Università degli Studi di Torino

PhD Programme in Complex Systems for Life Sciences

XXXIII cycle

Systematic characterization of human response to H1N1 influenza vaccination through the construction and integration of personalized transcriptome response profiles

Author

Carlo De Intinis

Supervisor

Michele Caselle

PhD Programme Co-ordinator

Michele De Bortoli

External supervisors

Emilio Siena
Duccio Medini

Contents

List of Figures

List of Tables

Transparency statement

Carlo De Intinis is a PhD student at the University of Turin, with a scholarship from GlaxoSmithKline Biologicals SA. Emilio Siena is an employee of the GSK group of companies.

The clinical study was independently conducted by the National Institute of Health (NIH), in accordance with the Declaration of Helsinki. The study is registered on clinicaltrials.gov with the reference ID NCT01191853. Data was collected from the Gene Expression Omnibus public database (Accession ID: GSE47353).

Acknowledgements

Emilio Siena
For being not only my tutor, but a true mentor during this experience.

Duccio Medini • *Michele Caselle*
For the invaluable feedback about this work and beyond.

Margherita Bodini • *Denise Maffione* • *Laurane De Mot*
Margherita Coccia • *Alessandro Brozzi*
For contributing to this work.

Roberta Marino • *Stefano Carello* • *Ottavia Prunas* • *Olivia Lanzoni*
Sara Sorgi • *Andrea Cavallone* • *Mattia Dalsass*
For being excellent colleagues and good friends during this experience.

Gianfranco De Intinis • *Amelia Di Pasquantonio*
For all the support that a son could ever ask and more.

Emanuela Berto
For the best years of my life.

Abstract

Gene expression data is commonly used in vaccine studies to identify genes exhibiting shifts in expression level between treatment and control cohorts. However, group-wise analyses of perturbations on a complex system, such as the human immune system, may not be suitable to fully characterize the impact of a treatment, as captured by the diversity of responses exhibited by individual subjects. Suboptimal responses from a subset of subjects may be determined by both heritable and non-heritable influences, as well as demographic and socioeconomic differences. Group-wise methods are optimal for capturing signals that differentiate treated from non-treated subjects, but they are not sensitive enough for detecting changes diversified across individuals, where variance and noise can hide the signal. Additionally, modulations occurring in a small proportion of subjects are also generally lost.

In this work we developed and tested an analysis framework for assessing *individual transcriptome response profiles* in the context of repeated-experiment studies (where the same subjects are repeatedly probed across multiple time points) and benchmarked it on a non-adjuvanted H1N1 vaccination dataset [1]. We built up from the work of Menche and colleagues [2], where the authors demonstrated the feasibility of the single subject approach in determining differentially expressed genes.

This approach succeeded in providing a higher amount of information compared to group-wise based comparisons. We highlighted extensive heterogeneity in the peripheral blood transcriptional response to vaccination and we were able to capture biologically meaningful signals despite a significant noise component in the data. Finally, we described the individual transcriptome response profiles via *pools of vaccination-induced genes* and studied their association with functional antibody responses. Overall, we propose a new analysis pipeline, that is generally applicable to repeated-measure experiments, and provide evidence that it can help to generate new and useful insights.

1 Introduction

This chapter briefly introduces key concepts regarding the human immune system in the context of the response to a vaccine stimulus and we provide a high-level overview of the most typically applied investigative techniques. Influenza and its vaccines are also described, with particular emphasis on the 2009 H1N1 pandemic. We introduce the limitations of canonical group-wise studies in the context of human immunology and strategies to overcome these limitations to gain new insights into the mode of action of vaccines and the vaccine-induced immunological responses.

1.1 Systems Biology

Systems Biology is the comprehensive, quantitative analysis of the multiple components of a biological system, and their interactions, over time [3]. Such **complex systems** are characterized by *emergent properties*: characteristics that are not reflected by the single components, nor can be appreciated by the independent study of the individual components. An example of an emergent property is the *robustness* of the system, which can arise from positive and negative feedback loops and/or redundant functions (i.e. biological pathways participating to the same biological function). The practice of systems biology involves collection, integration and analysis of multiple sets of biological data collected from different sources across distinct time points and hierarchical levels. The integration of multiple datasets also enables for a much higher flexibility, opening to numerous options for data analysis approaches and allowing to address questions that were left unanswered. This work represents an example of this last concept.

1.2 The human immune system

The human immune system is a highly complex biological system [4]. It encompasses multiple organs and its mechanisms of action are determined by a number of cells and molecules which interact at different levels and time scales. Its physiological role is the defence against exogenous molecules and pathogens, while at the same time avoiding the insurgence of a response against the host's self components. The responses of the immune system can be broadly categorized into innate and adaptive. The **innate response** is non-specific and it is quickly activated upon detection of pathogens and extraneous substances. In contrast, the **adaptive response** is a specific response, tailored to a precise pathogen and enhanced by repeated exposition to the antigen.

1.2.1 Innate immune response

The innate immunity is the host's first line of defence against infections. The response, quick and non-specific, is achieved through physical barriers, specialized cells and immunoactive molecules [5].

The **epithelium** is the first line of defence. It acts as a passive barrier against pathogens, but it also produces antimicrobial peptides. Its functions are also complemented by mucus, saliva and local bacterial flora. When a pathogen succeeds in overcoming this first barrier, it is faced by innate immune cells, which are able to sense this through *pattern recognition receptors* (PRRs), which recognize the evolutionarily conserved patterns of microorganisms [6].

This initial recognition, mediated by mast cells and macrophages, leads to **inflammation** which is characterized by the production of a variety of inflammatory mediators (cytokines, chemokines etc.) which lead to the recruitment of leukocytes (mainly neutrophils) to the affected tissue [7]. Monocytes are able to recognize danger signals through PRRs. They can phagocytose and present antigens, secrete chemokines and differentiate into macrophages and Dendritic Cells (DCs) [8]. DCs express receptors able to detect microbial molecules and act as Antigen Presenting Cells (APCs), leading to the promotion of the adaptive response (see Section 1.2.2) by interfacing the innate and adaptive arms of immunity [9]. Some examples of mediators are the Tumor Necrosis Factor (TNF) and components of the Interleukin-1 family, which recruit and activate leukocytes [10] [11]. Neutrophils then kill the foreign agents by releasing reactive oxygen species (ROS), reactive nitrogen species and other effectors from their granules [7]. The inflammatory response then results in the phagocytosis of infectious agents by the macrophages, the apoptosis of compromised host cells mediated by Natural Killer (NK) cells and the release of anti-inflammatory and reparative cytokines [12].

Another central component of the innate immune system is the **complement system**. It consists of a number of small proteins that circulate in the blood as inactive precursors. Upon activation, these proteins enhance the inflammatory response through the production of pro-inflammatory molecules and are involved both in lysis and phagocytosis processes due to their antigen opsonization activities [13]. The complement system also acts as a bridge between innate and adaptive immune response by stimulating T-lymphocytes [14].

1.2.2 Adaptive immune response

Adaptive immunity is the host's second line of response against pathogens. It is highly specific to given antigens, adaptable and characterized by the development of immunological memory: such memory allows to quickly and efficiently respond to a threat after having previously encountered it [15].

Lymphocytes B and **T** are the key players in this response and are respectively responsible for humoral and cellular immunity. Naïve B and T cells develop in the bone marrow and in the thymus respectively. When an exogenous molecule (antigen) is detected, T cells quickly migrate in secondary lymphoid organs, where they are presented with the antigen by Antigen Presenting Cells (APCs: Dendritic Cells, Macrophages, B cells). Specifically, the *Major Histocompatibility Complex II* (MHC class II) is used by APCs to present lysosomal degradation products to **CD4$^+$ T cells** which orchestrate the specific immune response [16]. Concurrently, MHC class I-antigen complexes stimulate **CD8$^+$ T cells**, which are able to kill infected cells [15]. A subset of these CD8$^+$ T cells will exhibit long-term survival: these **Memory T cells** elicit an enhanced response upon subsequent encounters with the antigen [17].

Naïve **B cells** are localized in the lymph nodes' follicles. Here, they are activated by interactions with the T cells through the stimulation of the CD40 receptor (ligated by the CD40L expressed by T helper cells). Alternatively, the bacterial lipopolysaccharides (LPS) or repetitive bacterial antigens (such as flagellin, which cross-links with the membrane-bound IgGs) may activate the B cell [18]. In both cases, short-lived plasma cells that produce low-affinity antibodies are generated. Additionally, if the B cells are stimulated by a T cell, they enter the germinal centres to differentiate into high-affinity plasma cells or Memory B cells. In the germinal centre, B cells undergo somatic hypermutation, which can generate high or low affinity antigen receptors. If a low affinity receptor can not be rescued into a high affinity state through the affinity maturation process, the entire cell undergoes apoptosis [19]. Subsequently, B cells with high affinity receptors undergo class switching, which causes the expression of membrane-bound IgGs or IgEs instead of IgMs. **High-affinity plasma cells** migrate in the extrafollicular regions of secondary lymphoid organs and in the bone marrow, where they secrete antibodies. **Memory B cells** survive for years, during which they recirculate between lymph nodes and spleen and are able to elicit a strong and efficient response to subsequent exposures to the antigen [20].

1.3 Vaccines

A vaccine is a biological preparation of a microbial antigen, often associated to adjuvants, which is administered to induce protection against microbial infections [21]. This protection is achieved by inducing an adaptive immune response against the antigen, leading to the development of immunological memory. The adjuvant's role is instead to enhance the immunological response. Several kinds of techniques are used to obtain different vaccine types.

- *Attenuated vaccines* - They contain a live virus which through repeated *in vitro* culture in sub-optimal conditions is rendered non-pathogenic for humans. Usually they show high efficacy, there is however the remote risk for retro-mutation into a pathogenic form in the human host.

- *Inactivated vaccines* - Composed of previously live pathogens, which are killed through chemical means, temperature or radiation. They are safer than attenuated vaccines, although they produce a weaker response.

- *Toxoid vaccines* - These are aimed at bacterial infections where the toxin produced by the bacteria is the main cause of pathogenicity. Toxins themselves can be rendered harmless and administered in this new toxoid form.

- *Subunit vaccines* - These are composed not from an entire microorganism, but from antigens which best stimulate the host's immune system, such as proteins, extracts of infected tissues or polysaccharides. They generally elicit a weaker response, so these are often adjuvanted. Another method for increasing immunogenicity is to *conjugate* polysaccharides to proteins.

- *Synthetic vaccines* - The antigen is obtained not through purification, instead recombinant DNA technologies are used to produce it in synthetic form.

- *Viral vector vaccines* - A non cytopathic virus is used for the expression of an antigen in the subject (which is infected with the virus itself). This induces a strong immune response, but safety concerns are still limiting the application of this technology.

- *DNA vaccines* - These work via a plasmid which contains a cDNA that codes for an antigen. The idea is that some APC will take up the plasmid and express the antigen themselves, causing an immune response. However these vaccines have not demonstrated high efficacy.

Additional experimental techniques, such as RNA-based vaccines, exist. These have shown higher immunogenicity than DNA vaccines in humans, although proof of efficacy is still missing and non trivial side effects were observed [22] [23] [24].

1.3.1 Vaccines development

Creating a vaccine is a long and complex process, which can take years or even decades. However, vaccines are typically the most effective intervention to control infectious diseases [25]. Initially, a **pre-clinical** study needs to be conducted: at this stage the vaccine is researched and developed. This is followed by the **clinical** stage, composed of four main phases and an additional fifth phase in some cases, where assessments on safety and efficacy are conducted on volunteers [26].

- *Phase I* - A limited number of subjects are tested, usually in a dose escalation study design, to establish the safety of the investigational vaccine for use in humans.
- *Phase II* - several hundreds of subjects are tested to evaluate effectiveness and to confirm the dosing ranges. Common short-term side effects and risks are also evaluated.
- *Phase III* - several hundreds to thousands of subjects are typically tested in these expanded trials. These studies gather additional information regarding effectiveness and safety. They are also needed to evaluate the overall risk-benefit relationship of the vaccine. If these studies are successful, the drug can be approved for marketing.
- *Phase IV* - these studies are performed after marketing approval. They are used to elucidate the incidence of adverse effects and to determine the drug's effect on morbidity of mortality. These studies may also be used to investigate a previously not described population (such as children). They can also be market-oriented comparison studies against competitor products.
- *Phase V* - some pharmaceutical companies may want to consider a further study phase, aimed at new indications for the drug, novel formulations or different dosage forms.

1.4 Influenza

Influenza is a human infectious respiratory disease primarily caused by genus A and B influenza viruses. Symptoms, which typically resolve within a few days, include runny nose, fever and myalgia. However the infection may also result in more severe complications such as systemic infection, pneumonia and even death. These unfavorable outcomes are more likely in specific communities, like elderly and immune compromised individuals [27] [28] [29]. All influenza strains are enveloped, negative-sense single-strand RNA viruses. To date, four genera of influenza virus are able to infect vertebrates: A, B, C and D [30]. Of these, only the first three are able to infect humans, although influenza D virus antibodies were found in humans with occupational contact with cattle [31]. Human influenza A and B viruses cause seasonal epidemics, while the A strains are the only influenza viruses known to cause pandemics. A and B strains contain eight RNA segments, which in the A strain encode 11 proteins: RNA polymerase subunits, Haemagglutinin (HA, glycoprotein, facilitates viral entry), Neuraminidase (NA, glycoprotein facilitates viral release), Nucleoprotein (NP), Matrix protein (M1), Membrane protein (M2), Nonstructural protein (NS1) and Nuclear export protein (NEP) [27] [32].

Influenza A viruses (IAV) are divided into subtypes based on the two surface proteins, HA and NA: the name of the A virus will be determined by these two antigenic subtypes (i.e. H1N1, H3N2). They have a variety of hosts. Influenza B viruses (IBV) cause symptoms similar to IAV, however IBV exhibits limited subtype diversity and slower antigenic drift (see Section 1.4.1), these two characteristics may be the reason for the narrower host range of IBV, with humans acting as the primary reservoir [30]. Influenza C virus (ICV) causes infections in children with typically mild symptoms. ICV is much more antigenically stable than IAV and serological studies indicate that humans develop antibodies against the virus during early life on a worldwide scale [33].

1.4.1 Influenza vaccines

Influenza viruses, especially A strains, cause 3-5 million severe cases and 290.000-650.000 deaths worldwide each year, not counting the four pandemics occurred in the last 100 years (1918, 1957, 1968 and 2009) [34] [35]. To date, annual vaccination represents the most effective intervention for protection against this disease [36]. However, a universal vaccine for influenza does not exist. The mechanisms of antigenic diversity shown by the virus lead to a persistent threat to global health: antigenic drift and antigenic shift are the cause of seasonal and pandemic manifestations respectively [35]. **Antigenic drift** is the result of point mutations in the hemagglutinin (HA) and neuraminidase (NA) viral genes, the two immunodominant components of the influenza virus outer envelope, leading to seasonal influenza outbreaks generally each year. **Antigenic shift** instead occurs unpredictably when novel influenza A viruses, previously able to infect animals, acquire the capacity to infect humans. This is known as a *spillover* event and it can happen through spontaneous genomic mutations or genetic reassortment of an animal strain with a human one, an event favoured by the segmented nature of the viral genome.

Obtaining an effective seasonal vaccine is a challenge in itself: the virus strains must be selected at least 6 months prior to commercialization since the vast majority of doses are produced by growing the virus in eggs. In this time frame a significant strain drift is possible, leading to antigenic mismatch and reduced vaccine effectiveness [35]. Low effectiveness can also be caused by the manufacturing process, since influenza can mutate during the growth phase in the eggs [37].

1.4.2 The pandemic H1N1 2009 influenza virus

Influenza A H1N1 caused the first influenza pandemic of the 21$^{\text{th}}$ century. It was the result of re-assortment between avian, human and swine influenza viruses and affected 214 countries in which the vast majority of the population had little to no immunity [38]. The seasonal influenza vaccines of the time were also shown as non beneficial to any age group with respect to an increase in cross-reactive antibodies, due to the antigenic distance of the pandemic strain from the seasonal one [39]. This highlighted the need for the development of a new ad hoc vaccine to cover this pandemic strain and to adopt new approaches such as adjuvants and specific administration schedules with the aim of increasing the breadth of coverage [39]. In the end, both adjuvanted and non-adjuvanted vaccines were developed. A recent meta-analysis highlighted an overall vaccine effectiveness (adjusted against hospitalization) of 61% (95% CI 14-82%) with higher effectiveness in lower age groups [40], while a previous study highlighted a possible benefit to be gained from different dosages in certain population groups [41].

Past experience highlights two fundamental concepts regarding influenza vaccination. First, a unique vaccine formulation may not be suited for the entire population. Second, the mechanisms through which the vaccines achieved protection are still not fully understood and further investigational studies are desirable. There is the need for an in-depth analysis of each vaccinated subject in order to better understand which immunological mechanisms are involved in the response to vaccination. This would be an impossible task with conventional methods, but recent bioinformatic approaches and in general the *Systems Vaccinology* field of research can be an important aid to this goal.

1.5 Systems Vaccinology

In Section 1.1 we defined Systems Biology as the comprehensive, quantitative analysis of the multiple components of a biological system, and their interactions over time. **Systems Vaccinology** can be defined as the application of Systems Biology methodologies to the vaccine discovery, research and development process. The rationale behind this approach is to overcome the hurdles posed by certain pathogens (i.e. influenza viruses), for which vaccine design is difficult and/or improved efficacy is desired [42]. The main goal of such approaches is the investigation of multiple immune parameters before and following vaccination, with the aim to better understand the immune response, gather insights into the mode of action of investigational vaccines and guide their selection and/or improvement. Systems Vaccinology studies involve the extensive use of *omics* technologies, such as sequencing of DNA and RNA across a multitude of time points, ranging from the pre-vaccinated state to even many months after vaccination. Typically, these data are also integrated with information obtained from conventional immunological assays (i.e. neutralization assays, ELISA, ELISPOT) and/or intrinsic factors (i.e. age, gender, population). The immune system in the context of a vaccine study is an excellent candidate for such approaches: vaccination is an orchestrated manipulation of the system, which can be achieved across multiple subject without incurring into ethical problems.

A series of system biology studies have been successful in identifying immune perturbations triggered by diverse influenza vaccine formulations such as live attenuated influenza vaccine (LAIV [43]), trivalent inactivated influenza vaccine (TIV [44]) and adjuvanted influenza vaccines (MF59, AS03 [45]). In addition, researchers were able to identify molecular signatures capable of predicting the serological response to influenza vaccination [46] [47] [2].

1.5.1 Variability across individuals

The goal of many systems vaccinology studies has been that of identifying biomarkers of vaccine-mediated protection or immunogenicity. Those are vaccine-induced molecular or cellular signatures that are able to prospectively determine if the vaccination will be successful in a given subject. The biomarkers identified to date, however, are able to provide a limited classification accuracy of 80-85% [48] [44] and no clear casual relationship between innate immune activation and adaptive immune responses have been identified so far. This is, in part, due to the high heterogeneity found in the human population, which leads to different responses to the same vaccine [49]. **Heterogeneity** is explained by both *heritable* (genetic variation) and *non-heritable influences* (such as symbiotic and pathogen microbes), the latter being the most influential factor [50]. However, the current demographic and socioeconomic changes of the past century are diversifying humans even more than before. We currently live in a world where age, nutritional status and incidence of chronic infections and inflammatory disorders are influencing the vaccination response of entire populations.

Antibody levels and their functionality still represent the best indicators of efficacy for many vaccines; systems biology had limited success in identifying novel correlates of protection. It is also known that a IFN-γ-mediated transcriptional response and the presence of antigen-specific T cells can be correlated with increased protection [51], however a clear distinction between protected and non-protected individuals is still not achieved [49]. Peripheral Blood Mononuclear Cells (PBMCs) are widely used in vaccinology studies, but they represent an indirect source of information. Probing the immune system in tissues other than blood could yield useful information. Specifically, characterization of tissue-resident memory T cells could inform about protection, since these cells are crucial for local immunity and recall responses [52]. However these kind of studies require invasive techniques which are harder to execute due to ethical and practical limitations.

Given the intrinsic heterogeneity of the immunological responses at the molecular, cellular and systemic level, classical group-wise analysis may not be the most suited approach to assess the responses triggered by vaccination in a given pool of individuals. This because group-wise methods can capture signals that are consistently found in most of the tested subjects, while they may not be sensitive enough for capturing subtle changes, for which variance and noise can hide the signal. In addition, modulations occurring in a small portion of the tested subjects could also be lost due to the nature of these methods.

Menche and coworkers, recently published a study in which they attempt an individual-subjects based analysis approach to better understand the molecular mechanisms underlying multiple systemic diseases [2]. We believe that the application of a similar approach to vaccine studies can improve our understanding of vaccination responses. Single-subject analysis could help overcome subject heterogeneity while being at the same time reliable, since the immune system is relatively stable over time within individual subjects [50].

1.5.2 Describing subject heterogeneity (Menche et al. 2017)

In canonical case-versus-control groups transcriptional studies, the aim is generally to obtain differentially expressed genes (DEGs) between the two cohorts. The underlying assumption is that such genes would point to biological mechanisms which if perturbed would lead to different phenotypic manifestations. These DEGs could then be used as biomarkers to diagnose a disease. However, multiple sources of heterogeneity can lead to inconsistencies across studies. This can be caused by intrinsic technical variability in the analysis techniques and methods, variable gene expression levels (from genetic factors, SNPs, copy number variations, non genetic factors etc.) and inherent stochasticity of the biological processes.

In 2017, Menche and colleagues [2] observed that very few genes are consistently modulated across all subjects in a specific cohort. Specifically, the analyzed healthy and diseased cohorts for asthma, Parkinson and Huntington diseases. They devised a computational method to obtain individual subject gene expression profiles, with the goal of acquiring predictive disease-associated gene pools. Although the observed perturbation profiles were highly heterogeneous, a statistically significant overlap of these pools with the individual transcriptome profiles led Menche et al. to correctly associate DEGs with a disease diagnosis in 80-100% of cases.

The individual perturbation expression profiles (PEEPs) were obtained through a Z-score which captured by how many standard deviations a gene expression level in a single diseased subject was distant from a reference distribution derived from healthy subjects. Through descriptive statistics, pathway enrichment analyses and the generation of disease-specific gene pools, the researchers were able to achieve the following results.

- Often DEGs were perturbed only in a small fraction of PEEPs.
- There were highly significant similarities between PEEPs of case subjects, which were absent in healthy subjects. Similarity was measured by observing how many genes were commonly modulated across pairwise subject combinations for each specific cohort. These similarities could not be attributed to a few common DEGs, but were the result of more complex patterns of pairwise overlaps.
- Pathway enrichment analysis showed a degree of homogeneity at the pathway level, leading the researchers to conclude that different perturbations within a molecular pathway could lead to similar phenotypic outcomes.
- The PEEPs value in providing useful information was proved by classifying subjects as healthy or diseased with comparable performances to state-of-the-art machine learning approaches. The PEEP concept was shown as complementary to these methods, being easily investigable through techniques such as enrichment analyses. The PEEPs could also be used as a measure of heterogeneity of the different diseases.

1.6 Aim of our work

In this work we developed and tested an analysis framework for **assessing individual transcriptome response profiles in the context of repeated-experiment studies**. In this type of investigation, the same subjects are repeatedly probed across multiple time points. This is the standard approach of most Phase I vaccine clinical studies, thus we expect our method to be readily generalizable. Our method is in part based in the work from Menche et al. ([2], see Section 1.5.2), where the authors demonstrated the single subject approach in determining differentially expressed genes.

We examined a dataset [1] of subjects simultaneously vaccinated with seasonal and pandemic 2009 H1N1 vaccines. We built up from the work of Menche and colleagues [2] by implementing a set of measures to assess the robustness and increase the confidence in the results obtained. We also sought to reduce the noise introduced by the stochasticity of biological processes. For this reason, we generated **categorical** individual transcriptome response profiles, in which the genes could only assume values of up/down modulation or no regulation. We investigated the feasibility of such an approach and developed a novel method for pathway enrichment analysis. We also sought to generate core **gene pools** which could be associated in a statistically significant manner to the vaccination event. From these pools we sought to determine which genes were driving the differences in the immunological response observed across subjects.

1.7 Main findings

- Our analysis highlighted extensive heterogeneity in the peripheral blood transcriptional response to vaccination.
- The *individual transcriptome response profiles* that we generated were able to capture biologically meaningful signals despite an important noise component in the data.
- We were able to describe the transcriptome response via *pools of vaccination induced genes* and gain information regarding the association of such responses with functional antibody response.

2 Dataset

In the following sections, we outline the dataset used for this study. Data types, applied data extraction and data pre-processing techniques are also presented.

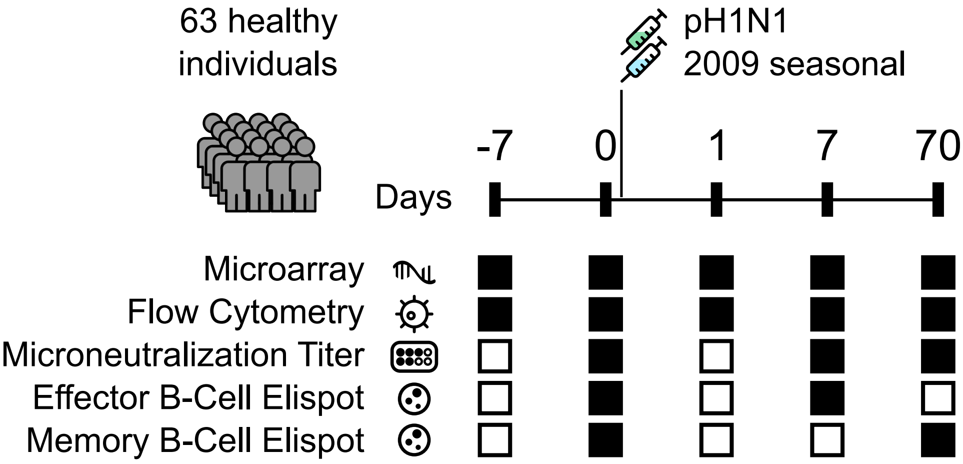
2.1 Tsang et al. 2014

The dataset we used comes from a vaccination study conducted by Tsang et al. [1], describing 63 healthy subjects (see Figure 2.1 for study outline). Subjects were 37 females and 26 males, with a mean age of 31, ranging from 21 to 62 years old (CIQR: [1] 23-38). Two non adjuvanted subunit vaccines were concomitantly administered to each subject: the 2009 Fluvirin seasonal influenza vaccine from Novartis and the 2009 H1N1 pandemic vaccine from Sanofi-Aventis. The data files are deposited in NCBI's Gene Expression Omnibus [53] [54] and are accessible through GEO Series accession number GSE47353.

2.1.1 Transcriptome data

The **transcriptome** is the ensemble of genetic code that is transcribed into RNA molecules. By studying it, researches hope to translate genomic sequence information into functional biological mechanisms, with the aim of better understanding the biological system under investigation [55]. **Microarrays**, initially developed for DNA mapping [56], are utilized to measure the expression level of thousands of genes within a mRNA sample [57]. The typical use of this technology involves the comparison of different biological conditions (i.e. health vs disease). DNA hybridization is the core

[1] Central InterQuartile Range



Days	-7	0	1	7	70
Microarray	■	■	■	■	■
Flow Cytometry	■	■	■	■	■
Microneutralization Titer	□	■	□	■	■
Effector B-Cell Elispot	□	■	□	■	□
Memory B-Cell Elispot	□	■	□	□	■

Figure 2.1: Study design of the Tsang et al. 2017 dataset indicating blood collections and analyses performed.

mechanism of action of microarrays: two DNA strands hybridize if they are complementary to each other (following the Watson-Crick rules). The probes (small sequences of DNA) are attached to the inert surface of the microarray and these will hybridize with the target sequences obtained from the retro-transcribed RNA sample, which are labeled (usually with a fluorescent dye). Each probe is of known sequence and location upon the array. Subsequently, laser-induced fluorescent imaging is used to generate an image. The amount of fluorescence is measured at each location of the array, allowing the quantification of the gene expression levels [55]. Importantly, these experiments do not provide data on the true concentrations of RNA, but are useful to compare expression levels among conditions and genes [57].

