Coexpression Analysis of Human Genes Across Many Microarray Data Sets

Jclub 04.06.2007

Priit Adler

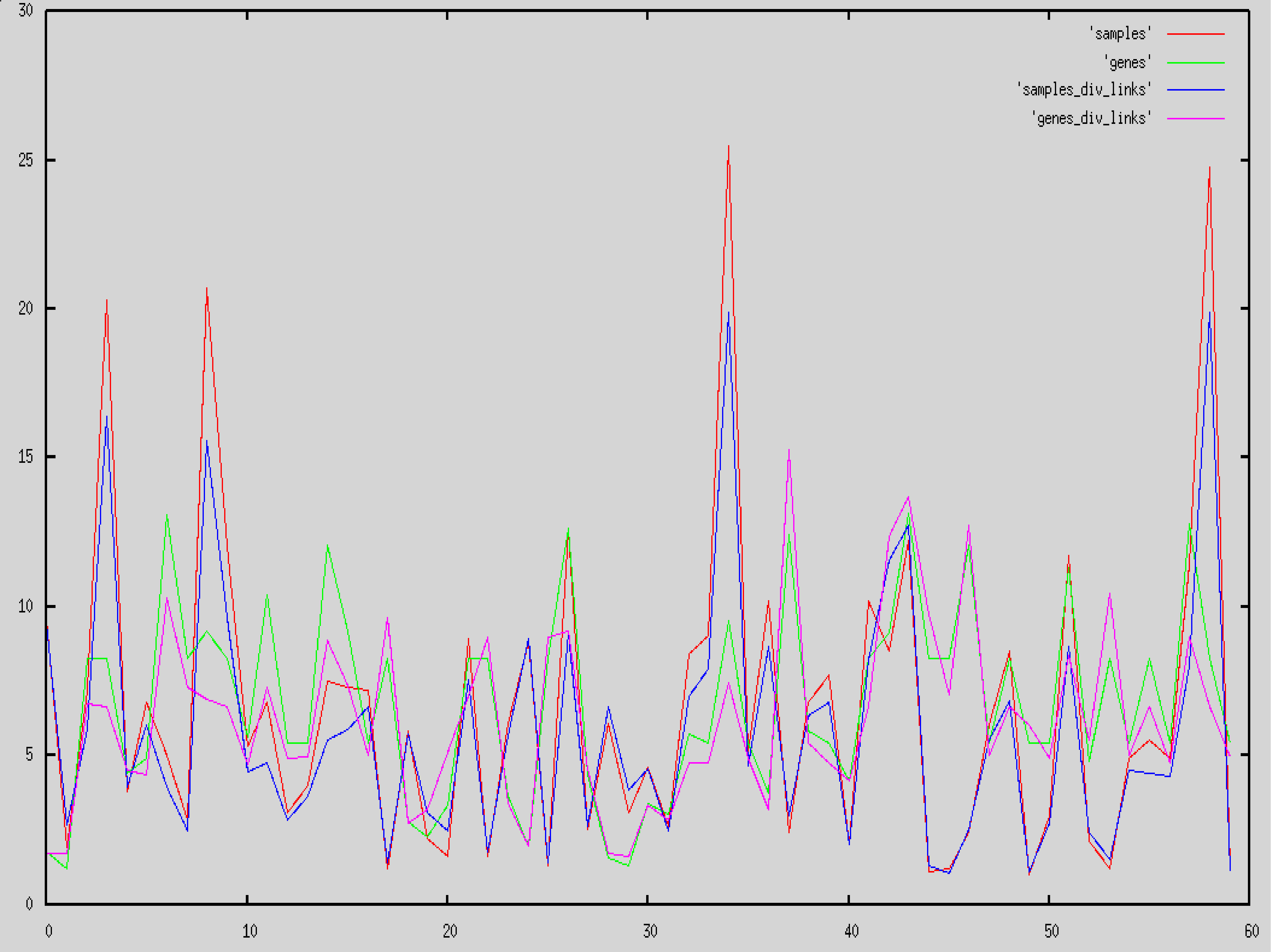
Coexpression Analysis of Human Genes Across Many Microarray Data Sets

Homin K. Lee, Amy K. Hsu, Jon Sajdak, Jie Qin and Paul Pavlidis

Genome Res. 2004 Jun;14(6):1085-94.

