protocols.io

VERSION 3

JUN 28, 2023

Protocol status: Working
We use this protocol and it's working

Protein network analysis links the NSL complex to Parkinson's disease via mitochondrial and nuclear biology – Protein-protein interaction data to functional enrichment analysis V.3

Katie Kelly[1,2], c.manzoni[3,2], Patrick Lewis[4,1,2], Helene Plun-Favreau[1,2]

[1]Department of Neurodegenerative Disease, UCL Institute of Neurology, Queen Square, London WC1N 3BG, UK;
[2]Aligning Science Across Parkinson's (ASAP) Collaborative Research Network, Chevy Chase, MD, 20 815, USA.;
[3]Department of Pharmacy, University of Reading, Reading, RG6 6AX, UK;
[4]The Royal Veterinary College, Royal College Street, London NW1 0TU, UK
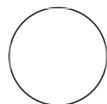
Katie Kelly

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.

Created: Jun 28, 2023

Last Modified: Jun 28, 2023

PROTOCOL integer ID:
84155

Keywords: Parkinson's
disease, NSL complex,
Mitophagy, In silico, Protein-
protein interaction (PPI),
Mito-CORE network
Interactome, ASAPCRN

Downloading and merging the Protein-Protein Interaction (...

1   ▪ All code can be found here : 10.5281/zenodo.7875447.

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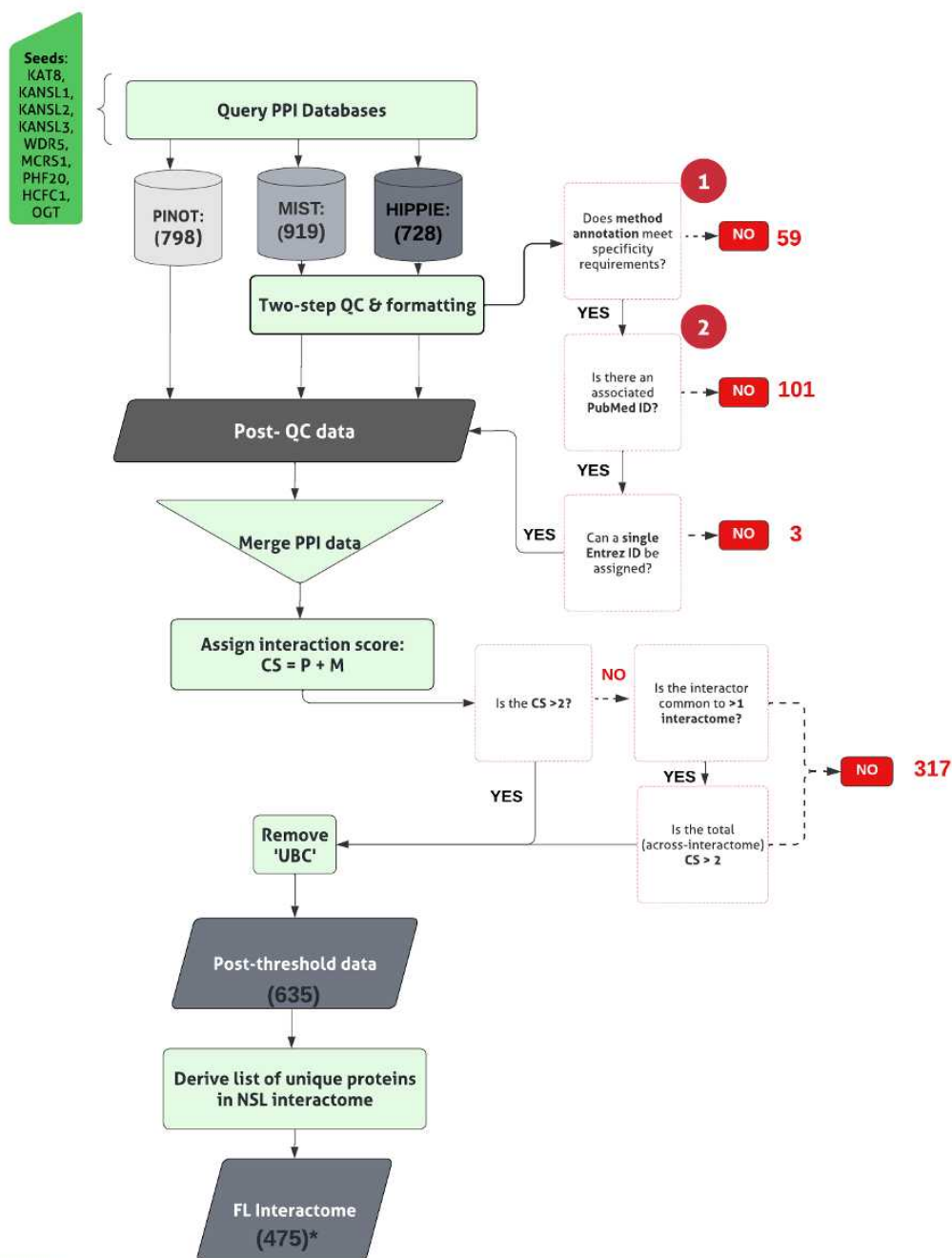
Figure 1. W-PPI-NA pipeline. Generating the first layer interactome of the NSL complex. The 'Seeds' are the nine members of the NSL complex. Circled numbers (1 & 2) indicate the two stages of quality control (QC) applied. Numbers provided in brackets indicate total number of interactions/interactors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

