

Construction of malaria gene expression network using partial correlations

Raya Khanin and Ernst Wit

Department of Statistics

University of Glasgow, UK

www.stats.gla.ac.uk/~raya/suppldata.html

The analytical objective

- Construct gene expression network of *P.falciparum*
- Study global topological structure of constructed network
- **Motivation:** Obtain clues on putative roles of genes with unknown functions based on their position in network
- 60% of genes lack sequence similarity with any other organism
- 65% of annotated genes encode proteins of unknown functions

Co-expression networks

- **Two genes are linked if their standard correlation is higher than threshold** (*Bergmann et al, 2004; van Noort et al, 2004*):
- **Results**
 - a few hubs with many links
 - many nodes with a few links
 - correlation between essentiality (lethality) and connectivity of a gene

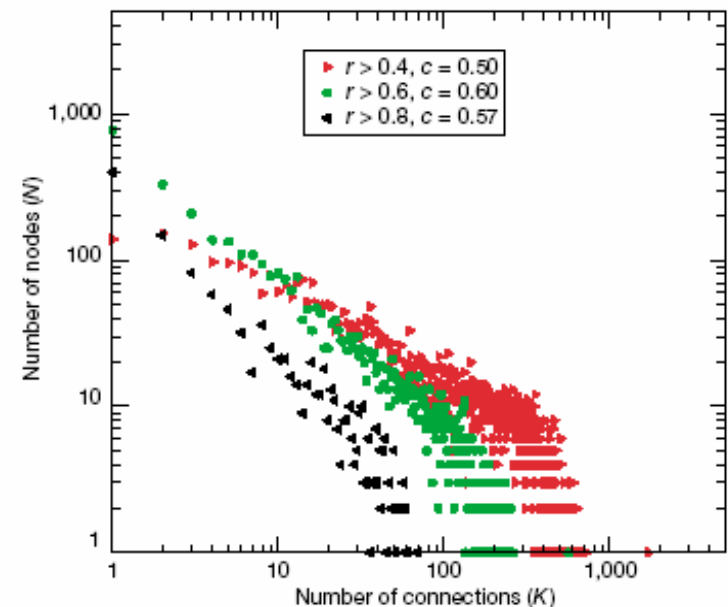


Fig 1 | Distribution of connections per node in the coexpression network. Nodes are genes and connections are defined by coexpression of two genes, resulting in a network. The number of nodes (N) with a certain number of connections (k) in the coexpression network is shown, where coexpression is defined by a correlation in expression pattern higher than 0.4 (right-pointing arrows), 0.6 (circles) or 0.8 (left-pointing arrows). The distributions at thresholds 0.6 and 0.8 are scale free with an exponent $\gamma \approx 1$.

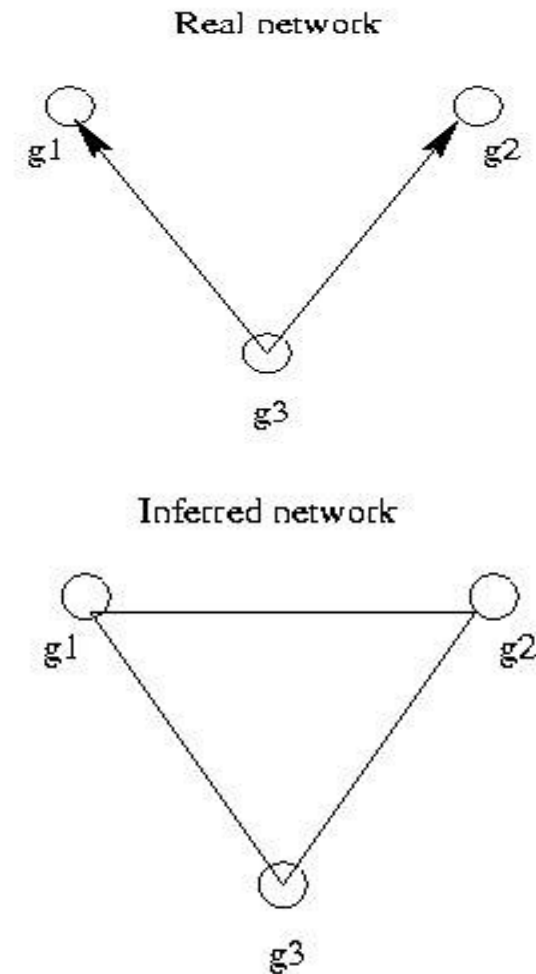
How scale-free?

