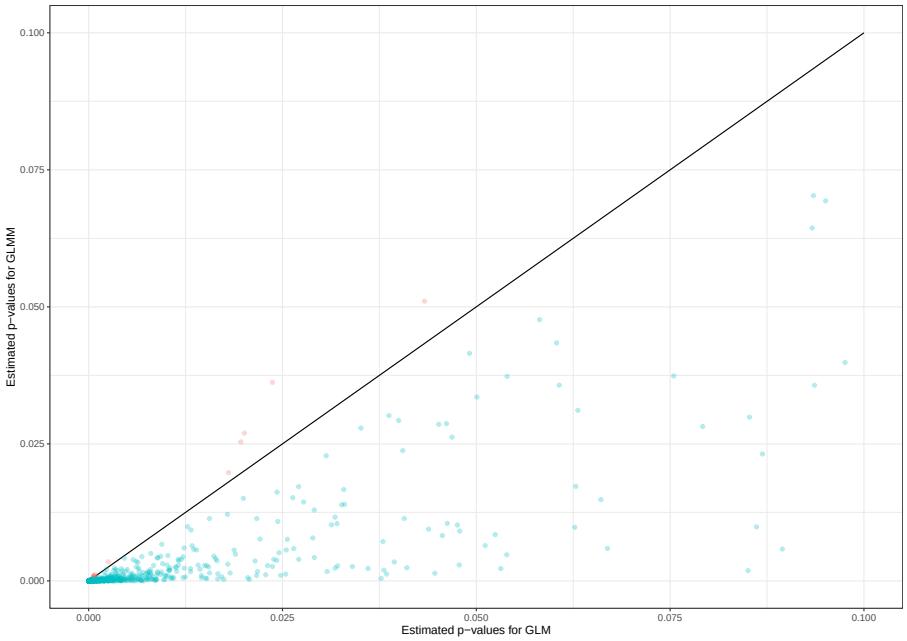
Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

Sebastian Ø. Ankill

# Statistical methods for analysis of gene expression count data applied to a Crohn's disease dataset

Master's thesis in Industrial Mathematics
Supervisor: Mette Langaas
Co-supervisor: Atle van Beelen Granlund
June 2022




NTNU
Norwegian University of
Science and Technology

Sebastian Ø. Ankill

# Statistical methods for analysis of gene expression count data applied to a Crohn's disease dataset

Master's thesis in Industrial Mathematics
Supervisor: Mette Langaas
Co-supervisor: Atle van Beelen Granlund
June 2022

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

# Preface

I would like to thank my supervisor Mette Langaas and co-supervisor Atle van Beelen Granlund for great guidance throughout the master's degree. I would also like to thank Joakim Wais for moral support and Oskar Elfmark for proofreading.

**Abstract**

This thesis is motivated by the need to use valid and powerful statistical models and methods in the analysis of gene expression count data in medicine, in particular when multiple observations from the same individual are present in the data. Analysis is performed on gene expression measurement of tissue samples from patients with Crohn's disease. General statistical models and methods is presented to explain the underlying theory of two established modern pipelines for analysis of gene expression count data. The two methods are limma-voom, which is based on a linear model, and DESeq2, which is based on a generalized linear model. An addition of a random intercept for modelling correlated data is studied for both models. The linear mixed model is already included for limma-voom, but DESeq2 does not support the use of a generalized linear mixed model. In a simulation study generalized linear models and generalized linear mixed models are fit to data simulated by generalized linear mixed model, and the model fit is evaluated. The generalized linear mixed model is found to be more powerful than the generalized linear model with this setup, and alternatives to model correlation for DESeq2 is discussed. The generalized linear model is found to be conservative for highly correlated data in the simulation. When the linear mixed model and the generalized linear mixed model are fit to the dataset for analysis a correlation is estimated for both models. Their fits are compared to the fit for the limma-voom and DESeq2 pipelines. Some additional genes are found to be statistically significant for the models that model correlated data compared to the ones that do not, but these are not the same for both models. All models give decent fit for the provided dataset, and genes of interest are presented with comparisons across models. A top list with consensus statistically significant genes for all four models for a given contrast are presented, and multiple of the genes are recognized as previously found potentially relevant genes for Crohn's disease. Alternate approaches and further extension of our generalized linear mixed model are discussed.

**Sammendrag**

Denne oppgaven er motivert av behovet for å bruke gyldige og sterke statistiske modeller og metoder for analyse av genuttrykksdata i medisin, spesielt når flere observasjoner fra samme individ er til stede i data. Det er utført analyse av genuttrykk telledata fra vevsprøver fra pasienter med Crohns sykdom. Generelle statistiske modeller og metoder er presentert for å forklare den underliggende teorien til to etablerte moderne metoder for analyse av genuttrykk telledata. De to metodene heter limma-voom, som er basert på en lineær modell, og DESeq2, som er basert på en generalisert lineær modell. Et tillegg av tilfeldig skjæringspunkt for modellering av korrelerte data er studert for begge modellene. Den lineære blandede modellen er allerede inkludert for limma-voom, mens DESeq2 støtter ikke bruken av en generalisert lineær blandet modell. I en simuleringsstudie tilpasses generaliserte lineære modeller og generaliserte lineære blandede modeller til data simulert av en generalisert lineær blandet modell, og modelltilpasning er evaluert. Den generaliserte lineære blandede modellen er funnet til å være sterkere enn den generaliserte lineære modellen med dette oppsettet, og alternativer til å modellere korrelasjon for DESeq2 er diskutert. Den generaliserte lineære modellen er funnet til å være konservativ for svært korrelerte data i simuleringen. Når den lineære blandede modellen og den generaliserte lineære blandede modellen er tilpasset datasettet for analyse estimeres en korrelasjon for begge modellene. Deres tilpasning er sammenlignet til limma-voom og DESeq2. Noen ekstra gener er funnet til å være statistisk signifikante for modellene som modellerer korrelerte data, sammenlignet med de som ikke gjør det, men disse er ikke de samme for begge modellene. Alle modellene gir gode tilpasninger for det gitte datasettet, og gener av interesse er presentert med sammenligninger på tvers av modeller. En toppliste med konsensus statistisk signifikante gener for alle fire modellene for en gitt kontrast er presentert, og flere av genene har tidligere blitt funnet til å potentielt være relevante for Crohns sykdom. Potensielle endringer for metoder og videre utvidelse av vår generaliserte lineære blandede modell er diskutert.

# Contents

# Abbreviations

**BH** Benjamini-Hochberg step-down method

**CD** Crohn's disease

**CDF** cumulative distribution function

**DESeq2** differential gene expression analysis based on the negative binomial distribution

**DNA** deoxyribonucleic acid

**EDM** exponential dispersion model

**FDR** false discovery rate

**GLM** generalized linear model

**GLMM** generalized linear mixed model

**IBD** inflammatory bowel disease

**LFC** log fold change

**limma** linear models for microarray and RNA-Seq data

**LM** linear model

**LMM** linear mixed model

**MA** mean average

**MAP** maximum a posteriori

**ML** maximum likelihood

**RNA** ribonucleic acid

**RNA-Seq** Ribonucleic acid sequencing

**UC** ulcerative colitis

**voom** variance modeling at the observational level

# 1 Introduction

This master thesis is motivated by the need to use valid and powerful statistical models and methods in the analysis of gene expression count data in medicine, in particular when multiple observations are from the same individual is present in the data. Crohn's disease is an inflammatory bowel disease which can inflame different sections of the intestine, both colon and ileum. Another inflammatory bowel disease, ulcerative colitis, can only inflame the colon. This makes us question if there is a difference between the Crohn's disease observed in colon and the one observed in ileum. We specifically want to compare tissue that is visibly inflamed with tissue that is not, from patients affected with Crohn's disease. With this design we might observe a difference in gene expression that makes the disease active in the affected tissue, and whether it is dormant or not existing in visually unaffected tissue. To measure whether Crohn's disease is dormant or not present in unaffected tissue we also need to compare against tissue samples from people that never have been observed with an inflammatory bowel disease. This will be our healthy reference group.

From a medical perspective our main interest is to examine if there is a significant difference in gene expression count data for unaffected and affected tissue between colon and ileum, while controlling for the gene expression in healthy individuals. Additionally we want to identify genes where this difference is present.

From a statistical point of view the main challenge is modeling the data, such that powerful analysis can be performed. We will particularly examine how to model correlated data. There is also a massive multiple testing problem, where we need to asses how we can control some chosen error measure.

In Section 2 we will introduce some general genetic concepts and explain how we obtain gene expression count data from a patients tissue sample. We will then describe what the disease we are modeling does to the body, and introduce the data that we later will analyze. In Section 3 we will bring up general statistical concept relevant to performing hypothesis testing for gene expression count data. In Section 4 we will go into depth of how two different pipelines, limma-voom and DESeq2, can be used to model the dataset presented. In Section 5 we will observe how correlation influences a negative binomial regression model and examine an extension of the principles used by DESeq2. The patient data will be analyzed in Section 6, before we discuss the relevance of our results and alternative choices that could have been made to obtain the results 7. And finally, we will conclude in Section 8.

Section 2, 3.1 - 3.4, 4.1 are revised versions with a different focus from Ankill (2021). Section 6 is inspired by Ankill (2021), where we now will examine two additional statistical models. The code for simulations in Section 5, and analysis in Section 6 can be found in Ankill (2022).

For notation boldface will be used to denote vectors. Further hat will be used to denote the estimated value of a variable.

# 2 Biological background

To get an idea of how we can statistically analyse the functionality of a specific cell we want to go into detail of where what we analyze come from. We also want to make it apparent that the reads we have available are dependent on the DNA from the sample holder, resources available in the cell and randomness. Then we will observe the dataset that originates from this process and its characteristics.

## 2.1 DNA

DNA is the foundation of life, it consists of the genetic building blocks that makes every organism unique. A single strand of DNA consists of a backbone of alternating five-carbon sugar (deoxyribose) and a phosphate. Each sugar in this backbone is connected to one of four different

nucleotides, also called bases. These are adenine, guanine, cytosine and thymine. DNA in its natural state is double stranded, where the strands are connected by hydrogen bonds between the bases. Each base has a complementary base that it connects to, adenine connects to thymine and guanine connects to cytosine. The sugar in the backbone of DNA consists of five carbons and three oxygens. When we number the carbons 1', 2' and so on, then the phosphates are connected to the 3'- and the 5'-carbon ends, which defines the orientation of a single strand of DNA (Datta and Nettleton 2014, Chap. 1).

## 2.2 Gene expression

DNA exists in the cells in the human body, and as a consequence of different cells having different functions, not all parts of the DNA are relevant for all cells. Relevant parts of the DNA in a cell is transcribed to messenger RNA (mRNA) and proteins are created from the sequence of bases in the mRNA. mRNA is quite similar to DNA in construction, but it has a ribose backbone instead of the deoxyribose used in DNA. Both ribose and deoxyribose are sugars, difering only in ribose having a OH group instead of a hydrogen in a specific position. A backbone can be interpreted as a longer string of the same sugar connected to each other, where each sugar is connected to a protein. It also has uracil as the complementary base of adenine instead of thymine, and is in most cases single stranded. The mRNA gets translated to proteins, and this process is essential in the central dogma of molecular biology, stating that mRNA is created from DNA through transcription, which in turn acts as a template for protein translation. In this thesis we will focus on the analysis of mRNA, the type of RNA that encodes proteins. A gene is defined as a specific region within the DNA that carries the code needed to form a mRNA that can be translated to a specific protein, or another functional unit within the cell. The transcription is initiated by a set of proteins, called the transcription machinery, binding to the DNA. This machinery then commences to temporarily split the two strands of DNA. This happens at a specific sequence of bases called a gene promoter region, which is an indicator of where to start the transcription of a specific gene. It will then continue transcribing by connecting the complementary sequence of the DNA until it hits a specific stop sequence. The genes transcribed in a specific cell at a specific state can give a good snapshot of a cells activity in that particular state. By analyzing all mRNAs within a sample, we get a good indication of the cellular mechanisms that are at play within the sample. This analysis is called whole genome gene expression (WGGE) analysis, and today there are several methods available for such analyses. In this thesis the method RNA sequencing (RNA-Seq) was used to perform WGGE.

## 2.3 RNA sequencing

The RNA was extracted using the mirVANA miRNA isolation kit, the same procedure as was done in Granlund et al. (2013). To analyze the RNA it first has to be chemically or enzymatically fragmented to smaller sizes. This RNA is then used to create single stranded cDNA, which is chemically similar to DNA, but since it is recoded from a segment of RNA it is much smaller. Then a double stranded cDNA is created by complementing the bases from the first strand, while also adding an adapter sequence to the ends of the cDNA for later use. PCR (polymerase chain reaction) uses parts of the adapter sequence as primers to amplify the cDNA, as the extracted amount is not enough for raw detection. The data we will analyze is from the Illumina platform (Illumina, Inc. 2022). There the cDNA to go through Bridge amplification before it is sequenced by synthesis, which means single strands are being read by fluorescently labeled deoxynucleoside triphosphate (dNTP) when the complementary strand is generated. During the sequence a probabilistic model determines which bases that are complemented to the sequence by synthesis, this is called base calling and creates a digital representation of the original RNA sample. After having obtained the raw RNA reads we want to map the reads to a reference genome, and register count data for different genes. This count data can be considered as counts of how much the different genes were expressed in the extracted sample, and is called gene expression count data (Van den Berge et al. 2019). A general overview of how we obtain the digital representation of the DNA can be found in Figure 1.

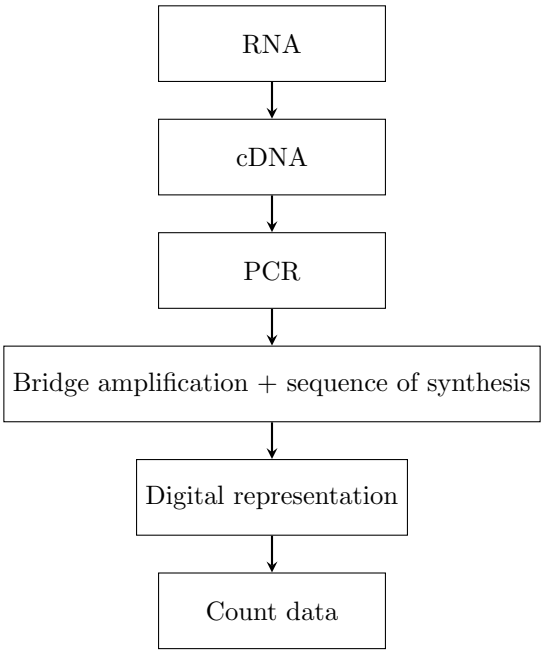



Figure 1: An overview of the process from extracted RNA to count data based on gene expression.

## 2.4 Inflammatory bowel disease

Inflammatory bowel disease (IBD) is a term used to describe diseases in the digestive tract that involve chronic inflammation. The two most prominent diseases covered by this umbrella term are ulcerative colitis (UC) and Crohn's disease (CD). While UC mainly affects the colon, CD can affect any part of the gastrointestinal (GI) tract. Treatment is mostly dependent on location of the inflammation, where CD in colon is treated similarly as UC. UC is only found in colon, whilst CD can be found in both colon and ileum. Figure 2 shows where in the colon CD and UC is most commonly found.

We will only do analysis based on CD in this project, and will therefore not go into more detail about UC. CD can affect any part of the gastrointestinal tract and tissue in close proximity. CD can result in tissue fibrosis, affecting the thickness of the intestinal wall to the point of it blocking the flow of digested content. Another serious complication is creation of fistulas, forming abnormal connections either between different parts of the gastrointestinal tract, from the GI to other organs, or from the GI tract and to the skin (The Crohn's and Colitis Foundation of America (CCFA) 2014).

## 2.5 Crohn's disease dataset

Table 1: Characterization of samples for Crohn's disease dataset. $H$ is our healthy group, $U$ corresponds to unaffected, which means Crohn's disease has been observed in the patient but not in the tissue the gene expression is based on. $A$ corresponds to affected, which mean Crohn's disease has been observed in the tissue the gene expression is based on. Colon and Ileum are two different sections of the digestive tract.

| | H | U | A |
|---|---|---|---|
| Colon | 20 | 24 | 20 |
| Ileum | 14 | 13 | 16 |

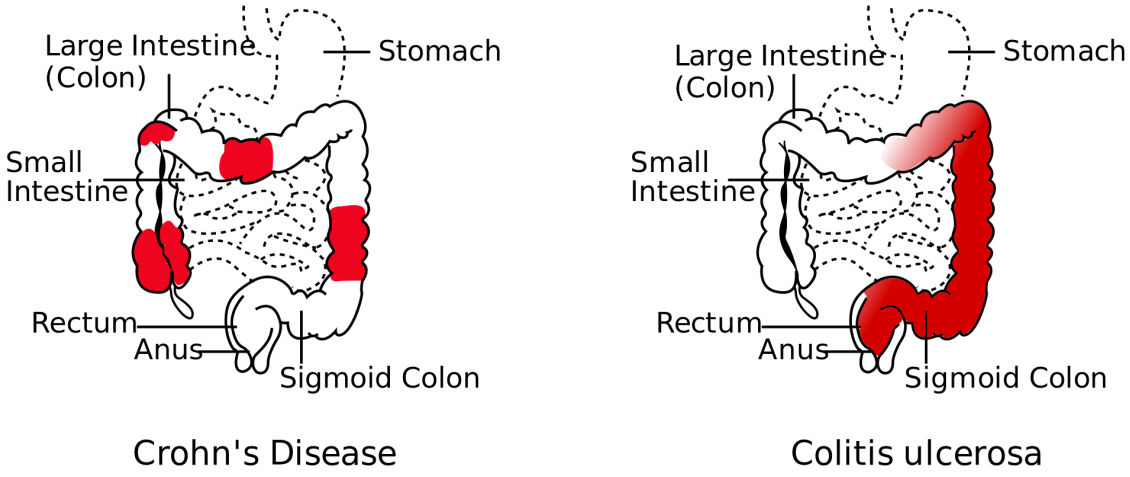



Figure 2: Figure of colon, large intestine, and ileum, small intestine. With focus on colon, as both Crohn's disease and ulcerative colitis can manifest there. The figure depicts the most common areas inflamed, marked by red, for Crohn's disease and ulcerative colitis in the colon. Colitis ulcerosa is a different name for ulcerative colitis.
**Source:** https://commons.wikimedia.org/wiki/File:Crohn%27s_Disease_vs_Colitis_ulcerosa.svg, licensed under CC BY-SA 3.0.

Table 2: Number of observations for all the patients in the Crohn's disease dataset.

| Number of observations | 1 | 2 |
|---|---|---|
| Number of patients | 77 | 15 |

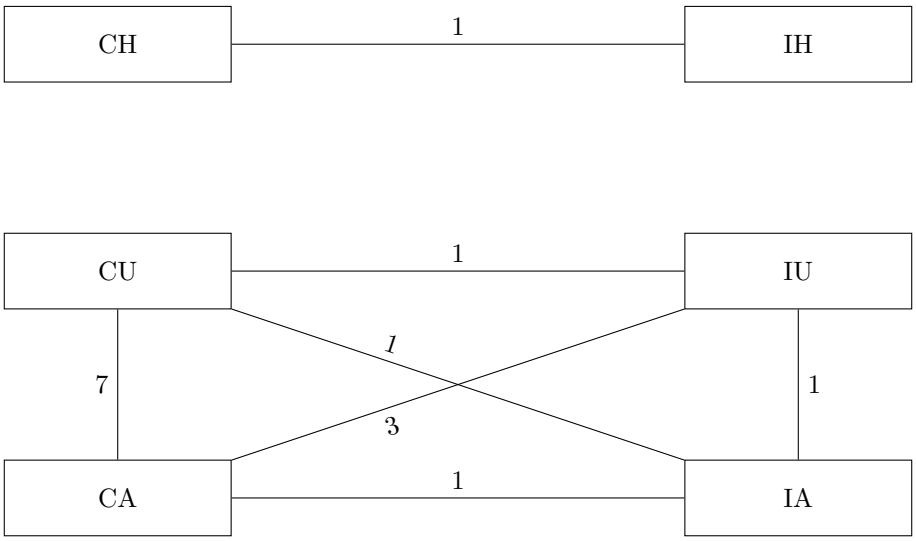



Figure 3: Figure that shows where and how many extractions were done for patients with multiple extractions.

The dataset of interest originates from tissue samples from patients with Crohn's disease, ulcerative colitis and healthy individuals. As we will only do analysis based on Crohn's disease we do not include the data concerning ulcerative colitis in our dataset. The observations are from colon, noted as C, and ileum, noted as I. The patients were either healthy, noted as H, unaffected, noted as U, or affected, noted as A. Healthy means that neither Crohn's disease or ulcerative colitis have been observed in that individual. Unaffected implies that Crohn's disease has been observed in the patient, but not in the tissue where the sample was extracted. Affected means that the sample originates from a tissue where Crohn's disease was observed.

The gene expression count data is on the form of an integer per gene per individual, the counts range from 0 to 2000000, where the mean of all counts is 360 before any filtering is done. It is common for gene expression count to have high variation across individuals due to the randomization inherent in amplification and different underlying genetics.

The number of observations for the combinations of where in the intestine the tissue was extracted from and observed inflammation are shown in Table 1. Some patients have multiple observations, the number of these can be seen in Table 2. In Figure 3 we can observe from where the pairs of observations from the patients with multiple observations originates. We should note that these tables are found after the removal of some patients due to reasons specified in Section 6.2. We found this to be the most convenient way, as then we only have one set of figures and they correspond to the final dataset used for finding results.

# 3 Statistical background

In this chapter we will examine the requirements for a model to be considered a generalized linear model (GLM), and show that the normal distribution satisfies these requirements. We will also show that the negative binomial response distribution satisfies these requirements under certain assumptions. We will cover how to estimate parameters for a GLM and how we can test hypotheses based on these estimates. The general concept of multiple testing and how this can be used to control false discovery rate will also be presented. Linear mixed models (LMM) will be introduced and their concepts explained. We also want to introduce ridge regression as a regularization technique and the bayesian interpretation of this. Finally we want to show how analysis of residuals can be done for the normal distribution and through quantile residuals for GLM.

## 3.1 Generalized linear model

A GLM is a model where the response, $Y$, is modeled a distribution that comes from the EDM (exponential dispersion model) family. These can be both discrete and continuous models, which makes this framework quite versatile.

GLMs consist of a random and a systematic component. The random component of an EDM has a probability distribution of the form

$$\mathcal{P}(y;\theta,\phi) = a(y,\phi)\exp\left\{\frac{y\theta-\kappa(\theta)}{\phi}\right\},$$

where

- $\theta$ is the canonical parameter,
- $\kappa(\theta)$ is the cumulant function,
- $\phi > 0$ is the dispersion parameter, and
- $a(y,\phi)$ is a normalizing function.

With this parameterization we can find the mean and variance by

$$E[Y] = \mu = \frac{d\kappa(\theta)}{d\theta}, \quad \text{Var}[Y] = \phi V(\mu) = \phi \frac{d^2\kappa(\theta)}{d\theta^2}, \tag{1}$$

where $V(\mu) = \frac{d\mu}{d\theta}$ is defined as the variance function. It can be shown that the variance function uniquely determines the distribution of the EDM, as it determines $\kappa(\theta)$ up to an additive constant (Dunn and Smyth 2018, Chap. 5.3).

Let $(x_i, o_i, y_i)$ be an observation triple for observation $i$, $i = 1, 2, \ldots, n$. The systematic component of a GLM is assumed to be a observation specific linear predictor, $\eta_i = o_i + \boldsymbol{x}_i^T \boldsymbol{\beta}$, where $\boldsymbol{x}_i^T$ is a known vector of covariates for the observed response including 1 for the intercept term, $\boldsymbol{\beta}$ is our vector of unknown regression coefficients for the different covariates, and $o_i$ is a known offset. The linear predictor is linked to the mean through a link function, $g(\mu) = \eta$. A special variant of the link function is the canonical link function, that is when $g(\mu) = \theta = \eta$. The link function is a strictly monotone and differentiable function, this ensures that it has an inverse, the response function $h(\eta) = \mu$, and can be differentiated (Dunn and Smyth 2018, Chap. 5.5).

When fitting the GLM we assume:

- All responses originate from the same process, with only the specified effects.
- The correct link function is chosen.
- All factors in the linear predictor follow the correct scale.
- The distribution chosen to model the problem is correct.
- The dispersion parameter is constant between responses.
- The responses are independent of each other.

It is important to note that these are never exactly true, but deviation from the model assumptions can have implications of the sensitivity of the conclusions. Therefore it is good practice to check the model assumptions to improve the model or check its validity (Dunn and Smyth 2018, Chap. 8).

### 3.1.1 Normal distribution

For the normal distribution with the response $Y$, we can write the probability distribution as

$$\begin{aligned}
\mathcal{P}(y; \mu, \sigma) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(y-\mu)^2}{2\sigma^2} \right\} \\
&= \frac{e^{-\frac{y^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \exp\left\{ \frac{y\mu - \mu^2/2}{\sigma^2} \right\},
\end{aligned}$$

where

$$\theta = \mu, \quad \kappa(\theta) = \frac{\theta^2}{2}, \quad \phi = \sigma^2, \quad a(y, \phi) = \frac{\exp\left\{ \frac{-y^2}{2\phi} \right\}}{\sqrt{2\pi\phi}}.$$

Inserting this into Equation (1) gives us the expected value $\mu$ and the variance $\sigma^2$. From this we can observe that a normal distributed GLM with canonical link becomes the multiple linear regression model. Parameters of the multiple linear regression model can be estimated by the method of maximum likelihood, or equivalently ordinary least squares. This is not an iterative method and is therefore preferred over the general solving procedure for a GLM, to be outlined in Section 3.2.1, as it is both faster and not dependent on a user specific convergence tolerance.

### 3.1.2 Negative binomial distribution

For the negative binomial distribution, which can also be considered a gamma-Poisson mixture with response $Y$, we can write the probability distribution as

$$\begin{aligned}\mathcal{P}(y;\mu,\alpha) &= \frac{\Gamma(y+1/\alpha)}{y!\Gamma(1/\alpha)}\left(\frac{1/\alpha}{\mu+1/\alpha}\right)^{1/\alpha}\left(\frac{\mu}{\mu+1/\alpha}\right)^{y} \\ &= \frac{\Gamma(y+1/\alpha)\alpha^{y}}{y!\Gamma(1/\alpha)}\exp\left\{y\ln\left(\frac{\mu}{1+\alpha\mu}\right)+\frac{1}{\alpha}\ln\left(\frac{1}{1+\alpha\mu}\right)\right\},\end{aligned}$$

where

$$\theta = \ln\left(\frac{\mu}{1+\alpha\mu}\right), \quad \kappa(\theta) = -\frac{1}{\alpha}\ln\left(1-\alpha e^{\theta}\right), \quad \phi = 1, \quad a(y,\phi) = \frac{\Gamma(y+1/\alpha)\alpha^{y}}{y!\Gamma(1/\alpha)}.$$

Note that we assume $\alpha$ is a known constant dispersion parameter, this can be considered as a constant coefficient of variation from the gamma mixing distribution. This also keeps the negative binomial as a single parameter exponential dispersion model. The canonical link becomes $g(\mu) = \ln\left(\frac{\mu}{1+\alpha\mu}\right)$ with our parameterization. This parameterization was motivated by De Jong, Heller et al. (2008, page 38-39). Inserting this into Equation (1) gives us,

$$E[Y] = \mu, \quad \text{Var}[Y] = \mu + \alpha\mu^2. \tag{2}$$

## 3.2 Parameter estimation

Let $(x_i, o_i, y_i)$ be an observation triple, and assume we have observed $i = 1, 2, \ldots, n$ independent observation pairs. We will first present maximum likelihood estimation to find relevant effects for different data. Then least squares estimation will be presented, as a special case of maximum likelihood when the GLM follows a normal distribution with canonical link. Then ridge regression will be introduced to reduce the estimated bias of our least square estimates.

### 3.2.1 Maximum likelihood estimation

To estimate the regression parameters we want to maximize the log-likelihood for our EDM,

$$l(\beta_0,\ldots,\beta_{p-1},\phi;y) = \sum_{i=1}^{n}\log\mathcal{P}(y_i;\mu_i,\phi).$$

When using canonical link functions the log-likelihood of the GLM is always concave, which makes the maximum likelihood (ML) estimate unique. In most cases estimation of the regression parameters must be done numerically through an iterative scheme until convergence. In the next section we will examine the normal EDM with canonical link function where the estimators for the regression parameter is found on the closed form. One way to do this is using the Fisher scoring algorithm, which requires computation of the score function and the expected Fisher information. When finding estimates for the score function and the expected Fisher information we will first consider a single element, and then generalize to matrix notation. Element $j$ of the score function can be written out as

$$\begin{aligned}U(\beta_j) &= \frac{\partial l(\beta_0,\ldots,\beta_{p-1},\phi;y)}{\partial\beta_j} = \frac{\partial\prod\log\mathcal{P}(y_i;\mu_i,\phi)}{\partial\theta}\frac{d\theta}{d\mu_i}\frac{d\mu_i}{d\eta_i}\frac{\partial\eta_i}{\partial\beta_j} \\ &= \sum_{i=1}^{n}\frac{(y_i-\mu_i)}{\phi}\frac{1}{V(\mu_i)}\frac{d\mu_i}{d\eta_i}x_{ji} = \frac{1}{\phi}\sum_{i=1}^{n}W_i\frac{d\eta_i}{d\mu_i}(y_i-\mu_i)x_{ji},\end{aligned}$$

where $V(\mu) = \frac{d\mu}{d\theta}$ is called the variance function and the different $W_i = \frac{1}{V(\mu_i)(d\eta_i/d\mu_i)^2}$ are called the working weights. Element $(j,k)$ of the expected Fisher information becomes

$$\begin{aligned}\mathcal{I}(\boldsymbol{\beta})_{jk} &= E\left[-\frac{\partial U(\beta_j)}{\partial \beta_k}\right] = E\left[-\sum_{i=1}^{n}\left(\frac{1}{V(\mu_i)}\frac{d\mu_i}{d\eta_i}x_{ji}\frac{\partial}{\partial\beta_k}\frac{(y_i-\mu_i)}{\phi} + \frac{(y_i-\mu_i)}{\phi}\frac{\partial}{\partial\beta_k}\frac{1}{V(\mu_i)}\frac{d\mu_i}{d\eta_i}x_{ji}\right)\right] \\ &= \frac{1}{\phi}E\left[\sum_{i=1}^{n}\frac{1}{V(\mu_i)}\frac{d\mu_i}{d\eta_i}x_{ji}\frac{\partial\mu_i}{\partial\eta_i}\frac{\partial\eta_i}{\partial\beta_k}\right] = \frac{1}{\phi}\sum_{i=1}^{n}W_i x_{ji}x_{ki}.\end{aligned}$$

This can be generalized to matrix notation,

$$\mathbf{U} = \frac{1}{\phi}X^T W M(\mathbf{y}-\boldsymbol{\mu}), \quad \mathcal{I} = \frac{1}{\phi}X^T W X, \tag{3}$$

where $W$ is the diagonal matrix of the working weights, $W_i$, and $M$ is the diagonal matrix of the derivatives of the link function, $d\eta_i/d\mu_i$. Then the Fisher scoring algorithm takes the form

$$\hat{\boldsymbol{\beta}}^{(r+1)} = \hat{\boldsymbol{\beta}}^{(r)} + \mathcal{I}\left(\hat{\boldsymbol{\beta}}^{(r)}\right)^{-1} U\left(\hat{\boldsymbol{\beta}}^{(r)}\right), \tag{4}$$

where the superscript $(r)$ denotes the estimate for the $r$-th iteration of the scheme. Note that $\phi$ cancels out for the estimation of $\hat{\boldsymbol{\beta}}$. After convergence an estimate of the covariance matrix for our parameters is the inverse of the expected Fisher information, $\text{Var}[\hat{\boldsymbol{\beta}}] = \mathcal{I}^{-1}$, where we can estimate the overdispersion, $\phi$, as $\hat{\phi} = \frac{1}{n-p}\sum_{i=1}^{n}\frac{(y_i-\hat{\mu}_i)^2}{V(\hat{\mu}_i)}$, where $n$ is the total number of observations and $p$ is the number of parameters estimated (Dunn and Smyth 2018, Chap. 6.8).

This is how we will estimate parameters for the negative binomial. For the normal distribution with a canonical link function the maximum likelihood estimator is the same as the least squares estimator and can be written on closed form.

### 3.2.2 Bayesian statistics

The purpose of Bayesian statistics is to combine the information we have about the data to be collected before doing analysis, the prior information, with the information we can observe from the data itself, the sample information. The prior information is usually defined in terms of a probability distribution on the parameter we want to estimate, and a hyperparameter the parameter depends on, this collection of parameters will be denoted as $\theta$. This probability distribution is defined by what the user expects of all parameters of $\theta$, and is denoted as $\pi(\theta)$. The sample information is the likelihood of the distribution that the data is assumed to follow, given that our parameters follows the assumed prior, which is written as $f(y|\theta)$. Assuming the sample information follows a continuous distribution and combining the prior and sample information gives us the posterior distribution,

$$\pi(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{\int f(y|\theta)\pi(\theta)d\theta},$$

which can be considered the conditional distribution of $\theta$ given the data available, $y$. Alternatively the integral becomes a sum if the sample information is expected to follow a discrete distribution. This approach is particularly powerful when doing inference for different sets of data with similar structure. Then one can construct a prior that is based on all samples, while the frequentist approach would need to use a pooled estimate, do separate estimates for each set or create a larger pipeline for estimation (Berger 2013, Chap. 1, 4).

Empirical Bayes is a branch of Bayesian statistics that uses the data available to estimate hyperparameters for our prior, instead of specifying a prior before observing the data.

### 3.2.3 Ordinary least squares and weighted least squares estimation

For the special case where the GLM follows a normal distribution with canonical link function the parameter estimates have a closed form solution. The method for solving this case is called ordinary least squares. Since this procedure is not iterative it is preferable over an iterative scheme.

If we assume that $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N(\mathbf{0}, I\sigma^2)$. The parameter estimators then becomes $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{Y}$ and $s^2 = \frac{(\mathbf{Y} - X\hat{\boldsymbol{\beta}})^T (\mathbf{Y} - X\hat{\boldsymbol{\beta}})}{n-p}$. Here $s^2$ is preferred over the ML estimate, where $s^2$ is the restricted maximum likelihood estimate for $\sigma^2$. In general the restricted maximum likelihood estimator is less biased than the ML estimator. Restricted maximum likelihood is not central in this presentation and will not be covered, more information can be found in Fahrmeir et al. (2007, Chap. 3.2). The ordinary least squares estimates is found under the assumption that the errors are uncorrelated and homoscedastic (Fahrmeir et al. 2007, Chap. 3.2).

Let us now examine an extended model where we assume a variance of the form $\text{Var}[\boldsymbol{\epsilon}] = W$. The main motivation for this change of variance is to model off diagonal variance elements, in other words covariance between responses. When $W$ is known we can transform our variables such that the errors are uncorrelated and homoscedastic. This can be done by *weighted least squares*. The drawback of weighted least squares is the the covariance matrix must be known. If we assume the covariance matrix to be $W$, then weighted least squares does ordinary least squares on the linear system,

$$W^{-1/2}\boldsymbol{Y} = W^{-1/2} X\boldsymbol{\beta} + W^{-1/2}\boldsymbol{\epsilon}.$$

Here $W^{-1/2}$ is not uniquely determined as the matrix only needs to satisfy $W^{-1/2}(W^{-1/2})^T = W^{-1}$. One method for finding $W^{-1/2}$ is the spectral decomposition.

