



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





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

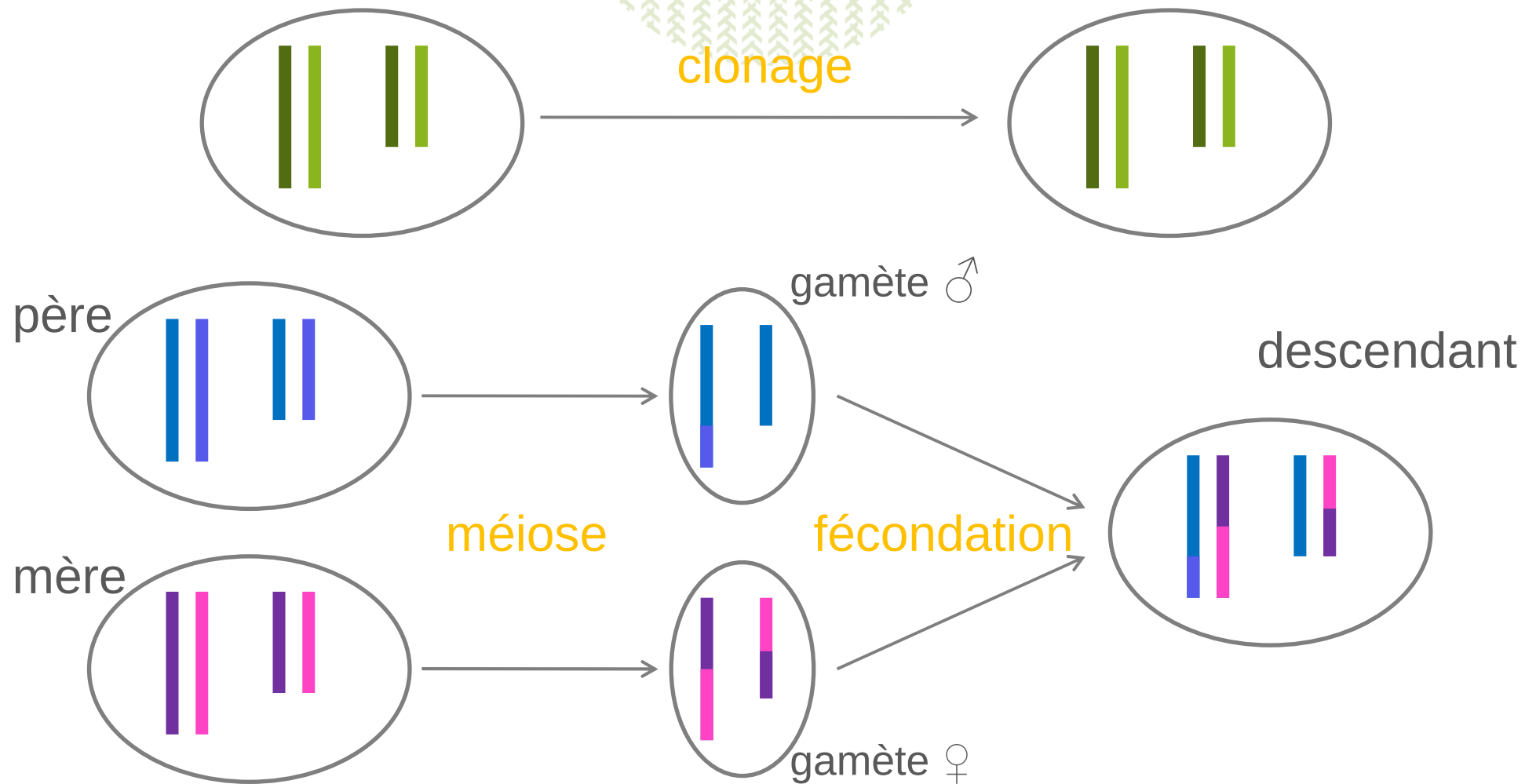
transmission par clonage

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

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

↖ alea de méiose

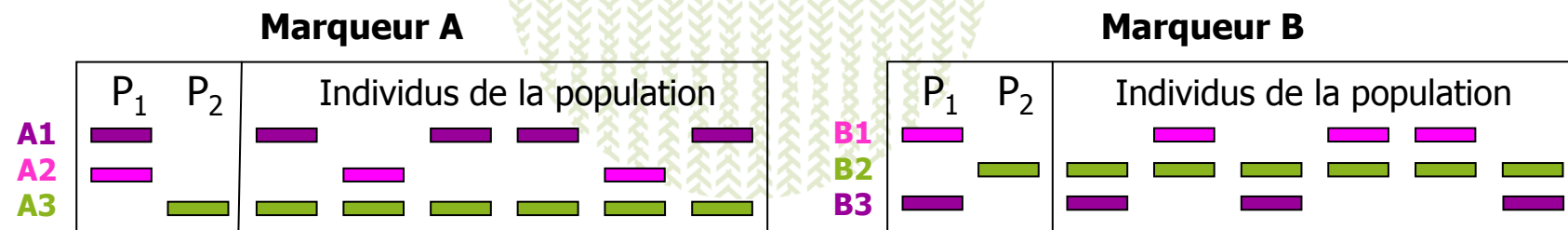


Sélection assistée par marqueurs (SAM)

