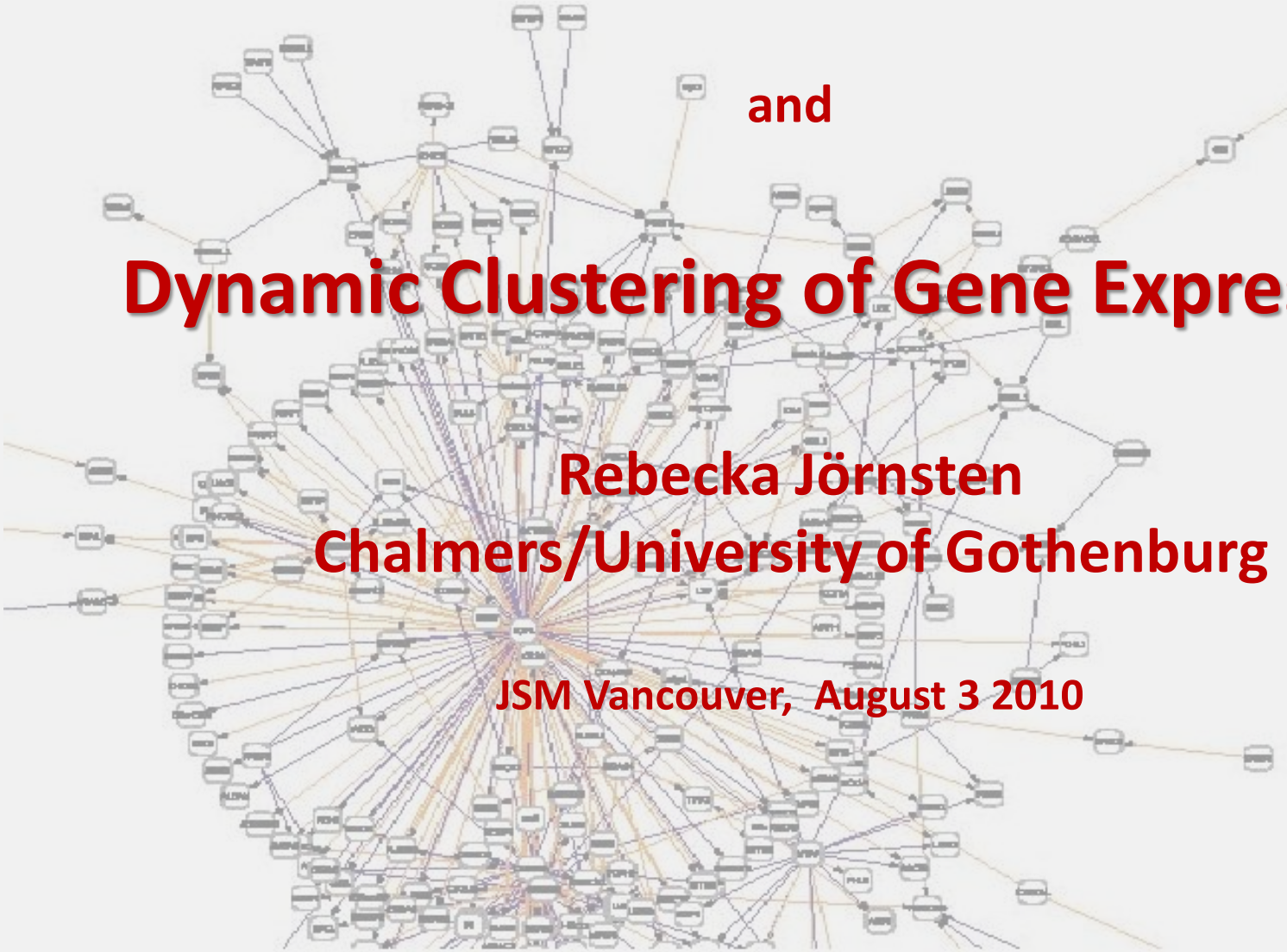
Endogenous Perturbation Analysis of Cancer



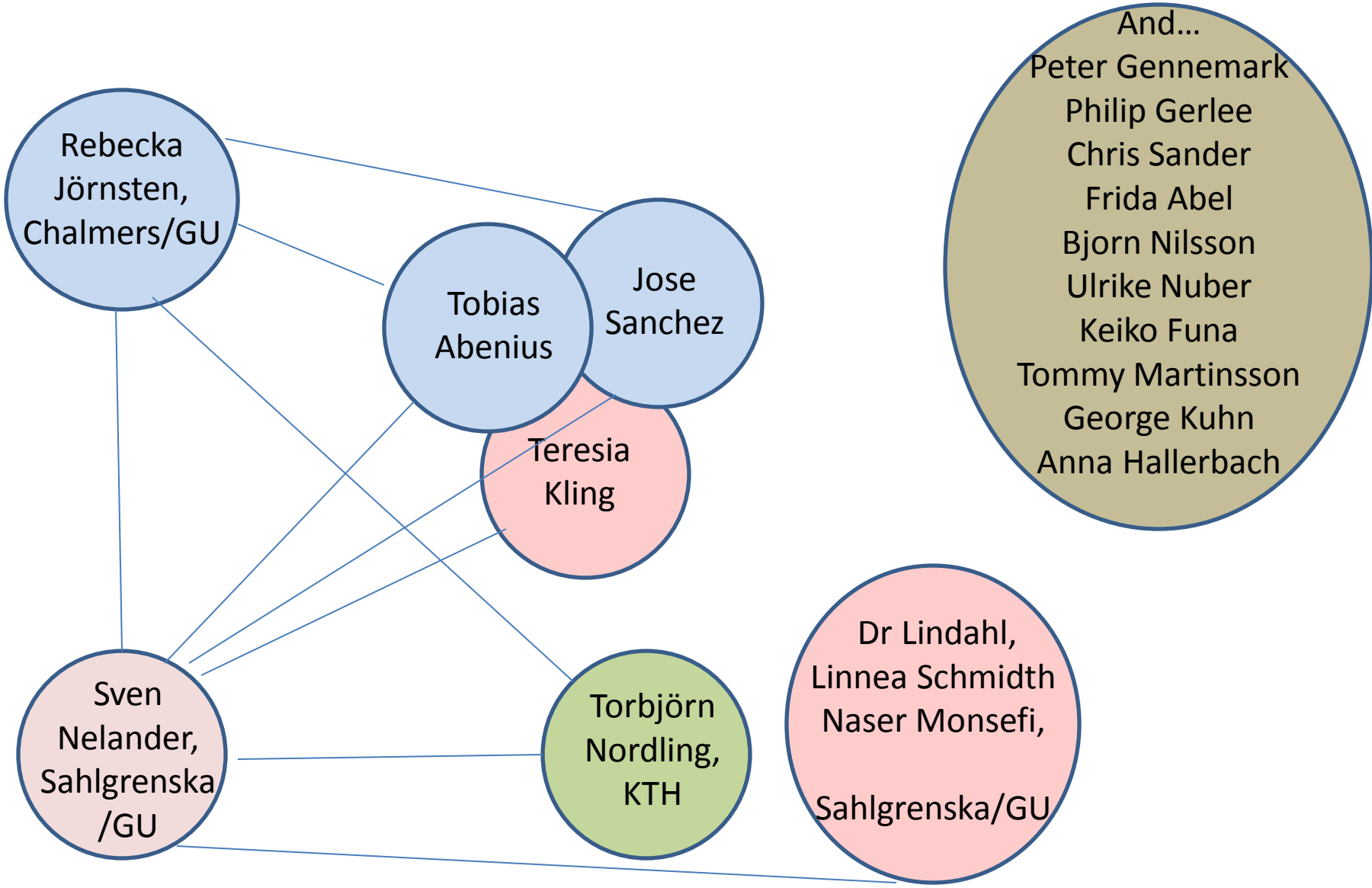
and

Dynamic Clustering of Gene Expression

Rebecka Jörnsten
Chalmers/University of Gothenburg

JSM Vancouver, August 3 2010

Some ongoing projects with ….



Part 1: Endogeous Perturbation Analysis of Cancer

Goals: * Develop computational tools to construct predictive models of cancer pathways.
* Predict key regulators & downstream targets

Network and pathway modeling:

•Experimental approach: knock out individual genes /transcripts & observe how the system responds.

CON: $$$, limited data, PRO: precise Q investigated

•Reverse engineering: uses no perturbations, e.g. network construction based on (partial)correlation of gene expression.

CON: limited information, PRO: lots of data available

Network and pathway modeling:

•Here, we view each tumor's

DNA copy number aberration (CNA) profile as a system perturbation

that simultaneously affects multiple genes,

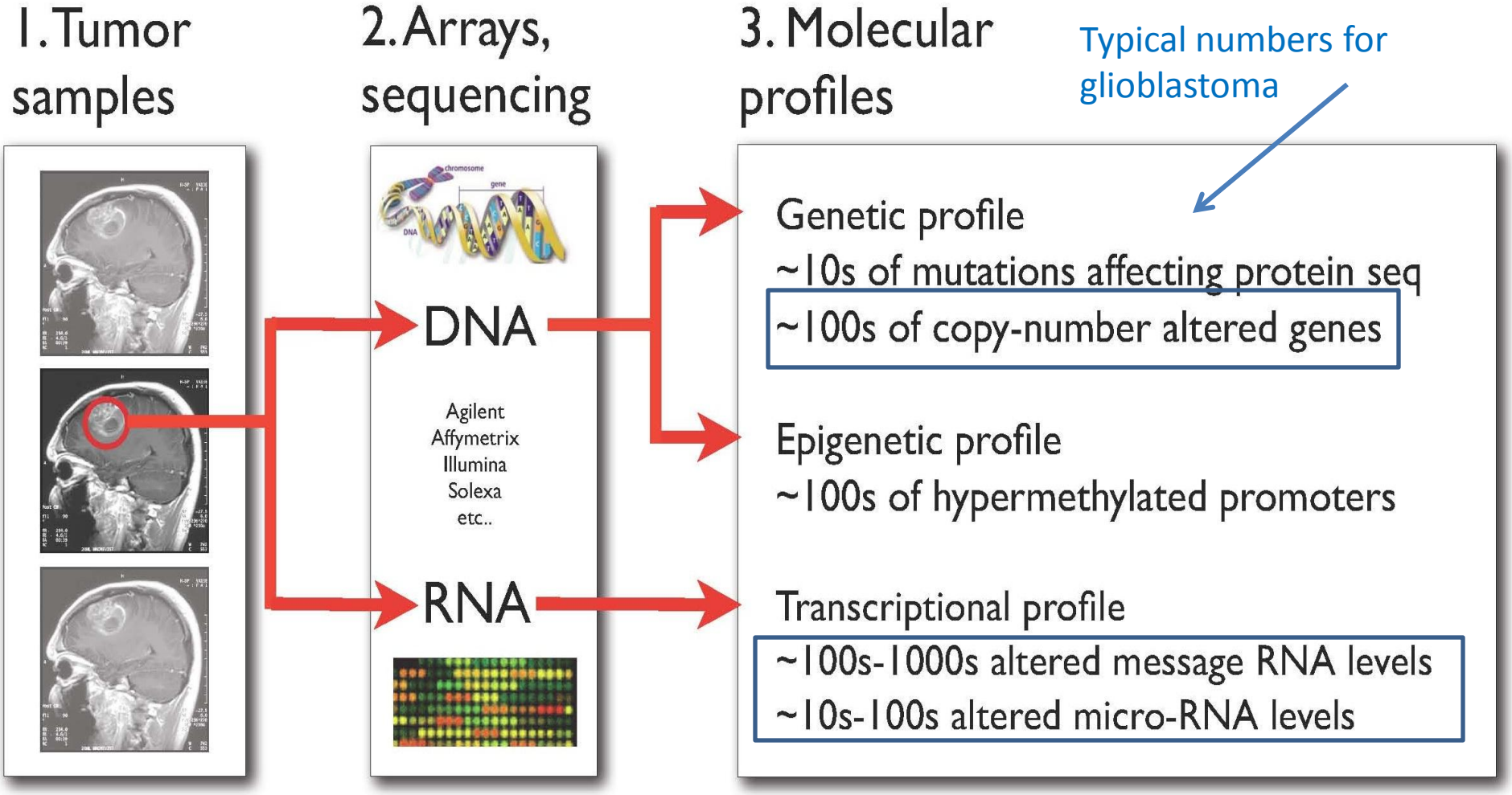
•and the corresponding

mRNA profile as the steady-state response to that perturbation.

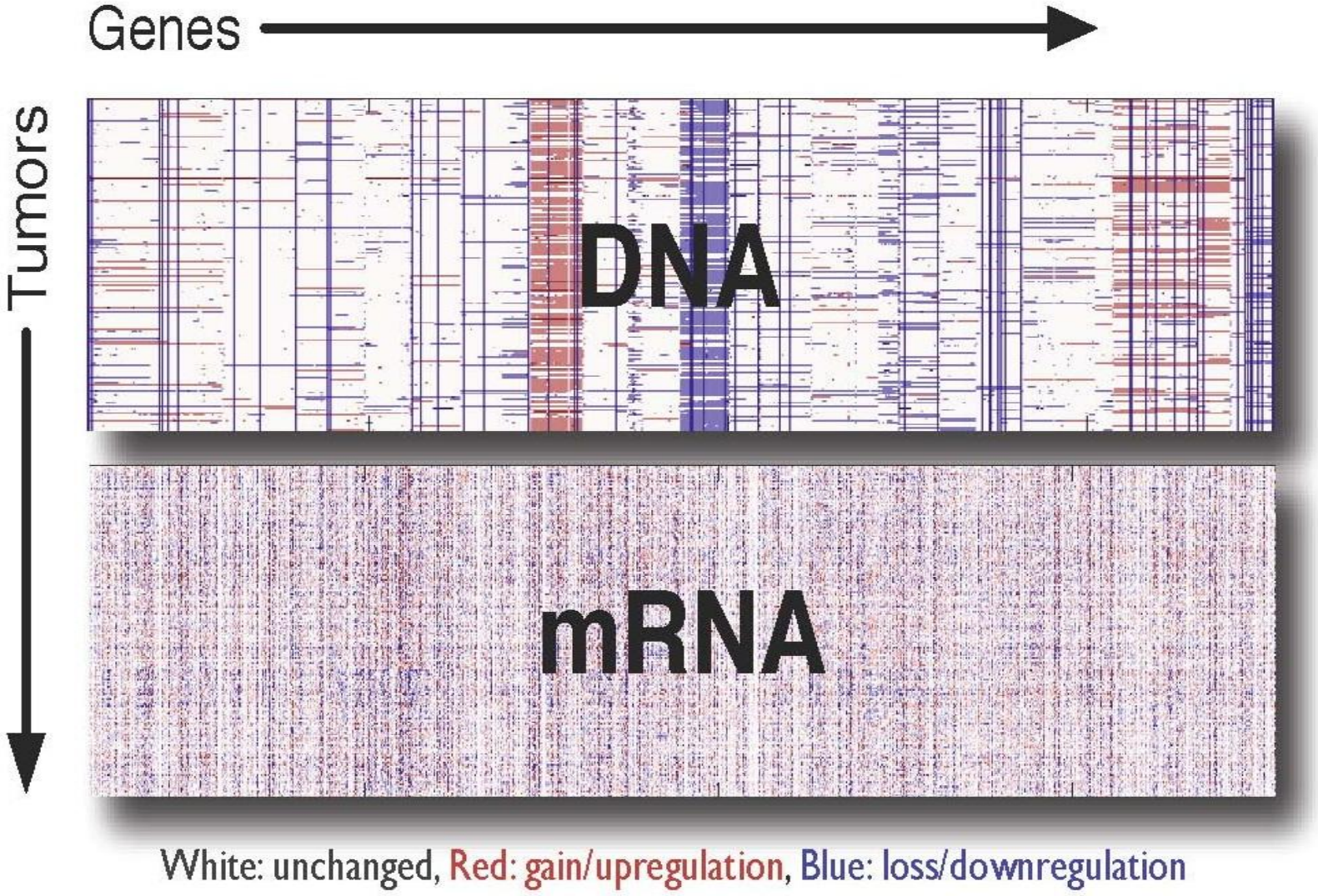
PRO: large-scale and high-quality experimental CNA and mRNA expression data available.

CON: we don't get to choose the perturbations.

Data: CNA and mRNA data from 180 glioblastoma tumors from the Cancer Genome Atlas Consortium (MSKCC, Broad, and other centers).



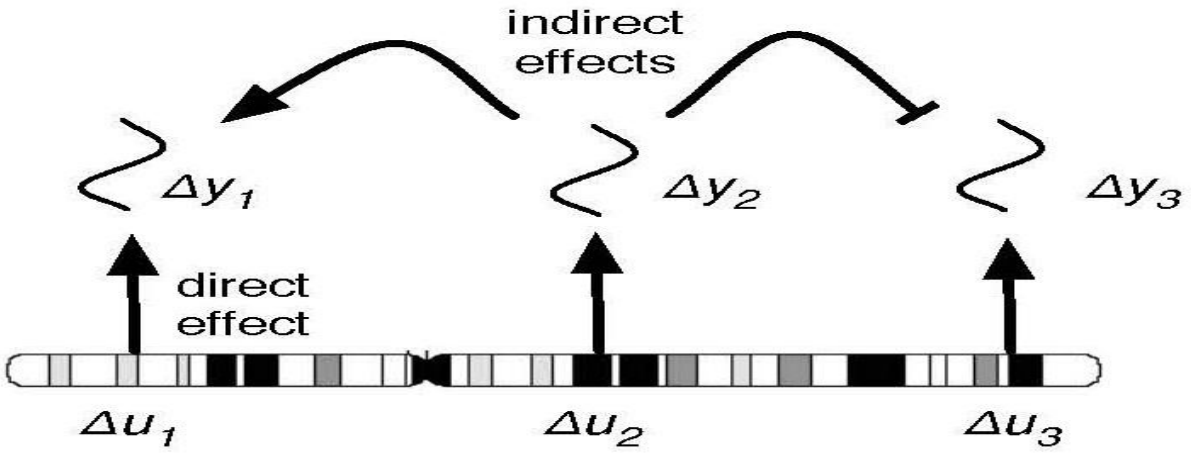
Data: CNA and mRNA data from 180 glioblastoma tumors from the Cancer Genome Atlas Consortium (MSKCC, Broad, and other centers).



Model with multiplicative effects

$$\frac{dy_i}{dt} = u_i \alpha_i \prod_j y_j^{w_{ij}} - \beta_i \prod_j y_j^{v_{ij}}$$

- where y_i is the mRNA expression of transcript i, and u_i the copy number.
- Elements w_{ij} and v_{ij} denote the effect of transcript j on i during synthesis/degradation respetively.
- The α_i and β_i denote "environmental effects (non-CNA perturbations)"



Steady-state solution

- We take CNA profiles from two tumors, u and $\tilde{u}$, and likewise for mRNA y, $\tilde{y}$.
- Define $\Delta u_i = \log(u_i) - \log(\tilde{u}_i)$, $\Delta y_i = \log(y_i) - \log(\tilde{y}_i)$
- At steady-state we can write:

$$\Delta u_i + \sum_j (w_{ij} - v_{ij}) \Delta y_i +$$

$$+(\log(\alpha_i) - \log(\tilde{\alpha}_i)) - (\log(\beta_i) - \log(\tilde{\beta}_i)) = 0$$

- We denote the *direct* causal influence of transcript j on i by $a_{ij} = w_{ij} - v_{ij}$.
- The term involving αs and βs by γ_i.

