Inter-Faculty Master Program on
Complex Systems and Networks
School of Mathematics
School of Biology
School of Geology
School of Economics
Aristotle University of Thessaloniki

Master Thesis

Title:

Network analysis applications in RNA-seq Data

Εφαρμογές ανάλυσης δικτύων σε δεδομένα αλληλούχισης νέας γενιάς

Vagiona Aimilia-Christina

Supervisor: Sgardelis Stefanos, Professor AUTH

Co-supervisors:

Fotis E. Psomopoulos, Researcher INAB CERTH

Spyros Petrakis, Researcher INAB CERTH

Thessaloniki, November 2019

Διατμηματικό Πρόγραμμα Μεταπτυχιακών Σπουδών στα
Πολύπλοκα Συστήματα και Δίκτυα
Τμήμα Μαθηματικών
Τμήμα Βιολογίας
Τμήμα Γεωλογίας
Τμήμα Οικονομικών Επιστήμων
Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

Μεταπτυχιακή Διπλωματική Εργασία

Τίτλος:

Network analysis applications in RNA-seq Data

Εφαρμογές ανάλυσης δικτύων σε δεδομένα αλληλούχισης νέας γενιάς

Βαγιωνά Αιμιλία-Χριστίνα

ΕΠΙΒΛΕΠΩΝ: Σγαρδέλης Στέφανος, Καθηγητής Α.Π.Θ.

Εγκρίθηκε από την Τριμελή Εξεταστική Επιτροπή την

……	……	……
Σ. Σγαρδέλης	Φ. Ψωμόπουλος	Σ. Πετράκης
Καθηγητής Α.Π.Θ.	Ερευνητής, ΙΝΕΒ ΕΚΕΤΑ	Ερευνητής, ΙΝΕΒ ΕΚΕΤΑ

Θεσσαλονίκη, Νοέμβριος 2019

……..
Βαγιωνά Αιμιλία - Χριστίνα
Πτυχιούχος Βιολόγος Α.Π.Θ.

Abstract

Next Generation Sequencing has created a huge amount of data - data that has internal dependencies and interactions. There are currently many tools that allow the primary analysis of NGS data. In this diploma thesis, a tool in R was constructed which allow: (a) the identification of correlations between different genes in transcriptional data, and (b) the analysis of differences in protein interaction networks of human disease models.

The polyglutamine (polyQ) neurodegenerative disease spinocerebellar ataxia type 1 (SCA1) is a lethal and progressive disorder caused by CAG expansions in the ataxin-1 (*ATXN1*) gene. Mutant *ATXN1* containing more than 39 CAG repeats encodes the production of a pathogenic protein with an abnormal 3-dimensional conformation. The misfolded protein forms inclusions within the nuclei of neurons and sequesters other nuclear proteins, as well. As a result, proteins in the inclusions, including ataxin-1, lose their normal function, an event that causes cytotoxicity and leads to cell necrosis.

Here, we aim in the identification of disease modules within protein interaction networks and molecular mechanisms of dysfunctions that are related to SCA1 progression. To this end, we analyzed RNA-seq data from a cell and a mouse model of SCA1 at three discrete time points of protein aggregation and compared them with similar data from the cerebellum of a SCA1 patient containing polyQ inclusions. We show that the pathways protein digestion and absorption, ECM-receptor interaction (cells-mice) and PI3K-Akt (cells-mice-patient) signaling are commonly dysregulated in all datasets.

Keywords: RNA-seq analysis, SCA1, Differential expression analysis, Protein interaction network

Περίληψη

Οι τεχνολογίες αλληλούχησης νέας γενιάς (Next Generation Sequencing) έχουν δημιουργήσει ένα τεράστιο όγκο δεδομένων – δεδομένα τα οποία ενέχουν εσωτερικές εξαρτήσεις και αλληλεπιδράσεις. Αυτή τη στιγμή υπάρχουν πολλά εργαλεία που επιτρέπουν την πρωταρχική ανάλυση δεδομένων NGS. Στα πλαίσια αυτής της διπλωματικής κατασκευάστηκε ένα εργαλείο στην R το οποίο επιτρέπει: (α) την ανεύρεση συσχετίσεων μεταξύ διαφορετικών γονιδίων σε μεταγραφικά δεδομένα και (β) την ανάλυση των διαφορών σε πρωτεϊνικά δίκτυα αλληλεπίδρασης από μοντέλα ανθρώπινων νοσημάτων.

Η νωτιαίο-παρεγκεφαλιδική αταξία τύπου 1 (SCA1) είναι μια θανατηφόρα νευροεκφυλιστική ασθένεια πολυγλουταμίνης (polyQ) που προκαλείται από την επέκταση τρινουκλεοτιδίων CAG στο γονίδιο της ataxin-1 (*ATXN1*). Η μεταλλαγμένη *ATXN1* περιέχει περισσότερες από 39 επαναλήψεις CAG και κωδικοποιεί την παραγωγή μιας παθολογικής πρωτεΐνης με λανθασμένη τρισδιάστατη διαμόρφωση. Η παθολογική πρωτεΐνη σχηματίζει έγκλειστα στον πυρήνα των νευρώνων τα οποία περιλαμβάνουν και άλλες πυρηνικές πρωτεΐνες. Αυτό έχει ως αποτέλεσμα την απώλεια της φυσιολογικής τους λειτουργίας, συμπεριλαμβανομένου και της ATXN1, γεγονός που προκαλεί κυτταροτοξικότητα και νέκρωση των κυττάρων.

Γι’ αυτό το σκοπό αναλύσαμε δεδομένα RNA-seq από ένα κυτταρικό και ένα ζωικό μοντέλο της ασθένειας σε τρεις διακριτές φάσεις της πρωτεϊνικής συσσωμάτωσης και τα συγκρίναμε με αντίστοιχα δεδομένα από την παρεγκεφαλίδα ενός ασθενή που περιέχει πρωτεϊνικά έγκλειστα polyQ. Μονοπάτια όπως η πέψη και απορρόφηση πρωτεϊνών, η αλληλεπίδραση μεταξύ της εξωκυτταρικής μήτρας και των υποδοχέων της (κυτταρικό και ζωικό μοντέλο), καθώς και το σηματοδοτικό μονοπάτι PI3K-Akt είναι απορυθμισμένα και στις τρεις ομάδες δεδομένων RNA-seq (κυτταρικό, ζωικό μοντέλο και ασθενής).

Λέξεις κλειδιά: Ανάλυση RNA-seq, SCA1, Ανάλυση διαφορικής έκφρασης, Πρωτεϊνικό δίκτυο αλληλεπίδρσης.

Context

Table of figures

Table of tables

Acknowledgments

I would like to express my sincere thanks to all my supervisors for the crucial guidance and encouragement for the integration of this work. Special thanks to Fotis Psomopoulos and to Spyros Petrakis, for their thorough and meaningful help throughout the course of this work. Also, I would like to thank my friends and my family for their continuous support.

1. Introduction

1.1 NGS analysis

Next generation sequencing (NGS) (Metzker, 2010) has become a valuable tool for life sciences and research projects, studying molecular biology, evolutionary biology, metagenomics and oncology. This technology has widened our understanding on epigenetics and is widely used in several experimental setups (Schorderet 2016). Compared to microarray technologies, NGS provides higher resolution data and more precise measurement of transcripts levels for studying gene expression. Furthermore the downstream analysis of RNA sequencing (RNA-seq) data, indicates the variability in gene expression between different samples (Wang et al. 2019).

NGS provides an enormous amount of complex data, making the extraction of the relevant experimental information time consuming and challenging for researchers (Backman and Girke 2016). Furthermore, several companies have developed so far different sequencing platforms.

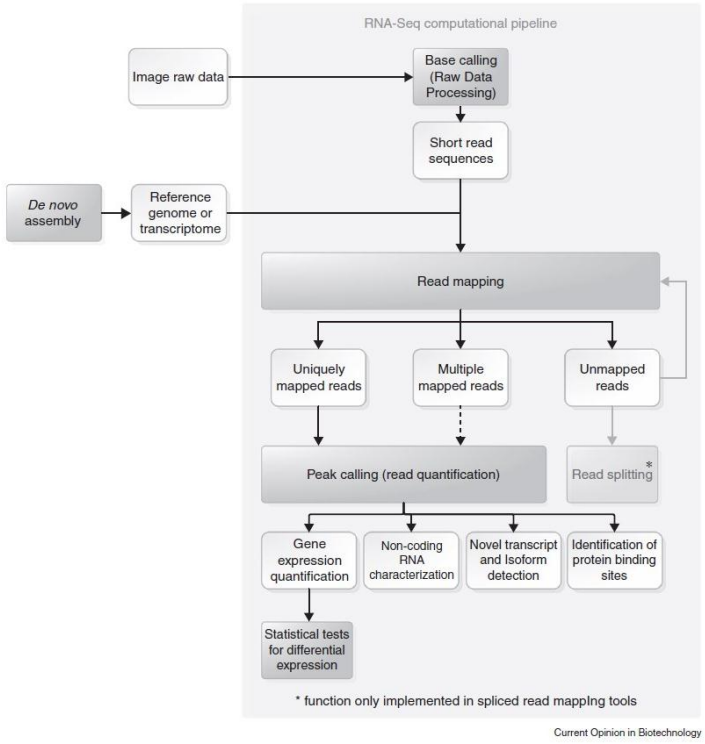

Figure 1: Computational pipeline for next-generation sequencing data (Mutz, 2013)

The pipeline for the analysis of RNA-seq data (Fig. 1) consists of four steps, provided that the genome or the transcriptome of the reference organism has been

already sequenced. First, raw image data need to be converted into short read sequences, which are subsequently aligned to the reference genome or transcriptome. The number of mapped reads is counted and gene expression levels are calculated by peak calling algorithms. Finally, differential gene expression is determined using statistical tests (Mutz, 2013).

The main experimental aim in RNA-seq is the identification of differentially expressed genes (DEG), which is the final step during the downstream data analysis. DEGs vary significantly in their expression levels between two sets of samples and are either up- or down-regulated. They can be further classified according to their biological process, molecular function, cellular localization or biological pathway in which they participate. Eventually, this allows the functional and biological interpretation of the experimental results (Sultan and Zubair 2019).

1.2 Poly-Q diseases

Polyglutamine diseases (polyQ) are a family of neurodegenerative disorders that are caused by CAG trinucleotide expansions in various genes. This mutation produces a pathogenic protein that contains a longer polyQ chain than the wild type protein (Orr 2012a). As a result, the mutant protein adopts a different conformation which is accompanied by a loss of its functionality (Siska, Koliakos, and Petrakis 2015). These prototypical protein misfolding disorders include Huntington disease, spinobulbar muscular atrophy, dentatorubral-pallidoluysian atrophy and several spinocerebellar ataxias (Shao and Diamond, 2007). There are approximately 30 different types of SCA identified to date, but the causative mutations have been identified for only half of them. Six SCAs, including the more prevalent SCA1, SCA2, SCA3, and SCA6 along with SCA7 and SCA17 are caused by expansion of a CAG repeat that encodes a polyglutamine tract in the affected protein (Orr 2012b).

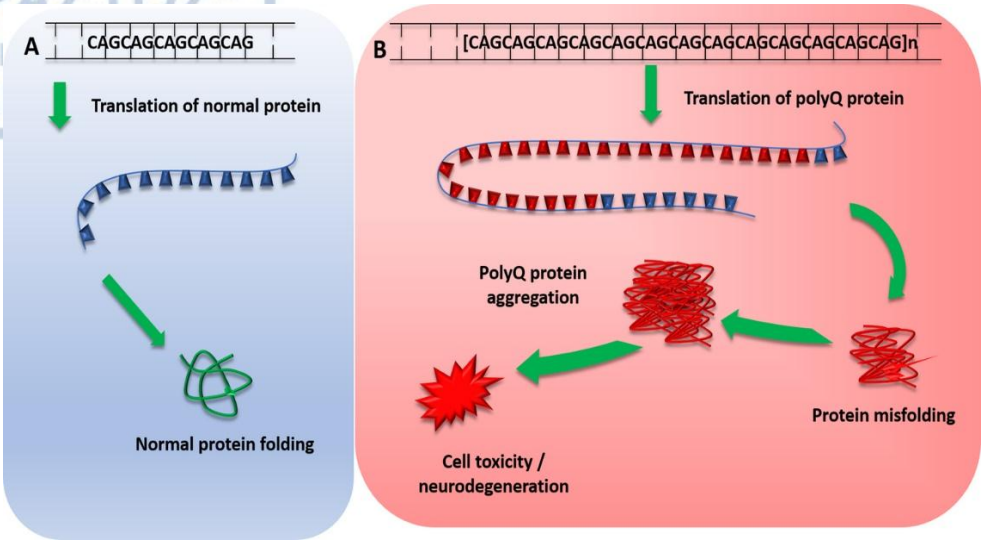

Figure 2: Effect of polyQ repeats in protein folding. A) Translation of a polyQ gene with a normal repeat range, produces protein with a proper folding. B) Pathogenic polyQ repeat expansions lead to production of a pathogenic protein with an expanded track that is misfolded (Sullivan et al. 2019)

1.2.1 Spinocerebellar ataxia type 1 (SCA1)

The polyQ disease SCA1 is a lethal, progressive, autosomal dominant neurodegenerative disorder caused by a CAG expansion in the ataxin-1 (*ATXN1*) gene (Orr et al. 1993). SCA1 patients typically display loss of coordination of the limbs and trunk, unstable gait, dysarthria, and nystagmus (Klockgether, 2011). Symptons typically manifest in midlife and worsen over the next 10–15 years; there is no available therapy to delay the onset or slow the progression of the disease (Zoghbi et al. 1988). A prominent and consistent SCA1 pathological feature is the loss of the cerebellar Purkinje cells (PCs) (Koeppen, 2005).

SCA1 is caused by an expanded CAG repeat in the *ATXN1* gene. Expansions longer than 39 CAG repeats are pathologic. Longer repeats generally result in an earlier onset of disease. However, CAT interruptions in the polyQ track of *ATXN1* can alter the penetrance and aggressiveness of the disease (Menon et al. 2013). Despite the fact that polyQ expansions are considered the main cause of pathology, other factors also affect the progression of SCA1 (Srinivasan and Shakkottai 2019) Numerous studies have shown that ATXN1 interacts with transcription regulators, RNA splicing factors, and other nuclear receptors that drive cerebellar Purkinje cell dysfunction (Lam et al. 2006; Tsuda et al. 2005)

Mutant ATNX1 misfolds into an abnormal 3-dimensional conformation and forms protein inclusions within the nuclei of neurons. As a result, ATXN1 loses its normal function, an event that damages cells and leads to cell necrosis. It is still unclear why polyQ-expanded ATXN1 inclusions are mainly found in the brain and the spinal cord (central nervous system). Cerebellar neurons that coordinate movement are particularly sensitive to ATXN1 aggregation. Their gradual dysfunction and loss causes the characteristic symptoms of SCA1 (Matilla-Dueñas, Goold, and Giunti 2008).

1.2.2 Molecular Mechanisms of Neurodegeneration

ATXN1 is located in both the cytoplasm and nucleus; the wild-type protein is able to translocate between these two subcellular compartments (Fig 3). The dynamics of ATXN1 cellular trafficking is altered by the expansion of the polyQ tract (Krol et al. 2008). Although mutant ATXN1 is able to enter the nucleus, its ability to be translocate back into the cytoplasm is dramatically reduced (Irwin et al. 2005). ATXN1 interacts with several proteins including the transcription regulators, Capicua (Lam et al. 2006), Gfi-1 (Tsuda et al. 2005), and the Rora–Tip60 complex (Serra et al. 2006). ATXN1 also interacts with RNA-splicing factors, such as RBM17 (Lim et al. 2008) and U2AF65 (de Chiara et al. 2009).

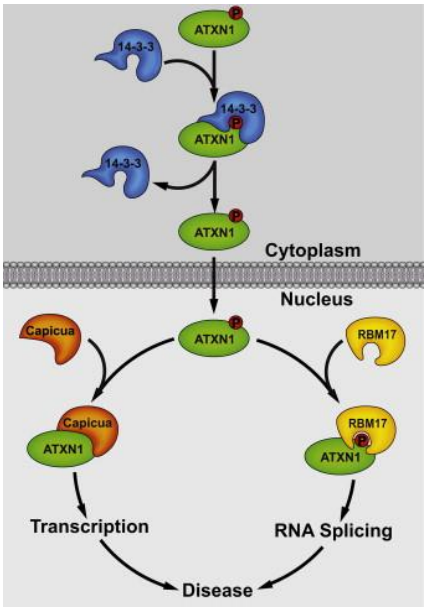

Figure 3.Contribution of ATXN1 S776 phosphorylation in SCA1 pathogenesis. ATXN1 in the nucleus interacts with either the transcriptional repressor Capicua or the RNA splicing factor RBM17. Phosphorylated ATXN1 interacts stronger with RBM17 affecting RNA splicing (Orr 2012a)

Several lines of evidence indicate that the C-terminal domain of ATXN1, plays a key role in its function and is associated with SCA1 pathogenesis. Ser776, immediately adjacent to the nuclear localization sequence (NLS) in the C-terminal of the protein, is an endogenous phosphorylation site (Emamian et al. 2003). Phosphorylation of S776 stabilizes ATXN1 and may regulate its interaction with other proteins such as the phospho-serine/phospho-threonine binding protein 14-3-3 (Chen et al. 2003), a signal transduction regulat (Morrison 2009), the splicing factors RBM17 (Lim et al. 2008) and the transcriptional repressor Capicua.

