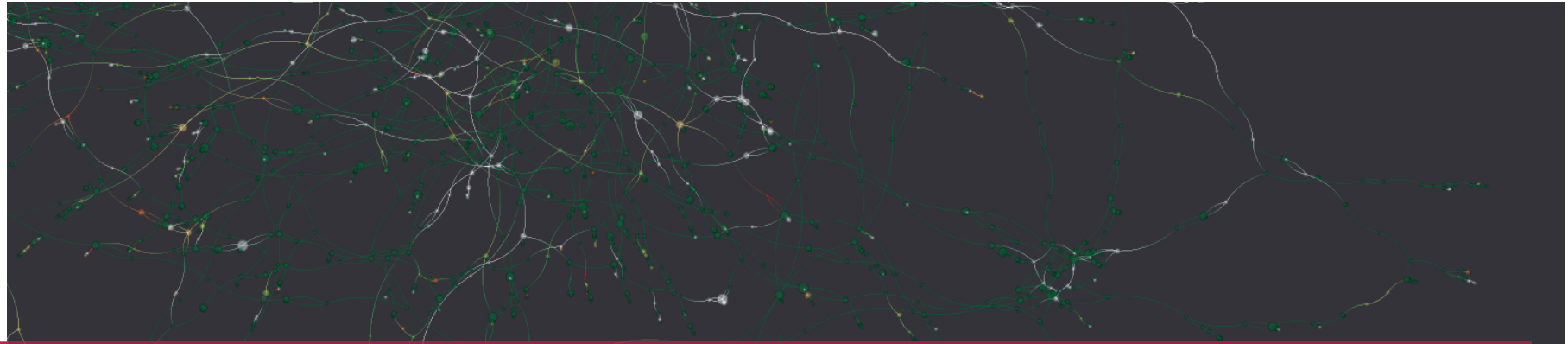
Universität Potsdam

HPI Hasso Plattner Institut
Digital Engineering · Universität Potsdam



Trends in Bioinformatics:
Causal Inference on Gene Expression Data

Philipp Bode

WiSe 18/19

Causal Inference on Gene Expression Data:
Recap: Motivation

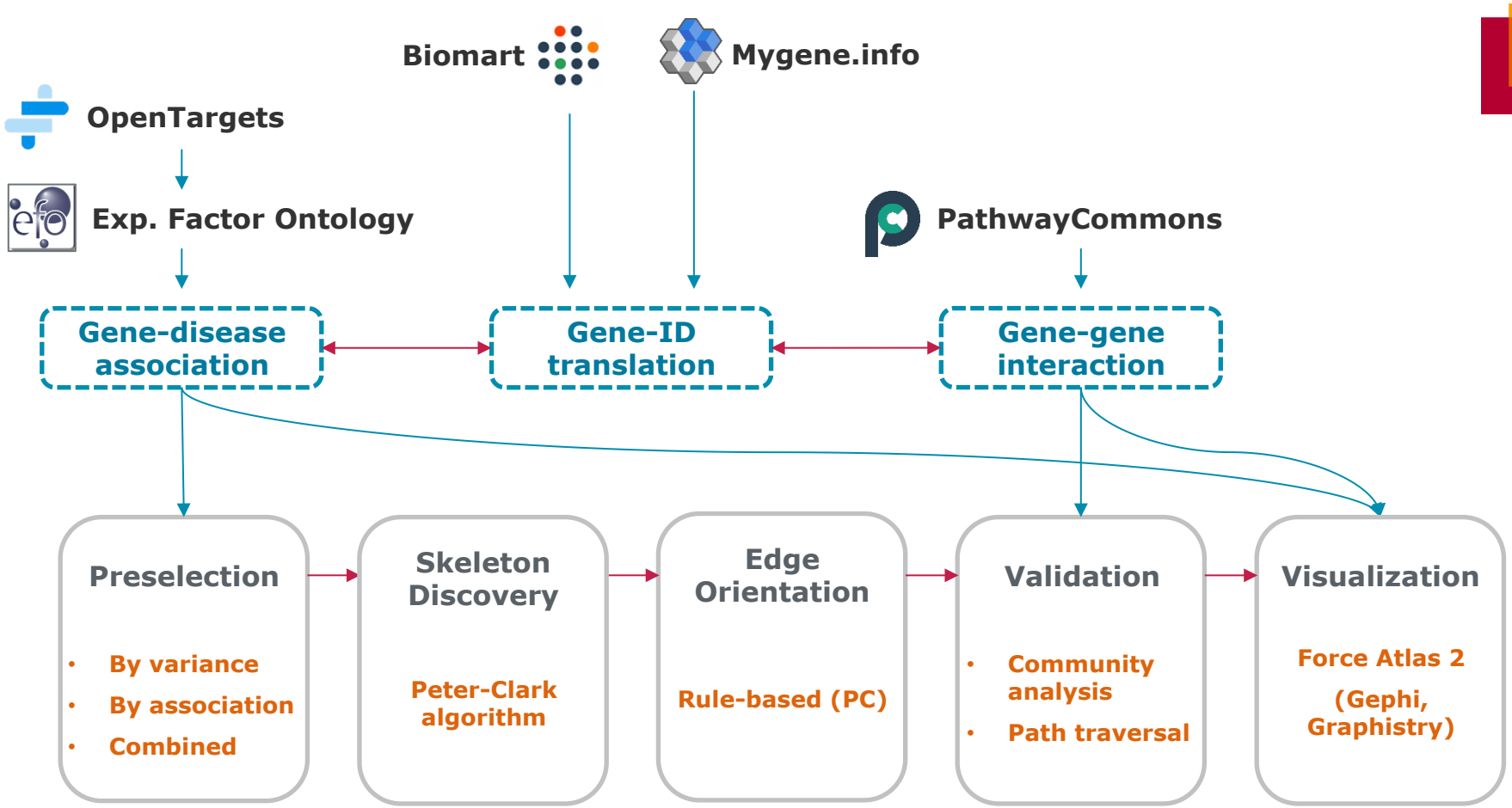
- TCGA RNAseq data
- Estimate gene regulatory networks from expression data:
  - Insights into transcription processes in cancerous cells[1]
- Network inference for now mostly restricted to low-dimensional data due to computational complexity[2]
- Show applicability of constraint-based causal structure learning on high-dimensional, real-world datasets

Causal Inference on Gene Expression Data:
Recap: Challenges

- Feasibility of constraint-based learning approach:
  - High dimensionality: 35K genes
  - Density of underlying causal graph
- (Most probably) many non-linear dependencies
  - Conditional independence tests computationally expensive[4]

Causal Inference on Gene Expression Data: Datasets

- Samples grouped by cancer type:
  - Glioblastoma Multiforme
  - Thyroid Carcinoma
  - Head and Neck Squamous Cell Carcinoma
  - Breast Invasive Carcinoma


Biomart
Mygene.info
OpenTargets
Exp. Factor Ontology
PathwayCommons
Gene-disease association
Gene-ID translation
Gene-gene interaction
Preselection
By variance
By association
Combined
Skeleton Discovery
Peter-Clark algorithm
Edge Orientation
Rule-based (PC)
Validation
Community analysis
Path traversal
Visualization
Force Atlas 2
(Gephi, Graphistry)

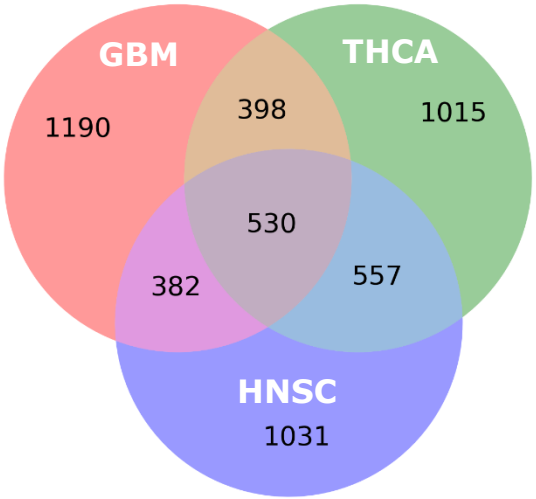
Gene Preselection

- Preselect genes sorted by:
  - Expression variance
  - Disease association score
  - Combined

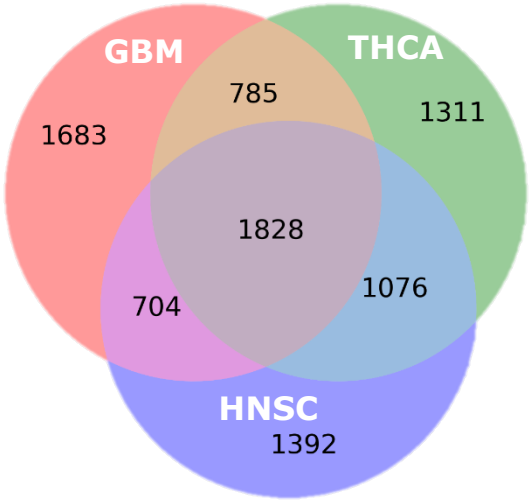
Gene Preselection

- Overlap of genes selected by variance:

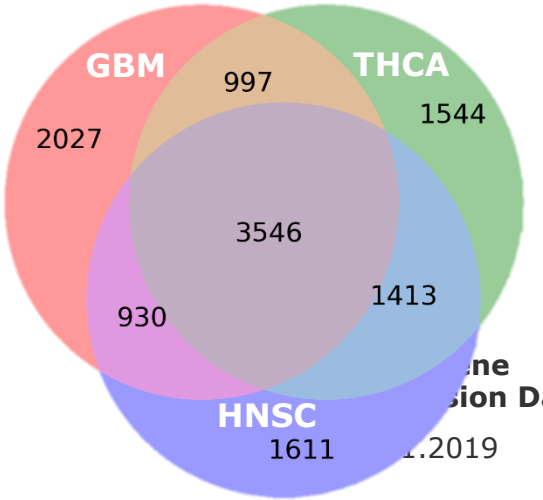
2500 genes



5000 genes



7500 genes

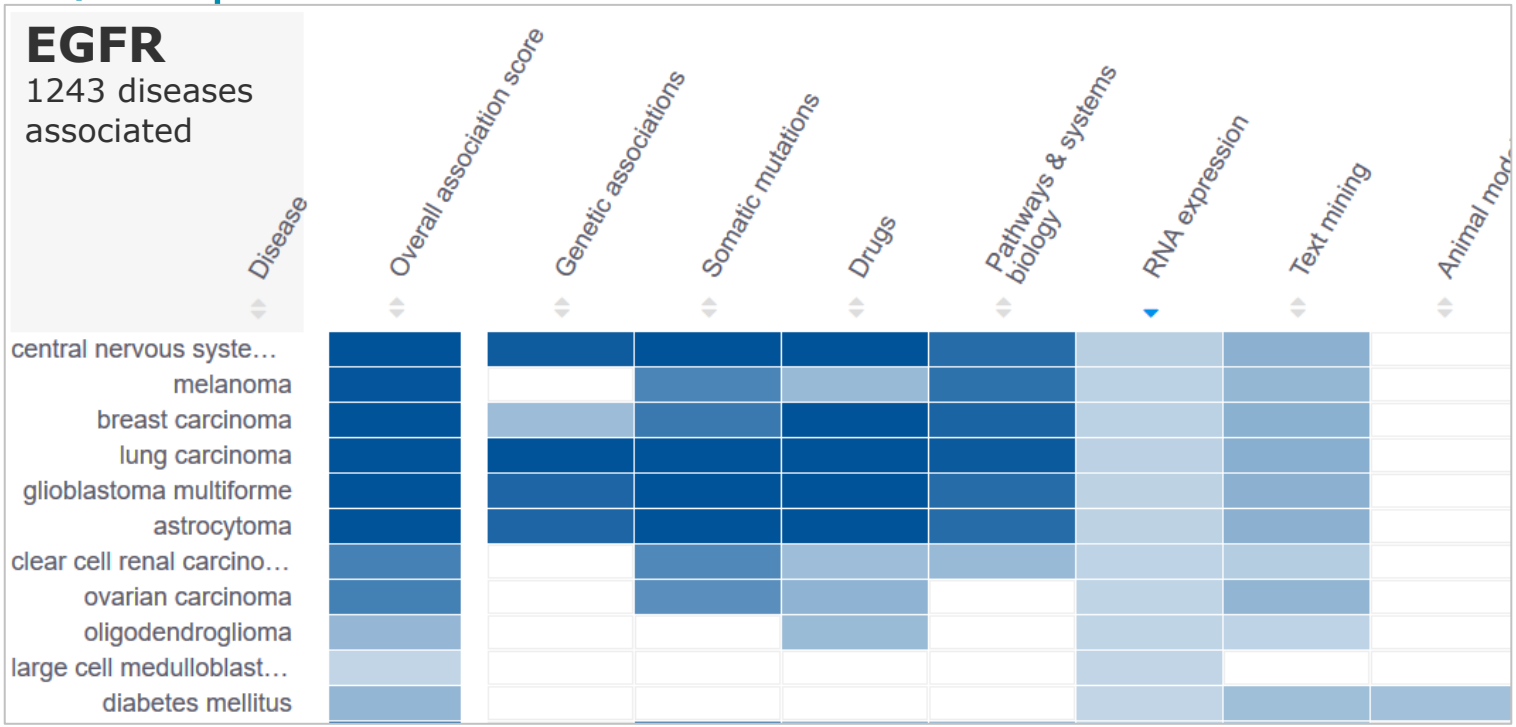


ne
sion Data
1.2019

Gene Preselection

- Gene-disease association score:
  - OpenTargets: Comprehensive aggregation database

Gene Preselection



CI on Gene Expression Data

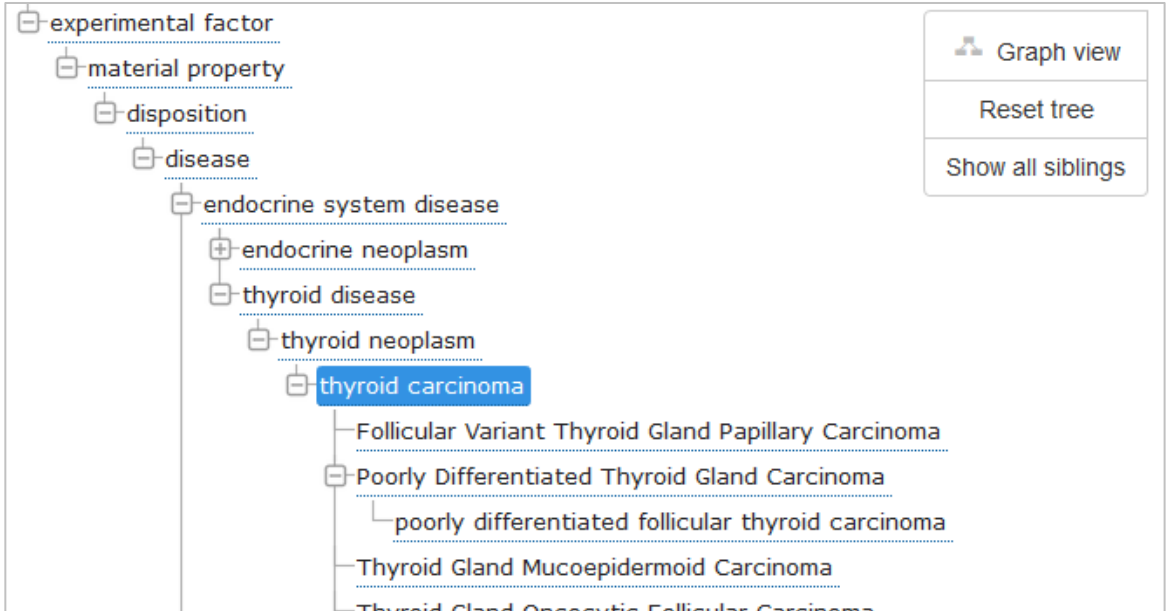
22.01.2019

Gene Preselection

- Gene-disease association score:
  - OpenTargets: Comprehensive aggregation database
- Scores assigned to Experimental Factor Ontology entities

Gene Preselection

- Gene-disease association score:
  - OpenTargets: Comprehensive aggregation database
- Scores assigned to Experimental Factor Ontology entities

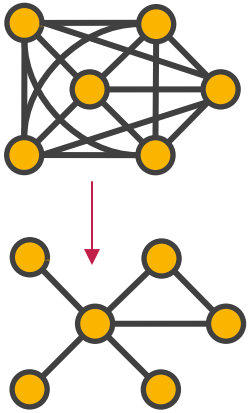


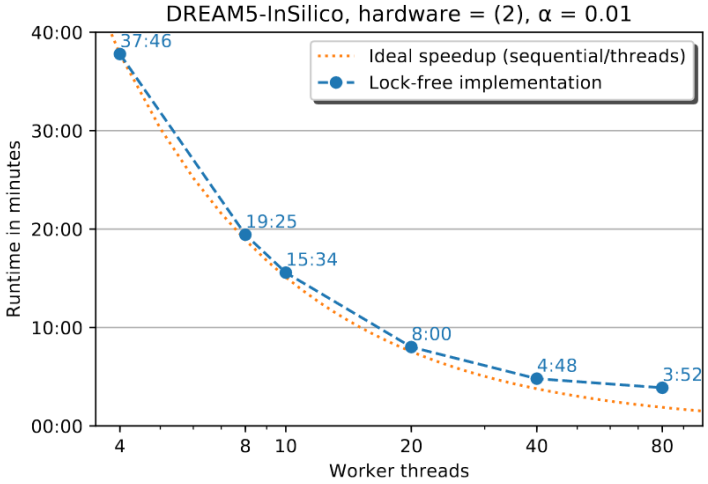
Gene Preselection

- Gene-disease association score:
  - OpenTargets: Comprehensive aggregation database
- Scores assigned to Experimental Factor Ontology entities
- Process OpenTargets JSON dump against EFO subclasses

Skeleton Discovery

- Lock-free, heavily parallelized Peter-Clark skeleton discovery





Skeleton Discovery

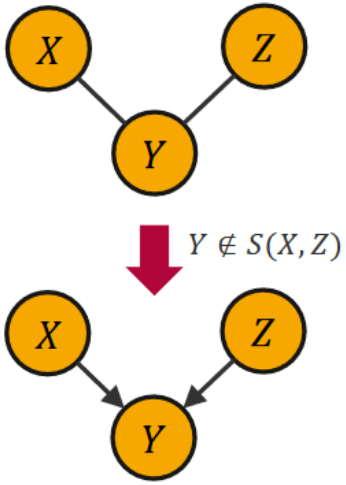
- Lock-free, heavily parallelized Peter-Clark skeleton discovery
- Persist skeleton edges and separation sets for edge orientation

Dataset	Genes	Samples	Runtime (s)	Skeleton Edges
TCGA-GBM	2500	161	26	2115
TCGA-THCA	2500	560	222	4949
TCGA-HNSC	2500	544	215	5372
TCGA-BRCA	2500 - 7000	1216	---	---
TCGA-GBM	5000	161	102	4212

threads = 64, p = 0.05

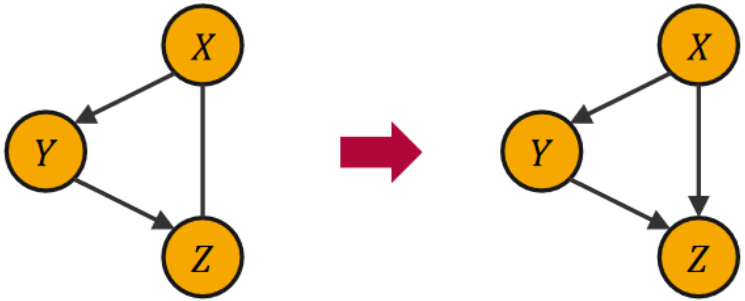
Edge Orientation[3]

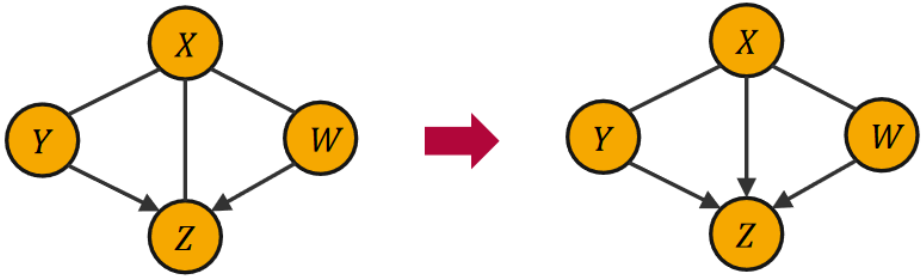
- Rule-based directing of edges:
  1. Determine *v*-Structures



Edge Orientation[3]

- Rule-based directing of edges:
  1. Determine *v*-Structures
  2. Iterate: Apply rules to avoid new *v*-Structures and circles



