universität wien

# MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Analysis of Gene Expression Data for Precision Medicine in Ulcerative Colitis“

verfasst von / submitted by

Marlene Steiner, BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2023 / Vienna 2023

| | |
|---|---|
| Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet: | A 066 645 |
| Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet: | Masterstudium Data Science |
| Betreut von / Supervisor: | Univ.-Prof. Mag. Dr. Thomas Rattei |

# Danksagung

Zuerst möchte ich mich von Herzen bei meinen Eltern bedanken, die mir dieses Studium ermöglicht haben. Ihr habt mir alle Türen offengehalten und meine Entscheidungen unterstützt. Ihr habt mit mir mitgefiebert, mitgefühlt und falls es mal nötig war, mich ermutigt an mich zu glauben. Danke!

Einen besonderen Dank gilt an Klemens Vierlinger vom Austrian Institute of Technology (AIT), der mich bei diesem Projekt begleitet hat. Seine Expertise und der gemeinsame Austausch waren sehr wertvoll und haben die Analyse geprägt. Ich möchte mich außerdem bei Thomas Rattei für die Betreuung dieser Masterarbeit bedanken. Seine konstruktiven Kommentare haben mir geholfen die Arbeit zu verbessern.

Besonders möchte ich mich bei meinem Freund Christoph bedanken. Er hat mich nicht nur emotional unterstützt, sondern war immer für mich da, um mir in schwierigen Zeiten den Rücken zu stärken und gute Zeiten gemeinsam zu zelebrieren. Seine Unterstützung und Ermutigungen haben mir geholfen, große Herausforderungen zu meistern und diese Arbeit abzuschließen.

Ich bin außerdem dankbar für die einzigartige Gemeinschaft im Studium und die Freundschaften die daraus entstanden sind. Danke an Pavla, Elisabeth und Lena, die mit mir alle Höhen und Tiefen des Masters geteilt haben. Zusammen haben wir als erster Jahrgang das Data Science Masterprogramm gemeistert. Danke außerdem an Nina, Maggie, Lini und vielen mehr, die mich auf dem Weg unterstützt haben.

# Abstract

Precision medicine seeks to incorporate factors other than the disease diagnosis into treatment decision-making. With the development of new, increasingly rapid and cost-effective RNA and DNA sequencing methods, there are completely new possibilities in the field of precision medicine. The principle is to use genetic markers in treatment decision making to address the genetic predisposition of the patient. This requires research on genetic markers that are associated with treatment response. In this thesis, biomarkers for treatment response for different treatments in ulcerative colitis patients are searched for using a cross-study data analysis. For the past two decades, the development of new antibody-based drugs from the group of biologics has transformed treatment of ulcerative colitis. Some drugs have been better studied while others are still in their infancy. In this work, all publicly available gene expression data that include the response to biologics in ulcerative colitis patients are analysed. The analysis includes eight different datasets and five different drugs and placebo. First, for each drug, differentially expressed genes at baseline between responders and non-responders are analysed. Then, machine learning models are used to find and evaluate predictive genes for treatment response. Lastly, the change in gene expression over the course of treatment between responders and non-responders is analysed. These analyses provide a broad overview of the response of different therapies in Ulcerative Colitis at a gene expression level and provides results that can be further processed upon multiple levels to better understand and treat this complex disease.

# Kurzfassung

Personalisierte Medizin versucht abgesehen von der Krankheitsdiagnose andere Faktoren in die Entscheidungsfindung der Behandlung einzubeziehen. Mit der Entwicklung von neuen, zunehmend schnellen und kostengünstigen RNA und DNA-Sequenzierungsmethoden, gibt es völlig neue Möglichkeiten im Bereich der personalisierten Medizin. Das Prinzip ist, genetische Marker bei der Behandlungsentscheidung zu nutzen und so auf die genetischen Gegebenheiten der Patient*innen einzugehen. Dazu müssen Biomarker erforscht werden, die mit dem Ansprechen auf eine Behandlung in Verbindung stehen. In dieser Arbeit wird mithilfe einer studienübergreifenden Datenanalyse nach Biomarkern für oder gegen das Ansprechen auf verschiedene Medikamente bei Colitis Ulcerosa Patient*innen gesucht. Seit ungefähr zwanzig Jahren hat die Entwicklung von neuen, antikörper-basierte Medikamenten - aus der Gruppe der Biologika - die Behandlung von Colitis Ulcerosa verändert. Manche Medikamente wurden bereits besser erforscht, während andere noch am Anfang stehen. In dieser Arbeit werden alle öffentlich verfügbaren Genexpressionsdaten, die das Ansprechen auf Biologiker in Colitis Ulcerosa Patient*innen beinhalten analysiert. Die Analyse umfasst acht verschiedenen Datensätze und sechs verschiedene Medikamente. Zuerst werden für jedes Medikament differentiell exprimierte Gene vor der Behandlung zwischen Responder und Nicht-Responder getestet. Weiters werden Machine Learning Modelle verwendet, um prädiktive Gene für oder gegen das Ansprechen auf Medikamente zu finden und zu evaluieren. Zuletzt wird die Veränderung von Genexpression im Laufe der Behandlung zwischen Responder und Nicht-Responder analysiert. Diese Analysen verschaffen einen breit gefächerten Überblick über das Ansprechen verschiedener Therapien in Ulcerative Colitis auf Genexpressionsebene und liefert Ergebnisse, die auf mehreren Ebenen weiter prozessiert werden können, um diese komplexe Krankheit besser zu verstehen und behandeln zu können.

# Contents

# List of Tables

# List of Figures

# Acronyms

**CD** Crohn's Disease.

**CEL** Cell intensitiy file.

**cmp** counts per million.

**DE** Differential Expression.

**IBD** Inflammatory Bowel Disease.

**MM** Mismatch.

**NCBI** National Center for Biotechnology Information.

**NGS** Next Generation Sequencing.

**PCR** Polymerase Chain Reaction.

**PM** Perfect Match.

**PVCA** Principle Variance Componens Analysis.

**rma** robust multiarray average.

**rrCV** Repeated-repeated cross validation.

**UC** Ulcerative Colitis.

# Chapter 1

# Introduction

Patients suffering from the same disease, treated under the same conditions, may respond differently to the same treatment. This motivates the application of precision medicine, also known as personalised medicine, stratified medicine or individualized medicine. Instead of treating every patient with the same disease with the same treatment, the treatment with the best outcome for each individual patient should be chosen. Therefore, treatment response predictors have to be found. These predictors are then used to support clinicians in their decision making [44].

The concept of precision medicine is not new. Clustering patients by some visible factors and treating them according to the experienced best outcome for this group has been practiced for at least 2000 years [44]. However, the ability to collect and analyse molecular data opened up new possibilities for precision medicine. It allows group comparisons of patients with different treatment responses on a molecular level. To find significant response predictors, sample sizes of these groups need to be large enough. The development of high throughput methods such as microarrays or next generation sequencing methods enabled collecting large amounts of genetic data fast and relatively cost efficient [19].

Genetic precision medicine already achieved successes in treatment of cancer, the field is known as precision oncology. There, the main idea is to analyse the genetics of the cancer cells, identify mutations and select treatments that best tackle this specific type of cancer [44].

Ulcerative colitis is an immune mediated chronic disease. The course of disease is complex and medical treatment is not always effective. About 25% of ulcerative colitis patients need colectomy [7], which results in long term quality of life reduction [16]. In the past two decades, new treatment options based on target-specific monoclonal antibodies have been introduced. Multiple studies collected gene expres-

sion data in respect to treatment of these biologics. In this thesis, publicly available treatment response data of eight different studies was analysed. This includes six different treatments (Infliximab, Golimumab, Adalimumab, Etrolizumab, Vedolizumab and Placebo), two different data generation methods (microarrays and RNA-Seq) and two different timepoints (baseline, follow-up).

The main questions investigated in this thesis are:

1. Which genes are associated with the response to antibody based medication in ulcerative colitis patients? More precisely: Which genes are differently expressed between responders and non-responders before treatment?

2. How well can response to certain treatments be predicted prior to treatment and which set of genes are relevant for prediction?

3. How does gene expression change during therapy in responders and in non-responders? Which genes change significantly?

All datasets are analysed from the raw data coming from the lab, even when already preprocessed data was available. There are numerous options to preprocess microarray and RNA-Seq data and in this cross study analysis, all datasets are preprocessed in the same way. Combining multiple datasets has the benefit of more samples and therefore more robust results and more statistical power. But this comes with a cost: Merging multiple datasets from different microarray or RNA-Seq generations comes with the price of only being able to analyse genes that are present in all datasets and study wise batch effects which need to be corrected. Sample sizes in this analysis range from 19 to 143 samples and more than $15,000$ genes to be analysed. Gene expression analysis that start with this sample-feature ratio are usually hypothesis-generating experiments. They report effects that are most prominent in the dataset with the expectation of including some false positives that are dataset specific effects but not ground truth. With the set of genes resulting from these hypothesis-generating analysis, further studies can be conducted to evaluate genes found. Therefore, reporting more false positive results is not as fatal as missing true positives due to significance thresholds that are too strict.

When analysing gene expression data with the aim to compare different groups, there are typically two approaches. The univariate differential expression approach where for each gene one hypothesis test is performed. And the machine learning approach, where models are given a set of features and should find the most important predictors to distinguish groups themselves.

This thesis is divided into 6 chapters. After the introduction 1, a brief overview of Ulcerative Colitis and gene expression data is given in a theoretical background chapter 2. Where the data was searched, which studies were considered and which were finally included in the analysis is described in the data chapter 3. Additionally this chapter contains comparisons of all datasets, to find out if the studies included are comparable. In the methods section 4 the preprocessing of microarray as well as RNA-Seq data is described. Moreover, it includes the description of methods used for differential expression and prediction. Finally the software used for the analysis is listed. In the results and discussion chapter 5 each subsection is dedicated to one research question. Finally, the thesis aims and findings were concluded in chapter 6. In the appendix 6 supplement figures, toptables of the differential expression analysis and R software details are reported.

## 1.1 Related Work

A lot of effort already went into identifying the gene expression signature of UC by comparing tissue from healthy and affected individuals [57, 88, 50, 12]. This was also summarised in multiple meta analysis [49, 34].

For analysing gene expression data of treatment response in UC however, this is the largest cross-study analysis to my knowledge. In this thesis

- 8 different datasets
- 5 different treatments + placebo
- 2 different data generation methods
- and 2 different timepoints

are analysed. Sakaram et al. analysed treatment response of four Infliximab studies (EMTAB7604, GSE14580, GSE23597, GSE16879) [70]. All data sets they used are also considered in this thesis. They neither use multiple treatments nor multiple timepoints. Another difference is that they use gene expressions of Ulcerative Colitis and Crohn's disease patients together without separation. This approach is quite questionable, since it has been found that the two disease clearly differ on gene expression level as shown by Lawrence et al. [48].

# Chapter 2

# Theoretical Background

## 2.1 Ulcerative Colitis (UC)

Ulcerative Colitis (UC) manifests primarily as a chronic inflammation of the colon - the lower part of the digestive tract. UC and Crohn's Disease (CD) together form the group of Inflammatory Bowel Disease (IBD). UC differentiates from CD in several aspects. While the inflammation in UC is restricted to the colon, CD may affect the entire gastrointestinal tract [16]. The inflammation in UC usually starts at the rectum and continuously extends up to the entire colon. This causes painful diarrhoea and bleeding from the bowel [23]. UC is characterised by alternating periods of flares and remission. There are different treatment options, which reduce symptoms significantly, but so far no medically induced cure for the disease has been found. Life expectancy of UC patients is only minimally reduced, nevertheless the disease influences the quality of life enormously [23, 81].

UC is not a rare disease, in Europe are about 2 million people affected [11]. The disease occurs worldwide, however, prevalences are higher in more industrialised regions [23]. Figure 2.2 shows the results of a meta analysis which summarised available prevalence data between 1990-2016. The highest prevalence worldwide was reported in Norway with 505 cases per $100,000$ inhabitants. In Germany, the prevalence of UC reaches 322 per $100,000$. [56] [55]

The distribution of the age at the time of diagnosis does not show a clear peak. UC may occur at any age with the highest probability between the age of 20 and 50. Furthermore, UC affects men and women equally often. [23]

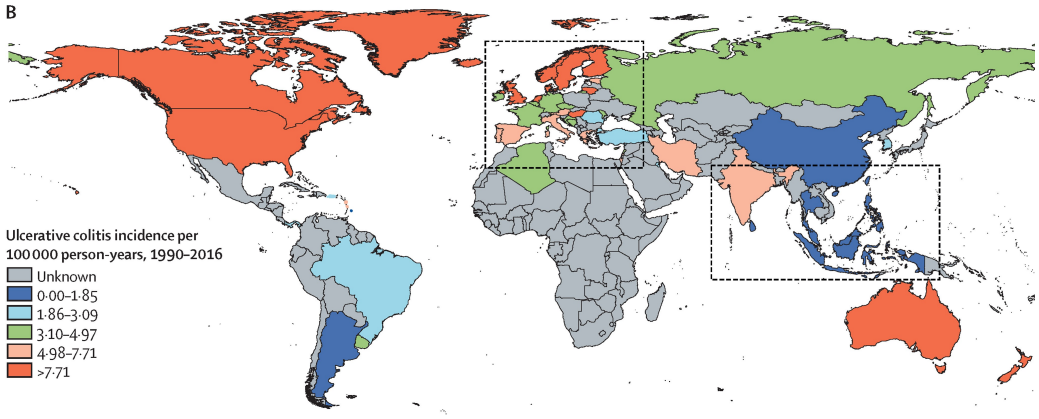



Figure 2.1: *Incidence of Ulcerative Colitis. [56]*

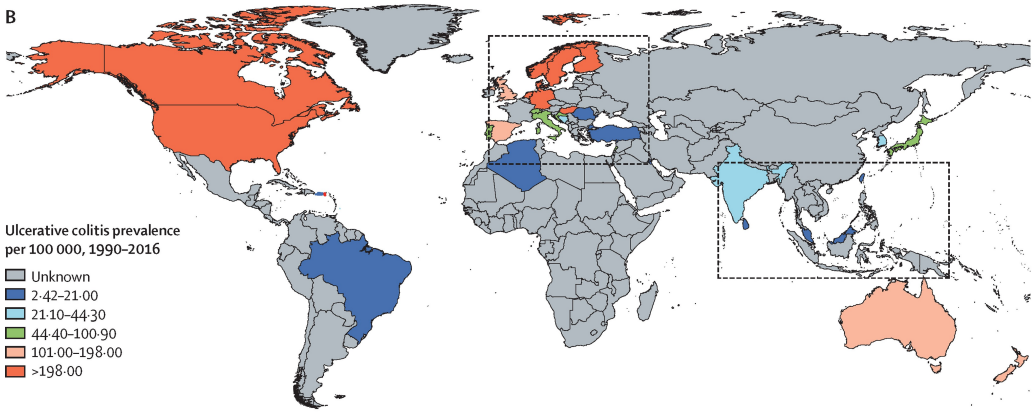



Figure 2.2: *Prevalence of Ulcerative Colitis. [56]*

### 2.1.1 Origin of UC

In UC symptoms are caused by the inflammation of the colon, which is caused by a disturbed immune system. It is assumed that the immune reaction of the mucosa in the colon to the microflora is inappropriate [53]. What causes this inappropriate reaction is not completely understood yet [23]. Most probably it is a combination of multiple risk factors. Genetic as well as environmental factors such as lifestyle and nutrition have been analysed [16, 23].

The study of identical twins shows that it is only to a little extend genetically determined. Given that one twin is diagnosed with UC, the probability that the other

identical twin is also affected by the disease is $6,3\%$. In comparison, this probability for CD is 67% [16].

The map of incidences (Figure 2.1) and prevalences (Figure 2.2) of UC around the world shows high incidence and prevalence rates in Europe, North America, Australia and New Zealand. Due to the regional differences the first assumption was a relationship between UC and the ethnic background. This assumption was analysed in migration studies. Based on multiple migration studies, Barreiro-de Acosta et al. concluded that people who migrated from a developing country to a country with higher prevalence of UC (e.g. western Europe), had an increasing risk of UC [51]. This indicates that not genetics, but more likely environmental factors are main risk factors for UC. The regions with high prevalences of UC have industrialization and high socioeconomic status in common. The lifestyle (e.g. eating habits, hygiene standards) and medical standards (e.g. Antibiotics) are risk factors of UC [23, 16]. Paradoxically UC is one of very few disease where smoking has a protective effect [23].

### 2.1.2 Symptoms, Course of disease

Symptoms such as abdominal pain, diarrhoea and bleeding from the bowel are caused by the inflammation of the colon. Furthermore, sometimes extraintestinal symptoms such as joint pain, eye infection or skin irritation occur [16]. UC is characterised by alternating flares and remission. Patients experience enormous fluctuations of symptoms within the course of disease. One flare-up can last several weeks to months. The remission phase, where the patient does not have any symptoms may last from months to years [20, 23].

The intensity of symptoms is graded in scores. The most commonly used score is the Mayo score, which includes various aspects of UC and takes values from 0 to 12, where 12 is the most severe degree. The mayo score consists of four subscores: rectal bleeding subscore, stool frequency subscore, physician assessment and endoscopy subscore. Each subscore is ranked between 0 and 3 and the sum of all subscores yields the mayo score[71]. A mayo score of 3-5 point is classified as severely active UC, a score of 6-10 as moderately active disease and a score of 11-12 as severely active disease. Many symptoms that occur during a flare are not specific to UC, hence the process of diagnosis can be long. Differential diagnosis have to be ruled out before with certainty diagnosing UC [16]. A cohort of 2100 UC patients reported an average diagnostic delay of two years with a standard deviation of 5.4 years [20].

#### Quality of life

Ulcerative colitis has an enormous impact on the quality of life [32, 20]. A study [20] surveying 2100 UC patients across 10 countries (including USA and Germany) reports that 65% of the patients felt like "UC controlled their life rather than them controlling their disease". 95% of the patients reported that they are taking prescription medication at the moment, including 44% of patients taking biologics. Patients have to use the bathroom on a good day on average 3.5 times and on a bad day on average 9.8 times. These symptoms affect not only their wellbeing but also their work. 74% of patients are employed and they missed on average 8.1 working days with a standard deviation of 16.7 days the past year. The majority of patients are mentally exhausted, despite current treatment.

### 2.1.3 Treatment

Since the etiology of UC is not fully understood, there are no causal treatment options. However there are symptomatic treatment options that manipulate the inflammatory processes at various levels [53]. During a flare, the goal is to induce remission and subsequently to maintain remission [15]. The main medical treatments for UC are

- 5-Aminosalicylic acid (5-ASA, Mesalazin)
- Corticosteroids
- Immunesuppressants
- Biologics

Moreover there are surgical treatment options. A complete removal of the colon (colectomy) may cure UC, since UC is restricted to the colon. However, the quality of life after a colectomy can be compared to the quality of life with an intermediate active UC [16]. Therefore, this procedure is only done if other treatment attempts failed and there is no prospect of improvement with medication only or if other complications occur [15, 64].

#### 5-Aminosalicylic acid

5-Aminosalicylic acid (5-ASA, Mesalazin) has an anti-inflammatory effect and has been discovered effective in mild to intermediate active UC [53].

Pure Mesalazin would be absorbed in the Jejenum. The desired location is however the colon, hence Mesalazin is either chemically modified to operate in the colon or is surrounded by a pH resistant shell, which opens up at the colonic pH level [53].

Compared to other medical treatments, 5-ASA tend to have the least side effects and are therefore often the first choice for treating UC [53].

### Corticosteroids

Corticosteroids are widely used due to their anti-inflammatory effect [53]. Also in IBD, corticosteroids are established treatment options. Since the introduction of corticosteroids as treatment for UC, the mortality during a severe flare-up has been reduced tremendously from 30% to 2% [23]. Due to many and severe side effects, treating UC with corticosteroids needs to be done cautiously. They should only be used if the patient did not respond to 5-ASA and in moderate to severe cases of UC [15]. Long term intake leads to severe side effects such as osteoporosis, Cushing-Syndrom, higher risk of susceptibility to infection, growth inhibiting during growth phase - just to name a few [53].

### Biologics

Biologics most recently changed treating UC. Treatments such as Infliximab, Adalimumab, Golimumab and Vedolizumab received market approvals for UC. [10].

Biologics differentiate from other medication in their production. Biologics are produced in living organisms such as bacteria. Complementary DNA of the desired protein is inserted into the organism which transcribes the DNA and translates it to proteins. These proteins are then harvested [53]. For IBD, biologics affecting the immune system, specifically antibody-based treatments have been found effective. Biologics have changed the treatment of inflammatory bowel disease in the past 20 years. Infliximab was the first biologic approved for treating CD and UC. Many other biologics have been analysed since and some of them have been found useful for treating UC. The nomenclature of these monoclonal antibodies is systematic [53]. The suffix indicates the type of antibody. The treatment ends with

- -ximab if it is a chimeric monoclonal antibody
- -zumab if it is a humanized monoclonal antibody
- -mumab if it is a human monoclonal antibody

Furthermore, if the two letters in front of the suffix are "li", treatment acts immunmodulatory [53]. This is the case for all biological treatments discussed in this context. All treatments and their categorization are listed in table 2.1.

Infliximab, Golimumab and Adalimumab are monoclonal antibodies binding to Tumornekrosisfactor$\alpha$ (TNF$\alpha$). TNF$\alpha$ is a cytokine that acts as messenger in local and systemic inflammation processes. It is produced by activated monocytes and macrophages in the human body[53]. Infliximab, Golimumab and Adalimumab all target TNF$\alpha$, however, there are differences in their pharmacokinetics, their production[24, 25, 26] and in binding sites at the TNF$\alpha$ complex[58]. Infliximab is injected intravenously, while Adalimumab and Golimumab are injected subcutaneously. Therefore the bioavailability differs in treatments. For the commercial production of these biologics, different hybridoma cells are used. Infliximab and Golimumab are produced in murine hybridoma cells by recombinant DNA technology[25, 26]. Adalimumab is produced in cell cultures of chinese hamster cells[24]. Despite those differences, all three antibodies bind to the TNF$\alpha$ complex, thereby inactivate the cytokine and suppress the inflammation.

Etrolizumab and Vedolizumab are humanized monoclonal antibodies which bind to $\beta$7 integrins. Specifically, Etrolizumab binds to $\alpha 4\beta 7$ and $\alpha \mathrm{E}\beta 7$ integrins and Vedolizumab only binds to the $\alpha 4\beta 7$ integrin. Integrins consisting of an $\alpha$ and a $\beta$ subunit, act as messenger between immune cells and the vascular endothelium [53].

| **Treatment** | **binding to** | **antibody type** |
|---|---|---|
| Infliximab | TNF$\alpha$ | chimeric monoclonal antibody |
| Golimumab | TNF$\alpha$ | human monoclonal antibody |
| Adalimumab | TNF$\alpha$ | human monoclonal antibody |
| Etrolizumab | $\alpha 4\beta 7$ and $\alpha \mathrm{E}\beta 7$ integrin | humanized monoclonal antibody |
| Vedolizumab | $\alpha 4\beta 7$ integrin | humanized monoclonal antibody |

Table 2.1: *Biologics analysed in this thesis*

## 2.2 Gene Expression Data

The process of gene expression is also known as the central dogma of molecular biology. Genetic code stored in the DNA is used to build functional proteins, the building blocks of all living organisms. This process is complex and many details have been investigated. The main concept however is that DNA is transcribed to messenger RNA (mRNA) inside the cells nucleus. Then, mRNA carries all necessary information outside the nucleus to the ribosomes, where mRNA is translated to proteins. This process can be analysed on different levels. The analysis of the genome - the code of DNA of an organism - is known as Genomics. The analysis of the transcriptome - the set of active mRNAs in a given tissue of a given timepoint - has been established as transcriptomics and accordingly the analysis of the proteome - the set of proteins expressed - is called proteomics. Figure 2.3 illustrates this relationship. [19, 75]

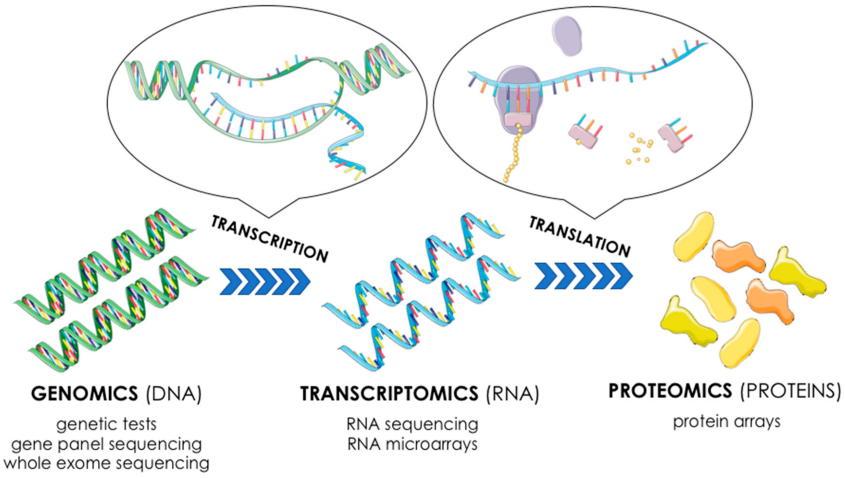



Figure 2.3: *Omics principle [75]*

Which level should be chosen for the analysis depends on the use case. For the analysis of treatment response in UC only transcriptomic data is available. Because gene expression is a complex process including e.g. gene regulation and DNA methylation, the analysis of the transcriptome - a later stage of the process, which represents the phenotype rather than the genotype - is often preferred for precision medicine. There are two well established methods for transcriptomics, RNA microarray and RNA-Seq. Microarrays revolutionised the analysis of gene expression in the early 2000s. It was for the first time possible to look at a whole landscape of ex-

pressions without a specific target [19]. RNA sequencing is the next generation and even more powerful than microarrays [38]. Both methods are still in use, however RNA-Seq overtook Microarrays recently.

In this thesis, data from Affymetrix microarrays as well as Illumina RNA-Seq are analysed. Both methods analyse the transcriptome of a biologic sample, e.g. a tissue or blood sample.

### 2.2.1 Affymetrix Microarrays

Affymetrix microarrays, also named Affymetrix GeneChip is a commercial, widely used microarray for the analysis of RNA or DNA. The foundation of Affymetrix microarrays was laid by Stephen Fodor in the late 1980 [33].

Affymetrix microarray chips measure the intensities of ten thousands of genetic sequences on a surface of about 1.3 $cm^2$ at the same time. The sequences interrogated - also called probes - are pre-defined. They represent parts of genes. Often multiple probes represent different parts of sequences of the same gene. A characteristic of Affymetrix microarrays is that probes consist of 25-base long oligonucleotides [33].

Affymetrix GeneChip analysis requires the following steps, which are described in the Affymetrix technical manual: After the biologic sample (e.g. blood, tissue) of interest is obtained, the RNA contained in the sample needs to be isolated. Then the RNA is reverse transcribed into complementary DNA (cDNA). The cDNA is transformed to biotin labeled cRNA via in vitro transcription reaction. For further analysis, the cRNA is fragmented, which means that RNA is separated into small sequences. The fragmented RNA is then hybridised to the Affymetrix array during a multiple hour incubation. In this process the fragments bind to the corresponding probe sets on the microarray platform. Subsequently, the array is washed, stained finally scanned [2]. There are detailed protocols provided by Affymetrix for every step. The scan delivers image files in .dat format. Finally, intensity values for each probe cell are calculated from the .dat images and stored in Cell intensitiy file (CEL) files. CEL files are the starting point of my analysis.

Affymetrix microarrays sometimes contain pairs of Perfect Match (PM) and Mismatch (MM) probes. Perfect match probes are the intensities measured from a 25-base long oligonucleotide. The idea of mismatch probes was to measure the amount of an unspecific signal, which can then be subtracted from the PM signal. Mismatch probes are measured the same way as others, just one base at the center of each oligonucleotide - the 13th position - is changed to a different base. However, this does not work as intended, since mismatch probes still include some of the signal[41]. Therefore, MM probes are usually discarded [33].

### 2.2.2 Illumina RNA Sequencing

Illumina sequencing methods belong to the category of next generation sequencing (NGS). NGS are massively parallelised sequencing methods, which produce huge amounts of sequences in a short time for a relatively low price. 2005, the so called Genome Analyzer brought sequencing to the next level compared to methods pre-NGS, which were part of the famous Human Genome Project. The Human Genome Project - the complete sequencing of the first genome - took 15 years and cost about three billion dollars [38]. In contrast, the Illumina HiSeq X system, which was released 2014, is able to sequence more than 45 human genomes in one day for about 1000$ per genome [38].

Compared to microarrays, RNA-Seq does not measure pre-defined intensities and is therefore not dependent on any species or pre-defined probes. RNA-Seq produces small nucleotide sequences, named reads [38].

With Illumina sequencing methods, DNA as well as RNA can be sequenced. The application of Illumina sequencing methods consist of four steps [38].

1. Library preparation: If RNA is analysed, first the complementary DNA (cDNA) is transcribed. In the next step the cDNA is randomly fragmented into short sequences. Then adapter ligations are applied to each sequence. These sequences are amplified using Polymerase Chain Reaction (PCR) and gel purified [38, 39].

2. Cluster Generation: The sequences prepared in the first step, are hybridized to the flow cell, a surface containing the complementary strands to the adapters added in step 1. During hybridisation, the sequence adapters bind to them. Then the sequences are amplified, such that clusters of the same sequences are on the flow cell. Note that the sequences are single stranded at this stage [38, 39].

3. Sequencing: In the sequencing step, the nucleotides of the reverse strands are added step by step. Each base is marked with a different fluorescent and when a base binds to the strand, the fluorescence is released. The emission of the whole cluster produces a measurable signal. The signal indicates which base was added and the base can be recorded in the read. This procedure is done nucleotide by nucleotide for all fragments simultaneously. The results are millions of reads stored in a data file, in the FASTQ format [38, 39]. In the so called FASTQ files, each read is recorded in 4 lines. The first line starts with an @ and contains the name of the read. The next line is the sequence content including a sequence of the letters A,T,G,C which stands for the bases

in the DNA or cDNA. The third line is a spacer e.g. a "+" and the fourth line contains the quality of each position of the read [14].

4. Data Processing: In this step data files are processed to extract meaning. This step is usually conducted by a bioinformatician and is the starting point of my analysis. How the reads are processed depends on the application. However, a common approach is to align the reads to some reference genome and count how many reads can be mapped to each gene. This results in a count matrix which is similar to the expression matrix in microarray data. The data processing of RNA-Seq data will be described more closely in section 4.2 in the Data and Methods chapter.

# Chapter 3

# Data

In this thesis, all publicly available gene expression datasets found in the context of treatment response in ulcerative colitis are analysed. This spans over eight datasets, six different treatments including placebo and two different data collection methods. The following sections describe the data selection process. How the datasets were queried, which studies were investigated more closely and which datasets were finally included in the analysis. Then, the sample data of all included studies are harmonised and compared on different levels.

## 3.1 Databank querys

The data repositories National Center for Biotechnology Information (NCBI) GEO (Gene Expression Omnibus)[54] and EBI Arrayexpress[22] were searched for suitable datasets with the search terms "Ulcerative Colitis"/"UC", "Inflammatory Bowel Disease"/"IBD", "treatment","response", and the names of various biologics (Infliximab, Adalimumab, Golimumab, Vedolizumab, Tofacitinib, Etrolizumab, Natalizumab, Ustekinumab). From this search, twelve studies were investigated more closely and finally eight studies were included in the analysis.

NCBI GEO is a publicly accessible databank for high-throughput gene expression data. Data is organised in three categories: *platforms*, *samples* and *series* [31]. Each entry in each category receives an accession number - a unique identifier.

- Platform entries contain array or sequencing platform information, e.g. which probes on the array correspond to which genes. Often commercial arrays are already registered and the platform entry may be used multiple times. The platform accession numbers start with "GPL".

- A sample record contains information about how the data of this sample was collected, where it was collected, how it was processed and the resulting data. The prefix of sample accession numbers is "GSM".
- A series record gathers multiple samples from one experiment and forms a dataset. Accession numbers of series start with "GSE".

EBI Arrayexpress is a data repository for functional genomics data. Data is organised in experiments. Experiments group samples from studies or publications. Whithin an experiment, metadata of the samples such as their origin and processing history is collected.

