May 2013

Statistical Investigation of the Immune Response in Non-Human Primate Models

Annika Laser
University of Wisconsin-Milwaukee

Statistical Investigation of the Immune Response in Non-human Primate Models

by

Annika Laser

A Thesis Submitted in
Partial Fulfillment of the
Requirements for the Degree of

Master of Science
in
Mathematics

at

The University of Wisconsin-Milwaukee
May 2013

ABSTRACT

Statistical Investigation of the Immune Response in Non-human Primate Models

by

Annika Laser

The University of Wisconsin-Milwaukee, 2013
Under the Supervision of Professors

The human immunodeficiency virus (HIV) was first detected more than 30 years ago. Since then, intensive research has been done to develop a broadly protective vaccine, though without success. Our goal is to unveil some features of the protective immunity in non-human primate lentiviral infections in order to emulate HIV-infection. Two primate species have been studied, rhesus macaques (Rh) (*Macaca mulatta*) and African green monkeys (Ag) (*Chlorocebus spp.*). Simian immunodeficiency virus (SIV) infection is non-pathogenic to Ag while Rh develop an AIDS-like illness. In this study, peripheral blood mononuclear cells (PBMC) from 8 Ag and 27 Rh were stimulated with phorbol merystate acetate and ionomycin to activate lymphocytes regardless of their specificity. We hypothesize that the immune response of the two species is fundamentally different resulting in the different reactions to SIV infection. CD4+ and CD8+ T-cells were investigated with respect to multiple surface markers and production of gamma-interferon

(IFN), tumor necrosis factor alpha (TNF) and interleukin two (IL2).

Additionally to principal component analysis, we tried a new approach by using exploratory factor analysis to reveal latent influences. We found differing relations for both Ag and Rh especially among cytokine secretion patterns. Based on our results, it is assumable that, besides their clear biological interaction, the TNF and IL2 are dependent on a latent factor in the Ag. However, this strong relation could not be found in Rh. Instead, TNF and IL2 seem to oppose each other for Rh because they are assigned to different latent factors.

Table of Contents

List of Figures

ACKNOWLEDGMENTS

First of all, I would like to express my gratitude to my advisors, Prof. Gabriella Pinter and Prof. Istvan Lauko. With their help I found a research topic that fit my interests and challenged me in different ways. As a result of their support, I was able to finish on time, a major obstacle that they helped me overcome.

Moreover, I want to thank Dr. Wail Hassan for his great cooperation and for providing the data, which were essential for the completion of this project. Also, I extend my gratitude to Prof. Chao Zhu for serving on my thesis committee.

Further, I would like to thank all those who made not only this thesis, but my entire study year in the US an unbelievable and unique experience. I would like to especially thank my roommates, for simply sharing a living space, Alexis H. Smith for staying up late, editing my thesis and being my professional English advisor and M. J. Wahl for discovering the US with me and for setting my mind to rest whenever needed and without whom I would never be where I am right now.

Last but not least, I would like to thank all my friends back home, whom I can meet all over the world and with whom I can stay in touch no matter where we are. Furthermore, I say thanks to my whole family, who keep believing that whatever I think is best for me actually is best for me, and who support every idea I come up. Mom, and dad, thank you for always trusting in me no matter what. Danke für einfach alles.

Milwaukee, May 2013

1 | Introduction

To date, there is no cure for the infection by the human immunodeficiency virus (HIV) ([34]). However, there are several ways of transmission. As often believed, not only gay, bisexual and other men who have sex with men (MSM) are affected, although these of 'all races and ethnicities remain the population most profoundly affected by HIV' ([34]). Still, 25% of new HIV infections in 2010 in the United States is accounted for by heterosexual contact. In 2012, an estimated approximately 50,000 people were diagnosed with HIV infection in the US. In the same year an estimated number of 32,000 people were diagnosed with the acquired immune deficiency syndrome (AIDS), being the final stage of the HIV-infection ([34]).

HIV exists with two types, HIV1 and HIV2. Usually, when it is simply referred to as HIV, the HIV1-type is being referred to. In contrast to other viruses, the human immunodeficiency virus does not randomly attack cells of their hosts' bodies, but it chooses cells of the immune system itself. Thus, while the virus replicates, more and more of these cells, the CD4 positive (CD4+) cells are affected, so that they cannot efficiently fight against the intruder. AIDS is called the last stage of the infection because at this point, the person's immune system is vastly damaged and thereby highly exposed to other diseases and certain cancers.

Nevertheless, there are medications available to retard the progression of HIV and improve the health status of the patients. However, the infected persons must be aware of their disease, which is not the case for one out of five HIV-infected person in the US. Additionally, today's available treatments have to be taken daily and only assure a slow

progression, not a cure of the disease being conducted.

Thus, extensive research is done to find out how to control HIV-infections and the implicated AIDS. One major goal is to find a vaccination against the virus itself.

In the following work, we will examine data on both rhesus monkeys (Rh) and African green monkeys (Ag). For non-human primates there exists an HIV-equivalent, the simian immunodeficiency virus (SIV) and, there exists a connection between HIV and SIV (see chapter 2 for more details). In this work we aim to perform a comperative study of the immunological response of the two primate species.

While rhesus monkeys develop an AIDS-like illness upon SIV-infection, just like humans upon HIV-infection, the African green monkeys seem to be resistant to the disease despite vigorous viral replication. Hence, we used the two species as model for protective and non-protective immune response. Cells from both Rh and Ag were examined being unstimulated and after stimulation by phorbol merystate acetate and ionomycin to stimulate lymphocytes regardless of their specificity.

As the reactions of both species upon SIV-infection are known, our hypothesis assumes the immune responses to be fundamentally different.

The statistical methods used aim to differentiate the two groups. On one hand, we therefore used a Mann-Whitney-U-test in order to detect significant differences between the groups. On the other hand, we applied multivariate statistical methods. These included principal component analysis as it was done in previous studies and additionally exploratory factor analysis.

From both African green and rhesus monkeys, cell populations were taken and investigated. As mentioned above, measurements of these cells were taken without stimulation and likewise after stimulation by phorbol merystate acetate and ionomycin . As the study was based on data from 35 specimens in total, 8 African green and 27 rhesus monkeys, any conditions of statistical method might not be sufficiently met.

While the Mann-Whitney-U-test points out statistically significant differences between the measurements taken for both Ag and Rh, the principal component analysis and the factor analysis detect relations among the variables. With the help of the latter, we are able to detect latent, hidden, factors influencing the variables. Thus, we might discover differences of these factors for the two species rather than only differing measurements for single variables.

Using these methods, we assume certain patterns to establish for the immune response of the cells of African green and rhesus monkeys.

In the first chapter, the biological background is explained. All variables occuring in the datasets are described and put into context. In the following part, the main statistical methods are described. Additionally, the factor analysis is explained in detail, especially the method used for the following evaluations. The datasets are given first with an evaluation for all the findings per dataset respectively. In the last chapter, all results are summed up to make a general assumption.

2 | Biological Background

The immune system of mammals is highly developed. African green monkeys (*Chlorecebus spp.*) [29, 28] and Rhesus monkeys (Rhesus macaques, *Macaca mulatta*) are often used for research on Human Immunodeficiency Virus (HIV) and the Acquired Immune Deficiency Syndrome (AIDS) [32, 31]. These two species unlike other non-human primates when infected by the Simian Immunodeficiency Virus (SIV), the primate equivalent to the HIV, present some useful and interesting features. Researchers know how these two types if primates react to the infection, so these two are studied to better understand why the animals react differently to the infection. During the acute phase, similarly for humans and primates, the virus is very active and continues to replicate along the entire course of infection. At an early stage, the infected subjects experience what is called clinical latency where the patients do not show symptoms of immune deficiency while viral replication proceeds. The initially still intact host immune system gets debilitated due to continuing CD4+ T-cell death (see below for further explanations).

While the Rhesus macaques (Rm) respond smilarly to humans (due to an increased destruction of the immune system AIDS may occur), the African green monkeys (Agm), in contrast, seem to be resistant to the virus [6]. They do not develop HIV-like diseases upon infection with SIV despite vigorous viral replication. Thus, Agm's are often used as control, non-pathogenic, groups whereas Rm's are the pathogenic 'probands' in different trials and studies.

HIV was first detected in the US at the beginning of the 1980s. It is assumed that the virus was in the country since the late 1960s [33]. There are different opinions on how the virus infected a human being the first time. The most believed one though is that SIV

of African monkeys mutated and crossed the species, causing HIV to occur in humans. How this crossing happened, is still highly debated. However, it was also discovered that there is a connection between Haiti and the first recorded HIV-infected patient in the US [30]. One reasonable explanation for this would be that the virus first made its way to Haiti, probably already via humans, and from there directly to the US. Still, these are just suggestions on certain discoveries but are neither proven nor accepted by all researchers.

It is important to understand that, despite disputes about the origin of HIV, these monkeys are chosen so often for experiments, because of the similarity between SIV and HIV and, especially, because of their differing reactions to the viral infection [7].

Before getting to the data itself, we have to clarify the context. Although we use monkeys in most experiments, our main focus is on humans.

Basically, the immune system consists of two parts, the innate and the adaptive immune system. The innate, also called unspecific, immune system is what keeps away any pathogen trying to invade our bodies. This includes mucosae, enzymes on the skin and in the gastrointestinal tract, and the first cellular defence. This includes granulocytes and macrophages which detect patterns of pathogens and destroy those. Similarly, the complement system, responsible for lysing bacteria, is part of the innate immune system. What these cells are specifically will be discussed below. However, if a particular pathogen resists all these barriers, another mechanism starts to work, the adaptive, or specific immune system. It is called 'system' because countless mechanisms have to work hand in hand to fight against the intruder. The main actors herein are cells, to be more specific, white blood cells, called leukocytes. In fact, though the majority of leukocytes are part of the innate system, still subgroups of leukocytes are the most important ones for the adaptive immune system.

Leukocytes can be divided into different subgroups. The clear majority form the granulocytes. The biggest white corpuscles though, size-wise, are the monocytes. These are equally important for both adaptive and innate immune system as they are progenitors

for the so called macrophages. Monocytes circulate in the blood and eventually leave the blood stream for tissues. There, they mature into macrophages. It can be said that macrophages are the tissue-form of monocytes. Additionally, there are so called natural killer cells (NKC), which simply 'kill' other cells. This means that both NKC as well as macrophages are able to perform phagocytosis. In simple words, one might say that they are able to digest pathogens. The pathogen is internalized by phagocytic cells and eventually, digested within that cell. Thus, NKC and macrophages play a huge part in the innate immune system. Macrophages are equally important for the adaptive immune system, because here they act as scavengers. While digesting pathogens, they display the antigen on their surface thereby activating other immune cells, especially lymphocytes.

Lymphocytes, also, are white blood cells. Two major groups are distinguished, B-cells and T-cells. These two groups are named for the origin of the cells themselves. T-cells maturate mainly in the thymus while B-cells come from the bone marrow. The latter produce antibodies against specific antigens. While B-cells recognize free antigens and are activated by their interaction, they also present this antigen on their surfaces. This helps to activate T-cells and other processes of the adaptive immune system.

Once activated, the production of antibodies may start. Roughly speaking, antibodies and antigens accumulate forming one complex that can be detected by e.g. macrophages and thus, be destroyed. B-cells can be activated right away. T-cells also work as mediators here, these are the so called T-helper cells (THs). Beside those, there are the T killer cells (TKs), or also called cytotoxic T-cells. As their name suggest Tks kill the pathogenic cells, i.e. virus-infected or similarly tumor cells, as soon as they can be detected as such, similar to macrophages in the adaptive immune system. Some authors additionally define a third group, inflammatory T cells. However, in the following, we will consider the usually defined two groups, THs and TKs. As stated above, both types of T-cells are activated by represented antigens on other cells' surfaces and most of them do need a second stimulator to work properly [14],[15]. Now the questions arise how this co-stimulation is carried out and how to distinguish all kinds of lymphocytes?

The answer lies in surface markers being tied to the membrane of the lymphocytes.

Most of those are proteins. To distinguish the vast amount of proteins, they are classified into clusters of differentiation (CD), which also give them their name. For example, all T-cells are CD3 positive (CD3+). If a a cell is not CD3+, it is not a T-cell. Similarly, TKs and THs are differentiated. While the first ones are CD8+ (and CD3+ as well as they are T-cells), the latter ones are CD4+ [10]. Co-stimulation of the T-cells also is based upon these surface proteins. If a T-cell displays a particular surface marker, it can be activated by the appropriate ligand. However, the measurements taken within an experiment usually concentrate on the amount of proteins on the cells' surfaces or cytokines being secreted by the cells rather than on the activations themselves due to technical limitations. Today's researchers often measure the so called mean fluorescence intensity (MFI) of particular proteins [8]. It should be stated that the MFI does not count the number of cells presenting a certain surface protein, but it gives information about its intensity. It is a semi-quantitative measurement of the amount of protein present per cell. Thus, by considering the MFI, we deal with a measure of protein expression in a population of cells.

Consider the following example. Given three cells, all of them called CD4+, because they have at least one CD4 surface marker, we now want to find out the MFI of CD3.

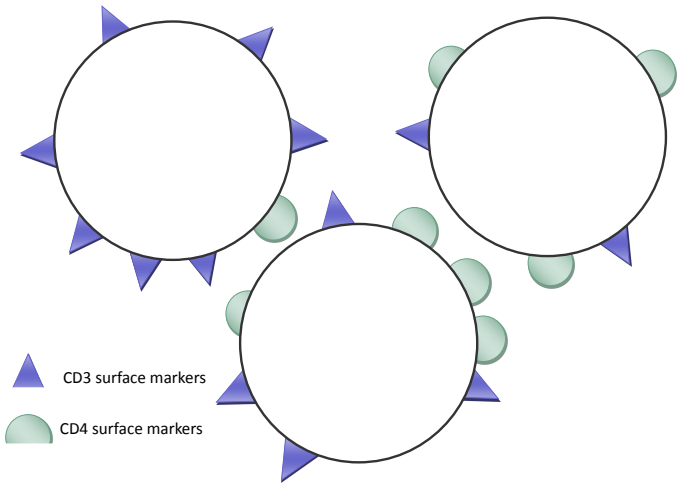


Figure 2.1: Illustration of surface markers

Figure 2.1 shows three cells, each of them having both CD4 surface markers and CD3 surface markers. Now, the MFI describes the average number of surface markers for CD3 on those given CD4+ cells. In our case it is

$$\frac{(7 + 4 + 3)}{21} = 66\% .$$

This can be done for other surface markers like CD28 or CD95 or even for CD4 surface markers on the CD4+ cells. In fact, being CD4+ does not say anything about the amount of CD4 expressed on the cell's surface.

Other properties of both T- and B-cells can also be defined by taking a look at the MFI of certain cytokines. First, both can develop into effectory and/or memory cells. We speak of effector cells whenever they were activated because of an intruder. Besides those, there are memory cells. This feature is one of the most important ones of the human immune system. Again, researchers did not come up with one unique satisfactory explanation. Most authors will state that those effector cells, which were ready to fight but could not do so because the intruder had already been removed successfully, become memory cells [7]. However, other opinions are that there either exists some kind of memory stem cells [23] or that e.g. CD8+ cells differentiate directly into a cell with properties of the central memory[22].

Here, another special feature comes into play, the 'dual personality of memory T cells' [21]. This distinction, again, can be drawn between their differing surface proteins. The so called central memory (CM) of T-cells, either CD4+ or CD8+, is characterized by a high proportion of CD95+ and CD28+ and a low count of CD45RA+, i.e. these cells are called CD45RA-, whereas cells of the effective memory (EM) show a rather low count of CD28, still a high proportion of CD95+ and can be either CD45RA+ or CD45RA-. Naïve cells are described as CD95- CD28- CD45RA+. It is obvious that there have to be differences in the function of cells of the EM, cells of the CM, and simply the naïve T/B-cells ,[22], [20], due to their differently expressed cytokines. However, the reason for those EM and CM cells to be called 'memory' is easy to understand. These cells, no matter

which kind, have the task to remember specific antigens. If the same antigen is displayed on any macrophage's surface, the human body knows how to react. This means, that the B-cells already know which antibody to produce, i.e. some memory B-cells just transform into effector cells right away. This property of the human immune system allows us to use vaccinations. Using an inoculant causes an unharmful infection by a certain virus, allowing the body to create a memory. In case of a 'real' infection, the cells are able to react more rapidly [13]. Within the data given, these three groups are distinguished on the one side, on the other side, in some cases an overall measurement for all CD4+/CD8+ cells was also taken.

As mentioned above, both kinds of cells need a co-stimulation in order to get activated [2]. This is, of course, also true when distinguishing them as naïve cells and EM and CM. Relevance, function, and interaction of these co-stimulators is highly debated as well and almost every study and researcher comes up with a new finding. Thus, in the following, we summarize the main tasks of the proteins measured in the data.
First, there is the CD95-ligand, which binds to CD95 or Fas [12]. It is also called the Fas death receptor in some literature [16]. It is known as this is the stopping protein, because it enhances the T-cell elimination, i.e. self-destruction or apoptosis, hopefully when the pathogen has been fought off successfully [23]. It was discovered, that CD8+ cells are more sensitive to CD95+ apoptosis [17]. Additionally, in the case of HIV, the infected cells express Fas, but they themselves are resistant, so that only uninfected cells are killed (see e.g. [18],[14],[15]). However, as with all cytokines and surface proteins, different assumptions can be found.

CD28 binds to B7, which works as a second signal for the T-cells to be activated [27]. CD45 acts similarly. As mentioned before, CD45 helps to distinguish EM and CM cells. It usually gets the suffix CD45RA or CD45R0 (for more details see e.g. [19]). Equivalently, CD28 helps to distinguish EM and CM. To simplify this relation, one might think of a distinction as the following:

There are many more CD's, however we only focus on the ones used in the data. Besides surface markers, there are other factors influencing the T-cells' work. One of these

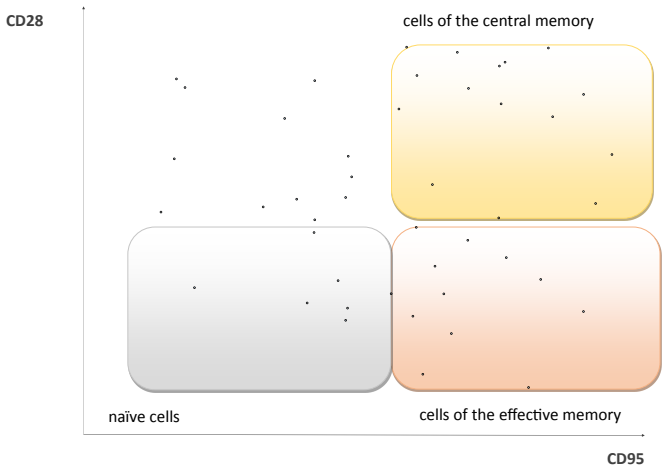


Figure 2.2: Distinction between naïve cells and cells of the central and effective memory

is Interleukin 2 (IL2), again a protein. It is secreted by T-cells and their ligand is also expressed on T-cells where they work as costimulators to support growth and proliferation [5],[3]. This is a way that especially antigen-specific T-cells try to ensure their survival. IL2 also induces Interferon-γ (IFN) production which, on its side, inhibits the proliferation of various pathogens by e.g. activating macrophages or inhibiting viral replication directly [27]. Thus, it has both immunostimulatory as well as immunomodulatory effects [4]. However, IFN-γ is required for the expression of IL2 receptors on the T-cells' surfaces [25]. That is why IFNγ is produced not only due to the presence of IL2 but also by NK cells and CD8 effector cells. NK cells, as well as macrophages and CD4+ cells also induce the production of the tumor necrosis factor α (TNF). TNF-α equally works immunomodulatory and inhibits viral replication. All these processes work hand in hand and sometimes in opposite directions. Mengozzi [2] for example speaks of a 'cross-regulatory phenomenon' among CD4-T-cells.

The measurement here works a bit differently then for the surface markers. As said, T-cells secrete IL2, IFN-γ and TNF-α. In order to determine how many of the cells are capable of producing either or more of these, first the T-cells are oppressed to secrete

these cytokines so that they can accumulate within the cell. This accumulation is needed to measure MFI.

After having stated all these connections within the immune system, it shall be reminded that the data given do measure the MFI only and not the 'real' activation of the ligands. Thus, there is always some speculation about how much effect the simple presence of receptors has.

HIV is today still said to be uncurable, because this specific virus infects T-Helper cells (CD4+ cells) (see e.g. [11], [9]). Thus in the process, the whole adaptive immune system is suspended step by step, causing an easy entrance for intruders. As soon as this destruction of the immune system by HIV becomes too strong, the 'disease' the patients have is called AIDS.

3 | Statistical Methods

3.1 General Approach

The data given are not at all what a statistician wishes to have. Very few subjects took part and many measurements were taken. For usual, well-tried and commonly-used statistical tests there is almost no chance to find out significant and clear results. In this thesis, a new approach is taken to deal with that large amount of variables.
In a previous project, that dealt with the same data sets, a principle component analysis (pca) was carried out to distinguish the Rhesus and African Green monkeys. We now picked up that idea, and additionally implemented a factor analysis, bringing potential subgroups to light.

The most important issue is to identify similarities and differences of these two methods. In the literature, most authors referring to the 'factor analysis' do not distinguish between the exploratory (efa) and confirmatory (cfa) factor analysis. However, what they are describing is usually the efa, which historically was called factor analysis or common factor analysis before its new confirmatory form was introduced within the last 40 years [1]. The term factor analysis itself leads to a huge discussion among authors. Whereas some strictly differentiate between the pca and fa, others call the pca to be a special of fa or vice versa. Additionally, the expression 'latent factor model' is either used equivalent to 'factor analysis' or again viewed as a separated field of interest [44]. Indeed, there are arguments for either opinion and in this work we neither support nor refuse any of them, but point out similarities and differences of both and their value for our data analysis.

Both the principle component analysis and factor analysis aim at reducing the dimension of the data given by linearly combining variables, sometimes called surface attributes [43]. This combination leads to either principal components explaining the data, called latent factors. The main difference lies in the interpretation and their way of explaining the data. The pca tries to explain all the variability within the data, i.e. the extracted factors or better called components completely explain variability [45]. In contrast, fa does not claim such a property. Indeed, in fa we assume common factors as well as specific (or unique) factors [1], leaving some variability unexplained or up to the single variable. The main aim of fa is to represent the covariance structure of the data, not to explain variability. Additionally, the pca orders the extracted components by importance, i.e. the first component usually explains most of the variability, the second component second most and so on. The factors obtained by fa, in contrast, only represent latent properties of the variables and do not reveal any information about either order or importance. Thus, fa is more comparable to clustering, which in fact is done by some authors. Still, huge differences can be seen, so that fa certainly can be claimed to be a field on its own, but still is highly connected to both pca and clustering methods.

Nevertheless, the efa 'is still regarded by many with a marked degree of skepticism' [1]. Its origin is found mainly in the field of social and behavioral research, some statisticians are in trouble with the form of input data. Usually, questionnaires with very subjective responses are used, so that justifying the questions' importance is quite difficult. Moreover, the results received from fa are ambiguous and highly depend on interpretation. As mentioned above, there is no order stated by fa, thus it is the task of the statistician to explain the importance of the outcomes.

Back to our data, another problem arises when dealing with fa. We are given only 35 specimens in total and especially when taking only the African green monkeys into account, these are only eight 'test subjects'. Any prior transformations of the data or preconditions that should be met, usually are very difficult to detect. For example, there is no chance to detect a normal distribution of any of the variables and also homogeneity

among individuals is only achieved by taking monkeys of the same species. Furthermore, when looking at the covariance matrices of the data, those are not even positive definite due to small computational errors. Thus, negative eigenvalues occur for these covariance matrices which is, in theory, impossible. We used the method 'nearPD' implemented in R in order to manipulate eigenvalues and thereby getting the nearest positive definite covariance matrix (see [47] for details on the function). This is a reason why one could argue that any factor analysis being done with the given data will not give any reasonable results.

Indeed, we do not claim to get any true or significant answers. The main focus of the work at hand is to get a new view on possible interactions and to try a new approach towards analyzing these kinds of data. Profound background knowledge in medicine and biology may then take advantage of the outcomes, give reason to correlations and support further investigations, which may lead to more significant statistical results.

As mentioned above, our main goal is to find out differences between the functioning of the immune systems of African Green monkeys and Rhesus monkeys. This is the reason why we are given the measurements of the cells in a natural state on the one hand, and on the other hand, the cell populations being stimulated by phorbol merystate acetate and ionomycin. Therefore, we carry out both principle component analysis and factor analysis for both types and for the data given with and without a stimulation. Moreover, a Mann-Whitney-U-test is conducted in either case in order to find out significant differences between the groups besides any 'latent' factors or components. We speak of a significant difference as soon as the achieved p-value is ≤ 0.5. The combination of all these information then leads to the new suggestions we want to make in that field.

All the computations were run by MATLAB and R. Performing the factor analysis was our main focus, so we start to describe the techniques we used for it first in a whole and in 3.2 more in detail. The function 'factoran' implemented in Matlab, carries out the factor analysis. It might be adjusted to the user's needs in many ways. For our purpose it suffices to use the simplest form, i.e. to use default parameters wherever it was possible. It was

not for the number of factors. This parameter has to be chosen by the user.

Factor analysis usually only returns a matrix of so called factor loadings. This matrix is of the size $n \times r$, where n is the number of variables and r the number of factors as determined ahead. Each variable now 'loads' onto one of the r factors. These loadings are usually numbers between 0 and 1, i.e. if a variable loads onto one factor with 1, this variable highly depends on this factor.

In order to find out groups of variables being dependent on one latent factor, a threshold has to be introduced. This threshold determines the limit down to which factor loadings can be seen as 'being part' of the factor.

Let us illustrate this circumstance with an example. If we had 10 variables and 2 factor and the factor loading matrix would be

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \quad . \tag{3.1}$$

Then we could easily say that the first five variables belong to the first factor and the last five variables clearly are part of the second factor. However, what happens if the factor loading matrix looks like

$$\begin{pmatrix} 0.9 & 0.9 & 0.9 & 0.9 & 0.9 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.9 & 0.9 & 0.9 & 0.9 & 0.9 \end{pmatrix} \quad ?$$

Still, we would make the same conclusion. But what about

$$\begin{pmatrix} 0.4 & 0.6 & 0.7 & 0.8 & 0.5 & 0.6 & 0.7 & 0.4 & 0.2 & 0.1 \\ 0.6 & 0.4 & 0.3 & 0.2 & 0.5 & 0.4 & 0.3 & 0.6 & 0.8 & 0.9 \end{pmatrix} \quad ?$$

Which conclusion should be drawn? This short demonstration might explain the use of our threshold. If the threshold is 0.5 then we cumulate all the variables having factor loadings of 0.5 and higher for a certain factor to one factor group. Thus, it can happen that one variable is determined by two or more latent factors or even none. By raising

the threshold, it becomes clear which variables still ‘stick’ together even if they have to load ‘more’ onto the factors. Having a matrix like (3.1) is utopian, so when the threshold would be set equal to 1, none of the variables will form a factor group. The term ‘factor group’ means a group of variables which mainly load onto one factor, wherein ‘mainly’ is determined by the discussed threshold.