The effects of the resulting transformation will be independent and have equal variance of 1, when $W$ is known. When we instead assume that the covariance matrix has a particular form we can estimate the variance not accounted for in our transformed data, $\text{Var}[W^{-1/2}\boldsymbol{\epsilon}] = s^2 I$, and compare $s^2$ to 1 to get a measure of fit. In other words, we can use weighted least squares with an assumed underlying covariance matrix to find an estimate for gene specific dispersion, which can be used to assert our overall fit (Fahrmeir et al. 2007, Chap. 4.1). We observe that the parameter estimates becomes

$$\hat{\boldsymbol{\beta}} = (X^T W^{-1} X)^{-1} X^T W^{-1} \mathbf{Y}, \quad s^2 = \frac{(\mathbf{Y} - X\hat{\boldsymbol{\beta}})^T W^{-1} (\mathbf{Y} - X\hat{\boldsymbol{\beta}})}{n-p}. \tag{5}$$

Observe that for ordinary least squares $s^2$ is an estimate of the variance after a model has been fit to the data. For weighted least squares $s^2$ is an estimate of the variance not captured in the weights, but after being transformed to consider an underlying covariance structure. Therefore the values of $s^2$ will not be directly comparable, as for ordinary least square it tries to directly capture all the variance, and for weighted least squares it tries to capture the variance after the expected variance has been captured.

### 3.2.4 Ridge regression

Ridge regression was introduced to give least squares regression a solution when the covariates are correlated, in other words when the covariates are non-orthogonal. If this correlation is high enough $X^T X$ will be close to singular, and its smallest eigenvalue will be close to 0. This can be solved by adding a ridge penalty parameter, $\lambda$, to the diagonal elements of $X^T X$. This is now commonly used when dealing with high dimensional design matrices, when there are more parameters to estimate than data available. Ridge regression changes our estimators to be biased to circumvent the problem of non-orthogonality. This effectively makes the model more robust (Marquardt and Snee 1975).

In a frequentist approach the ridge penalty parameter, $\lambda$, is a variable usually estimated by using cross-validation. An alternative way to consider ridge regression is to uses a zero centered normal prior on the coefficient vector before doing least squares, where the prior variance corresponds $\lambda$. If we were to use an empirical Bayes approach then the ridge penalty parameter is estimated by the data instead of by cross-validation. It is then not necessary for $\lambda$ to be a diagonal matrix with equal elements, but instead have observation specific elements corresponding to desired variance for each effect (Love, Huber and Anders 2014).

If we assume that $Y \sim N(X\boldsymbol{\beta}, \Sigma)$ then the ridge regression estimator takes the form

$$\hat{\boldsymbol{\beta}}_{ridge} = (X^T X + \lambda)^{-1} X^T y,$$

where $\lambda$ is a diagonal matrix estimated by empirical Bayes. We can further examine the expected value and variance of the ridge coefficient vector,

$$E[\hat{\boldsymbol{\beta}}_{ridge}] = (X^T X + \lambda)^{-1} X^T X \boldsymbol{\beta},$$

$$\text{Var}[\hat{\boldsymbol{\beta}}_{ridge}] = (X^T X + \lambda)^{-1} X^T \Sigma X (X^T X + \lambda I)^{-1}.$$

From this we can observe the bias commented on and the reduction in the variance of our estimator.

## 3.3 Hypothesis testing

In this subsection we want to introduce linear hypothesis and their asymptotic distribution, this will later be used to evaluate individual genes to evaluate if our observations are significant enough to have a potential effect to influence Crohn's disease. Further we will introduce concepts that generalizes the gene specific hypothesis to limit how many irrelevant genes that will be considered relevant for further analysis.

### 3.3.1 Testing linear hypothesis

For inference we want to consider the linear hypothesis $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{d}$ against $H_1 : \mathbf{C}\boldsymbol{\beta} \neq \mathbf{d}$, where $\mathbf{C}$ is the contrast matrix relating the coefficients we want to test. For GLMs it is intuitive to use the Wald statistic

$$w = (\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d})^T [\mathbf{C}\mathcal{I}^{-1}(\hat{\boldsymbol{\beta}})\mathbf{C}^T]^{-1} (\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}) \overset{\text{asymptotic}}{\sim} \chi^2_r, \tag{6}$$

where $r$ is the number of rows in $C$, this allows us to calculate approximate $p$-values (Fahrmeir et al. 2007, Chap. 5.1).

A $p$-value is found by comparing the test statistic to its asymptotic distribution, then use the probability that the observation was larger or equal to the observed when $H_0$ is true, $P(w \geq \chi^2_r | H_0)$. If all $p$-values are from models that follow $H_0$ then the distribution of $p$-values is asymptotically expected to follow a uniform distribution. This is a consequence of the $p$-value originating from the cumulative distribution function of the asymptotic test statistic (Goeman and Solari 2014).

An alternative when using a normally distributed parameter estimator, $\hat{\boldsymbol{\beta}}$, is the $t$-test. This comes from a special case when using the $F$-test for the normal EDM,

$$\frac{(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d})^T [\mathbf{C}\mathcal{I}^{-1}(\hat{\boldsymbol{\beta}})\mathbf{C}^T]^{-1} (\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d})/r}{s^2/\sigma^2} \sim \frac{\chi^2_r/r}{\chi^2_{n-p}/n-p} \sim F_{r,n-p}.$$

When $r = 1$, then this becomes

$$\frac{\hat{\beta}_j^2}{\widehat{\text{Var}[\hat{\beta}_j]}} \sim F_{1,n-p}.$$

Which by the taking the root becomes a $t$-test,

$$t_j = \frac{\hat{\beta}_j}{se_j} \sim t_{n-p},$$

where $se_j = \sqrt{\widehat{\text{Var}[\hat{\beta}_j]}}$ (Fahrmeir et al. 2007, Chap 3.1).

### 3.3.2 Multiple tesing

We would like to test many hypotheses of the type $\mathbf{C}\boldsymbol{\beta} = \mathbf{d}$. When we do this we would like to have a small as possible value of the expected ratio of true hypotheses rejected compared to the total number of rejections, which is called the false discovery rate (FDR). In our situation we want

Table 3: Possible errors to be committed when testing $m$ null hypotheses.

| | Declared non-significant | Declared significant | Total |
|---|---|---|---|
| True null hypotheses | $U$ | $V$ | $m_0$ |
| Non-true null hypotheses | $T$ | $S$ | $m - m_0$ |
| | $m - R$ | $R$ | $m$ |

to test $m$ different null hypothesis at the same time, where $m_0$ of them are true, and $R$ is the numbers of rejected hypotheses. This situation is summarized in Table 3, where $R$ is an observable random variable, and $U, S, T, V$ are unobservable random variables. We want to keep the FDR, $E[Q]$, as small as possible, where $Q$ is defined as

$$Q = \begin{cases} V/(V+S) & \text{if } V + S \neq 0 \\ 0 & \text{if } V + S = 0 \end{cases},$$

and can be considered the proportion of rejected true null hypotheses. When $V + S = 0$ it is impossible to reject a true null hypothesis, so it can be interpreted as the ratio being 0 (Benjamini and Hochberg 1995).

The Benjamini-Hochberg step-down method is a popular method for controlling the FDR. If we sort the $p$-values in ascending order such that $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(m)}$ and denote $H_{(i)}$ as the null hypothesis corresponding to $P_{(i)}$. Then the Benjamini-Hochberg step-down method claims:

$$\text{Let } k \text{ be the largest } i \text{ for which } P_{(i)} \leq \frac{i}{m} q^*,$$

then rejecting all $H_{(i)}, \quad i = 1, 2, \ldots, k$ controls the FDR at level $q^*$ (Benjamini and Hochberg 1995, Section. 3). This is also shown in Goeman and Solari (2014), if we can assume positive dependence through stochastic ordering.

One way to implement this in practice is by finding the $p$-values for the different hypotheses, then sorting them in ascending order, such that $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(m)}$. Then set the adjusted $p$-value of the largest $p$-value to its $p$-value, $P^*_{(m)} = P_{(m)}$, and iterate the scheme $P^*_{(i)} = \min(P_{(i)} \frac{m}{i}, P^*_{(i+1)})$ in descending order (Goeman and Solari 2014).

This scheme stays true to the method described in Benjamini and Hochberg (1995). When we want to control the FDR at level $q^*$, we compare all the adjusted $p$-values to $q^*$.

## 3.4 Linear mixed model

Linear mixed models (LMM) extend the linear predictor to include random effects, $\boldsymbol{\gamma}$. The objective of the random effect is to capture effects that are similar within a predefined cluster. We want to introduce a new setup for our responses, which makes the interpretation of cluster specific parameters more intuitive. Instead of the responses being a vector with length $n$, we want each cluster to have a vector of responses that belong to that cluster. We need to define a variable to assign which cluster we are examining, $i = 1, \ldots, q$, and a variable that assigns which observation of a cluster we are examining $j = 1, \ldots, n_i$. The clusters in our case captures which observations were from the same patient, and are of variable sizes. This interpretation changes our cluster specific covariate, $X_i$, to be a matrix with $n_i$ rows. Here we only want to consider a random intercept, but LMMs can include random effects that are dependent on observed covariates as well. As each cluster is a random sample from a larger group, the cluster specific parameter can be considered random with $\boldsymbol{\gamma} \sim N(\mathbf{0}, I_q \tau^2)$.

An alternative way of looking at $\boldsymbol{\gamma}$ is as a cluster specific error shared between measurements of the same cluster. One benefit of this extension is that the model uses the correlation from repeated observations, which is something that a normal GLM cannot.

If we start with a multiple regression model, $\mathbf{Y}_i \sim N(X_i\boldsymbol{\beta}, \sigma^2)$, and add a random intercept to the linear predictor, then the linear predictor becomes $\boldsymbol{\eta}_i = X_i\boldsymbol{\beta} + 1_{n_i}\gamma_i$. This model marginally

becomes $\mathbf{Y}_i \sim N(X_i\boldsymbol{\beta}, I_{n_i}\sigma^2 + I_{n_i}\tau^2)$, but if we know $\gamma_i$ the conditional distribution is $\boldsymbol{Y}_i|\gamma_i \sim N(X_i\boldsymbol{\beta} + 1_{n_i}\gamma_i, I_{n_i}\sigma^2)$.

This can be used to find the correlation between two observations of an individual by

$$\begin{aligned}
\text{Corr}[Y_{ij}, Y_{ij'}] &= \frac{\text{Cov}[Y_{ij}, Y_{ij'}]}{\text{Var}[Y_{ij}]^{1/2}\,\text{Var}[Y_{ij'}]^{1/2}} \\
&= \frac{\text{Cov}[\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \gamma_i + \epsilon_{ij}, \boldsymbol{x}_{ij'}^T\boldsymbol{\beta} + \gamma_i + \epsilon_{ij'}]}{\text{Var}[Y_{ij}]^{1/2}\,\text{Var}[Y_{ij'}]^{1/2}} \\
&= \frac{\text{Cov}[\gamma_i, \gamma_i]}{\text{Var}[Y_{ij}]^{1/2}\,\text{Var}[Y_{ij'}]^{1/2}} \\
&= \frac{\tau_0^2}{\sigma^2 + \tau_0^2},
\end{aligned}$$

where $j$ and $j'$ are from the same patient, but is not the same observation. One unfortunate consequence of this method is that the found correlation will always be positive in a LMM.

If we choose to ignore this correlation within our clusters our standard errors will be invalid. One example of this is if we examine two observations from two different different clusters, $a_1, a_2$ from one cluster and $b_1, b_2$ from another cluster. Assume that the correlation is the same within both clusters, $\text{Corr}[a_1, a_2] = \text{Corr}[b_1, b_2] = \rho_{ex}$, the observations in different clusters are uncorrelated, $\text{Corr}[a_i, b_j] = 0$ for $i, j = 1, 2$, and the variance for the two different clusters are equal, $\text{Var}[a_j] = \text{Var}[b_j] = \sigma_{ex}^2$. We can then observe that the variance of the effect when comparing observations from different clusters becomes

$$\begin{aligned}
\text{Var}\left[\frac{(a_1 + b_1) - (a_2 + b_2)}{2}\right] &= \frac{1}{4}\left(\text{Var}[a_1 + b_1] + \text{Var}[a_2 + b_2] - 2\,\text{Cov}[a_1 + b_1, a_2 + b_2]\right) \\
&= \frac{1}{4}\left(\text{Var}[a_1] + \text{Var}[a_2] + \text{Var}[b_1] + \text{Var}[b_2] - 2\,\text{Cov}[a_1, a_2] - 2\,\text{Cov}[b_1, b_2]\right) \\
&= \sigma_{ex}^2(1 - \rho_{ex}).
\end{aligned}$$

As the correlation is typically positive the variance for inference will be estimated as too small when estimating the effect across clusters (Agresti 2003, Chap. 9.1).

In our situation we will examine a mix of clustered and non clustered data. We believe that the correlated data will influence the final variance estimate, but less than in the above example. 28% of the data presented in Section 2.5 are from correlated clusters with two observations, whilst the rest are not from clusters, this can be observed from Table 2.

## 3.5 Model assessment

To analyze the response residuals we first need to make some distributional assumptions.

Assuming that our data comes from a normally distributed linear model, $\mathbf{y} \sim N(\boldsymbol{\mu}, \sigma^2 W)$. Then the responses are assumed to be normally distributed. The linear model can behave in an additive fashion as $y = \mu + \epsilon$, where $\epsilon$ is the error term and $\text{Var}[\epsilon] = \sigma^2$ (Section 3.2.3). As a consequence the error terms will be assumed to be normally distributed, and the residuals can be considered predictions of these error terms. We want to define these residuals, and start with the raw residuals which are defined as $\mathbf{r}_r = \mathbf{y} - \hat{\boldsymbol{\mu}}$. Here $\hat{\mu}_i$ is estimated from data, and is considered a random variable. If we examine the specific case of using weighted least squares estimation the variance of the raw residuals can then be written out to be,

$$\text{Var}[\mathbf{r}_r] = \sigma^2 W^{1/2}(I_n - H)W^{1/2},$$

where $H$ is called the hat matrix and defined as $H = W^{-1/2}X(X^TW^{-1}X)^{-1}X^TW^{-1}$. It is quite difficult to evaluate residuals with different variances, so we modify them to have constant variance by

$$\mathbf{r}^* = (I - H)^{-1/2}W^{-1/2}\mathbf{r}.$$

We can use this to define the standardized residuals as

$$\mathbf{r}' = \mathbf{r}^*/s,$$

where $s^2$ is our estimate for $\sigma^2$. This will be approximately standard normal, and will exactly follow a $t$-distribution with $n - p'$ degrees of freedom, where $p'$ is the number of parameters estimated (Dunn and Smyth 2018, Chap. 3).

One way to check the normality of the residuals is a Q-Q plot. Here the quantiles of the data is plotted against the quantiles of a standard normal distribution. Therefore the points will lay on a straight line in the Q-Q plot if the residuals have a normal distribution (Dunn and Smyth 2018, Chap 3).

For a GLM it is not as simple to directly analyse their residuals, so instead we need to use a transformation. In the case of DESeq2 we have a discrete distribution (negative binomial), which complicates this analysis even more. Our main goal now is to define a residual which is a standard normal variable with the same cumulative probability as $y$ has to being observed for the GLM (Dunn and Smyth 2018, Chap. 8). We now want to turn to quantile residuals for residual analysis.

### 3.5.1 Quantile residuals

Quantile residuals use a distribution based on the expected value for each response. We want to illustrate and explain this by an example, where we examine the negative binomial distribution. The procedure we will outline can be used for both discrete and continuous distributions, but for discrete it needs to use a randomization step to convert an interval into a value which is not necessary for a continuous distribution.

We first want to consider $\mathbf{y} \sim NB(\boldsymbol{\mu}, \alpha)$. Then we want to find the interval of cumulative probabilities this observation corresponds to, $(F_{NB}(y_i; \hat{\mu}_i, \alpha), F_{NB}(y_i + 1; \hat{\mu}_i, \alpha))$, here $F_{NB}$ is the CDF for the negative binomial distribution. We need to consider an interval of cumulative probabilities as we are taking into account a discrete distribution, where the CDF has discrete jumps between different responses. This is solved by a randomization step, where a draw from the uniform distribution based on the interval of cumulative probabilities decides the specific cumulative probability assigned to this residual, $u_i \sim U(F_{NB}(y_i; \hat{\mu}_i, \alpha), F_{NB}(y_i + 1; \hat{\mu}_i, \alpha))$. With this randomization step we find a random realization based on an interval of cumulative probabilities. We can further use this cumulative probability, $u_i$, by finding a value from a standard normal distribution with the same cumulative probability, $r_{Q\,i} = F_N^{-1}(u_i; \mu = 0, \sigma^2 = 1)$ and using this as our residual. These residuals are called quantile residuals and will have follow an exact standard normal distribution (Dunn and Smyth 2018, Chap. 8).

# 4 Statistical software for analysing count data

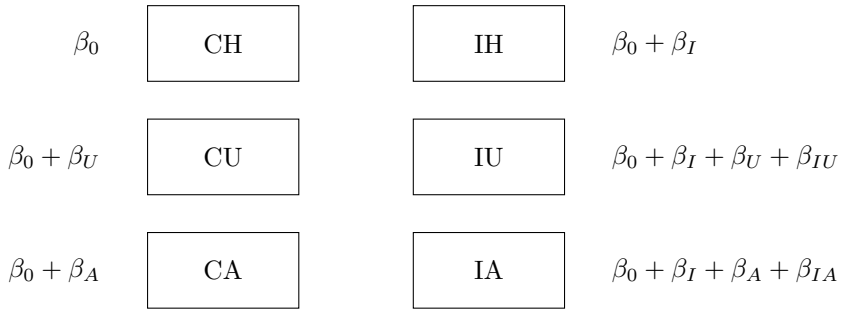



Figure 4: Overview of the different groups relevant for Crohn's disease and coefficients for the groups.

The data presented in Section 2.5 contains factors assigning each observation to one of two different types of tissue and three different states of inflammation, as well as a patient id and the gene

expression count values. We choose tissue to be a factor with two levels, $C$ for colon and $I$ for ileum, where colon is the reference level. Inflammation has three levels, $H$ for healthy, $U$ for unaffected and $A$ for affected. For inflammation we use healthy as our reference level. We can observe that the vector of coefficients then becomes

$$\boldsymbol{\beta} = [\beta_0 \quad \beta_U \quad \beta_A \quad \beta_I \quad \beta_{IU} \quad \beta_{IA}]^T.$$

Figure 4 shows the different groups of observations for Crohn's disease and shows which coefficients corresponds to different groups.

To analyze the data presented in Section 2.5 we want to introduce some well established statistical software. The data we have available consist of factors that tells us where the samples were extracted from and what state of inflammation the extracted sample had. These factors get turned into vectors of covariates which identify which observational group the observation belongs to, $\mathbf{x}_i$. With this design $\boldsymbol{\beta}$ can be estimated to model effects specific to the different observational groups of tissue with disease for each gene $g$. The purpose for the statistical software is to fit models to our data such that test can be performed to find genes which can explain the differences between the groups. The raw gene expression counts are denoted $r_{gi}$ and is for gene $g$ and observation $i$.

This vector of covariates $\mathbf{x}_i$ can be interpreted as an observation specific identifier for the different tissues. We can here use Figure 4 as reference to which effects corresponds to which group. One example for this is if a sample is extracted from Ileum and is unaffected (IU) then our vector of covariates would become $\mathbf{x}_i^T = [1\,1\,0\,1\,1\,0]$, then we can observe that $\mathbf{x}_i^T\boldsymbol{\beta} = \beta_0 + \beta_I + \beta_U + \beta_{IU}$. Our design matrix is defined as $X = [x_1\, x_2\, \ldots\, x_n]^T$

Both of these pipelines will result in estimates of log fold change (LFC) and the variance associated with this LFC. For both models this directly corresponds to the estimated effect of interest. The LFC has the interpretation that an increase of the covariate corresponding to the effect of interest will increase the expected gene count by a factor of $2^{LFC}$. In our case the covariates for relevant effects are the levels of inflammation and tissue. LFC has been used since since early micro array literature and is implemented to highlight results, but we will interpret this as the estimated effect of the relevant contrast.

## 4.1 Linear models for micro array data

Linear models for microarray analysis (Limma) is a Bioconductor R-package that has implemented a pipeline for analyzing RNA-sequencing (RNA-seq) and microarray data. One of its core features is that it borrows information for estimating variances across genes to deal with small sample sizes. When analyzing RNA-seq data limma transforms the gene counts to a log scale and assume that these are approximately normal (Ritchie et al. 2015).

Gene counts in general have large variances when the mean of counts for a gene is high, compared to the variances when the mean of counts is lower. The log counts instead give more similar variances than for the raw counts, but there is still an observable relationship between the variation between results and the mean log count of a gene. This can be modeled as a dependency and will be further explained in Section 4.1.3. This concept can be observed for the data described in Section 2.5 in Figure 19.

After the data has been log transformed, limma fits a linear model for each gene based on a design matrix to find estimates for coefficients and the estimated variance for each gene. Limma has three different procedures for modeling the mean-variance relationship for genes. These are a general pooled variance, a mean-variance trend called limma-trend, or a weight specific procedure called limma-voom, which is an acronym for 'variance modeling at the observational level'. The estimation of a general pooled variance is inferior as it generalizes the variance to one value instead of creating a relationship. Both limma-trend and limma-voom models a trend between the variance and the mean of the log counts, and can model correlation between observations. Both these methods have similar performance when the total counts from each patient is similar. When they are not limma-voom shows better performance than limma-trend, so we will only consider limma-voom (Law, Chen et al. 2014).

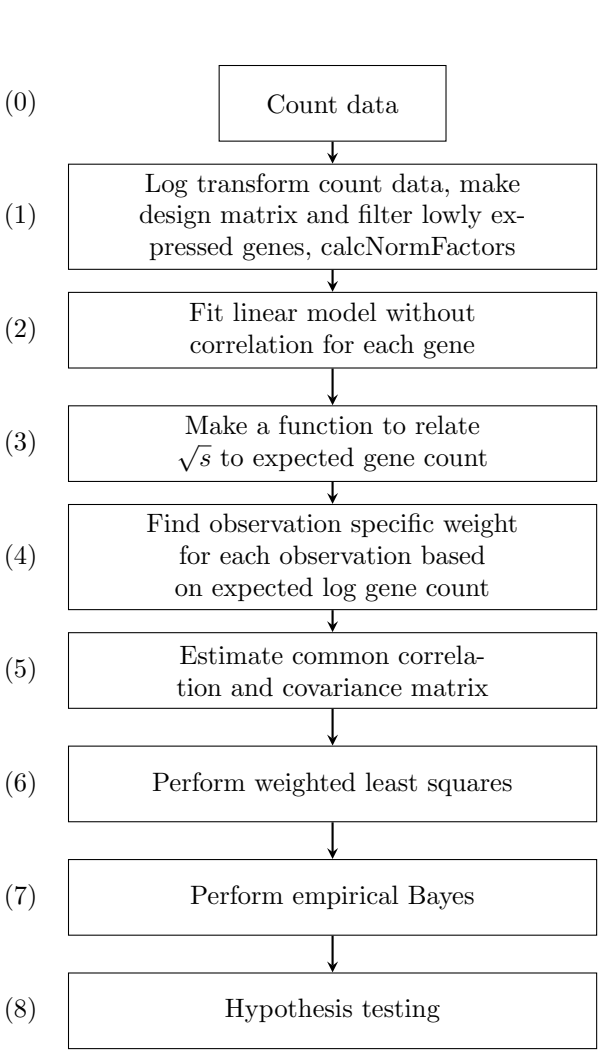



Figure 5: An overview of the limma-voom pipeline for finding differentially expressed genes for RNA-sequencing count data. This figure was inspired by explanation of the limma-voom pipeline from Law, Chen et al. (2014).

Limma-voom starts by estimating weights by fitting models for all genes and relating the estimated gene-wise variance to the gene-wise mean through weights. When the user specifies that correlation is present another method similar to LMM is used to estimate the gene specific correlation, before generalizing to a consensus correlation. Together the estimated weights and the consensus correlation is used to estimate a gene specific correlation matrix. Then this covariance matrix is used in weighted least squares to get an estimate of the effects and variance not accounted for. So limma-voom estimates the mean-variance relationship *before* we take into account the correlation. Empirical Bayes is then used to increase the power of our test before Wald and $t$-tests are used to do hypothesis testing for each gene (Law, Chen et al. 2014). All of these steps will be examined in further detail and are summarized in Figure 5.

In Block 5 limma-voom uses the found observation specific variances from an already fitted mean-variance trend to estimate the gene specific covariance matrix, and uses the variances that are not accounted for from the weighted least squares to measure the goodness of fit (Law, Chen et al. 2014).

### 4.1.1 Preprocessing

This section corresponds to Block 1 in Figure 5. We want to go a little further in depth on how limma does its preprocessing. We start by assuming we have $n$ different RNA samples, where multiple samples can be from the same patient or all from different patients, where we have recorded gene expression count data for $G$ different genes. This can be written as a matrix of gene counts, $r_{gi}$, for sample $i \in \{1, 2, \dots, n\}$ and gene $g \in \{1, 2, \dots, G\}$. We can then denote $R_i = \sum_{g=1}^{G} r_{gi}$ as the library size (total number of reads) for sample $i$, and use this to define log-CPM (counts per million) for each sample and gene,

$$y_{gi} = \log_2 \left( \frac{r_{gi} + 0.5}{R_i + 1.0} \cdot 10^6 \right). \tag{7}$$

The seemingly unmotivated constants ensure that $0 < \frac{r_{gi}+0.5}{R_i+1.0} < 1$, which avoids taking the log of zero and dividing by zero. The term in $r_{gi} + 0.5$ also has the pleasant effect of reducing the variability for lowly expressed genes (Law, Chen et al. 2014).

For filtering our intention is to remove genes that are believed to not have the potential to be relevant. For us the most intuitive solution is to automatically remove the irrelevant genes would be to remove genes that have little to no information available. That would imply that most of the samples have no information about the specific gene, which means low to no counts. A particular example would this could go wrong is for paneth cells which can be found in ileum but not colon. More than 50% of our data originates from colon, and if the paneth cells are of low concentration for all colon samples our filtering must require few counts for few observation to be able to not filter the paneth cells. So we remove genes that do not have a minimum log-CPM in a worthwhile number of samples, which is dependent upon the number of parameters we want to estimate. The minimum log-CPM required for the genes are dependent on the median library size, and is decided by the `filterByExpr` function from the edgeR package (Law, Alhamdoosh et al. 2016).

One problem when working with gene counts is that the library sizes can vary, one way to solve this is normalization. Another problem is that the number of transcripts in a sample is not necessarily the same. We want to highlight this by an example.

First consider two samples, A and B. The same number of transcripts that are present in A are also present in B. But B also has some the number of transcripts in A present for transcripts that A does not have. In this case B has twice as many transcripts as A has and after amplification then normalization with respect to library size would adjust sample A by a factor of 2. The intention of this example was to highlight that proportion of reads for a gene should be dependent on the whole sample and not just the sample size.

One way to account for different number of transcripts in a sample is by method of trimmed mean of M-values (TTM), this is implemented in edgeR through the function `calcNormFactors`. We

present the relevant equations for estimation of normalization factors,

$$\log_2(TTM_k^{(l)}) = \frac{\sum_{g\in G^*} w_{gk}^l M_{gk}^l}{\sum_{g\in G^*} w_{gk}^l}, \text{ where } M_{gk}^l = \frac{\log_2\left(\frac{r_{gl}}{R_l}\right)}{\log_2\left(\frac{r_{gk}}{R_k}\right)}, \text{ and } w_{gk}^l = \frac{R_k - r_{gk}}{R_k r_{gk}} + \frac{R_l - r_{gl}}{R_l r_{gl}},$$

and then explain how to interpret them. Just as previously $r_{gj}$ represents the observed gene expression count for gene $g$ and patient $j$, and $R_j$ corresponds to the library size for sample $j$. The interpretation of $M_{gk}^l$ is the log fold change of sample $k$ relative to sample $l$ for gene $g$. The weights, $w_{gk}^r$, can be interpreted as the inverse of the appropriate asymptotic variances, these can be found using the delta method. The weights, $w_{gk}^r$, are also for sample $k$ relative to sample $l$ for gene $g$. The sum considers genes in the set $G^*$, this set is a trimmed set of all genes, where 30% is trimmed based on the values of $M_{gk}^r$. This calculation requires $r_{gj} > 0$, and if that is not the case that sample is not considered for this gene (Robinson and Oshlack 2010). The normalization factor can be considered as a weighted trimmed mean based on the LFC between two samples. After normalization the majority of the count data will be more similar for each gene. We observe that this procedure only requires one sample as reference for all other samples.

### 4.1.2 Linear model

This section corresponds to Block 2 in Figure 5. We initially assume that the log-CPM approximately follows a normally distributed GLM, where all observations are independent, with a gene specific variance. For one observation this can be written as $y_{gi} \sim N(\boldsymbol{x}_i^T\boldsymbol{\beta}, \sigma_g^2)$. So after preprocessing limma-voom uses least squares estimation to get an estimate of the standard deviation for the different genes. Note that the assumption of independence is only used to give an estimate of gene specific variances, later we will use the dependence of observations from the same patient to estimate the covariance matrix.

### 4.1.3 Mean-variance relationship

In this section we will explain how Block 2−Block 4 for limma-voom in Figure 5 is done and comment on how this influences Block 6 in Figure 5. As previously mentioned probability distributions for count data generally have larger variances for larger counts. For the case of RNA-seq data it has been argued that the mean-variance relationship should be approximately quadratic (McCarthy, Chen and Smyth 2012). It is desirable for the coefficient of variation to decrease with the expected count and converge to a measure of biological variation as the expected count goes to infinity. In other words it is desired that the squared coefficient of variation should roughly follow $\frac{1}{\omega_{gi}} + \phi_g$, where $\omega_{gi} = E[r_{gi}]$.

If we assume the variance can be written on the form $\text{Var}[r_{gi}] = \omega_{gi} + \phi_g\omega_{gi}^2$, where $\phi_g$ is a dispersion parameter, we then get the log-counts per million $y_{gi} \approx \log_2(r_{gi}) - \log_2(R_i) + 6\log_2(10)$. Observe that $\text{Var}[r]$ has the same form as the variance for the negative binomial, see Section 3.1.2. As our analysis is conditional on $R$ it will be considered a constant in our analysis. We can then use $\text{Var}[Y_{gi}] \approx \text{Var}[\log_2(r_{gi})]$, and then writing the Taylor series around $\omega_{gi}$ we get $\log_2(r_{gi}) \approx \omega_{gi} + \frac{r_{gi}-\omega_{gi}}{\omega_{gi}} + o(r_{gi} - \omega_{gi})$. This can then be used to get an estimate of the variance, $\text{Var}[Y_{gi}] \approx \frac{\text{Var}[r_{gi}]}{\omega_{gi}^2} = \frac{1}{\omega_{gi}} + \phi_g$.

Limma-voom uses the fitted model to estimate the expected value, $\hat{\mu}_{gi} = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$, for each observation for each gene and residual standard error for each gene, $s_g^{(OLS)}$. We then need the average log-count value, which we find through

$$\tilde{r}_{gi} = \tilde{y}_g + \log_2(\tilde{R}_i) - \log_2(10^6),$$

where $\tilde{y}_g$ is the gene specific average of the log-CPM found in equation (7) and $\tilde{R}$ is the geometric mean of the library sizes plus one. It has been found that $\left(s_g^{(OLS)}\right)^{1/2}$ is roughly symmetrically

distributed, and is therefore used to estimate the mean-variance relationship. A locally weighted scatter plot smoothing (LOWESS) function is then fit based on $\left(s_g^{(OLS)}\right)^{1/2}$ as a function of $\tilde{r}$, we will call this function $lo()$. We can then use this function to find the precision weights to be used in limma-voom by $w_{gi} = lo(\tilde{r}_{gi})^{-4}$. Then the estimated weights and log-CPM values are sent into the empirical Bayes pipeline of limma (Law, Chen et al. 2014). The LOWESS curve for the dataset presented in Section 2.5 can be seen in Figure 19.

When the weights are estimated we want to estimate the effects by weighted least squares. The purpose of weights is to give an estimate for the covariance matrix, $\widehat{\text{Var}}[Y_g] = X^T(X^T\hat{V}_gX)^{-1}Xs_g^2$, where $\hat{V}_g$ is a diagonal matrix with precision weights for a specific gene, and $s_g^2$ is estimated by weighted least squares. Preferably we want $s_g^2$ to be estimated to 1, as then the precision weights have accounted for all the variance for that gene. All the gene specific variances, $s_g^2$, will be used to give a pooled estimate for a general prior for $s_g^2$, this will be covered further in 4.1.4.

### 4.1.4 Empirical Bayes analysis

This section corresponds to Block 7 in Figure 5. Limma-voom uses empirical Bayes to borrow information between genes to increase the power of our model, this will improve the estimates more the less observations we have. This subsection will be a short summary of the Bayesian setup as it is not our main focus, further elaborations and comments can be found in Smyth (2004).

If we assume that the coefficients found from the regression for each gene, $\hat{\boldsymbol{\beta}}$, is $\hat{\beta}_j|\beta_j, \sigma_g^2 \sim N(\beta_j, z_{gj}\sigma_g^2)$, where $Z_g\sigma_g^2$ is the covariance matrix for $\hat{\boldsymbol{\beta}}$, and $z_{gj}$ is the $j$-th diagonal element of $Z_g$. Observe that $\left(X^T\hat{V}_gX\right)^{-1}s_g^2$ is our estimate of $Z_g\sigma_g^2$, where $Z_g$ is not required to follow our assumptions for all genes. We also assume that $s_g^2|\sigma_g^2 \sim \frac{\sigma_g^2}{d_g}\chi^2_{d_g}$, where $d_g$ is the number of degrees of freedom for the residuals from the linear model for gene $g$.

Our main interest is to test hypotheses of the form

$$\begin{aligned} H_0 &: \mathbf{C}\boldsymbol{\beta} = \mathbf{0}, \\ H_1 &: \mathbf{C}\boldsymbol{\beta} \neq \mathbf{0}, \end{aligned} \tag{8}$$

and rank them in order of strength, as usually a number of the lowest $p$-values are followed up in further study. Our primary aim is therefore to rank the genes in order of evidence against $H_0$ compared to assigning adjusted $p$-values.

Limma-voom makes use of the parallel structure of gene counts by fitting the same hierarchical model to all genes. We assume prior information on $\sigma_g^2$ equivalent to a prior estimator $s_0^2$ with $d_0$ degrees of freedom,

$$\frac{1}{\sigma_g^2} \sim \frac{1}{d_0 s_0^2}\chi^2_{d_0}.$$

For limma-voom $s_0^2$ is a pooled variance estimate based on all gene specific estimates of variance not captured by the estimated covariance, $s_g^2$. With this hierarchical structure the posterior mean of $\sigma_g^2$ given $s_g^2$ can be written as

$$\tilde{s}_g^2 = \frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g}.$$

We can observe that the observed variances are shrunk towards a global mean dependent on the observed and prior degrees of freedom. Further detail in how this is fitted to specific observations can be found in Smyth (2004).

### 4.1.5 Test statistic

This section corresponds to Block 8 in Figure 5. With equation (8) the moderated $t$-statistic becomes

$$t_{gj} = \frac{\hat{\beta}_j}{\tilde{s}_g \sqrt{z_{gj}}},$$

and follows approximately a $t$-distributions with $d_g$ degrees of freedom. Here $z_{gj}$ is the estimated standard deviation for $\hat{\beta}_j$, found from the $j$-th diagonal element of $Z_g$.

Alternatively for testing a contrast, $C$, for more than one effect the Wald test statistic is used instead,

$$w = \frac{1}{\tilde{s}_g^2}(\mathbf{C}\hat{\boldsymbol{\beta}})^T[\mathbf{C}X^T\hat{V}_g X\mathbf{C}^T]^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}}),$$

which is assumed to be approximately chi-squared with $r$ degrees of freedom, which is the number of rows in $C$. For our analysis this will always be 1.

For controlling the FDR limma uses the Benjamini-Hochberg step-down method (BH) described in Section 3.3.2.

### 4.1.6 Correlated data

This section corresponds to Block 5 in Figure 5. In our data, to be analyzed in Section 6, some of our samples come from the same person, for this limma considers a correlation between samples from the same patient. We need to make the simplifying assumption that the correlation is common across genes, and common between duplicate observations, accounting for the statistical model. In practice this assumption makes our model more robust and as long as the correlation is sufficiently stable it will improve the overall model (Smyth, Michaud and Scott 2005).

