



Apprentissage du réseau de gènes du tournesol faut-il utiliser les réseaux Bayésiens ?

Simon de Givry

Journées Francophones sur les Réseaux Bayésiens et les Modèles Graphiques Probabilistes

11/10/2021

- **Contexte biologique**
- Méthodes d'apprentissage de la structure d'un réseau
- Résultats expérimentaux
- Conclusion et perspectives

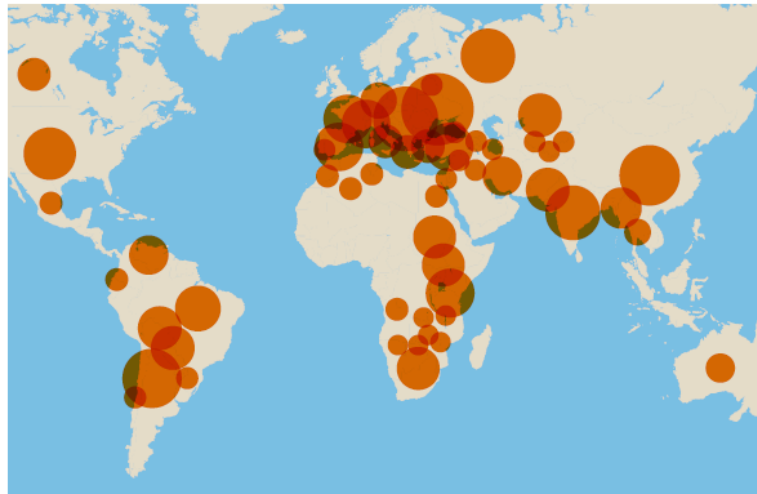
Projet SUNRISE 2012-2020



La France et l'Europe au coeur de la production internationale de tournesol

30 millions d'hectares
dans le monde,
avec 71 % des surfaces en Europe

39 millions de tonnes de graines
produites dans le monde,
avec 80 % de la production en Europe



données en 2019

Projet SUNRISE 2012-2020



The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution

Hélène Badouin^{1*}, Jérôme Guzy^{1*}, Christopher J. Grass^{1,2*}, Florent Murat³, S. Evan Staton², Ludovic Cottret¹, Christine Lelandais-Brière^{4,5}, Gregory L. Owens², Sébastien Carrère³, Baptiste Mayonade¹, Ludovic Legrand¹, Navdeep Gill², Nolan C. Kane^{2,6}, John E. Bowers², Sarel Hubner^{2,8,9}, Arnaud Belles¹⁰, Aurélie Bérard¹¹, Hélène Bergès¹⁰, Nicolas Blanchet¹, Marie-Claude Boniface¹, Dominique Brunel¹¹, Olivier Catrice¹, Nadia Chaidir^{2,12}, Clotilde Claude¹³, Cécile Donnadieu¹⁴, Thomas Faraut¹⁵, Ghislain Fievet¹, Nicolas Helmstetter¹⁰, Matthew King^{2,16}, Steven J. Knapp¹⁷, Zhao Lai^{18,19}, Marie-Christine Le Paslier¹¹, Yannick Lippi¹, Lolita Lorenzon¹, Jennifer R. Mandel²⁰, Gwenola Marage¹, Gwenaëlle Marchand¹,

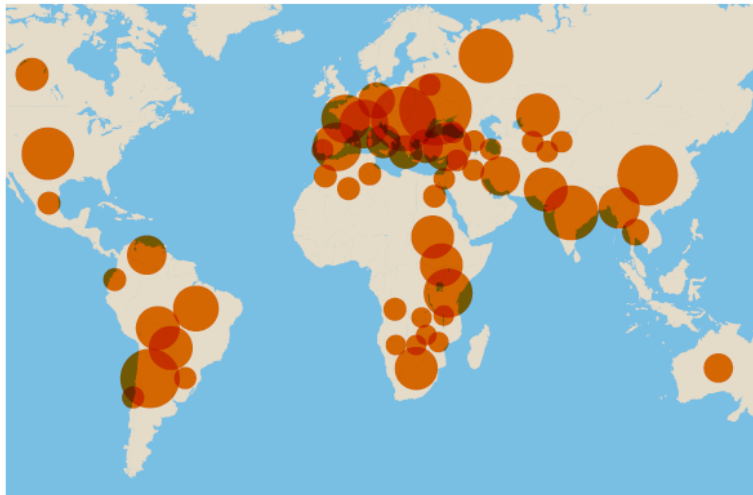


(Badouin et al., *Nature* 2017)

La France et l'Europe au coeur de la production internationale de tournesol

30 millions d'hectares dans le monde, avec 71 % des surfaces en Europe

39 millions de tonnes de graines produites dans le monde, avec 80 % de la production en Europe



données en 2019

Objectifs du projet

- 1) Améliorer la **production d'huile** dans des conditions adaptées au **changement climatique** et respectueuses de l'environnement
- 2) Comprendre les **bases génétiques et moléculaires** contrôlant la physiologie et le développement de la plante pour prédire les caractéristiques des **hybrides**
- 3) Développer pour l'ensemble de la filière des **outils et des méthodes** permettant de mieux maîtriser la culture



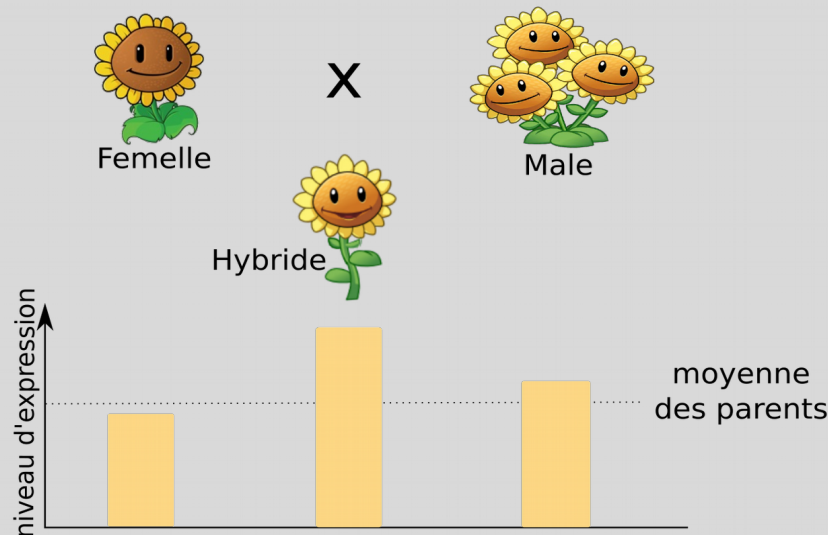
Apprendre le **réseau de régulation de gènes** permettant de comprendre la réponse du tournesol à la **sécheresse** en interaction avec l'**hétérosis**

Hétérosis

phénotype de l'**hybride**

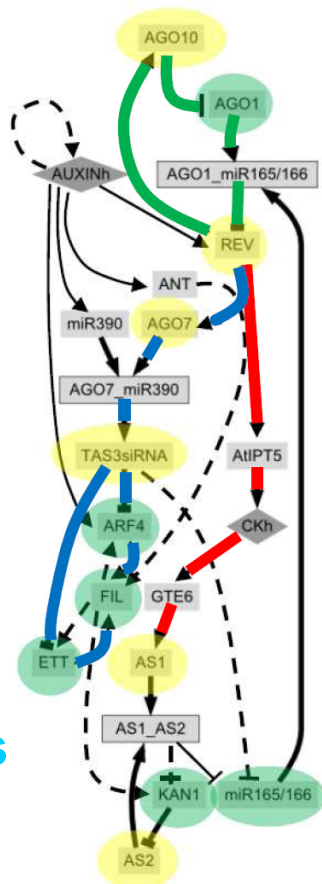
≠

moyenne du phénotype
des parents

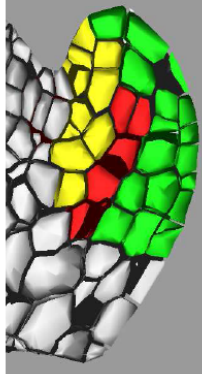


Mesures *RNAseq* sur 58050 gènes (Gody et al., 2020)

Exemple du réseau de développement de la fleur d'arabette

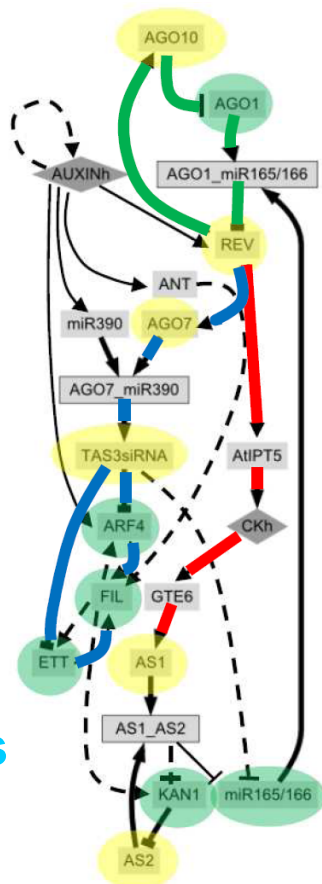


(Henry et al., *Molecular BioSystems* 2013)

	Zone 1 Adaxial	Zone 2 Vascular	Zone 3 Abaxial	
AGO1	1	1	1	
AGO10	1	0	0	
AGO7	1	0	0	
ANT	1	1	1	
ARF4	0	1	1	
AS1	1	1	0	
AS2	1	0	0	
ETT	0	1	1	
FIL	0	1	1	
KAN1	0	0	1	
MiR165/166	0	1	1	
MIR390	1	1	1	
REV	1	0	0	
TAS3	1	0	0	

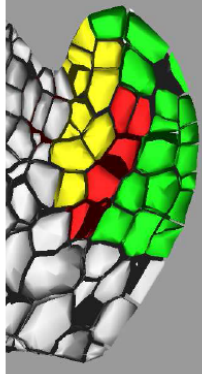
p= 15 gènes
37 régulations

Exemple du réseau de développement de la fleur d'arabette



$p =$ 15 gènes
37 régulations

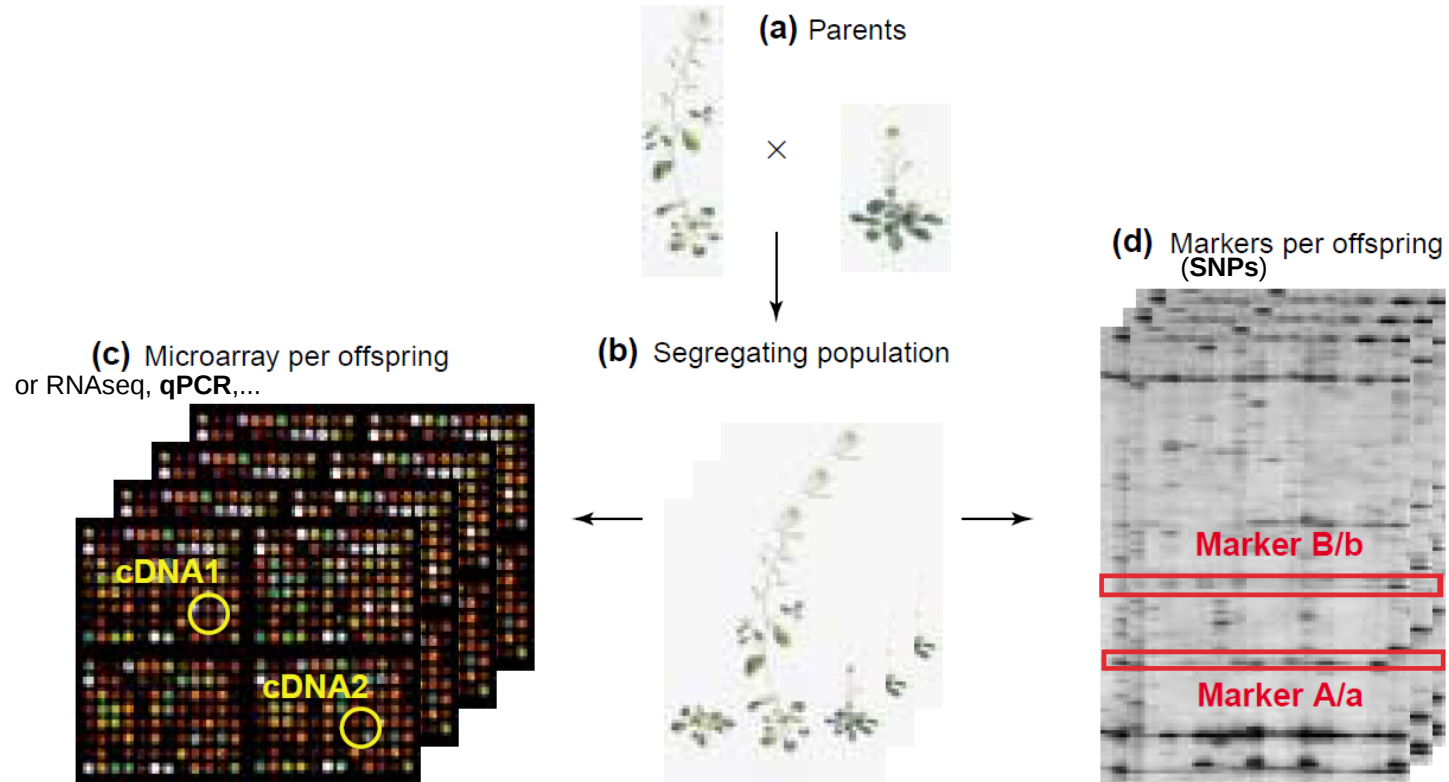
(Henry et al., *Molecular BioSystems* 2013)

