# NGS Analysis Report TYTO vs. ARIA CD362 ± / +

**Elaine Cullen[1*], Barry Digby[2+], Dr. Pilib Ó Broin[2], and Dr. Steve Elliman[1,]**

[1]Orbsen Therapeutics, Dangan Heights, Galway Business Park, Galway, H91 EFD0.
[2]Ó Broin Lab, School of Mathematics and Statistics, National University of Ireland, Galway, H91 TK33.
[*]Performed cell preparation, RNA extraction.
[+]Performed data analysis of samples.

## File/Sample renaming

Each bone marrow group has been renamed to facilitate creating model matrices for fitting covariates in the generalized linear model and to reduce clutter on plots. An explanation of the renamed files can be found below, and will be used for the rest of the report.

| | |
|---|---|
| Normal/Plastic Adherent | PA |
| CD362-/CD271+ ARIA | ARIA |
| CD362+/CD271+ ARIA | ARIA-CD362 |
| CD362+/CD271+ TYTO | TYTO |

**Table 1.** Description of sample names

H88, H89, H90 and H91 refer to the 4 donors in the study and cytokine/control samples will be explicitly stated in plots/results.

## Exploratory Data Analysis

Unsupervised clustering was performed on Cytokine and Control treated datasets in order to assess sample heterogeneity (plots generated are accessible at the end of the report). To generate the plots, read counts were normalized using DESeq2 median of ratios method whereby the geometric mean for each gene across all samples is calculated. The counts for a gene in each sample is then divided by this mean, and the median of these ratios in a sample is the size factor for that sample. This method corrects for RNA composition and library size. Normalized counts were transformed for visualization using $log_2(counts+1)$.

### Donor heterogeneity drives variation

To assess sample similarity, the pearson correlation was computed for each sample to assess linear correlation. Clustering was performed using sample correlation metrics to produce sample to sample heatmaps displayed in **Figure 21**. The plot shows clear clustering of each donor group (H88, H89, H90 & H91) in both Control and Cytokine treated samples, due to the magnitude of inherent genetic differences between the donors. Donor heterogeneity can thus be considered a confounding variable in the dataset, and must be included in the generalized linear model used by DESeq2 to separate the effect of donor variation on the explanatory variable. The notation in DESeq2 follows:

$$Design = (\sim Donor Variation + Cell Sorting Method)$$

### TYTO & PA cluster together & are distinct from ARIA/ARIA CD362 clusters

Sample to sample heatmaps offer poor resolution displaying secondary sources of variation, thus to assess clustering of cell sorting methods within donors, the pvclust package in R was employed to generate dendograms in **Figure 22**. Briefly, the package extracts 1000 bootstrap samples by randomly sampling the dataset. Hierarchical clustering is performed on each bootstrap copy and for each cluster the bootstrap probability (BP) and approximately unbiased (AU) probabilities are computed. The resulting bootstrap distribution is used to inform the final dendogram with measures of confidence given by BP. The dendogram in **Figure 22** shows ARIA/ARIA-CD362 sorted cells cluster together in 6/8 samples, implying that the correlation distance between these is negligible and the methods are very similar. TYTO sorted cells cluster together on the same branch as PA cells in 4/8 samples and are otherwise located in close proximity to PA samples on the dendogram tree. This suggests that TYTO is more similar to PA whilst being distinct from ARIA/ARIA-CD362 clusters.

### TYTO & PA separate from ARIA/ARIA-CD362 in latent space

Principal Components Analysis (PCA) was performed to determine the proportion of variance in each sample with respect to selected covariates 'cell sorting method', and 'donor'. Briefly, PCA works by reducing the number of highly correlated variables in the dataset to reduce dimensionality, whilst retaining the variability present within the datset. These uncorrelated variables representing the reduced dataset are referred to as principal components (PC).

Generally speaking, the first two principal components of a dataset explain a large proportion of the variance. By constructing an eigen correlation plot whereby each principal components correlation with 'cell sorting method' or 'donor', one can interrogate sample clustering within principal components using PCA plots. In both **Figure 23** and **Figure 24**, the eigen correlation plot is shown at the bottom of the page.

PC1 and PC2 in cytokine treated samples (**Figure 23**) show a strong correlation with donor (rep(licate) refers to donors), confirming the initial exploratory data analysis that donor variation explains the vast majority of variation in the dataset. This is further displayed in PCA plots (top left, top right) where PC1 vs. PC2 and PC2 vs PC3 show samples clustering according to their donor groups. The eigen correlation plot suggests that PC4 contains the explained variance between cell sorting methods (condition) at significance level of P-value <0.01. This can be seen in the PCA plots (center left, center right). In both plots TYTO and PA are stratified from ARIA/ARIA-CD362 samples, with the exception of one H88-PA sample. Control treated samples (**Figure 24**) follows the same trend as cytokine treated samples, where PC1 and PC2 explain the donor variation present in the dataset. However in the control treated samples PC4 variance has a stronger correlation with cell sorting methods (condition) denoted by the triple asterisk (P-value<0.001). Plotting PC4 against PC3 and PC2 (center left, center right) there is a much clearer separation of TYTO/PA vs. ARIA/ARIA-CD362. The overlap of TYTO/PA cells in the plot would suggest that TYTO and PA are quite similar. The same is true for ARIA/ARIA CD362, their overlapping latent space in PC4 suggests there is not much geometric distance between the two cell conditions.

## Methods

Differential gene tests were conducted using DESeq2 R package. DESeq2 fits negative binomial generalized linear models for each gene and uses the Wald test for significance testing, returning a table of differentially genes. Multiple testing correction was performed using independent hypothesis weighting (IHW) R package, with the significance threshold for genes set at P-value<0.05. Custom R scripts were used to parse the data and annotate the outputs using biomaRt.

For each comparison, differentially expressed genes were loaded into Cytoscape and coloured according to the direction of their Log 2 Fold Change. Solid blue/red colors were assigned if the Log 2 Fold Change exceeded $\pm 3$. Singletons were removed, and edges were created based on evidence of co-expression according to STRING DB. This method served as a means of filtering the differentially expressed genes list, returning putative *bona fide* expression patterns in the samples as inputs for Pathway Analysis.

---

Pathway Enrichment Categories: **GO**; Gene Ontology; Process (GO Biological Process), Component (GO Cellular Components), Function (GO Molecular Function), **UniProt**: database of protein sequence and functional information, **Pfam**: database of protein families and domains, **KEGG**: Kyoto Encyclopedia of Genes and Genomes (high level functionality of biological systems), **InterPro**: domain prediction in proteins, **Reactome**: curated peer reviewed biological pathways.

## Differential Expression Results

### TYTO vs PA (Cytokine treated)

Using the plastic adherent (PA) cells as a control reference, differentially expressed genes between TYTO and PA were detected using DESeq2. Applying a significance cut-off of P-value<0.05, a total of 156 differentially expressed genes were captured (70 upregulated, 86 downregulated). Fully annotated tables containing gene information and DESeq2 statistical metrics can be found at this github repository.

Co-expression analysis inferred by STRING-DB (**Figure 1**) reveals only a small subset of the differentially expressed genes are co-expressed. In agreement with the exploratory analysis results, the TYTO samples are quite similar to the PA samples, and thus will not produce a strong differentially expressed signal. The list of differentially expressed genes was used to perform pathway analysis in Cytoscape. Pathway analysis returned only 8 pathways, most of which are involved in plasma membrane, apical plasma membrane GO Cellular Components. A heatmap of the differentially expressed genes for TYTO vs. PA is given in **Figure 2**.

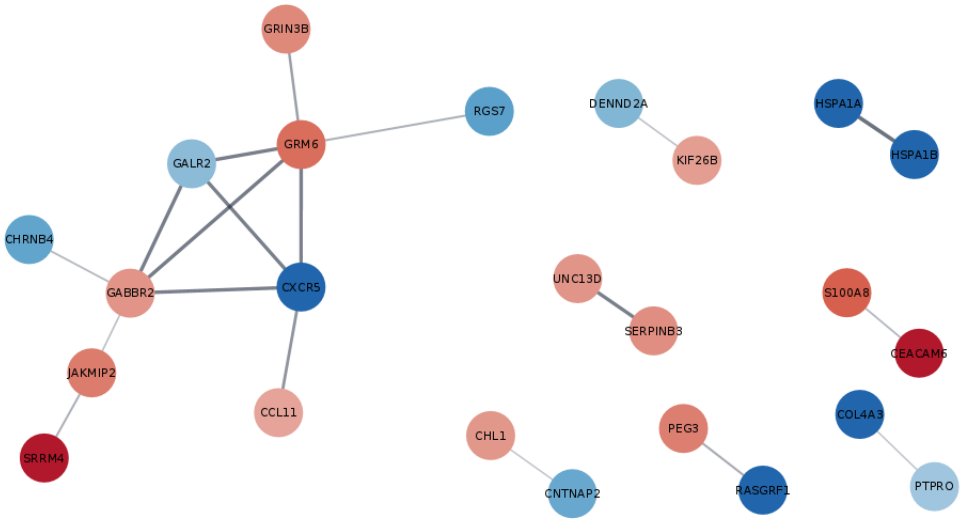


**Figure 1.** Protein-protein interactions inferred by STRING DB present in the list of differentially expressed genes returned by TYTO vs. PA. Node colors refer to the direction of Log 2 Fold Change (Blue: down, Red: Up). Edges between nodes signify a protein-protein interaction as defined by public databases curated by STRING DB

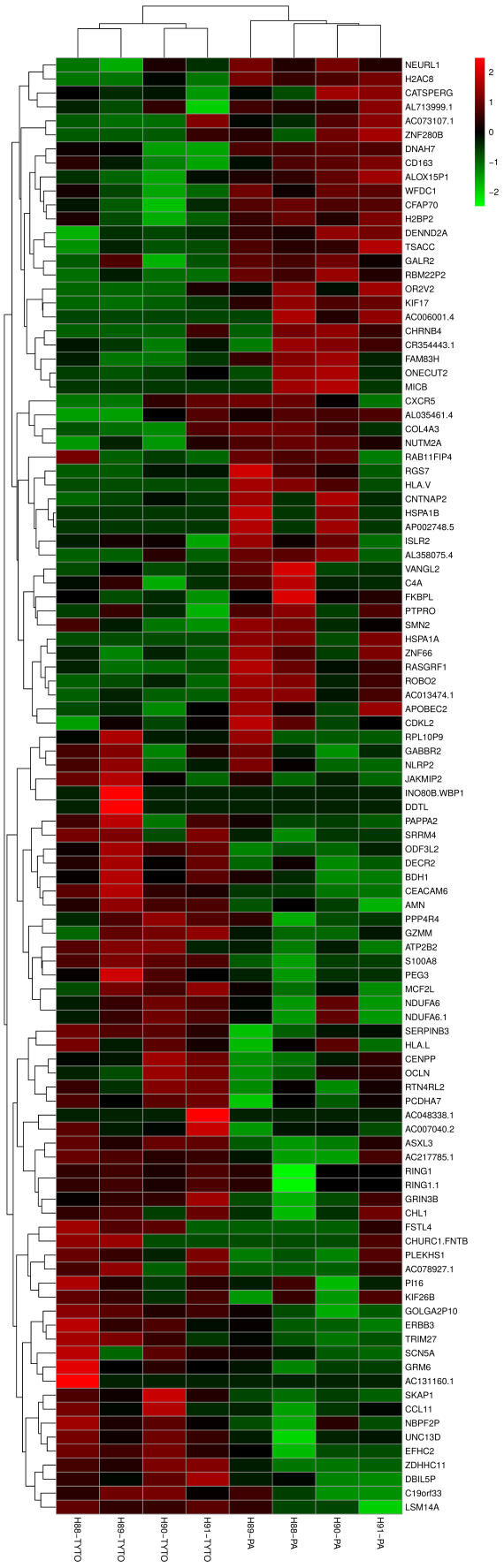


**Figure 2.** Heatmap of differentially expressed genes between Cytokine treated TYTO vs. PA samples. Expression values are scaled log2 normalized counts.

