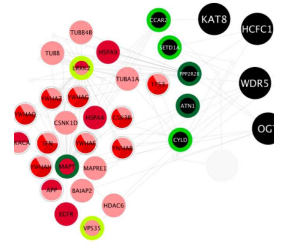




Dec 06, 2022

Version 1

In Silico analysis links the NSL complex to Parkinson's disease and the mitochondria – Protein-protein interaction data to functional enrichment analysis V.1



DOI

<https://dx.doi.org/10.17504/protocols.io.5qpvorb19v4o/v1>Katie Kelly^{1,2}, c.manzoni^{3,2}, Patrick Lewis^{4,1,2}, Helene Plun-Favreau^{1,2}¹Department of Neurodegenerative Disease, UCL Institute of Neurology, Queen Square, London WC1N 3BG, UK;²Aligning Science Across Parkinson's (ASAP) Collaborative Research Network, Chevy Chase, MD, 20 815, USA.;³Department of Pharmacy, University of Reading, Reading, RG6 6AX, UK;⁴The Royal Veterinary College, Royal College Street, London NW1 0TU, UK

Katie Kelly

Create & collaborate more with a free account

Edit and publish protocols, collaborate in communities, share insights through comments, and track progress with run records.

[Create free account](#)OPEN  ACCESS**DOI:** <https://dx.doi.org/10.17504/protocols.io.5qpvorb19v4o/v1>**External link:** <https://zenodo.org/record/7346957#.Y48jfYfP02x>

Protocol Citation: Katie Kelly, c.manzoni , Patrick Lewis, Helene Plun-Favreau 2022. In Silico analysis links the NSL complex to Parkinson's disease and the mitochondria – Protein-protein interaction data to functional enrichment analysis. **protocols.io** <https://dx.doi.org/10.17504/protocols.io.5qpvorb19v4o/v1>

License: This is an open access protocol distributed under the terms of the **Creative Commons Attribution License**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Protocol status: Working

We use this protocol and it's working

Created: October 26, 2022

Last Modified: January 09, 2023

Protocol Integer ID: 71840

Keywords: Parkinson's disease, NSL complex, Mitophagy, In silico, Protein-protein interaction (PPI), Mito-CORE network Interactome, ASAPCRN, relevance of the nsl mitochondrial interactome, mitochondrial interactome of the nsl complex, nsl mitochondrial interactome, protein interaction data to functional enrichment analysis, nsl complex to parkinson, mitochondrial interactome, protein interaction data, assay of gwas candidate, interactome of the nsl complex, pd gene, identifying common risk variant, parkinson, wide association study, protein interaction, nsl interactome, mitochondria, applying pd gene, functional enrichment analysis, relevance to pd progression, genome, bioinformatics approach, regulation of pink1, common risk variant, developed mitophagy, oriented nsl interactome, gwas candidate, enrichment analysis, nsl complex, gwa, mitophagy

Funders Acknowledgements:

The Michael J. Fox Foundation for Parkinson's Research (MJFF) and the Aligning Science Across Parkinson's (ASAP) initiative.

Grant ID: ASAP-000478

Masonic Charitable Foundation (PhD Prize Studentship)

Grant ID: -

Abstract

Whilst the majority (~90-95%) of PD cases are sporadic, much of our understanding of the pathophysiological basis of disease can be traced back to the study of rare, monogenic forms of disease. However, in the past decade, the availability of Genome-Wide Association Studies (GWAS) has facilitated a shift in focus, toward identifying common risk variants conferring an increased risk of developing PD across the population.

A recently developed mitophagy screening assay of GWAS candidates, has functionally implicated the non-specific lethal (NSL) complex, a chromatin remodeler, in the regulation of PINK1-mitophagy. Here, a bioinformatics approach has been taken to investigate the interactome of the NSL complex, to unpick its relevance to PD progression. The mitochondrial interactome of the NSL complex has been built, mining 3 separate repositories: PINOT, HIPPIE and MIST, for curated, literature-derived protein-protein interaction (PPI) data. A multi-layered approach has been taken to; i) build the 'mitochondrial' NSL interactome, applying PD gene-set enrichment analysis to explore the relevance of the NSL mitochondrial interactome to PD and, ii) build the PD-oriented NSL interactome, using functional enrichment, to uncover biological pathways underpinning the NSL /PD association.

