

# Computational Transcriptomic Profiling of Hamstring Muscle Contractures in Cerebral Palsy

Sofia Pietrini

## Abstract

Cerebral palsy (CP) is a neurodevelopmental disorder characterized by motor impairments and frequently accompanied by muscle contractures. Although not of genetic origin, these contractures are believed to arise from secondary muscle adaptations to upper motor neuron lesions. To uncover the molecular mechanisms underlying this process, we conducted a comprehensive transcriptomic analysis of 40 muscle biopsies (from 10 CP and 10 control individuals), using Affymetrix HG-U133A 2.0 microarrays.

Unsupervised techniques (PCA, K-means, hierarchical clustering) revealed distinct gene expression profiles between CP and control samples. Supervised classifiers (Random Forest, Lasso, LDA) achieved over 95% accuracy. Functional enrichment analyses using DAVID and g:Profiler indicated dysregulation in RNA processing, nuclear localization, cytoskeletal organization, and post-translational modifications. PathfindR network analysis further supported these findings and highlighted glutamatergic synapse signaling as a potentially critical pathway in CP pathophysiology.

This integrative systems biology approach reveals a multilayered molecular disruption in CP involving transcriptional misregulation, impaired protein homeostasis, and synaptic and cytoskeletal instability. These insights provide a foundation for advancing the understanding of CP-related muscle contractures.

## Introduction

Cerebral palsy (CP) is a disorder of the upper motor neuron (UMN) system that leads to a broad spectrum of movement impairments. It is caused by a non-progressive lesion in the developing brain and represents the most common motor disability in childhood, affecting approximately 3.6 per 1,000 children in the United States, a prevalence that has remained steady despite advances in neonatal care (1). While CP originates from a central nervous system lesion, its downstream effects on skeletal muscle are significant and complex. Interestingly, there are no known primary genetic mutations associated with CP, and its muscular phenotype arises as a consequence of impaired neural input, rather than intrinsic muscle defects.

A key secondary complication in CP is the development of muscle contractures, permanent muscle shortening that restricts joint movement. These contractures are thought to result from altered muscle adaptation to chronic changes in neural stimulation. Although considered an adaptive response, the mechanisms underlying contracture formation remain poorly understood. Emerging evidence suggests that transcriptional dysregulation in muscle plays a significant role in driving the pathological changes associated with increased passive stiffness and contracture development. Understanding these transcriptional changes could offer insight into how altered gene expression contributes to muscle remodeling and dysfunction in CP.

To investigate this, we analyzed RNA expression profiles from biopsies of two synergistic muscles, the gracilis and semitendinosus, which both contribute to knee flexion. A total of 40 microarrays were generated from 20 subjects: 10 individuals with CP and 10 age-matched control subjects, with separate microarrays run for each muscle biopsy. No genes showed a significant interaction between muscle type and disease status, indicating that both muscles undergo similar transcriptomic changes in response to CP. Hierarchical clustering, performed in a previous study (1), further confirmed that gracilis and semitendinosus biopsies from the same individual clustered closely together, suggesting that intra-subject variability between these muscles is minimal compared to inter-subject variability. From the same study also emerged that, despite expectations, biopsies did not cluster according to clinical severity scores, indicating that transcriptional profiles are not tightly linked to the phenotypic expression of severity in CP within this sample size.

This study aims to characterize the differential gene expression patterns associated with CP-related muscle contractures, and to identify key molecular pathways contributing to their development, with the ultimate goal of improving our understanding of the muscle-specific adaptations secondary to upper motor neuron injury.

## Materials and Methods

### Dataset

The dataset used in this analysis was retrieved from GEO database (2), it is accessible with the GEO accession number GSE31243. It contains transcriptomic data extracted from 40 muscle biopsies of 20 patients (10 with CP and 10 controls), processed through Affymetrix HG-U133A 2.0 technology, with a panel of 22277 genes. The samples were splitted in two groups: one including 20 biopsies, both of gracilis and semitendinosus, from patients with CP and the other including 20 biopsies obtained from control subjects.

### Methods

The analysis was carried out in an R environment, using the GEOquery library from the Bioconductor package (3) to retrieve the dataset.

Different methodologies, supervised and unsupervised, were employed to extract meaningful insights from the transcriptomic data. The initial exploratory analysis included visual representations, such as boxplots, and dimensionality reduction techniques like Principal Component Analysis (PCA). Given the result of the first boxplot (Figure S1), which was generated using the raw data, the dataset was transformed using a log2 transformation, to stabilize variance and normalize the data distribution across samples.

Several clustering algorithms were implemented, including K-means (4) and hierarchical clustering (5). K-means was initially executed with the default value of  $K=2$ , reflecting the binary nature of the dataset. For hierarchical clustering, different values of  $K$  were tested. As will be discussed later, it was ultimately found that a value of  $K = 5$  was more appropriate, as it allowed the outliers to be clustered separately with respect to the rest of the data points. Additionally, different linkage methods were evaluated to determine which offered the best discriminatory power, the final choice being Ward.D2.

The data pre-processing stage of the analysis was concluded with the use of the genefilter library, which was employed to filter out genes expressed in more than 20% of the samples. However, no genes were filtered out as a result of this process.

