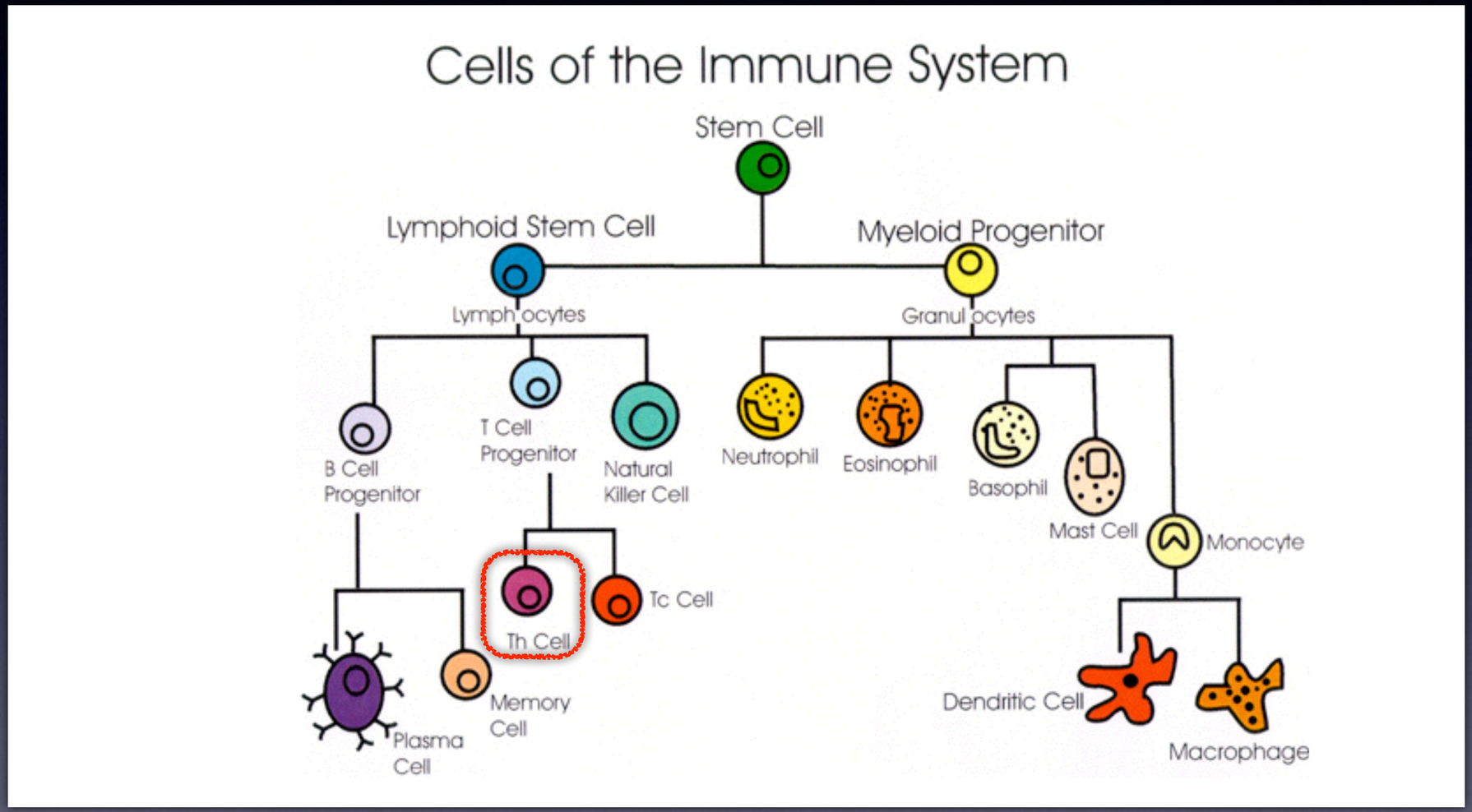
Conserved cross-species network modules elucidate Th17 T-cell differentiation in human and mouse

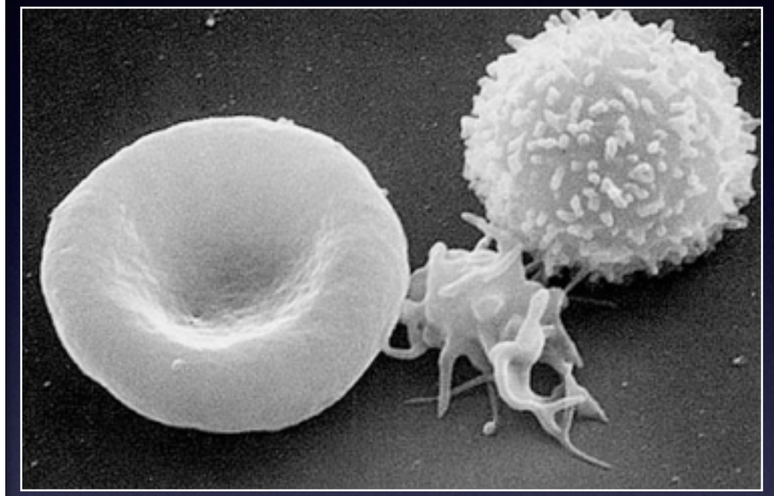
ECCB 14 BioNetVisA Workshop

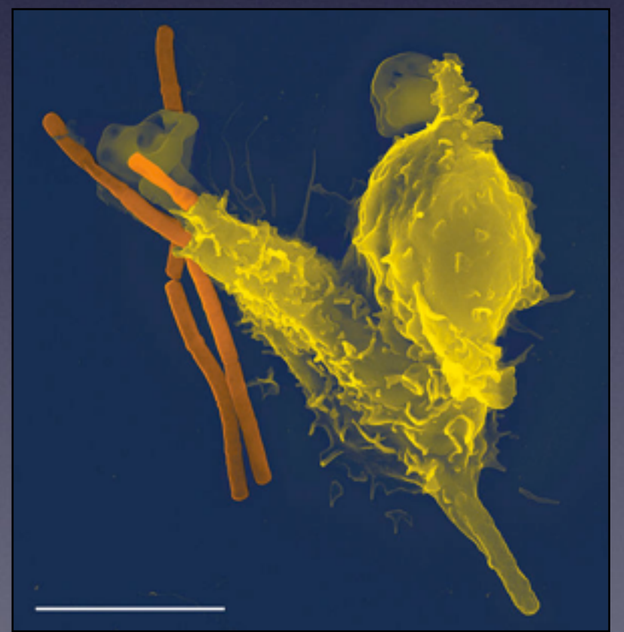
Mohammed El-Kebir, Hayssam Soueidan, Thomas Hume, Daniela Beisser, Marcus Dittrich, Tobias Müller, Guillaume Blin, Jaap Heringa, Macha Nikolski, Lodewyk F. A. Wessels, Gunnar W. Klau

"Trouble with mice is you always kill 'em."
— John Steinbeck, Of Mice and Men

T Cells

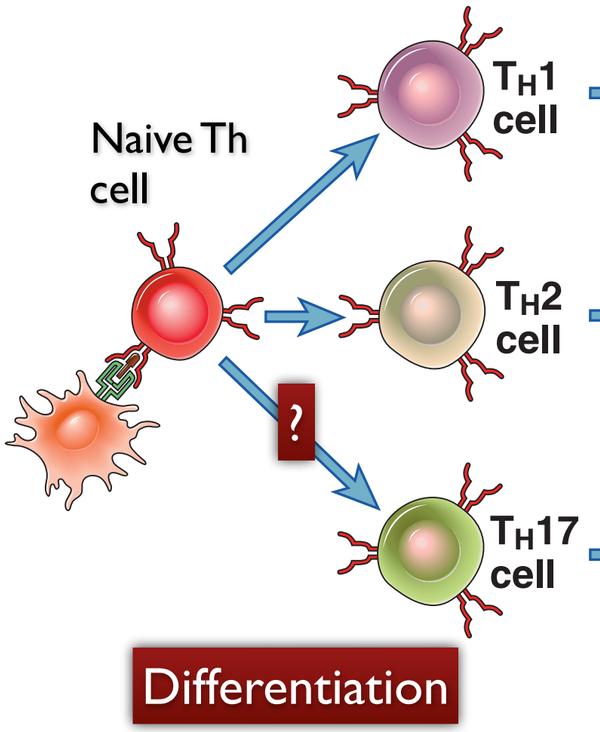






Wikimedia commons
Abbas, Cellular & Molecular Immunology, 7th ed

T Helper subsets



	Signature cytokines	Immune reactions	Host defense	Role in diseases
T_H1 cell	IFNγ	Macrophage activation; IgG production	Intracellular microbes	Autoimmune diseases; tissue damage associated with chronic infections
T_H2 cell	IL-4 IL-5 IL-13	Mast cell, eosinophil activation; IgE production; "alternative" macrophage activation	Helminthic parasites	Allergic diseases
T_H17 cell	IL-17A IL-17F IL-22	Neutrophilic, monocytic inflammation	Extracellular bacteria; fungi	Organ-specific autoimmunity

Abbas, Cellular & Molecular Immunology, 7th ed

Th17 cells: Clinical relevance

- Mucosal immunology : Th17 respond to bacterial and fungal antigens
- Th17 cells imbalance associated with several autoimmune diseases (Rheumatoid Arthritis, MS, psoriasis, lupus, CD)
- IL-17-deficient mice are more susceptible to the development of lung melanoma
- HIV infection specifically depletes Th17 population

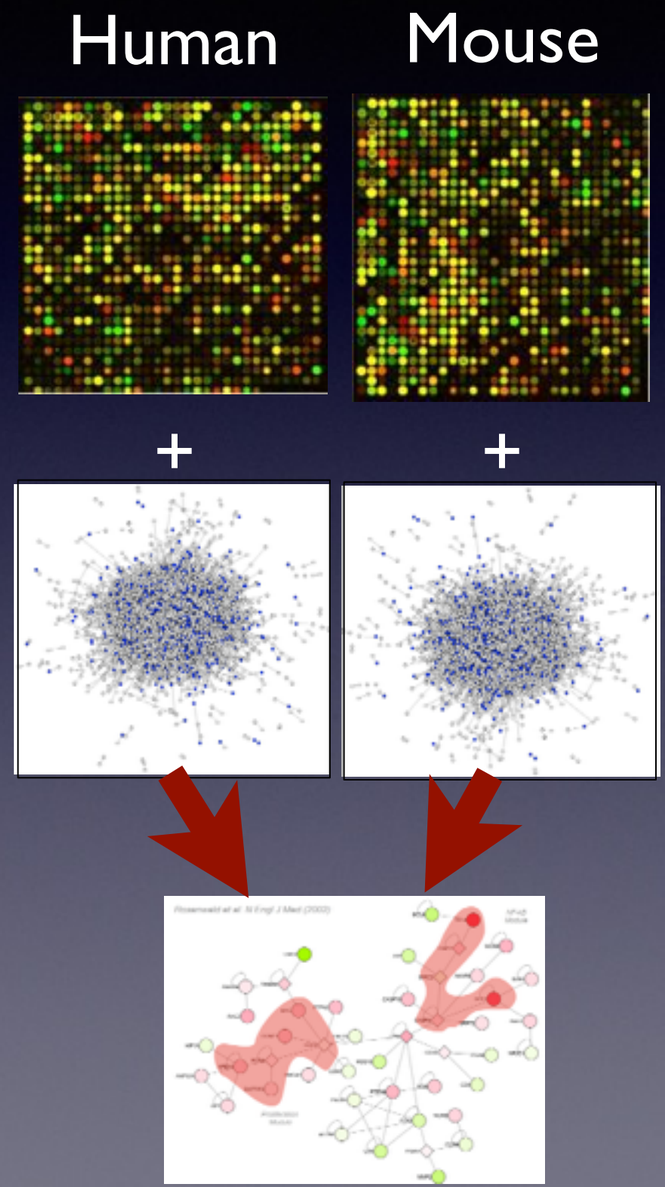
Main questions

- How to modulate Th17 response to self?
- What are the regulators of Th17 balance?
- What are the proteins and pathways responsible for proper differentiation of Th17 cells?

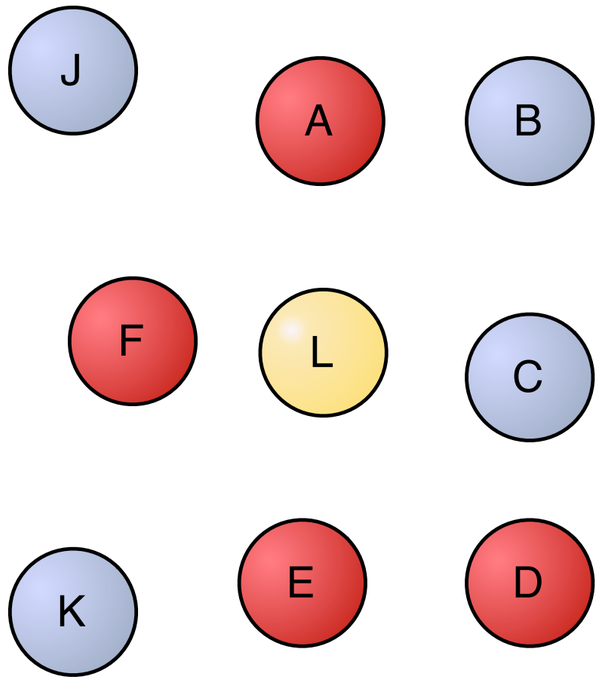
How well findings in mouse are transferrable to human immunology?

Our strategy

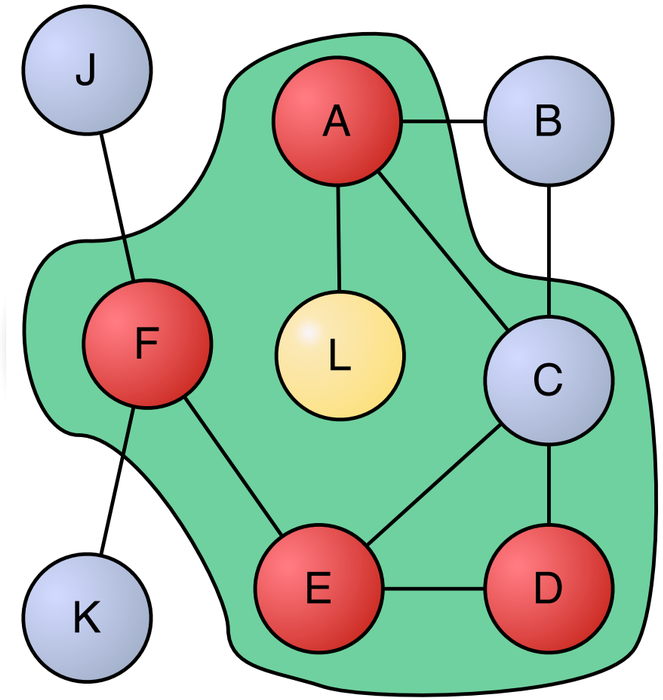
- Let's combine:
  - Human and Mouse Th17 differentiation transcriptomics data
  - Human and Mouse PPI networks
  - Orthology information between Human and Mouse
- Using an optimization framework
- To identify *conserved cross-species active modules*



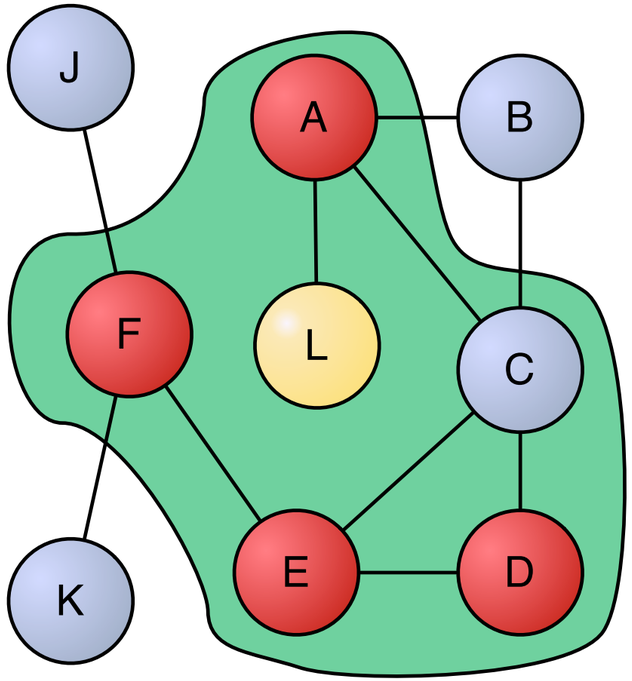
CCSAM:(1) Activity

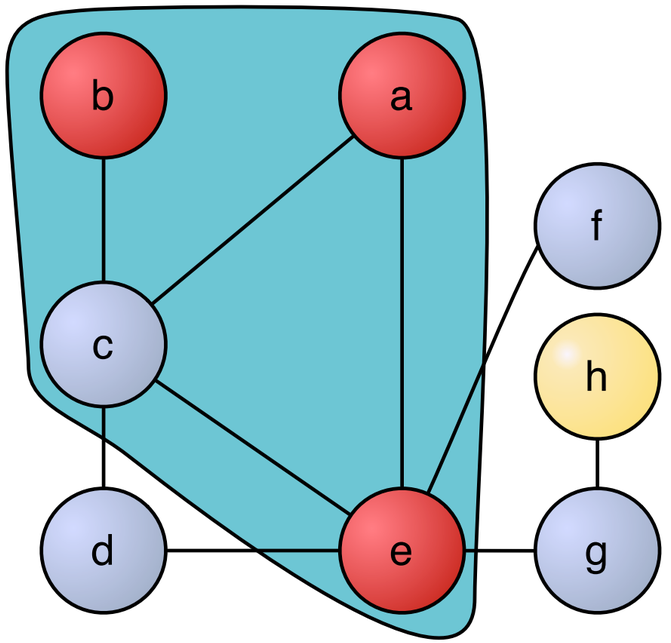


CCSAM:(11) Modularity

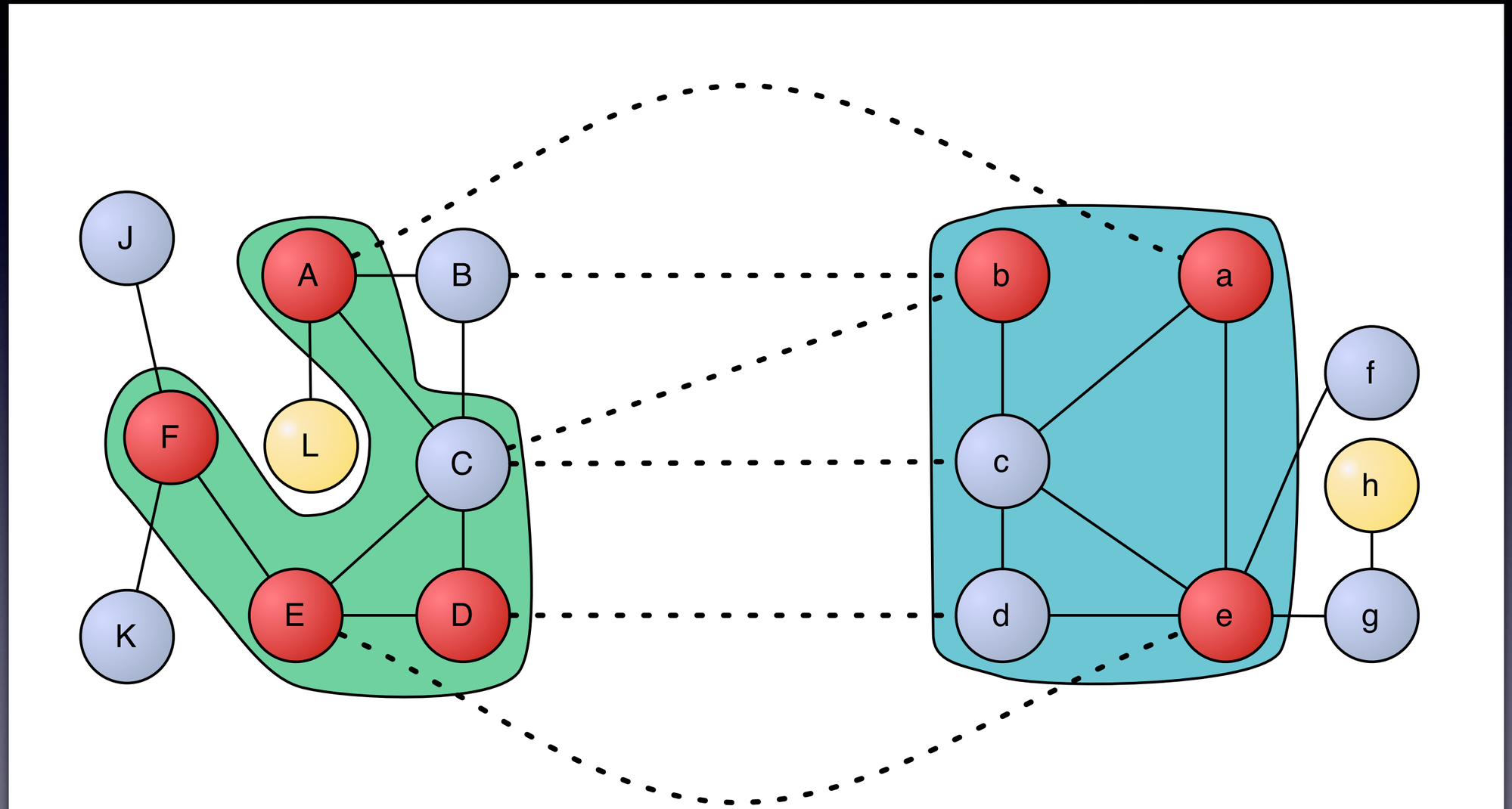


CCSAM:(11) Modularity^2





CCSAM:(III) Conservation



Today

- Material: Gene expression profiling & public datasets
- Method: Conserved active module
- Results: Modules identification
  - Regulation at 2h and 72h are well conserved
  - Overall dynamics is conserved
- Conclusions & future work

Today

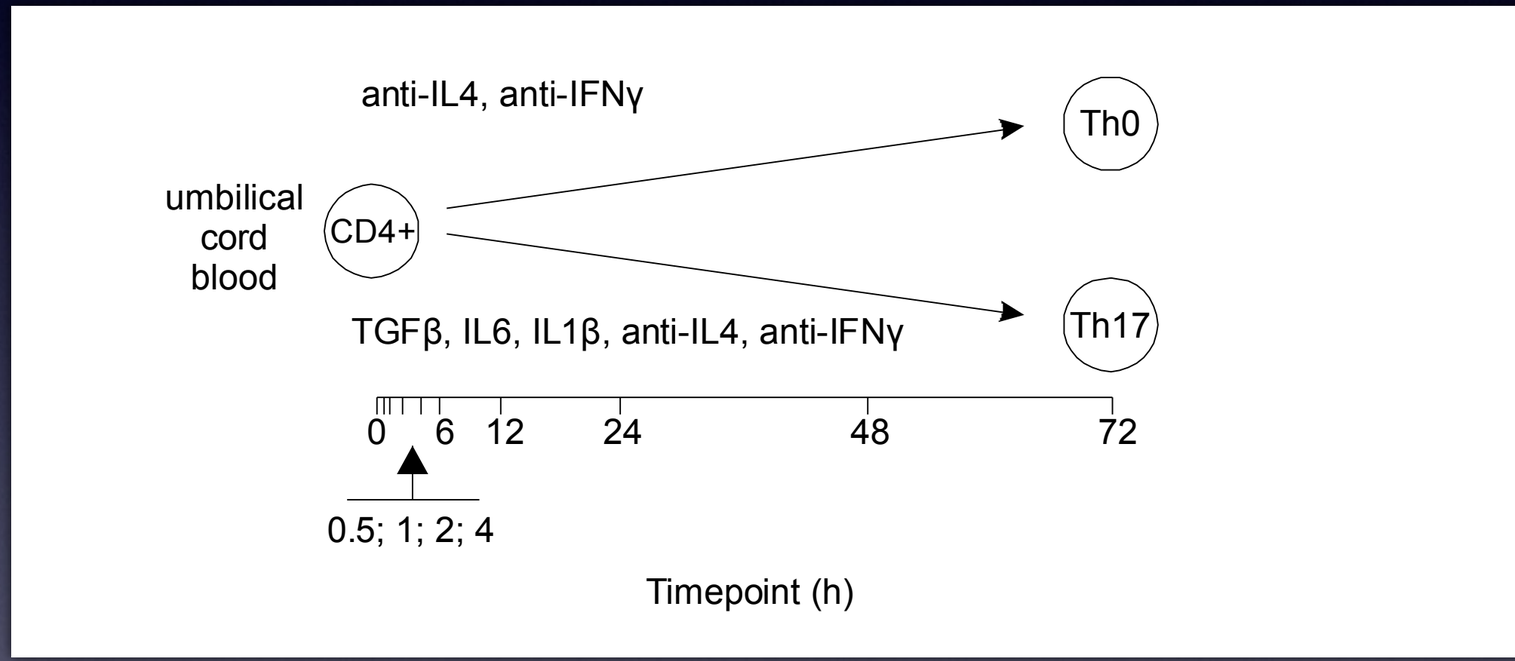
- Material: Gene expression profiling & public datasets
- Method: Conserved active module
- Results: Modules identification
  - Regulation at 2h and 72h are well conserved
  - Overall dynamics is conserved
- Conclusions & future work

M&M recipe:

- For each species individually:
  - Process RNA-Seq => Count matrix
  - Fit a GLM => estimated coefs
  - For each time point,
    - For each gene:
      - call for DE => p.value
      - Fit a BUM => activity score
  - Get PPI network & orthology relations

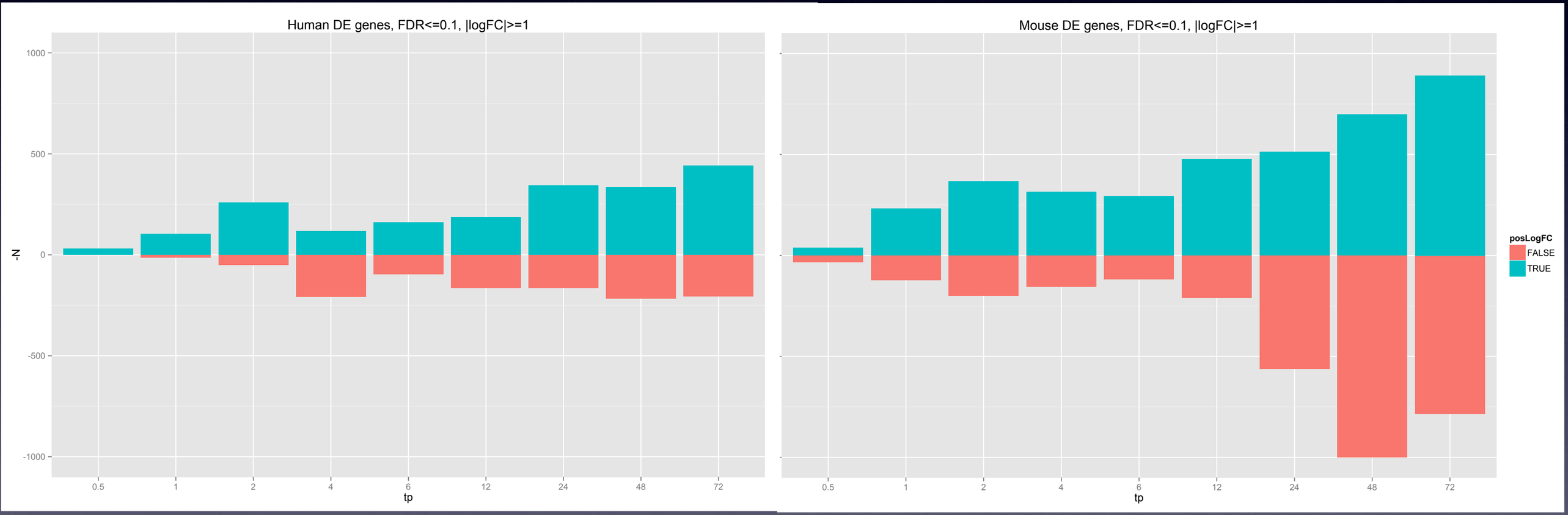
Transcriptional profiling of Human & Mouse Th17 cells

- Control Th0 vs Th17
- 9 time points, RNA-Seq
- Human: 14,338 mRNAs quantified in the first 72h
- Mouse: 11,751 mRNAs quantified in the first 72h
- Matched time points
- DE called with an edgeR GLM



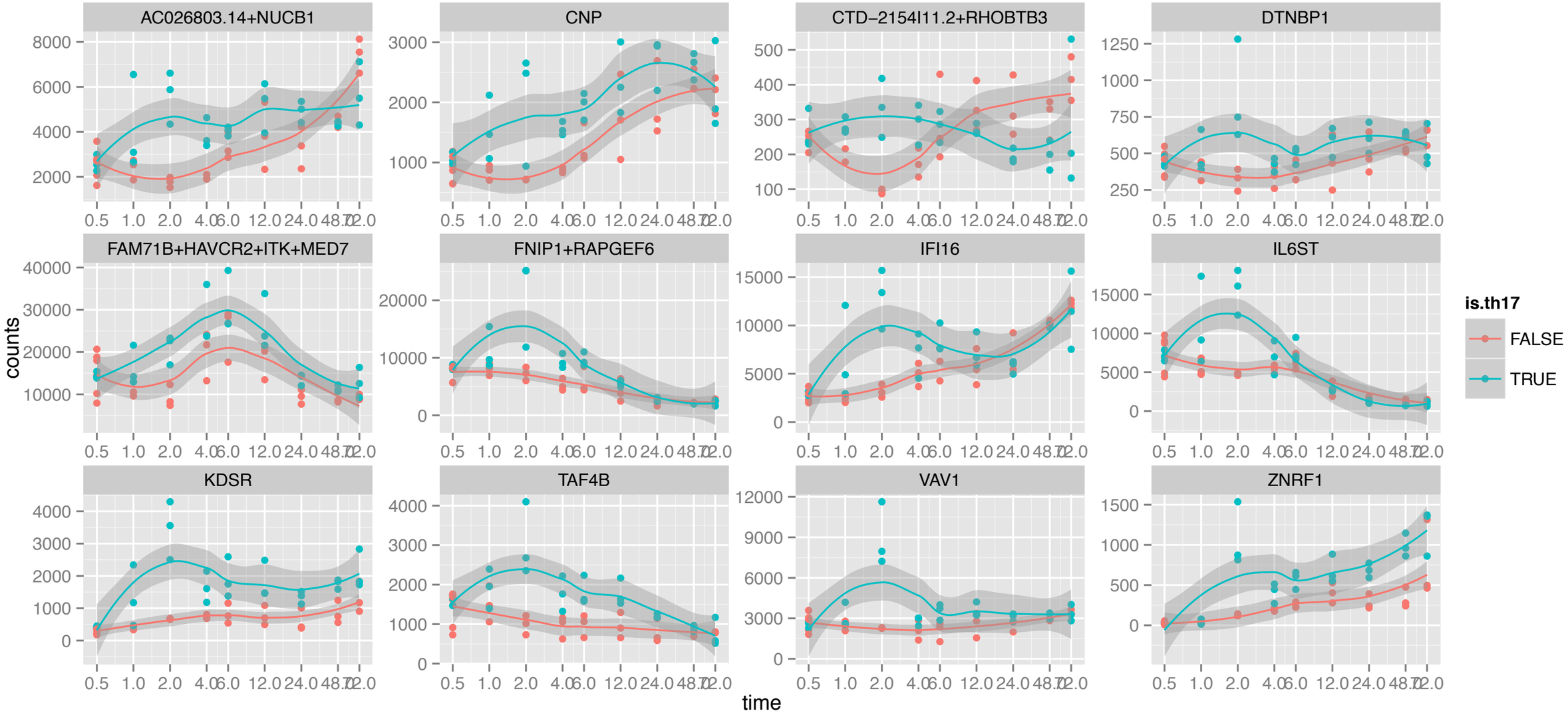
Tuomela et al., 2014

Transcription dynamics



- Biphasic in both species
- Seems “stronger” in the mouse samples
- earliest changes (1/2h!) visible

Contrasting: Genes DE at time 2

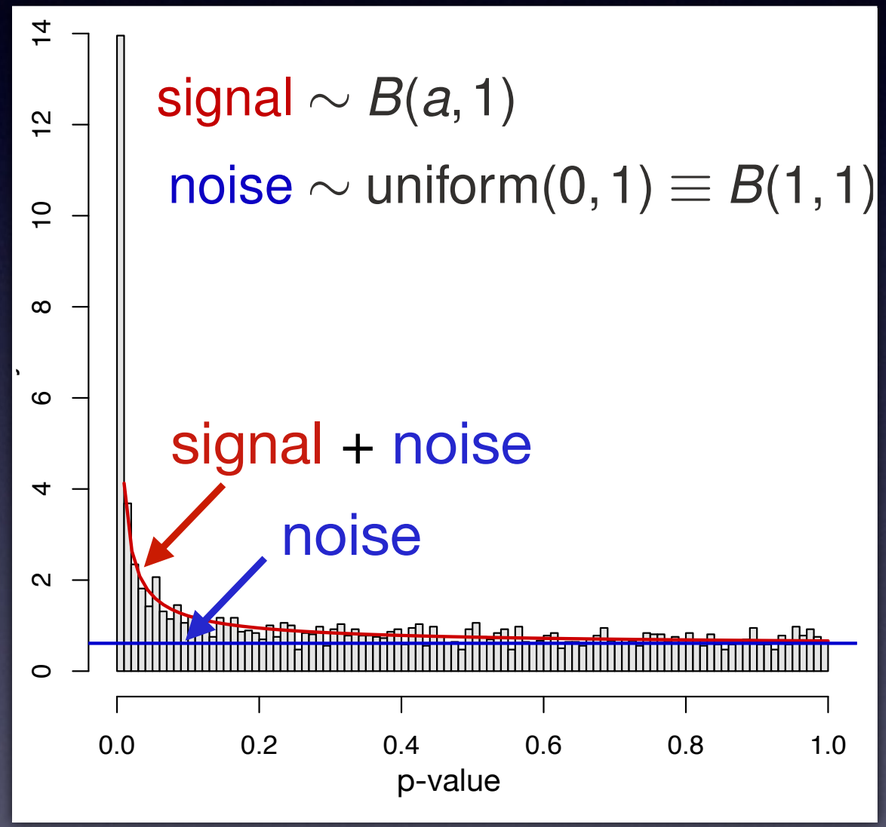


Scores: BUM Model

- Activity (positive and negative) scores derived from
  - p-values distribution
  - described with a beta-uniform mixture model;
  - controlling the FDR
  - using an adj. LL ratio:

$$s(x, \mathrm{FDR}) = \log \frac{\hat{a} x^{\hat{a}-1}}{\hat{a}\tau(\mathrm{FDR})^{\hat{a}-1}}$$

$$= (\hat{a} - 1)\left(\log(x) - \log(\tau(\mathrm{FDR}))\right)$$



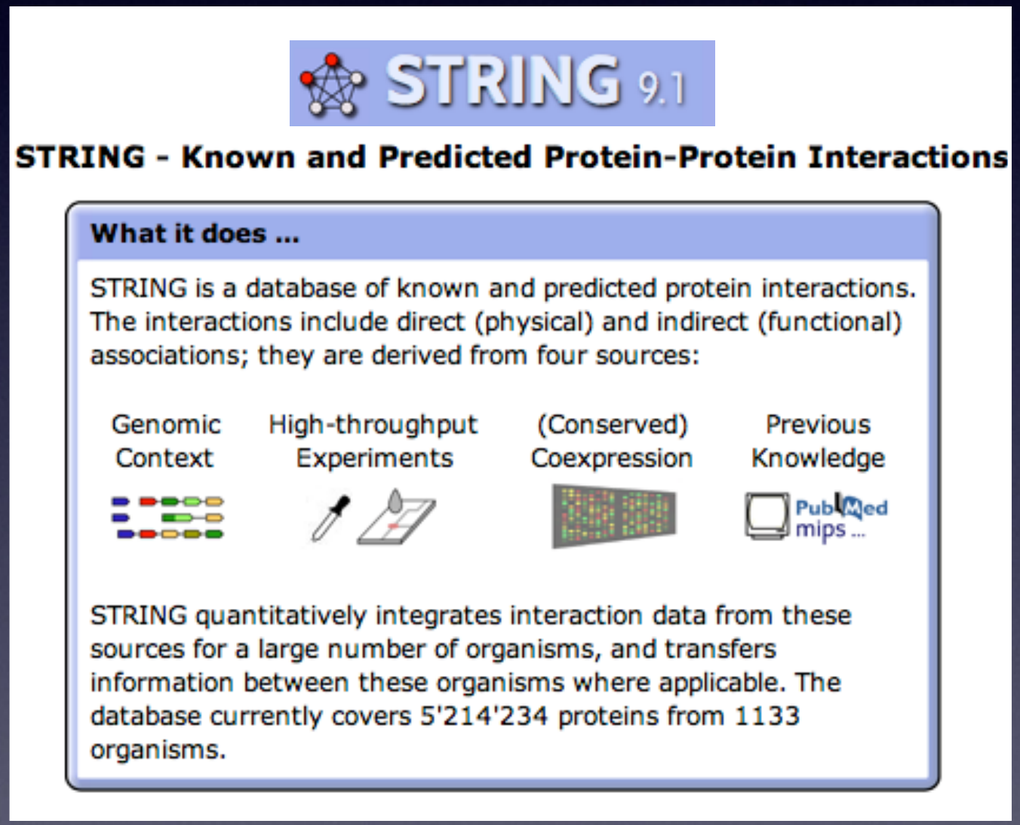
Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. Pounds 2003

Today

- Material: Gene expression profiling & public datasets
- Method: Conserved active module
- Results: Modules identification
  - Regulation at 2h and 72h are well conserved
  - Overall dynamics is conserved
- Conclusions & future work

PPI Networks

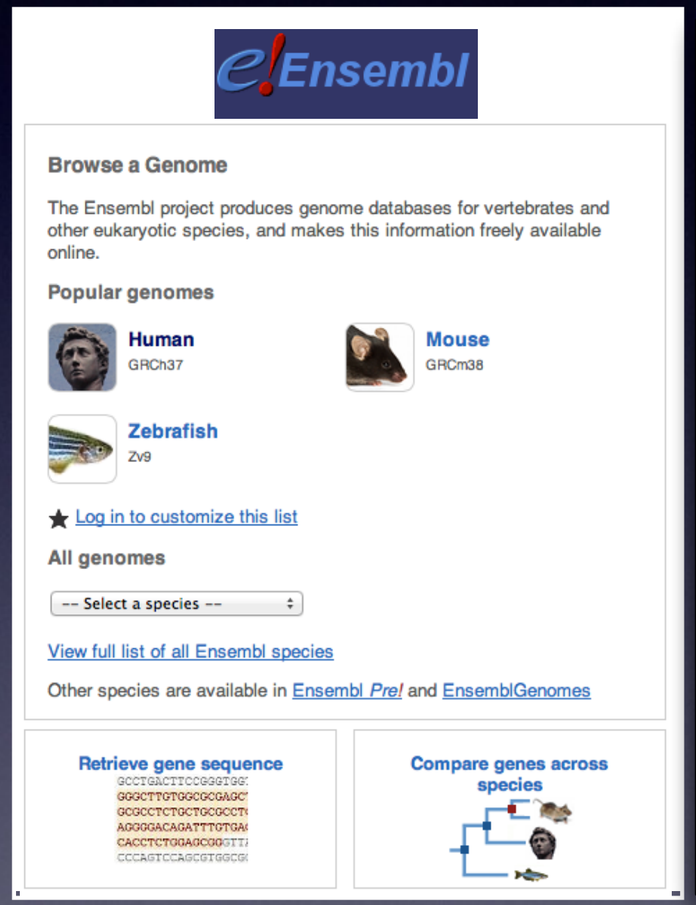
- Obtained from the STRING db
- Only kept *physical interactions*
- Mouse network: 12,121 nodes and 176,462 edges
- Human network : 14,280 nodes and 197,649 edges

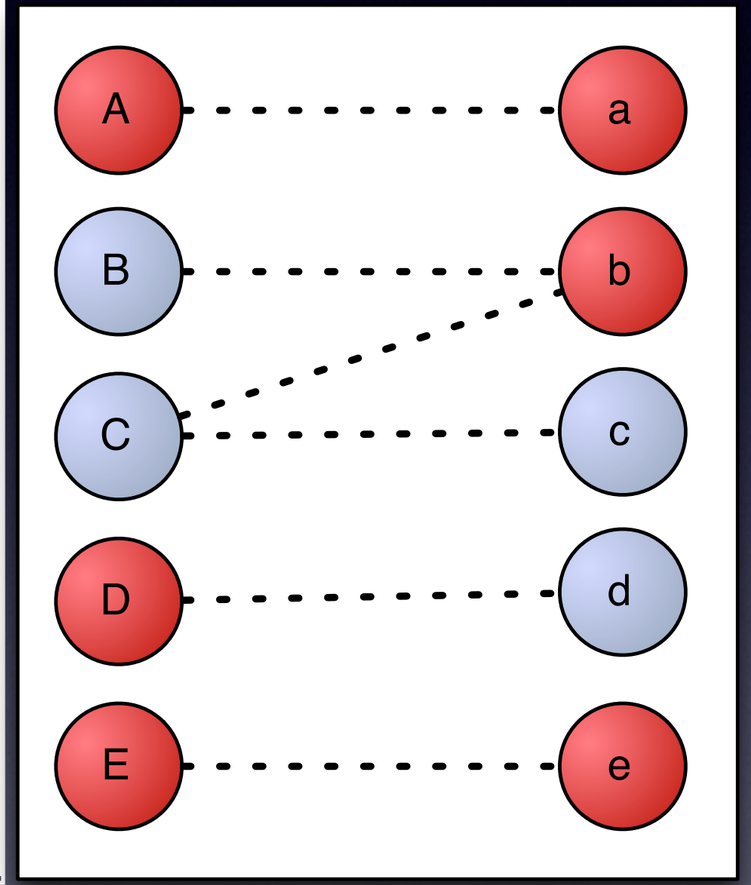


Orthology relations

- Obtained from ENSEMBL orthology
- Represented as a bi-partite graph M
- 85,640 human proteins
- 49,717 mouse proteins
- linked by 125,685 edges
- grouped in 16,736 bicliques

  (avg size of 8.08, median of 6, SD of 5.97)





M&M recipe (recap)

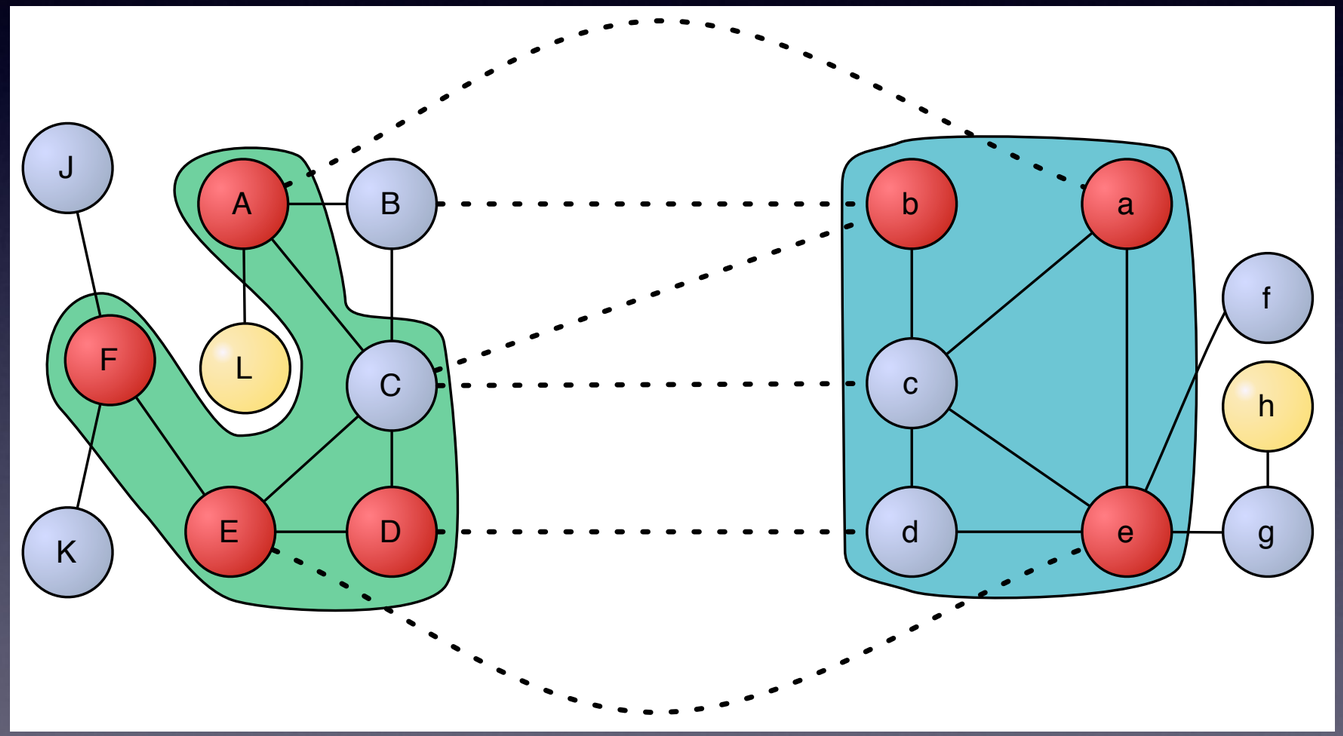
- For each species individually:
  - Process RNA-Seq => Count matrix
  - Fit a GLM => estimated coefs
  - For each time point,
    - For each gene:
      - call for DE => p.value
      - Fit a BUM => activity score
  - Get PPI network & orthology relations

Today

- Material: Gene expression profiling & public datasets
- Method: Conserved active module
- Results: Modules identification
  - Regulation at 2h and 72h are well conserved
  - Overall dynamics is conserved
- Conclusions & future work

Conserved active modules

- Formalized using a constraint modeling approach over boolean variables
- Constraints are linearized => MILP
- The MILP is then solved using CPLEX with a *branch-and-cut* algorithm



MILP:?

- A formulation, with:
  - an *objective function*
  - subject to *linear constraints*
  - where variables can be constrained to discrete domains ($\{0,1\}, \mathbb{N}$)
  - much harder than on $\mathbb{R}$
  - for which exact solutions can be found efficiently in practice

$$\max_{x_1, x_2} S_1 \cdot x_1 + S_2 \cdot x_2$$

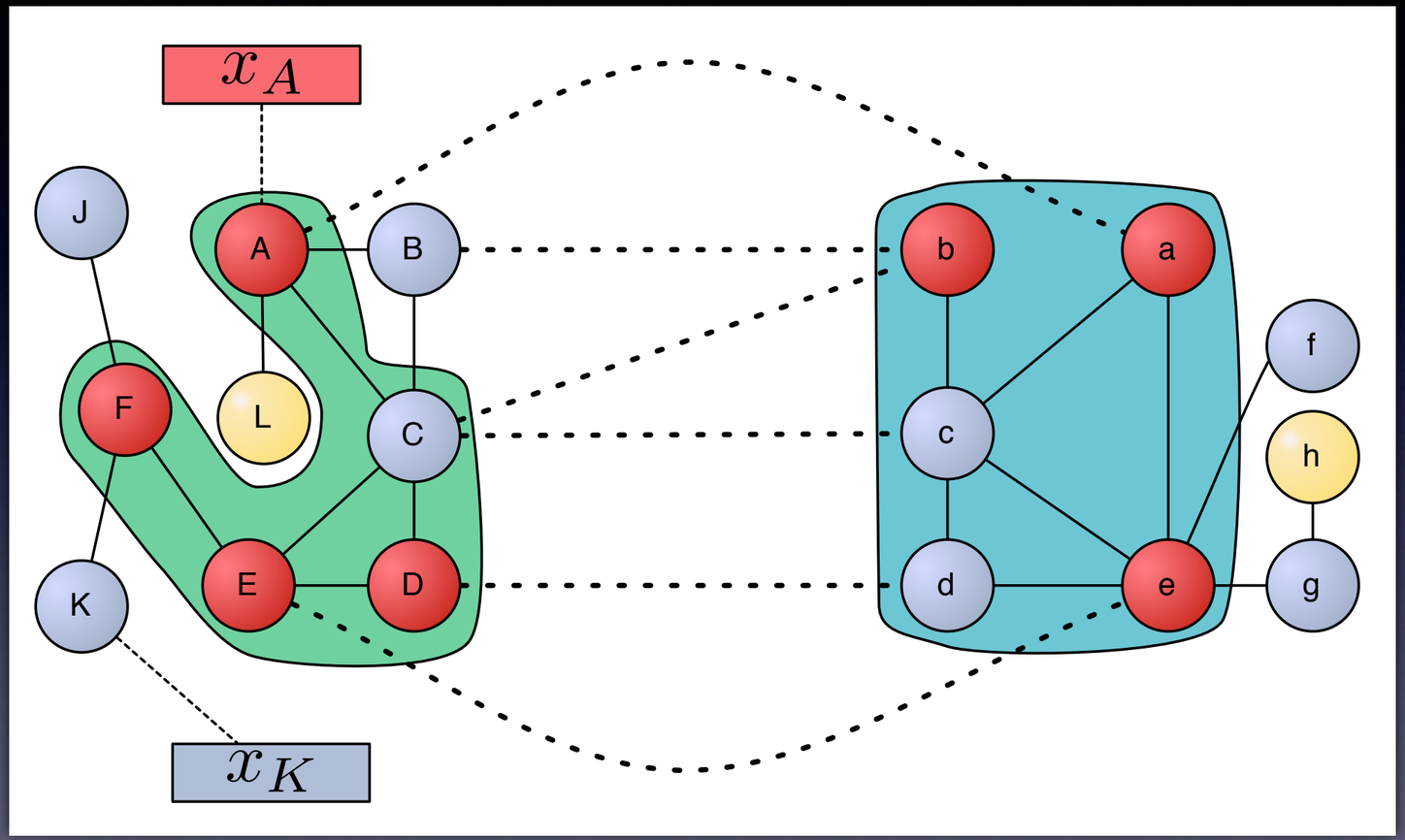
$$\text{Subject to: } x_1 + x_2 \leq L$$

$$F_1 \cdot x_1 + F_2 \cdot x_2 \leq F$$

$$P_1 \cdot x_1 + P_2 \cdot x_2 \leq P$$

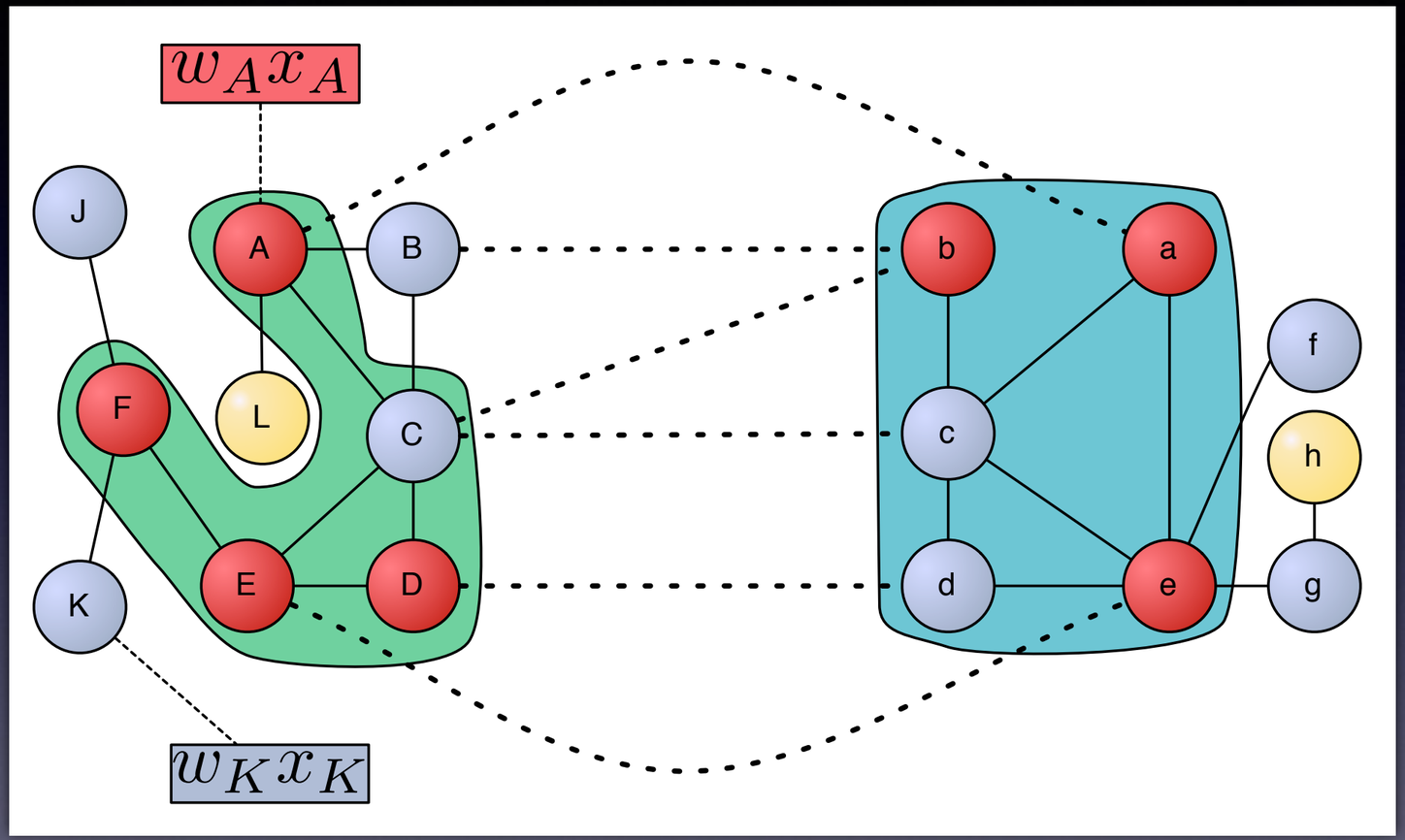
$$x_1 \geq 0, x_2 \geq 0$$

MILP: Variables



Boolean variables for nodes in solution

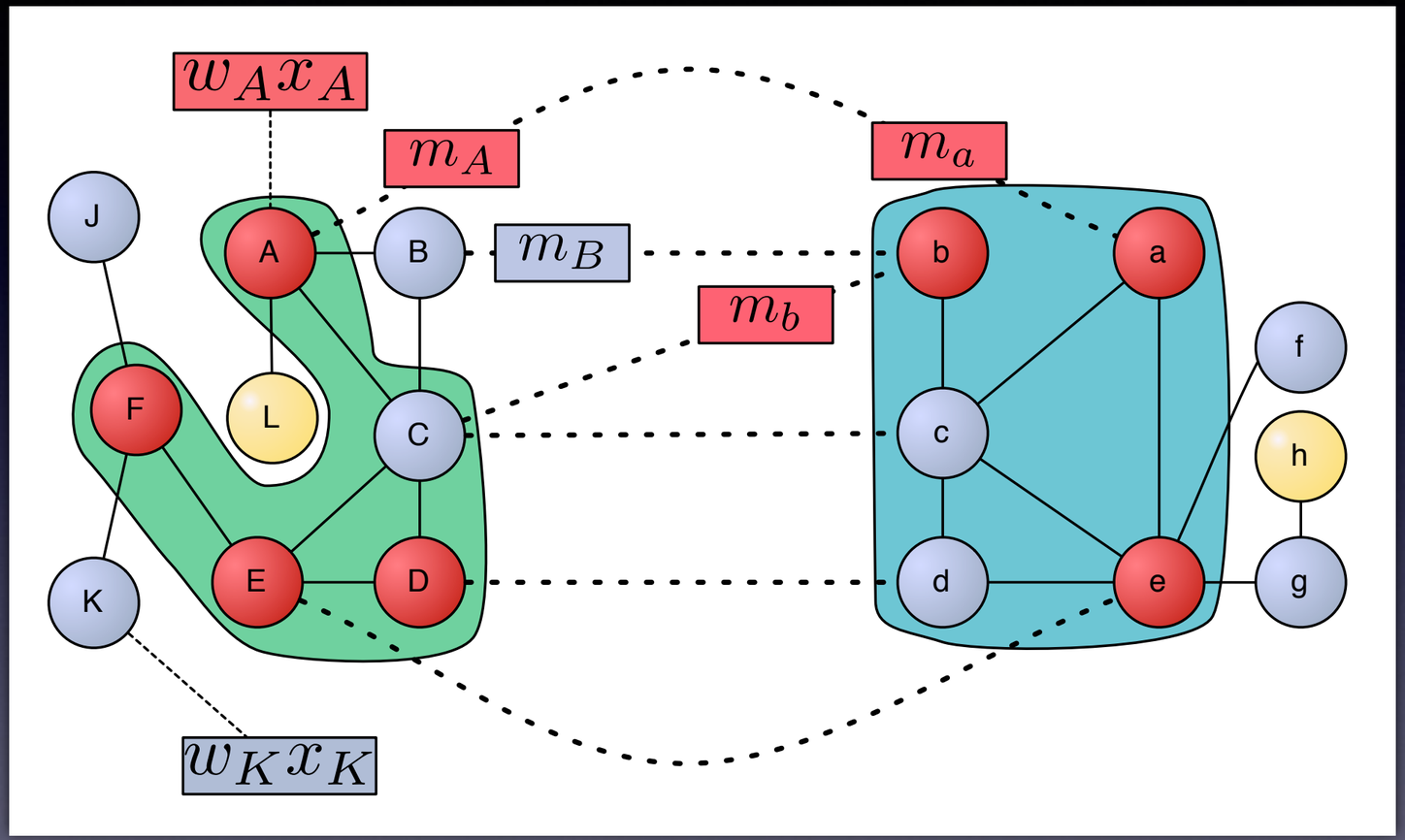
MILP: Variables



Weighted boolean variables for nodes in solution, objective function:

$$\max \sum_{v \in V_1 \cup V_2} w_v x_v$$

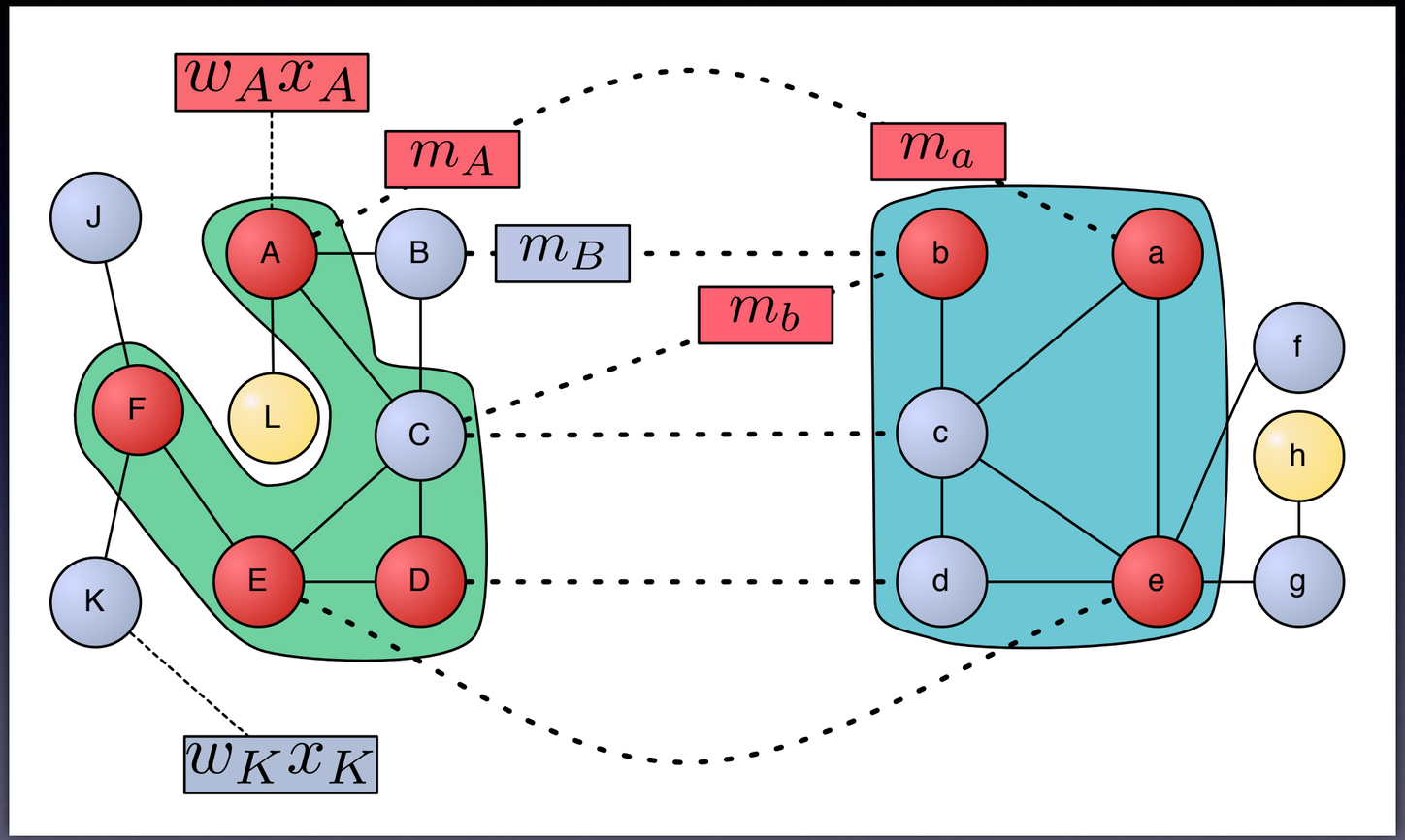
MILP:Variables



Boolean variables for conserved nodes

$$m_u = \max_{uv \in M} \{x_u x_v\}$$

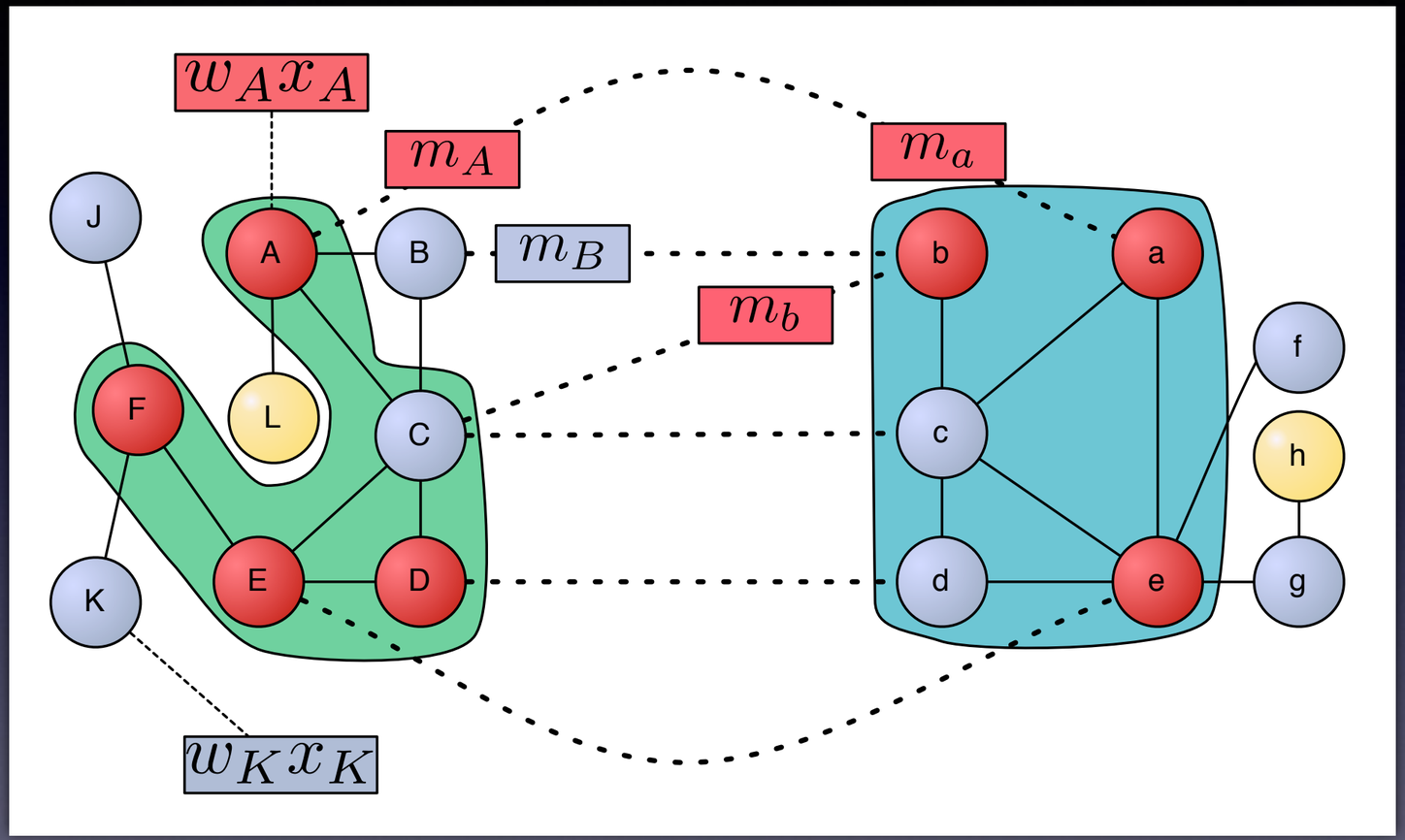
MILP: degree of conservation



constrained by having e.g more than $\alpha = 50\%$ of nodes being conserved

$$\sim \frac{\sum m_v}{\sum x_v} \geq 50\%$$

MILP: connectivity



- And satisfying the connectivity constraint:
  - Possibly an exponential number of constraints
  - Constraints added as needed during optimization

MILP: Formulation

$$\max \sum_{v \in V_1 \cup V_2} w_v x_v \tag{1}$$

$$\text{s.t. } m_u = \max_{uv \in M} \{x_u x_v\} \qquad u \in V_1 \tag{2}$$

$$m_v = \max_{uv \in M} \{x_u x_v\} \qquad v \in V_2 \tag{3}$$

$$\sum_{v \in V_1 \cup V_2} m_v \geq \alpha \sum_{v \in V_1 \cup V_2} x_v \tag{4}$$

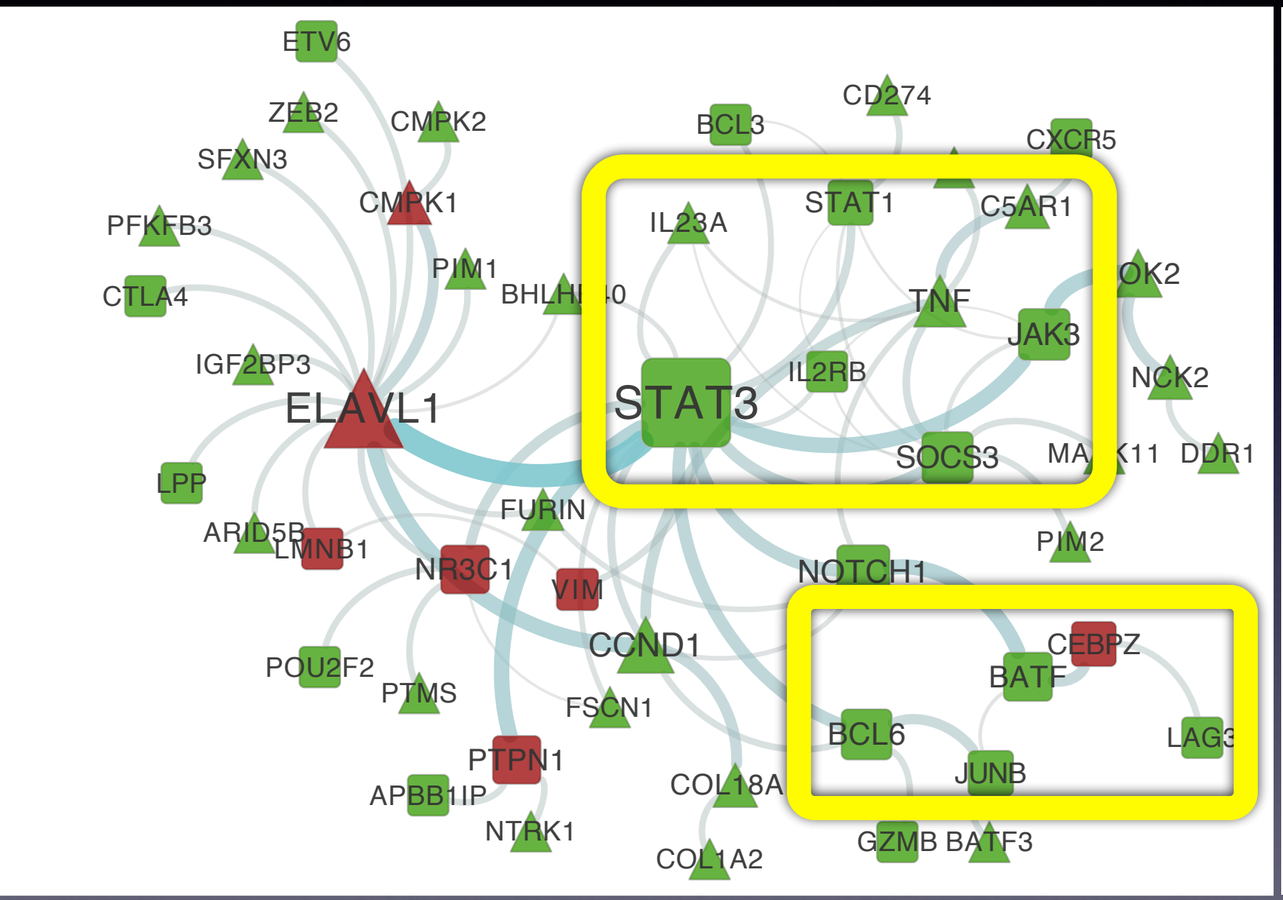
$$G_1[\mathbf{x}] \text{ and } G_2[\mathbf{x}] \text{ are connected} \tag{5}$$

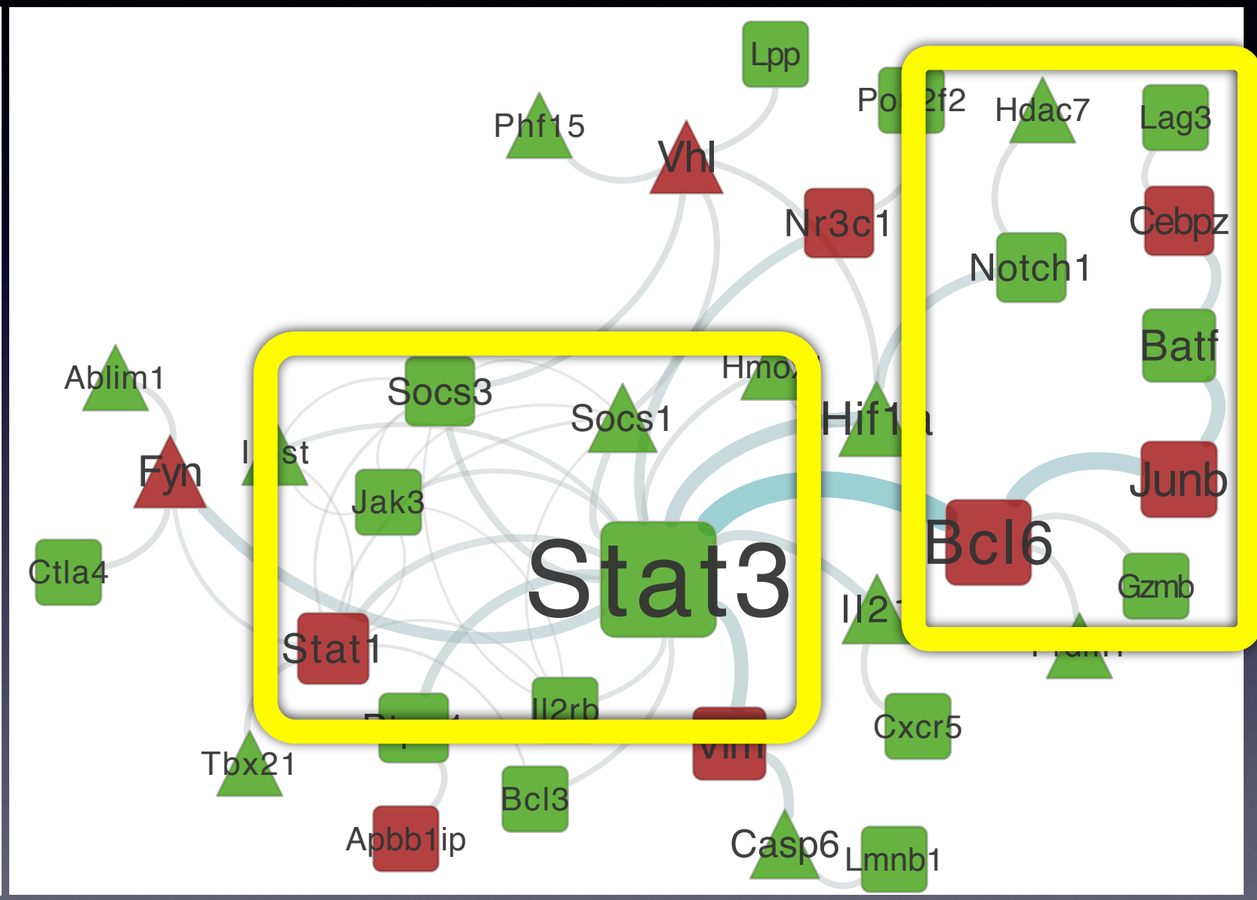
$$x_v, m_v \in \{0, 1\} \qquad v \in V_1 \cup V_2 \tag{6}$$

Today

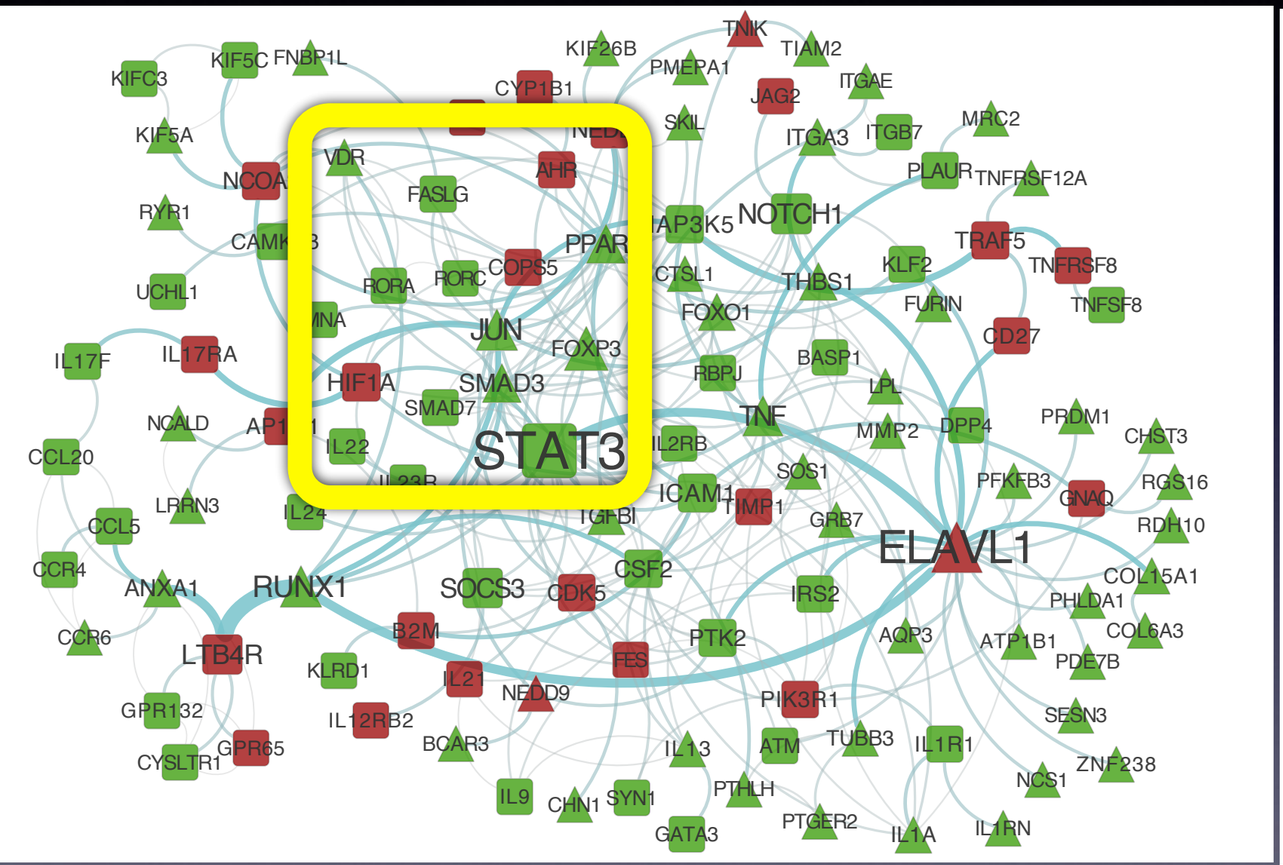
- Material: Gene expression profiling & public datasets
- Method: Conserved active module
- Results: Modules identification
  - Regulation at 2h and 72h are well conserved
  - Overall dynamics is conserved
- Conclusions & future work

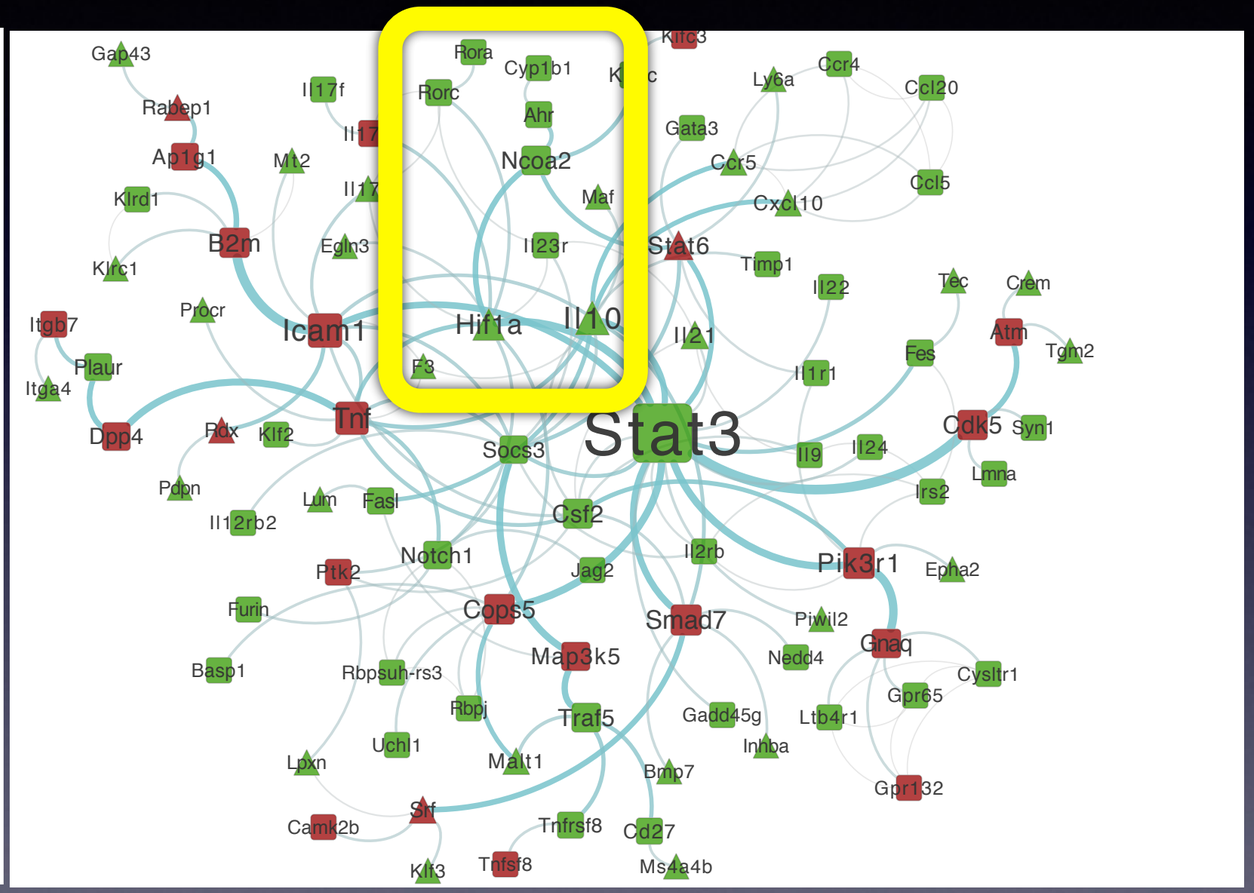
2h conserved module





72h conserved module

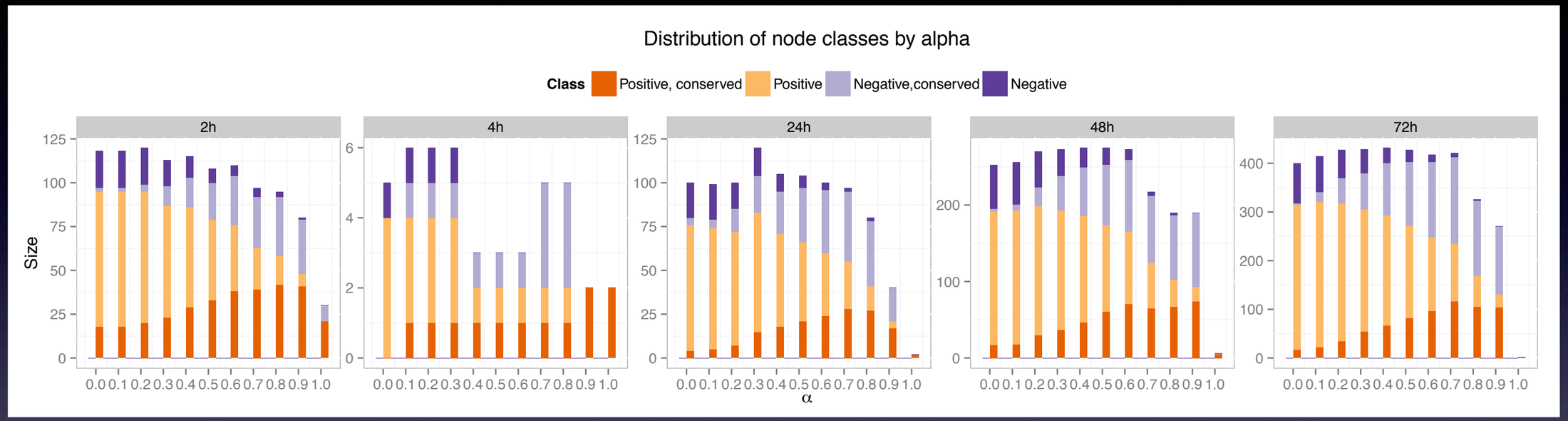


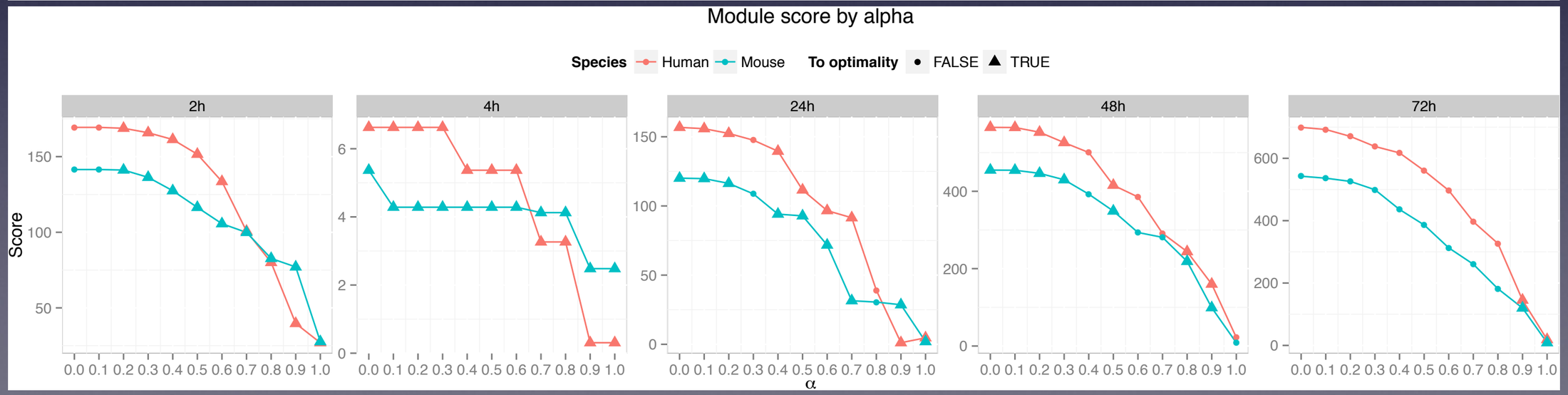


Today

- Material: Gene expression profiling & public datasets
- Method: Conserved active module
- Results: Modules identification
  - Regulation at 2h and 72h are well conserved
  - Overall dynamics is conserved
- Conclusions & future work

Overall dynamics and solutions





Today

- Material: Gene expression profiling & public datasets
- Method: Conserved active module
- Results: Modules identification
  - Regulation at 2h and 72h are well conserved
  - Overall dynamics is conserved
- Conclusions & future work

Conclusions

- Mouse and Human Th17 differentiation processes are well conserved during the first 72h
- Differentiation happens in two phases, very early (0h--4h) and late (12h--72h)
- We provide the first formulation of the *conserved active module problem* as well as an efficient MILP solver
- Code and recipes available there: http://software.cwi.nl/xheinz
- PS: We got the same results on an independent data set!

Future work

- Theoretical results on the computational complexity for specific network topologies
- Novel formulation for bi-conservation: Conservation between species and across time
- Non-supervised formulation: Clustering of samples based on conserved active modules

Thanks!

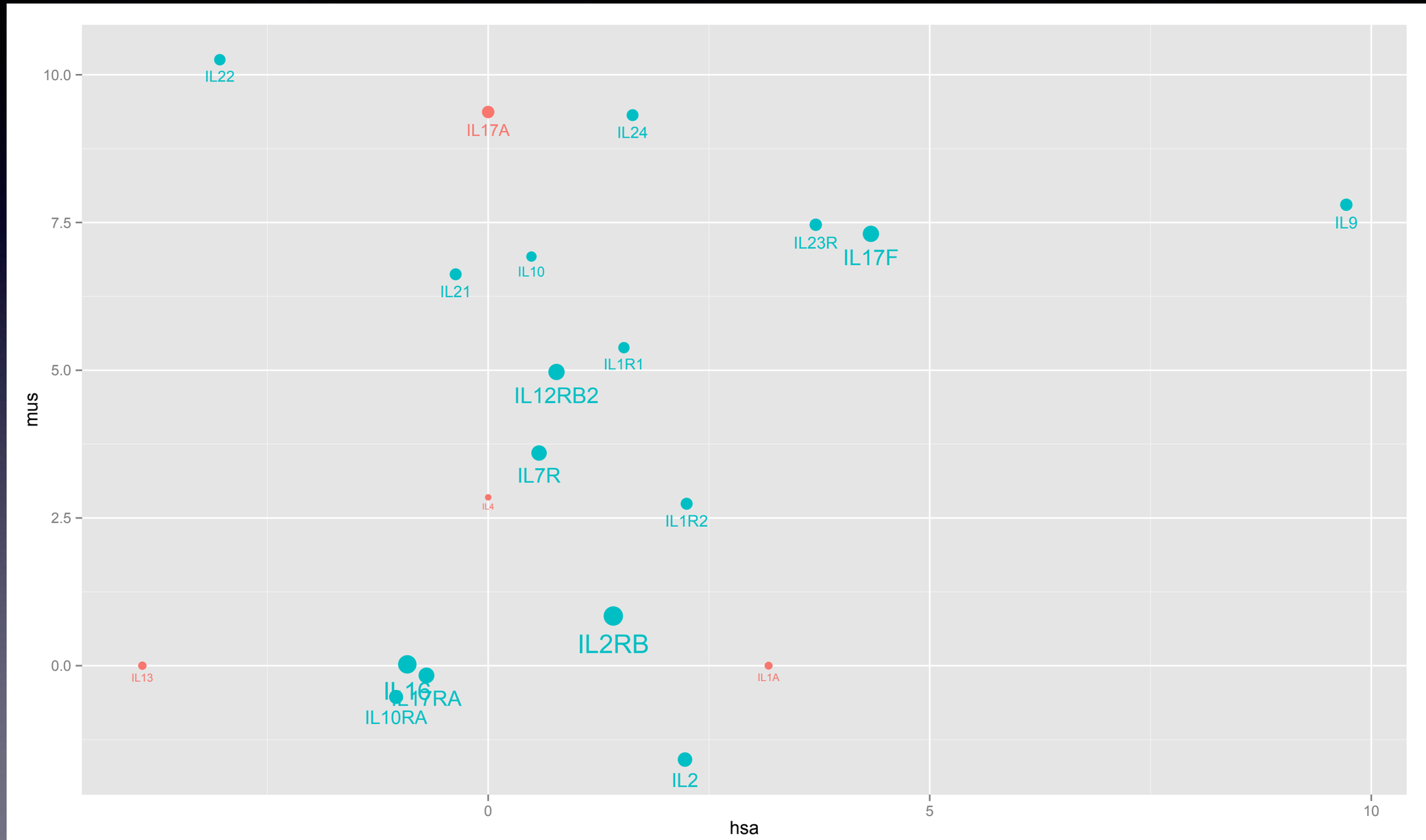
and see you @ poster 7!

CWI

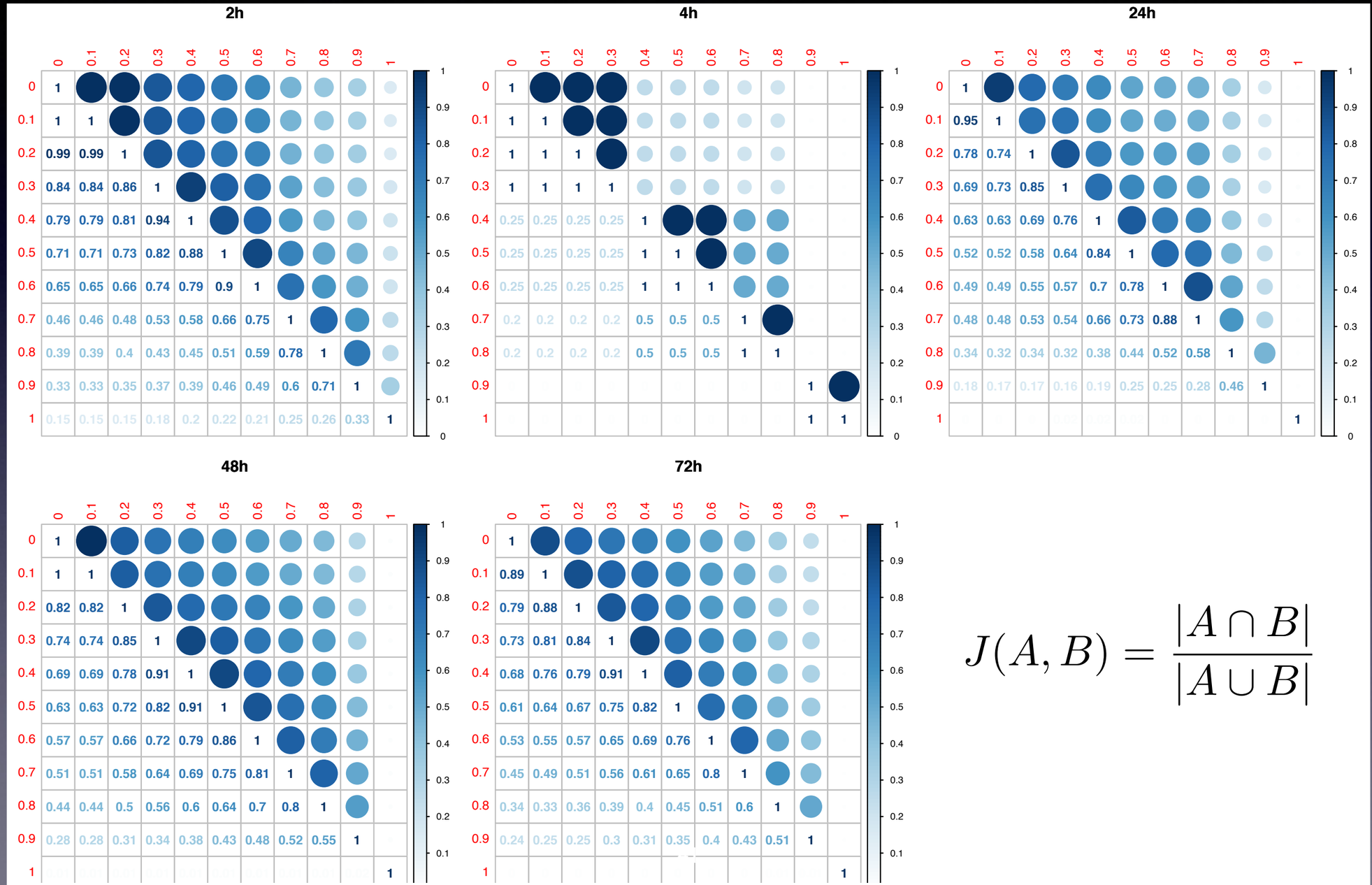
- Wessels group / NKI
- CWI for the expertise
- CBIB/CGFB for the environment (and computing power!)
- VU Centre for Integrative Bioinformatics
- ERASysBio for the funding
- Riitta Laheesma's group for the data
- And you for your attention... and for trying the tool: http://software.cwi.nl/xheinz

ERASysBio Plus

72h: What are the DE IL?

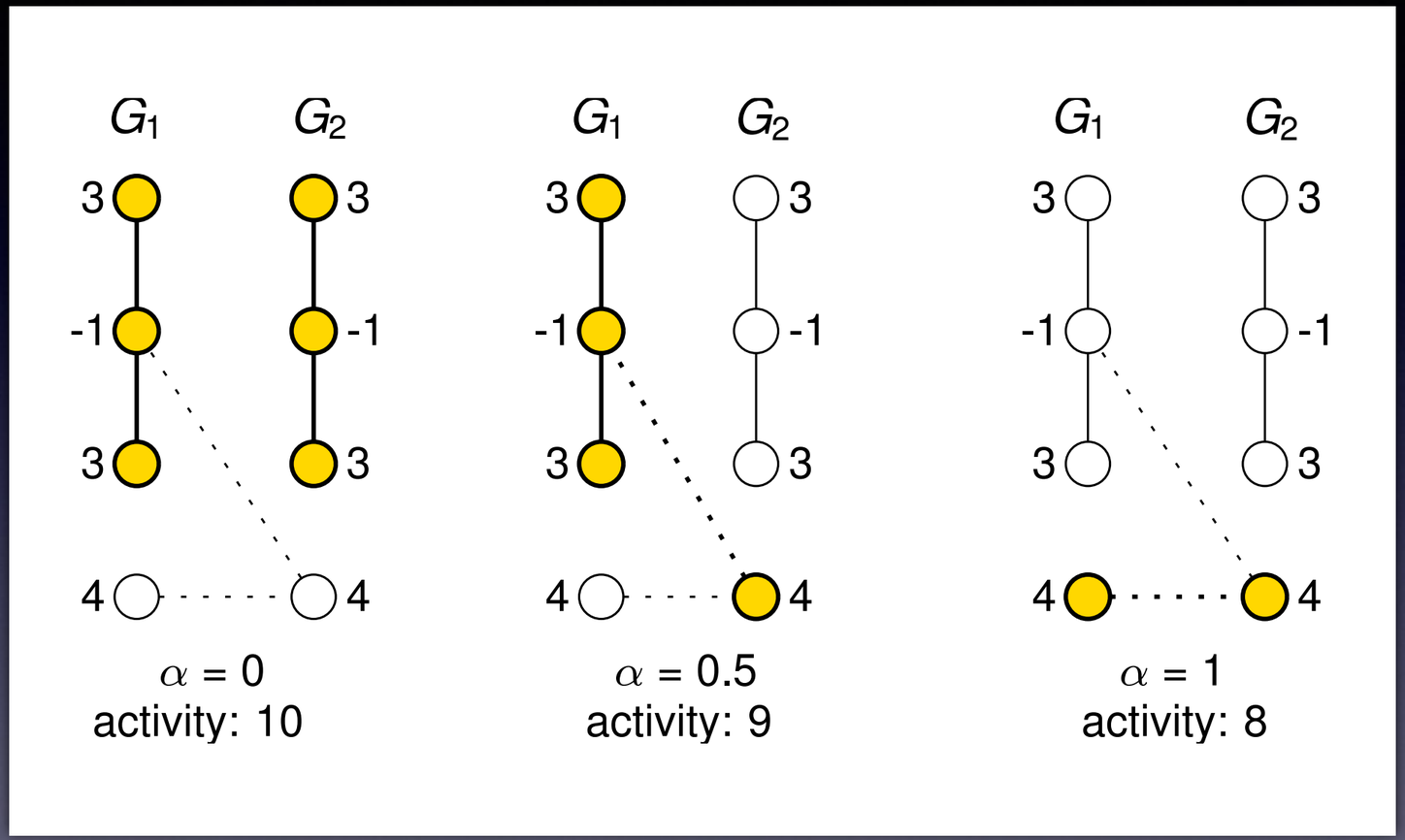


Influence of conservation on content



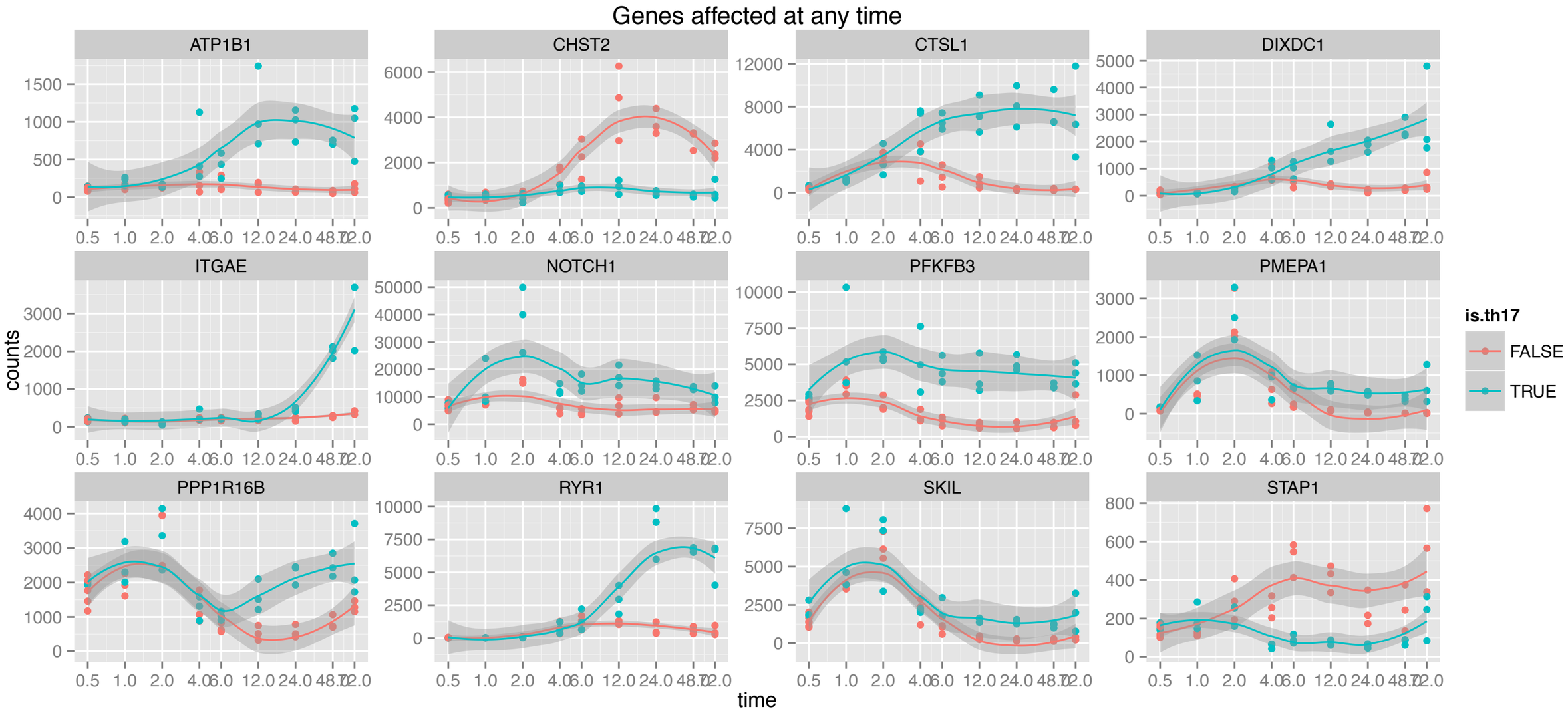
$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

MILP: Conservation tradeoff

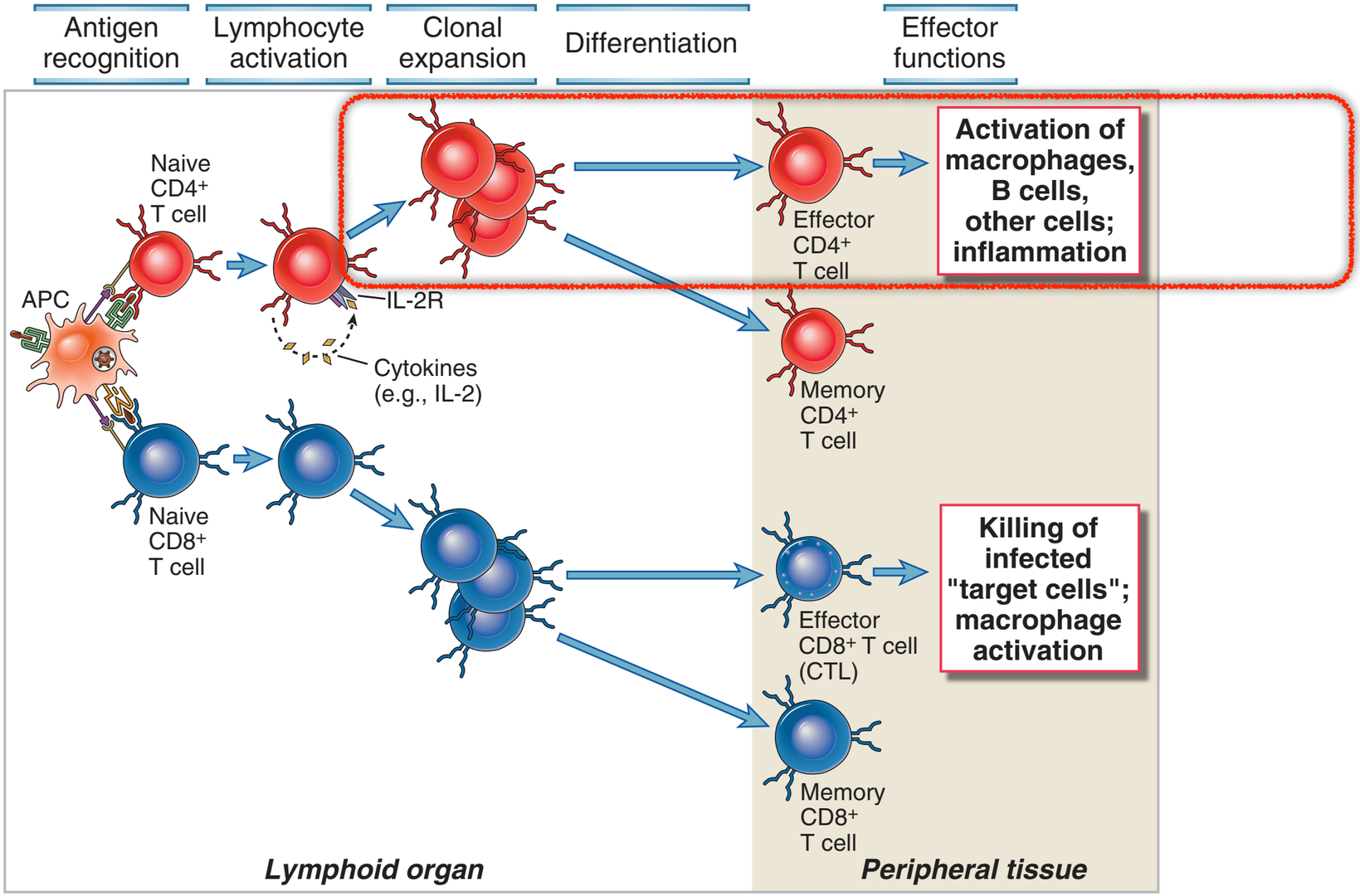


Controls tradeoff between overall activity and conservation

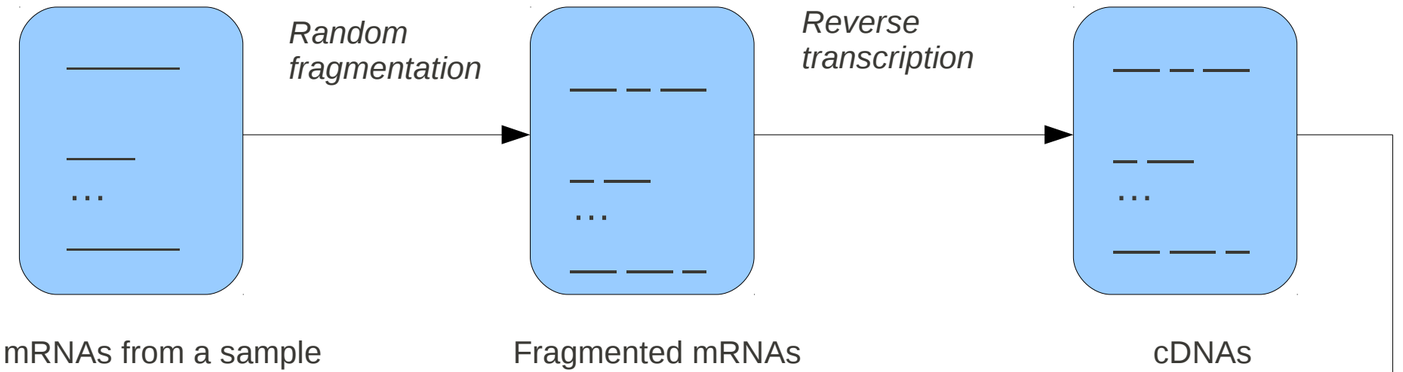
Genes affected at any time point



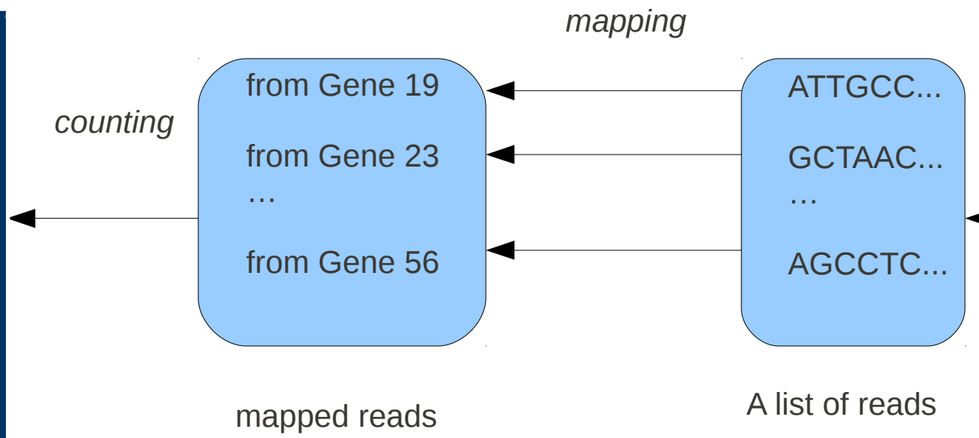
T cells response



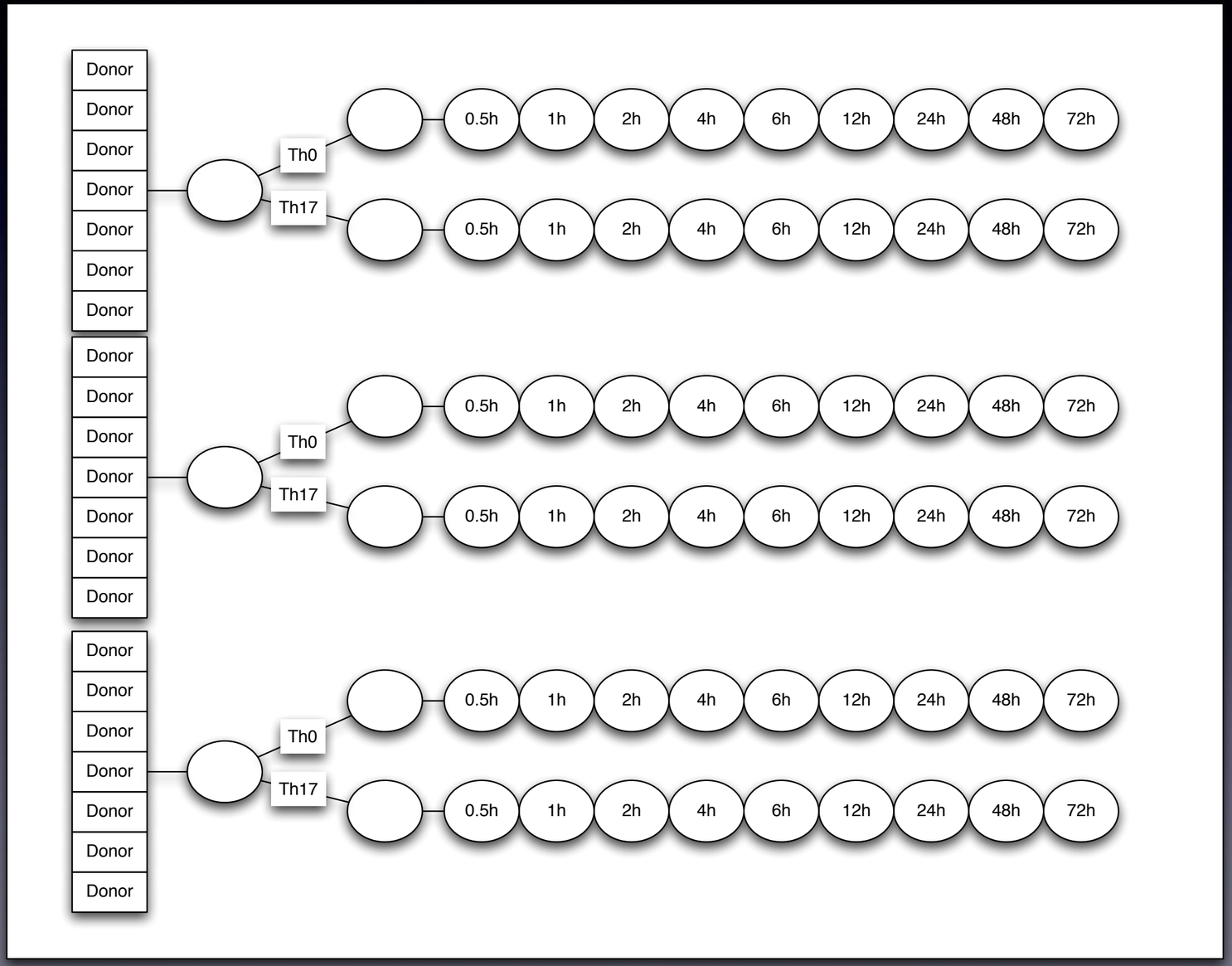
RNA-Seq



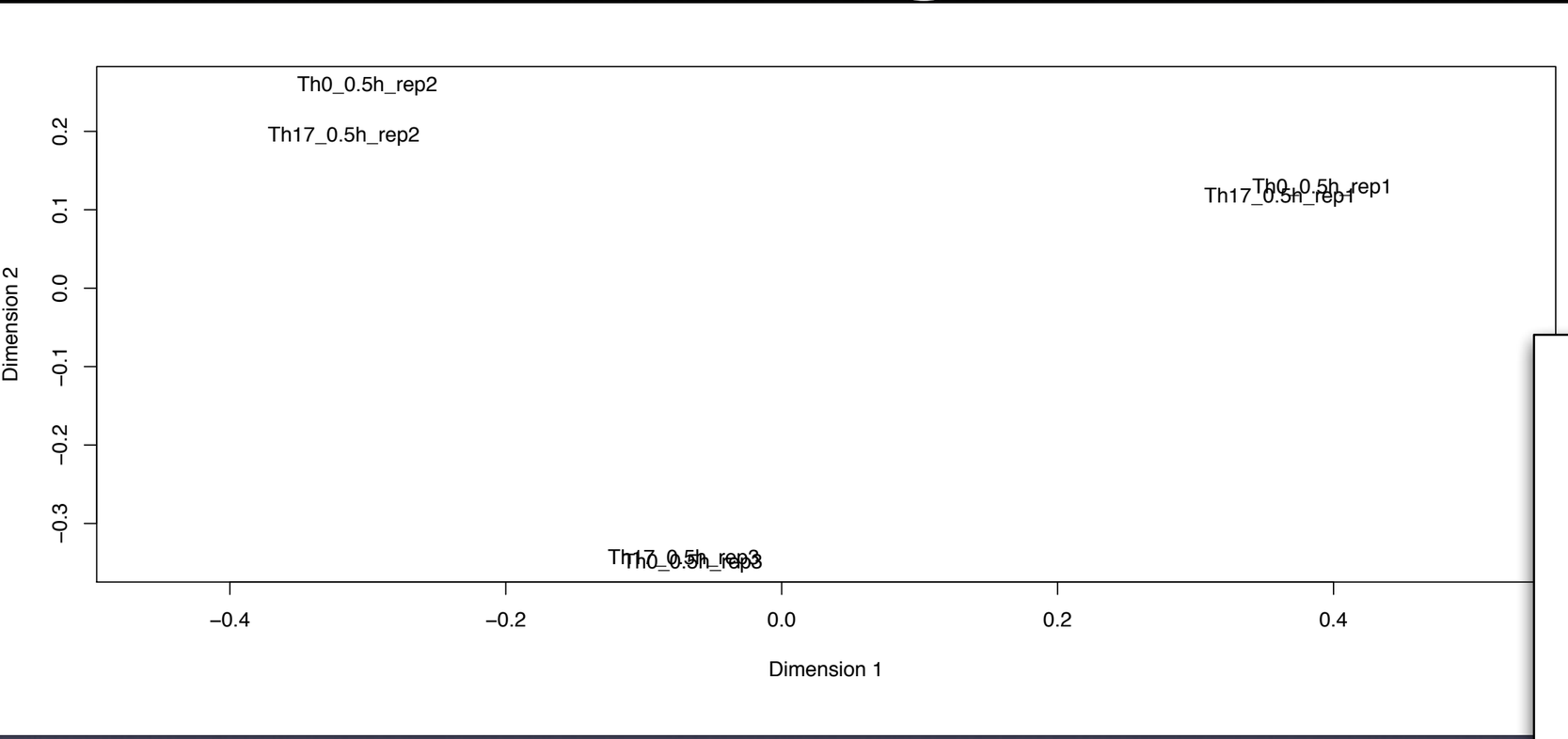
gene	Thp_rep1	Th0_0.5h_rep1	Th0_1h_rep1	Th0_2h_rep1
TAS2R42	0	0	0	0
RP11-140L24.3	6	7	1	6
MIR125B1	0	0	0	0
TRIM27	0	0	0	0
FBLN1	0	0	1	0
AC007919.19	0	0	0	0
EPS15L1	521	593	545	712
CTD-2334D19.1	0	0	0	0
CTA-714B7.5+CTA-714B7.6	12	11	12	10
ZBTB45	154	221	165	254
BAZ1A	2393	2987	3221	4764
FAM83C	0	0	0	0

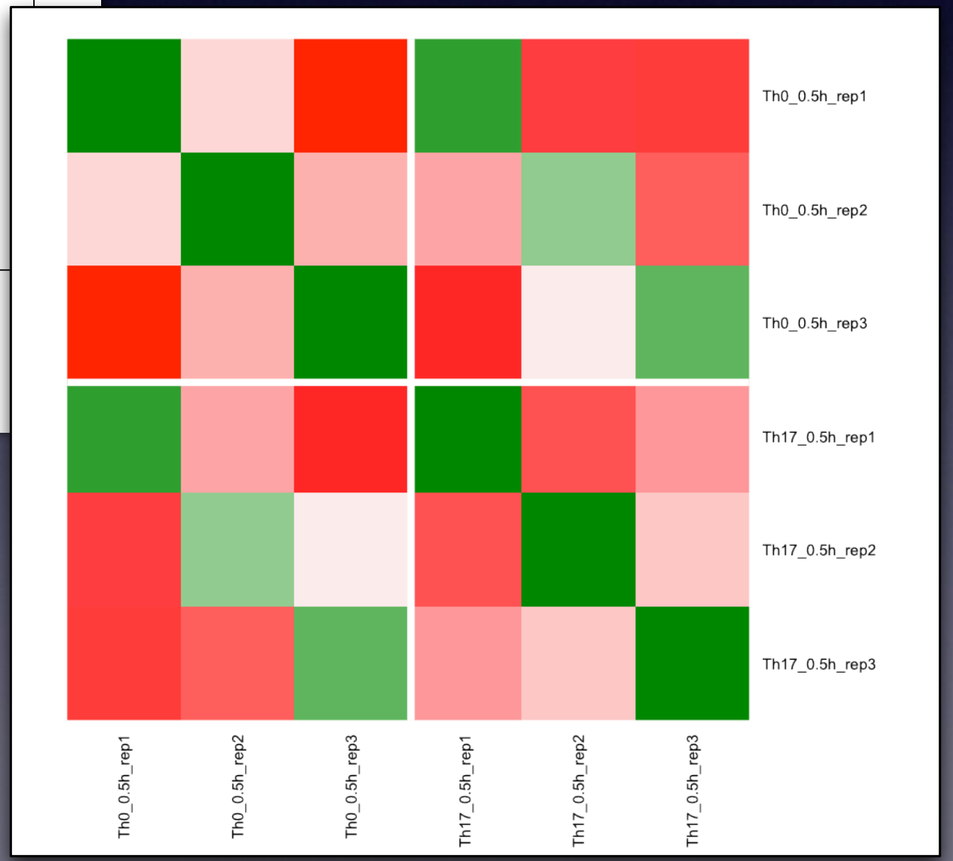


Human exp. design:

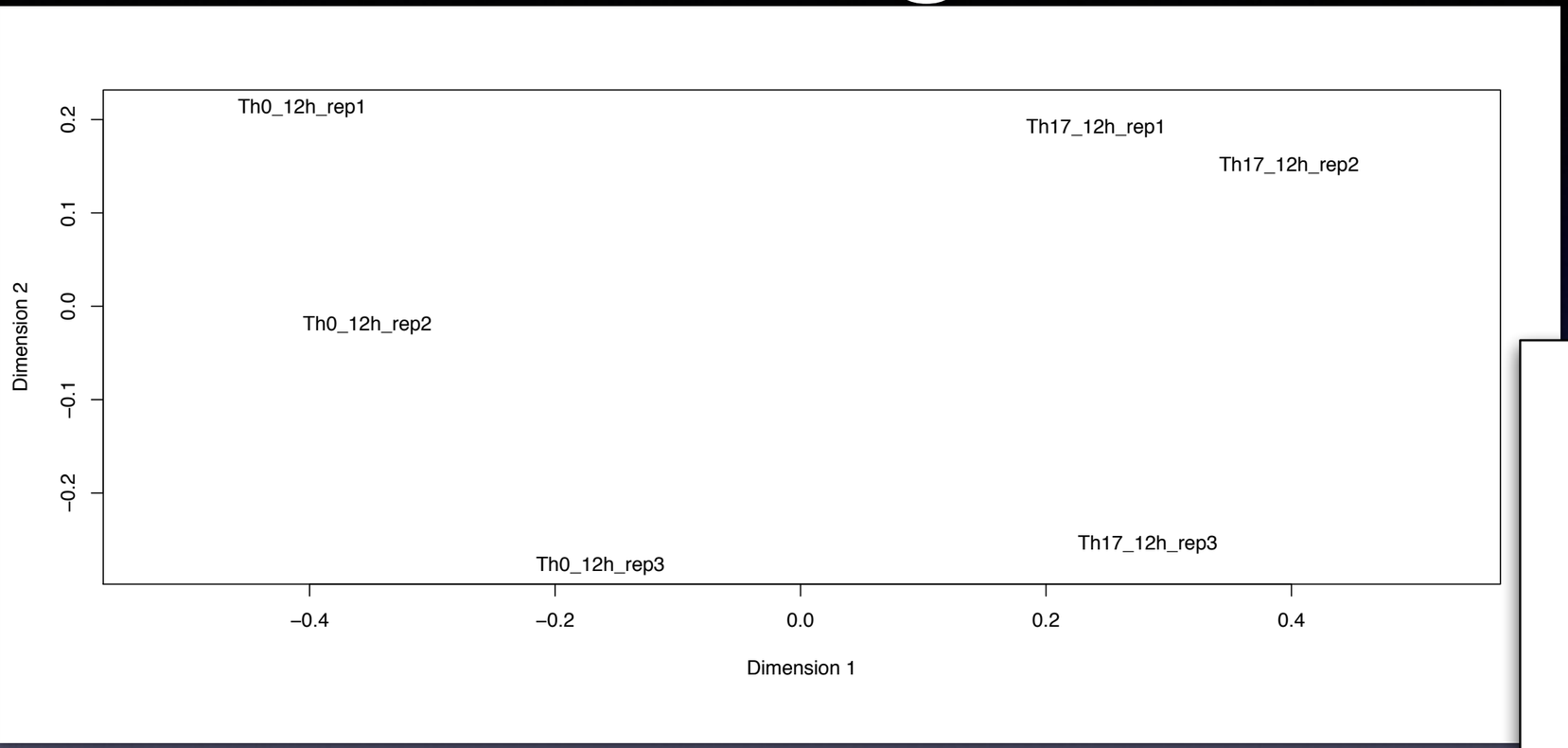


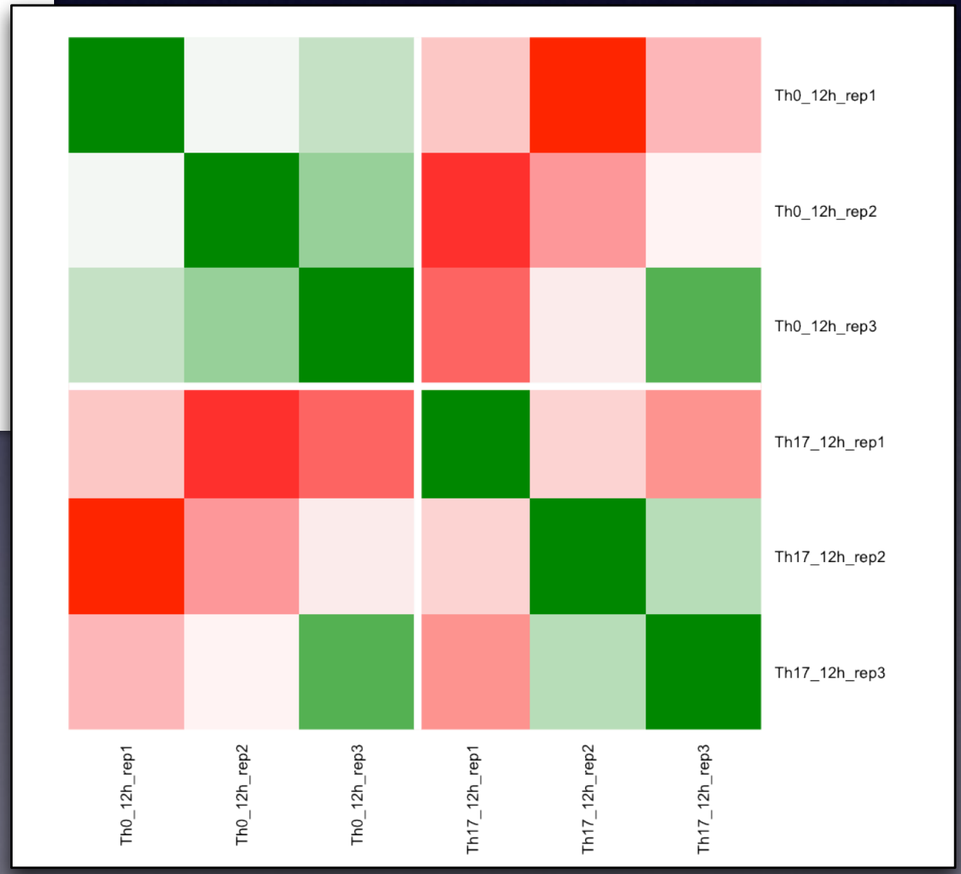
Strongest effects: 0.5h



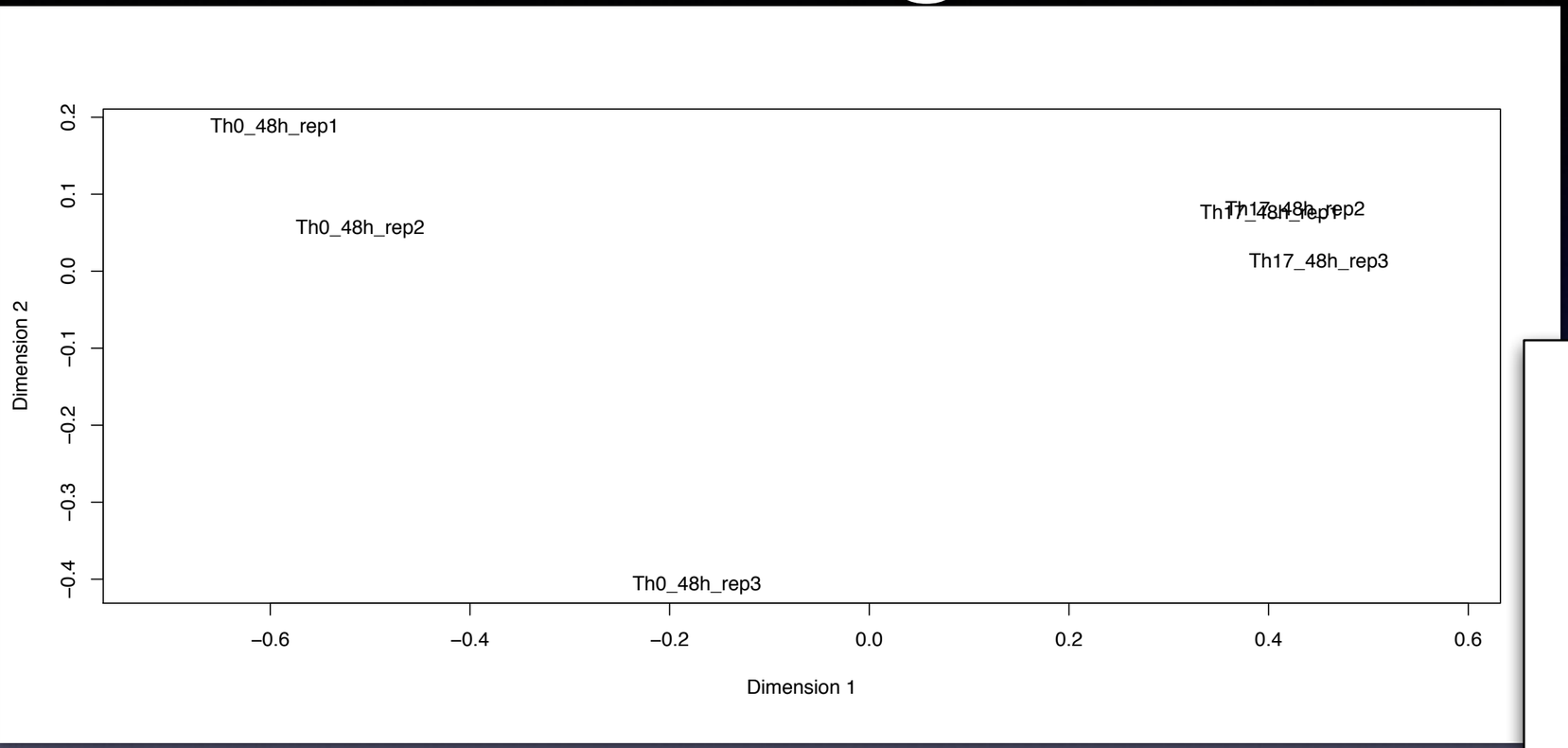


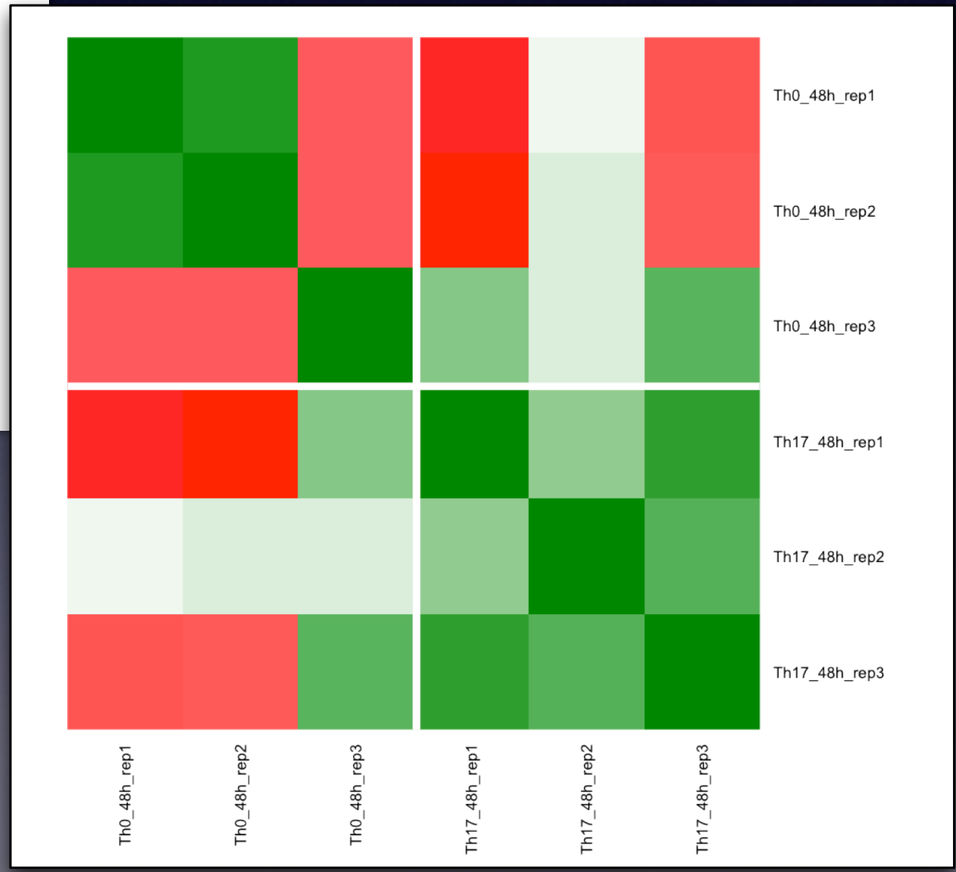
Strongest effects: 12h



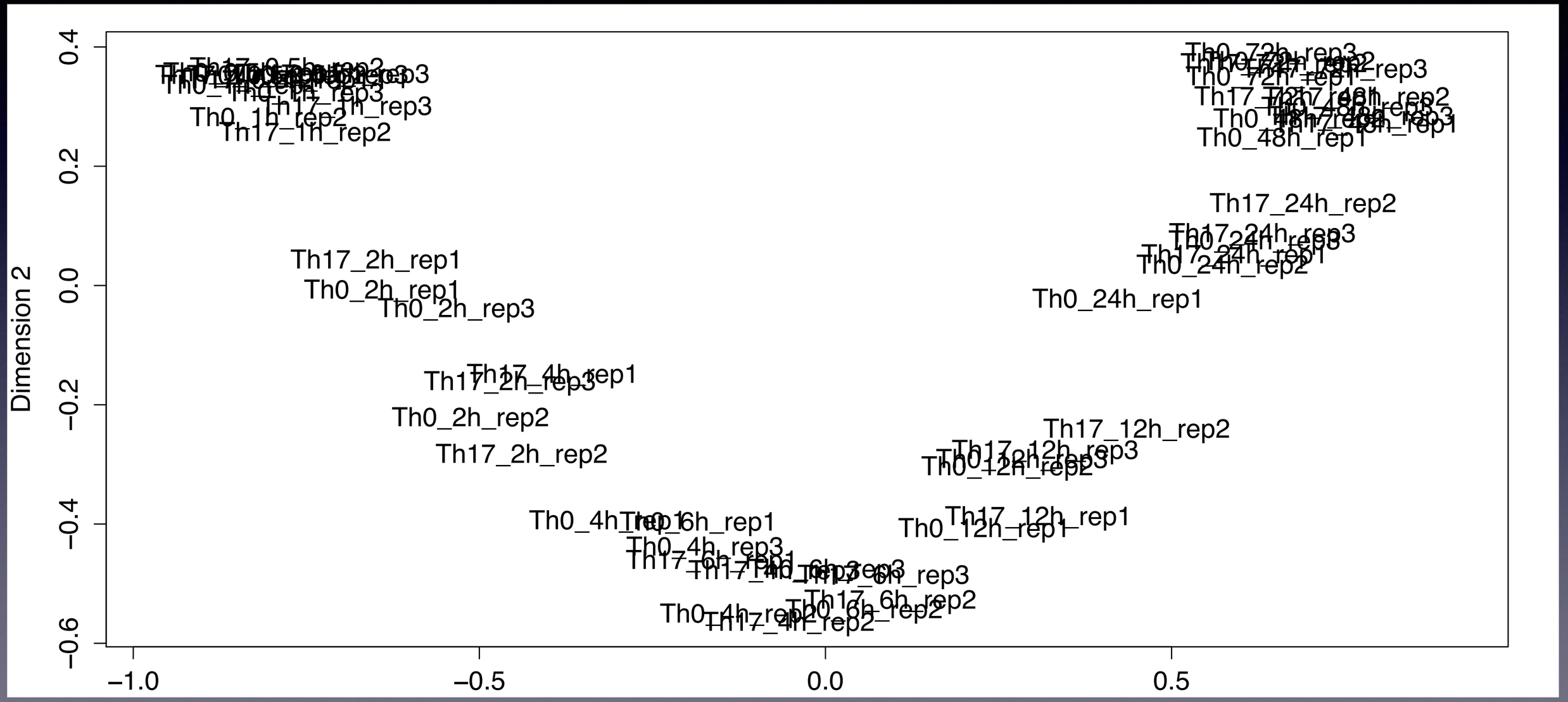


Strongest effects: 48h

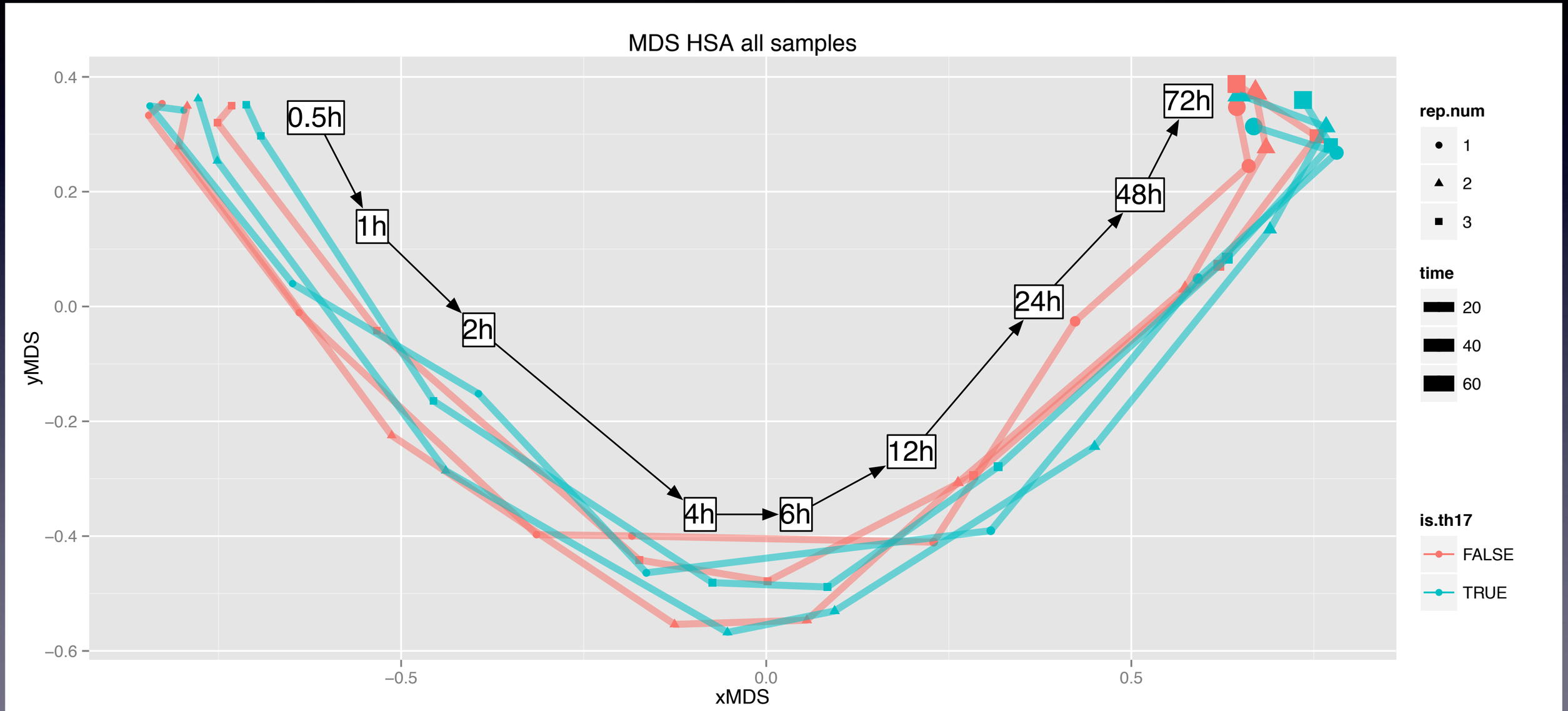




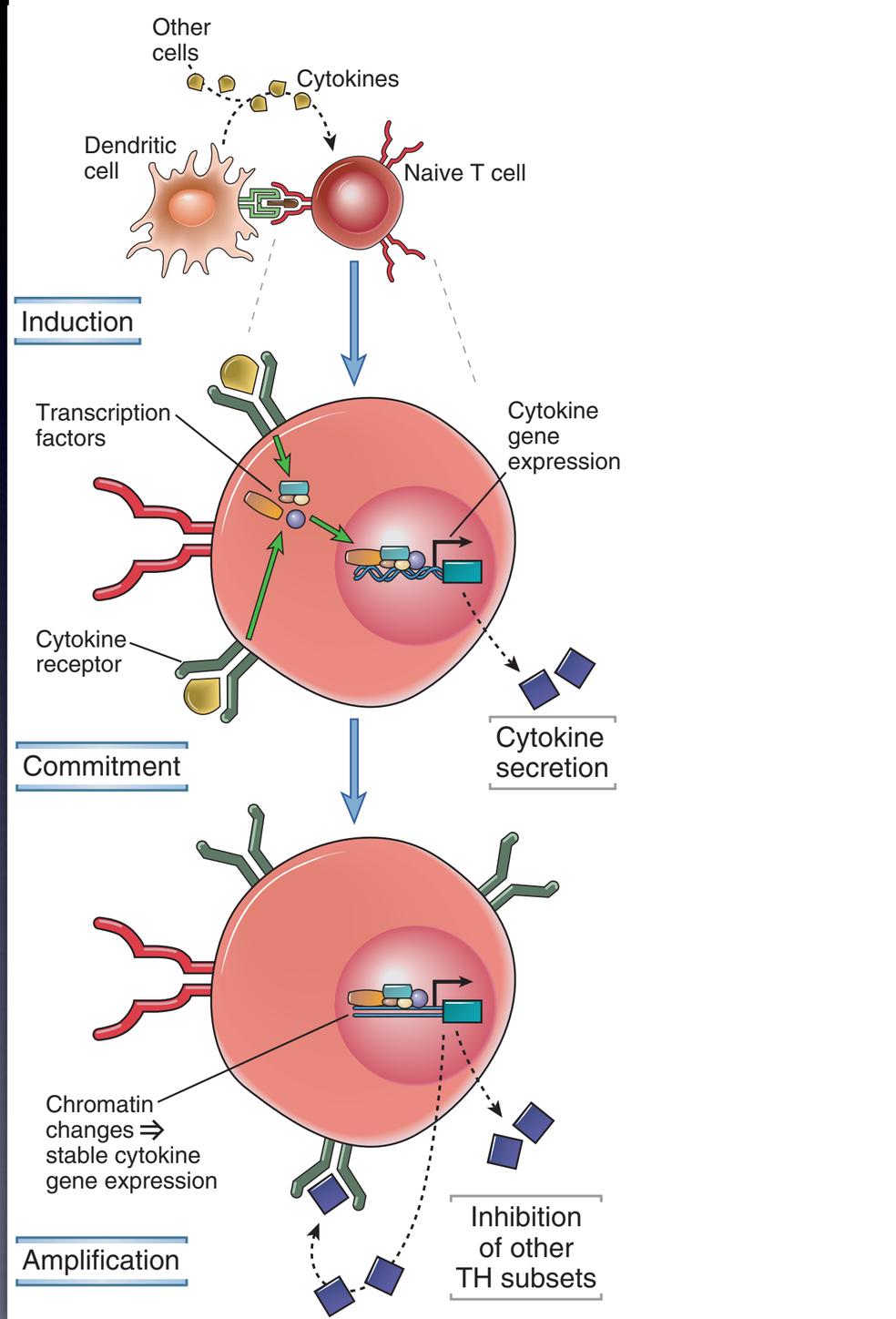
Between all samples



Between all samples

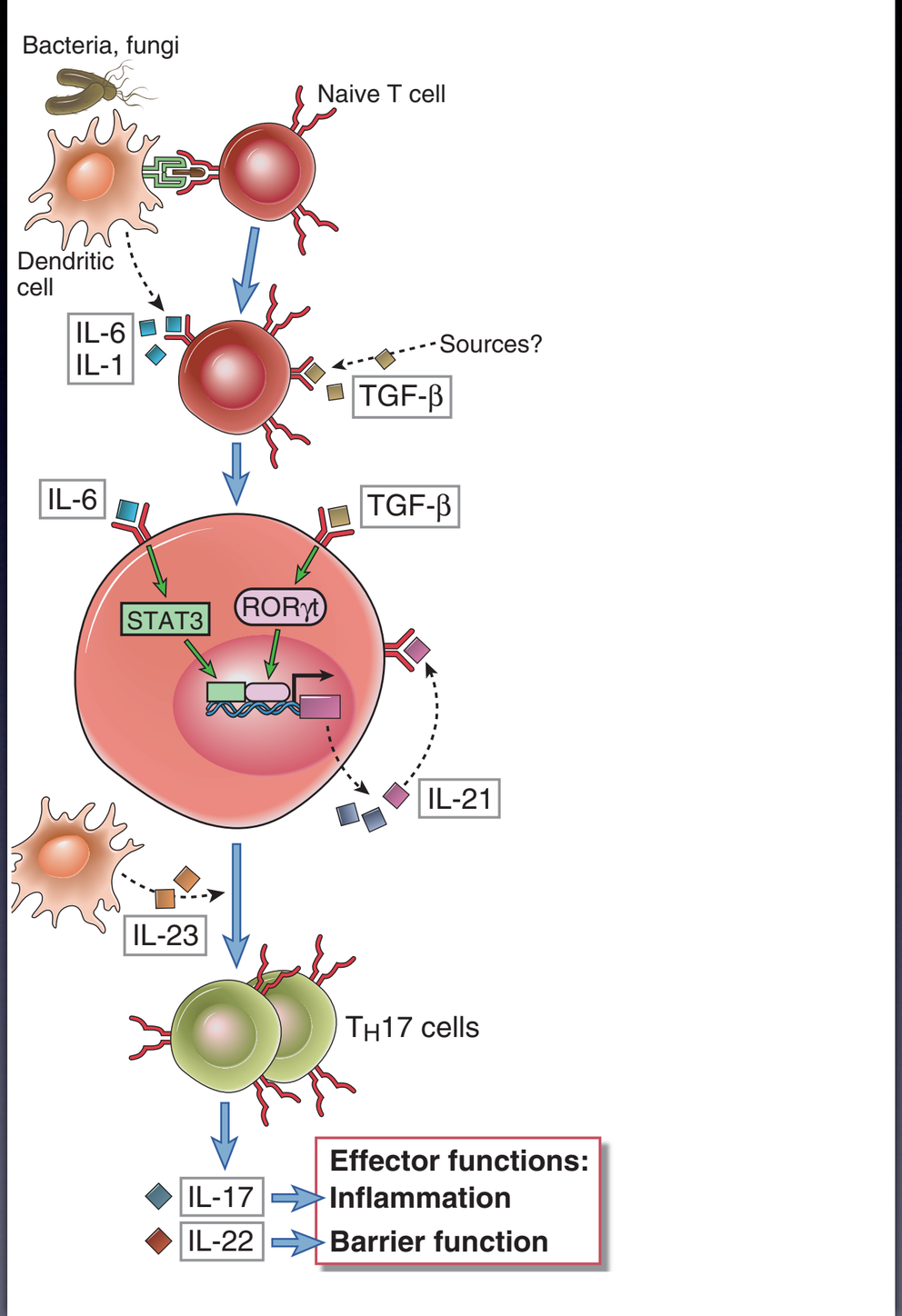


Th cells development & differentiation



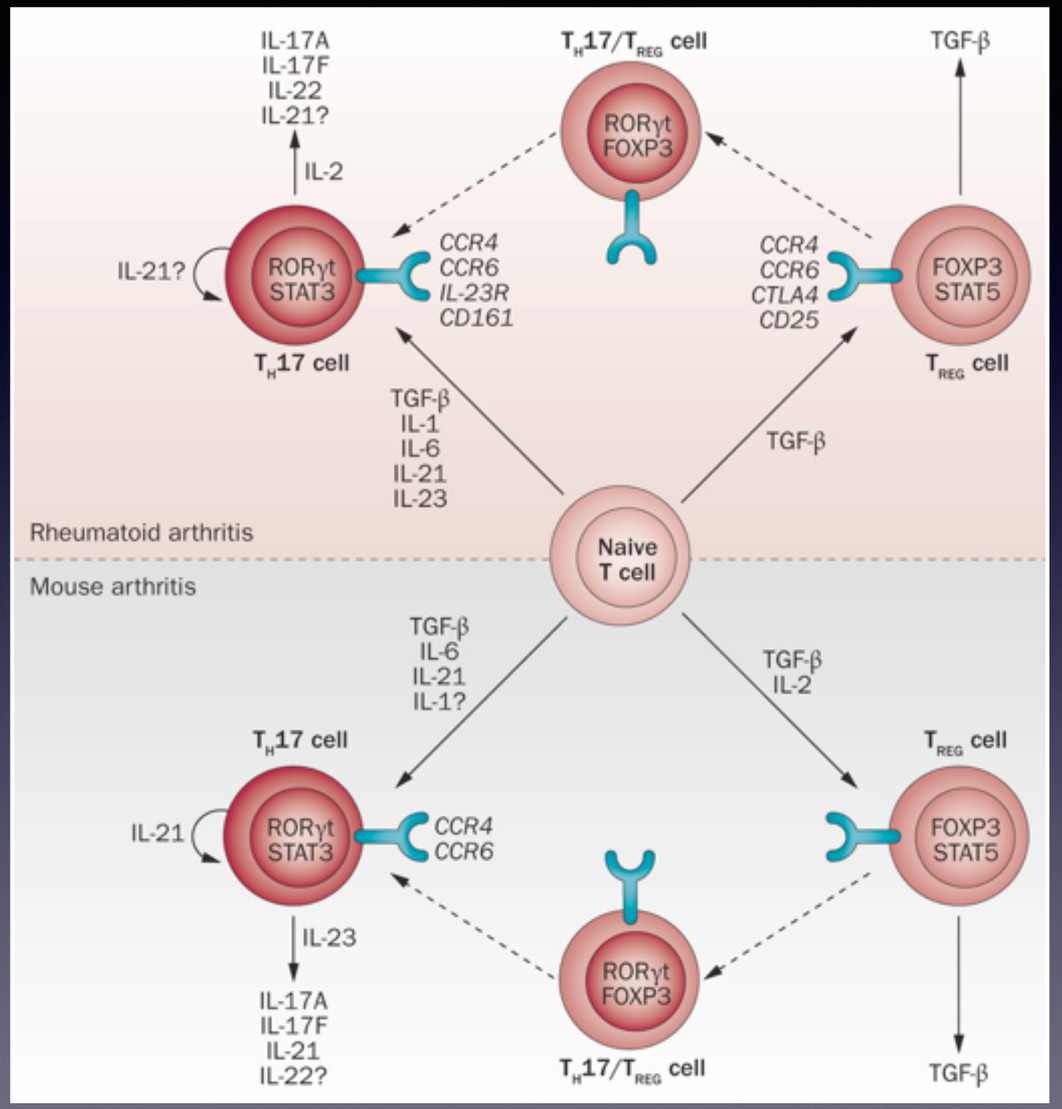
Abbas, Cellular & Molecular Immunology, 7th ed

Th17 cells development & differentiation



Mice vs Men

- Cells origin?
- Role of TGFB?
- Secreted cytokines?



Kobezda et al. (2014). Of mice and men: how animal models advance our understanding of T-cell function in RA. Nat. Reviews. Rheumatology

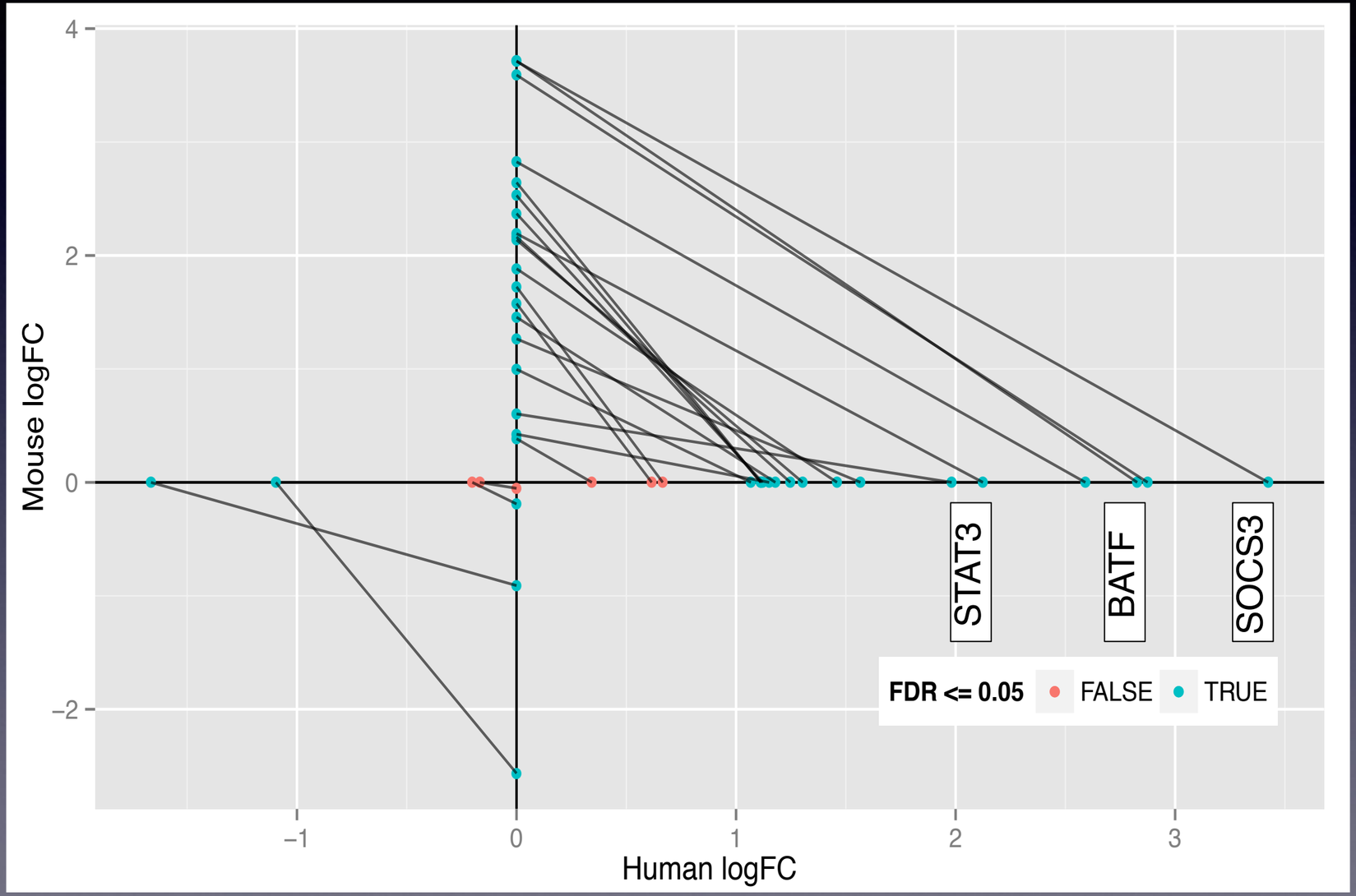
edgeR GLM

$$\log \mu_{gi} = \mathbf{x}_i^T \beta_g + \log N_i$$

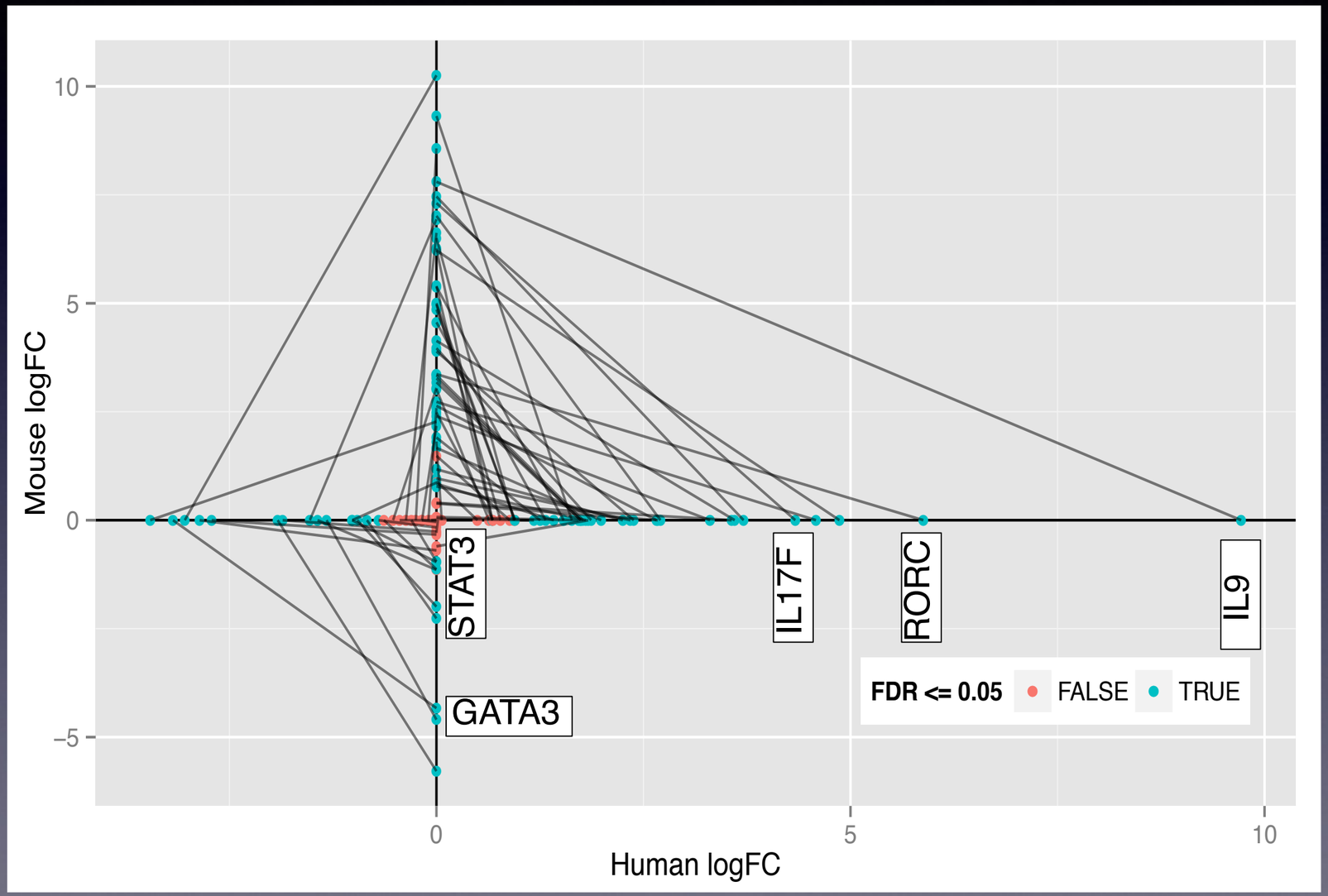
- We fit a model of the like :
- Here: mean ~ donor + time + treat:time
- Test for DE by contrasting <=> H0: treat:time ==0
  - LRT, compare models

counts ~ donor + time VS counts ~ donor + time + time:treat

2h dynamics of conserved module

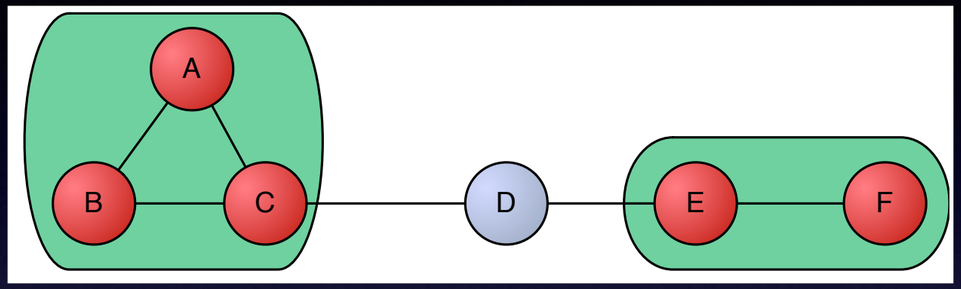


72h dynamics of conserved module



MILP: connectivity constraints

- For each connected component of the current solution S
- Determine its neighborhood not in S
- Formulate the two alternatives:
  - It's expanded towards other CCs
  - Or it'll be the final module (new y_v variables)



$$x_A \leq x_D + y_A + y_B + y_C$$
$$x_B \leq x_D + y_A + y_B + y_C$$
$$x_C \leq x_D + y_A + y_B + y_C$$

$$x_E \leq x_D + y_E + y_F$$
$$x_F \leq x_D + y_E + y_F$$

$$\text{with} \quad y_v \leq x_v \text{ and } \sum_{v \in V} y_v \leq 1;$$