Transcriptome data for this work was obtained from peripheral blood mononuclear cells (PBMCs). Fasting blood samples were drawn between 8 and 11 AM and were lysed and stored at -80°C within 30 min. Subsequently, RNA was extracted using miRNeasy (*QUIAGEN*). Samples were hybridized to *Affymetrix GeneChip Human Gene 1.0 ST Arrays.* Tsang and colleagues processed the data using APT from Affymetrix and removed batch effects via linear modeling.

2.1.2 Microneutralization Titer data

The **Microneutralization (MN) Assay** is based on the ability of serum antibodies to prevent infection in mammalian cells *in vitro* [58]. It can be used to assess baseline serostatus as well as the humoral response following natural infection or vaccination. Cells are stained and plated on a 96 well plate. Subsequently, in the quantitative form of the assay, serial two-fold dilutions of the serum sample are prepared and mixed with the virus. These dilutions are used to infect the plated cells. The **neutralization titer** is typically expressed as the reciprocal of the serum dilution showing at least 50% inhibition of infection in the cell culture [59].

Tsang and colleagues obtained virus-neutralizing titers for four strains (see Table 2.1) according to an assay based on the methods of the pandemic influenza reference laboratories of the Centers for Disease Control and Prevention described by Hancock et al. in 2009 [39]. For this specific work, reported titers were the highest dilution that completely suppressed viral replication.

Viral strain	**Description**
A/California/07/Swine	H1N1 pandemic of 2009 (pdm09)
A/Brisbane/59/07	H1N1 seasonal
A/Uruguay/716/07	H3N2, component in seasonal flu vaccine
B/Brisbane/60/2001	Seasonal influenza B strain

Table 2.1: Viral strains tested for pre-immunity

2.1.3 B-Cell ELISPOT assay

The **enzyme-linked immunospot (ELISPOT) assay** was first described in 1983 [60], with the basic methodology only undergoing minor modifications since then. Predominantly, the B-Cell assay is applied in the detection of B-Cell responses to infections and vaccinations [61], with the aim of determining the frequency of cytokine-secreting cells. While cytokines are generally the main proteins of interest, the assay itself is suitable for almost any secreted protein. In the assay, cells are cultured (in presence or absence of stimuli) on a surface coated with a specific antibody. Secreted cytokines are captured by the surface antibodies and can be detected via a detection antibody which can be either biotinilated and followed by a streptavidin-enzyme conjugate or can be directly conjugated to and enzyme. The substrate product then precipitates, with the end result being visible spots on the surface, each corresponding to a secreting cell.

Tsang and colleagues used the ELISPOT assay to determine total and influenza-specific IgG/A frequencies of antibody-secreting cells, using a protocol previously described (Ho et al., 2011 [62]).

3 Preliminary group-wise analysis for differentially expressed genes

In this chapter we describe a preliminary group-wise analysis aimed at uncovering differentially expressed genes (DEGs) from peripheral blood mononuclear cells (PBMCs) transcriptome data.

3.1 Preface on data analysis

The analyses presented in this work have been executed leveraging **R**, a language and environment for statistical computing and graphics [63]. Users can write scripts using the R programming language to add new functionalities or, in the case of this work, to define an entire analysis pipeline. Additional packages from the open-source, open-development software project **Bioconductor** [64] can be used to further expand the capabilities of R. When used, these packages will be referenced in the text.

3.2 Gene expression data pre-processing

We **downloaded** the gene expression dataset from a public repository (see Section 2.1) using the `GEOquery` R package [65] and obtained an *ExpressionSet* object [1] containing both the transcript abundance data and the metadata describing 292 samples and 33292 probes. We **annotated** probes and samples using the available metadata. Specifically, a sample is described by its anonymized subject ID and the time point of analysis, while probes are associated to the corresponding Gene Symbol (Entrez gene IDs). We

[1] One of the data classes of the Bioconductor project. Contains and combines standardized data structures from different sources to represent genomic data.

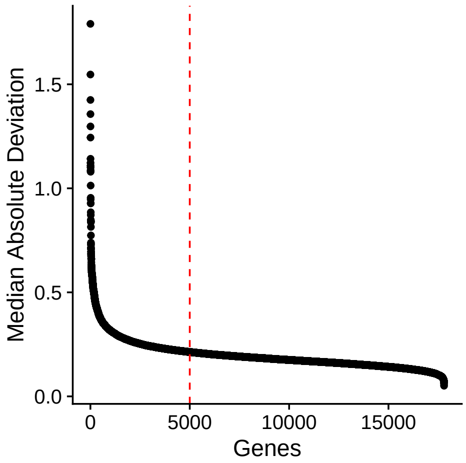


Figure 3.1: Gene-wise Median Absolute Deviation of expression data. The dashed line shows the cutoff at 5000 genes.

collapsed the **redundant probes** [2] of the microarrays using the Bioconductor package `limma` [66]; the expression values for the collapsed probes were computed as the geometric mean value of the individual probes' expression values. Since the dataset was already normalized and quality controls were executed by Tsang and colleagues, control probes were removed.

High-dimensional data, like microarray transcriptome data, are prone to type-I error (rejection of a true null hypothesis, or *false positive*). To mitigate this phenomenon, we reduced the data dimensionality by applying two **filtering steps**. First, we restricted our analysis to protein coding genes based on the gene type definitions downloaded from the Biomart database [67] as of Feb. 14th 2018; this reduced the dataset to 17798 transcripts. Then we computed the median absolute deviation [$\mathrm{MAD} = \mathrm{median}(|X_i - \tilde{X}|)$] of the genes across all samples (see Figure 3.1) and selected the 5000 genes with the highest MAD. These are the genes used for data analysis henceforth. The rationale for selecting genes with the highest MAD was that these genes display the highest variation in the expression values across the samples, making them ideal candidates to investigate the vaccination's effect on gene expression.

Albeit a total of 63 healthy subjects were enrolled in the study, not all transcriptomic sample runs were successful. This resulted in a reduced

[2] Probe sets that map to different regions of the same gene.

dataset size (see Table 3.1).

Time point	Available subjects
day -7	60
day 0	57
day 1	60
day 7	57
day 70	58

Table 3.1: Number of subjects with available transcriptome data

3.3 Group-wise analysis for differentially expressed genes

As a preliminary step, we executed a group-wise analysis on transcriptome data aimed at detecting differentially expressed genes. We considered days 1 and 7 as response time points and days -7, 0 and 70 as baselines. Day 70 was included in the baselines due to its similarity with the other two pre-vaccination time points, as reported by Tsang and colleagues. We calculated the mean differences of gene expression values for each combination of baseline-versus-response time points. Values were tested for significance using the Wilcoxon signed-rank test (p-value < 0.05, two-sided, paired) and the resulting p-values were corrected for multiple testing using the Benjamini Hochberg correction. Figure 3.2 *a*-*d* shows the result of the analysis relative to the day 0 baseline (see Supplementary Figures A.1 and A.2 for days -7 and 70 respectively).

In Figure 3.2 *e* we instead report the transcript abundance values of the CXCL10 gene for each subject in the dataset at days 0 and 1. This chemokine, also known as interferon γ-induced protein 10 (IP10) is a chemoattractant molecule capable of recruiting T lymphocytes into sites of tissue inflammation [68]. Previous vaccine studies reported CXCL10 as one of the most strongly upregulated genes in peripheral blood following vaccination, with a peak response typically occurring around 24 hours [69]. We also found the CXCL10-encoding mRNA level to be significantly upregulated 24 hours after immunization ($\log_2$ fold-change $= 0.96$, adjusted p-value $= 7.5 \times 10^{-7}$, two-tailed Wilcoxon signed-rank test). Despite the robust response observed at a group level, this gene was not modulated to an appreciable extent in

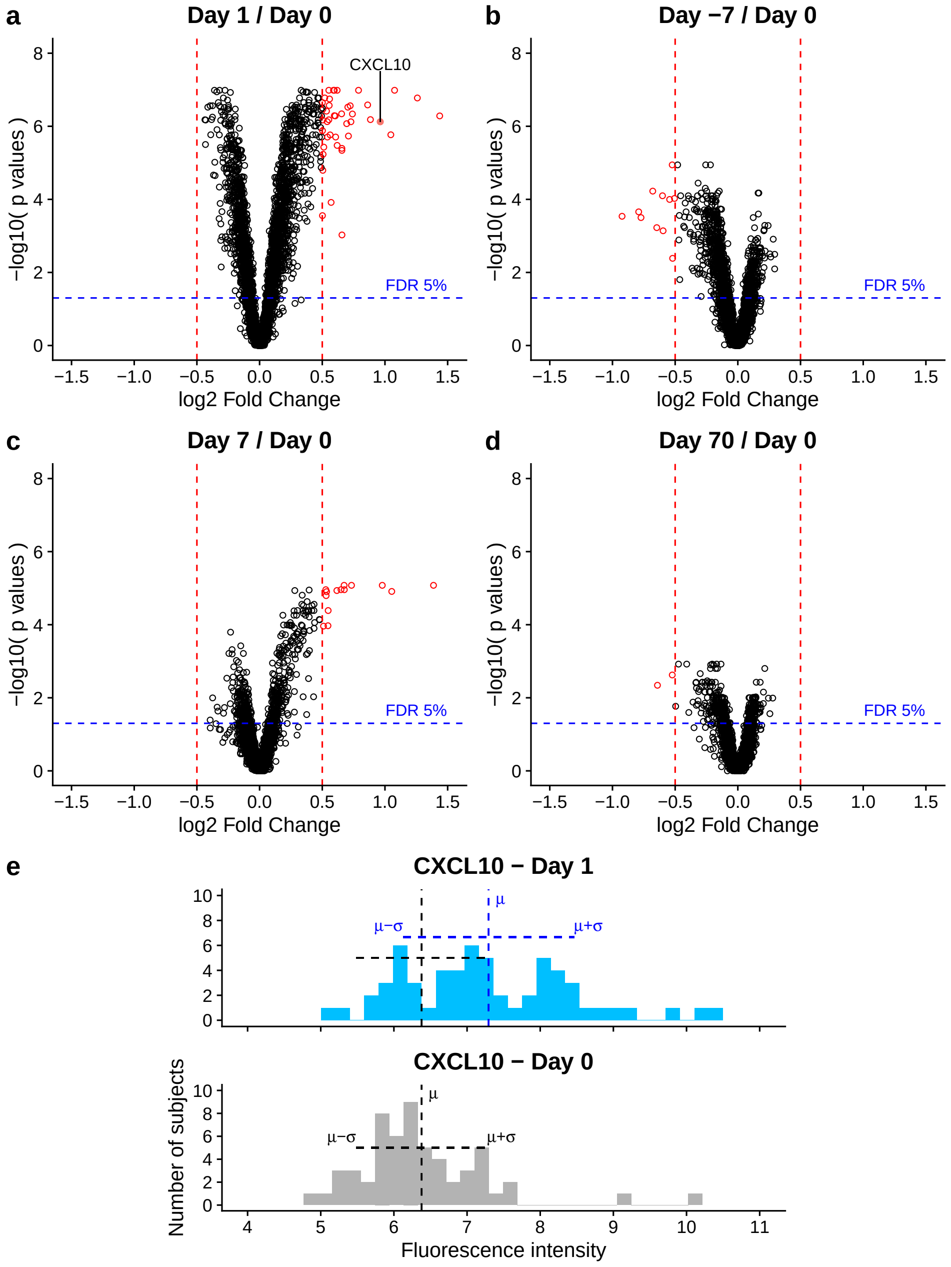


Figure 3.2: **a - d** Mean expression level differences of genes, with differentially expressed genes detected via group-wise analysis shown in red. The dashed lines show the cutoffs for calling a gene differentially expressed: $\log_2$ Fold Change ±0.5 (red lines), p-value < 0.05 (two-tailed, paired Wilcoxon signed-rank test, blue line). **e** comparison of the gene expression levels of the CXCL10 gene at day 1 and day 0.

every vaccinated subject. In half of the cases (30 out of 60), we found the transcript abundance to fall within ± one standard deviation from the average preimmune level. Conversely, for 5 subjects, CXCL10 baseline levels exceeded the 24 hours average level. Overall, this pattern suggest a considerable level of **intersubject heterogeneity** in the transcriptome responses to the influenza vaccines.

Generalizing, testing post-vaccination gene expression values against the day 0 baseline values resulted in 43 and 14 DEGs for days 1 and 7 respectively (absolute $\log_2$ fold-change from baseline ≥ 0.5, adjusted p-value < 0.05, two-tailed Wilcoxon signed-rank test). The assessment of transcripts abundance levels across different subjects revealed similar levels of heterogeneity as observed with CXCL10, suggesting that this characteristic is pervasive to most modulated genes (see Supplementary Tables B.1 and B.2).

3.4 Random effect removal from dataset

In order to take into consideration subject-specific deviations in transcript abundance levels (as opposed to vaccination-induced deviations), we applied a **linear mixed-effect model** for the analysis of the transcriptome profiles. Mixed effect modeling is a statistical procedure capable of describing the observed variance through three different components, the fixed effect, the random effect and the residual error. The *fixed effect* describes the variation induced by the treatment. The *random effect* describes the subject-specific effects on the values. The *residual error* quantifies the unexplained variance present in the data (noise). In the context of this study, the dataset contains repeated measurements of the same subjects across different time points. This characteristic is the main reason that led us to utilize a mixed effect model; a simpler linear model for instance would be unsuitable due to one of its assumptions being the total independence of the samples. With the *random effect* component of the mixed-effect model we are able to account for the fact that subject-specific effects are present in repeated-measurements data. Subjects may exhibit different gene expression values due to differences between subjects themselves which are not determined by the treatment. Since these subject-specific effects should be constant across all time points, by removing this component on a subject-wise basis, we can correctly

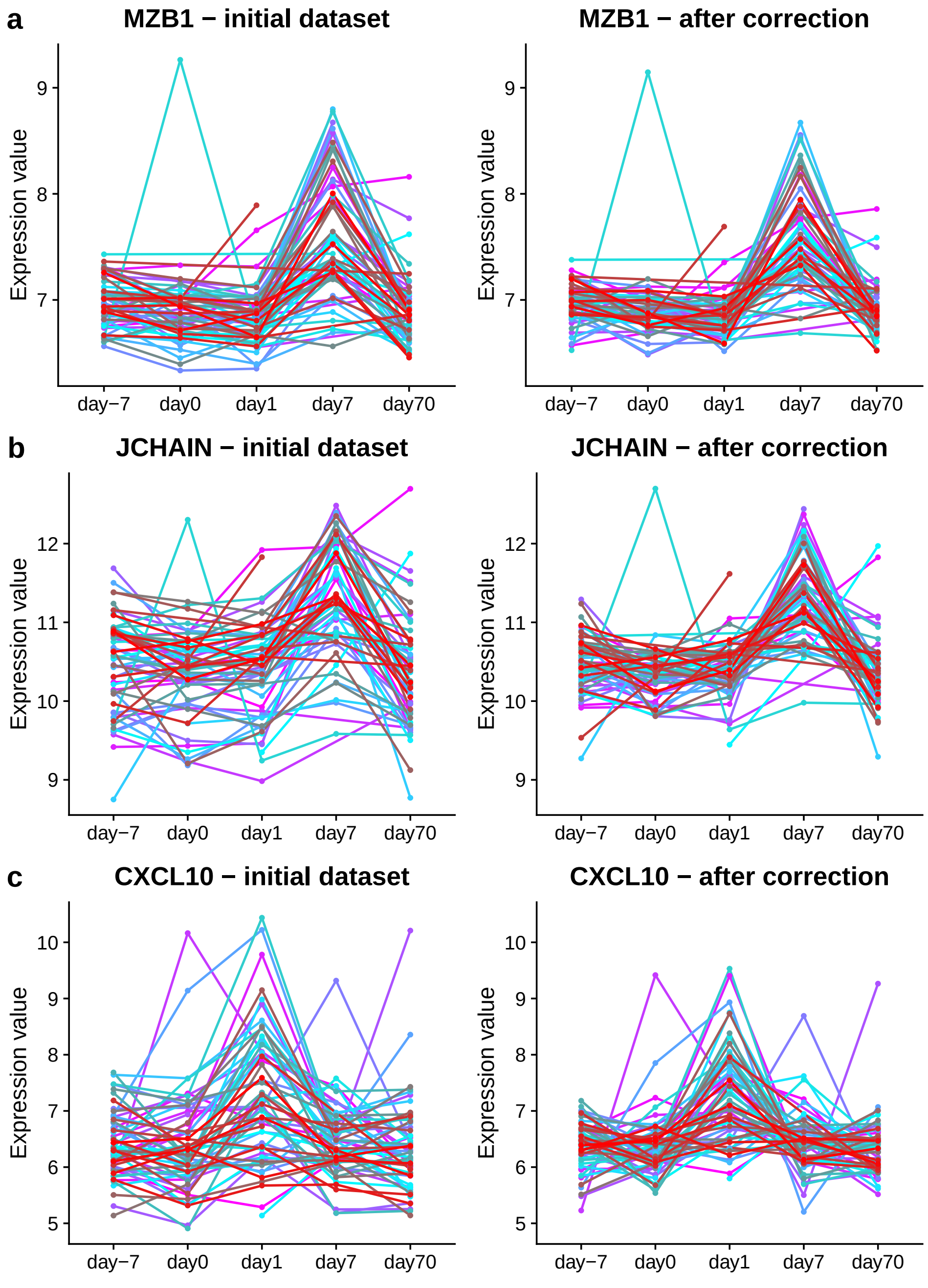


Figure 3.3: Examples of gene expression data values correction via linear mixed effect modeling.

compare gene expression values among subjects.

We used the R package `lme4` [70] to fit the model to the transcriptome data. We defined the model using the gene expression values as the dependent variable, the sample time point as the fixed effect and the subjects as the random effect. [3] We then subtracted the random effect that was calculated by the model from the gene expression values (see Figure 3.3).

[3] Model formula (in R language): `lmer(values ~ days + (1|subjects), data=DF)`

4 Analytical framework for deriving individual transcriptome response profiles

We present an analysis framework for deriving individual transcriptome response profiles in the context of repeated measures experiment data, where the same individuals are probed repeatedly across multiple time points. Our approach is able to derive individual profiles based on temporal differences in genes modulation occurring within each treatment group. We also report the exploratory results of our analysis and characterize the heterogeneity of the response profiles.

4.1 Framework

We developed our analytical framework using the R programming language [63]. Our method is based on a work from Menche and colleagues (see Section 1.5.2) and it compares the transcript abundance of each single gene, from a single subject and a specific time point, to a number of reference distributions generated by bootstrapping [1] the pre-immunization transcript abundance values for that same gene. Therefore, while the approach proposed by Menche et al. can generate individual profiles based on the comparison of different cohorts (control versus disease groups), our approach is able to derive individual profiles based on temporal differences in gene modulation occurring along time. This characteristic is particularly applicable

[1] **Bootstrap**: resampling technique used to estimate the statistics of a population by sampling multiple times a dataset. **Resampling**: any technique or instance of generating a new sample from an existing dataset.

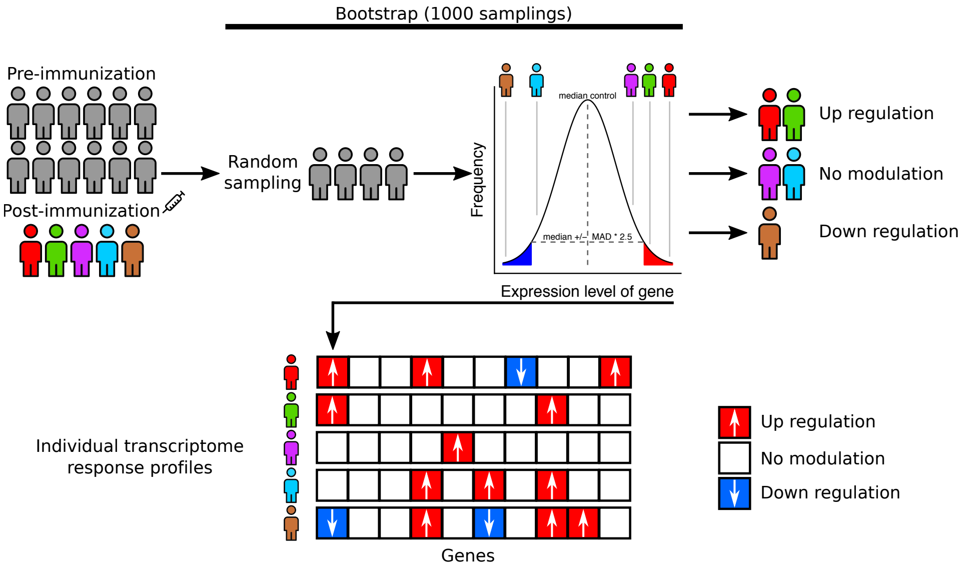


Figure 4.1: Framework for deriving individual transcriptome response profiles

in vaccine studies, where samples are usually collected and analyzed at different times following vaccination. Being a biological process, transcription has a stochastic component, which translates into gene expression variance. We devised a bootstrap approach to assess gene expression variance under normal conditions: therefore we aggregated subjects' data at day 0 and used it for the bootstrap procedure, as at this time point no treatment has been administered yet. This assessment of gene expression variance leads to more robust results. The bootstrap approach also allows to have a measure of confidence of the results. The main steps executed on each gene with our technique are outlined below and shown in Figure 4.1.

- We generate $n = 1000$ **control distributions** via bootstrap from the expression values of a single gene from pre-immunization data. n random samples are executed to generate n control distributions. Each of the samples occurs without replacement [2] and has size of $2/3$ of the total number of subjects available at day 0.

- For each bootstrap iteration, we derive thresholds t that we use to determine if a gene is modulated or not. The thresholds are set for

[2] In sampling without replacement, each unit has only one chance to be selected from the population.

each control distribution at ± 2.5 median absolute deviations (MAD, see Equation 4.1) from the median value $\tilde{X}$ (see Equation 4.2).

$$\mathrm{MAD} = \mathrm{median}(|X_i - \tilde{X}|) \tag{4.1}$$

$$t_{upregulation,downregulation} = \tilde{X} \pm \mathrm{MAD} * 2.5 \tag{4.2}$$

- We call a gene **modulated** for a single subject at a specific time point if its expression value is above (*up-regulation*) or below (*down-regulation*) the test's threshold value in at least 75% of the bootstrap iterations.

After these steps, we obtain **individual transcriptome response profiles**, which are arrays of categorical gene modulation data for each time point tested. Genes can be categorized as either *not modulated*, *up-regulated* or *down-regulated* and represented as 0, 1 and −1, respectively.

We applied this framework on the PBMC transcriptome data from Tsang and coworkers (see Section 2.1.1). We used the pre-vaccination (day 0) data to generate the control distributions and the post-vaccination time points as test data. Day 0 data was also tested in order to provide an estimate of the false positives rate.

4.2 Transcriptome responses to Influenza vaccination are highly heterogeneous

Initially, we compared the individual subject transcriptome profiles with the DEGs identified by the Wilcoxon test-based group-wise analysis (see Section 3.3) to test if our framework captured additional genes and if individual subjects exhibited differences in gene modulation. We also included samples from days -7 and 70 to be used as negative controls. Being collected before (day -7) and a long time after vaccination (day 70), these samples are expected to reflect no vaccine-specific effects and be indicative of the false discovery rate to be expected from the analysis. Despite some overlap, not all group-wise DEGs were found in the individual profiles (see Figure 4.2). Among the 43 DEGs found at day 1 with the group-wise analysis,

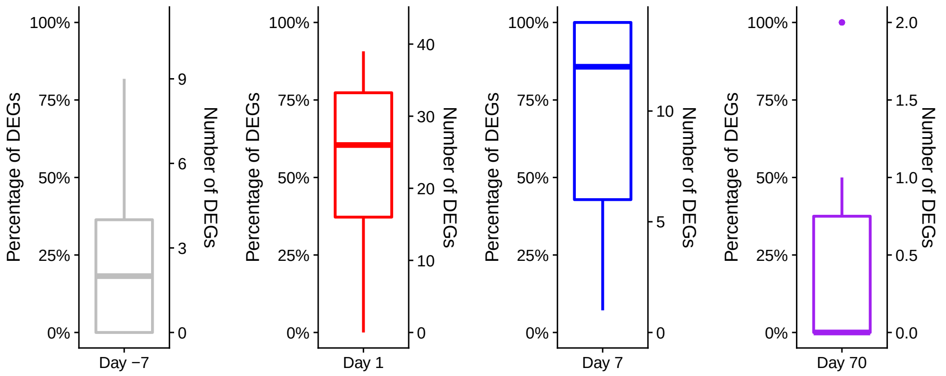


Figure 4.2: Differentially expressed genes from group-wise analysis that are also found within individual subject transcriptome profiles. *Right y-axes*: number of DEGs from group-wise analyses found in each individual transcriptome response profile. *Left y-axes*: percentage of DEGs from group-wise analysis found in each individual transcriptome response profile.

an average (median) of 60% (CIQR: [3] 37-77%) were also observed in the individual profiles, while at day 7 (14 DEGs) the overlap with the individual profiles averaged at 85% (CIQR: 42-100%). These values suggest that **peripheral blood transcriptome responses to influenza vaccination vary greatly across different subjects** and that group-wise analyses approaches are not suited to accurately describe this heterogeneity. The control time points showed an overall smaller overlap. For day -7 (11 DEGs): 18% (CIQR: 0-36%); for day 70 (2 DEGs): 0% (CIQR: 0-38%).

4.3 Individual-subjects response profiles provide a detailed characterization of the inter-subject response heterogeneity

We assessed the consistency of the transcriptome responses observed after vaccination across different individuals by comparing the individual subject response profiles obtained with our framework.

Figure 4.3 *a* represents the distributions of the number of modulated genes across different subjects and time points. Day 1 was the time point showing the highest amount of gene responsiveness as well as the highest

[3] Central InterQuartile Range

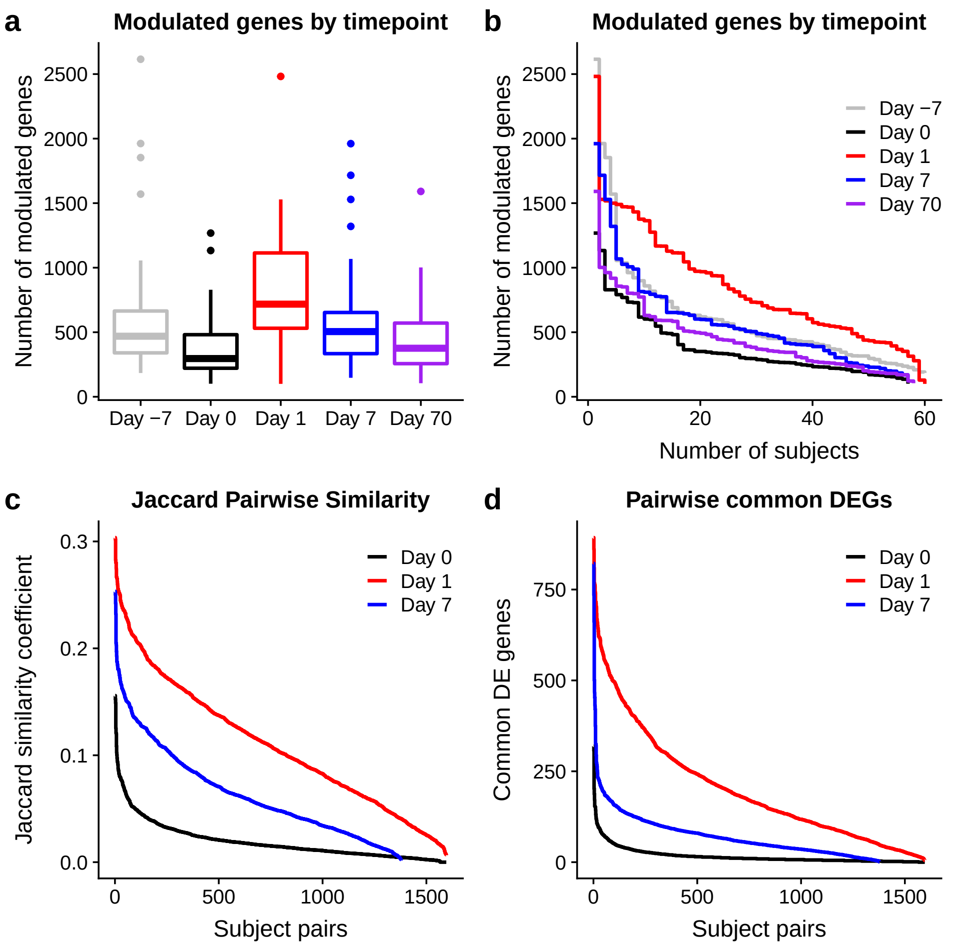


Figure 4.3: Heterogeneity among transcriptome profiles and quantification of DEGs. **a** Number of modulated genes identified in each subject across the different time points. **b** Reverse cumulative distributions of DEGs number across subjects. **c** Jaccard similarity coefficients of pairwise subjects combinations calculated using individual subject transcriptome profiles. **d** Number of shared DEGs among pairwise subjects combinations.