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](#)).

> **Note**
>
> Each resource permits 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.

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

5   Prior to merging the results for each interaction, files are parsed to identify the number of times the interaction was i) observed via a unique methodological technique and ii) reported in a unique publication.

5.1   Apply PINOT method grouping to the interactions downloaded from HIPPIE and MIST, to ensure consistency between the results from each database. To do so, download the 'Method conversion table' ; https://www.reading.ac.uk/bioinf/PINOT/PINOT_help.html#select from PINOT and convert methods according to the MI code.

> **Note**
>
> *Where the method code is not included within the PINOT method conversion table, it must be manually annotated by entering the MI code into OLS (the Ontology Lookup Service) (https://www.ebi.ac.uk/ols/index) and assigning a suitable method name from the PINOT conversion table (Supplementary table 2;* 10.5281/zenodo.7516685*)*

5.2   Parse files from each database to generate a separate dataframe containing 'publication' observations (for calculation of the publication score (PS)), and 'method' observations (for calculation of the method score (MS)). Unique observations for each interaction in each dataframe are allocated an individual row.

Thresholding the PPIs

6

Merge ' publication' observations and 'method' observation files. The number of rows occupied by each interaction corresponds to the number of observations. The CS for each interaction can be calculated calculated as:

$$CS = P + M$$

6.1   Apply a score threshold (CS > 2), to filter and remove lower confidence PPI data lacking reproducibility.

7   If interactions that failed to meet the threshold, interrogate further, to identify those interactors bridging >1 interactome.

8 For those interactors appearing within >1 interactome, apply a multi-interactome threshold represented by a $CS > 2$ across interactomes. Retain those meeting this multi-interactome threshold.

9 Combine those interactions meeting the single and multi- interactome threshold, to generate the *first layer* interactome.

10 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.

