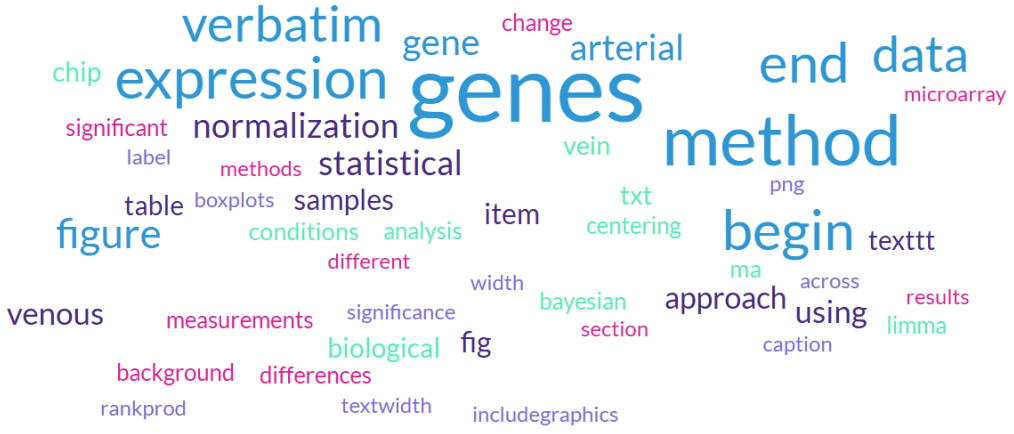
UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Statistical Methods in Genetics: Microarray Data



Alexandre Baptista, nº64506
Jéssica Soares, nº43356

Mestrado em Ciência de Dados
2024/2025

Contents

1 Introduction

When saphenous vein tissue (veins from the inner leg) is used for revascularization in patients operated from cardiac disease, in particular, coronary artery bypass graft surgery, this tissue is often subject to inflammatory and atherosclerosis processes resulting from the arteries failing to adapt to arterial conditions. By identifying differentially expressed genes in response to these varying arterial conditions we can better predict graft success or failure, as well as the associated pathways and targets that could contribute to the therapeutic inhibition of vein graft complications [1].

It is also important to understand how intimal hyperplasia develops in order to improve vein graft success. This condition is marked by abnormal smooth muscle growth and tissue buildup, and results in the narrowing of blood vessels. It is often triggered by endothelial damage caused during surgery. Ex vivo vein models allow for the detailed study of these mechanisms in a controlled setting, revealing disease-specific insights and testing treatments which help to improve patient outcomes [2].

In this study, researchers at a Brazilian cardiology lab analyzed how saphenous vein tissue responds to arterial versus venous conditions using cDNA microarrays. Since endothelial cells regulate inflammation and vascular function, identifying genes active under different hemodynamic forces helps uncover pathways linked to vein graft failure. Microarrays allowed high-throughput comparison of 2,994 genes, in the expectation of revealing differentially expressed targets. This approach—previously used to study cancer and development—highlights genes critical for endothelial function [3], and could offer valuable insights in aiding therapeutic discovery for cardiovascular diseases.

We will be normalizing data, and cross-validating our results across the Bayseian method of Lonnstedt and Speed, moderated t-statistics and the Rank Product method, to identify the genes response to arterial and venous blood flow conditions.

The LIMMA package, which provides an array of statistical methods for differential gene expression analysis, and the RankProd package, which contains the Rank Product (RP) method, are available through Bioconductor, an open-source platform dedicated to computational biology and bioinformatics research [5, 6].

The approach we will be following consists of:

Task	**Package used**
Compiling and normalizing the data	Limma
Applying the Bayesian Method developed by Lonnstedt and Speed	Limma
Applying the Moderated t-statistic	Limma
Applying the RankProd method	RankProd
Comparing the results obtained from the various methods	-

The code will be available in the file `MEG.rmd`.

2 Compiling and Normalizing the Data

For this analysis, we had three available datasets—`chip1.txt`, `chip2.txt`, and `chip3.txt`—each representing a separate patient. These files contain gene expression data from saphenous vein tissues exposed to arterial (Art) and venous (Ven) conditions, as well as corresponding background measurements (BgArt, BgVen).

Each file is composed of five columns:

- Art: Arterial regime expression levels;
- Ven: Venous regime expression levels;
- BgArt: Background values for arterial samples;
- BgVen: Background values for venous samples,
- Id: Gene identifiers.

We will compare how these 2,994 genes respond differently to arterial and venous blood flow conditions.

