UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escola Tècnica Superior d'Enginyeria Agronòmica i del Medi Natural

"EFFECT OF *CLOSTRIDIUM* TOXINS ON INTESTINAL EPITHELIAL CELL MORPHOLOGY"

FINAL DEGREE PROJECT IN BIOTECHNOLOGY

Wageningen University and Research

Systems and Synthetic biology
Food & Biobased Research

Student: Judith Cantó Santos

Supervisor in WUR: Jurriaan Mes

UPV Supervisor: Joaquín Cañizares Sales

Academic Year: 2016-2017

Valencia, July 2017

ABSTRACT

The thesis explained here is called the "*Effect of Clostridium Toxins on Intestinal Epithelial Cell Morphology". Clostridium difficile* is a gram-positive bacterium that causes bowel disorders in humans. They enter human body through food and target the intestinal epithelia to gain entry. They produce two toxins (Toxin A and Toxin B) that help them gain entry through the intestinal epithelial barrier. Caco-2 cells are commonly used as model cell systems to mimic intestinal enterocytes. Thus, it is being studied the effect of the clostridium toxins using Caco-2 and mouse cecum as intestinal epithelia models.

The aim of this research is to cross compare these model systems and focus on known pathways like the Rho GTPases signaling (from Ingenuity Pathway Analysis software) along with other important genes and pathways that affect cell morphology. In this regard, the work line is a literature search for genes affected by the effect of *Clostridium difficile* toxins. Then proceed to use R (statistical scripting language) and Bioconductor packages (maSigPro) to understand the changes occurring in these genes. 2 different datasets will be handled in total: one time points data (mice) and one Caco-2 gene microarray data.

Comparing the difference in expression of the genes under the effect of the toxins and under basal conditions after 2 hours and 6 hours in mice cecum and 4.5 hours in Caco2 have revealed the first regulators to be affected after toxins infection. Then, we proceed to relate these significant genes to their pathways to see which ones where the most affected. This systems approach allowed a broad view of the effect of *Clostridium difficile* infection in two model systems, mice and Caco2. Moreover, the comparison of mice cecum and Caco2 colon-like cells showed diverse toxins effects that are a consequence of differences between these systems.

KEY WORDS:

Clostridium difficile
Toxin A
Toxin B
Cell morphology
Gene expression
Epithelial cell

Student: Judith Cantó Santos

Supervisor in WUR: Jurriaan Mes

Supervisor in UPV: Joaquín Cañizares Sales

Valencia, July 2017

RESUMEN

El trabajo aquí descrito se denomina *"Efecto de las toxinas de Clostridium sobre la morfología de las células epiteliales intestinales*". *Clostridium difficile* es una bacteria gram positiva que causa trastornos intestinales en humanos. Entran en el cuerpo humano a través de los alimentos y se dirigen el epitelio intestinal para invadir los tejidos. Producen dos toxinas (Toxina A y Toxina B) que les ayudan a entrar a través de la barrera epitelial intestinal. Las células Caco-2 se usan comúnmente como sistemas de células modelo para imitar enterocitos intestinales. Así mismo, se está estudiando el efecto de las toxinas de *Clostridium* usando Caco-2 y ciego de ratón como modelos de epitelios intestinales.

El objetivo de esta investigación es comparar estos sistemas modelo y centrarse en las vías conocidas como la ruta de señalización de las Rho GTPasas (mediante Ingenuity Pathway Analysis software), junto con otros genes y vías importantes que afecten la morfología celular. Respecto a esto, la línea de trabajo consiste en realizar una búsqueda bibliográfica para obtener un conjunto de genes que estén afectados por los efectos de las toxinas de *Clostridium difficile.* A continuación, se procederá a utilizar R (lenguaje de escritura estadístico) y paquetes de Bioconductor (maSigPro) para comprender los cambios que se producen en estos genes. Se manejarán 2 conjuntos de datos diferentes en total: un conjunto de datos de puntos de tiempo (ratones) y datos de un microarray de genes de Caco-2.

Comparando la diferencia en la expresión de los genes bajo el efecto de las toxinas y en condiciones basales después de 2 horas y 6 horas en el ciego de ratones y 4,5 horas en Caco2, se han obtenido los primeros reguladores afectados por la infección de toxinas. El siguiente paso fue relacionar estos genes significativos con sus vías de señalización para ver cuáles son los más afectados. Este enfoque de sistemas permitió una visión amplia del efecto de la infección por *Clostridium difficile* en dos sistemas modelo, ratones y Caco-2. Por otra parte, la comparación del ciego de ratón con Caco2, que son células similares a las del colon, mostraron diversos efectos de las toxinas que son consecuencia de las diferencias entre estos sistemas modelo.

PALABRAS CLAVE:

Clostridium difficile
Toxina A
Toxina B
Morfología Celular
Expresión Génica
Célula Epitelial

Estudiante: Judith Cantó Santos

Tutor en WUR: Jurriaan Mes

Tutor en UPV: Joaquín Cañizares Sales

Valencia, Julio 2017

INDEX

GENERAL INDEX

INDEX OF FIGURES

INDEX OF TABLES

INDEX OF ABBREVIATIONS

DE - Differentially Expressed

RMA - Robust Multiarray Analysis

PCA - Principal Component Analysis

FC - Fold Change

IPA - Ingenuity Pathway Analysis

TcdA – Toxin A of *Clostridium difficile*

TcdB – Toxin B of *Clostridium difficile*

TcdAB – Toxin A and Toxin B of *Clostridium difficile*

TJ – Tight Junction

1. INTRODUCTION

1.1. *Clostridium difficile*

Clostridium difficile is a gram-positive bacterium that is part of the commensal microbiota of some individuals. *C. difficile* affects the colon because it is an anaerobic bacterium, and it is in this part of the intestine where the medium is anaerobic. Moreover, most of the gut microbiota is present in the colon, so *C. difficile* coexist with other species there. This bacterium produces spores, which can germinate in presence of glycine and cholate derivatives (Di Bella et al., 2016). These compounds are present in the colon but are processed by other bacteria from the intestine, preventing germination of *C. difficile* spores.

However, after treatment with antibiotics, the microbiota is destabilized, which prevents the metabolism of cholate by healthy bacteria and thus, favours *Clostridium difficile* spores' germination and overgrowth (Burns et al., 2010). Therefore, an increase in the population of *C. difficile* causes inflammation of the colon. The infection of *C. difficile* can trigger diarrhoea and, most importantly, pseudomembranous colitis. This infection can be treated with certain antibiotics that are effective for the bacterium (Moreno et al., 2013).

Clostridium difficile infection can be detected if an individual shows the following symptoms: watery diarrhoea (with several bowel movements a day for 2 or more days which may contain blood), fever, loss of appetite, nausea and abdominal pain. If these symptoms remain, *C. difficile* infection can become more serious and produce dehydration, electrolyte imbalances, low blood pressure, and even bowel perforation, kidney failure, or ultimately death (Moreno, et al., 2013).

The pathogenesis elicited by *Clostridium difficile* depends on two factors: a decrease in proportion of the gut microbiota after treatment with antibiotics and production of *C. difficile* toxins (high-molecular-weight toxins A and B) (Longo et al., 2015).

1.2. *Clostridium difficile* toxins

Most of *Clostridium difficile* strains produce two toxins: toxin A (TcdA), which is an enterotoxin, and toxin B (TcdB) (cytotoxin), produced by the genes tcdA and tcdB, respectively (Braun et al., 1996). *C. difficile* strains have a great genetic variability and are divided according to the types of toxins they synthesize, based on genetic variations of tcdA and tcdB genes (Di Bella et al., 2016). These genes have a multi-modular domain structure called ABCD model (A: biological activity; B: binding; C: cutting; D: delivery) (Jank, & Aktories, 2008). The role of each domain is:

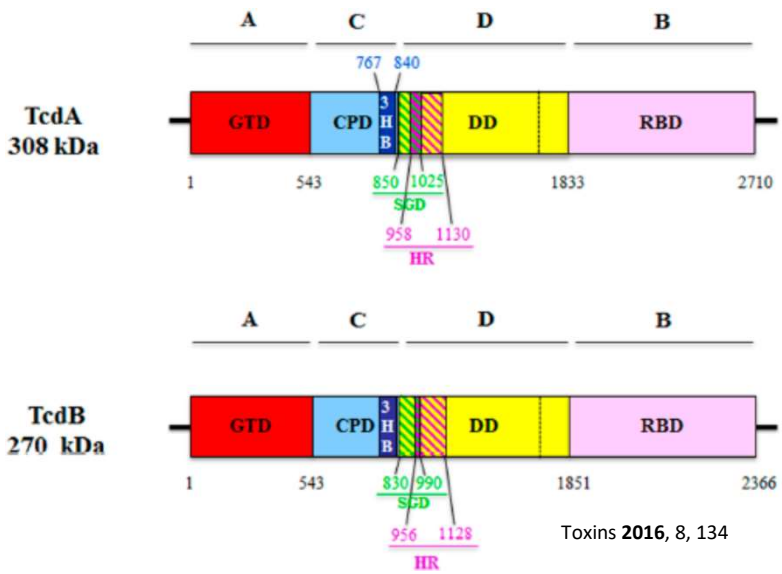
- A domain: corresponds to the N-terminal glucosyltransferase domain (GTD). It acts on the small Rho GTPases involved in regulation of the cytoskeleton.

- B domain, corresponding to the receptor binding domain (RBD), which consists of combined repeated oligopeptides (CROPs).

- C domain: related to the cysteine protease domain (CPD), which promotes the auto-catalytic cleavage of the toxins; and the three-helix bundle domain (3HB), so far identified in TcdA.

- D domain: corresponds to the delivery hydrophobic domain (DD), involved in the translocation of the toxins into the cytosol, as well as their binding to target cells. In this DD domain, there is

the small globular domain (SGD) in TcdA that corresponds to the minimal pore forming region (MPFR) in TcdB, and overlaps only partially to the hydrophobic region (HR) of TcdA and TcdB.



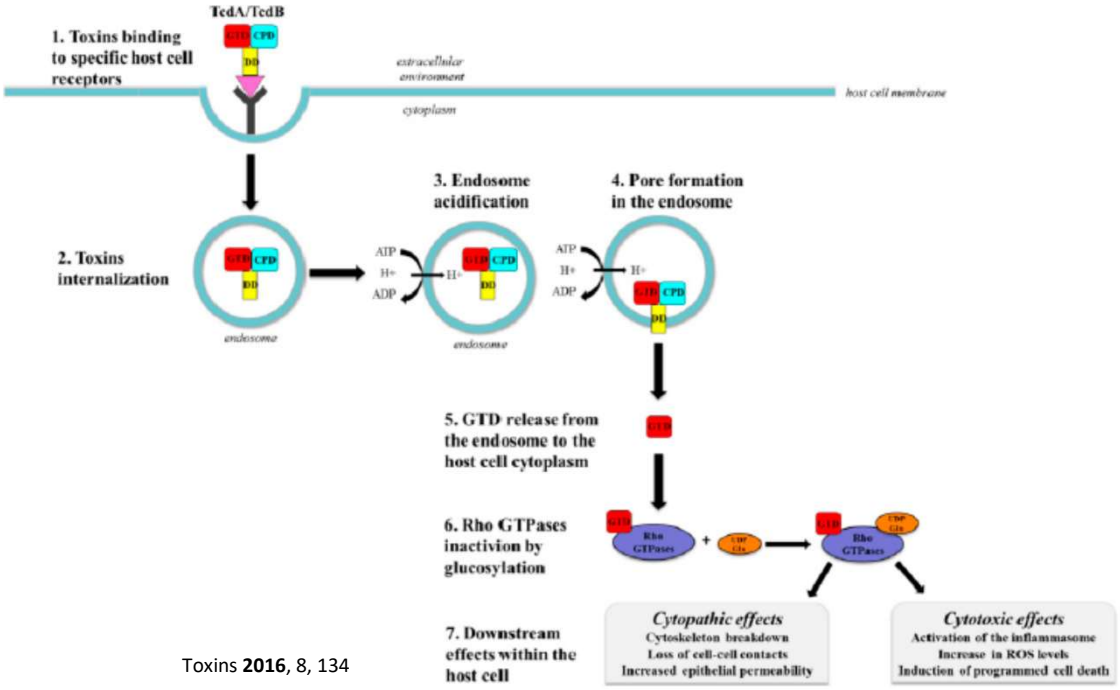
Toxins **2016**, 8, 134

Figure 1. Scheme of the multi-modular domain structure (ABCD model) of the genes encoding TcdA and TcdB *Clostridium difficile* toxins. ABDC model: A: biological activity; B: binding; C: cutting; D: delivery domains. (Di Bella et al., 2016)

The synthesis of TcdA and TcdB is controlled by environmental conditions and regulators, as nutrient availability, temperature, redox potential, etc. (Bouillaut et al., 2015). Generally, the transcription of tcdA and tcdB genes occurs when bacteria enter the stationary phase in which there is nutrient limitation or accumulation of growth inhibiting substances. Once expression of the toxin genes is induced, the toxin proteins accumulate inside the cell and are slowly released over the course of several hours. When reaching the intestine, *Clostridium difficile* enters a vegetative state, begins to spread, and starts to secrete TcdA and TcdB (Rupnik et al., 2009). However, the cellular toxicity induced by TcdA and TcdB not only affects intestinal cells, but can also cause systemic problems including ascites, pleural effusion, cardiopulmonary arrest, hepatic abscess, abdominal compartment syndrome, acute respiratory distress syndrome, and renal failure (Di Bella et al., 2016).

The pathogenic effects of CDI start when both TcdA and TcdB enter the cells and inactivate Rho GTPases, producing cytopathic and cytotoxic effects.

1.3. Molecular Mechanism of TcdA and TcdB toxins



Toxins **2016**, 8, 134

Figure 2. Toxins delivery into the host cell cytosol can be divided into seven main steps: (1) toxin binding to the host cell surface receptor; (2) toxins internalization through a receptor-mediated endocytosis; (3) endosome acidification; (4) pore formation; (5) GTD release from the endosome to the host cell cytoplasm; (6) Rho GTPases inactivation by glycosylation; and (7) downstream effects within the host cell, i.e., toxins-induced cytopathic and cytotoxic effects.
Colours represent: GTD: N-terminal glucosyltransferase domain (in red); CPD: cysteine protease domain (in cyan); DD: delivery domain (in yellow) (Di Bella et al., 2016).

The first step of *Clostridium difficile* infection mediated by TcdA and TcdB toxins is the binding of the toxins to the cellular surface of the host, and their further endocytosis. TcdA and TcdB are uptaken through a clathrin and dynamin-dependent mechanism, as their RBD (binding domain) recognizes receptors on the surface of the target cells in the host (Jank, & Aktories, 2008). Each toxin has a different surface receptor: Sucrase-isomaltase and the glycoprotein 96 (gp96) are related to TcdA; whereas the chondroitin sulfate proteoglycan 4 (CSPG4) and the poliovirus receptor-like 3 (PVRL3) are found in TcdB.

When toxins bind their receptors, they enter the cell through endocytosis. However, this endocytic vesicle needs to be more acidic to translocate the toxins to the cytosol. Acidification of the complex toxin-receptor allows a conformational change in the toxins, which is necessary to expose their hydrophobic regions and thus, facilitate the entrance into the cytosol. Once in the cytosol, the toxins carry out an autocatalytic cleavage that releases GTD domain (N-terminal glucosyltransferase domain). GTD targets Rho GTPases in the cytosol.