### ARIA vs. PA (Cytokine Treated)

Using the plastic adherent (PA) cells as a control reference, differentially expressed genes between ARIA and PA were detected using DESeq2. Applying a significance cut-off of P-value<0.05, a total of 136 differentially expressed genes were captured (65 upregulated, 75 downregulated). Fully annotated tables containing gene information and DESeq2 statistical metrics can be found at this github repository.

Co-expression analysis performed by STRING reveals the majority of the co-expressed genes are downregulated (upregulated in PA with respect to the ARIA sample), denoted by the blue nodes in **Figure 3**. To focus on upregulated ARIA genes, a seperate pathway analysis was conducted on upregulated genes alone. The results can be found at previously mentioned github repository under the file name 'aria_vs_pa_upregulated_pathways.csv'. The majority of the upregulated pathways were involved in striated muscle contraction, actin-mediated muscle contraction, filament sliding and sarcomeres. This collection of pathways suggest that the ARIA cells are expending ATP at an increased rate when compared to PA controls. Interestingly, Apoptosis was upregulated with genes CEACAM6, PEG3, EPB41L3, S100A8, ZNF385B, NLRP2 involved in the pathway at a statistical significance level of P-value 0.0241. A heatmap of the differentially expressed genes in the ARIA vs. PA comparison is available in **Figure 4**.

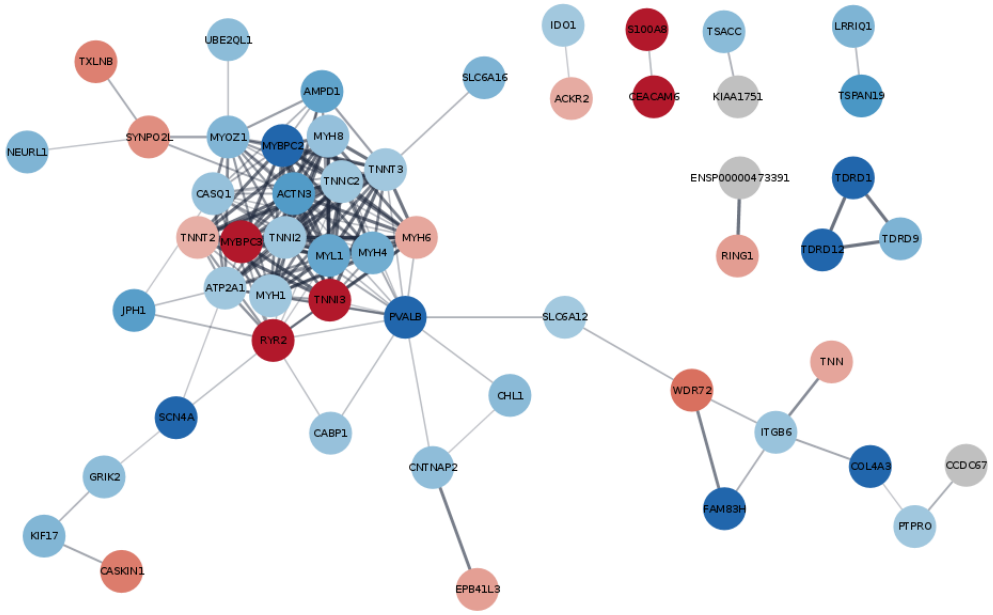


**Figure 3.** Protein-protein interactions inferred by STRING DB present in the list of differentially expressed genes returned by ARIA vs. PA. Node colors refer to the direction of Log 2 Fold Change (Blue: down, Red: Up). Edges between nodes signify a protein-protein interaction as defined by public databases curated by STRING DB

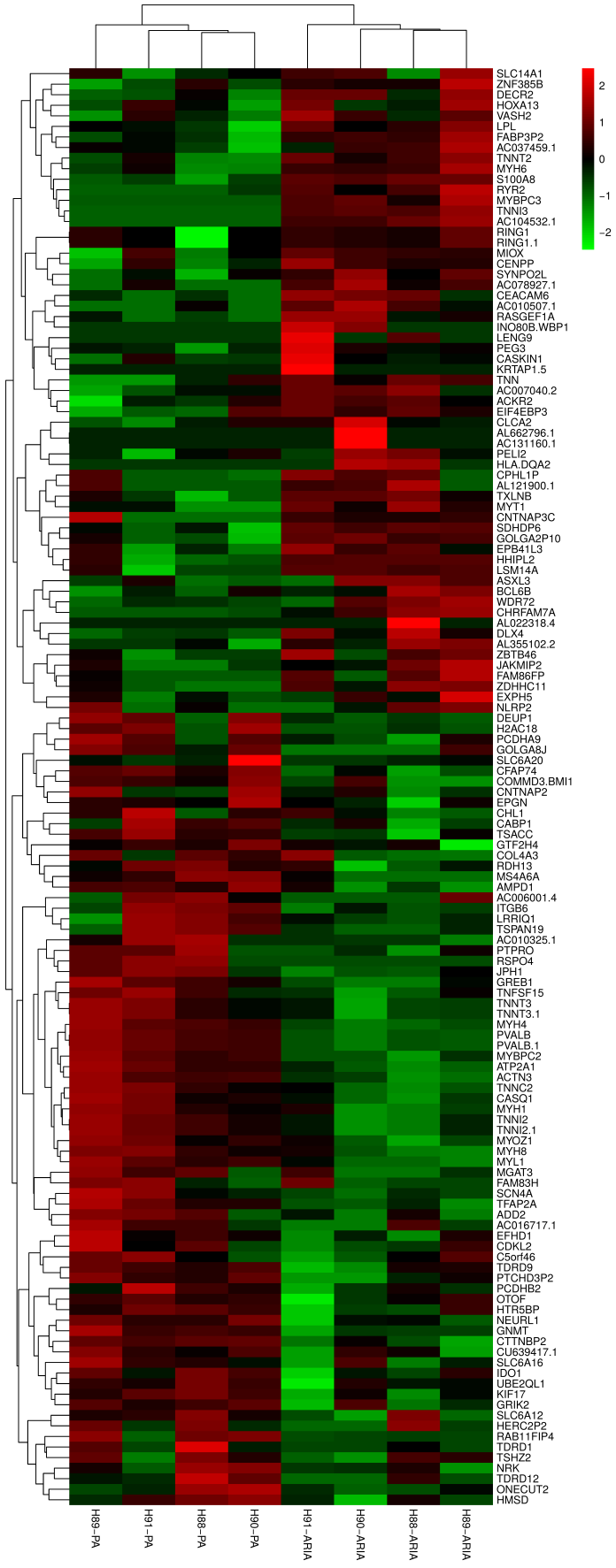


**Figure 4.** Heatmap of differentially expressed genes between Cytokine treated ARIA vs. PA samples. Expression values are scaled log2 normalized counts.

### ARIA-CD362 vs. PA (Cytokine Treated)

Using the plastic adherent (PA) cells as a control reference, differentially expressed genes between ARIA-CD362 and PA were detected using DESeq2. Applying a significance cut-off of P-value<0.05, a total of 130 differentially expressed genes were captured (66 upregulated, 64 downregulated). Fully annotated tables containing gene information and DESeq2 statistical metrics can be found at this github repository.

Co-expression analysis using Cytoscape revealed a large portion of the co-expressed genes are downregulated (as with ARIA, upregulated in PA with respect to ARIA-CD362). To interrogate the upregulated pathways, a separate pathway analysis was conducted, available at the previously mentioned github repository under the file name 'aria-cd362_vs_pa_upregulated_pathways.csv'. As with the ARIA population, the majority of the upregulated pathways were involved in troponin, myofilament, muscle contraction and striated muscle sliding. Yet again these pathways would suggest ARIA-CD362 is expending ATP at a higher rate than PA cells. Furthermore, the upregulation of HLA-DQA2 & HLA-DOB infer the ARIA-CD362 cells are employing Major Histocompatability Complex (MHC) genes as a defense against external stimuli or to dispense with the damaged, dying or infected cells. A heatmap of the differentially expressed genes is available at **Figure 6**.

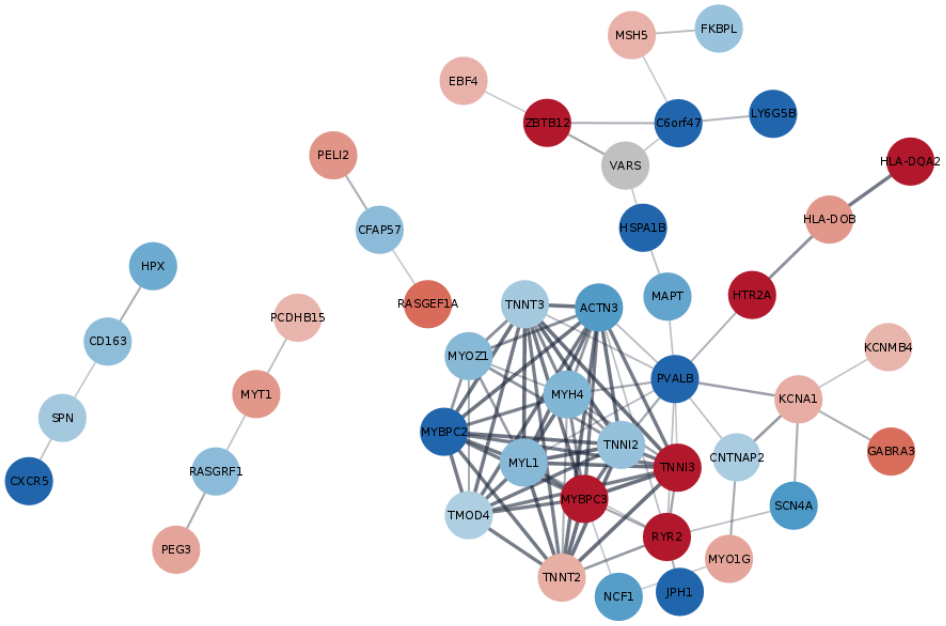


**Figure 5.** Protein-protein interactions inferred by STRING DB present in the list of differentially expressed genes returned by ARIA-CD362 vs. PA. Node colors refer to the direction of Log 2 Fold Change (Blue: down, Red: Up). Edges between nodes signify a protein-protein interaction as defined by public databases curated by STRING DB

