., et al. (2018). CD69 Enhances Immunosuppressive Function of Regulatory T-Cells and Attenuates Colitis by Prompting IL-10 Production. *Cell Death Dis.* 9, 905. doi: 10.1038/s41419-018-0927-9

Zhan, Y., and Cheers, C. (1995). Differential Induction of Macrophage-Derived Cytokines by Live and Dead Intracellular Bacteria *In Vitro*. *Infect. Immun.* 63, 720–723. doi: 10.1128/iai.63.2.720-723.1995

Conflict of Interest: Authors IM and AG are employed by Microbiotec srl.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

3.6 Final discussion

This work provided a mouse model to study underlying mechanisms of pneumococcal infection through the stimulation of splenocytes harvested from mice at different time points after infection. Our results demonstrated the importance of *in vitro* stimulating samples to observe the hosts' systemic responses and assess the memory generated by the infection. In the case of our model, a specific response could be observed from transcriptomics and cytokines concentration after stimulation of previously infected samples.

Transcriptomics and cytokines data analysis have proven to be useful techniques, providing a detailed characterization of this response at a molecular level. In particular, we observed the activation of biological pathways like antigen presentation, dendritic cells, cytokines and adaptive immune system. In addition, thanks to stimulation, an important difference was observed in the samples stimulated from day 4 after infection, with an important increase in the concentration of 21 from the 23 cytokines analyzed.

Finally, by performing data integration and feature selection, this work highlighted gene signatures and cytokines that could be used within the model to follow not only the pneumococcal infection process but possibly vaccination and antimicrobial therapies as well.

References

AlonsoDeVelasco, E., Verheul, A.F., Verhoef, J., Snippe, H., 1995. Streptococcus pneumoniae: virulence factors, pathogenesis, and vaccines. Microbiol. Rev. 59, 591–603. https://doi.org/10.1128/mr.59.4.591-603.1995

Berry, A.M., Paton, J.C., 1996. Sequence heterogeneity of PsaA, a 37-kilodalton putative adhesin essential for virulence of Streptococcus pneumoniae. Infect. Immun. 64, 5255–5262. https://doi.org/10.1128/iai.64.12.5255-5262.1996

Bogaert, D., de Groot, R., Hermans, P., 2004. Streptococcus pneumoniae colonisation: the key to pneumococcal disease. Lancet Infect. Dis. 4, 144–154. https://doi.org/10.1016/S1473-3099(04)00938-7

Briles, D.E., Paton, J.C., Mukerji, R., Swiatlo, E., Crain, M.J., 2019. Pneumococcal Vaccines. Microbiol. Spectr. 7. https://doi.org/10.1128/microbiolspec.GPP3-0028-2018

Brock, S.C., McGraw, P.A., Wright, P.F., Crowe, J.E., 2002. The Human Polymeric Immunoglobulin Receptor Facilitates Invasion of Epithelial Cells by *Streptococcus pneumoniae* in a Strain-Specific and Cell Type-Specific Manner. Infect. Immun. 70, 5091–5095. https://doi.org/10.1128/IAI.70.9.5091-5095.2002

Bronte, V., Pittet, M.J., 2013. The Spleen in Local and Systemic Regulation of Immunity. Immunity 39, 806–818. https://doi.org/10.1016/j.immuni.2013.10.010

Brown, J.S., Hammerschmidt, S., Orihuela, C.J., 2015. Streptococcus pneumoniae: molecular mechanisms of host-pathogen interactions. Elsevier/Academic Press, Amsterdam.

Coutinho, A., Möller, G., 1973. B Cell Mitogenic Properties of Thymus-independent Antigens. Nature. New Biol. 245, 12–14. https://doi.org/10.1038/newbio245012a0

Cundell, D.R., Gerard, N.P., Gerard, C., Idanpaan-Heikkila, I., Tuomanen, E.I., 1995. Streptococcus pneumoniae anchor to activated human cells by the receptor for platelet-activating factor. Nature 377, 435–438. https://doi.org/10.1038/377435a0

Deng, J.C., Standiford, T.J., 2005. The Systemic Response to Lung Infection. Clin. Chest Med. 26, 1–9. https://doi.org/10.1016/j.ccm.2004.10.009

Dockrell, D.H., Brown, J.S., 2015. Streptococcus pneumoniae Interactions with Macrophages and Mechanisms of Immune Evasion, in: Streptococcus Pneumoniae. Elsevier, pp. 401–422. https://doi.org/10.1016/B978-0-12-410530-0.00021-1

Ganaie, F., Saad, J.S., McGee, L., van Tonder, A.J., Bentley, S.D., Lo, S.W., Gladstone, R.A., Turner, P., Keenan, J.D., Breiman, R.F., Nahm, M.H., 2020. A New Pneumococcal Capsule Type, 10D, is the 100th Serotype and Has a Large *cps* Fragment from an Oral Streptococcus. mBio 11, e00937-20. https://doi.org/10.1128/mBio.00937-20

Goldblatt, D., Hussain, M., Andrews, N., Ashton, L., Virta, C., Melegaro, A., Pebody, R., George, R., Soininen, A., Edmunds, J., Gay, N., Kayhty, H., Miller, E., 2005. Antibody Responses to Nasopharyngeal Carriage of *Streptococcus pneumoniae* in Adults: A Longitudinal Household Study. J. Infect. Dis. 192, 387–393. https://doi.org/10.1086/431524

Hoe, E., Anderson, J., Nathanielsz, J., Toh, Z.Q., Marimla, R., Balloch, A., Licciardi, P.V., 2017. The contrasting roles of Th17 immunity in human health and disease: Th17 immunity in health and disease. Microbiol. Immunol. 61, 49–56. https://doi.org/10.1111/1348-0421.12471

Janoff, E.N., Rubins, J.B., Fasching, C., Charboneau, D., Rahkola, J.T., Plaut, A.G., Weiser, J.N., 2014. Pneumococcal IgA1 protease subverts specific protection by human IgA1. Mucosal Immunol. 7, 249–256. https://doi.org/10.1038/mi.2013.41

Kadioglu, A., Weiser, J.N., Paton, J.C., Andrew, P.W., 2008. The role of Streptococcus pneumoniae virulence factors in host respiratory colonization and disease. Nat. Rev. Microbiol. 6, 288–301. https://doi.org/10.1038/nrmicro1871

Kaplan, S.L., Barson, W.J., Lin, P.L., Romero, J.R., Bradley, J.S., Tan, T.Q., Hoffman, J.A., Givner, L.B., Mason, E.O., 2013. Early Trends for Invasive Pneumococcal Infections in Children After the Introduction of the 13-valent Pneumococcal Conjugate Vaccine. Pediatr. Infect. Dis. J. 32, 203–207. https://doi.org/10.1097/INF.0b013e318275614b

Khan, M.N., Sharma, S.K., Filkins, L.M., Pichichero, M.E., 2012. PcpA of Streptococcus pneumoniae mediates adherence to nasopharyngeal and lung epithelial cells and elicits functional antibodies in humans. Microbes Infect. 14, 1102–1110. https://doi.org/10.1016/j.micinf.2012.06.007

Koppe, U., Suttorp, N., Opitz, B., 2012. Recognition of Streptococcus pneumoniae by the innate immune system. Cell. Microbiol. 14, 460–466. https://doi.org/10.1111/j.1462-5822.2011.01746.x

Le Polain de Waroux, O., Flasche, S., Prieto-Merino, D., Edmunds, W.J., 2014. Age-Dependent Prevalence of Nasopharyngeal Carriage of Streptococcus pneumoniae before Conjugate Vaccine Introduction: A Prediction Model Based on a Meta-Analysis. PLoS ONE 9, e86136. https://doi.org/10.1371/journal.pone.0086136

Lebon, A., Verkaik, N.J., Labout, J.A.M., de Vogel, C.P., Hooijkaas, H., Verbrugh, H.A., van Wamel, W.J.B., Jaddoe, V.W.V., Hofman, A., Hermans, P.W.M., Ma, J., Mitchell, T.J., Moll, H.A., van Belkum, A., 2011. Natural Antibodies against Several Pneumococcal Virulence Proteins in Children during the Pre-Pneumococcal-Vaccine Era: the Generation R Study. Infect. Immun. 79, 1680–1687. https://doi.org/10.1128/IAI.01379-10

LeMessurier, K.S., Tiwary, M., Morin, N.P., Samarasinghe, A.E., 2020. Respiratory Barrier as a Safeguard and Regulator of Defense Against Influenza A Virus and Streptococcus pneumoniae. Front. Immunol. 11, 3. https://doi.org/10.3389/fimmu.2020.00003

Loughran, A.J., Orihuela, C.J., Tuomanen, E.I., 2019. *Streptococcus pneumoniae* : Invasion and Inflammation. Microbiol. Spectr. 7, 7.2.15. https://doi.org/10.1128/microbiolspec.GPP3-0004-2018

Malley, R., 2010. Antibody and cell-mediated immunity to Streptococcus pneumoniae: implications for vaccine development. J. Mol. Med. 88, 135–142. https://doi.org/10.1007/s00109-009-0579-4

Malley, R., Trzcinski, K., Srivastava, A., Thompson, C.M., Anderson, P.W., Lipsitch, M., 2005. $CD4^+$ T cells mediate antibody-independent acquired immunity to pneumococcal colonization. Proc. Natl. Acad. Sci. 102, 4848–4853. https://doi.org/10.1073/pnas.0501254102

Martin, F., Oliver, A.M., Kearney, J.F., 2001. Marginal Zone and B1 B Cells Unite in the Early Response against T-Independent Blood-Borne Particulate Antigens. Immunity 14, 617–629. https://doi.org/10.1016/S1074-7613(01)00129-7

Melegaro, A., Edmunds, W., Pebody, R., Miller, E., George, R., 2006. The current burden of pneumococcal disease in England and Wales. J. Infect. 52, 37–48. https://doi.org/10.1016/j.jinf.2005.02.008

Mitchell, A.M., Mitchell, T.J., 2010. Streptococcus pneumoniae: virulence factors and variation. Clin. Microbiol. Infect. 16, 411–418. https://doi.org/10.1111/j.1469-0691.2010.03183.x

Moberley, S., Holden, J., Tatham, D.P., Andrews, R.M., 2013. Vaccines for preventing pneumococcal infection in adults. Cochrane Database Syst. Rev. https://doi.org/10.1002/14651858.CD000422.pub3

Mureithi, M.W., Finn, A., Ota, M.O., Zhang, Q., Davenport, V., Mitchell, T.J., Williams, N.A., Adegbola, R.A., Heyderman, R.S., 2009. T Cell Memory Response to Pneumococcal Protein Antigens in an Area of High Pneumococcal Carriage and Disease. J. Infect. Dis. 200, 783–793. https://doi.org/10.1086/605023

Ogunniyi, A.D., LeMessurier, K.S., Graham, R.M.A., Watt, J.M., Briles, D.E., Stroeher, U.H., Paton, J.C., 2007. Contributions of Pneumolysin, Pneumococcal Surface Protein A (PspA), and PspC to Pathogenicity of *Streptococcus pneumoniae* D39 in a Mouse Model. Infect. Immun. 75, 1843–1851. https://doi.org/10.1128/IAI.01384-06

Olliver, M., Hiew, J., Mellroth, P., Henriques-Normark, B., Bergman, P., 2011. Human Monocytes Promote Th1 and Th17 Responses to Streptococcus pneumoniae. Infect. Immun. 79, 4210–4217. https://doi.org/10.1128/IAI.05286-11

Opitz, B., van Laak, V., Eitel, J., Suttorp, N., 2010. Innate immune recognition in infectious and

noninfectious diseases of the lung. Am. J. Respir. Crit. Care Med. 181, 1294–1309. https://doi.org/10.1164/rccm.200909-1427SO
Orihuela, C.J., Gao, G., Francis, K.P., Yu, J., Tuomanen, E.I., 2004. Tissue‐Specific Contributions of Pneumococcal Virulence Factors to Pathogenesis. J. Infect. Dis. 190, 1661–1669. https://doi.org/10.1086/424596
Papadatou, I., Tzovara, I., Licciardi, P., 2019. The Role of Serotype-Specific Immunological Memory in Pneumococcal Vaccination: Current Knowledge and Future Prospects. Vaccines 7, 13. https://doi.org/10.3390/vaccines7010013
Pichichero, M.E., 2017. Pneumococcal whole-cell and protein-based vaccines: changing the paradigm. Expert Rev. Vaccines 16, 1181–1190. https://doi.org/10.1080/14760584.2017.1393335
Pneumonia [WWW Document], n.d. URL https://www.who.int/news-room/fact-sheets/detail/pneumonia (accessed 5.6.22).
Pollard, A.J., Perrett, K.P., Beverley, P.C., 2009. Maintaining protection against invasive bacteria with protein–polysaccharide conjugate vaccines. Nat. Rev. Immunol. 9, 213–220. https://doi.org/10.1038/nri2494
Romagnani, S., 1999. Th1/Th2 Cells: Inflamm. Bowel Dis. 5, 285–294. https://doi.org/10.1097/00054725-199911000-00009
Rose, M.A., Schubert, R., Strnad, N., Zielen, S., 2005. Priming of Immunological Memory by Pneumococcal Conjugate Vaccine in Children Unresponsive to 23-Valent Polysaccharide Pneumococcal Vaccine. Clin. Vaccine Immunol. 12, 1216–1222. https://doi.org/10.1128/CDLI.12.10.1216-1222.2005
Tong, H.H., Blue, L.E., James, M.A., DeMaria, T.F., 2000. Evaluation of the Virulence of a *Streptococcus pneumoniae* Neuraminidase-Deficient Mutant in Nasopharyngeal Colonization and Development of Otitis Media in the Chinchilla Model. Infect. Immun. 68, 921–924. https://doi.org/10.1128/IAI.68.2.921-924.2000
Weinberger, D.M., Dagan, R., Givon‐Lavi, N., Regev‐Yochay, G., Malley, R., Lipsitch, M., 2008. Epidemiologic Evidence for Serotype‐Specific Acquired Immunity to Pneumococcal Carriage. J. Infect. Dis. 197, 1511–1518. https://doi.org/10.1086/587941
Weiser, J.N., Bae, D., Fasching, C., Scamurra, R.W., Ratner, A.J., Janoff, E.N., 2003. Antibody-enhanced pneumococcal adherence requires IgA1 protease. Proc. Natl. Acad. Sci. 100, 4215–4220. https://doi.org/10.1073/pnas.0637469100
Weiser, J.N., Ferreira, D.M., Paton, J.C., 2018. Streptococcus pneumoniae: transmission, colonization and invasion. Nat. Rev. Microbiol. 16, 355–367. https://doi.org/10.1038/s41579-018-0001-8
Zhang, J.-R., Mostov, K.E., Lamm, M.E., Nanno, M., Shimida, S., Ohwaki, M., Tuomanen, E., 2000. The Polymeric Immunoglobulin Receptor Translocates Pneumococci across Human Nasopharyngeal Epithelial Cells. Cell 102, 827–837. https://doi.org/10.1016/S0092-8674(00)00071-4
Zhang, Q., Bernatoniene, J., Bagrade, L., Paton, J.C., Mitchell, T.J., Hammerschmidt, S., Nunez, D.A., Finn, A., 2006. Regulation of Production of Mucosal Antibody to Pneumococcal Protein Antigens by T-Cell-Derived Gamma Interferon and Interleukin-10 in Children. Infect. Immun. 74, 4735–4743. https://doi.org/10.1128/IAI.00165-06

CHAPTER 4

Using Differential Expression analysis and Machine Learning algorithms to uncover molecular mechanisms of the rVSV-ZEBOV vaccine in three independent cohorts

4.1 Introduction

4.1.1 Ebola Virus and Ebola Virus Disease (EVD)

Since the first documented cases of disease caused by Ebola virus were described in 1976, the virus has re-emerged constantly in central Africa. Today, four different Ebola viruses are known to cause disease in humans: Sudan ebolavirus (SEBOV), Zaire ebolavirus (ZEBOV), Reston ebolavirus (REBOV) and Ivory Coast ebolavirus (ICEBOV), they were grouped in the ebolavirus genus, grouped under Filoviridae family (Feldmann et al., 2003). These viruses present a linear, negative sense, single-stranded RNA as genetic material, with approximately 19 kilobases of length (Bharat et al., 2012; Martin et al., 2017).

The Zaire ebolavirus is of particular interest since it is not only the most lethal but also responsible for many important epidemics, with fatality rates arriving up to 90% (Jadav et al., 2015). The Ebola Virus Disease (EVD) is characterized by unspecific symptoms including fever, headaches, muscle and joint pain, fatigue and gastrointestinal symptoms such as abdominal pain, diarrhea, and vomiting. However, the disease can rapidly evolve into a severe condition, causing external and internal bleeding, hypovolemic and septic shock and multiple organ failure (Chertow et al., 2014; Leligdowicz et al., 2016).

The Ebola Virus Disease (EVD) is considered a zoonotic disease and evidence that bats are carriers of EBOV and other filovirus have been increasing over time (Hayman et al., 2010; Koch et al., 2020; Leroy et al., 2005; Ogawa et al., 2015). An outbreak starts with the spillover from the animal reservoir to humans, with subsequent human-to-human transmission, especially through contact with body fluids such as blood, saliva, vomitus and stool (Dowell et al., 1999).

Due to this ability to spillover from animals and establish a human-to-human transmissible chain, EVD has the potential to cause thousands of deaths and serious sequelae in survivors, including persistent arthralgia, arthritis, and sight problems, besides the negative effects on the mental health of survivors and their relatives (Clark et al., 2015; Howlett et al., 2018; Tiffany et al., 2016). Therefore, vaccination is the best option to protect the population under increased risk and to contain new outbreaks (Kanapathipillai et al., 2014).

4.1.2 Vaccines against EVD

Given the situation of recurrent outbreaks in African countries, in 2014 there were different vaccine prototypes against EVD under development (Mohammadi, 2014), including subunit vaccines, non-replicant vectors, DNA vaccines and replication-competent vectors (Marzi and Feldmann, 2014). However, although some preclinical tests were carried out by then, it was only in response to the largest outbreak of EVD in 2014 that vaccine development was spurred towards clinical evaluation.

At that time, one of the vaccine candidates, rVSV-ZEBOV, had already demonstrated safety and efficacy in nonhuman primates (Jones et al., 2005). The rVSV-ZEBOV, today commercialized under the trade Ervebo, is a live, attenuated and replication-competent vaccine. This vaccine is based on a recombinant vesicular stomatitis virus (VSV), in which the glycoprotein from the VSV envelope was replaced by the Ebola Zaire's strain glycoprotein (Garbutt et al., 2004).

Huge scientific effort and financial support were conferred to accelerate the clinical development of this vaccine, which was demonstrated to be safe, immunogenic and protective in phase I, II and III clinical trials (Agnandji et al., 2016; Halperin et al., 2019; Henao-Restrepo et al., 2017; Heppner et al., 2017; Huttner et al., 2015; Regules et al., 2017). In this context, different consortiums were formed to further investigate the response to rVSV-ZEBOV vaccine. Among them, VSV-EBOVAC and VSV-EBOPLUS consortiums were established with the objective of identifying molecular mechanisms associated with the

vaccination, using, among other methods, transcriptomic analysis (Laverty and Meulien, 2019; Medaglini et al., 2015).

4.1.3 Machine Learning and Vaccinology in the Big-data era

As discussed in the initial chapter of this thesis, transcriptomic analysis and the Systems Vaccinology approach have enabled the characterization of the immune responses to different vaccines. Differential expression analysis, enrichment, and other downstream analyses are extremely important in this context. However, this type of data can be explored using other approaches, such as Machine Learning (ML).

Machine Learning (ML) is a branch of Artificial Intelligence that has been gaining importance in different areas, including biomedical research. ML can be defined as the ability to improve machine performance in specific tasks by identifying similarities in the data, and using this to infer information on a different data set. ML algorithms are capable of performing tasks and finding patterns in big data, which is considered impossible for human beings (Bench-Capon and Dunne, 2007).

ML has been allowing earlier and more accurate diagnosis, supporting the discovery of new subtypes of cancer, and making it possible to predict better treatments and outcomes (Goecks et al., 2020). With the rapid progress in this field and the exponential generation of data, personalized medicine is becoming more and more a part of our reality.

This represents a new era for the vaccinology field as well. The possibility of understanding vaccine immune responses, finding correlates of protection and uncovering mechanisms underlying immunogenicity and reactogenicity made scientists invest in this new approach. However, dealing with the amount of data produced is still a huge challenge, especially for biologists. In this context, tools that could help retrieve relevant information are of extreme importance.

4.1.4 The aim of this chapter

In this chapter, the transcriptomic responses to the rVSV-ZEBOV in different cohorts were compared by two distinct methods: the usual Differential Expression Analysis and a Machine Learning approach. The main objective was to understand the similarities and differences among the cohorts and compare the application of each methodology in two different scenarios: (i) high transcriptomics perturbation, represented by day 1 after vaccination, and (ii) low transcriptomics perturbation, represented by day 7 after vaccination.

Therefore, the first part of the Results section is dedicated to the comparison of the transcriptomic responses in the Swiss and North American cohorts, using a Differential Expression approach. The third part concerns a work developed in collaboration with the Computational Systems Biology Laboratory, within the VSV-EBOPLUS consortium. The selection of genes by Differential Expression analysis was compared with the Biological Feature Selection tool (BioFeatS), using the two cohorts cited above, and a third cohort conducted in Germany.

ISABELLE FRANCO MOSCARDINI
16/09/2023 19:17:47
Isabelle Franco moscardini

4.2 Methods

4.2.1 The cohorts

The details of the three cohorts used in this chapter are described in the results section of the material provided in the item 4.3.4 of this chapter.

4.2.2 Library Preparation, and Sequencing

Library preparation was performed with the Ion AmpliSeq™ Transcriptome Human Gene Expression Kit (Thermo Fisher Scientific), under the same conditions as the Swiss cohort, described in a previous publication (Santoro et al., 2021). Sequencing was performed with the Ion Proton Technology (Thermo Fisher Scientific). All the steps were performed following the manufacturer's instructions.

4.2.3 Data analysis

All the steps in data analysis were performed in R software (version 3.6.3). The edgeR package (Robinson et al., 2010) was used for differential expression analysis and genes with an adjusted p-value < 0.05 were classified as differentially expressed. Enrichment analyses were performed using the Blood Transcription Modules (BTM) database (Li et al., 2014), significance was assessed by the CERNO test or the hypergeometric test, both from tmod package (Weiner 3rd and Domaszewska, 2016). Venn Diagrams for the comparison of differentially expressed genes between the cohorts were built using the DiVenn webtool (Sun et al., 2019).

4.3 Results

4.3.1 Differences in the transcriptomic profile after vaccination in the two cohorts

The transcriptomic response to the ebola vaccine is already known to be marked by a robust perturbation in a gene expression level, stronger on the first day after vaccination and with a decreasing perturbation over time (Santoro et al., 2021), a pattern observed in both Swiss and North American cohorts. Differential expression analysis showed similar responses on days 1 and 2 after vaccination, with most of the differentially expressed genes (DEGs) being shared between the cohorts and a similar number of specific DEGs for each study (Figure 1).

However, on days 3 and 7 after vaccination, there are differences between the responses in each cohort. Day 3 is marked by specific genes up-regulated only in the Swiss cohort, while the genes specific for the North American cohort are mainly downregulated. On day 7, the North American cohort presented significantly less differentially expressed genes compared with the Swiss cohort. To understand the biological function of these DEGs, hypergeometric tests were performed by the tmod package, using the blood transcription module database.

The genes specifically activated in the Swiss cohort are related to an up-regulation of inflammatory responses, especially those linked to antiviral innate immunity and interferon, showing an extended innate response in comparison to the North American cohort. The latter, on the other hand, presented specific DEGs enriched to down-regulation of inflammatory processes linked to neutrophils and upregulation of CD4+ T cells responses, as shown in the plots in Figure 1.

