Large-scale network and prognostic analyses of gene expression and copy number aberration

Rebecka Jörnsten

Mathematical Sciences, University of Gothenburg/Chalmers

jornsten@chalmers.se

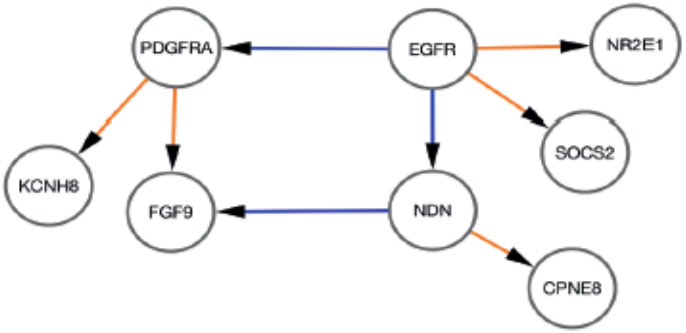
May 26, 2011

1. Data Integration - Network modeling
2. Modeling Steady-State mRNA levels as a function of Copy Number
   - CNA-driven network modeling
   - Estimation
   - Choice of regularization
3. Application
   - Analysis of glioblastoma
   - Validation
4. Software
5. Conclusion and Future work

Endogenous Perturbation analysis of Cancer

Goals

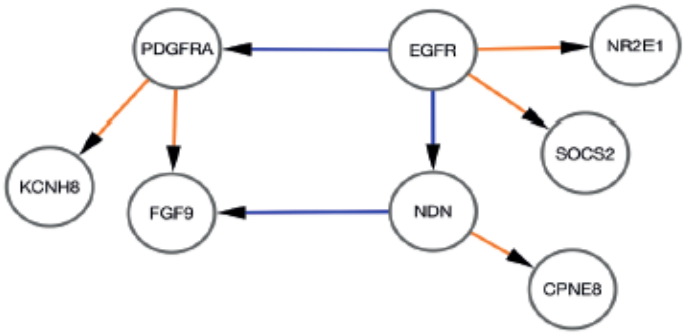
- Construct regulatory network and predictive models for cancer pathways
- Identify disease-specific key regulators and their targets
- Relate the network structure to patient survival



- Data now available at multiple levels (genetic seq, transcription, proteomics,...)

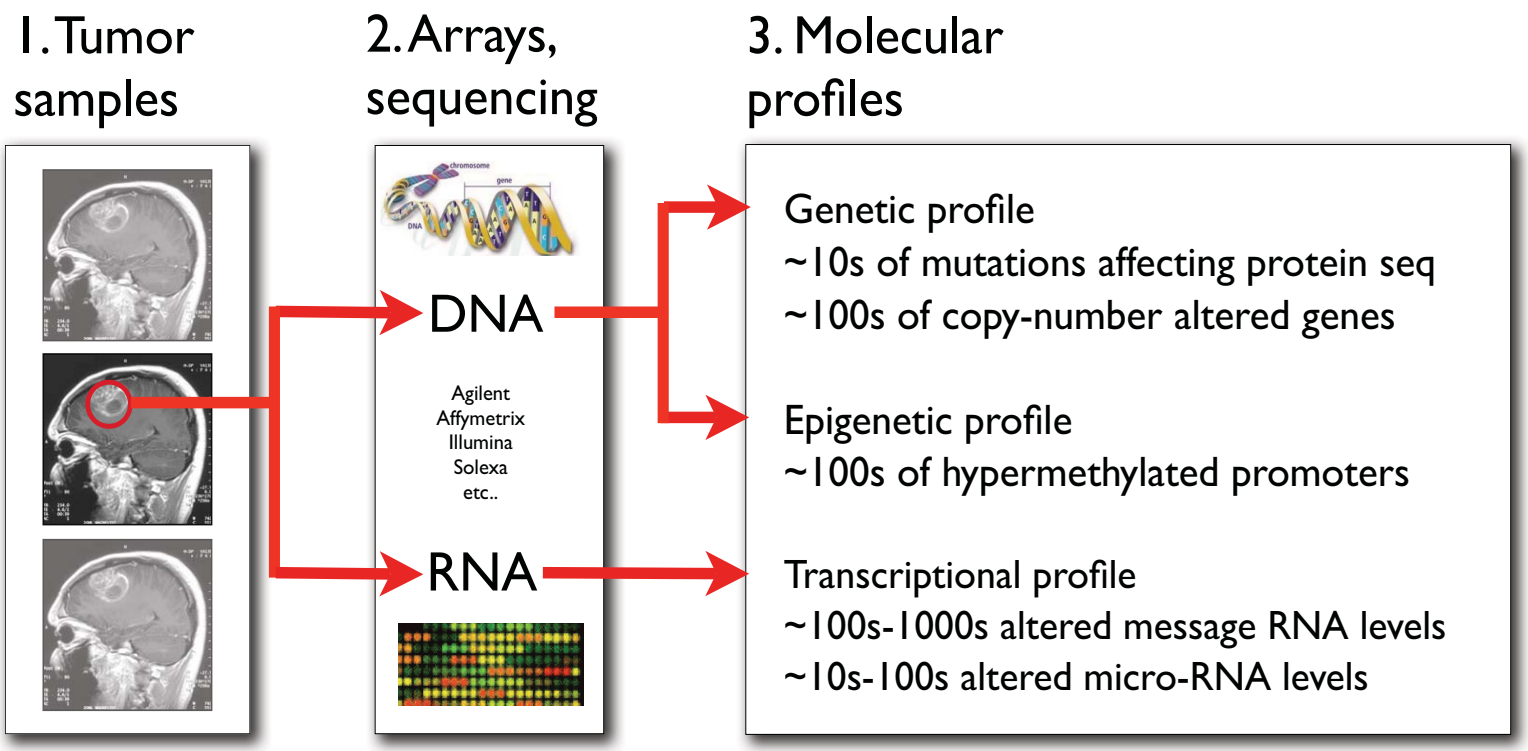
Endogenous perturbation

- We view each tumor's Copy Number Aberration (CNA) profile as a system perturbation that simultaneously affects multiple genes, and
- the mRNA profiles as the steady-state response to that perturbation

