

## RESEARCH ARTICLE

# Multiple Criteria Optimization (MCO): A gene selection deterministic tool in RStudio

Isis Narváez-Bandera<sup>1</sup>, Deiver Suárez-Gómez<sup>1</sup>, Clara E. Isaza<sup>1,2,3,4</sup>, Mauricio Cabrera-Ríos<sup>1,5\*</sup>

**1** Bioengineering Graduate Program, The Applied Optimization Group, University of Puerto Rico-Mayagüez, Mayagüez, Puerto Rico, **2** Public Health Program, Ponce Health Sciences University, Ponce, Puerto Rico, **3** Basic Sciences Department, Ponce Health Sciences University, Ponce, Puerto Rico, **4** Biology Department, University of Puerto Rico-Mayagüez, Mayagüez, Puerto Rico, **5** Industrial Engineering Department, The Applied Optimization Group, University of Puerto Rico-Mayagüez, Mayagüez, Puerto Rico

\* [mauricio.cabrera1@upr.edu](mailto:mauricio.cabrera1@upr.edu)



## Abstract

Identifying genes with the largest expression changes (gene selection) to characterize a given condition is a popular first step to drive exploration into molecular mechanisms and is, therefore, paramount for therapeutic development. Reproducibility in the sciences makes it necessary to emphasize objectivity and systematic repeatability in biological and informatics analyses, including gene selection. With these two characteristics in mind, in previous works our research team has proposed using multiple criteria optimization (MCO) in gene selection to analyze microarray datasets. The result of this effort is the MCO algorithm, which selects genes with the largest expression changes without user manipulation of neither informatics nor statistical parameters. Furthermore, the user is not required to choose either a preference structure among multiple measures or a predetermined quantity of genes to be deemed significant a priori. This implies that using the same datasets and performance measures (PMs), the method will converge to the same set of selected differentially expressed genes (repeatability) despite who carries out the analysis (objectivity). The present work describes the development of an open-source tool in RStudio to enable both: (1) individual analysis of single datasets with two or three PMs and (2) meta-analysis with up to five microarray datasets, using one PM from each dataset. The capabilities afforded by the code include license-free portability and the possibility to carry out analyses via modest computer hardware, such as personal laptops. The code provides affordable, repeatable, and objective detection of differentially expressed genes from microarrays. It can be used to analyze other experiments with similar experimental comparative layouts, such as micro-RNA arrays and protein arrays, among others. As a demonstration of the capabilities of the code, the analysis of four publicly-available microarray datasets related to Parkinson's Disease (PD) is presented here, treating each dataset individually or as a four-way meta-analysis. These MCO-supported analyses made it possible to identify MMP9 and TUBB2A as potential PD genetic biomarkers based on their persistent appearance across each of the case studies. A literature search confirmed the importance of these genes in PD and indeed as PD biomarkers, which evidences the code's potential.

## OPEN ACCESS

**Citation:** Narváez-Bandera I, Suárez-Gómez D, Isaza CE, Cabrera-Ríos M (2022) Multiple Criteria Optimization (MCO): A gene selection deterministic tool in RStudio. PLoS ONE 17(1): e0262890. <https://doi.org/10.1371/journal.pone.0262890>

**Editor:** Francisco J. Esteban, Universidad de Jaen, SPAIN

**Received:** December 10, 2020

**Accepted:** January 9, 2022

**Published:** January 27, 2022

**Peer Review History:** PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: <https://doi.org/10.1371/journal.pone.0262890>

**Copyright:** © 2022 Narváez-Bandera et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** All four microarray datasets relating to Parkinson's Disease (PD) are available from the Gene Expression Omnibus (GEO) repository (<http://www.ncbi.nlm.nih.gov/>)

geo). Accession numbers: GSE99039, GSE18838, GSE19587, and GSE57475.

**Funding:** The project described was supported by Award Number U54MD007579 from the National Institute on Minority Health and Health Disparities. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

**Competing interests:** The authors have declared that no competing interests exist.

## Introduction

The analysis of gene expression leads to insight into biological processes and identification of biomarkers, as well as characterization of differing responses to therapy by individuals. Some of the first high throughput experiments used to analyze gene expression were microarray experiments. These were and still are typically used in comparative experiments between case and control groups to identify differentially expressed genes. The data generated by experiments like this populate large public repositories such as Gene Expression Omnibus. Ideally, it would be possible to pool several of these experiments to carry out a meta-analysis and arrive at statistically more robust conclusions about potential biomarkers. However, the fact that many of these experiments are measured in different units and scales has often made simultaneous analysis difficult. Nevertheless, several techniques are used to discover biomarkers using microarrays [1, 2]; these range from traditional statistical models to more computationally complex machine learning methods. For instance, in [3–6], methods based on genetic algorithms, spearman correlation, relief-F, and joint mutual information, among others, are used to analyze microarray data with that purpose. However, the outputs of these methods, require the configuration of a varying number of parameters that significantly affect their results. This hampers both analysis objectivity and repeatability. To this end, our research group in Camacho et al [7] proposed the MCO approach as it appears in this manuscript, which can analyze microarrays and other -omics datasets relying on Pareto-optimality conditions. The MCO-based analysis is carried out without the assumption of underlying statistical distributions a priori, the selection of a threshold value, or the adjustment of parameters by the user. Moreover, MCO presents a ranking based on the simultaneous consideration of performance measures included in the analysis. Our succeeding works presented in Cruz-Rivera et al. [8], Lorenzo et al. [9], Isaza et al. [10], successfully developed and applied MCO for gene selection in Alzheimer's Disease, cervix cancer, and lung cancer, respectively. In the present work, MCO is fully automated using R. The resulting code maintains the nonparametric qualities of MCO and minimizes possible errors due to manual handling of data. Lead time to carry out analysis is also significantly decreased, making MCO a convenient and powerful tool to support the search for potential biomarkers. The MCO R tool can be accessed from the address [https://server-deiver.shinyapps.io/MCO\\_TURBO/](https://server-deiver.shinyapps.io/MCO_TURBO/).

## Design and implementation

### MCO algorithm

As discussed in this work, the MCO algorithm requires at least one treatment vs control, comparative, high-throughput, replicated experiment. Several microarray datasets follow this organization in the characterization of relative gene expression. The comparative layout allows obtaining PMs for every gene in the experiment to measure differences in relative expression. For example, one PM can be the absolute value of the difference of means between the two groups (treatment, control), another can be the absolute value of the difference of medians between them, and so on. It is doubtful that selecting genes solely using one PM at a time will coincide precisely with the resulting selection. This evidences that there exists conflicts between different PMs. In the case of MCO, each gene is represented through several PMs -very much like a coordinate system- and the final selection is made up of the genes showing the best possible compromises between all PMs. In mathematical terms, this refers to identifying the Pareto-efficient solutions of the associated multiple criteria optimization problem. The Pareto-efficient solutions, in turn, form the Pareto-efficient frontier of such problem, from here on also referred to simply as the frontier. There exists sufficiency in Pareto-optimality,

which means that for the genes that meet the Pareto-efficient conditions (i.e., those lying in the frontier), no other gene can be found in the experiment offering a better compromise between the PMs at hand. This confers certainty to the results.

