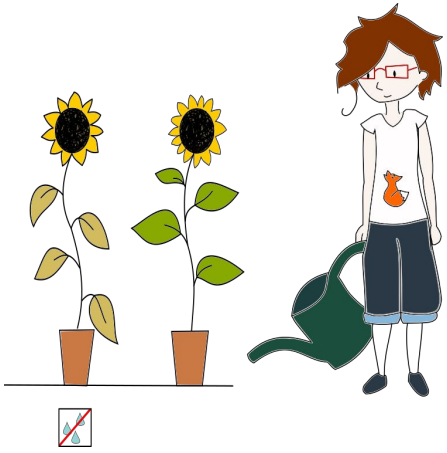
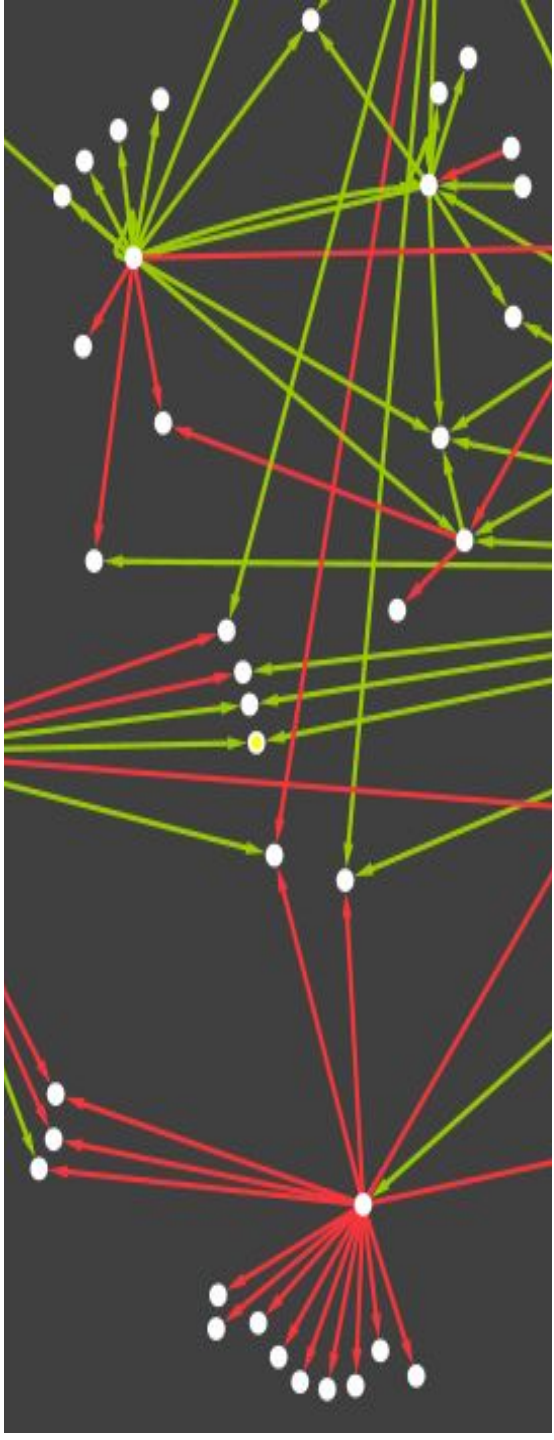
SUNRISE
UNE CULTURE POUR LE FUTUR

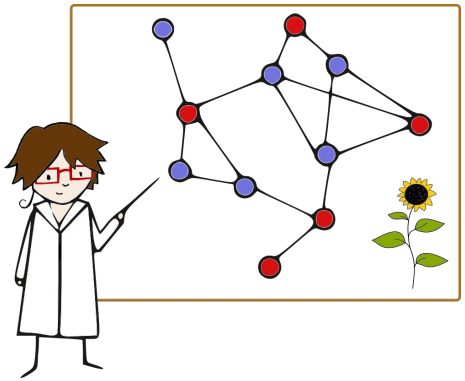
Choisir une méthode d'inférence adaptée à l'étude d'un processus biologique via la simulation de données

Lise Pomiès





Objectif



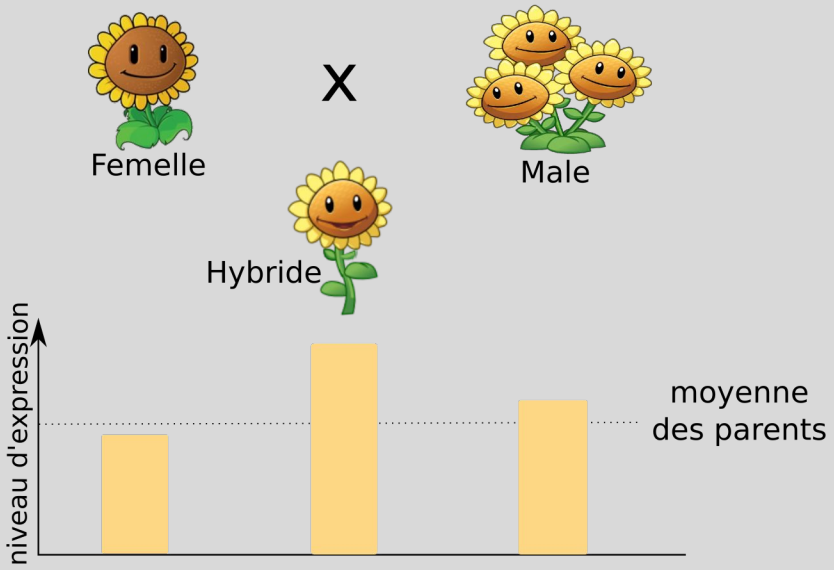
Inférer le **réseau de régulation de gènes** permettant de comprendre la réponse du **tournesol** à la **sécheresse** en interaction avec l'**hétérosis**

Hétérosis

phénotype hybride

≠

moyenne du phénotype des parents



2 expériences

13HP02

Environnement semi contrôlé

4 femelles
4 mâles
16 hybrides

Sélectionner *un pool de* ***gènes*** *impliqués dans la réponse à la sécheresse et dans l'hétérosis*

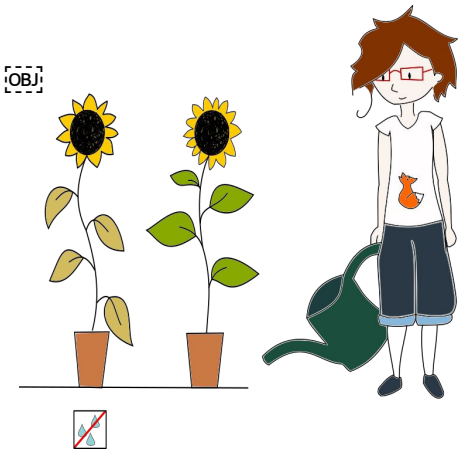
15EX05

Expérience en champs

435 hybrides
(issus de 72 parents)

Avoir suffisamment de ***données pour inférer*** *le réseau de régulation de gènes*

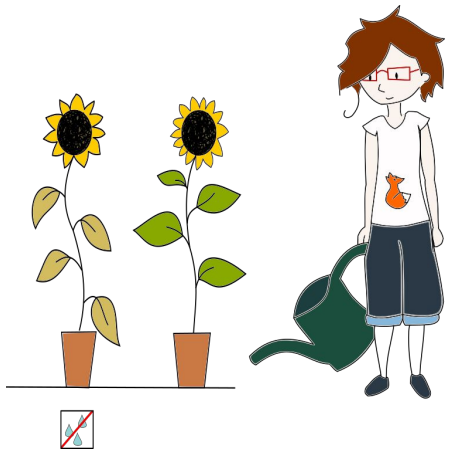
13HP02 : sélection des gènes



Génotypes	Traitements
4 femelles 4 mâles 16 hybrides	stress hydrique contrôle

→ Expression des gènes mesurés par **RNAseq** (**58 050 gènes**)

13HP02 : sélection des gènes

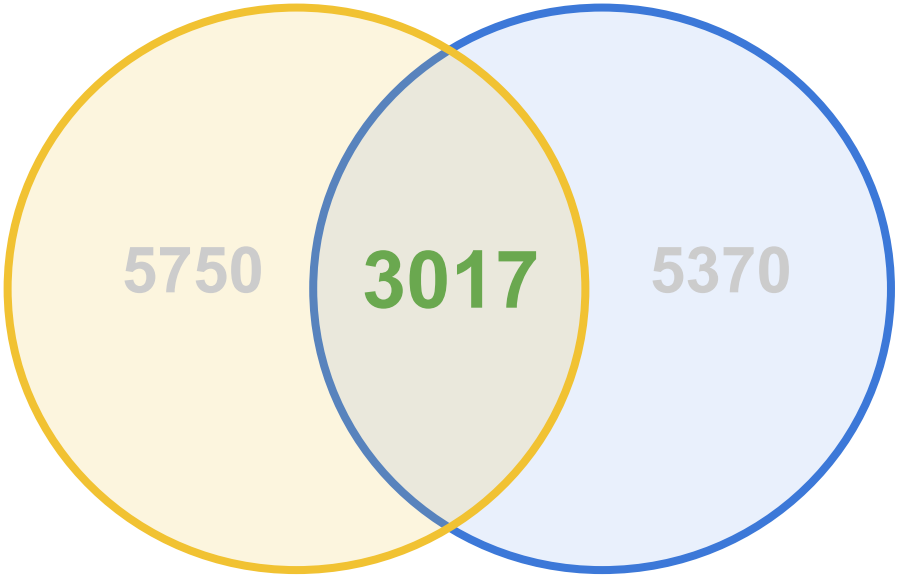


Génotypes	Traitements
4 femelles 4 mâles 16 hybrides	stress hydrique contrôle

→ Expression des gènes mesurés par **RNAseq** (**58 050 gènes**)

1. Identification des facteurs de transcription du tournesol

iTAK
détection de domaines protéiques spécifiques des facteurs de transcription



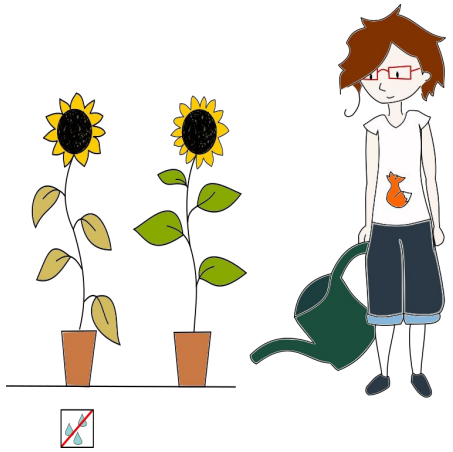
PlantTFCat
comparaison avec des facteurs de transcription connus chez A. thaliana

13HP02 : sélection des gènes

2. *Sélection d'un pool de facteurs de transcription*

Annotés avec des termes GO liés à la sécheresse

Différentiellement exprimés pour sécheresse x hétérosis ou sécheresse



13HP02 : sélection des gènes

2. Sélection d'un pool de facteurs de transcription

Annotés avec des termes GO liés à la sécheresse

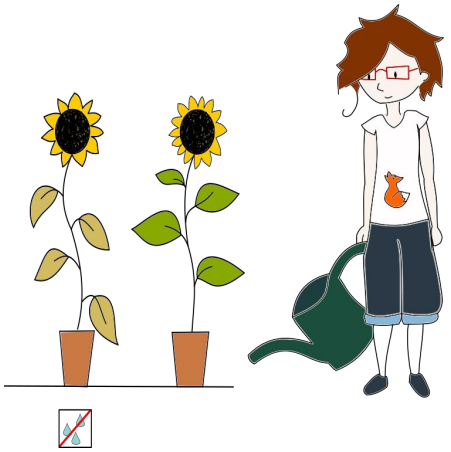
Différentiellement exprimés pour sécheresse x hétérosis ou sécheresse