Following data pre-processing, a comparison of machine learning algorithms was performed to identify predictive patterns of disease outcome. The methods included in the analysis were Random Forest (RF) (6), Lasso Regression (7) and Linear Discriminant Analysis (LDA) (8), all implemented with the help of the caret package, which also facilitated the split of the dataset into training and test sets.

The classifiers were trained and tested on a list of 652 differentially expressed genes, which were identified using a t-test with a p-value threshold of  $< 0.01$  (Benjamini-Hochberg correction).

Starting with RF, the model was trained, using the caret package, with 10-fold cross-validation to optimize generalization. The number of trees was set to 50 as a result of the error rate plot (Figure S2). Variable importance was extracted using the varImp() function, which ranks genes based on their contribution to the classification accuracy. The top 200 most important genes were selected and saved for further biological interpretation. The LDA model was also trained using 10-fold cross-validation.

To identify a spare set of discriminative genes Lasso regression was applied using the glmnet package. The Lasso model was fit to the gene expression data, then the regularization strength ( $\lambda$ ) was optimized via 10-fold cross-validation (cv.glmnet()), selecting the  $\lambda$  value that minimized deviance (lambda.min). The final model

was trained using caret with a tuning grid ( $\alpha = 1$  for LASSO,  $\lambda = \lambda_{\min}$ ) and family="binomial", parameter for binary outcomes, and the top 100 genes with the largest absolute coefficients were retained for downstream analysis.

The performances of the three algorithms were compared using the resamples() function.

Lastly, we employed an additional classification algorithm, SCUDO framework (9). The dataset was split into training and testing subsets using stratified sampling (createDataPartition()). In the training phase, we generated individual gene expression signatures using the scudoTrain function, selecting the top and bottom 25 most differentially expressed genes per sample. These signatures were used to build a similarity network among samples based on pairwise distances, and the resulting graph was visualized using scudoPlot(). To evaluate classification performance, the trained model was tested on the held-out test set using scudoTest(), followed by construction of a corresponding test network. Community detection was performed using the cluster\_spinglass algorithm to identify distinct sample groupings. Later, the model was also trained using 5-fold cross-validation, which tested combination of genes, selecting the optimal model by multiclass accuracy.

Gene ontology analysis was performed on the top 200 probes identified by Random Forest classification using two complementary approaches: DAVID v6.8 (10) for traditional functional annotation with emphasis on protein domains and post-translational modifications, and g:Profiler (2025 update) (11) for contemporary Gene Ontology (12) term enrichment across biological processes, molecular functions, and cellular components. For DAVID, we uploaded the list of 200 gene identifiers using the Homo sapiens genome as background and used the Functional Annotation Chart tool to identify enriched Gene Ontology terms, UniProt keywords, and protein domains.

In parallel, we performed enrichment analysis using g:Profiler, specifying the Homo sapiens genome as background and using default settings and g:SCS for multiple test correction. The top enriched terms across GO categories (Biological Process, Molecular Function, and Cellular Component), KEGG (13) pathways, Reactome pathways, and protein complexes were visualized using the g:Profiler interactive Manhattan plot and term table.

To identify functionally related gene networks and pathways, we performed network-based enrichment analysis using pathfindR (v2.4.0) (14) with KEGG database. The analysis was run with 5 iterations to ensure robustness, using GeneMania for gene-gene interaction filtering. Enriched pathways were hierarchically clustered based on shared genes and relationships were visualized as a term-gene network.

Network analysis was concluded with the employment of two other tools, STRING and EnrichNet, to further confirm previous findings.

## Results

### Exploratory Analysis

The dataset consisted of mRNA expression profiles of patients with Cerebral Palsy and typically developed patients, which were divided accordingly into two groups.

Raw gene expression data were visualized using boxplots to assess distribution of gene expression levels across samples. To

reduce skewness caused by high-expression outliers and to stabilize variance across genes, a log2 transformation was applied to the data prior to further analysis. After log2 transformation, boxplots were re-examined to confirm improved distributional uniformity across samples, indicating that the data were adequately normalized and suitable for downstream statistical analysis, as shown in Figure 1.

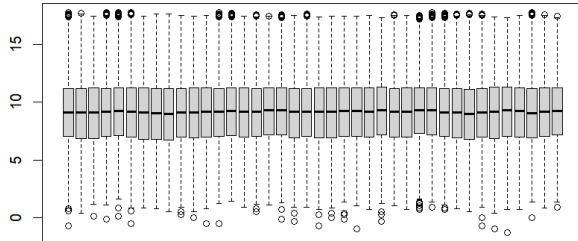


Fig. 1: Boxplot of initial values per sample (x-axis) after normalization with log2 transformation.

To explore the general patterns in the gene expression data and assess potential clustering between experimental groups, we performed PCA, which results were plotted (Figure 2). Since the 2D plot revealed distinct separation between the two groups, we decided not to produce a 3D plot. PC1, which accounts for 7.7% of the total variance in the dataset, effectively separates most Cerebral Palsy samples (outlined in blue) from control samples (outlined in orange).