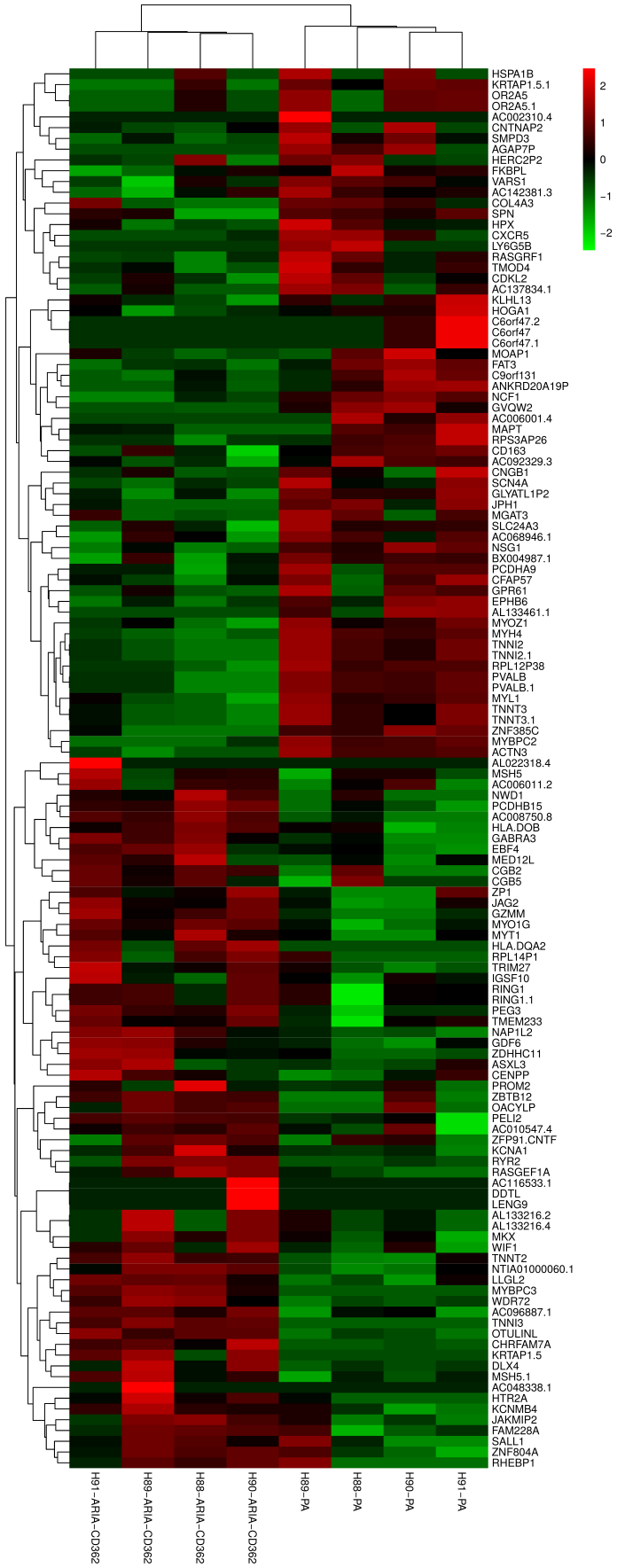


**Figure 6.** Heatmap of differentially expressed genes between Cytokine treated ARIA-CD362 vs. PA samples. Expression values are scaled log2 normalized counts.

### TYTO vs. ARIA (Cytokine treated)

Applying a significance cut-off of P-value<0.05, a total of 140 differentially expressed genes were captured (75 up-regulated, 65 down-regulated). Fully annotated tables containing gene information and statistical metrics from DESeq2 are available at this github repository. Analysis in Cytoscape revealed a small group of co-expressed genes according to STRING DB **Figure 7**.

Pathway analysis of the TYTO vs. ARIA comparison yielded surprising results. Given the results from the exploratory analysis, it was shown that TYTO and PA are more similar than ARIA and ARIA-CD362. With this in mind one would expect a diverse set of differentially expressed genes between TYTO vs. ARIA however this was not the case. TYTO cells were enriched for pathways that ARIA, ARIA-CD362 was enriched for with respect to PA. Sarcomere, myosin, fibronectin, muscle contraction and striated muscle sliding were present in upregulated pathways, available at the previously mentioned github repository. A heatmap of the differentially expressed genes is available at **Figure 8**.

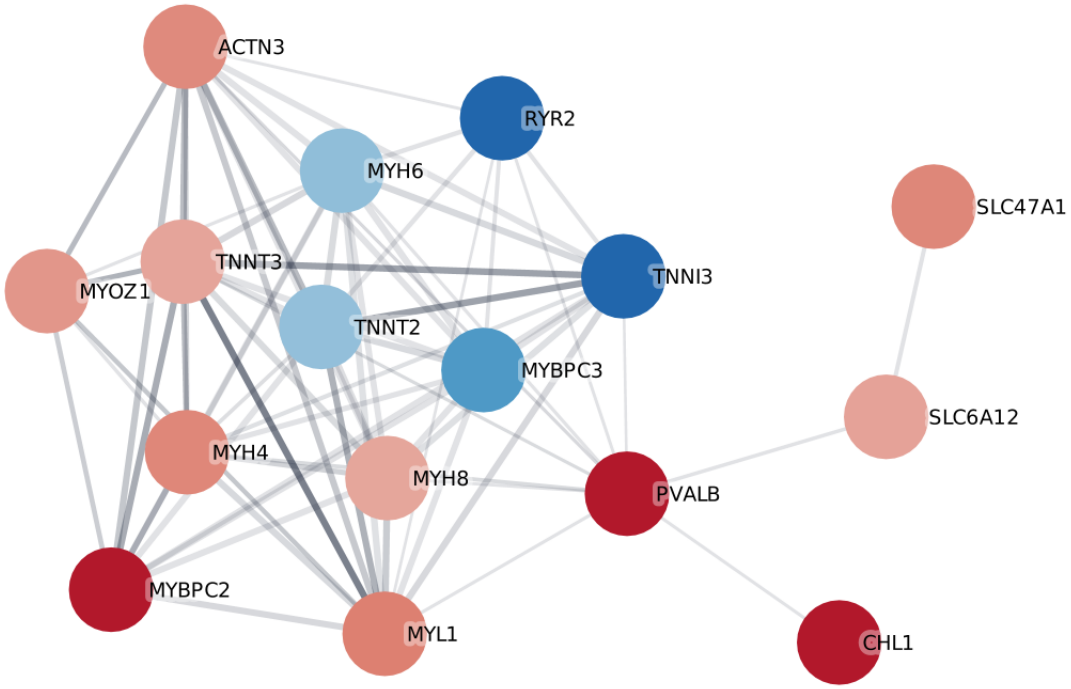


**Figure 7.** Protein-protein interactions inferred by STRING DB present in the list of differentially expressed genes returned by TYTO vs. ARIA. Node colors refer to the direction of Log 2 Fold Change (Blue: down, Red: Up). Edges between nodes signify a protein-protein interaction as defined by public databases curated by STRING DB

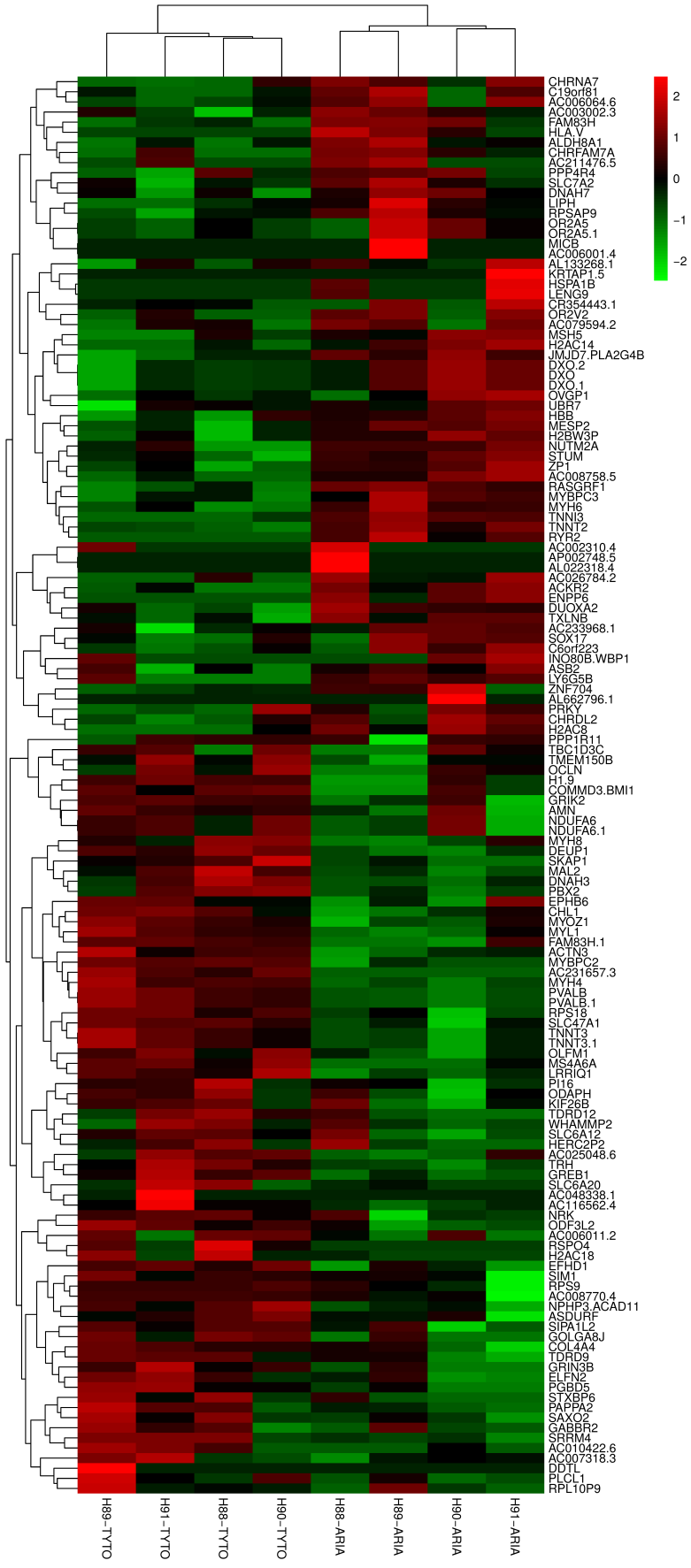


**Figure 8.** Heatmap of differentially expressed genes between Cytokine treated TYTO vs. ARIA samples. Expression values are scaled log2 normalized counts.

### TYTO vs. ARIA-CD362 (Cytokine treated)

Applying a significance cut-off of P-value<0.05, a total of 152 differentially expressed genes were returned using statistical significance threshold of P-value<0.05. Fully annotated tables containing gene information and statistical metrics from DESeq2 are available at this github repository.

