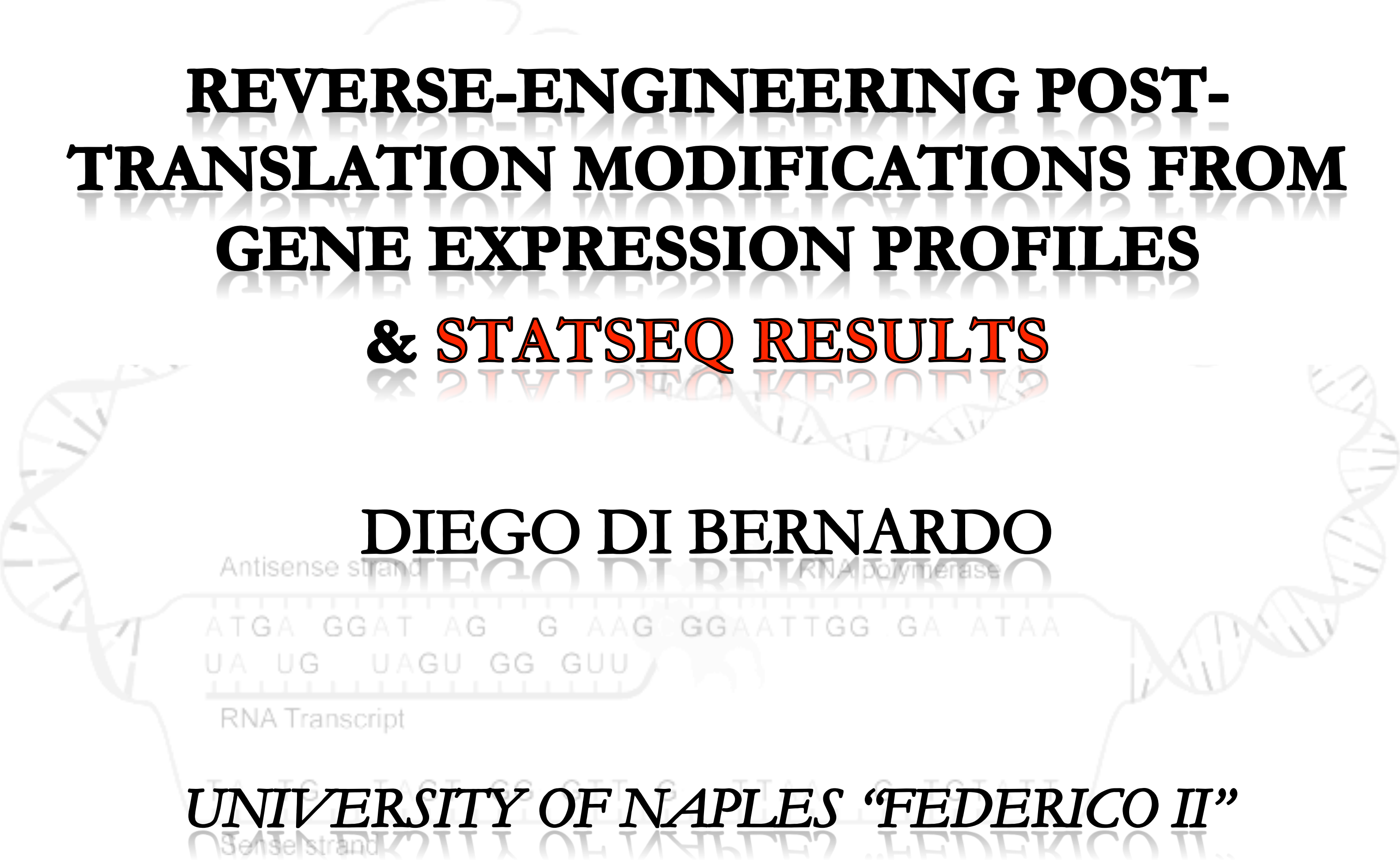
REVERSE-ENGINEERING POST-TRANSLATION MODIFICATIONS FROM GENE EXPRESSION PROFILES

& STATSEQ RESULTS



DIEGO DI BERNARDO

UNIVERSITY OF NAPLES "FEDERICO II"

tigem Telethon Institute of Genetics and Medicine TeleThon

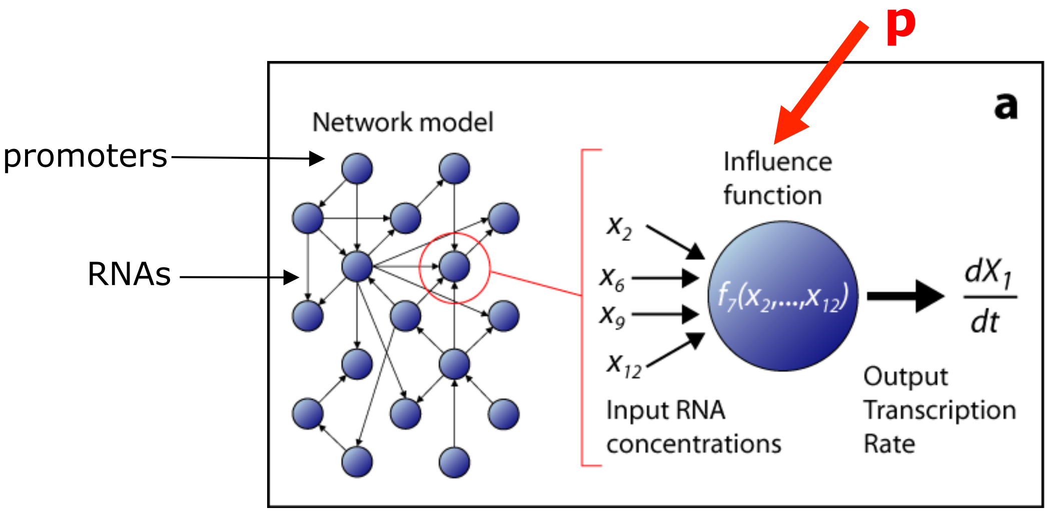
The StatSeq Dataset (just to make sure you remember)

- StatSeq consists of 72 datasets originated from 9 different "in silico" gene networks, each simulated under 8 different parameter settings such as population sizes, marker distances, and heritability.