We were able to present the results of the factor analysis and a given threshold in a more convenient way than simple matrices. The pictures created contain both information about the variables contained in the observed data set as well as about the groups of variables which, according to the given limit, load onto the same factor. As mentioned above, by changing this limit, we get a better sense of the magnitude of affiliation among the variables. Having them grouped still does not give us any knowledge about their importance.

This is the reason, why we used the principal component analysis as well as the Mann-Whitney-U-test. Whereas the factor analysis only detects groups of variables, which could be put together to form a score, the principal component analysis sorts the detected components by importance. Likewise, the Mann-Whitney-U-test identifies significant differences between both African green and Rhesus monkeys as well as differences between the stimulated and unstimulated measurements of the cell populations respectively.

Nonetheless, an interpretation made from all these investigations cannot be anything but subjective and may be one-sided. An in-depth analysis of all the methods mentioned as well as the biological explanation for the monkeys on one hand and the effects on humans on the other side would go far beyond the scope of this work and would not contribute to our main goal. Even though few test individuals are given, we still try to find a new approach of how to handle the large data set and with the help of these investigations make suggestions of what could be focused on when conducting further research on that topic. Finding factor groups which show certain patterns or even identify significant setups for either Ag and/or Rh would definitely help to distinguish the two species and

their specific immune responses. In contrast to Rh, SIV-infection is non-pathogenic to Ag, thus a new approach towards the research on HIV vaccines may be discovered. Indeed, this work does claim neither completeness nor precision. It only wants to contribute new ideas to be considered when researching on SIV or, consecutively, HIV.

3.2 Detailed Description of the Factor Analysis

In [1] both the principal component analysis and the factor analysis can be found under the headline 'CA- Independent Component Analysis', which 'seeks to uncover hidden variables in high-dimensional data' [1]. In the following, we will give a close-up view on the factor analysis as used by the function 'factoran' implemented in Matlab. Notations and definitions can be found in [1].

$$\mathbf{X} = \mathbf{AS} + \mathbf{e} \tag{3.2}$$

represents the linear mixing version of the noisy ICA model is given.
Here, $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{r \times m}$ is the matrix of unknown factor loadings, i.e. there are r variables and $m(\ll r)$ factors for n individuals. $\mathbf{S} \in \mathbb{R}^{m \times 1}$ and $\mathbf{e} \in \mathbb{R}^{r \times 1}$ displays the noise in the data.
Usually, $\mathbf{X}$ represents a data matrix with measurements $\mathbf{X}_1, \cdots, \mathbf{X}_n$ for n individuals. For simplicity, we assume that $\mathbf{X}$ is a data vector containing information on only one individual but for all r variables. Thus, we have

$$\mathbf{X} = \mathbf{X_1} = \begin{pmatrix} X_{11} \\ \vdots \\ X_{1r} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{r1} & a_{r2} & \cdots & a_{rm} \end{pmatrix} \cdot \begin{pmatrix} S_{11} \\ \vdots \\ S_{1m} \end{pmatrix} + \begin{pmatrix} e_{11} \\ \vdots \\ e_{1r} \end{pmatrix}$$

There are mainly two different approaches to factor analysis from this point on. The first one is the principal component method. This should not be confused with the principal component analysis, although both techniques rely on a least-squares approach. When using this method, there are no assumptions to be met by the data distribution-wise. In contrast, the maximum-likelihood factor analysis (MLFA) expects the data to follow a

multivariate Gaussian distribution. In our case, as pointed out before, there is no need to prove for normality due to the low number of specimens given. However, as it is done in several papers, one may argue that the assumption of normal distribution can be met in case there would have been more individuals. Moreover, we are not interested in confirming any hypotheses precisely in regard to a specific level of significance, but we want to speculate about coherence among variables. Thus, it is reasonable also for the data at hand to use the MLFA approach.

The first assumption made is that the m latent factors are multivariate normal distributed and are independent from the noise, so that we can state

$$\mathbf{S} \sim \mathcal{N}_m\left(\mathbf{0}, \mathbf{I_m}\right) \tag{3.3}$$

$$\mathbf{e} \sim \mathcal{N}_r\left(\mathbf{0}, \mathbf{\Phi}\right), \tag{3.4}$$

where $\mathbf{\Phi}$ is a diagonal matrix.
Due to (3.2) we know that $\mathbf{X}$ is also normal distributed such that

$$\mathbf{X} \sim \mathcal{N}_r\left(\mathbf{0}, \Sigma_{\mathbf{XX}}\right). \tag{3.5}$$

Proof.

$$\begin{aligned} \mathbb{E}\left(\mathbf{X}\right) &= \mathbb{E}\left(\mathbf{AS} + \mathbf{e}\right) \\ &= \mathbf{A} \cdot \mathbb{E}(\mathbf{S}) + \mathbb{E}(\mathbf{e}) \quad = \quad \mathbf{0} \end{aligned}$$

$$\begin{aligned} \mathrm{Cov}\left(\mathbf{X}\right) &= \mathrm{Cov}\left(\mathbf{AS} + \mathbf{e}\right) \\ &= \mathbf{A} \cdot \mathrm{Cov}(\mathbf{S}) \cdot \mathbf{A}^{\mathbf{T}} + \mathrm{Cov}(\mathbf{e}) \\ &= \mathbf{A}\mathbf{A}^{\mathbf{T}} + \mathbf{\Psi} \\ &= \mathbf{\Sigma_{XX}} \end{aligned} \tag{3.6}$$

□

This is equally true for the case of a data matrix $\mathbf{X}$ with n independent observation $\mathbf{X_1}, \cdots, \mathbf{X_n}$. Then it simply holds that $\mathbf{X_i} \sim \mathcal{N}_r(\mathbf{0}, \mathbf{\Sigma_{XX}})$.
Per definition (e.g. see [1]) it follows

$$\sum_{i=1}^{n} \mathbf{X_i X_i^T} = \mathbf{W} \sim Wishart_{\mathrm{r}}(\mathrm{n}, \mathbf{\Sigma_{XX}}) \quad . \tag{3.7}$$

The Wishart distribution describes the sum of squares of multivariate distributed random variables and thus can be seen as the multivariate equivalent to the Chi-squared distribution. However, the real covariance $\mathbf{\Sigma_{XX}}$ is usually not given, so it is estimated by

$$\begin{aligned} \hat{\mathbf{\Sigma}}_{\mathbf{XX}} &= \frac{1}{n} \cdot \sum_{i=1}^{n} (\mathbf{X_i} - \mathbb{E}(\mathbf{X_i})) \cdot (\mathbf{X_i} - \mathbb{E}(\mathbf{X_i}))^T \\ &= \frac{1}{n} \cdot \sum_{i=1}^{n} (\mathbf{X_i}) \cdot (\mathbf{X_i})^T \\ &= \frac{1}{n} \cdot \mathbf{X^T X} \end{aligned}$$

$$\overset{(3.7)}{\Longrightarrow} \quad n \cdot \hat{\mathbf{\Sigma}}_{\mathbf{XX}} \sim Wishart_{\mathrm{r}}(\mathrm{n}, \mathbf{\Sigma_{XX}}) \tag{3.8}$$

Instead of only dealing with one specimen, i.e. one data vector, we now consider the whole data matrix, such that the model can be expressed by

$$\mathbf{X}_{\mathrm{rxn}} = \mathbf{A}_{\mathrm{rxm}} \cdot \mathbf{S}_{\mathrm{mxn}} + \mathbf{e}_{\mathrm{rxn}}$$

$$\begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{r1} & X_{r2} & \cdots & X_{rn} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{r1} & a_{r2} & \cdots & a_{rm} \end{pmatrix} \times \begin{pmatrix} S_{11} & S_{12} & \cdots & S_{1n} \\ S_{21} & S_{22} & \cdots & S_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ S_{m1} & S_{m2} & \cdots & S_{mn} \end{pmatrix} + \begin{pmatrix} e_{11} & e_{12} & \cdots & e_{1n} \\ e_{21} & e_{22} & \cdots & e_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ e_{r1} & e_{m,2} & \cdots & e_{rn} \end{pmatrix}$$

$$= \mathbf{A} \cdot (\mathbf{s_1}, \cdots, \mathbf{s_n}) + (\mathbf{e_1}, \cdots, \mathbf{e_n}) \tag{3.9}$$

The probability density function for a $Wishart_{\mathrm{p}}(\mathrm{n}, \mathbf{V})$ -distributed random variable is given by

$$L(\mathbf{V}) = \frac{1}{2^{\frac{np}{2}} \cdot |\mathbf{V}|^{\frac{n}{2}} \cdot \mathbf{\Gamma}_p(\frac{n}{2})} \cdot |\mathbf{X}|^{\frac{n-p-1}{2}} \cdot \exp\left(-\frac{1}{2} \cdot \mathrm{tr}\left(\mathbf{V}^{-1}\mathbf{X}\right)\right) \tag{3.10}$$

From (3.8) and (3.10) we can conclude that the likelihood function depending on $\mathbf{\Sigma_{XX}}$, or equivalently, on $\mathbf{A}$ and $\mathbf{\Psi}$ is given by

$$L(\mathbf{A}, \mathbf{\Psi}) = \frac{1}{2^{\frac{nr}{2}} \cdot |\,\mathbf{\Sigma_{XX}}\,|^{\frac{n}{2}} \cdot \mathbf{\Gamma}_r(\frac{n}{2})} \cdot \left| n \cdot \mathbf{\hat{\Sigma}_{XX}} \right|^{\frac{n-r-1}{2}} \cdot \exp\left(-\frac{1}{2} \cdot \mathrm{tr}\,(\,\mathbf{\Sigma_{XX}^{-1}} \cdot n \cdot \mathbf{\hat{\Sigma}_{XX}}\,)\right) \tag{3.11}$$

The log-likelihood function is obtained by taking the logarithm of (3.11)

$$\begin{aligned} \log\left(L(\mathbf{A}, \mathbf{\Psi})\right) = & \frac{(n-r-1)}{2} \cdot \log|n \cdot \mathbf{\hat{\Sigma}_{XX}}| \\ & - \left(\frac{nr}{2} \cdot \log(2) + \frac{n}{2} \cdot \log|\mathbf{\Sigma_{XX}}| + \log\left(\mathbf{\Gamma}_r\left(\frac{n}{2}\right)\right)\right) \\ & - \frac{n}{2} \cdot \mathrm{tr}\left(\mathbf{\Sigma_{XX}^{-1}} \cdot \mathbf{\hat{\Sigma}_{XX}}\right) \end{aligned} \tag{3.12}$$

and can be reduced to

$$\begin{aligned} \log\left(L(\mathbf{A}, \mathbf{\Psi})\right) = & -\frac{n}{2} \cdot \log|\mathbf{\Sigma_{XX}}| - \frac{n}{2} \cdot \mathrm{tr}\left(\mathbf{\Sigma_{XX}^{-1}} \cdot \mathbf{\hat{\Sigma}_{XX}}\right) \\ = & -\frac{n}{2} \cdot \log|\mathbf{\Sigma_{XX}}| - \frac{n}{2} \cdot \mathrm{tr}\left(\mathbf{\hat{\Sigma}_{XX}} \cdot \mathbf{\Sigma_{XX}^{-1}}\right) \\ = & -\frac{n}{2} \cdot \log|\mathbf{A}\mathbf{A}^{\mathbf{T}} + \mathbf{\Psi}| - \frac{n}{2} \cdot \mathrm{tr}\left(\mathbf{\hat{\Sigma}_{XX}} \cdot (\mathbf{A}\mathbf{A}^{\mathbf{T}} + \mathbf{\Psi})^{-1}\right) \end{aligned} \tag{3.13}$$

when only taking into account the values influencing $\mathbf{A}$ and $\mathbf{\Psi}$. The task now is to maximize the log-likelihood as given in (3.13). This can be done by applying the EM-algorithm which comes from the area of clustering.

3.2.1 *EM-Algorithm*

Using the EM-Algorithm, it is assumed that the given data $\mathbf{X}$ have a certain distribution $p(\cdot, \Theta)$ with the notation

$$X \quad \sim \quad p(\cdot, \Theta) \quad . \tag{3.14}$$

In our case this distribution will be the normal distribution with $\Theta = (\mathbf{A}, \mathbf{\Psi})$, because of (3.5) and (3.6). The main idea of the EM-algorithm is to divide the given data $\mathbf{X}$ into an observed and a missing part. In our case, the row vectors of $\mathbf{S}$, the $\{\mathbf{s_i}\}$ will play the 'missing' role.
First the so called complete data-likelihood is defined by

$$\begin{aligned} \mathcal{L}(\Theta \mid \mathbf{X}) \quad &= \quad p(\mathbf{X} \mid \Theta) \\ &\left(= p(\mathbf{X} \mid \mathbf{A}, \mathbf{\Psi}) \right) \quad . \end{aligned}$$

Second, we define

$$\mathbf{X} \quad = \quad (\mathbf{X_{obs}^T}, \mathbf{X_{mis}^T})^T \quad . \tag{3.15}$$

Now, if the probability that a particular variable (or in our case a latent factor) depends only upon the observed values $\mathbf{X_{obs}}$ and not on the missing ones $\mathbf{X_{mis}}$, then we can integrate out the $\mathbf{X_{mis}}$ and define the observed data-likelihood by

$$\mathcal{L}_{\text{obs}}(\Theta \mid \mathbf{X_{obs}}) \quad = \int p(\mathbf{X_{obs}}, \mathbf{X_{mis}} \mid \Theta) \, \mathrm{d}\, \mathbf{X_{mis}} \quad . \tag{3.16}$$

Maximizing the likelihood results at this point in maximizing the oberserved data-likelihood

(3.16) with respect to Θ, or in our case, with respect to $\mathbf{A}$ and $\mathbf{\Psi}$.
Two steps follow. The expectation step (E) in which the expectation of the complete data-likelihood is computed conditioned on the observed data and the current parameter estimate, and, the maximization step (M) in which the current parameters are updated by maximizing the conditional expectation from the first step. We know that

$$\begin{aligned} p\left(\mathbf{X_{mis}} \mid \mathbf{X_{obs}}, \Theta\right) &= \frac{p\left(\mathbf{X_{obs}}, \mathbf{X_{mis}} \mid \Theta\right)}{p\left(\mathbf{X_{obs}} \mid \Theta\right)} \\ \Rightarrow \quad p\left(\mathbf{X_{obs}} \mid \Theta\right) &= \frac{p\left(\mathbf{X_{obs}}, \mathbf{X_{mis}} \mid \Theta\right)}{p\left(\mathbf{X_{mis}} \mid \mathbf{X_{obs}}, \Theta\right)} \quad . \end{aligned} \tag{3.17}$$

Then the observed data log-likelihood is obtained by

$$\begin{aligned} l\left(\Theta \mid \mathbf{X_{obs}}\right) &= \mathcal{L}\left(\Theta \mid \mathbf{X_{obs}}\right) \\ &\overset{(3.17)}{=} \log\left(p\left(\mathbf{X_{obs}} \mid \Theta\right)\right) \\ &= \log\left(p\left(\mathbf{X_{obs}}, \mathbf{X_{mis}} \mid \Theta\right)\right) - \log\left(p\left(\mathbf{X_{mis}} \mid \mathbf{X_{obs}}, \Theta\right)\right) \\ &= l\left(\Theta, \mathbf{X}\right) - \log\left(p\left(\mathbf{X_{mis}} \mid \mathbf{X_{obs}}, \Theta\right)\right) \quad . \end{aligned} \tag{3.18}$$

It is $l(\Theta, \mathbf{X})$ the complete data log-lilkelihood and $\log\left(p\left(\mathbf{X_{mis}} \mid \mathbf{X_{obs}}, \Theta\right)\right)$ the part of the complete data log-likelihood due to the missing data [1].
Next we define the following expectations conditioned on $p\left(\mathbf{X_{mis}} \mid \mathbf{X_{obs}}, \Theta'\right)$ where Θ' is a current value of Θ :

$$\begin{aligned} Q\left(\Theta \mid \Theta'\right) &= \int l\left(\Theta \mid \mathbf{X}\right) \cdot p\left(\mathbf{X_{mis}} \mid \mathbf{X_{obs}}, \Theta'\right) \mathrm{d} : \mathbf{X_{mis}} \\ &= \mathbb{E}\left(l\left(\Theta \mid \mathbf{X}\right) \mid \mathbf{X_{obs}}, \Theta'\right) \end{aligned} \tag{3.19}$$

$$H(\Theta \mid \Theta') = \int \log\left(p\left(\mathbf{X}_{\mathbf{mis}} \mid \mathbf{X}_{\mathbf{obs}}, \Theta\right)\right) \cdot p(\mathbf{X}_{\mathbf{mis}} \mid \mathbf{X}_{\mathbf{obs}}, \Theta')\, \mathrm{d}\, \mathbf{X}_{\mathbf{mis}}$$
$$= \mathbb{E}\left(\log\left(p\left(\mathbf{X}_{\mathbf{mis}} \mid \mathbf{X}_{\mathbf{obs}}, \Theta\right)\right) \mid \mathbf{X}_{\mathbf{obs}}, \Theta'\right) \quad . \tag{3.20}$$

It yields that

$$l(\Theta \mid \mathbf{X}_{\mathbf{obs}}) = Q(\Theta \mid \Theta') - H(\Theta \mid \Theta') \quad . \tag{3.21}$$

As we are about to maximize the log-likelihood, (3.21) should increase with each step of the iteration.
Let

$$h(\mathbf{X}_{\mathbf{mis}}) = \frac{p\left(\mathbf{X}_{\mathbf{mis}} \mid \mathbf{X}_{\mathbf{obs}}, \Theta\right)}{p\left(\mathbf{X}_{\mathbf{mis}} \mid \mathbf{X}_{\mathbf{obs}}, \Theta'\right)} \tag{3.22}$$

and we observe

$$\begin{aligned}
H(\Theta \mid \Theta') - H(\Theta' \mid \Theta') &\overset{(3.20)}{=} \mathbb{E}\left(\log\left(p\left(\mathbf{X}_{\mathbf{mis}} \mid \mathbf{X}_{\mathbf{obs}}, \Theta\right)\right) \mid \mathbf{X}_{\mathbf{obs}}, \Theta'\right) \\
&\quad - \mathbb{E}\left(\log\left(p\left(\mathbf{X}_{\mathbf{mis}} \mid \mathbf{X}_{\mathbf{obs}}, \Theta'\right)\right) \mid \mathbf{X}_{\mathbf{obs}}, \Theta'\right) \\
&\overset{(3.22)}{=} \mathbb{E}\left(\frac{p\left(\mathbf{X}_{\mathbf{mis}} \mid \mathbf{X}_{\mathbf{obs}}, \Theta\right)}{p\left(\mathbf{X}_{\mathbf{mis}} \mid \mathbf{X}_{\mathbf{obs}}, \Theta'\right)} \,\middle|\, \mathbf{X}_{\mathbf{obs}}, \Theta'\right) \\
&= \mathbb{E}\left(\log\left(h\left(\mathbf{X}_{\mathbf{mis}}\right)\right) \,\middle|\, \mathbf{X}_{\mathbf{obs}}, \Theta'\right) \\
&\overset{\log x \le x-1}{\le} \mathbb{E}\left(h\left(\mathbf{X}_{\mathbf{mis}}\right) \mid \mathbf{X}_{\mathbf{obs}}, \Theta'\right) - 1 \\
&\overset{(3.22)}{=} 0 \\
\Rightarrow \quad H(\Theta \mid \Theta') &\le H(\Theta' \mid \Theta')
\end{aligned} \tag{3.23}$$

Thus, we can show

$$
\begin{aligned}
& l\left(\Theta^{m+1} \mid \mathbf{X_{obs}}\right) - l\left(\Theta^{m} \mid \mathbf{X_{obs}}\right) \\
& \qquad \overset{(3.21)}{=} Q\left(\Theta^{m+1} \mid \Theta^{m}\right) - H\left(\Theta^{m+1} \mid \Theta^{m}\right) - \left(Q\left(\Theta^{m} \mid \Theta^{m}\right) - H\left(\Theta^{m} \mid \Theta^{m}\right)\right) \\
& \qquad = Q\left(\Theta^{m+1} \mid \Theta^{m}\right) - Q\left(\Theta^{m} \mid \Theta^{m}\right) + \underbrace{\left(H\left(\Theta^{m} \mid \Theta^{m}\right) - H\left(\Theta^{m+1} \mid \Theta^{m}\right)\right)}_{\overset{(3.23)}{\geq} 0} \\
& \qquad \geq 0
\end{aligned}
$$

where the last inequality derives from the assumption that the new parameter Θ^{m+1} is found by the EM-algorithm in order to get $Q(\Theta^{m+1} \mid \Theta^{m}) > Q(\Theta^{m} \mid \Theta^{m})$. This shows that at each iteration, the observed log-likelihood function increases and it can even be shown that there exisst a convergence at least to a local maximum [1]. However, there are apparently some disadvantages coming with this method, for example a slow convergence rate in case of a large missing part [1].

3.2.2 *EM-Algorithm used in Factor Analysis*

As mentioned before, the $\{\mathbf{s_i}\}$ are supposed to be multivariate $\mathcal{N}_m(\mathbf{0}, \mathbf{I_m})$ -distributed and independent from the $\{\mathbf{e_i} = \mathbf{X_i} - \mathbf{A}\mathbf{s_i}\} \sim \mathcal{N}_r(\mathbf{0}, \Psi)$. If the $\{\mathbf{s_i}\}$ were observed then the complete data likelihood would be given by their joint distribution

$$
\begin{aligned}
CompLik = & \prod_{i=1}^{n} \frac{1}{(2\pi)^{\frac{r}{2}} \cdot |\mathbf{\Psi}|^{\frac{1}{2}}} \cdot \exp\left(-\frac{1}{2} \cdot (\mathbf{e_i} - \mathbf{0})^T \mathbf{\Psi}^{-1} (\mathbf{e_i} - \mathbf{0})\right) \\
& \cdot \frac{1}{(2\pi)^{\frac{m}{2}} \cdot |\mathbf{I_m}|^{\frac{1}{2}}} \cdot \exp\left(-\frac{1}{2} \cdot (\mathbf{s_i} - \mathbf{0})^T \mathbf{I_m}^{-1} (\mathbf{s_i} - \mathbf{0})\right) \\
= & \left((2\pi)^r \cdot \mathbf{\Psi}\right)^{-\frac{n}{2}} \cdot \exp\left(-\frac{1}{2} \cdot \left(\left(\mathbf{X_i} - \mathbf{A}\mathbf{s_i}\right) \mathbf{\Psi}^{-1} \left(\mathbf{X_i} - \mathbf{A}\mathbf{s_i}\right)\right)\right)
\end{aligned}
$$

$$\cdot ((2\pi)^m)^{-\frac{1}{2}} \cdot \exp\left(-\frac{1}{2} \sum_{i=1}^{n} \mathbf{s_i}^T \mathbf{s_i} \right) \quad . \tag{3.24}$$

The complete data log-likelihood is then given by

$$\begin{aligned} log(Complik) = \quad & -\frac{n}{2} \left(\log\left((2\pi)^r\right) + \log |\, \mathbf{\Psi} \,| \right) - \frac{1}{2} \left(\sum_{i=1}^{n} (\, \mathbf{X_i} - \mathbf{A s_i} \,)\, \mathbf{\Psi}^{-1}\, (\, \mathbf{X_i} - \mathbf{A s_i} \,) \right) \\ & - \frac{n}{2} \left(\log \left(2\pi \right)^m \right) - \frac{1}{2} \sum_{i=1}^{n} \mathbf{s_i}^T \mathbf{s_i} \quad . \end{aligned} \tag{3.25}$$

According to the *EM-Algorithm* described in 3.2.1, we have to maximize $\mathbb{E}(log(Complik))$. We already know that $\{\mathbf{s_i}\} \sim \mathcal{N}_m(\mathbf{0}, \mathbf{I_m})$ and $\{\mathbf{X_i}\} \sim \mathcal{N}_r(\mathbf{0}, \mathbf{\Sigma_{XX}})$ with $\mathbf{\Sigma_{XX}} = \mathbf{A A^T} + \mathbf{\Psi}$ (see (3.6)). Then the distribution of the latent factors $\{\mathbf{s_i}\}$ conditioned on the observed data $\{\mathbf{X_i}\}$ as well as $\mathbf{A}$ and $\mathbf{\Psi}$ is again normal with parameter μ^*_{rx1} and Σ^*_{rxr} (see [1], *chapter 3*) where

$$\begin{aligned} \mu^* &= \mu_{\mathbf{s_i}} + \Sigma_{\mathbf{s_i X_i}} \Sigma^{-1}_{\mathbf{X_i X_i}} \cdot (\mathbf{X_i} - \mu_{\mathbf{X_i}}) \\ \Sigma^* &= \Sigma_{\mathbf{s_i s_i}} - \Sigma_{\mathbf{s_i X_i}} \Sigma^{-1}_{\mathbf{X_i X_i}} \Sigma_{\mathbf{X_i s_i}} \quad , \end{aligned}$$

and

$$\begin{aligned} \mu_{\mathbf{s_i}} &= \mathbf{0} \\ \mu_{\mathbf{X_i}} &= \mathbf{0} \\ \Sigma_{\mathbf{s_i s_i}} &= \mathbf{I_r} \\ \Sigma_{\mathbf{s_i X_i}} &= \mathbf{A} \\ \Sigma_{\mathbf{X_i X_i}} &= \mathbf{A A^T} + \mathbf{\Psi} \end{aligned}$$

It thus simply holds

$$(\mathbf{s_i} \mid \mathbf{X_i}, \mathbf{A}, \mathbf{\Psi}) \sim \mathcal{N}_r\, (\delta \mathbf{X}_i, \mathbf{\Lambda}). \tag{3.26}$$

with

$$\delta = \mathbf{A}^T(\mathbf{A}\mathbf{A}^T + \mathbf{\Psi})^{-1} \tag{3.27}$$

$$\mathbf{\Lambda} = \mathbf{I}_r - \mathbf{A}^T(\mathbf{A}\mathbf{A}^T + \Psi)^{-1}\mathbf{A}. \tag{3.28}$$