The interaction of ATXN1 with RBM17 is enhanced by the polyQ expansion but is dramatically suppressed in the presence of the phosphorylation-resistant ATXN1-A776, independently of the length of the polyQ tract (Lim et al. 2008). These data indicate that phosphorylation of serine 776 is critical for the strength of this interaction.

The gain of function of the ATXN1–CIC complex leads to neurodegeneration but also plays an important role to normal brain development and is essential for survival. Loss of this complex causes a spectrum of neurobehavioral phenotypes (hyperactivity, intellectual disability and social-behavioral deficits) (Lu et al. 2017). Also, a recent study shows that the interaction of ATXN1 with CIC is the major driver of toxicity in SCA1. Data from gain and loss of function models and SCA1 patients indicate that ATXN1-CIC complex is crucial for the observed toxicity while loss of CIC in the cerebellum does not result in the degeneration of the Purkinje cells to (Rousseaux et al. 2018).

1.2.3 Disease models

1.2.3.1 Mouse models

In order to gain insight into the pathogenesis of the SCA1, transgenic mice expressing the human *ATXN1* gene with either a normal or an expanded CAG tract have been generated. Mice expressing the normal ATXN1 had normal Purkinje cells, while transgenic animals with the mutant ATXN1 developed ataxia and Purkinje cell degeneration. These results indicate that a neurodegeneration mouse model can be established simply by introducing CAG repeat expansions in a wild-type protein

(Burright et al. 1995). Thus, SCA1 transgenic mice provide a tool to identify pathways associated with SCA1pathogenesis.

Recently, RNA-seq analysis in the cerebellum of SCA1 transgenic mice has demonstrated gene expression changes that are implicated in disease progression or the systemic response against it. Gene networks from SCA1 mice with progressive Purkinje cell loss were constructed and compared to healthy or ataxic mice that lack a progressive Purkinje cell loss (Ingram et al. 2016). These data indicate that e.g. the Cck gene is protective against the progression of SCA1.

1.2.3.2 Cell model

Animal models do not indicate molecular changes at the cellular level that are caused by the gradual aggregation of the mutant polyQ protein. Therefore, there is a need for cell models that would indicate molecular mechanisms of dysfunction potentially causing the disease phenotype in mice. A cell model of intranuclear protein aggregation was generated by the inducible overexpression ATXN1(Q82) in human mesenchymal stem cells. These cells are resistant to the cytotoxic effects of the mutant protein and allow the detailed study of its aggregation (Laidou *et al*., unpublished data).

In this study, we assessed the similarity of this SCA1 cell model with SCA1 transgenic mice. We also compared them with a human SCA1 cerebellum at the end-stage of the disease, containing polyQ inclusions. The potential implications of the commonly identified molecular changes for the pathogenesis of SCA1 are discussed.

2. Materials and Methods

2.1 Dataset summary

The expression level of genes from three different groups (Human, Mice and Cells) were used for analysis. Expression levels were measured in ‘‘fragments per kilobase of exon model per million mapped reads’’ (FPKM) (Trapnell et al. 2010).

2.1.1 Dataset of a SCA1 mouse model

In order to identify the differentially expressed genes that changed over time, the dataset of mice contains the FPKM abundance in six samples. Control in three time poins: week 5 (FPKM_W5_Q82), week 12 (FPKM_W12_Q82), week 28 (FPKM_W28_Q82) and SCA1 trangenic mice in three time points: week 5 (FPKM_W5_FVB), week12 (FPKM_W12_FVB) and week 28 (FPKM_W28_FVB). The accession number for the RNA-seq data reported in this paper is GEO: GSE75778 (Ingram et al. 2016b).

DIOPT (DRSC Integrative Ortholog Prediction Tool) tool was used to map orthologous genes among mice and human (DIOPT; http://www.flyrnai.org/diopt). DIOPT is a program that integrates ortholog predictions from 11 commonly used orthology tools (Hu et al. 2011). Human orthologs of mice genes were found based on Rank score and the list included only those genes with Rank score= High

2.1.2 Dataset of a human SCA1 patient

The second dataset includes RNA-seq data from post-mortem human cerebellum of a 74-year-old female SCA1 patient and an age-/sex-matched healthy individual. All tissues were obtained from the MRC London Neurodegenerative Diseases Brain Bank. Gene expression levels measured in FPKM (*FPKM.IZ_TR_184_S2. Control* and FPKM.IZ_TR_185_S3. SCA1) (Laidou *et al*. unpublished data).

2.1.3 Dataset of cell model

The SCA1 cell dataset comes from cells from human mesenchymal cells (MSCs) inducibly overexpressing polyQ-expanded ATXN1. This model reproducibly generates large nuclear inclusions. Gene expression levels were measured in FPKM

from four samples: at day 0 (FPKM_D0), day 2 (FPKM_D2), day 5 (FPKM_D5) and day 10 (FPKM_D10) (Laidou *et al*. unpublished data).

2.2 Differential Expression Analysis

Gene expression (abundance) was measured by FPKM values for every transcript and was associated with an individual gene. Each sample was measured in triplicates with the exception of human cerebellum. The expression level of each gene is the mean of the FPKM of its triplicates. T-test was applied to compare FPKM levels between the triplicates of a sample in mice and cells datasets and only genes showing a consistent expression (p-value<0.05) were used for further analysis. Gene expression in the human dataset was normalized using the GFOLD tool (Feng et al. 2012). A total of 12 samples were selected to obtain the gene expression patterns. The samples that were used for the analysis are listed in table 1.

Table 1: Samples for Differential Expression Analysis

Group	Time point	Samples/Control	Samples/Patient
Human	-	*FPKM.IZ_TR_184_S2. Control*	FPKM.IZ_TR_185_S3. SCA1
Mice	Week 5	FPKM_W5_FVB	FPKM_W5_Q82
	Week 12	FPKM_W12_FVB	FPKM_W12_Q82
	Week 28	FPKM_W28_FVB	FPKM_W28_Q82
Cells	Day 2	FPKM_D0	FPKM_D2
	Day 5	FPKM_D0	FPKM_D5
	Day 10	**FPKM_D0**	**FPKM_D10**

For the analysis, differential gene expression was calculated as the FC (Fold Change) of a SCA1 sample versus its respective control in each time point

$$FC=\frac{FPKM\ g\ (SCA1)}{FPKM\ g\ (control)}$$

FC data were log2 normalized and genes with $|\log_2 FC| > 0.5$ were considered as DEGs. A gene with a positive log2Fold value was considered as upregulated whereas a negative log2Fold marks down-regulated genes. Differential expression

analysis was performed in the R version 3.6.1 (RStudio Team 2016) programming software.

DEGs obtained at day 2 in cells, were compared to the DEGs at week 5 in mice. Similarly, DEGs at day 5 in cells, were compared to the DEGs at week 12 in mice and DEGs at day 10 in cells, were compared to the DEGs at week 28 in mice and to DEGs in human. Also, DEGs in human, were compared only to DEGs at week 28 in mice (Table 2).

Table 2: Comparisons details in Differential Expression Analysis

Comparisons for the Differential Expression Analysis
DEGs at day 2 in cells ~ DEGs at week 5 in mice
DEGs at day 5 in cells ~ DEGs at week 12 in mice
DEGs at day 10 in cells ~ DEGs at week 28 in mice
DEGs at day 10 in cells ~ DEGs at week 28 in mice ~ DEGs in human

In each comparison the common up- or -down regulated genes were used for the construction of heatmaps and principal component analysis. The clustering was performed in ggplot2 (version 3.2.1). PCA analysis was performed in ggbiplot (version 0.55) of R packages (version 3.6.1).

2.3 Functional Enrichment Analysis

To identify dysfunctional pathways associated with SCA1 pathogenesis, pathway enrichment analysis was performed using dysregulated genes from each comparison using the enrichR package (version 2.1). EnrichR provides an R interface to all 'Enrichr' databases. 'Enrichr' is a web-based tool for analyzing gene sets and returns any enrichment of common annotated biological features (Kuleshov et al. 2016). Up and down regulated gene lists were evaluated for significant enrichment against the KEGG database. KEGG is a database resource for understanding high-level functions and utilities of the biological system, including the cell, the organism and the ecosystem (Kanehisa and Goto, 2008).

Enriched pathways were selected and ranked by the combined score, as calculated by the EnrichR platform. KEGG pathways with a p-value less than 0.05 that are related to SCA1 were used for the construction of the Protein Protein Interaction Network.

2.4 Construction of Protein Interactions Networks

The protein products of genes that participate in the commonly dysregulated pathways per comparison were used for the construction of a protein-protein interaction network (PPI network) using the String database (Szklarczyk et al. 2017) in Cytoscape 3.7.2 version (Shannon 2003) . Only genes expressed in the nervous system (score of 4.8 using the relenat tissue filter)(Santos et al. 2015) and high confidence interactions (score of 0.950) were used. Unconnected nodes were deleted.

2.5 Network analysis

The Cystoscope plugin Network Analyzer (Assenov et al. 2008) was used to compute the centrality parameters of the network. We extracted genes based on four criteria: a) Degree centrality (DC), b) Betweenness centrality (BC), c) Closeness centrality (CC) and d) Clustering coefficient centrality (CU)

2.5.1 Degree centrality (DC)

All the lines connected by a node are called the degree of the node. The more connections between a node and other nodes, the greater the node. This indicates that the node is important for the network. The degree centrality of a node i, is defined as (Nieminen 1974):

$$C_i^{\mathrm{D}} = \frac{k_i}{N-1} = \frac{\sum_{j \in \mathbf{G}} a_{ij}}{N-1}$$

where N is a set of nodes, K is a set of edges and ki is the degree of node i.

2.5.2 Betweenness centrality (BC)

BC is the number of the shortest pathways of all node pairs through the node in a network, and the times of one node serves as the bridge of the shortest pathway between the other two nodes (Zhang 2018). BC refers to the frequency of node i appearing at nodes j and k. The standard formula is (Freeman 1978)

$$C_i^{\mathrm{C}} = (L_i)^{-1} = \frac{N-1}{\sum_{j \in \mathbf{G}} d_{ij}},$$

where i ≠ j ≠ k, gjk is the number of the shortest pathways between nodes j and k, gjk(i) is the number of the shortest pathways containing i, N is the number of nodes, denominator is two times the logarithmic number of nodes except node i in protein interaction network.

2.5.3. Closeness centrality (CC)

Also known as tightness centrality, it is based on the calculation of the average shortest pathway length of a node and all other nodes. CC is calculated as the reciprocal of the sum of the length of the shortest paths between the node and all other nodes in the graph. Thus, the more central a node is, the closer it is to all other nodes (Zhang 2018). The standard formula is (Sabidussi 1966).

$$C_i^{\mathrm{B}} = \frac{1}{(N-1)(N-2)} \sum_{j \in \mathbf{G}, j \neq i} \sum_{k \in \mathbf{G}, k \neq i, k \neq j} n_{jk}(i)/n_{jk},$$

where i ≠ j, dij is the shortest pathway between nodes i and j. If the connections between the node i and other nodes are very short, the more centrally located.

2.5.4 Clustering coefficient centrality (CU)

According to graph theory, clustering coefficient represents the degree of aggregation of nodes in a graph. It is the ratio of adjacent points pairs directly to all neighboring points in the neighboring points of the node (Watts and Strogatz 1998). The formula is defined as (Wasserman & Faust 1994):

$$\bar{C} = \frac{1}{n}\sum_{i=1}^{n} C_i.$$

where n represents the number of edges between all neighbors of node i.

3. Results

3.1 Differential expression analysis

3.1.1 Differential expression analysis between Cells-Day 2 and Mice-Week 5

To detect common gene expressions changes in the different datasets, we compared the profiles of mice at week 5 and cells at day 2. A total of 357 genes in mice at week 5 were selected, based on their consistent expression levels in biological triplicates (p-value <0.05 and |Log2fc|>0.5). These include 105 up-regulated and 252 down-regulated genes. Similarly, 687 genes were selected from the cells dataset at day 2 consisting 279 up-regulated and 408 downregulated genes.

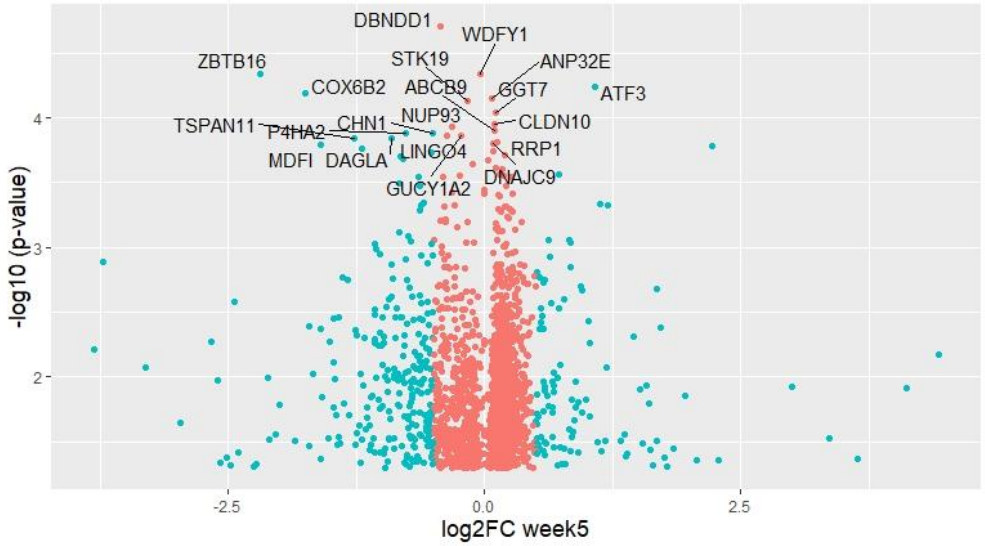

Figure 4: Volcano plot of genes (p-value < 0.05) in the mice RNA-seq dataset at week 5

The log2 fold change is represented on the x-axis, and negative log of p-values is represented on the y-axis of the volcano plot (Fig 4). Each point represents one gene, in the mice dataset at week 5. DEGs with |log2FC| > 0.5 are shown in blue while, nonsignificant genes are shown as red points. The top 20 significant genes are labeled in the volcano plot.

Figure 5 shows the volcano plot of the cell dataset at day 2 (p-value < 0.05). DEGs with |log2FC| > 0.5 are shown in blue, while nonsignificant genes in red. The top 20 significant genes are labeled in the volcano plot.

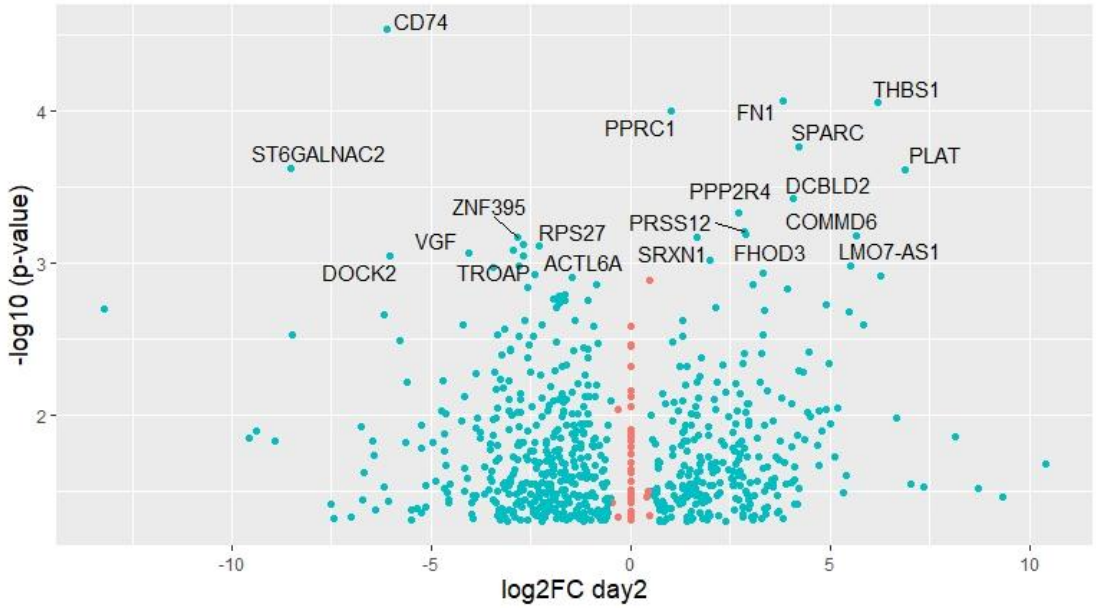

Figure 5: Volcano plot of genes (p-value < 0.05) in the cells RNA-seq dataset at day 2

A total of 18 genes were differentially expressed in both datasets (mice at week 5 and cells at day 2) as shown in Venn diagram (Fig 6). Commonly dysregulated genes are shown in Table 3.