Both, NCBI GEO and EBI Arrayexpress support MIAME and MINSEQE compliant data submissions. MIAME (Minimum Information About a Microarray Experiment) [9] and MINSEQE (Minimum Information about a high-throughput nucleotide SEQuencing Experiment) [74] are guidelines that describe what information and data is needed to interpret and reproduce the results of Microarray and Next Generation Sequencing (NGS) analysis, respectively.

## 3.2 Studies investigated

This cross study analysis aims to gather and analyse all publicly available data containing the following information:

- transcriptomic data (Microarray or RNA-Seq) available for every patient
- treatment response to antibody based treatments.

Table 3.1 lists all studies investigated more closely, their sample origin, the method and platform used to collect gene expression data and whether the studies were included in the analysis or not. Two studies E-GEOD-12251 and GSE16879 were not included due to overlapping or redundant observations with already included studies. E-GEOD-14580 is a subset of GSE16879. E-GEOD-14580 contains observations before treatment only, while GSE16879 also contains observations after treatment. Therefore the latter was chosen for this analysis. Both, E-GEOD-12251 and E-GEOD-23597 collected samples from the Active Ulcerative Colitis Trial (ACT) 1 an 2 [69]. To check if these samples are independent, Manhattan distances between expression values of all samples were computed and clear sample pairs were detected. This indicated some patients were in both cohorts and that the datasets are not independent. Therefore, E-GEOD-12251 was excluded from this analysis. E-GEOD-21231 was not included because the treatment is not antibody based. Furthermore,

samples in the study were taken from blood, a different matrix as all other studies investigated. Which genes are expressed depends a lot on the matrix and therefore tissue and blood samples cannot be merged. GSE150961 contains a pediatric ulcerative colitis cohort. This study does not define treatment response similarly to other studies as e.g. endoscopic remission after 4-12 weeks, but the dataset contains information whether patients progressed to colectomy after one year or not. Besides the different target variable definition, the study moreover does not contain any antibody based treatments [37].

E-MTAB-7845 was excluded at the end of the pipeline, because evidence of a systematic error in the data provided in Arrayexpress was found. Close investigations suggest that sample information (e.g. the response group) and gene expression data are not correctly assigned. Unfortunately, the authors have not yet replied to these findings and therefore results from this dataset had to be excluded. Reasoning is described in detail in a separate section 5.1.4.

Of the studies included in the analysis, there are five Affymetrix microarray datasets and three Illumina RNA-Seq datasets.

| **Accession number** | **Sample origin** | **Method** | **Platform** | **Included in the analysis** | **Citation** |
|---|---|---|---|---|---|
| E-GEOD-12251 | tissue | Microarray | | no[1] | [5] |
| E-GEOD-14580 | tissue | Microarray | | no[2] | [6] |
| E-GEOD-21231 | blood | Microarray | | no[3] | [43] |
| E-GEOD-23597 | tissue | Microarray | Affy HG U133+ 2.0 | yes | [77] |
| E-GEOD-72819 | tissue | RNA-Seq | Illumina HiSeq 2000 | yes | [76] |
| E-MTAB-7604 | tissue | RNA-Seq | Illumina HiSeq 4000 | yes | [79] |
| E-MTAB-7845 | tissue | RNA-Seq | Illumina HiSeq 4000 | yes* | [78] |
| GSE16879 | tissue | Microarray | Affy HG U133+ 2.0 | yes | [4] |
| GSE73661 | tissue | Microarray | Affy HG 1.0 ST | yes | [3] |
| GSE92415 | tissue | Microarray | Affy HG U133+ PM | yes | [71] |
| GSE150961 | tissue | RNA-Seq | | no[4] | [37] |
| GSE212849 | tissue | Microarray | Affy HG U133+ 2.0 | yes | [60] |

[1] samples overlap with samples from E-GEOD-23597
[2] is a subset of GSE16879
[3] treatment not antibody based
[4] treatment not antibody based and response defined as whether a colectomy was necessary or not.
* excluded at the end of the pipeline, see results section5.1.4

Table 3.1: *Overview of considered studies*

## 3.3 Included studies

More details of the datasets included in the analysis are displayed in table 3.2. The accession numbers, the disease, the number of samples and patients included in the analysis, the timepoints of sample collection and the treatments are listed for each dataset. Some datasets include multiple timepoints of the same patients while others only include gene expression data from biopsies before treatment (time = 0). Treatments investigated are Infliximab, Adalimumab, Golimumab, Etrolizumab, Vedolizumab and Placebo and some observations are from healthy control groups, which did not receive a treatment. Not all observations included in the datasets are useful for this analysis. Thus, observations are filtered in course of the analysis. The filtering steps are described in section 4.1.3. Before proceeding with the analysis, the comparability of these studies, conducted at various institutions needs to be checked.

| **Accession** | **Disease** | **# samples** | **# patients** | **Time in weeks**[1] | **Treatment** |
|---|---|---|---|---|---|
| E-GEOD-23597 | UC | 113 | 48 | 0, 8, 30 | Infliximab, Placebo |
| E-GEOD-72819 | UC | 70 | 70 | 0 | Etrolitumab |
| E-MTAB-7604 | UC | 19 | 19 | 0 | Adalimumab, Infliximab |
| E-MTAB-7845* | UC, CD | 47 | 47 | 0 | Vedolizumab |
| GSE16879 | UC,CD,Control | 133 | 73 | 0, 4, 6 | Infliximab, Control |
| GSE73661 | UC, Control | 175 | 78 | 0, 4, 8, 9, 12, 14, 16, 18, 26, 28, 32, 38, 52 | Infliximab, Control, Vedolizumab, Placebo |
| GSE92415 | UC, Control | 183 | 98 | 0,6 | Golimumab, Control, Placebo |
| GSE212849 | UC | 84 | 84 | 0 | Golimumab |
| sum | | 804 | 497 | | |

[1] Time between treatment and sample collection. If time = 0 the sample was taken prior to treatment i.e. at baseline.
* excluded at the end of the pipeline, see results section5.1.4

Table 3.2: *Overview of included datasets*

## 3.4 Harmonisation

To check if the studies included are comparable, various aspects need to be considered. This includes study designs regarding the treatment and response as well as gene expression data collection methods. In particular, the treatment dose, the treatment duration, which patients are included in the study, how response is defined, when response was evaluated, how gene expression data was collected, which tissue was collected and which generation of microarray or RNA-Seq analysis was used to generate the data is analysed. Table 3.3 lists treatment dose and repetition information and table 3.4 contains response information of each dataset. This information was either extracted from the dataset or from the corresponding papers (see citations in table 3.1).

The dose of Infliximab is either 5 or 10 mg/kg, although there was no information of treatment dose in GSE73661. The two Golimumab datasets GSE92415 and GSE212849 have different dosages. The first dose of Golimumab in GSE92415 varied between 100, 200 or 400mg while in GSE212849 all patients got 200mg at the beginning. In both studies the second dose was given after two weeks and consisted of half of the first dose. In GSE212849 patients received a third dose after 6 weeks. In most studies treatment was given/repeated at weeks 0,2 and 6. The severity of UC can be graded with the Mayo score and its subscores (see section 2.1.2). The Mayo score ranges from 0 to 12 and its four subscores can take values from 0 to 3. The higher the score the more severe the symptoms. Apart from E-MTAB-7845 and GSE73661, all patients included in the studies had a Mayo score of at least 5 or an endoscopic subscore larger or equal to 2 before treatment. The inclusion criterion of E-MTAB-7845 was an endoscopically proven active disease, without a score classification and for GSE73661 the inclusion criterion is unknown. Summarising, the datasets include UC patients with severe symptoms that were endoscopically proven. Response was defined similarly in the studies. E-MTAB-7604, E-MTAB-7845, GSE16879, GSE73661 and GSE212849 defined response as "endoscopy subscore 0 or 1". The response definition of E-GEOD-72819 "Mayo score $\leq 2$ with no individual subscore $> 1$" is slightly stricter than the majority response definition. E-GEOD-23597 and GSE92415 define response as "Decrease of at least 3 points of Mayo score and a decrease of bleeding score of at least 1". This definition differs from the others as it does not focus mainly on the endoscopic subscore. It is less strict than the response definition used in the majority of studies. Because Mayo scores are not available in all datasets, the response definition cannot be unified. However, considering all response definitions, it can be concluded that patients classified as responders, had a significant improvement of symptoms.

| Accession | Treatment dose | Repetition in weeks |
|---|---|---|
| E-GEOD-23597 | Infliximab: 5/10mg/kg | $0, 2, 6$ |
| E-GEOD-72819 | Etrolizumab: 100/300mg | $0, 2, 4, 8$ |
| E-MTAB-7604 | Adalimumab: $160, 80, 40$mg | $0, 2, 3-8$ |
| | Infliximab: 5mg/kg | $0, 2, 6$ |
| E-MTAB-7845 | Vedolizumab: 300mg | $0, 2, 6$ |
| GSE16879 | Infliximab: 5mg/kg | 0 or $0, 2, 6$ |
| GSE73661 | Infliximab: NA | NA |
| | Vedolizumab: NA | NA |
| GSE92415 | Golimumab: | $0, 2$ |
| | first dose: 100/200/400mg | |
| | second dose: half of first dose | |
| GSE212849 | Golimumab: $200, 100, 100/50$mg | $0, 2, 6$ |

Table 3.3: *Comparison of datasets: treatment dose and repetition*

| Accession | TRT | Score before trt | Response definition | Time[1] |
|---|---|---|---|---|
| E-GEOD-23597 | IFX | Mayo: $6-12$ | Decrease of at least 3 points of Mayo score and a decrease of bleeding score of at least 1 | 8 |
| E-GEOD-72819 | EMB | Mayo: $\geq 5$ | Mayo score $\leq 2$ with no individual subscore $> 1$ | 10 |
| E-MTAB-7604 | ADM, | Endoscopy subscore $\geq 2$ | Endoscopy subscore | 8 |
| | IFX | | 0 or 1 | 14 |
| E-MTAB-7845 | VDZ | Endoscopically proven active disease | Endoscopy subscore 0 or 1 | 14 |
| GSE16879 | IFX | Mayo: $6-12$ | Endoscopy subscore 0 or 1 | 4 or 6 |
| GSE73661 | IFX, | NA | Endoscopy subscore 0 or 1 | $4-6$ |
| | VDZ | NA | | 6 |
| GSE92415 | GLM | Mayo: $6-12$ and Endoscopy subscore $\geq 2$ | Decrease of at least 3 points of Mayo score and a decrease of bleeding score of at least 1 | 6 |
| GSE212849 | GLM | Mayo: $6-12$ and Endoscopy subscore $\geq 2$ | Endoscopy subscore 0 or 1 | 6 |

[1] time of response evaluation in weeks after treatment

Table 3.4: *Comparison of datasets: inclusion criterion, response definition and time*

Besides differences in treatment and response there are different data generation methods. In table 3.2 the platforms used to extract gene expression information are listed. There are three different Affymetrix microarray types and two different

Illumina RNA-Seq generations. Microarray and RNA-Seq both measure gene expression, however the methods are fundamentally different (see section 2.2) and therefore the microarray and RNA-Seq data is preprocessed and analysed separately and then compared on a fold change level. In microarray datasets treatments overlap. Therefore, the datasets are merged to one large dataset. The probe ids are not identical in different Affymetrix generations, so the probe ids are converted to gene symbol ids using biomart [21]. If multiple probe ids map to the same gene symbol, the average intensity of probes is taken. If one probe id maps to multiple genes, gene symbols are concatenated to one id. RNA-Seq datasets are analysed individually, because there are no overlaps in treatments.

# Chapter 4

# Methods

Microarrays and RNA-Seq both measure gene expression, however the methods are fundamentally different (see Section 2.2) and therefore the data is processed and analysed separately. In this section the preprocessing of microarray and RNA-Seq data is described. Since the preprocessing steps depend a lot on the data, some visualisations of the data are also included in this section. Then, the methods used in differential expression analysis and prediction analysis are explained. Finally, the main software used in the analysis is listed.

## 4.1 Microarray data preprocessing workflow

This section describes the steps applied to get from raw CEL files in each datasets to one merged dataset that can be used for further analysis such as differential expression (section 4.3) and prediction (section 4.4). The microarray data preprocessing workflow is based on the workflow described by Bernd Klaus in [45]. CEL files are imported, robust multiarray average (rma) is applied, datasets are merged and Combat batch correction is performed. Throughout the analysis, quality checks are conducted using suitable visualisations and descriptive statistics.

### 4.1.1 R Object Expression Set

Gene expression data contains information upon multiple levels. On the one hand, there are expression values and on the other, there is metadata of the samples and metadata of the genes. An especially useful object to store and analyse this data in R is the expression set object from the biobase package [36]. It enables data manipulation such as subsetting, filtering and merging of datasets. An expression set

consists of three elements, the main element is the expression matrix (exprs), a matrix containing the gene expression values with one row for each gene and one column for each sample. The second element, the phenotype data (pData) provides room for sample and patient details. Each sample is assigned to one row and arbitrarily many features can be added as columns. Information about the genes are stored in the third element, the feature Data (fData) [36]. Instead of a non-descriptive id, one may add a gene description or a different gene identifier. There is a number of gene identifier classes, e.g. NCBIs Gene Symbols, Entrz ID, EBIs Ensemnbl ID. After downloading the CEL files from ArrayExpress and GEO, respectively (see 3.1), all datasets were rearranged into the expression set format. Therefore, the phenotype data needs to be downloaded and relevant information extracted.

### 4.1.2 Background correction, normalisation and summarisation

Figure 4.1 shows boxplots of raw gene expressions for each array. The majority of gene expression values is 0 or just above 0. Therefore, the boxplots are skewed at 0 and mainly very highly expressed genes are visible. Moreover, it provides an overview of the range of expression values. Study GSE212849 for example contains expressions above 40000, while the maximum expression value of GSE92415 is about 24000.

In general, Affymetrix microarray data is noisy due to inevitable differences in sample preparation in the lab, differences in chip manufacturing or processing[33]. The smallest deviation of sample processing, e.g. at which time of the day or at which atmospheric ozon level the array was processed, can influence results [42, 29]. To overcome the noisy, skewed and differently scaled expressions, data transformations are necessary. Irizarry et.al. introduced a summary measure the so called robust multiarray average (RMA)[41]. It consists of four steps, the background correction, quantile normalisation, log2 transformation and summarisation.

Background correction aims to extract the true signal from the intensities measured and thereby remove the so called background noise. This method assumes that the intensities consist of the true signal which is exponentially distributed and the background noise which is normally distributed. The estimation of the true signal is calculated as the expected value of the true signal given the measured intensities[41].

Quantile normalisation unifies the distribution of probe intensities of each array. It aims to remove variation in arrays from the same dataset. The variation originates from the above mentioned, inevitable differences in sample preparation in the lab, differences in chip manufacturing or processing. The average distribution of expres-

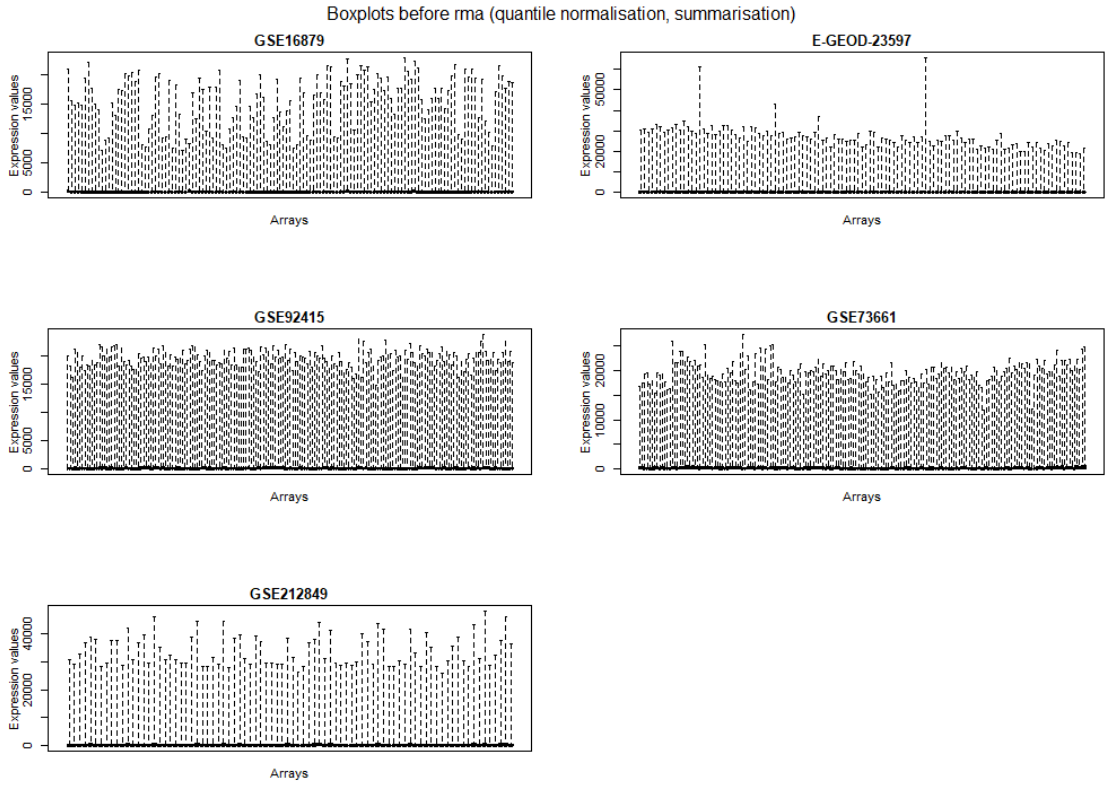



Figure 4.1: *Boxplots of raw microarray data*
*Each boxplot represents the expression values of one array.*

sions is calculated and subsequently the distribution of intensities of each array are adjusted to this average distribution[33].

The expressions are log2 transformed to remove the large range of expressions where most data is skewed at 0 and some observations are very large. This data is better represented by a log transformed scale.

Some Affymetrix chips perform repeated measurements. By summarising them on a probeset level, one probe is represented by one value. The summarisation algorithm in RMA only uses PM probes and neglects MM probes (see 2.2.1) [33].

The RMA algorithm was applied to all five microarray datasets. The processed expressions are shown in Figure 4.2 and 4.3 in the upper graphs "Expressions before batch correction". RMA transformed expressions that were initially in the range between 0 and a couple of thousands (see raw data in Figure 4.1) to a range between

0 and 15.

### 4.1.3 Sample Filtering

Depending on the analysis, different observations need to be filtered. Table 3.2 contains summary statistics of all available data. Some datasets include not only UC patients, but also CD patients or healthy control observations. This analysis focuses on UC patients only, therefore CD and control are excluded. Some datasets contain placebo data. Placebo observations are useful as controls for the analysis of treatment response and therefore included in the analysis and treated as individual treatment group.

The first two research questions focus on gene expression data at baseline. For these analyses observations before treatment (at time = 0) are filtered. The third research question aims to analyse the change in gene expression during therapy - using "follow up-baseline" data. Therefore, observations of the same patient before and after treatment are required. Due to data availability, one timepoint after treatment - namely the first observation at least 4 and at most 12 weeks after treatment - is chosen. From this point, baseline and change during treatment data are analysed separately.

### 4.1.4 Dataset Merging

Filtering splits the analysis into two parts. The analysis of gene expression at baseline and the change during treatment data composed of baseline and follow up observations.

Five microarray datasets contain observations at baseline. These are merged to one large expression set. Therefore the intersection of gene symbols evaluated by different chip generations is taken. A dataset of $17,787$ different gene symbols remains.

Four datasets contain observations before and after treatment. These are merged to another large expression set including $17,787$ gene symbols for the change during treatment analysis.

### 4.1.5 Batch correction

Boxplots (Figure 4.2 and Figure 4.3), results of PCAs (Figure 4.4 and Figure 4.5) and PVCAs (Figure 4.6 and Figure 4.7) showed clear batch effects in both merged datasets, baseline and change during treatment. These batch effects introduce noise to the data and need to be corrected before further analysis.

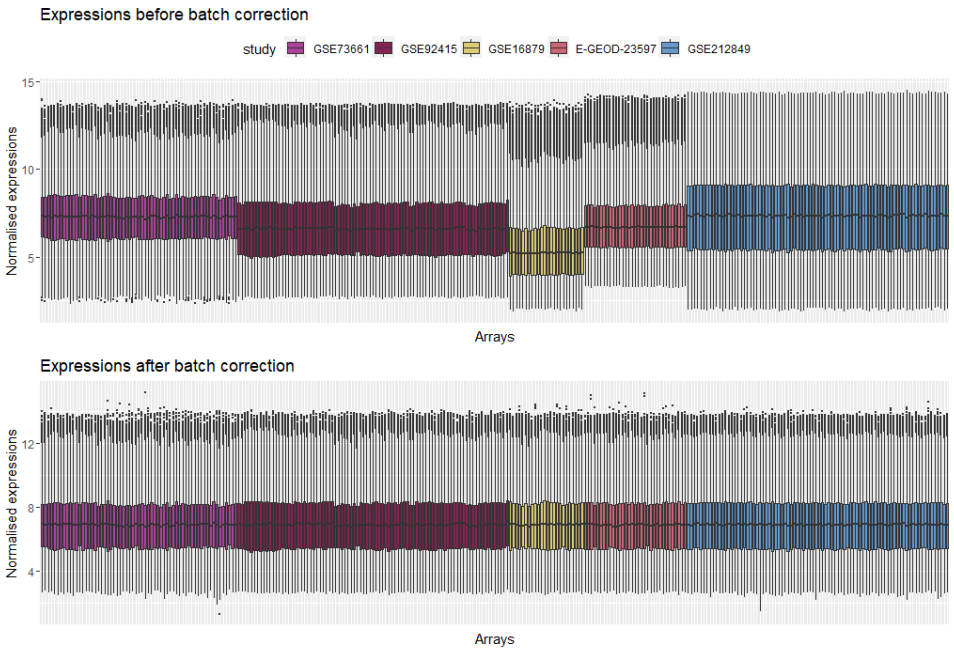



Figure 4.2: *Boxplots of expressions by arrays before and after batch correction. Data: baseline*

There are different methods to address batch effects. One method, which has worked well in the past is `ComBat`, which adjusts batch effects using Empirical Bayes [33]. This method was proposed by Johnson et.al. in 2007 [42]. It assumes that batch effects affect genes within batches in similar ways. Furthermore it requires normalised data, no missing values in the expression matrix, the model matrix and a known variable to adjust for. `ComBat` fits additive and multiplicative batch effects for each batch and each gene using Empirical Bayes. Empirical Bayes estimates do not use any arbitrary prior distribution. It uses the data to estimate the parameters of the prior distribution. In the case of `Combat` a normal distribution is used for the additive batch effect and an inverse gamma distribution is used for the multiplicative term. The parameters are estimated using the method of moments [42].

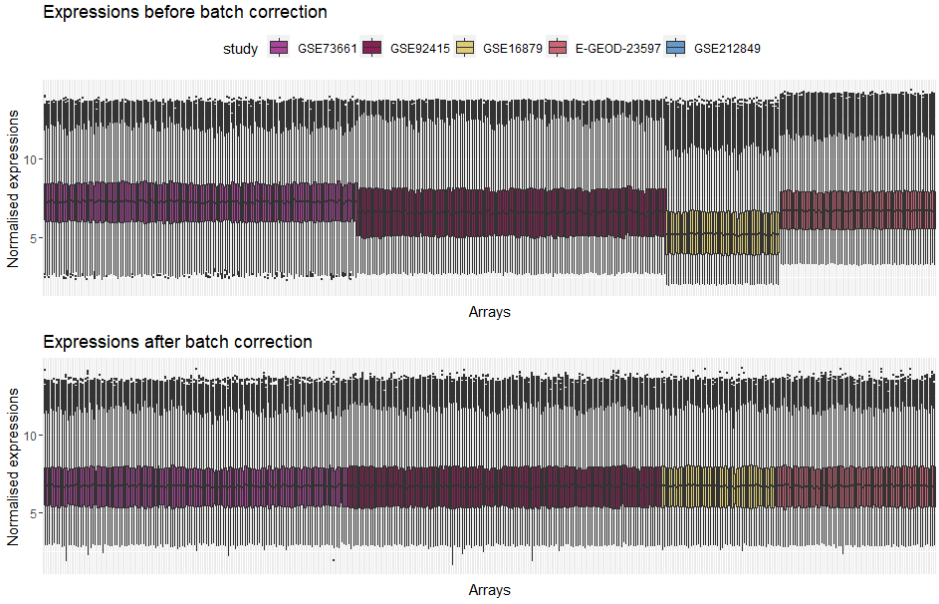



Figure 4.3: *Boxplots of expressions by arrays before and after batch correction. Data: baseline and follow up*

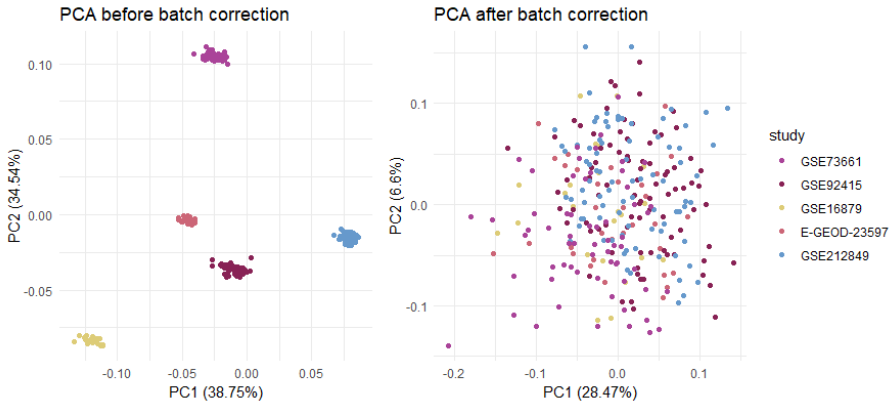



Figure 4.4: *First two principle components before and after batch correction. Data: baseline*

For batch correction the R function `ComBat` from the sva package is used. The plots of the first two principle components before and after batch correction (Figure 4.4 and Figure 4.5) show that after batch correction neither study batches nor other

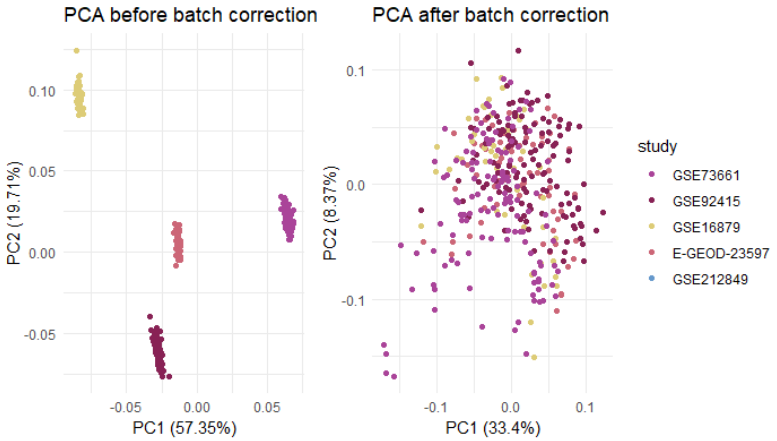



Figure 4.5: *First two principle components before and after batch correction. Data: baseline and follow up*

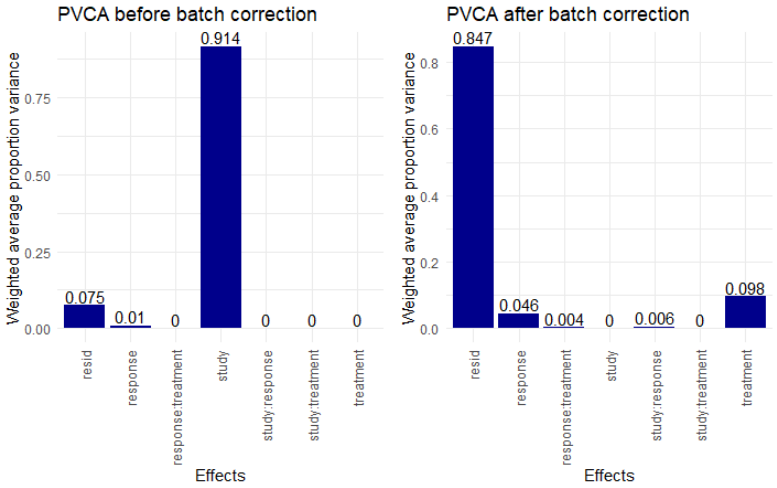



Figure 4.6: *Results of PVCA before and after batch correction. Data: baseline*

obviously visible batches remain. The results of the Principle Variance Componens Analysis (PVCA) shown in Figure 4.6 and Figure 4.7 display how much of the variance is explained by which variable. Before batch correction 91.4% and 93.3% of the variance is explained by the variable study. This effect is completely eliminated after batch correction. In return other variables such as response (4.6% and 8.5%) and

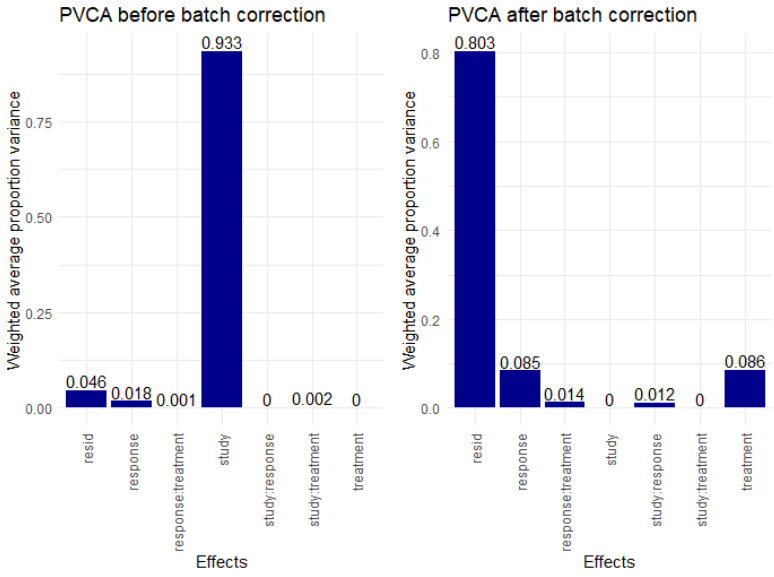



Figure 4.7: *Results of PVCA before and after batch correction. Data: baseline and follow up*

treatment(9.8% and 8.6%) gain explaining power and a large part 84.7% and 80.3% can not be explained by any of these variables and appears in the residual term. By removing the noise of the study bias, variables of interest emerge as explaining variables.

## 4.2 RNA-Seq data preprocessing workflow

The RNA-Seq data preprocessing workflow describes how fastq files are processed to counts and how counts are processed such that they can enter differential expression (section 4.3) and prediction (section 4.4). Processing of fastq files is based on the 2019 biocore lecture on RNA-Seq [14]. First, a quality check of the raw fastq files is done. Since no adapters to trim are found in the quality control, the reads are aligned to the human reference genome and genes are counted with STAR. After a quality control of the read alignment, the counts are analysed. The analysis of counts is orientated on the limma userguide [73]. Low counts are excluded from the data, the logarithm of counts per million (cmp) are calculated and normalised.

### 4.2.1 Fastqc

To check the quality of the raw reads in the fastq files, the program FastQC[30] was used. FastQC generates one report for every sample. The program MultiQC[28] subsequently summarises all FastQC outputs to one report. Quality control of sequencing data is important to learn more about the data, its composition and quality. FastQC calculates basic statistics such as number of reads or read lengths. It also summarises the sequence qualities and reports the proportion of the four bases per read position. This helps to find anomalies in the data. Before sequencing adapters are added to the RNAs (section 2.2.2), which are sequenced but not needed for further analysis. In the quality check the adapter sequences are located and specified and can be removed from the reads in a separate trimming step.

The multiqc reports generated for the datasets E-GEOD-72819, E-MTAB-7604 and E-MTAB-7845 showed a good quality in reads. Table 4.1 shows output plots from the quality control. In Illumina sequencing it is often the case that read quality decreases the longer the read gets due to cumulative errors. All three datasets used an Illumina HiSeq sequencing method and all three datasets contain relatively short reads (50 and 51 bases). Due to the short read length, the problem of reduced base quality as the sequence gets longer, is hardly present in the datasets. Only the first one, E-GEOD-72819 shows a small quality reduction over read length. FastQC reports a notable sequence duplication in all three datasets. Furthermore, there are patterns visible in the base composition heatmaps. The heatmaps show the base composition of each sample across the read length. A grey color means that all bases are represented approximately equally. If one color dominates, it means that one base appears more often than others. In all three datasets, the first 12 bases show a similar pattern of base imbalance. This pattern is common and is most

probably caused by a random priming step. In theory, random priming should result in random primers, but in practice some primers are favoured during the priming step[72]. Although this produces a bias in reads, this bias is present in every sample and therefore will not affect comparisons. Simon Andrews argues about the idea of trimming the affected sequences, that this still would not solve the problem since the same bias is contained after the first 12 bases[72]. MultiQC reported that no adapter sequences were found and therefore trimming is not necessary. The reads are without further processing aligned to the human genome.

