University of Crete
School of Sciences and Technology
Department of Mathematics and Applied Mathematics

Gender Bias in Clinical Studies: A Statistical Approach

Diploma Thesis

Mikela Daphne Loizou Manske
tem2206

Supervisor: Pavlos Pavlidis, Associate Professor, Department of Biology UoC and Affiliated Researcher, ICS-FORTH

Heraklion 2025

Acknowledgements

I would like to express my sincere gratitude to my supervising professor Pavlos Pavlidis for his guidance and encouragement throughout this research. His own interest has been a source of motivation, inspiring me to push forward.

I am also deeply grateful to my family and friends for their support and motivation. They often remind me how beautiful life is, helping me maintain perspective throughout this journey.

Thank you all for your contributions to this achievement.

Mikela Loizou,

March 2025.

Abstract

This thesis aims to investigate the impact of gender bias in clinical research, with a focus on differences in gene expression. The historical exclusion of females in clinical trials and drug testing has led to significant disparities in healthcare outcomes. To explore this issue, this thesis statistically analyzes gender-based differences in gene expression across various conditions. A few basic theoretical concepts are explained before presenting the methodology used. Using datasets from the GEO database, hypothesis testing is conducted separately for each dataset. Linear models, applied through the limma package in R, identify significantly differentially expressed genes between genders. The results are presented and visualized, highlighting the extent of gender-specific variations in gene expression.

Keywords: Medical research, gender bias, gene expression, differentially expressed gene (DEG), statistical bias, microarray, t-test, linear model, limma, design matrix, contrast matrix, means model, mean-reference model, factor, level, empirical Bayes, GEO datasets

Contents

Lists of Figures and Tables

1. Introduction

1.1 Gender Bias in medical research

Gender bias[1] in medicine has long been a serious, but overlooked, issue, leading to critical consequences for women’s health. Historically, medical research has been focused on the male body as the default, leading to misconceptions of the female body and underrepresentation of female subjects in clinical trials. This phenomenon creates a gender data gap that may affect drug development, disease understandings and treatment protocols. This chapter focuses on how gender bias in medicine can lead to misdiagnoses, ineffective treatments and increased health risks.

The historical view of male bodies as the standard, disregarding the female body greatly affects female health. Researchers tend to exclude females from studies or underrepresent them, leading to insufficient data based on gender. This phenomenon is being cultivated and reproduced using two main arguments: The first one consists of the notion that the female body is just a deviation from the male body (only a few kilograms less), failing to address the differences in metabolization of drugs, body structure, hormonal influences, social roles etc. The second argument is related to the menstrual cycle and the hormones produced by it. Researchers suggest that females should participate in clinical trials only when they are in the early follicular phase of their menstrual cycle because that’s when their hormone levels are at their lowest, impacting the drug effects less. In addition to that, pregnancy is almost entirely excluded from medical research. But in real life, pregnancies and menstrual cycles exist throughout all of their phases, thus, leaving the results of trials partially not reliable. *(Criado Perez, 2019)*

[1] The terms 'gender' and 'sex' have distinct meanings, with 'sex' referring to biological characteristics and 'gender' encompassing social and cultural roles. However, in this thesis, these terms will be used interchangeably, following the convention of many biomedical studies that refer to sex-based biological differences as 'gender differences' in gene expression.

1.1.1 Exclusion and underrepresentation of females

This systemic bias against the female anatomy has led to the exclusion or underrepresentation of females in clinical studies, trials and even textbooks *(Plataforma SINC, 2008)*. Numerous examples can be found where females were either underrepresented or excluded from studies, specifically in studies conducted before the 80s, when the problem of gender bias was first assessed. Some large and significant studies from before 1980 that were the object of critique included “(a) the Physicians’ Health Study of the effects of aspirin on cardiovascular disease, in which 22,071 men and 0 women physicians were enrolled; (b) the Multiple Risk Factor Intervention Trial (MRFIT), a randomized trial conducted from 1973 to 1982 to evaluate correlations among blood pressure, smoking, cholesterol, and coronary heart disease in 12,866 men and 0 women; and (c) the National Institute on Aging’s Baltimore Longitudinal Study of Aging, extending from 1958 to 1975 , which excluded female subjects, despite the fact that women constitute two-thirds of the population over age 65”. *(Schiebinger, 2003)*

A more modern example of underrepresentation of females in studies is a 2016 review of the inclusion of women in US HIV research. It found that females made up 19.2% of participants in antiretroviral studies, 38.1% in vaccination studies and 11.1% in studies to find a cure *(Curno et al, 2016)*. Another example is that the first research into cardiovascular diseases was conducted only on males, and females continue to be underrepresented, making up 25% of participants across 31 landmark trials for congestive heart failure *(Cristiana Vitale et al., 2017).* In cases where studies do include female participants, the results are often not sex-disaggregated, impeding the ability of conducting meta-analyses. *(Criado Perez, 2019, p. 179; Beery and Zucker, 2011)*

This underrepresentation can manifest itself even in studies with animals or cells. A 2007 paper found that 90% of pharmacological articles described male-only studies. *(Criado Perez, 2019, p. 177)*. A 2014 paper found that 22% of studies on animals did not specify sex, and those who did, included 80% only males *(Yoon et al., 2014)*. As for cell studies, a 2011 review of 10 cardiovascular journals found that when sex was specified in the studies, 69% reported using only male cells *(Yoon et al., 2014)*. A 2007 analysis of 645 cardiovascular clinical trials found that only 24% reported sex-specific results *(Yoon et al., 2014)*

1.1.2 Consequences of gender bias in medical research

The exclusion or underrepresentation of females in studies overlooks the massive differences the female body has from the male. Aside from the obvious differences in reproductive organs, the menstrual cycle and pregnancy, research has found differences in “every tissue and organ system in the human body” *(Marts and Keitt, 2004, cited by Criado Perez, 2019, p. 171)*. For example, differences can be found in the metabolization process of drugs, the mechanical workings of the heart, lung capacity, the immune system and even in our cells. Females also experience different symptoms in cardiovascular diseases, HIV, Parkinson's and behavioral disorders like ADHD, autism and Asperger's syndrome. *(Criado Perez, 2019)*

The consequences of gender bias in medicine are a serious issue which should not be underestimated. Due to the lack of data, women are often prescribed medication without it being appropriately researched on women in terms of dosage, efficacy and side effects. This can lead to the ineffectiveness of drugs or adverse drug reactions (ADRs), with both of them being potentially fatal. For example, drugs that are often prescribed for high blood pressure, may lower men’s mortality from heart attack but increase cardiac-related deaths for women *(Schiebinger, 2003)*. Another indicator of the seriousness of this issue is that in 2014 the FDA (Food and Drug Administration) released data on ADR reports showing that between 2004-2013 more than 2 million women experienced ADR’s compared to less than 1.3 million men *(Keating and Millman, 2014)*. The exclusion of participants due to their menstrual cycle has also been shown to harm the efficacy of drugs. Menstrual cycle impacts have been found in antipsychotics, antihistamines, antibiotic treatments and heart medication *(Zopf et al., 2008)*. Even medical devices and medical advice and guidelines can be negatively impacted by male-centered research *(Criado Perez, 2019, p. 181)*.

Another direct outcome is that gender bias leads to poor understanding and interpretation of symptoms which can cause misdiagnoses in women. Oftentimes treatment is delayed or denied in women, with their symptoms being dismissed as psychological or simply not identified. A notable example of how dangerous these differences in symptoms are, is the heart attack: Women often don't have the "classic" heart attack symptoms like chest and left-arm pains but rather have symptoms like stomach pain, breathlessness, nausea and fatigue *(Khamis, Ammari and Mikhail, 2016, cited by Criado Perez, 2019; Lear, 2014)*. Even behavioral disorders may be left undiagnosed. For example, it was widely accepted that boys are about four times more likely to have autism *(Szalavitz, 2016)*. But a new study suggests that there are more girls living with autism than previously assumed, they just weren't diagnosed due to gender differences in the socialization of kids. *(Mifsud, 2016, cited by Criado Perez, 2019)*

1.1.3 Attempts and other biases

There have been attempts to include females in medical research. In 1993 the US passed the National Institute of Health Revitalization Act, which made it illegal to not include women in federally funded clinical trials, but this requirement did not come into effect until 2016 which is when the US introduced the requirement that the data in trials must be disaggregated and analyzed by sex. *(Office of Research on Women's Health, 2021; Schiebinger, 2003)*. Unfortunately, this does not apply to non NIH-funded research, and even NIH-funded research is not complying entirely with its own rules. In 1997, the US Government Accounting Office (GAO) released a report which criticized the NIH for having "no readily accessible source of data on the demographics of NIH study populations" making it hard to determine whether the NIH was enforcing its own rules or not. *(Criado Perez, 2019, p. 183)*

Biases in medicine can also appear in other terms, for example in racial, socioeconomic and age-related forms (Vitale et al., 2017). These overlapping biases often amplify the effects of the data gap, producing even more significant consequences. For example, in the US African-American women are 243% more likely than white women to die from pregnancy or childbirth-related issues. *(Tucker et al., 2007, cited in Criado Perez, 2019, p. 201)*

1.1.4 Conclusion

Gender bias remains a significant problem in medical research with serious consequences. The exclusion of females in studies, the focus on the male body as the default and the failure to account for sex-specific differences in disease presentation and drug metabolism all contribute to misdiagnoses, ineffective treatments and higher mortality rates in women. Despite the issue being addressed for many years now, institutions have been slow to adjust their guidelines in order to prevent it. To close the gender data gap a shift towards gender inclusive studies and sex-disaggregated data analysis is required. Medical research must adopt statistical methodologies that can account for these variations. Only by acknowledging and actively addressing the issue, the medical community can move towards a more inclusive and effective healthcare system, suitable for all.

1.2 Gender bias and gene expression

As previously mentioned, sex differences can be found in almost every tissue and organ system of the human body. Gene expression could not possibly remain unaffected by these differences. Gene expression studies provide significant molecular-level insights on how diseases may manifest themselves differently in males and females. They can reveal sex-specific biological variations that can impact disease susceptibility, drug responses and treatment efficacy. By applying statistical analyses on gene expression datasets, identifying differentially expressed genes between male and female samples is made possible, providing quantitative proof of those biological sex differences. Gene expression analysis can also address the question of “which are the genes that are differentially expressed between males and females”, providing insight on the relationship of these genes and known variations in disease prevalence and response to treatments. *(Darolti and Mank, 2023)*

Overlooking these differences by not sex-disaggregating the results of biomedical studies, contributes to misdiagnoses, ineffective treatments and unrecognized risk factors. Unfortunately, many large-scale genomic datasets, such as those used in this thesis from the Gene Expression Omnibus (GEO), often do not have their data analyzed, or even categorized by gender. An indicator of this lack of categorization is the fact than when searching GEO for suitable datasets for this thesis, only 0.241% (74 out of 30702) were found to have gender as a subset variable type in “homo sapiens” and “expression profiling by array” datasets.

2. Theoretical Framework

2.1 Bias

2.1.1 Statistical Bias

In statistics, *bias* is a process which produces results that differ systematically from the truth, leading to the reduced accuracy of estimates or conclusions. It is caused by flaws in data collection, measurement, analysis, or assumptions and can happen at any stage of inference. Among analytic clinical studies, there are four main categories of bias: *selection bias, classification bias, confounding bias, and publication bias.*

Selection bias occurs when the sample chosen for analysis is not representative of the population, often due to systematic differences in how individuals are chosen for the study. For example, let’s suppose a study on the effects of a drug excludes patients who dropped out, possibly due to the side effects. The results will be skewed towards showing that the drug is safer than it actually is.

Classification bias, also called measurement bias, occurs when there are errors in the categorization or measurement of variables. For example, in a study on the effects of smoking on lung cancer, participants who smoke occasionally are misclassified as non-smokers, leading to underestimation of the results of smoking on lung cancer.

Confounding bias occurs when the relationship between an exposure and an outcome is distorted by a third parameter (a confounder) that is associated with both the exposure and the outcome, making it difficult to determine the true effect of the exposure. For example, let’s assume a study finds that people who exercise more tend to have better mental health. However, income could be a confounder because individuals with higher income tend to have more access to gyms, trainers etc. and more free time to exercise, but also have better mental health due to reduced stress caused by financial problems. *(Lambert, 2011)*

Publication bias occurs when studies with positive or significant results are more likely to be published than those with negative or inconclusive findings. For example, suppose multiple studies research if a new drug is effective. If only the studies showing a significant benefit are published, it may be incorrectly assumed that the drug is highly effective.

2.1.2 Gender bias

More specifically, *gender bias* refers to the unequal representation of different genders in the research design, participant selection or data analysis, leading to results that may not be accurate for all genders. Gender bias can lead to misdiagnoses, ineffective treatments or missed health risks, particularly in underrepresented groups. It mostly manifests itself in the forms of classification bias, if gender is not even recorded as a variable that could influence the results, in the form of selection bias if some genders are underrepresented or excluded from the participant selection, without the study explicitly stating it, or in the form of publication bias if research related to gender differences is selectively published based on the nature of the findings. However, it can also arise in the form of confounding bias, if the gender itself is the confounding factor by influencing both the exposure and the outcome. Although gender bias is typically found in the form of selection or classification bias, this thesis is going to statistically analyze gender bias that arises in the form of confounding bias, exactly due to the lack of information regarding gender present in most studies.

2.2 Genes

2.2.1 Gene expression

Genes are sequences of nucleotides found in chromosomes, which are thread-like structures formed of DNA. Genes are the fundamental unit of heredity and are passed down from parents to offspring. Some genes serve as instructions for protein synthesis, while others regulate the activity of other genes.

Gene expression is the process by which a gene’s information is used in the synthesis of a functional gene product such as proteins or non-coding RNA. *Gene expression levels* quantify how much a gene is being expressed at a specific time through measuring the amount of mRNA produced. When a gene has higher expression levels than expected (i.e. compared to a baseline or the control levels) it is referred to as *upregulated*. Similarly, if a gene has lower expression levels than expected it is referred to as *downregulated*. If a gene is significantly upregulated or downregulated between two experimental conditions it is declared as *differentially expressed*. This thesis focuses on finding such differentially expressed genes (DEGs). *(MedlinePlus, 2025)*

2.2.2 Measurement of gene expression levels

There are many tools and technologies that can be utilized to measure gene expression levels, but the two most common are *Microarrays* and *RNA-sequencing* (RNA-seq). Although RNA-seq is a newer technology and has a lot of advantages (ability to detect novel transcripts, wider dynamic range, higher specificity and sensitivity etc.), microarrays were the standard technology used for gene expression profiling for many years. Therefore, the majority of widely accessible datasets (specifically GEO datasets) are in the form of microarray. *(Illumina, 2025)*

2.2.3 Microarrays

The Microarray method for measuring gene expression levels involves a glass or silicone chip with arrays in tiny spots of less than 200 micrometers in diameter, each spot containing a DNA probe complementary to a specific gene. DNA microarray analysis consists of the following steps:

1. Messenger RNA (mRNA) is extracted from the cells of interest and converted do complementary DNA (cDNA)
2. cDNA is labeled with fluorescent dye
3. The cDNA is applied to the microarray chip
4. When the fluorescent labeled cDNA binds to the complementary probes, the gene is shown to be active. The intensity of the fluorescence signal corresponds to the expression level of the gene *(Stoakes, 2019)*

2.2.4 Representation of gene expression levels

Gene expression levels are usually represented in the form of a *nxm* matrix, where n is the number of genes and m the number of samples. Each entry of the matrix represents the expression level of a specific gene for a specific sample. *(UCDavis Bioinformatics Training, 2019)*

2.2.5 Variance of gene expression

It is of interest to take a look at the variance of gene expression. Highly expressed genes often have a higher variance *(Griffiths et al., 2009)*. This phenomenon is often referred to as *heteroscedasticity*, which means that for a given expression level, a lot of variation in the amount of variance is observed *(Harvard Chan Bioinformatics Core, 2025).* This can cause statistical tests to have skewed variance estimates, especially when the sample size is small. For this reason, it is convenient to use the limma package of R which utilizes an *empirical Bayes* approach. The empirical Bayes model stabilizes variance estimates across genes to improve the reliability of differential expression analysis. More detailed information about this approach is going to be provided in chapters 3.4.7 and 3.4.8.

3. Methodology

3.1 Hypothesis testing

In this study, the basic goal is to determine whether there are statistically significant gender differences to be found in each one of the datasets used. The two competing hypotheses can be formulated as follows:

Null Hypothesis H_0 :

Gender has no significant effect on differential gene expression between control and disease.

Alternative Hypothesis H_1 :

Gender has a significant effect on differential gene expression between control and disease.

To examine these hypotheses, statistical tests will be conducted on gene expression data to compare results between genders. The criteria for statistical significance is set at a p-value threshold of 0.05. If the p-value is less than 0.05, the null hypothesis H_0 can be rejected providing evidence for the alternative hypothesis H_1. If the p-value is equal or greater than 0.05, the null hypothesis H_0 can not be rejected indicating that there is insufficient evidence for the alternative hypothesis H_1.

3.2 Method selection

Extracting biological information from microarray data, requires appropriate statistical methods. In this thesis, two of those are used and analyzed: the t-test and limma. Those two methods differ both in the approach and in the end results.

3.3 The t-test

The student's t-test is the most simple and commonly applied statistical method for finding differentially expressed genes. It requires more than one samples that are classified in two groups. Gene expression levels are then collected for each gene and every sample and the t-test compares directly the mean gene expression levels between the two groups.

For simplicity, an example for the t-test conducted for only one gene is generated. The same method can be applied to any number of genes, but multiple testing correction is crucial to control for false positives.

Suppose the gene expression values across two groups, e.g. control and treatment.

Let

$x_1, x_2, \ldots, x_n$ be the expression values for the gene in samples which belong to the control group and

$y_1, y_2, \ldots, y_m$ be the expression values of the gene in samples which belong to the treatment group.