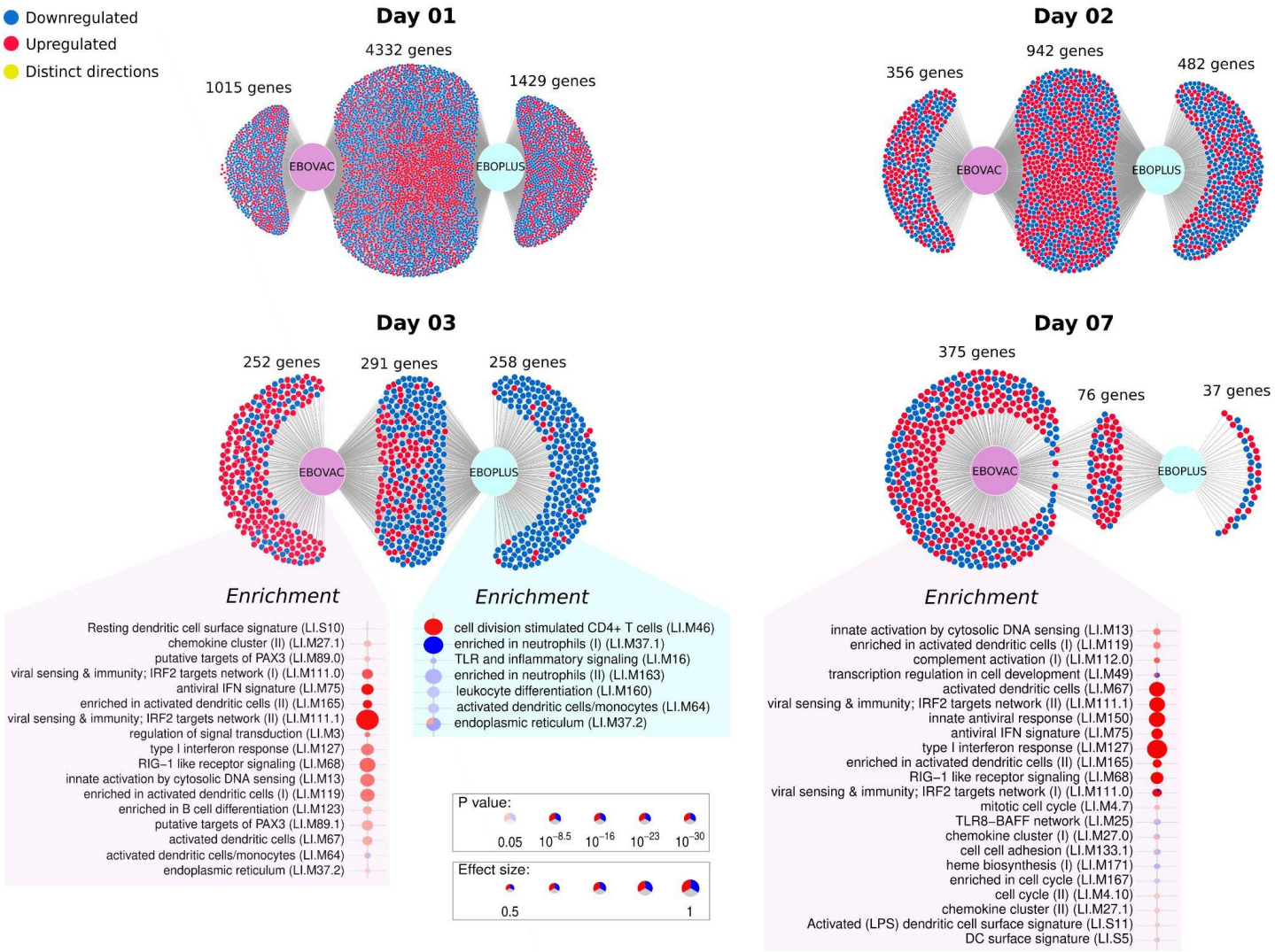


Figure 1. Differentially Expressed Genes (DEGs) for each cohort and time point. Diagrams representing the DEGs that are specific or shared between the two cohorts at each time point. Blue circles represent downregulated genes while red represent upregulated ones. Yellow circles represent genes that present a different direction in each of the cohorts. Gene Set analyses were performed to address the biological function of the genes unique to each cohort on days 3 and 7.

ISABELLE FRANCO MOSCARDINI

16/09/2023 19:22:13

4.3.2 Pathway analysis: apparent prolonged interferon response in the Swiss cohort and earlier T cell response in the North American cohort

Isabelle Franco moscardini

The Swiss and North American cohorts presented a very consistent profile of activated modules in the enrichment analysis performed with the CERNO test, using the genes ranked by their adjusted p-value. The supplementary figure 1 represents the most significant activated modules for each time point and cohort (enrichment p value < 0.01).

Here, we divided these enrichment results into innate and adaptive immune responses, highlighting the main differences between the two studies.

Innate responses

The strong antiviral response mediated through Interferon was already described for the rVSV-ZEBOV vaccine. In both North America and Swiss cohorts, there is an important up-regulation of interferon and inflammatory cytokines genes (Figure 2). On days 3 and 7, these innate antiviral and inflammatory responses are still substantial in the Swiss cohort, but less significant in the North American cohort.

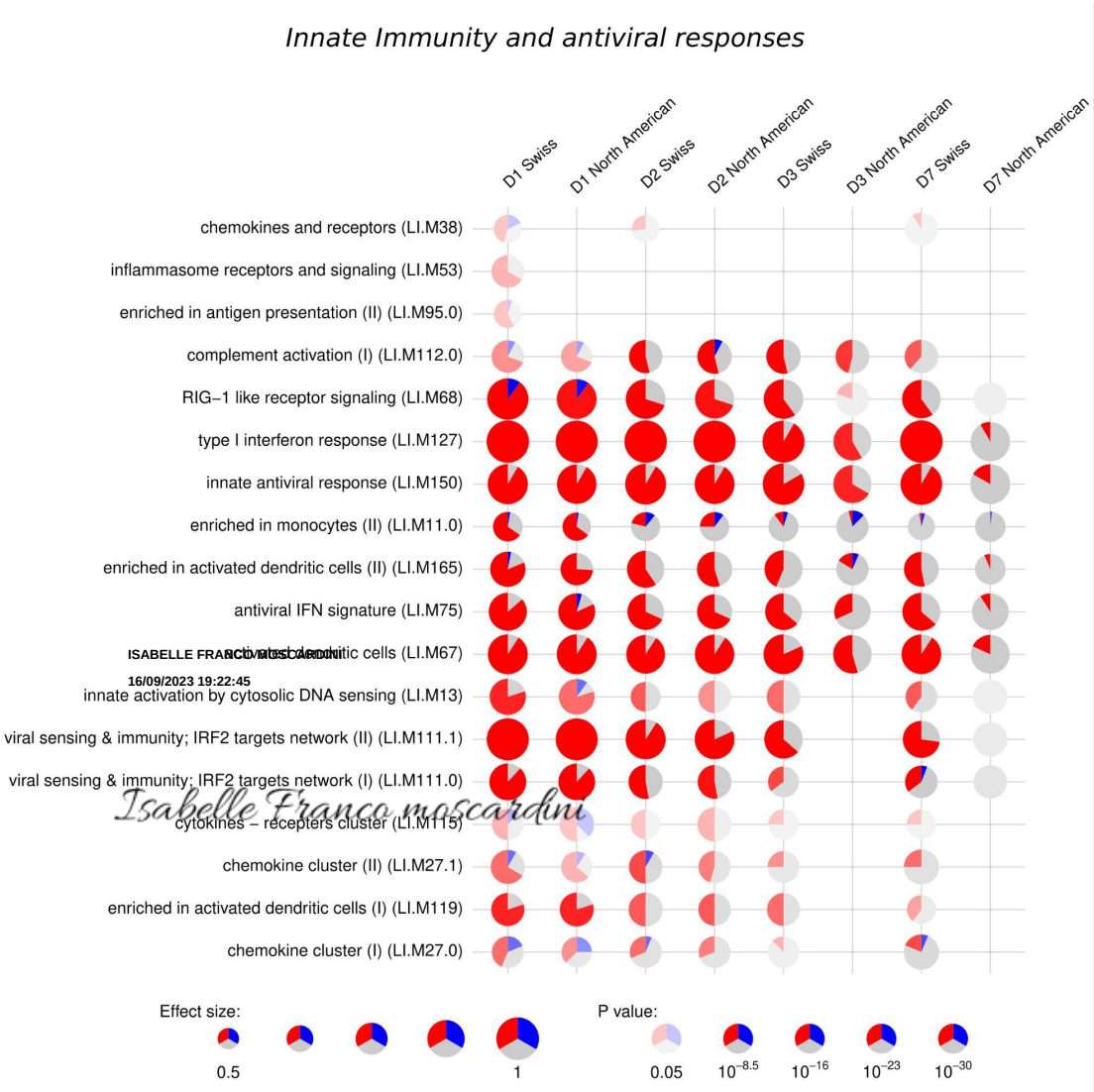


Figure 2. Modules of the innate immunity and antiviral response. Enrichment analysis highlighting pathways with a stronger and/or prolonged activation in the Swiss cohort, compared to the North American cohort. This pattern was observed in different modules related to innate immunity, antiviral response, activated dendritic cells and cytokines/chemokines.

Modules related to chemokines and viral sensing are not enriched anymore from day 3 in the North American cohort. In contrast, modules linked to interferon, innate antiviral responses, and dendritic cells have a much less significant activation on days 3 and 7, compared to the Swiss cohort. This fact is also explained by the discrepancy in the number of differentially expressed genes.

ISABELLE FRANCO MOSCARDINI
15/08/2022 18:33:17

Isabelle Franco moscardini

Adaptive responses

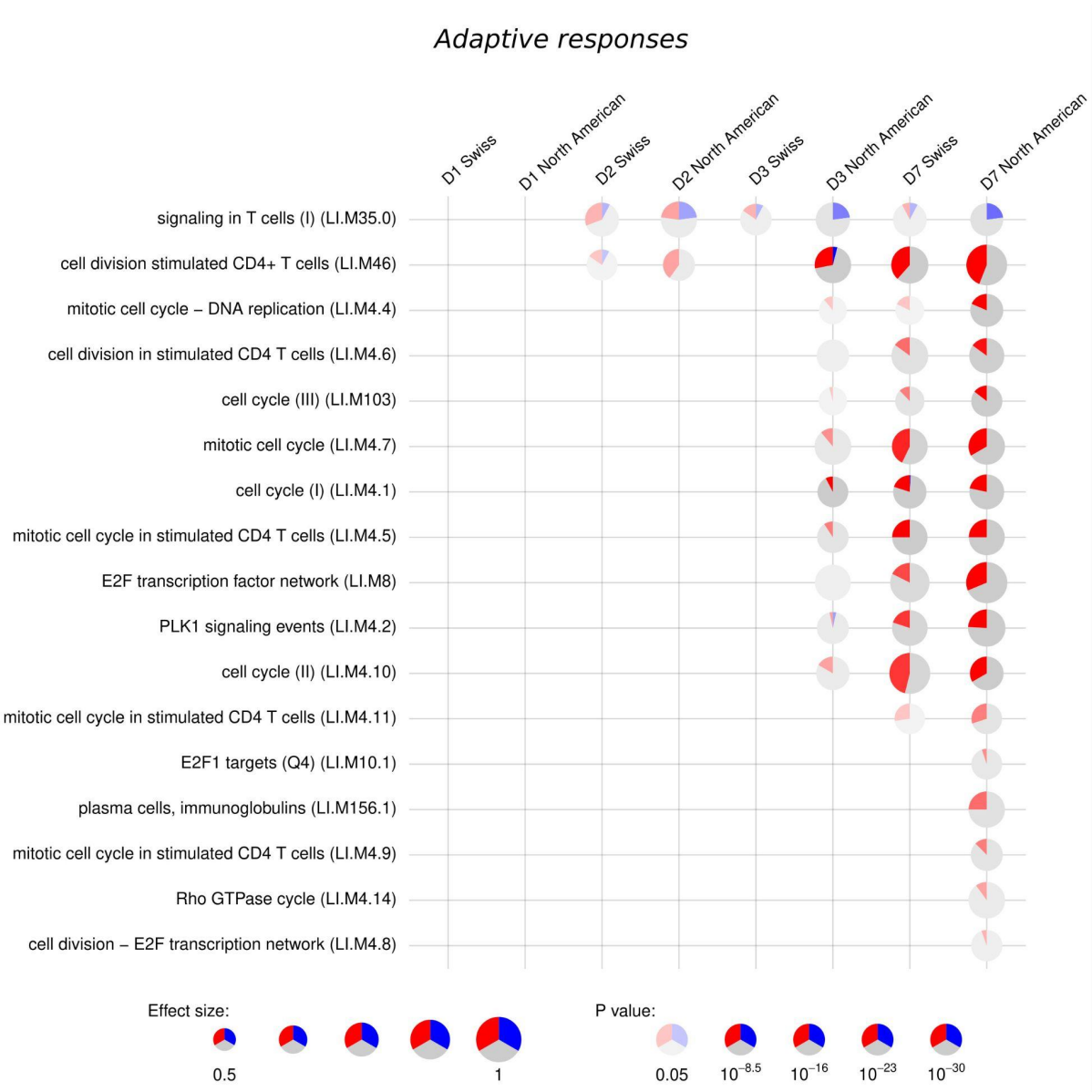


Figure 3. Modules of the adaptive immune system. Differences in the adaptive responses were also seen in the enrichment analysis. The North American cohort presents an early activation of T Cell and cell cycle modules, already at day 3 post vaccination. At day 7, the North American cohort also presents unique activation of modules linked to adaptive responses.

While the Swiss cohort presents an apparent longer innate antiviral response, the North American cohort shows an apparent earlier T cell activation, as modules related to T CD4 cells and cell cycle are enriched from day 3 (Figure 3). Moreover, on day 7, the module "plasma cells, immunoglobulins" is enriched only in the North American cohort, together with another T CD4 module.

4.3.4 Prioritizing the importance of biological components within High Throughput data: a machine learning approach

Prioritizing the importance of biological components within High Throughput data: a machine learning approach

Isabelle Franco Moscardini[1†*], Leandro Yukio Mano Alves[2†], Patrícia Conceição Gonzalez Dias Carvalho (Patricia Gonzalez-Dias)[5,6], Thiago Dominguez Crespo Hirata[2], Thomaz Lüscher Dias[2], Ana Paula Barbosa do Nascimento[2,], Alice Gerlini[1], Francesco Santoro[7], Donata Medaglini[7], Gianni Pozzi[7], VSV-EBOVAC and VSV-EBOPLUS Consortia, Helder Takashi Imoto Nakaya (Helder I Nakaya)[3,4*]

Affiliations:

[1] Microbiotec srl, Siena, Italy

[2] Department of Clinical and Toxicological Analyses, School of Pharmaceutical Sciences, University of São Paulo, São Paulo, Brazil;

[3] Scientific Platform Pasteur-University of São Paulo, São Paulo, Brazil;

[4] Hospital Israelita Albert Einstein, São Paulo, Brazil.

[5] Oxford Vaccine Group, University of Oxford

[6] Department of Clinical Sciences, Liverpool School of Tropical Medicine, Liverpool, United Kingdom.

[7] Laboratory of Molecular Microbiology and Biotechnology (LAMMB), Department of Medical Biotechnologies, University of Siena, Italy.

†These authors have contributed equally to this work.

*To whom correspondence should be addressed

ABSTRACT

The use of High-Throughput technologies to characterize immune responses to infections and vaccination has been rising fast, producing an unprecedented amount of data. As new computational methods emerge and demand for personalized medicine increases, machine learning algorithms gain space in biomedical research. In this study we share a Machine Learning approach to deal with massive datasets, aiming for a robust framework that can be applied to different types of omic data. The transcriptomic profile after vaccination with rVSV-ZEBOV was evaluated in three independent cohorts using two different approaches. The traditional Differential Expression Analysis of the RNA-sequencing data workflow was compared to a new workframe based on feature selection and machine learning algorithms, denominated as the Biological Feature Selection tool (BioFeatS).

When transcriptomic perturbations are high, BioFeatS can select a summarized list of genes that allows one to study the biological processes focusing on a gene level analysis, prioritizing features that better distinguish the class of the samples. When the perturbation is less protuberant, BioFeatS is able to retrieve genes not found via Differential Expression analysis, and bring more consistency in the Gene Set analysis. Genes specifically selected by BioFeatS have also brought insights on the biological processes taking place 7 days after vaccination.

The process of selecting genes by Feature Selection and Machine Learning algorithms has shown to diverge from Differential Expression analysis. Therefore, to bring meaningful biological insights and extract a more robust biological signature, these methodologies could be applied in a complementary way.

Keywords: Machine Learning, Feature selection, Ebola Vaccine, Omic data analysis

INTRODUCTION

The Zaire ebolavirus is a filovirus known to cause frequent outbreaks of Ebola Virus Disease (EVD) in west and equatorial Africa (Malvy et al., 2019). EVD is characterized by high mortality rates, as a result of an intense systemic inflammatory response that can lead to multiple organ impairment (Jacob et al., 2020). In view of the largest known outbreak in west Africa in 2014, efforts were combined aiming at the clinical development of an effective vaccine against the disease. A recombinant live-attenuated vaccine candidate, based on the vesicular stomatitis virus expressing the glycoprotein of the Zaire ebolavirus, had proved safe and highly efficacious in nonhuman primates challenged with ebola virus (Jones et al., 2005).

Different clinical trials were carried out to assess safety and efficacy of this vaccine in humans. Particularly, the cohort conducted in Geneva, Switzerland, provided data on gene expression, proteins, metabolites, cytokines, and microRNAs. Two other cohorts, one conducted in the United States and another in Germany have collected gene expression data as well (Rechtien et al., 2017; Santoro et al., 2021a).

High-throughput technologies provide the possibility of evaluating the expression of thousands of different features in different biological layers, revolutionizing the way we study organisms and diseases. Simultaneously, computational techniques have emerged to handle these new data and, despite the increasing availability of methods, scientists still cope with extracting meaningful biological information from these experiments.

With the challenge of analyzing and integrating different biological layers, arises the need for a structured framework that could deal with the massive amount of generated data. The use of Machine Learning algorithms in biomedical sciences is a way to deal with this challenge and it has been driving advances in different areas, from diagnosis to treatment adherence, contributing to the evolution of precision medicine (Goecks et al., 2020).

OMIC technologies provide us with complex datasets, commonly presenting hundreds or thousands of features, which usually greatly outnumber the sample size. The high-dimensional datasets can affect the performance of the Machine Learning algorithms in

different ways. For instance, multiple correlated features can deceive the algorithm training step since they bring redundant information in the model (James et al., 2013). Also, the presence of irrelevant features can result in overfitting of the model in the training process, impacting the model performance in unseen datasets (Gonzalez-Dias et al., 2020).

With the purpose of overcoming these potential issues, we created the Biological Feature Selection tool (BioFeatS). BioFeatS can deal with large data sets thanks to its robust approach, which includes using three different methodologies for feature selection, creating a list of consistent features among these three methods, ordering the consistent list using Random Forest algorithm and then evaluating the ability of the final list of features in discriminating different classes by using a combination of different Machine Learning algorithms.

Hence, the purpose of this work is to retrieve meaningful biological information on the effect of Ervebo® vaccine from a systems biology perspective and to introduce BioFeatS, the framework we employed to address the challenges of working with massive datasets.

Isabelle Franco moscardini

METHODS

Study Design

This work relied on data from three different cohorts conducted in different countries. The detailed study design for the German and Swiss cohorts were already described in previous publications by Agnandji et al (Agnandji et al., 2016) and Santoro et al (Santoro et al., 2021a), respectively. The North American cohort was a double-blind, randomized, and controlled trial conducted to test the immunogenicity and safety of rVSVΔG-ZEBOV-GP vaccine. All participants were healthy adults aged from 18 to 65 years and received a dose of 2×10^7 pfu.

RNA sequencing data collection and preprocess

The collection, sequencing and preprocess of data from the German cohort was described in a previous work by Rechtien et al (Rechtien et al., 2017). For the North American cohort, the RNA extraction, library preparation and RNA-sequencing were performed following the same protocols of the Swiss cohort, described by Santoro et al (Santoro et al., 2021a).

Differential Expression Analysis and Gene Set analysis

Low expressed genes were filtered out in all the three studies by maintaining only genes that had more than 1 count per million in at least ten samples. Differential expression (DE) analysis was conducted in the R software environment, using the edgeR package (Robinson et al., 2010) and genes with a False Discovery Rate (FDR) of less than 0.05 were considered differentially expressed. Gene set analysis was performed using the tmod package and the blood transcription modules (BTM) (Li et al., 2014b; Weiner 3rd and Domaszewska, 2016). The hypergeometric function was used to assess the significance of the enriched modules.

BioFeatS Algorithm

The BioFeatS identifies and ordered list of features that best distinguishes the outcome of 2 or more groups. The model selects the features based on the combination of the following feature selection methods: Pearson's correlation, Kbest and Recursive Feature Elimination (RFE). After generating a list of features for each method, a unified list is produced by selecting features from the intersection of 2 out the 3 methods. From this list, the BioFeatS generates the ranking importance obtained from the Random Forest algorithm and then removes the features with an importance value equal to 0. BioFeatS' algorithm subsets the dataset with the selected features and trains 4 different machine learning (K-fold = 10) algorithms: Support Vector Machine (SVM), k-Nearest Neighbors (kNN), Naive Bayes and AdaBoost Classifier. Thereafter, the tool generates as output a table with the values of F1-score, area under the curve (AUC), accuracy, and precision, obtained from each model with the selected features. Also, provided in the output file are the median and harmonic mean calculated from all methods and metrics. Additionally, BioFetS can also identify outlier samples based on the number of times the samples are wrongly classified by the algorithm after being tested 4 times with a k-fold of 10. A detailed version on the BioFeatS methodology is provided in the supplementary materials (Supplementary file 1)

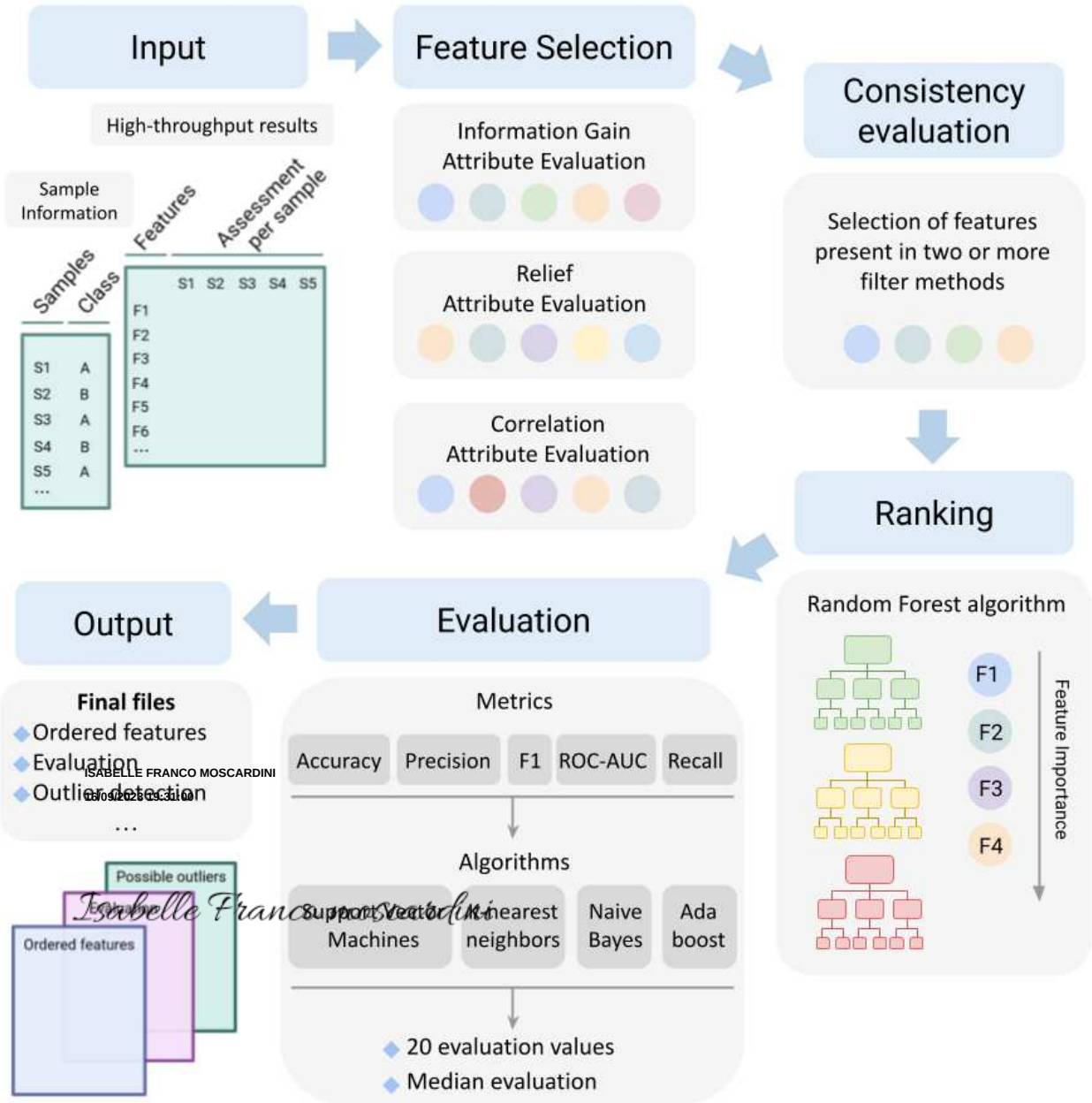


Figure 1. BioFeatS general workflow. The input data consists of two tables, one being the result of a high-throughput experiment, with the variables measured in the first column and the other columns representing these values for each sample. The second table should inform the class of each sample. Data is filtered by three different methods of feature selection, each method providing a distinct list of features. Features that are present in two or three of these lists are selected and ranked by their importance through Random Forest algorithm. Features with importance zero in the Random Forest model are removed. Performance evaluation is assessed by four Machine Learning algorithms, using four different metrics. The final output includes: (i) a list of the selected features ordered by their importance, (ii) the values of the metrics for each method and the median and harmonic mean of these values, and (iii) a table indicating samples that are consistently misclassified.