Downloading the Protein-Protein Interaction (PPI) Data

- All code can be found here : [v1.0.0 W-PPI-NA_NSL](#)
The pipeline to derive the **first layer** interactome can be found in Figure 1.

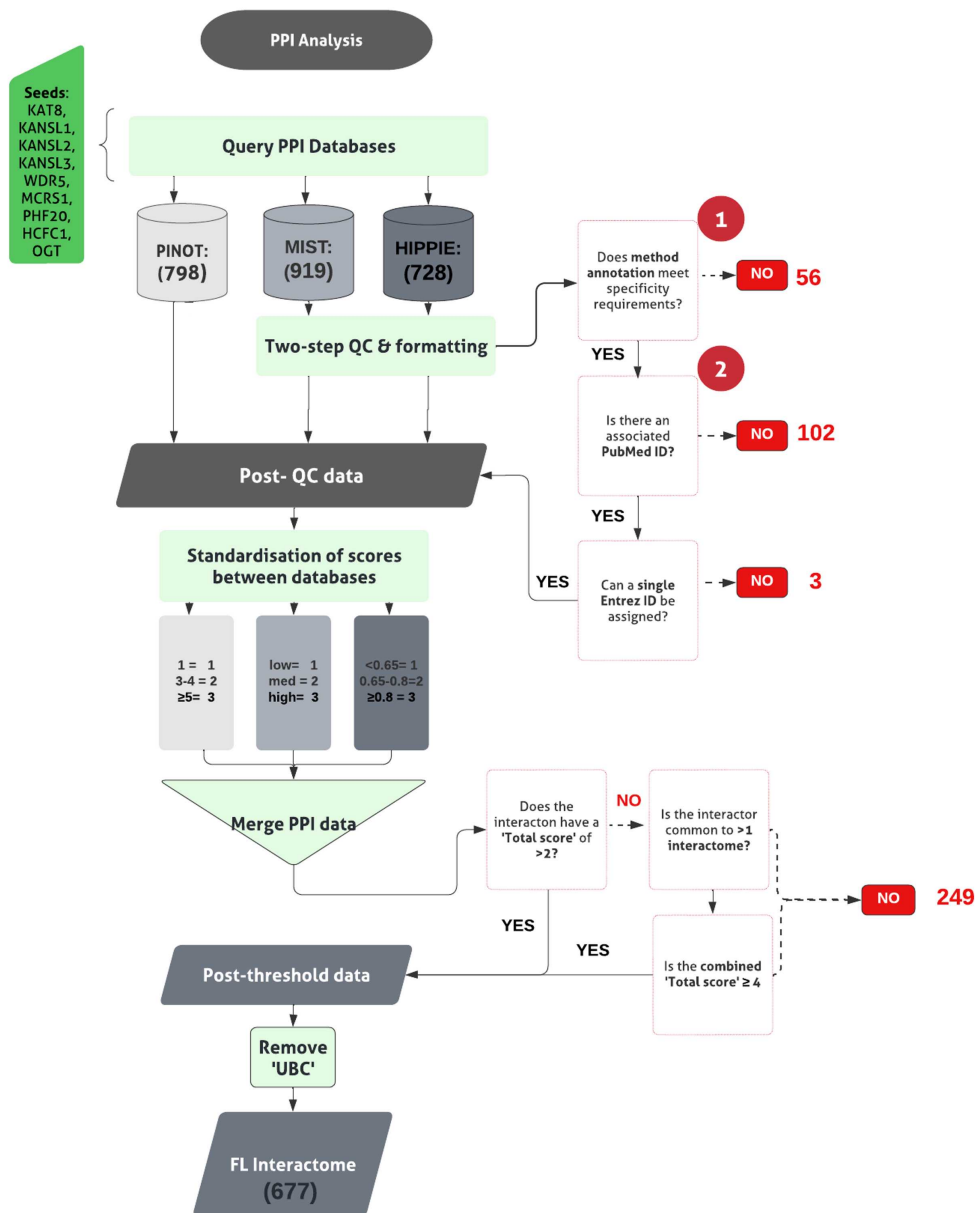


Figure 1. W-PPI-NA pipeline. Generating the first layer interactome of the NSL complex. The 'Seeds' are the nine members of the NSL complex. Circled numbers (1 & 2) indicate the two stages of quality control (QC) applied. Numbers provided in brackets indicate total number of interactions retained at each stage.