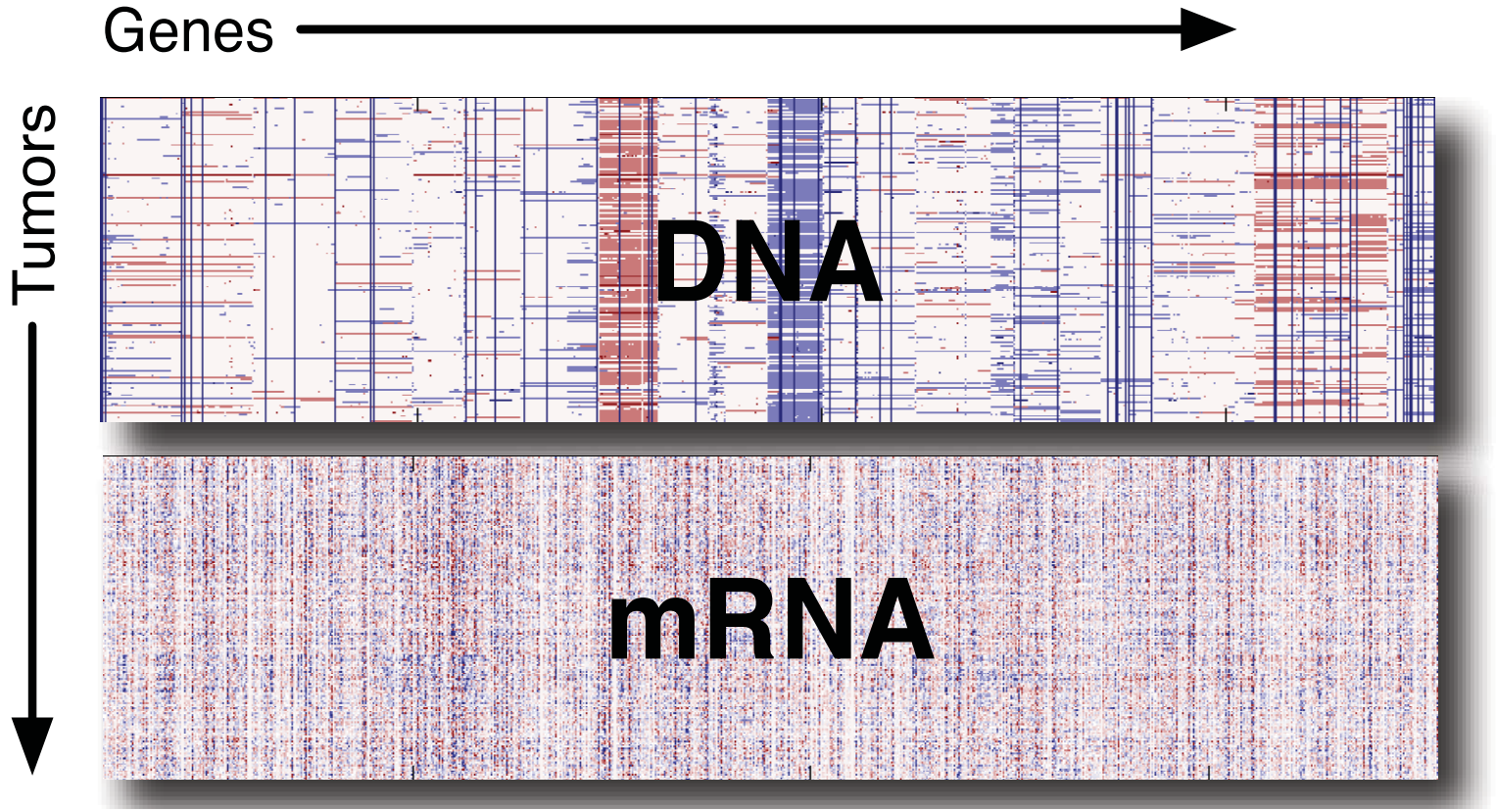


- CNAs tend to appear in a patient-specific manner - ideal for network construction
- Ongoing projects like TCGA means massive amounts of mRNA/CNA data are available

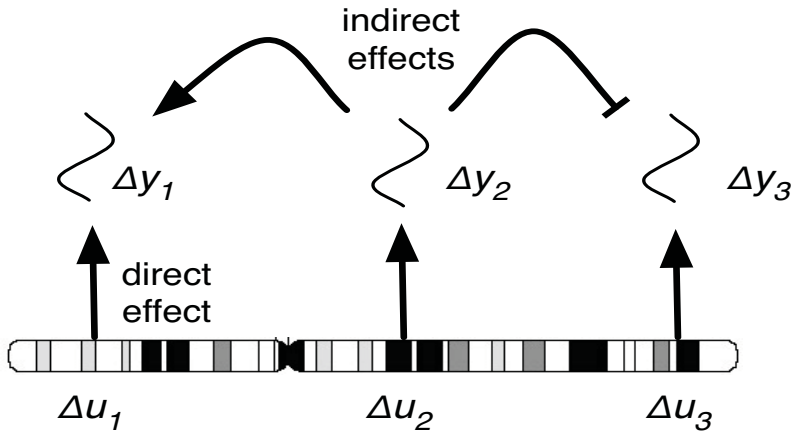
Data - 186 tumors from the Cancer Genome Atlas consortium (TCGA)



Data - 186 tumors from the Cancer Genome Atlas consortium (TCGA)



A model for CNA-mRNA interaction



After some manipulation, we can write

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

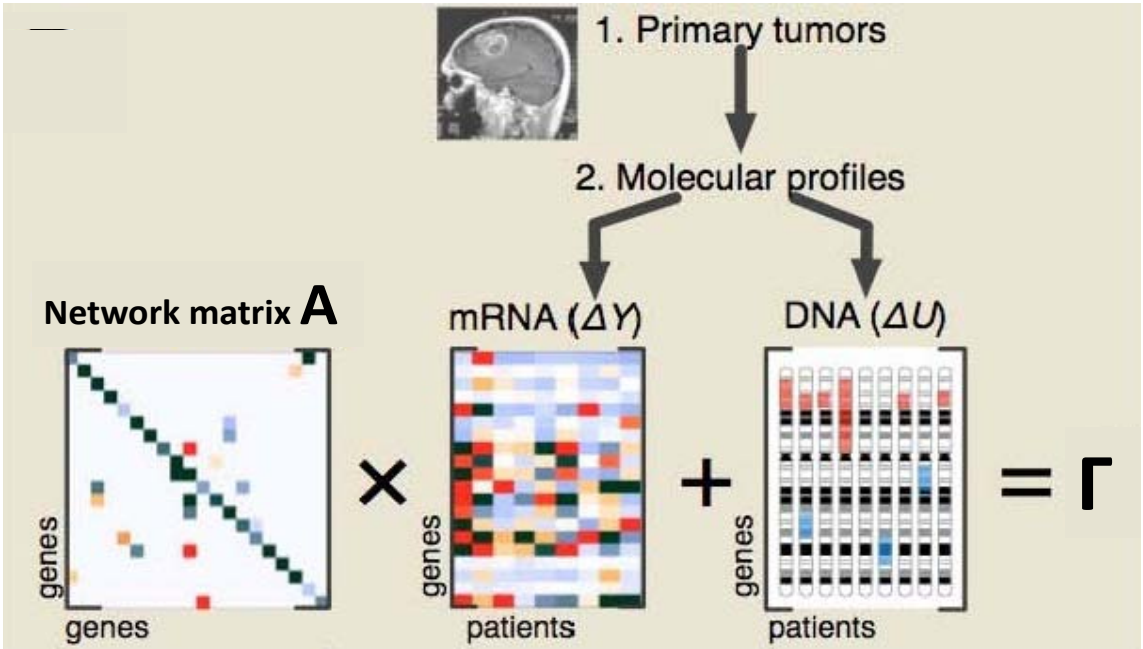
where **y** is the mRNA levels and **u** the CNA.

- The elements of A, a_{ij}, capture the *direct* causal influence of transcript j on i.
- Γ is the 'noise' (non-CNA specific effect on mRNA).

Transcriptional network

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

- A is called the *transcriptional network*.
- Γ includes all the tumor-specific differences, unmeasured environmental effects, SNPs etc.