3. Sélection de gènes cibles

Non facteur de transcription

Les plus différentiellement exprimés sécheresse x heterosis

10 gènes cibles sélectionnés



13HP02 : sélection des gènes

2. *Sélection d'un pool de facteurs de transcription*

Annotés avec des termes GO liés à la sécheresse

Différentiellement exprimés pour sécheresse x hétérosis ou sécheresse

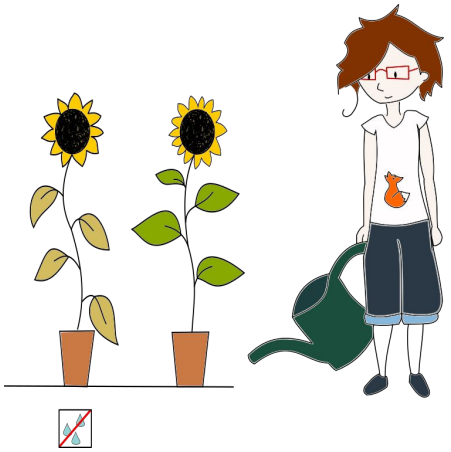
3. *Sélection de gènes cibles*

Non facteur de transcription

Les plus différentiellement exprimés sécheresse x hétérosis

10 gènes cibles sélectionnés

liste de 180 gènes d'intérêts
(FT majoritairement)



15EX05 : collecte des données

15EX05 Expérience en champs

mesure de l'expression **qPCR (fluidigm)**	**180 gènes** (variables) **435 hybrides** (individus)	**ratio favorable** à l'inférence de réseau de régulation de gènes
Mesure **SNP (Illumina Hi-Seq)**	**72 parents** (hybrides déduits)	

15EX05 : collecte des données

15EX05 Expérience en champs

mesure de l'expression **qPCR (fluidigm)**	**180 gènes** (variables) **435 hybrides** (individus)	**ratio favorable** à l'inférence de réseau de régulation de gènes
Mesure **SNP (Illumina Hi-Seq)**	**72 parents** (hybrides déduits)	

/!\ hybrides → individus non-indépendants /!
Méthodes actuelles pour individus indépendants

15EX05 : collecte des données

15EX05 Expérience en champs

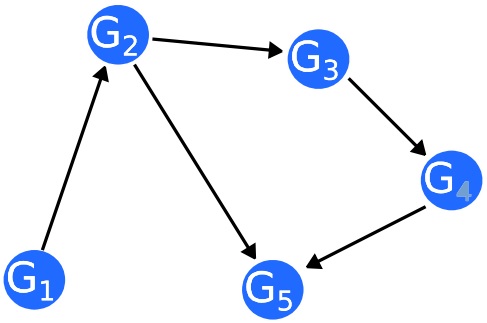
mesure de l'expression **qPCR (fluidigm)**	**180 gènes** (variables) **435 hybrides** (individus)	**ratio favorable** à l'inférence de réseau de régulation de gènes
Mesure **SNP (Illumina Hi-Seq)**	**72 parents** (hybrides déduits)	

/!\ hybrides → individus non-indépendants /!
Méthodes actuelles pour individus indépendants

Tester différentes méthodes d'inférence de réseau

Tester méthodes d'inférence

vrai réseau biologique



Condition d'expression

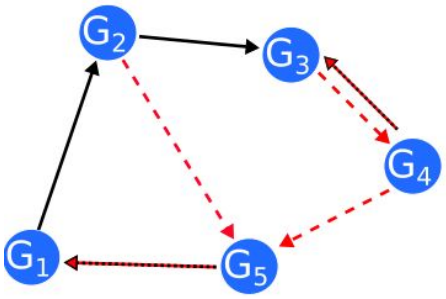
Gène du réseau

	c1	c2	c3	c4	c5	c6	c7
g1							
g2							
g3							
g4							
g5							

Niveaux d'expression associés aux gènes du réseau

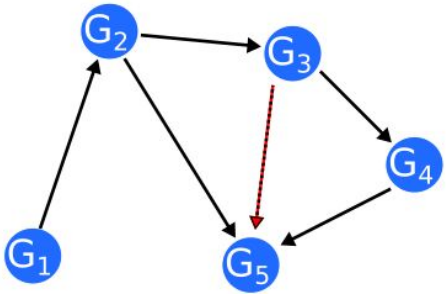
utilisation de différentes méthodes de reconstruction de réseau

Méthode 1



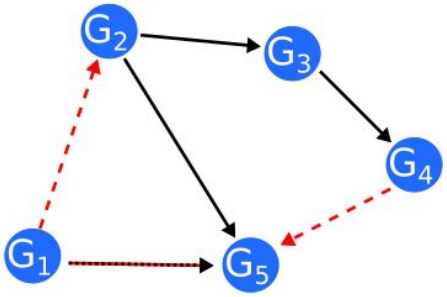
2 arcs faux
3 arcs manquants

Méthode 2



1 arc faux

Méthode 3



1 arc faux
2 arcs manquants

→ *La méthode n°2 donne les meilleurs résultats*

Tester méthodes d'inférence

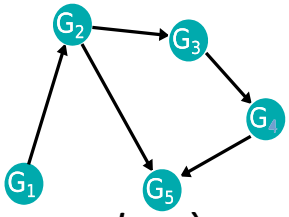
Problème : On ne **connaît pas** le **vrai réseau** (même partiellement)

→ impossible de comparer les méthodes d'inférence

Tester méthodes d'inférence

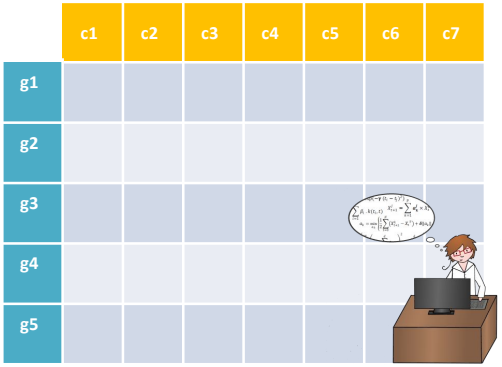
Problème : On ne **connaît pas** le **vrai réseau** (même partiellement)
→ impossible de comparer les méthodes d'inférence

Tester les méthodes sur un **jeu de données artificiel** mais avec des **caractéristiques proches**



Réseau de gènes artificiel

Simulation d'expérience

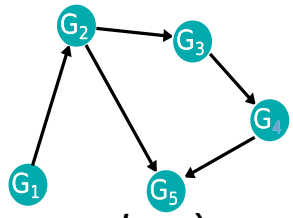


Données simulées

Tester méthodes d'inférence

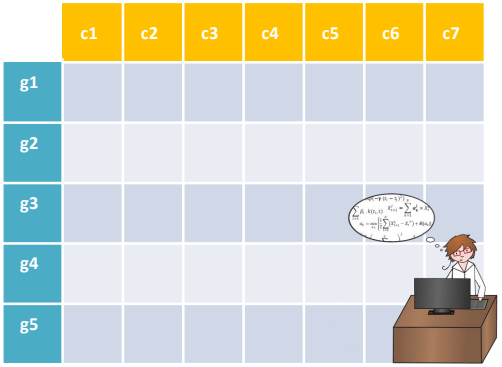
Problème : On ne **connaît pas** le **vrai réseau** (même partiellement)
→ impossible de comparer les méthodes d'inférence

Tester les méthodes sur un **jeu de données artificiel** mais avec des **caractéristiques proches**



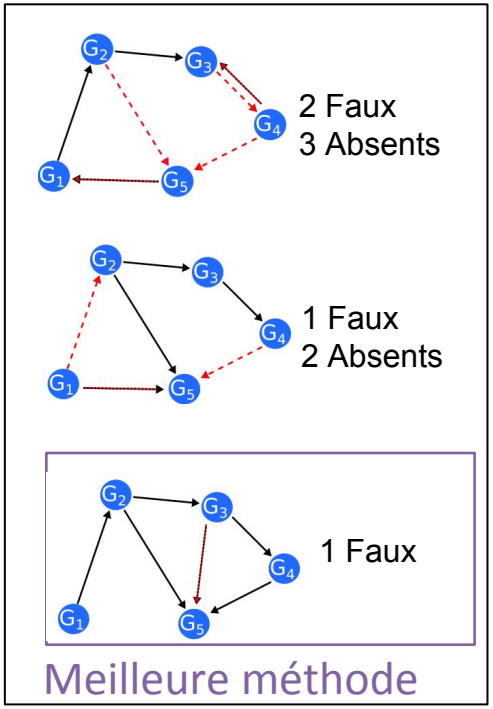
Réseau de gènes artificiel

Simulation d'expérience



Données simulées

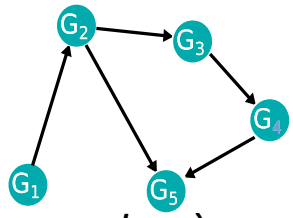
Test méthode d'inférence



Tester méthodes d'inférence

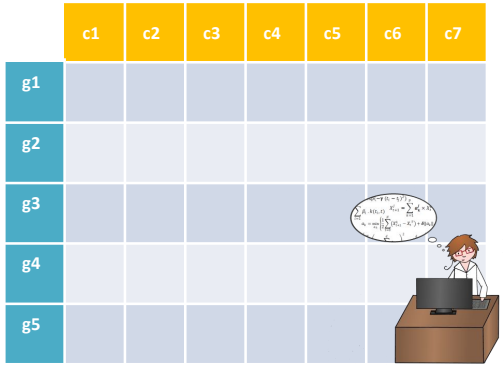
Problème : On ne **connaît pas** le **vrai réseau** (même partiellement)
→ impossible de comparer les méthodes d'inférence

Tester les méthodes sur un **jeu de données artificiel** mais avec des **caractéristiques proches**



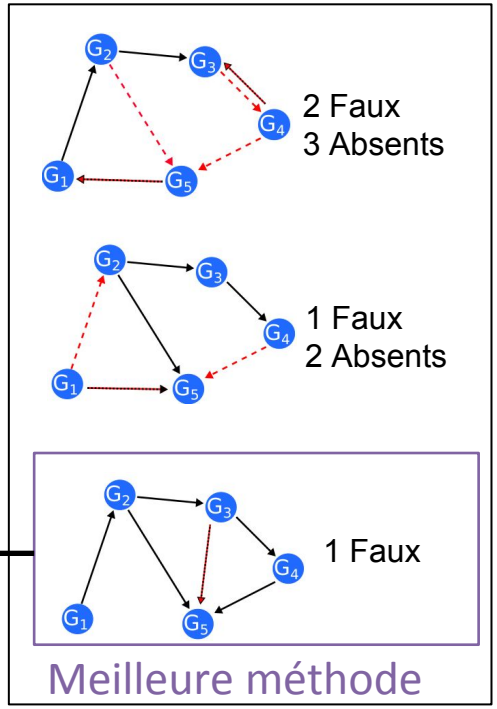
Réseau de gènes artificiel

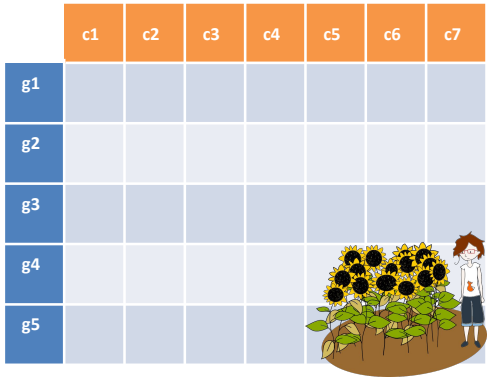
Simulation d'expérience



Données simulées

Test méthode d'inférence





Jeu de données 15EX05

Inférence
du réseau de gènes
sécheresse x hétérosis

Jeu de données artificiel

Réseau de gènes artificiel

Créer un **réseau artificiel** le plus proche possible de notre réalité biologique → Utiliser les données disponibles pour une **espèce proche** → *A. thaliana* plante modèle Nombreuses interactions entre les gènes mises en évidence