It is of interest to determine if the compared groups have the same mean gene expression or not. Hypothesis testing is constructed as follows:

Null Hypothesis H_0 :

The mean expression of the gene is the same in both groups $\bar{x}=\bar{y}$

Alternative Hypothesis H_1 :

The mean expression of the gene differs between the groups $\bar{x}\neq\bar{y}$

The t-statistic is then calculated as:

$$t=\frac{\bar{x}-\bar{y}}{\sqrt{\frac{s_x^2}{n}+\frac{s_y^2}{m}}}$$

where $\bar{x},\bar{y}$ are the mean gene expressions, s_x^2, s_y^2 are the variances and *n, m* are the number of samples in control group and treatment group respectively.

Using the t-distribution table, the p-value can be obtained by comparing it to the significance interval which usually is 95%. A small p-value (<0.05) indicates that the null hypothesis can be rejected, meaning that the gene is statistically significantly differentially expressed.

Gene expression datasets often contain thousands of different genes, and therefore require a correction on the false positive values that may arise due to multiple testing. Some corrections that can be applied to adjust the p-value are the FDR (False Discovery Rate) and the Bonferroni correction. *(Cui and Churchill, 2003)*

3.4 The limma method

3.4.1 Introduction to limma

In contrast to the t-test, the *limma* (linear models for microarray data) method uses linear models to represent the relationship between gene expression and the parameters that may influence it. Limma fits a linear model for each gene across all samples using the formula:

$$Y = X\beta + \varepsilon$$

where Y is the matrix of gene expression levels (genes as rows and columns as samples), X is the design matrix which represents the experimental conditions (further information about the design matrix can be found in section 3.4.5), β are the parameters representing the effect of each condition and ε is the error term.

A separate linear model is fitted to each gene estimating β and the variance for each condition. The variance, which may differ vastly among genes, is stabilized through the empirical Bayes method and hypothesis tests are performed on the coefficients β. The resulting p-values are adjusted for multiple testing using the FDR method. More detailed information about the theoretical framework and the methods of limma are going to be presented in the next chapter. *(Law et al., 2020)*

3.4.2 Theoretical concepts of limma

Before going into a more detailed explanation of the theoretical framework and methods of limma, it is necessary to first clarify key terms and definitions. From now on, gene expression levels will often be referred to as *response variables*, while the factors influencing them will often be referred to as *explanatory variables*. Explanatory variables explain the variation in gene expression and the response variables respond to the effects of explanatory variables.

A *Statistical model* is used to describe the relationship between the explanatory variable(s) and the response variable. The behavior of the model is defined by the *model parameters* which are estimable numerical values which describe the direction and magnitude with which the explanatory variable(s) affect the response variable.

There are two types of explanatory variables: *covariates* and *factors*. Covariates contain continuous or discrete numerical values of quantitative measurements. For example, age could be a covariate with values 0.2, 8.4, 26, 44.5 and so forth, each number indicating the years a person has lived. Factors, on the other hand, contain discrete categorical values. Age could also be treated as a factor if each numerical value was classified as `young` or `old`. The unique values of a factor are referred to as *levels* of the factor. In this example the levels are `young` and `old`. For clarity, the levels of a factor will be presented in a different font, as demonstrated in the examples above.

Due to the structure of most medical research datasets, only factors will be used. Therefore, from this point forward, all explanatory variables will be assumed to be factors. In this case, it is of interest to determine the gene expression difference between levels. There are two basic models that accomplish this estimation: *the means model* and *the mean-reference model*. The means model calculates the mean gene expression of a level and the mean-reference model directly calculates the gene expression difference between levels. Below is a more detailed description of how those models work. *(Law et al., 2020)*

3.4.3 Means model

The means model describes the relationship between the explanatory variables and the response variable using the following model:

$$expression = \beta 1 * \texttt{level1} + \beta 2 * \texttt{level2}$$

where $\beta 1$ is the observed mean gene expression for `level1` and $\beta 2$ is the observed mean gene expression for `level2`. The levels take turns taking the values 1 or 0. When `level1` is equal to 1, `level2` is mandatorily equal to 0, and vice versa. That is because the categorization of samples is mutually exclusive, meaning that a sample can either be categorized as `level1` or `level2`. This way, when estimating the gene expression for `level1`, this gives:

$$expression = \beta 1$$

When estimating the gene expression for `level2`, this gives:

$$expression = \beta 2$$

Basically, this model returns the mean gene expression of each level independently.

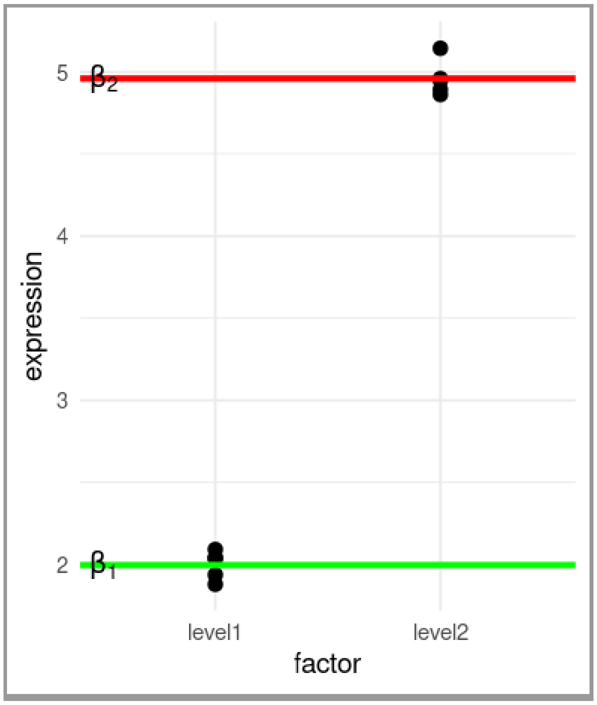


Figure 1: Means model for estimation of gene expression.
Generated using R.

Whilst the mean gene expression for each level may come in useful, it is usually of interest to calculate the gene expression difference between levels. Those differences can be calculated using linear combinations of the parameters, referred to as *contrasts*. For example, when estimating the gene expression difference between `level1` and `level2` the contrast to be used would be *(1,-1)*, which calculates *β1-β2*. Further information about the construction of contrasts is going to be presented in chapter 3.4.5. *(Law et al., 2020)*

3.4.4 Mean-reference model

The mean-reference model directly calculates the gene expression difference between levels using a specific level as a reference (hence the name mean-reference). Such a model is parameterized for the mean gene expression of the reference level and the rest of the levels are parameterized relative to the reference. It models the expected gene expression of each level as follows:

$$expression = \beta 1 + \beta 2 * \texttt{level2}$$

where $\beta 1$ is the observed mean gene expression for `level1` and $\beta 2$ is the observed difference in mean gene expressions between `level1` and `level2`. As in the means model, `level2` may only take values 0 (when estimating the mean for `level1`) or 1 (when estimating the mean difference between `level1` and `level2`). This way the gene expression of `level1` can be estimated as follows:

$$expression = \beta 1$$

When estimating the gene expression for `level2`, this gives :

$$expression = \beta 1 + \beta 2$$

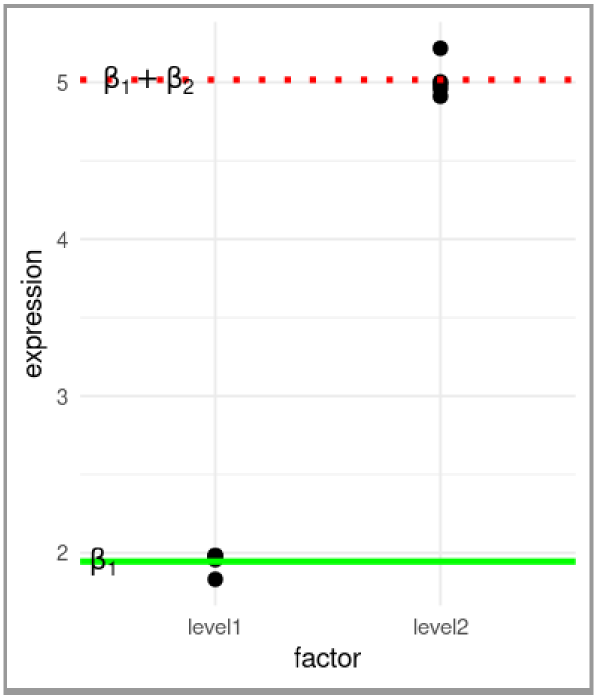


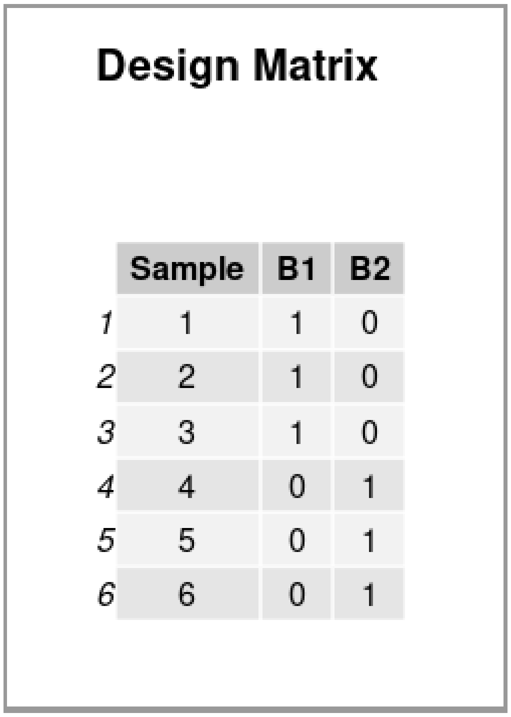
Figure 2: Mean-reference model for estimation of gene expression.
Generated using R.

These models are mathematically equal and yield the same results, but they differ in parametrization. It is up to each individual to determine which model to use. In the analysis of real data in this thesis, both models were used in order to confirm the validity of the process. *(Law et al., 2020)*

3.4.5 Design and contrast matrices

When coding for a means or a mean-reference model, design and contrast matrices are useful tools. In this section, the process of constructing those two matrices with and without an intercept term is going to be presented.

Design Matrices are a mathematical representation of the form of a statistical model, defining the relationships between samples. Each row of the design matrix is associated with a sample of the dataset. When working with a means model, each column of the design matrix is associated with the model parameters. An intercept term needs to be added when working with a mean-reference model in order to represent the mean expression for the reference group. In this case, each additional column of the matrix represents the effect relative to the reference group.

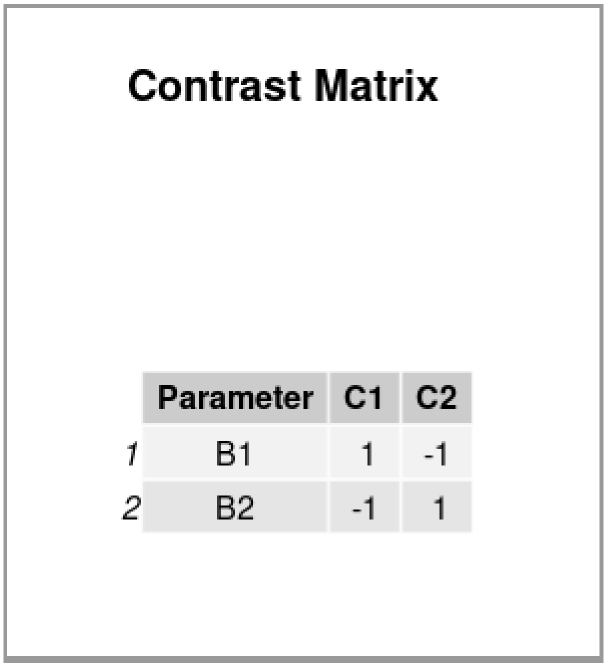


Design Matrix

	Sample	B1	B2
1	1	1	0
2	2	1	0
3	3	1	0
4	4	0	1
5	5	0	1
6	6	0	1

Figure 3: Example design matrix.
Generated using R.

Contrast matrices are used to specify the comparisons of interest in the model used, based on the design matrix. Each column of the contrast matrix corresponds to a comparison of interest and each row corresponds to a model parameter. When using a means model, the design matrix has no intercept term, therefore the group means are directly being modeled, and each contrast explicitly compares these groups. When working with a mean-reference model, the design matrix has an intercept term, therefore each contrast is relative to the reference group. *(Law et al., 2020)*



Contrast Matrix

	Parameter	C1	C2
1	B1	1	-1
2	B2	-1	1

Figure 4: Example contrast matrix.

Generated using R.

3.4.6 Additive vs. interaction effect in limma

In studies with multiple factors it is of importance to presume the relationship between factors that influence gene expression. An *additive model* is used when factors are assumed to have independent effects on gene expression, whereas an *interactive model* is applied when one factor is believed to modify the effect of another. The limma package enables explicit testing for interaction effects, allowing for the determination of the most appropriate model. Based on the research question and findings of this thesis, it was deemed appropriate to assume an interaction between the factors.

Additive effect assumes that the influence of multiple factors on gene expression is independent of each other. The total effect is the sum of each individual effect. Using a means model, the assumptions of the additive effect can be stated as follows:

$$expression = \beta 1 * \texttt{level1} + \beta 2 * \texttt{level2}$$

where *β1* is the effect of `level1` and *β2* is the effect of `level2`.

Interaction effect assumes that the level of influence of a factor is dependent on the level of another factor. The total effect is no longer the sum of each individual effect. Using a means model, the assumptions of the interaction effect can be stated as follows:

$$expression = \beta 1 * \texttt{level1} + \beta 2 * \texttt{level2} + \beta 3 * (\texttt{level1}\ X\ \texttt{level2})$$

where *β1* is the effect of `level1`, *β2* is the effect of `level2`, and *β3* is the interaction term, which captures the combined effects of `level1` and `level2`. *(Law et al., 2020)*

3.4.7 Empirical Bayes

The *empirical Bayes* approach is mostly used when dealing with a large number of observed values, as in the data of this thesis, which typically contains over 20,000 genes and 20 samples. The key difference between an empirical Bayes approach and other Bayesian models lies within the estimation of the prior probability distribution. Empirical Bayes estimates the prior distribution from the observed data, in contrast to standard Bayesian methods for which the prior distribution is fixed before any data is observed.

The empirical Bayes method shares information across observed values in order to get a better estimate of the prior distribution, especially when the number of samples is small. Basically, the method can be summarized to the following two steps:

1. Estimation of the overall distribution of the data.
2. Using this distribution as a prior to estimate each average.

3.4.8 Gaussian example of empirical Bayes

Below is a Guassian example of the empirical Bayes method, which aims to simplify the understanding of the approach:

Suppose a dataset with p genes and n number of samples (observations of each gene). Let x_{ij} be the *ith* sample of the *jth* gene. The goal is to estimate the mean gene expression of each gene, $\mu_j = \mathbb{E}[x_{ij}]$ for $j=1,\ldots,p$

Consider the following two-step hierarchical model for this dataset:

$$x_{ij} \sim N(\mu_j, \sigma^2), \quad \mu_j \sim N(\mu_0, \sigma_0^2),$$

where μ_0 and σ_0^2 are the parameters of the prior distribution. For simplicity, let $\sigma^2 = 1.$

Instead of specifying μ_0 and σ_0^2 based on prior knowledge, they are estimated using the available data.

Specifically:

- The sample means $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ are computed for each gene.
- The observed sample means and their variance across features are used to estimate the overall mean μ_0 and the variance of the prior distribution σ_0^2 .

Then, the posterior for μ_j is:

$$\mu_j | x_{1j}, \ldots, x_{nj} \sim N\left(\frac{n\bar{x}_j + \mu_0}{n + 1/\sigma_0^2}, \frac{1}{n + 1/\sigma_0^2}\right),$$

where μ_0 and σ_0^2 are derived empirically from the data itself. *(Casella, 1985)*

3.5 T-test vs. limma

Initially, the t-test was chosen as the differential expression analysis method; however, it was soon disregarded after a comparison with limma, which was found to be significantly superior.

As previously mentioned, gene expression variance differs vastly among low and high expressed genes. The t-test estimates variances for each gene individually which can lead to skewed variance estimates, especially when the sample size is small. Limma on the other hand, uses the empirical Bayes approach which shrinks the individual gene variances (towards a common value, thus stabilizing the estimates.

Another factor which influenced the decision is the scale of the used datasets. High-throughput data (microarrays, RNA-seq) requires testing thousands of genes simultaneously. The t-test approach applies a test to each gene individually, followed by a post-hoc correction for multiple testing, while limma uses a linear model framework, which handles multiple comparisons at once. Limma also provides a better control over false positives by adjusting p-values for multiple testing.

When there are more than two groups in an experiment, which was mostly the case for the used datasets, a more flexible design is needed. The t-test is limited to comparing two groups directly, while limma has the ability to handle more complex experimental designs using a design matrix to incorporate each factor to the linear model.

Another factor of interest is the statistical power of tests. The t-test relies only on the data of each individual gene, while limma shares information across genes. Therefore, especially when the sample size is low, limma has more statistical power than the t-test.

Several other factors contribute to limma's superiority, which will not be further discussed, including its preprocessing and normalization capabilities, adaptability to other forms of data (such as RNA-seq), its outputs, and the tools it provides for visualization. *(Jeanmougin et al., 2010), (Smyth, 2004)*

4. Application

4.1 Example dataset

4.1.1 Setup of example data

For easier understanding of the model applied, a virtual matrix was set up, representing real life data. The virtual matrix created, contains values of numbers chosen from a normal distribution with mean 100 and a standard deviation of 10 ordered in 100 rows and 40 columns, each row representing a gene, and each column representing a sample.

Two factors were also created for this example: gender and treatment, gender taking values `female` and `male` and treatment taking values `control` and `treatment`. The factors were set up in such a way that the first 20 samples are `female` and the following 20 samples are `male`. As for the factor treatment, samples 1 to 10 and 21 to 30 are `control` samples and samples 11 to 20 and 31 to 40 are `treatment` samples. In R, this can be achieved by creating the vector variables gender and treatment as factors, with their levels specified.