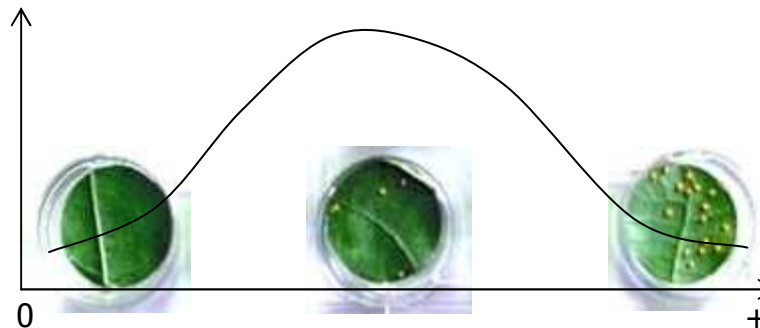
utiliser des marqueurs pour (mieux) prédire les valeurs génétiques (additives) des individus candidats à la sélection

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

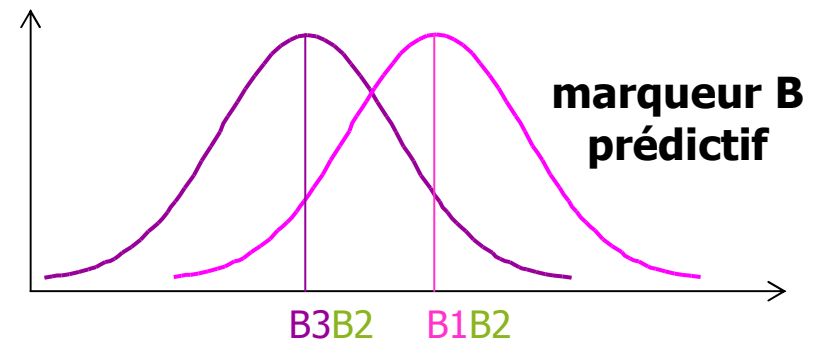
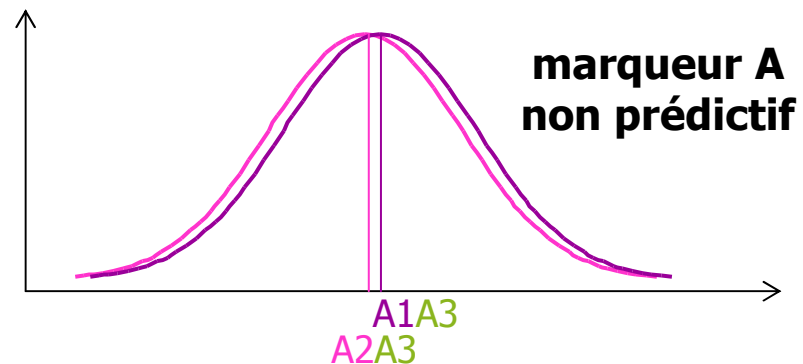
marqueur : caractère héritable qui permet de caractériser un génotype indépendamment de l'âge et de l'environnement



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





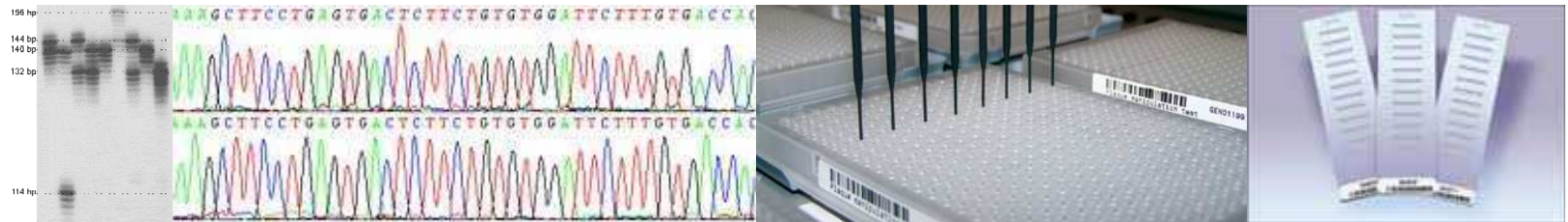
Sélection assistée par marqueurs (SAM)
utiliser des marqueurs pour (mieux) prédire les
valeurs génétiques (additives) des individus
candidats à la sélection

