Context-based Analysis of NGS data on Complex Human Diseases: Atherosclerosis

GoBI: Block and Project Phase - Alexander Fastner, Franziska Koller (Group 7)

Introduction

Complex diseases such as Atherosclerosis require thorough analysis and understanding for an effective and efficient treatment. Atherosclerosis is a chronic disease of the arterial wall that forms plaques inside the arteries which slows down and hinders blood flow. Often coronary arteries are affected. Due to critical blood and thereby oxygen shortages, atherosclerosis can lead to heart attacks and strokes. Especially disruptured plaques that start to clot inside the arteries block the blood flow and thereby the oxygen supply in important cells.

Up to now it is not entirely clear what causes atherosclerosis. Researchers have shown that toxins such as nicotine, permanent high blood pressure (hypertension), raised cholesterol and obesity contribute to the risk. These factors supposedly initiate damages to the inner arterial cell wall when present over a long period of time. The damaged arterial wall enables e.g. cholesterol to accumulate at the arterial walls.

Macrophages (white blood cells and part of the immune system) play an important role in the immune response to the plaques inside the arteries.
M1 (classically activated) macrophages are known to encourage inflammation and M2 (alternatively activated) macrophages decrease inflammation and encourage tissue repair. Macrophages are involved in nonspecific defense as well as the initiation of specific defense mechanisms of the cell against virus infections and other pathogens. Phagocytosis is the process by which a cell, in this case a macrophage, engulfs and digests viruses and pathogens. [1] [2]
In atherosclerosis macrophages digest the accumulated lipoproteins at the inner arterial wall.
After their cell death macrophages stay at the arterial wall as foam cells (=dead macrophages filled with cholesterol). Instead of decreasing the plaque at the arterial wall they unfortunately contribute to it. As the plaque keeps growing smooth muscle

[1] M. Amine Bouhlel et al.: PPARγ Activation Primes Human Monocytes into Alternative M2 Macrophages with Anti-inflammatory Properties, *Cell Metabolism* , 2007, https://doi.org/10.1016/j.cmet.2007.06.010

[2] Wynn, T., Chawla, A. & Pollard, J. Macrophage biology in development, homeostasis and disease. *Nature* **496,** 445–455 (2013). https://doi.org/10.1038/nature12034

cells from the outer arterial wall start to migrate towards the plaque and engulf it to shield it from the bloodstream and further growth.

However, if the plaque now filled with lipoproteins and other cellular material (e.g. calcium) ruptures it will start to clot and possibly block the entire bloodstream through the artery. [3]

The exact trigger of the accumulation of lipoproteins that form the plaques inside the arteries still leaves many open questions. [4]

RNA sequencing (RNA-seq) is one of the most efficient approaches to biological and medical questions at hand. Gene expression levels in different cell types can be analyzed and give detailed insight to biological processes and molecular functions.

In the following we examine two data sets with different conditions and separately for M1 and M2. The effects of the treatments and the differences of M1 and M2 cells will be analyzed in detail. Additional materials are available in the attached RMarkdown files.

Analysis of Data Set project3

The data set “project3” contains the data of 24 samples (4 groups with each 6 replicates) and includes 33,260 mouse genes. The analyzed cells are Bone-marrow-derived macrophages (BMDM).

With this data set the effects of **Let-7b** deletion on macrophage transcriptomes can be analyzed.

Let-7b is a miRNA (microRNA) that is part of several processes.

In general miRNAs are short non-coding RNAs involved in post-transcriptional gene regulation. Usually they are part of the RNA-induced silencing complex (RISC). The RISC recognizes target mRNAs by pairing them with the miRNA. This can lead to translational inhibition or destabilization of the target mRNA. [5]

The names of microRNAs usually include a number which indicates the order of the naming, in this case Let-6 was named before Let-7. The letters following the number such as ‘b’ in Let-7b refer to a high similarity between the microRNAs with the additional lower case letters.

[3] https://www.youtube.com/watch?v=g3kDdg8r6NY

[4] University of Virginia Health System. "Fundamental beliefs about atherosclerosis overturned: Complications of artery-hardening condition are number one killer worldwide." ScienceDaily. ScienceDaily, 6 July 2015. <www.sciencedaily.com/releases/2015/07/150706123730.htm>.