Co-expression analysis reveled a coherent set of upregulated, co-expressed genes in TYTO. The results were strikingly similar to TYTO vs. ARIA, a testament of how similar ARIA, ARIA-CD362 are in the dataset. The upregulated pathways suggest cells are undergoing actin-mediated muscle contraction in TYTO vs. ARIA, available at the previously mentioned github repository. A heatmap of the differentially expressed genes is available at **Figure 10**.

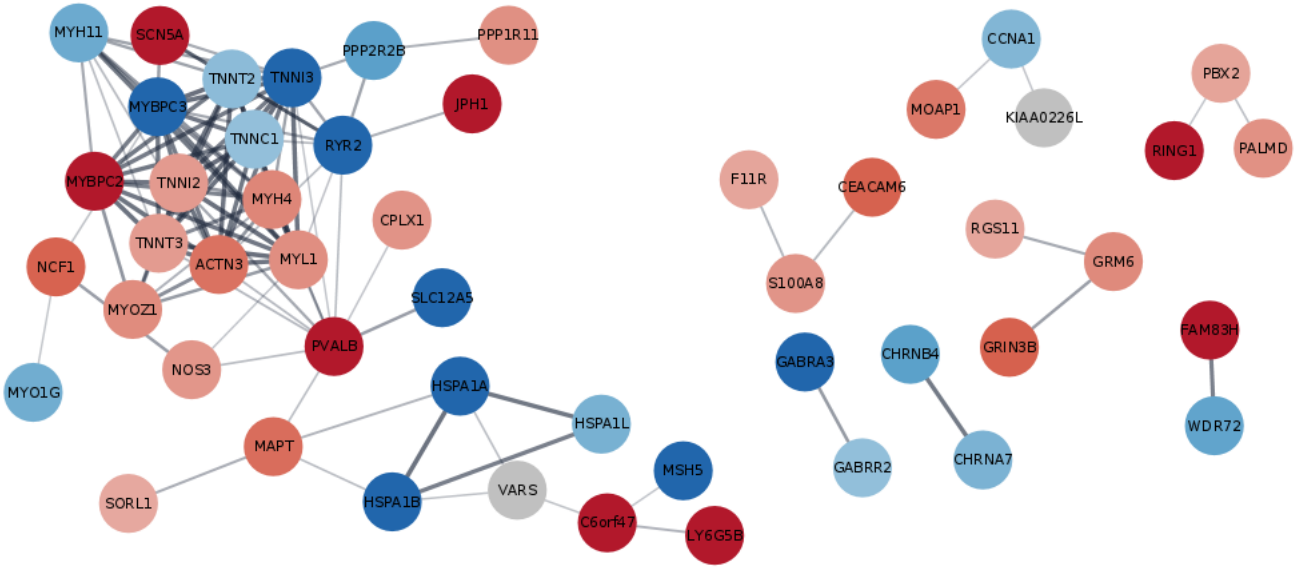


**Figure 9.** Protein-protein interactions inferred by STRING DB present in the list of differentially expressed genes returned by TYTO vs. ARIA-CD362. Node colors refer to the direction of Log 2 Fold Change (Blue: down, Red: Up). Edges between nodes signify a protein-protein interaction as defined by public databases curated by STRING DB.

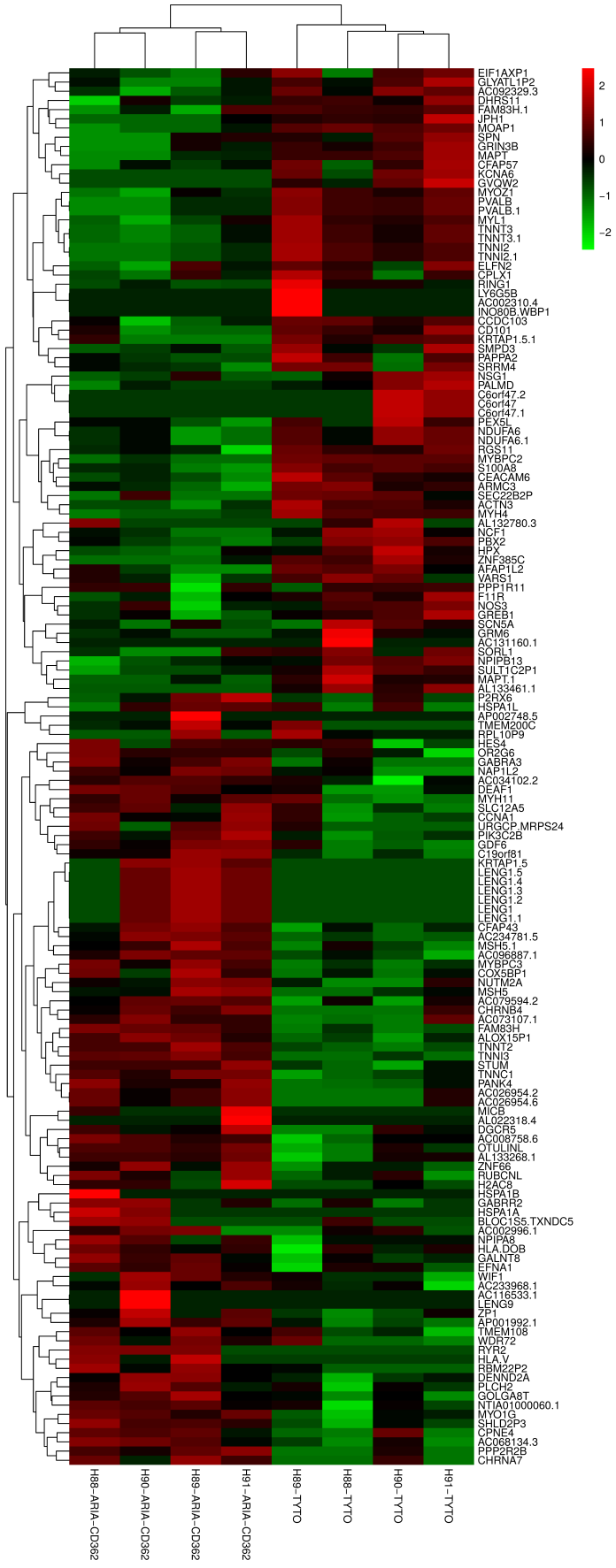


**Figure 10.** Heatmap of differentially expressed genes between Cytokine treated TYTO vs. ARIA-CD362 samples. Expression values are scaled log2 normalized counts.

### TYTO vs. PA (Control Treated)

Applying a significance cut-off of P-value<0.05, a total of 156 differentially expressed genes were captured (70 upregulated, 86 downregulated). Fully annotated tables containing gene information and statistical metrics from DESeq2 are available at this github repository.

Co-expression analysis in Cytoscape revealed no clear signal of upregulated or downregulated genes. A heterogenous network was generated as can be seen in **Figure 11** confounding pathway analysis. Pathway analysis was performed on the upregulated genes in TYTO with respect to PA and is available at the previously mentioned github repository. A small number of upregulated pathways were returned, with all of them involved in enzyme reactions in the cell or cell membrane. A heatmap of differentially expressed genes is available at **Figure 12**.

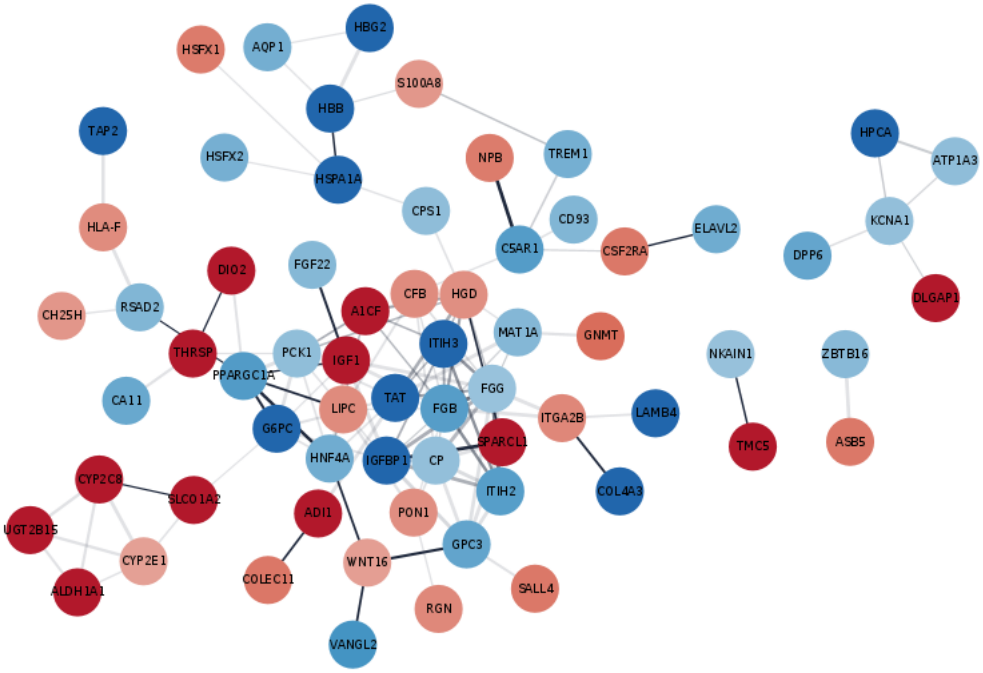


**Figure 11.** Protein-protein interactions inferred by STRING DB present in the list of differentially expressed genes returned by TYTO vs. PA. Edges between nodes signify a protein-protein interaction as defined by public databases curated by STRING DB

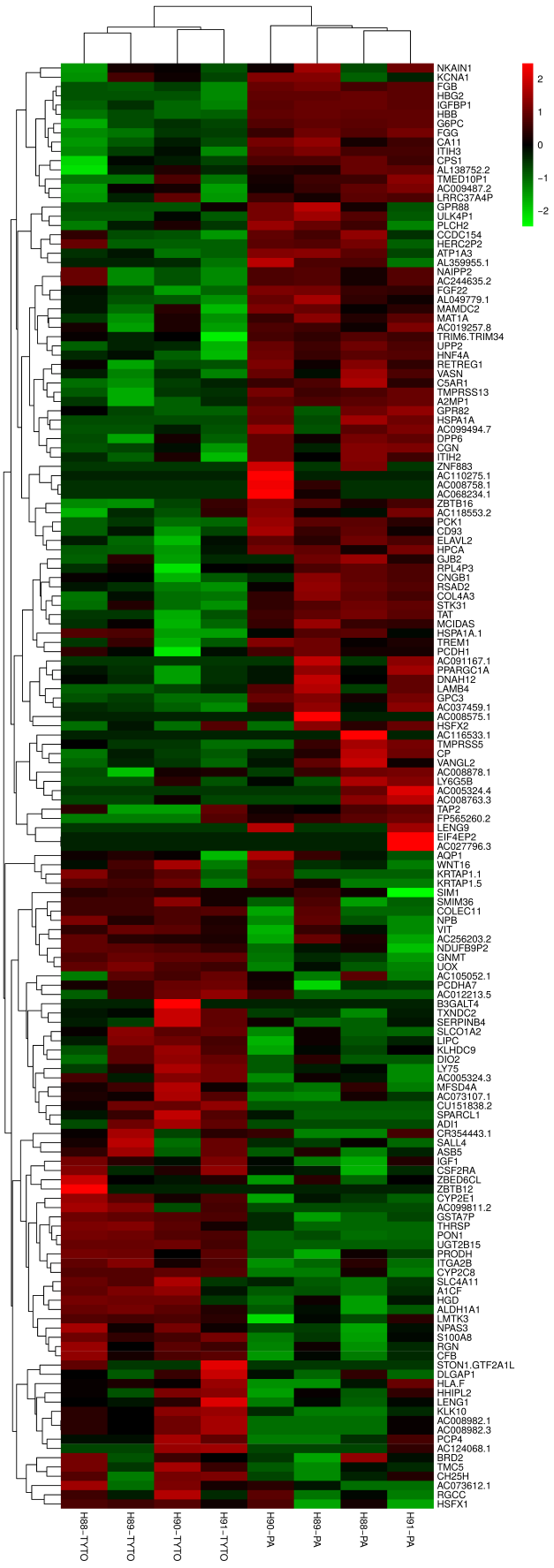


**Figure 12.** Heatmap of differentially expressed genes between Control treated TYTO vs. PA samples. Expression values are scaled log2 normalized counts.