2.1 Compiling the Data

In this microarray experiment, the data files (`chip1.txt`, `chip2.txt` and `chip3.txt`) contain the expression levels of the genes from the tissue subjected to the arterial regime (Art), the genes from the tissue subjected to the venous regime (Ven), and their respective background values (BgArt and BgVen).

By using `read.table` and renaming our columns, we get to better control how the data is imported, ensuring consistency across files and prepares the data for the next steps (normalization, background correction). This way, we map the original column names ("Art", "Ven") to the standard names ("R", "G") that limma expects for two-color array data.

```
read_microarray <- function(file) {
df <- read.table(file, header = TRUE)
colnames(df) <- c("ID", "R", "G", "Rb", "Gb")
return(df)}
```

Once the columns are mapped, the files are imported into dataframes named c1, c2 and c3:

```
c1 <- read_microarray("chip1.txt")
c2 <- read_microarray("chip2.txt")
c3 <- read_microarray("chip3.txt")
```

The data is then compiled into an RGlist:

```
dd <- new("RGList", list(
  R = cbind(c1$R, c2$R, c3$R),
  G = cbind(c1$G, c2$G, c3$G),
  Rb = cbind(c1$Rb, c2$Rb, c3$Rb),
  Gb = cbind(c1$Gb, c2$Gb, c3$Gb),
))

colnames(dd$R) <- colnames(dd$G) <- colnames(dd$Rb) <- colnames(dd$Gb) <- c("c1", "c2", "c3")
```

2.2 Normalizing the Data

2.2.1 Background Correction

Before we can compare gene expression levels between arterial and venous samples, we first need to clean up the data. The microarray captures two types of signals:

- The real signal from actual gene expression
- Background "noise" from random molecules sticking to the chip

Background correction involves adjusting both the arterial (red) and venous (green) channel values using their respective background measurements (BgArt and BgVen) before doing any normalization. This is a crucial first step that ensures that when we compare the arterial vs venous samples, we're only looking at real biological differences - not just random noise or errors from the experiment.

```
dd_corrected <- backgroundCorrect(dd, method="normexp")
```

We consider the normexp (normal-exponential convolution) method to be the most adequate for our use-case since it's the most commonly used for two-color microarray background correction.

2.2.2 Data Normalization

The data in the provided files is not normalized - and therefore we can not rightfully compare the intensities recorded as some variables contribute more than other to the model estimation - so we will use the limma package to center and scale the intensities of `chip1.txt`, `chip2.txt` and `chip3.txt`. The normalization will be performed within arrays and between arrays.

Normalization will be done in two stages:

- Within-array normalization corrects dye biases and spatial artifacts (e.g., using LOESS to balance the red (R) and green (G) channels).

- Between-array normalization (e.g., quantile normalization) aligns intensity distributions across all samples to make them comparable.

To ensure our microarray data is reliable for downstream analysis, we must verify that normalization effectively removes technical biases while preserving biological signals. This is where MA plots and boxplots become essential diagnostic tools.

Boxplots show how data is spread out using five key values: the smallest data point (minimum), the 25th percentile (Q1), the middle value (median), the 75th percentile (Q3), and the largest data point (maximum). These help us see if data is balanced across samples.

MA-plots compare two important measurements for each gene:

1. The ratio of red to green intensities (M-value);
2. The average intensity of both colors (A-value).

Normalization Within Arrays In our analysis comparing arterial and venous gene expression patterns from saphenous vein samples, we needed to properly calibrate the two-color microarray measurements. This calibration step centers the log-ratio differences between arterial (red channel) and venous (green channel) conditions around zero for each patient's data, ensuring the biological comparisons aren't distorted by technical variations in the experimental measurements. We evaluated two different approaches to achieve this balance:

Method "none":

```
MA.raw <- normalizeWithinArrays(dd, method="none")
```

Method "loess":

```
MA.with <- normalizeWithinArrays(dd.correct,method="loess")
```

The first approach makes no adjustments, providing a baseline for comparison. The second method uses sophisticated curve-fitting to account for measurement inconsistencies that vary by intensity level.

Each method has particular strengths depending on the data characteristics. The unadjusted approach serves only as a reference point to show how much correction was needed. The curve-fitting approach generally works best for most biological comparisons as it handles subtle but important variations in the measurements.

We compared the results using boxplots and MA-plots (Figure 1 and Figure 2) to see which method performed best. The method "none" showed high variability between samples, with uneven spreads and many outliers. In contrast, the Loess method significantly reduced that variability, producing more balanced distributions centered around zero.