	Zone 1 Adaxial	Zone 2 Vascular	Zone 3 Abaxial	
AGO1	1	1	1	
AGO10	1	0	0	
AGO7	1	0	0	
ANT	1	1	1	
ARF4	0	1	1	
AS1	1	1	0	
AS2	1	0	0	
ETT	0	1	1	
FIL	0	1	1	
KAN1	0	0	1	
MIR165/166	0	1	1	
MIR390	1	1	1	
REV	1	0	0	
TAS3	1	0	0	

Nombre de graphes dirigés possibles : $2^{p(p-1)}$

$p=15 : \sim 10^{63}$, $p=17 : \sim 10^{81}$

Protocole des données : génétique génomique

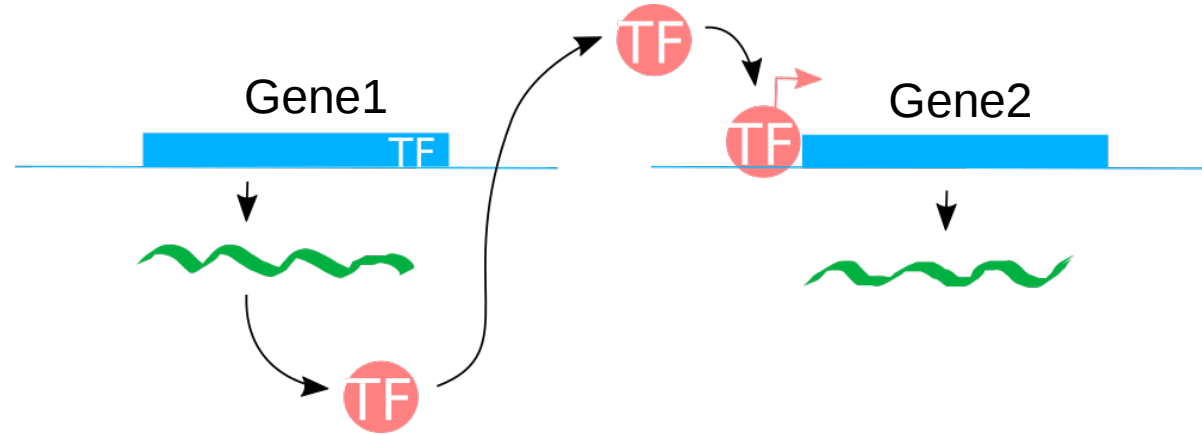


(Jansen & Nap, *TRENDS in Genetics* 2001)

Sélection d'un nombre restreint de gènes (<180)

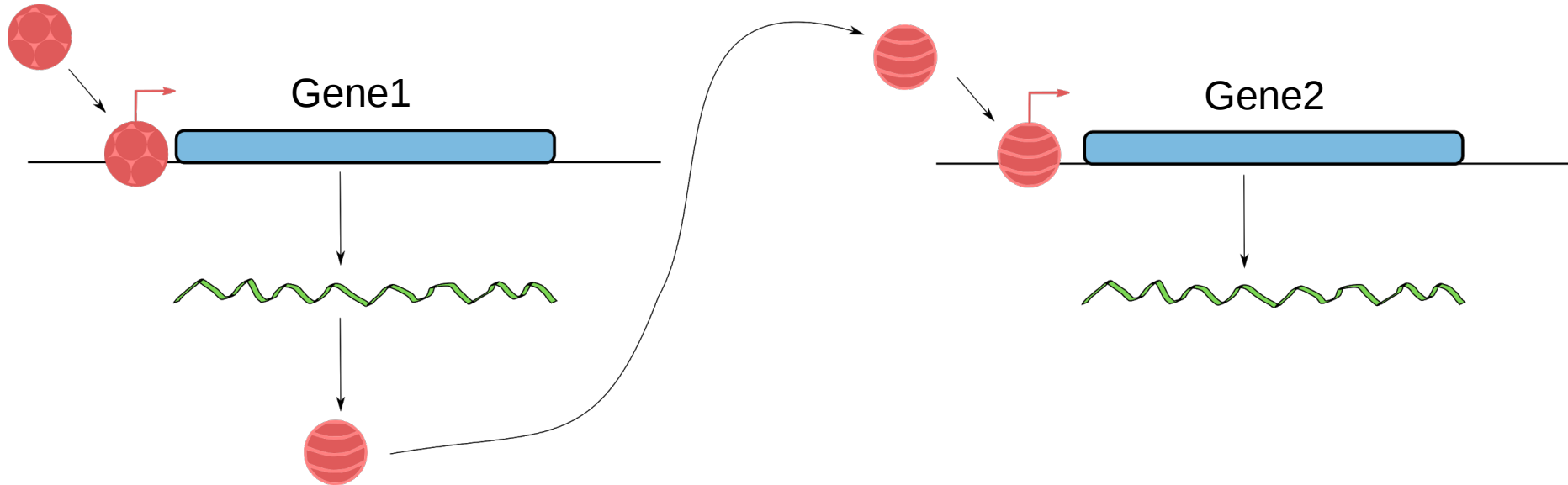
- Différentiellement exprimés en condition sécheresse x hétérosis
- **Facteurs de transcription** (*Transcription Factor TF*)

TF: gène capable de réguler l'expression d'autres gènes



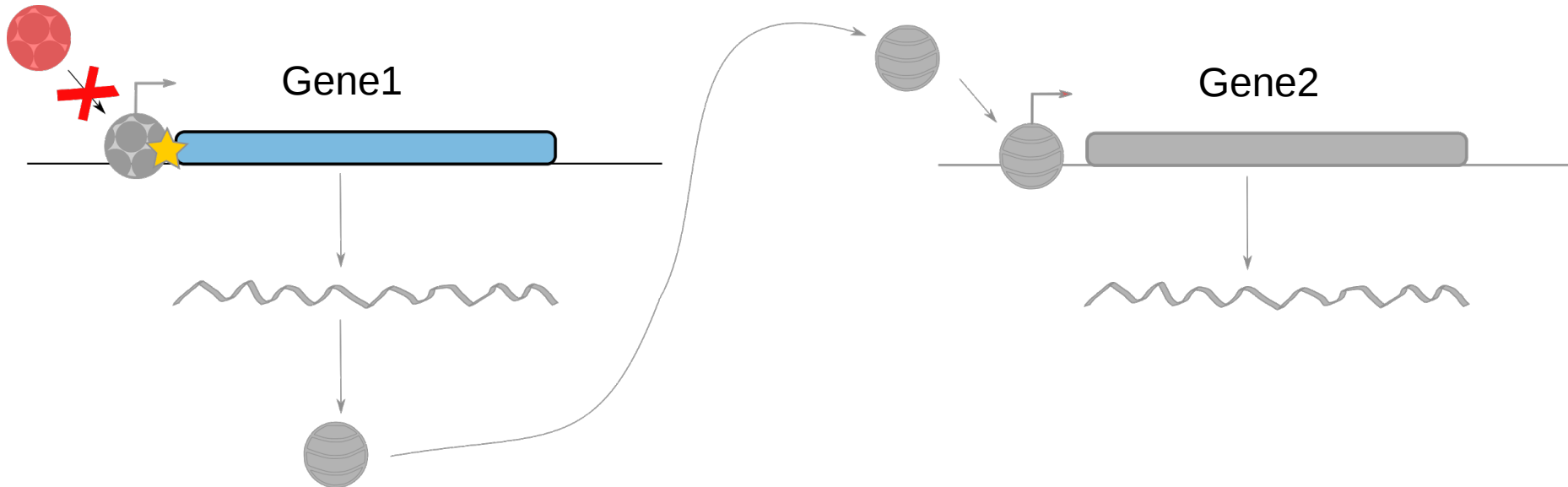
Single Nucleotide Polymorphisme (SNP)

- Mutation ponctuelle du génome (G ou A, C ou T)



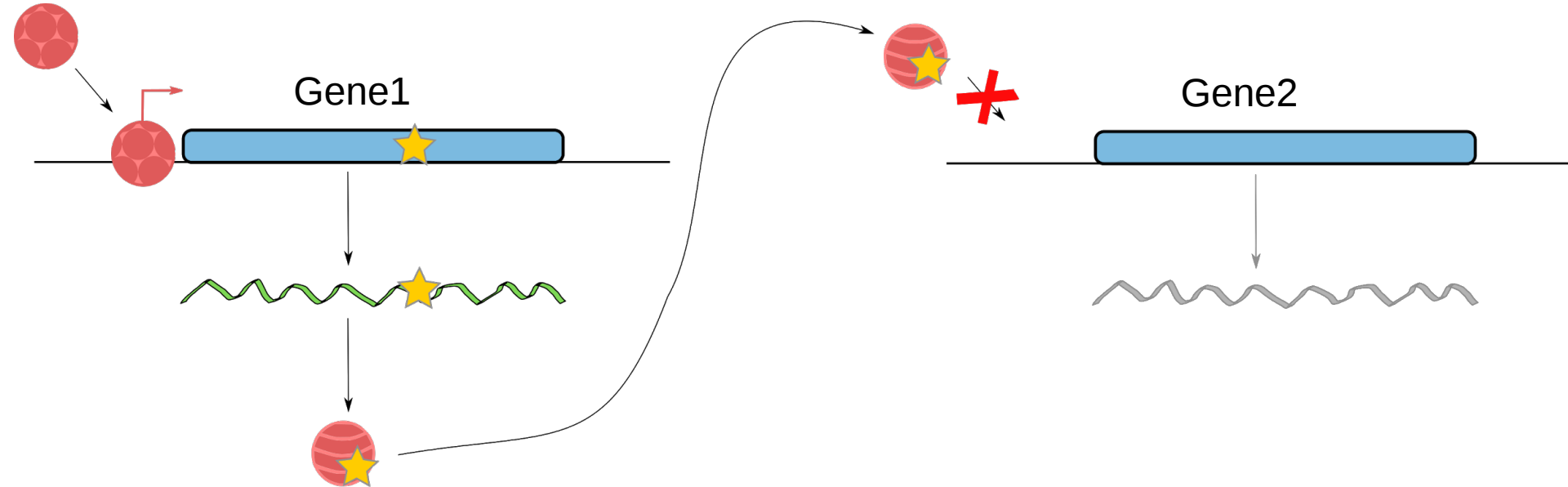
Single Nucleotide Polymorphisme (SNP)

- Mutation ponctuelle du génome
- 2 effets possibles sur l'expression des gènes (**cis** ou **trans**)



Single Nucleotide Polymorphisme (SNP)

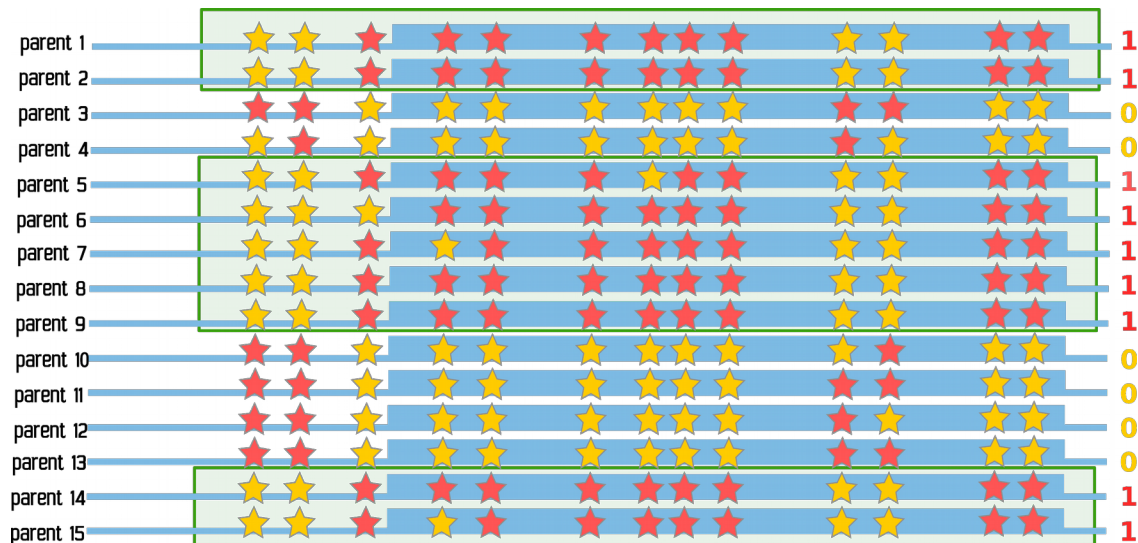
- Mutation ponctuelle du génome
- 2 effets possibles sur l'expression des gènes (cis ou **trans**)



Haplotype parental (SNPs fusionnés par gène)



Haplotype parental (SNPs fusionnés par gène)



Clustering (k-medoid) en 2 classes

0 : haplotype sauvage
1 : haplotype muté

2 haplotypes parentaux

=> **génotype** de l'hybride

$$M_i \in \{0, 1, 2\}$$

Apprentissage du réseau à partir de données mesurées ou simulées

Données mesurées en champs

$n=353$ hybrides *F1* de tournesol
(pool de 36 x 36 parents)

$p=173$ gènes *FT* stress x hétérosis
- mesures de l'expression par *qPCR*
- mesures des génotypes (SNPs)



(Penouilh et al., 2020)

Données simulées

100 jeux issus du même réseau
(activation/inhibition aléatoire,..)