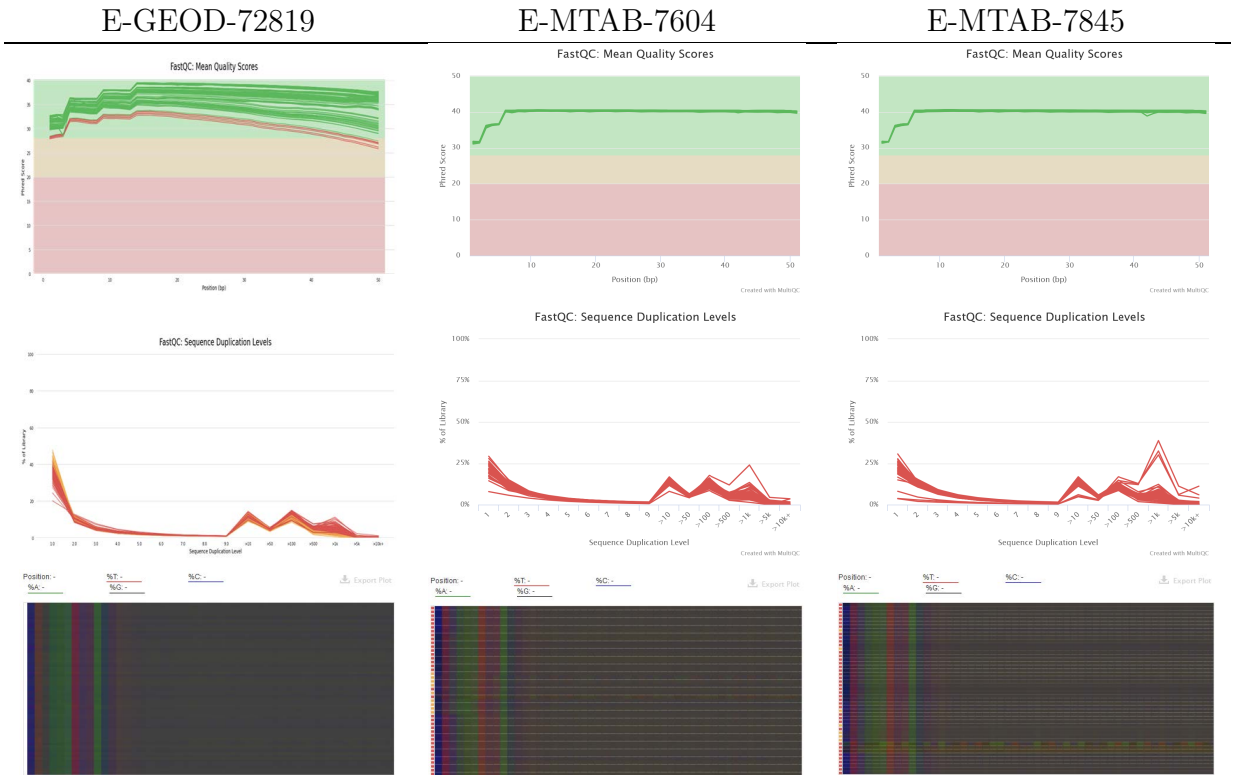



Table 4.1: *RNA-Seq Quality Control: Mean quality scores, sequence duplication levels and a heatmap of base composition of the reads*

### 4.2.2 Read Alignment

Aligning the reads to the human genome (homo sapiens GRCh38.98) was performed with STAR (Spliced Transcripts Alignment to a Reference)[18]. The alignment of millions of reads to a genome is a computationally challenging task. To perform this

task with reasonable cost, efficient algorithms and solutions have been developed. Which algorithm is used, often depends on the computational resources available. STAR was developed to align RNA-Seq data and performs read mapping with high accuracy. It outperforms other algorithms in speed, however it is memory intensive[1]. STAR can process single stranded or double stranded sequencing data[17].

STAR uses a suffix array index to speed up the matching process. For each read it searches the longest sequence of the read that matches to a part of the reference genome. This sequence is called the MMP the Maximal Mappable Prefix[1]. Unless the whole read matches a part in the genome, an MMP of the remaining part of the read is searched. This splitting of reads accounts for splicing and RNA-fusion[40].

STAR was executed with the following options:

```
#star index
STAR    --runMode genomeGenerate
        --genomeDir Star_index_49bp/Star_index_50bp
        --genomeFastaFiles Homo_sapiens.GRCh38.dna.primary_assembly.fa
        --sjdbGTFfile Homo_sapiens.GRCh38.98.chr.gtf
#star
 STAR   --runThreadN 16
         --genomeDir Star_index_49bp/Star_index_50bp
         --readFilesIn $i
         --readFilesCommand zcat
         --outSAMtype BAM SortedByCoordinate
         --quantMode GeneCounts
```

The read alignment worked well for all three datasets. In Figure 4.8 the alignment qualities of all samples are summarised. In all datasets, the proportion of uniquely mapped reads is around 80%. Some reads (about 20%) are mapped to multiple loci and almost 0% are unmapped or mapped to too many loci. This distribution of read mappings is expected for RNA-Seq data. Uniquely mapped reads are used for counting[17].

### 4.2.3 Counting Reads

The next step is to count the reads that have been aligned to the human reference genome. Therefore, the number of reads uniquely mapped to each gene is counted. STAR optionally performs counting while aligning reads to the genome[17]. A read is counted if it overlaps one gene and not counted if it overlaps more than one gene. The total counts of each sample counted by STAR is shown in Figure 4.9. The sum

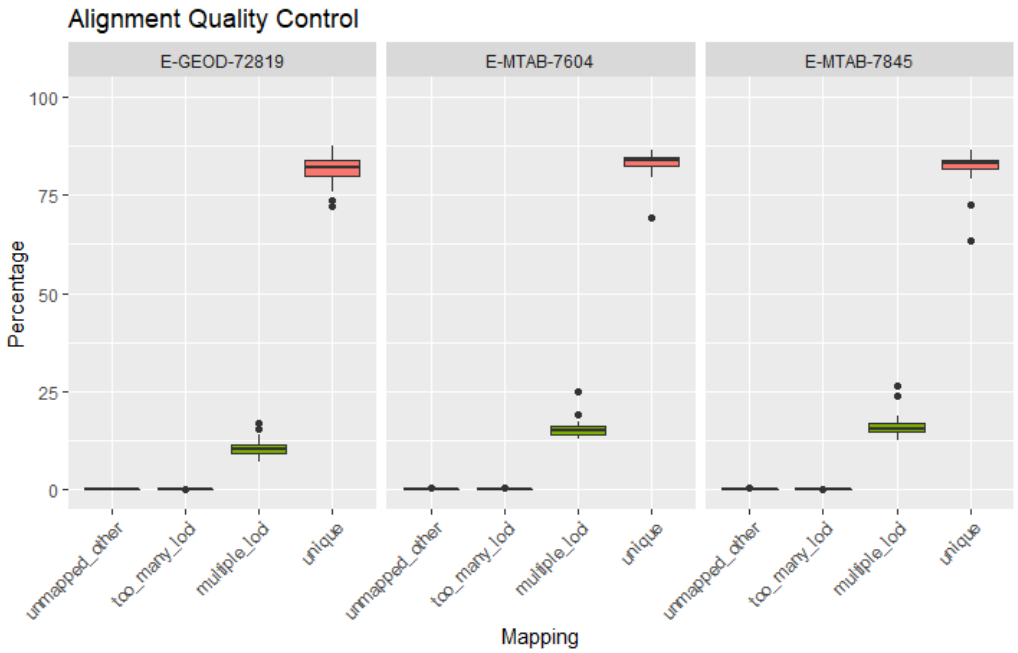



Figure 4.8: *Alignment quality control*

of all counts of a sample is also known as sequencing depth or library size of the sample.

Before counts can enter differential expression or prediction analysis, they need to be preprocessed. Genes with no or very few counts throughout all samples are excluded, the counts are transformed into log counts per million (logCPM) and TMM (trimmed mean of log-ratios) scaling accounts for library size differences [68].

### 4.2.4 Exclusion of low counts

For the exclusion of low counts the R function `filterByExpr` from the EdgeR package was used. Kept genes must have a total count of at least fifteen over all samples. Additionally, there have to be at least ten reads in $n$ samples, where $n$ is the number of samples in the response group with fewer samples. The number of features remaining are: Infliximab+Adalimumab: 16070, Etrolizumab: 18102 Vedolizumab: 15847.

These genes have sufficiently large counts, such that differential expression can be detected.

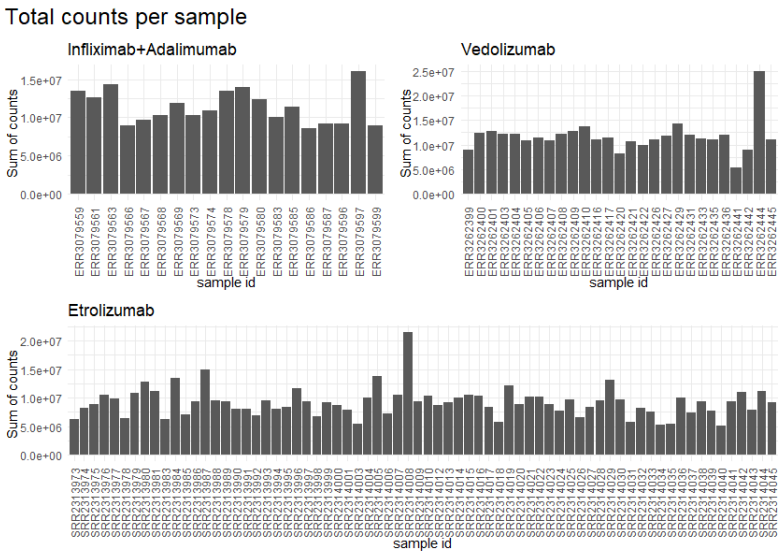



Figure 4.9: *Sequencing depth: Sum of all counts per sample*

### 4.2.5 Processing of counts

Further processing of counts is necessary to enable the comparisons between samples and to minimize technical biases in the data. For example, different sequencing depths impact comparisons between samples [27]. Variability in library sizes between samples occur due to differences in processing of samples, such as the number of molecules sequenced [27]. As there are slight differences in sequencing depths in the data (see Figure 4.9), counts can not directly be compared. The limma user's guide suggests to use TMM normalisation and cpm for comparative studies [73]. Therefore, the R functions `calcNormFactors` and `cmp` from the edgeR package are applied before further analysis of differential expression or prediction. If one sample has twice as many reads as another sample, it would be intuitive to simply scale the number of reads per gene by library sizes. Robinson and Oshlack argue that this approach is too simple for biologic processes as this approach neglects differences in compositions of RNA samples i.e. differentially expressed genes in different samples [68]. They propose TMM normalisation, which accounts for library size as well as differences in compositions of RNA samples.

## 4.3 Differential Expression (DE)

The detection of differentially expressed genes between predefined groups is a common analysis for gene expression data [33]. Differential expression methods typically compare the expression values between groups for each gene separately. This generates a large multiple testing problem which is discussed in section 4.3.1. Another burden of this univariate approach is the individual analysis of genes. The dynamics of gene expression in cells is complex. Genes often act together in so called pathways. Moreover, some genes are regulated by other genes and therefore directly or indirectly proportional to each other. Summarised, gene expression values are not independent. Despite the fact that this univariate approach neglects these complex dynamics of gene expression, it is one of the most commonly used analysis of gene expression data. Göhlmann and Talloen (2009, p154-155, [33]) reason this as follows:

> The incomplete approach of testing single genes independently for differential expression works remarkably well in practice. The main reason is because of the curse of high dimensionality of microarray data. Because there are typically so many groups and subgroups of correlated and interacting genes present in a single microarray, a multivariate approach that tries to incorporate most or all covariates becomes highly complex. And such a model has two drawbacks: it is difficult or impossible to interpret, and it is likely to overfit thereby impending its generalizability. In a complex and high-dimensional context, a simple solution will often guide the researcher to the clearest and most informative patterns of the data.

In this analysis, there are about $17,000$ genes in the microarray and RNASeq data. Depending on the treatment, $19-143$ samples are available. This feature observation ratio is very sensitive and requires special methods. There is always the connection between effect size and sample sizes needed to find this effect significant. Although the generation of gene expression data becomes less expensive every year, the cost and effort of gene expression data collection restricts sample sizes. Restricted sample sizes mean less power, which means that weak effects are more difficult to detect. With differential expression, the most prominent differences in gene expression data between responders and non-responders can be detected.

For differential expression in microarray data multiple strategies are found in the literature. One of the most commonly used method are gene wise linear models, with an empirical Bayes moderated variance for hypothesis testing. This concept was applied to the microarray data for each treatment separately using the functions `lmFit`,

`eBayes` and `topTable` from the limma[66] R package. This pipeline of functions performs moderated t-tests for each gene between responders and non-responders. It allows a weight matrix and a custom design matrix as input. There was no custom weight matrix used and as a contrast for the design matrix *responseYes-responseNo* was defined. The pipeline returns a toptable, a ranked gene list including p-values, adjusted p-values and fold changes from the moderated t-tests. Adjusted p-values are Benjamin Hochberg adjusted p-values and the log2 fold change of a gene is the difference between the average expressions of responders and non-responders on a log2 scale. According to the defined contrasts, it follows that genes with a negative log fold change have a smaller mean expression in responders than in non-responders and these genes are therefore called down regulated genes in responders. Analogously, genes with a positive log fold change have a larger mean expression in responders than in non-responders and these genes are called up regulated genes in responders.

The moderated t-tests applied in this analysis can be described as follows: In the moderated t-tests, the null hypothesis that the means of expressions of responders and non-responders are equal is tested against the alternative that the means are unequal. The moderated t-test assumes normally distributed expressions within genes. Standard t-tests and moderated t-test differ in the standard deviation used in the test statistic and consequently in the distribution the teststatistic follows under the null hypothesis. The variance estimates are adjusted using an empirical Bayes model such that variances are shifted towards a common variance across all genes. The test statistic of one hypothesis test for one gene is given in the following equation 4.1.

$$T_{moderated} = \frac{\bar{x}_{response} - \bar{x}_{noresponse}}{\tilde{s}/\sqrt{n}} \tag{4.1}$$

$$\tilde{s}^2 = \frac{s^2 d + s_0^2 d_0}{d + d_0} \tag{4.2}$$

where $\bar{x}$ is the mean expression of a gene for the specific group. Moreover, $n$ is the sample size and $\tilde{s}^2$ in equation 4.2 the empirical Bayes moderated variance. This estimate weights the originally estimated variance of the tested gene $s^2$ and the global variance over all genes $s_0^2$ according to the degrees of freedom $d$ of the original t-test and the degrees of freedom from the prior distribution $d_0$. This smoothens the variances and therefore penalizes those genes with low variance estimates [33]. Under the null hypothesis, the test statistic follows a t-distribution with $d + d_0$ degrees of freedom [62]. By including the variance estimates of other genes in the variance estimate used one test, a more stable inference especially with low sample sizes is provided [33].

To analyse the RNA-Seq data similarly to the microarray data, limma-trend was chosen over alternative DE methods such as DESeq or limma-voom. Slight adjustments to the moderated t-tests used in the microarray data have to be made such that the model fits better the properties of RNA-Seq data. Law et al. developed two methods for dealing with RNA-Seq data: *limma-trend* and *voom*[47]. If sequencing depth is similar between samples, both methods have performed equally well. As sequencing depths of the three RNA-Seq datasets are very similar, (see Figure 4.9) and limma-trend is more similar to the method applied on microarray data, limma-trend was used. Variances of RNA-Seq data vary with count size and therefore a shift to a global variance as in moderated t-test would not be accurate[47]. Instead, limma-trend shifts variances to a mean-variance trend curve using empirical Bayes. The mean-variance trend curve is a curve fitted on the count size and variance[47].

Results of differential expression are presented in volcano plots, heatmaps and toptables. Results of multiple Differential expression analysis are compared with venn diagrams and scatterplot-correlation matrices of log2 fold changes.

### 4.3.1 Multiple testing

In univariate differential expression analysis, as many hypothesis test as there are genes are performed. As multiple tests with the same hypothesis are calculated, multiple testing needs to be kept in mind. If for instance $20,000$ tests with a significance level of 0.05 are conducted, 1000 false positives are expected. In the toptables resulting from differential expression analysis, Benjamin Hochberg p-values [8] are reported. In the volcano plots original p-values are used for better comparison between treatments. Especially for low sample sizes, effects might not be large enough to detect significant genes when correcting for multiple testing. As a threshold for gene selection a p-value smaller 0.05 and a log2 fold change smaller $-1$ or larger 1 is chosen. This threshold selects only those genes that have large fold changes i.e. large differences between groups.

It should be noted, that this is an hypothesis-generating analysis. Starting from a couple thousands of genes, this analysis reports the most promising genes, which can be evaluated in follow up studies. In these kinds of analysis, false discoveries are not as fatal as missed discoveries [33].

## 4.4 Prediction

A contrary approach to differential expression are multivariate prediction models to classify responders and non-responders. Machine learning models can be used to detect a set of genes that classify response to a certain treatment. Machine learning models can also be used to evaluate the predictive power of a preselected set of genes. With the microarray data, a set of predictive genes is found using prediction models and with the RNA-Seq data the genes found in the microarray analysis are evaluated.

To find predictive genes from the full set of genes, genes need to be prefiltered before handing them to a machine learning model.

### 4.4.1 Prefiltering

After all preprocessing steps, the Microarray data contains about 17.000 genes. Sample sizes per treatement are ranging from 36 to 143 samples. Before a model can be fit to the data, unsupervised filtering is necessary to decrease dimensionality and exclude features that are unlikely to be good predictors. If the expressions of a gene do not vary much between any sample, they will not vary between responders and non-responders and this gene will not be a good predictor. Therefore, a filter that does not use the response information, but only the expression values is applied to select a subset of genes. The idea is to extract those genes that vary much and that contain many extreme expression values. In this analysis the range between the 5% and the 95% quantile is calculated, and genes with the largest values are selected. Besides the inter 5-95 quantile range also a variance and an interquartile filter were explored. Setting the threshold of how many genes are selected is difficult due to the following tradeoff: If the cutoff is very strict, only a few genes make it to the modeling step which might introduce a bias and important features might be neglected. If the cutoff is too loose, there might be too many features and too much noise for the model. Therefore, different thresholds were considered with the aim to select as many genes as possible without loss of the models predictive performance.

### 4.4.2 Repeated-repeated cross validation (rrCV)

To build and evaluate robust models, repeated-repeated cross validation (rrCV) is used. An overview of the concept is given in Figure 4.10. The main idea is to repeat repeated cross validation on different splits of test and training data.

This method incorporates properties of gene expression data:

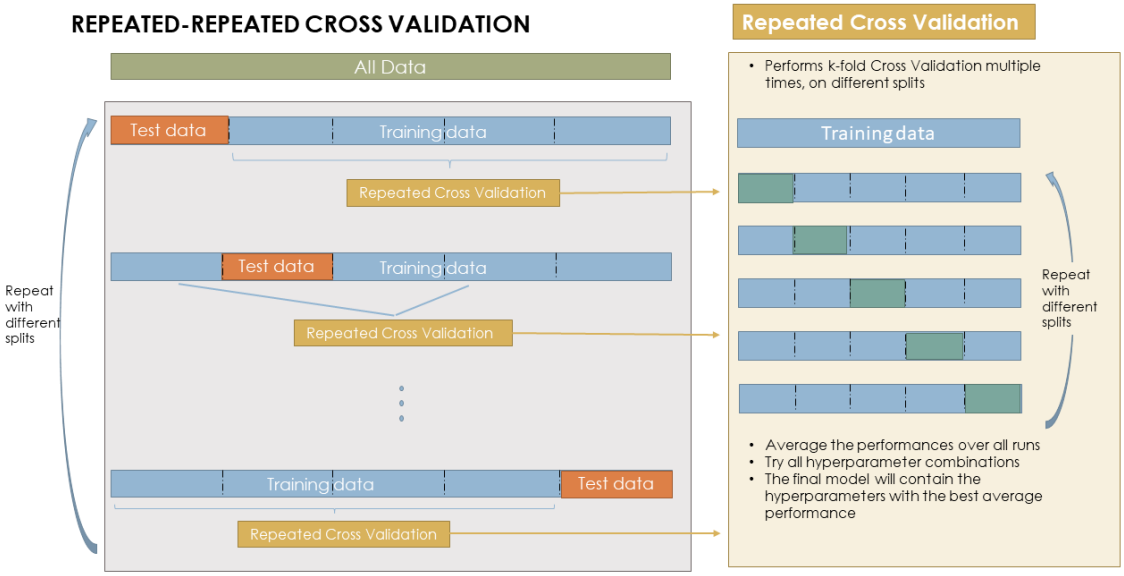



Figure 4.10: *Repeated-Repeated Cross Validation*
*The full data is split into folds (test folds). Each fold is once used as test data and the remaining folds are used as training data. With the training data repeated CV is performed, which returns a model that is evaluated on the test data. This is repeated on different splits (test repeats). Therefore, $\#test\_folds \times \#test\_repeats$ many models are returned. RrCV takes test_folds, test_repeats, train_folds, train_repeats and the models parameters as inputs.*

- Since there are more features than observations in gene expression data and there is a high risk of overfitting. *Cross validation* trains and evaluates models on different data and therefore avoids overfitting.
- The smaller the dataset, the higher the danger of including a split bias. *Repeated cross validation* performs multiple splits and therefore yields a more robust model.
- The relationships of gene expressions are complex. This complexity might not be captured by one single model. To generate multiple models and still evaluate the models using an independent test set, *repeated-repeated cross validation* is useful. An additional advantage of rrCV is that it provides a distribution of confidence for each predictor, extracted from all models.

Even after prefiltering genes, there are usually more features than observations.

Thus, a model that can be applied in this setting has to be chosen. In this analysis, glmnet also known as elastic net[89] and random forest are used. Therefore the implementation of the R caret package [46] is used.

Both models glmnet and random forest have the advantage of providing a variable importance measure for each predictor. The variable importance of the glmnet is calculated as follows: The absolute value of the coefficient estimates are taken and then mapped to a range between 0 and 100. The variable importance of the random forest is a post hoc analysis. An out of bag performance is fitted for each tree of the forest. Then the values of one variable are permuted and the out of bag performance is measured again for every tree. The variable importance of this one variable results from the differences between the performances before and after permuting the values of the variable [52]. A drastic reduction of performance suggests that the variable is important in the model.

## 4.5 Software

The main software used for the analysis was R v4.2.1 [63]. For data wrangling tidyverse v1.3.1 [87] including dplyr v1.0.9 [86], tidyr v.1.2.0 [84], readr v2.1.2 [85] and stringr v1.4.0 [83] was used. Visualisations were constructed with ggplot2 v3.3.6 [82] and patchwork v1.1.2 [61] For gene expression data libraries from the Bioconductor v3.15 collection of bioinformatics libraries were useful. For data processing and analysing limma v3.52.2 [66], edgeR v3.38.1 [67], oligo v1.60.0 [13] and Biobase v2.56.0 [35] were mainly used.

For handling different gene IDs, biomaRt v2.52.0 [21], AnnotationDbi v1.58.0 [59] and EnsDb.Hsapiens.v79 v.2.99.0 [65] were used.

Models were built with caret v6.0-93 [46].

RNA-Seq raw data was processed with FastQC v0.11.9[30], MultiQC v1.9 [28] and STAR v2.7.6a [17].

The full list of R libraries used is provided in the appendix D.

# Chapter 5

# Results and Discussion

In this chapter, each section is dedicated to one research question. First, differentially expressed genes between responders and non-responders at baseline are analysed. Second, response prediction at baseline is calculated. Third, the change of gene expression during treatment in responders and non-responders for different treatments is evaluated by differential expression.

Data was processed as described in the data preprocessing workflows for microarrays 4.1 and RNA-Seq 4.2, respectively. Both preprocessing workflows prepare the data for differential expression and prediction. Table 5.1 shows baseline and table 5.2 shows follow up - baseline sample sizes remaining after preprocessing.

| | Treatment | Samples | Response | No response |
|---|---|---|---|---|
| Microarray | Infliximab | 72 | 34 | 38 |
| | Golimumab | 143 | 53 | 90 |
| | Vedolizumab | 40 | 9 | 31 |
| | Placebo | 36 | 14 | 22 |
| RNA-Seq | Infliximab+Adalimumab | 19 | 2+4 | 6+7 |
| | Etrolizumab | 70 | 12 | 58 |
| | Vedolizumab* | 27 | 16 | 11 |
| | sums | 291+116 | 110+34 | 181+82 |

* excluded at the end of the pipeline, see section 5.1.4

Table 5.1: *Sample sizes at baseline*

| | Treatment | Samples | Response | No response |
|---|---|---|---|---|
| Microarray | Infliximab | 72 | 34 | 38 |
| | Golimumab | 42 | 24 | 18 |
| | Vedolizumab | 40 | 9 | 31 |
| | Placebo | 31 | 11 | 20 |
| | sums | 185 | 78 | 107 |

Table 5.2: *Sample sizes follow up-baseline*

## 5.1 Differentially expressed genes in responders and non-responders at baseline

Differential expression for microarray and RNA-Seq data was performed using the limma package (see section 4.3). Groups of responders and non-responders were compared for each treatment and method (microarray and RNA-Seq) separately. The individual results of each differential expression analysis are presented as volcano plots and selected genes as expression heatmaps. Furthermore, toptables including the top selected genes are printed in the appendix B. To compare the results of all groups, log2 fold changes are compared in a scatterplot-correlation matrix. This comparison showed unexpected results in the Vedolizumab RNA-Seq dataset (E-MTAB-7845). Further investigation (see section 5.1.4) revealed a systematic error in this dataset and therefore its results are not reliable and not presented.

### 5.1.1 Microarray

For each treatment group comparisons between responders and non-responders are calculated individually. The microarray data contains four treatments: Infliximab, Golimumab, Vedolizumab and Placebo. The volcano plots of all four group comparisons are shown in Figure 5.1. Volcano plots are scatterplots which show the log2 fold changes on the x-axis and the -log10 p-values of all hypothesis tests conducted on the y-axis. Genes with a very low or very high log2 fold change and a low p-value are most interesting. Therefore, the genes in the left and right upper corners of the volcano plots are selected with the threshold of

$$|\text{log2 fold change}| > 1 \text{ AND p-value} < 0.05 \tag{5.1}$$

For better comparability between treatments the original p-value rather than the adjusted p-values are used. However, in the toptables in the appendix B the original p-values and the BH adjusted p-values are included. The volcano plots show

that Infliximab has the largest differences in gene expression between responders and non-responders. P-values of Infliximab are higher than p-values of Golimumab, even though there are more samples for Golimumab. The low fold changes in Golimumab might be explained by the slightly different definition of the response variable for the two datasets included. Here, a separate DE analysis might have been a better approach. For placebo both, fold changes and p-values are low and only two genes pass the threshold (see equation 5.3). This indicates that there is no general non progression signature, but rather treatment specific response signatures before treatment. DE of Vedolizumab data shows p-values as low as in the placebo analysis, however fold changes are larger and therefore more genes pass the threshold. The low p-values are possibly due to lower sample size in the Vedolizumab data. Another observation that emerges from the volcano plots is that the treatments Infliximab and Vedolizumab show more down regulated genes than up regulated genes in the responder group. This effect is interesting, as this suggest that if certain genes are expressed less at baseline, the patient will more likely respond to these treatments.

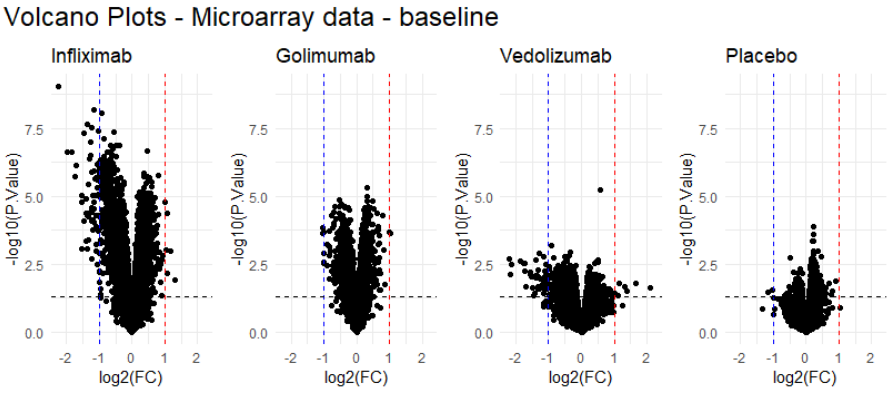



Figure 5.1: *Volcano Plot - Microarray data - baseline*
*The dashed lines mark the gene selection thresholds for p-values and fold changes*

Expressions of genes passing the threshold are visualised in heatmaps in Figure 5.2. Each row contains expressions of one gene and each column expressions of one sample. Genes and samples are ordered according to unsupervised hierarchical clustering. The according dendrograms are shown at the border of the heatmap. Moreover, study and response dimensions are added in the sample annotations. The study annotation shows that there is no clustering by study groups. Differential expression analysis worked best for Infliximab. The heatmap shows one branch in the dendrogram which predominantly contains responders and another which predominantly contains non-responders. The heatmap of the Golimumab analysis shows

three sample clusters, where the cluster in the center contains predominantly responders. In the Vedolizumab heatmap, a clear cluster on the right, which solely contains non-responders is visible. In general, for this analysis detecting non responder signatures is more interesting than responder signatures, because not trying a treatment even though the patient would respond, is more fatal than trying a treatment even though they would not respond. Here again Mayo scores would be interesting to incorporate, to see if this non responder cluster consists of non-responders who did not have much improvement of symptoms compared to baseline.

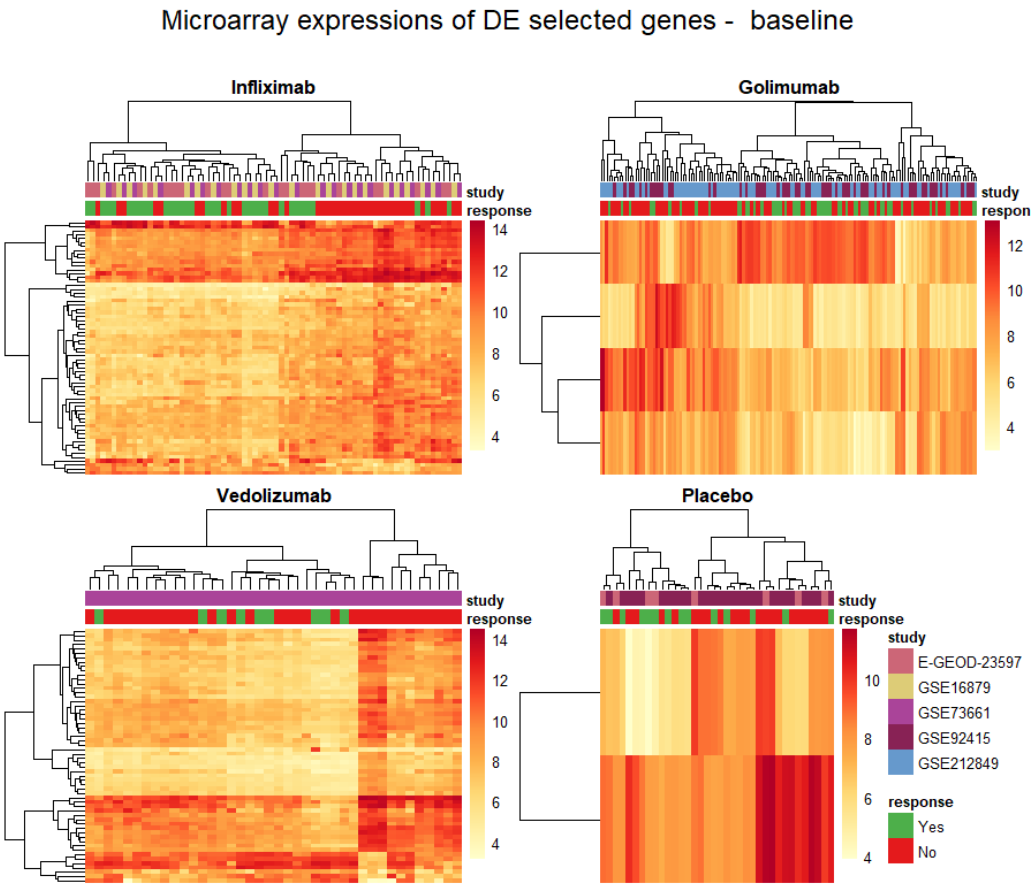



Figure 5.2: *Heatmap of DE selected genes - Microarray data - baseline*

In Figure 5.3, a heatmap of expressions of the union of all genes selected by differential expression is shown. In this plot additional annotations are added. For the columns a treatment information is added and for the rows an indication in which treatment datasets the specific gene was selected as differentially expressed.

