Laboratory for Statistical Artificial InteLligence & INtegrative Genomics

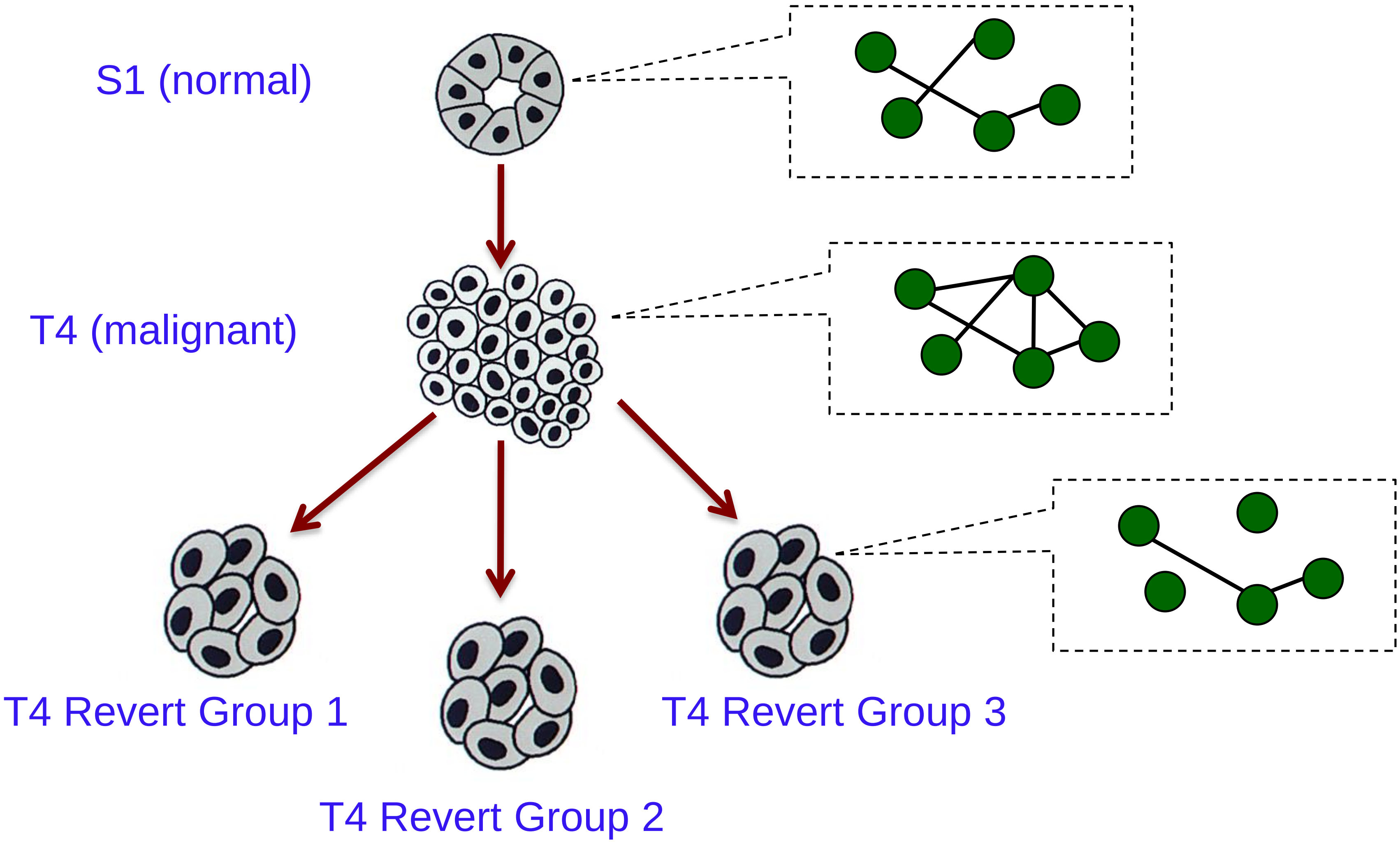
TREEGL: Reverse Engineering Tree-Evolving Gene Networks Underlying Developing Biological Lineages

Ankur P. Parikh*, Wei Wu*, Ross E. Curtis, Eric P. Xing

Carnegie Mellon University

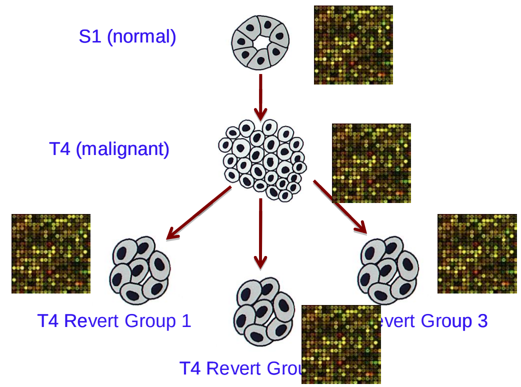
University of Pittsburgh

Progression and Reversion of Breast Cancer cells

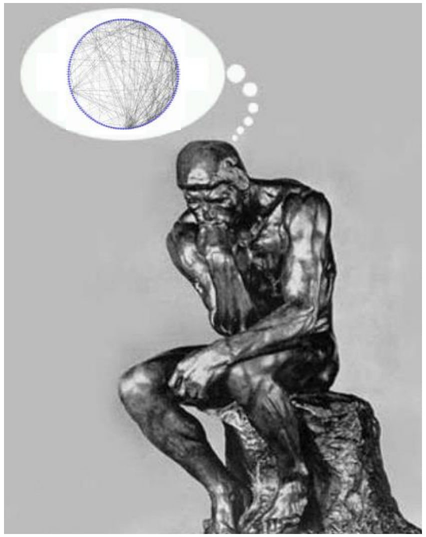


Existing Work

- Pool samples, and infer a single network



- or estimate cell-line specific network independently



- We assume:
  - The network evolves, and therefore are related
  - we need to **INFER** the **Lineage of Networks** from as few as ONE microarray per cell line

$$\mathcal{D} = \{x_1^i, \ldots, x_p^i\}_{i=1}^n \quad \Rightarrow \quad G_1, \ldots, G_n$$

Our Approach

- A sparse regression approach to **jointly** estimating all the networks in the genealogy (which we call ***Treegl***)
- L1 penalty enforces sparseness
- Total variation penalty penalizes differences among adjacent cells in the genealogy, but also allows for sharp differences

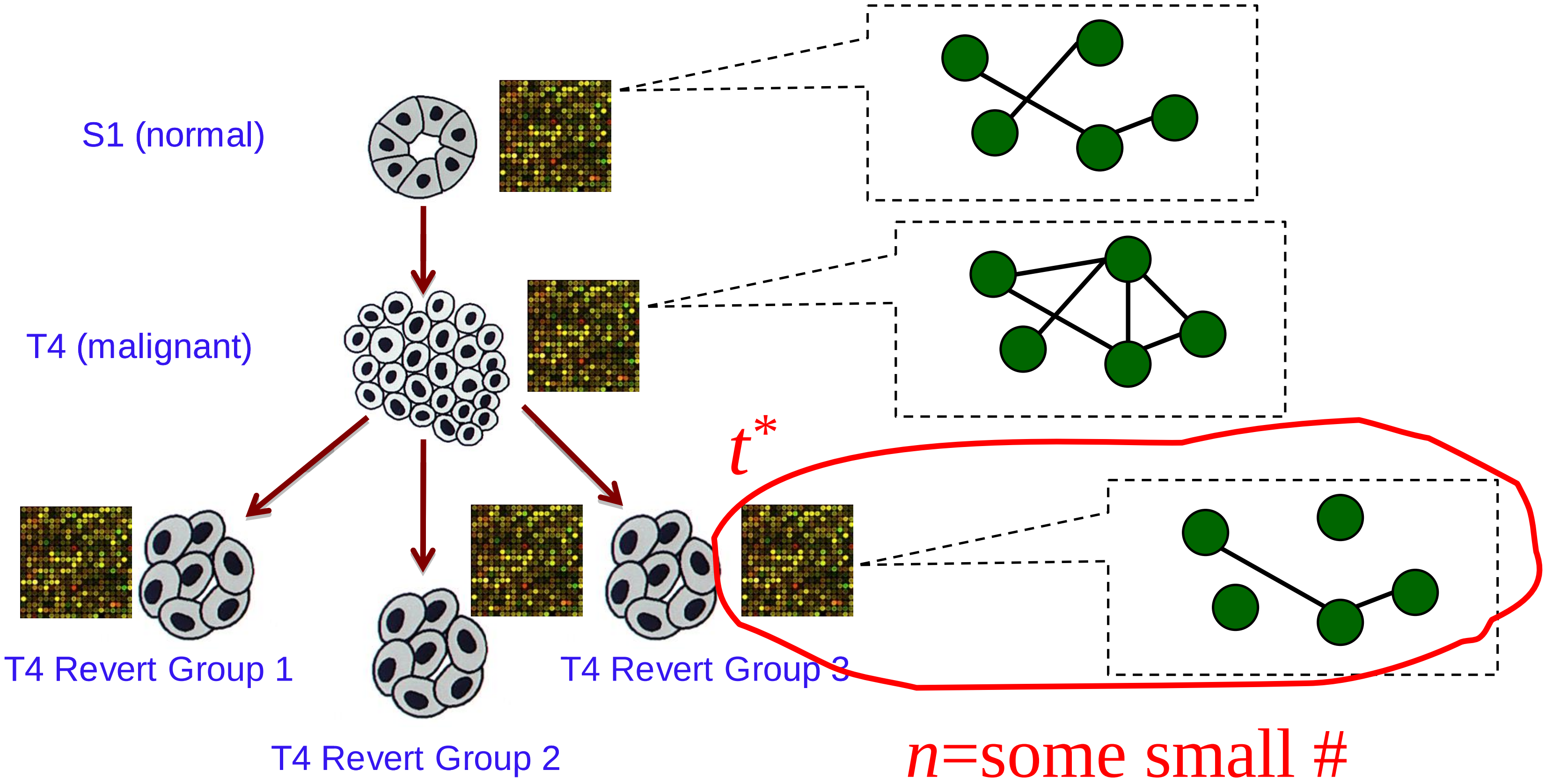
Outline

- **Theory and Algorithm**
  - Sparsity and the LASSO
  - Neighborhood Selection for Network Reconstruction
  - Our algorithm: Treegl

- **Breast Cancer Progression and Reversal Analysis**
  - Description of Data
  - Overview of Recovered Networks
  - Interactions among GO groups
  - GO analysis

Theory and Algorithm

Reverse engineer lineage-specific "rewiring" gene networks



Challenges

- Very small sample size
  - observations are scarce and costly
- Noisy data
- Large dimensionality of the data (~10^4 genes)
  - # variables >> # of samples
  - least squares regression fails!
  - complexity regularization is required
- And now the data are non-iid since underlying probability distribution is changing !

Sparsity

- One common assumption to make **sparsity**.

- **Makes biological sense:** Genes are only assumed to interface with small groups of other genes.