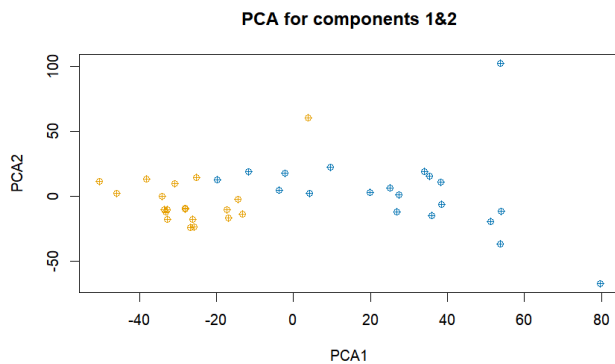


Fig. 2: 2D PCA plot highlighting the effective separation of most Cerebral Palsy samples (outlined in blue) from control samples (outlined in orange).

### Data Clustering Analysis

Results from K-means clustering show an overall great separation of the samples in the two groups, as shown in Figure 3. The

labels assigned to each point, which can be observed in the plot, correspond to the ground truth. Overall, most of the control subjects were correctly clustered in the orange cluster, while most of the CP subjects were grouped in the blue one. Similar to what could be observed in the PCA plot, the data points close to the center are the only ones misclassified. Hierarchical

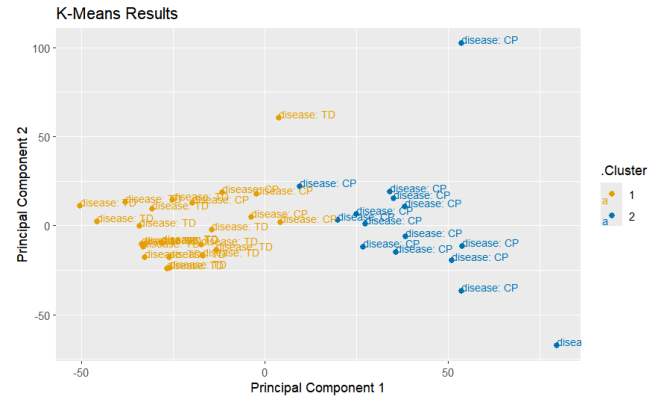


Fig. 3: Plot of the result of K-means clustering. The labels assigned to each point correspond to the ground truth. An overall great separation can be clearly observed, with few data points being misclassified.

clustering was then employed to further explore the structure of the data. Various linkage methods were tested, including average, complete, Ward.D, and Ward.D2. Among these, only the last two were able to produce balanced clusters and achieve an almost perfect separation between the two study groups, as shown in Figure 4. The figure clearly supports the findings of previous analyses: most subjects are correctly grouped according to the two conditions, with a small subset of five patients misclassified by various clustering techniques. In this case, four CP patients are grouped in the first cluster, while only one control subject is misclassified. Since the algorithm was unable to fully distinguish between the two groups, we chose a value of  $K=5$ , so to create five separate clusters. This value was chosen as it was the smallest value that allowed for the problematic subjects to be clustered separately.

### Supervised Analysis

Three supervised learning algorithms, Lasso regression, Random Forest and Linear Discriminant Analysis were evaluated using 10-fold cross-validation to classify samples into control and cerebral palsy groups. All methods demonstrated high accuracy ( $>95\%$ ), as it is shown in Figure 5, with Lasso achieving perfect classification on the dataset, suggesting strong linear separability between groups. Both RF and LDA showed comparable performance, with identical variability across cross-validation folds, though RF was prioritized for downstream analysis due to its ability to rank genes by importance, providing interpretable biological insights and generate stable feature rankings for enrichment analysis.

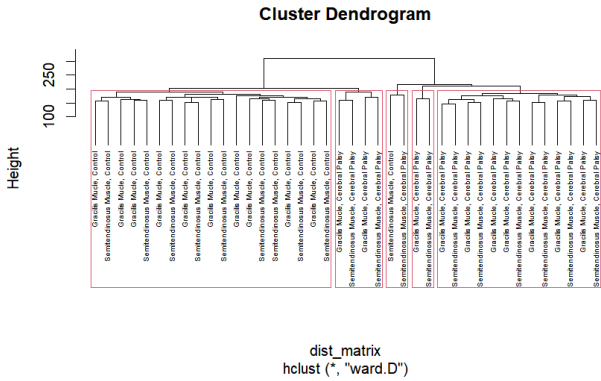


Fig. 4: Cluster Dendrogram resulting from the application of hierarchical clustering technique with Ward.D linkage method. Most of the subjects are correctly grouped together based on their condition. The decision of creating five distinct clusters, despite the binary nature of the dataset, was made to isolate the misclassified subjects into small clusters.

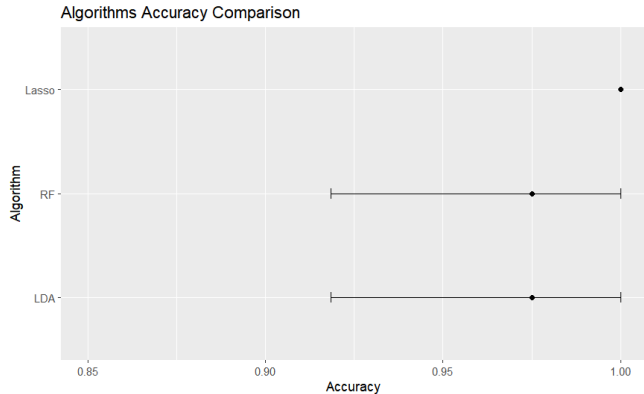


Fig. 5: Comparison of accuracies of supervised algorithms for classification, assessed using 10-fold cross-validation. All the three methods present high accuracy, with Lasso regression being a perfect classification method for the dataset. For RF and LDA the variability is the same, around 1.25, still the classification done with RF was chosen to extract a ranked gene list to move forward with enrichment analysis.