- For each of the 72 datasets there are two matrices:
  - i) the gene expression matrix
  - ii) the genotype matrix which represents the mutated genes.

- The problem is to identify the network topology from the data (reverse-engineering).

To solve the problem: Network Inference by Regression (NIR)



$$dx_1/dt = a_2\, x_2 + a_6\, x_6 + a_9\, x_9 + a_{12}\, x_{12} \;\; + p$$

NIR requires knowledge of the perturbed gene in each experiment but it recovers a DIRECTED NETWORK

Gardner, di Bernardo et al, Science, 2003; Cantone et al, Cell, 2009 – code @ http://dibernardo.tigem.it

For steady-state data the eqs. become:

$$dx_i /dt = a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{iN}x_N + p$$

$$\Downarrow$$

$$0 = a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{iN}x_N + p$$

$$\Downarrow$$

$$a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{iN}x_N = -p$$

A solution can be obtained by linear regression:

- We can solve one gene at a time by writing the eq. for a gene *i* in experiment *1*:

$$a_{i1}x_{11} + a_{i2}x_{21} + \ldots + a_{iN}x_{N1} = -p$$

Assuming we over-express one gene at a time, then we will obtain N experiments. E.g. if we perturbed gene *i* in the 2nd experiment:

$$\begin{aligned}
a_{i1}x_{11} + a_{i2}x_{21} + \ldots + a_{iN}x_{N1} &= 0\\
a_{i1}x_{12} + a_{i2}x_{22} + \ldots + a_{iN}x_{N2} &= \mathbf{1}\\
\vdots \qquad \vdots \qquad\qquad \vdots \quad &\quad \vdots\\
a_{i1}x_{1N} + a_{i2}x_{2N} + \ldots + a_{iN}x_{NN} &= 0
\end{aligned}$$

NxN

How gene 1 regulates gene i

Perturbation vector p

Gene N in all M experiment

This is solved by linear regression with variable selection and assuming *a sparse network*, i.e. **genes (N)<exps (M)**

$$\boxed{[a_{i1}\ldots a_{iN}]^T = X^{-1}\mathbf{p}}$$

Application to StatSeq data:

- i) the gene expression matrix = X

- ii) the genotype matrix which represents the mutated genes =P

  - Assuming that the mutated genes cause a change in expression of the target genes.

  - Assuming a sparse network, i.e. each gene is connected at most to 10 other genes, so that the $\mathbf{a}_i$ vector is of dimension 10.

Results: it works better that MI/Correlation methods.

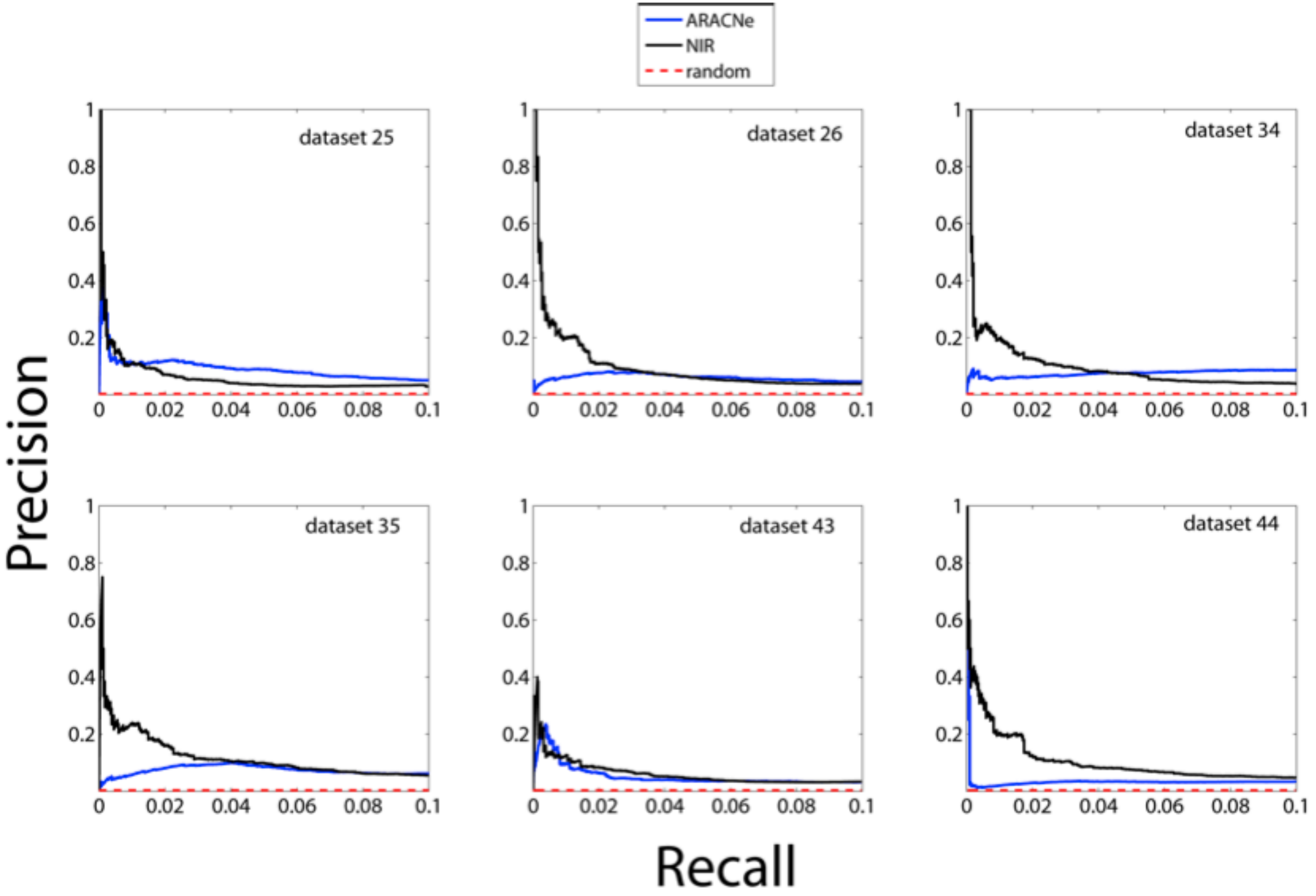


Fig. 3 Precision-Recall curve at 10% of Recall for NIR and ARACNe algorithms. The Precision (TP/(TP+FP)) vs. Recall (TP/(TP+FN)) curve at 10% of Recall for NIR (black line) and ARACNe (blue line) algorithms. Only the first two type of each datasets composed by 1000 genes have been used. The dashed line represents the precision of the random algorithm.

