Universiteit Antwerpen

Multimodal Data Integration for Predictive Modelling of Measles Vaccine Response with Cross-Vaccine Marker Validation

Elias Dams

Promotor: Dr. Pieter Meysman
Supervisor: Fabio Affaticati

University of Antwerp
Faculty of Science

2024-2025

Submitted in fulfilment of the requirements for the degree of
Master in Computer Science: AI & Data Science

June 2025

Contents

1 Introduction **8**

2 Background **9**
2.1 Background in Biology 9
2.1.1 Immune System Overview 9
2.1.2 Antibody Titers 11
2.2 Background in Computer science 12

3 Data Exploration and Preprocessing **13**
3.1 Label Assignment Strategy 13
3.2 Antibody Titer Trajectories 13
3.2.1 By Response Label 13
3.3 Correlation Analysis Within Individual Datasets 15
3.3.1 Methodology 15
3.3.2 Cytokine Data 16
3.3.3 Cytometry Data 17
3.3.4 TCR Metrics 17
3.3.5 RNA Data 17
3.4 Integrated Correlation Analysis 18

List of Figures

2.1 Diagram of Innate and Adaptive Immunity 10

3.1 Antibody Titer Trajectories by Response Label 14
3.2 Antibody Titer Trajectories by Original Label (Quadrant) 14
3.3 Correlation heatmap of the cytokine data using Ward.D2 clustering. . . . 16
3.4 Correlation heatmap of the cytometry data using Ward.D2 clustering. . . 17

List of Tables

List of Acronyms

- TCR: T-cell receptor

Summary

Acknowledgements

Abstract

CHAPTER 1

Introduction

Vaccination is one of the most cost-effective strategies to prevent infectious diseases; however, individuals often exhibit very different responses to the same vaccine. This thesis focuses on predicting responses to the measles vaccine using a data-driven, multimodal approach. The ultimate goal is to develop predictive models that not only forecast individual vaccine responses but also validate key immune markers across different vaccines. This process is known as cross-vaccine marker validation.

From a computer science perspective, my work leverages machine learning techniques to integrate diverse types of biological data. In this study, I combine measurements of immune system responses, such as cytokine levels and cell counts obtained from cytometry, with molecular profiling data, including antibody titers and T cell receptor (TCR) sequences. For instance, antibody titers, which quantify the concentration of specific antibodies in the blood, serve as a direct indicator of the immune system's ability to neutralise pathogens. Cytokine levels provide early signals about the body's readiness to respond, while TCR sequencing reveals which T cell clones are active and specific to the vaccine antigen.

Initially, I work with measles data from a study comprising 40 samples. While this dataset provides a valuable starting point, its relatively small size poses challenges in terms of statistical power and model generalisability. Furthermore, integrating heterogeneous data types introduces additional difficulties, such as differing scales, potential noise, and missing values, all of which must be carefully addressed during data preprocessing and feature selection.

TODO: info about the study my data is coming from.

In the following chapters, I will detail the processes of data exploration, feature selection, model building, and validation that underpin this multimodal approach. This work aims not only to improve our ability to predict vaccine responses but also to enhance our understanding of the immune markers that are most relevant for effective immunisation across different vaccines.

CHAPTER 2

Background

To understand the work presented in this thesis, it is essential to have a basic grasp of concepts from both immunology and data science. As a computer scientist, my approach is mainly data-driven, focusing on extracting, integrating, and analysing various types of biological data. However, a foundational understanding of the underlying biological processes is critical to meaningfully interpret the results and validate the predictive models developed in this research.

2.1 Background in Biology

2.1.1 Immune System Overview

First of all, I would like to give an overview of how the immune system works. The immune system can be roughly divided into three main parts, as depicted in Figure 2.1:

Physical barriers (top)
Physical barriers such as the skin and mucous membranes form the body's first line of defence by preventing most pathogens from entering.