Experiment buildup

- **finding repeating co-expression links from different expression data sets.**
- 60 data sets
- find co-expression links in every data set independently
  - does dataset size and number of links correlate ?? (next slide)
  - using only statistically significant links
- count repeating links in different data sets
- use only links found in more than tree data sets (statistical issue!)

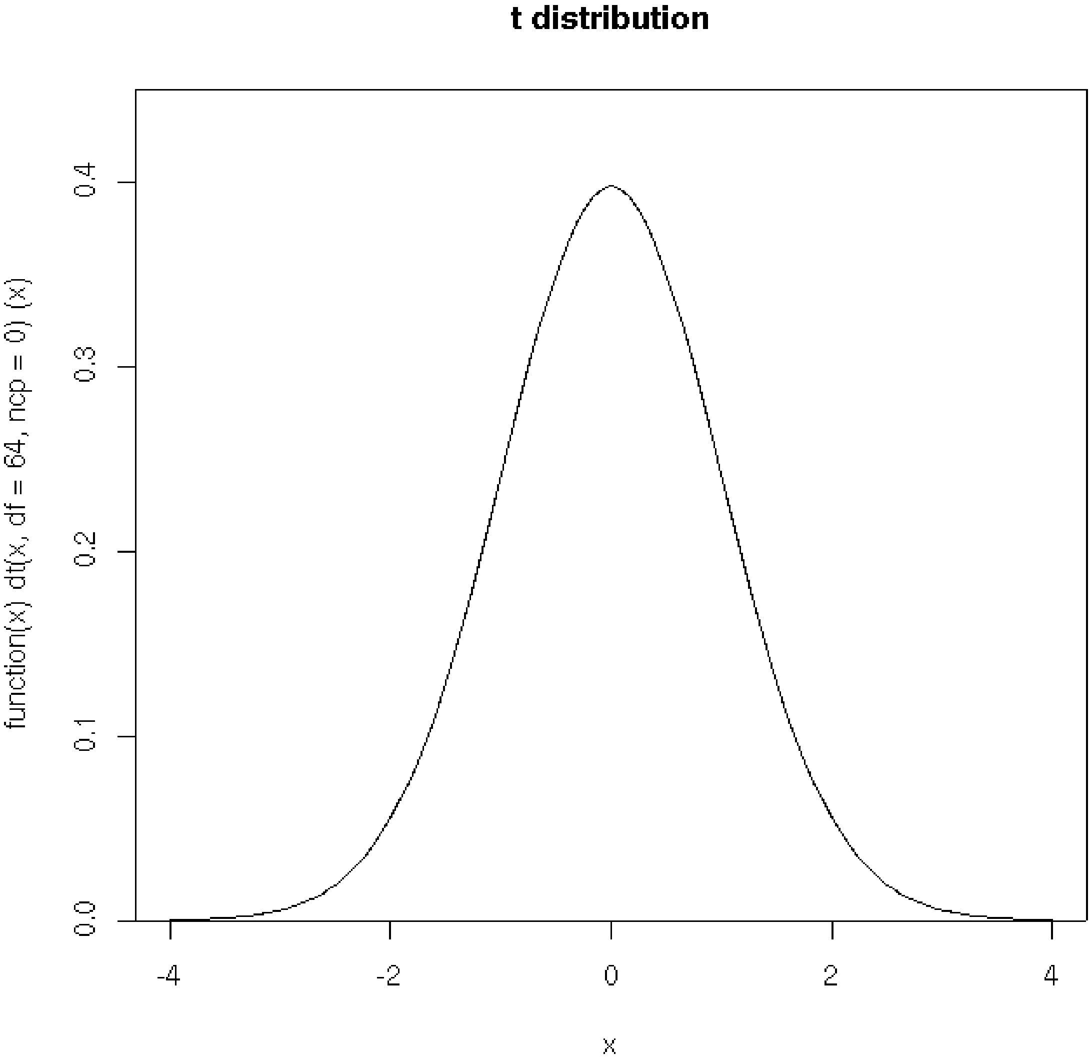


'samples'
'genes'
'samples_div_links'
'genes_div_links'
30
25
20
15
10
5
0
0
10
20
30
40
50
60