Time point	p-value	adjusted p-value (BH)	-$\log_{10}$ adjusted p-value
Day 0 VS Day 1	5.494e-08	5.494e-07	6.26
Day 1 VS Day 70	1.113e-06	5.567e-06	5.254
Day 0 VS Day 7	9.286e-05	3.095e-04	3.509
Day -7 VS Day 0	2.162e-04	5.406e-04	3.267
Day 1 VS Day 7	4.734e-04	9.469e-04	3.024
Day -7 VS Day 70	1.776e-03	2.960e-03	2.529
Day -7 VS Day 1	4.449e-03	6.356e-03	2.197
Day 7 VS Day 70	2.078e-02	2.598e-02	1.585
Day 0 VS Day 70	3.660e-01	4.066e-01	0.391
Day -7 VS Day 7	6.984e-01	6.984e-01	0.156

Table 4.1: Wilcoxon signed-rank tests of time point comparisons of individual subject transcriptome profiles, ordered by adjusted p-value.

amount of variance across subjects. The individual profiles had an average of 717 modulated genes, with a CIQR ranging from 531 to 1114 (see Figure 4.3 *a*). In general, the other time points showed a smaller variance, although the difference between 1st and 3rdrd quartiles were still approximately two-fold (CIQR day 7: 334-653, CIQR day -7: 340-664, CIQR day 70 256-571). These results suggest that the transcriptome modulation across different individuals is highly heterogeneous.

To test the sensitivity of our approach in capturing vaccine-specific signals, we compared the number of modulated genes across different time points. The expectation is to observe fewer differences among baseline time points (day -7 VS day 0, day 70 VS day 0) compared to time points following the vaccination. Indeed, we found the comparisons between the baseline time points and days 1 and 7 to be the most significant among all possible combinations of time points (see Table 4.1). Day 1 individual profiles had significantly more modulation when compared to other time points (Figure 4.3 *b* and Table 4.1, Wilcoxon signed-rank test [4]). Contrarily to this behaviour, day 7 profiles demonstrated marginally superior modulation when compared to day 70, but were not significantly different from day -7.

Additionally, a relevant information regarding individual response pro-

[4] Non-parametric statistical hypothesis test used to compare two related samples (in this case the relation is due to the repeated measurements on the samples). The samples do not need to be normally distributed. We use the test to assess whether two dependent samples were selected from populations having the same distribution.

files is how frequently the various genes are found to be modulated across different subjects. Surprisingly, this analysis revealed that no single gene was found to be consistently modulated across every subject. This was observed for both day 1 and day 7. At the same time, a substantial proportion of modulated genes were found to be specific to individual subjects, specifically 185 and 300 at days 1 and 7 respectively. We observed the most consistent response to be at day 1 (see Figure 4.3 *b*). At this time point, for example, the number of genes fond to be modulated across 20 or more subjects was 587. At day 7 the same value dropped to 120 genes. Pre-vaccination time points had 84 and 0 of these genes (days -7 and 0 respectively), while we observed 8 genes at day 70.

Next, we investigated the similarity of individual transcriptome response profiles. Based on all possible pairwise combinations of subjects we computed the Jaccard similarity coefficient (see Equation 4.3) and the number of commonly modulated genes for each of these subject pairs (see Figure 4.3 *c,d*). The Jaccard similarity coefficient has a finite range of values: $0 \leq J(A,B) \leq 1$, where 0 represents a total absence of similarity and 1 represents identical gene modulations between two subjects (see Figure 4.4).

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{4.3}$$

We observed a median Jaccard coefficient of 0.014 (CIQR: 0.007-0.024) at the pre-immunization time point. These values increased 7- and 5-fold at days 1 and 7 respectively; the day 1 median was of 0.1 (CIQR: 0.06-0.15), while the day 7 median corresponded to 0.055 (CIQR: 0.03-0.09). The similarity exhibited by the subject profiles at the post-vaccination time points was significantly higher than day 0 (p-values day 1 vs day 0 and day 7 vs day $0 < 2.2 \times 10^{-16}$, two-sided Kolmogorov-Smirnov [5] test). These small similarities across the individual profiles confirm the substantial heterogeneity across individual transcriptome response profiles.

To confirm that the observed signals were contributed by vaccination, as opposed to random signals, we compared the overlap across the individual profiles at days 0, 1 and 7, using the Fisher exact test (see Figure 4.5 and

[5] Non-parametric test of equality of one dimensional distributions, used to compare two samples.

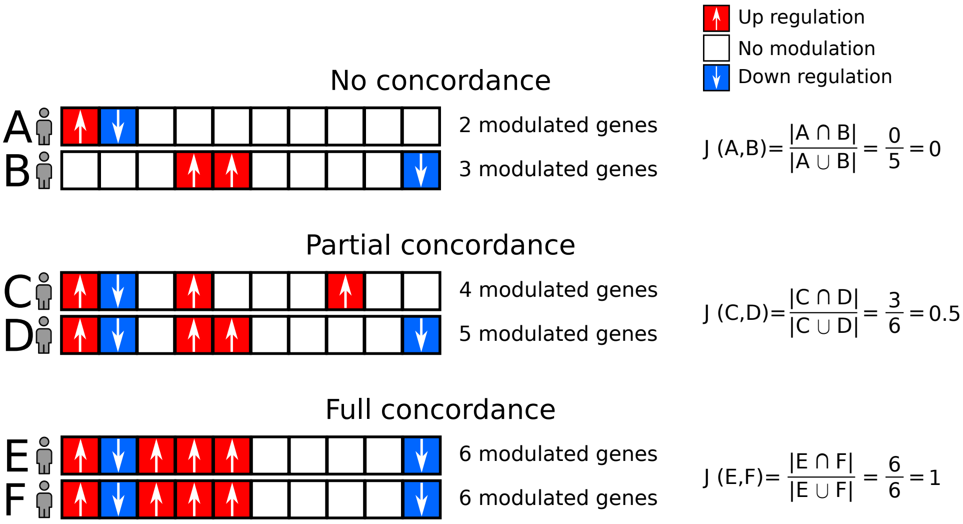


Figure 4.4: Examples of Jaccard similarity coefficient values obtained from pairwise subject combinations. Each square represents a gene from an individual subject transcriptome profile.

Table 4.2). The rationale was that day 0 responses, which do not include the vaccination effect, should display a significantly smaller overlap.

	Subject A	Subject B
Non shared DEGs	#	#
Shared DEGs	#	#

Table 4.2: Example of contingency table generated for each pairwise combination of subjects.

Leveraging the same pairwise combinations of subjects that we previously used to calculate the Jaccard similarity coefficients, we computed the contingency tables for the Fisher tests (see Table 4.2). At day 0 none of the comparisons had significant overlap (Benjamini-Hochberg adjusted p-value ≤ 0.05), whereas at the post-vaccination time points (days 1 and 7) responses were significantly associated to vaccination in 90% and 74% of the cases.

Overall, the collected evidences show that the individual response profiles generated using our framework were able to capture vaccination-specific responses and provided supplemental information when compared to conventional group-wise analyses. Moreover, we have observed substantial intersubject heterogeneity in the peripheral blood transcriptome response to the subunit influenza vaccines. Despite the statistically significant overlaps

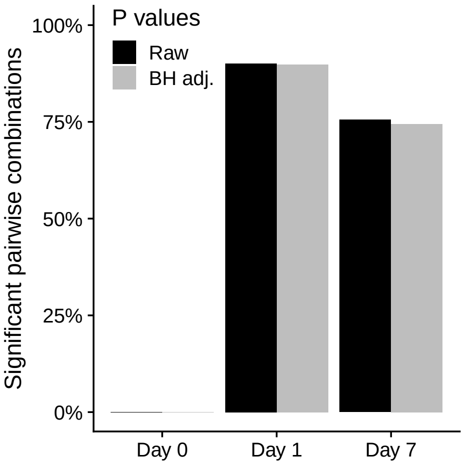


Figure 4.5: Percentages of pairwise subject combinations whose gene overlap is statistically significant according to the Fisher exact test. We report raw and adjusted p-values (Benjamini-Hochberg correction for multiple testing) in black and gray respectively.

in gene modulations, the number of genes commonly modulated across individuals was generally low and no genes were found to be modulated across all subjects at a specific time point.

4.4 Individual transcriptome response profiles are independent from age and gender

We examined subjects' age and gender (see Section 2.1) in order to understand if individual transcriptome response profiles were influenced by these intrinsic characteristics. To achieve this we generated networks of subjects for days 1 and 7, where the position of subjects in the network was determined by the degree of similarity in the transcriptome response previously measured via Jaccard similarity coefficients (i.e. subjects with similar transcriptome response profiles appear to be geometrically close in the network; see Section 4.3). The expectation is to be able to visually identify clusters of subjects which may arise if the subject's intrinsic characteristics influence the transcriptome response.

We used the R package `RCy3` [71] and the software `Cytoscape` [72] to generate and visualize **networks** where *nodes* represent subjects, with color

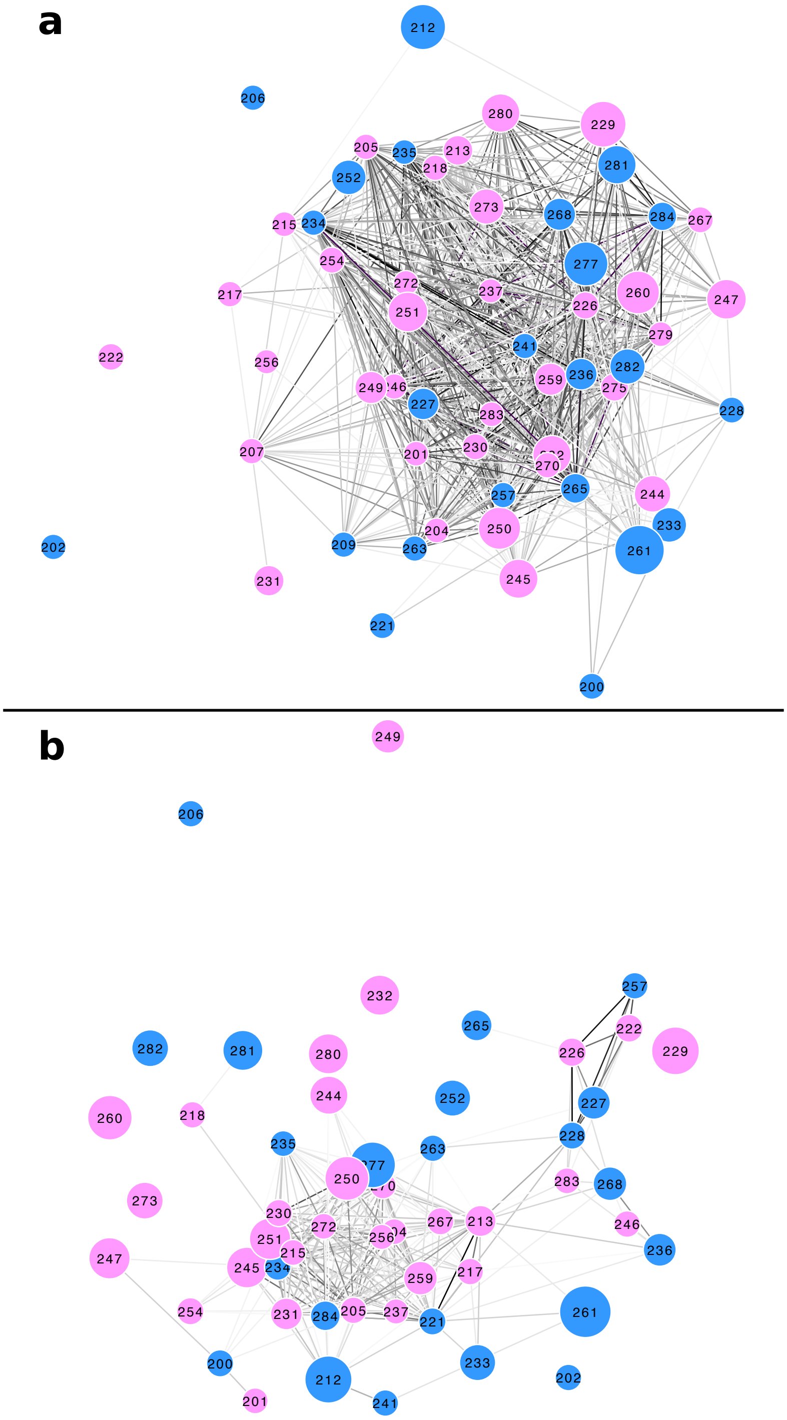


Figure 4.6: Networks for days 1 and 7 (**a** and **b** respectively). *Nodes* are color-coded by gender (cyan and pink for males and females respectively) and size-coded by age (smaller for younger). *Edges* are color-coded by Jaccard similarity coefficient (higher values are darker, values below 0.1 are hidden).

and size corresponding to gender and age, respectively. *Edges* of the graphs were defined using the Jaccard similarity coefficient of each subject pair and were color-coded using the value of the coefficients themselves (darker color for higher coefficients). Nodes were spatially organized using the *Perfuse Force Directed Layout*, using the Jaccard coefficients as weights.

The networks for days 1 and 7 are shown in Figure 4.6. We were unable to visually identify subject stratification either by gender or age in relation to the similarity of the transcriptome response, in accordance to what was already observed by Tsang and colleagues [1]. However, we previously observed a higher similarity in subjects' transcriptome responses at day 1 (higher Jaccard similarity coefficients of pairwise subject combinations, see Section 4.3). This higher similarity is shown in the graphical representation of the network by the darker color of the edges connecting the subjects and by the tighter grouping of the nodes (see Figure 4.6).

4.5 Investigation of the robustness of the observed individual subject response profiles

In this work we describe the results of an implementation of our framework based on a **non-parametric test** for determining differentially expressed genes from bootstrapped control distributions. The thresholds for calling a gene modulated have been defined as ±2.5$*$ *median absolute deviations* from the median value of the control distribution. However to understand the degree of stability of the results, during development we tested a total of four approaches.

- We compared the non-parametric test for differential expression to a **parametric** one, based on t-distribution parametrization. For the parametric test, thresholds are set to ±1.96$*$ *standard deviations* from the mean gene expression value.
- We tested the **influence of outlying values** on determining differentially expressed genes. Therefore, we repeated both the parametric and the non-parametric tests after removing the outliers.

To determine the number of outlier genes for each subject, we counted

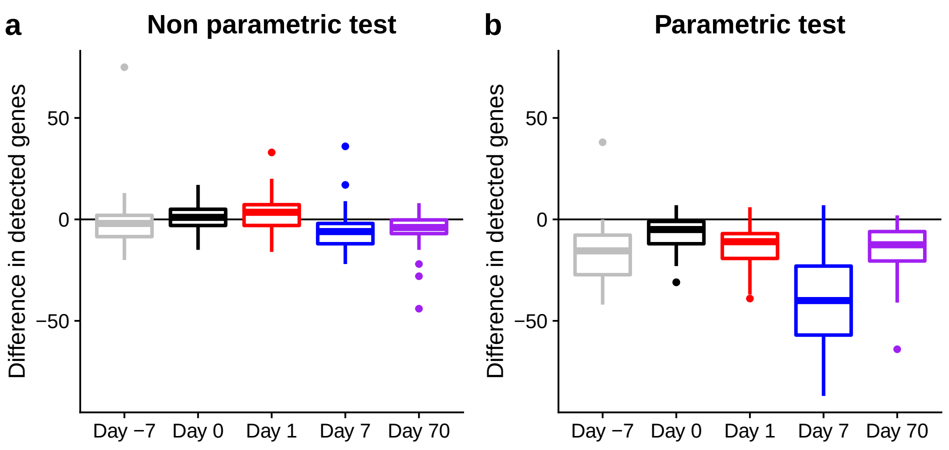


Figure 4.7: Difference in number of DEGs detected by our framework when removing the subject with the highest number of outlying genes (each data point is a subject). In **a** the *median* of the control distributions is used to determine the thresholds for calling a gene modulated, in **b** the *mean* is used.

the occurrences of genes which exhibited a distance higher than 3 standard deviations from the mean distribution of the gene itself across every subject at every specific time point. For the day 0 (pre-vaccination) time point, the number of outlier genes identified across the different subjects ranged from 3 to 18 (median = 7), with a peak of 164 genes in a single subject. After removing this last subject with the highest number of outlier genes, both the parametric and the non-parametric tests for detecting differentially expressed genes were repeated.

The number of DEGs identified by the non-parametric tests, within the individual response profiles, remained fairly close to the original values when outliers were removed (see Figure 4.7). Deviations from the original approach ranged from −7 to +3 genes, with an average of −2 genes not being detected after outlier removal. Considering the median number of DEGs detected in an individual across all time points was 467, the 2 genes differences represents a 0.43% discrepancy. We therefore conclude that the non-parametric approach that we adopted showed to be robust to the presence of outliers. This was in direct contrast with the results obtained from the parametric tests (median across time points of −13, CIQR: −27,−6). This effect was especially evident 7 days post-vaccination: with the non-parametric test we observed a median decrease of 6 genes (CIQR: −12,−2), while the parametric test resulted in an average of 40 less genes detected (CIQR: −57,−23). These results reinforced our decision to present data coming from a non-parametric test without exclusion of subjects from the analysis.

5 Functional characterization of individual subject response profiles

In this section we present the functional characterization of the individual subject response profiles via enrichment analysis on canonical pathways. We also present a novel computational approach aimed at understanding if the pathway enrichment is observed by chance or it is caused by an underlying biological signal.

5.1 Pathway Enrichment Analysis

We wanted to understand whether, despite the heterogeneity in the transcriptome response observed across subjects, there are any biological processes being activated in a more consistent manner in response to vaccination. To this end, we performed a pathway enrichment analysis to functionally characterize each individual subject response profile.

We **obtained the pathways** from the Molecular Signatures Database [73] [74] [75] by downloading the Curated gene sets (*C2 collection*) of *Canonical pathways* [1] (collection version 6.1, 1329 gene sets as of June 21th, 2018). Gene sets in this collection are curated from online pathway databases [2] and biomedical literature, with many sets being also contributed by individual domain experts. We **pre-processed pathway data** by removing gene symbols described in the pathways which were not found among the 5000 genes that were included in our study (see Section 2.1). This led to 2369 gene symbols in common between subject data and pathways (which originally

[1] available at http://software.broadinstitute.org/gsea/msigdb/collections.jsp

[2] BioCarta, KEGG, Matrisome Project, Pathway Interaction Database, Reactome, SigmaAldrich, Signaling Gateway, SuperArray SABiosciences.

included 8904 genes). Furthermore we restricted our analysis to pathways for which we could map between 10 and 200 genes. This led to 473 gene sets to be tested via pathway enrichment analysis (see Table 5.1).

Database	**Raw**	**Processed**
Reactome	674	231
Kegg	186	124
PID	196	88
Biocarta	217	13
Others	56	17
Total	1329	473

Table 5.1: Number of pathways from C2 collection before and after the two pre-processing steps.

Pathway enrichment was assessed through the Fisher exact test and applying the following threshold: p-value ≤ 0.05, one-sided, with Benjamini Hochberg correction for total number of tested pathways across all subjects. The contingency matrix was designed to assess whether the number of modulated genes within a given pathway was higher than random expectation (see Table 5.2). By observing the gene modulations of a subject in relation to a specific pathway, we can understand if there is a significantly higher proportion of modulated genes associated to a specific pathway when compared to the total number of modulated genes within a specific subject.

	Pathway gene	**Other genes**
Modulated in subject	**a**: common	**b**: only subject
Not modulated in subject	**c**: only pathway	**d**: not modulated

Table 5.2: Example of contingency table generated for testing a pathway in a single subject. **a** Genes modulated in the subject which belong to the pathway. **b** Genes modulated in the subject which do not belong to the pathway. **c** Genes not modulated in the subject which belong to the pathway. **d** Genes not modulated in the subject which do not belong to the pathway.

Using this approach we could derive a list of enriched pathways for each subject. The level of conservation of each biological function could then be assessed as the proportion of subjects in which the relative pathways were enriched. Figure 5.1 lists the top 10 pathways found to be enriched with the highest frequency across time points (see Supplementary Tables B.3 - B.7

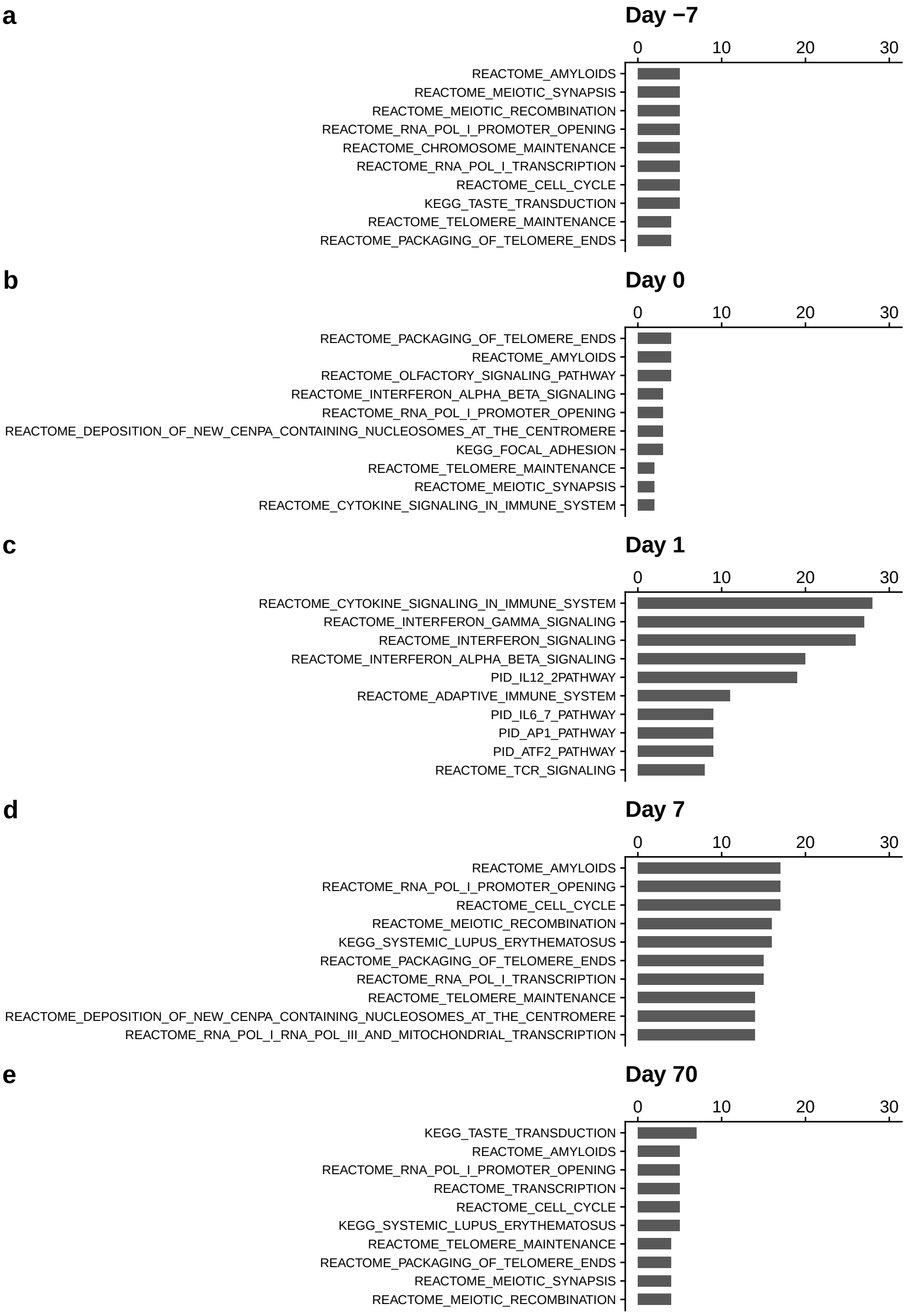


Figure 5.1: Top ten pathways showing enrichment in the highest number of subjects across all time points. *X-axis*: number of subjects in which a pathway is significantly enriched. Days -7, 0 and 70 should be regarded as negative controls.

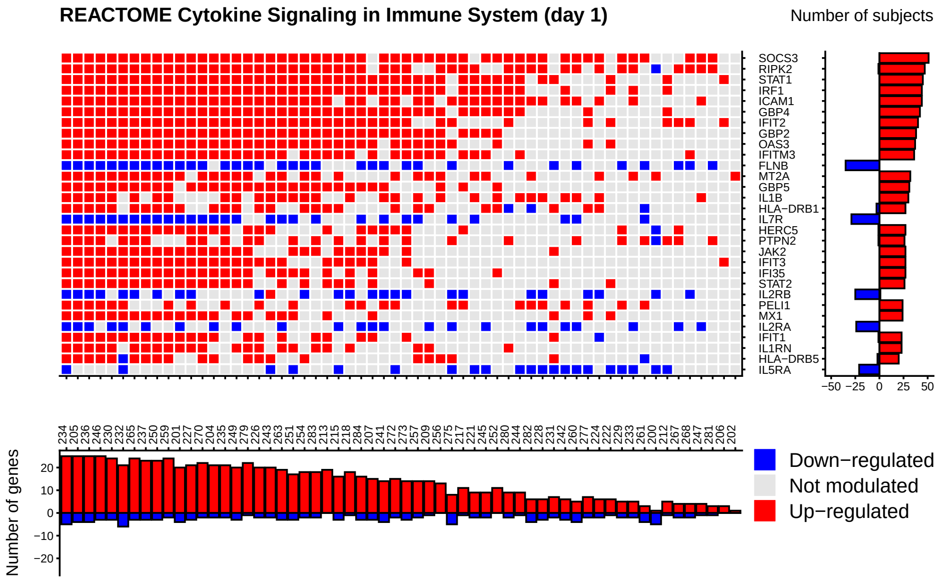


Figure 5.2: Top 30 DEGs (out of 93) of a pathway. Each row corresponds to a gene and each column to a subject. On the right: number of gene modulations observed across all subjects. On the bottom: number of modulated genes for each subject across the pathway.

for the complete list of all enriched pathways). Consistently with what was observed at the gene level, also the canonical pathway activation was variable across the study population. The pathway found as the most frequently enriched at 24 hours, *cytokine signaling in immune system*, was enriched in less than 50% of the subjects (28 out of 60, see Figure 5.1 *c*). Figure 5.2 represents a more detailed example of how gene modulation within a representative pathway varied greatly across the different vaccinated subjects.

Collectively, the pattern of enrichment observed **24 hours** post vaccination pointed to the activation of cytokine signaling pathways, including Interferon-γ and Interleukins 12 and 6. IFN-γ is both a marker of $\mathrm{T_H1}$, CD4, CD8 and Natural Killer cells and an antiviral mediator which, in the context of influenza induces the *MxA* protein, an inhibitor of Influenza A replication [76]. IL-12 is a T cell-stimulating factor and reduces the IL-4 mediated suppression of IFN-γ production [77]. IL-6 is a marker and mediator of ongoing inflammation mediated by the innate immune cells (dendritic cells, macrophages). In the context of an influenza A viral infection it has been shown to be required for viral clearance [78]; and it also plays an important role in modulating and switching the innate immune response towards an adaptive response [79]. **Seven days** post vaccination, we observed less consistent responses: pathways were found to be enriched in less than 30% of the subjects (up to 17 out of 57). Signals pointed to B cell proliferation signals and plasmablasts activity. B cells are key players of the cellular immunity and their replication leads to the generation of high-affinity plasma cells, which secrete high-affinity antibodies, and memory B cells, which confer immunological memory and strong responses to subsequent exposures to the same antigen [18] [20]. Plasmablasts are the B-cell lineage precursors of non-dividing plasma cells. These are antibody-secreting cells which retain membrane bound antibodies and migratory potential [80].

