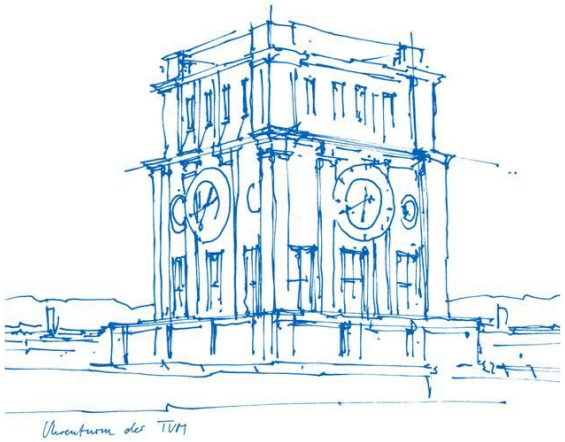
Data Science in Systems Biology
TUM School of Life Sciences
Technical University of Munich

T9 - Network-medicine-based drug repurposing: Nextflow pipeline for drug repurposing

Markus List

[BC]^2 2025

Basel, 08.09.2025



Structure

- ~ 30 min **Pipeline introduction**
  - Motivation
  - Introduction to Nextflow
  - Pipeline overview and some analysis insights
  - Hands-on session overview
- ~ 15 min **Coffee break**
- ~ 30 min **Hands-on session**
  - Run the pipeline on Huntington's Disease data
  - Output interpretation

Motivation

Disease modules

Disease proteins do not act in isolation

Example **Huntington's Disease**:

- Neurodegenerative disorder caused by mutation in the HTT gene

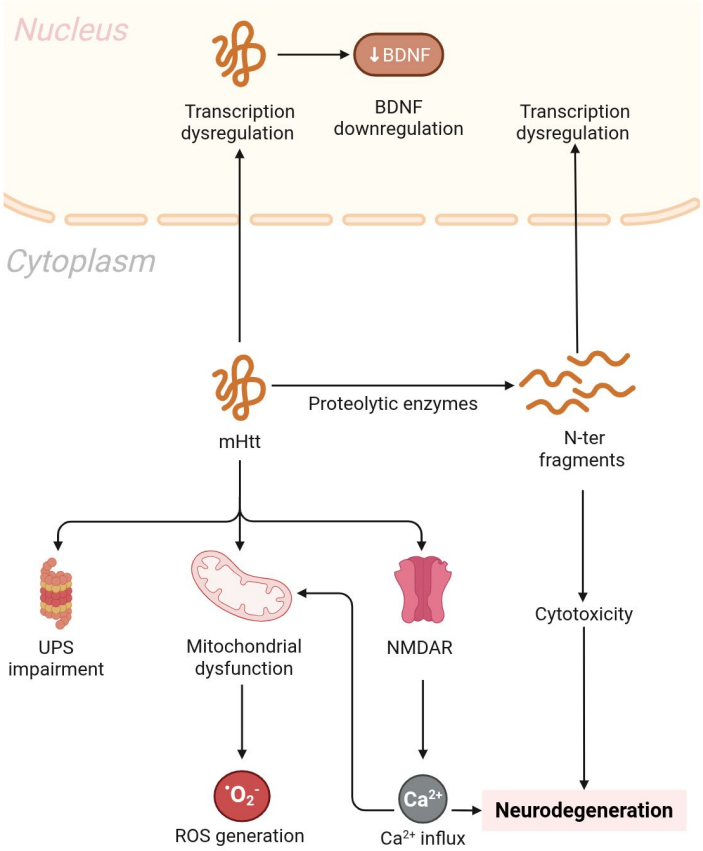
mHtt

Disease modules

Disease proteins do not act in isolation

Example **Huntington's Disease**:

- Neurodegenerative disorder caused by mutation in the HTT gene
- Mutant protein (mHtt) disrupts various cellular processes



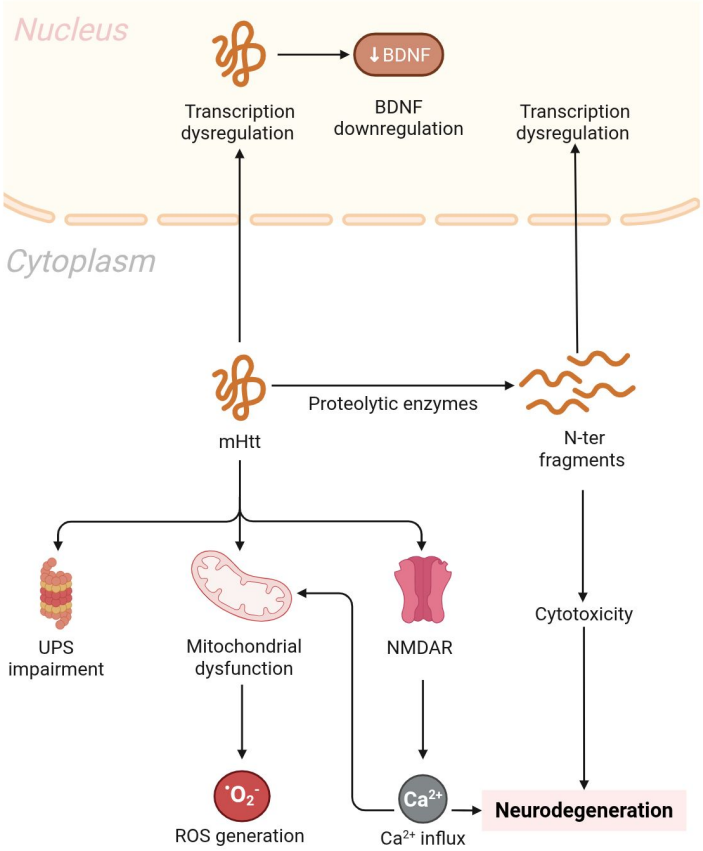
Modified with permission from *Maritan, M. (2025) BioRender.*

Disease modules

Disease proteins do not act in isolation

Example **Huntington's Disease**:

- Neurodegenerative disorder caused by mutation in the HTT gene
- Mutant protein (mHtt) disrupts various cellular processes
- Exact mechanism incompletely understood



Modified with permission from *Maritan, M. (2025) BioRender.*

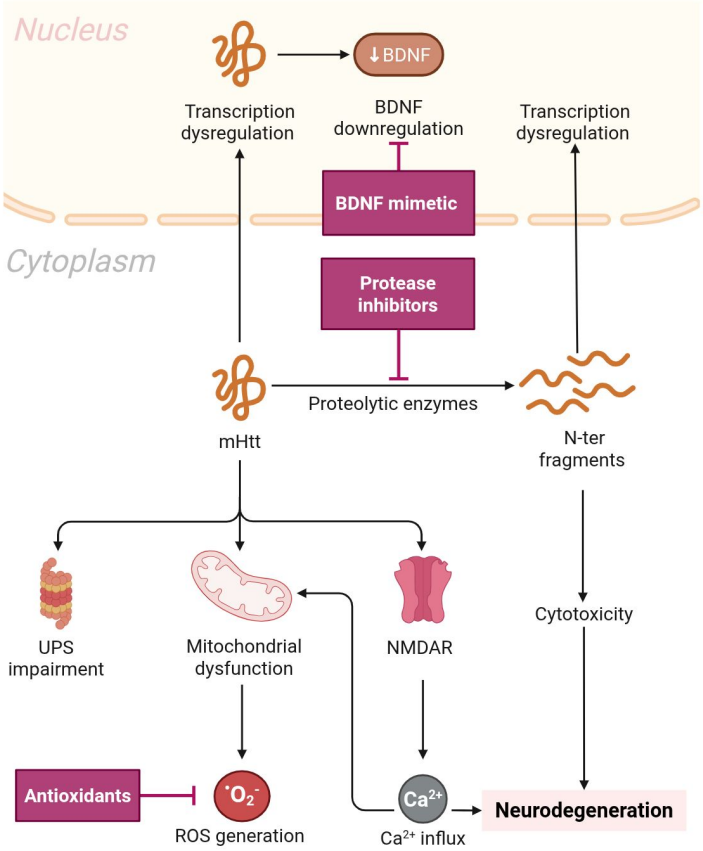
Disease modules

Disease proteins do not act in isolation

Example **Huntington's Disease**:

- Neurodegenerative disorder caused by mutation in the HTT gene
- Mutant protein (mHtt) disrupts various cellular processes
- Exact mechanism incompletely understood

➔ Understanding the mechanism is key to guiding treatment research



Modified with permission from *Maritan, M. (2025) BioRender.*

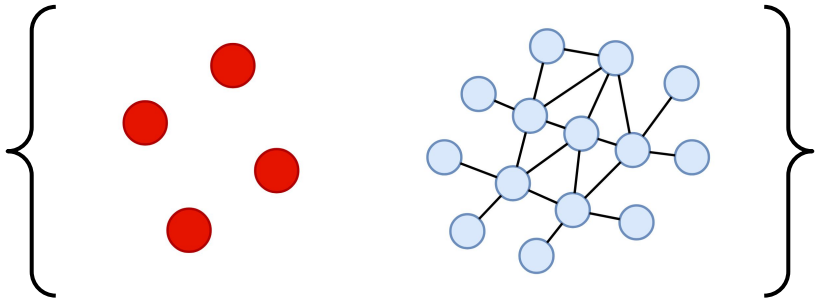
Disease module inference

Disease module inference

Seed nodes

Disease-associated genes/proteins from

- Databases
- Literature
- Experiments



Interaction network

Usually protein-protein interactions (PPIs)

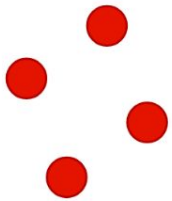
Disease module inference

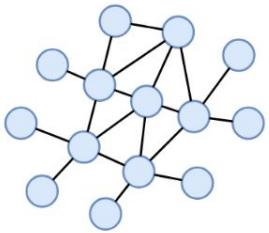
Seeds

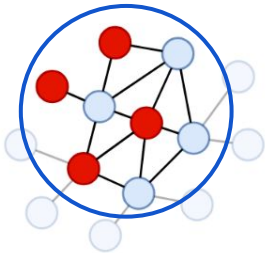
Interaction network

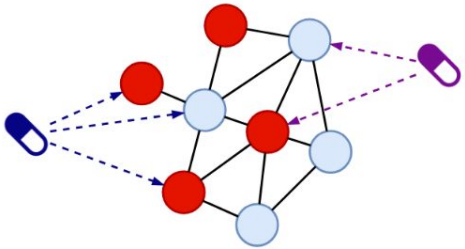
Disease module

Drug prioritization / repurposing



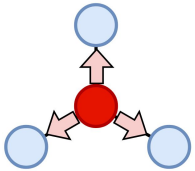




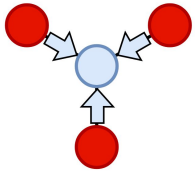


Algorithms for disease module inference

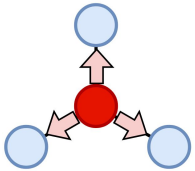
1st Neighbors:
Adds all nodes that directly interact with a seed node