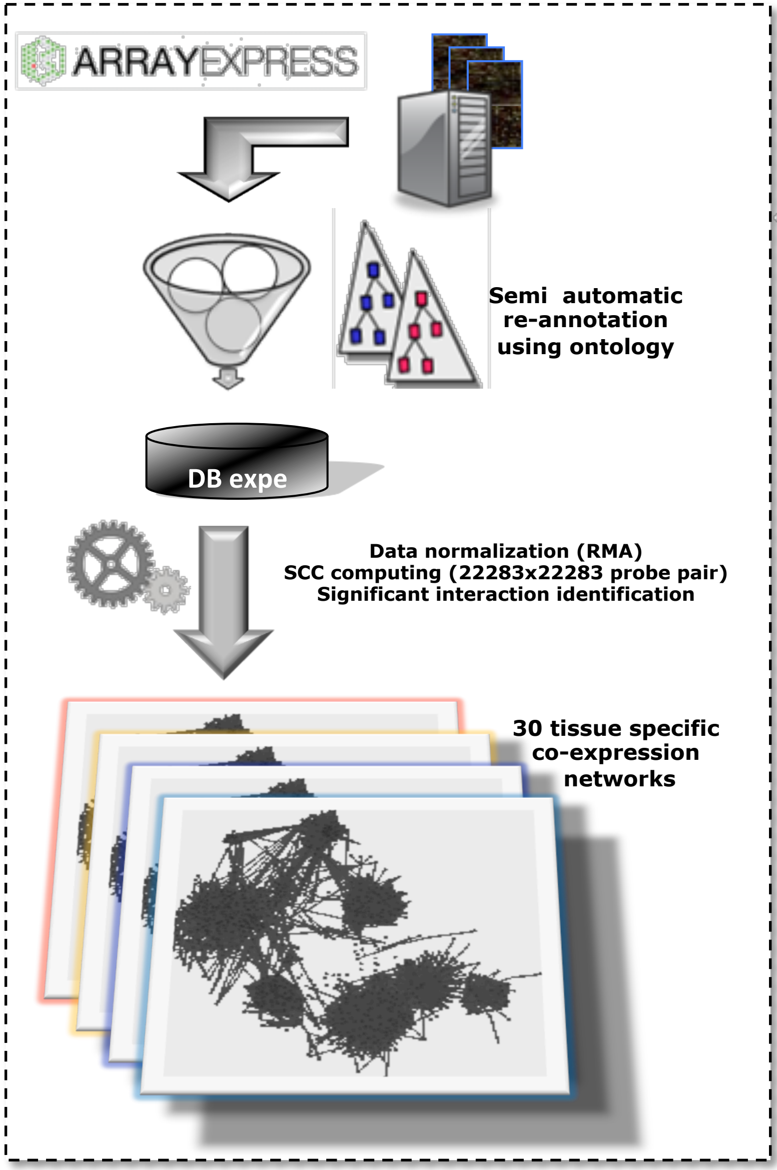
Part II

Differential Network Analysis for the identification of condition-specific pathway activity and regulation

Gennaro Gambardella

Gambardella G. et al, Bioinformatics, under review

Overview of the reverse-engineering strategy (very simple):



Semi automatic re-annotation using ontology

Data normalization (RMA)
SCC computing (22283x22283 probe pair)
Significant interaction identification

30 tissue specific co-expression networks

We built a **database** containing re-annotated microarray experiments for tissues and cell type for HUMAN.