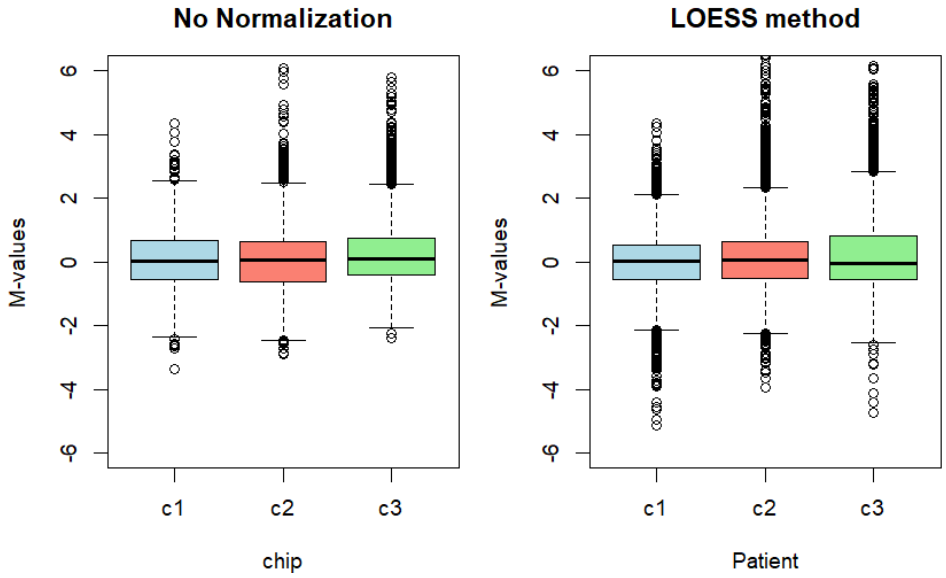


Figure 1: Boxplots of the different distribution of log-intensities within arrays, using the normalization methods aforementioned. Before normalization, medians and spreads often vary between samples due to technical variability; after normalization, boxes should align, reflecting consistent distributions.

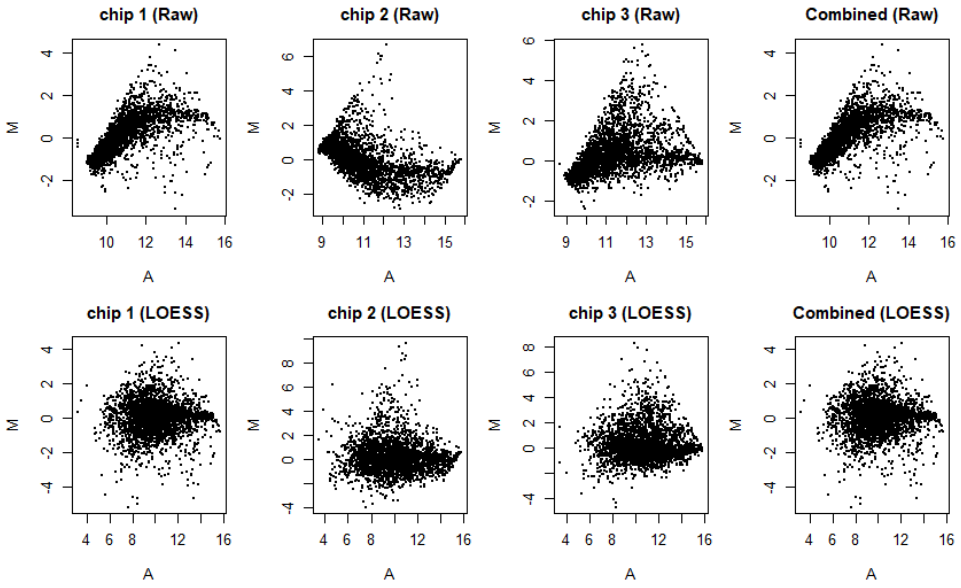


Figure 2: MA-plots for the aforementioned normalization methods within arrays. MA plots visualize the log-ratio of intensities (M = log(R/G)) against the average log-intensity (A = $\frac{1}{2}$log(R×G)) for each gene. Before normalization, we expect asymmetric scatter around M=0 due to technical biases; after normalization, points should symmetrically cluster around M=0, confirming successful dye bias removal.

The MA-plots revealed similar patterns. The raw data had strong curved trends, suggesting uneven dye effects or technical biases. The Loess method flattened these curves effectively, with points evenly spread around the zero line, indicating successful correction.