[5] “MIRLET7B Gene – MircorRNA Let-7b”. GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=MIRLET7B. Accessed on 2022/03/13.

Sometimes microRNAs are denoted with a -3p or -5p suffix. This means that they either come from the 3'-arm or the 5'arm of the pre-miRNA. [6]

Table 1 gives an overview of the 4 groups where KO_M1 and KO_M2 are the groups with the Let-7b deletion and WT_M1 and WT_M2 are the wild type groups that function as control groups from two different mice.

Group name	**Condition**	**mouse_id**
KO_M1	Let-7b deletion	335.73
KO_M2	Let-7b deletion	335.73
WT_M1	wild type	330.97
WT_M2	wild type	330.97

Table 1: overview of groups in the data set "project3"

We extracted the gene id of the gene with the maximum number of counts per sample. The data was taken from *project3_raw.readcounts.csv*. Interestingly in all 24 samples only 4 different genes have the maximum number of counts. These genes are:

gene_id	gene_name	gene_length	additional info
ENSMUSG00000024661.7	Fth1[7]	4512	Ferritin heavy chain (ferroxidase enzyme)
ENSMUSG00000029580.14	Actb[8]	3640	Beta actin
ENSMUSG00000050708.16	Ftl1[9]	1923	ferritin light polypeptide
ENSMUSG00000106106.3	CT010467.1[10]	213445	/

Table 2: information on genes with the maximum number of counts

We decided to analyze the number of counts per group. This gives insight into how well

[6] Wikipedia contributors. MicroRNA. Wikipedia, The Free Encyclopedia. February 19, 2022, 18:27 UTC. Available at: https://en.wikipedia.org/wiki/MicroRNA#Nomenclature. Accessed March 12, 2022.
[7] "Fth1". NCBI. https://www.ncbi.nlm.nih.gov/genome/gdv/browser/gene/?id=14319
[8] "Actb". NCBI. https://www.ncbi.nlm.nih.gov/genome/gdv/browser/gene/?id=11461
[9] "Ftl1". NCBI. https://www.ncbi.nlm.nih.gov/gene/14325
[10] "CT010467.1". NCBI. https://www.ncbi.nlm.nih.gov/nuccore/CT010467.1/

the different groups can be compared and should be kept in mind when drawing conclusions. A group with a significantly higher number of counts is not necessarily more important than other groups; the reason most likely lies in technical and biological sequencing circumstances.