Innate Immunity (right)
In case a pathogen still crosses the barriers, innate immunity comes into action. This defence is rapid and non-specific. Think of macrophages and neutrophils as cells that engulf invaders through a process called phagocytosis. Eosinophils and other granulocytes also attack pathogens or initiate inflammatory responses. Natural killer cells (NK cells) are also part of innate immunity and can directly destroy infected or abnormal cells. Although this response is very rapid, it does not recognise pathogens in the same specific way as the next branch. [2]

Adaptive Immunity (left)
Acquired or adaptive immunity is the "slower but more targeted" defence. B cells are an important part here. They are responsible for producing antibodies, which are pro-

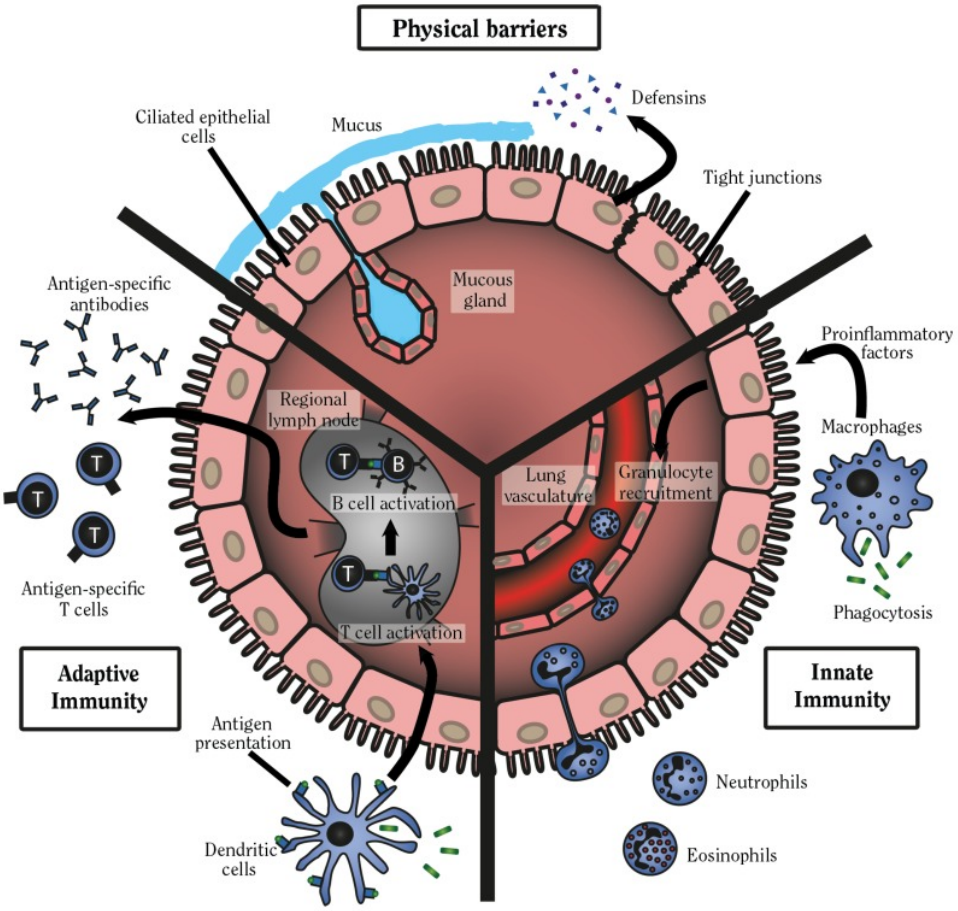

Figure 2.1: Diagram showing physical barriers, innate immune cells (e.g., macrophages, dendritic cells, natural killer cells) and adaptive immune components (B and T cells) working together. Reproduced from Figure 10.5 in *The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General* (2014) [1].