The correlation matrix, $V$, is assumed to have $\rho$ in cells that indicate the same patient, 1s on the diagonal, and otherwise 0's. In the case of limma-voom this a little more complex, as each observation does not necessarily have the same variance. So the variance matrix will instead be $W_g = Q_g V Q_g^T$, where $Q_g$ is a $(n \times 1)$ vector consisting of the square root of the weights found in Section 4.1.3. Finding the general $\rho$ is a more complex process. First limma finds a $\rho_g$ for each gene by fitting a gamma generalized linear mixed model and extracting the correlation from the proportion of variance assigned to the random effect (Venables and Ripley 2002). This method can measure a negative correlation as well as positive, compared to the simpler scheme of LMM explained in Section 3.4. To assure that the covariance matrix for samples extracted from the same patient is positive semidefinite we need to assume that $\frac{1}{1-m} + 0.1 \leq \rho_g \leq 0.99$. All estimated $\rho_g$ that does not satisfy this requirement will be changed to the closest endpoint of this interval. The $\rho_g$-s are then transformed thorough the inverse hyperbolic tangent, $\theta_g = \tanh^{-1}(\rho_g)$. With these transformed values we can find a trimmed mean, where the largest and smallest 15% are trimmed to make the model more robust. We denote $\theta_{(1)} \leq \theta_{(2)} \leq \ldots \theta_{(G)}$, and then trimmed mean can be written as

$$\hat{\theta} = \frac{\sum_{i=3G/20}^{17G/20} \theta_{(i)}}{3G/10}. \tag{9}$$

Then the general $\rho$ is found by transforming the found mean through the hyperbolic tangent function, $\rho = \tanh(\hat{\theta})$ (Smyth, Michaud and Scott 2005).

Our model will have a covariance matrix $W_g$, where $\mathbf{Y}_g \sim N(X\hat{\boldsymbol{\beta}}, W_g)$. Here $W_g = Q_g V Q_g^T$, where $Q_g$ is a $(n \times 1)$ vector consisting of the weights found in Section 4.1.3. This system can be run through weighted least square to find estimators $\hat{\boldsymbol{\beta}}$ and $\hat{s}_g^2$. Observe that the covariance matrix for limma-voom will be gene specific because of weights and not correlation, as only the consensus correlation will be used in the covariance matrix.

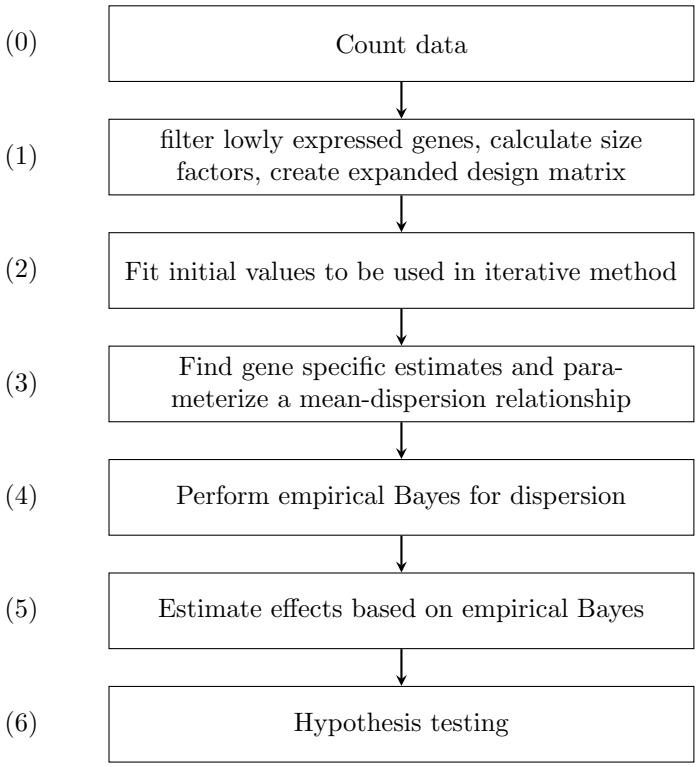



Figure 6: Overview of the standard DESeq2 pipeline for finding differentially expressed genes for RNA-sequencing count data. This figure was inspired by the explanation of the DESeq2 pipeline from Love, Huber and Anders (2014).

## 4.2 DESeq2

DESeq2 is a Bioconductor R-package that implements a pipeline for differential gene expression analysis based on the negative binomial distribution. The negative binomial distribution is used to model the count data in a GLM before empirical Bayes methods are used to improve estimates for effects and dispersion. The DESeq2 framework does not model correlation between different samples or estimate precision weights to reduce the relevance of poor RNA samples. In these situations the normal distribution based on a log-transformation, as used in the limma framework, is more attractive (Law, Chen et al. 2014).

The use of negative binomial model is motivated as follows. Intuitively we can consider the Poisson model for gene count data, where due to the randomness from the prepossessing the reads are independent of each other. This by itself is not enough to account for all the variability in the data. Instead we want to use a mixture model, the Gamma-Poisson model, also known as the negative binomial distribution. It turns out that this model is versatile enough to fit our data (Holmes and Huber 2018, Chap. 8).

### 4.2.1 Preprocessing

This section corresponds to Block 1 in Figure 6. DESeq2 uses raw counts in its pipeline, so little preprocessing is necessary. For preprocessing DESeq2 does normalization through size factors and removes lowly expressed genes. By removing lowly expressed genes we mean removing genes that for most of the samples have low counts, which can be considered as having little information about the gene expression. We are not likely to observe anything statistically significant due to the high dispersion for lowly expressed count data, and therefore remove it. This method is the same as the one used for limma in Section 4.1.1, the `filterByExpr` function from edgeR (Smyth 2016).

The purpose of normalization is to remove a systematic bias based on the sequencing depth.

DESeq2 does this through estimation of size factors from an observation specific median of the ratios of observed counts,

$$z_i = \text{median}_{g:r_g^R \neq 0} \frac{r_{gi}}{r_g^R}, \quad r_g^R = \left( \prod_{i=1}^{m} r_{gi} \right)^{1/m}.$$

It is preferable to use a geometric mean, as outliers will have less of an effect than if we were to use an arithmetic mean. It will also end up as closer to 1 then the arithmetic mean, which is preferable in this case. This will later be used as an offset in our model to correct for the differences between samples.

DESeq2 uses an expanded design matrix, which has a different variable for each level of the different factors, that means that the reference level for a factor is given its own parameter instead of being absorbed into the intercept of the model. This expanded design matrix does not necessarily have full rank so we use a normal zero centered prior for regularization, such that the effects gets unique solutions. The expanded design matrix can be used together with a regularization method to give bias reduced estimates for all levels for a factor. This regularization ends up being the same as using ridge regression for a specific shrinkage parameter $\lambda$ decided by the data.

### 4.2.2 Negative binomial generalized linear model

This section corresponds to Block 2 in Figure 6. DESeq2 is based on the negative binomial distribution as introduced in Section 3.1.2, where it uses the $\log_2$ link function, $\eta_i = \log_2(\mu_i) - \log_2(z_i)$, where $\log_2(z_i)$ is an offset variable and $z_i$ is defined in Section 4.2.1.

DESeq2 uses a concept based on Cooks distance to determine if an observation should be labelled an outlier or not. Cook's distance can be considered a distance which measures the influence a sample has on the fitted coefficients for a gene (Holmes and Huber 2018). An outlier is categorized by comparing the Cook's distance to Fisher distribution, where the Cook's distance depends on the Pearson residual and the hat matrix, $H = W^{-1/2}X(X^TW^{-1}X)^{-1}X^TW^{-1/2}$, for a given sample (Love, Huber and Anders 2014). If an observation is classified as an outlier it would be replaced by a trimmed mean over all values.

### 4.2.3 Mean-variance relationship

This section corresponds to Block 3 and Block 4 in Figure 6. The negative binomial distribution has an inherent mean-variance relationship, as can be observed from the variance of a negative binomial variable, see equation (2). More strength can be gained by borrowing information across genes in the form of a mean-variance relationship. It has been observed that the ML dispersion estimates, $\hat{\alpha}_g^{gw}$, also should have a dependency on the mean (Anders, Reyes and Huber 2012). DESeq2 models the dispersion as dependent on the average expression strength,

$$\bar{\mu}_g = \frac{1}{n} \sum_{i=1}^{n} \frac{r_{gi}}{z_i}, \tag{10}$$

where outliers are removed. It is assumed that this trend can be written on the form, $\alpha_{tr}(\bar{\mu}_g) = \frac{\alpha_1}{\bar{\mu}_g} + \alpha_0$. To estimate this takes multiple steps, where it initially estimates the effects using the maximum likelihood, as explained in Section 3.2.1. This method requires the use of a one parameter GLM, so it uses a moment of method estimate for the dispersion, $\alpha_g^{gw}$, and treats it as a constant. Then the effects are estimated and used to improve the estimates of the dispersions by using a modified profile likelihood for a better gene-wise estimate of the dispersion. This has the effect of reducing the bias of our gene-wise dispersion estimates. With the improved dispersion estimates the hyperparameters $\alpha_0$ and $\alpha_1$ can be estimated to get a mean variance relationship, before estimating the effects (Love, Huber and Anders 2014).

The final dispersion estimate is found using empirical Bayes. It has been observed by Wu, Wang and Wu (2013) that a log normal prior fits the dispersion distribution for the trend for typical

count data. With the trend we can estimate a variance for the log normal prior from the data. This variance can be split into two contributions, where the first is a measure of variance for the logarithmic dispersion around the trend whilst the second arises from the log normal sampling distribution for the dispersion estimator. With this prior and the trend we can find maximum a poseriori (MAP) estimates for all genes. Further detail about the procedure can be found in Love, Huber and Anders (2014).

Now that we have an initial gene-wise estimate and a MAP estimate for the dispersion we want to compare them and evaluate if using the MAP estimate is reasonable. First we find a standard deviation based on the median absolute deviation between the log of the trend and the initial gene-wise estimate,

$$s_{tr} = \text{median}_g \left( |\log \alpha_g^{gw} - \log \alpha_{tr}(\bar{\mu}_g)| \right).$$

If the log difference between the gene-wise dispersion estimate and the estimated trend is more than two times the estimated standard deviation, $\log \alpha_i^{gw} > \log \alpha_{tr}(\bar{\mu}_g) + 2s_{tr}$, then the gene-wise estimate is used instead of the MAP estimate. DESeq2 claims that the MAP estimate is then unreasonable and the adjustment would only lead to false positive test results (Love, Huber and Anders 2014).

### 4.2.4 Effect estimation

This section corresponds to Block 5 in Figure 6. For effects DESeq2 uses a normal zero centered prior with a variance, $\hat{\sigma}_r^2$, for each effect in the expanded design matrix. We will not cover how to estimate $\hat{\sigma}_r^2$, but a general outline can be found in Love, Huber and Anders (2014). We will go further into detail on how coefficients for the different effects are estimated. Assuming that we have found $\hat{\sigma}_r^2$ for all the different effects in the expanded design matrix we can extend the concepts introduced in 3.2.1 to use a zero centered normal prior. In the empirical Bayes setup we want to maximize the log-likelihood with an regularizing term, so we only need to manipulate the first and second derivative of the log-likelihood to implement this into our scheme. The first and second derivative of the log-likelihood can be found in Equation (3), where the score function is the first derivative and the expected Fisher information is the negative of the double derivative. The term we want to add is of the form

$$-\frac{\lambda}{2}\hat{\boldsymbol{\beta}}^{*T}\hat{\boldsymbol{\beta}}^{*},$$

where $\lambda$ denotes a diagonal matrix with elements $\lambda_{rr} = \hat{\sigma}_r^2$ (Love, Huber and Anders 2014). The star, which will be used for both $\boldsymbol{\beta}$ and $X$, denote that they use the expanded design matrix. We can then change the score function from Equation (3) to include the term from the prior,

$$\mathbf{U} = \frac{1}{\phi} X^{*T} W M (\mathbf{r}_g - \boldsymbol{\mu}) - \lambda \boldsymbol{\beta}^{*}.$$

We also change the expected Fisher from Equation (3) to include the term from the prior,

$$\mathcal{I} = \frac{1}{\phi} X^T W X + \lambda.$$

Then taking into account the extended design matrix the iterative scheme becomes

$$\hat{\boldsymbol{\beta}}^{*(r+1)} = \left( \frac{1}{\phi} X^{*T} W^{(r)} X^{*} + \lambda \right)^{-1} \frac{1}{\phi} X^{*T} W^{(r)} \left( X^{*} \hat{\boldsymbol{\beta}}^{*(r)} + M(\mathbf{r}_g - \boldsymbol{\mu}^{(r)}) \right),$$

where $\mathbf{r}_g$ is the gene specific vector of counts. We can now use the link function $\eta_i = \log_2 \left( \frac{\mu_i}{z_i} \right)$ and observe that $M_i = \frac{\partial \eta_i}{\partial \mu_i} = \frac{1}{\mu_i \ln(2)}$ and $\phi = 1$. A more detailed derivation is given in Appendix A.1. The scheme can then be rewritten to

$$\hat{\boldsymbol{\beta}}^{*(r+1)} = \left( X^{*T} W^{(r)} X^{*} + \lambda \right)^{-1} X^{*T} W^{(r)} Z^{(r)},$$

where $Z^{(r)}$ is a diagonal matrix with elements $Z_{ii}^{(r)} = \log_2 \frac{\mu_i^{(r)}}{z_i} + \frac{r_{gi} - \mu_i^{(r)}}{\mu_i^{(r)} \ln(2)}$. We observe that this is similar to the iteratively reweighted ridge regression algorithm with a given $\lambda$.

After the publication of Love, Huber and Anders (2014) DESeq2 has changed their standard pipeline from using a prior on the effects for effect estimation. Now the maximum likelihood estimates are used in the standard pipeline and shrinkage of the estimators is optional and done during the hypothesis testing. This is done by using a more general function, `LFCShrink` that is part of the edgeR library Love, Anders and Huber (2022). This change was motivated by the shrinkage resulting from the normal prior sometimes being too strong (Love, Anders and Huber 2022). For our analysis we will use the pipeline described in Love, Huber and Anders (2014) and use the older routine with ridge like regularization.

### 4.2.5 Wald test

This section corresponds to Block 6 in Figure 6. For the Wald statistic we use Equation (6), where we replace the expected Fisher information with the estimated variance of our effects. This is the same concept as for Section 3.3.1, but using a Bayesian (ridge) approach. The variance becomes,

$$\text{Var}\left[\hat{\boldsymbol{\beta}}^{*(r+1)}\right] = \left(X^{*T} W^{(r)} X^* + \lambda\right)^{-1} X^{*T} W^{(r)} X^* \left(X^{*T} W^{(r)} X^* + \lambda\right)^{-1}.$$

The derivation of the variance can be found in Appendix A.2.

# 5 Negative binomial generalized linear mixed model

In this section we will introduce the generalized linear mixed effects models (GLMM), which will be explained by concepts from LMM and GLM. We will examine how a negative binomial GLM can be fitted based on data generated by a negative binomial GLMM, and how we can interpret the resulting model. This will be observed through simulations, where we will compare fits of GLMM and GLM to data simulated based on a GLMM model.

In this section we will consider the negative binomial GLM and GLMM with a log link, where $e$ is the base instead of 2. Previously we have considered the log link with 2 as the base for the interpretation of log-fold change, but equations become simpler with $e$ as the base. The difference between the bases is just a factor, so if we want to change our estimates from using $e$ as the log base to 2 we only need to use a scaling factor, $e^{\boldsymbol{x}_i^T \boldsymbol{\beta}} = 2^{\log_2(e) \boldsymbol{x}_i^T \boldsymbol{\beta}}$. The resulting test statistics, for both GLM and GLMM, will be the same regardless of the base used, as the constant scaling factor $\log_2(e)$ will be present in both effect estimate and variance estimate, but the effect estimate and variance estimate will be scaled based on the choice of log base.

To define the GLMM we will compare it to the GLM and highlight differences. Similarly as for the GLM the GLMM consists of a random component and a systematic component. The definition for the random component for the GLMM is the same as the one used in Section 3.1, so the main difference between a GLM and GLMM is in the systematic component.

Before we consider the systematic component we want to rearrange our data to a similar structure to the one used in Section 3.4. Where for each cluster we have vectors for the responses and offsets, and a matrix for the covariates. For the systematic component we want to consider a cluster triple $(X_i, \boldsymbol{o}_i, \boldsymbol{y}_i)$ for cluster $i$, $i = 1, 2, \ldots, q$. The systematic component is assumed to be an cluster specific linear predictor, $\boldsymbol{\eta}_i = \mathbf{o}_i + X_i \boldsymbol{\beta} + \mathbf{1}_{n_i} \gamma_i$, where $X_i$ is a known matrix of covariates. With this interpretation $\boldsymbol{\gamma}$ is a vector of cluster specific random intercepts. GLMM can also fit random effects, then $1_{n_i}$ needs to be changed to a cluster specific design matrix and $\gamma_i$ needs to be changed to a vector of effects (Fahrmeir et al. 2007, Chap. 7). We will only use random intercepts.

The only difference in the systematic component between the GLM and GLMM is that GLMM uses a random component, $\gamma_i$, in the linear predictor. Observe that the linear predictor is the same as what is used for LMM, so in other words GLMM is just a GLM that the random effect extension of the linear predictor used in LMM.

## 5.1 Motivation for using a generalized linear mixed effects model

GLM models the mean of response to be a function of the given covariates, but this is not necessarily the case if multiple observations are from the same patient. Intuitively one would assume that the molecules in an extracted sample is dependent on the DNA of the patient it is extracted from. And therefore two observations from the same patient are believed to be correlated. We can consider observations from the same patient to be part of the same cluster. When multiple draws are from the same cluster they can no longer be considered independent and corrections must be considered to get an accurate model. Limma-voom solves this by estimating a consensus correlation and alters the expected correlation matrix to account for this, see Section 4.1.6. In the case of a negative binomial GLM the variance is dependent on the mean which makes direct alteration of the correlation matrix difficult. In this case we turn to a negative binomial GLMM. To decide if a mixed model is a good model for the data we can consider the estimated correlation between observations. This can be used to tell if it is even necessary to use a GLMM or if the GLM gives a decent enough fit. Our hypothesis is that if we were to ignore the correlation between our data, similarly for the LMM example in Section 3.4, the variance estimate of effects between clusters would be too small. Thus, we will now study the correlation structure of the negative binomial GLMM.

## 5.2 Negative binomial covariance structure

Our data has multiple observations from the same person, but DESeq2 has no inherent way of modeling the correlation of these. One intuitive solution to this is to use a GLMM instead of a GLM, where we use a random intercept for the different patients the samples are extracted from. Due to the high variance of the negative binomial distribution we believe the model might have problems estimating the random intercepts accurately.

Consider a negative binomial GLMM with link function $\boldsymbol{\eta}_i = \log(\boldsymbol{\mu}_i)$, where $\boldsymbol{\eta}_i = \mathbf{o}_i + X_i\boldsymbol{\beta} + \mathbf{1}_{n_i}\gamma_i$ and $\boldsymbol{\gamma}$ is a random intercept following $N(\mathbf{0}, I_q\tau^2)$, and $\mathbf{o}_i$ is a known offset. We can use the law of total expectation and the law of total covariance to get marginal estimates of the mean and variance for this GLMM (Agresti 2003, Chap. 9.4). Both the law of total expectation and covariance consist of two layers, one inner which is conditional on the random intercept and one outer layer that uses the expected values found in the inner layers as constants. The law of total expectation and covariance are provided as a Theorem in Appendix B in Fahrmeir et al. (2007).

As we will consider a log link with normally distributed random intercepts we want to introduce the log-normal distribution. If $Z$ is a normally distributed variable with expected value $E[Z] = \mu$ and variance $\text{Var}[Z] = \sigma^2$, then $e^Z$ is log-normally distributed with mean $E[e^Z] = e^{\mu+\sigma^2/2}$ and variance $\text{Var}[e^Z] = e^{\sigma^2-1}e^{2\mu+\sigma^2}$. We will also need the expected square of a log-normally distributed variable, $E\left[\left(e^Z\right)^2\right] = e^{2\mu+2\sigma^2}$ (Johnson, Kotz and Balakrishnan 1994, Chap. 14).

We will consider $\gamma_i$, which is normally distributed with $E[\gamma_i] = 0$ and $\text{Var}[\gamma_i] = \tau^2$. The expected gene expression count in cluster $i$ and from observation $j$ becomes,

$$E[Y_{ij}] = E[E[Y_{ij}|\gamma_i]] = E[e^{o_{ij}+\boldsymbol{x}_{ij}^T\boldsymbol{\beta}+\gamma_i}] = e^{o_{ij}+\boldsymbol{x}_{ij}^T\boldsymbol{\beta}}E[e^{\gamma_i}] = e^{o_{ij}+\boldsymbol{x}_{ij}^T\boldsymbol{\beta}}e^{\tau^2/2}. \tag{11}$$

If we assume that we have two observations from cluster $i$, from individual $j$ and $j'$ the covariance becomes,

$$\text{Cov}(Y_{ij}, Y_{ij'}) = E[\text{Cov}(Y_{ij}, Y_{ij'}|\gamma_i)] + \text{Cov}(E[Y_{ij}|\gamma_i], E[Y_{ij'}|\gamma_i]).$$

We observe that the expected value of the covariance when the random effect is given will be 0, when $j$ and $j'$ are different observations, which results in

$$\begin{aligned}\text{Cov}(Y_{ij}, Y_{ij'}) &= \text{Cov}(E[Y_{ij}|\gamma_i], E[Y_{ij'}|\gamma_i]) = \text{Cov}(e^{o_{ij}+\boldsymbol{x}_{ij}^T\boldsymbol{\beta}+\gamma_i}, e^{o_{ij'}+\boldsymbol{x}_{ij'}^T\boldsymbol{\beta}+\gamma_i}) \\ &= e^{o_{ij}+\boldsymbol{x}_{ij}^T\boldsymbol{\beta}}\,\text{Cov}(e^{\gamma_i}, e^{\gamma_i})e^{o_{ij'}+\boldsymbol{x}_{ij'}^T\boldsymbol{\beta}} \\ &= e^{o_{ij}+o_{ij'}+(\boldsymbol{x}_{ij}^T+\boldsymbol{x}_{ij'}^T)\boldsymbol{\beta}}e^{\tau^2}(e^{\tau^2}-1).\end{aligned} \tag{12}$$

In the case where $j$ and $j'$ is the same observation we end up with the variance for that observation

$$\begin{aligned}
\text{Var}[Y_{ij}] &= E[\text{Cov}(Y_{ij}, Y_{ij}|\gamma_i)] + e^{2o_{ij}+2\boldsymbol{x}_{ij}^T\boldsymbol{\beta}}e^{\tau^2}(e^{\tau^2}-1) \\
&= E[e^{o_{ij}+\boldsymbol{x}_{ij}^T\boldsymbol{\beta}+\gamma_i} + \alpha_{GLMM}e^{2o_{ij}+2\boldsymbol{x}_{ij}^T\boldsymbol{\beta}+2\gamma_i}] + e^{2o_{ij}+2\boldsymbol{x}_{ij}^T\boldsymbol{\beta}}e^{\tau^2}(e^{\tau^2}-1) \\
&= e^{o_{ij}+\boldsymbol{x}_{ij}^T\boldsymbol{\beta}}E\left[e^{\gamma_i}\right] + \alpha_{GLMM}e^{2o_{ij}+2\boldsymbol{x}_{ij}^T\boldsymbol{\beta}}E\left[e^{2\gamma_i}\right] + e^{2o_{ij}+2\boldsymbol{x}_{ij}^T\boldsymbol{\beta}}e^{\tau^2}(e^{\tau^2}-1) \\
&= e^{o_{ij}+\boldsymbol{x}_{ij}^T\boldsymbol{\beta}}e^{\tau^2/2} + \alpha_{GLMM}e^{2o_{ij}+2\boldsymbol{x}_{ij}^T\boldsymbol{\beta}}e^{2\tau^2} + e^{2o_{ij}+2\boldsymbol{x}_{ij}^T\boldsymbol{\beta}}e^{\tau^2}(e^{\tau^2}-1).
\end{aligned} \tag{13}$$

The correlation can then be found by inserting $\text{Cov}[Y_{ij}, Y_{ij'}]$, $\text{Var}[Y_{ij}]$, and $\text{Var}[Y_{ij'}]$ from (12) and (13) into

$$\text{Corr}[Y_{ij}, Y_{ij'}] = \frac{\text{Cov}(Y_{ij}, Y_{ij'})}{\sqrt{\text{Var}[Y_{ij}]}\sqrt{\text{Var}[Y_{ij'}]}}. \tag{14}$$

## 5.3 Estimation for generalized linear mixed effects model

A GLMM can be considered a two stage model, where in the first stage the model is conditional on the random effects, $\gamma_i$, the response is supposed to follow a GLM. This is often called the conditional model. In this case the observations are assumed to be independent with a cluster specific known offset, the linear predictor then takes the form $g(E[\boldsymbol{Y}_i|\gamma_i]) = \boldsymbol{\eta}_i = X_i\boldsymbol{\beta} + \mathbf{1}_{n_i}\gamma_i$. In the second stage $\gamma_i$ are assumed to be independent draws from $\boldsymbol{\gamma} \sim N(\mathbf{0}, I_q\tau^2)$. We let $f(\boldsymbol{\gamma}; I\tau^2)$ denote the normal probability density function, with mean $\mathbf{0}$ and variance $I\tau^2$, and let $f(\boldsymbol{Y}_i|\gamma_i; \boldsymbol{\beta_g})$ denote the conditional probability mass function of $\boldsymbol{Y}_i$ given $\gamma_i$. The marginal distribution of $Y$ after integrating out $\boldsymbol{\gamma}$ is considered the likelihood for a GLMM,

$$\ell(\boldsymbol{\beta}, I\tau^2; Y) = \prod_{i=1}^{q} \int f(\boldsymbol{Y}_i|\gamma_i; \boldsymbol{\beta}) f(\gamma_i; \tau^2)\, d\gamma_i. \tag{15}$$

In most cases this is approximated numerically through Gauss-Hermite quadrature, Monte Carlo methods or Laplace approximation. Then this is maximized to find our parameter estimates, usually through Newton-Rapson (Agresti 2003, Chap. 9.5).

Here we only consider normal random effects. Rarely the normality assumptions for the random effects can be checked. This raises the question if misspecification of the distribution for the random effects can harm the model. However selecting an incorrect distribution for the random effect does not tend to alter the bias of the estimators for those effects. Even though the random effects can give different predicted values, the overall accuracy of prediction tends to be similar. One should note that the between cluster effects can be sensitive to the distribution choice of random effects, when the variance for random effects are assumed constant but actually depends on the covariate values (Agresti 2003, Chap. 9.5).

We can observe that the marginal distribution in Equation (15) is similar to a Bayesian posterior distribution, with a zero centered normal prior on the cluster specific effects.

When the GLMM is underdispersed, when the estimated variance of the model is less than the mean, or the estimated Fisher information is close to singular can make the GLMM not converge. In the case where only the random effects do not converge the GLMM returns a model with $\tau = 0$ (Bolker 2022). We have used the function function `glmer.nb` from the lme4 package (*glmer.nb: Fitting Negative Binomial GLMMs* 2022).

## 5.4 Simulations

The core of DESeq2 is a negative binomial GLM, where data is assumed to be independent. In the case of correlated data (for example from the same person), there is as far as we know, no solution for DESeq2. Our question is if it is possible to replace the negative binomial GLM with the negative binomial GLMM in DESeq2. DESeq2 is a complex pipeline that models relations across many genes for RNA-Seq, and we would like to start at a more basic level by examining RNA-Seq

for a single gene. We want to accomplish this by simulating a dataset based on a negative binomial GLMM and fit both a GLMM and GLM to observe how well a GLM can deal with correlated data on a gene-wise level.

An overview of how we want to present this can be seen from this list:

1. First we want to get a better idea of how a negative binomial GLMM differs from a GLM for the conditional and marginal distribution, particularly how the mean and variance is influenced by this change, in Section 5.4.1.
2. Further we want to comment on parameter choices we want to use for a gene-wise simulation based the GLMM, where they are based on the parameter estimates from the DESeq2 analysis on the dataset presented in Section 2.5, this will be presented in Section 5.4.2.
3. Then we want to do a gene specific simulation under $H_0$, where our goal is to observe how accurate a GLM can estimate an underlying GLMM for data similar to the dataset introduced in Section 2.5, which will be presented in Section 5.4.3.
4. Finally we want to compare these gene-wise simulations to similar ones done under $H_1$, which will be presented later in Section 5.4.3.

### 5.4.1 General observations for a negative binomial GLMM

For the simulations we want a simple model for the fixed linear predictor, with one intercept and a two level factor, $\boldsymbol{\beta} = [\beta_0\, \beta_1]^T$. Initially we want our effect to be 0 such that our test follows $H_0 : \beta_1 = 0$, to be able to asses the validity of the Wald test for GLM and GLMM. The data models clusters with $n_i$ observations through the linear predictor, $\boldsymbol{\eta}_i = X\boldsymbol{\beta} + 1_{n_i}\gamma_i$. Each individual have their own random effect, $\gamma_i$, regardless if they have one or two observations. For our effects we want equal number of observations with and without the effect. For our data simulation we choose to use a dispersion parameter, $\alpha$, based on the ones estimated from DESeq2 for our data. Specifically we choose to use the median from the DESeq2 analysis, $\alpha = 0.25$.

First we examine the conditional distribution for a random selection of random effects for the negative binomial distribution. This will be done by observing the conditional probability mass function for a specific observation. This is shown in Figure 7 where the negative binomial GLM probability mass functions are plotted in blue, compared to a conditional model where $\gamma_i = 0$ is plotted in red. The figure was produced using the parameters $\alpha = 0.25, \boldsymbol{\beta} = [3, 0]^T, \tau = 0.3$. When $\gamma_i = 0$ we end up with a conditional model without a random intercept, which corresponds to a GLM with the same parameters.

We want to simulate a dataset with $n = 200$ observations, where we want clusters to have different random effects. We want to split our data into different clusters, where a defined number of clusters, $m$, have two observations and the rest of the observations, $n - 2m$, come from separate clusters. This can also be considered as a proportion of correlated data, which we will denote as $m^* = \frac{2m}{n}$. Note that in our case one observation is only dependent on one other observation, otherwise it is independent. Before we do simulations we see that both Equation (11) and (13) tells us that both the expected value and variance of the marginal model will increase when the variance of the random effects increases.

To get a better intuition of the marginal distribution we want to draw random effects and draw an observation from the conditional distribution to simulate the marginal distribution, where each observation is weighted by the inverse of the total number of draws. In Figure 8, 20000 observations were drawn with $\alpha = 0.25, \boldsymbol{\beta} = [3, 0]^T, \tau = 0.3$ in blue compared to the same model without random effects, $\tau = 0$, in red. We can observe the anticipated increase of the expected value and standard deviation from Equation (11) and (13) in both Figure 8 and Table 4.

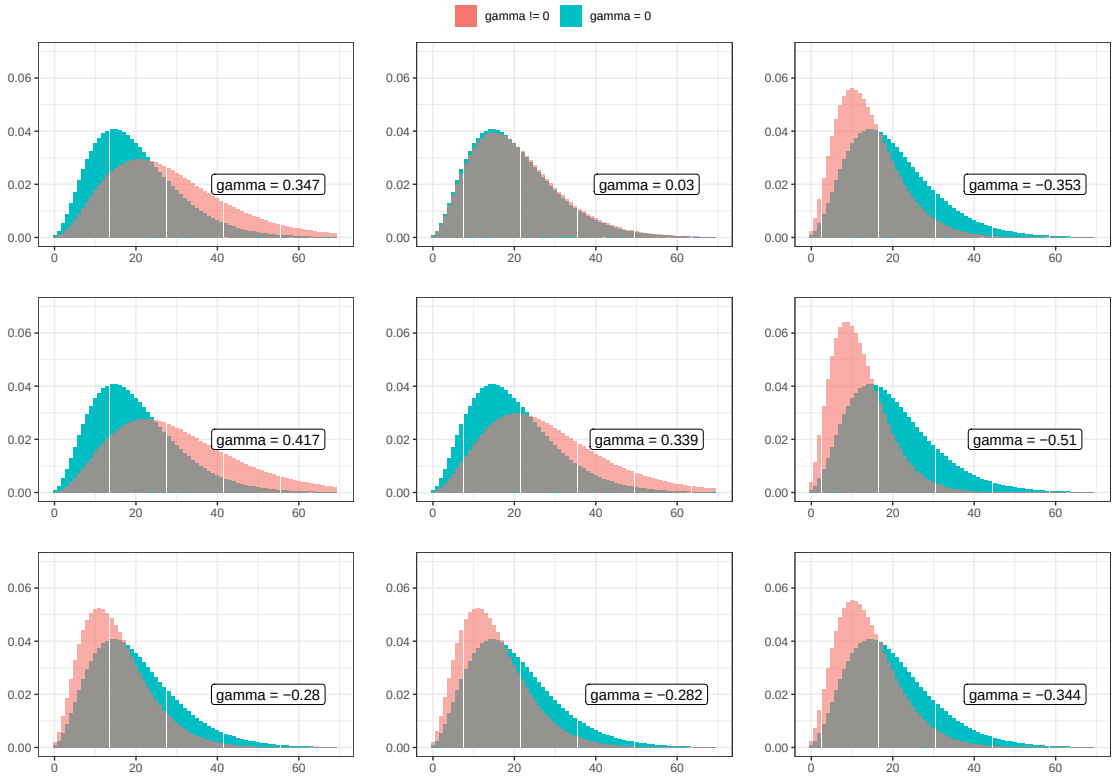



Figure 7: Nine different conditional negative binomial probability mass functions, for different simulated $\gamma_i$, in turquoise with a conditional mass function for $\gamma_i = 0$ in red for reference. The probability mass functions follow the model explained in Section 3.1.2 with $\alpha = 0.25, \boldsymbol{\beta} = [3, 0]^T, \tau = 0.3$.