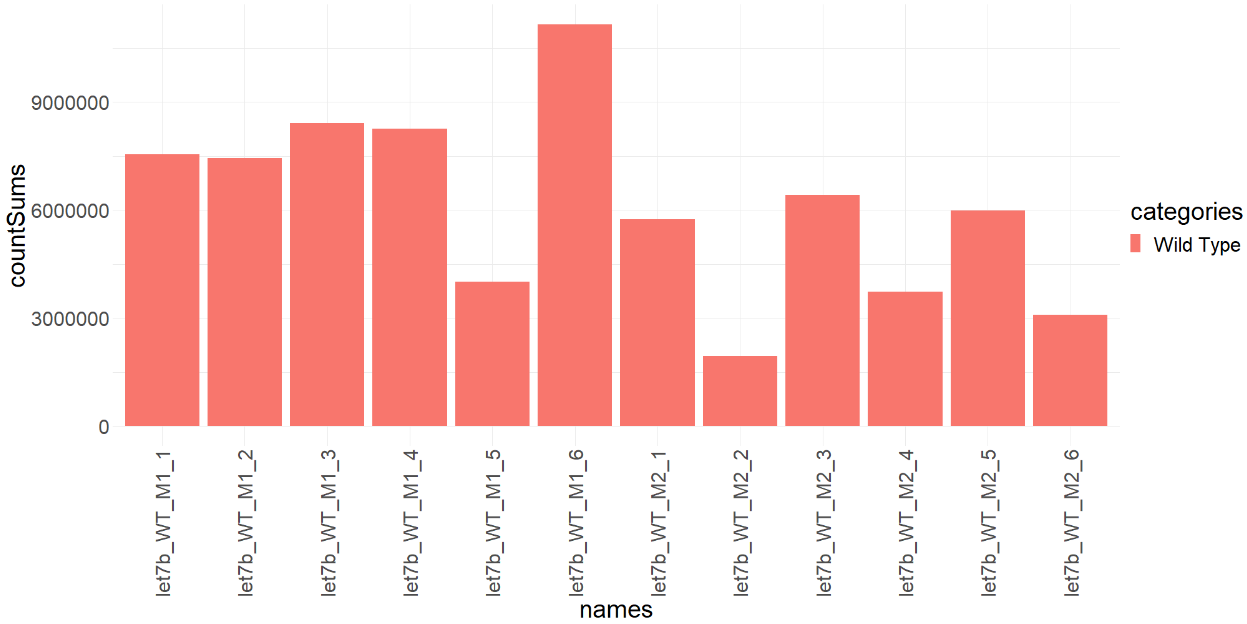


Figure 1: Count distribution of wild type samples in the project3 data set

This plot shows the count distribution for all replicates of the wild type condition.
For each replicate the number of counts was summed up over all genes and visualized in the bars.The x-axis shows the sum of all counts, the y-axis shows all 12 groups of the wild type condition.
Replicate 6 of M1 (let7b_WT_M1_6) has the highest number of counts with over 11 million (11 152 346). Replicate 2 of M2 (let7b_WT_M2_2) only has ~2 million read counts (1 947 233).

Analysis with DESeq2

We conducted a differential gene expression analysis based on the negative binomial distribution with DESeq2 [11].

It gives an overview of the differentially expressed genes in the data set. It identifies the up-regulated and down-regulated genes based on the number of counts.

[11] Love MI, Huber W, Anders S (2014). “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.” *Genome Biology*, **15**, 550. doi: 10.1186/s13059-014-0550-8.

In project3 **1003** genes are up-regulated and **1018** genes are down-regulated from a total number of **33,260** genes.

Up-regulated genes were defined as a log2Foldchange > 2 and a p-value < 0.05. Down-regulated genes have a log2Foldchange < -2 and a p-value < 0.05.