Here, $\mathbf{A}$ and $\mathbf{\Psi}$ will be the values which have to be updated in the M-step of the EM-algorithm later. In order to update those by maximizing $\mathbb{E}(log(Complik))$, the following statistics shall suffice:

$$C_{XX} = \sum_{i=1}^{n} \mathbf{X_i}\mathbf{X_i}^T \tag{3.29}$$

$$C_{XS} = \sum_{i=1}^{n} \mathbf{X_i}\mathbf{s_i}^T \tag{3.30}$$

$$C_{SS} = \sum_{i=1}^{n} \mathbf{s_i}\mathbf{s_i}^T \tag{3.31}$$

Additionally we notice the following for the expected values conditioned on $\mathbf{A}, \mathbf{\Psi}$ and the observed data $\{\mathbf{X_i}\}$

$$\begin{aligned}
C^*_{XX} = \quad & \mathbb{E}\left(C_{XX} \mid \{\mathbf{X_i}\}, \mathbf{A}, \mathbf{\Psi}\right) = \mathbb{E}\left(\sum_{i=1}^{n} \mathbf{X_i}\mathbf{X_i}^T \mid \{\mathbf{X_i}\}, \mathbf{A}, \mathbf{\Psi}\right) \\
= \quad & C_{XX} \\
C^*_{XS} = \quad & \mathbb{E}\left(C_{XS} \mid \{\mathbf{X_i}\}, \mathbf{A}, \mathbf{\Psi}\right) = \mathbb{E}\left(\sum_{i=1}^{n} \mathbf{X_i}\mathbf{s_i}^T \mid \{\mathbf{X_i}\}, \mathbf{A}, \mathbf{\Psi}\right) \\
= \quad & \sum_{i=1}^{n} \mathbf{X_i}\, \mathbb{E}\left(\mathbf{s_i} | \{\mathbf{X_i}\}, \mathbf{A}, \mathbf{\Psi}\right)^T \\
= \quad & \sum_{i=1}^{n} \mathbf{X_i}\, (\delta \mathbf{X_i})^T \\
= \quad & C_{XX}\delta^T \\
C^*_{SS} = \quad & \mathbb{E}\left(C_{SS} \mid \{\mathbf{X_i}\}, \mathbf{A}, \mathbf{\Psi}\right) = \mathbb{E}\left(\sum_{i=1}^{n} \mathbf{s_i}\mathbf{s_i}^T\right)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} \mathbb{E}\left(\mathbf{s_i}\mathbf{s_i}^T \mid \{\mathbf{X_i}\}, \mathbf{A}, \mathbf{\Psi}\right) \\
&= \sum_{i=1}^{n} \left(\mathrm{Var}\left(\mathbf{s_i}\mathbf{s_i}^T \mid \{\mathbf{X_i}\}, \mathbf{A}, \mathbf{\Psi}\right) + \mathbb{E}\left(\mathbf{s_i} \mid \{\mathbf{X_i}\}, \mathbf{A}, \mathbf{\Psi}\right) \cdot \mathbb{E}\left(\mathbf{s_i} \mid \{\mathbf{X_i}\}, \mathbf{A}, \mathbf{\Psi}\right)^T\right) \\
&= \sum_{i=1}^{n} \left(\mathbf{\Lambda} + (\delta \mathbf{X}_i) \cdot (\delta \mathbf{X}_i)^T\right) \\
&= n \cdot \mathbf{\Lambda} + \delta \cdot \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^T \delta^T \\
&= n \cdot \mathbf{\Lambda} + \delta C_{XX} \delta^T .
\end{aligned}
$$

The above definitions are used in the E-step to obtain the conditioned expectation of the complete data likelihood. Let c be a constant to simplify the following computations.

$$
\begin{aligned}
\mathbb{E}\left(log(Complike)\right) &\overset{(3.24)}{=} c - \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}\left(\left(\mathbf{X_i} - \mathbf{A}\mathbf{s_i}\right)^T \mathbf{\Psi}^{-1} \left(\mathbf{X_i} - \mathbf{A}\mathbf{s_i}\right) \Big| \mathbf{X_i} \right) \\
&\quad - \frac{n}{2} \cdot \log\left(|\mathbf{\Psi}^{-1}| \right) - \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}\left(\mathbf{s_i}\mathbf{s_i}^T \mid \mathbf{X_i}\right) \\
&= c - \frac{1}{2} \sum_{i=1}^{n} \mathbf{X_i}^T \mathbf{\Psi}^{-1} \mathbf{X_i} - \frac{1}{2} \sum_{i=1}^{n} \mathbf{A}^T \mathbf{\Psi}^{-1} \mathbb{E}\left(\mathbf{s_i}\mathbf{s_i}^T \mid \mathbf{X_i}\right) \mathbf{A} \\
&\quad + \sum_{i=1}^{n} \mathbf{X_i}^T \mathbf{\Psi}^{-1} \mathbf{A}\, \mathbb{E}\left(\mathbf{s_i} \mid \mathbf{X_i}\right) \\
&= \rho
\end{aligned}
$$

Setting the derivate of ρ with respect to both $\mathbf{A}$ and $\mathbf{\Psi}$ equal to zero and solving for $\mathbf{A}$ and $\mathbf{\Psi}$ respectively will deliver their estimators.
Starting with $\mathbf{A}$ it holds

$$\frac{\delta \rho}{\delta \mathbf{A}} = -\sum_{i=1}^{n} \mathbf{\Psi}^{-1}\mathbf{A}\,\mathbb{E}\left(\mathbf{s_i}\mathbf{s_i}^T \mid \mathbf{X_i}\right) + \sum_{i=1}^{n} \mathbf{\Psi}^{-1}\mathbf{X_i}\,\mathbb{E}\left(\mathbf{s_i} \mid \mathbf{X_i}\right) = 0$$

$$\Rightarrow \sum_{i=1}^{n} \mathbf{\Psi}^{-1}\mathbf{A}\,\mathbb{E}\left(\mathbf{s_i}\mathbf{s_i}^T \mid \mathbf{X_i}\right) = \sum_{i=1}^{n} \mathbf{\Psi}^{-1}\mathbf{X_i}\,\mathbb{E}\left(\mathbf{s_i} \mid \mathbf{X_i}\right)$$

$$\Rightarrow \hat{A} = \sum_{i=1}^{n} \mathbf{X_i}\,\mathbb{E}\left(\mathbf{s_i} \mid \mathbf{X_i}\right) \cdot \left(\sum_{i=1}^{n} \mathbf{A}\,\mathbb{E}\left(\mathbf{s_i}\mathbf{s_i}^T \mid \mathbf{X_i}\right)\right)^{-1}$$

$$= C^*_{XS} \cdot C^{*\,-1}_{SS} \quad . \tag{3.32}$$

Equally, for $\mathbf{\Psi}$ we get

$$\frac{\delta \rho}{\delta \mathbf{\Psi}} = \frac{n}{2} \cdot \mathbf{\Psi} - \frac{1}{2}\sum_{i=1}^{n} \mathbf{X_i}\mathbf{X_i}^T + \sum_{i=1}^{n} \mathbf{A}\,\mathbb{E}\left(\mathbf{s_i} \mid \mathbf{X_i}\right)\mathbf{X_i}^T$$

$$- \frac{1}{2}\sum_{i=1}^{n} \mathbf{A}\,\mathbb{E}\left(\mathbf{s_i}\mathbf{s_i}^T \mid \mathbf{X_i}\right)\mathbf{A}^T = 0$$

$$\Rightarrow \frac{n}{2} \cdot \mathbf{\Psi} = \frac{1}{2}\sum_{i=1}^{n} \mathbf{X_i}\mathbf{X_i}^T - \sum_{i=1}^{n} \mathbf{A}\,\mathbb{E}\left(\mathbf{s_i} \mid \mathbf{X_i}\right)\mathbf{X_i}^T$$

$$+ \frac{1}{2} \cdot \mathbf{A} \sum_{i=1}^{n} \mathbb{E}\left(\mathbf{s_i}\mathbf{s_i}^T \mid \mathbf{X_i}\right)\mathbf{A}^T$$

$$\overset{(3.32)}{=} \frac{1}{2}\sum_{i=1}^{n} \mathbf{X_i}\mathbf{X_i}^T - \mathbf{A}\,(C^*_{XS})^T$$

$$+ \frac{1}{2} \cdot C^*_{XS}\, C^{*\,1}_{SS} \cdot C^*_{SS} \cdot (C^*_{XS}\, C^{*\,-1}_{SS})^T$$

$$\Rightarrow n \cdot \mathbf{\Psi} = \sum_{i=1}^{n} \mathbf{X_i}\mathbf{X_i}^T - 2\,\mathbf{A}\,(C^*_{XS})^T + C^*_{XS} \cdot C^{*\,-1}_{SS}(C^*_{XS})^T$$

$$\Rightarrow \quad n \cdot \mathbf{\Psi} = \sum_{i=1}^{n} \mathbf{X_i} \mathbf{X_i}^T - 2\,\mathbf{A}\,(C^*_{XS})^T + \mathbf{A} \cdot (C^*_{XS})^T$$
$$= C^*_{XX} - \mathbf{A}\,(C^*_{XS})^T$$

$$\Rightarrow \quad \hat{\mathbf{\Psi}} = diag\,\{\, C^*_{XX} - \mathbf{A}\,(C^*_{XS})^T \,\} \quad . \tag{3.33}$$

As we assume $\mathbf{\Psi}$ to be a diagonal matrix, we use the diagonal constraint in the last step. In the M-step we then use the following regression estimates,

$$\hat{\mathbf{A}} = C^*_{XS} C^{*-1}_{SS}$$

$$\hat{\mathbf{\Psi}} = diag\{C^*_{XX} - C^*_{XS} C^{*-1}_{SS} C^{*T}_{XS}\} \quad .$$

Summing up the $EM - Algorithm$ used in maximum-likelihood factor analysis originates the following algorithm (adapted from [1]).

EM-Algorithm

Step 1 Take initial guesses $\hat{\mathbf{A}}_0$ and $\hat{\mathbf{\Psi}}_0$ for the parameters $\hat{\mathbf{A}}$ and $\hat{\mathbf{\Psi}}$ respectively.

Step 2 EM-Algorithm

Step 2.1 E-step Compute

$$C_{XX} = \sum_{i=1}^{n} \mathbf{X_i} \mathbf{X_i}^T$$

$$C^{(\mathrm{k}-1)}_{XS} = C_{XX} \cdot \delta^T_{k-1}$$
$$C^{(\mathrm{k}-1)}_{SS} = \delta_{(\mathrm{k}-1)} \cdot C_{XX} \cdot \delta^T_{(\mathrm{k}-1)} + n \cdot \Lambda_{(\mathrm{k}-1)}$$

$$\text{with} \quad \delta_{(k-1)} = \hat{\mathbf{A}}^T_{(k-1)} \left(\hat{\mathbf{A}}_{(k-1)} \hat{\mathbf{A}}^T_{(k-1)} + \hat{\mathbf{\Psi}}^T_{(k-1)}\right)^{-1}$$
$$\Lambda_{(k-1)} = \mathbf{I}_r - \delta_{(k-1)} \hat{\mathbf{A}}_{(k-1)}$$

Step 2.2 M-step Update the estimators,

$$\hat{\mathbf{A}}^T_{(k)} \longleftarrow C_{XS}^{(k-1)} \left(C_{SS}^{(k-1)}\right)^{-1}$$
$$\hat{\mathbf{\Psi}}^T_{(k)} \longleftarrow diag\left\{C_{XX} - C_{XS}^{(k-1)} \left(C_{SS}^{(k-1)}\right)^{-1} \left(C_{XS}^{(k-1)}\right)^T\right\}$$

Step 3 Stop when convergence has been attained.

4 | Results

4.1 Description of the Data

The data sets given contain information on the cell popluations of 35 monkeys, 8 being African Green monkeys and 27 Rhesus monkeys. The original data included computed variables which obviously have to correlate with other ones, so these were deleted prior to further investigations. A detailed list can be found in appendix 5.

The data sets given are determined by either focusing on CD8 positive cells or CD4 positive cells. Both these types of cells occur as subgroups of T cells, which are on their part CD3 positive. In this study, researchers looked at all those cells being positive for either CD4 or CD8 and, then, took the mean fluorescence intensity of CD28, CD45RA, CD3, CD95 and CD4 or CD8. For the CD4+ and CD8+ cells similar measurements were taken, so for simplicity and lack of space we only concentrated on CD4+ cells for the following description.

Three files are given comprising different information on the individuals. First of all, each individual appears twice, because all examinations were done with cells being in a usual state of health and another time, the cell populations were investigated after being stimulated by phorbol merystate acetate and ionomycin. As we already know that African Green monkeys are resistant to the onset of the disease caused by SIV infection, whereas Rhesus monkeys develop AIDS-like illness upon SIV infection, the happenings during the stimulation compared to the basic case are of interest.

Hence, it might be possible to discover fundamental differences between their im-

mune response. Consequently, these findings may bring us a step closer to understand the reaction of the immune system of Ag upon SIV infection, thus revealing hints on potential vaccination possibilities.

In all datasets, both unstimulated and stimulated cell populations of Ag and Rh were investigated. However, in the following, when it is referred to measurements of or simply referred to Ag and Rh, we actually mean the investigations made upon the appropriate cell population.

In each file, there are at least three of the following four main groups. First, there are all CD4+ cells taken together, second, only the naïve CD4+ cells are observed, third, the CD4+ cells belonging to the effective memory (EM) are taken into account and fourth, the CD4+ cells from the central memory (CM) form one group. Within each of these groups, the mean fluorescence intensity (MFI) of certain surface markers and the proportion of the cytokines gamma IFN (IFN), TNF-alpha (TNF) and Interleukin 2 (IL2) is measured, as well as the proportion of these groups themselves and the proportion of granulocytes, lymphocytes and monocytes in the whole sample. What kind of surface markers and specifications of the CD4+ cells examined depends on the data set and, of course, varies for each of them. Thus, many different variables accumulate when taking into account all the data given per individual. Bearing in mind the low number of individuals, this might not be the best point of departure for any statistical analysis.

4.2 Evaluation of the Datasets

Evaluating the given data was the aim of this work. However, it is also the most difficult one. As stated many times before, there is not one solution and not the one answer we are looking for. Instead, we will describe some of the observations made using the statistical methods described before. Finding distinct patterns for Ag and Rh is a quite ambitious goal. However, the combination of multiple statistical methods might give new directions for the research on HIV. In the following, all computations and figures are done with either the software MATLAB or R. We will go through each dataset separately first and finally,

we connect all the findings.

Explanation of Plots

Boxplots Giving an overview of the range and location of the parameter values, boxplots are especially used to get a first impression of the data.

Gscatter-plot : As parameters, the scores of a previous performed principal component analysis (pca) are passed. This is done so that the coordinates of the data, after having applied the pca, are displayed in the new coordinate system. We use gscatter-plots to compare the cell populations of Ag and Rh respectively being stimulated and unstimulated, and to compare the stimulated and unstimulated cell populations of Rh and Ag respectively.

Biplot After having performed a pca, both coefficients and scores are illustrated in one coordinate system in a biplot. The direction and length of each vector represents how each variable contributed to the principal components. Biplots occur with or without labels for single variables. Having variables labeled is only helpful considering the most outer, i.e. the most influential variables, due to the non-legibility of the most inner ones. Biplots can tell us which variables seem to be most affecting to the principal components.

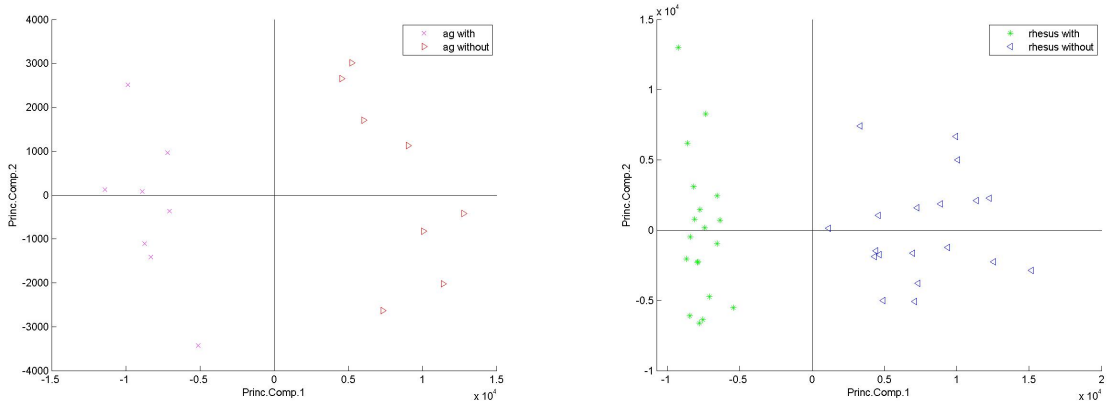
Image of grouped variables All variables available for a dataset are represented in one image ordered content-wise. Colors are expressing groups generated by a previous performed factor analysis. For certain analyses, the third 'group' represents signficant differences between the stimulated and unstimulated cell populations of the observed species.

CD4

This first dataset contains information about the surface markers expressed by CD4+ cells. A whole list and explanation of the variables in the dataset '$RM\_AGM\_PMA + I\_CD4$'

can be found in appendix (5). In figure 4.1 you can see all these variables (except for the 'Animal ID' and 'stim').

The principal component analysis (pca) carried out for both Africa Green (Ag) and Rhesus monkeys (Rh) (or indeed, rather for the measurements taken from their cell populations) as well as for the cell populations being stimulated (stim) or unstimulated (unstim) showed a clear separation for the according groups ((4.2),(4.3)). For the Rh, in the gscatter-plot, the data points for the stimulated population are found on the left hand side of the y-axis, while those without stimulation concentrate on the right side. These two plotted principal components explain 90% of the variance and from 4.2 we can say that especially the first component contributes to that distinction (89% of the variance is explained by the first component). For the Ag even 93% of the variance are explained by the first principal component which can be seen in 4.2.



(a) cells of Ag

(b) cells of Rh

Figure 4.2: Gscatter-plots for cell populations of Ag and Rh in the **CD4**-dataset

However, it might be obvious that a distinction between stimulated and unstimulated cell populations can be made. The distinction between the cell populations of Ag and Rh is more important. We need at least two principal components to explain 87% (unstim) or 85% (stim) and at least three principal components to explain more than 90% of the variance within the data. Still, even the gscatter-plot of the first two components displays a difference between the two species in both the stimulated and unstimulated case (4.3).

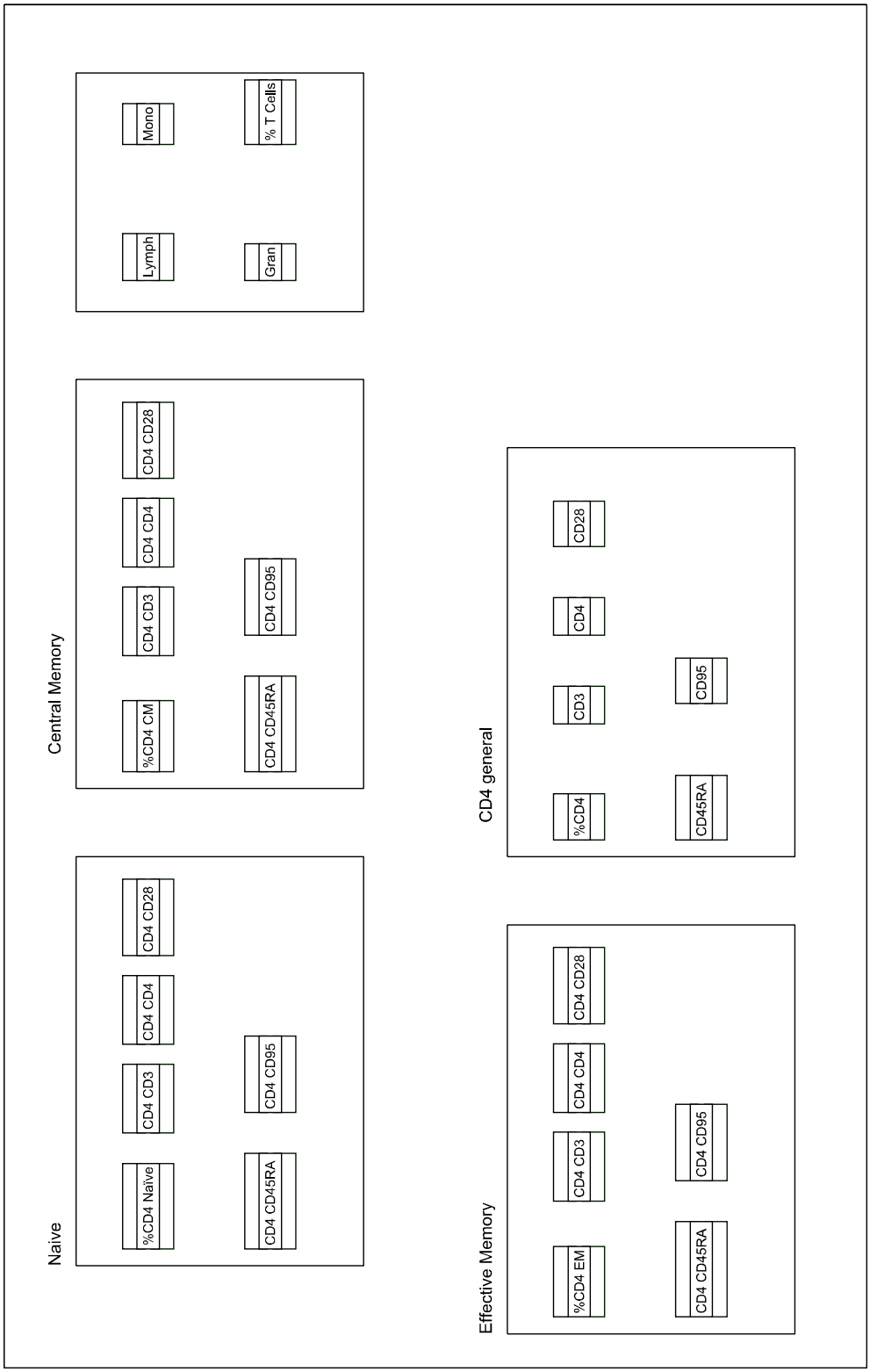
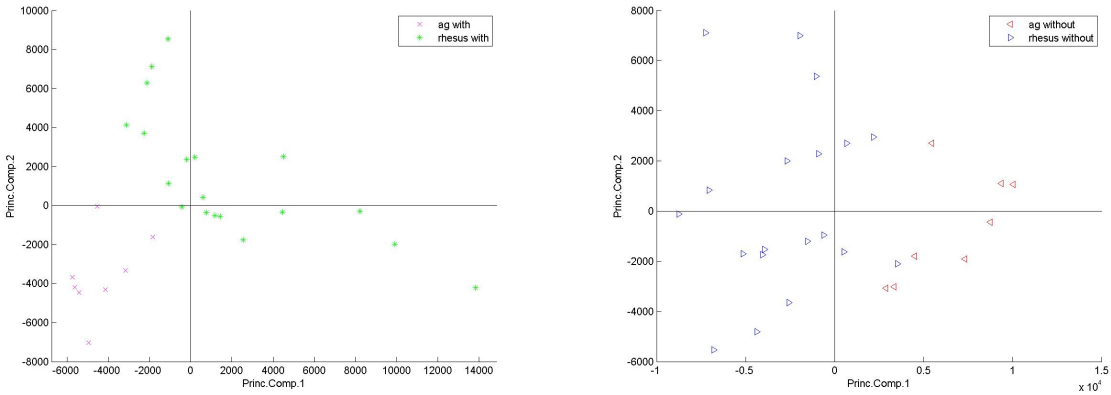


Figure 4.1: Variables in **CD4**-dataset



(a) stimulated cells

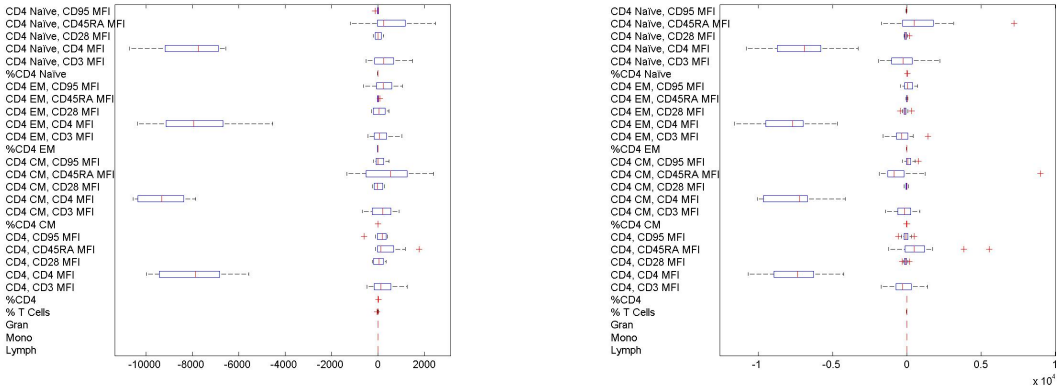
(b) unstimulated cells

Figure 4.3: Gscatter-plots for the stimulated and unstimulated cell populations for the **CD4**-dataset

Thus, we can assume that there are differences between the expressions of surface markers measured for the cell populations of Ag and Rh whether the specimens were stimulated or not. Processing a Mann-Whitney-U-test, we indeed obtained multiple variables being significant different between Ag (stim/ unstim) and Rh (stim/unstim) respectively. Additionally, we performed a factor analysis (fa) to attempt to enclose differing groupings. As described in chapter 3.2, fa was done with increasing thresholds.

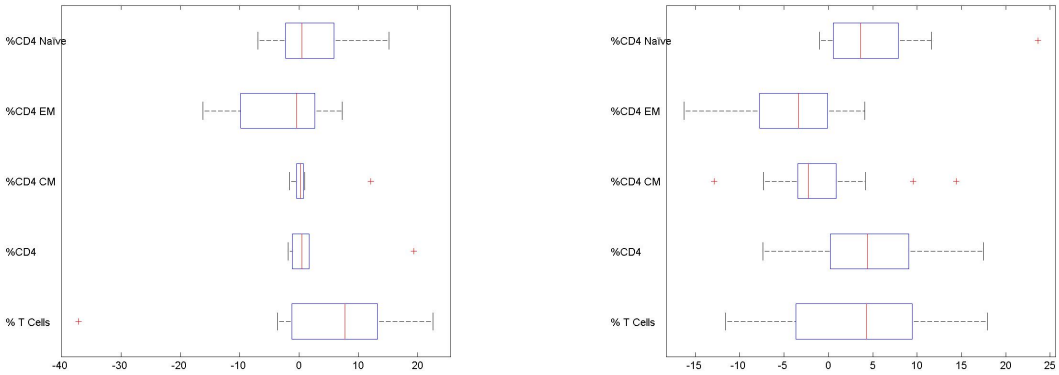
Taking a look at the unstimulated case first, the grouping for Ag and Rh is strikingly similar. One group consists of (see(1)(2))
‘CD4, CD3 MFI’, ‘CD4 Naïve, CD3 MFI’, ‘CD4 EM, CD3 MFI’, ‘CD4, CD28 MFI’, ‘CD4 Naïve, CD28 MFI’, ‘CD4 CM, CD28 MFI’, ‘CD4 EM, CD28 MFI’, ‘CD4, CD4 MFI’, ‘CD4 Naïve, CD4 MFI’, ‘CD4 CM, CD4 MFI’, ‘CD4 EM, CD4 MFI’.

It is worth noting that all these variables show significant differences between Ag and Rh in the unstimulated case using the Mann-Whitney-U-test. For all ‘CD3 MFI’ the Rh have higher values while for all ‘CD28 MFI’ and ‘CD4 MFI’ their values are lower than the ones of Ag (4.4).