Table 4: Table of theoretical and estimated values for both expected value and standard deviation for the responses from a simulated marginal distribution of a negative binomial GLMM. The conditional distribution at $\gamma_i = 0$ is included to observe the difference between including and excluding the random effect. The negative binomial model and parameters are explained in Section 3.1.2, in this table we used the parameters $\alpha = 0.25, \boldsymbol{\beta} = [3, 0]^T, \tau = 0.5$ where 50 random effects have two observations and 100 random effects have one observation, in other words $m^* = 0.5$, for this simulation.

| | Expected value | Standard deviation |
|---|---|---|
| Conditional $\gamma_i = 0$ | 20.08 | 11.00 |
| Theoretical | 22.76 | 18.33 |
| Estimated | 22.65 | 18.34 |

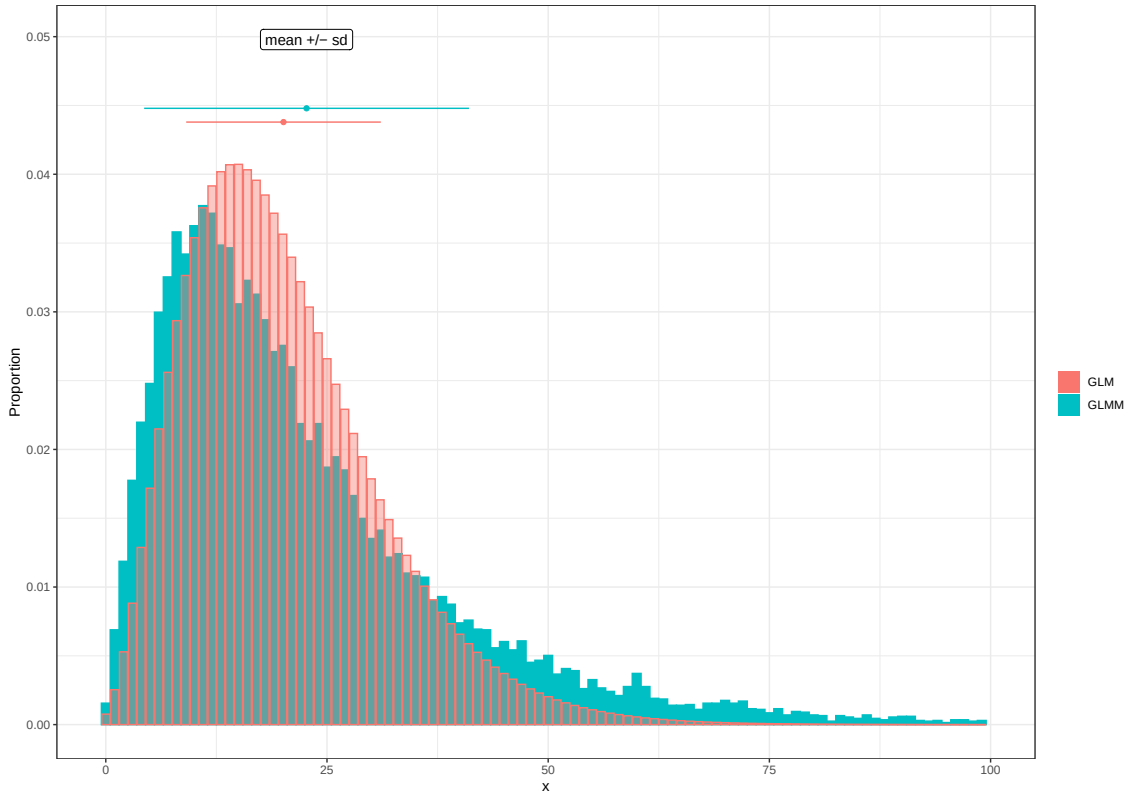



Figure 8: Marginal GLMM distribution with data estimated from a negative binomial GLMM using log link. The negative binomial model and parameters are explained in Section 3.1.2, in this figure we use the parameters $\alpha = 0.25, \boldsymbol{\beta} = [3, 0]^T, \tau = 0.5$ where 50 clusters have two observations and 100 clusters have one observation. The blue figure shows a simulated marginal distribution for comparison to a model without mixed effects. The red figure shows the probability mass function of a GLM with the same parameters as the GLMM, in other words if $\tau = 0$, for reference. The overlap of the two probability functions, which are red and blue, becomes grey.

If we want to simulate data following a negative binomial GLMM, then a GLM fit to the marginal distribution would need to estimate a new $\alpha$, as the one used for simulating the data does not account for the increase in dispersion from the random effects. When the GLM needs to estimate a parameter that is known for the GLMM, $\alpha$, we also want the GLMM to estimate it to not give another unfair advantage in the fitting procedure. Further, note that the final dispersion estimate for DESeq2 is in most cases a combination of the regularized maximum likelihood estimate and the estimated mean-variance relationship. Initially we want to do a gene specific simulation, where we do not have the option to borrow information across genes. Therefore there is not a direct relation between the $\alpha$ estimated in the simulations and the final DESeq2 dispersion estimate used for our data.

GLM does not allow person specific normal distributed effect, which GLMM is based on, this makes them difficult to compare. We observe from both Figure 8, Table 4 and Equation (13) that if we were to use the same parameter estimates for the GLM and the GLMM the GLMM would estimate a higher expected value and variance. In our case where we want to base the data on a GLMM, then the GLM cannot use any of the input parameters for its estimates. We want GLMM to also estimate all the parameters that were used for input to not give a more unfair advantage for GLMM for model fit.

To summarize we observed through equations, figures and tables that the random intercept adds another layer of variance that needs to be accounted for. As a consequence of considering a log-link the expected value and variance of the marginal model increased. This requires the GLM we fit to this dataset to alter the intercept and dispersion estimate to account for the additional layer of variance in the marginal model.

### 5.4.2 Model setup for simulation

We now shift our focus to the use of the Wald test for GLM and GLMM, and study the cumulative distribution function of $p$-values for both simulated GLM and GLMM and compare their values at the 0.05 level. The distribution of $p$-values are expected to be uniform if the Wald test gives exact $p$-values. We will study this by comparing the CDF of $p$-values for GLM and GLMM, and for reference compare to the CDF of a uniform distribution bound by 0 and 1. If the cumulative distribution of $p$-values is above the CDF of the uniform distribution the test is optimistic, that means it finds more tests to be statistically significant than desired at that specific level. Otherwise if it is below the test is conservative, and finds less tests to be statistically significant than desired at that specific level. We would prefer a conservative test, as even though an optimistic test is more powerful it does not hold its level of significance. For our different simulations we will keep $\boldsymbol{\beta}$ and $\alpha$ the same and vary the variance of the random effects, $\tau$, and the number of clusters with two observations, $m$. When both $\boldsymbol{\beta}$ and $\alpha_{GLMM}$ are held constant, then changing $\tau$ is effectively changing the correlation between observations from different clusters, which can be observed for Equation (12) and (13). We would expect the GLMM to give a better fit the more observations we had that were correlated, and GLM should get worse with an increase in correlated observations. We would also expect that an increase in correlation would worsen the fit for GLM. The most important thing for our test is that it holds its level of significance, as it is better for a test to be conservative compared to optimistic.

The parameters to be used in the simulation are motivated from the analysis of the Crohn's data presented in Section 2.5, and analyzed in Section 6. For these data limma-voom found the consensus correlation to be estimated to approximately 0.1. In the negative binomial simulations we therefore want to include a correlation similar to this in value. We also want to observe higher correlations to investigate if the GLM fit still will be good. The mean of counts after filtering is $e^{4.5} \approx 90$ and the median of estimated $\alpha$ is 0.25. For our simulation we have chosen to use $\beta_0 = 5.3, \quad \alpha = 0.25$ for our parameters. These values were inspired by gene $PDE6A$, which is one of the top genes for a contrast of interest, the base mean and estimated log fold change for this gene can be found in Table 5 in Appendix C.3. We want to vary the proportion of correlated data, $m^* = \{0.3, 0.6, 1\}$. We also want to vary the variance of the random effect, $\tau = \{0.2, 0.3, 0.4\}$, which corresponds to $\text{Corr}[Y_{ij}, Y_{ij'}] \approx \{0.13, 0.25, 0.37\}$.

### 5.4.3 gene-wise simulations

To further examine the difference between GLM and GLMM we want to simulate 1000 different datasets based on a GLMM with the predefined parameters, then fit both a GLM and GLMM to each dataset and compare all parameter estimates. We want to compare plots of cumulative $p$-values, to check if the models are conservative or optimistic under $H_0$. Note that each $p$-value is based on the Wald test, which is dependent on the effect estimate, dispersion estimate and the mean of counts through the mean dependent variance. A plot of cumulative $p$-values with this setup is found in Figure 9. Here we observe that the GLMM estimates are generally closer to the exact uniform CDF than the GLM estimates, which is what we would expect. The cumulative $p$-values of the GLMM fluctuates around the uniform CDF, which we will call the exact line, so dependent on the level of significance chosen the test can be either conservative or optimistic. This is not a desired trait, but could be acceptable if the difference between the estimated level and exact significance level is negligible. Further we can observe that the GLM Wald test grows more conservative the higher the within group correlation and the higher proportion of correlated data becomes. From Figure 9 it looks like the GLM Wald test gives a similar distribution of $p$-values as the GLMM Wald test when the proportion of correlated data is low, regardless of the level of correlation. Intuitively we assumed that this addition of a cluster specific effect would worsen the fit regardless of the proportion of correlated data, but it seems like the GLM still gives a good estimate of the effect and variance of the effect. We should comment that our Wald test appears conservative for the GLM when the dataset is generated to follow a GLMM. Note that the GLM we model for gene-wise estimates is not directly comparable to the GLM in DESeq2, as it does not use any information from the mean variance relationship. Of main interest is how many percent of the hypothesis tests that are found to be statistically significant under $H_0$ with a significance level of 5%. This will tell us how many false positives we have found and if this is consistent with the level of significance we have chosen. We also want to compare the difference between the dispersion estimates for the models and the effects.

The following figures are generated:

- Figure 10: Here the intercepts, $\beta_0$, of the GLM and GLMM from the simulations are plotted.
- Figure 11: Here the effects, $\beta_1$, of the GLM and GLMM from the simulations are plotted.
- Figure 12: Here the dispersion estimate, $\alpha$, of the GLM and GLMM from the simulations are plotted.
- Figure 13: Here the standard deviation of the random effects, $\tau$, of the GLMM from the simulations are plotted.

To start with we want to observe the estimates of the intercept, which can be seen in Figure 10. Both the dispersion and intercept are assumed to be higher for the GLM, as it tries to estimate the marginal model and cannot model random effects, whilst the GLMM we expect to have estimates similar to the parameters used for simulation. The estimates of intercept seems to be constant for different proportions data that are correlated, $m^*$, whilst it varies with a change in correlation. The value that was used in simulations is $\beta_0 = 5.3$. There seems to be an increase of the intercept estimated for the GLM when the correlation, which is the same as $\tau$ when other variables are held constant, increases the estimated intercept. This makes sense when we consider the marginal expected value from Equation (11), where the expected value is higher regardless of the presence of effect, so it needs to be captured by the intercept. It is interesting to observe that the estimated intercept for the GLMM decreases with an increase of the correlation. This decrease is taking the estimated value further and further away from the value used to simulate the datasets. This seems to be the case for all the different proportions of correlated data, $m^*$, as well, we do not know why this is the case.

With similar logic we hypothesise that the effect estimates will be similar for both models under $H_0$, and $H_1$. We can observe that this is the case under $H_0$ from Figure 11. The effect estimate seems to have the same distribution independent on both the correlation and the proportion of correlated data, $m^*$.

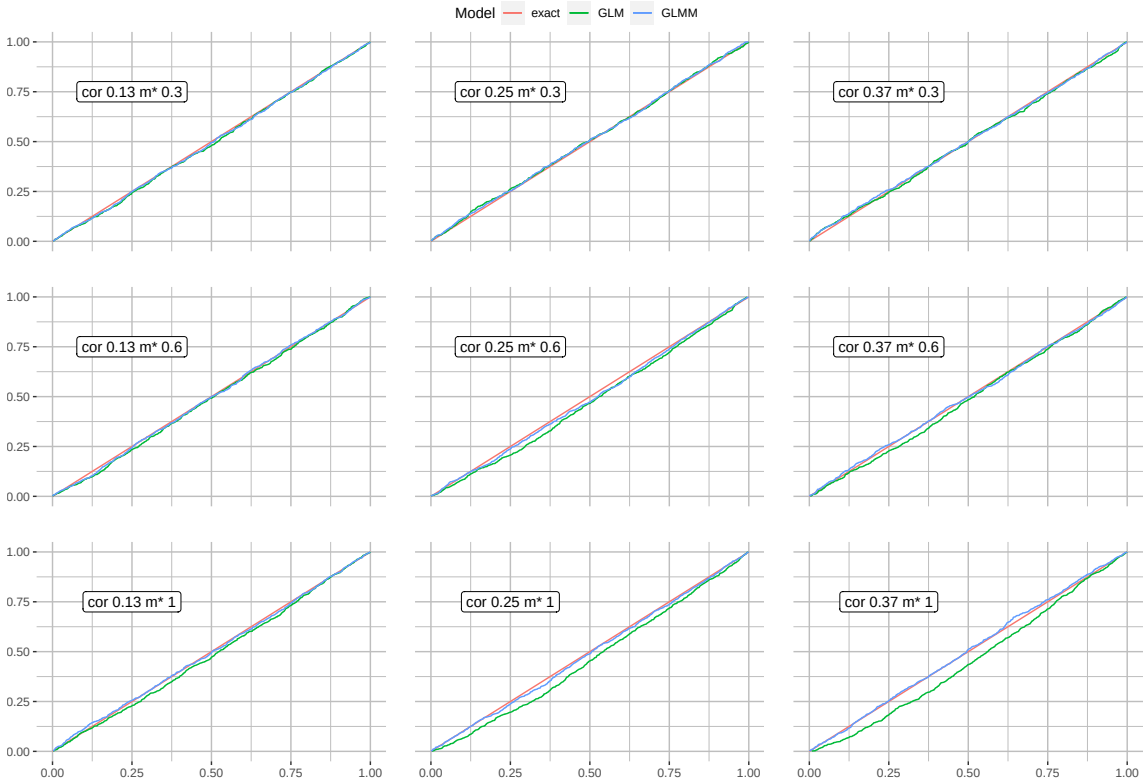



Figure 9: Cumulative $p$-values from a GLM and GLMM Wald test, fitted to data produced following the assumptions of the GLMM under $H_0$. 1000 models were fit to different datasets produced by the negative binomial distribution with log link, where $\beta_0 = \log(200)$, $\alpha = 0.25$, and an effect $\beta_1 = 0$. These are plotted against the CDF of a uniform distribution, in red, which we call exact. The cumulative $p$-values from the GLM is plotted in green and the GLMM in blue. Both the $x$- and $y$- axis go from 0 to 1.

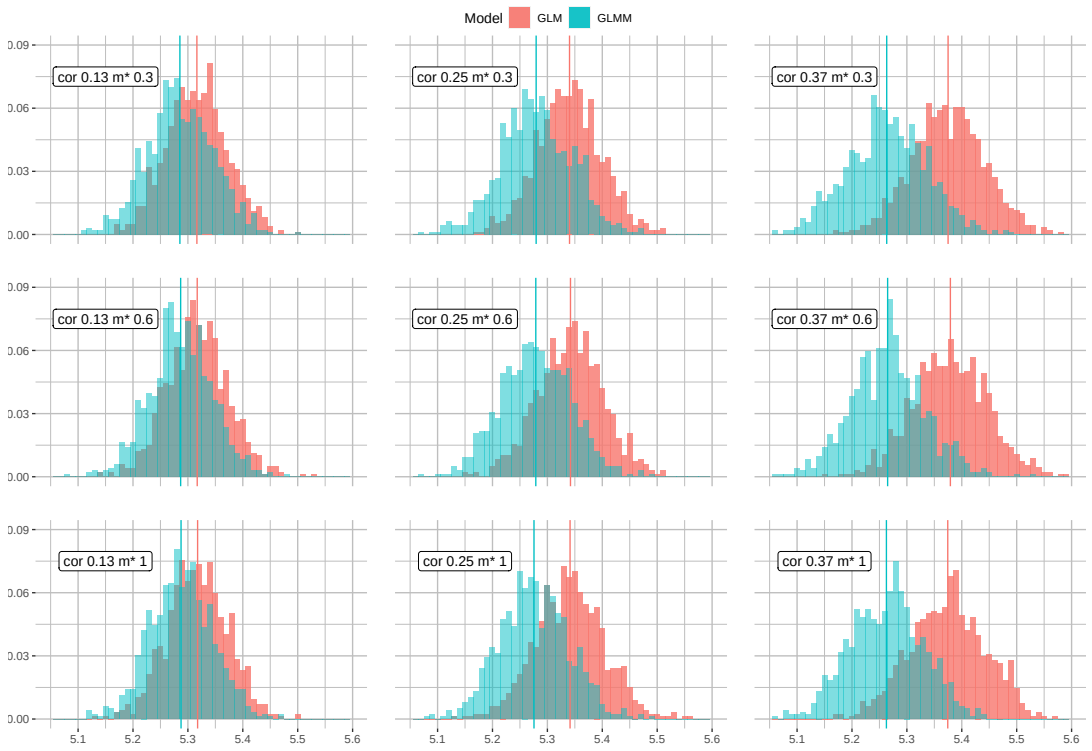



Figure 10: Histogram of the estimated intercepts from simulation for different correlations and proportions of correlated data. The estimated intercepts from the GLM is plotted in red and the GLMM in blue. The vertical lines represents the mean of the estimated intercepts for the two different models.

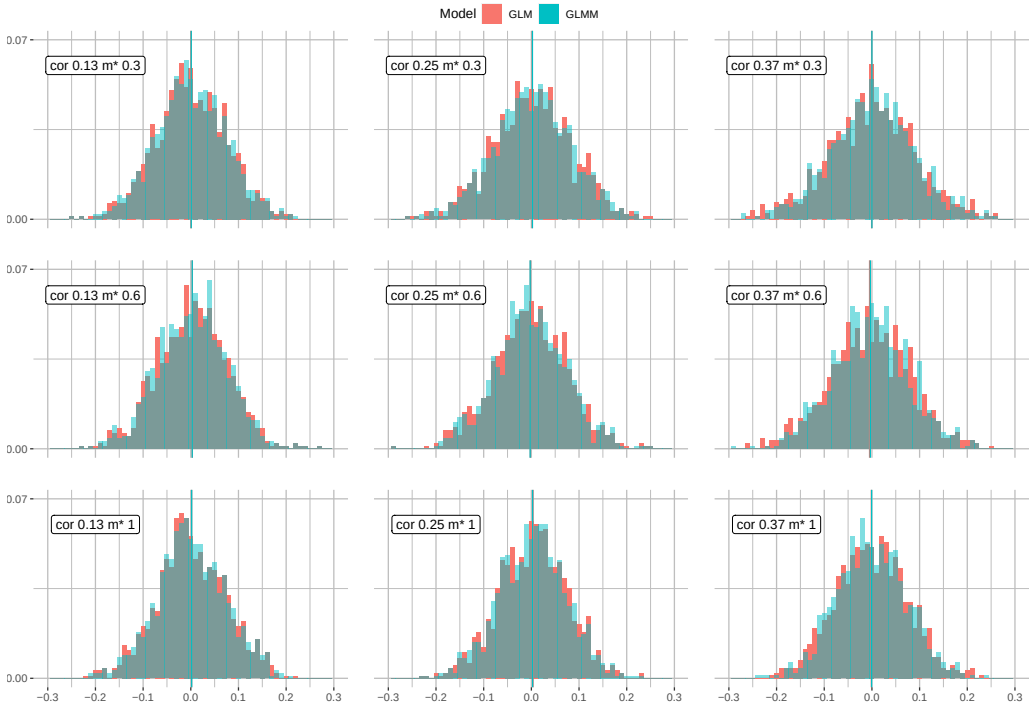



Figure 11: Histogram of the estimated effects from simulation for different correlations and proportions of correlated data. The estimated effects from the GLM is plotted in red and the GLMM in blue. The vertical lines represents the mean of the estimated intercepts for the two different models. These vertical lines are estimated to similar values, so it can be hard to observe that they are both present.

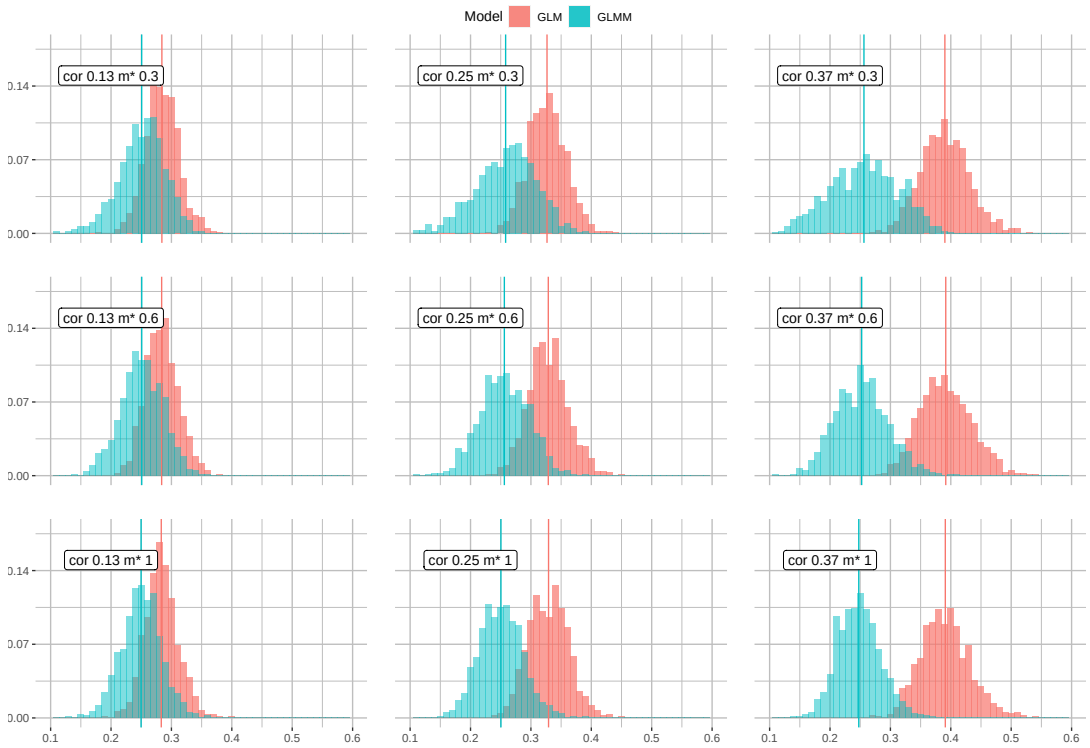



Figure 12: Histogram of the estimated dispersion parameter from simulation for different correlations and proportions of correlated data. The estimated dispersion parameter from the GLM is plotted in red and the GLMM in blue. The vertical lines represents the mean of the estimated dispersion parameters for the two different models.

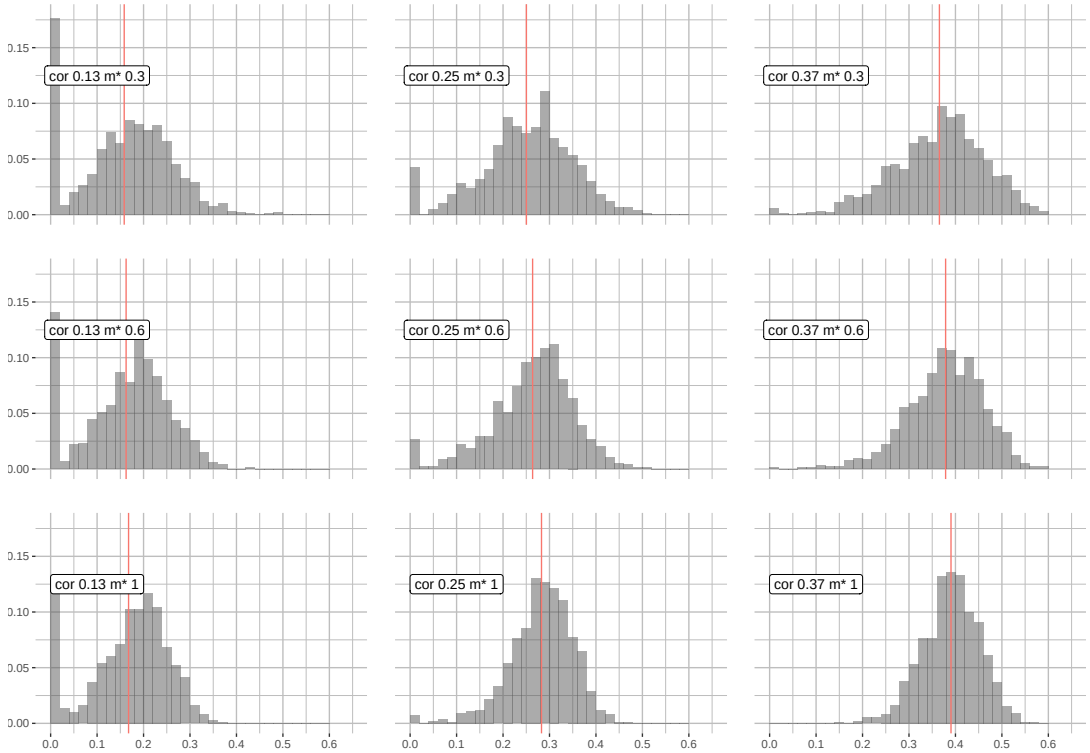



Figure 13: Histogram of the estimated $\tau$ parameter from simulation for different correlations and proportions of correlated data. The vertical line represents the mean of the estimated $\tau$ for the GLMM.

We turn to the dispersion estimates, which can be observed in Figure 12. We commented on expecting the dispersion estimate to increase for GLM with an increase in correlation, which seem to be the case. The mean of dispersion estimates for the GLMM is approximately the value used in simulation, $\alpha = 0.25$, for all the models. Further we can observe that mean of the dispersion estimates seem to be similar for the different levels for proportion of correlated data. But the distribution of the dispersion estimates seems to have more variation for both models the smaller the proportion of correlated data is. There does not seem to be a similar relation for GLM, where both the mean and variation of the estimates seems to be constant when the proportion of correlated data changes.

We also want to examine a plot of the standard deviation of the random effects used in the model, $\tau$, which can be seen in Figure 13. This is only relevant for the GLMM. Similarly to the dispersion estimate for the GLMM, the variance of the estimates seems to decrease with an increase in the proportion of correlated data. For lower correlations there also seem to be a zero inflation of $\tau$, this is probably a convergence issue. The values of $\tau$ used in the simulations were $\tau = \{0.2, 0.3, 0.4\}$, and the mean seem to converge towards the value used for simulation with an increase in the proportion of correlated data.

We want to focus on the proportion of rejected null hypothesis with a significance level of 5%. A plot of the proportion of rejected null hypothesis for all significance levels can be found in Figure 9, this also has the interpretation of emprical CDF for estimated $p$-values. In general GLMM rejects approximately 5% of the hypothesis under the null hypothesis, but for some combination of $m^*$ and correlation, the number is smaller for some and larger for others. GLM generally rejects less than 5% of the hypothesis under the null hypothesis. We provide a confidence interval for the proportion of observed rejected hypothesis for both the GLM and GLMM for each of the the 9 combinations of $m^*$ and correlation for a significance cut-off at 5%. To do this we model the number, $b$, of rejected hypothesis as originating from the binomial distribution, $b \sim Bin(N = 1000, p = 0.05)$. We observe that $E\left[\frac{b}{N}\right] = p$ and $\text{Var}\left[\frac{b}{N}\right] = \frac{p(1-p)}{N}$. Then the boundaries of our confidence interval for proportion of rejected hypothesis under the null hypothesis with a significance level of 5% becomes $\hat{b} \pm 1.96\sqrt{\frac{0.05 \cdot 0.95}{1000}}$. There are four proportions of rejected hypothesis that do not include the true value. Two of them are when correlation is set to 0.37 and $m^* = 0.3$, where both of the models reject too many hypothesis. For GLM the boundaries for the confidence interval becomes $0.0659 \pm 0.0135$, and for GLMM they become $0.0660 \pm 0.0135$. The other two are for the GLM, where the model is conservative, both cases for $m^* = 1$, with correlation 0.27 and 0.37. When the correlation is 0.27 the boundaries of the interval becomes $0.0283 \pm 0.0135$, and when the correlation is 0.37 the boundaries of the interval becomes $0.0295 \pm 0.0135$. Overall GLMM gives a decent fit whilst GLM tends to be conservative when the correlation is high.

We did the same procedure under $H_1$, where the effect was set to $\beta_1 = 2$. Most of the estimation ended up quite similar and plots of these can be found in Appendix B. One thing we want to point out is that both under $H_0$ and $H_1$ the effect, $\beta_1$, was estimated to be similar for both the GLM and GLMM. We hypothesize that this is because the difference between the GLMM and the marginal model that GLM estimates is captured in the intercept and dispersion estimates. This is convenient as the effect estimates will end up similar, but the variance of the effect will vary. This variance will influence if the effect is found to be statistically significant.

With the effect being similar under the null and the alternative hypothesis we hypothesize that the general $p$-values will be lower for GLMM than for GLM. This is generally the case in Figure 14, where we only plot a comparison of $p$-values when the correlation is set to 0.37 and $m^* = 1$. Here 98% of the estimated $p$-values for GLMM are lower than for GLM, under the alternative hypothesis. Since the effects are similar to the ones estimated under the null hypothesis this means the overall effect variance is higher for GLM than GLMM under the alternative hypothesis.

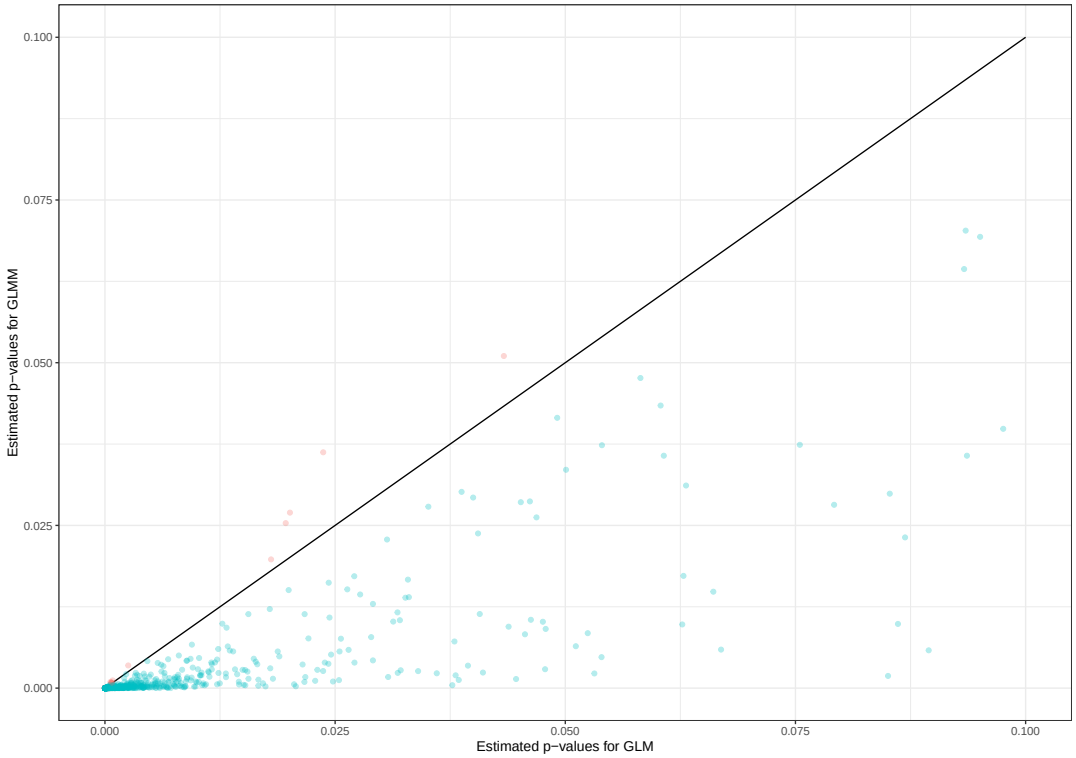



Figure 14: Comparison of $p$-values from GLMM and GLM based on simulated data under the alternate hypothesis. Blue points marks when GLMM estimates a lower $p$-value than GLM, and red marks when GLM estimates a lower $p$-value than GLM. The black line represents where $p$-values would be equal for the GLM and GLMM.

Overall the main difference between a GLM and GLMM fit to data generated from a GLMM with only random intercept is the estimated intercept,$\beta_0$, and the dispersion parameter, $\alpha$. As GLM cannot capture the random intercept it absorbs the effect they have with the link function into the intercept estimate. Further GLM needs to estimate the additional dispersion from the random intercept as part of the dispersion estimate, and this seems to get worse the more of the data that is correlated. We have observed that when correlation is estimated to be low, or when the proportion of correlated data is low GLM and GLMM results in similar fits.

# 6 Gene expression analysis of Crohn's disease data

We now turn to the Crohn's disease dataset presented in Section 2.5. In this section we will study three contrast of interest. We will fit four models to the data introduced, two based on limma-voom, where one takes into account the correlation and one does not, one based on DESeq2 and one GLMM for comparison. All the models except for GLMM estimates a mean-variance relationship, therefore our hypothesis is that the GLMM will not fit the data as well as the other models. A general overview of pipelines for our analysis can be found in Figure 15.

## 6.1 Contrasts of interest

The data presented in Section 2.5 contains gene expression counts from two different types of tissue and three different states of disease, that is 6 groups in total. A general introduction to the groupings and the effects we assigned to each group was given in Section 4, and illustrated in Figure 4. For convenience we will again state that the gene specific vector of coefficients is

$$\boldsymbol{\beta}_g = [\beta_0 \quad \beta_U \quad \beta_A \quad \beta_I \quad \beta_{IU} \quad \beta_{IA}]^T.$$

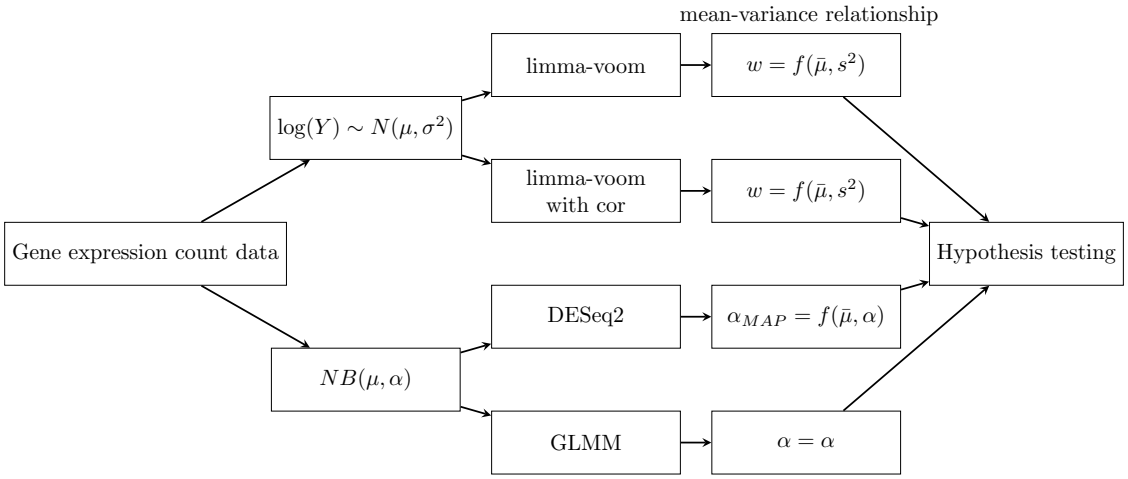



Figure 15: Overview of core assumptions made for the data and pipelines used. Observe that GLMM is the only model that does not have a mean-variance relationship, whilst the other models base it on the average count for each gene. Both limma-voom with correlation and GLMM takes the correlation into account, whilst limma-voom and DESeq2 does not. Both limma-voom and DESeq2 borrows information across genes to estimate a mean-variance relationship, but for our GLMM we use the ML estimate of the dispersion only.

Figure 4 shows the different groups of observations for Crohn's disease and clarifies which coefficients corresponds to the different groups. Observe that both effects of the reference states, healthy and colon, are absorbed into the intercept when using classical design matrices. This is not the case for the extended design matrix, which is used in DESeq2, where they would have their own effects. The extended design matrix is used so that the same regularization is applied to the reference levels as to the other factor levels.

Since Crohn's disease can inflame both colon and ileum we are interested in comparing gene expression between the tissues to observe if the disease behaves differently in different tissues. This will be done by comparing effects between unaffected and affected tissue for one tissue type, and then compare if there is any significant difference between the comparisons for colon and ileum. In other words this is a comparison of a comparison, or a comparison between unaffected and affected tissue correcting for tissue type. This is our main contrast of interest.

Another interesting comparison would be for tissue from healthy patients against unaffected tissue corrected for tissue type from Crohn's disease patients. This comparison could help us observe if there is an underlying difference in tissue with the disease expressed and tissue without. This difference could be explained as either the disease manipulating the surrounding tissue or as a specific gene expression that makes you more prone to the disease. If this is the case we would expect some of the gene expressions connected to Crohn's disease to be more or less active for unaffected tissue than for healthy tissue. What initiates IBD is not known, so our comments about how it could behave is educated speculation (The Crohn's and Colitis Foundation of America (CCFA) 2014). This will be our second contrast of interest.

Our third contrast will be to compare the affected tissue against healthy tissue and compare if there is any difference for colon and ileum. Another way to look at the contrast between unaffected and affected is to first consider two different contrasts where affected is compared to healthy corrected for tissue and unaffected is compared to healthy corrected for tissue before comparing both contrast against each other. In other words using healthy as a reference state in comparisons before comparing the contrast of interest.