- **Makes statistical sense:** Learning is now feasible in high dimensions with small sample size

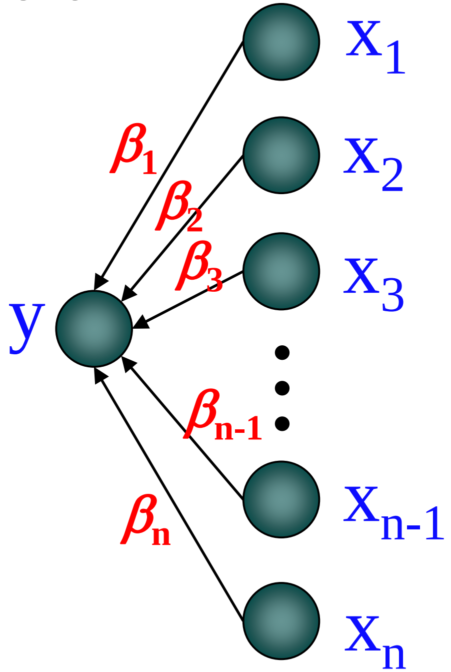
Sparsity: In a mathematical sense

- Consider least squares linear regression problem:
- Sparsity means most of the beta's are zero.

$$\hat{\boldsymbol{\beta}} = \mathrm{argmin}_{\beta} ||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2$$

subject to:

$$\sum_{j=1}^{p} \mathbb{I}[|\beta_j| > 0] \leq C$$



- But this is not convex!!! Many local optima, computationally intractable.

L1 Regularization (LASSO) [Tibshirani 1996]

- A convex relaxation.

Constrained Form

$$\hat{\boldsymbol{\beta}} = \mathrm{argmin}_{\beta} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2$$

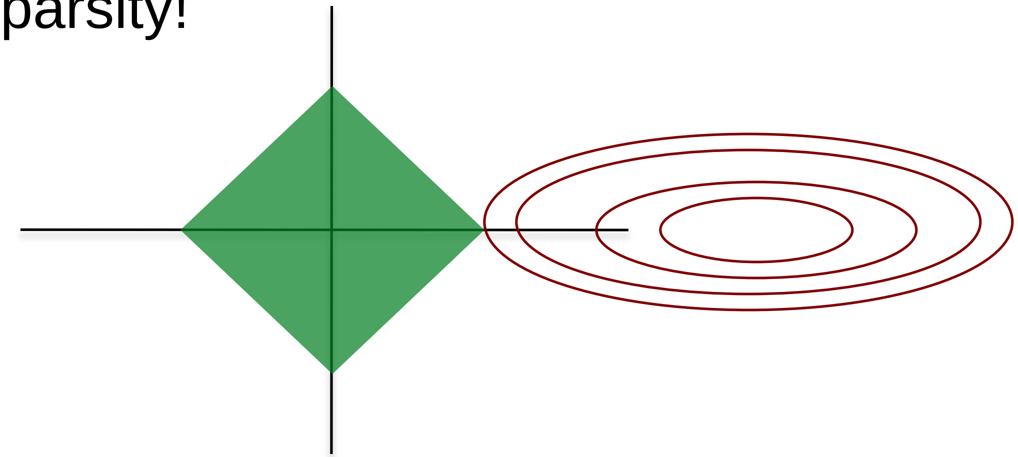
subject to:

$$\sum_{j=1}^{p} |\beta_j| \leq C$$

Lagrangian Form

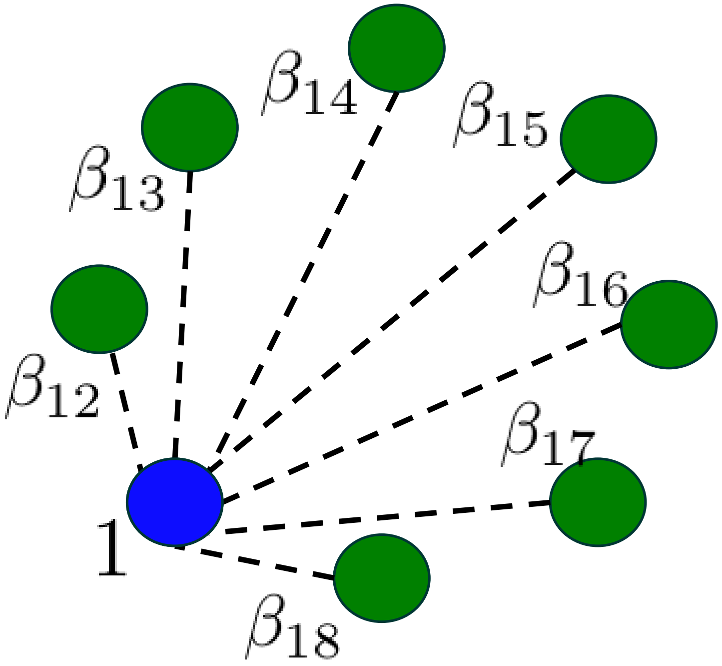
$$\hat{\boldsymbol{\beta}} = \mathrm{argmin}_{\beta} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|_1$$

- Still enforces sparsity!



Network Learning with the Graphical LASSO [Meinshausen and Buhlmann 2006]

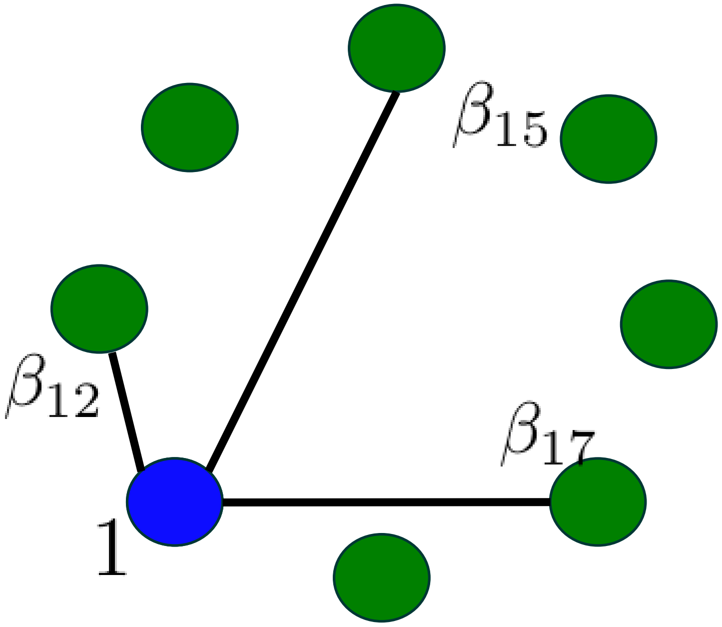
- Perform neighborhood selection



Network Learning with the Graphical LASSO

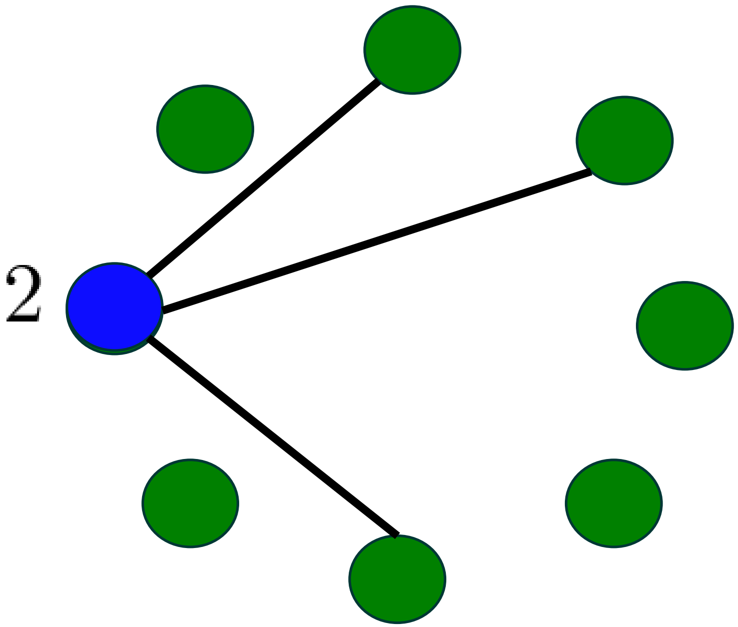
- Use the LASSO to select the neighborhood of each node

$$\hat{\boldsymbol{\beta}}_{\mathbf{1}} = \operatorname{argmin}_{\beta_1} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_{\mathbf{1}}\|^2 + \lambda \|\boldsymbol{\beta}_{\mathbf{1}}\|_1$$



Network Learning with the Graphical LASSO

- Repeat this for every node

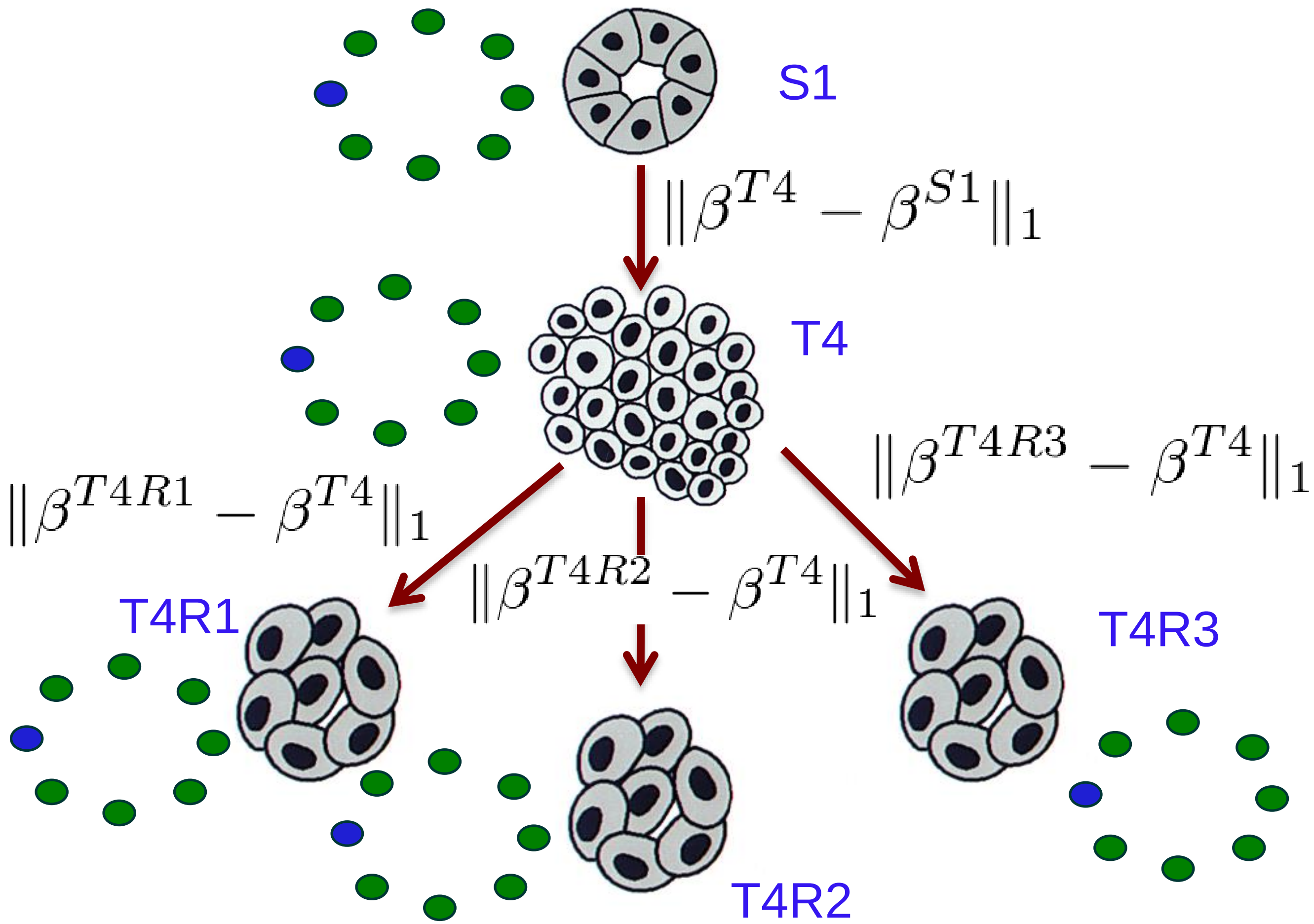


But this can only estimate one network....