### ARIA vs. PA (Control Treated)

Applying a significance cut-off of P-value<0.05, a total of 208 differentially expressed genes were captured (72 upregulated and 136 downregulated). Fully annotated tables containing gene information and statistical metrics from DESeq2 are available at this github repository. Whilst detecting perturbed pathways in this comparison using Cytoscape **Figure 13**, it is clear that PA genes (blue nodes signify downregulated in ARIA when compared to PA, i.e upregulated in PA) provide a much clearer biological signal. The upreglated genes in ARIA were mostly singletons, and no pathway analysis results were returned.

Upregulated pathways in PA cells can be found at the previously mentioned github repository. PA cells were enriched for coagulation, cell adhesion, wound healing and fibrin clot formation pathways. A heatmap of the differentially expressed genes are available at **Figure 14**.

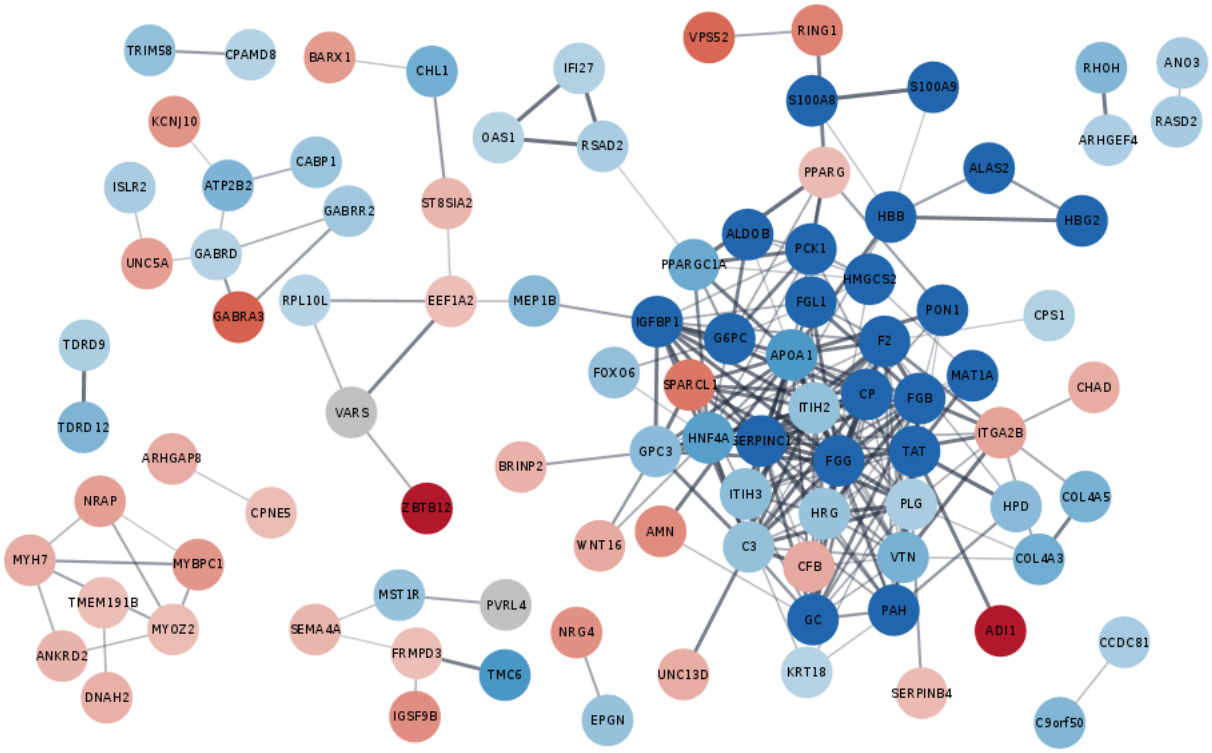


**Figure 13.** Protein-protein interactions inferred by STRING DB present in the list of differentially expressed genes returned by ARIA vs. PA. Edges between nodes signify a protein-protein interaction as defined by public databases curated by STRING DB

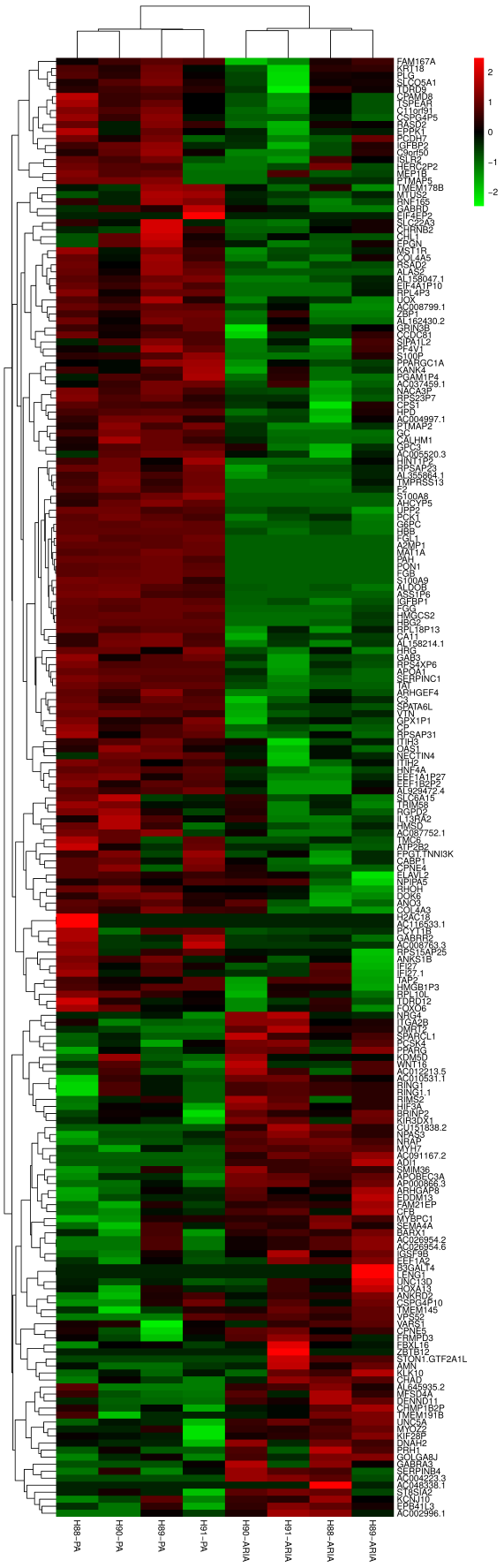


**Figure 14.** Heatmap of differentially expressed genes between Control treated ARIA vs. PA samples. Expression values are scaled log2 normalized counts

### ARIA-CD362 vs. PA (Control Treated)

Applying a significance cut-off of P-value<0.05, a total of 195 differentially expressed genes were captured (162 upregulated, 33 downregulated). Fully annotated tables containing gene information and statistical metrics from DESeq2 are available at this github repository.

Co-expression analysis of the differentially expressed genes favoured expression in PA cells. A separate pathway analysis of upregulated genes in ARIA-CD362 revealed only 6 pathways, most associated with muscle filament sliding, a recurring theme with this dataset. The entire pathway analysis results (favouring PA upregulated genes) is available at the following link. A heatmap of the differentially expressed genes is available at **Figure 16**.

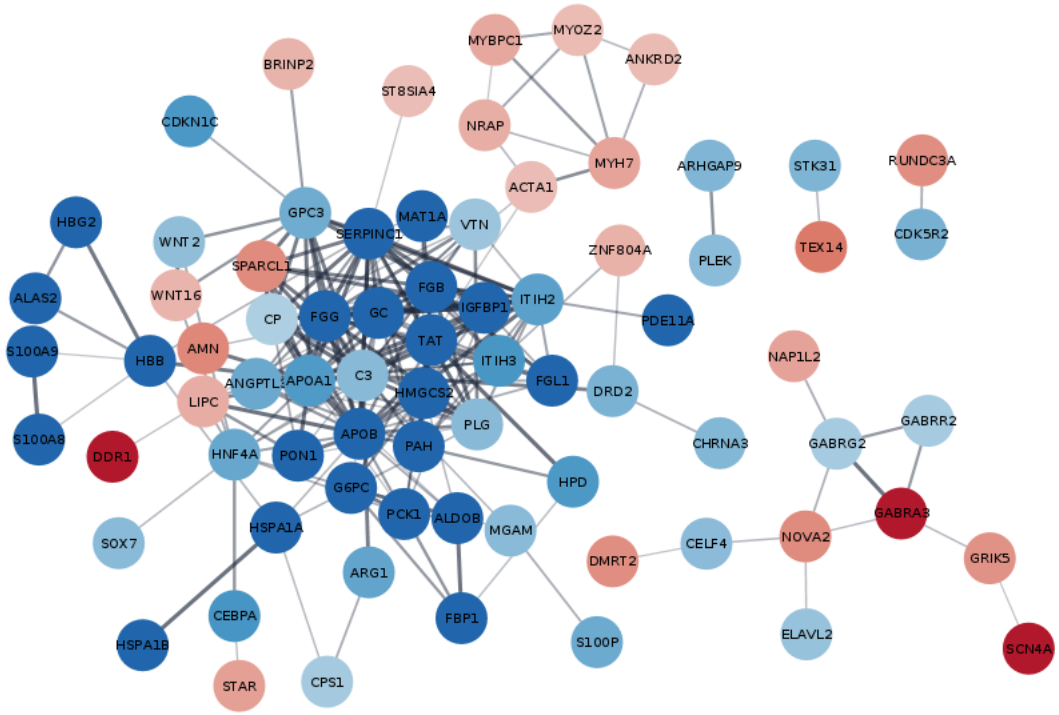


**Figure 15.** Protein-protein interactions inferred by STRING DB present in the list of differentially expressed genes returned by ARIA-CD362 vs. PA. Edges between nodes signify a protein-protein interaction as defined by public databases curated by STRING DB