11 Generate the list of unique interactors within the *first layer interactome*

Note

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

Generating the Mito-CORE Network

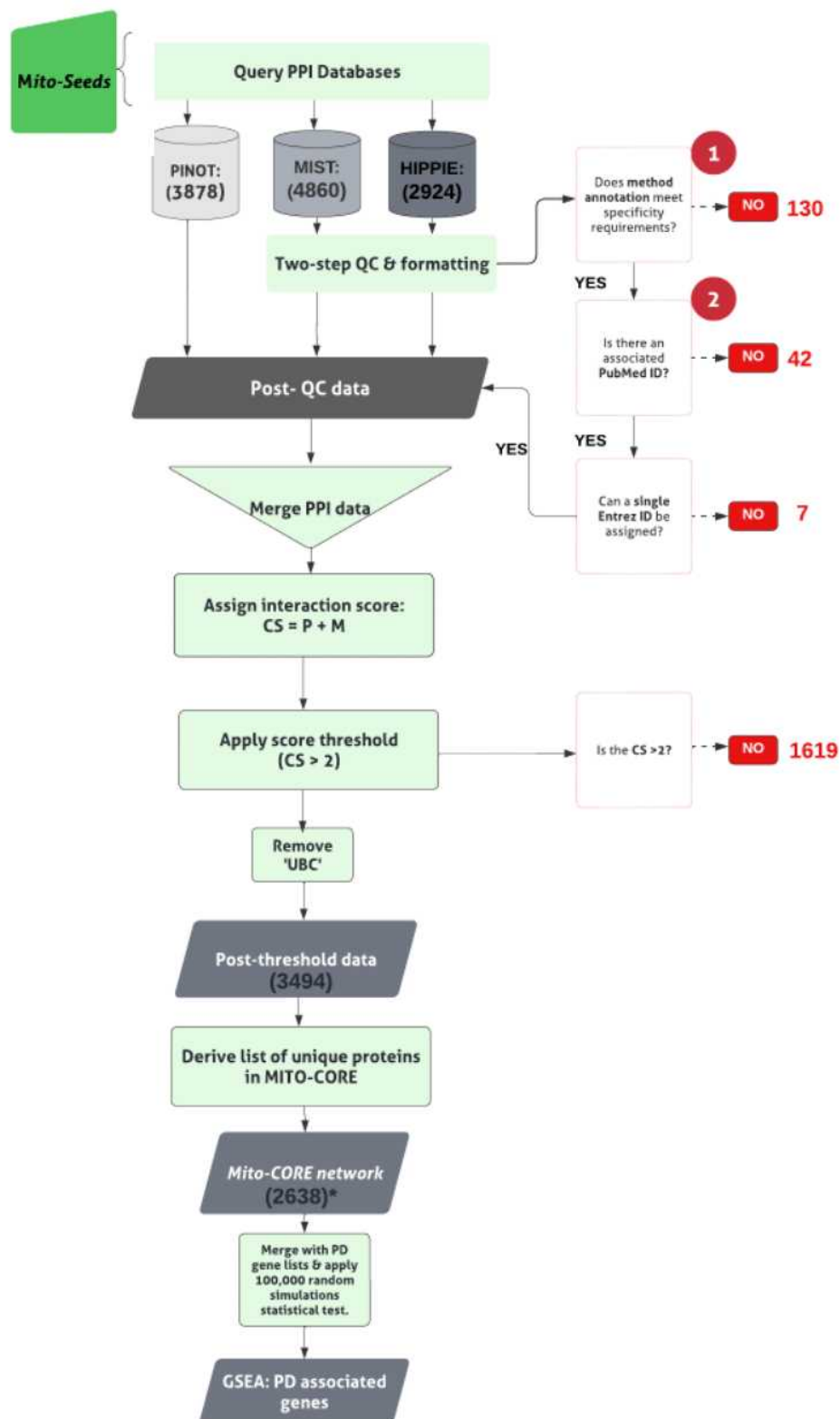12 The pipeline to derive the **Mito-CORE network** can be found in *Figure 2.*

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 layermembers 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/interactors retained at each stage. *Mito-seeds + second layer interactors (-NSL seeds).

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

> **Note**
>
> 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** *(Supplementary table 4;* *10.5281/zenodo.7516685*)

14 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.

15 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)

16   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** (Supplementary table 3; 10.5281/zenodo.7516685).