teins that bind to specific antigens (foreign substances) to neutralise or mark them for destruction. The level of these antibodies in the blood is often measured as "antibody titers". Higher titers generally indicate a stronger immune response. T cells are also crucial and have several roles. They help coordinate the immune response (often referred to as "helper T cells") and can directly kill infected cells (cytotoxic T cells). T cell receptors (TCRs) are highly specific and can be sequenced to understand which T cell clones are active in response to a vaccine. A major advantage of adaptive immunity is that it "learns" from previous exposure, allowing for much faster and more powerful immune responses in the event of repeated infection. This ability to form memory also underlies how vaccines work. [2]

Together, these three pillars provide a robust defence system that is able to successfully ward off or fight off most infections. It is also this dynamic between innate and adaptive immunity that determines the degree of vaccine response. the innate branch prepares the way, while the adaptive branch provides targeted antibodies and memory cells. This is important to understand because my models integrate elements of both the innate and adaptive immune systems. For example, innate markers such as cytokine levels and certain cell counts (measured via cytometry) can provide early signals about

the body's general readiness to respond. Meanwhile, adaptive markers, such as TCR sequences of T cells, directly reflect the specific immune response that leads to antibody production after vaccination. Combining this data gives us a more complete picture of the immune landscape, allowing me to better predict how effectively an individual will respond to the measles vaccine.

2.1.2 Antibody Titers

In addition to assessing the cellular components of the immune response, it is crucial to quantify the functional output of the adaptive immune system, namely antibody production. Antibody titers provide a quantitative measurement of the concentration of specific antibodies in the blood, and essentially reflect the "signal strength" of the immune response against a particular antigen. High titers usually mean that the immune system is actively fighting off the pathogen by targeting and clearing it effectively. In contrast, low titers indicate a weaker response and thus less robust protection.

To determine antibody titers in the laboratory, serial dilution tests are often used. In these tests, a serum sample is progressively diluted and each dilution tests its ability to bind to the target antigen. The titer is defined as the highest dilution at which antibodies can still be detected. This method allows a practical estimate of the antibody concentration in the original sample.

In my research thesis, antibody titers serve as an important biomarker to label vaccine response categories. By categorising individuals based on their antibody titers, I can distinguish between strong and weak responders. This classification is essential for developing and validating predictive models because it provides a clear, quantifiable endpoint that reflects the effectiveness of the measles vaccine in eliciting a protective immune response.

2.2 Background in Computer science

TODO: info about the used computer science techniques.

CHAPTER 3

Data Exploration and Preprocessing

In this chapter, I describe how the measles antibody titer data was prepared and how response labels were assigned. **TODO:** ...

3.1 Label Assignment Strategy

The original dataset included antibody titer measurements at four time points (Day0, Day21, Day150, and Day365), as well as detailed qualitative classifications for each subject (e.g. `responder`, `no response - high ab` and `no response - low ab`). However, these classifications were too specific for the initial stages of modeling. Consequently, I reduced the labels to a simple, two-class scheme (e.g. `responder` and `non-responder`) to create a more straightforward prediction task that can be refined later if necessary.

3.2 Antibody Titer Trajectories

3.2.1 By Response Label

Figure 3.1 shows the antibody titer trajectories over four time points, separated by each subject's final response label (`response` or `no response`). On the x-axis, we have the days at which titers were measured (0, 21, 150, 365). The y-axis represents the titer level. Each line corresponds to a single subject's progression across these time points. From this visualization, it is evident that "responders" generally exhibit a marked increase in titers between Day 0 and Day 21 (and sometimes up to Day 150), whereas "non-responders" show either a small rise or a plateau. This distinction validates the simplified two-class approach used for model building. In contrast, Figure 3.2 displays the antibody titer trajectories based on the original, more granular labels (i.e., `responder`, `no response - high ab`, and `no response - low ab`).