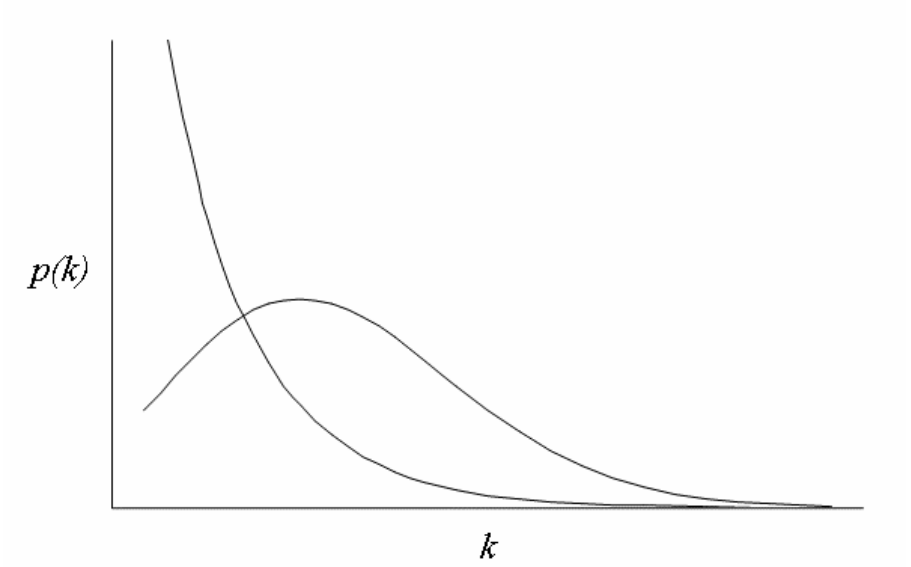


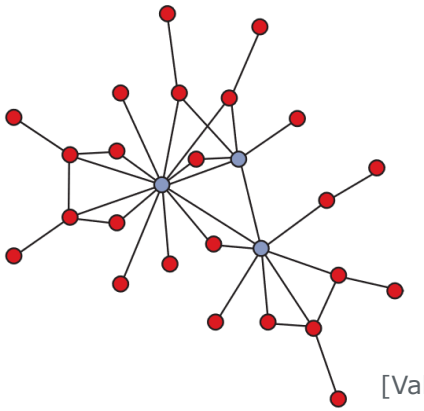
Edge Orientation[3]

- Rule-based directing of edges:
  1. Determine *v*-Structures
  2. Iterate: Apply rules to avoid new *v*-Structures and circles
- Implemented in *pcalg* R-package:
  - 2500 genes: ~ 1.5h
  - 5000 genes: ~ 12h
- Between 70-90% directed edges

Validation

- Gene Regulatory Networks - Scale-free networks? [Val1]
- For "large" k: $P(k) \sim k^{-\gamma}$ with $2 < \gamma < 3$
- Proposed to be evolutionary favorable [Val2]

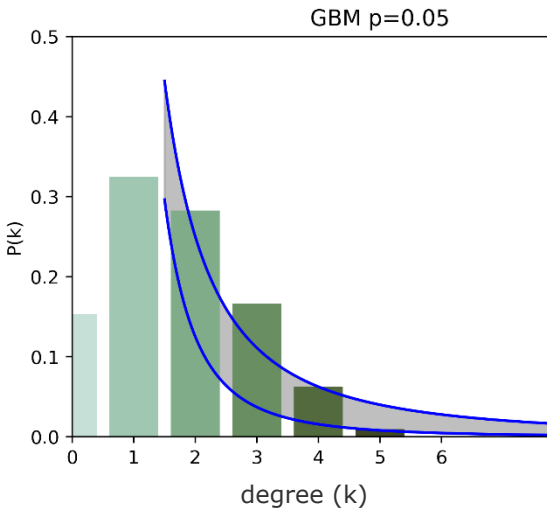


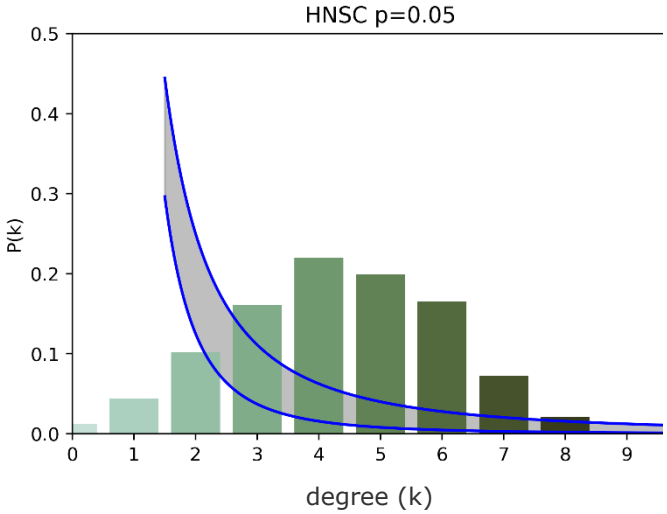


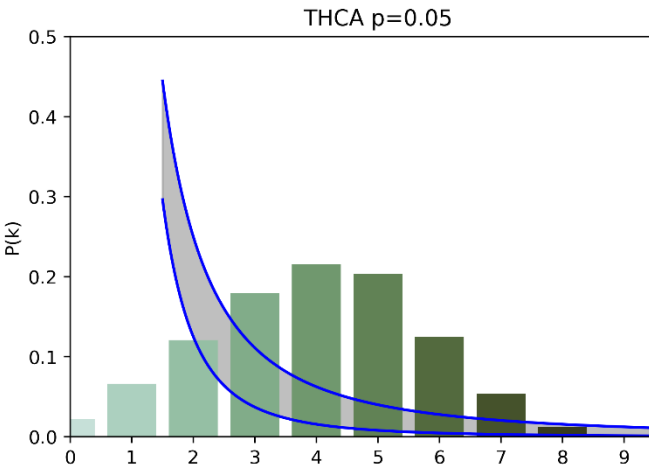
[Val1]

Validation

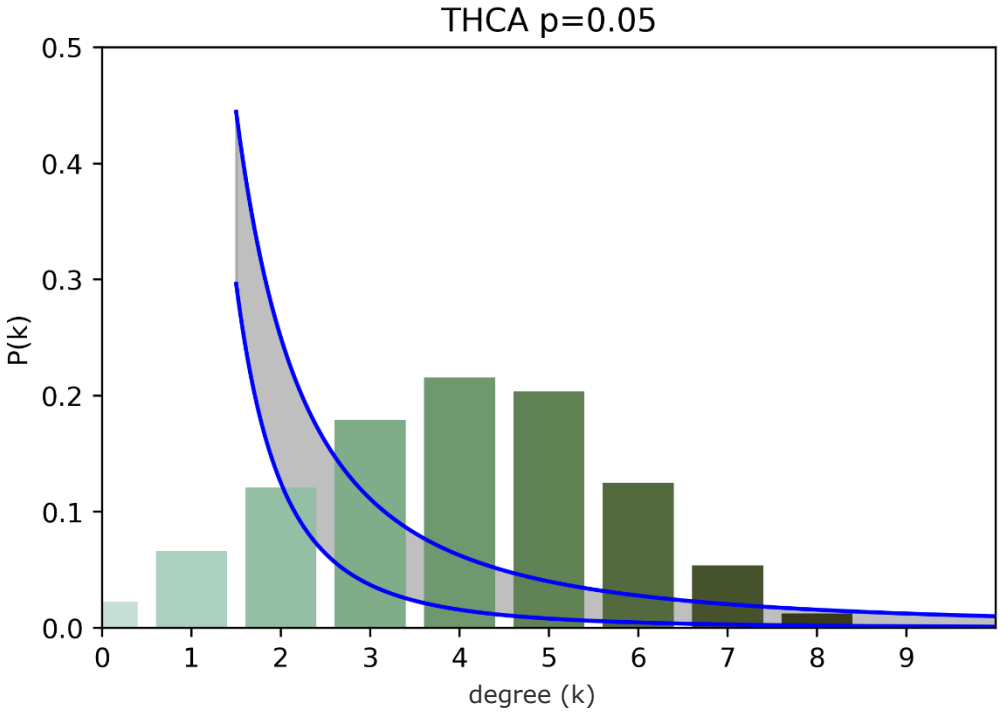
- Degree distribution of resulting graphs