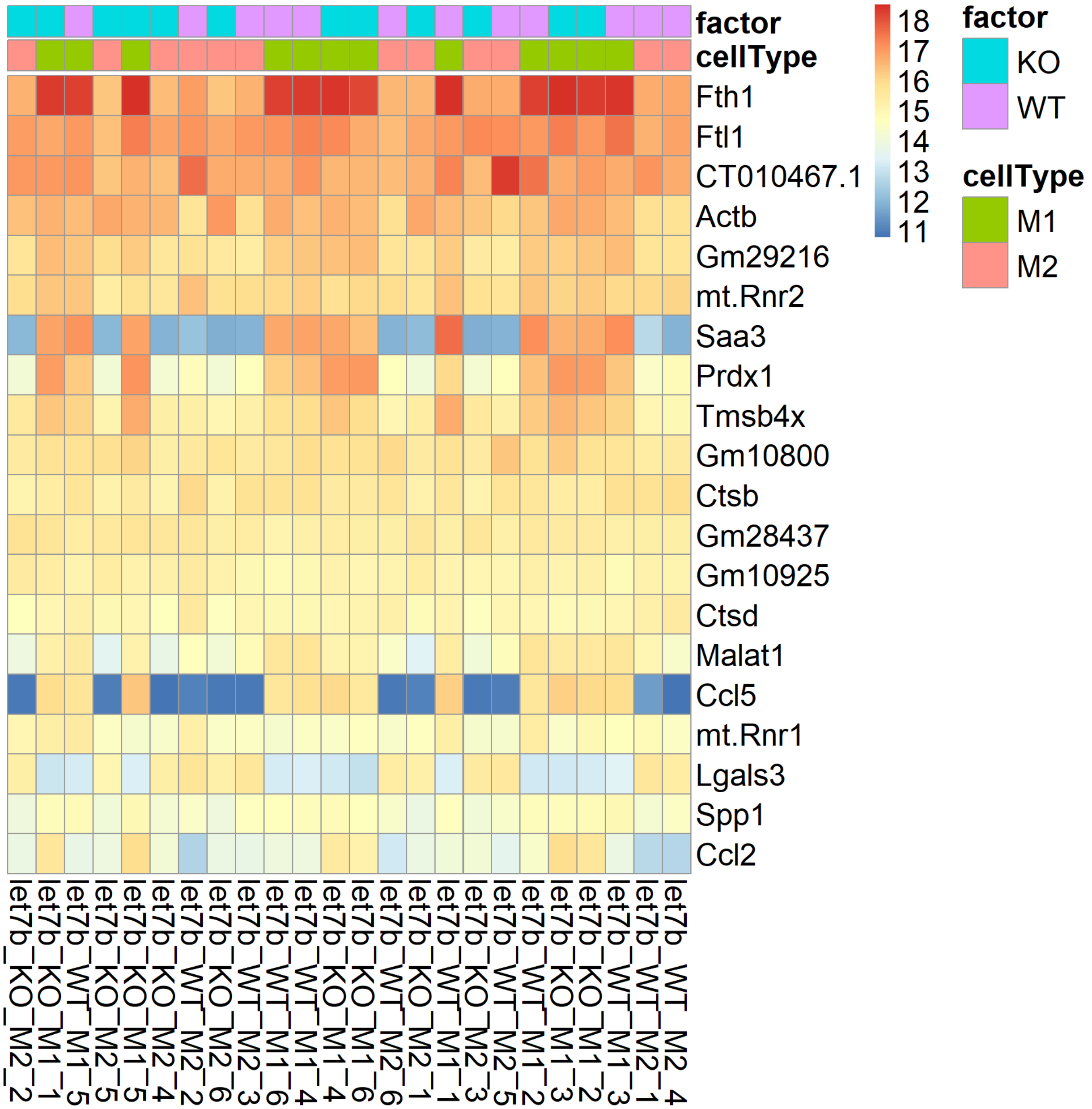


Figure 2: Heatmap for the top 20 genes of project3. We differentiated between the macrophage type M1 (green) and M2 (pink) and between the condition wild type (purple) and the Let-7b deletion (blue). Each square represents the level of gene expression, red represents an up-regulation of the gene, blue represents a down-regulation of the gene.

One can see a strong clustering by cell type (M1/M2) and a weak clustering by WT/KO. With this in mind we support this hypothesis with a PCA. The Fth1 gene shows a strong up-regulation for all M1 macrophage samples. The Ccl5 gene in contrast is significantly lower expressed in the M2 samples.

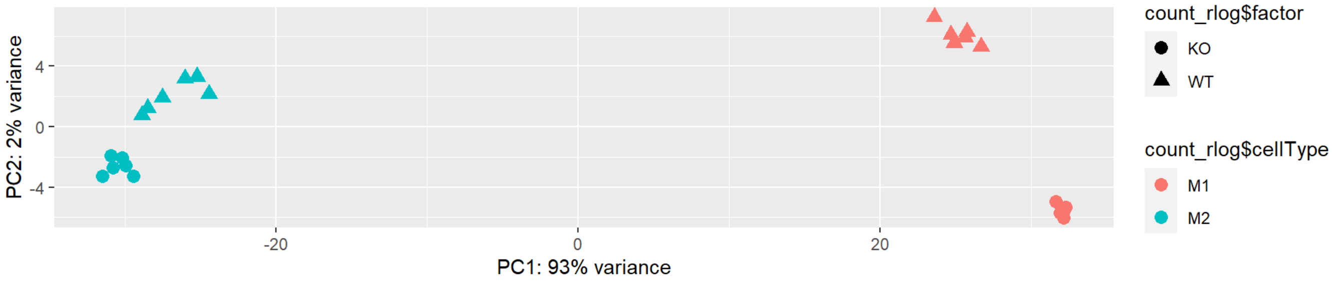


Figure 3: Principal Component Analysis (PCA) analyzing the variance induced by the two factors of M1/M2 and KO/WT. The samples are plotted with the color and shape of their cell type and factor. The X and Y axis show the amount variance in input factors and show clusters of various samples.

There is a clear clustering of both M1 and M2 and smaller internal subgroups for KO and WT. As we also see in the heatmaps the difference in M1 and M2 is much more significant than that between WT/KO.