- We need to learn a whole genealogy of networks.
- Too few samples to learn each network independently
- How to ``share information” among the samples of different cell types while still exposing sharp differences?

The Total Variation Penalty

Penalize differences between networks of adjacent cell types



Our Method: Tree-Guided Graphical Lasso (Treegl)

RSS for all cell types

$$\hat{\boldsymbol{\beta}}^{(1)}, \ldots, \hat{\boldsymbol{\beta}}^{(N)} = \operatorname{argmin}_{\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(N)}} \sum_{n=1}^{N} \|\mathbf{Y}^{(n)} - \mathbf{X}^{(n)} \boldsymbol{\beta}^{(n)}\|^2 + \lambda_1 \sum_{n=1}^{N} \|\boldsymbol{\beta}^{(n)}\|_1 + \lambda_2 \sum_{n=2}^{N} \|\boldsymbol{\beta}^{(n)} - \boldsymbol{\beta}^{\pi(n)}\|_1$$

sparsity

Sparsity of difference

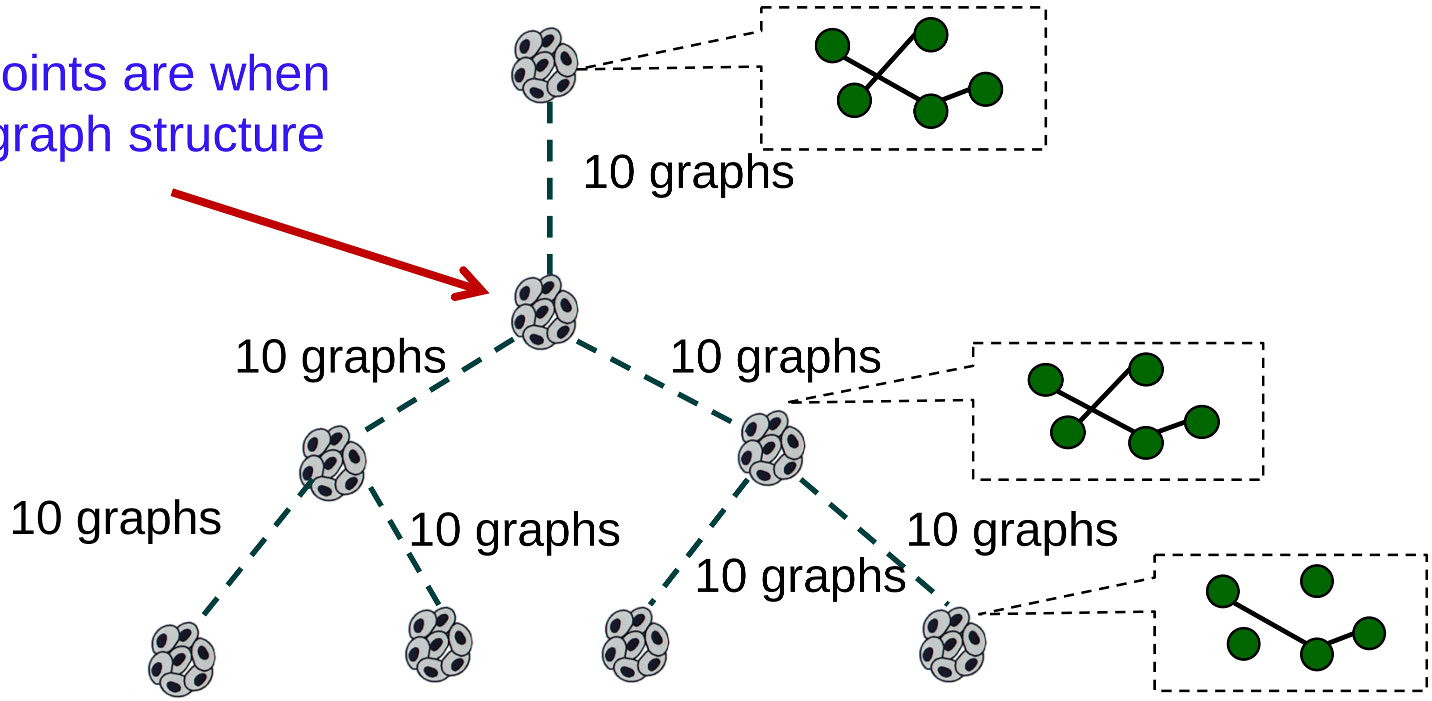
Optimization

- Loss function is convex
- Used **CVX** – MATLAB package for convex optimization
- For large scale problems, the proximal accelerated gradient method of Chen et al. (2011) can be used

Simulation Framework

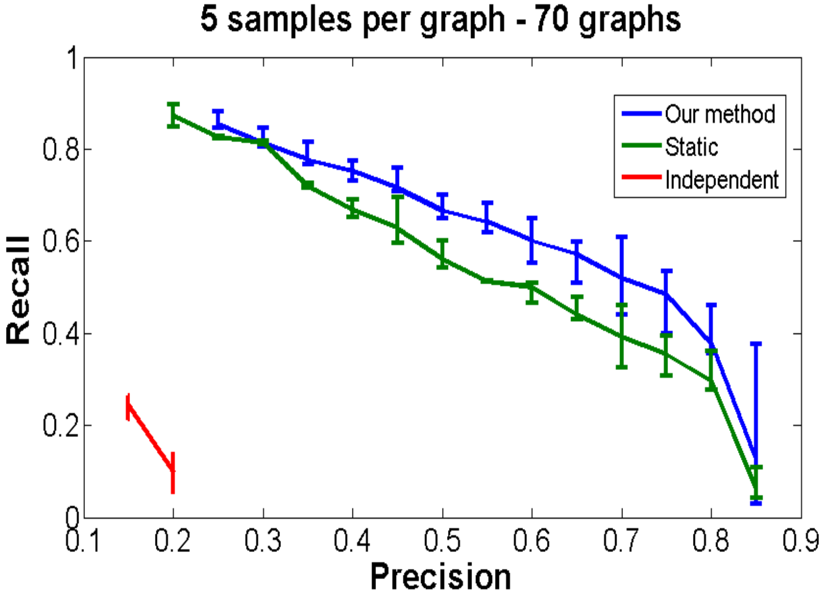
Randomly generate 70 graphs with the following genealogy.

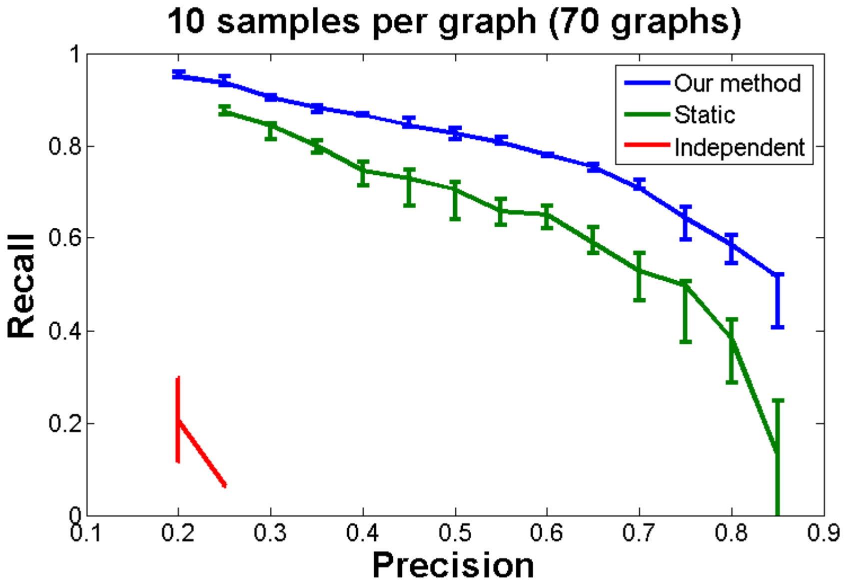
Branch points are when the true graph structure changes



The algorithm does **not** know a priori which graphs are the same and which aren't.