- Proposed model: **Scale-free network**
 - It indicates the absence of a typical node in the network
 - Scale-free networks are characterized by a **power-law** distribution: $P(k) \sim k^{-\gamma}$

 - We found MLE γ for 10 published interaction datasets
 - By performing *goodness-of-fit tests based on chi-squared distribution*, we concluded
- all networks significantly differ from scale-free behaviour.**

Limitations of co-expression networks approach

- Overestimates the number of connections: not only nodes with direct connections but also nodes with indirect connections are included:
- If threshold is not high enough, some connections are left out.
- If threshold is too low, the number of random connections increases.

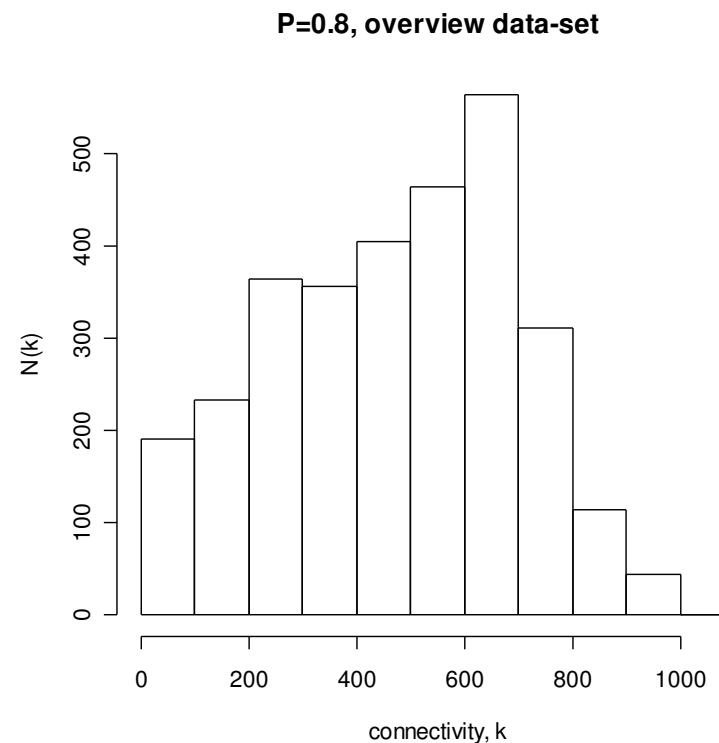


P.falciparum datasets

- **Overview dataset** (3048 genes) from the complete intraerythrocytic developmental cycle (46 time-points)
 - remove genes with more than 50% missing values
 - impute other missing values using R-package `impute()`
 - average the values for multiple oligonucleotides
- **Validation dataset** (2234 genes) from human and mosquito stages of malaria parasite cycle (9 time-points; *Le Roch et al, Science, 2003*; dataset was used for clustering gene expression profiles)

Limitations of co-expression networks approach to malaria dataset

- Trying to impose sparseness results in a **very high threshold values**, p : $\langle k \rangle = 50$, $p = 0.935$ and $\langle k \rangle = 30$, $p = 0.95$. These values of p are too high and many links will not be included.
- For $p = 0.8$, the constructed network is **not sparse**, $\langle k \rangle = 470$, and the network **topology is different** from other known networks.