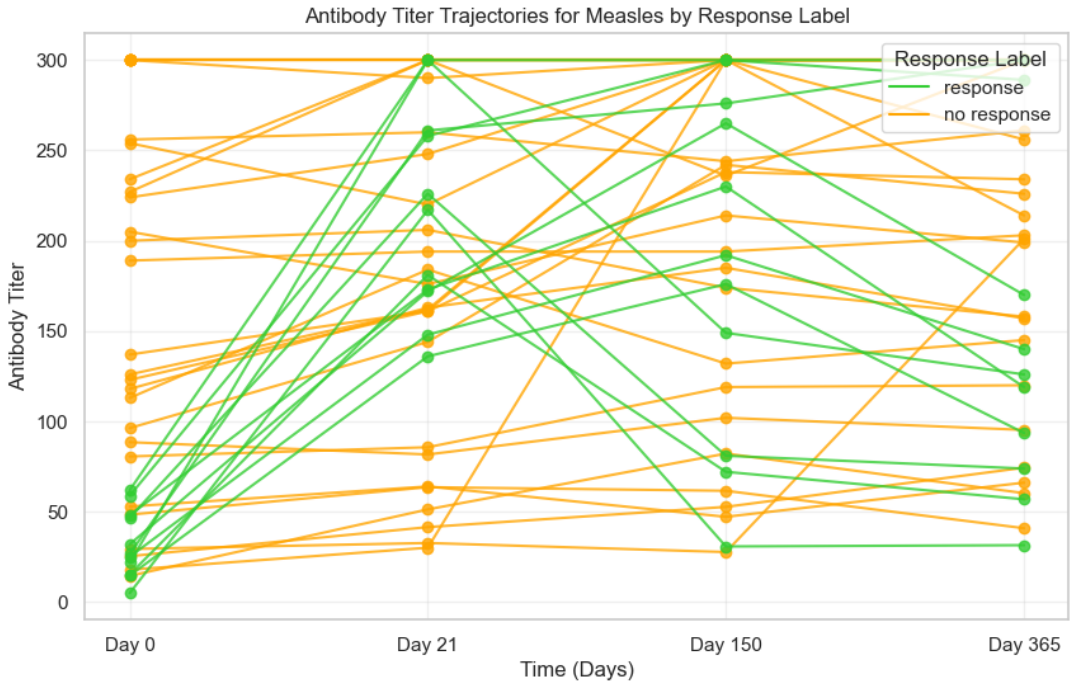

Figure 3.1: Antibody titer trajectories for each subject, colored by final response label. Subjects labeled as `response` are shown in green, while those labeled as `no response` are shown in orange.