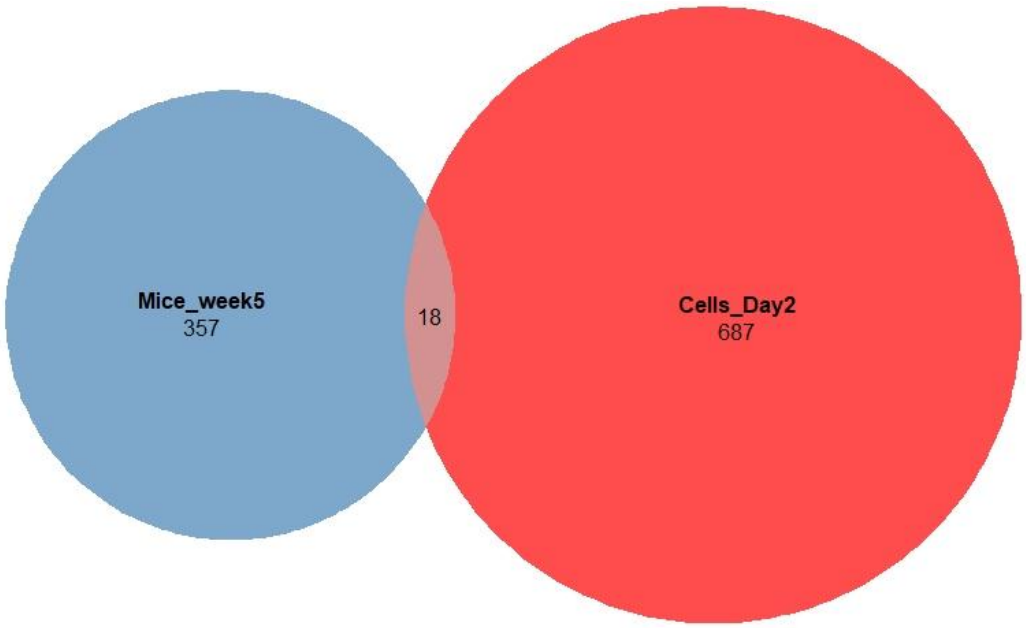

Figure 6: Venn diagram indicates 18 overlapping genes between DEGs in mice at week 5 and cells at day 2

Table 3: List of the common dysregulated genes between week 5 in mice and day2 in cells

Gene symbol	Log2FCcells_D2	LOG2FCmice_w5
GREB1	-8,93002	-0,77085
CRLF1	-6,67426	-0,92048
ATF3	-3,99286	1,074221
CREG1	-3,38309	-0,79244
MPZ	-2,89119	-1,5086
PER1	-2,46202	0,538488
TSPAN18	-2,39366	0,618492
COL18A1	-0,61665	-0,93243
IGF2BP1	0,710022	3,643578
AKR1B1	1,400455	0,6668
SPATS2L	1,581184	-0,53053
MMP14	2,2899	0,553661
PRSS12	2,85587	-1,45611
FRMD6	2,895461	0,742413
BCAR1	2,97111	-0,56847
IGFBP5	3,065207	-0,54587
PTGER2	3,359382	-1,24782
THBS1	6,199589	0,504296

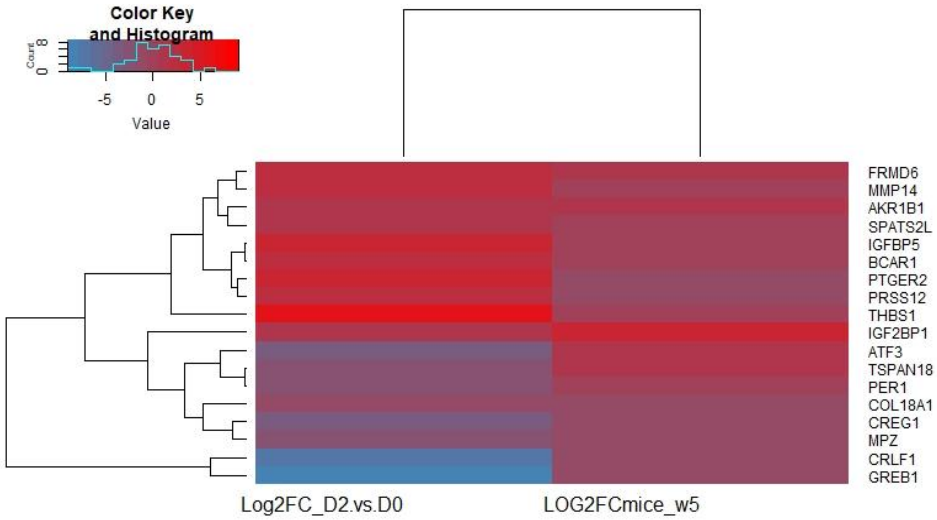

Figure 7: Heatmap showing the log2FC in the expression of the 18 overlapping genes in mice at week 5 and cells at day 2

Five genes were down-regulated (GREB1, CRLF1, CREG1, MPZ, COL18A1) while five genes were up-regulated (IGF2BP1, AKR1B1, MMP14, FRMD6, THBS1) in both datasets as show in Figure 7.

Principal component analysis (PCA) indicates that the overlapping genes (n=18) distinctively cluster into two different categories, suggesting that they have a different gene expression pattern in the two datasets (Fig 8).

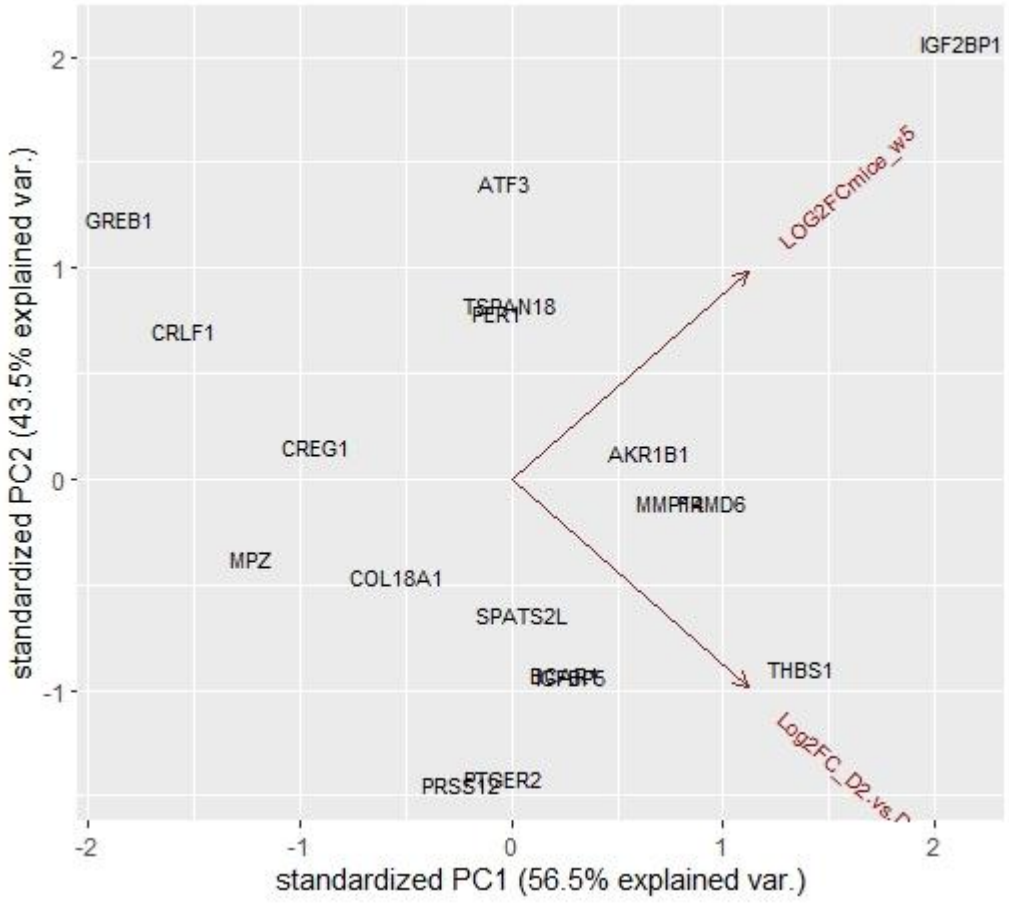

Figure 8: Principal Component Analysis (PCA) of gene expression profiles from mice at week 5 and cells at day 2

3.1.2 Differential expression analysis between Cells-Day 5 and Mice-Week 12

Next, we compared the RNA-seq datasets of mice at week 12 and cells at day 5 and studied gene expression patterns. 1204 genes were selected from the mice dataset at week 5 based on their consistent expression in the biological triplicates (p-value <0.05 and |Log2FC|>0.5). 398 genes were up-regulated and 593 were down-regulated. Similarly, 789 genes were selected from cells dataset, consisting 303 up-regulated and 486 down-regulated. Volcano plots of these two gene groups are shown in Fig 9 and Fig 10 and the top 20 dysregulated are labeled. DEGs with |log2FC| > 0.5 are shown in blue while, nonsignificant genes are shown in red.

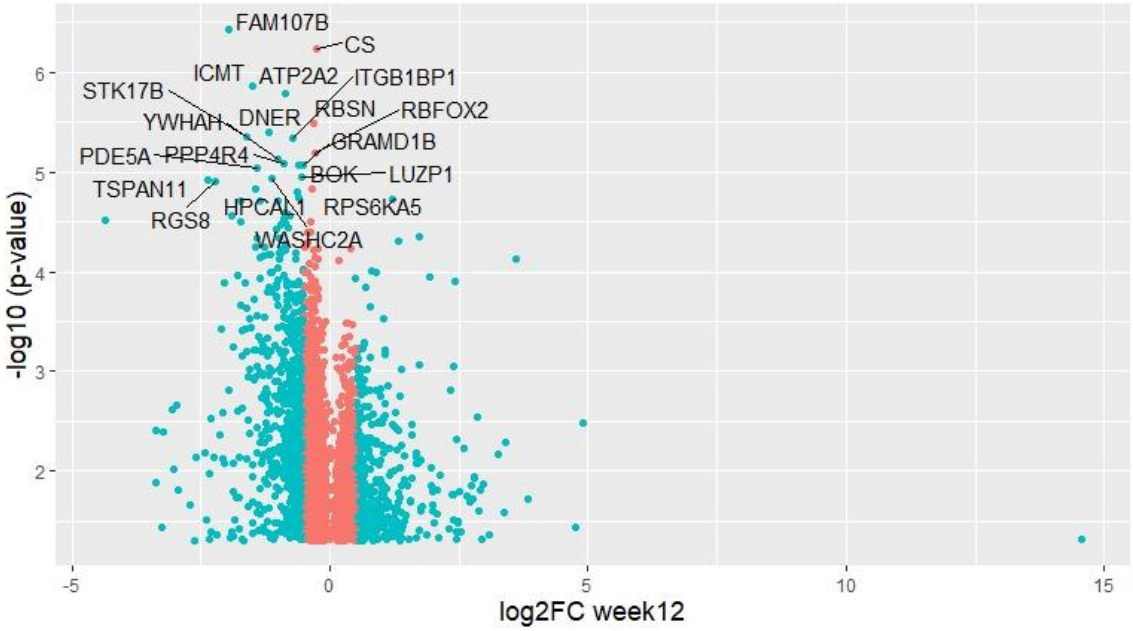

Figure 9: Volcano plot of genes (p-value < 0.05) in the mice RNA-seq dataset at week 12

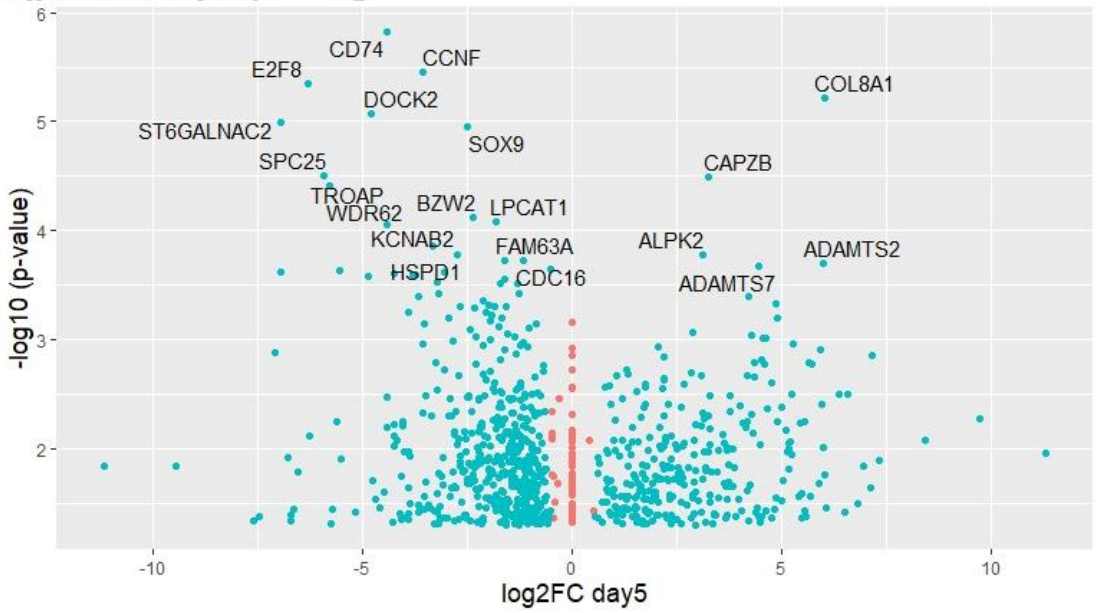

Figure 60 Volcano plot of genes (p-value < 0.05) in cells RNA-seq dataset at day 5

A Venn diagram indicates that 53 genes are common between DEGs of mice at week 12 and DEGs of cells at day 5 (Fig 11). Overlapping genes are listed in Table 4.

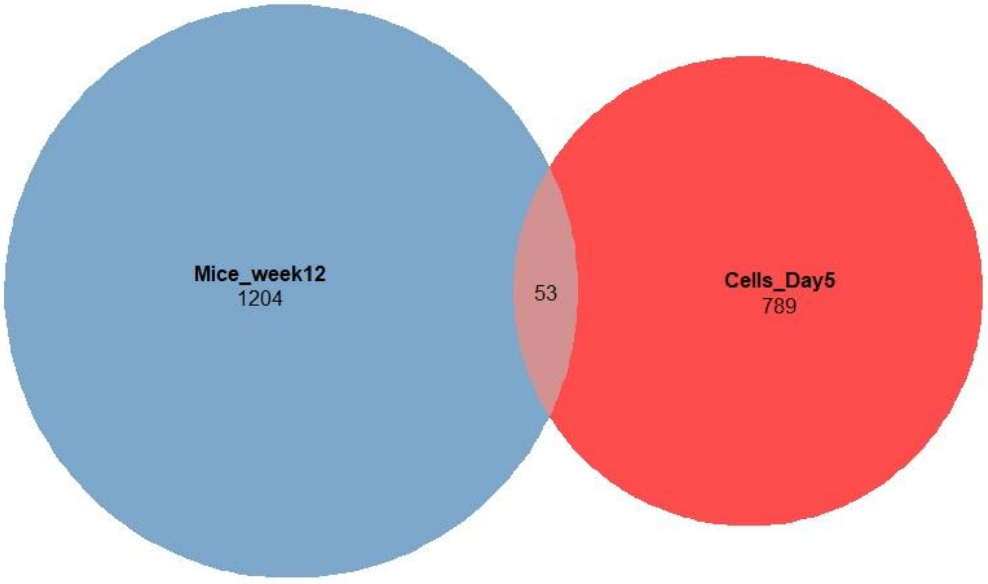

Figure 71: Venn diagram indicates 53 overlapping genes between DEGs of mice at week 12 and cells at day 5

Table 4: List of the common dysregulated genes between mice at week 12 and cells at day 5

Gene symbol	Log2FCcells_D5	LOG2FCmice_w12	Gene symbol	Log2FCcells_D5	LOG2FCmice_w12
ACOT9	3,344038	-0,55874	MICALL2	-0,85744	0,700209
ADAMTS1	2,223245	-0,87229	MMP14	2,93528	0,706291
ALDH2	4,857237	0,590464	MXRA7	2,041899	1,578055
ATF3	-2,84088	1,967727	P4HA2	2,352431	-0,63494
BRWD3	-1,96821	-0,64848	PAXBP1	-3,14028	-0,62975
CAMK2A	6,928887	-1,1581	PDPN	1,675112	0,653105
CCND1	1,024861	0,617102	PLCB2	-7,47617	1,108255
CD44	1,062387	0,642703	PLCD3	2,891969	-0,57104
CD74	-4,43277	0,915891	PLEKHG4	-3,20301	-0,75041
CLSTN2	2,238265	-0,94799	PRR11	-2,50013	1,151382
COL18A1	1,660514	-2,04348	RCN3	5,170932	0,631953
COL1A2	6,012376	0,54037	RNASEH2C	-1,51229	0,798496
COL5A1	4,820899	-2,42979	RPS17	-1,24759	-0,70547
COL6A2	3,700745	0,89672	RPS27	-1,07739	-0,75195
CREG1	-1,86688	-0,905	SDC1	1,325948	0,755294
CSRP2	5,223708	0,594577	SLC25A37	-1,33823	-0,51211
DGKZ	-1,79097	-0,83163	SLC6A17	-7,60689	-0,54365
EMILIN1	3,238149	1,060482	SVEP1	4,748111	-0,8229
FKBP10	2,679717	0,518734	TAX1BP3	2,22202	0,549636
GNAI1	4,462322	-0,54067	TEAD3	2,003282	0,590779
GPC6	2,993134	-0,52998	THBS1	5,715139	0,788769
GTF3C3	-1,25205	-0,58349	TMEM70	-1,65618	-0,53306
HHIP	-2,54431	-0,76642	TNS3	2,878453	-0,61856
IGFBP5	5,161975	-1,43226	TOMM6	-1,60017	0,624994
ITGA1	4,255396	-0,52111	TYMS	-2,06716	1,034065
LXN	4,126967	0,503225	VPS13B	-0,68987	-0,54891
LYPLA1	-0,93064	-0,60209			

As shown in Fig 12, the 53 overlapping genes clustered into two subgroups. A total of 17 genes were up-regulated and 14 were down-regulated. Principal component analysis (PCA) was applied to explore relationships in gene expression among the samples. According to PCA, the samples from mice and cells are separated, indicating the differences on gene expression (Fig 13).