When comparing different ways to adjust the data within each sample, we chose the approach that best balanced and standardized the measurements. Looking at the distribution plots, both methods tested removed systematic biases effectively. However, the Loess method stood out because it reduced

variability in the boxplots and eliminated intensity-dependent biases in the MA-plots, making it the best choice for our study.

Normalization Between Arrays In this normalization step, the data will be adjusted to make sure the measurements are comparable across all samples. Since the Loess approach worked best for balancing individual samples, we used the same method to standardize measurements between different samples.

```
MA.bet2 <- normalizeBetweenArrays(MA.with, method="scale")
boxplot(MA.bet2$M,names=colnames(MA.bet2$M),col=rainbow(8))
```

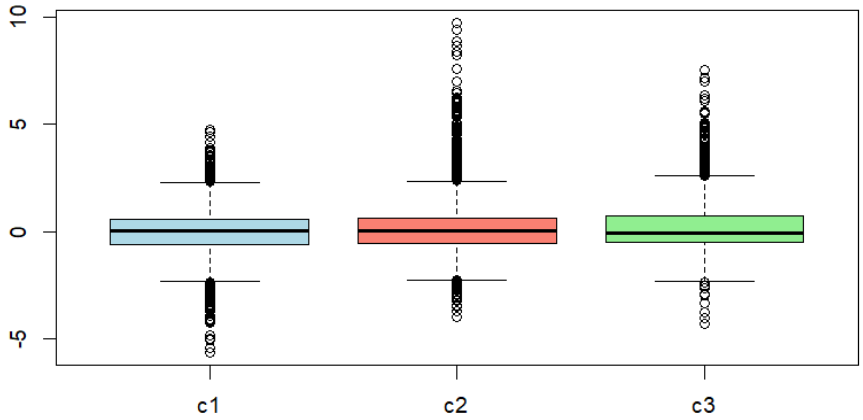


Figure 3: Boxplots of the different distribution of log-intensities between arrays, using the Loess method

3 Applying the Bayesian Method of Lonnstedt and Speed

To identify genes with meaningful expression changes between conditions, we use a statistical approach that accounts for variability in the data. This method fits a mathematical model to each gene's measurements across samples, then applies adjustments to improve reliability. The key output helps pinpoint which genes show true biological differences rather than random noise.

```
model <- lmFit(MA.bet2$M)
fit <- eBayes(model)
top_genes <- topTable(fit, coef = 1, adjust.method = "fdr", p.value = 0.05)
```

The method `eBayes()` applies empirical Bayes moderation: it takes raw variances from `lmFit()` and stabilizes extreme variances towards a common value by borrowing information from all the genes. This is crucial in our study since we only have 3 samples, which are bound to result in unreliable variances per gene. Through the visualization of the variance shrinkage we can confirm that the variances are stabilized.

```
plotSA(fit, main = "Variance Shrinkage via Empirical Bayes")
```

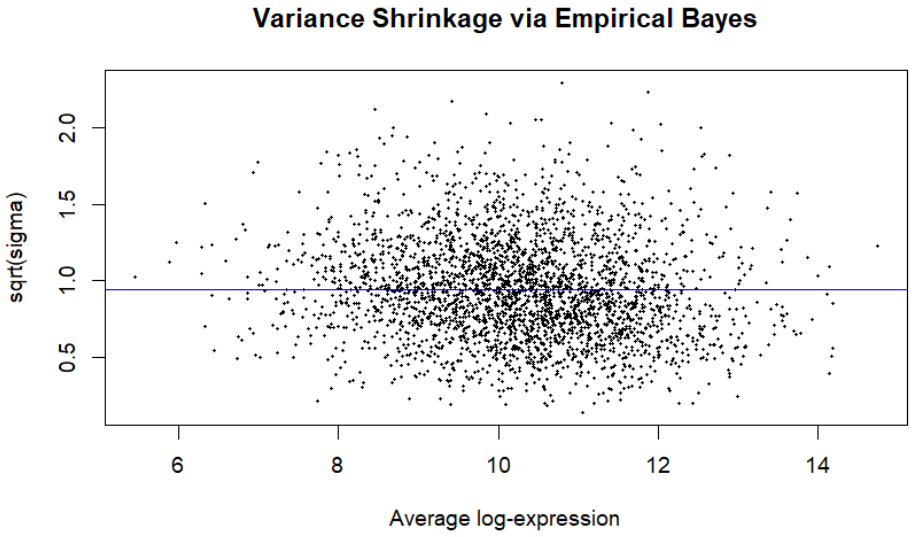


Figure 4: Shrinkage stabilized extreme variances via Empirical Bayes, which can be confirmed by the clustering of data points around the blue line.