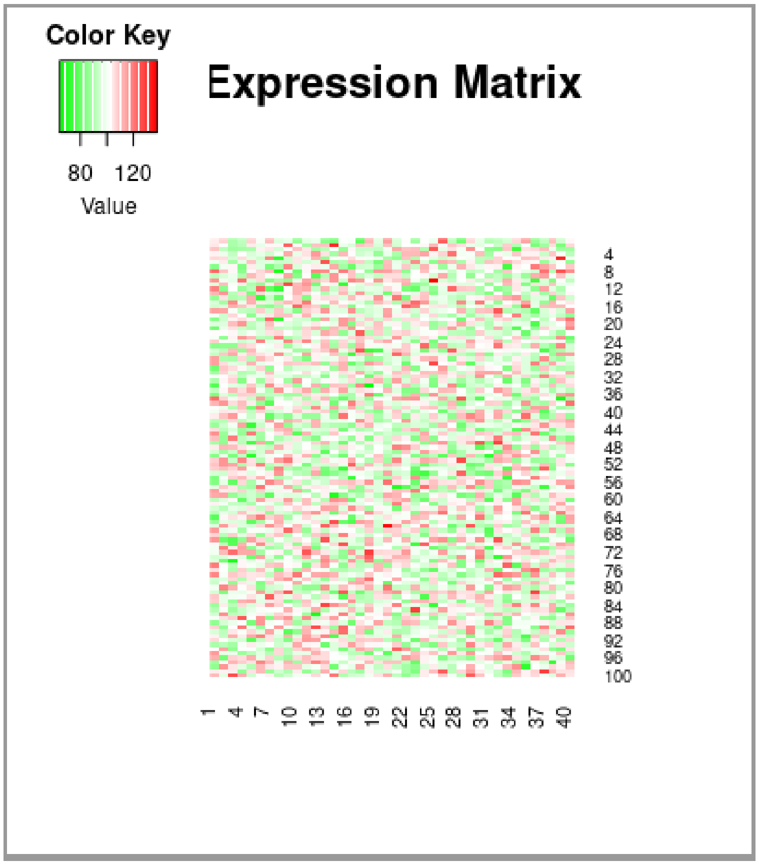


Figure 5: Example expression matrix.

Generated using R.

To accurately simulate real data, which doesn't follow a normal distribution strictly, following adjustments were made to the matrix:

- Values for genes in row 20 to 60 for samples in columns 11 to 20 (`female treatment`) were upregulated by 10.
- All values for genes in columns 21 to 40 (`male`) were upregulated by 10.

- Values for genes in row 40 to 80 and for samples in columns 31 to 40 (`male treatment`) were upregulated by another 25.

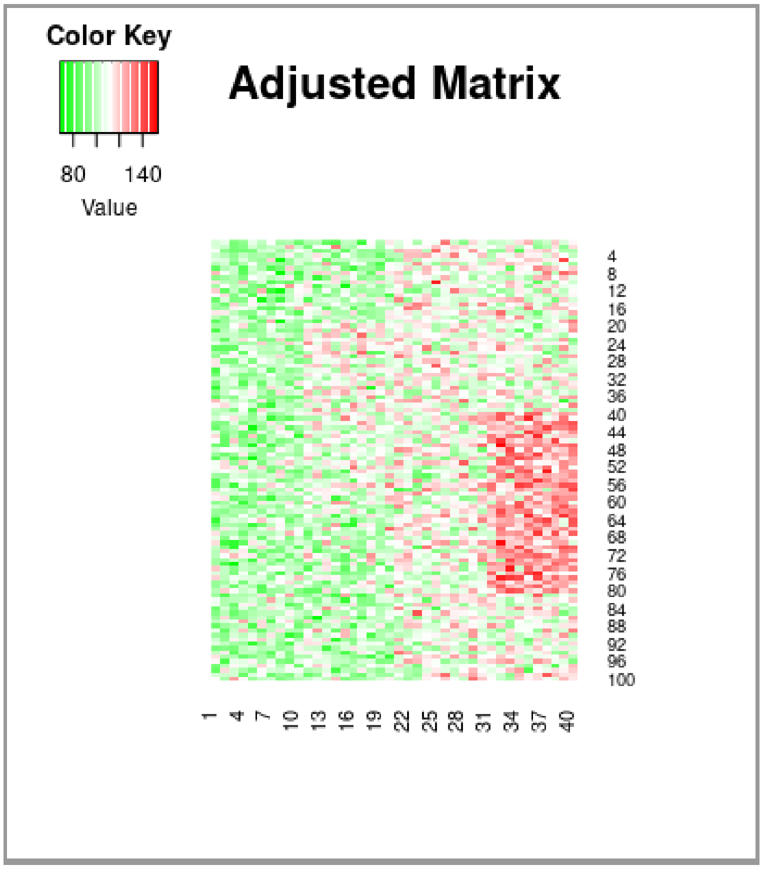


Figure 6: Adjusted example expression matrix.

Generated using R.

These tweaks have the following results in the example data:

- `Female control` samples have all genes expressed at about 100.
- `Females` who received `treatment` have genes 20 to 60 expressed at about 110.
- All `males` have all genes expressed at about 110.
- `Males` who received `treatment` have genes 40 to 80 expressed at about 135.

This way, treatment affects `male` samples differently in two ways. Firstly, by upregulating specific genes more than in `female` samples and secondly, by upregulating different genes than those in `female` samples. The way the virtual data was set up, should lead to the result that when investigating the effect of gender on treatment, the most differentially expressed genes come from rows 60 to 80.

4.1.2 Application on example

Differentially expressed genes of the created data can be detected by applying a means model or a mean-reference model using limma. This process involves the following steps:

1. A design matrix is constructed in order to model the effects of gender and treatment on gene expression.
2. A contrast matrix is constructed to specify the comparisons of interest.
3. A linear model is applied to each gene to estimate the effects of gender and treatment.
4. Empirical Bayes adjustment is used in order to improve variance estimation across genes.
5. Statistically significant differentially expressed genes are extracted from the results.

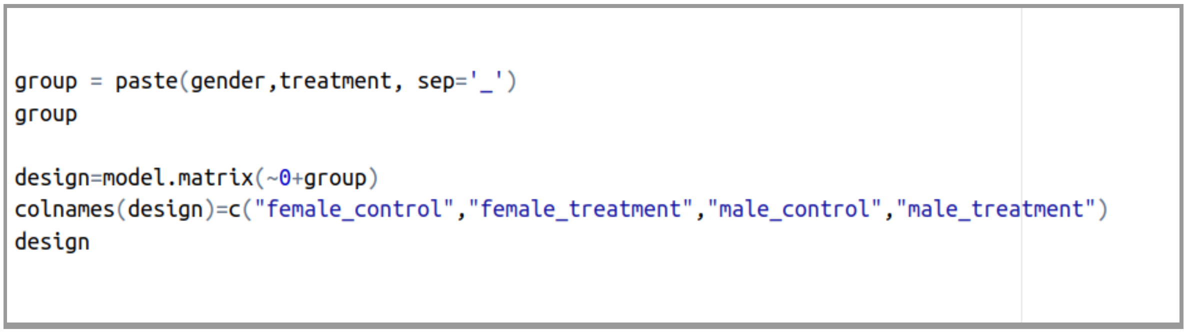
4.1.3 Differences between means and mean-reference model

Both the means model and the mean-reference model are going to be applied, to demonstrate that they both yield the same results and that either one of those can be of use. There are a few small but significant differences between the two models.

The first difference is whether the grouping of factors is required. When using a means model it is helpful to group the factors, i.e. create a new list variable consisting of characters that each represent a sample and specify the factor levels of this sample, which is going to be used when setting up the design matrix. The means model treats each different category as a separate entity. Grouping the factors allows us to indirectly capture interaction effects in order to ensure that all unique combinations are represented. When using a mean-reference model this grouping of factors isn't necessary because the interaction term automatically captures the interaction effect.

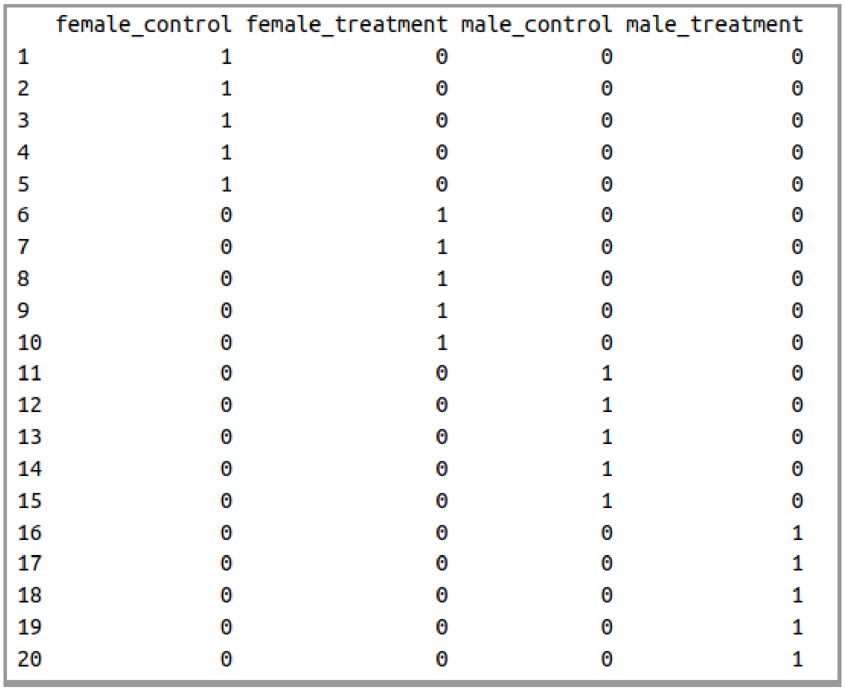
The second distinction lies within the construction of the design matrix. As previously explained, the mean-reference model uses an intercept term, while the means model does not. When coding the design matrix for the means model, the previously grouped factors can be used directly, and each column of the design matrix represents a group category (~0 + group). In contrast, the mean-reference model uses the factors directly and sets one factor level as a reference (+ gender * treatment). In this case, the first column of the design matrix represents the reference level, the next columns represent the individual level deviation from the reference level and the last ones represent the difference between the interaction effect of levels and the reference level. The interaction effect of levels is denoted with ":" (e.g. `male:treatment`). Below is the code used in order to set up the design matrices and their R console outputs, when using a means model and a mean-reference model, respectively.

Means model



```
group = paste(gender,treatment, sep='_')
group

design=model.matrix(~0+group)
colnames(design)=c("female_control","female_treatment","male_control","male_treatment")
design
```

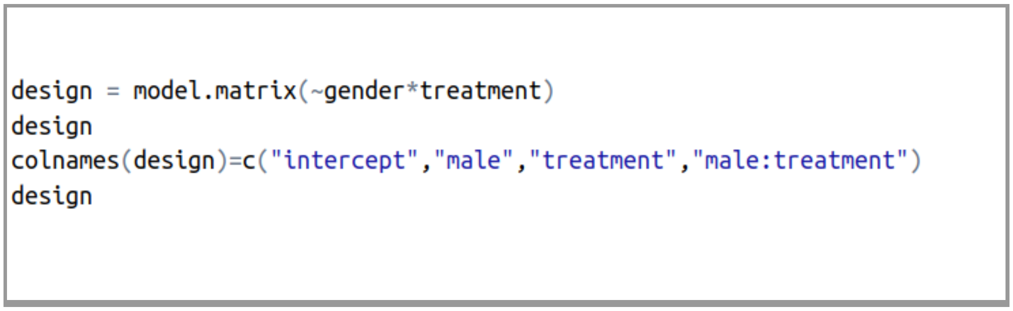
Figure 7: R code for grouping of factors and construction of design matrix using a means model.



```
   female_control female_treatment male_control male_treatment
1               1                0            0              0
2               1                0            0              0
3               1                0            0              0
4               1                0            0              0
5               1                0            0              0
6               0                1            0              0
7               0                1            0              0
8               0                1            0              0
9               0                1            0              0
10              0                1            0              0
11              0                0            1              0
12              0                0            1              0
13              0                0            1              0
14              0                0            1              0
15              0                0            1              0
16              0                0            0              1
17              0                0            0              1
18              0                0            0              1
19              0                0            0              1
20              0                0            0              1
```

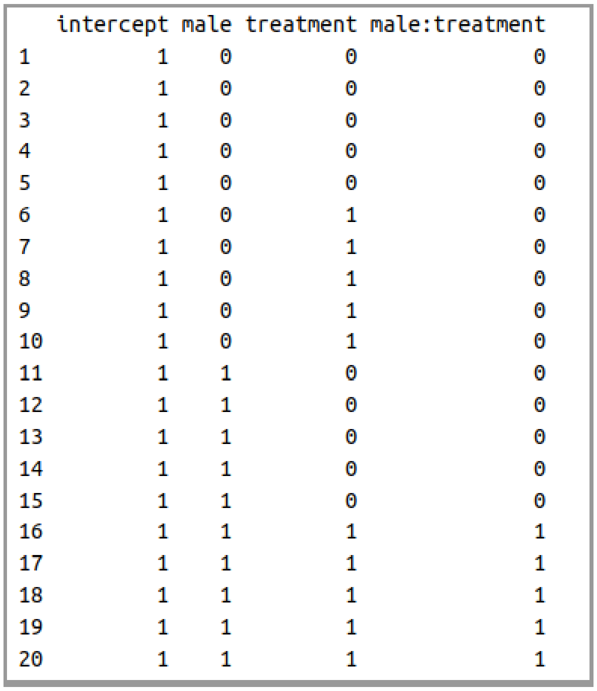
Figure 8: R console output for design matrix of 20 samples using a means model.

Mean-reference model:



```
design = model.matrix(~gender*treatment)
design
colnames(design)=c("intercept","male","treatment","male:treatment")
design
```

Figure 9: R code for construction of design matrix using a mean-reference model.



```
   intercept male treatment male:treatment
1          1    0         0              0
2          1    0         0              0
3          1    0         0              0
4          1    0         0              0
5          1    0         0              0
6          1    0         1              0
7          1    0         1              0
8          1    0         1              0
9          1    0         1              0
10         1    0         1              0
11         1    1         0              0
12         1    1         0              0
13         1    1         0              0
14         1    1         0              0
15         1    1         0              0
16         1    1         1              1
17         1    1         1              1
18         1    1         1              1
19         1    1         1              1
20         1    1         1              1
```

Figure 10: R console output for design matrix of 20 samples using a mean-reference model.

The last difference falls in the range of setting up the contrast matrix. When using a means model, comparisons of interest can be directly modeled using subtractions between groups (e.g. `female_treatment - female_control`). Using a mean-reference model, each group already represents a contrast in relation to the reference level, therefore effects can be extracted directly (e.g. `male:treatment`).

4.1.4 Interaction effect

When using the mean-reference model, each column of the design matrix in the example captures:

- Intercept: the `female control` (no `treatment`) group
- Male: the difference of `males` from `females` (if there was no interaction between gender and treatment)
- Treatment: the effect of `treatment`

- Male:treatment : how treatment and gender interact

Effectively, the Male:treatment models how the `treatment` effect differs for `males` compared to `females`.

Therefore, as described in chapter 3.4.4, the mean gene expression for each individual group can be modeled as follows:

- `Female control` = Intercept
- `Female treatment` = Intercept + Treatment
- `Male control` = Intercept + Male
- `Male treatment` = Intercept + Treatment + Male + Male:Treatment

When comparing the effect of treatment between `females` and `males`, this gives:

```
( Female Control - Female Treatment ) - (Male Control - Male Treatment)
                              =
          Intercept - Intercept - Treatment - Intercept - Male
            + Intercept + Treatment + Male + Male:Treatment
                              =
                        Male:Treatment
```

Therefore, the interaction effect “male:treatment” can be directly used in the contrast matrix.

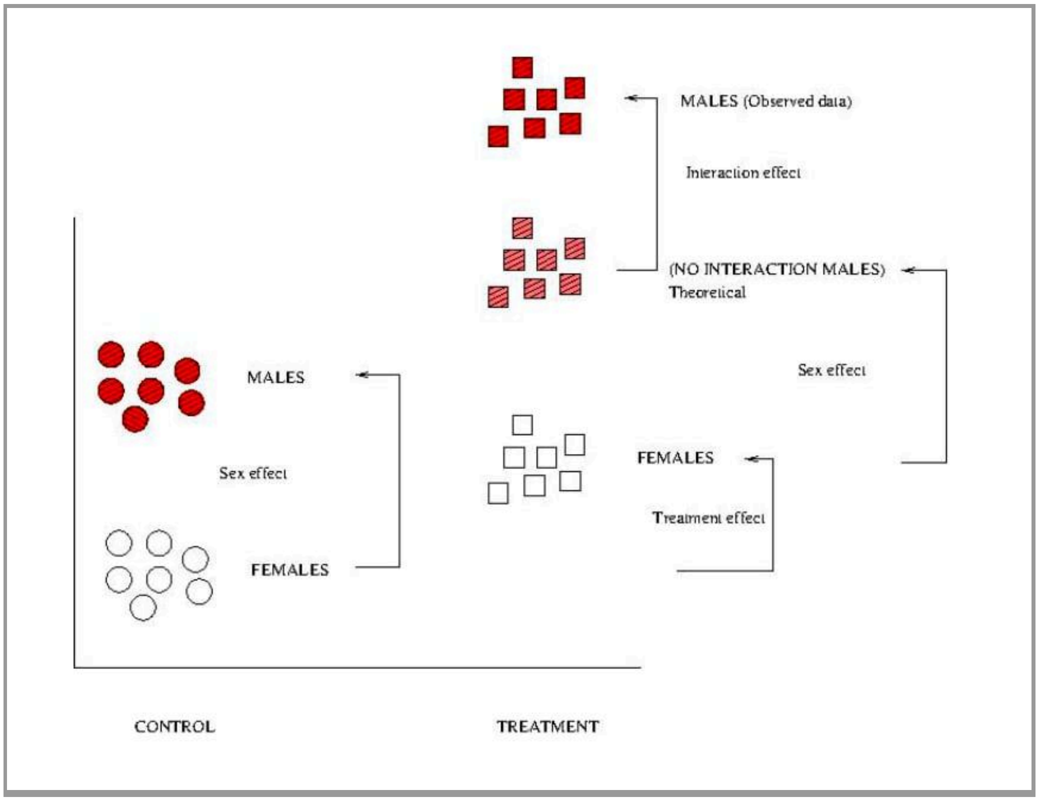
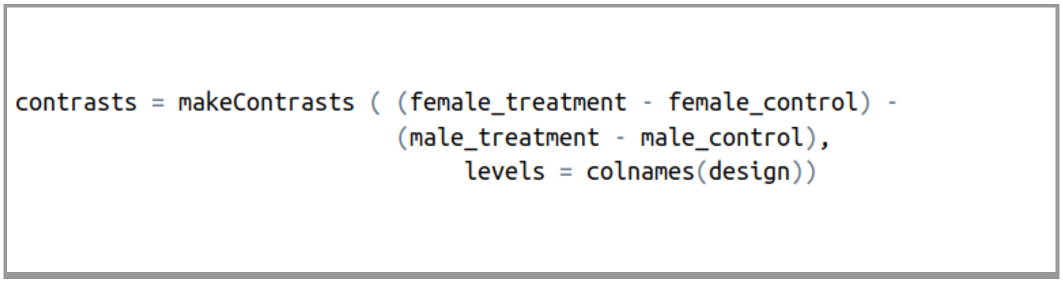


Figure 11: Interaction effect.