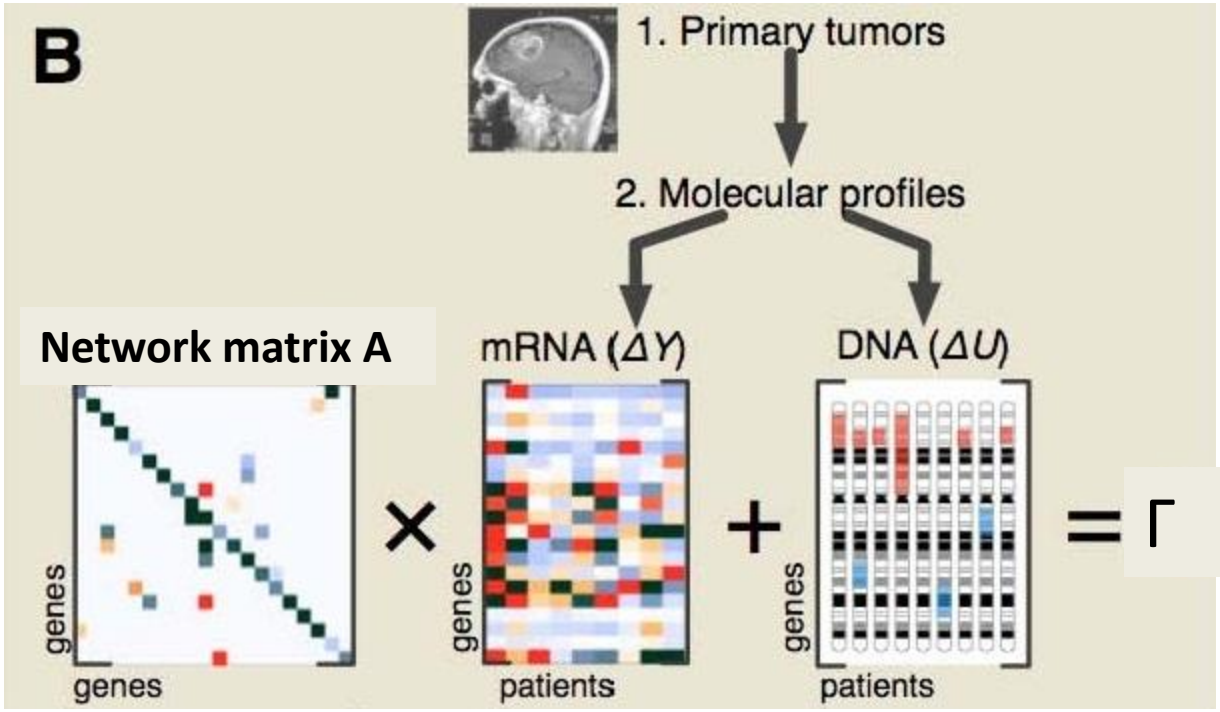
For the whole system we can write

$$A\Delta\mathbf{y} + \Delta\mathbf{u} + \Gamma = 0$$

Steady-state solution

$$A\Delta\mathbf{y} + \Delta\mathbf{u} + \Gamma = 0$$

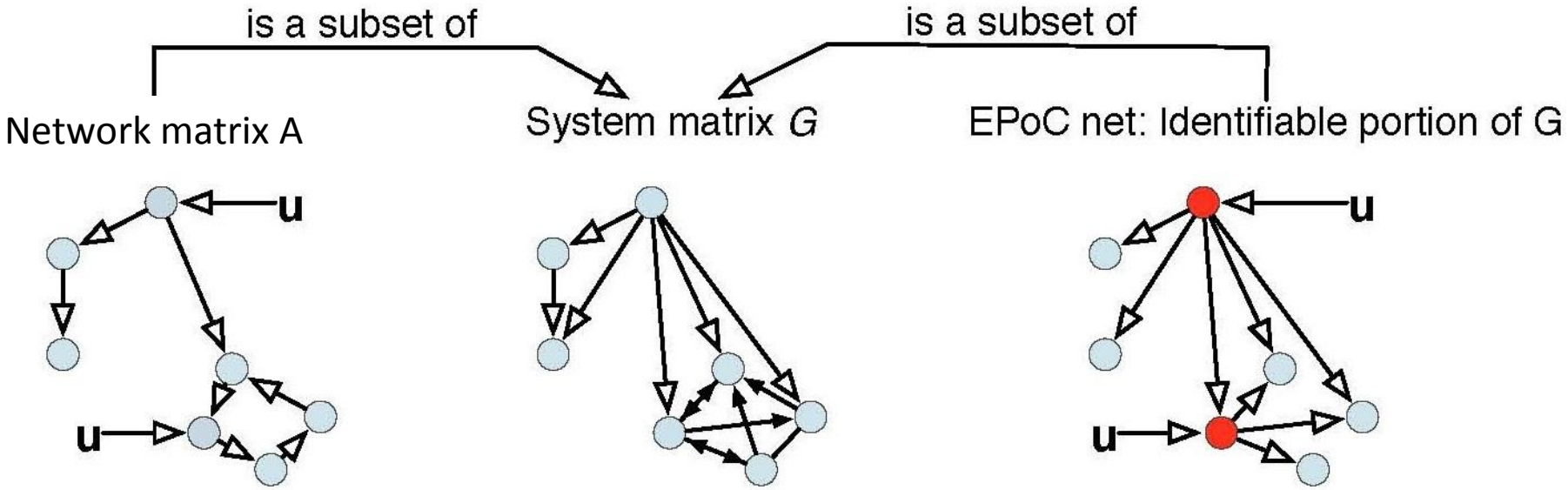
- A is called the *network matrix*, our parameters of interest.
- Γ includes all the tumor-specific differences, unmeasured environmental effects, SNPs etc.



Alternative Representation

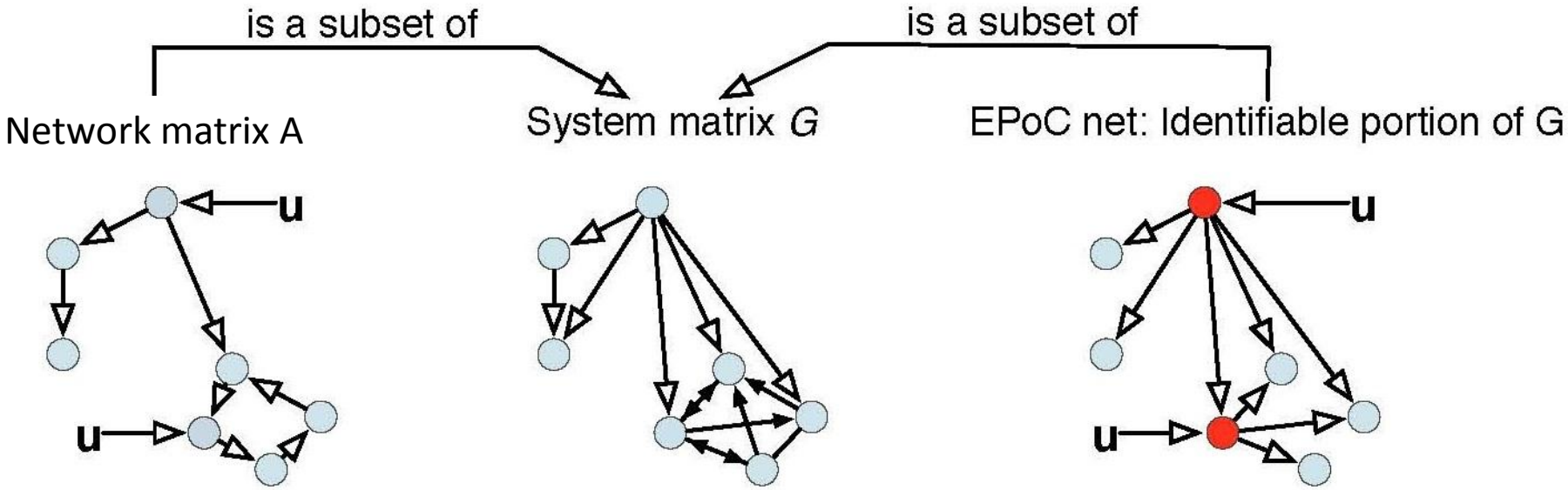
$$\Delta \mathbf{y} = G\Delta \mathbf{u} + \Gamma'$$

- $G = -A^{-1}$ is called the *system matrix*.
- G includes both direct and indirect effects, but this representation is more in line with the biological dogma of causal order AND in this formulation the input is much less correlated which helps with estimation.



We tried estimating A directly vs G.

Estimation of G leads to more stable and reproducible results.



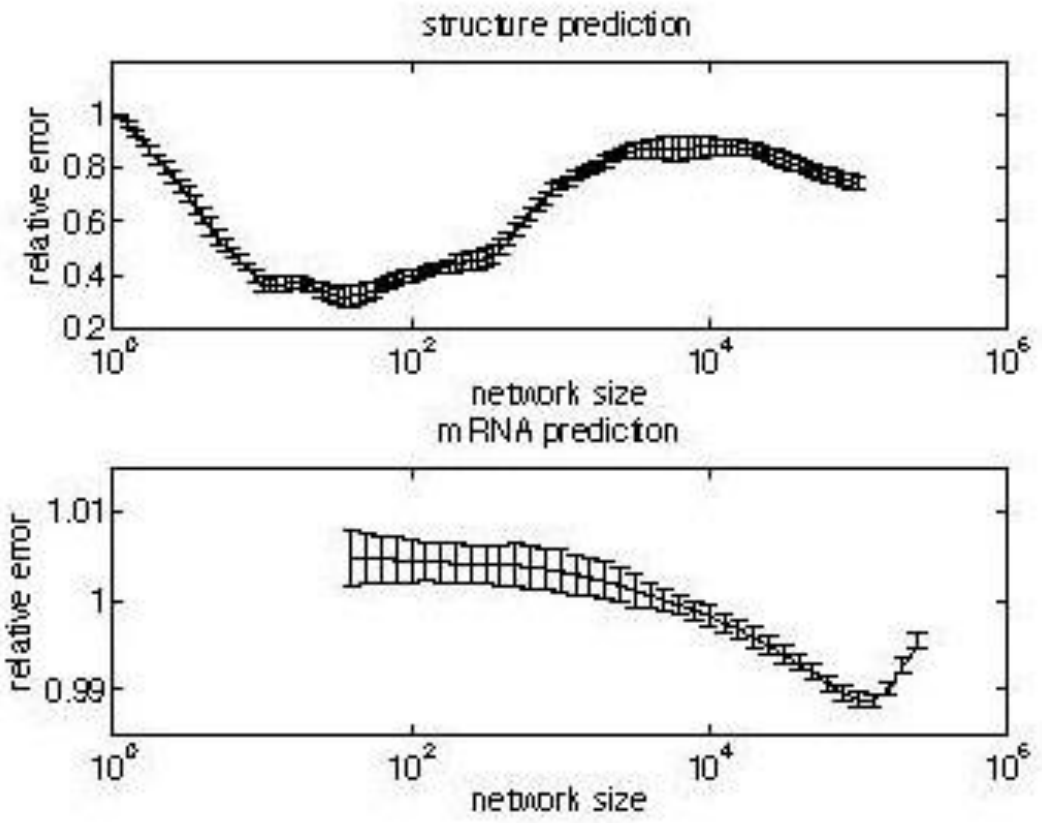
Estimating G -> A.