## MCO tool

The present work describes the development of an open-source code in RStudio of the MCO algorithm. The tool permits to identify, in a single run, the first  $F$  frontiers in an analysis. The user can specify  $F$  to create a hierarchy of genes organized in succeeding frontiers. The MCO tool was designed to support: (1) individual analysis of microarray datasets using two or three PMs, and (2) meta-analysis using two to five different datasets with one PM from each dataset (see Fig 1). It should be noted that the MCO algorithm could handle more datasets in a meta-analysis; however, to keep the computational cost low, the MCO tool is limited to a maximum of five datasets. The application of MCO results in sets of genes organized in  $F$  frontiers with decreasing levels of significance. In both options, the MCO algorithm follows three steps (Algorithm 1). In the first step, the PMs are selected from a predefined list, including the median, the mean, the mode, the third quartile, or a quartile of interest to the analyst. When the analysis is performed for an individual dataset, choosing between two or three PMs is possible. In the case of a meta-analysis, the default PM is the median from each dataset. The difference of the selected PMs between cases and controls is calculated for each gene. The absolute value function is further applied to then be subjected to a linear transformation aimed to set up a minimization MCO problem. In the second step, the MCO tool allows the user to divide the dataset into  $S$  groups (parallelism) to address RAM constraints when using hardware with modest capabilities. The MCO tool proceeds to find the local frontier of each group. The genes in each one of the  $S$  local frontiers are analyzed together to find the global Pareto-efficient frontier. This second step returns the genes with the best possible balances among the PMs to be minimized and are the ones identified as potential biomarkers. The third step entails applying the MCO algorithm  $F$  times, each time removing the previous frontier. This approach returns genes organized hierarchically in  $F$  frontiers.

**Algorithm 1:** Pseudocode of the MCO tool, the procedure to carry out the selection of the first  $F$  Pareto-efficient frontiers.  $D_k$  represents the  $K$  datasets to be analyzed, in this implementation  $K$  falls between 1 and 5.  $PM_c$  is the number of performance measures used to quantify relative expression changes between treatment and control.  $F$  is the number of successive frontiers to be presented in the analysis. The output is presented as a list of genes, hierarchically organized by frontier number.

**Input:**

$D_k \leftarrow$  Number of datasets |  $D_k \leftarrow (X_m, Y_n) \mid k \in 1, 2, \dots, K;$

$X_m \in$  samples-  $m = 1, 2, \dots, M;$

$Y_n \in$  gene sets-  $n = 1, 2, \dots, N;$

$PM_c \leftarrow$  Number of PMs |  $c = 1, 2, \dots, C;$

( $c =$  PMs in conflict);

$S \leftarrow$  Number of groups in which the genes are split |  $s = 1, 2, \dots, S;$

$F \leftarrow$  Number of frontiers |  $F = 1, 2, \dots, f;$

**for**  $D_k \leftarrow 1$  to  $K$  **do**

**for**  $PM \leftarrow 1$  to  $c$  **do**

**for**  $s \leftarrow 1$  to  $S$  **do**

**for**  $f \leftarrow 1$  to  $F$  **do**

        MCO function ( $PM_c$ )

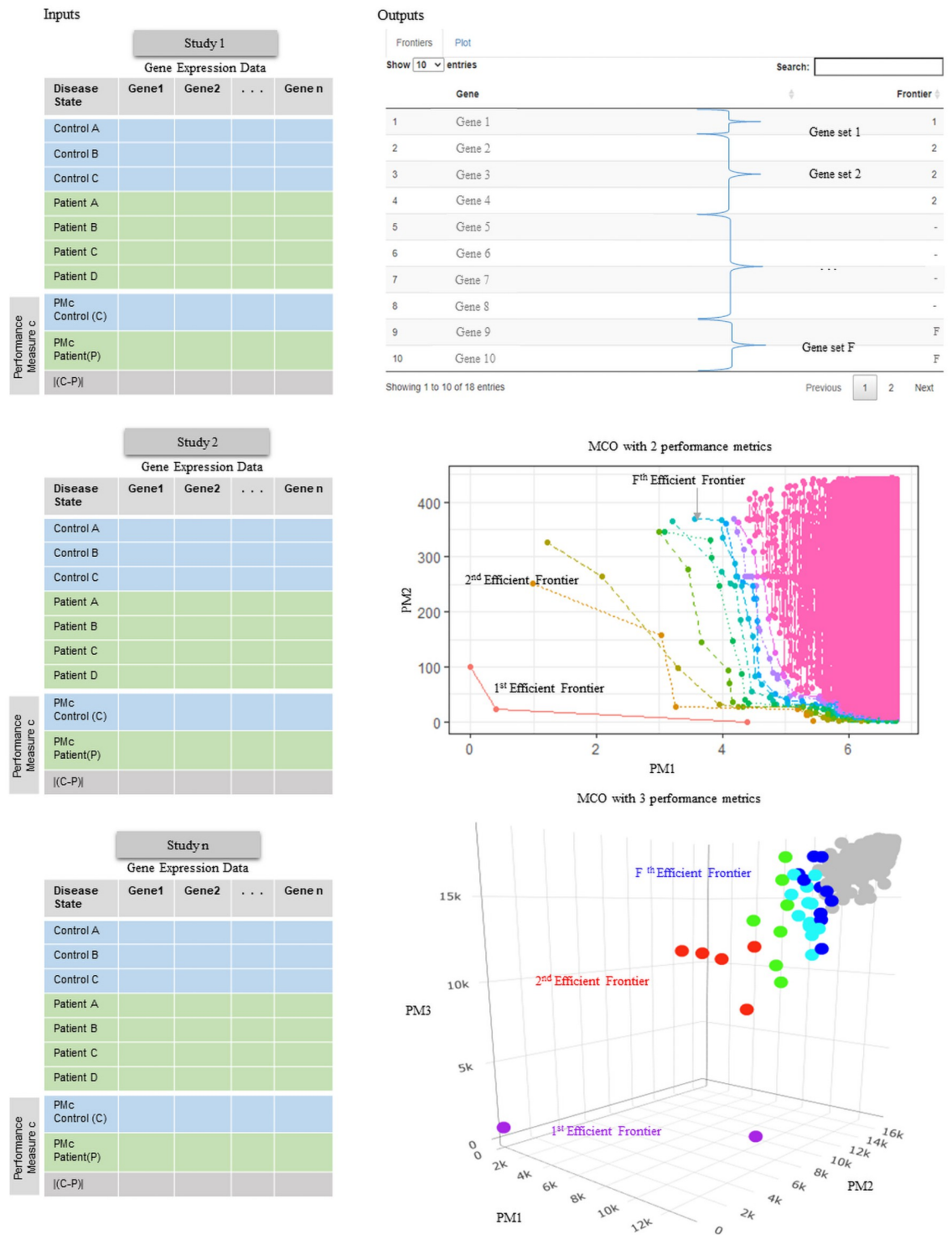
**end**

**end**

**end**

**end**

**Output:** Genes sets for each frontiers  $F$



**Fig 1. Overview of the MCO tool.** Gene expression datasets can be analyzed individually or by combining several datasets. The PMs can be generated for individual analysis or meta-analysis of up to five datasets simultaneously. The MCO results can then be visualized by the function plotMCO.

<https://doi.org/10.1371/journal.pone.0262890.g001>

Table 1. List of PD studies from GEO.

