RNA-seq analysis of multiple ApoE -/- mice datasets reveals characteristic expression profiles in atherosclerosis

Group 3: Sonja Stockhaus and Mahima Arunkumar

Abstract:

One of the most common causes of death in the world is a chronic degenerative disease called atherosclerosis. It is a complex and very dangerous cardiovascular disease, which typically progresses over long periods of time. In order to understand the underlying characteristic expression profiles of atherosclerosis, we investigated the effect of two microRNAs and of LPS-stimulation on macrophages in ApoE -/- mice. Regarding both microRNAs we found that only one dataset showed reliable results which is the dataset where *let7b* was deleted. Here we saw that the knockout has an impact on the regulation of inflammation. We also found cell type-specific differences between macrophage types M1 and M2 which were also visible in the transcriptional effects caused by the knockout. The analysis of the deletion of *miR-147* was not reliable as the replicates were very inconsistent. Regarding the LPS-time course dataset we found that even after short time intervals the reaction to LPS was rather strong as a very large number of transcripts, including immunoregulatory (cytokines, chemokines, etc.) genes, were altered significantly. However, we also found some inconsistencies which made us question if LPS is a perfect inflammation model after all.

Keywords: Atherosclerosis, ApoE -/- mice, *let7b*, *miR-147*, macrophages, LPS-stimulation, biomarker

1. Introduction

Atherosclerosis is a disease of the arteries that involves the build-up of plaque in the artery walls. This can not only disturb blood flow, there is also the danger of plaque rupture which can lead to thrombosis and in the worst case to death through myocardial infarction or stroke [1]. Pathogenesis of atherosclerosis involves the migration of monocytes into the subendothelial space and differentiation to macrophages [1]. They take up oxidized low-density lipoprotein (oxLDL) which transforms them into so-called “foam-cells”, containing lipid droplets [1]. Other cells, like smooth muscle cells (SMC) also move into the area of the building plaque which is covered by a fibrous cap [1]. This accumulation of cells (also called atherosclerotic lesion) can narrow the arterial lumen and rupture when it gets too unstable, potentially leading to thrombosis [1].

There are different subtypes of macrophages that play a role in atherosclerosis. They are called M1 and M2 and are specialized macrophages that perform different tasks in the immune system. The M1 type is also known as “inhibitor” or “killer” type as it is responsible for killing pathogens [2]. This involves the production of reactive oxygen species and secretion of pro-inflammatory cytokines [2]. M2 macrophages on the other hand are known for healing and growth promoting functions like tissue repair and cell proliferation [2]. For example, they show an increased production of ornithine (synthesized by the enzyme arginase), which is a precursor of collagen that is an important structural element of the extracellular matrix [2].

To understand the cause and disease progression mechanisms of Atherosclerosis with the goal of finding a cure as well as effective prevention techniques, we need a suitable model organism for this disease. It turns out that atherosclerosis-prone apolipoprotein E-deficient (ApoE−/−) mice display poor lipoprotein clearance with subsequent accumulation of cholesterol ester-enriched particles in the blood, which promote the development of atherosclerotic plaques [3]. These atherosclerotic lesions seem to be morphologically identical to those found in humans [3]. Therefore, the introduction of this animal model of atherosclerosis has truly revolutionized our understanding of the atherogenic process and has allowed the mouse to quickly become the most popular mammalian model of atherosclerosis to date.

Here, we investigated the effect of two microRNAs and LPS-stimulation on macrophages of atherosclerotic mice. ApoE -/- mice were fed with high cholesterol diet in order to induce atherosclerosis. On the one hand, we looked at the impact of LPS-stimulation on mural macrophages by analyzing time-series RNA-Seq data. On the other hand, we examined the effect that the deletion of specific micro-RNAs, namely *let7b* and *miR-147*, has on macrophages derived from the bone-marrow or atherosclerotic lesions.

Micro-RNAs are non-coding RNA molecules of small length. They take part in gene regulation by attaching to mRNA and thereby blocking translation. Like this, the expression of micro-RNA can downregulate other genes without interfering their transcription.

Here, the transcriptome of samples was measured by using the so-called prime-seq protocol. It is a type of RNA-sequencing that was developed to reduce noise originating from amplification [4]. In order to do so, the cDNA from every sample is not only labelled with a sample-specific barcode, but every cDNA molecule is also labelled with a sequence called unique molecular identifier (UMI). This means that when the gene counts are determined, all reads with the same UMI can be counted as one, since they originate from the same cDNA and were copied during polymerase chain reaction [5]. However, two distinct transcripts result in two distinct cDNA molecules with therefore different UMIs, which means that two reads with different UMIs are counted separately. Like this, the UMI-counts provide a quantification of gene expression in the sample cells. In praxis, some mistakes are introduced during amplification and there are also sequencing errors. This is why not only reads with exact same UMI-sequences can be collapsed to one count, one can also align the UMI-sequences and collapse very similar ones [5].

In contrast to other RNA-Seq protocols, it does not make sense to correct UMI counts for gene length because the sequenced molecules are not fragmented. Therefore, the effect that long genes get more counts than short genes as they lead to longer transcripts and more fragments does not play a role in prime-seq. On the other hand, it might be interesting to compute Reads Per Million (RPM) values when comparing different samples. Here, one normalizes for the fact that a gene gets more counts when there are more counts overall, so the count values of the samples are divided by a per million scaling factor, to bring them on the same scale. This factor is the sum of all gene-counts in a sample divided by one million.

We know that inflammation plays a key role in the development of atherosclerosis. We also know that lipopolysaccharides (LPS) can increase the morbidity and mortality of atherosclerosis-associated cardiovascular disease. LPS, the major outer surface membrane component of Gram-negative bacteria, mimics a systemic infection quite effectively. We can make use of this and administer this endotoxin to ApoE-deficient mice. The LPS increase the atherosclerotic lesion size and the titer of plasma antibodies directed against oxidized low-density lipoprotein, thus aggravate the disease in those mice. Therefore, inhibition of LPS-induced inflammation could be of clinical value as it might represent a useful treatment for atherosclerosis by manipulating immunological effectors.

2. Materials and Methods

Deletion of let7b

Two atherosclerotic mice were used to investigate the role of the micro-RNA *let7b*. In one of them, the *let7b*-gene was knocked out. Then, bone-marrow derived macrophages (BMDM) were obtained and activated in vitro leading to differentiation into either the subtype M1 or M2. This resulted in four sample groups: macrophage subtype 1 or 2 from the wildtype (WT) or knock-out (KO) mouse. Each group consists of six replicates, which makes an overall number of 24 samples.

For all samples, the dataset contains UMI-counts for 33,260 genes. We computed the mean of UMI-counts for every gene and considered only those genes expressed, that have a mean of at least 1. The reason for this is that one UMI-count means that one transcript of the gene was measured but in order for our dataset to be interesting for other researchers focusing on a specific gene, we argue that the gene should have a certain expression level which we chose to characterize by demanding a detected expression in on average every sample. 14,528 genes satisfy this condition. A gene that is not among the expressed genes is *Xist*. When looking at the UMI-counts, it becomes clear that there is no difference between the samples of the two different mice leading to the conclusion that both mice are male. The reason for this is that *Xist* expression indicates female mice as in them, one of the two X-chromosomes is silenced to achieve equal dosages between males and females. *Xist* is an involved in X-chromosome silencing.

Furthermore, it might be interesting to know the highest expressed genes in our dataset. In order to assess this, we computed the top 5% of the genes with the highest UMI-counts for each sample and then united the resulting sets to get one gene set for the entire dataset. The resulting gene set contains 2,595 genes. Since 5% of 33,260 genes correspond to 1,663 genes, we already see that the gene sets of the 24 samples must have big overlaps.

Supplementary figure 1 [Supplem 1] provides an overview of the different biotypes of the genes defined in the annotation file Mus_musculus.GRCm38.75.gtf. We see the fraction of RPM-counts that belong to a gene of the 7 most frequent biotypes. The distribution of all samples looks very similar and the majority of counts belongs to a protein coding gene.

Deletion of miR-147

This dataset consists of 12 samples where the transcriptome of macrophages derived from atherosclerotic lesions was measured with the prime-seq protocol. All of the samples were taken from ApoE -/- mice that were fed a high-cholesterol diet. There are two groups of six replicates each, one group containing WT samples, the other KO samples derived from mice with a knock-out of *miR-147*.

For each sample, the expression of 34,187 genes is given in UMI-counts. However, the first two genes are called *EGFP1* and *EGFP2*, these are no mouse genes but markers that were used during the experiment and therefore not relevant to our analysis which is why they can be removed. Again, the active (mean of UMI-counts at least 1) and highest expressed genes (in the top 5% of at least one sample) can be assessed. Here, the resulting numbers of genes are 16,793 for the active genes and 3,071 for the highest expressed genes. 5% of 34,185 is about 1,709 which means the size of the set of highest expressed genes is about 1.8-times as big. In this dataset, *Xist* is among the expressed genes leading to the conclusion that the mice used were female.

Again, one can analyze the distribution of the biotypes of genes where the RPM-counts belong to. [Supplem 2] shows the resulting bar plot including the seven most frequent biotypes. Here, one can see that the majority of counts belongs to a protein-coding gene. However, there is bigger variation between the samples than in the *let7b*-dataset [Supplem 1].

LPS-time-course dataset

Experiments for the LPS – time course data set were carried out on ApoE-deficient mice. At timepoint 0h, the used ApoE-deficient mice were healthy, and their macrophages were extracted, stimulated with LPS and monitored for the following timepoints: 1h, 2h and 4h after LPS stimulation. Each timepoint also has information for four replicates. The generated data was used for subsequent RNA-seq analysis.

However, since we were given raw FASTQ and BAM files for the LPS-time-course dataset which originated from 4 different mappers, namely STAR, hisat, contextmap and tophat2, we first had to convert them to a count-file using featureCounts prior to conducting our analysis on this dataset. The reference genome which was used was Mus_musculus.GRCm38.75.gtf from ensemble. We compared all 4 mappers and found that they were very similar which is why we only used the counts from STAR for all downstream analysis [Supplem 3].

[Supplem 4] provides an overview of the different biotypes of the genes defined in the annotation file used.

As soon as we produced the count-file for this LPS-time-course dataset we analyzed differentially expressed genes (DEGs), conducted pathway and enrichment analysis as well as time-course analysis. Additionally, we looked at alternative splicing patterns for a specific gene: Apolipoprotein B48 Receptor (*ApoB48R*). This gene is known to play a pivotal role in the pathogenesis of atherosclerosis [1]. Also, we investigated cytokines, pro -and anti-inflammatory signatures as well as multiple biomarkers specific to atherosclerosis.

3. Results

Deletion of let7b

We transformed the UMI counts of the dataset to RPM-values and performed Principal Component Analysis (PCA) on them. Figure 1 shows the result. We see that the first two principal components only explain 26.6% of the variance of the original data, which is not very much. However, the sample groups cluster very well. The first principal component (PC1) correlates with the difference between macrophage type (M1 or M2), while the second principal component (PC2) correlates with the difference between genotypes (WT or KO). The explained variance of PC1 is more than three times as high as the one of PC2. This hints that the difference between the two macrophage types is bigger than the effects that can be observed after the knock-out of *let7b*. This is why we also separated the dataset according to M1/M2 for the following analyses.

Figure 1 also illustrates that the group with the biggest distances within the cluster is wildtype and M2, suggesting that the replicates in this group were least consistent.

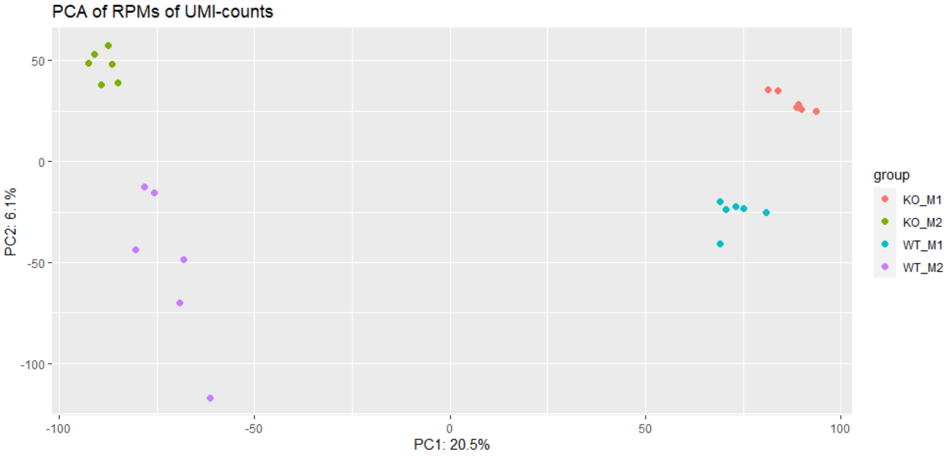


Figure 1: Scatterplot of the first two principal components after PCA of the RPM-values derived from the UMI-counts in the *let7b*-dataset. Groups are defined by genotype (WT or KO of *let7b*) and macrophage type (M1 or M2). Each group consists of six replicates.

Like we have seen above, there seems to be a considerable difference between the macrophage types M1 and M2. In the following, we wanted to examine this difference by using only the WT samples from the *let7b*-deletion dataset. The samples were taken from the bone marrow of the same mouse, only the activation of the macrophages was different. We performed differential expression analysis on the UMI-counts using DESeq2 [Supplem 5]. This resulted in 5,169 significant genes (adjusted p-value ≤ 0.01). 2,569 of them are downregulated in M2-samples (log-2-foldchange < 0), the remaining 2,600 are upregulated M2-samples.