To visualize the relationship between gene expression changes and their statistical significance, we will be creating a volcano plot. This type of scatter plot displays each gene as a single dot, with its position showing both the magnitude of gene expression change (horizontal axis) and the statistical significance of that change (vertical axis). This visualization helps quickly identify genes with both large and statistically reliable expression changes.

```
volcanoplot(fit,highlight=100,
main="Volcano Plot - Genes with Highest Differential Expression", cex=0.2)
```

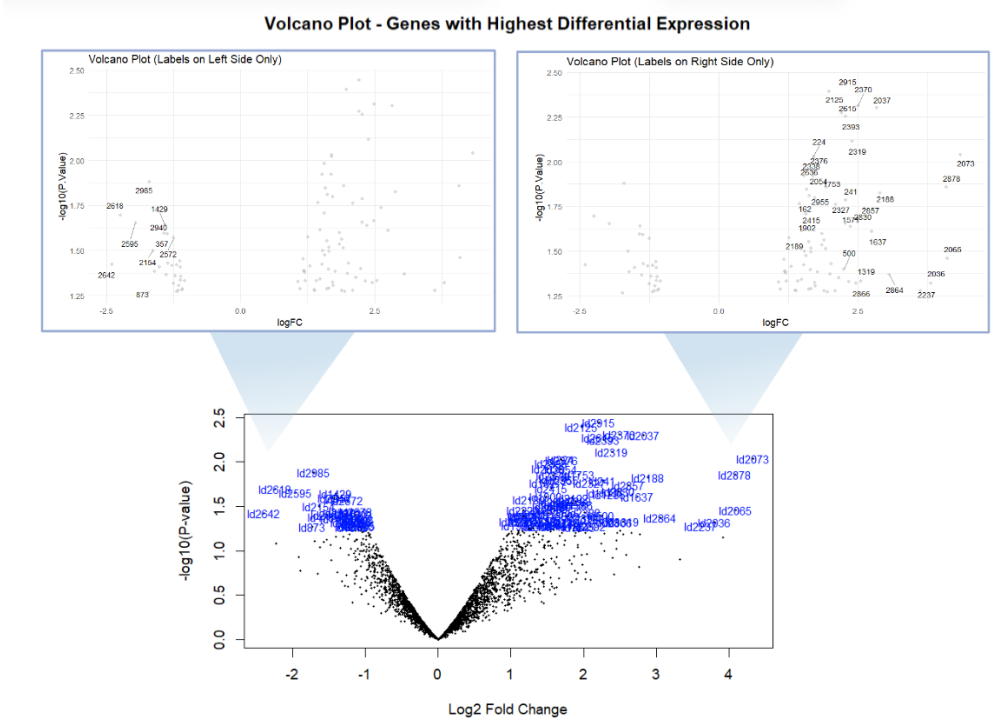


Figure 5: Representation of the volcano plot for 100 genes using a fitted Bayesian Model. This plot presents the -log10(p-values) versus the log2(fold change) of each gene. The ggrepel package was used to repel overlapping text labels and to create a zoomed version of the left and right sides.

The volcano plot highlights genes with strong expression differences and high confidence levels, appearing in the top left or top right regions. Genes positioned on the right (log2 fold change ¿ 0) show increased expression, while those on the left (log2 fold change ¡ 0) show decreased expression. Central and lower-positioned genes reflect changes too small to be statistically meaningful. Our results reveal more genes with increased expression than decreased.

Next, we ranked genes by their statistical confidence (log-odds of differential expression) and listed the top 100 most significant candidates.

```
table1 <- topTable(fit, adjust.method="BH", sort.by="B", number=100)
table1
```

The standard threshold for significance in these analyses was set at log-odds values greater than zero. However, when examining the sorted results, no genes met this criterion. The adjusted p-values for all genes remained consistently high (approximately 0.9 when transformed via -log10), suggesting none showed statistically meaningful differences. This outcome indicates that, based on our data and the chosen analytical approach, we lack sufficient evidence to identify any genes with significant expression changes between the tested conditions and we are then unable to reject the null hypothesis.