- We use an L1-penalty and estimate the rows of G

- Penalty parameter? We investigate two goals:

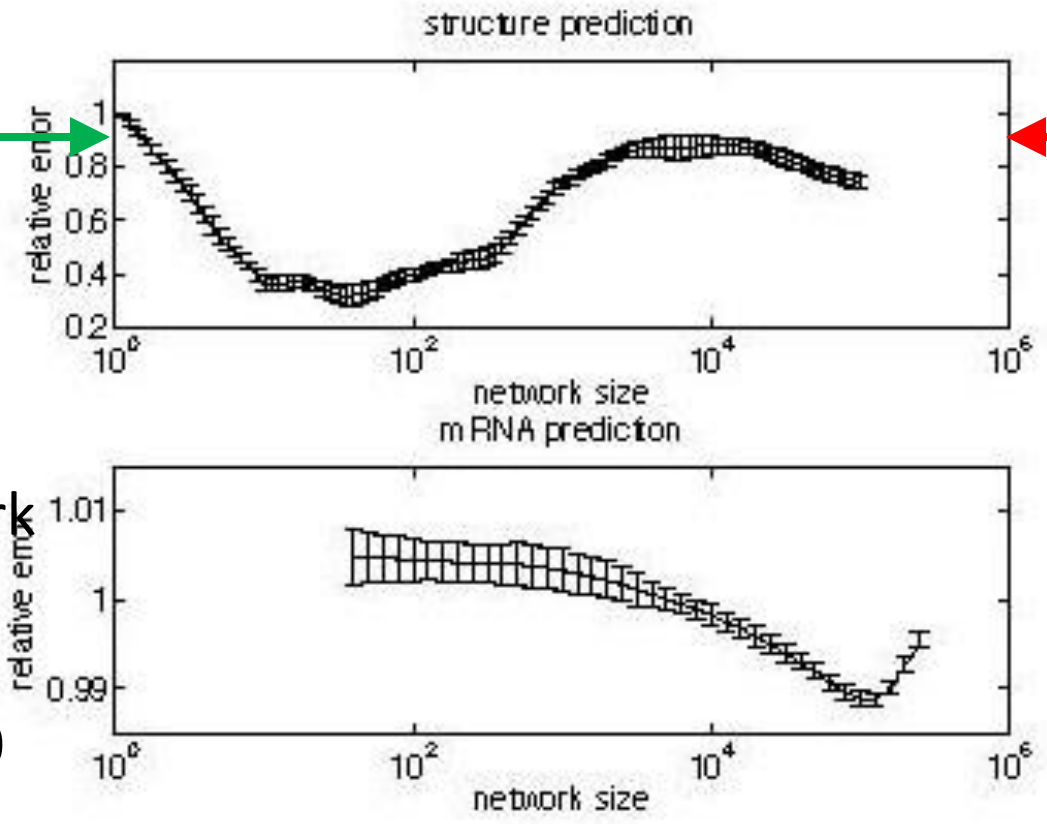
  * network reproducibility

  * predictive power



Estimating G -> A.

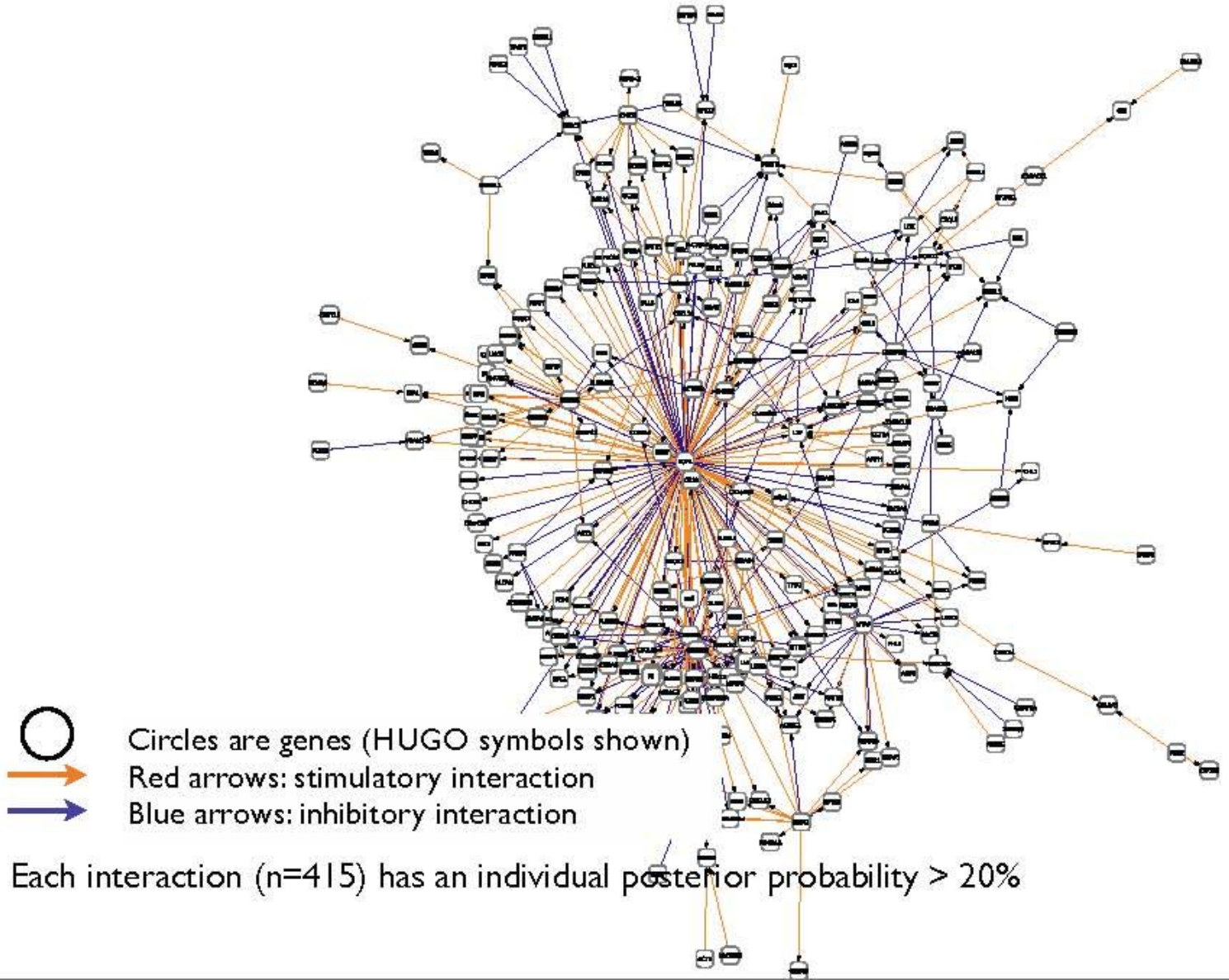
To choose the penalty parameter, we resample tumor pairs repeatedly and investigate
a) network agreement, b) mRNA prediction.



For structure prediction – stable network with ~500 interactions among 10000 genes

For mRNA prediction – network with a large number of interactions (~3/per gene).

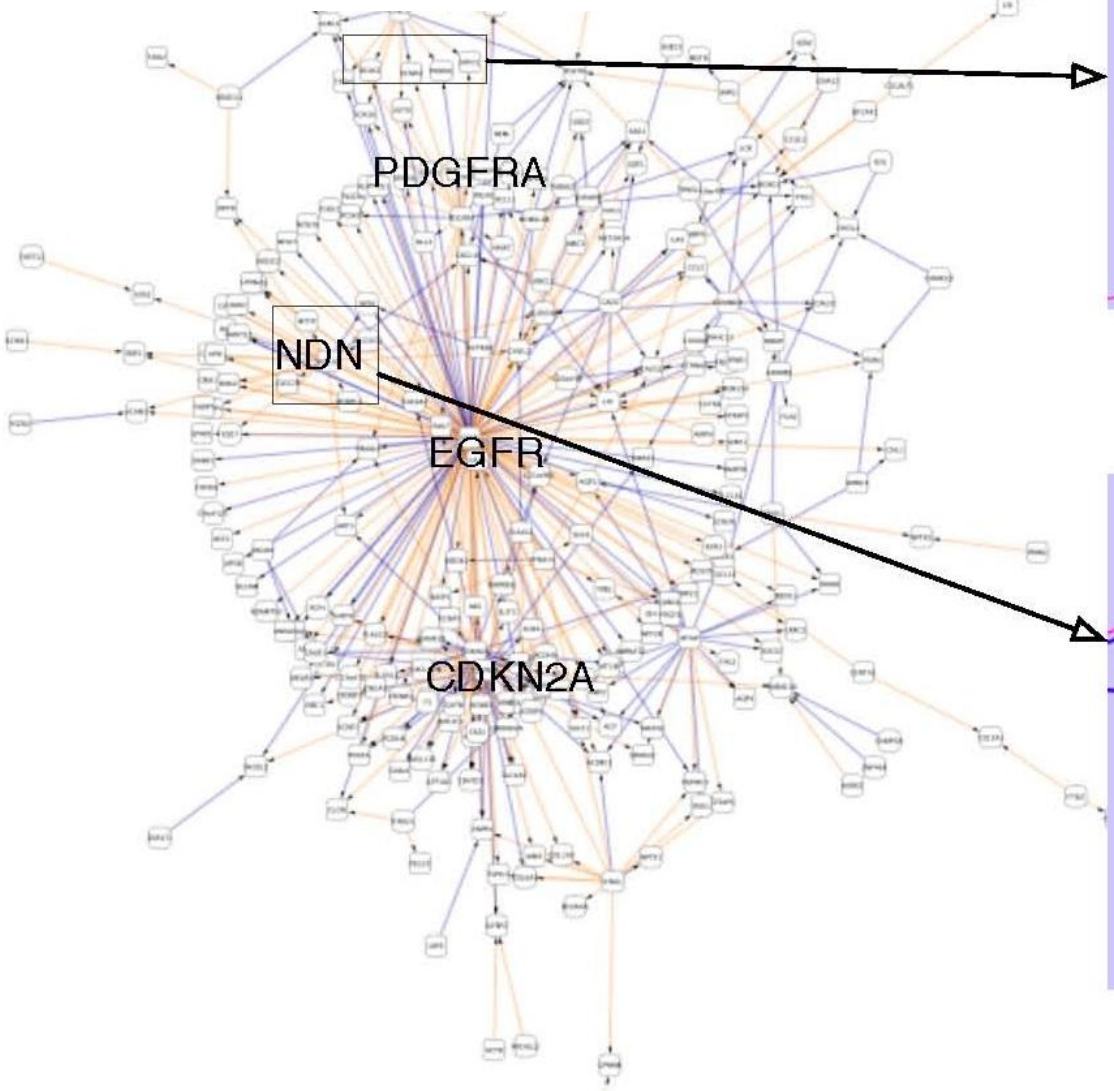
The EPoC model of grade IV glioblastoma is based on 146 patients from the Cancer Genome Atlas study



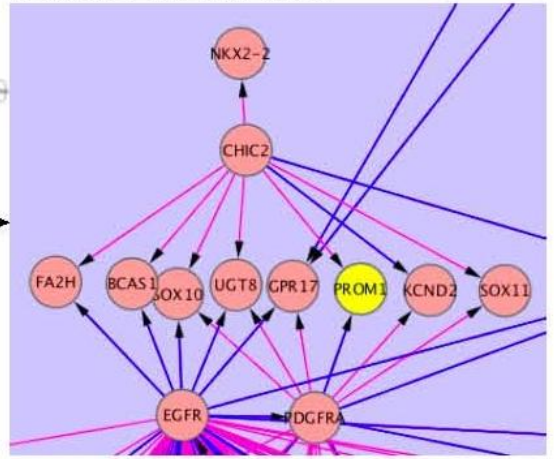
Circles are genes (HUGO symbols shown)
Red arrows: stimulatory interaction
Blue arrows: inhibitory interaction