The Random Forest classifier identified 500 most influential genes for distinguishing between sample groups, ranked by their variable importance scores (Figure 6). After analyzing the plot, the top 200 genes (importance scores ranging from 0.10 to 0.05) were selected for downstream analysis. The step-like pattern observed in the importance plot, particularly around the 100th and 200th positions, can possibly suggest a natural threshold between core discriminative genes (top 100), secondary contributors (top 200) and background signal, or simply could be plateau regions representing groups of genes with similar predictive power.

Due to its high accuracy, the top 100 genes selected by LASSO regression were chosen to perform network analysis, prioritizing Lasso over Random Forest. This was mostly done because of Lasso's sparse selection property which enables it

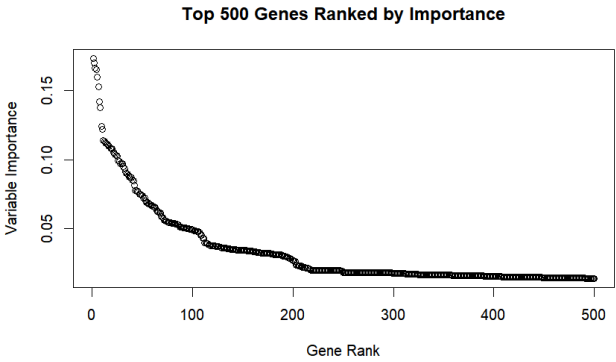


Fig. 6: Top 500 most influential genes for classification, ranked for their variable importance in RF classification. In the plot it is clear how, among all the genes plotted, the first 200 are sufficient to move forward with downstream analysis. Also, a step-like pattern can be observed.

to inherently identifying a minimal set of non-redundant genes by shrinking weakly associated features to zero, reducing noise in network construction. Also, Lasso's linear coefficients reflect conditional dependencies between genes and the outcome, better capturing biologically interpretable relationships for pathway-centric analyses.

Concluding the supervised analysis part of the study, SCUDO analysis yielded a distinct clustering of cerebral palsy and control samples in both the training and test networks. Figure 7 represents the test network, where samples were arranged into two visually distinct groups: control subjects (cyan) and CP patients (red), suggesting that the SCUDO-derived signatures generalized well to unseen data. The clear separation between the two conditions indicates the presence of robust disease-associated transcriptomic patterns. Classification accuracy assessed on the test set (0.8) confirmed high discriminatory performance, consistent with findings from other supervised models such as Random Forest and Lasso. In Figure 8 the same test network was further analyzed using the spinglass community detection algorithm. This revealed two tightly connected clusters, visually highlighted in blue and red. The modular structure of the network and the minimal overlap between groups reflect the biological consistency and predictive strength of the SCUDO approach.

However, when trained with 5-fold cross validation, SCUDO yielded an unexpected accuracy value of 0.3. Hence, this classification method was discarded for downstream analysis.

### Functional Analysis

Functional annotation of the top 200 Random Forest-selected genes was performed by DAVID, the results are shown in Figure 9.

The analysis revealed significant enrichment in genes localized to the nucleus, and cytosol, suggesting a strong involvement of nuclear and cytoplasmic regulatory components in CP pathophysiology. Enriched molecular functions included RNA binding, protein binding, and transcription coactivator activity, highlighting alterations in transcriptional and post-transcriptional regulation in CP.

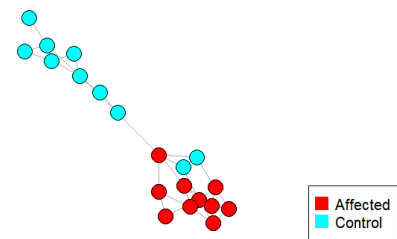


Fig. 7: Network visualization of test samples generated with SCUDO based on transcriptomic similarity. Nodes represent individual samples; colors indicate clinical groups (red = affected, cyan = control). The spatial separation of clusters reflects the ability of SCUDO-trained gene signatures to distinguish between CP and control samples.

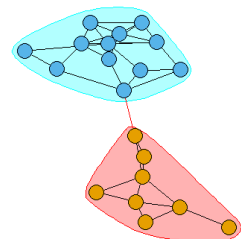


Fig. 8: Community detection of SCUDO test network using the spinglass algorithm. Two distinct communities were identified and highlighted (blue and red shaded regions), corresponding closely to control and CP sample groups.

In addition, several post-translational modifications were significantly overrepresented, including phosphorylation, ubiquitin conjugation, acetylation, and methylation, which are known to regulate protein function and signaling pathways.

Complementary results were obtained using g:Profiler. In particular, key enriched terms included protein binding and cytoskeleton consistent with the DAVID findings. Additional significant annotations included focal adhesion, cortical cytoskeleton, and sarcolemma, indicating alterations in structural and cytoskeletal dynamics in CP patients. These enrichments collectively suggest that cerebral palsy may involve dysregulation of nuclear function and structural cellular components.