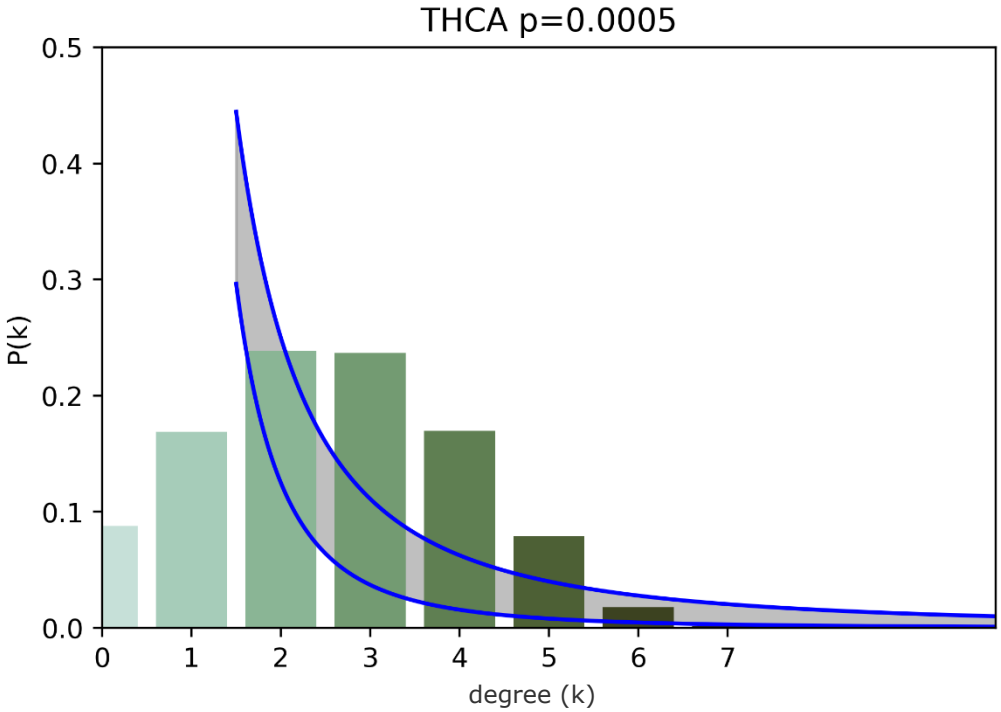
Validation



CI on Gene Expression Data

22.01.2019

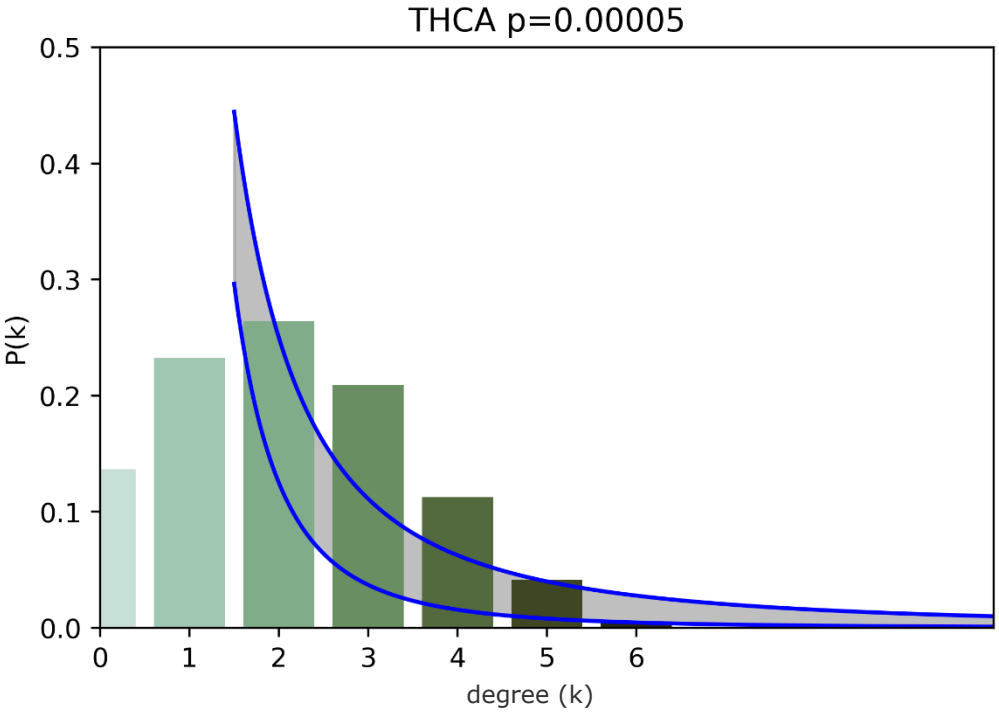
Validation



CI on Gene Expression Data

22.01.2019

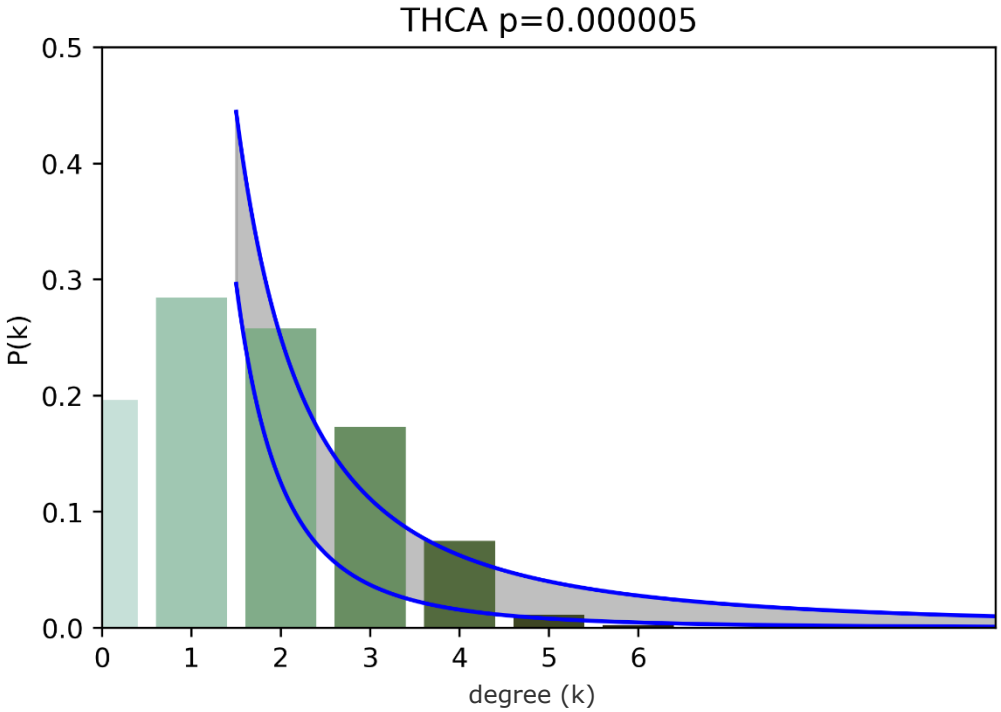
Validation



CI on Gene Expression Data

22.01.2019

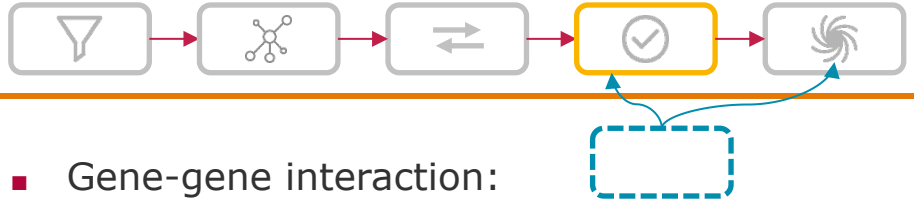
Validation



CI on Gene Expression Data

22.01.2019

Validation



- Gene-gene interaction:
  - Co-Expression (COXPRESdb, GeneFriends) of limited usefulness
  - PathwayCommons: Meta-database for pathway information
    - 2.374.707 total binary interactions
    - 1.161.796 for all TCGA genes
    - 30.000-50.000 on preselected genes
- How to validate graphs?

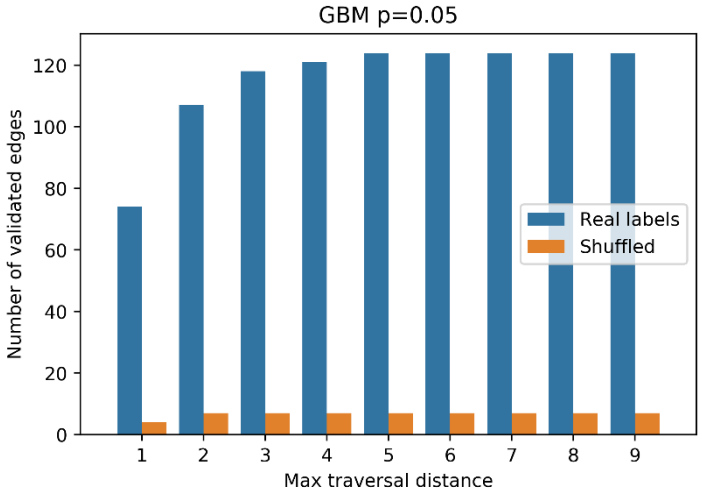
Validation: Communities

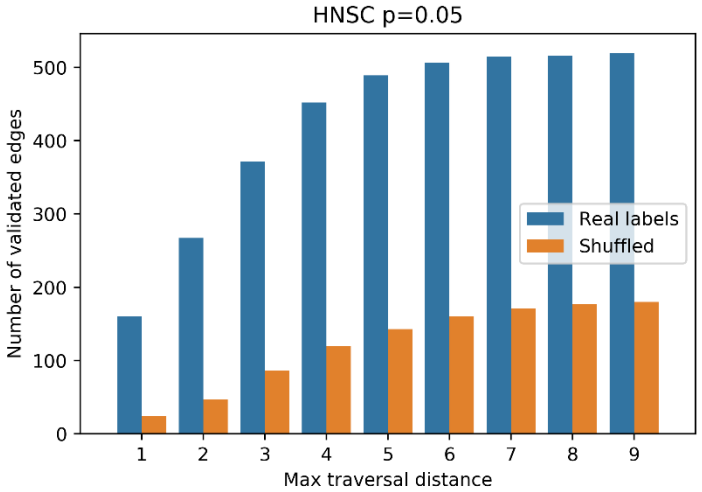
- Detect community structures with InfoMap-approach [Val3]
- Validate within the communities

- But: Boundaries seem to be too arbitrary

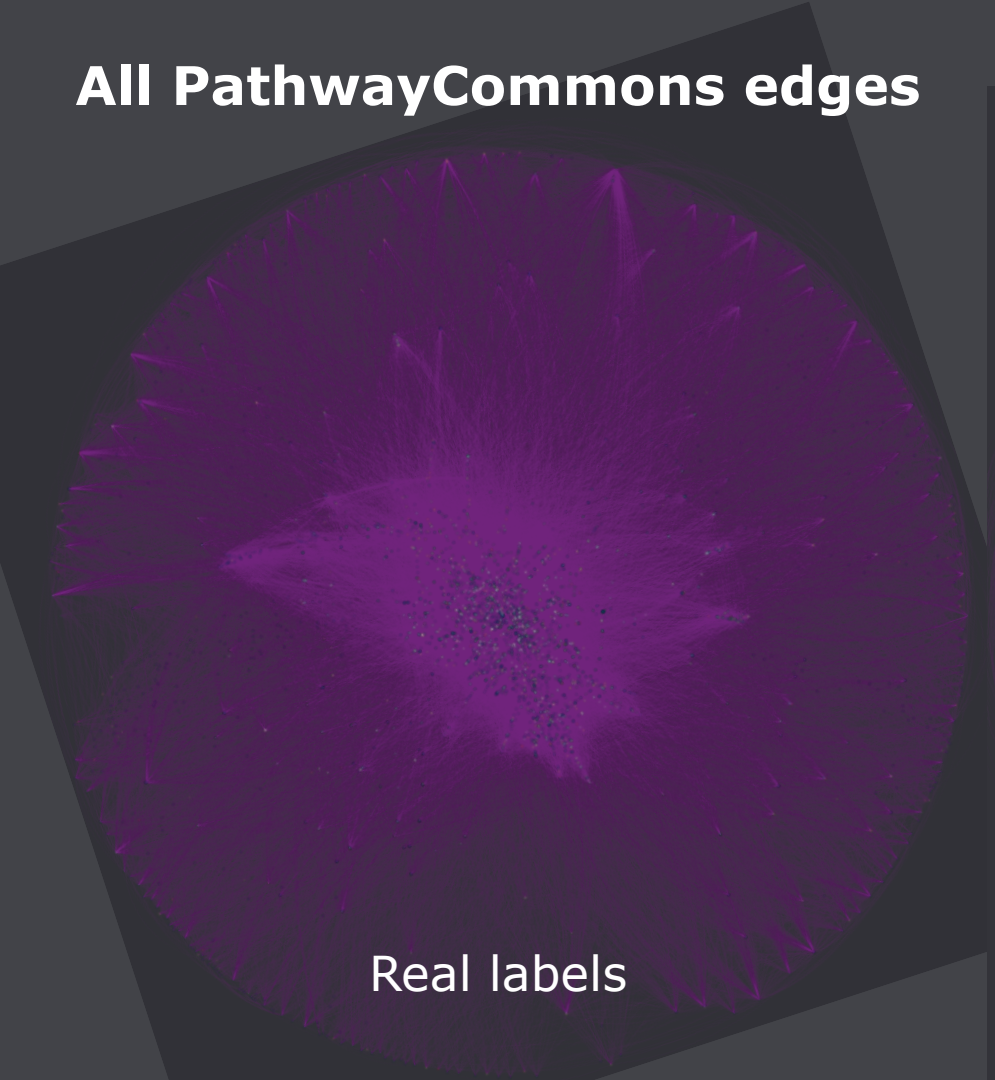
Validation: Neighbourhood Traversal

- Check close neighborhood of every node against PathwayCommons
- Compare to random labelling