Each interaction (n=415) has an individual posterior probability > 20%

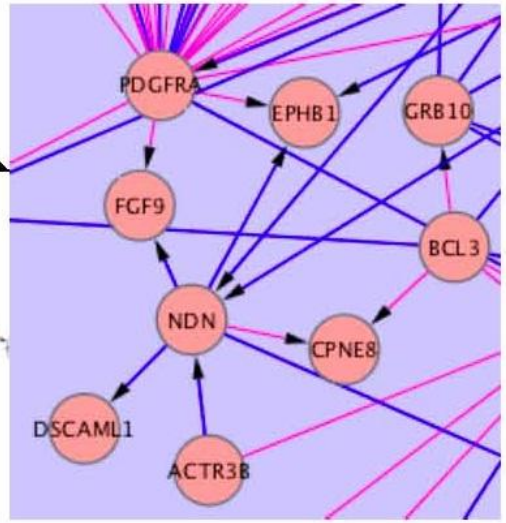
The overall network contains a number of well-established regulators and markers



Stem cell module

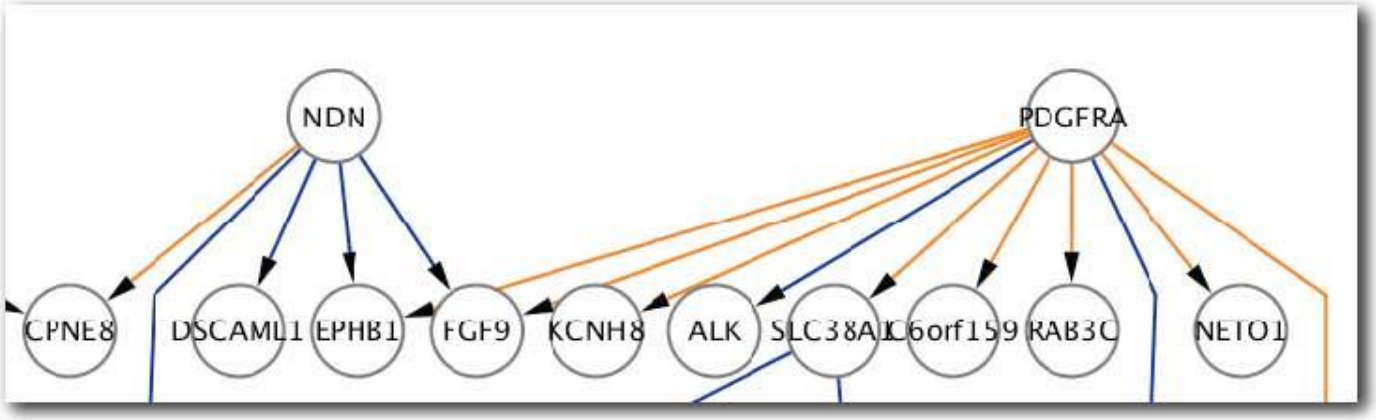


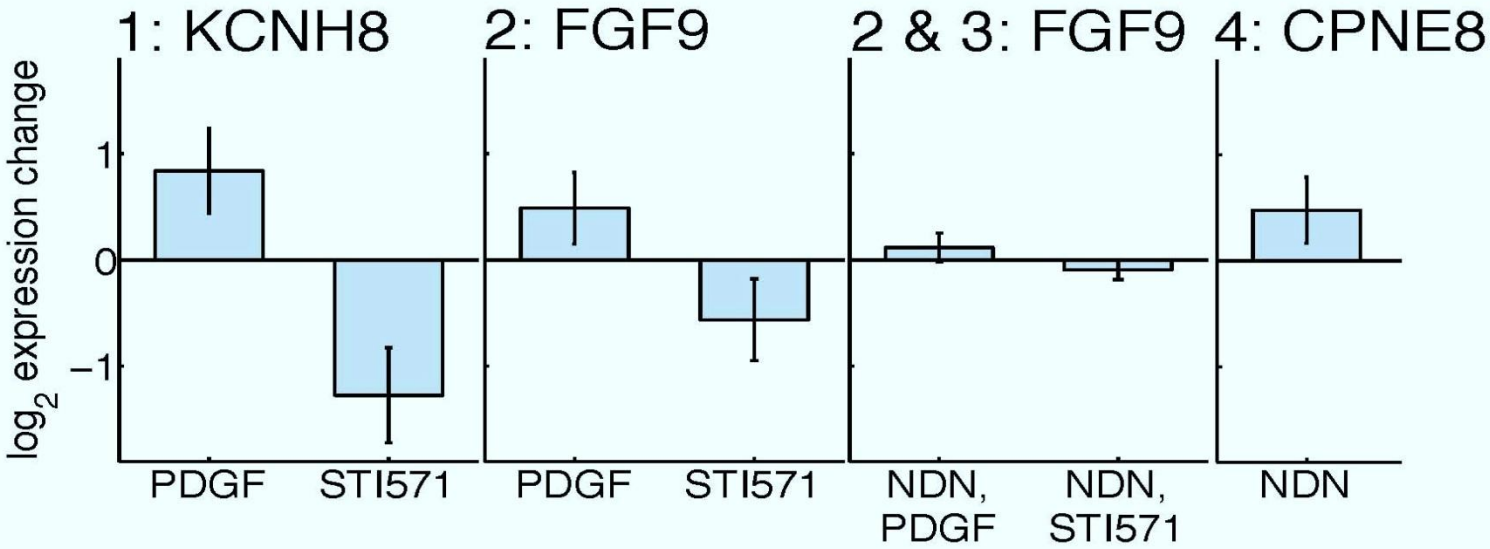
NDN as a hub



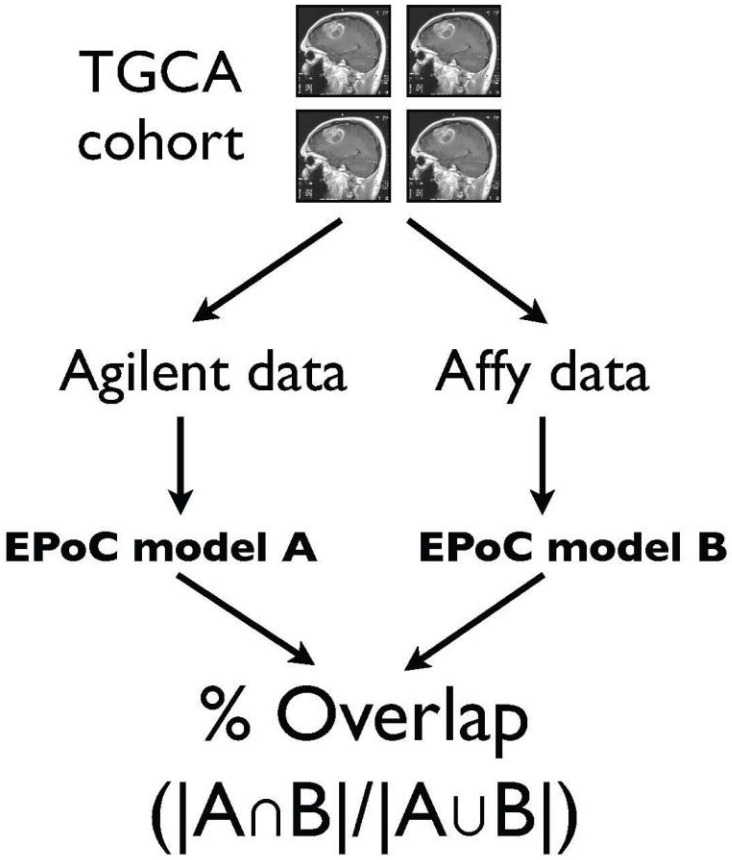
Testing hub genes: is NDN a previously overlooked glioblastoma tumor suppressor?

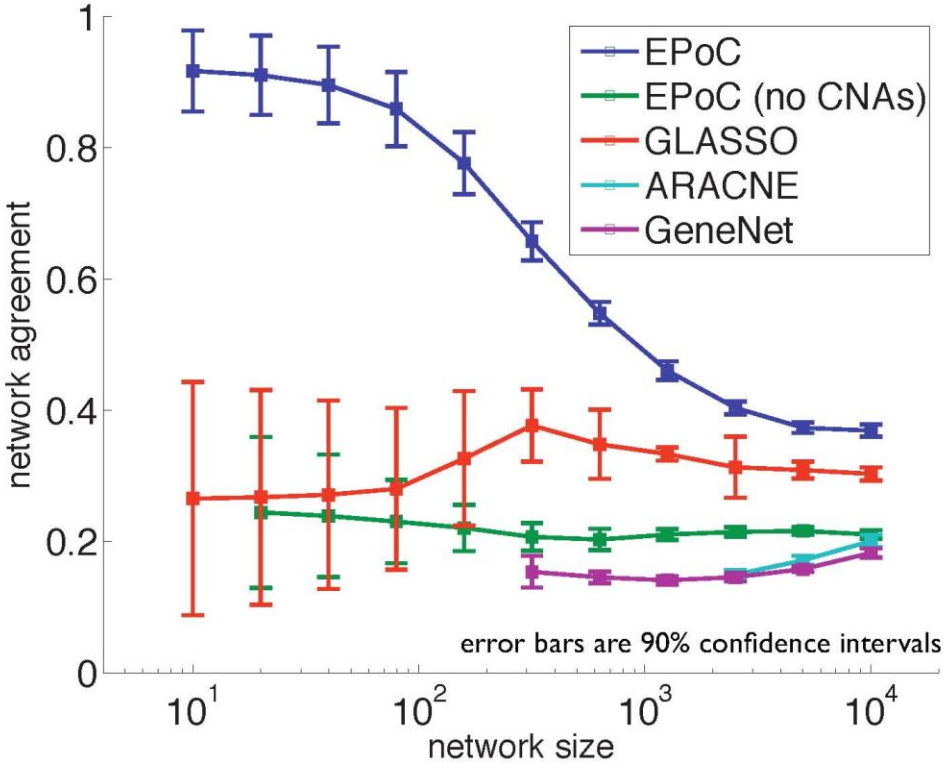
Nelander lab, together with Dr. Linda Lindahl at Sahlgrenska





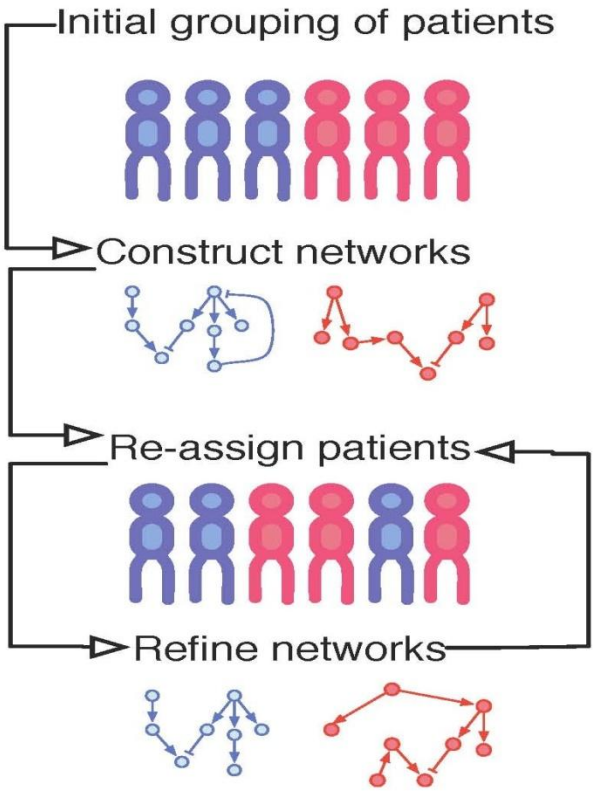
EPoC is faster and more consistent than a set of alternative methods



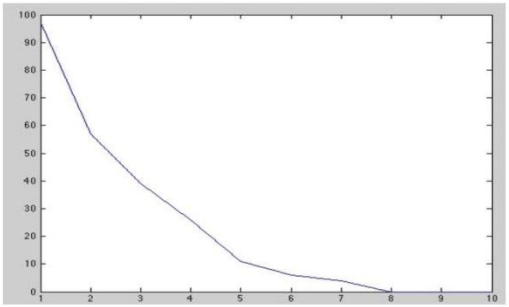


Part 2: Dynamic clustering of gene expression.