The response annotation shows more responders in the cluster on the left than in the cluster on the right. Unfortunately, the Mayo scores at response time are not available for all patients. Especially the scores of non-responder would be interesting, as the group of non-responder might be quite variable. Ranging from patients who did not respond to the treatment at all to patients whose symptoms improved but who just have not met the criterion to be classified as a responder. Over all 291 samples shown in the heatmap, there is one small cluster of samples especially prominent: In the three main branches in the sample dendrogram the branch in the middle. It consists only of non-responders and samples originate from different studies and treatments. The pattern in the heatmap shows that almost all expressions are different to the remaining samples. This cluster would be interesting to analyse further if more patient data was available. Whether these patients have something else in common (e.g. another disease) or this cluster could contain general non responder signatures.

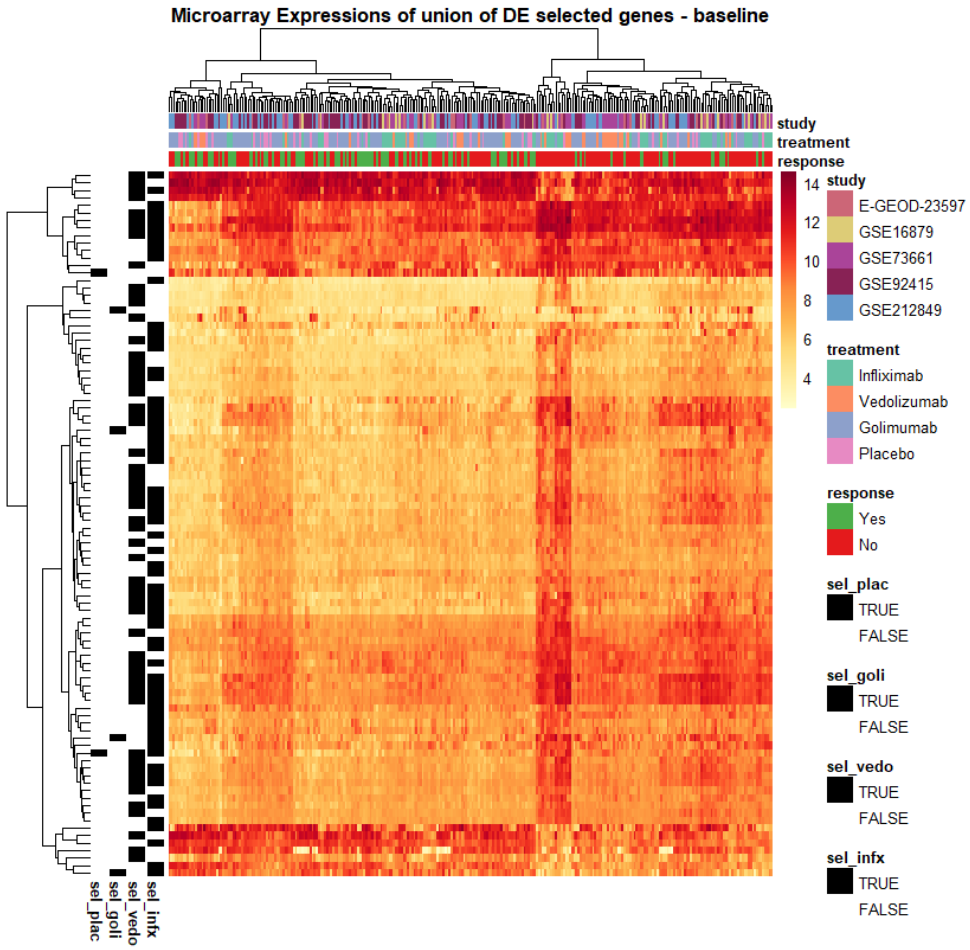



Figure 5.3: *Heatmap of union of DE selected genes - Microarray data - baseline*

### 5.1.2 RNA-Seq

For the RNA-Seq data differential expression between responders and non-responders is calculated for Infliximab and Adalimumab, Etrolizumab and Vedolizumab. Infliximab and Adalimumab originate from the same study (E-MTAB-7604). There are only two responder and six non responder samples of Infliximab and four responder and seven non responder samples of Adalimumab. The two treatments are not split due to the low sample size and because both treatments target the same cytokine TNF$\alpha$. Differences and similarities of the treatments are elaborated in section 2.1.3. The results of Vedolizumab are not shown due to incorrect data (see section 5.1.4).

Figure 5.4 shows the volcano plots of the RNA-Seq data. The fold changes in Etrolizumab are very low and only few genes pass the threshold. Genes are selected with the same thresholds for log2 fold change and p-value as in the microarray analysis:

$$|\text{log2 fold change}| > 1 \text{ AND p-value} < 0.05 \tag{5.2}$$

The volcano plot of Infliximab and Adalimumab shows larger fold changes. The effect of more down than up regulated genes in responder seen in the Infliximab results of microarray data, is a bit weaker in the RNASeq data but still visible. While a couple of genes have very high fold changes, the left arm of the volcano plot is larger. There are more genes below a log2 fold change of -1 than there are genes above a log2 fold change of 1.

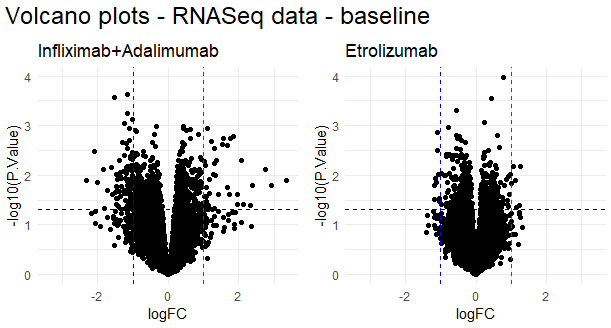



Figure 5.4: *Volcano Plot - RNA-Seq data - baseline*
*The dashed lines mark the gene selection thresholds for p-values and fold changes*

Expressions of genes passing this threshold are visualised in heatmaps in Figure 5.5. In both treatments a cluster of non-responders is detectable. In the TNF$\alpha$ inhibitor group, the cluster on the left contains all but one response samples and

only one no response sample. In the Etrolizumab dataset, only 12 out of 70 samples are from responders. Most responder samples are clustered together in the heatmap.

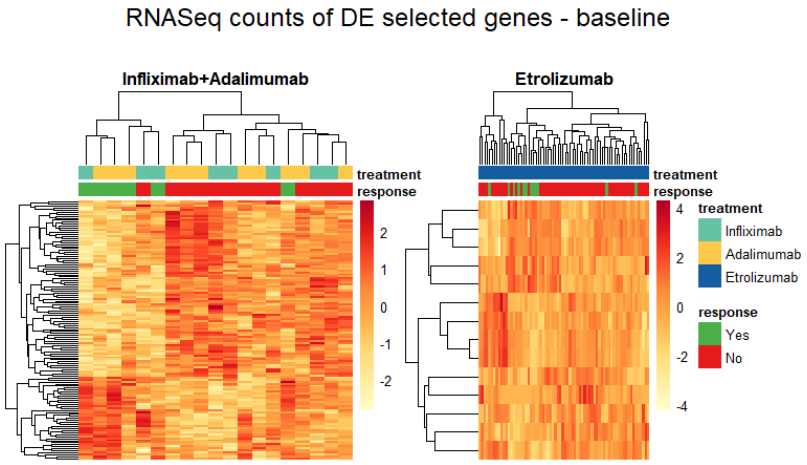



Figure 5.5: *Heatmap of DE selected genes - RNA-Seq data - baseline Values are scaled by row for better readability*

### 5.1.3 Comparing all differential expression results

To compare all differential expression results, log2 fold changes are plotted against each other in a scatterplot matrix 5.6. The figure also includes distributions of log2 fold changes on the diagonal and Pearson correlations between all combinations. For each gene the log2 fold change is the difference between the average expressions of responders and non-responders on a log2 scale. E-MTAB-7845 stands out with throughout negative correlations between its fold changes and fold changes of other datasets. Correlations between fold changes of Vedolizumab RNA-Seq and Vedolizumab Microarray are especially interesting. They have the lowest correlation over all, with -0.592 and are clearly indirectly proportional. The corresponding scatterplot shows: Many underrepresented genes in the responder group in the microarray data are overrepresented in the responder group in the RNA-Seq data. And the other way around, many overrepresented genes in the responder group in the microarray data are underrepresented in the responder group in the RNA-Seq data. This led to further investigations of the dataset E-MTAB-7845 (see section 5.1.4). All other datasets have positively correlated fold changes. Especially the microarray data shows high positive correlations between treatments. In the last row, the scatterplots of the placebo group fold changes are shown. The fold changes in the

placebo group are in general very low, and (except for E-MTAB-7845) fold changes from other treatments are positively correlated with the placebo results.

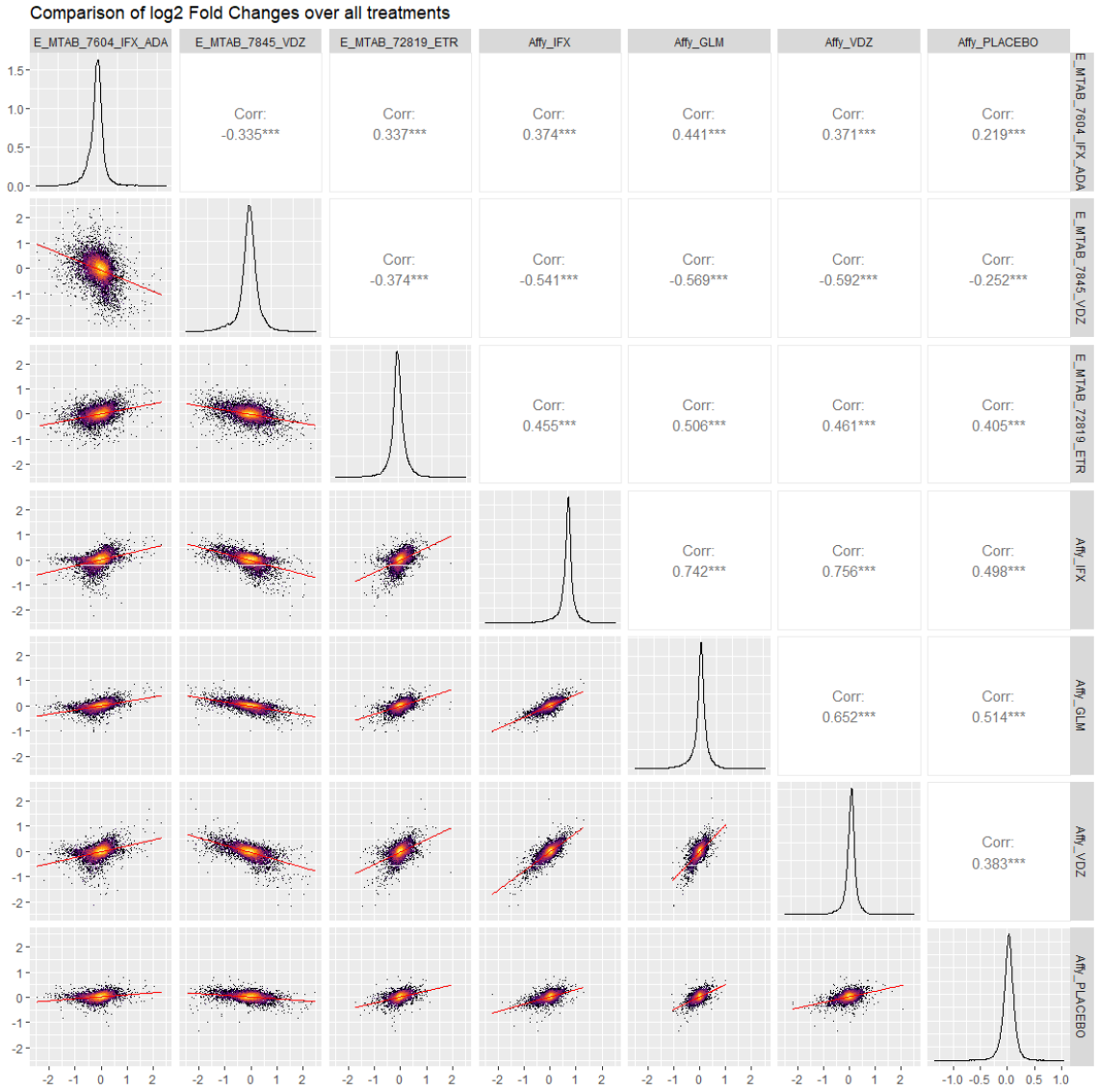



Figure 5.6: *Scatterplot and correlation matrix of log2 fold changes*
*The red lines in the scatterplots are linear regressions fit to the depicted log2 fold changes. Moreover, the density of dots is illustrated by a color scale where yellow represents the highest density and black the lowest*

### 5.1.4 Exclusion of E-MTAB-7845

The comparison of log2 fold changes of responders and non-responders in Figure 5.6 shows that the Vedolizumab RNA-Seq data (E-MTAB-7845) follows an opposite trend as all other treatment response analysis groups. This initiated a plausibility analysis of E-MTAB-7845. The authors of the data, Verstockt et al. analysed the same data in their publication [78]. Reproducing the results from the paper was tried with the data downloaded from ArrayExpress. On the one hand, fold changes reported in the paper were compared with fold changes from this analysis. On the other hand, counts from the beginning of the pipeline were compared (see Figure A.1 in the appendix). The boxplots of 5 specific genes grouped by response show a major difference between counts in the two response groups in the data used in the publication and the data downloaded from ArrayExpress. Thus, the data neither coincides with the publication nor with the other datasets from this meta analysis. This suggests that the sample data (including the response variable) in ArrayExpress is not correctly assigned to the RNA-Seq fastq files. I contacted the authors and have not yet received a reply. Therefore, this dataset is excluded from the results.

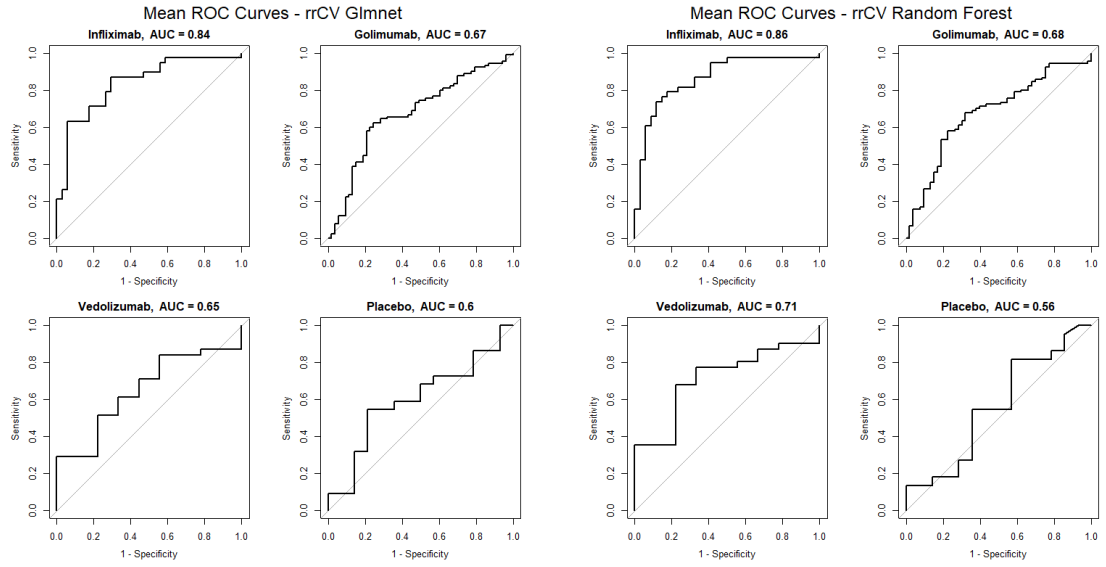
## 5.2 Response Prediction

A complementary approach to the univariate gene-wise group comparison are multivariate machine learning models. For microarray data, models are built and evaluated using rrCV (see section 4.4). For RNA-Seq data, sample sizes - especially when splitting them into response groups - are not large enough to effectively filter predictive genes from a large set of features. Therefore, the data was used to evaluate the genes found with the microarray data.

### 5.2.1 Microarray

For each treatment, features were prefiltered using the inter 5-95 quantile range as described in section 4.4.1. All genes with an inter 5-95 quantile range larger than 3.5 were forwarded to the modeling step with rrCV (see section 4.4.2). For Infliximab 164, for Vedolizumab 204, for Golimumab 192 and for Placebo 164 genes passed this threshold. RrCV was performed with 10 test repeats, 5 test folds, 5 train repeats and 5 train folds. This returned 50 models and each one of these models was evaluated on a test fold it has not seen before. The hyperparameter values tried for the glmnet were $\alpha = 0, 0.2, 0.4, 0.6, 0.8, 1$ and $\lambda = 0.5, 0.6, 0.7, 0.8, 0.9, 1$ and for the random forest $mtry = 2, 20, 38, 56, 74, 92, 110, 128, 146, 164$. Both, glmnet and random forest

optimised the area under the ROC curve. The ROC curves resulting from the test datasets were averaged to one ROC curve and printed in Figure 5.7. The glmnet and the random forest show similar results. With both models, the prediction of the response to Infliximab worked best, with an AUC of 0.84 and 0.86 respectively. Golimumab achieved an AUC of 0.67 in the glmnet and 0.68 in the random forest rrCV. Prediction of the response to Vedolizumab was also difficult. The AUC of the glmnet is 0.65 and of the random forest 0.71. For placebo, response was not possible to predict (AUC 0.6 and 0.56).




(a) *Glmnet* (b) *Random Forest*

Figure 5.7: *ROC curves from rrCV glmnet and rrCV random forest*

The distributions of variable importances over all models are shown in Figure 5.8 for glmnet and Figure 5.9 for random forest. IL13RA2 (interleukin 13 receptor subunit alpha 2) has a significant role for the response prediction of Infliximab. This gene was found useful to predict response to Infliximab in studies that used subsets of the data as used in this analysis [70], but also in studies that used different data [80]. In half of the models returned from rrCV it is the gene with a variable importance of 100% i.e. the most important variable. Other genes important for the prediction of response to Infliximab are TNFRSF11B (TNF receptor superfamily member 11b), STC1 (Stanniocalcin 1), WNT5A (Wnt Family Member 5A) and CCL2 (C-C motif chemokine ligand 2). For Golimumab, which is also a TNF$\alpha$ blocker, IL13RA was one of the 4 genes selected in DE analysis and it appeared as 24th most important

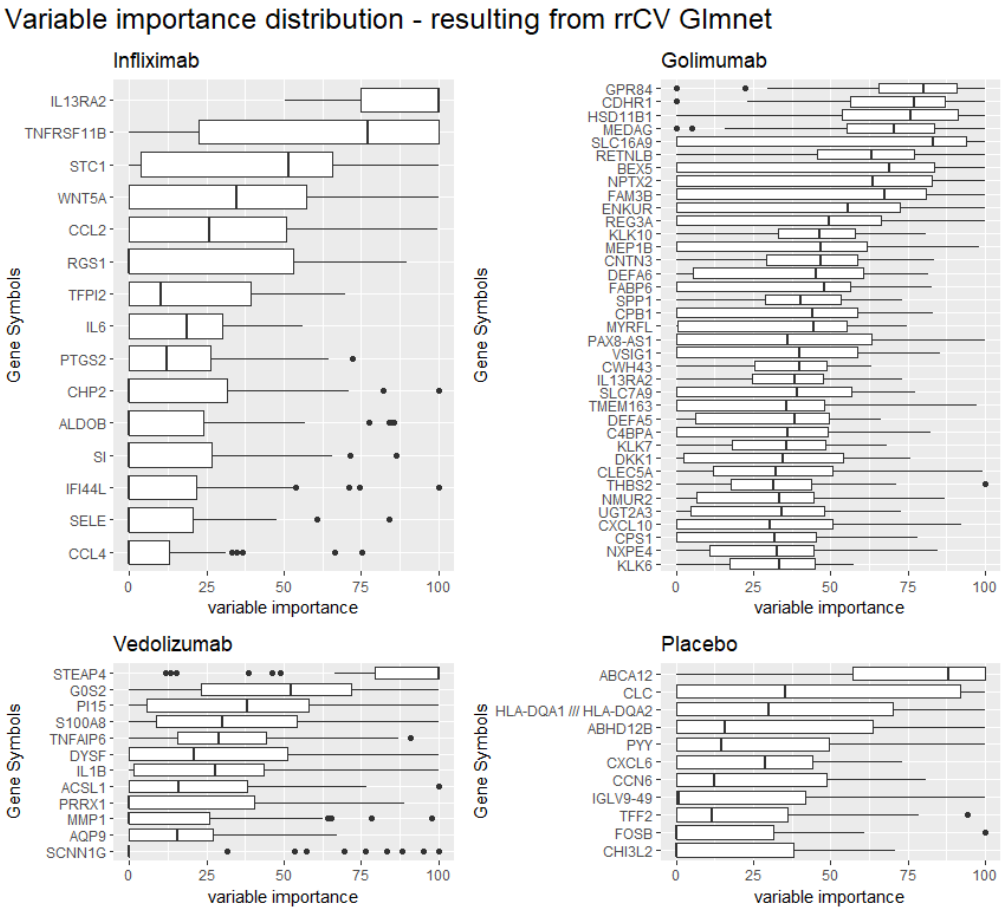



Figure 5.8: *Distribution of variable importances in glmnet prediction models*

gene in prediction analysis with the glmnet.

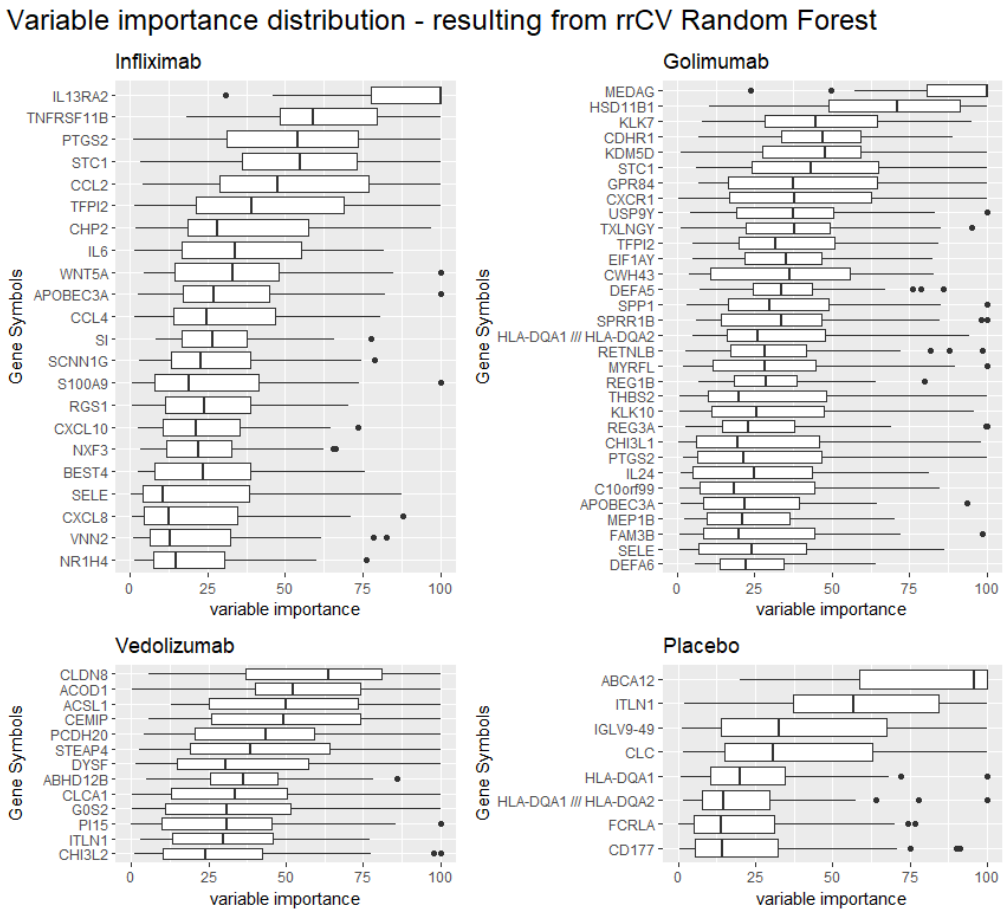



Figure 5.9: *Distribution of variable importances in random forest prediction models*

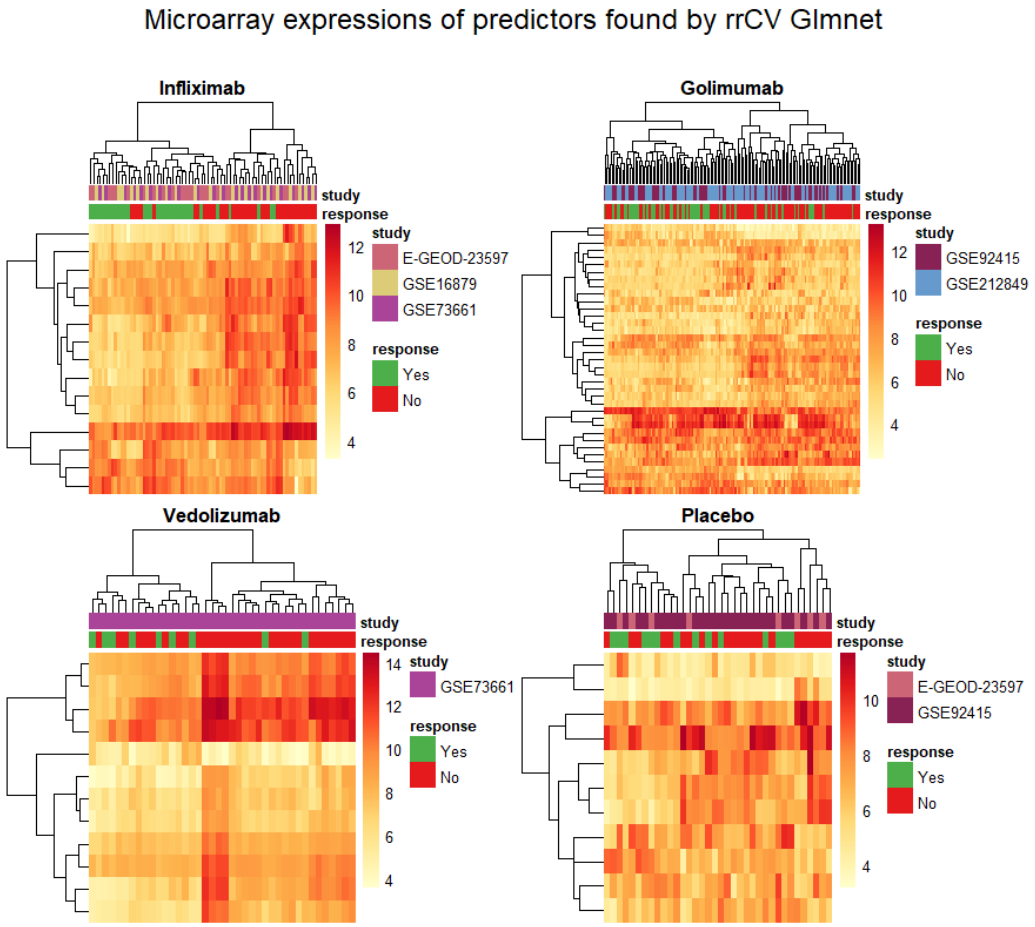



Figure 5.10: *Heatmap of genes selected by variable importance of prediction models*

Heatmaps of the genes with the highest variable importance are shown in Figure 5.10. The heatmap of the most important genes in Infliximab shows a nice clustering of responders and non-responders. In the multivariate setting, more predictors were found for Golimumab than in the univariate setting. In the heatmaps, one of the two main clusters contains more responders and the other contains more non-responders. The most important genes in the prediction of Golimumab response are GPR84(G protein-coupled receptor 84), CDHR1(cadherin related family member 1), HSD11B1(hydroxysteroid 11-beta dehydrogenase 1), MEDAG(mesenteric estrogen dependent adipogenesis) and SLC16A9(solute carrier family 16 member 9).

The prediction models are able to find some genes that were also found by differential expression analysis. Figure 5.11 shows overlaps of genes found with differential

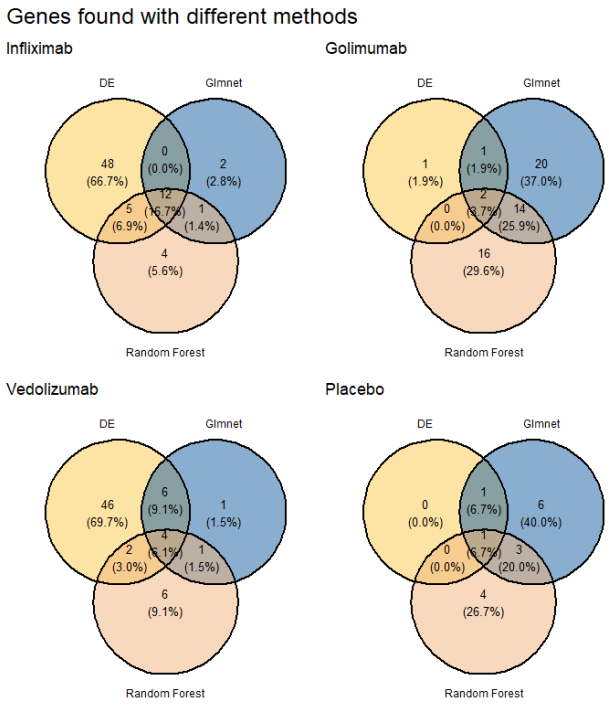



Figure 5.11: *Overlaps between genes found with different methods*

expression, rrCV glmnet and rrCV random forest. There are 12 genes overlapping for all three methods on the Infliximab data. For Golimumab, three out of four genes found with DE were also found by prediction models. For Vedolizumab, twelve genes found with DE were also found with glmnet or random forest and all DE selected genes in the placebo group were found by glmnet. DE and prediction independently selected the most important genes in the data starting from about $17,000$ genes. These large overlaps are remarkable and that this many genes were found with both of these distinct methods is a proof for both methods.

### 5.2.2 RNA-Seq

The RNASeq datasets with the treatments Infliximab+Adalimumab contains only 19 observations and the smaller responder group consists of only 6 observations. Due to these low sample sizes, the dataset was not used to find new connections, but to evaluate genes found with the Infliximab microarray data. To evaluate genes found for the response to Vedolizumab on the microarray data, originally the RNASeq Vedolizumab dataset E-MTAB-7845 was planned to be used. Unfortunately this dataset had to be excluded from the analysis (see section 5.1.4). Instead, the RNA-Seq Etrolizumab data is used. Etrolizumab targets $\alpha 4\beta 7$ and $\alpha E\beta 7$ integrins while Vedolizumab targets the $\alpha 4\beta 7$ integrin only. The intersection of genes found with differential expression and glmnet prediction from the microarray data were evaluated by fitting and evaluating glmnets with leave one out cross validation (LOOCV) to the new data. The heatmaps in Figure 5.12 show counts of the genes evaluated. The resulting ROC curves are shown in Figure 5.13. The ROC curve on the left side shows that these 12 genes ("IL13RA2","TNFRSF11B", "STC1", "WNT5A", "IL6", "PTGS2", "CCL2", "TFPI2", "SELE", "RGS1", "CCL4", "CHP2") can be used to predict response to TNF$\alpha$ blockers Infliximab and Adalimumab. Table 5.3 shows the confusion matrices split into the two treatment groups. For Infliximab, all but one sample were classified correctly while Adalimumab had four misclassifications. This yields an accuracy of 0.875 for Infliximab and 0.63 for Adalimumab.

| | predicted non-responder | predicted responder |
|---|---|---|
| observed non-responder | 5 | 1 |
| observed responder | 0 | 2 |

(a) *Infliximab*

| | predicted non-responder | predicted responder |
|---|---|---|
| observed non-responder | 6 | 1 |
| observed responder | 3 | 1 |

(b) *Adalimumab*

Table 5.3: *Confusion matrices prediction*

The genes TNFRSF11B, STC1, PTGS2 and IL12RA2 are part of the 5 gene signature found in by Arijs et. al. [5]. In their study they used two datasets (GSE14580 and GSE12251) that were also used in this analysis (see table 3.1). The same genes appear in the top ranking in their and in my univariate as well as multivariate analysis even though in my analysis a third microarray dataset was included. This is on the one hand a proof of concept and on the other hand shows that results are reproducible. Additionally these genes perform well in the evaluation on an independent RNA-Seq dataset, which proves the robustness of these genes. The experiment of

trying to predict the response to Etrolizumab with the results from the microarray Vedolizumab results did not work. The ROC curve 5.13 shows that no classification threshold for the prediction of response separates the two groups well.