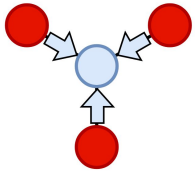
ROBUST:
Connects the seeds using diverse prize-collecting Steiner trees
Bernett et al. *Bioinformatics* (2022)



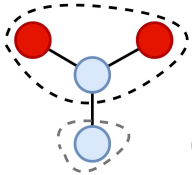
DIAMOnD:
Iteratively expands the seeds integrating the node with the most significant connectivity
Ghiassian et al. *PLoS Computational Biology* (2015)



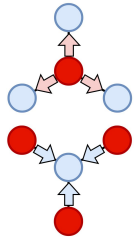
ROBUST (bias aware):
Works like ROBUST but penalizes the inclusion of overstudied genes/proteins
Sarkar et al. *Bioinformatics* (2023)



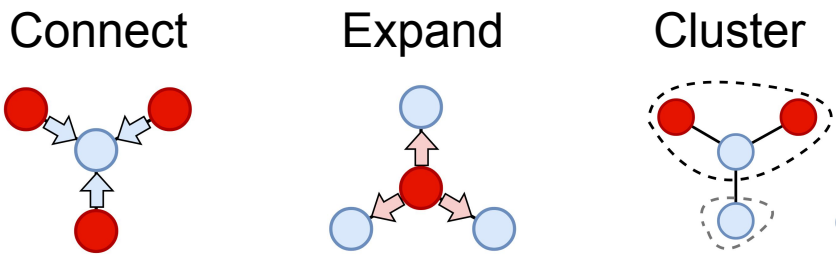
DOMINO:
Partitions the network into clusters and returns those which are enriched with seeds
Levi et al. *Molecular Systems Biology* (2021)



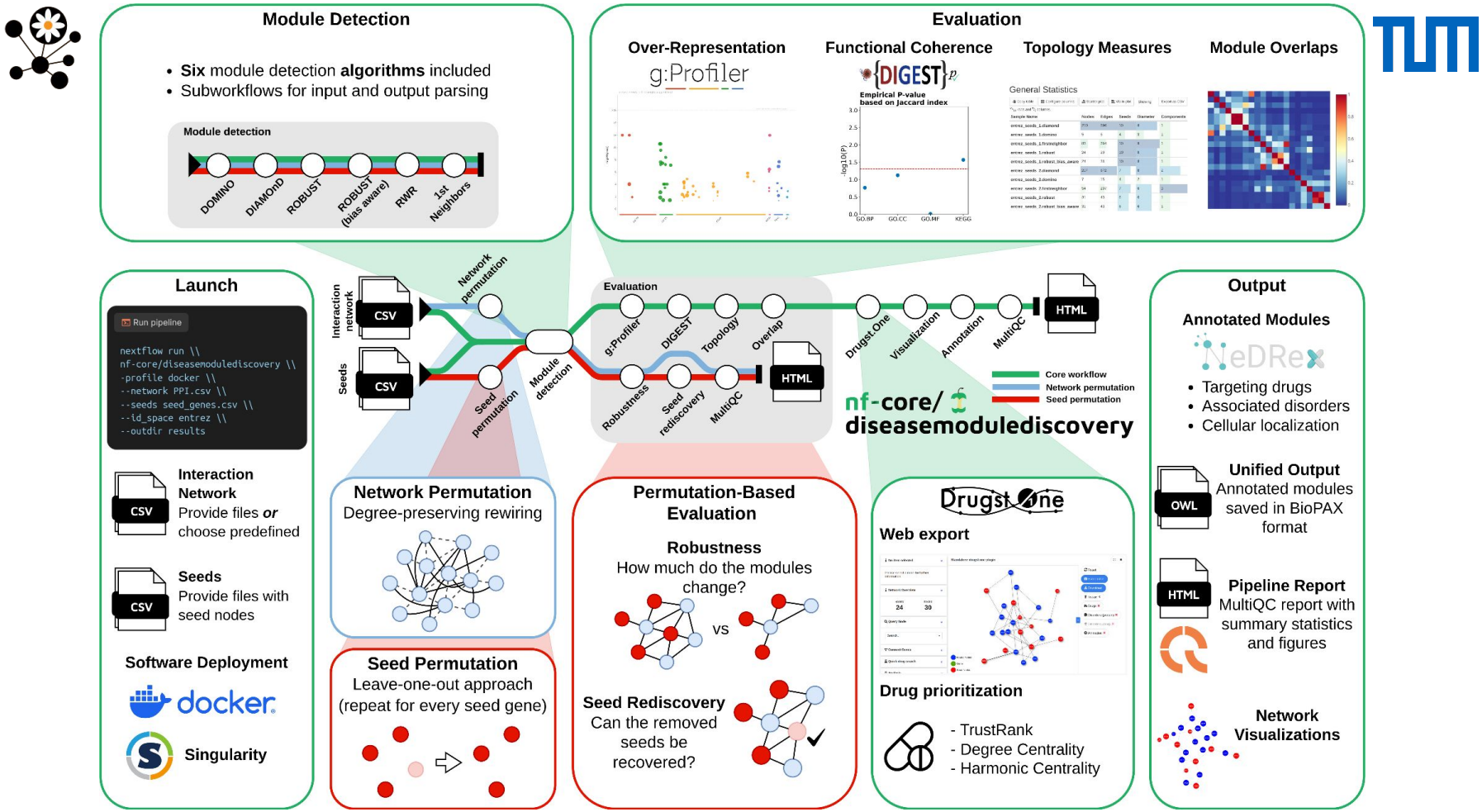
Random Walk with Restart (RWR): Expands the seeds using network propagation until all seeds are connected



Methods use different modeling approaches



Evaluating disease modules is challenging


Module Detection
• Six module detection algorithms included
• Subworkflows for input and output parsing
Module detection
DOMINO
DIAMOnD
ROBUST
ROBUST (bias aware)
RWR
1st Neighbors
Evaluation
Over-Representation
g:Profiler
Functional Coherence
DIGEST
Empirical P-value based on Jaccard index
Topology Measures
General Statistics
Module Overlaps
Launch
Run pipeline
nextflow run \\
nf-core/diseasemoduleddiscovery \\
-profile docker \\
--network PPI.csv \\
--seeds seed_genes.csv \\
--id_space entrez \\
--outdir results
Interaction Network
Provide files or choose predefined
Seeds
Provide files with seed nodes
Software Deployment
docker
Singularity
Interaction network
Seeds
CSV
Network permutation
Seed permutation
Module detection
Evaluation
g:Profiler
DIGEST
Topology
Overlap
Robustness
Seed rediscovery
MultiQC
HTML
Drugst.One
Visualization
Annotation
MultiQC
HTML
Core workflow
Network permutation
Seed permutation
nf-core/diseasemoduleddiscovery
Network Permutation
Degree-preserving rewiring
Seed Permutation
Leave-one-out approach
(repeat for every seed gene)
Permutation-Based Evaluation
Robustness
How much do the modules change?
vs
Seed Rediscovery
Can the removed seeds be recovered?
Drugst.One
Web export
Drug prioritization
- TrustRank
- Degree Centrality
- Harmonic Centrality
Output
Annotated Modules
NeDRex
• Targeting drugs
• Associated disorders
• Cellular localization
OWL
Unified Output
Annotated modules saved in BioPAX format
HTML
Pipeline Report
MultiQC report with summary statistics and figures
Network Visualizations

Nextflow

nextflow

Nextflow is a workflow management language used to automate scientific analysis pipelines

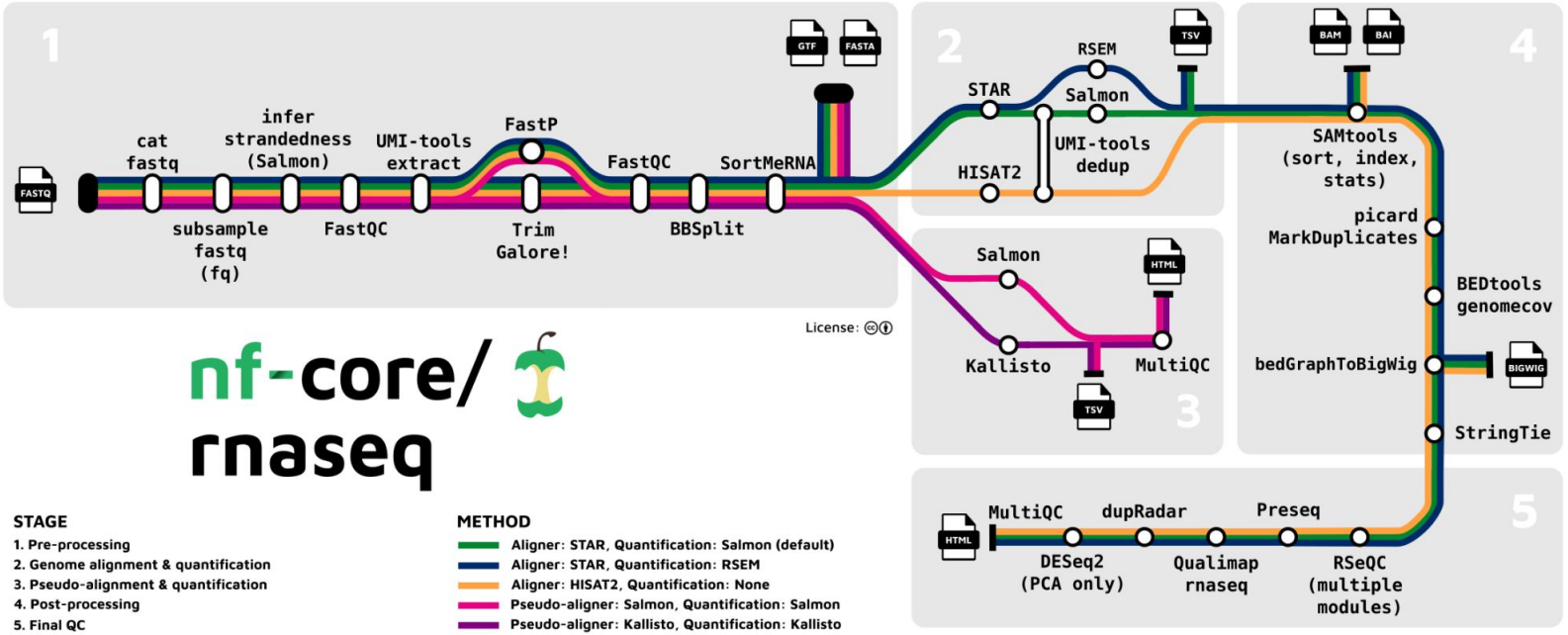


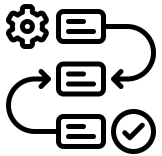
Image source: https://raw.githubusercontent.com/nf-core/rnaseq/3.14.0/docs/images/nf-core-rnaseq_metro_map_grey.png