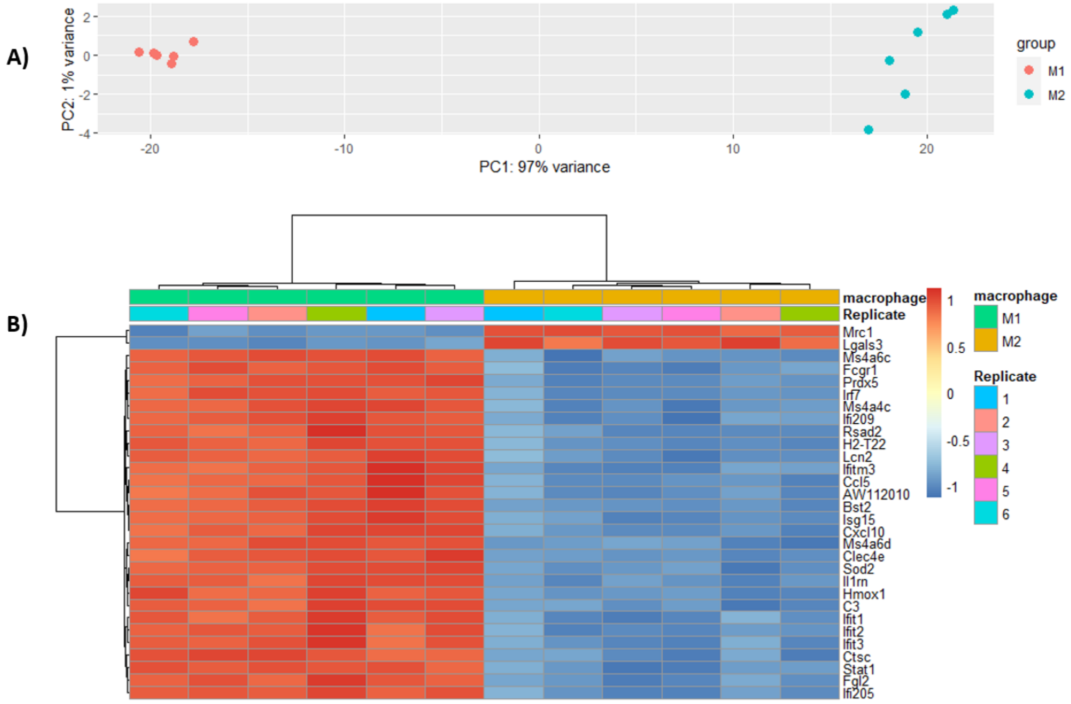


Figure 2: Differences between macrophage types in the wildtype samples from the *let7b*-deletion dataset. A) PCA-plot of the rlog-transformed UMI-counts of the wildtype samples. B) Heatmap of the rlog-transformed UMI-counts of the wildtype samples for the 30 most significant genes according to the DESeq2-analysis.

The wildtype samples cluster very well according to the macrophage type (Figure 2A). The first principal component correlates with the difference between M1 and M2 and it captures 97% of the variance in the data. This gives us reason to assume that most differences between the wildtype samples come from the different macrophage type and samples of the same type are very similar. The rlog-normalized UMI-counts for the most significant genes also show a drastic difference between M1 and M2 samples while there is few variation within the two groups (Figure 2B).

We then performed enrichment analysis to get an idea of the functions of the significant genes. We computed enriched gene sets of the Gene Ontology (for all three namespaces: biological process, cellular component, biological function) and enriched KEGG pathways. The supplementary table [Supplem 6] shows the ten most significant (according to adjusted p-value) gene sets/pathways for every category. For example, the gene set "positive regulation of cytokine production" (from GO: BP) is enriched. 139 of the 200 significant genes in this set are downregulated in M2, the remaining 61 genes are upregulated. Since macrophages produce cytokines as messenger which are involved in inflammatory processed. Like described above, M1 and M2 macrophages have quite different effects which is why it is not surprising that there are a lot of differences in the cytokine production.

In the introduction, we already described that M2 macrophages show an increased production of ornithine, synthesized by the enzyme arginase. The gene *Arg1* codes for an arginase and our analysis shows that it is upregulated in M2, supporting the expectation. Some more interesting genes here are *Igf1* and *Pdgfa*, coding for growth factors which are upregulated in M2. Again, this fits our expectations. *Mrc1* is a marker for M2 macrophages [6] and agreeing with this, it is upregulated in our M2 wildtype samples. Also, some pro-inflammatory proteins like Nfkb and Csf1 seem to be decreased in M2 as the genes *Nfkb2* and *Csf1* are downregulated there.

Overall, we see a considerable difference between the expression patterns of M1 and M2 wildtype macrophages. The analysis supports the expectation that M2 cells are rather growth-promoting and healing while reducing inflammation, in comparison to M1 cells.

Next, we analyzed the effect of *let7b*-deletion in M1 and M2 macrophages. First, we applied DESeq2 to find genes that differ in their expression between the two conditions [Supplem 7, Supplem 8]. Like explained above, we separated the dataset into M1 and M2 samples before. For M1, the analysis resulted in 1,500 significant genes (adjusted p-value ≤ 0.01), for M2 we found 1,152 genes. Table 1 gives an overview of these two significant gene sets. We see that there is an overlap of 339 significant genes, however, not all of them mean a consistent effect on both macrophage types, as 68 of them are upregulated in one type and downregulated in the other type. Therefore, they indicate opposite effects of the *let7b*-deletion in M1 and M2. 271 genes are significant with the same direction of regulation in M1 and M2, which suggests a similar effect of the knock-out. Furthermore, 1,161 genes are exclusively significant in M1 and 813 in M2. They also hint at cell-type specific effects of the *let7b*-deletion.

genes in KO:	**downregulated in M1**	**upregulated in M1**	**not significant in M1 but in M2**	
downregulated in M2	204	52	367	623
upregulated in M2	16	67	446	529
not significant in M2 but in M1	434	727	0	1161
	654	846	813	2313

Table 1: Overview of the sets of significant genes in M1 and M2 and how many genes are up- or downregulated in the knock-out-samples.

Normally, micro-RNAs downregulate their targets which is why we would expect to see upregulation after a knock-out. But the knock-out of micro-RNAs can also lead to a downregulation of genes, for example when a gene targeted by the micro-RNA downregulates the expression of a third gene. Then, the target of the micro-RNA appears upregulated when the micro-RNA is missing and its increased levels lead to more inhibition of the third gene.

We used TargetScanMouse [7] to predict the targets of *let7b*, which gave us a list of 1,076 genes. Then we compared them with the lists of significant genes for M1 and M2. The Venn diagram [Supplem 9] shows that 926 of the predicted target genes are significant in neither the M1 samples nor the M2 samples. A possible reason for that is that these genes are not expressed at a level sufficient for detecting significant differences in M1 and M2. Another interesting thing to mention here, is that there is an overlap of size 107 between the TargetScan prediction and the significant genes in M1. However, 35 of these genes are downregulated in the KO-samples of M1. For M2, the size of the intersection gene set is 69, out of which 42 genes are downregulated in the KO-samples of M2. This is very surprising, as TargetScan tries to predict direct targets of micro-RNAs and as described above, one would expect them to be upregulated in a knock-out sample. Interactions between targeted genes are a possible explanation for this, for example one gene restraining the transcription of the other and therefore balancing or even outdoing the effect of the micro-RNA deletion. All in all, this observation nevertheless means that the prediction of TargetScan cannot be verified very well with our dataset.

Next, we performed enrichment analysis using DAVID (Database for Annotation, Visualization, and Integrated Discovery). The goal here was to find enriched gene sets that represent biological processes/cellular components/molecular functions (from the Gene Ontology) and KEGG pathways. Here, enriched means that a gene set contains a lot of genes from a provided list of genes (in our case lists of significant genes). An enriched gene sets contains more significant genes than what can be expected by chance (probability of seeing this many genes by chance is the p-value). It helps to find out which processes in the cell are affected by the *let7b*-deletion.

Firstly, we used all genes that are significant in M1 and M2 but have different fold-change directions as input (68 genes). This resulted in only 10 significant gene sets (meaning a Benjamini-corrected p-value ≤ 0.01) which are all rather "unspecific" (like nucleus, cytosol, cytoplasm or DNA binding). 5 of the significant sets belong to cellular component (from GO), the rest to molecular function (from GO). Overall, the result does not give us precise hints to the difference between M1 and M2 concerning the effects of *let7b*-deletion.

To get an idea of the effect of the *let7b*-deletion in general, we repeated the enrichment analysis with DAVID for the significant gene lists of M1 and M2 separately. Again, a lot of "unspecific" gene sets like cytoplasm turned out to be significant. However, here we had much larger gene lists (1,500 for M1 and 1,152 for M2) which resulted in a lot more significant gene sets (Benjamini-corrected p-value ≤0.01: 725 sets for M1, 521 sets for M2). The intersect of the sets of significant gene sets contains 308 gene sets. Among them is the KEGG pathway "Fluid shear stress and atherosclerosis" (mmu05418) which involves the relationship of atherosclerosis and blood flow. Here, we have in vitro activated BMDMs and therefore certainly no blood flow, but we can assume that *let7b*-deletion has an impact on this pathway in vivo.

Like we saw above, our analysis yields a lot of significant genes. *Malat1* for example is downregulated in both M1 and M2 cells. Research has shown that reduced levels promote the formation of atherosclerotic lesions [8]. Some other genes (*Cd200r1*, *Gramd1b*, *Cd5l*) are also downregulated and normally have atheroprotective effects (limit inflammation, transport cholesterol out of the cell, protect macrophages against apoptosis through oxLDL). Therefore, the downregulation in KO-samples would indicate worse atherosclerosis. On the other hand, there are genes like *Ccr2* and *Wdfy1* that are both downregulated in M1 and M2. *Wdfy1* promotes inflammation and it has been shown that the knock-out of *Ccr2* decreases lesion area [1]. Furthermore, the genes *Cd36* and *Ctss* are downregulated in M2 cells. *Cd36* codes for a protein that acts as a receptor for oxLDL and brings it into the cell, thereby promoting foam cell formation. *Ctss* codes for a protein that promotes smooth muscle cell migration [9]. These downregulation of these two genes (at least in M2) suggest an atheroprotective effect of *let7b*-deletion.

Deletion of miR-147

Principal Component Analysis can help again to reduce the dimensions of the dataset and get an overview of the samples. Figure 3 shows the scatterplot of PC1 against PC2 after PCA of RPM-counts derived from the UMI-counts. It gets clear that the replicates of the two groups do not cluster which means they might be very inconsistent. Two samples (one from each group) seem to be especially far away from the others, however, for the remaining samples, there is still no visible clustering. The overall variation of the original data explained by PC1 and PC2 together is 31.3%. Again, this is not very much, but in Figure 1, it becomes clear that for the *let7b*-dataset, 26.6% were enough to capture a difference between macrophage types and genotype.

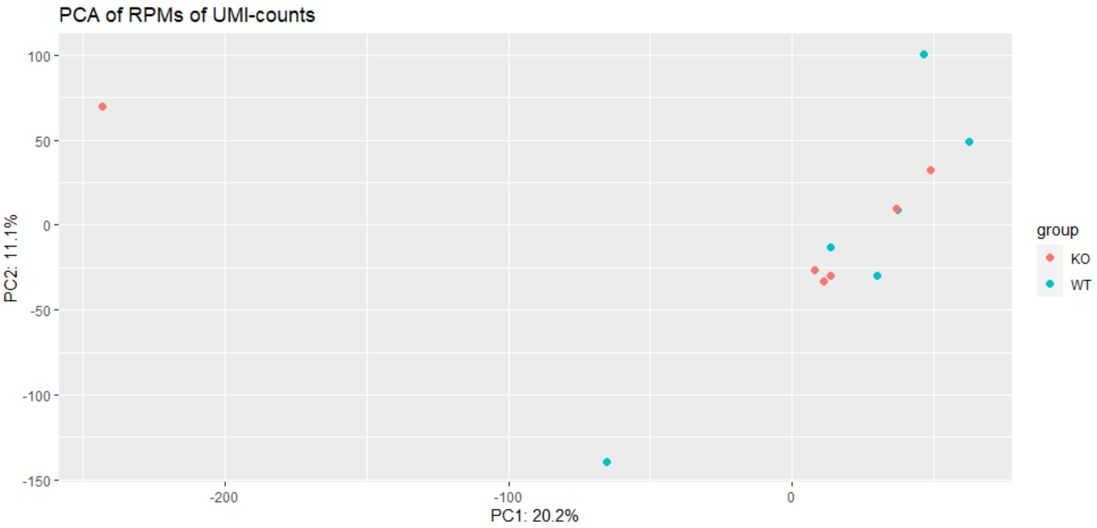


Figure 3: Scatterplot of the first two principal components after PCA of the RPM-values derived from the UMI-counts in the *miR-147* dataset. Groups are defined by genotype (WT or KO of *miR-147*) with six replicates each.

Figure 3 already illustrates that the results of the analyses with the *miR-147*-dataset are rather shaky because the measurements do not seem to capture the two conditions observed here. It is possible that this is due to inconsistencies or other flaws in the experiments, given the assumption that the deletion of *miR-147* has a systematic effect on macrophages.

Next, we tried to find single genes showing a consistent difference between KO and WT in their expression. However, using DESeq2 resulted in no significant genes for a significance level of 0.01 (smallest adjusted p-value: ≈ 0.09). Like described above, micro-RNAs inhibit gene expressions by binding to other RNA molecules. Using TargetScanMouse [7], we obtained a list of 20 possible target genes of *miR-147*. 15 of them were among the genes we considered expressed for the dataset. One would expect them to be upregulated in the KO samples as an inhibitory factor is missing there. However, as Figure 4 shows, the samples do not cluster according to the genotype and the expected difference between WT and KO cannot be observed in the regularized logarithm (rlog) transformed UMI-counts.

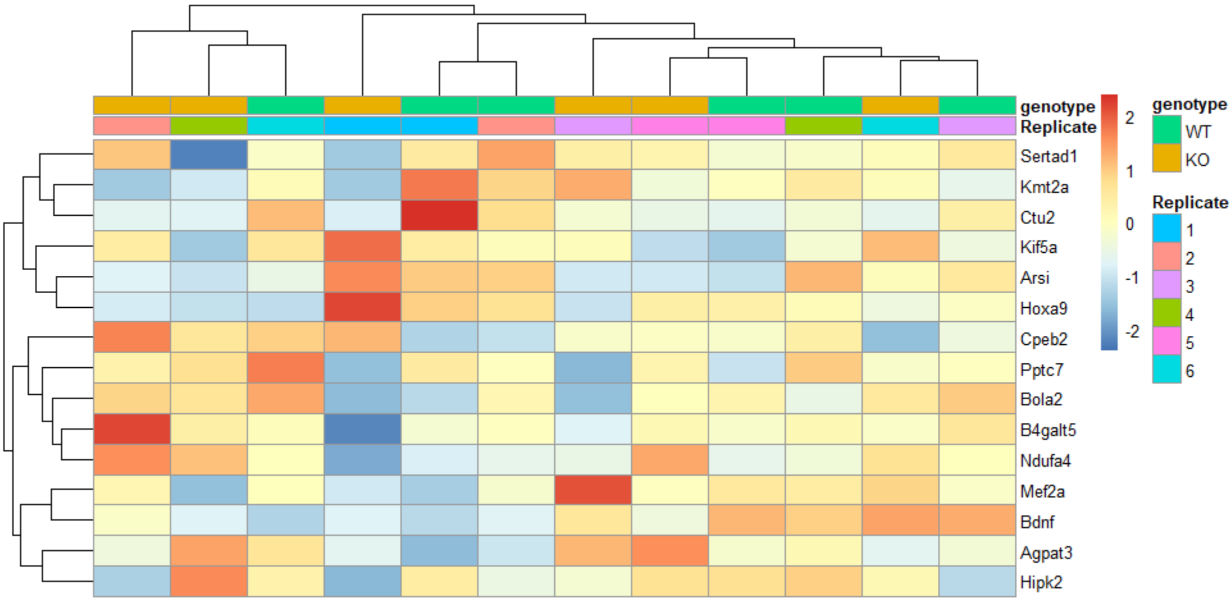


Figure 4: Heatmap of the rlog-transformed UMI-counts of the *miR-147* deletion dataset. Only the 20 genes predicted to be targets of *miR-147* were selected for the heatmap. Data is centered and scaled by row.