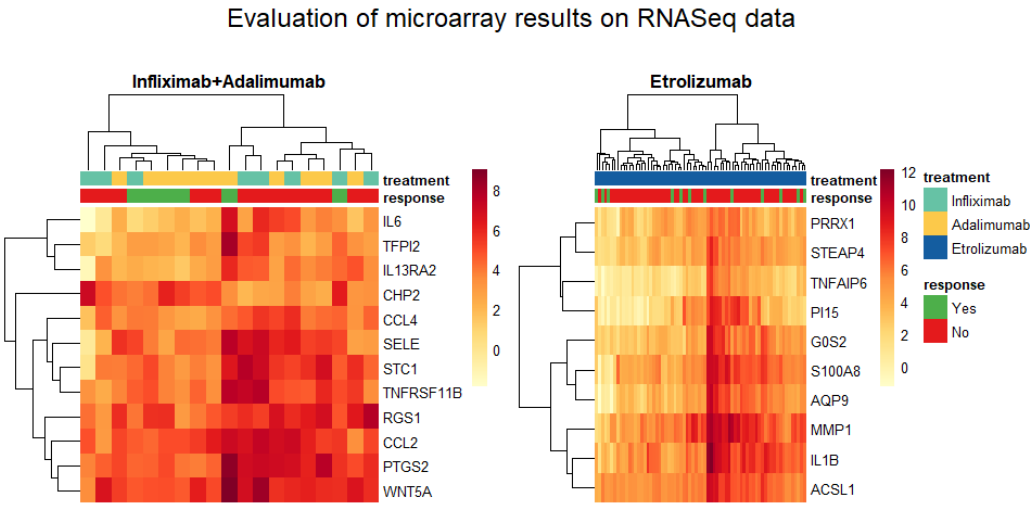



Figure 5.12: *Heatmap of RNA-Seq count values of genes found in the Microarray differential expression and prediction analysis*

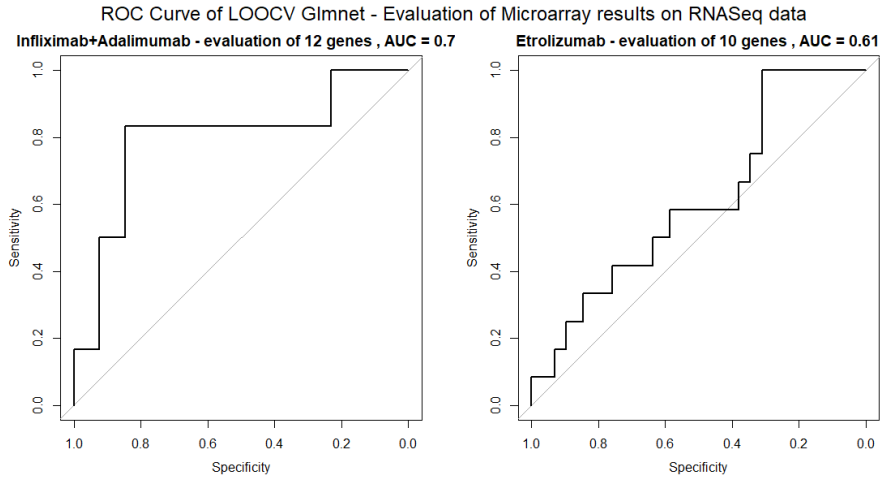



Figure 5.13: *ROC curves of genes found with Microarray data evaluated on RNA-Seq data*

## 5.3 Gene expression change during treatment

The third analysis focuses on the change of gene expression during therapy between responders and non-responders. Given one sample at baseline (before treatment) and - of the same patient - one sample at follow up (after treatment), the differences of these two samples can be compared in responders and non-responders. Differential expression is calculated with the "follow up - baseline" differences. Volcano plots, heat maps and overlaps of selected genes summarise the results. Moreover, the toptables of the analysis are added in the appendix C. This analysis shows the effect the treatment has on gene expression as well as the effect the healing process has on gene expression.

The sample sizes of the data used in this analysis are summarised in table 5.2. Using the quantile normalised and batch corrected data, the difference of expression values "follow up - baseline" is calculated for each patient. On these differences, gene wise hypothesis tests are applied (see section 4.3). A positive fold change means that over all patients the expression of this gene increased more in the responder group than in the non responder group during treatment. Accordingly, a negative fold change can be interpreted as a larger decrease of expressions during treatment in the responder group compared to the non responder group. The volcano plots from this analysis (Figure 5.14) show larger -log10 p-values and more extreme log2 fold changes than the previous analysis. From this follows that the treatment and healing effect is stronger than the effect of response before treatment. Infliximab shows the greatest effect, followed by the Vedolizumab results. Here also the placebo group shows differentially expressed genes. This is not surprising, because in this data on the one hand the placebo effect and on the other hand the change of gene expression during healing is captured.

Genes are selected with the same thresholds for log2 fold changes and p-values as in the before treatment differential expression analysis:

$$|\text{log2 fold change}| > 1 \text{ AND p-value} < 0.05 \tag{5.3}$$

The "follow up - baseline" expression differences of the selected genes are presented in heatmaps in Figure 5.16. The hierarchical clustering of samples shows three main clusters for Infliximab, where the one on the right predominantly contains responder and the others predominantly contains non responder. In Golimumab, a nice clustering of responder is visible in the left branch of the dendrogram.

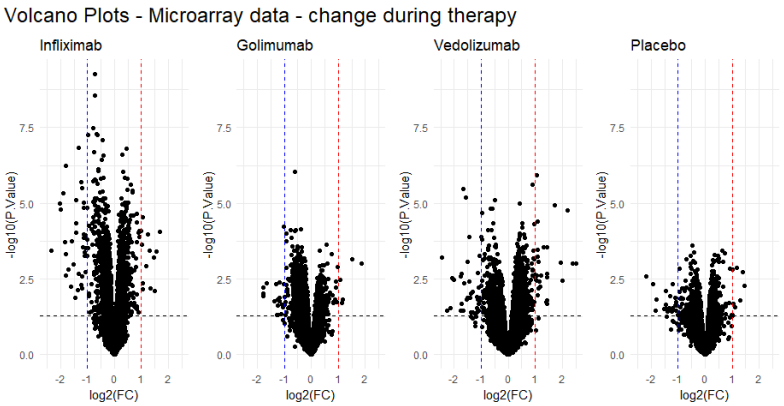



Figure 5.14: *Volcano Plot - Microarray - change during treatment*
*The dashed lines mark the gene selection thresholds for p-values and fold changes*

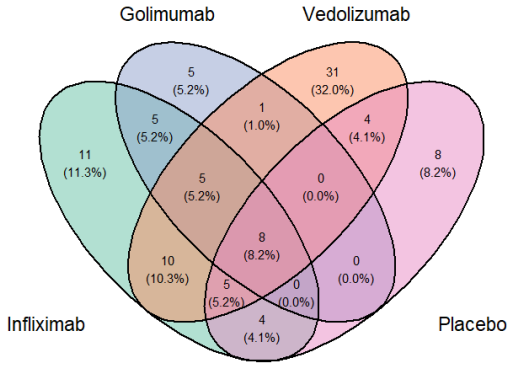



Figure 5.15: *Overlapps of genes found for the change during treatment for different treatments*

The Venn diagram in Figure 5.15 shows large overlaps of selected genes between treatments. 21 out of 25 genes found in the placebo group were also found in at least one other treatment. 8 genes have been selected for all treatments. This shows the overall healing effect rather than treatment specific effects.

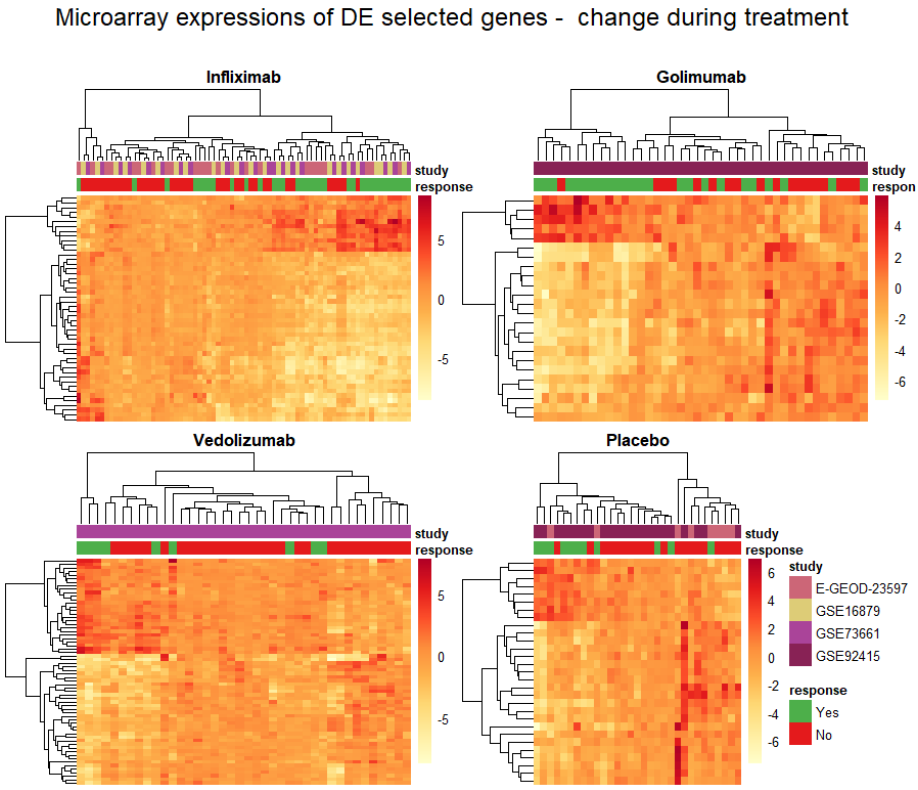



Figure 5.16: *Heatmap of DE selected genes - Microarray - change during treatment Colors represent the differences follow up - baseline*

Figure 5.17 shows the union of DE selected genes except for the genes selected for placebo. This highlights the treatment effect rather than the healing effect. In the heatmap two large sample clusters are visible. The cluster on the left mainly contains responders and the cluster on the right mainly contains non-responders. Looking at the patterns inside the heatmap, four squares are detectable: Two clear sample groups and two clear gene groups.

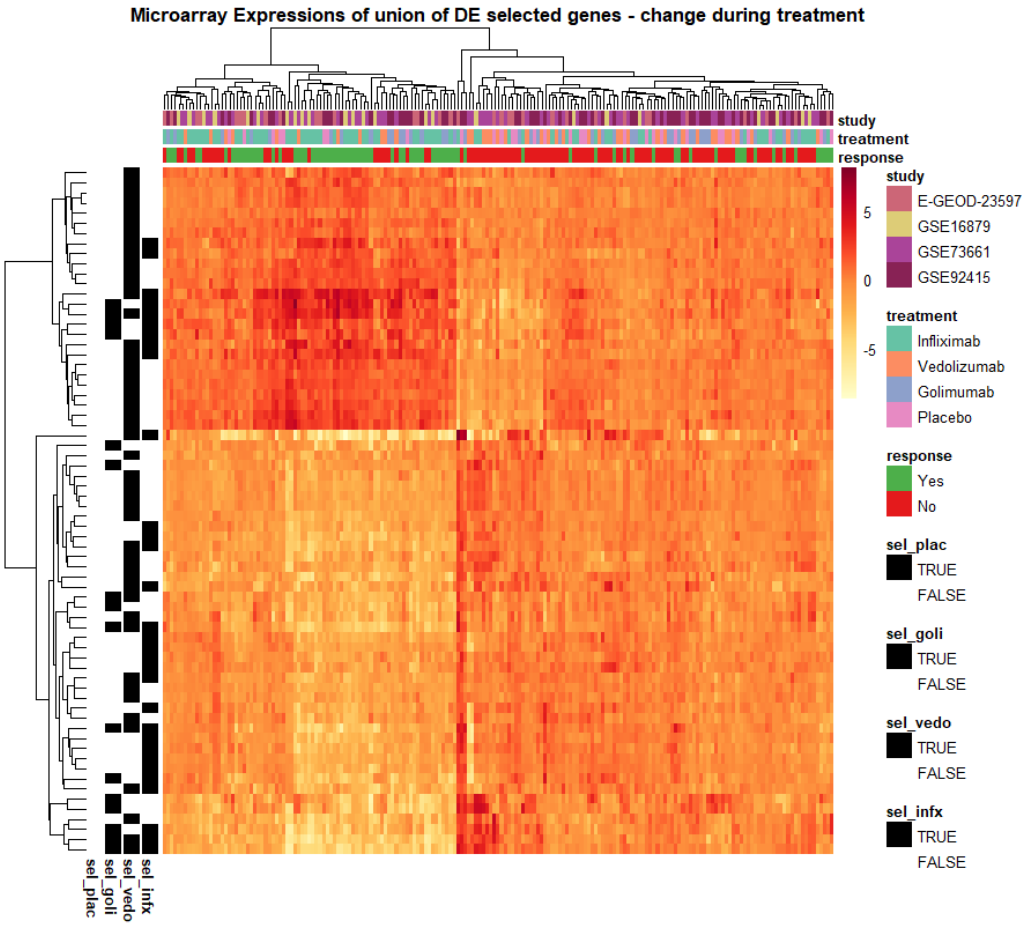



Figure 5.17: *Heatmap of union of DE selected genes - Microarray data - change during treatment*
*The union of DE selected genes without the genes selected in the placebo group are shown.*

# Chapter 6

# Conclusion

This thesis' aim was to summarise open gene expression data on treatment response in UC. Moreover, it was focused on novel treatments consisting of monoclonal antibodies. One part of the analysis dealt with data collected before treatment and the other part dealt with the change of gene expression before and after treatment. Throughout the thesis, samples of responder and non responder of each treatment were compared. This comparison was approached by univariate differential expression analysis and multivariate machine learning models.

Eight datasets with the accession numbers E-GEOD-23597, E-GEOD-72819, E-MTAB-7604, E-MTAB-7845, GSE16879, GSE73661, GSE92415, GSE212849 from gene expression databanks were found suitable. They all contain gene expression data of colon tissue before and sometimes after treatment of some monoclonal antibody treatment. The treatment groups are Infliximab, Golimumab, Adalimumab, Vedolizumab, Etrolizumab and Placebo.

A large part of the analysis consisted of preprocessing the datasets starting with raw data, such that all datasets are preprocessed in the same way. Microarray and RNA-Seq had to be preprocessed separately and differently, because the data is fundamentally different. At the very end of the analysis, results of differential expression were compared over all treatments and microarray/RNA-Seq methods. This comparison has revealed that one dataset (E-MTAB-7845) behaves differently than the others. After digging deeper, it turns out that there must be an error in the data in ArrayExpress. Meta analysis like this thesis depend on publicly available data and such problems are a drawback on such meta-analyses. Another limitation is that not much patient data was available in the databanks. For instance mayo scores were not available in the majority of samples which restricted this analysis to only consider a binary response variable that was defined by the datasets authors.

Fortunately, the criteria for the inclusion of patients in the studies and the response variable are defined the same or similarly in almost all studies. Heatmaps of a set of selected genes often show some separation between responders and non responders, but this separation is not 100%. It would be interesting to include the mayo scores, if they were available. A patient whose symptoms improved after treatment but who just not met the criterion to be classified as responder might have different gene expressions than a patient whose symptoms did not improve at all over treatment, but both are classified as non responder. These differences had to be neglected due to missing information. Perhaps, a continuous score such as the mayo score might be better to model treatment response or at least useful for the interpretation of the results. Gene expression data analysis often lacks in reproducibility. Especially for low sample sizes a split into two groups might not only measure the effect between the groups it was split into e.g. response and non response but also other invisible or unknown effects that distinguished these two groups of patients. The less known about the patients the less can be corrected for other variables in the models.

Another observation made in the course of this analysis was that machine learning models require more sample sizes than differential expression and if they should be trained and evaluated on different data, a decent amount of data per group is required. The prediction on microarray Infliximab data worked well with 34 response and 38 non-response observations. Lower sample sizes might work as well if effect sizes are large enough.

For Infliximab, samples from four different datasets were available. Infliximab is the first biologic used to treat UC and is therefore the most explored treatment in this context. In this thesis, robust genes to predict the response to Infliximab before treatment were found independently by univariate differential expression analysis and multivariate prediction analysis on microarray data. These twelve genes "IL6", "TFPI2", "IL13RA2", "CHP2", "CCL4", "SELE", "STC1", "TNFRSF11B", "RGS1", "CCL2", PTGS2" and "WNT5A" were evaluated on an RNA-Seq dataset. The models built with these genes showed good performance on independent observations with an AUC = 0.7. Some of these genes are known to be part of inflammatory processes and have been found in the context of response to Infliximab in Ulcerative Colitis. This thesis' results are further analysed with pathway and network analysis to extract more meaning of the results. Moreover the results will function as reference for the new data collected in course of the EU Immuniverse project.

# Bibliography


[1] Harvard Chan Biognmeninformatics Core (HBC). *Introduction to RNA-Seq using high-performance computing - ARCHIVED*. Accessed: 2023-25-02. URL: %5Curl%7Bhttps://hbctraining.github.io/Intro-to-rnaseq-hpc-02/lessons/03_alit.html%7D.

[2] Affymetrix. "GeneChip - Expression Analysis Technical Manual". In: (2005-2009).

[3] Ingrid Arijs et al. "Effect of vedolizumab (anti-$\alpha 4\beta 7$-integrin) therapy on histological healing and mucosal gene expression in patients with UC". In: *Gut* 67.1 (2018), pp. 43–52.

[4] Ingrid Arijs et al. "Mucosal gene expression of antimicrobial peptides in inflammatory bowel disease before and after first infliximab treatment". In: *PloS one* 4.11 (2009), e7984.

[5] Ingrid Arijs et al. "Mucosal gene signatures to predict response to infliximab in patients with ulcerative colitis". In: *Gut* 58.12 (2009), pp. 1612–1619.

[6] Ingrid Arijs et al. "Mucosal gene signatures to predict response to infliximab in patients with ulcerative colitis". In: *Gut* 58.12 (2009), pp. 1612–1619.

[7] Simon P Bach and Neil J Mortensen. "Ileal pouch surgery for ulcerative colitis". In: *World journal of gastroenterology: WJG* 13.24 (2007), p. 3288.

[8] Yoav Benjamini and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995), pp. 289–300.

[9] Alvis Brazma et al. "Minimum information about a microarray experiment (MIAME)—toward standards for microarray data". In: *Nature genetics* 29.4 (2001), pp. 365–371.

[10] Kassenärztliche Bundesvereinigung. "Biologika - Colitis Ulcerosa". In: 4 (2020).

[11] Johan Burisch et al. "The burden of inflammatory bowel disease in Europe". In: *Journal of Crohn's and Colitis* 7.4 (2013), pp. 322–337.

[12] Rebecca Carey et al. "Activation of an IL-6: STAT3-dependent transcriptome in pediatric-onset inflammatory bowel disease". In: *Inflammatory bowel diseases* 14.4 (2008), pp. 446–457.

[13] Benilton S Carvalho and Rafael A Irizarry. "A Framework for Oligonucleotide Microarray Preprocessing". In: *Bioinformatics* 26.19 (2010), pp. 2363–7. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btq431.

[14] Luca Cozzuto, Julia Ponomarenko, and Sarah Bonnin. *Biocore RNAseq course 2019 https://biocorecrg.github.io/RNAseq_course_2019/*. Accessed: 2023-25-02.

[15] Baumgart Daniel C. et al. *Biologic Therapy for Inflammatory Bowel Disease*. UNI-MED Verlag AG, 2017.

[16] DCCV e.V. Deutsche Morbus Crohn / Colitis ulcerosa Vereinigung et al. *Chronisch entzündliche Darmerkrankungen*. S. Hirzel Verlag Stuttgart, 2006.

[17] Alexander Dobin. *STAR manual 2.7.6a*. Accessed: 2023-26-02. 2020. URL: %5Curl%7Bhttps://gensoft.pasteur.fr/docs/STAR/2.7.6a/STARmanual.pdf%7D.

[18] Alexander Dobin et al. "STAR: ultrafast universal RNA-seq aligner". In: *Bioinformatics* 29.1 (2013), pp. 15–21.

[19] Sorin Draghici. *Statistics and data analysis for microarrays using R and bioconductor*. CRC Press, 2016.

[20] Marla C Dubinsky et al. "Ulcerative Colitis Narrative Global Survey Findings: The Impact of Living With Ulcerative Colitis—Patients' and Physicians' View". In: *Inflammatory bowel diseases* 27.11 (2021), pp. 1747–1755.

[21] Steffen Durinck et al. "BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis". In: *Bioinformatics* 21.16 (2005), pp. 3439–3440.

[22] EBI. *ArrayExpress*. Accessed: 2023-08-03. URL: %5Curl%7Bhttps://www.ebi.ac.uk/biostudies/arrayexpress%7D.

[23] Stange Eduard F. *Ulcerative colitis - Crohn's disease*. UNI-MED Verlag AG, 2017.

[24] EMA. "Humira Product Information: https://www.ema.europa.eu/en/documents/product-information/humira-epar-product-information_de.pdf". In: ().

[25] EMA. "Remicade Product Information: https://www.ema.europa.eu/en/documents/product-information/remicade-epar-product-information_de.pdf". In: ().

[26] EMA. "Simponi Product Information: https://www.ema.europa.eu/en/documents/product-information/simponi-epar-product-information_de.pdf". In: ().

[27] Ciaran Evans, Johanna Hardin, and Daniel M Stoebel. "Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions". In: *Briefings in bioinformatics* 19.5 (2018), pp. 776–792.

[28] Philip Ewels et al. "MultiQC: summarize analysis results for multiple tools and samples in a single report: https://multiqc.info/". In: *Bioinformatics* 32.19 (2016), pp. 3047–3048.

[29] Thomas L Fare et al. "Effects of atmospheric ozone on microarray data quality". In: *Analytical chemistry* 75.17 (2003), pp. 4672–4675.

[30] *FastQC*. Accessed: 2023-25-02. June 2015. URL: %5Curl%7Bhttps://qubeshub.org/resources/fastqc%7D.

[31] GEO. *Gene Expression Omnibus: Overview = https://www.ncbi.nlm.nih.gov/geo/info/overview.html*. Accessed: 2022-12-12.

[32] Subrata Ghosh and Rod Mitchell. "Impact of inflammatory bowel disease on quality of life: Results of the European Federation of Crohn's and Ulcerative Colitis Associations (EFCCA) patient survey". In: *Journal of Crohn's and Colitis* 1.1 (2007), pp. 10–20.

[33] Hinrich Gohlmann and Willem Talloen. *Gene expression studies using Affymetrix microarrays*. Chapman and Hall/CRC, 2009.

[34] Atle van Beelen Granlund et al. "Whole genome gene expression meta-analysis of inflammatory bowel disease colon mucosa demonstrates lack of major differences between Crohn's disease and ulcerative colitis". In: *PloS one* 8.2 (2013), e56818.

[35] W. Huber et al. "Orchestrating high-throughput genomic analysis with Bioconductor". In: *Nature Methods* 12.2 (2015), pp. 115–121. URL: http://www.nature.com/nmeth/journal/v12/n2/full/nmeth.3252.html.

[36] Wolfgang Huber et al. "Orchestrating high-throughput genomic analysis with Bioconductor". In: *Nature methods* 12.2 (2015), pp. 115–121.

[37] Jeffrey S Hyams et al. "Clinical and biological predictors of response to standardised paediatric colitis therapy (PROTECT): a multicentre inception cohort study". In: *The Lancet* 393.10182 (2019), pp. 1708–1720.

[38] Illumina. "An introduction to Next-Generation Sequencing Technology". In: (2017).

[39] Illumina. *Illumina Sequencing by Synthesis*. Illumina. 2017. URL: https://www.youtube.com/watch?v=fCd6B5HRaZ8&t=1s.

[40] Illumina. *Illumina Support: STAR*. Accessed: 2023-25-02. URL: %5Curl%7Bhttps://support.illumina.com/help/BS_App_RNASeq_Alignment_OLH_1000000006112/Content/Source/Informatics/STAR_RNAseq.htm%7D.

[41] Rafael A Irizarry et al. "Exploration, normalization, and summaries of high density oligonucleotide array probe level data". In: *Biostatistics* 4.2 (2003), pp. 249–264.

[42] W Evan Johnson, Cheng Li, and Ariel Rabinovic. "Adjusting batch effects in microarray expression data using empirical Bayes methods". In: *Biostatistics* 8.1 (2007), pp. 118–127.

[43] Boyko Kabakchiev et al. "Gene expression changes associated with resistance to intravenous corticosteroid therapy in children with severe ulcerative colitis". In: *PLoS One* 5.9 (2010), e13085.

[44] Ali Khodadadian et al. "Genomics and transcriptomics: the powerful technologies in precision medicine". In: *International Journal of General Medicine* 13 (2020), p. 627.

[45] Bernd Klaus. "An end to end workflow for differential gene expression using Affymetrix microarrays". In: *F1000Research* 5 (2016).

[46] Max Kuhn. *caret: Classification and Regression Training*. R package version 6.0-93. 2022. URL: https://CRAN.R-project.org/package=caret.

[47] Charity W Law et al. "voom: Precision weights unlock linear model analysis tools for RNA-seq read counts". In: *Genome biology* 15.2 (2014), pp. 1–17.

[48] Ian C Lawrance, Claudio Fiocchi, and Shukti Chakravarti. "Ulcerative colitis and Crohn's disease: distinctive gene expression profiles and novel susceptibility candidate genes". In: *Human Molecular Genetics* 10.5 (2001), pp. 445–456.

[49] Bryan Linggi et al. "Meta-analysis of gene expression disease signatures in colonic biopsy tissue from patients with ulcerative colitis". In: *Scientific reports* 11.1 (2021), pp. 1–12.

[50] V Lorén et al. "ANP32E, a protein involved in steroid-refractoriness in ulcerative colitis, identified by a systems biology approach". In: *Journal of Crohn's and Colitis* 13.3 (2019), pp. 351–361.

[51] Barreiro-de Acosta Manuel and Domínguez-Muñoz J. Enrique. "Migratory Movements and the Risk of Inflammatory Bowel Disease". In: *World Gastroenterology News* 20.1 (2015).

[52] Kuhn Max. *The caret package*. Accessed: 2023-26-02. 2019. URL: %5Curl%7Bhttps://topepo.github.io/caret/index.html%7D.

[53] Freissmuth Michael, Offermanns Stefan, and Böhm Stefan. *Pharmakologie und Toxikologie - Von den molekularen Grundlagen zur Pharmakotherapie*. Springer-Verlag, 2016.

[54] NCBI. *Gene Expression Omnibus*. Accessed: 2023-08-03. URL: %5Curl%7Bhttps://www.ncbi.nlm.nih.gov/geo/%7D.

[55] Siew C Ng et al. "Supplement to: Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies". In: *The Lancet* 390.10114 (2017), pp. 2769–2778.

[56] Siew C Ng et al. "Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies". In: *The Lancet* 390.10114 (2017), pp. 2769–2778.

[57] J$\varphi$rgen Olsen et al. "Diagnosis of ulcerative colitis before onset of inflammation by multivariate modeling of genome-wide gene expression data". In: *Inflammatory bowel diseases* 15.7 (2009), pp. 1032–1038.

[58] Masatsugu Ono et al. "Structural basis for tumor necrosis factor blockade with the therapeutic antibody golimumab". In: *Protein Science* 27.6 (2018), pp. 1038–1046.

[59] Hervé Pagès et al. *AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor*. R package version 1.58.0. 2022. URL: https://bioconductor.org/packages/AnnotationDbi.

[60] Polychronis Pavlidis et al. "Cytokine responsive networks in human colonic epithelial organoids unveil a molecular classification of inflammatory bowel disease". In: *Cell Reports* 40.13 (2022), p. 111439.

[61] Thomas Lin Pedersen. *patchwork: The Composer of Plots*. R package version 1.1.2. 2022. URL: https://CRAN.R-project.org/package=patchwork.

[62] Belinda Phipson et al. "Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression". In: *The annals of applied statistics* 10.2 (2016), p. 946.

[63] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2022. URL: https://www.R-project.org/.

[64] Shakespeare R., Green J., and Turner J. *Gastrointestinal Medicine - core science, medicine and surgery in one book*. JP Medical, 2017.

[65] Johannes Rainer. *EnsDb.Hsapiens.v79: Ensembl based annotation package*. R package version 2.99.0. 2017.

[66] Matthew E Ritchie et al. "limma powers differential expression analyses for RNA-sequencing and microarray studies". In: *Nucleic Acids Research* 43.7 (2015), e47. DOI: 10.1093/nar/gkv007.

[67] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". In: *bioinformatics* 26.1 (2010), pp. 139–140.

[68] Mark D Robinson and Alicia Oshlack. "A scaling normalization method for differential expression analysis of RNA-seq data". In: *Genome biology* 11.3 (2010), pp. 1–9.

[69] Paul Rutgeerts et al. "Infliximab for induction and maintenance therapy for ulcerative colitis". In: *New England Journal of Medicine* 353.23 (2005), pp. 2462–2476.

[70] Suraj Sakaram et al. "A Multi-mRNA Prognostic Signature for Anti-TNF$\alpha$ Therapy Response in Patients with Inflammatory Bowel Disease". In: *Diagnostics* 11.10 (2021), p. 1902.

[71] William J Sandborn et al. "Subcutaneous golimumab induces clinical response and remission in patients with moderate-to-severe ulcerative colitis". In: *Gastroenterology* 146.1 (2014), pp. 85–95.

[72] Andrews Simon. *Positional sequence bias in random primed libraries*. Accessed: 2023-26-02. 2016. URL: %5Curl%7Bhttps://sequencing.qcfail.com/articles/positional-sequence-bias-in-random-primed-libraries/%7D.

[73] Gordon K. Smyth et al. "limma: Linear Models for Microarray and RNA-Seq Data User's Guide". In: *Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research* (2015).

[74] FGED Society. *MINSEQE https://www.fged.org/projects/minseqe/*. Accessed: 2022-12-18.

[75] Stanislaw Supplitt et al. "Current achievements and applications of transcriptomics in personalized cancer medicine". In: *International Journal of Molecular Sciences* 22.3 (2021), p. 1422.

[76] Gaik W. Tew et al. "Association Between Response to Etrolizumab and Expression of Integrin E and Granzyme A in Colon Biopsies of Patients With Ulcerative Colitis". In: *Gastroenterology* 150.2 (2016), 477–487.e9. ISSN: 0016-5085. DOI: https://doi.org/10.1053/j.gastro.2015.10.041. URL: https://www.sciencedirect.com/science/article/pii/S0016508515015747.

[77] Gary Toedter et al. "Gene expression profiling and response signatures associated with differential responses to infliximab treatment in ulcerative colitis". In: *Official journal of the American College of Gastroenterology— ACG* 106.7 (2011), pp. 1272–1280.

[78] Bram Verstockt et al. "Expression levels of 4 genes in colon tissue might be used to predict which patients will enter endoscopic remission after vedolizumab therapy for inflammatory bowel diseases". In: *Clinical Gastroenterology and Hepatology* 18.5 (2020), pp. 1142–1151.

[79] Bram Verstockt et al. "Low TREM1 expression in whole blood predicts anti-TNF response in inflammatory bowel disease". In: *EBioMedicine* 40 (2019), pp. 733–742.

[80] Bram Verstockt et al. "Mucosal IL13RA2 expression predicts nonresponse to anti-TNF therapy in Crohn's disease". In: *Alimentary Pharmacology & Therapeutics* 49.5 (2019), pp. 572–581.