Di Tommaso, Paolo, et al. *Nature biotechnology* 35.4 (2017): 316-319.

nextflow

***Nextflow** is a workflow management language used to automate scientific analysis pipelines*

Orchestration



- Passes input and output files between different processes
- Runs processes in parallel whenever possible
- Tracks intermediate results → allows resuming a cached workflow in case of failure

Reproducibility

- Nextflow automatically deploys all required software
- Various container solutions supported

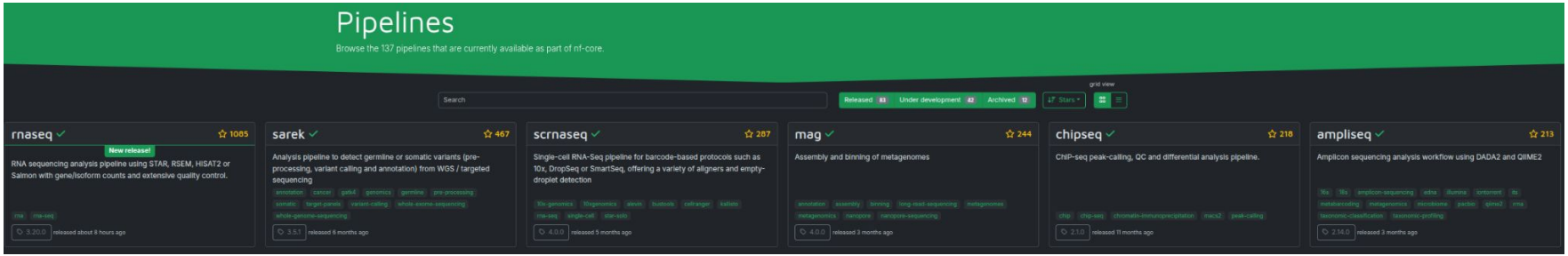
Portability

- Supports execution on local machines, cloud platforms, and various workload schedulers
- Examples:

Di Tommaso, Paolo, et al. *Nature biotechnology* 35.4 (2017): 316-319.

***nf-core** provides a collection of community-curated, best-practice Nextflow pipelines*

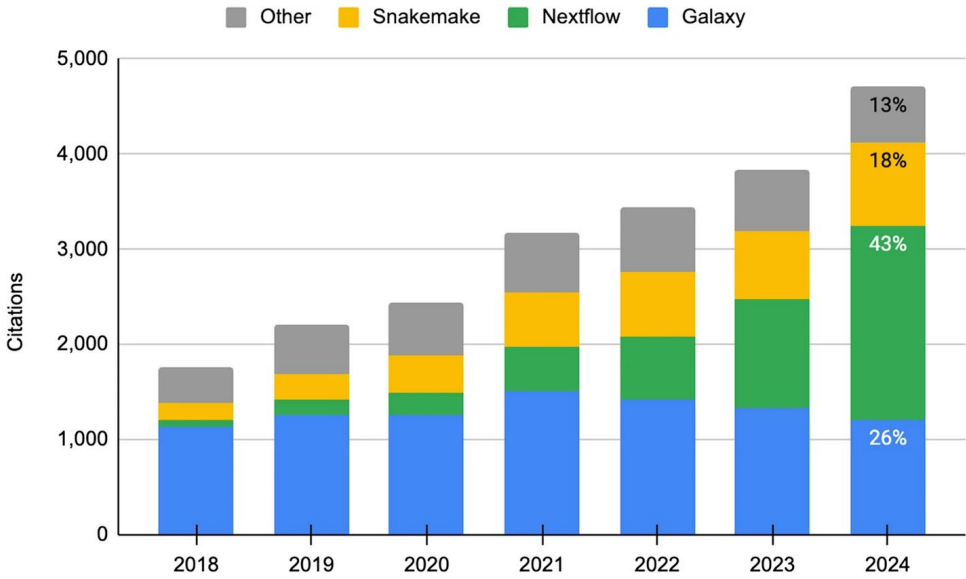
- Pipelines for most standard workflows (RNA-seq, variant calling, ChIP-seq, …)
- Cooperation instead of duplication
  - Avoids redundant pipelines
  - Extend and maintaining one pipeline per analysis type
- Input, output, and usage documentation
- Automated continuous integration tests for all pipelines



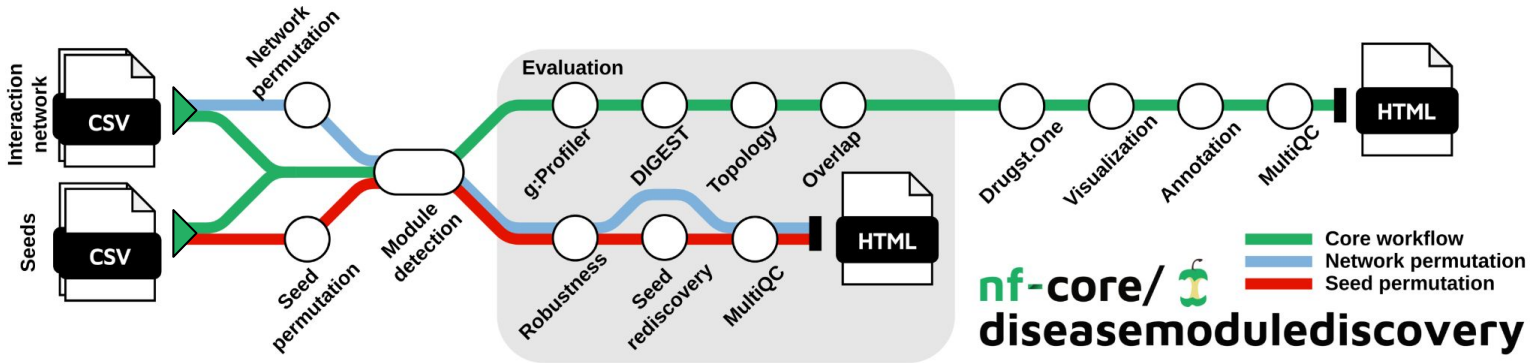
Ewels, P. A., Peltzer, A., et al. (2020). *Nature Biotechnology*, *38*(3), 276–278.

Nextflow and nf-core

- **Nextflow**: Currently most used bioinformatics pipeline framework
- Popularity in large parts thanks to nf-core
- **nf-core**: Community developing high-quality pipelines for different use cases, mostly in bioinformatics



Pipeline overview




```
Run pipeline

nextflow run \
nf-core\diseasemodulediscovery \
-profile docker \
--network PPI.csv \
--seeds seed_genes.csv \
--id_space entrez \
--outdir results
```

Software Deployment

docker

Singularity

CSV

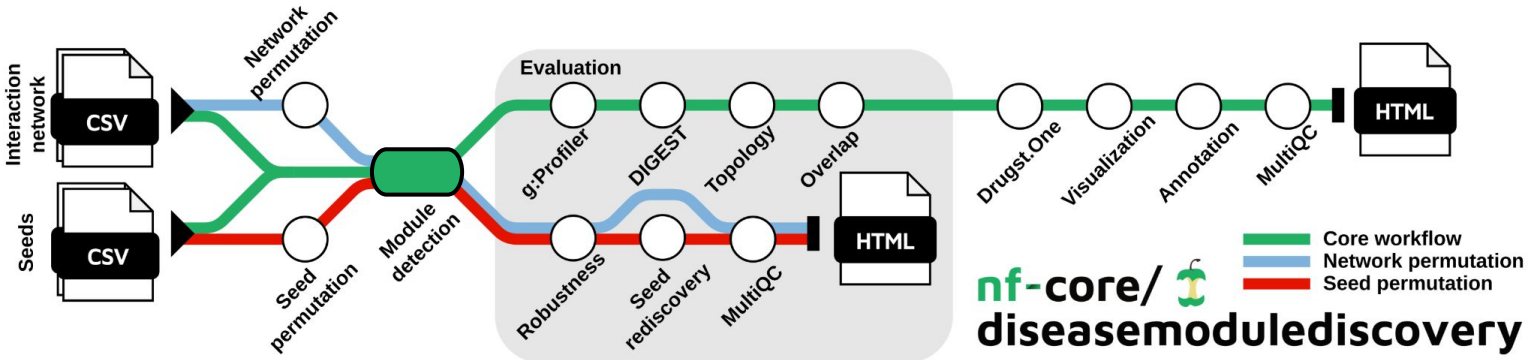
Interaction Network

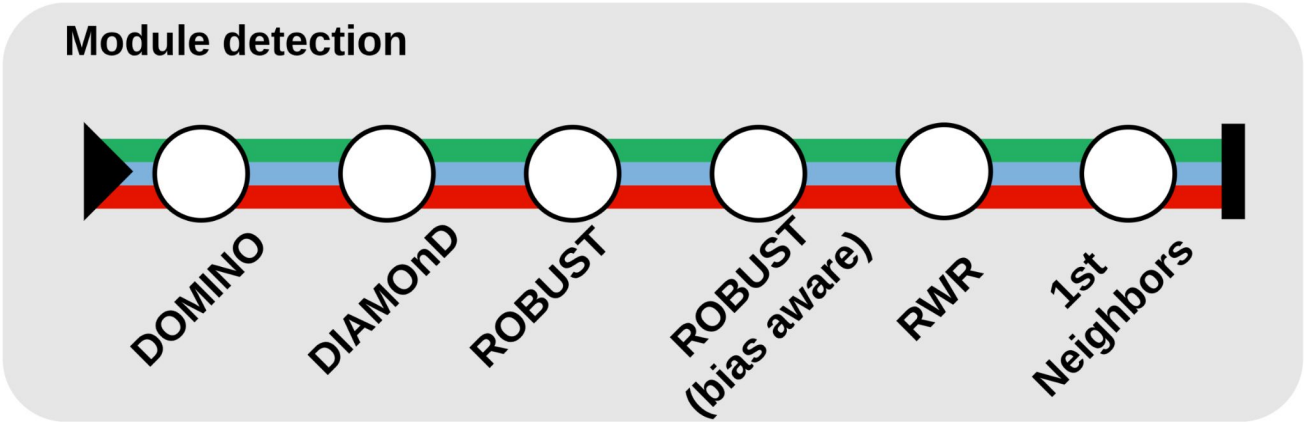
Provide files ***or*** choose predefined