CNA-driven network

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

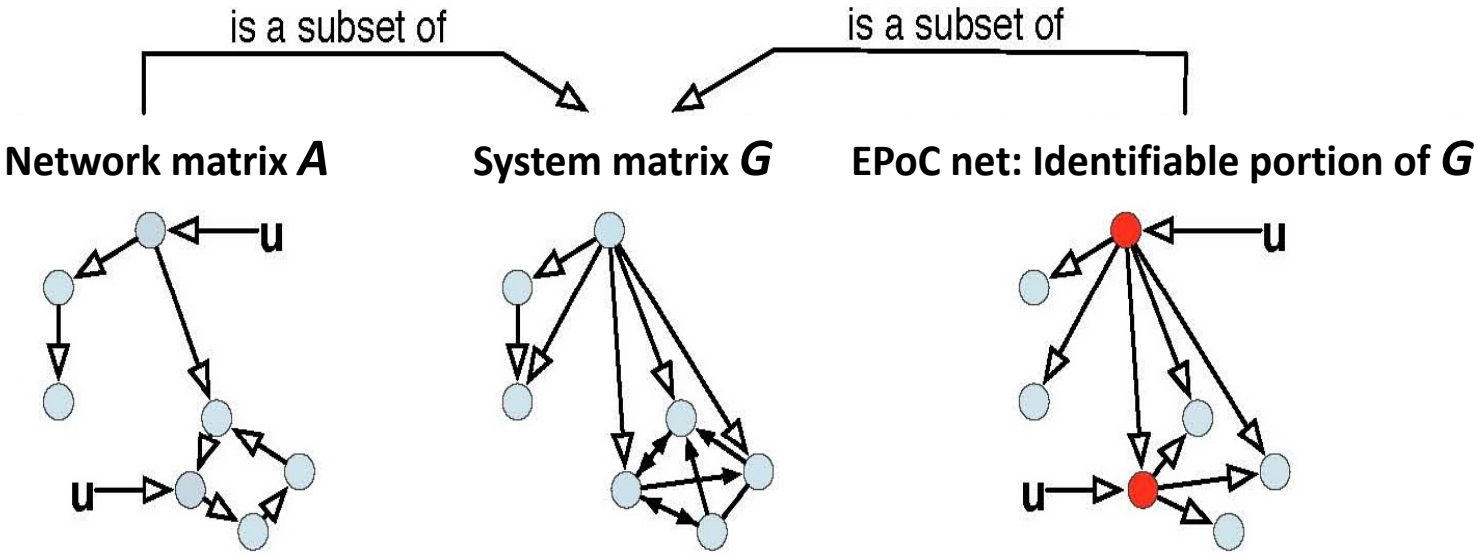
- $G = -A^{-1}$ is called the *system matrix*, or *CNA-driven network*.
- G represents the *system gain*: where the genetic variation (system input) shows up as amplified (positive or negative) signal in the mRNA expression (system output).
- G thus detects transcriptional modules under CNA control
- whereas A contains both direct effects of the CNA as well as mRNA-mRNA regulation that may be non-disease specific.

A vs G

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

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

- A is difficult to estimate due to strong correlations between mRNAs in related pathways.
- The correlation structure among the CNAs is much lower, making robust estimation of G an easier task.



Estimating G

- The system matrix G is $n \times n$, where $n = 10000+$ genes.
- We have only $T = 186$ tumors/samples.
- Need to use regularized estimation techniques.

L1 penalized regression

$$\min_G \|\Delta Y - G\Delta U\|_F^2 + \lambda \sum_{i \neq j} \mid G[i,j] \mid$$

- λ is the regularization parameter that controls the degree of sparsity (number of non-zeroes) in G.
- We don't penalize the diagonal elements of G since direct CNA-mRNA for each gene is assumed to always be present.

Choosing λ - the network size

- We want the networks to be interpretable and sparse
- No "spurious" edges should be included
- How do we validate the results?

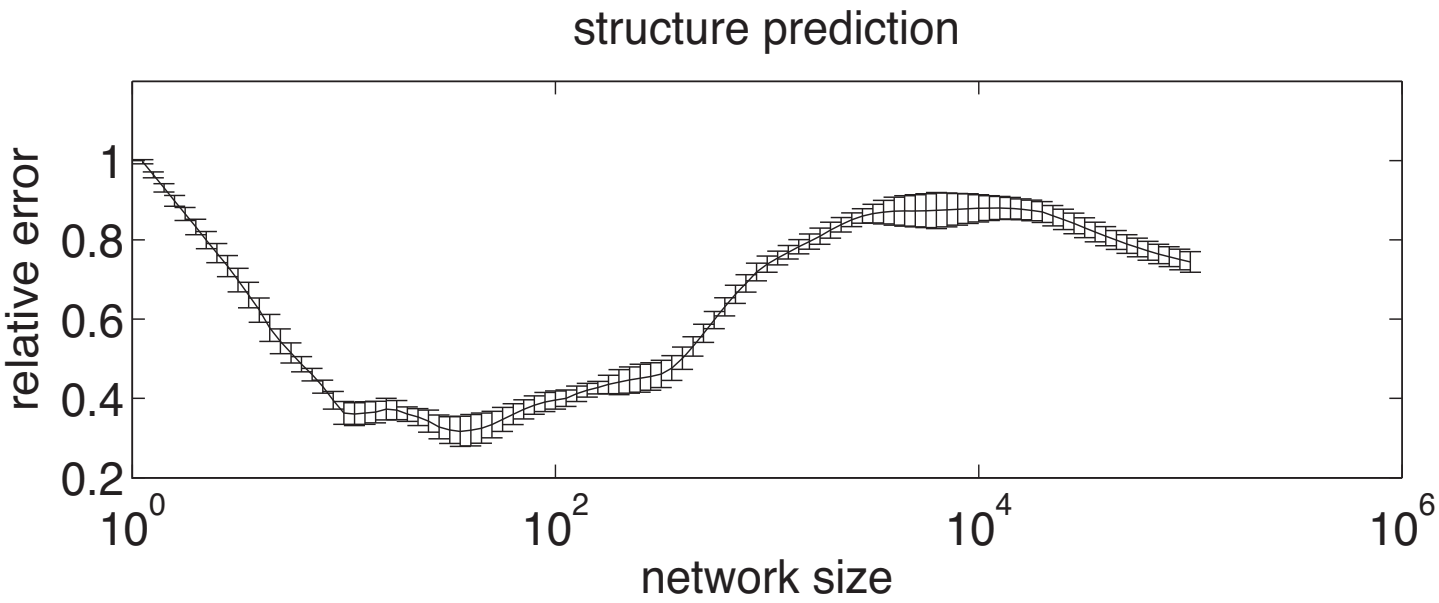
Validation statistics

We consider two different evaluation measures;

- Network structure consistency - Kendall's W
- mRNA prediction - minimum average prediction error
- A robust, final network is obtained from repeated bootstrap simulations - only edges that appear consistently across bootstrap samples are kept in the final network model.

