Inferring Boolean Networks from Single-Cell Human Embryo Datasets

Mathieu Bolteau[1]
Jérémie Bourdon[1] Laurent David[2] Carito Guziolowski[1]

[1]Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France

[2]Nantes Université, CHU Nantes, INSERM, Center for Research in Transplantation and Translational Immunology, UMR 1064, F-44000 Nantes, France

ISBRA 2023

Wednesday, October 11th 2023

Motivations

Need to **better understand preimplantation development**

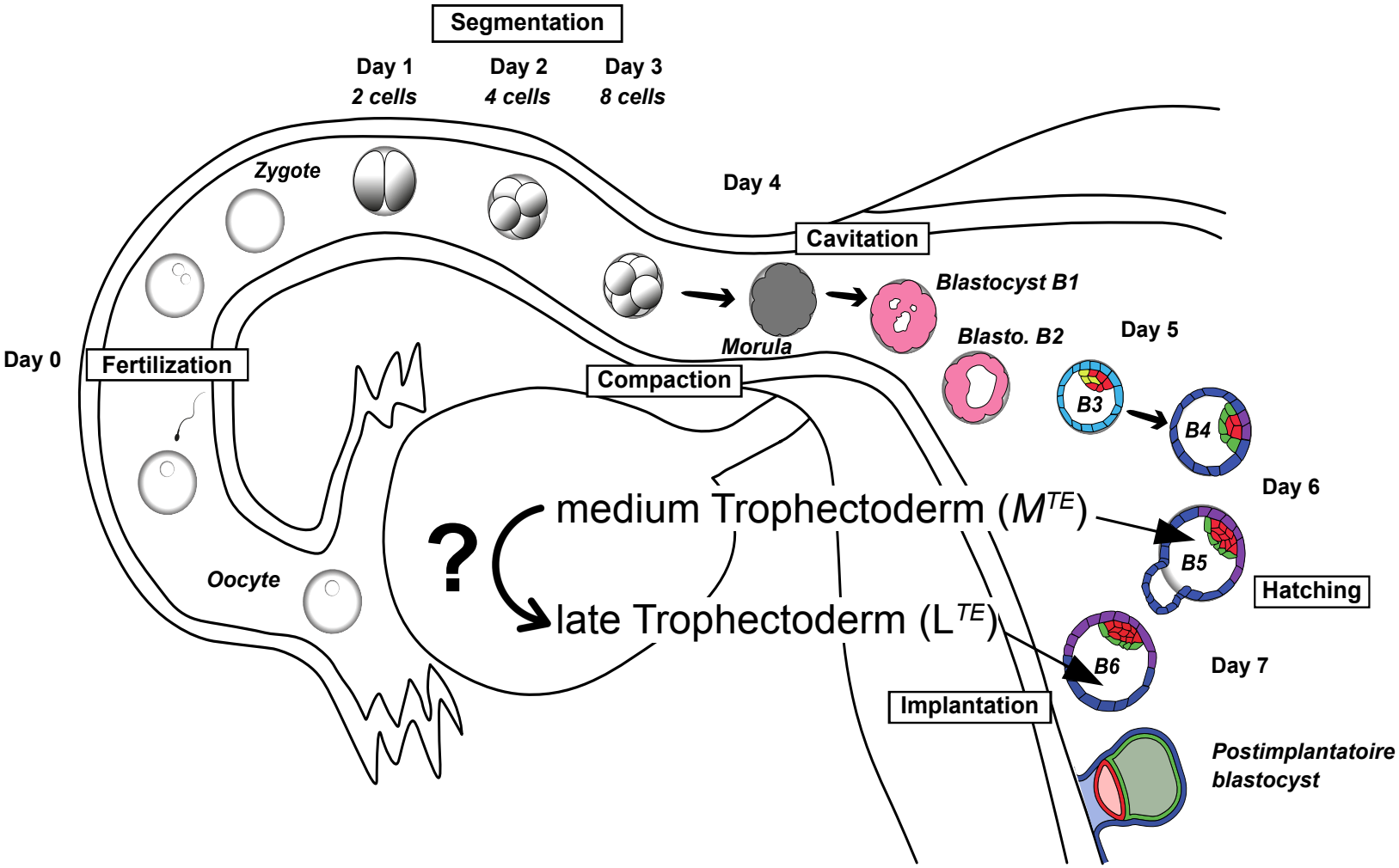
Motivations

Need to **better understand preimplantation development**

Research on human embryos is **limited** (experiments, law, ethics)