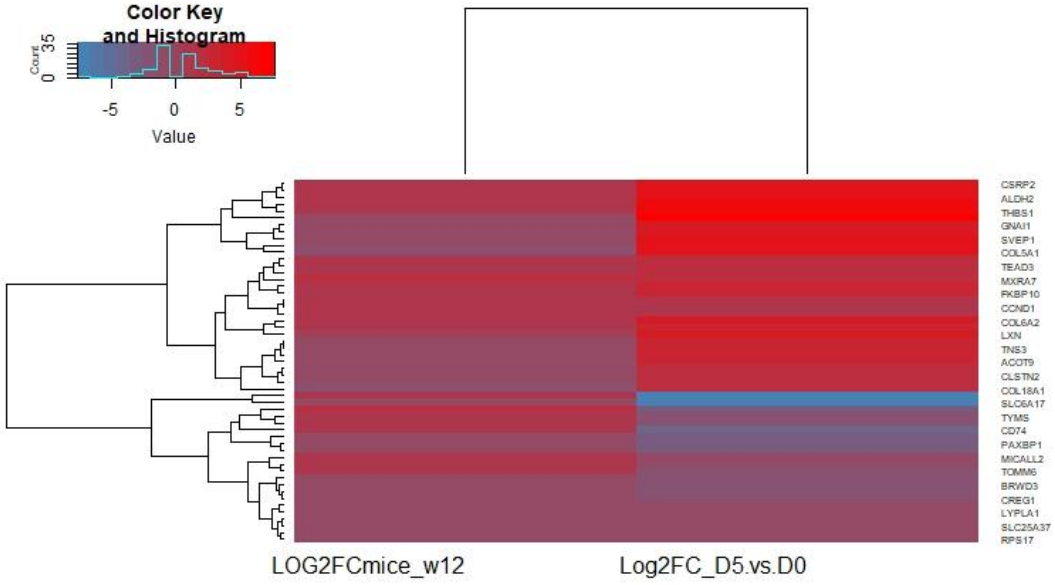

Figure 82: Heatmap showing the log2FC expression of the 53 overlapping genes in mice at week 12 and cells at day 5

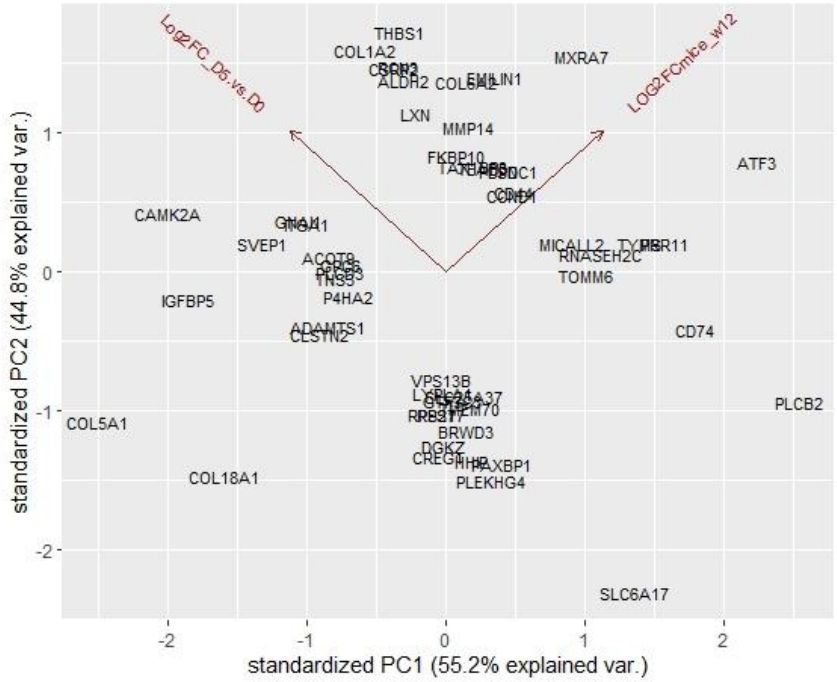

Figure 93: Principal Component Analysis (PCA) of gene expression profiles from mice at week 12 and cells at day 5

3.1.3 Differential expression analysis between Cells-Day 10 and Mice-Week 28

Next, we assessed the similarity in gene expression that might be related with the final stages of the disease by comparing mice at week 28 and cells at day 10. After applying a filtering approach for consistent expression among the experimental triplicates in both samples (p-value < 0.05 and |Log2FC|>0.5), we identified 1063 genes in mice, of which 470 were up-regulated and 593 were down-regulated. A total of 801 genes were selected from the cell dataset, consisting of 307 up- regulated and 494 down-regulated genes.

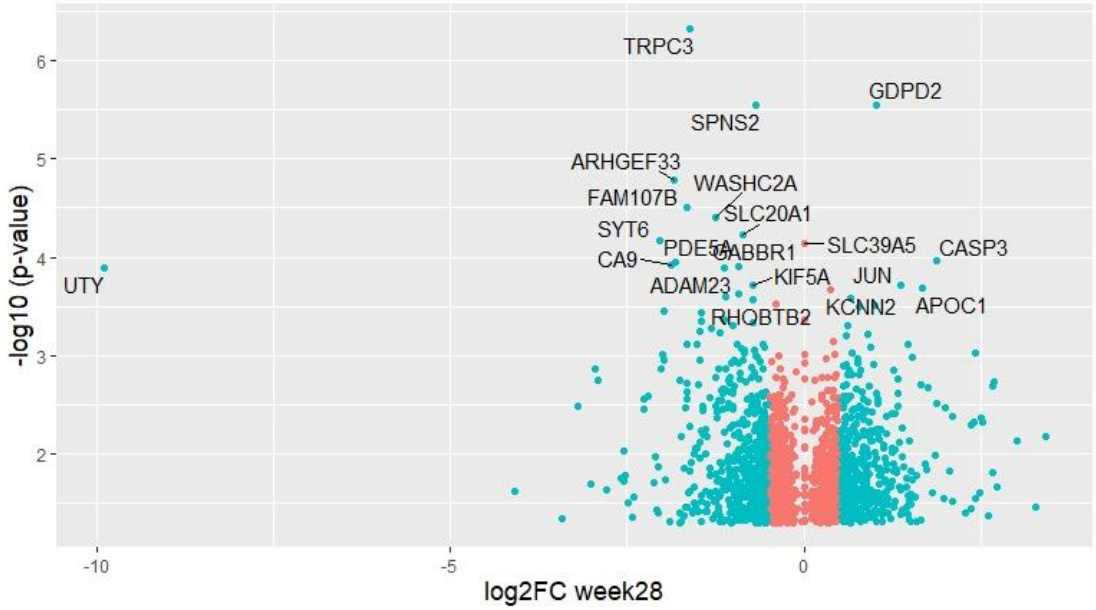

Figure 104: Volcano plot of genes (p-value < 0.05) in the mice RNA-seq dataset at week 28

The volcano plots for both samples (Fig 14-15), indicated that the majority of genes were down-regulated. DEGs with a |log2FC| > 0.5 are shown in blue while, nonsignificant genes are shown in red. The top 20 significant genes are labeled in the plots. However, none of them top were shared between the two datasets.

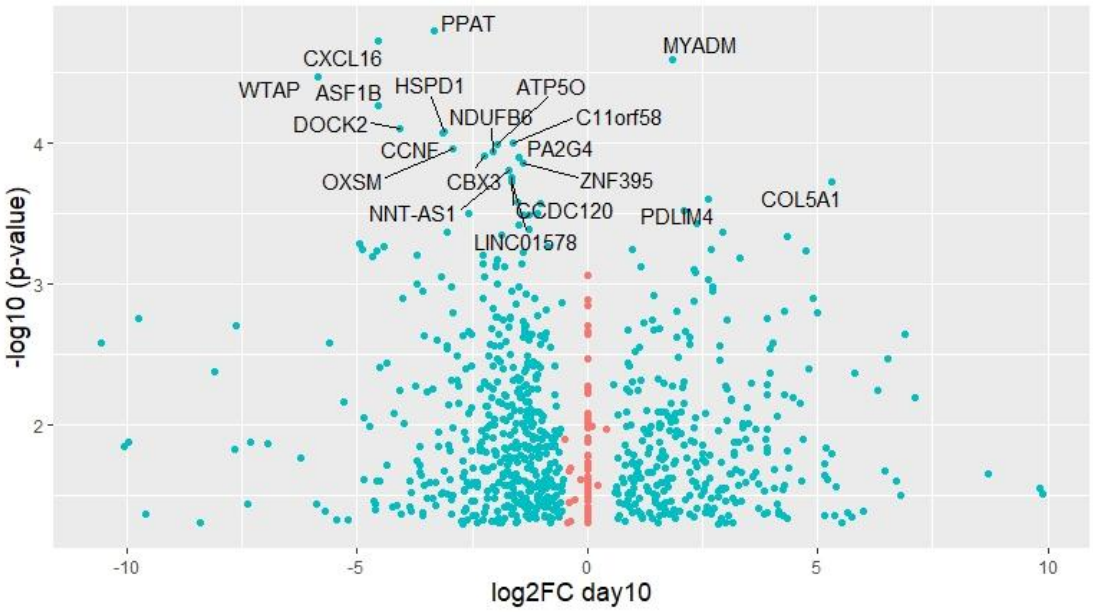

Figure 115: Volcano plot of genes (p-value < 0.05) in cells RNA-seq dataset at day 10

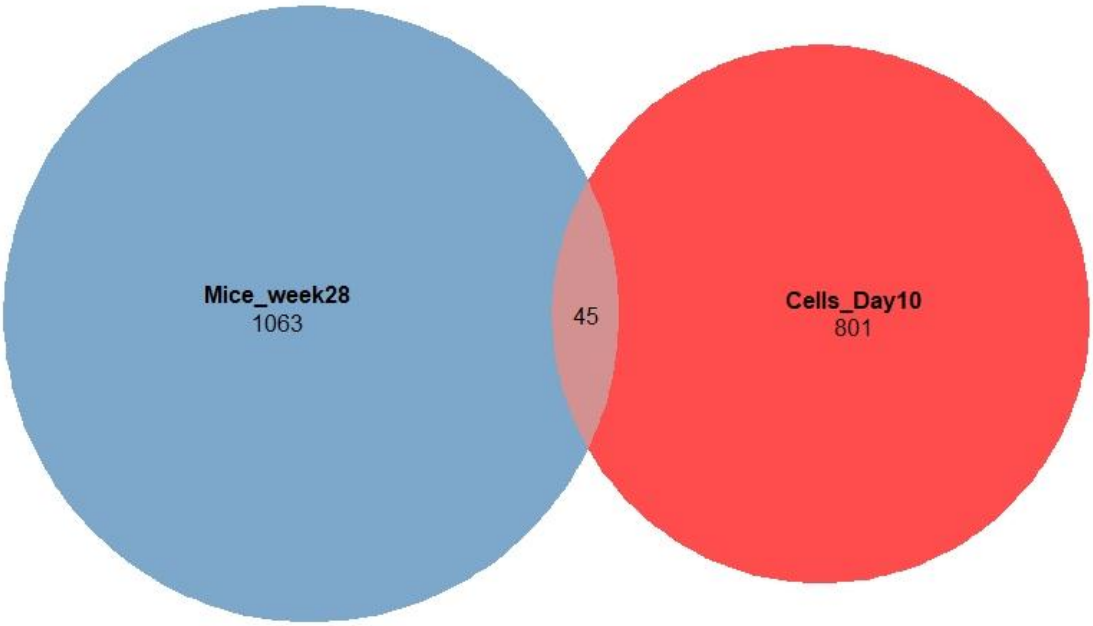

Figure 126: Venn diagram indicates 45 overlapping genes between DEGs of mice at week 28and cells at day 10

A Venn diagram (Fig 16) shows that the dataset of mice at week 28 and cells at day 10 share 45 DEG. These genes are listed in Table 5.

Table 5: List of the common dysregulated genes between mice at week 28 and cells at day 10

Gene symbol	Log2FCcells_D10	FCmice_w28	Gene symbol	Log2FCcells_D10	FCmice_w28
ALDH2	4,01995	1,526289	MASP1	-1,19948	1,509169
ALPK2	1,922516	0,465245	MICALL2	-1,28233	1,420941
ARHGAP35	0,697986	0,68924	MT1X	-2,44932	2,814508
ARL4D	-0,85719	1,899774	NUP93	-1,48011	0,655368
ATF3	-3,49685	2,592387	NUPR1	1,546291	2,15288
BCAR1	2,345642	0,532016	OGT	-2,12476	0,662158
CCDC120	-1,63841	1,421447	OPN3	-1,81258	0,532433
CD74	-3,71045	2,424684	PAXBP1	-2,66235	0,613639
CLDN11	9,81759	1,77024	PLEKHG4	-1,79173	0,568401
COL16A1	4,159534	1,78999	RNASEH2A	-2,65635	1,471403
COL5A1	5,302192	0,166235	RPS13	-2,0745	1,642933
COPZ2	2,288441	2,238164	SDC1	1,561245	1,766427
COTL1	6,300548	1,431714	SERPINE2	1,889727	1,548375
CREG1	-1,97608	0,488844	SLC1A5	1,168965	1,929327
CYGB	-1,92884	1,528308	SLC20A1	1,616554	0,55206
CYR61	1,452894	1,881745	SLC25A36	-2,15903	0,657641
DUSP4	-3,56701	0,380268	TMEM119	4,236712	1,648185
FKBP10	2,682255	1,634965	TNC	1,609077	0,108788
HHIP	-1,61916	0,580852	TRIM37	-3,05601	0,706137
HMGB2	-3,64112	1,792593	TRIM62	0,903225	2,175274
ITM2C	1,862302	1,582011	VGLL3	1,651481	0,362141
LAPTM5	-9,7438	1,57986	WDR62	-4,43234	1,623979
LXN	3,443682	1,657066			

A heatmap was generated from the common DEGs between the two datasets (Fig 17). The Pearson correlation was used to compute distances between genes and samples. Each column corresponds to a dataset and each row to a specific gene. As shown, 22 common genes were up-regulated in both mice and cells while no common gene was down-regulated in both datasets.

The PCA plot showed that the common genes from each dataset, clustered separately, which indicates that their expression pattern in cells and mice was different (Fig 18).