Network analysis

The PathfindR network analysis revealed several significantly enriched pathways in cerebral palsy patients, as shown in Figure 11. The most prominent pathway was the glutamatergic

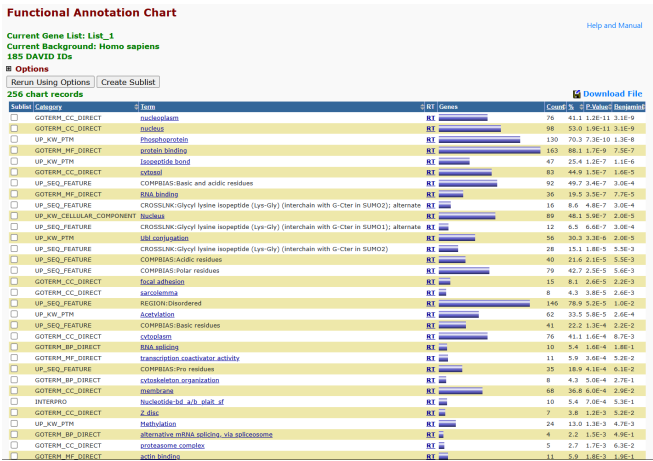


Fig. 9: Functional annotation chart of top 200 genes identified by Random Forest, using DAVID. Each row represents an enriched functional term across various annotation categories. Terms such as RNA binding, and phosphoprotein are among the most significantly enriched, with adjusted p-values (Benjamini correction) shown in the rightmost column. The bar graphs indicate the number and percentage of genes associated with each term, emphasizing a strong enrichment in nuclear and RNA-related processes.



Fig. 10: Functional enrichment analysis of the top 200 genes using g:Profiler. Each dot represents an enriched term, plotted by significance along the y-axis. The lower panel provides the top 12 enriched terms, with details including the source, GO term ID, name, and adjusted p-values. Notably, terms such as protein binding, cytoskeleton organization and focal adhesion were highly enriched, highlighting biological processes and cellular structures potentially relevant to cerebral palsy pathophysiology.

synapse, suggesting alterations in neuronal signaling. Other key pathways included bacterial invasion of epithelial cells hippo signaling pathway and leukocyte transendothelial migration. These pathways showed strong statistical significance, with fold enrichment values ranging from 5 to 20. The analysis identified 2-4 key genes associated with each pathway.

Notably, the proteasome and RNA-related pathways (spliceosome) from the PathfindR analysis showed overlap with the functional

enrichment results, which also identified significant involvement of protein degradation and RNA binding mechanisms. Similarly, the adherens junction pathway aligns with the cytoskeletal organization terms found in the functional enrichment analysis. This is visually reflected in the term-gene graph (Figure 12), where

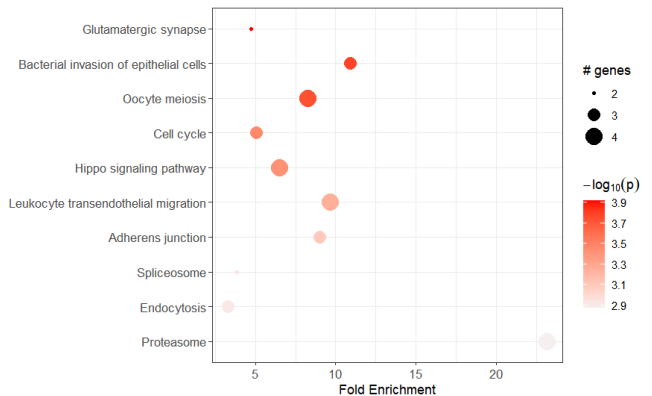


Fig. 11: Pathway enrichment network analysis of differentially expressed genes using PathFindR. The network visualization highlights significantly enriched pathways and their interactions. Key pathways include glutamatergic synapse, bacterial invasion of epithelial cells Hippo signaling, and leukocyte transendothelial migration. Fold enrichment values (5-20) are indicated for select pathways.

enriched pathways are connected to individual up- and down-regulated genes. Several pathways, such as hsa04520 (Adherens junction), hsa04670 (Leukocyte transendothelial migration), and hsa04110 (Cell cycle regulation), act as network hubs, each linked to multiple differentially expressed genes. Moreover, proteasome-related terms (e.g., hsa03050 and hsa03040) are supported by a cluster of consistently upregulated genes.

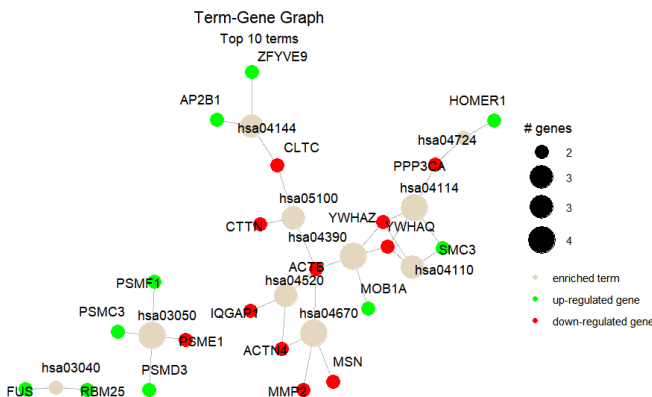


Fig. 12: Term-gene interaction network generated by PathfindR showing the top 10 significantly enriched pathways (beige nodes) and their associated differentially expressed genes. Genes are colored based on regulation status: green (upregulated), red (downregulated).