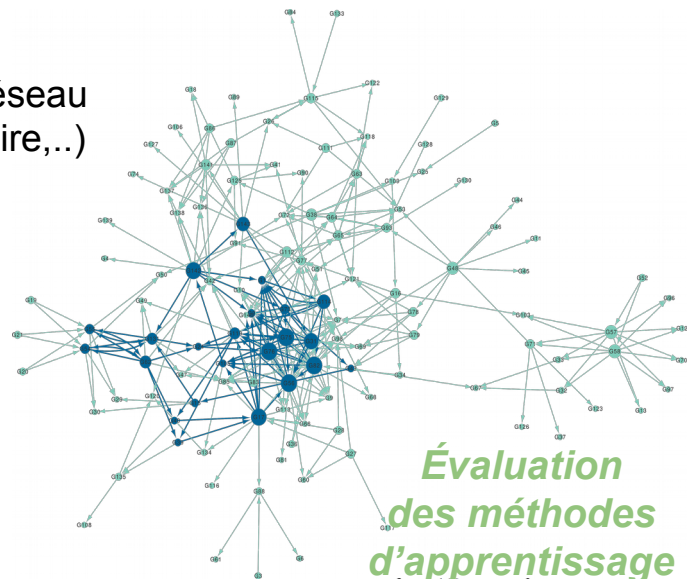
$n=463$ hybrides

$p=143$ gènes

313 arêtes

Prend en compte

- **dépendance entre individus et hétérosis**
- héritabilité
- forte connectivité
- duplication ancestrale du génome



Évaluation
des méthodes
d'apprentissage



Simulateur d'expression de gènes

SysGenSIM (*DREAM5 Systems Genetics Challenge*)

(Pinna et al., *Bioinformatics* 2011)

$$\frac{dG_g}{dt} = \underset{\substack{\text{cis-effect of gene g SNP(c)} \\ \text{on gene g expression}}}{Z_g^c} \cdot \underset{\substack{\text{basal transcription} \\ \text{rate of gene g}}}{V_g} \cdot \underset{\substack{\text{SNP(c) noise}}}{\theta_g^{syn}} \cdot \prod_k \left(1 - \underset{\substack{\text{role of gene k} \\ \text{on gene g} \\ \in \{-1; 0; 1\}}}{A_{kg}} \frac{\underset{\substack{\text{[mRNA] of} \\ \text{gene k}}}{G_k^{h_{kg}}}}{G_k^{h_{kg}} + \left(\underset{\substack{\text{trans-effect of SNP(t) of k on} \\ \text{it activity (more or less} \\ \text{efficient regulator)}}}{K_{kg}/Z_k^t} \right)^{h_{kg}}} \right) - \underset{\substack{\text{degradation rate} \\ \text{constant of gene g}}}{\lambda_g} \cdot \theta_g^{deg} \cdot G_g$$

degradation noise

[mRNA] of gene g

for each gene k

Simulateur d'expression de gènes

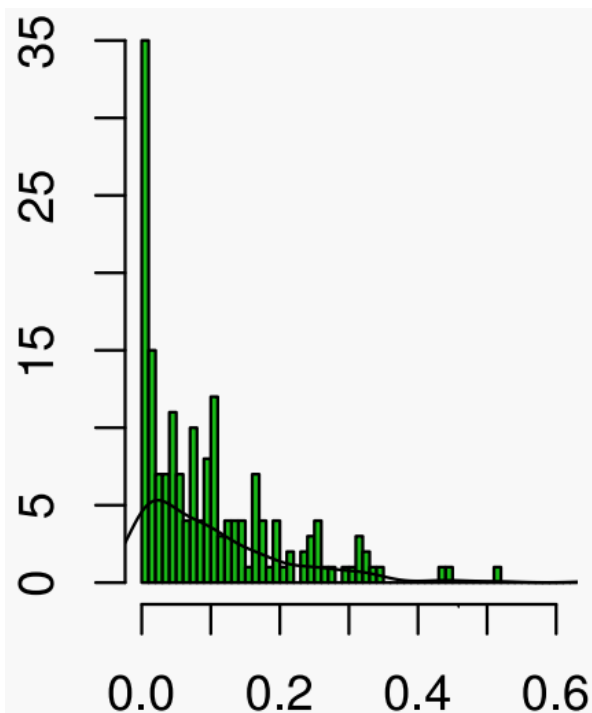
SysGenSIM (*DREAM5 Systems Genetics Challenge*)

(Pinna et al., *Bioinformatics* 2011)

$$\frac{dG_g}{dt} = \underbrace{Z_g^c}_{\substack{\text{cis-effect of gene g SNP(c) on gene g expression} \\ \uparrow \\ \text{hétérosis}}} \cdot \underbrace{V_g}_{\substack{\text{basal transcription rate of gene g}}} \cdot \underbrace{\theta_g^{syn}}_{\substack{\text{SNP(c) noise}}} \cdot \prod_k \left(1 - \underbrace{A_{kg}}_{\substack{\text{role of gene k on gene g} \\ \in \{-1; 0; 1\}}} \frac{G_k^{h_{kg}}}{\underbrace{G_k}_{\substack{\text{[mRNA] of gene k}}} + \left(\underbrace{K_{kg}/Z_k^t}_{\substack{\text{trans-effect of SNP(t) of k on its activity (more or less efficient regulator)} \\ \uparrow \\ \text{hétérosis}}} \right)^{h_{kg}}} \right) - \underbrace{\lambda_g}_{\substack{\text{degradation rate constant of gene g}}} \cdot \underbrace{\theta_g^{deg}}_{\substack{\text{degradation noise}}} \cdot \underbrace{G_g}_{\substack{\text{[mRNA] of gene g}}}$$

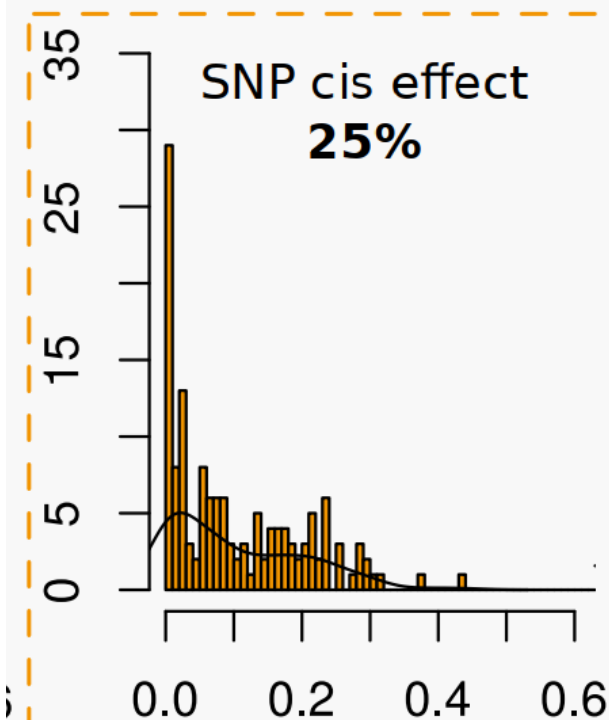
Héritabilité

% de la variance de l'expression d'un gène expliqué par les haplotypes parentaux



données réelles

heritability



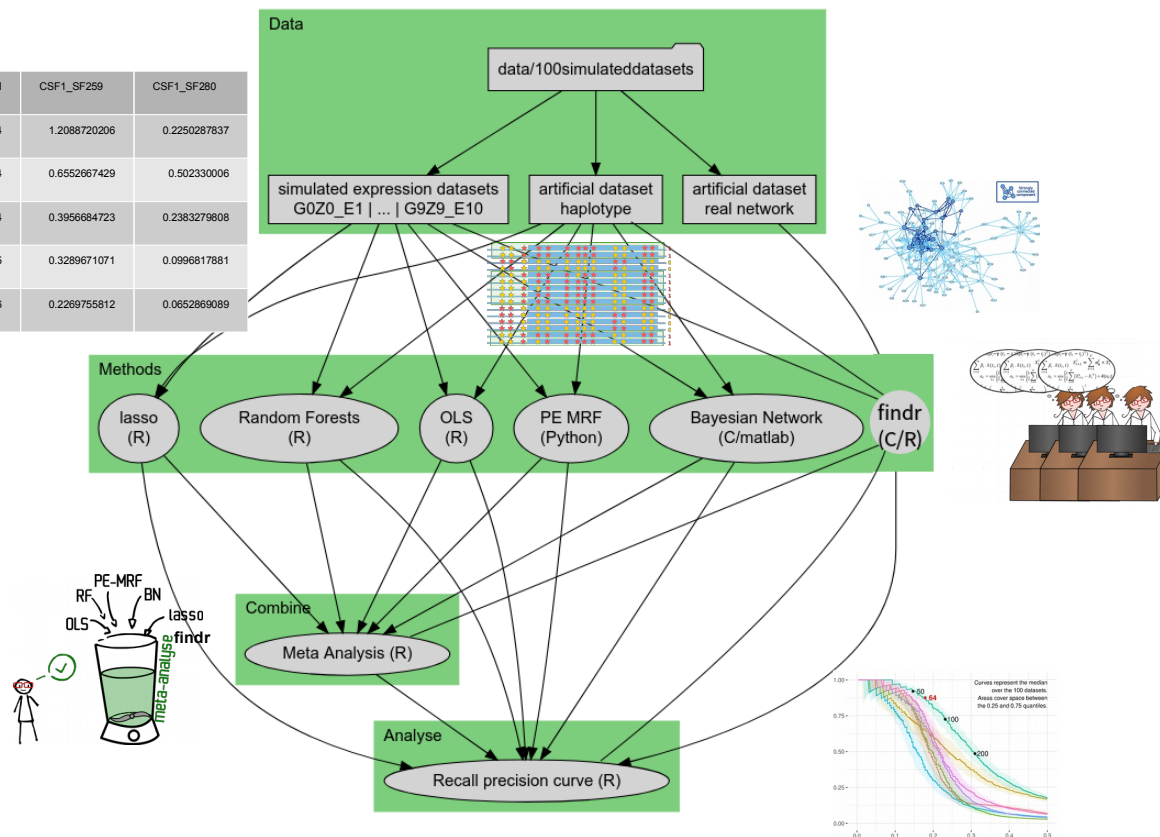
heritability

données simulées

- Contexte biologique
- **Méthodes d'apprentissage de la structure d'un réseau**
- Résultats expérimentaux
- Conclusion et perspectives

Workflow sur forgemia

	CSF1_HAS54	CSF1_PNS1	CSF1_RA36RM	CSF1_SF259	CSF1_SF280
HanXRQChr01 g0009121	0.3138026576	0.2674776948	-0.133306404	1.2088720206	0.2250287837
HanXRQChr01 g0014581	0.3917365876	0.4649870023	0.2314454684	0.6552667429	0.502330006
HanXRQChr01 g0020591	0.1927627968	0.3493922634	0.2196945394	0.3956684723	0.2383279808
HanXRQChr01 g0029831	0.0591202083	0.1515971127	0.1262509555	0.3289671071	0.0996817881
HanXRQChr01 g0030841	0.1058883472	0.1938359848	0.0966813296	0.2269755812	0.0652869089