Above the contrasts are explained using a sample from the same tissue as reference before comparing across tissues. There is also another interpretation that leads to the same contrast and that is to compare a level of disease, i.e. unaffected, across tissue samples and then compare it to a different level of disease, i.e. healthy or affected. In other words you first find what genes are differently expressed between unhealthy in ileum and in colon. Then find a similar comparison

for a different level of disease, for example healthy, and then compare the two comparisons. This equivalence can easily be observed from writing out the contrasts, $(IU - CU) - (IH - CH) = (IU - IH) - (CU - CH) = \beta_{IU}$.

In short, our three contrasts of interest and corresponding parameters are:

- The difference between $CH$ and $CU$ compared to the difference between $IH$ and $IU$, $(IU - IH) - (CU - CH) = \beta_{IU}$,
- the difference between $CH$ and $CA$ compared to the difference between $IH$ and $IA$, $(IA - IH) - (CA - CH) = \beta_{IA}$, and
- the difference between $CU$ and $CA$ compared to the difference between $IU$ and $IA$, $(IA - IU) - (CA - CU) = \beta_{IA} - \beta_{IU}$.

The different levels and which coefficients is used for different levels can be observed in Figure 4. Our main focus will be to compare unaffected against affected tissue between colon and ileum, $\beta_{IA} - \beta_{IU}$. Figures for the other contrasts can be found in Appendix C. Table of the consensus statistically significant genes for limma-voom taking and not taking into account the correlation, DESeq2 and GLMM can be found in Appendix C.3 for all three contrasts. The R code for gene expression analysis can be found in the file "analysis pipeline.R" in the GitHub repository Ankill (2022).

## 6.2 Preprocessing and quality control

We first examined the distribution of library sizes (total counts recorded for a patient), where our dataset mainly had between 16 and 24 million genes counted, but one had 306740 and was therefore removed. The distribution of library sizes for the data used in the analysis, after this removal, can be observed in Figure 16.

Secondly we want to remove data from genes with little to no information, this is done by removing genes that mainly have observations of zero activity for most individuals. This filtering is done by the `filterByExpr` function in the edgeR library (Smyth 2016). This function decides how many individuals need to have more than a minimum proportion of a gene to be considered for analysis. For our data we require a minimum of 13 observations to have more than 0.48 CPM to be considered for analysis. These values are based on our design matrix and are found by the `filterByExpr` function. This filtering considers 58003 genes, where only 27581 remains after filtering. Filtering generally increases the power of a statistical test, as it influences how many tests will be performed for the multiple testing procedure. It is therefore important that the filtering criterion is independent of the test statistic under the null hypothesis. Otherwise it would invalidate the assumptions for any multiple testing procedure, which we use to find adjusted $p$-values (Bourgon, Gentleman and Huber 2010).

Before fitting a linear model we make a MDS (multidimensional scaling) plot, with the aim to perform quality check of the data. This plot is implemented by both limma and DESeq2, and gives a two-dimensional overview of similarities between higher dimensional data (Wickelmaier 2003). The plot can be observed in Figure 17. From this plot we can observe a separation between observations between different tissues in the leading log fold change dimension. There does not seem to be a large difference between unaffected and healthy for both tissues in the second leading fold change dimension. There does however seem to be a gap between the group of healthy and unaffected, and affected. For ileum we can also observe that the unaffected and healthy groups have an overlap with affected, something that is not the case for colon. From this plot alone we believe that there will be few statistically significant gene differences between unaffected and healthy, and many when comparing affected to any of the two other states when correcting for tissue.

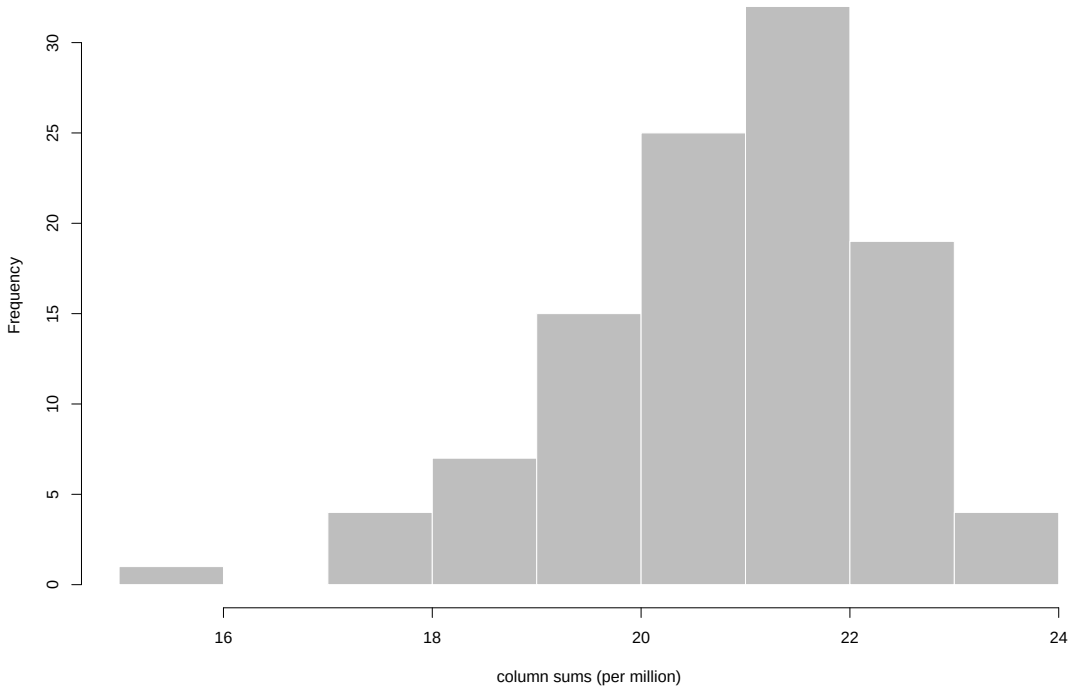



Figure 16: Histogram of the sum of counts across all genes, library sizes, for different patients.

We have previously commented on a MDS plot of the same dataset (Ankill 2021), where two outliers were found. After consulting the co-supervisor about the two outliers their origin was found to be questionable and they removed from the dataset to not produce unnecessary noise. The two observations are therefore not present in this MDS plot, nor in the presentation of data in Section 2.5.

In Figure 18 we show a heatmap of a set of genes. How these genes were selected is explained in detail here in Section 6. The heatmap can help us see genes and samples which behave in a similar fashion. The clustering distance for both genes and samples is found by correlation, and the clustering was done by the `pheatmap` function from the pheatmap package. We observe that there are many similarities for clusters across samples, which tells us that multiple genes have similar relations for gene expression. This could be an indication that the genes are part of a larger biological process, or it could be random chance.

## 6.3 Mean-variance relationship

For limma-voom the relationship between the average expression of an observation and the weights can be observed in Figure 19. Correlation is estimated after the fitting of the weights, so the weights are the same for the model that takes and does not take correlation into consideration. We observe that there are few values with low average log expression and standard deviation, which tells us that our filtering was sufficient or too strict, otherwise a lot of observations with low average log expression and standard deviation would be present. As mentioned in Section 4.1.3 we expect the squared coefficient of variation to roughly follow $\frac{1}{\mu} + \phi$. The gene-wise variance estimates seem to roughly follow a mean-variance trend, which is captured well by the fitted curve. As a consequence we believe the empirical Bayes procedure with the mean-variance trend improves our estimates.

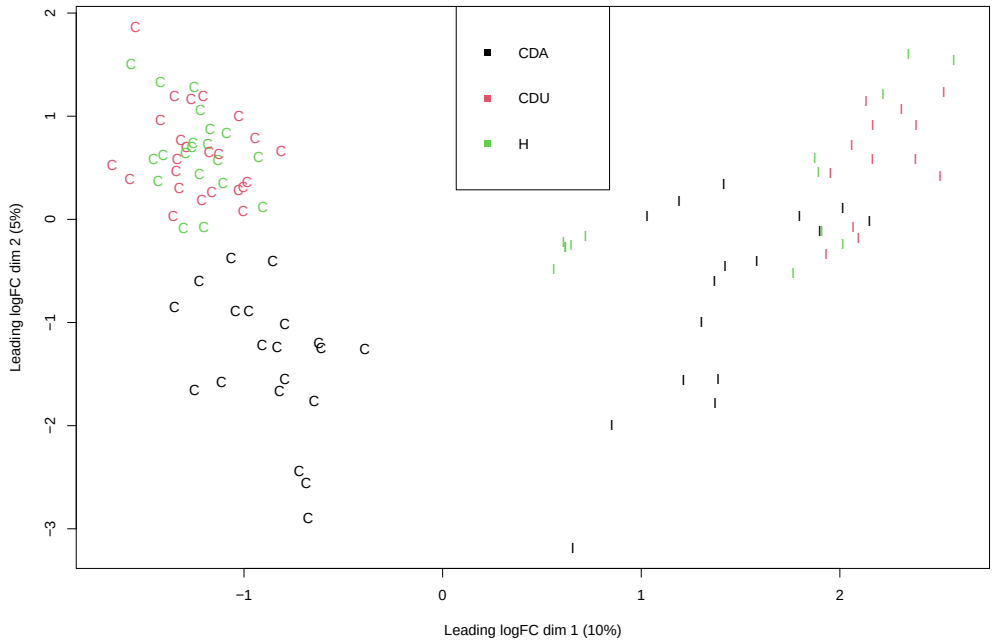



Figure 17: Plot of samples in multidimensional scaling, I denotes that the sample is from ileum and C denotes that it is from colon. In the legend the color used is explained, where black is used for affected, red is used for unaffected and green is used for healthy. The abbreviation CDA is Crohn's disease (CD) in combination with affected tissue (A) and CDU is CD in combination with unaffected tissue (U). Both limma-voom and DESeq2 does exploratory analysis on the log scale, where DESeq2 uses a regularized log transformation of the data.

The variance not accounted for by the weights for the model taking the correlation into account can be observed in Figure 20 and Figure 21 for the model that does not take the correlation into account. This measurement is mentioned for weighted least squares in Section 3.2.3. It should be noted that the two figures have few differences, which should make their results similar. Ideally the blue line would be at exactly 1 and the residuals should be distributed around the blue line. We can observe that the line is close to 1, which tells us that the assumptions for our weights are approximately fulfilled. The residuals seem to have a higher spread for higher average log-expressions, but they seem to be mostly centered around the mean line. There might be a slight trend of decreasing variance for an increase in average log-expression, but compared to Figure 19 it a massive improvement.

For DESeq2 the mean-variance relationship can be found in Figure 22. This plot includes the gene-wise estimates, the mean-variance trend and the MAP estimate for the dispersion. We expect the trend to follow $\alpha_{tr}(\bar{\mu}) = \frac{\alpha_1}{\bar{\mu}} + \alpha_0$, which seems to be satisfied. We observe that the MAP dispersion estimate lies between the gene estimate and the mean-variance line, which is what we expected. There are also some gene estimates that are large enough to be considered outliers and do not use the mean-variance relationship for the final estimate, these are circled in blue around the gene-wise estimate. From this plot alone the majority of the gene-wise dispersion estimates seem to roughly follow the estimated mean-variance relationship. From this plot alone we would consider the underlying assumptions of the mean-variance relationship for DESeq2 to be fulfilled. We have chosen in this plot to only include 1000 randomly chosen genes, as it is hard to interpret a plot consisting of three different levels for 27581 different genes.

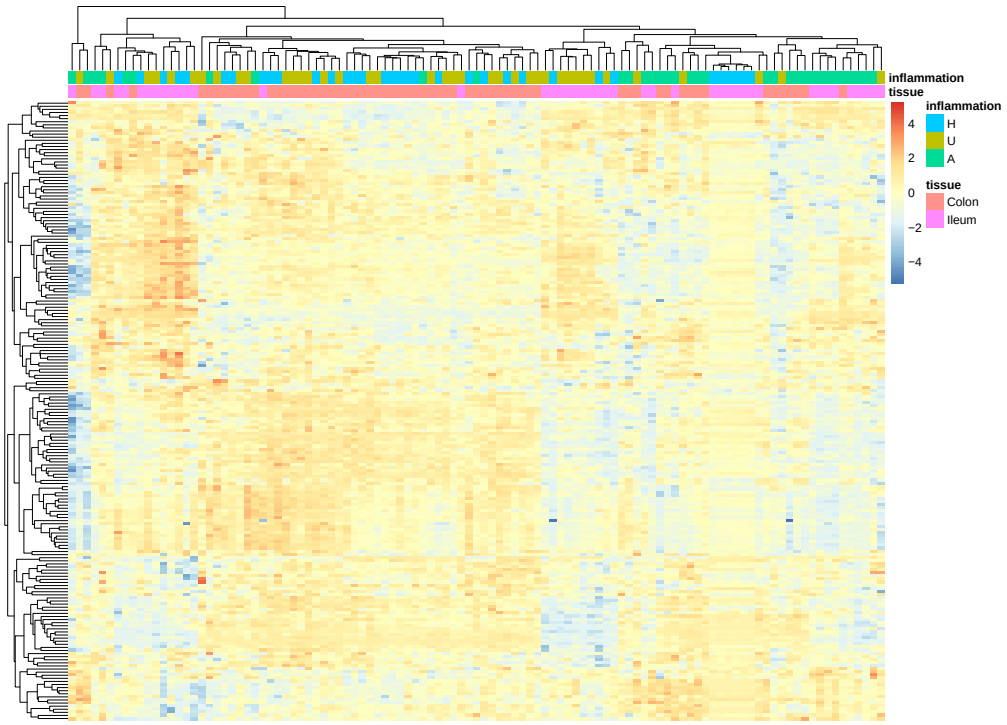



Figure 18: Heatmap for the 200 most statistically significant genes for DESeq2 for the contrast $\beta_{IA} - \beta_{IU}$. The rows represents different genes and the columns represents different samples. No names are given in this plot, as it would only add unnecessary information. The rows measure the difference from the mean scaled with respect to standard deviation for the gene. The clustering is done with respect to correlation for both rows and columns.

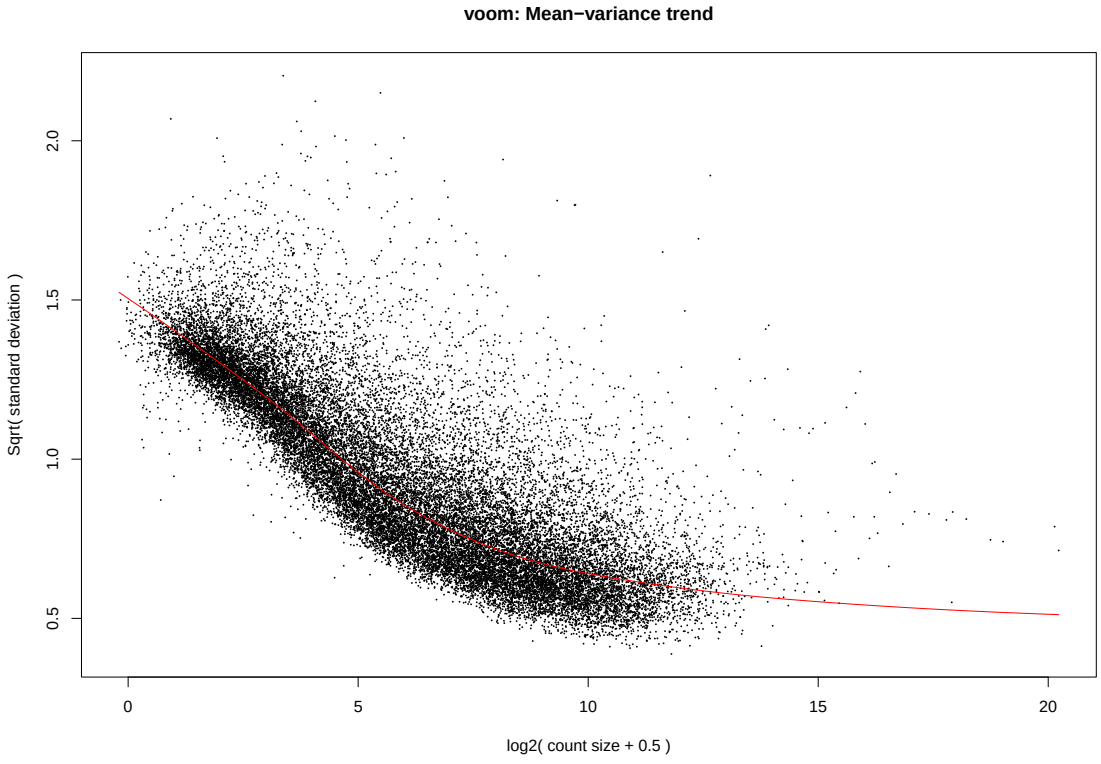



Figure 19: Plot for mean-variance trend for limma-voom, where the square root of the standard deviation is plotted against the average log counts for genes, used to estimate observation weights.

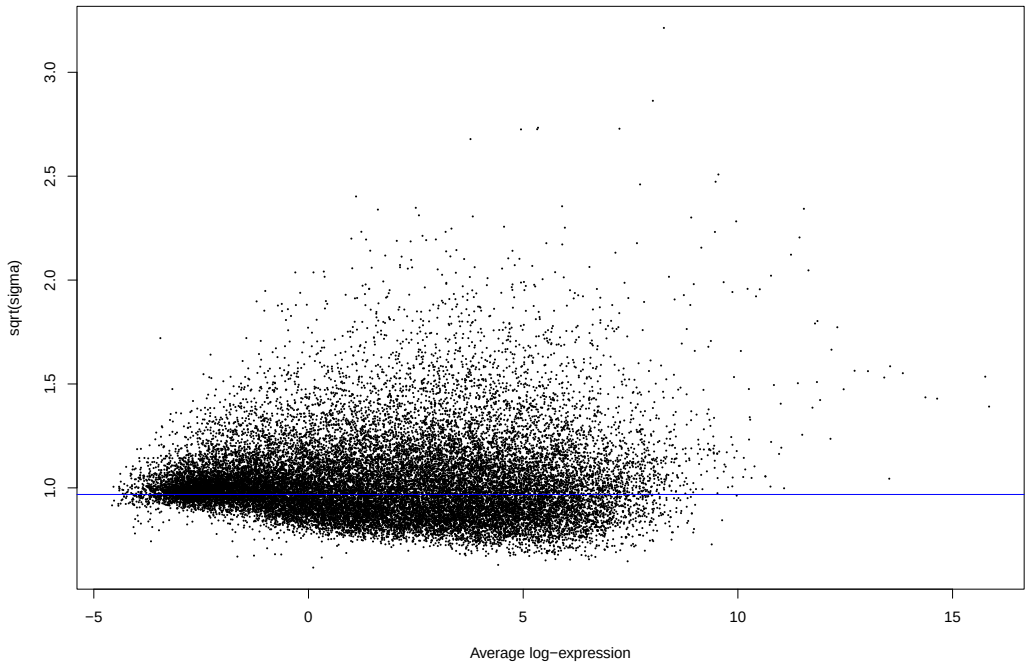



Figure 20: Plot of the variance not accounted for in weights, after modeling the mean-variance relationship for limma-voom when accounting for correlation. The black points represents the square root of the standard deviation for a gene after weighted transformation, and the blue line is the average across all genes.

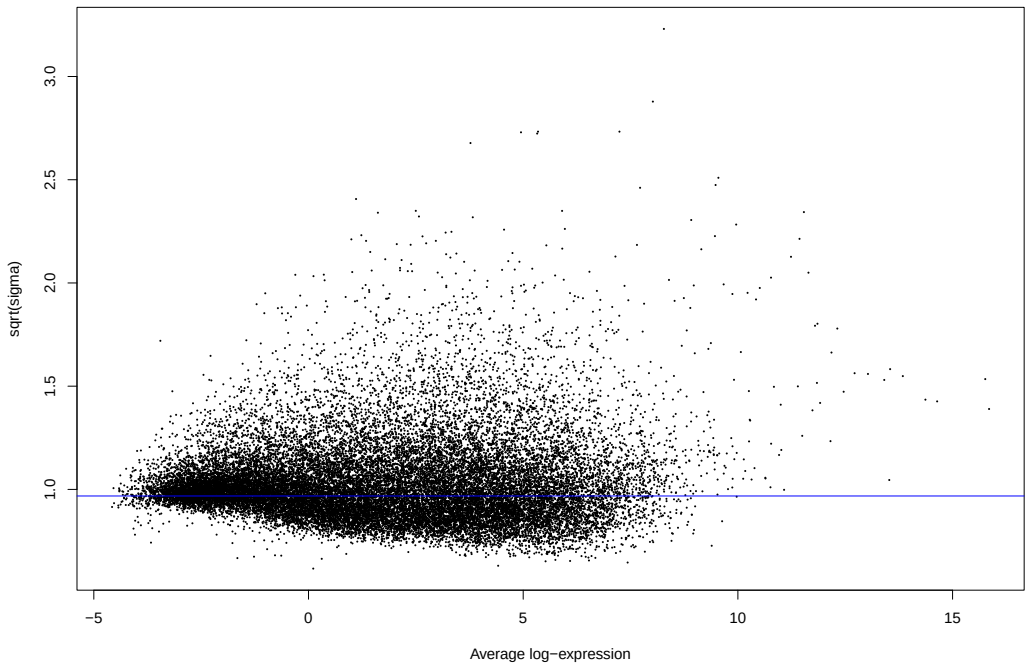



Figure 21: Plot of the variance not accounted for in weights, after modeling the mean-variance relationship for limma-voom when not modeling correlation. The black points represents the square root of the standard deviation for a gene after weighted transformation, and the blue line is the average across all genes.

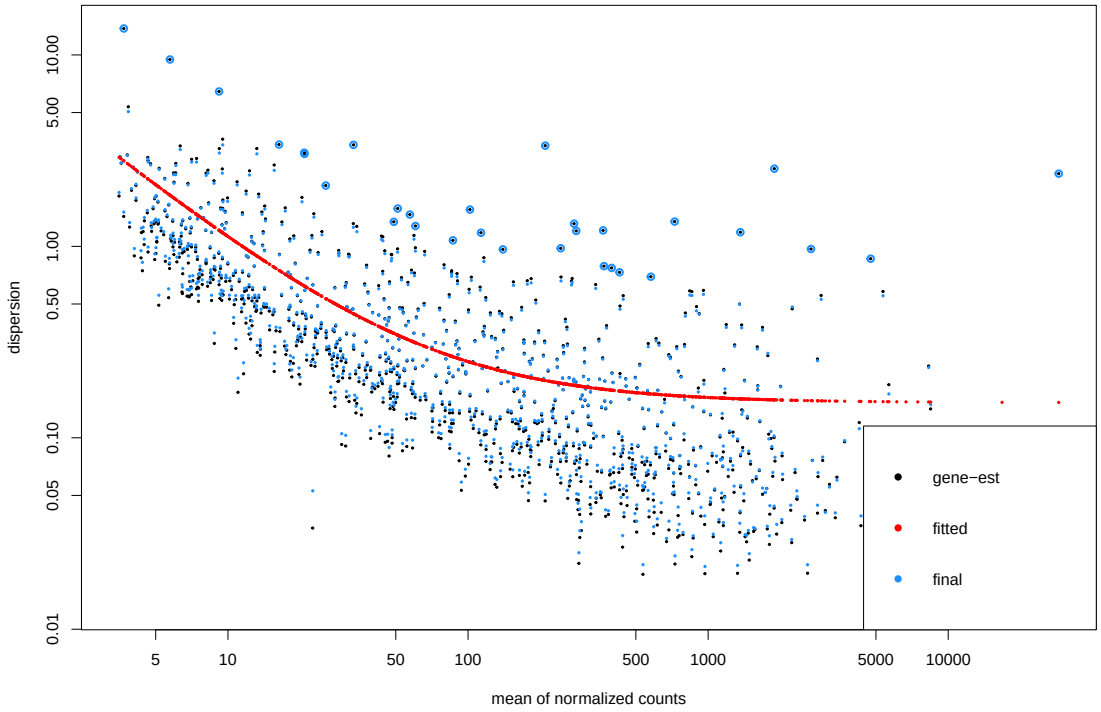



Figure 22: Plot of mean-variance trend for 1000 random genes used in DESeq2, where the dispersion parameter, $\alpha_{GLMM}$, is plotted against the mean of normalized counts. Black markers represent the gene-wise estimates, the red markers represents the overall mean-variance trend, and the blue marker is the final MAP estimate used further in the pipeline. When the gene-wise estimate has a blue circle around it the gene-wise estimate is considered an outlier and used instead of the MAP estimate.

A plot of the mean values against the dispersion estimate for GLMM can be found in Figure 23. This is a similar plot to Figure 22, and if uncorrelated count data has an underlying mean-variance relationship it is reasonable to assume that correlated count data also has it. In other words we expect to see a trend for the dispersion estimates similar to the observed from DESeq2, which is not taken advantage of in our model. In the figure there seems to be an underlying trend, where the dispersion parameter decreases with an increase of the mean of normalized counts. We can observe that in the middle of the figure there seems to be a more dense cloud of observations. We suspect that our model would have been improved by modeling a mean-variance relationship, but the underlying trend is not as obvious as it is for limma-voom.

All four models seems to have an observable relation of between the variance/dispersion gene-wise estimates. Both limma-voom models and DESeq2 tries to capture this trend to improve the models, whilst GLMM does not. Limma-voom and DESeq2 seems to mostly capture the trend of the data. Limma-voom plots the distribution of the variance after accounting for the mean-variance trend, where the fit is improved by the relationship. The estimated trends for limma-voom and DESeq2 were similar to the trends generally observed for gene expression count data.

## 6.4 Correlation

For limma-voom the consensus correlation across genes is estimated to 0.098, where only 30 out of 107 observations are correlated. This correlation was found from Equation (9). In Figure 24 we can observe the distribution of the correlation through its inverse hyperbolic tangent values.

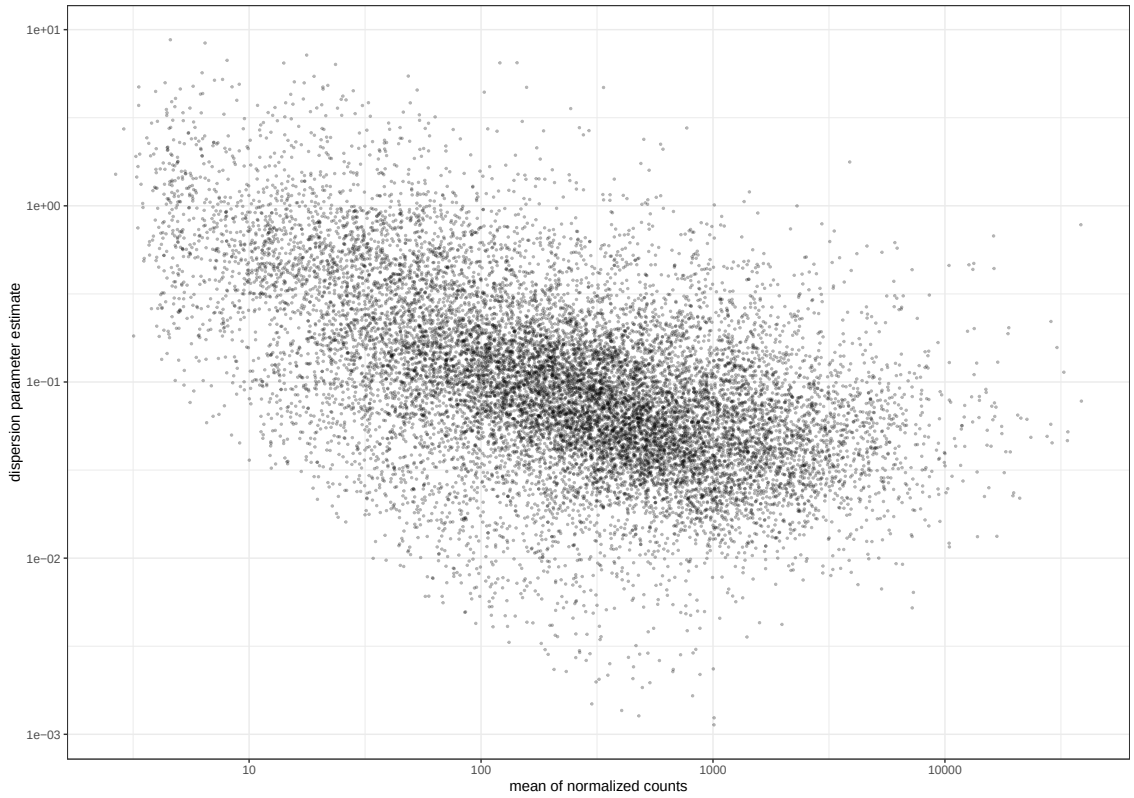



Figure 23: Plot of mean-variance relationship for GLMM, where the dispersion parameter $\alpha$ is plotted against the mean of normalized counts. Both the $x$- and $y$-axis are shown on the log scale.

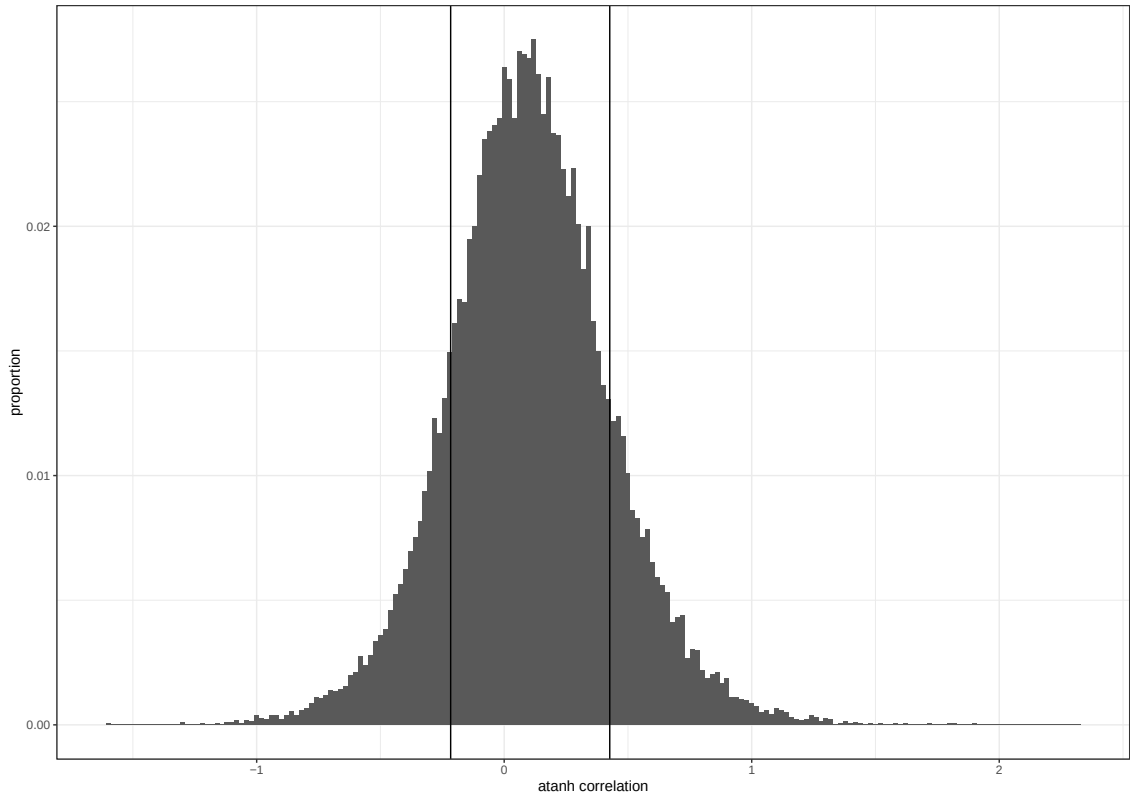



Figure 24: Arctanh correlation for limma-voom with vertical lines marking the outer 30% of the values. Observations in the outer 30% of the values are not consider for the trimmed mean.

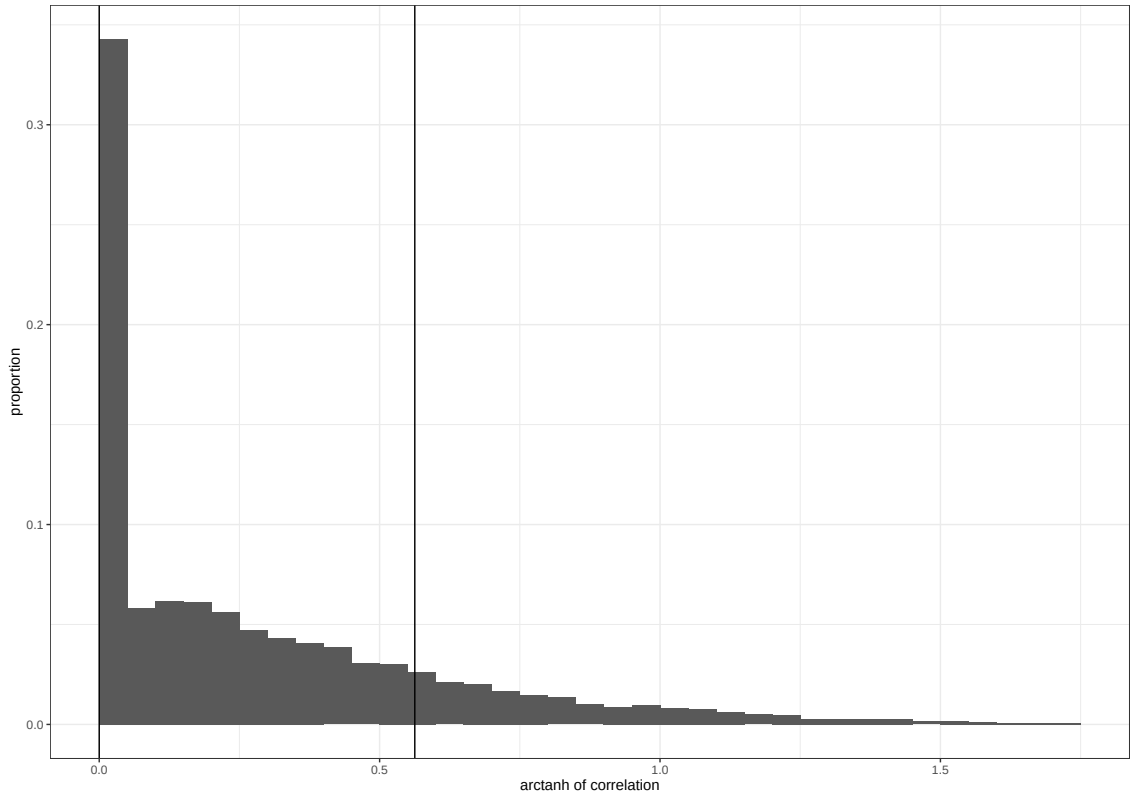



Figure 25: Arctanh correlation for GLMM with vertical lines marking the outer 30% of the values. Observations in the outer 30% of the values are not consider for the trimmed mean.

For GLMM a consensus correlation across genes is estimated to 0.21. This consensus correlation was found from Equation (9), where the correlations was found from Equation (14). If limma-voom had used the LMM procedure mentioned for estimation of correlation for a normal distribution, then it would also only be non-negative. From Figure 25 we can observe 0 correlation for more than 30% of the responses, GLMM in this case defaulted to a GLM model (the variance of the random effects were estimated to approximately 0). The GLMM correlation can only be positive, unlike the procedure used on limma-voom, as a consequence the correlation for GLMM is expected to be higher than for limma-voom. Intuitively negative correlation does not make sense for observations from the same patient, but multiple gene-wise correlation factors were estimated as negative for limma-voom, which can be observed in Figure 24. Limma-voom estimates a consensus correlation and uses it for its gene-wise effect estimate. This is not the case for GLMM, where each gene finds its own correlation and the consensus correlation is created after the models have been fit and is more to present the overall dependency within the models.

There were some of the GLMM models that did not converge. DESeq2 avoids convergence problems by a prior on the effects, as mentioned in Section 4.2.4. This makes the iterative scheme converge faster when the Fisher information is close to singular. We should mention that when the Fisher information is close to singular we do not expect any effects to be relevant, but removing the non convergent genes from the multiple testing procedure would have an impact on our multiple testing procedure. Therefore we chose to assign a $p$-value of 1 to all genes where the model did not converge.