RESULTS

The three transcriptomics cohorts

Three different cohorts were used to test BioFeatS' performance in identifying biological signatures for the response to Ervebo® vaccine (Figure 2). The Swiss and the German cohorts already have publications associated with their transcriptomic data (Rechtien et al., 2017; Santoro et al., 2021a). All of the cohorts have collected samples for transcriptomics analysis before vaccination (Day 0) and 1, 3 and 7 days after vaccination.

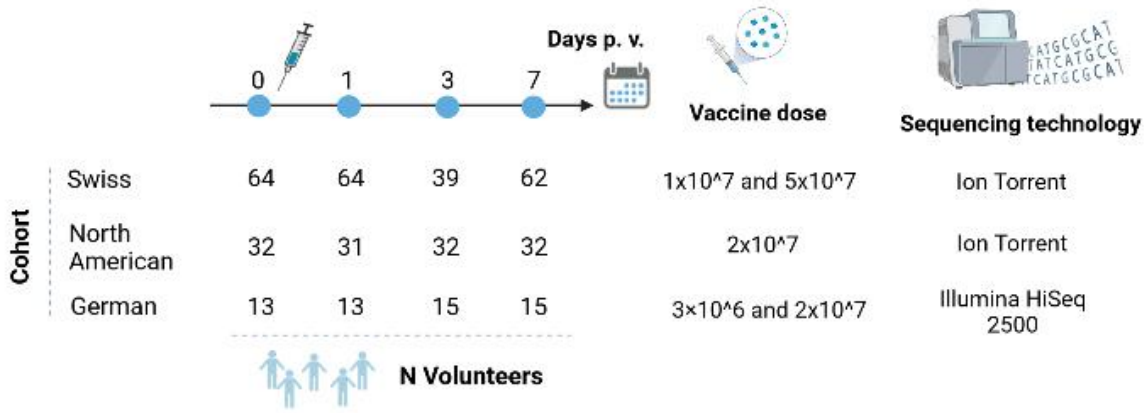


Figure 2. The sample sizes and cohorts' characteristics. Three independent cohorts studying the response to rVSV-ZEBOV vaccine were compared in this study Four time points were in common among the cohorts: day 0 (before vaccination) and days 1, 3 and 7 after vaccination. The figure displays the number of volunteers at each time point for each cohort. Besides the number of volunteers, vaccine dose and sequencing technology also varied among the cohorts.

BioFeatS summarizes the transcriptomic response to Ervebo® vaccine in different cohorts

Differential expression (DE) analysis has been the standard approach for assessing information provided by transcriptomics essays. However, in some cases the perturbations are very strong, like the responses one day after Ervebo® vaccination, leading to a very high number of differentially expressed genes (DEGs). In this context, it is a challenge to spot the main features linked to this response. While DE analysis led to the selection of a thousands of features, especially on day 1 after vaccination, BioFeatS was able to detect dozens of

features, summarizing the perturbation of the three cohorts in 6 features: HERC6, AGRN, IFI35, KIAA1958, IFIT1, and OAS3.

It is important to highlight that, even when selecting genes that are also differentially expressed (DEGs), the selected by BioFeatS are not necessarily the most statistically significant DEGs. For instance, the third most important gene in BioFeatS selection for the North American cohort would be in the 340 position in the list of DEGs ranked by their score (defined by -log10 of the adjusted p value multiplied by the log2 of the Fold-Change). This fact highlights the distinction in the prioritization process employed by DE analysis and by feature selection coupled with ML algorithms, which can lead to distinct gene lists.

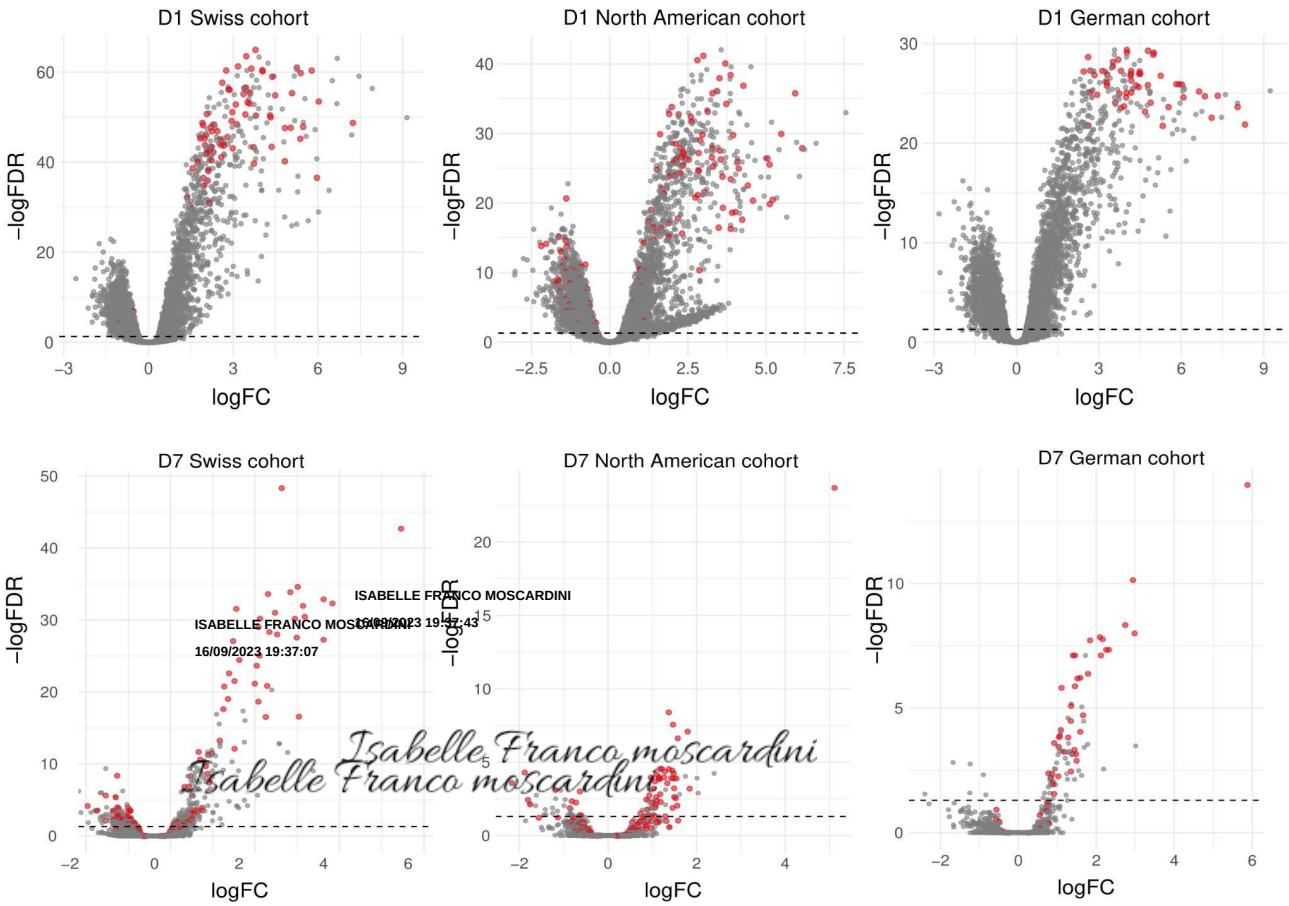


Figure 3. Genes selected by BioFeatS. Volcano plot for each cohort at days 1 and 7 post vaccination, genes selected by BioFeatS analysis are highlighted in red. The dashed line represents the significance threshold of adjusted p value < 0.05.

BioFeatS selected a smaller number of genes at day 1 after vaccination, which were not necessarily the ones with the highest values of log2 Fold-Change or adjusted - logp value in the DE analysis. However, these genes have a high contribution to the classification of the groups, in this case before and one day after vaccination (Figure 3). On the other hand, day 7 after vaccination presents a much less protuberant response in the transcriptomic level. In this case, BioFeatS was able to detect features that were not considered Differentially Expressed, but again, are considered important for the distinction of the groups.

The features specifically selected by BioFeatS tool at day 7 after vaccination

At day 7 post vaccination with Ervebo®, among the genes identified in BioFeatS, the German cohort presented the highest similarity to DE analysis (40/57, 70.2%), followed by Swiss (103/172, 59.9%) and North American (56/199, 28.1%) cohorts.

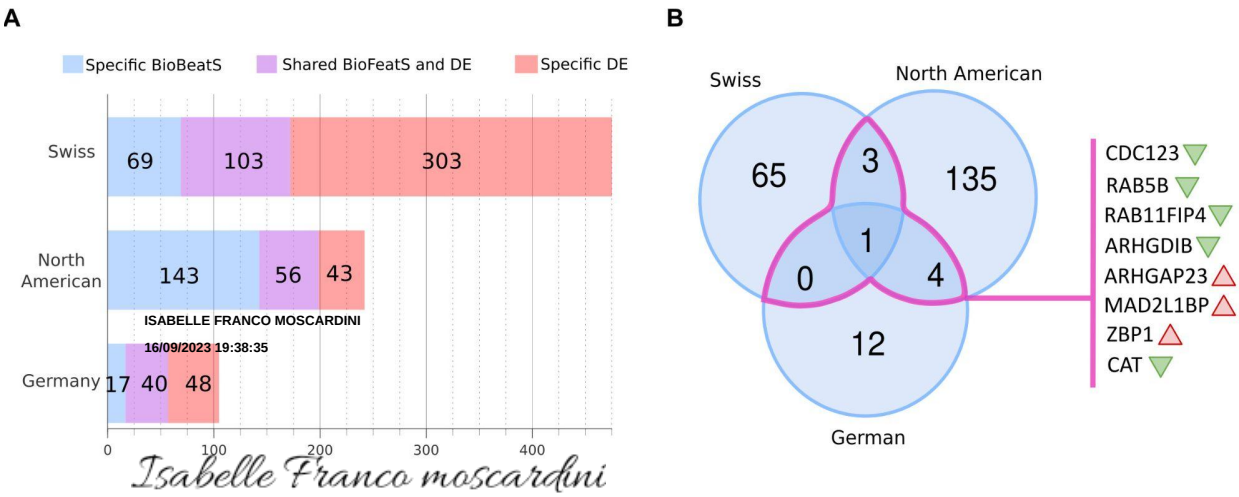


Figure 4. Genes selected by BioFeatS and by Differential Expression Analysis at day 7 after vaccination. **A.** Number of genes selected specifically by BioFeatS (blue) or DE analysis (red) and the intersection between them (purple). **B.** Venn Diagram of the specific genes selected by BioFeatS in at least two of the three cohorts and their direction compared to day 0 (before vaccination) - green triangles identify genes less expressed, while red triangles represent genes with higher expression at day 7.

While some of the identified genes are shared between the methods, all the cohorts presented genes selected only via BioFeatS. In the Swiss cohort, 69 genes were uniquely

identified by BioFeatS, for the North American and German cohorts, these numbers were 143 and 17 respectively. When selecting the specific common features in at least two of the cohorts, a signature of 8 genes is revealed. The features ARHGAP23, MAD2L1BP and ZBP1 are differentially expressed at day 1 in all of the three datasets, indicating that the difference captured by BioFeatS at day 7 is probably a remnant response. The gene CAT was down-regulated on day 3 in two of the three cohorts. The other features included CDC123, RAB5B, RAB11FIP4 and ARHGDIB which were not identified as differentially expressed at day 1 or 3 after vaccination.

The gene ZBP1 encodes a protein that plays a role in the innate immune response by binding to foreign DNA and inducing type-I interferon production, having an important role in different viral infections such as Influenza (Kuriakose et al., 2018), Hepatitis B (Farci et al., 2010), COVID-19 (Sims et al., 2013) and cytomegalovirus (Vermijlen et al., 2010). The genes RAB11FIP4 and RAB5B are involved in endosome membrane recycling and plasma membrane to endosome transport, respectively. RAB5B is involved in antigen processing and presentation**.** MAD2L1BP participates in the coordination of cell cycle events, and it is linked to the E2F transcription factor network. This gene was also up-regulated in naïve primary human B-lymphocytes infected with Epstein-Barr virus (Mrozek-Gorska et al., 2019), and proteins from this virus can interact with the protein encoded by MAD2L1BP (Li et al., 2015).

ARHGAP23 and ARHGDIB are members of the Rho family of proteins. Interestingly, ARHGAP23 was identified as a required host factor for the entry of VSV in the cells (Panda et al., 2011), it is upregulated at day 01 in the DE analysis and over expressed at day 7 by BioFeatS. CAT gene encodes a Catalase, which is involved in the response to oxidative stress. BioFeatS identified this gene as less expressed 7 days after vaccination. In fact, viral infections can inhibit the activity of catalase, such as the Newcastle disease virus in chickens (Subbaiah et al., 2011), and influenza (Celestino et al., 2018; Yamada et al., 2012) and herpes simplex virus type 2 (Sartori et al., 2012) in mice.

ISABELLE FRANCO MOSCARDINI
16/09/2023 19:39:38
Isabelle Franco moscardini

Finally, CDC123 was not selected in the DE analysis in any of the three cohorts, but was selected by BioFeatS in all of them. This gene is a chaperone needed to the assembly of the eukaryotic translation initiation factor 2 complex (eIF2), positively regulating the translation initiation. Since the phosphorylation of eIF2 leads to inhibition of the host's translation, this complex has a role described in different viral infections, including vesicular stomatitis virus (Connor and Lyles, 2005) and Ebola virus (Strong et al., 2008). CDC123 was also downregulated in dendritic cells infected with Newcastle Disease Virus (Zaslavsky et al., 2010) and human bronchial epithelial cells infected with human metapneumovirus (Bao et al., 2008).

Features selected with BioFeatS tool on day 7 after vaccination reveals a consistent response among the cohorts

When the transcriptomic responses are less expressive, DE analysis might not be able to identify a robust signature. To understand if BioFeatS could contribute to the characterization of the immune responses to the rVSV-ZEBOV vaccine, we compared the gene set analysis of features selected by BioFeatS with those selected by DE analysis, using the blood transcription modules database (BTM). Interestingly, enrichment analysis of features selected by BioFeatS (14/30, 46.6%) showed a higher consistency among the cohort then the DE analysis (17/48, 35.4%).

BioFeatS has identified modules in the North American cohort that were also present in the other studies but were not identified in DE analysis, such as the "*viral sensing & immunity"; "IRF2 targets network - (I) and (II)"*, "*type I interferon response"*, "*antiviral IFN signature"*, and *"enriched in activated dendritic cells (II)"*. Moreover, the module "*plasma cells, immunoglobulins"* was identified by BioFeatS in the Swiss and North American cohorts but only in the North American cohort in the DE analysis. On the other hand, modules related to cell cycle, that were consistent among cohorts, were not catched in features selected by BioFeatS in the Swiss cohort. Importantly, modules enriched by genes

selected by BioFeatS reflect biological signatures capable of best distinguishing two different groups.

Figure 5. BTM Modules from BioFeatS and Differential Expression Analysis at day 7 after vaccination. **A.** Venn diagram with the number of pathways activated by DE analysis (left) and BioFeatS (right), from each cohort: USA (blue), Switzerland (red), Germany (green). **B.** Pathways enriched by BioFeatS and DE Analysis in each cohort on day 7.

DISCUSSION

In the pathway-level analysis that have guided transcriptomics and other OMICs analysis from the beginning, a higher number of features is usually seen in a positive light. This concept has been changing in the past years with the advances in precision medicine, and the dissemination of machine learning and data integration methods, which focus more in a gene-level perspective.

Here we have shown that Differential Expression Analysis and Machine Learning algorithms select genes in a distinct manner, depending on the level of transcriptomic perturbation in the data set. In our example, day 7 after vaccination with rVSV-ZEBOV vaccine presented a lower perturbation compared to day 1, but both situations could benefit from the BioFeatS framework. At day 1 BioFeatS was able to highlight the features, among thousands of Differentially Expressed Genes, that best classify samples. At day 7, BioFeatS was able to retrieve different genes from DE analysis and a more consistent activation of Blood Transcriptional Modules.

Since one day after vaccination the transcriptomic perturbation is very high, the signature identified by BioFeatS is contained in the signature identified by the differentially expressed genes, which varied between 4688 and 5361 genes, depending on the cohort. The 6 genes consistently identified by BioFeatS summarizes the context of a viral infection. HERC6 for example is part of the antiviral response to VSV and Ebola infections, and it is correlated with protection against Marburg virus in a postexposure (rVSV)-MARV vaccine study. OAS3, IFI35 and IFIT1 are Interferon induced proteins and involved in the immune responses against viruses. AGRN and KIAA1958, despite not having a clear role described in viral infections, are up-regulated in B cells infected with Epstein-Barr virus (Mrozek-Gorska

et al., 2019). Besides their biological role, these features were selected for having the best classification attributes in distinguishing the two groups.

The identification of genes related to translation, such as CDC123, might indicate biological processes taking place within infected cells. In fact, viruses have the ability to selectively inhibit host translation, as well as infected cells might shut down protein synthesis as an antiviral strategy (Gale et al., 2000; Schneider and Mohr, 2003). Particularly, the Vesicular Stomatitis Virus (VSV) is known to inhibit host gene expression at multiple levels, including transcription, nuclear cytoplasmic transport, and translation, in order to suppress antiviral responses in infected cells (Ahmed and Lyles, 1998; Rajani et al., 2012). Among the specific genes selected by BioFeatS, CDC123 is linked to translation, and in the Swiss cohort, for instance, BioFeatS specifically spotted differences in the expression of many genes linked to translation and transcription, which were less expressed at day 7 post vaccination, including ZNF33A, RYBP, POLR2C, PABPC1 and EIF3L.

Collectively, our results suggest that BioFeatS was able to specifically highlight cellular processes that may have resulted from vaccination with the recombinant VSV, such as the perturbations in transcription, endosome transport and translation, besides the up-regulation of host's factors that are important for the entry of VSV in the cells, process that seem to be important for the classification on samples at day 7 post vaccination.

Despite studying the same vaccine and collecting data at the same time points after vaccination, these cohorts differ in many aspects. Starting by the location of the clinical trial and the number of volunteers. The Swiss cohort, for instance, presents 2 times the number of volunteers of the North American cohort and 4 times the number of the German study. The vaccine dose is another important factor that differed not only among cohorts but also inside the Swiss and German cohorts. Finally, different sequencing technologies were used to assess changes in gene expression. These differences can affect both BioFeatS and DE analysis, leading to distinct results, and therefore being a limitation in integrating different clinical trials. Even with these relevant differences between the cohorts, Biofeats identified biological

pathways in a very consistent way. For instance, BioFeatS was able to identify signatures of antiviral responses and dendritic cell activation in all three cohorts, while DE analysis did not find these signatures in the North American cohort.

Moreover, BioFeatS was built with the purpose to be a flexible tool for different data types. Since the same dataset analyzed by different methodologies can select distinct lists of features, the possibility of using a single tool, robust enough to be applied to different data, can facilitate the process of data integration. Our tool uses machine learning algorithms that have been employed in different OMIC data analysis, including metabolomics (Liebal et al., 2020), proteomics (Suvarna et al., 2021) and cytokines (Kimita et al., 2022; Saharan et al., 2021).

The quality of the predictive models depends on the quality of the data set used in the analysis. Therefore, it is highly recommended to perform pre-processing steps to assess the overall data quality, existence of missing data and outliers, and perform an appropriate normalization. A careful examination of BioFeatS' output is also important, since it displays the algorithm performance and can aid in identifying samples that may have been misclassified. Finally, a well processed dataset endows a richer range of insights provided by the BiofeatS algorithm.

Conflict of Interests

Author IM is employed by Microbiotec srl.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgments

The authors thank all participants in the cohort studies. This work was supported by grants from the Innovative Medicines Initiative 2 Joint Undertaking under the VSV-EBOVAC (grant number 115842) and VSV-EBOPLUS (grant number 116068) projects within the Innovative Medicines Initiative Ebola+ program. Conduction of the North American trial was funded in part with Federal funds from the Department of Health and Human Services; Office of the Assistant Secretary for Preparedness and Response; Biomedical Advanced Research and Development Authority, under contract number HHSO100201500002C.

IM received a PhD fellowship under the Marie Sklodowska-Curie actions (MSCA) – Innovative Training Networks (ITN), Project VacPath (Novel vaccine vectors to resist pathogen challenge) grant agreement No 812915 funded by the European Union's Horizon 2020 research and innovation programme.

ISABELLE FRANCO MOSCARDINI
16/09/2023 20:01:30

Members of the VSV-EBOVAC Consortium

Selidji T Agnandji, Rafi Ahmed, Jenna Anderson, Floriane Auderset, Philip Bejon, Luisa Borgianni, Jessica Brosnahan, Annalisa Ciabattini, Olivier Engler, Mariëlle C Haks, Ali M Harandi, Donald Gray Heppner, Alice Gerlini, Angela Huttner, Peter G Kremsner, Donata Medaglini, Thomas P Monath, Francis M Ndungu, Patricia Njuguna, Tom H M Ottenhoff, David Pejoski, Mark Page, Gianni Pozzi, Francesco Santoro, and Claire-Anne Siegrist.

Isabelle Franco moscardini

Members of the VSV-EBOPLUS Consortium

Selidji T Agnandji, Luisa Borgianni, Annalisa Ciabattini, Sheri Dubey, Michael Eichberg, Olivier Engler, Alice Gerlini, Patricia Conceição Gonzalez Dias Carvalho, Mariëlle C Haks, Ali M Harandi, Angela Huttner, Peter G Kremsner, Kabwende Lumeka, Donata Medaglini, Helder I Nakaya, Sravya S Nakka, Essone P Ndong, Tom H M Ottenhoff, Gianni Pozzi, Sylvia Rothenberger, Francesco Santoro, Claire-Anne Siegrist, Suzanne van Veen, Eleonora Vianello.

REFERENCES

Agnandji, S.T., Huttner, A., Zinser, M.E., Njuguna, P., Dahlke, C., Fernandes, J.F., Yerly, S., Dayer, J.-A., Kraehling, V., Kasonta, R., Adegnika, A.A., Altfeld, M., Auderset, F., Bache, E.B., Biedenkopf, N., Borregaard, S., Brosnahan, J.S., Burrow, R., Combescure, C., Desmeules, J., Eickmann, M., Fehling, S.K., Finckh, A., Goncalves, A.R., Grobusch, M.P., Hooper, J., Jambrecina, A., Kabwende, A.L., Kaya, G., Kimani, D., Lell, B., Lemaître, B., Lohse, A.W., Massinga-Loembe, M., Matthey, A., Mordmüller, B., Nolting, A., Ogwang, C., Ramharter, M., Schmidt-Chanasit, J., Schmiedel, S., Silvera, P., Stahl, F.R., Staines, H.M., Strecker, T., Stubbe, H.C., Tsofa, B., Zaki, S., Fast, P., Moorthy, V., Kaiser, L., Krishna, S., Becker, S., Kieny, M.-P., Bejon, P., Kremsner, P.G., Addo, M.M., Siegrist, C.-A., 2016. Phase 1 Trials of rVSV Ebola Vaccine in Africa and Europe. N. Engl. J. Med. 374, 1647–1660. https://doi.org/10.1056/NEJMoa1502924

Ahmed, M., Lyles, D.S., 1998. Effect of Vesicular Stomatitis Virus Matrix Protein on Transcription Directed by Host RNA Polymerases I, II, and III. J. Virol. 72, 8413–8419. https://doi.org/10.1128/JVI.72.10.8413-8419.1998

Bao, X., Sinha, M., Liu, T., Hong, C., Luxon, B.A., Garofalo, R.P., Casola, A., 2008. Identification of human metapneumovirus-induced gene networks in airway epithelial cells by microarray analysis. Virology 374, 114–127. https://doi.org/10.1016/j.virol.2007.12.024

Celestino, I., Checconi, P., Amatore, D., De Angelis, M., Coluccio, P., Dattilo, R., Alunni Fegatelli, D., Clemente, A.M., Matarrese, P., Torcia, M.G., Mancinelli, R., Mammola, C.L., Garaci, E., Vestri, A.R., Malorni, W., Palamara, A.T., Nencioni, L., 2018. Differential Redox State Contributes to Sex Disparities in the Response to Influenza Virus Infection in Male and Female Mice. Front. Immunol. 9, 1747. https://doi.org/10.3389/fimmu.2018.01747

Connor, J.H., Lyles, D.S., 2005. Inhibition of Host and Viral Translation during Vesicular Stomatitis Virus Infection. J. Biol. Chem. 280, 13512–13519. https://doi.org/10.1074/jbc.M501156200

Farci, P., Diaz, G., Chen, Z., Govindarajan, S., Tice, A., Agulto, L., Pittaluga, S., Boon, D., Yu, C., Engle, R.E., Haas, M., Simon, R., Purcell, R.H., Zamboni, F., 2010. B cell gene signature with massive intrahepatic production of antibodies to hepatitis B core antigen in hepatitis B virus-associated acute liver failure. Proc. Natl. Acad. Sci. U. S. A. 107, 8766–8771. https://doi.org/10.1073/pnas.1003854107

Gale, M., Tan, S.-L., Katze, M.G., 2000. Translational Control of Viral Gene Expression in Eukaryotes. Microbiol. Mol. Biol. Rev. 64, 239–280. https://doi.org/10.1128/MMBR.64.2.239-280.2000

Goecks, J., Jalili, V., Heiser, L.M., Gray, J.W., 2020. How Machine Learning Will Transform Biomedicine. Cell 181, 92–101. https://doi.org/10.1016/j.cell.2020.03.022

Gonzalez-Dias, P., Lee, E.K., Sorgi, S., de Lima, D.S., Urbanski, A.H., Silveira, E.L., Nakaya, H.I., 2020. Methods for predicting vaccine immunogenicity and reactogenicity. Hum. Vaccines Immunother. 16, 269–276. https://doi.org/10.1080/21645515.2019.1697110

Jacob, S.T., Crozier, I., Fischer, W.A., Hewlett, A., Kraft, C.S., Vega, M.-A. de L., Soka, M.J., Wahl, V., Griffiths, A., Bollinger, L., Kuhn, J.H., 2020. Ebola virus disease. Nat. Rev. Dis. Primer 6, 13. https://doi.org/10.1038/s41572-020-0147-3

James, G., Witten, D., Hastie, T., Tibshirani, R. (Eds.), 2013. An introduction to statistical learning: with applications in R, Springer texts in statistics. Springer, New York.