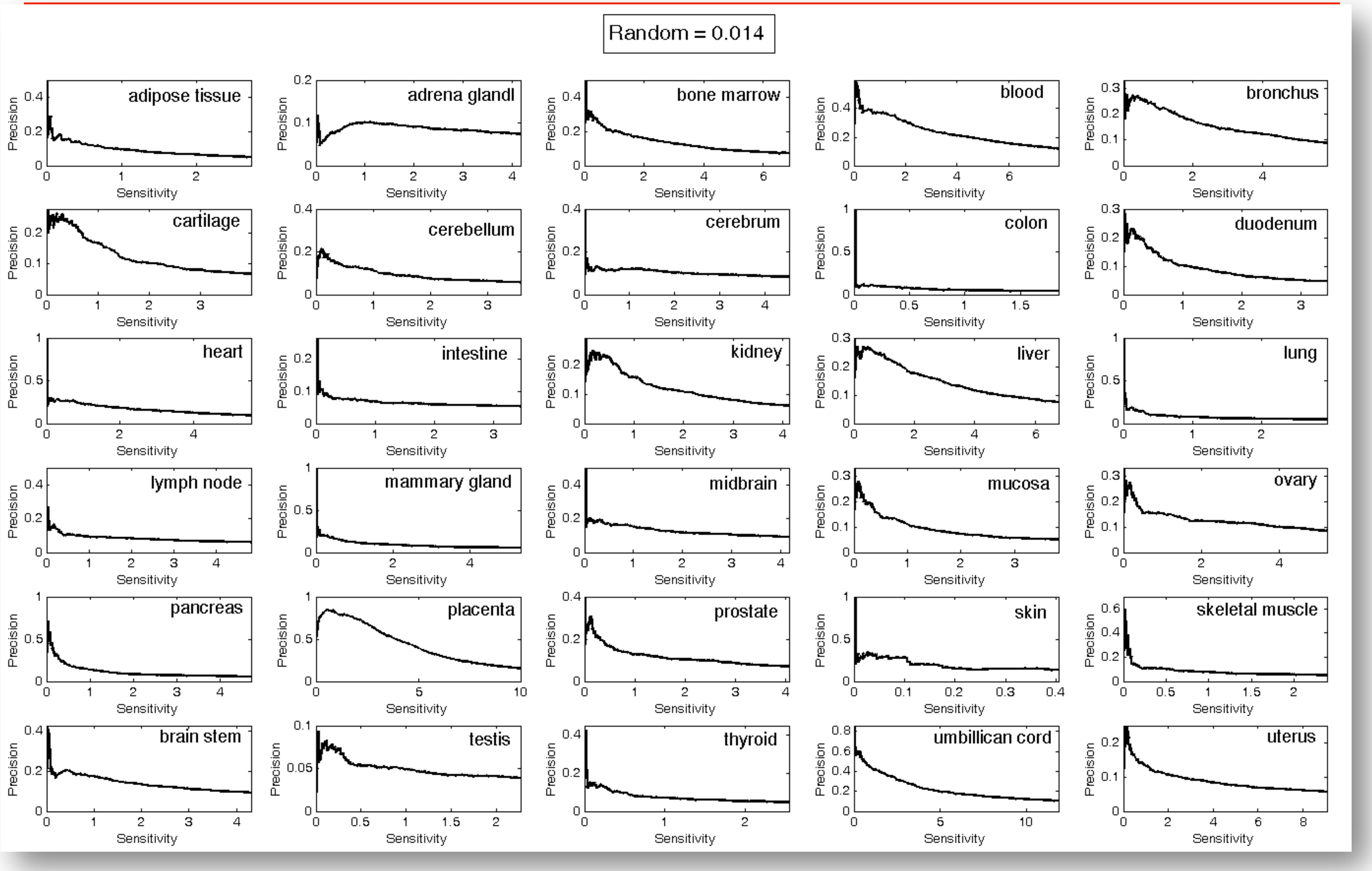
2930 HUMAN microarray hybridizations

22,283 transcripts for each platform.

30 tissue specific co-expression networks using the **S**pearman **C**orrelation **C**oefficient *SCC*.

1. Adipose Tissue
2. Adrenal Gland
3. Bone Marrow
4. Blood
5. Bronchus
6. Cartilage
7. Cerebellum
8. Cerebrum
9. Colon
10. Duodenum
11. Heart
12. Intestine
13. Kidney
14. Liver
15. Lung
16. Lymph Nodes
17. Mammary Gland
18. Mid Brain
19. Mucosa
20. Ovary
21. Pancreas
22. Placenta
23. Prostate
24. Skin
25. Skeletal Muscle
26. Brain Stem
27. Testis
28. Thyroid
29. Umbellican
30. Uterus

Results: Co-expression networks, structure & validation



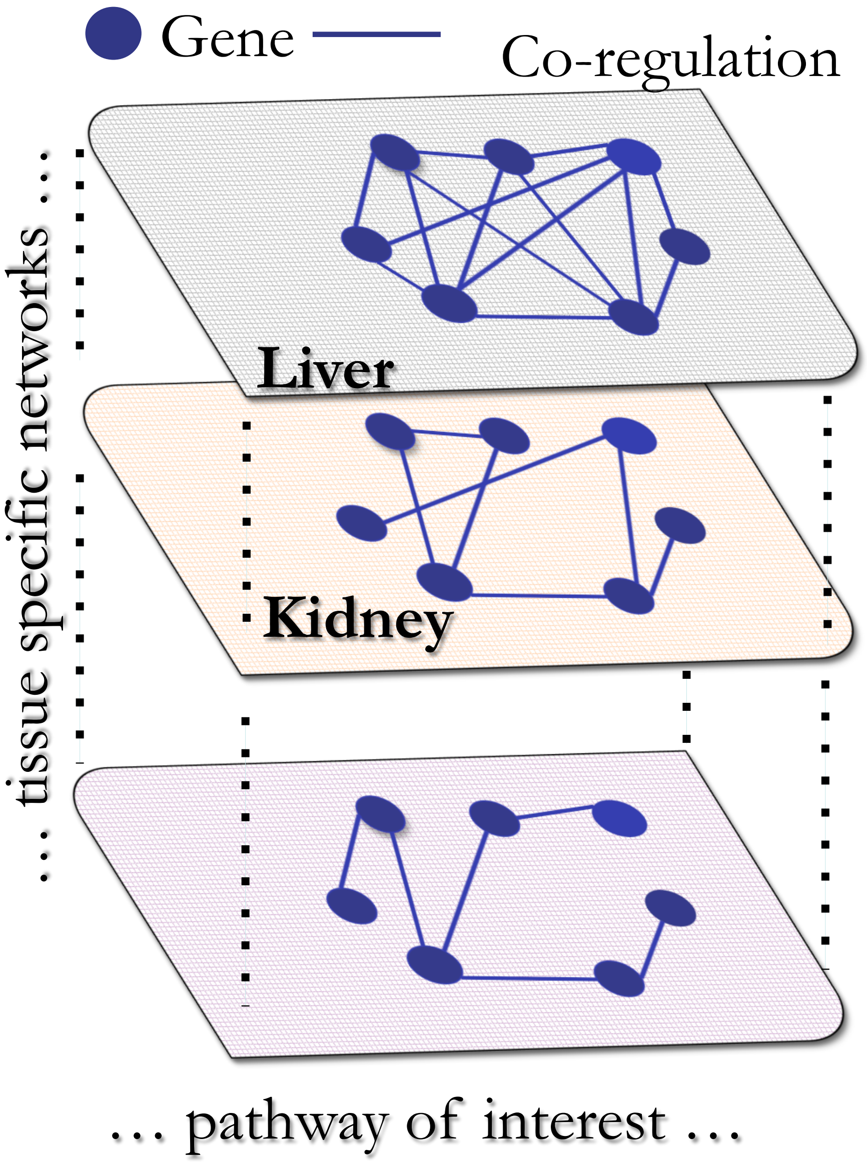
The Golden standard is a mainly composed of about **80,000** experimentally validate interactions **from Reactome database**.