(a) African green monkeys (b) rhesus monkeys

Figure 4.4: Boxplots for the difference between unstimulated and stimulated cells in the **CD4**-dataset



(a) African green monkeys (b) rhesus monkeys

Figure 4.5: Boxplots for the difference of the proportions between unstimulated and stimulated cells in the **CD4**-dataset

Actually, the latter one, that Ag compared to the Rh have a higher count of `CD4+ MFI` in all groups we observed (which were the cells of the effective memory, cells of the central memory, naïve cells and all CD4+ cells of the sample), even when being unstimulated. This large amount of around 15,000 drops by 9,000 when the cell population

is stimulated. Similarly, the initial amount of approximately 12,000 for the Rh drops by around 9,000 after stimulation. However, the propoprtions of the CD4+ cells belonging to either the central or effective memory, or naïve cells, remains about the same (4.5).

This means that in the basic, unstimulated case, cell of Ag's already have a higher MFI of CD4 among the CD4+ cells i.e. on average there are more surface markers for CD4.

However, 'CD4 CM, CD3 MFI' is not significantly different between the two species and, also, is not assigned to any of the groups. This may explain that even though it does not differ too much between Ag and Rh, it still has to be looked at separately. A second group includes '%CD4 Naïve', '%CD4 EM','%CD4' and 'CD4 CD45RA MFI' for both Ag and Rh, but it rather is related with '%CD4 EM' for the Ag and additionally with 'CD4, CD95 MFI' for the Rh ((2),(1)).
A third group might consist out of 'CD4 EM, CD45RA MFI' and 'CD4 CD45RA MFI'. Allowing a higher threshold, additionally 'CD4 CM, CD45RA MFI' and 'CD4 Naïve, CD45RA MFI' join this group ((2),(1)).

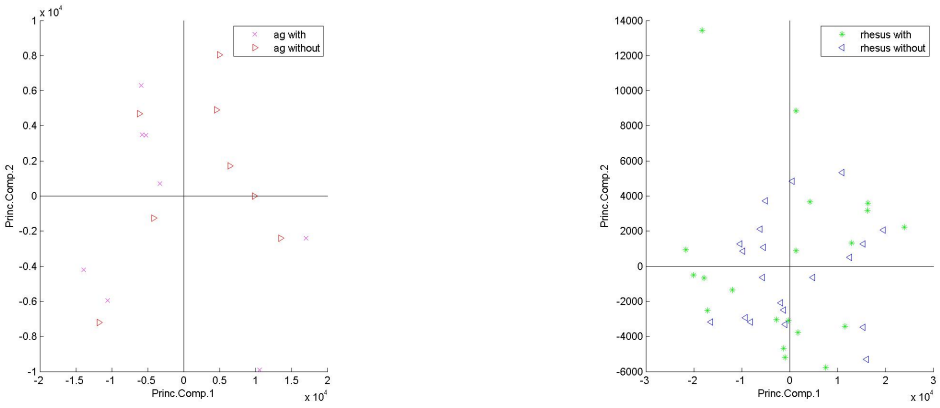
Considering now the stimulated case, these groups change even though the change for each parameter respectively is similar for both Ag and Rh (4.4). A considerable decrease can be seen for 'CD4, CD4 MFI', 'CD4 Naïve, CD4 MFI', 'CD4 CM, CD4 MFI' and 'CD4 EM, CD4 MFI' as well as a slight decrease in 'CD4 CM, CD45RA MFI' and 'CD4 Naïve, CD45RA MFI'(4.4).

Nevertheless, the first group from the unstimulated sample splits up. While all the 'CD4, CD28 MFI' stay together only with the 'CD4, CD3 MFI' for the Rh (4), they equally do so with the 'CD4, CD4 MFI' for the Ag. The 'CD4, CD3 MFI' for the Ag now form a new group together with the 'CD4 EM, CD45RA MFI'(3). Almost all 'CD4, CD4 MFI' for the Rh however are not allotted to any group. A second group for the Rh also include all 'CD4, CD45RA MFI', 'CD4 EM, CD95 MFI' as well as 'CD4 CM, CD95MFI'(5).

CD8

Here, again, we are dealing with the surface markers. Now, we examine CD8+ T-cells. After evaluating the outcome for CD8+ cells, we will go on and compare the results of CD4+ and CD8+ cells and their expression of surface markers. A complete list and explanation of the variables in the dataset '*RM_AGM_PMA + I_CD*8' can be found in appendix (5). These are almost the same as for the **CD4**-data, only lacking the number of lymphocytes, monocytes and granolucytes, compared to figure 4.1.

In contrast to the **CD4**-dataset, the pca within Ag and Rh does not reveal a clear distinction between the stimulated and unstimulated cell populations when taking a look at the first two principal components in (4.6) even though they explain 87% of the variance.

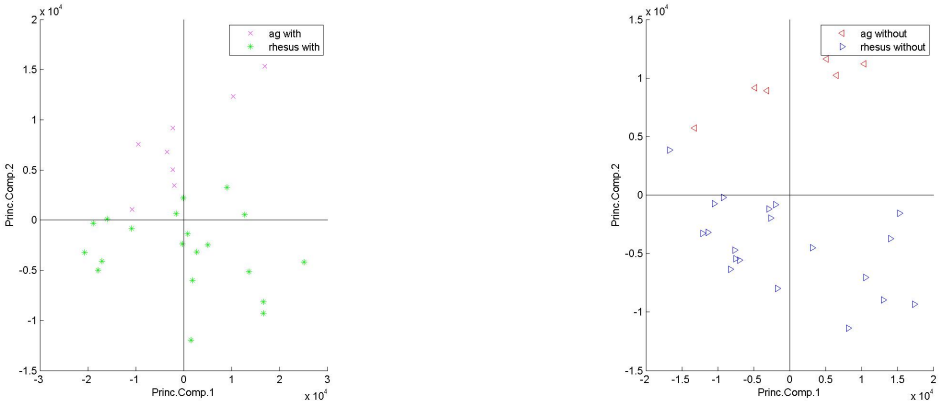


(a) cells of Ag

(b) cells of Rh

Figure 4.6: Gscatter-plots for cell populations of Ag and Rh in the **CD8**-dataset

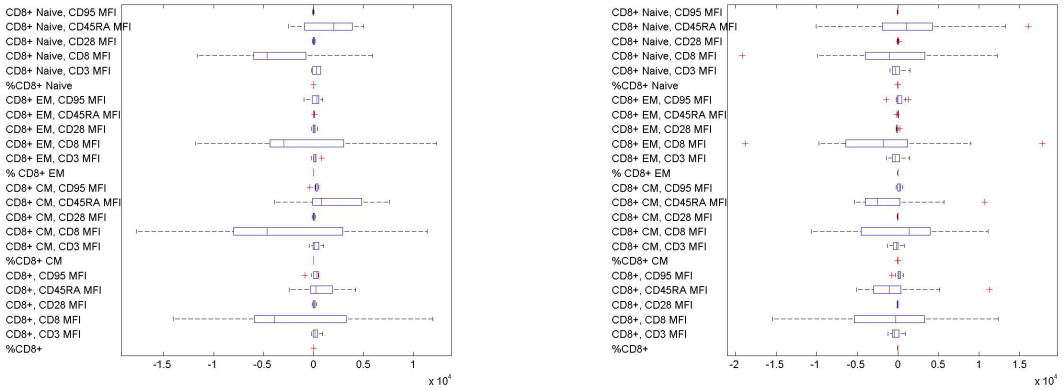
However, pca's performed for cell populations with and without stimulation respectively, shown in (4.7), reveals a palpable separation, with the first two components accounting for 89% (Ag) and 82% (Rh).



(a) stimulated cells (b) unstimulated cells

Figure 4.7: Gscatter-plots for the stimulated and unstimulated cell populations for the **CD8**-dataset

Taking a look at the changes from unstimulated to stimulated populations in (4.8), noticeable bars are the ones for all `'CD8 MFI'` as well as `'CD8 CM CD45RA MFI'` and `'CD8 Naïve, CD45RA MFI'` and additionally `'CD8 CD45RA MFI'` for Rh.



(a) African green monkeys (b) rhesus monkeys

Figure 4.8: Boxplots for the difference between unstimulated and stimulated cells in the **CD8**-dataset

Likewise, the direction of change is intriguing. While the values for `'CD8 MFI'` are

all decreasing for Ag, this is not true for 'CD8 CM, CD8 MFI' for the Rh. Equally, 'CD8 CM CD45RA MFI' and 'CD8 Naïve, CD45RA MFI' both increase with regard to the Ag, while this only holds for the latter for the Rh. Here, both 'CD8 CM CD45RA MFI' and 'CD8, CD45RA MFI' decrease. Still, in contrast to the **CD4**-dataset, for the CD8+ a clear statement cannot be made about CD8+ MFI. For the Ag, these amounts may decrease by up to 5000, but for the Rh the average stays in between -1000 and +1000.

These discrepancies result in ambiguous factor groups. For example in the unstimulated case, we detect one group for the Rh consisting of 'CD8, CD8 MFI', 'CD8 Naïve, CD8 MFI', 'CD8 CM, CD8 MFI' and 'CD8 EM, CD8 MFI' as well as 'CD8 Naïve, CD45RA MFI', 'CD8, CD45RA MFI' and 'CD8 CM, CD45RA MFI' being joined by 'CD8, CD3 MFI', 'CD8 Naïve, CD3 MFI', 'CD8 CM, CD3 MFI' and 'CD8 EM, CD3 MFI'. On the contrary for the ag, the former ones ('CD8 EM, CD8 MFI', 'CD8 CM, CD8 MFI', 'CD8, CD8 MFI') are combined with 'CD8, CD95 MFI', 'CD8 Naïve, CD45RA MFI' as well as '%CD4 CM', '%CD4 EM' and '%CD4'. Meanwhile, all the 'CD3 MFI' go mainly along with 'CD8 Naïve, CD45RA MFI' and 'CD8 CM, CD45RA MFI' and 'CD8 EM, CD95 MFI' and 'CD8, CD95 MFI'. For both Rh and Ag almost all 'CD28 MFI' form a separate group (Ag without 'CD28 CM, CD3 MFI', Rh joined by '%CD8', 'CD8 Naïve, CD95 MFI' and '%CD8 Naïve') (see (7) and (6)).

In the stimulated case the 'CD8 MFI' remain as one group just as in the unstimulated case. However, now they are joined by 'CD8, CD45RA MFI', 'CD8 Naïve, CD45RA MFI', 'CD8 Naïve, CD3 MFI', 'CD8 EM, CD3 MFI', 'CD8 EM, CD95 MFI' and even more variables for a lower threshold for the Rh (8) and for the Ag those are joined by all 'CD28 MFI' (10). Despite that, the last group of the Rh in the unstimulated case splits up to form one group out of, amongst other, '% CD8','% CD8 CM','% CD8 EM' and '% CD8 Naïve' and another group out of 'CD8 Naïve, CD28 MFI', 'CD8 CM, CD28 MFI' and 'CD8, CD28 MFI' (9). For the Ag, the group containing 'CD3 MFI' is joined by almost all 'CD95 MFI' and 'CD8 EM, CD45RA MFI'.

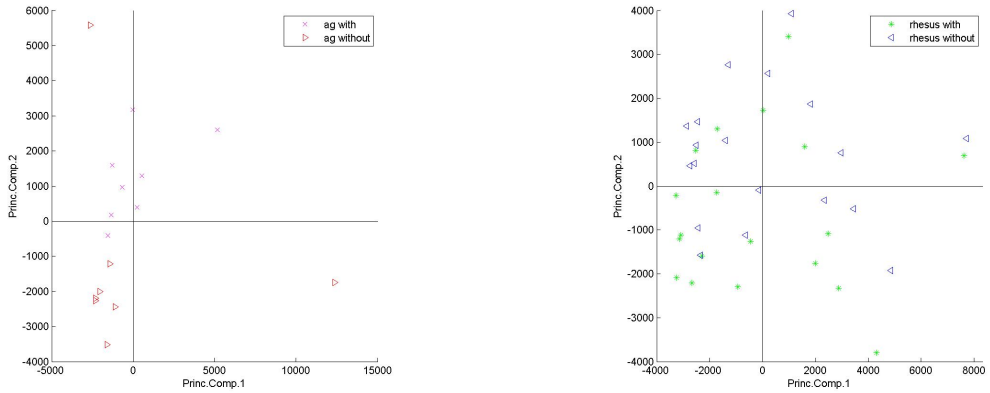
Both the **CD4**- and the **CD8**-dataset give information about the surface markers on CD4+ and CD8+ cells, respectively. For the CD4+ cells, it seems as if a certain pattern exists describing which surface markers of the EM, CM, etc. interact with each other. On the opposite, it appears to be much more difficult to find such a general scheme for the CD8+ cells.

Next, we examine cytokines within, again, both CD8+ and CD4+ cells. These are no surface markers but are secreted by the cells. How the measurements are proceeded is described in (2).

CD4 cytokines

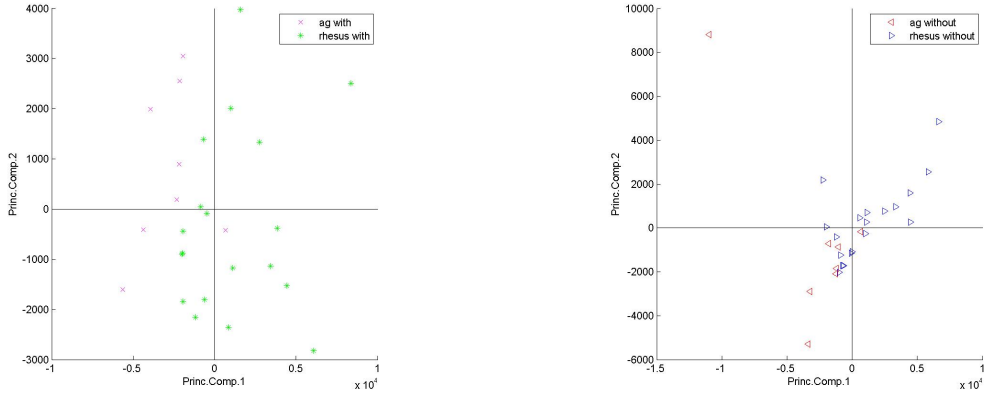
The dataset '*RM_AGM_PMA+I_CD4_Cytokines*' contains numerous variables listed in appendix (5). In contrast to the surface markers examined in the **CD8**- and **CD4**-dataset, we know consider cytokines which are secreted by CD4+ cells in the present dataset or by CD8+ cells in the following one. These are the tumor necrosis factor α (TNF), interleukin 2 (IL2) and interferon α (IFN).

Principal component analyses performed for cells of Rh and of Ag as well as for the unstimulated and stimulated cell populations do not reveal any enlightening information ((4.9),(4.10)). To achieve more than 90% to be explained by the variance we would have to take into account at least four principal components respectively.



(a) cells of Ag (b) cells of Rh

Figure 4.9: Gscatter-plots for cell populations of Ag and Rh in the **CD4 cytokines**-dataset



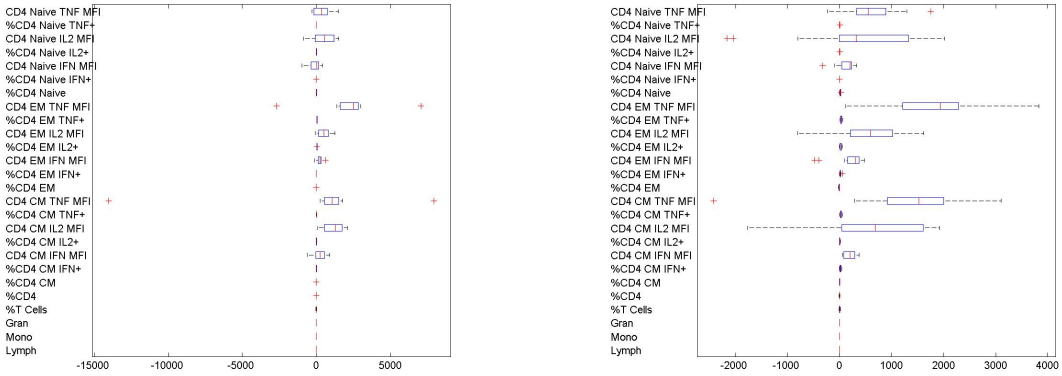
(a) stimulated cells (b) unstimulated cells

Figure 4.10: Gscatter-plots for the stimulated and unstimulated cell populations for the **CD4 cytokines**-dataset

The formation of groups is quite difficult to compare. The differences seem to outnumber the similarities when attempting this challenge. The groups given for a threshold of 0.5 and 3 factors are given in ((11),(13),(12),(14)).

Nevertheless, the changes taking place when the individuals are stimulated are more or less equivalent for both species. Thus, `'CD4 EM'`, `'CD4 CM'` and `'CD4 Naïve TNF`

MFI', 'CD4 EM', 'CD4 CM' and 'CD4 Naïve IL2 MFI', 'CD4 EM', 'CD4 CM' and 'CD4 Naïve IFN MFI' all undergo the most noticeable changes (4.11).



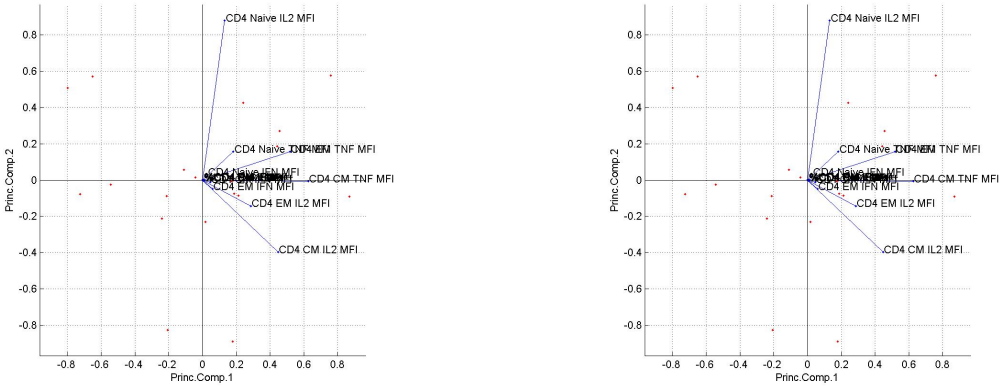
(a) African green monkeys

(b) rhesus monkeys

Figure 4.11: Boxplots for the difference between unstimulated and stimulated cells in the **CD4 cytokines**-dataset

Instead of trying to find any connection between the factor groups, we took the differences between stimulated and unstimulated cell population of Ag and Rh, respectively, and treated those as new datasets.

Again, a pca could not show any clustering when taking both Rh and Ag together. However, pca's for Ag and Rh separately, when considering only those variables identified by the Mann-Whitney-U-test to be signficantly different, may provide an argument. While the first two principal components explain 97% of the variance for the Ag, it needs at least four components for the Rh to achieve this number. The most important influences for the Rh are given by 'CD4 CM IL2 MFI', 'CD4 CM TNF MFI', 'CD4 EM TNF MFI' and additionally for the Ag by 'CD4 Naïve IL2 MFI' and can be seen in (4.12).



(a) African green monkeys

(b) rhesus monkeys

Figure 4.12: Biplot of pca performed for significant variables among the Ag and Rh in the **CD4 cytokines**-dataset

As a next step, we only took the significant different variables for Ag and Rh into account and used factor analysis. The first important thing to notice is the different amount of those parameters. These are more for the Rh than for Ag, and it contains all measurements of the effective memory for both and for the Rh all
measurements of the central memory and more than the Ag regarding the Naïve CD4+ cells (see the turquoise group in (16) and (15)).

Still, some peculiarities may be detected. For the Ag all values displaying `'%IL2'` and `'%TNF'` form one group even though the threshold has to be rather small to include `'CD4 CM %TNF'` (16). Meanwhile, for Rh such a pattern cannot be observed. Whereas all `'%TNF'` are accumulated in one group when employing a low threshold, this cannot be stated for either `'%IL2'` or `'%IFN'` (15)).

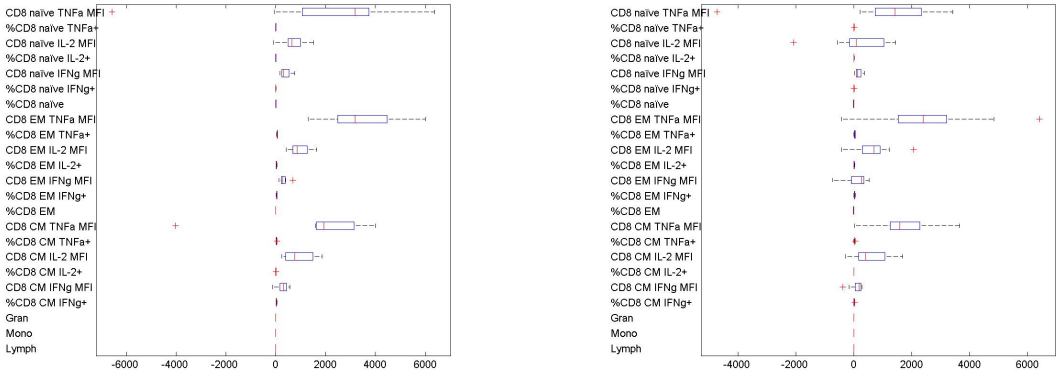
CD8 cytokines

The dataset '*RM_AGM_PMA+I_CD8_Cytokines*' comprises information on the same variables as '*RM_AGM_PMA+I_CD4_Cytokines*' except for CD8+ cells and the lack of

'%T Cells', '%CD4' and '%CD4 CM'. The pca's for the whole dataset show only slight separations when looking at the first two principal components only. In fact, for each pca at least three (pca comparing stimulated cell populations of Ag and Rh) or even four (pca comparing unstimulated cell populations of Ag and Rh, pca comparing stimulated and unstimulated cell populations of Ag, pca comparing stimulated and unstimulated cell populations of Rh) principal components to explain at least 90% of the variance within the data. Analyzing the group forming process, once more, is not easy. Particularly, when observing the data without stimulation there is no obvious rule how to distinguish the cell groups of Ag's and Rh's. The only visible similarity is that the measurements of the effective memory are partly connected to those of the central memory and to those of the naïve cells, though in different ways for the two species (17),(18).

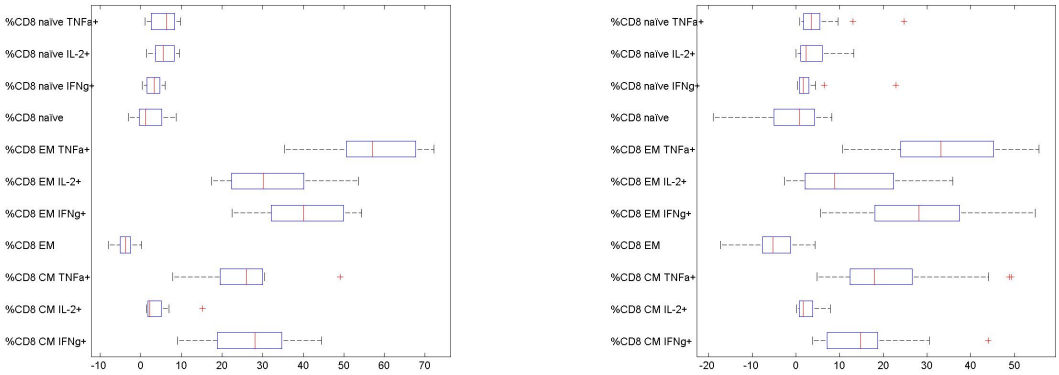
Notwithstanding, the groups in the stimulated case show certain analogies. Both for cell populations of Rh's and Ag's all the 'CD8 % IFN' and 'CD8 % TNF' are all (for a low threshold) in one group. In the same way, 'CD8 % IL2', 'CD8 IFN MFI', 'CD8 IL2 MFI' and 'CD8 TNF MFI' can be found together as shown in (19) and (20).

Observing the boxplots illustrates the change within the variables between the stimulated and unstimulated cell populations it is noteworthy that the percentage of IFN, TNF and IL2 do not change (4.14), however their absolute number equipollent changes dramatically (4.13).



(a) African green monkeys (b) rhesus monkeys

Figure 4.13: Boxplots for the difference between unstimulated and stimulated cells in the **CD8 cytokines**-dataset

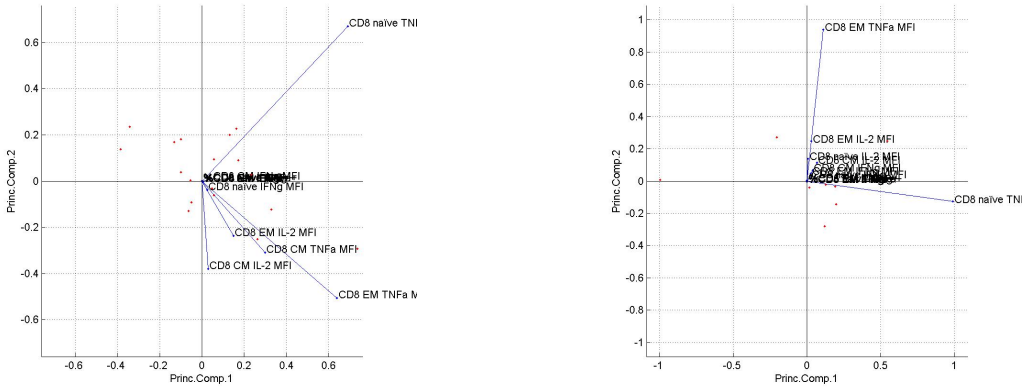


(a) African green monkeys (b) rhesus monkeys

Figure 4.14: Boxplots for the difference of the proportions between unstimulated and stimulated cells in the **CD8 cytokines**-dataset

Comparing the differences between the stimulated and unstimulated populations for each Ag's and Rh's discloses results akin to the **CD4 cytokines**-dataset. However, in the case at hand, all values for `'CD8 TNF MFI'`, `'CD8 IL2 MFI'` and `'CD8 IFN MFI'` increase and, on average, more for the Rh's than for the Ag's.

Considering only the significantly different variables for each Rh's and Ag's, which are especially for the Rh's almost all variables, the pca's supports this suggestions. The most influential variables here are `'CD8 EM TNF MFI'` and `'CD8 Naïve TNF MFI'`(4.15).