We consider TargetScanMouse to be a reliable predictor for micro-RNA target genes, which means it is surprising for us that there is no clustering in the heatmap from Figure 4. Like described above, the reason could be an erroneous dataset. One could suppose that *miR-147* is not expressed in macrophages from atherosclerotic lesions or that maybe its target genes are not or only slightly expressed in this cell type. However, in this case all samples would be replicates of the same experiment and we observed that the replicates are not very consistent. This would again suggest a problem in the data, likely originating from the experiment. On the other hand, we have seen for the deletion of *let7b*, that the target prediction did not work as expected either because a few predicted target genes were downregulated in the knock-out samples.

LPS-time-course

To measure differential gene expression in the LPS-time-course dataset, DESeq2 with the default parameters was used for the comparison of all timepoints. We filtered all genes out where the count value for all samples was zero and detected 14242 active genes in our dataset. From this we found 6742 significantly differentially expressed genes (DEGs) for a threshold set at p-adjusted ≤ 0.01. Thus, these DEGs show a significant difference in expression between our various time points in the LPS-time-course dataset.

In the next step, we wanted to compare the two most popular tools for differential expression analysis: DESeq2 and edgeR. DESeq2 uses a geometric normalization strategy, whereas edgeR is based on the weighted mean of the log-ratios. We used edgeR with the default parameters and genes were labelled as significantly differentially expressed if their p-values were below the threshold of p-adjusted ≤ 0.01. The result is: DESeq2 found 6742 DEGs, whereas edgeR only found 1595 DEGs.

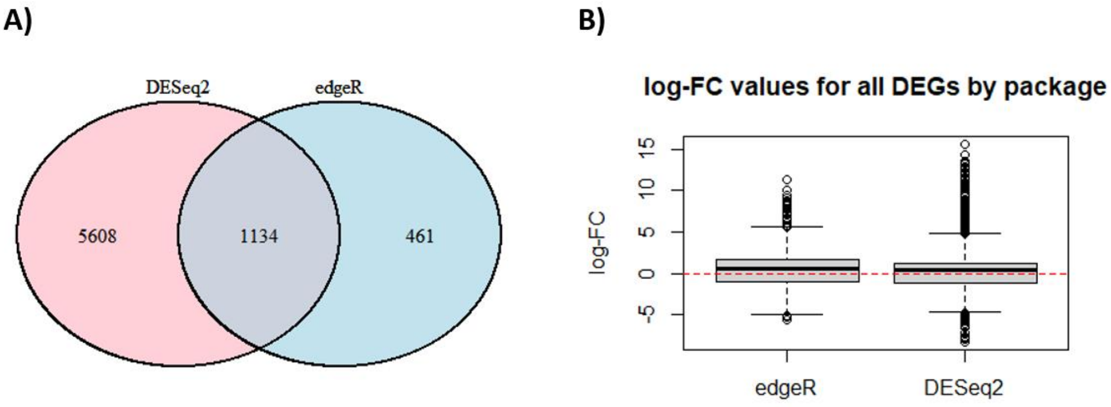


Figure 5: A) Venn diagram showing the number of significantly differentially expressed genes resulting from DESeq2 and edgeR. **B)** Boxplot showing the log-fold change values for all DEGs for edgeR as well as DESeq2.

From Figure 5A we can see that while there is a large overlap between both techniques, DESeq2 finds a total of > 5000 genes that edgeR does not denote as DEGs. Besides the difference in the number of DEGs, there is also a slight difference in the calculated log-fold change values. Figure 5B shows - for all DEGs - the distribution of the log-FC values. It seems that edgeR shows a little bit more of value range then DESeq2. This could possibly be attributed to their different normalization techniques.

To functionally annotate the most significant genes, gene ontology analysis was performed using DAVID. Gene ontology and enrichment analysis was analyzed using clusterProfiler. The p-value cutoff was set at 0.01 and the q-value cutoff was 0.1. Figure 6 shows the resulting GO biological pathways:

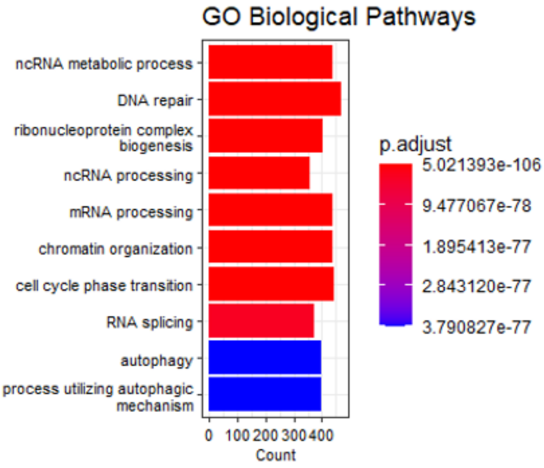
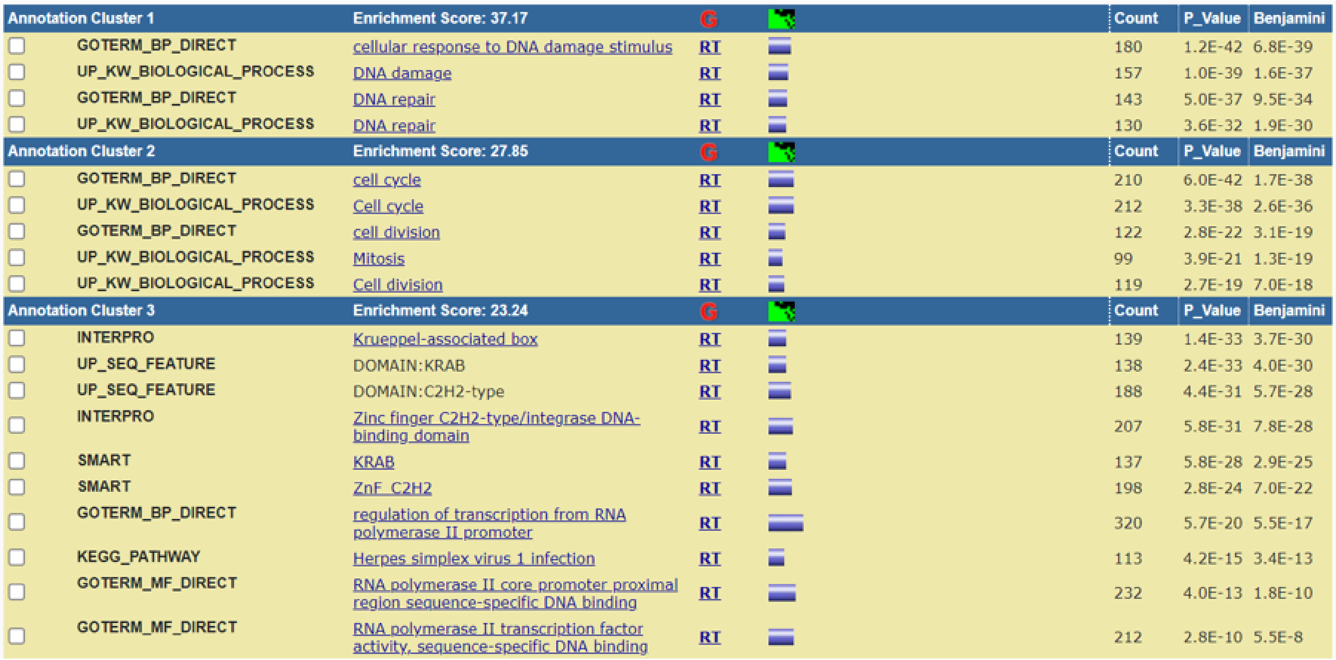


Figure 6: Bar plot showing the top 10 GO biological pathways from clusterProfiler.



					Count	P_Value	Benjamini
Annotation Cluster 1		Enrichment Score: 37.17	G		Count	P_Value	Benjamini
□	GOTERM_BP_DIRECT	cellular response to DNA damage stimulus	RT		180	1.2E-42	6.8E-39
□	UP_KW_BIOLOGICAL_PROCESS	DNA damage	RT		157	1.0E-39	1.6E-37
□	GOTERM_BP_DIRECT	DNA repair	RT		143	5.0E-37	9.5E-34
□	UP_KW_BIOLOGICAL_PROCESS	DNA repair	RT		130	3.6E-32	1.9E-30
Annotation Cluster 2		Enrichment Score: 27.85	G		Count	P_Value	Benjamini
□	GOTERM_BP_DIRECT	cell cycle	RT		210	6.0E-42	1.7E-38
□	UP_KW_BIOLOGICAL_PROCESS	Cell cycle	RT		212	3.3E-38	2.6E-36
□	GOTERM_BP_DIRECT	cell division	RT		122	2.8E-22	3.1E-19
□	UP_KW_BIOLOGICAL_PROCESS	Mitosis	RT		99	3.9E-21	1.3E-19
□	UP_KW_BIOLOGICAL_PROCESS	Cell division	RT		119	2.7E-19	7.0E-18
Annotation Cluster 3		Enrichment Score: 23.24	G		Count	P_Value	Benjamini
□	INTERPRO	Krueppel-associated box	RT		139	1.4E-33	3.7E-30
□	UP_SEQ_FEATURE	DOMAIN:KRAB	RT		138	2.4E-33	4.0E-30
□	UP_SEQ_FEATURE	DOMAIN:C2H2-type	RT		188	4.4E-31	5.7E-28
□	INTERPRO	Zinc finger C2H2-type/integrase DNA-binding domain	RT		207	5.8E-31	7.8E-28
□	SMART	KRAB	RT		137	5.8E-28	2.9E-25
□	SMART	ZnF_C2H2	RT		198	2.8E-24	7.0E-22
□	GOTERM_BP_DIRECT	regulation of transcription from RNA polymerase II promoter	RT		320	5.7E-20	5.5E-17
□	KEGG_PATHWAY	Herpes simplex virus 1 infection	RT		113	4.2E-15	3.4E-13
□	GOTERM_MF_DIRECT	RNA polymerase II core promoter proximal region sequence-specific DNA binding	RT		232	4.0E-13	1.8E-10
□	GOTERM_MF_DIRECT	RNA polymerase II transcription factor activity, sequence-specific DNA binding	RT		212	2.8E-10	5.5E-8

Table 2: Table showing the top 3 annotation clusters from a total of 326 clusters from DAVID.

From Figure 6 and Table 2 we can see some general GOs, for example DNA repair or cell cycle phase transition, to be significantly enriched. Considering that we also have an overall huge number of significantly differentially expressed genes, we can say that after LPS stimulation cells go through a lot of systemic, general changes.

One of the first steps in investigating how LPS stimulation behaves over time could be to look at possible clusters resulting from dimensionality reduction methods like Principal Component Analysis (PCA).

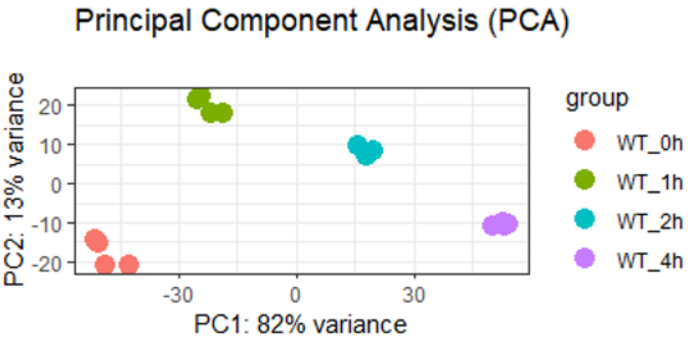


Figure 7: PCA plot showing 4 distinct clusters for the 4 timepoints.

From Figure 7 we can see that PC1 characterizes the time progression since we can see 4 distinct clusters corresponding to each timepoint. Thus, after each timepoint the samples are less like the initial timepoint at 0 hour.

Moreover, we visualized the top 10 significantly differentially expressed genes in our LPS-time-course dataset.

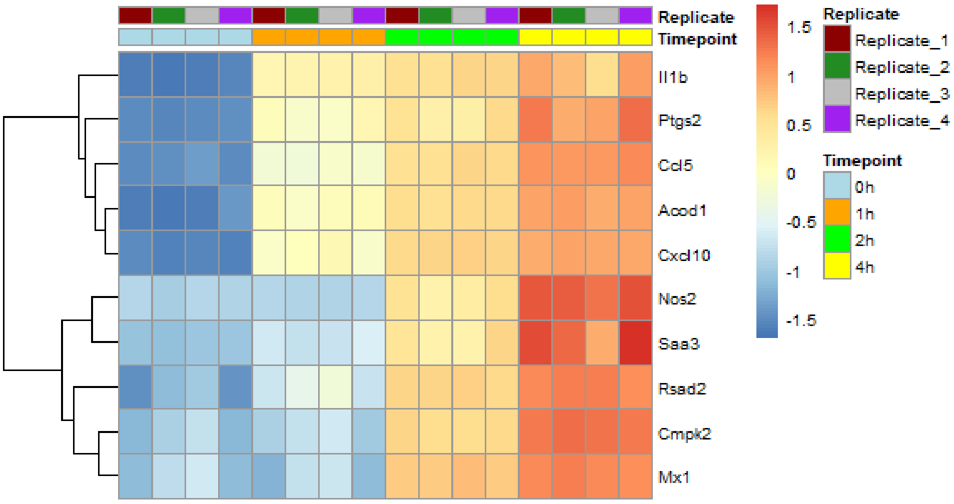


Figure 8: Heatmap showing the top 10 differentially expressed genes for LPS-time-course dataset which are also significant for a p-adjusted value of 0.01. We used the rlog-normalized counts and scaled the genes to z-score.