STRING and EnrichNet further reinforced the functional themes identified in the transcriptomic data. STRING analysis produced an extensive interaction network in which K-means clustering identified three distinct modules (Figure 13). The largest cluster (red) represented general cellular functions, while the two smaller clusters were functionally specialized, the green one enriched in RNA-binding proteins and ribonucleoproteins, and the blue one composed entirely of ubiquitin-related proteins. Similarly, EnrichNet pathway analysis using KEGG revealed

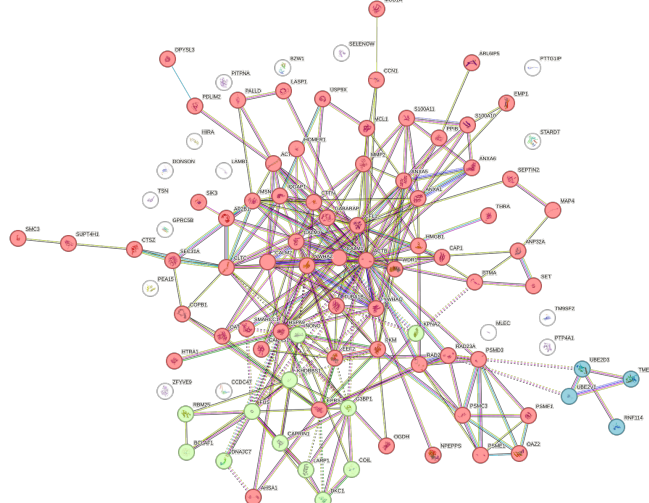


Fig. 13: Protein interaction network generated by STRING visualized with K-means clustering ( $k = 3$ ). The red cluster corresponds to proteins involved in general cellular processes. The green cluster is enriched in RNA-binding proteins and ribonucleoproteins, while the blue cluster consists exclusively of ubiquitin-related proteins.

that the top five ranked pathways by gene similarity included proteasome, nucleotide excision repair, adherens junction, and infection-related pathways such as pathogenic *Escherichia coli* infection and bacterial invasion of epithelial cells, as it is shown in Figure 14.

### Discussion

The transcriptomic analysis conducted in this study reveals a robust distinction between muscle tissue samples from children with Cerebral Palsy (CP) and typically developing (control) individuals. Both unsupervised and supervised approaches consistently confirmed that the two conditions exhibit distinct gene expression profiles, underscoring the presence of a transcriptional signature associated with CP-related muscle contractures.

Exploratory analysis using PCA demonstrated a clear separation between CP and control samples along the first principal component, which accounted for 7.7% of the total variance. This separation was further reinforced by clustering methods: K-means successfully grouped most samples according to their clinical condition, and hierarchical clustering with Ward linkage methods produced similarly accurate groupings. Interestingly, a small








Annotation (pathway/process) ▾	Significance of network distance distribution (XO-Score) ▾	Significance of overlap (Fisher-test, q-value) ▾	Dataset size (uploaded gene set) ▾	Dataset size (pathway gene set) ▾	Dataset size (overlap) ▾
<b>Pathogenic <i>Escherichia coli</i> infection</b>					
 <a href="#">compute graph visualization</a> <a href="#">see mapped genes</a>	0.8010	0.0025	90	52	5 (show)
<b>Proteasome</b>					
 <a href="#">compute graph visualization</a> <a href="#">see mapped genes</a>	0.7728	0.0116	90	43	4 (show)
<b>Nucleotide excision repair</b>					
 <a href="#">compute graph visualization</a> <a href="#">see mapped genes</a>	0.3542	0.6051	90	43	2 (show)
<b>Bacterial invasion of epithelial cells</b>					
 <a href="#">compute graph visualization</a> <a href="#">see mapped genes</a>	0.3326	0.2569	90	68	3 (show)
<b>Adherens junction</b>					
 <a href="#">compute graph visualization</a> <a href="#">see mapped genes</a>	0.3106	0.2578	90	72	3 (show)

Fig. 14: Top five ranked pathways by gene similarity identified by EnrichNet pathway analysis using KEGG.

number of samples near the decision boundaries were consistently misclassified across techniques, suggesting subtle heterogeneity within or across clinical groups.

The performance of supervised classification models, conducted on a subset of 652 differentially expressed genes, which were identified using a t-test with a p-value threshold of  $< 0.01$ , further confirmed the strong discriminative signal in the dataset. Lasso regression achieved perfect classification accuracy, indicating near-complete linear separability between the two groups. Random Forest and LDA also yielded accuracies exceeding 95%, with low variability across cross-validation folds. Despite Lasso's superior classification performance, Random Forest was selected for feature ranking due to its ability to quantify variable importance and facilitate biologically meaningful gene prioritization, while Lasso was chosen for network analysis.

Functional enrichment analysis highlighted widespread dysregulation of genes involved in nuclear and cytoplasmic processes in CP, suggesting that disruptions in RNA processing and protein modification pathways may underlie aspects of the disease. This supports the growing recognition that CP, while primarily caused by early brain injury, also exhibits molecular changes similar to those seen in other neurodevelopmental and neuromuscular disorders.

