Sélection génomique, WGDs, Réseaux

Charles-Elie Rabier

ISEM, Institut des Sciences de l'Evolution de Montpellier
LIRMM, Laboratoire d'informatique, de Robotique et de Microélectronique

Première partie : Sélection Génomique

Prédiction en sélection génomique

Brigitte Mangin

LIPM, Laboratoire des Interactions Plantes et Microorganismes, Toulouse

Simona Grusea

Institut National des Sciences Appliquées de Toulouse

Plan

Plan

Introduction

Sélection Génomique

- Objectif = prédire les valeurs génétiques des candidats à la sélection
- Plus besoin de détecter les QTLs ! ! !

Nouvelles technologies de sequençage

- Milliers de marqueurs disponibles $\Rightarrow$ tous les QTLs fortement corrélés (fort Déséquilibre de liaison) avec au moins un marqueur

Outils statistiques

- Tous les marqueurs analysés simultanément $\Rightarrow$ Régression sur tout le génome

Problème de haute dimension

$K >> n$

Genomic Selection

Training population

$$Y = \mu + \underbrace{f(M, \theta)}_{g} + \varepsilon$$

← Model

$$\hat{\theta}$$

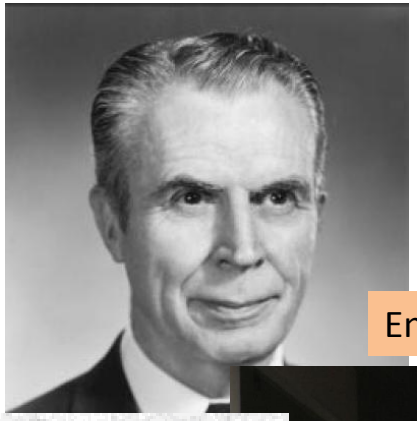
← Estimation method

Test population (selection candidates)

$$\underbrace{f(M, \hat{\theta})}_{\hat{g}}$$

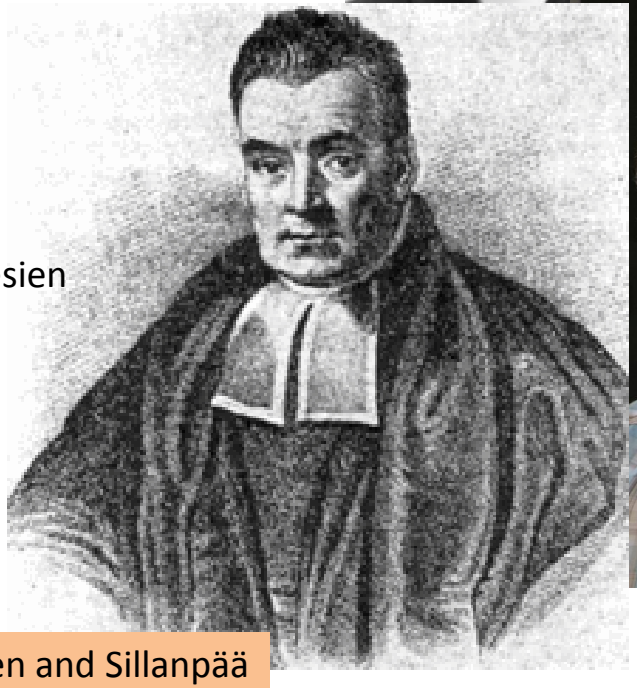
← Prediction

Cadres statistiques

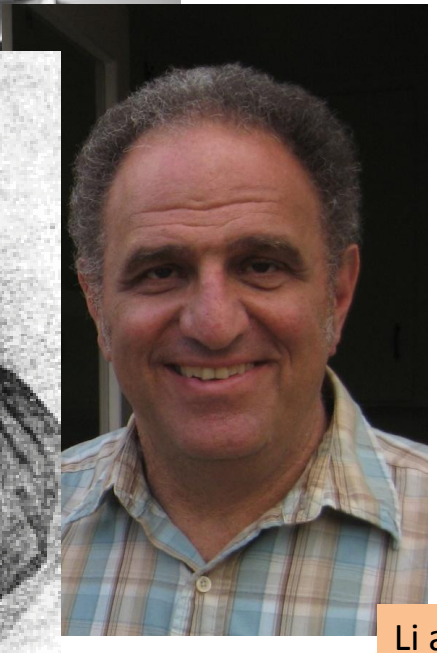


C.R. Henderson
Modèle mixte

Endelman



T. Bayes
Modèle bayésien



R. Tibshirabi
Régression pénalisée

Li and Sillanpää

Kärkkäinen and Sillanpää

Classement

- En général, le classement des méthodes est Bayes $\geq$ Régressions pénalisées $>$ Modèle mixte
- les méthodes ont moins d'influence que la densité de marquage, la taille de la population Training, l'héritabilité
- ou que la "distance" entre la population Training et la population TEST

Cet exposé : focus sur GBLUP, RRBLUP, Ridge (Pénalité L2)

Modèle causal vs modèle de prédiction

Objectif : Prédiction en grande dimension

Modèle causal (Q QTLs)

$\theta^\star$ effets QTLs, $M^\star$ matrice de génotypes aux QTLs,
n individus d'entrainement

$$Y = M^\star \theta^\star + e$$

où $Y = (Y_1, ..., Y_n)'$, $\theta^\star = (\theta_1^\star, ..., \theta_Q^\star)'$, $e \sim N(0, \sigma_e^2 I_n)$

Modèle Bayésien de prédiction (K marqueurs, où $K >> n$)

θ effets marqueurs, M matrice de génotypes aux marqueurs des "Training"

$$Y = M\theta + \varepsilon$$

où $Y = (Y_1, ..., Y_n)'$, $\theta = (\theta_1, ..., \theta_K)' \sim N(0, \sigma_\theta^2 I_K)$, $\varepsilon \sim N(0, \sigma_\varepsilon^2 I_n)$, $\varepsilon_j \perp\!\!\!\perp \theta_k$

Apprentissage $\hat{\theta} = \mathbb{E}(\theta \mid Y) = M' (MM' + \lambda I_n)^{-1} Y = (M'M + \lambda I_K)^{-1} M' Y$

Régression Ridge (Pénalité L2) avec $\lambda = \sigma_\varepsilon^2 / \sigma_\theta^2$

Introduction d'un individu TEST

- Soit un individu TEST numéroté $n+1$

$$Y_{n+1} = M^{\star}_{n+1}\ \theta^{\star} + e_{n+1} \quad \text{où} \quad e_{n+1} \sim N(0, \sigma^2_e)$$
$$\text{et} \quad M^{\star}_{n+1} \text{ génotypes aux QTLs de l'individu } n+1$$

- Prédiction de la valeur phénotypique

$$\hat{Y}_{n+1} = M_{n+1}\hat{\theta} = M_{n+1}M'\left(MM' + \lambda I_n\right)^{-1} Y$$
$$= M_{n+1}\left(M'M + \lambda I_K\right)^{-1} M'Y$$

$\Rightarrow$ Critère d'accuracy

$$\rho = \frac{\text{Cov}\left(\hat{Y}_{n+1},\ Y_{n+1}\right)}{\sqrt{\text{Var}\left(\hat{Y}_{n+1}\right)\text{Var}\left(Y_{n+1}\right)}} \quad \text{avec } M_{n+1} \text{ aléatoire et } M \text{ fixe}$$

Plan

Formule générale pour l'accuracy

Théorème (R., Barre, ..., Mangin, Plos One 2016)

Conditionnellement à M et $M^\star$,

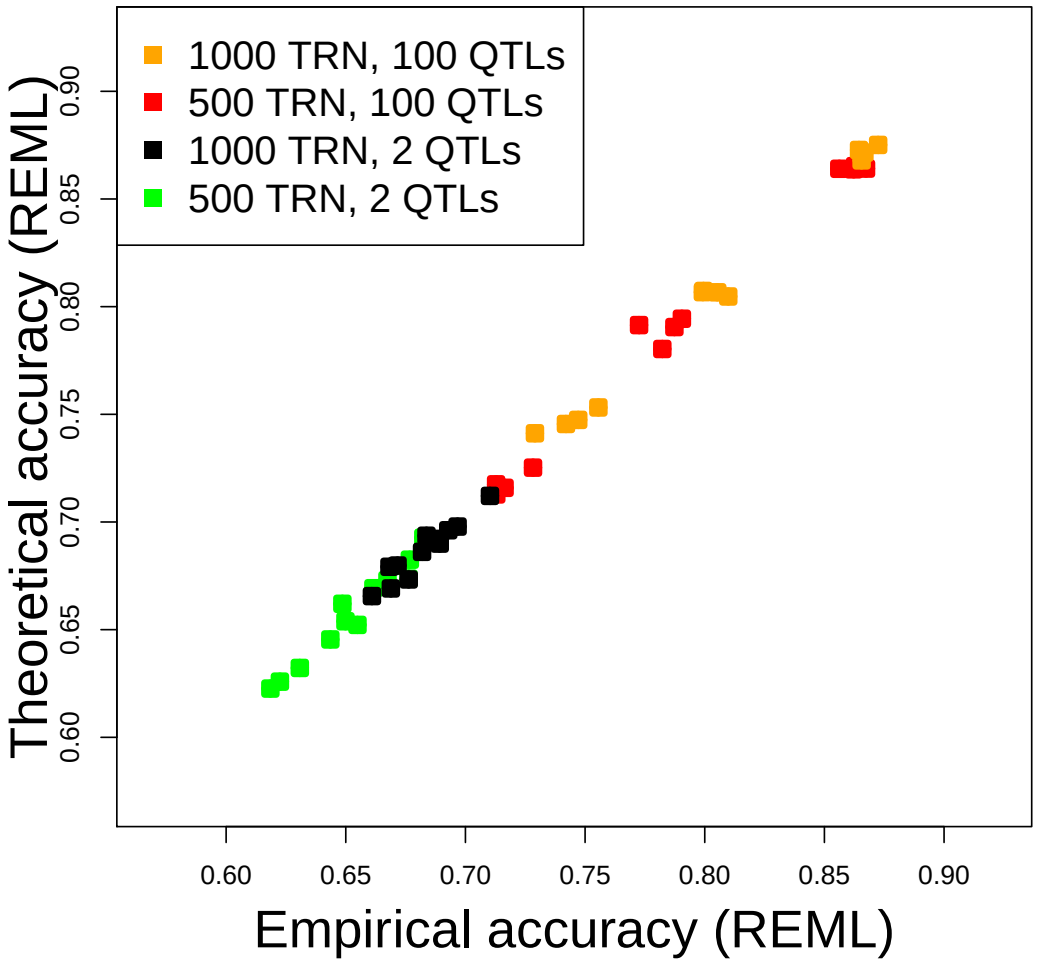
$$\rho = \frac{\theta^{\star\prime}\ \mathbb{E}\left(M_{n+1}^{\star\prime} M_{n+1}\right) M' V^{-1} M^\star \theta^\star}{\left\{\sigma_e^2 \mathbb{E}\left(\left\|M_{n+1} M' V^{-1}\right\|^2\right) + \theta^{\star\prime} M^{\star\prime} V^{-1} M \mathrm{Var}\left(M'_{n+1}\right) M' V^{-1} M^\star \theta^\star\right\}^{1/2} \Omega^{1/2}}$$

où $V = MM' + \lambda I_n$ et $\Omega = \mathrm{Var}\left(M_{n+1}^\star \theta^\star\right) + \sigma_e^2$

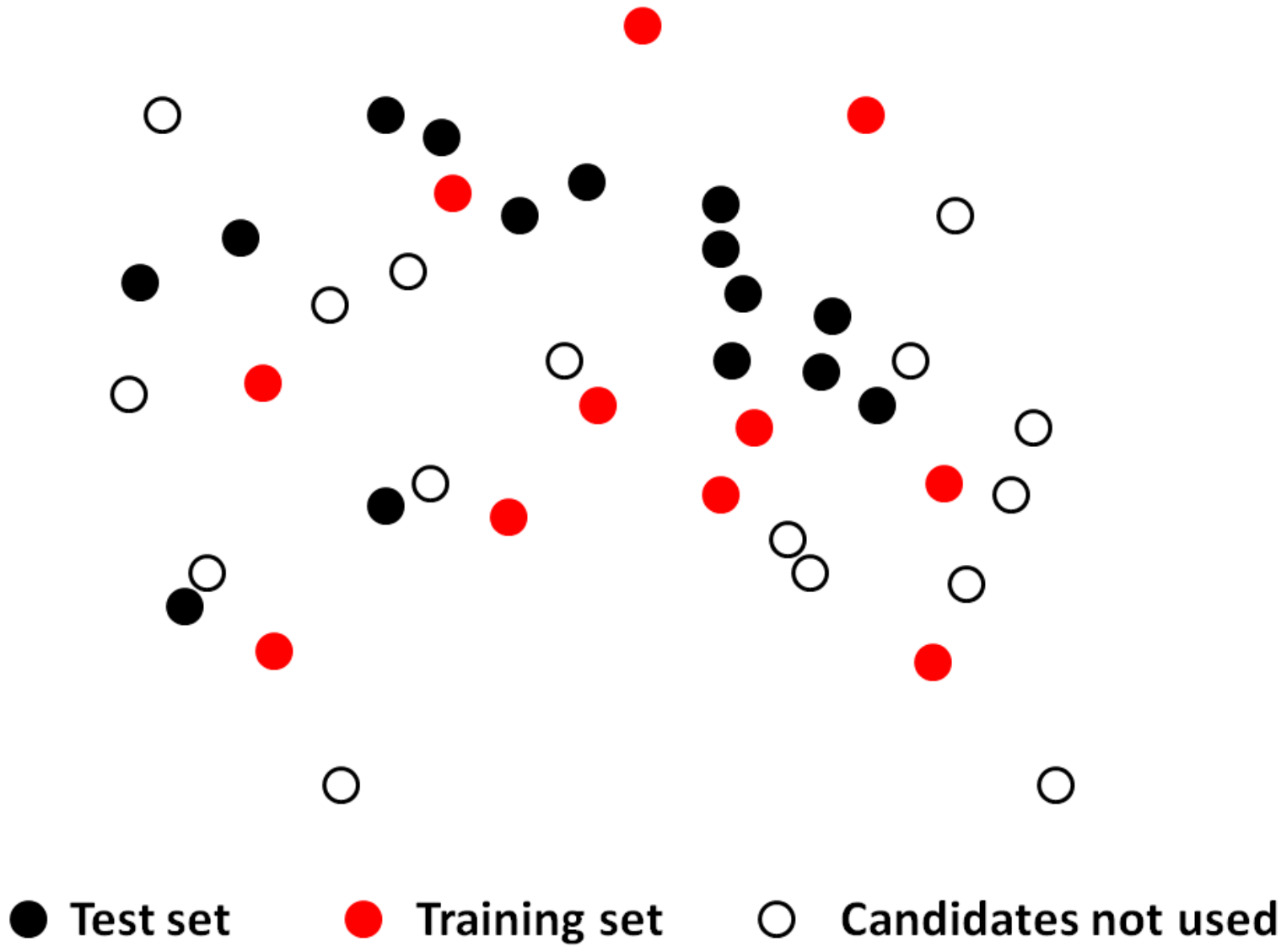
Facteurs agissant sur l'accuracy :

- Colonne q de $M' V^{-1} M^\star$: DL (utilisant la métrique V) entre chaque marqueur et le QTL q
- $\mathbb{E}\left(\left\|M_{n+1} M' V^{-1}\right\|^2\right)$: similarité entre TRN et TEST
- $\mathrm{Var}\left(M'_{n+1}\right)$: matrice de covariance de taille $K \times K$, contenant l'ensemble des DL classiques du TEST

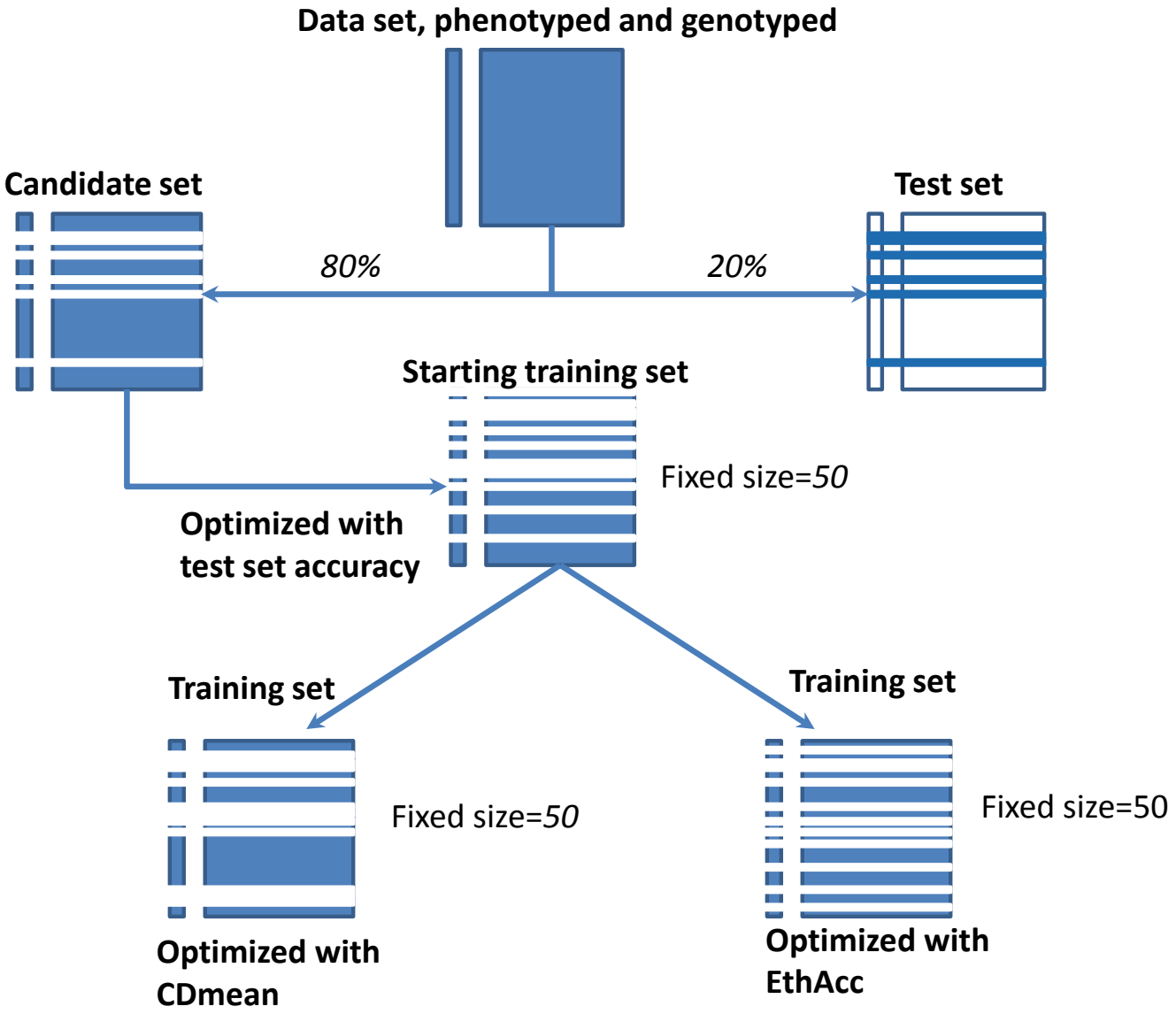
Les QTLs sont situés sur quelques marqueurs (DL parfait)



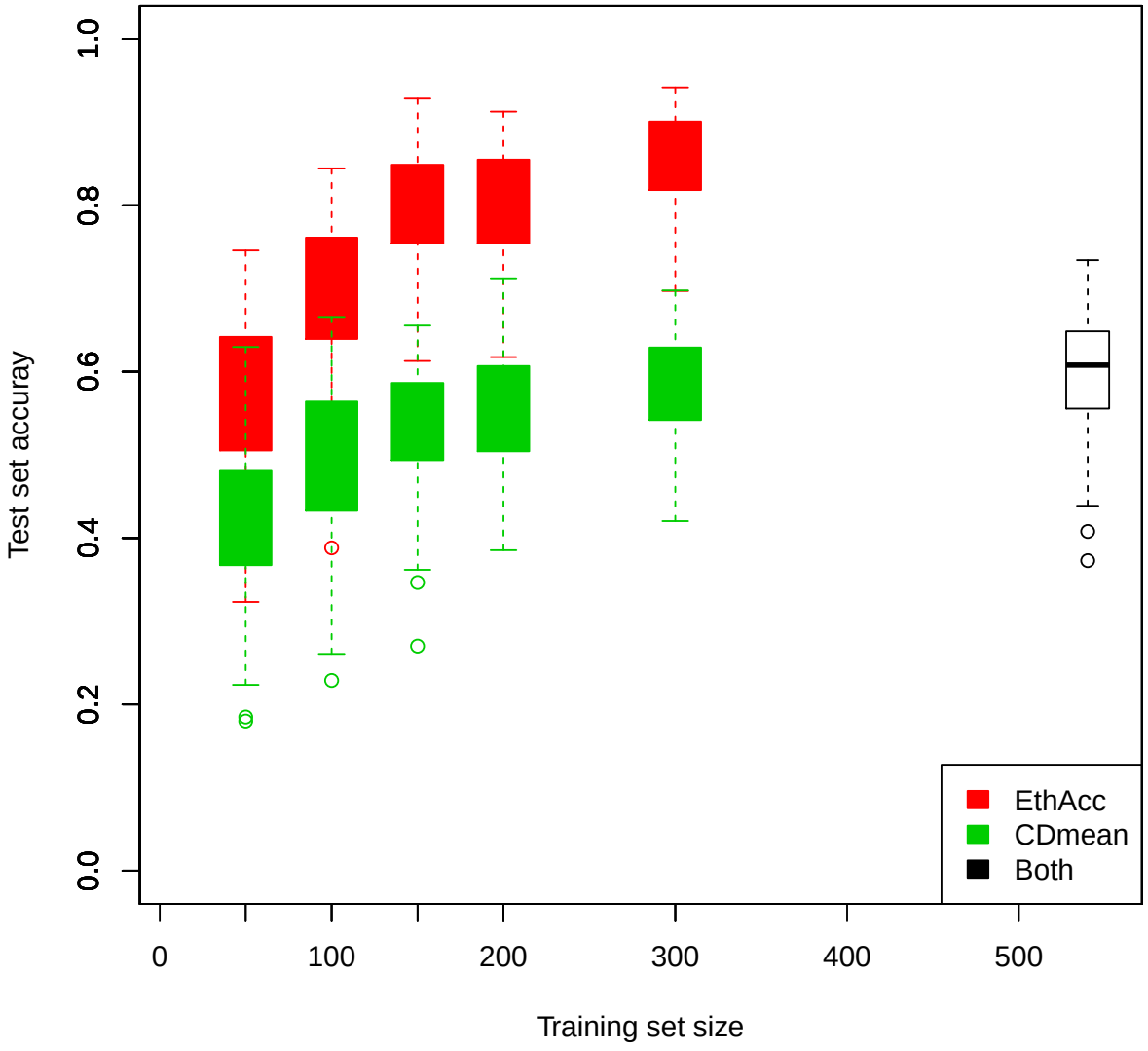
Plan

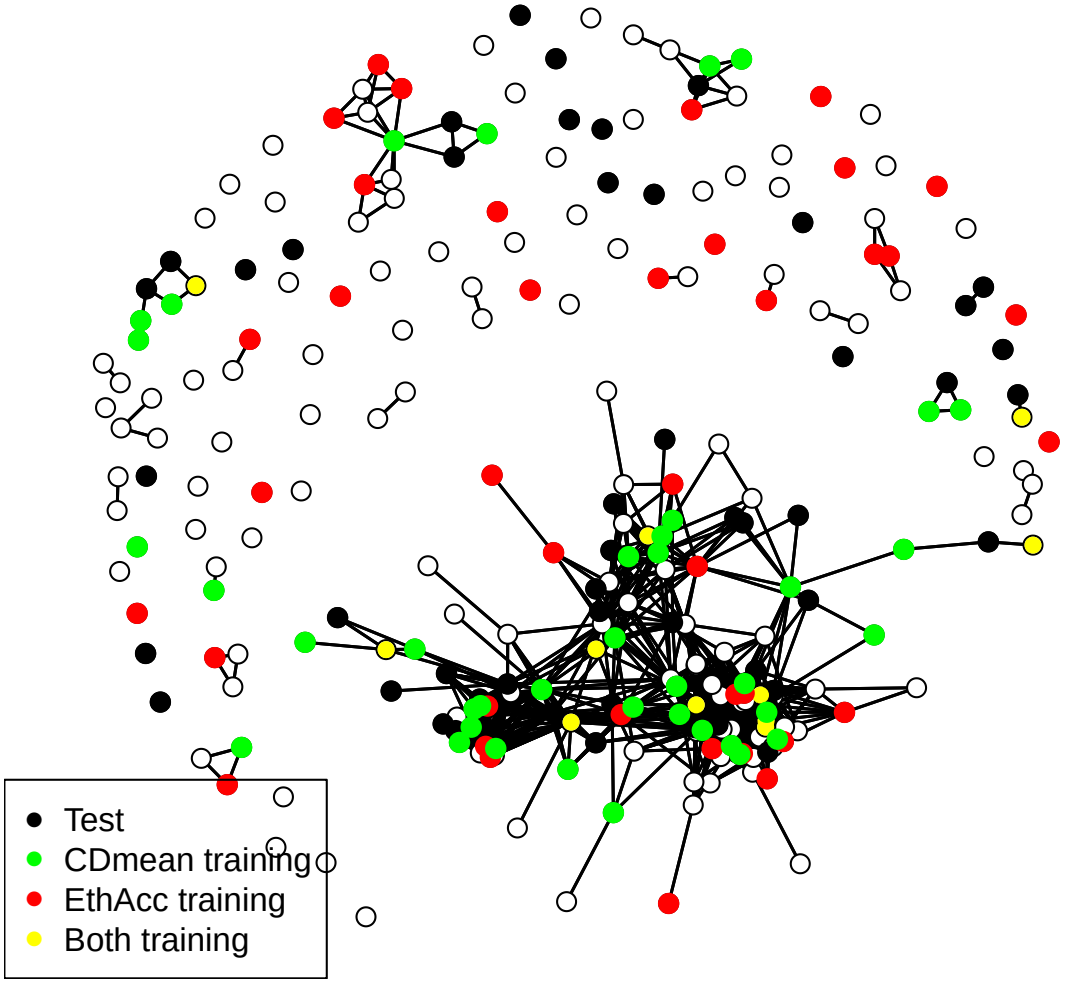

Test set
Training set
Candidates not used

Training set optimization : EthAcc versus CDmean (Rincent et al, Genetics 2012)



Training set optimization (Sugar beet) : EthAcc versus CDmean





- CDmean optimization gave an accuracy of 0.07
- EthAcc optimization gave an accuracy of 0.76

Plan

Pour aller plus loin ...

Théorème (R., Barre, ..., Mangin, Plos One 2016)

Conditionnellement à M et $M^\star$,

$$\rho = \frac{\theta^{\star\prime}\ \mathbb{E}\left(M_{n+1}^{\star\prime} M_{n+1}\right) M' V^{-1} M^\star \theta^\star}{\left\{\sigma_e^2 \mathbb{E}\left(\|M_{n+1} M' V^{-1}\|^2\right) + \theta^{\star\prime} M^{\star\prime} V^{-1} M \mathrm{Var}\left(M_{n+1}'\right) M' V^{-1} M^\star \theta^\star\right\}^{1/2} \Omega^{1/2}}$$

où $V = MM' + \lambda I_n$ et $\Omega = \mathrm{Var}\left(M_{n+1}^\star \theta^\star\right) + \sigma_e^2$.

Décomposition SVD de M

$$M = U\ D\ W'$$

où

- D matrice diagonale de taille $r \times r$, de plein rang
- U matrice de taille $n \times r$, telle que $U'U = I_r$
- W matrice de taille $p \times r$, telle que $W'W = I_r$

En injectant la décomposition SVD dans notre formule

Par abus de notation :

- $\theta^\star$ vecteur sparse de dimension K contenant les effets QTLs

Théorème (R., Mangin, Grusea (en révision pour Scand J. Statistics))

$$\hat{\rho} \geq \frac{\|WW'\theta^\star\|^2 \ \min \frac{d_s^4}{d_s^2+\lambda}}{\sqrt{\sigma_e^2 \ r + \ \|WW'\theta^\star\|^2 \ \max d_s^2} \quad \sqrt{\|WW'\theta^\star\|^2 \ \max d_s^2 + \sigma_e^2}}$$

$$\hat{\rho} = \frac{\sum_{s=1}^{r} \frac{d_s^4}{d_s^2+\lambda} \left\| W^{(s)} W^{(s)\prime} \theta^\star \right\|^2}{\left(\sigma_e^2 \ \sum_{s=1}^{r} \frac{d_s^4}{(d_s^2+\lambda)^2} + \sum_{s=1}^{r} \frac{d_s^6}{(d_s^2+\lambda)^2} \|W^{(s)} W^{(s)\prime} \theta^\star\|^2 \right)^{1/2} \left(\sum_{s=1}^{r} d_s^2 \|W^{(s)} W^{(s)\prime} \theta^\star\|^2 + \sigma_e^2 \right)^{1/2}}$$

On aimerait avoir $\sigma_e^2 \ \sum_{s=1}^{r} \frac{d_s^4}{(d_s^2+\lambda)^2}$ petit, et $\sum_{s=1}^{r} \frac{d_s^4}{d_s^2+\lambda} \left\| W^{(s)} W^{(s)\prime} \theta^\star \right\|^2$ grand

Vers une amélioration du RRBLUP, GBLUP, Ridge

Idée : considérer un espace de dimension plus faible

Rappel : $U = \left(U^{(1)}, \ldots, U^{(r)}\right)$ base orthonormale de l'espace engendré par les colonnes de M.

On choisit $\tilde{r}$ colonnes de U.

Soit l'estimateur

$$\tilde{\theta} = M' V^{-1} \tilde{U} \tilde{U}' Y \quad \text{où} \quad \tilde{U} = \left(U^{\sigma(1)}, \ldots, U^{\sigma(\tilde{r})}\right)$$

où $\tilde{U}\tilde{U}' Y$ est la projection de Y sur $Vect\left\{U^{\sigma(1)}, \ldots, U^{\sigma(\tilde{r})}\right\}$.

$\Rightarrow$ Prédictions a l'aide de $\tilde{\theta}$

Accuracy basée sur ce nouvel estimateur

Théorème (R., Mangin, Grusea (en révision pour Scand J. Statistics))

$$\tilde{\rho} \geq \frac{\left\| \tilde{W}\tilde{W}'\theta^\star \right\|^2 \min\limits_{1\leq s\leq \tilde{r}} \frac{d^4_{\sigma(s)}}{d^2_{\sigma(s)}+\lambda}}{\sqrt{\sigma_e^2\,\tilde{r} + \left\| \tilde{W}\tilde{W}'\theta^\star \right\|^2 \max\limits_{1\leq s\leq \tilde{r}} d^2_{\sigma(s)}} \sqrt{\left\| \tilde{W}\tilde{W}'\theta^\star \right\|^2 \max\limits_{1\leq s\leq \tilde{r}} d^2_{\sigma(s)} + \sigma_e^2}}$$

$$\tilde{\rho} = \frac{\sum\limits_{s=1}^{\tilde{r}} \frac{d^4_{\sigma(s)}}{d^2_{\sigma(s)}+\lambda} \left\| \tilde{W}^{(\sigma(s))}\tilde{W}^{(\sigma(s))\prime}\theta^\star \right\|^2}{\left(\sigma_e^2 \sum\limits_{s=1}^{\tilde{r}} \frac{d^4_{\sigma(s)}}{(d^2_{\sigma(s)}+\lambda)^2} + \sum\limits_{s=1}^{\tilde{r}} \frac{d^6_{\sigma(s)}}{(d^2_{\sigma(s)}+\lambda)^2} \left\| \tilde{W}^{(\sigma(s))}\tilde{W}^{(\sigma(s))\prime}\theta^\star \right\|^2 \right)^{1/2} (\Omega)^{1/2}}$$

$$\text{avec } \Omega = \sum_{s=1}^{r} d^2_{\sigma(s)} \left\| W^{(\sigma(s))}W^{(\sigma(s))\prime}\theta^\star \right\|^2 + \sigma_e^2$$

Dans quelles conditions améliore-t-on l'accuracy ?

- Estimateur Ridge $\hat{\theta}$ basé sur toutes les colonnes de U
  - accuracy $\hat{\rho}$, prédiction $\hat{Y}_{n+1}$
- Nouvel estimateur $\tilde{\theta}$ basé sur $\tilde{r}$ colonnes de $U \Rightarrow \tilde{\beta}$
  - accuracy $\tilde{\rho}$, prédiction $\tilde{Y}_{n+1}$
- Complémentaire $\vec{\theta}$ de notre nouvel estimateur basé sur les $r - \tilde{r}$ colonnes restantes de U
  - accuracy $\vec{\rho}$, prédiction $\vec{Y}_{n+1}$

$\tilde{\rho} \geq \hat{\rho}$ si et seulement si

$$\frac{\widehat{\text{Cov}}\left(\tilde{Y}_{n+1}, Y_{n+1}\right)}{\widehat{\text{Cov}}\left(\vec{Y}_{n+1}, Y_{n+1}\right)} \geq \frac{\widehat{\text{Var}}\left(\tilde{Y}_{n+1}\right)}{\widehat{\text{Var}}\left(\vec{Y}_{n+1}\right)} \left(1 + \sqrt{1 + \frac{\widehat{\text{Var}}\left(\vec{Y}_{n+1}\right)}{\widehat{\text{Var}}\left(\tilde{Y}_{n+1}\right)}}\right).$$

Un exemple où θ appartient à l'espace engendré par les lignes de M

- Chromosome de longueur 1M
- 1000 marqueurs
- $\theta = 0.3W^{(1)} + 0.3W^{(2)} + 0.3W^{(3)}$
- $\tilde{r}$ et les colonnes de U choisies par validation croisée
- 100 TESTS

σ_e^2	n	Méthode	200 générations	400 générations
1	500	$\hat{\rho}$	0.7550	0.6721
		$\hat{\hat{\rho}}$	0.7810	0.7041
	800	$\hat{\rho}$	0.7487	0.7037
		$\hat{\hat{\rho}}$	0.7728	0.7312
9	500	$\hat{\rho}$	0.3370	0.2623
		$\hat{\hat{\rho}}$	0.3809	0.3079
	800	$\hat{\rho}$	0.3317	0.2904
		$\hat{\hat{\rho}}$	0.3734	0.3330

Remerciements

Simona Grusea (INSA), Brigitte Mangin (LIPM)

INSA TOULOUSE

Lipm Laboratoire Interactions Plantes Micro-organismes

Collaborateurs sur CROPDL

- Muriel Tavaud / Jacques David (SupAgro Montpellier)
- Gilles Charmet / Delphine Ly / François Balfourier (INRA Clermont Ferrand)
- Philippe Barre (INRA Lusignan)/ Torben Asp (Aarhus university, Danemark)

INRA SCIENCE & IMPACT SelGen

MIA TOULOUSE

Deuxième partie : Duplications Entières du Génome

Probabilistic approaches for detecting and locating whole genome duplications

Charles-Elie Rabier, Tram Ta, Cécile Ané

UW - Madison
Departments of Statistics and of Botany

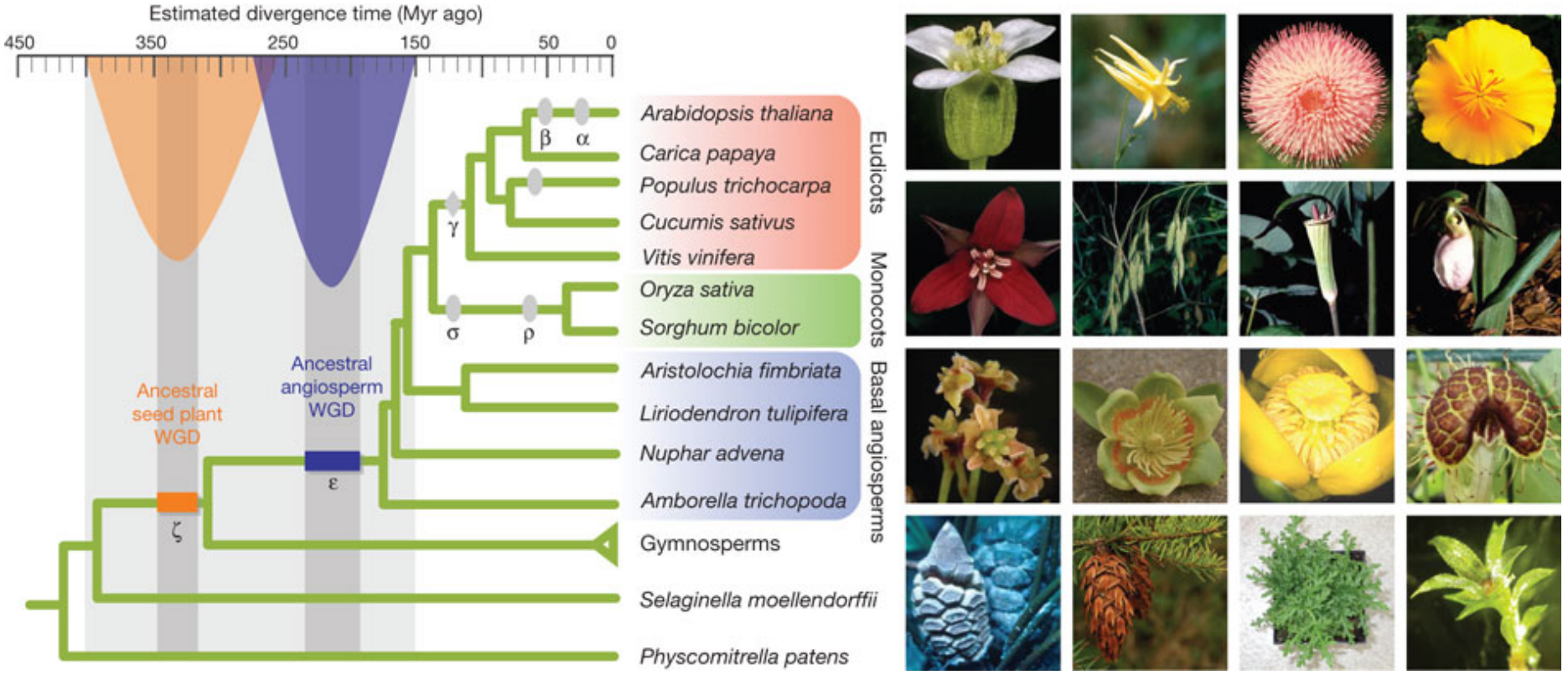
Whole Genome Duplication (WGD)

"Ancestral polyploidy in seed plants and angiosperms", Jiao et al. (Nature 2009)

"Whole-genome duplication followed by gene loss and diploidization has long been recognized as an important evolutionary force in animals, fungi and other organisms, especially plants"

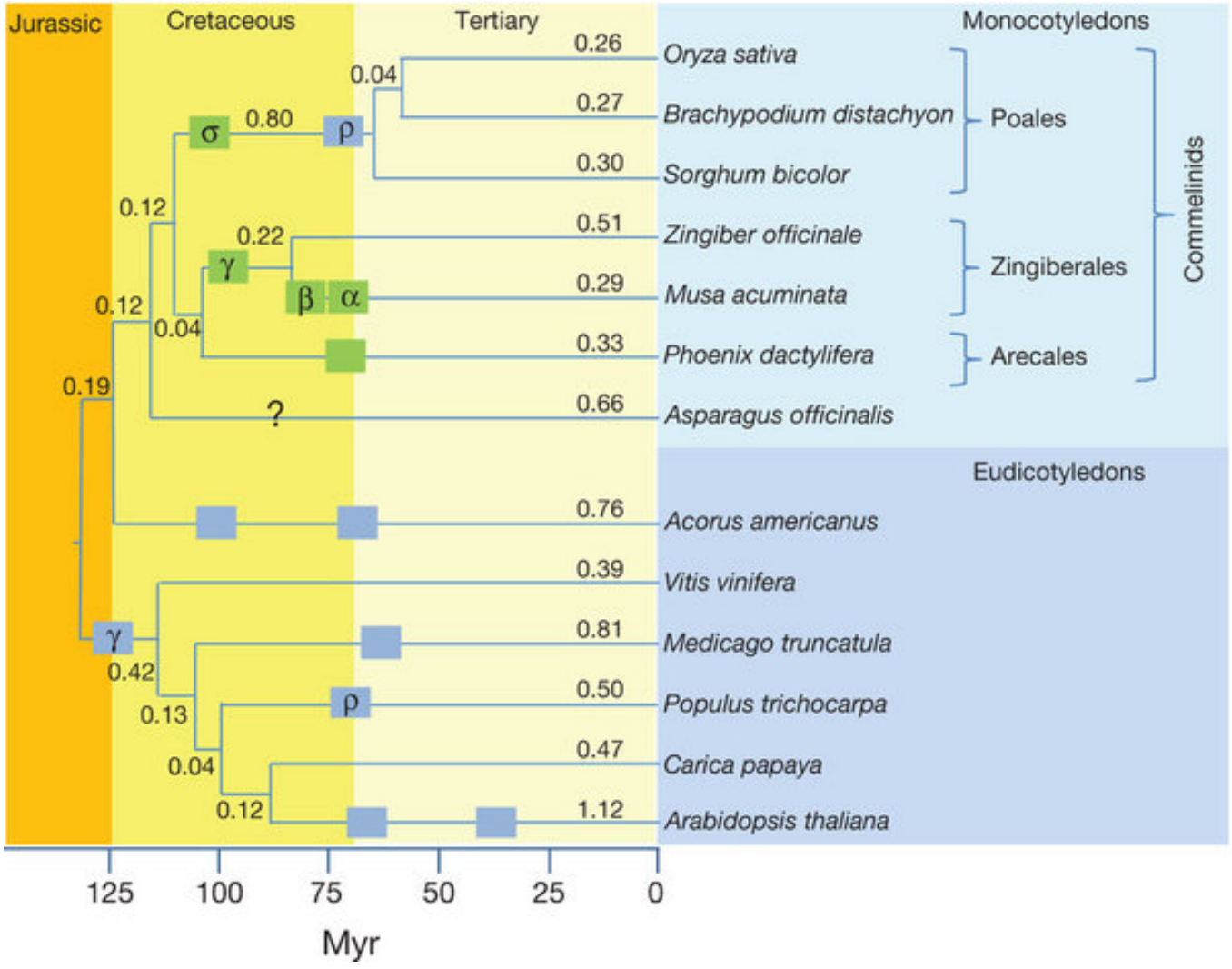
WGD in seed plants and angiosperms

Jiao et al. (Nature 2009)



WGD in bananas

D'Hont et al. (Nature 2012)



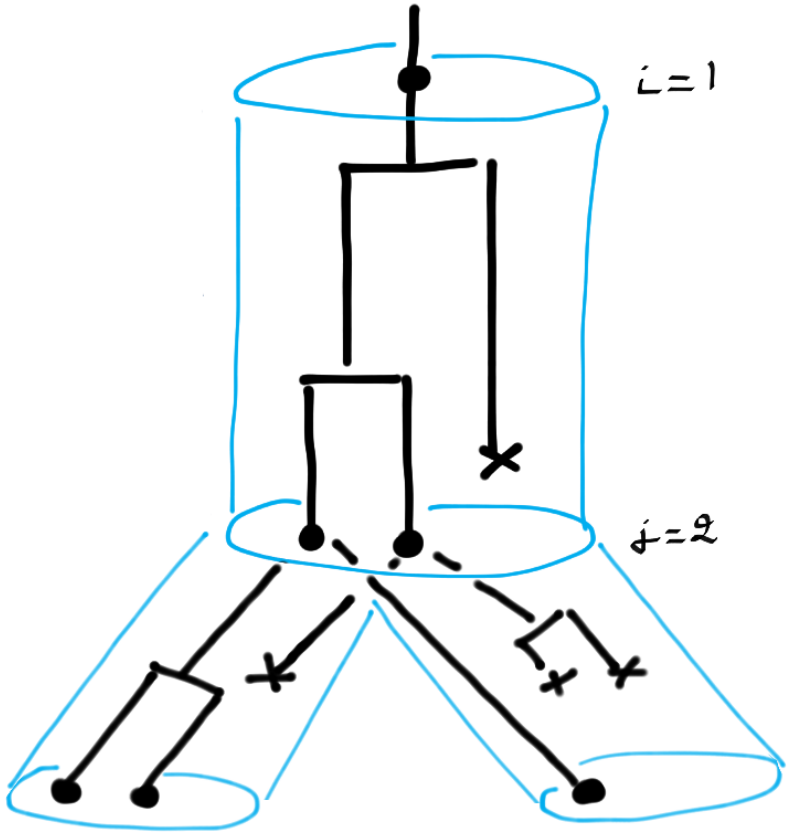
Probabilistic model for gene family evolution

- phylogenetic framework : multiple species
- probabilistic model to avoid ad-hoc filtering of families or nodes
- requires : genes clustered into families. No synteny.

Birth-death model for small-scale events, and
WGD model for large-scale events.

$$\text{likelihood} = \prod_{\text{families } f} \text{likelihood}(f)$$

Birth - death process for small scale events



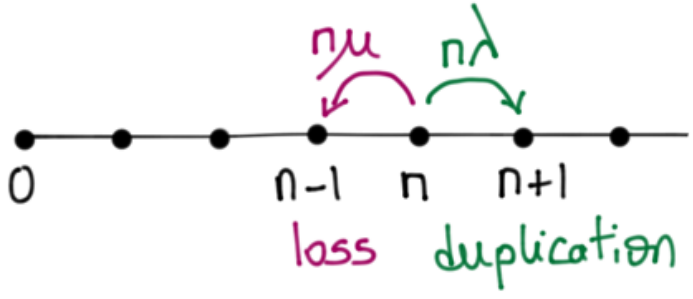
Data
(1 family)

gene counts: X=2 X=1

sequences: a1=AACC b1=AATT

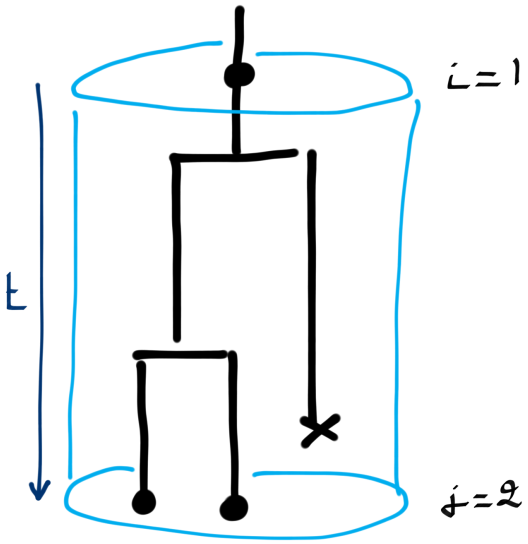
a2=AACT

Birth rate λ, death rate μ



Likelihood of gene counts, birth - death process

λ, μ : birth & death rates



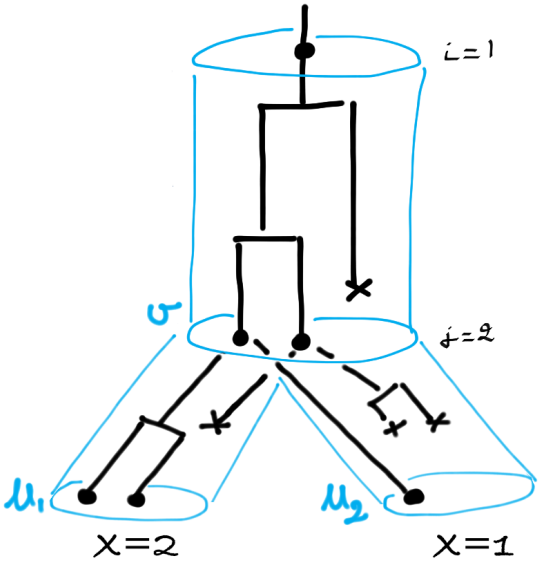
$$p_t(i,j) = \mathbb{P}\left(X_t = j | X_0 = i\right)$$

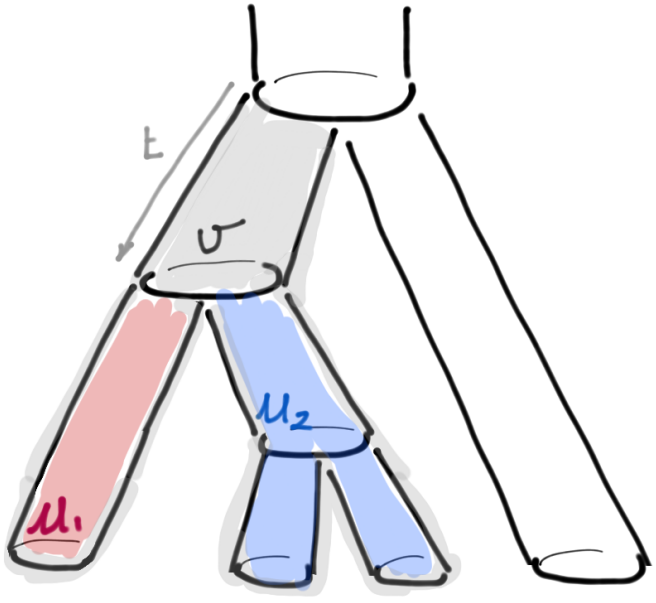
$$p_t(1,0) = \gamma_t = \frac{\mu\left(e^{(\lambda-\mu)t}-1\right)}{\lambda e^{(\lambda-\mu)t}-\mu},$$
$$p_t(1,1) = (1-\gamma_t)(1-\psi_t) \text{ with } \psi_t = \frac{\lambda}{\mu}\gamma_t$$

$$p_t(i,j) = \sum_{k=0}^{i\wedge j} \binom{i}{k} \binom{i+j-k-1}{i-1} \gamma_t^{i-k} \psi_t^{j-k} (1-\gamma_t-\psi_t)^k$$

Bailey (1964)

Likelihood of gene counts, birth - death process





Conditional likelihood $L_v(i)$ at node v : probability of gene count data below v given $X = i$ at parent of v, calculated recursively :

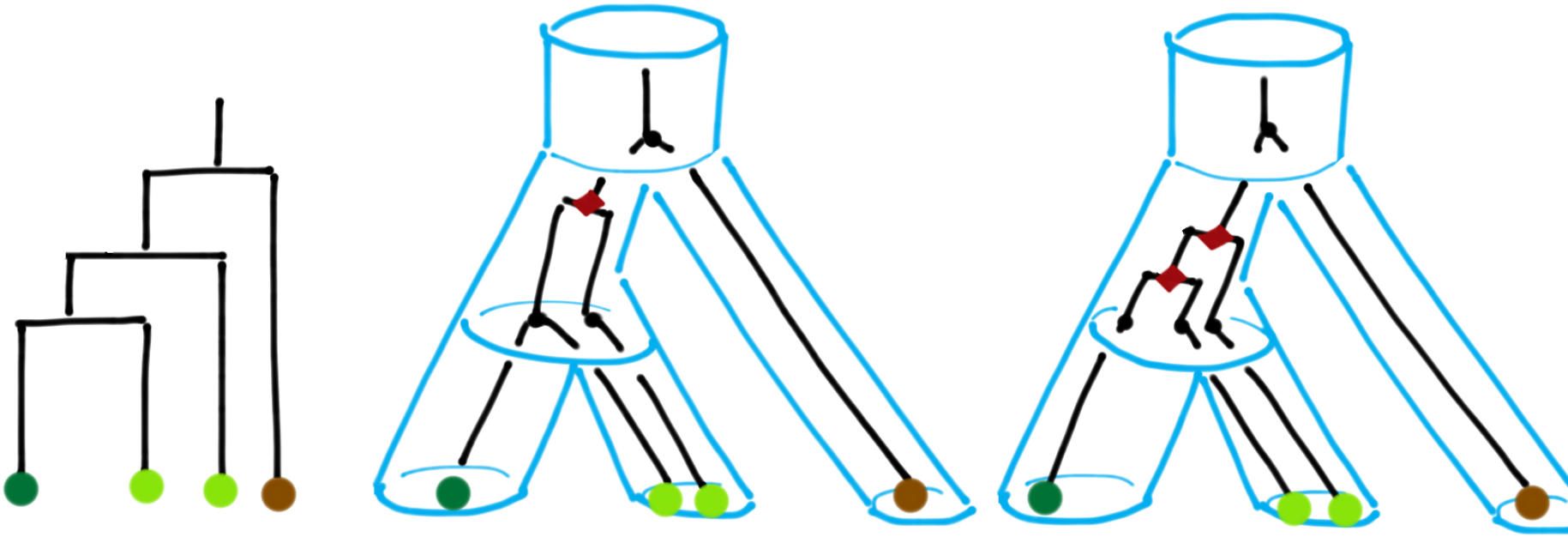
$$L_v(i) = \sum_j p_t(i,j) L_{u_1}(j) L_{u_2}(j)$$

Geometric prior π for # at the root :

$$\text{likelihood} = \sum_j \pi(j) L_{u_1}(j) L_{u_2}(j)$$

or Csűrös & Miklós (2009)

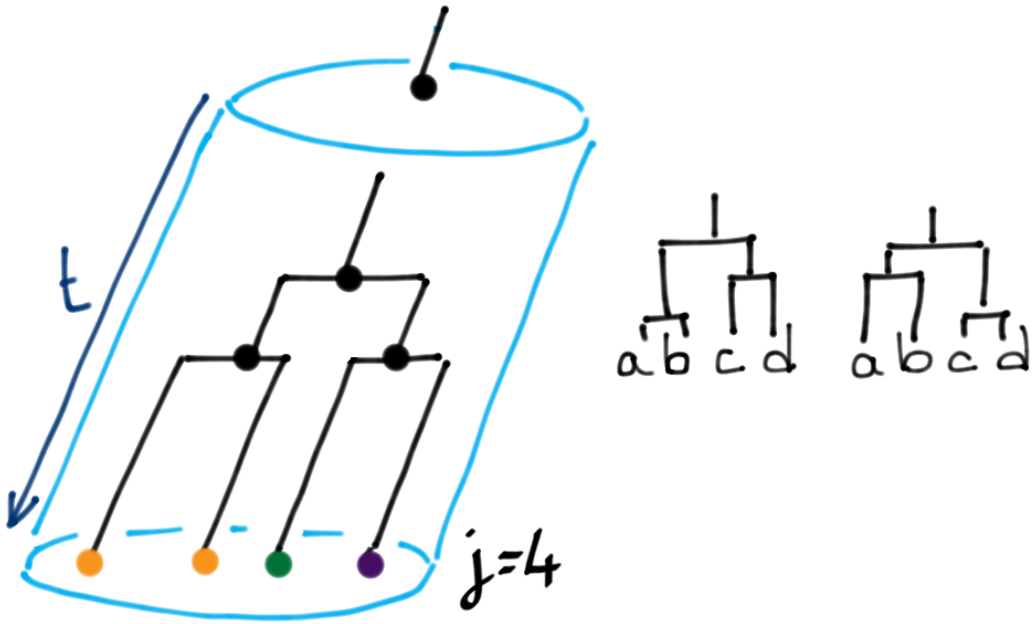
Likelihood of gene tree reconciliations, BD process



Problem 1 : each gene tree has many ”reconciliations” : to map gene tree inside species tree.

Likelihood of gene tree reconciliations, BD process

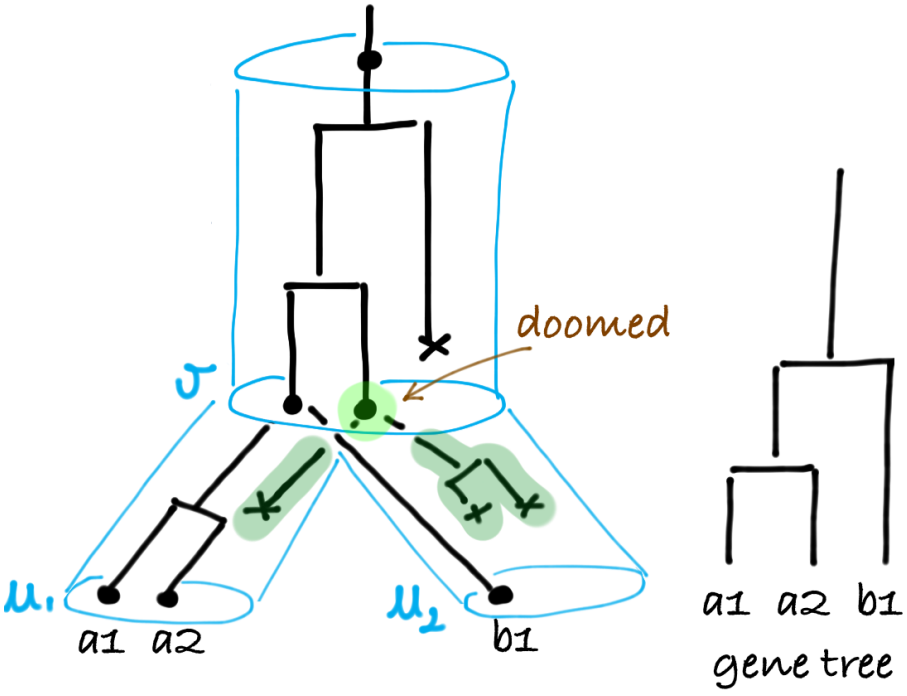
Problem 2 : labels



For a reconciled subtree within a ‘slice’, *j* tips, 3 colors

Arvestad et al. (2009), Rasmussen & Kellis (2011)

Likelihood of gene trees reconciliations, BD process

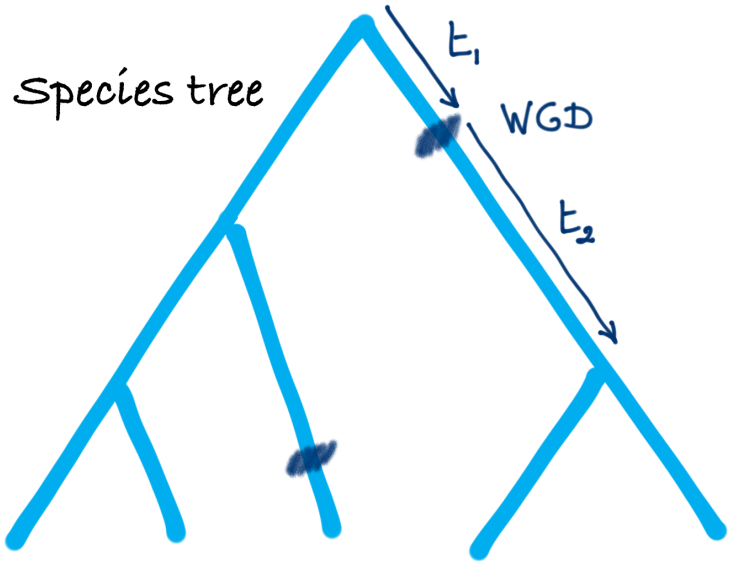


Problem 3 : gene trees lack doomed lineages

d_v : probability that a lineage starting at node v leaves no descendent (or : is doomed). Recursively :

$$d_v = \left(\sum_j p_{t_1}(1,j) d_{u_1}^j \right) \left(\sum_j p_{t_2}(1,j) d_{u_2}^j \right)$$

WGD model for large-scale events



At the WGD :

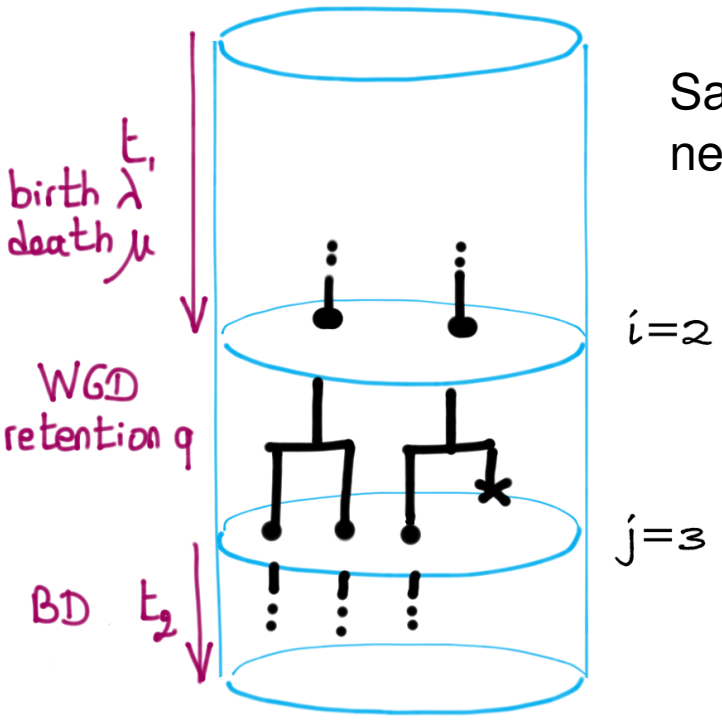
- each gene is duplicated
- second copy lost immediately with probability $1 - q$.

Each WGD has its own retention rate q, to explain :

- Large-scale events
- fragmentation : tendency to lose the extra copy, increased background loss rate shortly after WGD
- extension to whole genome triplications

Rabier, Ta, Ané (2014)

Likelihood : birth-death + WGD model



Same recursive algorithm through the tree, but new transition probabilities along WGD edges :

$$p_{\mathrm{WGD}}(i,j) = \binom{i}{j-i} q^{j-i}(1-q)^{2i-j}$$

$$(i \leq j \leq 2i)$$

Two methods to detect WGDs

Using gene counts only :

- **fast** ($<$ 1s)
- exact likelihood
- optimize λ, μ and separate q's at each WGD
- but : **limited** information

R package `WGDgc`

Two methods to detect WGDs

Using full sequences :

- **rich** information and model, but
- **slow** (e.g. 1h/family) : integrate over tree, reconciliation, branch lengths (gene-specific and species-specific rates).
- approximate likelihood
  - search over gene trees, but most parsimonious reconciliation.
  - new algorithm to find MP reconciliation with WGDs
- fixed $\hat{\lambda}, \hat{\mu}$

C++ program `spimapWGD`, based on `SPIMAP` (Rasmussen & Kellis 2011)

If you are interested in the gene tree ...

Some notations

- S : species tree
- D : data (ie. alignment data)
- T : gene tree topology
- ℓ : branch length
- R : reconciliation

Bayesian framework

- $\mathbb{P}(T, R|S)$: topology prior
- $\mathbb{P}(\ell|T, R, S)$: branch length prior
- $\mathbb{P}(T, R, \ell|D, S)$: posterior

$\Rightarrow$ Markov Chain Monte Carlo (Hasting Metropolis) to estimate posterior distribution $\mathbb{P}(T, R, \ell|D, S)$

Approximate versus exact likelihood

Exact Likelihood

$$\mathbb{P}(D|S) = \sum_{T,R} \int_I \mathbb{P}(D, I, T, R|S)$$

$$= \sum_{T,R} \int_I \mathbb{P}(D|I, T, S)\ \mathbb{P}(I|T, R, S)\ \mathbb{P}(T, R|S)$$

Approximate Likelihood

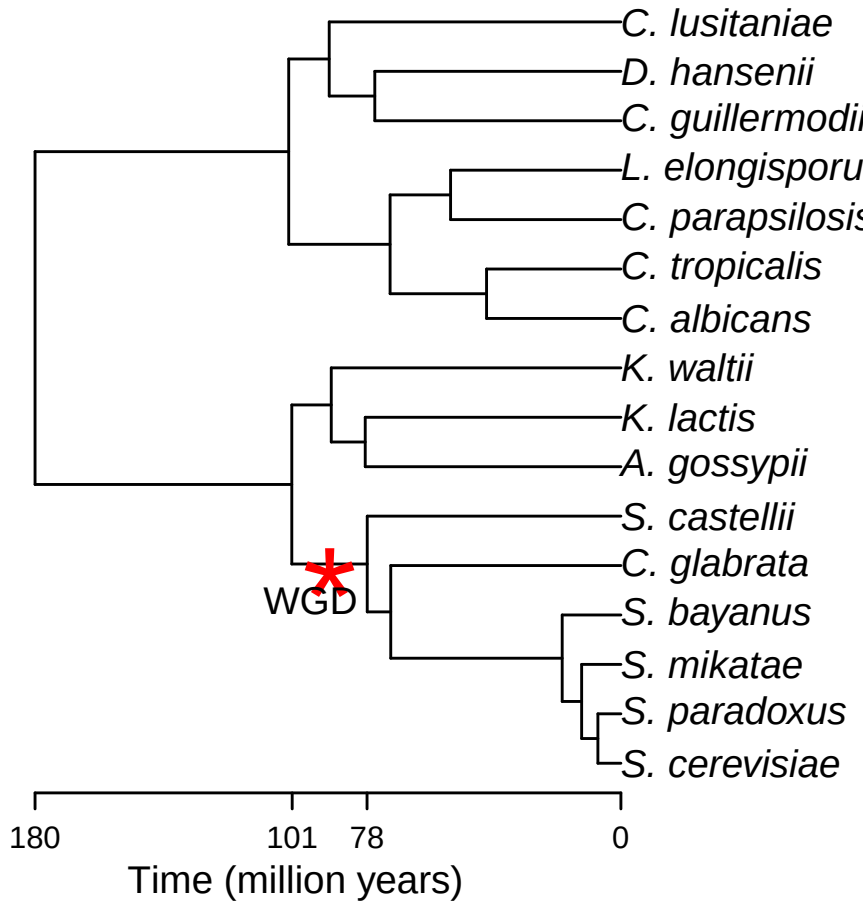
$$\mathbb{P}(D|S) \approx \mathbb{P}(D, \hat{\ell}, \hat{T}, \hat{R}|S)$$

with $\hat{\ell}$, $\hat{T}$, $\hat{R}$ maximum a posteriori estimators of ℓ, T, R given the data

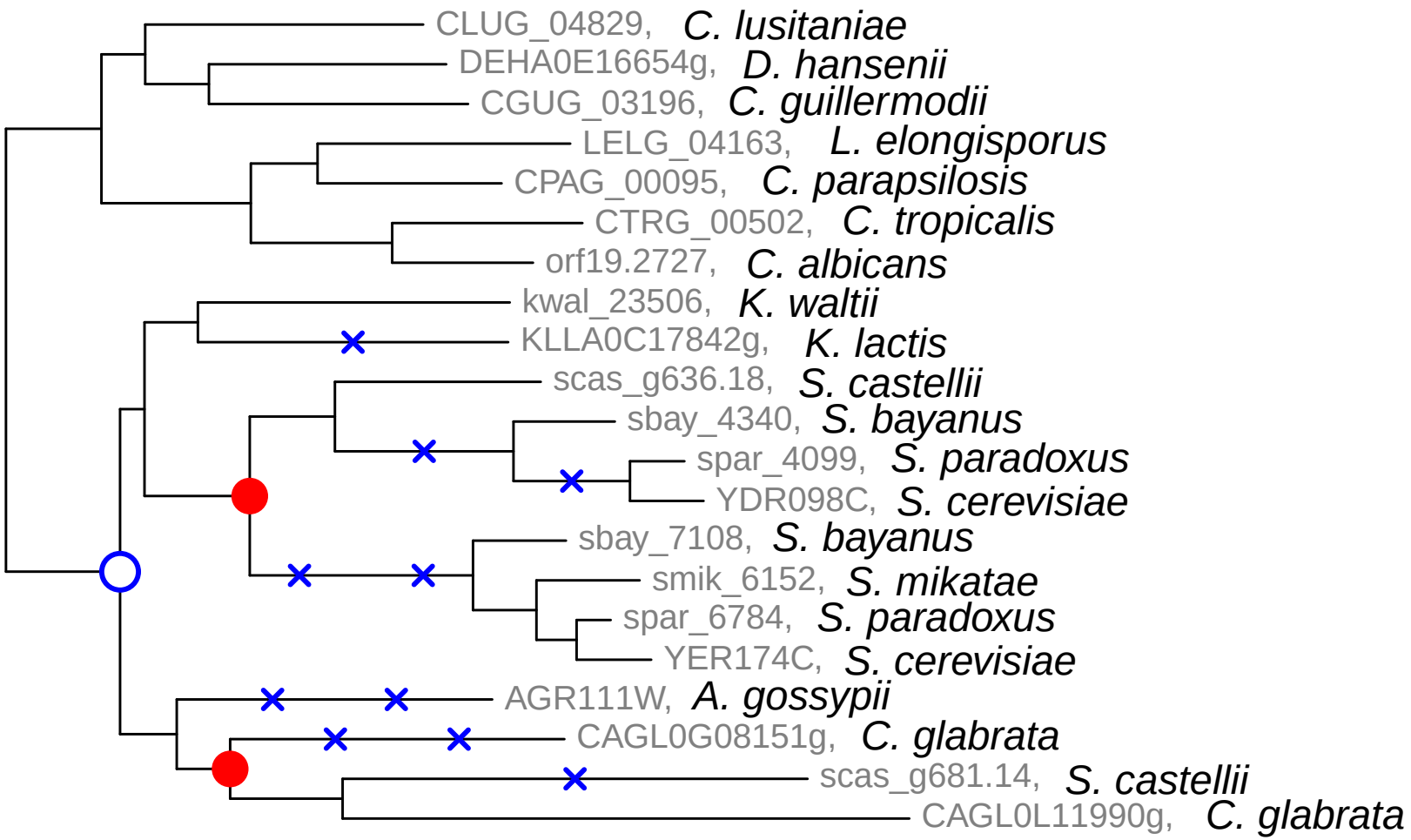
Yeast genome evolution

Kellis et al. (2004), from synteny on *Kluyveromyces waltii* and *S. cerevisiae* : ”12% of the paralogous gene pairs were retained in each doubly conserved synteny block”

- 9209 gene families (Butler et al 2009)
- filter : 3932 families with $\geq$ 1 gene in both *Candida* and *Saccharomyces* subclades



A phylogenetic tree of gene family 1306

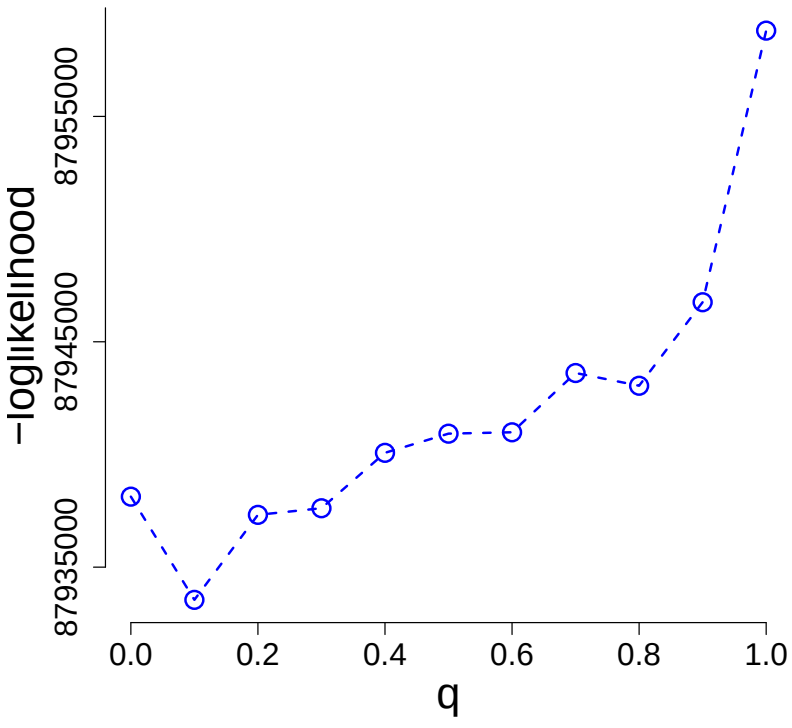


2 duplications at the WGD (red circles), 0 loss at the WGD

1 duplication, 10 losses (blue crosses)

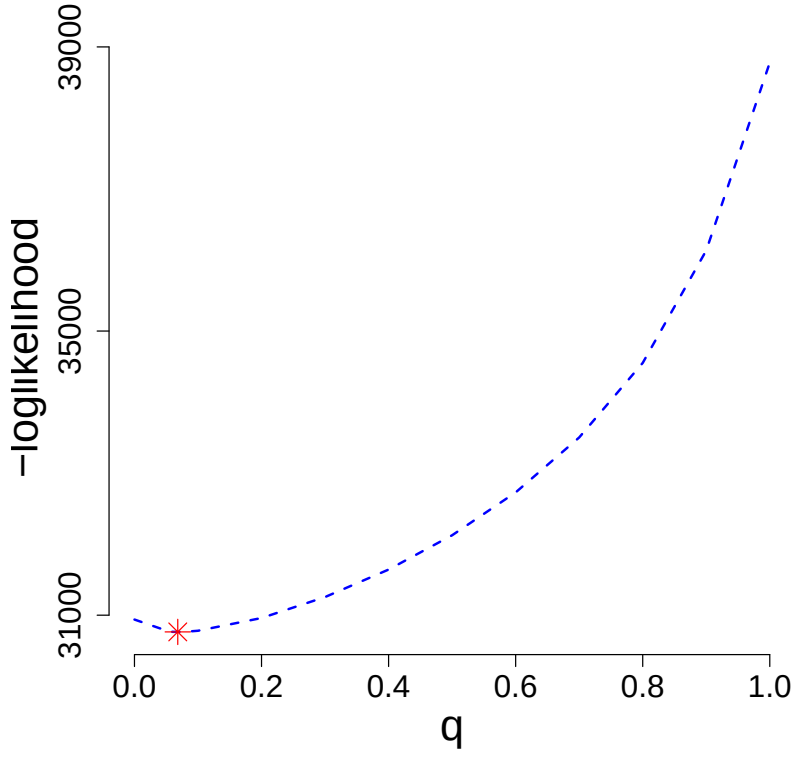
Testing the Yeast WGD

from sequences



LRT : 9159.5

from gene counts

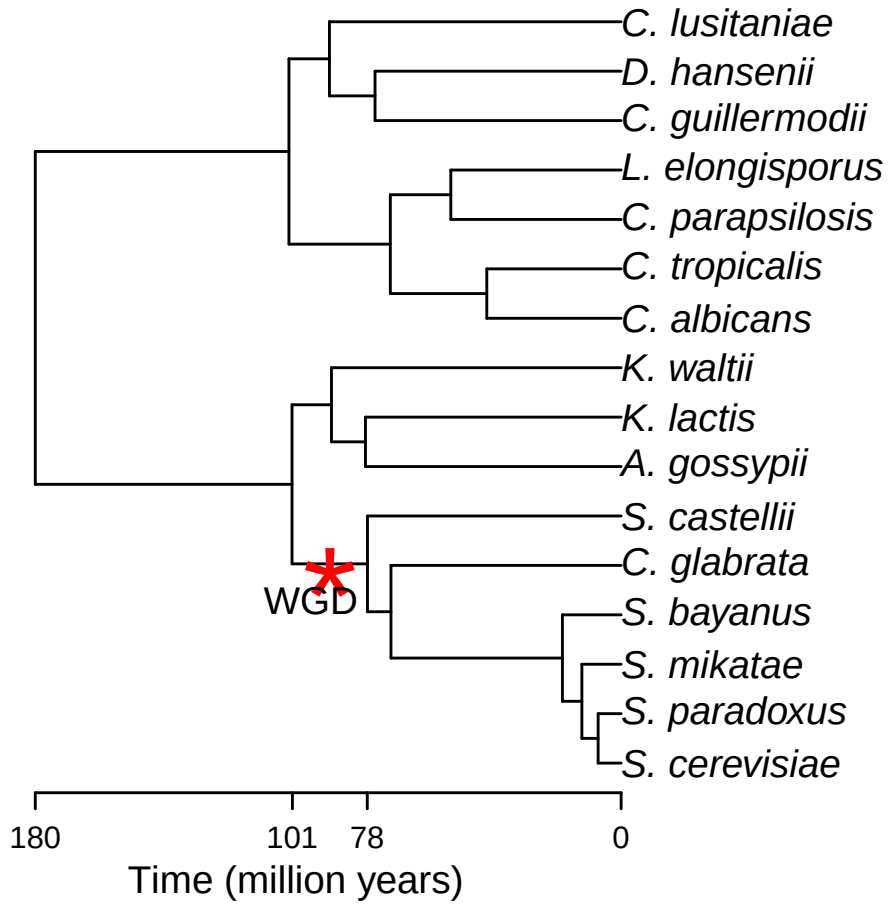


LRT : 348.1

retention rate : $\hat{q} = 6.81\%$, in $[0.058, 0.079]$ with 95% confidence

Yeast WGD timing

$\hat{t} = 0$: immediately before speciation,
$\hat{t} \leq 5.04$ My with 95% confidence.



Thanks to

Cécile Ané
Tram Ta

Matt Rasmussen
Bill Taylor

DEB-0949121

Troisième partie : Réseaux phylogénétiques

Apport des approches phylogénétiques pour expliquer l'origine des génomes mosaïques, exemple chez le Riz

Charles-Elie Rabier, Vincent Berry, Fabio Pardi et Céline Scornavacca

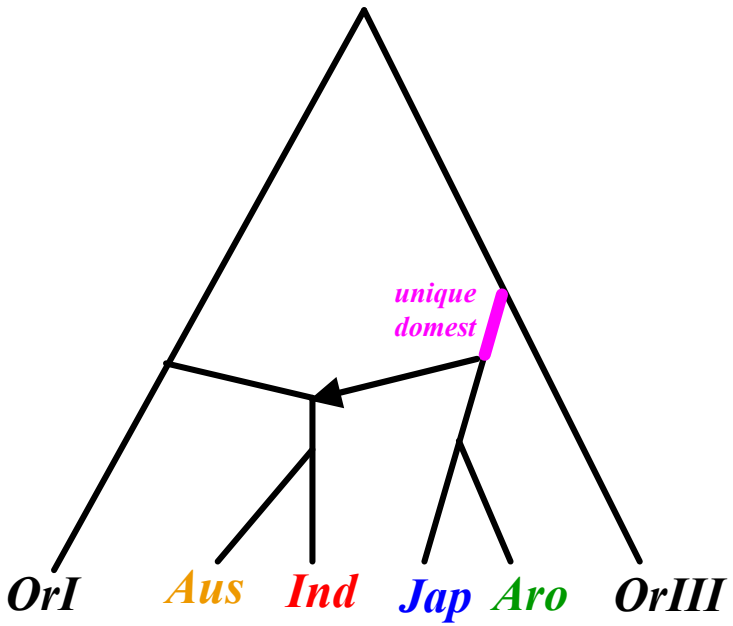
ISEM, Institut des Sciences de l'Evolution de Montpellier
LIRMM, Laboratoire d'informatique, de Robotique et de Microélectronique
Genome Harvest

Jean-Christophe Glaszmann, Joao Santos

AGAP, Amélioration Génétique et Adaptation des Plantes, CIRAD

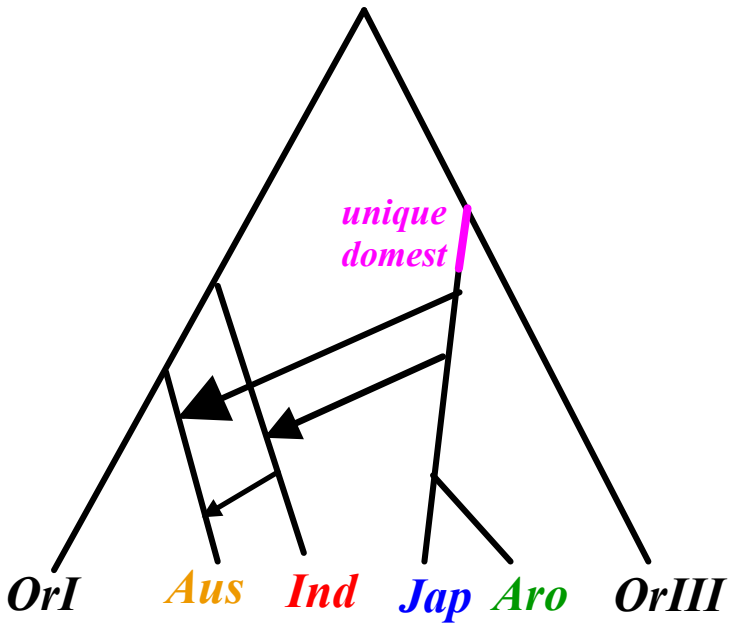
Quelques thèses sur la domestication

- Huang et al. (Nature, 2012) : japonica domestiqué à partir d'un riz sauvage dans le sud de la Chine, puis croisé à un sauvage dans le sud est de l'Asie, générant indica



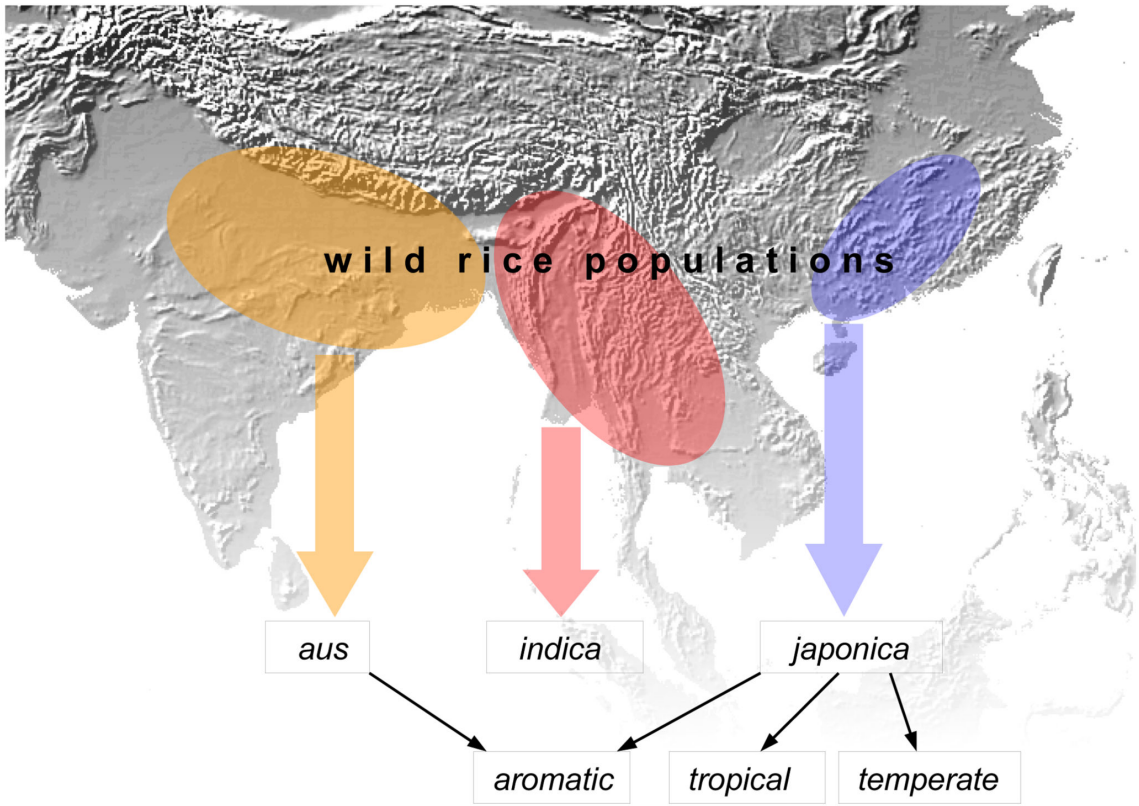
Quelques thèses sur la domestication

- Choi et al. (MBE, 2017) soutiennent aussi un seul évènement de domestication (japonica). Introgression par hybridation de japonica et proto-indica et proto-aus, générant indica et aus



Quelques thèses sur la domestication

- Civan et al. (Nature Plants, 2015) : indica, japonica et aus domestiqués séparément dans différentes parties d'Asie



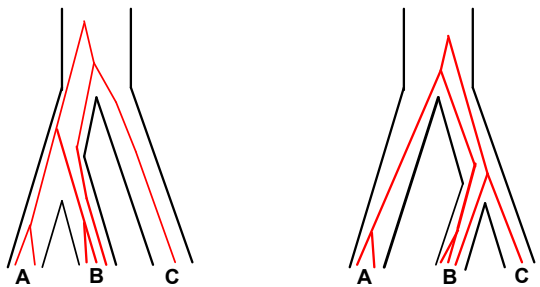
Notre approche méthodologique

On s'intéresse à un modèle qui, outre le tri de lignées, considère explicitement les mutations et hybridation. Modélisation Bayésienne plus fine.

Nos pistes :

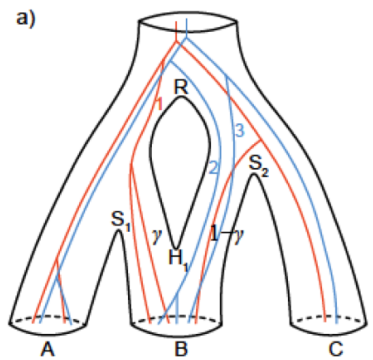
1. Inférence d'arbres d'espèces + arbres résumés en réseaux phylogénétiques

SNAPP (Bryant et al. 2012, MBE) + SplitsTree



2. Inférence directe de réseaux

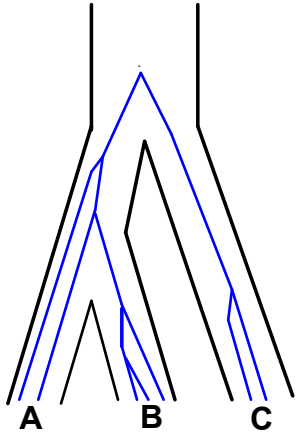
Extension de SNAPP aux réseaux



Logiciel SNAPP pour l'inférence Bayésienne d'arbres (Bryant et al. 2012, MBE)

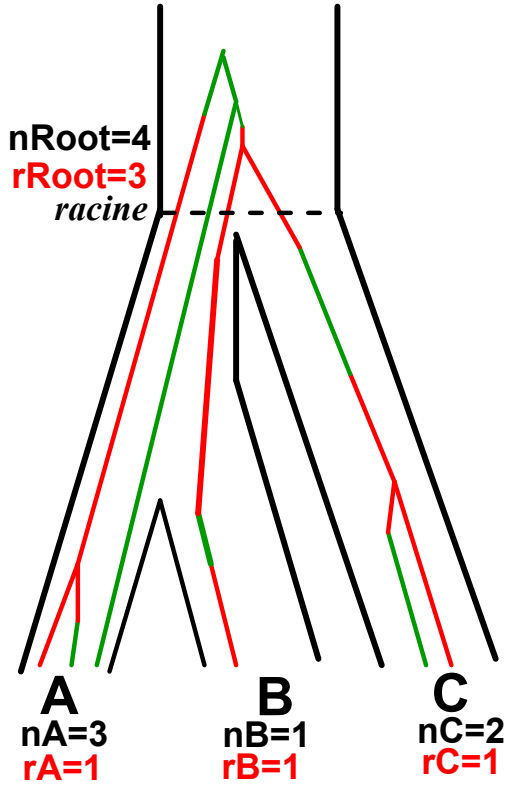
- Marqueurs bialléliques (SNPs) indépendants sachant l'arbre d'espèces

- Modélisation de l'arbre de locus (backward)
  - Processus de coalescence évoluant à l'intérieur d'un arbre d'espèces (MultiSpecies Coalescent)
  - Processus autorisant la discordance entre arbres de locus et arbres d'espèces (tri de lignées incomplet)



Les mutations interviennent au cours du temps

- Modélisation des séquences (forward)
  - mutation (rouge $\leftrightarrow$ vert) : modèle markovien évoluant le long des branches de l'arbre de locus
  - u : taux de mutation rouge $\rightarrow$ vert
  - v : taux de mutation vert $\rightarrow$ rouge



- V.a. : rRoot, nRoot, rA, rB, rC
- pas d'aléa dans nA, nB, nC
- Data=(rA, rB, rC)
- Vraisemblance : $\mathbb{P}(Data \mid S)$

Calcul de vraisemblance dans un arbre (1)

$$
\begin{aligned}
&\mathbb{P}(Data) \\
&= \sum_i \sum_j \mathbb{P}(Data \mid Count, n_{root} = i, r_{root} = j)\, \mathbb{P}(n_{root} = i, r_{root} = j \mid Count) \\
&= \sum_i \sum_j \mathbb{P}(Data \mid Count, n_{root} = i, r_{root} = j)\, \mathbb{P}(r_{root} = j \mid n_{root} = i) \\
&\qquad \times \mathbb{P}(n_{root} = i \mid Count)
\end{aligned}
$$

Calcul de vraisemblance dans un arbre (2)

- $\mathbb{P}(n_{root} = i \mid Count)$ calculé récursivement en remontant dans le temps (postorder)

  Tavaré (Theor Pop Biol, 1984), Watterson (Theor Pop Biol, 1984), Takahata and Nei (Genetics, 1985) ...

- $\mathbb{P}(Data \mid Count, n_{root} = i, r_{root} = j)$ calculé récursivement en remontant dans le temps (postorder)

  Slatkin (Genetics, 1996) vs. Griffiths and Tavaré (Springer, 1997)

- $\mathbb{P}(r_{root} = j \mid n_{root} = i)$ calculé par
  - la loi Binomiale : $\mathbb{P}(r_{root} = j \mid n_{root} = i) = C_i^j p^j (1-p)^{i-j}$
  - la loi $\beta(\theta, \theta)$ sur le paramètre p de la Binomiale :
    $\mathbb{P}(r_{root} = j \mid n_{root} = i) = C_i^j B(j+\theta, i-j+\theta)/B(\theta, \theta)$

- Astuces afin de raccourcir les calculs : Vraisemblances partielles...

La statistique Bayésienne dans SNAPP

- S : arbre d'espèces (topologie, longueurs de branches, tailles de populations)
- X_i : alignements pour le locus i
- G_i : arbre de locus pour le locus i
- m loci

$$\mathbb{P}(S|X_1,\ldots,X_m) \propto \left(\prod_{i=1}^{m}\int_{\psi}\mathbb{P}(X_i|G_i)\mathbb{P}(G_i|S)dG_i\right)P(S)$$
$$\propto \mathbb{P}\left(Data \mid S\right)P(S)$$

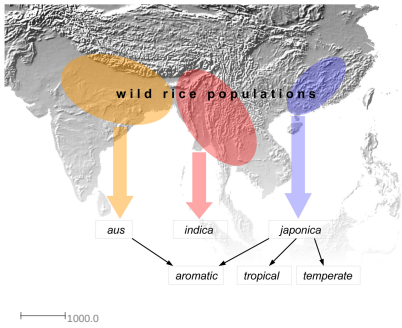
SNAPP intègre sur tous les arbres de locus

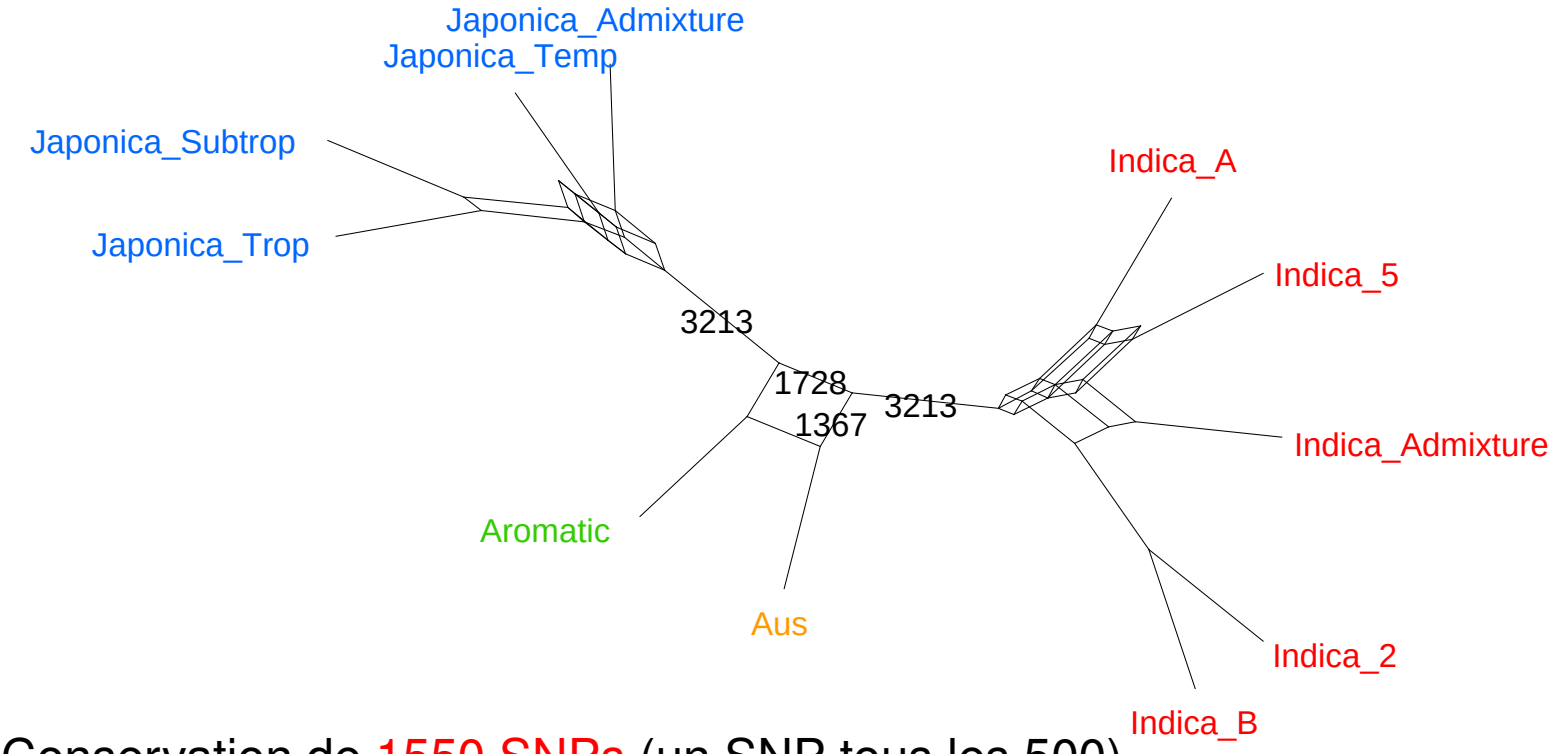
Calcul de la *prior* P(S) par le processus de naissances

$\Rightarrow$ Markov Chain Monte Carlo (MCMC) afin d'estimer la distribution à posteriori de $\mathbb{P}(S|X_1,\ldots,X_m)$

Implémenté dans BEAST

Chromosome 6 (données J. Santos, J-C. Glaszmann)



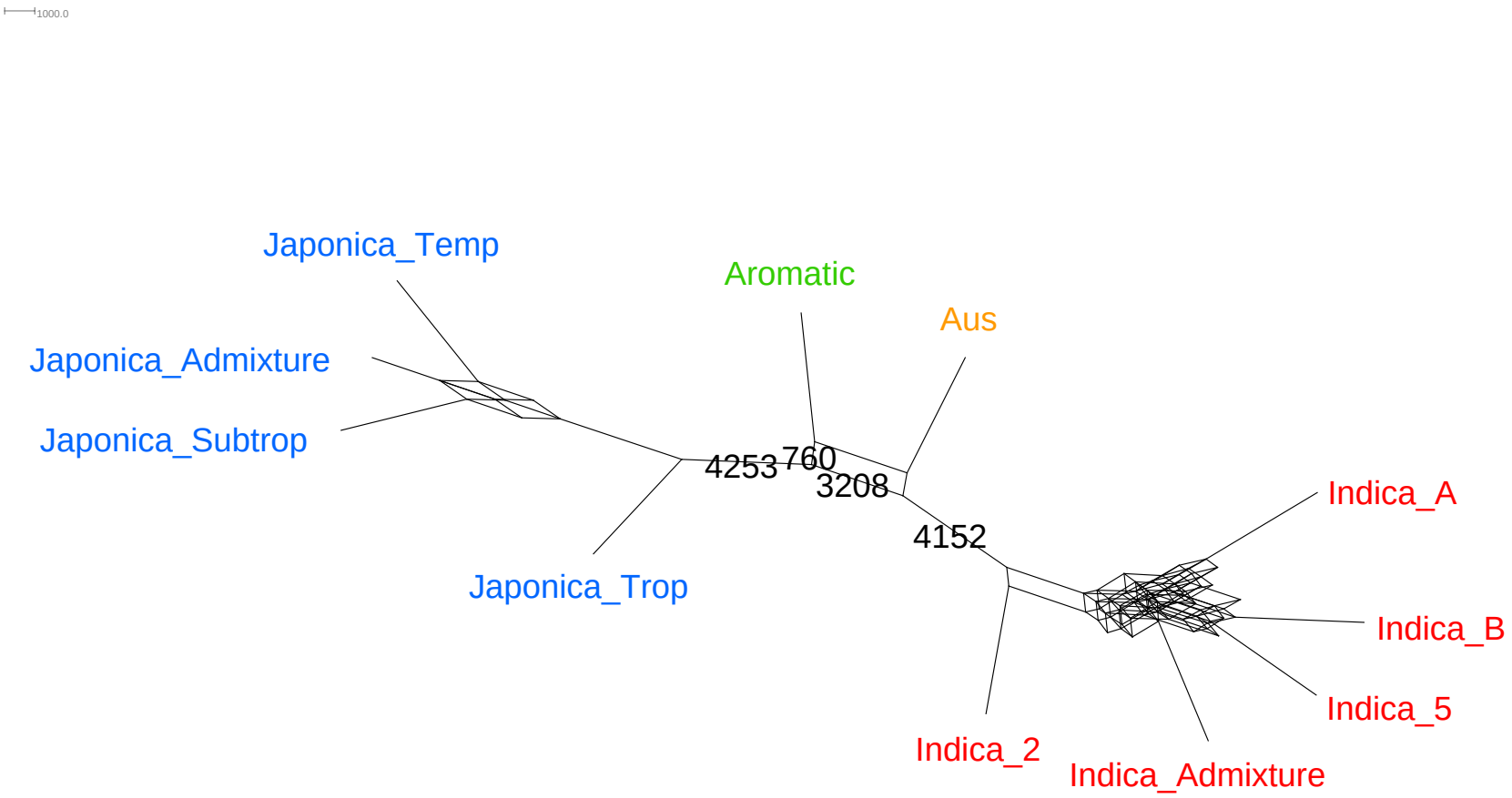


Conservation de 1550 SNPs (un SNP tous les 500)

Chromosome 10 (données J. Santos, J-C. Glaszmann)

Conservation de 1089 SNPs (un SNP tous les 500)

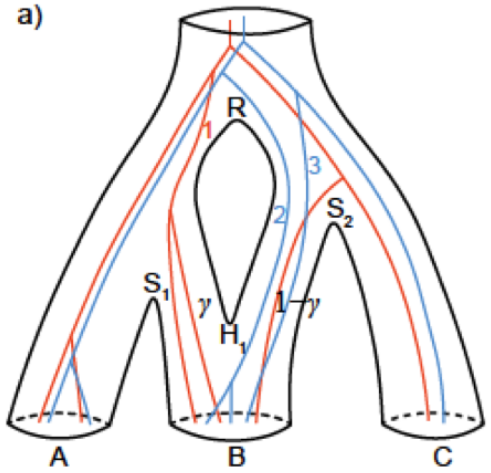
- JDD2 (1er SNP= 50ème SNP du chromosome 10)



Simulateur basé sur un réseau (Genome Harvest)

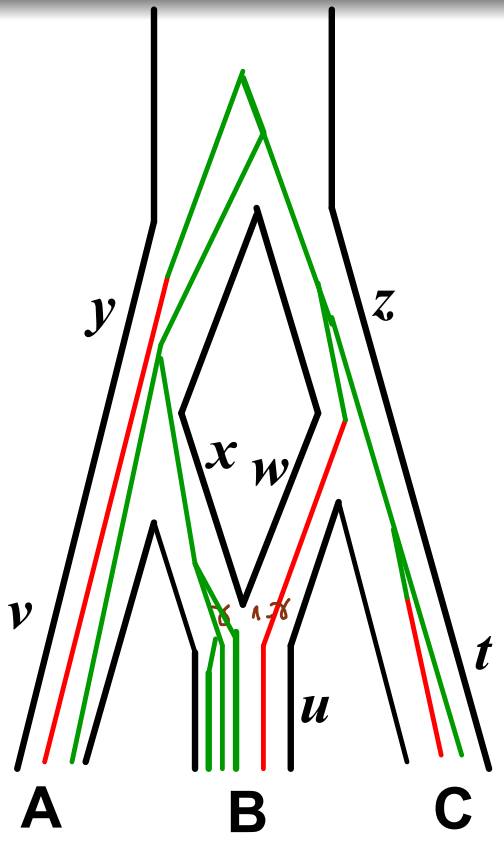
SNAPPSimNet construit sur la base du simulateur SNAPPSim de Bryant et al. (2012)

- Génération d'arbres de locus évoluant à l'intérieur d'un réseau selon un processus de coalescence



- Snapp est fortement attiré par un scénario sous-jacent au réseau

Piste 2 : une méthode Bayésienne directe d'inférence de réseaux



$Data_z$: proportion de rouge/vert dans les espèces sous la branche z
$Data_y$: proportion de rouge/vert dans les espèces sous la branche y

$Data_{zT}$ et $Data_{yT}$ ne sont plus indépendantes...
$Data_{zT}$ et $Data_{yT}$ comprennent les allèles rouges et verts de l'espèce hybride B !!!
On ne peut plus effectuer le produit des probabilités
$\mathbb{P}(Data_{zT}) \times \mathbb{P}(Data_{yT})$!!!

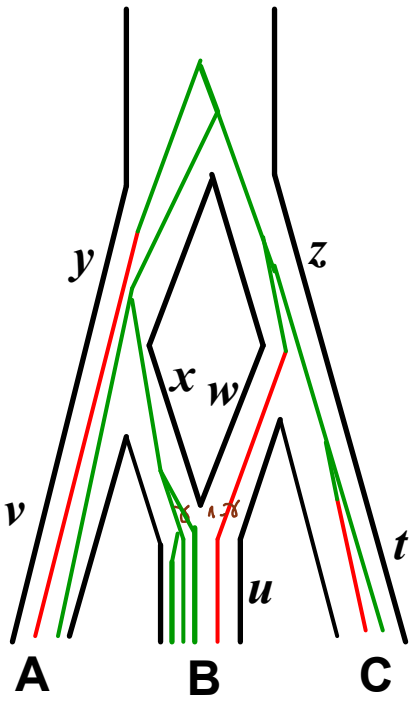
Calcul de la vraisemblance dans un réseau

Pour raccourcir, *Count* est implicite dans les probabilités ...

$$
\begin{aligned}
&\mathbb{P}\left(Data\right) \\
=&\sum_i \sum_j \mathbb{P}\left(Data \mid n_{root} = i, r_{root} = j\right) \mathbb{P}\left(r_{root} = j \mid n_{root} = i\right) \\
&\mathbb{P}\left(n_{root} = i\right) \\
=&\sum_i \sum_j \sum_{i'} \sum_{j'} \mathbb{P}\left(Data_{zT}\, Data_{yT} \mid n_{yT} = i', n_{zT} = i - i', r_{yT} = j',\right. \\
&\left. r_{zT} = j - j'\right) \mathbb{P}\left(r_{yT} = j',\ r_{zT} = j - j' \mid n_{yT} = i', n_{zT} = i - i', r_{root} = j\right) \\
&\mathbb{P}\left(n_{yT} = i', n_{zT} = i - i' \mid n_{root} = i\right) \mathbb{P}\left(r_{root} = j \mid n_{root} = i\right) \mathbb{P}\left(n_{root} = i\right)
\end{aligned}
$$

On ne peut plus effectuer le produit des probabilités
$\mathbb{P}\left(Data_{zT}\right) \times \mathbb{P}\left(Data_{yT}\right)$!!!

Notre algorithme : calcul des lois jointes



Quantités calculées successivement

(1) $\mathbb{P}\left(Data_{uT} \mid n_{uT}, r_{uT}\right)$

(2) $\mathbb{P}\left(Data_{xB}\, Data_{wB} \mid n_{xB}, r_{xB}, n_{wB}, r_{wB}\right)$

(3) $\mathbb{P}\left(Data_{xT}\, Data_{wB} \mid n_{xT}, r_{xT}, n_{wB}, r_{wB}\right)$

(4) $\mathbb{P}\left(Data_{xT}\, Data_{wT} \mid n_{xT}, r_{xT}, n_{wT}, r_{wT}\right)$

(5) $\mathbb{P}\left(Data_{vT} \mid n_{vT}, r_{vT}\right)$

(6) $\mathbb{P}\left(Data_{yB}\, Data_{wT} \mid n_{yB}, r_{yB}, n_{wT}, r_{wT}\right)$

(7) $\mathbb{P}\left(Data_{tT} \mid n_{tT}, r_{tT}\right)$

(8) $\mathbb{P}\left(Data_{yB}\, Data_{zB} \mid n_{yB}, r_{yB}, n_{zB}, r_{zB}\right)$

(9) $\mathbb{P}\left(Data_{yT}\, Data_{zB} \mid n_{yT}, r_{yT}, n_{zB}, r_{zB}\right)$

(10) $\mathbb{P}\left(Data_{yT}\, Data_{zT} \mid n_{yT}, r_{yT}, n_{zT}, r_{zT}\right)$

(11) $\mathbb{P}\left(Data \mid n_{root}, r_{root}\right)$

- Optimisation combinatoire des calculs en cours (dimension des matrices)

Conclusion

- Implémentation de la méthode Bayésienne pour les réseaux (en cours)
- L'inférence de réseau est un sujet compétitif : Tanja Stadler (ETH Zurich), Luay Nakhleh (Rice University, USA). Notre approche devrait être plus performante sur des réseaux aux nombreuses hybridations
- Jusqu'alors travail sur le riz $\rightarrow$ autre plante d'intérêt ? Genome Harvest : Banane, Citrus, Caféier, Riz, Tomate, Canne à sucre ...
- Afin de comprendre l'histoire des riz cultivés, nécessité de disposer de riz sauvages, à l'instar de Choi et al. (MBE, 2017), Wang et al (Genome Research, 2017) ...
- SNAPP disponible sur http ://snapp.otago.ac.nz

Descriptif des simulations

- Population panmictique évoluant pendant 30, 50 or 70 generations
- Taille de la population : 100 TESTS +
  - $n = 500$ Trainings
  - $n = 1000$ Trainings
- Espèce haploïde
- Recombination modélisée selon Haldane (processus de Poisson d'intensité 1)
- Longueur du génome $L = 1M$
- Marqueurs équidistants
- Densité de marqueurs : 100 SNPs, 1000 SNPs, 5000 SNPs ou 10000 SNPs
- Nombre de QTLs
  - 2 QTLs à 3cM et 80cM avec effet $+1$ et -2
  - 100 QTLs équidistants avec effet $+0.15$

Un software intégrant différentes formules pour l'accuracy en selection génomique

ShinyGPAS par Morota (Université de Nebraska Lincoln)
disponible sur https ://chikudaisei.shinyapps.io/shinygpas/

Formules implémentées :

- Daetwyler et al. (Plos One 2008, Genetics 2010)
- Goddard et al. (Genetica 2009, Journal Of Animal Breeding And Genetics 2011)
- Nous :) (Plos One, 2016)
- de los Campos et al. (Plos Genetics, 2013)

$$\rho \leq h\sqrt{\{1-(1-b)^2\}\,h^2}$$

b : coeff de régression entre relations génomiques aux marqueurs et relations génomiques aux génomes causaux

- Karaman et al. (Plos One, 2016), $\tilde{h}^2$ proportion de variance expliquée par les marqueurs

$$\rho = \tilde{h}\sqrt{\frac{n\tilde{h}^2}{n\tilde{h}^2 + M_e}}$$

De nouvelles proxies utilisant les effets QTLs

En injectant l'estimation des causaux dans notre formule (LASSO, Adaptative LASSO, Group LASSO) ...

$$\hat{\rho} = \frac{\hat{\theta}^{\star\prime}\ \mathbb{E}\left(M_{n+1}^{\star\prime} M_{n+1}\right) M' V^{-1} \hat{M}^{\star} \hat{\theta}^{\star}}{\left\{\sigma_e^2 \mathbb{E}\left(\left\|M_{n+1} M' V^{-1}\right\|^2\right) + \hat{\theta}^{\star\prime} \hat{M}^{\star\prime} V^{-1} M \mathrm{Var}\left(M_{n+1}'\right) M' V^{-1} \hat{M}^{\star} \hat{\theta}^{\star}\right\}^{1/2} \Omega^{1/2}}$$

où $V = MM' + \lambda I_n$ et $\Omega = \mathrm{Var}\left(M_{n+1}^{\star} \theta^{\star}\right) + \sigma_e^2$

$\mathbb{E}\left(M_{n+1}^{\star\prime} M_{n+1}\right)$, $\mathbb{E}\left(\left\|M_{n+1} M' V^{-1}\right\|^2\right)$ et $\mathrm{Var}\left(M_{n+1}'\right)$ peuvent être estimés à l'aide :

- uniquement des Trainings $\rightarrow$ estimation de l'accuracy avant le génotypage des TESTS ... $\hat{\rho}_{before}$
- à la fois des Trainings et TESTS $\rightarrow$ estimation de l'accuracy après le génotypage des TESTS ... $\hat{\rho}_{after}$

Régressions pénalisées

LASSO (Tibshirani, JRSSB 1996)

$$\hat{\theta}_{LASSO} = \text{argmin} \, \| Y - X\theta \|^2 + \lambda \sum_{k=1}^{K} |\theta_k|$$

Adaptative LASSO (Zou, JASA 2006)

$$\hat{\theta}_{ADLASSO} = \text{argmin} \, \| Y - X\theta \|^2 + \lambda \sum_{k=1}^{K} w_k \, |\theta_k|$$

Group LASSO (Yuan and Lin, JRSSB 2006)

sparsité par groupes, L nombre de groupes
n_ℓ nombre de marqueurs dans le groupe ℓ

$$\hat{\theta}_{GPLASSO} = \text{argmin} \left\| Y - \sum_{\ell=1}^{L} X_\ell \vec{\theta}_\ell \right\|^2 + \lambda \sum_{\ell=1}^{L} \sqrt{n_\ell} \left\| \vec{\theta}_\ell \right\|^2$$

Accuracy moyenne (100 simulations)

- LD parfait
- 100 QTLs avec le même effet $+0.15$
- 2 QTLs à 3cM et 80cM avec effets $+1$ and -2
- Training et TESTS basés sur 50 générations
- 1000 marqueurs, 500 Trainings, 100 TESTS

Méthode	100 QTLs	2 QTLs
Emp. Acc.	0.8143 (0.0036)	0.6683 (0.0053)
$\hat{\rho}_{before}(\theta^{\star})$	0.8086 (0.0001)	0.6597 (0.0001)
$\hat{\rho}_{before}(\hat{\theta}^{\star}_{LASSO})$	0.7635 (0.0017)	0.5354 (0.0031)
$\hat{\rho}_{before}(\hat{\theta}^{\star}_{ADLASSO})$	0.7627 (0.0012)	0.6488 (0.0017)
$\hat{\rho}_{before}(\hat{\theta}^{\star}_{GPLASSO})$	0.7581 (0.0014)	0.5471 (0.0029)
$\hat{\rho}_{after}(\theta^{\star})$	0.8045 (0.0019)	0.6576 (0.0021)
$\hat{\rho}_{after}(\hat{\theta}^{\star}_{LASSO})$	0.7502 (0.0026)	0.5347 (0.0037)
$\hat{\rho}_{after}(\hat{\theta}^{\star}_{ADLASSO})$	0.7489 (0.0023)	0.6454 (0.0027)
$\hat{\rho}_{after}(\hat{\theta}^{\star}_{GPLASSO})$	0.7479 (0.0024)	0.5495 (0.0034)

Predictive ability : ρ

There are several predictive ability estimators

- assuming the mixed model is correct
  - based on coefficient of determination (CD) :

$$\hat{\rho}^{\text{CD}} = \frac{h}{n_{\text{test}}} \sum_i \sqrt{\text{CD}(g_{\text{test},i})}$$

  - based on prediction error variance (PEV) :

$$\hat{\rho}^{\text{PEV}} = \frac{h}{n_{\text{test}}} \sum_i \sqrt{1 - \frac{\text{PEV}(g_{\text{test},i})}{\sigma_g^2}}$$

- assuming a fixed linear QTL model is known and correct and using the mixed model as ”instrumental” model : the theoritical accuracy Rabier et al., PlosOne, 2016

Thank you for your attention.

A special thank to Fanny Bonnafous, Prune Pegot-Espagnet, Charles-Elie Rabier, Ellen Goudemand, Renaud Rincent

Equipe de Tanja Stadler (ETH Zurich)

Statistique Bayésienne dans le cadre d'un réseau

- N : réseau phylogénétique
- X_i : alignements pour le SNP i
- G_i : arbre de gènes pour le SNP i
- m SNPs

$$\mathbb{P}(N, G_1, \ldots, G_m | X_1, \ldots, X_m) \propto \left(\prod_{i=1}^{m} \mathbb{P}(X_i|G_i)\mathbb{P}(G_i|N) \right) \mathbb{P}(N)$$

Calcul de l'a priori P(N) par un processus de naissance/hybridation

$\Rightarrow$ Markov Chain Monte Carlo afin d'estimer la distribution à posteriori de $\mathbb{P}(N, G_1, \ldots, G_m | X_1, \ldots, X_m)$.

Ils n'intègrent pas sur tous les arbres de gènes

Zhang et al (MBE, Décembre 2017)

Equipe de Luay Nakhleh (Université de Rice, USA)

- Wen et al. (Plos Genetics, 2016)
  - Arbres de gènes inférés lors d'un étape préliminaire
  - Données = arbres de gènes !!!

$$\mathbb{P}(N|G_1,\ldots,G_m) \propto \left(\prod_{i=1}^{m} \mathbb{P}(G_i|N)\right) \mathbb{P}(N)$$

- Zhu et al. (Plos Computational Biology, Janvier 2018)
  - Données = alignements
  - Intégration sur tous les arbres de gènes ...
  - Algorithme inspiré de SNAPP mais très couteux en temps de calcul

$$\mathbb{P}(N|X_1,\ldots,X_m) \propto \left(\prod_{i=1}^{m} \int_{\psi} \mathbb{P}(X_i|G_i)\mathbb{P}(G_i|S)dG_i\right) \mathbb{P}(N)$$