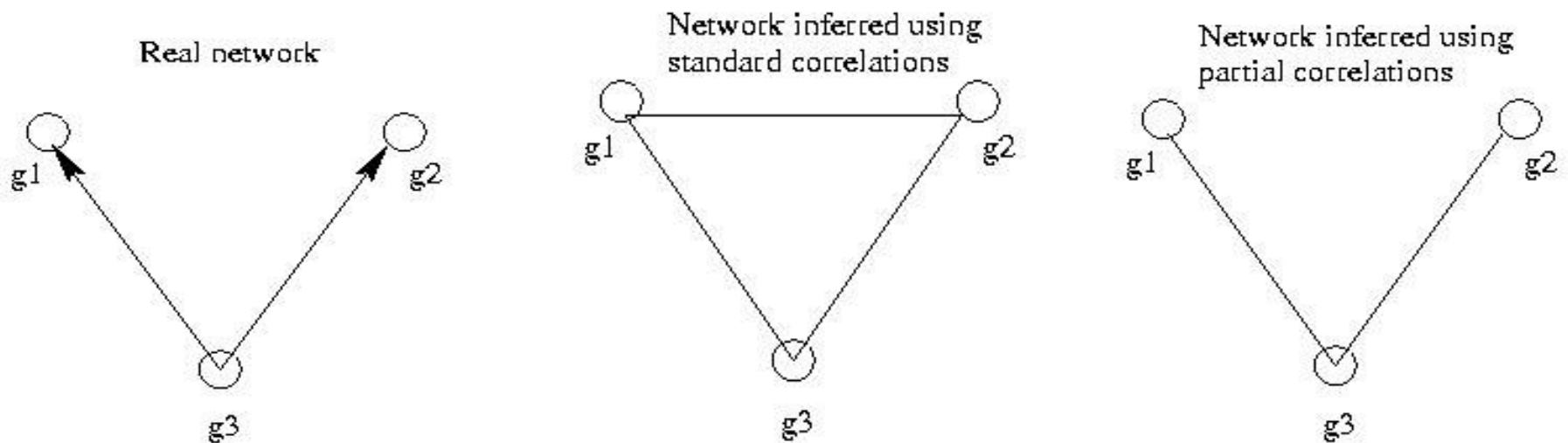


Using partial correlations

- We propose to use partial correlations to filter the more likely links from a larger set of potential links with high correlations.
- Partial correlation of genes i and j with respect to all other genes whose effect is removed (fixed) is given by

$$r_{ij} = \frac{\omega_{ij}}{\sqrt{\omega_{ii} \omega_{jj}}}$$

$\Omega = P^{-1} = \omega_{ij}$ is the inverse of correlation matrix.



Other methods based on partial correlations

- Partial correlations have been used in Graphical Gaussian Modelling
- First-order partial correlations (*Wille et al, 2004*)
- Second-order partial correlations (*de la Fuente et al, 2004*)
for each gene pair they consider effect of a third gene (or a pair of genes) separately; the edge is drawn when the pair-wise correlation is not the effect of any of other genes.
- Full-order partial correlations (*Schafer and Strimmer, 2004*)
developed estimators of partial correlations for small samples and fitted network using FDR.

Methodology

- Genes i and j are connected if their **standard and partial correlations** are higher than their respective cut-off values:

$$i \leftrightarrow j : p_{ij} \geq p \ \& \ r_{ij} \geq r$$

- Pearson correlation matrix P for small samples is degenerate and pseudo-inverse of correlation matrix was used

Schafer and Strimmer, 2004. `cor2pcor()` function from R-package GeneTS: <http://www.stat.uni-muenchen.de/~strimmer/genets/>