In particular, the misregulation of RNA-binding proteins (RBPs), a hallmark of many motor neuron diseases (MNDs) and muscular dystrophies (15), appears to be relevant in CP as well. These proteins, including various hnRNP family members, are critical for RNA splicing, transport, and stability (15). Their dysfunction can lead to abnormal alternative splicing and stress granule formation, contributing to disease pathology. Although CP is not a genetic disorder in the classical sense, transcriptomic parallels with conditions like myotonic dystrophy suggest a shared molecular foundation involving RBP-related regulatory networks (16).

Further transcriptomic findings point to disruption in key developmental pathways, essential for cytoskeletal organization and neural connectivity. Consistent with this, muscle tissue in children with CP shows structural and functional abnormalities from an early age (17).

Together, these findings reinforce a model of CP as a condition that spans both neurological and muscular domains, with secondary,

progressive changes in muscle biology that may be amenable to therapeutic intervention (18).

Complementary to the enrichment analysis performed with DAVID and g:Profiler, the network analysis using PathfindR, STRING, and EnrichNet offered deeper insights into the molecular mechanisms underlying cerebral palsy. PathfindR successfully confirmed several pathways identified in earlier analyses, including those related to RNA processing (spliceosome), protein degradation (proteasome), and cytoskeletal regulation (adherens junction, focal adhesion). Additionally, the glutamatergic synapse pathway emerged as a highly enriched term, aligning with the broader hypothesis of altered synaptic signaling in CP and neurodevelopmental dysfunctions.

STRING network clustering identified distinct modules, including RNA-binding proteins and ubiquitin-related factors, further supporting transcriptional and proteostatic disruption. EnrichNet's gene similarity ranking emphasized pathways such as proteasome, bacterial invasion, and adherens junctions, in addition to themes of immune activation, structural remodeling, and impaired protein homeostasis.

While the network analysis revealed enrichment in pathways such as oocyte meiosis and Hippo signaling, current literature provides limited or no direct evidence linking these pathways to cerebral palsy pathogenesis. Additionally, the enrichment of pathways related to bacterial invasion and immune activation, identified by both PathfindR and EnrichNet, is consistent with growing evidence that chronic inflammation plays a sustained role in CP, particularly among preterm infants (19).

## Conclusion and future perspectives

In this study we conducted a comprehensive transcriptomic characterization of muscle tissue from individuals with cerebral palsy, revealing a multi-layered molecular pathology that extends beyond the initial upper motor neuron lesion. Through a combination of unsupervised clustering, supervised classification, enrichment analysis, and network-based interpretation, we identified consistent and biologically meaningful alterations in gene expression. These changes point to disruptions in RNA processing, post-translational protein regulation, and cytoskeletal organization, core processes essential for both neuronal function and muscle integrity. Convergent findings across DAVID, g:Profiler, and PathfindR highlighted shared pathways such as the proteasome, spliceosome, and glutamatergic synapse, suggesting that synaptic signaling and protein homeostasis may play a prominent role in CP.

Moving forward, integrating transcriptomic data with proteomic, epigenetic, and single-cell analyses could provide a more granular understanding of how these pathways operate across different cell types and developmental stages. Functional validation of key targets may inform biomarker development or guide intervention strategies aimed at mitigating long-term disability in affected individuals.

## References

1. Lucas R Smith, Henry G Chambers, Shankar Subramaniam, and Richard L Lieber. Transcriptional abnormalities of hamstring muscle contractures in children with cerebral palsy. *PLoS ONE*, 7(8):e40686, 2012. Epub 2012 Aug 16. doi: 10.1371/journal.pone.0040686.