Rho GTPases are glycosylated by TcdA and TcdB. When Rho proteins become glycosylated they are inactivated, promoting the pathogenetic effects of *Clostridium difficile* infection. As Rho GTPases mediate several signal transduction pathways, their glycosylation binds Rho proteins to cell membranes irreversibly and prevents Rho-GTPases dependent signalling. In concrete, those Rho proteins inactivated by TcdA and TcdB are: RhoA, RhoB, RhoC, RhoG, Rac1, Rac2, Rac3,

Cdc42, and TC10 (Papatheodorou, et al. 2010). The inhibition of the interaction between Rho proteins and their effectors produced cytopathic and cytotoxic effects (Di Bella et al., 2016).

The cytopathic effects are observed as morphological changes, as actin cytoskeleton is disrupted and thus, it affects tight and adherent junctions, loss of cell-cell contacts, and increased epithelial permeability, all of which are probably the cause of diarrhoea (Chen et al., 2015). Furthermore, reduced cell adherence causes apoptosis and cell loss, due to limited epithelial cell renewal and inhibition of cell proliferation. Consequently, cell cycle progression and actin-dependent cytokinesis are suppressed.

Regarding cytotoxic effects, they are the result of RhoA inactivation and upregulation of the pro-apoptotic early gene product RhoB, involved in the regulation of programmed cell death. The cytotoxic effects are also associated with the activation of the inflammasome by the glycosylated RhoA, which is probably the cause inflammation and colitis induced by C. difficile (Ng et al., 2010).

The downstream effects of *C. difficile* toxins have been analysed through different approaches. One approach is through a systems perspective. This way, some experiments investigated the transcriptional profile of *Clostridium difficile*, TcdA and TcdB, host cells and colon of mice under the effect of *C. difficile* toxins. Janvilisri *et al.* (2010), analyse the transcriptional profile of host cells and *C. difficile*, to understand the host-pathogen interactions*,* by comparing the changes in transcription that occur in *Clostridium difficile* and human colorectal epithelial Caco-2 cells during infection. They found differentially expressed genes that encode molecules potentially involved in CDI (Janvilisri et al., 2010).

Another experiment conducted by D'Auria *et al.* (2012), was particularly interested in the effects of TcdA and TcdB in the host. They observed the transcriptional profile of human colonic epithelial HCT-8 cells infected with TcdA and TcdB *in vitro* at 2h, 6h and 24h (D'Auria et al., 2012). They found pathways implicated in the impairment of the epithelial barrier, what promotes CDI pathogenesis (Kim et al., 2010). The same research group perform a study of the effect of TcdA and TcdB *in vivo* (D'Auria et al., 2013). The idea of the experiment was to analyse the genome-wide responses after an intracecal injection of toxin into mice over 16-h time course (measurements taken at 2h, 6h and 16h) to link cellular and physiological responses.

2. AIM

In this research, the aim was to uncover the initial interactions between *Clostridium difficile* toxins and host cells. Two datasets were used to see the effect of TcdA and TcdB in gene expression and physiological changes of the host: the first corresponds to microarray data from an intracecal toxin infection in mice (D'Auria et al., 2013), and the second, to human colorectal epithelial Caco-2 cells under toxin infection.

Mice dataset contain data from 2 hours, 6 hours and 16 hours after TcdA, TcdB and TcdAB infection. In this experiment, we will use 2h and 6h time-points under TcdAB infection to look into the first steps of the infection and the genes and pathways that become affected earlier. The idea is to distinguish which genes are differentially expressed at 2h, and if their expression is maintained or returns to basal levels at 6h. Thus, we will conduct a gene expression analysis of 2h and 6h mice dataset, followed by a pathway analysis that will aid to identify the first downstream effects triggered by TcdAB.

Regarding Caco2 data, these cells were infected with *C. difficile* toxins for 4.5 hours. The idea is to compare their differentially expressed genes, obtained after gene expression analysis, and their most significant pathways to the ones obtained in the cecum of mice.

Comparison of genes and pathways from two different datasets from different organisms (mice and human) is expected to provide information about the mechanism of action of the toxins at the first moments of the infection and might help in the search for target molecules with a key role in the pathogenesis of TcdAB infection.

3. MATERIALS AND METHODS

3.1. Sources of microarray data

Two data sets were analysed: mice data belonged to the Department of Biomedical Engineering and Department of Medicine of the University of Virginia (Charlottesville, Virginia, USA); and Caco2 cells to Wageningen Food & Biobased Research of Wageningen University (Wageningen, The Netherlands).

Regarding mice data, they were infected with 20 ng/ml of TcdA and TcdB during 16h (D'Auria et al., 2013). Gene expression was measured at 2h, 6h and 16h by hybridizing the mRNA extracted from the cecum of mice into Affymetrix mouse genome 430 2.0 GeneChip. The data obtained was placed in NCBI Gene Expression Omnibus (GEO) repository under the accession GSE44091 (Barrett et al., 2013) .

On the other hand, Caco-2 cells were cultured for 7 days, as by that time Caco2 cells resemble colon epithelium, and exposed to Toxin A and Toxin B for 4.5 h. A gene expression analysis to obtain the differentially expressed genes in Caco2 was performed in Wageningen Food & Biobased Research Centre (Wageningen University) in a former research. This analysis was done by the student Architha Ellappalayam, as part of her Master Thesis called "Comparison of HCT8, Caco2 and mouse in the light of *Clostridium* toxin studies". Thus, in this research the data from the gene expression in Caco2 was taken and used for pathway analysis. This means that Caco2 data will be considered in the pathway analysis, but is not shown in the gene expression analysis.

3.2. Processing of data files

The first data files processed were those from mice (GSE44091). Using R-scripting language, Cell Intensity Files (CEL files) were extracted from GSE44091 data file and raw data was processed. Packages taken from Bioconductor aided the procedure. In concrete, the packages "affy" and "limma" allow to extract and normalize the data (Gentleman et al., 2004).

One of the first steps of the analysis, consists in converting the probes from the arrays into genes. For this, BrainArray database was used to download mapping files that were provided when the raw data was read (using "affy") (MICROARRAY LAB, 2017). Normalization of the data was performed by RMA (Robust Multi-Array Average Analysis) normalization (from "affy"), which includes background correction (to remove non-specific signal), between-array normalization (to uniformize the data within an experiment) and reporter summarization (extract a gene expression value from the probes that target its transcripts). RMA ensures that only present, perfect-match (PM) probes are considered, discarding mismatch probes.

The quality of the raw data before and after normalization was examined by plotting boxplots (affyPLM package) (Hahne et al.,2010). Furthermore, hierarchical clustering of the data was performed to observe if the groups of samples occur because of type of toxin (TcdA or TcdB) or time (2h, 6h, 16h) (or both) (Bell, 2007). Moreover, a Principal Component Analysis aided to examine the data already normalized and observe the correlation of the variables in the plot.

To read the expression data in an easier way, a design matrix was built. For each toxin and time-point, there were expression values that corresponded to each. With this matrix, we performed a pooled t-test using the linear model. A contrast matrix was then formed by comparing each toxin and time-point to its control. This matrix was again fitted to the linear model, and after the empirical Bayesian approach was applied to it as well.

Differentially expressed genes were considered according to the p-value. A p-value lower than 0.01 was applied to distinguish differentially-expressed genes. The results were observed with volcano plots (fold change vs p-value). In order to avoid false positives as a result of multiple t-tests, the p-value was adjusted using Benjamin and Hochberg's false discovery rate (FDR). However, if the p-value was corrected for mice genes after 2h of toxin exposure, there were no differentially expressed genes considered. Therefore, we used raw p-values without adjusting for multiple comparisons both for mice genes at 2h and at 6h. We did not consider DE genes at 16h in this research. Eventually, differentially expressed genes were placed into Excel files.

3.3. Gene annotation

The differentially expressed genes were identified with their Affymetrix ID. They were transformed to Entrez Gene ID before they were submitted to DAVID. Entrez Gene is a gene-specific database from the National Center for Biotechnology Information (NCBI) which contains records from completely-sequenced genomes (Maglott et al., 2011). For each gene, Entrez Gene forms unique identifiers (GeneID), which are used to integrate information of their nomenclature, descriptions, gene-products, etc.

The genes associated with GeneID were inserted in DAVID. This is a Database for Annotation, Visualization and Integrated Discovery (DAVID) (Huang et al., 2009). The objective is to provide functional annotation tools to be aware of the biological meaning of large list of genes. To do this, it allows to identify enriched biological terms (like GO terms), cluster related annotation genes or visualize them in pathway maps, convert gene identifiers from one type to another, etc.

The list with our differentially expressed genes were placed into the Gene Accession Conversion Tool in DAVID, to assign them a DAVID identifier, species to which they belong and a gene name. This aided us to check if the genes were indeed from mice (*Mus musculus*).

3.4. Orthologue transformation

The further step was to transform our mice genes into human ones. To do so, we needed a database with orthologues. In Ensembl (Aken et al., 2016), a genome browser for vertebrate genomes which annotates genes, computes multiple alignments, predicts regulatory function and collects disease data, we found the tool BioMart (Smedley et al., 2015). This is an open data source management system that provides an interface to biomedical databases from different research groups worldwide.

One of BioMart's databases is Mouse Genome Informatics (MGI) (Shaw, 2016). It contains information about mouse genome features, locations, alleles, and orthologs, hosted by Jackson Laboratory, in US. We used MouseMine (Motenko et al., 2015) to access mouse data from MGI, as it is a query form from MGI that allows to export data in text files. MouseMine is powered by InterMine framework (Smith et al., 2012), which permits the integration and analysis of complex biological data from many different biological data sources and formats.

Using MouseMine, we introduced the list of genes we obtained from DAVID, to create a list in this data source. With this list, we searched for the human orthologues of our mice genes via HumanMine (Dessimoz et al., 2012), another database from InterMine. HumanMine integrates *Homo sapiens* and *Mus musculus* genomic data. We obtained a table with the orthologues of the mice genes.

To check these orthologues from HumanMine, we introduce their primary gene identifier in DAVID (as for the mouse genes), to assess if they were human orthologues and if their annotated function matches in both databases. We also compare the orthologues with the ones they obtained in the research were mice data was taken (D'Auria et al., 2013).

3.5. Pathway Analysis

Using all mouse genes under the effect of TcdA and TcdB with their corresponding fold change and p-value at 2h and at 6h, we perform a Core Analysis with Ingenuity Pathway Analysis (IPA) software from Qiagen (Hilden, Germany). For a detailed description of Ingenuity Pathways Analysis, visit https://www.qiagenbioinformatics.com/. This tool analyses, integrates and interprets data from 'omics experiments, including microarrays experiments.

The idea is to find biological meaning of these large data sets to further identify targets or biomarkers involved. To do this, IPA compares the datasets we introduced with the information of the Ingenuity Pathway Knowledge Base, a data repository containing biological interactions and functional annotations from many databases, to identify model relationships between genes, proteins, cells, drugs and diseases. These modelled relationships are manually reviewed to increase their accuracy and provide links to their primary literature.

A key feature of IPA is that it contains full content of information for Human, Mouse and Rat. If the data introduced belongs to other species, it is mapped to orthologues of one of these three species using HomoloGene (NCBI). HomoloGene is an automated tool from NCBI that identifies putative homology groups from annotated gene sets of a wide range of completely sequenced eukaryotic genomes.

The Core Analysis of all mouse genes was done under a p-value threshold of 0.01. The aim of the Core Analysis is to identify relationships, functions, mechanisms or pathways significant to a dataset. It was based on an expression analysis, based on the expression log ratio.

The Core Analysis predicted canonical pathways and upstream regulators involved and their activation or inhibition state. Although we introduced the differentially expressed mice genes at 2h and at 6h under the effect of TcdA, TcdB and TcdAB, we will refer only to the results obtained under TcdAB effects. The Canonical Pathways check if the molecules of the dataset introduced belonged to pre-defined pathways, and if so, predict their activation or inhibition by calculating an activity score called z-score. If z-score is positive, a pathway is activated, if it is negative, a pathway is suppressed. P-value of the whole pathway is also calculated and we use a pathway p-value lower of 0.05 to consider a pathway to be significant. Moreover, in these pathways, upstream regulators are involved. After identifying the main regulators affecting the observed gene expression changes, a network of these regulators and targets is displayed, to hypothesize the effect of these molecules in gene regulatory networks.

Caco2 cells with their genes, p-values and expression log ratios were already placed in IPA by a previous research conducted in the same department. Thus, we directly took the results from the Canonical Pathways of the Core Analysis to compare them with the pathways analysed in mice.

4. RESULTS AND DISCUSSION

4.1. Gene Expression Analysis

The first part of the research consisted in a gene expression analysis. Datafiles were processed and normalized using R studio. A wide range of functions and images aided us to find the differentially expressed genes under TcdAB infection in mice at 2h and at 6h (code in annex).

The starting point of our data analysis was to perform an RMA normalization (Robust Multi Array Average). Considering that an expression analysis was going to be done, normalizing the expression values was necessary to ensure a proper comparison between treated and control samples. Thus, box plots were made in R studio, both for unnormalized and normalized mice samples (Figure 3).

The raw data (Figure 3) showed a great variability among data samples. This is the result of background noise, which affects the quality of the data. After RMA normalization, the variability is uniform across the data samples, ensuring a proper removal of the background noise.

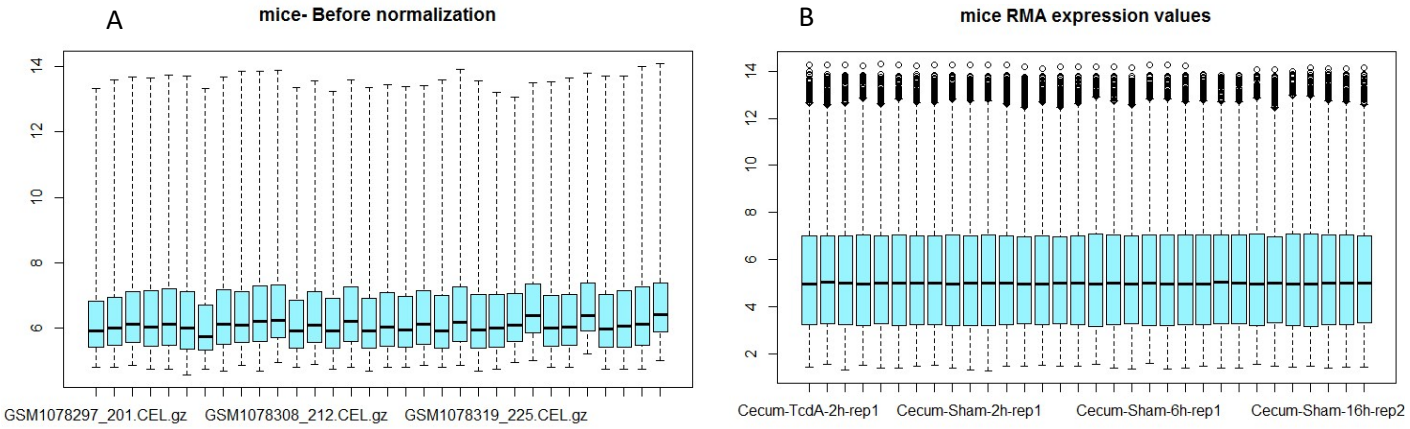


Figure 3. Box plots. In A, a representation of the original data without any RMA normalization. In B, RMA- normalized data.

4.1.1. Clustering of the data

Once the 32 mice samples were normalized, the following step was to see how similar they were among each other and how they cluster in groups. To do this, a Principal Component Analysis (PCA) was performed (Figure 4). The samples corresponded to 3 time-points of toxin exposure (2h, 6h and 16h) and 4 types of toxin samples (TcdA, TcdB, TcdAB and Sham). There were some replicates: 3 replicates for TcdA (at 2h, 6h, 16h), TcdB (at 2h, 6h,16h), TcdAB (at 2h) and Sham (at 2h, 6h); 1 replicate for TcdAB at 6h and 4 replicates for Sham at 16h. The absence of more replicates at TcdAB 6h and TcdAB 16h was due to death of the mice during the experiment.