17   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

18   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

19   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 a library of 19,947 genes. 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

20 The pipeline to derive the **PD-CORE network** can be found in *Figure 3.*
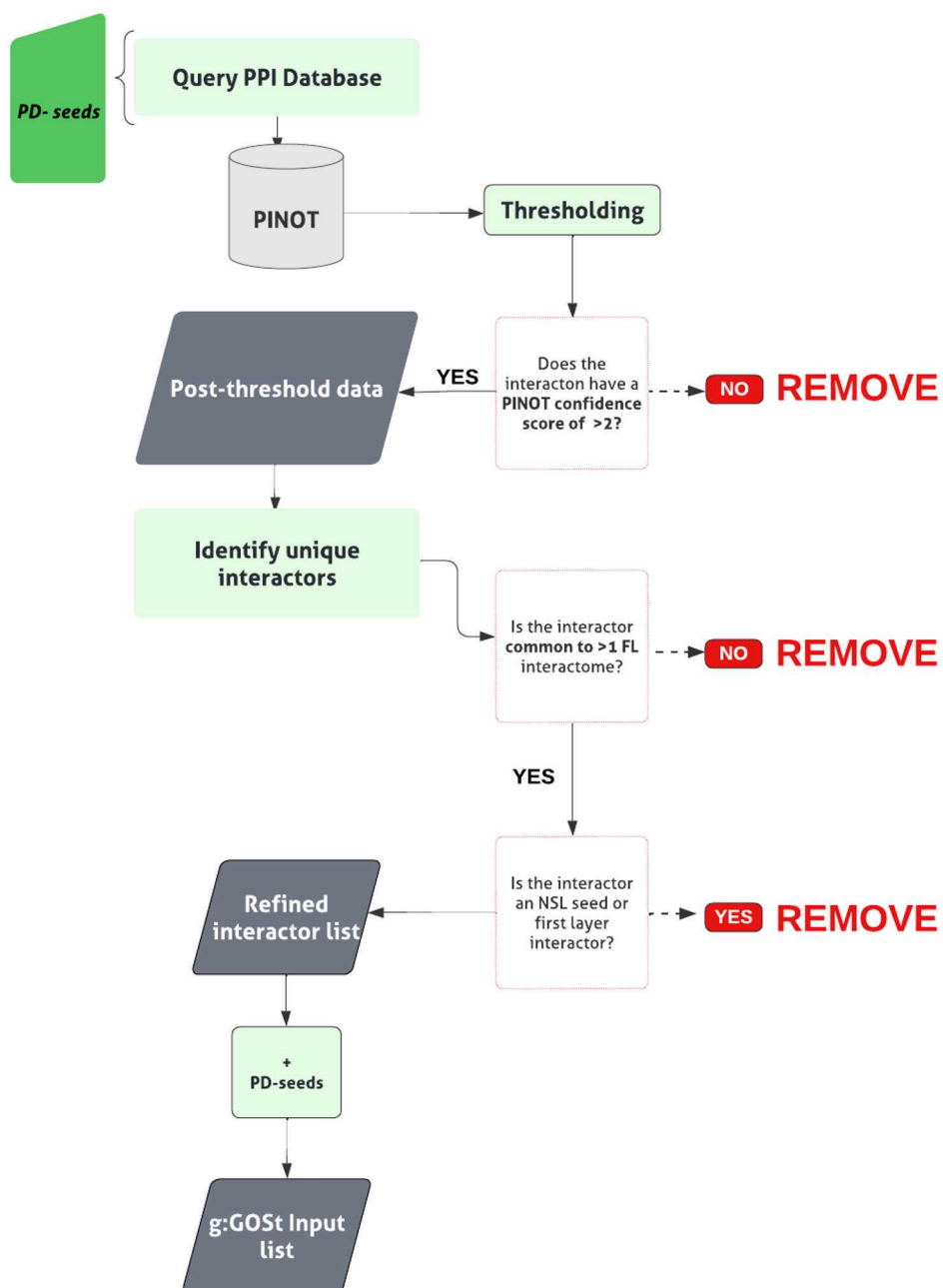
Figure 3. W-PPI-NA pipeline. The 'PD-seeds' refers to the PD associated first layer members.