(a) African green monkeys (b) rhesus monkeys

Figure 4.15: Biplot of pca performed for significant variables among the Ag and Rh in the **CD8 cytokines**-dataset

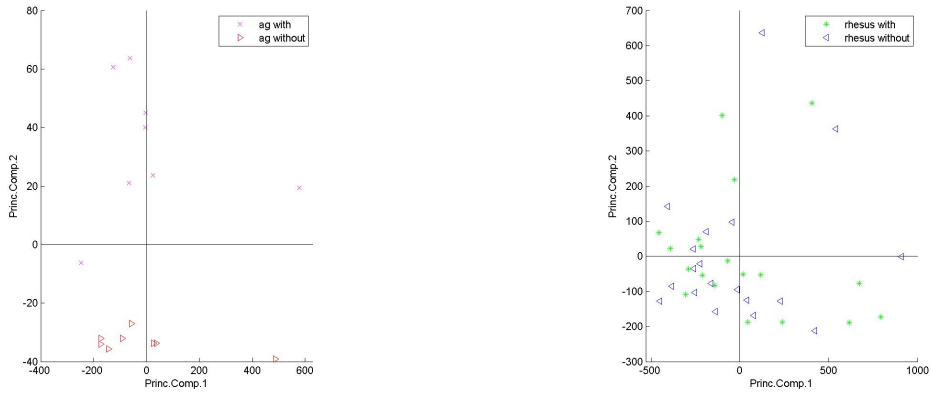
Even within the grouping, a separation of the percentages to the according MFI values can be noticed. Using the method of factor analysis, we recognize a division of `'% IL2'` and `'%IFN'` for both Ag's and Rh's and additionally, for the Rh's all the `'%IFN'` and `'%TNF'` accumulate (22),(21).

Looking back to the **CD4 cytokines**-dataset and the groups for their significant different variables, we might assume that a distinction between `'IL2'` values on one hand and `'TNF'` values on the other hand has to be made for both CD4+ and CD8+ cells for cell populations of the African Green as well as of Rhesus monkeys. Still, the role of `'IFN'` remains unclear as it ends up with `'%Il2'` for the Ag's in the **CD4 cytokines**-dataset but appears together in one group with `'%IFN'` for the Rh's in both the **CD4 cytokines**- and **CD8 cytokines**-dataset.

CD4 boolean

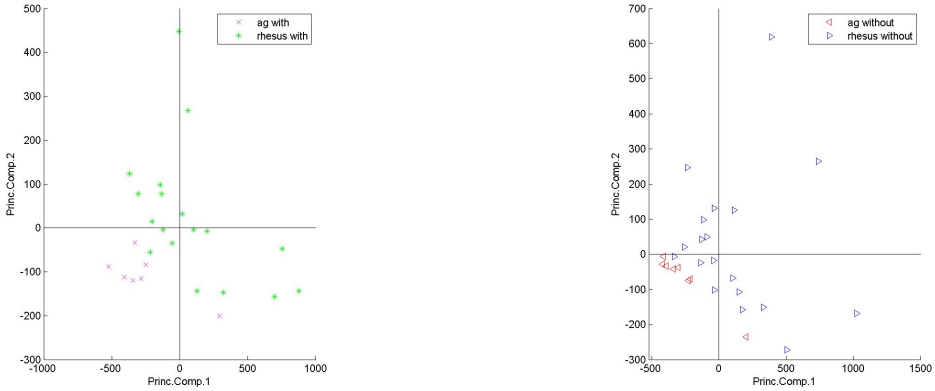
This dataset, '*RM_AGM_PMA+I_CD4_Boolean*', contains again information about the Tumor Necrosis Factor α (TNF), Interleukin 2 (IL2) and Interferon γ (IFN). This time, for each combination of these three factors it is accounted for. That means those cells being positive for one and negative for the other two or vice versa as well as all other combinations are measured. Indeed, the number of cells among the CD4+ cells being negative for all three IFN, TNF and IL2 outnumbers any other combination. Thus, the case `'TNF- IL2- IFN-'` is not considered and proprtions are taken of the remainder.

First, we examine the outcome of principal component analyses. For the cell populations of Ag's and Rh's a clear separation between the stimulated and unstimulated population can be seen neither for Ag's nor for Rh's (4.16). In both cases, the first two principal components explain about 98% of the variance. Similarly, performing the pca for the stimulated and unstimulated measurements gives 98% of explanation for the first two components. Here, the gscatter-plots show a distinction of Ag and Rh in (4.17) and those two first principal components even account for 99% of the variance within the data. Following the detailed results given by pca might be a good approach for further investigation on this particular dataset.



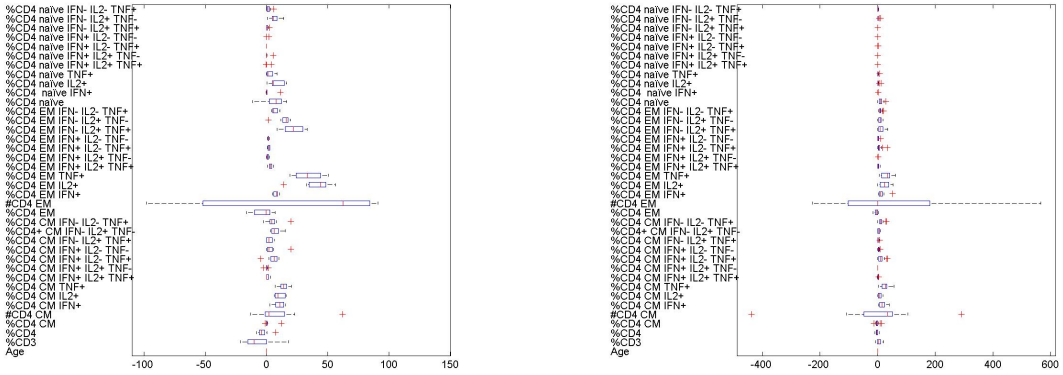
(a) cells of Ag (b) cells of Rh

Figure 4.16: Gscatter-plots for cell populations of Ag and Rh in the **CD4 boolean**-dataset



(a) stimulated cells (b) unstimulated cells

Figure 4.17: Gscatter-plots for the stimulated and unstimulated cell populations for the **CD4 boolean**-dataset

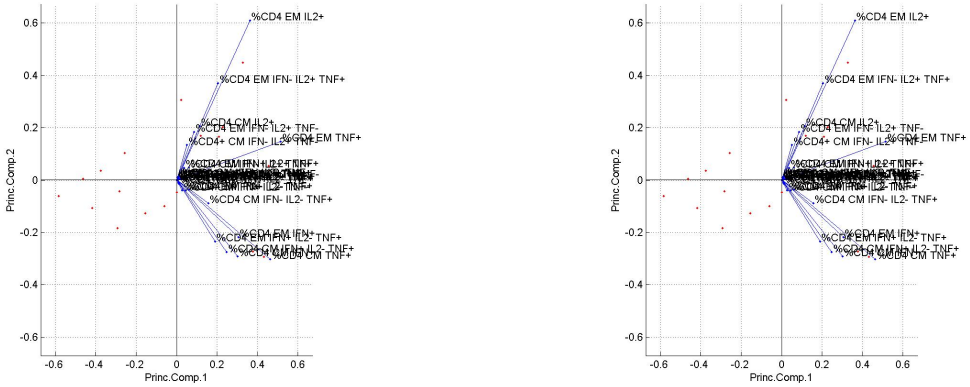


(a) African green monkeys (b) rhesus monkeys

Figure 4.18: Boxplots for the difference between unstimulated and stimulated cells in the **CD4 boolean**-dataset

Given the boxplots in 4.18, no obvious changes take place for the Rh, except for a wide range of change for `'#EM'` and `'# CM'`. In contrast, for the Ag's an increase in all `'% CM / EM  IFN / Il2 / TNF'` can be detected. The enormous gain of the percentage of these pathogen-reducing factors certainly should be examined further. With their help, there might be a way to differentiate Ag and Rh and learn why Ag do not get an AIDS-like disease upon SIV-infection. When it comes to the grouping, no such obvious observance can be made for the stimulated cell populations. In the unstimulated case, however, the main difference between cell populations of Ag's and Rh's lies in the separation of the cells of the central memory and the ones of the effective memory (23),(24). Whereas almost all cells of the CM are accumulated into one group for the Rh, a kindred grouping happens for the EM cells of the Ag.

Next, the significant different variables for the cells of each Rh's and Ag's are studied. The pca's bring up an explanation of 89% (Ag) and 93% (Rh) when considering the first three principal components. The main actors herein for the Ag cells are the proportions of TNF, IL2 and IFN within the EM and CM cells and for Rh's the main contribution seems to be made by `'%CD3'`, followed by `'% EM IL2'` (4.19).



(a) African green monkeys

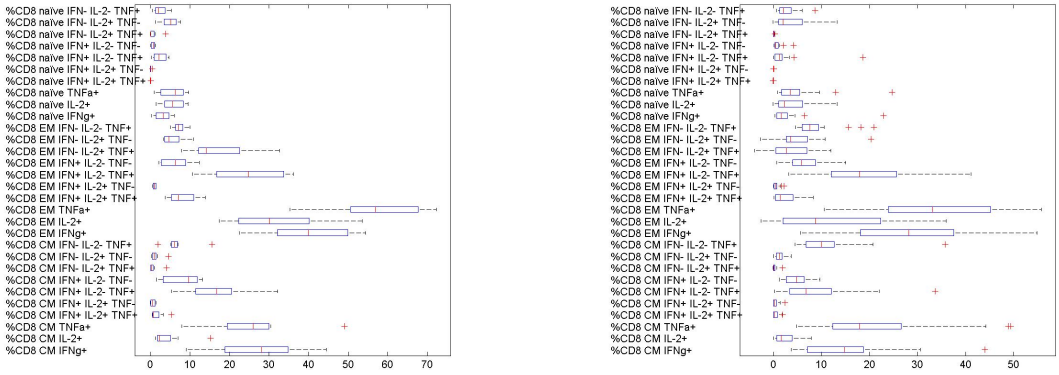
(b) rhesus monkeys

Figure 4.19: Biplot of pca performed for significant variables among the Ag and Rh in the **CD4 Boolean**-dataset

While for the cell population of the Rh almost all variables unveil significant differences between the stimulated and unstimulated case, there are only a few naïve cells but, all EM and almost all CM cells from the Ag emulate the Rh. The groups generated by the factor analysis indeed do project a certain setup. For the Ag, especially `'% TNF'` and
`'% Il2'` amount to one group (26) whereas for the Rh one group consists out of almost all naïve parameters joined by `'% CM TNF'` and `'% EM TNF'` and a second group conjoins `'% CM IL2'` and `'% EM Il2'` with those proportions of cells being positive for IL2 (e.g. `'%CD4 EM IFN- IL2+ TNF+'`) (25).

CD8 boolean

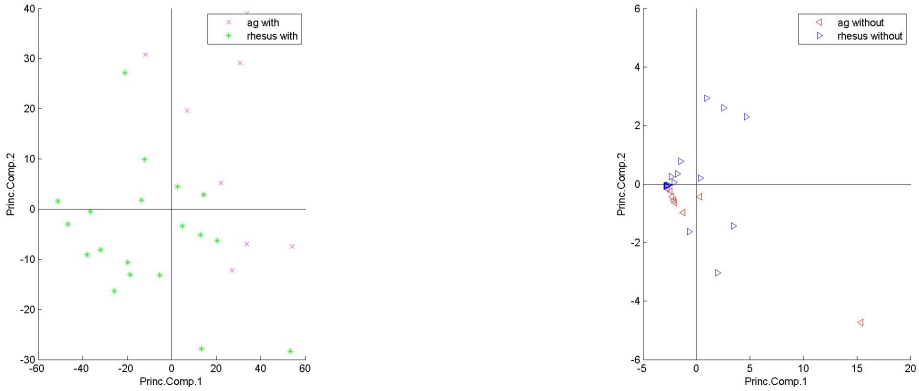
While in the **CD4 boolean**-dataset only few changes occur when the cell populations of Ag and Rh were stimulated, this is very different in the dataset '*RM_AGM_PMA+I_CD8_Boolean*'. Here, all values increase, and in explicit numbers, values for cells of Ag rise even more than those of Rh (4.20).



(a) African green monkeys (b) rhesus monkeys

Figure 4.20: Boxplots for the difference between unstimulated and stimulated cells in the **CD8-Boolean**-dataset

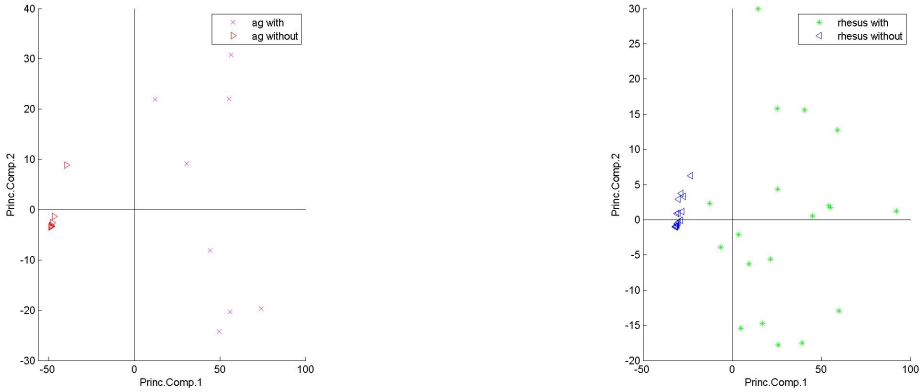
The highest growth occurs for `'%IFN'`, `'%TNF'` and `'%IL2'` equally for cells of Rh and Ag. Thus, the difference of the two immune systems might not be found regarding the CD8+ cell population. Considering the pca's among stimulated and unstimulated cell populations, a distinction may be made with help of the first two components in the unstimulated case, which justify 97% , and with the help of at least 3 components in the stimulated case explaining about 89% . Still, gscatter-plots picturing the first two components shed light on the difference between Ag's and Rh's (4.21).



(a) stimulated cells (b) unstimulated cells

Figure 4.21: Gscatter-plots for the stimulated and unstimulated cell populations for the **CD8-Boolean**-dataset

A definite discrepancy between stimulated and unstimulated cell populations for both Ag's and Rh's is apparent and depicted by the gscatter-plots 4.22.



(a) cells of Ag (b) cells of Rh

Figure 4.22: Gscatter-plots for cell populations of Ag and Rh in the **CD8-Boolean**-dataset

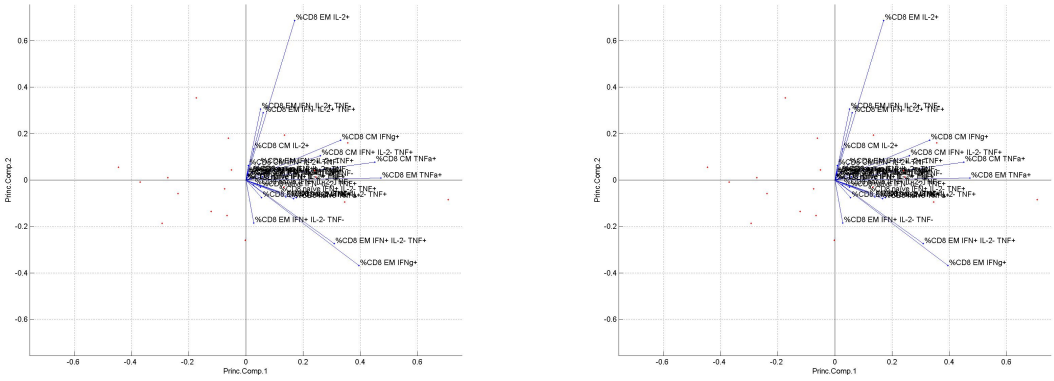
The factor analysis even generates similar groups for both Ag and Rh in the stimulated case. Compared to the uneven distribution of variables to groups in the **CD4 boolean**-dataset in the stimulated case, this might be a slight relation within CD8+ cells. However,

relations that contribute to discover similarities instead of differences.

One of the factor groups contains ‘% CD8 EM IL2’, ‘% CD8 CM IL2’ and associated parameters such as ‘%CD4 EM IFN- IL2+ TNF+’, ‘%CD4 EM IFN+ IL2+ TNF+’ and ‘%CD4 EM IFN+ IL2+ TNF-’. A second group is dominated by ‘% TNF’ and matching variables, however for the Rh’s in connection with ‘% IFN’ but for the Ag’s the latter form a separate group together with their corresponding values. All ‘%CD4 IFN- IL2- TNF+’ compose a single group for the Rh’s while ‘%TNF’ belongs to another group. Bearing in mind the boxplots 4.20, it is self-explanatory that almost all variables show significant differences when measured within stimulated and unstimulated cell populations.

Similar, but not equivalent to the **CD4 boolean**-dataset ‘%EM IL2’, ‘%EM IFN’, ‘%EM TNF’, ‘%CM TNF’ and ‘%CM IFN’ play a mayor role for the pca of Ag in which the first two component already account for 90% of the variance (4.23).

For the Rh a different picture reveals. The same variables as for the Ag in this dataset contribute to the principle components (4.23), however it needs four components to achieve at least a 94% -explanation of the variance. The grouping shows a somewhat clear distinction of all ‘%IL2’ to ‘%TNF’ and ‘%IFN’ for both Ag’s and Rh’s.



(a) African green monkeys

(b) rhesus monkeys

Figure 4.23: Biplot of pca performed for significant variables among the Ag and Rh in the **CD8 boolean**-dataset

Furthermore, one might suggest a certain relation between naïve cells and cells of the central memory for both species as certain values of these two belong to the same group even for a low threshold.

Summing up the evaluations of the **CD4 boolean**-dataset and the **CD8 boolean**-dataset, one might not recognize a specific outcome. However, some particularities are noticeable. First, the change of the actual values of the variables differs greatly between the two datasets, giving an indication of the different roles CD4+ and CD8+ cells play in the immune response. Second, for the Ag a certain distinction of `'%IL2'` from both `' %TNF'` and `'%IFN'` can be identified, albeit this is only true for `'%IL2'` of the cells of the effective and central memory but not for the naïve ones considering the cells of the Rhesus monkeys.

5 | Conclusion

There were two kinds of cells studied, CD4+ T cells and CD8+ T cells (see (2) for detailed explanations). For both of these cell types that play a crucial role in the immune system, two important contributing parts were examined. First, we observed surface markers on CD4+ and CD8+ cells. Second, we observed cytokines which are usually produced and secreted by CD4+ and CD8+ cells.

For the former, the following distinctive features were observed.
Looking at CD4+ cells, the simple comparison between the parameters of the unstimulated and stimulated cell populations shows a specific feature. While the value of `CD4+ MFI` dropped equivalently for both Rh and Ag, the baseline amount for Ag was much higher (4.4). Similarly, the groups generated by factor analysis (fa) changed for both species from the unstimulated to the stimulated case.

For the unstimulated cell populations, we can see similar groupings for both cells of Ag and of Rh. Even though the variables including `CD28 MFI`, `CD4 MFI` and `CD3 MFI` are significantly different between Rh and Ag in the unstimulated case, they still form one group for each species (1), (2).

Considering the stimulated cell populations however, reveals a different image. For the Rh all the `CD3 MFI` seem to indicate the same latent factor as all `CD28 MFI` (5), while for the Ag the latter seem to be related to `CD4 MFI`, and the values for `CD3 MFI` form rather a separate group, at least separated from `CD4 MFI` (see (3)). This separation also happens for the stimulated Rh cell population. The `CD4 MFI` are not even assigned to any

group anymore.

For the **CD8**-dataset such a pattern is not observable. Even the comparison of values for the stimulated and unstimulated cell populations gives limited prospects for future studies. The only noticeable difference between Ag and Rh might be, that while the values of `CD8 MFI` increase for CM, EM and naïve cells for the Ag, this is only true for the two latter ones when considering the cells of Rh. `'CD8+ CM, CD8 MFI'` even shows a slight decrease on average for the Rh. The same phenomenon can be observed for `'CD8+ CM, CD45RA MFI'` and `'CD8+, CD45RA MFI'` (see (4.8)).

Taking these first two datasets into account, we can suggest, that clarifying the role of CD4+ cells calls for subsequent research. Here, patterns can be observed. For the Rh, for example, there might be a third parameter influencing both CD28+ and CD3+ surface markers, while another latent factor might give information about the relation between CD28+ and CD4+ surface markers. As said before, for CD8+ cells no such general idea was detected in order to find differences between the two species, but, there might be one to reveal similarities.

Moreover, cytokines being produced and secreted by both CD4+ and CD8+ cells were analyzed. To summarize all the findings for the associated datasets, a general trend can be noted for the Ag. Groupings for the stimulated and unstimulated cell populations as well as groupings for the change being undergone from unstimulated to stimulated cells, reveal a certain basic pattern. It seems as if, for both CD4+ and CD8+ cells, the amount and proportion of IL2 and TNF are predicated by one latent factor respectively. Even though their relation is not quite obvious when taking high thresholds, a more general assumption can be made, saying that IL2 can be strictly separated from IFN regarding the cells of Ag. Even though the biological relation between all three cytokines cannot be denied, the performance of an exploratory factor analysis did not bring them into one group being dependent on one major hidden factor. This can be seen for the **CD4 cytokines**-dataset in figure (16) as well as for the **CD8 cytokines**-dataset in figures (19) and

(22) and for the **CD8 boolean**-dataset and **CD4 boolean**-dataset in the figures (30), (32) and (26).

In contrast, such a clear structure cannot be seen for the cells of Rh. The only repeating occurrence is that variables containing TNF-α (especially `% TNF`) seem to form one group generated by fa. While for the Ag the parameters for IL2 sometimes were joined by variables presenting TNF-α, this does not happen for Rh. In fact, these two relatively oppose each other, so that they rather occur in different groups (see especially (15) and (28)), but still are not mutually exclusive. Another difference from the Ag can be found in the correspondence of TNF and IFN. These two and associated parameters are noticed to end up in the same factor groups on several occasions, but still not in general (see (20), (29)).

Unquestionably, there are more relations which can be detected from the results of the performed methods. However, the different behaviour of the three cytokines TNF-α, IL2 and IFN-γ stated can be equally certified by the analyzes. Although much research was done on the correlation between HIV and IL2 (([38])), TNF ([39]) and IFN ([36]) respectively, there is little information given on their interaction (see e.g. ([40]), ([25])) within infected individuals. We can state that there might be a certain connection between these cytokines that differs between pathogenic (Rh) and non-pathogenic (Ag) populations. Thus, it can be an evidence for not only latent factors but some other reasonable explanation on how the infection by SIV, and similarly HIV, is handled by Ag and how HIV might be defeated in order to prevent AIDS.

Bibliography

[1] Alan J. Izenman, *Modern Multivariate Statistical Techniques - Regression, Classification, and Manifold Learning.* Springer Science+Business Media, LLC, New York, 2008.

[2] M. Mengozzi, M. Malipatolla et al., *Naïve CD4 T cells inhibit CD28-costimulazed R5 HIV replication in CD4 T cells* published in 'PNAS', vol.98, September 25 2001

[3] C. S. Subauste, *CD154 and Type-1 Cytokine Response: From Hyper IgM Syndrome to Human Immunodeficiency Virus Infection.* published in 'The Journal of Infectious Diseases', 185(Suppl1), 2002.

[4] W. F. Ng et al., *Human CD4+CD25+ cells: a naturally occuring population of regulatory T cells.* published in 'Blood' by The American Society of Hermatology, Vol 98 Number 9, 2001.

[5] B. Jacquelin et al., *Nonpathogenic SIV infection of African green monkeys induces a strong but rapidly controlled type I IFN response.* published in 'The Journal of Clinical Investigation', Volume 119 Number 12, 2009.

[6] M. R. Reynolds et al., *Macaques Vaccinated with Simian Immunodeficiency Virus SICmac239 Δ nef Delay Acquisition and Control Replication after Repeated Low-*

Dose Heterologous SIV Challenge. published in 'The Journal of Virology', Vol 84, 2010.

[7] E. Holznagel et al., *Immunological changes in simian immunodeficiency virus (SIV_{agm})-infected African green monkeys (AGM): expanded cytotoxic T lymphocyte, natural killer and B cell subsets in the naturl host of SIV_{agm}.* published in 'Journal of General Virology', 83 GB , 2002.

[8] UCFlow, *http://ucflow.blogspot.com/2009/04/what-is-mfi.html.* published by The University of Chicago Flow Cytometry Facility , 2009.

[9] *http://labtestsonline.org/understanding/analytes/cd4/tab/test.* ,published by by American Association for Clinical Chemistry , last reviewed 2012.

[10] M. B. Gross , *http://omim.org/entry/186940.* published by 'Online Mendelian Inheritance in Man (OMIM)' , 2011.

[11] Y. Pacheco et al, *Despite an impaired response to IL-7, CD4+EM T cells from HIV-positive patients proliferate normally in response to IL-15 and its superagonist, RLI.* published by 'AIDS' , 2011.

[12] eBioscience, Inc. , *Human CD & Other Cellular Antigens Antibodies for multicolor flow cytometry, functional assays and immunohistochemistry.* http://www.ebioscience.com/resources/human-cd-chart.htm#human-cd-chart, 03/2013.

[13] Hua W. et al., *Central memory CD4 cells are an early indicator of immune reconstitution in HIV/AIDS patients with anti-retroviral treatment.* published in 'Immunological Investigations', Vol 41(1), 2012.

[14] K. Tassiopoulos et al., *CD28-negative CD4+ and CD8+ T cells in antiretroviral therapy-naive HIV-infected adults enrolled in adult clinical trials group studies.* published in 'The Journal of Infectious Diseases' , Vol 205(11) , 2012 .

[15] J. Gamberg et al., *Lack of CD28 expression on HIV-specific cytotoxic T lymphocytes is associated with disease progression.* published in 'Immunology and cell biology',Vol 82(1) ,2004 .

[16] L. Chen at al. , *CD95 promotes tumour growth.* published in 'Nature', Vol 465, 2010.

[17] Y.M. Mueller et al., *Increased CD95/Fas-induced apoptosis of HIV-specific CD8(+) T cells.* published in Immunity', Vol 15(6), 2001.

[18] S.P. Aries , *Fas (CD95) expression on CD4+ T cells from HIV-infected patients increases with disease progression.* published in The Journal of Molecular Medicine, Berlin'(Germany)' , 1995.

[19] *http://www.pathologyoutlines.com/topic/cdmarkerscd45ra.html.* published by PathologyOutlines.com, Inc., Michigan, last revised 2012.

[20] F. Sallusto et al. , *Two subsets of memory T lymphocytes with distinct homing potentials and effector functions.* published in 'Nature' ,Vol 401, 1999.

[21] C. R. Mackay , *Dual personality of memory T cells.* published in 'Nature', Vol 401, 1999.

[22] M. Berard and D.F. Tough, *Qualitative differences between naïve and memory T cells.* published in 'Immunology' , Vol 106(2) , 2002.

[23] T. Willinger et al. , *Molecular signatures distinguish human central memory from effector memory CD8 T cell subsets.* published in 'The Journal of Immunology',Vol 175(9) , 2005.

[24] M. Ramaswamy et al., *Specific elimination of effector memory CD4p T cells due to enhanced Fas signaling complex formation and association with lipid raft microdomains.* published in 'Cell Death and Differentiation', Vol 18, 2011.

[25] Prof Dr H Ibelgaufts, *http://www.copewithcytokines.de/cope.cgi?key=IFN-gamma.* Cytokines & Cells Online Pathfinder Encyclopedia, Version 31.4, 2013 .

[26] M. Croft, *The role of TNF superfamily members in T-cell function and diseases.* published in 'Nature Reviews Immunology', Vol 9(4), 2009.

[27] T.N. Golovina et al., *CD28 costimulation is essential for human T regulatory expansion and function.* published in 'The Journal of Immunology', Baltimore, 2008.