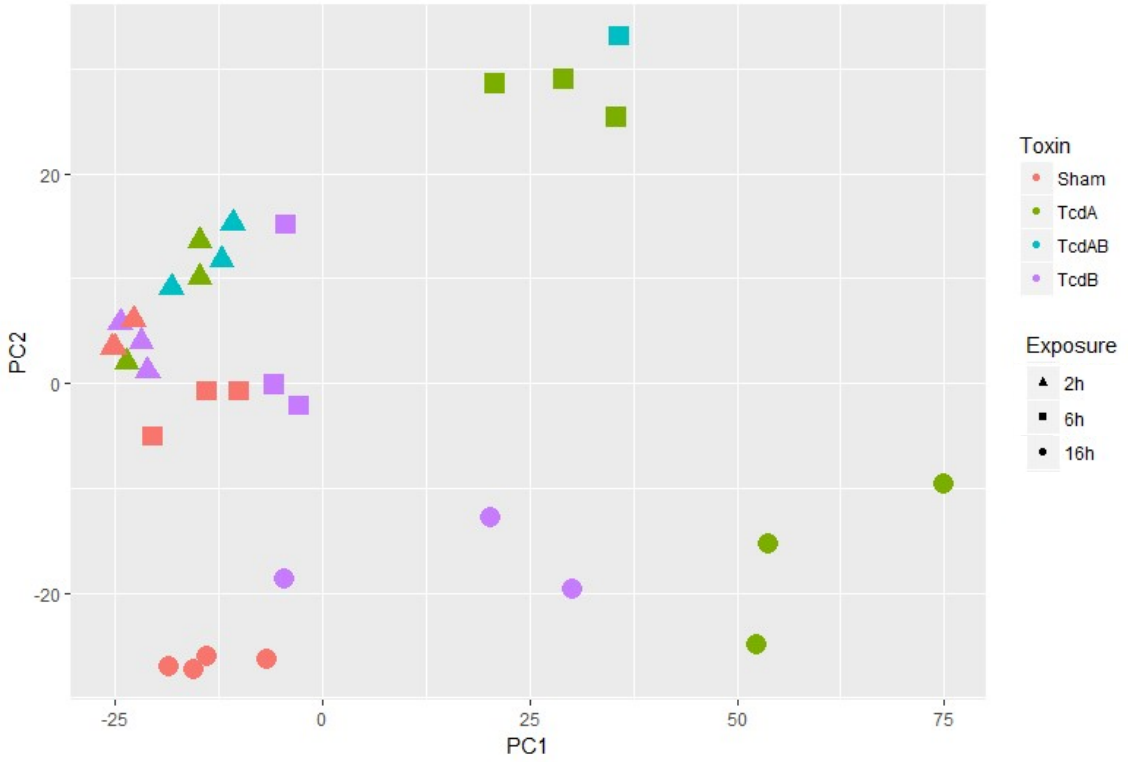


Figure 4. Principal Component Analysis of Mice samples. Each type of toxin is represented in a different colour and each time point of exposure to toxin is represented in a different shape.

From the PCA plot, 3 clusters of samples are observed. Samples corresponding to 2h data and Sham and TcdB at 6h are grouped in the middle left part of the plot. This means that in the first hours of toxin exposure the difference with control samples is small and not visible in the PCA plot. In the top of the plot, TcdA and TcdAB at 6h are together, showing a difference from the behaviour of TcdB and Sham at 6h. Regarding, 16h data, the samples are spread in the bottom part of the plot, with significant differences among Sham, TcdB and TcdA at 16h. Consequently, the higher the time exposed to toxins, the higher the difference between treated and control samples, and TcdA seems to have a more significant role than TcdB.

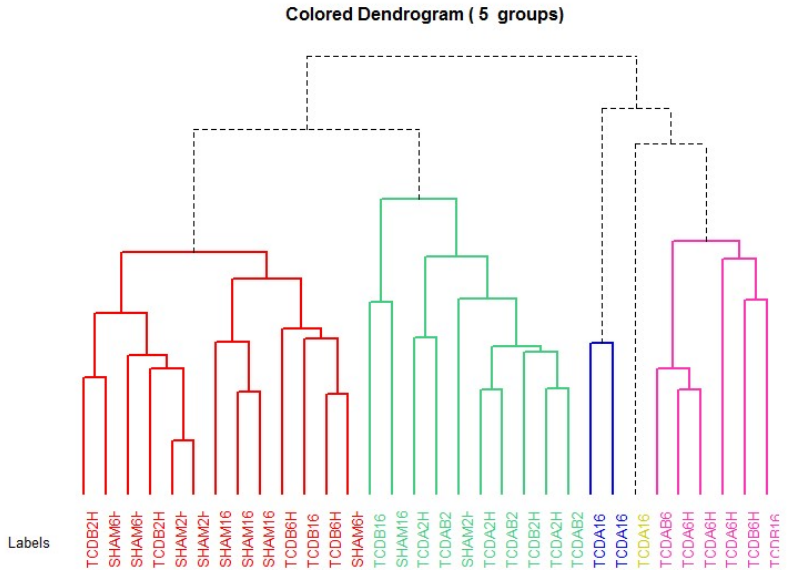


Figure 5. Dendogram representing a Hierarchical Clustering of mice samples.

The Hierarchical Clustering (Figure 5) represents similar results with the PCA plot. The first cluster from the left contains samples from TcdB and Sham at 2h, 6h and 16h. The second cluster (in green) represents 2h data (TcdA, TcdAB and 1 replicate of Sham 2h and TcdB 2h), plus one replicate from TcdB 16h and Sham 16h. The following 3 samples (in blue and yellow) are formed by TcdA 16h, followed by the last cluster of TcdA and TcdAB at 6h. As seen in the PCA plot, TcdA shows a greater difference from control samples than TcdB.

4.1.2. Representation of Differentially Expressed (DE) genes

During data analysis, the intensity values of the microarray probes, were converted into expression values of the genes with their corresponding p-value. Expression values were log2-converted to embrace more values within a range in expression. From these expression values, fold change was calculated, which is a ratio of the expression of a gene under toxin exposure divided by the expression in the control sample of the same gene. If the ratio is higher of 1.5 or lower of -1.5, the gene considered is upregulated or downregulated, respectively. Regarding the p-value, this is a statistical value which calculates the probability that a value is what it is by chance. The lower the p-value, the less probable it is a random value and thus, it is more significant to the results.

From this point of the analysis we have only considered the effects of TcdAB in mice at 2h and at 6h. To obtain the differentially expressed genes, we considered a fold change of 1.5 and a p-value lower than 0.01. These restrictions gave 267 DE genes in mice under TcdAB expression at 2h. DE genes at 2h are represented in a volcano plot (Figure 6).

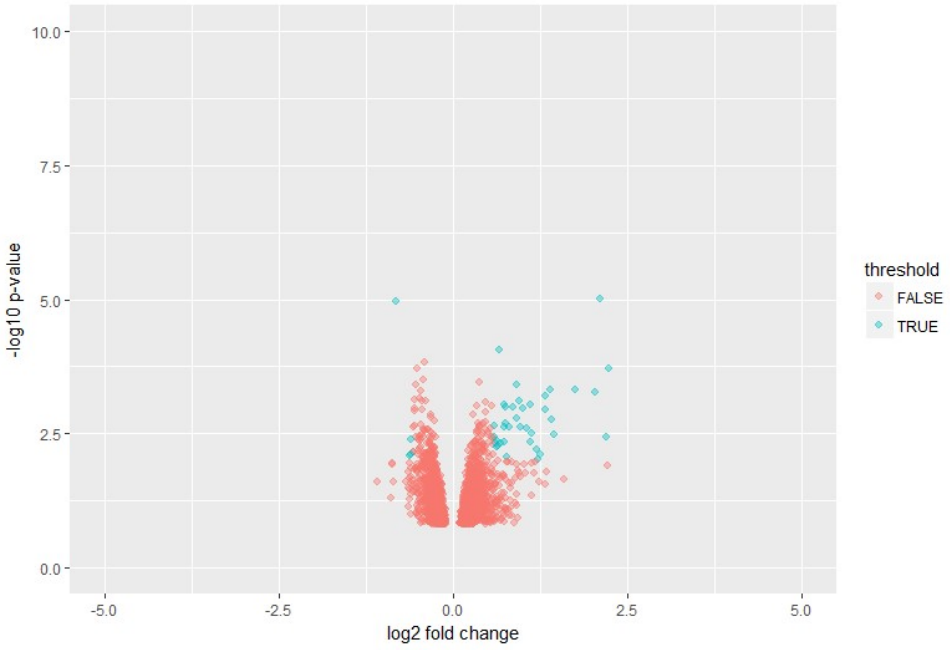


Figure 6. Volcano plot of DE genes under TcdAB infection in mice for 2h. In blue, DE genes with p-value < 0.01. Genes in the left part are downregulated, in the right part they are upregulated.

Besides the volcano plot, a heatmap was made with each toxin type and time point compared to its control sample. We clustered the 267 DE genes found in mice TcdAB 2h and we compared the expression of these 267 genes with the values in the other samples (Figure 7).

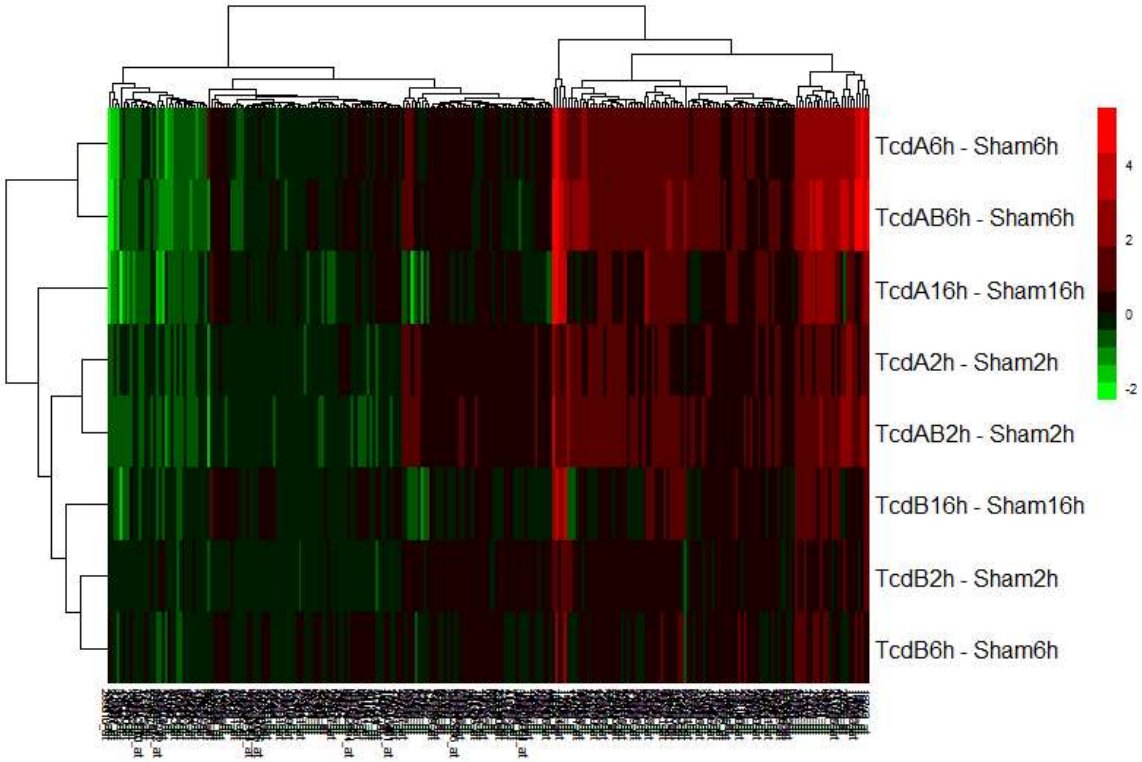


Figure 7. Heatmap with 267 DE genes (from TcdAB 2h with p-value<0.01) compared to the expression values of these genes in the other toxins and time points.

From the heatmap, there are visible differences among the samples. First, TcdB clusters together at the three time points (2h, 6h and 16h) and shows a lower degree of expression than both TcdA and TcdAB. TcdA and TcdAB at 2h have similar expression and it is lower than the expression of TcdAB at 6h and TcdA at 6h and 16h. These results are consistent with the PCA plot (Figure 4) and the dendogram (Figure 5). Again, we see that TcdA and TcdAB have a similar expression and it is higher than TcdB. In concrete, our focus is in TcdAB at 2h and at 6h. We expect a higher number of DE genes at 6h and thus, more significant changes in the pathways.

4.2. Pathways Analysis

At this point DE genes for mice under TcdAB infection for 2h are known. Thus, the following step is to relate these DE genes to their pathways and see in which way they affect the responses triggered by the host towards *C. difficile* toxins. To do so, Ingenuity Pathway Analysis (IPA) software was used. Indeed, we performed a Core Analysis, to identify pathways significant to our Mice TcdAB 2h dataset considering the expression log ratios of the genes (fold change). From this analysis, we checked the most significant canonical pathways and upstream regulators involved and their activation or inhibition state. The level of activity of the pathways is measure by the z-score, a statistical parameter that predicts if the pathway is activated or suppressed considering the expression ratios of their regulators.

The same procedure was followed to check to which pathways DE genes from TcdAB 6h and Caco2 cells (4.5h, 7days) correlate. Therefore, here we introduced DE genes in Caco2 cells in IPA to analyse the pathways to which these genes belong.

4.2.1. Number of Significant Pathways

The first step after the core analysis was to compare the number of canonical pathways with a p-value lower than 0.05 for the three datasets (TcdAB 2h, TcdAB 6h and Caco2). The number of significant pathways is represented in a Venn's diagram (Figure 8).

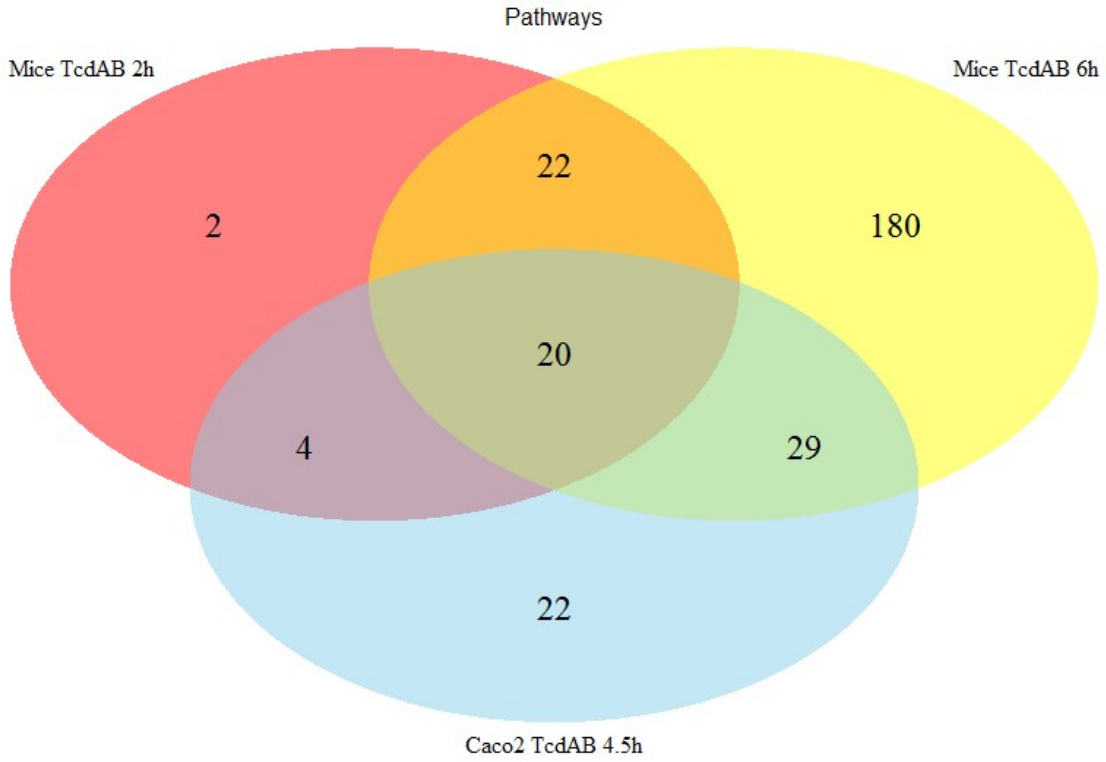


Figure 8. Venn's diagram with pathways which p-value < 0.05, for TcdAB 2h, TcdAB 6h and Caco2.