From Figure 8, it is obvious that the cells are actively responding to the LPS-stress factor. The inflammatory response shows drastic differences even after 1 hour for all top 10 genes which are known to be related to inflammation: Interleukin 1 Beta (*Il1β*), Prostaglandin-Endoperoxide Synthase 2 (*Ptgs2*), C-C Motif Chemokine Ligand 5 (*Ccl5*), Aconitate Decarboxylase 1 (*Acod1*), C-X-C Motif Chemokine Ligand 10 (*Cxcl10*). We can further categorize Serum Amyloid A3 (*Saa3*) and Cytidine/Uridine Monophosphate Kinase 2 (*Cmpk2*) into one group as both genes are known to participate in monocyte differentiation. The last major group is related to antiviral or antimicrobial response. This is not surprising as the samples were stimulated with LPS. The genes are: Nitric Oxide Synthase 2 (*Nos2*), Radical S-Adenosyl Methionine Domain Containing 2 (*Rsad2*), MX Dynamin Like GTPase 1 (*Mx1*). All gene functions were inferred from genecards.org [10].

As atherosclerosis is a complex inflammatory disease, there are many influential biomarkers that contribute to the disease's progression. Most of our biomarkers are inflammatory cytokines. Cytokines are regulators of host responses to infection, immune responses, and inflammation. Some cytokines act to make disease worse (proinflammatory cytokines), whereas others serve to reduce inflammation and promote healing (anti-inflammatory cytokines) [11].

We first looked at chemokines. Chemokines can stimulate the migration of cells, most notably white blood cells (leukocytes). Chemokines like C-C Motif Chemokine Ligand 3 (*Ccl3*), C-C Motif Chemokine Ligand 4 (*Ccl4*), C-C Motif Chemokine Ligand 5 (*Ccl5*) and C-X-C Motif Chemokine Ligand 10 (*Cxcl10*) are involved in all stages of atherosclerosis with various roles [12 - 14].

Next, we looked at following pro-inflammatory signatures:

- Interleukin-1 beta (*Il1β*) signals through the MAPK pathway by binding to Il-1Rl (receptor) and activates transcription factors (TF) and Nuclear Factor of Kappa Light Polypeptide Gene Enhancer In B-Cells Inhibitor (*NF- κb*) which results in pro-inflammatory gene expression. In atherosclerosis, *Il1β* is considered an important contributor in all stages of the disease. Specifically, it increases adhesion molecule expression, vascular permeability, and smooth muscle cell (SMC) proliferation [13, 14].

- Vascular Endothelial Growth Factor A (*Vegfa*) also mediates inflammation, angiogenesis and vascular permeability [13, 14].

- Tumor Necrosis Factor A (*TNFα*) is a pro-inflammatory cytokine and it is mainly exported by monocytes and macrophages. TNF signals through the TNF receptor (TNFR). TNFR activates *NF- κb* which is responsible for cell survival, proliferation, inflammation, and immune regulation. *TNFα* is also involved in production of chemokines and cytokines, expression of adhesion molecules, recruitment of leukocytes, induction of smooth muscle cell proliferation and lipid metabolism [13, 14].

- Interferon gamma (*Ifnγ*) has many pro-atherogenic properties. *Ifnγ* signaling activates T-cells, macrophages, and natural killer (NK)-cells [13, 14].

- Interleukin 12 (*Il12*) is a T-cell growth factor and is often initiated by monocyte or macrophage activation by oxLDL [13, 14].

- Interleukin 18 (*Il18*) induces cytokines and chemokines and significantly contributes to chronic inflammation at the site of lesion development. *Il18* studies have shown that the presence of *Il18* co-localizes with macrophages in atherosclerotic lesions and potentially also causes plaque destabilization. It is absent in healthy arterial regions though [13, 14].

Then, we looked at following anti-inflammatory signatures:

- Interleukin 2 (*Il2*) is a known angiogenic factor and has shown to be increased in coronary artery disease and stable angina patients but not for acute coronary syndrome [12, 14].

- Interleukin 10 (*Il10*) is a lymphokine which suppresses the expression of *Ifnγ*, *TNFα* and T-cell proliferation [12, 14].

- Transforming Growth Factor Beta (*Tgfβ*) inhibits macrophage expression and proinflammatory cytokine synthesis [12, 14].

- Interleukin 4 (*Il4*) promotes Th2 lymphocyte development and inhibits LPS-induced proinflammatory cytokines synthesis [12, 14].

- Interleukin 13 (*Il13*) shares homology with *Il4* and thus also attenuates macrophage functions [12, 14].

- Interleukin 6 (*Il6*) is a glycoprotein produced by macrophages. It stimulates SMC proliferation [12, 14].

Figure 9 shows how the chemokines and the pro -and anti-inflammatory markers behave in our LPS-time-course data:

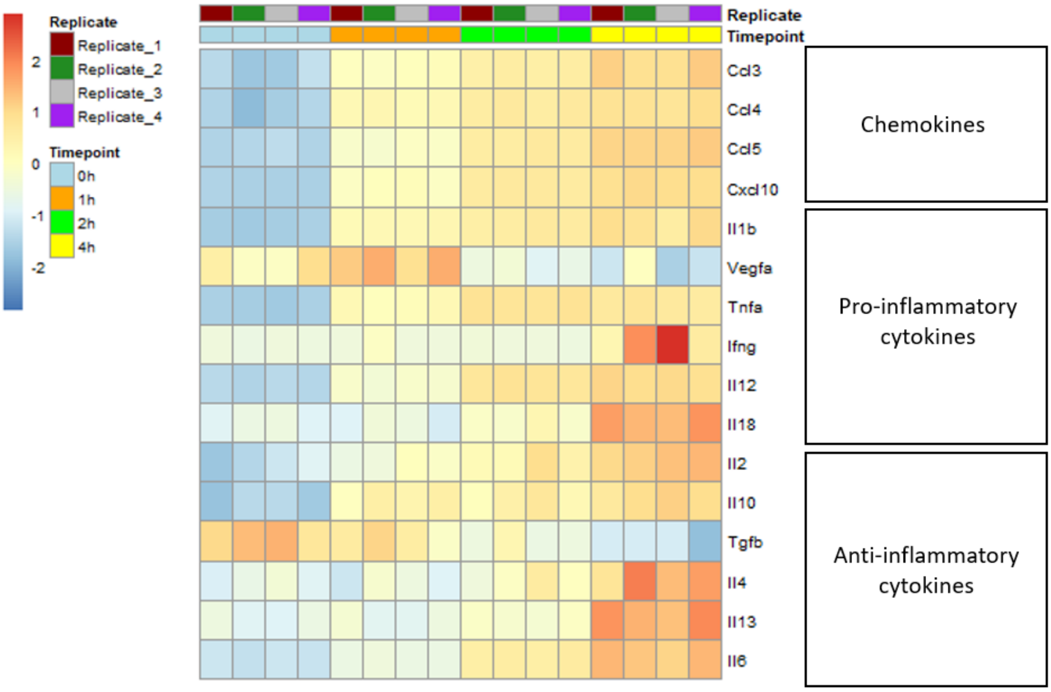


Figure 9: Heatmap showing biomarker genes. Specifically, chemokines, pro -and anti-inflammatory genes for each replicate and timepoint.

From Figure 9 we can see that all the chemokines seem to show a very similar expression pattern since in all cases the expression increases gradually as time progresses. When we look at the pro-inflammatory markers, we generally also see a similar increase in expression with time. However, we also see one exception. The gene *Vegfa* seems to decrease in expression with time. On the other hand, almost all the anti-inflammatory signatures behave like pro-inflammatory cytokines, in the sense that their expression increases with time as well. There is one exception though: *Tgfb*.

These results could make us wonder if LPS is a "perfect" inflammation model after all. We would expect all pro-inflammatory genes to increase with time and all anti-inflammatory genes to decrease with time. Even though, this is sometimes the case, it is not true for all instances.

In the next step, we looked at alternative splicing. It is known that nearly all multi-exon genes are alternatively spliced [15]. This allows a single gene to generate multiple RNA isoforms that give rise to different protein isoforms, which consequently drive phenotypic complexity [15]. There are countless diseases, including atherosclerosis, that are associated with alternative splicing dysregulation [15].

The search for newer therapeutic targets for the prevention of atherosclerosis and lowering of lipid levels has led to new potential targets, some of which aim to harness the power of alternative splicing [15]. In this paper, we addressed the question if alternative splicing of Apolipoprotein B48 Receptor (*ApoB48R*) may play a role in the development of atherosclerosis. *ApoB48R* is a macrophage receptor that binds to the apolipoprotein B48 of dietary triglyceride-rich lipoproteins [15]. However, if overwhelmed with elevated plasma triglyceride, *ApoB48R* may contribute to foam cell formation and atherothrombogenesis [15].

To check if *ApoB48R* shows interesting results for our LPS-time-course dataset, we generated a plot using the DEXSeq package.

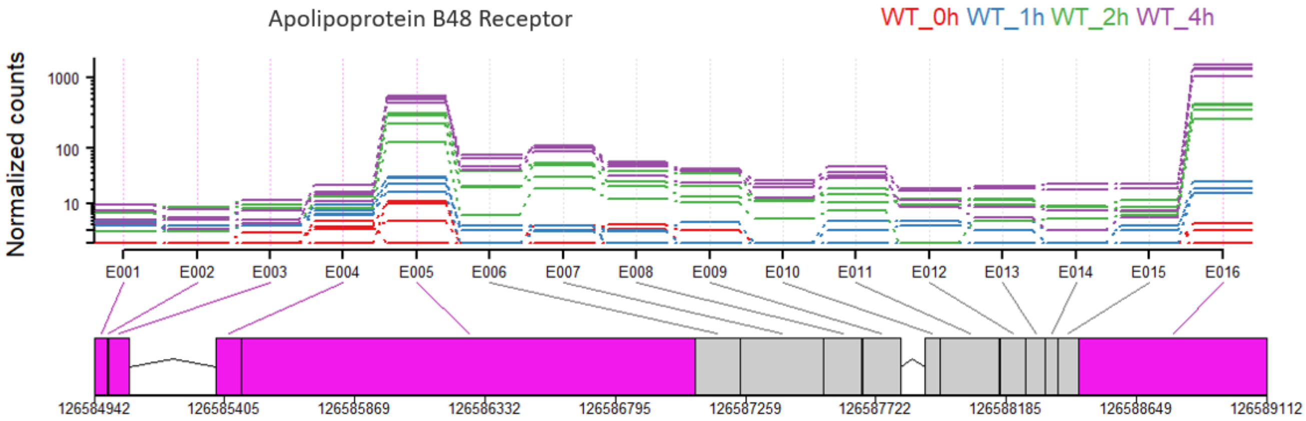


Figure 10: Fitted alternative splicing expression of *ApoB48R* with normalized count values of each exon in each of the samples displaying all 4 timepoints. Shown in purple are those exons that showed significant differential exon usage.

We can tell by looking at Figure 10 that we do see alternative splicing happening, but it does not change as time progresses since for each timepoint the order of the samples remains mostly same. We have 5 exons which show significant differential exon usage. The gene associated with 5 transcripts. We can clearly see that for the LPS-time-course dataset the expression of *ApoB48R* increases dramatically with progression of time.

4. Discussion & Conclusion

miRNA knockouts

The *let7b*-deletion dataset shows a considerable difference between the macrophage types M1 and M2 that is consistent throughout the replicates. However, in this experiment, BMDMs were activated in vitro and some research has shown that there is a difference between in vitro and in vivo activated macrophages [16]. This means that the differences we found here are not necessarily transferable to in vivo macrophages from atherosclerotic lesions [16]. It would be interesting to assess the difference between these macrophages to estimate how reliable our results are for actual atherosclerosis.

Our analysis has shown that the *let7b*-deletion leads to a lot of significant differences compared to the wildtype. We have even seen different effects in M1 and M2 macrophages. Some pathways and genes indicate that *let7b*-deletion has an influence on atherosclerosis in vivo. Some genes seem to have contrary effects, which is why it would be interesting to find out which genes are primarily affected by the knock-out and distinguish them from the significant genes that are only a secondary result (for example regulated by a target of *let7b*). Like stated above, we again would have to make sure that our results translate to in vivo conditions. A possible follow-up experiment for this would be to look at the severity of atherosclerosis in wildtype and *let7b*-knock-out mice.

Like explained above, the *miR-147*-deletion dataset does not allow for very robust conclusions on the effect of the *miR-147*-knock-out as the samples are inconsistent and show no systematic differences between wildtype and knock-out. However, research has shown that the knock-out of *miR-147* leads to an increased inflammatory response in murine macrophages when some of their Toll-like receptors (TLRs) are activated [17]. The authors suggest that the micro-RNA prevents the inflammatory response from becoming too severe [17]. Therefore, it would definitely be interesting to try to use this effect in the treatment of atherosclerosis. Here, the data does not clearly show the results described in the paper, like described above, maybe the reason is an erroneous experiment.

LPS-time-course

In summary, using RNA-seq, we looked at the effects of LPS macrophage stimulation over 4 timepoints. We looked at the alternative splicing patterns of *ApoB48R* and found clear increase in expression with progression of time. One could try to study the interaction and function of these various isoforms, specifically in the atherosclerosis context. It would also be interesting to see this in an experimental setting if we really could reduce blood lipid levels after inducing alternative splicing in form of exon skipping. This way, we would be reducing the levels of the full-length isoform which is why this approach might be of clinical value if the experimental results match the hypothesis.

Furthermore, our study demonstrates that even after short time intervals the reaction to LPS was rather strong as a very large number of transcripts, including immunoregulatory (cytokines, chemokines, etc.) genes, were altered significantly. A good approach in investigating biomarkers in a progressive disease such as atherosclerosis that may take decades to show symptoms would be to identify separate markers associated with the different stages of the disease. However, we also had to question if LPS is a perfect inflammation model after all as most of the pro -and anti-inflammatory markers we investigated showed expected results, however, this was not the case for all markers we used.

In the future, this study could be extended to include data from other high-dimensional surveys, such as microRNA, ChIP-seq, and proteomics, to provide further insight into the global gene regulation processes that occur in LPS-induced atherosclerosis. This could help our understanding of atherosclerosis and other complex diseases so that finding a cure or effective prevention methodologies will become a likely possibility.

6. External tools

The following tools were used via the web-interface and not in R.

DAVID: Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nature Protoc. 2009;4(1):44-57.

CIBERSORTx: Newman, A.M., Steen, C.B., Liu, C.L. et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat Biotechnol 37, 773–782 (2019). https://doi.org/10.1038/s41587-019-0114-2

7. Supplement

Table 3 provides an overview of the R-scripts.