Jones, S.M., Feldmann, H., Ströher, U., Geisbert, J.B., Fernando, L., Grolla, A., Klenk, H.-D., Sullivan, N.J., Volchkov, V.E., Fritz, E.A., Daddario, K.M., Hensley, L.E., Jahrling, P.B., Geisbert, T.W., 2005. Live attenuated recombinant vaccine protects nonhuman primates against Ebola and Marburg viruses. Nat. Med. 11, 786–790. https://doi.org/10.1038/nm1258

Kimita, W., Bharmal, S.H., Ko, J., Petrov, M.S., 2022. Identifying endotypes of individuals after an attack of pancreatitis based on unsupervised machine learning of multiplex cytokine profiles. Transl. Res. J. Lab. Clin. Med. S1931-5244(22)00166–9. https://doi.org/10.1016/j.trsl.2022.07.001

Kuriakose, T., Zheng, M., Neale, G., Kanneganti, T.-D., 2018. IRF1 Is a Transcriptional Regulator of ZBP1 Promoting NLRP3 Inflammasome Activation and Cell Death during Influenza Virus Infection. J. Immunol. 200, 1489–1495. https://doi.org/10.4049/jimmunol.1701538

Li, R., Liao, G., Nirujogi, R.S., Pinto, S.M., Shaw, P.G., Huang, T.-C., Wan, J., Qian, J., Gowda, H., Wu, X., Lv, D.-W., Zhang, K., Manda, S.S., Pandey, A., Hayward, S.D., 2015. Phosphoproteomic Profiling Reveals Epstein-Barr Virus Protein Kinase Integration of DNA Damage Response and Mitotic Signaling. PLOS Pathog. 11, e1005346. https://doi.org/10.1371/journal.ppat.1005346

Li, S., Rouphael, N., Duraisingham, S., Romero-Steiner, S., Presnell, S., Davis, C., Schmidt, D.S., Johnson, S.E., Milton, A., Rajam, G., Kasturi, S., Carlone, G.M., Quinn, C., Chaussabel, D., Palucka, A.K., Mulligan, M.J., Ahmed, R., Stephens, D.S., Nakaya, H.I., Pulendran, B., 2014. Molecular signatures of antibody responses derived from a systems biology study of five human vaccines. Nat. Immunol. 15, 195–204. https://doi.org/10.1038/ni.2789

Liebal, U.W., Phan, A.N.T., Sudhakar, M., Raman, K., Blank, L.M., 2020. Machine Learning Applications for Mass Spectrometry-Based Metabolomics. Metabolites 10, E243. https://doi.org/10.3390/metabo10060243

Malvy, D., McElroy, A.K., de Clerck, H., Günther, S., van Griensven, J., 2019. Ebola virus disease. The Lancet 393, 936–948. https://doi.org/10.1016/S0140-6736(18)33132-5

Mrozek-Gorska, P., Buschle, A., Pich, D., Schwarzmayr, T., Fechtner, R., Scialdone, A., Hammerschmidt, W., 2019. Epstein–Barr virus reprograms human B lymphocytes immediately in the prelatent phase of infection. Proc. Natl. Acad. Sci. 116, 16046–16055. https://doi.org/10.1073/pnas.1901314116

Panda, D., Das, A., Dinh, P.X., Subramaniam, S., Nayak, D., Barrows, N.J., Pearson, J.L., Thompson, J., Kelly, D.L., Ladunga, I., Pattnaik, A.K., 2011. RNAi screening reveals requirement for host cell secretory pathway in infection by diverse families of negative-strand RNA viruses. Proc. Natl. Acad. Sci. 108, 19036–19041. https://doi.org/10.1073/pnas.1113643108

Rajani, K.R., Pettit Kneller, E.L., McKenzie, M.O., Horita, D.A., Chou, J.W., Lyles, D.S., 2012. Complexes of Vesicular Stomatitis Virus Matrix Protein with Host Rae1 and Nup98 Involved in Inhibition of Host Transcription. PLoS Pathog. 8, e1002929. https://doi.org/10.1371/journal.ppat.1002929

Rechtien, A., Richert, L., Lorenzo, H., Martrus, G., Hejblum, B., Dahlke, C., Kasonta, R., Zinser, M., Stubbe, H., Matschl, U., Lohse, A., Krähling, V., Eickmann, M., Becker, S., Thiébaut, R., Altfeld, M., Addo, M., Agnandji, S.T., Krishna, S., Kremsner, P.G., Brosnahan, J.S., Bejon, P., Njuguna, P., Addo, M.M., Becker, S., Krähling, V., Siegrist, C.-A., Huttner, A., Kieny, M.-P., Moorthy, V., Fast, P., Savarese, B., Lapujade, O., 2017. Systems Vaccinology Identifies an Early Innate Immune Signature as a Correlate of Antibody Responses to the Ebola Vaccine rVSV-ZEBOV. Cell Rep. 20, 2251–2261. https://doi.org/10.1016/j.celrep.2017.08.023

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

Saharan, S.S., Nagar, P., Creasy, K.T., Stock, E.O., Feng, J., Malloy, M.J., Kane, J.P., 2021. Machine learning and statistical approaches for classification of risk of coronary artery disease using plasma cytokines. BioData Min. 14, 26. https://doi.org/10.1186/s13040-021-00260-z

Santoro, F., Donato, A., Lucchesi, S., Sorgi, S., Gerlini, A., Haks, M., Ottenhoff, T., Gonzalez-Dias, P., Consortium, Vsv-Ebovac, Consortium, Vsv-Eboplus, Nakaya, H., Huttner, A., Siegrist, C.-A., Medaglini, D., Pozzi, G., 2021. Human Transcriptomic Response to the VSV-Vectored Ebola Vaccine. Vaccines 9, 67. https://doi.org/10.3390/vaccines9020067

Sartori, G., Pesarico, A.P., Pinton, S., Dobrachinski, F., Roman, S.S., Pauletto, F., Rodrigues, L.C., Prigol, M., 2012. Protective effect of brown Brazilian propolis against acute vaginal lesions caused by herpes simplex virus type 2 in mice: involvement of antioxidant and anti-inflammatory mechanisms: PROTECTIVE EFFECT OF BROWN BRAZILIAN

PROPOLIS. Cell Biochem. Funct. 30, 1–10. https://doi.org/10.1002/cbf.1810
Schneider, R.J., Mohr, I., 2003. Translation initiation and viral tricks. Trends Biochem. Sci. 28, 130–136. https://doi.org/10.1016/S0968-0004(03)00029-X
Sims, A.C., Tilton, S.C., Menachery, V.D., Gralinski, L.E., Schäfer, A., Matzke, M.M., Webb-Robertson, B.-J.M., Chang, J., Luna, M.L., Long, C.E., Shukla, A.K., Bankhead, A.R., Burkett, S.E., Zornetzer, G., Tseng, C.-T.K., Metz, T.O., Pickles, R., McWeeney, S., Smith, R.D., Katze, M.G., Waters, K.M., Baric, R.S., 2013. Release of severe acute respiratory syndrome coronavirus nuclear import block enhances host transcription in human lung cells. J. Virol. 87, 3885–3902. https://doi.org/10.1128/JVI.02520-12
Strong, J.E., Wong, G., Jones, S.E., Grolla, A., Theriault, S., Kobinger, G.P., Feldmann, H., 2008. Stimulation of Ebola virus production from persistent infection through activation of the Ras/MAPK pathway. Proc. Natl. Acad. Sci. 105, 17982–17987. https://doi.org/10.1073/pnas.0809698105
Subbaiah, K.C.V., Raniprameela, D., Visweswari, G., Rajendra, W., Lokanatha, V., 2011. Perturbations in the antioxidant metabolism during Newcastle disease virus (NDV) infection in chicken: Protective role of vitamin E. Naturwissenschaften 98, 1019–1026. https://doi.org/10.1007/s00114-011-0855-3
Suvarna, K., Biswas, D., Pai, M.G.J., Acharjee, A., Bankar, R., Palanivel, V., Salkar, A., Verma, A., Mukherjee, A., Choudhury, M., Ghantasala, S., Ghosh, S., Singh, A., Banerjee, A., Badaya, A., Bihani, S., Loya, G., Mantri, K., Burli, A., Roy, J., Srivastava, A., Agrawal, S., Shrivastav, O., Shastri, J., Srivastava, S., 2021. Proteomics and Machine Learning Approaches Reveal a Set of Prognostic Markers for COVID-19 Severity With Drug Repurposing Potential. Front. Physiol. 12, 652799. https://doi.org/10.3389/fphys.2021.652799
Vermijlen, D., Brouwer, M., Donner, C., Liesnard, C., Tackoen, M., Van Rysselberge, M., Twité, N., Goldman, M., Marchant, A., Willems, F., 2010. Human cytomegalovirus elicits fetal gammadelta T cell responses in utero. J. Exp. Med. 207, 807–821. https://doi.org/10.1084/jem.20090348
Weiner 3rd, J., Domaszewska, T., 2016. tmod: an R package for general and multivariate enrichment analysis (preprint). PeerJ Preprints. https://doi.org/10.7287/peerj.preprints.2420v1
Yamada, Y., Limmon, G.V., Zheng, D., Li, N., Li, L., Yin, L., Chow, V.T.K., Chen, J., Engelward, B.P., 2012. Major Shifts in the Spatio-Temporal Distribution of Lung Antioxidant Enzymes during Influenza Pneumonia. PLoS ONE 7, e31494. https://doi.org/10.1371/journal.pone.0031494
Zaslavsky, E., Hershberg, U., Seto, J., Pham, A.M., Marquez, S., Duke, J.L., Wetmur, J.G., tenOever, B.R., Sealfon, S.C., Kleinstein, S.H., 2010. Antiviral Response Dictated by Choreographed Cascade of Transcription Factors. J. Immunol. 184, 2908–2917. https://doi.org/10.4049/jimmunol.0903453

Supplementary File 1

Biological Feature Selection (BioFeatS)

The Biological Feature Selection Tool (BioFeatS) was developed to select the best features for classifying biological groups. This tool can be applied for different types of data, especially for those provenient from high-throughput technologies, which contains large amounts of features. BioFeatS is based on the Machine Learning approach divided into three distinct stages: i) feature selection; ii) ranking and; iii) evaluation. The BioFeatS combines different selection methods to choose only the most informative features (e.g. genes, proteins, microRNAs, cytokines etc.), then proceeds with ranking, assessing their importance for the machine learning model and, finally, evaluating the selected features in the trained models, by using distinct machine learning algorithms.

Selection of Attributes

Feature Selection techniques aim to select a subset of features of greater relevance for the construction of the predictive model (1). The central premise when using Feature Selection techniques is that most datasets contain redundant or irrelevant Features for the learning of the algorithm and, therefore, can be removed without leading to loss of information in the model (2). This provides benefits such as the reduction of overfitting and training time, as well an increased accuracy of the model (1, 2).

In this sense, for the development of the feature selection stage of BioFeatS, three techniques of Features Selection were used, namely:

I. Pearson correlation: verifies the absolute value of the Person correlation between the response variable and the numerical Features of the data set (3). For BioFeatS we have established an N number of Features with the highest correlation;
II. kBest: selects resources according to the highest scoring k (4). For BioFeatS the amount of Features selected corresponds to the number N established.
III. Recursive Feature Elimination (RFE): selects features recursively considering sets of Features increasingly smaller. First, an estimator is trained in the initial set of Features and the importance of each Feature is obtained (for BioFeatS Support Vector Regression is used). Then, the less important Features are eliminated from the current set (5). The procedure is recursively repeated in the obtained set until the N number of Features is reached.

After execution, each Feature Selection technique provides a list of Features of greater relevance according to the employed methodology. From this, a single list is generated with the intersection of Features present in at least two of the three techniques.

It is important to highlight that, for the calculation of the N number of Features used in BioFeatS in Feature Selection techniques, the total number of Features contained in the single list after the execution of the Features intersection is taken into account. The calculation of the number N is obtained by the following equation:

$$\frac{NF}{N} > 0.5$$

where NF is the amount of Features from the single list and N is the value selected for the Feature Selection techniques. In order to drastically reduce the number of features, the number N initially receives the value of 100. If the equation is not satisfied, N is then incremented by 100 until the condition of the equation 1 is satisfied. Therefore the output list of features (NF) will be at least 50% of the total amount of the input. In case the input file has less than 100 features, all features are selected instead of N, and the equation is not used.

Ranking

In decision trees, each node is a condition of how to divide values into a single resource. Such a condition is based on impurity, which in the case of classification problems is entropy and, for regression problems, is variance. In this sense, when training a decision tree it is possible to calculate how much each resource contributes to reduce the weighted impurity (6). The ordering step of BioFeatS is based on this logic, and for the calculation of Feature importance, the Random Forest algorithm is used, which uses the average of the decrease of impurity on the trees. Thus, after the feature selection step, the most relevant Features of the database are ordered according to their importance. It is worth noting that in this ordering stage, the features that have values of importance equal to zero are removed from the final list of Features.

Evaluation

Using different algorithms BioFeatS evaluates the quality of the chosen Features. These algorithms were selected due to the great diversity of their components, which means that they have different methodologies and they use different mathematical approaches to learn and classify the samples. In this way, it is possible to define a more generalized machine learning model. The algorithms chosen are:

I. Support Vector Machine (SVM): establishes a hyperplane in an N-dimensional space (N - number of resources) that distinctly sorts data points (7).
II. k-Nearest Neighbors (kNN): uses the proximity of the data to perform the classification/prediction on the grouping of an individual data point (8).
III. Naive Bayes: uses the probabilistic paradigm to perform classification tasks, based on the Bayes theorem (9).
IV. AdaBoost Classifier: uses joint learning methods (meta-learning), using an iterative approach to learn from "weak" classifier errors and turn them into strong classifiers (10).

For the performance analysis of the selected algorithms, the methodology of Experimental Planning and Evaluation (11) is used. Moreover, in BioFeatS, a k-fold cross-validation is used, with k = 10, being k-1 for training and the rest for testing. Thus, it is possible to measure the error estimate more accurately, since the average value estimate tends to a real zero error rate as it increases n, which is usually the case for small sets of examples (11).

Finally, the evaluation of each algorithm is assessed by four commonly used classification metrics: Area Under the ROC Curve (AUC), Precision, Accuracy and the F1-score, which is a combination of Precision and Recall metrics. The final output of the model includes the value of each metric for each algorithm evaluated and the final mean and harmonic mean of these values.

References

1. Zebari, R., Abdulazeez, A., Zeebaree, D., Zebari, D., & Saeed, J. (2020). A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. *Journal of Applied Science and Technology Trends*, *1*(2), 56-70.

2. Zhu, Y., Ma, J., Yuan, C., & Zhu, X. (2022). Interpretable learning based dynamic graph convolutional networks for Alzheimer's disease analysis. *Information Fusion*, *77*, 53-61.

3. Liu, Y., Mu, Y., Chen, K., Li, Y., & Guo, J. (2020). Daily activity feature selection in smart homes based on pearson correlation coefficient. *Neural Processing Letters*, *51*(2), 1771-1787.

4. Dissanayake, K., & Md Johar, M. G. (2021). Comparative Study on Heart Disease Prediction Using Feature Selection Techniques on Classification Algorithms. *Applied Computational Intelligence and Soft Computing*, *2021*.

5. Han, Y., Huang, L., & Zhou, F. (2021). A dynamic recursive feature elimination framework (dRFE) to further refine a set of OMIC biomarkers. *Bioinformatics*, *37*(15), 2183-2189.

6. Hasanin, T., Khoshgoftaar, T. M., Leevy, J., & Seliya, N. (2019, April). Investigating random undersampling and feature selection on bioinformatics big data. In *2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService)* (pp. 346-356). IEEE.

7. Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, *408*, 189-215.

8. Tabares-Soto, R., Orozco-Arias, S., Romero-Cano, V., Bucheli, V. S., Rodríguez-Sotelo, J. L., & Jiménez-Varón, C. F. (2020). A comparative study of machine learning and deep learning algorithms to classify cancer types based on microarray gene expression data. *PeerJ Computer Science*, *6*, e270.

9. Sen, P. C., Hajra, M., & Ghosh, M. (2020). Supervised classification algorithms in machine learning: A survey and review. In *Emerging technology in modelling and graphics* (pp. 99-111). Springer, Singapore.

10. Almustafa, K. M. (2020). Prediction of heart disease and classifiers' sensitivity analysis. BMC bioinformatics, 21(1), 1-18.

11. Mano, L. Y., Faiçal, B. S., Gonçalves, V. P., Pessin, G., Gomes, P. H., de Carvalho, A. C., & Ueyama, J. (2020). An intelligent and generic approach for detecting human emotions: a case study with facial expressions. *Soft Computing*, *24*(11), 8467-8479.

4.4 Final discussion

The differences in the transcriptomic responses observed between the Swiss and North American cohorts can be due to different factors. The cohorts presented distinct vaccine doses and gender balance, two factors that could impact both the innate and adaptive responses (Flanagan et al., 2017; Rechtien et al., 2017). Moreover, the number of volunteers in each study was also different.

When comparing these responses with a third cohort, using two different methodologies, we see that the differences highlighted in the DE analysis are less evident when using BioFeatS, with some of the modules related to the innate responses being present in all of the three cohorts and the "plasma cells, immunoglobulins" module activated also in the Geneva cohort. Therefore, selecting features by Differential Expression analysis or by a combined Machine learning approach lead to distinct selection of genes, which also implicate distinct enrichment results .

The BioFeatS framework was built with the aim of bringing meaningful biological information while also selecting the best features in distinguishing two different classes, and this methodology has shown to provide a consistent result among independent cohorts. Moreover, BioFeatS could also facilitate data integration processes, since it can be applied to different OMIC data.

In brief, the methodology's choice should take in consideration particularities of each data set and the aims of each project, but our data suggests that a combination of DE analysis with a Machine Learning approach could bring a more robust and consistent signature.

References

Agnandji, S.T., Huttner, A., Zinser, M.E., Njuguna, P., Dahlke, C., Fernandes, J.F., Yerly, S., Dayer, J.-A., Kraehling, V., Kasonta, R., Adegnika, A.A., Altfeld, M., Auderset, F., Bache, E.B., Biedenkopf, N., Borregaard, S., Brosnahan, J.S., Burrow, R., Combescure, C., Desmeules, J., Eickmann, M., Fehling, S.K., Finckh, A., Goncalves, A.R., Grobusch, M.P., Hooper, J., Jambrecina, A., Kabwende, A.L., Kaya, G., Kimani, D., Lell, B., Lemaître, B., Lohse, A.W., Massinga-Loembe, M., Matthey, A., Mordmüller, B., Nolting, A., Ogwang, C., Ramharter, M., Schmidt-Chanasit, J., Schmiedel, S., Silvera, P., Stahl, F.R., Staines, H.M., Strecker, T., Stubbe, H.C., Tsofa, B., Zaki, S., Fast, P., Moorthy, V., Kaiser, L., Krishna, S., Becker, S., Kieny, M.-P., Bejon, P., Kremsner, P.G., Addo, M.M., Siegrist, C.-A., 2016. Phase 1 Trials of rVSV Ebola Vaccine in Africa and Europe. N. Engl. J. Med. 374, 1647–1660. https://doi.org/10.1056/NEJMoa1502924

Bench-Capon, T.J.M., Dunne, P.E., 2007. Argumentation in artificial intelligence. Artif. Intell. 171, 619–641. https://doi.org/10.1016/j.artint.2007.05.001

Bharat, T.A.M., Noda, T., Riches, J.D., Kraehling, V., Kolesnikova, L., Becker, S., Kawaoka, Y., Briggs, J.A.G., 2012. Structural dissection of Ebola virus and its assembly determinants using cryo-electron tomography. Proc. Natl. Acad. Sci. 109, 4275–4280. https://doi.org/10.1073/pnas.1120453109

Chertow, D.S., Kleine, C., Edwards, J.K., Scaini, R., Giuliani, R., Sprecher, A., 2014. Ebola Virus Disease in West Africa — Clinical Manifestations and Management. N. Engl. J. Med. 371, 2054–2057. https://doi.org/10.1056/NEJMp1413084

Clark, D.V., Kibuuka, H., Millard, M., Wakabi, S., Lukwago, L., Taylor, A., Eller, M.A., Eller, L.A., Michael, N.L., Honko, A.N., Olinger, G.G., Schoepp, R.J., Hepburn, M.J., Hensley, L.E., Robb, M.L., 2015. Long-term sequelae after Ebola virus disease in Bundibugyo, Uganda: a retrospective cohort study. Lancet Infect. Dis. 15, 905–912. https://doi.org/10.1016/S1473-3099(15)70152-0

Dowell, S.F., Mukunu, R., Ksiazek, T.G., Khan, A.S., Rollin, P.E., Peters, C.J., for the Commission de Lutte contre les Epidémies à Kikwit, 1999. Transmission of Ebola Hemorrhagic Fever: A Study of Risk Factors in Family Members, Kikwit, Democratic Republic of the Congo, 1995. J. Infect. Dis. 179, S87–S91. https://doi.org/10.1086/514284

Feldmann, H., Jones, S., Klenk, H.-D., Schnittler, H.-J., 2003. Ebola virus: from discovery to vaccine. Nat. Rev. Immunol. 3, 677–685. https://doi.org/10.1038/nri1154

Flanagan, K.L., Fink, A.L., Plebanski, M., Klein, S.L., 2017. Sex and Gender Differences in the Outcomes of Vaccination over the Life Course. Annu. Rev. Cell Dev. Biol. 33, 577–599. https://doi.org/10.1146/annurev-cellbio-100616-060718

Garbutt, M., Liebscher, R., Wahl-Jensen, V., Jones, S., Möller, P., Wagner, R., Volchkov, V., Klenk, H.-D., Feldmann, H., Ströher, U., 2004. Properties of Replication-Competent Vesicular Stomatitis Virus Vectors Expressing Glycoproteins of Filoviruses and Arenaviruses. J. Virol. 78, 5458–5465. https://doi.org/10.1128/JVI.78.10.5458-5465.2004

Goecks, J., Jalili, V., Heiser, L.M., Gray, J.W., 2020. How Machine Learning Will Transform Biomedicine. Cell 181, 92–101. https://doi.org/10.1016/j.cell.2020.03.022

Halperin, S.A., Das, R., Onorato, M.T., Liu, K., Martin, J., Grant-Klein, R.J., Nichols, R., Coller, B.-A., Helmond, F.A., Simon, J.K., 2019. Immunogenicity, Lot Consistency, and Extended Safety of rVSVΔG-ZEBOV-GP Vaccine: A Phase 3 Randomized, Double-Blind, Placebo-Controlled Study in Healthy Adults. J. Infect. Dis. 220, 1127–1135. https://doi.org/10.1093/infdis/jiz241

Hayman, D.T.S., Emmerich, P., Yu, M., Wang, L.-F., Suu-Ire, R., Fooks, A.R., Cunningham, A.A., Wood, J.L.N., 2010. Long-Term Survival of an Urban Fruit Bat Seropositive for Ebola and Lagos Bat Viruses. PLoS ONE 5, e11978. https://doi.org/10.1371/journal.pone.0011978