GEO accession	Year	Platform	Probe sets	Genes	Control (Male/Fem)	Condition (Male/Fem)
GSE99039	2017	GPL570	54675	23520	212 (70/142)	191 (101/90)
GSE18838	2010	GPL5175	316919	17326	11 (6/5)	17 (13/4)
GSE19587	2010	GPL571	22277	13515	10 (6/4)	12 (6/6)
GSE57475	2015	GPL6947	49576	25146	49 (26/23)	93 (62/31)

<https://doi.org/10.1371/journal.pone.0262890.t001>

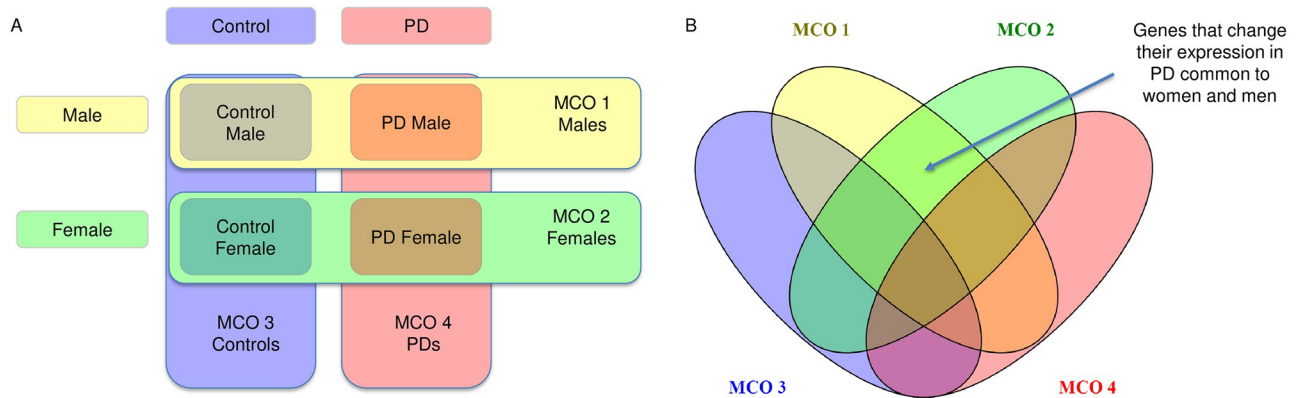
## Results

### MCO tool—Application

To demonstrate how the MCO tool works, four microarray datasets relating to Parkinson's Disease (PD) were selected from the Gene Expression Omnibus (GEO) repository. These are GSE99039 [11], GSE18838 [12], GSE19587 [13], and GSE57475 [14] (<http://www.ncbi.nlm.nih.gov/geo>). In the aggregation of the four datasets, there are 282 control samples (108 male/174 female) and 313 PD samples (182 male/131 female). Table 1 lists their characteristics. These datasets were selected using the following query filters: (1) 'Parkinson's' was used as a keyword, (2) the type of dataset was defined as 'expression profiling by array', (3) the organism was 'Homo sapiens', and (4) the gender information was set to 'available'. The latter was included since it has been considered relevant to differentiate PD profiles [15]. Following these criteria, each dataset contained four distinct groups arising from the intersection of sex and type of sample: MaleControl, MalePD, FemaleControl, and FemalePD.

Two types of analysis are presented: individual analysis of datasets and meta-analysis of multiple datasets. For the first type, we treated each of the four datasets individually. For the second type, we meta-analyzed three datasets simultaneously and four datasets simultaneously. Six analysis instances result in this manner: four individual analyses and two meta-analyses. In each instance, the goal was to detect genes with significant relative expression changes through MCO to characterize and infer their biological meaning in PD. MCO requires that at least two PMs are identified to work. On each of the four individual analysis instances, two PMs were used: the absolute value of the difference of the means and the absolute value of the difference of the medians between the pair of groups under comparison. The MCO analyses in these instances were, then, all 2-dimensional (2D). On the other hand, in the meta-analysis involving three datasets, there were three PMs: the absolute value of the difference of medians between the pair of groups under comparison from each dataset. The MCO analyses involved here are 3D. Finally, in the last meta-analysis instance involving four datasets, there were four PMs: the absolute value of the difference of medians between the pair of groups under comparison from each dataset, resulting in an MCO analysis that is 4D. On each of the six analysis instances, MCO was applied four times (Fig 2A): MCO 1 compares the pair of groups MaleControl—MalePD; MCO 2 compares the groups FemaleControl—FemalePD; MCO 3 compares the groups MaleControl—FemaleControl; and MCO 4, the groups MalePD—FemalePD. The genes of interest were those with biomarking potential solely for PD. Following set theory -see Venn diagram in Fig 2B this results in selecting the genes in the intersection of MCO 1 and MCO 2 that are not in MCO 3 or MCO 4.

One last difference is that, in the four individual analysis instances, the number of frontiers,  $F$ , was set to 10. In the two meta-analyses, it was set to 1 to keep results manageable. The results of the six analysis instances can be identified by defining: number of datasets in the analysis, the number of PMs (dimensions), and the last four digits of the identifiers of the datasets involved. Explicitly, the instances are: 1–2D-9039, 1–2D-8838, 1–2D-9587, 1–2D-7475, 3–3D-9039/8838/9587, 4–4D-9039/8838/9587/7475. The genes of interest for the four individual analysis instances can be consulted in Table 2. For illustration purposes, Fig 3 shows the



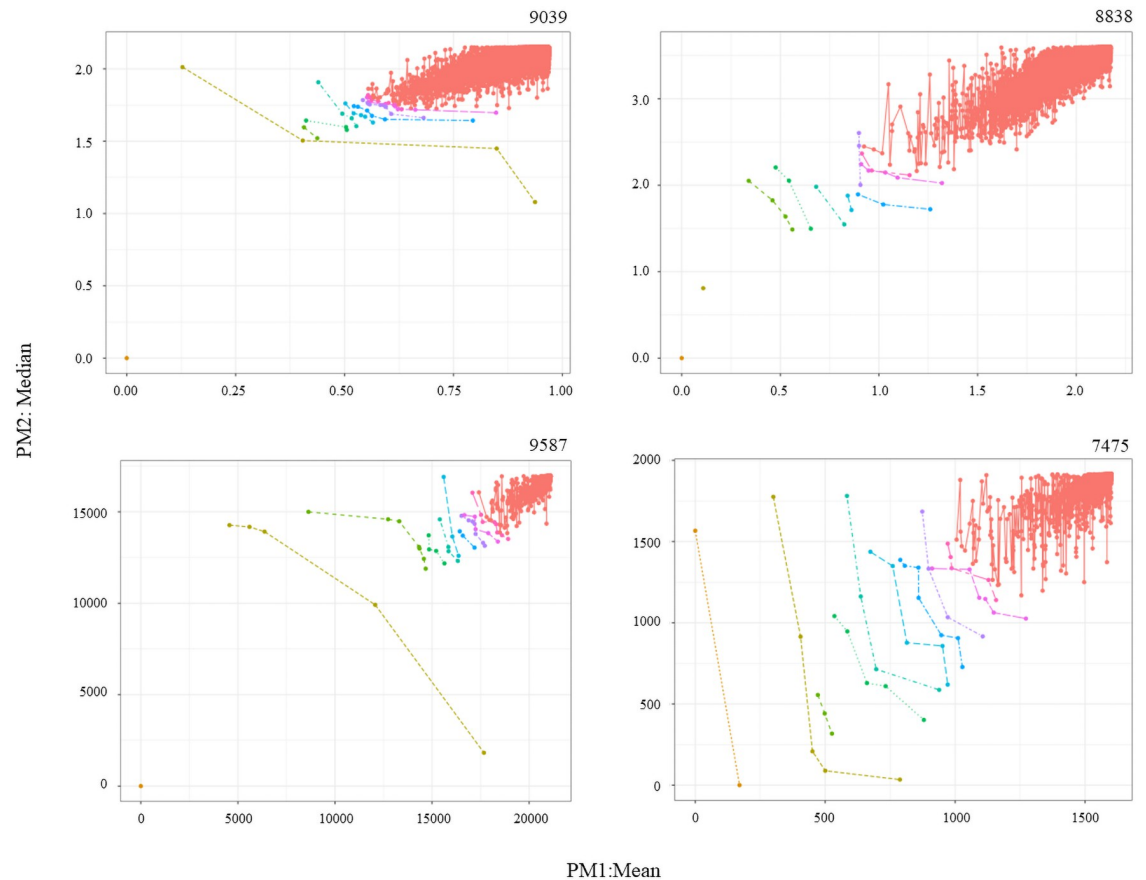
**Fig 2. The general scheme for all six analysis instances.** (A) MCO 1 compares the pair of groups MaleControl—FemaleControl; MCO 2 compares the groups FemaleControl—FemalePD; MCO 3 compares the groups MaleControl—FemaleControl; and MCO 4, the groups MalePD-FemalePD; (B) The genes of interest were those with biomarking potential solely for PD. Following set theory, this selects the genes in the intersection of MCO 1 and MCO 2 that are not in MCO 3 or MCO 4.