5.2 Evaluating pathways enrichment through a permutation test

We tested whether the pathway enrichment was indicative of a true biological signal or if it was observed by chance. This could be achieved by applying the following bootstrap-based procedure:

- We generate pairwise subject combinations to calculate the Jaccard similarity coefficients based on the number of commonly modulated genes between subjects. This calculation is executed including only the n genes belonging to the pathway.
- For each subject combination, we generate 1000 Jaccard control coefficients by random sampling n genes 1000 times (where n represents the number of genes in the pathway). The bootstrap is executed on all the genes studied in this work, excluding the genes belonging to the pathway currently being examined ($5000 - n$)
- We consider a pathway as significantly enriched for a subject pair if its Jaccard coefficient is higher than the 95% quantile of the control values.

The rationale for this was that genes belonging to the same functional pathway typically show more coordinated modulation when compared to randomly assorted genes. We executed this analysis first by restricting the subjects to those who showed enrichment in a pathway (see Figures 5.3 and 5.4 for days 1 and 7 respectively) and then including every subject in the calculation, regardless of pathway enrichment (see Supplementary Figures A.3 and A.4).

Overall, we observed that despite the substantial response heterogeneity across different subjects, there was a convergence towards more consistent modulation for genes participating to biological processes involved in the vaccination response. The Jaccard similarity coefficients generated from true pathway data were significantly higher (above the 95% Confidence Interval control threshold) when compared to reference control distributions generated through a permutation test. Because of this observation we regarded the results of the pathway enrichment analysis as reliable indicators of the underlying biological processes. As an example, the pathways which were found as the most frequently enriched 24 hours after vaccination (*Cytokine Signaling in Immune System* and *Interferon Gamma Signaling*) displayed values well in excess of the 95% Confidence Interval thresholds, an observation valid not only when looking at subjects which displayed enrichment in those pathways (see Figure 5.3), but also when considering the complete cohort (see Supplementary Figure A.3).

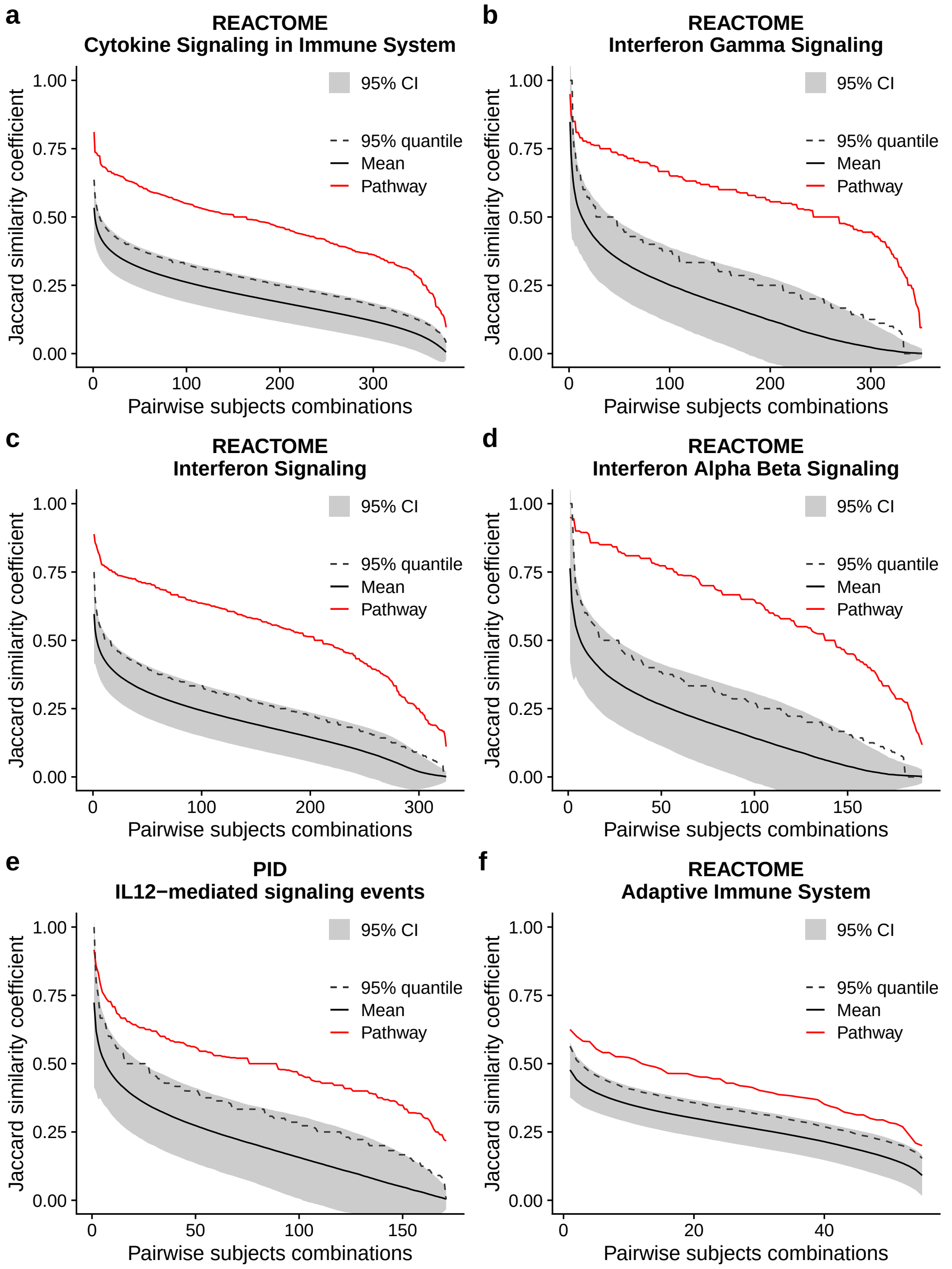


Figure 5.3: Most frequently enriched pathways within **day 1** individual subject transcriptomes. The analysis is executed only on subject pairs showing enrichment for the specific pathway. The Jaccard similarity coefficients associated to the pathway genes are shown in red. The mean of the 1000 control Jaccard indices obtained through the bootstrap procedure is shown in black. The 95% quantile of the bootstrap values is shown through a dashed line. The shadow in gray represents the 95% Confidence Interval of the bootstrap values distribution.

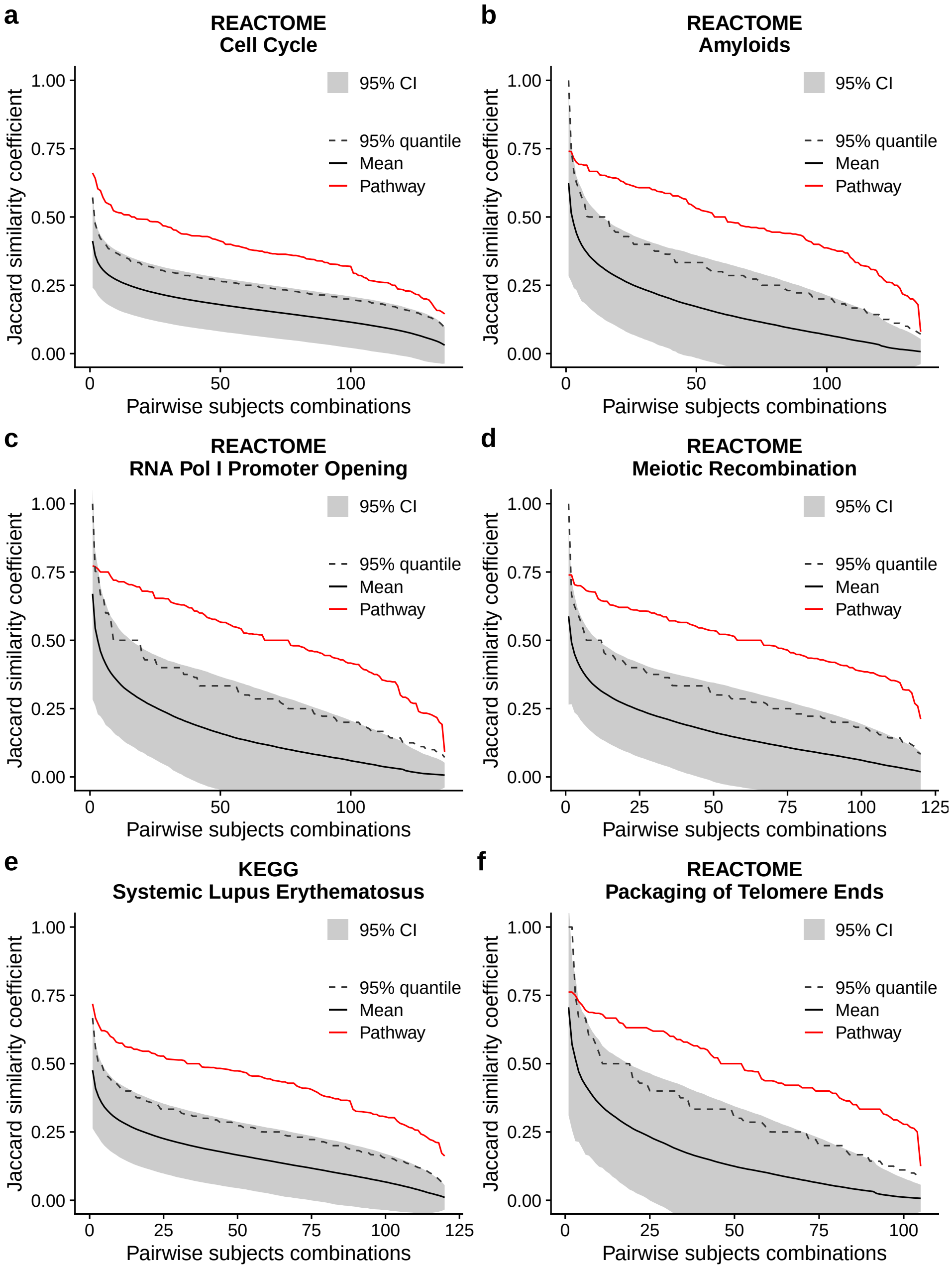


Figure 5.4: Most frequently enriched pathways within **day 7** individual subject transcriptomes. The analysis is executed only on subject pairs showing enrichment for the specific pathway. The Jaccard similarity coefficients associated to the pathway genes are shown in red. The mean of the 1000 control Jaccard indices obtained through the bootstrap procedure is shown in black. The 95% quantile of the bootstrap values is shown through a dashed line. The shadow in gray represents the 95% Confidence Interval of the bootstrap values distribution.

6 Characterization of vaccination-specific modulated gene pools

In this chapter we describe our approach to obtain pools of genes showing a consistent modulation in response to vaccination. This is done by comparing the number of subjects in which genes are differentially expressed with the random expectation of observing gene modulations across the dataset. We also functionally characterize these genes taking into account the seroresponse of the subjects.

6.1 Definition of the pool of vaccination-induced genes

With our framework we established gene modulation comparing the transcript abundance of individual genes against a reference distribution. As such, this approach may be affected by type-I error, which translates into the wrongful detection of modulated genes. To identify the pool of genes that can be assumed to be modulated by vaccination, we computed the probability of a modulated gene to appear in multiple subjects by random chance and compared it to the frequencies of modulation observed in the dataset. Based on this comparison, genes appearing with higher frequency than randomly expected were assumed to be modulated in response to the vaccination. This method was in part inspired by the work from Menche and colleagues (see Section 1.5.2).

We determined the number X of subjects in which a gene must be modulated in order to be included in our vaccination-induced gene pool. To use

X as a threshold for inclusion, we had to establish how many modulations of the same gene can be observed across two or more subjects due to random expectation. In the *null model*, each subject can show g modulated genes drawn randomly out of G total genes. The probability for one gene to be modulated in k out of n subjects can then be calculated using the binomial distribution, where $p = \frac{g}{G}$

$$f(k; n, p) = \Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{6.1}$$

Consequently, it is possible to calculate the number of subjects per gene (and obtain the relative histogram) using the mean number of genes observed by multiplying $G \times (f; n, p)$. Below, we outline the individual steps used to execute this analysis in our work. The calculations are executed separately for each time point.

- The R function `rbinom(n, size, prob)` is used to generate 1000 binomial distributions with the following parameters:
  - the number of observations n is set to 5000 (number of genes under examination)
  - the number of trials (*size*) is set to the number of available subjects at the specific time point under analysis
  - the probability of success on each trial (*prob*) is set to the mean value of modulated genes across subjects divided by the total number of genes
- Each of the 1000 distributions is used to infer the minimum number of subjects X_{min} in which the number of shared modulated genes across subjects is zero.
- The threshold X for inclusion in the vaccination-induced gene pool is defined by selecting the 95^{th} percentile of the computed X_{min} values. This allows to identify genes with a probability of being modulated across multiple subjects of less than 5%.

The applied combinatorial method resulted in the following thresholds: $X_{\text{day 0}} = 16$, $X_{\text{day 1}} = 25$, $X_{\text{day 7}} = 19$, $X_{\text{day 70}} = 17$, meaning that to be included in the vaccination-induced gene pool, a gene must be modulated in more than 16, 25, 19 and 17 subjects for days 0, 1, 7 and 70, respectively. A gene exceeding these thresholds can be reasonably attributed to the vaccine-induced gene response.

This allowed to identify a core pool of 341 genes for day 1 and 135 genes for day 7 (see Figure 6.1 and Supplementary Tables B.8 - B.9). We applied the same test to day 0 response profiles (observed before vaccine administration) as negative control and we did not identify vaccination-associated genes. In accordance to our previous observations, also within these gene pools the response appeared to be more robust at day 1 compared to day 7. At day 1 we observed a median number of subjects in which a gene was modulated of 30 (50% of the 60 subjects, CIQR: [1] 27-35), whereas at day 7 the median number was of 25 (44% of the 57 subjects, CIQR: 21-30). In accordance with our previous observations, modulation of the two vaccination-induced gene pools was heterogeneous across the cohort: while some subjects displayed almost complete modulation across the pool, others showed limited response.

24 hours post vaccination we found modulated genes which pointed to inflammation and antiviral responses, in accordance to what we previously observed through pathway enrichment analysis (see Section 5.1). As an example, we observed IRF1 and STAT1 (modulated in 45 and 44 subjects out of 60, respectively), which are involved in regulatory pathways that enhance the expression of mononuclear chemokines [81]. IRF1 is an interferon regulatory factor which sustain transcriptional response. Its transcription is activated both by INF-γ activated STAT1 and TNF-activated NFκB, which bind to their respective promoters of IRF1 [82]. SOCS3 was observed as upregulated with the highest frequency across the cohort (51 subjects out of 60). It is induced by various cytokines (such as IL-6, IL-10 and INF-γ) and acts as a suppressor of cytokine signaling. It is induced during influenza infection to regulate the production of pro-inflammatory cytokines. Influenza A virus infection can also actively enhance upregulation of SOCS3, contributing to excessive production of IL-6 during infection, aiding viral replication [78]. The previously described CXCL10 gene (see Section 3.3) was also observed in the day 1 gene pool. The modulated genes found in the day 7 pool were consistent with the signals observed through pathway enrichment

[1] Central InterQuartile Range

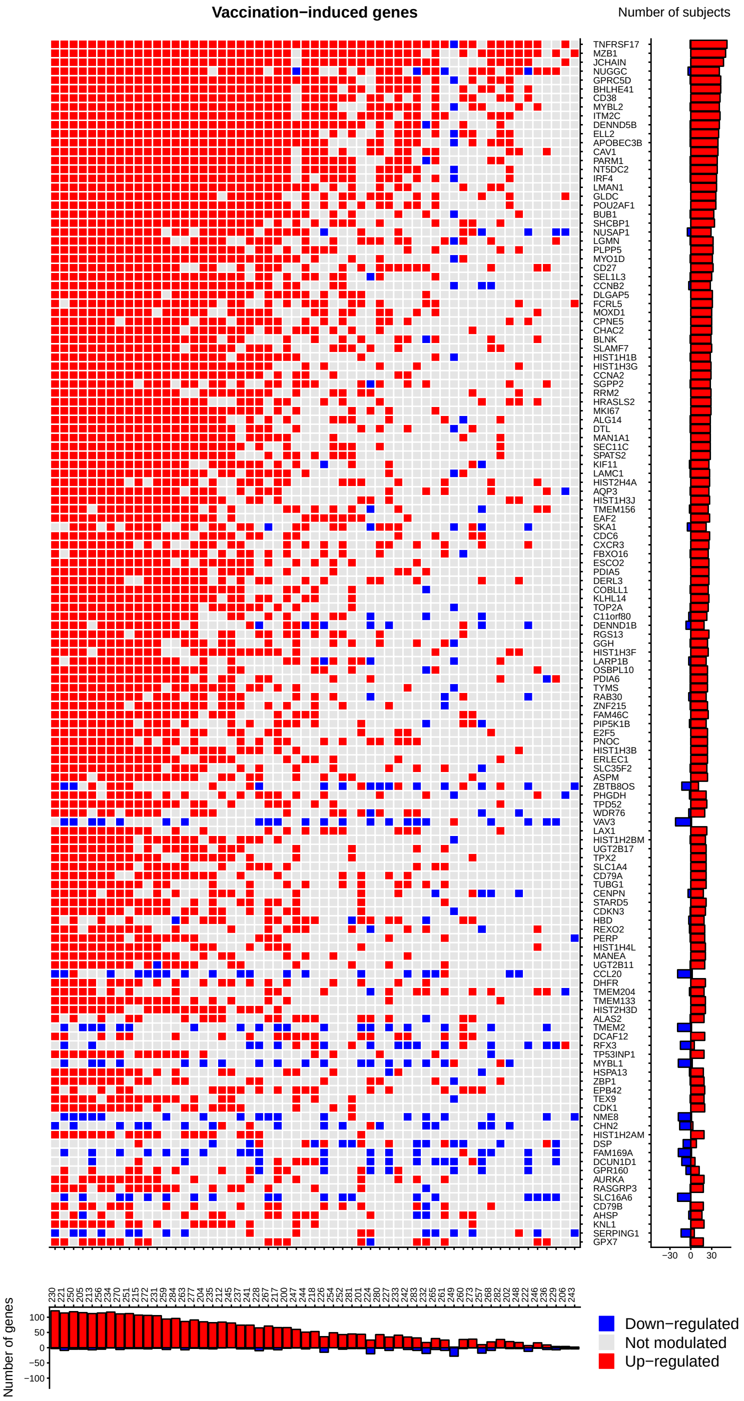


Figure 6.1: Vaccination-induced gene pool (day 7, 135 genes). Each row corresponds to a gene and each column to a subject. On the right: number of gene modulations observed across all subjects. On the bottom: number of modulated genes for each subject across the gene pool.

analysis. Briefly, we observed signals pointing to B and T cell proliferation, plasmablasts activity and signals pointing to immunoglobulin production (further details will be described in Section 6.3).

6.2 Classification of subjects into high- and low responders

In line with the approach from Tsang et al. [1], we stratified the study subjects into *high* and *low responders to vaccination*, based on their day 70 sera ability to neutralize the virus. This was done using the influenza virus MicroNeutralization (MN) assay data (see Section 2.1.2). Tsang and colleagues [1] defined high and low responders by selecting the subset of subjects which were below the 20$^{\text{th}}$ and above the 80$^{\text{th}}$ percentile mark of the metric. We chose to follow a different approach in order to retain the majority of the subjects in the classification process: subjects were split into the two classes using the median value of MN titers (3.71); subjects with MN titers above the median value were classified as *high responders*, while those with MN titers below the median were classified as *low responders.* Taking into account the limited study data availability of MN titers, we could classify and include in subsequent analyses 48 and 46 subjects from days 1 and 7 respectively, out of the 60 and 57 total subjects which we previously observed through transcriptome and pathway enrichment analyses.

6.3 Functional analysis of the day 7 vaccination-induced gene pool

Due to previous knowledge [15] [21] [46] [1] regarding the expected time frame of the adaptive immune responses, we focused our analysis on the day 7 vaccination-induced gene pool and tried to identify potential genes whose response could be predictive of the functional antibody response to vaccination. Starting from the 135 vaccination-induced genes we were able to identify a **subpool** of 66 genes that were specifically modulated in the *high responder* subject class (see Supplementary Table B.8). This was done using the Fisher exact test (one sided, p-value < 0.05, Benjamini Hochberg correction) using the contingency Table 6.1. By observing the modulations of a gene in relation to the subject class we can test if there is a significantly higher proportion of modulated genes within a specific class with respect to the total number of modulations observed across all subjects.

	Modulated gene	**Non modulated gene**
High responders	a	c
Low responders	b	d

Table 6.1: Contingency table used to test for genes associated to the *high responder* subject class via Fisher test. A table is generated for each of the 135 genes from the vaccination-induced day 7 gene pool. **a - b** Instances in which a gene is modulated in the *high* or *low responder* subject class, respectively. **c - d** Instances in which a gene is not modulated in the *high* or *low responder* subject class, respectively. The sum of **a**, **b**, **c** and **d** is the total number of subjects.

Clear transcriptional signals coherent with the immune response, specifically the B-cell mediated humoral response, could be detected in the **complete day 7 vaccination-induced gene pool** of 135 genes. We observed upregulation of TNFRSF17 in 52 out of 57 subjects. This is known as the B-cell maturation antigen, which is predominantly expressed on differentiated B cells and mediates the survival of plasma cells involved in long-term immunity [83]. MZB1 was also detected as positively modulated in 50 subjects. This gene encodes for the protein of the same name which regulates IgM assembly and secretion [84] [85].

Regarding the **high-responder subpool** of 66 genes obtained through the Fisher test shown above, we observed up regulation of the JCHAIN gene in 47 subjects, this encodes the joining chain of multimeric IgA and IgM [86] and it is transcribed during B- and T-Cell lymphopoiesis [87]. Also modulated (40 subjects) is the ELL2 elongation factor for RNA polymerase II. This factor drives the production of mRNA specific to immunoglobulin secretion [88] by enhancing both polyadenylation and exon skipping in the gene encoding the immunoglobulin heavy-chain complex [89]. We detected up regulation of ERLEC1, an endoplasmatic reticulum lectin that functions in N-glycan [2] recognition [90] and protects immature polypeptides from degradation [91]. POU2AF1 (also positively modulated) restricts the ubiquitous

[2] N-linked Glycosylation: biosynthetic process that regulates the maturation of proteins through the secretory pathway.

activity of its promoter in B cells [92] and it is required for the VDJ recombination of immunoglobulin κ genes [93]. We observed both CD38 and IRF4 genes upregulated in 42 and 37 subjects respectively. The first orchestrates Th1 response, migration and survival [94], while the second sustains CD8$^+$ T cell differentiation and cell expansion [95] and more broadly coordinates T helper cell fate [96]. Lastly, CAV1 was up regulated in 39 subjects. In B cells it controls the distribution of B cell antigen receptors on the cell surface [97], while in T cells it regulates T cell receptor signal strength and T-cell differentiation into alloreactive T cells [98].

7 Association of transcriptome data with humoral immune response

In this section we show how the magnitude of vaccine-induced transcriptome modulation, defined as the proportion of modulated genes in each subject, is associated with the seroresponse assessed 70 days after vaccination. We also provide evidences suggesting that latent variables, like pre-existing immunity to the vaccine, age and gender do not have a major impact on the modulation of the peripheral blood transcriptome. Finally, we describe the construction of a Machine Learning model which allowed to identify genes that are predictive of subjects' immune response to vaccination.

7.1 Day 7 transcriptome response is associated with the functional antibody response

In this part of the study we investigated whether the transcriptome response to the vaccination could be linked to subjects' seroconversion observed 70 days post vaccination. Starting with the pools of vaccination-induced genes (defined in Chapter 6), we tested whether the proportion of modulate genes within each subject was correlated to the day 70 functional antibody response. The percentage of modulated genes for each subject was used as as predictor of the antibody response. Day 1 and day 7 transcriptome profiles were analyzed separately. Antibody responses, defined as the Influenza MicroNeutralization (MN) titers measured 70 days post vaccination, were used to stratify the study subjects into two discrete classes: *high* and *low responders*. As outlined in Section 6.2, subjects with MN titers above the median

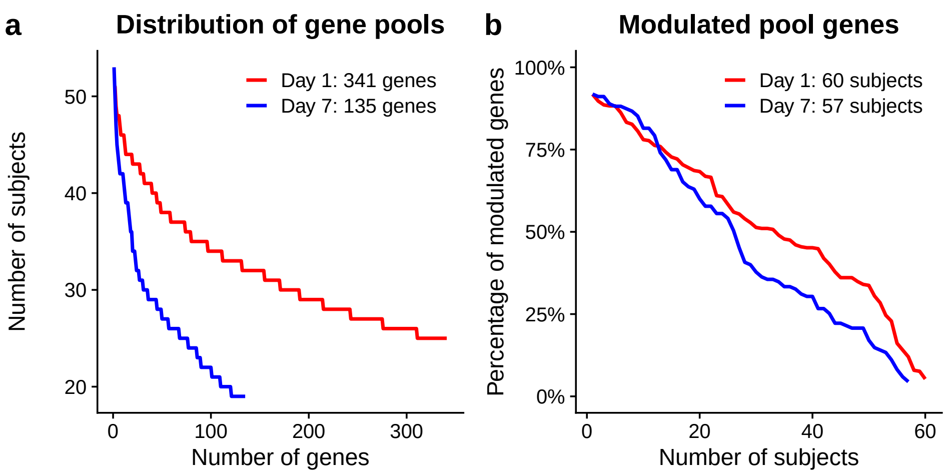


Figure 7.1: **a** Reverse cumulative distributions of the number of subjects showing modulation of the genes belonging to the day 1 and day 7 pools. Y-axis: subjects in which n genes (on the X-axis) are observed. **b** Percentage of modulation observed in the pools across each subject. In both plots, subjects are ranked in a decreasing Y-axis value manner, disregarding subject identity across time points.

value of the cohort were classified as *high responders*, while those with MN titer values below the median were classified as *low responders*.

Figure 7.1 shows the number and percentage of genes belonging to the vaccination-induced gene pools which were found modulated in the individual subject transcriptome profiles. Day 1 individual subject response profiles showed an overall higher number of modulated genes compared to the day 7 transcriptome responses. The number of modulated genes, at day 1 was of 175 (median number of genes across the individual response profiles) out of 341 (CIQR: 123-247), whereas day 7 profiles showed a median of 54 modulated genes out of 135 (CIQR: 24-93, see Figure 7.1 *a*). In relative terms, this translated in 51% modulated genes (median; CIQR: 36-72%) for day 1 and 41% modulated genes (median; CIQR: 25-68%) for day 7 (see Figure 7.1 *b*).

We compared the magnitude of transcriptome modulation, defined as the percentage of differentially expressed genes in the vaccination-induced gene pools, with the *high* and *low responder* subject classification. This highlighted differences between day 1 and day 7 transcriptome responses; while day 1 individual response profiles did not differ significantly between

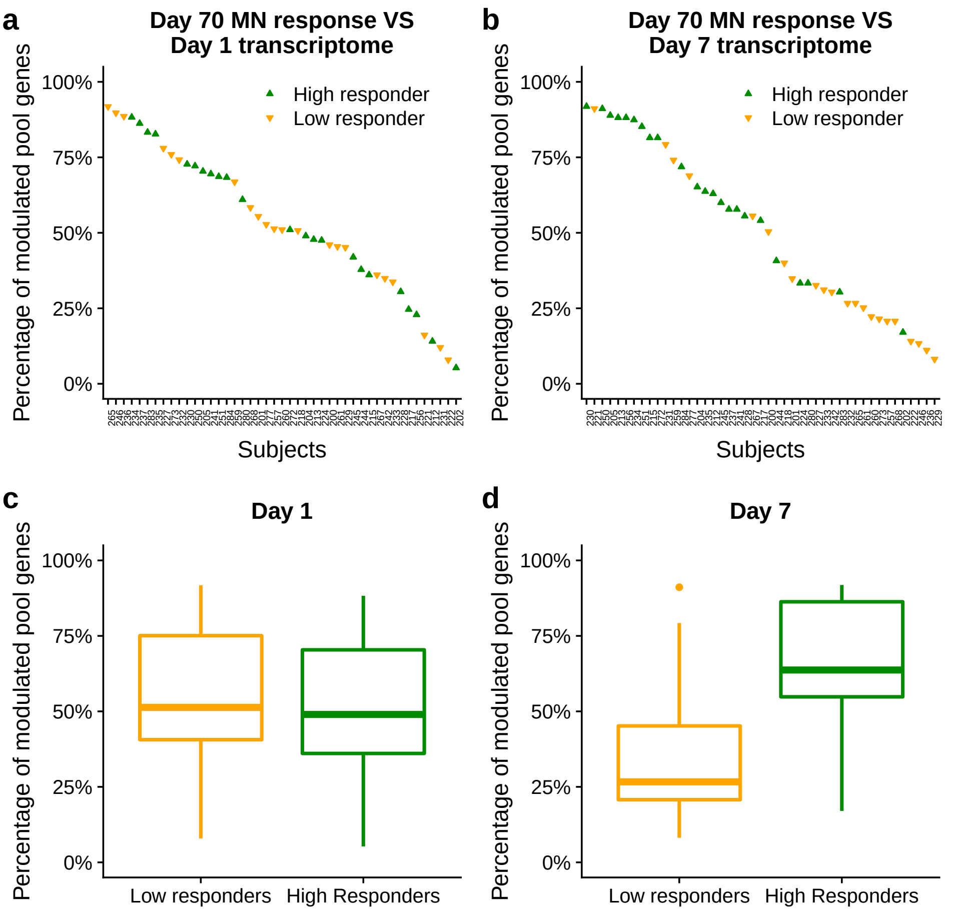


Figure 7.2: **a-b** Percentage of modulation in the day 1 and day 7 vaccination-induced gene pools shown in relation to the immunological response classification based on day 70 Influenza MicroNeutralization titers. **c-d** Distributions of vaccination-induced gene pools modulation percentages grouped by immunological response classification.