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

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

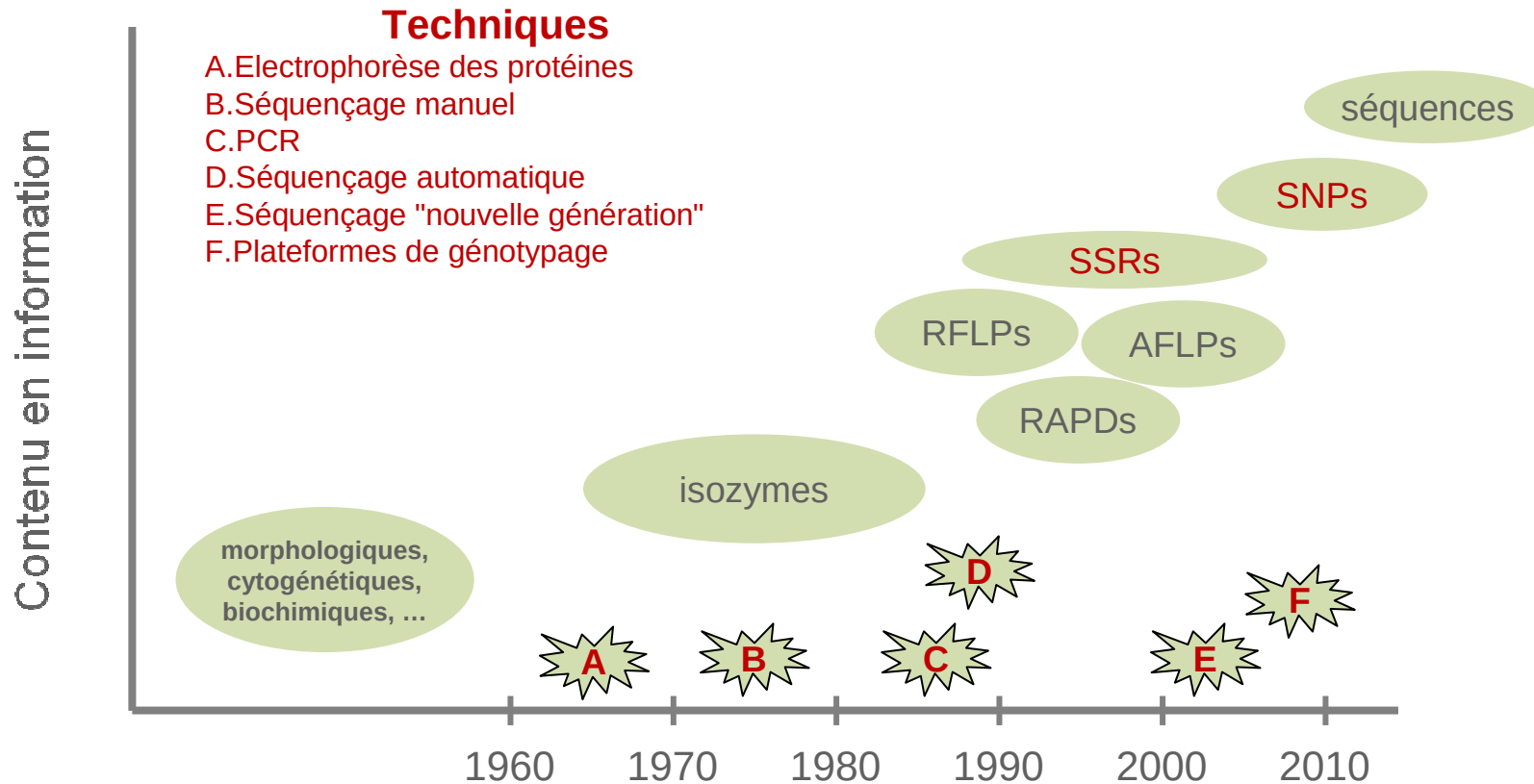


14 ème journée "CaSciModOT"
30 juin 2011 - Tours

ALIMENTATION
AGRICULTURE
ENVIRONNEMENT



Histoire des techniques de marquage



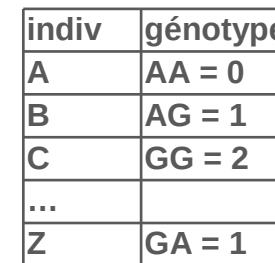
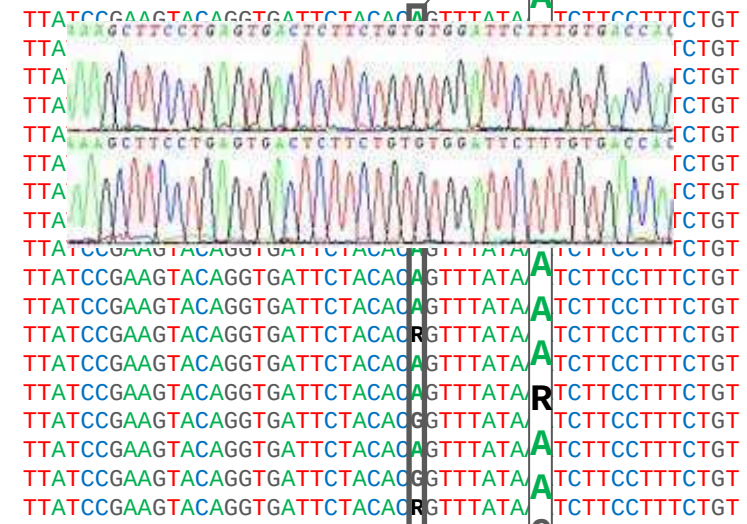
d'après CTGN CAP,
Genomics in tree breeding and forest
ecosystems Short Course 2009