2 Collect PPIs for NSL seeds using 3 different web-based tools;

1) **PINOT** (Version 1.1 with lenient filter option) (Protein Interaction Network Online Tool) (Tomkins, Ferrari et al. 2020, DOI: <http://dx.doi.org/10.1186/s12964-020-00554-5>)

2) **HIPPIE** with no threshold on interaction score (Human Integrated Protein-Protein Interaction rEference) (Alanis-Lobato, Andrade-Navarro et al. 2017 ; DOI: <https://doi.org/10.1093/nar/gkw985>; RRID:SCR_014651).

3) **MIST v5.0** (Molecular Interaction Search Tool) (Hu, Vinayagam et al. 2018 ; DOI: [10.1093/nar/gkx1116](https://doi.org/10.1093/nar/gkx1116)).

Note

Each resource permitted interrogation of a selection of IMEx consortium: <https://www.imexconsortium.org/> (IMEx - The International Molecular Exchange Consortium ;RRID:SCR_002805) associated repositories, to obtain literature-derived, curated PPI data.

3 PPI data obtained using MIST and HIPPIE are subjected to quality control (QC), QC steps 1 & 2 (already integrated within the PINOT pipeline) to remove low quality data.

Note

In Excel, i) **QC1** :Entries lacking “*interaction detection method*” annotation, or ii) **QC2**: a PubMed ID, are removed.

Note

Data downloaded from PINOT is parsed in R, to retain and rename only relevant dataframe columns (code found in file *1.3. Standardisation of Score (GitHub)*).

4 Formatting between the output files is standardized and interactors' IDs are converted to the approved EntrezID, UniprotID and HGNC gene name.

**Note**

To do this, we merge the lists of interactions, with a *Gene dictionary in R*. This is a complete list of 19,947 genes (obtained by processing the file downloaded from HGNC: [(https://www.genenames.org/download/statistics-and-files/) on August 2019 (Tomkins, Ferrari et al. 2020 ; DOI: http://dx.doi.org/10.1186/s12964-020-00554-5)]).

- 5 Where 'UBC', a ubiquitin moiety, is identified as an interactor within the *first layer*, review the supporting publication.

Note

Ubiquitin is understood to be conjugated to proteins as a 'flag' for degradation. As such, we suggest it might introduce non-specific protein interactions into the analysis.

- 6 Where 'UBC', a ubiquitin moiety, is identified as an interactor within the *first layer*, review the supporting publication.

Note

Ubiquitin is understood to be conjugated to proteins as a 'flag' for degradation. It might introduce non-specific protein interactions into the analysis.

- 7 Merge interaction data, across the 3 databases to generate a single file for each seed's interactome.

Note

A 'reference' protein list was first generated, for each interactome. This contained the list of unique interactors found for each seed, across all three databases. Then, interaction lists for each interactome, obtained from HIPPIE, MIST and PINOT, could be merged (code found in file 1.3. *Standardisation of Score (GitHub)*).

Merging and Thresholding the PPIs

- 8 Calculate the total score (CS_T) for each interaction the (CS_T) was calculated as:

$$CS_T = CS_P + CS_M + CS_H$$

Note

The (CS_T) ranges therefore from a min = 1 (PPI reported in only 1 database with low confidence), to a max = 9 (PPI reported by the 3 databases, always with high confidence).

9

Apply an arbitrary score threshold ($CS_T > 2$), to filter and remove lower confidence PPI data lacking reproducibility.