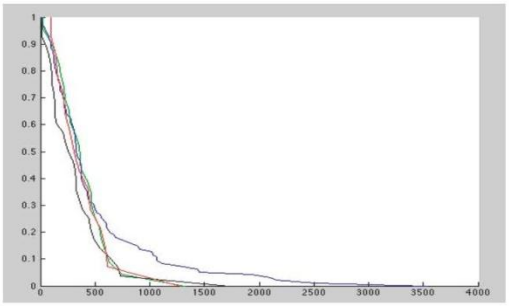
- A. Extension of Part 1 – underway.



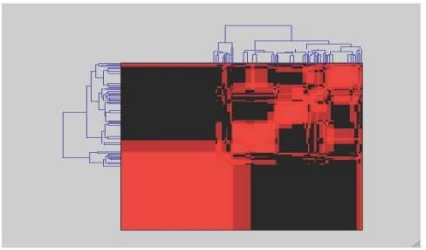
Convergence



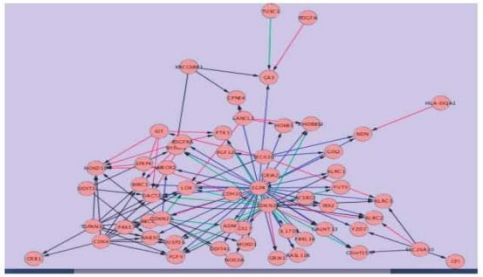
Network-based clustering predicts survival



Robust co-clusters



Cluster specific networks



Part 2: Dynamic clustering of gene expression.

- B. Mixture modeling with Glasso

Previous work by Wei Pan et al. 2009

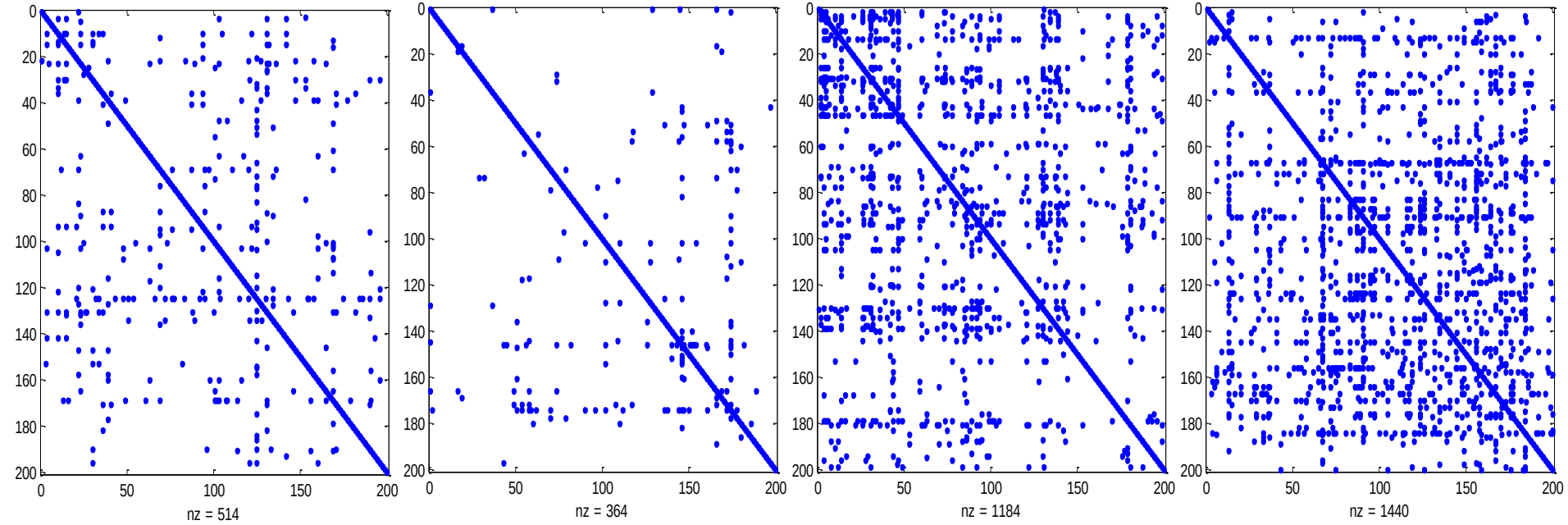
Here,

* We investigate the selection of penalty parameters for different clusters.

* We view the mixture component InvCov estimates as providing a dynamic bi-clustering.

Example: 4 clusters – InvCov with non-zero entries highlighted.

* Stable gene cluster – any edges present across all tumor clusters.
* Dynamic gene cluster – any edges that are present for only a subset of tumor clusters.



Dynamic clustering of gene expression.

Modelselection problem – sparsity of clusters
- number of clusters
one can affect the other…

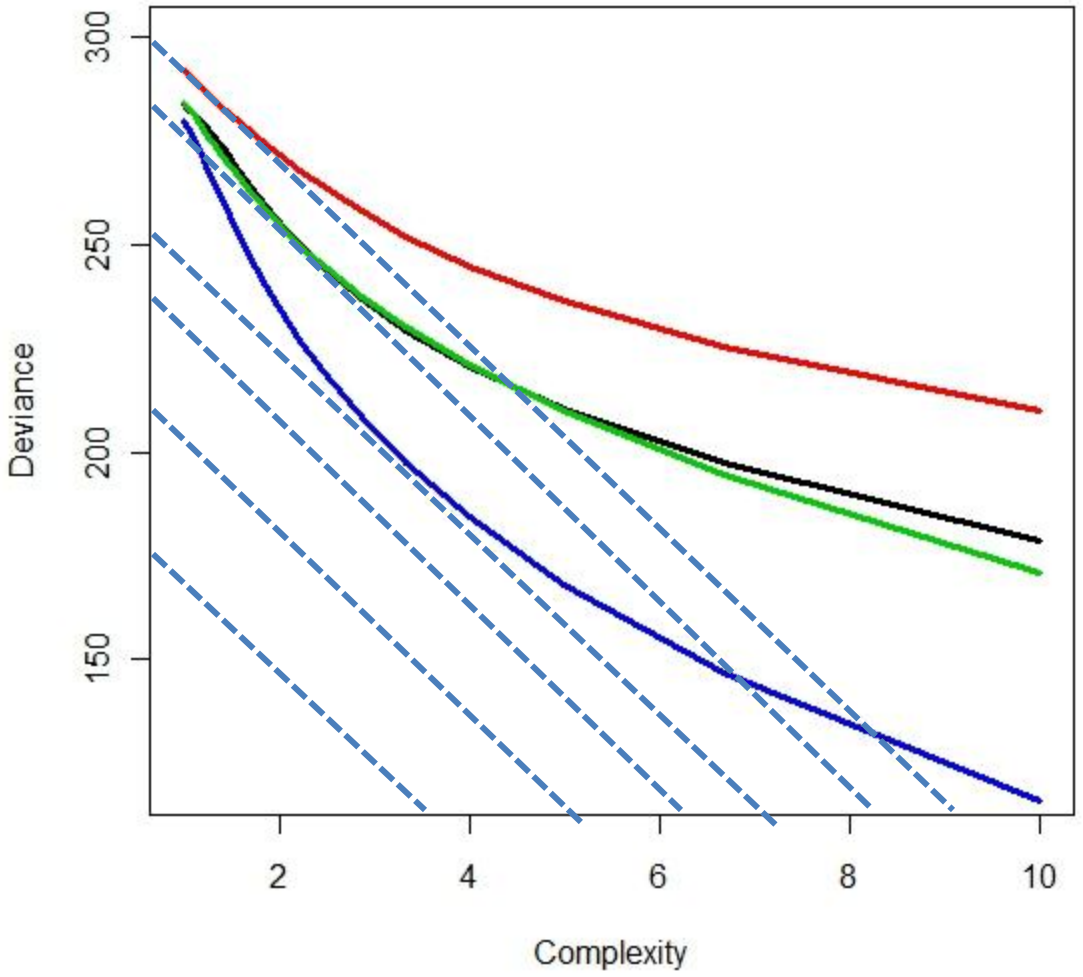
- How to penalize the complexity of different clusters?

- Here, we build on Jornsten 2009 using Rate-Distortion to select penalty parameters for each cluster.

Dynamic clustering of gene expression.

Rate-Distortion :

trade-off
between
Complexity and
fit (Deviance)
is the same for
all clusters.



Dynamic clustering of gene expression.