Simulation Results





Exploring the Progression and Reversion of Breast Cancer cells

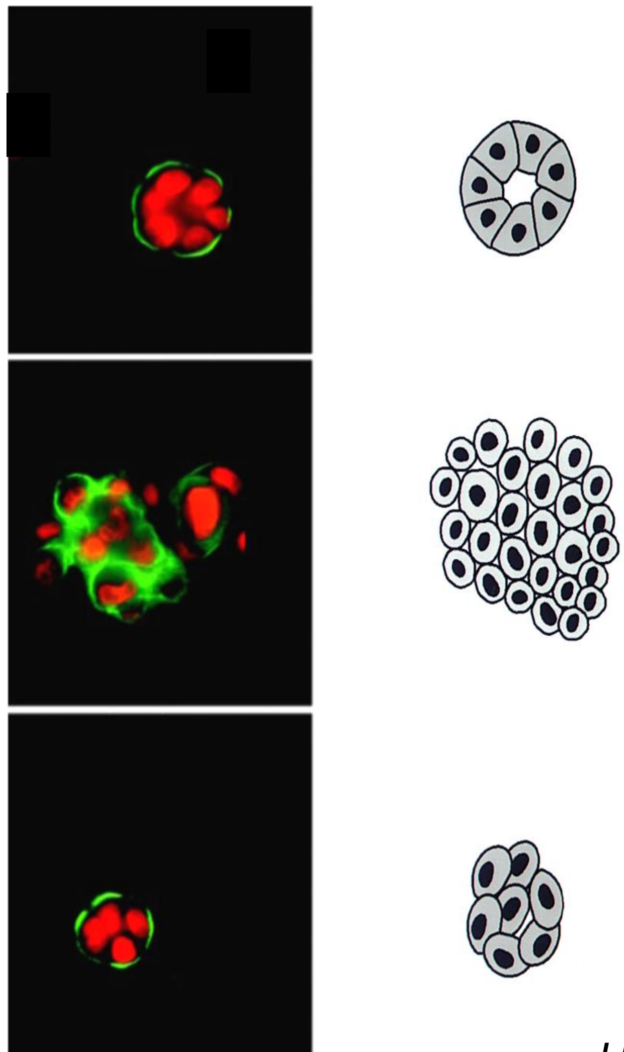
Breast Cancer Progression Series

S1

T4

inhibitors of signaling pathways

T4R

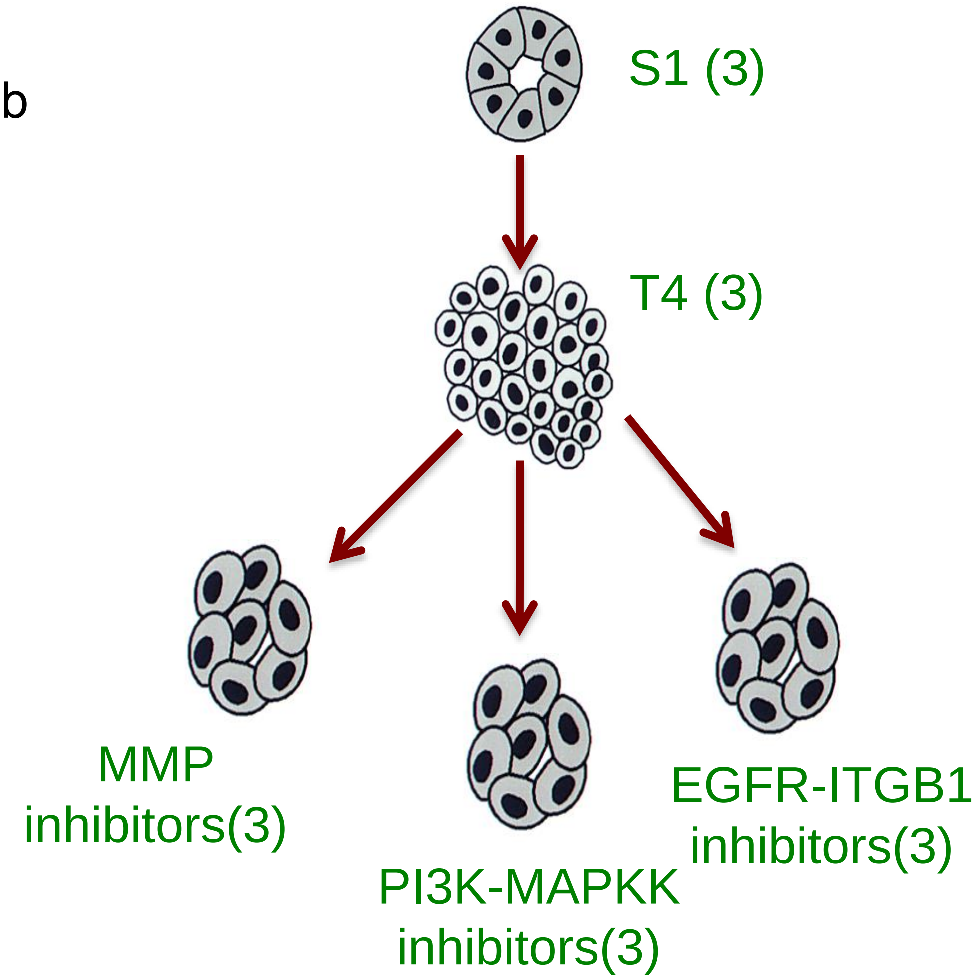


Dr. Mina Bissell, Berkeley

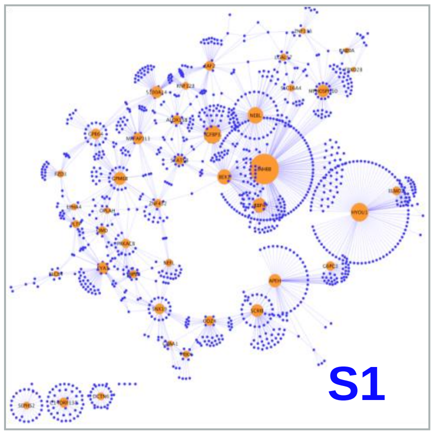
Hong, et. al. JCB 164(4): 603-612

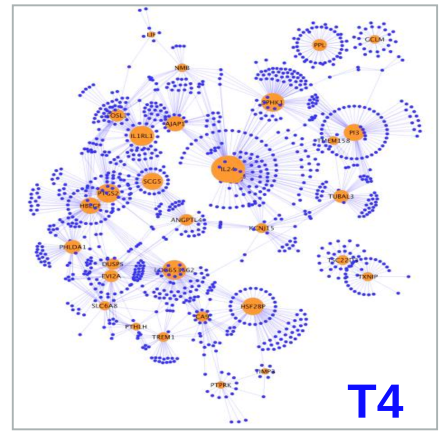
Microarray Dataset Details

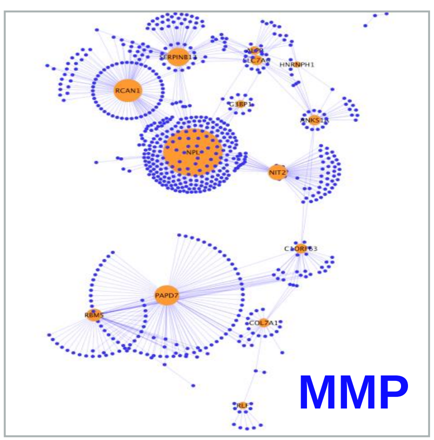
- Obtained from Dr. Mina Bissell's lab at LBNL
- Small sample size dataset (15 arrays in total)
- Merge data to increase the power of the network analysis
  (3 samples in each group)