Six méthodes d'apprentissage testées



	Modèle	Raisonnement	Score d'une arête	Code	Temps CPU	Spécificités
Findr	Tests statistiques	triplets (M_i, E_i, E_j)	probabilité a posteriori	R / C	rapide ~1min	calculs analytiques
Lasso	Régression linéaire pénalisée	gène par gène	50 bootstraps + grille de 100 valeurs du paramètre de pénalisation λ	R	rapide ~2min	
RandomForest	Arbres de régression	gène par gène	1000 bootstraps + arbres aléatoires (→ score d'importance)	R	rapide ~2min	utilisation des bootstraps pour le calcul du score d'importance
OLS	Régression par moindres carrés pénalisée	gène par gène	1 - p-value	R	lent ~2h	matrice d'apparement (dépendances entre individus)
PE-MRF	Champ de Markov (graphe non orienté)	global	50 bootstraps + 100 valeurs λ	Python	lent ~15h	modèle hybride continu/discret
BayesianNetwork	Réseau Bayésien (graphe orienté acyclique)	global	50 bootstraps + 50 valeurs λ	C++ / shell	lent 20h	méthode exacte

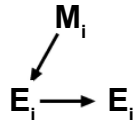
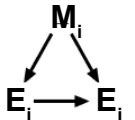
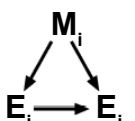
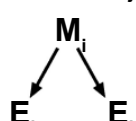
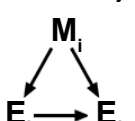
chaque
méthode
fournit une
**liste
d'arêtes
triées**



**Méta-analyse
sur les rangs**

Findr: multiple likelihood ratio tests

(Wang & Michoel, *PLOS comp. bio.* 2017)

Test ID	Test name	Null (hypothesis)		Alternative (hypothesis)	Selected hypothesis
0	Correlation	E_i	E_j	$E_i \text{ — } E_j$	Alternative
1	Primary (Linkage)	M_i	E_i	$M_i \rightarrow E_i$	Alternative
2	Secondary (Linkage)	M_i	E_j	$M_i \rightarrow E_j$	Alternative
3	(Conditional) Independence			Null	
4	Relevance			Alternative	
5	Controlled			Alternative	

- M_i : génotype du gène i
 E_i : expression du gène i
(hyp. : distributions gaussiennes)

- Probabilité a posteriori**
gène $i \rightarrow$ gène j

$$P = 0.5 (P_2 P_5 + P_4)$$

- Package R findr

Least Absolute Shrinkage and Selection Operator (Lasso)



- Régression linéaire **pénalisée** (*hyp. : distributions gaussiennes*)

$$\hat{\boldsymbol{\theta}}^{lasso} = \arg \min_{\boldsymbol{\theta}} || Y - \mathbf{X}\boldsymbol{\theta} ||_{\ell_2}^2 + \lambda || \boldsymbol{\theta} ||_{\ell_1}$$

$\mathbf{Y}$: E_j , expression du gène j , $\mathbf{X}$: expressions des autres gènes $E_{i \neq j}$ et génotypes M_i

- Grille $\boldsymbol{\Lambda}$ de 100 valeurs pour la pénalisation λ de 0 à max. (aucun régresseur sélectionné)
- Résolue avec le package R *glmnet* (LAR)

Ordinary Least Square (OLS)



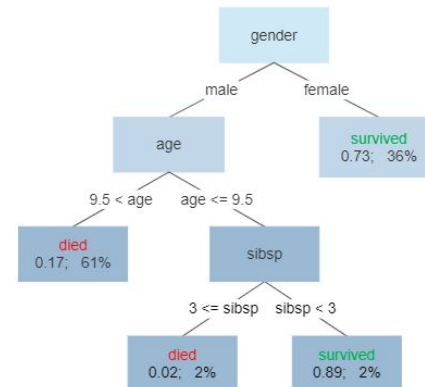
- Régression linéaire simple (*hyp.* : $p \ll n$)
- Prise en compte des dépendances entre hybrides via une **matrice d'apparentement**
- Deux modèles de régression séparés pour apprendre $M_i \rightarrow E_j$ et $E_i \rightarrow E_j$
- Résolus avec les packages R *mlmm.gwas* et *ASReml-R*

Random Forest



- Régression non-linéaire par arbres de décisions
- K = 1000 arbres de décisions avec aléas
 - **Bootstrap** : tirage aléatoire avec remise des échantillons
 - Sélection parmi 1/3 des régresseurs choisis aléatoirement
- Résolue avec le package R *randomForest*

Survival of passengers on the Titanic

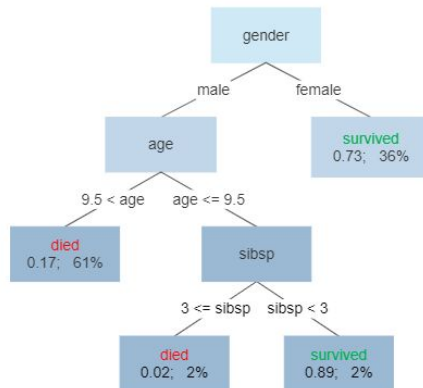


Source wikipedia

Random Forest

- Régression non-linéaire par arbres de décisions
- $K = 1000$ arbres de décisions avec aléas
 - Bootstrap : tirage aléatoire avec remise des échantillons
 - Sélection parmi 1/3 des régresseurs choisis aléatoirement
- Résolue avec le package R *randomForest*
- **Score d'importance** d'un régresseur
 - Permutation de la valeur du régresseur dans les échantillons non utilisés par le bootstrap ($n/3$)
 - Score = différence d'erreur de régression avec ou sans permutation

Survival of passengers on the Titanic



Source wikipedia

Pairwise Exponential Markov Random Field (PE-MRF)

(Park et al., *PMLR* 2017)

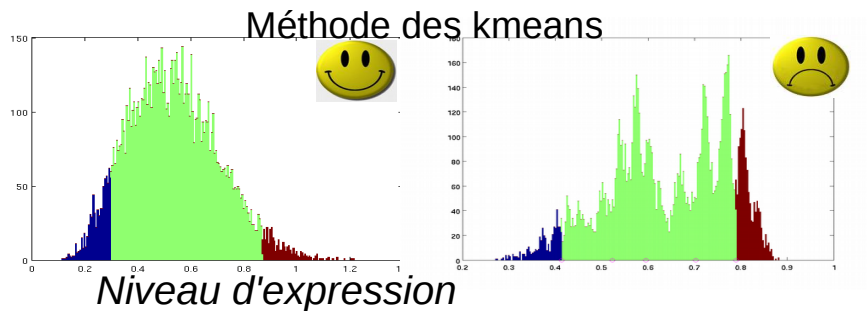
(Brouard et al., *In Proc. of Constraint Programming* 2020)



- **Champ de Markov** : modèle graphique probabiliste **non-orienté**
 - probabilité jointe factorisée par des produits de fonctions à une ou deux variables aléatoires (au plus $2p^2$ fonctions binaires)
 - modèle hétérogène : distributions gaussiennes pour les expressions et discrètes pour les génotypes
- Méthode par maximisation de la vraisemblance pénalisée
 - Approximation convexe de la fonction de partition (normalisation)
 - Grille Λ de 100 valeurs pour la pénalisation λ avec une norme ℓ_1/ℓ_2
- Résolue avec l'algorithme *Alternating Direction Method of Multipliers* (ADMM)

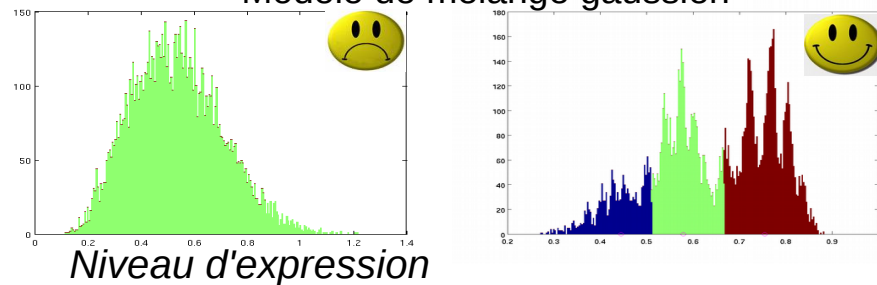
Réseau Bayésien : modèle graphique probabiliste **orienté**

- Variables aléatoires **discrètes** : discrétisation par bootstrap en au plus **3 valeurs**



+

Modèle de mélange gaussien



(Vandel, *PhD* 2012)

Réseau Bayésien : modèle graphique probabiliste **orienté**

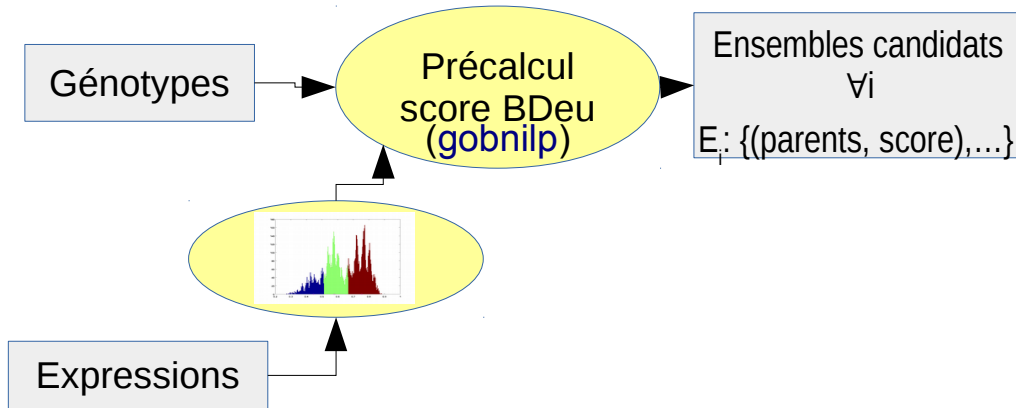
- Variables aléatoires discrètes
- Probabilité jointe factorisée par un produit de $2p$ probabilités conditionnelles
- Graphe dirigé acyclique (*Directed Acyclic Graph (DAG)*)
- Apprentissage de la structure par maximisation de la vraisemblance marginale

NP-dur

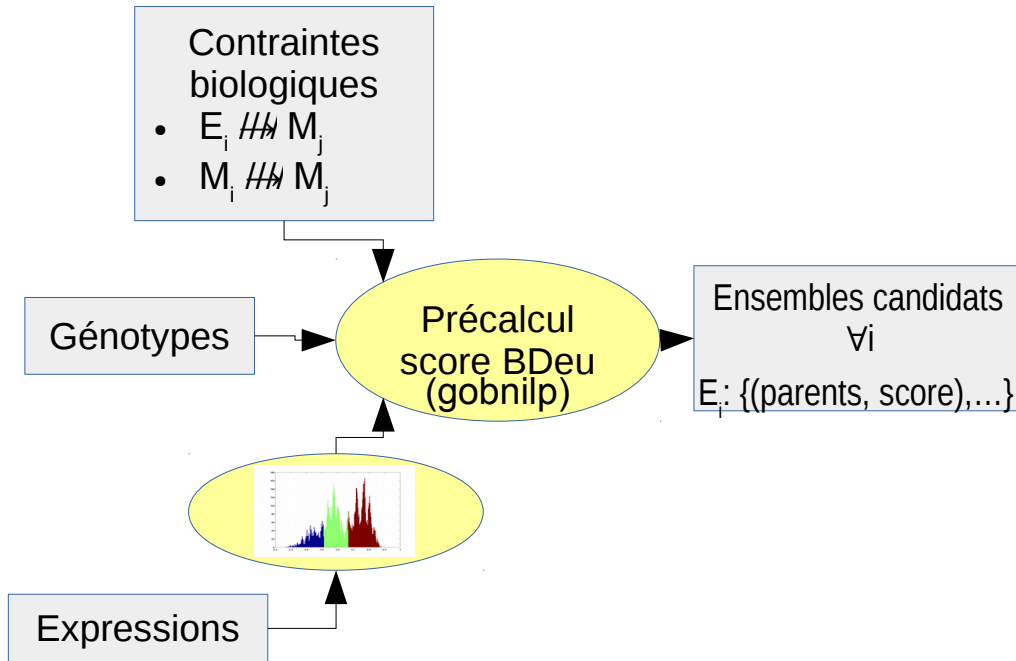
- Fonction de score globale *Bayesian Dirichlet equivalent uniform (BDeu, ESS λ)*
- Décomposable en fonctions locales par variable
 - **Précalcul des scores locaux, au pire 2^{2p-1}**
 - si $n=463$ alors au plus 4 parents (*BIC/BDeu*, Campos & Ji, *JMLR* 2011)
 - En pratique, limite fixée à **2 parents max**

($p=143$: ~1min pour 2 parents, 1h pour 3 & ~2M scores locaux à $\lambda=10^{-7}$, >100h pour 4)