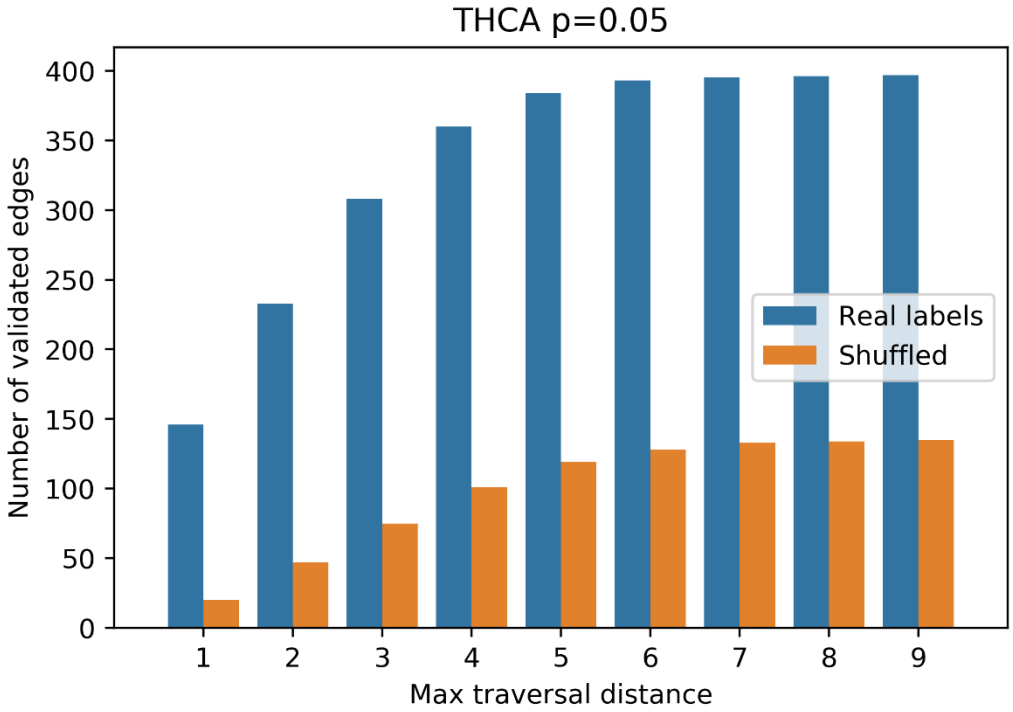




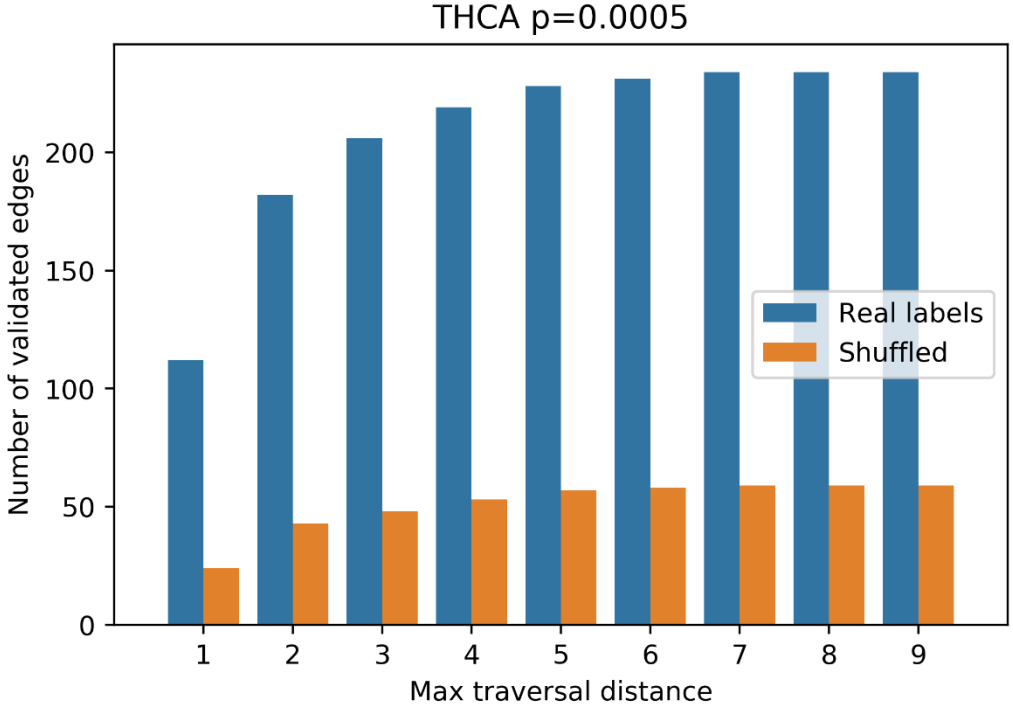
All PathwayCommons edges



Validation


THCA p=0.05
Number of validated edges
0
50
100
150
200
250
300
350
400
1
2
3
4
5
6
7
8
9
Max traversal distance
Real labels
Shuffled