4.1.5 Contrasts and linear fitting for each model

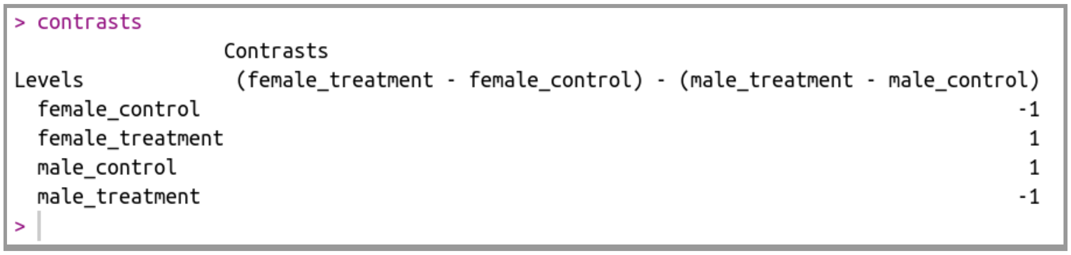
In order to test whether the treatment affects `female` samples differently than `male` samples, distinct comparisons were specified with each model used. After setting up the contrasts, linear models were applied using the `lmFit` and the `contrasts.fit` functions of limma. The results were then corrected using the empirical Bayes correction. Below is the code used in order to construct the contrast matrices, their R console outputs, the code used for applying the linear models and the topTable outputs of the results, when using a means model and a mean-reference model, respectively.

Means Model:



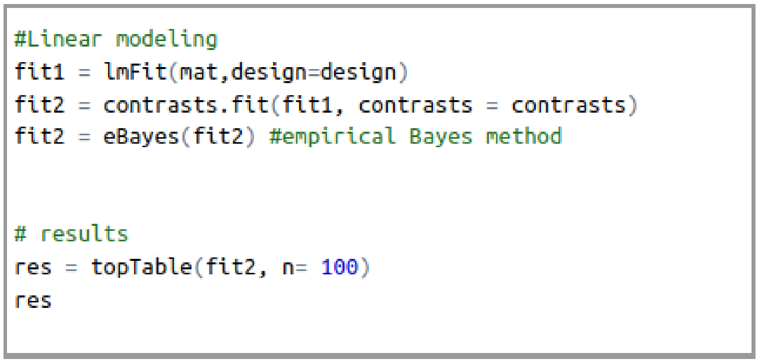
```
contrasts = makeContrasts ( (female_treatment - female_control) -
                            (male_treatment - male_control),
                                 levels = colnames(design))
```

Figure 12: R code for construction of contrast matrix using a means model.



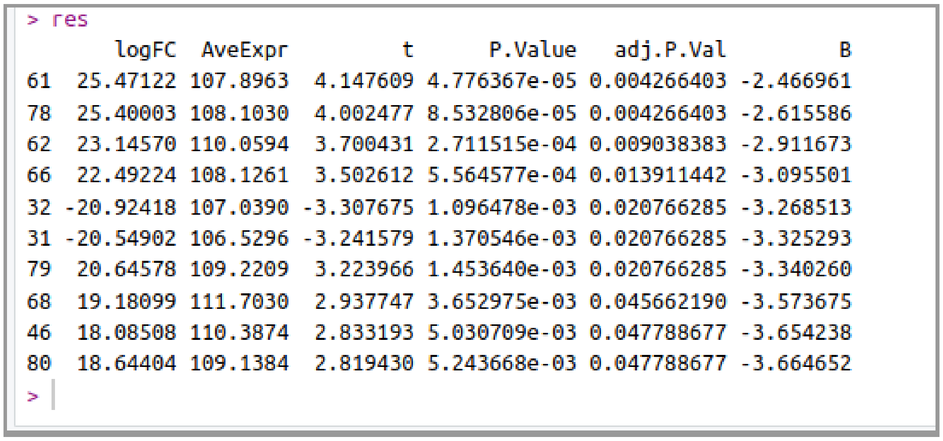
```
> contrasts
                  Contrasts
Levels             (female_treatment - female_control) - (male_treatment - male_control)
  female_control                                                                      -1
  female_treatment                                                                     1
  male_control                                                                         1
  male_treatment                                                                      -1
> 
```

Figure 13: R console output of contrast matrix using a means model.



```
#Linear modeling
fit1 = lmFit(mat,design=design)
fit2 = contrasts.fit(fit1, contrasts = contrasts)
fit2 = eBayes(fit2) #empirical Bayes method


# results
res = topTable(fit2, n= 100)
res
```

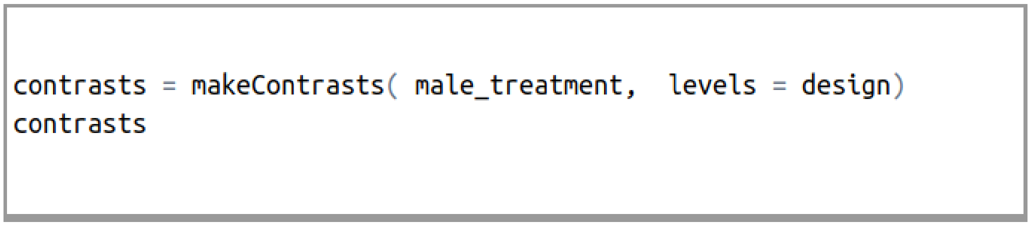
Figure 14: R code for fitting the linear models using a means model.



```
> res
        logFC  AveExpr         t      P.Value   adj.P.Val         B
61  25.47122 107.8963  4.147609 4.776367e-05 0.004266403 -2.466961
78  25.40003 108.1030  4.002477 8.532806e-05 0.004266403 -2.615586
62  23.14570 110.0594  3.700431 2.711515e-04 0.009038383 -2.911673
66  22.49224 108.1261  3.502612 5.564577e-04 0.013911442 -3.095501
32 -20.92418 107.0390 -3.307675 1.096478e-03 0.020766285 -3.268513
31 -20.54902 106.5296 -3.241579 1.370546e-03 0.020766285 -3.325293
79  20.64578 109.2209  3.223966 1.453640e-03 0.020766285 -3.340260
68  19.18099 111.7030  2.937747 3.652975e-03 0.045662190 -3.573675
46  18.08508 110.3874  2.833193 5.030709e-03 0.047788677 -3.654238
80  18.64404 109.1384  2.819430 5.243668e-03 0.047788677 -3.664652
> 
```

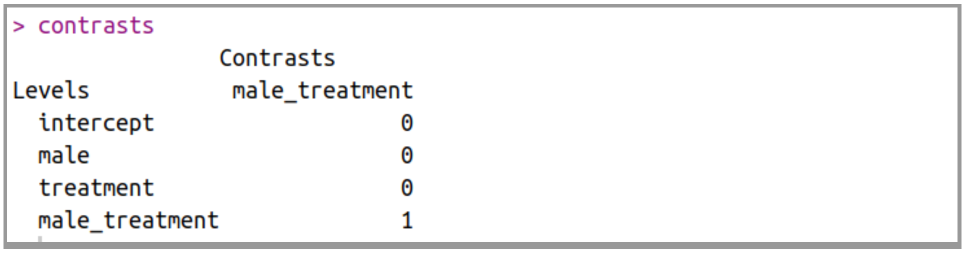
Figure 15: R console output of topTable of fitted results using a means model.

Mean-reference Model:



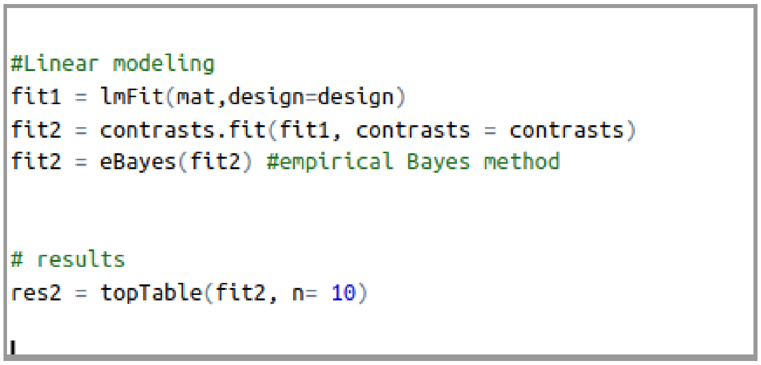
```
contrasts = makeContrasts( male_treatment,  levels = design)
contrasts
```

Figure 16: R code for construction of contrast matrix using a mean-reference model.



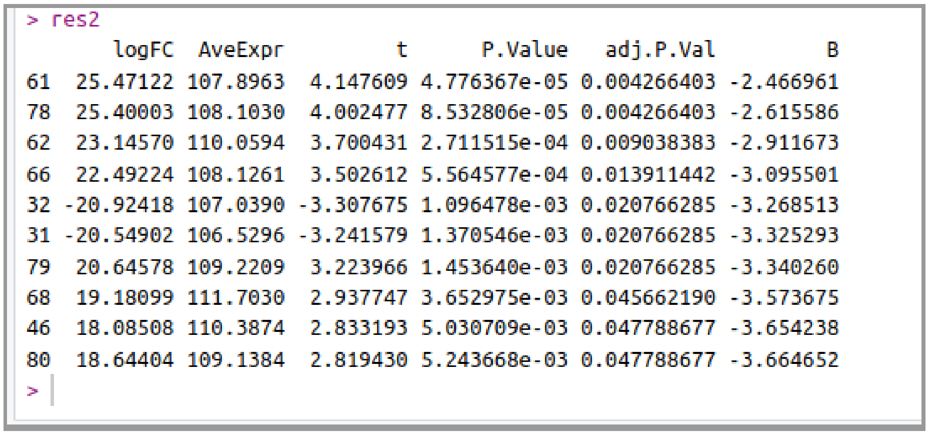
```
> contrasts
                Contrasts
Levels            male_treatment
  intercept                    0
  male                         0
  treatment                    0
  male_treatment               1
```

Figure 17: R console output of contrast matrix using a mean-reference model.



```
#Linear modeling
fit1 = lmFit(mat,design=design)
fit2 = contrasts.fit(fit1, contrasts = contrasts)
fit2 = eBayes(fit2) #empirical Bayes method


# results
res2 = topTable(fit2, n= 10)
```

Figure 18: R code for fitting the linear models using a means model.



```
> res2
       logFC  AveExpr         t      P.Value   adj.P.Val         B
61  25.47122 107.8963  4.147609 4.776367e-05 0.004266403 -2.466961
78  25.40003 108.1030  4.002477 8.532806e-05 0.004266403 -2.615586
62  23.14570 110.0594  3.700431 2.711515e-04 0.009038383 -2.911673
66  22.49224 108.1261  3.502612 5.564577e-04 0.013911442 -3.095501
32 -20.92418 107.0390 -3.307675 1.096478e-03 0.020766285 -3.268513
31 -20.54902 106.5296 -3.241579 1.370546e-03 0.020766285 -3.325293
79  20.64578 109.2209  3.223966 1.453640e-03 0.020766285 -3.340260
68  19.18099 111.7030  2.937747 3.652975e-03 0.045662190 -3.573675
46  18.08508 110.3874  2.833193 5.030709e-03 0.047788677 -3.654238
80  18.64404 109.1384  2.819430 5.243668e-03 0.047788677 -3.664652
> 
```

Figure 19: R console output for topTable of fitted results using a mean-reference model.

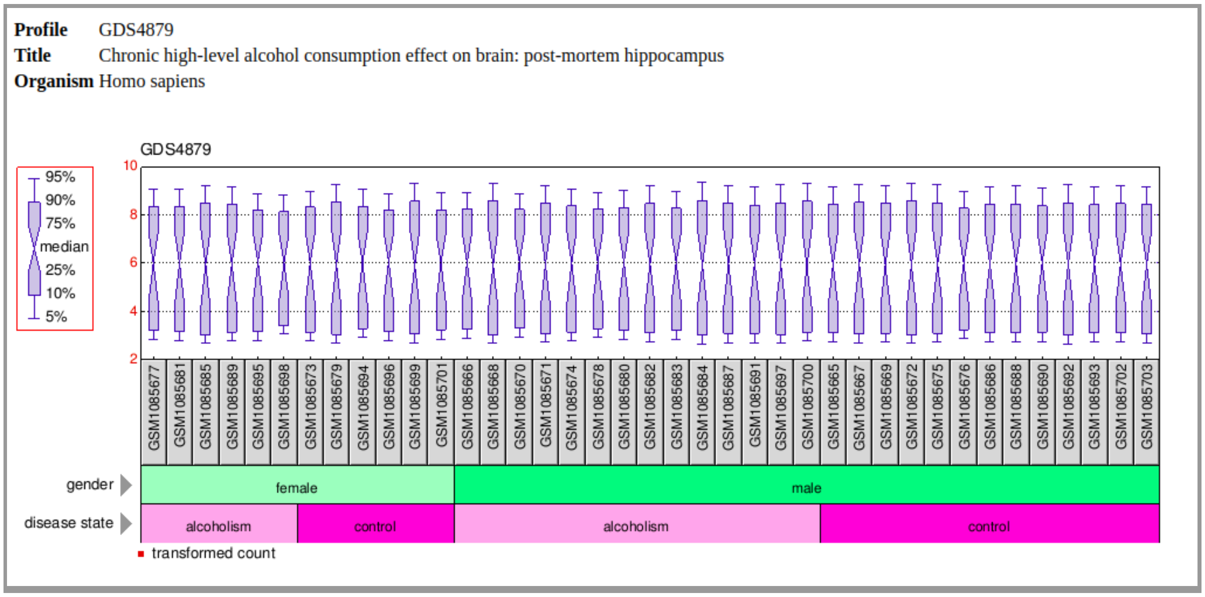
The “logFC” column in figures 15 and 19 both capture the interaction effect, i.e. how the `treatment` effect differs in `males` compared to `females`.

Using a threshold of adjusted p-value < 0.05, both comparisons yielded the same results. 15 genes were found to be differentially expressed, 11 of those stemming from rows 60 to 80, as previously speculated.

4.2 Datasets

This study analyzed gene expression microarray datasets of various diseases and conditions downloaded from the Gene Expression Omnibus (GEO). Microarray datasets in GEO often include multiple experimental conditions represented by subsets, for example disease state, age, gender etc. GEO is an international public database which archives and freely distributes comprehensive sets of microarray, next-generation sequencing and other forms of high-throughput functional genomic data, submitted by the scientific community. It is managed by the National Center for Biotechnology Information (NCBI) which is part of the National Library of Medicine (NLM), a branch of the National Institute of Health (NIH). *(NCBI, n.d.)*

The search query used to find suitable datasets for this study, included data for "homo sapiens", in the form of "expression profiling by array" (microarray) and with "gender" as a subset. This query gave an output of 74 results, from which 30 were selected. The 30 datasets were chosen based on relevance with this study and the least amount of complexity, because excessively complicated datasets with many factors could lead to unreliable results.



Profile GDS4879
Title Chronic high-level alcohol consumption effect on brain: post-mortem hippocampus
Organism Homo sapiens

Figure 20: Gene expression graph from GEO dataset (GDS4879, NCBI GEO). Source: https://www.ncbi.nlm.nih.gov/geo/tools/profileGraph.cgi?ID=GDS4879

Data downloaded from GEO is in the form of a .soft file (text), containing first various information about the dataset (name, date, institute, sample size, factors etc.) before revealing the data itself. The data is represented in the form of an *(n+1)x(m+2)* matrix, where n is the number of genes and m the number of samples. The first row contains the sample names and the first and second column represent ID_REF (Reference identifier) and IDENTIFIER (Gene identifier) respectively. Reference identifier is the probe ID used on the microarray platform and gene identifier maps the probe ID to a known gene or transcript, typically containing gene symbols or gene IDs. *(NCBI, n.d.)* Excluding the top row and the first two columns, each entry of the rest of the matrix represents the expression level of a specific gene for a specific sample.

Below is a list of the datasets used by this study. Links to each study will be provided in Appendix A. at the end of this thesis.