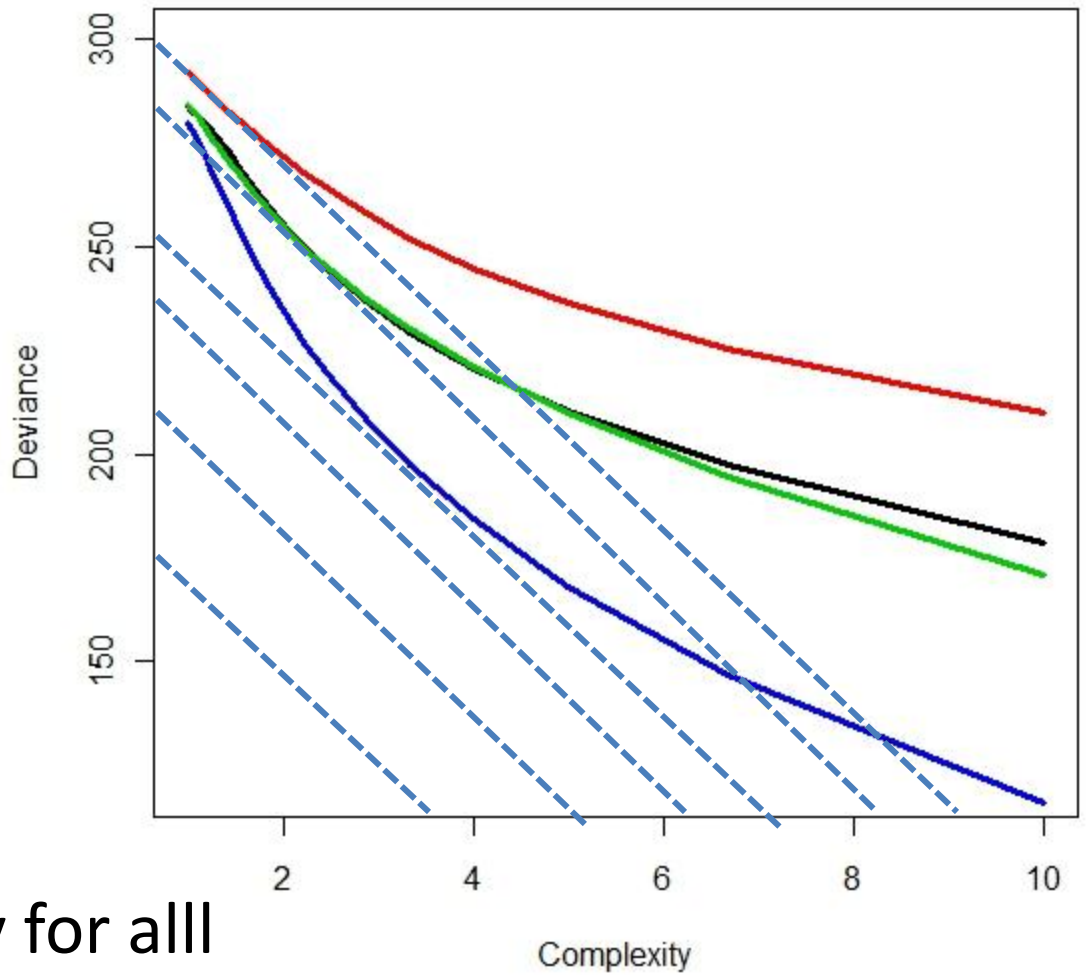
Rate-Distortion :

Single tuning-parameter search:

The Slope

vs.

Having to search over a multivariate grid.



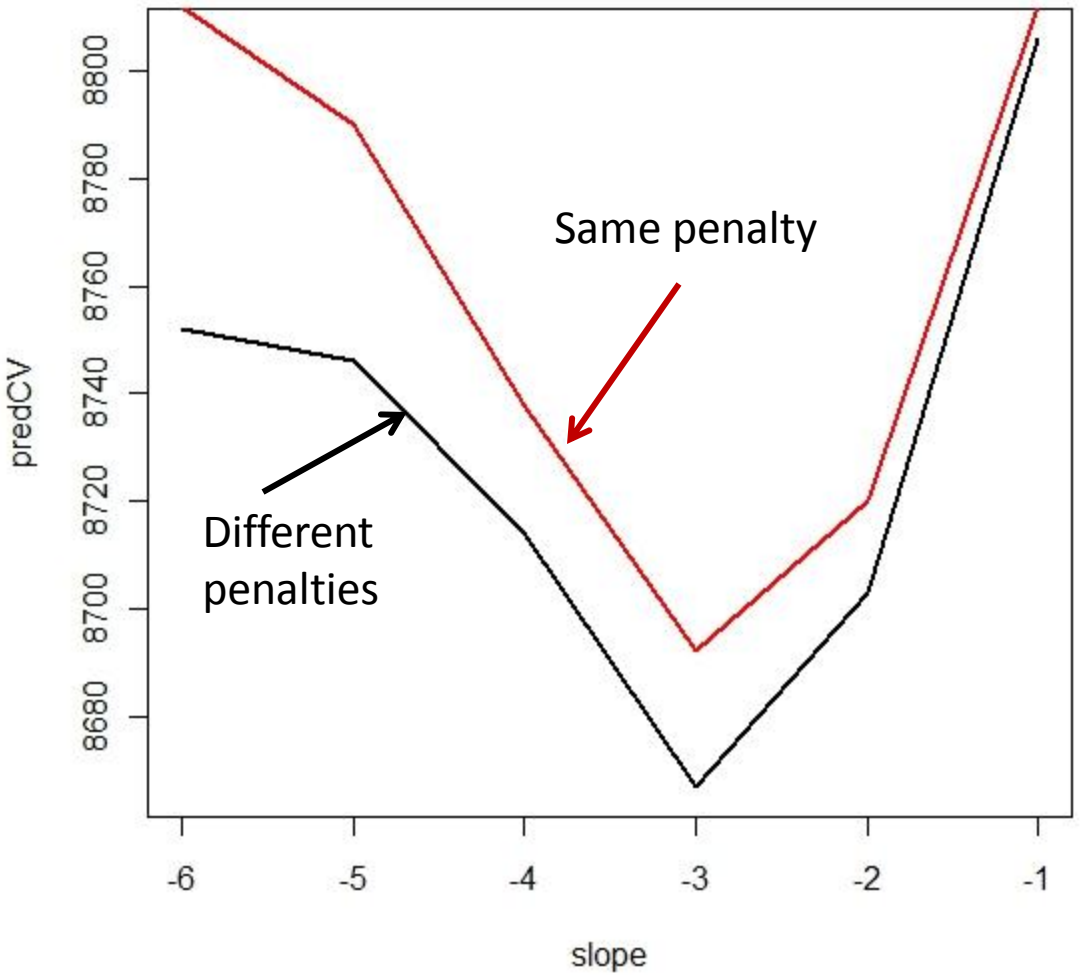
Performance much improved cmp using same penalty for alll

Dynamic clustering of gene expression.

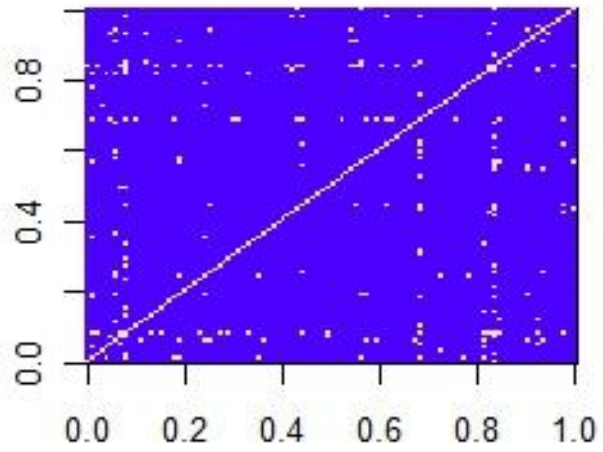
Algorithm (Given K number of clusters)

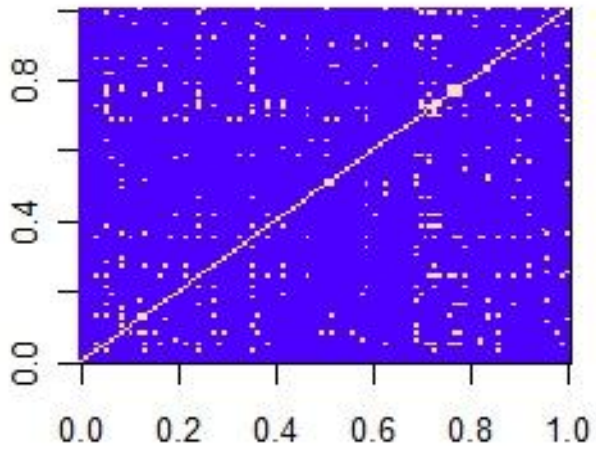
* For Slope Constraint S = Smin,…,Smax
  - For B random splits into TrainData and TestData
  - On the TrainData…
    - Repeat until convergence:
      - ❖ Run EM-Glasso with small common penalty
      - ❖ For k=1,…,K – for cluster k
        - ▪ Run Glasso with different penalties and compute the RD-curve
        - ▪ Pick point on curve with slope S and ID corresponding penalty
      - ❖ Run EM-Glasso with different penalties
  - Compute the CV predicted likelihood on the TestData
* Choose the Slope Constraint S_{opt} that minimized predCV
* Run EM-Glasso on full data with selected penalties.

- The prediction CV error is minimized for 4 clusters.
- The prediction CV is smaller when using different penalties for the different clusters

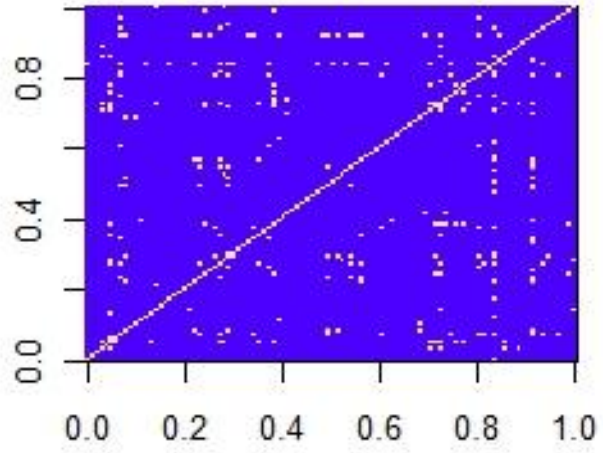


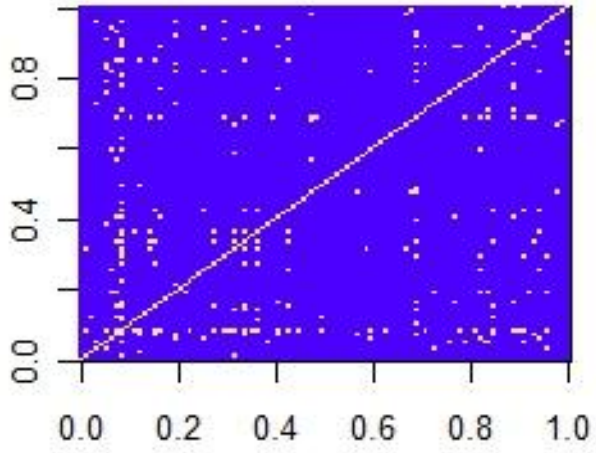
The InvCov complexity varies for the 4 tumor clusters.