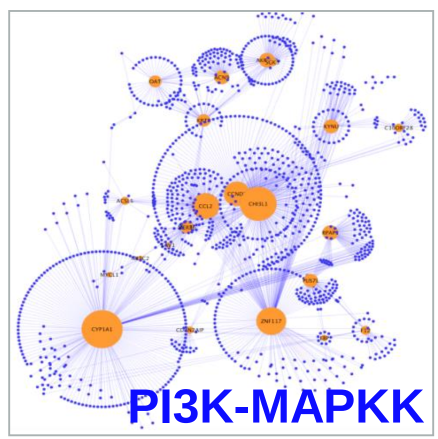


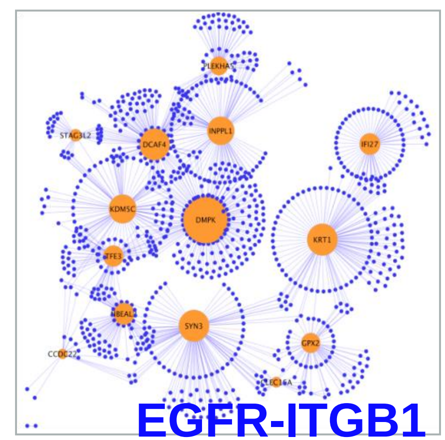
Results Overview



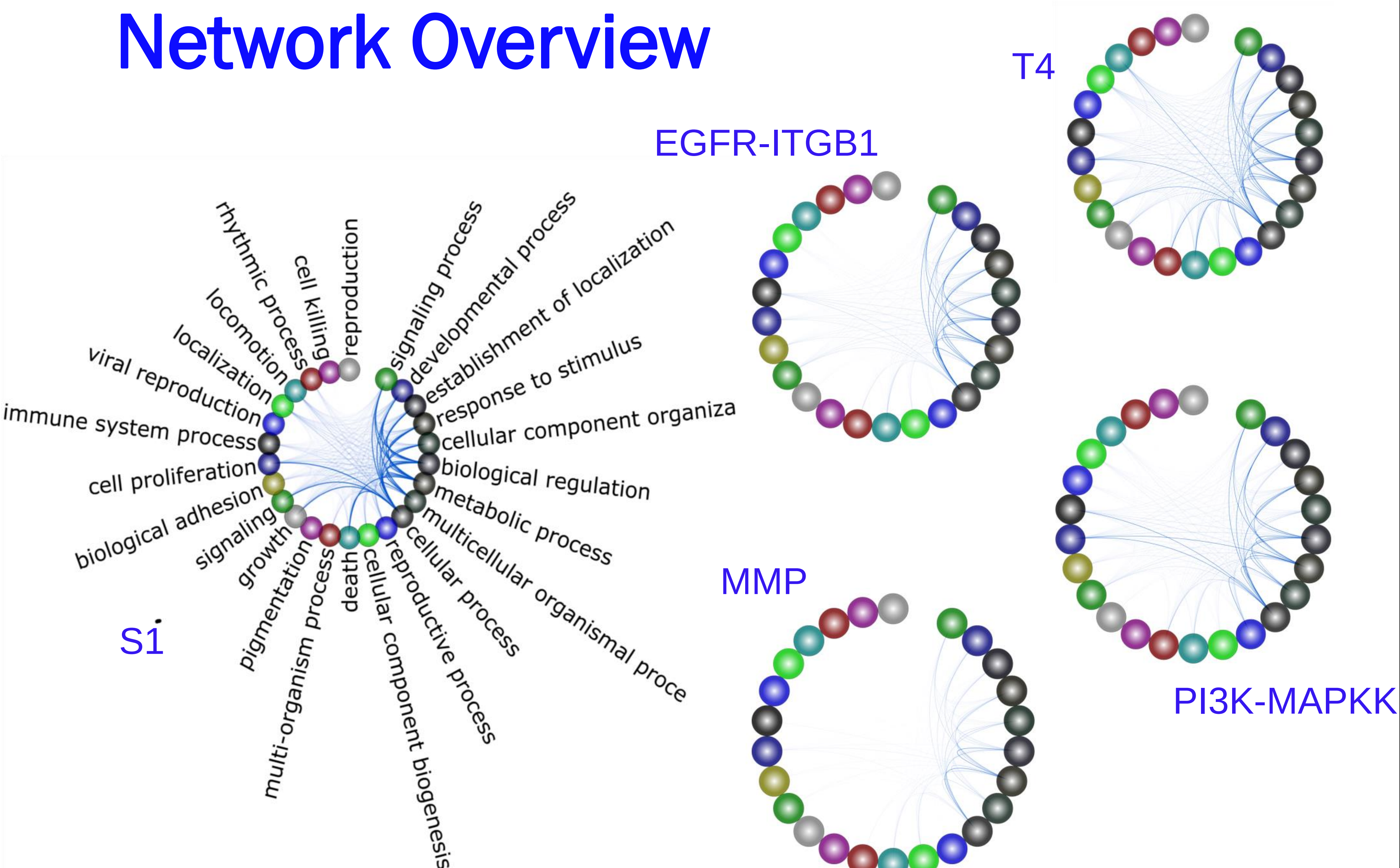






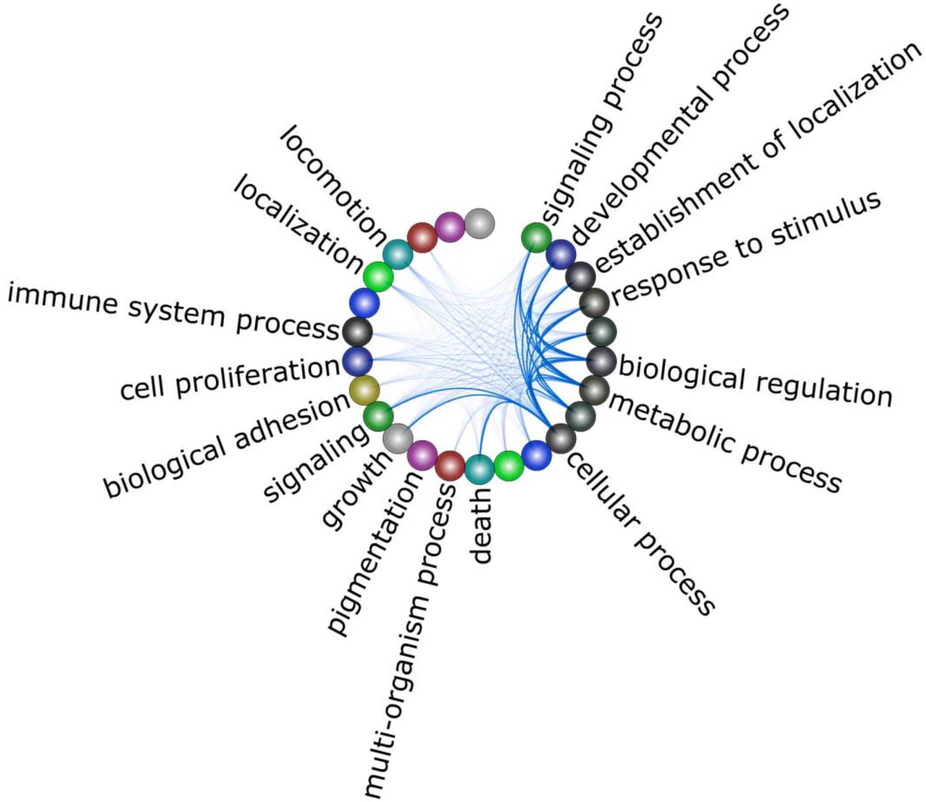


Network Overview

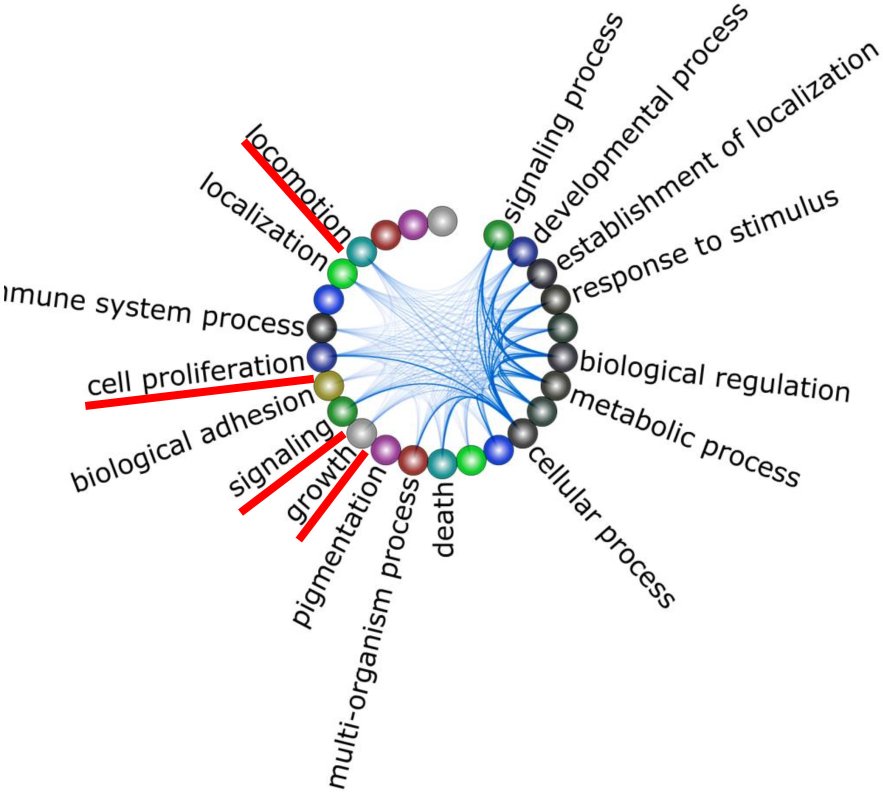


Interactions – Biological Processes

S1 cells

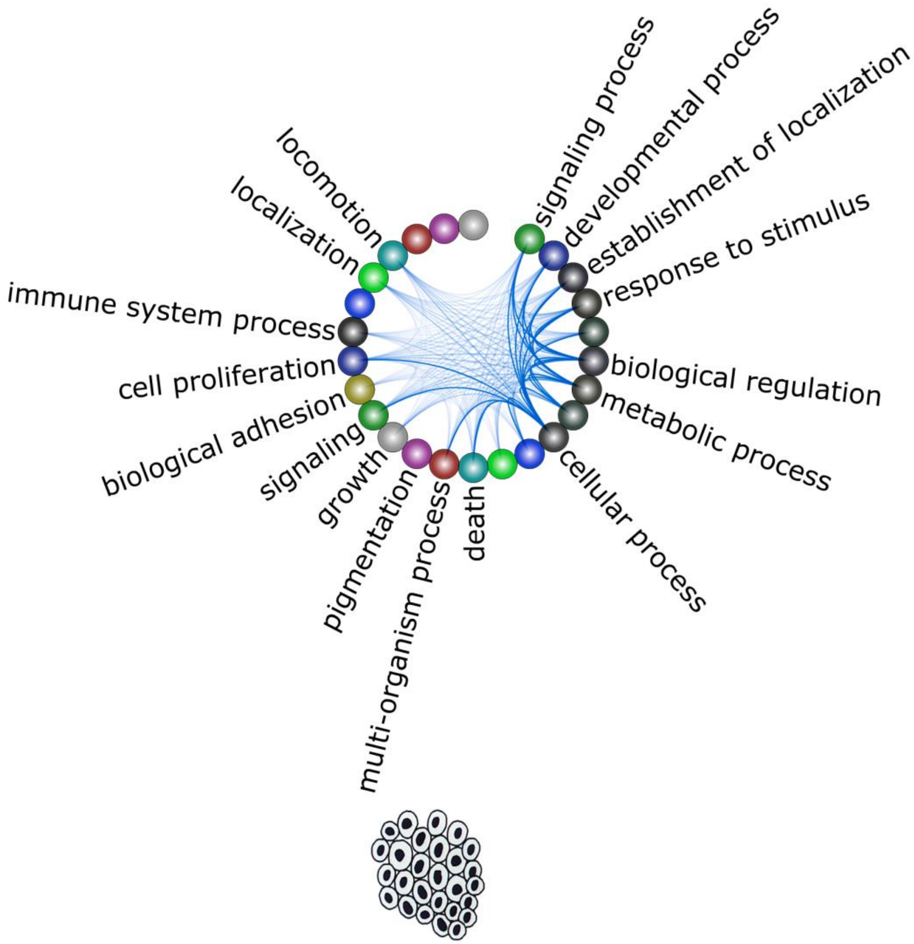


T4 cells: Increased Cell Proliferation, Growth, Signaling, Locomotion

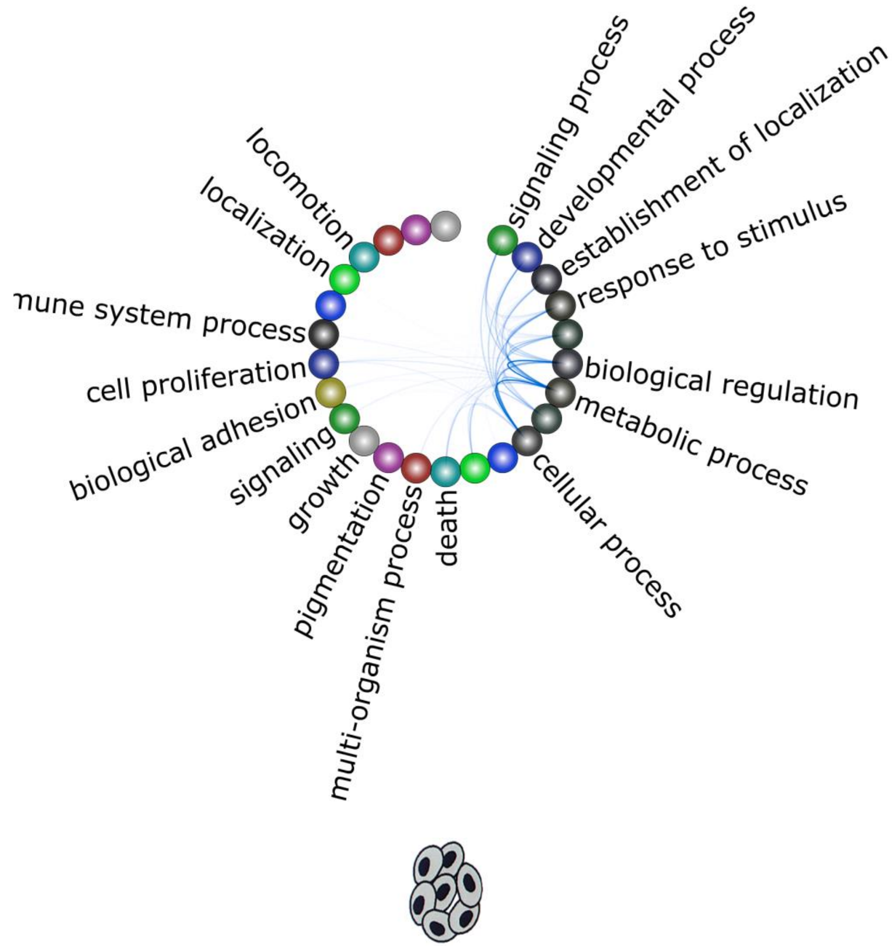