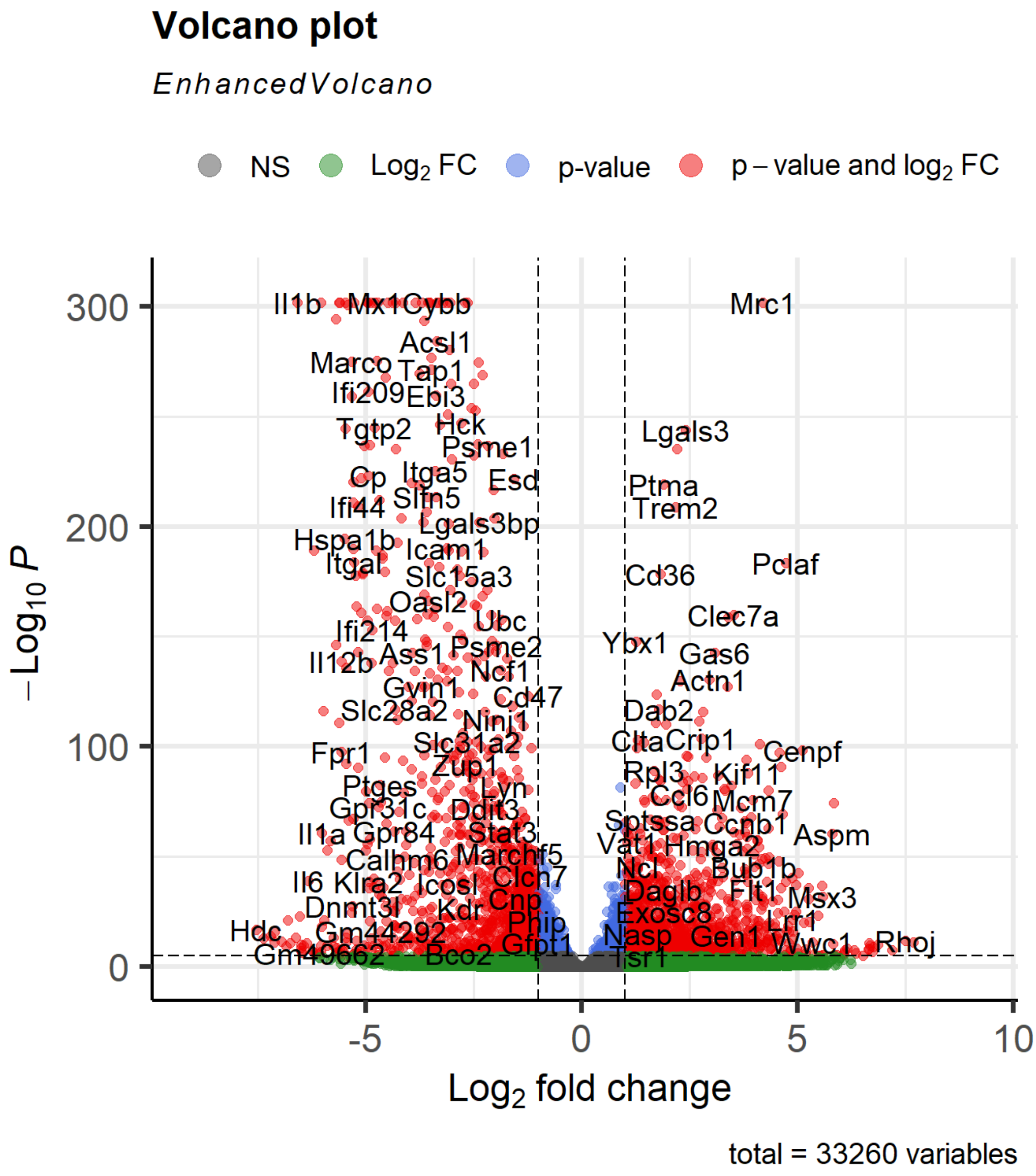


Figure 4: Volcano plot for RNA-seq count data of project3. The y-axis shows the -log10(p-value) and the x-axis the log2foldchange, both computed with DESeq2. Significant genes have a p-value<0.05 and a log2foldchange < -2 or >2. The red points represent the significant genes, the green blue and gray points represent the insignificant genes.

