Sélection génomique: quelles données, comment les traiter ... ou l'alphabet Bayésien des généticiens

Helene Muranty, Leopoldo Sanchez Rodriguez

Sélection génomique : quelles données, comment les traiter … ou l'alphabet Bayésien des généticiens

Hélène Muranty, Leopoldo Sanchez

Unité Amélioration, Génétique et Physiologie Forestières

14 ème journée "CaSciModOT"
30 juin 2011 - Tours

ALIMENTATION
AGRICULTURE
ENVIRONNEMENT

INRA

Plan

> Notions de génétique quantitative

> Données de génotypage

> Structure de populations

> Traitement des données : évaluer les valeurs génétiques des candidats à la sélection

Les modèles de la génétique quantitative

ce qu'on mesure

une forme particulière d'un génome

Phénotype = Génotype x Environnement

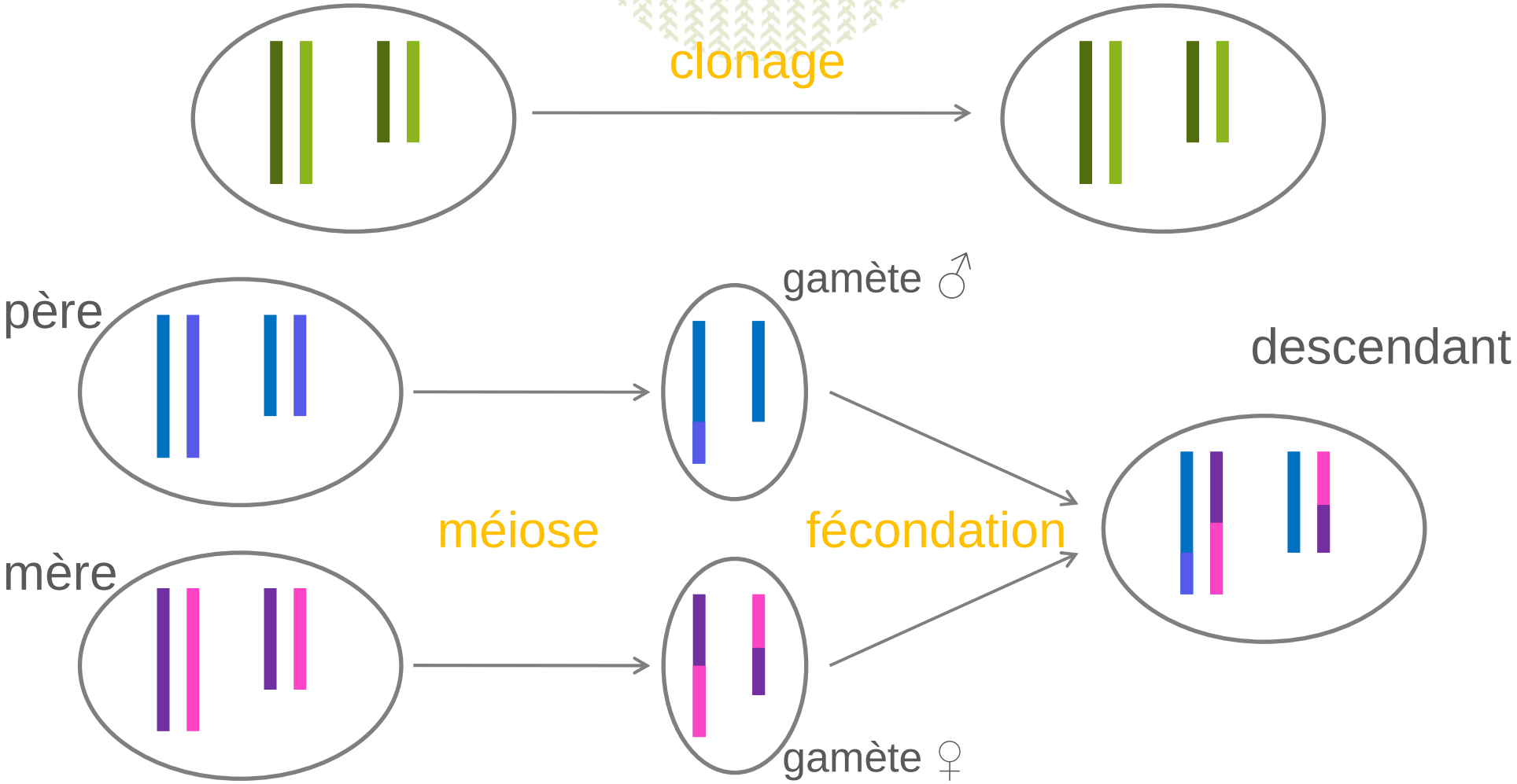
transmission par clonage

Phénotype = Génotype + Environnement

Génotype = Additivité + Dominance + Epistasie

transmission par descendance

Les bases de la génétique



Les bases de la génétique quantitative

> valeur génétique : effet moyen d'un génotype sur un caractère

> effet moyen d'un allèle : espérance de la valeur génétique quand l'allèle est présent dans le génotype

> valeur génétique additive d'un génotype : somme des effets moyens des allèles qui le constituent

> un parent transmet, en espérance, la moitié de sa valeur génétique additive

$$A_I = \tfrac{1}{2}(A_P + A_M) + e_I$$

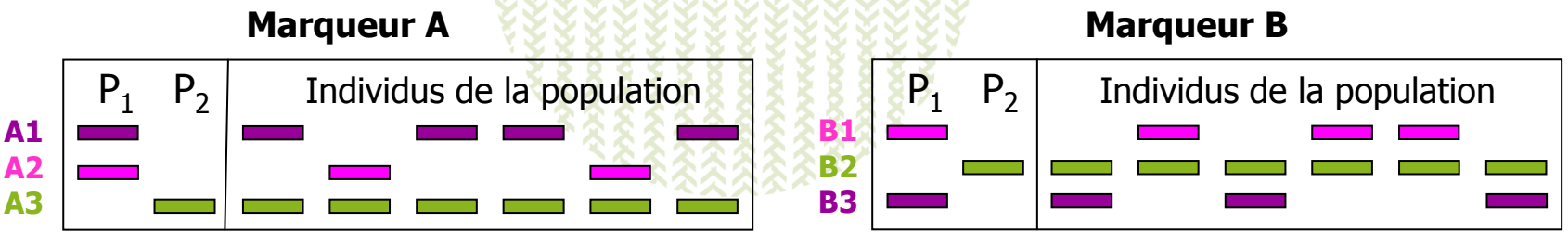
↖ **alea de méiose**

Sélection assistée par marqueurs (SAM)