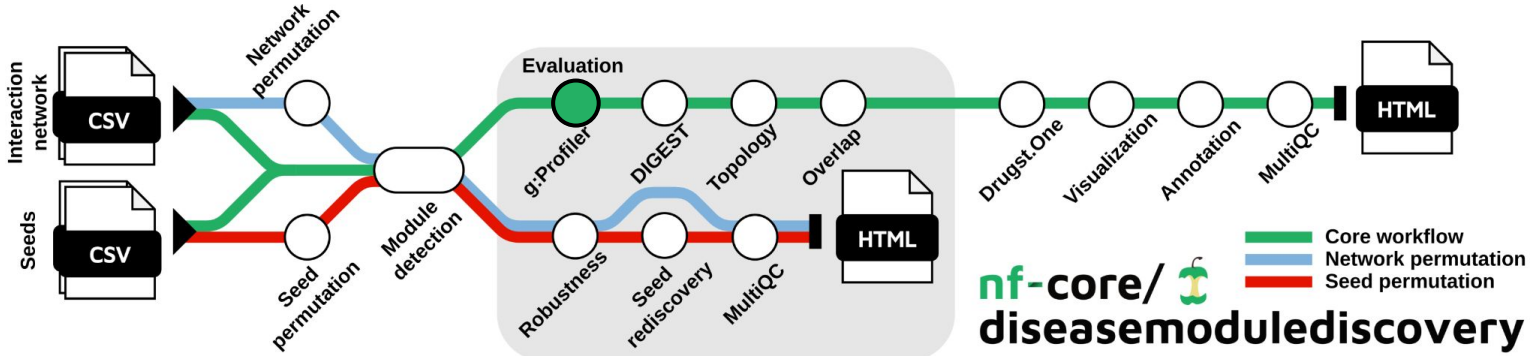
CSV

Seeds

Provide files with seed nodes

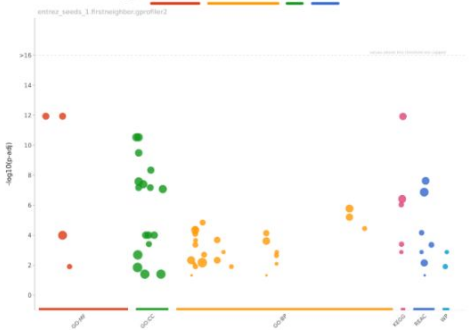

Interaction network
CSV
Seeds
CSV
Network permutation
Seed permutation
Module detection
Evaluation
g:Profiler
DIGEST
Topology
Overlap
Robustness
Seed rediscovery
MultiQC
HTML
Drugst.One
Visualization
Annotation
MultiQC
HTML
Core workflow
Network permutation
Seed permutation
nf-core/
diseasemodulediscovery


Module detection
DOMINO
DIAMOnD
ROBUST
ROBUST (bias aware)
RWR
1st Neighbors

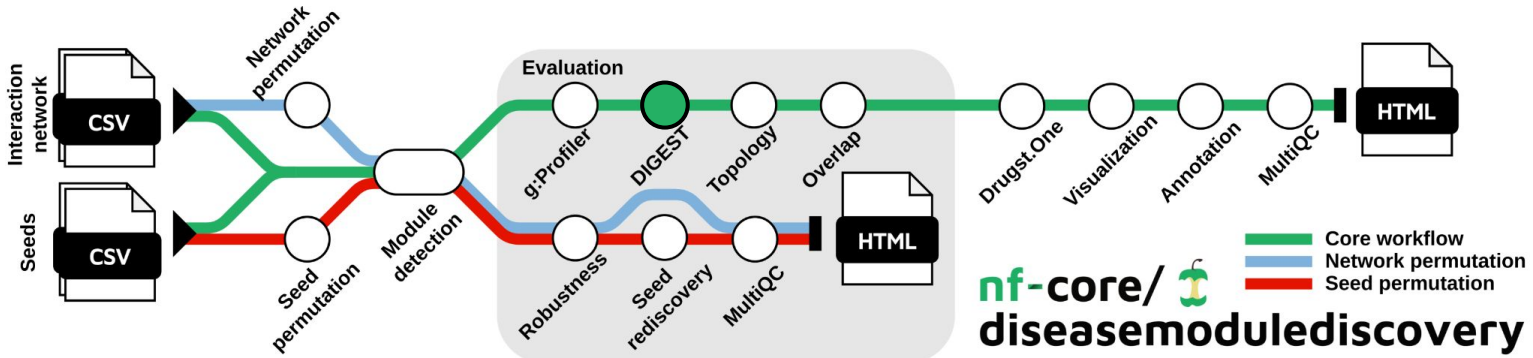


Over-Representation

g:Profiler

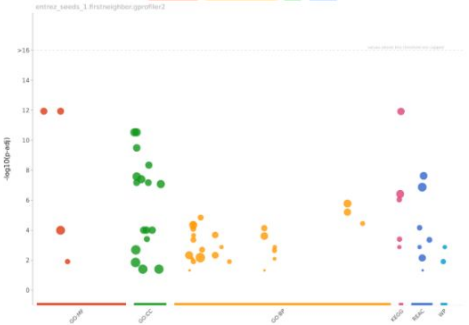


Kolberg et al. *F1000Research* (2020)



Over-Representation

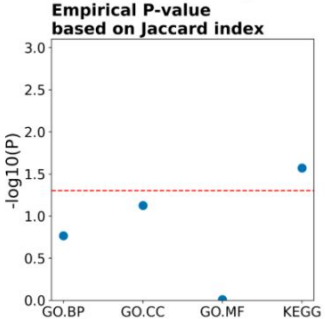
g:Profiler



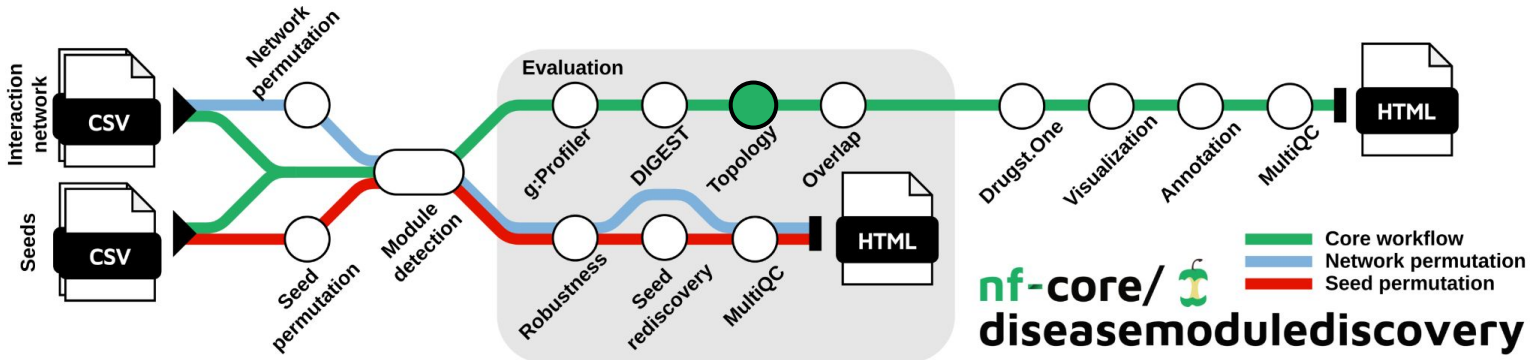
Kolberg et al. *F1000Research* (2020)

Functional Coherence

{DIGEST}p

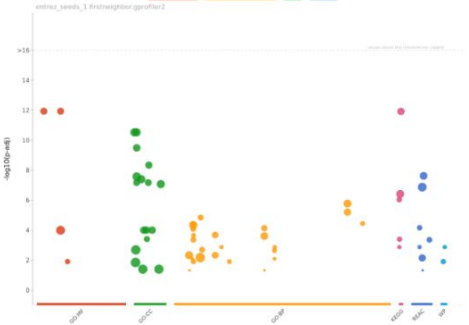


Adamowicz et al.
Briefings in Bioinformatics (2022)



Over-Representation

g:Profiler

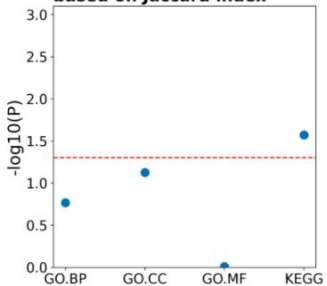


Kolberg et al. *F1000Research* (2020)

Functional Coherence

{DIGEST}p

Empirical P-value based on Jaccard index



Adamowicz et al.
Briefings in Bioinformatics (2022)

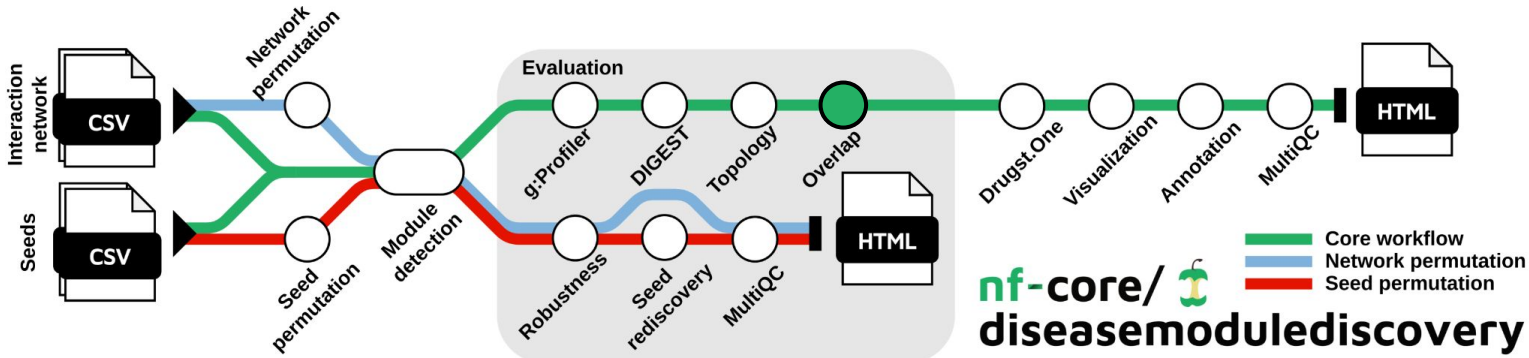
Topology Measures

General Statistics

Copy table | Configure columns | Scatter plot | Violin plot | Showing | Export as CSV

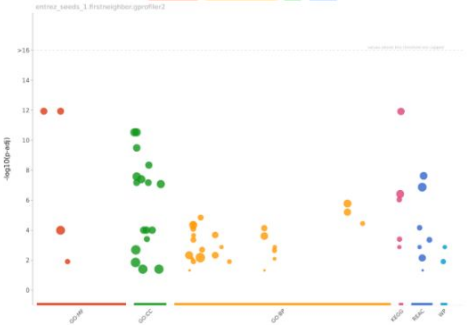
10/10 rows and 6/6 columns.