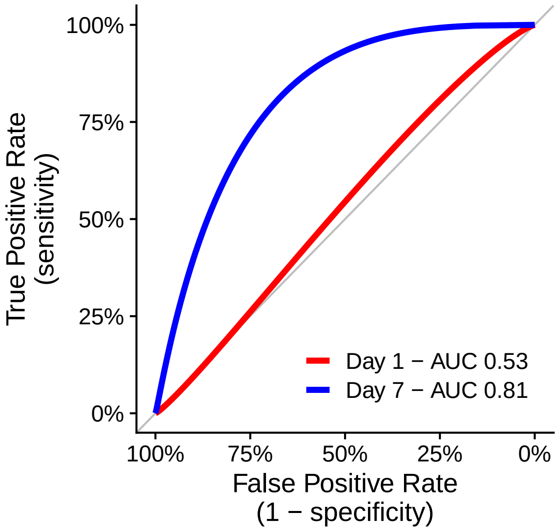


Figure 7.3: Smoothed ROC curves of the immunological response classification predictions inferred using the modulation percentages of the vaccination-induced gene pools calculated from the individual subject transcriptome response profiles.

the two subject classes, day 7 profiles showed a remarkably higher modulation within the *high responders* subjects (see Figure 7.2). Specifically, at day 7 the median modulation of the vaccination-induced gene pool in subjects classified as *high responders* was 63% (CIQR: 55-86%), while in *low responders* we observed a median value of 27% (CIQR: 20-45%). The difference was statistically significant (day 7 p-value = 0.002, day 1 p-value = 0.964, two-sided Kolmogorov-Smirnov tests, Benjamini-Hochberg correction for multiple testing). This trend was not observed at day 1, in which *high responders* displayed a median modulation of 49% (CIQR: 36-70%) and *low responders* exhibited a similar centrality of 51% (CIQR: 41-75%).

Based on these observations, we next assessed the ability of day 7 responses to stratify the subjects into *high* and *low responders* (see Figure 7.3). We built the ROC curve [1] and calculated the AUC [2] using the R package pROC [100]. The percentage of modulated genes within each individual subject was chosen as predicting variable, while the subject class (*high*/*low*

[1] Receiver Operating Characteristic curve. It is a graphical representation that illustrates the diagnostic ability of a binary classifier [99].

[2] Area Under the Curve: measure of the ability of the test to correctly discriminate between the binary classes, with a value of 1 being a perfect classification.

responder) obtained from day 70 MN titer data was the target variable to predict. We observed that day 7 transcriptome data was able to discriminate between the *high* and *low responder* subject classes, whereas day 1 data was not (AUCs 0.81 and 0.53 respectively). This result was in agreement with the general notion that day 7 blood-derived signatures reflect the establishment of the humoral immune response.

7.2 Subjects' age, gender and pre-vaccination status have no influence on whole blood transcriptome responses

We investigated whether the intrinsic variables (age, gender and pre-existing immunity) had an effect on the whole blood transcriptome response. For this purpose, we observed the proportion of modulated genes in the vaccination-induced gene pools [3] in relation to the data from the intrinsic variables.

We were unable to observe any relevant association between subjects' **age** and the percentage of gene pool modulation, both at days 1 and 7 ($\rho =$ -0.09, -0.17; p-value = 0.49, 0.44 for days 1 and 7 respectively, Spearman's Rank Correlation Coefficient, see Supplementary Figure A.5 a-b). Similarly, also subjects' **gender** was found not to be associated with gene modulation (ROC's AUCs = 0.57, 0.62 for days 1 and 7 respectively, see Supplementary Figure A.5 c). We examined **pre-existing immunity** (see Section 2.1.2) in relation to the proportion of modulated genes in the vaccination-induced gene pools. We stratified subjects based on their pre-existing immune status using MicroNeutralization titers data. Subjects with values below the limit of detection were defined as *not pre-immunized* to a viral strain and all other subjects as *pre-immunized* (see Table 7.1). Throughout the data we observed a prevalence of pre-existing immunity to the seasonal influenza strains (A/Brisbane, A/Uruguay, B/Brisbane), while the majority of the subjects did not exhibit pre-existing immunity to the pandemic strain (A/California). We attempted to predict the *pre-immunized* and *not pre-immunized* subject classes using the percentages of vaccination-induced gene pool modulation at days 1 and 7 (see Supplementary Figure A.6). The percentage of pool modulation was unable to predict the pre-immune status of the subjects:

[3] See Chapter 6 for gene pool definition.

Viral strain	No	Yes	Day 1 AUC	Day 7 AUC
A/California/07/Swine	42	15	0.57	0.54
A/Brisbane/59/07	17	40	0.51	0.59
A/Uruguay/716/07	7	50	0.59	0.71
B/Brisbane/60/2001	19	38	0.58	0.75

Table 7.1: Pre-existing immunity to viral strains based on Microneutralization Titer data and ROC AUCs obtained by associating the percentage of modulation of the vaccination-induced gene pools to pre-existing immunity data. **Yes**/**No**: presence/absence of pre-existing immunity. **Day 1 AUC** / **Day 7 AUC**: AUCs obtained when the percentages of modulation observed in the vaccination-induced gene pools were associated to the pre-existing immunity.

ROCs for both days 1 and 7 exhibited AUC values between 0.51 and 0.75 (see Table 7.1 and Supplementary Figure A.6). It is important to note that due to class imbalance [4] a high enough AUC would still not have been a sufficient reason to claim the association between the two parameters.

7.3 Identification of candidate genes involved in the magnitude of the immunological response

We have already described how the percentage of modulated genes which belong to the day 7 vaccination-induced gene pool was associated with the serological response observed 70 days post vaccination (see Section 7.1). To further examine this association, we built a Machine Learning model. This enabled us to test the ability of the transcriptome response to predict antibody response and to gather more detailed information on which genes are the most informative. This aids the formulation of biological hypotheses regarding the mode of action of the vaccine.

[4] This occurs when the class distribution is imbalanced. To avoid many machine learning algorithms to have low predictive accuracy for the infrequent class, a cost-sensitive approach must be employed. In addition, undersampling/oversampling of frequent/infrequent classes may be required.

7.3.1 Random Forest

Random Forest (RF) is a machine learning approach based on **decision tree learning**. A decision tree is a flow-chart-like structure with the following main components:

- *node*: represents a test on an attribute (i.e. is a gene modulated?)
- *branch*: originates from the node, corresponds to the outcome of a test on an attribute (i.e. the gene is modulated / the gene is not modulated)
- *leaf*: terminal node of a decision tree. It holds a class label, which is the decision taken by the tree after executing all the tests that led to that particular leaf (i.e. the subject is a high / low responder).

A tree is built by splitting the training data (which constitutes the root node of the tree) into random subsets, this is called **bagging**. The splits of the tree are then defined using a random subset of the classification features (the genes in our study, see Figure 7.4 a).

The **RF algorithm** [101] generates multiple individual decision trees (the forest, see Figure 7.4 b). Each tree is generated by randomly selecting subsets of the training data (bagging, described above). For each node of the trees a predefined number of all predictors is randomly sampled in order to choose the best split from among those features [102]. In doing this, the algorithm assigns a score to the features, identifying those with higher predictive power. This process is what defines **feature importance**: the RF algorithm estimates the importance of a feature by assessing how much prediction error increases when data for that feature is permuted, leaving all the others unchanged [102]. In the context of this work, these scores allow to identify those genes which higher discriminant (or predictive) power.

The multiple trees generated by the algorithm operate as an ensemble: where a single tree would perform poorly in generalizing the classification problem, an ensemble is able to overcome this limitation. A large number of trees operate as a committee in the classification problem and the final prediction of the model will be based on the *majority vote* obtained from the whole forest.

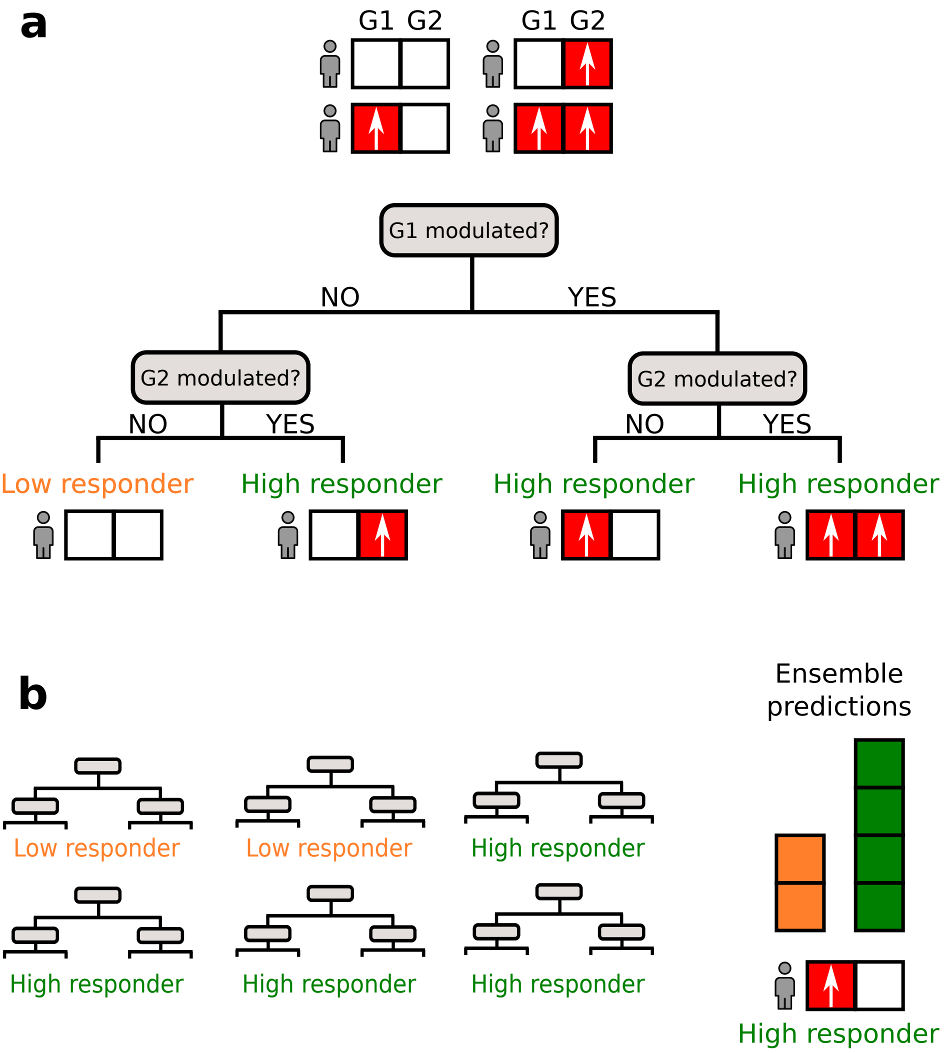


Figure 7.4: **a** The various trees of the Random Forest are generated using training data by splitting subjects using a random subset of features (genes). The final classification is already known from the training data. **b** An ensemble of trees is used to classify subjects from the test data. The *majority vote* of the whole forest determines the final outcome.

7.3.2 Guided Random Forest

In Guided Random Forest (GRF), first a regular RF is built. Next, another run of the algorithm is executed, with the difference that the features are weighted using the importance scores calculated for each feature by the regular RF [103]. The features that are less important in the regular RF are penalized in the GRF. The results from this model are expected to be relevant for the classification problem, but not necessarily non-redundant. Relevancy consists in the fact that the identified genes will be those able to discriminate between subject classes. Non-redundancy stems from the fact that two different genes might predict with the same predictive power the subject class, but the GRF is not able to define if those genes are both needed to predict the subject class, or if only one of them (and which one) is needed.

7.3.3 Guided Random Forest model construction

We decided to utilize the Guided Random Forest approach due to its reported superior performance compared to RF ([103], also observed by us). We used the implementation found in the `RRF` R package [103] [104] [105]. Genes from the individual response profiles were utilized as predictive variables by considering their modulation status (*modulated* vs *non modulated*) and not taking into consideration the direction of the modulation (*up-regulation* vs *down-regulation*). The robustness of the prediction was assessed using a 5-fold cross-validation. This was achieved by splitting the dataset into five stratified [5] subsets (with comparable proportion of *high* and *low responders*) using the `caret` R package [106]. Within each cross-validation run, four datasets were iteratively used for training the models, while the remaining one was used for testing (see Figure 7.5).

To build the five GRF models, we initially performed a **parameters optimization** procedure. Through this procedure we were able to define the optimal number of trees needed for each GRF and the adequate number of features (genes) to use for each split along the trees. By running a test

[5] Stratified sampling: the population is divided into homogeneous subgroups before sampling. Each subgroup is then sampled by a fraction which is proportional to that of the total population. This allows in our case to have the same proportion of *high* and *low responders* in each of the five splits.

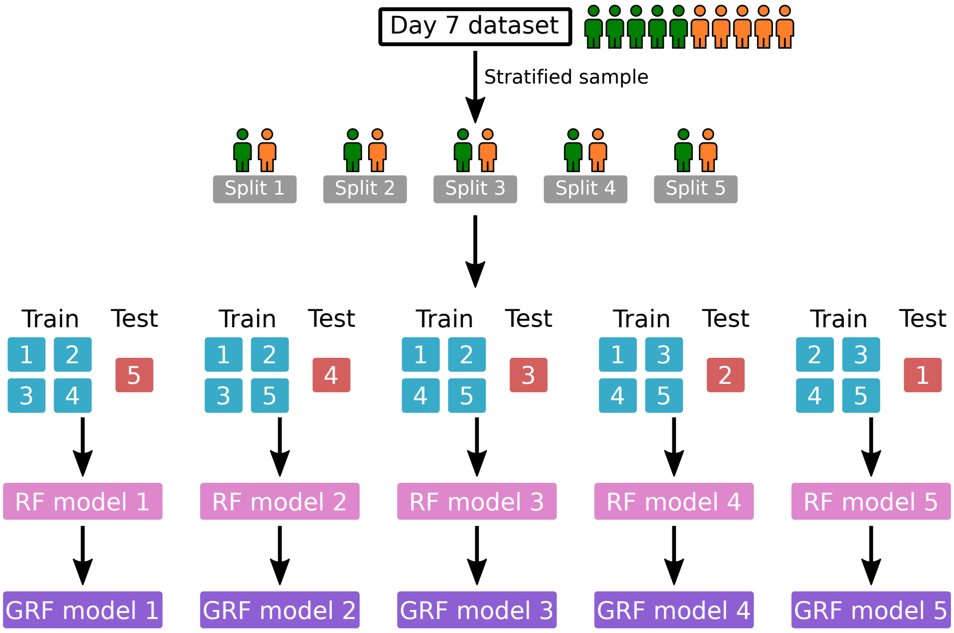


Figure 7.5: 5-fold cross-validated Guided Random Forest approach. From the full day 7 dataset a 5-fold stratified sample is executed. The resulting five splits are used to iteratively train and test five Random Forests. The Guided Random Forests are then obtained by weighting the features with the coefficients obtained from the Random Forests.

GRF we found 3×10^4 trees to be the optimal value to achieve stabilization of the Out Of Bag (OOB) error [6] without incurring into overfitting across all five models. The number of features to randomly sample at each split of the decision trees was set as the square root of the total number of features (135 genes from the day 7 vaccination-induced gene pool, $mtry = \sqrt{135} \simeq 12$). This corresponded to the default suggested value. Subsequently, we screened a range of these values ($1 \leq mtry \leq 40$) by building multiple models and evaluating the mean OOB error. The lowest errors were observed at values 4, 9 and 21 (mean OOB error of 19%, 22% and 26% respectively). In the main text we present the results obtained with $mtry = 4$, while the runs with $mtry = 9$ and 12 are shown in the Supplementary Figures A.7 and A.8. We obtained five GRF models with a mean classification accuracy of 80% (100%, 80%, 78%, 67% and 78% across models, with AUCs of 100%, 72%,

[6] Out-of-bag error. Since during the training procedure the RF algorithm generates each tree using a random sample of the original subjects (bagging), it is possible to test for errors in model prediction by using the subjects which were not selected to build that particular tree. OOB is the mean prediction error on each training sample [107].

70%, 75% and 70% respectively, see Figure 7.6 a-b).

Feature selection was based on the *mean decrease Gini* [7] as a measure of feature importance. [8] We calculated the median value of the *mean decrease Gini* for each gene across the five GRF models to rank the features. To understand which genes were driving the classification of the GRFs, we examined the top 10 informative features of each model and assessed their stability across models (see Figure 7.6 c-d). 3 of these genes were shared across all five models, 6 genes were observable in four out of five models and 1 was found in three models. GLDC was one of the most informative genes in driving subject classification across models as it was consistently observed in all GRF models. It encodes a mitochondrial glycine decarboxylase, which confers survival advantage to T cells in hypoxic environments such as sites of infection [109] [110]. HIST1H3G is a replication-dependent histone protein which peaks in expression during the S-phase [111] and it was found consistently in all five models. We also observed ZBP1 (4 out of 5 GRFs), a cytosolic DNA sensor which induces the DNA-mediated interferon response [112] and APOBEC3B (3 out of 5 models), a cytosine deaminase induced by viral infection [113]. Among these genes we also observed both IRF4 and POU2AF1 (4 out of 5 models, we described their functions Section 6.3). Looking at these genes, there are evident signals pointing at B and T cell replication and interaction in the context of an immune response to infection. Additionally, the POU2AF1 gene is required for the VDJ recombination of immunoglobulin κ genes [92]. This could be indicative of an immunoglobulin affinity maturation process which is primarily observed within the *high responders* group. Overall, the results obtained through our machine learning approach suggests that the highest humoral immune response of *high responder* subjects, measured 70 days post vaccination, could be the result

[7] The mean decrease in Gini coefficient is a way of measuring feature importance. The more the accuracy of the random forest decreases due to random assignment (or exclusion) of a feature, the more the feature itself is regarded as important. Features with a large mean decrease in accuracy (higher Gini coefficients) are regarded as important for the classification problem [108].

[8] It has been argued by Strobl and colleagues [108] that suboptimal predictor features may be artificially preferred when using the Gini index as a selection criterion. This problem occurs for features with a high number of categories when compared to other features, with different scales of measurement and if continuous features are present (these are preferred over the categorical features). In this study we are not concerned by these problems because we are not using continuous features and because our number of possible categories are all equal, since data is all discretized into 0 or 1 (absence/presence of modulation respectively).

of a more robust B cell activation and affinity-maturation process occurring as early as 7 days after vaccine administration. This is in line with what has been proposed by other independent studies [15] [21] [46] [1].

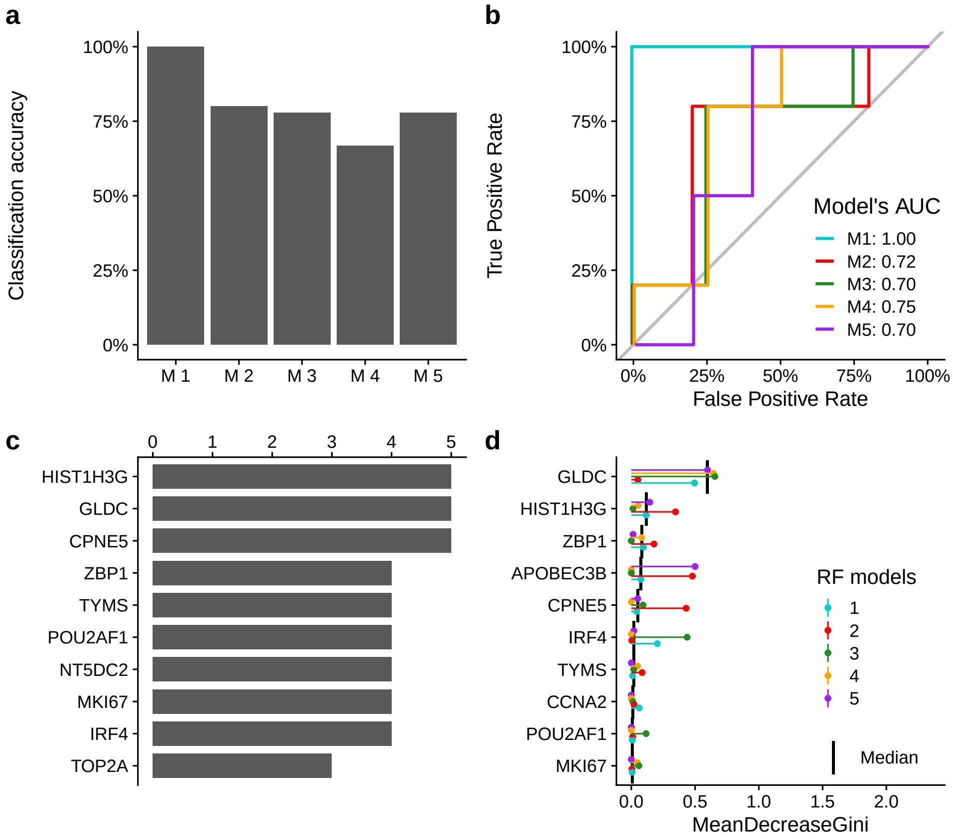


Figure 7.6: Guided Random Forest results ($mtry = 4$). **a** Classification accuracy of the models (percentage of correctly classified subjects). **b** ROC curves and AUCs of the models. **c** Number of occurrences across the five models of the most informative genes. From each model, the top 10 most informative features are reported. The y-axis indicates the number of models in which a gene was observed in the top 10. **d** Top 10 most informative genes across GRF model ordered by median of the *mean decrease Gini* coefficients (higher values correspond to higher feature importance).

8 Discussion

It is increasingly believed that group-wise analyses of perturbations on a complex system may not be suitable to fully capture the range of responses exhibited by individual subjects. In the context of vaccinations, suboptimal responses from a subset of subjects may be determined by both heritable and non-heritable influences. However, demographic and socioeconomic differences could render age classes or even entire populations not able to gain a suitable degree of protection from a particular vaccine formulation. Group-wise methods are optimal for capturing signals that are consistently found in most tested subjects, but they are not sensitive enough for detecting subtle changes, where variance and noise can hide the signal. Additionally, modulations occurring in a small proportion of tested subjects are also generally lost.

The **aim of our work** was to extract, from vaccine response data, a higher amount of information compared to canonical group-wise methods. For this purpose, we developed a new data analysis framework for the generation and functional characterization of individual subject transcriptome response profiles and benchmarked it on a non-adjuvanted H1N1 vaccination dataset. We were in part inspired by Menche et al. [2], which developed an approach to describe subject heterogeneity among cohorts of disease-vs-healthy subjects. In this work we instead described variability in the same cohort across multiple time points. The individual transcriptome response profiles resulting from our pipeline are arrays of categorical gene modulation data and as such we believe these objects to be easily interpretable and comparable across not only different experimental conditions, but also distinct studies.

As an example we have shown the CXCL10 gene, which was among the most strongly modulated genes in response to the two subunit influenza vaccines, to be inconsistently modulated across subjects, a property shared by every gene in the dataset. With our approach it was possible to accurately assess its response variance as well as identify the subset of subjects in which this gene was modulated. Generalizing, it was possible to decipher the **transcriptome response pattern of each subject** and to observe characteristics that were not captured by conventional group-wise analyses.

This result was achieved by developing a **bioinformatic pipeline** aimed at deriving individual transcriptome profiles based on temporal differences in genes' modulation. The pipeline analyzes transcriptome data on a gene-by-gene basis to derive which genes show a temporal behaviour indicative of a perturbation induced by a treatment (in the context of our study, influenza vaccine administration). By generating a control distribution of gene expression values from the whole study subjects' data prior to vaccination, we could assess the variance of the transcript abundance across subjects and define threshold values which a gene should exceed in order to be considered modulated by the treatment. However, gene expression is a biological process and as such it has a degree of stochasticity associated to it. This characteristic can lead to discover not only modulated genes, but also a number of false positives. We compensated for this by implementing a bootstrap step in our pipeline. The bootstrap allows to repeat the test for each gene multiple times and provides a quantitation of how reliable the call of a modulated gene is. After evaluating the transcriptome response as deviations from multiple reference distributions, we continued the analysis by taking into consideration the discrete modulation status of a gene. Meaning that each gene was considered as up-regulated, down-regulated or non-modulated, without taking into consideration the actual quantitative changes in transcript abundance values. This allowed us to employ analysis techniques only suited for categorical data. Finally, such categorical data may be easily compared across studies.

By analyzing the **individual subject transcriptome profiles** obtained with our pipeline, we found the day 1 responses to be more consistent than those observed at day 7, as indicated by the higher Jaccard similarity among subjects. This temporal decrease in similarity may be interpreted as the result of cumulative differences that add up along the chain of immunological events triggered by vaccination. Despite significant overlaps across the individual perturbation profiles, no gene was found to be constitutively expressed in all vaccinated subjects. This last consideration reinforces the hypothesis that group-wise analysis approaches may not fully capture modulations that occur in a subset of analyzed subjects and it suggests that such events occur regularly and, as such, should be accounted for.

We wanted to understand whether, despite the observed heterogeneity in our transcriptome profiles, there were any biological processes being consistently activated in a in response to vaccination. Thus, we examined our profiles in relation to metabolic molecular **pathways**. Menche and colleagues [2] hypothesized that differentially expressed genes in the same biological pathway may lead to similar phenotypical outcomes in the context of diseases, regardless of the identity of the pathway's gene. They reached this conclusion by observing higher similarity in pathway enrichment across subjects when compared to individual gene perturbations. From the results of our work, this seems to be applicable also in the context of immune system's perturbations. Overall we observed that despite the substantial response heterogeneity across different subjects, there was a convergence towards more consistent modulation for genes participating to biological processes involved in the vaccination response. From this analysis we observed the activation of cytokine signaling pathways (including INF-γ, IL2 and IL6) 24 hours post vaccination and cell proliferation and activation signals (arguably a signature of plasmablasts activity) seven days after the vaccine administration.

Subsequently, we wanted to determine whether our gene modulation data could be associated with the seroresponse observed through the MicroNeutralization titers data obtained 70 days following vaccination. Above we have described our supervised approach, which relied on predefined pathway annotations, to obtain signals that could point to the biological functions which may have been activated by the vaccination and may lead to the development of adaptive immunity. However, we wanted to precisely define which genes were modulated by the vaccines, so we opted for a non-supervised approach. Genes were selected based on their behaviour within the different time points of the study, without relying on external information. We obtained **vaccination-induced gene pools**: groups of genes whose modulation could be reasonably asserted to the vaccine-induced stimulus. We found that the magnitude of transcriptional response (number of significantly modulated genes within the gene pool), measured 7 days post-immunization, was associated (AUC = 0.81) with the magnitude of the humoral antibody response to vaccination, a property which was not shared with the day 1 gene pool. This could be justified by the fact that at 24 hours we prevalently observed signals from the innate immune response, whereas after a week there were clear signals from the adaptive response, which in turn could lead to the generation of functional antibodies. Albeit this last process might be influenced by the innate response, we could not find any apparent link between the two responses. As stated, the pools were generated without a priori knowledge of gene function. Nonetheless, they included genes involved in the innate and the adaptive immune responses at days 1 and 7 respectively, which reinforced our confidence in this unsupervised approach. Some of the genes (i.e. TNFRSF17 and MZB1) were modulated across subjects, whereas others (i.e. JCHAIN, ELL2, ERLEC1) showed up regulation mainly in *high responders*. All subjects displayed B cell activation and antibody secretion signals, in addition *high responders* presented transcriptional activities bound to VDJ recombination, glycan and immunoglobulin biosynthesis, T cells differentiation and cell expansion. In Section 1.2.2 we outlined the processes which are involved in the generation of an adaptive immune response, which can lead to the development of immunological memory, the end goal of vaccination. The signals that we observed from the day 7 vaccination induced gene pool pointed exactly to these processes: Lymphocytes B and T were activated and signs of interaction between these cells could be observed. Such interaction can lead to the generation of high-affinity plasma cells and memory B cells. In this study, observing such signals demonstrates that our method correctly captured the effects of two influenza vaccines of known efficacy. In other studies, the presence/absence of these signals could point to the efficacy or inefficacy of a vaccine in generating immunological memory. Additionally, in this particular study it is reasonable to assume that *low responders* could be reacting less to the vaccination event due to pre-exposition to the influenza viruses targeted. Although Tsang and colleagues [1] specify that most of the cohort was naïve for the pandemic virus, the same could not be said for the seasonal virus. Since both vaccines for pandemic and seasonal H1N1 viruses were administered to all study subjects, it was not possible to discriminate between a lack of response intensity due to vaccine ineffectiveness, pre existing immunity or the potential cross-reactivity of memory B cells [39] [114].