Plusieurs bases de données regroupent ces interactions :

AtPID	protéine - protéine FT → gène	*experimentation* *text mining (bibliographie)*
AtTFIN-1	FT - FT	*experimentaion*
AtRegNet	FT → gène	*experimentaion*
PlantRegMap	FT → gène	*prediction* *experimentaion*

Jeu de données artificiel

Réseau de gènes artificiel

Créer un **réseau artificiel** le plus proche possible de notre réalité biologique → Utiliser les données disponibles pour une **espèce proche** → *A. thaliana* plante modèle Nombreuses interactions entre les gènes mises en évidence

Plusieurs bases de données regroupent ces interactions :

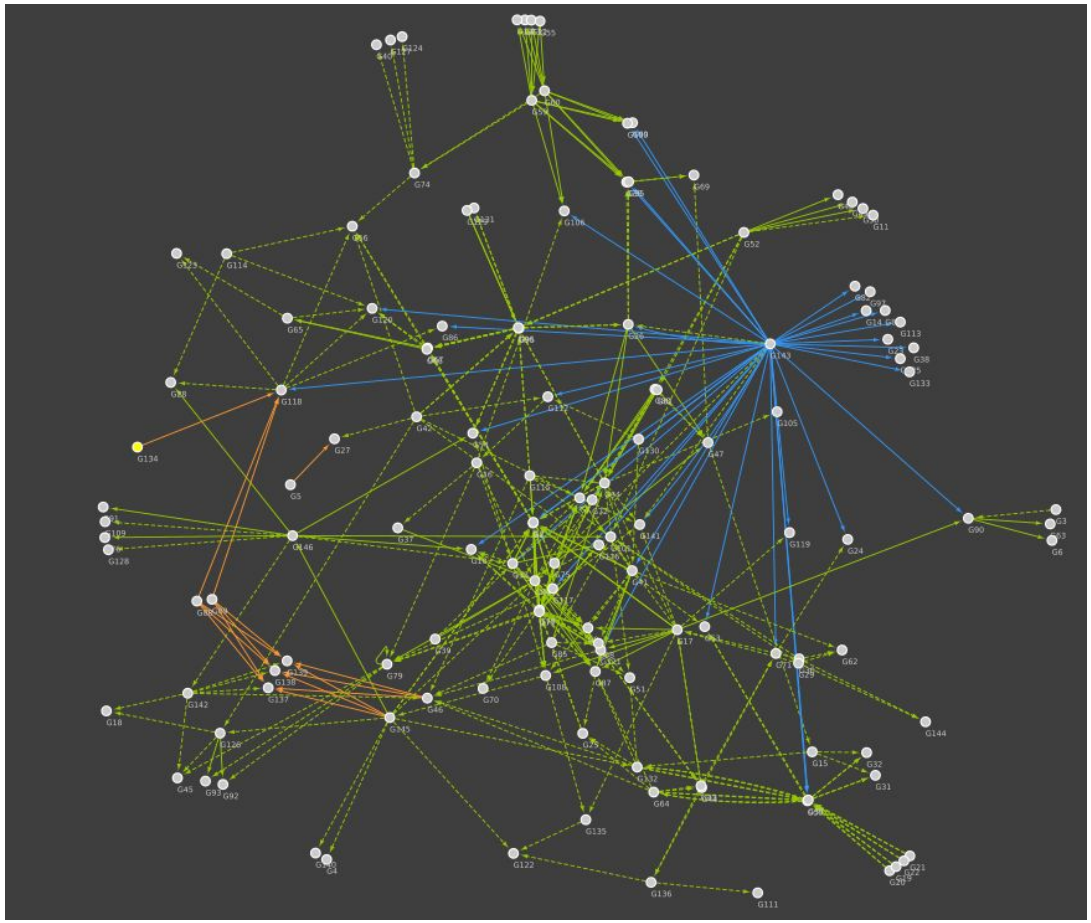
AtPID	protéine - protéine **FT → gène**	*experimentation* *text mining (bibliographie)*
AtTFIN-1	FT - FT	*experimentaion*
AtRegNet	**FT → gène**	*experimentaion*
PlantRegMap	**FT → gène**	*prediction* *experimentaion*

→ Sélection des relations de type régulation (prédiction + expérimentation)

Jeu de données artificiel

Réseau de gènes artificiel

Pour les homologues *A. thaliana* des gènes tournesol sélectionnés, relations de régulations décrites dans les bases de données (expérimentation + prédiction)



AtRegNet (16)

AtPID (36)

PlantRegMap (312)

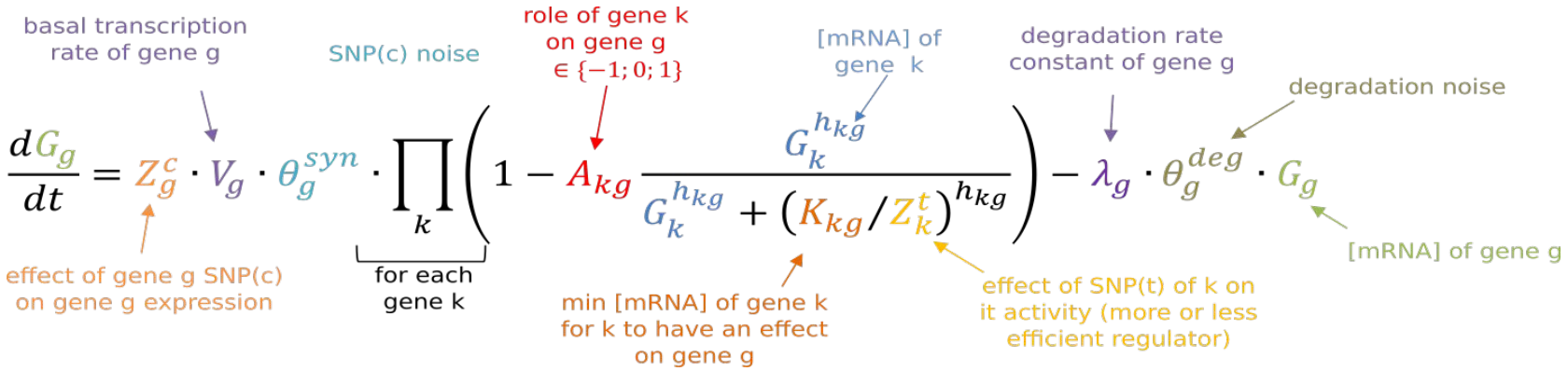
- - - - Prédiction (62%)

Jeu de données artificiel

Simulateur de données

SysGenSIM

Simulateur d'expression de gènes en contexte génétique-génomique

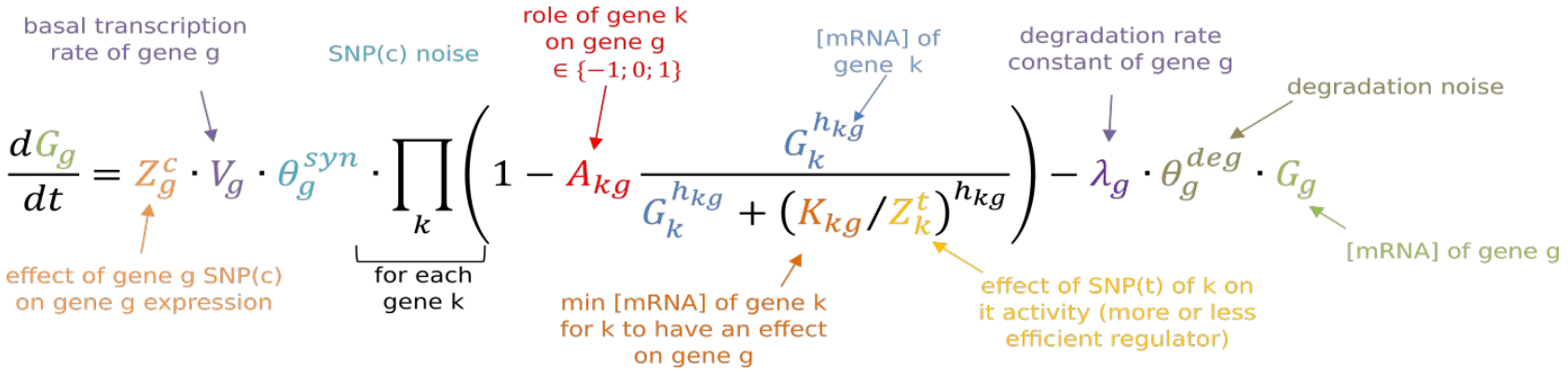


Jeu de données artificiel

Simulateur de données

SysGenSIM

Simulateur d'expression de gènes en contexte génétique-génomique



Problème : SysGenSIM adapté aux RIL (individus homozygotes)

→ ***Déterminer des haplotypes pour chacun des gènes***

→ ***Adapter SysGenSIM aux hybrides***

→ ***Choisir des valeurs pour les différents paramètres***

Jeu de données artificiel

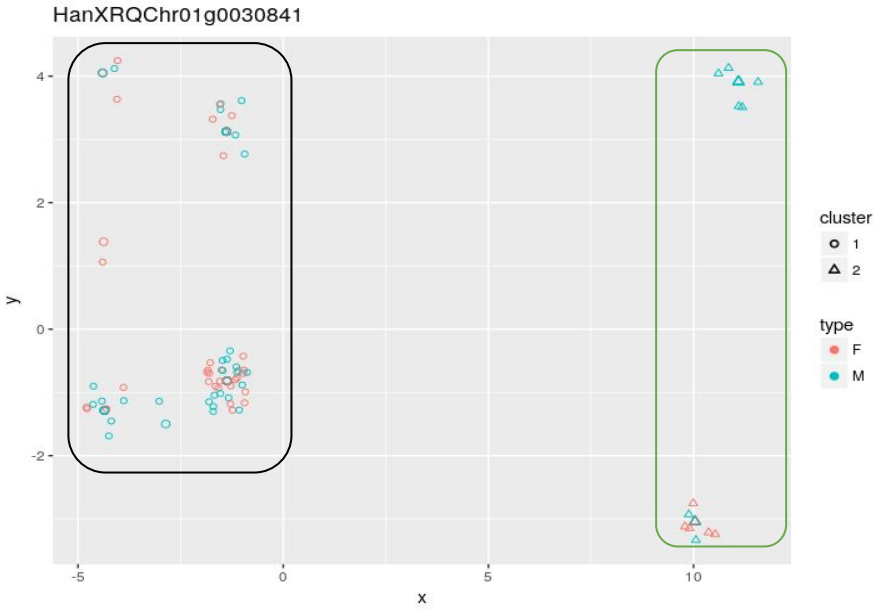
Déterminer des haplotypes

Jeu de données artificiel

Déterminer des haplotypes

parents avec les mêmes SNP

Clustering pour déterminer 2 haplotypes possibles (chez les parents)