<https://doi.org/10.1371/journal.pone.0262890.g002>

**Table 2. Genes of interest from individual analysis instances 1–2D-9039, 1–2D-8838, 1–2D-9587, and 1–2D-7475 and the references supporting their roles in Parkinson’s Disease (PD) or neurodegenerative diseases (ND).**

Individual Analysis Instance	Gene symbol	Gene name	Reference related to	
			PD	ND
1–2D-9039	TUBB2A	Tubulin Beta 2A Class Iia	[16–18]	[19]
	CFD	Complement Factor D	[20]	[21]
	PTGDS	Prostaglandin D2 Synthase	[16]	
	LRRN3	Leucine Rich Repeat Neuronal 3	[16, 22]	
	ANXA3	Annexin A3		[23, 24]
	GPR97	G Protein-Coupled Receptor 97	[25]	
	PRKCD	Protein Kinase C Delta	[26, 27]	
	MMP9	Matrix Metalloproteinase 9	[28–30]	
	PGLYRP1	Peptidoglycan Recognition Protein 1	[31]	
	SPI1	Spi-1 Proto-Oncogene	[22]	
1–2D-8838	ND6	NADH Dehydrogenase Subunit 6	[32]	
	GTF2B	General Transcription Factor IIB		
	RPL18	Ribosomal Protein L18		
	TAGLN2	Transgelin 2		
	TMEM14B	Transmembrane Protein 14B		
	GABARAP	GABA(A) Receptor-Associated Protein	[18]	
	HIST1H1E	Histone Cluster 1		
1–2D-9587	OPHN1	Oligophrenin 1		
1–2D-7475	OAZ1	Ornithine Decarboxylase Antizyme 1		
	EEF1A1	Elongation Factor 1-Alpha 1	[33, 34]	
	ARHGDIB	Rho GDP Dissociation Inhibitor Beta	[35]	
	HBD	Hemoglobin Subunit Delta	[36]	
	CFD	Complement Factor D	[20]	[21]
	UBA52	Ubiquitin Carboxyl Extension Protein 52	[18]	

<https://doi.org/10.1371/journal.pone.0262890.t002>



**Fig 3. MCO 1: MalesPD Vs MaleControl.** Graphical representations for MCO 1 associated to individual analysis instances 1–2D-9039, 1–2D-8838, 1–2D-9587, and 1–2D-7457. Solutions toward the origin (0,0) are more significant.

<https://doi.org/10.1371/journal.pone.0262890.g003>

graphical results for MCO 1 on each dataset, while Fig 4 does the same for MCO 2. The complete lists identifying ten frontiers on each dataset can be found in the S2-S5 Tables in S1 File.

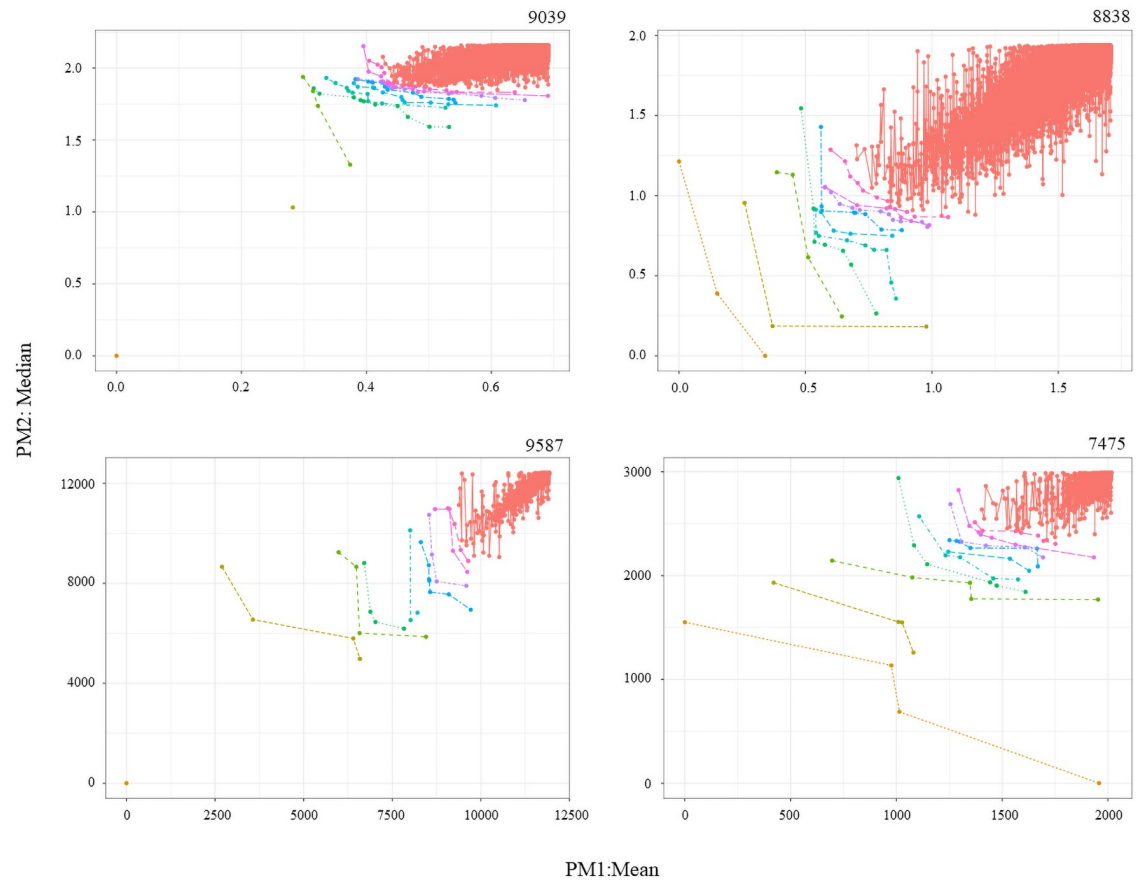
Table 3 contains the genes of interest for the two meta-analysis instances. In 3–3D-9039/8838/9587, four genes were identified as potential biomarkers: MMP9, RPS11, TUBA1B, and TUBB2A. Fig 5 shows the results for MCO 1 and MCO 2 for illustration purposes for this instance. The complete lists for MCO 1 through MCO 3 can be found in the S6 Table in S1 File.