Next, we further investigated if the previously described transcriptome data of the day 7 vaccination-induced gene pool was able to predict the magnitude of the antibody response, referred to through the categorical *high* and *low responder* subject classes. We executed feature selection on a Machine Learning model to determine which genes were driving the **magnitude in serological response** observed 70 days post-vaccination. We observed signals pointing to B and T cell replication and interaction in the context of viral infection as well as evidences of immunoglobulin generation, although components of the innate response were also observed (i.e. the ZBP1 gene, which is involved in the DNA-mediated interferon response). The identified genes did not exhibit the highest values of Fold-Change when examined via canonical case versus control group analysis, but rather they were mainly (although not exclusively) modulated in *high responders*. Again, this highlighted the need to approach these kind of biological questions with methods able to leverage individual subject data.

Finally, we tested whether latent factors like pre-vaccination serological status, age and gender data could have an impact on individual transcriptome response profiles. We were not able to identify distinct drivers of differential vaccination response in the **pre-vaccination and intrinsic characteristics** of study subjects. Gender was not a relevant factor, as well as pre-existing immunity and age (in accordance to Tsang and colleagues [1]). Regarding pre-existing immunity, there was a lack of ability in identifying differences in the transcriptome responses due to the concurrent vaccination with two distinct vaccine preparations (seasonal and pandemic H1N1 vaccines), as well as the possible lack of efficacy in a fraction of subjects and potential cross-reactivity. Age is a known factor for differential vaccine response, however this particular study was conducted on adults ranging from 21 to 62 years old, with no representatives from the most extreme age groups (< 18 and > 65 years old).

In conclusion:

- We have developed a bioinformatic pipeline to define individual transcriptome response profiles from vaccinated individuals. This approach was successful in providing complementary information with respect to group-wise based comparisons. Group-wise comparisons measure the average shift of a response to a treatment, whereas our method provides a new dimension of information, represented by the frequency and identity of subjects in which a response occurs. Human peripheral blood transcriptome response to influenza vaccination was found to be highly heterogeneous.

- Despite the heterogeneity, we could identify core functional signatures in the PBMC transcriptome data, suggesting that genes involved in biological processes related to the response to vaccinations respond more robustly. At day 1 we observed Cytokine signaling pathways, including IFN-γ, IL2, IL6, while seven days post vaccination we could identify cell proliferation and activation signals (plasmablasts activity) and genes involved in VDJ recombination and immmunoglobulin production.
- Day 7 transcriptome responses displayed B and T cell proliferation signals, plasmablasts activity and signals pointing to immunoglobulin production. The amount of these responses was found to be associated with functional antibody response (AUC = 81%). The GLDC and HIST1H3G genes, respectively involved in the conferral of survival advantage to T cells in sites of infection and in cell replication, were identified to be the most informative in the classification of high and low responders.

The purpose of this study is to better characterize and understand human responses to vaccination and use these information to drive the improvement and optimization of new and established vaccines. The ability to detect which genes are predictive of an optimal seroresponse to vaccination and to determine the fraction of subjects in which these genes are modulated, enables researchers to assess the fraction of subjects for which a vaccine is expected to be effective, leading to understand whether different formulations could be more effective in distinct populations. We also envision an extension of this methodology to multiple datasets with the aim of detecting shared vaccination-induced signatures. This method was designed to be easily generalizable to all repeated-measure experiments and adaptable to different omics technologies. Additionally, the approach can be deployed in cohort-based studies with the aim of comparing the cohort-specific behaviours in response to treatments or in the context of diseases. Our pipeline requires little computational resources and can be fully executed on consumer grade hardware, which is what was done during this entire work.

Bibliography

[1] J. S. Tsang, P. L. Schwartzberg, Y. Kotliarov, A. Biancotto, Z. Xie, R. N. Germain, E. Wang, M. J. Olnes, M. Narayanan, H. Golding, S. Moir, H. B. Dickler, S. Perl, and F. Cheung, "Global analyses of human immune variation reveal baseline predictors of postvaccination responses," *Cell*, vol. 157, no. 2, pp. 499–513, 2014.

[2] J. Menche, E. Guney, A. Sharma, P. J. Branigan, M. J. Loza, F. Baribaud, R. Dobrin, and A.-L. Barabási, "Integrating personalized gene expression profiles into predictive disease-associated gene pools," *npj Systems Biology and Applications*, vol. 3, no. 10, 2017.

[3] D. E. Zak and A. Aderem, "Systems biology and innate immunity," *Immunological reviews*, vol. 227, pp. 264–282, 2009.

[4] E. Ahmed and A. Hashish, "On modelling the immune system as a complex system," *Theory in Biosciences*, vol. 124, pp. 413–418, 2006.

[5] G. Gasteiger, A. D'osualdo, D. A. Schubert, A. Weber, E. M. Bruscia, and D. Hartl, "Cellular Innate Immunity: An Old Game with New Players," *Journal of Innate Immunity*, vol. 9, pp. 111–125, 2017.

[6] S. E. Degn and S. Thiel, "Humoral Pattern Recognition and the Complement System," *Scandinavian Journal of Immunology*, vol. 78, pp. 181–193, 2013.

[7] R. Medzhitov, "Origin and physiological roles of inflammation," *Nature*, vol. 454, no. 7203, pp. 428–435, 2008.

[8] S. Chiu and A. Bharat, "Role of monocytes and macrophages in regulating immune response following lung transplantation," *Current Opinion in Organ Transplantation*, vol. 21, no. 3, pp. 239–245, 2016.

[9] G. J. Clark, N. Angel, M. Kato, J. A. López, K. MacDonald, S. Vuckovic, and D. N. Hart, "The role of dendritic cells in the innate immune system," *Microbes and Infection*, vol. 2, no. 3, pp. 257–272, 2000.

[10] J. Šedý, V. Bekiaris, and C. F. Ware, "Tumor necrosis factor superfamily in innate immunity and inflammation," *Cold Spring Harbor Perspectives in Biology*, vol. 7, no. a016279, 2015.

[11] C. A. Dinarello, "Overview of the IL-1 family in innate inflammation and acquired immunity," *Immunological Reviews*, vol. 281, pp. 8–27, 2018.

[12] C. N. Serhan and J. Savill, "Resolution of inflammation: The beginning programs the end," *Nature Immunology*, vol. 6, no. 12, pp. 1191–1197, 2005.

[13] J. R. Dunkelberger and W. C. Song, "Complement and its role in innate and adaptive immune responses," *Cell Research*, vol. 20, pp. 34–50, 2010.

[14] K. Li, K. J. Anderson, Q. Peng, A. Noble, B. Lu, A. P. Kelly, N. Wang, S. H. Sacks, and W. Zhou, "Cyclic AMP plays a critical role in C3a-receptor mediated regulation of dendritic cells in antigen uptake and T-cell stimulation," *Blood*, vol. 112, no. 13, pp. 5084–5094, 2008.

[15] F. A. Bonilla and H. C. Oettgen, "Adaptive immunity," *Journal of Allergy and Clinical Immunology*, vol. 125, pp. S33–S40, 2010.

[16] N. Germic, Z. Frangez, S. Yousefi, and H. U. Simon, "Regulation of the innate immune system by autophagy: monocytes, macrophages, dendritic cells and antigen presentation," *Cell Death and Differentiation*, vol. 26, no. 4, pp. 715–727, 2019.

[17] N. P. Restifo and L. Gattinoni, "Lineage relationship of effector and memory T cells," *Current Opinion in Immunology*, vol. 25, no. 5, pp. 556–563, 2013.

[18] M. McHeyzer-Williams, S. Okitsu, N. Wang, and L. McHeyzer-Williams, "Molecular programming of B cell memory," *Nature Reviews Immunology*, vol. 12, no. 1, pp. 24–34, 2012.

[19] D. Y. Tsai, K. H. Hung, C. W. Chang, and K. I. Lin, “Regulatory mechanisms of B cell responses and the implication in B cell-related diseases,” *Journal of Biomedical Science*, vol. 26, no. 1, pp. 1–13, 2019.

[20] S. L. Nutt, P. D. Hodgkin, D. M. Tarlinton, and L. M. Corcoran, “The generation of antibody-secreting plasma cells,” *Nature Reviews Immunology*, vol. 15, no. 3, pp. 160–171, 2015.

[21] A. K. Abbas, A. H. Lichtman, and P. Shiv, *Immunologia cellulare e molecolare*. Elsevier, 6th ed., 2011.

[22] N. Pardi, M. J. Hogan, F. W. Porter, and D. Weissman, “mRNA vaccines-a new era in vaccinology,” *Nature Reviews Drug Discovery*, vol. 17, no. 4, pp. 261–279, 2018.

[23] K. Bahl, J. J. Senn, O. Yuzhakov, A. Bulychev, L. A. Brito, K. J. Hassett, M. E. Laska, M. Smith, Ö. Almarsson, J. Thompson, A. M. Ribeiro, M. Watson, T. Zaks, and G. Ciaramella, “Preclinical and Clinical Demonstration of Immunogenicity by mRNA Vaccines against H10N8 and H7N9 Influenza Viruses,” *Molecular Therapy*, vol. 25, no. 6, pp. 1316–1327, 2017.

[24] M. Alberer, U. Gnad-Vogt, H. S. Hong, K. T. Mehr, L. Backert, G. Finak, R. Gottardo, M. A. Bica, A. Garofano, S. D. Koch, M. Fotin-Mleczek, I. Hoerr, R. Clemens, and F. von Sonnenburg, “Safety and immunogenicity of a mRNA rabies vaccine in healthy adults: an open-label, non-randomised, prospective, first-in-human phase 1 clinical trial,” *The Lancet*, vol. 390, no. 10101, pp. 1511–1520, 2017.

[25] R. Rappuoli and A. Aderem, “A 2020 vision for vaccines against HIV, tuberculosis and malaria,” *Nature*, vol. 473, no. 7348, pp. 463–469, 2011.

[26] S. C. Chow and J. P. Liu, *Design and Analysis of Clinical Trials: Concepts and Methodologies*. Wiley, 3rd ed., 2013.

[27] F. Krammer, G. J. D. Smith, R. A. M. Fouchier, M. Peiris, K. Kedzierska, P. C. Doherty, P. Palese, M. L. Shaw, J. Treanor, R. G. Webster, and A. G. Sastre, “Influenza,” *Nature Reviews*, vol. 4, no. 3, 2018.

[28] S. A. Sellers, R. S. Hagan, F. G. Hayden, and W. A. Fischer, "The hidden burden of influenza: A review of the extra-pulmonary complications of influenza infection," *Influenza and other Respiratory Viruses*, vol. 11, no. 5, pp. 372–393, 2017.

[29] W. W. Thompson, D. K. Shay, E. Weintraub, L. Brammer, N. Cox, L. J. Anderson, and K. Fukuda, "Mortality Associated With Influenza and Respiratory Syncytial Virus in the United States," *JAMA*, vol. 289, no. 2, pp. 179–186, 2003.

[30] S. Su, X. Fu, G. Li, F. Kerlin, and M. Veit, "Novel Influenza D virus: Epidemiology, pathology, evolution and biological characteristics," *Virulence*, vol. 8, no. 8, pp. 1580–1591, 2017.

[31] S. K. White, W. Ma, C. J. McDaniel, G. C. Gray, and J. A. Lednicky, "Serologic evidence of exposure to influenza D virus among persons with occupational contact with cattle," *Journal of Clinical Virology*, vol. 81, pp. 31–33, 2016.

[32] E. Ghedin, N. A. Sengamalay, M. Shumway, J. Zaborsky, T. Feldblyum, V. Subbu, D. J. Spiro, J. Sitz, H. Koo, P. Bolotov, D. Dernovoy, T. Tatusova, Y. Bao, K. St George, J. Taylor, D. J. Lipman, C. M. Fraser, J. K. Taubenberger, and S. L. Salzberg, "Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution," *Nature*, vol. 437, no. 7062, pp. 1162–1166, 2005.

[33] N. Salez, J. Mélade, H. Pascalis, S. Aherfi, K. Dellagi, R. N. Charrel, F. Carrat, and X. de Lamballerie, "Influenza C virus high seroprevalence rates observed in 3 different population groups," *Journal of Infection*, vol. 69, no. 2, pp. 182–189, 2014.

[34] F. Krammer, "The human antibody response to influenza A virus infection and vaccination," *Nature Reviews Immunology*, vol. 19, no. 6, pp. 383–397, 2019.

[35] C. I. Paules and A. S. Fauci, "Influenza Vaccines: Good, but We Can Do Better," *Journal of Infectious Diseases*, vol. 219, no. Suppl 1, pp. S1–S4, 2019.

[36] A. E. Fiore, T. M. Uyeki, K. Broder, L. Finelli, G. L. Euler, J. A. Singleton, J. K. Iskander, P. M. Wortley, D. K. Shay, J. S. Bresee,

and N. J. Cox, “Prevention and control of influenza with vaccines: recommendations of the Advisory Committee on Immunization Practices (ACIP),” *MMWR Recomm Rep.*, vol. 59, 2010.

[37] S. J. Zost, K. Parkhouse, M. E. Gumina, K. Kim, S. D. Perez, P. C. Wilson, J. J. Treanor, A. J. Sant, S. Cobey, and S. E. Hensley, “Contemporary H3N2 influenza viruses have a glycosylation site that alters binding of antibodies elicited by egg-adapted vaccine strains,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 114, no. 47, pp. 12578–12583, 2017.

[38] S. Jones, S. Nelson-Sathi, Y. Wang, R. Prasad, S. Rayen, V. Nandel, Y. Hu, W. Zhang, R. Nair, S. Dharmaseelan, D. V. Chirundodh, R. Kumar, and R. M. Pillai, “Evolutionary, genetic, structural characterization and its functional implications for the influenza A (H1N1) infection outbreak in India from 2009 to 2017,” *Scientific Reports*, vol. 9, no. 1, pp. 1–10, 2019.

[39] K. Hancock, V. Veguilla, X. Lu, W. Zhong, E. N. Butler, H. Sun, F. Liu, L. Dong, J. R. DeVos, P. M. Gargiullo, T. L. Brammer, N. J. Cox, T. M. Tumpey, and J. M. Katz, “Cross-reactive antibody responses to the 2009 pandemic H1N1 influenza virus,” *New England Journal of Medicine*, vol. 361, no. 20, pp. 1945–1952, 2009.

[40] L. E. Lansbury, S. Smith, W. Beyer, E. Karamehic, E. Pasic-Juhas, H. Sikira, A. Mateus, H. Oshitani, H. Zhao, C. R. Beck, and J. S. Nguyen-Van-Tam, “Effectiveness of 2009 pandemic influenza A(H1N1) vaccines: A systematic review and meta-analysis,” *Vaccine*, vol. 35, no. 16, pp. 1996–2006, 2017.

[41] W. W. Busse, S. P. Peters, M. J. Fenton, H. Mitchell, E. R. Bleecker, M. Castro, S. Wenzel, S. C. Erzurum, A. M. Fitzpatrick, W. G. Teague, N. Jarjour, W. C. Moore, K. Sumino, S. Simeone, S. Ratanamaneechat, M. Penugonda, B. Gaston, T. M. Ross, S. Sigelman, J. R. Schiepan, D. J. Zaccaro, C. J. Crevar, D. M. Carter, and A. Togias, “Vaccination of patients with mild and severe asthma with a 2009 pandemic H1N1 influenza virus vaccine,” *Journal of Allergy and Clinical Immunology*, vol. 127, no. 1, pp. 130–137.e3, 2011.

[42] R. H. Raeven, E. van Riet, H. D. Meiring, B. Metz, and G. F. Kersten, "Systems vaccinology and big data in the vaccine development chain," *Immunology*, vol. 156, no. 1, pp. 33–46, 2018.

[43] H. I. Nakaya, J. Wrammert, E. K. Lee, L. Racioppi, S. Marie-Kunze, W. N. Haining, A. R. Means, S. P. Kasturi, N. Khan, G. M. Li, M. McCausland, V. Kanchan, K. E. Kokko, S. Li, R. Elbein, A. K. Mehta, A. Aderem, K. Subbarao, R. Ahmed, and B. Pulendran, "Systems biology of vaccination for seasonal influenza in humans," *Nature Immunology*, vol. 12, no. 8, pp. 786–795, 2011.

[44] H. I. Nakaya, T. Hagan, S. S. Duraisingham, E. K. Lee, M. Kwissa, N. Rouphael, D. Frasca, M. Gersten, A. K. Mehta, R. Gaujoux, G. M. Li, S. Gupta, R. Ahmed, M. J. Mulligan, S. Shen-Orr, B. B. Blomberg, S. Subramaniam, and B. Pulendran, "Systems Analysis of Immunity to Influenza Vaccination across Multiple Years and in Diverse Populations Reveals Shared Molecular Signatures," *Immunity*, vol. 43, no. 6, pp. 1186–1198, 2015.

[45] C. Parra-Rojas, V. von Messling, and E. A. Hernandez-Vargas, "Adjuvanted influenza vaccine dynamics," *Scientific Reports*, vol. 9, no. 73, pp. 1–15, 2019.

[46] K. L. Bucasas, L. M. Franco, C. A. Shaw, M. S. Bray, J. M. Wells, D. Niño, N. Arden, J. M. Quarles, R. B. Couch, and J. W. Belmont, "Early patterns of gene expression correlate with the humoral immune response to influenza vaccination in humans," *Journal of Infectious Diseases*, vol. 203, no. 7, pp. 921–929, 2011.

[47] G. Obermoser, S. Presnell, K. Domico, H. Xu, Y. Wang, E. Anguiano, L. A. Thompson-Snipes, R. Ranganathan, B. Zeitner, A. Bjork, D. Anderson, C. Speake, E. Ruchaud, J. Skinner, L. Alsina, M. Sharma, H. Dutartre, A. Cepika, E. Israelsson, P. Nguyen, Q. A. Nguyen, A. C. Harrod, S. M. Zurawski, V. Pascual, H. Ueno, G. T. Nepom, C. Quinn, D. Blankenship, K. Palucka, J. Banchereau, and D. Chaussabel, "Systems scale interactive exploration reveals quantitative and qualitative differences in response to influenza and pneumococcal vaccines," *Immunity*, vol. 38, no. 4, pp. 831–844, 2013.

[48] D. Furman, V. Jojic, B. Kidd, S. Shen-Orr, J. Price, J. Jarrell, T. Tse, H. Huang, P. Lund, H. T. Maecker, P. J. Utz, C. L. Dekker, D. Koller, and M. M. Davis, "Apoptosis and other immune biomarkers predict influenza vaccine responsiveness," *Molecular Systems Biology*, vol. 9, no. 659, pp. 1–14, 2013.

[49] R. Rappuoli, E. Siena, and O. Finco, "Will systems biology deliver its promise and contribute to the development of new or improved vaccines?: Systems biology views of vaccine innate and adaptive immunity," *Cold Spring Harbor Perspectives in Biology*, vol. 10, 2017.

[50] P. Brodin and M. M. Davis, "Human immune system variation.," *Nature reviews. Immunology*, vol. 17, pp. 21–29, 2017.

[51] H. I. Nakaya, E. Clutterbuck, D. Kazmin, L. Wang, M. Cortese, S. E. Bosinger, N. B. Patel, D. E. Zak, A. Aderem, T. Dong, G. Del Giudice, R. Rappuoli, V. Cerundolo, A. J. Pollard, B. Pulendran, and C.-A. Siegrist, "Systems biology of immunity to MF59-adjuvanted versus nonadjuvanted trivalent seasonal influenza vaccines in early childhood," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 113, no. 7, pp. 1853–1858, 2016.

[52] S. N. Mueller and L. K. Mackay, "Tissue-resident memory T cells: Local specialists in immune defence," *Nature Reviews Immunology*, vol. 16, pp. 79–89, 2016.

[53] R. Edgar, M. Domrachev, and A. E. Lash, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Research*, vol. 30, no. 1, pp. 207–210, 2002.

[54] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis, and A. Soboleva, "NCBI GEO: Archive for functional genomics data sets - Update," *Nucleic Acids Research*, vol. 41, no. D1, pp. 991–995, 2013.

[55] D. D. Dalma-Weiszhausz, J. Warrington, E. Y. Tanimoto, and C. G. Miyada, "The Affymetrix GeneChip® Platform: An Overview," *Methods in Enzymology*, vol. 410, no. 06, pp. 3–28, 2006.

[56] J. D. Hoheisel, “Microarray technology: Beyond transcript profiling and genotype analysis,” *Nature Reviews Genetics*, vol. 7, no. 3, pp. 200–210, 2006.

[57] A. L. Tarca, R. Romero, and S. Draghici, “Analysis of microarray experiments of gene expression profiling,” *American Journal of Obstetrics and Gynecology*, vol. 195, no. 2, pp. 373–388, 2006.

[58] C. P. Verschoor, P. Singh, M. L. Russell, D. M. Bowdish, A. Brewer, L. Cyr, B. J. Ward, and M. Loeb, “Microneutralization assay titres correlate with protection against seasonal influenza H1N1 and H3N2 in children,” *PLoS ONE*, vol. 10, no. 6, pp. 7–13, 2015.

[59] WHO, “Serological Diagnosis of Influenza By Microneutralization Assay,” *WHO guideline on MN test*, pp. 1–25, 2010.

[60] C. C. Czerkinsky, L.-A. Nilsson, H. Nygren, O. Ouchterlony, and A. Tarkowski, “A Solid-Phase Enzyme-Linked Immunospot (ELISPOT) Assay for Enumeration of Specific,” *Journal of Immunological Methods*, vol. 65, pp. 109–121, 1983.

[61] J. M. Walker, *Western Blotting Methods and Protocols.* 2015.

[62] J. Ho, S. Moir, W. Wang, J. G. Posada, W. Gu, M. T. Rehman, R. Dewar, C. Kovacs, M. C. Sneller, T.-w. Chun, D. A. Follmann, and A. S. Fauci, “Enhancing effects of adjuvanted 2009 pandemic H1N1 influenza A vaccine on memory B-cell responses in HIV-infected individuals,” *AIDS*, vol. 25, no. 3, pp. 295–302, 2011.

[63] R Core Team, *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2020.

[64] W. Huber, V. J. Carey, R. Gentleman, S. Anders, M. Carlson, B. S. Carvalho, H. C. Bravo, S. Davis, L. Gatto, T. Girke, R. Gottardo, F. Hahne, K. D. Hansen, R. A. Irizarry, M. Lawrence, M. I. Love, J. Macdonald, V. Obenchain, A. K. Oleś, H. Pagès, A. Reyes, P. Shannon, G. K. Smyth, D. Tenenbaum, L. Waldron, and M. Morgan, “Orchestrating high-throughput genomic analysis with Bioconductor,” *Nature Publishing Group*, vol. 12, no. 2, pp. 115–121, 2015.

[65] S. Davis and P. S. Meltzer, “GEOquery: A bridge between the Gene Expression Omnibus (GEO) and BioConductor,” *Bioinformatics*, vol. 23, no. 14, pp. 1846–1847, 2007.

[66] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth, “limma powers differential expression analyses for RNA-sequencing and microarray studies,” *Nucleic Acids Research*, vol. 43, no. 7, 2015.

[67] F. Cunningham, P. Achuthan, W. Akanni, J. Allen, M. R. Amode, I. M. Armean, R. Bennett, J. Bhai, K. Billis, S. Boddu, C. Cummins, C. Davidson, K. J. Dodiya, L. Gil, T. Grego, L. Haggerty, A. Gall, C. Garc, E. Haskell, T. Hourlier, O. G. Izuogu, S. H. Janacek, T. Juettemann, M. Kay, M. R. Laird, I. Lavidas, Z. Liu, C. Marug, J. E. Loveland, T. Maurel, A. C. Mcmahon, B. Moore, J. Morales, J. M. Mudge, M. Nuhn, D. Ogeh, A. Parker, A. Parton, M. Patricio, A. I. Abdul, B. M. Schmitt, H. Schuilenburg, D. Sheppard, H. Sparrow, E. Stapleton, M. Szuba, K. Taylor, G. Threadgold, A. Thormann, A. Vullo, B. Walts, A. Winterbottom, A. Zadissa, M. Chakiachvili, A. Frankish, S. E. Hunt, M. Kostadima, N. Langridge, F. J. Martin, M. Muffato, E. Perry, M. Ruffier, D. M. Staines, S. J. Trevanion, B. L. Aken, A. D. Yates, D. R. Zerbino, and P. Flicek, “Ensembl 2019,” *Nucleic Acids Research*, vol. 47, no. November 2018, pp. 745–751, 2019.

[68] J. H. Dufour, M. Dziejman, M. T. Liu, J. H. Leung, T. E. Lane, and A. D. Luster, “IFN-gamma-Inducible Protein 10 (IP-10; CXCL10)-Deficient Mice Reveal a Role for IP-10 in Effector T Cell Generation and Trafficking,” *The Journal of Immunology*, vol. 168, no. 7, pp. 3195–3204, 2002.

[69] O. Sobolev, E. Binda, S. O’Farrell, A. Lorenc, J. Pradines, Y. Huang, J. Duffner, R. Schulz, J. Cason, M. Zambon, M. H. Malim, M. Peakman, A. Cope, I. Capila, G. V. Kaundinya, and A. C. Hayday, “Adjuvanted influenza-H1N1 vaccination reveals lymphoid signatures of age-dependent early responses and of clinical adverse events,” *Nature Immunology*, vol. 17, no. 2, pp. 204–213, 2016.

[70] D. Bates, M. Mächler, B. M. Bolker, and S. C. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–47, 2015.

[71] J. A. Gustavsen, S. Pai, R. Isserlin, B. Demchak, and A. R. Pico, "Rcy3: Network biology using Cytoscape from within R," *F1000Research*, vol. 8, no. 1774, pp. 1–20, 2019.

[72] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks," *Genome Research*, vol. 13, pp. 2498–2504, 2003.

[73] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences*, vol. 102, pp. 15545–15550, oct 2005.

[74] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, and J. P. Mesirov, "Molecular signatures database (MSigDB) 3.0," *Bioinformatics*, vol. 27, no. 12, pp. 1739–1740, 2011.

[75] A. Liberzon, C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. P. Mesirov, and P. Tamayo, "The Molecular Signatures Database Hallmark Gene Set Collection," *Cell Systems*, vol. 1, no. 6, pp. 417–425, 2015.

[76] D. A. Chesler and C. S. Reiss, "The role of IFN-γ in immune responses to viral infections of the central nervous system," *Cytokine and Growth Factor Reviews*, vol. 13, no. 6, pp. 441–454, 2002.

[77] H. Zheng, Y. Ban, F. Wei, and X. Ma, *Regulation of Cytokine Gene Expression in Immunity and Diseases*, vol. 941. 2016.

[78] S. Liu, R. Yan, B. Chen, Q. Pan, Y. Chen, J. Hong, L. Zhang, W. Liu, S. Wang, and J. L. Chen, "Influenza virus-induced robust expression of SOCS3 contributes to excessive production of IL-6," *Frontiers in Immunology*, vol. 10, no. AUG, pp. 1–15, 2019.

[79] C. A. Hunter and S. A. Jones, “IL-6 as a keystone cytokine in health and disease,” *Nature Immunology*, vol. 16, no. 5, pp. 448–457, 2015.