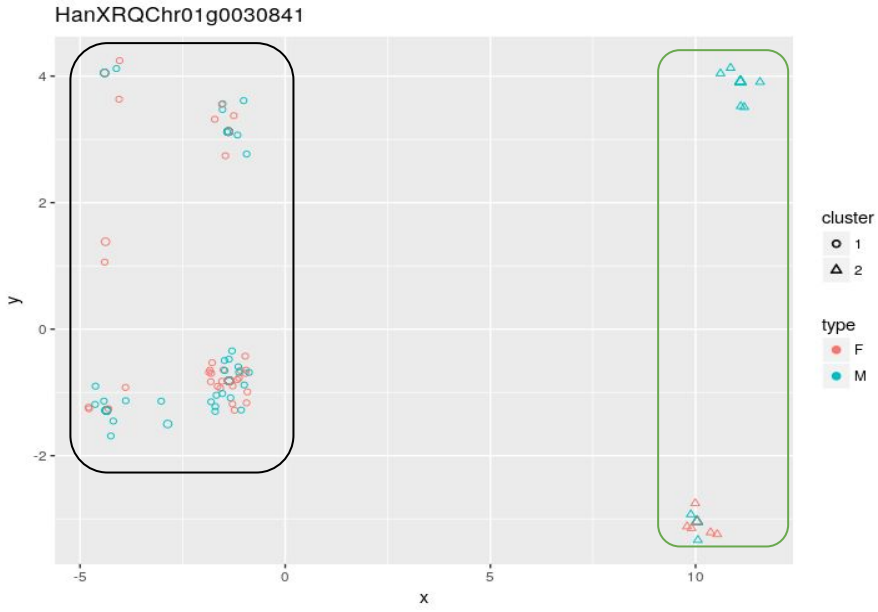


Jeu de données artificiel

Déterminer des haplotypes

parents avec les mêmes SNP

⇨ Clustering pour déterminer 2 haplotypes possibles (chez les parents)



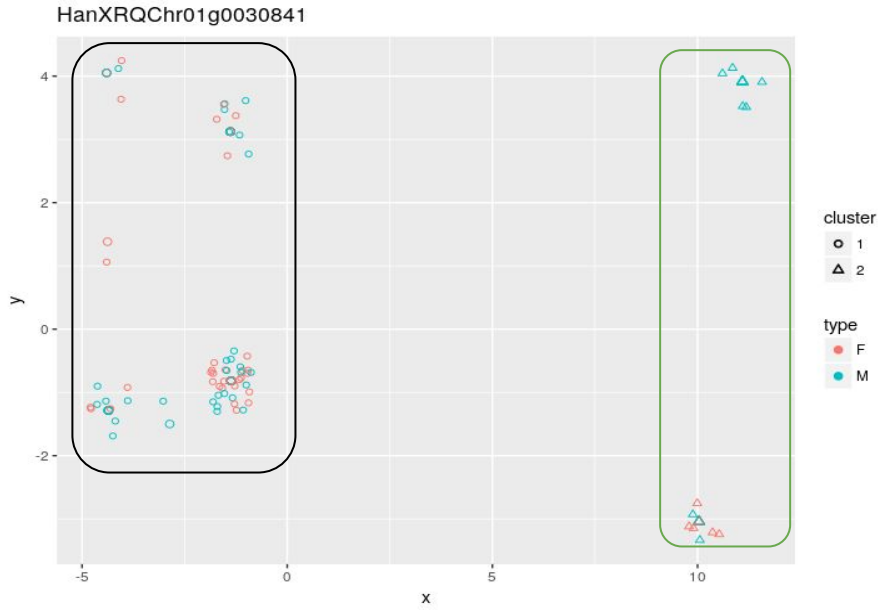
	état allélique
parents	0 sauvage 1 muté
hybride	0 homozygote sauvage 1 hétérozygote (sauvage-muté) 2 homozygote muté

Jeu de données artificiel

Déterminer des haplotypes

parents avec les mêmes SNP

→ Clustering pour déterminer 2 haplotypes possibles (chez les parents)

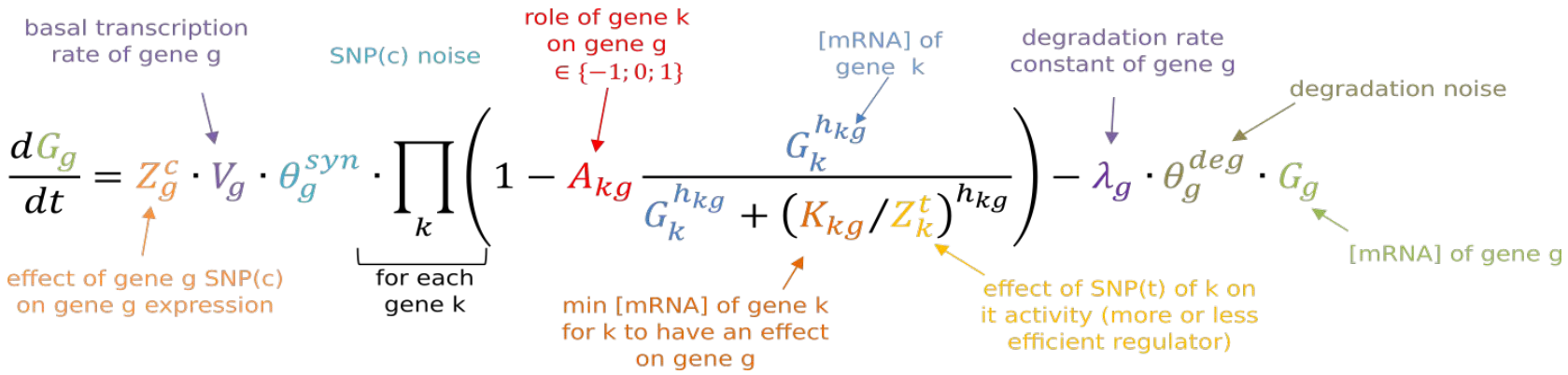


	état allélique
parents	0 sauvage 1 muté
hybride	0 homozygote sauvage 1 hétérozygote (sauvage-muté) 2 homozygote muté

→Tableau genotype x SNP muté et/ou sauvage

Jeu de données artificiel

Adapter SysGenSIM aux hybrides



La matrice des Z : genotype x gène → SNP muté ou sauvage (déterminée avant)

SysGenSIM : fonctionne avec des RIL → mutation portée par les deux allèles

wt	1
m	0.75

Dans notre cas, modification de SysGenSIM pour que Z prennent les valeurs

wt - wt	1
m - m	0.75
wt - m	0.75 mutated dominance (10%) 0.87 additive effect (80%) 1 wt dominance (10%)

Jeu de données artificiel

Paramètres de SysGenSIM

1. *Network topology*

G1	G2
G2	G3
G2	G4
G4	G3

liste des arêtes orientés
non typées :
induction / répression

% induction / répression de l'expression à fixer
→ utilisation d'**information biologique**

AtRegNet : FT → gène (experimental)

Jeu de données artificiel

Paramètres de SysGenSIM

1. *Network topology*

G1	G2
G2	G3
G2	G4
G4	G3

liste des arêtes orientés
non typées :
induction / répression

% induction / répression de l'expression à fixer
→ utilisation d'**information biologique**

AtRegNet : FT → gène (experimental)

2. *Genotype parameters*

Matrice état SNP

	H1	H2	H3	H4	H4
G1	1	1	1	1	0
G2	0	0	0	1	1
G3	0	1	0	0	1
G4	1	1	0	0	0

% de SNP avec un effet cis ou un effet trans

Jeu de données artificiel

Paramètres de SysGenSIM

1. *Network topology*

G1	G2
G2	G3
G2	G4
G4	G3

liste des arêtes orientés
non typées :
induction / répression

% induction / répression de l'expression à fixer
→ utilisation d'**information biologique**

AtRegNet : FT → gène (experimental)

2. *Genotype parameters*

Matrice état SNP

	H1	H2	H3	H4	H4
G1	1	1	1	1	0
G2	0	0	0	1	1
G3	0	1	0	0	1
G4	1	1	0	0	0

% de SNP avec un effet cis ou un effet trans

3. *Model Parameters*

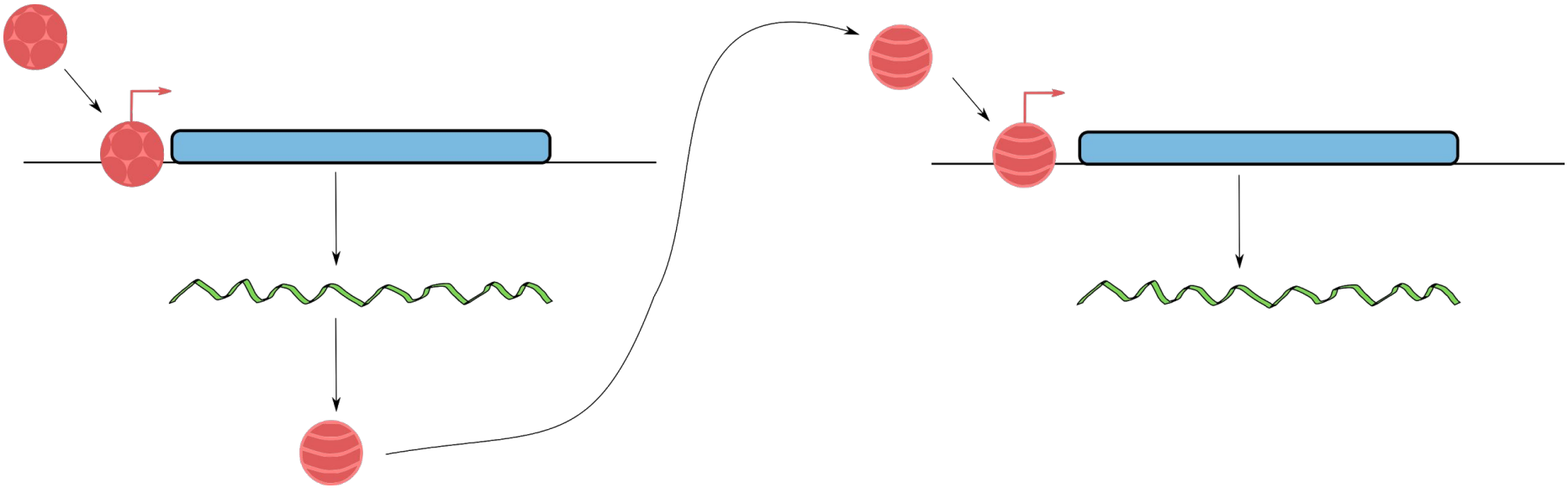
- Basal transcription rate
- Cooperativity coefficient
- Basal degradation rate
- Basal level for activity