14 ème journée "CaSciModOT"
30 juin 2011 - Tours

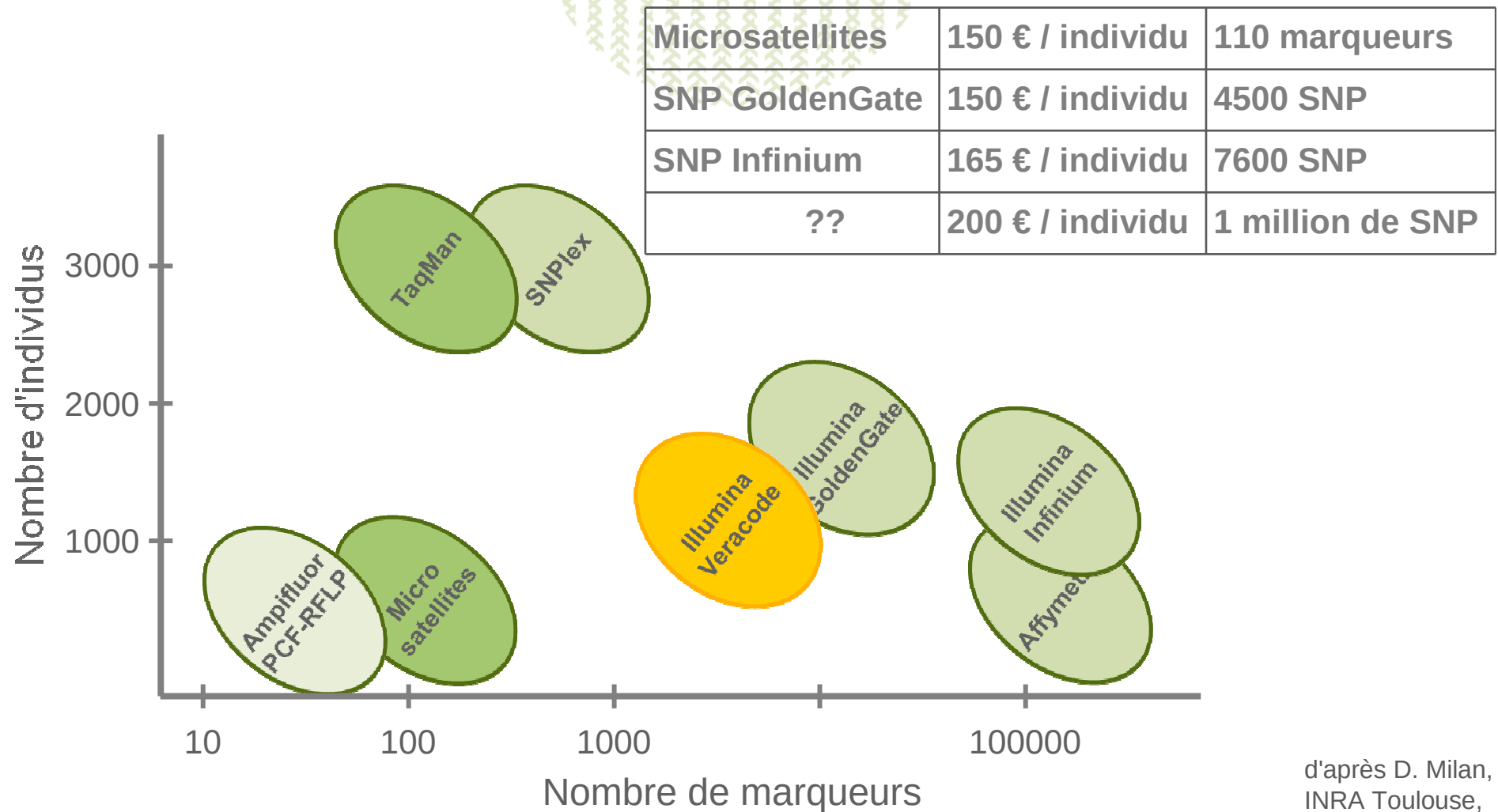
ALIMENTATION
AGRICULTURE
ENVIRONNEMENT



observé et codé



Une idée des volumes de données en génotypage



d'après D. Milan,
INRA Toulouse,
2008



Puces de génotypage SNP commerciales

> bovins : 54 k SNP -> 777 k SNP