Criteria for choosing cut-off parameters

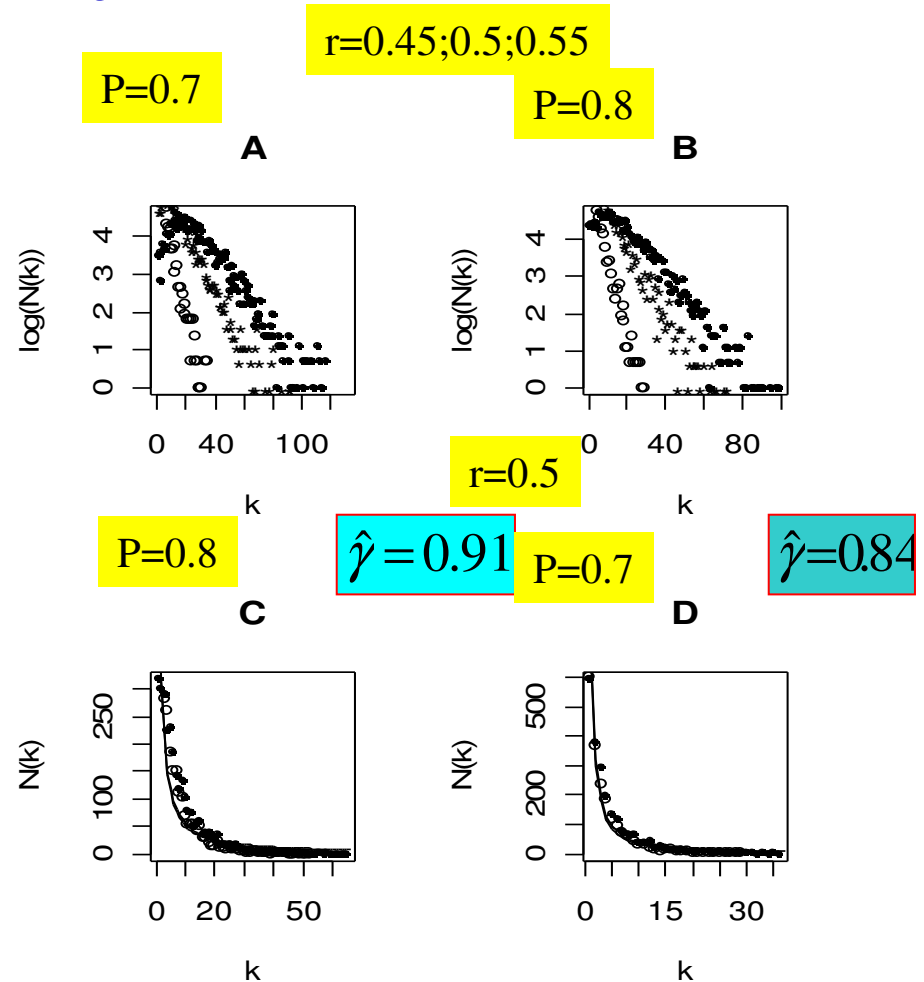
Choose cut-off parameters p and r to satisfy four criteria:

- **Small-world property:** clustering coefficient C is much higher than that of random network ≈ 0.005 . (C is measure of extent that genes, connected to a specific gene, are linked among themselves)
- **Network sparseness:** average connectivity $\langle k \rangle$ of order 10-30.
- **Connectivity drop-off rate:** power exponent: $\hat{\gamma} \in (0.5, 2)$
- **Scale-free chi-squared statistic** (as low as possible)

Results: connectivity distribution

- Topologies of constructed networks are consistent with other reported networks: a **few hubs and many genes with few links**.
- Qualitatively, topology does not depend on exact values within a region:

$$0.45 \leq r \leq 0.6, 0.7 \leq p \leq 0.8$$
- Values outside this region result in other types of network topology.