par défaut

Jeu de données artificiel

Paramètres de SysGenSIM

Effet cis ou trans des SNP



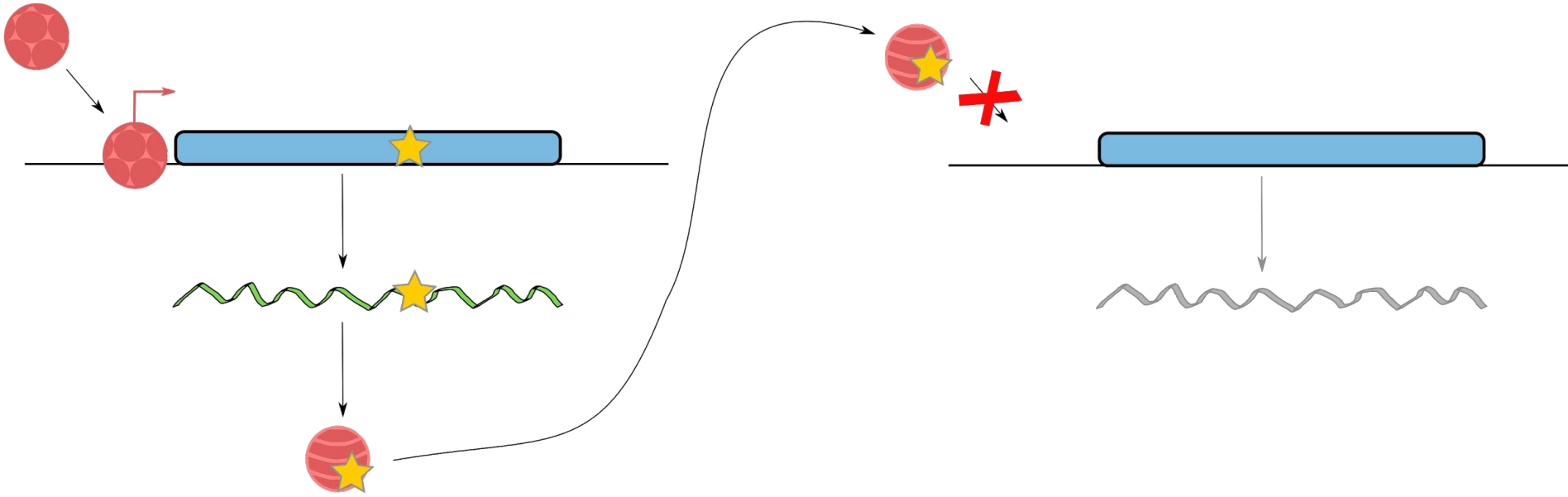
Jeu de données artificiel

Paramètres de SysGenSIM

Effet cis ou trans des SNP

SNP avec un effet trans :

Le SNP modifie l'activité et l'efficacité de la protéine issu du gène portant le SNP



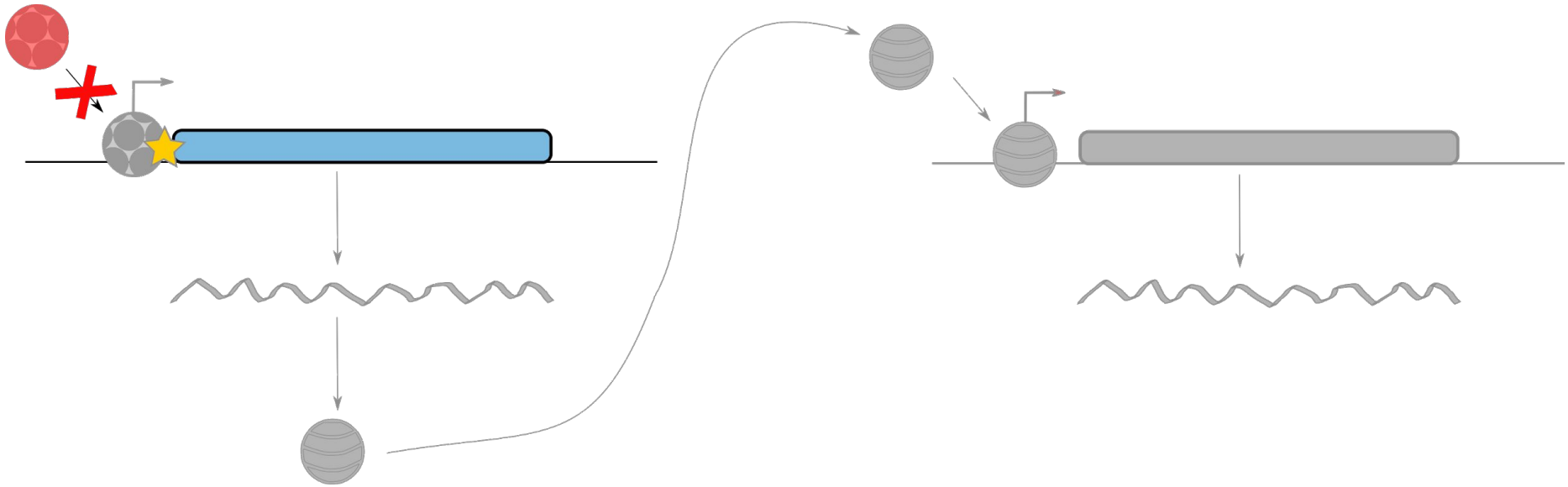
Jeu de données artificiel

Paramètres de SysGenSIM

Effet cis ou trans des SNP

SNP avec un effet cis :

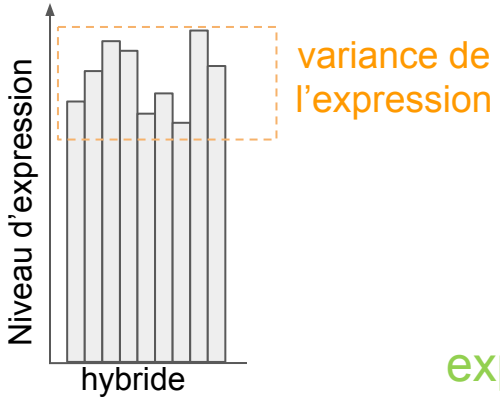
Le SNP modifie le taux de transcription du gène portant le SNP



→ Quel information utiliser pour choisir le % de SNP avec un effet cis ou trans ?

Jeu de données artificiel

Ajuster les paramètres



héritabilité : % de la variance de l'expression d'un gène expliqué par le génotype des parents

expression moyenne du gène

expression du gène pour l'hybride *fm*

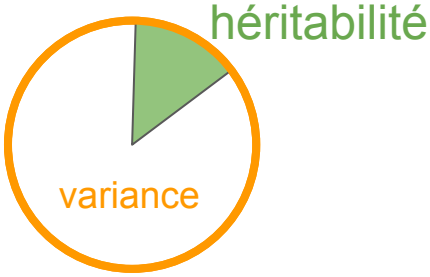
$$y_{fm} = \mu + F_f + M_m + \epsilon_{fm}$$