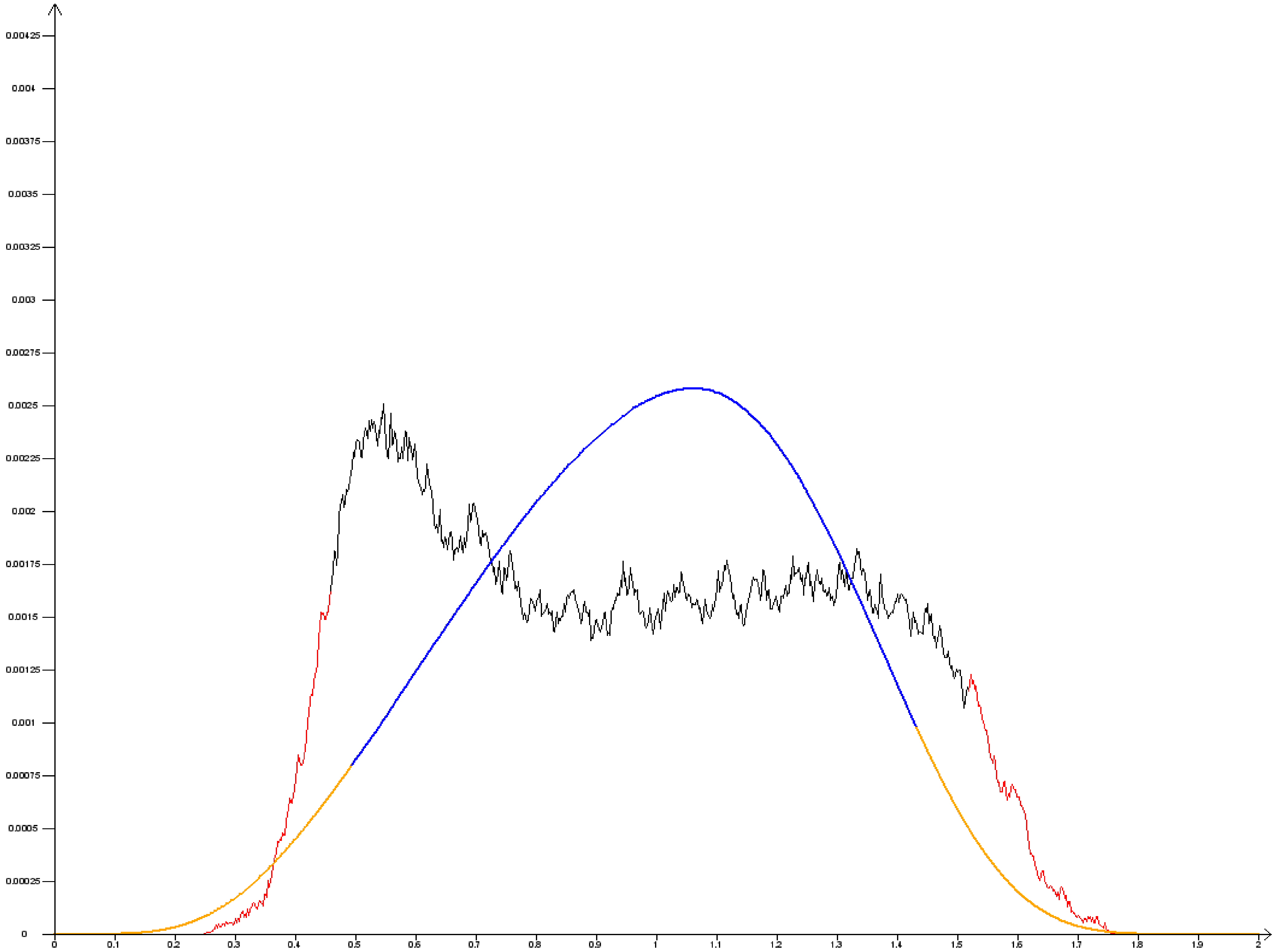
Co-expression link identification

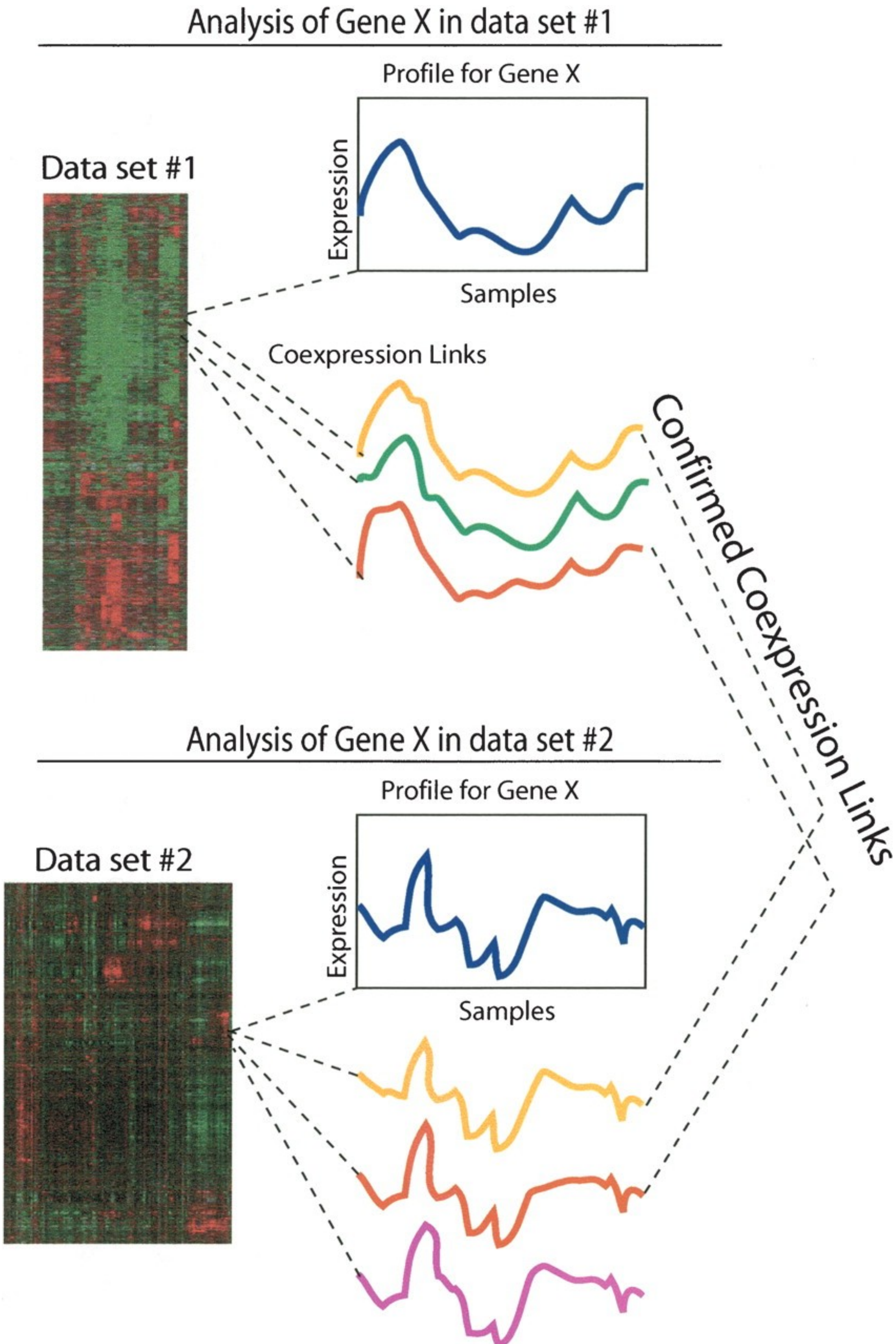
- Using *t* distribution with *n* - 2 degrees of freedom, where *n* is number of samples in data set.
- ***Null* hypothesis** - two genes have nothing in common if the correlation between them corresponds to above distribution.
- *P*-values were corrected using Bonferroni correction with α=0.01
- In addition only links among 0.5 % top or bottom were concidered for further study.
- alltogether 9.7 M “raw links”

t distribution



- *t* distribution, with *n* - 2 degrees of freedom
- *n* = 66 (average from data sets in article)


0.00425
0.004
0.00375
0.0035
0.00325
0.003
0.00275
0.0025
0.00225
0.002
0.00175
0.0015
0.00125
0.001
0.00075
0.0005
0.00025
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8
1.9
2


Analysis of Gene X in data set #1
Profile for Gene X
Data set #1
Expression
Samples
Coexpression Links
Confirmed Coexpression Links
Analysis of Gene X in data set #2
Profile for Gene X
Data set #2
Expression
Samples