2. Tanya Barrett, Stephen E. Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F. Kim, Maxim Tomashevsky, Kimberly A. Marshall, Katherine H. Phillippy, Patti M. Sherman, Michelle Holko, Andrey Yefanov, Hyeseung Lee, Naigong Zhang, Cynthia L. Robertson, Nadezhda Serova, Sean Davis, and Alexandra Soboleva. Ncbi geo: archive for functional genomics data sets—update. *Nucleic Acids Research*, 41(D1):D991–D995, 11 2012. [arXiv:https://academic.oup.com/nar/article-pdf/41/D1/D991/3678141/gks1193.pdf](https://academic.oup.com/nar/article-pdf/41/D1/D991/3678141/gks1193.pdf), doi:10.1093/nar/gks1193.
3. Bioconductor Core. An overview of projects in computing for genomic analysis. Technical Report 1, Bioconductor Core, La Jolla, CA, November 2002. Biocore Technical Report. URL: <https://www.bioconductor.org/help/publications/tech-reports/relProjTR.pdf>.
4. J. A. Hartigan and M. A. Wong. A k-means clustering algorithm. *JSTOR: Applied Statistics*, 28(1):100–108, 1979.
5. Stephen C. Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.
6. Leo Breiman. Random forests. 45(1):5–32. doi:10.1023/A:1010933404324.
7. R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, 58:267–288, 1996.
8. David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003. URL: <http://portal.acm.org/citation.cfm?id=944937>, doi:http://dx.doi.org/10.1162/jmlr.2003.3.4-5.993.
9. Mario Lauria, Petros Moyseos, and Corrado Priami. Scudo: a tool for signature-based clustering of expression profiles. *Nucleic Acids Research*, 43(W1):W188–W192, 05 2015. [arXiv:https://academic.oup.com/nar/article-pdf/43/W1/W188/7476247/gkv449.pdf](https://academic.oup.com/nar/article-pdf/43/W1/W188/7476247/gkv449.pdf), doi:10.1093/nar/gkv449.
10. B.T. Sherman, G. Panzade, T. Imamichi, and W. Chang. David ortholog: an integrative tool to enhance functional analysis through orthologs. *Bioinformatics*, 40(10):btæ615, Oct 2024. doi:10.1093/bioinformatics/btæ615.
11. Liis Kolberg, Uku Raudvere, Ivan Kuzmin, Jaak Vilo, and Hedi Peterson. gprofiler2— an r package for gene list functional enrichment analysis and namespace conversion toolset g:profiler. *F1000Research*, 9 (ELIXIR)(709), 2020. R package version 0.2.3.
12. Paul D. Thomas, David Ebert, Arun Muruganujan, Tomonari Mushayahama, Laurent P. Albou, and Huaiyu Mi. Panther: Making genome-scale phylogenetics accessible to all. *Protein Science*, 31(1):8–22, Jan 2022. doi:10.1002/pro.4218.
13. Minoru Kanehisa and Susumu Goto. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, Jan 2000. doi:10.1093/nar/28.1.27.
14. Ege Ulgen, Ozan Ozisik, and Osman Ugur Sezerman. pathfinder: An r package for comprehensive identification of enriched pathways in omics data through active subnetworks. *Frontiers in Genetics*, 10:858, 2019. URL: <https://doi.org/10.3389/fgene.2019.00858>.
15. Faye Ibrahim, Tatsuaki Nakaya, and Zissimos Mourelatos. Rna dysregulation in diseases of motor neurons. *Annual Review of Pathology: Mechanisms of Disease*, 7:323–352, Oct 2012. Epub ahead of print 2011 Oct 24. doi:10.1146/annurev-pathol-011110-130307.
16. Faye Ibrahim, Tatsuaki Nakaya, and Zissimos Mourelatos. Rna dysregulation in diseases of motor neurons. *Annual Review of Pathology: Mechanisms of Disease*, 7:323–352, Oct 2012. Epub ahead of print 2011 Oct 24. doi:10.1146/annurev-pathol-011110-130307.
17. Ayşe Gülşen Doğan and İhsan Çetin. Changes in the ubiquitination system in children with cerebral palsy. *Journal of Contemporary Medicine*, 13(4):652–656, 2023. URL: <https://dergipark.org.tr/en/pub/jcm/issue/80353/1296330>, doi:10.16899/jcm.1296330.
18. C. J. Walsh, J. Batt, M. S. Herridge, S. Mathur, G. D. Bader, P. Hu, P. Khatri, and C. C. dos Santos. Comprehensive multi-cohort transcriptional meta-analysis of muscle diseases identifies a signature of disease severity. *Scientific Reports*, 12(1):11260, Jul 2022. doi:10.1038/s41598-022-15003-1.
19. Madison C. B. Paton, Megan Finch-Edmondson, Russell C. Dale, Michael C. Fahey, Claudia A. Nold-Petry, Marcel F. Nold, Alexandra R. Griffin, and Iona Novak. Persistent inflammation in cerebral palsy: Pathogenic mediator or comorbidity? a scoping review. *Journal of Clinical Medicine*, 11(24):7368, Dec 2022. doi:10.3390/jcm11247368.



## Supplementary Figures

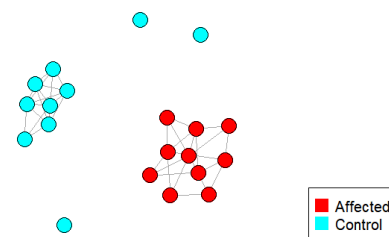
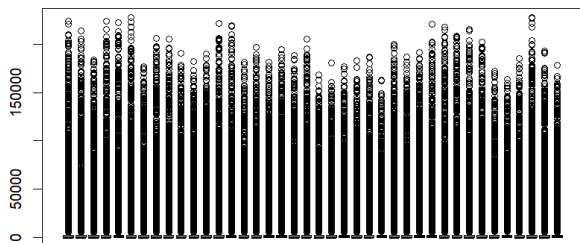


Fig. S3: Network visualization of training samples generated with SCUDO based on transcriptomic similarity.

Fig. S1: Boxplot of initial values per sample (x-axis) before normalization with log2 transformation.

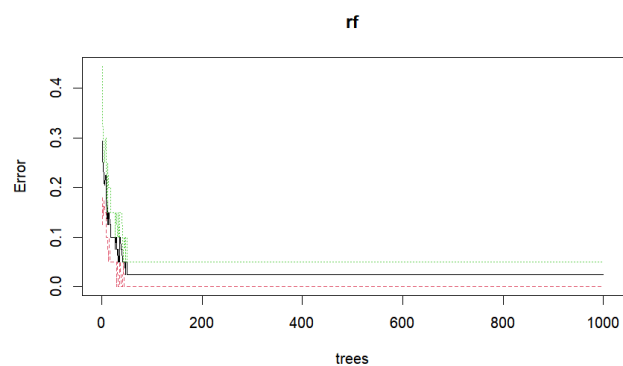


Fig. S2: Random Forest error rate as a function of the number of trees in the ensemble.