effet aléatoire femelle

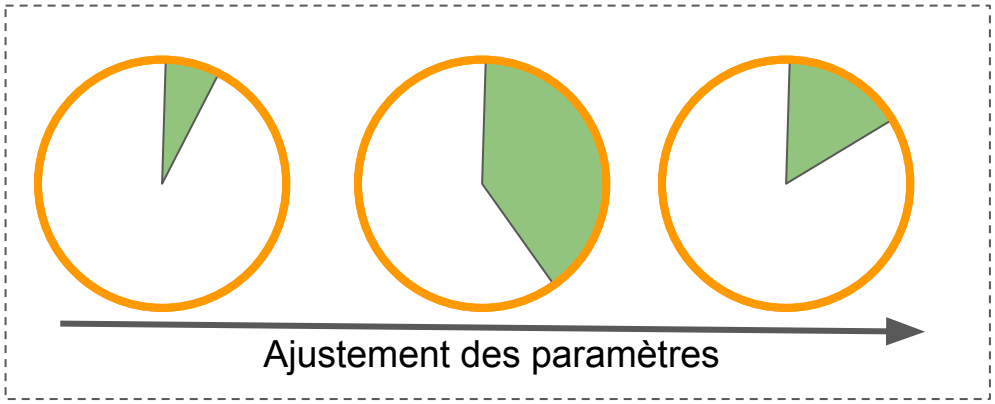
effet aléatoire mâle

Vrai jeu de données

15EX05 fluidigm



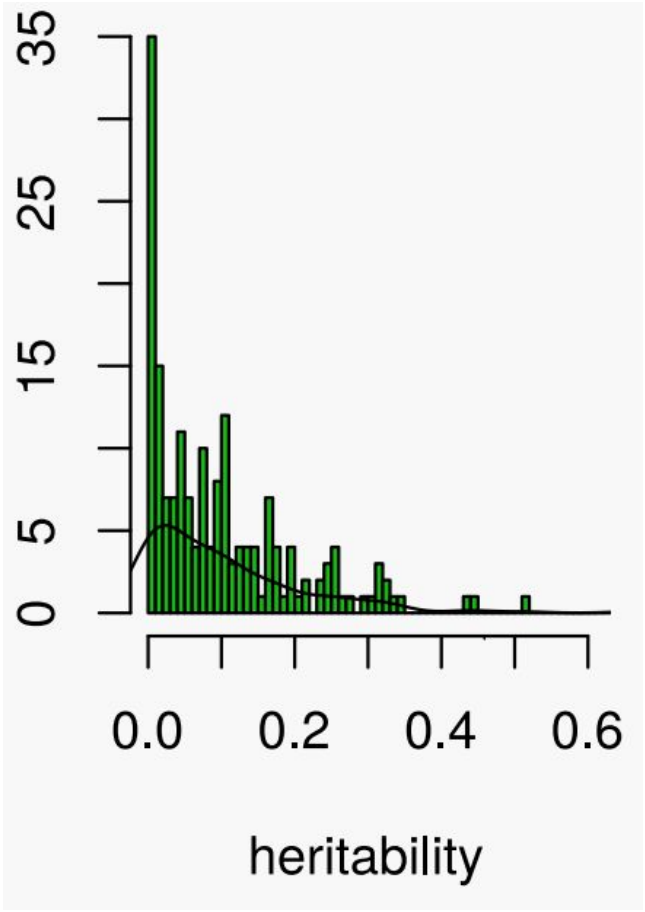
Jeu de données artificiel



Jeu de données artificiel

Ajuster les paramètres

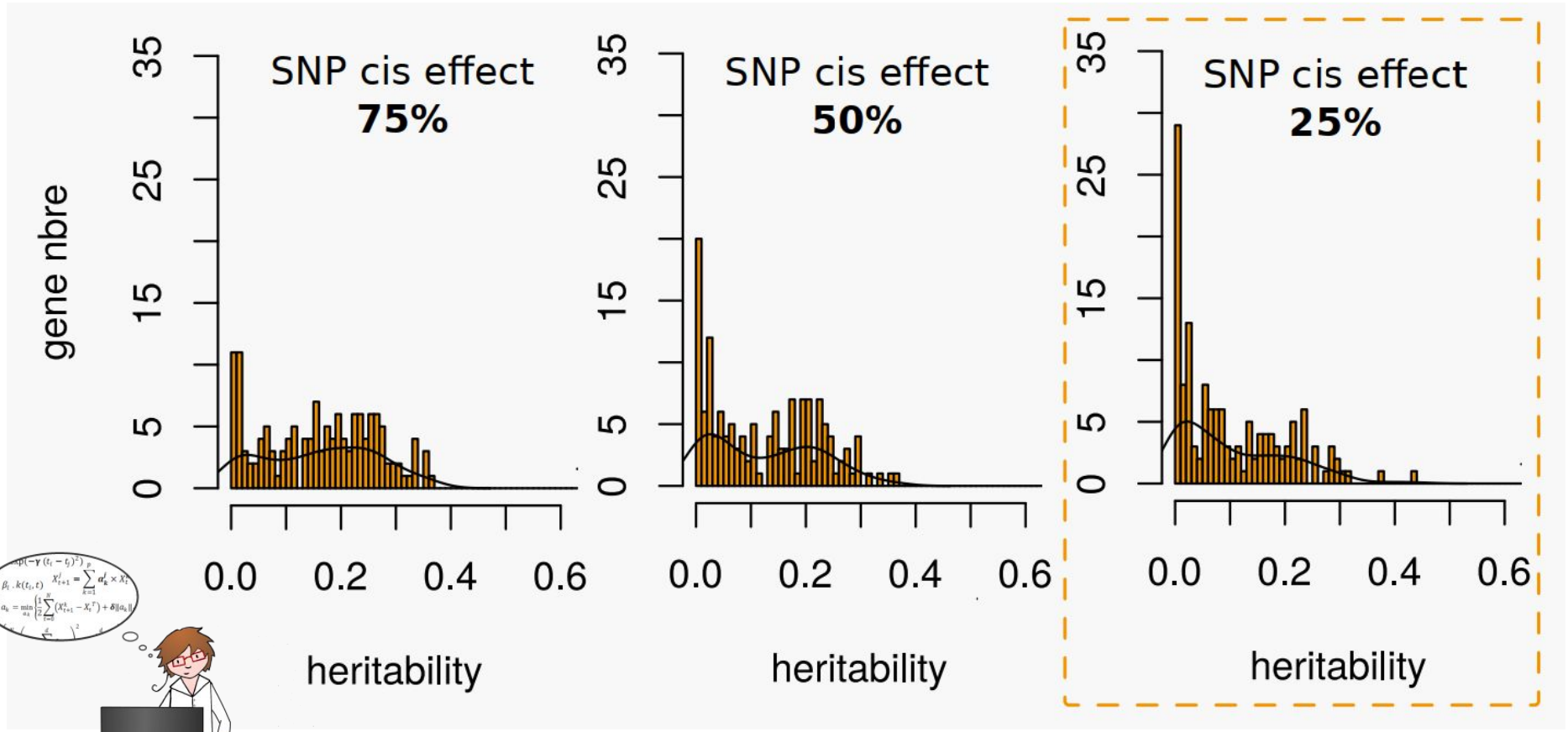
Distribution heritabilté 15EX05



Jeu de données artificiel

Ajuster les paramètres

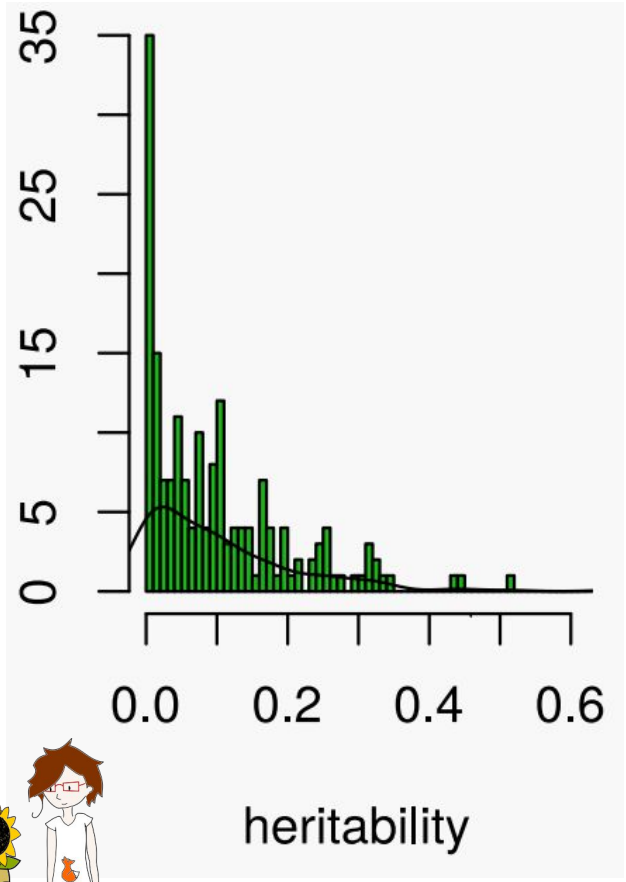
Distribution héritabilité données simulées (SysGenSIM)

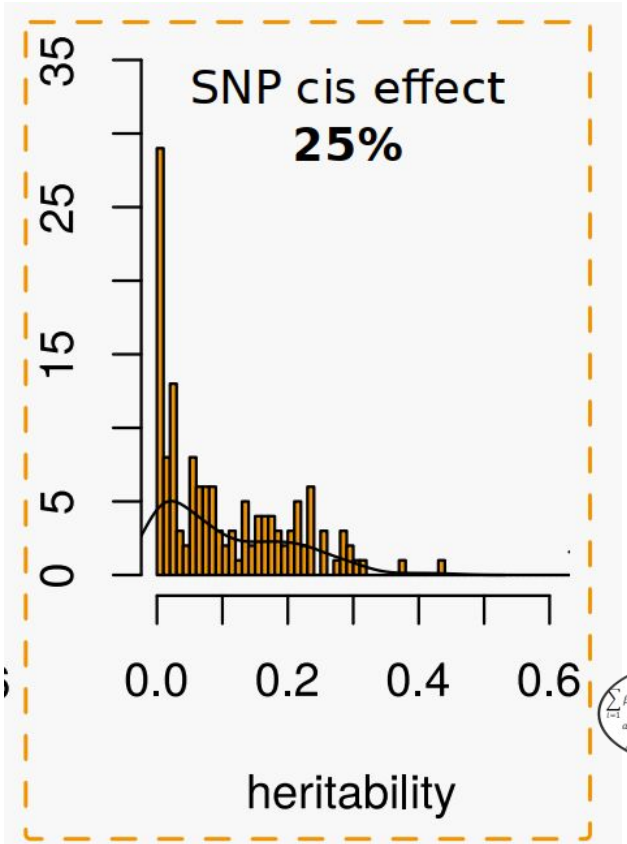


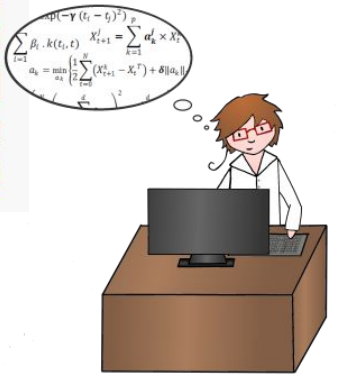
Jeu de données artificiel

Ajuster les paramètres

Comparaison héritabilité 15EX05 - SysGenSIM data

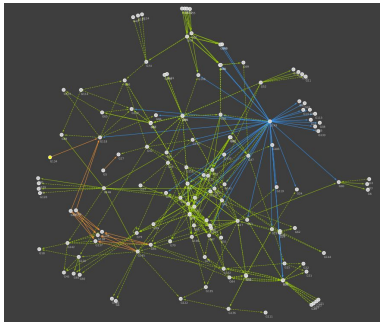






Jeu de données artificel

Réseau



144 gènes
313 arêtes

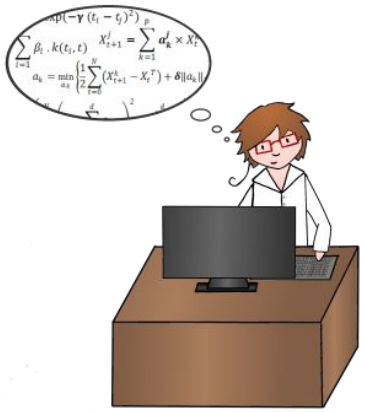
Données pour **463 génotypes**

- mesure d'expressions

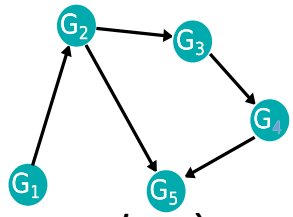
données quantitative
1 valeur d'expression par gène x génotype

- mesure de SNP

données qualitative (0,1,2)
1 valeur de SNP par gène x génotype

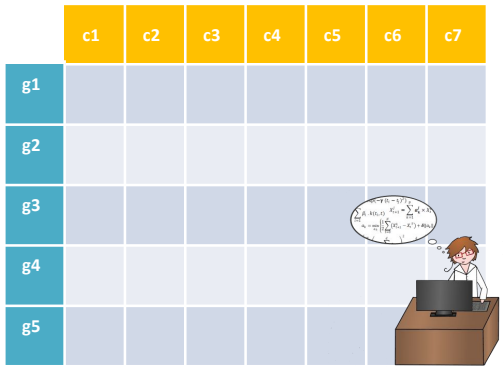


Test de méthodes d'inférence



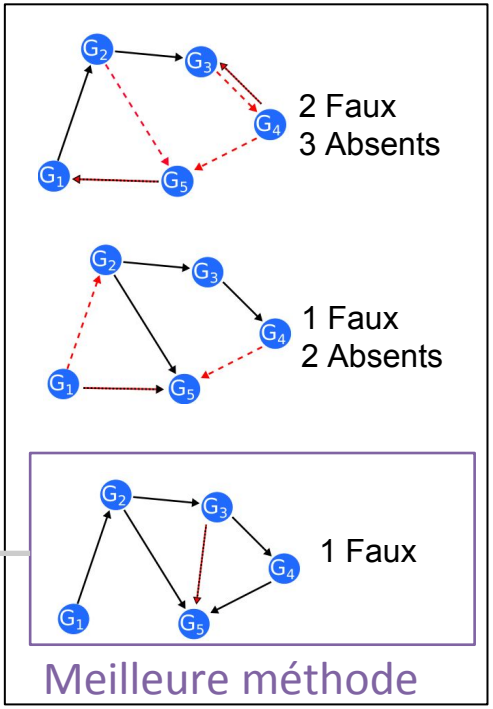
Réseau de gènes artificiel

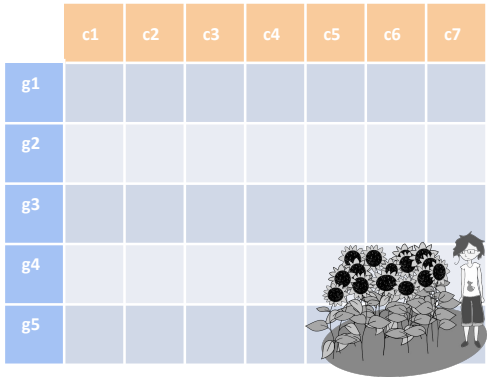
Simulation d'expérience



Données simulées

Test méthode d'inférence



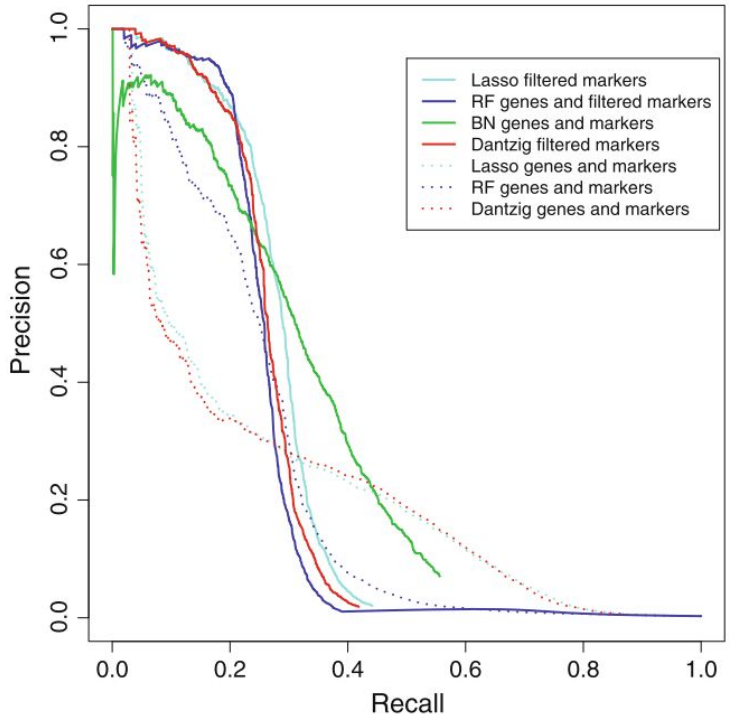


Jeux de données 15EX05

Inférence
du réseau de gènes
sécheresse x hétérosis

Test de méthodes d'inférence

Challenge DREAM5 (2011) :
Inférence de réseau en contexte Genetical-Genomics
Mesures de l'expression + mesures de SNP (données artificielles)