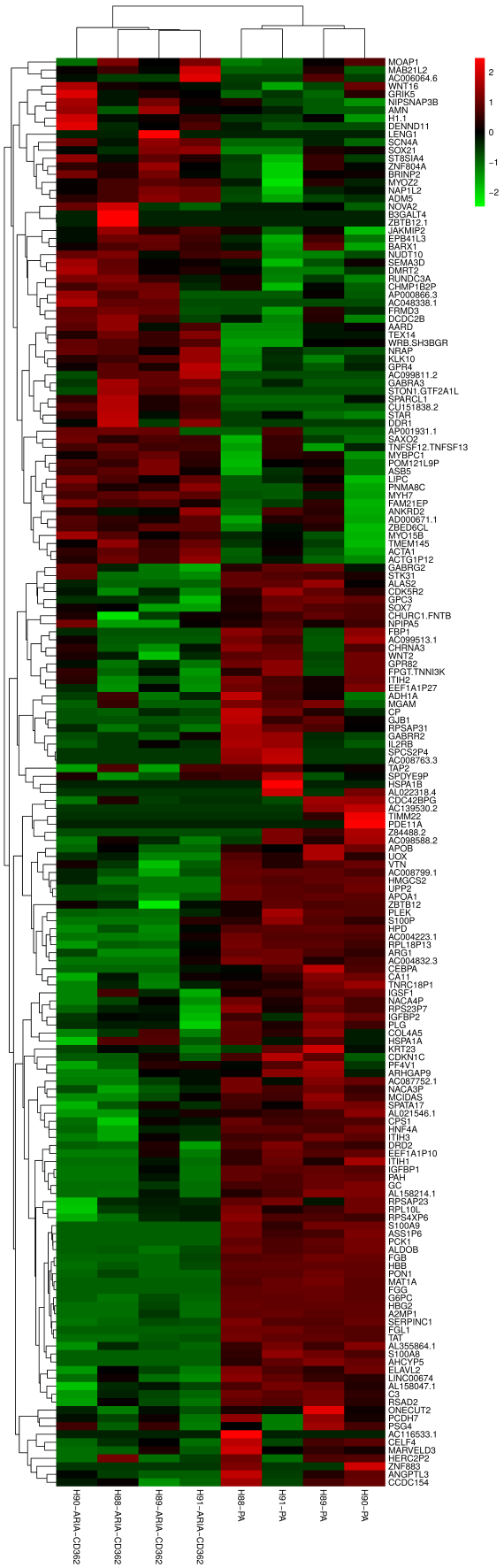


**Figure 16.** Heatmap of differentially expressed genes between Control treated ARIA-CD362 vs. PA samples. Expression values are scaled log2 normalized counts

### TYTO vs. ARIA (Control Treated)

Applying a significance cut-off of P-value<0.05, a total of 178 differentially expressed genes were captured (63 upregulated, 115 downregulated). Fully annotated tables containing gene information and statistical metrics from DESeq2 are available at this github repository. Filtering by co-expressed genes resulted in a network primarily populated by upregulated genes **Figure 17**. This is reflected in the heatmap in **Figure 18**, most of the upregulated genes are present in TYTO samples.

Due to the lack of genes expressed by ARIA in this comparison, only upregulated pathways in TYTO were considered. Close to 300 upregulated pathways were returned, suggesting TYTO cells have a strong, biologically relevant signal when compared to ARIA cells. The pathways enriched are mainly centered around fibrin clot formation, blood coagulation, cell-cell adhesion and a variety of cellular defensive mechanisms.

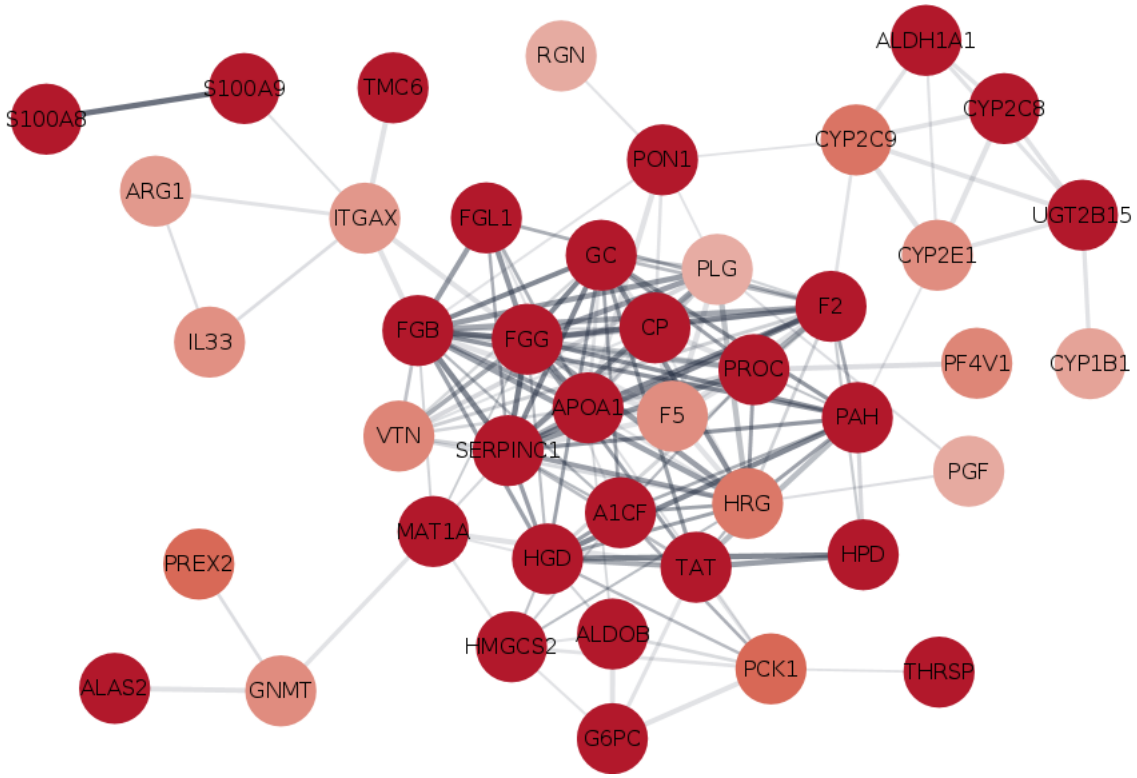


**Figure 17.** Protein-protein interactions inferred by STRING DB present in the list of differentially expressed genes returned by TYTO vs. ARIA. Edges between nodes signify a protein-protein interaction as defined by public databases curated by STRING DB

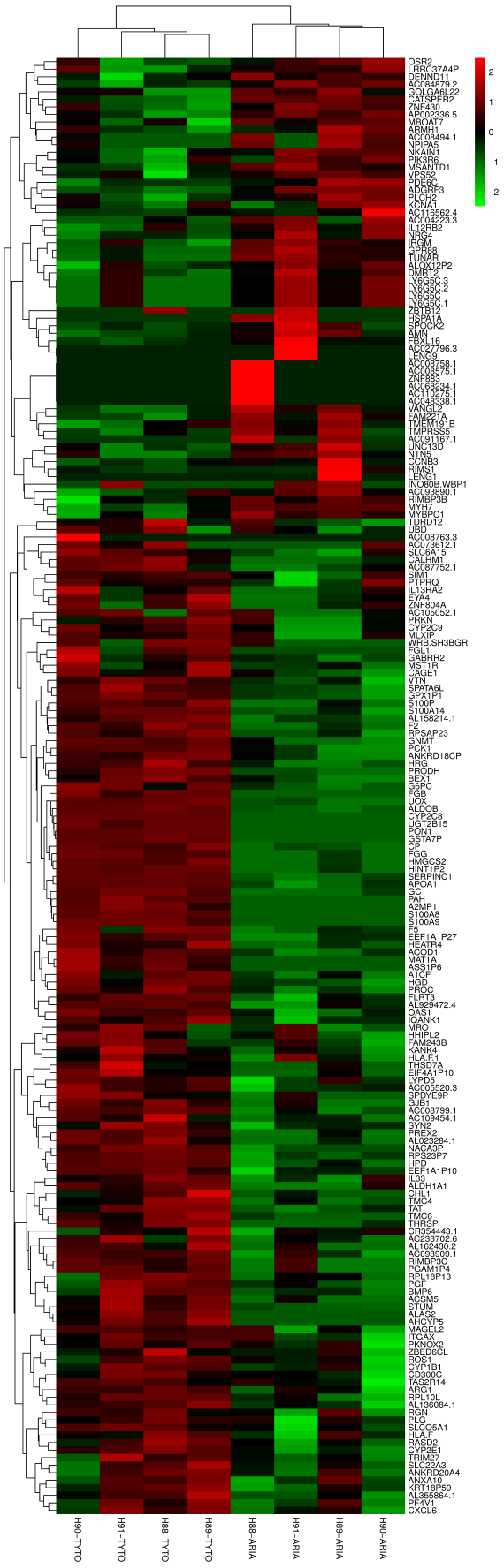


**Figure 18.** Heatmap of differentially expressed genes between Control treated TYTO vs. ARIA samples. Expression values are scaled log2 normalized counts

### TYTO vs. ARIA-CD362 (Control Treated)

Applying a significance cut-off of P-value<0.05, a total of 197 differentially expressed genes were captured (126 upregulated, 71 downregulated). Fully annotated tables containing gene information and statistical metrics from DESeq2 are available at this github repository.

Similarly to TYTO vs. ARIA (again demonstrating the similarity of ARIA and ARIA-CD362 in this dataset), the TYTO vs. ARIA-CD362 comparison is dominated by the upregulated genes in TYTO, as can be seen in the co-expression network in **Figure 19**. Pathway analysis reveals upregulation in several biochemical pathways such as gluconeogenesis, carboxylic acid/amino acid catabolic processes, lipoprotein clearance and a multitude of signalling pathways, possibly mediated by MAPK signalling. A heatmap of the differentially expressed genes is available at **Figure 20**.