Henao-Restrepo, A.M., Camacho, A., Longini, I.M., Watson, C.H., Edmunds, W.J., Egger, M., Carroll, M.W., Dean, N.E., Diatta, I., Doumbia, M., Draguez, B., Duraffour, S., Enwere, G., Grais, R., Gunther, S., Gsell, P.-S., Hossmann, S., Watle, S.V., Kondé, M.K., Kéïta, S., Kone, S., Kuisma, E., Levine, M.M., Mandal, S., Mauget, T., Norheim, G., Riveros, X., Soumah, A., Trelle, S., Vicari, A.S., Røttingen, J.-A., Kieny, M.-P., 2017. Efficacy and effectiveness of an rVSV-vectored vaccine in preventing Ebola virus disease: final results from the Guinea ring vaccination, open-label, cluster-randomised trial (Ebola Ça Suffit!). The Lancet 389, 505–518. https://doi.org/10.1016/S0140-6736(16)32621-6

Heppner, D.G., Kemp, T.L., Martin, B.K., Ramsey, W.J., Nichols, Richard, Dasen, E.J., Link, C.J., Das, Rituparna, Xu, Z.J., Sheldon, E.A., Nowak, T.A., Monath, T.P., Heppner, D., Kemp, T., Martin, B., Ramsey, W., Nichols, R, Dasen, E., Fusco, J., Crowell, J., Link, C., Creager, J., Monath, T., Das, R, Xu, Z., Klein, R., Nowak, T., Gerstenberger, E., Bliss, R., Sheldon, E., Feldman, R., Essink, B.J., Smith, W., Chu, L., Seger, W., Saleh, J., Borders, J., Adams, M., 2017. Safety and immunogenicity of the rVSVΔG-ZEBOV-GP Ebola virus vaccine candidate in healthy adults: a phase 1b randomised, multicentre, double-blind, placebo-controlled, dose-response study. Lancet Infect. Dis. 17, 854–866. https://doi.org/10.1016/S1473-3099(17)30313-4

Howlett, P.J., Walder, A.R., Lisk, D.R., Fitzgerald, F., Sevalie, S., Lado, M., N'jai, A., Brown, C.S., Sahr, F., Sesay, F., Read, J.M., Steptoe, P.J., Beare, N.A.V., Dwivedi, R., Solbrig, M., Deen, G.F., Solomon, T., Semple, M.G., Scott, J.T., 2018. Case Series of Severe Neurologic Sequelae of Ebola Virus Disease during Epidemic, Sierra Leone. Emerg. Infect. Dis. 24, 1412–1421. https://doi.org/10.3201/eid2408.171367

Huttner, A., Dayer, J.-A., Yerly, S., Combescure, C., Auderset, F., Desmeules, J., Eickmann, M., Finckh, A., Goncalves, A.R., Hooper, J.W., Kaya, G., Krähling, V., Kwilas, S., Lemaître, B., Matthey, A., Silvera, P., Becker, S., Fast, P.E., Moorthy, V., Kieny, M.P., Kaiser, L., Siegrist, C.-A., 2015. The effect of dose on the safety and immunogenicity of the VSV Ebola candidate vaccine: a randomised double-blind, placebo-controlled phase 1/2 trial. Lancet Infect. Dis. 15, 1156–1166. https://doi.org/10.1016/S1473-3099(15)00154-1

Jadav, S., Kumar, A., Ahsan, M., Jayaprakash, V., 2015. Ebola Virus: Current and Future Perspectives. Infect. Disord. - Drug Targets 15, 20–31. https://doi.org/10.2174/1871526515666150320162259

Jones, S.M., Feldmann, H., Ströher, U., Geisbert, J.B., Fernando, L., Grolla, A., Klenk, H.-D., Sullivan, N.J., Volchkov, V.E., Fritz, E.A., Daddario, K.M., Hensley, L.E., Jahrling, P.B., Geisbert, T.W., 2005. Live attenuated recombinant vaccine protects nonhuman primates against Ebola and Marburg viruses. Nat. Med. 11, 786–790. https://doi.org/10.1038/nm1258

Kanapathipillai, R., Henao Restrepo, A.M., Fast, P., Wood, D., Dye, C., Kieny, M.-P., Moorthy, V., 2014. Ebola Vaccine — An Urgent International Priority. N. Engl. J. Med. 371, 2249–2251. https://doi.org/10.1056/NEJMp1412166

Koch, L.K., Cunze, S., Kochmann, J., Klimpel, S., 2020. Bats as putative Zaire ebolavirus reservoir hosts and their habitat suitability in Africa. Sci. Rep. 10, 14268. https://doi.org/10.1038/s41598-020-71226-0

Laverty, H., Meulien, P., 2019. The Innovative Medicines Initiative −10 Years of Public-Private Collaboration. Front. Med. 6, 275. https://doi.org/10.3389/fmed.2019.00275

Leligdowicz, A., Fischer, W.A., Uyeki, T.M., Fletcher, T.E., Adhikari, N.K.J., Portella, G., Lamontagne, F., Clement, C., Jacob, S.T., Rubinson, L., Vanderschuren, A., Hajek, J., Murthy, S., Ferri, M., Crozier, I., Ibrahima, E., Lamah, M.-C., Schieffelin, J.S., Brett-Major, D., Bausch, D.G., Shindo, N., Chan, A.K., O'Dempsey, T., Mishra, S., Jacobs, M., Dickson, S., Lyon, G.M., Fowler, R.A., 2016. Ebola virus disease and critical illness. Crit. Care 20, 217. https://doi.org/10.1186/s13054-016-1325-2

Leroy, E.M., Kumulungui, B., Pourrut, X., Rouquet, P., Hassanin, A., Yaba, P., Délicat, A., Paweska, J.T., Gonzalez, J.-P., Swanepoel, R., 2005. Fruit bats as reservoirs of Ebola virus. Nature 438, 575–576. https://doi.org/10.1038/438575a

Li, S., Rouphael, N., Duraisingham, S., Romero-Steiner, S., Presnell, S., Davis, C., Schmidt, D.S., Johnson, S.E., Milton, A., Rajam, G., Kasturi, S., Carlone, G.M., Quinn, C., Chaussabel, D., Palucka, A.K., Mulligan, M.J., Ahmed, R., Stephens, D.S., Nakaya, H.I., Pulendran, B., 2014. Molecular signatures of antibody responses derived from a systems biology study of five human vaccines. Nat. Immunol. 15, 195–204. https://doi.org/10.1038/ni.2789

Martin, B., Canard, B., Decroly, E., 2017. Filovirus proteins for antiviral drug discovery: Structure/function bases of the replication cycle. Antiviral Res. 141, 48–61. https://doi.org/10.1016/j.antiviral.2017.02.004

Marzi, A., Feldmann, H., 2014. Ebola virus vaccines: an overview of current approaches. Expert Rev. Vaccines 13, 521–531. https://doi.org/10.1586/14760584.2014.885841

Medaglini, D., Harandi, A.M., Ottenhoff, T.H.M., Siegrist, C.-A., Consortium, V.-E., Agnandji, S.T., Ahmed, R., Anderson, J., Auderset, F., Borgianni, L., Brosnahan, J., Ciabattini, A., Engler, O.,

Haks, M.C., Heppner, G., Gerlini, A., Kremsner, P.G., Leib, S., Monath, T., Ndungu, F., Njuguna, P., Page, M., Pozzi, G., Rappuoli, R., Santoro, F., 2015. Ebola vaccine R&D: Filling the knowledge gaps. Sci. Transl. Med. 7. https://doi.org/10.1126/scitranslmed.aad3106

Mohammadi, D., 2014. International community ramps up Ebola vaccine effort. The Lancet 384, 1658–1659. https://doi.org/10.1016/S0140-6736(14)61788-8

Ogawa, H., Miyamoto, H., Nakayama, E., Yoshida, R., Nakamura, I., Sawa, H., Ishii, A., Thomas, Y., Nakagawa, E., Matsuno, K., Kajihara, M., Maruyama, J., Nao, N., Muramatsu, M., Kuroda, M., Simulundu, E., Changula, K., Hang'ombe, B., Namangala, B., Nambota, A., Katampi, J., Igarashi, M., Ito, K., Feldmann, H., Sugimoto, C., Moonga, L., Mweene, A., Takada, A., 2015. Seroepidemiological Prevalence of Multiple Species of Filoviruses in Fruit Bats (*Eidolon helvum*) Migrating in Africa. J. Infect. Dis. 212, S101–S108. https://doi.org/10.1093/infdis/jiv063

Rechtien, A., Richert, L., Lorenzo, H., Martrus, G., Hejblum, B., Dahlke, C., Kasonta, R., Zinser, M., Stubbe, H., Matschl, U., Lohse, A., Krähling, V., Eickmann, M., Becker, S., Thiébaut, R., Altfeld, M., Addo, M., Agnandji, S.T., Krishna, S., Kremsner, P.G., Brosnahan, J.S., Bejon, P., Njuguna, P., Addo, M.M., Becker, S., Krähling, V., Siegrist, C.-A., Huttner, A., Kieny, M.-P., Moorthy, V., Fast, P., Savarese, B., Lapujade, O., 2017. Systems Vaccinology Identifies an Early Innate Immune Signature as a Correlate of Antibody Responses to the Ebola Vaccine rVSV-ZEBOV. Cell Rep. 20, 2251–2261. https://doi.org/10.1016/j.celrep.2017.08.023

Regules, J.A., Beigel, J.H., Paolino, K.M., Voell, J., Castellano, A.R., Hu, Z., Muñoz, P., Moon, J.E., Ruck, R.C., Bennett, J.W., Twomey, P.S., Gutiérrez, R.L., Remich, S.A., Hack, H.R., Wisniewski, M.L., Josleyn, M.D., Kwilas, S.A., Van Deusen, N., Mbaya, O.T., Zhou, Y., Stanley, D.A., Jing, W., Smith, K.S., Shi, M., Ledgerwood, J.E., Graham, B.S., Sullivan, N.J., Jagodzinski, L.L., Peel, S.A., Alimonti, J.B., Hooper, J.W., Silvera, P.M., Martin, B.K., Monath, T.P., Ramsey, W.J., Link, C.J., Lane, H.C., Michael, N.L., Davey, R.T., Thomas, S.J., 2017. A Recombinant Vesicular Stomatitis Virus Ebola Vaccine. N. Engl. J. Med. 376, 330–341. https://doi.org/10.1056/NEJMoa1414216

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

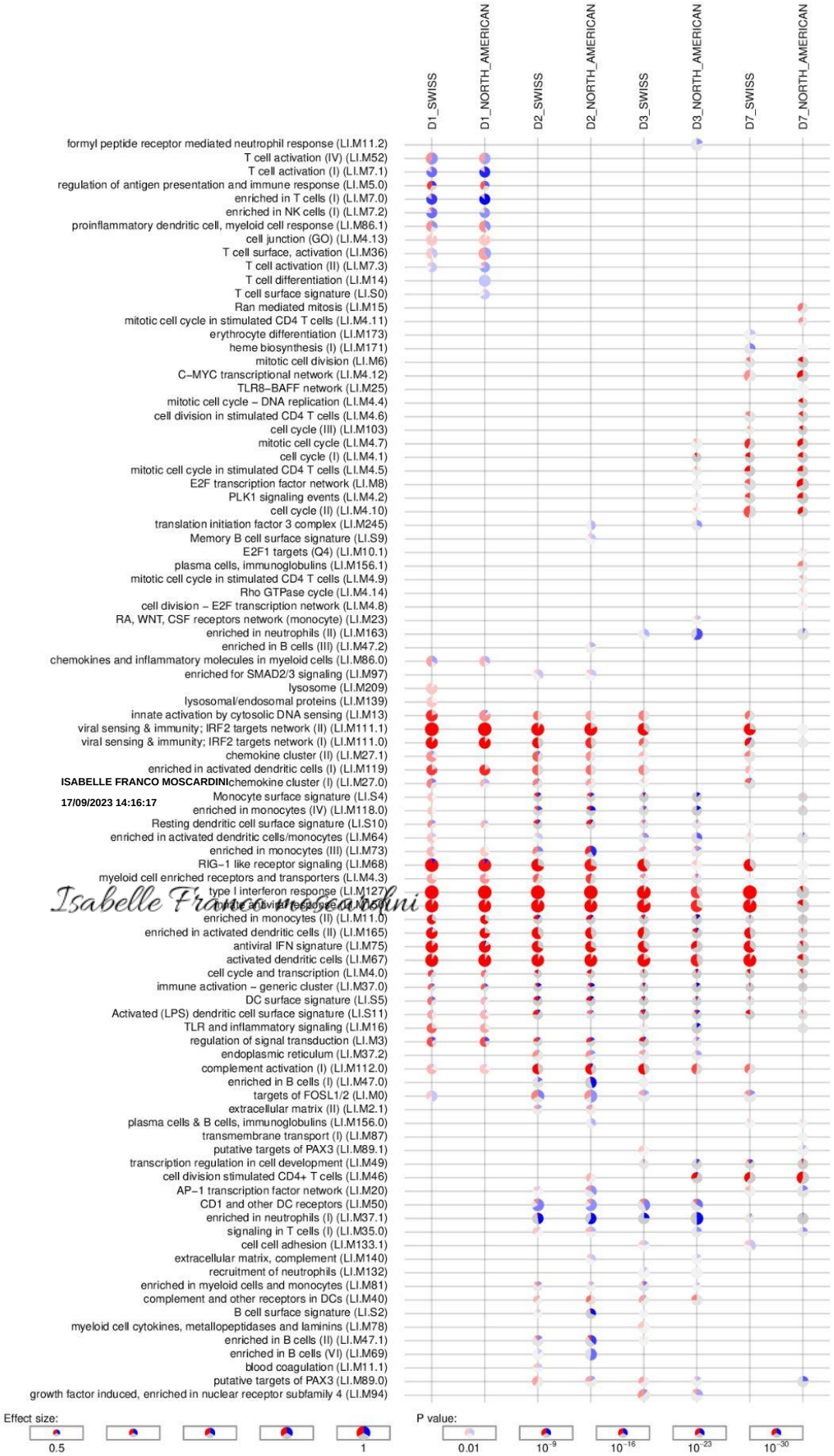
Santoro, F., Donato, A., Lucchesi, S., Sorgi, S., Gerlini, A., Haks, M., Ottenhoff, T., Gonzalez-Dias, P., Consortium, Vsv-Ebovac, Consortium, Vsv-Eboplus, Nakaya, H., Huttner, A., Siegrist, C.-A., Medaglini, D., Pozzi, G., 2021. Human Transcriptomic Response to the VSV-Vectored Ebola Vaccine. Vaccines 9, 67. https://doi.org/10.3390/vaccines9020067

Sun, L., Dong, S., Ge, Y., Fonseca, J.P., Robinson, Z.T., Mysore, K.S., Mehta, P., 2019. DiVenn: An Interactive and Integrated Web-Based Visualization Tool for Comparing Gene Lists. Front. Genet. 10, 421. https://doi.org/10.3389/fgene.2019.00421

Tiffany, A., Vetter, P., Mattia, J., Dayer, J.-A., Bartsch, M., Kasztura, M., Sterk, E., Tijerino, A.M., Kaiser, L., Ciglenecki, I., 2016. Ebola Virus Disease Complications as Experienced by Survivors in Sierra Leone. Clin. Infect. Dis. 62, 1360–1366. https://doi.org/10.1093/cid/ciw158

Weiner 3rd, J., Domaszewska, T., 2016. tmod: an R package for general and multivariate enrichment analysis (preprint). PeerJ Preprints. https://doi.org/10.7287/peerj.preprints.2420v1

Supplementary Figure 1



Transcriptional Analysis After Mucosal Priming by a Recombinant Vaccine Vector *Streptococcus gordonii* Expressing the MOMP Chlamydial antigen Reveals Enrichment of Specific Immune Pathways and identifies a Signature Correlated with Antibody Titers

Isabelle Franco Moscardini [1†], Bar Philosof [2†], Elena Pettini[2], Francesco Santoro[2], Alice Gerlini[1], Gianni Pozzi[2],

[1] Microbiotec srl, Siena, Italy

[2] Laboratory of Molecular Microbiology and Biotechnology (LAMMB), Department of Medical Biotechnologies, University of Siena, Italy.

† These authors have contributed equally to this work

ISABELLE FRANCO MOSCARDINI

Isabelle Franco moscardini

ABSTRACT

Mucosal surfaces are particularly vulnerable to infection, and vaccines targeting these areas would be of great importance. However, understanding the responses and protective mechanisms induced by mucosal vaccines have been challenging, particularly with regards to the genital tract, impacting the development of this type of vaccines. The transcriptomic analyses and Systems Biology approach emerges as potential tools to better study these immunological mechanisms and find new correlates of immunogenicity.

Streptococcus gordonii is a Gram-positive bacterium, member of the human oral microbiome and has been studied as a vaccine vector for different antigens and and sites of immunization. In this work we studied the transcriptomic response to the vaginal colonization with the Wild-Type or a recombinant *S. gordonii* expressing the CTH522 protein, a multivalent antigen composed of regions of the major outer membrane protein (MOMP) of *Chlamydia trachomatis*.

Combining the intravaginal immunizations and a subcutaneous boost with the CTH522 protein, we had access to the systemic responses through the transcriptomic analysis of the splenocytes from mice primed with either the WT or recombinant strain. The priming with the recombinant bacteria modulated different biological processes, including the activity of IL-1 and IL-2 signaling networks, the activation of transcription factors and T cell modules and the expression of genes like Ccl3. Moreover, a signature of genes correlated with the antibody response was identified, implicating the Interferon type I pathway, the cell cycle activity and genes like Ccl3 and Il18bp.

INTRODUCTION

Streptococcus gordonii is a Gram-positive bacterium and a member of the human oral microbiome, that can be genetically manipulated to express heterologous antigens based on chromosomal integration of a donor construct (1), and used as a vaccine delivery vector. We have previously shown that *S. gordonii* vectors expressing various antigens, and delivered in mucosal tissues, could protect from lethal toxin challenge (2) and activate different immune compartments such as antibody production and T-cell proliferation (3–9). In vaginal delivery to both mice and non-human primates, recombinant *S. gordonii* vectors successfully colonized the vaginal tract, induced antibody production both locally and systemically and resolved vaginal yeast infection by *Candida albicans* (10–12). *S. gordonii* vector was also found to be safe in a phase I clinical trial when administered nasally (13).

The ability of mucosal vaccination or infection to stimulate a systemic immune response is greatly dependent on the type of mucosal tissue (14). The vaginal mucosa is unique in its characteristics as it contains both type I and type II mucosal tissues, with their respective immunological features (15). Numerous studies have demonstrated that vaginal immunization can induce systemic cellular and humoral responses (16–20). However, despite these features it remains an underexplored route of immunization compared to other mucosal tissues such as the nasal and oral, due to their presence in both sexes, the simplicity of administration and decades of experience with pharmaceuticals delivery (21). Additionally, the vaginal tract is considered a complex site for antigen processing and requires adequate adjuvants (22–25).

Given the complexity of this route of immunization, new approaches may be useful for better understanding the response generated after vaccination. Systems Vaccinology presents itself as a possibility to assess the complexity behind the immune mechanisms that lead to protection (26). Different vaccines had their responses characterized through this approach, including Influenza (27,28), Yellow Fever (29) and Ebola (30,31).

In the context of mucosal vaccines, Systems Vaccinology has been pointed as promising to support the search for correlates of protection and provide important insights into the underlying mechanisms driving protective immunity (26). Vaccines against Tuberculosis (32), Influenza (33) and enteric diseases such as cholera and enterotoxigenic *Escherichia coli* (34), have recently had important mechanisms elucidated thanks to this approach. Moreover, important advances in understanding the mechanisms behind protection against simian immunodeficiency virus (SIV) infection were also achieved. This systems approach allowed the identification of early blood transcriptional signatures that correlate with antigen-specific antibody responses in vaginal secretions (35), and has contributed to the identification of synergic factors between host restriction factors and vaccine-induced immune responses (36).

Widely used in Systems Biology studies, transcriptomics offers the possibility of following and comparing the gene expression in various conditions, allowing the study of the biological processes after vaccination. In the present work, we examined the *in vitro* transcriptomic signature observed in splenocytes harvested from mice intravaginally immunized with either WT or recombinant *S. gordonii* strain expressing the *Chlamydia trachomatis* (C.t) multivalent MOMP antigen, CTH522 (37–39), and then boosted with the purified CTH522 protein.

Here, we show that vaginal priming with a recombinant *S. gordonii* expressing on its surface the CTH522 molecule modulates the transcriptomic response of splenocytes stimulated *in vitro* with the CTH522 protein. Additionally, we identified a gene signature correlated with anti-CTH522 IgG levels. Lastly, we demonstrate that the boosting schedule, at three or six months after the priming, influences the immune modules activated upon antigen reencounter. Our analysis showed that vaginal colonization with recombinant *S. gordonii* induced persistent and noticeable changes in the transcriptomic response of *in vitro* stimulated splenocytes to the antigen.

METHODS

Mice

Seven-weeks old female BALB/C mice from Charles River (Lecco, Italy) were housed under specific pathogen-free conditions in the animal facility of the Laboratory of Molecular Microbiology and Biotechnology (LA.M.M.B.), Department of Medical Biotechnologies at University of Siena, and treated according to national guidelines (Decreto Legislativo 26/2014). Experiments were planned and conducted utilizing the three R's principles (Reduce, Replace and Refine), which included environmental enrichment and nesting, veterinary oversight, numbers reflecting statistical significance, and the use of anesthesia followed by cervical dislocation for the sacrifice. All animal studies were approved by the Ethics Committee "Comitato Etico Locale dell'Azienda Ospedaliera Universitaria Senese" and the Italian Ministry of Health (number 1004/2015-PR on September 22, 2015).

Experimental Design

Mice were intravaginally (IVAG) primed three times on weeks 0, 1 and 2 with either Wild-Type (GP1295) or recombinant (FR368) *S. gordonii*. Three or six-months after the priming, mice were subcutaneously boosted with 5µg of purified unadjuvanted CTH522 protein. The transcriptomic response was characterized 10 days after boosting in *in vitro* stimulated splenocytes, while the induction of CTH522- specific IgG serum response was evaluated at the same time point on serum samples.

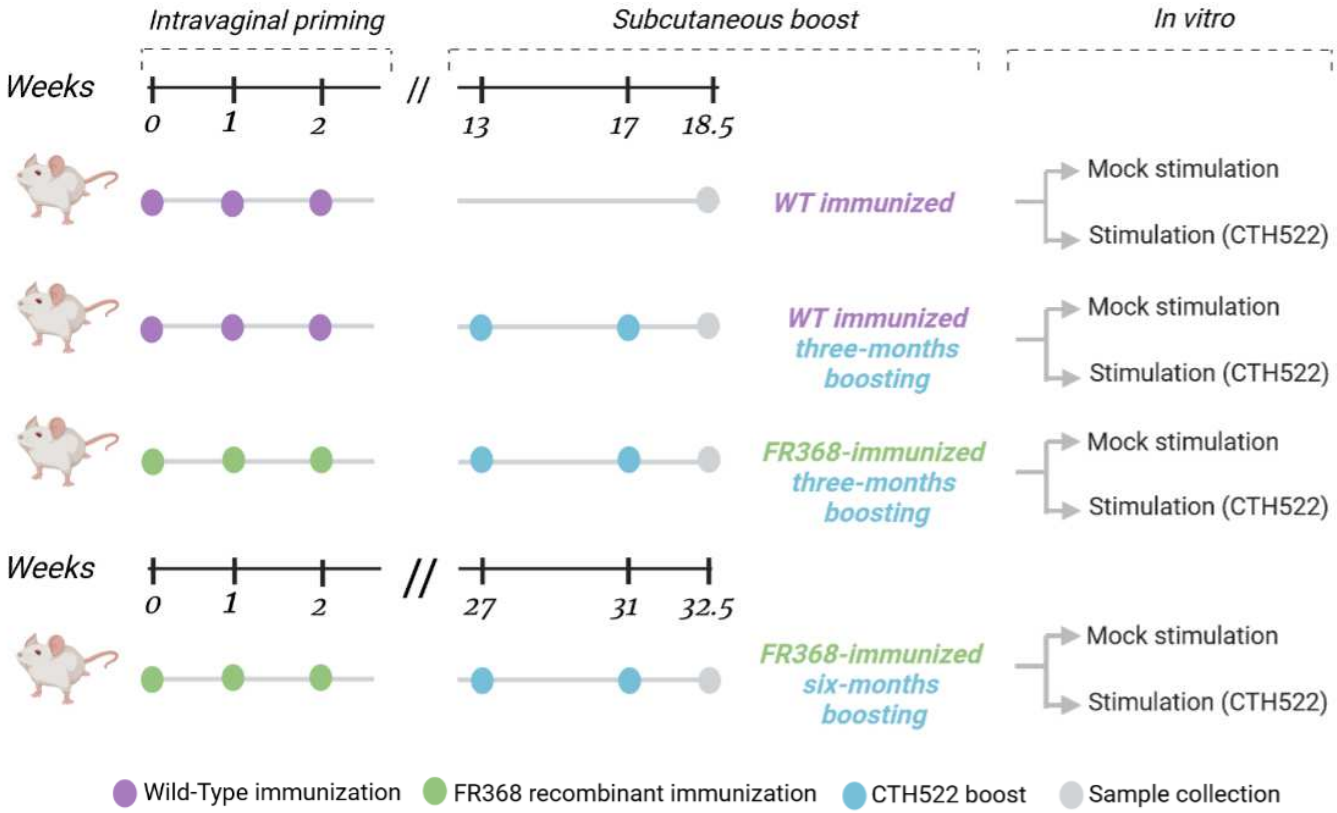


Figure 1. Experimental Design. Groups of 6 mice were intravaginally primed with Wild-Type (WT) or recombinant (FR368) *S. gordonii*. Three primings were performed on weeks 0, 1 and 2. For the boosted groups, subcutaneous administrations of the CTH522 protein were performed on weeks 13 and 17 for the three-months boosting schedule and on weeks 27 and 31 for the six-months boosting schedule. Spleens and blood were collected 10 days after the final boost and splenocytes were seeded in the presence (Stimulation) or absence (Mock-stimulation) of the purified CTH522. Blood samples were used for the assessment of IgG response by ELISA.