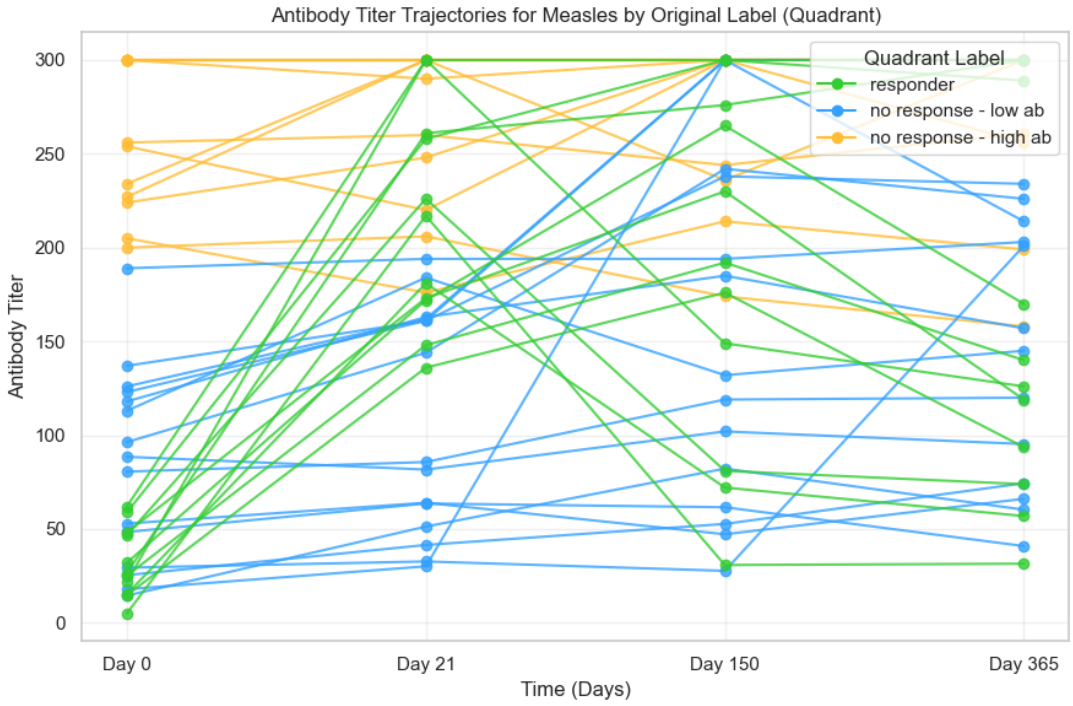

Figure 3.2: Antibody titer trajectories for each subject, colored by the original qualitative labels (quadrant). This view highlights the detailed categorization of responses, offering insights into the nuances of each subject's titer progression.

The simplified `response_label` variable is used throughout the thesis as the target for classification tasks. This choice reduces complexity, addressing both data imbalance and interpretability concerns. While the dataset initially contained more nuanced labels (e.g., distinguishing "no response - high ab" from "no response - low ab"), combining them into a single "no response" category improves the feasibility of training robust machine learning models given the limited sample size.

3.3 Correlation Analysis Within Individual Datasets

In this chapter, I explore the interrelationships among the variables in the different datasets. The data consist of five distinct datasets:

- Cytokines
- Cytometry
- clonal breadth (TCR metrics)
- clonal depth (TCR metrics)
- RNA data

Because each dataset captures different aspects of the immune response, I begin by investigating correlations within each individual dataset. Subsequently, I concatenate the datasets to perform an integrated correlation analysis. This approach allows me to uncover both within-modality and cross-modality relationships.

3.3.1 Methodology

TODO: Check if right... In my analysis, I use a Weighted Gene Co-expression Network Analysis (WGCNA) framework to identify modules of highly correlated features. First, I compute a correlation matrix from the data, which quantifies the pairwise relationships between features. This correlation matrix is then transformed into a distance matrix by taking one minus the absolute correlation value, ensuring that strongly correlated features are considered close. Instead of relying on a fixed linkage method like Ward.D2, WGCNA integrates hierarchical clustering with network analysis to detect modules, or clusters, of co-expressed features. In the implementation, I use the `flashClust` package for efficient hierarchical clustering, and then apply a dynamic tree cut procedure to define the modules. The modules are further visualized by assigning each a unique color using WGCNA's labeling functions and displaying the resulting dendrogram and heatmap plots. This approach captures complex correlation patterns, effectively groups similar features. Moreover, by merging datasets from different modalities, we can explore both within-modality and cross-modality relationships, ultimately facilitating more effective downstream machine learning tasks.