[81] Magnus Hofrenning Wanderås et al. "Predictive factors for a severe clinical course in ulcerative colitis: results from population-based studies". In: *World journal of gastrointestinal pharmacology and therapeutics* 7.2 (2016), p. 235.

[82] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN: 978-3-319-24277-4. URL: https://ggplot2.tidyverse.org.

[83] Hadley Wickham. *stringr: Simple, Consistent Wrappers for Common String Operations*. R package version 1.4.0. 2019. URL: https://CRAN.R-project.org/package=stringr.

[84] Hadley Wickham and Maximilian Girlich. *tidyr: Tidy Messy Data*. R package version 1.2.0. 2022. URL: https://CRAN.R-project.org/package=tidyr.

[85] Hadley Wickham, Jim Hester, and Jennifer Bryan. *readr: Read Rectangular Text Data*. R package version 2.1.2. 2022. URL: https://CRAN.R-project.org/package=readr.

[86] Hadley Wickham et al. *dplyr: A Grammar of Data Manipulation*. R package version 1.0.9. 2022. URL: https://CRAN.R-project.org/package=dplyr.

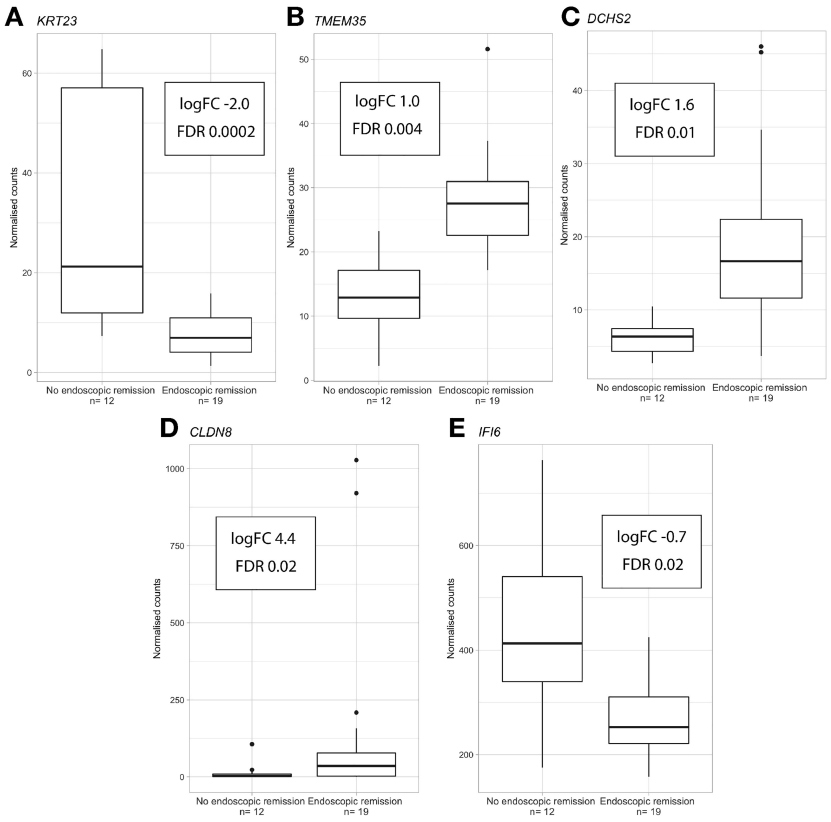
[87] Hadley Wickham et al. "Welcome to the tidyverse". In: *Journal of Open Source Software* 4.43 (2019), p. 1686. DOI: 10.21105/joss.01686.

[88] Xinmei Zhao et al. "Mobilization of epithelial mesenchymal transition genes distinguishes active from inactive lesional tissue in patients with ulcerative colitis". In: *Human molecular genetics* 24.16 (2015), pp. 4615–4624.

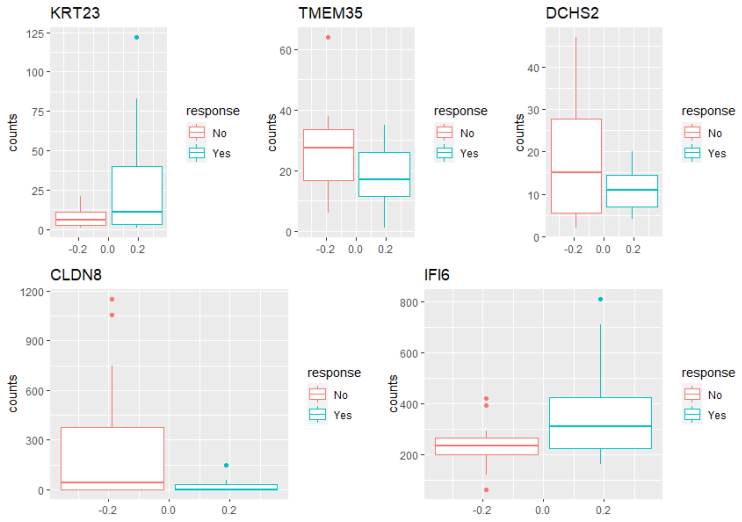
[89] Hui Zou and Trevor Hastie. "Regularization and variable selection via the elastic net". In: *Journal of the royal statistical society: series B (statistical methodology)* 67.2 (2005), pp. 301–320.

# Appendix A

# Supplemental figures




(a) *Boxplots in the paper [78]*




(b) *Boxplots from the data*

Figure A.1: *Comparison of data in the paper and data downloaded from ArrayExpress. Boxplots of specific genes grouped by response. On the left side are counts of non responders and on the right side counts of responders*

# Appendix B

# Toptables baseline

| SYMBOL | logFC | AveExpr | P.Value | adj.P.Val |
|---|---|---|---|---|
| IL13RA2 | -2.23 | 7.75 | 8.88e-10 | 1.58e-05 |
| ACSL4 | -1.16 | 9.32 | 6.71e-09 | 5.18e-05 |
| KLHL5 | -0.92 | 6.96 | 8.74e-09 | 5.18e-05 |
| TNFRSF11B | -1.36 | 7.34 | 2.23e-08 | 9.93e-05 |
| FGF7[1] | -1.24 | 8.18 | 3.00e-08 | 1.05e-04 |
| LY96 | -1.01 | 9.83 | 3.95e-08 | 1.05e-04 |
| SNAPC1 | -0.54 | 7.15 | 4.16e-08 | 1.05e-04 |
| STC1 | -1.47 | 7.76 | 4.71e-08 | 1.05e-04 |
| GLIPR1 | -0.87 | 8.61 | 7.43e-08 | 1.47e-04 |
| WNT5A | -1.28 | 8.61 | 9.73e-08 | 1.73e-04 |
| RHOQ RHOQP1 RHOQP2 RHOQP3 | -0.67 | 8.13 | 1.26e-07 | 1.82e-04 |
| OSGIN2 | -0.39 | 6.75 | 1.27e-07 | 1.82e-04 |
| IKBIP | -0.49 | 6.02 | 1.33e-07 | 1.82e-04 |
| SAMHD1 | -0.58 | 8.74 | 1.82e-07 | 2.27e-04 |
| MGAT4B | 0.47 | 7.57 | 2.21e-07 | 2.27e-04 |
| IL6 | -1.96 | 7.55 | 2.24e-07 | 2.27e-04 |
| PTGS2 | -1.83 | 7.83 | 2.30e-07 | 2.27e-04 |
| DENND5A | -0.74 | 9.24 | 2.39e-07 | 2.27e-04 |
| ST8SIA4 | -0.70 | 6.55 | 2.43e-07 | 2.27e-04 |
| ASAP1 | -0.60 | 8.16 | 2.79e-07 | 2.48e-04 |
| INHBA | -1.26 | 6.97 | 3.13e-07 | 2.65e-04 |
| SNX10 | -0.86 | 8.84 | 3.28e-07 | 2.65e-04 |
| PRNP | -0.50 | 8.57 | 3.45e-07 | 2.67e-04 |
| CEP170 | -0.64 | 7.24 | 4.13e-07 | 2.95e-04 |
| CCL2 | -1.31 | 10.17 | 4.14e-07 | 2.95e-04 |
| EDNRB | -0.81 | 7.99 | 4.50e-07 | 2.99e-04 |
| NIBAN1 | -0.96 | 8.76 | 4.54e-07 | 2.99e-04 |
| TFPI | -0.75 | 6.69 | 4.93e-07 | 3.01e-04 |
| SRGN | -0.88 | 11.14 | 5.08e-07 | 3.01e-04 |
| RGS5 | -0.76 | 7.68 | 5.12e-07 | 3.01e-04 |
| RBMS1 | -0.68 | 9.83 | 5.25e-07 | 3.01e-04 |
| SAMSN1 | -1.01 | 8.22 | 5.89e-07 | 3.27e-04 |
| CSGALNACT2 | -0.60 | 7.82 | 6.20e-07 | 3.34e-04 |
| TLR1 | -1.05 | 7.66 | 6.59e-07 | 3.37e-04 |
| STX11 | -0.62 | 6.71 | 6.62e-07 | 3.37e-04 |
| EVI2A | -0.81 | 7.88 | 6.90e-07 | 3.41e-04 |
| GPR183 | -0.96 | 8.93 | 7.52e-07 | 3.53e-04 |
| TFPI2 | -1.70 | 6.19 | 7.54e-07 | 3.53e-04 |
| PLAU | -0.96 | 8.21 | 8.05e-07 | 3.59e-04 |
| ADGRL2 | -0.54 | 5.78 | 8.08e-07 | 3.59e-04 |
| CLIC2 | -0.68 | 7.18 | 8.98e-07 | 3.89e-04 |
| ROBO1 | -0.87 | 9.63 | 9.46e-07 | 3.97e-04 |
| SLC16A6 | -1.01 | 7.08 | 9.59e-07 | 3.97e-04 |
| CEBPB | -0.66 | 10.43 | 1.04e-06 | 4.19e-04 |
| DEGS1 | -0.45 | 8.39 | 1.07e-06 | 4.20e-04 |
| NR3C1 | -0.66 | 8.52 | 1.12e-06 | 4.20e-04 |
| AKR1B1 | -0.82 | 9.75 | 1.15e-06 | 4.20e-04 |
| GBP5 | -1.14 | 8.38 | 1.17e-06 | 4.20e-04 |
| DRAM1 | -0.59 | 9.38 | 1.18e-06 | 4.20e-04 |
| HGF | -0.72 | 5.95 | 1.19e-06 | 4.20e-04 |

| SYMBOL | logFC | AveExpr | P.Value | adj.P.Val |
|---|---|---|---|---|
| IL11 | -1.22 | 6.39 | 1.28e-06 | 4.27e-04 |
| SELE | -1.73 | 7.90 | 1.89e-06 | 5.13e-04 |
| TNC | -1.09 | 7.47 | 1.99e-06 | 5.13e-04 |
| MIR155 MIR155HG | -1.11 | 8.03 | 2.22e-06 | 5.29e-04 |
| LRRK2 | -1.05 | 8.01 | 2.58e-06 | 5.59e-04 |
| MNDA | -1.34 | 8.62 | 3.54e-06 | 6.55e-04 |
| CCL8 | -1.02 | 8.37 | 4.90e-06 | 7.55e-04 |
| SERPINE1 | -1.05 | 7.17 | 4.97e-06 | 7.55e-04 |
| RGS1 | -1.03 | 8.09 | 8.08e-06 | 9.68e-04 |
| VNN2 | -1.20 | 8.54 | 8.93e-06 | 1.01e-03 |
| NCF2 | -1.02 | 8.64 | 8.98e-06 | 1.01e-03 |
| CHI3L1 | -1.54 | 10.69 | 9.03e-06 | 1.01e-03 |
| CCL4 | -1.10 | 8.40 | 9.28e-06 | 1.02e-03 |
| CXCL6 | -1.39 | 8.05 | 1.22e-05 | 1.22e-03 |
| IL24 | -1.53 | 7.55 | 1.55e-05 | 1.34e-03 |
| HSD11B2 | 1.02 | 9.28 | 1.56e-05 | 1.34e-03 |
| G0S2 | -1.14 | 9.65 | 1.60e-05 | 1.35e-03 |
| MMP3 | -1.52 | 11.08 | 1.62e-05 | 1.35e-03 |
| S100A9 | -1.22 | 9.90 | 2.68e-05 | 1.85e-03 |
| SELL | -1.17 | 9.76 | 2.86e-05 | 1.94e-03 |
| CCL3 CCL3L1 CCL3L3 | -1.24 | 8.85 | 3.77e-05 | 2.20e-03 |
| FCGR3A FCGR3B | -1.25 | 7.93 | 3.87e-05 | 2.22e-03 |
| CXCL8 | -1.41 | 9.62 | 4.26e-05 | 2.32e-03 |
| CHP2 | 1.06 | 8.14 | 4.34e-05 | 2.33e-03 |
| CCN1 | -1.08 | 7.62 | 4.86e-05 | 2.46e-03 |
| CXCR2 | -1.35 | 8.07 | 5.15e-05 | 2.52e-03 |
| PI15 | -1.34 | 6.18 | 5.38e-05 | 2.56e-03 |
| IL1B | -1.21 | 10.55 | 7.70e-05 | 2.98e-03 |
| TNFAIP6 | -1.46 | 7.85 | 7.77e-05 | 2.98e-03 |
| APOBEC3A | -1.24 | 7.35 | 8.28e-05 | 3.10e-03 |
| BCL2A1 | -1.16 | 9.79 | 8.69e-05 | 3.17e-03 |
| CXCL10 | -1.08 | 9.93 | 1.30e-04 | 3.88e-03 |
| MMP10 | -1.20 | 8.81 | 1.75e-04 | 4.56e-03 |
| TREM1 | -1.20 | 7.45 | 1.77e-04 | 4.59e-03 |
| IL1A | -1.04 | 6.95 | 1.96e-04 | 4.90e-03 |
| S100A8 | -1.34 | 11.50 | 4.09e-04 | 7.57e-03 |
| CXCL5 | -1.40 | 8.51 | 4.51e-04 | 8.04e-03 |
| MMP1 | -1.10 | 11.53 | 5.63e-04 | 9.24e-03 |
| PROK2 | -1.35 | 8.40 | 8.40e-04 | 1.19e-02 |
| S100A12 | -1.55 | 8.29 | 8.54e-04 | 1.20e-02 |
| ITLN1 | 1.03 | 11.78 | 9.37e-04 | 1.28e-02 |
| OSM | -1.10 | 7.54 | 1.09e-03 | 1.40e-02 |
| RETNLB | 1.18 | 8.14 | 1.09e-03 | 1.40e-02 |
| ACOD1 | -1.00 | 5.42 | 1.27e-03 | 1.53e-02 |
| AQP9 | -1.22 | 8.45 | 2.08e-03 | 2.14e-02 |
| SPP1 | -1.05 | 8.28 | 3.37e-03 | 2.96e-02 |
| CXCL11 | -1.00 | 7.38 | 6.22e-03 | 4.44e-02 |
| CLCA1 | 1.09 | 11.26 | 7.17e-03 | 4.88e-02 |
| DEFA5 | 1.30 | 9.39 | 1.24e-02 | 6.90e-02 |

[1] /// FGF7P1 /// FGF7P2 /// FGF7P3 /// FGF7P4 /// FGF7P5 /// FGF7P6 /// FGF7P7 /// FGF7P8

Table B.1: *Toptable - baseline - Microarray - Infliximab*
*Toptables contain the first 50 DE genes according to the p-value and the selected genes passing the threshold: $|log2\ fold\ change| > 1$ AND p-value $< 0.05$.*

| SYMBOL | logFC | AveExpr | P.Value | adj.P.Val |
|---|---|---|---|---|
| DPF3 | 0.31 | 6.28 | 4.78e-06 | 3.28e-02 |
| PTPRO | 0.33 | 5.86 | 1.04e-05 | 3.28e-02 |
| PRR16 | -0.54 | 6.25 | 1.34e-05 | 3.28e-02 |
| MAP2K6 | 0.49 | 8.45 | 1.46e-05 | 3.28e-02 |
| SH2D6 | 0.31 | 5.71 | 1.47e-05 | 3.28e-02 |
| TSPAN2 | -0.47 | 5.78 | 1.94e-05 | 3.28e-02 |
| PAPPA | -0.52 | 6.54 | 2.27e-05 | 3.28e-02 |
| TMTC1 | -0.36 | 5.85 | 2.30e-05 | 3.28e-02 |
| PTX3 | -0.68 | 5.60 | 2.43e-05 | 3.28e-02 |
| HOXB9 | 0.36 | 7.87 | 2.48e-05 | 3.28e-02 |
| SNCA | -0.26 | 6.29 | 2.52e-05 | 3.28e-02 |
| SHISAL1 | -0.35 | 6.51 | 2.53e-05 | 3.28e-02 |
| GPRIN2 | 0.43 | 7.53 | 2.55e-05 | 3.28e-02 |
| BAIAP2-DT | 0.24 | 7.49 | 2.58e-05 | 3.28e-02 |
| AIFM3 | 0.52 | 7.11 | 2.97e-05 | 3.31e-02 |
| NMNAT2 | -0.23 | 5.07 | 3.20e-05 | 3.31e-02 |
| KIAA0232 | 0.22 | 8.65 | 3.23e-05 | 3.31e-02 |
| ZNF662 | 0.33 | 7.63 | 3.58e-05 | 3.31e-02 |
| LILRA3 | -0.70 | 6.91 | 3.95e-05 | 3.31e-02 |
| VIPR1 | 0.62 | 8.77 | 4.18e-05 | 3.31e-02 |
| FZD10 | -0.52 | 5.61 | 4.18e-05 | 3.31e-02 |
| GUCY1B2 | 0.26 | 5.39 | 4.73e-05 | 3.31e-02 |
| CDHR1 | 0.79 | 6.48 | 4.92e-05 | 3.31e-02 |
| MEDAG | -0.75 | 6.48 | 4.95e-05 | 3.31e-02 |
| SERTAD4 | 0.41 | 5.80 | 5.03e-05 | 3.31e-02 |
| SDK1 | -0.48 | 6.47 | 5.16e-05 | 3.31e-02 |
| PDGFRA | -0.41 | 8.10 | 5.23e-05 | 3.31e-02 |

| SYMBOL | logFC | AveExpr | P.Value | adj.P.Val |
|---|---|---|---|---|
| FANK1 | 0.37 | 7.31 | 5.36e-05 | 3.31e-02 |
| RGMB | 0.38 | 7.15 | 5.39e-05 | 3.31e-02 |
| THBD | -0.50 | 7.68 | 6.30e-05 | 3.59e-02 |
| MYCL | 0.20 | 6.19 | 6.63e-05 | 3.59e-02 |
| VEGFC | -0.39 | 6.79 | 6.68e-05 | 3.59e-02 |
| CCDC3 | -0.62 | 7.56 | 6.97e-05 | 3.59e-02 |
| FAM114A2 | -0.18 | 7.53 | 7.04e-05 | 3.59e-02 |
| LDLRAD3 | -0.46 | 8.35 | 7.17e-05 | 3.59e-02 |
| PLEKHA4 | -0.23 | 6.20 | 7.27e-05 | 3.59e-02 |
| HOXB5 | 0.32 | 7.11 | 7.64e-05 | 3.67e-02 |
| HOXA2 | 0.32 | 6.44 | 8.06e-05 | 3.70e-02 |
| MTARC1 | 0.34 | 8.41 | 8.42e-05 | 3.70e-02 |
| S100A3 | -0.37 | 5.93 | 8.42e-05 | 3.70e-02 |
| TEX43 | 0.36 | 4.12 | 8.60e-05 | 3.70e-02 |
| TFPI | -0.43 | 6.60 | 9.04e-05 | 3.70e-02 |
| ATP13A4 | 0.33 | 5.21 | 9.10e-05 | 3.70e-02 |
| ACVRL1 | 0.27 | 9.34 | 9.15e-05 | 3.70e-02 |
| HSD11B1 | -0.69 | 6.18 | 9.59e-05 | 3.72e-02 |
| UQCRC1 | 0.33 | 9.34 | 1.00e-04 | 3.72e-02 |
| PRDX4 | -0.25 | 10.50 | 1.01e-04 | 3.72e-02 |
| SLC39A2 | 0.40 | 4.48 | 1.02e-04 | 3.72e-02 |
| CBLIF | 0.46 | 5.09 | 1.04e-04 | 3.72e-02 |
| FBXO25 | 0.18 | 8.11 | 1.05e-04 | 3.72e-02 |
| SPP1 | -1.05 | 8.00 | 1.46e-04 | 4.18e-02 |
| IL13RA2 | -1.04 | 6.83 | 2.21e-04 | 4.18e-02 |
| RETNLB | 1.02 | 8.44 | 2.33e-04 | 4.26e-02 |
| ANXA10 | -1.02 | 6.34 | 2.58e-03 | 8.35e-02 |

Table B.2: *Toptable - baseline - Microarray - Golimumab*
*Toptables contain the first 50 DE genes according to the p-value and the selected genes passing the threshold:* $|log2\ fold\ change| > 1$ *AND p-value* $< 0.05$.

| SYMBOL | logFC | AveExpr | P.Value | adj.P.Val |
|---|---|---|---|---|
| PLCG2 RN7SKP176 | 0.57 | 8.70 | 5.92e-06 | 1.05e-01 |
| ALPL | -0.92 | 6.06 | 6.56e-04 | 7.55e-01 |
| ATP11A | -0.35 | 6.49 | 1.13e-03 | 7.55e-01 |
| STEAP4 | -1.30 | 6.95 | 1.30e-03 | 7.55e-01 |
| UBASH3B | -0.38 | 6.27 | 1.52e-03 | 7.55e-01 |
| PDE10A | -0.49 | 4.91 | 1.72e-03 | 7.55e-01 |
| SLC2A3 | -1.16 | 8.31 | 1.73e-03 | 7.55e-01 |
| FADS1 | -0.84 | 7.70 | 1.77e-03 | 7.55e-01 |
| S100A8 | -2.19 | 11.16 | 1.96e-03 | 7.55e-01 |
| L3MBTL3 | -0.43 | 6.95 | 2.24e-03 | 7.55e-01 |
| TRNP1 | 0.51 | 6.49 | 2.30e-03 | 7.55e-01 |
| G0S2 | -1.40 | 9.61 | 2.30e-03 | 7.55e-01 |
| TNFAIP6 | -1.87 | 7.43 | 2.52e-03 | 7.55e-01 |
| LRP12 | -0.47 | 4.96 | 2.52e-03 | 7.55e-01 |
| KIRREL1 | -0.46 | 6.52 | 2.54e-03 | 7.55e-01 |
| FCRLB | 0.43 | 5.72 | 2.73e-03 | 7.55e-01 |
| HTR4 | 0.52 | 5.09 | 2.75e-03 | 7.55e-01 |
| PSMA8 | 0.40 | 4.69 | 2.85e-03 | 7.55e-01 |
| IL1B | -1.77 | 9.81 | 2.90e-03 | 7.55e-01 |
| ADAMTS9 | -0.55 | 6.85 | 2.90e-03 | 7.55e-01 |
| PI15 | -1.72 | 6.44 | 3.04e-03 | 7.55e-01 |
| KCNJ15 | -1.07 | 5.53 | 3.16e-03 | 7.55e-01 |
| AQP9 | -2.14 | 8.08 | 3.36e-03 | 7.55e-01 |
| ITPR2 | -0.34 | 7.49 | 3.38e-03 | 7.55e-01 |
| ZC3H7B | -0.18 | 7.81 | 3.49e-03 | 7.55e-01 |
| ADGRL2 | -0.71 | 6.34 | 3.55e-03 | 7.55e-01 |
| MSC | -0.58 | 5.93 | 3.71e-03 | 7.55e-01 |
| MAP4K4 | -0.57 | 8.05 | 3.72e-03 | 7.55e-01 |
| PSG6 | 0.24 | 4.23 | 3.73e-03 | 7.55e-01 |
| BCR | -0.29 | 7.60 | 3.90e-03 | 7.55e-01 |
| DYSF | -0.99 | 8.22 | 4.03e-03 | 7.55e-01 |
| FCAR | -0.60 | 5.03 | 4.09e-03 | 7.55e-01 |
| LINC00643 | 0.26 | 3.45 | 4.10e-03 | 7.55e-01 |
| VWF | -0.64 | 8.03 | 4.12e-03 | 7.55e-01 |
| PARP16 | 0.27 | 8.15 | 4.23e-03 | 7.55e-01 |
| STK10 | -0.40 | 8.20 | 4.31e-03 | 7.55e-01 |
| ANKRD30B | -0.56 | 4.45 | 4.35e-03 | 7.55e-01 |
| SGTB | -0.42 | 5.98 | 4.45e-03 | 7.55e-01 |
| ITGAX | -0.74 | 7.34 | 4.58e-03 | 7.55e-01 |
| THBS1 | -0.77 | 8.37 | 4.63e-03 | 7.55e-01 |
| CR1 | -0.99 | 6.22 | 4.63e-03 | 7.55e-01 |
| DENND2A | -0.44 | 7.89 | 4.68e-03 | 7.55e-01 |
| HSPG2 | -0.47 | 8.34 | 4.85e-03 | 7.55e-01 |
| LOXL2 | -0.59 | 6.93 | 4.85e-03 | 7.55e-01 |
| STX11 | -0.68 | 6.59 | 4.88e-03 | 7.55e-01 |
| ADAM12 | -0.71 | 5.79 | 4.96e-03 | 7.55e-01 |
| CXCR1 | -1.34 | 6.72 | 5.11e-03 | 7.55e-01 |
| ARID5B | -0.36 | 6.58 | 5.12e-03 | 7.55e-01 |
| DNAJC2 | -0.23 | 7.11 | 5.21e-03 | 7.55e-01 |
| NLRP3 | -0.66 | 4.93 | 5.24e-03 | 7.55e-01 |
| ACSL1 | -1.08 | 9.00 | 5.59e-03 | 7.55e-01 |
| FPR1 | -1.07 | 6.44 | 5.83e-03 | 7.55e-01 |
| PRRX1 | -1.12 | 7.04 | 5.85e-03 | 7.55e-01 |
| CXCL8 | -1.71 | 9.00 | 5.90e-03 | 7.55e-01 |
| MMP1 | -1.56 | 11.04 | 6.36e-03 | 7.55e-01 |
| S100A12 | -2.16 | 8.22 | 7.21e-03 | 7.55e-01 |
| FFAR2 | -1.09 | 6.19 | 7.39e-03 | 7.55e-01 |
| PTGS2 | -1.55 | 7.42 | 7.86e-03 | 7.55e-01 |
| VNN2 | -1.37 | 8.05 | 7.98e-03 | 7.55e-01 |
| BCL2A1 | -1.36 | 9.40 | 8.41e-03 | 7.55e-01 |
| FCGR1A FCGR1BP FCGR1CP | -1.05 | 7.50 | 8.63e-03 | 7.55e-01 |
| IL6 | -1.53 | 6.90 | 8.90e-03 | 7.55e-01 |
| INHBA | -1.11 | 6.67 | 9.26e-03 | 7.55e-01 |
| MGAM | -1.28 | 7.32 | 1.02e-02 | 7.55e-01 |
| S100A9 | -1.40 | 9.62 | 1.03e-02 | 7.55e-01 |
| PLAU | -1.01 | 8.16 | 1.11e-02 | 7.55e-01 |
| NCF2 | -1.07 | 8.46 | 1.13e-02 | 7.55e-01 |
| IL1A | -1.24 | 6.05 | 1.15e-02 | 7.55e-01 |
| MMP9 | -1.18 | 9.23 | 1.20e-02 | 7.55e-01 |
| FPR2 | -1.13 | 5.63 | 1.22e-02 | 7.55e-01 |
| GLT1D1 | -1.04 | 6.33 | 1.22e-02 | 7.55e-01 |
| PLEK | -1.18 | 8.07 | 1.24e-02 | 7.55e-01 |
| LRRK2 | -1.04 | 7.86 | 1.44e-02 | 7.55e-01 |
| TNC | -1.09 | 7.52 | 1.52e-02 | 7.55e-01 |
| CLCA1 | 1.66 | 11.04 | 1.59e-02 | 7.55e-01 |
| MME | -1.08 | 7.20 | 1.59e-02 | 7.55e-01 |
| LAMP3 | -1.06 | 9.69 | 1.66e-02 | 7.55e-01 |
| FABP6 | 1.13 | 5.38 | 1.68e-02 | 7.55e-01 |
| COL12A1 | -1.07 | 7.86 | 1.78e-02 | 7.55e-01 |
| CSF3R | -1.21 | 7.20 | 1.99e-02 | 7.55e-01 |
| ITLN1 | 1.32 | 10.94 | 2.09e-02 | 7.55e-01 |
| MMP3 | -1.68 | 10.52 | 2.15e-02 | 7.55e-01 |
| IL24 | -1.31 | 7.40 | 2.17e-02 | 7.55e-01 |
| MNDA | -1.09 | 8.48 | 2.19e-02 | 7.55e-01 |
| SELE | -1.48 | 7.94 | 2.35e-02 | 7.55e-01 |
| SELL | -1.18 | 9.44 | 2.39e-02 | 7.55e-01 |
| TREM1 | -1.22 | 7.14 | 2.43e-02 | 7.55e-01 |
| UGT2B17 | 2.08 | 8.50 | 2.44e-02 | 7.55e-01 |
| OSM | -1.27 | 7.11 | 2.49e-02 | 7.55e-01 |
| CLEC4D | -1.04 | 5.71 | 2.58e-02 | 7.55e-01 |
| CXCL6 | -1.31 | 8.42 | 2.68e-02 | 7.55e-01 |
| GUCA2A | 1.38 | 8.91 | 3.23e-02 | 7.55e-01 |
| CXCR2 | -1.14 | 7.70 | 3.69e-02 | 7.55e-01 |
| SLC26A3 | 1.01 | 10.72 | 3.70e-02 | 7.55e-01 |
| PROK2 | -1.43 | 8.38 | 3.72e-02 | 7.55e-01 |
| C10orf99 | 1.06 | 10.35 | 4.13e-02 | 7.55e-01 |
| REG3A | 1.11 | 7.88 | 4.67e-02 | 7.55e-01 |
| CXCL13 | -1.46 | 9.46 | 4.67e-02 | 7.55e-01 |

Table B.3: *Toptable - baseline - Microarray - Vedolizumab*
*Toptables contain the first 50 DE genes according to the p-value and the selected genes passing the threshold:* $|log2\ fold\ change| > 1$ *AND p-value* $< 0.05$.