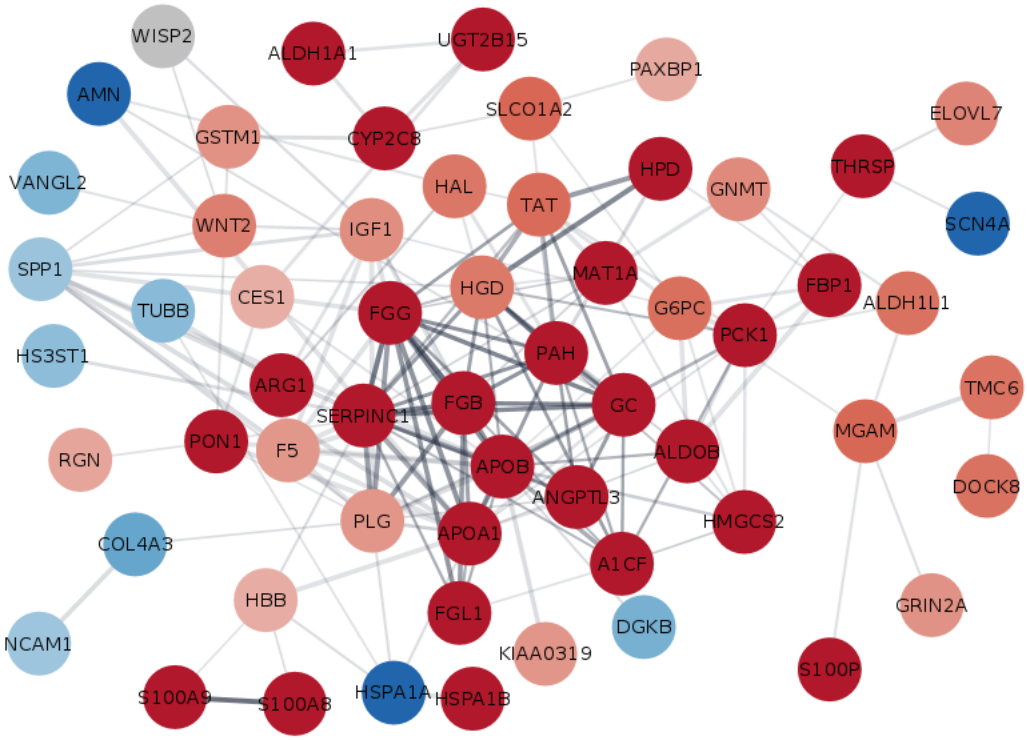


**Figure 19.** Protein-protein interactions inferred by STRING DB present in the list of differentially expressed genes returned by TYTO vs. ARIA-CD362. Node colors refer to the direction of Log 2 Fold Change (Blue: down, Red: Up). Edges between nodes signify a protein-protein interaction as defined by public databases curated by STRING DB.

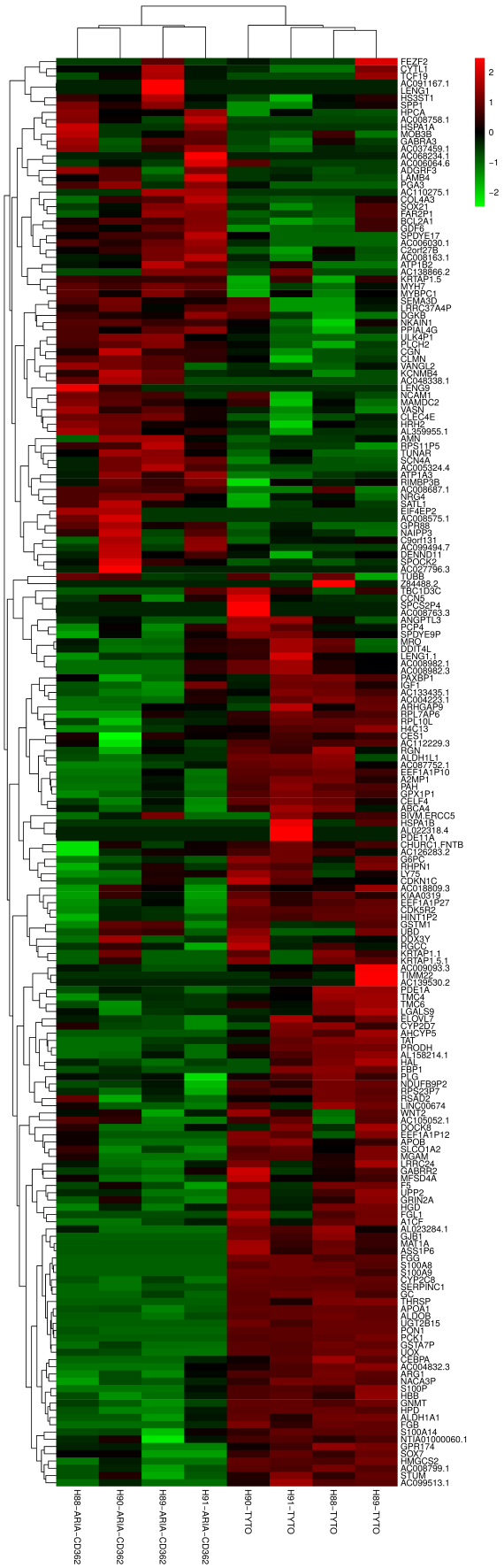


**Figure 20.** Heatmap of differentially expressed genes between Control treated TYTO vs. ARIA-CD362 samples. Expression values are scaled log2 normalized counts

## Discussion

Exploratory data analysis proved useful for identifying large sources of variation in the data, and to assess the degree of variation in the dataset owing to cell types (ARIA, ARIA-CD362, PA and TYTO). In principal component 4 it was shown that the samples cluster according to their cell sorting methods, however it is important to note that PC4 accounts for only $\sim$5% of variation in the dataset. Given these results by the exploratory analysis, it was always going to be a challenge to extract a meaningful signal in each comparison made.

The Cytokine treated population suggested protoplasm tension in the ARIA cells (both ARIA and ARIA-CD362) with the large majority of upregulated pathways involved in actin mediated muscle sliding, sarcomeres and myosin formation. Comparisons of TYTO vs. the plastic adherent control yielded no meaningful signals. After comparing each cell type to the control, they were then compared to each other. TYTO vs. ARIA / ARIA-CD362 returned upregulated genes in actin mediated muscle sliding for the TYTO cells, similarly to ARIA vs. PA comparisons.

The Control treated population had extremely weak ARIA, ARIA-CD362 signals. Each comparison they were involved in was dominated by the other cell type in pathway analysis. TYTO samples had a strong signal involving cell adhesion and wound healing pathways when compared to ARIA samples, and a small set of results in the TYTO vs. PA comparison further confirmed the similarity of TYTO and PA.