To get a better idea of similarities between the two correlation estimates we wanted to compare the proportion of correlation that was estimated to be less than 0.05 for both limma-voom and GLMM. Limma-voom estimated 0.44 of the correlations to be less than 0.05, whilst GLMM estimated 0.34 of the correlation to be less than 0.05. This could result from GLMM generally estimating higher correlations, or GLMM having convergence problems given that negative correlation exist between two obs from the same individual for the same gene.

## 6.5 Model fit

In Figure 26, 27, 28 and 29 we plot the gene specific mean after normalization, on the $x$-axis, against the $\log_2$ fold change in the contrast of interest, on the $y$-axis. For limma-voom the gene specific mean after normalization corresponds the average log expression, $\tilde{y}_g$, the gene specific average of the log-CPM found in Equation (7). For DESeq2 and GLMM the gene specific mean after normalization corresponds to $\bar{\mu}$, which is given in Equation (10). This type of plot is called a mean average (MA) plot, and is directly implemented in both limma and DESeq2. Both axis are log transformed, the $y$-axis in the form of the log fold change and the $x$-axis in the form of powers of 10. Under $H_0$ we would expect the log fold change to be centered around 0, with a variance reducing with an increase in average log-expression/mean of normalized counts. For most higher counts less log fold change is required for a gene to be considered statistically significant. This is a consequence of only highly expressed genes having enough information to express a statistically significant difference when the difference is small. This relation is not absolute, as the mean-variance relationship is part of a mixture for the final gene-wise variance estimate. This seems to be satisfied in general for the MA plots presented here and in Appendix C.1. Note that this plot shows all genes at once, and can therefore not be used to say something about the model for a specific gene. If we wanted to check the assumptions for a gene specific model we could use residual analysis to check whether the model assumptions are fulfilled or not.

We want to do residual analysis for one gene in Figure 30. The gene the analysis was done on was chosen by the lowest sum of adjusted $p$-values for all four models for the main contrast of interest, $\beta_{IA} - \beta_{IU}$. This corresponds to gene $SLC16A1$. Here we observe a Q-Q plot for the different models, this can be used to assess the fit of the model. The residuals for DESeq2 and GLMM were found by using the quantile residuals as explained in Section 3.5. As both DESeq2 and GLMM uses normalization factors the normalized count data is not necessarily integers anymore, and were transformed to be integers by the `round` function in base R. The DESeq2 and GLMM looks to have great fits, while both limma-voom models seem to be slightly light tailed. Light tailed means that the distribution has less extreme observed responses than expected. We should not read too much into this tail as the model will still give a decent fit, but we do trust DESeq2 and GLMM to give a better fit for this gene. Note that this residual analysis is only for one of 27581 genes that was not filtered, and does not necessarily represent the fit for other genes.

It should be mentioned that for some genes GLMM did not converge. One reason for this could be that the Fisher information matrix is close to singular. One way to solve this would be to consider a regularization technique, for example ridge regression, which is something DESeq2 does for effect estimation. This is a reason for why we should not blindly trust the results from the GLMM.

## 6.6 Significance tests

We want to analyze the distribution of $p$-values together with the analyzed MA plots to give an idea of the overall fit for the different pipelines. First we look at limma-voom, where Figure 31 shows the distribution of estimated $p$-values for all genes sent through the pipeline. We can observe that the resulting distributions of $p$-values are quite similar, and the resulting distribution looks like a mixture of a uniform distribution and an exponential distribution. An underlying uniform distribution is desired.

Figure 32 shows the distribution of estimated $p$-values for DESeq2. This plot is quite similar to the one for limma-voom with and without accounting for correlation, but the proportion of $p$-values that is between 0 and 5% is significantly higher for DESeq2 and GLMM. We can observe that both models with correlation has a higher proportion of $p$-values smaller than 5% compared to the model that makes the same distributional assumptions without correlation.

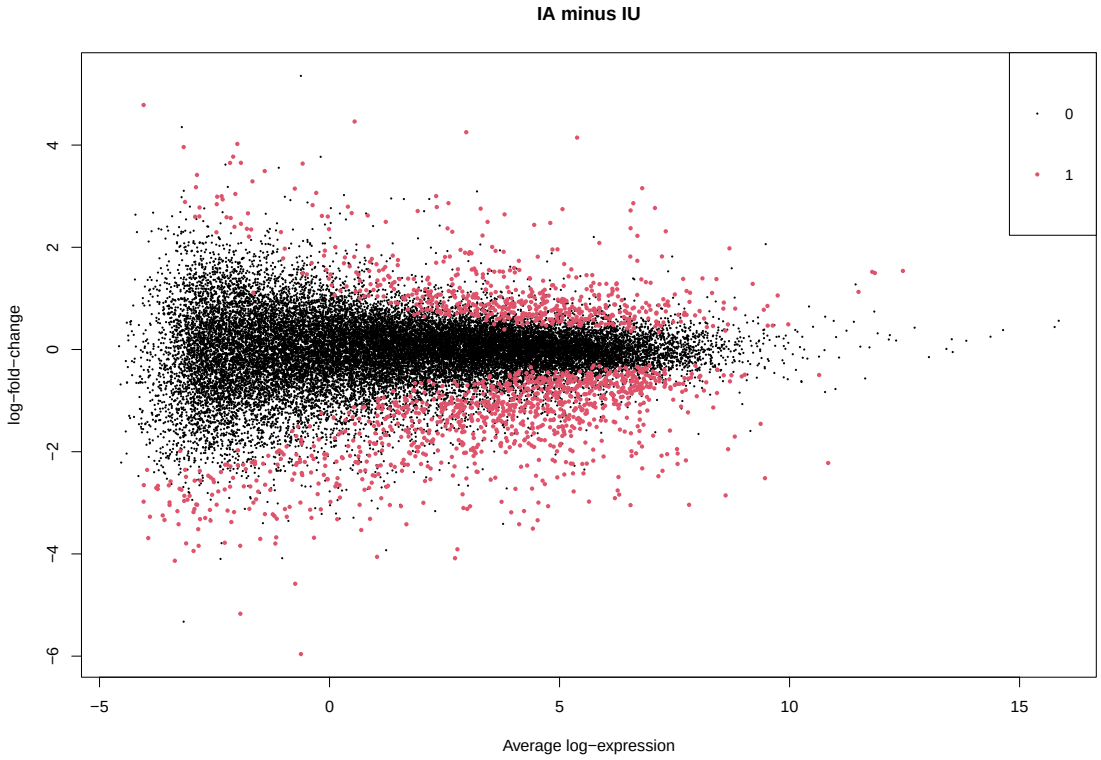



Figure 26: MA plot for limma-voom when not accounting for correlation for the contrast $\beta_{IA} - \beta_{IU}$. This plot was generated using the `plotMA` function in the limma package.

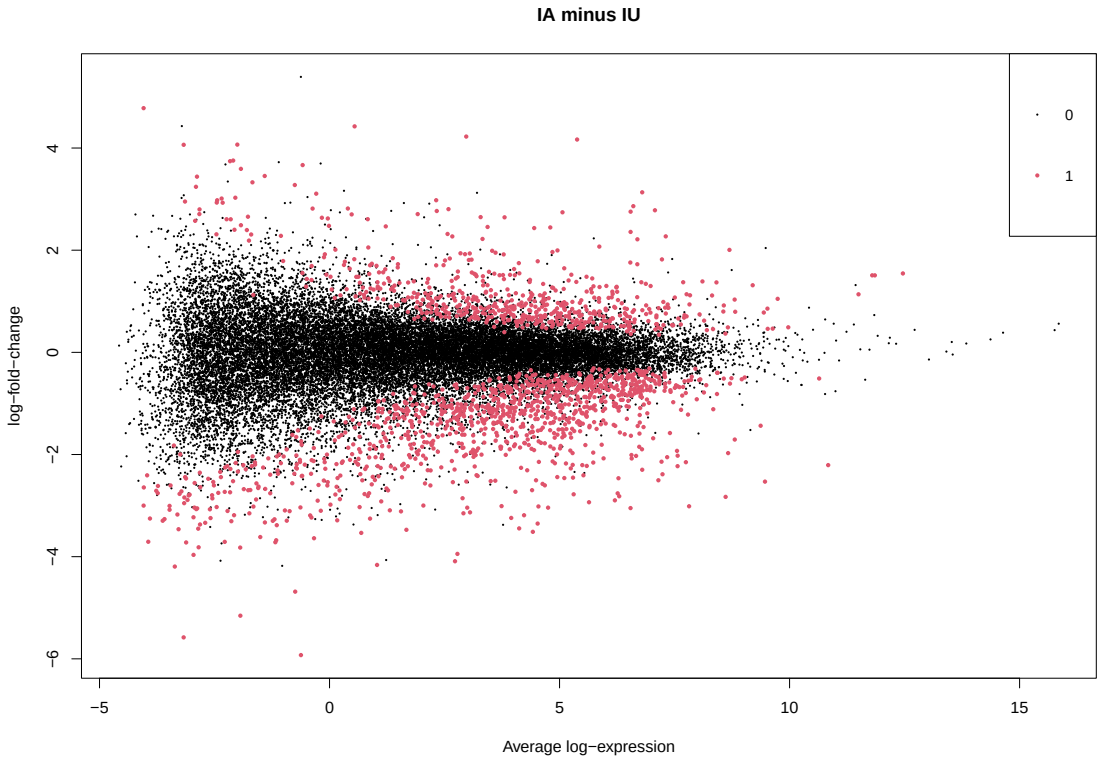



Figure 27: MA plot for limma-voom when accounting for correlation for the contrast $\beta_{IA} - \beta_{IU}$. This plot was generated using the `plotMA` function in the limma package.

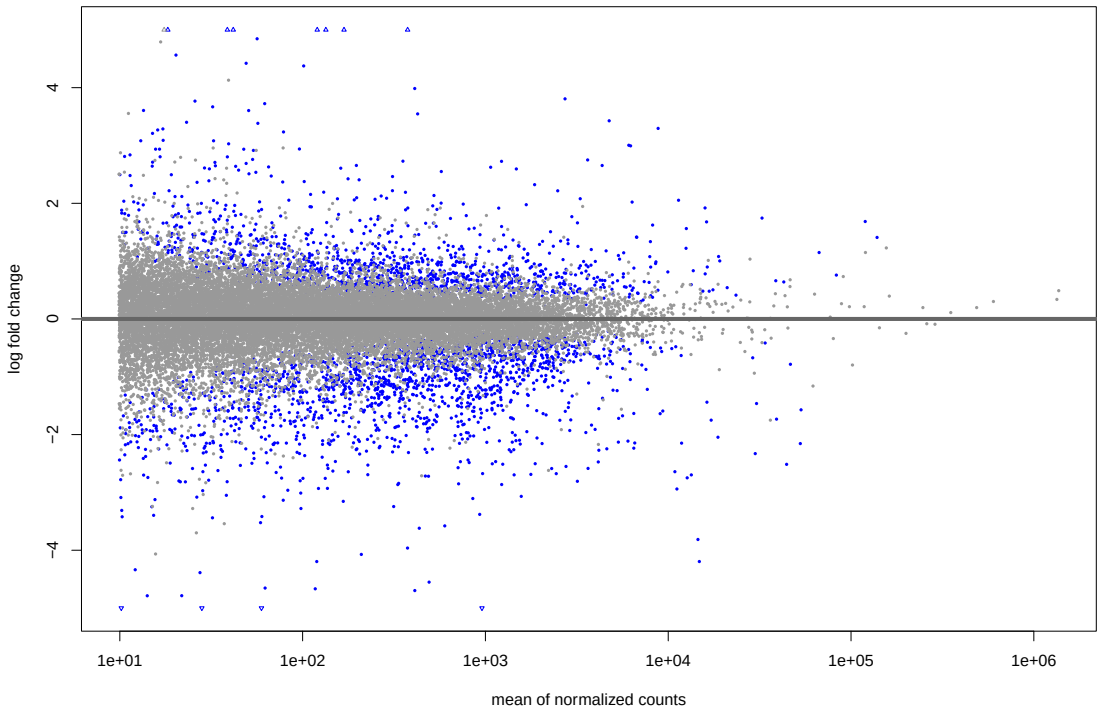



Figure 28: MA plot for DESeq2 for the contrast $\beta_{IA} - \beta_{IU}$. This plot was generated using the `plotMA` function in the DESeq2 package.


Figure 29: MA plot for GLMM for the contrast $\beta_{IA} - \beta_{IU}$. R code is found in Ankill (2022).

Now that the models fitted the data have been observed to be decent, but not perfect, we want to compare the final results across the models. One way this can be done is to create a Venn diagram for all statistically significant genes and compare which genes are statistically significant using a FDR-cutoff of 0.05 between the different models. For further analysis we would be most comfortable with examining the consensus agreement between the models. This Venn diagram can be observed in Figure 33. We see that both DESeq2 and our GLMM find a lot more genes to be statistically significant compared to the other pipelines. It looks like there are more similarities between the underlying distribution assumptions, normal of log counts for limma-voom and negative binomial for DESeq2 and GLMM, than for the models that do and do not account for correlation. If we had only considered the established pipeline, not regarded the GLMM results as valid, we would have found 1413 common statistically significant genes for the contrast $\beta_{IA} - \beta_{IU}$ $(1015 + 398)$. We should comment that 1413 statistically significant genes for a specific contrast with the BH-procedure, as that represents 5% of the genes we consider after initial filtering.

Similar genes are found to be relevant for the effect $\beta_{IA}$, see Figure 51. But there are overall fewer genes that are found to be relevant. In Figure 50 one to two genes are found to be relevant for limma-voom, whilst DESeq2 and GLMM both find more than 100 to be relevant. This is not a good sign for DESeq2 and GLMM, if we believe unaffected and healthy tissue to be similar for most genes. If our intuition is correct then DESeq2 and GLMM gives lower $p$-values than they should.

The top 20 genes of our gene ranking list, where the ranking is based on the sum of adjusted $p$-values for all four models, can be found in Table 5. For most of the shown genes the LFC estimates are similar for the models that use the same distributional assumptions. There is a lot of variation in the base mean in the gene ranking, and that the majority of found genes have a negative effect. Multiple of the genes found have previously been found to be a potentially relevant gene for IBD, especially the GLA genes (Ahmad, Marshall and Jewell 2006).

When comparing Table 5 against Table 6 and 7 in Appendix C.3 we find one gene in the top list for both $\beta_{IA} - \beta_{IU}$ and $\beta_{IU}$, $ANO6$. The top list for $\beta_{IA} - \beta_{IU}$ and $\beta_{IA}$ have 6 genes in common. Initially we believed there is little difference between unaffected and healthy tissue samples. This is reflected in the fact that only one gene is shared in the top list with $\beta_{IA} - \beta_{IU}$ and the number of genes believed to be statistically significant in Figure 50 in Appendix C.2. We anticipated more similarities between the top lists of $\beta_{IA} - \beta_{IU}$ and $\beta_{IA}$, so there there needs to be more of a difference between variance or effect estimates for healthy and unaffected corrected for tissue than expected. The gene we did residual analysis on, $SLC16A1$, is in the top list for both $\beta_{IA} - \beta_{IU}$ and $\beta_{IA}$.

Throughout this chapter we have discussed potential strengths and weaknesses for all of our models. We have seen that the fit for the different models seem to be good both in general and for one specific gene. The genes found to be statistically significant for all of our models will be a robust set of genes that could be used for further research. The top 20 of these can be seen in Table 5.

Table 5: Top 20 genes across models for the Crohn's data set (based on ranking by the sum of the adjusted $p$-value for all models) for the contrast $\beta_{IA} - \beta_{IU}$.

| names | base.mean | voom.LFC | voom.cor.LFC | DESeq2.LFC | GLMM.LFC |
|---|---|---|---|---|---|
| SLC16A1 | 1068 | 2.746 | 2.742 | 2.624 | 2.624 |
| PMP22 | 3642 | -2.758 | -2.76 | -2.585 | -2.585 |
| HLA-DMA | 1469 | -2.175 | -2.157 | -2.105 | -2.1 |
| HLA-DRA | 13377 | -2.855 | -2.829 | -2.697 | -2.699 |
| RHOBTB2 | 1137 | -2.445 | -2.461 | -2.467 | -2.467 |
| CREB3L2 | 4990 | -1.23 | -1.218 | -1.232 | -1.203 |
| HLA-DMB | 730 | -2.722 | -2.698 | -2.468 | -2.48 |
| SLC7A7 | 765 | -1.907 | -1.914 | -1.821 | -2.04 |
| AQP3 | 2440 | -2.778 | -2.779 | -2.676 | -2.712 |
| MDK | 2893 | -2.031 | -2.028 | -1.929 | -1.931 |
| ALDOC | 709 | -2.727 | -2.729 | -2.573 | -2.584 |
| ZNF512B | 420 | -1.793 | -1.802 | -1.861 | -1.858 |
| ANO6 | 952 | -1.655 | -1.671 | -1.628 | -1.701 |
| SLC6A20 | 1571 | -3.341 | -3.352 | -3.068 | -3.093 |
| HLA-DRB1 | 1773 | -2.976 | -2.934 | -2.688 | -2.688 |
| AGT | 227 | -2.834 | -2.856 | -2.528 | -2.633 |
| PDE6A | 184 | 1.57 | 1.582 | 1.536 | 1.687 |
| SUCLG2 | 377 | 0.935 | 0.936 | 0.967 | 0.966 |
| CD2 | 318 | -2.549 | -2.52 | -2.204 | -2.277 |
| SGMS1 | 413 | -1.136 | -1.143 | -1.117 | -1.117 |

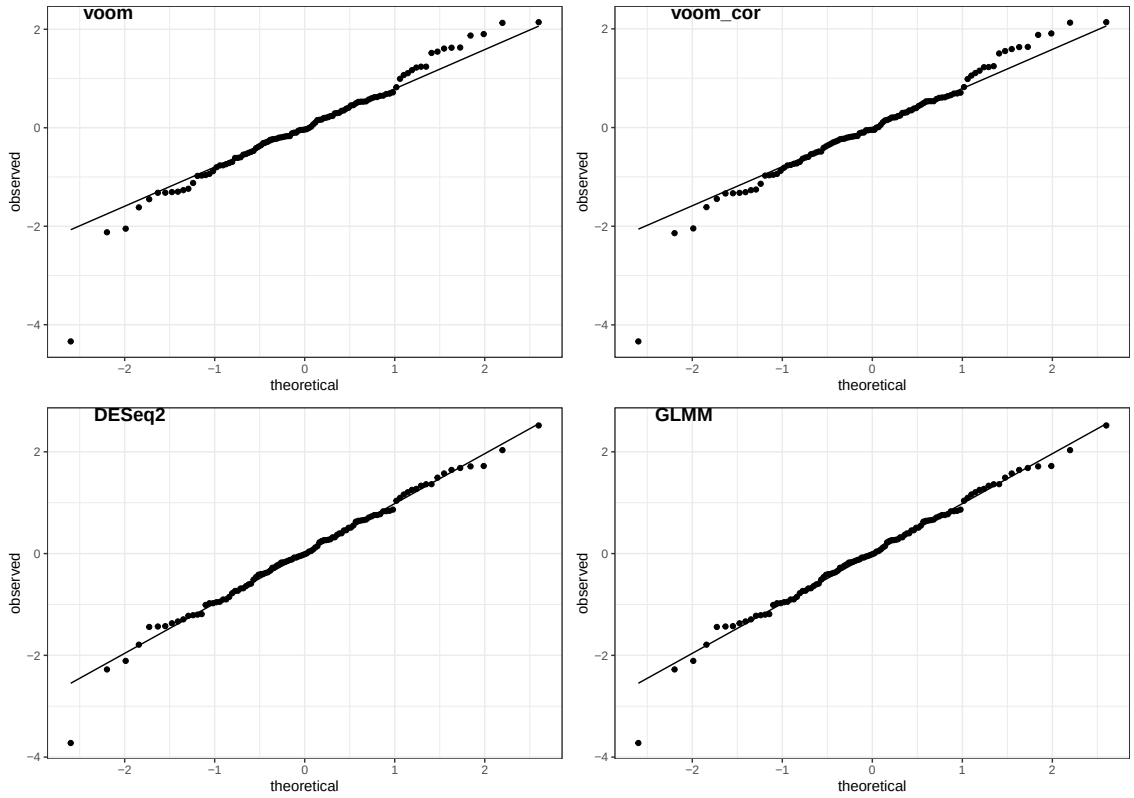



Figure 30: Q-Q Plot of the residuals for all models. Quantile residuals were used for observed standardized normal residuals for DESeq2 and GLMM. Theoretical residuals are the corresponding expected values for a standard normal variable.

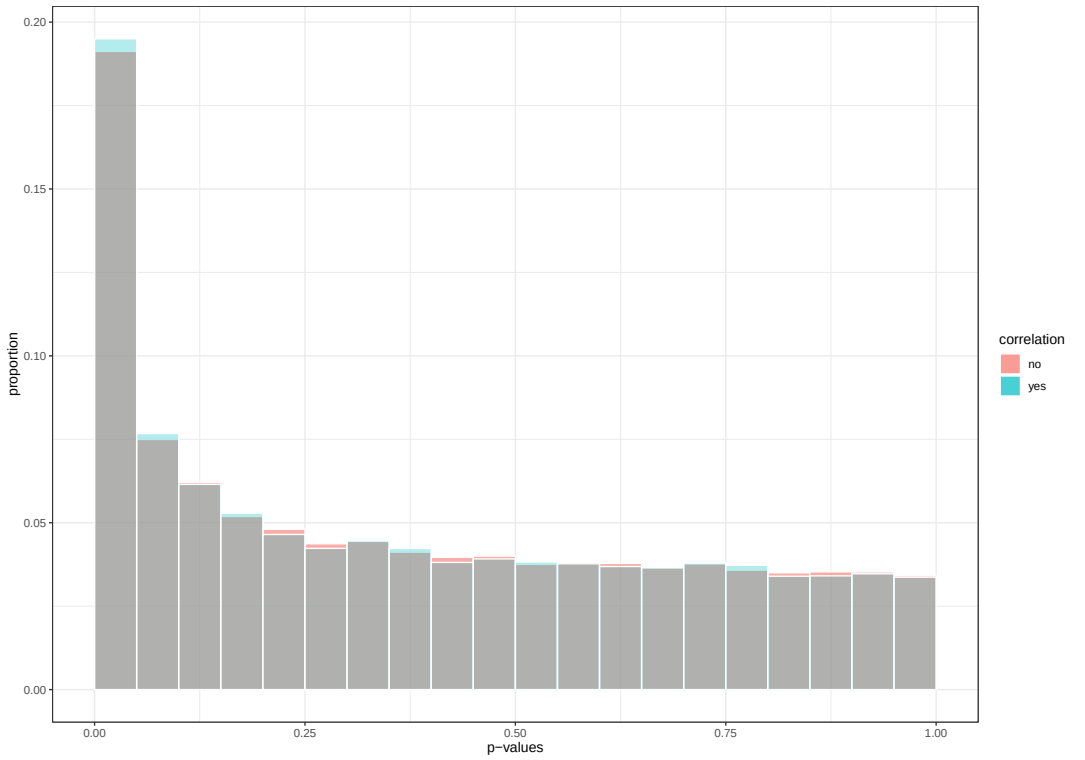



Figure 31: $P$-values for the contrast $\beta_{IA} - \beta_{IU}$ from limma-voom with and without assuming the observations to be correlated. The $p$-values that do not account for the correlation are plotted in red, whilst the ones that do are plotted in blue, and when both colors are present the mixture becomes grey.

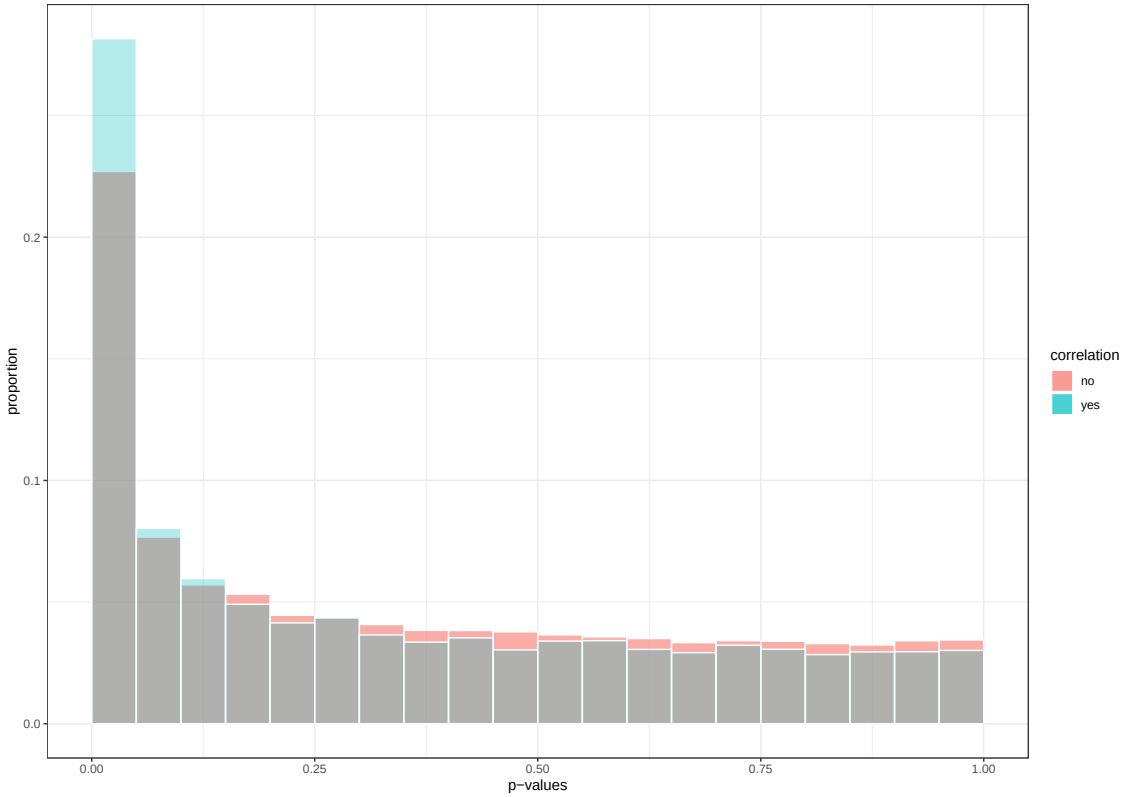



Figure 32: $P$-values for the contrast $\beta_{IA} - \beta_{IU}$ from DESeq2 is plotted in red and the GLMM in blue. When both are present the color mixture becomes grey.

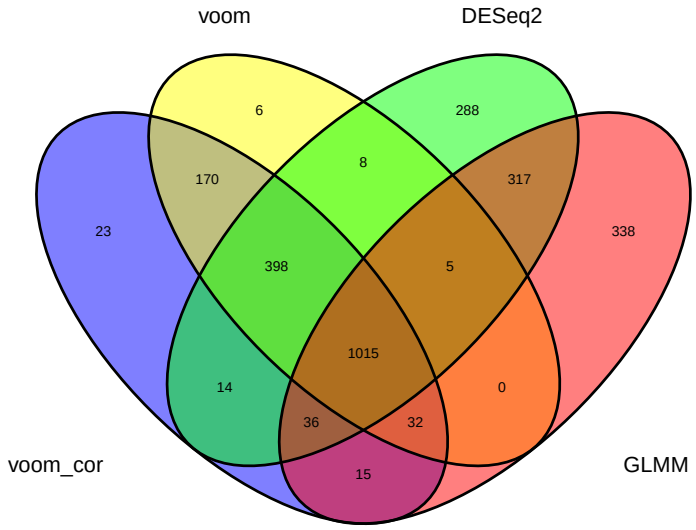



Figure 33: Venn diagram for statistically significant genes for the contrast $\beta_{IA} - \beta_{IU}$, with a FDR significance level of 5%, between limma-voom when and when not accounting for correlation, DESeq2 and a GLMM. In this figure limma-voom is denoted as voom, and limma-voom with correlation is denoted as voom_cor. This figure was produced using the `ggvenn` function from the ggvenn package.

# 7 Discussion

**Filtering** Filtering is a central topic of discussion for gene expression count data analysis. Common practices vary from using the implemented filterByExpr function (as done in Law, Alhamdoosh et al. (2016)), manually selecting a cutoff count number or CPM that at least a selected number of samples satisfy (as done in Chen, Lun and Smyth (2016)), and removing genes where the average count is higher than a specified number (as done in Love, Anders, Kim et al. (2015)). Preferably we want our filtering criteria to be;

- independent from the test statistic under the null hypothesis,
- correlated to the test statistic under the alternative hypothesis, and
- to not change the dependence structure between the test before filtering.

See Love, Anders and Huber (2022) and Bourgon, Gentleman and Huber (2010).

**Batch effects** The analyses we have presented in Section 6 do all assume that the data is of the type assumed by the model. One challenge not discussed in Section 6 is potential batch effects. Leek et al. (2010) defines batch effects as follows: "Batch effects are sub groups of measurements that have qualitatively different behaviour across conditions and are unrelated to the biological or scientific variables in a study." These could be from for example artifacts from the overall extraction procedure. When these are not known or cannot be directly modeled it is more appropriate to use a function that tries to estimate these effects empirically, for example surrogate variable analysis (SVA). These methods work best when the biological variables of interest are not highly confounded with the batch effects (Leek et al. 2010). Other established functions for the same purpose are remove unwanted variation (RUV), and probabilistic estimation of expression residuals (PEER). For larger datasets this practice is encouraged by the one of the authors of the DESeq2 package (Love 2022). This practice is also mentioned in the vignettes for DESeq2, where it is encouraged to correct for unwanted variation when it is present in the data (Love, Anders and Huber 2022).

**Batch effects for our dataset** We should mention that when empirically estimating batch effects it should capture the correlation between samples from the same patient in some way. So this might help DESeq2 solve the problem of not being able to model variation by capturing parts of random effects as fixed effects. We analyzed our data with SVA for limma-voom, where it estimated 22 significant surrogate variables. After SVA was implemented the consensus correlation was estimated to half of its original value. For a FDR of 5% and the main contrast of interest, $\beta_{IA} - \beta_{IU}$, 409 genes were found to be relevant instead of 1703. 22 significant surrogate variables is more than we expected, and this could come from the surrogate variables trying to estimate the underlying correlation. It would be interesting to observe how the surrogate variables would have behaved if person was included as a fixed effect. Further work would be to include SVA into the pipeline and observe how it alters the fit, and observe the differences between the correlated and uncorrelated model.

**Normalization** Normalization is another central topic for discussion for gene expression count data analysis. Scaling with respect to library size was the old standard. However for many biological applications this alone is too simple. We also need to consider the sample composition. If one gene is highly expressed for a sample it would reduce all other gene counts that are normalized with respect to library size. Limma-voom solves this by considering TTM normalization, whilst DESeq2 only considers the library size through size factors. In essence these are quite similar and the main difference is that TTM normalization uses a trimmed mean, and size factors uses the median.

**Correlation for models** In our analysis we have examined models that do and do not account for correlation. For limma-voom we used a LMM for estimating a consensus correlation before

using weighted least squares with respect to the consensus correlation to estimate effects. For GLM we used a GLMM for correlation, where the random effects were supposed to capture the patient specific correlation. These both follow similar assumptions, but the mean dependency of the negative binomial GLMM is more complex. In the case of our dataset the limma-voom pipeline with correlation found approximately 4% more genes to be statistically significant than the limma-voom pipeline that did not account for correlation. By including the correlation for limma-voom the analysis appears to have become more powerful. For limma-voom with correlation there are articles documenting the pipeline controlling the FDR at a chosen level for large enough sample sizes. This is also true for DESeq2, where it in general has higher power (Bi and Liu 2016).

**Simulations** In Section 5 we have observed how a GLM and GLMM would fit models for data generated from a GLMM. In general we observe that GLM is generally conservative, but has lower power, while GLMM has higher power and is also generally conservative. Out of nine simulations only one did not give valid $p$-values compared to three for GLM. Due to our focus on the FDR we would rather want a conservative model, than an optimistic one. Further work could include modeling a mean variance relationship for GLMM to make the dispersion estimates more robust, and borrow information across genes to improve the gene-wise estimate.

**Multiple observations from one individual** It should also be considered what multiple samples extracted from one individual should be interpreted as. Here we have considered correlation as long as the sample is from the same individual, but this might not be the case when the samples are from two different tissues. From our understanding of genetics the gene expression measure in tissue from the same patient should be correlated. However if a gene is only active in one of the tissues we would not expect that gene to be correlated across the tissues. This is something that could be considered in further work.

**Verifications of findings** The ultimate goal of this thesis is to produce a list of potentially relevant genes that highlight the difference in behaviour for Crohn's disease between colon and ileum. As follow up work to this thesis we would want to examine the genes by comparing our results to other publications, and cluster genes from the same process together to get an idea of which processes influence Crohn's disease.

# 8 Conclusion

In this thesis we aimed to identify differentially expressed genes for linear contrasts from gene expression count data. This was done through two established pipelines for gene expression count data, limma-voom and DESeq2. A key element of this thesis is modelling multiple observations from the same individual. Mixed effects models were introduced to account for correlation. We created a pipeline for generalized linear mixed models inspired by DESeq2, whilst for limma-voom a linear mixed model was already implemented. Through simulations this new pipeline was found to be more powerful than GLM under the null hypothesis, where multiple simulations were run for one gene. Our simple pipeline has to be improved upon to be able to compete against established pipelines, where it requires features such as regularization to make the estimates more robust and a way to borrow information across genes.

Based on the dataset analysed in this thesis multiple genes were identified as relevant with an FDR of 0.05. This analysis was done by two different underlying assumptions, where both had two models, one that accounted for correlation and one that did not. The fit for all four models were found to be good, both through MA plot and a Q-Q plot for one gene. To get a robust result we presented the genes all models found to be statistically significant as our result, where the top 20 are presented in Table 5. This consensus list takes into account different models and pipelines, and we believe it should be a good reference point for further studies of Crohn's disease.

# References


Agresti, Alan (2003). *Categorical data analysis*. John Wiley & Sons.

Ahmad, Tariq, Sara E Marshall and Derek Jewell (2006). «Genetics of inflammatory bowel disease: the role of the HLA complex». In: *World journal of gastroenterology: WJG* 12.23, p. 3628.

Anders, Simon, Alejandro Reyes and Wolfgang Huber (2012). «Detecting differential usage of exons from RNA-seq data». In: *Nature Precedings*, pp. 1–1.

Ankill, Sebastian Ø (2021). «Statistical methods for analysis of gene expression count data from patients with Crohn's disease». In: Available upon request.

— (2022). *"Collection of R Scripts for analysis of gene expression count data"*. URL: https://github.com/Sebank/gene-analysis.

Benjamini, Yoav and Yosef Hochberg (1995). «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, pp. 289–300.

Berger, James O (2013). *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media.

Bi, Ran and Peng Liu (2016). «Sample size calculation while controlling false discovery rate for differential expression analysis with RNA-sequencing experiments». In: *BMC bioinformatics* 17.1, pp. 1–13.

Bolker, Ben (2022). *GLMM FAQ*. URL: https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html (visited on 01/04/2022).

Bourgon, Richard, Robert Gentleman and Wolfgang Huber (2010). «Independent filtering increases detection power for high-throughput experiments». In: *Proceedings of the National Academy of Sciences* 107.21, pp. 9546–9551.

Chen, Yunshun, Aaron TL Lun and Gordon K Smyth (2016). «From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline». In: *F1000Research* 5.

Datta, Somnath and Dan Nettleton (2014). *Statistical analysis of next generation sequencing data*. Springer.

De Jong, Piet, Gillian Z Heller et al. (2008). «Generalized linear models for insurance data». In: *Cambridge Books*.

Dunn, Peter K and Gordon K Smyth (2018). *Generalized linear models with examples in R*. URL: https://www.huber.embl.de/msmb/ (visited on 14/02/2022).

Fahrmeir, Ludwig et al. (2007). *Regression*. Springer.

*glmer.nb: Fitting Negative Binomial GLMMs* (2022). URL: https://www.rdocumentation.org/packages/lme4/versions/1.1-29/topics/glmer.nb (visited on 22/04/2022).