1. Cigarette smoke effect on the oral mucosa
2. Chronic high-level alcohol consumption effect on brain: post-mortem hippocampus
3. Type 2 diabetic and insulin-resistant but normoglycemic cohorts: skeletal muscle
4. Subcutaneous and visceral adipose tissue from lean or obese children

5. Postmortem Alzheimer's disease brains: Hisayama study
6. Non-syndromic cleft lip and palate: dental pulp stem cell cultures
7. Skeletal muscles from men and women of various ages
8. Uveal melanoma
9. Pituitary gonadotrope tumors
10. Bariatric surgery effect on obesity-related type 2 diabetes: whole blood
11. Hair follicles from males and females
12. Ulcerative colitis
13. Aging brain: frontal cortex expression profiles at various ages
14. Severe asthma: bronchial epithelial cell
15. Age effect on normal adult brain: frontal cortical region
16. Male and female in vitro cultured placental cell types
17. Systemic lupus erythematosus and arthritides: synovial biopsies
18. Schizophrenia: postmortem superior temporal cortex
19. Early and late onset colorectal cancers
20. Induced pluripotent stem cell-derived neurons from schizophrenia patients
21. Type 2 diabetic pancreatic islets of Langerhans
22. Type 2 diabetes and role of hepatokines
23. Morbidly obese insulin-resistant patients: omental and subcutaneous adipose tissue
24. Schizophrenia: postmortem anterior prefrontal cortex
25. Obesity-associated insulin resistance independent of BMI: omental and subcutaneous adipose tissues
26. Effect of lifestyle on Moroccan Amazighs: peripheral blood leukocytes
27. "Moderate stage Huntington's disease lymphocytes
28. Resistance exercise training effect on the skeletal muscle
29. Parkinson's disease: substantia nigra
30. Alzheimer blood mononuclear cells

4.3 Application on real data

4.3.1 Data preprocessing

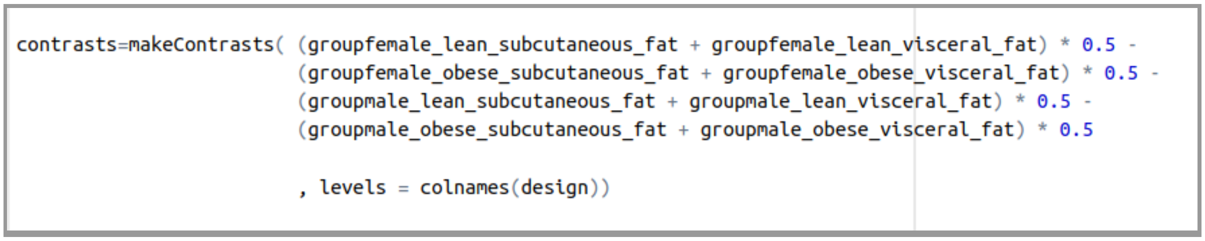
Before analyzing the data in order to extract differentially expressed genes with the limma package of R, some preprocessing was done. After downloading the .soft file from the GEO database, the data matrix described in the previous chapter was extracted into a .clean text file. For each subset (factor) of the dataset, a separate .csv file was created, listing the samples belonging to that factor. They were organized in such a way that each file was named after the factor and the factor level it represented. This arrangement is beneficial for the creation of factors in the main code. The code used to generate the .csv files is presented in Appendix C. at the end of this thesis.

In some cases, where a factor had many levels, those were grouped in order to make the process less complex. For example, dataset number 21 had three factors: gender, disease state and age. Factor gender had 2 levels, `female` and `male`, factor disease state had also 2 levels `control` and `type 2 diabetes`, but factor age had 12 levels each level indicating the specific age of each sample. Those 12 levels were grouped into 3, by categorizing each age into `young`, `middle` or `old`. Each one of those levels was then handled like a separate dataset.

4.3.2 Application of main code

Although most GEO datasets were already normalized, there were a few cases in which they were not. The dataset_value_type in the .clean file is set to “transformed count” when the data is normalized or to “count” when it is not. Each dataset was additionally examined for normalization through visualization with boxplots. In cases where the data wasn’t already normalized, a log2 transformation was applied.

Utilizing the previously mentioned arrangement of the subset .csv files, factors where created using the `factor` function of R. To verify the validity of the application, both the means model and the mean-reference model where used while creating the design and contrast matrices, following the procedure described in the example. It was carefully decided which contrasts to create for each dataset, so that results would explicitly represent the effect of gender on each study. In cases where datasets contained factors that were not of direct interest, those were averaged during the process of constructing the contrast matrices. For example, dataset number 4 had three factors: gender, disease state and tissue. Since the factor tissue was not of direct interest for the analysis conducted, it was averaged when constructing the contrast matrix. The process of averaging the factor tissue takes the following form in terms of coding:



```
contrasts=makeContrasts( (groupfemale_lean_subcutaneous_fat + groupfemale_lean_visceral_fat) * 0.5 -
                         (groupfemale_obese_subcutaneous_fat + groupfemale_obese_visceral_fat) * 0.5 -
                         (groupmale_lean_subcutaneous_fat + groupmale_lean_visceral_fat) * 0.5 -
                         (groupmale_obese_subcutaneous_fat + groupmale_obese_visceral_fat) * 0.5

                         , levels = colnames(design))
```

Figure 21: R code for averaging factors during the construction of contrasts.

After defining the appropriate contrasts, linear models were applied using the `lmFit` and the `contrasts.fit` functions of limma. The results were then corrected using the empirical Bayes correction. Each dataset was tested for differentially expressed genes between genders with a threshold of adjusted p-value < 0.5. Identifiers (gene IDs) of genes meeting this criterion and their respective adjusted p-values were extracted into a .csv file. *(Bioconductor, 2025)*

In order to filter out DEGs between genders that are not disease-related, the above process was repeated, this time extracting differentially expressed genes between control and condition for each dataset. The intersection of DEGs between genders and between control and condition for each dataset was found and extracted into a .csv file.

5. Results

Prior to applying the limma model, a two-sided t-test was performed on each dataset. The results from the t-tests and the limma method varied because of the differences described in chapter 3.5.

As described in chapter 3.1 the two competing hypotheses for testing for significantly differentially expressed genes between genders are:

Null Hypothesis H_0 :

Gender has no significant effect on differential gene expression between control and disease.

Alternative Hypothesis H_1 :

Gender has a significant effect on differential gene expression between control and disease.

Each dataset was treated as an independent study, where gender differences in gene expression were analyzed. Therefore, hypothesis testing was conducted separately for each dataset, yielding the following results:

- Significantly differentially expressed genes between genders and between control and disease conditions were found in 11 out of the total 30 datasets (36.7%). This indicates that for those datasets the null hypothesis (H_0) can be rejected, providing evidence for the alternative (H_1).

- No significantly differentially expressed genes between genders and between control and disease conditions were found in the remaining 19 datasets (63.5%). For those datasets, the null hypothesis (H_0) can not be rejected, providing insufficient evidence for the alternative hypothesis (H_1).

Although many of the datasets providing results towards the alternative hypothesis (H_1), suggest a pattern, this cannot be interpreted as a global result. However, it is of interest to examine the genes identified as significantly differentially expressed.

5.1 DEGs between genders and between control and disease conditions

A total of 677 genes were found to be significantly differentially expressed between genders and control and disease conditions in the 11 datasets. Of those, 11 (1.6%) appeared in more than one dataset tested. Below is a list of the top 15 genes ranked by frequency, along with the number of times they appeared. A link to the complete list of differentially expressed genes ranked by frequency is provided in Appendix B. at the end of this thesis.

Rank	Gene ID	Frequency
1	V11	2
2	V114	2
3	V12	2
4	V128	2
5	V13	2
6	V14	2

7	V15	2
8	V150	2
9	V154	2
10	V16	2
11	V18	2
12	208045_at	1
13	AB072904	1
14	ABCC3	1
15	ABI3BP	1

Table 1: List of top 15 DEGs based on frequency and their corresponding frequencies.

5.2 Gene expression profiling

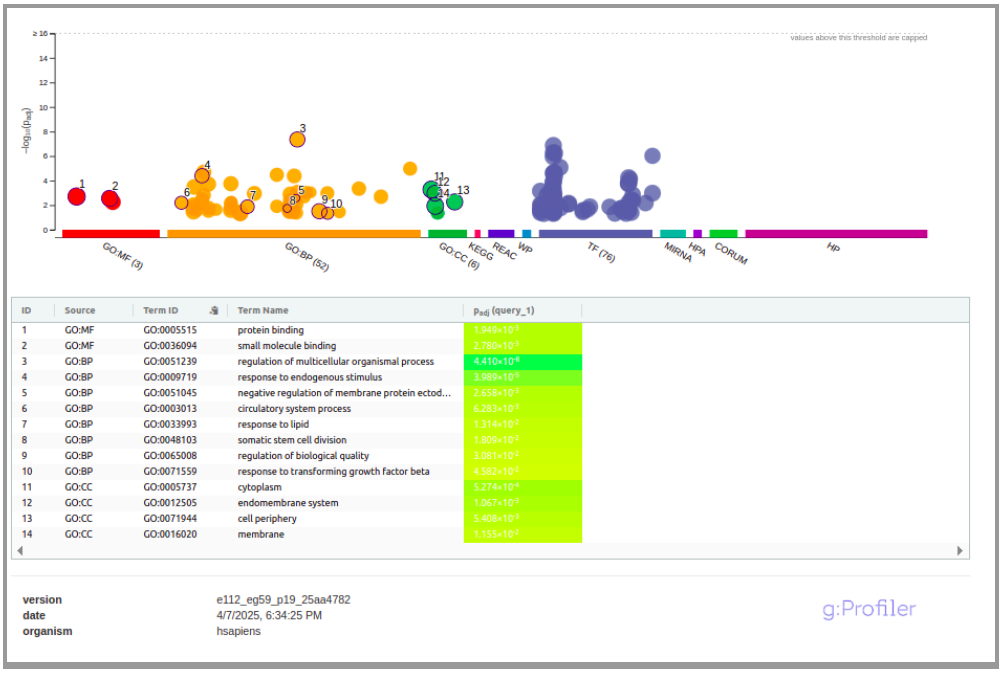
The above mentioned list of intersected DEGs was inputted into an online gene profiling tool *(Kolberg et al., 2023)* and the results were analyzed through the use of AI. *(OpenAI, 2025)*. Using the below screenshot as a prompt, the AI tool gave the following response:

The top functional enrichment results of the intersecting DEGs are visualized in Figure X, highlighting significant overrepresentation in molecular function (MF), biological process (BP), and cellular component (CC) gene ontology categories. Notably, the most enriched terms included *protein binding* (GO:0005515) and *small molecule binding* (GO:0036094), alongside *regulation of multicellular organismal process* (GO:0051239), and *response to endogenous stimulus* (GO:0009719), suggesting a central role for these genes in regulatory signaling pathways and hormone responsiveness (Ashburner et al., 2000; The Gene Ontology Consortium, 2021).

The biological processes also revealed key involvement in *somatic stem cell division* (GO:0048103), *response to lipid* (GO:0033993), and *transforming growth factor beta response* (GO:0071559), all processes known to be influenced by sex hormones and critical to development and disease progression (Arnold et al., 2016; Oliva et al., 2020). This supports the idea that sex-biased gene regulation may contribute to differential susceptibility and progression of certain diseases (Klein and Flanagan, 2016).

Cellular component terms such as *membrane* (GO:0016020), *endomembrane system* (GO:0012505), and *cytoplasm* (GO:0005737) further underscore a localization of these gene products to key signaling and trafficking compartments (Uhlén et al., 2015), which are vital for mediating extracellular signals and cellular communication.

The plot clearly illustrates that while molecular function and cellular localization are fairly specific, the breadth and density of enriched biological processes suggest these intersecting genes contribute to a complex, multifactorial regulatory landscape, likely modulated by both sex and disease state. (Gershoni and Pietrokovski, 2017; Ghosh and Klein, 2017).



ID	Source	Term ID	Term Name	p_{adj} (query_1)
1	GO:MF	GO:0005515	protein binding	1.949×10^{-3}
2	GO:MF	GO:0036094	small molecule binding	2.780×10^{-3}
3	GO:BP	GO:0051239	regulation of multicellular organismal process	4.410×10^{-8}
4	GO:BP	GO:0009719	response to endogenous stimulus	3.989×10^{-5}
5	GO:BP	GO:0051045	negative regulation of membrane protein ectod...	2.658×10^{-3}
6	GO:BP	GO:0003013	circulatory system process	6.283×10^{-3}
7	GO:BP	GO:0033993	response to lipid	1.314×10^{-2}
8	GO:BP	GO:0048103	somatic stem cell division	1.809×10^{-2}
9	GO:BP	GO:0065008	regulation of biological quality	3.081×10^{-2}
10	GO:BP	GO:0071559	response to transforming growth factor beta	4.582×10^{-2}
11	GO:CC	GO:0005737	cytoplasm	5.274×10^{-4}
12	GO:CC	GO:0012505	endomembrane system	1.067×10^{-3}
13	GO:CC	GO:0071944	cell periphery	5.408×10^{-3}
14	GO:CC	GO:0016020	membrane	1.155×10^{-2}

version	e112_eg59_p19_25aa4782
date	4/7/2025, 6:34:25 PM
organism	hsapiens

Figure 22: Screenshot of results provided by gProfiler after inputting the list of DEGs.

6. Conclusion

According to the presented results, significantly differentially expressed genes between genders were found in 11 out of 30 datasets. These findings suggest a pattern that is somewhat in accordance with the statements of the introduction in chapter 1.1 of this thesis. There are sex differences to be found in every part of the human body, including gene expression. Overlooking these differences by not sex-disaggregating the results, contributes to misdiagnoses, ineffective treatments and unrecognized risk factors.

The major difficulty in the gene expression analysis of this thesis was finding GEO datasets that included gender as a factor. This lack of gender specific data in GEO datasets, perfectly exhibits the essence of the issue addressed in this thesis: Biomedical research is not gender specific. The absence of gender specific data in studies, limits the capacities of studies investigating sex specific differences. Filling this absence would enhance the reliability and applicability of gene expression studies while also enabling more meta-analyses.

Beyond biological sex, there are also significant limitations in other kinds of diversity of study populations. Persons of color, ethnic minorities, trans, intersex and elderly individuals are rarely included in biomedical research. This lack of diverse data contributes to gaps in knowledge, leading to disparities in healthcare for these underrepresented groups.

In order to close these data gaps, future research needs to focus on greater inclusion of diverse populations and on better metadata annotation. Actively addressing these issues creates a more inclusive and therefore effective-for-all healthcare system.

References

Ashburner, M. et al. (2000) ‘Gene ontology: tool for the unification of biology’, *Nature Genetics*, **25**(1), pp. 25–29.

Arnold, A.P. et al. (2016) ‘Sex chromosomes and sex differences in gene expression’, *Nature Reviews Genetics*, **17**(12), pp. 707–718.

Beery, A.K. and Zucker, I., (2011). ‘Sex bias in neuroscience and biomedical research’. *Neuroscience & Biobehavioral Reviews,* **35**(3), pp.565-572. doi: https://doi.org/10.1016/j.neubiorev.2010.07.002. *Available at: https://www.sciencedirect.com/science/article/abs/pii/S0149763410001156?via%3Dihub (Accessed: 9 March 2025)*

Bioconductor (2025) *limma: User's Guide*. Available at: https://www.bioconductor.org/packages/devel/bioc/vignettes/limma/inst/doc/usersguide.pdf (Accessed: 9 March 2025)

Casella, G. (1985) ‘An Introduction to Empirical Bayes Data Analysis’, *The American Statistician*, **39**(2), pp. 83–87. Available at: http://www.jstor.org/stable/2682801 (Accessed: 9 March 2025)

Curno, M. J., Rossi, S., Hodges-Mameletzis, I., Johnston, R., Price, M. A., & Heidari, S. (2016). ‘A Systematic Review of the Inclusion (or Exclusion) of Women in HIV Research: From Clinical Studies of Antiretrovirals and Vaccines to Cure Strategies’. *Journal of acquired immune deficiency syndromes (1999)*, ***71***(2), 181–188. https://doi.org/10.1097/QAI.0000000000000842 Available at: https://pubmed.ncbi.nlm.nih.gov/26361171/ (Accessed: 9 March 2025)

Criado Perez, C. (2019) *Invisible women: exposing data bias in a world designed for men*. London: Chatto & Windus

Cui, X. and Churchill, G.A. (2003) ‘Statistical design and hypothesis testing of cDNA microarray experiments’, *Genome Biology*, **4**(4), pp. 210. doi:10.1186/gb-2003-4-4-210. Available at: https://genomebiology.biomedcentral.com/articles/10.1186/gb-2003-4-4-210 (Accessed: 9 March 2025)

Darolti, I. and Mank, J.E. (2023) ‘Sex-biased gene expression at single-cell resolution: cause and consequence of sexual dimorphism’, *Evolution Letters*, **7**(3), pp. 148–156. doi: 10.1093/evlett/qrad013. Available at: https://academic.oup.com/evlett/article/7/3/148/7119960 (Accessed: 9 March 2025)

Gershoni, M. and Pietrokovski, S. (2017) ‘The landscape of sex-differential transcriptome and its consequent selection in human adults’, *BMC Biology*, **15**(1), p. 7.

Ghosh, S. and Klein, R.S. (2017) ‘Sex Drives Dimorphic Immune Responses to Viral Infections’, *The Journal of Immunology*, **198**(5), pp. 1782–1790.

Griffiths, R., Morris, J., Jones, M., Smith, P. and Thomas, D. (2009) ‘Understanding gene expression in cancer: insights from microarray analysis’, *Journal of Cancer Research*, **65**(12), pp. 1234–1245. doi:

10.1111/j.1467-9868.2008.00690.x. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC2740938/ (Accessed: 9 March 2025)

Harvard Chan Bioinformatics Core (2025) ‘Count modeling in RNA-seq analysis’, *HBC Training Modules*. Available at: https://hbctraining.github.io/Training-modules/planning_successful_rnaseq/lessons/count_modeling.html (Accessed: 9 March 2025)

Illumina (2025) ‘RNA-Seq vs. Arrays: Key Advantages of RNA Sequencing’, *Illumina*. Available at: https://emea.illumina.com/science/technology/next-generation-sequencing/beginners/advantages/rna-seq-vs-arrays.html (Accessed: 6 March 2025)

Jeanmougin, M., de Reynies, A., Marisa, L., Paccard, C., Nuel, G. and Guedj, M. (2010) ‘Should we abandon the t-test in the analysis of gene expression microarray data: a comparison of variance modeling strategies’, *PLoS One*, **5**(9), e12336. doi: 10.1371/journal.pone.0012336. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC2933223/ (Accessed: 9 March 2025)

Keating, D. and Millman, J. (2014) ‘Bad medicine: The awful drug reactions Americans report’, *The Washington Post*. Available at: https://www.washingtonpost.com/news/wonk/wp/2014/06/07/bad-medicine-the-awful-drug-reactions-americans-report/ (Accessed: 9 March 2025)

Khamis, R. Y., Ammari, T., & Mikhail, G. W. (2016). ‘Gender differences in coronary heart disease’. *Heart (British Cardiac Society)*, ***102***(14), 1142–1149. https://doi.org/10.1136/heartjnl-2014-306463 Available at: https://heart.bmj.com/content/102/14/1142 (Accessed 9 March 2025)

Klein, S.L. and Flanagan, K.L. (2016) ‘Sex differences in immune responses’, *Nature Reviews Immunology*, **16**(10), pp. 626–638.