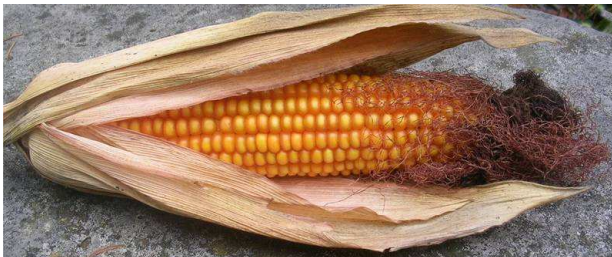
> porcs : 62 k SNP

> ovins : 54 k SNP

> poulets : 57 k SNP

> riz : 44 k SNP

> maïs : 55 k SNP





Structure de populations

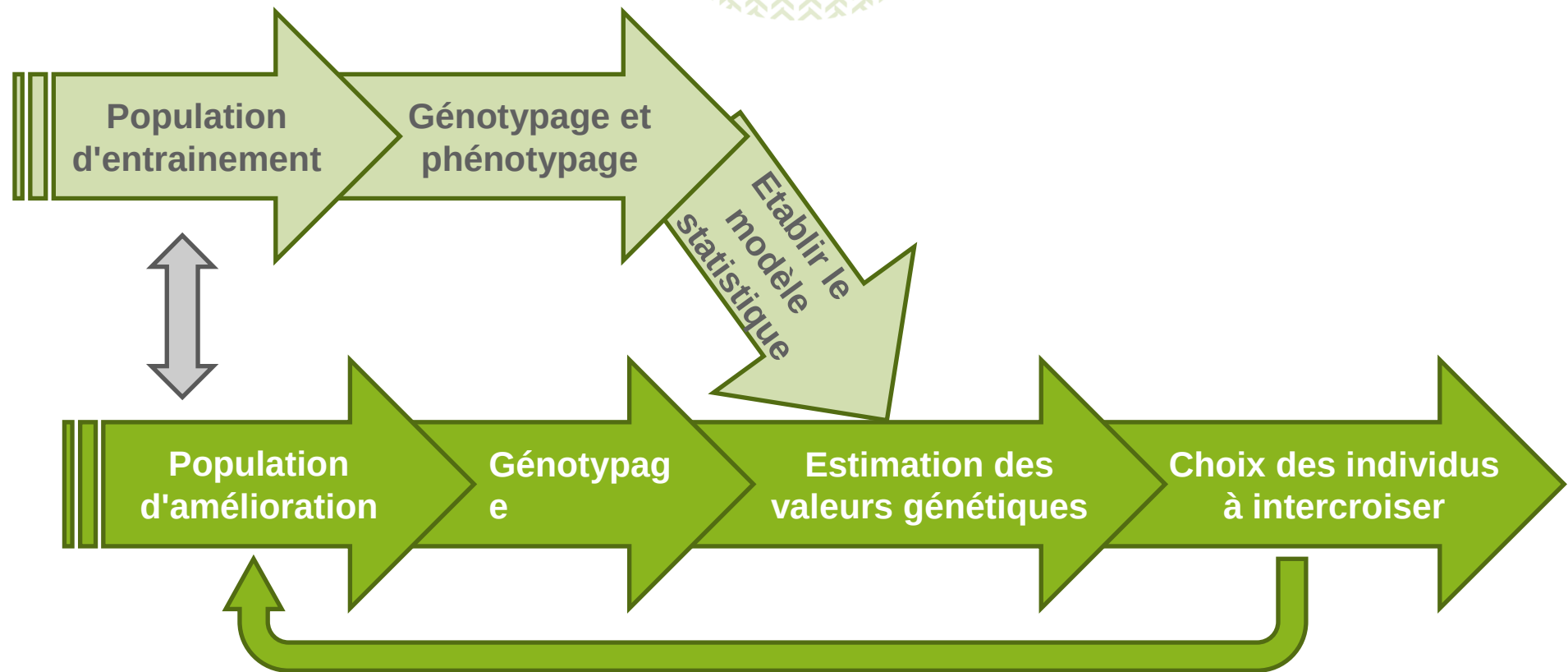


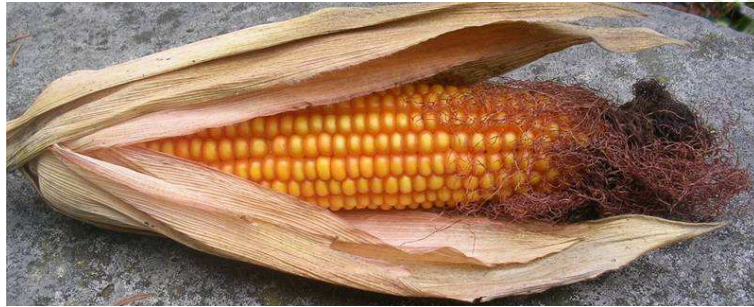
14 ème journée "CaSciModOT"
30 juin 2011 - Tours