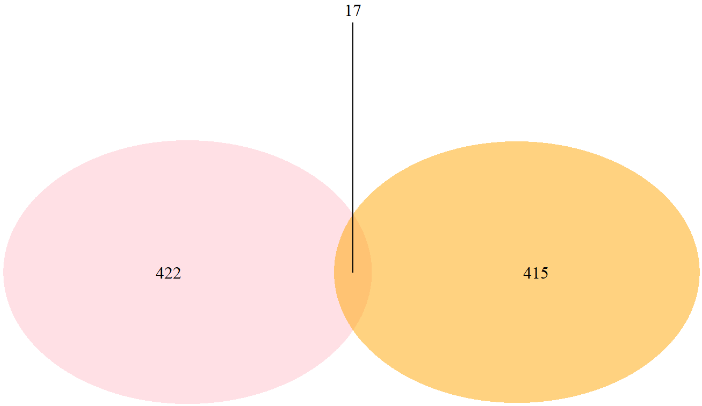


Figure 5: The Venn-Diagram visualizes the number of significant genes in M1 (422 in pink) and in M2 (415 in orange). The intersection (i.e. the number of genes that is significant in M1 *and* M2 is 17).

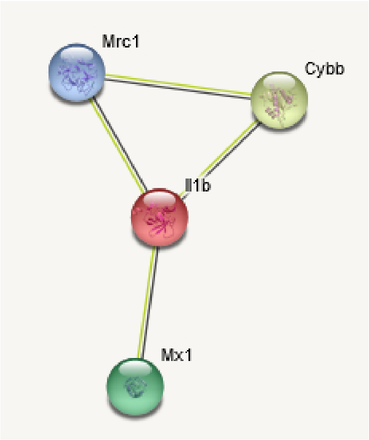


Figure 6: Interaction (Co-Expression) of Top4 genes(Mrc1, Cybb, Il1b and Mx1) from volcano plot visualized with STRING. [12]

The volcano plot in Figure 4 visualizes the significant genes of the project3 data set (in red).

The top 4 significant genes are known to be co-expressed (analyzed with STRING, Figure 4). The interaction of Mrc1 and Il1b could be relevant for atherosclerosis as Mrc1 is usually involved in endocytosis by macrophages and Il1b is a proinflammatory cytokine. Macrophages play an important role in atherosclerosis since they digest

[12] Szklarczyk D*, Gable AL*, Nastou KC, Lyon D, Kirsch R, Pyysalo S, Doncheva NT, Legeay M, Fang T, Bork P‡, Jensen LJ‡, von Mering C‡.
The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets .
Nucleic Acids Res. 2021 Jan 8;49(D1):D605-12.

cholesterol in the arteries but remain at the arterial walls as foam cells and contribute to the atherosclerotic plaques.

Gene Set Analysis with DAVID

The Database for Annotation, Visualization and Integrated Discovery (DAVID) can be used to analyze large lists of genes. Multiple functional annotation tools help to understand the biological background and meaning of selected gene sets.

The up-regulated and down-regulated significant genes that were identified with DESeq2 from the project3 data set were analyzed with the Functional Annotation Clustering from DAVID.

Clusters in DAVID are defined as a “group of terms having similar biological meaning due to sharing similar gene members”. The background in our case was Mus musculus as provided by DAVID.

The up-regulated genes resulted in 145 DAVID clusters.

The cluster with the highest enrichment score (65.67) includes 122 genes and is characterized as a biological process that is part of the cell cycle, especially the cell division.

The down-regulated genes resulted in 121 clusters.

The cluster with the highest enrichment score (63.98) includes 119 genes and is also characterized as a biological process, in this case involved in immune system responses.

Especially the exact development of immune system responses are relevant when working with chronic diseases such as atherosclerosis. A problem of atherosclerosis is that the immune response involves macrophages which digest toxins of the arterial blood flow but at the same time contribute to the plaques by accumulating as foam cells at the inner arterial wall.