| SYMBOL | logFC | AveExpr | P.Value | adj.P.Val |
|---|---|---|---|---|
| CPNE6 | 0.23 | 5.53 | 1.27e-04 | 9.90e-01 |
| MRPS12 | 0.22 | 6.38 | 2.41e-04 | 9.90e-01 |
| ACTL10 | 0.24 | 6.54 | 4.36e-04 | 9.90e-01 |
| FNDC11 | 0.20 | 6.18 | 4.48e-04 | 9.90e-01 |
| RNU1-123P TPPP3 | 0.36 | 6.70 | 1.02e-03 | 9.90e-01 |
| RHOXF2 RHOXF2B | 0.23 | 4.24 | 1.08e-03 | 9.90e-01 |
| HES7 | 0.23 | 5.26 | 1.32e-03 | 9.90e-01 |
| LDHC | 0.48 | 5.21 | 1.74e-03 | 9.90e-01 |
| RASA4 RASA4B RASA4CP | -0.49 | 8.51 | 1.84e-03 | 9.90e-01 |
| TAS1R1 | 0.16 | 4.73 | 2.01e-03 | 9.90e-01 |
| RRP9 | 0.23 | 7.03 | 2.07e-03 | 9.90e-01 |
| EXOSC3 | 0.21 | 7.52 | 2.19e-03 | 9.90e-01 |
| SLC47A2 | 0.27 | 4.74 | 2.37e-03 | 9.90e-01 |
| PPP1R3F | 0.16 | 6.02 | 2.40e-03 | 9.90e-01 |
| CACNG7 | 0.21 | 5.01 | 2.44e-03 | 9.90e-01 |
| LINC02873 | 0.20 | 3.96 | 2.65e-03 | 9.90e-01 |
| SCARA5 | 0.31 | 7.19 | 2.85e-03 | 9.90e-01 |
| CRYBB2 | 0.22 | 5.31 | 3.23e-03 | 9.90e-01 |
| SYN3 | 0.18 | 4.97 | 3.35e-03 | 9.90e-01 |
| HCN2 | 0.14 | 5.30 | 3.49e-03 | 9.90e-01 |
| CRH | 0.16 | 4.45 | 3.57e-03 | 9.90e-01 |
| ADORA1 | 0.16 | 6.08 | 3.63e-03 | 9.90e-01 |
| GZMM | 0.20 | 6.33 | 3.83e-03 | 9.90e-01 |
| CSN1S1 | 0.31 | 3.89 | 3.83e-03 | 9.90e-01 |
| CGB1 CGB2 CGB3 CGB5 CGB7 CGB8 | 0.22 | 6.16 | 3.87e-03 | 9.90e-01 |

| SYMBOL | logFC | AveExpr | P.Value | adj.P.Val |
|---|---|---|---|---|
| SLC39A12 | 0.19 | 3.39 | 4.30e-03 | 9.90e-01 |
| BOP1 | 0.20 | 7.16 | 4.49e-03 | 9.90e-01 |
| TIMM50 | 0.20 | 7.51 | 4.57e-03 | 9.90e-01 |
| ZNF80 | -0.17 | 3.40 | 4.58e-03 | 9.90e-01 |
| PA2G4 | 0.22 | 9.23 | 4.67e-03 | 9.90e-01 |
| PLXNA4 | 0.14 | 5.36 | 4.88e-03 | 9.90e-01 |
| MCM3 | 0.24 | 8.76 | 4.90e-03 | 9.90e-01 |
| LIN54 | 0.22 | 5.98 | 4.95e-03 | 9.90e-01 |
| CDT1 | 0.28 | 6.63 | 5.01e-03 | 9.90e-01 |
| ANTKMT | 0.21 | 7.50 | 5.08e-03 | 9.90e-01 |
| C16orf92 | 0.21 | 4.61 | 5.18e-03 | 9.90e-01 |
| GINS2 | 0.25 | 6.83 | 5.22e-03 | 9.90e-01 |
| TH | 0.17 | 4.85 | 5.29e-03 | 9.90e-01 |
| ISOC2 | 0.34 | 8.35 | 5.43e-03 | 9.90e-01 |
| RRM1 | 0.27 | 8.49 | 5.61e-03 | 9.90e-01 |
| TP53I11 | 0.20 | 5.97 | 6.00e-03 | 9.90e-01 |
| NTHL1 | 0.28 | 7.69 | 6.13e-03 | 9.90e-01 |
| STEAP2 | -0.38 | 8.58 | 6.46e-03 | 9.90e-01 |
| NDUFV3 | 0.24 | 8.30 | 6.47e-03 | 9.90e-01 |
| POLG2 | -0.25 | 7.19 | 6.58e-03 | 9.90e-01 |
| C20orf27 | 0.15 | 6.37 | 6.62e-03 | 9.90e-01 |
| LCE2B | 0.17 | 4.76 | 6.64e-03 | 9.90e-01 |
| C20orf173 | 0.19 | 4.56 | 6.67e-03 | 9.90e-01 |
| SNCB | 0.16 | 5.52 | 6.77e-03 | 9.90e-01 |
| SCN5A | 0.25 | 5.17 | 6.80e-03 | 9.90e-01 |
| HLA-DQA1 HLA-DQA2 | -1.02 | 9.11 | 2.82e-02 | 9.90e-01 |
| CXCL6 | -1.16 | 7.00 | 3.36e-02 | 9.90e-01 |

Table B.4: *Toptable - baseline - Microarray - Placebo*
*Toptables contain the first 50 DE genes according to the p-value and the selected genes passing the threshold:* $|log2\ fold\ change| > 1$ *AND p-value* $< 0.05$.

| SYMBOL | logFC | AveExpr | P.Value | adj.P.Val |
|---|---|---|---|---|
| BMP7 | -1.15 | 3.75 | 2.37e-04 | 8.48e-01 |
| PLA2G5 | -1.52 | -0.10 | 2.76e-04 | 8.48e-01 |
| CLECL1 | -1.17 | 0.92 | 5.59e-04 | 8.48e-01 |
| TMEM116 | -1.03 | 1.86 | 7.58e-04 | 8.48e-01 |
| SYPL2 | -1.24 | 0.09 | 9.08e-04 | 8.48e-01 |
| HTT | -0.34 | 6.44 | 1.04e-03 | 8.48e-01 |
| KIFAP3 | 0.42 | 5.07 | 1.05e-03 | 8.48e-01 |
| DLEC1 | -1.10 | 0.47 | 1.12e-03 | 8.48e-01 |
| GNG5 | 1.12 | 2.30 | 1.12e-03 | 8.48e-01 |
| NUDCD2 | 0.42 | 4.43 | 1.14e-03 | 8.48e-01 |
| MESDC1 | 0.63 | 5.16 | 1.17e-03 | 8.48e-01 |
| MPLKIP | 0.52 | 4.22 | 1.24e-03 | 8.48e-01 |
| FRG1B | -0.94 | 1.05 | 1.27e-03 | 8.48e-01 |
| DNASE1L3 | -0.90 | 5.60 | 1.32e-03 | 8.48e-01 |
| PRC1 | 0.86 | 1.16 | 1.50e-03 | 8.48e-01 |
| IKBKE | -0.39 | 4.81 | 1.53e-03 | 8.48e-01 |
| PHGR1 | 1.84 | 7.18 | 1.69e-03 | 8.48e-01 |
| C20orf24 | 1.56 | 1.61 | 1.78e-03 | 8.48e-01 |
| KRT10 | 0.62 | 3.76 | 1.78e-03 | 8.48e-01 |
| C4orf48 | 1.75 | 0.73 | 1.83e-03 | 8.48e-01 |
| VSIG10L | -1.09 | 0.62 | 2.01e-03 | 8.48e-01 |
| TMEM238 | 1.21 | 1.51 | 2.05e-03 | 8.48e-01 |
| IER3IP1 | 0.39 | 5.05 | 2.15e-03 | 8.48e-01 |
| HDC | -1.30 | 2.68 | 2.18e-03 | 8.48e-01 |
| SZT2 | -0.44 | 5.26 | 2.24e-03 | 8.48e-01 |
| IGSF9B | -1.20 | 2.45 | 2.31e-03 | 8.48e-01 |
| ABHD5 | 0.45 | 5.23 | 2.37e-03 | 8.48e-01 |
| ADAT3 | 1.54 | 0.77 | 2.44e-03 | 8.48e-01 |
| FABP2 | 1.67 | 4.46 | 2.51e-03 | 8.48e-01 |
| WDTC1 | -0.38 | 5.92 | 2.67e-03 | 8.48e-01 |
| TMEM128 | 0.41 | 3.63 | 2.81e-03 | 8.48e-01 |
| TMEM160 | 1.31 | 1.37 | 2.87e-03 | 8.48e-01 |
| PM20D1 | -1.15 | -0.47 | 2.94e-03 | 8.48e-01 |
| RAB2A | 0.41 | 7.25 | 3.26e-03 | 8.48e-01 |
| OSTC | 0.35 | 6.31 | 3.31e-03 | 8.48e-01 |
| DUSP5 | -0.70 | 6.64 | 3.32e-03 | 8.48e-01 |
| RP11-11N5.1 | -2.10 | 0.53 | 3.40e-03 | 8.48e-01 |
| RP11-455F5.3 | -1.02 | -0.22 | 3.45e-03 | 8.48e-01 |
| MPP2 | -0.92 | 1.00 | 3.65e-03 | 8.48e-01 |
| PKNOX2 | -0.87 | 1.70 | 3.73e-03 | 8.48e-01 |
| IMP3 | 0.32 | 4.19 | 3.75e-03 | 8.48e-01 |
| CCDC136 | -1.25 | 0.55 | 3.98e-03 | 8.48e-01 |
| SLC37A2 | -0.70 | 3.64 | 4.09e-03 | 8.48e-01 |
| C19orf44 | -0.65 | 1.79 | 4.11e-03 | 8.48e-01 |
| RP11-147L13.15 | -0.97 | 1.22 | 4.18e-03 | 8.48e-01 |
| HAP1 | -1.08 | 0.10 | 4.19e-03 | 8.48e-01 |
| ACTL10 | 0.92 | 0.74 | 4.46e-03 | 8.48e-01 |
| GAB2 | -0.45 | 4.89 | 4.49e-03 | 8.48e-01 |
| EMILIN3 | -1.08 | 0.22 | 4.50e-03 | 8.48e-01 |
| EEF1A1P1 | 0.65 | 1.60 | 4.58e-03 | 8.48e-01 |
| IFI30 | -1.01 | 0.22 | 4.83e-03 | 8.48e-01 |
| ENDOG | 1.35 | 0.10 | 5.03e-03 | 8.48e-01 |
| LRRC26 | 2.07 | 0.15 | 5.04e-03 | 8.48e-01 |
| SIGLEC6 | -1.37 | 0.85 | 5.10e-03 | 8.48e-01 |
| COMTD1 | 1.03 | 3.48 | 5.28e-03 | 8.48e-01 |
| BRINP3 | 1.32 | 0.52 | 5.89e-03 | 8.48e-01 |
| MESP1 | 1.06 | -0.10 | 6.28e-03 | 8.48e-01 |
| CTC-425O23.5 | -1.06 | -0.12 | 6.38e-03 | 8.48e-01 |
| KCNMB1 | -1.16 | 2.70 | 6.56e-03 | 8.48e-01 |
| HOXD11 | -1.62 | 0.49 | 7.01e-03 | 8.48e-01 |
| ITIH5 | -1.20 | 5.21 | 7.13e-03 | 8.48e-01 |
| NXPH3 | -1.20 | 0.85 | 7.41e-03 | 8.48e-01 |
| UGT2B17 | 2.77 | 5.55 | 7.54e-03 | 8.48e-01 |
| SV2B | -1.17 | 1.22 | 7.68e-03 | 8.48e-01 |
| EVX2 | -1.85 | -0.19 | 7.69e-03 | 8.48e-01 |
| LDLRAD2 | -1.01 | 0.46 | 7.85e-03 | 8.48e-01 |
| FAM131B | -1.04 | 0.44 | 7.95e-03 | 8.48e-01 |
| MYH11 | -1.27 | 6.89 | 8.45e-03 | 8.48e-01 |
| CYP27C1 | -1.18 | 0.48 | 8.47e-03 | 8.48e-01 |
| C4B | -1.36 | 2.41 | 8.94e-03 | 8.48e-01 |
| AHNAK2 | -1.18 | 3.48 | 9.29e-03 | 8.48e-01 |
| APOA1 | 1.29 | 0.77 | 9.51e-03 | 8.48e-01 |
| CNTN1 | -1.48 | 0.76 | 9.64e-03 | 8.48e-01 |
| RP11-273B20.1 | -1.01 | 0.24 | 1.01e-02 | 8.48e-01 |
| PCSK1N | 1.33 | 0.17 | 1.03e-02 | 8.48e-01 |
| IGHV3OR16-13 | -1.43 | 0.81 | 1.03e-02 | 8.48e-01 |
| HOXD10 | -1.36 | 1.56 | 1.08e-02 | 8.48e-01 |
| EGR3 | -1.20 | 4.03 | 1.16e-02 | 8.48e-01 |
| FADS2 | -1.32 | 5.24 | 1.18e-02 | 8.48e-01 |
| HOXD12 | -1.45 | -0.53 | 1.19e-02 | 8.48e-01 |
| SYNPO2 | -1.10 | 5.91 | 1.22e-02 | 8.48e-01 |
| TPGS1 | 1.16 | 0.47 | 1.23e-02 | 8.48e-01 |
| HOXD13 | -2.33 | 2.47 | 1.25e-02 | 8.48e-01 |
| FCGR2C | -1.15 | 1.33 | 1.28e-02 | 8.48e-01 |
| MTND4P12 | 3.36 | 5.14 | 1.28e-02 | 8.48e-01 |
| CXCR4 | -1.16 | 6.00 | 1.37e-02 | 8.48e-01 |
| MIR646HG | -1.04 | -0.35 | 1.42e-02 | 8.48e-01 |
| HLA-DQB2 | -2.02 | 2.65 | 1.42e-02 | 8.48e-01 |
| SLC18A2 | -1.02 | 0.51 | 1.46e-02 | 8.48e-01 |
| FAM163A | -1.02 | 0.91 | 1.56e-02 | 8.48e-01 |
| JPH2 | -1.07 | 2.10 | 1.59e-02 | 8.48e-01 |
| CA1 | 2.93 | 5.04 | 1.59e-02 | 8.48e-01 |
| PLN | -1.11 | 2.93 | 1.61e-02 | 8.48e-01 |
| REG3A | 2.40 | 5.20 | 1.62e-02 | 8.48e-01 |
| ADAMTS8 | -1.15 | 1.84 | 1.65e-02 | 8.48e-01 |
| TMEM178B | -1.46 | 1.03 | 1.67e-02 | 8.48e-01 |
| CADM3 | -1.33 | 2.63 | 1.71e-02 | 8.48e-01 |
| MS4A12 | 1.66 | 5.68 | 1.79e-02 | 8.48e-01 |
| PRIMA1 | -1.08 | 2.91 | 1.79e-02 | 8.48e-01 |
| APOBEC1 | 1.08 | 2.09 | 1.80e-02 | 8.48e-01 |
| TFAP2A | -1.17 | 0.01 | 1.90e-02 | 8.48e-01 |
| SEMA3D | -1.09 | 1.88 | 1.94e-02 | 8.48e-01 |
| MYBPC2 | -1.45 | 0.04 | 2.04e-02 | 8.48e-01 |
| NPHS1 | -1.72 | -0.27 | 2.07e-02 | 8.48e-01 |
| SERPINA5 | -1.25 | 1.43 | 2.07e-02 | 8.48e-01 |
| SLC1A2 | -1.40 | 0.39 | 2.11e-02 | 8.48e-01 |
| KEL | -1.49 | -0.11 | 2.25e-02 | 8.48e-01 |
| COL19A1 | -1.24 | 0.43 | 2.30e-02 | 8.48e-01 |
| CLDN3 | 1.15 | 6.70 | 2.37e-02 | 8.48e-01 |
| LONRF2 | -1.15 | -0.12 | 2.38e-02 | 8.48e-01 |
| GFRA1 | -1.06 | 3.13 | 2.41e-02 | 8.48e-01 |
| DEFA6 | 1.95 | 3.38 | 2.41e-02 | 8.48e-01 |
| SIGLEC14 | -1.05 | 2.36 | 2.45e-02 | 8.48e-01 |
| CLU | -1.59 | 7.30 | 2.46e-02 | 8.48e-01 |
| EPHB6 | -1.07 | 2.45 | 2.47e-02 | 8.48e-01 |
| CASC9 | 1.43 | 0.29 | 2.47e-02 | 8.48e-01 |
| UGT2B15 | 1.75 | 2.12 | 2.49e-02 | 8.48e-01 |
| PPP1R14C | 1.09 | 3.81 | 2.66e-02 | 8.48e-01 |
| RASGEF1A | -1.04 | 1.45 | 2.68e-02 | 8.48e-01 |
| CHI3L2 | -1.37 | 4.60 | 2.70e-02 | 8.48e-01 |
| IGFBP2 | 1.03 | 5.56 | 2.73e-02 | 8.48e-01 |
| RNF150 | -1.07 | 2.60 | 2.83e-02 | 8.48e-01 |
| NPTX1 | -1.44 | 1.27 | 2.91e-02 | 8.48e-01 |
| CCL3L3 | -1.47 | 1.52 | 2.96e-02 | 8.48e-01 |
| TYRP1 | -1.21 | 1.06 | 3.24e-02 | 8.48e-01 |
| MS4A2 | -1.03 | 2.41 | 3.28e-02 | 8.48e-01 |
| DCC | -1.10 | 0.20 | 3.36e-02 | 8.48e-01 |
| NA | 1.09 | 0.55 | 3.38e-02 | 8.48e-01 |
| LY86-AS1 | -1.01 | -0.30 | 3.45e-02 | 8.48e-01 |
| CWH43 | 1.07 | 2.65 | 3.90e-02 | 8.48e-01 |
| PRELP | -1.47 | 3.14 | 3.90e-02 | 8.48e-01 |
| RP11-223C24.1 | 1.01 | 0.12 | 3.91e-02 | 8.48e-01 |
| SORCS1 | -1.07 | 2.38 | 3.91e-02 | 8.48e-01 |
| TMIGD1 | 2.12 | 2.88 | 3.93e-02 | 8.48e-01 |
| BEST4 | 1.95 | 3.45 | 3.94e-02 | 8.48e-01 |
| SHISA3 | -1.14 | 0.33 | 3.97e-02 | 8.48e-01 |
| DMBT1 | 1.13 | 9.34 | 4.03e-02 | 8.48e-01 |
| RELN | -1.15 | 1.67 | 4.05e-02 | 8.48e-01 |
| OTOP2 | 2.34 | 1.40 | 4.14e-02 | 8.48e-01 |
| IGKV6-21 | -1.41 | 3.55 | 4.19e-02 | 8.48e-01 |
| LMO3 | -1.06 | 0.56 | 4.19e-02 | 8.48e-01 |
| ALDH1A2 | 1.13 | 4.21 | 4.22e-02 | 8.48e-01 |
| FAM135B | -1.04 | 1.79 | 4.31e-02 | 8.48e-01 |
| TMEM82 | 1.07 | 1.76 | 4.43e-02 | 8.48e-01 |
| CCL4L1 | -1.18 | 2.49 | 4.48e-02 | 8.48e-01 |
| SCN7A | -1.07 | 2.87 | 4.65e-02 | 8.48e-01 |
| P2RY12 | -1.21 | 0.18 | 4.71e-02 | 8.48e-01 |
| TTC24 | -1.15 | 0.06 | 4.71e-02 | 8.48e-01 |
| EYA2 | 1.29 | 1.72 | 4.78e-02 | 8.48e-01 |
| LY6D | -1.80 | 0.99 | 4.78e-02 | 8.48e-01 |
| FABP1 | 1.03 | 9.05 | 4.81e-02 | 8.48e-01 |
| CXCR5 | -1.33 | 0.13 | 4.85e-02 | 8.48e-01 |
| FTH1 | 1.14 | 6.17 | 4.85e-02 | 8.48e-01 |
| MT-TM | 1.06 | -0.03 | 4.86e-02 | 8.48e-01 |
| CD72 | -1.03 | 3.50 | 4.94e-02 | 8.48e-01 |

Table B.5: *Toptable - baseline - RNA-Seq - Infliximab+Adalimumab*
*Toptables contain the first 50 DE genes according to the p-value and the selected genes passing the threshold:* $|log2\ fold\ change| > 1$ *AND p-value* $< 0.05$.

| SYMBOL | logFC | AveExpr | P.Value | adj.P.Val |
|---|---|---|---|---|
| KRT8P33 | 0.77 | -0.06 | 1.06e-04 | 9.83e-01 |
| SLC5A3 | 0.42 | 4.89 | 2.78e-04 | 9.83e-01 |
| NPTN-IT1 | -0.55 | 0.37 | 4.86e-04 | 9.83e-01 |
| ARHGAP18 | 0.24 | 5.08 | 8.84e-04 | 9.83e-01 |
| CARD6 | -0.77 | 3.86 | 1.11e-03 | 9.83e-01 |
| BEND3P1 | -1.09 | -0.02 | 1.37e-03 | 9.83e-01 |
| NRG4 | 0.61 | 0.46 | 1.55e-03 | 9.83e-01 |
| CABP4 | -0.55 | 0.38 | 1.57e-03 | 9.83e-01 |
| RP11-46C24.7 | -0.48 | 0.83 | 1.63e-03 | 9.83e-01 |
| EIF4A1 | -0.52 | 1.26 | 1.80e-03 | 9.83e-01 |
| CTC-559E9.8 | -0.42 | 0.15 | 1.95e-03 | 9.83e-01 |
| AC093495.4 | -0.30 | 3.12 | 2.15e-03 | 9.83e-01 |
| CD163L1 | 0.68 | 5.00 | 2.79e-03 | 9.83e-01 |
| CD82 | -0.57 | 6.34 | 2.83e-03 | 9.83e-01 |
| CD99P1 | -0.43 | 1.48 | 3.10e-03 | 9.83e-01 |
| MVB12B | 0.36 | 3.61 | 3.14e-03 | 9.83e-01 |
| RP11-115D19.1 | -1.09 | 1.15 | 3.15e-03 | 9.83e-01 |
| RUVBL2 | 0.23 | 5.27 | 3.44e-03 | 9.83e-01 |
| NUTM2A-AS1 | -0.41 | 3.63 | 3.78e-03 | 9.83e-01 |
| LINC00346 | -0.77 | 1.28 | 3.80e-03 | 9.83e-01 |
| EPHB1 | -0.84 | 1.18 | 3.88e-03 | 9.83e-01 |
| HCG27 | -0.63 | 1.20 | 4.03e-03 | 9.83e-01 |
| RP11-58K22.5 | -0.57 | 3.60 | 4.04e-03 | 9.83e-01 |
| ARMCX6 | -0.27 | 2.67 | 4.09e-03 | 9.83e-01 |
| RP1-34M23.5 | -0.85 | 0.43 | 4.10e-03 | 9.83e-01 |
| LCORL | 0.20 | 4.21 | 4.27e-03 | 9.83e-01 |
| PWWP2A | 0.24 | 4.37 | 4.28e-03 | 9.83e-01 |
| RP4-610C12.4 | -0.70 | 0.02 | 4.28e-03 | 9.83e-01 |
| RP11-1149O23.2 | -0.37 | 1.54 | 4.32e-03 | 9.83e-01 |
| WDR45 | -0.30 | 4.08 | 4.39e-03 | 9.83e-01 |
| KIAA1614-AS1 | -0.45 | -0.06 | 4.39e-03 | 9.83e-01 |

| SYMBOL | logFC | AveExpr | P.Value | adj.P.Val |
|---|---|---|---|---|
| ANKRD27 | 0.22 | 4.93 | 4.51e-03 | 9.83e-01 |
| CH507-9B2.5 | -0.66 | 1.46 | 4.84e-03 | 9.83e-01 |
| LINC01232 | -0.53 | 2.03 | 5.38e-03 | 9.83e-01 |
| ZKSCAN2 | -0.34 | 1.85 | 5.55e-03 | 9.83e-01 |
| ECH1 | 0.41 | 5.41 | 5.58e-03 | 9.83e-01 |
| RP11-299J3.8 | -0.19 | 4.47 | 5.66e-03 | 9.83e-01 |
| RP11-286N22.8 | -0.33 | 1.37 | 5.98e-03 | 9.83e-01 |
| RPL13AP25 | 0.34 | 2.60 | 6.13e-03 | 9.83e-01 |
| CTD-2215L10.1 | 0.54 | 1.18 | 6.24e-03 | 9.83e-01 |
| RP11-572M11.4 | -0.41 | 0.61 | 6.28e-03 | 9.83e-01 |
| ZNF574 | 0.22 | 3.36 | 6.33e-03 | 9.83e-01 |
| ADCY9 | 0.38 | 4.85 | 6.43e-03 | 9.83e-01 |
| BRINP3 | 1.26 | -0.13 | 6.71e-03 | 9.83e-01 |
| LRP4 | 1.05 | 4.55 | 6.72e-03 | 9.83e-01 |
| DYM | -0.15 | 4.80 | 6.86e-03 | 9.83e-01 |
| AGA | -0.27 | 4.17 | 6.88e-03 | 9.83e-01 |
| MAN2A2 | -0.21 | 5.76 | 7.01e-03 | 9.83e-01 |
| BTAF1 | -0.23 | 5.55 | 7.03e-03 | 9.83e-01 |
| RP1-232L22__B.1 | -0.64 | 0.82 | 7.43e-03 | 9.83e-01 |
| RP1-55C23.7 | -1.07 | 1.22 | 9.45e-03 | 9.83e-01 |
| HLA-V | -1.14 | 0.86 | 1.16e-02 | 9.83e-01 |
| EYA2 | 1.13 | 1.16 | 1.28e-02 | 9.83e-01 |
| SERPINA9 | -1.13 | -0.50 | 1.40e-02 | 9.83e-01 |
| VNN2 | -1.18 | 4.19 | 1.59e-02 | 9.83e-01 |
| NAMPTP1 | -1.05 | 3.31 | 2.98e-02 | 9.83e-01 |
| SULT1C2 | -1.17 | 2.79 | 3.04e-02 | 9.83e-01 |
| PTHLH | -1.05 | 0.74 | 3.38e-02 | 9.83e-01 |
| PHGR1 | 1.23 | 7.96 | 4.06e-02 | 9.83e-01 |
| LRRC26 | 1.15 | 2.18 | 4.52e-02 | 9.83e-01 |

Table B.6: *Toptable - baseline - RNA-Seq - Etrolizumab*
*Toptables contain the first 50 DE genes according to the p-value and the selected genes passing the threshold: $|log2\ fold\ change| > 1$ AND p-value $< 0.05$.*

# Appendix C

# Toptables change during treatment

| SYMBOL | logFC | AveExpr | P.Value | adj.P.Val |
|---|---|---|---|---|
| BACE2 | -0.72 | -0.35 | 5.37e-10 | 9.55e-06 |
| APIP | -0.73 | -0.21 | 2.76e-09 | 2.45e-05 |
| TMC5 | -0.78 | -0.23 | 3.34e-08 | 1.64e-04 |
| ASRGL1 | -0.66 | -0.23 | 5.13e-08 | 1.64e-04 |
| ASS1 | -0.96 | -0.35 | 5.50e-08 | 1.64e-04 |
| GMDS | -0.63 | -0.22 | 5.53e-08 | 1.64e-04 |
| RTEL1 RTEL1-TNFRSF6B | -0.45 | -0.32 | 8.02e-08 | 2.04e-04 |
| C2CD4A | -1.32 | -0.60 | 1.46e-07 | 3.02e-04 |
| CLGN | 0.44 | 0.12 | 1.53e-07 | 3.02e-04 |
| SLCO4A1 | -0.76 | -0.57 | 2.00e-07 | 3.55e-04 |
| TTC5 | 0.28 | 0.12 | 2.48e-07 | 3.80e-04 |
| S100A11 | -0.39 | -0.36 | 2.56e-07 | 3.80e-04 |
| SOD1 | -0.47 | -0.17 | 3.74e-07 | 5.12e-04 |
| DMBT1 | -1.81 | -1.08 | 5.95e-07 | 7.56e-04 |
| SORD SORD2P | -0.60 | -0.28 | 7.68e-07 | 9.11e-04 |
| IFITM3 | -0.60 | -0.48 | 8.63e-07 | 9.30e-04 |
| CAB39L | 0.32 | 0.10 | 8.89e-07 | 9.30e-04 |
| TBC1D2 | -0.41 | -0.21 | 1.44e-06 | 1.32e-03 |
| SSPN | 0.37 | 0.21 | 1.47e-06 | 1.32e-03 |
| TEAD4 | -0.49 | -0.24 | 1.52e-06 | 1.32e-03 |
| CYP27A1 | 0.49 | 0.23 | 1.58e-06 | 1.32e-03 |
| GABRE | -0.73 | -0.12 | 1.67e-06 | 1.32e-03 |
| HK2 | -0.71 | -0.27 | 1.70e-06 | 1.32e-03 |
| SLC5A1 | -0.78 | -0.28 | 1.91e-06 | 1.42e-03 |
| ABCA12 | -1.22 | -0.58 | 2.03e-06 | 1.45e-03 |
| LRG1 | -0.62 | -0.40 | 2.27e-06 | 1.56e-03 |
| PAG1 | 0.61 | 0.15 | 2.43e-06 | 1.60e-03 |
| C4BPB | -1.21 | -1.05 | 3.21e-06 | 2.04e-03 |
| SYNC | 0.47 | 0.25 | 3.58e-06 | 2.20e-03 |
| BCHE | 0.68 | 0.22 | 3.74e-06 | 2.22e-03 |
| PDIA4 | -0.44 | -0.37 | 4.02e-06 | 2.31e-03 |
| DEPDC7 | 0.67 | 0.50 | 4.32e-06 | 2.39e-03 |
| CRELD2 | -0.52 | -0.42 | 4.43e-06 | 2.39e-03 |
| DUOXA2 | -1.89 | -1.52 | 4.69e-06 | 2.45e-03 |
| RNF145 | -0.40 | -0.25 | 6.67e-06 | 3.23e-03 |
| DPY19L1 | -0.36 | -0.20 | 6.77e-06 | 3.23e-03 |
| LPCAT1 | -0.71 | -0.53 | 6.86e-06 | 3.23e-03 |
| ADAMTSL3 | 0.24 | 0.11 | 6.89e-06 | 3.23e-03 |
| CASP1 | -0.67 | -0.35 | 7.08e-06 | 3.23e-03 |
| HPSE | -0.54 | -0.35 | 7.68e-06 | 3.42e-03 |
| SLC6A14 | -1.42 | -1.04 | 8.00e-06 | 3.43e-03 |
| LIPG | -0.72 | -0.21 | 8.22e-06 | 3.43e-03 |
| ABCB6 | -0.43 | -0.25 | 8.30e-06 | 3.43e-03 |
| TPK1 | -0.80 | -0.43 | 8.96e-06 | 3.57e-03 |
| PDZK1IP1 | -1.13 | -0.88 | 9.40e-06 | 3.57e-03 |

| SYMBOL | logFC | AveExpr | P.Value | adj.P.Val |
|---|---|---|---|---|
| TXNIP | 0.43 | 0.01 | 9.55e-06 | 3.57e-03 |
| PIK3R1 | 0.38 | 0.24 | 9.58e-06 | 3.57e-03 |
| PLCD1 | 0.51 | 0.33 | 9.63e-06 | 3.57e-03 |
| SAA1 SAA2 SAA2-SAA4 | -2.03 | -1.46 | 1.04e-05 | 3.74e-03 |
| CCDC88B | -0.35 | -0.21 | 1.05e-05 | 3.74e-03 |
| CASP5 | -1.15 | -0.76 | 1.38e-05 | 4.08e-03 |
| XDH | -1.01 | -0.57 | 1.42e-05 | 4.14e-03 |
| TNIP3 | -2.00 | -1.75 | 1.56e-05 | 4.30e-03 |
| IL1R2 | 1.05 | 0.33 | 2.93e-05 | 6.14e-03 |
| NOS2 | -1.41 | -1.13 | 4.44e-05 | 7.81e-03 |
| AQP8 | 1.67 | 1.30 | 8.59e-05 | 1.12e-02 |
| C4BPA | -1.42 | -1.35 | 9.58e-05 | 1.18e-02 |
| CFI | -1.14 | -0.87 | 1.07e-04 | 1.24e-02 |
| PNLIPRP2 | 1.23 | 0.86 | 1.07e-04 | 1.24e-02 |
| SLC26A4 | -1.07 | -0.28 | 1.12e-04 | 1.27e-02 |
| LCN2 | -1.13 | -0.83 | 1.40e-04 | 1.43e-02 |
| DUOX2 | -1.57 | -1.22 | 1.77e-04 | 1.56e-02 |
| SERPINB7 | -1.85 | -1.65 | 1.85e-04 | 1.57e-02 |
| ABCG2 | 1.03 | 0.73 | 2.30e-04 | 1.66e-02 |
| CXCL1 | -1.17 | -1.36 | 2.85e-04 | 1.85e-02 |
| AZGP1 | -1.07 | -0.79 | 3.05e-04 | 1.91e-02 |
| HMGCS2 | 1.31 | 1.02 | 3.57e-04 | 2.08e-02 |
| REG1B | -2.32 | -1.91 | 3.70e-04 | 2.11e-02 |
| SLC26A2 | 1.56 | 1.24 | 3.86e-04 | 2.13e-02 |
| VNN1 | -1.12 | -0.93 | 4.09e-04 | 2.23e-02 |
| S100A8 | -1.81 | -2.17 | 4.81e-04 | 2.49e-02 |
| ANXA1 | -1.01 | -1.06 | 5.30e-04 | 2.59e-02 |
| MMP12 | -1.10 | -1.36 | 5.47e-04 | 2.64e-02 |
| UGT2A3 | 1.45 | 0.92 | 6.20e-04 | 2.83e-02 |
| DEFB4A DEFB4B | -1.53 | -2.03 | 1.07e-03 | 3.93e-02 |
| MT1M | 1.19 | 0.93 | 1.15e-03 | 4.15e-02 |
| SERPINB3 | -1.70 | -1.75 | 1.53e-03 | 4.71e-02 |
| FAM3B | -1.06 | -1.12 | 1.89e-03 | 5.35e-02 |
| TCN1 | -1.33 | -1.72 | 2.08e-03 | 5.70e-02 |
| MMP3 | -1.79 | -2.15 | 2.40e-03 | 6.15e-02 |
| SERPINB4 | -1.21 | -1.30 | 2.63e-03 | 6.47e-02 |
| GUCA2B | 1.06 | 0.93 | 4.29e-03 | 8.33e-02 |
| CXCL8 | -1.23 | -1.65 | 5.00e-03 | 9.14e-02 |
| REG1A | -1.60 | -1.76 | 5.65e-03 | 9.87e-02 |
| IDO1 | -1.21 | -1.43 | 6.37e-03 | 1.05e-01 |
| PCK1 | 1.35 | 1.64 | 6.53e-03 | 1.07e-01 |
| GUCA2A | 1.04 | 0.81 | 6.83e-03 | 1.10e-01 |
| MMP1 | -1.37 | -1.76 | 7.26e-03 | 1.13e-01 |
| CLDN8 | 1.48 | 1.42 | 7.72e-03 | 1.16e-01 |
| CHI3L1 | -1.44 | -1.90 | 1.28e-02 | 1.49e-01 |