[28] *http://www.gmi.oeaw.ac.at/research-groups/magnus-nordborg/population-genetics-of-african-green-monkeys.* published by 'Gregor Mendel Institute of Molecular Plant Biology', Vienna, 2013.

[29] *http://www.iucnredlist.org/details/4233/0.* hosted by 'International Union for Conservation of Nature and Natural Resources', 2013, cited from IUCN 2012. IUCN Red List of Threatened Species. Version 2012.2.

[30] *http://www.avert.org/origin-aids-hiv.htm.* hosted by 'Avert', UK, downloaded April 2013.

[31] *Primate Factsheets: Rhesus macaque (Macaca mulatta) Taxonomy, Morphology, & Ecology . <http://pin.primate.wisc.edu/factsheets/entry/rhesus_macaque>.* published by Cawthon Lang KA, 2005.

[32] *http://animals.nationalgeographic.com/animals/mammals/rhesus-monkey/.* hosted by National Geographic Society, 2013.

[33] *http://www.pharmazeutische-zeitung.de/index.php?id=4019.* published by Govi-Verlag, 2013, originally printed in 2007.

[34] *http://www.cdc.gov/hiv/resources/factsheets/us.htm.* hosted by the Centers for Disease Control and Prevention CDC, last reviewed Feb,2013.

[35] Diane V. Havlir et al., *Serum Interleukin-6 (IL-6), IL-10, Tumor Necrosis Factor (TNF) Alpha, Soluble Type II TNF Receptor, and Transforming Growth Factor Beta Levels in Human Immunodeficiency Virus Type 1-Infected Individuals with Mycobacterium avium Complex Disease.* published in the 'Journal of Clinical Microbiology, January 2001, Vol 39(1)

[36] Simone C. Zimmerli et al., *HIV-1-specific IFN-/IL-2-secreting CD8 T cells support CD4-independent proliferation of HIV-1-specific CD8 T cells.* published in PNAS, May 2005, Vol.102(20)

[37] I. Sereti et al. ,IL-2Űinduced CD4+ T-cell expansion in HIV-infected patients is associated with long-term decreases in T-cell proliferation. published in 'Blood' by The American Society of Hermatology in 2004, Vol.104(3)

[38] Laura Sivitz, *IL-2 Immunotherapy Fails to Benefit HIV-Infected Individuals Already Taking Antiretrovirals.* published by the National Institute of Allergy and Infectious Diseases (NIAID), Feb. 2009.

[39] Salazar-Gonzalez JF et al. ,*Relationship of plasma HIV-RNA levels and levels of TNF-alpha and immune activation products in HIV infection.* published in the 'Clinical Immunology and Immunopathology', Juli 1997, Vol.84(1).

[40] *http://www.hprd.org/diseases?hprd_id=00957&isoform_id=00957_1&isoform_name=.* hosted by Johns Hopkins University and the Institute of Bioinformatics, last modified 2005.

[41] Z. Ghahramani & G. E. Hinton, *The EM Algorithm for Mixtures of Factor Analyzers.* Technical Report CRG-TR-96-1, May 1996, Toronto (Canada).

[42] *http://www.cytokine-index.com/showProtein.php?proteinId=434.* published by PeproTech, 2012.

[43] L. R. Tucker, R. C. MacCullum, *Exploratory Factor Analysis.* University of Illinois, Ohio State University, 1997.

[44] D. J. Bartholomew, M. Knott, *Latent Variable Models and Factor Analysis* Kendall's Library of Statistics 7, Oxford University Press Inc., New York, 1999.

[45] K. Backhaus, B. Erichson, W. Plinke, R. Weiber, *Multivariate Analysemethoden - Eine anwendungsorientierte Einführung.* Springer-Verlag, 1996.

[46] K. Überla, *Faktorenanalyse.* Springer-Verlag, Ulm, 1971.

[47] *http://stat.ethz.ch/R-manual/R-devel/library/Matrix/html/nearPD.html.* hosted by the Swiss Federal Institute of Technology, Zurich, 2013.

Appendix: Datasets and Figures

CD4

The variables given in the dataset '*RM_AGM_PMA + I_CD*4 are

Animal ID Identification number for each individual

Stim Status of Stimulation (None or PMA+I)

Lymph Number of lymphocytes in the sample

Mono Number of monocytes in the sample

Gran Number of granolucytes in the sample

% T Cells Proportion of T cells in the sample

%CD4 Proportion of CD4+ cells (among the T cells ???)

CD4, CD3 MFI MFI of CD3 among the CD4+ cells

CD4, CD4 MFI MFI of CD4 among the CD4+ cells

CD4, CD28 MFI MFI of CD28 among the CD4+ cells

CD4, CD45RA MFI MFI of CD45RA among the CD4+ cells

CD4, CD95 MFI MFI of CD95 among the CD4+ cells

%CD4 CM Proportion of CD4+ cells of the Central Memory (among the CD4+ cells)

CD4 CM, CD3 MFI MFI of CD3 among the CM of the CD4+ cells

CD4 CM, CD4 MFI MFI of CD4 among the CM of the CD4+ cells

CD4 CM, CD28 MFI MFI of CD28 among the CM of the CD4+ cells

CD4 CM, CD45RA MFI MFI of CD45RA among the CM of the CD4+ cells

CD4 CM, CD95 MFI MFI of CD95 among the CM of the CD4+ cells

%CD4 EM Proportion of CD4+ cells of the Effective Memory (among the CD4+ cells)

CD4 EM CD3 MFI MFI of CD3 among the EM of the CD4+ cells

CD4 EM, CD4 MFI MFI of CD4 among the EM of the CD4+ cells

CD4 EM, CD28 MFI MFI of CD28 among the EM of the CD4+ cells

CD4 EM, CD45RA MFI MFI of CD45RA among the EM of the CD4+ cells

CD4 EM, CD95 MFI MFI of CD95 among the EM of the CD4+ cells

%CD4 Naïve Proportion of the naïve CD4+ cells (among the CD4+ cells)

CD4 Naïve, CD3 MFI MFI of CD3 among the Naive CD4+ cells

CD4 Naïve, CD4 MFI MFI of CD4 among the Naive CD4+ cells

CD4 Naïve, CD28 MFI MFI of CD28 among the Naive CD4+ cells

CD4 Naïve, CD45RA MFI MFI of CD45RA among the Naive CD4+ cells

CD4 Naïve, CD95 MFI MFI of CD95 among the Naive CD4+ cells

Figure 1: Factor groups for unstimulated cells of Rhesus monkeys with 3 factors and threshold 0.5 in the **CD4**-dataset

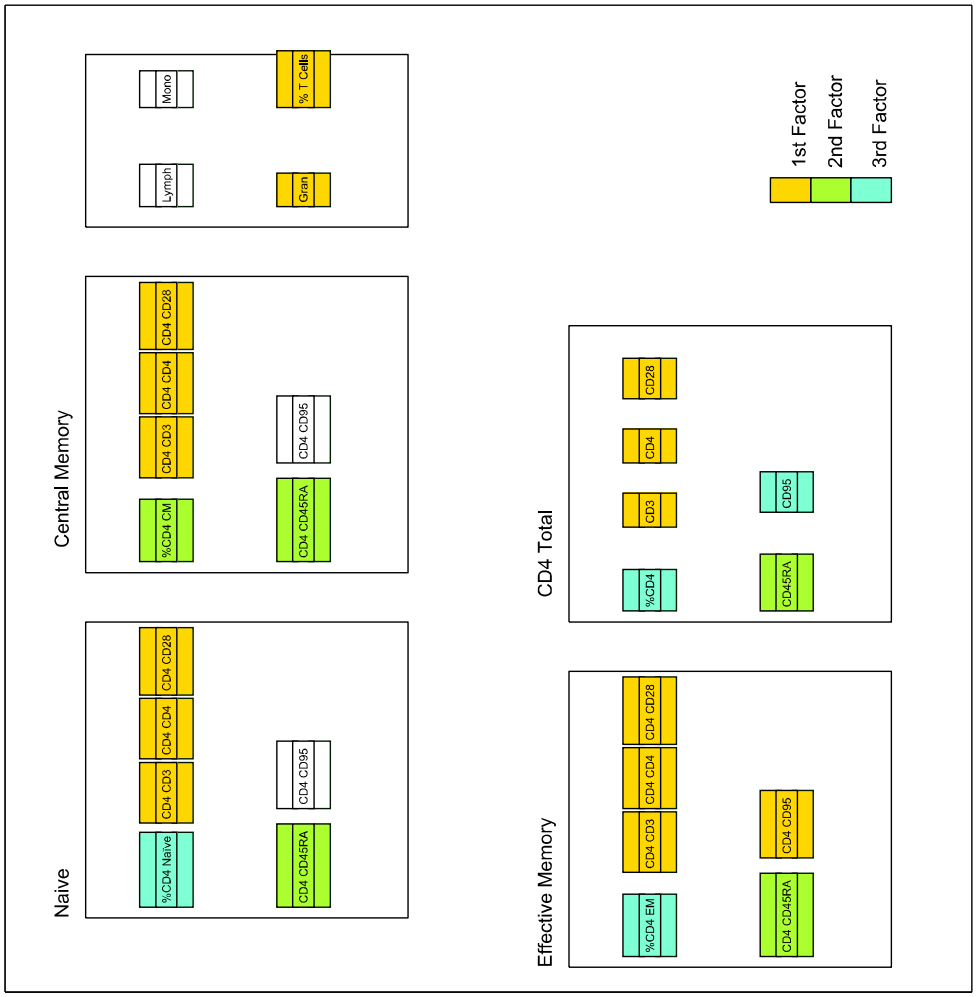


Figure 2: Factor groups for unstimulated cells of African green monkeys with 3 factors and threshold 0.6 in the **CD4**-dataset

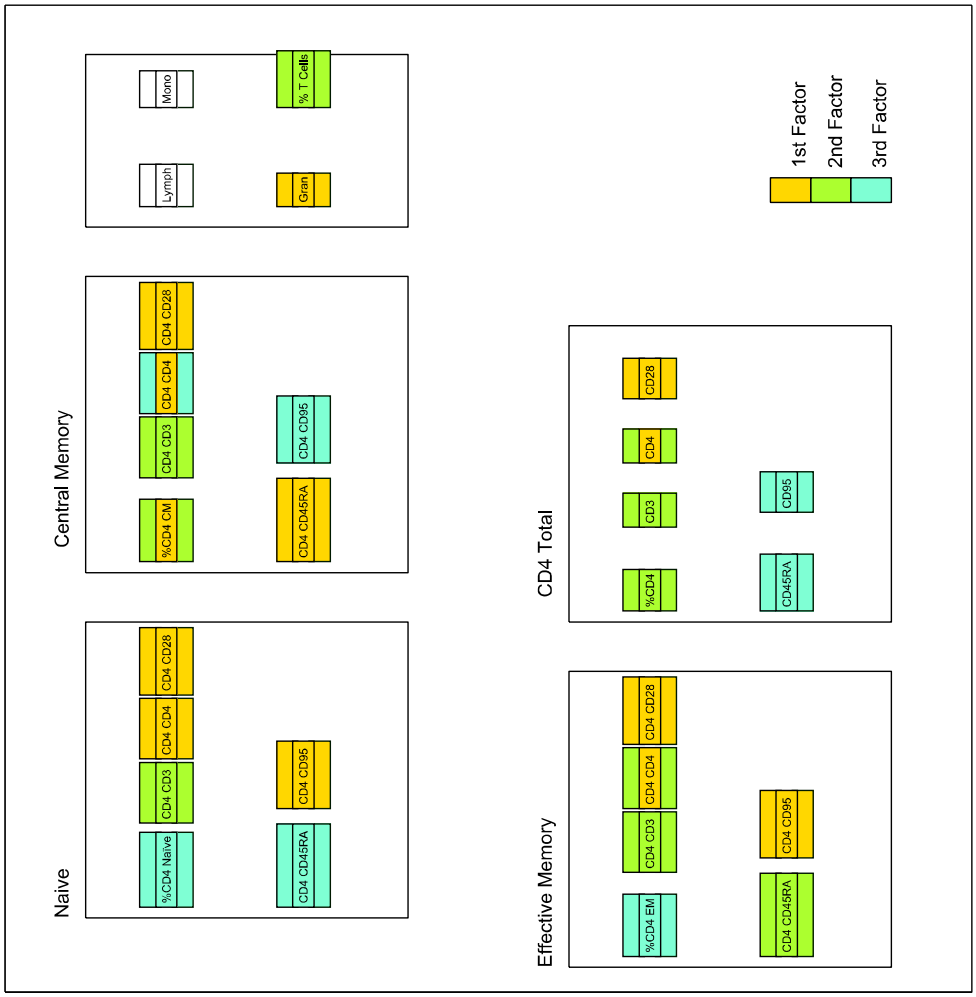


Figure 3: Factor groups for stimulated cells of African green monkeys with 3 factors and threshold 0.5 in the **CD4**-dataset

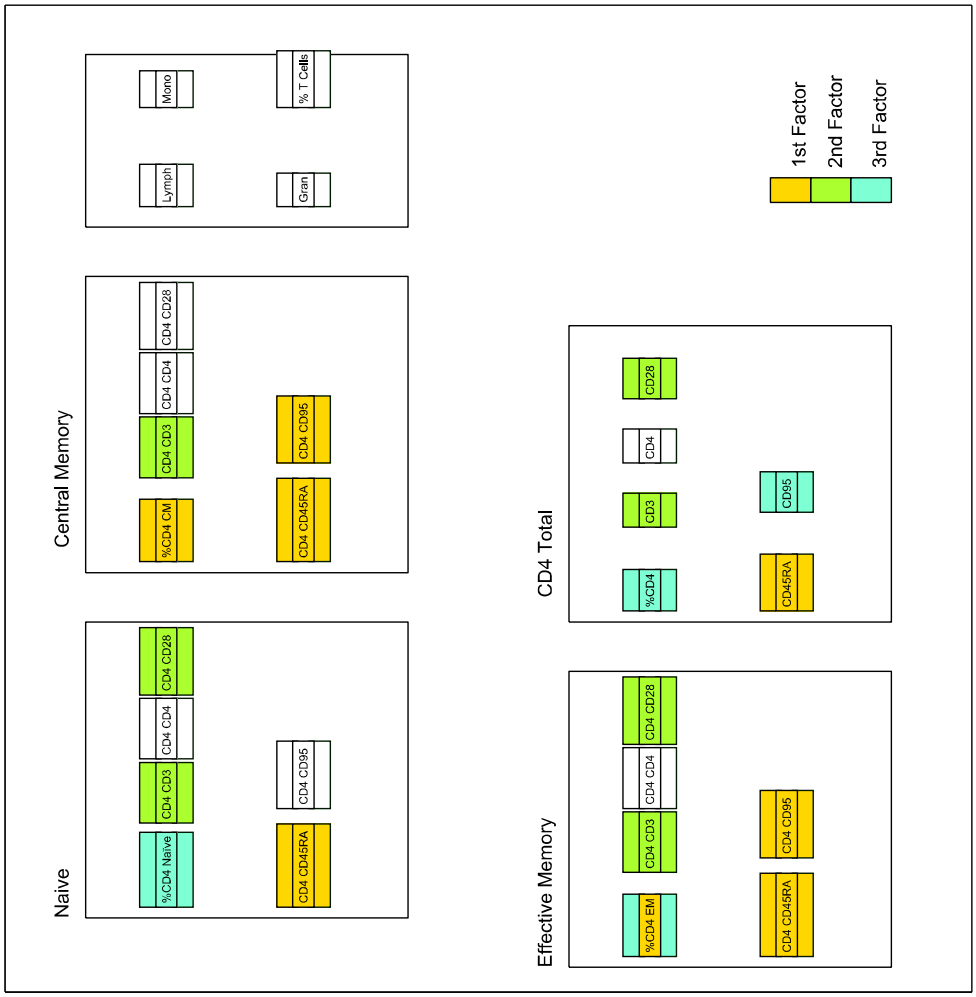


Figure 4: Factor groups for stimulated cells of Rhesus monkeys with 3 factors and threshold 0.6 in the **CD4**-dataset

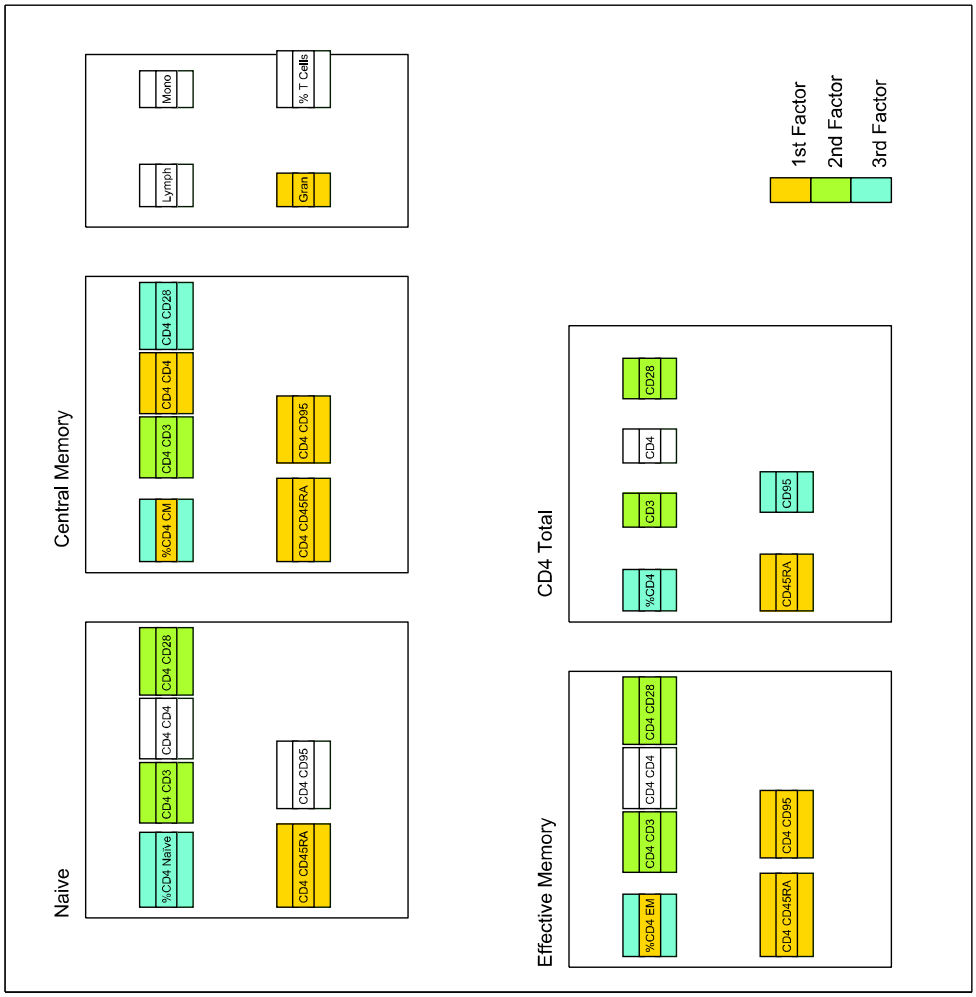


Figure 5: Factor groups for stimulated cells of Rhesus monkeys with 3 factors and threshold 0.5 in the **CD4**-dataset

CD8

The variables given in the dataset '*RM_AGM_PMA + I_CD*8 are

Animal ID and Stim Identification number for each individual and status of Stimulation (None or PMA+I)

%CD8 Proportion of CD8+ cells (among the T cells ???)

CD8, CD3 MFI MFI of CD3 among the CD8+ cells

CD8, CD8 MFI MFI of CD8 among the CD8+ cells

CD8, CD28 MFI MFI of CD28 among the CD8+ cells

CD8, CD45RA MFI MFI of CD45RA among the CD8+ cells

CD8, CD95 MFI MFI of CD95 among the CD8+ cells

%CD8 CM Proportion of CD8+ cells of the Central Memory (among the CD8+ cells)

CD8 CM, CD3 MFI MFI of CD3 among the CM of the CD8+ cells

CD8 CM, CD8 MFI MFI of CD8 among the CM of the CD8+ cells

CD8 CM, CD28 MFI MFI of CD28 among the CM of the CD8+ cells

CD8 CM, CD45RA MFI MFI of CD45RA among the CM of the CD8+ cells

CD8 CM, CD95 MFI MFI of CD95 among the CM of the CD8+ cells

%CD8 EM Proportion of CD8+ cells of the Effective Memory (among the CD8+ cells)

CD8 EM CD3 MFI MFI of CD3 among the EM of the CD8+ cells

CD8 EM, CD8 MFI MFI of CD8 among the EM of the CD8+ cells

CD8 EM, CD28 MFI MFI of CD28 among the EM of the CD8+ cells

CD8 EM, CD45RA MFI MFI of CD45RA among the EM of the CD8+ cells

CD8 EM, CD95 MFI MFI of CD95 among the EM of the CD8+ cells

%CD8 Naïve Proportion of the naïve CD8+ cells (among the CD8+ cells)

CD8 Naïve, CD3 MFI MFI of CD3 among the Naive CD8+ cells

CD8 Naïve, CD8 MFI MFI of CD8 among the Naive CD8+ cells

CD8 Naïve, CD28 MFI MFI of CD28 among the Naive CD8+ cells

CD8 Naïve, CD45RA MFI MFI of CD45RA among the Naive CD8+ cells

CD8 Naïve, CD95 MFI MFI of CD95 among the Naive CD8+ cells .

Figure 6: Factor groups for unstimulated cells of Rhesus monkeys with 3 factors and threshold 0.5 in the **CD8**-dataset

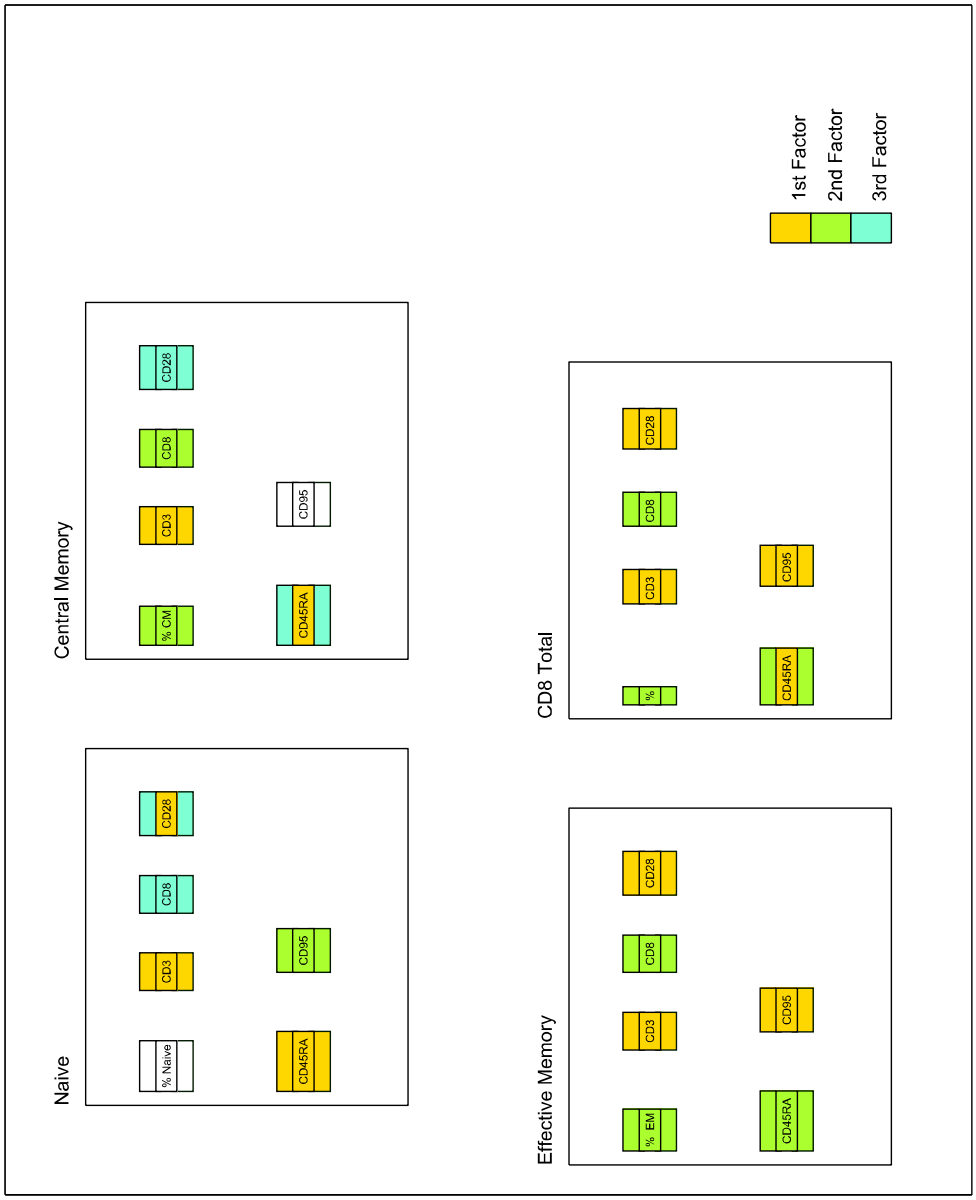


Figure 7: Factor groups for unstimulated cells of African green monkeys with 3 factors and threshold 0.5 in the **CD8**-dataset

Figure 8: Factor groups for stimulated cells of Rhesus monkeys with 3 factors and threshold 0.7 in the **CD8**-dataset