In 4–4D-9039/8838/9587/7475, 10 genes of interest were identified: EEF2, MMP9, CFD, DAZAP2, MYL6, ARHGDI1B, RPL18, RPS11, CD81, and TUBB2A, as shown in Table 3. The complete lists for MCO 1 through MCO 4 can be found in the S7 Table in S1 File.

The Venn diagrams in Fig 6 show the overlap within the meta-analysis instances 3–3D-9039/8838/9587 in (A) and 4–4D-9039/8838/9587/7475 in (B). Notably, three genes (MMP9, RPS11, and TUBB2A) were identified in both instances. In addition, two out of these three, MMP9 and TUBB2A, were also identified in instance 1–2D-9039.

## Discussion

In the literature, our team performed a series of queries combining the name of each of the ten genes and either ‘Parkinson’s Disease’ or ‘Neurodegenerative’ to look for related biological



**Fig 4. MCO 2: FemalesPD Vs FemaleControl.** Graphical representations for MCO 2 associated to individual analysis instances 1–2D-9039, 1–2D-8838, 1–2D-9587, and 1–2D-7475. Solutions toward the origin (0,0) are more significant.

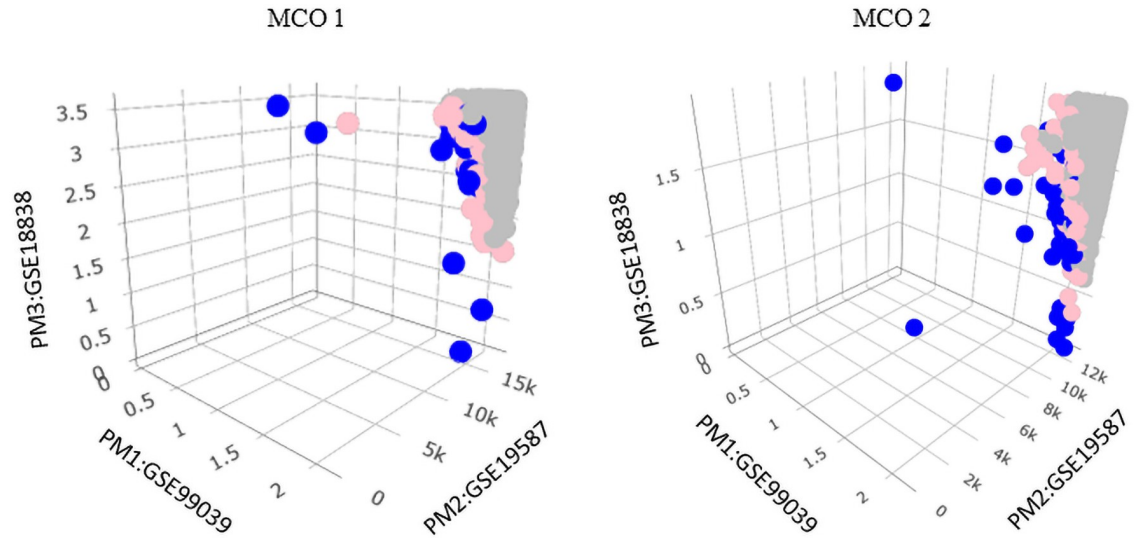
<https://doi.org/10.1371/journal.pone.0262890.g004>

**Table 3. Genes of interest for meta-analysis instances 3–3D-9039/8838/9587 and 4–4D-9039/8838/9587/7475 and the references supporting their roles in Parkinson’s Disease (PD) or neurodegenerative diseases (ND).**

Meta-Analysis Instance	Gene symbol	Gene name	Reference related to	
			PD	ND
3–3D-9039/8838/9587	MMP9	Matrix Metallopeptidase 9	[28–30]	
	RPS11	Ribosomal Protein S11	[37]	
	TUBA1B	Tubulin Alpha 1b, K-ALPHA-1	[35]	
	TUBB2A	Tubulin Beta 2A Class IIa	[16–18]	[19]
4–4D-9039/8838/9587/7475	EEF2	Eukaryotic Translation Elongation Factor 2	[38, 39]	
	MMP9	Matrix Metallopeptidase 9	[28–30]	
	CFD	Complement Factor D	[20]	[21]
	DAZAP2	DAZ Associated Protein 2		
	MYL6	Myosin Light Chain 6	[40]	
	ARHGDIB	Rho GDP Dissociation Inhibitor Beta	[35]	
	RPL18	Ribosomal Protein L18		
	RPS11	Ribosomal Protein S11	[37]	
	CD81	CD81 Molecule	[41]	
	TUBB2A	Tubulin Beta 2A Class IIa	[16–18]	[19]

<https://doi.org/10.1371/journal.pone.0262890.t003>

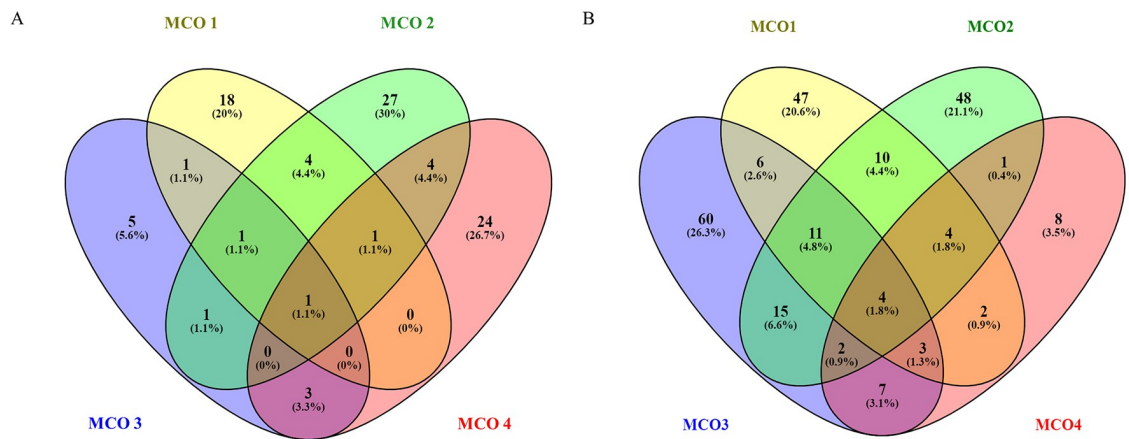




**Fig 5. Meta-analysis instance 3–3D-9039/8838/9587.** Graphical representations for MCO 1 and MCO 2 are associated with this instance. Solutions toward the origin (0,0,0) are more significant.

<https://doi.org/10.1371/journal.pone.0262890.g005>

evidence. This process found out that 9 out of 10 genes (MMP9, RPS11, TUBB2A, EEF2, CFD, DAZAP2, MYL6, ARHGDI1B, RPL18, and CD81) identified in the meta-analysis instances have appeared in the literature as related to PD or neurodegenerative conditions. For instance, MMP9, a protein-coding gene, appears in 17 articles describing a direct association with PD. In [30], MMP9 was identified as a potential marker for Lewy body disorders, i.e. Parkinson’s Disease. In [16], using the LASSO algorithm with 10-fold cross-validation cycles, TUBB2A was one of the 25 genes selected as differentially expressed mRNAs for PD. Furthermore, in [17], it is argued that TUBB2A is a molecular biomarker for PD in the blood, supporting



Common genes between:	Meta-analysis Instance:													
	3-3D-9039/8838/9587							4-4D-9039/8838/9587/7475						
MCO1-MCO2	MMP9	RPS11	TUBA1B	TUBB2A	EEF2	MMP9	CFD	DAZAP2	MYL6	ARHGDI1B	RPL18	RPS11	CD81	TUBB2A
MCO3-MCO4	DDX3Y	VBPI	RPS4Y1		RPS4X	DDX3Y	RPL22	RPS18	MAL	RPS25	RPL35			
MCO1-MCO2-MCO3	TUBA1A				ACTB	PI3	TUBA1A	ALPL	TUBA1B	CST3	PSAP	TUBA1C	CD74	TYROBP
MCO1-MCO2-MCO4	CFL1				RPL41	RPL11	EEF1A1	OAZ1						
MCO1-MCO2-MCO3-MCO4	RPS24				CFL1	RPS24	HBB	PTGDS						

**Fig 6. Venn diagrams of meta-analysis instances 3–3D-9039/8838/9587 in (A) and 4–4D-9039/8838/9587/7475 in (B).** The accompanying table lists the genes in each intersection.