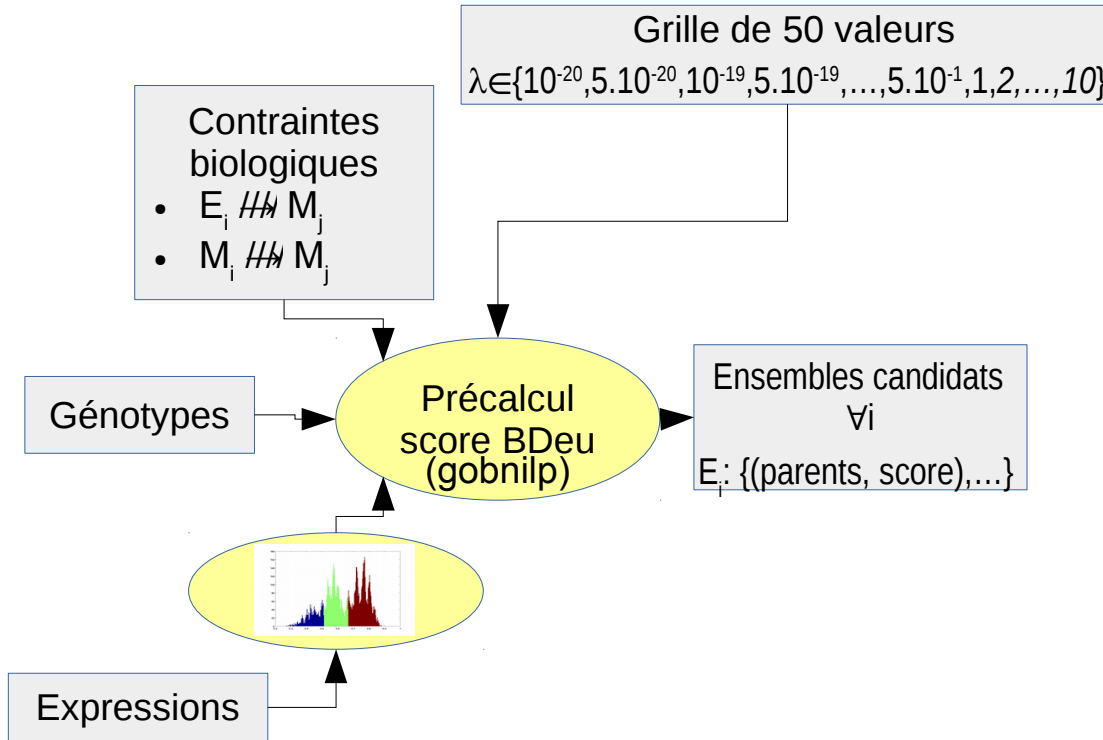
Bayesian Network



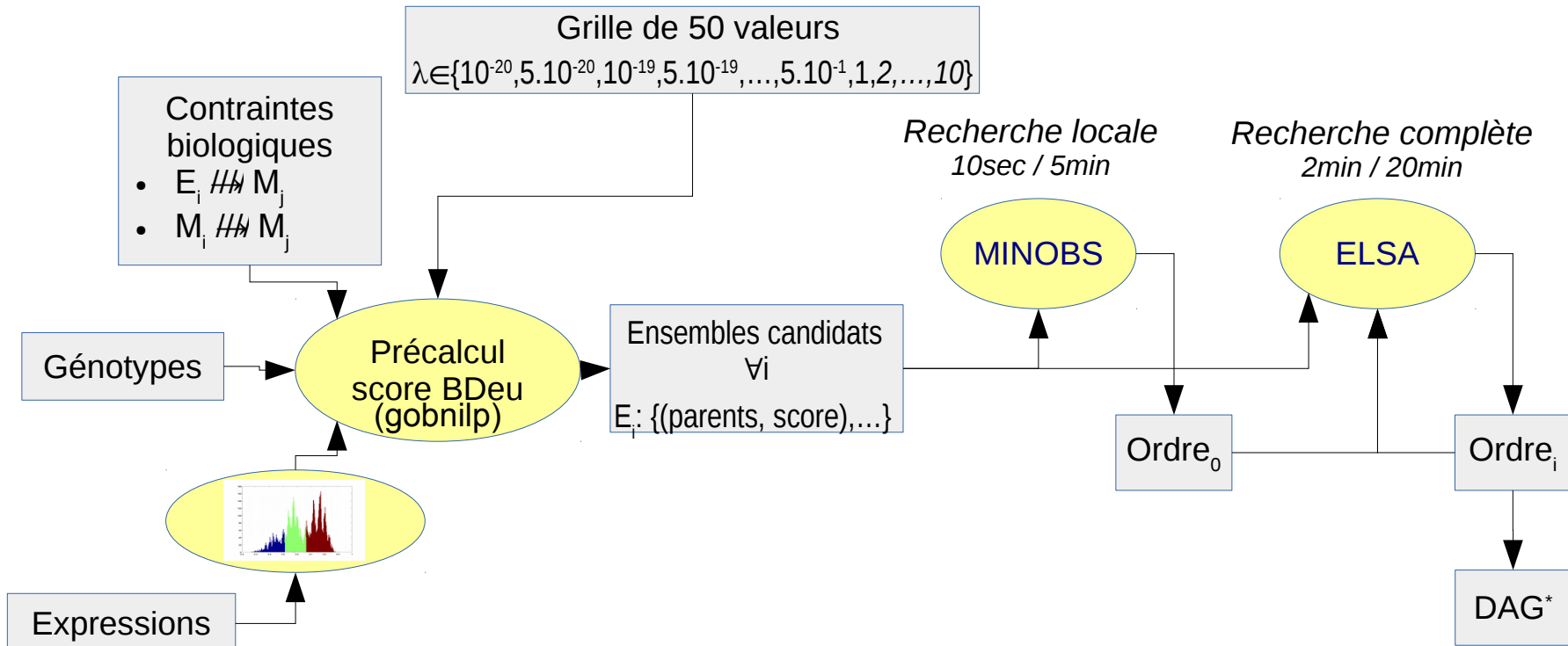
Bayesian Network



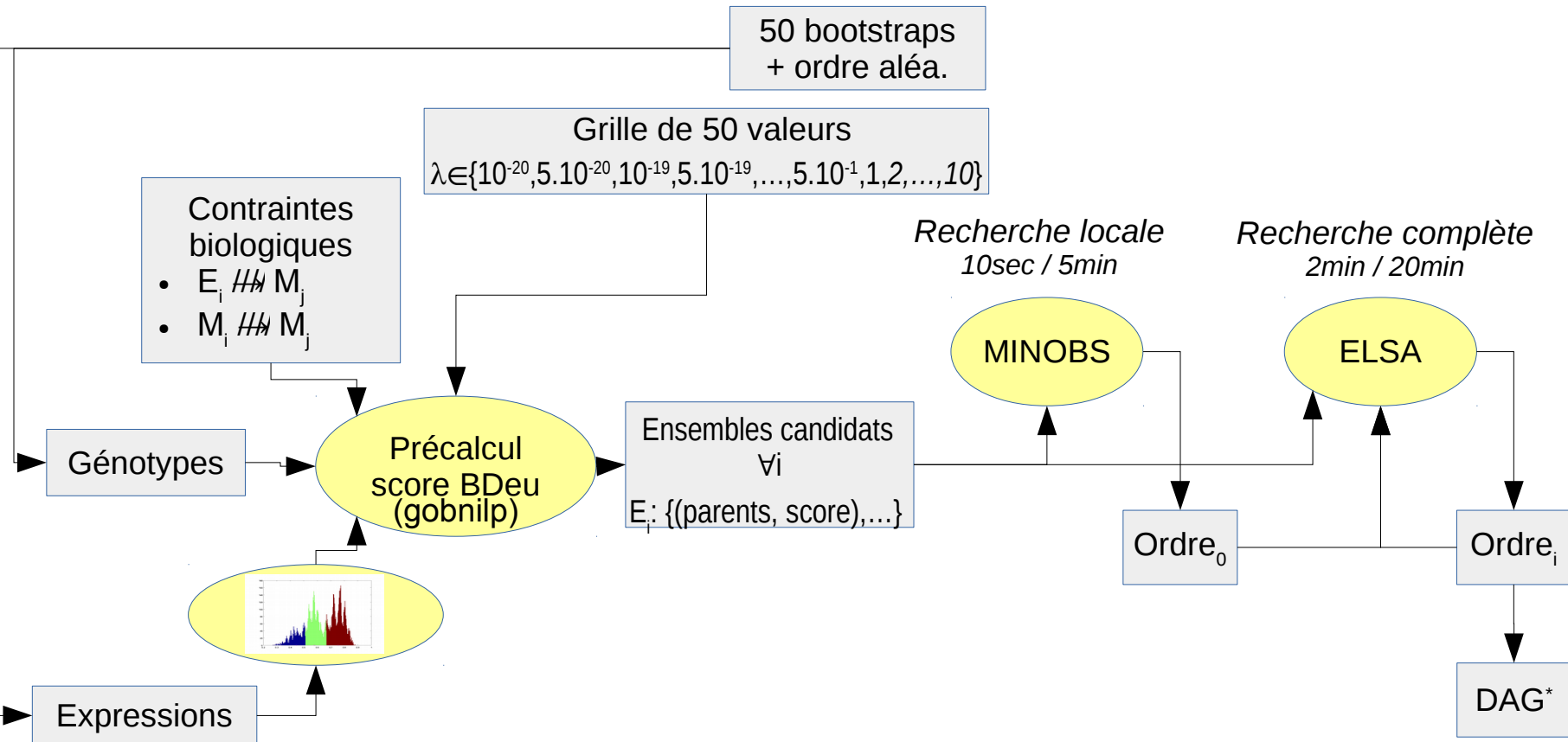
Bayesian Network



Bayesian Network

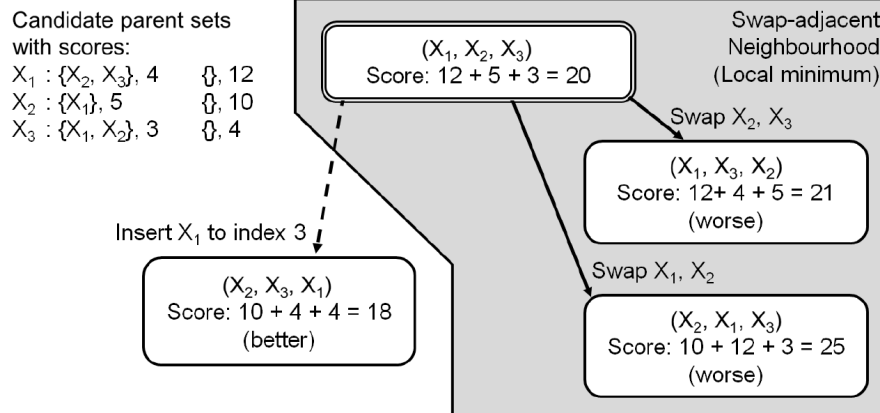


Bayesian Network



Recherche locale itérative dans l'espace des ordres

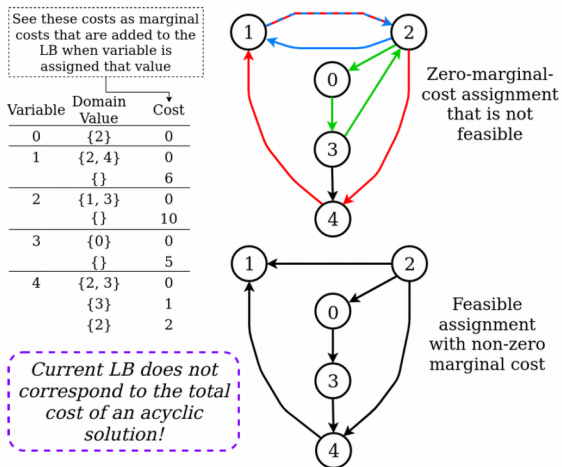
- Ordre des variables $\rightarrow$ meilleure structure compatible (facile à trouver avec scores locaux précalculés)
- Voisinage : insertion et 2-swap



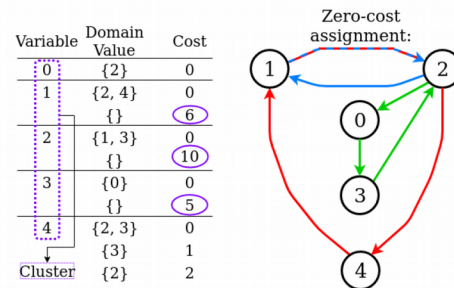
- Algorithme génétique (20 individus, hill-climbing with tabu list, order-based crossover, mutation)

Recherche complète dans l'espace des ordres

- Algorithme de séparation et évaluation
- Redémarrage à chaque meilleur ordre trouvé



Acyclicity checker using only zero-cost values returns $\{0, 1, 2, 3, 4\}$ as the violated cluster. Resulting cluster cut:

$$v_1 = \{\} \quad \vee \quad v_2 = \{\} \quad \vee \quad v_3 = \{\}$$


Transfer the minimum of $c_{v_1=\{\}} = 6$, $c_{v_2=\{\}} = 10$, $c_{v_3=\{\}} = 5$, which is 5, to the LB.

Score commensurable d'une arête

- **Bootstrap** : B = 50 jeux de données obtenus par tirage aléatoire avec remise des échantillons

$$w_{ij}^m = \frac{1}{B} \frac{1}{|\Lambda|} \sum_{b=1}^B \sum_{\lambda \in \Lambda} \alpha_{ij}^m(b, \lambda) \quad \alpha_{ij}^m(b, \lambda) = 1 \text{ si } E_i \rightarrow E_j \text{ prédite}$$

$$u_{ij}^m = \frac{1}{B} \frac{1}{|\Lambda|} \sum_{b=1}^B \sum_{\lambda \in \Lambda} \beta_{ij}^m(b, \lambda) \quad \beta_{ij}^m(b, \lambda) = 1 \text{ si } M_i \rightarrow E_j \text{ prédite}$$

$m \in \{\text{Lasso, BayesNet, PE-MRF}\}$, sauf pour

- Findr : score w = score u = probabilité a posteriori d'une arête
 - RandomForest : score = score d'importance
 - OLS : score = 1 – p-valeur
- Ranking : tri des arêtes suivant leur score (w et u) $\rightarrow$ rang moyen $r \in \{0,1\}$

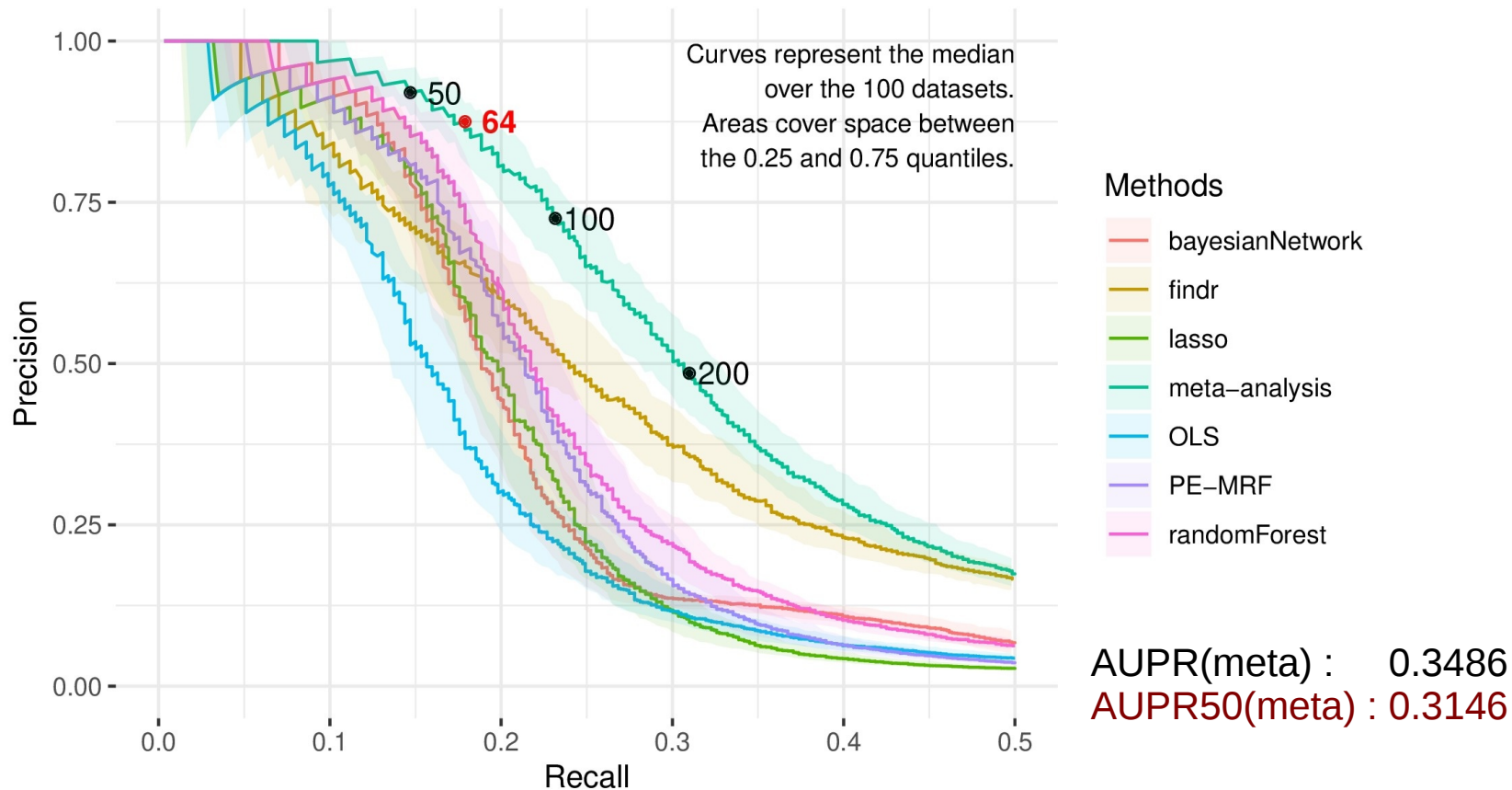
Fisher's inverse chi-square test

$$S_{ij} = 1 - \prod_{m=1}^6 (1 - \mathbf{r}_{ij}^m)$$

S_{ij} : score final de la régulation du gène i vers le gène j

- Contexte biologique
- Méthodes d'apprentissage de la structure d'un réseau
- **Résultats expérimentaux**
- Conclusion et perspectives

Qualité de l'apprentissage sur données simulées



Contribution des méthodes à la méta-analyse



AUPR50		Individual contribution to meta-analysis
	Single method	All - All except one
findr	0.2594	3.48%
randomForest	0.2377	0.91%
PE-MRF	0.2194	-1.39%
geneBayesNet	0.2142	1.3%
lasso	0.2024	-0.89%
OLS	0.1758	-0.34%
Meta-analysis		
All methods	0.3146	
All / {lasso, PE-MRF}	0.3308	
{findr, lasso, randomForest}	0.3259	

démo

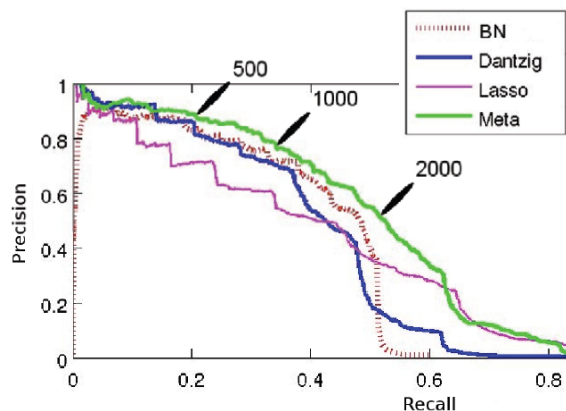
Challenge DREAM5...

and DREAM6



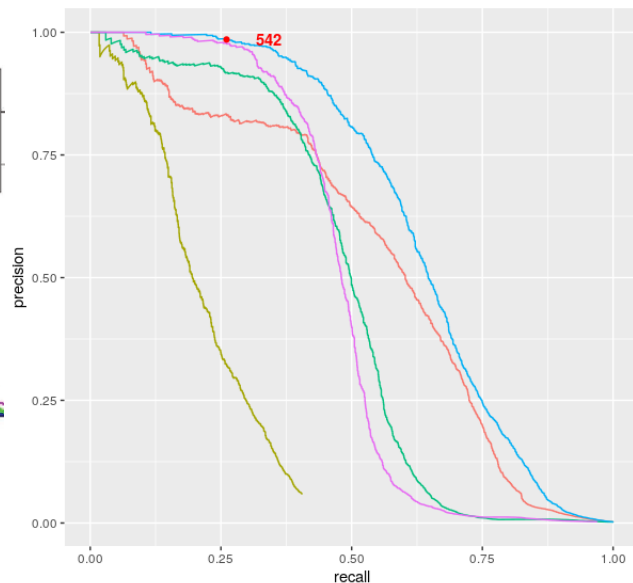
AUPR(findr): 0.547 (Wang&Michoel, 2017)

AUPR(GENIE3): 0.58
(Huynh-Thu et al., 2013)



AUPR(meta): 0.482
(Vignes et al., 2011)

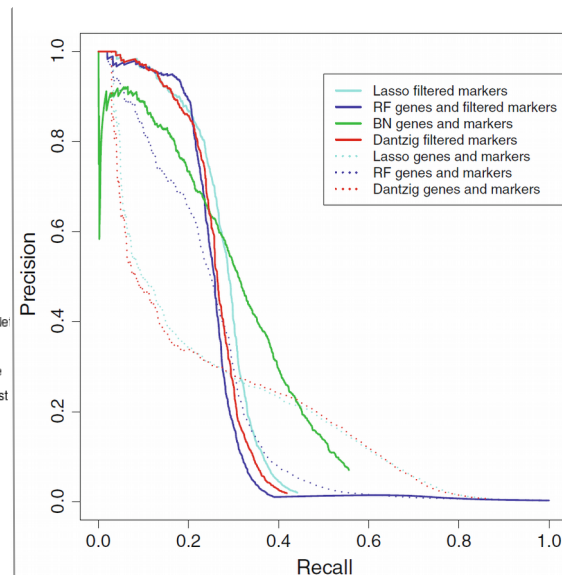
RP curves by method, on DREAM5_SysGenA999 Network 1 dataset



DREAM5
Net1

AUPR(meta): 0.65

$n=999$ indiv., $p=1000$ genes, 2048 edges



DREAM6

Net5-Conf2-DS34-900SH
900 individuals, 1000 genes

AUPR(BN): 0.305
(Allouche et al., 2014)

ELSA versus MINOBS



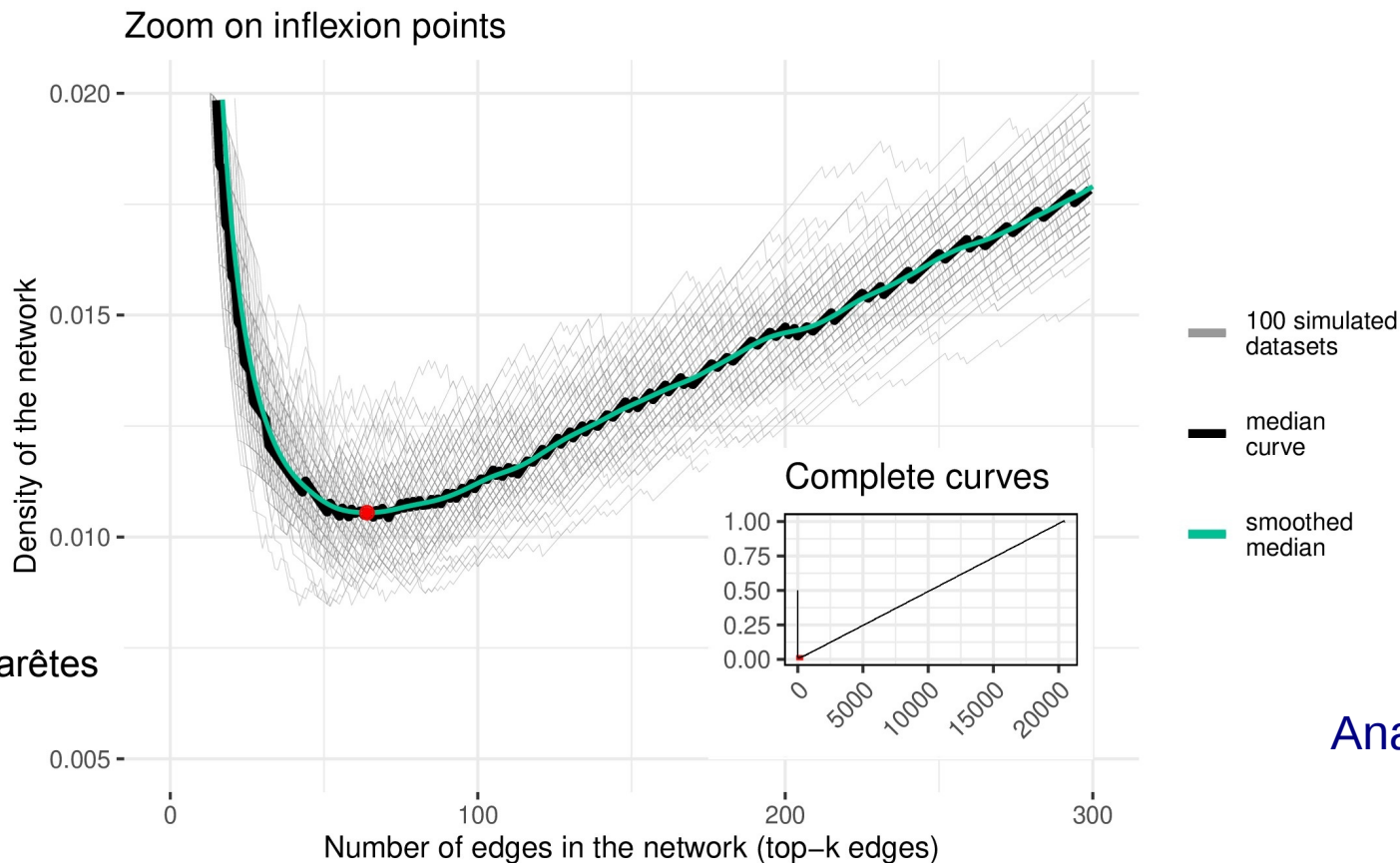
	2p variables	scores	Temps max. MINOBS/ ELSA	Temps CPU total	ELSA > MINOBS	ELSA optimum
Données simulées*	286	11K	10s / 2min	1h	23,3%	97,6%
Données réelles ⁺	346	117K	5min / 15min	20h	35,4%	75,6%
<i>Données réelles étendues^τ</i>	698	327K	<i>15min / 5min</i>	<i>2j</i>	22,4%	44,2%
DREAM5 ^τ	2000	277K	1h / -	20j (13j scores)	0 %	0%

* moyennes sur 100 * 2500 runs (Xeon 2,5GHz 20-core server)

+ moyennes sur 2250 runs (λ max = 5) (Xeon 3GHz 20-core server)

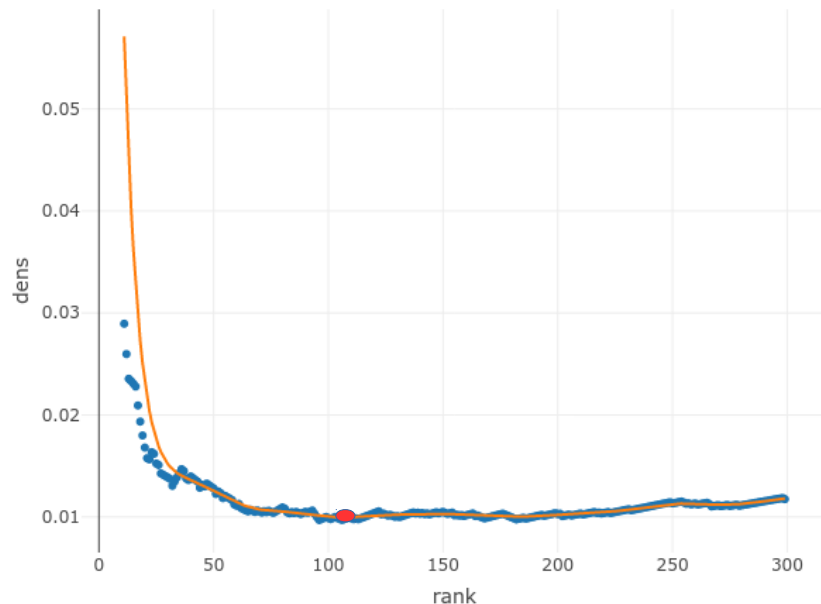
^τ moyennes sur 2250 runs (λ max = 5) (Xeon 2,5GHz 20-core server)

Sélection du nombre d'arêtes qui minimise la densité du sous-graphe obtenu à partir des sommets connectés

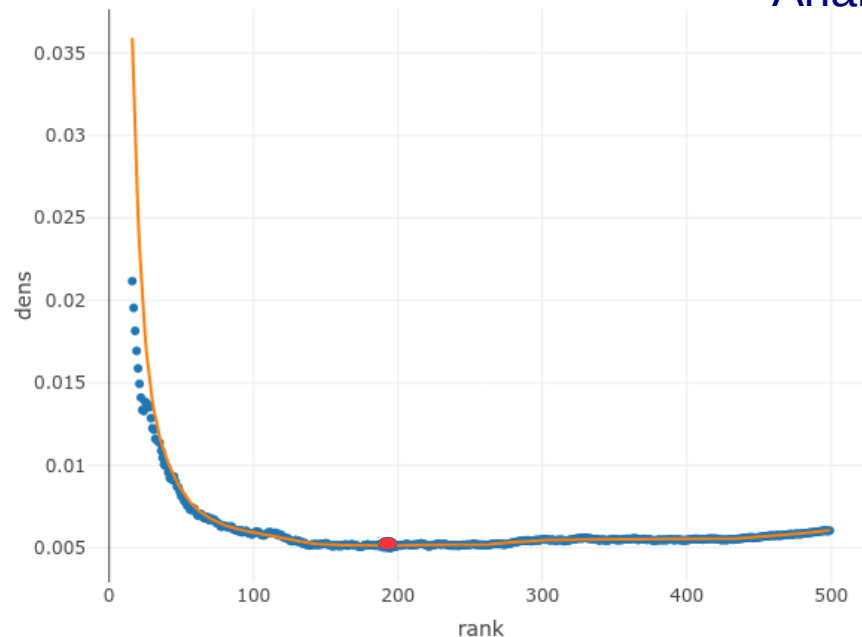


Analyse

Analyse



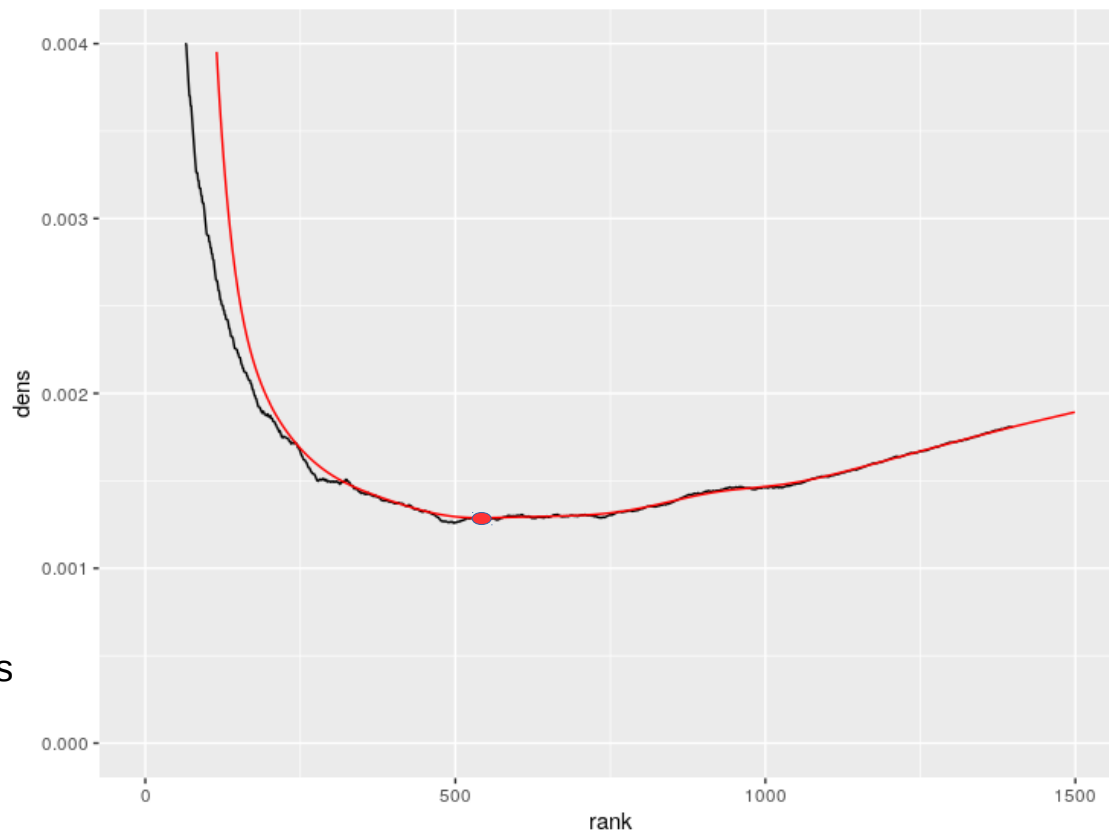
Données réelles
minimum à 106 arêtes



Données réelles étendues
minimum à 192 arêtes

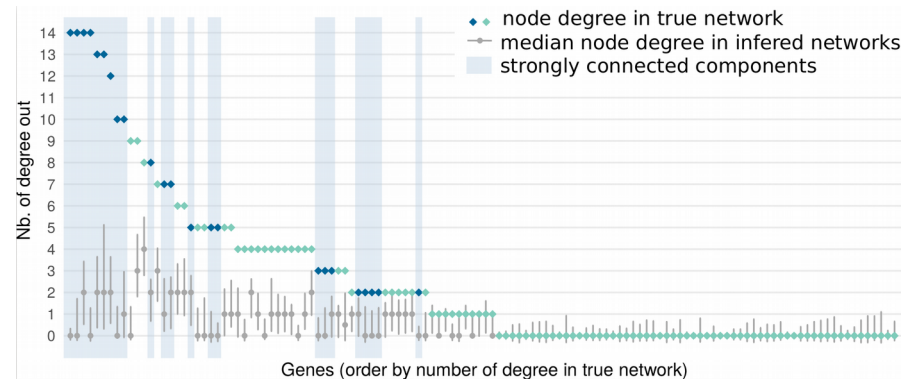
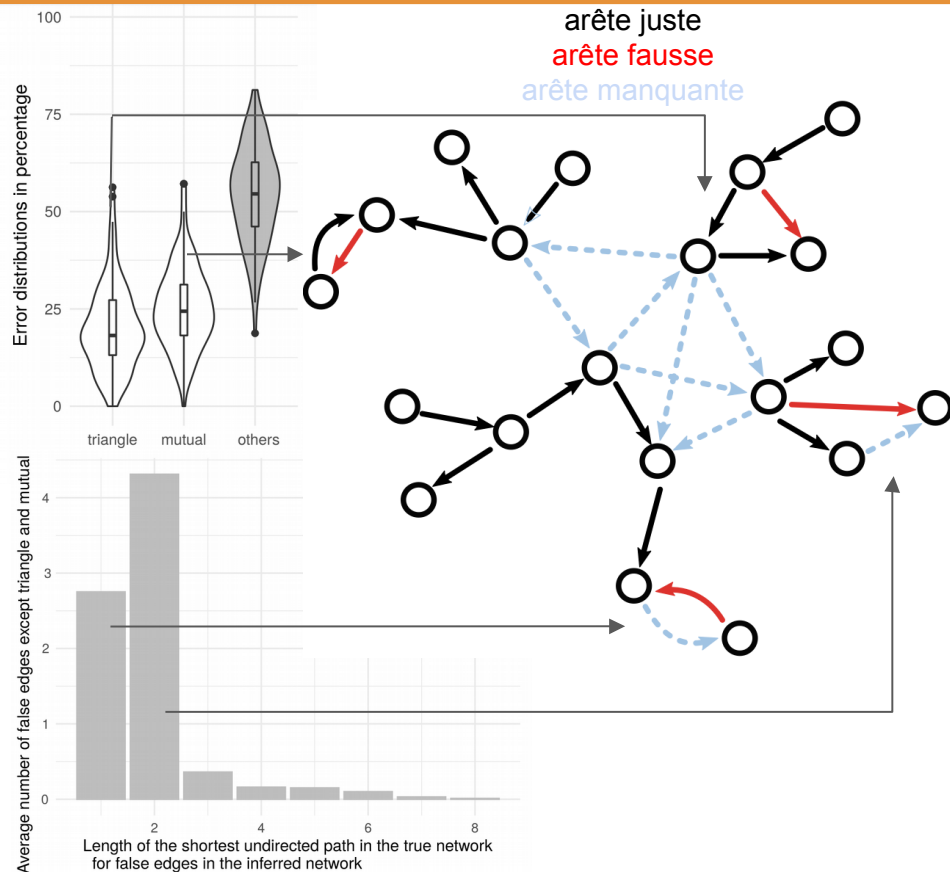
Analyse de la densité

Density of the Infered network by rank, DREAM5_SysGenA999 Network 1 dataset



minimum à 542 arêtes

Analyse des erreurs dans les réseaux appris

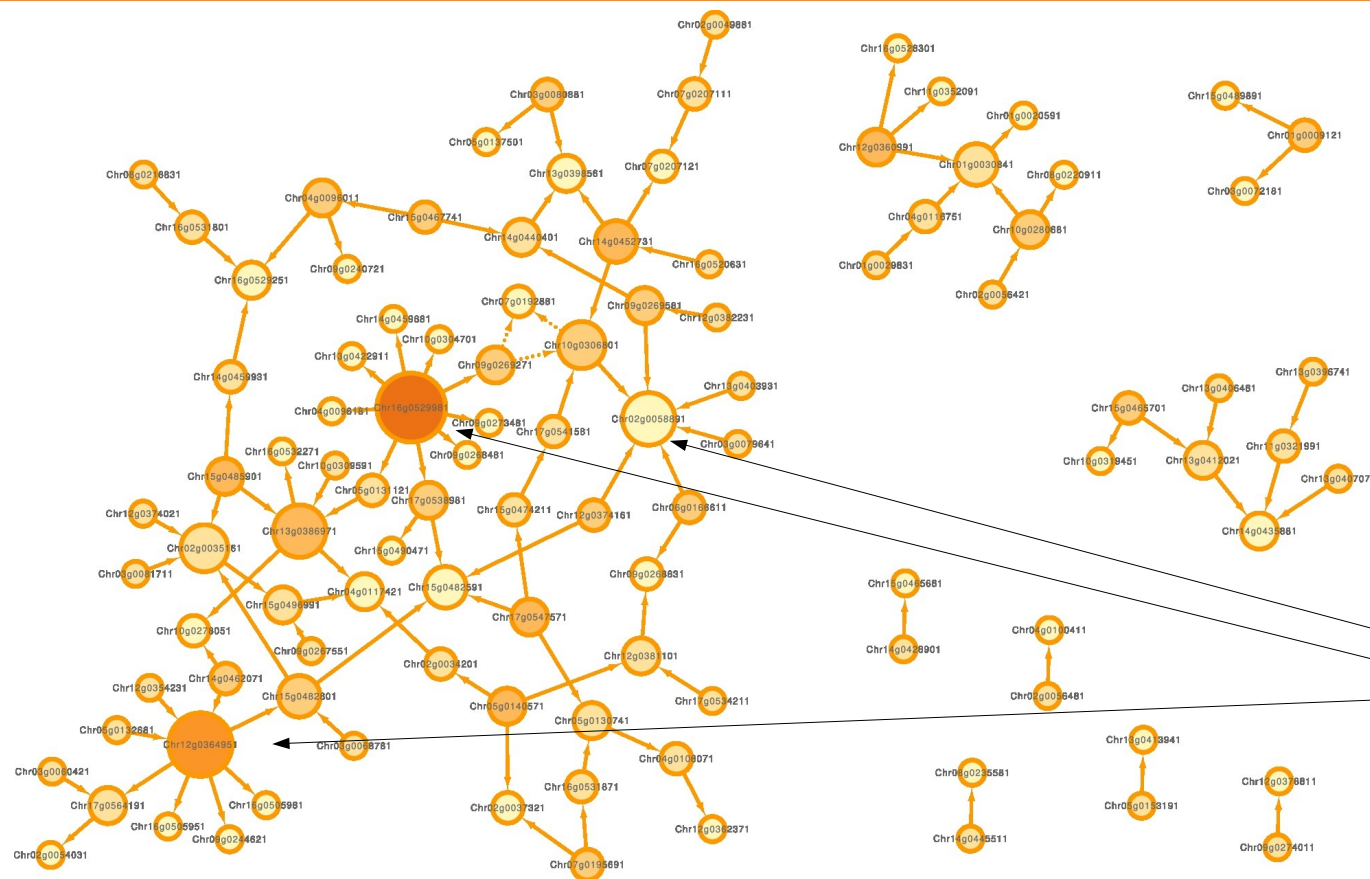


50% d'erreurs : triangles & arêtes doubles

Autres erreurs : mauvaise orientation ou manque de gène(s) intermédiaire(s)

Réseau faiblement connecté

Réseau appris sur données réelles du tournesol



105 gènes
connectés par
106 *régulations*

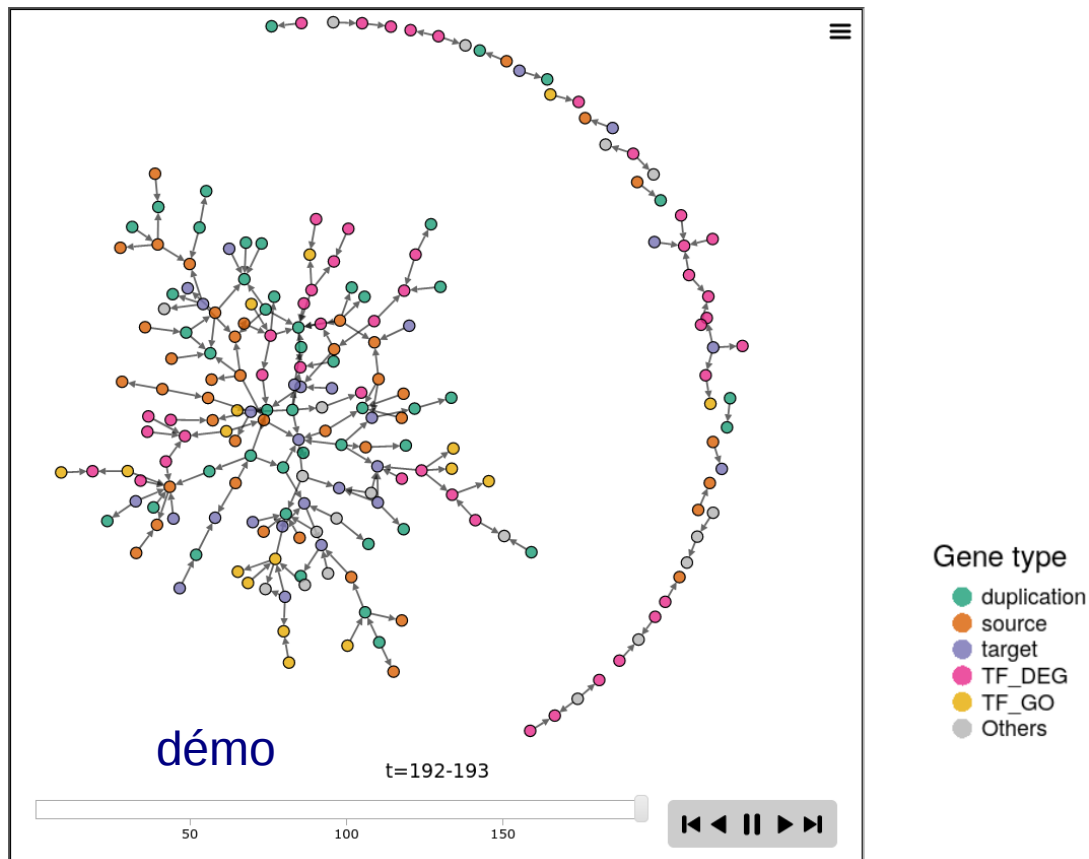
3 hubs liés au
stress abiotique
dans l'arabette

Nouvelle liste de 180 gènes à partir du réseau précédent

Sélection des nouveaux gènes

- sources potentielles du réseau
- cibles potentielles du réseau
- paralogues des gènes du réseau

démo (données 2018)



- Contexte biologique
- Méthodes d'apprentissage de la structure d'un réseau
- Résultats expérimentaux
- **Conclusion et perspectives**

- Modèles graphiques probabilistes
 - Faible amélioration sur les données simulées
 - Probabilité d'une arête versus probabilité d'un graphe (bootstrap, grille λ ?)
 - Méthode exacte dans 75% des cas sur notre jeu de données réelles
 - Limitation de l'approche à 2 parents max. (temps/espace du précalcul)
- Réseau de régulation du tournesol
 - Large composante connexe mais reste faiblement connecté (proche d'un arbre)
 - Orientation des arêtes grâce à la génétique (génotypes)
 - Analyse et validation biologique sur le jeu de données étendues

References



- D. Allouche, C. Cierco-Ayrolles, S. de Givry, G. Guillermin, B. Mangin, T. Schiex, J. Vandel, and M. Vignes. A panel of learning methods for the reconstruction of gene regulatory networks in a systems genetics context. In *Gene Network Inference*, pages 9–31. Springer, 2013.
- H Badouin, et al. The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. *Nature* 546.7656 : 148-152, 2017.
- C Brouard, S de Givry, and T Schiex. Pushing data in cp models using graphical model learning and solving. In Proc. of *CP-20*, p. 881-827, Louvain-la-neuve, Belgium, 2020.
- de Campos, C. P., & Ji, Q. Efficient structure learning of Bayesian networks using constraints. *Journal of Machine Learning Research*, 12, 663-689, 2011.
- J. H. Friedman, T. Hastie, and R. Tibshirani. glmnet: lasso and elastic-net regularized generalized linear models. R package, 2010.
- L. Gody, H. Duruflé, N. Blanchet, C. Carré, L. Legrand, B. Mayjonade, S. Muñoz, L. Pomiès, S. de Givry, N. B. Langlade, et al. Transcriptomic data of leaves from eight sunflower lines and their sixteen hybrids under water deficit. *OCL Oilseeds and fats crops and lipids*, 27:48, 2020.
- A Henry, F Monéger, A Samal and O C. Martin. Network function shapes network structure: the case of the Arabidopsis flower organ specification genetic network. *Molecular BioSystems*, 9, 1726-1735, 2013.
- V. A. Huynh-Thu, L. Wehenkel, and P. Geurts. Gene regulatory network inference from systems genetics data using tree-based methods. In *Gene Network Inference*, pages 63–85. Springer, 2013.
- Jansen, R C., and J-P Nap. Genetical genomics: the added value from segregation. *TRENDS in Genetics* 17.7: 388-391, 2001.
- C Lee and P van Beek. Metaheuristics for Score-and-Search Bayesian Network Structure Learning. *Proceedings of the 30th Canadian Conference on Artificial Intelligence*, Edmonton, Alberta, May, 2017.
- Y. Park, D. Hallac, S. Boyd, and J. Leskovec. Learning the network structure of heterogeneous data via pairwise exponential markov random fields. In *Artificial Intelligence and Statistics*, pages 1302–1310. *PMLR*, 2017.
- C. Penouilh-Suzette, L. Pomiès, H. Duruflé, N. Blanchet, F. Bonnafous, R. Dinis, C. Brouard, L. Gody, C. J. Grassa, X. Heudelot, M. Laporte, M. Larroque, G. Marage, B. Mayjonade, B. Mangin, S. de Givry, and N. B. Langlade. RNA expression dataset of 384 sunflower hybrids in field condition. *OCL Oilseeds & fats crops & lipids*, 27:36, 2020.
- Pinna, Andrea, et al. Simulating systems genetics data with SysGenSIM. *Bioinformatics* 27.17: 2459-2462, 2011.
- L Pomiès, C Brouard, H Duruflé, E Maigné, C Carré, L Gody, F Trösser, G Katsirelos, B Mangin, N B. Langlade, and S de Givry. Gene Regulatory Network Inference Methodology for Genomic and Transcriptomic Data Acquired in Genetically Related Heterozygote Individuals. *Submitted*, 2021.
- F. Trösser, S. de Givry, and G. Katsirelos. Improved acyclicity reasoning for bayesian network structure learning. In Proc. of *IJCAI-21*, Montreal, Canada, 2021.
- J Vandel. Learning gene regulatory networks in genetical genomics experiments. *PhD*, INRA Toulouse, 2012.
- M. Vignes, J. Vandel, D. Allouche, N. Ramadan, C. Cierco, T. Schiex, B. Mangin, and S. de Givry. Gene Regulatory Network Reconstruction Using Bayesian Networks, the Dantzig Selector, the Lasso and Their Meta-Analysis. *PLoS ONE*, 6(12), 2011.
- L. Wang and T. Michael. Efficient and accurate causal inference with hidden confounders from genome-transcriptome variation data. *PLoS comp. biol.*, 13(8):e1005703, 2017.

Remerciements



MIA-Paris

Lise Pomiès, Céline Brouard, Auxanne Calmont, Clément Carré, George Katsirelos, Elise Maigné, Fulya Trösser



Harold Duruflé, Louise Gody, Caroline Laffray, Brigitte Mangin, Charlotte Penouilh, Marco Moroldo, Nicolas Langlade

expression
moyenne du
gène

effet aléatoire
male

$$y_{fm} = \mu + GCA_f + GCA_m + \epsilon_{fm} \quad (\text{GCA model})$$

$$y_{fm} = \mu + F_f + M_m + \epsilon_{fm} \quad (\text{FM model})$$

$$y_{fm} = \mu + F_f + M_m + I_{fm} + \epsilon_{fm} \quad (\text{FMI model})$$

expression du
gène pour
l'hybride *fm*

effet aléatoire
femelle

effet aléatoire
interaction