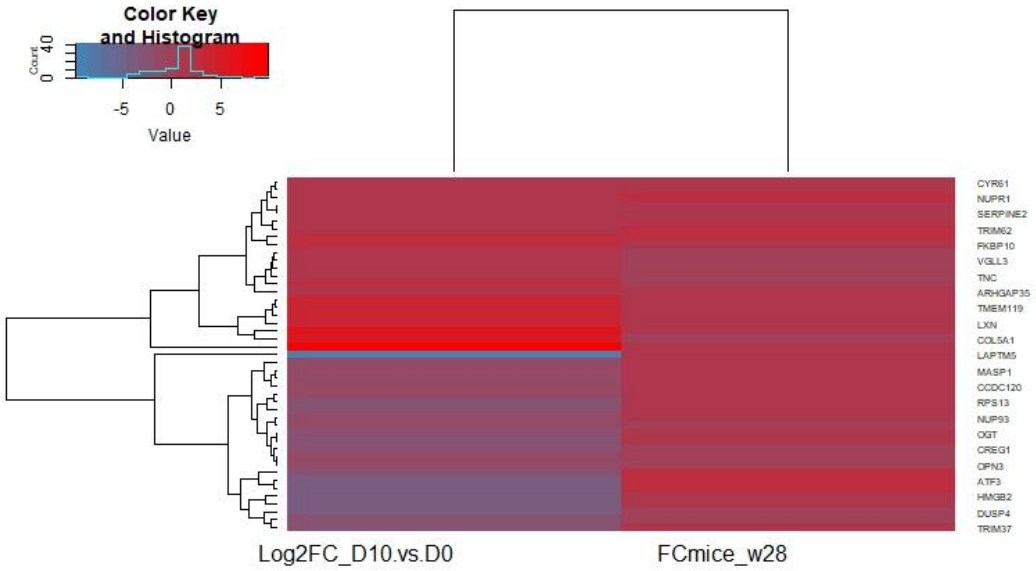

Figure 137: Heatmap showing the log2FC expression of the 45 overlapping genes mice at week 28 and cells at day 10

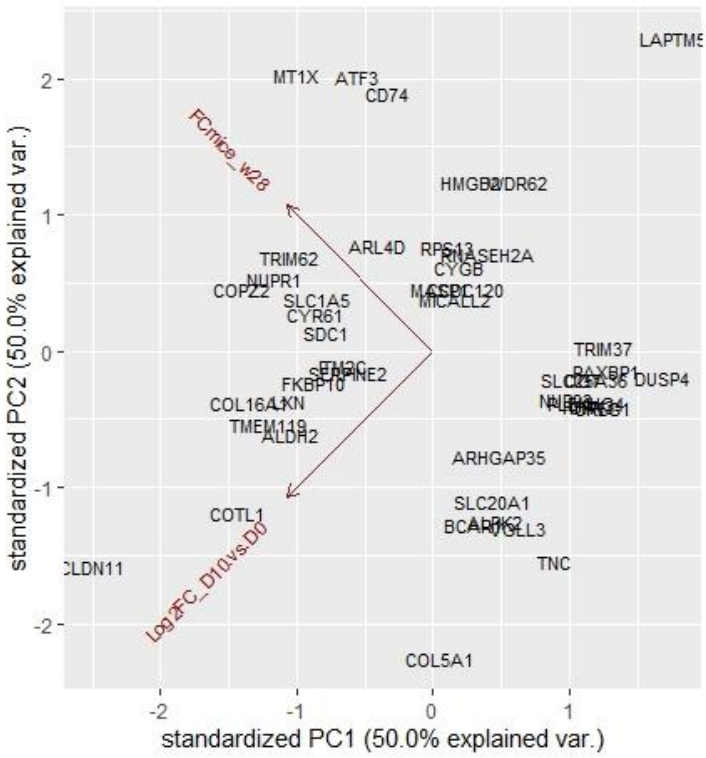

Figure 148: Principal Component Analysis (PCA) of gene expression profiles from mice at week 28 and cells at day 10

3.1.4 Differential expression analysis between Cells-Day 10, Mice-Week 28 and post-mortem human cerebellum

Finally, we compared the RNA-seq datasets from mice and cells at the third time point with the RNA-seq data from a human SCA 1 patient at the end stage of the disease. Due to the lack of biological replicates, gene expression levels in the human tissue were normalized using the GFOLD tool. In total, 2.683 genes were selected with |log2fc| > 0.5. The majority of the genes were down-regulated as also observed in the cells and mice datasets. Only 791 were up-regulated as shown in Fig 19.

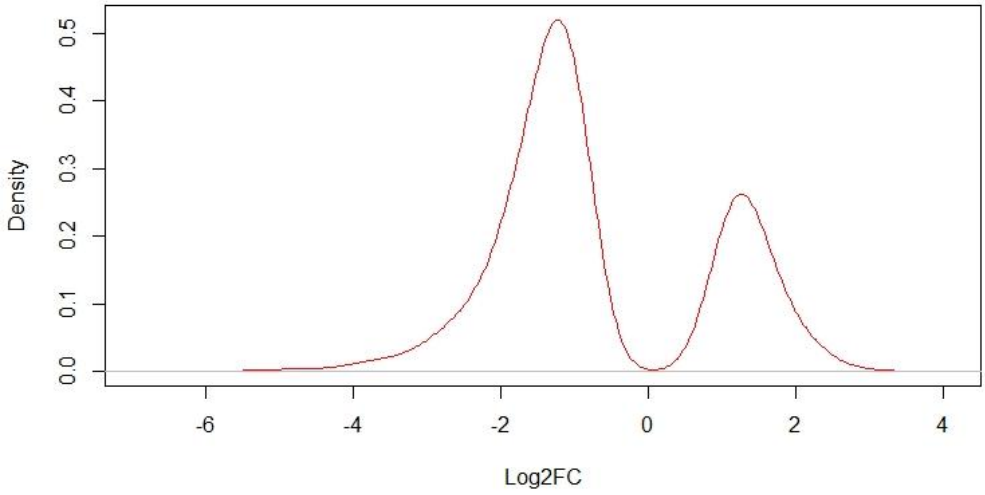

Figure 1915: Density plot of Log2fc in the human RNA-seq-dataset after GFOLD normalization

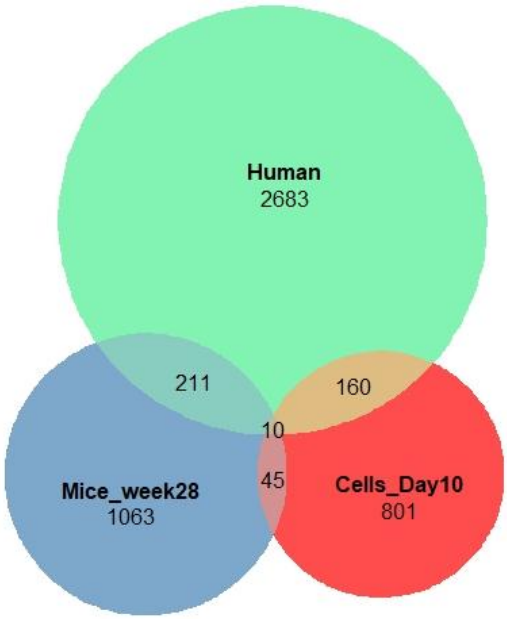

Figure 160: Venn diagram showing 10 overlapping genes between DEGs of mice at week 28, cells at day 10 and human cerebellum at the end stage of the disease

Table 6: List of the common dysregulated genes between mice at week 28, cells at day 10 and human cerebellum at the end stage of disease

Gene symbol	Log2FCcells_D10	LOG2FCmice_w28	Log2FChuman
ALDH2	4,01995	0,610028	-0,90557
BCAR1	2,345642	-0,91046	-0,93997
COTL1	6,300548	0,517743	-1,2512
CYGB	-1,92884	0,611935	-2,32192
ITM2C	1,862302	0,66176	-1,99668
MICALL2	-1,28233	0,506847	-1,16066
MT1X	-2,44932	1,492883	-1,55837
OGT	-2,12476	-0,59475	1,22999
TRIM37	-3,05601	-0,50198	0,858714
TRIM62	0,903225	1,121197	-1,11581

Venn diagram in Fig 20 shows the total number of DEGs per dataset. The common dysregulated genes in all datasets are also listed in Table 6.

The heatmap in Fig 21 shows the expression values of the 10 common genes between the mice, cells and human datasets. None of them was dysregulated at the same direction in all three datasets, based on Pearson's correlations coefficients. PCA plot (Fig 22) indicates that the datasets cluster into three different categories, indicating that their gene expression pattern was different.

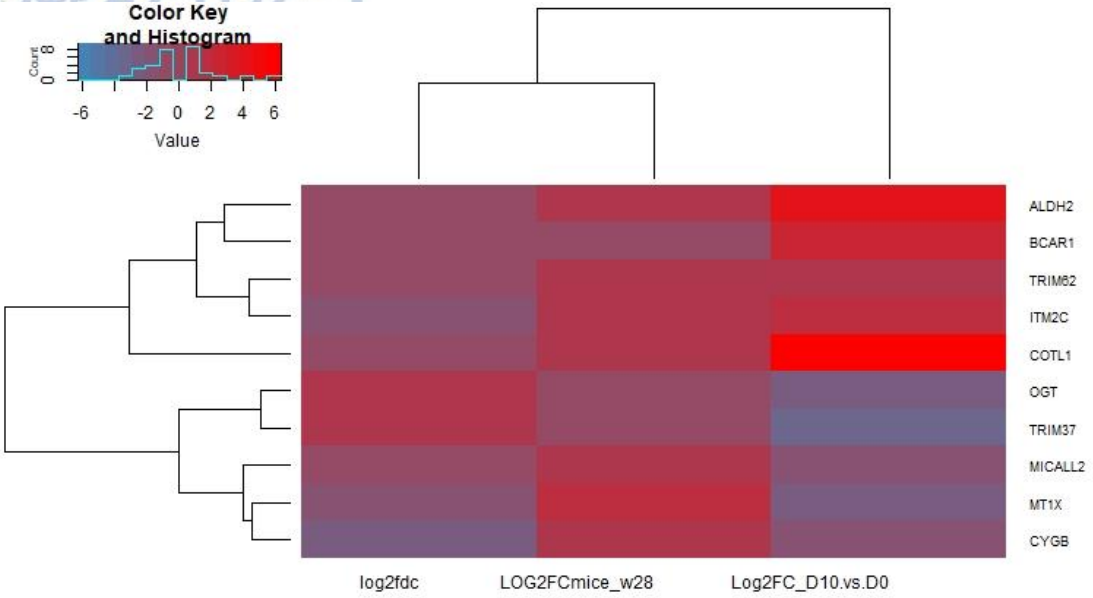

Figure 171: Heatmap showing the log2FC expression of the 10 overlapping mice at week 28, cells at day 10 and human SCA1 cerebellum at the end stage of disease

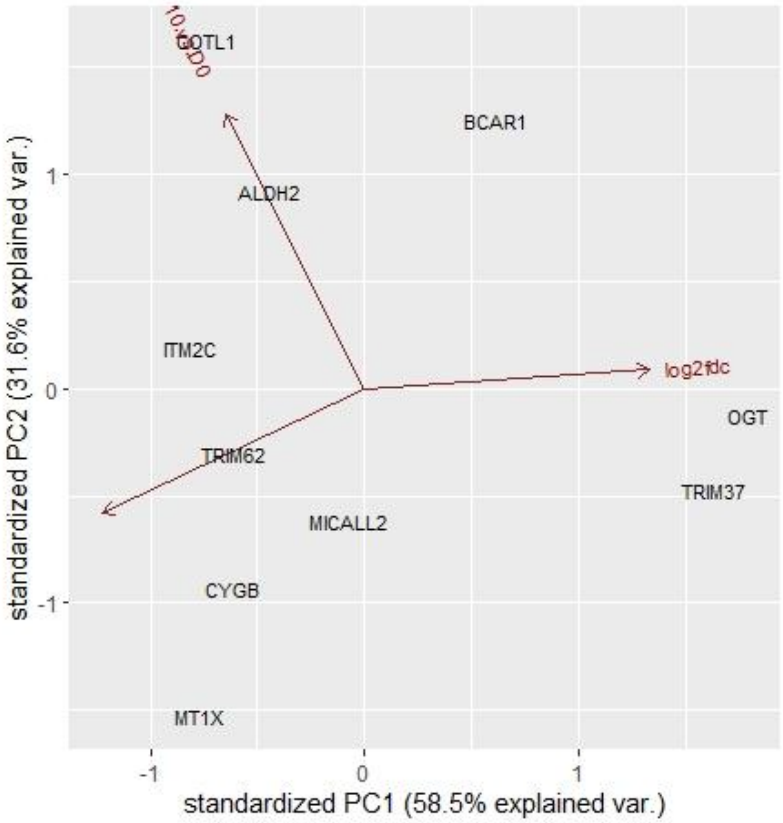

Figure 182: Principal Component Analysis (PCA) of gene expression profiles from mice at week 28, cells at day 10 and human SCA1 cerebellum at the end stage of disease

3.2. Functional enrichment analysis

Very few common genes were identified in the comparisons between the different datasets. We therefore, attempted to identify common dysfunctional pathways by performing pathway enrichment analysis in each individual dataset using the significantly DEGs per time point. Then, the components of the common dysregulated pathways per comparison were used for the construction of perturbed protein Interaction networks.

3.2.1 Enrichment analysis in Cells-Day 2 and in Mice-Week 5

The 687 DEGs from the cells dataset at day 2 were categorized using the KEGG database. This analysis identified 28 pathways (p-value < 0.05) including Ribosome, ECM-receptor interaction, Alzheimer's disease, Focal adhesion and PI3K-Akt signaling pathway. Similarly, the 357 DEGs from the mice at week 5 were categorized in 18 pathways (p-value < 0.05) including: Aldosterone synthesis and secretion, Circadian entrainment, Protein digestion and absorption, Renin secretion, Cholinergic synapse, ECM-receptor interaction, Mucin type O-Glycan biosynthesis, PI3K-Akt signaling pathway. The two datasets shared three common dysregulated pathways, namely Protein digestion and absorption, ECM-receptor interaction and PI3K-Akt signaling pathway. Table 7 shows the common pathways, the number of identified components of the pathways and the p-value of the enrichment analysis in each dataset.

Table 7: Common dysregulated pathways between cells at day 2 and mice at week 5

	CELLS		MICE	
Term	overlap	p-value	overlap	p-value
Protein digestion and absorption	8/90	0,011802818	6/90	0,005421201
ECM-receptor interaction	14/82	6,59884E-07	5/82	0,015525499
PI3K-Akt signaling pathway	26/341	0,000123607	12/341	0,019696666

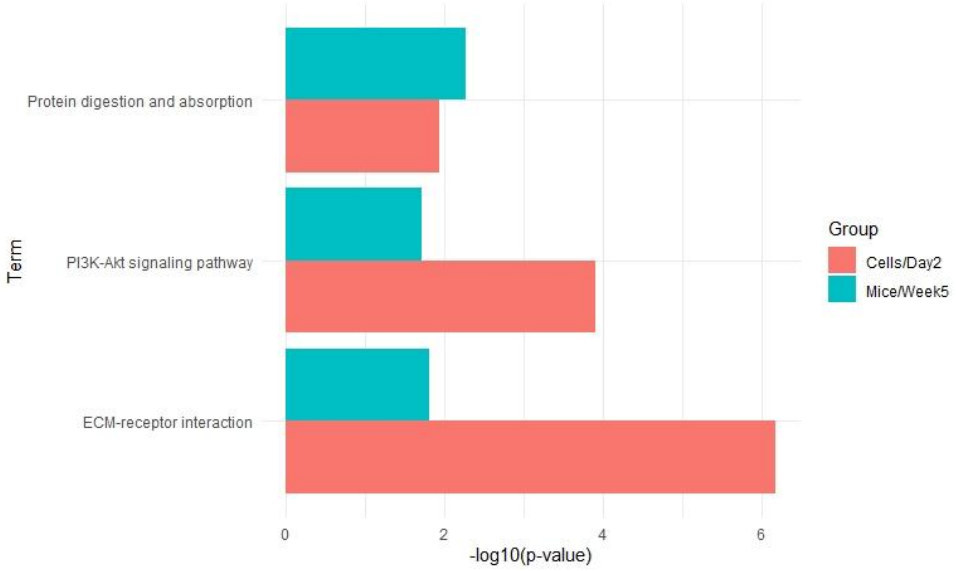

Figure 193: Barplot showing the common dysregulated pathways in cells at day 2 (red color) and mice at week 5 (blue color)

Figure 23 shows the common dysregulated pathways in two datasets. Pathways are shown on the x-axis while, -log of p-values on the y-axis.

3.2.2 Enrichment analysis in Cells-Day 5 and in Mice-Week 12

A total of 789 DEGs from cells at day 5 were categorized in 36 pathways ($p<0.05$), including ECM-receptor interaction, Focal adhesion, Protein digestion and absorption, Proteoglycans in cancer, Regulation of actin cytoskeleton, PI3K-Akt signaling pathway ,Ribosome, DNA replication, Fatty acid biosynthesis, and Cell cycle. Likewise, the 1204 DEGs from mice at week 12 were categorized in 58 pathways ($p<0.05$), including: Calcium signaling pathway, Ribosome, ECM-receptor interaction, Neuroactive ligand-receptor interaction, Focal adhesion, Alzheimer's disease, Rap1 signaling pathway, Protein digestion and absorption, PI3K-Akt signaling pathway and Parkinson's disease. The two datasets share eight common dysregulated pathways as it is shown in Table 8 and Fig 24.