In silico predictive model of human embryonic development

Human embryonic development



Adapted from Meistermann, *Ph.D. thesis*, 2020

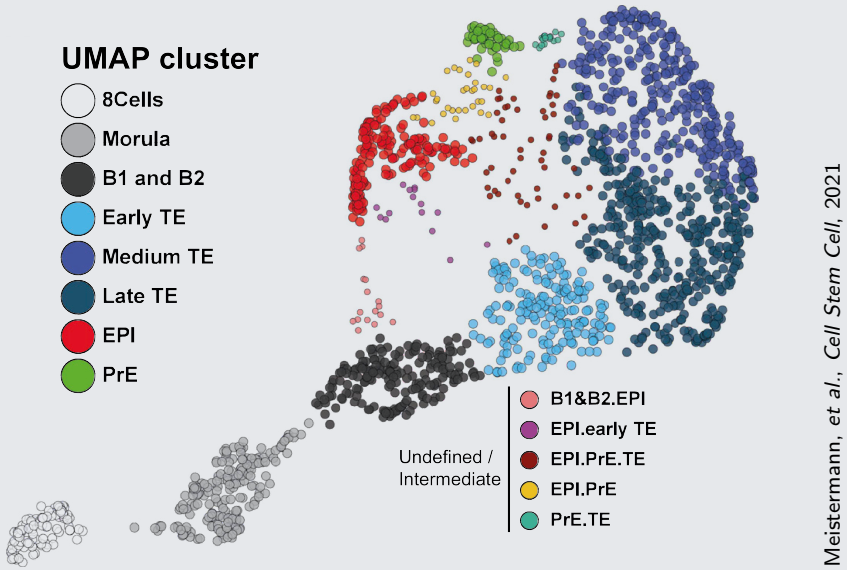
Background [Meistermann, *et al.*, *Cell Stem Cell*, 2021]

scRNAseq data from multiple stage embryos

Expression of $\sim 20,000$ genes in $\sim 1,700$ cells from 128 multi-stage embryos

Previous results (in house)

- **Clustering** of cells
- Identification of **gene modules** $\rightarrow$ 438 transcription factors (TFs)
- **Pseudotime** evolution of cells at different developmental stages



Goal: Boolean models of embryonic developmental stages

Challenges

- Single cell data specifities: sparsity and redundancy
- High dimensional data: $\sim 20,000$ genes for $\sim 1,700$ cells
- Unavailable perturbations