- **We use $p=0.7$, $r=0.5$:**
 $\langle k \rangle = 15$, $\max(k) = 101$, $\langle C \rangle = 0.2$

Validation of constructed network

- **Permutation Test**
 - Independent permutation of components of each gene profile
 - Recomputing correlation and partial correlation matrices
 - Establishing a link if the thresholding conditions are satisfied
 - 100 permutation tests resulted in 200 p-values=0.01 with the rest being zero
 - FDR procedure with 10% control level resulted in all links found by thresholding procedure from overview dataset being significant
- **Proof-of-principle results**

Connectivity and essentiality

- Top **66 hubs** of the network constructed from the **overview dataset** ($p=0.7, r=0.5$):
 - 13 with no annotation, 7 on plastid genome
 - **7 genes** are known to have the cell essential functions in cell growth and/or maintenance, metabolism, energy pathways, biosynthesis
 - 35% percent of all annotated genes encode proteins with identifiable function (~16 genes)
 - **8 genes** are either conserved or have homologues to proteins in other organisms
- Top **66 hubs** constructed from **validation dataset** ($p=0.8, r=0.5$) contain **20** (virtually all annotated genes in the list) with essential cell functions
- 50% of **66 hubs** (excluding plastid) are in the 6% of genes that were found to be common to all four stages of the parasite life cycle (*Florens et al, 2002*)