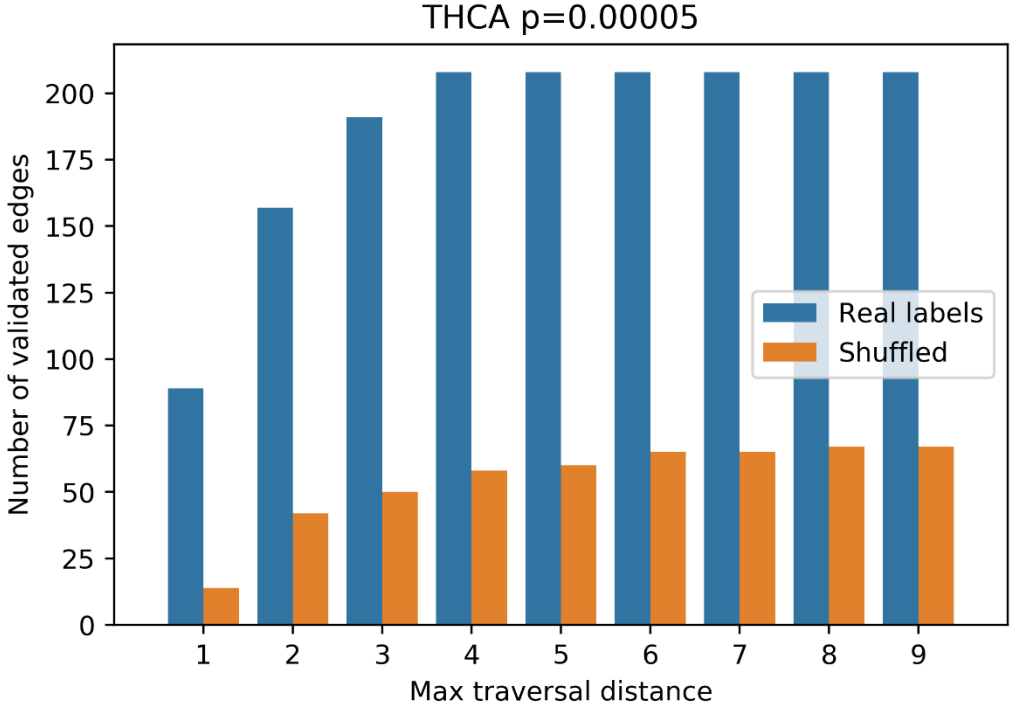
Validation



CI on Gene Expression Data

22.01.2019

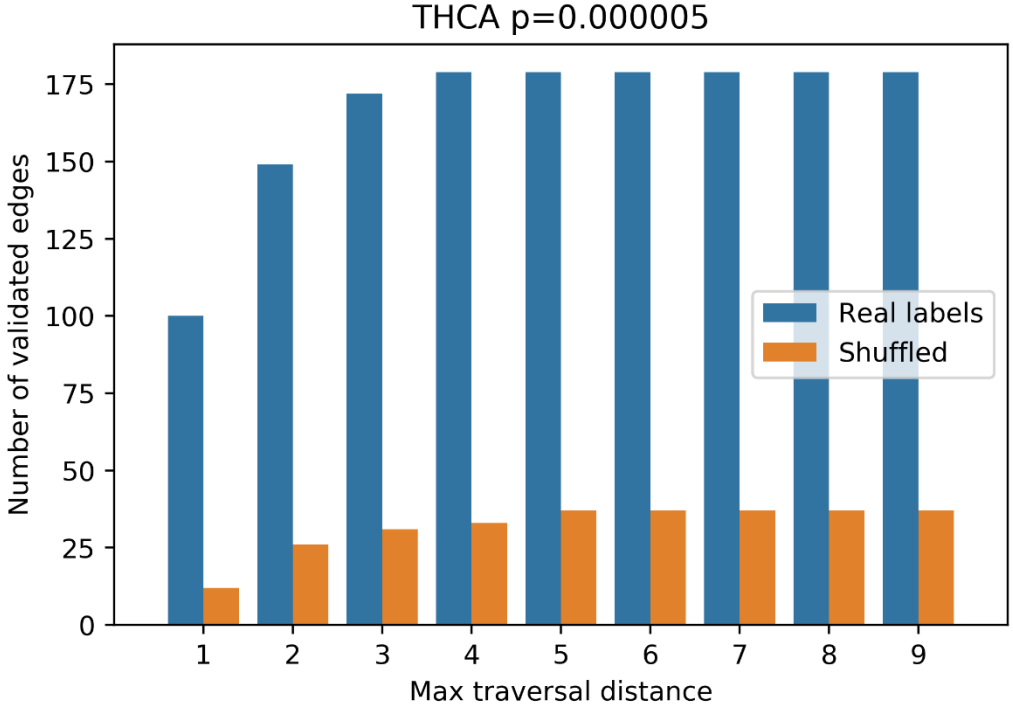
Validation



CI on Gene Expression Data

22.01.2019

Validation

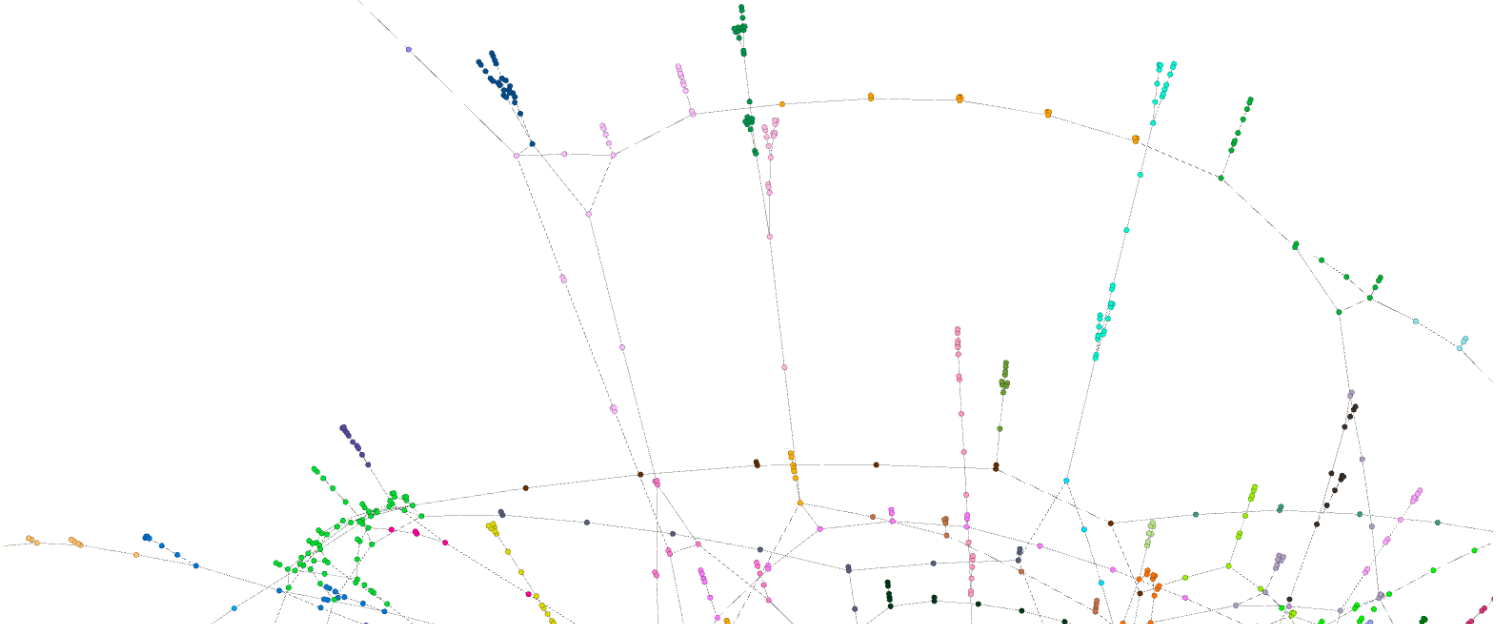

THCA p=0.000005
Number of validated edges
0
25
50
75
100
125
150
175
1
2
3
4
5
6
7
8
9
Max traversal distance
Real labels
Shuffled

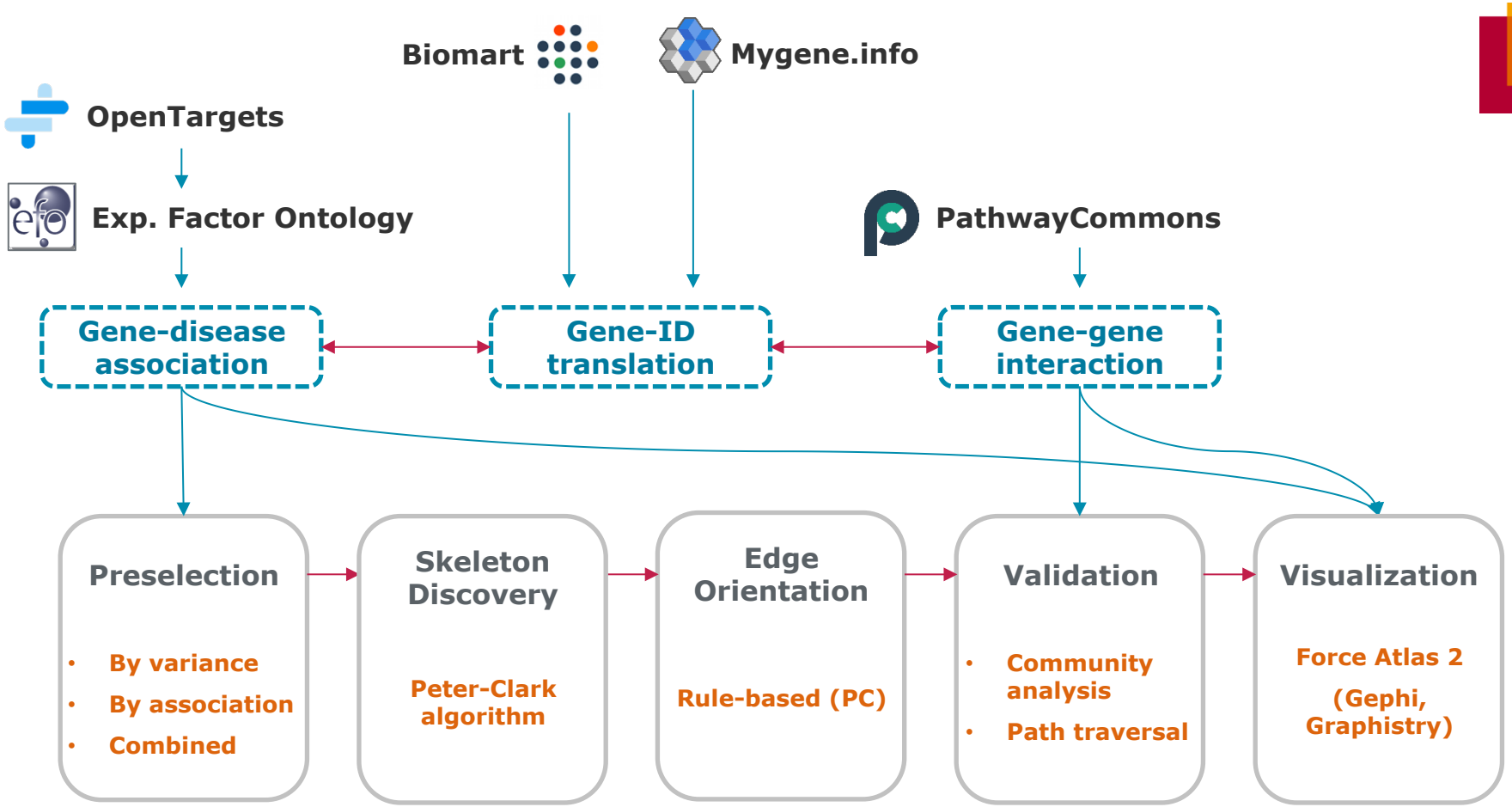
Validation: Evaluation

- Accuracy could probably be increased with incorporation of non-linear independence test
- Estimation of strengths of causal effects not yet included
- Even without preprocessing, results show biological significance
- Next steps: Explore existing configurations further:
  - Gene preselect count, method
  - p-value
  - Rembrandt GBM dataset

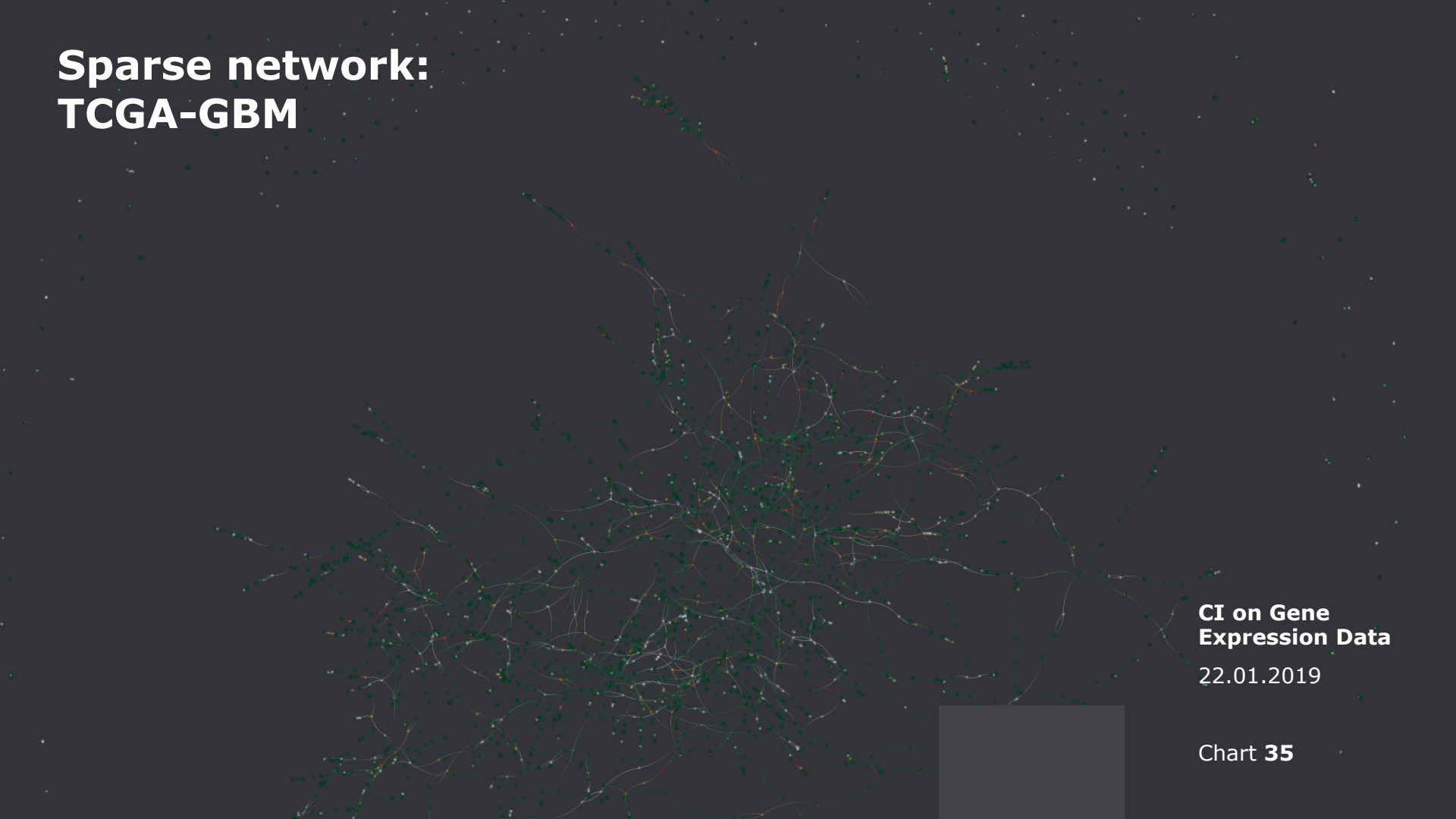
Visualization

- Force Atlas 2: Force-directed graph layout
- Gephi: Open-Source client with streaming interface
- Graphistry: Proprietary API for in-browser view