- 10 Merge interaction data, across the 3 databases to generate a single file for each seed's interactome. For each interactor the (CS_T) was calculated as:

Note

For an interaction to have a $CS_T = 3$, it must be reported either with low confidence across all 3 databases, or with moderate or high confidence in a single database. In the case that the interaction is reported with low confidence across 3, we reason that it has at a minimum passed the stringent QC of PINOT, and thus have retained the interaction.

- 11 If interactions that failed to meet the threshold, interrogate further, to identify those interactors bridging >1 interactome.
- 12 For those interactors appearing within >1 interactome, apply a multi-interactome threshold represented by a $CS_T \geq 4$ across interactomes. Retain those meeting this multi-interactome threshold.
- 13 Combine all seed specific interaction lists, to obtain the *first layer interactome*.
- 14 Generate the list of unique interactors within the *first layer interactome* (code found in file 1.3. *Standardisation of Score (GitHub)*).

**Note**

A single column within the multi-column dataframe will be retained (Interactor Entrez ID). Duplicates will be removed.

- 15 Where 'UBC', a ubiquitin moiety, is identified as an interactor within the *first layer*, review the supporting publication. Unless the interaction being studied is specific, remove.

Note

Ubiquitin is understood to be conjugated to proteins as a 'flag' for degradation. As such, we suggest it might introduce non-specific protein interactions into the analysis.

Generating the Mito-CORE Network

- 16 The pipeline to derive the ***Mito-CORE network*** can be found in *Figure 2*.