Immunizations

Following estrous cycle synchronization with subcutaneously delivered 0.1 mg of β-estradiol 17-valerate (#E1631, Sigma-Aldrich) resuspended in ethanol and diluted in olive oil, mice were immunized three times on weeks 0, 1, and 2 by the intravaginal route (IVAG) with 10^9 CFU in a volume of 20µL PBS of either Wild-Type (GP1295) or recombinant (FR368) *S. gordonii* bacterial vector expressing the vaccine antigen CTH522 (Philosof et al. 2022, unpublished). On weeks 13 and 17 (three-months boosting) or 27 and 31 (six-months boosting), mice were subcutaneously boosted with either CTH522 protein (5µg/mouse)

administered in a volume of 100µl/mouse in NaCl 0.9% (Fresenius Kabi, Italy) or Saline (NaCl 0.9%). Mice were sacrificed ten days after the second boost (week 18.5 or 32.5).

Sample collection and cell preparation

Blood samples were taken from individual mice by cardiac puncture at day ten post second boost upon sacrifice. Samples were incubated for 30 min at room temperature and then centrifuged at 1,200 x *g* for 10 min. Sera were collected and stored at -20°C until analysis by ELISA. Spleens were mashed onto 70 µm nylon screens (Sefar Italia, Italy) and washed two times in RPMI medium (#BE12-167F, Lonza, Belgium) supplemented with 100 U/ml penicillin/streptomycin (#P0781, Sigma-Aldrich) and 10% fetal bovine serum (#10082, Gibco, USA). Samples were treated with red blood cells lysis buffer according to manufacturer instruction (#00-4300-54, eBioscience, USA) and quantified.

ISABELLE FRANCO MOSCARDINI
17/09/2023 14:30:17

ELISA Isabelle Franco moscardini

Serum CTH522-specific IgG levels were determined by enzyme-linked immunosorbent assay (ELISA). Flat bottomed Maxisorp microtitre plates (Nunc, Denmark) were coated with CTH522 (1 µg/ml) overnight at 4°C in a volume of 100µl/well. Plates were washed and blocked with 200µl/well of PBS containing 1% BSA (Sigma-Aldrich) for 2 hours at 37°C. Serum samples were added and titrated in three to five-fold dilutions in 100 µl/well diluent buffer. After incubation for 2 hours at 37°C samples were washed and incubated with the Alkaline-Phosphatase-conjugated goat anti-mouse IgGs (IgG1 #1070-04, IgG2a #1080-04, IgG2b #1090-04, Anti-IgG3 #1100-04 diluted 1:1,500 for total IgG in 100 µl/well and developed by adding 200µl/well of 1 mg/ml AP substrate (#P5994, Sigma-Aldrich). The optical density was recorded using Multiskan FC Microplate Photometer (Thermo Scientific). Positive controls were included in all assays as follows: anti-IgG coating (1:1000, #1010-01),

IgG standards (#0107-01), anti-MOMP rabbit serum (SSI, Denmark) and anti-rabbit IgG-AP (#4050-04).

Splenocytes *in vitro* stimulation

10^6 splenocytes from each mouse were suspended in 100 µl of cRPMI and plated in a 96-well plate in 5 replicates per mouse per condition. For stimulation, 100 µl of 10 µg/mL CTH522 were added to each well, at a final concentration of 5 µg/mL. In mock samples 100 µl of cRPMI were added. Cells were incubated for 6 hours at 37°C with 5% CO_2. Replicates were pooled down together in a same-sample same-condition manner, centrifuged for 10 min at 500g 4°C and supernatant was discarded. Cell pellets were resuspended in 50 µl lysis buffer RA1 (#740955 NucleoSpin kit, Machery-Nagel) and flash frozen in liquid nitrogen. Samples were stored in -80°C until library preparation. RNA was extracted from frozen samples (#740955 NucleoSpin kit, Machery-Nagel), quantified using Qubit RNA quantification kit per manufacturer's instructions and Quality controlled using Agilent Bioanalyzer RNA 6000 nano kit (#5067-1511, Agilent) per manufacturer's instructions.

ISABELLE FRANCO MOSCARDINI
17/09/2023 14:30:48

Illumina sequencing

Isabelle Franco moscardini

Libraries were prepared using Stranded mRNA prep kit (Illumina, USA) according to manufacturer's instructions, using 60 ng of total RNA input per sample with dual indexing. Pooled libraries were sequenced on a single run of an Illumina NovaSeq 6000 instrument with 100 bp single end reads. Base calling was performed using Illumina's basespace FASTQ Generation pipeline. Basecalled reads were transferred on a local server and trimmed with Trimmomatic to remove low quality bases (Q treshold=20) at the beginning of the read and within the read using a sliding window size of 5 nucleotides with a required quality of 4.

Trimmed reads under 36 nucleotides in length were discarded. Trimmed reads were aligned to the mouse reference transcriptome using STAR and reads were counted using HTSeq.

Data Analysis

All the transcriptomics data analyses were carried out in R software, version 4.1.2 (2021-11-01), running under Windows 10 x64. Scripts can be found on Github at the following link: https://github.com/IsaMoscardini/VacPath/tree/main/Src.

Low expressed genes were filtered out and counts were normalized by log2CPM (40) to assess data variability by Principal Component Analysis (PCA), performed using the mixOmics package (41). Differential Expression Analysis was performed using the DESeq2 package (42). Genes with an adjusted p value less than 0.05 were considered Differentially Expressed (DEG). The *Wild-Type immunized no boosting Mock-stimulated* group was used as a baseline for the overall comparison between groups and the *Wild-type three-months boosted stimulated* group was used as baseline for the comparison with the *FR368-primed three- and six-months boosting stimulated* schedules.

Mouse Ensembl genes were converted into human gene Symbol (HGNC) by biomaRt (43,44) and enrichment analysis was carried out using the Blood Transcription Modules (BTM) database (45) and the CERNO test from the tmod package (46), using genes ranked by the adjusted p-value.

Correlation analysis was performed in log2CPM normalized data, after filtering low expressed genes. Gene expression values were correlated with the log2 of IgG titers using Spearman correlation. The Uniprot database (47), the web tool EnrichR (48,49), and the MSigDB gene set database (50,51) were used to access biological information about these genes.

RESULTS

Data variability

Principal Component Analysis (PCA) was performed to assess intra- and inter-group variability (Figure 2). The baseline group (WT saline mock-stimulated) presented a small intra-group variability, clustering together (in grey). The other immunization schedules presented a higher variability and samples had higher dispersion. The *in vitro* stimulation led samples to spread towards the lower part of the graph, driven by genes negatively correlated to the second component, including Irf8, Socs3, Stat3, Socs1 and Il21 (data not shown). No outliers were detected.

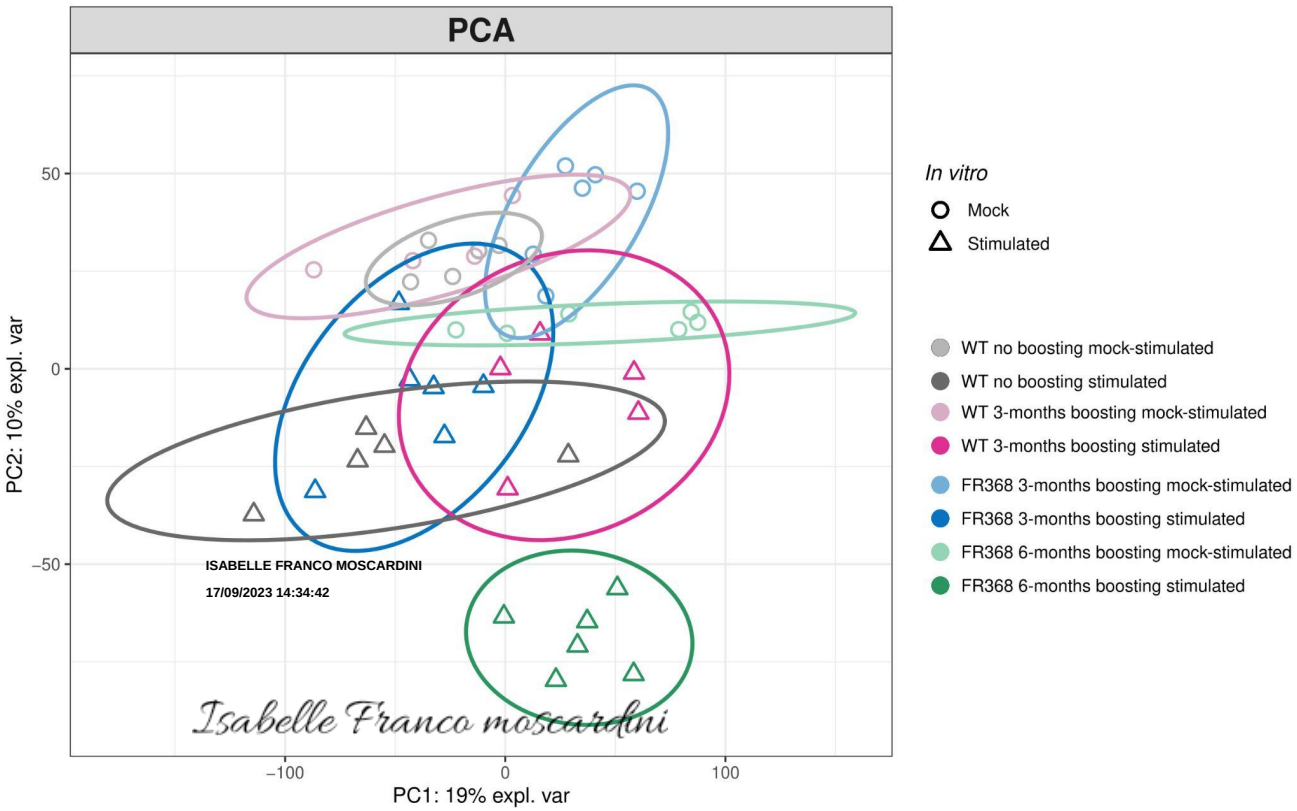


Figure 2. Principal Component Analysis. To evaluate gene expression data distribution and the presence of possible outliers, PCA analysis was performed by mixOmics package, using the normalized expression values. 19% of the total variance is explained by the first component and 10% by the second component, which seems important for the distinction between stimulated and mock-stimulated samples.

***In vitro* stimulation with the purified CTH522 generates a strong signal and provides insight into the host's response to different immunization schemes.**

To understand the transcriptomic changes driven by each *in vivo* immunization scheme and by the *in vitro* stimulation process, we performed a Differential Expression (DE) analysis with the DESeq2 package, using the *Wild-Type no boosting mock-stimulated* group as baseline. The total number of differentially expressed genes (DEGs) for each group is provided in table 1.

Immunization Schedule	Intravaginal priming	Boost (time after priming)	In vitro Stimulation	Up-regulated genes	Down-regulated genes
WT no boosting stimulated	Wild-Type	No	Yes	379	308
WT 3-months boosting mock-stimulated	Wild-Type	Yes (3 months)	No	0	0
WT 3-months boosting stimulated	Wild-Type	Yes (3 months)	Yes	345	185
FR368 3-months boosting mock-stimulated	Recombinant FR368	Yes (3 months)	No	131	82
FR368 3-months boosting stimulated	Recombinant FR368	Yes (3 months)	Yes	243	64
FR368 6-months boosting mock-stimulated	Recombinant FR368	Yes (6 months)	No	782	352
FR368 6-months boosting stimulated	Recombinant FR368	Yes (6 months)	Yes	1387	1284

Table 1. The number of Differentially Expressed Genes (DEGs) for each immunization schedule. Differential Expression Analysis was performed using the DESeq2 package and genes with an adjusted p-value of less than 0.05 were considered differentially expressed. The WT no boosting mock-stimulated group was used as baseline.

The *in-vitro* stimulation process leads to a transcriptomic perturbation that is slightly higher in samples that never encountered the antigen before (WT no boosting stimulated) compared to the WT and FR368 three-months boosting schedules. The exception is the FR368 six-months boosting schedule, which presents a higher number of differentially expressed genes compared to the other groups. These DEGs are mainly related to metabolism, especially glucose metabolism, and in a smaller number to the interferon pathway (data not shown).

Following the DE analysis, Gene Set analysis was performed using a Functional Class Scoring method and the Blood Transcription Modules (BTM) database. A strong signal was generated in *in vitro* stimulated samples, regardless of the immunization scheme, characterized by the up-regulation of DNA and viral sensing, innate antiviral pathways, and

Interferon and chemokines responses (Figure 3). Although these modules were shared, differences between groups were observed. For example, the activation of the *proinflammatory cytokines and chemokines* module was stronger in the FR368-immunized mice subjected to the three-months boosting schedule, owing to a higher fold-change of genes like TNF, CCL3 and CCRL2. On the other hand, the *signaling in T cells (I)* module was stronger in the FR368-immunized mice in the six-months boosting schedule, driven by genes differentially expressed only in this condition, such as IFNG, EGR1, JUNB, PRF1 and TNFRSF9.

In addition to the observed common signature, the stimulation process allowed the identification of specific modules enriched only in the FR368-immunized groups. Both three- and six-months boosting schedules activated modules linked to T cells, cell cycle, putative targets of PAX3 and the AP-1 transcription factor network. Specific DEGs for these two immunization schedules include CD83, ATF3, LEF1 and CCL3. Moreover, modules related to dendritic cells and monocytes were specifically activated in the three-months boosting schedule, by unique DEGs such as TNFRSF1B, GRINA, SLC7A11 and IL36G. On the other hand, the *cytokines − receptors cluster* and different T cell modules were unique to the six-months boosting schedule, enriched by DEGs specific for this immunization schedule, like CSF2, GZMA, GATA3, IL12RB1 and NLRC3.

The mock-stimulated groups did not bring relevant biological information, highlighting the importance of the *in vitro* stimulation process for assessing responses through DE and Gene Set Analyses.

Figure 3. Gene Set Enrichment Analysis. The significance of the activation of each Blood Transcription Module for each group was assessed through a multivariate enrichment analysis. A

strong antiviral response is detected in the *in vitro* stimulated groups, while specific signatures distinguish different immunization schedules, such as the FR368-primed groups.

Key transcriptional differences in response to *in vitro* stimulation in splenocytes from recombinant or WT-immunized mice

To further explore the transcriptional differences driven by the priming with the recombinant *S. gordonii*, a distinct DE analysis was carried out comparing *in vitro* stimulated samples from the FR368-immunized three-months and six-months boosting schedules to the WT-immunized group. In total, 46 differentially expressed genes were found in common between the two FR368-immunized groups, 16 downregulated, 13 upregulated and 17 genes presenting different directions in each group.

Genes that belong to the IL-1 and IL-2 signaling network, or that are activated by these cytokines, were up-regulated in both schedules, including Il1b, Mapkapk2, Nampt, Peli1, Plcd3, Psmb1, Pten and Stk17b. Moreover, the Neutrophil cytosol factor 1 (Ncf1) is described as coexpressed with Il1b and it was also found to be up-regulated.

In fact, after *in vitro* stimulation, all groups showed an increase in the Il1b gene when compared to *WT no boosting mock-stimulated* samples, as seen in the first DE analysis. This increase is not only much more pronounced in FR368-primed animals, but also followed by the increase in other genes related to these pathways,

Correlation between gene expression values and serum IgG titers highlights genes and biological pathways possibly linked to increased antibody production.

To investigate genes whose expression could be linked to the antibody response, normalized gene expression values were correlated with the IgG titers. Since the *in vitro* stimulation led to a strong effect in the gene expression observed in the Gene Set analysis, the correlation with the IgG titers was performed with the gene expression profile of the 17 mock-stimulated samples that had *in vivo* contact with the antigen (WT three-months

boosting, FR368 three-months boosting and FR368 six-months boosting). These samples, obtained at the same time point as the measurements of antibody titers, represent a more biologically relevant sample than the stimulated ones for the correlation with serum IgG titers.

The expression of 553 genes was found significantly correlated (p-value < 0.05) with the log2 of the IgG titers in the serum. Looking at the biological function, these genes enriched for pathways related to Myc targets V1 (cell proliferation pathway), interferon alpha response, mTORC1 signaling and IL2/STAT5 signaling, activated mainly by genes negatively correlated with IgG titers (Figure 4A). Many of the significantly correlated genes are linked to the interferon pathway, especially type I, as represented in the network of Figure 4B. Genes such as Azi2, Trim25, Pum2 and Eftud2 encode proteins that activate this pathway and were negatively correlated with the total IgG titers, as well as Jak1, a kinase that plays a major role in the interferon signal transduction. On the other hand, genes like Yipf2, which inhibits the cGAS-STING signaling, were positively correlated with the IgG titers, suggesting that indeed the interferon signaling pathway may be less expressed in samples with higher titers.

Figure 5A displays the barplot indicating the log2 of the IgG titers for each sample, colored by immunization groups and a heatmap with z-score of normalized expression values for the 50 genes most significantly correlated with IgG titers. Among the top negatively correlated features, there are genes involved in the interferon type I network, such as Azi2, Phb2 and Eftud2, but also Anp32b, that might participate in the regulation of adaptive immune responses. On the other hand, among the top positive correlated genes, there are Ccl3 (an important chemoattractant), Il18bp (the encoder of the natural antagonist of IL-18) and Pi4kb (involved in Golgi-to-plasma membrane trafficking).

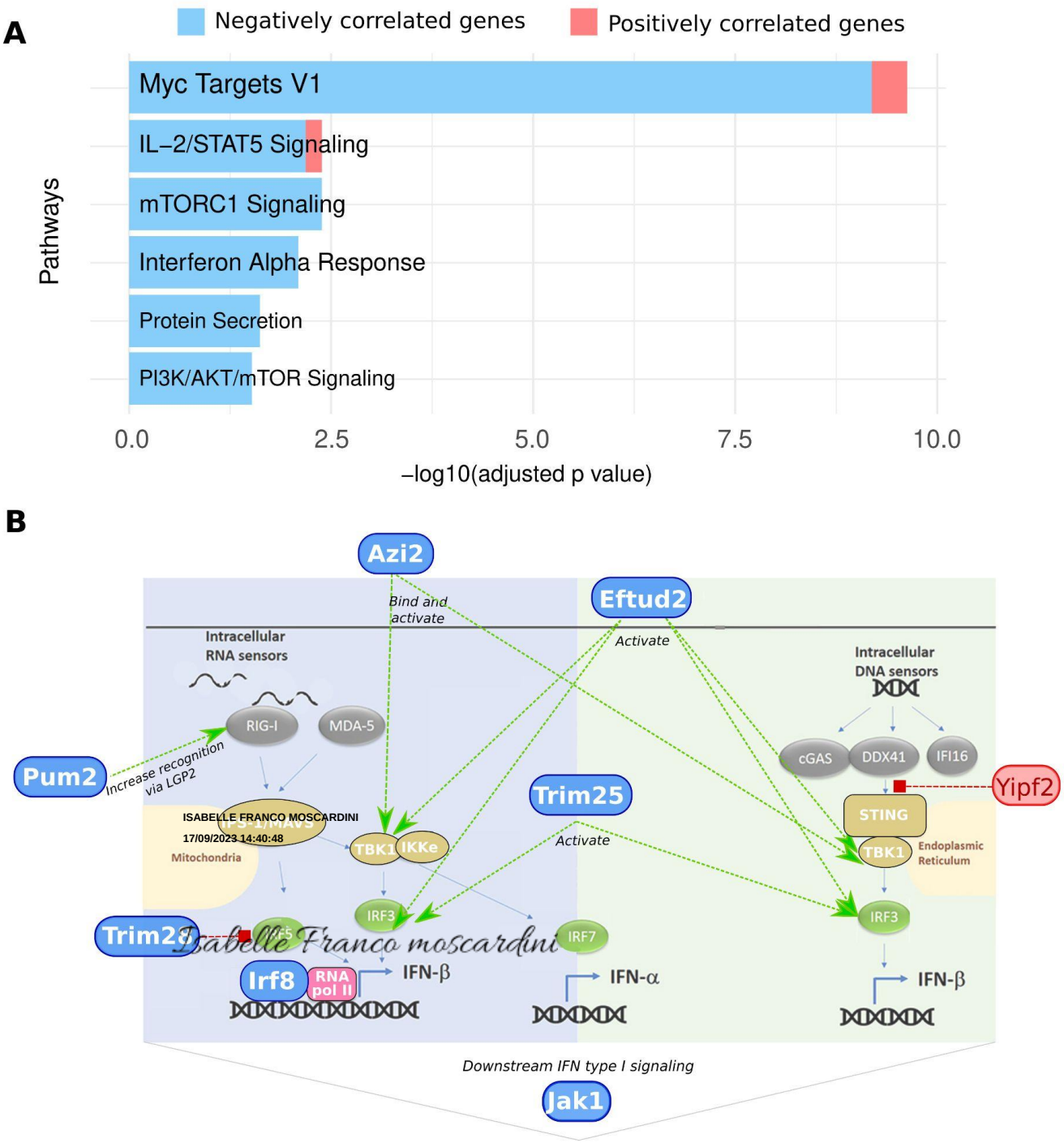


Figure 4. Correlation of gene expression and serum IgG titers. **A.** Gene set analysis of the 553 genes significantly correlated with serum IgG levels. Performed with enrichR web tool using the MSigDB Hallmark 2020 database. **B.** Interferon type I activation network. The 7 genes found negatively correlated with the IgG titers are represented in blue while the gene positively correlated with the IgG titers is represented in red. Green arrows indicate activation of the protein/pathway, while the red arrow represents the inhibition of the protein/pathway. The figure was adapted from Jefferies CA, 2019 (52). The link between the genes and the Interferon pathways was found in previously published works (52–60).