Recount3

Recount3 is a publicly available collection and database for biological and medical datasets that also provides several tools and resources to analyze the data at hand. It includes a total of 316,443 human and 416,803 mouse run accessions (individual datasets) collected from the Sequence Read Archive (SRA) and other public projects such as The Genotype-Tissue Expression (GTEx) project. [13] These human and mouse data sets are RNA sequencing (RNA-seq) samples that were processed by the specifically developed Monorail analysis pipeline. The goal of the resource is to grant easy and efficient access to relevant and publicly available data sets and studies and combine them in a single database.

Recount3 provides several online tools as well as R/bioconductor packages that can be applied on datasets from recount3 and also on private data which can then easily be analyzed, compared and combined with any other recount3 dataset.

Data sets are available for exon-exon junctions, genes and also other features. Users can also process their own data with the Monorail pipeline and make it available on recount3 for others.

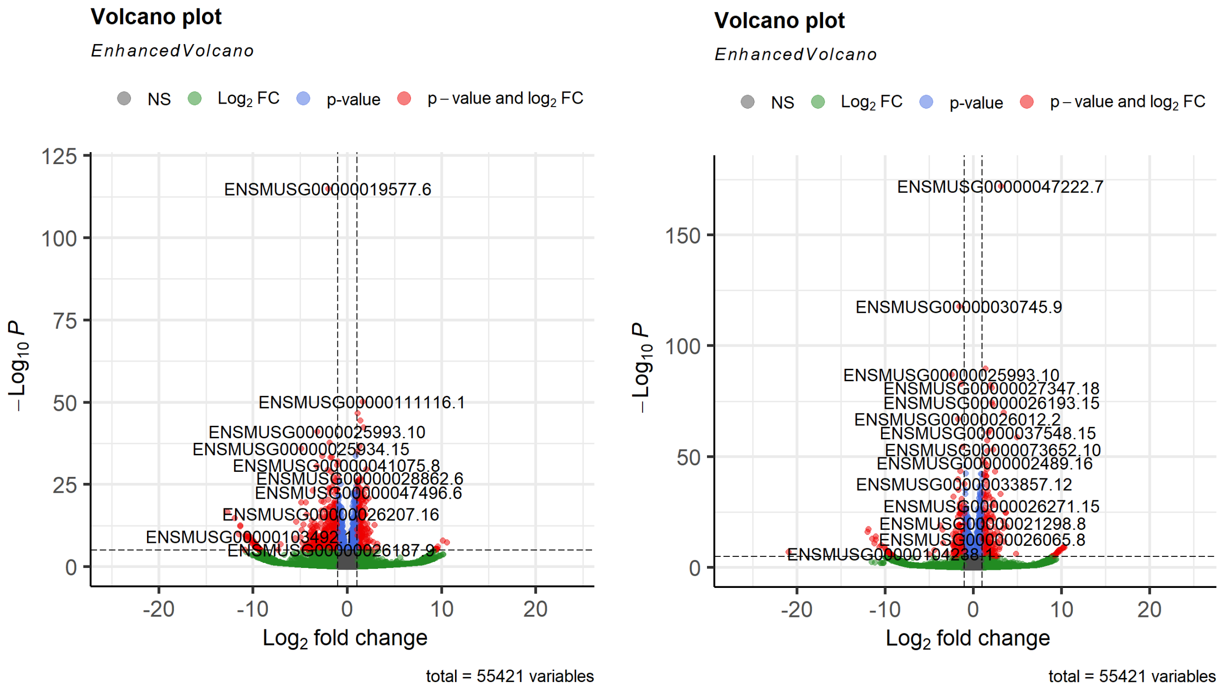
Analysis of dataset from recount3

We selected the RNA-seq data set from the study SRP136558. The data set was originally used to study gene expression changes in adipose tissue macrophages (ATM1 and ATM2) and adipocyte progenitors (AP) following cold exposure or FGF21-mimetic antibody administration.

It includes 24 samples. We focussed on the 12 samples that were generated from the macrophages M1 and M2 and treated with cold exposure.

The analysis of the adipose tissue macrophages could be relevant for our hypotheses on the dataset project3 as obesity is one of the major risk factors for atherosclerosis.