[80] U. Klein and R. Dalla-Favera, “Germinal centres: Role in B-cell physiology and malignancy,” *Nature Reviews Immunology*, vol. 8, no. 1, pp. 22–33, 2008.

[81] D. Venkatesh, T. Ernandez, F. Rosetti, I. Batal, X. Cullere, F. W. Luscinskas, Y. Zhang, G. Stavrakis, G. García-Cardeña, B. H. Horwitz, and T. N. Mayadas, “Endothelial TNF Receptor 2 Induces IRF1 Transcription Factor-Dependent Interferon-β Autocrine Signaling to Promote Monocyte Recruitment,” *Immunity*, vol. 38, no. 5, pp. 1025–1037, 2013.

[82] C. V. Ramana, M. Chatterjee-Kishore, H. Nguyen, and G. R. Stark, “Complex roles of Stat1 in regulating gene expression,” *Oncogene*, vol. 19, no. 21, pp. 2619–2627, 2000.

[83] C. M. Coquery and L. D. Erickson, “Regulatory roles of the tumor necrosis factor receptor BCMA,” *Critical Reviews in Immunology*, vol. 32, no. 4, pp. 287–305, 2012.

[84] E. van Anken, F. Pena, N. Hafkemeijer, C. Christis, E. P. Romijn, U. Grauschopf, V. M. Oorschot, T. Pertel, S. Engels, A. Ora, V. Lastun, R. Glockshuber, J. Klumperman, A. J. Heck, J. Luban, and I. Braakman, “Efficient IgM assembly and secretion require the plasma cell induced endoplasmic reticulum protein pERp1,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 40, pp. 17019–17024, 2009.

[85] H. Flach, M. Rosenbaum, M. Duchniewicz, S. Kim, S. L. Zhang, M. D. Cahalan, G. Mittler, and R. Grosschedl, “Mzb1 protein regulates calcium homeostasis, antibody secretion, and integrin activation in innate-like B cells,” *Immunity*, vol. 33, no. 5, pp. 723–735, 2010.

[86] M. J. Niles, L. Matsuuchi, and M. E. Koshland, “Polymer IgM assembly and secretion in lymphoid and nonlymphoid cell lines: evidence that J chain is required for pentamer IgM synthesis.,” *Proceedings of the National Academy of Sciences*, vol. 92, no. 7, pp. 2884–2888, 1995.

[87] F. E. Bertrand, G. L. Billips, G. L. Gartland, H. Kubagawa, and H. W. Schroeder, "The J Chain Gene Is Transcribed During B and T Lymphopoiesis in Humans," *The Journal of Immunology*, vol. 156, pp. 4240–4244, 1996.

[88] K. S. Park, I. Bayles, A. Szlachta-McGinn, J. Paul, J. Boiko, P. Santos, J. Liu, Z. Wang, L. Borghesi, and C. Milcarek, "Transcription Elongation Factor ELL2 Drives Ig Secretory-Specific mRNA Production and the Unfolded Protein Response," *The Journal of Immunology*, vol. 193, no. 9, pp. 4663–4674, 2014.

[89] K. Martincic, S. A. Alkan, A. Cheatle, L. Borghesi, and C. Milcarek, "Transcription elongation factor ELL2 directs immunoglobulin secretion in plasma cells by stimulating altered RNA processing," *Nature Immunology*, vol. 10, no. 10, pp. 1102–1109, 2009.

[90] E. Maverakis, K. Michiko Shimoda, M. Gershwin, F. Patel, R. Wilken, S. Raychaudhuri, R. Ruhaak, and C. Lebrilla, "Glycans In The Immune System and The Altered Glycan Theory of Autoimmunity: A Critical Review," *Journal of Autoimmunity*, vol. 57, pp. 1–13, 2015.

[91] T. Fujimori, Y. Kamiya, K. Nagata, K. Kato, and N. Hosokawa, "Endoplasmic reticulum lectin XTP3-B inhibits endoplasmic reticulum-associated degradation of a misfolded α1-antitrypsin variant," *FEBS Journal*, vol. 280, no. 6, pp. 1563–1575, 2013.

[92] S. Massa, S. Junker, K. Schubart, G. Matthias, and P. Matthias, "The OBF-1 gene locus confers B cell-specific transcription by restricting the ubiquitous activity of its promoter," *European Journal of Immunology*, vol. 33, no. 10, pp. 2864–2874, 2003.

[93] R. Casellas, M. Jankovic, G. Meyer, A. Gazumyan, Y. Luo, R. G. Roeder, and M. C. Nussenzweig, "OcaB is required for normal transcription and V(D)J recombination of a subset of immunoglobulin κ genes," *Cell*, vol. 110, no. 5, pp. 575–585, 2002.

[94] L. Frasca, G. Fedele, S. Deaglio, C. Capuano, R. Palazzo, T. Vaisitti, F. Malavasi, and C. M. Ausiello, "CD38 orchestrates migration, survival, and Th1 immune response of human mature dendritic cells," *Blood*, vol. 107, no. 6, pp. 2392–2399, 2006.

[95] S. Yao, B. F. Buzo, D. Pham, L. Jiang, E. J. Taparowsky, M. H. Kaplan, and J. Sun, “Interferon regulatory factor 4 sustains CD8+T cell expansion and effector differentiation,” *Immunity*, vol. 39, no. 5, pp. 833–845, 2013.

[96] V. Krishnamoorthy, S. Kannanganat, M. Maienschein-Cline, S. L. Cook, J. Chen, N. Bahroos, E. Sievert, E. Corse, A. Chong, and R. Sciammas, “The IRF4 Gene Regulatory Module Functions as a Read-Write Integrator to Dynamically Coordinate T Helper Cell Fate,” *Immunity*, vol. 47, no. 3, pp. 481–497, 2017.

[97] S. Minguet, K. Kläsener, A. M. Schaffer, G. J. Fiala, T. Osteso-Ibánez, K. Raute, I. Navarro-Lérida, F. A. Hartl, M. Seidl, M. Reth, and M. A. Del Pozo, “Caveolin-1-dependent nanoscale organization of the BCR regulates B cell tolerance,” *Nature Immunology*, vol. 18, no. 10, pp. 1150–1159, 2017.

[98] A. Schönle, F. A. Hartl, J. Mentzel, T. Nöltner, K. S. Rauch, A. Prestipino, S. A. Wohlfeil, P. Apostolova, A. K. Hechinger, W. Melchinger, K. Fehrenbach, M. C. Guadamillas, M. Follo, G. Prinz, A. K. Ruess, D. Pfeifer, M. A. Del Pozo, A. Schmitt-Graeff, J. Duyster, K. I. Hippen, B. R. Blazar, K. Schachtrup, S. Minguet, and R. Zeiser, “Caveolin-1 regulates TCR signal strength and regulatory T-cell differentiation into alloreactive T cells,” *Blood*, vol. 127, no. 15, pp. 1930–1939, 2016.

[99] T. Fawcett, “An introduction to ROC analysis,” *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[100] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J. C. Sanchez, and M. Müller, “pROC: An open-source package for R and S+ to analyze and compare ROC curves,” *BMC Bioinformatics*, vol. 12, no. 77, 2011.

[101] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, pp. 5–32, 2001.

[102] A. Liaw and M. Wiener, “Classification and Regression by randomForest,” *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[103] H. Deng and G. Runger, “Gene selection with guided regularized random forest,” *Pattern Recognition*, vol. 46, no. 12, pp. 3483–3489, 2013.

[104] H. Deng and G. Runger, “Feature selection via regularized trees,” *The 2012 International Joint Conference on Neural Networks (IJCNN)*, 2012.

[105] H. Deng, “Guided Random Forest in the RRF Package,” *CoRR*, vol. 1306.0237, pp. 1–2, 2013.

[106] M. Kuhn, “Building Predictive Models in R Using the caret Package,” *Journal of Statistical Software*, vol. 28, no. 5, 2008.

[107] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning.* 2013.

[108] C. Strobl, A. L. Boulesteix, A. Zeileis, and T. Hothorn, “Bias in random forest variable importance measures: Illustrations, sources and a solution,” *BMC Bioinformatics*, vol. 8, no. 25, 2007.

[109] M. D. Buck, D. O’Sullivan, and E. L. Pearce, “T cell metabolism drives immunity,” *Journal of Experimental Medicine*, vol. 212, no. 9, pp. 1345–1360, 2015.

[110] E. L. Pearce and E. J. Pearce, “Metabolic pathways in immune cell activation and quiescence,” *Immunity*, vol. 38, no. 4, pp. 633–643, 2013.

[111] W. F. Marzluff, P. Gongidi, K. R. Woods, J. Jin, and L. J. Maltais, “The human and mouse replication-dependent histone genes,” *Genomics*, vol. 80, no. 5, pp. 487–498, 2002.

[112] A. Takaoka, Z. Wang, M. K. Choi, H. Yanai, H. Negishi, T. Ban, Y. Lu, M. Miyagishi, T. Kodama, K. Honda, Y. Ohba, and T. Taniguchi, “DAI (DLM-1/ZBP1) is a cytosolic DNA sensor and an activator of innate immune response,” *Nature*, vol. 448, no. 7152, pp. 501–505, 2007.

[113] J. L. McCann, M. M. Klein, E. M. Leland, E. K. Law, W. L. Brown, D. J. Salamango, and R. S. Harris, “The DNA deaminase APOBEC3B interacts with the cell-cycle protein CDK4 and disrupts CDK4-mediated nuclear import of Cyclin D1,” *Journal of Biological Chemistry*, vol. 294, no. 32, pp. 12099–12111, 2019.

[114] G.-M. Li, C. Chiu, J. Wrammert, M. McCausland, S. F. Andrews, N.-Y. Zheng, J.-H. Lee, M. Huang, X. Qu, S. Edupuganti, M. Mulligan, S. R. Das, J. W. Yewdell, A. K. Mehta, P. C. Wilson, and R. Ahmed, "Pandemic H1N1 influenza vaccine induces a recall response in humans that favors broadly cross-reactive memory B cells," *Proceedings of the National Academy of Sciences*, vol. 109, no. 23, pp. 9047–9052, 2012.

A | Supplementary Figures

A.1 Group-wise analysis for DEGs (baseline: day -7)

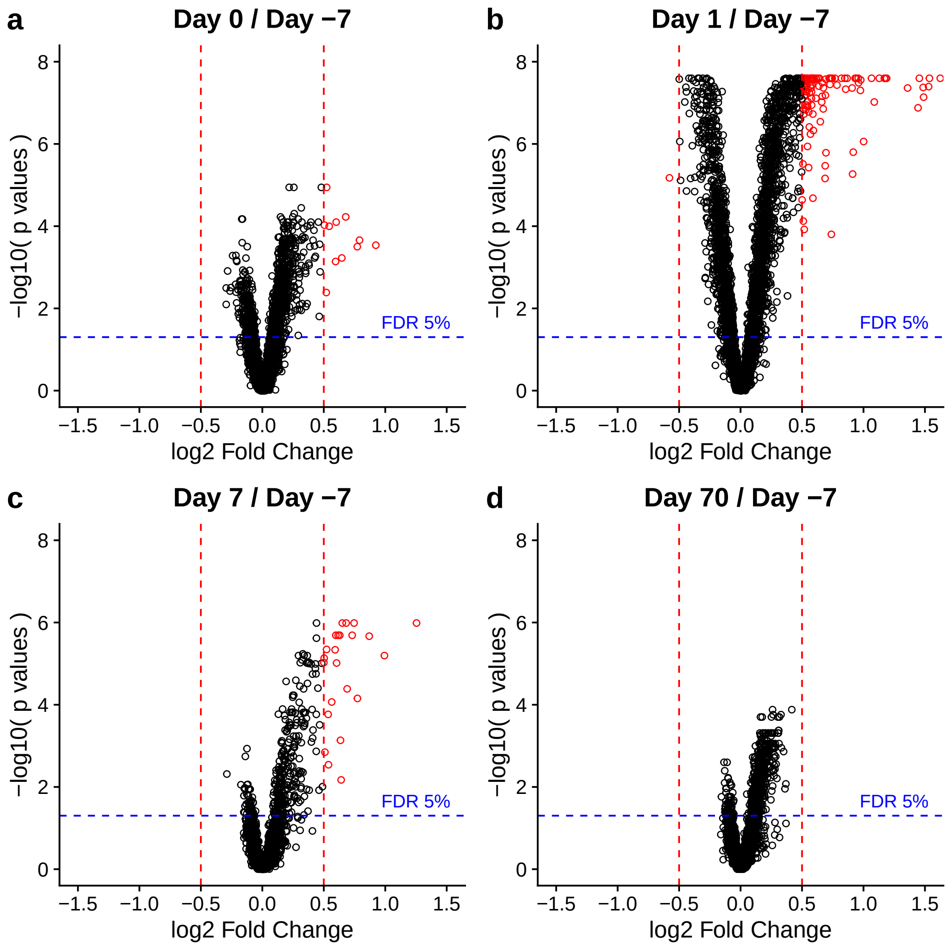


Figure A.1: Mean expression level differences of genes, with differentially expressed genes detected via group-wise analysis shown in red. The dashed lines show the cutoffs for calling a gene differentially expressed: $\log_2$ Fold Change ±0.5 (red lines), p-value < 0.05 (two-tailed, paired Wilcoxon signed-rank test, blue line).

A.2 Group-wise analysis for DEGs (baseline: day 70)

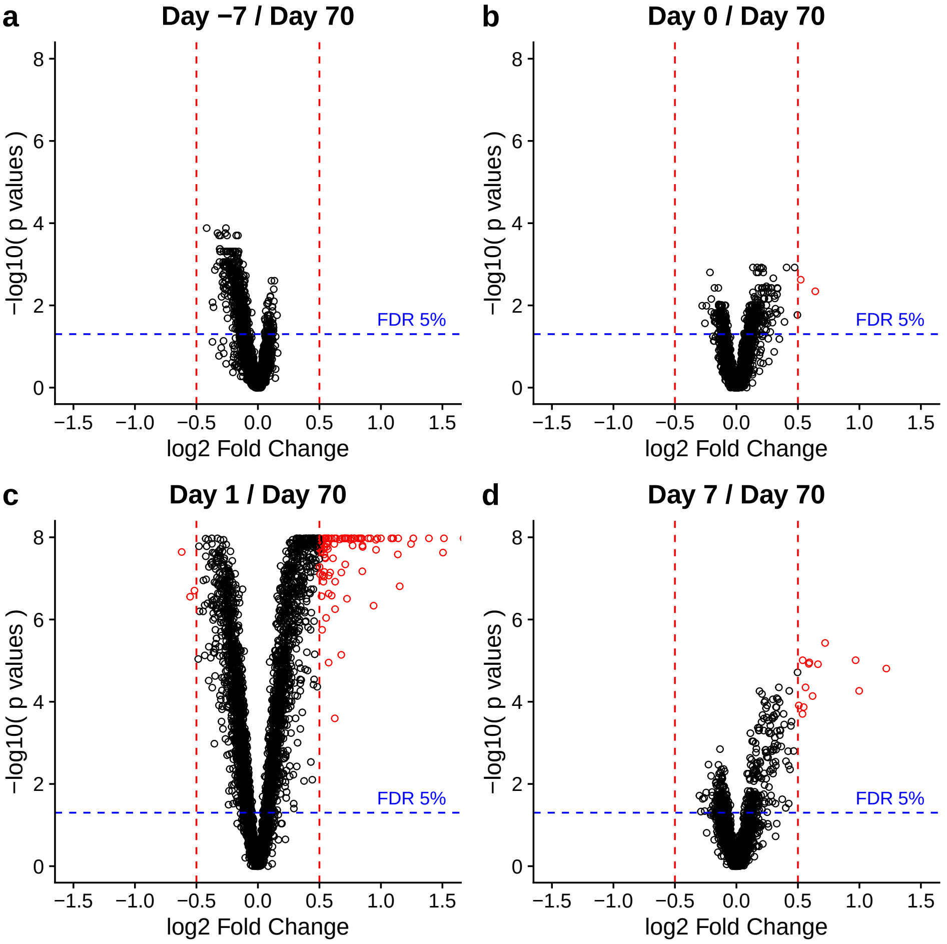


Figure A.2: Mean expression level differences of genes, with differentially expressed genes detected via group-wise analysis shown in red. The dashed lines show the cutoffs for calling a gene differentially expressed: $\log_2$ Fold Change ±0.5 (red lines), p-value < 0.05 (two-tailed, paired Wilcoxon signed-rank test, blue line).

A.3 Permutation test on pathways (day 1)

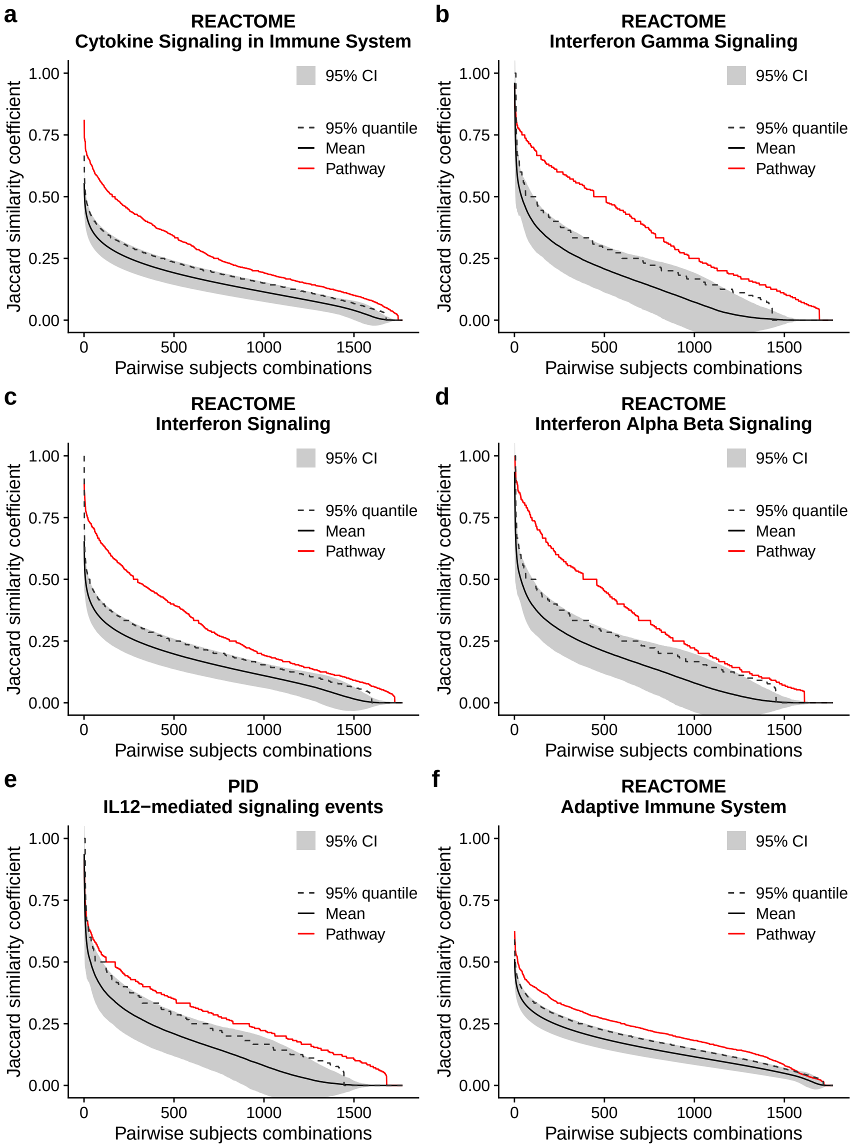


Figure A.3: Most frequently enriched pathways within **day 1** individual subject transcriptomes. The analysis is executed on all pairwise subject combinations, regardless of pathway enrichment being present in the subjects. The Jaccard similarity coefficients associated to the pathway genes are shown in red. The mean of the 1000 control Jaccard indices obtained through the bootstrap procedure is shown in black. The 95% quantile of the bootstrap values is shown through a dashed line. The shadow in gray represents the 95% Confidence Interval of the bootstrap values distribution.

A.4 Permutation test on pathways (day 7)

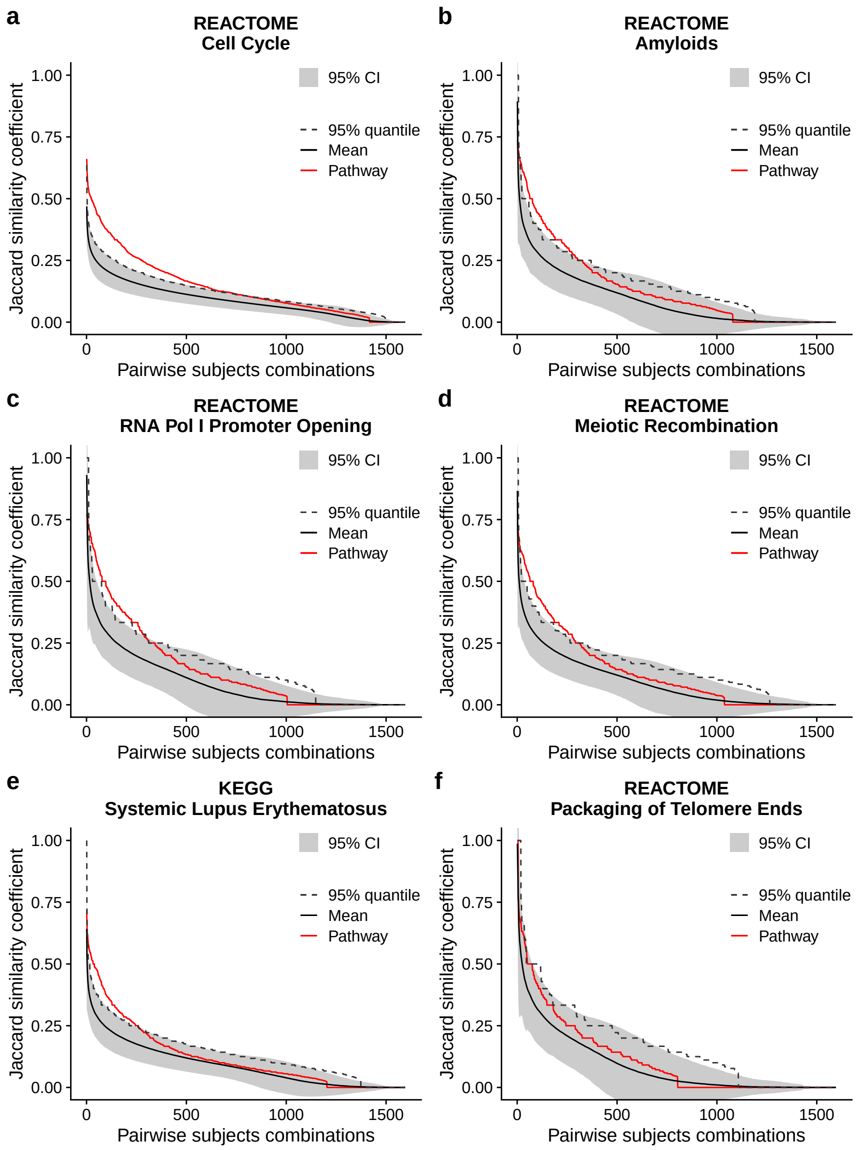


Figure A.4: Most frequently enriched pathways within **day 7** individual subject transcriptomes. The analysis is executed on all pairwise subject combinations, regardless of pathway enrichment being present in the subjects. The Jaccard similarity coefficients associated to the pathway genes are shown in red. The mean of the 1000 control Jaccard indices obtained through the bootstrap procedure is shown in black. The 95% quantile of the bootstrap values is shown through a dashed line. The shadow in gray represents the 95% Confidence Interval of the bootstrap values distribution.

A.5 Gene pool modulation VS intrinsic variables

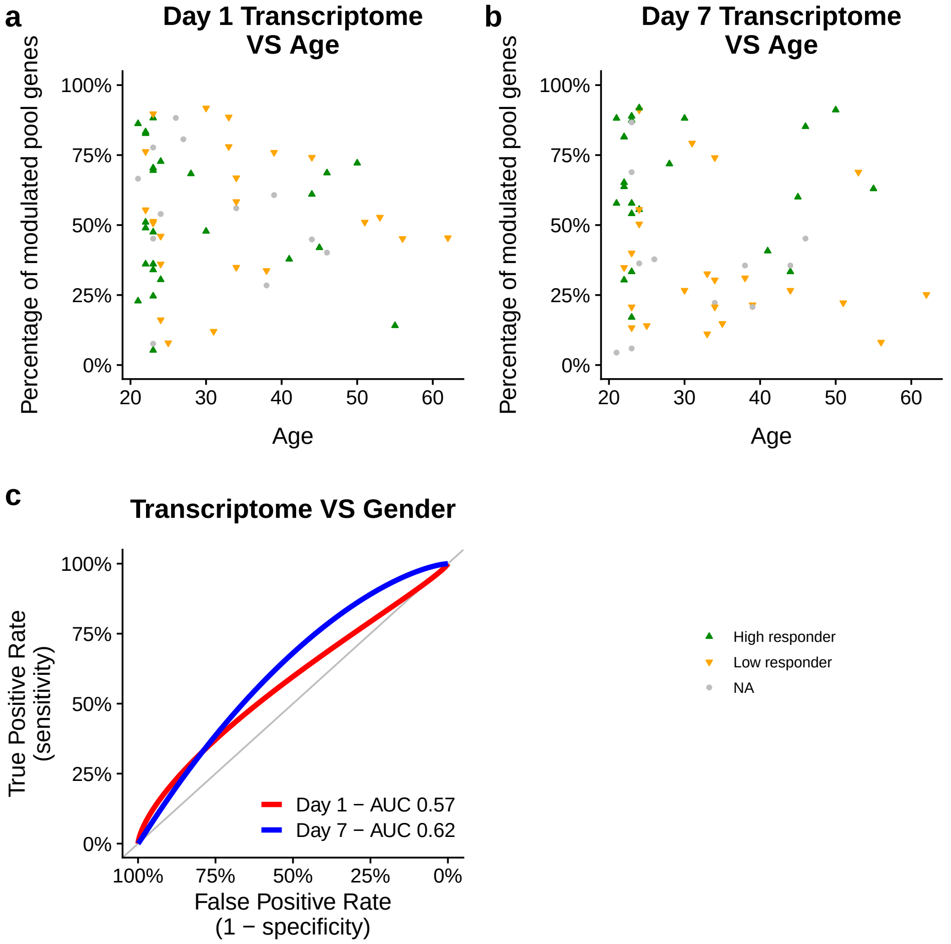


Figure A.5: **a-b** Percentages of modulated pool genes (y-axis) plotted in relation to subject's age (x-axis). Information on functional antibody response (high/low responders) is encoded via color and shape of points. Spearman correlation for days 1 and 7 respectively: ρ = -0.09, -0.11; p-value = 0.49, 0.44 **c** ROC curve of subject's gender inferred using the percentage of modulation of the vaccination-induced gene pools calculated from the individual subject transcriptome response profiles.