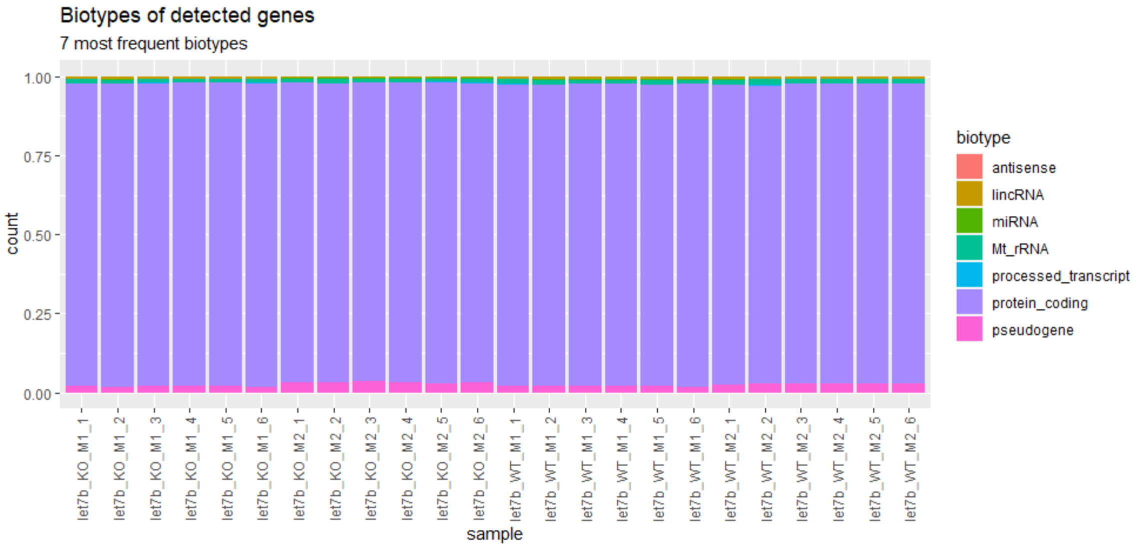
Filename	**Explanation**
ds4_targetScan.R	*miR-147*-deletion: Overview of the TargetScan predicted genes, attempted DESeq2-analysis
ds4_restrictedAnalysis.R	*miR-147*-daletion: DESeq2-analysis with only KO 1, 3, 6 and WT 1, 3, 6
ds3_let7b_deletion.R	*let7b*-deletion: DESeq2-analysis separated by macrophage type (M1, M2), comparison of results and with TargetScan prediction
ds3_macrophageTypes.R	*let7b*-deletion: only WT samples. DESeq2-analysis for difference between M1 and M2. Enrichment analysis with clusterProfiler
expressedGenes.R	Determine active and highest expressed genes for *let7b*- and *miR-147*-deletion datasets
expressedBiotypes.R	Biotype analysis of *let7b*- and *miR-147*-deletion datasets
overviewPCA.R	PCA of UMI-counts for *let7b*- and *miR-147*-deletion datasets
cibersortx_preparation.R	Preparation of reference data for CIBERSORTx

ds5_LPS_analysis.R	LPS-time-course: Run of featureCounts, Mapper comparisons (STAR, hisat, contextmap, tophat2), Deseq2-analysis for STAR data for difference in all timepoints, generated all resulting tables and plotted gene expression data, biotype analysis, basic statistic plots, pathway and pathway impact and enrichment analysis, Deseq2 and edgeR comparison, biomarker analysis and time-course analysis
dexseq_run.R	Alternative Splicing with Dexseq for multiple genes (among them *ApoB48R*)
load_SubreadOutput.R	Helper-script for alternative splicing
recount3.R	Implemented access to multiple data sets: got gene-, exon-, junction- and bp-level count data for a dataset (SRP134109), get region data from a set of data sets using snapcount), computed appropriate fold changes for conditions from the counts for the SRP134109 dataset
snapcount.R	Get a dataset (SRP134109) from recount3, plots for the BigWig-files (for specific genes)
sessionInfo.txt	R-session information for all packages we used and their version

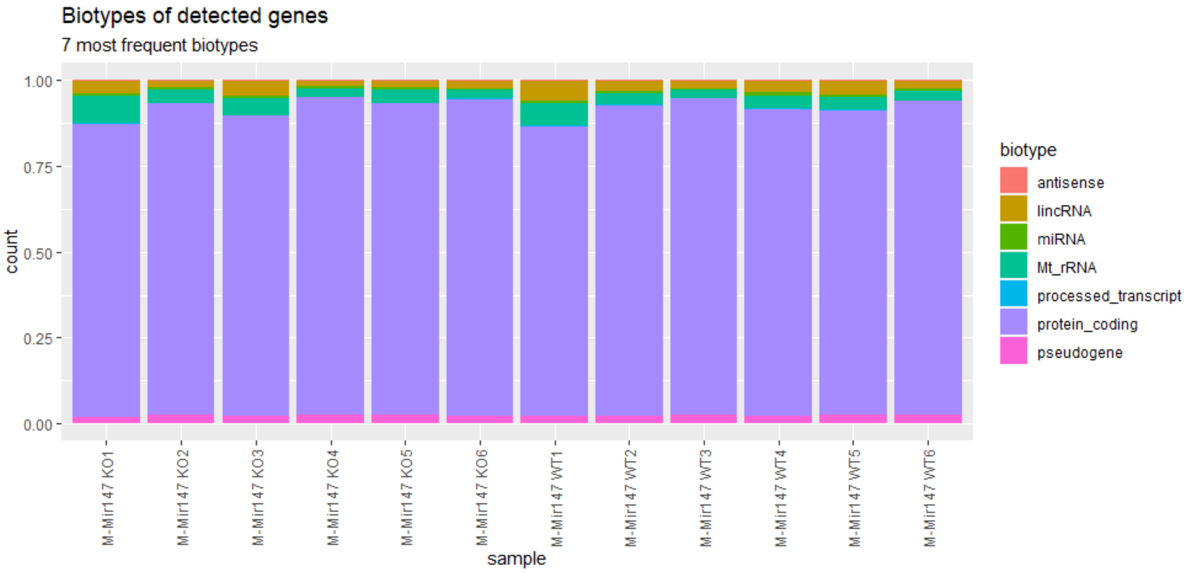
Table 3: Overview of all resulting files and R-scripts used for generating all results and plots.

Full tables for all three datasets containing information on the set of expressed, differentially expressed, alternatively spliced genes, as well as sets of active and differentially enriched genesets and pathways can also be found in the directory supplementary_data on the server.

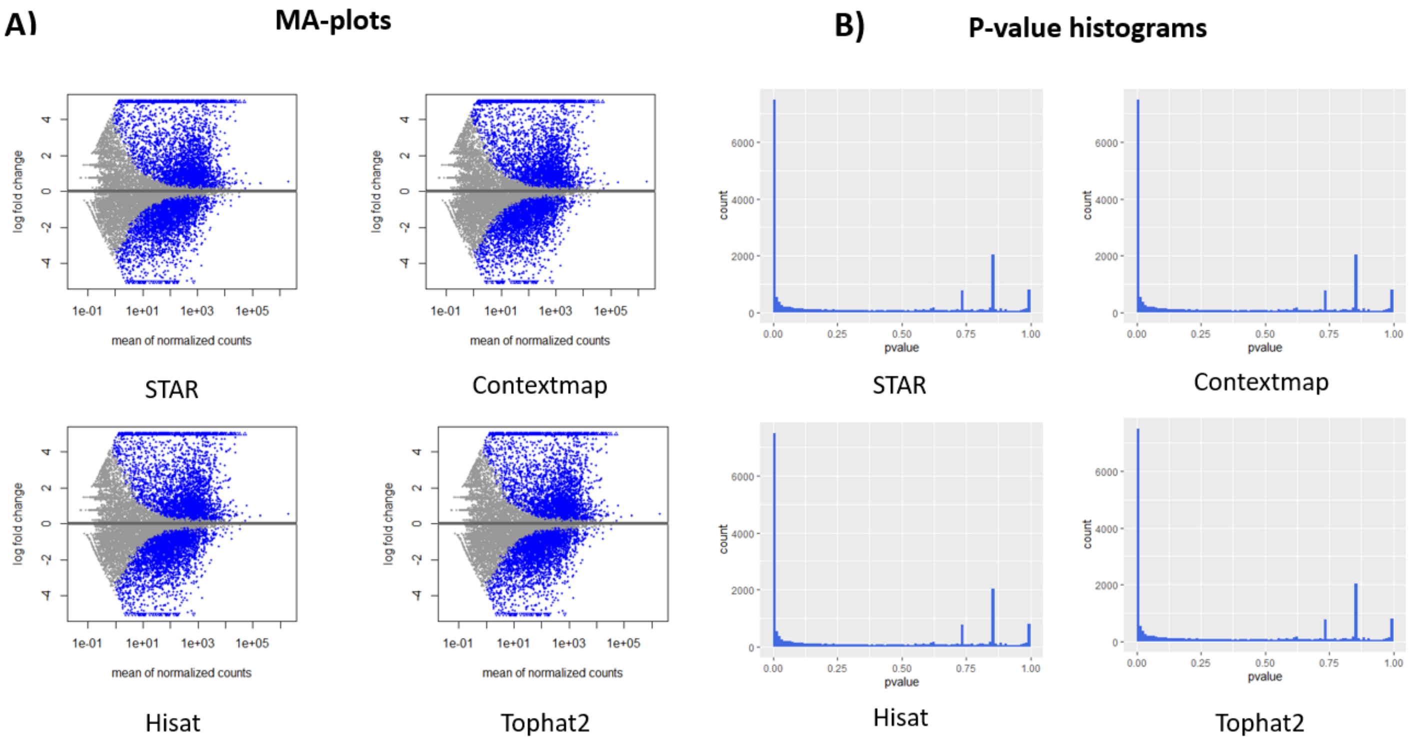
The following plots and figures have been referenced in the main text of this paper.



Supplem 1: Fractions of counts belonging to genes of the 7 most frequent biotypes for each sample. Sample names: let7b_KO are mice with a knock-out of the micro-RNA *let7b*, wild type is let7b_WT. M1/M2 means cell-type (macrophage subtype 1 or 2), the last number denotes the replicate number.

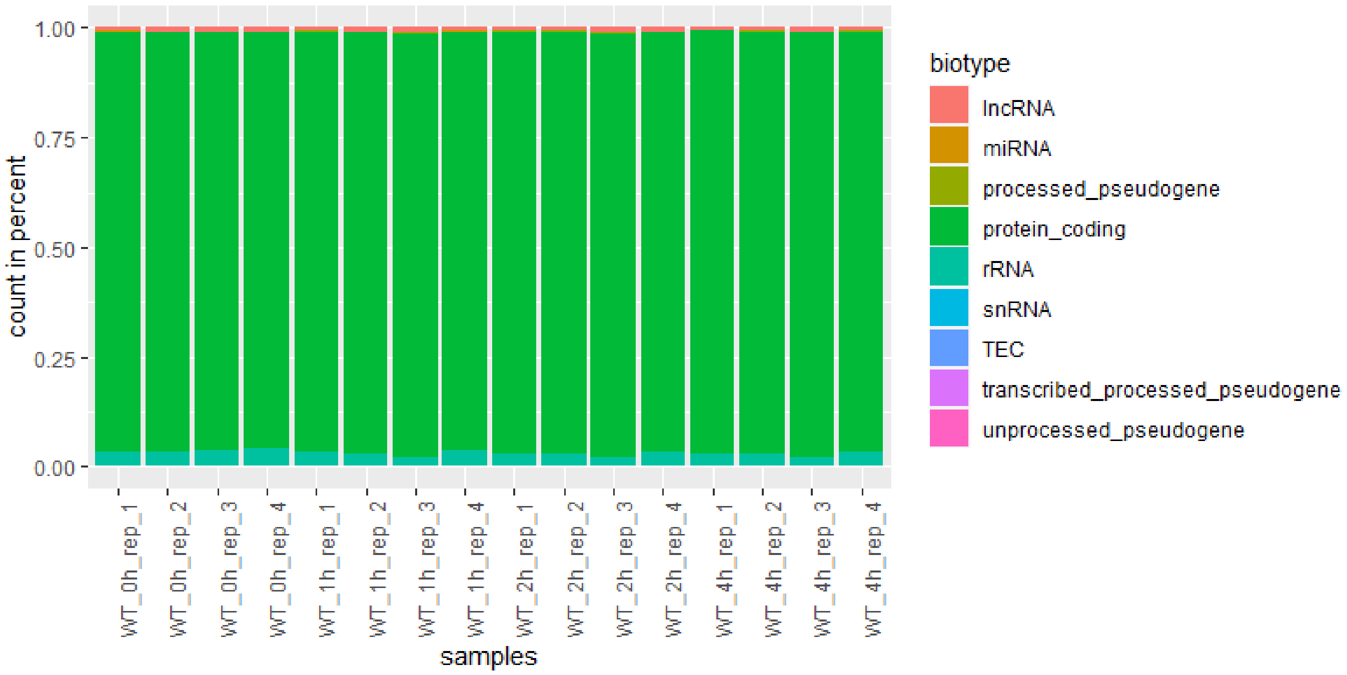


Supplem 2: Fractions of counts belonging to genes of the 7 most frequent biotypes for each sample. Names of the samples: M-Mir147 stands for the micro-RNA *miR-147*, WT/KO means the sample came from wildtype or knockout (of *miR-147*) mice, the numbers in the back denote the replicates.