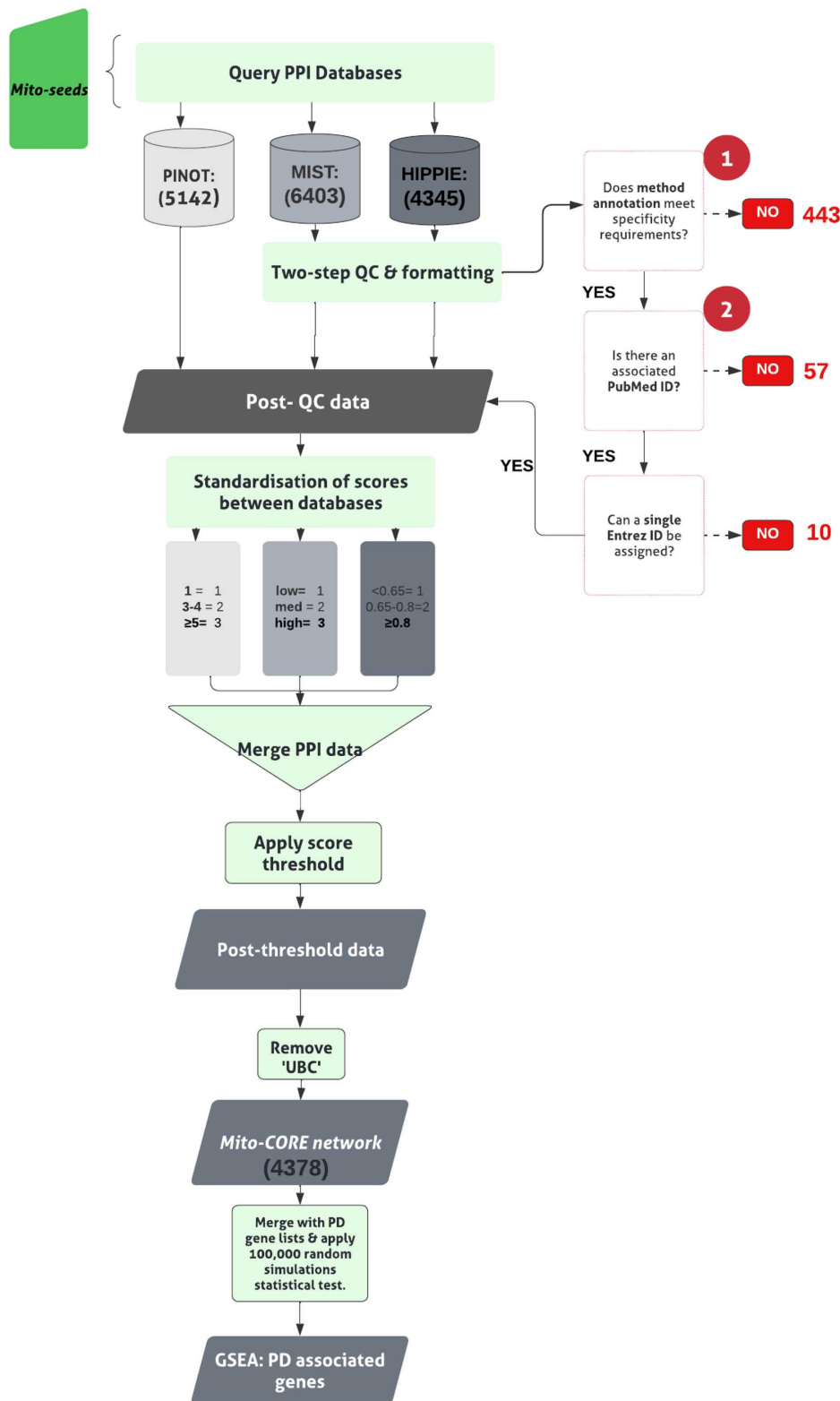


Figure 2. W-PPI-NA pipeline. Building the *Mito-CORE network*, and application of PD Gene-set enrichment analysis (GSEA). 'Mito-seeds' refers to the mitochondrial first layer members of the NSL interactome. Circled numbers (1 & 2) indicate the two stages of quality control (QC) applied . Numbers provided in brackets indicate total number of interactions retained at each stage. * score threshold is applied as described in the

pipeline in Figure 1, after the 'Merge PPI data' step.

- 17 First, prioritise members of the *first layer* with mitochondrial annotation (- OGT, since it was a seed to derive the *first layer* interactome). Here, these are termed '*Mito seeds*'.

Note

Here, proteins with mitochondrial annotation are obtained via 2 independent inventories:
i) **AmiGO2 encyclopedia** (AmiGO (RRID:SCR_002143)), to derive experimentally determined mitochondrial protein lists. Two accession terms were used: **GO:0005759**, to obtain proteins annotated to the "mitochondrial matrix" and **GO:0031966** for proteins annotated to the "mitochondrial membrane". In both cases, 'Homo sapiens' should be specified as the search organism.

ii) the Human **MitoCarta3.0 dataset** (MitoCarta (RRID:SCR_018165)) to retrieve proteins for which a Mitochondrial Targeting Sequence (MTS) has been identified.

Convert interactors' IDs to the approved EntrezID, UniprotID and HGNC gene name using the *Gene dictionary*. Remove proteins with nonunivocal conversions to these 3 identifiers.

Combine i) with ii) to generate the mitochondrial genes list.

- 18 Merge each list of mitochondrial proteins with the *first layer* interactome, to find overlaps. The overlaps represent members of the mitochondrial interactome for the NSL complex. (code found in file *1.5. Enrichment Analyses: Mitochondrial Proteins(GitHub)*).
- 19 Input *mito seeds* into all three PPI tools, to obtain the *second layer*. The *NSL seeds* together with the *Mito seeds*, and *second layer* interactors form the complete **Mito-CORE network**.

Gene Set Enrichment Analysis (GSEA)

- 20 Conduct GSEA for PD associated genes by comparing the members of the interactome under investigation (*first layer* alone or *complete Mito-CORE network*) to a list of 180 unique PD associated genes;

Note

The PD associated gene list is generated by consulting 3 publicly accessible resources:

i) PanelApp v 1.68 diagnostic grade genes (green annotations) for PD and Complex Parkinsonism (Martin, Williams et al. 2019)(**Gene Panel: Parkinson's Disease and Complex Parkinsonism (Version 1.108)**).

ii) the latest GWAS meta-analysis (Nalls, Blauwendraat et al. 2019).

To each of the gene lists above, convert interactors' IDs to the approved EntrezID, UniprotID and HGNC gene name using the Gene dictionary. Remove proteins with nonunivocal conversions to these 3 identifiers.

iii) a list of 15 genes associated with Mendelian PD, obtained from a recent W-PPI-NA (Ferrari, Kia et al. 2018).