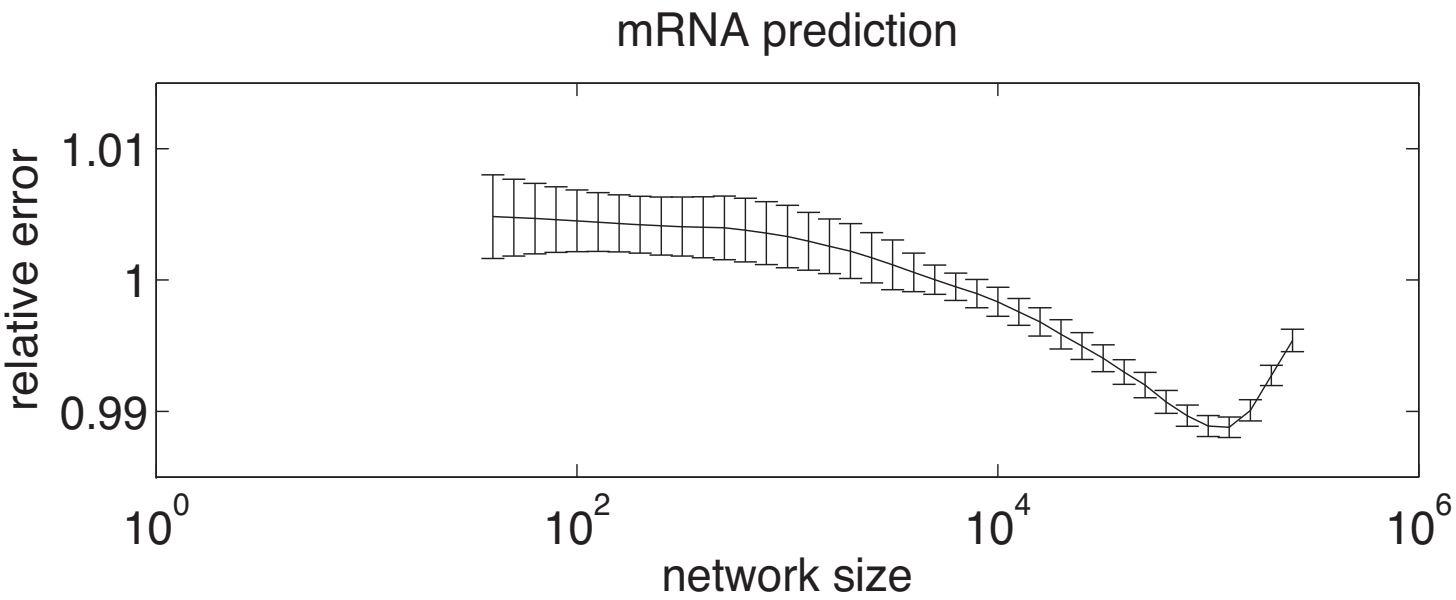
Network consistency

- Compare networks obtained from two random data subsets using Kendall's W - more appropriate correlation measure when any number of distinct outcomes (edge weights) can occur.
- W is 1 if all networks agree, and 0 if the network agreement is essentially random.
- Most consistent networks contain $\sim$ 400 edges.



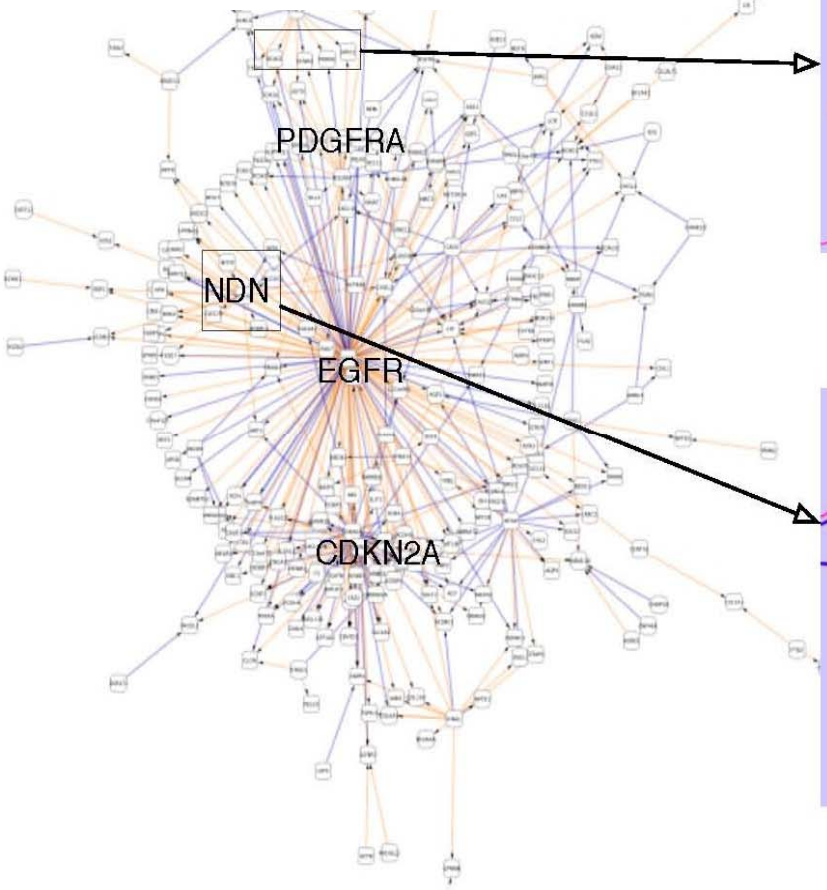
mRNA prediction

- We use $k = 10$-fold cross-validation and pseudo-bootstrap.
- We construct the network leaving out 1/10th of the data
- We use the network model to predict the mRNA levels from the CNA levels on the leave-out data.
- Minimum mRNA prediction errors for networks with ~ 10000 edges.

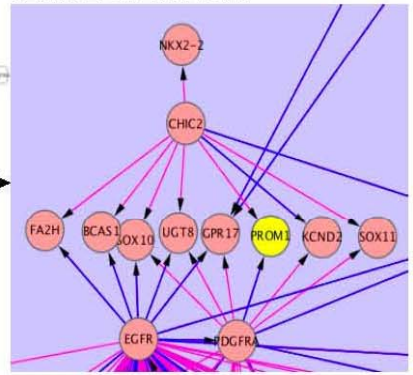


Glioma network analysis

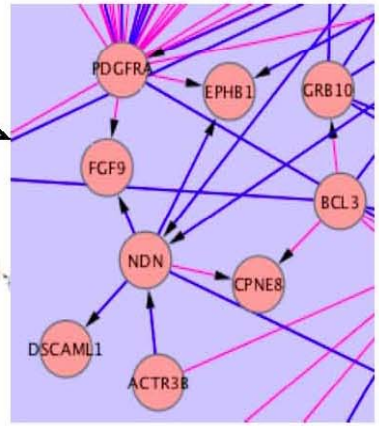
The overall network contains a number of well-established regulators and markers