Goeman, Jelle J and Aldo Solari (2014). «Multiple hypothesis testing in genomics». In: *Statistics in medicine* 33.11, pp. 1946–1978.

Granlund, Atle van Beelen et al. (2013). «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, e56818.

Holmes, Susan Huber and Wolfgang Huber (2018). *Modern statistics for modern biology*. Cambridge University Press.

Illumina, Inc. (2022). *NGS library preparation*. URL: https://www.illumina.com/techniques/sequencing/ngs-library-prep.html (visited on 19/04/2022).

Johnson, Norman L, Samual Kotz and Narayanaswamy Balakrishnan (1994). *Continuous univariate distributions, volume 1*. Wiley.

Law, Charity W, Monther Alhamdoosh et al. (2016). «RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR». In: *F1000Research* 5.

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

Leek, Jeffrey T et al. (2010). «Tackling the widespread and critical impact of batch effects in high-throughput data». In: *Nature Reviews Genetics* 11.10, pp. 733–739.

Love, Michael I, Simon Anders, Vladislav Kim et al. (2015). «RNA-Seq workflow: gene-level exploratory analysis and differential expression». In: *F1000Research* 4.

Love, Michael I, Wolfgang Huber and Simon Anders (2014). «Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2». In: *Genome biology* 15.12, pp. 1–21.

Love, Michael I., Simon Anders and Wolfgang Huber (2022). *Analyzing RNA-seq data with DESeq2*. URL: https://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html (visited on 09/02/2022).
Love, Mike (2022). *Thread for high samples differential expression analysis*. URL: https://twitter.com/mikelove/status/1513468597288452097 (visited on 25/04/2022).
Marquardt, Donald W and Ronald D Snee (1975). «Ridge regression in practice». In: *The American Statistician* 29.1, pp. 3–20.
McCarthy, Davis J, Yunshun Chen and Gordon K Smyth (2012). «Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation». In: *Nucleic acids research* 40.10, pp. 4288–4297.
Ritchie, Matthew E et al. (2015). «limma powers differential expression analyses for RNA-sequencing and microarray studies». In: *Nucleic acids research* 43.7, e47–e47.
Robinson, Mark D and Alicia Oshlack (2010). «A scaling normalization method for differential expression analysis of RNA-seq data». In: *Genome biology* 11.3, pp. 1–9.
Smyth, Gordon (2016). *filterByExpr: Filter Genes By Expression Level*. URL: https://rdrr.io/bioc/edgeR/man/filterByExpr.html (visited on 15/12/2021).
Smyth, Gordon K (2004). «Linear models and empirical bayes methods for assessing differential expression in microarray experiments». In: *Statistical applications in genetics and molecular biology* 3.1.
Smyth, Gordon K, Joëlle Michaud and Hamish S Scott (2005). «Use of within-array replicate spots for assessing differential expression in microarray experiments». In: *Bioinformatics* 21.9, pp. 2067–2075.
The Crohn's and Colitis Foundation of America (CCFA) (2014). *The facts about inflammatory bowel diseases*. URL: https://www.crohnscolitisfoundation.org/sites/default/files/legacy/assets/pdfs/ibdfactbook.pdf (visited on 15/12/2021).
Van den Berge, Koen et al. (2019). «RNA sequencing data: Hitchhiker's guide to expression analysis». In: *Annual Review of Biomedical Data Science* 2, pp. 139–173.
Venables, William N and Brian D Ripley (2002). *mixedModel2: Fit Mixed Linear Model with 2 Error Components*. URL: https://www.rdocumentation.org/packages/statmod/versions/1.4.36/topics/mixedModel2 (visited on 15/12/2021).
Wickelmaier, Florian (2003). «An introduction to MDS». In: *Sound Quality Research Unit, Aalborg University, Denmark* 46.5, pp. 1–26.
Wu, Hao, Chi Wang and Zhijin Wu (2013). «A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data». In: *Biostatistics* 14.2, pp. 232–243.

# A Statistical derivation of formulas

## A.1 GLM with normal prior simplifies to reweighted ridge regression for a GLM

Here we will show how the given link function and a penalty term changes the MLE estimate to a ridge like shrinkage algorithm. We start by finding the relevant equations, namely equation (3) and (4), which we write out for convenience. We also write out the definition of the $M$ matrix, which is defined is Section 3.2.1 and the link function used with its offset defined in Section 4.2.2. For Ridge we also want to consider an expanded design matrix, which uses a different identifier for each level of a factor. This changes both $\beta$ and $X$, and are denoted $\beta^*$ and $X^*$.

$$\hat{\boldsymbol{\beta}}^{*(r+1)} = \hat{\boldsymbol{\beta}}^{*(r)} + \mathcal{I}\left(\hat{\boldsymbol{\beta}}^{*(r)}\right)^{-1} U\left(\hat{\boldsymbol{\beta}}^{*(r)}\right)$$

$$\mathcal{I}\left(\hat{\boldsymbol{\beta}}^{*(r)}\right) = \frac{1}{\phi} X^{*T} W^{(r)} X^*, \quad U\left(\hat{\boldsymbol{\beta}}^{*(r)}\right) = \frac{1}{\phi} X^{*T} W^{(r)} M^{(r)} (\mathbf{r}_g - \boldsymbol{\mu}^{(r)})$$

$$M_{ii}^{(r)} = \frac{\partial \eta_i^{(r)}}{\partial \mu_i^{(r)}} = \frac{1}{\mu_i^{(r)} \ln(2)}$$

$$\eta_i^{(r)} = \log_2 z_i + x_i^{*T} \hat{\boldsymbol{\beta}}^{*(r)} = \log_2 \mu_i^{(r)}$$

When considering a penalty term it will be added to the log likelihood, and the derivations will be a separate term compared to the score function and the expected Fisher information. For the score function we want to add the derivative of the prior,

$$\frac{\partial}{\partial \hat{\boldsymbol{\beta}}^{*(r)}} \left( -\frac{\lambda}{2} \hat{\boldsymbol{\beta}}^{*(r)T} \hat{\boldsymbol{\beta}}^{*(r)} \right) = -\lambda \hat{\boldsymbol{\beta}}^{*(r)}.$$

For the expected Fisher information we want use the changed score function instead of the original. This results in a term of the form,

$$-\frac{\partial}{\partial \hat{\boldsymbol{\beta}}^{*(r)}} \left( -\lambda \hat{\boldsymbol{\beta}}^{*(r)T} \right) = \lambda.$$

Then the iterative scheme becomes,

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^{*(r+1)} &= \hat{\boldsymbol{\beta}}^{*(r)} + \mathcal{I}\left(\hat{\boldsymbol{\beta}}^{*(r)}\right)^{-1} U\left(\hat{\boldsymbol{\beta}}^{*(r)}\right), \\
\hat{\boldsymbol{\beta}}^{*(r+1)} &= \hat{\boldsymbol{\beta}}^{*(r)} + \left(\frac{1}{\phi} X^{*T} W^{(r)} X^{*} + \lambda\right)^{-1} \left(\frac{1}{\phi} X^{*T} W^{(r)} M^{(r)} (\mathbf{r}_g - \boldsymbol{\mu}^{(r)}) - \lambda \hat{\boldsymbol{\beta}}^{*(r)}\right), \\
\hat{\boldsymbol{\beta}}^{*(r+1)} &= \left(\frac{1}{\phi} X^{*T} W^{(r)} X^{*} + \lambda\right)^{-1} \left(\left(\frac{1}{\phi} X^{*T} W^{(r)} X^{*} + \lambda\right) \hat{\boldsymbol{\beta}}^{*(r)} + \frac{1}{\phi} X^{*T} W^{(r)} M^{(r)} (\mathbf{r}_g - \boldsymbol{\mu}^{(r)}) - \lambda \hat{\boldsymbol{\beta}}^{*(r)}\right), \\
\hat{\boldsymbol{\beta}}^{*(r+1)} &= \left(\frac{1}{\phi} X^{*T} W^{(r)} X^{*} + \lambda\right)^{-1} \left(\frac{1}{\phi} X^{*T} W^{(r)} X^{*} \hat{\boldsymbol{\beta}}^{*(r)} + \frac{1}{\phi} X^{*T} W^{(r)} M^{(r)} \left(\mathbf{r}_g - \boldsymbol{\mu}^{(r)}\right)\right), \\
\hat{\boldsymbol{\beta}}^{*(r+1)} &= \left(\frac{1}{\phi} X^{*T} W^{(r)} X^{*} + \lambda\right)^{-1} \frac{1}{\phi} X^{*T} W^{(r)} \left(X^{*} \hat{\boldsymbol{\beta}}^{*(r)} + M^{(r)} \left(\mathbf{r}_g - \boldsymbol{\mu}^{(r)}\right)\right).
\end{aligned}$$

The last term, $\left(X^{*} \hat{\boldsymbol{\beta}}^{*(r)} + M^{(r)} \left(\mathbf{r}_g - \boldsymbol{\mu}^{(r)}\right)\right)$, becomes a diagonal matrix, Z, with elements $Z_{ii}^{(r)} = \log_2 \frac{\mu_i^{(r)}}{z_i} + \frac{r_{gi} - \mu_i^{(r)}}{\mu_i^{(r)} \ln(2)}$. We can now observe that the iterative scheme is of the same form as a reweighted ridge regression for a GLM,

$$\hat{\boldsymbol{\beta}}^{*(r+1)} = \left(\frac{1}{\phi} X^{*T} W^{(r)} X^{*} + \lambda\right)^{-1} \frac{1}{\phi} X^{*T} W^{(r)} Z^{(r)}.$$

## A.2 Variance for Wald for negative binomial GLM with normal prior

Our intention is to find the variance for our estimated effects after convergence, from the iterative scheme,

$$\hat{\boldsymbol{\beta}}^{*(r+1)} = \left(\frac{1}{\phi} X^{*T} W^{(r)} X^{*} + \lambda\right)^{-1} \frac{1}{\phi} X^{*T} W^{(r)} Z^{(r)}.$$

We denote the converged estimate as being of iteration $n$. We start by rewriting the variance to isolate the variance of $Z^{(n)}$,

$$\begin{aligned}
\operatorname{Var}\left[\hat{\boldsymbol{\beta}}^{*(n)}\right] &= \operatorname{Var}\left[\left(\frac{1}{\phi} X^{*T} W^{(n)} X^{*} + \lambda\right)^{-1} \frac{1}{\phi} X^{*T} W^{(n)} Z^{(n)}\right] \\
&= \left(\frac{1}{\phi} X^{*T} W^{(n)} X^{*} + \lambda\right)^{-1} \frac{1}{\phi} X^{*T} W^{(n)} \operatorname{Var}\left[Z^{(n)}\right] \left(\left(\frac{1}{\phi} X^{*T} W^{(n)} X^{*} + \lambda\right)^{-1} \frac{1}{\phi} X^{*T} W^{(n)}\right)^{T} \\
&= \frac{1}{\phi^2} \left(\frac{1}{\phi} X^{*T} W^{(n)} X^{*} + \lambda\right)^{-1} X^{*T} W^{(n)} \operatorname{Var}\left[Z^{(n)}\right] W^{(n)} X^{*} \left(\frac{1}{\phi} X^{*T} W^{(n)} X^{*} + \lambda\right)^{-1}.
\end{aligned}$$

Then we want to consider the elements of $Z^{(n)}$, where due to its diagonal nature we consider the

diagonal elements separately.

$$\begin{aligned}
\text{Var}\left[Z_{ii}^{(n)}\right] &= \text{Var}\left[\log_2 \frac{\mu_i^{(n)}}{z_i} + \frac{R_{gi} - \mu_i^{(n)}}{\mu_i^{(n)} \ln(2)}\right] \\
&= \text{Var}\left[\frac{R_{gi}}{\mu_i^{(n)} \ln(2)}\right] = \frac{1}{\left(\mu_i^{(n)} \ln(2)\right)^2} \text{Var}[R_{gi}] \\
&= \frac{1}{\left(\mu_i^{(n)} \ln(2)\right)^2} \left(\mu_i^{(n)} + \left(\mu_i^{(n)}\right)^2 \alpha_i\right) \\
&= \left(W_{ii}^{(n)}\right)^{-1}
\end{aligned}$$

The variance of $R_{gi}$ is the variance from the distribution assumed of $R_{gi}$ to follow, in this case the negative binomial distribution. The variance of the negative binomial distribution is found in equation (2). Here $\mu_i^{(n)}$ is considered a constant and does not contribute to the variance of $Z_{ii}^{(n)}$. In the case of the negative binomial the working weights become,

$$\begin{aligned}
W_{ii}^{(n)} &= \frac{1}{V\left(\mu_i^{(n)}\right)\left(\frac{\partial \eta_i}{\partial \mu_i^{(n)}}\right)^2} \\
&= \frac{1}{\frac{\partial \mu_i^{(n)}}{\partial \theta}\left(\frac{\partial \eta_i}{\partial \mu_i^{(n)}}\right)^2} \\
&= \frac{1}{\left(\mu_i^{(n)} + \left(\mu_i^{(n)}\right)^2 \alpha_i\right)\left(\frac{\partial \eta_i}{\partial \mu_i^{(n)}}\right)^2} \\
&= \frac{1}{\left(\mu_i^{(n)} + \left(\mu_i^{(n)}\right)^2 \alpha_i\right)\left(\frac{\partial}{\partial \mu_i^{(n)}} \log_2 \mu_i^{(n)}\right)^2} \\
&= \frac{1}{\left(\mu_i^{(n)} + \left(\mu_i^{(n)}\right)^2 \alpha_i\right)\left(\frac{1}{\mu_i^{(n)} \ln(2)}\right)^2}.
\end{aligned}$$

We can then return to the total variance estimate,

$$\text{Var}\left[\hat{\boldsymbol{\beta}}^{*(n)}\right] = \left(X^{*T} W^{(n)} X^* + \lambda\right)^{-1} X^{*T} W^{(n)} X^* \left(X^{*T} W^{(n)} X^* + \lambda\right)^{-1},$$

which is equal to the result presented without derivation in Love, Huber and Anders (2014).

# B Simulation plots under the alternative hypothesis

These plots correspond to the alternative hypothesis introduced in Section 5.4.3. The parameter values used in this setup were $\boldsymbol{\beta} = [5.3\, 2]$, $\alpha = 0.25$, $\tau = \{0.2, 0.3, 0.4\}$ and $m^* = \{0.3, 0.6, 1\}$.

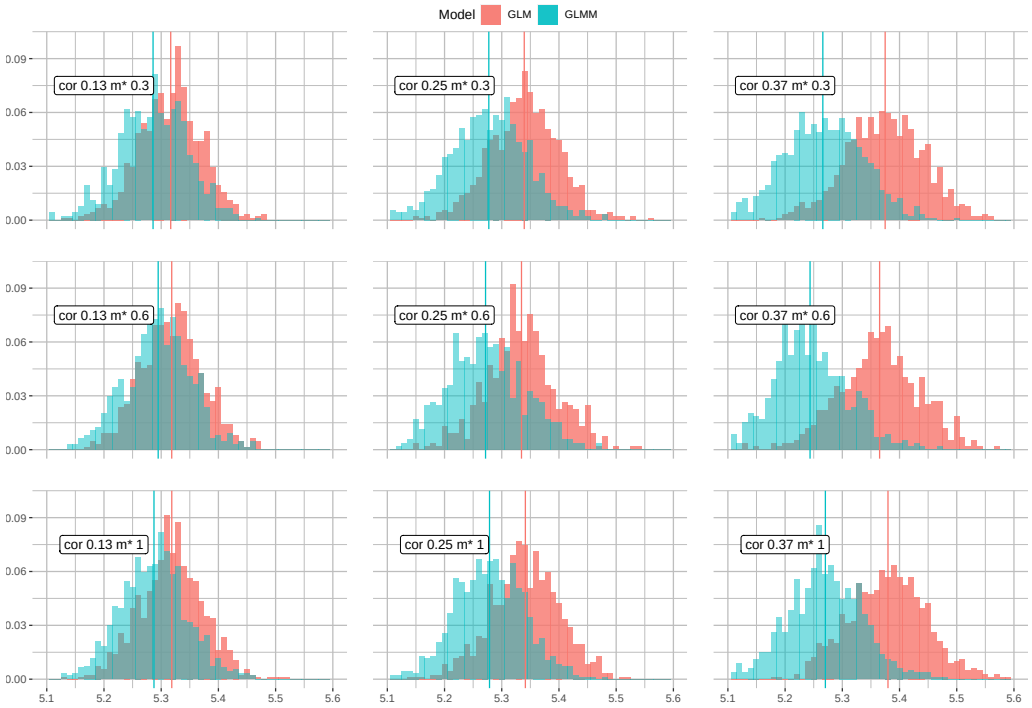



Figure 34: Histogram of the estimated intercept, $\beta_0$, from simulation for different correlations and proportion of data that considers correlation. The estimated effects from the GLM is plotted in red and the GLMM in blue. The vertical lines represents the mean of the intercepts for the two different models.

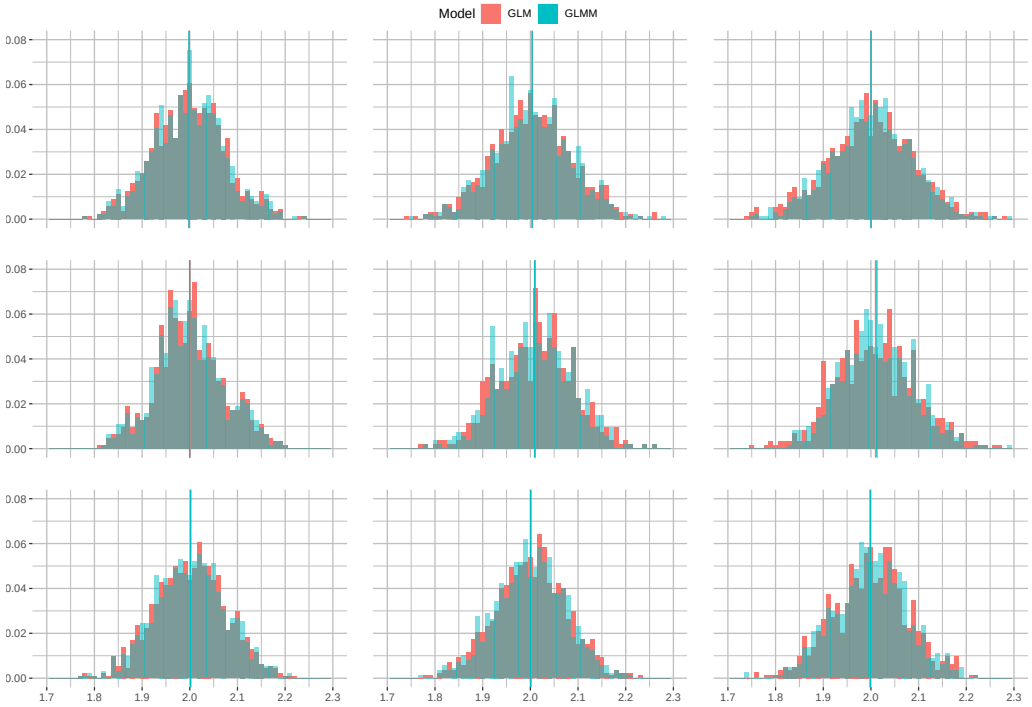



Figure 35: Histogram of the estimated effects, $\beta_1$, from simulation for different correlations and proportion of data that considers correlation. The estimated effects from the GLM is plotted in red and the GLMM in blue. The vertical lines represents the mean of the intercepts for the two different models.

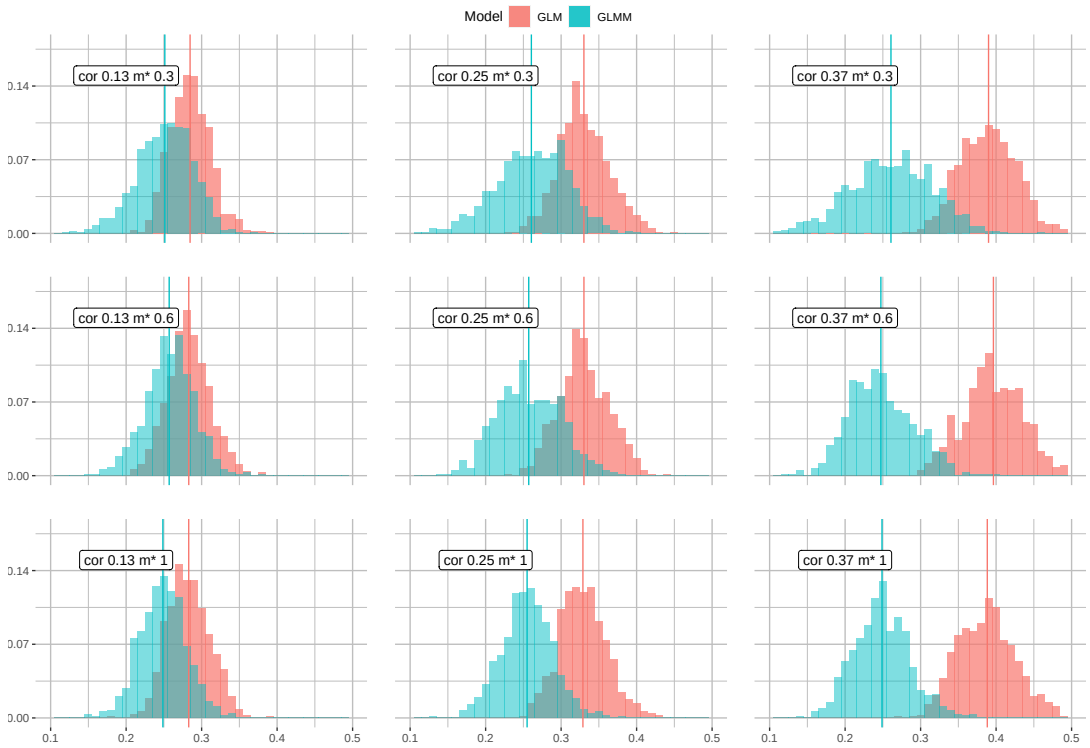



Figure 36: Histogram of the estimated dispersion parameter from simulation for different correlations and proportion of data that considers correlation. The estimated dispersion parameter from the GLM is plotted in red and the GLMM in blue. The vertical lines represents the mean of the dispersion parameters for the two different models.

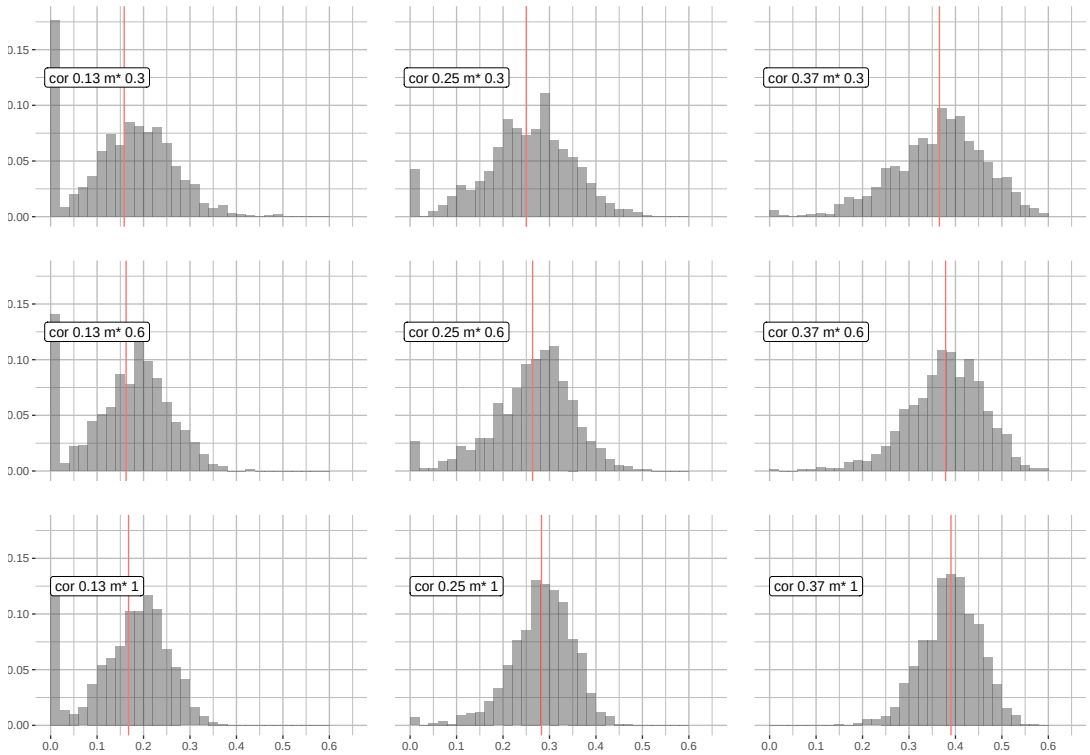



Figure 37: Histogram of the estimated $\tau$ parameter from simulation for different correlations and proportion of data that considers correlation. The vertical line represents the mean of the estimated $\tau$ for the GLMM.

# C Additional results

The plots presented are of the two contrasts of interest, $\beta_{IU}$ and $\beta_{IA}$. These plots are of the same nature as the ones introduced in Section 6, and based on our Crohn's disease dataset. In particular MA plots, distribution of $p$-values, Venn diagrams of $p$-values for different models and top lists for the contrasts $\beta_{IU}$ and $\beta_{IA}$.

## C.1 Model fit

Here MA plots for the contrasts $\beta_{IU}$ and $\beta_{IA}$ will be presented.

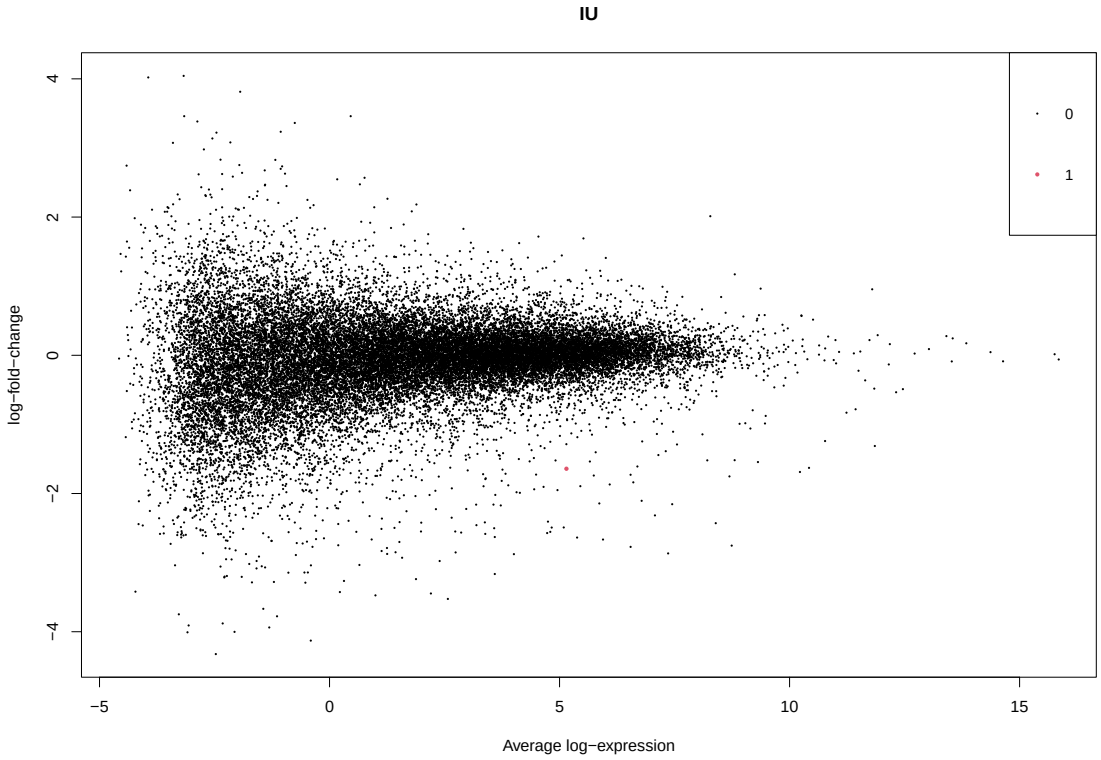



Figure 38: MA plot for limma-voom when not considering correlation for the effect $\beta_{IU}$. This plot was generated using the `plotMA` function from the limma package.

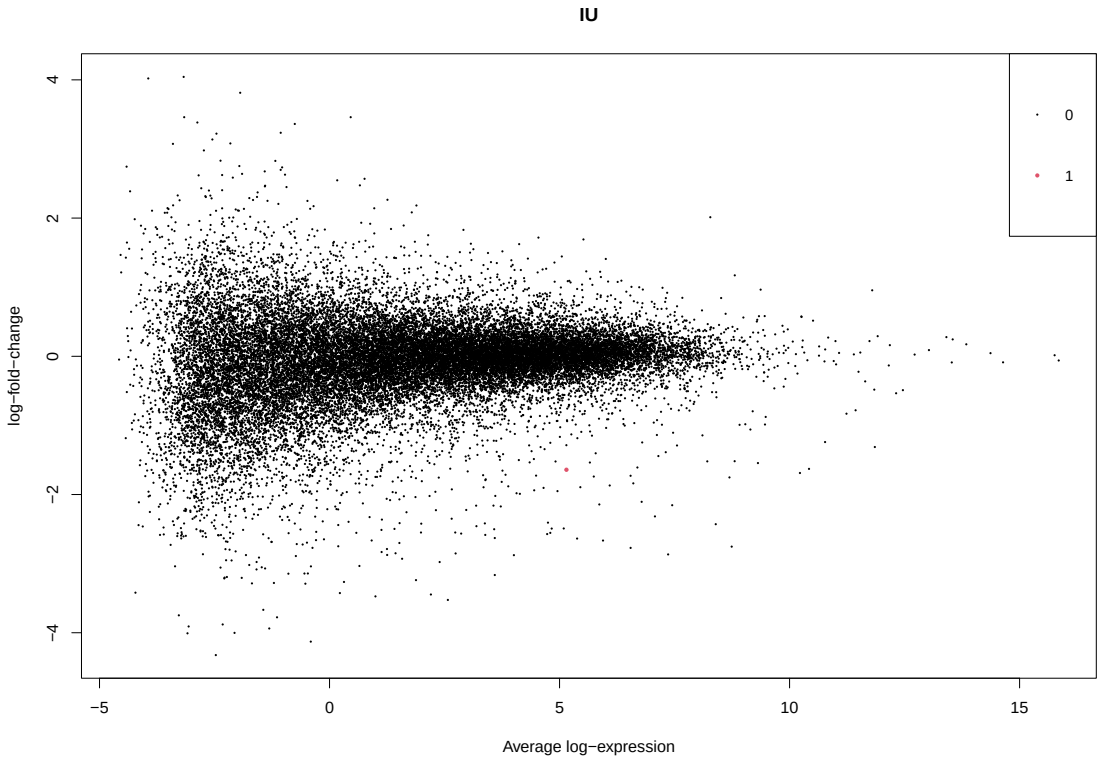



Figure 39: MA plot for limma-voom when considering correlation for the effect $\beta_{IU}$. This plot was generated using the `plotMA` function from the limma package.

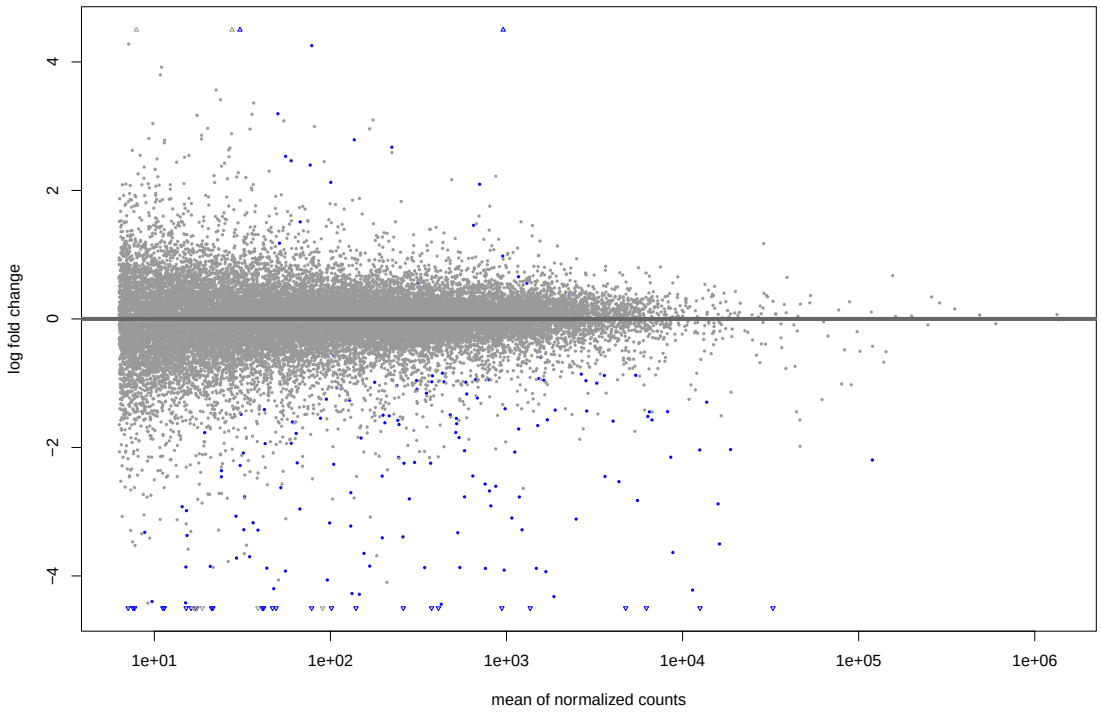



Figure 40: MA plot for DESeq2 for the effect $\beta_{IU}$. This plot was generated using the `plotMA` function from the DESeq2 package.

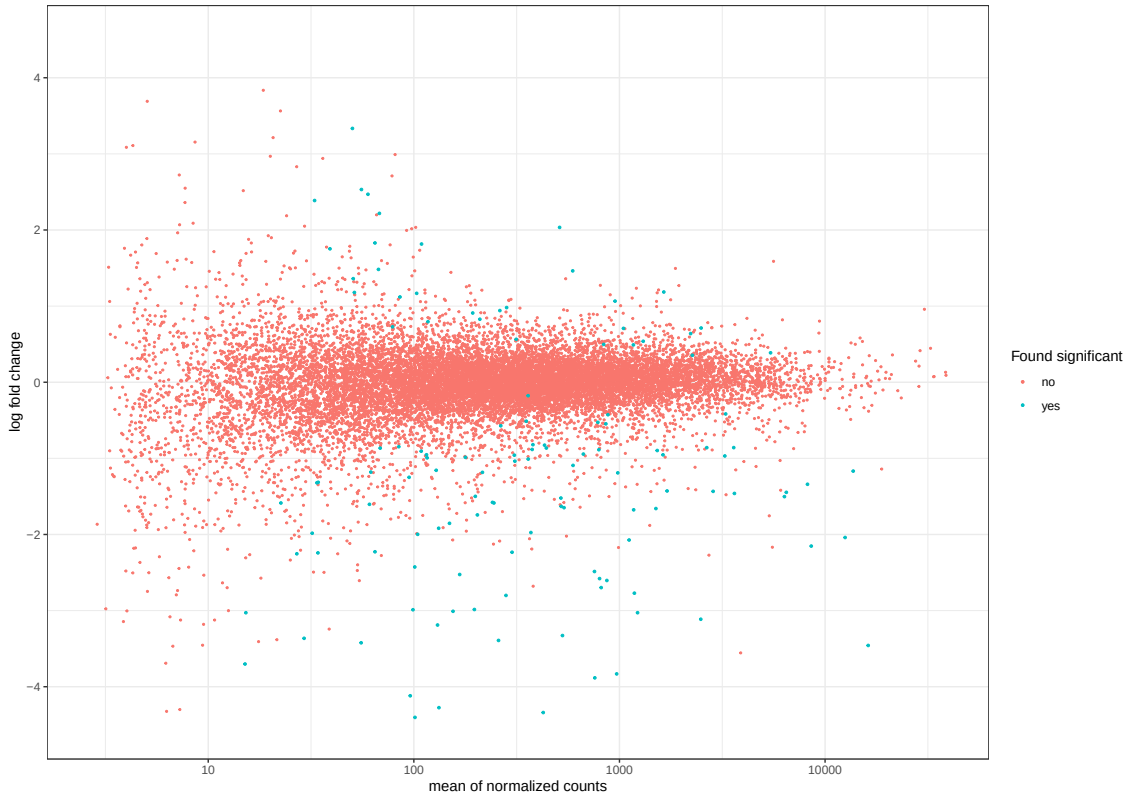



Figure 41: MA plot for DESeq2 for the effect $\beta_{IU}$. R code is found in Ankill (2022).