Sample Name	Nodes	Edges	Seeds	Diameter	Components
entrez_seeds_1.diamond	210	696	10	8	1
entrez_seeds_1.domino	5	6	4	3	1
entrez_seeds_1.firstneighbor	83	264	10	9	1
entrez_seeds_1.robust	24	29	10	8	1
entrez_seeds_1.robust_bias_aware	24	31	10	8	1
entrez_seeds_2.diamond	207	672	7	8	2
entrez_seeds_2.domino	7	15	4	2	1
entrez_seeds_2.firstneighbor	54	207	7	6	3
entrez_seeds_2.robust	31	43	6	6	1
entrez_seeds_2.robust_bias_aware	31	43	6	6	1



Over-Representation

g:Profiler

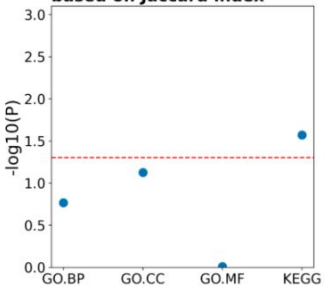


Kolberg et al. *F1000Research* (2020)

Functional Coherence

DIGEST

Empirical P-value based on Jaccard index



Adamowicz et al.
Briefings in Bioinformatics (2022)

Topology Measures

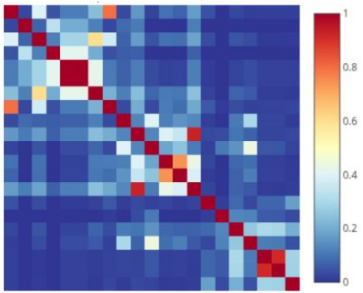
General Statistics

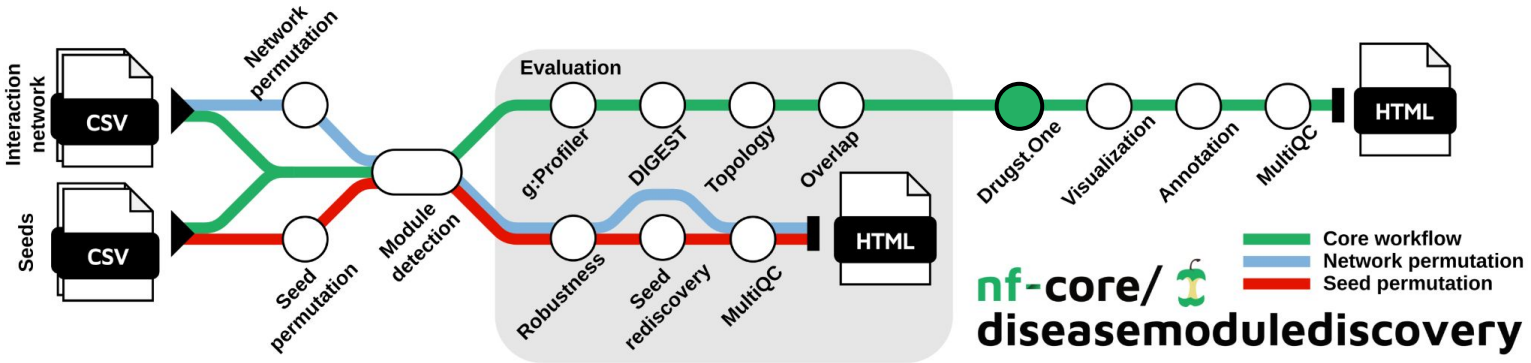
Copy table | Configure columns | Scatter plot | Violin plot | Showing | Export as CSV

10/10 rows and 6/6 columns.

Sample Name	Nodes	Edges	Seeds	Diameter	Components
entrez_seeds_1.diamond	210	696	10	8	1
entrez_seeds_1.domino	5	6	4	3	1
entrez_seeds_1.firstneighbor	83	264	10	9	1
entrez_seeds_1.robust	24	29	10	8	1
entrez_seeds_1.robust_bias_aware	24	31	10	8	1
entrez_seeds_2.diamond	207	672	7	8	2
entrez_seeds_2.domino	7	15	4	2	1
entrez_seeds_2.firstneighbor	54	207	7	6	3
entrez_seeds_2.robust	31	43	6	6	1
entrez_seeds_2.robust_bias_aware	31	43	6	6	1

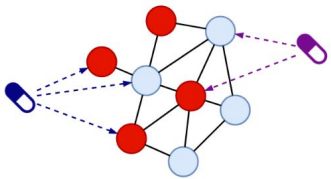
Module Overlaps





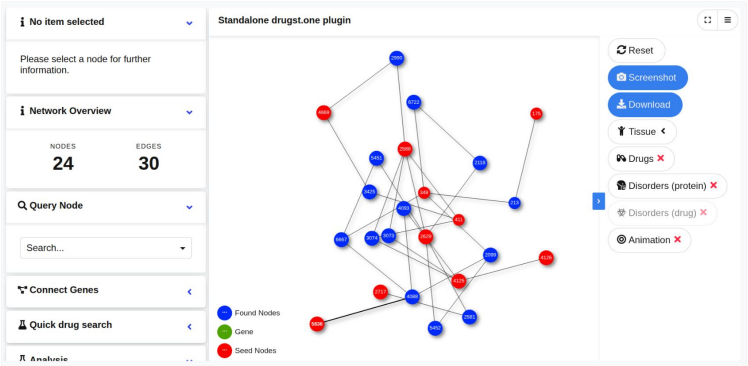
Drugst.One

Drug Prioritization

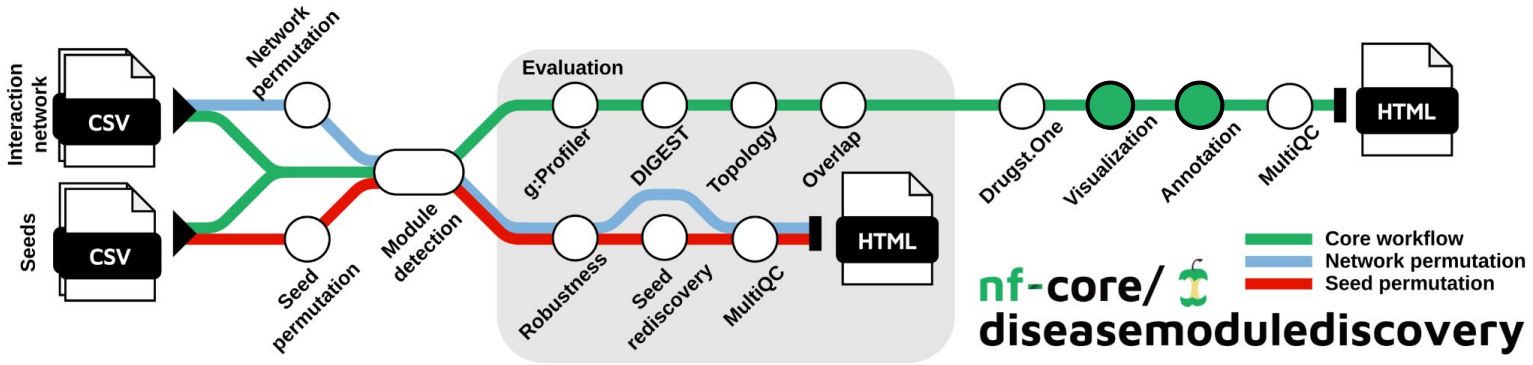


- TrustRank
- Degree Centrality
- Harmonic Centrality

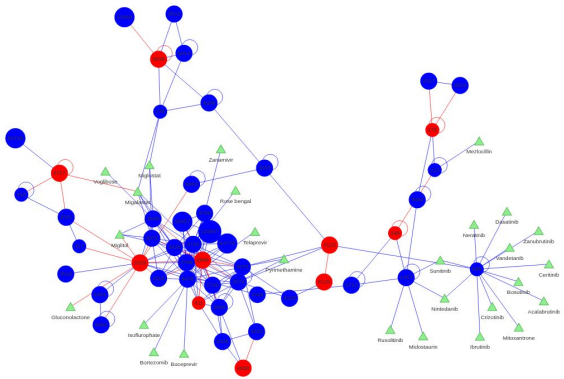
Web Export



Maier et al. *Nucleic Acids Research* (2024)



Network Visualization

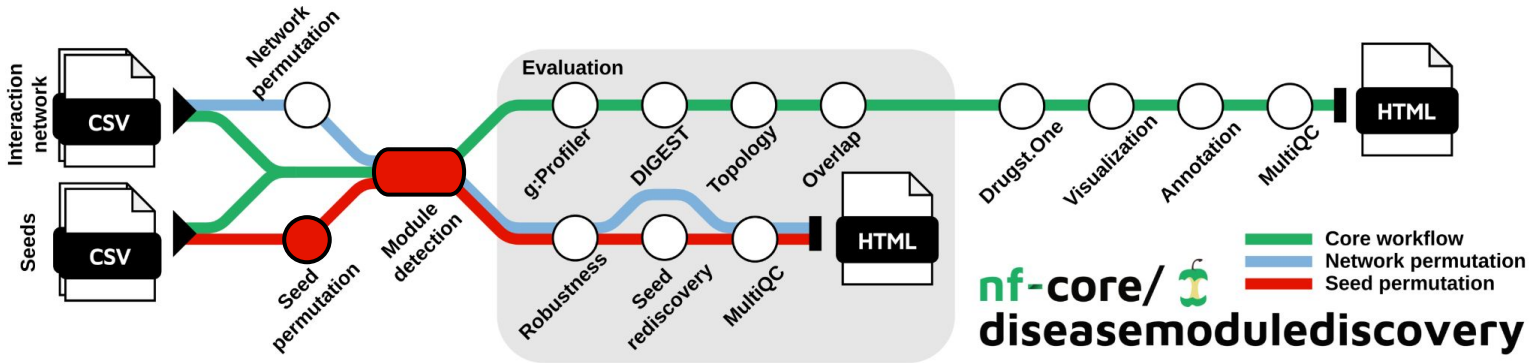


Module Annotation

NeDRex

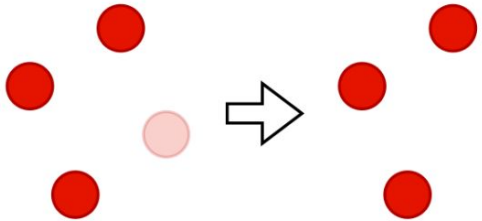
- Targeting drugs
- Side effects
- Disorders
- Cellular component