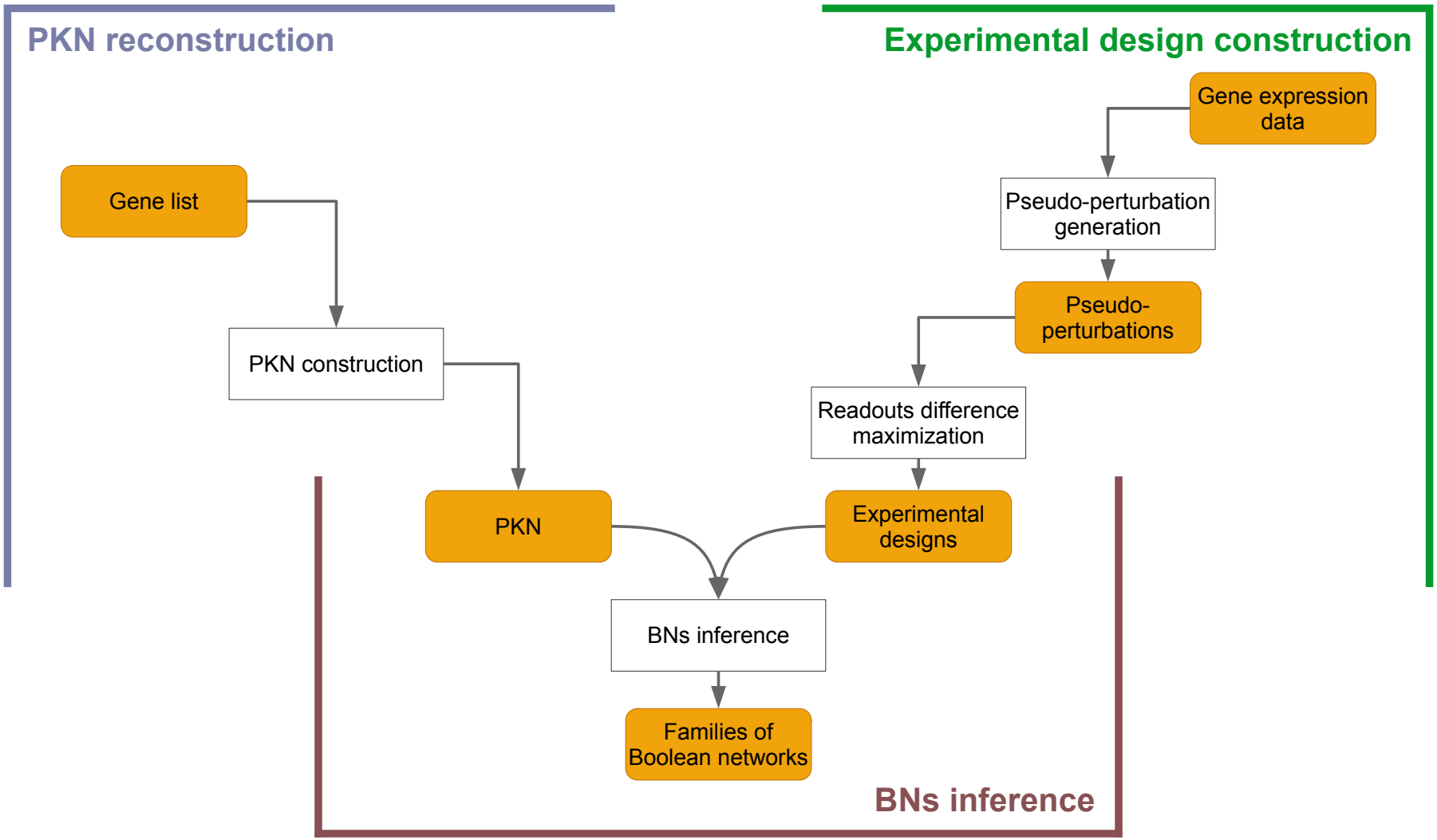
Goal: Boolean models of embryonic developmental stages