<https://doi.org/10.1371/journal.pone.0262890.g006>

**Table 4. Pathways enriched for the genes of interest identified in meta-analysis.**

ID	Description	GeneRatio	BgRatio	pvalue	p.adjusted	qvalue	Gene	Count
R-HSA-156902	Peptide chain elongation	3/9	89/10856	4.32E-05	0.003275522	0.001607	EEF2/ RPL18/ RPS11	3
R-HSA-156842	Eukaryotic Translation Elongation	3/9	93/10856	4.93E-05	0.003275522	0.001607	EEF2/ RPL18/ RPS11	3
R-HSA-166658	Complement cascade	2/9	58/10856	9.86E-04	0.028553514	0.014011	CFD/ CD81	2
R-HSA-2766	Translation	3/9	291/10856	1.42E-03	0.028553514	0.014011	EEF2/ RPL18/ RPS11	3
R-HSA-192823	Viral mRNA Translation	2/9	89/10856	2.30E-03	0.028553514	0.014011	RPL18/ RPS11	2
R-HSA-2682334	EPH-Ephrin signaling	2/9	92/10856	2.46E-03	0.028553514	0.014011	MMP9/ MYL6	2

<https://doi.org/10.1371/journal.pone.0262890.t004>

similar assertions in [18]. The authors in [17] demonstrated that TUBB2A in reduced expression reasonably predicted PD as a blood biomarker via a meta-analysis of 11 datasets from GEO (8 from substantia nigra and 3 from blood samples) with further validation analyzing mRNA expression levels in the blood of 50 sporadic PD patients and 50 control subjects. In agreement with known biology, TUBB2A was one of the top identified genes from both meta-analysis instances (Table 3). TUBB2A is a strong candidate for more in-depth explorations at the experimental level for PD. The fact that MMP9 and TUBB2A appeared in the results of instances 3–3D-9039/8838/9587 and 4–4D-9039/8838/9587/7475 as well as in 1–2D-9039, supports the sensitivity of our method to detect genes that play a crucial role in the disease under study. Besides this supporting biological evidence on the efficacy of the MCO tool, the code is also computationally efficient as it can meta-analyze up to five datasets simultaneously. The largest instance presented here, 4–4D-9039/8838/9587/7475, took around 10 minutes to process for any of the MCO 1 thru MCO 4 analyses on a personal laptop with a 2.90 GHz Intel Core i7 CPU and 16G RAM.

The ten genes of interest identified in meta-analysis instance 4–4D-9039/8838/9587/7475 were the subject of gene ontology (GO) enrichment analysis using ReactomePA R package [42] and Enrichr tool [43]. Table 4 lists the results from enrichPathway, and the Table 5 the results from Enrichr.

**Table 5. GO biological process.**

Term	P-value	Overlap_genes
Translation (GO:0006412)	0.000137136	[EEF2, RPL18, RPS11]
Cellular macromolecule biosynthetic process (GO:0034645)	0.000423823	[EEF2, RPL18, RPS11]
SRP-dependent cotranslational protein targeting to membrane (GO:0006614)	0.000880239	[RPL18, RPS11]
Cytoplasmic translation (GO:0002181)	0.000939488	[RPL18, RPS11]
Cotranslational protein targeting to membrane (GO:0006613)	0.000959656	[RPL18, RPS11]
Protein targeting to ER (GO:0045047)	0.001150535	[RPL18, RPS11]
Nuclear-transcribed mRNA catabolic process, nonsense-mediated decay (GO:0000184)	0.001382294	[RPL18, RPS11]
neutrophil degranulation (GO:0043312)	0.001462415	[CFD, EEF2, MMP9]
neutrophil activation involved in immune response (GO:0002283)	0.001497690	[CFD, EEF2, MMP9]
neutrophil mediated immunity (GO:0002446)	0.001524499	[CFD, EEF2, MMP9]

<https://doi.org/10.1371/journal.pone.0262890.t005>

Having a short solution list made it possible to perform an in-depth literature search for each gene. The information added to the ontology analyses and found in which pathways the unlisted genes could be included. From the solution list: EEF2, RPL18, RPS11 code for proteins involved in protein synthesis (translation). It has been reported in various studies that protein synthesis is affected in PD and that some ribosomal proteins have expression changes in the condition [38, 44, 45], and eEF2 (the protein product of EEF2) has been reported to be expressed less in PD [45]. The DAZAP2 gene product, not included in the ontology analysis, could also be associated with translation (protein synthesis). The association is through the DAZAP2 gene product role in stress granules (SGs) that enclose different translation system components when the cell is under stress [46]. The DAZAP2 gene product also participates in translation through interactions with eukaryotic initiation factor 4G (<https://www.genecards.org/cgi-bin/carddisp.pl?gene=DAZAP2>). The ARHGDIB and TUBB2A gene products have roles in cytoskeletal organization and dynamics. The expression of genes for proteins involved in cytoskeleton dynamics, such as tubulin, changes in PD [44]. MMP9 and CD81 gene products are involved in cell motility and extracellular matrix dynamics, MMP9 expression changes in PD, and amyotrophic lateral sclerosis [47].

## Supporting information

**S1 File.** Contains a comparative study of MCO Vs CFS, IG and eBayes gene selection methods and all the supporting tables and figures.  
(ZIP)

## Author Contributions

**Conceptualization:** Clara E. Isaza, Mauricio Cabrera-Ríos.

**Data curation:** Isis Narváez-Bandera, Deiver Suárez-Gómez.

**Formal analysis:** Isis Narváez-Bandera.

**Funding acquisition:** Clara E. Isaza.

**Investigation:** Isis Narváez-Bandera, Deiver Suárez-Gómez, Mauricio Cabrera-Ríos.

**Methodology:** Isis Narváez-Bandera.

**Project administration:** Clara E. Isaza, Mauricio Cabrera-Ríos.

**Resources:** Mauricio Cabrera-Ríos.

**Software:** Isis Narváez-Bandera, Deiver Suárez-Gómez.

**Supervision:** Clara E. Isaza, Mauricio Cabrera-Ríos.

**Validation:** Isis Narváez-Bandera, Deiver Suárez-Gómez.

**Visualization:** Isis Narváez-Bandera, Deiver Suárez-Gómez.

**Writing – original draft:** Isis Narváez-Bandera.

**Writing – review & editing:** Isis Narváez-Bandera, Clara E. Isaza, Mauricio Cabrera-Ríos.

## References

1. Tyagi V, Mishra A. A survey on different feature selection methods for microarray data analysis. *International Journal of Computer Applications*. 2013; 67(16). <https://doi.org/10.5120/11482-7181>