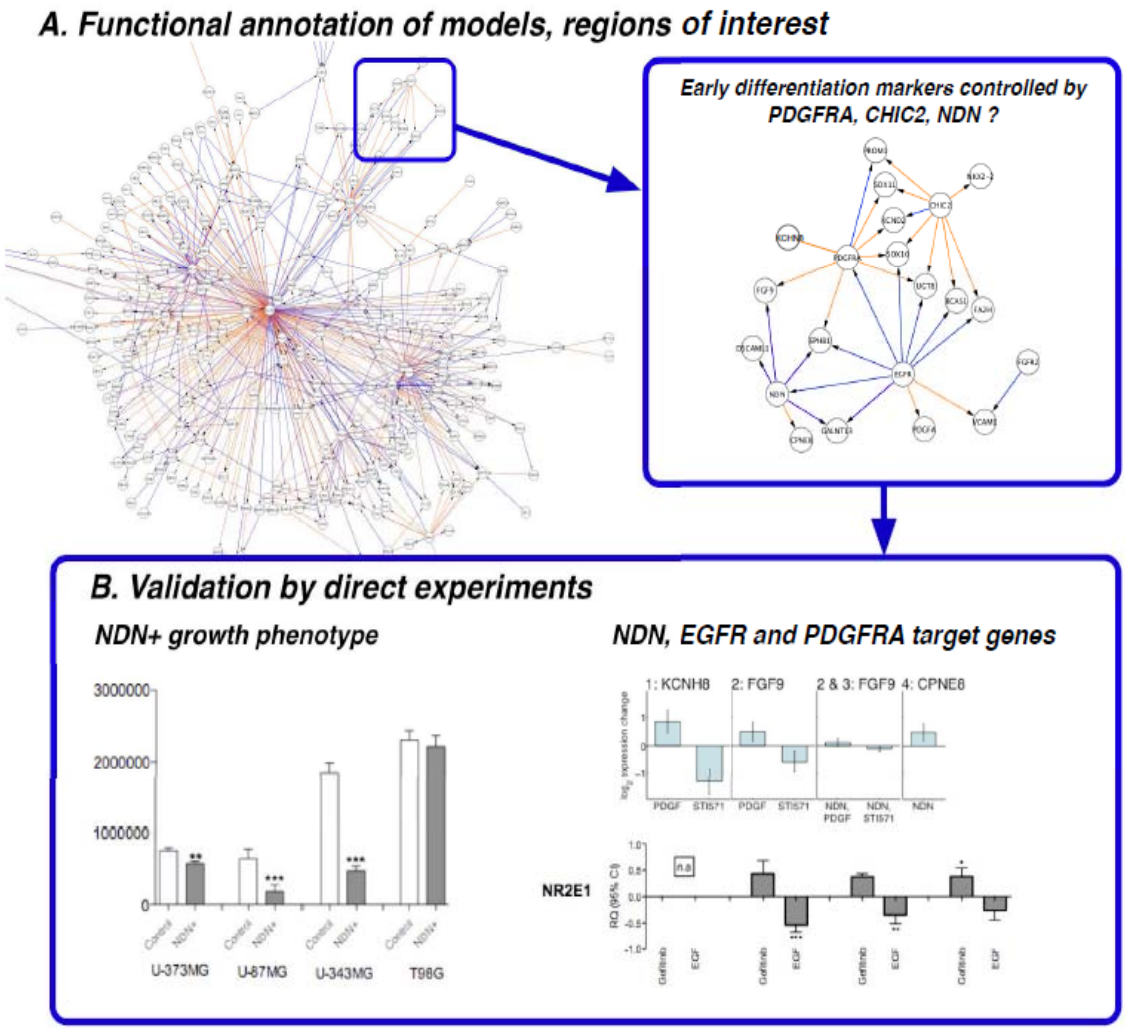
Stem cell module



NDN as a hub

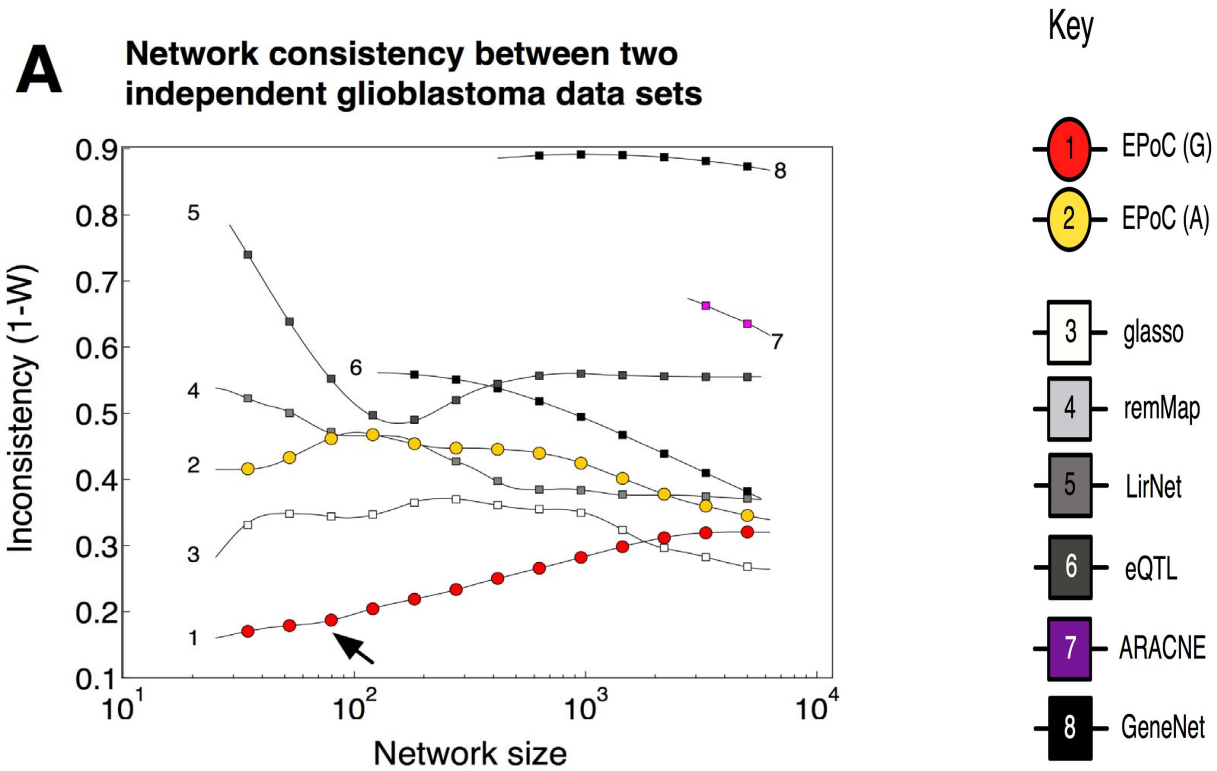


Experimental validation

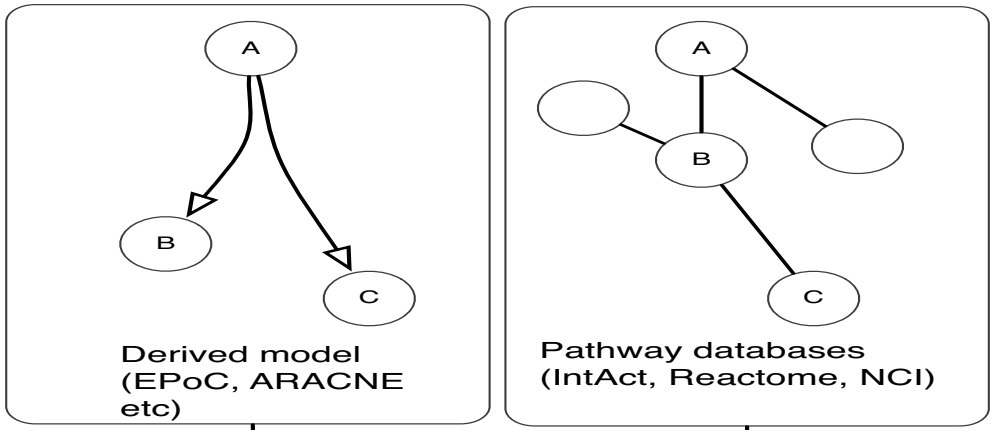


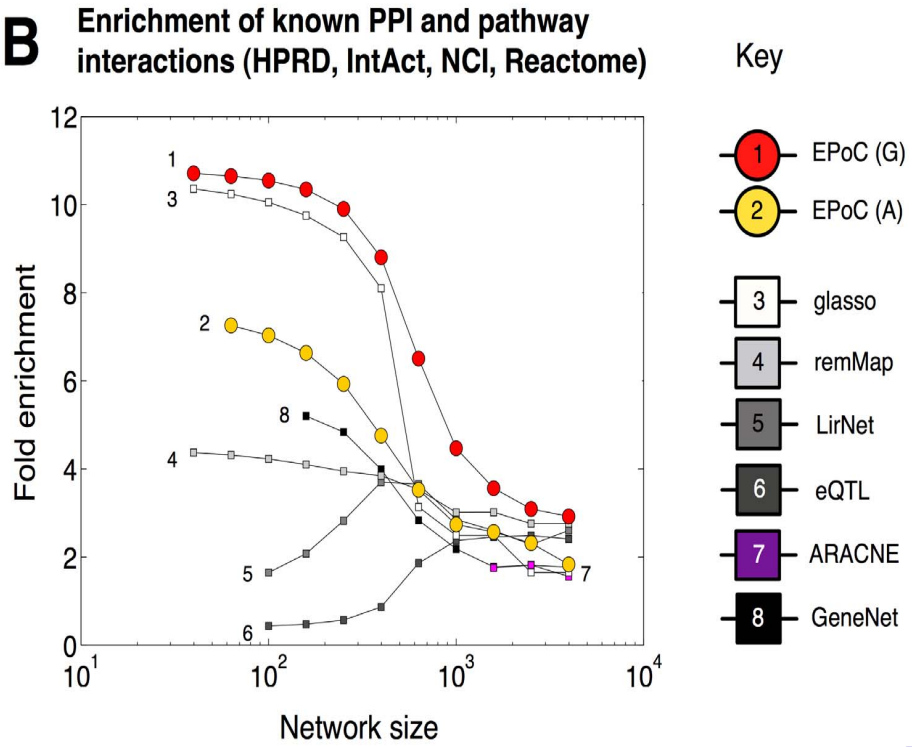
Network consistency

We compare network consistency across network sizes and between competing methods.