Challenges

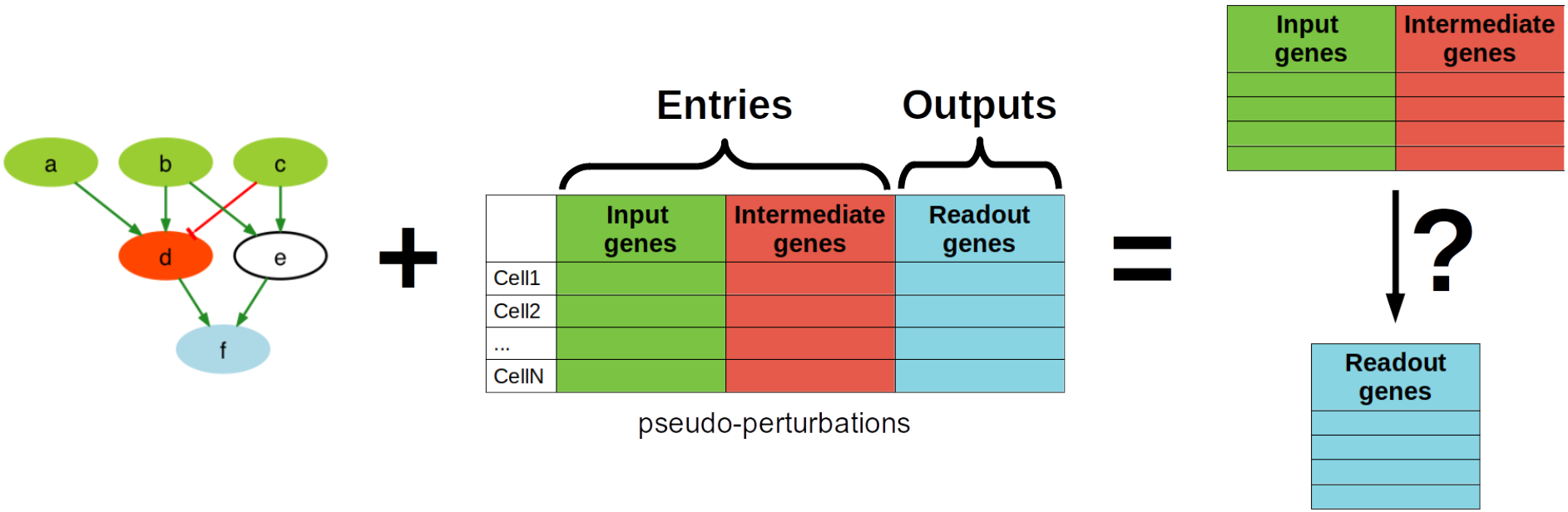
- Single cell data specifities: sparsity and redundancy
- High dimensional data: $\sim 20,000$ genes for $\sim 1,700$ cells
- Unavailable perturbations

Proposed solution

- Distinguish between **two developmental stages**
- Build **families of network models** for each stage
- Identify **regulatory mechanisms** that differentiate both models and representing multiple cells
- Application on medium (M^{TE}) and late (L^{TE}) trophectoderm stages


PKN reconstruction
Experimental design construction
Gene expression data
Gene list
Pseudo-perturbation generation
Pseudo-perturbations
PKN construction
Readouts difference maximization
PKN
Experimental designs
BNs inference
Families of Boolean networks
BNs inference