From Venn's diagram, the trend observed is that the number of significant pathways increases with time (less number in TcdAB 2h, followed by Caco2 4.5h and Mice TcdAB 6h). Moreover, almost all pathways in TcdAB 2h are also remarkable in TcdAB 6h (42 out of 48 of TcdAB 2h pathways). Among the three datasets, there are 20 pathways that are significant among all three. Another interesting fact is that there are 4 pathways in TcdAB 2h and Caco2 4.5h, but not

in Mice TcdAB 6h, which could be due to a decrease in their activity at later time points. It is noteworthy to mention that the pathways here represented have a p-value lower than 0.05, but their z-score (activation score) was not checked yet. Thus, if some pathways have a z-score of 0 (no change in activity) or their activation score is not known, they will not be considered for our further analysis.

4.2.2. Activation State of the Pathways

Pathways with p-value lower than 0.05 were plotted in a heatmap (Figure 9). The aim of this representation was to observe the direction of expression (activation or suppression) of the pathways for each dataset.

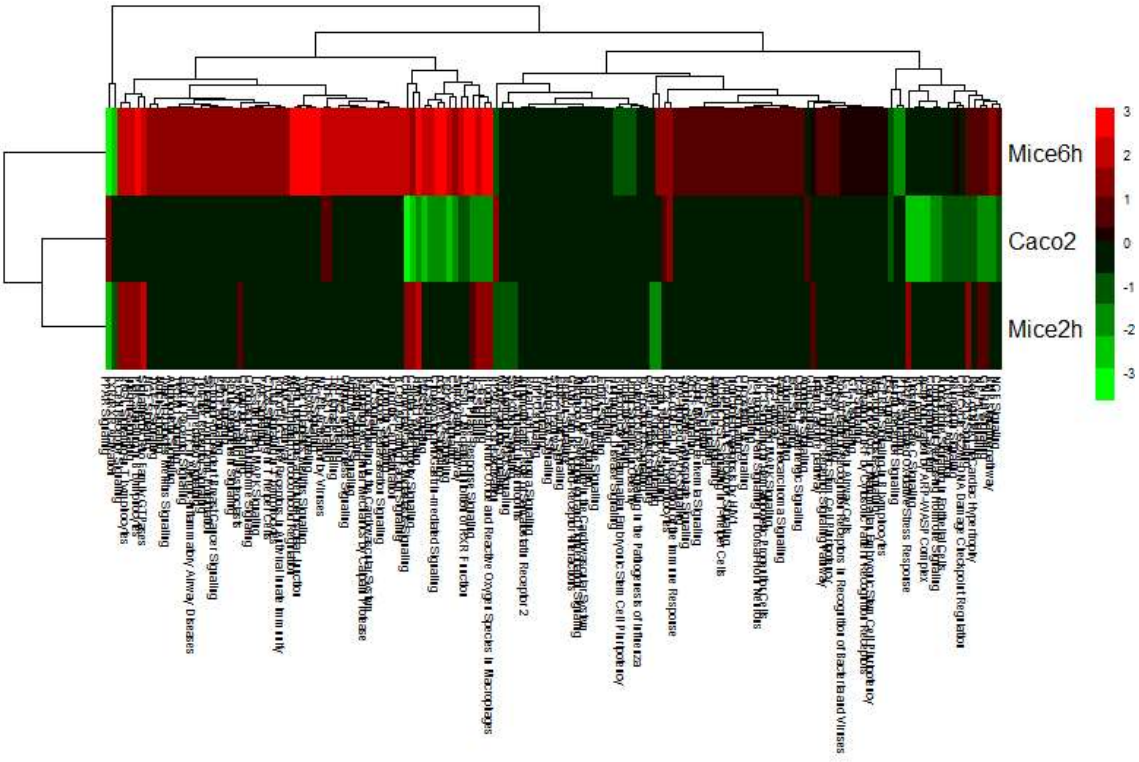


Figure 9. Heatmap with pathways which p-value<0.05 for TcdAB 2h, TcdAB 6h and Caco2.

In the heatmap, Mice TcdAB 6h is the dataset with higher number of significant pathways, as seen in Venn's diagram (Figure 8). Most of them are overactivated (in red). On the contrary, Mice TcdAB 2h shows the lowest number of relevant pathways, also as appeared in Venn's diagram (Figure 8), but still there are more overactivated than suppressed pathways. However, in Caco2, the majority of the pathways are suppressed and follow the opposite direction of expression than mice, in which the same pathways are activated. Moreover, some pathways appear in black in the three datasets, which means that the activation score (z-score) is 0 or unknown. These pathways will not be considered later on. The number of pathways activated and suppressed for each dataset are represented in table 1.

Table 1. Pathways with p-value < 0.05 and a negative (suppressed, in green) or positive (activated, in red) z-score for each dataset.

Num Pathways	Mice 2h	Mice 6h	Caco2 4.5h
Negative z-score	7	41	**80**
Positive z-score	14	**123**	15

At this point the number of significant pathways and their direction of activation was known. The next step was to analyse the name of the pathways and their z-scores, to check if there are pathways that follow the same direction of expression for the three datasets (TcdAB 2h, TcdAB 6h and Caco2). We represented the z-scores in table 2, but we only mention the pathways with a z-score in at least 2 of the 3 datasets. Pathways with a z-score in only one of the samples were removed as no comparison with the pathways from the other samples was possible.

Table 2. Pathways with their z-scores in TcdAB 2h, TcdAB 6h and Caco2 are represented. Highlighted in yellow are the pathways which follow the same direction of expression (same sign of the z-score) in Caco2 and Mice.

Significant pathways after TcdAB infection	**z- score Caco2**	**z- score Mice 2h**	**z- score Mice 6h**
Colorectal Cancer Metastasis Signaling	-3.5	1	1.732
ERK/MAPK Signaling	-2.887	0	2.111
ERK5 Signaling	-2.333	0	2.183
NRF2-mediated Oxidative Stress Response	-2.324	1.342	0
Cardiac Hypertrophy Signaling	-2.324	1.633	1.298
IL-6 Signaling	-2.111	1.342	2.777
HMGB1 Signaling	-2.111	2.236	2.401
ILK Signaling	-2.111	0	2.343
CXCR4 Signaling	-2.111	0	1.512
Cholecystokinin/Gastrin-mediated Signaling	-1.941	0	2.711
p38 MAPK Signaling	-1.941	-0.447	2.558
IL-8 Signaling	-1.732	1	2
NF-ÎºB Signaling	-1.732	0.447	0.714
Production of Nitric Oxide and Reactive Oxygen Species in Macrophages	-1.732	1.633	2.598
BMP signaling pathway	-1.633	0	1.069
GNRH Signaling	-1.508	0.447	0.756
Acute Phase Response Signaling	-1.508	0.816	3.042
Neuregulin Signaling	-1.414	0	0.243
p53 Signaling	-1.414	0	-1.279
Sumoylation Pathway	-1.414	0	2.236
NGF Signaling	-1.414	0	0.87
GÎ±12/13 Signaling	-1.265	1	0.962
Role of NFAT in Cardiac Hypertrophy	-1.155	0	0.707
LPS/IL-1 Mediated Inhibition of RXR Function	-0.905	0	3
HGF Signaling	-0.707	0	1.512
TGF-Î² Signaling	-0.378	0	1.342
IL-1 Signaling	-0.378	0	2.065
NF-ÎºB Activation by Viruses	0.378	0	2.043

TNFR1 Signaling	0.447	0	2.132
Death Receptor Signaling	0.632	0	1.095
CD27 Signaling in Lymphocytes	1	0	1.291
PPAR Signaling	1.414	-2.236	-3.651
PPARÎ±/RXRÎ± Activation	1.508	-1	-1.183
Protein Kinase A Signaling	0	-1.633	1.05
Neurotrophin/TRK Signaling	0	-1	-0.426
LXR/RXR Activation	0	-1	-2.837
SAPK/JNK Signaling	0	-1	-0.447
RANK Signaling in Osteoclasts	0	0.447	1.569
GÎ±q Signaling	0	0.447	0.18
PKCÎ¸ Signaling in T Lymphocytes	0	1	2.117
Integrin Signaling	0	1	2.082
PI3K Signaling in B Lymphocytes	0	1.342	2.921
B Cell Receptor Signaling	0	1.342	1.769
Signaling by Rho Family GTPases	0	2	2.271

There are 44 pathways with a z-score in more than one dataset. From them, only 6 have the same direction of expression (sign of the z-score) in Caco2 and Mice under TcdAB infection (Table 2, in yellow). This is consistent with the heatmap (Figure 9), in which in Caco2 the majority of the pathways were suppressed as opposed to mice. Moreover, most of the pathways are related to response to stimuli and stress activities. This makes sense as both Mice and Caco2 cells are dealing with bacterial toxin infection, so they need to activate pathways which sense defence mechanism to overcome the infection.

After comparing the z-scores for Mice TcdAB 2h, Mice TcdAB 6h and Caco2, we looked into the regulators of each of the pathway. The objective was to check if the molecules implicated in the pathways were upregulated or downregulated, or both, and if there is a correlation between the expression of the regulators of a pathway in the three datasets. In Table 3, the most upregulated (in red) and downregulated (in green) molecules are mentioned for each pathway.

From Table 3, the over and under-expressed regulators for each pathway in TcdAB 2h, TcdAB 6h and Caco2 do not show a high degree of similarity. However, the most upregulated gene in Mice TcdAB 2h is c-Jun in almost all the pathways, while in Caco2 it is downregulated. C-Jun is a transcription factor that activates under stress stimuli. Their responsive genes mediate the response triggered by proinflammatory cytokines and other defence responses. A higher expression of c-Jun in Mice TcdAB 2h will lead to a higher degree of inflammation than in Caco2, in which this transcription factor is suppressed. In Mice TcdAB 6h, c-Jun is not overactivated any more, but instead, c-Fos is upregulated at this time point. C-Jun and c-Fos interact in the nucleus to form Ap-1 (Activator protein 1) transcription factor, which regulates gene expression under a variety of stimuli like cytokines, growth factors, stress and bacterial and viral infections. The decrease in expression of c-Jun could be to favour the expression of c-Fos and thus, promote Ap-1 transcription of genes.

Table 3. Representation of the most activated (in red) and suppressed (in green) molecules for each pathway considered in Table 2. Blank spaces refer to z-scores that are 0 for a dataset.

Significant Pathways after TcdAB infection	**DE molecules in Caco2**	**DE molecules in Mice 2h**	**DE molecules in Mice 6h**
Colorectal Cancer Metastasis Signaling	Frizzled, WNT, RAS, KRAS, TGFβ, TNFR, IL6R, RHO, c-JUN, iNOS, PTGER		TNFα, IL6, c-Fos, TLR, MLH1, RRβ, Gγ, PKA, PI3K, LRP, Frizzled
ERK/MAPK Signaling	CAS, RAS, Integrin, MKP, CREB		c-Fos, PKA, PI3K, PP1/PP2A, VRK2, BAD, CREB
ERK5 Signaling	LIF, RAS, Gα12/13, CREB, 14-3-3		c-Fos, FRA-1, 14-3-3, BAD, MEK5, CREB
NRF2-mediated Oxidative Stress Response	PERK, ERK, Jun, Small MAP, CBP/p300, HSP22/40/90	PERK, HSP22/40/90, ASK1	FRA1, PKC, PI3K, CAT, CBP/p300, HSP22/40/90
Cardiac Hypertrophy Signaling	c-Jun, TGFβ, IL6R, RAS, RHO		IL6, ADRα, Gγ, Gα, PI3K, AC, PKA, Gaq11
IL-6 Signaling	IL6R, TNFR, IL1R, c-Jun	SOCS3, C-Jun	TNFα, IL6, SOCS3, NF-IL6, c-Fos, CD14, IL1, c-Raf
HMGB1 Signaling	c-JUN, RHO, RAS, TNFR1, IL1R	c-JUN, TPA	c-Fos, TNFα, ICAM1, TPA, PI3K, HAT
ILK Signaling	Rac, TNFR, c-JUN, iNOS, BMP2, RHO		c-Fos, TNFα, TNFR, HIF1α, PP2, PI3K, BMP2
CXCR4 Signaling	c-JUN, RHO, RAS, AC, Gα		c-FOS, EGR1, Gγ, Gα, PI3K, PKC, AC
Cholecystokinin/Gastrin-mediated Signaling	c-JUN, RHO, RAS		c-Fos, TNFα, ICER, IL1, PKC, c-RAF
p38 MAPK Signaling	TGFβ, IL1R, TNFR/Fas, MKP1/5, MKK6	MSK1, MKP1/5, TNFR/Fas	TIFA, p38 MAPKβ, MKK3, TNFR/Fas, MAX, FADD, IL1
IL-8 Signaling	CXCL1, Rho, AP-1, pro-HB-EGF		pro-HB-EGF, ICAM-1, Gγ, PKC, PI3K, PLD, RAB11FIP2
NF-IkB Signaling	BMP2/4, RAS, IL1R/TLR, TNFR		TNFα, A20, BMP2/4, BAFF, PI3K, IL1, IL1R/TLR, PKCζ, IKK, FADD
Production of Nitric Oxide and Reactive Oxygen Species in Macrophages	TNFR, Rho, AP1, iNOS	Rho, TNFR, AP1, MEKK	TNFα, TNFR, PPARα, CAT, PKC, PP1, PP2a/PP2a, PI3K
BMP signaling pathway	BMP, Ras, JUN	JUN, PITX2	PITX2, BMP, c-Raf, CBP, PKA
GNRH Signaling	EGR1, c-Jun	c-Jun	EGR-1, c-Fos, Gaq/11, PKC
Acute Phase Response Signaling	c-Jun, FGA, FGB, IL6R, IL1R, TNFR, Ras	Ask, SOCS3, c-Jun	TNFα, IL6, c-Fos, HP, NF-IL6, SOCS3, IKBα, IL1, ECSIT, AMBP, GCR, c-Raf
Neuregulin Signaling		EGFR ligand, ERBB4 ligand, GRB7, Ralt, Ras	Ralt, ERBB2, PI3K p85, PKC, Erbin
p53 Signaling	c-Jun, p300, Teap, NOXA		HIF1A, TRIM29, Apaf1, Teap, SIRT, p300, PI3K
Sumoylation Pathway	c-Myb, RhoGDI, Rho, Ap1	MDM2, Ask1	IKBα, c-Myb, Sirt1, GR