Overlap with pathway databases



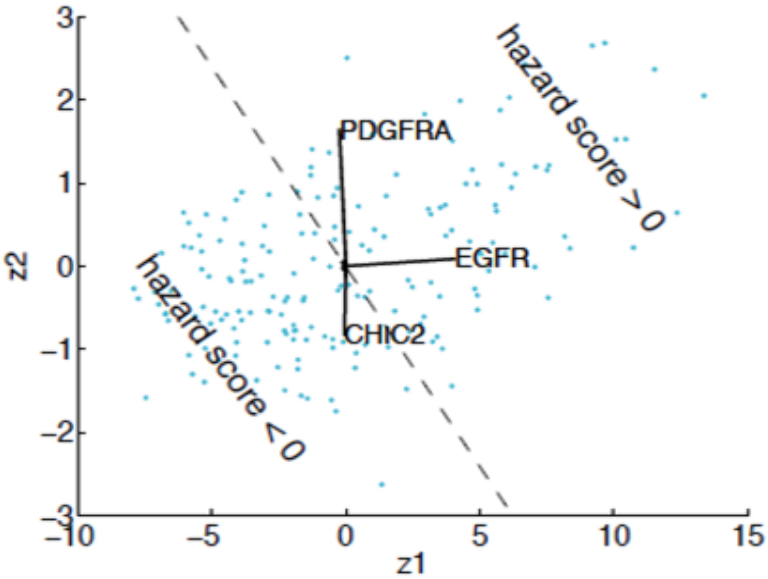


- We map networks to pathway repositories HPRD, Reactome, Intact, and NCI-nature.
- Compare pathway links to the shortest paths in networks.
- EPoC-G is clearly enriched for short or direct paths cmp other methods.

Decomposition of G

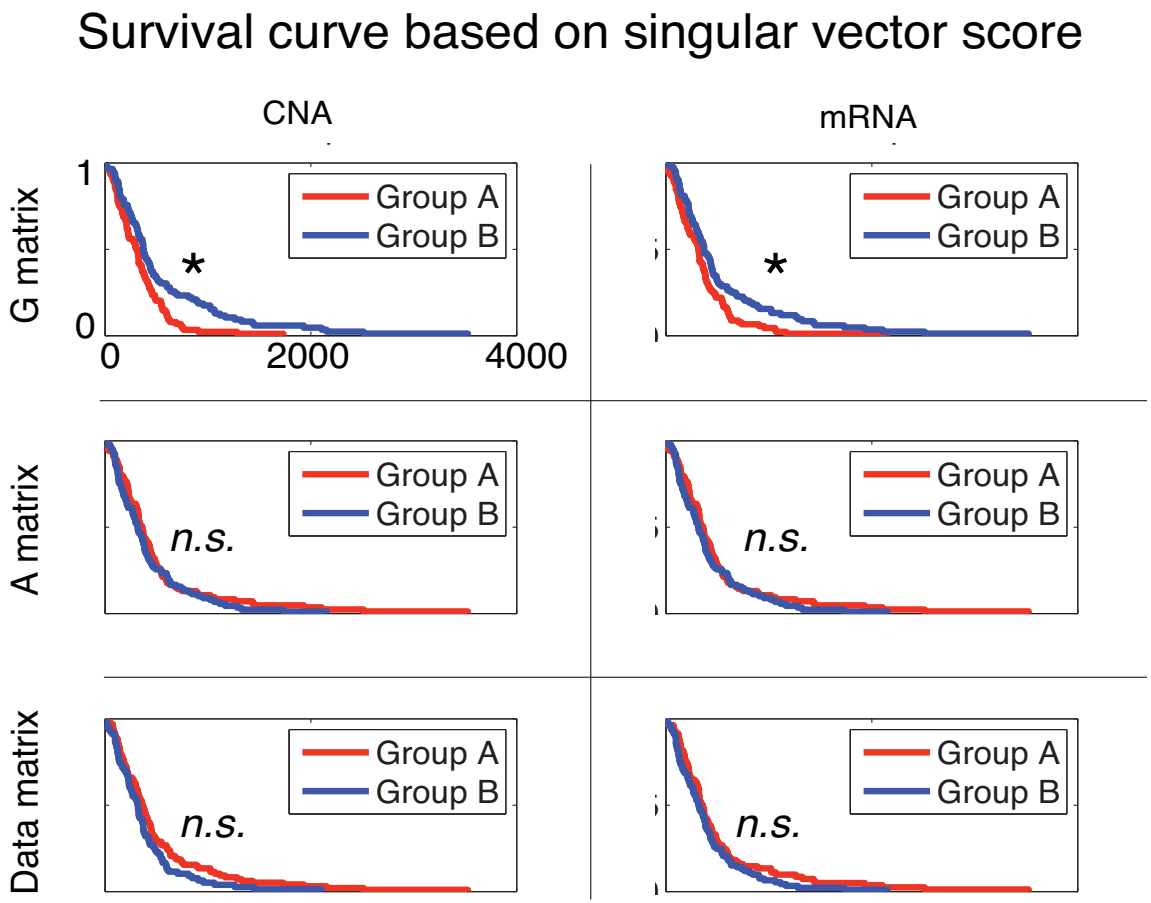
- Genomic data = 10000 by 186, Survival data = 1 by 186 - how relate?
- The SVD decomposition of $G = C\Lambda D^T$ has the following meaning:
  - leading columns of D are directions of CNA perturbations that are amplified by the system.
  - leading columns of C are directions of mRNA transcripts most affected by the directions in D.
- Write $\Delta Y = G\Delta U = C\Lambda D^T \Delta U$
- $\rightarrow C^T \Delta Y = \Lambda D^T \Delta U$
- Projecting mRNA onto columns of C = output, Projecting CNA onto D = input, Λ = amplification.
- Projection scores = same or similar dimensionality as survival data.