ALIMENTATION
AGRICULTURE
ENVIRONNEMENT



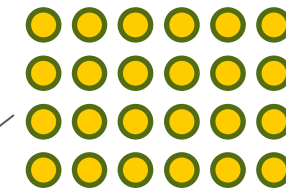
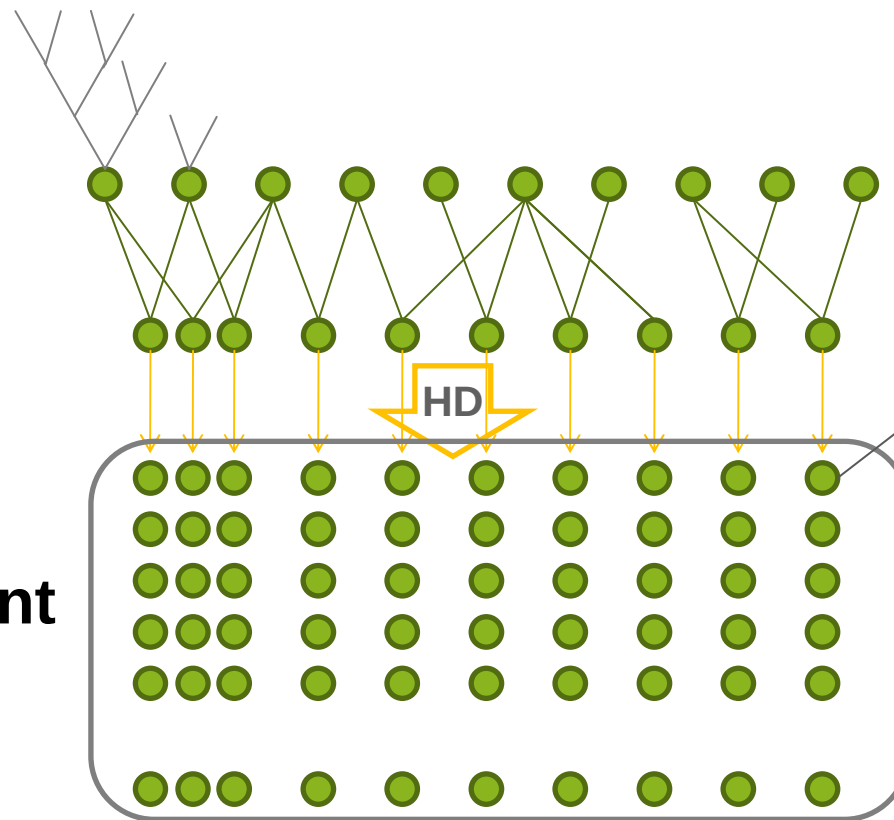
Le modèle de la sélection génomique





1 testeur ●

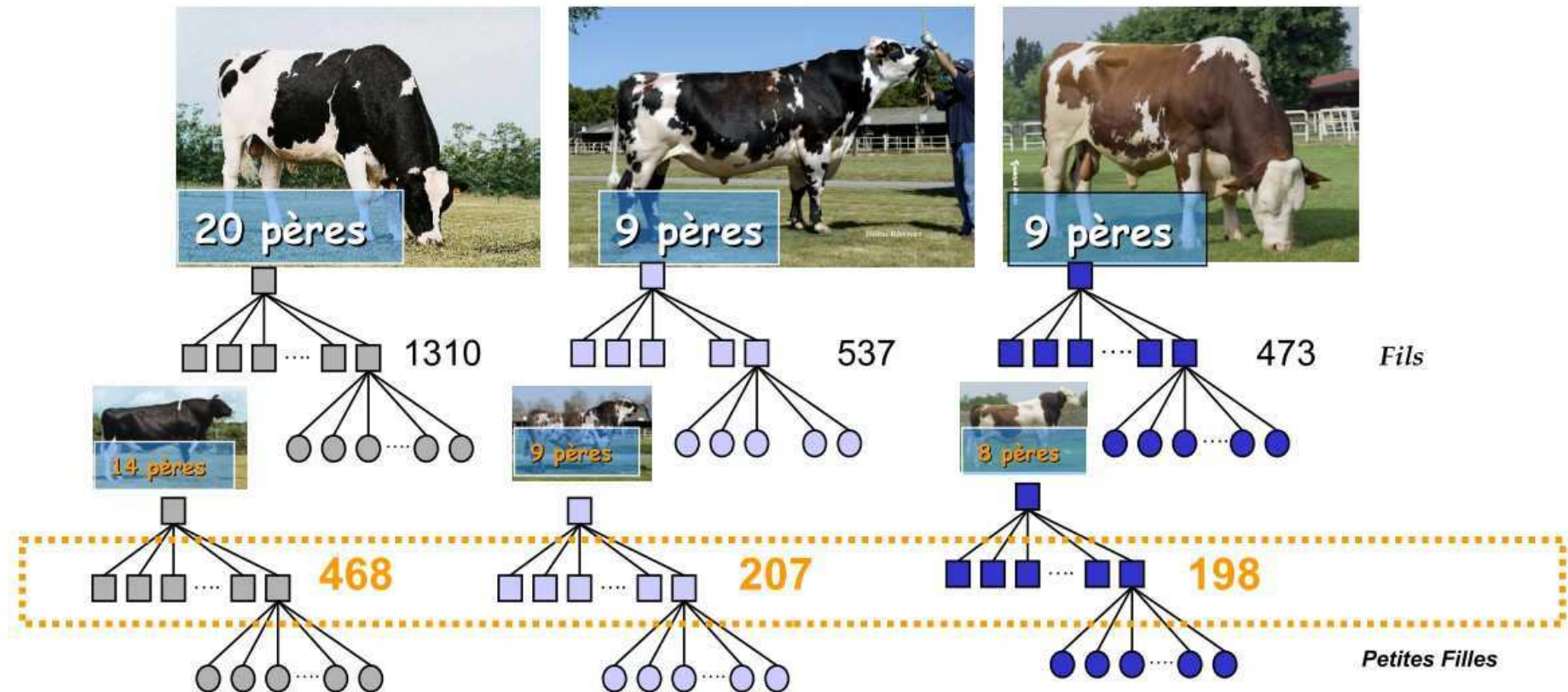
29 lignées
36 familles
**population
d'entraînement
1380 lignées**