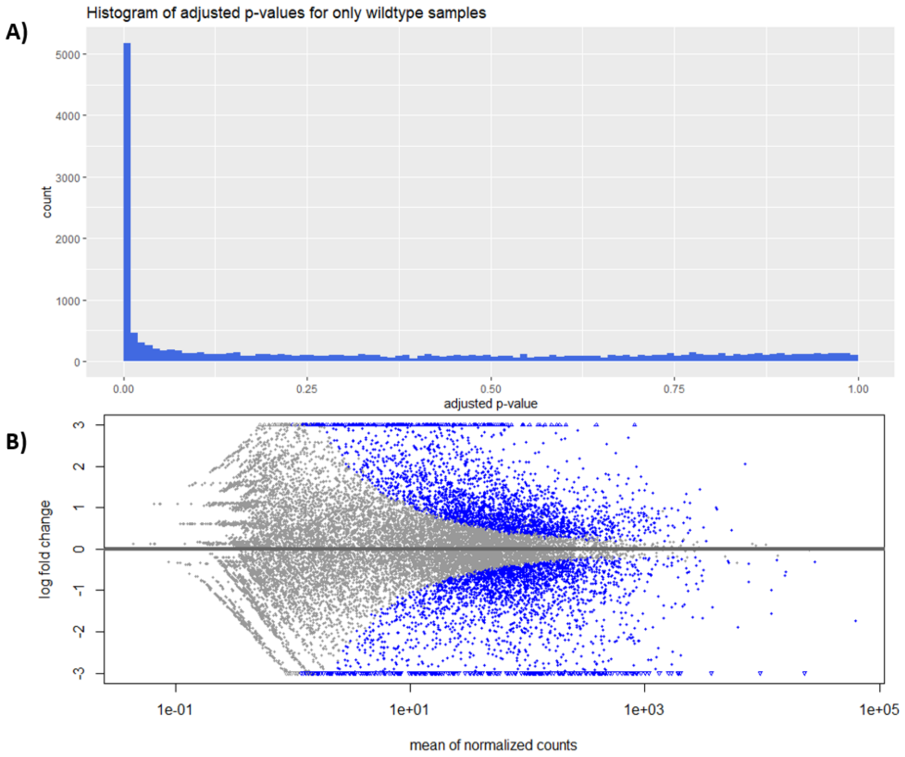
Supplem 3: A) All 4 MA-plots for 4 different mappers: STAR, contextmap, hisat and tophat2. B) All p-value histograms (resulting from Dese2) for all 4 mappers: STAR, contextmap, hisat and tophat2.

All 4 mappers seem to be very robust as all of them deliver highly similar results.



Supplem 4: Stacked bar plot showing the percentage of counts for each detected biotype for each sample in the LPS-time-course dataset.

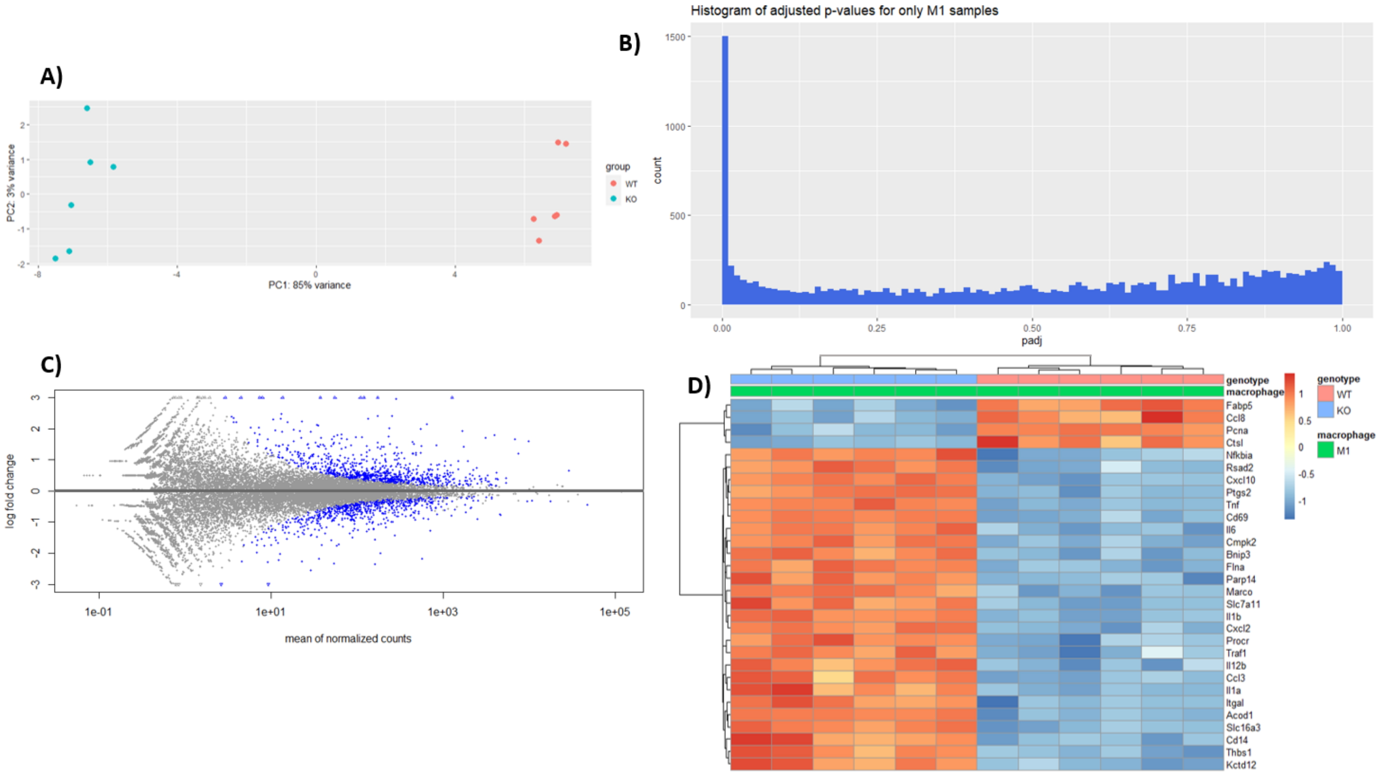
From this we can see that our samples seem very consistent and all of them have less than 10 percent of contamination with biotypes other than protein-coding mRNA.



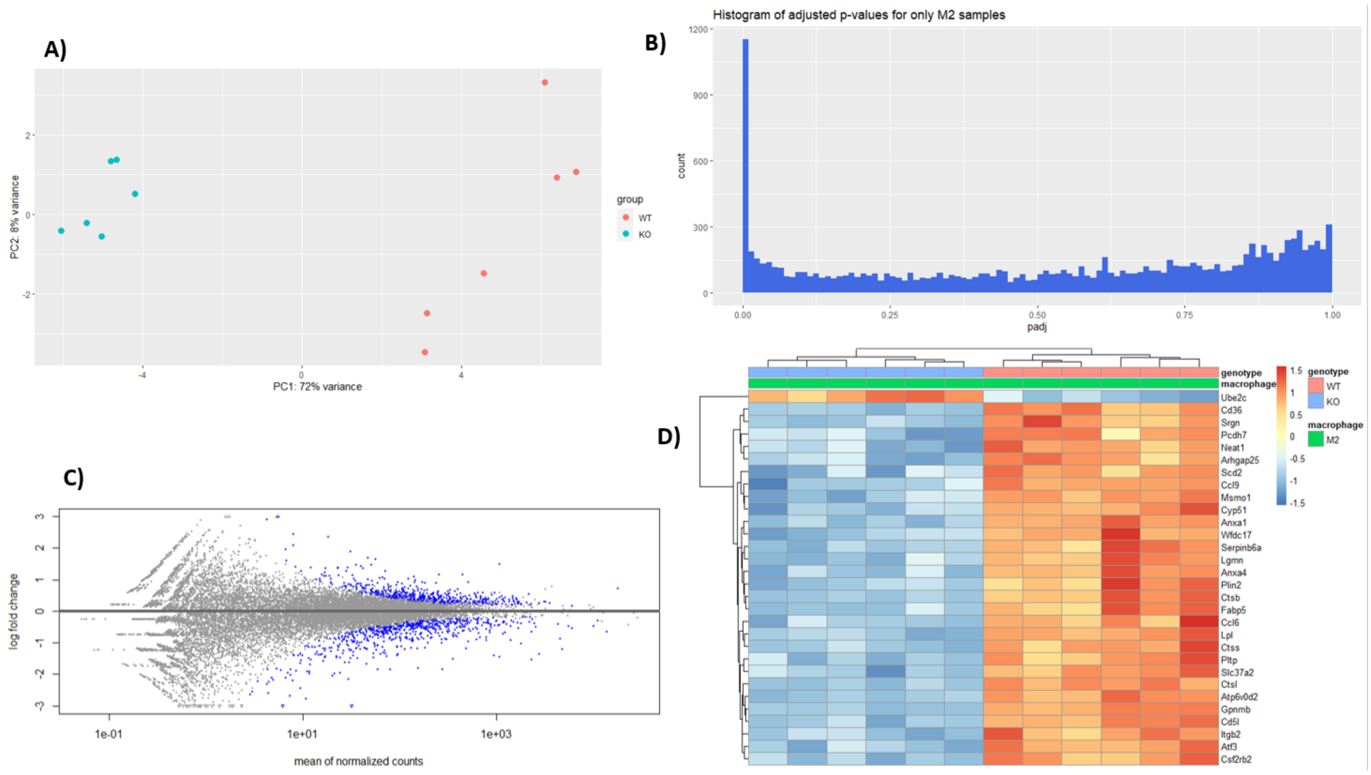
Supplem 5: histogram of adjusted p-values and MA-plot (alpha = 0.01) for the DESeq2 analysis on only the wildtype samples from the *let7b*-deletion dataset.

GO: biological process	GO: cellular component	GO: molecular function	KEGG pathway
ribonucleoprotein complex biogenesis	chromosomal region	structural constituent of ribosome	Coronavirus disease - COVID-19
positive regulation of cytokine production	chromosome, centromeric region	ubiquitin-like protein ligase binding	Cell cycle
symbiotic process	kinetochore	enzyme activator activity	Epstein-Barr virus infection
ribosome biogenesis	lytic vacuole	ubiquitin protein ligase binding	Parkinson disease
DNA replication	lysosome	double-stranded RNA binding	Proteasome
chromosome segregation	spindle	single-stranded DNA binding	Spliceosome
response to virus	nuclear envelope	ATPase activity	Amyotrophic lateral sclerosis
DNA repair	ribosomal subunit	transcription coregulator activity	Ribosome
viral process	ribosome	rRNA binding	Human T-cell leukemia virus 1 infection
positive regulation of response to external stimulus	cytosolic ribosome	helicase activity	Salmonella infection

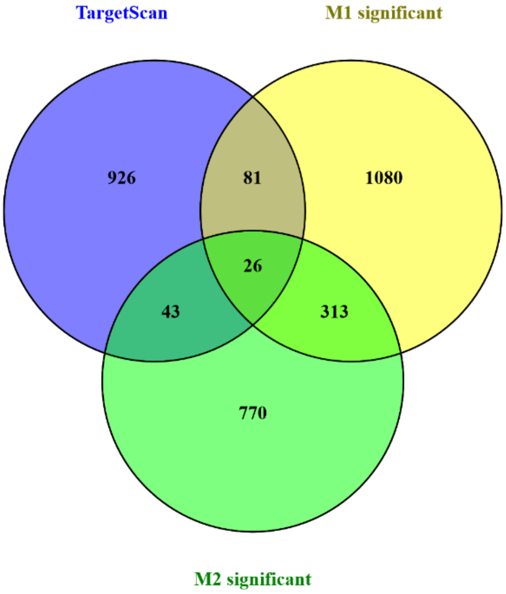
Supplem 6: 10 most significant gene sets/pathways according to adjusted p-value of enrichment analysis for significant genes from DESeq2 (*let7b*-deletion dataset, only wildtype samples).



Supplem 7: overview plots for the DESeq2 analysis of the M1-samples from the *let7b*-deletion dataset. **A)** PCA-plot of rlog-transformed UMI-counts. **B)** Histogram of the adjusted p-values. **C)** MA-plot (alpha = 0.01). **D)** Heatmap of the rlog-transformed UMI-counts of the 30 most significant genes (according to the adjusted p-value).



Supplem 8: overview plots for the DESeq2 analysis of the M2-samples from the *let7b*-deletion dataset. **A)** PCA-plot of rlog-transformed UMI-counts. **B)** Histogram of the adjusted p-values. **C)** MA-plot (alpha = 0.01). **D)** Heatmap of the rlog-transformed UMI-counts of the 30 most significant genes (according to the adjusted p-value).



Supplem 9: Venn diagram of genes predicted by TargetScan to be targets of *let7b* (blue), significant genes in M1 (yellow) and significant genes in M2 (green)

The following plots and figures have not been referenced in the main text of this paper but also contain some interesting information and results.

We looked at how certain pro-inflammatory markers behave in other data sets obtained from the recount3 resource.

Recount3 is a resource which consists of RNA-seq gene, exon, and exon-exon junction counts as well as coverage bigWig files from various studies for human and mouse respectively. Since the revolution of high-throughput expression measurement, massive amounts of data representing a plethora of conditions, diseases, species, and measurement techniques are produced. The advantage recount3 is, however, that all data was processed in a uniform manner using the Monorail pipeline.

Regarding this SRP134109-dataset from recount3, the authors investigated the role of interleukin-1β (*Il1β*) in advanced stages of atherosclerosis. They performed intervention studies on smooth muscle cell (SMC) lineage by tracing mice with advanced atherosclerosis using anti- *Il1β* or immunoglobulin (*IgG*) control antibodies. From this experiment they found that *Il1β* promotes an atheroprotective distribution of smooth muscle cells and macrophages in late-stage murine atherosclerotic lesions.

We picked 5 pro-inflammatory genes from our LPS-time-course dataset and applied them to the SRP134109-dataset. We wanted to check if the *Il1β*-inhibition in this study was effective and reduced overall pro-inflammatory response or not.

Thus, Figure 11 depicts the 5 pro-inflammatory genes we chose to investigate for this dataset:

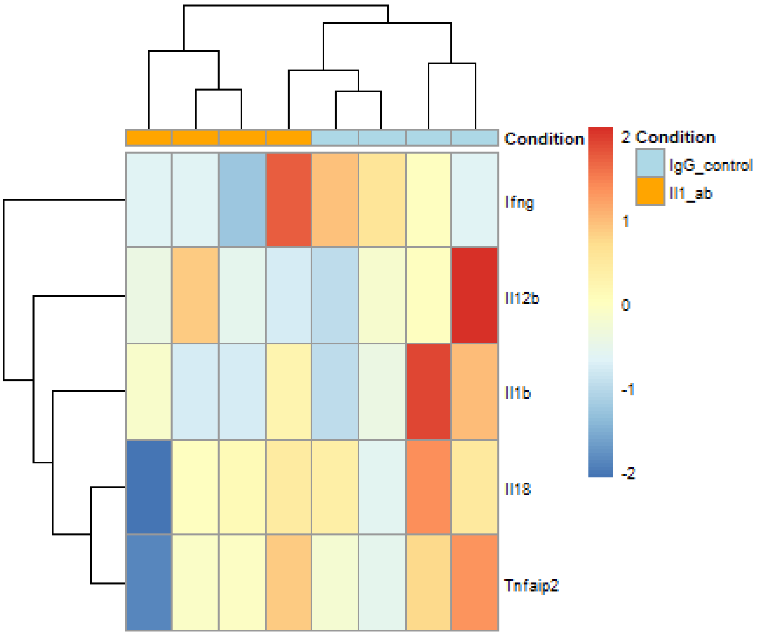


Figure 11: Heatmap showing 5 pro-inflammatory markers for SRP134109 dataset with Il1_ab (*Il1b* - inhibition) and *IgG*-control samples.