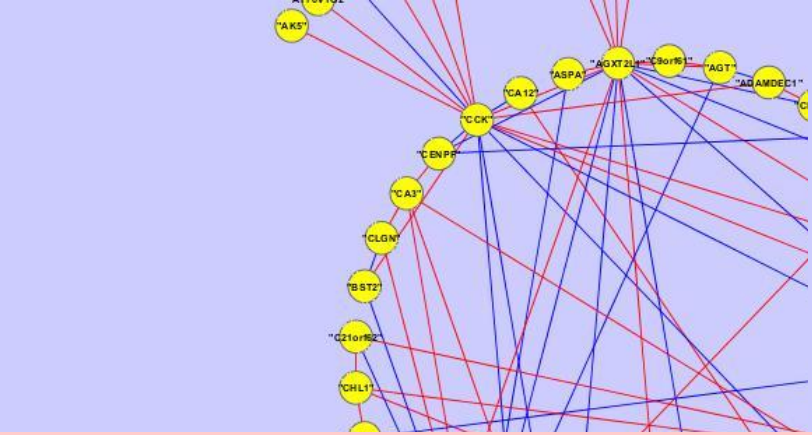


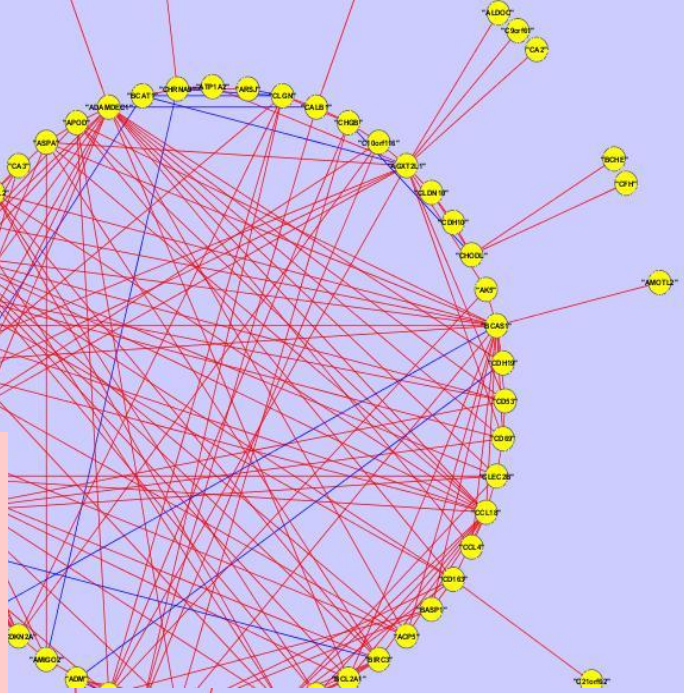


…. But some gene-gene interactions appear to be present in all or the majority of clusters, others are unique.

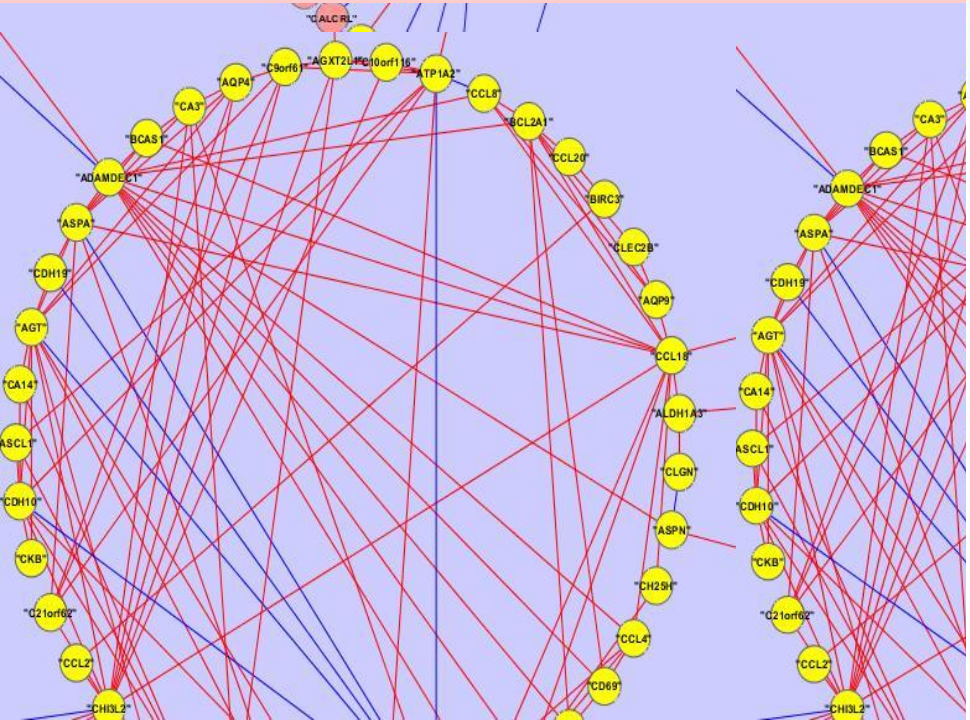








The 4 tumor clusters overlap in part with consensus clusters found by Verhaak et al, 2009



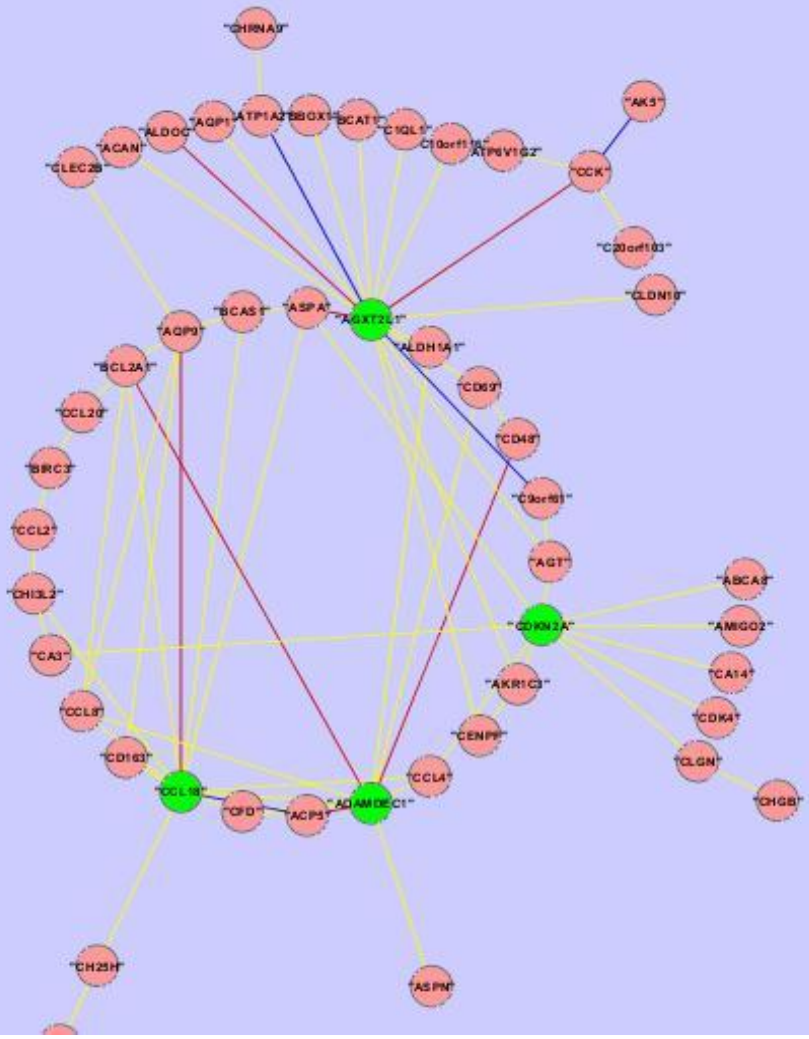
Cell PRESS

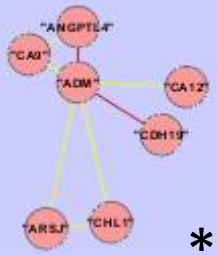
Cancer Cell
Article

Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in *PDGFRA*, *IDH1*, *EGFR*, and *NF1*

Roel G.W. Verhaak,[1,2,17] Katherine A. Hoadley,[3,4,17] Elizabeth Purdom,[7] Victoria Wang,[8] Yuan Qi,[4,5] Matthew D. Wilkerson,[4,5] C. Ryan Miller,[4,6] Li Ding,[9] Todd Golub,[1,10] Jill P. Mesirov,[1] Gabriele Alexe,[1] Michael Lawrence,[1,2] Michael O'Kelly,[1,2] Pablo Tamayo,[1] Barbara A. Weir,[1,2] Stacey Gabriel,[1] Wendy Winckler,[1,2] Supriya Gupta,[1] Lakshmi Jakkula,[11] Heidi S. Feiler,[11] J. Graeme Hodgson,[12] C. David James,[12] Jann N. Sarkaria,[13] Cameron Brennan,[14] Ari Kahn,[15] Paul T. Spellman,[11] Richard K. Wilson,[9] Terence P. Speed,[7,16] Joe W. Gray,[11] Matthew Meyerson,[1,2] Gad Getz,[1] Charles M. Perou,[3,4,8] D. Neil Hayes,[4,5,*] and The Cancer Genome Atlas Research Network

[1]The Eli and Edythe L. Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, MA 02142, USA
[2]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02115, USA
[3]Department of Genetics
[4]Lineberger Comprehensive Cancer Center
[5]Department of Internal Medicine, Division of Medical Oncology
[6]Department of Pathology and Laboratory Medicine
University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
[7]Department of Statistics
[8]Group in Biostatistics
University of California, Berkeley, CA 94720, USA
[9]The Genome Center at Washington University, Department of Genetics, Washington University School of Medicine, St. Louis, MO 63108, USA
[10]Department of Pediatric Oncology, Center for Cancer Genome Discovery, Dana-Farber Cancer Institute, Boston, MA 02115, USA
[11]Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[12]Department of Neurological Surgery, University of California, San Francisco, CA 94143, USA
[13]Department of Radiation Oncology, Mayo Clinic, Rochester, MN 55905, USA
[14]Department of Neurosurgery, Memorial Sloan-Kettering Cancer Center, New York, NY 10065, USA
[15]SRA International, Fairfax, VA 22033, USA
[16]Walter and Eliza Hall Institute, Parkville, Victoria 3052, Australia
[17]These authors contributed equally to the work
*Correspondence: hayes@med.unc.edu
DOI 10.1016/j.ccr.2009.12.020





*Some oncogenes appear as hubs across all clusters.
*Some hubs are particular to a subset of clusters.

*Some interactions are stable across all or some clusters.

We are in the process of investigating this further for a larger set of genes, and regularizing toward similar magnitude and/or sign of the interaction across clusters .

Conclusion & Future Work

- EPoC scales to networks ~10000 nodes.
- Performance better than other popular methods.
- Extension: *to subclasses of tumors and class-specific networks.
  *to include multiple data sources, e.g. methylation.

- Dynamic clustering – using Glasso to detect tumor subclasses and dynamic/stable gene clusters.
- Extension to larger and multiple data sets + investigating more avenues of regularization between/within clusters.

Acknowledgements.