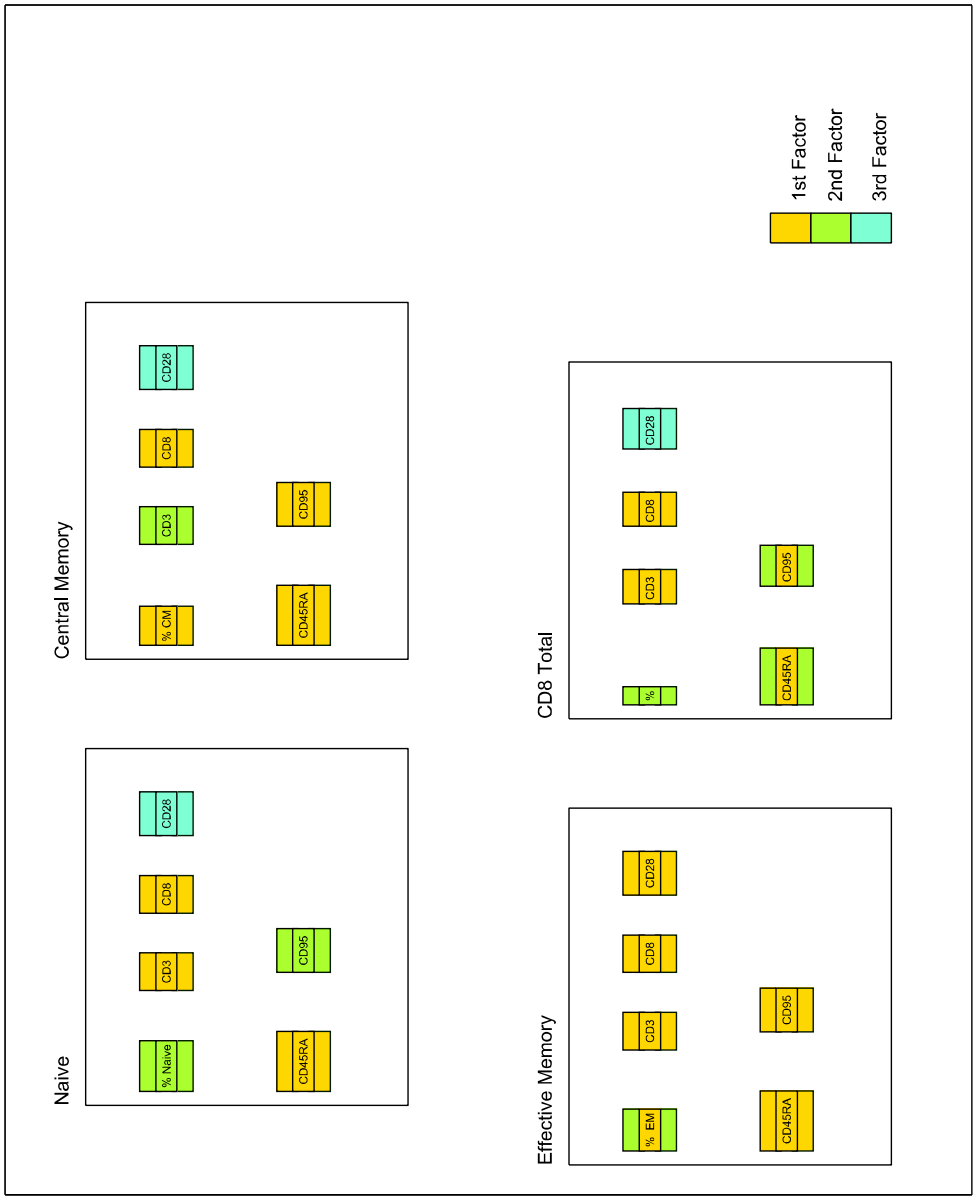


Figure 9: Factor groups for stimulated cells of Rhesus monkeys with 3 factors and threshold 0.5 in the **CD8**-dataset

Figure 10: Factor groups for stimulated cells of African green monkeys with 2 factors and threshold 0.5 in the **CD8**-dataset

CD4 cytokines

The variables given in the dataset '*RM_AGM_PMA + I_CD*4*_Cytokines* are

Lymph Number of lymphocytes in the sample

Mono Number of monocytes in the sample

Gran Number of granolucytes in the sample

%T Cells Proportion of T-cells in the sample

%CD4 Proportion of CD4+ in the sample

%CD4 CM Proportion of CD4+ cells of the central memory (among the CD4+ cells)

%CD4 CM IFN+ Proportion of the CD4+ IFN+ cells within the central memory

CD4 CM IFN MFI MFI of IFN among the cells of the central memory

%CD4 CM IL2+ Proportion of the CD4+ IL2+ cells within the central memory

CD4 CM IL2 MFI MFI of IL2 among the cells of the central memory

%CD4 CM TNF+ Proportion of the CD4+ TNF+ cells within the central memory

CD4 CM TNF MFI MFI of TNF among the cells of the central memory

%CD4 EM Proportion of CD4+ cells of the effective Memory (among the CD4+ cells)

%CD4 EM IFN+ Proportion of the CD4+ IFN+ cells within the effective memory

CD4 EM IFN MFI MFI of IFN among the cells of the effective memory

%CD4 EM IL2+ Proportion of the CD4+ IL2+ cells within the effective memory

CD4 EM IL2 MFI MFI of IL2 among the cells of the effective memory

%CD4 EM TNF+ Proportion of the CD4+ TNF+ cells within the effective memory

CD4 EM TNF MFI MFI of TNF among the cells of the effective memory

%CD4 Naïve Proportion of the naïve CD4+ cells (among the CD4+ cells)

%CD4 Naïve IFN+ Proportion of the CD4+ IFN+ cells among the naïve cells

CD4 Naïve IFN MFI MFI of IFN among the naïve cells

%CD4 Naïve IL2+ Proportion of the CD4+ IL2+ cells among the naïve cells

CD4 Naïve IL2 MFI MFI of IL2 among the naïve cells

%CD4 Naïve TNF+ Proportion of the CD4+ TNF+ cells among the naïve cells

CD4 Naïve TNF MFI MFI of TNF among the naïve cells

Figure 11: Factor groups for unstimulated cells of African green monkeys with 3 factors and threshold 0.5 in the **CD4 cytokines**-dataset

Figure 12: Factor groups for unstimulated cells of Rhesus monkeys with 3 factors and threshold 0.5 in the **CD4 cytokines**-dataset

Figure 13: Factor groups for stimulated cells of African green monkeys with 3 factors and threshold 0.5 in the **CD4 cytokines**-dataset

Figure 14: Factor groups for stimulated cells of Rhesus monkeys with 3 factors and threshold 0.5 in the **CD4 cytokines**-dataset

Figure 15: Factor groups for significant different variables of stimulated and unstimulated cells of Rhesus monkeys with 2 factors and threshold 0.5 in the **CD4 cytokines**-dataset

Figure 16: Factor groups for significant different variables of stimulated and unstimulated cells of African green monkeys with 2 factors and threshold 0.6 in the **CD4 cytokines**-dataset

CD8 cytokines

The variables given in the dataset '*RM_AGM_PMA + I_CD*8*_Cytokines* are

Lymph Number of lymphocytes in the sample

Mono Number of monocytes in the sample

Gran Number of granolucytes in the sample

%CD8 CM IFNg+ Proportion of the CD8+ IFN+ cells within the central memory

CD8 CM IFNg MFI MFI of IFN among the cells of the central memory

%CD8 CM IL-2+ Proportion of the CD8+ IL2+ cells within the central memory

CD8 CM IL-2 MFI MFI of IL2 among the cells of the central memory

%CD8 CM TNFa+ Proportion of the CD8+ TNF+ cells within the central memory

CD8 CM TNFa MFI MFI of TNF among the cells of the central memory

%CD8 EM Proportion of CD8+ cells of the effective Memory (among the CD8+ cells)

%CD8 EM IFNg+ Proportion of the CD8+ IFN+ cells within the effective memory

CD8 EM IFNg MFI MFI of IFN among the cells of the effective memory

%CD8 EM IL-2+ Proportion of the CD8+ IL2+ cells within the effective memory

CD8 EM IL-2 MFI MFI of IL2 among the cells of the effective memory

%CD8 EM TNFa+ Proportion of the CD8+ TNF+ cells within the effective memory

CD8 EM TNFa MFI MFI of TNF among the cells of the effective memory

%CD8 Naïve Proportion of the naïve CD8+ cells (among the CD8+ cells)

%CD8 Naïve IFNg+ Proportion of the CD8+ IFN+ cells among the naïve cells

CD8 Naïve IFNg MFI MFI of IFN among the naïve cells

%CD8 Naïve IL-2+ Proportion of the CD8+ IL2+ cells among the naïve cells

CD8 Naïve IL-2 MFI MFI of IL2 among the naïve cells

%CD8 Naïve TNFa+ Proportion of the CD8+ TNF+ cells among the naïve cells

CD8 Naïve TNFa MFI MFI of TNF among the naïve cells

Figure 17: Factor groups for unstimulated cells of African green monkeys with 3 factors and threshold 0.5 in the **CD8 cytokines**-dataset

Figure 18: Factor groups for unstimulated cells of Rhesus monkeys with 3 factors and threshold 0.5 in the **CD8 cytokines**-dataset

Figure 19: Factor groups for stimulated cells of African green monkeys with 2 factors and threshold 0.5 in the **CD8 cytokines**-dataset

Figure 20: Factor groups for stimulated cells of Rhesus monkeys with 2 factors and threshold 0.5 in the **CD8 cytokines**-dataset

Figure 21: Factor groups for significant different variables of stimulated and unstimulated cells of Rhesus monkeys with 2 factors and threshold 0.6 in the **CD8 cytokines**-dataset

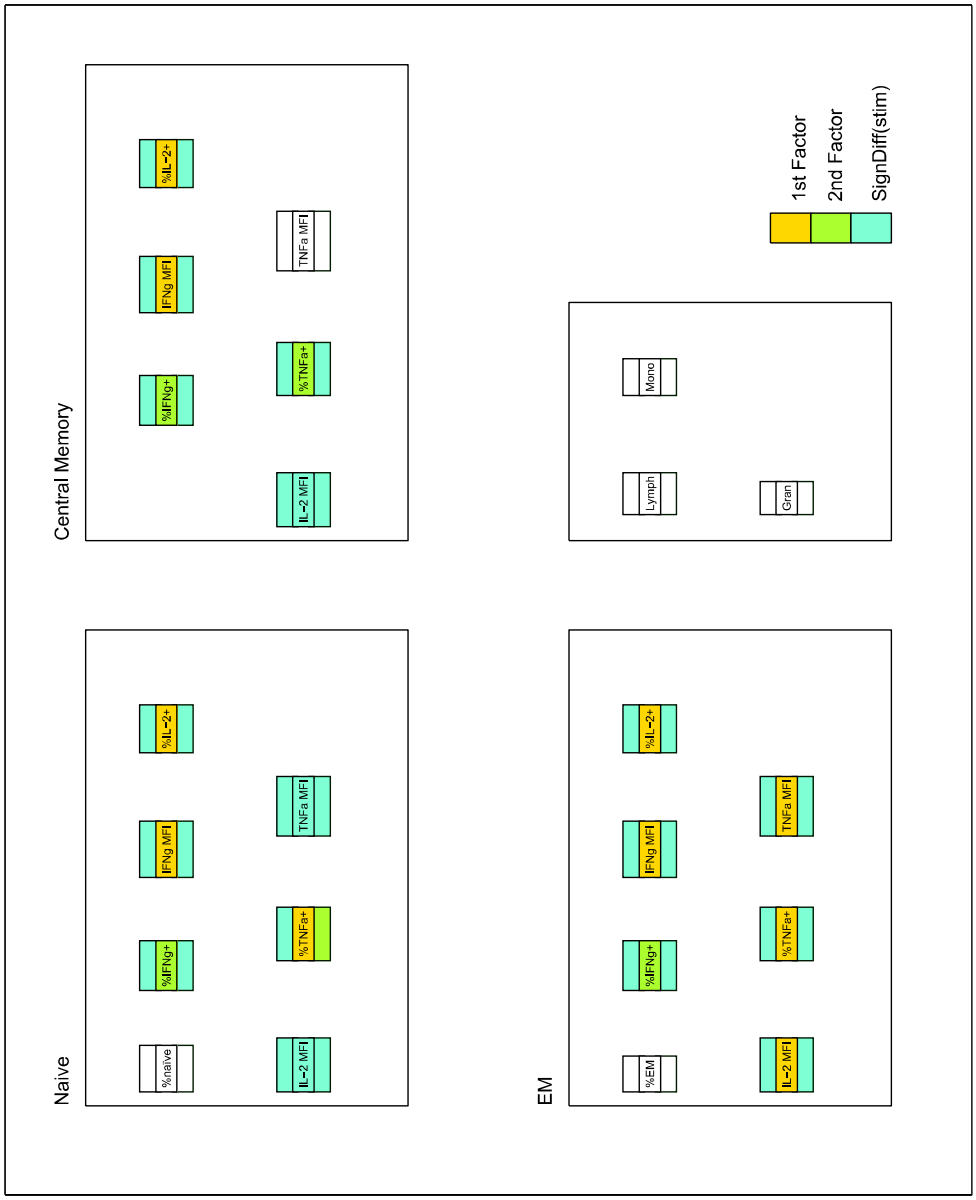


Figure 22: Factor groups for significant different variables of stimulated and unstimulated cells of African green monkeys with 2 factors and threshold 0.6 in the **CD8 cytokines**-dataset

CD4 boolean

The variables given in the dataset '*RM_AGM_PMA+I_CD4_Boolean* are

Sex gender of the individuals (male/female)

Age Age of the individuals

%CD3 Proportion of CD3+ cells in the sample

%CD4 Proportion of CD4+ cells in the sample

%CD4 CM Proportion of CD4+ cells of the central memory (among the CD4+ cells)

#CD4 CM Actual number of CD4+ cells of the central memory

%CD4 CM IFN+ Proportion of the CD4+ IFN+ cells within the central memory

%CD4 CM IL2+ Proportion of the CD4+ IL2+ cells within the central memory

%CD4 CM TNF+ Proportion of the CD4+ TNF+ cells within the central memory

%CD4 CM IFN+ IL2+ TNF+ Proportion of IFN+ IL2+ TNF+ cells within the central memory

%CD4 CM IFN+ IL2+ TNF- Proportion of IFN+ IL2+ TNF- cells within the central memory

%CD4 CM IFN+ IL2- TNF+ Proportion of IFN+ IL2- TNF+ cells within the central memory

%CD4 CM IFN+ IL2- TNF- Proportion of IFN+ IL2- TNF- cells within the central memory

%CD4 CM IFN- IL2+ TNF+ Proportion of IFN- IL2+ TNF+ cells within the central memory

%CD4 CM IFN- IL2+ TNF- Proportion of IFN- IL2+ TNF- cells within the central memory

%CD4 CM IFN- IL2- TNF+ Proportion of IFN- IL2- TNF+ cells within the central memory

%CD4 EM Proportion of CD4+ cells of the effective memory (among the CD4+ cells)

#CD4 EM Actual number of CD4+ cells of the effective memory

%CD4 EM IFN+ Proportion of the CD4+ IFN+ cells within the effective memory

%CD4 EM IL2+ Proportion of the CD4+ IL2+ cells within the effective memory

%CD4 EM TNF+ Proportion of the CD4+ TNF+ cells within the effective memory

%CD4 EM IFN+ IL2+ TNF+ Proportion of IFN+ IL2+ TNF+ cells within the effective memory

%CD4 EM IFN+ IL2+ TNF- Proportion of IFN+ IL2+ TNF- cells within the effective memory

%CD4 EM IFN+ IL2- TNF+ Proportion of IFN+ IL2- TNF+ cells within the effective memory

%CD4 EM IFN+ IL2- TNF- Proportion of IFN+ IL2- TNF- cells within the effective memory

%CD4 EM IFN- IL2+ TNF+ Proportion of IFN- IL2+ TNF+ cells within the effective memory

%CD4 EM IFN- IL2+ TNF- Proportion of IFN- IL2+ TNF- cells within the effective memory

%CD4 EM IFN- IL2- TNF+ Proportion of IFN- IL2- TNF+ cells within the effective memory

%CD4 naïve Proportion of the naïve CD4+ cells (among the CD4+ cells)

%CD4 Naïve IFN+ Proportion of the CD4+ IFN+ naïve cells (among the naïve cells)

%CD4 Naïve IL2+ Proportion of the CD4+ IL2+ naïve cells (among the naïve cells)

%CD4 Naïve TNF+ Proportion of the CD4+ TNF+ naïve cells (among the naïve cells)

%CD4 Naïve IFN+ IL2+ TNF+ Proportion of the CD4+ IFN+ IL2+ TNF+ naïve cells (among the naïve cells)

%CD4 Naïve IFN+ IL2+ TNF- Proportion of the CD4+ IFN+ IL2+ TNF+ naïve cells (among the naïve cells)

%CD4 Naïve IFN+ IL2- TNF+ Proportion of the CD4+ IFN+ IL2- TNF+ naïve cells (among the naïve cells)

%CD4 Naïve IFN+ IL2- TNF- Proportion of the CD4+ IFN+ IL2- TNF- naïve cells (among the naïve cells)

%CD4 Naïve IFN- IL2+ TNF+ Proportion of the CD4+ IFN- IL2+ TNF+ naïve cells (among the naïve cells)

%CD4 Naïve IFN- IL2+ TNF- Proportion of the CD4+ IFN- IL2+ TNF- naïve cells (among the naïve cells)

%CD4 Naïve IFN- IL2- TNF+ Proportion of the CD4+ IFN- IL2- TNF+ naïve cells (among the naïve cells)

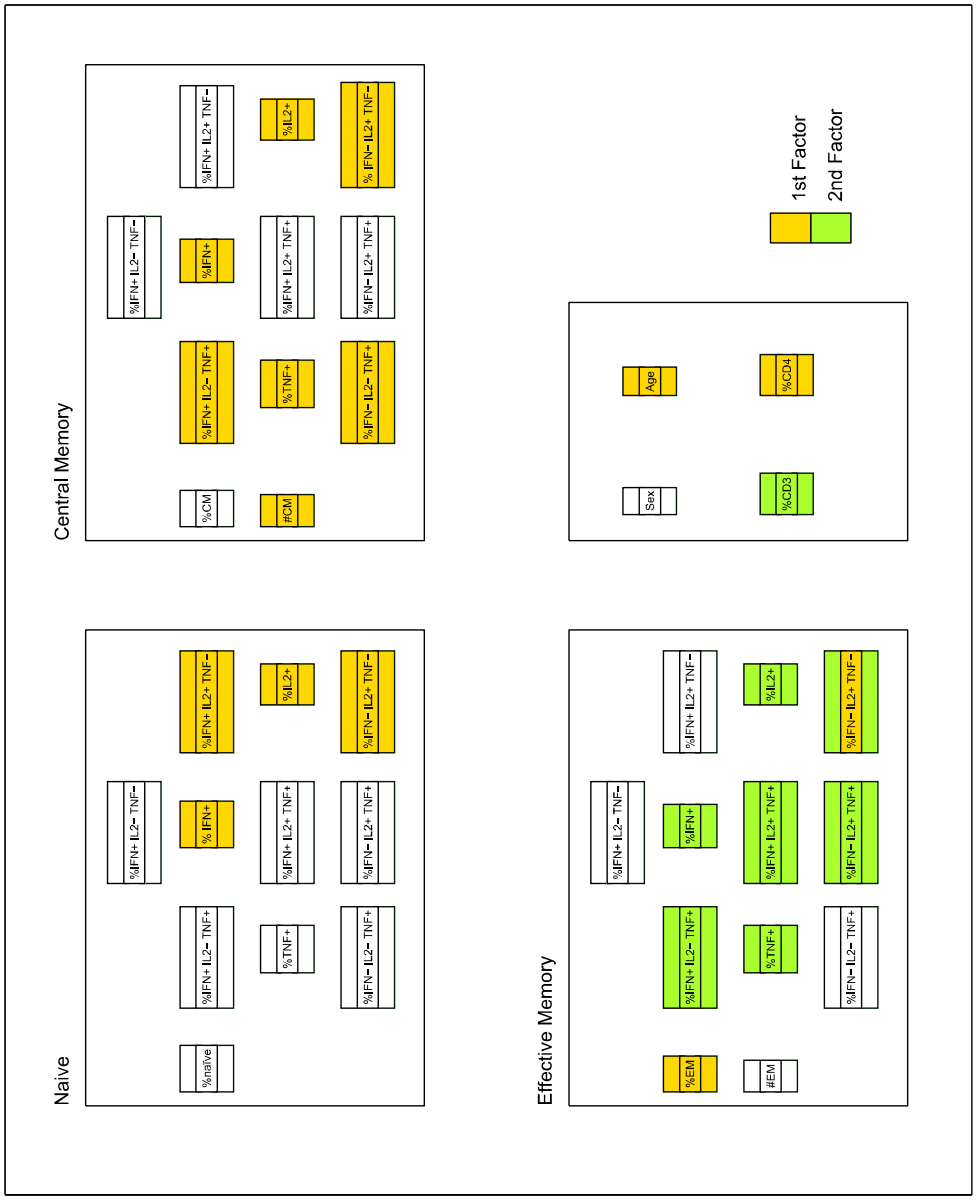


Figure 23: Factor groups for unstimulated cells of African green monkeys with 2 factors and threshold 0.5 in the **CD4 boolean**-dataset

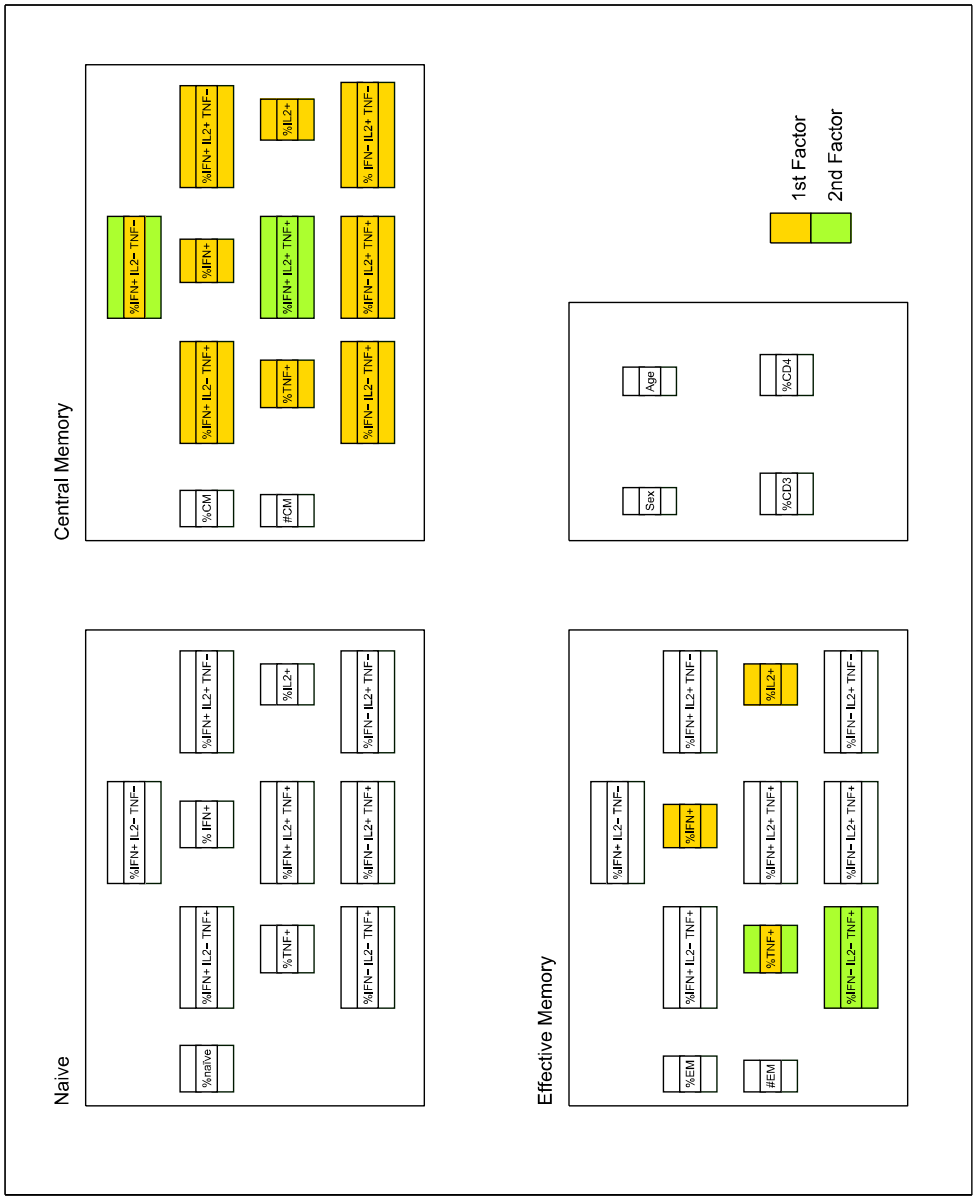


Figure 24: Factor groups for unstimulated cells of Rhesus monkeys with 2 factors and threshold 0.5 in the **CD4 boolean**-dataset

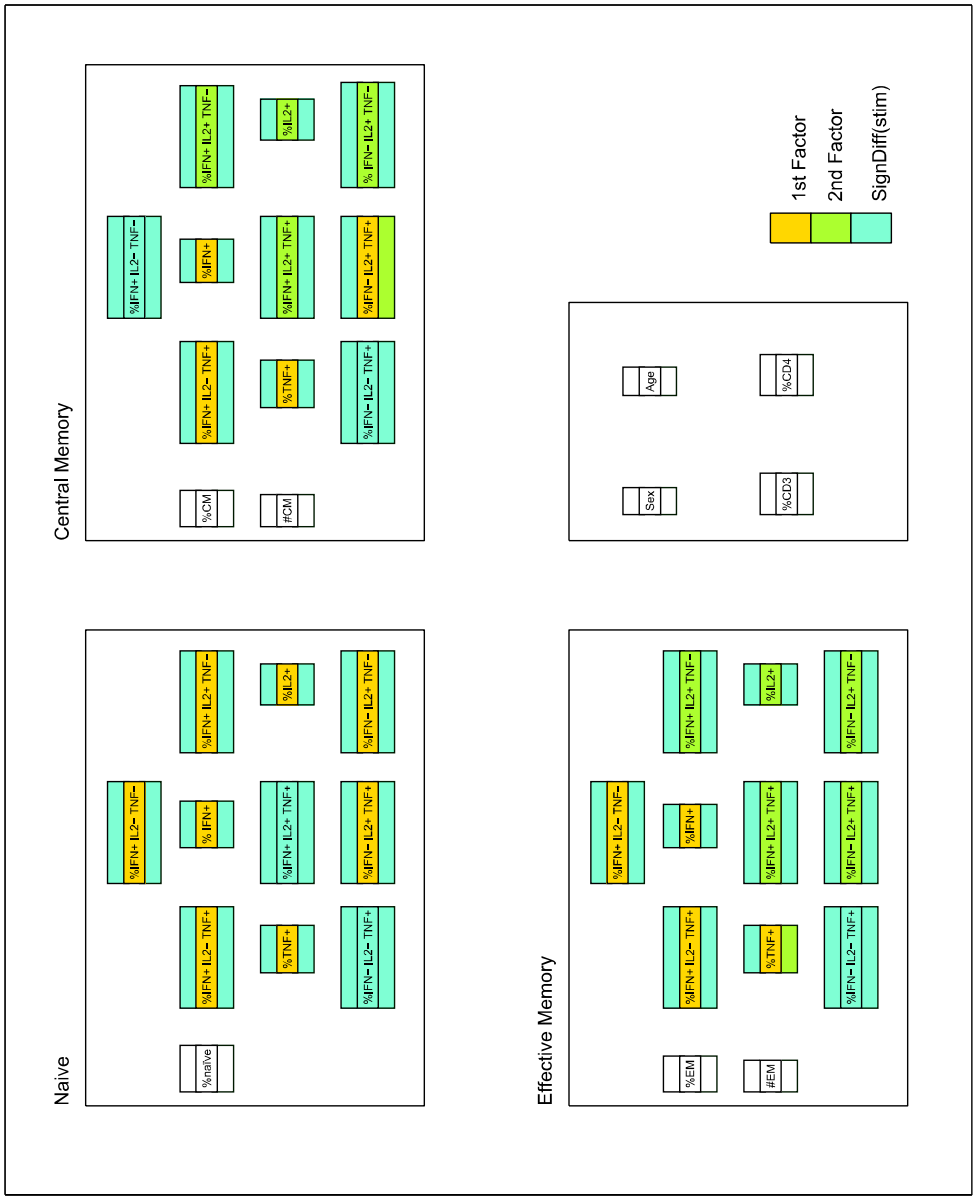


Figure 25: Factor groups for significant different variables of stimulated and unstimulated cells of Rhesus monkeys with 2 factors and threshold 0.5 in the **CD4 boolean**-dataset