1380 hybrides
évalués en 7 lieux
pour rendement
et matière sèche

1152 marqueurs SNP

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

Copyright © 2009 by the Genetics Society of America
DOI: 10.1534/genetics.109.103952

Additive Genetic Variability and the Bayesian Alphabet

Daniel Gianola,^{*,†,‡,§} 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_e^2 . Let

$$\xi = [\sigma_1^2, \sigma_2^2, \dots, \sigma_k^2]$$

Full conditional for σ_j^2 :

$$\begin{aligned} f(\sigma_j^2 | \mathbf{y}, \mu, \alpha, \xi_{-j}, \sigma_e^2) &\propto f(\mathbf{y}, \mu, \alpha, \xi, \sigma_e^2) \\ &\propto f(\mathbf{y} | \mu, \alpha, \xi, \sigma_e^2) f(\alpha_j | \sigma_j^2) f(\sigma_j^2) f(\mu, \alpha_{-j}, \xi_{-j}, \sigma_e^2) \\ &\propto (\sigma_j^2)^{-1/2} \exp\left\{-\frac{\alpha_j^2}{2\sigma_j^2}\right\} (\sigma_j^2)^{-(2+\nu_\alpha)/2} \exp\left\{\frac{\nu_\alpha S_\alpha^2}{2\sigma_j^2}\right\} \\ &\propto (\sigma_j^2)^{-(2+\nu_\alpha+1)/2} \exp\left\{-\frac{\alpha_j^2 + \nu_\alpha S_\alpha^2}{2\sigma_j^2}\right\} \end{aligned}$$

27/51

14 ème journée "CaSciModOT"
30 juin 2011 - Tours

ALIMENTATION
AGRICULTURE
ENVIRONNEMENT



Etablir le modèle statistique

Y : phénotype de n individus

X : matrice des génotypes des n individus pour p marqueurs

$p \gg n$

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

modèles additifs

Bayes A, B, C

LASSO

régression ridge

...

apparemment 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

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

LASSO

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

$$\sigma_j^2 \sim \text{Exp}(\lambda^2)$$

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

Bayes A

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

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

$\Leftrightarrow$

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

Bayes B

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

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

Bayes C π

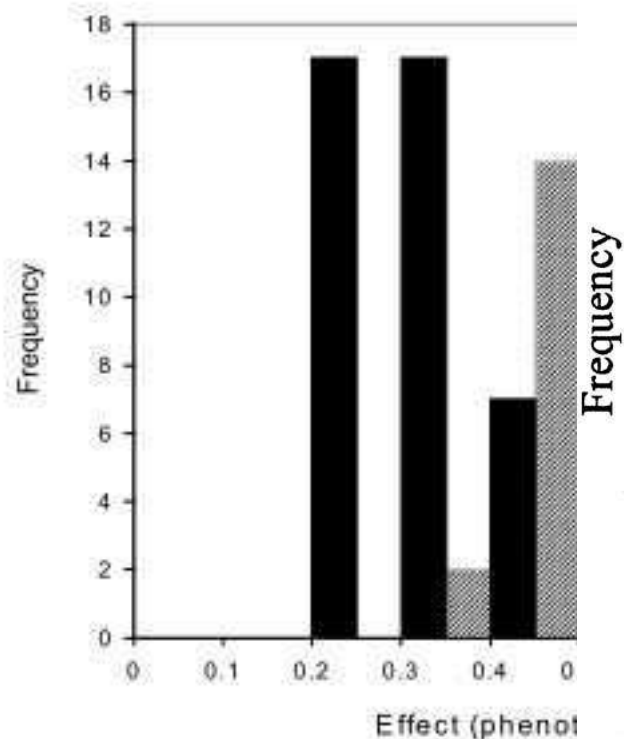
$$(\alpha_j | \pi, \sigma_j^2) \begin{cases} \sim N(0, \sigma_j^2) & \text{proba}(1-\pi) \\ \mathbf{C} = 0 & \text{proba } \pi \end{cases}$$