Interactions – Biological Processes

T4 cells



MMP-T4R cells: Significantly reduced interactions

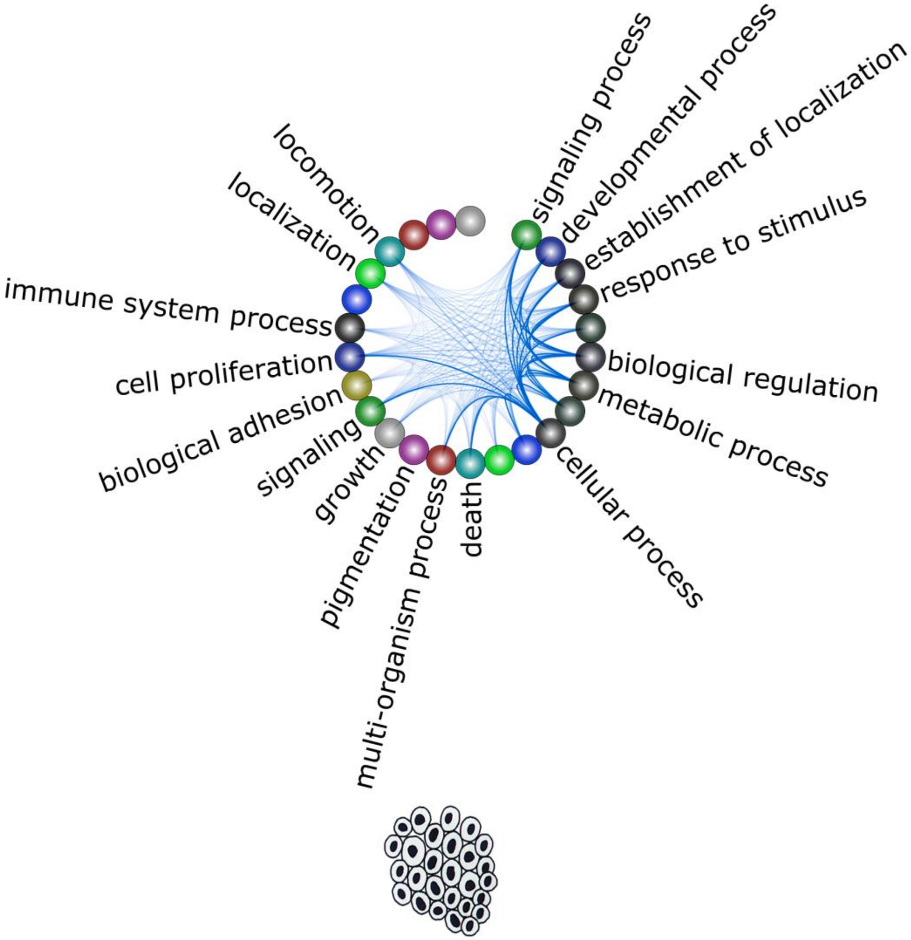


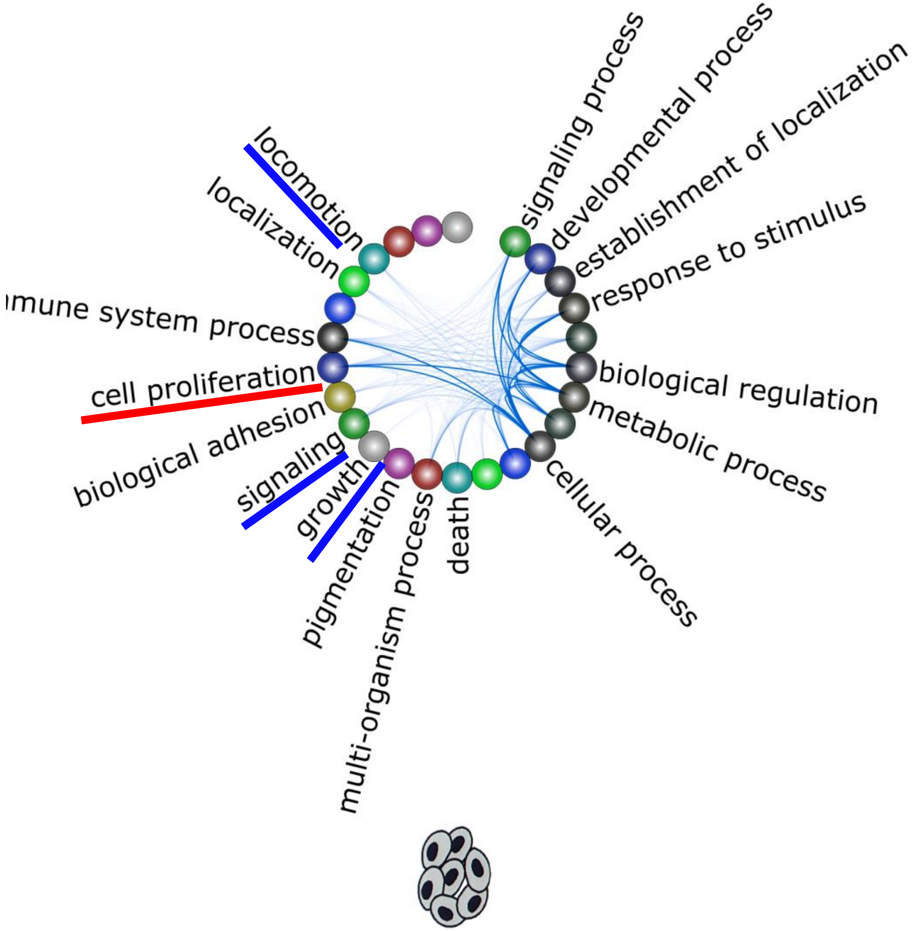
SAILING LAB
Laboratory for Statistical Artificial InteLligence & INtegrative Genomics

Interactions – Biological Processes

T4 cells

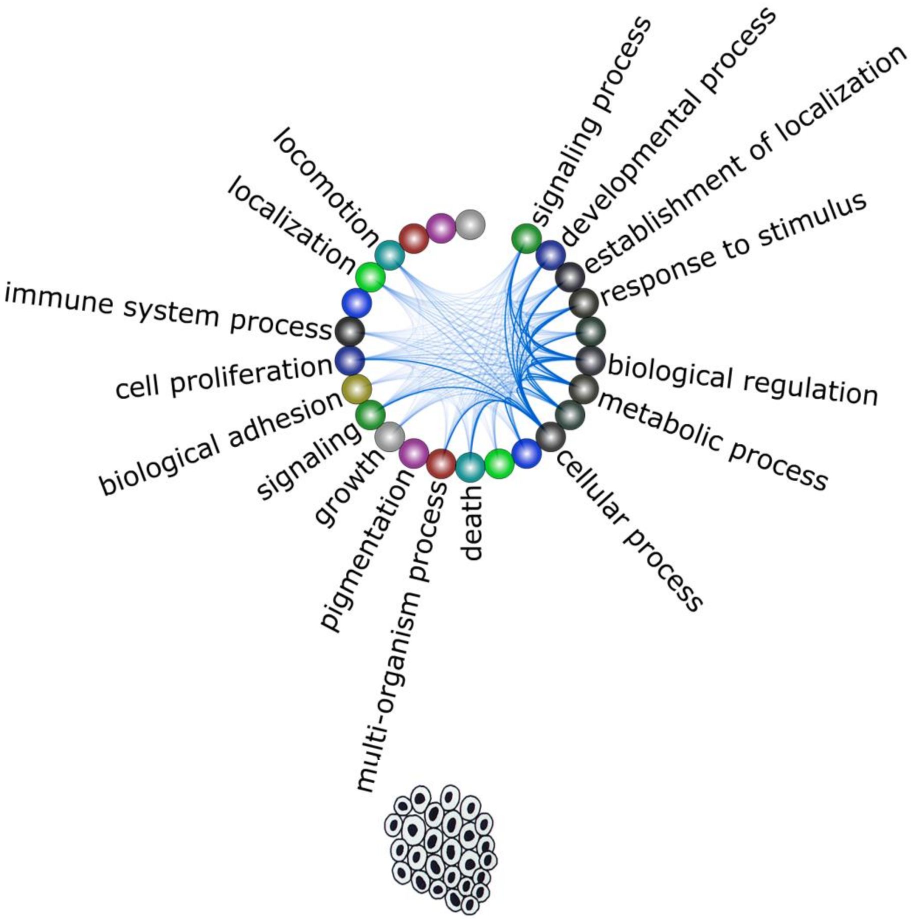
PI3K-MAPKK-T4R: Reduced Growth, Locomotion and Signaling