Kolberg, L., Raudvere, U., Kuzmin, I., Adler, P., Vilo, J. and Peterson, H. (2023) ‘g:Profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update)’, *Nucleic Acids Research*, doi: 10.1093/nar/gkad347 Available at: https://biit.cs.ut.ee/gprofiler/gost (Accessed 18 April 2025)

Lambert, J. (2011) ‘Statistics in brief: how to assess bias in clinical studies?’, *Clinical Orthopaedics and Related Research*, **469**(6), pp. 1794–1796. doi:10.1007/s11999-010-1538-7. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3094617/ (Accessed: 9 March 2025)

Law, C.W., Zeglinski, K., Dong, X., Alhamdoosh, M., Smyth, G.K. and Ritchie, M.E. (2020) ‘A guide to creating design matrices for gene expression experiments’, *F1000Research*, **9**, p. 1444. doi: 10.12688/f1000research.27893.1. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC7873980/ (Accessed: 9 March 2025)

Lear, M.W. (2014) ‘Women’s atypical heart attacks’, *The New York Times*. Available at: https://www.nytimes.com/2014/09/28/opinion/sunday/womens-atypical-heart-attacks.html?_r=0 (Accessed: 9 March 2025)

MedlinePlus (2025) ‘What is a gene?’, *MedlinePlus Genetics*. Available at: https://medlineplus.gov/genetics/understanding/basics/gene/ (Accessed: 9 March 2025)

Mifsud, M. (2016) ‘Girls with autism spectrum disorder : missed diagnosis or misdiagnosis?’, *University of Malta.* Available at: https://www.um.edu.mt/library/oar/handle/123456789/15597 (Accessed: 9 March 2025)

NCBI (n.d.) ‘GEO FAQ: What is GEO?’, *National Center for Biotechnology Information*. Available at: https://www.ncbi.nlm.nih.gov/geo/info/faq.html#what (Accessed: 9 March 2025)

NCBI (n.d.) ‘GEO Platform Information’, *National Center for Biotechnology Information*. Available at: https://www.ncbi.nlm.nih.gov/geo/info/platform.html (Accessed: 9 March 2025).

Office of Research on Women's Health (2021) ‘Including women and minorities in clinical research: Background’, *Office of Research on Women's Health, National Institutes of Health.* Available at: https://orwh.od.nih.gov/including-women-and-minorities-in-clinical-research-background (Accessed: 9 March 2025)

Oliva, M. et al. (2020) ‘The impact of sex on gene expression across human tissues’, *Science*, **369**(6509).

Plataforma SINC. (2008). ‘Medical Textbooks Use White, Heterosexual Men As A 'Universal Model''. *ScienceDaily*. Available at: www.sciencedaily.com/releases/2008/10/081015132108.html (Accessed: 9 March 2025)

Schiebinger, L. (2003) ‘Women’s health and clinical trials’, *Journal of Clinical Investigation*, **112**(7), pp. 973–977. doi: 10.1172/JCI19993. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC198535/ (Accessed: 9 March 2025)

Smyth, G.K. (2004) ‘Linear models and empirical Bayes methods for assessing differential expression in microarray experiments’, *Statistical Applications in Genetics and Molecular Biology*, **3**, Article 3. doi: 10.2202/1544-6115.1027. Available at: https://pubmed.ncbi.nlm.nih.gov/16646809/ (Accessed: 9 March 2025)

Stoakes, Shelley Farrar. 2019. How Do Microarrays Work?. *News-Medical,* Available at: https://www.news-medical.net/life-sciences/How-Do-Microarrays-Work.aspx. (Accessed: 9 March 2025)

Szalavitz, M. (2016) ‘Autism: It’s different in girls’, *Scientific American*. Available at: https://www.scientificamerican.com/article/autism-it-s-different-in-girls/ (Accessed: 9 March 2025)

The Gene Ontology Consortium (2021) ‘The Gene Ontology resource: enriching a GOld mine’, *Nucleic Acids Research*, **49**(D1), pp. D325–D334.

Tucker, M.J., Berg, C.J., Callaghan, W.M. and Hsia, J. (2007) ‘The Black-White disparity in pregnancy-related mortality from 5 conditions: differences in prevalence and case-fatality rates’*, American Journal of Public Health,* **97**(2), pp. 247–251. doi: 10.2105/AJPH.2005.072975. Available at: https://pubmed.ncbi.nlm.nih.gov/17194867/ (Accessed: 9 March 2025)

UCDavis Bioinformatics Training (2019) ‘Expression matrix generation and data reduction in single-cell RNA sequencing’, *UCDavis Bioinformatics Training*. Available at: https://ucdavis-bioinformatics-training.github.io/2019-single-cell-RNA-sequencing-Workshop-UCD_UCSF/data_reduction/Expression_Matrix.html (Accessed: 9 March 2025)

Uhlén, M. et al. (2015) ‘Tissue-based map of the human proteome’, *Science*, **347**(6220).

Vitale, C., Fini, M., Spoletini, I., Lainscak, M., Seferovic, P., & Rosano, G. M. (2017). ‘Under-representation of elderly and women in clinical trials’. *International journal of cardiology*, ***232***, 216–221. https://doi.org/10.1016/j.ijcard.2017.01.018 Available at: https://pubmed.ncbi.nlm.nih.gov/28111054/ (Accessed: 9 March 2025)

Yoon, D. Y., Mansukhani, N. A., Stubbs, V. C., Helenowski, I. B., Woodruff, T. K., & Kibbe, M. R. (2014). ‘Sex bias exists in basic science and translational surgical research’. *Surgery*, ***156***(3), 508–516. https://doi.org/10.1016/j.surg.2014.07.001 Available at: https://pubmed.ncbi.nlm.nih.gov/25175501/ (Accessed: 9 March 2025)

Zopf, Y., Rabe, C., Neubert, A., Gassmann, K. G., Rascher, W., Hahn, E. G., Brune, K., & Dormann, H. (2008). ‘Women encounter ADRs more often than do men’. *European journal of clinical pharmacology*, ***64***(10), 999–1004. https://doi.org/10.1007/s00228-008-0494-6 Available at: https://pubmed.ncbi.nlm.nih.gov/18604529/ (Accessed: 9 March 2025)

Appendix A. Datasets

1. Cigarette smoke effect on the oral mucosa
   https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS3709

2. Chronic high-level alcohol consumption effect on brain: post-mortem hippocampus
   https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS4879

3. Type 2 diabetic and insulin-resistant but normoglycemic cohorts: skeletal muscle
   https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS3884

4. Subcutaneous and visceral adipose tissue from lean or obese children
   https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS4276

5. Postmortem Alzheimer's disease brains: Hisayama study
   https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS4758

6. Non-syndromic cleft lip and palate: dental pulp stem cell cultures
   https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS5071

7. Skeletal muscles from men and women of various ages
   https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS4858

8. Uveal melanoma
   https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS4281

9. Pituitary gonadotrope tumors
   https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS4275

10. Bariatric surgery effect on obesity-related type 2 diabetes: whole blood
    https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS3881

11. Hair follicles from males and females
    https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS2356

12. Ulcerative colitis
    https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS2014

13. Aging brain: frontal cortex expression profiles at various ages
    https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS707

14. Severe asthma: bronchial epithelial cell
    https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS5037

15. Age effect on normal adult brain: frontal cortical region
    https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS5204

16. Male and female in vitro cultured placental cell types
    https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS5016

17. Systemic lupus erythematosus and arthritides: synovial biopsies
    https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS4195

18. Schizophrenia: postmortem superior temporal cortex
    https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS4522

19. Early and late onset colorectal cancers
    https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS5232

20. Induced pluripotent stem cell-derived neurons from schizophrenia patients
    https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS3938

21. Type 2 diabetic pancreatic islets of Langerhans
    https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS3882

22. Type 2 diabetes and role of hepatokines
    https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS3883

23. Morbidly obese insulin-resistant patients: omental and subcutaneous adipose tissue
    https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS3781

24. Schizophrenia: postmortem anterior prefrontal cortex
    https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS4523

25. Obesity-associated insulin resistance independent of BMI: omental and subcutaneous adipose tissues
    https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS3962

26. Effect of lifestyle on Moroccan Amazighs: peripheral blood leukocytes
    https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS4283

27. "Moderate stage Huntington's disease lymphocytes
    https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS2887

28. Resistance exercise training effect on the skeletal muscle
    https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS4894

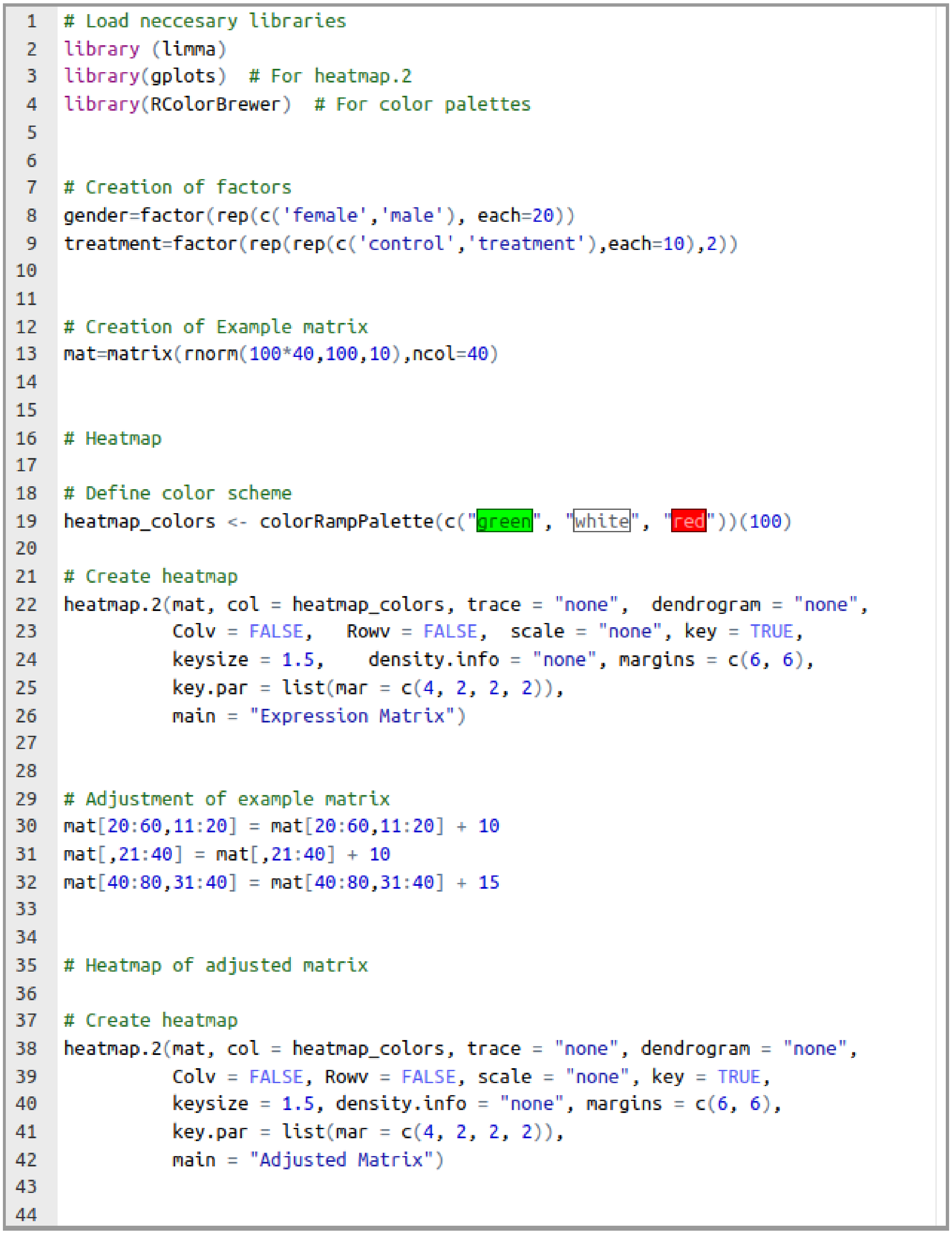
29. Parkinson's disease: substantia nigra
    https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS2821

30. Alzheimer blood mononuclear cells
    https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS2601

Appendix B. Results List

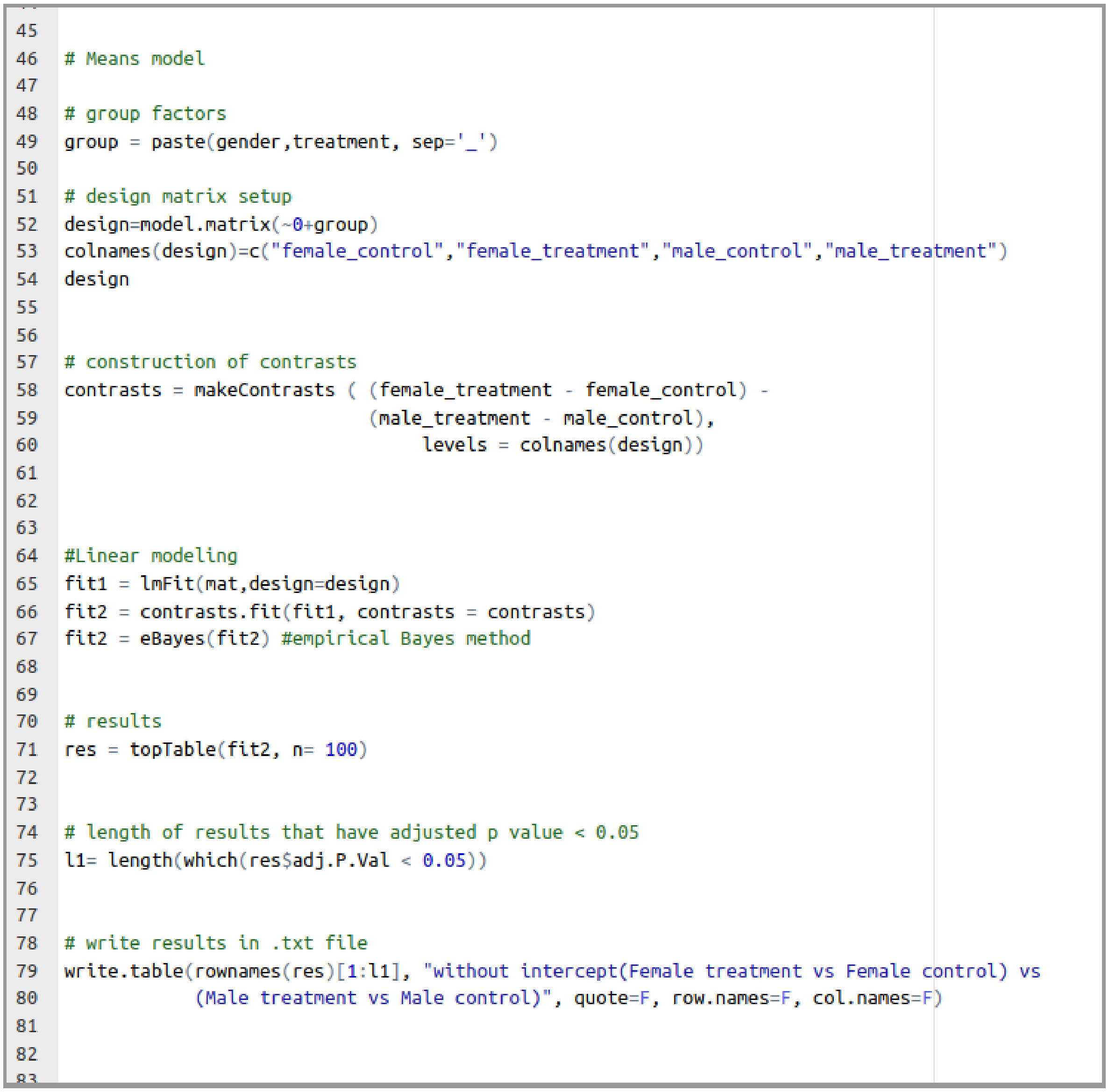
1. Descending list of differentially expressed genes between genders and between control and disease conditions, ranked by frequency, along with the number of times they appeared in datasets:
   intersection_all_DEGS_counted