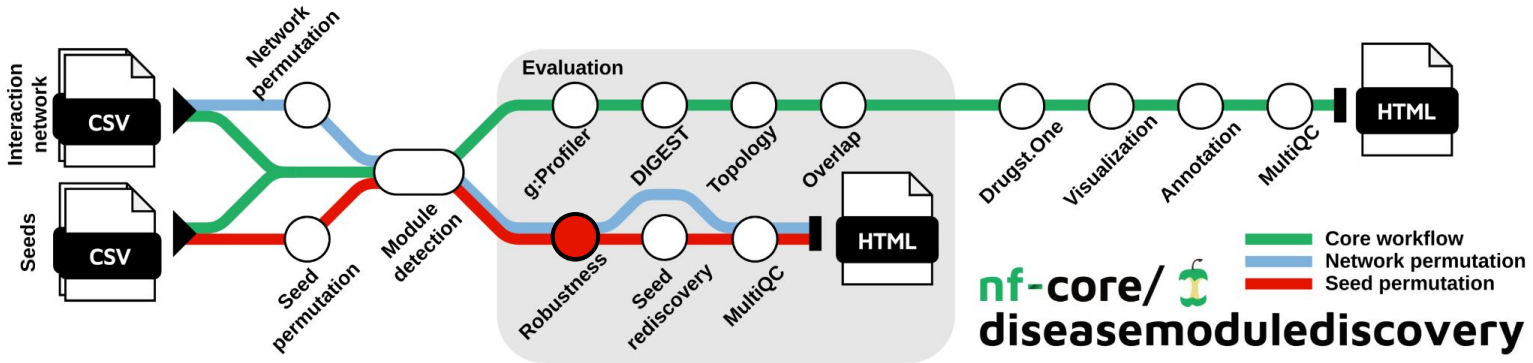
Sadegh et al. *Nature Communications* (2021)



Seed Permutation

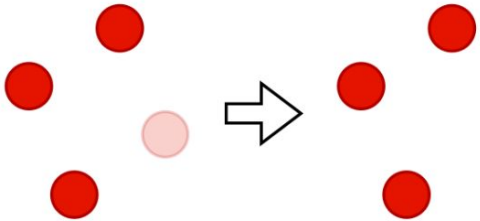
Leave-one-out approach
(repeat for every seed gene)





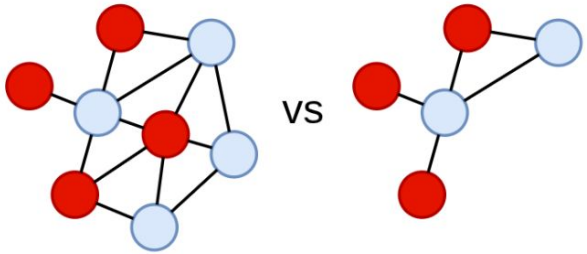
Seed Permutation

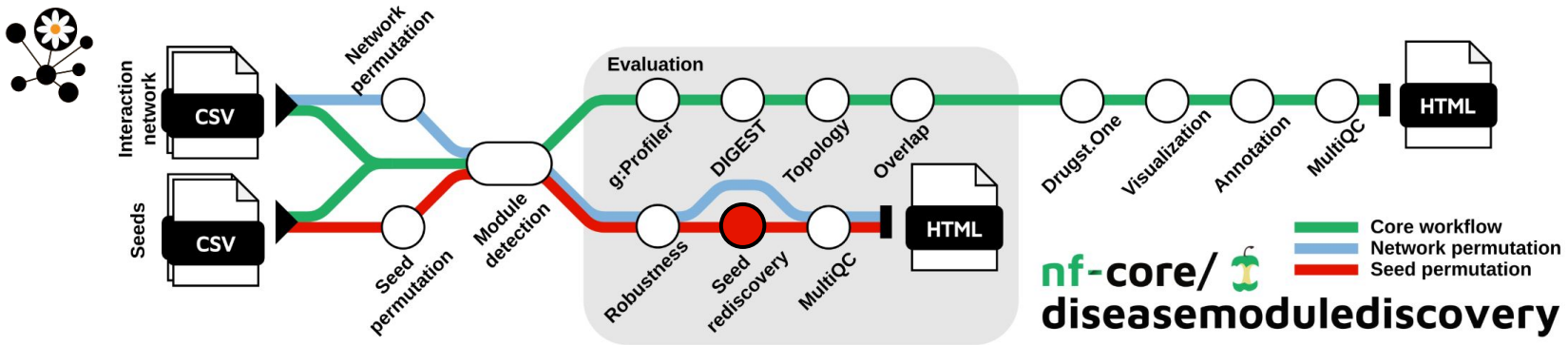
Leave-one-out approach
(repeat for every seed gene)



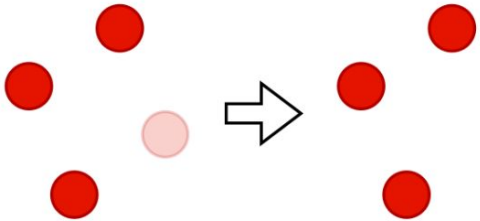
Robustness

How much do the modules change?

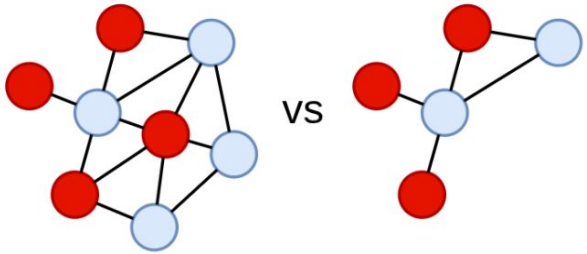




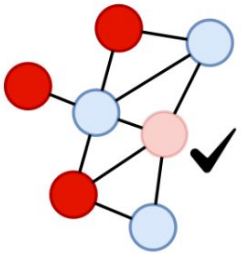
Seed Permutation
Leave-one-out approach
(repeat for every seed gene)

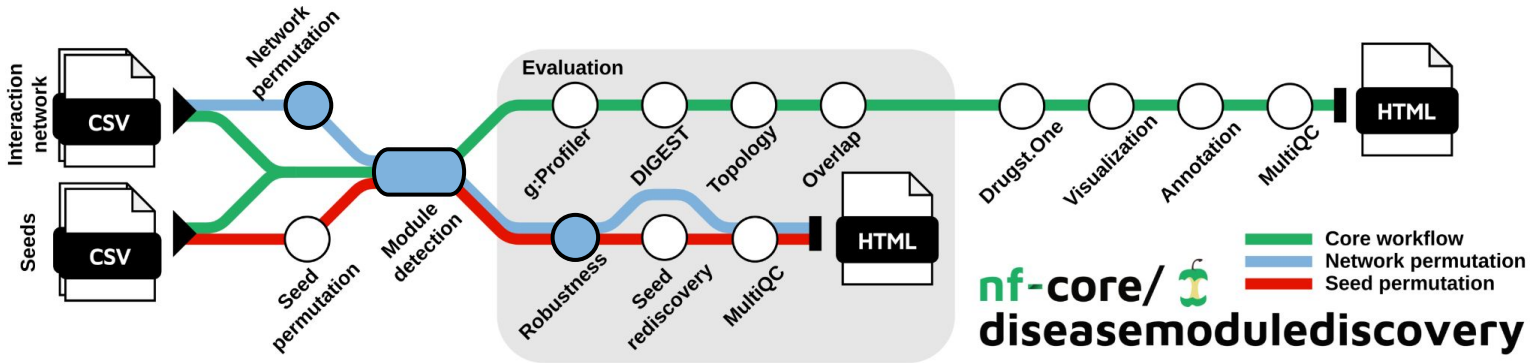


Robustness
How much do the modules change?



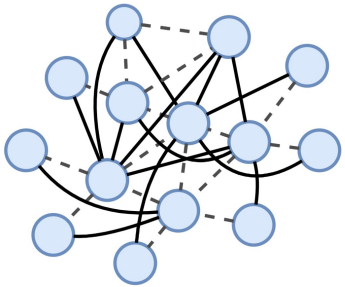
Seed Rediscovery
Can the removed seed genes be recovered?





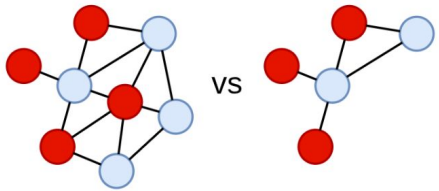
Network Permutation

Degree-preserving rewiring

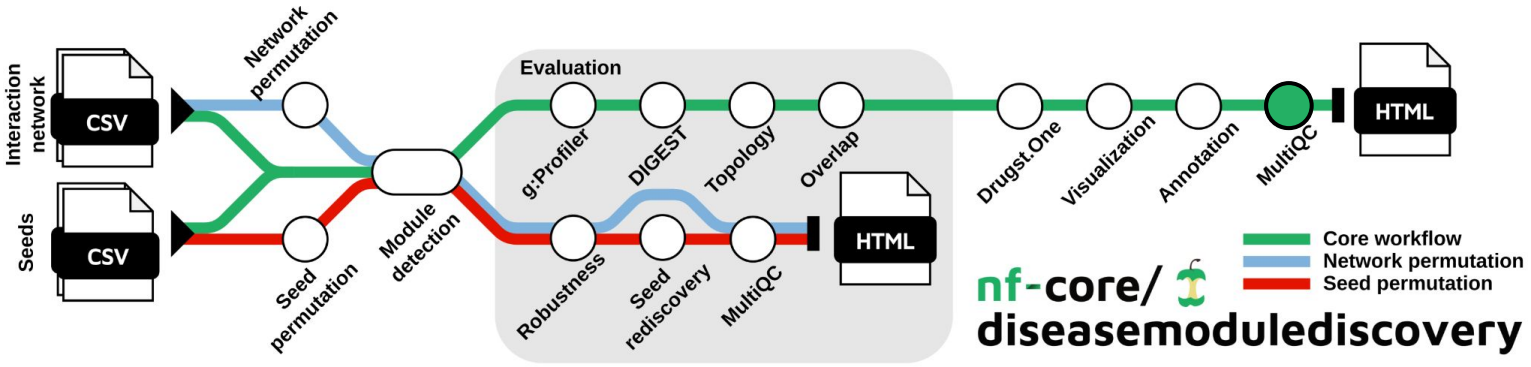


Module detection methods are prone to learning from node degrees, instead of individual interactions.

➔ If module remains largely **the same,** it is **based on node degrees**.