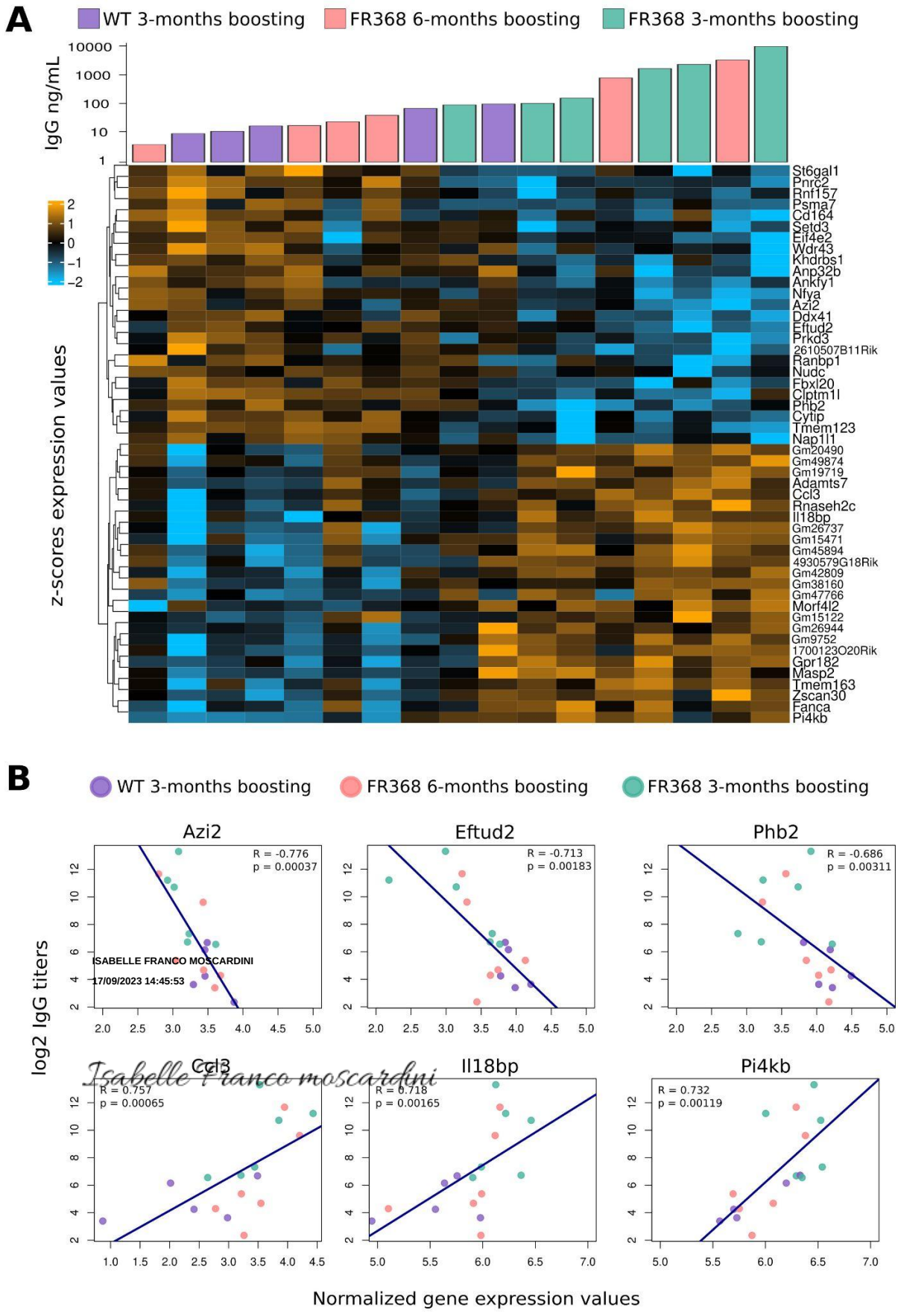


Figure 5. Top 50 genes correlated with serum IgG titers. **A.** Barplot displaying the IgG titers (ng/mL) in the serum, bars are colored by immunization schedule, followed by the heatmap representing the z-score of the normalized expression values of the 50 genes with the lowest p-values for the correlation with IgG. **B.** Dot plots for 6 of the top correlated genes, displaying the normalized expression value on the x-axis and the log2 of the IgG titers on the y-axis. Dots are colored by immunization schedule, as the bars in panel A.

DISCUSSION

In this work, we leverage transcriptomic analysis to study the host's response to different immunization schedules. Mice were intravaginally primed with either Wild-Type or a recombinant *S. gordonii* expressing the CTH522 protein, a multivalent chlamydial antigen containing regions of MOMP from *C. trachomatis* serovars D, E, F and G found to be safe and immunogenic in a phase I clinical trial (39), and then received two subcutaneous boosts of the unadjuvanted CTH522 protein. Splenocytes were collected 10 days after the final boost and seeded in the presence or absence of the CTH522 protein. The *in vitro* stimulation process is important to characterize the response generated by a local infection (61). In the present study, we took advantage of this method to assess the systemic response generated by the intravaginal priming and the subsequent subcutaneous boosts.

The stimulation process allowed us to retrieve information on the differences between intravaginally priming mice with the wild-type (WT) or with the recombinant bacteria expressing the *C. trachomatis* antigen (FR368). In the enrichment analysis, compared to the baseline group, the T cell signaling pathway was significantly activated only in FR368-primed samples, as well as the AP-1 transcription factor network module. The AP-1 transcription factor network is involved in different bacterial and viral infections (62,63), including in *Chlamydia trachomatis* (64) and *Chlamydia pneumoniae* (65) infections, in the latter case being a key factor of the host's response, regulating inflammatory mediators like IL6, IL8, and IFN.

Differences between the FR368-primed and the WT-primed groups after stimulation have also emphasized the role of the IL-1 and IL-2 signaling pathways in the response generated by different priming schedules. IL-1 is commonly mentioned for its roles in innate immunity and inflammation, however, this cytokine plays essential roles in the bridging of adaptive responses (66). IL-1 has been shown to improve the differentiation of naive T cells through the regulation of DC activation (67–69) and to enhance the persistence and response of memory cells when administered together with the antigen (70). In addition, IL-1 is known for its role in favoring the differentiation of CD4+ T cells during priming, acting as a driver of Th17 responses - a key mediator of mucosal immunity (71). Therefore, the increase in this pathway after stimulation of FR368-primed samples could indicate an important mechanism of this immunization schedule.

The Interferon type I signaling network was negatively correlated with the IgG titers in the mock-stimulated samples. However, whether the antigen encounter *in vivo* upon subcutaneous boosting leads to an activation of the interferon pathways, as seen in the *in vitro* stimulated samples, requires further investigations. Since our samples were collected 10 days after the final boost, the detected transcriptomic responses are probably related to downstream processes occurring as a consequence of the boosting and may not be tightly linked to the response to the CTH522 protein itself.

The Ccl3 gene was differentially expressed after stimulation only in FR368-primed samples, both in three- and six-months boosting schedules. Interestingly, this gene was also positively correlated with the serum IgG. Ccl3, also known as macrophage inflammatory protein 1 alpha (MIP-1α), is an important chemoattractant secreted by various cells, including fibroblasts, epithelial cells, lymphocytes, resident and recruited monocytes, and macrophages (72,73). Besides acting as an attractant for immune cells like monocytes (74,75) and natural killer cells (76), CCL3 has a well described role in the migration of dendritic cells (77–79) and lymphocytes (80).

In a study examining the immunogenicity of an adenovirus-based vaccine vector, co-expression of CCL3 with the retroviral antigens increased vaccine protection from infection by enhancing neutralizing antibody titers and virus-specific CD4+ T cell responses (81). Indeed, previous studies suggest that CCL3 can enhance humoral and cellular responses in both mucosal and systemic immunity. A study comparing the nasal administration of Chicken egg albumin (OVA) in the presence or absence of CCL3 found enhanced systemic antibody responses marked by higher levels of IgM and all the IgG subtypes. The CD4+ T cells in Peyer patches, cervical lymph nodes and spleens of mice immunized in the presence of CCL3 exhibited marked increases in OVA-specific proliferative responses. Moreover, CCL3 promoted mucosal and systemic CD8+ CTL responses (82).

It has been suggested that CCL3 is an important cytokine for sustaining and amplifying a previously primed T-cell response (83). Our data suggest that intravaginal immunization with the recombinant bacteria expressing the CTH522 antigen induces Ccl3 expression upon antigen reencounter through *in vitro* stimulation.

Taken together, our data suggest that the *in vivo* site and context of antigen encounter modulate the transcriptomic signature of *in vitro* stimulated splenocytes. We demonstrated a differential activation of inflammatory pathways genes, which was associated with higher systemic antibody response. Moreover, we have shown that Ccl3 is a marker of the recall responses in mice primed with the recombinant *S. gordonii,* and it is associated with the IgG titers after immunization, being a possible biomarker of vaccine response.

References

Abraham, S., Juel, H.B., Bang, P., Cheeseman, H.M., Dohn, R.B., Cole, T., Kristiansen, M.P., Korsholm, K.S., Lewis, D., Olsen, A.W., McFarlane, L.R., Day, S., Knudsen, S., Moen, K., Ruhwald, M., Kromann, I., Andersen, P., Shattock, R.J., Follmann, F., 2019. Safety and immunogenicity of the chlamydia vaccine candidate CTH522 adjuvanted with CAF01 liposomes or aluminium hydroxide: a first-in-human, randomised, double-blind, placebo-controlled, phase 1 trial. Lancet Infect. Dis. 19, 1091–1100. https://doi.org/10.1016/S1473-3099(19)30279-8

Beninati, C., Oggioni, M.R., Boccanera, M., Spinosa, M.R., Maggi, T., Conti, S., Magliani, W., De Bernardis, F., Teti, G., Cassone, A., Pozzi, G., Polonelli, L., 2000. Therapy of mucosal candidiasis by expression of an anti-idiotype in human commensal bacteria. Nat. Biotechnol. 18, 1060–1064. https://doi.org/10.1038/80250

Ben-Sasson, S.Z., Hu-Li, J., Quiel, J., Cauchetaux, S., Ratner, M., Shapira, I., Dinarello, C.A., Paul, W.E., 2009. IL-1 acts directly on CD4 T cells to enhance their antigen-driven expansion and differentiation. Proc. Natl. Acad. Sci. 106, 7119–7124. https://doi.org/10.1073/pnas.0902745106

Castanier, C., Zemirli, N., Portier, A., Garcin, D., Bidère, N., Vazquez, A., Arnoult, D., 2012. MAVS ubiquitination by the E3 ligase TRIM25 and degradation by the proteasome is involved in type I interferon production after activation of the antiviral RIG-I-like receptors. BMC Biol. 10, 44. https://doi.org/10.1186/1741-7007-10-44

Chen, E.Y., Tan, C.M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G.V., Clark, N.R., Ma'ayan, A., 2013. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics 14, 128. https://doi.org/10.1186/1471-2105-14-128

Ciabattini, A., Giomarelli, B., Parigi, R., Chiavolini, D., Pettini, E., Aricò, B., Giuliani, M.M., Santini, L., Medaglini, D., Pozzi, G., 2008a. Intranasal immunization of mice with recombinant Streptococcus gordonii expressing NadA of Neisseria meningitidis induces systemic bactericidal antibodies and local IgA. Vaccine 26, 4244–4250. https://doi.org/10.1016/j.vaccine.2008.05.049

Ciabattini, A., Pettini, E., Andersen, P., Pozzi, G., Medaglini, D., 2008b. Primary Activation of Antigen-Specific Naive CD4 $^{+}$ and CD8 $^{+}$ T Cells following Intranasal Vaccination with Recombinant Bacteria. Infect. Immun. 76, 5817–5825. https://doi.org/10.1128/IAI.00793-08

Ciabattini, A., Pettini, E., Arsenijevic, S., Pozzi, G., Medaglini, D., 2010. Intranasal immunization with vaccine vector Streptococcus gordonii elicits primed CD4+ and CD8+ T cells in the genital and intestinal tracts. Vaccine 28, 1226–1233. https://doi.org/10.1016/j.vaccine.2009.11.021

Çuburu, N., Graham, B.S., Buck, C.B., Kines, R.C., Pang, Y.-Y.S., Day, P.M., Lowy, D.R., Schiller, J.T., 2012. Intravaginal immunization with HPV vectors induces tissue-resident CD8+ T cell responses. J. Clin. Invest. 122, 4606–4620. https://doi.org/10.1172/JCI63287

Culley, F.J., Pennycook, A.M.J., Tregoning, J.S., Hussell, T., Openshaw, P.J.M., 2006. Differential Chemokine Expression following Respiratory Virus Infection Reflects Th1- or Th2-Biased Immunopathology. J. Virol. 80, 4521–4527. https://doi.org/10.1128/JVI.80.9.4521-4527.2006

Czerkinsky, C., Holmgren, J., 2010. Topical immunization strategies. Mucosal Immunol. 3, 545–555. https://doi.org/10.1038/mi.2010.55

Danforth, J.M., Strieter, R.M., Kunkel, S.L., Arenberg, D.A., VanOtteren, G.M., Standiford, T.J., 1995. Macrophage Inflammatory Protein-1α Expression in Vivo and in Vitro: The Role of Lipoteichoic Acid. Clin. Immunol. Immunopathol. 74, 77–83. https://doi.org/10.1006/clin.1995.1011

Dey, N., Liu, T., Garofalo, R.P., Casola, A., 2011. TAK1 regulates NF-ΚB and AP-1 activation in airway epithelial cells following RSV infection. Virology 418, 93–101. https://doi.org/10.1016/j.virol.2011.07.007

Difabio, S., 1998. Vaginal immunization of Cynomolgus monkeys with Streptococcus gordonii expressing HIV-1 and HPV 16 antigens. Vaccine 16, 485–492.

https://doi.org/10.1016/S0264-410X(97)80002-3
Domingos-Pereira, S., Derré, L., Warpelin-Decrausaz, L., Haefliger, J.-A., Romero, P., Jichlinski, P., Nardelli-Haefliger, D., 2014. Intravaginal and Subcutaneous Immunization Induced Vaccine Specific CD8 T Cells and Tumor Regression in the Bladder. J. Urol. 191, 814–822. https://doi.org/10.1016/j.juro.2013.08.009
Driggers, P.H., Ennist, D.L., Gleason, S.L., Mak, W.H., Marks, M.S., Levi, B.Z., Flanagan, J.R., Appella, E., Ozato, K., 1990. An interferon gamma-regulated protein that binds the interferon-inducible enhancer element of major histocompatibility complex class I genes. Proc. Natl. Acad. Sci. 87, 3743–3747. https://doi.org/10.1073/pnas.87.10.3743
Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., Huber, W., 2005. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinforma. Oxf. Engl. 21, 3439–3440. https://doi.org/10.1093/bioinformatics/bti525
Durinck, S., Spellman, P.T., Birney, E., Huber, W., 2009. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat. Protoc. 4, 1184–1191. https://doi.org/10.1038/nprot.2009.97
Falcone, V., Mihm, D., Neumann-Haefelin, D., Costa, C., Nguyen, T., Pozzi, G., Ricci, S., 2006. Systemic and mucosal immunity to respiratory syncytial virus induced by recombinant *Streptococcus gordonii* surface-displaying a domain of viral glycoprotein G. FEMS Immunol. Med. Microbiol. 48, 116–122. https://doi.org/10.1111/j.1574-695X.2006.00130.x
Gjertsson, I., Hultgren, O.H., Collins, L.V., Pettersson, S., Tarkowski, A., 2001. Impact of transcription factors AP-1 and NF-κB on the outcome of experimental Staphylococcus aureus arthritis and sepsis. Microbes Infect. 3, 527–534. https://doi.org/10.1016/S1286-4579(01)01408-3
Guo, Z., Zhang, M., An, H., Chen, W., Liu, S., Guo, J., Yu, Y., Cao, X., 2003. Fas ligation induces IL-1β-dependent maturation and IL-1β-independent survival of dendritic cells: different roles of ERK and NF-κB signaling pathways. Blood 102, 4441–4447. https://doi.org/10.1182/blood-2002-11-3420
Jefferies, C.A., 2019. Regulating IRFs in IFN Driven Disease. Front. Immunol. 10, 325. https://doi.org/10.3389/fimmu.2019.00325
Kasturi, S.P., Kozlowski, P.A., Nakaya, H.I., Burger, M.C., Russo, P., Pham, M., Kovalenkov, Y., Silveira, E.L.V., Havenar-Daughton, C., Burton, S.L., Kilgore, K.M., Johnson, M.J., Nabi, R., Legere, T., Sher, Z.J., Chen, X., Amara, R.R., Hunter, E., Bosinger, S.E., Spearman, P., Crotty, S., Villinger, F., Derdeyn, C.A., Wrammert, J., Pulendran, B., 2017. Adjuvanting a Simian Immunodeficiency Virus Vaccine with Toll-Like Receptor Ligands Encapsulated in Nanoparticles Induces Persistent Antibody Responses and Enhanced Protection in TRIM5α Restrictive Macaques. J. Virol. 91, e01844-16. https://doi.org/10.1128/JVI.01844-16
Kaushal, D., Foreman, T.W., Gautam, U.S., Alvarez, X., Adekambi, T., Rangel-Moreno, J., Golden, N.A., Johnson, A. M.F., Phillips, B.L., Ahsan, M.H., Russell-Lodrigue, K.E., Doyle, L.A., Roy, C.J., Didier, P.J., Blanchard, J.L., Rengarajan, J., Lackner, A.A., Khader, S.A., Mehra, S., 2015. Mucosal vaccination with attenuated Mycobacterium tuberculosis induces strong central memory responses and protects against tuberculosis. Nat. Commun. 6, 8533. https://doi.org/10.1038/ncomms9533
Kim-Anh Le Cao, F.R., 2018. mixOmics. https://doi.org/10.18129/B9.BIOC.MIXOMICS
Kotloff, K.L., Wasserman, S.S., Jones, K.F., Livio, S., Hruby, D.E., Franke, C.A., Fischetti, V.A., 2005. Clinical and Microbiological Responses of Volunteers to Combined Intranasal and Oral Inoculation with a *Streptococcus gordonii* Carrier Strain Intended for Future Use as a Group A Streptococcus Vaccine. Infect. Immun. 73, 2360–2366. https://doi.org/10.1128/IAI.73.4.2360-2366.2005
Kozlowski, P.A., Aldovini, A., 2019. Mucosal Vaccine Approaches for Prevention of HIV and SIV Transmission. Curr. Immunol. Rev. 15, 102–122. https://doi.org/10.2174/1573395514666180605092054
Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins,

S.L., Jagodnik, K.M., Lachmann, A., McDermott, M.G., Monteiro, C.D., Gundersen, G.W., Ma'ayan, A., 2016. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 44, W90–W97. https://doi.org/10.1093/nar/gkw377

Kwant, A., Rosenthal, K.L., 2004. Intravaginal immunization with viral subunit protein plus CpG oligodeoxynucleotides induces protective immunity against HSV-2. Vaccine 22, 3098–3104. https://doi.org/10.1016/j.vaccine.2004.01.059

Letvin, N.L., Rao, S.S., Montefiori, D.C., Seaman, M.S., Sun, Y., Lim, S.-Y., Yeh, W.W., Asmal, M., Gelman, R.S., Shen, L., Whitney, J.B., Seoighe, C., Lacerda, M., Keating, S., Norris, P.J., Hudgens, M.G., Gilbert, P.B., Buzby, A.P., Mach, L.V., Zhang, J., Balachandran, H., Shaw, G.M., Schmidt, S.D., Todd, J.-P., Dodson, A., Mascola, J.R., Nabel, G.J., 2011. Immune and Genetic Correlates of Vaccine Protection Against Mucosal Infection by SIV in Monkeys. Sci. Transl. Med. 3. https://doi.org/10.1126/scitranslmed.3002351

Li, S., Rouphael, N., Duraisingham, S., Romero-Steiner, S., Presnell, S., Davis, C., Schmidt, D.S., Johnson, S.E., Milton, A., Rajam, G., Kasturi, S., Carlone, G.M., Quinn, C., Chaussabel, D., Palucka, A.K., Mulligan, M.J., Ahmed, R., Stephens, D.S., Nakaya, H.I., Pulendran, B., 2014. Molecular signatures of antibody responses derived from a systems biology study of five human vaccines. Nat. Immunol. 15, 195–204. https://doi.org/10.1038/ni.2789

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., Tamayo, P., 2015. The Molecular Signatures Database Hallmark Gene Set Collection. Cell Syst. 1, 417–425. https://doi.org/10.1016/j.cels.2015.12.004

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdottir, H., Tamayo, P., Mesirov, J.P., 2011. Molecular signatures database (MSigDB) 3.0. Bioinformatics 27, 1739–1740. https://doi.org/10.1093/bioinformatics/btr260

Lietz, R., Bayer, W., Ontikatze, T., Johrden, L., Tenbusch, M., Storcksdieck genannt Bonsmann, M., Überla, K., Dittmer, U., Wildner, O., 2012. Codelivery of the Chemokine CCL3 by an Adenovirus-Based Vaccine Improves Protection from Retrovirus Infection. J. Virol. 86, 1706–1716. https://doi.org/10.1128/JVI.06244-11

Lillard, J.W., Singh, U.P., Boyaka, P.N., Singh, S., Taub, D.D., McGhee, J.R., 2003. MIP-1α and MIP-1β differentially mediate mucosal and systemic adaptive immunity. Blood 101, 807–814. https://doi.org/10.1182/blood-2002-07-2305

Lindell, D.M., Standiford, T.J., Mancuso, P., Leshen, Z.J., Huffnagle, G.B., 2001. Macrophage Inflammatory Protein 1α/CCL3 Is Required for Clearance of an Acute *Klebsiella pneumoniae* Pulmonary Infection. Infect. Immun. 69, 6364–6369. https://doi.org/10.1128/IAI.69.10.6364-6369.2001

Logerot, S., Figueiredo-Morgado, S., Charmeteau-de-Muylder, B., Sandouk, A., Drillet-Dangeard, A.-S., Bomsel, M., Bourgault-Villada, I., Couëdel-Courteille, A., Cheynier, R., Rancez, M., 2021. IL-7-Adjuvanted Vaginal Vaccine Elicits Strong Mucosal Immune Responses in Non-Human Primates. Front. Immunol. 12, 614115. https://doi.org/10.3389/fimmu.2021.614115

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

Luci, C., Hervouet, C., Rousseau, D., Holmgren, J., Czerkinsky, C., Anjuère, F., 2006. Dendritic Cell-Mediated Induction of Mucosal Cytotoxic Responses following Intravaginal Immunization with the Nontoxic B Subunit of Cholera Toxin. J. Immunol. 176, 2749–2757. https://doi.org/10.4049/jimmunol.176.5.2749

Luft, T., Jefford, M., Luetjens, P., Hochrein, H., Masterman, K.-A., Maliszewski, C., Shortman, K., Cebon, J., Maraskovsky, E., 2002. IL-1β Enhances CD40 Ligand-Mediated Cytokine Secretion by Human Dendritic Cells (DC): A Mechanism for T Cell-Independent DC Activation. J. Immunol. 168, 713–722. https://doi.org/10.4049/jimmunol.168.2.713

Mantovani, A., Dinarello, C.A., Molgora, M., Garlanda, C., 2019. Interleukin-1 and Related Cytokines in the Regulation of Inflammation and Immunity. Immunity 50, 778–795.

https://doi.org/10.1016/j.immuni.2019.03.012
McKay, P.F., Mann, J.F.S., Pattani, A., Kett, V., Aldon, Y., King, D., Malcolm, R.K., Shattock, R.J., 2017. Intravaginal immunisation using a novel antigen-releasing ring device elicits robust vaccine antigen-specific systemic and mucosal humoral immune responses. J. Controlled Release 249, 74–83. https://doi.org/10.1016/j.jconrel.2017.01.018
Medaglini, D., Ciabattini, A., Cuppone, A.M., Costa, C., Ricci, S., Costalonga, M., Pozzi, G., 2006. In Vivo Activation of Naive CD4 $^{+}$ T Cells in Nasal Mucosa-Associated Lymphoid Tissue following Intranasal Immunization with Recombinant *Streptococcus gordonii*. Infect. Immun. 74, 2760–2766. https://doi.org/10.1128/IAI.74.5.2760-2766.2006
Medaglini, D., Ciabattini, A., Spinosa, M.R., Maggi, T., Marcotte, H., Oggioni, M.R., Pozzi, G., 2001. Immunization with recombinant Streptococcus gordonii expressing tetanus toxin fragment C confers protection from lethal challenge in mice. Vaccine 19, 1931–1939. https://doi.org/10.1016/S0264-410X(00)00434-5
Medaglini, D., Pozzi, G., King, T.P., Fischetti, V.A., 1995. Mucosal and systemic immune responses to a recombinant protein expressed on the surface of the oral commensal bacterium Streptococcus gordonii after oral colonization. Proc. Natl. Acad. Sci. 92, 6868–6872. https://doi.org/10.1073/pnas.92.15.6868
Mitchell, D.A., Batich, K.A., Gunn, M.D., Huang, M.-N., Sanchez-Perez, L., Nair, S.K., Congdon, K.L., Reap, E.A., Archer, G.E., Desjardins, A., Friedman, A.H., Friedman, H.S., Herndon II, J.E., Coan, A., McLendon, R.E., Reardon, D.A., Vredenburgh, J.J., Bigner, D.D., Sampson, J.H., 2015. Tetanus toxoid and CCL3 improve dendritic cell vaccines in mice and glioblastoma patients. Nature 519, 366–369. https://doi.org/10.1038/nature14320
Moscardini, I.F., Santoro, F., Carraro, M., Gerlini, A., Fiorino, F., Germoni, C., Gholami, S., Pettini, E., Medaglini, D., Iannelli, F., Pozzi, G., 2022. Immune Memory After Respiratory Infection With Streptococcus pneumoniae Is Revealed by in vitro Stimulation of Murine Splenocytes With Inactivated Pneumococcal Whole Cells: Evidence of Early Recall Responses by Transcriptomic Analysis. Front. Cell. Infect. Microbiol. 12, 869763. https://doi.org/10.3389/fcimb.2022.869763
Mottram, L., Lundgren, A., Svennerholm, A.-M., Leach, S., 2020. Booster vaccination with a fractional dose of an oral cholera vaccine induces comparable vaccine-specific antibody avidity as a full dose: A randomised clinical trial. Vaccine 38, 655–662. https://doi.org/10.1016/j.vaccine.2019.10.050
Nakaya, H.I., Wrammert, J., Lee, E.K., Racioppi, L., Marie-Kunze, S., Haining, W.N., Means, A.R., Kasturi, S.P., Khan, N., Li, G.-M., McCausland, M., Kanchan, V., Kokko, K.E., Li, S., Elbein, R., Mehta, A.K., Aderem, A., Subbarao, K., Ahmed, R., Pulendran, B., 2011. Systems biology of vaccination for seasonal influenza in humans. Nat. Immunol. 12, 786–795. https://doi.org/10.1038/ni.2067
Narita, R., Takahasi, K., Murakami, E., Hirano, E., Yamamoto, S.P., Yoneyama, M., Kato, H., Fujita, T., 2014. A Novel Function of Human Pumilio Proteins in Cytoplasmic Sensing of Viral Infection. PLoS Pathog. 10, e1004417. https://doi.org/10.1371/journal.ppat.1004417
Neote, K., DiGregorio, D., Mak, J.Y., Horuk, R., Schall, T.J., 1993. Molecular cloning, functional expression, and signaling characteristics of a C-C chemokine receptor. Cell 72, 415–425. https://doi.org/10.1016/0092-8674(93)90118-A
Nguyen, N.D.N.T., Olsen, A.W., Lorenzen, E., Andersen, P., Hvid, M., Follmann, F., Dietrich, J., 2020. Parenteral vaccination protects against transcervical infection with Chlamydia trachomatis and generate tissue-resident T cells post-challenge. Npj Vaccines 5, 7. https://doi.org/10.1038/s41541-020-0157-x
Oggioni, M.R., Medaglini, D., Romano, L., Peruzzi, F., Maggi, T., Lozzi, L., Bracci, L., Zazzi, M., Manca, F., Valensin, P.E., Pozzi, G., 1999. Antigenicity and Immunogenicity of the V3 Domain of HIV Type 1 Glycoprotein 120 Expressed on the Surface of Streptococcus gordonii. AIDS Res. Hum. Retroviruses 15, 451–459. https://doi.org/10.1089/088922299311204