Learning predictive models



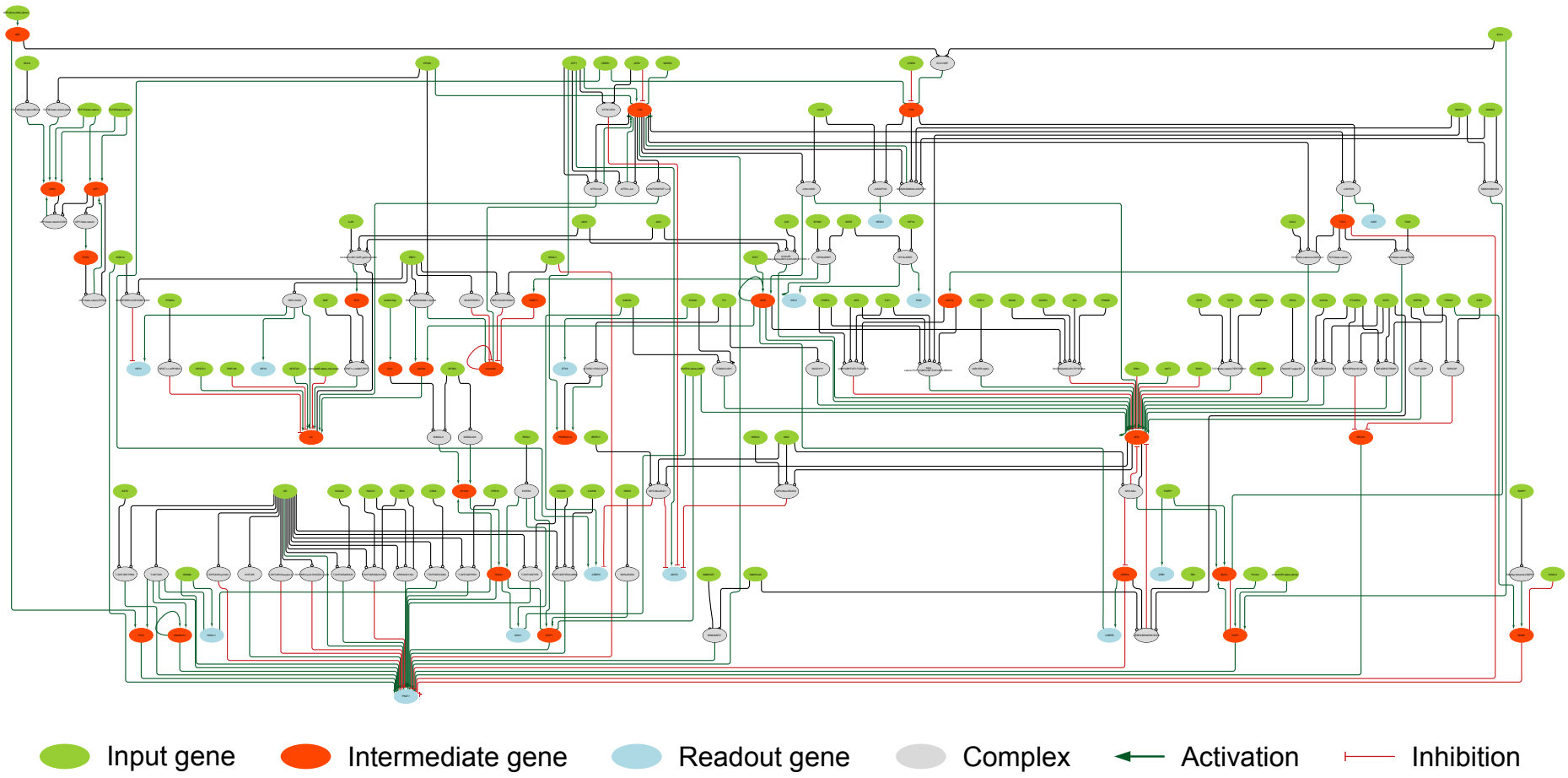
Prior-Knowledge Network

Experimental design

Predictive model

- Signed and directed causal interactions among genes
- Gene expression for a developmental stage

- From 438 TFs
- 233 nodes : inputs (85), intermediates (36), readouts (19)
- 369 edges



Method

Idea

Extract pseudo-perturbation experiments from scRNAseq data given the PKN structure (Step 1)

Data preprocessing

- Binarization of input + intermediate genes. Basic approach: gene is expressed (1) if at least 2 reads are present; else it is absent (0).
- Normalization of readout genes.

Pseudo-perturbation generation

Inputs + intermediates | Readouts

Cell	A	B	C	D	E	F	G	H	Class
1	1	1	0	1	0	0.8	0.4	0.6	A
2	1	0	1	0	0	0.2	0.5	0.3	A
3	1	1	0	1	1	0.8	0.3	0.9	A
4	1	1	1	0	1	0.6	0.1	0.2	B
5	1	0	0	1	1	0.7	0.8	0.5	B
6	0	0	1	0	1	0.7	0.2	0.3	B