Figure 26: Factor groups for significant different variables of stimulated and unstimulated cells of African green monkeys with 2 factors and threshold 0.5 in the **CD4 boolean**-dataset

CD8 boolean

The variables given in the dataset '*RM_AGM_PMA+I_CD8_Boolean* are

%CD8 CM IFN+ Proportion of the CD8+ IFN+ cells within the central memory

%CD8 CM IL2+ Proportion of the CD8+ IL2+ cells within the central memory

%CD8 CM TNF+ Proportion of the CD8+ TNF+ cells within the central memory

%CD8 CM IFN+ IL2+ TNF+ Proportion of IFN+ IL2+ TNF+ cells within the central memory

%CD8 CM IFN+ IL2+ TNF- Proportion of IFN+ IL2+ TNF- cells within the central memory

%CD8 CM IFN+ IL2- TNF+ Proportion of IFN+ IL2- TNF+ cells within the central memory

%CD8 CM IFN+ IL2- TNF- Proportion of IFN+ IL2- TNF- cells within the central memory

%CD8 CM IFN- IL2+ TNF+ Proportion of IFN- IL2+ TNF+ cells within the central memory

%CD8 CM IFN- IL2+ TNF- Proportion of IFN- IL2+ TNF- cells within the central memory

%CD8 CM IFN- IL2- TNF+ Proportion of IFN- IL2- TNF+ cells within the central memory

%CD8 EM IFN+ Proportion of the CD8+ IFN+ cells within the effective memory

%CD8 EM IL2+ Proportion of the CD8+ IL2+ cells within the effective memory

%CD8 EM TNF+ Proportion of the CD8+ TNF+ cells within the effective memory

%CD8 EM IFN+ IL2+ TNF+ Proportion of IFN+ IL2+ TNF+ cells within the effective memory

%CD8 EM IFN+ IL2+ TNF- Proportion of IFN+ IL2+ TNF- cells within the effective memory

%CD8 EM IFN+ IL2- TNF+ Proportion of IFN+ IL2- TNF+ cells within the effective memory

%CD8 EM IFN+ IL2- TNF- Proportion of IFN+ IL2- TNF- cells within the effective memory

%CD8 EM IFN- IL2+ TNF+ Proportion of IFN- IL2+ TNF+ cells within the effective memory

%CD8 EM IFN- IL2+ TNF- Proportion of IFN- IL2+ TNF- cells within the effective memory

%CD8 EM IFN- IL2- TNF+ Proportion of IFN- IL2- TNF+ cells within the effective memory

%CD8 Naïve IFN+ Proportion of the CD8+ IFN+ naïve cells (among the naïve cells)

%CD8 Naïve IL2+ Proportion of the CD8+ IL2+ naïve cells (among the naïve cells)

%CD8 Naïve TNF+ Proportion of the CD8+ TNF+ naïve cells (among the naïve cells)

%CD8 Naïve IFN+ IL2+ TNF+ Proportion of the CD8+ IFN+ IL2+ TNF+ naïve cells (among the naïve cells)

%CD8 Naïve IFN+ IL2+ TNF- Proportion of the CD8+ IFN+ IL2+ TNF+ naïve cells (among the naïve cells)

%CD8 Naïve IFN+ IL2- TNF+ Proportion of the CD8+ IFN+ IL2- TNF+ naïve cells (among the naïve cells)

%CD8 Naïve IFN+ IL2- TNF- Proportion of the CD8+ IFN+ IL2- TNF- naïve cells (among the naïve cells)

%CD8 Naïve IFN- IL2+ TNF+ Proportion of the CD8+ IFN- IL2+ TNF+ naïve cells (among the naïve cells)

%CD8 Naïve IFN- IL2+ TNF- Proportion of the CD8+ IFN- IL2+ TNF- naïve cells (among the naïve cells)

%CD8 Naïve IFN- IL2- TNF+ Proportion of the CD8+ IFN- IL2- TNF+ naïve cells (among the naïve cells)

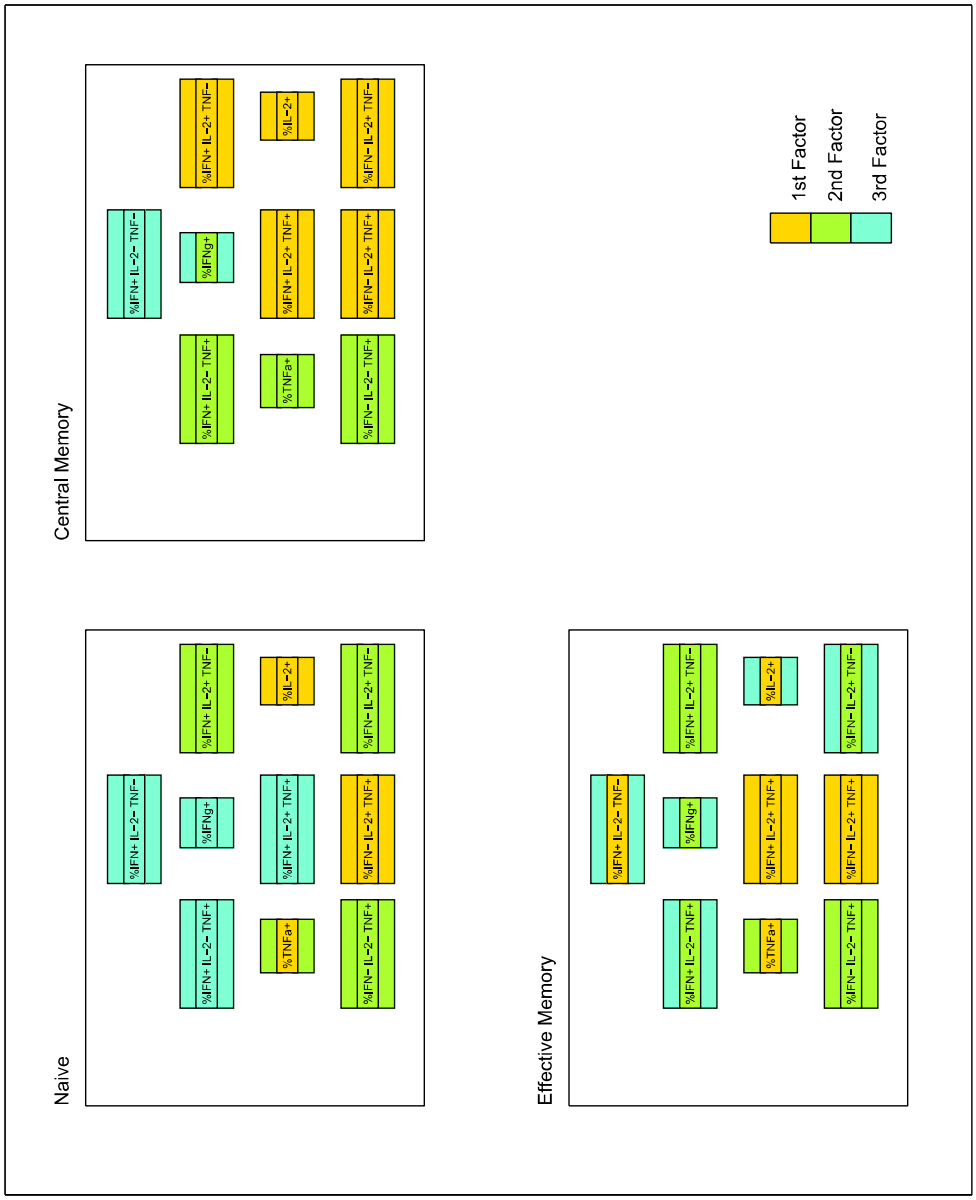


Figure 27: Factor groups for stimulated cells of African green monkeys with 3 factors and threshold 0.5 in the **CD8 boolean**-dataset

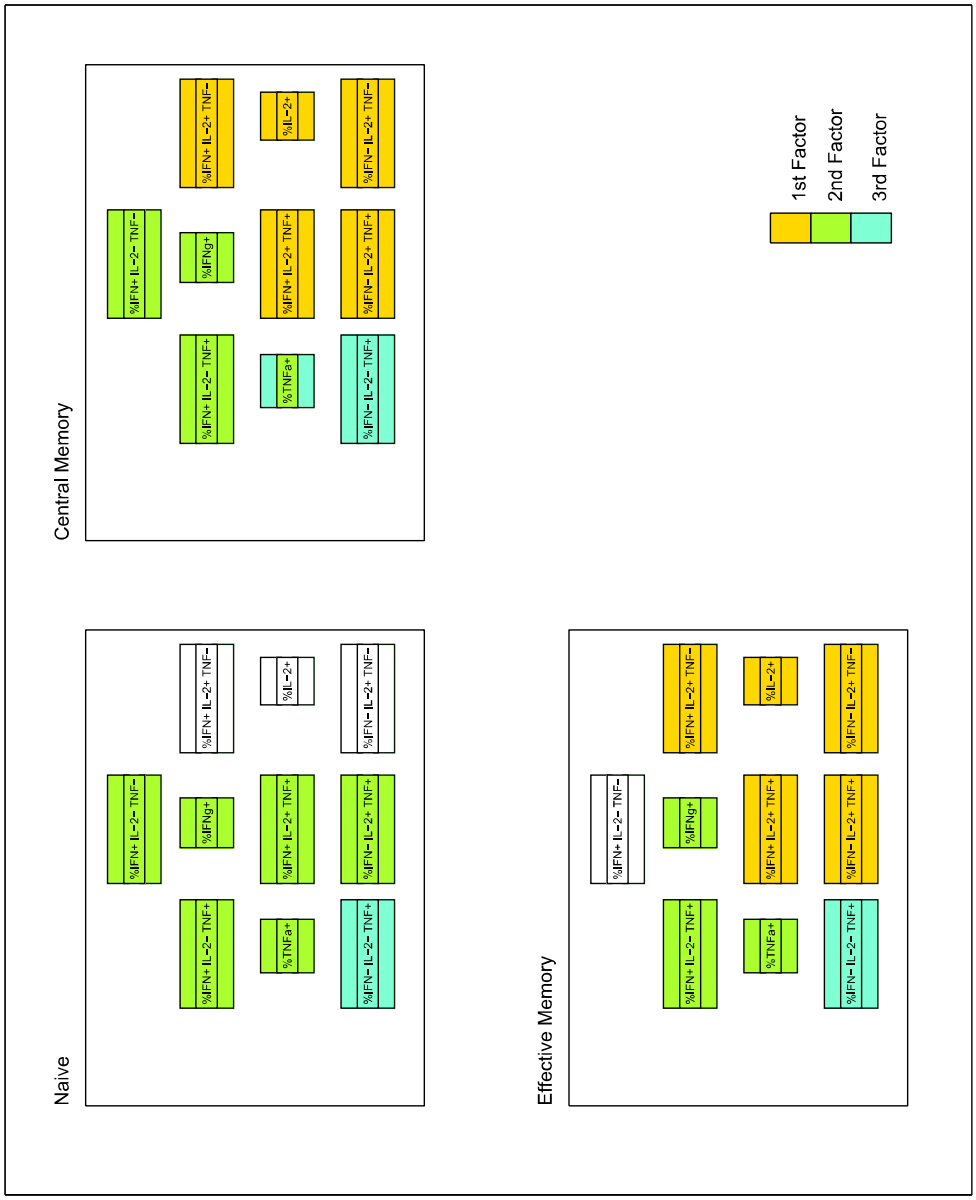


Figure 28: Factor groups for stimulated cells of Rhesus monkeys with 3 factors and threshold 0.5 in the **CD8 boolean**-dataset

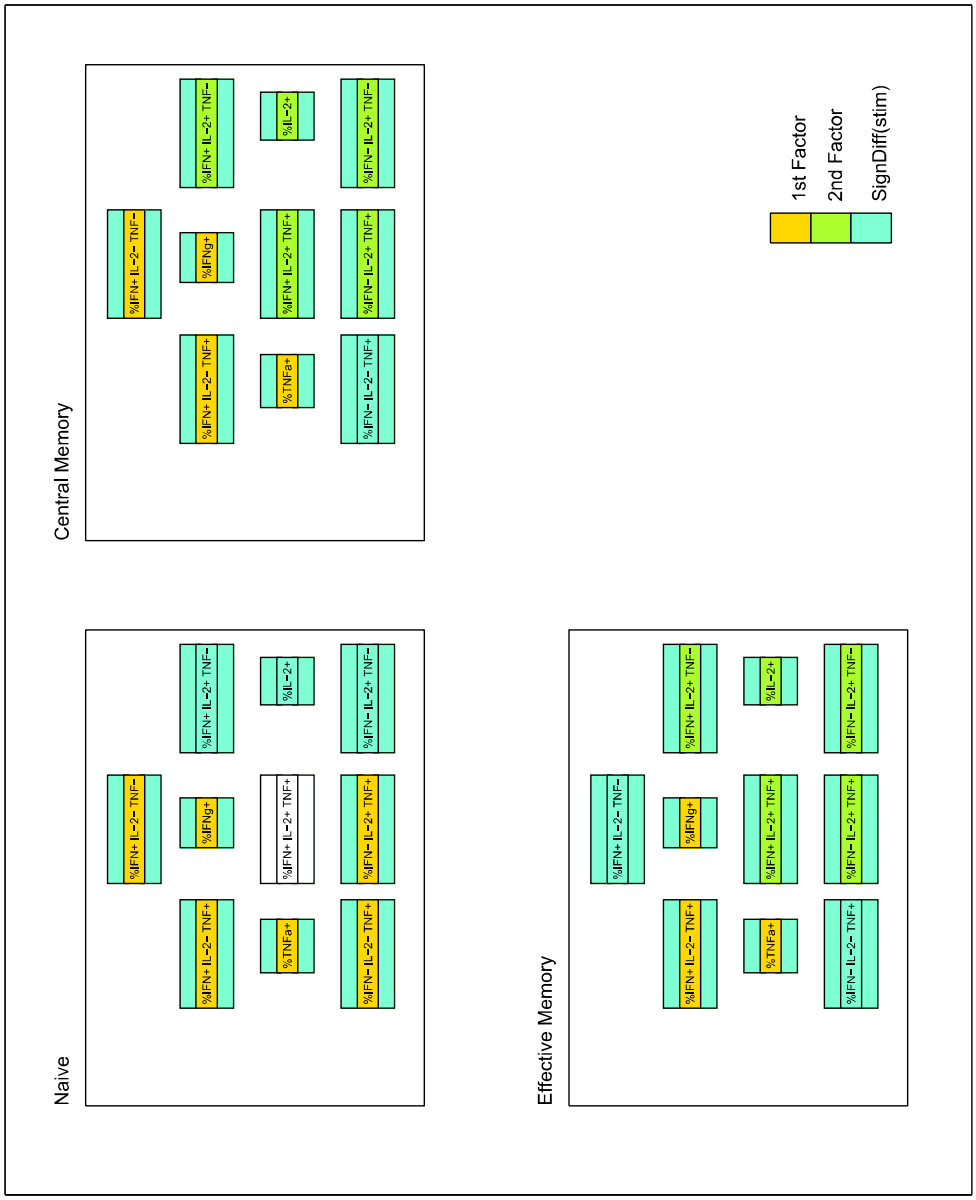


Figure 29: Factor groups for significant different variables of stimulated and unstimulated cells of Rhesus monkeys with 2 factors and threshold 0.5 in the **CD8 boolean**-dataset

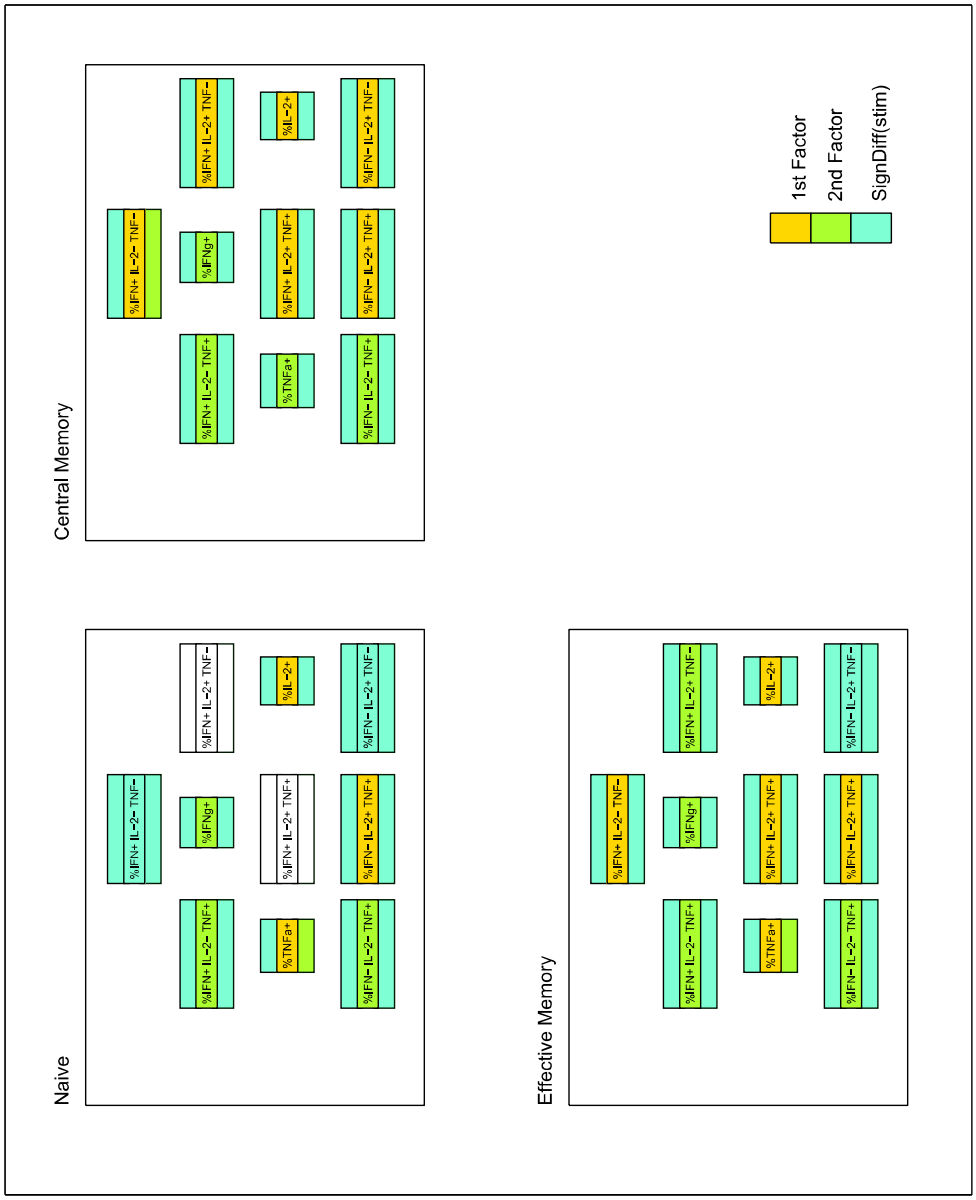


Figure 30: Factor groups for significant different variables of stimulated and unstimulated cells of African green monkeys with 2 factors and threshold 0.5 in the **CD8 boolean**-dataset

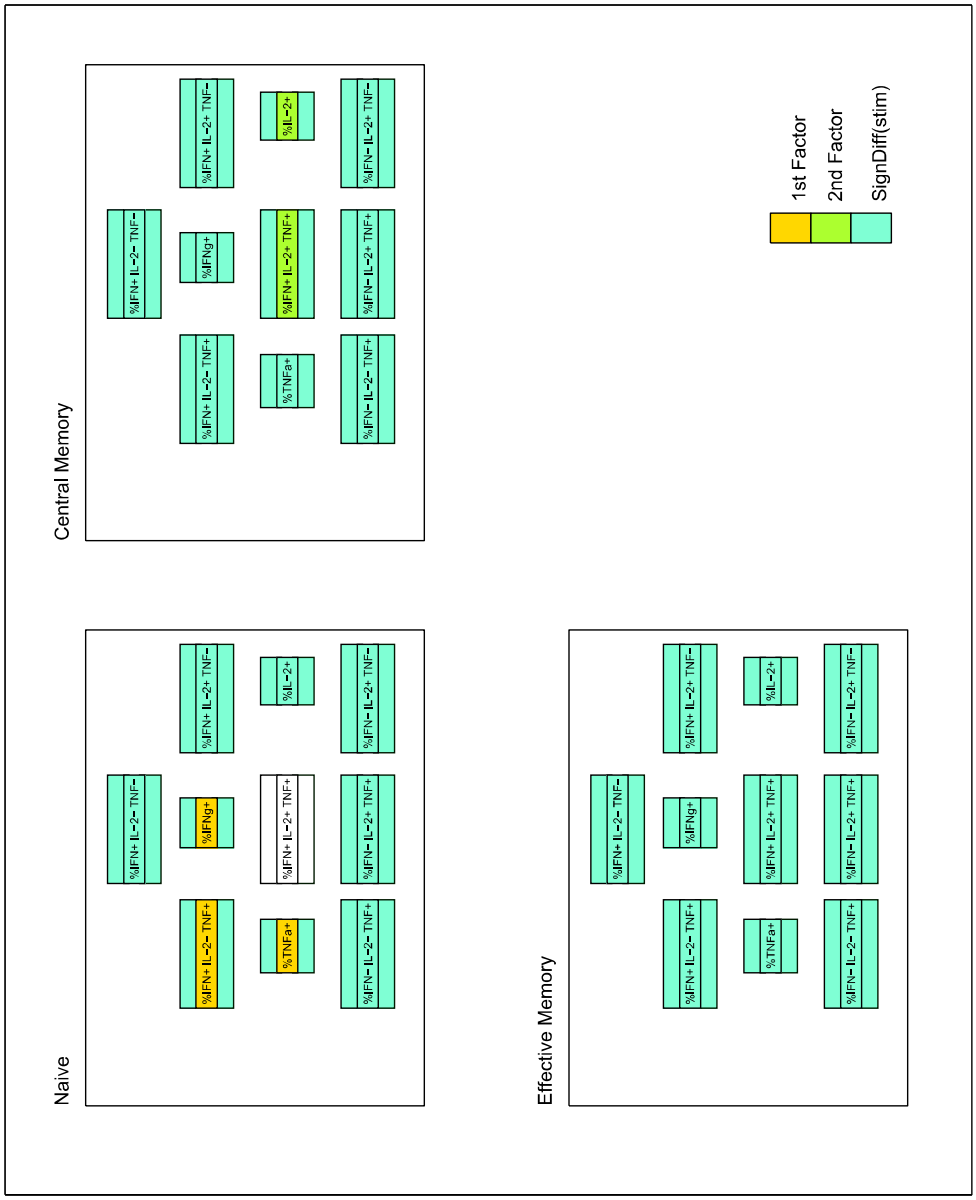


Figure 31: Factor groups for significant different variables of stimulated and unstimulated cells of Rhesus monkeys with 2 factors and threshold 0.95 in the **CD8 boolean**-dataset

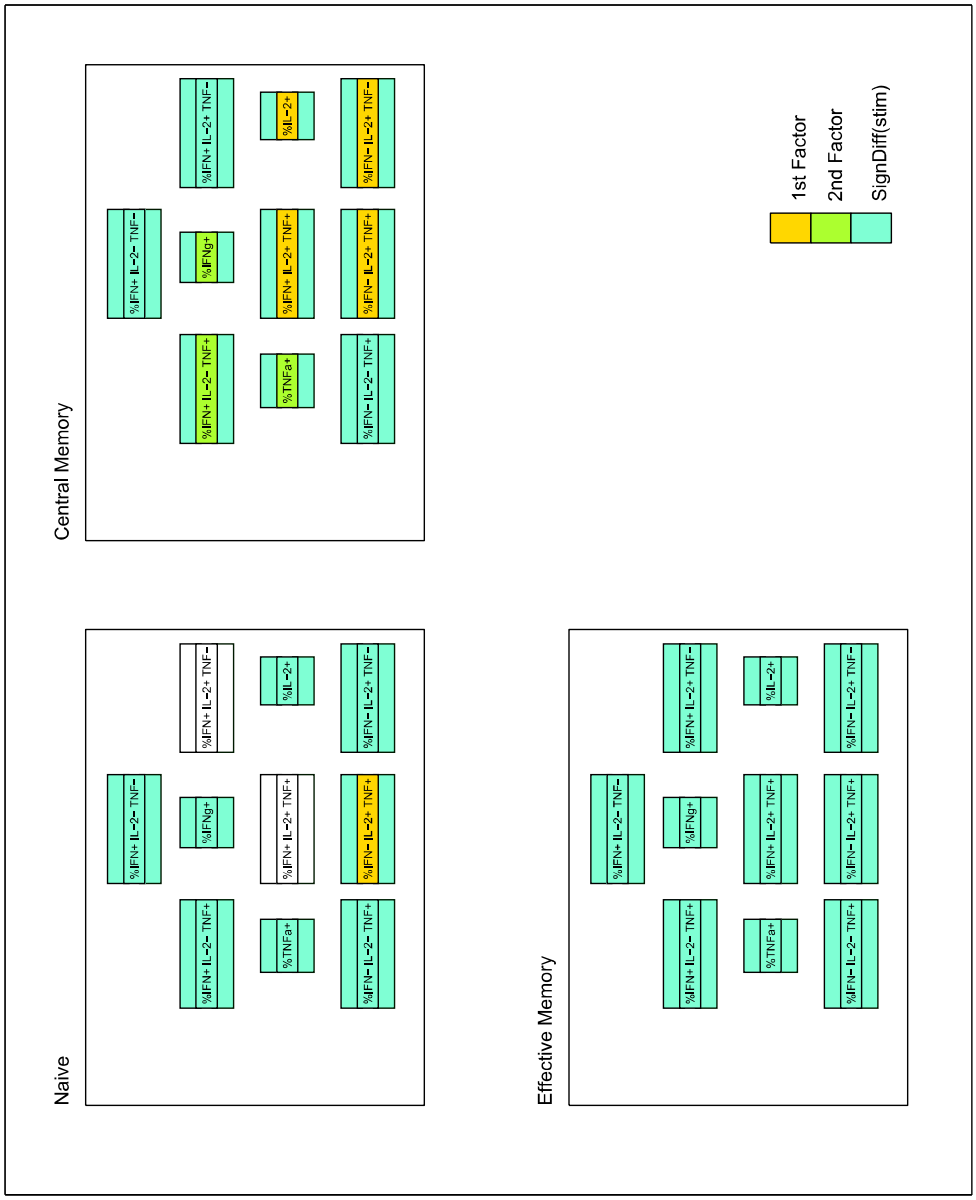


Figure 32: Factor groups for significant different variables of stimulated and unstimulated cells of African green monkeys with 2 factors and threshold 0.95 in the **CD8 boolean**-dataset