Combine the genes from i, ii, and iii to generate a PD associated genes list.

- 21 Merge the list of 180 PD associated genes with the list of unique (*first layer / Mito-CORE network*) interactors, to find overlaps between the two lists. The overlaps represent PD associated proteins within the direct interactome/mitochondrial interactome for the NSL complex (code found in file *1.6. Enrichment Analyses: PD-associated genes (GitHub)*).
- 22 Repeat the above step with the list of 15 Mendelian PD genes, to ascertain enrichment of this more stringent list.

Note

Intersections between the *first layer* and the PD-associated gene list will be termed '*PD-seeds*'.

Statistical Evaluation via Random Networks Simulation

- 23 Use an '100,000 random simulations' test of significance to validate statistical significance of overlaps of PD genes with the *first layer* and complete **Mito-CORE network** (code found in file *100,000 Random Simulations testing (GitHub)*).

Note

100,000 random genes, equivalent in length to *first layer/complete Mito-CORE network*, are obtained using the R random sampling function, from the *Gene dictionary*. Running the code compares each random list to the PD associated gene list, keeping track of the matches. The code then allows comparison of the distribution of random matches to the real number of experimental matches and , via the *p*-norm function. A *p*-value for the enrichment is returned.

Generating the PD-CORE Network

24 The pipeline to derive the **PD-CORE network** can be found in *Figure 3*.

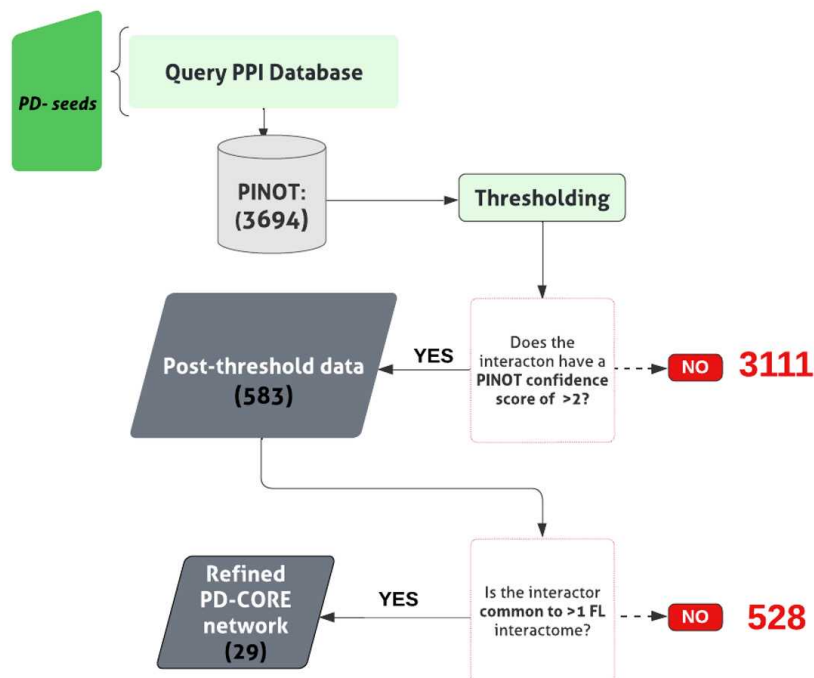


Figure 3. W-PPI-NA pipeline. 'PD-seeds' refers to the PD associated *first layer* members. Numbers provided in brackets indicate total number of interactions/interactors retained at each stage.

- 25 Apply an arbitrary confidence threshold of '*CSp* >2', eliminating data with just a single publication and method from the downstream analysis (code found in file 1.7 *Functional enrichment analysis (GitHub)*).
- 26 Once again, convert interactors' IDs to the approved EntrezID, UniprotID and HGNC gene name using the *Gene dictionary*.



27 Remove proteins with nonunivocal conversions to these 3 identifiers.

28 To remove background noise, keep only members of the *second layer* bridging >1 PD seed within the *PD-CORE* network.

Note

This step removes protein interactors that are private to 1 PD seed only.