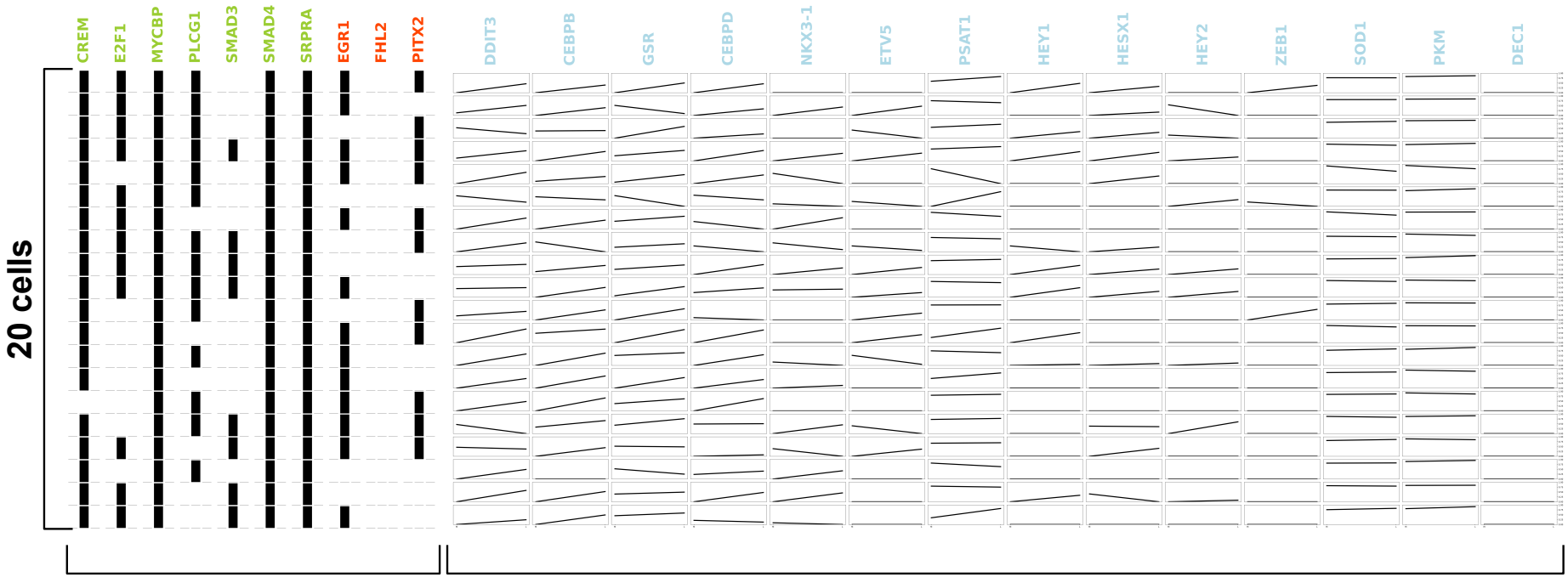
Logic program in ASP (k=3)

Cell	A	C	D	F	G	H	Class
1	1	0	1	0.8	0.4	0.6	A
2	1	1	0	0.2	0.5	0.3	A

Cell	A	C	D	F	G	H	Class
5	1	0	1	0.7	0.8	0.5	B
4	1	1	0	0.6	0.1	0.2	B

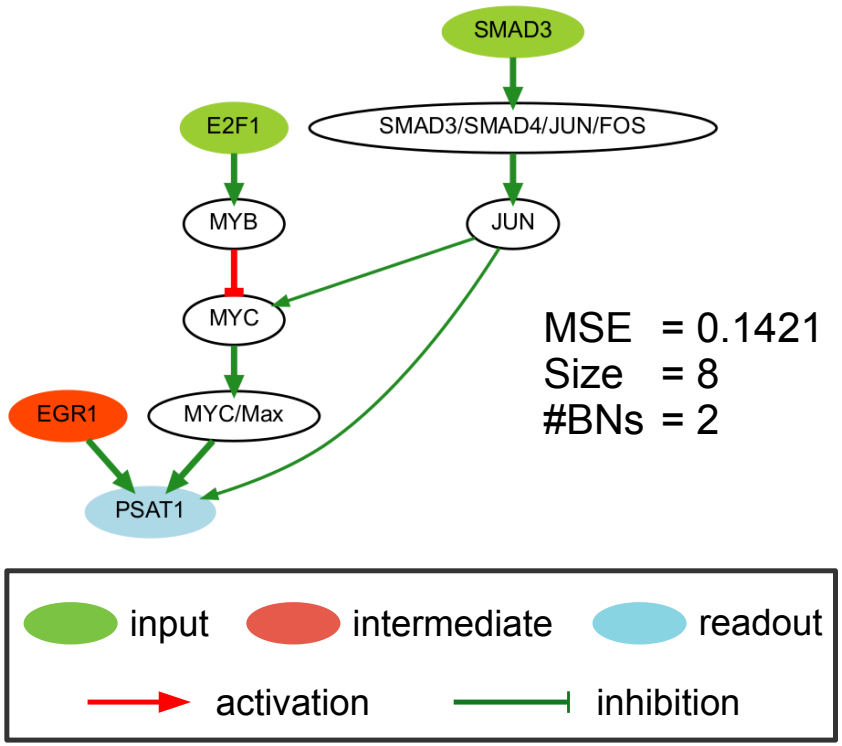
- 3 selected genes: A, C, D ($k = 3$)
- Matching cells: (1,5), (2,4) ← pseudo-perturbations
- Different guaranteed pseudo-perturbation vectors
- Optimal number of matching cells: 2