Oggioni, M.R., Pozzi, G., 1996. A host-vector system for heterologous gene expression in Streptococcus gordonii. Gene 169, 85–90. https://doi.org/10.1016/0378-1119(95)00775-X
Oh, J.Z., Ravindran, R., Chassaing, B., Carvalho, F.A., Maddur, M.S., Bower, M., Hakimpour, P., Gill, K.P., Nakaya, H.I., Yarovinsky, F., Sartor, R.B., Gewirtz, A.T., Pulendran, B., 2014. TLR5-Mediated Sensing of Gut Microbiota Is Necessary for Antibody Responses to Seasonal Influenza Vaccination. Immunity 41, 478–492. https://doi.org/10.1016/j.immuni.2014.08.009
Olive, A.J., Haff, M.G., Emanuele, M.J., Sack, L.M., Barker, J.R., Elledge, S.J., Starnbach, M.N., 2014. Chlamydia trachomatis-Induced Alterations in the Host Cell Proteome Are Required for Intracellular Growth. Cell Host Microbe 15, 113–124. https://doi.org/10.1016/j.chom.2013.12.009
Olsen, A.W., Follmann, F., Erneholm, K., Rosenkrands, I., Andersen, P., 2015. Protection Against *Chlamydia trachomatis* Infection and Upper Genital Tract Pathological Changes by Vaccine-Promoted Neutralizing Antibodies Directed to the VD4 of the Major Outer Membrane Protein. J. Infect. Dis. 212, 978–989. https://doi.org/10.1093/infdis/jiv137
Pannaraj, P.S., da Costa-Martins, A.G., Cerini, C., Li, F., Wong, S.-S., Singh, Y., Urbanski, A.H., Gonzalez-Dias, P., Yang, J., Webby, R.J., Nakaya, H.I., Aldrovandi, G.M., 2022. Molecular alterations in human milk in simulated maternal nasal mucosal infection with live attenuated influenza vaccination. Mucosal Immunol. https://doi.org/10.1038/s41385-022-00537-4
Pettini, E., Prota, G., Ciabattini, A., Boianelli, A., Fiorino, F., Pozzi, G., Vicino, A., Medaglini, D., 2013. Vaginal Immunization to Elicit Primary T-Cell Activation and Dissemination. PLoS ONE 8, e80545. https://doi.org/10.1371/journal.pone.0080545
Pulendran, B., 2020. Systems Biological Approaches for Mucosal Vaccine Development, in: Mucosal Vaccines. Elsevier, pp. 753–772. https://doi.org/10.1016/B978-0-12-811924-2.00045-6
Querec, T.D., Akondy, R.S., Lee, E.K., Cao, W., Nakaya, H.I., Teuwen, D., Pirani, A., Gernert, K., Deng, J., Marzolf, B., Kennedy, K., Wu, H., Bennouna, S., Oluoch, H., Miller, J., Vencio, R.Z., Mulligan, M., Aderem, A., Ahmed, R., Pulendran, B., 2009. Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans. Nat. Immunol. 10, 116–125. https://doi.org/10.1038/ni.1688
Ran, Y., Xiong, M., Xu, Z., Luo, W., Wang, S., Wang, Y.-Y., 2019. YIPF5 Is Essential for Innate Immunity to DNA Virus and Facilitates COPII-Dependent STING Trafficking. J. Immunol. 203, 1560–1570. https://doi.org/10.4049/jimmunol.1900387
Rechtien, A., Richert, L., Lorenzo, H., Martrus, G., Hejblum, B., Dahlke, C., Kasonta, R., Zinser, M., Stubbe, H., Matschl, U., Lohse, A., Krähling, V., Eickmann, M., Becker, S., Thiébaut, R., Altfeld, M., Addo, M., Agnandji, S.T., Krishna, S., Kremsner, P.G., Brosnahan, J.S., Bejon, P., Njuguna, P., Addo, M.M., Becker, S., Krähling, V., Siegrist, C.-A., Huttner, A., Kieny, M.-P., Moorthy, V., Fast, P., Savarese, B., Lapujade, O., 2017. Systems Vaccinology Identifies an Early Innate Immune Signature as a Correlate of Antibody Responses to the Ebola Vaccine rVSV-ZEBOV. Cell Rep. 20, 2251–2261. https://doi.org/10.1016/j.celrep.2017.08.023

Rhoades, E.R., Cooper, A.M., Orme, I.M., 1995. Chemokine response in mice infected with Mycobacterium tuberculosis. Infect. Immun. 63, 3871–3877. https://doi.org/10.1128/iai.63.10.3871-3877.1995

Ricci, S., Medaglini, D., Rush, C.M., Marcello, A., Peppoloni, S., Manganelli, R., Palú, G., Pozzi, G., 2000. Immunogenicity of the B Monomer of *Escherichia coli* Heat-Labile Toxin Expressed on the Surface of *Streptococcus gordonii*. Infect. Immun. 68, 760–766. https://doi.org/10.1128/IAI.68.2.760-766.2000
Robinson, M.D., McCarthy, D.J., Smyth, G.K., 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140. https://doi.org/10.1093/bioinformatics/btp616
Ryzhakov, G., Randow, F., 2007. SINTBAD, a novel component of innate antiviral immunity, shares a TBK1-binding domain with NAP1 and TANK. EMBO J. 26, 3180–3190. https://doi.org/10.1038/sj.emboj.7601743

Santoro, F., Donato, A., Lucchesi, S., Sorgi, S., Gerlini, A., Haks, M., Ottenhoff, T., Gonzalez-Dias, P., Consortium, Vsv-Ebovac, Consortium, Vsv-Eboplus, Nakaya, H., Huttner, A., Siegrist, C.-A., Medaglini, D., Pozzi, G., 2021. Human Transcriptomic Response to the VSV-Vectored Ebola Vaccine. Vaccines 9, 67. https://doi.org/10.3390/vaccines9020067

Sozzani, S., Luini, W., Borsatti, A., Polentarutti, N., Zhou, D., Piemonti, L., D'Amico, G., Power, C.A., Wells, T.N., Gobbi, M., Allavena, P., Mantovani, A., 1997. Receptor expression and responsiveness of human dendritic cells to a defined set of CC and CXC chemokines. J. Immunol. Baltim. Md 1950 159, 1993–2000.

Sozzani, S., Sallusto, F., Luini, W., Zhou, D., Piemonti, L., Allavena, P., Van Damme, J., Valitutti, S., Lanzavecchia, A., Mantovani, A., 1995. Migration of dendritic cells in response to formyl peptides, C5a, and a distinct set of chemokines. J. Immunol. Baltim. Md 1950 155, 3292–3295.

Stary, G., Olive, A., Radovic-Moreno, A.F., Gondek, D., Alvarez, D., Basto, P.A., Perro, M., Vrbanac, V.D., Tager, A.M., Shi, J., Yethon, J.A., Farokhzad, O.C., Langer, R., Starnbach, M.N., von Andrian, U.H., 2015. A mucosal vaccine against *Chlamydia trachomatis* generates two waves of protective memory T cells. Science 348, aaa8205. https://doi.org/10.1126/science.aaa8205

Taub, D.D., Conlon, K., Lloyd, A.R., Oppenheim, J.J., Kelvin, D.J., 1993. Preferential Migration of Activated $CD4^+$ and $CD8^+$ T Cells in Response to MIP-1α and MIP-1β. Science 260, 355–358. https://doi.org/10.1126/science.7682337

Taub, D.D., Sayers, T.J., Carter, C.R., Ortaldo, J.R., 1995. Alpha and beta chemokines induce NK cell migration and enhance NK-mediated cytolysis. J. Immunol. Baltim. Md 1950 155, 3877–3888.

The UniProt Consortium, Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., Alpi, E., Bowler-Barnett, E.H., Britto, R., Bursteinas, B., Bye-A-Jee, H., Coetzee, R., Cukura, A., Da Silva, A., Denny, P., Dogan, T., Ebenezer, T., Fan, J., Castro, L.G., Garmiri, P., Georghiou, G., Gonzales, L., Hatton-Ellis, E., Hussein, A., Ignatchenko, A., Insana, G., Ishtiaq, R., Jokinen, P., Joshi, V., Jyothi, D., Lock, A., Lopez, R., Luciani, A., Luo, J., Lussi, Y., MacDougall, A., Madeira, F., Mahmoudy, M., Menchi, M., Mishra, A., Moulang, K., Nightingale, A., Oliveira, C.S., Pundir, S., Qi, G., Raj, S., Rice, D., Lopez, M.R., Saidi, R., Sampson, J., Sawford, T., Speretta, E., Turner, E., Tyagi, N., Vasudev, P., Volynkin, V., Warner, K., Watkins, X., Zaru, R., Zellner, H., Bridge, A., Poux, S., Redaschi, N., Aimo, L., Argoud-Puy, G., Auchincloss, A., Axelsen, K., Bansal, P., Baratin, D., Blatter, M.-C., Bolleman, J., Boutet, E., Breuza, L., Casals-Casas, C., de Castro, E., Echioukh, K.C., Coudert, E., Cuche, B., Doche, M., Dornevil, D., Estreicher, A., Famiglietti, M.L., Feuermann, M., Gasteiger, E., Gehant, S., Gerritsen, V., Gos, A., Gruaz-Gumowski, N., Hinz, U., Hulo, C., Hyka-Nouspikel, N., Jungo, F., Keller, G., Kerhornou, A., Lara, V., Le Mercier, P., Lieberherr, D., Lombardot, T., Martin, X., Masson, P., Morgat, A., Neto, T.B., Paesano, S., Pedruzzi, I., Pilbout, S., Pourcel, L., Pozzato, M., Pruess, M., Rivoire, C., Sigrist, C., Sonesson, K., Stutz, A., Sundaram, S., Tognolli, M., Verbregue, L., Wu, C.H., Arighi, C.N., Arminski, L., Chen, C., Chen, Y., Garavelli, J.S., Huang, H., Laiho, K., McGarvey, P., Natale, D.A., Ross, K., Vinayaka, C.R., Wang, Q., Wang, Y., Yeh, L.-S., Zhang, J., Ruch, P., Teodoro, D., 2021. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 49, D480–D489. https://doi.org/10.1093/nar/gkaa1100

Van Den Eeckhout, B., Tavernier, J., Gerlo, S., 2021. Interleukin-1 as Innate Mediator of T Cell Immunity. Front. Immunol. 11, 621931. https://doi.org/10.3389/fimmu.2020.621931

Wang, A., Al-Kuhlani, M., Johnston, S.C., Ojcius, D.M., Chou, J., Dean, D., 2013. Transcription factor complex AP-1 mediates inflammation initiated by *Chlamydia pneumoniae* infection: AP-1 mediates *Chlamydia pneumoniae* inflammation. Cell. Microbiol. 15, 779–794. https://doi.org/10.1111/cmi.12071

Weiner 3rd, J., Domaszewska, T., 2016. tmod: an R package for general and multivariate enrichment analysis (preprint). PeerJ Preprints. https://doi.org/10.7287/peerj.preprints.2420v1

Wesa, A., Galy, A., 2002. Increased production of pro-inflammatory cytokines and enhanced T cell responses after activation of human dendritic cells with IL-1 and CD40 ligand. BMC Immunol.

3, 14. https://doi.org/10.1186/1471-2172-3-14
Woodrow, K.A., Bennett, K.M., Lo, D.D., 2012. Mucosal Vaccine Design and Delivery. Annu. Rev. Biomed. Eng. 14, 17–46. https://doi.org/10.1146/annurev-bioeng-071811-150054
Xu, D., 2008. Protein tyrosine phosphatases in the JAK/STAT pathway. Front. Biosci. Volume, 4925. https://doi.org/10.2741/3051
Xu, H., Cai, L., Hufnagel, S., Cui, Z., 2021. Intranasal vaccine: Factors to consider in research and development. Int. J. Pharm. 609, 121180. https://doi.org/10.1016/j.ijpharm.2021.121180
Zhao, W., Wang, L., Zhang, M., Wang, P., Yuan, C., Qi, J., Meng, H., Gao, C., 2012. Tripartite Motif-Containing Protein 38 Negatively Regulates TLR3/4- and RIG-I–Mediated IFN-β Production and Antiviral Response by Targeting NAP1. J. Immunol. 188, 5311–5318. https://doi.org/10.4049/jimmunol.1103506
Zhu, C., Xiao, F., Hong, J., Wang, K., Liu, X., Cai, D., Fusco, D.N., Zhao, L., Jeong, S.W., Brisac, C., Chusri, P., Schaefer, E.A., Zhao, H., Peng, L.F., Lin, W., Chung, R.T., 2015. EFTUD2 Is a Novel Innate Immune Regulator Restricting Hepatitis C Virus Infection through the RIG-I/MDA5 Pathway. J. Virol. 89, 6608–6618. https://doi.org/10.1128/JVI.00364-15

ISABELLE FRANCO MOSCARDINI
17/09/2023 15:01:02

Isabelle Franco moscardini

CHAPTER 6

Final discussion and Conclusions

Ebola virus, *Streptococcus pneumoniae* and *Chlamydia trachomatis* are three pathogens that researchers have put their efforts to fight in the past decades. Despite the huge advances in diagnosis, vaccination, and treatment, we still have important gaps to be covered in order to definitely overcome them.

To date, most gene expression studies aimed to deepen understanding of biology at the level of mechanisms and pathways, especially regarding the immune responses, and it is hoped that this knowledge will support different fields, ranging from rational vaccine design to identification of specific diagnostic and prognostic biomarkers of infection and vaccination. In this context, Systems Biology can be a powerful tool to explore the immune response while preserving, as much as possible, the complexity of the biological systems.

The main objective of this thesis was to leverage RNA-sequencing technology and computational methods to contribute new knowledge in the responses to *S. pneumoniae* infection and to Ebola and *C. trachomatis* vaccination. We have established a model to study a systemic response to *S. pneumoniae* lung infection, which permitted us to identify evidence of an early recall immune response in the spleens of mice intranasally infected with this pathogen. Genes and cytokines involved in this process were characterized, suggesting an involvement of both innate and adaptive branches of the immune system. Our *in-vitro* stimulation model has also shown to be valuable in detecting a previous infection by the expression values of only eleven genes.

ISABELLE FRANCO MOSCARDINI
17/09/2023 15:02:39
Isabelle Franco moscardini

Technology, Computation and Biology walk together in Systems Biology. During the course of this thesis, many available methods were used and a new framework based on Feature Selection and Machine Learning algorithms was built to deal with High-throughput data. This tool was applied to understand the differences and similarities between cohorts studying the rVSV-ZEBOV vaccine, against Ebola virus.

Systems Biology can also help address important challenges, such as the understanding of the responses to mucosal vaccines. In this work we characterized the transcriptomic profile of mice immunized with different schedules, including the intravaginal priming with the wild-type or a recombinant *S. gordonii* expressing the CTH522 protein, a *C. trachomatis* antigen. Besides the differences observed in the antibody response, the transcriptomic profile showed that the intravaginal priming modulates the systemic responses to the CTH522 protein and genes correlated with the IgG titers in the serum were identified.

This work has focused on understanding immune responses to infections and vaccines using transcriptomic analysis and Systems Biology, leveraging different methodologies to analyze and integrate high-throughput data. These approaches have shown to contribute to research at different levels, from animal models to clinical trials, in a diverse range of vaccines and infections.

ISABELLE FRANCO MOSCARDINI
17/09/2023 15:03:20

Isabelle Franco moscardini

APPENDIX

Publications

- Moscardini IF, Santoro F, Carraro M, Gerlini A, Fiorino F, Germoni C, Gholami S, Pettini E, Medaglini D, Iannelli F, Pozzi G. Immune Memory After Respiratory Infection With *Streptococcus pneumoniae* Is Revealed by *in vitro* Stimulation of Murine Splenocytes With Inactivated Pneumococcal Whole Cells: Evidence of Early Recall Responses by Transcriptomic Analysis. Front Cell Infect Microbiol. 2022 Jun 20;12:869763. doi: 10.3389/fcimb.2022.869763. PMID: 35795182; PMCID: PMC9251119.

Presentations

- **VacPath Annual Meeting 2020:** "Using Systems Biology approach to understand immune responses to vaccines"
- **VSV-EBOPLUS Annual Meeting 2020:** "Comparison of transcriptomic response to rVSV-ZEBOV in the Geneva cohort and in the North American cohort"
- **VacPath Annual Meeting 2021:** "A transcriptomic approach to study recall immune responses"
- **VSV-EBOPLUS Annual Meeting 2021:** "Prioritizing the importance of biological components within High Throughput data: a machine learning approach"
- **VacPath Annual Meeting 2022:** "Using Systems Biology and Machine Learning to uncover gene signatures of infection and vaccination"

ISABELLE FRANCO MOSCARDINI
17/09/2023 15:04:58

Isabelle Franco Moscardini

Brazilian 🇧🇷 28/11/1994

📍Strada di Marciano, 41 - Siena (SI) - 53100, Italy

📞 +39 3394007179 ✉️ *isabelle.moscardini@gmail.com*

linkedin.com/in/isabelle-franco-moscardini-91134b149

https://github.com/IsaMoscardini

Summary

PhD Candidate in Bioinformatics and MSCA ESR. My project consists in understanding the immune response to infections and vaccines through a systems biology approach. I have a Bachelor's degree in Pharmacy and Biochemistry from the University of São Paulo (USP). I have previous experience with Molecular Biology and internships in pharmacovigilance and Medical and Scientific Information at Sanofi.

Education

PhD in Medical Biotechnologies (Oct 2019 - Oct 2022)
VacPath Project - University of Siena and Microbiotec srl, Siena - Italy

Bachelor in Pharmacy and Biochemistry (Feb 2013 - Aug 2019)
University of São Paulo, São Paulo - Brazil

ISABELLE FRANCO MOSCARDINI
17/09/2022 15:05:31

Sciences de la Vie (Jan 2017 - Jul 2017)
Exchange program - Sorbonne University, Paris - France

Experience

Isabelle Franco moscardini

PhD Student in Medical Biotechnologies (Oct 2019 - Oct 2022)
PhD student at VacPath program, supported by the Marie Sklodowska-Curie actions (MSCA).

- Application of OMICs data analysis, data integration, biomarker discovery and coexpression analysis to characterize molecular mechanisms of pneumococcal infection and vaccines against *C. trachomatis* and HIV

Part of VSV-EBOPLUS Consortium in collaboration with Computational Systems Biology Laboratory

- Use of OMIC data integration and Machine Learning methods to characterize the signatures of the immune response to the rVSV-ZEBOV Ebola vaccine, correlating RNA sequencing with immunogenicity and reactogenicity data
- Study the effect of rVSV-ZEBOV vaccine on the expression of long non-coding RNAs and their relation with protein-coding genes

Research internship - Computational Systems Biology Laboratory (Aug 2017 - Nov 2018)
Bachelor's thesis project: Comparative Analysis of Signaling Pathways Involved in Malaria Infection

- Use public data to perform a metanalysis of transcriptomic studies (Microarray), identifying molecular biomarkers of Malaria infection
- Application of different bioinformatics tools to characterize the consistent response found among the studies

Intern in Medical and Scientific Information - Sanofi (Dec 2018 - Jun 2019)
Working with different interface areas including Pharmacovigilance, Regulatory affairs, Call center and Quality.

- External assistance to healthcare professionals and consumers, clarifying questions based on the scientific literature and Internal assistance in the search of scientific articles and references for

company materials

- Participation in local and Global projects, including the implementation of a global Medical Information system in the Brazilian affiliate
- Creation and improvement of operational procedures and training of employers and service providers

Intern in Pharmacovigilance - Sanofi (Dec 2017 - Nov 2018)

- Assistance in process control and analysis of pharmacovigilance data from digital media and market research
- Interface with vendors and training of digital media and market research agencies in pharmacovigilance procedures
- Assistance in signal detection analysis

Research internship - Molecular Biology and Microbial Ecology Laboratory (Aug 2014 - Oct 2016)
Project: "Linocin M18 Gene Expression and Characterization in Burkholderia seminalis"

- Characterization of Linocin M18 gene expression at different conditions. Trained in different molecular biology techniques such as real-time PCR, electrophoresis, DNA and RNA extraction, bacterial cloning and basics of Protein Structure Modeling
- FAPESP scholarship (Oct/2015 - Oct/2016)

Published Articles

Moscardini IF, Santoro F, Carraro M, Gerlini A, Fiorino F, Germoni C, Gholami S, Pettini E, Medaglini D, Iannelli F, Pozzi G. Immune Memory After Respiratory Infection With *Streptococcus pneumoniae* Is Revealed by *in vitro* Stimulation of Murine Splenocytes With Inactivated Pneumococcal Whole Cells: Evidence of Early Recall Responses by Transcriptomic Analysis. Front Cell Infect Microbiol. 2022 Jun 20;12:869763. doi: 10.3389/fcimb.2022.869763. PMID: 35795182; PMCID: PMC9251119.

Courses

- Vaccinology Course (Final mark 18/20)
  - *Institut Pasteur - 7 February to 4 March 2022*
- Research Writing in the Sciences
  - *INASP - 6 April to 17 May 2021*

Other Activities

- Professor's assistant in Physiopathology I, Immunodiagnostics, Bioinformatics applied to Health Sciences and Molecular Biology
- Student Representative and participation in college entities

Key Skills

ISABELLE FRANCO MOSCARDINI
Informatics
17/09/2023 15:06:19

- Programming: R, Python
- Windows and Linux environments
- Code collaboration/Version Control (Git & Github)
- Intermediate skills in Microsoft Office package
- SPSS and Minitab software

Isabelle Franco Moscardini

Personal skills

- Excellent teamwork and communication skills
- Self-directed learning skills
- Problem-solving

Languages

🇬🇧 English (Fluent)

🇫🇷 French (Delf B2)

🇮🇹 Italian (Intermediate)

🇧🇷 Portuguese (Native Speaker)

Acknowledgments

I have started this thesis saying that "*Systems biology is about putting together rather than taking apart, integration rather than reduction"*. In fact, I believe science is about putting things together, because everything we learn is somehow thanks to the knowledge shared by others. Science is about sharing and discussing, it is about integrating what you know with other people's knowledge.

In this way, I would like to thank everyone in the Laboratory of Molecular Microbiology and Biotechnology and in Microbiotec srl for having me during these three years. I especially thank Professor Gianni Pozzi, Professor Francesco Santoro and Alice Gerlini, for trusting me for this position and for guiding me in my projects throughout this PhD. I also thank Caterina and Valentina for their kindness and assistance. Additionally, this work would not be possible without the support of the European Commission, which financed the VacPath research program.

I also thank Professor Helder Nakaya and people from the Computational Systems Biology Laboratory in Brazil, for introducing me into bioinformatics, for our great collaborations in the VSV-EBOPLUS consortium, but especially for all the technical and moral support I received during these years. A special thanks to Fernando, Patricia and Thiago for their unconditional help and for being amazing friends.

Just as science is about sharing, so is life. I am lucky to have amazing friends and a family that supported my education from the beginning and that continuously kept pushing me forward, even thousands of kilometers away. My endless gratitude to my mother and father for always encouraging me to go further, and to my sister and best friend, Aline. Thanks Cecilia and Lucas, my dear friends that are always cheering me up and thanks Simone, for all the bioinformatics discussions (and the great pizzas). And I could not finish without thanking my dear colleague and friend Bar, and his wife Anna, for making my life lighter, even during the pandemics.