NGF Signaling	TRAF4, PKCδ, ERK5	NGF, MEKK	PI3K, PKCζ, TrkA
GÎ±12/13 Signaling	IKK, PAR, RAS, C-JUN		IKB, LPAR, VAV, PI3K
Role of NFAT in Cardiac Hypertrophy	AKAP5, TGFβ, RAS, LIF		IL6, Na^{+}/Ca^{2+} exchanger, PKC, Gγ, PI3K, HDAC
LPS/IL-1 Mediated Inhibition of RXR Function	CYP2A, c-Jun, TNFR		TNFα, CD14, PPAR, PGC-1α, PGC-1β, LRH-1, CAT, RMO, FATP
HGF Signaling	c-Fos, c-Jun	c-Jun, c-Fos, MEKK	IL6, c-Fos, PKC, PI3K
TGF-β Signaling	c-Jun, AP-1, BMR2/4/7, TGFβ		TMEPAI, BMR2/4/7, Type I Receptor, Type II Receptor, VDR, IRF7, PITX2
IL-1 Signaling	c-Jun, IL1R1, Gα, AC		c-Fos, Gγ, Gα, AC, PKA
NF-IkB Activation by Viruses	IKK, MEKK1, RAS, PKC		IkB, NF-kB, PI3K, PKC, IKK, c-RAF
TNFR1 Signaling	TNFR1, c-Jun	c-Jun, A20	TNFα, A20, APAF, c-Fos, RAIDD, FADD, IKK, cIAP
Death Receptor Signaling	TL1, TNFR1		TNFα, TNFR1, TNFR2, APAF, FADD, RAIDD, PARP
CD27 Signaling in Lymphocytes	c-Jun	c-JUN, MEKK	c-Fos, APAF
PPAR Signaling	c-JUN, RAS, TNFR1, IL1, NCOA, NR0B2	c-JUN	TNFα, c-Fos, PPARα, NCOA
PPARα/RXRα Activation	MKK3/6, IKK, MEF2C, c-JUN, IGFβ, AC, RAS, IL1R, NR0B2, BCL3		BCL3, IL6, IkBα, PPARα, PGC-1, Gαq
Protein Kinase A Signaling		PTP, PDE, PYG	TGFBR2, SMAD3, CREM, 14-3-3, Troponin I, Adducin, MLCK, PP1, PKC, PDE, PKA, AC, Hh, Gγ, PLC
Neurotrophin/TRK Signaling		c-JUN, ASK1, Neurotrophin	c-FOS, PI3K
LXR/RXR Activation		LDLR, IDOL	TNFα, CD14, IL6, TLR3, IL1, ACC
SAPK/JNK Signaling		c-JUN, DUSP8, ASK1	GADD45, DUSP8, PI3K, Gβγ, GCKs, FADD
RANK Signaling in Osteoclasts		c-Jun, MEKK, c-Cbl	c-Fos, IKK, PI3K, c-Raf
Gαq Signaling		PLD	Gαq, Gγ, IKK, PI3K, c-Raf, PKC, PLD
PKCÎ¸ Signaling in T Lymphocytes		c-Jun, MEKK	c-FOS, VAV, PI3K, IKK
Integrin Signaling		AND-34	ACK, ZYX, TSPAN, PI3K, MLCK, RAP, c-RAF
PI3K Signaling in B Lymphocytes		c-JUN, CBL, BLK	C3, IL4R, c-FOS, VAV, aPKC, IKK.
B Cell Receptor Signaling		c-Jun, MEKK	EGR1, PI3K, VAV, RAP
Signaling by Rho Family GTPases		c-Jun	c-Fos, Gα, Gγ, MLCK, PAR6, aPKC, PI3K

4.2.3. Analysis of the Pathways

At this point we analysed the information of all pathways present in TcdAB 2h, TcdAB 6h and Caco2 in a broad sense. We know the number and the name of the pathways implicated at each time point and the most DE regulators. Therefore, the next step is to look thoroughly into the most remarkable pathways for each dataset. To do this, we are going to separate the analysis in two parts:

- First part: Compare the level of activation in the pathways between Mice at 2h and Mice at 6h under TcdAB infection.

- Second part: Compare the level of activation in the pathways between Mice (2h and 6h) and Caco2 cells under TcdAB infection.

The objectives of dividing the analysis in two parts are to determine differences in time of the activation of the regulators in the pathways (Mice data at 2h and at 6h, first part), and to cross-compare the effect of the toxins in samples from different model systems (mice and Caco2, second part).

4.2.3.1. Comparison of Mice Datasets under Toxins' Effect

In order to compare the activation of the pathways between Mice TcdAB 2h and Mice TcdAB 6h, we have to consider the pathways which are activated or suppressed in both datasets (absolute z-score > 0). Considering Table 2, in which there are 44 pathways that contain a significant z-score in two of the three samples, there are 24 pathways with z-score in Mice TcdAB 2h and Mice TcdAB 6h. They are plotted in a heatmap (Figure 10) to observe the direction of activation.

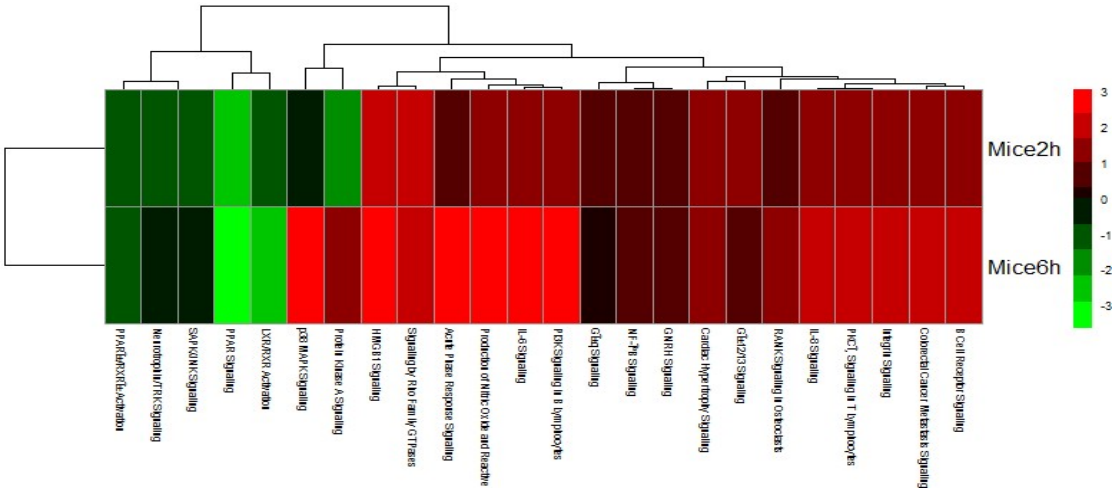


Figure 10. Heatmap with pathways which p-value<0.05 and |z-score| > 0 for TcdAB 2h and TcdAB 6h in mice.

The trend visualized in the heatmap is that in Mice under TcdAB infection for 6h, the activation or the suppression of the pathways is more remarkable than at 2h. Thus, although the pathways plotted are significant in both datasets, we expect a higher effect in the regulators of the pathways at 6h than at 2h. This is consistent with the heatmap in Figure 9.

Regarding we analysed all of the 24 pathways in both mice 2h and 6h, we are going to limit the explanation to the most significant pathways (Figure 11).

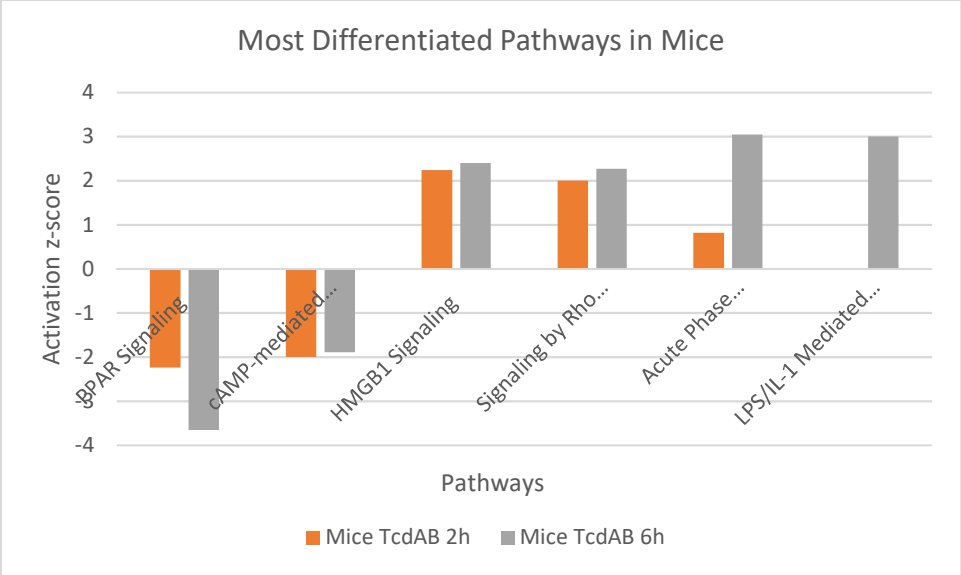


Figure 11. Graph of the pathways with higher z-score (in absolute value) in Mice TcdAB 2h and Mice TcdAB 6h vs their activation z-score.

From the figure, we can conclude that the suppressed pathways ("PPAR Signalling" and "cAMP-mediated Signaling") are similar in Mice TcdAB at 2h and at 6h, but the most activated pathways differ. In fact, the most activated pathways in Mice TcdAB 2h ("HMGB1 Signaling" and "Signaling by Rho Family GTPases") retain almost the same level of expression in Mice TcdAB 6h, but the most activated in Mice 6h ("Acute Phase Response Signalling" and "LPS/IL-1 Mediated Inhibition of RXR Function") were fairly or not activated at 2h.

"PPAR Signalling" and "Signaling by Rho Family GTPases" will be explained in the cross comparison between mice and Caco2 data. Here we will discuss the regulation of "HMGB1 Signaling" and "Acute Phase Response Signalling", the most activated pathways in Mice TcdAB 2h and Mice TcdAB 6h, respectively. We decided to highlight these pathways because of their connection to other pathways that are directly involved in the cytopathic and cytotoxic effects mediated by the toxins.

4.2.3.1.1. HMGB1 Signalling

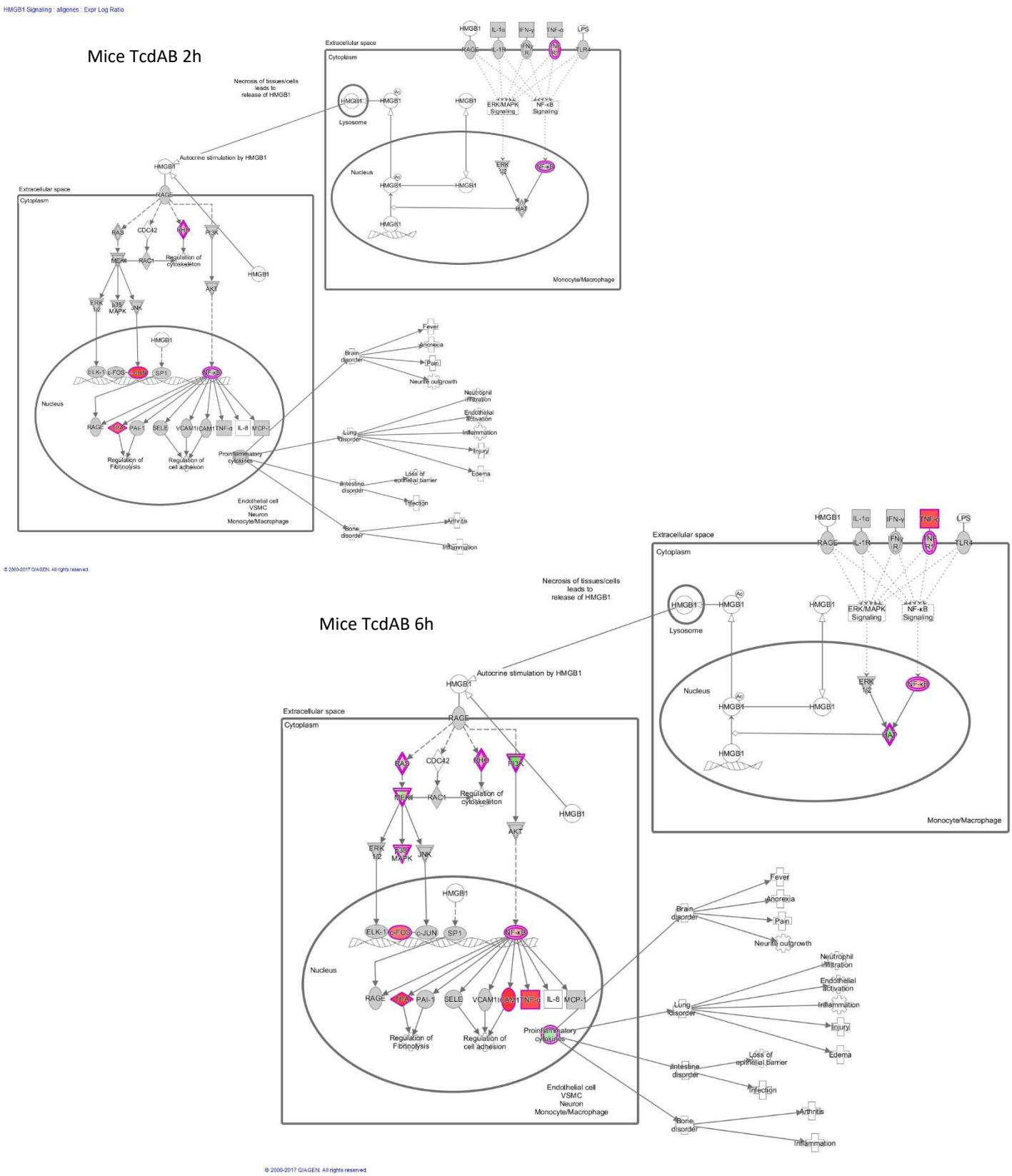


Figure 12. HMHB1 Signalling Pathway in Mice TcdAB 2h (above) and Mice TcdAB 6h (below). Upregulated molecules are coloured in red, suppressed molecules in green.

High Mobility Group-B1 (HMGB1) is a DNA binding protein involved in the regulation of nucleosome structure and gene transcription. Its secretion occurs after stimulation with endotoxins (by macrophages, monocytes, and pituicytes). HMGB1 mediates strong, direct bactericidal effects, so as a result, this is one of the first pathways to become activated after a bacterial invasion (Bianchi & Agresti, 2005).

When IL-1, TNF-α, IFN-γ, LPS or HMGB1 ligands bind their receptors (IL-1R, TNFR1, TLR4, IFN-γR and RAGE respectively) NF-κB and MAPK pathways activate. Activated MAPKs and NF-KB enter the nucleus and switch on histone acetylases (HATs) or inhibition of deacetylases. This triggers HMGB1 acetylation.

HMGB1 can activate cell surface receptors on various cell types through ERK1/2 and JNK and p38; PI3K and Akt; the transcription factors NF-κB and Sp1 as well as Rac1 and CDC42. These activations result in the release of a proinflammatory cytokines and chemokines: TNF-α, IL-1α, IL-1β, IL-1Ra, IL-6, IL-8 and MCP1, upregulation of adhesion molecules: ICAM1 and VCAM1, RAGE, and HMGB1 itself (Klune & Dhupar, 2008).