DIfferential Network Analysis can elucidate tissue-specific pathways

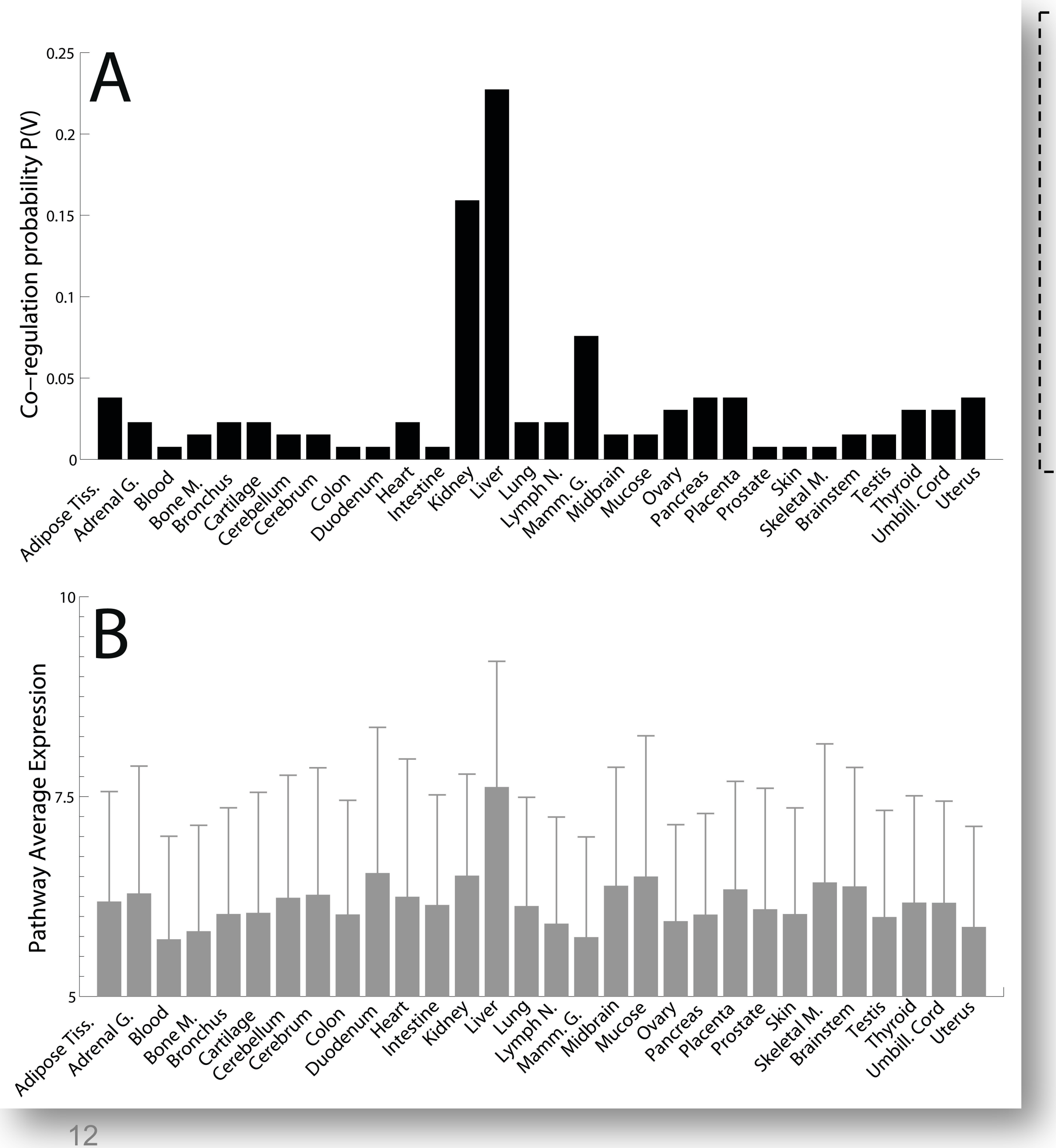
- We developed a network-based algorithm, **DINA**, which is able to identify sets of genes which are significantly co-regulated only in specific conditions.

- The algorithm stars:
  1. **with a set of M genes** and a **set of N networks**.
  2. quantifies how variable **the co-regulation probability** is across the N networks using **an entropy-based measure (H).**

- Its significance is estimated using a Permutation Test.



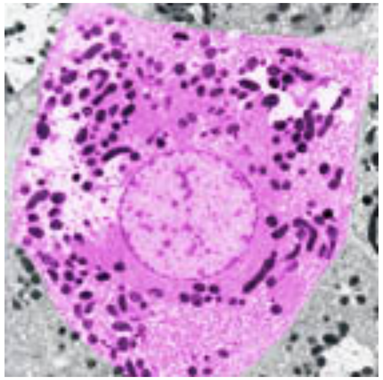
Results: Application to 187 KEGG pathways, the top significant pathways



In order to test whether DINA was, indeed, able to identify tissue-specific pathways we used the full manually curated list of **187 KEGG pathways** from MsigDb.

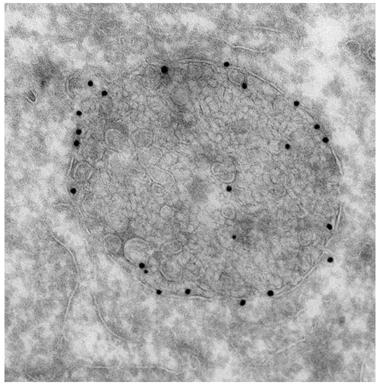
- The **Glycine, serine and threonine metabolism** is present only in liver and kidney.
- **Using only the expression level of the genes in the pathway we would have not obtained the correct answer.**

DINA is able to detect dysregulated pathways in disease

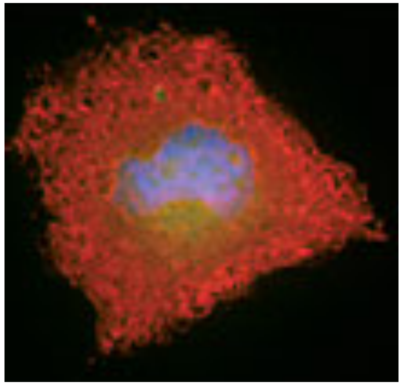


Primary hepatocytes

[wt p53]



HepG2
(initial)
[wt p53]



Huh7
(severe)
[**mt p53**]

Hepatocarcinoma cell lines: a simple model of HCC progression

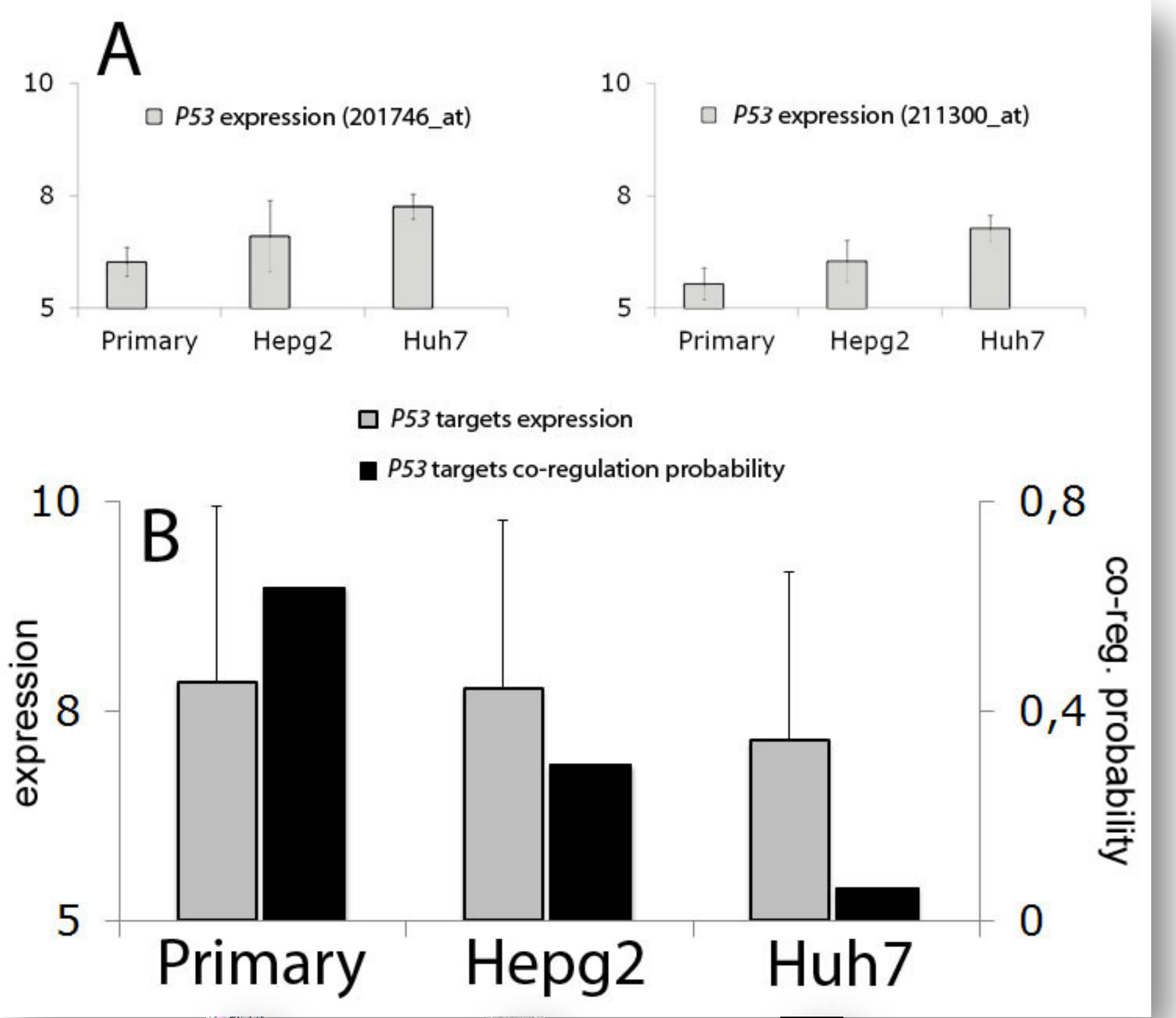
1. Primary human hepatocytes

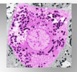
2. HepG2 cell lines (initial stage)

3. Huh7 cell lines (severe)

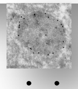
DINA is able to detect dys-regulated pathways in disease

We selected 34 bona fide targets of **p53** [1] and checked for their co-expression in the HCC cell lines.

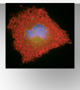




(wt)
wt p53



(initial)
wt p53

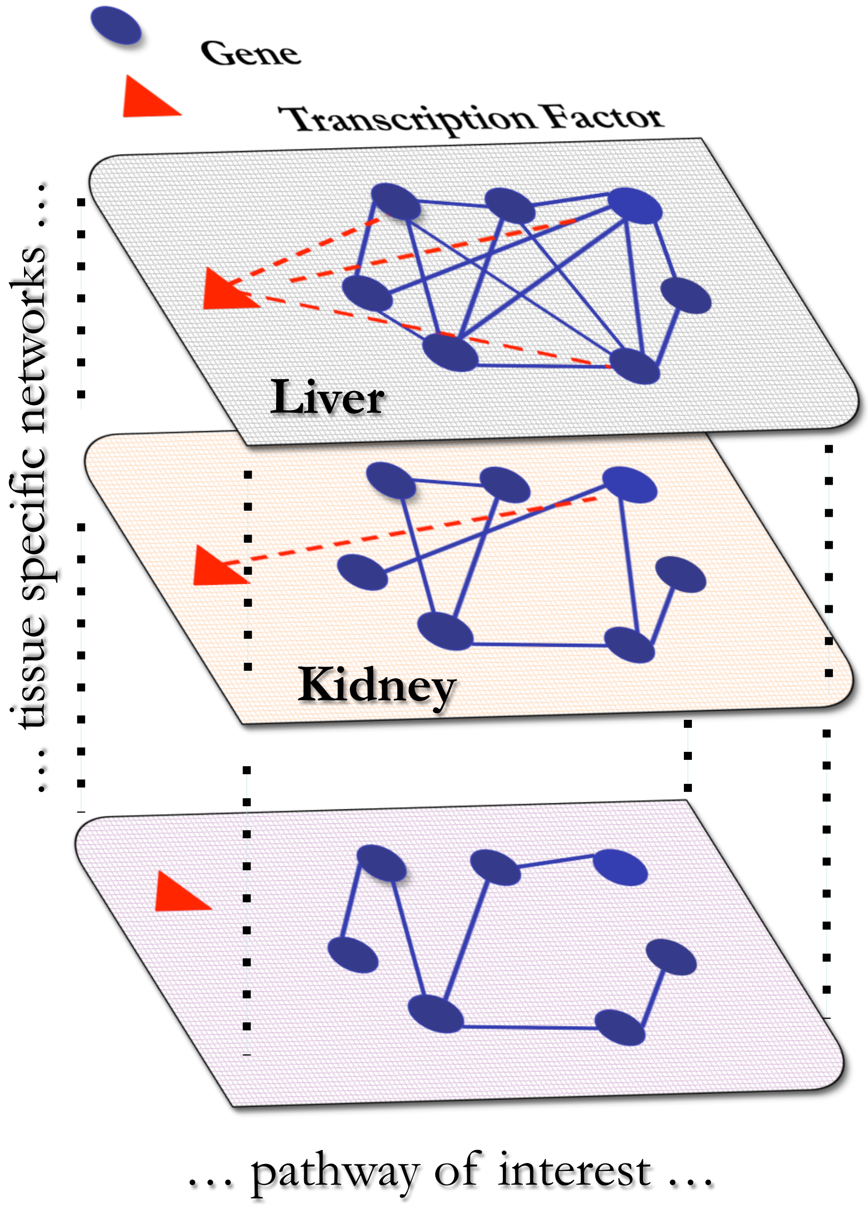


(severe)
mt p53

[1] Lim *et al.* (2007) The p53 knowledgebase: an integrated information resource for p53 research. *Oncogene, Mar 8;26(11):1517-21.*

DIfferential Network Analysis (DINA) for the identification of TFs

- We computed, for a total of 1358 verified TFs, the number of edges connecting each TF to the enzymes in the selected pathway in each of the 30 TSCN.
  - We selected those TFs that were significantly differentially co-expressed with the enzymes across the tissues using the exact Fisher test.



DIfferential Network Analysis (DINA) for the identification of TFs

Symbol	Name	Role	Citations
NR1H4	nuclear receptor subfamily 1, group H, member 4	activator	[45, 81, 82]
ESRRG	estrogen-related receptor gamma	activator	[82, 83]
TRPS1	trichorhinophalangeal syndrome I	inhibitor	–
NR1I3	nuclear receptor subfamily 1, group I, member 3	activator	[47, 48, 82]
HNF4A	hepatocyte nuclear factor 4, alpha	activator	[49, 82]
ZNF394	zinc finger protein 394	inhibitor	–
TBR1	T-box, brain, 1	activator	–
DAB2	disabled homolog 2, mitogen-responsive phosphoprotein	activator	–
DIP2C	disco-interacting protein 2 homolog C (Drosophila)	activator	–
TRIM15	tripartite motif-containing 15	activator	–
ASB9	ankyrin repeat and SOCS box-containing 9	activator	–
YEATS2	YEATS domain containing 2	inhibitor	–
SIRT4	sirtuin 4	activator	[50–52]

TABLE LEGEND

Bold: *genes encoding proteins with known TF activity.*

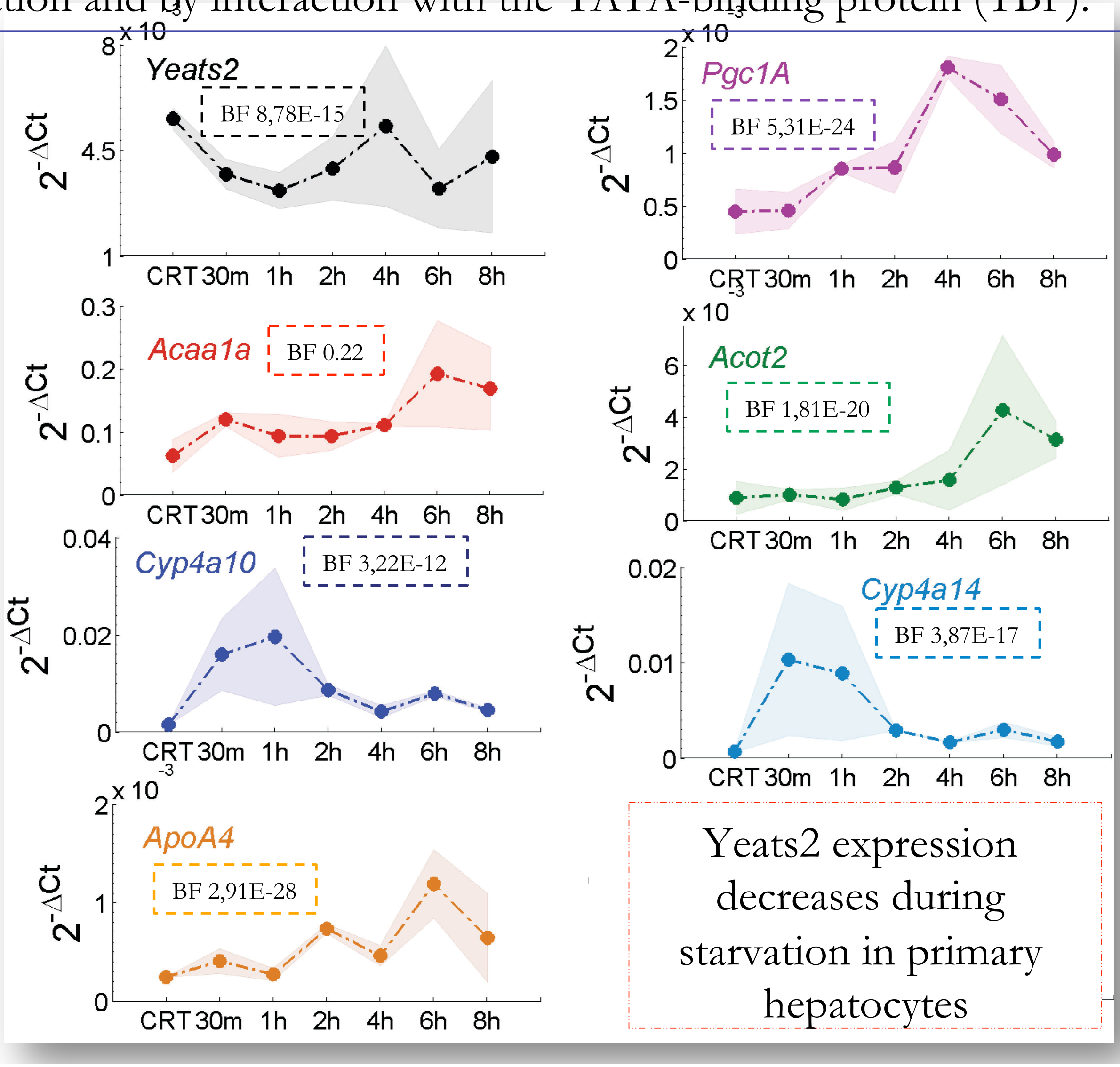
No Bold: *genes encoding protein indirectly acting on transcription*

- For each of the 9 metabolic pathways previously identified as tissue-specific, we identified the regulators shared by the majority (i.e. 7 out of 9) of metabolic pathways.

- Very little is known about YEATS2 function. Recently, it has been demonstrated to interact with the ATAC complex (Ada-Two-A-Containing)

Yeats2 as a novel regulator of metabolic gene expression

YEATS2 has been proposed to participate to the ATAC (Ada-Two-A-Containing) complex. ATAC, together with SAGA (Spt-Ada-Gcn5-Acetyl-Transferase), is able to modulate transcription, both by chromatin modification and by interaction with the TATA-binding protein (TBP).



Yeats2 expression decreases during starvation in primary hepatocytes

Thanks to Nicoletta Moretti

Conclusion II

- We hypothesized that genes belonging to a tissue-specific pathway are actively co-regulated, and hence co-expressed, only in specific tissues where the pathway is active, but not in others, independently of their absolute level of expression.

- We proposed an approach (DINA) based on quantifying the variability in the co-regulation probability and gene topology across tissues or conditions.

- We showed that this approach can be succesfully usend to elucidate tissue specific pathway and regulators.

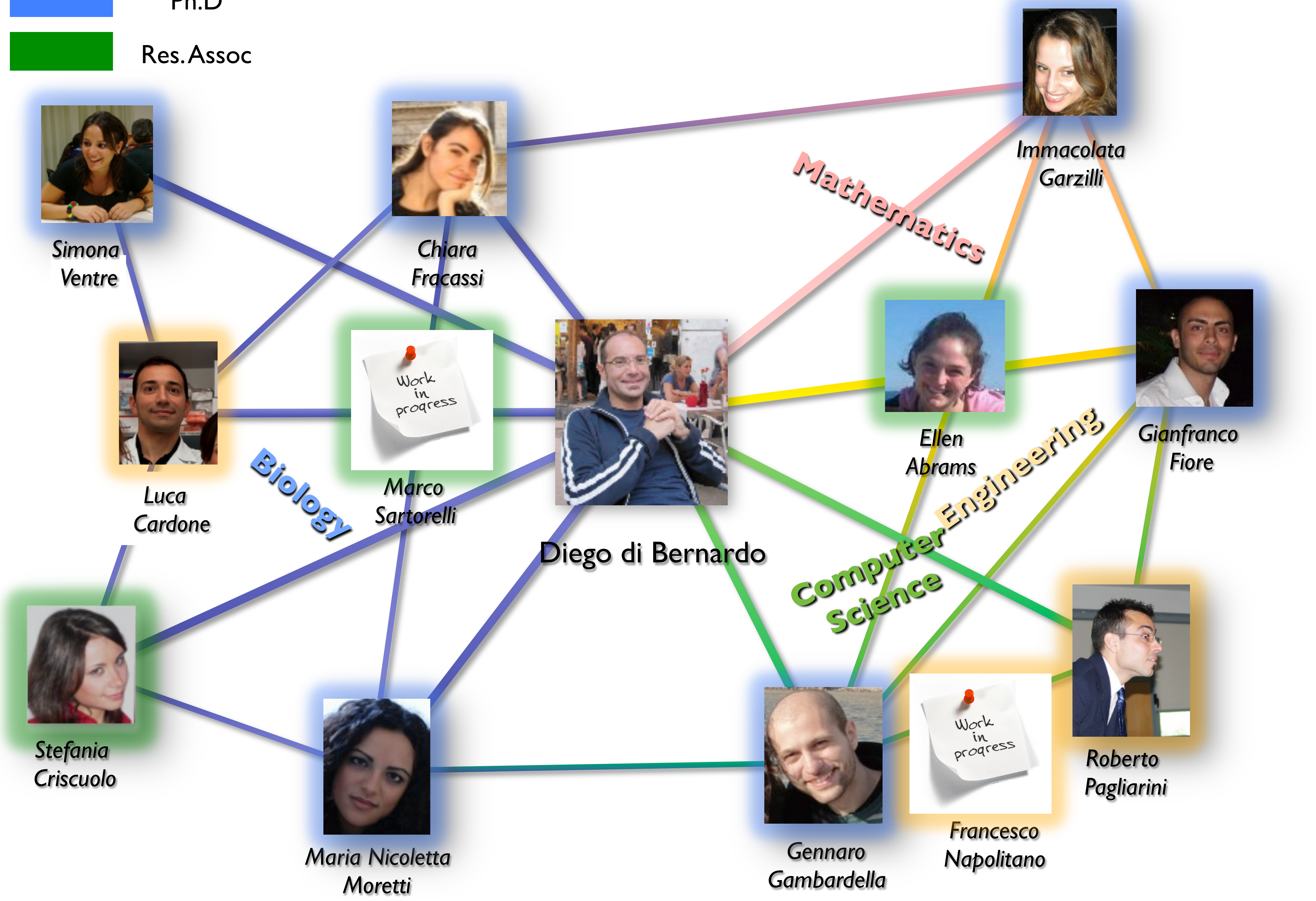
- We showed that DINA is also able to identify dysregulated pathway in disease.

Web tool availabe at http://dina.tigem.it

🟧 PostDoc

🟦 Ph.D

🟩 Res. Assoc



Immacolata Garzilli

Simona Ventre

Chiara Fracassi

Mathematics

Luca Cardone

Marco Sartorelli

Diego di Bernardo

Ellen Abrams

Gianfranco Fiore

Biology

Engineering

Computer Science

Stefania Criscuolo

Maria Nicoletta Moretti

Gennaro Gambardella

Francesco Napolitano

Roberto Pagliarini