4 Applying the Moderated t-Statistic

In genomic studies with small sample sizes, standard t-statistics can suffer from high variability in variance estimates, reducing the power to detect true differential expression. To address this, a moderated t-statistic—implemented in the LIMMA package—applies empirical Bayes methods to stabilize variance estimates by "borrowing information" across genes, effectively shrinking gene-specific variances toward a common pooled estimate and increasing reliability under limited replicates [4].

The moderated t-statistic follows a classical t-distribution with adjusted degrees of freedom, allowing significance testing without relying on non-parametric approximations [4].

When identifying genes with significant expression changes, an alternative method involves analyzing the adjusted t-statistics. Using the same ranking function, we can extract a sorted list of the most promising candidates, ordered by the strength of their modified t-values (sorted using the "t" parameter). This approach highlights genes with the most reliable and pronounced log-ratio differences.

```
table2 <- topTable(fit, adjust.method="BH", resort.by="t", number=100)
table2
```

The analysis revealed the top 100 genes to be significant genes, with all adjusted p-values remaining below 0.05 as well as all other conventional thresholds. The identified genes with differential expression remain the same that were obtained while using the Bayesian Method of Lonnstedt and Speed, presented in the previous section.

5 Applying the RankProd Method

5.1 Assessing Statistical Significance

The RankProd method offers an alternative statistical method for detecting significant genes. This non-parametric statistical test performs a rank product analysis on the same log-transformed, normalized dataset, providing a different perspective on our dataset. It is able to detect genes that are consistently upregulated (or downregulated) in replicate experiments and performs well even in reduced sample sizes [6].

It allows to determine each gene's significance level and for flexible control of the false-detection rate [6]. As this method is being used in conjunction with limma, we considered it was acceptable to be more lenient in the cutoff point. In this case, we are allowing 60% of genes to be false positives and therefore to provide a less reliable list. This list will then be cross-validated with the top 100 genes found using the limma method.

```
aa <- RP(MA.bet2$M, c1 = c(1, 1, 1), logged = TRUE)
Best_genes_RP <- topGene(aa, cutoff = 0.6, method = "pfp")
Best_genes_RP
```

We were only able to find significant genes at very high FDR cutoffs (0.6-0.9). This can be due to several factors such as:

1. Subtle differences in gene expression between arterial and venous regimes (Figure 6);
2. Limited statistical power due to a sample of 3 patients, considering that the RankProd method requires consistent evidence across all replicates;
3. Potential patient-to-patient variability surpassing differences in gene expression.

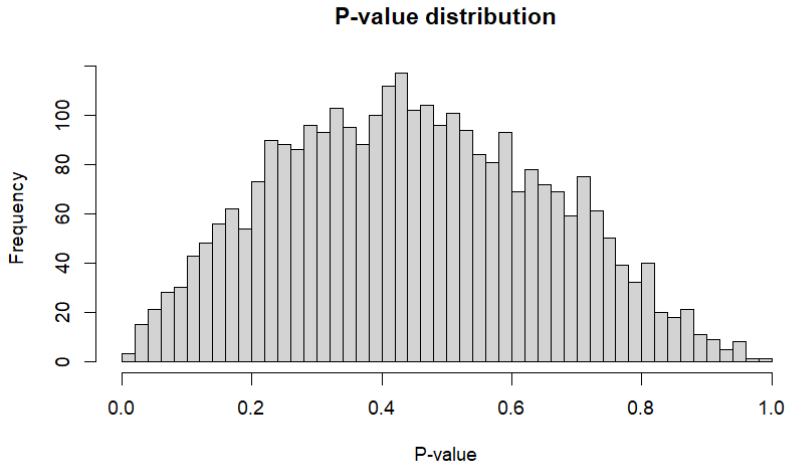


Figure 6: Most genes show similar expression between arterial and venous conditions. Only a small fraction of genes are truly differentially expressed and a sample size of 3 has limited power to detect small but real differences.

The results show that the significant genes found with RankProd exhibit low PFP values and 83% of the genes detected with RankProd at FDR ¡ 0.6 are present in the top 100 list of differentially expressed genes obtained from the limma method.