Table 8: Common dysregulated pathways between cells at day 5 and mice at week 12

	CELLS		MICE	
Term	overlap	p-value	overlap	p-value
Ribosome	38/137	4,1454E-22	17/137	0,003367322
ECM-receptor interaction	18/82	2,26798E-09	12/82	0,003378597
Focal adhesion	24/202	1,43813E-06	21/202	0,009694847
PI3K-Akt signaling pathway	29/341	8,69097E-05	30/341	0,022386223
Protein digestion and absorption	12/90	0,000205078	11/90	0,01826595
Alzheimer's disease	15/168	0,002640443	18/168	0,011868882
Rap1 signaling pathway	16/211	0,009370345	21/211	0,015346232
Parkinson's disease	11/142	0,024534867	14/142	0,045101274

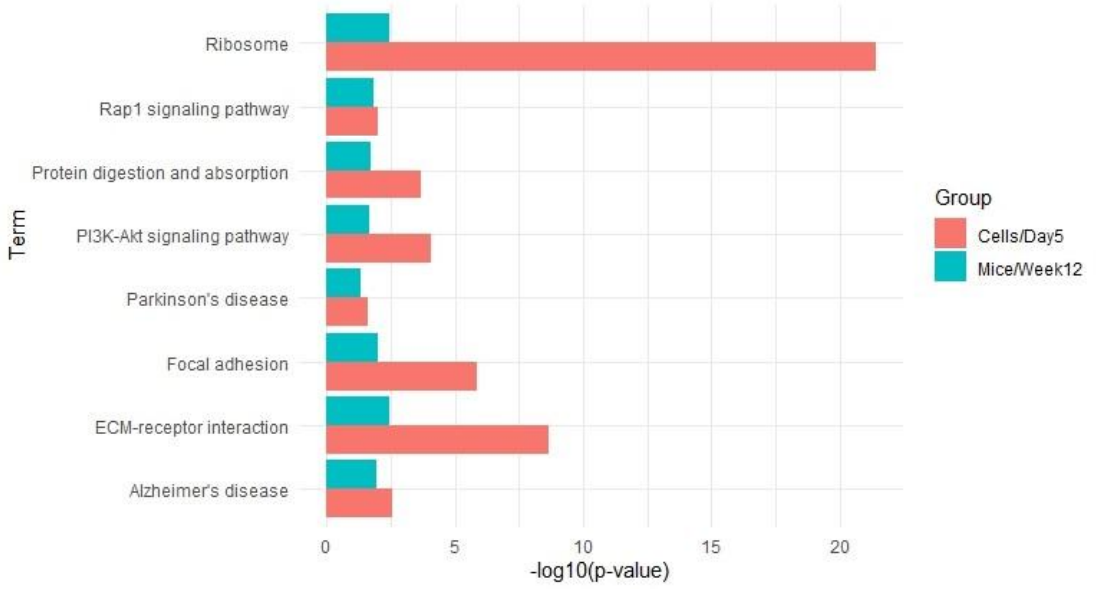

Figure 204: Barplot showing the common dysregulated pathways incells at day 5 (red color) and mice at week 12 (blue color)

3.2.3 Enrichment analysis in Cells-Day 10 and in Mice-Week 28

The 801 DEGs from the cells dataset at day 10 participate in 32 pathways ($p<0.05$) including Ribosome, ECM-receptor interaction, Focal adhesion, Alzheimer's disease, PI3K-Akt signaling pathway, Cell cycle, Rap1 signaling pathway, Parkinson's disease, AGE-RAGE signaling pathway.1.063 DEGs from mice at week 28 were categorized in 32 pathways containing: Rap1 signaling pathway, Regulation of actin cytoskeleton, Phospholipase D signaling pathway, AGE-RAGE signaling pathway in diabetic complications, PI3K-Akt signaling pathway, ECM-receptor interaction and Focal adhesion. Table 9 shows the common pathways, the number of identified components of the pathway and the p-value of the analysis in each dataset. Figure 25 shows the common dysregulated pathways on the x-axis while -log of p-values on the y-axis.

Table 9: Common dysregulated pathways between cells at day 10 and mice at week 28

	CELLS		MICE	
Term	overlap	p-value	overlap	p-value
AGE-RAGE signaling pathway	9/101	0,019702126	11/101	0,018299382
ECM-receptor interaction	13/82	2,08937E-05	9/82	0,029735026
Focal adhesion	21/202	6,22556E-05	17/202	0,041428432
PI3K-Akt signaling pathway	26/341	0,001305403	27/341	0,025640402
Protein digestion and absorption	8/90	0,027403072	9/90	0,049522938
Rap1 signaling pathway	16/211	0,010838329	22/211	0,001997927
Regulation of actin cytoskeleton	18/214	0,002434057	22/214	0,00238481

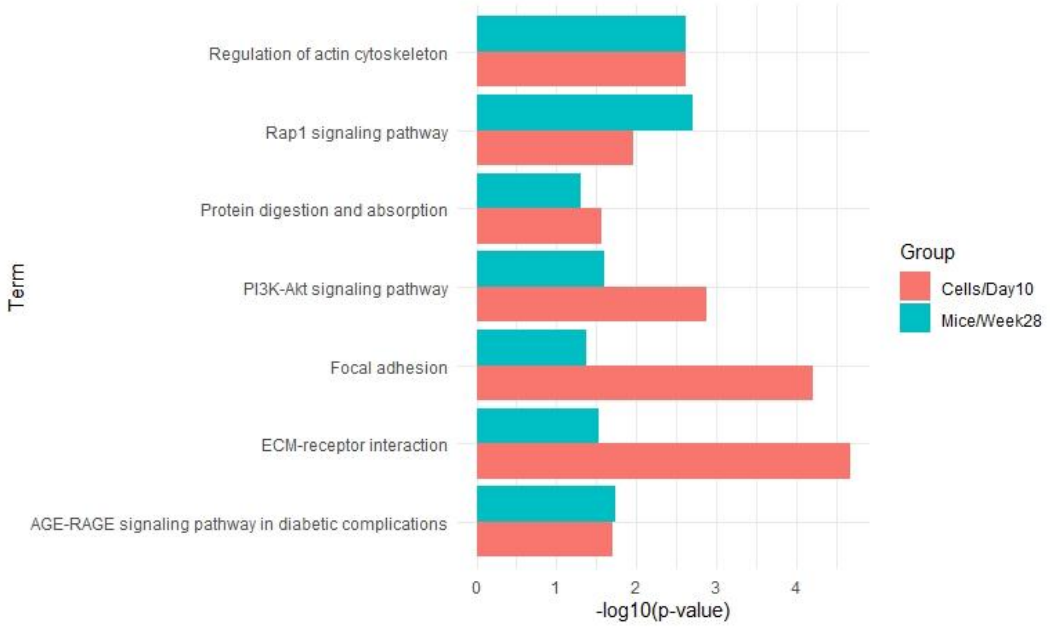

Figure 215: Barplot showing the common dysregulated pathways in cells at day 10 (red color) and mice at week 28 (blue color)

3.2.4 Enrichment analysis in Cells-Day 10, in Mice-Week 28 and in Human

To further explore the biological significance of DEGs in mice at week 28 and cells at day 10 datasets, we analyzed the 2.683 DEGs in human SCA1 cerebellum at the end stage of the disease. The analysis identified 94 pathways ($p<0.05$) including: MAPK signaling pathway, Focal adhesion, Regulation of actin cytoskeleton, Neurotrophin signaling pathway, Glutamatergic synapse, PI3K-Akt signaling pathway, Rap1 signaling pathway, AGE-RAGE signaling pathway in diabetic complications and cAMP signaling pathway. The common dysfunctional pathways between the three datasets are listed in Table 10 while, the bar plot (Fig. 26) shows these pathways on x-axis and -log10 of p-values on y-axis.

Table 10: Common dysregulated pathways between cells at day 10, mice at week 28 and in human SCA1 cerebellum at the end stage of the disease

	CELLS		MICE		HUMAN	
Term	overlap	p-value	overlap	p-value	overlap	p-value
AGE-RAGE signaling pathway	9/101	0,019702126	11/101	0,018299382	26/101	0,000659867
Focal adhesion	21/202	6,22556E-05	17/202	0,041428432	55/202	1,36445E-07
PI3K-Akt signaling pathway	26/341	0,001305403	27/341	0,025640402	68/341	0,000461597
Rap1 signaling pathway	16/211	0,010838329	22/211	0,001997927	46/211	0,000522729
Regulation of actin cytoskeleton	18/214	0,002434057	22/214	0,00238481	53/214	5,24569E-06

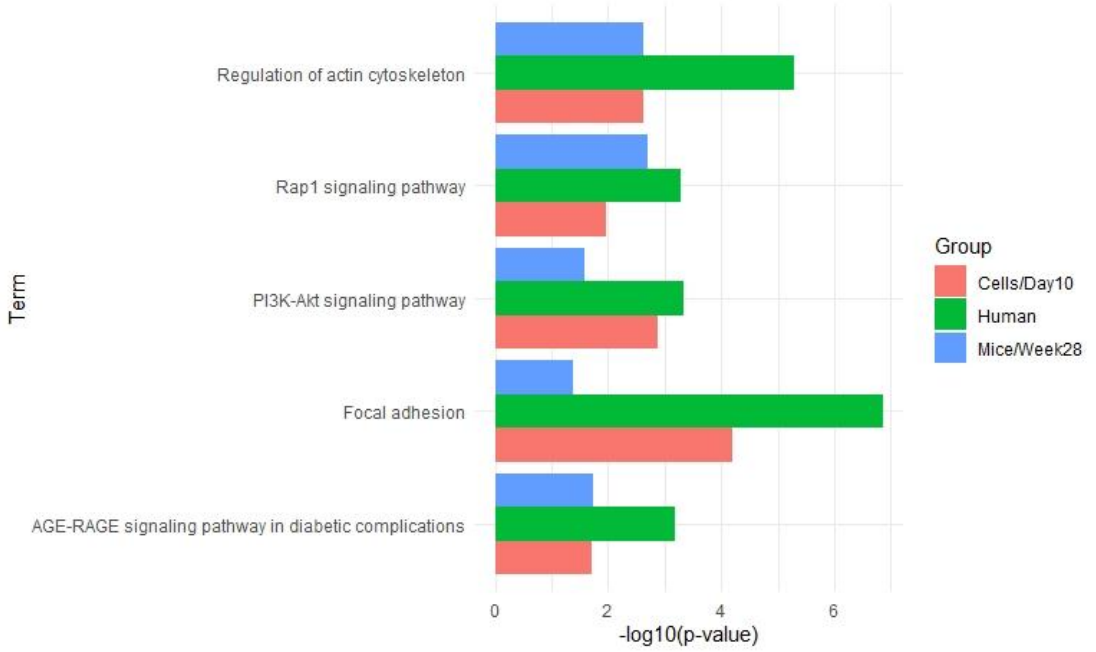

Figure 226: Barplot showing the common dysregulated pathways in cells at day 10 (red color), mice at week 28 (blue color) and human SCA1 cerebellum at the end stage of disease (green color)

3.3 Construction of Protein Interaction Networks

PPIs networks provide a tool to study system biology and may give insights into the cellular molecular mechanism that are affected in a disease condition. In the present study, genes that participate in the commonly dysregulated pathways per comparison were used for the construction of protein interaction networks (shown below).

3.3.1 Protein Interaction Network at early stage of protein aggregation

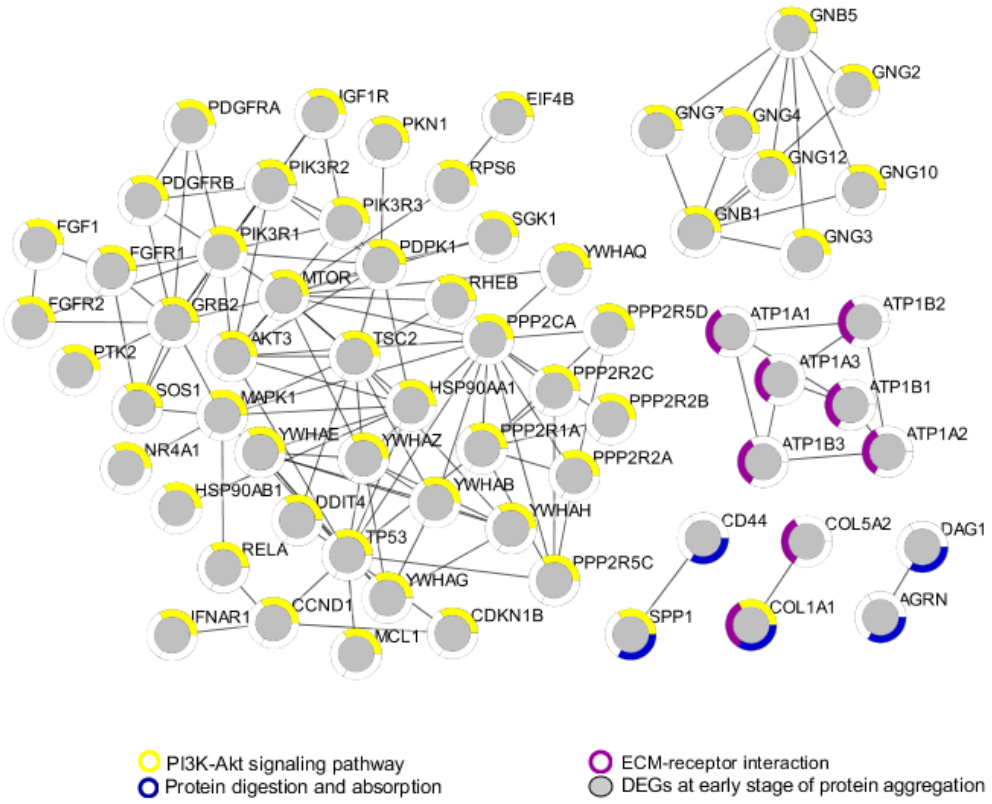

Figure 237: Protein Interaction Network of at early stage of protein aggregation

The PPI network at the early stage of protein aggregation in cells and mice (cell dataset at day 2 and mouse dataset at week 5) consists of 65 nodes (genes) and 131 edges (Fig. 27). The pathway in which each gene participates is shown in color. Their majority are components of PI3K-Akt signaling pathway.

3.3.2 Protein Interaction Network at middle stage of aggregation

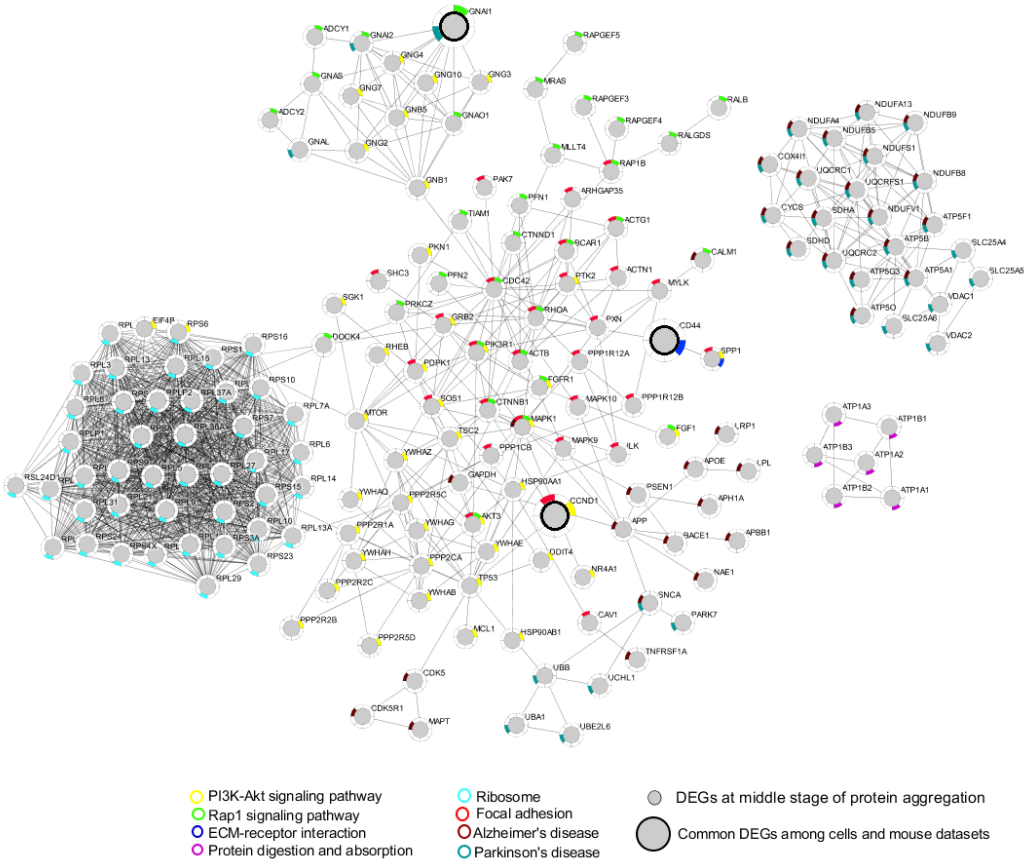

Figure 248: Protein interaction network at middle stage of aggregation

Figure 28 shows the PPI network at the middle stage of protein aggregation (cells at day 5 and mice at week 12). It consists of 177 nodes (genes) and 1241 edges. The bold nodes represent the common DEGs of the network in the cells and mouse RNA-seq datasets. Thus, CCND1, CD44, GNAI1 genes are commonly dysregulated among cells and mouse datasets. CCND1 genes encodes the cyclin D1 protein which participates in PI3K-Akt signaling pathway and Focal adhesion pathways and is up-regulated in both cells and mice datasets. CD44 participates in ECM-receptor interaction pathway and is commonly up-regulated in our data. GNAI1 is component of Parkinson's disease and Rap1 signaling pathways.

3.3.3 Protein Interaction Network at late stage of protein aggregation (cell and mouse datasets)

Figure 259: Protein Interaction Network at late stage of protein aggregation (cell and mouse datasets)

At the late stage of protein aggregation, the PPI network consists of 143 nodes (genes) and 402 edges (Fig 29). The common DEGs in cell and mouse datasets are presented as bold. These are: BCAR1 and AGHGAP35 which are both up-regulated in our data and are part of the Focal adhesion and Regulation of actin cytoskeleton pathways.