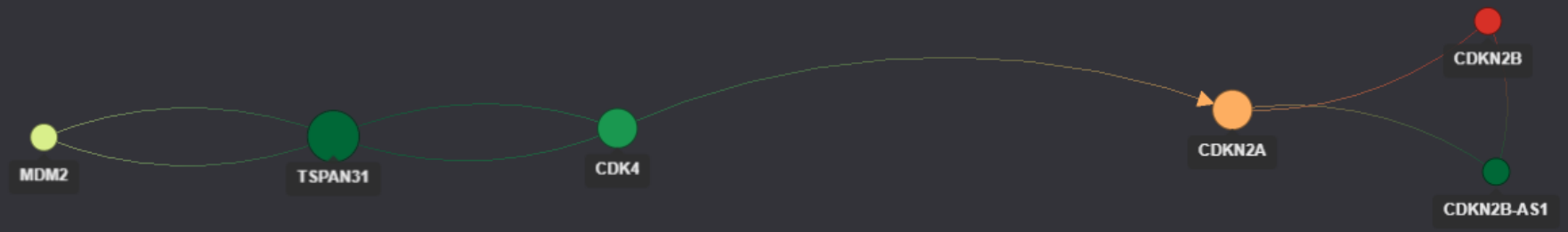


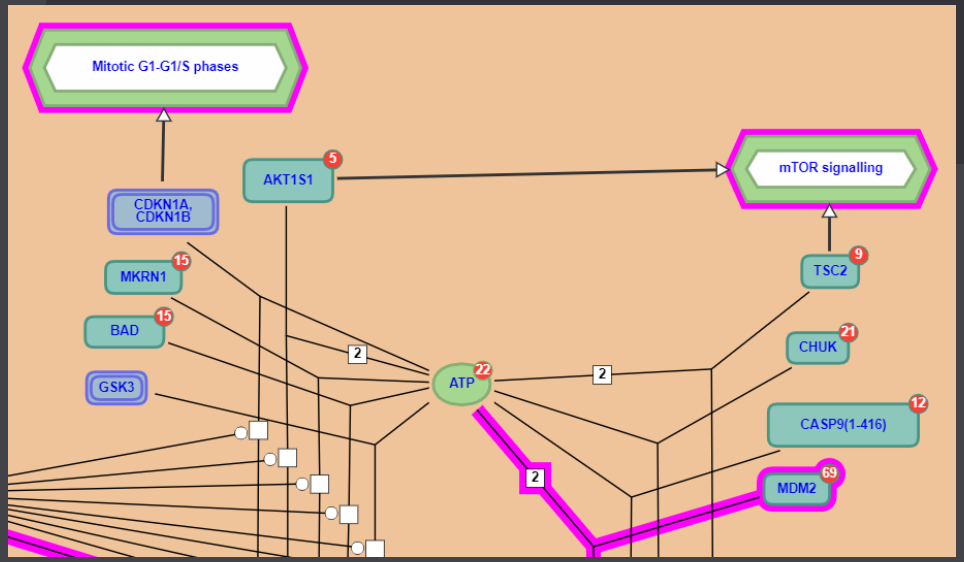

Biomart
Mygene.info
OpenTargets
Exp. Factor Ontology
PathwayCommons
Gene-disease association
Gene-ID translation
Gene-gene interaction
Preselection
By variance
By association
Combined
Skeleton Discovery
Peter-Clark algorithm
Edge Orientation
Rule-based (PC)
Validation
Community analysis
Path traversal
Visualization
Force Atlas 2
(Gephi, Graphistry)