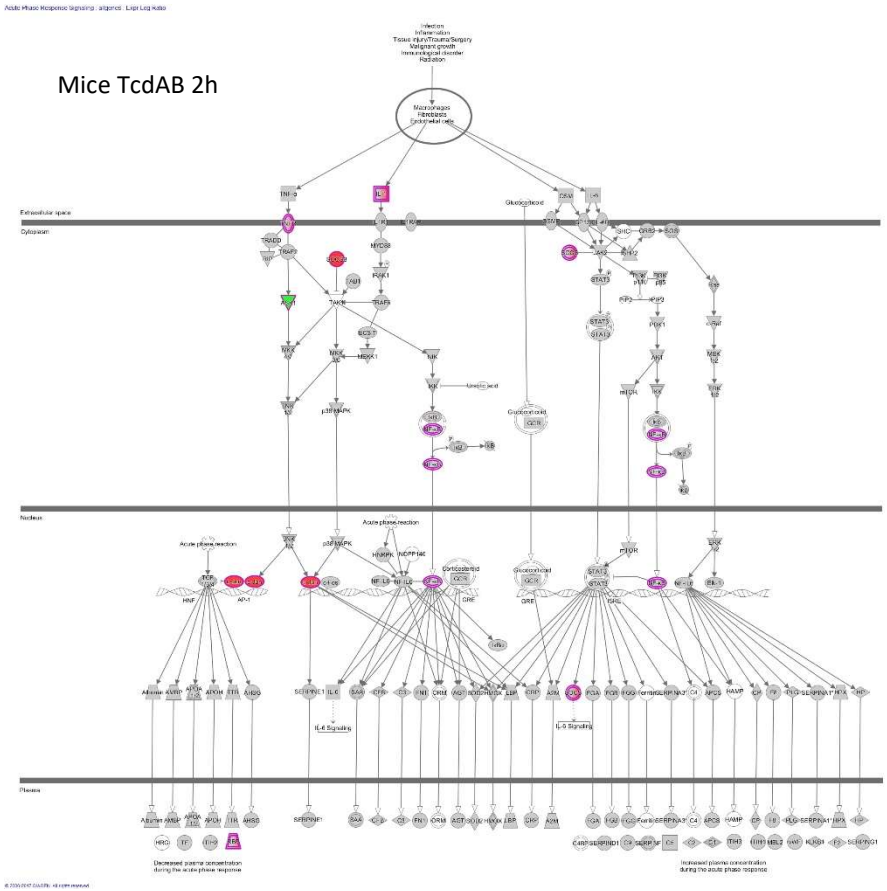
HMGB1 signals to the cell motility system by activating CDC42, Rac1 and Rho. At the lamellopodia, which is the area of actin microfilament formation in the plasma membrane, HMGB1 interacts with the extracellular matrix and membrane receptors, affecting cell motility and metastasis.

Regarding HMGB1 pathway after 2h and 6h of TcdAB infection in mice (Figure 12), the first molecules to become activated in Mice TcdAB 2h are NF-KB, c-JUN and TPA. NF-KB and c-JUN are both transcription factors, so their responsive genes appear upregulated in Mice TcdAB 6h. Thus, TNF-α, TPA, ICAM1 and c-Fos are upregulated at 6h. TNF-α is a proinflammatory cytokine that increase their expression in response to NF-KB. TPA, also upregulated at 2h, is responsible of dissolving blood clots, that might appear when a cluster of bacteria forms. ICAM1 is an adhesion molecule, so its upregulation is involved in cell-to-cell contacts, that are disrupted by bacterial toxins. Regarding c-Fos, it becomes upregulated when c-JUN returns to basal levels. In fact, this happens because c-Fos is an NF-KB responsive gene, so its expression is delayed from c-Jun's expression. Both transcription factors are needed to activate AP1 transcription factor. Thus, their expression must be regulated to avoid excessive expression of either c-Jun or c-Fos without the expression of the other.

At 6h, HAT, PI3K and proinflammatory cytokines start to be suppressed. This might be a negative feedback mechanism, to avoid excessive HBGB1 signalling once the bacterial invasion is sensed, and other pathways have already become activated.

4.2.3.1.2. Acute Phase Response Signalling

Mice TcdAB 2h



Mice TcdAB 6h

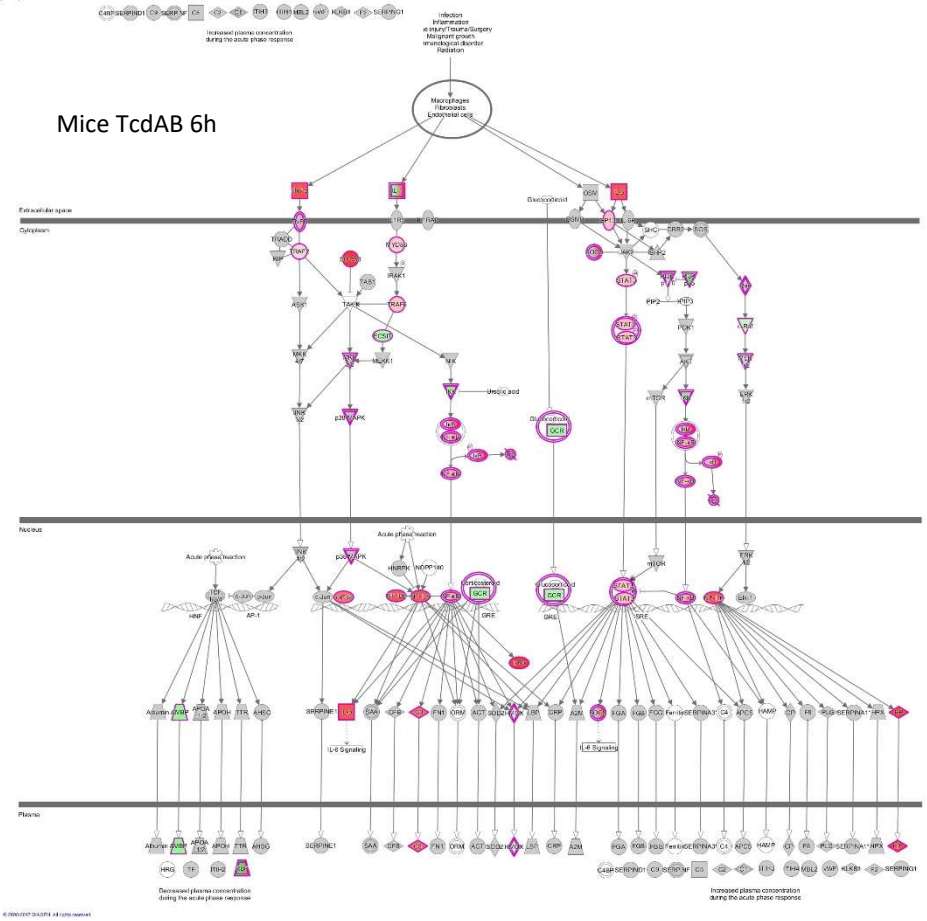


Figure 13. Acute Phase Response Signalling Pathway in Mice TcdAB 2h (above) and Mice TcdAB 6h (below). Upregulated molecules are coloured in red, suppressed molecules in green.

The acute phase response is a rapid inflammatory response that provides protection against microorganisms using non-specific defense mechanisms. There are proteins that activate the acute phase response (positive acute phase response proteins) and other that suppress this response (negative acute phase response). Therefore, negative acute phase response proteins are downregulated when this pathway is activated. Many of these proteins bind hormones, such as glucocorticoids. Thus, downregulation of these proteins increases the number of hormones available in plasma. This effect is visible in Mice TcdAB 6h, where GCR (glucocorticoid receptors) are downregulated. Moreover, GCR complex to the GRE site of the DNA to act as promoters. Their suppression leads to a lower expression of their responsive genes, which are involved in tight junction sealing (Balda et Matter, 2009). Thus, less GCR induces less TJs and increase epithelial permeability.

Positive acute phase response proteins become upregulated when acute phase response is elicited, but their higher levels in plasma are visible 4-5 hours after the stimulus. Thus, upregulated molecules are seen in Mice TcdAB 6h but not in Mice TcdAB 2h. Upregulated positive acute phase proteins remain in plasma for 24 up to 48 hours. They are important in the first stages of an infection because they opsonize and trap microorganisms, activate the complement, neutralize harmful enzymes and modulate the immune response (Gruys, et al. 2005).

At 2 hours, acute phase response starts with the upregulation of SOCS3 and c-JUN and the downregulation of ASK1. c-JUN appears to be the first overstimulated regulator. However, it binds the AP1 promoter in the DNA, which responsive genes are related to the negative acute phase response. Thus, SOCS3 and ASK1 might mediate in reducing c-JUN overactivation. ASK1 is a kinase involved in JNK pathway that activates c-JUN. Its downregulation reduces the signaling in this pathway, and therefore, the upregulation of c-JUN. Parallelly, SOCS3 becomes upregulated and inhibits TAK kinase, involved in JNK and p38 MAPK pathways. Both pathways are involved in the activation of c-JUN, so less TAK activation leads to a decrease in c-JUN. Moreover, c-JUN returns to its basal expression at 6h, but SOCS3 remains upregulated at that time.

At 6h, IL6 is upregulated, as a result of TNFα upregulation, that activates NF-IL6 transcription factor through NIK pathway. NF-IL6 is upregulated, promoting the expression of IL6. Furthermore, IL6 is the first activator of JAK/STAT pathway, thus enhancing the expression of positive acute phase proteins. Remarkably, glucocorticoids are able to release IL6 into the circulation. Overall, IL6 is an important mediator of the acute phase response.

4.2.3.2. Comparison of Mice and Caco2 Datasets under Toxins' Effect

After the analysis of the toxins' effect in the cecum of mice, we decided to cross-compare the results with pathways obtained in Caco2. Considering the pathways in Table 2, there are 33 out of 44 pathways with a z-score in Caco2 and in one of the samples in mice (either 2h, 6h or both).

We plot the three most suppressed and the three most activated pathways in Caco2 (lowest and highest z-score, respectively) and we compared them to the z-scores of these pathways in mice (Figure 14).

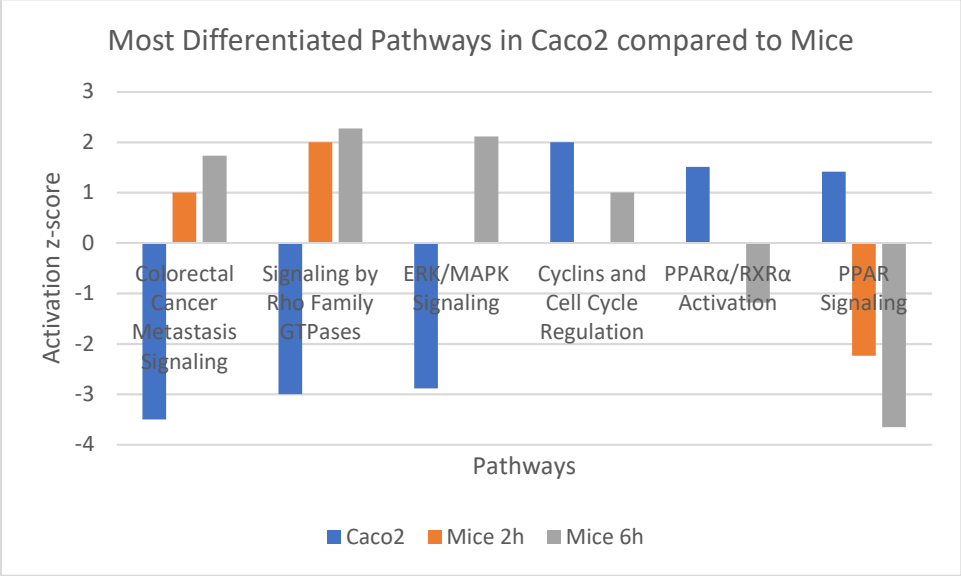
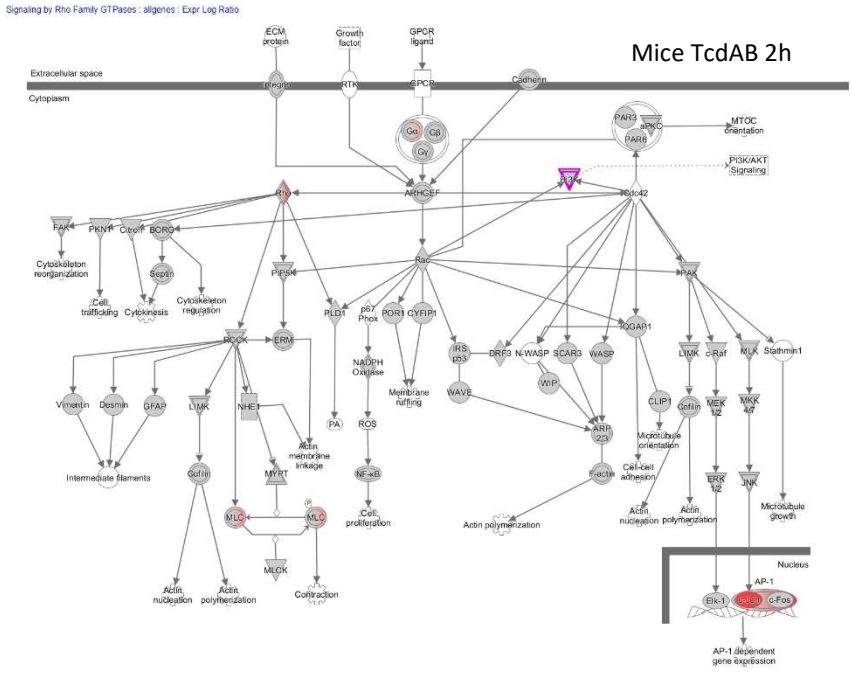
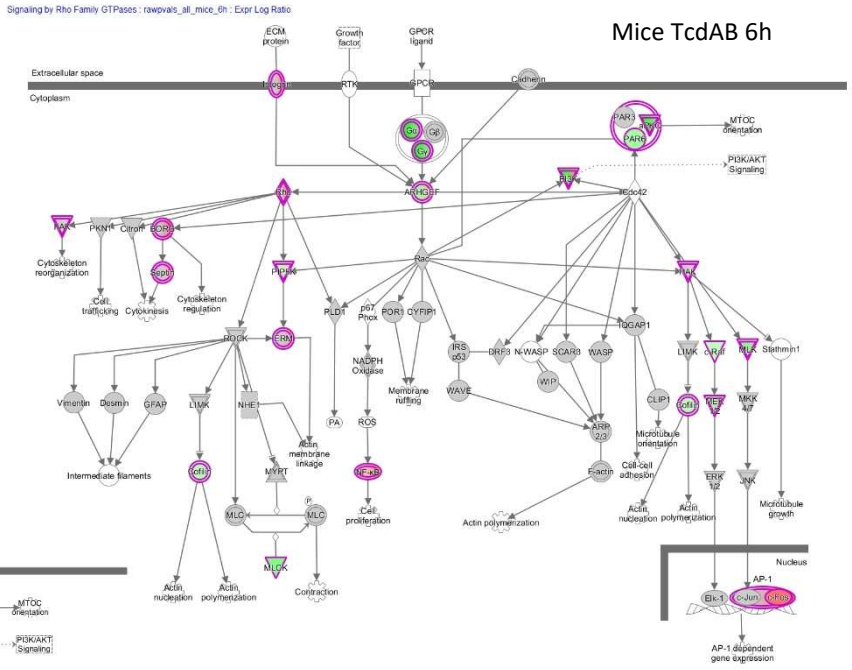


Figure 14. Graph of the most activated and suppressed pathways in Caco2 (higher and lower z-scores) vs their activation z-score in Caco2, Mice TcdAB 2h and Mice TcdAB 6h.