Appendix C. Code



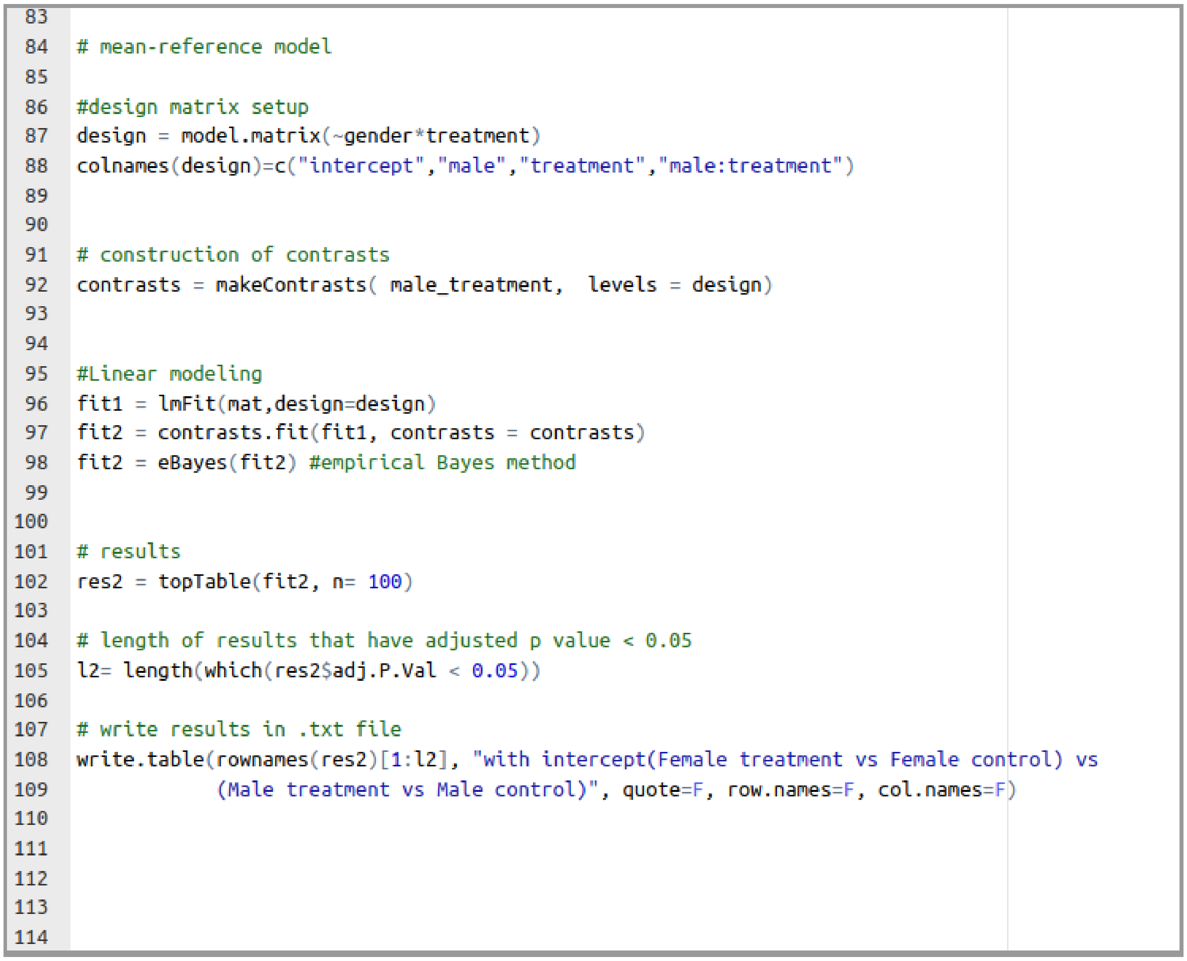
```r
# Load neccesary libraries
library (limma)
library(gplots)  # For heatmap.2
library(RColorBrewer)  # For color palettes


# Creation of factors
gender=factor(rep(c('female','male'), each=20))
treatment=factor(rep(rep(c('control','treatment'),each=10),2))


# Creation of Example matrix
mat=matrix(rnorm(100*40,100,10),ncol=40)


# Heatmap

# Define color scheme
heatmap_colors <- colorRampPalette(c("green", "white", "red"))(100)

# Create heatmap
heatmap.2(mat, col = heatmap_colors, trace = "none",  dendrogram = "none",
          Colv = FALSE,   Rowv = FALSE,  scale = "none", key = TRUE,
          keysize = 1.5,    density.info = "none", margins = c(6, 6),
          key.par = list(mar = c(4, 2, 2, 2)),
          main = "Expression Matrix")


# Adjustment of example matrix
mat[20:60,11:20] = mat[20:60,11:20] + 10
mat[,21:40] = mat[,21:40] + 10
mat[40:80,31:40] = mat[40:80,31:40] + 15


# Heatmap of adjusted matrix

# Create heatmap
heatmap.2(mat, col = heatmap_colors, trace = "none", dendrogram = "none",
          Colv = FALSE, Rowv = FALSE, scale = "none", key = TRUE,
          keysize = 1.5, density.info = "none", margins = c(6, 6),
          key.par = list(mar = c(4, 2, 2, 2)),
          main = "Adjusted Matrix")

```

Figure 23: R code for example. Part 1/3.



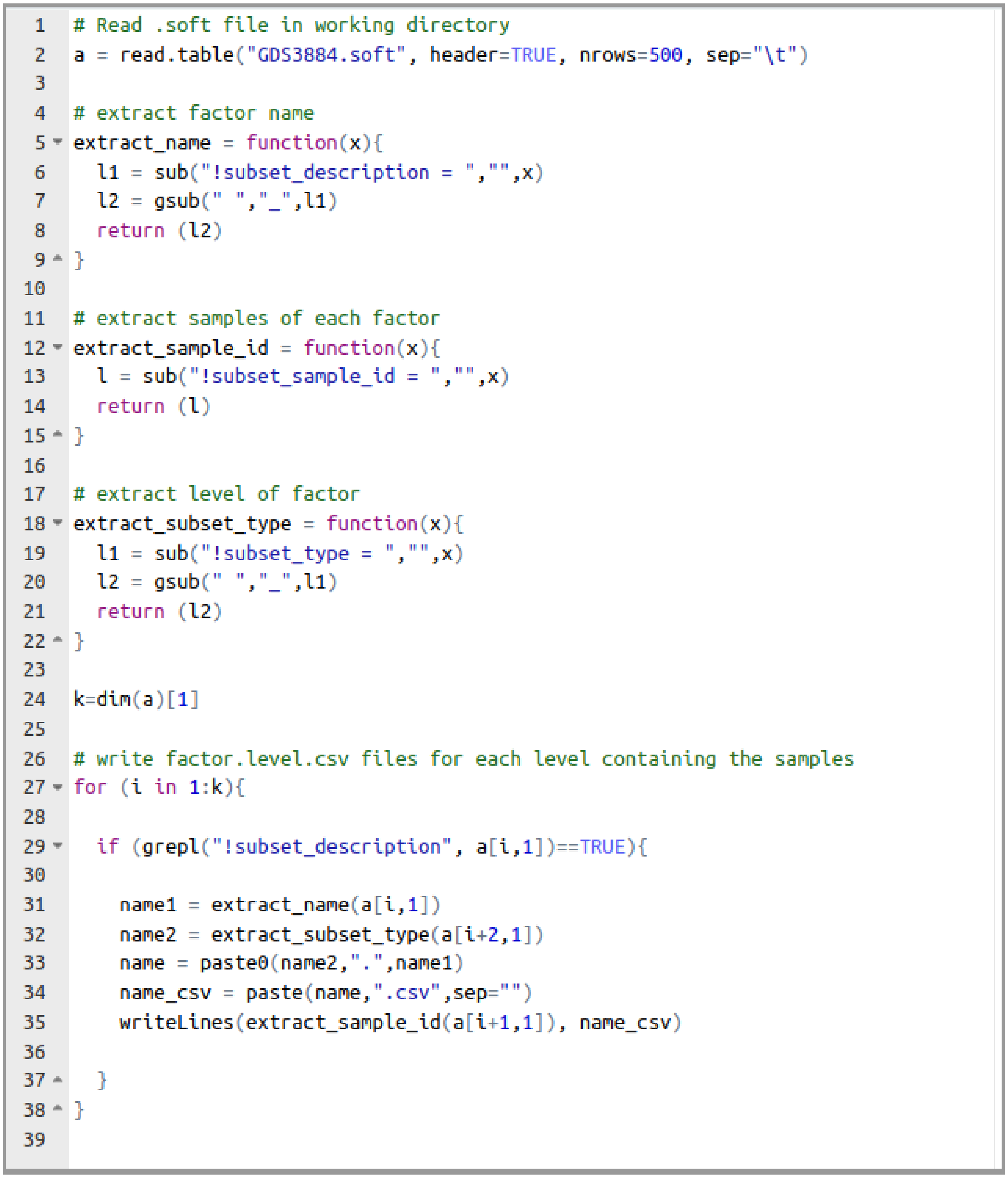
```r

# Means model

# group factors
group = paste(gender,treatment, sep='_')

# design matrix setup
design=model.matrix(~0+group)
colnames(design)=c("female_control","female_treatment","male_control","male_treatment")
design


# construction of contrasts
contrasts = makeContrasts ( (female_treatment - female_control) -
                            (male_treatment - male_control),
                                 levels = colnames(design))



#Linear modeling
fit1 = lmFit(mat,design=design)
fit2 = contrasts.fit(fit1, contrasts = contrasts)
fit2 = eBayes(fit2) #empirical Bayes method


# results
res = topTable(fit2, n= 100)


# length of results that have adjusted p value < 0.05
l1= length(which(res$adj.P.Val < 0.05))


# write results in .txt file
write.table(rownames(res)[1:l1], "without intercept(Female treatment vs Female control) vs
            (Male treatment vs Male control)", quote=F, row.names=F, col.names=F)


```

Figure 24: R code for example. Part 2/3.



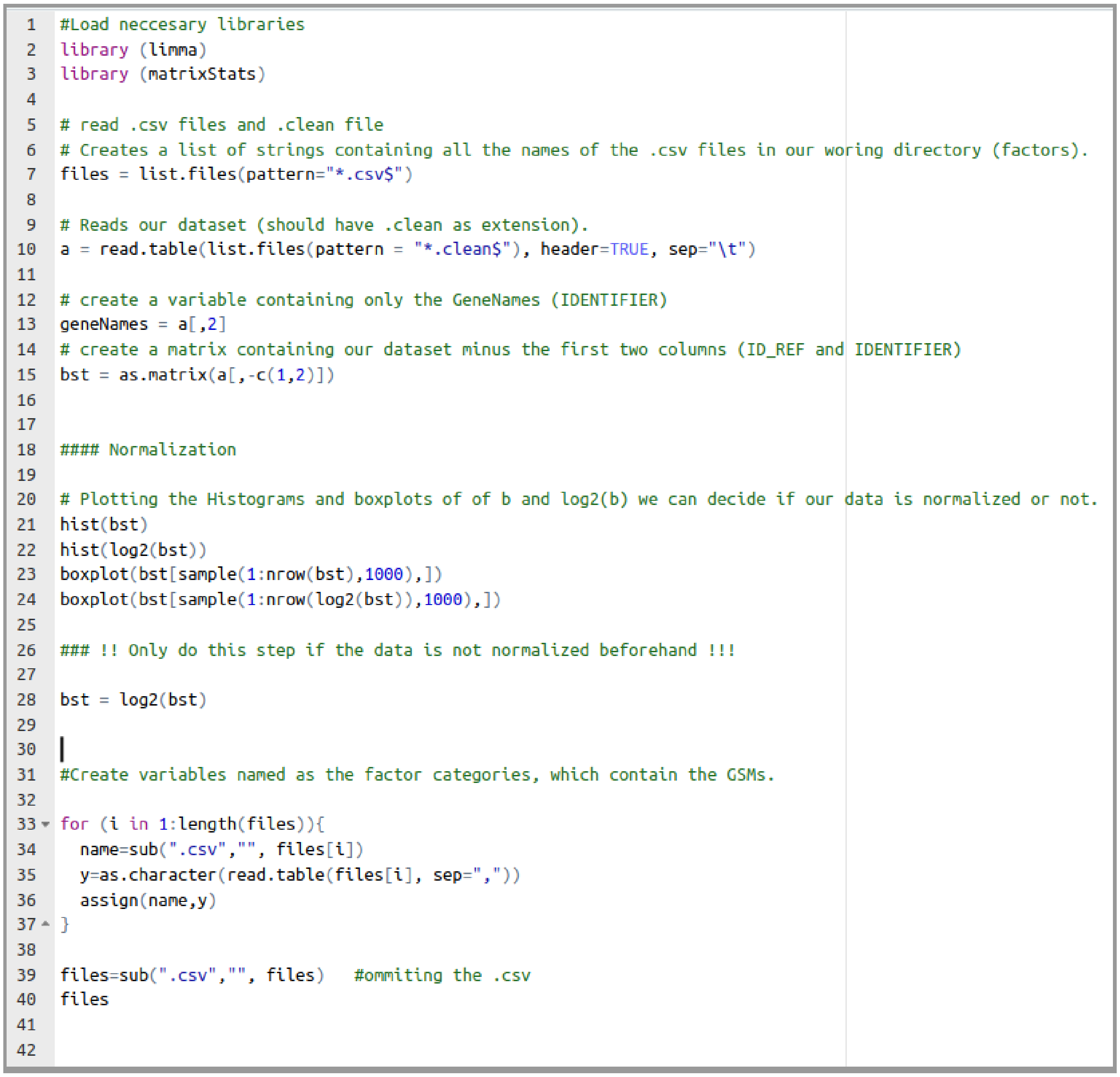
```r

# mean-reference model

#design matrix setup
design = model.matrix(~gender*treatment)
colnames(design)=c("intercept","male","treatment","male:treatment")


# construction of contrasts
contrasts = makeContrasts( male_treatment,  levels = design)


#Linear modeling
fit1 = lmFit(mat,design=design)
fit2 = contrasts.fit(fit1, contrasts = contrasts)
fit2 = eBayes(fit2) #empirical Bayes method


# results
res2 = topTable(fit2, n= 100)

# length of results that have adjusted p value < 0.05
l2= length(which(res2$adj.P.Val < 0.05))

# write results in .txt file
write.table(rownames(res2)[1:l2], "with intercept(Female treatment vs Female control) vs
            (Male treatment vs Male control)", quote=F, row.names=F, col.names=F)





```

Figure 25: R code for example. Part 3/3.



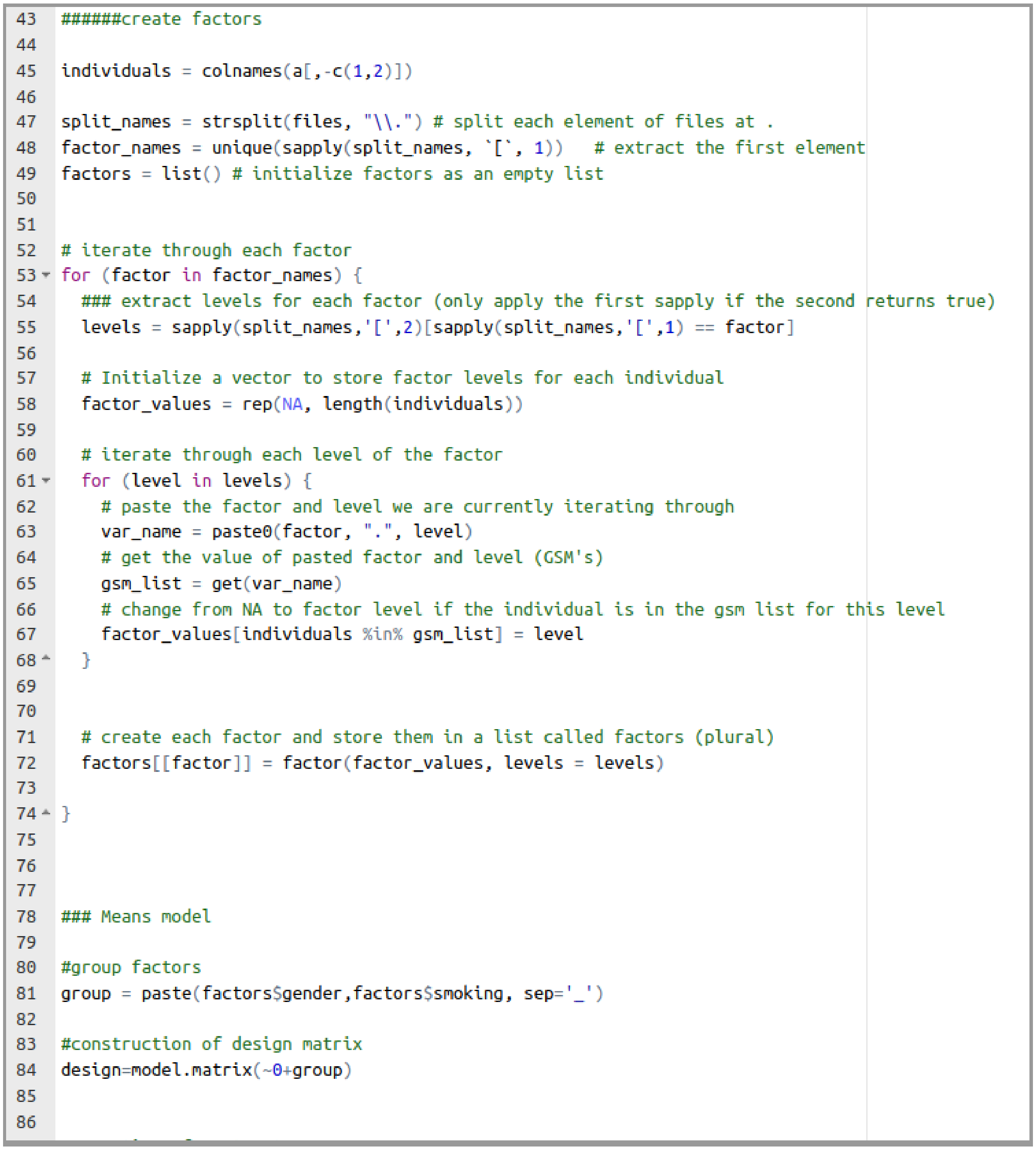
```r
# Read .soft file in working directory
a = read.table("GDS3884.soft", header=TRUE, nrows=500, sep="\t")

# extract factor name
extract_name = function(x){
  l1 = sub("!subset_description = ","",x)
  l2 = gsub(" ","_",l1)
  return (l2)
}

# extract samples of each factor
extract_sample_id = function(x){
  l = sub("!subset_sample_id = ","",x)
  return (l)
}

# extract level of factor
extract_subset_type = function(x){
  l1 = sub("!subset_type = ","",x)
  l2 = gsub(" ","_",l1)
  return (l2)
}

k=dim(a)[1]

# write factor.level.csv files for each level containing the samples
for (i in 1:k){

  if (grepl("!subset_description", a[i,1])==TRUE){

    name1 = extract_name(a[i,1])
    name2 = extract_subset_type(a[i+2,1])
    name = paste0(name2,".",name1)
    name_csv = paste(name,".csv",sep="")
    writeLines(extract_sample_id(a[i+1,1]), name_csv)

  }
}
```

Figure 26: R code for creation of .csv files for each level.