Decomposition of G



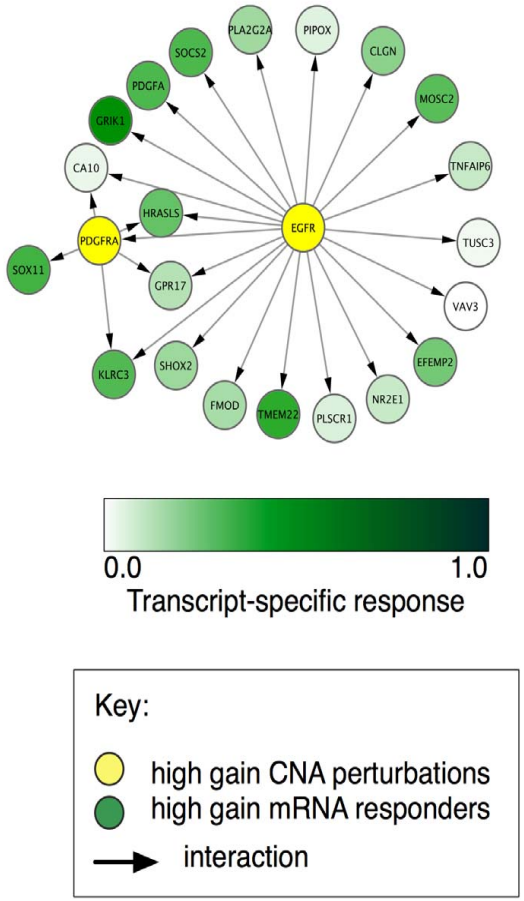
- $C^T \Delta Y = \Lambda D^T \Delta U$
- Consider the leading projections (first columns of C and D).
- mRNA profiles of individual patients are projected onto C: $Z_y = C^T \Delta Y$ and CNA profiles are projected by $Z_u = D^T \Delta U$
- Compare the survival of the patients using these projected scores.

Decomposition of G vs A and data



Decomposition of G

- We can color-code the networks using the leading SVD components.
- Identifies disease driving perturbations and their targets



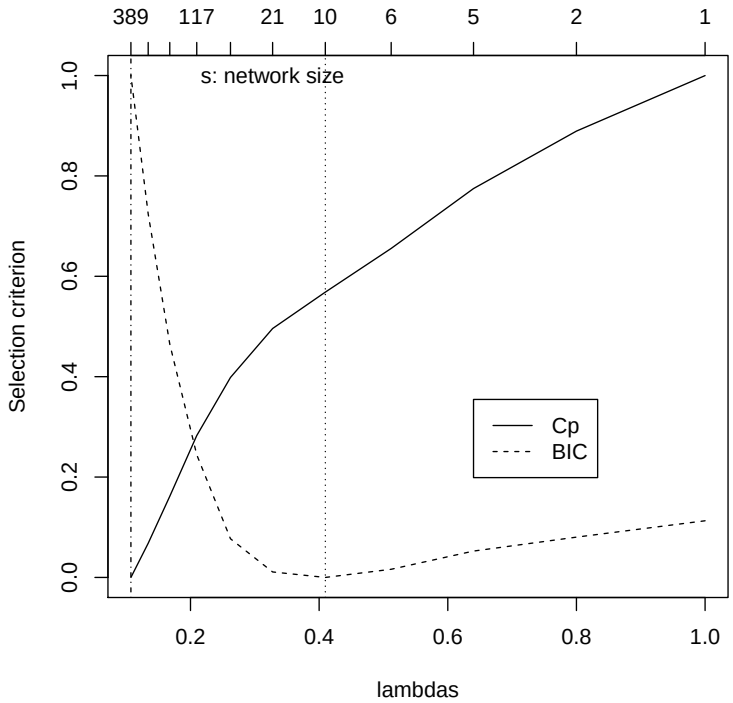
EPoC R package

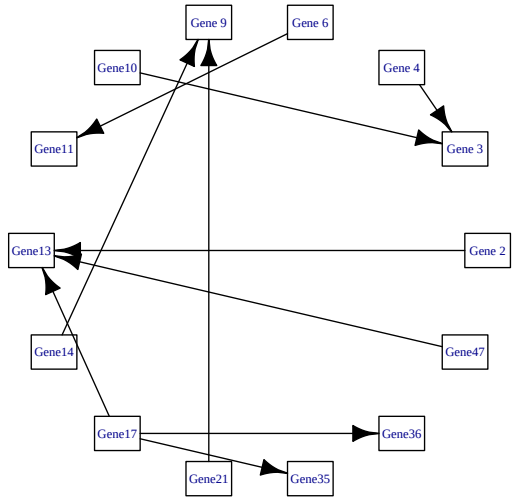

```
> install.packages('epoc')
> library(epoc)
> G <- epocG(y,u)
> summary(G)
Call:
epocG(Y = y, U = u)

Models:
```

	R2	Cp	BIC	RSS	links
lambda=1	0.0783	10088.244	95.4035	8526.760	2
lambda=0.8	0.0943	9920.825	-46.5724	8379.271	5
lambda=0.512	0.1097	9760.655	-176.1815	8236.298	9
lambda=0.4096	0.1186	9681.121	-188.2727	8154.101	18
lambda=0.3277	0.1345	9575.735	-23.0230	8006.640	56
lambda=0.2621	0.1606	9517.561	820.2891	7765.579	180
lambda=0.2097	0.1552	9897.528	2517.8707	7815.453	357
lambda=0.1678	0.1735	10021.138	3945.4843	7646.480	523

EPoC R package

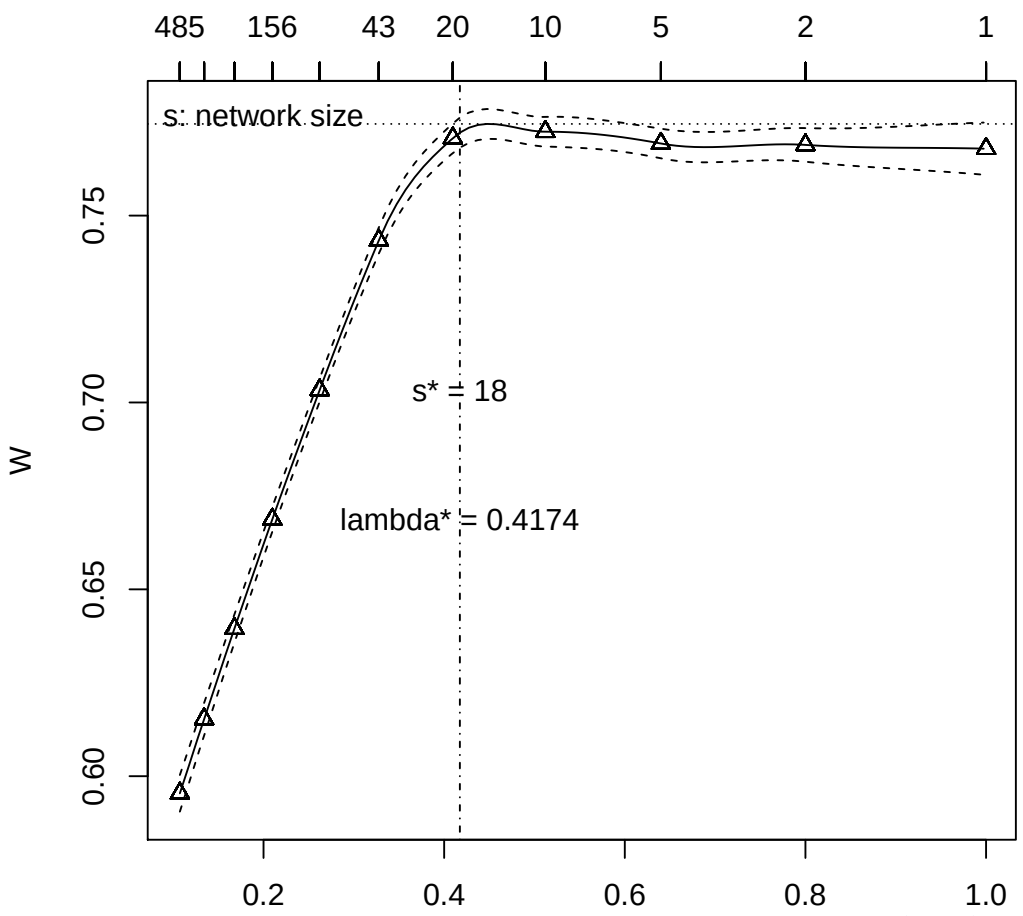
```
> plot.modelsel(G)
> plot(G, k = which.min(G$BIC))
```