Lazareva et al. *Briefings in Bioinformatics* (2021)



multiqc

- HTML pipeline report
- Summarizes modules and evaluation results
- Execution details and software versions

Ewels et al. *Bioinformatics* (2016)

Available PPI networks

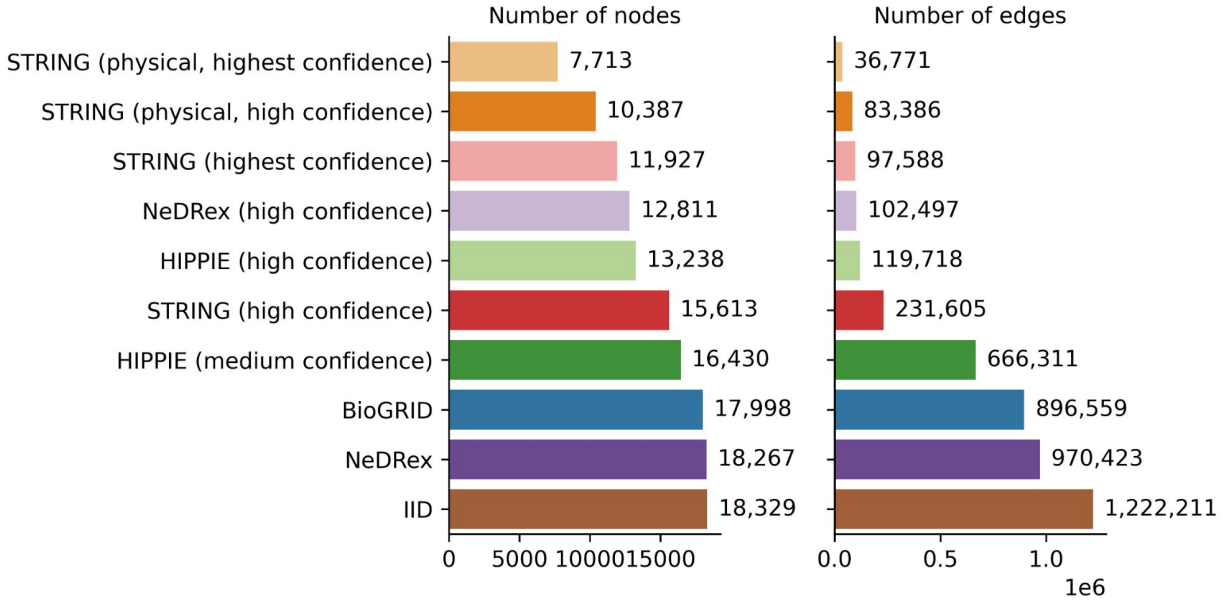
Pipeline provides selection of PPI networks in different ID spaces

PPI networks:

- BioGRID
- IID
- HIPPIE (2 variants)
- NeDRex (2 variants)
- STRING (4 variants)

ID spaces:

- Uniprot proteins
- Ensembl genes
- Entrez genes
- Gene symbols



Output

Most **output files are named** according to the seed file, network file, and method used, for example:
`<seed_file_name>.<network_file_name>.<method_name>.tsv`

A typical run directory has the following **structure**:

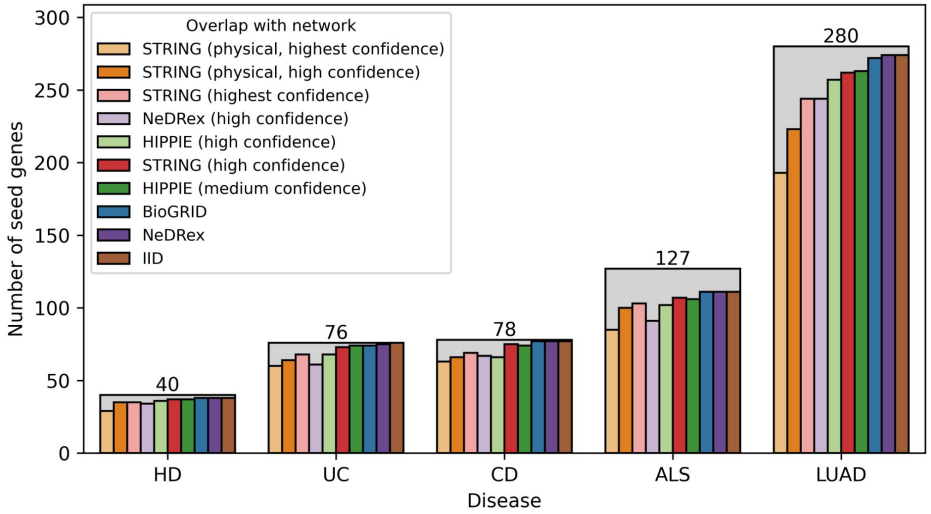
```
├── {outdir}
│   ├── input                                  Pre-processed input files
│   ├── modules                                Disease modules in different formats
│   ├── modules_visualized                     Network visualizations of the modules
│   ├── evaluation                             Module evaluation results
│   │   ├── digest                                 Functional coherence analysis
│   │   ├── gprofiler                          Over-representation analysis
│   │   ├── network_permutation                    Network permutation analysis
│   │   └── seed_permutation                       Seed set permutation analysis
│   ├── drug_prioritization                    Drug rankings
│   ├── modules_visualized_with_drugs          Network visualizations including drugs
│   ├── multiqc                                Pipeline summary report
│   ├── pipeline_info                      Run statistics (memory usage, etc…)
│   └── reports                                Additional reports
└── work                                       Cached intermediate files
```

Some results

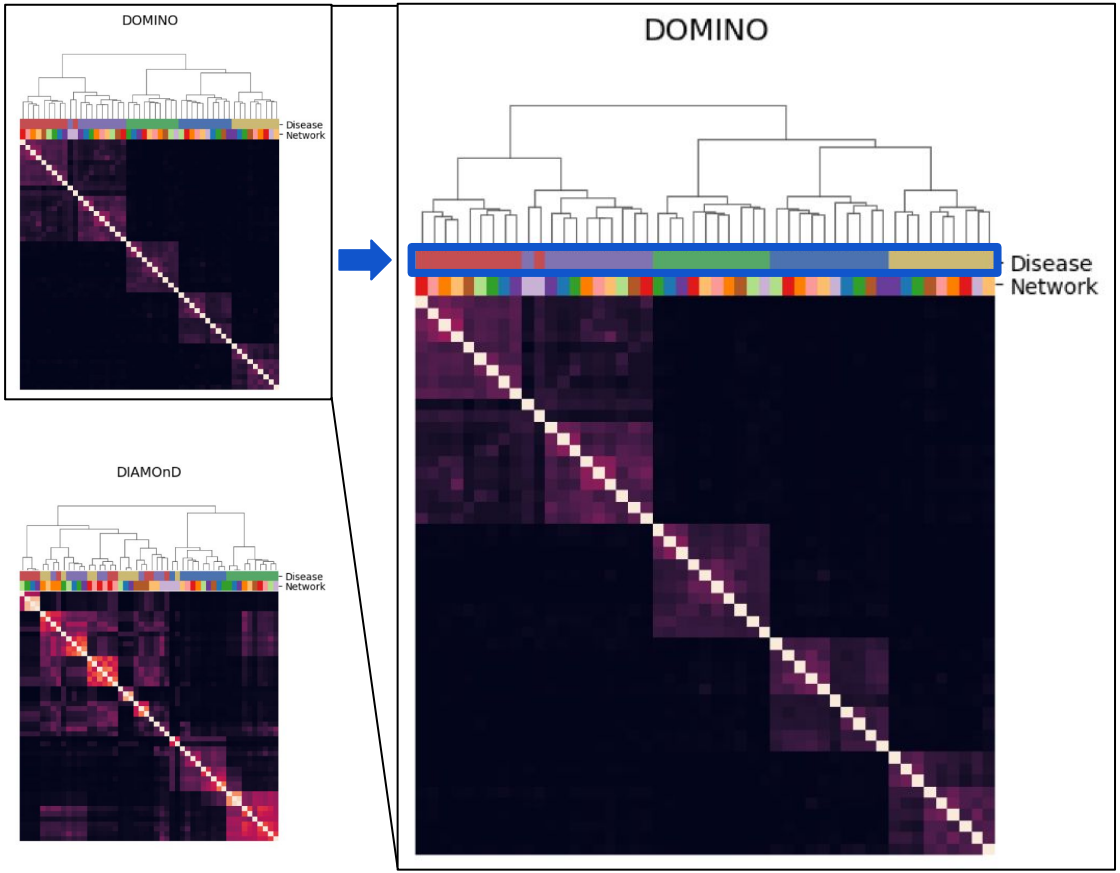
Pipeline demonstration

***10** networks x **5** diseases = **50** input combinations*

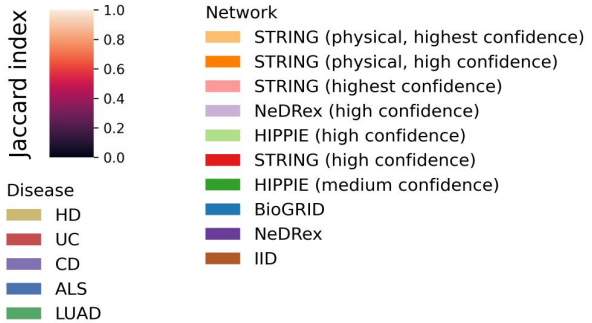
- Huntington's disease (**HD**)
- Ulcerative colitis (**UC**)
- Crohn's disease (**CD**)
- Amyotrophic lateral sclerosis (**ALS**)
- Lung adenocarcinoma (**LUAD**)



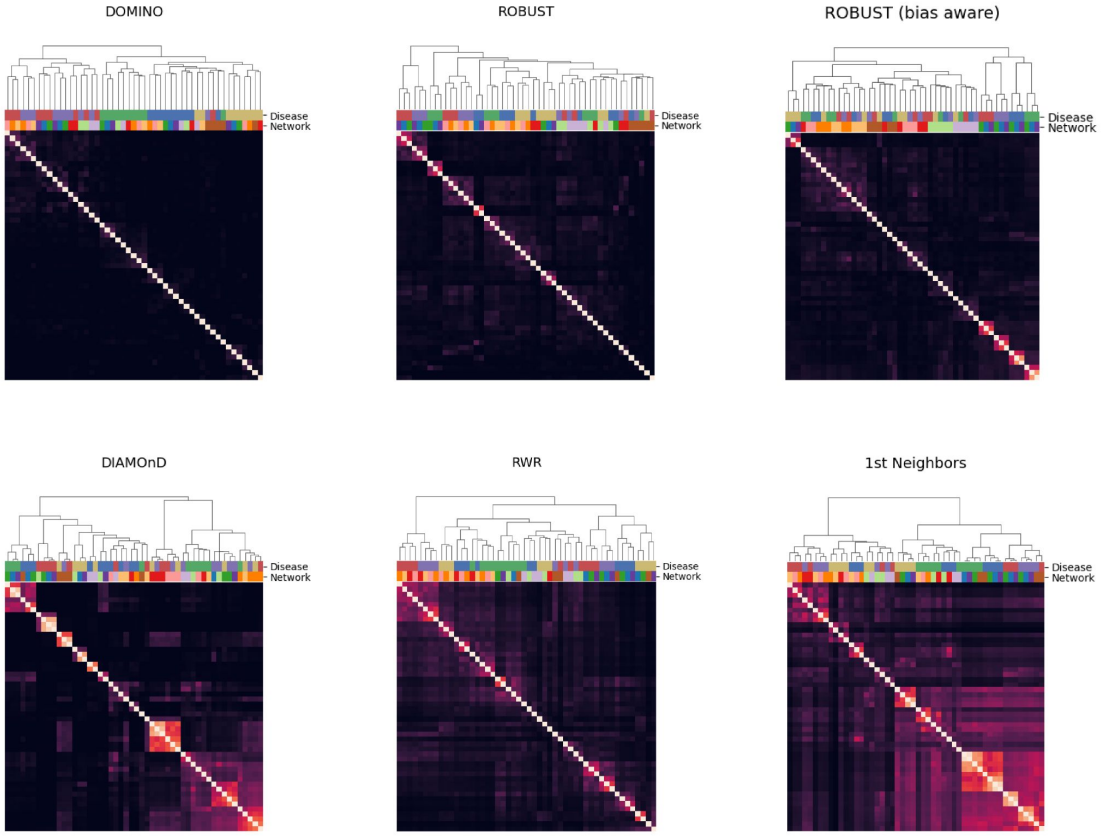
Module overlaps



DOMINO, ROBUST**, and **ROBUST (bias aware)
yield disease-specific modules



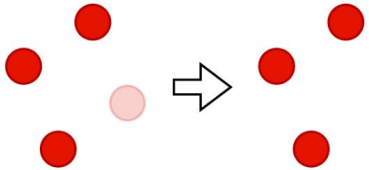
Module overlaps – excluding the seeds…



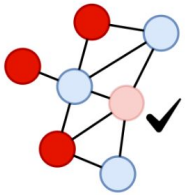
However, *the disease specificity primarily arises from the seed nodes…*