In the plot, the trend is that the most and the least activated pathways in Caco2 show an opposite direction of expression in mice. Moreover, half of them are not even activated in mice at 2h. There's only one pathway with the same direction of expression (Cyclins and Cell Cycle Regulation). Therefore, our attention will be focused in the pathways: "Signalling by Rho Family GTPases" and "PPAR Signalling". These pathways were among the most significant ones also in the mice comparison (Figure 11).

4.2.3.2.1. Signalling by Rho Family GTPases





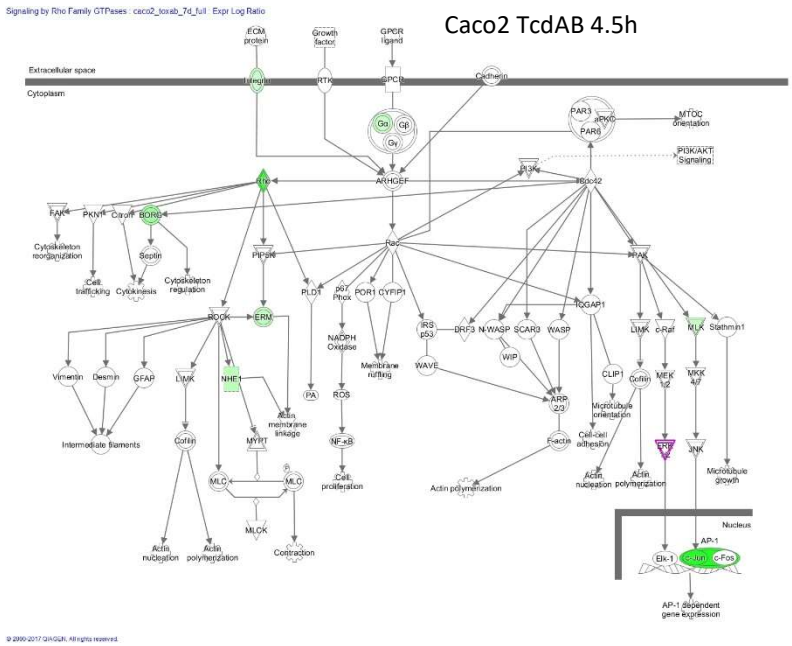


Figure 15. Signalling by Rho Family GTPases Pathway in Mice TcdAB 2h (above), Mice TcdAB 6h (middle) and Caco2 TcdAB 4.5h (below). Upregulated molecules are coloured in red, suppressed molecules in green.

Rho Family GTPases are small GTP-binding proteins that become activated by growth factors, cytokines, adhesion molecules, hormones and integrins. Signaling pathways that are regulated by these GTPase family play an important role in several pathological conditions, including cancer, inflammation, and bacterial infections (Etienne-Manneville & Hall, 2002). Rho signalling cascade is related to HMGB1 signalling, because Rho GTPases sense HMGB1 signals.

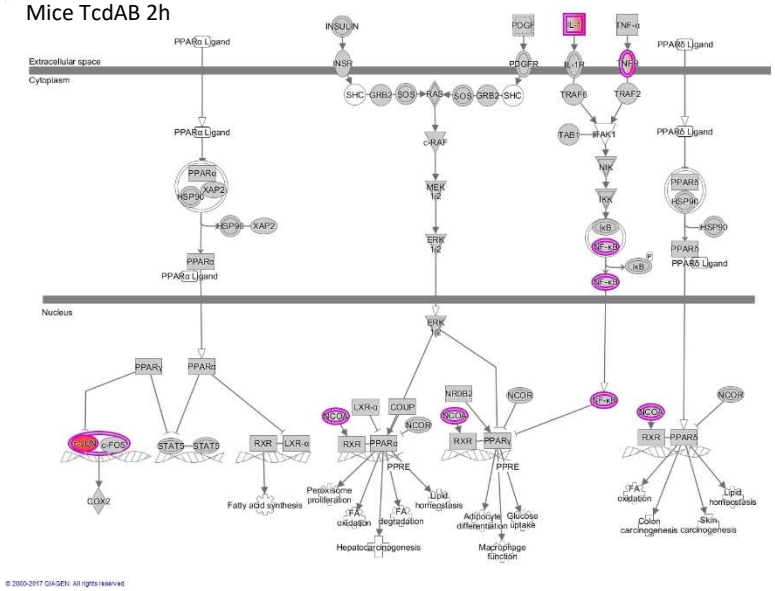
The importance of Rho GTPase signalling in the regulation of biological processes is remarkable, as they are involved in reorganization of the actin cytoskeleton, transcriptional regulation, vesicle trafficking, morphogenesis, neutrophil activation, phagocytosis, mitogenesis, apoptosis and tumorigenesis.

In mammals, the GTPase family currently consists of three subfamilies: Rho, Rac and Cdc42. Each one controls the formation of a distinct cytoskeletal element in mammalian cells. Most important in our research is Rho, that regulates bundling of actin filaments into stress fibers and the formation of focal adhesion complexes (Ridley, 2015).

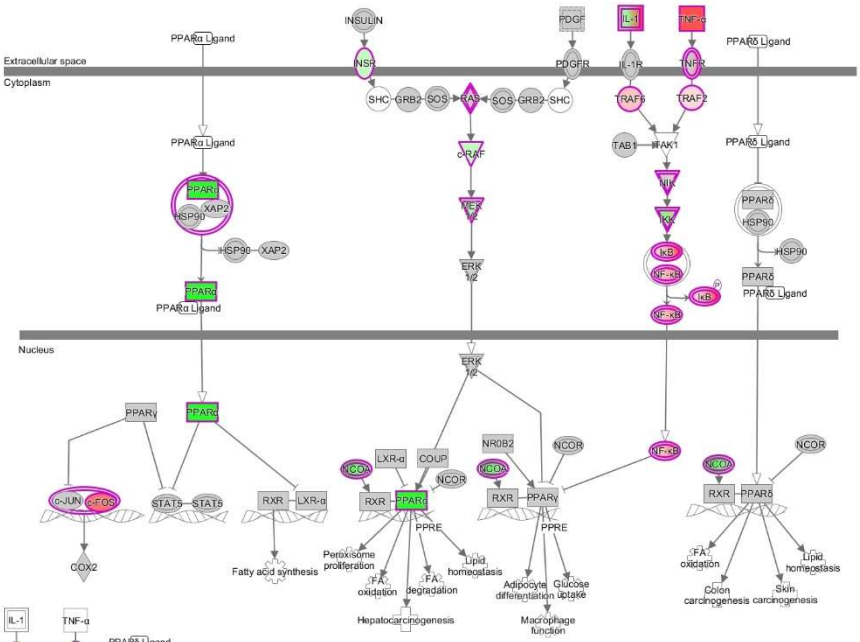
After infection with TcdAB toxins, Rho GTPases cascade changes its basal activation and expression. At 2h, the only molecule highly activated in the pathway is c-Jun, besides the positive z-score of the overall pathway. At 6h, some kinases are slightly suppressed, probably to decrease the transcription of Jun responsive genes and return to homeostasis (decrease inflammation) after the first stages of bacterial infection. One of these kinases suppressed is MLCK, a kinase that regulates tight junction permeability. Suppression of MLCK is related to the loosening of TJ in the intestinal epithelial barrier (Balda et Matter, 2009). In Caco2 the overall pathway is suppressed.

4.2.3.2.2. PPAR Signalling

Mice TcdAB 2h



Mice TcdAB 6h



Caco2 TcdAB 4.5h

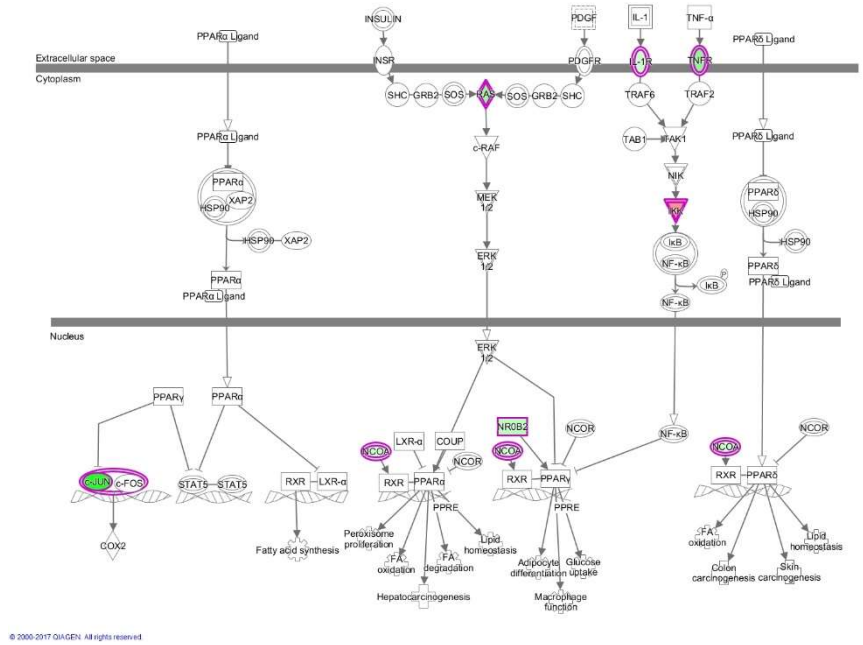


Figure 16. PPAR Signalling Pathway in Mice TcdAB 2h (above), Mice TcdAB 6h (middle) and Caco2 TcdAB 4.5h (below). Upregulated molecules are coloured in red, suppressed molecules in green.

PPARs (Peroxisome Proliferator-Activated Receptors) are nuclear receptor proteins that act as transcription factors to regulate metabolic pathways. They induce genes that affect fatty acid metabolism, peroxisome proliferation, colon and hepatocarcinogenesis. PPAR function is supressed by cytokines, growth factors and insulin (Michalik et al., 2006).

There are three types of PPAR receptors. PPARα blocks STAT5 transcription factor and LXRα. LXRα blockage inhibits fatty acid biosynthesis but favours fatty acids degradation. This receptor is activated by MAPK. PPARγ inhibits STAT5, c-Jun and c-Fos and its inhibited itself by MAPK and NF-KB stimulated by TNFα and IL1. PPARδ stimulates fatty acids oxidation.

At 2h in mice, there is an increasing IL1 and TNFR expression, which then leads to a higher expression in Mice 6h, that increases the level of suppression of the pathway. Later, at 6h, expression of TNFα inhibits PPARα. This suppression is not visible in Mice TcdAB 2h nor Caco2, as TNFα is not overexpressed. Also in mice at TcdAB 6h, NCOA is suppressed, less RXR activation and thus, less PPARα activation. Suppression of PPARα leads to a decreased fatty acids biosynthesis and oxidation.

5. CONCLUSIONS

The effect of *Clostridium difficile* toxins was analysed using a systems approach. Comparing the effects of TcdAB at 2h and at 6h in mice cecum did not have the same influence than in Caco2 colon-like cells after 4.5 hours of toxin infection. This approach is informative to analyse the genes and pathways most affected by *Clostridium* toxins. We observed that most of the pathways that popped-up in the analysis contained shared genes which activation or expression were affected. This was the case of some transcription factors (c-Jun, c-Fos, NF-kB), some kinases (Rho, Ras, PI3K, PKA) and target genes (inflammatory cytokines as TNFα, IL-1). Thus, these pathways should not only be considered as a whole, in fact, we should look inside their regulators and their upstream and downstream effects.

Differences in the activation of mice and Caco2 pathways were in part, due to the intrinsic differences in both model systems. Whereas in mice we were considering cecum tissue *in vivo*, in human epithelial Caco2 cell line we were considering colon-like cells cultured *in vitro* for 7 days. Although *in vitro* and *in vivo* experiments show different results, it is significant that we still saw some overlap in the pathways affected (e.g. Rho GTPases Signalling, PPAR Signalling, etc.) and some differentially expressed genes (c-Jun, NF-kB, ETC.).

In order to improve the research, we could have compared mice cecum DE genes with mice intestinal enterocytes (*in vitro* system). This way, we could remove the effect of other cells in mice cecum tissue expression. Indeed, it is probable that our results not only had intestinal epithelial cells from mice but also other type of cells, like immune cells, that could affect the degree of activation of the pathways.

Considering the significant pathways obtained, Rho GTPases Signalling is the most affected by the *Clostridium difficile* toxins infection. Upstream of this pathway, HMGB1 signalling sensed and sent signals to the GTPases that elicit actin cytoskeleton disruption and eventually, tight junction disruption and changes in cell morphology. Therefore, Tight Junction Signalling pathway was also connected to Rho GTPases, and their disruption lead to impairment of the colon epithelia, promoting diarrhoea and inflammation of the colon. Regarding PPAR and Acute Phase Response Signalling, these are pathway that affect the overall tissue by promoting gene expression of metabolic pathways or inflammation, respectively. Indeed, this was found also in literature (Janvilisri et al., 2010), in which it was explained that TcdA and TcdB enter the cells through receptor-mediated endocytosis, inactivate Rho GTPases, what leads to disaggregation of the actin cytoskeleton and intestinal epithelial cell damage, what eventually increases permeability of tight junctions.

Due to the broad and diverse effect of the toxins in the systems, it is difficult to find a specific pathway or regulator to target for the regulation of TcdAB infection. Therefore, more research is needed insight the molecular effects of the toxins in the organism. These data could serve to identify diagnostic markers or targets to overcome host responses to *Clostridium difficile* toxins.

REFERENCES

AKEN, B. L., AYLING, S., BARRELL, D., CLARKE, L., CURWEN, V., FAIRLEY, S., ... & HOWE, K. (2016). The Ensembl gene annotation system. *Database*, *2016*, baw093.

BALDA, M. S., & MATTER, K. (2009). Tight junctions and the regulation of gene expression. *Biochimica Et Biophysica Acta (BBA)-Biomembranes*, *1788*(4), 761-767.

BARRETT T, WILHITE SE, LEDOUX P, EVANGELISTA C, KIM IF, TOMASHEVSKY M, MARSHALL KA, PHILLIPPY KH, SHERMAN PM, HOLKO M, YEFANOV A, LEE H, ZHANG N, ROBERTSON CL, SEROVA N, DAVIS S, SOBOLEVA A. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res. 2013 Jan;41(Database issue): D991-5.)

BELL, G., (2007). Analysis of Microarray Data. WIBR Microarray Course, Whitehead Institute, 97 pp.

BIANCHI, M. E., & AGRESTI, A. (2005). HMG proteins: Dynamic players in gene regulation and differentiation. *Current Opinion in Genetics and Development*, *15*(5 SPEC. ISS.)

BOUILLAUT, L., DUBOIS, T., SONENSHEIN, A. L., & DUPUY, B. (2015). Integration of metabolism and virulence in Clostridium difficile. *Research in microbiology*, *166*(4), 375-383.

BRAINARRAY, 2008. Microarray Lab, University of Michigan, viewed 1st April 2017. http://brainarray.mbni.med.umich.edu/brainarray/AboutUs/AboutUs.asp

BRAUN, V., HUNDSBERGER, T., LEUKEL, P., SAUERBORN, M., & VON EICHEL-STREIBER, C. (1996). Definition of the single integration site of the pathogenicity locus in Clostridium difficile. *Gene*, *181*(1), 29-38.

BURNS, D. A., HEAP, J. T., & MINTON, N. P. (2010). Clostridium difficile spore germination: an update. *Research in microbiology*, *161*(9), 730-734.

CHEN, S., SUN, C., WANG, H., & WANG, J. (2015). The role of Rho GTPases in toxicity of Clostridium difficile toxins. *Toxins*, *7*(12), 5254-5267.

D'AURIA, K. M., DONATO, G. M., GRAY, M. C., KOLLING, G. L., WARREN, C. A., CAVE, L. M., ... & HEWLETT, E. L. (2012). Systems analysis of the transcriptional response of human ileocecal epithelial cells to Clostridium difficile toxins and effects on cell cycle control. *BMC systems biology*, *6*(1), 2.