Plusieurs méthodes testées avec de bon résultats

- lasso
- random forest
- bayesian network

Fournissent des listes d'arêtes triées

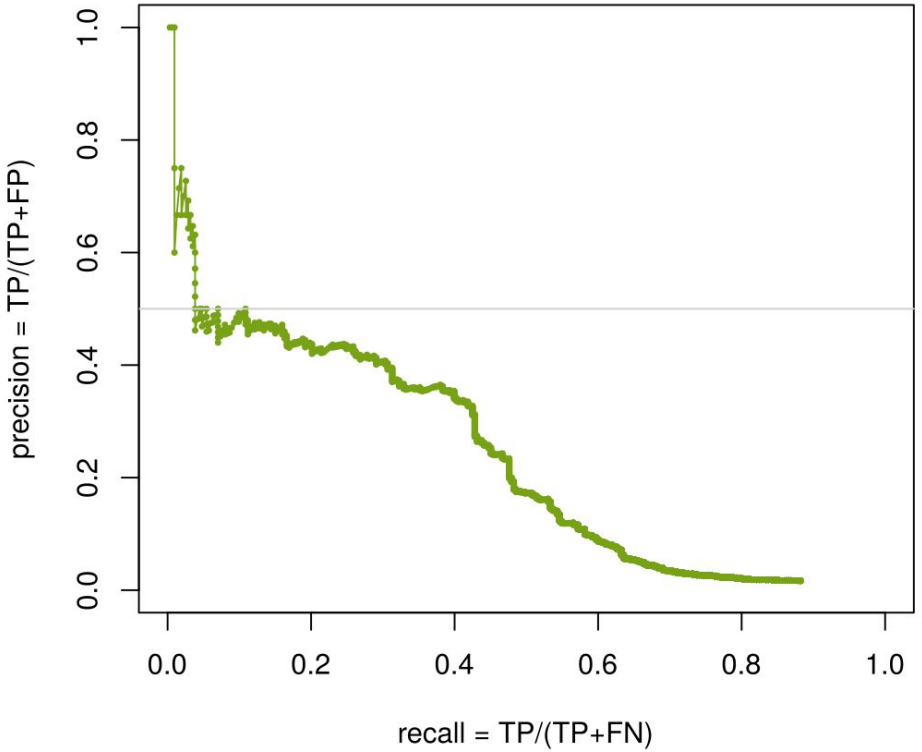
→ Est ce que ces méthodes donnent d'aussi bon résultats sur notre jeu de données artificiel ?

Test de méthodes d'inférence

Test des méthodes développées pour DREAM5 sur notre jeu de données artificiel

Lasso

data : expression + SNP
nb bootstrap : 50
nb lambda : 100



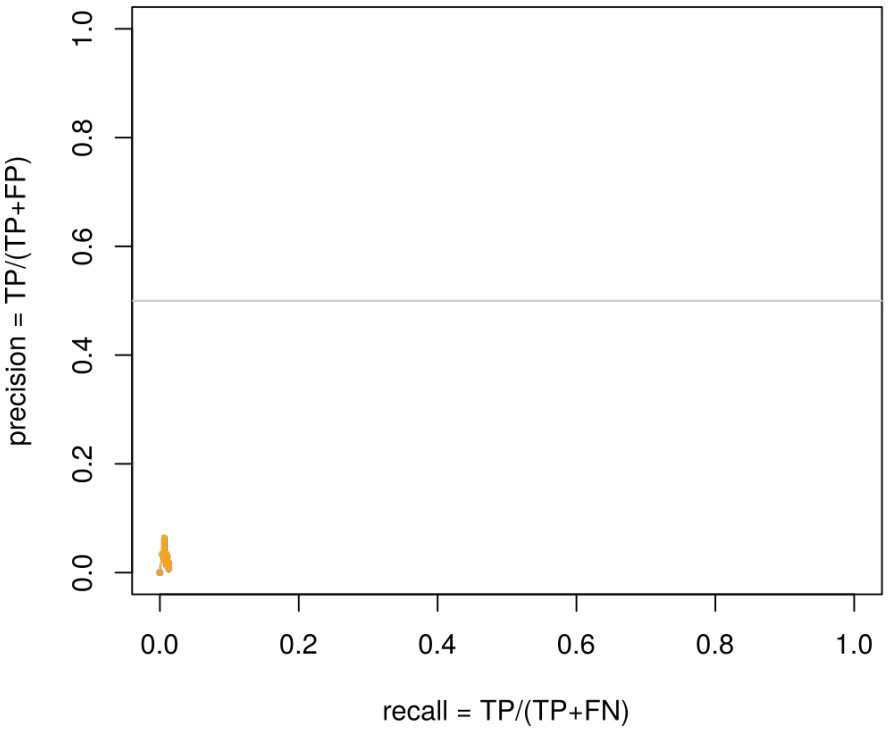
- problèmes d'orientations des arêtes
- 1ères arrêtes trouvées avec les données d'expression et de SNP
- résultats moins bon