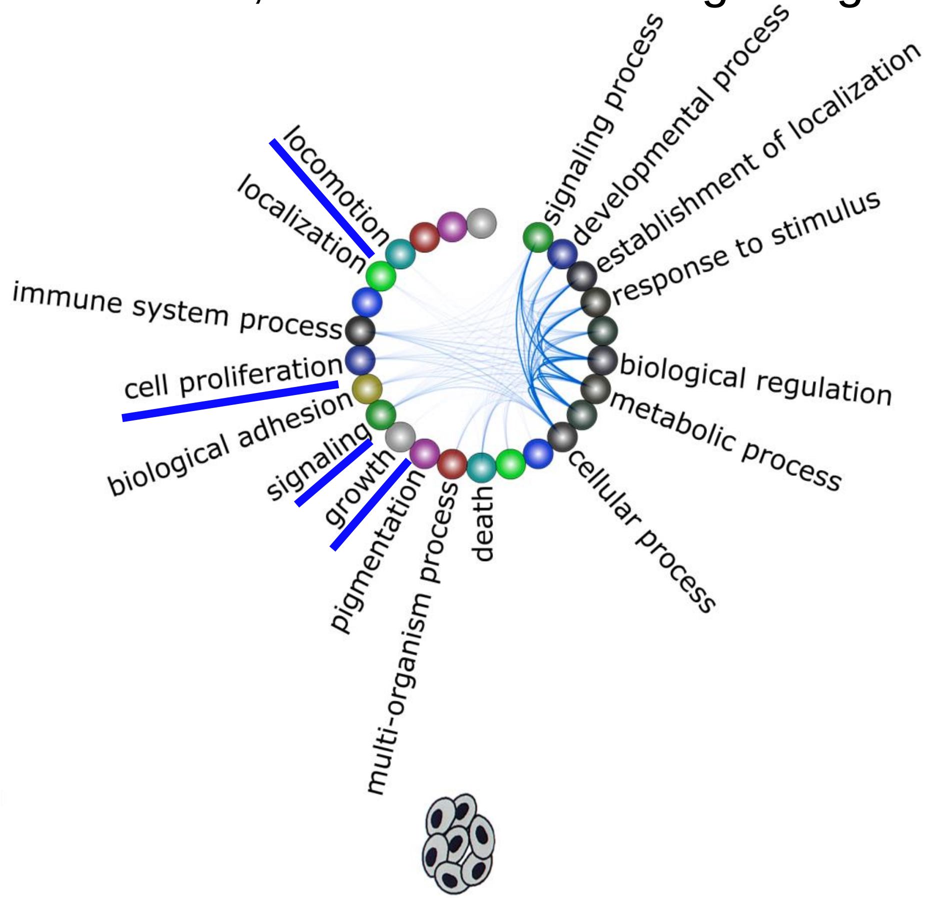


Interactions – Biological Processes

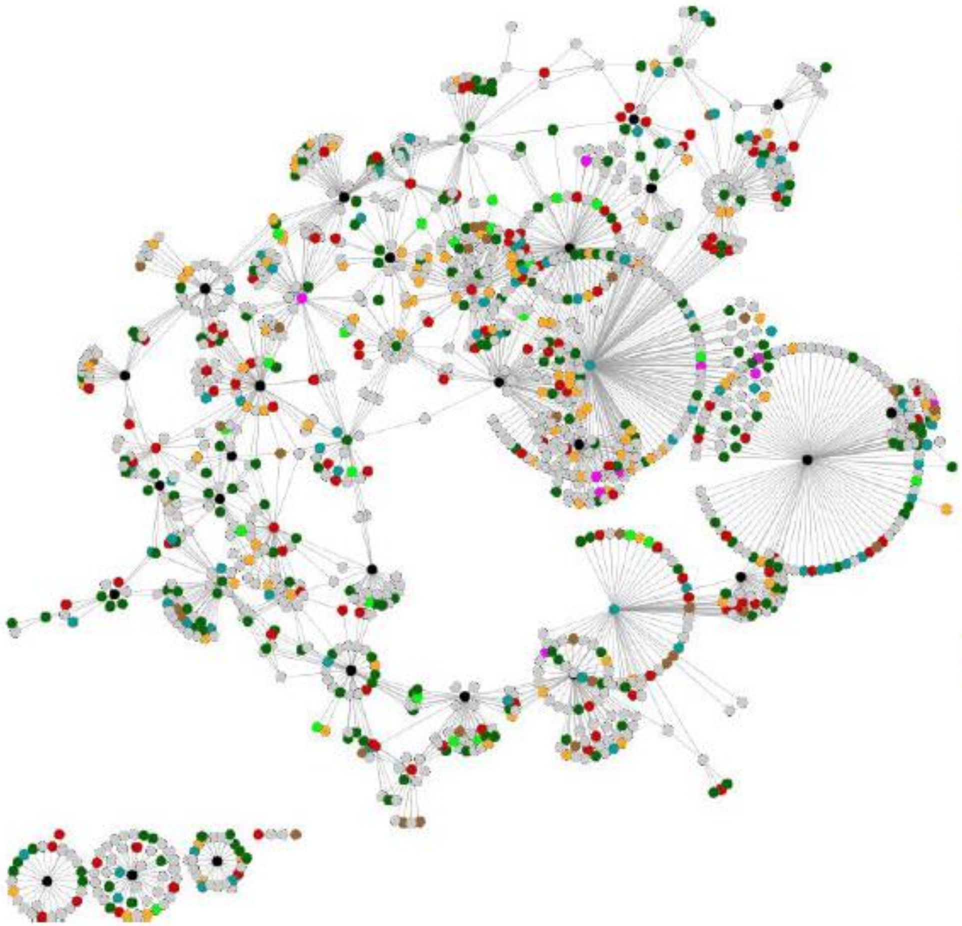
T4 cells



EGFR-ITGB1-T4R – Reduced Growth Proliferation, Locomotion and Signaling

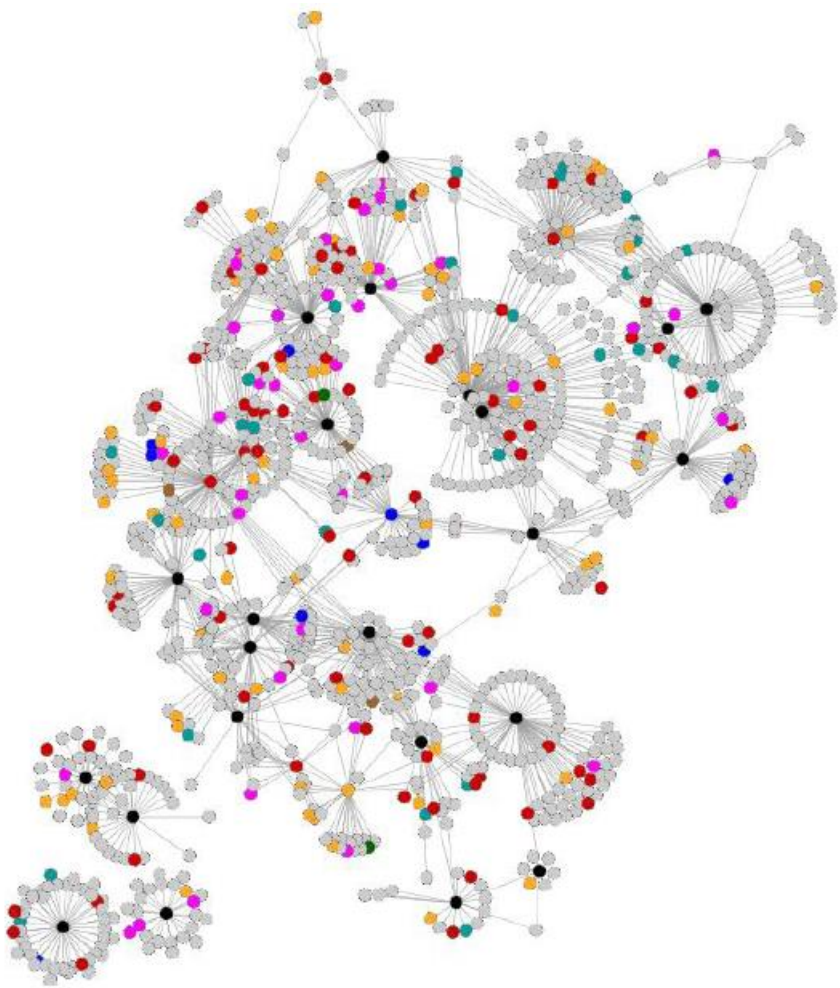


S1 Cells – GO Analysis



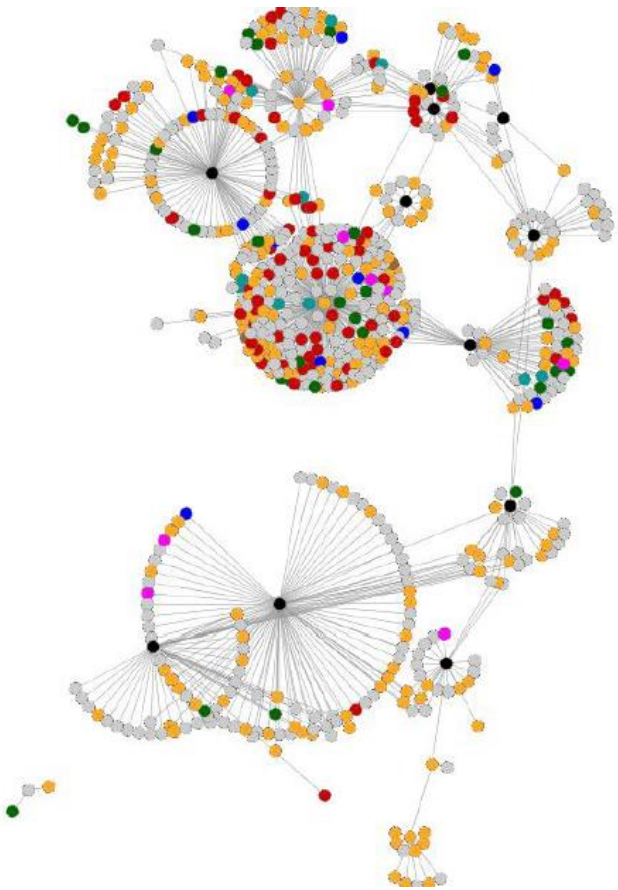
- biopolymer metabolic process
- cellular respiration
- RNA metabolic process
- mitochondrion
- ribosome
- DNA replication
- macromolecular metabolic process
- energy derivation by oxidation of organic compounds
- Hub

T4 cells – GO Analysis



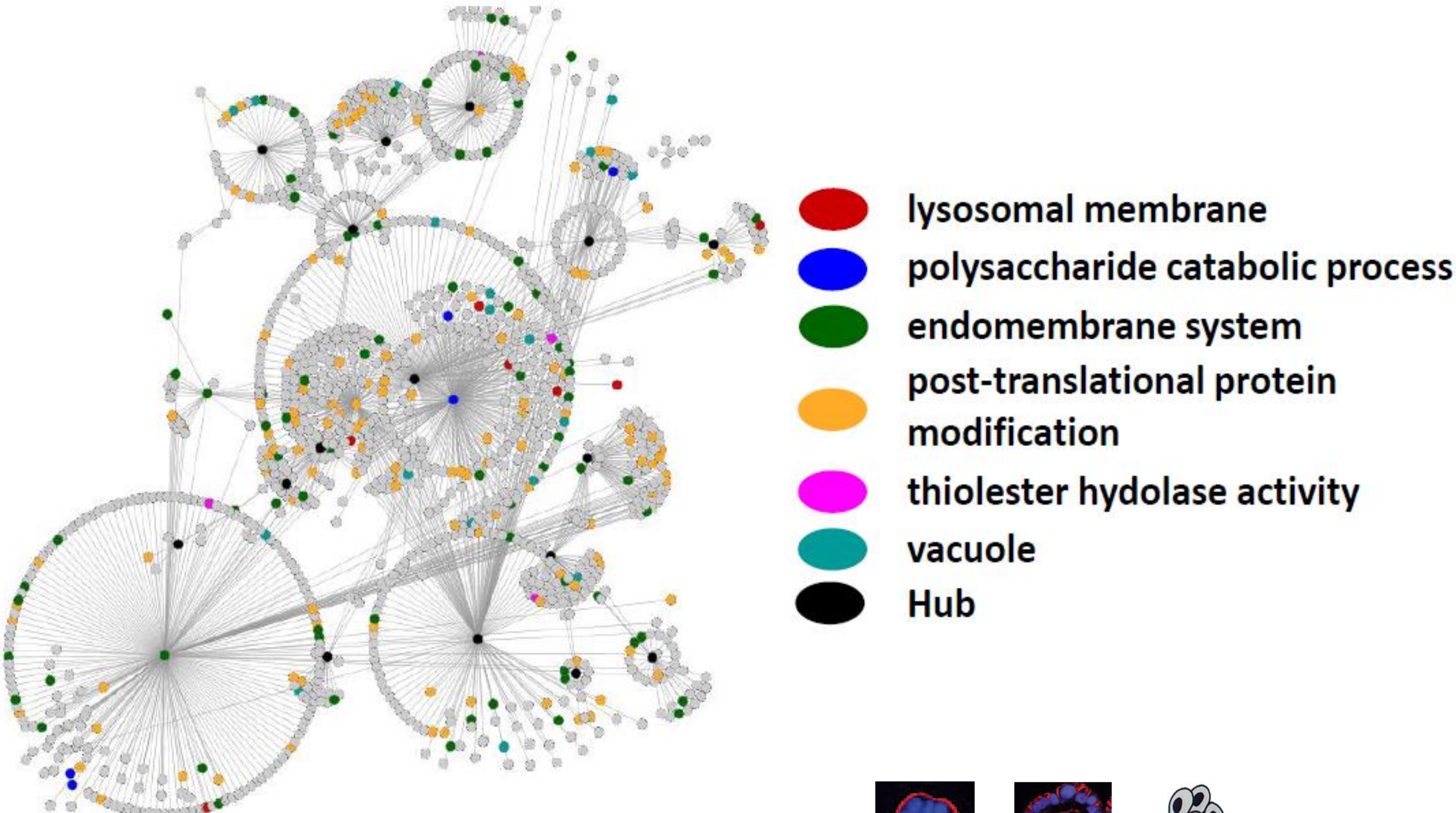
- cell proliferation
- angiogenesis
- blood vessel morphogenesis
- intracellular signaling cascade
- GTP binding
- actin binding
- growth factor activity
- Hub