21   Input PD *seeds* into PINOT to obtain the *second layer of the PD-CORE network.*

22   Apply an arbitrary confidence threshold 'CS >2', eliminating data with just a single publication and method from the downstream analysis

23   Once again, convert interactors' IDs to the approved EntrezID, UniprotID and HGNC gene name.

24   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.

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

Functional Enrichment Analysis

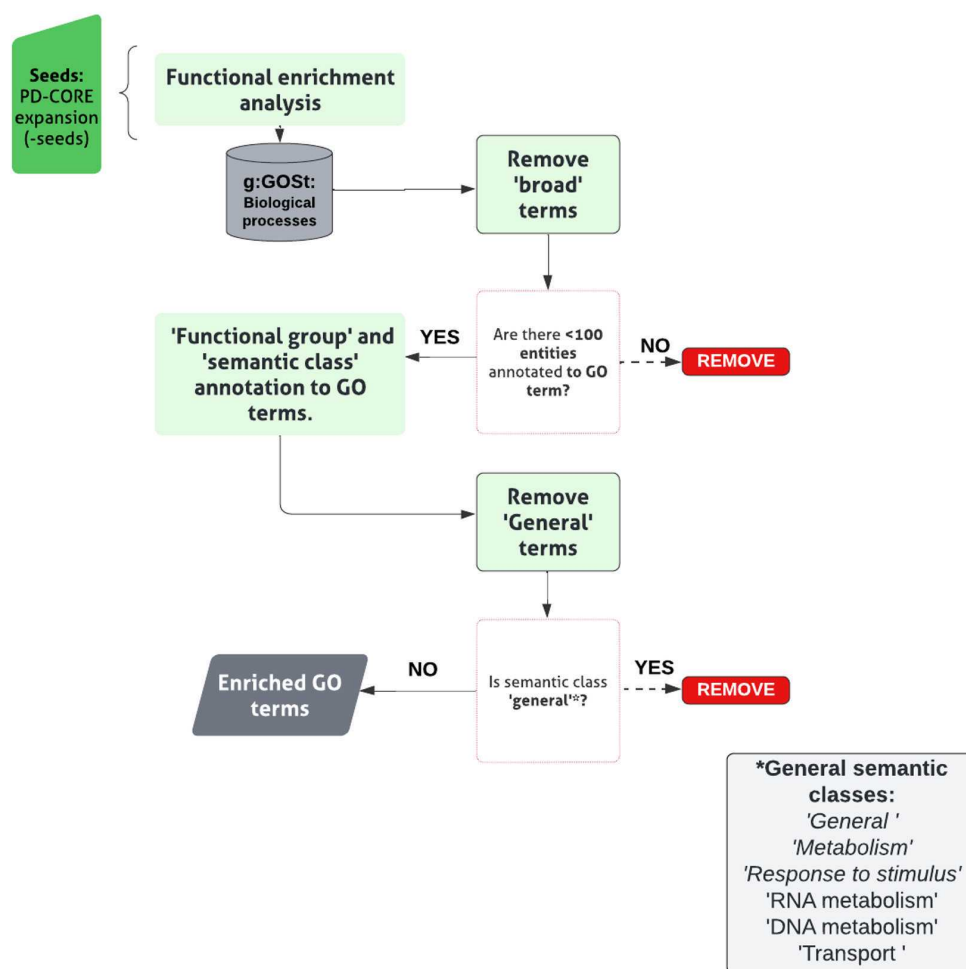**26**   The general pipeline for this analysis can be found in *Figure 4.*

Figure 4. Functional Enrichment general pipeline. The grey box indicates Semantic Classes (SCs) removed from the analysis, as they are classified as 'general'.

27 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).

28 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.

29 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.

30 Assign remaining terms to custom-made 'semantic classes'(SC), accompanied by a parent 'functional group'(FG) and discard generic terms.

> **Note**
>
> Assignment is manual.

31 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.

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