3.3.2 Cytokine Data

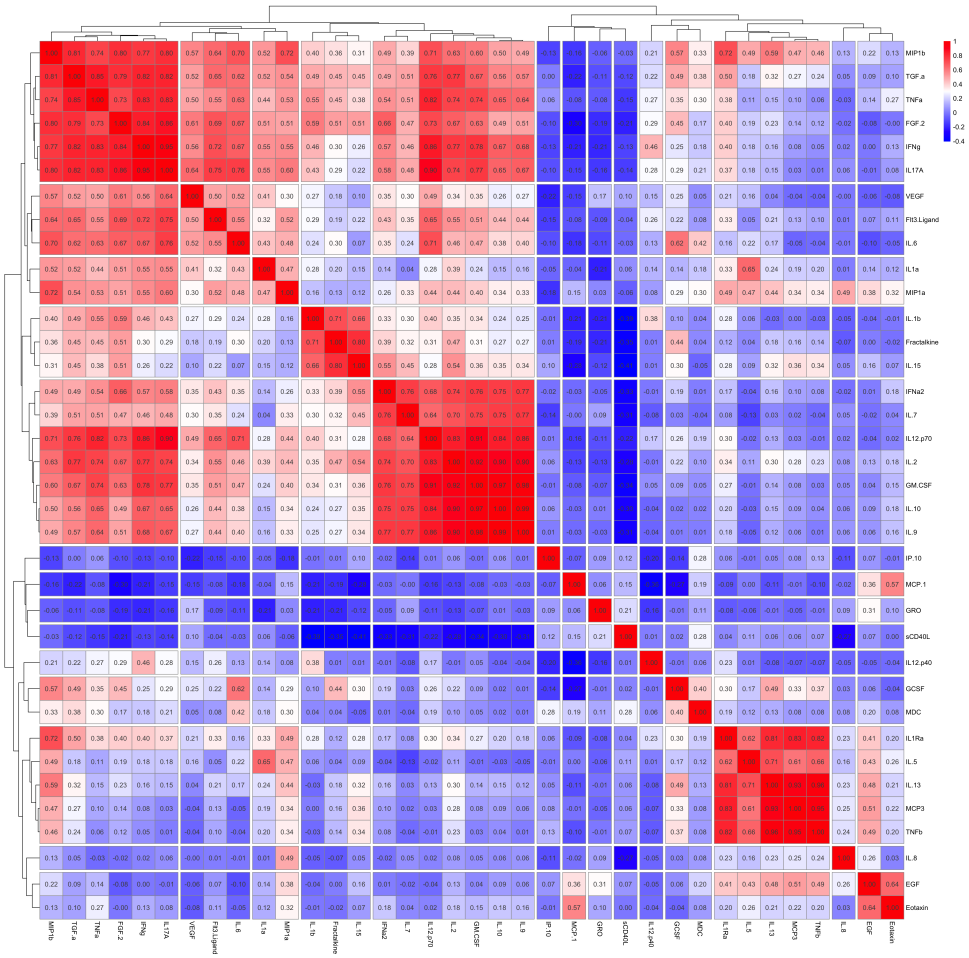

Figure 3.3: Correlation heatmap of the cytokine data using Ward.D2 clustering.

The following cytokine clusters were obtained from the WGCNA hierarchical clustering analysis:

- **Cluster 1:** MIP1β, TGF-α, TNF-α, FGF-2, IFN-γ, IL17A.
- **Cluster 2:** VEGF, Flt3 Ligand, IL-6.
- **Cluster 3:** IL1α, MIP1α.
- **Cluster 4:** IL-1β, Fractalkine, IL-15.
- **Cluster 5:** IFNα2, IL-7, IL12-p70, IL-2, GM-CSF, IL-10, IL-9.
- **Cluster 6:** IL1Ra, IL-5, IL-13, MCP3, TNFβ.
- **Cluster 7:** GCSF, MDC.
- **Cluster 8:** EGF, Eotaxin.