A.6 Gene pool modulation VS pre-vaccination MN titers

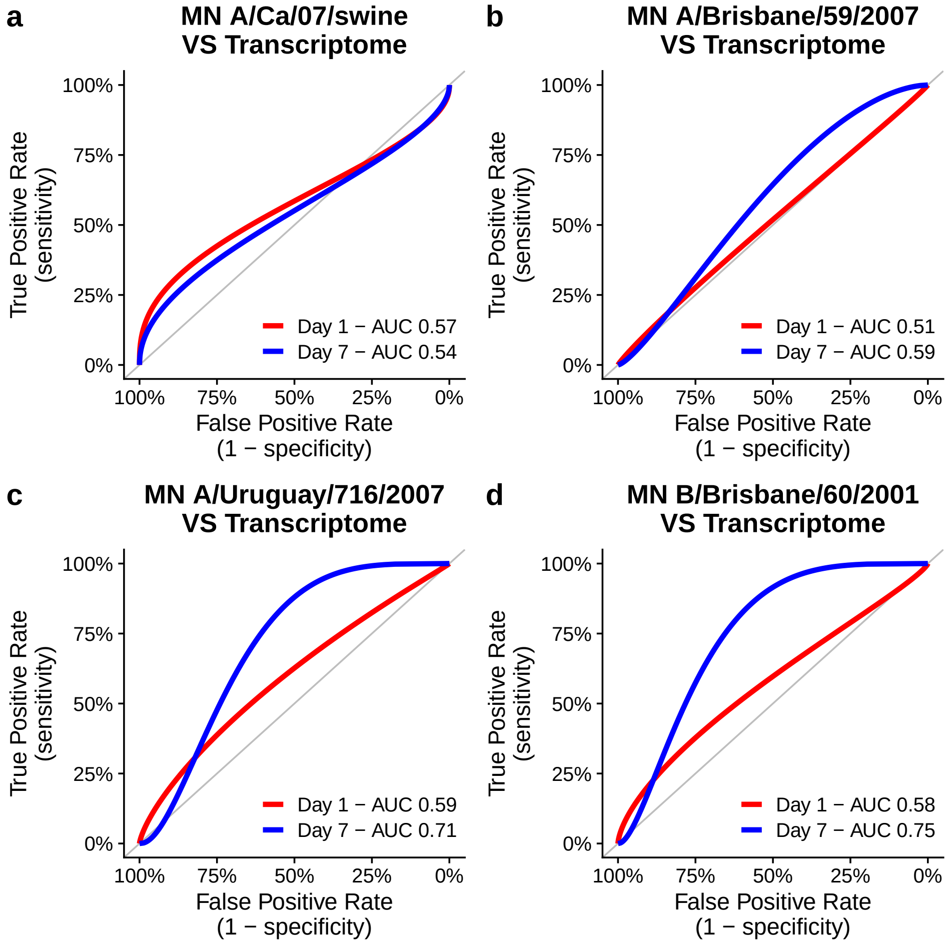


Figure A.6: ROC curves of the pre-vaccination status predictions inferred using the percentage of modulation of the vaccination-induced gene pools calculated from the individual subject transcriptome response profiles.

A.7 Guided Random Forest results (*mtry*=9)

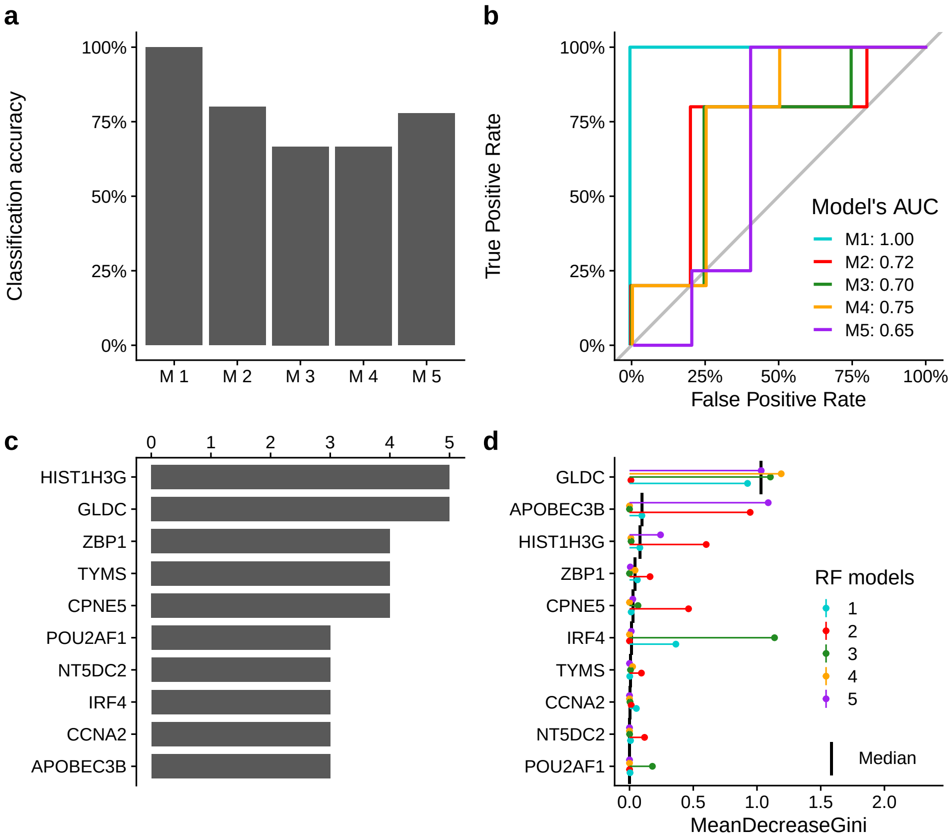


Figure A.7: Guided Random Forest results (*mtry* = 9). **a** Classification accuracy of the models (percentage of correctly classified subjects). **b** ROC curves and AUCs of the models. **c** Number of occurrences across the five models of the most informative genes. From each model, the top 10 most informative features are reported. The y-axis indicates the number of models in which a gene was observed in the top 10. **d** Top 10 most informative genes across GRF model ordered by median of the Mean Decrease Gini coefficients (higher values correspond to higher feature importance).

A.8 Guided Random Forest results (*mtry*=12)

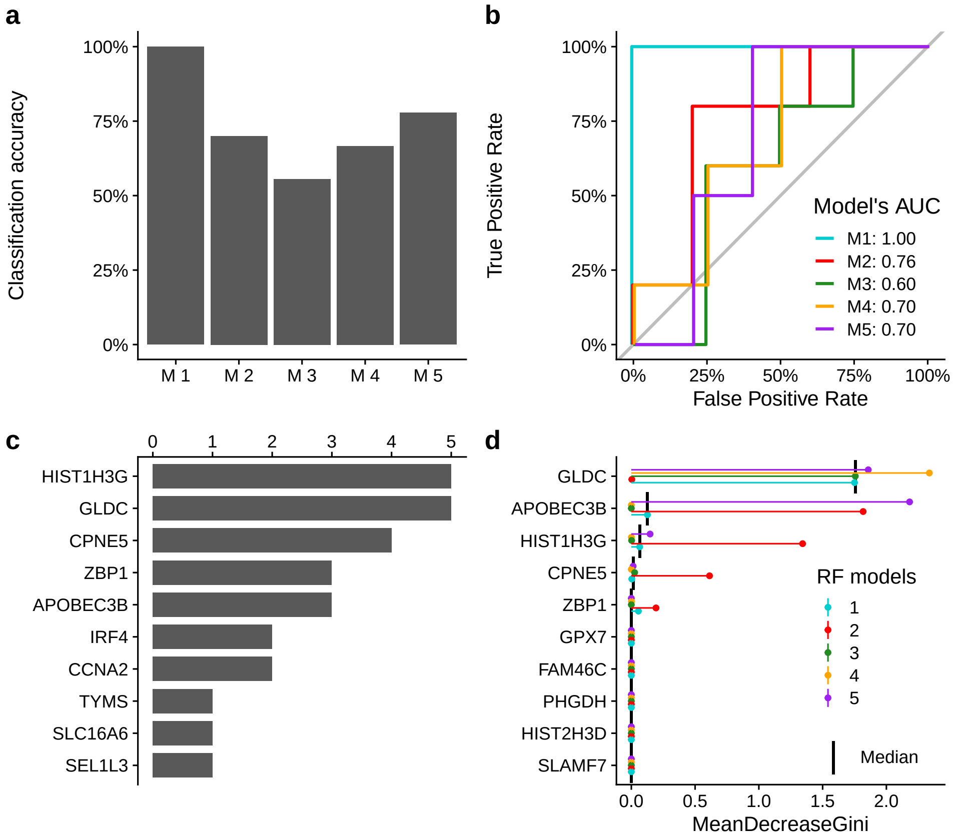


Figure A.8: Guided Random Forest results ($mtry = 12$). **a** Classification accuracy of the models (percentage of correctly classified subjects). **b** ROC curves and AUCs of the models. **c** Number of occurrences across the five models of the most informative genes. From each model, the top 10 most informative features are reported. The y-axis indicates the number of models in which a gene was observed in the top 10. **d** Top 10 most informative genes across GRF model ordered by median of the Mean Decrease Gini coefficients (higher values correspond to higher feature importance).

B | Supplementary Tables

B.1 DEGs from group-wise analysis (day 1)

Genes	**Mean Fold Change**	**-log$_{10}$ p-value (adjusted BH)**	**Day 1 value within 1 SD of day 0**	**Day 0 value higher than day 1 mean**
ANKRD22	1.437	6.284	25	2
HBEGF	1.258	6.775	9	5
G0S2	1.077	6.985	11	2
IL1B	1.047	5.768	18	6
CXCL10	0.962	6.124	30	5
PTGS2	0.884	6.183	19	7
SERPING1	0.862	6.585	22	2
SOCS3	0.789	6.985	10	3
NR4A3	0.74	6.337	13	5
B3GNT5	0.728	6.124	23	9
OSM	0.723	6.562	8	0
IDO1	0.709	5.733	28	2
IFIT2	0.704	6.523	29	6
NR4A2	0.695	6.072	19	8
AREG	0.657	3.027	34	15
RGS1	0.657	5.341	27	8
PLK2	0.657	5.396	16	4
CD83	0.654	6.341	11	5
FOSB	0.619	5.477	28	6
JUN	0.619	6.985	24	5
CA1	0.607	5.706	42	14
GBP4	0.604	6.284	18	2
KLF4	0.596	6.985	15	3
FFAR2	0.596	6.29	27	3
FOS	0.589	6.985	15	4
CXCL8	0.571	3.917	32	13

continued . . .

Genes	**Mean Fold Change**	**-log$_{10}$ p-value (adjusted BH)**	**Day 1 value within 1 SD of day 0**	**Day 0 value higher than day 1 mean**
PFKFB3	0.562	5.765	18	8
STAT1	0.558	6.747	18	2
LAP3	0.555	6.57	23	3
NAMPT	0.552	6.178	22	8
ATF3	0.552	6.985	10	0
CLEC6A	0.541	5.706	27	11
TRIB1	0.539	6.127	20	7
EPSTI1	0.534	6.415	28	7
RIPK2	0.52	6.777	10	2
CCL2	0.512	5.429	36	9
EREG	0.507	5.234	27	8
CCDC59	0.506	6.471	27	6
HCAR3	0.504	4.794	25	9
PMAIP1	0.503	6.172	12	2
NFIL3	0.502	5.879	19	7
EGR2	0.501	3.555	35	11
DDIT3	0.501	6.65	13	2

Table B.1: Differentially expressed genes identified from group-wise analysis (day 1 / day 0).

B.2 DEGs from group-wise analysis (day 7)

Genes	**Mean Fold Change**	**-log$_{10}$ p-value (adjusted BH)**	**Day 7 value within 1 SD of day 0**	**Day 0 value higher than day 7 mean**
TNFRSF17	1.387	5.078	17	1
GPRC5D	1.054	4.91	20	3
JCHAIN	0.979	5.078	16	1
MZB1	0.733	5.078	18	1
CD38	0.676	4.958	22	2
APOBEC3B	0.674	5.078	29	3
ITM2C	0.649	4.959	22	1
CAV1	0.616	4.935	26	1
DENND5B	0.547	4.388	29	3
KLHL14	0.546	3.972	27	4
ELL2	0.534	4.904	23	1
NT5DC2	0.531	4.796	26	2
MYBL2	0.528	4.959	29	2
HIST1H3B	0.51	3.961	30	6

Table B.2: Differentially expressed genes identified from group-wise analysis (day 7 / day 0).

B.3 Enriched pathways (day -7)

Pathway name	**Count**
KEGG taste transduction	5
REACTOME cell cycle	5
REACTOME rna pol i transcription	5
REACTOME chromosome maintenance	5
REACTOME rna pol i promoter opening	5
REACTOME meiotic recombination	5
REACTOME meiotic synapsis	5
REACTOME amyloids	5
KEGG systemic lupus erythematosus	4
PID ap1 pathway	4
REACTOME meiosis	4
REACTOME generic transcription pathway	4
REACTOME rna pol i rna pol iii and mitochondrial transcription	4
REACTOME deposition of new cenpa containing nucleosomes at the centromere	4
REACTOME packaging of telomere ends	4
REACTOME telomere maintenance	4
PID nfat tfpathway	3
PID atf2 pathway	3
PID hif1 tfpathway	3
REACTOME cell cycle mitotic	3

Table B.3: Pathways observed as significantly enriched (Fisher exact test $p \leq 0.05$, one-sided, with Benjamini Hochberg correction for total number of tested pathways across all subjects), alongside the number of subjects in which this observation was made.

B.4 Enriched pathways (day 0)

Pathway name	**Count**
REACTOME olfactory signaling pathway	4
REACTOME amyloids	4
REACTOME packaging of telomere ends	4
KEGG focal adhesion	3
REACTOME deposition of new cenpa containing nucleosomes at the centromere	3
REACTOME rna pol i promoter opening	3
REACTOME interferon alpha beta signaling	3
KEGG ribosome	2
KEGG olfactory transduction	2
KEGG asthma	2
KEGG systemic lupus erythematosus	2
REACTOME meiosis	2
REACTOME immunoregulatory interactions between a lymphoid and a non lymphoid cell	2
REACTOME srp dependent cotranslational protein targeting to membrane	2
REACTOME mhc class ii antigen presentation	2
REACTOME tcr signaling	2
REACTOME generation of second messenger molecules	2
REACTOME generic transcription pathway	2
REACTOME rna pol i transcription	2
REACTOME transcription	2

Table B.4: Pathways observed as significantly enriched (Fisher exact test $p \leq 0.05$, one-sided, with Benjamini Hochberg correction for total number of tested pathways across all subjects), alongside the number of subjects in which this observation was made.

B.5 Enriched pathways (day 1)

Pathway name	**Count**
REACTOME cytokine signaling in immune system	28
REACTOME interferon gamma signaling	27
REACTOME interferon signaling	26
REACTOME interferon alpha beta signaling	20
PID il12 2pathway	19
REACTOME adaptive immune system	11
PID atf2 pathway	9
PID ap1 pathway	9
PID il6 7 pathway	9
KEGG graft versus host disease	8
REACTOME tcr signaling	8
PID il23 pathway	7
REACTOME activation of nf kappab in b cells	7
REACTOME signaling by the b cell receptor bcr	7
REACTOME regulation of ornithine decarboxylase odc	7
REACTOME p53 dependent g1 dna damage response	7
REACTOME apc c cdh1 mediated degradation of cdc20 and other apc c cdh1 targeted proteins in late mitosis early g1	7
REACTOME scf beta trcp mediated degradation of emi1	7
KEGG proteasome	6
PID reg gr pathway	6

Table B.5: Pathways observed as significantly enriched (Fisher exact test $p \leq 0.05$, one-sided, with Benjamini Hochberg correction for total number of tested pathways across all subjects), alongside the number of subjects in which this observation was made.

B.6 Enriched pathways (day 7)

Pathway name	**Count**
REACTOME cell cycle	17
REACTOME rna pol i promoter opening	17
REACTOME amyloids	17
KEGG systemic lupus erythematosus	16
REACTOME meiotic recombination	16
REACTOME rna pol i transcription	15
REACTOME packaging of telomere ends	15
REACTOME meiosis	14
REACTOME rna pol i rna pol iii and mitochondrial transcription	14
REACTOME deposition of new cenpa containing nucleosomes at the centromere	14
REACTOME telomere maintenance	14
PID e2f pathway	13
REACTOME cell cycle mitotic	12
REACTOME transcription	12
REACTOME g0 and early g1	11
REACTOME meiotic synapsis	11
REACTOME chromosome maintenance	9
REACTOME asparagine n linked glycosylation	8
REACTOME mitotic g1 g1 s phases	6
REACTOME dna replication	6

Table B.6: Pathways observed as significantly enriched (Fisher exact test $p \leq 0.05$, one-sided, with Benjamini Hochberg correction for total number of tested pathways across all subjects), alongside the number of subjects in which this observation was made.

B.7 Enriched pathways (day 70)

Pathway name	**Count**
KEGG taste transduction	7
KEGG systemic lupus erythematosus	5
REACTOME cell cycle	5
REACTOME transcription	5
REACTOME rna pol i promoter opening	5
REACTOME amyloids	5
PID e2f pathway	4
REACTOME meiosis	4
REACTOME rna pol i transcription	4
REACTOME cell cycle mitotic	4
REACTOME g1 s transition	4
REACTOME mitotic g1 g1 s phases	4
REACTOME rna pol i rna pol iii and mitochondrial transcription	4
REACTOME chromosome maintenance	4
REACTOME deposition of new cenpa containing nucleosomes at the centromere	4
REACTOME meiotic recombination	4
REACTOME meiotic synapsis	4
REACTOME packaging of telomere ends	4
REACTOME telomere maintenance	4
KEGG cell cycle	3

Table B.7: Pathways observed as significantly enriched (Fisher exact test $p \leq 0.05$, one-sided, with Benjamini Hochberg correction for total number of tested pathways across all subjects), alongside the number of subjects in which this observation was made.

B.8 Vaccination-induced gene pool (day 1)

Gene Symbol and number of modulations across subjects are reported.

SOCS3	51	**PFKFB3**	41	**PPIF**	37	**FRMD3**	35
HBEGF	51	**DDX21**	41	**RCN1**	37	**KLHL15**	35
HAUS3	49	**RAB20**	41	**OAS3**	37	**NAP1L3**	35
NPC1	48	**PPP1R15A**	41	**CD69**	37	**PLEKHO1**	34
RIPK2	48	**GZF1**	41	**COQ10B**	37	**SDC3**	34
SAT1	48	**SIDT1**	41	**APOL4**	37	**DUSP10**	34
ANKRD22	47	**CBLB**	41	**BCL6**	37	**NR4A1**	34
SERPING1	46	**PIGA**	41	**STX11**	37	**PSMC1**	34
TUBA1A	46	**PYHIN1**	40	**SNX10**	37	**FAM103A1**	34
DDIT3	46	**IFIT2**	40	**RANBP6**	37	**BCL2A1**	34
CA5B	46	**FBXO33**	40	**PTCH1**	37	**SECTM1**	34
STAT1	45	**C2**	40	**SLC25A33**	36	**GADD45B**	34
ATP2B4	44	**NR4A3**	40	**FAM159A**	36	**DYSF**	34
ICAM1	44	**YOD1**	39	**IDI1**	36	**HLA-DQB2**	34
MTHFD2	44	**SCO2**	39	**IFITM3**	36	**ARL4A**	34
CCNL1	44	**B3GNT5**	39	**CXCL10**	36	**KLF10**	34
IRF1	44	**VEGFA**	39	**LMNB1**	36	**DNAJA1**	34
CDKN1A	44	**RNF19B**	38	**ATP1B1**	35	**NFIL3**	34
IFRD1	44	**GBP2**	38	**ZNF124**	35	**PDE4B**	33
ATF3	43	**UBE2D1**	38	**EPB41L3**	35	**SLAMF8**	33
HIST2H2BE	43	**IFIT5**	38	**FFAR3**	35	**RHOU**	33
EPSTI1	43	**DET1**	38	**PLAUR**	35	**PRKCQ**	33
PSME2	43	**LGALS3BP**	38	**BHLHE40**	35	**KLRD1**	33
WARS	43	**IKZF2**	38	**FLNB**	35	**ERP27**	33
OSM	43	**NFKBIZ**	38	**PLK2**	35	**TMEM116**	33
ODF3B	43	**LAP3**	38	**ETF1**	35	**ABHD13**	33
KLF4	43	**GK3P**	38	**IER3**	35	**C15orf39**	33
NLRP3	42	**G0S2**	37	**TWISTNB**	35	**B3GNT2**	33
GBP4	42	**CD84**	37	**PLEKHF2**	35	**PLEK**	33
PSTPIP2	42	**FCMR**	37	**TOX**	35	**KREMEN1**	33
FFAR2	42	**OTUD1**	37	**CTSL**	35	**YPEL1**	33

CMTM6	33	KIAA1143	31	SMNDC1	29	ICOS	28
PARP9	33	RIOK3	31	DENR	29	CMPK2	28
EREG	33	ZFP36	31	PTGDR	29	EPHA4	28
ITK	33	GNB4	31	TBX21	29	SMOX	28
ING3	33	CXCL11	31	SEC11C	29	IGF2BP2	28
IDO1	33	SERINC5	31	PMAIP1	29	HERC5	28
SLA	33	DUSP1	31	BCL2L11	29	LAMTOR3	28
RPF1	32	CYP51A1	31	CSRNP1	29	PHAX	28
ELL2	32	UGCG	31	COMMD8	29	TMEM167A	28
DUSP5	32	FRMPD3	31	SNCA	29	DEK	28
OAT	32	TENM1	31	IL7R	29	CCT6A	28
TMEM45B	32	NAMPT	30	CD83	29	RRS1	28
KLRG1	32	PRF1	30	SOD2	29	AAED1	28
KLRC4-KLRK1	32	MYOF	30	AGPAT4	29	EIF1AX	27
FOS	32	RBM4B	30	DNAJB9	29	SETSIP	27
LACTB	32	TWF1	30	PNPLA8	29	DNTTIP2	27
MT2A	32	UCHL3	30	TRIB1	29	NET1	27
SLFN13	32	ADGRG5	30	NCALD	29	TFAM	27
SNRPD1	32	JMJD6	30	ATP6AP2	29	IFIT3	27
PTPN4	32	RMND5A	30	HPRT1	29	HBD	27
ZC3H15	32	IL1B	30	RNF11	28	SLC15A3	27
MRPL44	32	PARP14	30	IFI44L	28	CCDC65	27
TGM2	32	LAMP3	30	LGR6	28	SLC2A14	27
TCN2	32	DTHD1	30	DBT	28	KLRC3	27
TYMP	32	CENPU	30	PRRG4	28	POPDC2	27
KBTBD8	32	ACSL1	30	CWC15	28	ZBTB1	27
EOMES	32	GMPR	30	CLEC6A	28	TNFAIP2	27
PSMD6	32	HLA-DRB1	30	DRAM1	28	CFL2	27
PAIP1	32	CAPZA2	30	UFM1	28	PSMA4	27
SBDS	32	ALDH1A1	30	SAV1	28	CENPN	27
GBP5	31	GAB3	30	THBS1	28	IFI35	27
PDGFD	31	TAF13	29	MAPK6	28	BECN1	27
OLR1	31	DENND2D	29	ADGRG1	28	PTPN2	27
KLRC2	31	FCRL3	29	CCR7	28	BCL3	27
CCDC59	31	PPA1	29	ADNP2	28	LBH	27
						CDC25B	27

APOL1	27	**STAT2**	26	**LTV1**	26	**SKAP1**	25
THOC7	27	**NAP1L1**	26	**PPP1R10**	26	**ACP1**	25
GOLIM4	27	**DNAJC3**	26	**BLVRA**	26	**PI3**	25
PTGER4	27	**EPB42**	26	**DOCK4**	26	**DTX3L**	25
SYNE1	27	**C16orf87**	26	**NDUFA5**	26	**CACNA2D2**	25
BPGM	27	**SLFN14**	26	**HOOK3**	26	**ENOPH1**	25
TMEM140	27	**DUSP3**	26	**HEMGN**	26	**CXCL9**	25
MYBL1	27	**ITGA2B**	26	**C1QB**	25	**SCARB2**	25
JAK2	27	**LILRA3**	26	**RGS1**	25	**RASGEF1B**	25
SCML1	27	**ASAP2**	26	**SYF2**	25	**DNAJC21**	25
UBXN10	26	**PDCL3**	26	**SLC30A1**	25	**RUNX2**	25
CLIC4	26	**MGAT4A**	26	**SHTN1**	25	**SCML4**	25
OR2W3	26	**DIP2A**	26	**RBM7**	25	**SGK1**	25
TMEM63A	26	**SAMSN1**	26	**LRRK2**	25	**AUTS2**	25
FAS	26	**ZBTB21**	26	**YBX3**	25	**FGL2**	25
FAM204A	26	**IL2RB**	26	**TAS2R19**	25	**MSR1**	25
PRDX3	26	**ZNF589**	26	**RFC3**	25	**TJP2**	25
SMCO4	26	**CNBP**	26	**SPRY2**	25	**CEP78**	25
CLSTN3	26	**SPCS3**	26	**MT1G**	25	**ALAS2**	25

B.9 Vaccination-induced gene pool (day 7)

Gene Symbol and number of modulations across subjects are reported.

TNFRSF17	53	MOXD1	30	ZNF215	25	TMEM204	21
MZB1	50	DTL	29	RAB30	25	DHFR	21
JCHAIN	47	ALG14	29	TYMS	25	CCL20	21
NUGGC	45	MKI67	29	PDIA6	25	UGT2B11	21
GPRC5D	44	HRASLS2	29	OSBPL10	25	MANEA	21
BHLHE41	43	RRM2	29	LARP1B	25	HIST1H4L	21
DENND5B	42	SGPP2	29	HIST1H3F	25	PERP	21
ITM2C	42	CCNA2	29	GGH	25	CDK1	20
MYBL2	42	HIST1H3G	29	PHGDH	24	TEX9	20
CD38	42	HIST1H1B	29	ZBTB8OS	24	EPB42	20
ELL2	41	LAMC1	28	ASPM	24	ZBP1	20
APOBEC3B	40	KIF11	28	SLC35F2	24	HSPA13	20
NT5DC2	39	SPATS2	28	ERLEC1	24	MYBL1	20
PARM1	39	SEC11C	28	HIST1H3B	24	TP53INP1	20
CAV1	39	MAN1A1	28	PNOC	24	RFX3	20
IRF4	38	CDC6	27	E2F5	24	DCAF12	20
LMAN1	37	SKA1	27	PIP5K1B	24	TMEM2	20
POU2AF1	36	EAF2	27	LAX1	23	ALAS2	20
GLDC	36	TMEM156	27	VAV3	23	GPX7	19
NUSAP1	34	HIST1H3J	27	WDR76	23	SERPING1	19
SHCBP1	34	AQP3	27	TPD52	23	KNL1	19
BUB1	34	HIST2H4A	27	REXO2	22	AHSP	19
LGMN	33	RGS13	26	HBD	22	CD79B	19
CD27	32	DENND1B	26	CDKN3	22	SLC16A6	19
MYO1D	32	C11orf80	26	STARD5	22	RASGRP3	19
PLPP5	32	TOP2A	26	CENPN	22	AURKA	19
FCRL5	31	KLHL14	26	TUBG1	22	GPR160	19
DLGAP5	31	COBLL1	26	CD79A	22	DCUN1D1	19
CCNB2	31	DERL3	26	SLC1A4	22	FAM169A	19
SEL1L3	31	PDIA5	26	TPX2	22	DSP	19
SLAMF7	30	ESCO2	26	UGT2B17	22	HIST1H2AM	19
BLNK	30	FBXO16	26	HIST1H2BM	22	CHN2	19
CHAC2	30	CXCR3	26	HIST2H3D	21	NME8	19
CPNE5	30	FAM46C	25	TMEM133	21		

B.10 Genes specific for high responder subject class (day 7)

Gene	p value	Gene	p value
APOBEC3B	0.001	UGT2B17	0.011
IRF4	0.001	HIST1H2AM	0.011
HIST1H3G	0.001	MAN1A1	0.014
GLDC	0.001	JCHAIN	0.017
MKI67	0.002	CDK1	0.018
POU2AF1	0.002	LMAN1	0.018
TOP2A	0.002	SLAMF7	0.022
TYMS	0.002	LAMC1	0.022
NT5DC2	0.002	DTL	0.022
CD38	0.002	HRASLS2	0.022
CCNA2	0.002	DENND5B	0.022
CPNE5	0.002	CCNB2	0.022
ITM2C	0.003	KLHL14	0.022
CHAC2	0.004	ERLEC1	0.022
SEC11C	0.004	PDIA6	0.022
ZBP1	0.005	DERL3	0.022
HIST2H3D	0.006	MANEA	0.022
ELL2	0.006	HIST1H1B	0.022
KIF11	0.006	MOXD1	0.022
DLGAP5	0.006	PNOC	0.022
SHCBP1	0.006	GGH	0.022
MYO1D	0.006	MYBL2	0.027
RRM2	0.006	PHGDH	0.037
SEL1L3	0.006	LAX1	0.037
HIST1H2BM	0.006	ASPM	0.037
HIST1H3B	0.006	KNL1	0.037
HIST1H3F	0.006	LARP1B	0.037
HIST1H3J	0.006	FAM46C	0.039
ESCO2	0.006	SLC35F2	0.039
GPRC5D	0.009	CDC6	0.039
CAV1	0.009	COBLL1	0.039
BUB1	0.011	FBXO16	0.039
TPX2	0.011	BHLHE41	0.048

Table B.8: Subpool of 66 genes specifically modulated in the *high responder* subject class. P-values are calculated via Fisher exact test, one sided, with Benjamini Hochberg correction for multiple testing.