2. Lyttleton O, Wright A, Treanor D, Quirke P, Lewis P. Extending the tissue microarray data exchange specification for inclusion of data analysis results. *Journal of pathology informatics*. 2011; 2. <https://doi.org/10.4103/2153-3539.78263> PMID: 21572505
3. Singh P, Shukla AK, Vardhan M. Hybrid approach for gene selection and classification using filter and genetic algorithm. *International Conference on Inventive Computing and Informatics (ICICI)*. IEEE, 2017;832–837.
4. Shukla AK, Tripathi D. Detecting biomarkers from microarray data using distributed correlation based gene selection. *Genes & genomics*. 2020;1–17. PMID: 32040771
5. Shukla AK, Singh P, Vardhan M. DNA Gene Expression Analysis on Diffuse Large B-Cell Lymphoma (DLBCL) Based on Filter Selection Method with Supervised Classification Method. *Computational Intelligence in Data Mining, Singapore*. 2019;783–792. [https://doi.org/10.1007/978-981-10-8055-5\\_69](https://doi.org/10.1007/978-981-10-8055-5_69)
6. Shukla AK, Pippal SK, Gupta S, Ramachandra-Reddy B, Tripathi D. Knowledge discovery in medical and biological datasets by integration of Relief-F and correlation feature selection techniques. *Journal of Intelligent & Fuzzy Systems*. 2020; 38(5):6637–6648 <https://doi.org/10.3233/JIFS-179743>
7. Camacho-Cáceres KI, Acevedo-Díaz JC, Pérez-Marty LM, Ortiz-Mojer M, Irizarry J, Cabrera-Ríos M, et al. Multiple criteria optimization joint analyses of microarray experiments in lung cancer: from existing microarray data to new knowledge. *Cancer medicine*. 2015; 4(12):1884–1900. <https://doi.org/10.1002/cam4.540> PMID: 26471143
8. Cruz-Rivera YE, Perez-Morales J, Santiago YM, Gonzalez VM, Morales L, Cabrera-Ríos M, et al. A Selection of Important Genes and Their Correlated Behavior in Alzheimer's Disease. *Journal of Alzheimer's Disease*. 2018; 65(1):193–205. <https://doi.org/10.3233/JAD-170799> PMID: 30040709
9. Lorenzo E, Camacho-Cáceres KI, Ropelewski A, Rosas JF, Ortiz-Mojer M, Pérez-Marty LM, et al. An optimization-driven analysis pipeline to uncover biomarkers and signaling paths: cervix cancer. *Microarrays*. 2015; 4(2):287–310. <https://doi.org/10.3390/microarrays4020287> PMID: 26388997
10. Isaza C, Rosas JF, Lorenzo E, Marrero A, Ortiz C, Ortiz MR, et al. Biological signaling pathways and potential mathematical network representations: biological discovery through optimization. *Cancer medicine*. 2018; 7(5):1875–1895. <https://doi.org/10.1002/cam4.1301> PMID: 29635835
11. Shamir R, Klein C, Amar D, Vollstedt EJ, Bonin M, Usenovic M, et al. Analysis of blood-based gene expression in idiopathic Parkinson disease. *Neurology*. 2017; 89(16):1676–1683 <https://doi.org/10.1212/WNL.0000000000004516> PMID: 28916538
12. Shehadeh LA, Yu K, Wang L, Guevara A, Singer C, Vance J, et al. SRRM2, a potential blood biomarker revealing high alternative splicing in Parkinson's disease. *PLoS One*. 2010; 8; 5(2):e9104. <https://doi.org/10.1371/journal.pone.0009104> PMID: 20161708
13. Lewandowski NM, Ju S, Verbitsky M, Ross B, Geddie ML, Rockenstein E, et al. Polyamine pathway contributes to the pathogenesis of Parkinson disease. *Proceedings of the National Academy of Sciences*. 2010; 107(39):16970–5. <https://doi.org/10.1073/pnas.1011751107>
14. Locascio JJ, Eberly S, Liao Z, Liu G et al. Association between  $\alpha$ -synuclein blood transcripts and early, neuroimaging-supported Parkinson's disease. *Brain*. 2015; 138(Pt 9):2659–71. <https://doi.org/10.1093/brain/awv202> PMID: 26220939
15. Picillo M, Nicoletti A, Fetoni V, Garavaglia B, Barone P, Pellecchia MT. The relevance of sex in Parkinson's disease: a review. *Journal of neurology*. 2017; 264(8), 1583–1607. <https://doi.org/10.1007/s00415-016-8384-9> PMID: 28054129
16. Jiang F, Wu Q, Sun S, Bi G, Guo L. Identification of potential diagnostic biomarkers for Parkinson's disease. *FEBS open bio*. 2019; 9(8), 1460–1468. <https://doi.org/10.1002/2211-5463.12687> PMID: 31199560
17. Zhang X, Hou NN, Liang YZ, Zhang HQ, Zhang CN, Ji SY, et al. Substantia nigra dysregulated gene spectra and blood biomarkers for Parkinson's disease. Available at SSRN 3426068. 2019.
18. Simunovic F, Yi M, Wang Y, Macey L, Brown LT, Krichevsky AM, et al. Gene expression profiling of substantia nigra dopamine neurons: further insights into Parkinson's disease pathology. *academic.oup.com. Brain*. 2009; 132(7), 1795–1809. <https://doi.org/10.1093/brain/awn323> PMID: 19052140
19. Sferra A, Fattori F, Rizza T, Flex E, Bellacchio E, Bruselles A, et al. Defective kinesin binding of TUBB2A causes progressive spastic ataxia syndrome resembling saccinopathy. *Human molecular genetics*. 2018; 27(11), 1892–1904 <https://doi.org/10.1093/hmg/ddy096> PMID: 29547997
20. Lee M, Wathier M, Love JA, McGeer E, McGeer PL. Inhibition of aberrant complement activation by a dimer of acetylsalicylic acid *Neurobiology of Aging*. 2015; 36(10), 2748–2756. PMID: 26248865
21. Natarajan R, Hagman S, Hämäläinen M, Leppänen T, Dastidar P, Moilanen E, Elovaara I. Adipsin is associated with multiple sclerosis: a follow-up study of adipokines. *Multiple sclerosis international*, 2015. <https://doi.org/10.1155/2015/371734> PMID: 26634156