3.3.3 Cytometry Data

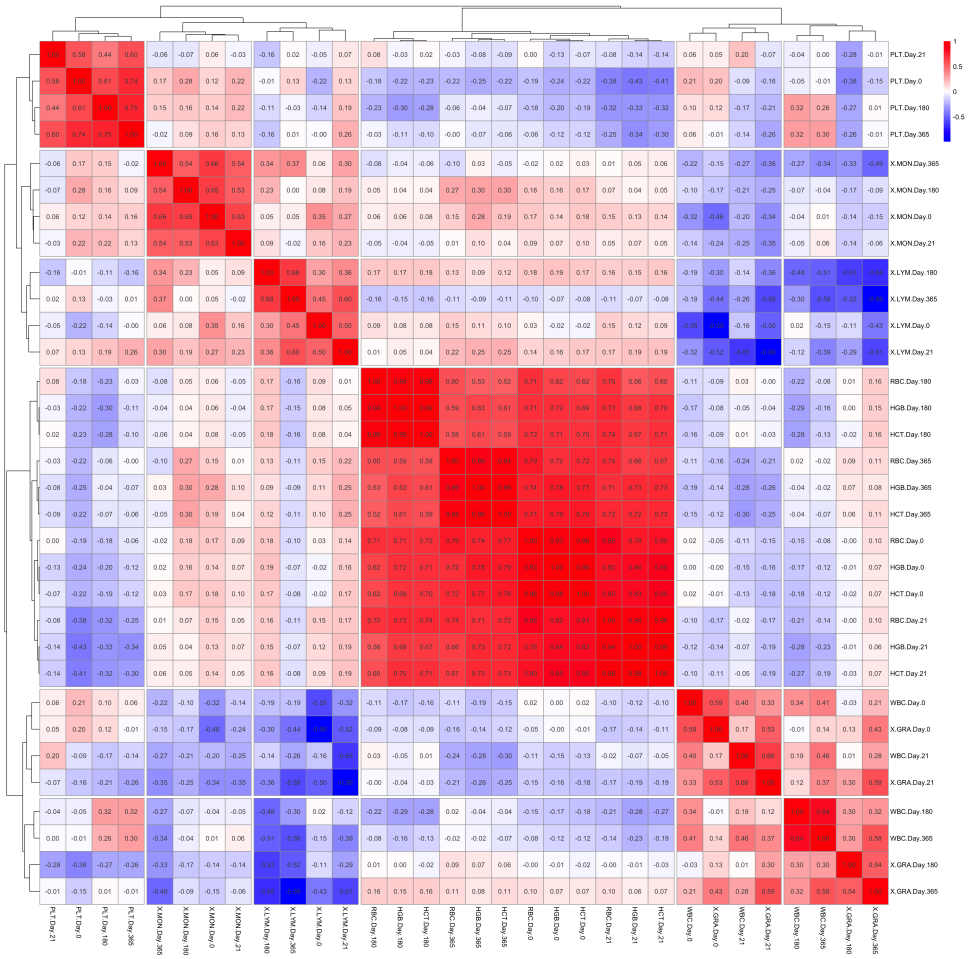

Figure 3.4: Correlation heatmap of the cytometry data using Ward.D2 clustering.

3.3.4 TCR Metrics

- **Cluster 1: (RBC, HGB, HCT)**
  RBC (red blood cell count), HGB (hemoglobin), and HCT (hematocrit) are all interrelated measures of erythrocyte mass and oxygen-carrying capacity. Because they essentially quantify the same biological component—namely, red blood cells in circulation—their values often move in tandem. At Day 1, the heatmap reveals high correlation coefficients (e.g., 0.89–0.99) among RBC, HGB, and HCT, and similar correlation patterns emerge at Day 0, Day 21, Day 180, and Day 365, reinforcing their consistent co-variation over time. Ward.D2 naturally groups these three parameters into a single cluster by minimizing within-cluster variance. Even when only Day 1 is considered, RBC, HGB, and HCT remain tightly linked, underscoring their close functional relationship.

3.3.5 RNA Data

TODO: ...

3.4 Integrated Correlation Analysis

After analyzing each dataset separately, I combine them into two integrated datasets. The first integrated dataset merges the Cytokine and Cytometry data, while the second integrates Cytokine, Cytometry, and RNA data. The second combined dataset will be used in later stages of the thesis. This integrated analysis provides insights into how variables from different data types (e.g., cytokine levels, cell counts, and RNA expression) correlate with one another. Although this approach increases the number of correlations observed, it also introduces challenges such as higher dimensionality and potential confounding effects.

TODO: ...

Bibliography

[1] *The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General*. Centers for Disease Control and Prevention, Atlanta, GA, 2014. Figure 10.5: Diagram of innate and adaptive immunity.

[2] C. A. Janeway, P. Travers, M. Walport, and M. J. Shlomchik. *Immunobiology (9th Edition)*. Garland Science, 2001. innate vs adaptive immunity.

[3] F. Murtagh and P. Legendre. Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion? *Journal of Classification*, 31(3):274–295, 2014.