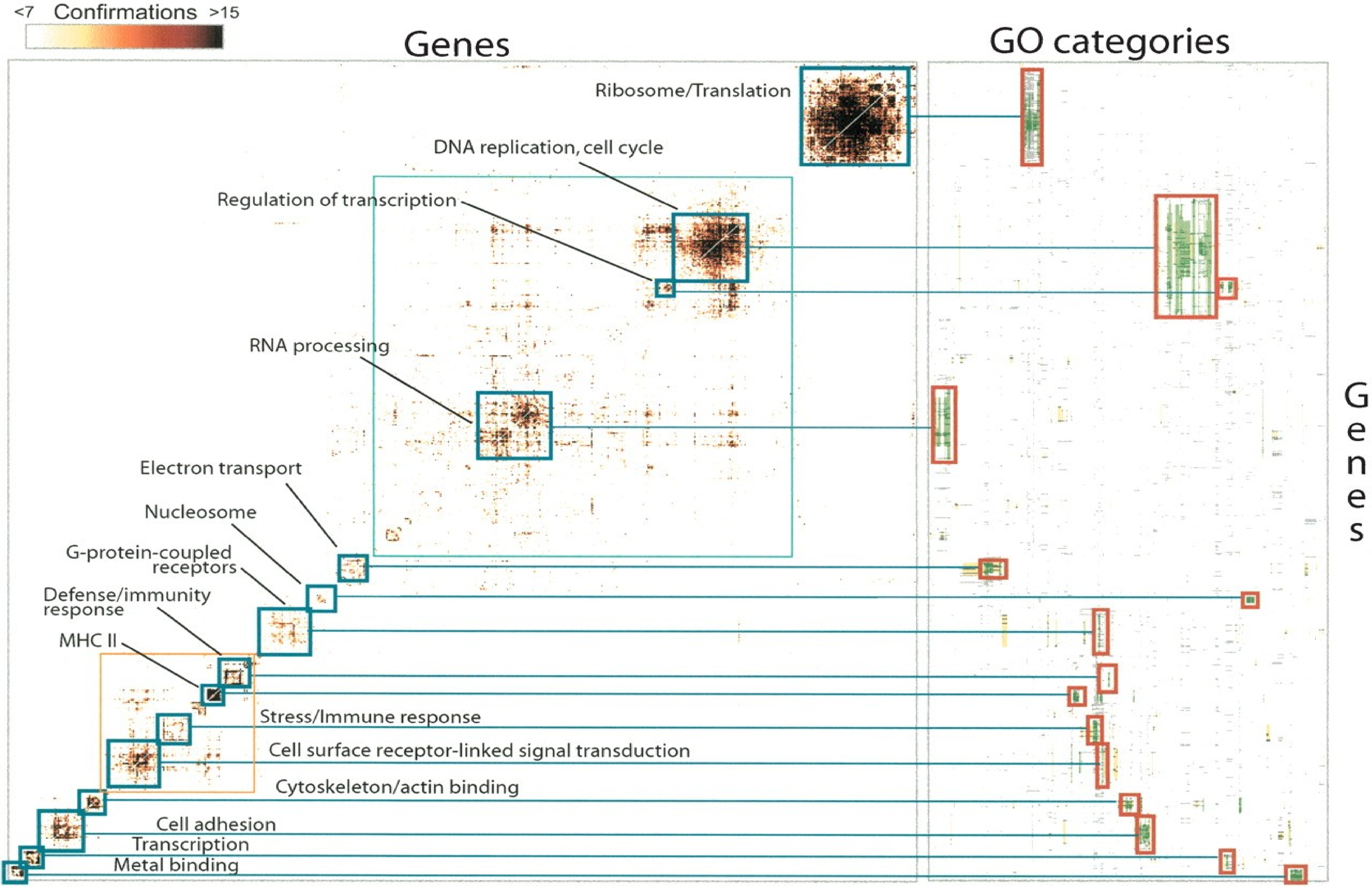
Link comfirmation

- “shuffled” data test
  - about the same number of links in each “shuffled” data set as origin
  - about the same number of positive and negative correlation
    - (positive correlations in real data more than negative)!
  - 100 such datasets
- the results:
  - **~ 5 % of “3+” links found in real data could also be found in randomized data**
  - ~ 24 % of “2+” links found in real data could also be found in randomized data