D'AURIA, K. M., KOLLING, G. L., DONATO, G. M., WARREN, C. A., GRAY, M. C., HEWLETT, E. L., & PAPIN, J. A. (2013). In vivo physiological and transcriptional profiling reveals host responses to Clostridium difficile toxin A and toxin B. *Infection and immunity*, *81*(10), 3814-3824.

DESSIMOZ, C., GABALDÓN, T., ROOS, D. S., SONNHAMMER, E. L., HERRERO, J., & QUEST FOR ORTHOLOGS CONSORTIUM. (2012). Toward community standards in the quest for orthologs. Bioinformatics, 28(6), 900-904.

DI BELLA, S., ASCENZI, P., SIARAKAS, S., PETROSILLO, N., & DI MASI, A. (2016). Clostridium difficile toxins A and B: Insights into pathogenic properties and extraintestinal effects. Toxins, 8(5), 134.

ETIENNE-MANNEVILLE, S., & HALL, A. (2002). Rho GTPases in Cell Biology. Nature, 420(6916), 629–635.

GENTLEMAN, R. C., CAREY, V. J., BATES, D. M., BOLSTAD, B., DETTLING, M., DUDOIT, S., ... & HORNIK, K. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, *5*(10), R80.

GRUYS, E., TOUSSAINT, M. J. M., NIEWOLD, T. A., & KOOPMANS, S. J. (2005). Acute phase reaction and acute phase proteins. Journal of Zhejiang University SCIENCE, 6B(11), 1045–1056.

HAHNE, F., HUBER, W., GENTLEMAN, R., & FALCON, S. (2010). Processing Affymetrix Expression Data, in *Bioconductor case studies*. Springer Science & Business Media, 25-38.

HOMOLOGENE. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 – [cited 1st May 2017]. Available from: https://www.ncbi.nlm.nih.gov/gene/

HUANG, D. W., SHERMAN, B. T., & LEMPICKI, R. A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic acids research, 37(1), 1-13.

HUANG, D. W., SHERMAN, B. T., & LEMPICKI, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols, 4(1), 44-57.

JANK, T., & AKTORIES, K. (2008). Structure and mode of action of clostridial glucosylating toxins: the ABCD model. *Trends in microbiology*, *16*(5), 222-229.

JANVILISRI, T., SCARIA, J., & CHANG, Y. F. (2010). Transcriptional profiling of Clostridium difficile and Caco-2 cells during infection. *Journal of Infectious Diseases*, *202*(2), 282-290.

KIM, M., ASHIDA, H., OGAWA, M., YOSHIKAWA, Y., MIMURO, H., & SASAKAWA, C. (2010). Bacterial interactions with the host epithelium. *Cell host & microbe*, *8*(1), 20-35.

KLUNE, J., & DHUPAR, R. (2008). HMGB1: Endogenous Danger Signaling. *Molecular Medicine*, *14*(7–8), 1.

LEFFLER, D. A., & LAMONT, J. T. (2015). Clostridium difficile infection. *New England Journal of Medicine*, *372*(16), 1539-1548.

MAGLOTT, D., OSTELL, J., PRUITT, K. D., & TATUSOVA, T. (2011). Entrez Gene: gene-centered information at NCBI. Nucleic acids research, 39(suppl 1), D52-D57.

MICHALIK, L., AUWERX, J., BERGER, J. P., CHATTERJEE, V. K., GLASS, C. K., GONZALEZ, F. J., … STAELS, B. (2006). International Union of Pharmacology. LXI. Peroxisome Proliferator-Activated Receptors. Pharmacol. Rev., 58(4), 726–741.

MORENO, M. A., FURTNER, F., & RIVARA, F. P. (2013). Clostridium difficile: a cause of diarrhea in children. JAMA pediatrics, 167(6), 592-592.

MOTENKO, H., NEUHAUSER, S. B., O'KEEFE, M., & RICHARDSON, J. E. (2015). MouseMine: a new data warehouse for MGI. Mammalian Genome, 26(7-8), 325-330.

NG, J., HIROTA, S. A., GROSS, O., LI, Y., ULKE–LEMEE, A., POTENTIER, M. S., ... & ARMSTRONG, G. D. (2010). Clostridium difficile toxin–induced inflammation and intestinal injury are mediated by the inflammasome. *Gastroenterology*, *139*(2), 542-552.

PAPATHEODOROU, P., ZAMBOGLOU, C., GENISYUEREK, S., GUTTENBERG, G., & AKTORIES, K. (2010). Clostridial glucosylating toxins enter cells via clathrin-mediated endocytosis. *PloS one*, *5*(5), e10673.

QIAGEN, 1984. Ingenuity Pathway Analysis, viewed 1st May 2017 https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis/

RIDLEY, A. J. (2015). Rho GTPase signalling in cell migration. Current Opinion in Cell Biology, 36, 103–112.

RUPNIK, M., WILCOX, M. H., & GERDING, D. N. (2009). Clostridium difficile infection: new developments in epidemiology and pathogenesis. Nature Reviews Microbiology, 7(7), 526-536.

SHAW, D. R. (2016). Searching the Mouse Genome Informatics (MGI) resources for information on mouse biology from genotype to phenotype. Current protocols in bioinformatics, 1-7.

SMEDLEY, D., HAIDER, S., DURINCK, S., PANDINI, L., PROVERO, P., ALLEN, J., ... & BARDOU, P. (2015). The BioMart community portal: an innovative alternative to large, centralized data repositories. Nucleic acids research, gkv350.

SMITH, R. N., ALEKSIC, J., BUTANO, D., CARR, A., CONTRINO, S., HU, F., ... & STEPAN, R. (2012). InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data. *Bioinformatics*, *28*(23), 3163-3165.

ANNEX

Annex 1. R code for Microarray Data Analysis

```
###

#Mice: GSE44091 dataset

#Name: Judith Cantó

#Code description - Code to load the GSE44091 dataset and find differentially expressed genes

###

#Installing the CDF files from BRAINARRAY

install.packages("../../2h-data mice/example/mouse4302mmentrezg.db_21.0.0.zip", repos = NULL, type = "source")

install.packages("../../2h-data mice/example/mouse4302mmentrezgcdf_21.0.0.zip", repos = NULL, type = "source")

install.packages("../../2h-data mice/example/mouse4302mmentrezgprobe_21.0.0.zip", repos = NULL, type = "source")

install.packages("../../2h-data mice/example/pd.mouse4302.mm.entrezg_21.0.0.zip", repos = NULL, type = "source")

install.packages("FactoMineR")

#Packages for microarray analysis are to be downloaded from BiocLite
source("https://bioconductor.org/biocLite.R")

biocLite()

#Downloading the required packages

biocLite("affy")

biocLite("limma")

biocLite("annotate")

biocLite("GEOquery")

biocLite("Biobase")

biocLite("genefilter")

biocLite("affyPLM")

biocLite("tools")

#Loading the libraries in the environment

library(affy)

library(limma)

library(annotate)
```

```
library(GEOquery)

library(affyPLM)

library(genefilter)

library(mouse4302mmentrezgcdf)

library(mouse4302mmentrezg.db)

library(mouse4302mmentrezgprobe)

library(pd.mouse4302.mm.entrezg)

library(FactoMineR)

library(Biobase)

library(tools)

require(ggplot2)

#Set the working directory

setwd(“C:\Users\J\Desktop\Bachelor thesis- Axis\2hdatamice\CEL files”)

#Uncompress the CEL files contained in GSE44091 dataset TAR file

library(R.utils)

untar("GSE44091_RAW.tar", files = NULL, list = FALSE, exdir = ".", compressed = NA, extras =
NULL, verbose = FALSE, restore_times =  TRUE, tar = Sys.getenv("TAR"))

#Read the CEL files from the folder

rm(list = ls())

lst = list.files("./")

dat = ReadAffy(filenames = lst, cdfname = "mouse4302mmentrezgcdf", verbose = TRUE)

#Performing RMA normalization on the raw data

eset.data = affy::rma(dat)

rma = as.data.frame(exprs(eset.data))

#Creating a file with the name of the CEL files and their phenotypic data

dim(rma)

x = colnames(rma)

colnames(rma) = sub(pattern = "\\_.*","",x)

colnames(rma)

x = colnames(rma)

y = read.csv("C:/Users/J/Desktop/Bachelor thesis- Axis/Judith/Pheno_data.csv", header = F,
row.names = 1)
```

```
colnames(rma) = y[x,1]

colnames(rma)

#Principal Component Analysis

write.table(rma, file="mouse_pca_data.txt", quote=F, sep="\t")

mouse_pca = read.table("mouse_pca_data.txt",sep = "\t", header= T, row.names = 1)

colnames(mouse_pc) <- c("Toxin", "Exposure")

PC.1 = PC$x[,c(1,2)]

PC.1 = cbind(PC.1,mouse_pc[row.names(PC.1),c(1,2)])

write.table(PC.1, file="pca_cluster.txt", quote=F, sep="\t") #I modified this in Excel

library(ggplot2)

# Scatter plot with multiple groups

# shape depends on exposure

pca_clust = read.table("pca_cluster.txt",sep = "\t", header= T, row.names = 1)

ggplot(pca_clust, aes(x=PC1, y=PC2, shape=Exposure, color= Toxin, size=4)) +  geom_point()

#####Quality control checks for the expression data######

#box plot of the original data without any RMA normalization

boxplot(dat, col= "cadetblue1",main="mice- Before normalization")

#box plot of the normalized data

boxplot(rma, col="cadetblue1",main="mice RMA expression values")

#creating hierarchical clustering to see effect of genes in conditions in mice

distance.mouse <- dist(t(rma), method="maximum")

clusters.mouse <- hclust(distance.mouse)

#Visualisation of the dendrogram

#load code of A2R function

source("http://addictedtor.free.fr/packages/A2R/lastVersion/R/code.R")

#colored dendrogram

op = par(bg = "white")

A2Rplot(clusters.mouse, k = 5, boxes = FALSE,bg = "white", col.up = "black", col.down =
c("red", "seagreen3", "blue", "yellow3", "maroon1", labels = TRUE))
```

#####Finding differentially expressed genes for mouse data ########

#Create a design matrix. The levels in the design matrix are checked by looking at samples

```
des = factor(c("TcdA2h","TcdA2h","TcdA2h","TcdB2h","TcdB2h","TcdB2h","TcdAB2h",
        "TcdAB2h","TcdAB2h","Sham2h","Sham2h","Sham2h","TcdA6h","TcdA6h",
        "TcdA6h","TcdB6h","TcdB6h","TcdB6h","TcdAB6h","Sham6h","Sham6h",
        "Sham6h","TcdA16h","TcdA16h","TcdA16h","TcdB16h","TcdB16h","TcdB16h",
        "Sham16h","Sham16h","Sham16h","Sham16h"))
design = model.matrix(~0+ des)
colnames(design) = sub("des","",colnames(design))
colnames(design)
```

#setting up the contrast matrix for the design matrix

```
conts = makeContrasts(TcdA2h - Sham2h, TcdB2h - Sham2h, TcdAB2h - Sham2h, TcdA6h -
Sham6h, TcdB6h - Sham6h, TcdAB6h - Sham6h, TcdA16h - Sham16h, TcdB16h - Sham16h,
levels = design)
```

#providing the data for limma analysis

#fit the linear model to the expression set

```
fit <- lmFit(rma, design)
```

#contrast matrix is now combined with the per-probeset linear model fit

```
fit2 <- contrasts.fit(fit, conts)
fit2 <- eBayes(fit2)
```

#Obtain the expression coefficients (log2 ratios)

```
lim= fit2$coefficients
dim(lim)
dim(rma)
```

#Obtain the p-values

```
pvals= fit2$p.value
```

 # Get the adjusted p-values

```
temp = p.adjust(pvals[,1],method = "hochberg")
```

#To view the count of differentially expressed genes (based only on p-value)

```
sum(temp<0.05)  # no DE genes if we adjust the p-value
sum(pvals[,3]<0.01) # we will work with raw p-values, we obtained 267 DE genes
```

#Make a volcano plot

```
require(ggplot2)

genelist<- toptable(fit2, coef=1, number = 2500, sort.by = "P")

genelist$threshold = as.factor(abs(genelist$logFC) > 0.58 & genelist$P.Value < 0.01)

ggplot(data=genelist, aes(x=logFC, y=-log10(P.Value), colour=threshold)) +

  geom_point(alpha=0.4, size=1.75) + xlim(c(-5, 5)) + ylim(c(0, 10)) + xlab("log2 fold change") +
ylab("-log10 p-value")
```

 # Heatmap with DE genes

```
new.coeff = lim[names(gene1.list),]

dim(new.coeff)

pacman::p_load(pheatmap)

myColor <- colorRampPalette(c("green", "black", "red"))(10)

myBreaks <- c(seq(min(new.coeff), 0, length.out=ceiling(10/2) + 1),
seq(max(new.coeff)/10, max(new.coeff), length.out=floor(10/2)))

pheatmap::pheatmap(t(new.coeff),  cluster_row = T, cluster_cols = T, color = myColor,
breaks=myBreaks, fontsize = 6.5, fontsize_row=12, fontsize_col = 6)
```

#Writing the differentially expressed genes onto a new table

```
Temp = row.names(pvals[which(pvals[,3]<0.01, arr.ind = T),])

gene.list = lim[Temp,3]
```

#Storing them in the output folder

```
write.csv(gene.list, "./gene_list.csv" ) # DE genes in Mice TcdAB 2h

write.csv(info, "./gene6_list.csv" ) #we repeat the procedure for DE genes in Mice TcdAB 6h
```

Pathway Analysis

#Venn's diagram of DE pathways at Mice 2h, 6h and Caco2 4.5h.

```
library(VennDiagram)

grid.newpage()

g= draw.triple.venn(area1 = 48, area2 = 251, area3 = 75, n12 = 42, n23 = 49, n13 = 24,
n123 = 20, category = c("Mice TcdAB 2h", "Mice TcdAB 6h", "Caco2 TcdAB 4.5h"), lty = "blank",
cex = 1.5, fill = c("red", "yellow", "skyblue"))

require(gridExtra)

grid.arrange(gTree(children=g), top="Pathways")
```

 # Heatmap with pathways which pvalue<0.05

```
pathway = read.csv ("path_heat.csv", header= F, row.names = 1)
```

```
pacman::p_load(pheatmap)

myColor <- colorRampPalette(c("green", "black", "red"))(10)

colnames(pathway) = c("Caco2", "Mice2h", "Mice6h")

myBreaks <- c(seq(min(pathway), 0, length.out=ceiling(10/2) + 1),
seq(max(pathway)/10, max(pathway), length.out=floor(10/2)))

pheatmap::pheatmap(t(pathway),  cluster_row = T, cluster_cols = T, color = myColor,
breaks=myBreaks, fontsize = 6.5, fontsize_row=12, fontsize_col = 6)
```

 # Heatmap with pathways which pvalue<0.05 and z-score in Mice 2h and Mice 6h

```
pathway = read.csv ("path_mice.csv", header= F, row.names = 1)

pacman::p_load(pheatmap)

myColor <- colorRampPalette(c("green", "black", "red"))(10)

colnames(pathway) = c("Mice2h", "Mice6h")

myBreaks <- c(seq(min(pathway), 0, length.out=ceiling(10/2) + 1),
seq(max(pathway)/10, max(pathway), length.out=floor(10/2)))

pheatmap::pheatmap(t(pathway),  cluster_row = T, cluster_cols = T, color = myColor,
breaks=myBreaks, fontsize = 6.5, fontsize_row=12, fontsize_col = 6)
```