22. Su L, Wang C, Zheng C, Wei H, Song X. A meta-analysis of public microarray data identifies biological regulatory networks in Parkinson's disease. *BMC medical genomics*. 2018; 11(1), 40. <https://doi.org/10.1186/s12920-018-0357-7> PMID: 29653596
23. Agrawal M, Biswas A. Molecular diagnostics of neurodegenerative disorders. *rontiers in molecular bio-sciences*. 2015; 2, 54. <https://doi.org/10.3389/fmolb.2015.00054> PMID: 26442283
24. Mastrokolias A, Ariyurek Y, Goeman JJ, Van-Duijn E, Roos RA, Van-Der-Mast RC et al. Huntington's disease biomarker progression profile identified by transcriptome sequencing in peripheral blood. *European Journal of Human Genetics*. 2015; 23(10), 1349–1356. <https://doi.org/10.1038/ejhg.2014.281> PMID: 25626709
25. Santiago JA, Potashkin JA. Integrative network analysis unveils convergent molecular pathways in Parkinson's disease and diabetes. *PloS one*. 2013; 8(12), e83940. <https://doi.org/10.1371/journal.pone.0083940> PMID: 24376773
26. Kanthasamy AG, Anantharam V, Zhang D, Latchoumycandane C, Jin H, Kaul S, et al. A novel peptide inhibitor targeted to caspase-3 cleavage site of a proapoptotic kinase protein kinase C delta (PKC $\delta$ ) protects against dopaminergic neuronal degeneration in Parkinson's disease models. *Free Radical Biology and Medicine*. 2006; 41(10), 1578–1589 <https://doi.org/10.1016/j.freeradbiomed.2006.08.016> PMID: 17045926
27. Gordon R, Singh N, Lawana V, Ghosh A, Harischandra DS, Jin H, et al. Protein kinase C $\delta$ . upregulation in microglia drives neuroinflammatory responses and dopaminergic neurodegeneration in experimental models of Parkinson's disease. *Neurobiology of disease*. 2016; 93, 96–114. <https://doi.org/10.1016/j.nbd.2016.04.008> PMID: 27151770
28. Lorenzl S, Calingasan N, Yang L, Albers DS, Shugama S, Gregorio J, et al. Matrix metalloproteinase-9 is elevated in 1-methyl-4-phenyl-1, 2, 3, 6-tetrahydropyridine-induced parkinsonism in mice. *Neuromolecular medicine* 2004; 5(2), 119–131. <https://doi.org/10.1385/NMM:5:2:119> PMID: 15075439
29. Singh D, Srivastava SK, Chaudhuri TK, Upadhyay G. Multifaceted role of matrix metalloproteinases (MMPs). *Front. Mol. Biosci.*, vol. 2, no. MAY, p. 19, May 2015.
30. Maetzler W, Deleersnijder W, Hanssens V, Bernard A, Brockmann K, Marquetand J, et al. GDF15/MIC1 and MMP9 cerebrospinal fluid levels in Parkinson's disease and Lewy body dementia. *PLoS ONE*. 2016; 11(3) <https://doi.org/10.1371/journal.pone.0149349> PMID: 26938614
31. Goldman SM, Kamel F, Ross GW, Jewell SA, Marras C, Hoppin JA, et al. Peptidoglycan recognition protein genes and risk of Parkinson's disease. *Movement Disorders*. 2014; 29(9), 1171–1180. <https://doi.org/10.1002/mds.25895> PMID: 24838182
32. Coskun P, Wyrembak J, Schriener SE, Chen HW, Marciniack C, LaFerla F, et al. A mitochondrial etiology of Alzheimer and Parkinson disease. *Biochimica et Biophysica Acta (BBA)-General Subjects*. 2012; 1820(5), 553–564. <https://doi.org/10.1016/j.bbagen.2011.08.008> PMID: 21871538
33. George G, Singh S, Lokappa SB, Varkey J. Gene co-expression network analysis for identifying genetic markers in Parkinson's disease—a three-way comparative approach. *Genomics*. 2019; 111(4), 819–830. <https://doi.org/10.1016/j.ygeno.2018.05.005> PMID: 29852216
34. Khwanraj K, Madlah S, Grataitong K, Dharmasaroja P. Comparative mRNA expression of eEF1A isoforms and a PI3K/Akt/mTOR Pathway in a cellular model of Parkinson's disease. *Parkinson's Disease*, 2016. <https://doi.org/10.1155/2016/8716016> PMID: 26981313
35. Kim JM, Lee KH, Jeon YJ, Oh JH, Jeong SY, Song IS, et al. Identification of genes related to Parkinson's disease using expressed sequence tags. *DNA research*. 2006; 13(6), 275–286. <https://doi.org/10.1093/dnares/dsl016> PMID: 17213182
36. Santiago JA, Potashkin JA. Blood transcriptomic meta-analysis identifies dysregulation of hemoglobin and iron metabolism in Parkinson' disease. *Frontiers in Aging Neuroscience*. 2017; 9, 73. <https://doi.org/10.3389/fnagi.2017.00073> PMID: 28424608
37. Moran LB, Graeber MB. Towards a pathway definition of Parkinson's disease: A complex disorder with links to cancer, diabetes and inflammation. *Neurogenetics*. 2008; 9(1), 1–13. <https://doi.org/10.1007/s10048-007-0116-y> PMID: 18196299
38. Garcia-Esparcia P, Hernández-Ortega K, Koneti A, Gil L, Delgado-Morales R, Castaño E, et al. Altered machinery of protein synthesis is region- and stage-dependent and is associated with  $\alpha$ -synuclein oligomers in Parkinson's disease. *Acta neuropathologica communications*. 2015; 3(1), 76. <https://doi.org/10.1186/s40478-015-0257-4> PMID: 26621506
39. Jan A, Jansonius B, Delaidelli A, An YA, Ferreira N, Smits LM, et al. Activity of translation regulator eukaryotic elongation factor-2 kinase is increased in Parkinson disease brain and its inhibition reduces alpha synuclein toxicity. *Acta neuropathologica communications*. 2018; 6(1), 54. <https://doi.org/10.1186/s40478-018-0554-9> PMID: 29961428

40. Rocha S, Freitas A, Guimaraes SC, Vitorino R, Aroso M, Gomez-Lazaro M. Biological Implications of Differential Expression of Mitochondrial-Shaping Proteins in Parkinson's Disease. *Antioxidants*. 2018; 7(1), 1. <https://doi.org/10.3390/antiox7010001>
41. Russo I, Bubacco L, Greggio E. Exosomes-associated neurodegeneration and progression of Parkinson's disease. *American journal of neurodegenerative disease*. 2012; 1(3), 217–225. PMID: [23383394](https://pubmed.ncbi.nlm.nih.gov/23383394/)
42. Yu G and He QY. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Molecular BioSystems*. 2016; 12(2):477–479. <https://doi.org/10.1039/C5MB00663E>
43. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC bioinformatics*. 2013; 14(1), 1–14. <https://doi.org/10.1186/1471-2105-14-128> PMID: [23586463](https://pubmed.ncbi.nlm.nih.gov/23586463/)
44. Fernandes HJ, Patikas N, Foskolou S, Field SF, Park JE, Byrne ML, et al. Single-cell transcriptomics of Parkinson's disease human in vitro models reveals dopamine neuron-specific stress responses. *Cell Reports*. 2020; 33(2), 108263. <https://doi.org/10.1016/j.celrep.2020.108263> PMID: [33053338](https://pubmed.ncbi.nlm.nih.gov/33053338/)
45. Kapur M, Monaghan CE, Ackerman SL. Regulation of mRNA translation in neurons—a matter of life and death. *Neuron*, 2017; 96(3), 616–637. <https://doi.org/10.1016/j.neuron.2017.09.057> PMID: [29096076](https://pubmed.ncbi.nlm.nih.gov/29096076/)
46. Kim JE, Ryu I, Kim WJ, Song OK, Ryu J, Kwon MY, et al. Proline-rich transcript in brain protein induces stress granule formation. *Molecular and cellular biology* 2008; 28(2), 803–813. <https://doi.org/10.1128/MCB.01226-07> PMID: [17984221](https://pubmed.ncbi.nlm.nih.gov/17984221/)
47. He X, Zhang L, Yao X, Hu J, Yu L, Jia H, et al. Association studies of MMP-9 in Parkinson's disease and amyotrophic lateral sclerosis. *PLoS One*. 2013; 8(9), e73777. <https://doi.org/10.1371/journal.pone.0073777> PMID: [24040066](https://pubmed.ncbi.nlm.nih.gov/24040066/)