6 Assessing Biological Relevance

We will now shift focus to identifying genes showing biologically meaningful patterns in their expression levels, as even subtle log-fold changes may still hold clinical relevance. To do this we examined the top 100 genes ranked by their fold change magnitude and adjusted p-values.

```
Best_genes_RP <- topGene(aa, num.gene = 100, method = "pfp")
Best_genes_RP$Table2 <-
Best_genes_RP$Table2[order(Best_genes_RP$Table2[,3], decreasing = TRUE), ]
Best_genes_RP
```

This allows us to explore which genes have the highest fold change and therefore the highest biological differences.

```
barplot(Best_genes_RP$Table2[1:15,3],
        names.arg=paste0("Id", Best_genes_RP$Table2[1:15,1]),
        las=2, main="Top 15 Genes: Fold Changes (Arterial/Venous)",
        ylab="Fold Change", col="steelblue")
abline(h=10, col="red", lty=2)
```

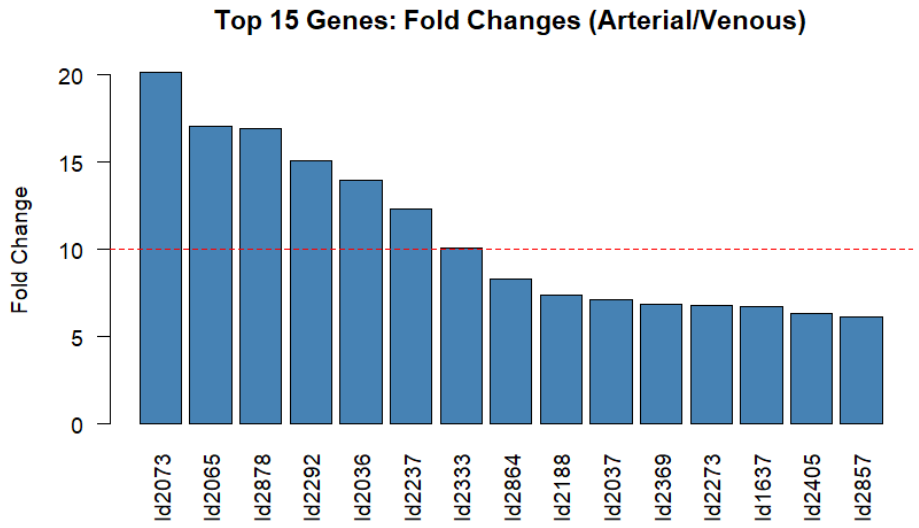


Figure 7: Genes with the highest biological impact. The first 5 genes have a pfp ¡ 0.3 with a FC ¿13. These genes have both large biological relevance and statistical significance.

7 Conclusion

By mitigating noise in variance estimation, LIMMA's package has allowed us to investigate differential gene expression in the microarray data provided, which is notable in a study with so few biological replicates. Similarly, the RankProd package allowed a non-parametric approach to identify differentially expressed genes, making this study reproducible and reliable for small sample sizes, where most genes show similar expression between arterial and venous conditions.

As the last step in our study, we will be comparing the outcomes from different statistical methods to see whether the top 100 genes remain consistent across the different approaches. This comparison focuses both on the statistical validity of these genes and on their biological relevance. Although our analysis was limited to three replicates, we considered a statistical significance of 0.05 and allowed a higher false discovery rate (FDR) to maximize detection sensitivity. This was a trade-off that we considered to be justified by the simultaneous use of Bayesian inference methods, which provided a broader gene pool for cross-validation.

Despite the anonymous gene identifiers (ID1–ID2994), the observed 10x to 20x increase in gene expression suggests strong biological relevance, and we would expect these genes to influence critical functional pathways (e.g. linked to revascularization and cardiac repair).

The Euler's diagram below (Figure 8) reveals a complete match between the genes identified using the Bayesian approach of Lönnstedt and Speed and those detected with the moderated t-statistic. Since they are based on the same statistical approach, with the first method using Bayesian probabilities and the second using empirical Bayes, it is expected that they would naturally identify the same pool of significant genes.

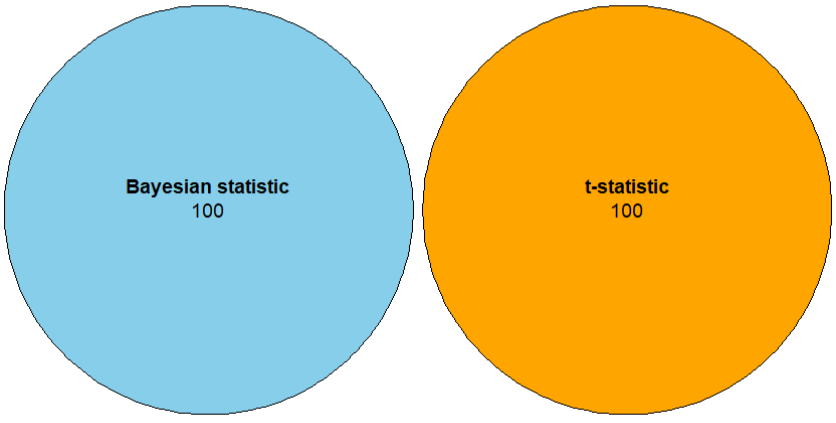


Figure 8: The 100 most significant genes identified through Bayesian method and moderated t-statistic are identical, ranked by their log-transformed statistical evidence and highest expression change.