3.3.4 Protein Interaction Network at late stage of protein aggregation (cell, mouse and human datasets)

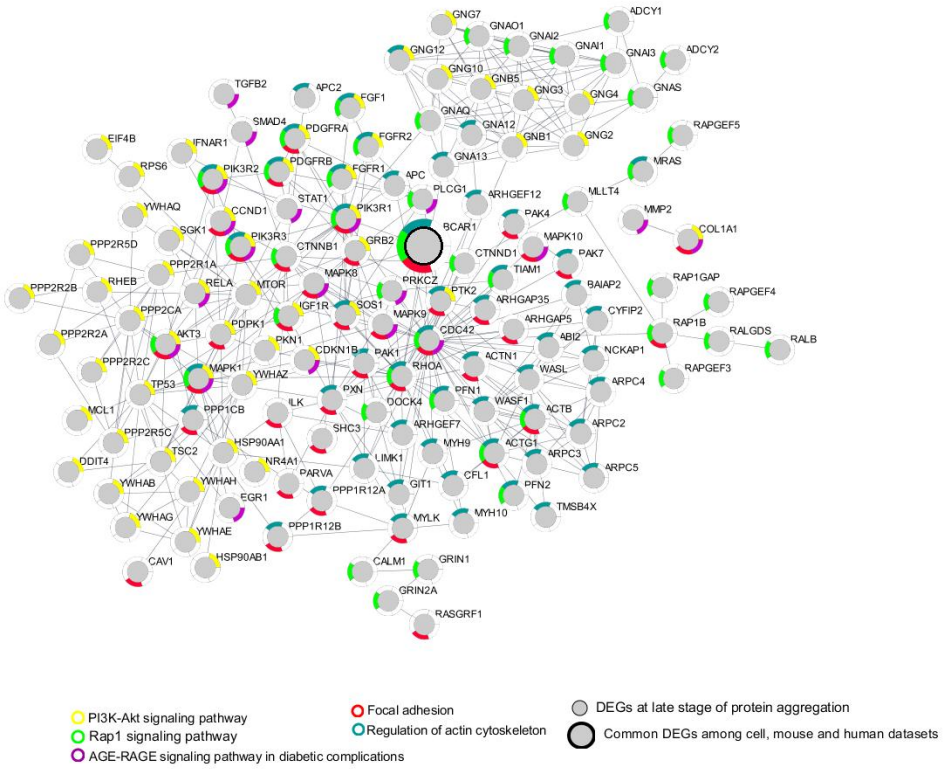

Figure 260: Protein Interaction Network at late stage of protein aggregation (cell, mouse and human datasets)

At the late stage of the disease, the PPI network of the commonly dysregulated pathways between cell, mouse and SCA1 patient, consists 167 nodes (genes) and 956 edges (Fig 30). BCAR1 (shown in bold) is a common DEG gene in all RNA-seq datasets (cells, mice, human). BCAR1 encodes an adaptor protein which participates in Rap1 signaling, Focal adhesion and Regulation of actin cytoskeleton pathways.

3.3.5 Protein Interaction Network at all early, middle and late stage of protein Aggregation

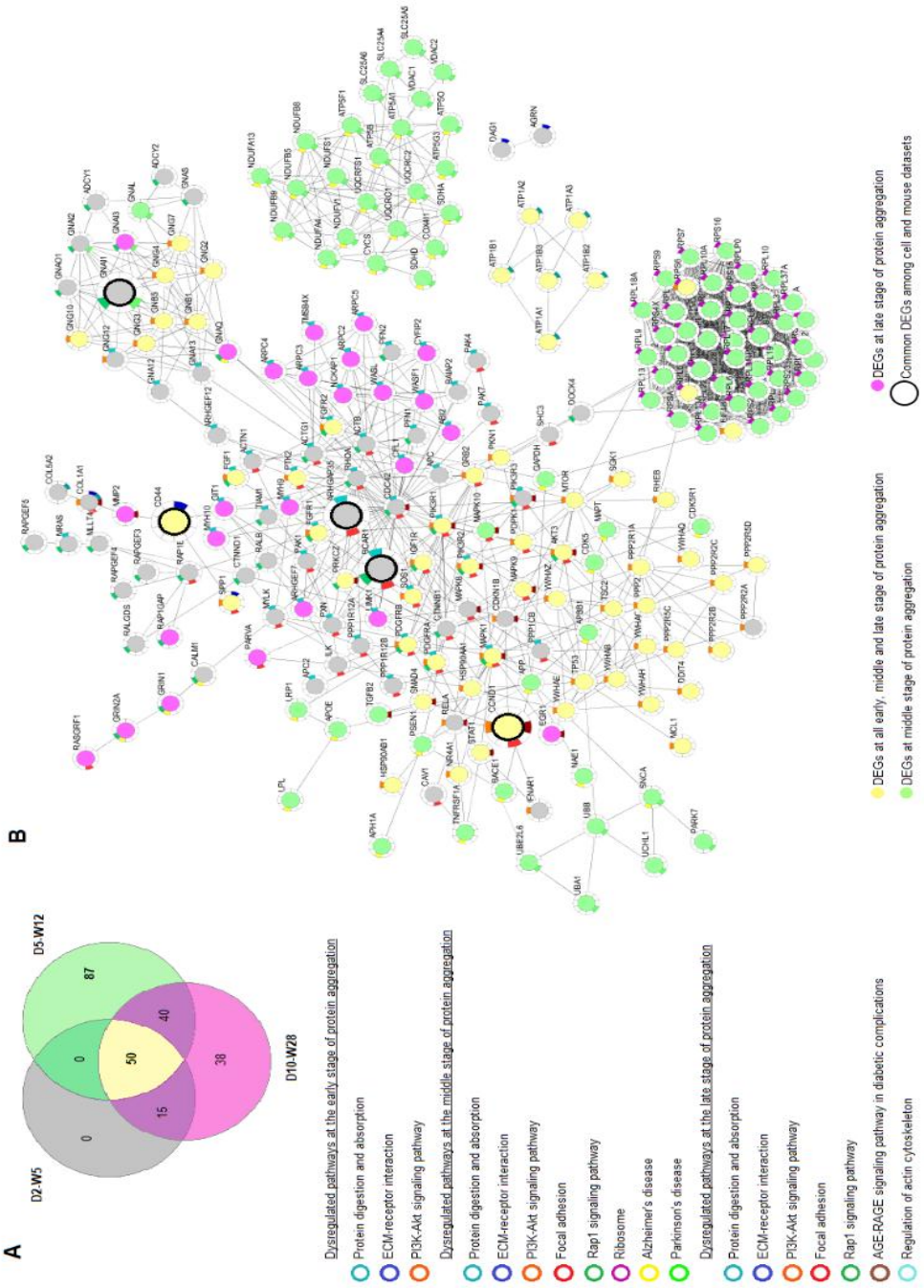

Figure 271: Protein Interaction Network at all early, middle and late stage of protein aggregation (cell and mouse datasets)

Figure 31 shows a total PPI network at early, middle and late stages of the protein aggregation among cell and mouse datasets. Yellow nodes represent the DEGs at all stages of protein aggregation, green nodes the DEGs at middle stage

and purple nodes the DEGs at the late stage of the protein aggregation. Common DEGs among network and RNA-seq datasets (cell and mouse) are presented as bold nodes. These genes are CCND1, CD44, GNAI1, BCAR1 and AGHGAP35.

3.3.6 Network analysis

3.3.6.1 Degree Centrality

The relatively high degree of a nodeindicates that this protein interacts with several proteins. RPS6, CDC42, RPL15 and RPS3 have the highest DC (Table 11). Most of the proteins with the highest DC are components of the Ribosome pathways.

Table 11: Top 10 nodes (proteins) with higher DC value

Gene Symbol	Degree Centrality	Gene Symbol	Degree Centrality
RPS6	45	RPL3	43
CDC42	44	RPL13A	42
RPL15	44	RPL37A	42
RPS3	44	RPL6	42
RPSA	43	RPS3A	42

3.3.6.2 Betweenness Centrality

The top five nodes are CDC42, ATP5B, ATP5A1, MTOR, and DOCK4 (Table 12), indicating that these proteins play a pivotal role in the network. CDC42, has the highest value/status of “mediator.” The shortest pathways of many proteins pass through CDC42, which regulates the flow of information through the network.

Table 12: Top 10 nodes (proteins) with higher BC value

Gene Symbol	Betweenness Centrality	Gene Symbol	Betweenness Centrality
CDC42	0,4064438	UQCRFS1	0,12572079
ATP5B	0,20698295	GNB1	0,10902259
ATP5A1	0,16889568	RHOA	0,10262978

MTOR	0,14118029	ATP1A1	0,1
DOCK4	0,12641788	ATP1A2	0,1

3.3.6.3. Closeness Centrality

Closeness Centrality indicates the degree of proximity between a protein and other proteins. AGRN and DAG1 have the largest CC value in this network (Table 13). These proteins are more closely associated with other proteins, have the average shortest pathway to other proteins as they are in the center at the center of the network. The CC value of ATP proteins are also high.

Table 13: Top 10 nodes (proteins) with higher CC value

Gene Symbol	Closeness Centrality	Gene Symbol	Closeness Centrality
AGRN	1	ATP1A3	0,71428571
DAG1	1	ATP1B1	0,71428571
UQCRFS1	0,71875	ATP1B2	0,71428571
ATP1A1	0,71428571	ATP1B3	0,71428571
ATP1A2	0,71428571	ATP5B	0,6969697

3.3.6.4 Clustering coefficient

The clustering coefficient represents the dense connection between some nodes. Node 1 is connected to the nodes 2 and 3, therefore there is a high possibility that nodes 2 and 3 are also connected. The CU value of several genes (e.g. SLC25A6, SDHD, APC, GIT1, and MYH10) is equal to 1 (Table 14). It shows that the two neighbors interact with each other, forming a group structure which is connected with each other closely.

Table 14: Top 10 nodes (proteins) with higher CU value

Gene Symbol	Clustering coefficient	Gene Symbol	Clustering coefficient
APC	1	PFN2	1
GIT1	1	BAIAP2	1
SLC25A6	1	PAK4	1

SDHD	1	PAK7	1
MYH10	1	YWHAQ	1

Table 15 lists the common DEGs among cell and mouse datasets and their centrality values. ARHGAP35, BCAR1, CCND1, CD44 are commonly up-regulated in RNA-seq datasets.

Table 15: Common DEGs among cell and mouse datasets and their centrality values

Gene Symbol	LOG2FC cell dataset	LOG2FC mouse dataset	Degree Centrality	Betweenness Centrality	Closeness Centrality	Clustering coefficient
ARHGAP35	0,697986	0,68924	3	0,00019704	0,32	0,66666667
BCAR1	2,345642	0,532016	7	0,00155489	0,34188034	0,61904762
CCND1	1,024861	0,617102	6	0,01002992	0,3030303	0,13333333
CD44	1,062387	0,642703	3	0,0307749	0,27322404	0
GNAI1	4,462322	-0,54067	11	0,00024472	0,25094103	0,67272727

4. Discussion

In the current study, we have analyzed RNA-seq datasets obtained from three SCA1-related samples, a cell and a mouse model and human SCA1 cerebellum. These groups were compared in different stages of protein aggregation in order to find similarities in terms of mechanisms that lead to the disease. Following the identification of significant DEGS, we constructed heatmaps and PCAs for each comparison, performed functional enrichment analysis to study the similarities between them and generated protein-protein interaction networks. The number of DEGs were different among the comparisons. The lowest and highest number of DEGs were identified in the comparison between mice at week 28, cells at day 10, human (n=10) and mice at week 12 and cells n day 5 respectively (n=53). This observation highlights that the number of the overlapping DEGs between samples was low.

In order to study disease progression, we found genes that are commonly dysregulated. ATF3 and CREG1 genes are commonly dysregulated at all stages of the protein aggregation. ATF3, a gene for activating transcription factor 3, is also was overexpressed in Huntington cell line (Liang et al. 2009), while CYGB is related with Huntington disease (Kocerha et al. 2013; Mattis et al. 2012). IGFBP 5, MMP14 and THBS1 genes were dysregulated at the early and middle stage of SCA1. Several lines of evidence suggest a down regulation of IGFBP 5 in two spinocerebellar ataxia (SCA) mouse models (for SCA1 and SCA7) (Sanz-Gallego et al. 2014), in mouse model (SCA17) (Friedman et al. 2007) and in Purkinje cells (Ramachandran et al. 2014). MMP14 gene was up-regulated in our results, and also dysregulated in Huntington cellular model systems (Bano et al. 2011). Genes such as ALDH2 and MICALL2 are dysregulated at the middle and late stage, while BCAR1, COTL1, CYGB, ITM2C, MT1X, OGT, TRIM37, TRIM62 genes are dysregulated only at the late stage of the disease among cell, mouse and human SCA1 patient datasets.

ATXN1 is a transcriptional regulator. Therefore, papers describe the transcriptional effect of polyQ-expanted ATXN1 in cerebellum. These studies have identified that various biological pathways, including glutamate signaling, calcium signaling, and long-term depression, are enriched in this tissue at different time-points (Crespo-Barreto et al. 2010; Cvetanovic et al. 2011; Gatchel et al. 2008; Serra et al. 2004). Here, we also aimed to determine the biological role of DEGs and identify commonly affected cellular processes in all datasets.

In our analysis, we identified that protein digestion and absorption, ECM-receptor interaction and PI3K-Akt signaling pathways were dysregulated in all time points between cells and mice datasets. The dysregulation of the PI3K-Akt pathway by polyQ inclusions is not unusual (Matilla-Dueñas et al. 2010). In fact, proteins oligomers were shown to dysregulate cell cycle through the PI3K-Akt pathway (Bhaskar et al. 2009) which promotes necrotic cell death (Wu et al. 2009). Descriptive studies of human neurodegenerative disorders and experimental studies of animal models of neurodegeneration have begun to define potential mechanisms of ECM disruption that can lead to synaptic and neuronal loss. Protein aggregation can be associated with ECM alterations that would result to co-deposition of ECM components. Those ECM alterations can result in loss of protective perineuronal nets (PNNs) and increased susceptibility to cell death (Bonneh-Barkay and Wiley 2009).

The ribosome, Alzheimer's and Parkinson's related pathways were dysregulated in the middle stage of protein aggregation (day5-week12). Ribosomal protein genes, are also highly expressed in Huntington mouse models (Carnemolla et al. 2009). In polyQ neurodegenerative diseases, the expanded CAG RNAs interact directly with nucleolin (NCL), a protein that regulates rRNA transcription (Tsoi et al. 2012). Regulation of actin cytoskeleton and AGE-RAGE signaling pathway were dysregulated at the end stage of the disease. As previous studies have shown RAGE is upregulated in the neurodegenerative process of Huntington disease and correlate with cell death (Deyts et al. 2009; Anzilotti et al. 2012), as huntingtin protein could bind to the RAGE leading to neuronal cell death. Dysregulation of actin dynamics plays a key role in neurodegenerative disorders (Eira et al. 2016). The actin cytoskeleton is strongly regulated by signaling pathways, namely by the Rho GTPase family. In Huntington disease huntingtin protein interacts with several players of the Rho GTPase signaling pathways (Tourette et al. 2014).

Protein Interaction networks provide a tool to study the cellular molecular mechanism that are affected in a disease condition The protein products of genes that participate in the commonly dysregulated pathways per comparison were used for the construction of a protein-protein interaction network During the last years, network studies have been applied to biological data indicating that the degree of connectivity is a key property of any network (Jeong et al. 2001). The most common approach to identify key nodes in a network is to search for the most connected nodes (hubs). The underlying assumption was that these hubs could be critical to explain the pathogenesis of diseases. In our results, CDC42 and genes that are parts

of the Ribosome pathways have the higher DC. Previous studies have shown that CDC42 appears to function as an initiator of neuronal cell death (Bazenet, Mota, and Rubin 1998) and it is involved in the pathology of Huntington's disease (Li and Li 2004), while ribosomal proteins have been shown to alter the aggregation of polyQ proteins in animal models (Williams and Paulson 2008; Nollen et al. n.d.)

Betweenness Centrality is another key indicator that demonstrates nodes which may be relevant in a network (Yu et al. 2007; Joy et al. 2005). In our data, CDC42, the protein with the higher DC, has also the higher BC, indicating that this node play an important role in the network. Proteins with higher CC values (e.g. AGRN, DAG1, UQCRFS1 and ATP proteins) are components of the three clusters in network. In the large cluster, the node with the higher CC is the CDC42 protein. Furthermore, within the interaction network, essential proteins also tend to be more cliquish (as determined from the clustering coefficient) (Yu et al. 2004). Proteins with high CU values are APC and GIT1 which are related with ESBM (Bott et al. 2016) and Huntington diseases (Goehler et al. 2004) respectively.