Table C.1: *Toptable - change during treatment - Microarray - Infliximab*
*Toptables contain the first 50 DE genes according to the p-value and the selected genes passing the threshold:* $|log2\ fold\ change| > 1$ *AND p-value* $< 0.05$.

| SYMBOL | logFC | AveExpr | P.Value | adj.P.Val |
|---|---|---|---|---|
| SLC47A1 | -0.62 | -0.19 | 8.86e-07 | 1.58e-02 |
| EGFL6 | -1.03 | -0.51 | 5.80e-05 | 2.38e-01 |
| LAIR1 | -0.37 | -0.23 | 7.26e-05 | 2.38e-01 |
| SLAMF8 | -0.77 | -0.47 | 7.57e-05 | 2.38e-01 |
| CLEC1A | -0.59 | -0.27 | 7.61e-05 | 2.38e-01 |
| CARD6 | -0.60 | -0.24 | 8.04e-05 | 2.38e-01 |
| FCRL5 | -0.92 | -0.55 | 9.66e-05 | 2.46e-01 |
| CD163 | -0.73 | -0.22 | 1.39e-04 | 3.08e-01 |
| CD38 | -0.93 | -0.53 | 1.84e-04 | 3.63e-01 |
| SLC13A2 | 0.56 | 0.22 | 2.38e-04 | 3.68e-01 |
| TENT5C | -0.52 | -0.25 | 2.62e-04 | 3.68e-01 |
| CTSK | -0.72 | -0.42 | 2.72e-04 | 3.68e-01 |
| RFX5 | -0.36 | -0.17 | 2.77e-04 | 3.68e-01 |
| FKBP11 | -0.51 | -0.29 | 3.26e-04 | 3.68e-01 |
| WAS | -0.55 | -0.23 | 3.59e-04 | 3.68e-01 |
| ERICH5 | 0.32 | 0.11 | 3.66e-04 | 3.68e-01 |
| NLRP7 | -0.64 | -0.36 | 3.69e-04 | 3.68e-01 |
| AMPD1 | -0.59 | -0.13 | 3.72e-04 | 3.68e-01 |
| KCNJ10 | -0.37 | -0.29 | 4.02e-04 | 3.73e-01 |
| CD4 | -0.47 | -0.17 | 4.85e-04 | 3.73e-01 |
| B3GALT5-AS1 | 0.76 | 0.33 | 4.86e-04 | 3.73e-01 |
| BIRC3 | -0.73 | -0.47 | 5.03e-04 | 3.73e-01 |
| SPON2 | -0.42 | -0.19 | 5.30e-04 | 3.73e-01 |
| DOK3 | -0.54 | -0.37 | 6.15e-04 | 3.73e-01 |
| HOXC9 | -0.31 | 0.05 | 6.42e-04 | 3.73e-01 |
| SIGLEC7 | -0.25 | -0.19 | 6.43e-04 | 3.73e-01 |
| RHBDF2 | -0.48 | -0.23 | 6.53e-04 | 3.73e-01 |
| TACSTD2 | -0.48 | -0.08 | 6.74e-04 | 3.73e-01 |
| CTSH | -0.60 | -0.34 | 7.18e-04 | 3.73e-01 |
| POU2AF1 | -0.64 | -0.17 | 7.21e-04 | 3.73e-01 |
| CLDN8 | 1.51 | 0.79 | 7.23e-04 | 3.73e-01 |
| TRIM69 | -0.39 | -0.16 | 7.31e-04 | 3.73e-01 |
| MAN2B1 | -0.46 | -0.13 | 7.53e-04 | 3.73e-01 |
| IL10RA | -0.50 | -0.22 | 7.57e-04 | 3.73e-01 |
| IL18BP | -0.28 | -0.12 | 7.74e-04 | 3.73e-01 |
| ADA2 | -0.57 | -0.23 | 8.03e-04 | 3.73e-01 |
| KRT1 | -0.33 | 0.08 | 8.06e-04 | 3.73e-01 |

| SYMBOL | logFC | AveExpr | P.Value | adj.P.Val |
|---|---|---|---|---|
| POLH | -0.24 | -0.06 | 8.07e-04 | 3.73e-01 |
| ISL2 | -0.34 | -0.18 | 8.91e-04 | 3.73e-01 |
| AQP8 | 1.85 | 0.55 | 9.73e-04 | 3.73e-01 |
| TMEM156 | -0.64 | -0.25 | 9.81e-04 | 3.73e-01 |
| PSAT1 | -0.55 | -0.22 | 9.81e-04 | 3.73e-01 |
| PER2 | 0.61 | 0.25 | 9.97e-04 | 3.73e-01 |
| FAM30A | -0.58 | -0.39 | 1.05e-03 | 3.73e-01 |
| MAP3K8 | -0.74 | -0.15 | 1.07e-03 | 3.73e-01 |
| FOXD2 | 0.41 | 0.18 | 1.09e-03 | 3.73e-01 |
| CPA6 | 0.46 | 0.21 | 1.10e-03 | 3.73e-01 |
| GJA4 | -0.44 | -0.14 | 1.13e-03 | 3.73e-01 |
| CD180 | -0.75 | -0.54 | 1.14e-03 | 3.73e-01 |
| ENTPD1 | -0.55 | -0.31 | 1.14e-03 | 3.73e-01 |
| GUCA2B | 1.06 | 0.19 | 3.27e-03 | 3.93e-01 |
| CXCL9 | -1.18 | -1.03 | 4.10e-03 | 3.93e-01 |
| CHI3L1 | -1.66 | -1.27 | 4.42e-03 | 4.00e-01 |
| C4BPA | -1.32 | -1.10 | 4.73e-03 | 4.00e-01 |
| ABCA12 | -1.02 | -0.61 | 4.79e-03 | 4.00e-01 |
| CXCL13 | -1.79 | -1.64 | 5.96e-03 | 4.18e-01 |
| DEFB4A DEFB4B | -1.77 | -1.45 | 9.49e-03 | 4.51e-01 |
| FDCSP | -1.79 | -1.49 | 1.12e-02 | 4.52e-01 |
| CXCL10 | -1.17 | -0.81 | 1.35e-02 | 4.56e-01 |
| SLC26A2 | 1.17 | 0.44 | 1.45e-02 | 4.64e-01 |
| DMBT1 | -1.24 | -0.87 | 1.50e-02 | 4.64e-01 |
| SAA1 SAA2 SAA2-SAA4 | -1.36 | -1.28 | 1.65e-02 | 4.67e-01 |
| TCN1 | -1.17 | -1.06 | 1.66e-02 | 4.67e-01 |
| ABCG2 | 1.14 | 0.45 | 1.88e-02 | 4.67e-01 |
| MMP1 | -1.00 | -0.65 | 2.16e-02 | 4.67e-01 |
| IDO1 | -1.08 | -0.77 | 2.25e-02 | 4.69e-01 |
| S100A8 | -1.23 | -1.40 | 2.40e-02 | 4.74e-01 |
| SPRR1B | -1.05 | -0.64 | 2.66e-02 | 4.81e-01 |
| DUOX2 | -1.10 | -0.84 | 4.44e-02 | 5.15e-01 |
| TNIP3 | -1.25 | -1.47 | 4.74e-02 | 5.20e-01 |
| MMP3 | -1.26 | -0.91 | 4.79e-02 | 5.20e-01 |

Table C.2: *Toptable - change during treatment - Microarray - Golimumab*
*Toptables contain the first 50 DE genes according to the p-value and the selected genes passing the threshold:* $|log2\ fold\ change| > 1$ *AND p-value* $< 0.05$.

| SYMBOL | logFC | AveExpr | P.Value | adj.P.Val |
|---|---|---|---|---|
| GLRA2 | 1.06 | 0.30 | 1.16e-06 | 1.93e-02 |
| NTRK2 | 0.90 | 0.28 | 2.41e-06 | 1.93e-02 |
| CFI | -1.65 | -0.65 | 3.25e-06 | 1.93e-02 |
| ACKR4 ACKR4P1 | -1.57 | -0.56 | 6.53e-06 | 2.85e-02 |
| UNC13D | -0.49 | -0.21 | 8.01e-06 | 2.85e-02 |
| MTPAP | 0.43 | 0.04 | 1.02e-05 | 2.93e-02 |
| PNLIPRP2 | 1.72 | 0.47 | 1.15e-05 | 2.93e-02 |
| UBE2L6 | -0.66 | -0.24 | 1.49e-05 | 3.00e-02 |
| TAP2 | -0.57 | -0.23 | 1.52e-05 | 3.00e-02 |
| MT1M | 2.21 | 0.07 | 1.69e-05 | 3.01e-02 |
| CFB | -0.97 | -0.33 | 2.13e-05 | 3.45e-02 |
| LGALS2 | 1.08 | 0.45 | 3.98e-05 | 5.57e-02 |
| TCIRG1 | -0.55 | -0.21 | 4.40e-05 | 5.57e-02 |
| UCK2 | 0.51 | 0.06 | 4.41e-05 | 5.57e-02 |
| CPA6 | 0.96 | 0.29 | 4.70e-05 | 5.57e-02 |
| GABRB2 | 0.52 | 0.02 | 6.29e-05 | 6.99e-02 |
| PSMB9 | -0.75 | -0.36 | 7.52e-05 | 7.66e-02 |
| TAP1 | -0.62 | -0.34 | 7.76e-05 | 7.66e-02 |
| CDKL2 | 0.44 | 0.06 | 9.84e-05 | 9.16e-02 |
| FRAS1 | 0.59 | 0.23 | 1.03e-04 | 9.16e-02 |
| APBA1 | 0.41 | 0.15 | 1.19e-04 | 9.76e-02 |
| KYNU | -1.45 | -0.59 | 1.26e-04 | 9.76e-02 |
| SAMD9L | -0.99 | -0.43 | 1.26e-04 | 9.76e-02 |
| SYT17 | 0.60 | 0.15 | 1.35e-04 | 9.78e-02 |
| RNF213 | -0.52 | -0.30 | 1.38e-04 | 9.78e-02 |
| SMARCE1 | 0.40 | 0.02 | 1.55e-04 | 1.02e-01 |
| ZBP1 | -0.98 | -0.34 | 1.55e-04 | 1.02e-01 |
| HMGA2 | 0.34 | 0.09 | 2.14e-04 | 1.34e-01 |
| RTEL1 RTEL1-TNFRSF6B | -0.47 | -0.16 | 2.32e-04 | 1.34e-01 |
| IFITM3 | -0.72 | -0.24 | 2.44e-04 | 1.34e-01 |
| TMEM63C | 0.67 | 0.09 | 2.54e-04 | 1.34e-01 |
| BTN3A1 | -0.51 | -0.12 | 2.56e-04 | 1.34e-01 |
| RBM28 | 0.37 | 0.06 | 2.65e-04 | 1.34e-01 |
| CDHR1 | 1.44 | 0.45 | 2.77e-04 | 1.34e-01 |
| CRELD2 | -0.46 | -0.15 | 2.80e-04 | 1.34e-01 |
| POGK | 0.56 | 0.05 | 2.82e-04 | 1.34e-01 |
| PADI2 | 1.31 | 0.36 | 2.86e-04 | 1.34e-01 |
| LPCAT1 | -0.74 | -0.34 | 2.95e-04 | 1.34e-01 |
| RRN3 | 0.62 | 0.03 | 3.01e-04 | 1.34e-01 |
| SLC25A33 | 0.52 | 0.05 | 3.02e-04 | 1.34e-01 |
| CRNKL1 | 0.45 | -0.06 | 3.17e-04 | 1.36e-01 |
| RNF139 | 0.47 | 0.12 | 3.20e-04 | 1.36e-01 |
| MAT2A | 0.49 | 0.14 | 3.35e-04 | 1.38e-01 |
| TRIB2 | -0.75 | -0.29 | 3.47e-04 | 1.39e-01 |
| MT1G | 1.08 | 0.00 | 3.53e-04 | 1.39e-01 |
| MT1E | 0.92 | 0.01 | 3.65e-04 | 1.39e-01 |
| ZBTB39 | 0.39 | 0.09 | 3.68e-04 | 1.39e-01 |
| CCDC115 | 0.52 | 0.09 | 3.85e-04 | 1.42e-01 |
| ATPAF1 | 0.44 | 0.10 | 3.98e-04 | 1.45e-01 |
| TRAK2 | 0.56 | 0.13 | 4.14e-04 | 1.47e-01 |
| APOL1 | -1.07 | -0.50 | 5.37e-04 | 1.57e-01 |
| MT1F | 1.09 | 0.05 | 5.81e-04 | 1.57e-01 |
| CHI3L1 | -2.47 | -0.86 | 6.09e-04 | 1.57e-01 |

| SYMBOL | logFC | AveExpr | P.Value | adj.P.Val |
|---|---|---|---|---|
| CXCL1 | -1.50 | -0.70 | 8.77e-04 | 1.58e-01 |
| ABCB1 | 1.12 | 0.33 | 8.78e-04 | 1.58e-01 |
| HMGCS2 | 1.95 | 0.60 | 9.02e-04 | 1.58e-01 |
| CLDN8 | 2.50 | 0.58 | 9.73e-04 | 1.60e-01 |
| PCK1 | 2.40 | 1.01 | 1.00e-03 | 1.60e-01 |
| LAX1 | -1.12 | -0.33 | 2.05e-03 | 2.00e-01 |
| MMP7 | -1.72 | -0.52 | 2.12e-03 | 2.01e-01 |
| UGT2A3 | 1.45 | 0.24 | 2.19e-03 | 2.02e-01 |
| MFSD4A | 1.01 | 0.30 | 2.21e-03 | 2.02e-01 |
| OGN | 1.22 | 0.46 | 2.38e-03 | 2.03e-01 |
| CNTN3 | 1.05 | 0.27 | 2.43e-03 | 2.05e-01 |
| SAA1 SAA2 SAA2-SAA4 | -1.77 | -1.00 | 2.51e-03 | 2.08e-01 |
| CHP2 | 1.45 | 0.23 | 2.78e-03 | 2.18e-01 |
| TCN1 | -2.09 | -1.03 | 2.90e-03 | 2.23e-01 |
| B3GALT5-AS1 | 1.02 | 0.39 | 3.26e-03 | 2.27e-01 |
| TNIP3 | -2.02 | -1.26 | 3.34e-03 | 2.27e-01 |
| SLC26A2 | 2.00 | 0.48 | 3.52e-03 | 2.27e-01 |
| IL1R2 | 1.05 | 0.10 | 3.68e-03 | 2.29e-01 |
| EXPH5 | 1.14 | 0.32 | 3.90e-03 | 2.29e-01 |
| CXCL11 | -1.44 | -0.83 | 3.90e-03 | 2.29e-01 |
| IDO1 | -1.65 | -1.19 | 4.35e-03 | 2.40e-01 |
| FCRL5 | -1.01 | -0.28 | 5.22e-03 | 2.49e-01 |
| BEST2 | 1.18 | 0.21 | 5.46e-03 | 2.55e-01 |
| MZB1 | -1.10 | -0.32 | 5.99e-03 | 2.69e-01 |
| SLC16A9 | 1.12 | 0.42 | 6.47e-03 | 2.73e-01 |
| VSIG1 | -1.32 | -0.65 | 8.39e-03 | 2.97e-01 |
| ALDOB | -1.37 | -0.50 | 9.21e-03 | 3.08e-01 |
| CKB | 1.08 | 0.32 | 1.08e-02 | 3.18e-01 |
| CXCL9 | -1.16 | -0.48 | 1.24e-02 | 3.30e-01 |
| SERPINA3 | -1.50 | -0.77 | 1.31e-02 | 3.34e-01 |
| C4BPA | -1.41 | -0.89 | 1.34e-02 | 3.36e-01 |
| C4BPB | -1.05 | -0.71 | 1.34e-02 | 3.36e-01 |
| ABCA12 | -1.13 | -0.55 | 1.47e-02 | 3.43e-01 |
| DUOXA2 | -1.54 | -0.84 | 1.52e-02 | 3.46e-01 |
| HEPACAM2 | 1.43 | 0.03 | 1.56e-02 | 3.48e-01 |
| IGKV4-1 | -1.05 | -0.38 | 1.65e-02 | 3.53e-01 |
| CWH43 | 1.39 | 0.56 | 1.85e-02 | 3.59e-01 |
| FAM3B | -1.09 | -0.67 | 2.08e-02 | 3.74e-01 |
| ADH1C | 1.44 | 0.53 | 2.09e-02 | 3.74e-01 |
| KLK6 | -1.10 | -0.43 | 2.10e-02 | 3.76e-01 |
| MMP3 | -1.72 | -0.93 | 2.14e-02 | 3.77e-01 |
| FCGR3A FCGR3B | -1.15 | -0.55 | 2.79e-02 | 4.05e-01 |
| REG1A | -2.15 | -1.83 | 2.87e-02 | 4.08e-01 |
| SERPINB3 | -1.76 | -0.93 | 3.39e-02 | 4.25e-01 |
| CCL18 | -1.09 | -0.51 | 3.47e-02 | 4.28e-01 |
| REG1B | -2.28 | -2.09 | 3.54e-02 | 4.30e-01 |
| S100A8 | -1.66 | -0.98 | 3.54e-02 | 4.30e-01 |
| DUOX2 | -1.30 | -0.70 | 3.93e-02 | 4.44e-01 |
| MMP10 | -1.16 | -1.07 | 3.95e-02 | 4.45e-01 |
| DEFB4A DEFB4B | -1.39 | -0.84 | 4.08e-02 | 4.49e-01 |
| SERPINB4 | -1.21 | -0.62 | 4.52e-02 | 4.70e-01 |

Table C.3: *Toptable - change during treatment - Microarray - Vedolizumb*
*Toptables contain the first 50 DE genes according to the p-value and the selected genes passing the threshold:* $|log2\ fold\ change| > 1$ *AND p-value* $< 0.05$.

| SYMBOL | logFC | AveExpr | P.Value | adj.P.Val |
|---|---|---|---|---|
| SEC24A | -0.48 | -0.05 | 2.41e-04 | 9.97e-01 |
| NCR3LG1 | 0.62 | 0.14 | 3.62e-04 | 9.97e-01 |
| GBE1 | -0.42 | -0.07 | 4.18e-04 | 9.97e-01 |
| GALNT8 | 0.74 | 0.08 | 4.59e-04 | 9.97e-01 |
| ZCCHC12 | 0.35 | 0.06 | 4.93e-04 | 9.97e-01 |
| UHRF1 | -0.54 | -0.17 | 5.27e-04 | 9.97e-01 |
| CLYBL | 0.47 | 0.15 | 5.40e-04 | 9.97e-01 |
| AMT | 0.50 | 0.16 | 6.56e-04 | 9.97e-01 |
| PSAT1 | -0.70 | -0.06 | 6.97e-04 | 9.97e-01 |
| FAAH | 0.46 | 0.14 | 7.28e-04 | 9.97e-01 |
| ALOX12B | 0.30 | 0.02 | 7.38e-04 | 9.97e-01 |
| PMCH | -0.43 | -0.12 | 7.90e-04 | 9.97e-01 |
| TRABD2A | 0.50 | 0.14 | 8.27e-04 | 9.97e-01 |
| RAP1GAP | 0.66 | 0.23 | 8.32e-04 | 9.97e-01 |
| TDRD5 | 0.33 | 0.02 | 8.42e-04 | 9.97e-01 |
| MCM10 | -0.38 | -0.13 | 9.35e-04 | 9.97e-01 |
| CDC6 | -0.54 | -0.12 | 9.83e-04 | 9.97e-01 |
| RCL1 | 0.27 | -0.07 | 1.11e-03 | 9.97e-01 |
| MELK | -0.53 | -0.13 | 1.29e-03 | 9.97e-01 |
| BEST2 | 1.18 | 0.59 | 1.32e-03 | 9.97e-01 |
| LIPG | -0.51 | 0.02 | 1.41e-03 | 9.97e-01 |
| UCA1 | -0.96 | -0.09 | 1.41e-03 | 9.97e-01 |
| CDHR1 | 0.93 | 0.49 | 1.43e-03 | 9.97e-01 |
| FNDC3B | -0.29 | -0.04 | 1.50e-03 | 9.97e-01 |
| CA7 | 1.03 | 0.25 | 1.54e-03 | 9.97e-01 |
| DNAJC3 | -0.39 | -0.06 | 1.74e-03 | 9.97e-01 |
| NCBP2-AS1 NCBP2AS2 | 0.32 | 0.03 | 1.80e-03 | 9.97e-01 |
| KCNJ11 | 0.25 | 0.05 | 1.80e-03 | 9.97e-01 |
| RETNLB | 1.39 | 0.44 | 1.88e-03 | 9.97e-01 |
| SPDL1 | -0.34 | -0.07 | 2.05e-03 | 9.97e-01 |
| PLEKHG3 | 0.33 | 0.12 | 2.06e-03 | 9.97e-01 |
| DHFR DHFRP1 | -0.53 | -0.12 | 2.07e-03 | 9.97e-01 |
| CALU | -0.32 | -0.12 | 2.21e-03 | 9.97e-01 |
| BTBD6 | 0.25 | 0.05 | 2.22e-03 | 9.97e-01 |
| PHTF1 | -0.32 | -0.06 | 2.27e-03 | 9.97e-01 |
| DEFB4A DEFB4B | -2.19 | -0.68 | 2.53e-03 | 9.97e-01 |
| CRACR2A | 0.43 | 0.16 | 2.58e-03 | 9.97e-01 |
| FAHD2A | 0.25 | 0.06 | 2.73e-03 | 9.97e-01 |

| SYMBOL | logFC | AveExpr | P.Value | adj.P.Val |
|---|---|---|---|---|
| MTFR2 | -0.33 | -0.07 | 2.78e-03 | 9.97e-01 |
| FAM189A1 | 0.50 | 0.14 | 2.86e-03 | 9.97e-01 |
| CHEK1 | -0.41 | -0.13 | 2.92e-03 | 9.97e-01 |
| ELL2 | -0.40 | -0.12 | 3.00e-03 | 9.97e-01 |
| JAZF1 | -0.39 | -0.06 | 3.02e-03 | 9.97e-01 |
| TRMT9B | 0.24 | 0.11 | 3.12e-03 | 9.97e-01 |
| PPP3R1 | -0.40 | 0.00 | 3.16e-03 | 9.97e-01 |
| WNK2 | 0.33 | 0.15 | 3.23e-03 | 9.97e-01 |
| ADK | -0.33 | -0.02 | 3.51e-03 | 9.97e-01 |
| PDIA6 | -0.28 | -0.04 | 3.51e-03 | 9.97e-01 |
| RTKN | 0.35 | 0.06 | 3.53e-03 | 9.97e-01 |
| GPIHBP1 | 0.28 | 0.02 | 3.54e-03 | 9.97e-01 |
| C4BPB | -1.19 | -0.29 | 4.52e-03 | 9.97e-01 |
| SERPINB3 | -1.92 | 0.30 | 4.54e-03 | 9.97e-01 |
| CLDN8 | 1.47 | 0.60 | 5.38e-03 | 9.97e-01 |
| TCN1 | -1.45 | -0.31 | 7.75e-03 | 9.97e-01 |
| C4BPA | -1.29 | -0.16 | 1.13e-02 | 9.97e-01 |
| VSIG1 | -1.09 | -0.01 | 1.33e-02 | 9.97e-01 |
| SERPINB7 | -1.81 | 0.19 | 1.51e-02 | 9.97e-01 |
| HEPACAM2 | 1.31 | 0.49 | 1.59e-02 | 9.97e-01 |
| PPBP | -1.07 | -0.09 | 1.80e-02 | 9.97e-01 |
| MT1M | 1.06 | 0.26 | 2.03e-02 | 9.97e-01 |
| SAA1 SAA2 SAA2-SAA4 | -1.47 | -0.30 | 2.16e-02 | 9.97e-01 |
| NOS2 | -1.05 | -0.24 | 2.29e-02 | 9.97e-01 |
| ABCA12 | -1.09 | -0.09 | 2.39e-02 | 9.97e-01 |
| CLCA1 | 1.10 | 0.27 | 2.62e-02 | 9.97e-01 |
| CXCL8 | -1.32 | -0.49 | 2.90e-02 | 9.97e-01 |
| IL1B | -1.20 | -0.12 | 2.96e-02 | 9.97e-01 |
| PROK2 | -1.44 | -0.34 | 3.00e-02 | 9.97e-01 |
| REG1A | -1.52 | -0.15 | 3.16e-02 | 9.97e-01 |
| TNIP3 | -1.27 | -0.41 | 3.37e-02 | 9.97e-01 |
| OSM | -1.09 | -0.24 | 3.49e-02 | 9.97e-01 |
| MMP3 | -1.84 | -0.84 | 3.60e-02 | 9.97e-01 |
| DUOXA2 | -1.24 | -0.06 | 3.73e-02 | 9.97e-01 |
| AQP9 | -1.38 | -0.33 | 3.78e-02 | 9.97e-01 |
| MMP7 | -1.01 | -0.41 | 4.72e-02 | 9.97e-01 |
| GUCA2A | 1.02 | 0.19 | 4.95e-02 | 9.97e-01 |

Table C.4: *Toptable - change during treatment - Microarray - Placebo*
*Toptables contain the first 50 DE genes according to the p-value and the selected genes passing the threshold: $|log2\ fold\ change| > 1$ AND p-value $< 0.05$.*

# Appendix D

# Software details: all R packages used

- R version 4.2.1 (2022-06-23 ucrt), `x86_64-w64-mingw32`
- Base packages: base, datasets, graphics, grDevices, grid, methods, parallel, stats, stats4, utils
- Other packages: affy 1.74.0, affycoretools 1.68.1, AnnotationDbi 1.58.0, AnnotationFilter 1.20.0, ArrayExpress 1.56.0, Biobase 2.56.0, BiocGenerics 0.42.0, BiocParallel 1.30.3, biomaRt 2.52.0, Biostrings 2.64.0, caret 6.0-93, DEFormats 1.24.0, DESeq2 1.36.0, doParallel 1.0.17, dplyr 1.0.9, e1071 1.7-11, edgeR 3.38.1, EnsDb.Hsapiens.v79 2.99.0, ensembldb 2.20.2, forcats 0.5.1, foreach 1.5.2, gdata 2.18.0.1, genefilter 1.78.0, GenomeInfoDb 1.32.2, GenomicFeatures 1.48.3, GenomicRanges 1.48.0, GEOquery 2.64.2, ggfortify 0.4.14, ggplot2 3.3.6, ggvenn 0.1.9, glmnet 4.1-4, gridExtra 2.3, IRanges 2.30.0, iterators 1.0.14, kableExtra 1.3.4, knitr 1.39, lattice 0.20-45, limma 3.52.2, Matrix 1.4-1, MatrixGenerics 1.8.1, matrixStats 0.62.0, mgcv 1.8-40, nlme 3.1-158, oligo 1.60.0, oligoClasses 1.58.0, openxlsx 4.2.5, patchwork 1.1.2, pheatmap 1.0.12, pROC 1.18.0, purrr 0.3.4, pvca 1.36.0, RColorBrewer 1.1-3, readr 2.1.2, readxl 1.4.0, S4Vectors 0.34.0, stringr 1.4.0, SummarizedExperiment 1.26.1, sva 3.44.0, tibble 3.1.7, tidyr 1.2.0, tidyverse 1.3.1, viridis 0.6.2, viridisLite 0.4.0, WriteXLS 6.4.0, XVector 0.36.0
- Loaded via a namespace (and not attached): affxparser 1.68.1, affyio 1.66.0, annotate 1.74.0, AnnotationForge 1.38.0, apcluster 1.4.10, assertthat 0.2.1, backports 1.4.1, base64enc 0.1-3, BiocFileCache 2.4.0, BiocIO 1.6.0, BiocManager 1.30.18, biovizBase 1.44.0, bit 4.0.4, bit64 4.0.5, bitops 1.0-7, blob 1.2.3, boot 1.3-28.1, broom 1.0.0, BSgenome 1.64.0, cachem 1.0.6, Category 2.62.0, caTools 1.18.2, cellranger 1.1.0, checkmate 2.1.0, class 7.3-20, cli 3.3.0, cluster 2.1.3, codetools 0.2-18, colorspace 2.0-3, compiler 4.2.1, crayon 1.5.1, curl 4.3.2, data.table 1.14.2, DBI 1.1.3, dbplyr 2.2.1, DelayedArray 0.22.0, deldir 1.0-6, dichromat 2.0-0.1, digest 0.6.29, doRNG 1.8.2, ellipsis 0.3.2, evaluate 0.15, fansi 1.0.3, fastmap 1.1.0, ff 4.0.7, filelock 1.0.2, foreign 0.8-82, Formula 1.2-4, fs 1.5.2, future 1.28.0, future.apply 1.9.1,

gcrma 2.68.0, geneplotter 1.74.0, generics 0.1.3, GenomeInfoDbData 1.2.8, GenomicAlignments 1.32.0, GGally 2.1.2, ggbio 1.44.1, Glimma 2.6.0, globals 0.16.1, glue 1.6.2, GO.db 3.15.0, GOstats 2.62.0, gower 1.0.0, gplots 3.1.3, graph 1.74.0, GSEABase 1.58.0, gtable 0.3.0, gtools 3.9.2.2, hardhat 1.2.0, haven 2.5.0, Hmisc 4.7-0, hms 1.1.1, htmlTable 2.4.1, htmltools 0.5.2, htmlwidgets 1.5.4, httr 1.4.3, hwriter 1.3.2.1, igraph 1.3.5, interp 1.1-3, ipred 0.9-13, jpeg 0.1-9, jsonlite 1.8.0, KEGGREST 1.36.3, KernSmooth 2.23-20, latticeExtra 0.6-30, lava 1.6.10, lazyeval 0.2.2, lifecycle 1.0.1, listenv 0.8.0, lme4 1.1-29, locfit 1.5-9.5, lubridate 1.8.0, magrittr 2.0.3, MASS 7.3-57, memoise 2.0.1, minqa 1.2.4, ModelMetrics 1.2.2.2, modelr 0.1.8, munsell 0.5.0, nloptr 2.0.3, nnet 7.3-17, OrganismDbi 1.38.1, parallelly 1.32.1, PFAM.db 3.15.0, pillar 1.8.0, pkgconfig 2.0.3, plyr 1.8.7, png 0.1-7, preprocessCore 1.58.0, prettyunits 1.1.1, prodlim 2019.11.13, progress 1.2.2, ProtGenerics 1.28.0, proxy 0.4-27, R.methodsS3 1.8.2, R.oo 1.25.0, R.utils 2.12.0, R6 2.5.1, rappdirs 0.3.3, RBGL 1.72.0, Rcpp 1.0.8.3, RCurl 1.98-1.7, recipes 1.0.1, ReportingTools 2.36.0, reprex 2.0.1, reshape 0.8.9, reshape2 1.4.4, restfulr 0.0.15, Rgraphviz 2.40.0, rjson 0.2.21, rlang 1.0.3, rmarkdown 2.14, rngtools 1.5.2, rpart 4.1.16, Rsamtools 2.12.0, RSQLite 2.2.14, rstudioapi 0.13, rtracklayer 1.56.1, rvest 1.0.2, scales 1.2.0, shape 1.4.6, splines 4.2.1, stringi 1.7.6, survival 3.3-1, svglite 2.1.0, systemfonts 1.0.4, tidyselect 1.1.2, timeDate 4021.104, tools 4.2.1, tzdb 0.3.0, utf8 1.2.2, VariantAnnotation 1.42.1, vctrs 0.4.1, vsn 3.64.0, WebGestaltR 0.4.4, webshot 0.5.3, whisker 0.4, withr 2.5.0, xfun 0.31, XML 3.99-0.10, xml2 1.3.3, xtable 1.8-4, yaml 2.3.5, zip 2.2.1, zlibbioc 1.42.0