From the heatmap in Figure 11 we can see a trend, namely that the *Il1β*-inhibition treatment did reduce overall pro-inflammatory response to a certain degree compared to the *IgG*-control samples. However, we do not see an overall drastic reduction in inflammation.

Thus, we can conclude that even though *Il1β* antibody treatment did show some interesting results which should be further investigated, it cannot be used as a treatment by itself to slow down the inflammatory processes in advanced atherosclerosis.

After that, we had a closer look at the base-pair level count data in the dataset SRP134109 from recount3. The website provides a link to a genome browser where we obtained the genomic regions for the five different biomarkers: *Cd40*, *Nfkb2*, *Ccl5*, *Cxcl10*, *Il1b* [18]. For these regions, we extracted the count data from the bigWig-files of every sample in the dataset and plotted it.

We see that some of the extracted regions (*Ccl5*, *Cd40*, *Il1b*) seem to contain longer untranslated regions upstream and downstream of the actually translated region, as the plot shows zero counts in the front and back. The other two plots (*Cxcl10*, *Nfkb2*) do not show these regions. There is one curve for every sample representing the counts at the respective positions. All curves are oscillating, while peaks mean high count levels. We interpret the peaks as exonic regions. The pattern originates from the structure of the genes. After transcription, some parts (introns) of the RNA molecule are cut out, leaving only the exons. During RNA-sequencing, reads that stem from such a processed mRNA do not map on the genome continuously, as they only cover the exonic regions. In *Nfkb2*, this pattern is especially well visible. Here, we also see that not all peaks have the same height. One reason for this is that during RNA-Seq only relatively short sequences are actually obtained, so a read does not cover a whole gene like *Nfkb2* that is (according to the plot) more than 7,000 base-pairs long. The fragments that are read during sequencing are sampled randomly so the distribution of the reads on a gene is not necessarily even. Another possible effect here is alternative splicing which means that the processing of an RNA molecule can vary. For example, an exon can be part of 50% of the mRNAs, while it is cut out in the remaining mRNAs. This means that we would expect this exon to have half as many counts as an exon that is contained in all mRNAs of the respective gene.

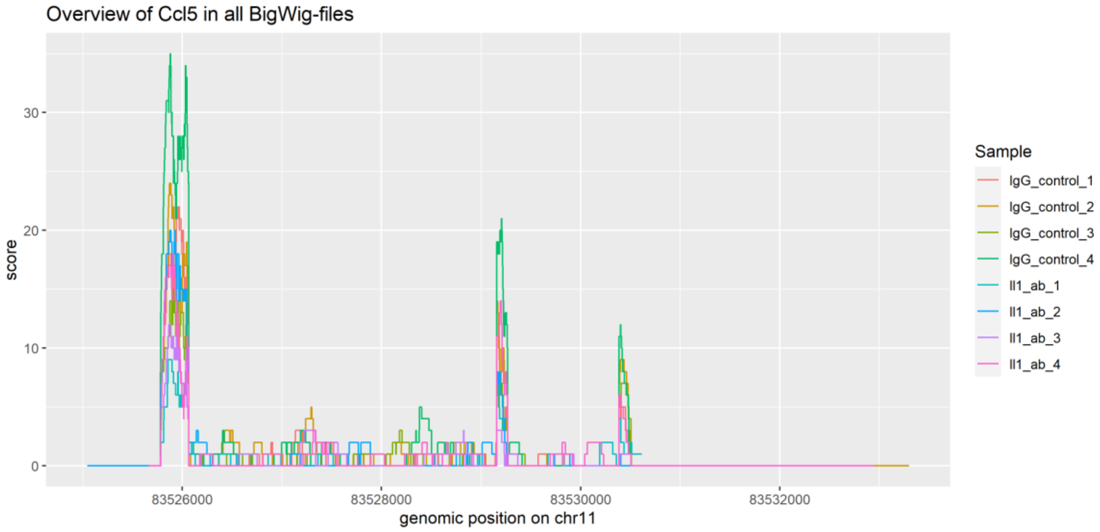


Figure 12: Overview of base-pair level count data for *Ccl5* in the dataset SRP134109 for all samples.

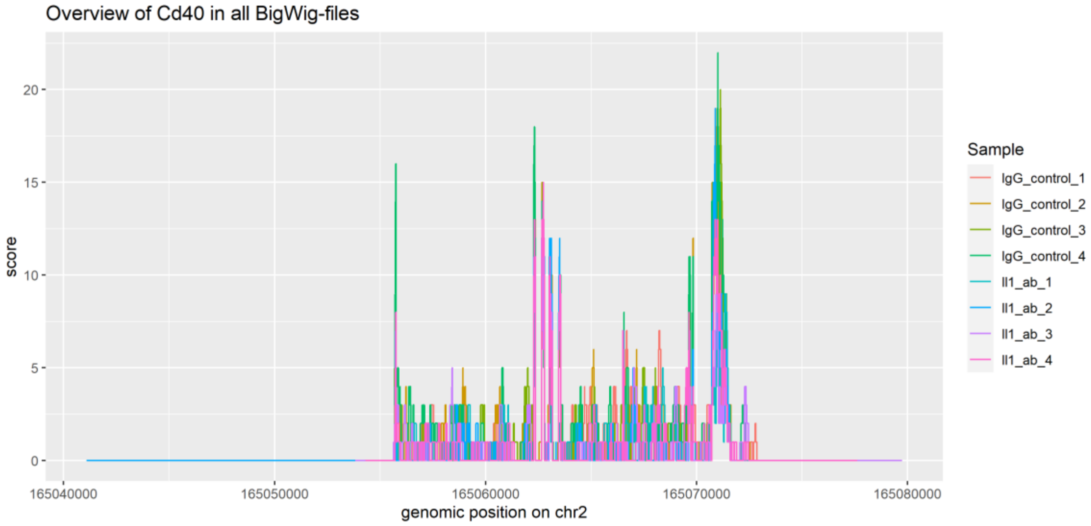


Figure 13: Overview of base-pair level count data for *Cd40* in the dataset SRP134109 for all samples.

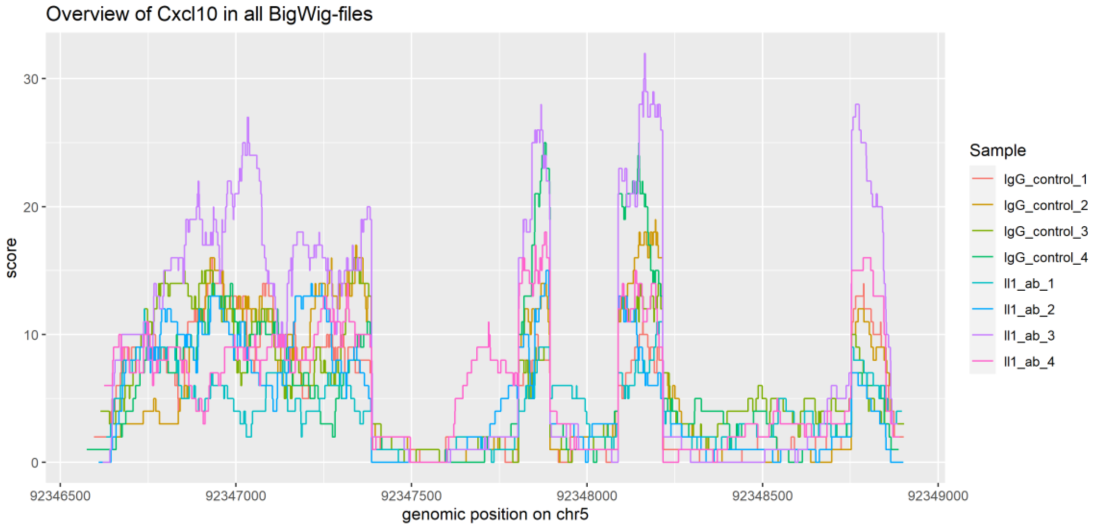


Figure 14: Overview of base-pair level count data for *Cxcl10* in the dataset SRP134109 for all samples.

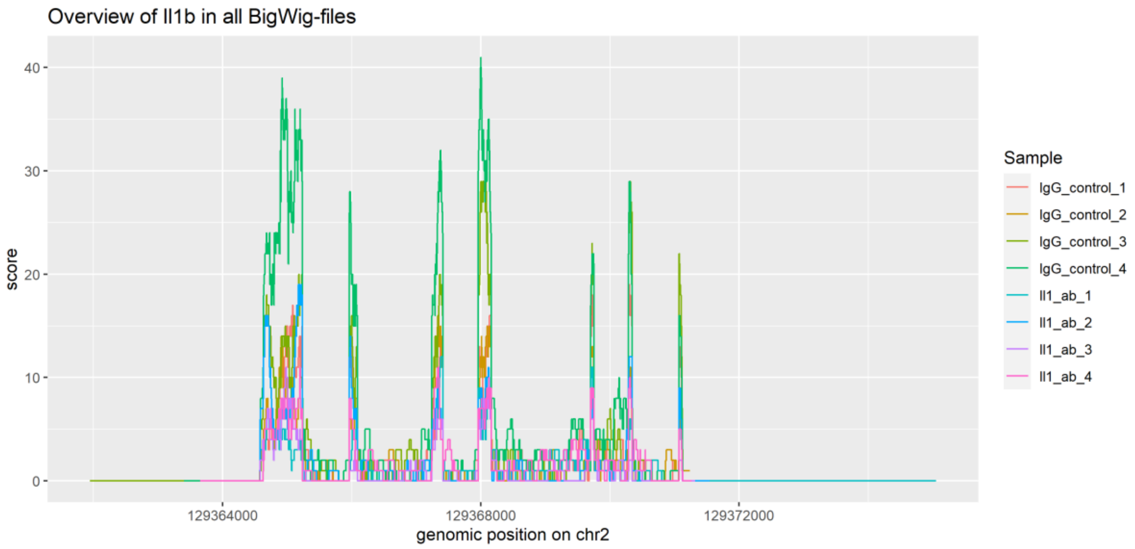


Figure 15: Overview of base-pair level count data for *Il1β* in the dataset SRP134109 for all samples.

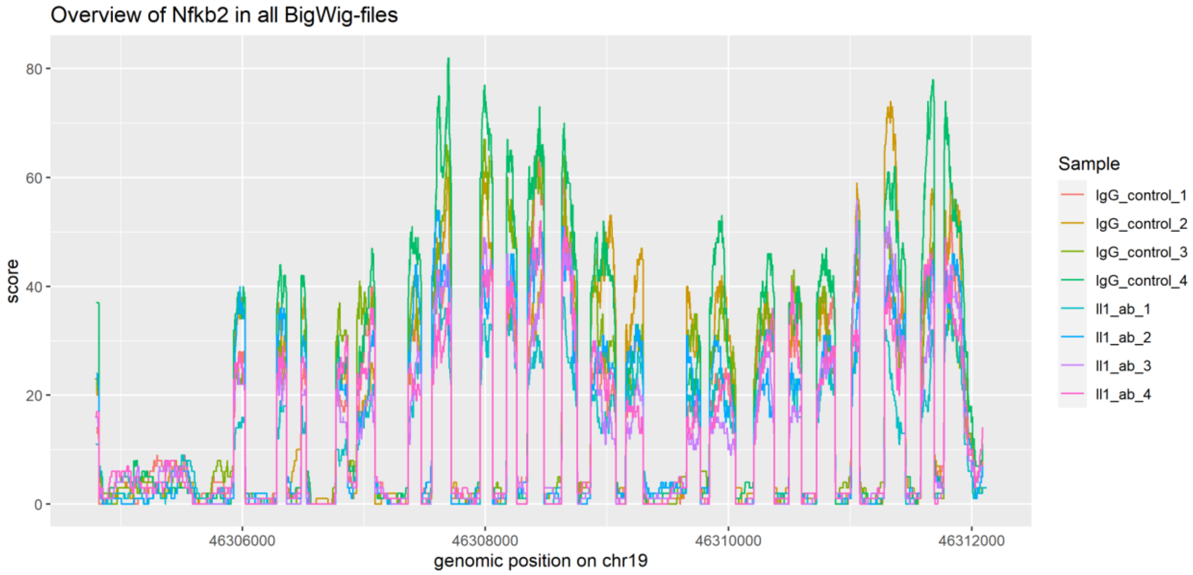


Figure 16: Overview of base-pair level count data for *Nfkb2* in the dataset SRP134109 for all samples.

We tried to perform deconvolution of our bulk-RNA-Seq samples from the *let7b*- and the *miR-147*-deletion datasets, in order to test is we could confirm that the samples only consisted of macrophages. For this, we first downloaded reference data from Tabula Muris Senis. We prepared the reference sample files which mainly consists of making sure that the row- and column-names are in the right format. For example, the columns have to be named according to the cell type of the single cell Seq sample which includes some data wrangling and the metadata. Our reference datasets were Marrow-10X_P7_2 from the droplet data for the *let7b*-deletion dataset, because the cells were from the bone marrow too (Consortium, The Tabula Muris (2017): Single-cell RNA-seq data from microfluidic emulsion. figshare. Dataset. https://doi.org/10.6084/m9.figshare.5715025.v1). For the *miR-147*-deletion dataset, we used the FACS data from the aorta because our samples were taken from atherosclerotic lesions (Consortium, Tabula Muris; Webber, James; Batson, Joshua; Pisco, Angela (2018): Single-cell RNA-seq data from Smart-seq2 sequencing of FACS sorted cells (v2). figshare. Dataset. https://doi.org/10.6084/m9.figshare.5829687.v8). Then, we used CIBERSORTx to generate signature matrices from these reference files. Finally, CIBERSORTx needs a mixture sample and a signature matrix to perform the deconvolution.