5. Conclusion

Our study focuses on the comparison of three RNA-seq datasets from SCA1 related samples; namely, a cell model, a mouse model and a human SCA1 cerebellum, in order to find similarities in terms of mechanisms that lead to the disease. Following the identification of significant DEGS, we constructed heatmaps and PCAs for each comparison, performed functional enrichment analysis and generated protein-protein interaction networks. The pathways that found to be commonly dysregulated among the datasets at all the stages of protein aggregation are: protein digestion and absorption, ECM-receptor interaction (cells, mice) and PI3K-Akt (cells, mice, human) signaling pathway. Further studies are required to examine the detailed molecular mechanisms, underlying the various biological effects of the components of these pathways in SCA1.

References

Anzilotti, Serenella, Carmela Giampà, Daunia Laurenti, et al.
2012 Immunohistochemical Localization of Receptor for Advanced Glycation End (RAGE) Products in the R6/2 Mouse Model of Huntington's Disease. Brain Research Bulletin 87(2–3): 350–358.

Assenov, Yassen, Fidel Ramírez, Sven-Eric Schelhorn, Thomas Lengauer, and Mario Albrecht
2008 Computing Topological Parameters of Biological Networks. Bioinformatics 24(2): 282–284.

Backman, Tyler W. H., and Thomas Girke
2016 SystemPipeR: NGS Workflow and Report Generation Environment. BMC Bioinformatics 17(1): 388.

Bano, D, F Zanetti, Y Mende, and P Nicotera
2011 Neurodegenerative Processes in Huntington's Disease. Cell Death & Disease 2(11): e228–e228.

Bazenet, C. E., M. A. Mota, and L. L. Rubin
1998 The Small GTP-Binding Protein Cdc42 Is Required for Nerve Growth Factor Withdrawal-Induced Neuronal Death. Proceedings of the National Academy of Sciences 95(7): 3984–3989.

Bhaskar, Kiran, Megan Miller, Alexandra Chludzinski, et al.
2009 The PI3K-Akt-MTOR Pathway Regulates Aβ Oligomer Induced Neuronal Cell Cycle Events. Molecular Neurodegeneration 4(1): 14.

Bonneh-Barkay, Dafna, and Clayton A. Wiley
2009 Brain Extracellular Matrix in Neurodegeneration. Brain Pathology 19(4): 573–585.

Bott, Laura C., Florian A. Salomons, Dragan Maric, et al.
2016 The Polyglutamine-Expanded Androgen Receptor Responsible for Spinal and Bulbar Muscular Atrophy Inhibits the APC/CCdh1 Ubiquitin Ligase Complex. Scientific Reports 6(1): 27703.

Burright, Eric N, H Brent Clark, Antonio Servadio, et al.
1995 SCA1 Transgenic Mice: A Model for Neurodegeneration Caused by an Expanded CAG Trinucleotide Repeat. Cell 82(6): 937–948.

Carnemolla, Alisia, Elisa Fossale, Elena Agostoni, et al.
2009 Rrs1 Is Involved in Endoplasmic Reticulum Stress Response in Huntington Disease. Journal of Biological Chemistry 284(27): 18167–18173.

Chen, Hung-Kai, Pedro Fernandez-Funez, Summer F. Acevedo, et al.
2003 Interaction of Akt-Phosphorylated Ataxin-1 with 14-3-3 Mediates Neurodegeneration in Spinocerebellar Ataxia Type 1. Cell 113(4): 457–468.

de Chiara, Cesira, Rajesh P. Menon, Molly Strom, Toby J. Gibson, and Annalisa Pastore

2009 Phosphorylation of S776 and 14-3-3 Binding Modulate Ataxin-1 Interaction with Splicing Factors. Yue Feng, ed. PLoS ONE 4(12): e8372.

Crespo-Barreto, Juan, John D. Fryer, Chad A. Shaw, Harry T. Orr, and Huda Y. Zoghbi
2010 Partial Loss of Ataxin-1 Function Contributes to Transcriptional Dysregulation in Spinocerebellar Ataxia Type 1 Pathogenesis. Gregory S. Barsh, ed. PLoS Genetics 6(7): e1001021.

Cvetanovic, Marija, Jay M Patel, Hugo H Marti, Ameet R Kini, and Puneet Opal
2011 Vascular Endothelial Growth Factor Ameliorates the Ataxic Phenotype in a Mouse Model of Spinocerebellar Ataxia Type 1. Nature Medicine 17(11): 1445–1447.

Deyts, Carole, Beatriz Galan-Rodriguez, Elodie Martin, et al.
2009 Dopamine D2 Receptor Stimulation Potentiates PolyQ-Huntingtin-Induced Mouse Striatal Neuron Dysfunctions via Rho/ROCK-II Activation. Howard E. Gendelman, ed. PLoS ONE 4(12): e8287.

Eira, Jessica, Catarina Santos Silva, Mónica Mendes Sousa, and Márcia Almeida Liz
2016 The Cytoskeleton as a Novel Therapeutic Target for Old Neurodegenerative Disorders. Progress in Neurobiology 141: 61–82.

Emamian, Effat S., Michael D. Kaytor, Lisa A. Duvick, et al.
2003 Serine 776 of Ataxin-1 Is Critical for Polyglutamine-Induced Disease in SCA1 Transgenic Mice. Neuron 38(3): 375–387.

Feng, Jianxing, Clifford A. Meyer, Qian Wang, et al.
2012 GFOLD: A Generalized Fold Change for Ranking Differentially Expressed Genes from RNA-Seq Data. Bioinformatics 28(21): 2782–2788.

Freeman, Linton C.
1978 Centrality in Social Networks Conceptual Clarification. Social Networks 1(3): 215–239.

Friedman, Meyer J, Anjali G Shah, Zhi-Hui Fang, et al.
2007 Polyglutamine Domain Modulates the TBP-TFIIB Interaction: Implications for Its Normal Function and Neurodegeneration. Nature Neuroscience 10(12): 1519–1528.

Gatchel, J. R., K. Watase, C. Thaller, et al.
2008 The Insulin-like Growth Factor Pathway Is Altered in Spinocerebellar Ataxia Type 1 and Type 7. Proceedings of the National Academy of Sciences 105(4): 1291–1296.

Goehler, Heike, Maciej Lalowski, Ulrich Stelzl, et al.
N.d. A Protein Interaction Network Links GIT1, an Enhancer of Huntingtin Aggregation, to Huntington's Disease: 13.

Hu, Yanhui, Ian Flockhart, Arunachalam Vinayagam, et al.
2011 An Integrative Approach to Ortholog Prediction for Disease-Focused and Other Functional Studies. BMC Bioinformatics 12(1): 357.

Iida, M., K. Sahashi, N. Kondo, et al.

2017 Akt Signaling Pathway Is Dysregulated in Polyglutamine Diseases. Journal of the Neurological Sciences 381: 209.

Ingram, Melissa, Emily A.L. Wozniak, Lisa Duvick, et al.
2016a Cerebellar Transcriptome Profiles of ATXN1 Transgenic Mice Reveal SCA1 Disease Progression and Protection Pathways. Neuron 89(6): 1194–1207.
2016b Cerebellar Transcriptome Profiles of ATXN1 Transgenic Mice Reveal SCA1 Disease Progression and Protection Pathways. Neuron 89(6): 1194–1207.

Irwin, Stuart, Mark Vandelft, Deborah Pinchev, et al.
2005 RNA Association and Nucleocytoplasmic Shuttling by Ataxin-1. Journal of Cell Science 118(1): 233–242.

Jeong, H., S. P. Mason, A.-L. Barabási, and Z. N. Oltvai
2001 Lethality and Centrality in Protein Networks. Nature 411(6833): 41–42.

Joy, Maliackal Poulo, Amy Brock, Donald E. Ingber, and Sui Huang
2005 High-Betweenness Proteins in the Yeast Protein Interaction Network. Journal of Biomedicine and Biotechnology 2005(2): 96–103.

Kanehisa, Minoru, and Susumu Goto
N.d. KEGG: Kyoto Encyclopedia of Genes and Genomes: 4.

Kocerha, Jannet, Yuhong Liu, David Willoughby, et al.
2013 Longitudinal Transcriptomic Dysregulation in the Peripheral Blood of Transgenic Huntington's Disease Monkeys. BMC Neuroscience 14(1): 88.

Krol, Hilde A., Przemek M. Krawczyk, Klazien S. Bosch, et al.
2008 Polyglutamine Expansion Accelerates the Dynamics of Ataxin-1 and Does Not Result in Aggregate Formation. Mark Cookson, ed. PLoS ONE 3(1): e1503.

Kuleshov, Maxim V., Matthew R. Jones, Andrew D. Rouillard, et al.
2016 Enrichr: A Comprehensive Gene Set Enrichment Analysis Web Server 2016 Update. Nucleic Acids Research 44(W1): W90–W97.

Lam, Yung C., Aaron B. Bowman, Paymaan Jafar-Nejad, et al.
2006 ATAXIN-1 Interacts with the Repressor Capicua in Its Native Complex to Cause SCA1 Neuropathology. Cell 127(7): 1335–1347.

Li, Shi-Hua, and Xiao-Jiang Li
2004 Huntingtin–Protein Interactions and the Pathogenesis of Huntington's Disease. Trends in Genetics 20(3): 146–154.

Liang, Yideng, Haibing Jiang, Tamara Ratovitski, et al.
2009 ATF3 Plays a Protective Role against Toxicity by N-Terminal Fragment of Mutant Huntingtin in Stable PC12 Cell Line. Brain Research 1286: 221–229.

Lim, Janghoo, Juan Crespo-Barreto, Paymaan Jafar-Nejad, et al.
2008 Opposing Effects of Polyglutamine Expansion on Native Protein Complexes Contribute to SCA1. Nature 452(7188): 713–718.

Lu, Hsiang-Chih, Qiumin Tan, Maxime W C Rousseaux, et al.
2017 Disruption of the ATXN1–CIC Complex Causes a Spectrum of Neurobehavioral Phenotypes in Mice and Humans. Nature Genetics 49(4): 527–536.

Matilla-Dueñas, Antoni, Robert Goold, and Paola Giunti
2008 Clinical, Genetic, Molecular, and Pathophysiological Insights into Spinocerebellar Ataxia Type 1. The Cerebellum 7(2): 106–114.

Matilla-Dueñas, Antoni, Ivelisse Sánchez, Marc Corral-Juan, et al.
2010 Cellular and Molecular Pathways Triggering Neurodegeneration in the Spinocerebellar Ataxias. The Cerebellum 9(2): 148–166.

Menon, Rajesh P., Suran Nethisinghe, Serena Faggiano, et al.
2013 The Role of Interruptions in PolyQ in the Pathology of SCA1. Christopher E. Pearson, ed. PLoS Genetics 9(7): e1003648.

Morrison, Deborah K.
2009 The 14-3-3 Proteins: Integrators of Diverse Signaling Cues That Impact Cell Fate and Cancer Development. Trends in Cell Biology 19(1): 16–23.

Nieminen, Juhani
1974 On the Centrality in a Graph. Scandinavian Journal of Psychology 15(1): 332–336.

Nollen, Ellen A A, Susana M Garcia, Gijs van Haaften, et al.
N.d. Genome-Wide RNA Interference Screen Identifies Previously Undescribed Regulators of Polyglutamine Aggregation: 6.

Orr, Harry T.
2012a Cell Biology of Spinocerebellar Ataxia. The Journal of Cell Biology 197(2): 167–177.
2012b SCA1—Phosphorylation, a Regulator of Ataxin-1 Function and Pathogenesis. Progress in Neurobiology 99(3): 179–185.

Ramachandran, Pavitra S, Ryan L Boudreau, Kellie A Schaefer, Albert R La Spada, and Beverly L Davidson
2014 Nonallele Specific Silencing of Ataxin-7 Improves Disease Phenotypes in a Mouse Model of SCA7. Molecular Therapy 22(9): 1635–1642.

Rousseaux, Maxime W.C., Tyler Tschumperlin, Hsiang-Chih Lu, et al.
2018 ATXN1-CIC Complex Is the Primary Driver of Cerebellar Pathology in Spinocerebellar Ataxia Type 1 through a Gain-of-Function Mechanism. Neuron 97(6): 1235-1243.e5.

Sabidussi, Gert
1966 The Centrality Index of a Graph. Psychometrika 31(4): 581–603.

Santos, Alberto, Kalliopi Tsafou, Christian Stolte, et al.
2015 Comprehensive Comparison of Large-Scale Tissue Expression Datasets. PeerJ 3: e1054.

Sanz-Gallego, Irene, Francisco J Rodriguez-de-Rivera, Irene Pulido, Ignacio Torres-Aleman, and Javier Arpa

2014 IGF-1 in Autosomal Dominant Cerebellar Ataxia - Open-Label Trial. Cerebellum & Ataxias 1(1): 13.

Schorderet, Patrick
2016 NEAT: A Framework for Building Fully Automated NGS Pipelines and Analyses. BMC Bioinformatics 17(1): 53.

Serra, Heliane G., Courtney E. Byam, Jeffrey D. Lande, et al.
2004 Gene Profiling Links SCA1 Pathophysiology to Glutamate Signaling in Purkinje Cells of Transgenic Mice. Human Molecular Genetics 13(20): 2535–2543.

Serra, Heliane G., Lisa Duvick, Tao Zu, et al.
2006 RORα-Mediated Purkinje Cell Development Determines Disease Severity in Adult SCA1 Mice. Cell 127(4): 697–708.

Shannon, P.
2003 Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Research 13(11): 2498–2504.

Shao, J., and M. I. Diamond
2007 Polyglutamine Diseases: Emerging Concepts in Pathogenesis and Therapy. Human Molecular Genetics 16(R2): R115–R123.

Siska, Evangelia K., George Koliakos, and Spyros Petrakis
2015 Stem Cell Models of Polyglutamine Diseases and Their Use in Cell-Based Therapies. Frontiers in Neuroscience 9.
http://journal.frontiersin.org/Article/10.3389/fnins.2015.00247/abstract, accessed July 9, 2019.

Srinivasan, Sharan R., and Vikram G. Shakkottai
2019 Moving Towards Therapy in SCA1: Insights from Molecular Mechanisms, Identification of Novel Targets, and Planning for Human Trials. Neurotherapeutics.
http://link.springer.com/10.1007/s13311-019-00763-y, accessed August 13, 2019.

Sullivan, Roisin, Wai Yan Yau, Emer O'Connor, and Henry Houlden
2019 Spinocerebellar Ataxia: An Update. Journal of Neurology 266(2): 533–544.

Sultan, Ghazala, and Dr Swaleha Zubair
2019 BIOINFORMATICS APPROACHES FOR BIG DATA ANALYTICS IN PRECISION MEDICINE: AN OVERVIEW: 10.

Szklarczyk, Damian, John H Morris, Helen Cook, et al.
2017 The STRING Database in 2017: Quality-Controlled Protein–Protein Association Networks, Made Broadly Accessible. Nucleic Acids Research 45(D1): D362–D368.

Tourette, Cendrine, Biao Li, Russell Bell, et al.
2014 A Large Scale Huntingtin Protein Interaction Network Implicates Rho GTPase Signaling Pathways in Huntington Disease. Journal of Biological Chemistry 289(10): 6709–6726.

Trapnell, Cole, Brian A Williams, Geo Pertea, et al.

2010 Transcript Assembly and Quantification by RNA-Seq Reveals Unannotated Transcripts and Isoform Switching during Cell Differentiation. Nature Biotechnology 28(5): 511–515.

Tsoi, H., T. C.-K. Lau, S.-Y. Tsang, K.-F. Lau, and H. Y. E. Chan
2012 CAG Expansion Induces Nucleolar Stress in Polyglutamine Diseases. Proceedings of the National Academy of Sciences 109(33): 13428–13433.

Tsuda, Hiroshi, Hamed Jafar-Nejad, Akash J. Patel, et al.
2005 The AXH Domain of Ataxin-1 Mediates Neurodegeneration through Its Interaction with Gfi-1/Senseless Proteins. Cell 122(4): 633–644.

Wang, Tianyu, Boyang Li, Craig E. Nelson, and Sheida Nabavi
2019 Comparative Analysis of Differential Gene Expression Analysis Tools for Single-Cell RNA Sequencing Data. BMC Bioinformatics 20(1): 40.

Watts, Duncan J, and Steven H Strogatz
1998 Collective Dynamics of 'Small-World' Networks 393: 3.

Williams, Aislinn J., and Henry L. Paulson
2008 Polyglutamine Neurodegeneration: Protein Misfolding Revisited. Trends in Neurosciences 31(10): 521–528.

Yu, Haiyuan, Dov Greenbaum, Hao Xin Lu, Xiaowei Zhu, and Mark Gerstein
2004 Genomic Analysis of Essentiality within Protein Networks. Trends in Genetics 20(6): 227–231.

Yu, Haiyuan, Philip M Kim, Emmett Sprecher, Valery Trifonov, and Mark Gerstein
2007 The Importance of Bottlenecks in Protein Networks: Correlation with Gene Essentiality and Expression Dynamics. PLoS Computational Biology 3(4): 9.

Zoghbi, Huda Y., Marilyn S. Pollack, Leslie A. Lyons, et al.
1988 Spinocerebellar Ataxia: Variable Age of Onset and Linkage to Human Leukocyte Antigen in a Large Kindred. Annals of Neurology 23(6): 580–584.