Test de méthodes d'inférence

Test des méthodes développées pour DREAM5 sur notre jeu de données artificiel

Random forest

data : expression + SNP
nb tree : 100 000

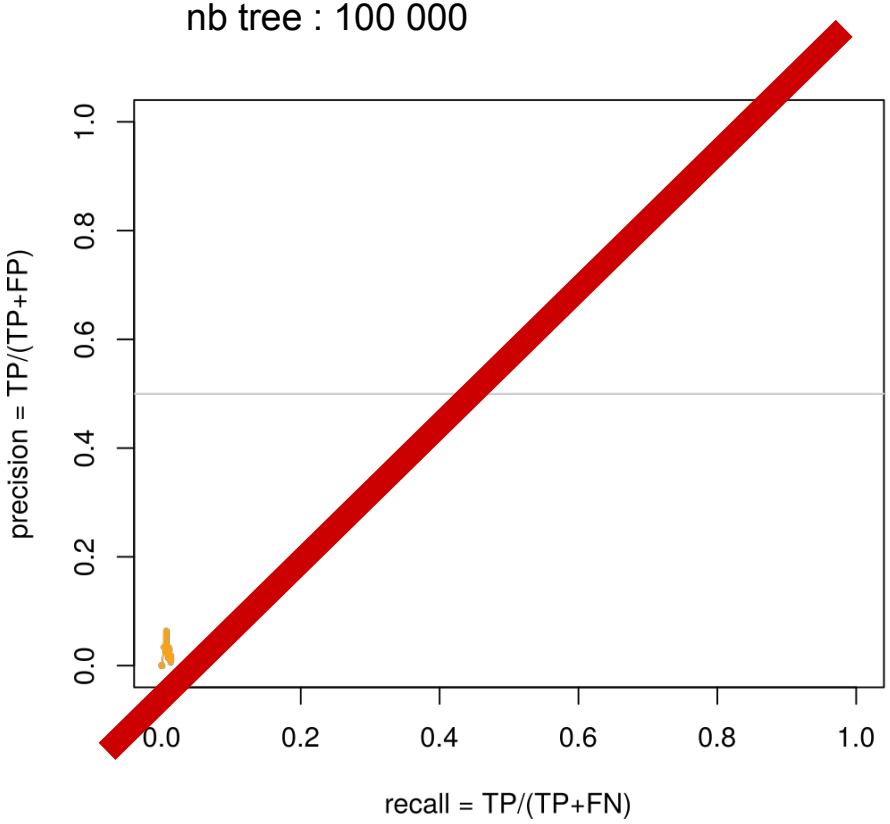


Test de méthodes d'inférence

Test des méthodes développées pour DREAM5 sur notre jeu de données artificiel

Random forest

data : expression + SNP
nb tree : 100 000

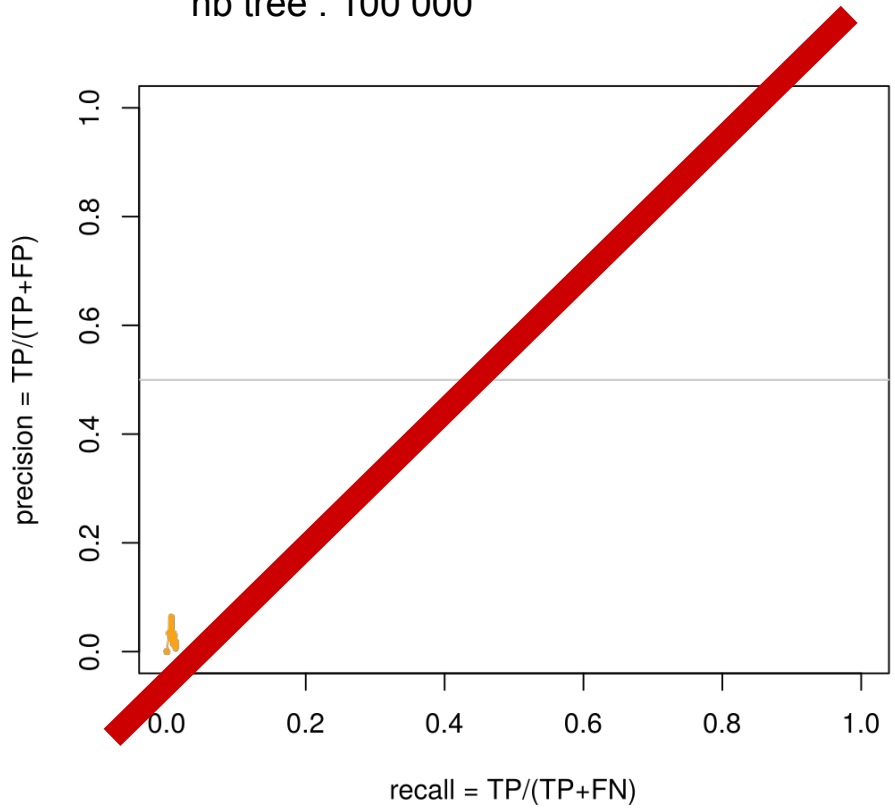


Test de méthodes d'inférence

Test des méthodes développées pour DREAM5 sur notre jeu de données artificiel

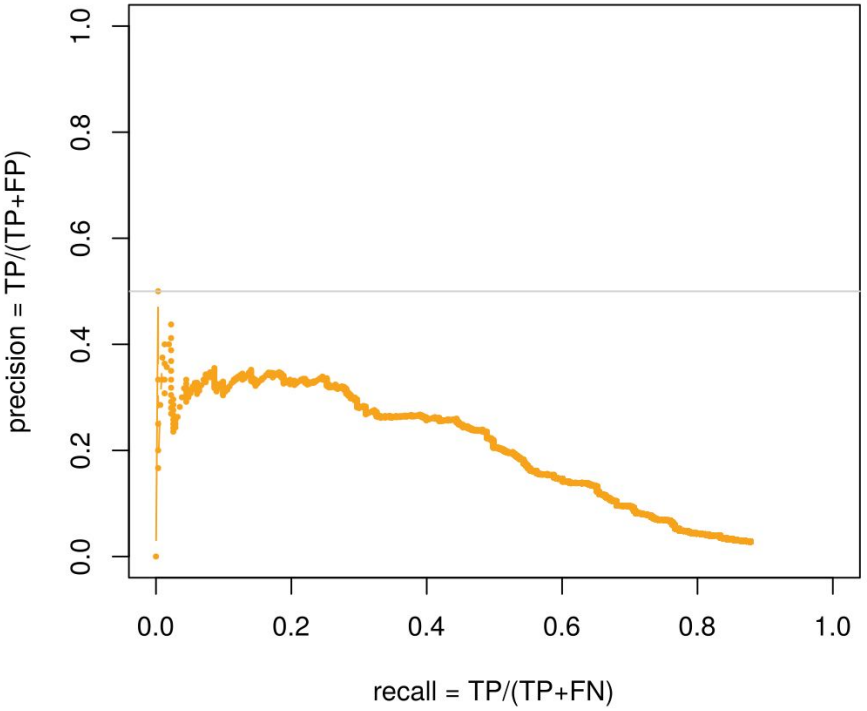
Random forest

data : expression + SNP
nb tree : 100 000



Random forest

data : expression
nb tree : 100 000



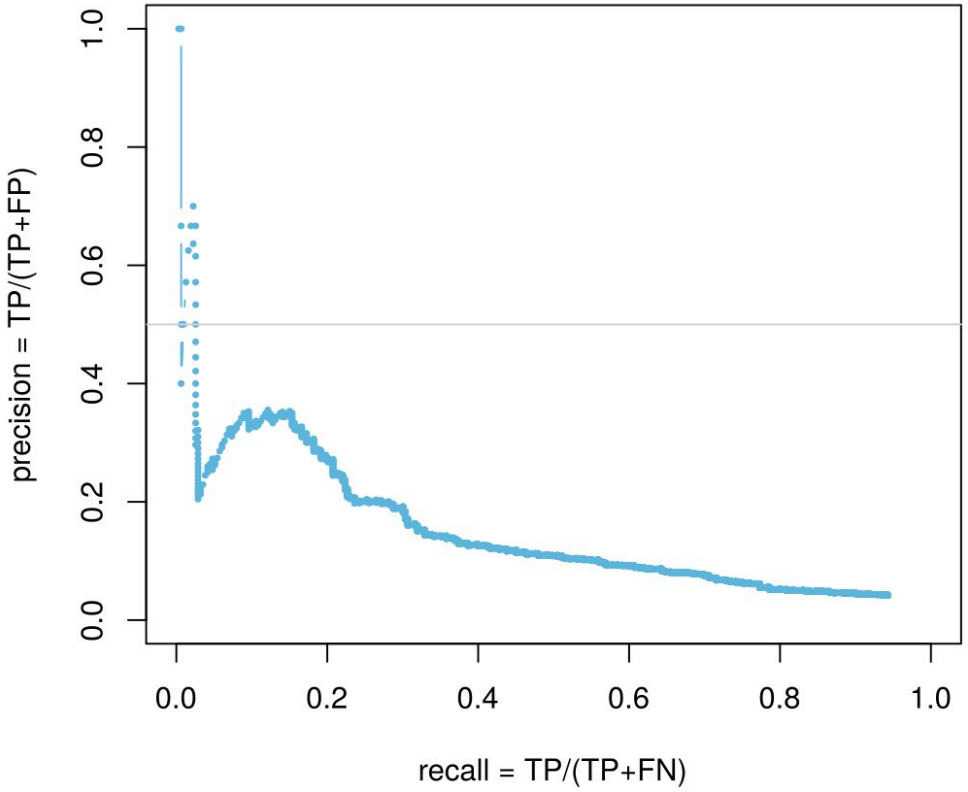
Résultats moins bon que pour les jeux de données test fournis

Test de méthodes d'inférence

Exact structure learning Bayesian Network

data : expression (discretized in 3 values) + SNP
parents par noeud : 2 au maximum
nb bootstrap : 100
equivalent sample size : varie entre 10^{-16} et 10

outil : **GOBNILP**



- 1ères arrêtés SNP et expression
- problèmes d'orientation des arêtes

Test de méthodes d'inférence

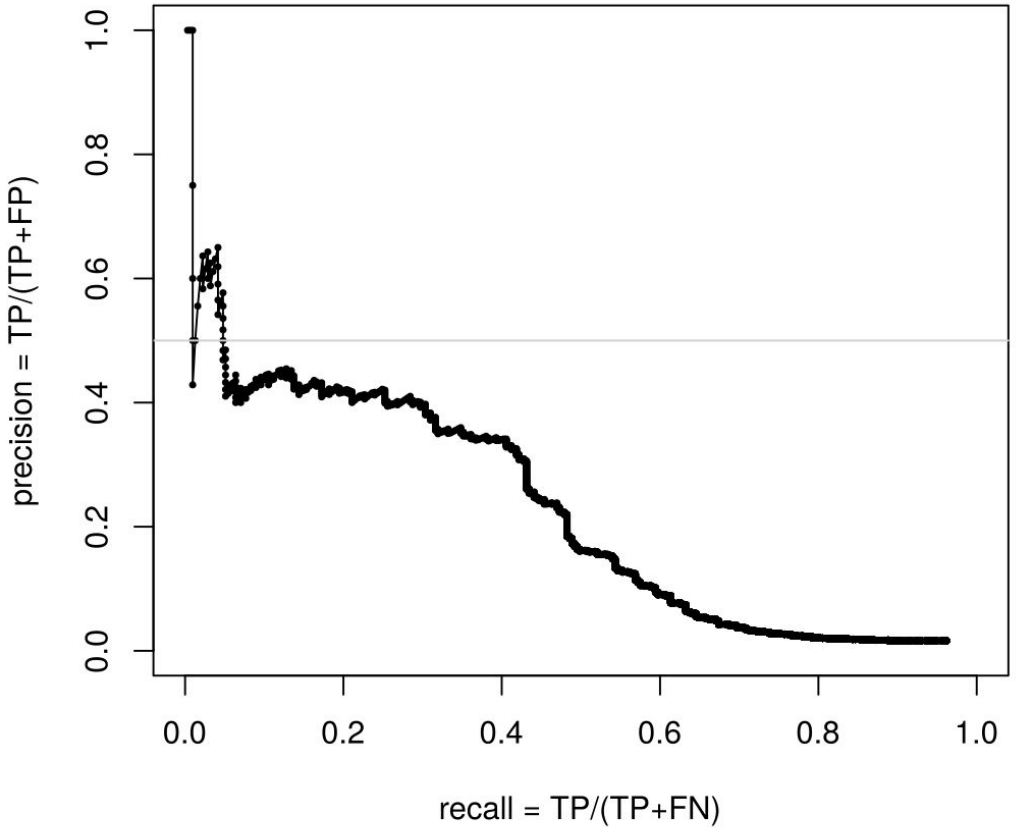
Peu de méthodes testées, et méthodes imparfaites

Résultats moyen voir mauvais pour ces différentes méthodes

→ *combiner les résultats obtenus par une **méta-analyse***

$$S_{ij} = \sum_{m \in \mathcal{M}} \log\left(1 - r_{ij}^{m}\right)$$

$$r_{ij} = 1 - \exp\left(S_{ij}\right)$$

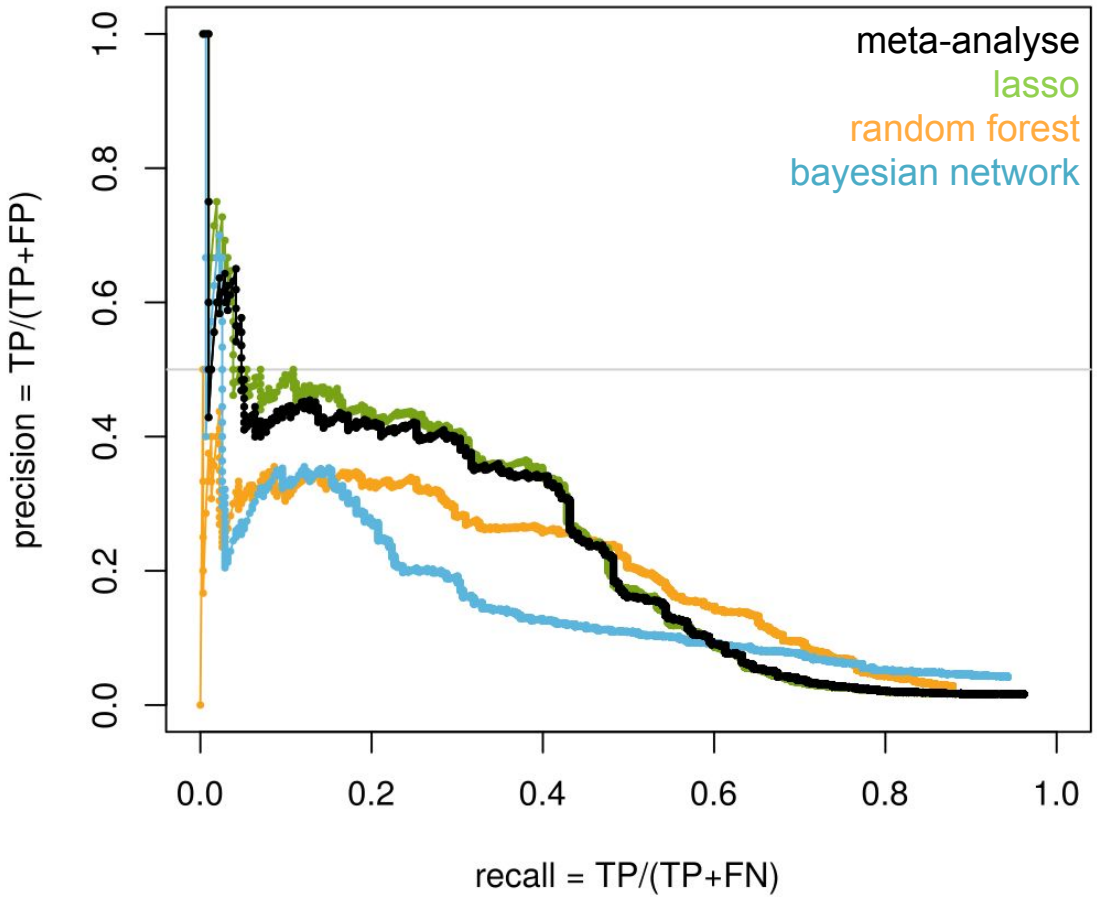


Test de méthodes d'inférence

Peu de méthodes testées, et méthodes imparfaites

Résultats moyen voir mauvais pour ces différentes méthodes

→ *combiner les résultats obtenus par une **méta-analyse***



Conclusions et Perspectives

- Mise en évidence de **difficultés** pour les approches développées lors de DREAM5
  - Plus **faible contribution** de l'information des marqueurs (**SNP**)
  - **Connectivité** assez **élevée** du réseau artificiel (nombreux hubs reliés entre eux)

- Tester d'autres approches de reconstruction
  simples corrélations, modèles graphiques de type champ de Markov ou réseau Bayésien à variables mixtes discrètes et continues, *etc.*

- Application au **jeu de données 15EX05**
  - Validation du réseau prédit par de la bibliographie et la recherche de motifs de facteurs de transcription
  - Proposer des gènes supplémentaires à mesurer en qPCR pour compléter le réseau

Merci de votre attention

Si vous voulez que je teste votre méthode

lise.pomies@inra.fr