29 The NSL seeds together with the PD seeds, and the non-private *second layer* interactors from the complete ***PD-CORE network***.

Functional Enrichment Analysis

30 The general pipeline for this analysis can be found in *Figure 4*.

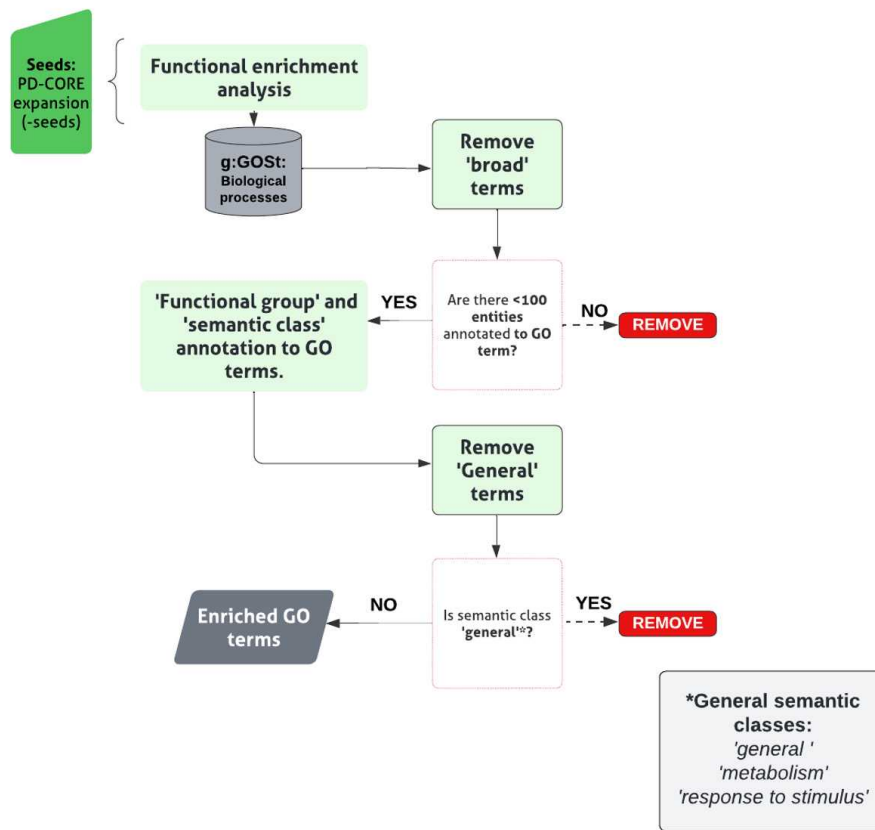


Figure 4. General pipeline for functional enrichment analysis. Grey box indicates Semantic Classes (SCs) removed from the analysis, as they are classified as ‘general’.

- 31 Assess enrichment of particular biological processes within the *PD-CORE* network, members (- NSL seeds), by inputting into the **g:Profiler search tool**, g:GOST (G:Profiler ; Ashburner, Ball et al. 2000, Gene Ontology 2021; RRID:SCR_006809).
- 32 Conduct enrichment for GO terms associated with ‘Biological Processes (BPs)’ only, with all other analysis settings left unadjusted, generating a list of enriched GO:BP terms.
- 33 Apply a threshold to the list of enriched GO:BP terms, to retain those with term size <100 thus effectively removing ‘broad’ GO:BP terms. (code found in file *1.7 Functional enrichment analysis (GitHub)*).
- 34 Assign remaining terms to custom-made ‘semantic classes’(SC), accompanied by a parent ‘functional group’(FG).

**Note**

Assignment is manual.

35 Discard generic terms (classified in the semantic classes of: General , Metabolism, and Response to Stimulus) from further analysis.

36 Pool GO:BP terms contributing to each semantic class to identify the list of proteins within the network contributing to the enrichment of that specific semantic class.

Note

The lowest *p*-value of all GO terms associated with a single semantic class is selected, to represent enrichment of the semantic class.

37 The final list of semantic classes, within each functional group represents those enriched within the network.