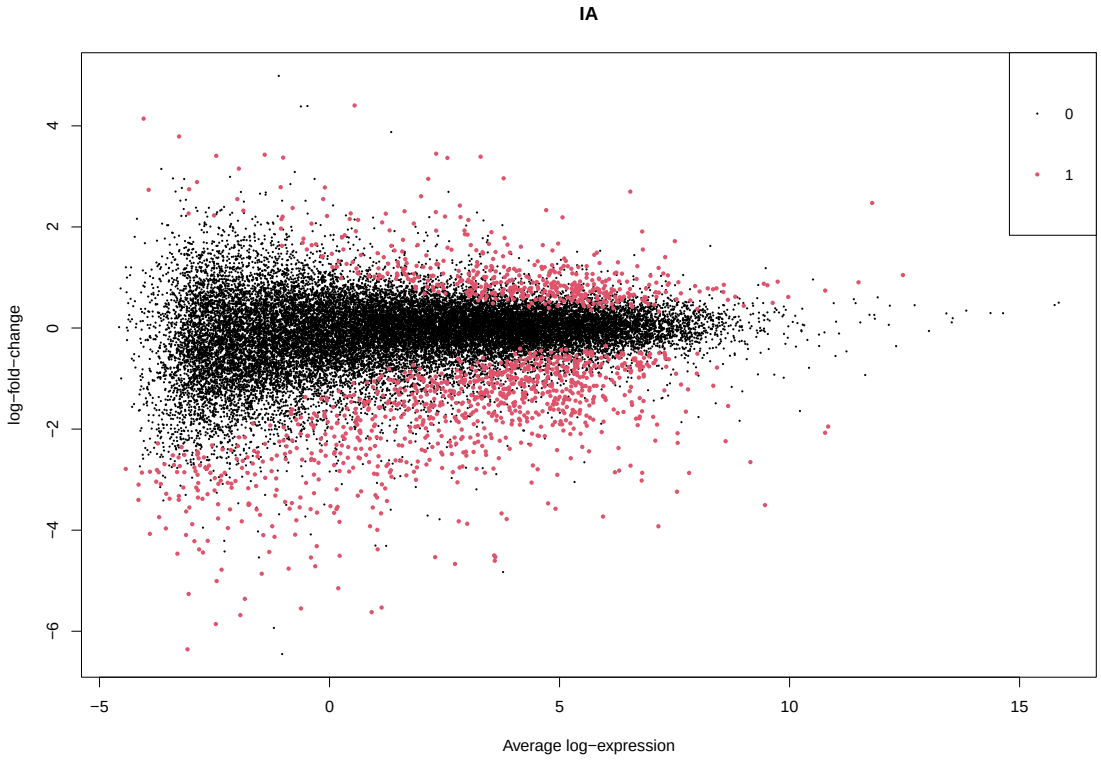



Figure 42: MA plot for limma-voom when not considering correlation for the effect $\beta_{IA}$. This plot was generated using the `plotMA` function from the limma package.

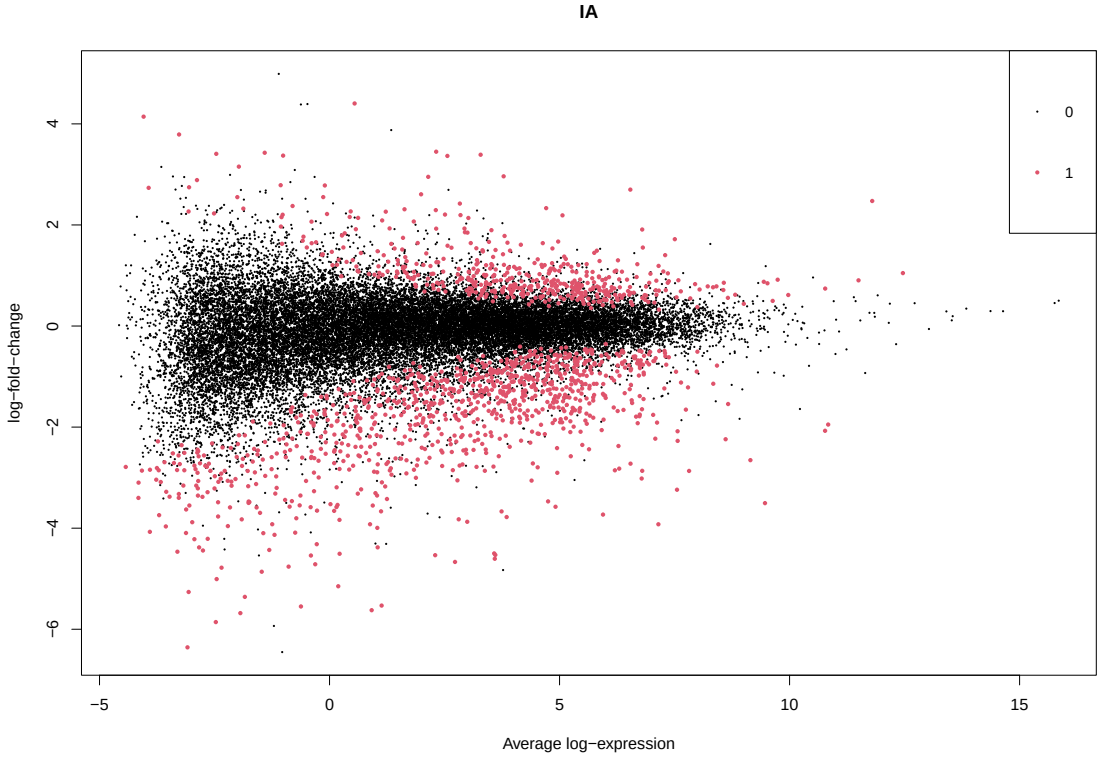



Figure 43: MA plot for limma-voom when considering correlation for the effect $\beta_{IA}$. This plot was generated using the `plotMA` function from the limma package.

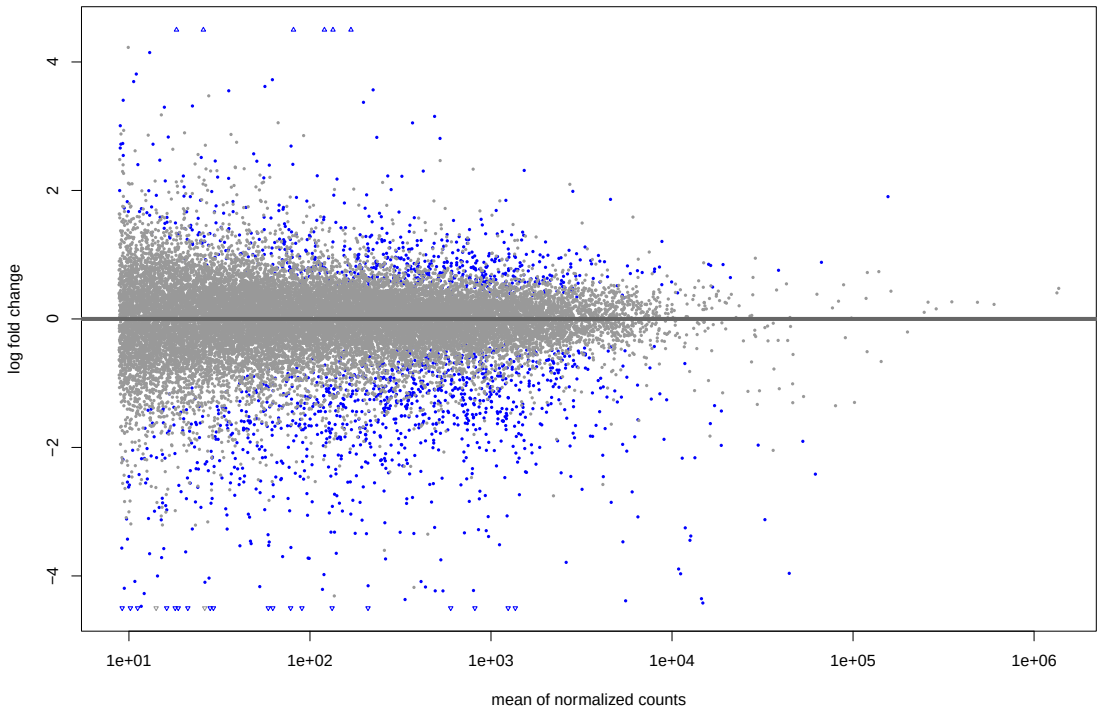



Figure 44: MA plot for DESeq2 for the effect $\beta_{IA}$. This plot was generated using the `plotMA` function from the DESeq2 package.

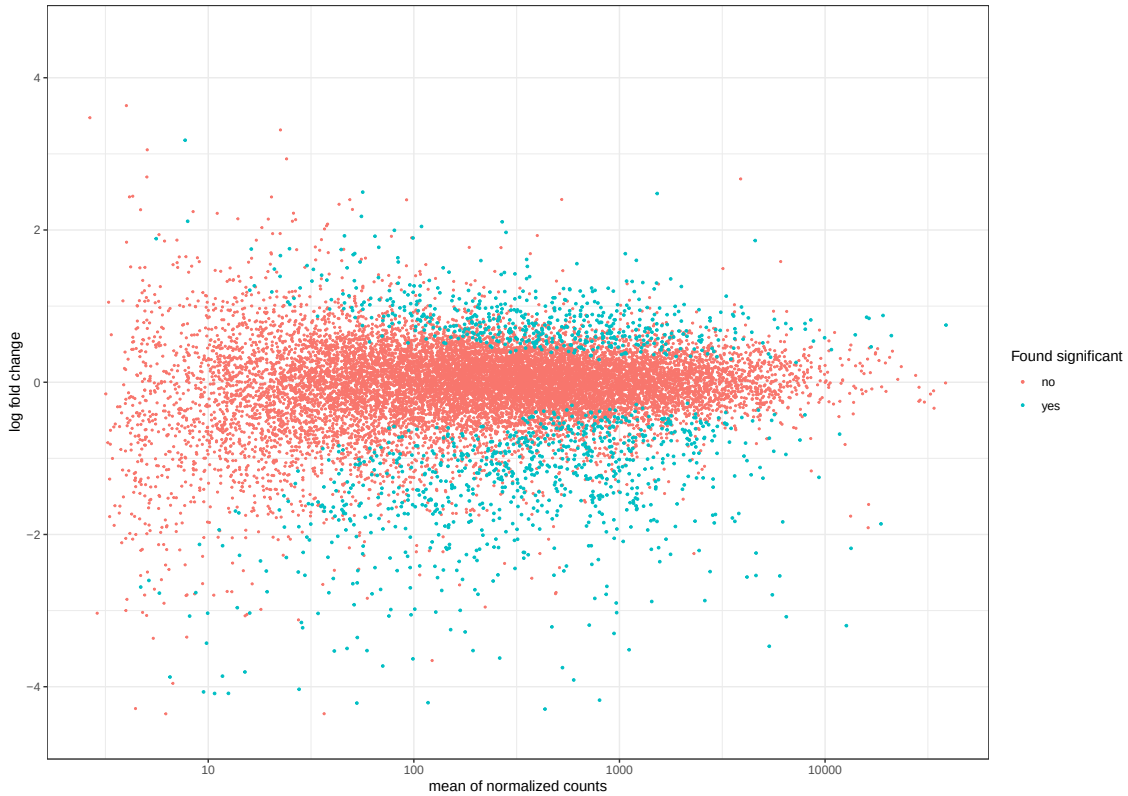



Figure 45: MA plot for DESeq2 for the effect $\beta_{IU}$. R code is found in Ankill (2022).

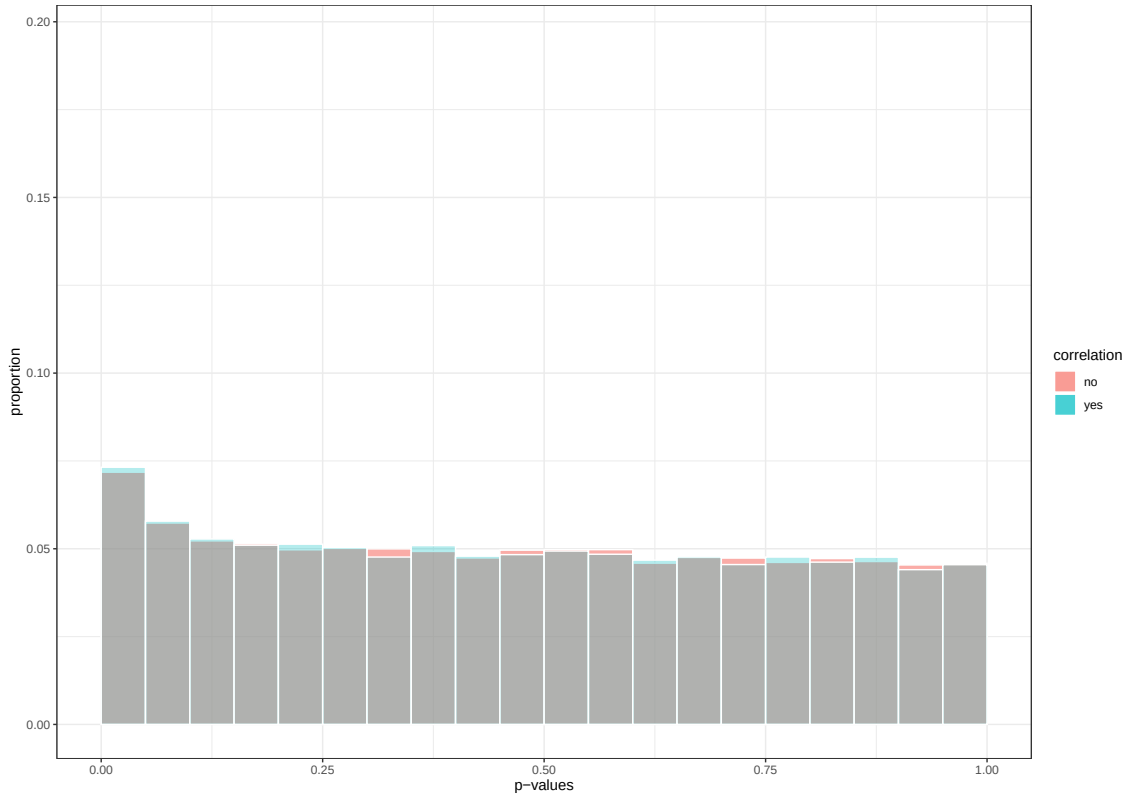



Figure 46: Distribution of $p$-values for limma-voom when considering the effect $\beta_{IU}$. The red corresponds to the pipeline that does not consider correlation for patients, whilst the blue does. When both are present the plot becomes grey.

## C.2 P-values

Here the distribution of $p$-values and venn diagrams for genes found to be statistically for the different models are presented for the contrasts $\beta_{IU}$ and $\beta_{IA}$.

## C.3 Results

Here top tables for the contrasts $\beta_{IU}$ and $\beta_{IA}$ are presented.

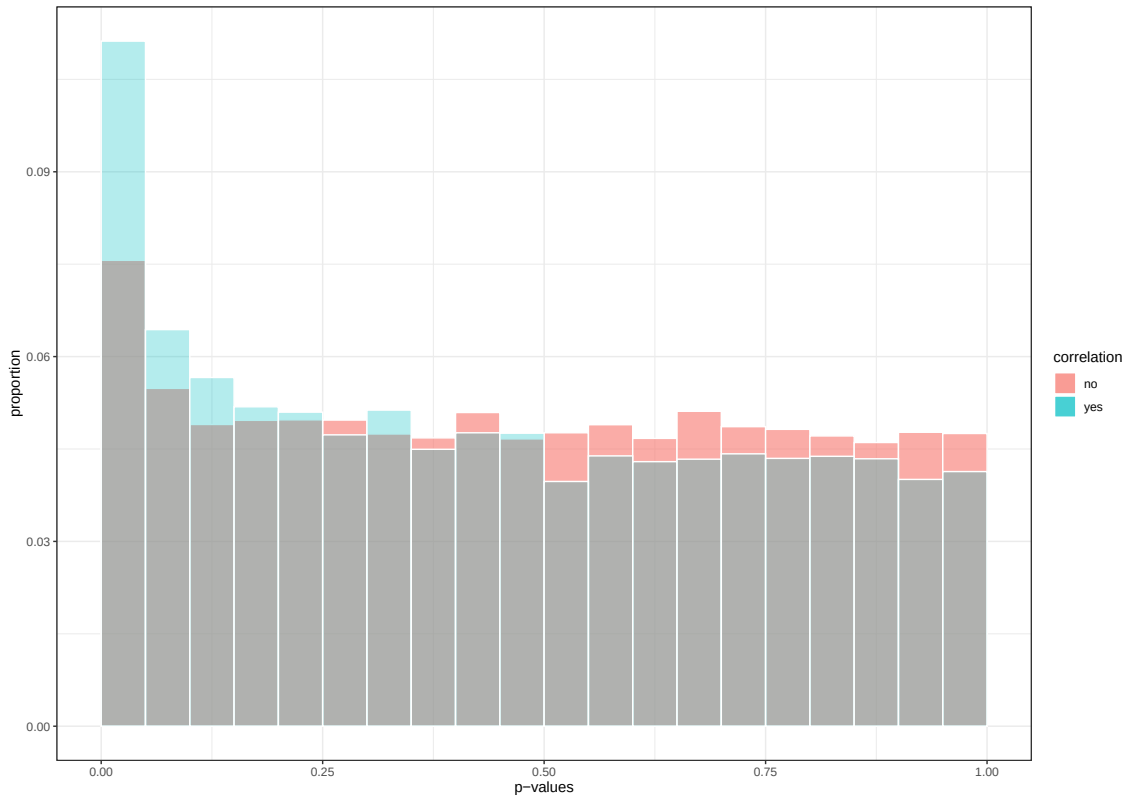



Figure 47: Distribution of $p$-values for DESeq2 and GLMM when considering the effect $\beta_{IU}$. The red corresponds to $p$-values from DESeq2, whilst the blue corresponds to $p$-values from GLMM. When both are present the plot becomes grey.

Table 6: Top 20 genes across models for the Crohn's data set (based on ranking by adjusted $p$-value for all models) for the contrast $\beta_{IU}$.

| names | base.Mean | voom.LFC | voom.cor.LFC | DESeq2.LFC | GLMM.LFC |
|---|---|---|---|---|---|
| SLC9A2 | 1172 | -1.64 | -1.64 | -1.71 | -1.68 |
| PARM1 | 12515 | -1.52 | -1.52 | -2.04 | -2.04 |
| COL6A5 | 55 | 3.46 | 3.41 | 2.53 | 2.53 |
| ANO6 | 952 | 1.08 | 1.09 | 0.979 | 1.07 |
| CA2 | 6224 | -2.77 | -2.79 | -4.83 | -4.83 |
| CD24P4 | 1182 | -1.95 | -1.99 | -2.77 | -2.77 |
| AIFM3 | 370 | -1.72 | -1.74 | -2.25 | -1.97 |
| KCNN4 | 203 | -1.67 | -1.68 | -1.62 | -1.74 |
| GOLM1 | 1628 | -0.824 | -0.821 | -0.954 | -0.95 |
| PIGZ | 1705 | -1.40 | -1.39 | -1.57 | -1.43 |
| ZBTB12 | 314 | 0.552 | 0.554 | 0.561 | 0.563 |
| PLCD3 | 3624 | -1.73 | -1.71 | -2.45 | -1.46 |
| HSPA5P1 | 11 | -4.01 | -3.92 | -9.28 | -9.01 |
| PRKACB | 667 | -0.871 | -0.859 | -0.943 | -0.94 |
| TFCP2L1 | 2485 | -2.14 | -2.13 | -3.12 | -3.11 |
| SATB2-AS1 | 868 | -1.99 | -2.03 | -2.60 | -2.60 |
| B3GALT1 | 197 | -2.42 | -2.44 | -3.41 | -2.99 |
| FFAR4 | 300 | -1.90 | -1.89 | -2.23 | -2.23 |
| LEFTY1 | 130 | -2.85 | -2.82 | -3.22 | -3.19 |
| FER1L4 | 537 | -1.54 | -1.53 | -1.85 | -1.65 |

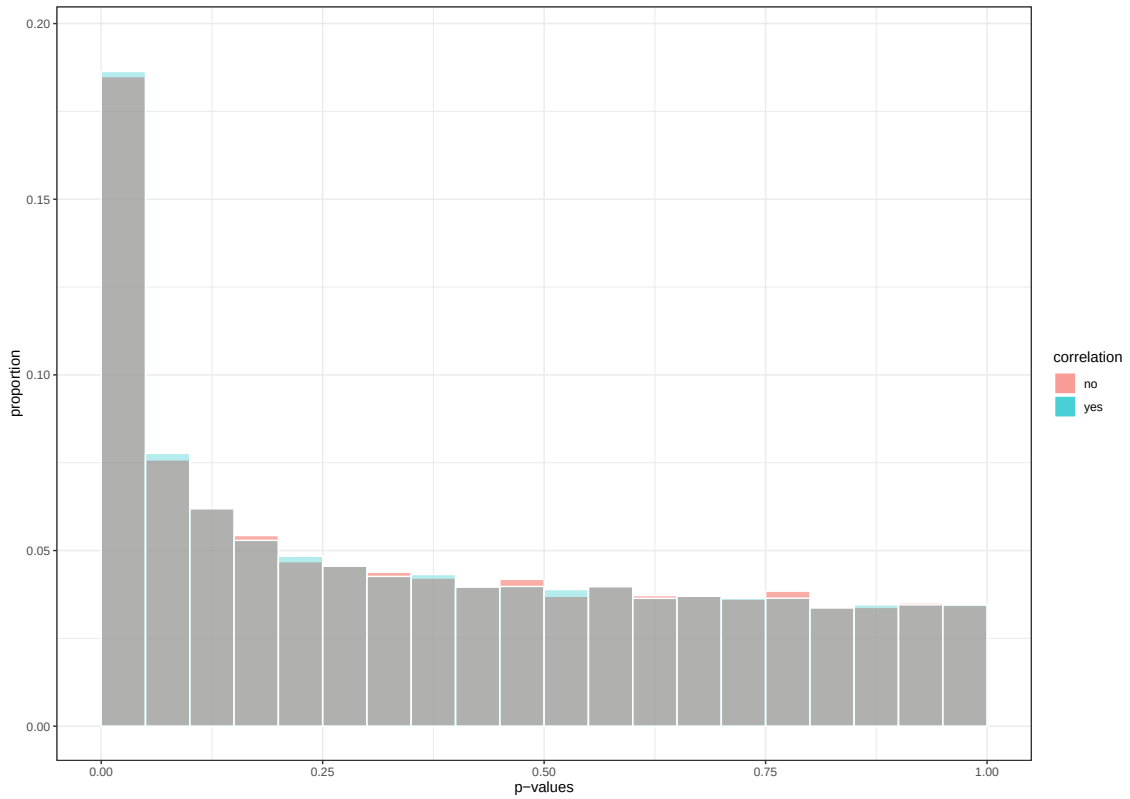



Figure 48: Distribution of $p$-values for limma-voom when considering the effect $\beta_{IA}$. The red corresponds to the pipeline that does not consider correlation for patients, whilst the blue does. When both are present the plot becomes grey.

Table 7: Top 20 genes across models for the Crohn's data set (based on ranking by adjusted $p$-value for all models) for the contrast $\beta_{IA}$.

| names | base.mean | voom.LFC | voom.cor.LFC | DESeq2.LFC | GLMM.LFC |
|---|---|---|---|---|---|
| C10orf99 | 6481 | -3.239 | -3.238 | -3.081 | -3.081 |
| CREB3L2 | 4990 | -1.214 | -1.214 | -1.293 | -1.26 |
| HLA-DMA | 1469 | -1.814 | -1.813 | -1.805 | -1.806 |
| HSPA5P1 | 11 | -6.358 | -6.298 | -8.108 | -8.08 |
| UBE2L6 | 1043 | -2.099 | -2.097 | -2.049 | -2.091 |
| HLA-DRA | 13378 | -2.24 | -2.231 | -2.16 | -2.182 |
| SLC16A1 | 1068 | 2.19 | 2.19 | 1.689 | 1.689 |
| AK3 | 535 | 1.21 | 1.214 | 1.229 | 1.23 |
| PI3 | 815 | -4.53 | -4.514 | -5.394 | -5.242 |
| TRIM40 | 965 | -2.904 | -2.887 | -3.077 | -2.9 |
| APOL1 | 471 | -1.828 | -1.818 | -1.793 | -1.795 |
| FZD7 | 185 | -1.863 | -1.888 | -1.839 | -1.839 |
| CYBC1 | 2437 | -0.922 | -0.927 | -0.914 | -0.932 |
| MDK | 2893 | -1.793 | -1.793 | -1.85 | -1.848 |
| ACAD10 | 602 | 0.8 | 0.797 | 0.823 | 0.803 |
| ENSG00000254488 | 53 | -2.883 | -2.846 | -2.704 | -2.636 |
| DPP7 | 614 | -1.136 | -1.146 | -1.11 | -1.11 |
| HLA-DMB | 730 | -2.09 | -2.087 | -2.159 | -2.148 |
| KLHL29 | 289 | -1.127 | -1.131 | -1.132 | -1.132 |
| ODF3B | 744 | -1.513 | -1.524 | -1.42 | -1.435 |

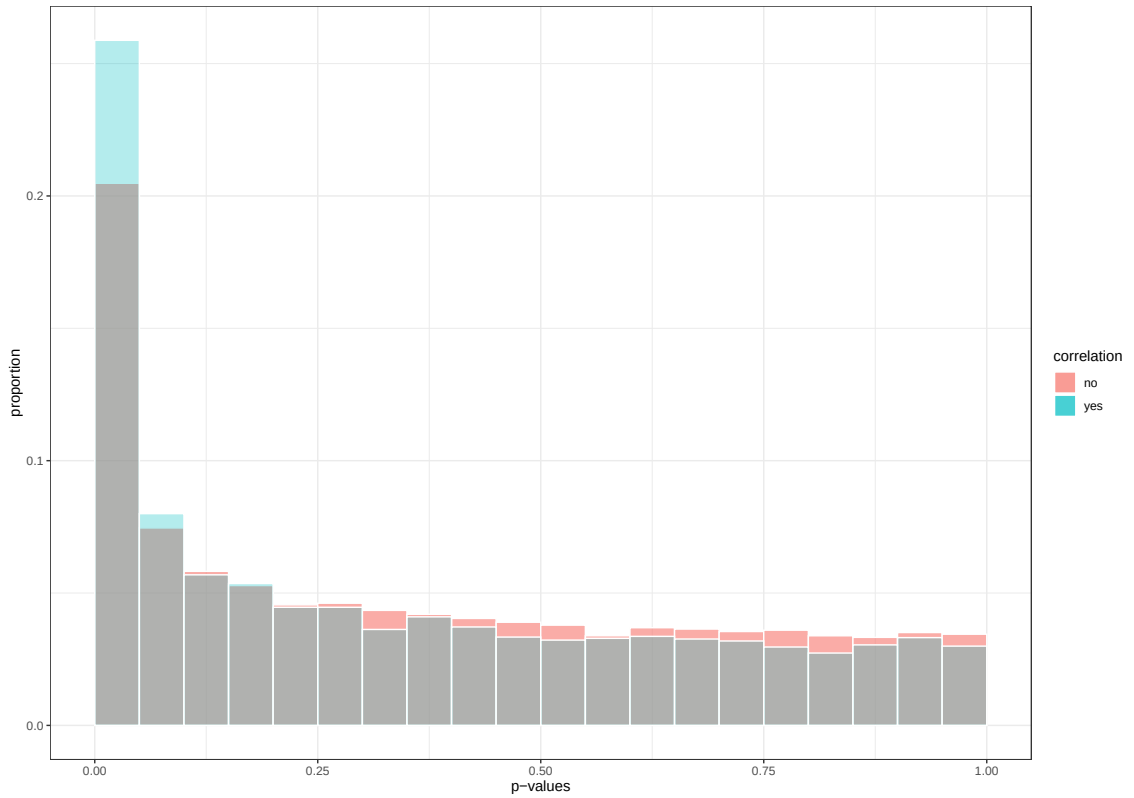



Figure 49: Distribution of $p$-values for DESeq2 and GLMM when considering the effect $\beta_{IA}$. The red corresponds to $p$-values from DESeq2, whilst the blue corresponds to $p$-values from GLMM. When both are present the plot becomes grey.

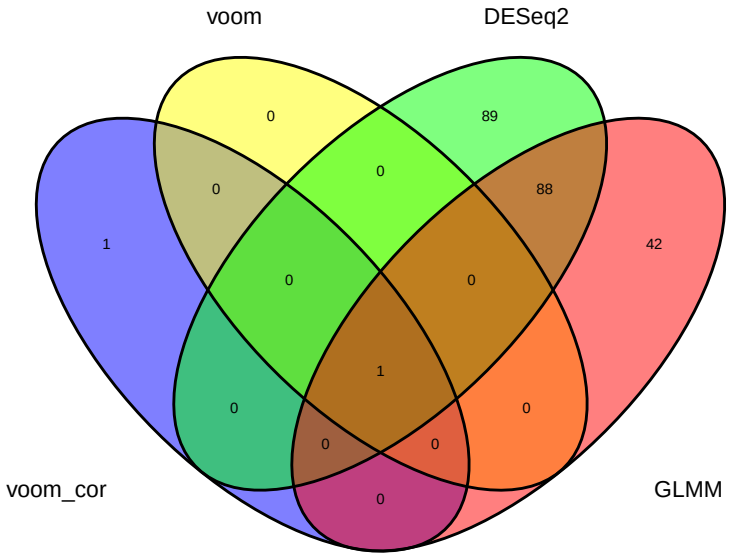



Figure 50: Venn diagram for genes considered significant for the effect $\beta_{IU}$, with a significance level of 5% for adjusted $p$-value, between limma-voom with and without considering correlation, DESeq2 and a GLMM. In this figure limma-voom is denoted as voom, and limma-voom with correlation is denoted as voom_cor. This figure was produced using the `ggvenn` function from the ggvenn package.

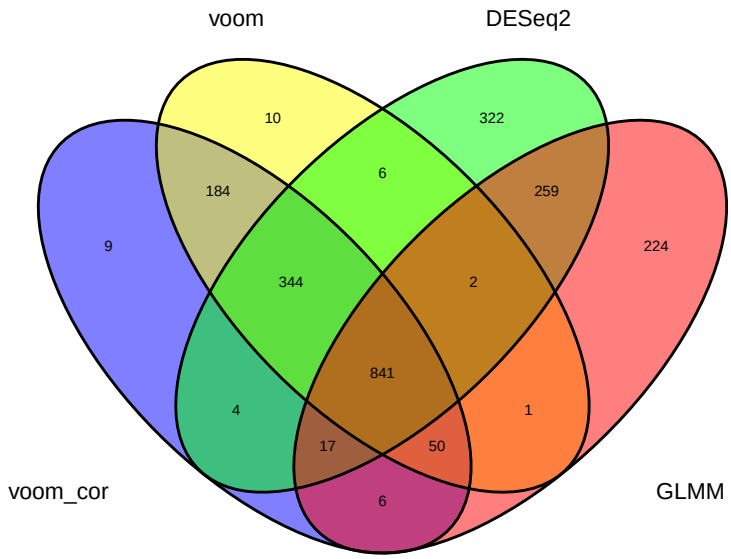
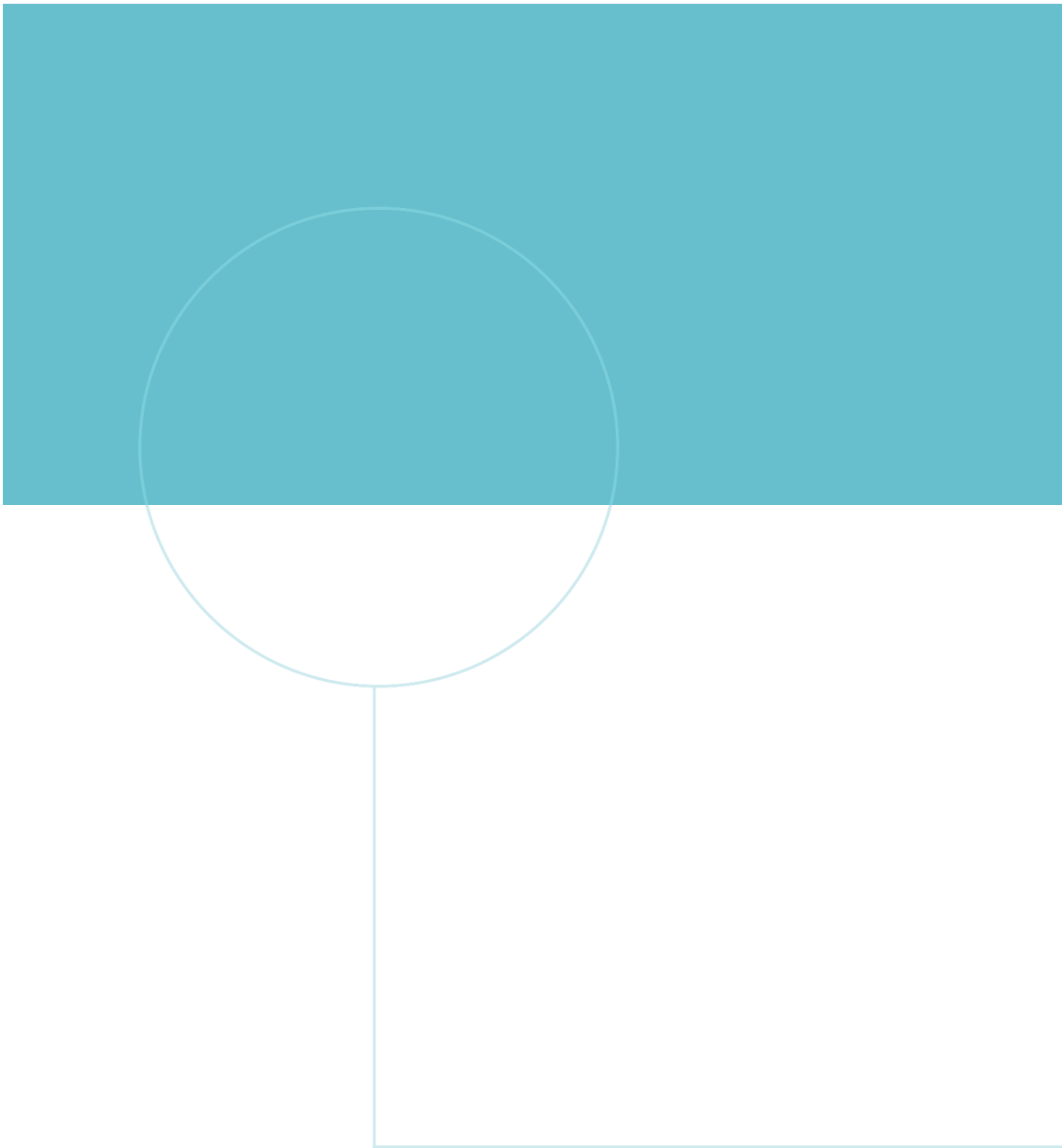



Figure 51: Venn diagram for genes considered significant for the effect $\beta_{IA}$, with a significance level of 5% for adjusted $p$-value, between limma-voom with and without considering correlation, DESeq2 and a GLMM. In this figure limma-voom is denoted as voom, and limma-voom with correlation is denoted as voom_cor. This figure was produced using the `ggvenn` function from the ggvenn package.



NTNU

Norwegian University of
Science and Technology