When comparing results from the Bayesian approach with the Rank Product analysis, we observe only a small overlap in differentially expressed genes, as shown in Figure 9.

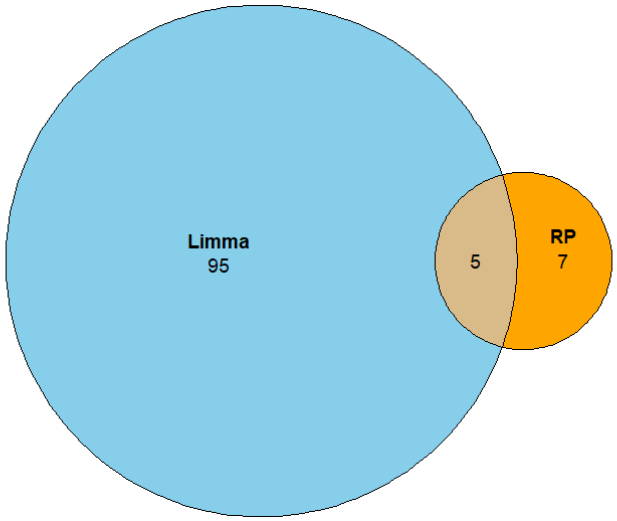


Figure 9: Overlap of the most significant genes identified through Bayesian method and RankProduct method, ranked by their log-transformed statistical evidence and highest expression change. Out of the 7 genes found by the RankProduct method, we considered those that simultaneously possess high differential expression and a low rate of false positives, leaving us with 5 genes.

Considering the results obtained through cross-validation, when considering both statistical significance and biological relevance, we would expect that these 5 genes would be important for understanding vascular adaptation and potential therapeutic targets, contributing to the decrease of vein graft complications and better patient outcomes.

References

[1] McQueen, L.W.; Ladak, S.S.; Layton, G.R.; Wozniak, M.; Solomon, C.; El-Dean, Z.; Murphy, G.J.; Zakkar, M. Spatial Transcriptomic Profiling of Human Saphenous Vein Exposed to Ex Vivo Arterial Haemodynamics—Implications for Coronary Artery Bypass Graft Patency and Vein Graft Disease. *Int. J. Mol. Sci.* 2024, 25, 10368. https://doi.org/10.3390/ijms251910368

[2] Haron, N.A.; Ishak, M.F.; Yazid, M.D.; Vijakumaran, U.; Ibrahim, R.; Raja Sabudin, R.Z.A.; Alauddin, H.; Md Ali, N.A.; Haron, H.; Ismail, M.I.; et al. Exploring the Potential of Saphenous Vein Grafts Ex Vivo: A Model for Intimal Hyperplasia and Re-Endothelialization. *J. Clin. Med.* 2024, 13, 4774. https://doi.org/10.3390/jcm13164774

[3] Ho, M., Yang, E., Matcuk, G., Deng, D., Sampas, N., Tsalenko, A., Tabibiazar, R., Zhang, Y., Chen, M., Said Talbi, Ho, Y.D., Wang, J., Tsao, P.S., Amir Ben-Dor, Zohar Yakhini, Bruhn, L. and Quertermous, T. (2003). Identification of endothelial cell genes by combined database mining and microarray analysis. *Physiological Genomics*, 13(3), 249–262. https://doi.org/10.1152/physiolgenomics.00186.2002.

[4] Wang G, Muschelli J, Lindquist MA. Moderated t-tests for group-level fMRI analysis. *Neuroimage*. 2021 Aug 15;237:118141. doi: 10.1016/j.neuroimage.2021.118141. Epub 2021 May 4. PMID: 33962000; PMCID: PMC8295929.

[5] Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* 2004;5(10):R80. doi: 10.1186/gb-2004-5-10-r80. Epub 2004 Sep 15. PMID: 15461798; PMCID: PMC545600.

[6] Breitling, R., Armengaud, P., Amtmann, A., & Herzyk, P. (2004). Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS letters*, 573(1-3), 83–92. https://doi.org/10.1016/j.febslet.2004.07.055