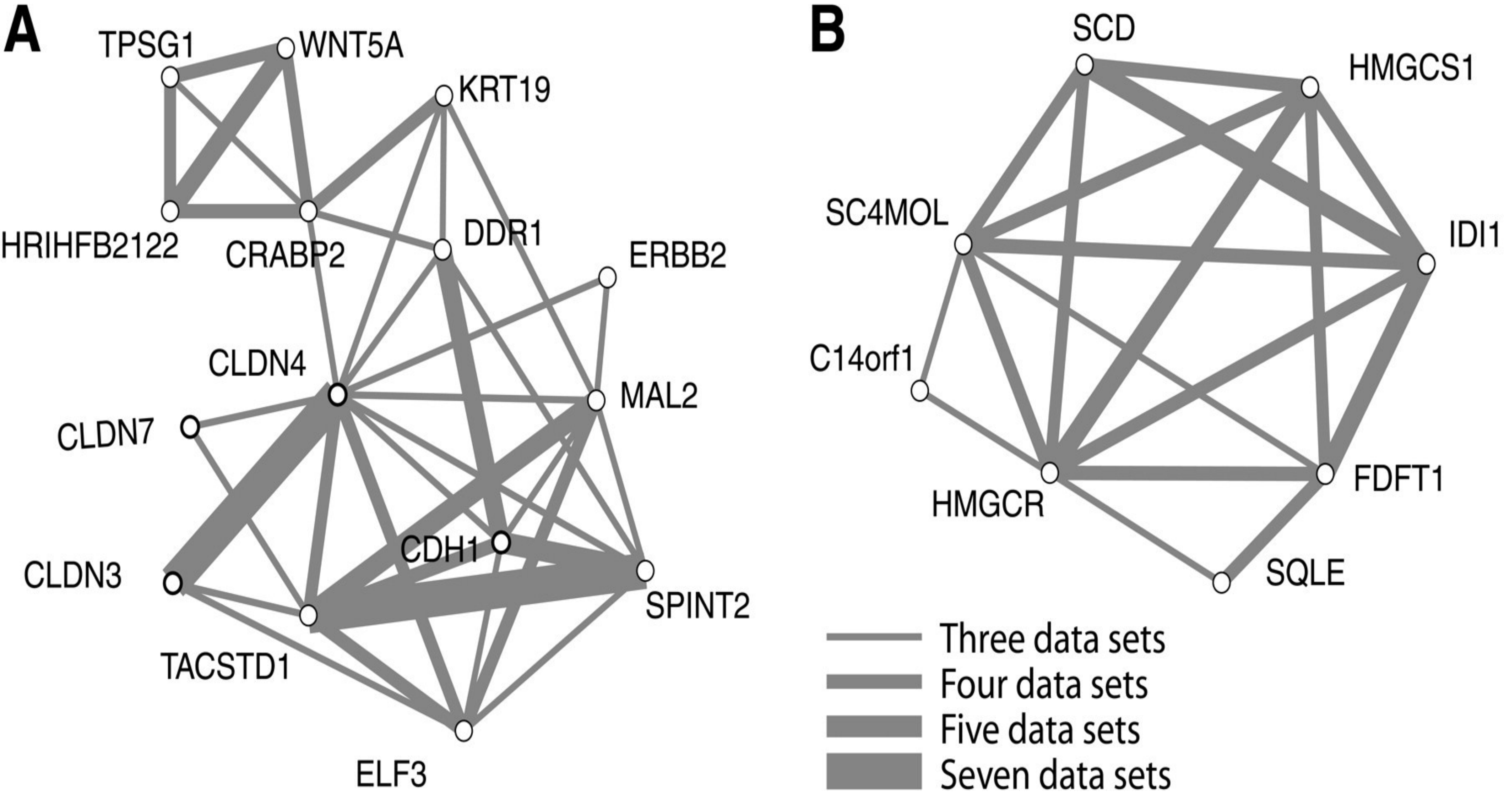
results

- from 9.7 M only 220 k of links were found in “3+” data sets
- 8805 genes of 14.172 compared genes have atleast one “3+” link (60 %)
- also a GO analysis, but not used for confimation, therefore not interestin in this article ...
- For cluster analysis only “7+” links were used
  - 720 genes
  - 10 089 links

results


<7
Confirmations
>15
Genes
GO categories
Ribosome/Translation
DNA replication, cell cycle
Regulation of transcription
RNA processing
Electron transport
Nucleosome
G-protein-coupled receptors
Defense/immunity response
MHC II
Stress/Immune response
Cell surface receptor-linked signal transduction
Cytoskeleton/actin binding
Cell adhesion
Transcription
Metal binding
Genes

To visualize smaller network bits



For discussion !

- Does this kind of statistical schema is good enough ?
- Could there be some better, improved, schema ?
  - Instead of *t* distribution use the distribution of **all distances** in given data set
  - or is it not statistical ?

tnx

- any other questions ?
- comments ?
- suggestions ?