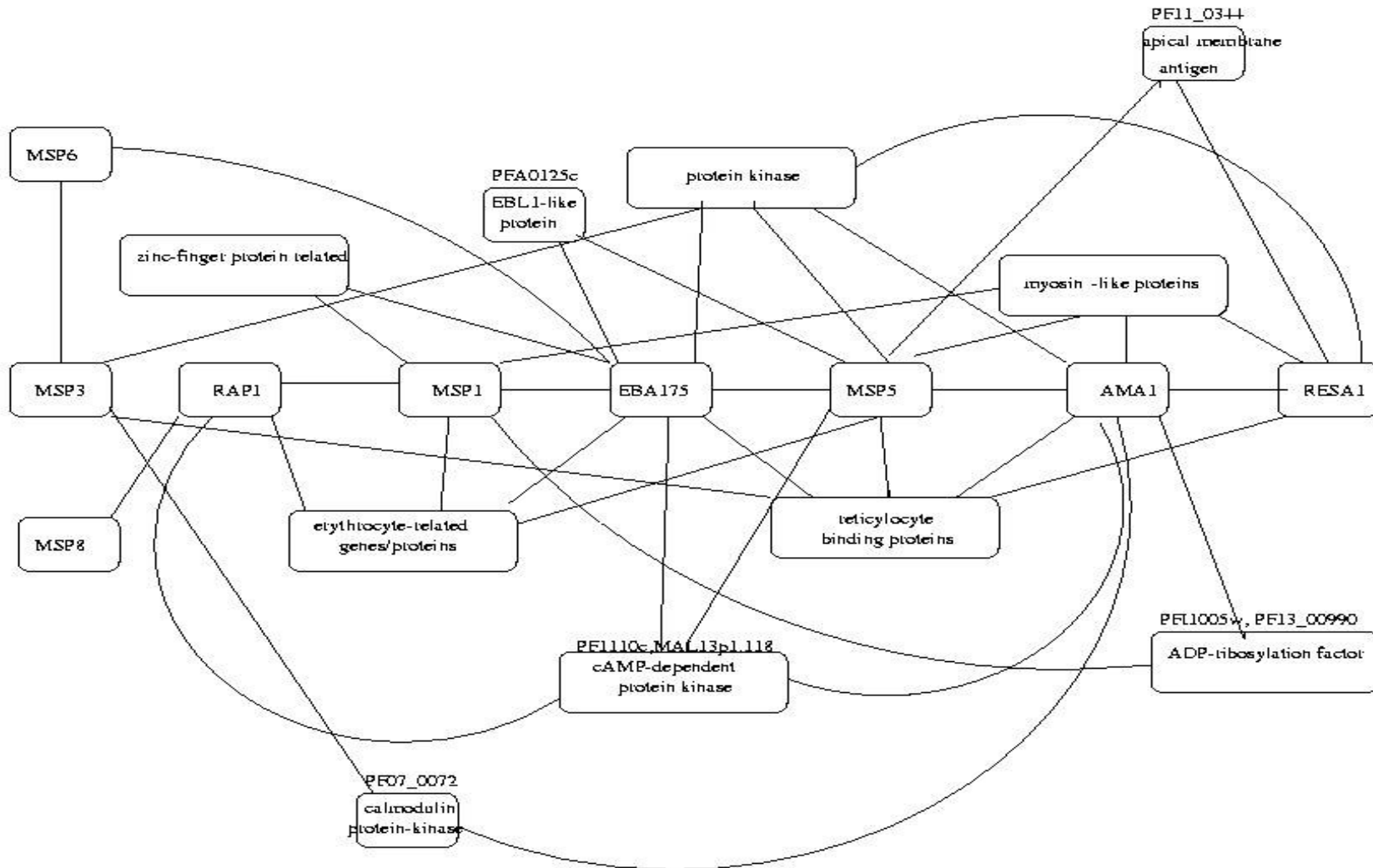
Gene with unknown functionalities

How **25 hubs with unknown functions** clustered in the validation dataset of *Le Roch et al (2003)*:

- 10 genes belong to cluster 13; 5 genes belong to cluster 12, 5 genes belong to cluster 15:
 - Clusters 12,13 are mainly involved in **cell-cycle regulation** and progression to trophozoite stage
 - Cluster 15 contains genes with roles in **cell invasion** that are under evaluation as blood-stage vaccine
- According to Le Roch et al (2003) “genes from the clusters 12,13 may represent potential targets for drugs focused on disruption of the trophozoite stage, while additional candidate vaccine antigens could come from yet uncharacterized genes of the cluster 15.”

Hubs with unknown functionalities warrant further investigation

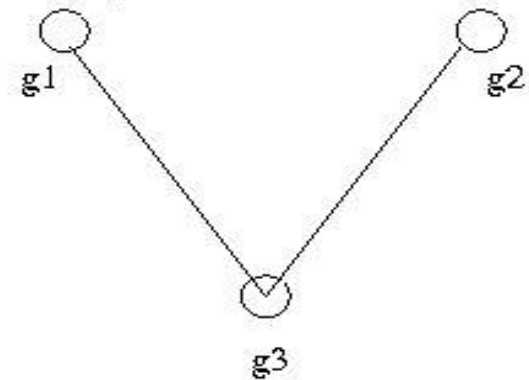
Major candidates for vaccination



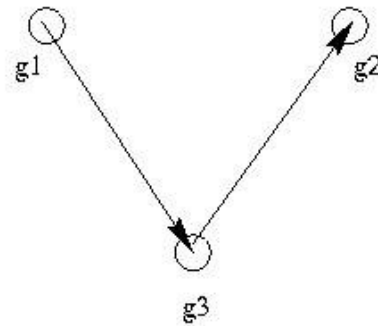
Limitations of our approach

- Link between two genes does not imply **causality** (undirected network)
- Network fitting methods should be based on **multiple testing procedures**.
- **Machine learning techniques** could be a viable alternative.

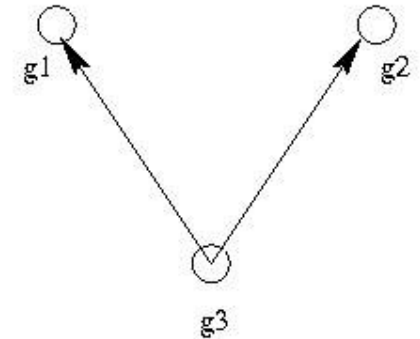
Network inferred using partial correlations



Real network



Real network



Conclusions

- The constructed network is a small world networks with topology similar to other studied networks and hubs being enriched by essential genes
- Biological conclusions from network look promising.
- More information
www.stats.gla.ac.uk/~raya/suppldata.html