[13] Wilks, C., Zheng, S.C., Chen, F.Y. et al. recount3: summaries and queries for large-scale RNA-seq expression and splicing. Genome Biol 22, 323 (2021). https://doi.org/10.1186/s13059-021-02533-6



Figures 7 & 8:

Volcano Plots of significant genes in M1 and M2 respectively. Significant genes are shown in red and are defined as those which have a p-value < 0.05 and a Log2 fold change > 2 or < -2.

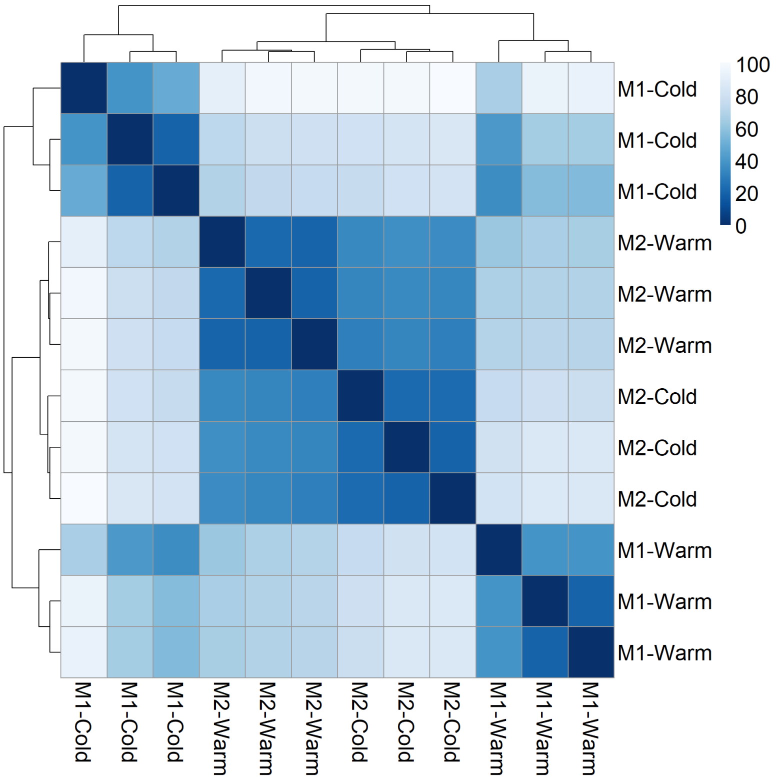


Figure 9:
This sample-sample heatmap shows the distances between the samples as clusters. Of note here is the clear clustering of M1 and M2 groups with smaller but still noticeable similarities for warm and cold.

The volcano plots in Figure 7 (for all M1 samples) and Figure 8 (for all M2 samples) show that the recount3 dataset includes several genes that are differentially expressed, either up-regulated or down-regulated. The distance heatmap in Figure 9 visualizes a clear clustering between M1 and M2. The samples with either cold- or warm exposure do not show such a clear clustering.

Conclusion and Outlook

blitzGSEA

blitzGSEA is an algorithm for efficient gene set enrichment analysis.
It uses the same running sum statistic as the GSEA algorithm. GSEA is one of the most popular tools for computing statistical tests that compare differential expression between gene sets.
In contrast to GSEA blitzGSEA does not use permutation tests to compare differential expression in the annotated gene sets. Instead blitzGSEA uses gamma distributions to approximate the enrichment score probabilities.
This leads to an improvement in the performance and a better approximation of very small p-values. Small p-values are especially a problem as multiple testing correction methods such as Bonferroni and Benjamini-Hochberg fail to provide good results for very small p-values. [14]
For further analysis blitzGSEA could be applied and the results could be compared to the output of the gene set enrichment analysis with DAVID and ClusterProfiler.

From our analysis on the Project 3 and SRP136558 datasets we conclude that due to the significantly stronger clustering of M1 and M2 as opposed to the various analyzed factors that a separated analysis is more effective. The variation in measured data is mostly attributable to the differences in M1 and M2 although clear secondary factor clusters are also identifiable for both datasets.

[14] Alexander Lachmann, Zhuorui Xie, Avi Ma'ayan, blitzGSEA: efficient computation of gene set enrichment analysis through gamma distribution approximation, *Bioinformatics*, 2022;, btac076, https://doi.org/10.1093/bioinformatics/btac076