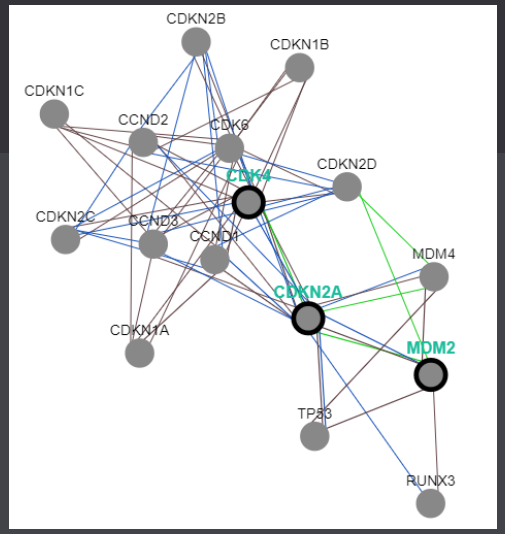
Sparse network: TCGA-GBM



TCGA-GBM: TP53 proto-oncogenes





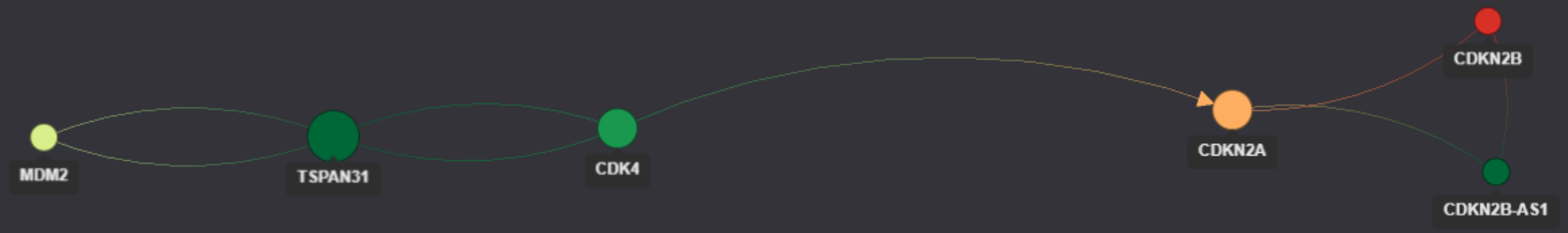


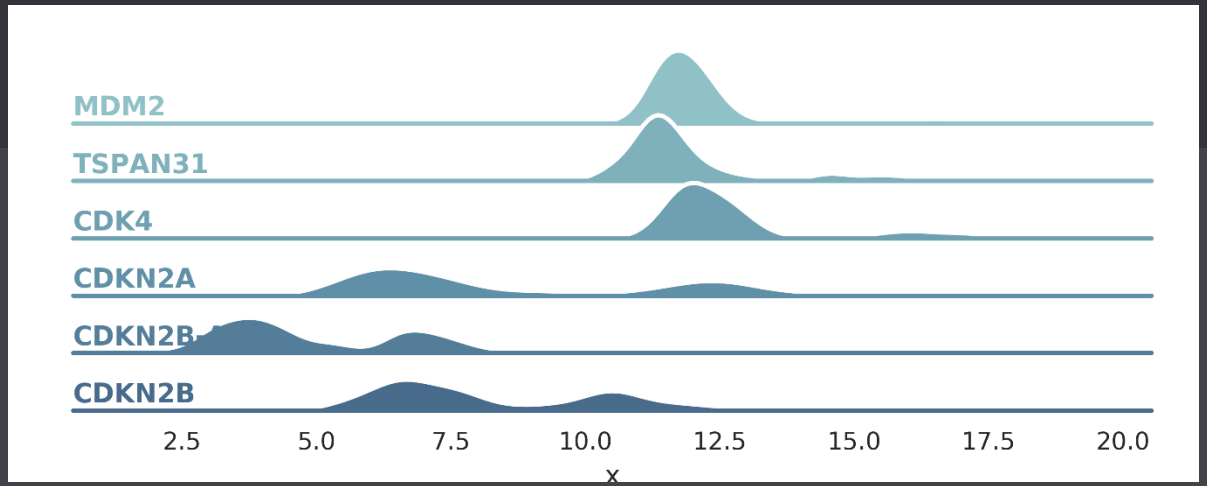
CI on Gene Expression Data

22.01.2019

Chart **36**

TCGA-GBM: TP53 proto-oncogenes



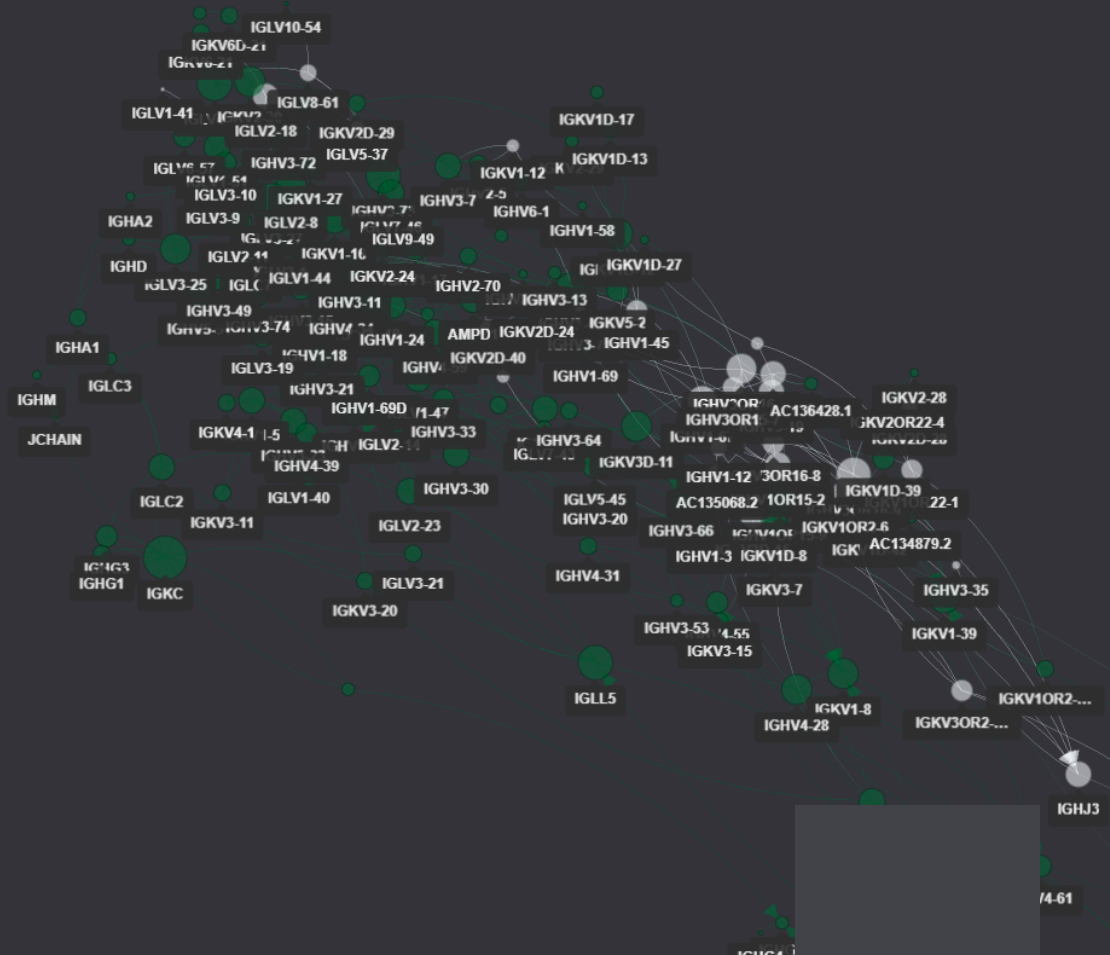


CI on Gene Expression Data

22.01.2019

Chart **37**

TCGA-THCA: Immunoglobulin-Cluster



CI on Gene Expression Data

22.01.2019

Chart **38**

Sources

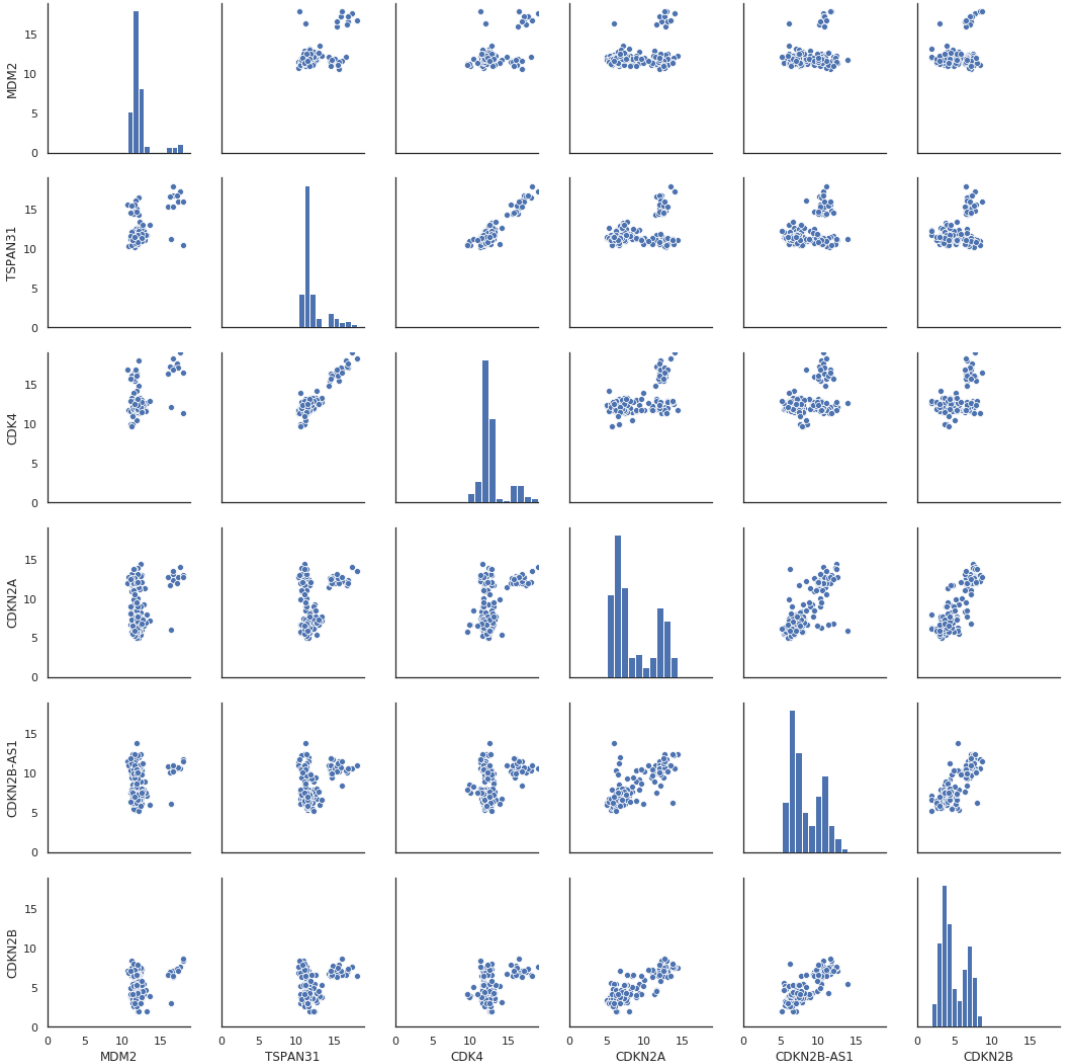
- [1] ICGC-TCGA DREAM Somatic Mutation Calling - RNA Challenge https://www.synapse.org/#!Synapse:syn2813589/wiki/401435
- [2] Le, Thuc Duy, Taosheng Xu, Lin Liu, Hu Shu, Tao Hoang, and Jiuyong Li. "ParallelPC: An R Package for Efficient Causal Exploration in Genomic Data." In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 207-218. Springer, Cham, 2018.
- [3] Causal Inference – Theory and Applications: https://hpi.de/plattner/teaching/archive/summer-term-2018/causal-inference-theory-and-applications.html
- [4] Ramsey, Joseph D. "A scalable conditional independence test for nonlinear, non-Gaussian data." *arXiv preprint arXiv:1401.5031* (2014).
- [Val1] Barabasi, Albert-Laszlo, and Zoltan N. Oltvai. "Network biology: understanding the cell's functional organization." *Nature reviews genetics* 5, no. 2 (2004): 101.
- [Val2] Leclerc, Robert D. "Survival of the sparsest: robust gene networks are parsimonious." *Molecular systems biology* 4, no. 1 (2008): 213.
- [Val3] Bohlin, Ludvig, Daniel Edler, Andrea Lancichinetti, and Martin Rosvall. "Community detection and visualization of networks with the map equation framework." In *Measuring Scholarly Impact*, pp. 3-34. Springer, Cham, 2014.
- [Viz1] https://reactome.org/PathwayBrowser/#/R-HSA-1257604&FLG=MDM2

Datasets: Composition

Project	Total	Tumor Primary	Normal tissue
TCGA-GBM	161	156	5
TCGA-THCA	560	502	58
TCGA-HNSC	544	500	44

Edge Orientation: Results

```
GBM p=0.05 (2834, 1396, 719, 0.660047281323877)
GBM_5000 p=0.05 (5632, 2792, 1420, 0.6628679962013295)
HNSC p=0.05 (5786, 4958, 414, 0.922933730454207)
HNSC p=0.0005 (3823, 2623, 600, 0.8138380390940118)
THCA p=0.05 (5395, 4503, 446, 0.909880783996767)
THCA p=0.0005 (3800, 2574, 613, 0.8076561029181047)
THCA p=0.00005 (3243, 1927, 658, 0.7454545454545455)
THCA p=0.000005 (2768, 1456, 656, 0.6893939393939394)
```

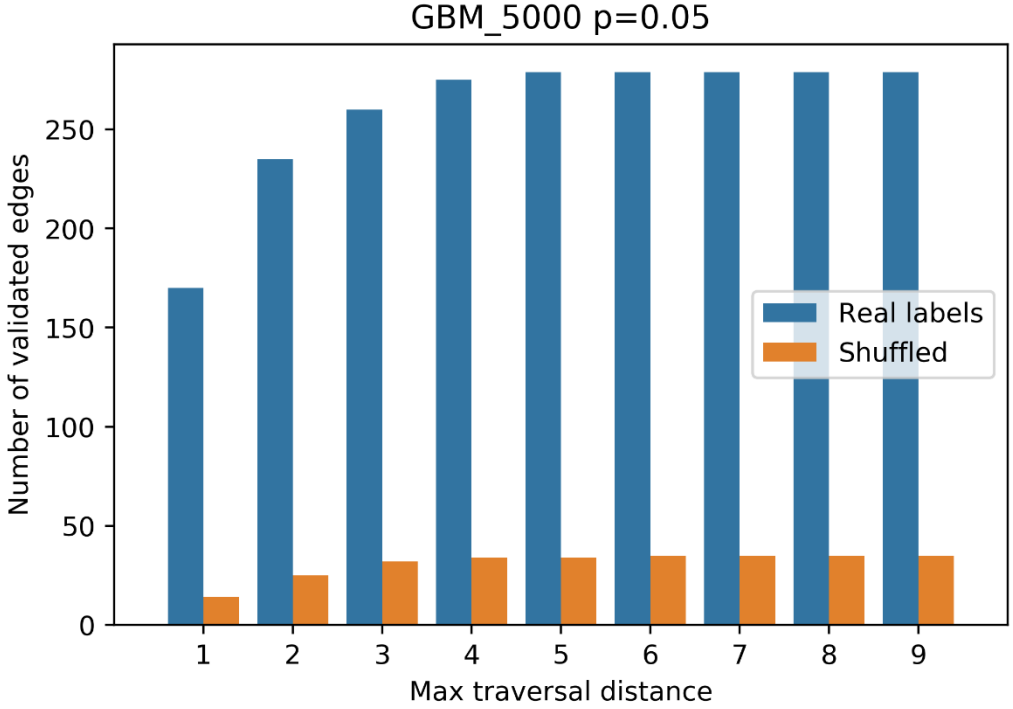


CI on Gene Expression Data

22.01.2019

Chart **42**

Validation



CI on Gene Expression Data

22.01.2019

Analysis Flow