## Exploratory Data Analysis Figures

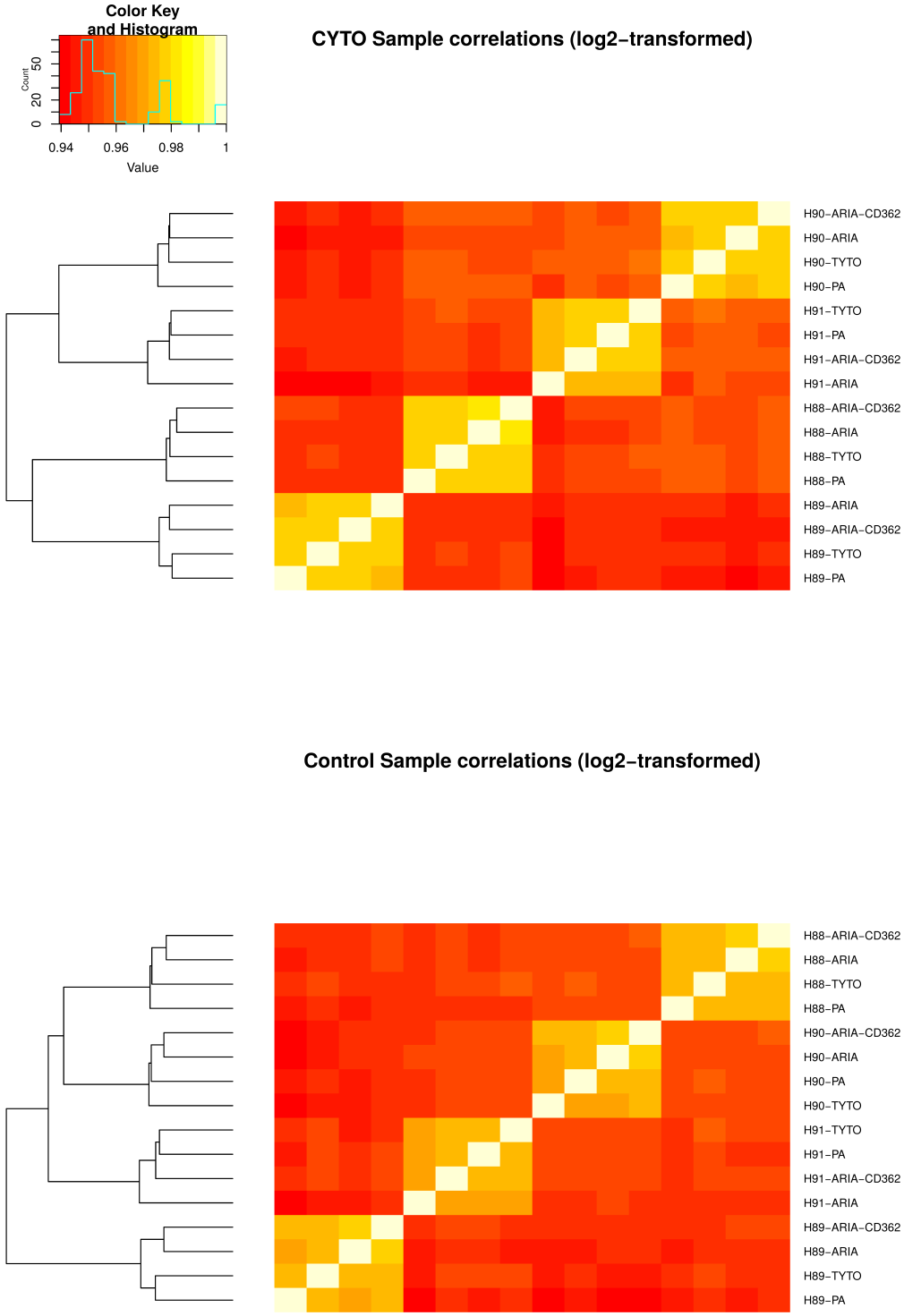


**Figure 21.** Sample to Sample heatmaps for cytokine (top) and control (bottom) treated cells.

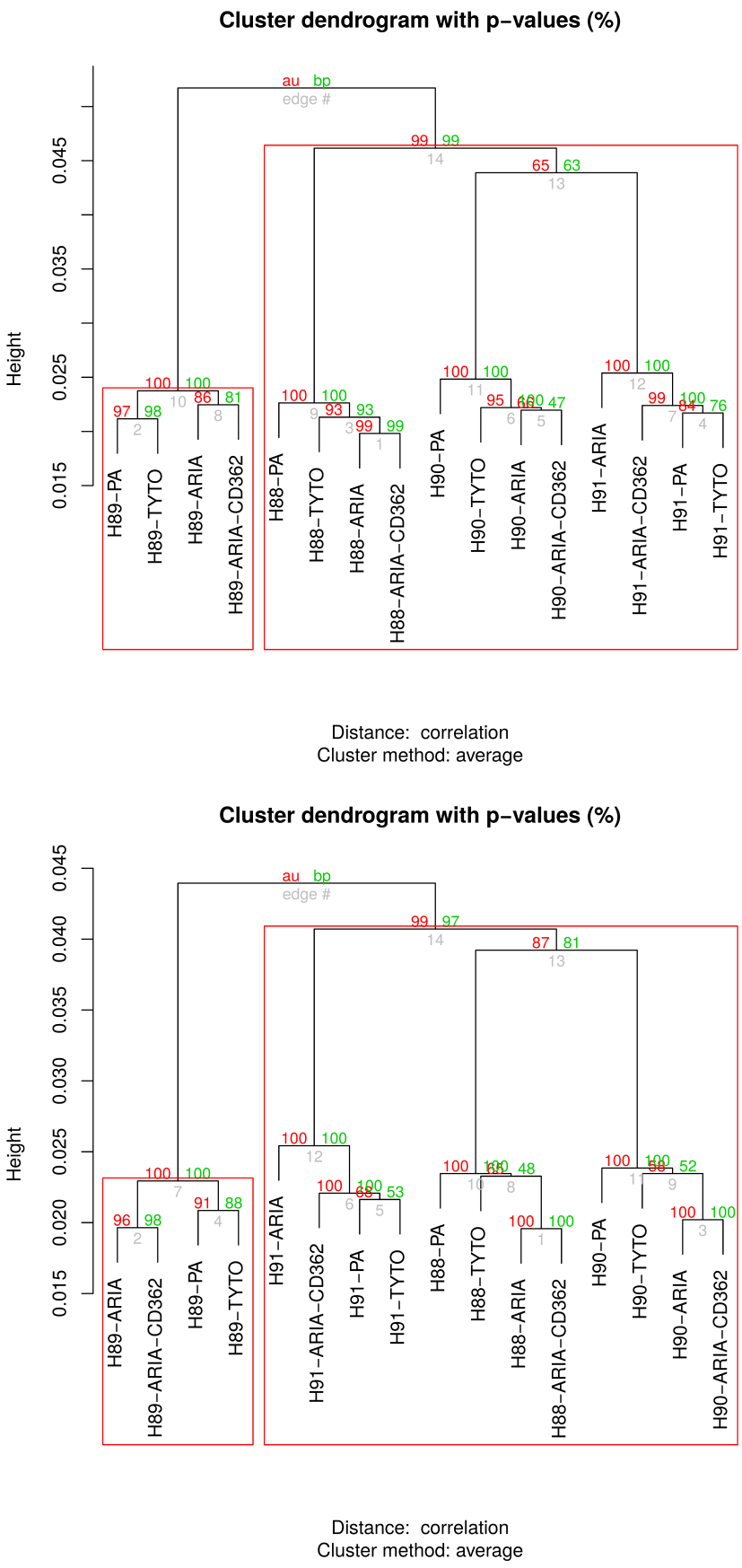


**Figure 22.** Hierarchical clustering of samples. Cytokine (top) and Control (bottom) treated samples.

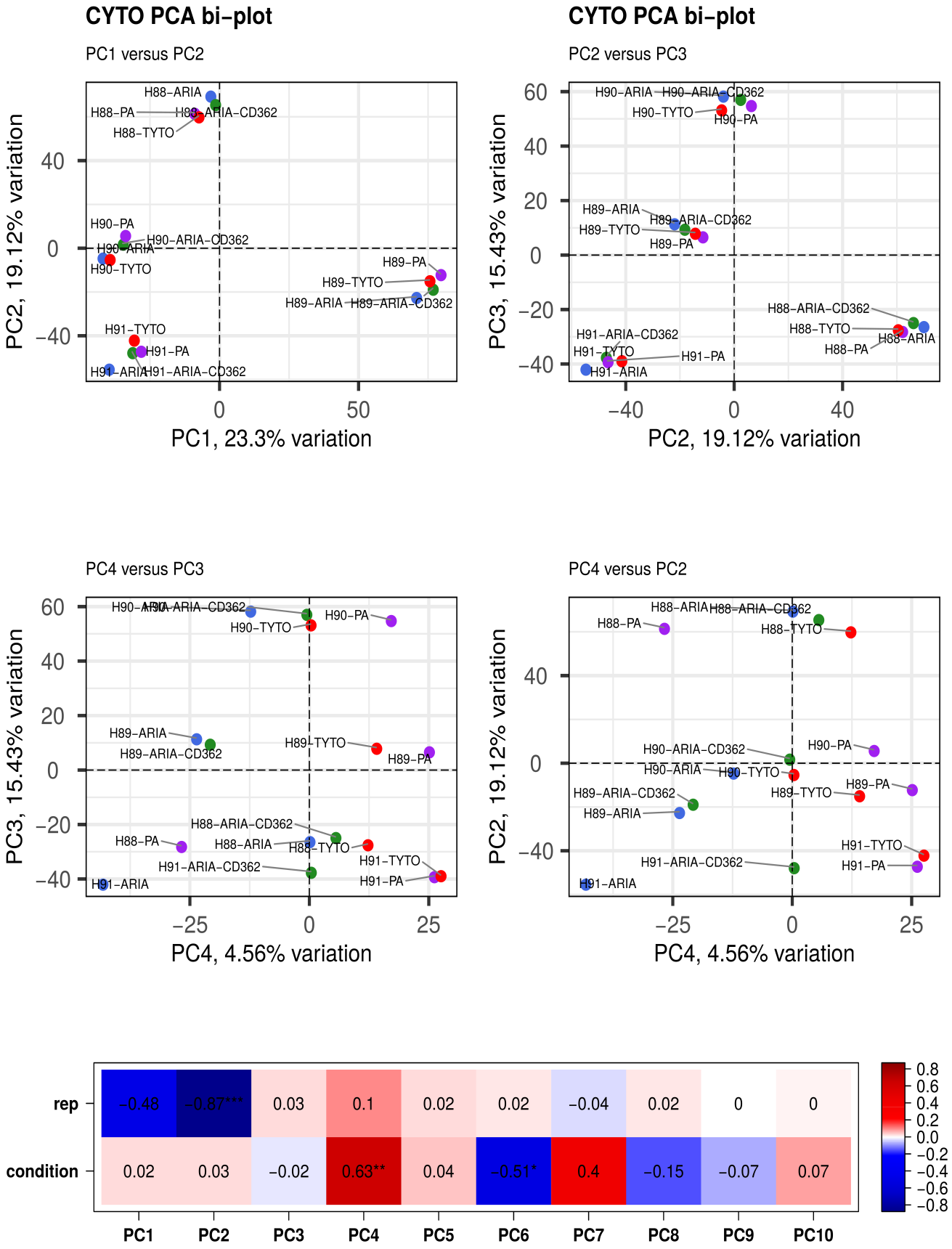


**Figure 23.** PCA Analysis of Cytokine treated samples. Beginning top left: PC1 vs. PC2, PC2 vs. PC3, PC4 vs PC3, PC4 vs. PC2, eigen correlation plot displaying the relationship of principal components with covariates. (*) P-value<0.05, (**) P-value<0.01, (***) P-value<0.001.

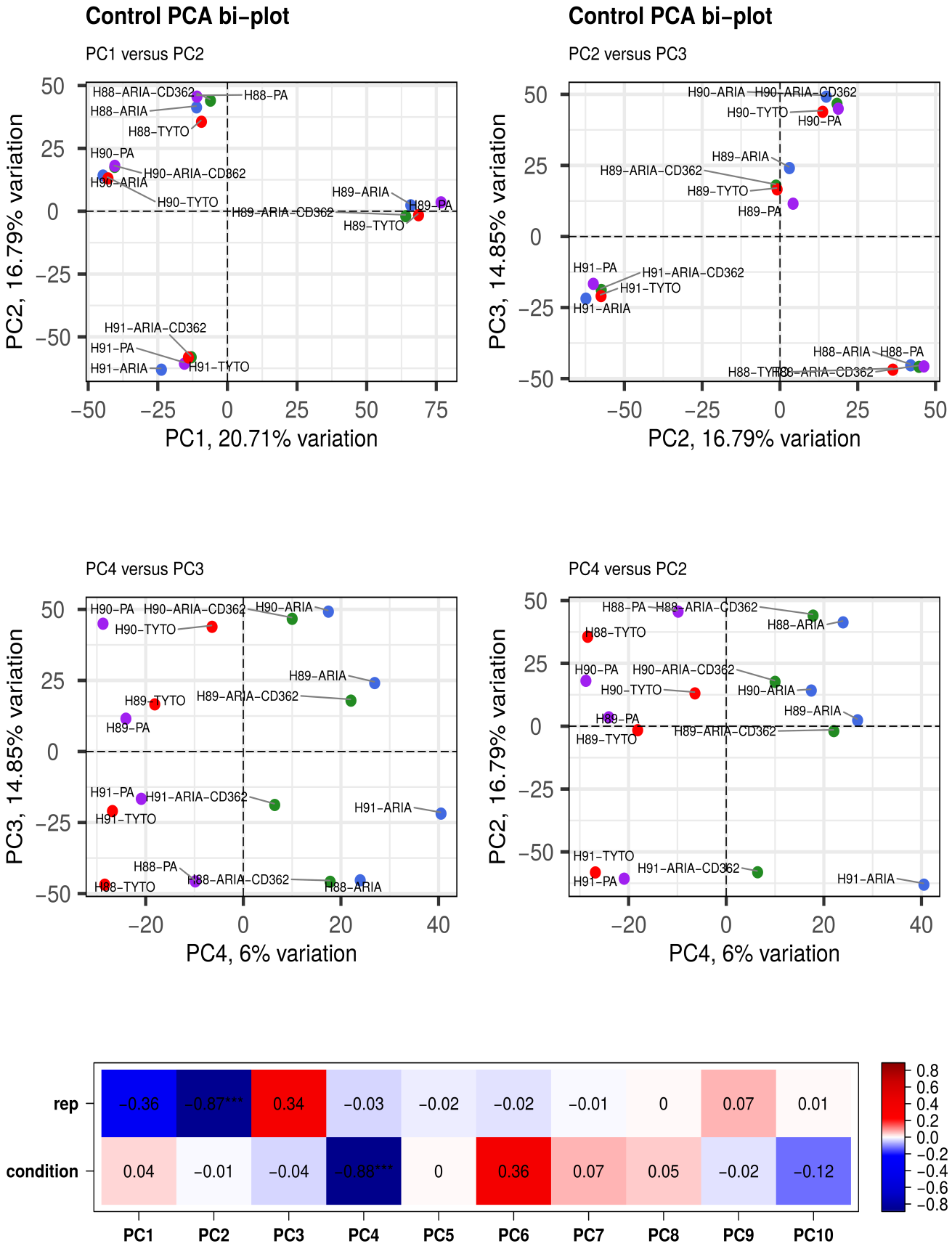


**Figure 24.** PCA Analysis of Control treated samples. Beginning top left: PC1 vs. PC2, PC2 vs. PC3, PC4 vs PC3, PC4 vs. PC2, eigen correlation plot displaying the relationship of principal components with covariates. (*) P-value<0.05, (**) P-value<0.01, (***) P-value<0.001.