Seed rediscovery

Seed Permutation
Leave-one-out approach
(repeat for every seed gene)

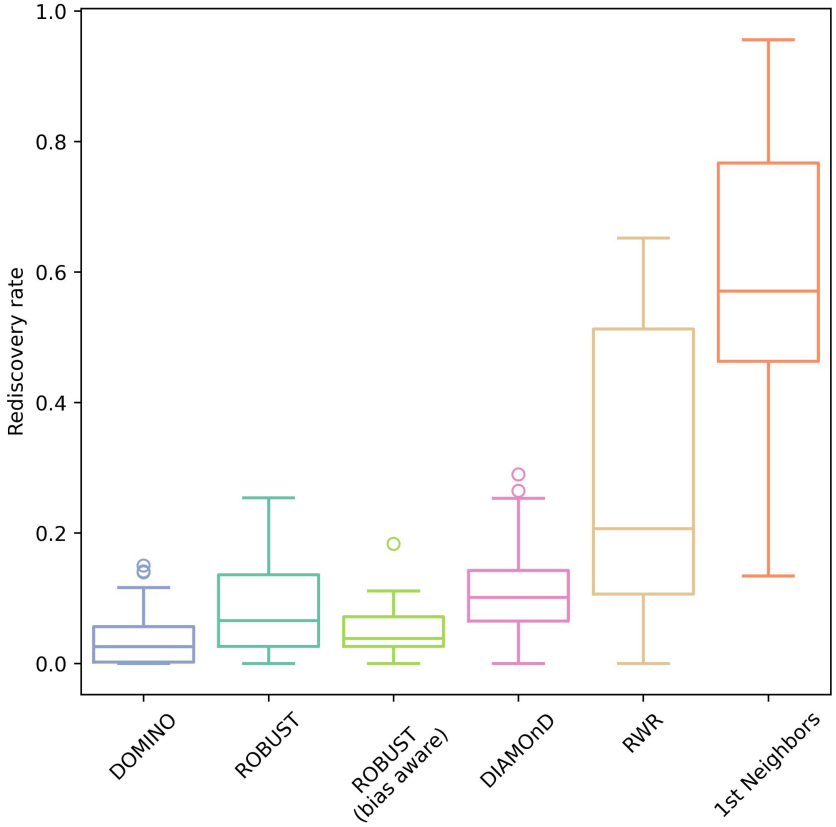


Seed Rediscovery
Can the removed seed genes
be recovered?



Including all direct interactors is the most reliable way to recover individual seeds

***However**, there can be a lot of direct interactors…*

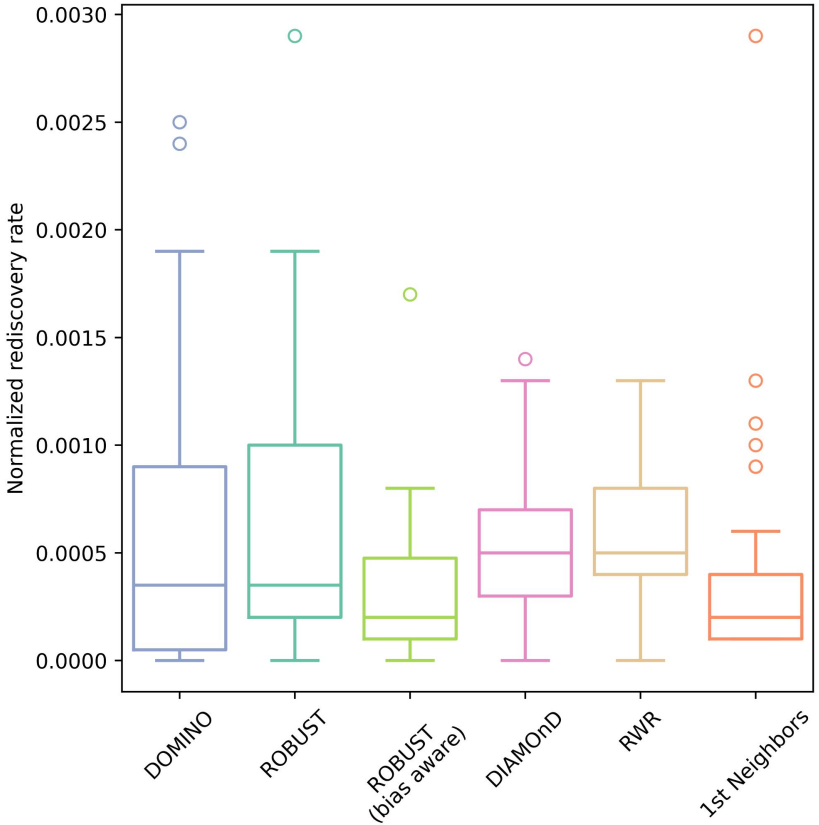


Seed rediscovery – normalized…

Normalized rediscovery rate *is divided by the number of nodes in the module*

Differences *between tools are less prominent*

DIAMOnD *and* ***RWR*** *yield highest median recovery*

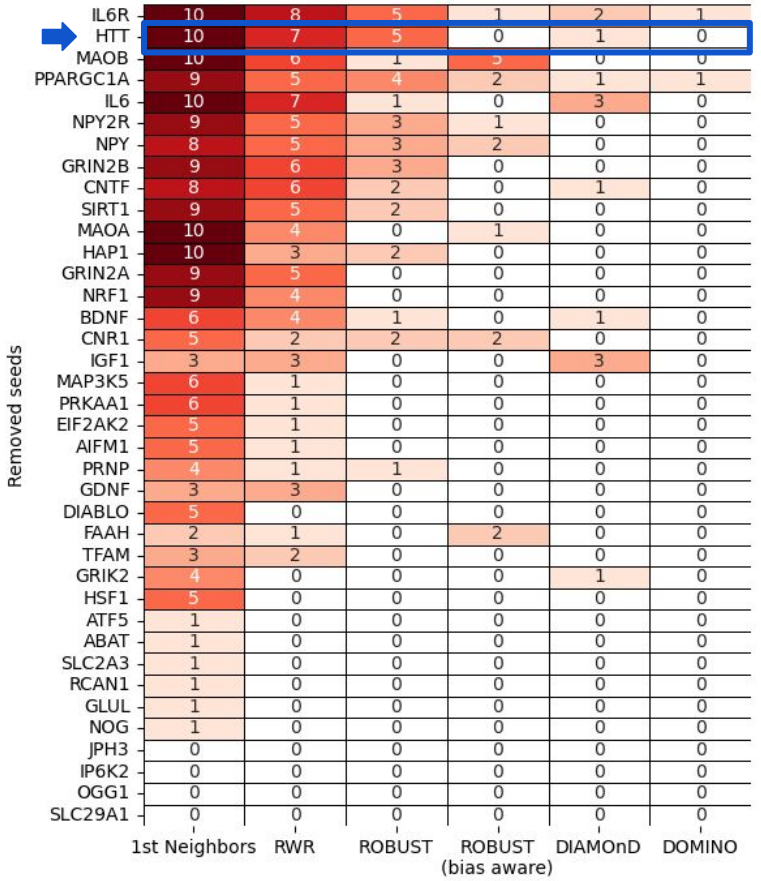


Seed rediscovery – gene level

Can individual highly important genes be rediscovered?

***HTT** is among the most frequently rediscovered seeds in Huntington's Disease*

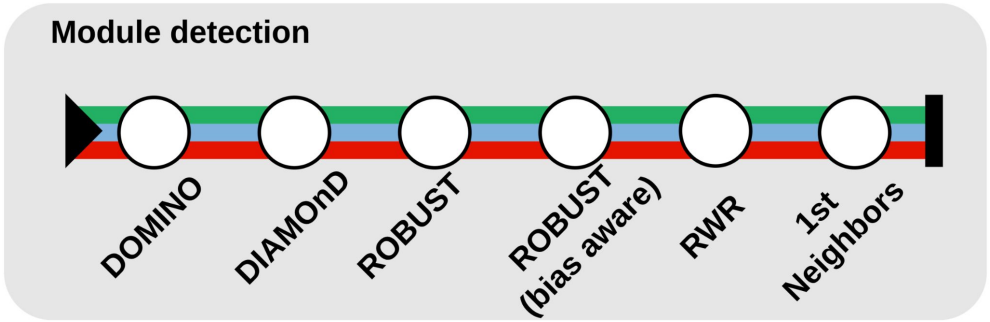
***But:** Most methods cannot rediscover it reliably*



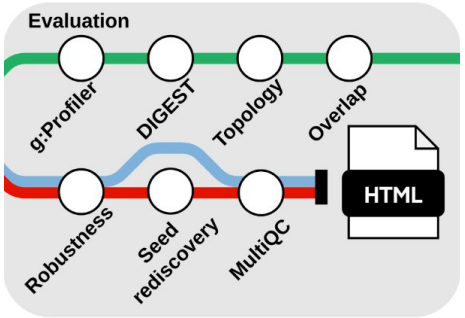
Summary

Pipeline *offers easy access to…*

Disease module discovery methods



Evaluation framework



Systematically **compare** methods

Modular **expandable** design

Track **progress** in the field

Acknowledgments

Developers and co-authors

Johannes Kersting
Lisa Spindler
Joaquim Aguirre-Plans
Chloé Bucheron
Quirin Manz
Fernando Delgado-Chaves
Tanja Pock
Mo Tan
Cristian Nogales
Andreas Maier
Jan Baumbach
Jörg Menche
Emre Guney

RePo4EU
DaiSyBio
nf-core
universität wien
STALICLA

Funded by the European Union

Hands-on

Hands-on

Focuses on:

- Running the pipeline on the Huntington's Disease data obtained earlier
- Working with the pipeline output and MultiQC report

Instructions are in the **`Session3_nextflow_pipeline`** folder:
https://github.com/REPO4EU/network_medicine_drug_repurposing_tutorial/tree/main/Session3_nextflow_pipeline

Best to work with it directly through the GitHub web view: