# Combining co-expression and co-location for gene network inference in porcine muscle development in late gestation

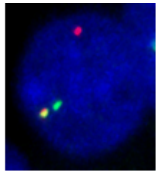
Laurence Liaubet, Nathalie Vialaneix

October 14th, 2019 - NETBIO

Unités MIAT & GenPhySE, INRA Toulouse

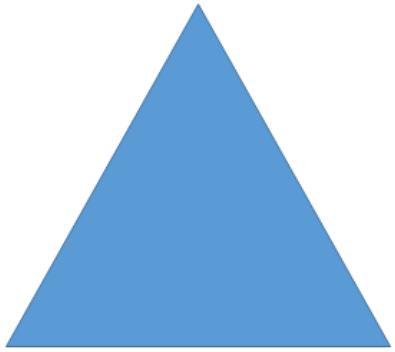
# General context, an interdisciplinary network

A fundamental molecular question, the regulation of gene expression
An ethical and economic breeding problem, survival at birth
A statistical question, modelization of gene co-expression with adding biological information

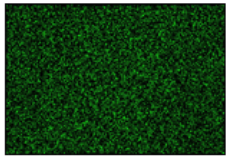


INRA/GenPhySE,
Yvette Lahbib-Mansais and Co.

Cytogenetics



Functional genomics



INRA/GenPhySE,
Laurence Liaubet and Co.

Statistics

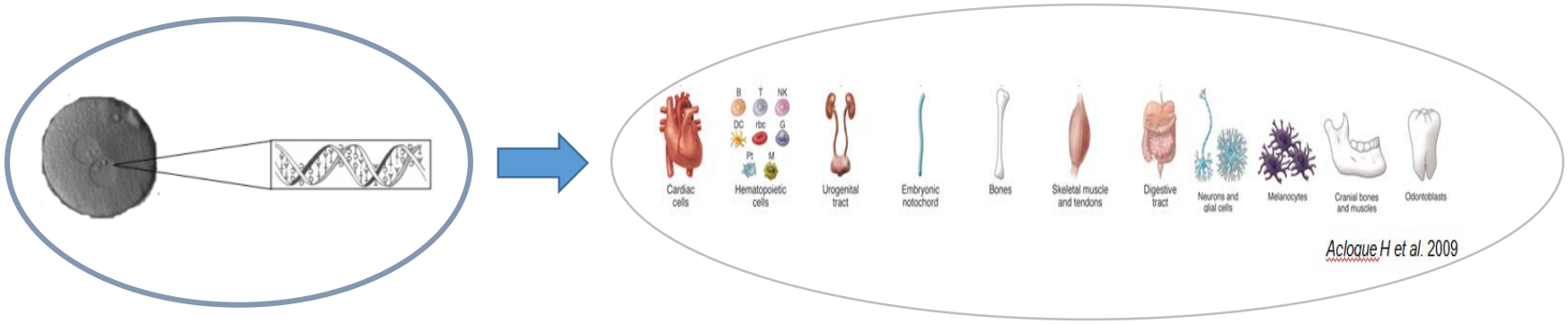
$$\operatorname{argmin}_{\beta_j} \sum_{i=1}^{n} \left(x_i^j - \beta_j^T x_i^{-j}\right)^2 + \lambda \|\beta_j\|_{\epsilon_1} + \mu \left( \sum_{\beta_{jl} \text{ of type 1}} (\beta_{jl} \pm a)^2 + \sum_{\beta_{jl} \text{ of type 2}} (\beta_{jl})^2 \right)$$

INRA/MIA-T,
Nathalie Vialaneix and Co.

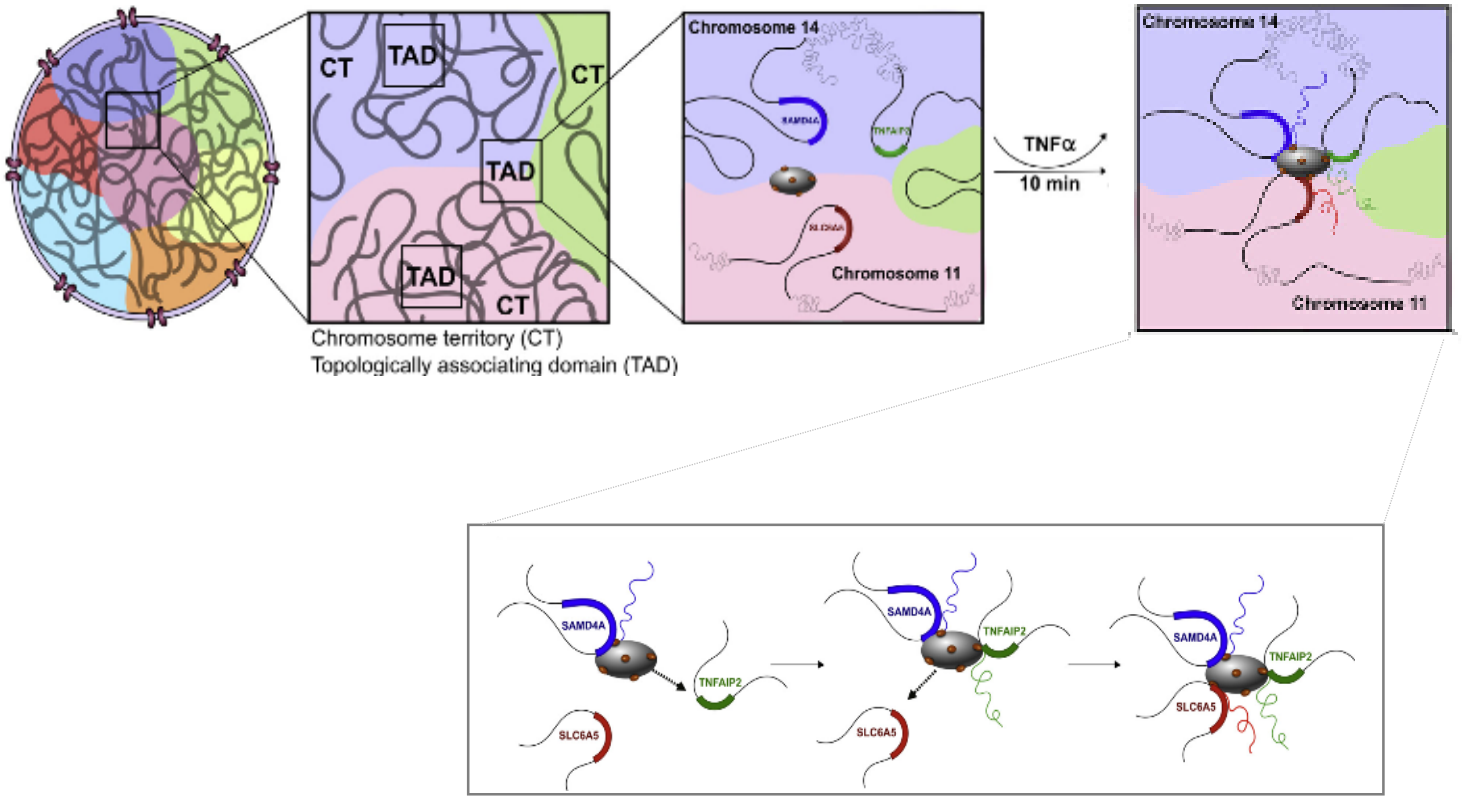
# Biological context: a fundamental molecular question

Interpretation of phenotypes not explained only by genetic approach.
The same genome sequence produce a wide range of differentiated cells.
Modulation of gene expression: Epigenetic marks and chromatin regulation, cis and trans regulation ➔ 3D nuclear topography……

Response to physiological context: growth, health, reproduction, adaptation



# Biological context: a fundamental molecular question




Hierarchical Transcription in a Multigene Complex

Fanucchi et al. Cell 2013

# Biological context: an increased mortality at birth

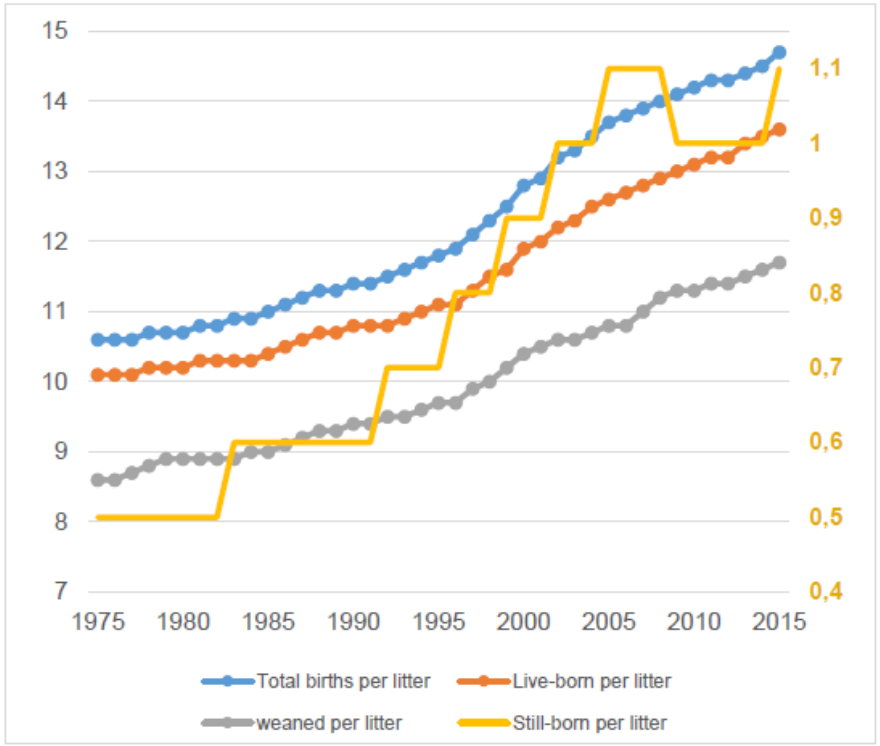



*Figure 1. Evolution of average number of piglets per litter in France from 1975 to 2015. The data used to build this graph were collected and treated by the GTTT (Technical Management of Sow Herds) of the French Porc and Pig Institut (IFIP).*

The selection for more prolificacy and meat production has been accompanied by a substantial increase in mortality of piglets at birth

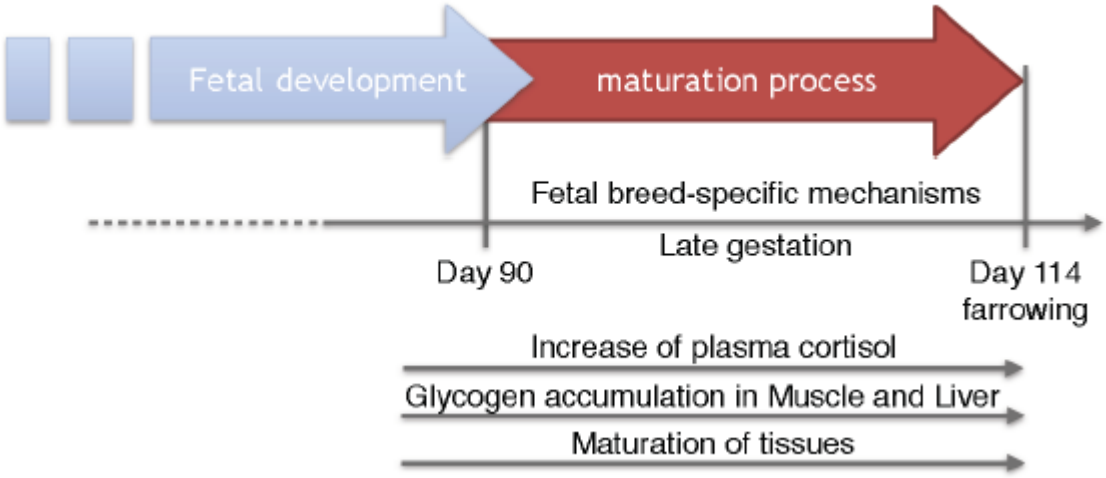
14 % of newborns died between birth and weaning

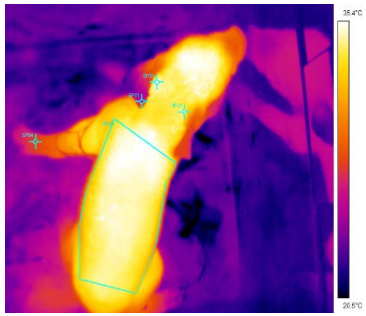
Peak of mortality in the first two days after birth ⟺ maturity

Thèse de Maria Marti-Marimon, 2018

# Specific mechanisms during late gestation in pigs

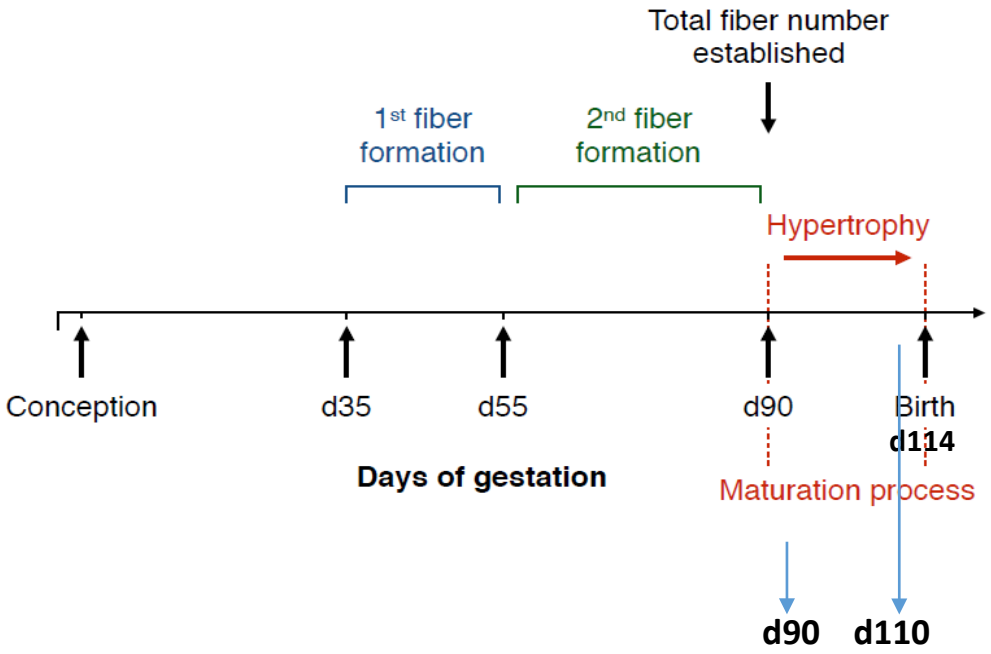
Maturity = plain development allowing survival at birth



Wilson et al., 1998 ; Biensen et al., 1998 ; Leenhouwers et al., 2002 ; Canario, 2006 ; Thèse de Valentin Voillet, 2016

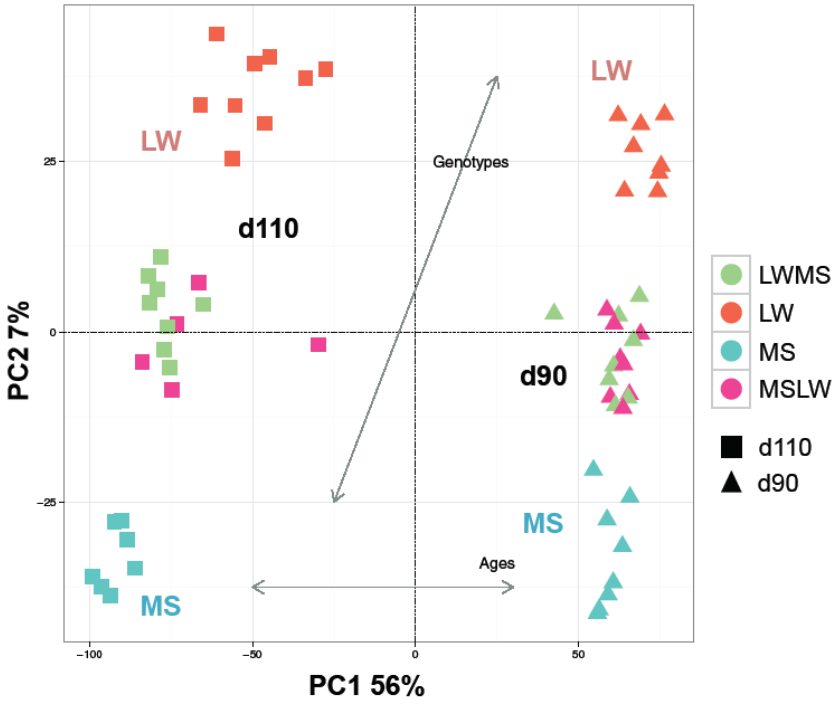
# Specific mechanisms during late gestation in pigs – muscle tissue



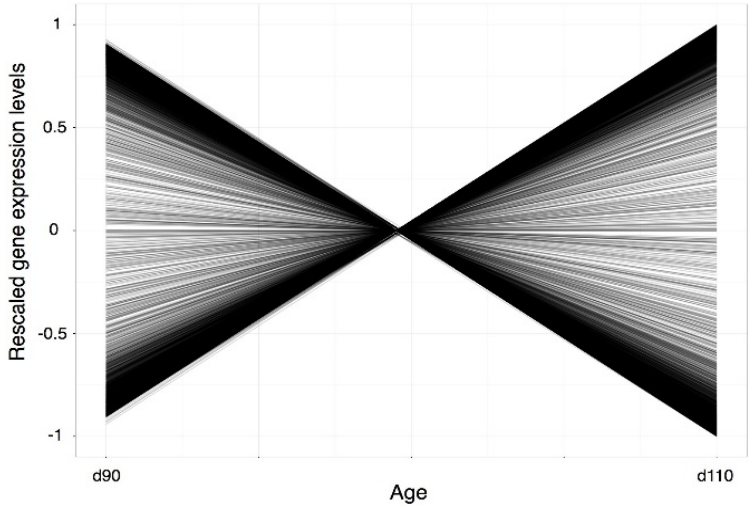

ANR project PORCINET (2010)
Systems biology of piglet maturity

Sampling of *longissimus* (n=459) ➔ Transcriptomic analysis (n = 64)

# A dramatic switch of gene expression occurred in late gestation



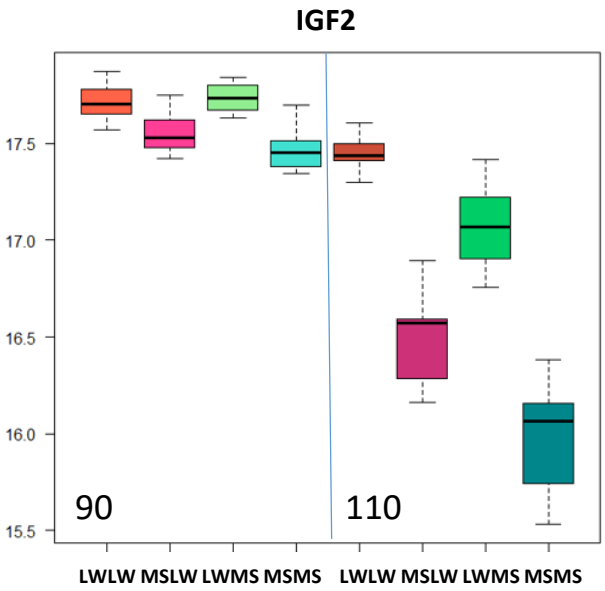

PCA without variables selection




5,167 genes differential between 90 and 110 dg in the fetal muscle (Bonferroni 1%)

With 1,131 DEGs for age x genotype (maturity) found in Voillet et al. (2014)

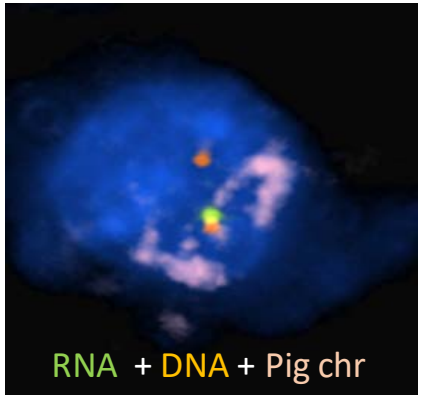
# IGF2, a gene of fundamental importance in pig muscle development

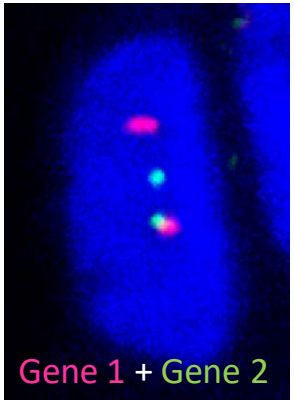



IGF2 (pat), *insulin-like growth factor 2*

Pig ➔ QTL for adiposity muscle mass
Human ➔ fetal growth and intrauterine growth restriction

Co-expression = co-regulation? = nuclear co-localization?




Marti-Marimon et al., 2018.

1. Network inference with GGM

2. Coming back to our problem: gene expression and FISH experiments

# In short, what are we looking for?

Hypothetizing that **co-expression is related to co-location**:

- have an automated process to automatically find relevant pairs of genes for which co-location can be tested (because FISH experiments are time consumming and targeted experiments)

- improve network inference using co-location information

# 1. Network inference with GGM

# Framework

Data: large scale gene expression data

$$\underbrace{\begin{matrix} \text{individuals} \\ n \end{matrix} \left\{ X = \begin{pmatrix} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & X_i^j & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix} \right.}_{\text{variables (gene expressions), } p}$$

here: micro-array experiment, $n = 61$ (gestational age: 90 days) and $p = 13,855$ uniquely annotated genes

# Framework

Data: large scale gene expression data

$$\underbrace{\left\{ X = \begin{pmatrix} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & X_i^j & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix} \right.}_{\text{variables (gene expressions), } p}$$

$$\text{individuals } n$$

here: micro-array experiment, $n = 61$ (gestational age: 90 days) and $p = 13,855$ uniquely annotated genes

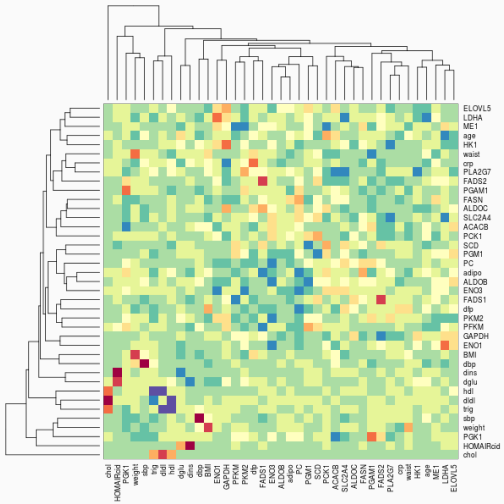
What we want to obtain: a network with

- nodes: genes;
- edges: large and direct co-expression between two genes (track transcription regulations)

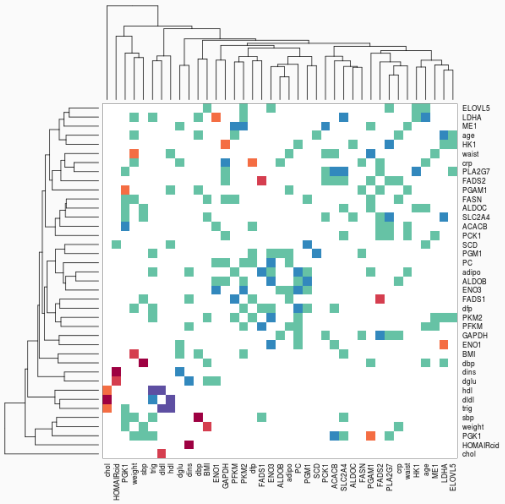
# Using *correlations*: relevance network

**Butte and Kohane (1999, 2000)**

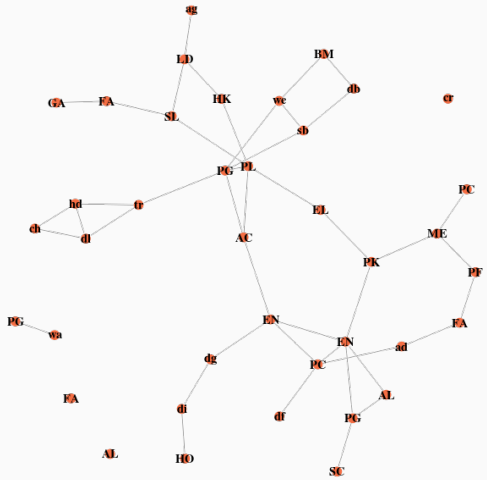
First (naive) approach: calculate correlations between expressions for all pairs of genes, threshold the smallest ones and build the network.



"Correlations"

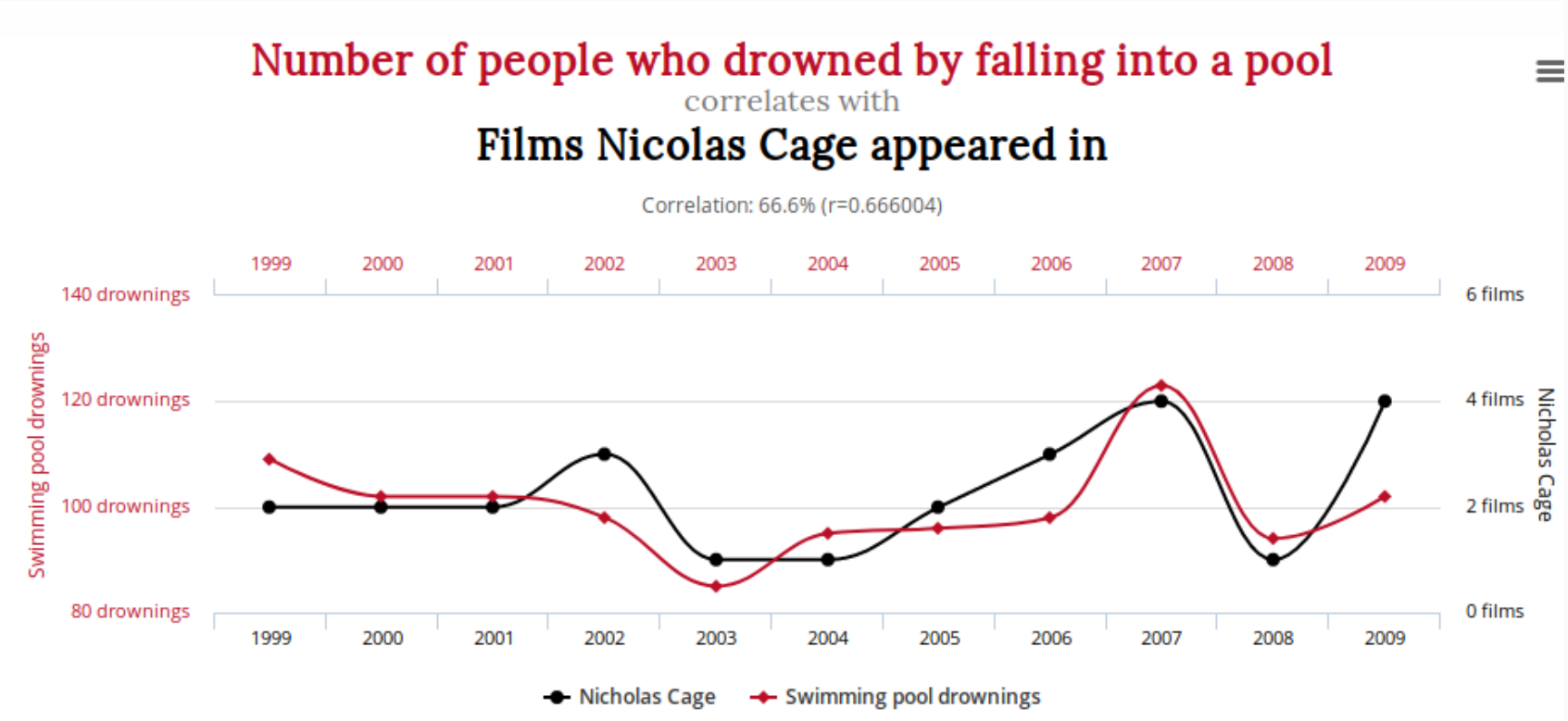


Thresholding

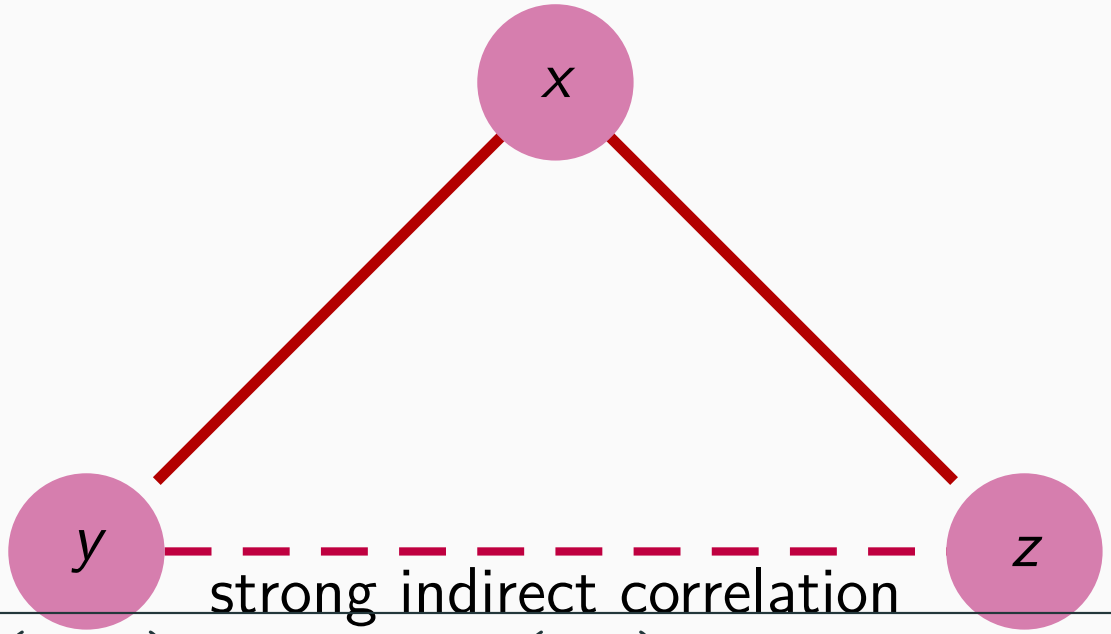


Graph

# But correlation is not causality...

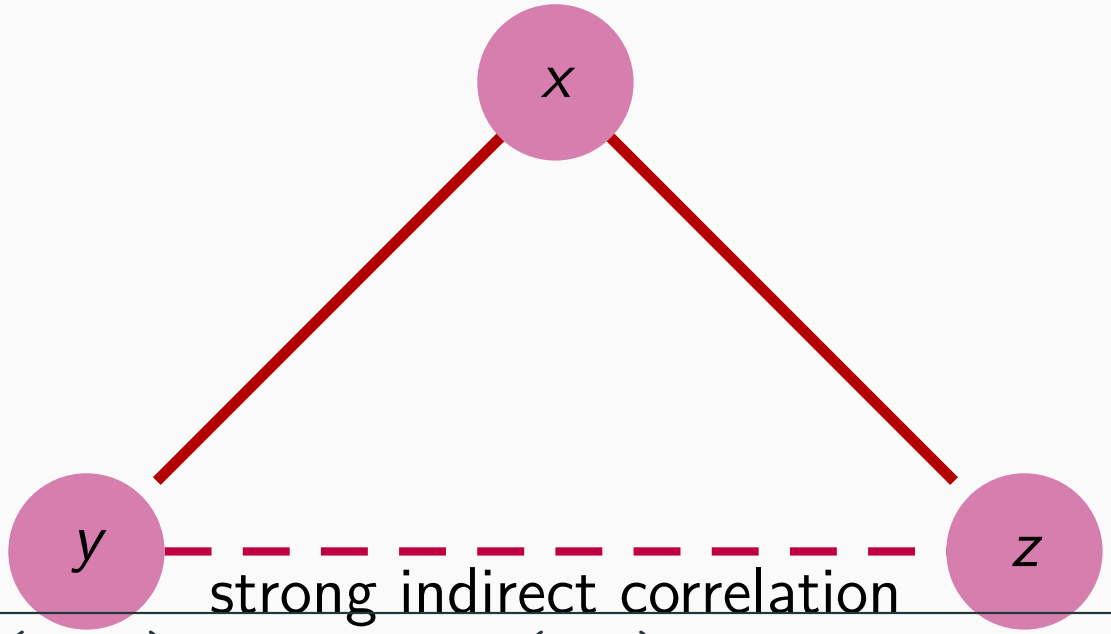


# But correlation is not causality...



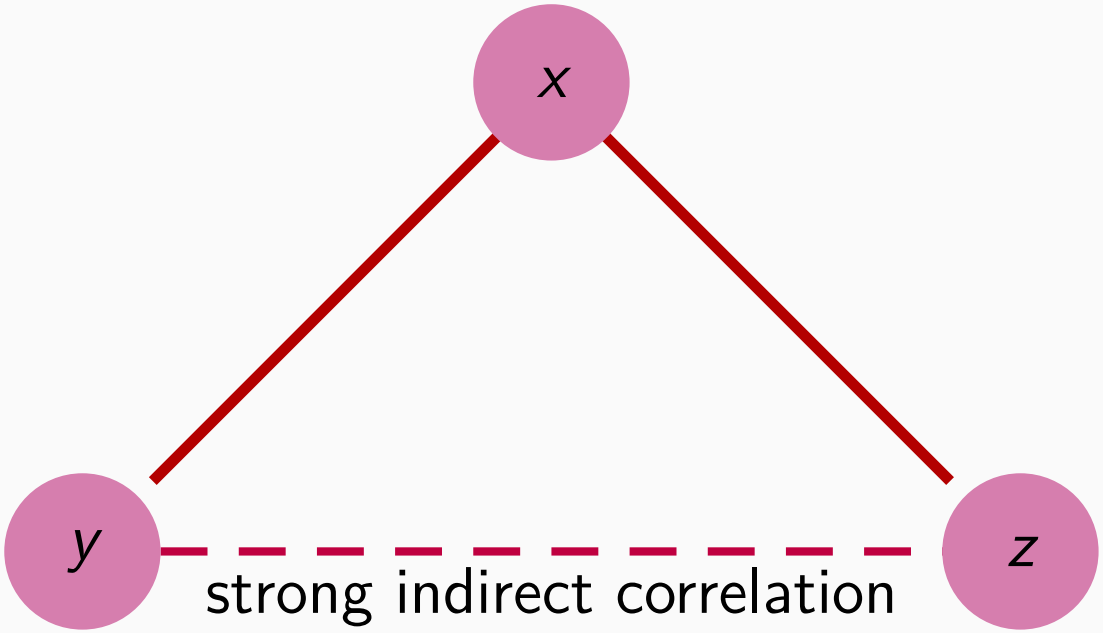

```
set.seed(2807); x <- runif(100)
y <- 2*x+1+rnorm(100,0,0.1); cor(x,y); [1] 0.9988261
z <- 2*x+1+rnorm(100,0,0.1); cor(x,z); [1] 0.998751
cor(y,z); [1] 0.9971105
```

# But correlation is not causality...




```
set.seed(2807); x <- runif(100)
y <- 2*x+1+rnorm(100,0,0.1); cor(x,y); [1] 0.9988261
z <- 2*x+1+rnorm(100,0,0.1); cor(x,z); [1] 0.998751
cor(y,z); [1] 0.9971105
# Partial correlation
cor(lm(y~x)$residuals,lm(z~x)$residuals) [1]
-0.1933699
```

# But correlation is not causality...




Networks are built using partial correlations, i.e., correlations between gene expressions knowing the expression of all the other genes (residual correlations).

# Gaussian Graphical Model (GGM)

$(X_i)_{i=1,\dots,n}$ are i.i.d. Gaussian random variables $\mathcal{N}(0, \Sigma)$ (gene expression); then

$$j \longleftrightarrow j' (\text{genes } j \text{ and } j' \text{ are linked}) \Leftrightarrow \mathbb{C}\text{or}\left(X^j, X^{j'} | (X^k)_{k \neq j,j'}\right) \neq 0$$

# Gaussian Graphical Model (GGM)

$(X_i)_{i=1,\dots,n}$ are i.i.d. Gaussian random variables $\mathcal{N}(0, \Sigma)$ (gene expression); then

$$j \longleftrightarrow j' (\text{genes } j \text{ and } j' \text{ are linked}) \Leftrightarrow \mathbb{C}\text{or}\left(X^j, X^{j'} | (X^k)_{k \neq j,j'}\right) \neq 0$$

$\mathbb{C}\text{or}\left(X^j, X^{j'} | (X^k)_{k \neq j,j'}\right) \simeq \left(\Sigma^{-1}\right)_{j,j'} \Rightarrow$ find the partial correlations by means of $(\widehat{\Sigma}^n)^{-1}$.

# Gaussian Graphical Model (GGM)

$(X_i)_{i=1,\dots,n}$ are i.i.d. Gaussian random variables $\mathcal{N}(0,\Sigma)$ (gene expression); then

$$j \longleftrightarrow j' (\text{genes } j \text{ and } j' \text{ are linked}) \Leftrightarrow \mathbb{C}\text{or}\left(X^j, X^{j'} | (X^k)_{k\neq j,j'}\right) \neq 0$$

$\mathbb{C}\text{or}\left(X^j, X^{j'} | (X^k)_{k\neq j,j'}\right) \simeq \left(\Sigma^{-1}\right)_{j,j'} \Rightarrow$ find the partial correlations by means of $(\widehat{\Sigma}^n)^{-1}$.

Problem: $\Sigma$ is a $p$-dimensional matrix (with $p$ large) and $n$ is small compared to $p \Rightarrow (\widehat{\Sigma}^n)^{-1}$ is a poor estimate of $\Sigma^{-1}$!

# Sparse approaches

Relation between partial correlation and LM: if $S = \Sigma^{-1}$ and writing

$$X^j = \beta_j^\top X^{-j} + \epsilon$$

we have: $\beta_{jj'} = \frac{s_{jj'}}{s_{jj}}$. So edges (non zero partial correlations) also correspond to coefficients different to zero in the $p$ regression models above (for $j = 1, \ldots, p$).

# Sparse approaches

Relation between partial correlation and LM: if $S = \Sigma^{-1}$ and writing

$$X^j = \beta_j^\top X^{-j} + \epsilon$$

we have: $\beta_{jj'} = \frac{s_{jj'}}{s_{jj}}$. So edges (non zero partial correlations) also correspond to coefficients different to zero in the $p$ regression models above (for $j = 1, \ldots, p$).

To ensure sparsity of $\beta_j$: Meinshausen and Bühlmann (2006)

$$\text{argmin}_{\beta_j} \sum_{i=1}^{n} \left( x_i^j - \beta_j^\top x_i^{-j} \right)^2 + \lambda \|\beta_j\|_{\ell_1}$$

# Including prior knowledge in this model

Suppose that we have some clues that:

- for some pairs $(j, j')$, an edge is likely to occur between $j$ and $j'$
- for some pairs $(j, j')$, it is likely that there is no edge between $j$ and $j'$

# Including prior knowledge in this model

Suppose that we have some clues that:

- for some pairs $(j, j')$, an edge is likely to occur between $j$ and $j'$
- for some pairs $(j, j')$, it is likely that there is no edge between $j$ and $j'$

then, we want to drive $\beta_{jj'}$

- toward $\pm a$ with $a$ some positive value (the sign is that of the correlation $\mathrm{Cor}(X^j, X^{j'})$)
- toward 0

# Including another penalty in the model

$$\text{argmin}_{\beta_j} \sum_{i=1}^{n} \left( x_i^j - \beta_j^\top x_i^{-j} \right)^2 + \lambda \|\beta_j\|_{\ell_1} + \underbrace{\mu \left( \sum_{j' \text{ of type 1}} (\beta_{jj'} \pm a)^2 + \sum_{j' \text{ of type 2}} (\beta_{jj'})^2 \right)}_{\text{smooth penalty for co-localized (or not) pairs}}$$

# Including another penalty in the model

$$\text{argmin}_{\beta_j} \sum_{i=1}^{n} \left( x_i^j - \beta_j^\top x_i^{-j} \right)^2 + \lambda \|\beta_j\|_{\ell_1} + \underbrace{\mu \left( \sum_{j' \text{ of type 1}} (\beta_{jj'} \pm a)^2 + \sum_{j' \text{ of type 2}} (\beta_{jj'})^2 \right)}_{\text{smooth penalty for co-localized (or not) pairs}}$$

In practice:

- $a = 1$ (after scaling of gene expressions)
- $\lambda$ chosen with stability selection based on bootstrap for a fixed $\mu$
- $\mu$ chosen as the minimum value recovering exactly prior knowledge

# 2. Coming back to our problem: gene expression and FISH experiments

# Starting point

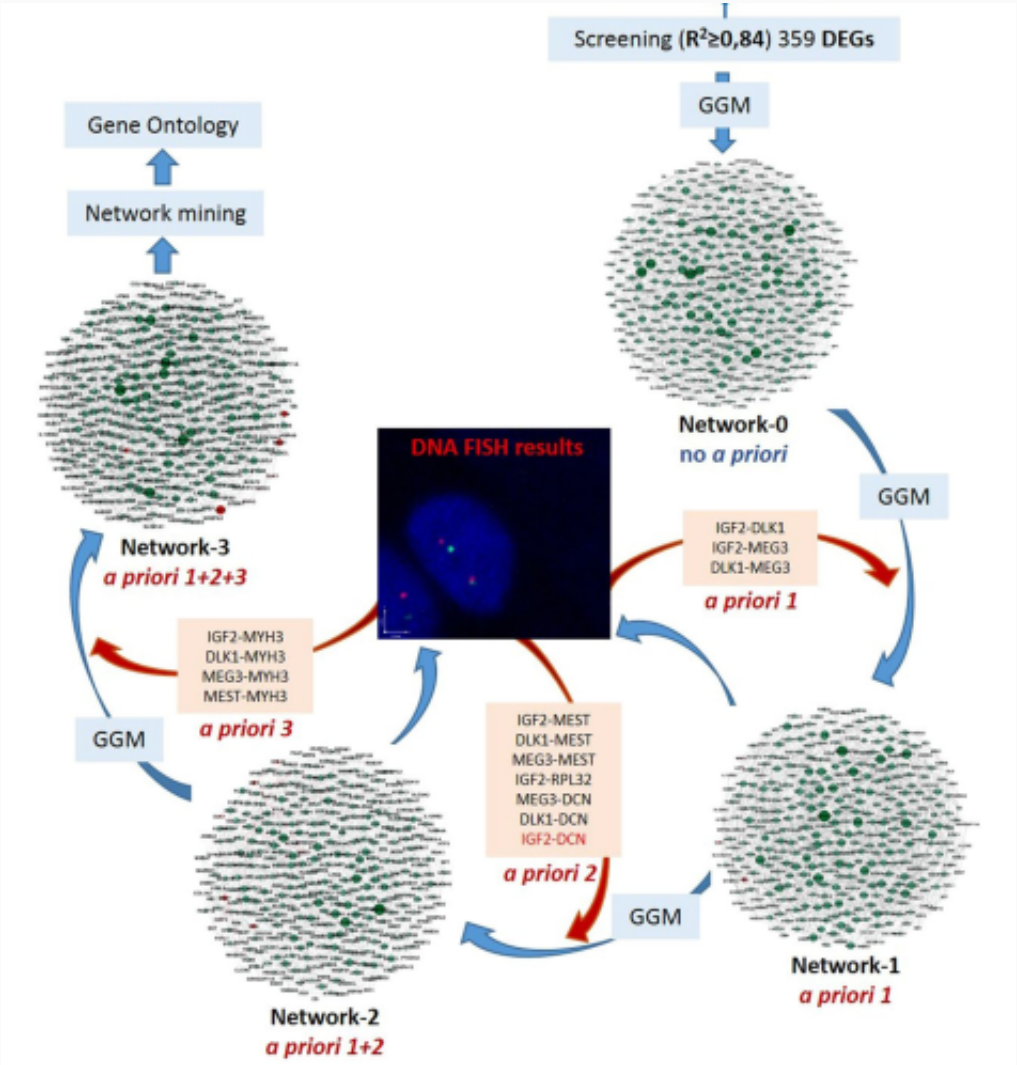
- restricted to a list of genes likely involved in foetal development (1,131 DEGs between 90 and 110 days of gestation as found in Voillet et al. (2014))

# Starting point

- restricted to a list of genes likely involved in foetal development (1,131 DEGs between 90 and 110 days of gestation as found in Voillet et al. (2014))

- started from an even more restricted list including genes of interest (*IGF2*, *DLK1* and *MEG3*) and the genes highly correlated to these genes ($p = 359$ genes at the end)

# Iterative process: from co-location to network and conversely



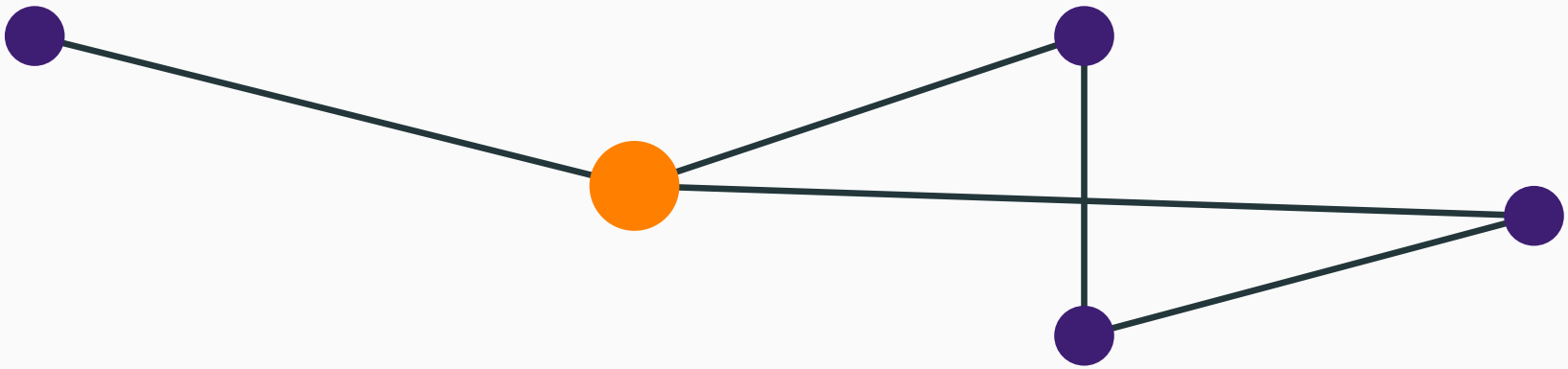
# What to do with these networks? Network mining...

1. Node importance
2. Clustering of nodes (and comparison of clustering with NMI)
3. GO analysis

# Node importance

For every node, computation of:

- degree (number of edges afferent to a given node)
- betweenness centrality measure

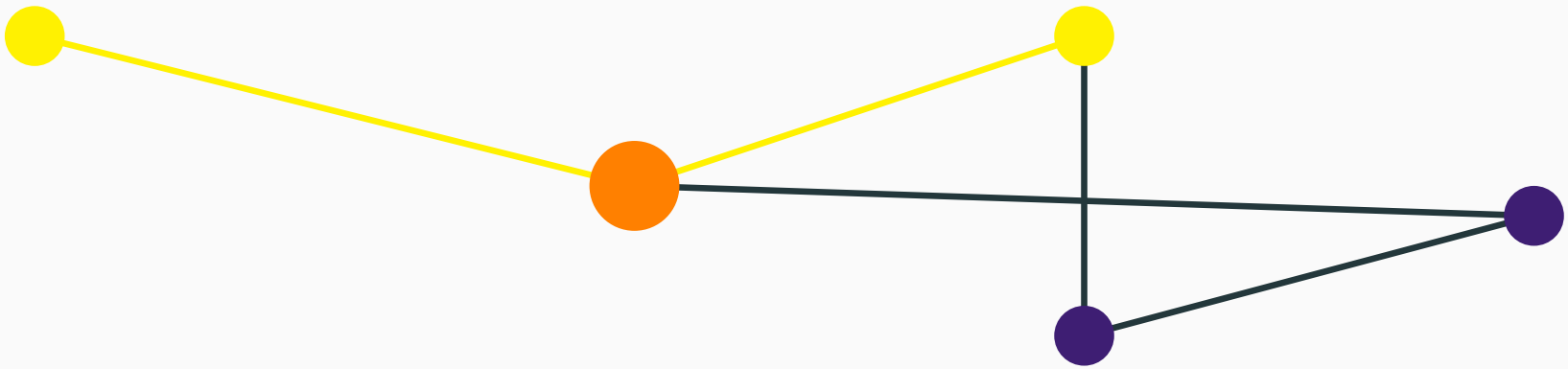


The orange node's degree is equal to 2, its betweenness to 4.

# Node importance

For every node, computation of:

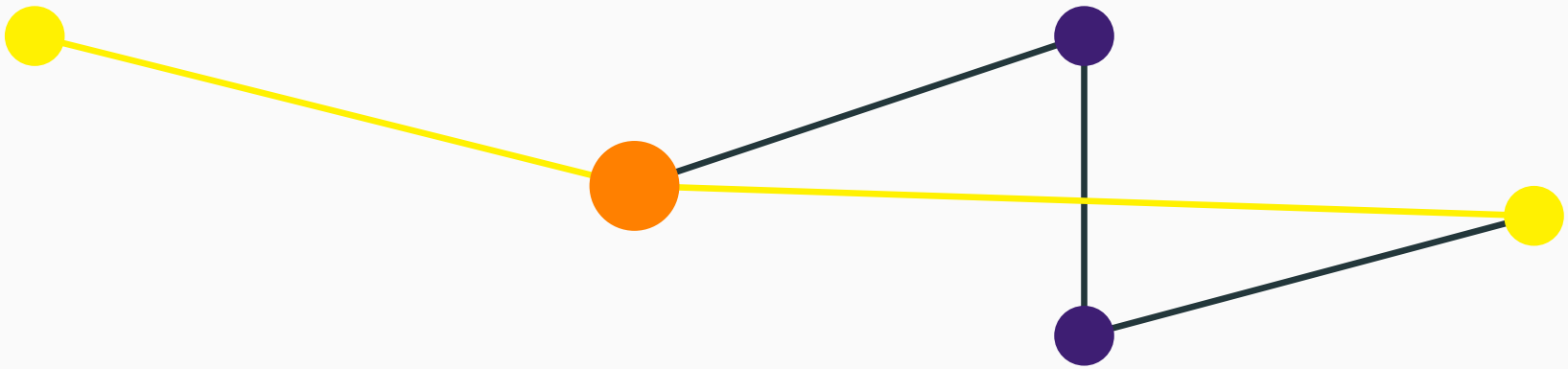
- degree (number of edges afferent to a given node)
- betweenness centrality measure



# Node importance

For every node, computation of:

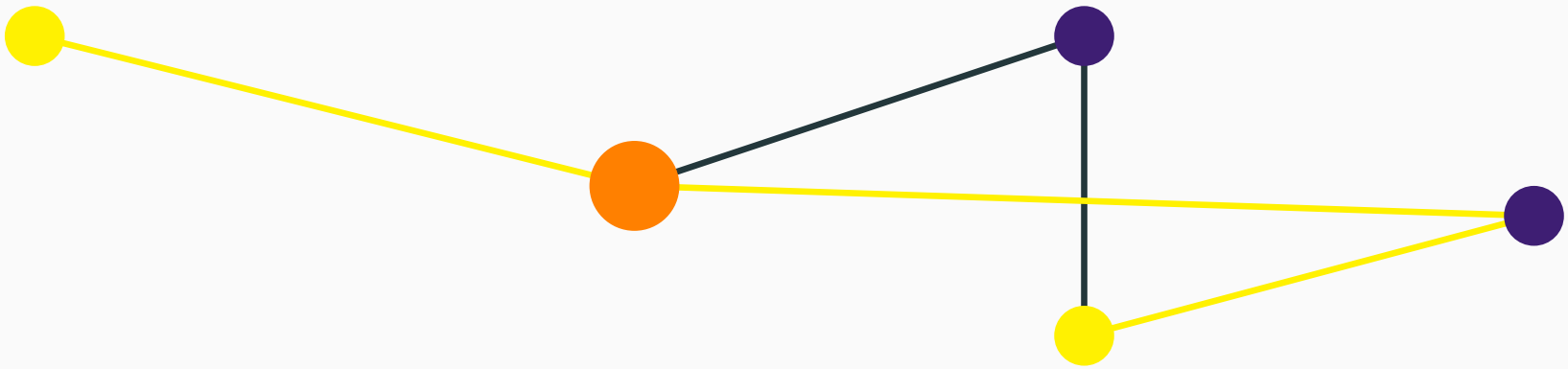
- degree (number of edges afferent to a given node)
- betweenness centrality measure



# Node importance

For every node, computation of:

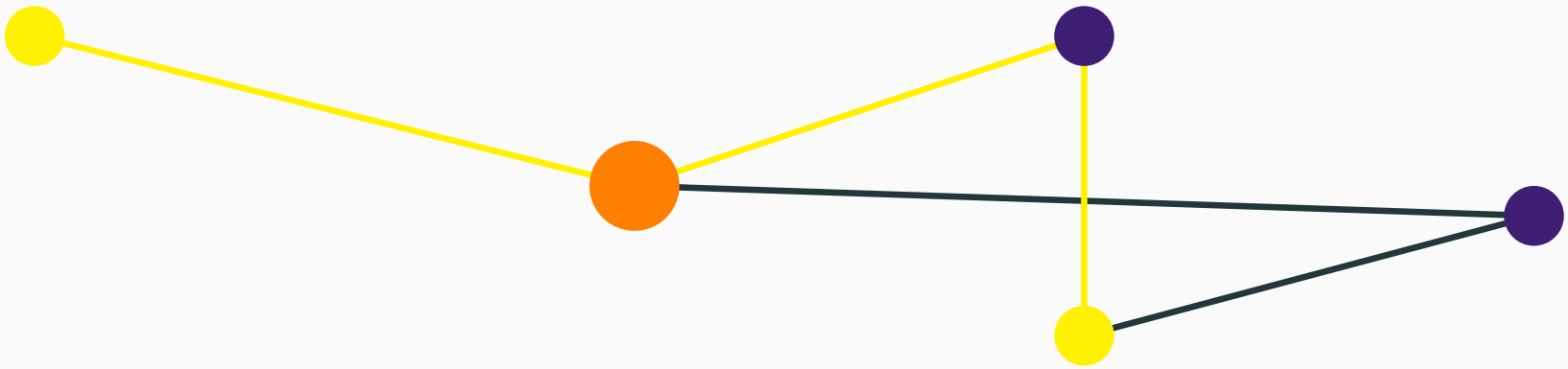
- degree (number of edges afferent to a given node)
- betweenness centrality measure



# Node importance

For every node, computation of:

- degree (number of edges afferent to a given node)
- betweenness centrality measure



# Find clusters by modularity optimization

The modularity Newman and Girvan (2004) of the partition $(\mathcal{C}_1, \ldots, \mathcal{C}_K)$ is equal to:

$$\mathcal{Q}(\mathcal{C}_1, \ldots, \mathcal{C}_K) = \frac{1}{2m} \sum_{k=1}^{K} \sum_{x_i, x_j \in \mathcal{C}_k} (W_{ij} - P_{ij})$$

with $P_{ij}$: weight of a "null model" (graph with the same degree distribution but no preferential attachment):

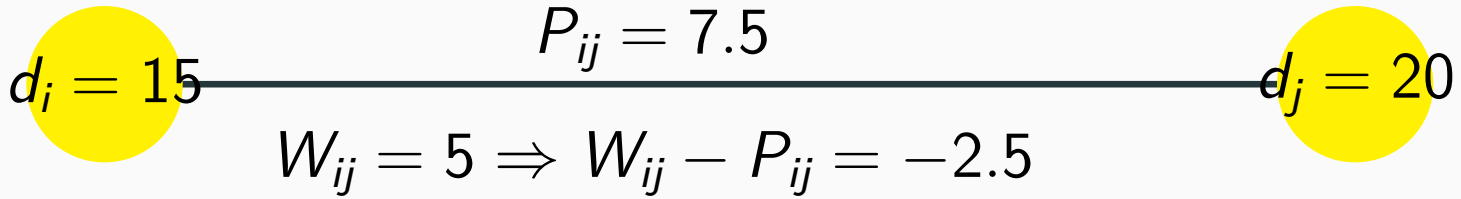
$$P_{ij} = \frac{d_i d_j}{2m}$$

with $d_i = \frac{1}{2} \sum_{j \neq i} W_{ij}$.

# Interpretation of the modularity

A good clustering should maximize the modularity:

- $\mathcal{Q} \nearrow$ when $(x_i, x_j)$ are in the same cluster and $W_{ij} \gg P_{ij}$
- $\mathcal{Q} \searrow$ when $(x_i, x_j)$ are in two different clusters and $W_{ij} \gg P_{ij}$ $(m = 20)$

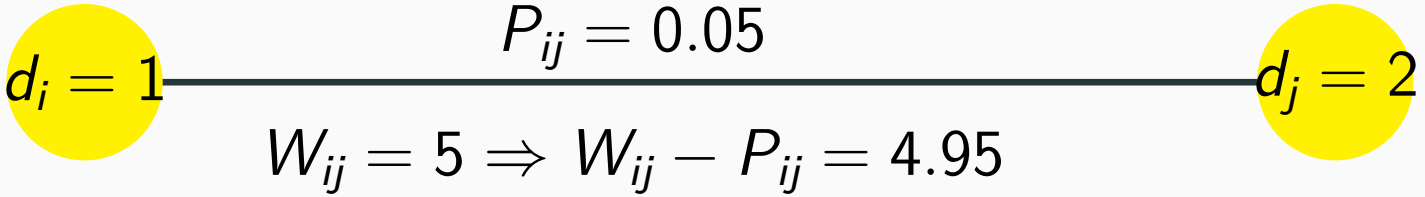



$i$ and $j$ in the same cluster decreases the modularity

# Interpretation of the modularity

A good clustering should maximize the modularity:

- $\mathcal{Q} \nearrow$ when $(x_i, x_j)$ are in the same cluster and $W_{ij} \gg P_{ij}$
- $\mathcal{Q} \searrow$ when $(x_i, x_j)$ are in two different clusters and $W_{ij} \gg P_{ij}$ $(m = 20)$




$i$ and $j$ in the same cluster increases the modularity

# Interpretation of the modularity

A good clustering should maximize the modularity:

- $\mathcal{Q} \nearrow$ when $(x_i, x_j)$ are in the same cluster and $W_{ij} \gg P_{ij}$
- $\mathcal{Q} \searrow$ when $(x_i, x_j)$ are in two different clusters and $W_{ij} \gg P_{ij}$
- Modularity
  - helps separate hubs
  - is not an increasing function of the number of clusters: useful to choose the relevant number of clusters

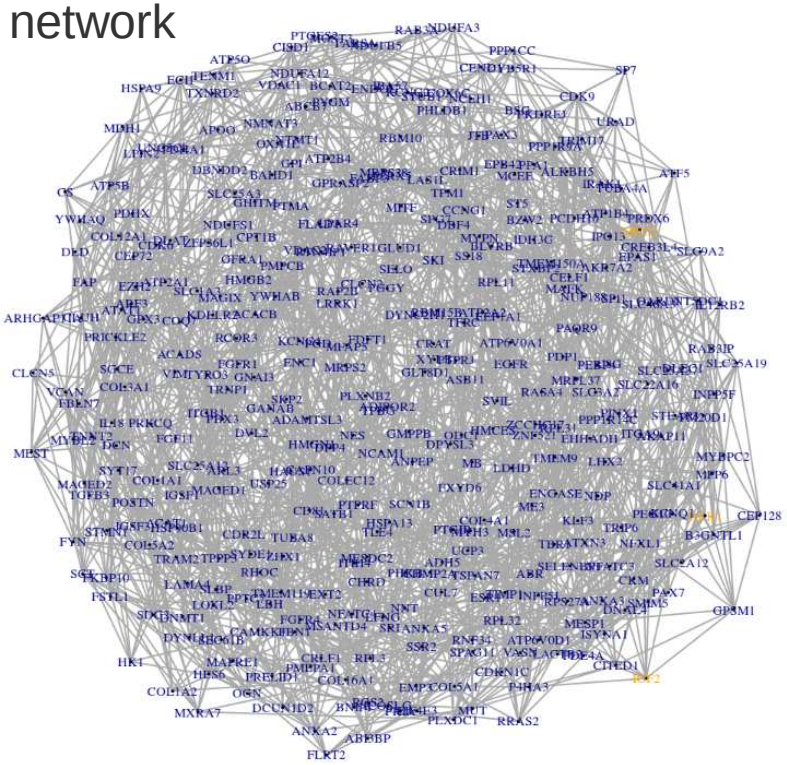
Approximate optimization with the Louvain algorithm Blondel et al. (2008) (among others)

# Network inference iteration and 3D FISH validations

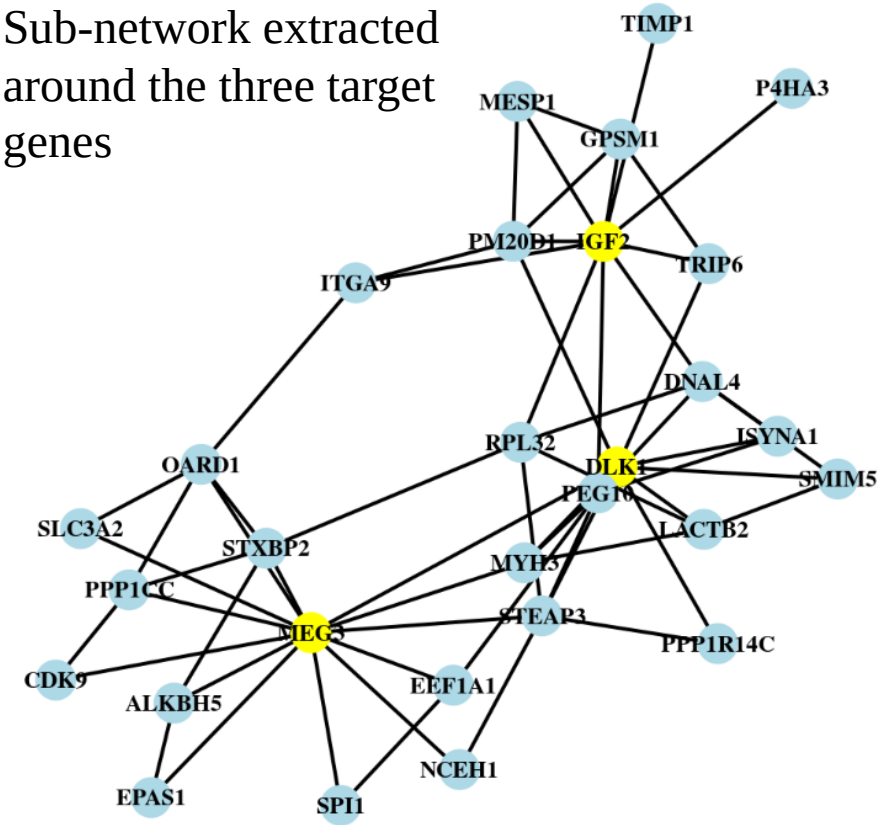
359 DEGs were selected for being highly correlated with *IGF2*, *DLK1* and *MEG3* ($R^2 \geq 0.84$)

➔ Network 0 to 3 with 359 nodes
➔ Network 0 without *a priori,* 2,279 edges (density: 3.55%)

Full network



Sub-network extracted around the three target genes

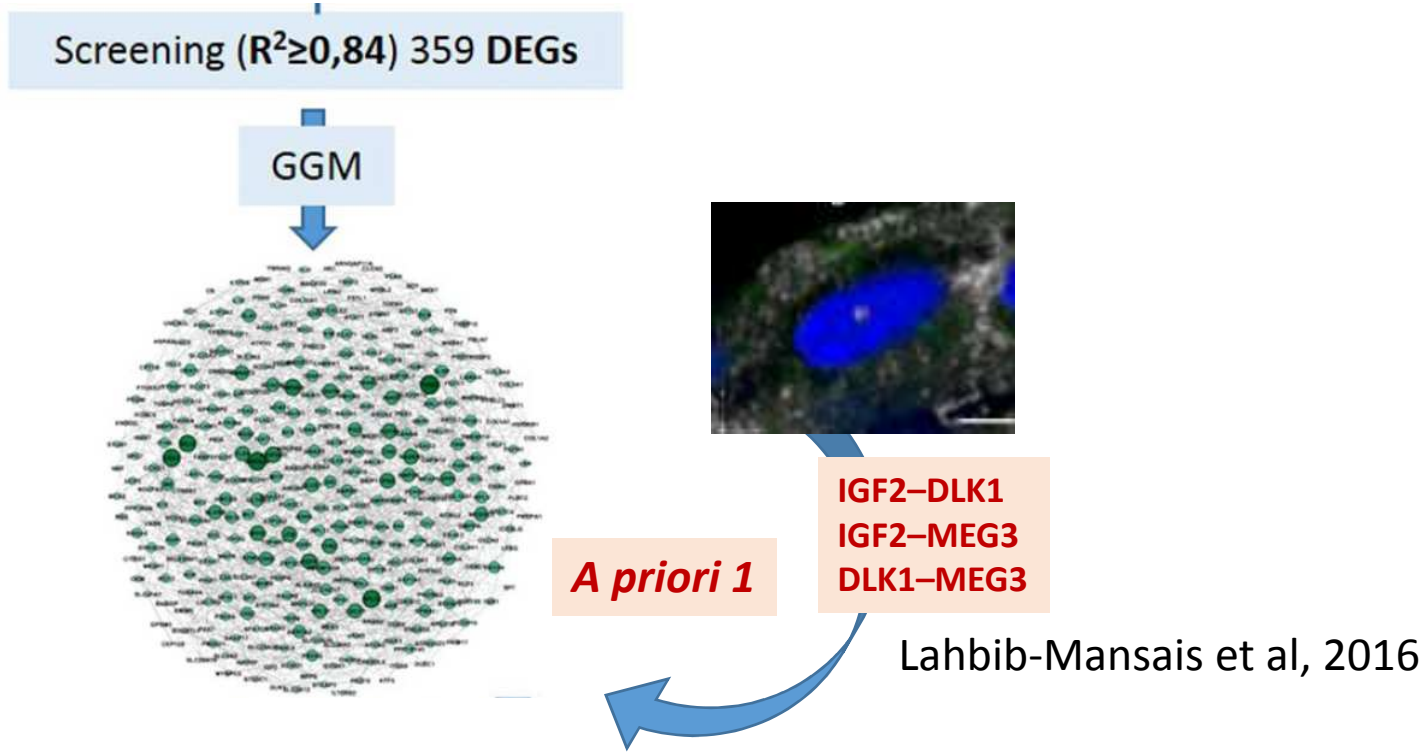



Marti-Marimon et al., 2018.

# Network inference iteration and 3D FISH validations

359 DEGs were selected for being highly correlated with *IGF2*, *DLK1* and *MEG3* ($R^2 \geq 0.84$)

➔ Network 0 to 3 with 359 nodes
➔ Network 0 without *a priori*, 2,279 edges (density: 3.55%)

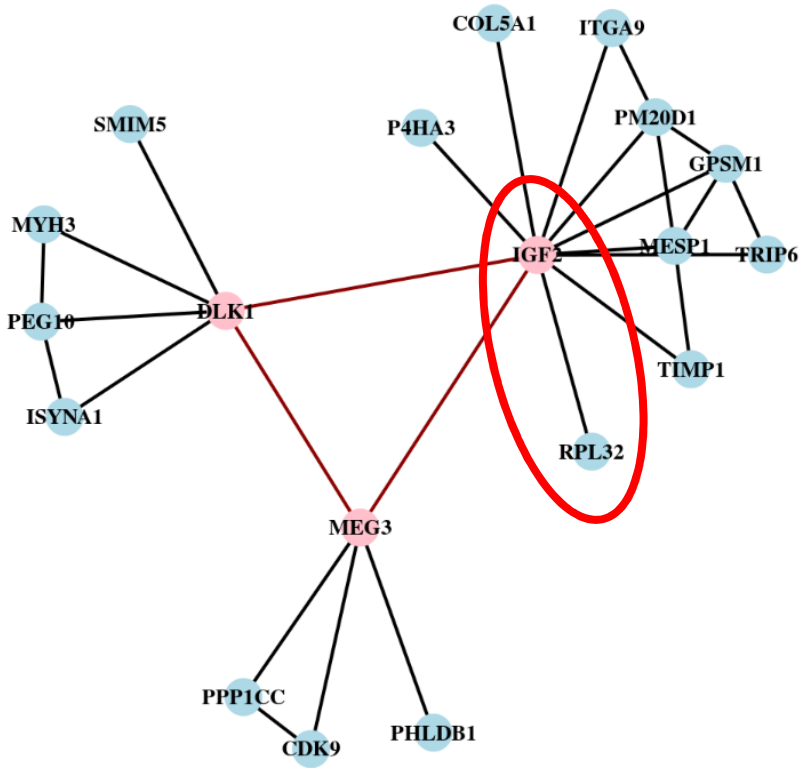



Marti-Marimon et al., 2018.

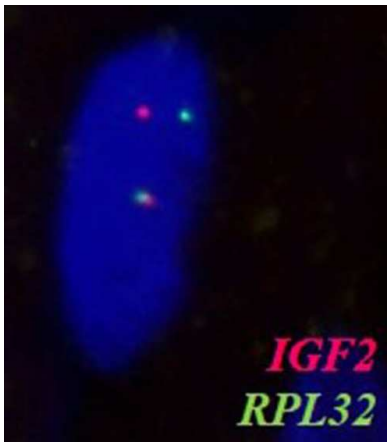
# Network inference iteration and 3D FISH validations

359 DEGs were selected for being highly correlated with *IGF2*, *DLK1* and *MEG3* ($R^2 \geq 0.84$)

➔ Network 0 to 3 with 359 nodes
➔ Network 0 without *a priori,* 2,279 edges (density 3.55%)
➔ Network 1 with triple co-localization of *IGF2*, *DLK1* and *MEG3,* 2,250 edges (density 3.50%)




➔ Test FISH 3D



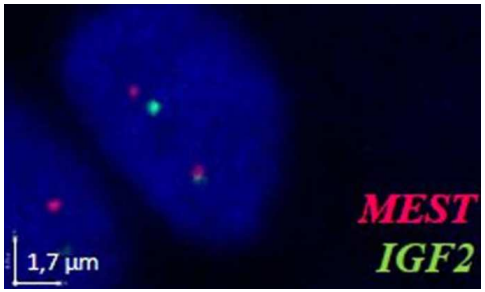

*IGF2* and *RPL32* were associated in 20% of the analysed nuclei (threshold 10%)

Marti-Marimon et al., 2018.

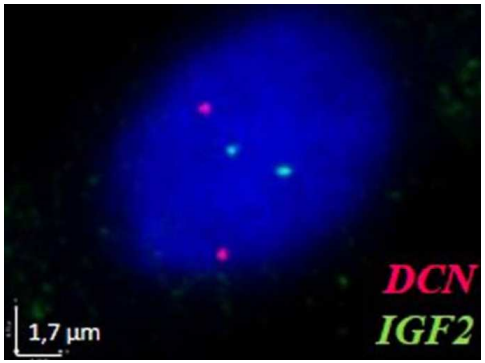
# Network inference iteration and 3D FISH validations

359 DEGs were selected for being highly correlated with *IGF2*, *DLK1* and *MEG3* ($R^2 \geq 0.84$)

➔ Network 0 to 3 with 359 nodes
➔ Network 0 without *a priori,* 2,279 edges and density 3.55%
➔ Network 1 with co-localization of *IGF2*, *DLK1* and *MEG3,* 2,250 edges and density 3.50%
➔ Network 2 with test of *MEST* and *DCN* associations, 2,091 edges and density 3.25%

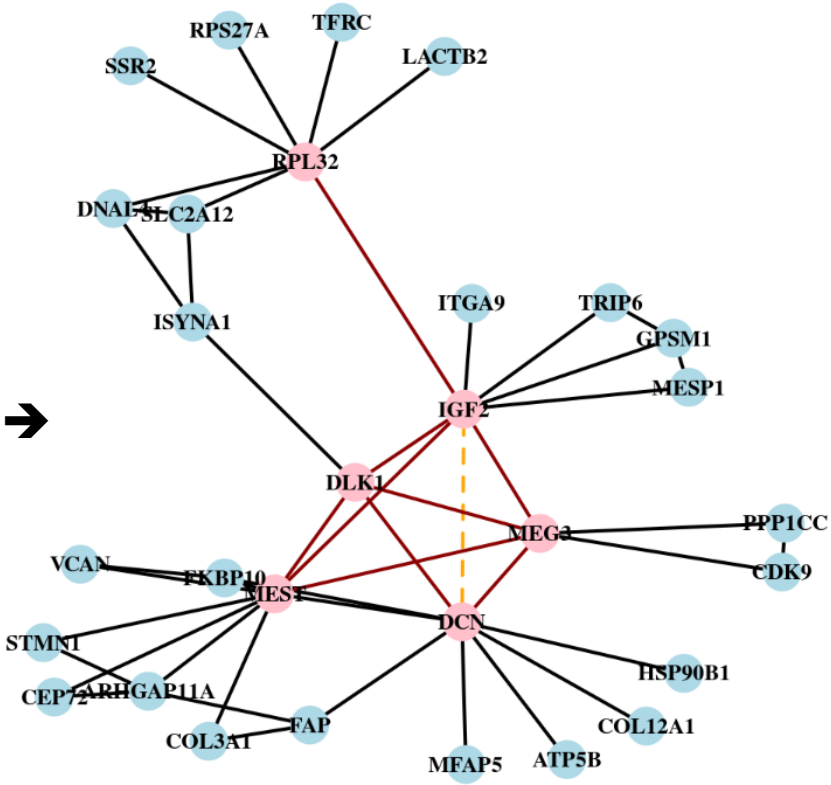



34% (+)




10% (-)

➔ Test FISH 3D
Network inference ➔



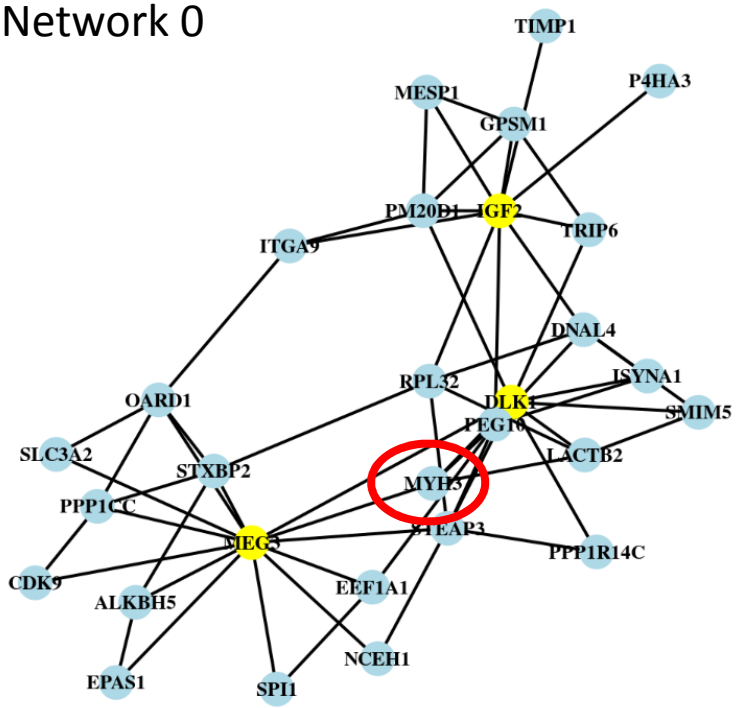

Marti-Marimon et al., 2018.

# Network inference iteration and 3D FISH validations

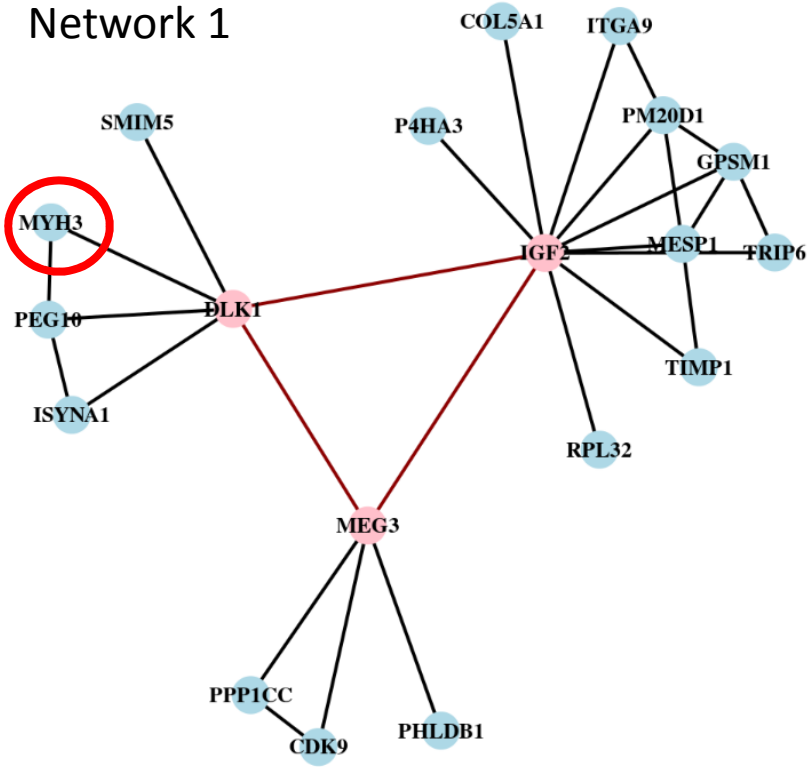
359 DEGs were selected for being highly correlated with *IGF2*, *DLK1* and *MEG3* ($R^2 \geq 0.84$)

➔ Network 3 with test co-localization with MYH3 (ntw 0 and 1)

Network 0




Network 1



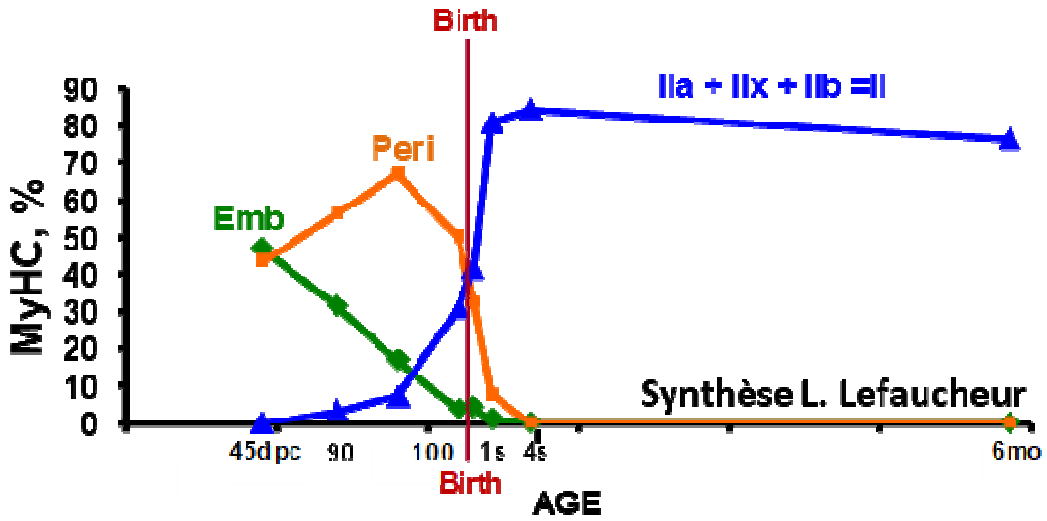

Marti-Marimon et al., 2018.

# Network inference iteration and 3D FISH validations

359 DEGs were selected for being highly correlated with *IGF2*, *DLK1* and *MEG3* ($R^2 \geq 0.84$)

➔ Network 3 with test co-localization with *MYH3* (ntw 0 and 1)

MYH3 = Embryonic myosin, excellent biomarker of muscle maturity (Voillet et al., 2018)
No functional link known with IGF2!

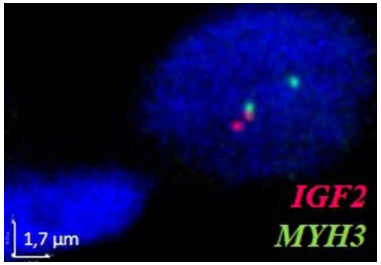



Marti-Marimon et al., 2018.

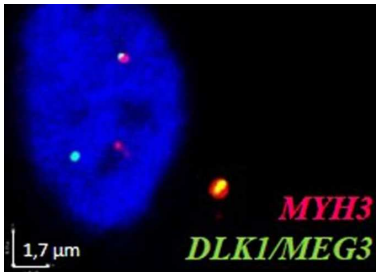
# Network inference iteration and 3D FISH validations

359 DEGs were selected for being highly correlated with *IGF2*, *DLK1* and *MEG3* ($R^2 \geq 0.84$)

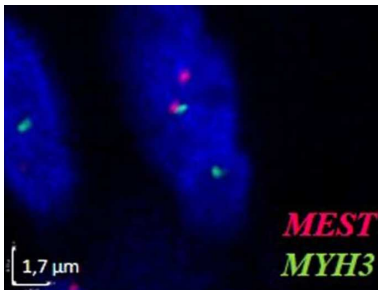
➔ Network 3 with test co-localization with *MYH3* (ntw 0 and 1), 2,091 edges and density 3.25%




52% (+)



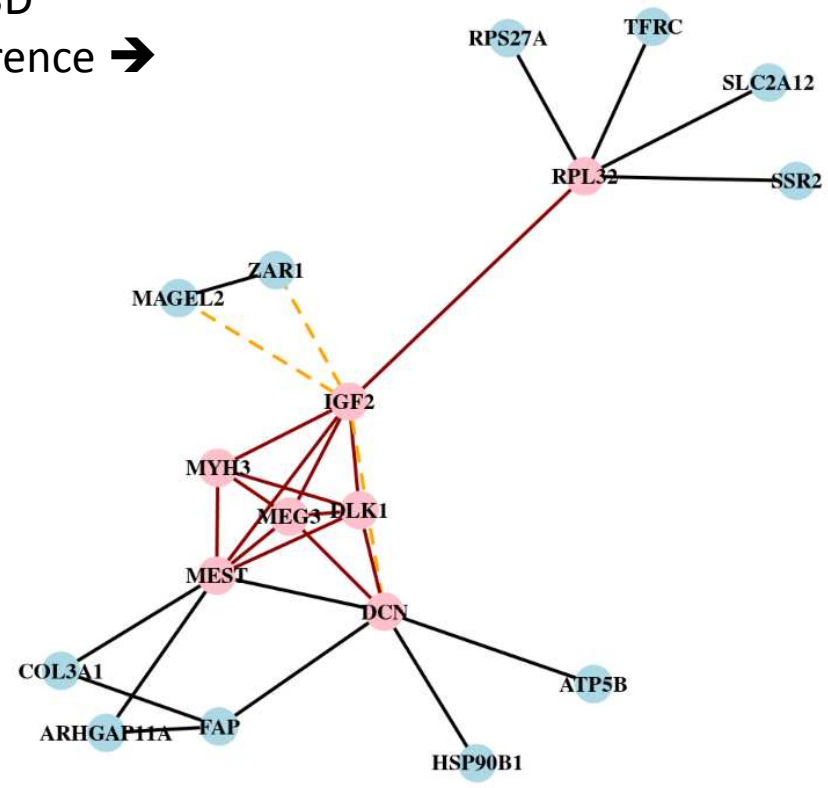

45% (+)




26% (+)

Marti-Marimon et al., 2018.

➔ Test FISH 3D
Network inference ➔



# Network mining (network structure with key genes)

The **degree** of a node is the number of edges afferent to this gene. High degree genes are connected to many other genes (hub).

The **betweenness** of the node  is the number of shortest paths between pairs of genes in the network that pass through that gene. High-betweenness genes are central and more likely to disconnect the network if removed.

# Network mining (network structure with key genes)

| | Network 0 | | Network 1 | | Network 2 | | Network 3 | | Comparison between Network 0 and Network 3 (% of variation) | |
|---|---|---|---|---|---|---|---|---|---|---|
| gene symbol | degree | betweeness | degree | betweeness | degree | betweeness | degree | betweeness | Degree | betweeness |
| *ADIPOR2* | 15 | 646,65 | 14 | 487,32 | 15 | 628,78 | 14 | 660,97 | -7 | 2 |
| *AKR7A2* | 19 | 492,63 | 17 | 436,71 | 15 | 474,10 | 14 | 291,90 | -26 | -41 |
| *CD81* | 17 | 551,17 | 18 | 616,7 | 15 | 478,76 | 17 | 600,58 | 0 | 9 |
| *CRAT* | 19 | 716,24 | 15 | 518,26 | 16 | 738,30 | 14 | 573,58 | -26 | -20 |
| *DCN* | 16 | 438,86 | 18 | 560,83 | 9 | 288,82 | 6 | 357,74 | -63 | -18 ↘ |
| *DLK1* | 10 | 103,52 | 6 | 81,7 | 5 | 74,22 | 5 | 24,13 | -50 | -77 ↘ |
| *DPP4* | 15 | 568,91 | 16 | 672,01 | 15 | 674,94 | 15 | 597,87 | 0 | 5 |
| *EGFR* | 16 | 624,92 | 12 | 375,87 | 12 | 385,35 | 11 | 354,78 | -31 | -43 |
| *GHITM* | 16 | 578,58 | 17 | 588,76 | 16 | 592,35 | 14 | 496,63 | -13 | -14 |
| *GLUD1* | 13 | 575,69 | 13 | 553,28 | 12 | 574,48 | 12 | 586,27 | -8 | 2 |
| *IGF2* | 10 | 118,26 | 11 | 231,09 | 8 | 260,58 | 7 | 622,44 | -30 | 426 ↗ |
| *LPAR4* | 14 | 464,31 | 17 | 644,76 | 18 | 812,81 | 16 | 798,82 | 14 | 72 |
| *MEG3* | 13 | 282,32 | 5 | 55,75 | 6 | 120,18 | 5 | 24,13 | -62 | -91 ↘ |
| *MESP1* | 12 | 228,49 | 14 | 320,34 | 14 | 483,27 | 14 | 775,31 | 17 | 239 |
| *MEST* | 13 | 148,2 | 12 | 121,44 | 10 | 345,69 | 7 | 385,27 | -46 | 160 ↘ |
| *MRPS28* | 16 | 743 | 15 | 743,29 | 16 | 953,42 | 15 | 796,14 | -6 | 7 |
| *MYH3* | 14 | 610,73 | 14 | 656,6 | 11 | 455,62 | 4 | 0,00 | -71 | -100 ↘ |
| *NMNAT3* | 17 | 562,63 | 18 | 664,84 | 16 | 473,55 | 17 | 573,15 | 0 | 2 |
| *RAVER1* | 16 | 613,84 | 16 | 665,73 | 16 | 696,35 | 16 | 745,66 | 0 | 21 |
| *RPL32* | 18 | 717,96 | 15 | 557,65 | 7 | 149,80 | 5 | 243,11 | -72 | -66 ↘ |
| *SELO* | 18 | 692,52 | 14 | 438,35 | 14 | 459,46 | 15 | 587,32 | -17 | -15 |
| *SYDE1* | 15 | 436,75 | 17 | 530,29 | 14 | 459,66 | 18 | 745,52 | 20 | 71 |
| *TFRC* | 15 | 595,1 | 15 | 534,83 | 13 | 437,38 | 17 | 846,81 | 13 | 42 |
| *TYRO3* | 20 | 785,95 | 18 | 659,9 | 16 | 603,94 | 17 | 700,03 | -15 | -11 |
| *YWHAB* | 20 | 670,22 | 17 | 470,35 | 17 | 538,41 | 17 | 547,17 | -15 | -18 |

Marti-Marimon et al., 2018.

# Network clustering

To analyse the evolution of the network structure from Network 0 to Network 3, clustering of the genes was performed on each network.

Normalized mutual information (NMI) measure the similarity between two clusterings. The value is comprised between 0 and 1 and is equal to 1 when the two clusterings are identical.

➔ clusterings become more consistent when introducing new biological information in each network inference iteration

| | Network 0 | Network 1 | Network 2 | Network 3 |
|---|---|---|---|---|
| Network 0 | 1 | **0.3893** | 0.3381 | 0.3244 |
| Network 1 | 0.3893 | 1 | **0.4007** | 0.3923 |
| Network 2 | 0.3381 | 0.4007 | 1 | **0.4152** |
| Network 3 | 0.3244 | 0.3923 | 0.4152 | 1 |

Marti-Marimon et al., 2018.

# Network clustering: Networks 0 and 3 were analysed in depth to search for any correspondence between clusters

Pairwise contingency tables between clusterings. Percentage of genes for each cluster in Network 0 found in each cluster of Network 3. In bold and red, the most resembling values between clusters.

| | | Clusters in Network 3 | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 |
| Clusters in Network 0 | 1 | **64,10** | 7,69 | 7,69 | 2,56 | 7,69 | 10,26 |
| | 2 | 8,77 | **68,42** | 0,00 | 1,75 | 19,30 | 1,75 |
| | 3 | 14,89 | 0,00 | **65,96** | 19,15 | 0,00 | 0,00 |
| | 4 | 3,92 | 1,96 | 11,76 | **82,35** | 0,00 | 0,00 |
| | 5 | 34,09 | 6,82 | 4,55 | 11,36 | 43,18 | 0,00 |
| | 6 | 3,57 | 17,86 | 14,29 | 32,14 | 0,00 | 32,14 |
| | 7 | 11,11 | 38,89 | 0,00 | 11,11 | 38,89 | 0,00 |
| | 8 | 0,00 | 48,72 | 33,33 | 10,26 | 5,13 | 2,56 |
| | 9 | 5,56 | 11,11 | 16,67 | 27,78 | 38,89 | 0,00 |

Marti-Marimon et al., 2018.

# Functional enrichment analysis: Gene Ontology Biological Process

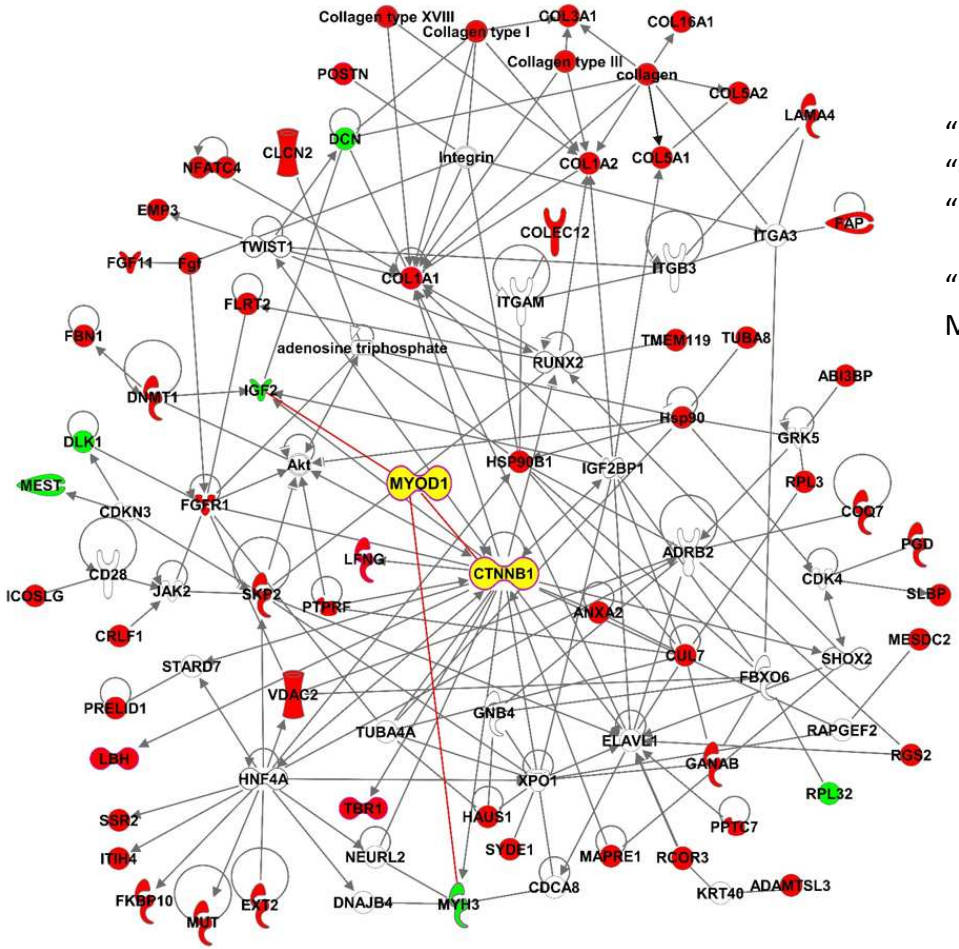
Functional enrichment analysis based on Gene Ontology was performed using the web tool Webgestalt (WEB-based GEne SeT AnaLysis Toolkit)

| | Network 0 - Cluster 1 | Network 3 - Cluster 1 | |
|---|---|---|---|
| **GOBP Terms** | **FDR** | **FDR** | **Target** |
| Extracellular structure | 5,76E-05 | **1,14E-08** | **DCN** |
| Cellular response to organonitrogen compound | **6,80E-04** | 1,16E-02 | **IGF2** |
| Reponse to transformaing growth factor beta | **2,35E-03** | 1,24E-01 | |
| Multicellular organism metabolic process | 2,35E-03 | 3,05E-03 | |
| Skin development | **3,18E-03** | 1,44E-01 | |
| Neuron migration | **2,82E-02** | 4,37E-01 | |
| Regulation of neuron projection development | **3,07E-02** | 4,93E-01 | |
| Mesoderm development | | 1,24E-01 | **MEST** |
| Muscle organ development | | 8,35E-01 | **MYH3** |
| Notch signaling pathway | | 5,56E-01 | **DLK1** |
| Collagen fibril organization | 1,10E-04 | 1,02E-05 | |
| | **Network 0 - Cluster 8** | **Network 3 - Cluster 2** | |
| **GOBP Terms** | **FDR** | **FDR** | |
| Generation of precursor metabolites and energy | 1,64E-02 | **1,32E-07** | |
| Oxidation-reduction process | 7,25E-03 | **5,63E-09** | |
| Energy derivation by oxidation of organic compounds | 8,17E-03 | **1,88E-06** | |
| Cellular respiration | 8,17E-03 | **2,65E-07** | |

Marti-Marimon et al., 2018.

# Functional enrichment analysis: reconstructed network of genes in cluster 1 of Network 3 with Ingenuity Pathway Analysis (IPA)

IPA proposed to connecting 49 (82%) out of 60 genes in a network.
*MYOD1* and *CTNNB1* were identified by upstream regulator analysis as potential transcriptional factors for a group of genes including *IGF2* and *MYH3.*




"Cell Morphology", 14 genes, *p*-value = 1.75e-08
"Quantity of cells", 31 genes, *p*-value = 2.48e-09
"Morphology of connective tissue cells", 8 genes, *p*-value = 1.27e-04

"Formation of muscle", 10 genes, *p*-value = 2.98e-05, involved IGF2 and MYH3 together with CTNNB1 and MYOD1.

Marti-Marimon et al., 2018.
Villa-Vialaneix et al., 2013

# Conclusions

- 82% of edges in Network 0 were conserved in Network 3
- The most important genes in Network 0 were among those showing the highest values of betweenness and degree in Network 3.

➔ Not major disturbances in the network structure

- In the local analysis, the NMI value revealed that the clusters resembled one another more with each new network inferred.
- Four out of six clusters in the final network conserved more than 62% of genes in the corresponding clusters of Network 0.

- IGF2-MEST, (DLK1/MEG3)-MEST, (DLK1/MEG3)-DCN, that were observed to be connected in co-expression networks in other studies.
- *DLK1*, *MEG3*, *RPL32*, *MEST*, *DCN* and *MYH3* were less connected with the rest of the other genes in Network 3 but not *IGF2.*

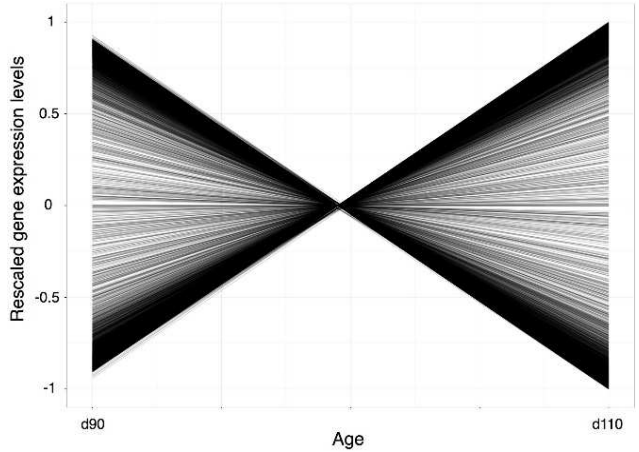
- No previous association between IGF2 and MYH3, even though the two genes are known to be involved in muscle development ➔ *overexpression and accumulation of β-catenin in the nuclei of differentiating murine myoblasts results in higher MyoD activation and Myhc induction* (Ramazzotti et al, 2016)

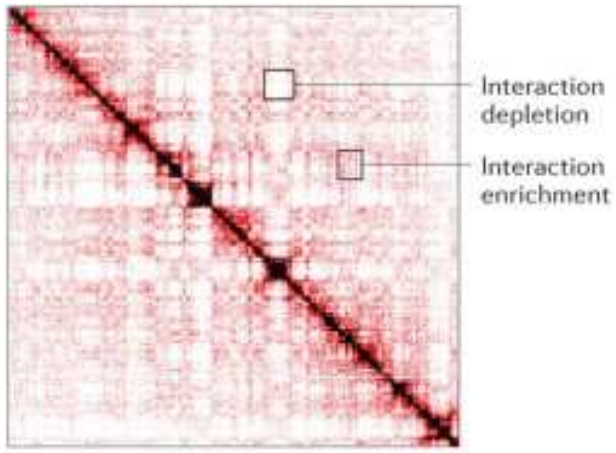
Marti-Marimon et al., 2018.

# Conclusions - Perspectives

- What is published and what is not…
- Intermediate modelling is retained as valuable information on robust or non-robust interactions ➔ currently, new interactions are being tested by FISH 3D

- Dramatic change in gene expression at the end of gestation ➔ Search of interaction whole genome (Maria Marti-Marimon thesis)




**Whole genome interaction Maps**
3D Chromosome conformation capture
**Hi-C** in progress

# Thank you for your attention!

Blondel, V., Guillaume, J., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communites in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, P10008:1742–5468.

Butte, A. and Kohane, I. (1999). Unsupervised knowledge discovery in medical databases using relevance networks. In *Proceedings of the AMIA Symposium*, pages 711–715.

Butte, A. and Kohane, I. (2000). Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. In *Proceedings of the Pacific Symposium on Biocomputing*, pages 418–429.

Meinshausen, N. and Bühlmann, P. (2006). High dimensional graphs and variable selection with the Lasso. *Annals of Statistic*, 34(3):1436–1462.

Newman, M. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review, E*, 69:026113.

Voillet, V., SanCristobal, M., Lippi, Y., Martin, P. G., Iannuccelli, N., Lascor, C., Vignoles, F., Billon, Y., Canario, L., and Liaubet, L. (2014). Muscle transcriptomic investigation of late fetal deelopment identifies candidate genes for piglet maturity. *BMC Genomics*, 15:797.