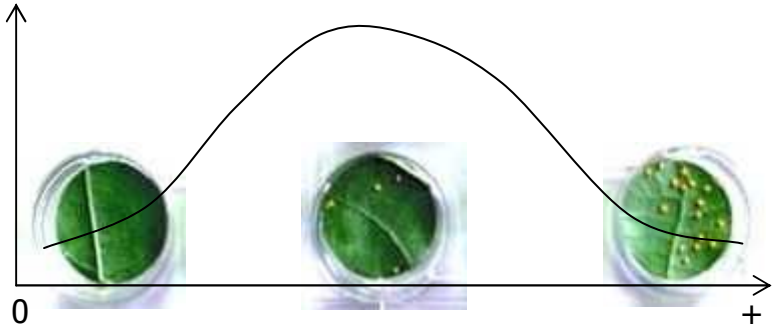
utiliser des marqueurs pour (mieux) prédire les valeurs génétiques (additives) des individus candidats à la sélection

$$A_i = \sum_{QTL} x_i q_q + e_i$$

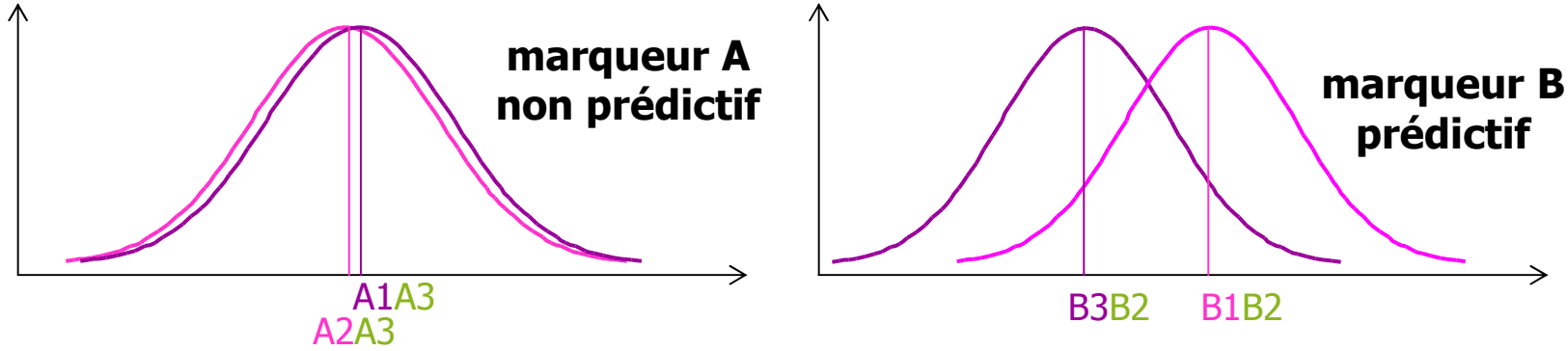
marqueur : caractère héritable qui permet de caractériser un génotype indépendamment de l'âge et de l'environnement



Caractère mesuré
dans la population
(ex : résistance à la rouille
chez le peuplier)



Distribution des performances en fonction du génotype au marqueur A ou B



Sélection assistée par marqueurs (SAM)

utiliser des marqueurs pour (mieux) prédire les valeurs génétiques (additives) des individus candidats à la sélection

$$A_i = \sum_{QTL} x_i q_q + e_i$$

Sélection génomique = suite de la SAM, avec

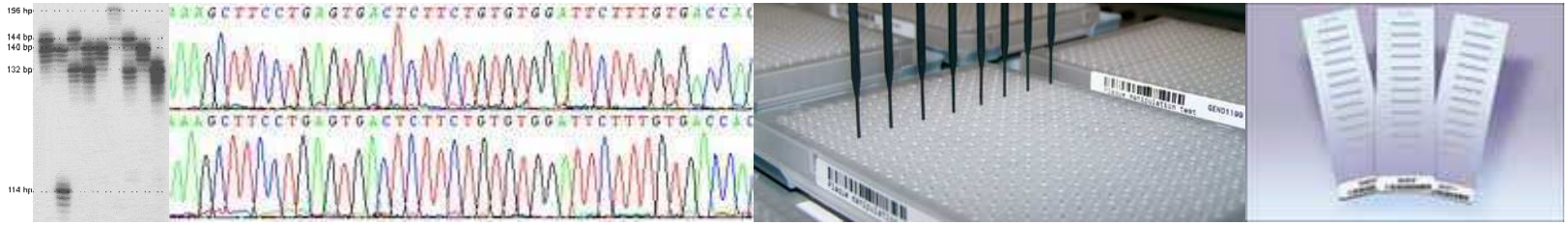
> plus de marqueurs,

> une autre façon d'utiliser le phénotypage

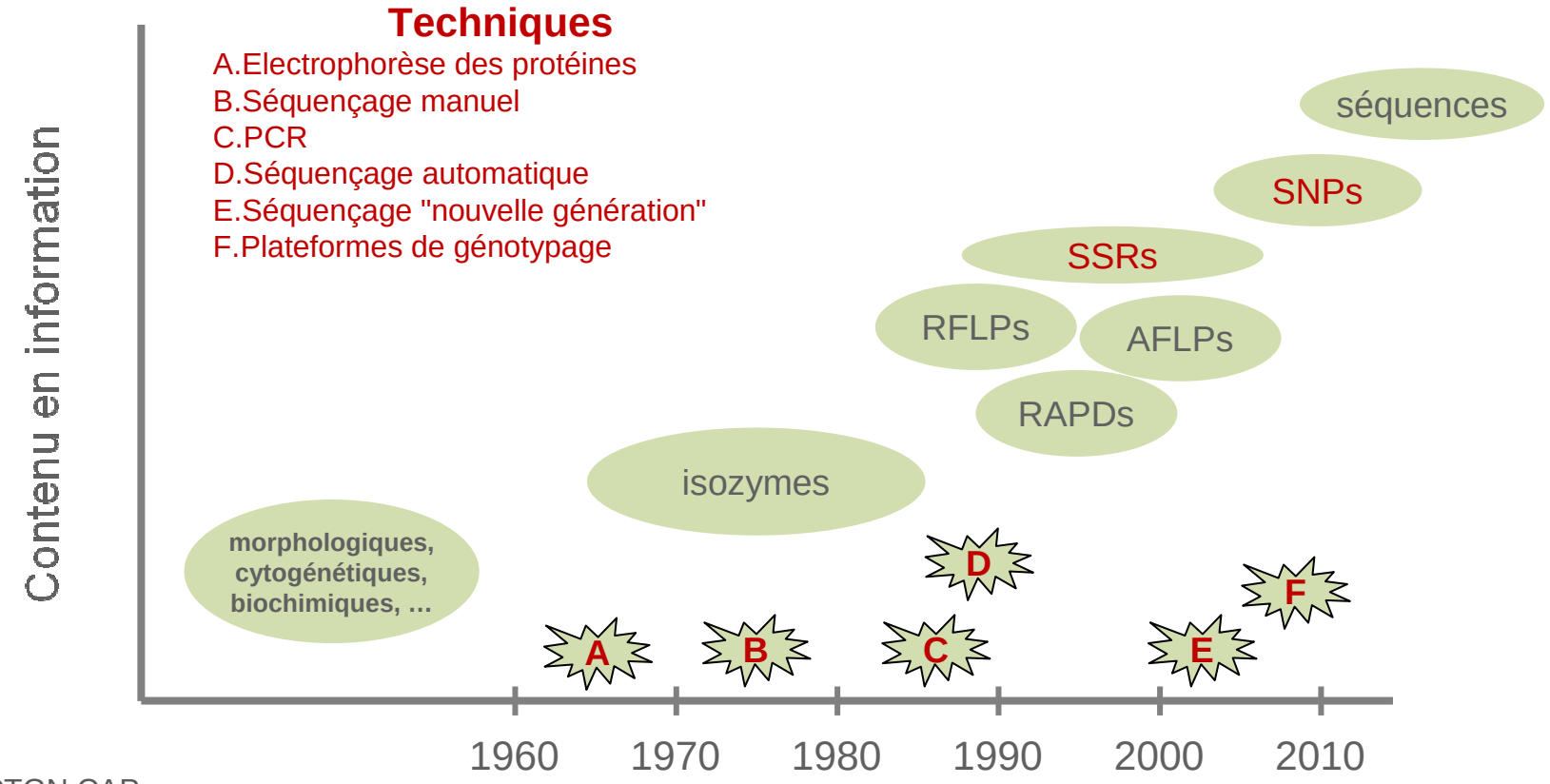
> des méthodes statistiques mieux adaptées

Meuwissen et al 2001

Données de génotypage



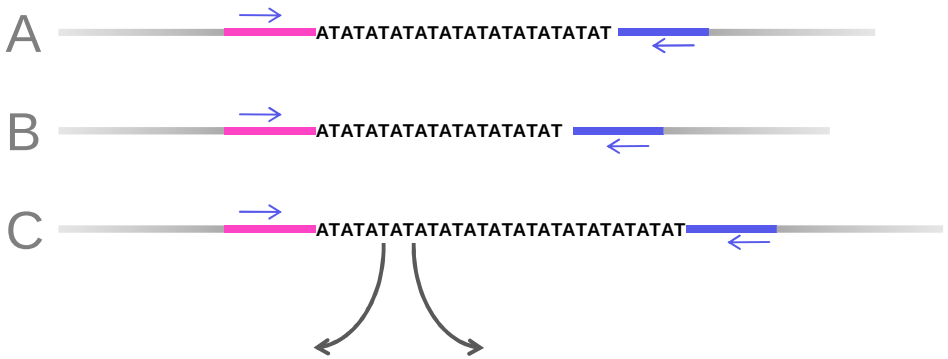
Histoire des techniques de marquage

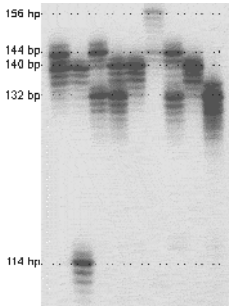


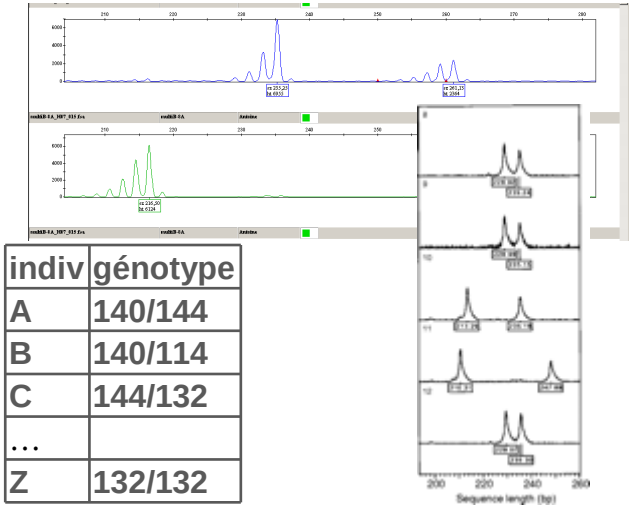
d'après CTGN CAP,
Genomics in tree breeding and forest ecosystems Short Course 2009

Polymorphisme observé et codage des données

Microsatellites / SSRs







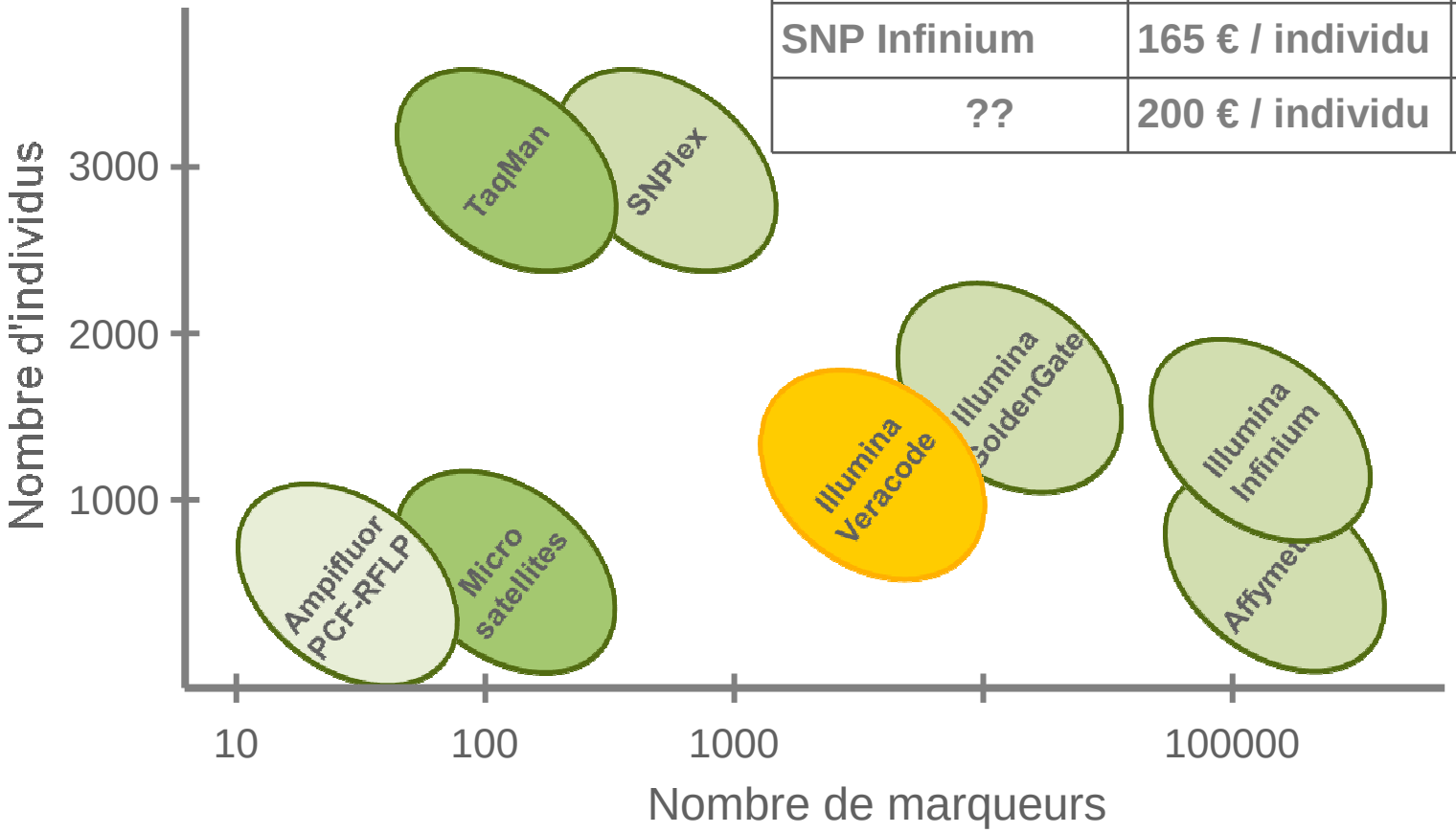
indiv	génotype
A	140/144
B	140/114
C	144/132
…	
Z	132/132

SNPs

indiv	génotype
A	AA = 0
B	AG = 1
C	GG = 2
…	
Z	GA = 1

Une idée des volumes de données en génotypage

Microsatellites	150 € / individu	110 marqueurs
SNP GoldenGate	150 € / individu	4500 SNP
SNP Infinium	165 € / individu	7600 SNP
??	200 € / individu	1 million de SNP



d'après D. Milan, INRA Toulouse, 2008

Puces de génotypage SNP commerciales

> bovins : 54 k SNP -> 777 k SNP

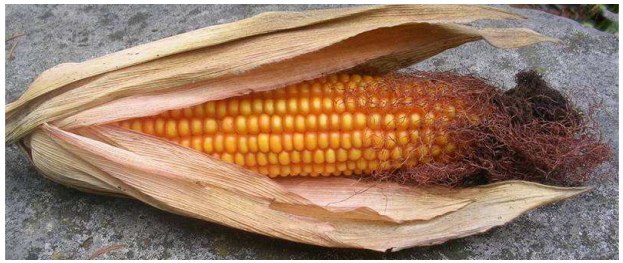
> porcs : 62 k SNP

> ovins : 54 k SNP

> poulets : 57 k SNP

> riz : 44 k SNP

> maïs : 55 k SNP



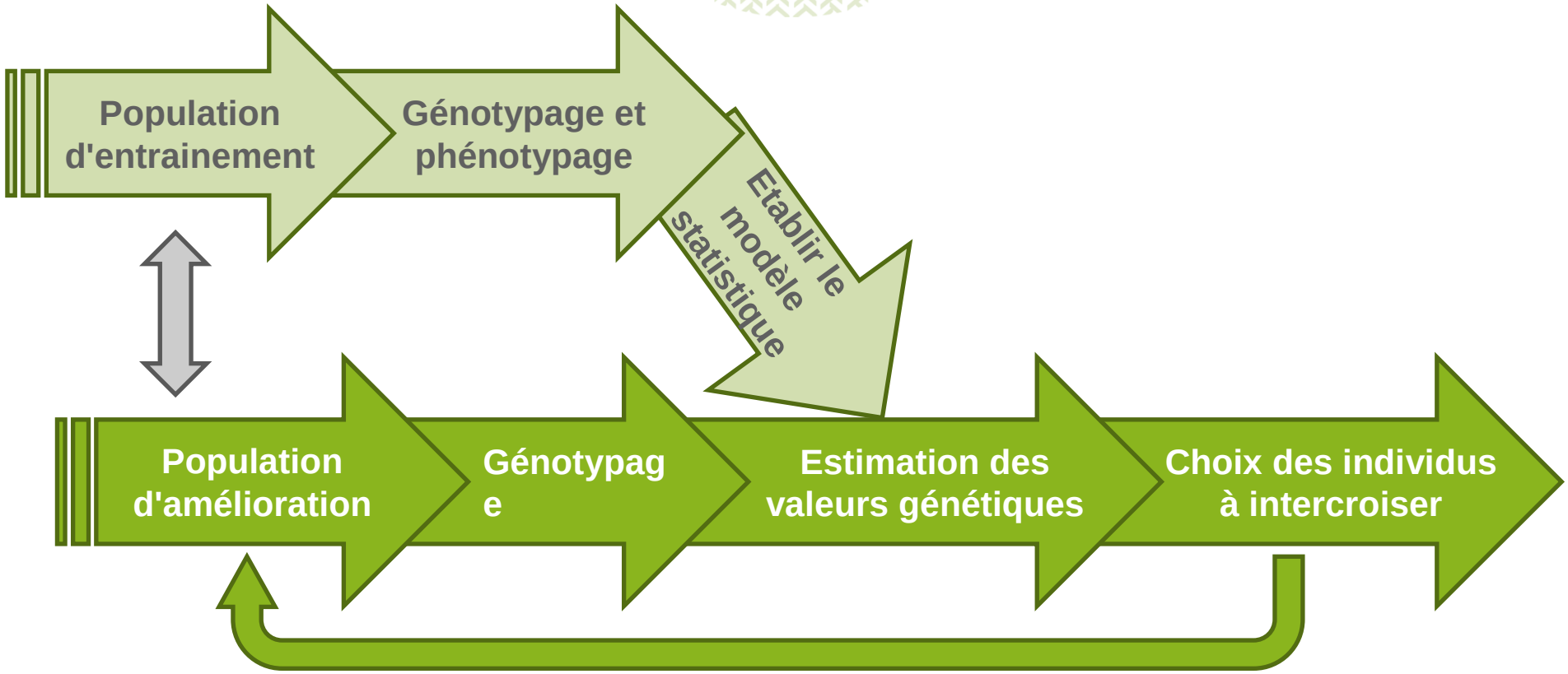
Structure de populations

14 ème journée "CaSciModOT"
30 juin 2011 - Tours

ALIMENTATION
AGRICULTURE
ENVIRONNEMENT

INRA

Le modèle de la sélection génomique



d'après T. Albrecht *et al.*,
TAG online 20/04/2011

1 testeur

29 lignées

36 familles

HD

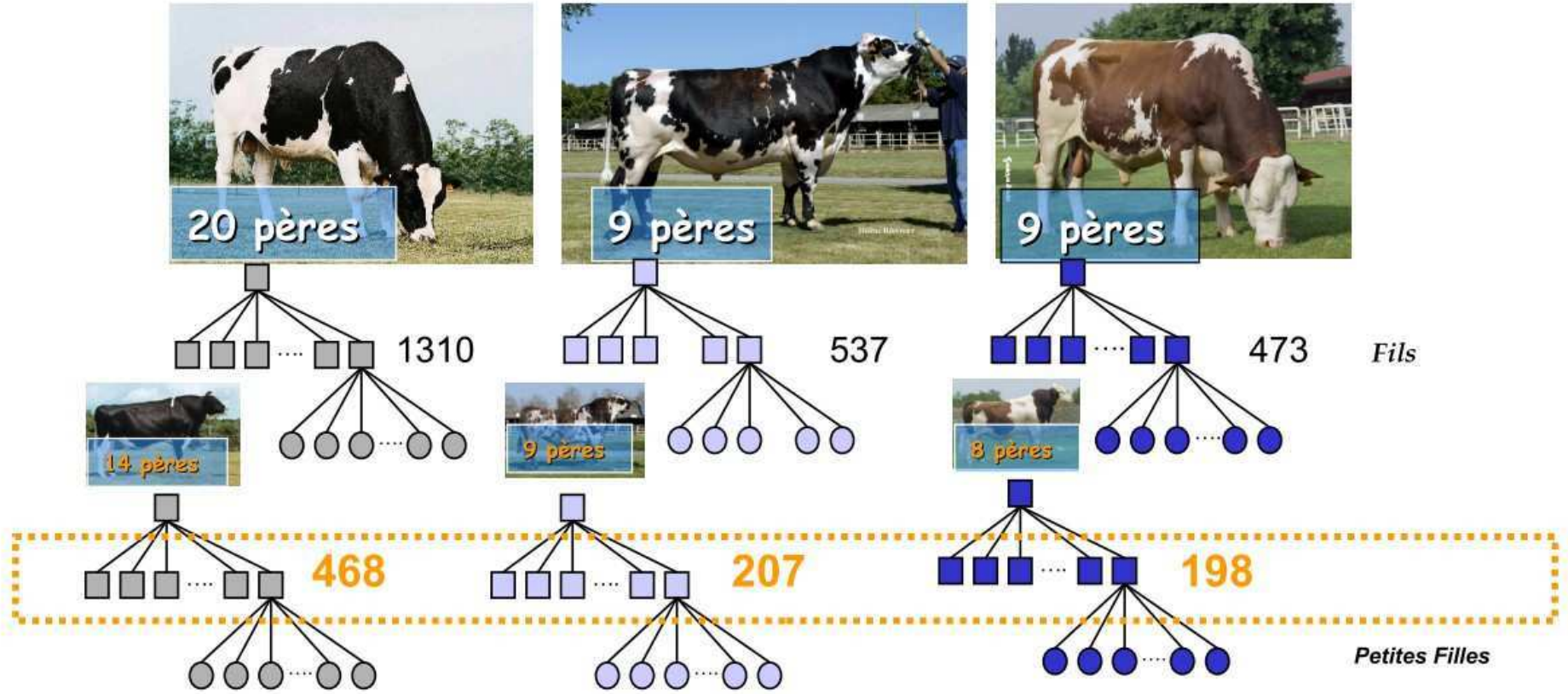
population d'entrainement 1380 lignées

1380 hybrides évalués en 7 lieux pour rendement et matière sèche

1152 marqueurs SNP

d'après F. Guillaume *et al.*,
Institut de l'Elevage, 2009

Test de la sélection génomique en bovins laitiers en France



> environ 3200 taureaux génotypés pour 54k SNP
> évalués sur descendance pour 25 caractères / production laitière

Traitement des données

DOI: 10.1534/genetics.109.103952

Additive Genetic Variability and the Bayesian Alphabet

Daniel Gianola,*,†,‡,1 Gustavo de los Campos,* William G. Hill,§ Eduardo Manfredi‡ and Rohan Fernando**

Full conditional for single-site Gibbs

Full conditionals are the same as in the "Normal" model for μ, α_j, and σ^2_e. Let

$$\boldsymbol{\xi} = [\sigma^2_1, \sigma^2_2, \ldots, \sigma^2_k]$$

Full conditional conditional for σ^2_j:

$$f(\sigma^2_j|\mathbf{y}, \mu, \boldsymbol{\alpha}, \boldsymbol{\xi}_{j\_}, \sigma^2_e) \propto f(\mathbf{y}, \mu, \boldsymbol{\alpha}, \boldsymbol{\xi}, \sigma^2_e)$$

$$\propto f(\mathbf{y}|\mu, \boldsymbol{\alpha}, \boldsymbol{\xi}, \sigma^2_e) f(\alpha_j|\sigma^2_j) f(\sigma^2_j) f(\mu, \boldsymbol{\alpha}_{j\_}, \boldsymbol{\xi}_{j\_}, \sigma^2_e)$$

$$\propto (\sigma^2_j)^{-1/2} \exp\{-\frac{\alpha^2_j}{2\sigma^2_j}\} (\sigma^2_j)^{-(2+\nu_\alpha)/2} \exp\{\frac{\nu_\alpha S^2_\alpha}{2\sigma^2_j}\}$$

$$\propto (\sigma^2_j)^{-(2+\nu_\alpha+1)/2} \exp\{\frac{\alpha^2_j + \nu_\alpha S^2_\alpha}{2\sigma^2_j}\}$$

27/51

Etablir le modèle statistique

Y : phénotype de *n* individus

X : matrice des génotypes des *n* individus pour *p* marqueurs

p >> n

$$Y = \mu + f(X) + \varepsilon$$

modèles additifs
Bayes A, B, C
LASSO
régression ridge
...

apparentement global
matrice d'apparent^t
génomique
$G = (XX' - g_0) / g_1$

autres modèles
RKHS regression
...

Modèles additifs

$$Y = \mu + \sum_{j=1}^{p} x_j \alpha_j + \varepsilon$$

ridge regression
BLUP

$$\left(\alpha_j \middle| \sigma^2\right) \sim N\left(0, \sigma^2\right)$$

R ?

Bayes A

$$\left(\alpha_j \middle| \sigma_j^2\right) \sim N\left(0, \sigma_j^2\right)$$

$$\sigma_j^2 \sim \nu_\alpha S_{\nu_\alpha}^2 \chi_{\nu_\alpha}^{-2}$$

A

$$\Leftrightarrow$$

$$\alpha_j \sim t\left(0, S_{\nu_\alpha}^2, \nu_\alpha\right)$$

Bayes B

$$\left(\alpha_j \middle| \pi, \sigma_j^2\right) \begin{cases} \sim N\left(0, \sigma_j^2\right) & proba(1-\pi) \\ = 0 & proba\,\pi \end{cases}$$

B

$$\sigma_j^2 \sim \nu_\alpha S_{\nu_\alpha}^2 \chi_{\nu_\alpha}^{-2}$$

LASSO

$$\left(\alpha_j \middle| \sigma_j^2\right) \sim N\left(0, \sigma_j^2\right)$$

$$\sigma_j^2 \sim Exp\left(\lambda^2\right)$$

L

$$\lambda^2 \sim \Gamma(\delta, r)$$

Bayes Cπ

$$\left(\alpha_j \middle| \pi, \sigma_j^2\right) \begin{cases} \sim N\left(0, \sigma_j^2\right) & proba(1-\pi) \\ = 0 & proba\,\pi \end{cases}$$

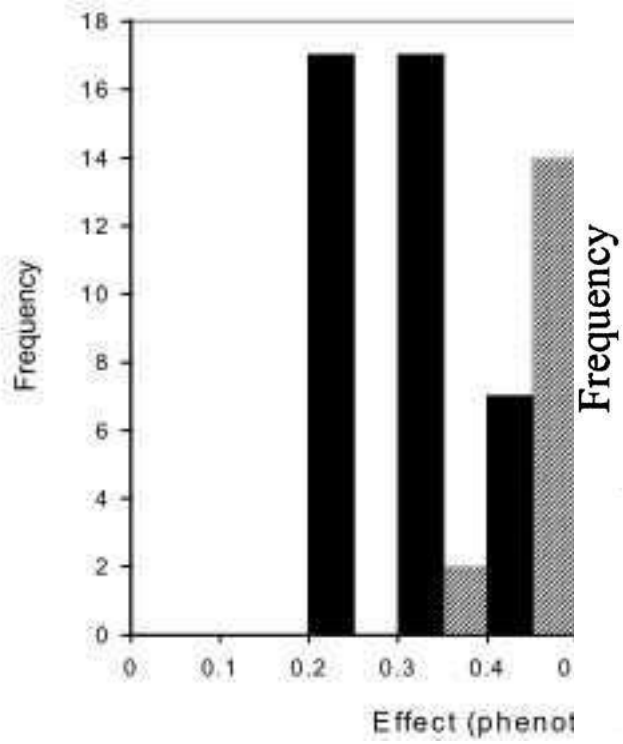
C

$$\sigma_j^2 \sim \nu_\alpha S_{\nu_\alpha}^2 \chi_{\nu_\alpha}^{-2}$$

$$\pi \sim Uniforme(0,1)$$

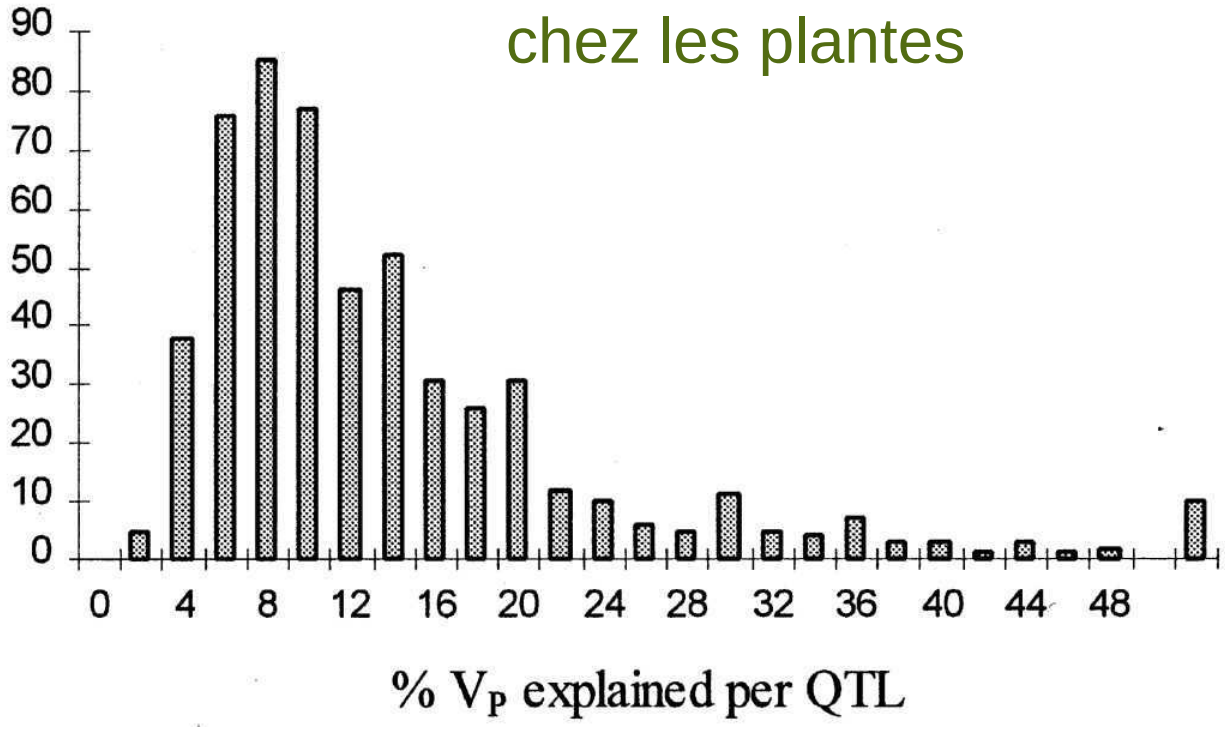
Justification des distributions de mélange Bayes B et Bayes Cπ

chez les animaux



d'après Hayes et Goddard (2001)

chez les plantes



d'après Kearsey et Farquhar (1998)

Apparentement global

modèle animal "classique"

$$Y = \mu + Zu + \varepsilon$$

effet "animal"

$$Var(u) = \mathbf{A}\,\sigma^2_a$$

matrice d'apparentement calculée à partir du pedigree

modèle GBLUP

$$Y = \mu + Zg + \varepsilon$$

valeur génétique

$$Var(g) = \mathbf{G}\,\sigma^2_a$$

matrice de similarité calculée avec les marqueurs

x_i=0/1/2 $w_i = x_i - 2f_i$

$$\mathbf{G} = \frac{\mathbf{WW'}}{2\sum f_i(1-f_i)}$$

ou

D diagonale

$$\mathbf{D}_{ii} = \frac{1}{p[2f_i(1-f_i)]}$$

G=WDW'

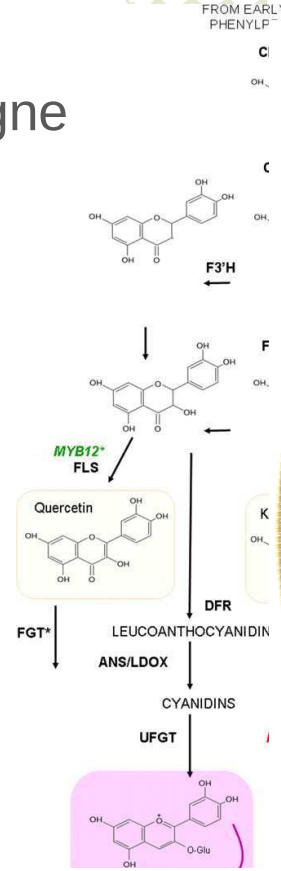
d'après VanRaden (2008)

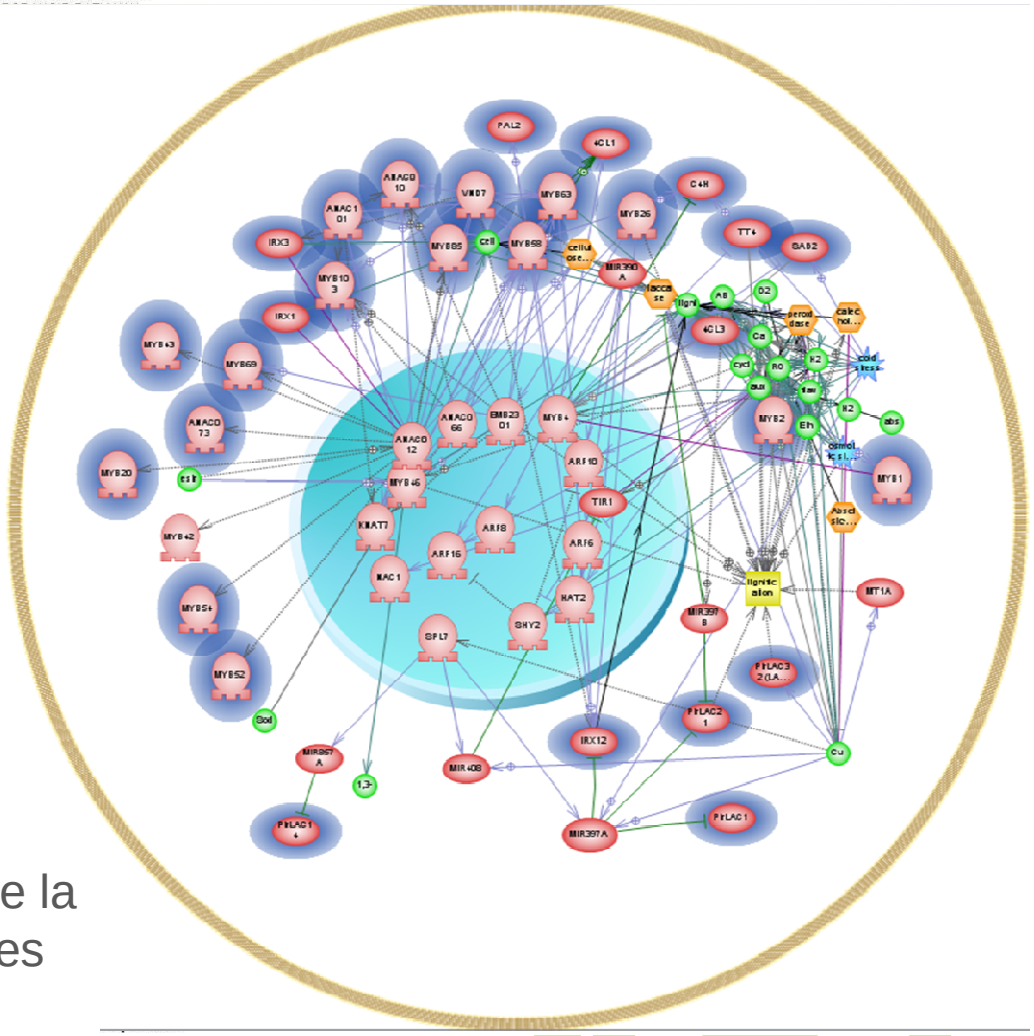
Besoin de modèles non-additifs

Voie métabolique des anthocyanines chez la vigne

ENZYMES

RÉGULATEURS





réseau de régulation de la biosynthèse des lignines
d'après JC Leplé

"Ecole Gianola"

Autres modèles

reproducing kernel Hilbert spaces (RKHS) regression

$Y = \mu + Zg + \varepsilon$

$g \sim N(0, \mathbf{K}\sigma^2_g)$

K matrice "noyau"

$$\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\varphi \times d_{ij})$$

avec $d_{ij} = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2$

Bayesian nonparametric "regression" with Dirichlet process priors

	BB			Bb			bb		
	CC	Cc	cc	CC	Cc	cc	CC	Cc	cc
AA	5	3	3	3	2	2	3	2	2
Aa	3	2	2	2	1	1	2	1	1
aa	2	1	1	1	0	0	1	0	0

Bayesian regularization of artificial neural networks (BRANN)

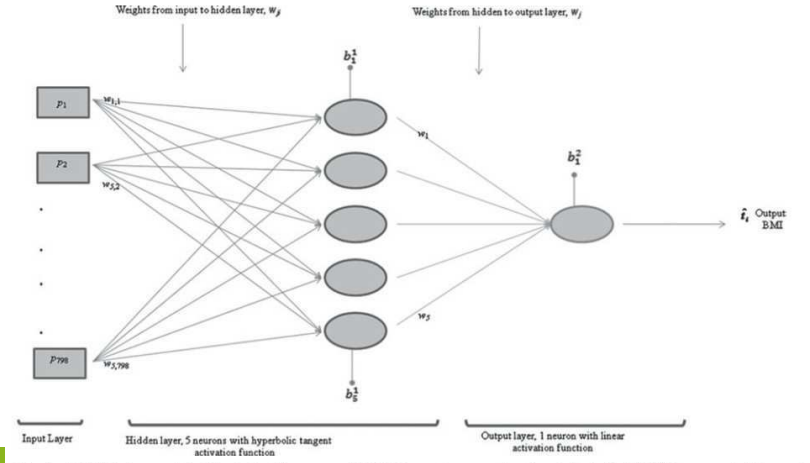


Fig. 1. ANN design used in this study. There were 798 SNP genotypes used as inputs (p_{ij}). Each SNP is connected to up to five neurons via coefficients w_{jk} (j denotes neuron, k denotes SNP). Each hidden and output neuron has a bias parameter b_j^{l0}, j denotes neuron, l denotes layer).

Machine learning procedures

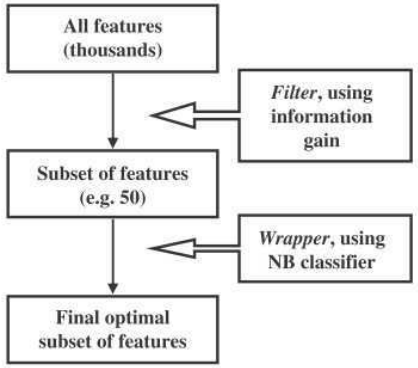


Figure 1 The combined filter + wrapper feature selection approach. The subset of features obtained from the filter step, having a size that is largely reduced from the original input set, is fed into a wrapper step, which uses a naï ve Bayesian (NB) classifier.

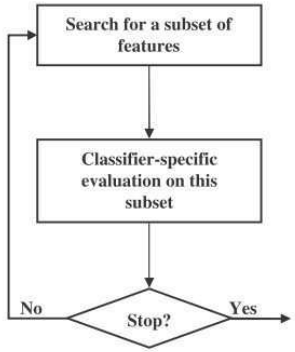


Figure 2 Outline of the wrapper selection procedure. The usefulness of a newly found feature subset is measured by a predefined classifier. If the stopping criterion (usually a threshold of prediction accuracy) is met, search stops and the subset is accepted. Otherwise, a new search is conducted.

Comparer les modèles statistiques

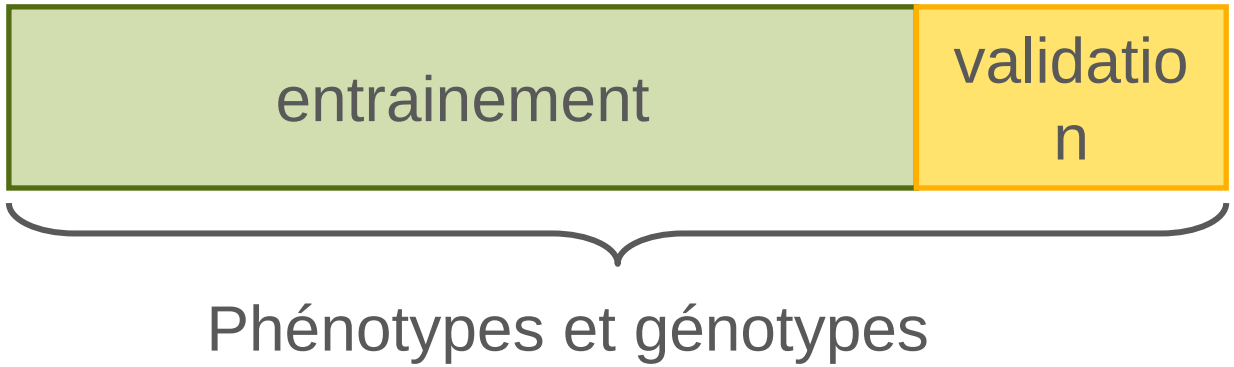
> Simulations

Comparer valeurs génétiques prédites / simulées

Problème : imaginer le modèle génétique

> Vraies données

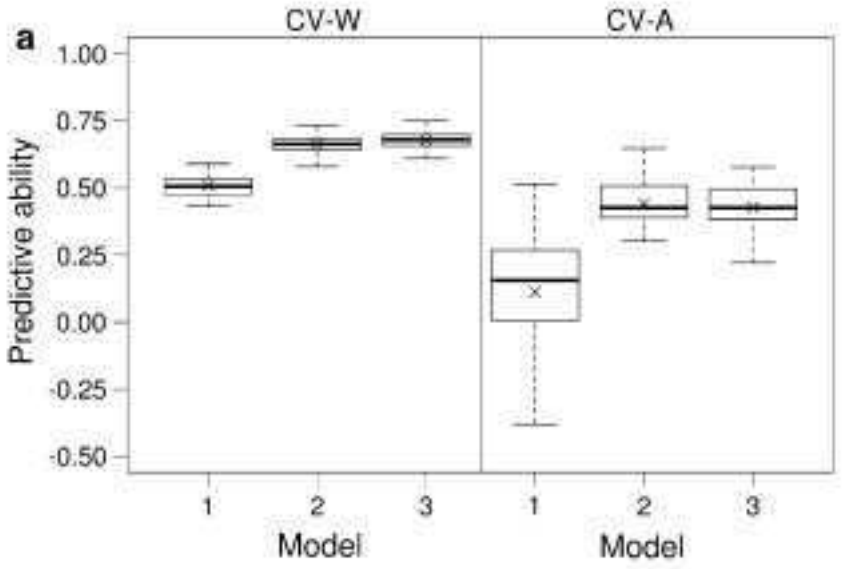
Validation croisée : partager population d'entraînement en deux parties : une / entraînement, autre/validation



Comparer les modèles statistiques

> Validation croisée

d'après T. Albrecht *et al.*,
TAG online 20/04/2011



1 : pedigree seulement
2: apparentement global par les marqueurs
3: 1 + 2

CV-W : échantillonnage intra-famille
CV-A : échantillonnage entre familles

En pratique, quels outils ?

> des bibliothèques sous 

- BLR (de los Campos et Perez Rodriguez, 2010) : ridge regression, LASSO
- glmnet (Friedmann et al, 2011) : LASSO, Elastic net GLM
- mixOmics (Dejean et al, 2011) : sparse PLS, …

> d'autres programmes

- GS3 (Legarra et al, 2010) : Bayes Cπ, GBLUP

et sûrement pleins d'autres encore …

Conclusion, perspectives

> une nouvelle façon d'envisager la sélection => révolution dans certaines filières (bovins laitiers)

> volumes de données : augmentations à prévoir

- **nombre d'individus : consortiums internationaux**
- **marqueurs : issus du séquençage direct**

> traitement des données : modèles

- **modèles additifs : plus de 2 classes d'effets**
- **modèles non additifs**

> traitement des données : temps de calcul

- **modèles bayésiens : approximations pour réduire temps de calcul / MCMC**

Merci de votre attention

14 ème journée "CaSciModOT"
30 juin 2011 - Tours

ALIMENTATION
AGRICULTURE
ENVIRONNEMENT