MMP-T4R cells – GO Analysis

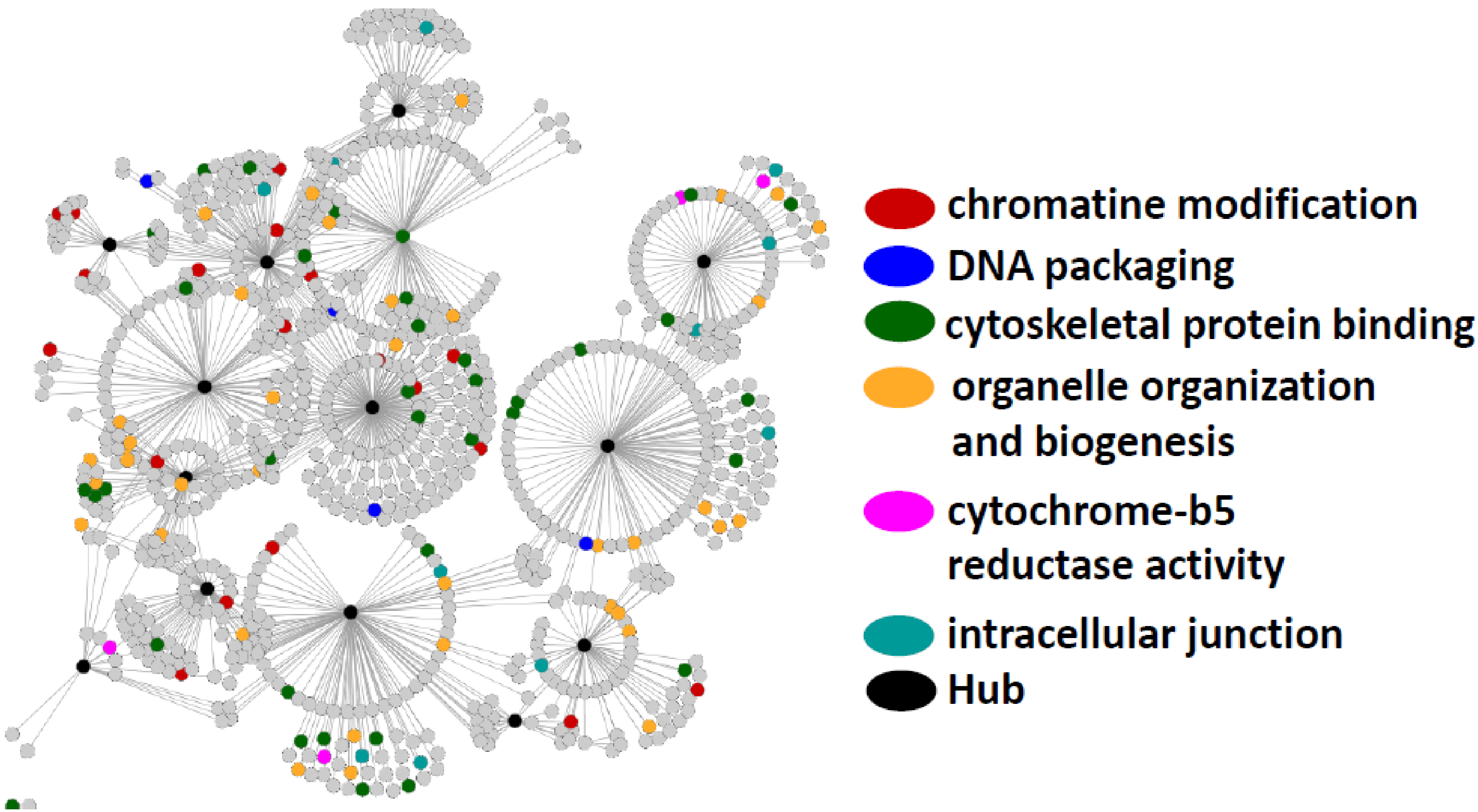


- mitochondrion
- fatty acid metabolic process
- membrane enclosed lumen
- primary metabolic process
- nuclear transport
- cofactor metabolic process
- oxidative phosphorylation
- Hub

PI3K-MAPKK-T4R cells – GO analysis

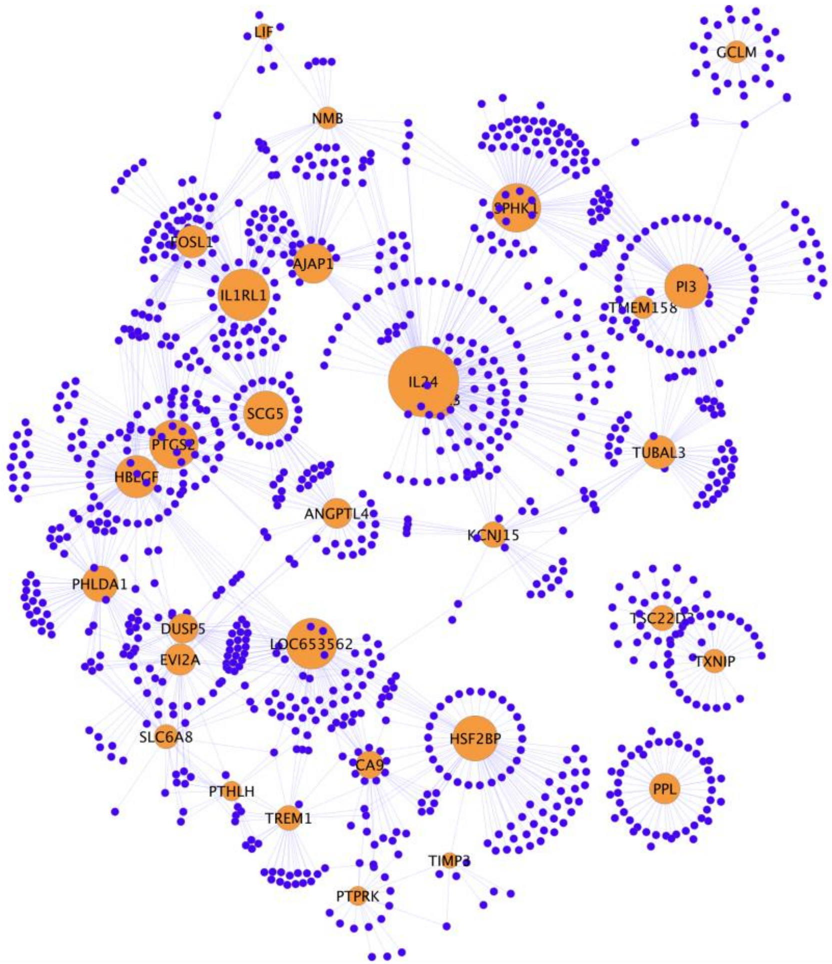


EGFR-ITGB1-T4R cells – GO Analysis

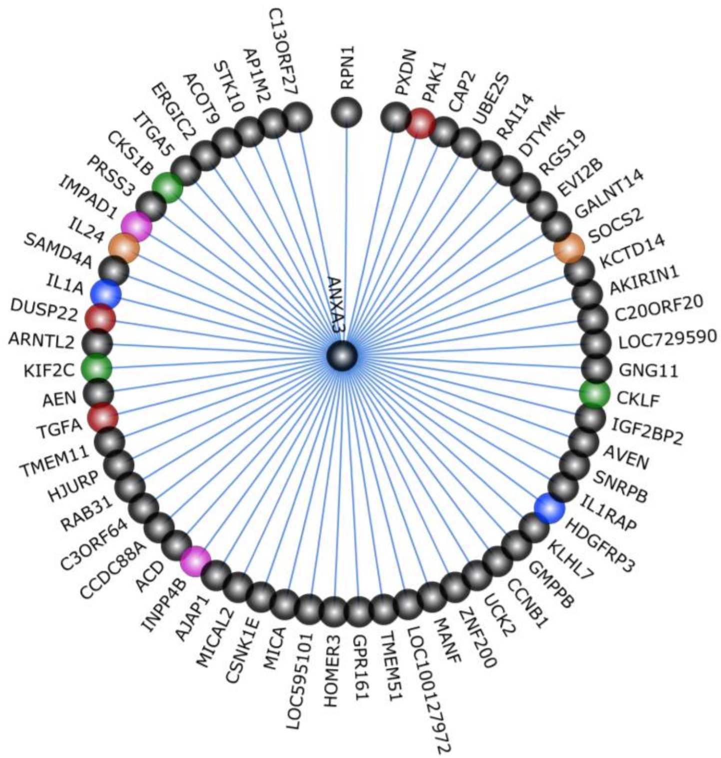


Identification of Potential Drug Targets

Hubs in T4 Network



ANXA3 Subnetwork

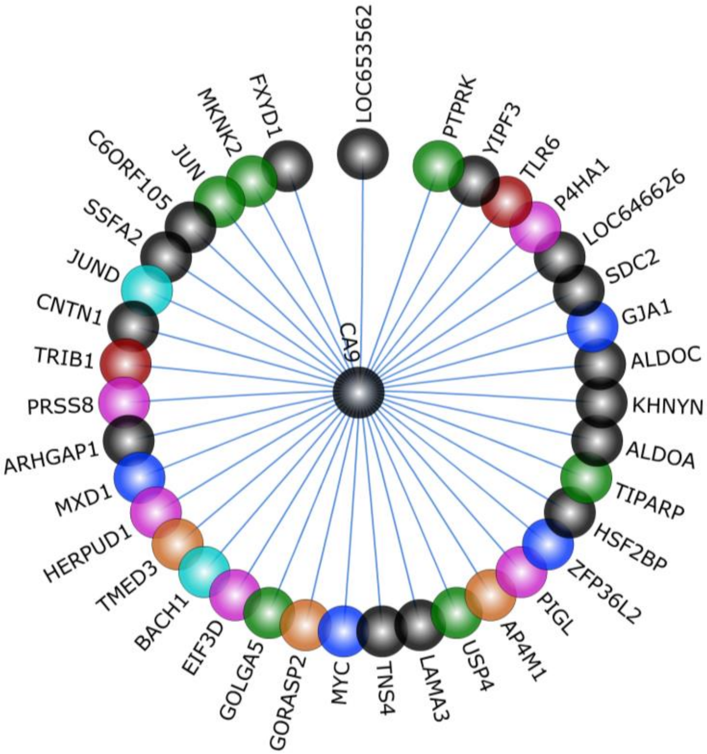


- **regulation of MAP kinase activity**
- **growth factor activity**
- **cell proliferation**
- **cytokine activity**
- **phosphoric monoester hydrolase activity**

Description: Encodes a protein belonging to the annexin family, and is known to play a role in the **regulation of cell growth** and is thought to be a **biomarker of cancer** (Jung et al., 2010).

Laboratory for Statistical Artificial InteLligence & INtegrative Genomics

CA9 Subnetwork



- regulation of MAP kinase activity
- cell proliferation
- post-translational protein modification
- golgi apparatus part
- protein metabolic process
- transcription factor activity

Description: Encodes carbonic anhydrase IX. It has been implicated in **cell proliferation**, and **renal cell carcinoma** (Jubb et al., 2004).

Conclusion

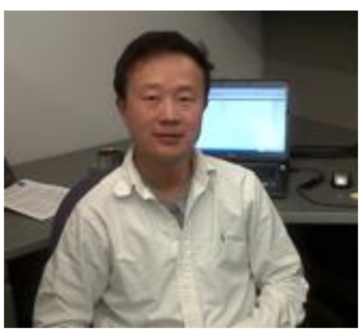
- We present a method to learn a collection of networks over a genealogy.
  - This allows us to efficiently integrate information across samples while still exposing sharp differences
- We perform an analysis of breast cancer cells using our algorithm.
  - Functional analysis shows that our method is producing biologically valid results.
  - Our approach may help biologists better decipher networks specific to various breast cancer cells
  - Thus providing better treatment for personalized medicine

Acknowledgements

Dr. Mina Bissell, Berkeley

ISMB Travel Fellowship 2011

NSF



Dr. Ren Xu, Kentucky

NIH

Alfred P. Sloan Fellowship to EPX