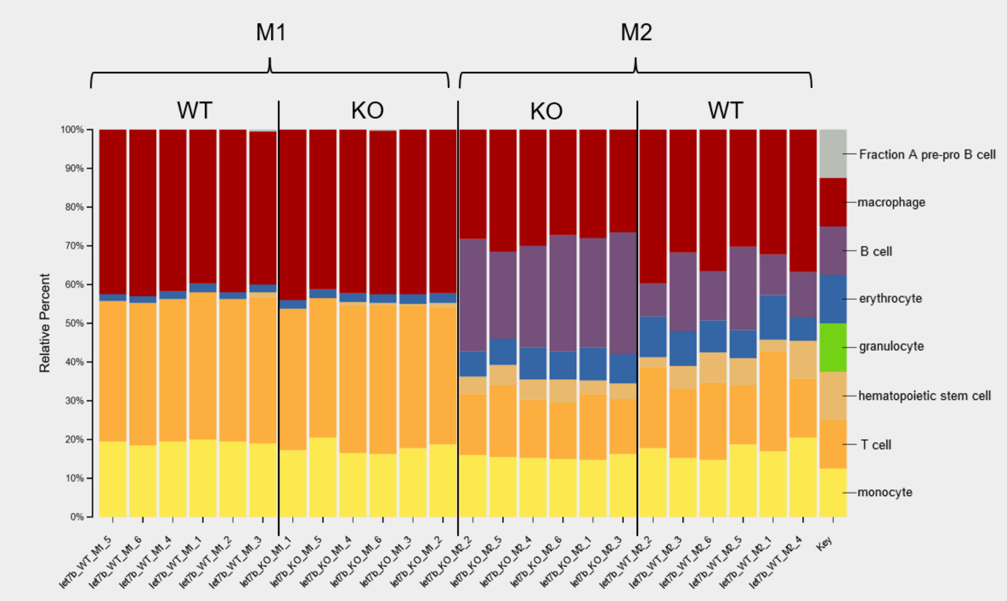


Figure 17: resulting cell type compositions of the samples in the *let7b*-deletion dataset after performing decomposition with CIBERSORTx. As reference data, we used scRNA-Seq data from Tabula Muris Senis (droplet Marrow-10X_P7_3)

Figure 17 shows the result of the cell type decomposition with CIBERSORTx for the *let7b*-deletion dataset. One can see a difference between the M1 and M2 samples, which is interesting, as we did not expect a systematic difference in the cell type composition. However, we do not know how (and how well) the experimentalists made sure that there were only macrophages in the sample. For that reason, and because decomposition is quite a hard problem that relies heavily on the used reference data, we think that the result of our decomposition is not very reliable. Possibly, the visible difference between the M1 and M2 samples is not due to different sample composition but arises from the differences between the cell types and is "misinterpreted" by CIBERSORTx.

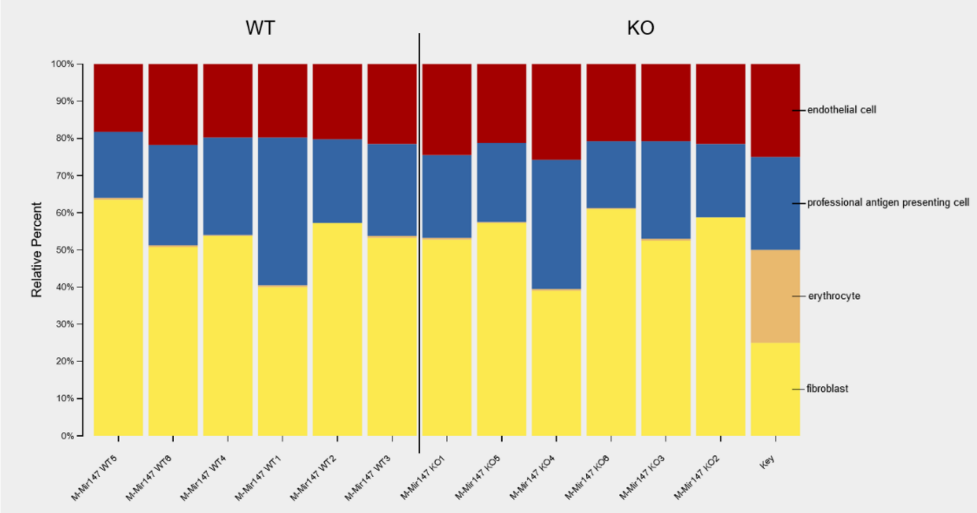


Figure 18: resulting cell type compositions of the samples in the *miR-147*-deletion dataset after performing decomposition with CIBERSORTx. As reference data, we used scRNA-Seq data from Tabula Muris Senis (FACS, aorta-counts)

We also performed decomposition analysis with CIBERSORTx on the *miR-147*-deletion dataset. This time, we used reference data from the aorta, however, it does not include macrophages as extra cell type. The resulting decomposition can be seen in figure 18. One could expect macrophages to be among the professional antigen presenting cells (in the Cell Ontology, they are a subclass), however, not all macrophages present antigens and there are also other antigen presenting cell types. This is why we believe the result from this decomposition is even less reliable than the one from figure 17. Again, we do not know how well the cell types were separated during the experiment. Macrophages from atherosclerotic lesions were microdissected with a laser.

Analysis of miR-147-dataset restricted to only some samples

We have seen that the replicates of the *miR-147*-deletion dataset are inconsistent, which is why there is no gene that shows a significant difference between the conditions WT and KO. Sometimes, single replicates are flawed because of the experiment which is why one might still be able to draw some conclusions when using only a subset of the data.

By looking at the TargetScan predictions for *miR-147* and plotting the normalized counts for the target genes, we determined 3 replicates for each group that should be removed to see a significant difference between WT and KO. Therefore, we repeated the DESeq2-analysis without the samples KO1, KO3, KO6, WT1, WT3 and WT6. This time, the analysis resulted in five significant genes (adjusted p-value ≤ 0.01): *Ndufa4* (which is also among the predicted targets), *Fabp4*, *Malat1*, *Col6a1* and *Cald1*.

In figure 19 D, we see that the significant genes lead to a correct clustering of the WT and KO samples. On the other hand, figures A and B show that the samples do still not cluster together in the PCA and that most adjusted p-values are very high. This shows that the analysis does not find systematic differences between WT and KO.

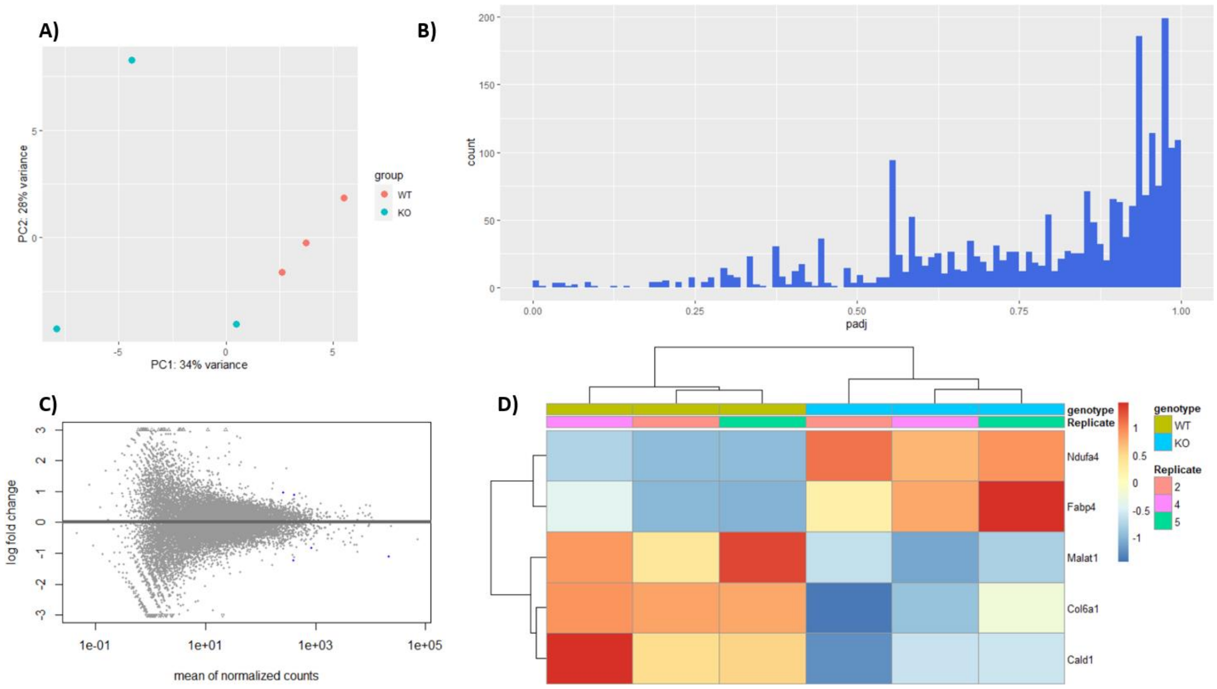


Figure 19: Results of the DESeq2-analysis of the samples KO 2, 4, 5 and WT 2, 4, 5 from the *miR-147*-deletion dataset. **A)** PCA of the rlog-normalized UMI-counts. **B)** Histogram of the adjusted p-values. **C)** MA-plot with alpha = 0.01. **D)** Heatmap of the rlog-normalized counts (centered and scaled by row) for the five significant genes.

In the end, leaving some samples out of the analysis led to more significant genes, because they were consistent for this subset of samples. However, the PCA-plot (figure 19 A) does not suggest that the chosen samples are consistent and cluster well enough for a robust analysis. Furthermore, choosing only the parts of a dataset that lead to significant results is highly questionable, especially when there is no good reason to do so (for example the experimentalists say there was a problem with a sample). This is why we consider the result of the analysis above to be not very robust.

Nevertheless, it might be interesting to investigate the role of the significant genes in atherosclerosis and their potential as possible therapeutic targets.

References

[1] Hopkins PN: Molecular Biology of Atherosclerosis. American Physiological Society, Physiological Reviews. Published 2013 July 01. doi:10.1152/physrev.00004.2012

[2] Mills CD, Ley K: M1 and M2 Macrophages: The Chicken and the Egg of Immunity. Innate Immun 2014;6:716-726. doi: 10.1159/000364945

[3] Lo Sasso, Giuseppe et al. “The Apoe(-/-) mouse model: a suitable model to study cardiovascular and respiratory diseases in the context of cigarette smoke exposure and harm reduction.” Journal of translational medicine vol. 14,1 146. 20 May. 2016, doi:10.1186/s12967-016-0901-1

[4] Janjic A, Wange LE, Bagnoli JW, Geuder J, Nguyen P, Richter D, Vieth B, Vick B, Jeremias I, Ziegenhain C, Hellmann I, Enard W: Prime-seq, efficient and powerful bulk RNA-sequencing. bioRxiv. Published 2021 January 01. doi:10.1101/2021.09.27.459575. URL: http://biorxiv.org/content/early/2021/09/28/2021.09.27.459575.abstract (last access: 14.03.2022)

[5] Parekh S, Ziegenhain C, Vieth B, Enard W, Hellmann I: zUMIs - A fast and flexible pipeline to process RNA sequencing data with UMIs. Gigascience. 2018 Jun 1;7(6):giy059. doi: 10.1093/gigascience/giy059. PMID: 29846586; PMCID: PMC6007394

[6] Wei Y, Corbalán-Campos J, Gurung R, Natarelli L, Zhu M, Exner N, Erhard F, Greulich F, Geißler C, Uhlenhaut NH, Zimmer R, Schober A: Dicer in Macrophages Prevents Atherosclerosis by Promoting Mitochondrial Oxidative Metabolism. Circulation, 138(18), 2007-2020. Published 2018. doi:10.1161/CIRCULATIONAHA.117.031589. URL: https://www.ahajournals.org/doi/abs/10.1161/CIRCULATIONAHA.117.031589 (last access: 14.03.2022)

[7] URL: https://www.targetscan.org/mmu_80/, Release 8.0 from September 2021. Last access: 12.03.2022. Whitehead Institute for Biomedical Research

[8] Cremer S, Michalik KM, Fischer A, Pfisterer L, Jaé N, Winter C, Boon RA, Muhly-Reinholz M, John D, Uchida S, Weber C, Poller W, Günther S, Braun T, Li DY, Maegdefessel L, Perisic Matic L, Hedin U, Soehnlein O, Zeiher A, Dimmeler S. Hematopoietic Deficiency of the Long Noncoding RNA MALAT1 Promotes Atherosclerosis and Plaque Inflammation. Circulation. 2019 Mar 5;139(10):1320-1334. doi: 10.1161/CIRCULATIONAHA.117.029015. Erratum in: Circulation. 2019 Jul 16;140(3):e161. PMID: 30586743

[9] Ni H, Xu S, Chen H, Dai Q. Nicotine Modulates CTSS (Cathepsin S) Synthesis and Secretion Through Regulating the Autophagy-Lysosomal Machinery in Atherosclerosis. Arterioscler Thromb Vasc Biol. 2020 Sep;40(9):2054-2069. doi: 10.1161/ATVBAHA.120.314053. Epub 2020 Jul 9. PMID: 32640907

[10] URL: https://www.genecards.org/, Version 5.8, Last access: 13.03.2022

[11] Cavaillon JM. Pro- versus anti-inflammatory cytokines: myth or reality. Cell Mol Biol (Noisy-le-grand). 2001 Jun;47(4):695-702. PMID: 11502077

[12] URL: https://www.sinobiological.com/resource/cytokines/inflammatory-cytokines, Last access: 13.03.2022

[13] Soeki T, Sata M. Inflammatory Biomarkers and Atherosclerosis. Int Heart J. 2016;57(2):134-9. doi: 10.1536/ihj.15-346. Epub 2016 Mar 11. PMID: 26973275

[14] Brown, Todd M, and Vera Bittner. “Biomarkers of atherosclerosis: clinical applications.” Current cardiology reports vol. 10,6 (2008): 497-504. doi:10.1007/s11886-008-0078-1

[15] Hasimbegovic, Ena et al. “Alternative Splicing in Cardiovascular Disease-A Survey of Recent Findings.” Genes vol. 12,9 1457. 21 Sep. 2021, doi:10.3390/genes12091457

[16] Orecchioni M, Ghosheh Y, Pramod AB, Ley K: Macrophage Polarization: Different Gene Signatures in M1(LPS+) vs. Classically and M2(LPS−) vs. Alternatively Activated Macrophages. Frontiers in Immunology, Vol 10. Published 2019 May 24. doi: 10.3389/fimmu.2019.01084, https://www.frontiersin.org/article/10.3389/fimmu.2019.01084

[17] Liu G, Friggeri A, Yang Y, Park YJ, Tsuruta Y, Abraham E. miR-147, a microRNA that is induced upon Toll-like receptor stimulation, regulates murine macrophage inflammatory responses. Proceedings of the National Academy of Sciences, Published 2009. 106(37):15819-15824. doi:10.1073/pnas.0901216106

[18] URL: https://genome.ucsc.edu/cgi-bin/hgTracks?db=mm10&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtPosition=&position=chr2%3A165055627%2D165072948&hgsid=1300517349_0kxJU6xVWSUCHns358oYje2wV86g, last access: 14.03.2022