EPoC R package

```
> W <- epoc.validation(type = "concordance", y, u
      + , repl = 20)
> plot(W)
```

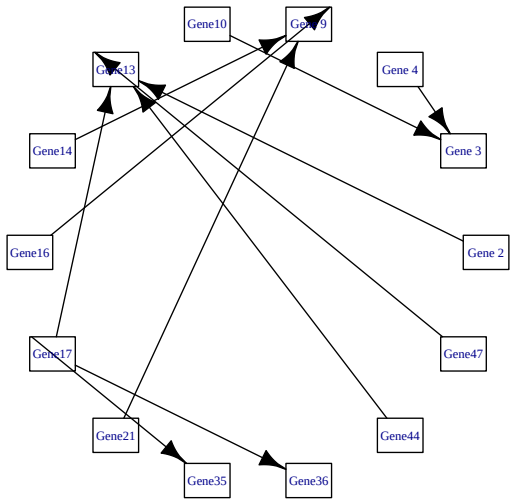


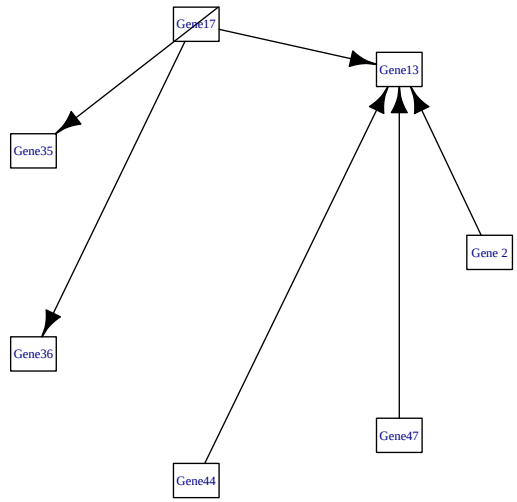
EPoC R package

`epoc.bootstrap, epoc.final,..`

We arrive at a final network G which we can use for survival analysis.

```
> G.svd <- epoc.svd(G.final, C = 3, numload = c(10, 10,
+     10))
> epoc.svdplot(G.svd, C = 1)
```



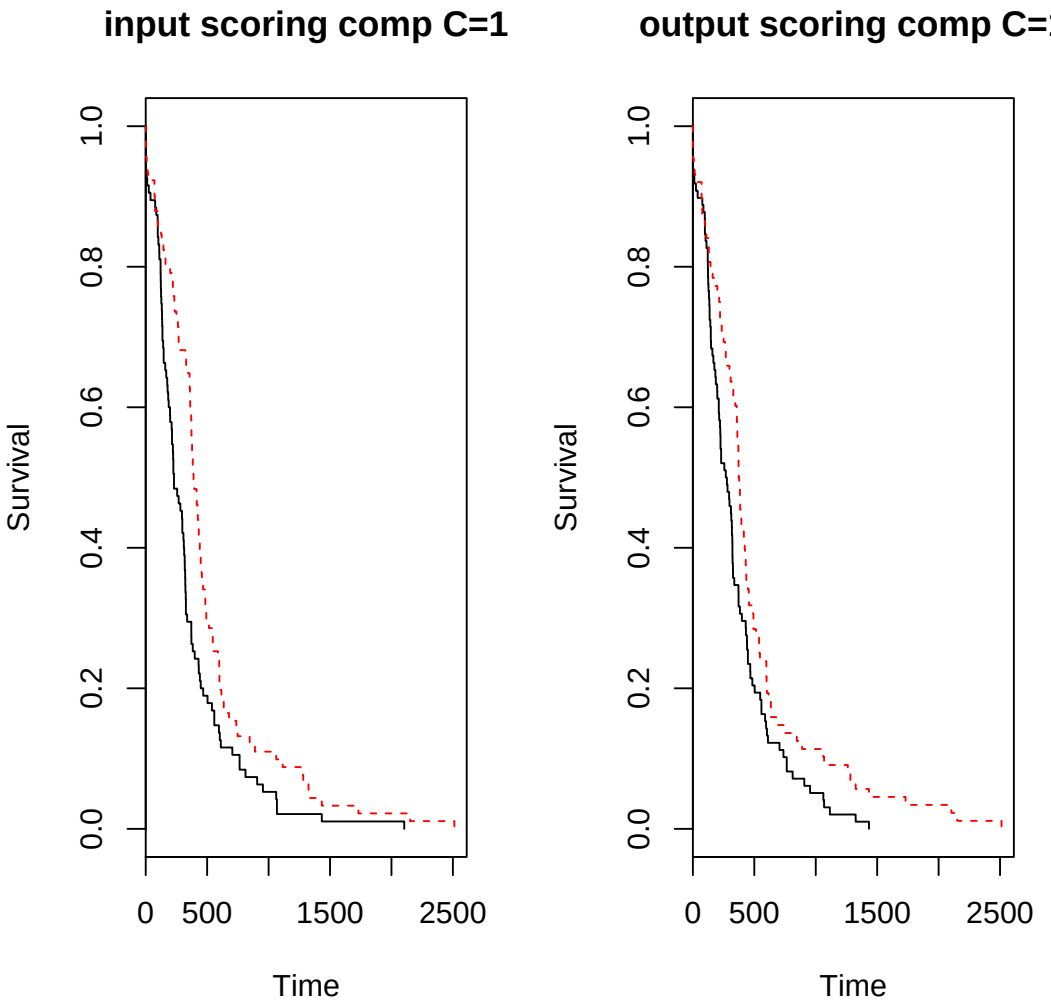


EPoC R package


```
> G.surv <- epoc.survival(G.svd, y, u, surv, C = 1, type = "G")
> summary(G.surv)
           In
Call:
survdiff(formula = Surv(surv) ~ sign(sc.in))
                 N Observed Expected (O-E)^2/E (O-E)^2/V
sign(sc.in)=-1 98       98     78.1      5.09      9.12
sign(sc.in)=1  88       88    107.9      3.68      9.12

 Chisq= 9.1  on 1 degrees of freedom, p= 0.00253

           Out
Call:
survdiff(formula = Surv(surv) ~ sign(sc.out))
                  N Observed Expected (O-E)^2/E (O-E)^2/V
sign(sc.out)=-1 96       96     72.1      7.93      13.7
sign(sc.out)=1  90       90    113.9      5.02      13.7

 Chisq= 13.7  on 1 degrees of freedom, p= 0.000209
```

EPoC R package

```
> plot(G.surv)
```



Conclusions

- EPoC scales to 10000 genes and produces stable network estimates
- Attained network models exhibit good agreement with pathway databases
- Experimental validation of novel hubs identify interesting therapeutic targets
- The EPoC network provides clinical stratification into long- and short-term survival, whereas competing methods do not.

Future Work

- Extend EPoC to tumor subtype identification (promising results already)
- Common and subtype specific network modules (ongoing work with PhD students in my group)
- Include multiple data sources (e.g. miRNA, methylation)
- Methodological work - supervised prognostic network estimation

Acknowledgements

- Sven Nelander
- Tobias Abenius
- Teresia Kling, Linnea Schmidt, Bodil Nordlander, Erik Johansson, Torbjörn Nordling, Chris Sander, Björn Nilsson, Peter Gennemark, Keiko Funa, Linda Lindahl
- Cancerfonden, Barncancerfonden, Vetenskapsradet, BioCare, Sahlgrenska-CMR, NB-CNS