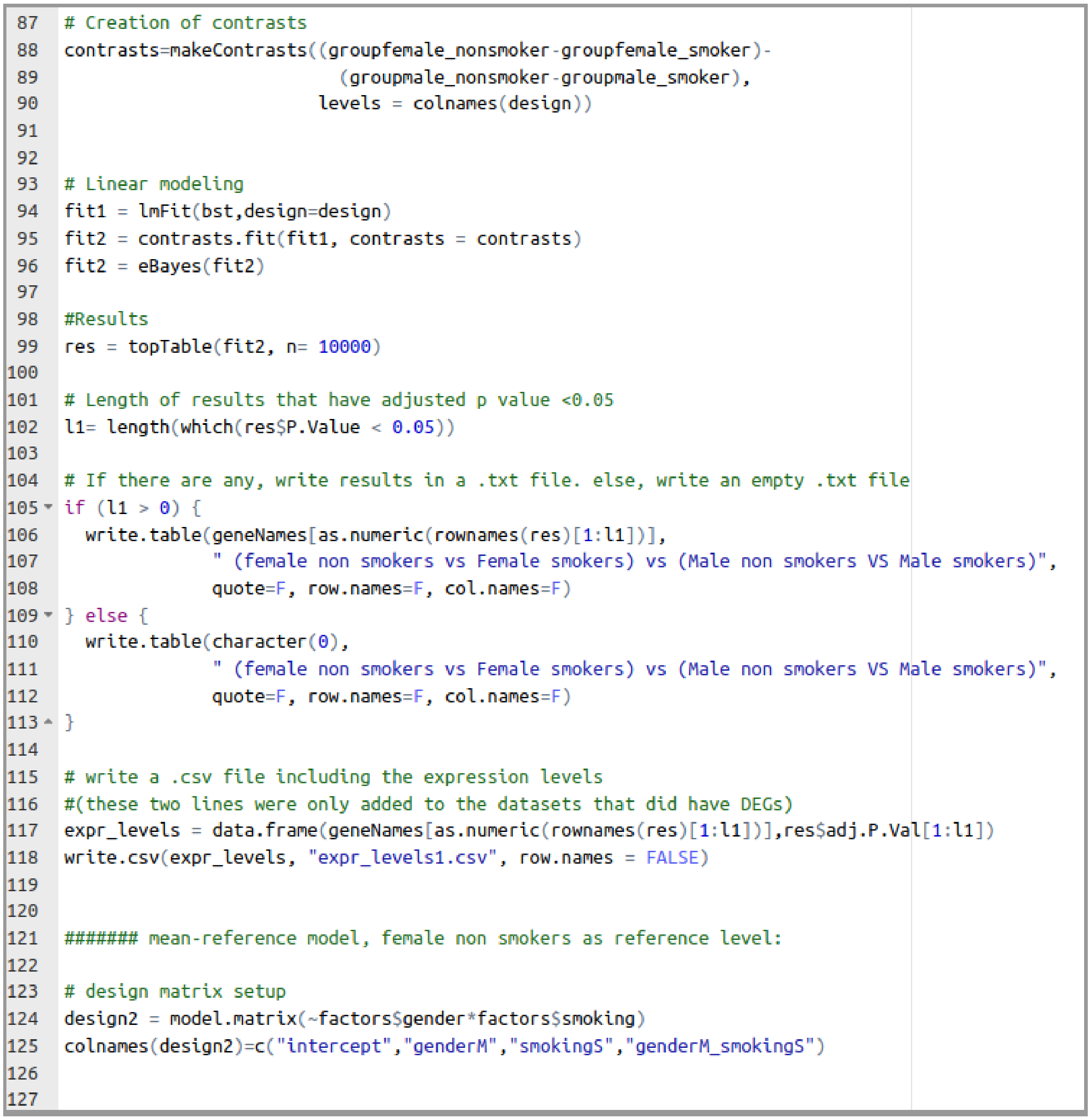
```r
#Load neccesary libraries
library (limma)
library (matrixStats)

# read .csv files and .clean file
# Creates a list of strings containing all the names of the .csv files in our woring directory (factors).
files = list.files(pattern="*.csv$")

# Reads our dataset (should have .clean as extension).
a = read.table(list.files(pattern = "*.clean$"), header=TRUE, sep="\t")

# create a variable containing only the GeneNames (IDENTIFIER)
geneNames = a[,2]
# create a matrix containing our dataset minus the first two columns (ID_REF and IDENTIFIER)
bst = as.matrix(a[,-c(1,2)])


#### Normalization

# Plotting the Histograms and boxplots of of b and log2(b) we can decide if our data is normalized or not.
hist(bst)
hist(log2(bst))
boxplot(bst[sample(1:nrow(bst),1000),])
boxplot(bst[sample(1:nrow(log2(bst)),1000),])

### !! Only do this step if the data is not normalized beforehand !!!

bst = log2(bst)


#Create variables named as the factor categories, which contain the GSMs.

for (i in 1:length(files)){
  name=sub(".csv","", files[i])
  y=as.character(read.table(files[i], sep=","))
  assign(name,y)
}

files=sub(".csv","", files)   #ommiting the .csv
files


```

Figure 27: R code for application on each dataset. Part 1/4.



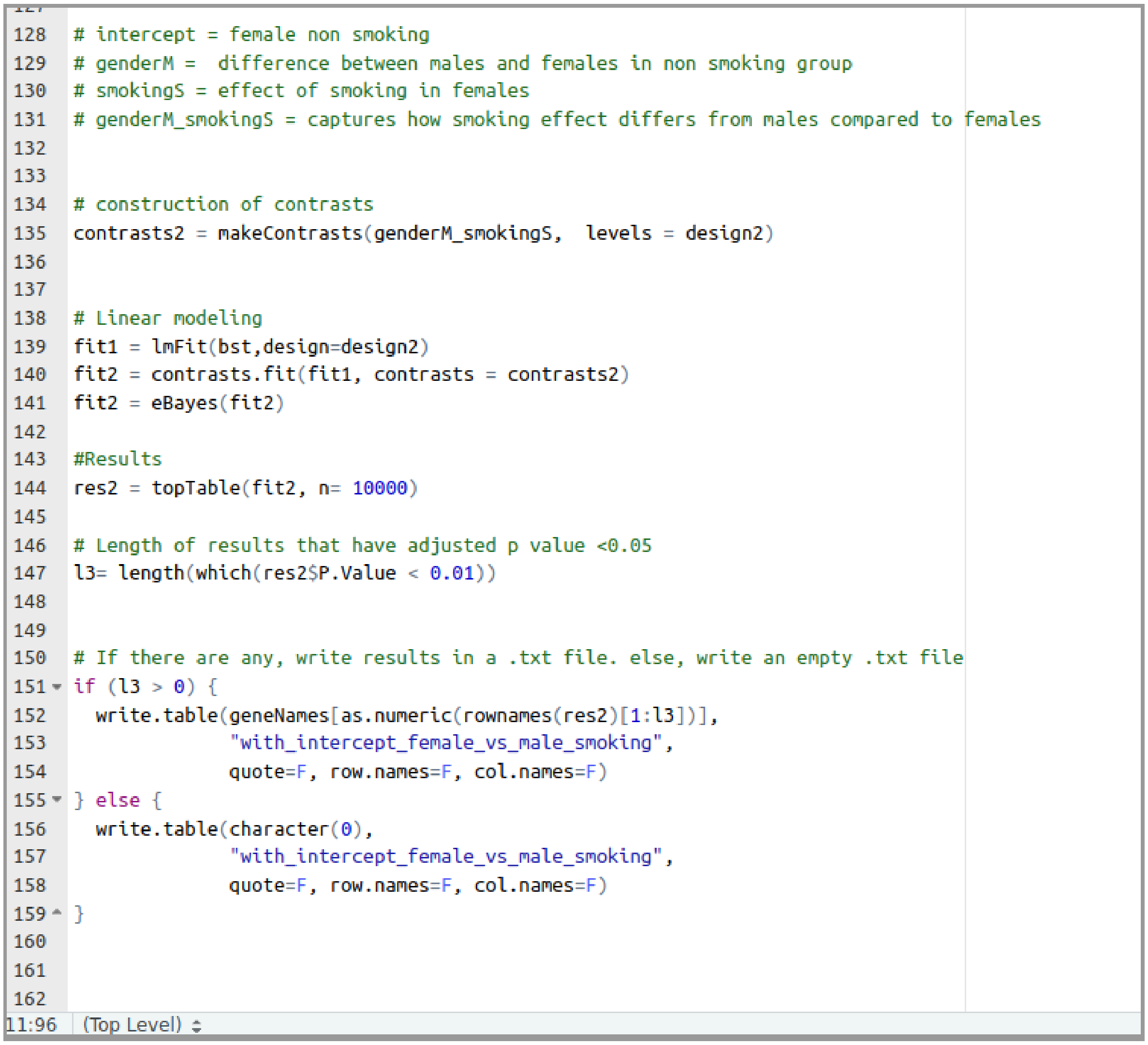
```r
######create factors

individuals = colnames(a[,-c(1,2)])

split_names = strsplit(files, "\\.") # split each element of files at .
factor_names = unique(sapply(split_names, `[`, 1))   # extract the first element
factors = list() # initialize factors as an empty list


# iterate through each factor
for (factor in factor_names) {
  ### extract levels for each factor (only apply the first sapply if the second returns true)
  levels = sapply(split_names,'[',2)[sapply(split_names,'[',1) == factor]

  # Initialize a vector to store factor levels for each individual
  factor_values = rep(NA, length(individuals))

  # iterate through each level of the factor
  for (level in levels) {
    # paste the factor and level we are currently iterating through
    var_name = paste0(factor, ".", level)
    # get the value of pasted factor and level (GSM's)
    gsm_list = get(var_name)
    # change from NA to factor level if the individual is in the gsm list for this level
    factor_values[individuals %in% gsm_list] = level
  }


  # create each factor and store them in a list called factors (plural)
  factors[[factor]] = factor(factor_values, levels = levels)

}



### Means model

#group factors
group = paste(factors$gender,factors$smoking, sep='_')

#construction of design matrix
design=model.matrix(~0+group)


```

Figure 28: R code for application on each dataset. Part 2/4.



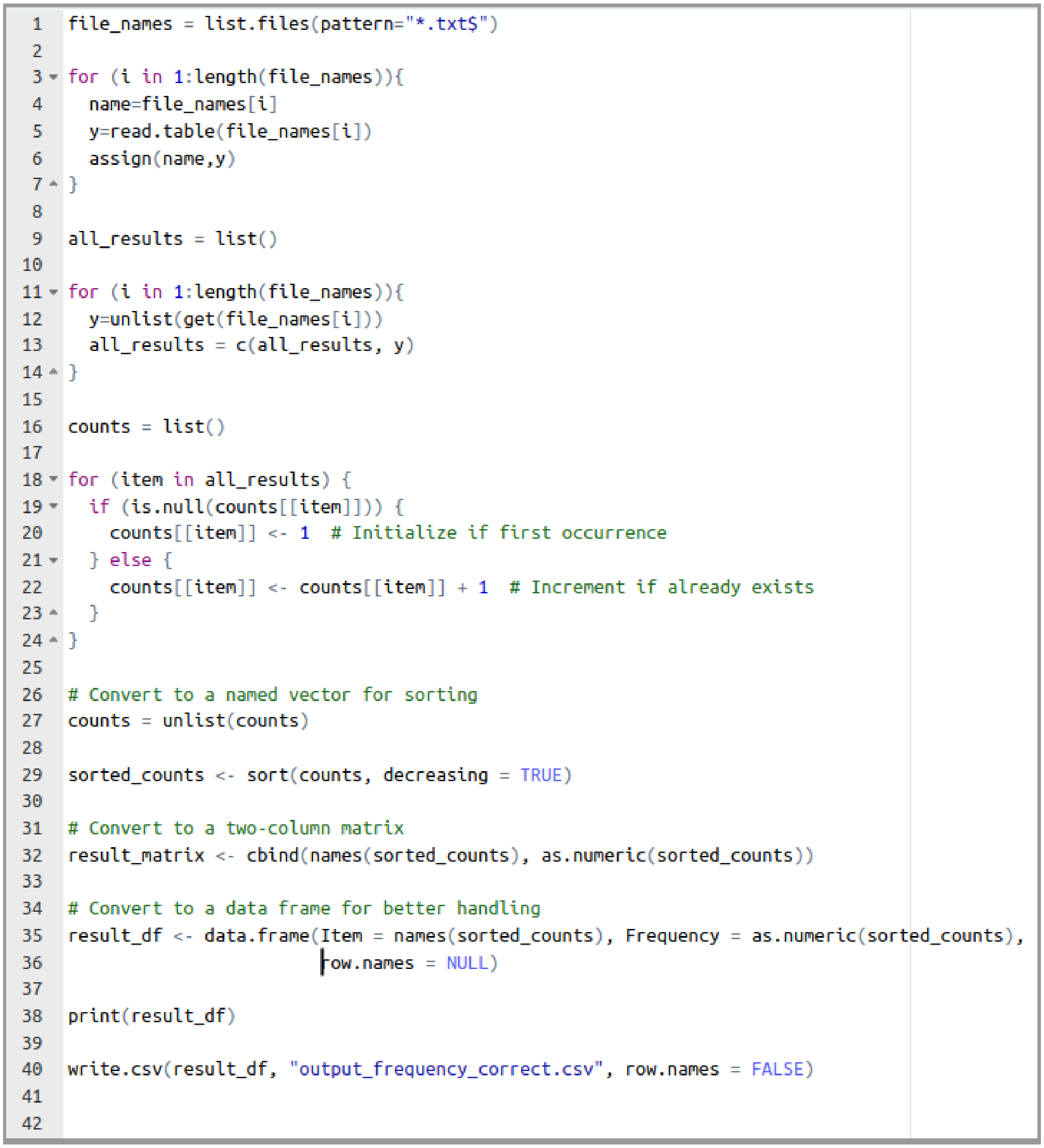
```
# Creation of contrasts
contrasts=makeContrasts((groupfemale_nonsmoker-groupfemale_smoker)-
                          (groupmale_nonsmoker-groupmale_smoker),
                        levels = colnames(design))


# Linear modeling
fit1 = lmFit(bst,design=design)
fit2 = contrasts.fit(fit1, contrasts = contrasts)
fit2 = eBayes(fit2)

#Results
res = topTable(fit2, n= 10000)

# Length of results that have adjusted p value <0.05
l1= length(which(res$P.Value < 0.05))

# If there are any, write results in a .txt file. else, write an empty .txt file
if (l1 > 0) {
  write.table(geneNames[as.numeric(rownames(res)[1:l1])],
              " (female non smokers vs Female smokers) vs (Male non smokers VS Male smokers)",
              quote=F, row.names=F, col.names=F)
} else {
  write.table(character(0),
              " (female non smokers vs Female smokers) vs (Male non smokers VS Male smokers)",
              quote=F, row.names=F, col.names=F)
}

# write a .csv file including the expression levels
#(these two lines were only added to the datasets that did have DEGs)
expr_levels = data.frame(geneNames[as.numeric(rownames(res)[1:l1])],res$adj.P.Val[1:l1])
write.csv(expr_levels, "expr_levels1.csv", row.names = FALSE)


####### mean-reference model, female non smokers as reference level:

# design matrix setup
design2 = model.matrix(~factors$gender*factors$smoking)
colnames(design2)=c("intercept","genderM","smokingS","genderM_smokingS")

```

Figure 29: R code for application on each dataset. Part 3/4.



```r
# intercept = female non smoking
# genderM =  difference between males and females in non smoking group
# smokingS = effect of smoking in females
# genderM_smokingS = captures how smoking effect differs from males compared to females


# construction of contrasts
contrasts2 = makeContrasts(genderM_smokingS,  levels = design2)


# Linear modeling
fit1 = lmFit(bst,design=design2)
fit2 = contrasts.fit(fit1, contrasts = contrasts2)
fit2 = eBayes(fit2)

#Results
res2 = topTable(fit2, n= 10000)

# Length of results that have adjusted p value <0.05
l3= length(which(res2$P.Value < 0.01))


# If there are any, write results in a .txt file. else, write an empty .txt file
if (l3 > 0) {
  write.table(geneNames[as.numeric(rownames(res2)[1:l3])],
              "with_intercept_female_vs_male_smoking",
              quote=F, row.names=F, col.names=F)
} else {
  write.table(character(0),
              "with_intercept_female_vs_male_smoking",
              quote=F, row.names=F, col.names=F)
}



```

Figure 30: R code for application on each dataset. Part 4/4.



```r
file_names = list.files(pattern="*.txt$")

for (i in 1:length(file_names)){
  name=file_names[i]
  y=read.table(file_names[i])
  assign(name,y)
}

all_results = list()

for (i in 1:length(file_names)){
  y=unlist(get(file_names[i]))
  all_results = c(all_results, y)
}

counts = list()

for (item in all_results) {
  if (is.null(counts[[item]])) {
    counts[[item]] <- 1  # Initialize if first occurrence
  } else {
    counts[[item]] <- counts[[item]] + 1  # Increment if already exists
  }
}

# Convert to a named vector for sorting
counts = unlist(counts)

sorted_counts <- sort(counts, decreasing = TRUE)

# Convert to a two-column matrix
result_matrix <- cbind(names(sorted_counts), as.numeric(sorted_counts))

# Convert to a data frame for better handling
result_df <- data.frame(Item = names(sorted_counts), Frequency = as.numeric(sorted_counts),
                        row.names = NULL)

print(result_df)

write.csv(result_df, "output_frequency_correct.csv", row.names = FALSE)
```

Figure 31: R code for extraction of frequency list of results.