Reconstructed exprimental design

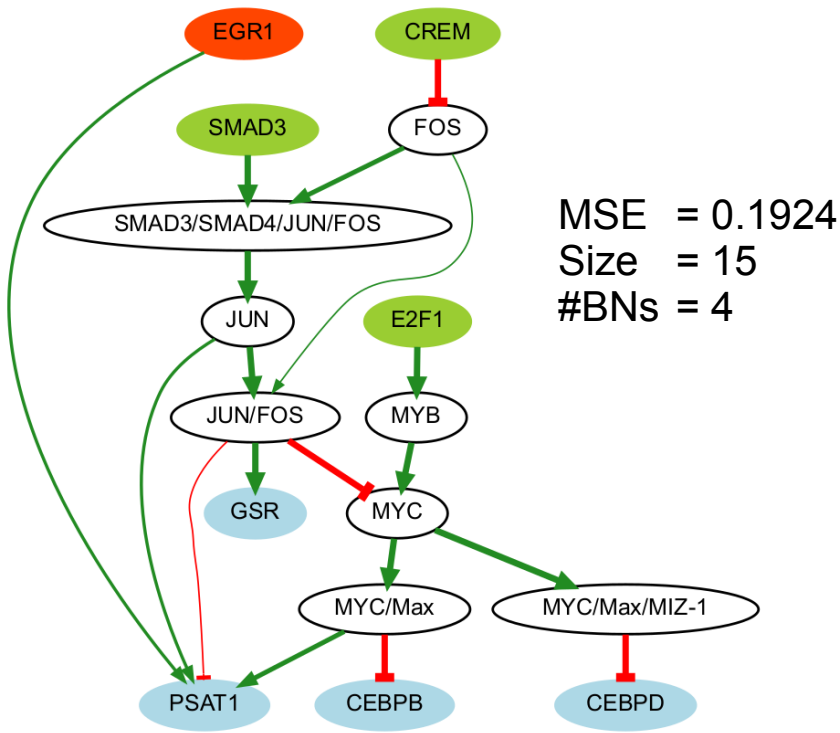


Inferred BN families using Caspo [Guziolowski *et al.*, *Bioinformatics*, 2013]

Medium TE



Late TE



- Greater BNs variability for L^{TE} $\rightarrow$ Gain of function
- L^{TE} seems more unstable (number of BNs) $\rightarrow$ transition from L^{TE} to another stage

Conclusion

Pseudo-perturbation generation

- Algorithm to select cells and genes to generate pseudo-perturbations
  $\rightarrow$ 20 pseudo-perturbations in 65h
- Expression of 10 genes across 20 cells are representative of the cell populations (*e.g.* 75% in M^{TE} and 89% in L^{TE})
- Our method deals with single cell data and its specificities (redundancy and sparsity)

General method

- A method that learns Boolean networks of 2 stages using scRNAseq data and PKN
- Mechanisms of TF-gene regulations distinguishing 2 developmental stages
- Complementarity with the state of the art
  - Boolean models without using perturbations
  - Method taking into account the diverse states of cell population

Aknowledgements

- Jérémie Bourdon @LS2N, Nantes University
- Carito Guziolowski @LS2N, Centrale Nantes
- Laurent David @CR2TI, Nantes University Hospital, Nantes University
- ANR AIBY4 & ANR BOOSTIVF