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

$$\pi \sim \text{Uniforme}(0, 1)$$

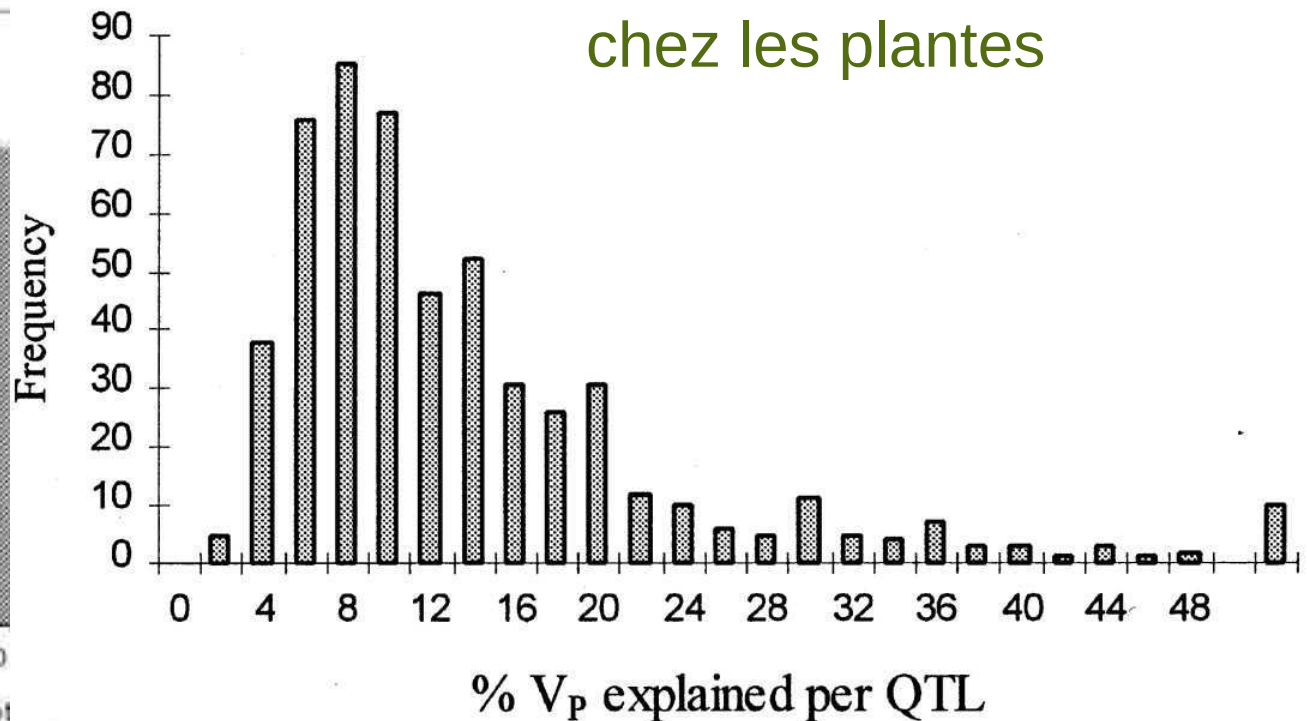
Justification des distributions de mélange Bayes B et Bayes C π

chez les animaux



d'après Hayes et Goddard (2001)

chez les plantes



d'après Kearsey et Farquhar (1998)

Apparemment global

modèle animal "classique"

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

effet "animal" 

$$\text{Var}(u) = \mathbf{A} \sigma_a^2$$

matrice d'apparentement
calculée à partir du pedigree

modèle GBLUP

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

valeur génétique 

$$\text{Var}(g) = \mathbf{G} \sigma_a^2$$

matrice de similarité
calculée avec les marqueurs

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

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

ou

$$\mathbf{D} \text{ diagonale}$$
$$D_{ii} = \frac{1}{p[2f_i(1-f_i)]}$$
$$\mathbf{G} = \mathbf{W}\mathbf{D}\mathbf{W}'$$

d'après VanRaden (2008)

le modèles nor

FROM EARLY PRODUCTS OF THE
PHENYL

C

OH₂

5

"Ecole Gianola"

Autres modèles

reproducing kernel Hilbert spaces
(RKHS) regression

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

$\mathbf{K}$ matrice "noyau"

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

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

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)

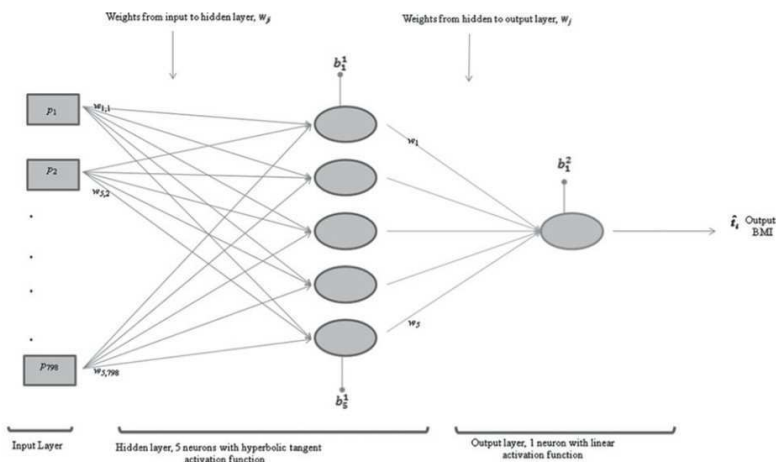


Fig. 1. ANN design used in this study. There were 798 SNP genotypes used as inputs (p_j). Each SNP is connected to up to five neurons via coefficients w_{jk} (j denotes neuron, k denotes SNP). Each hidden and output neuron has a bias parameter b_j^l , j denotes neuron, l denotes layer.

Machine learning procedures

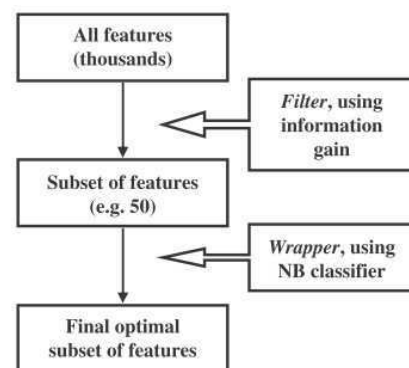


Figure 1 The combined filter + wrapper feature selection approach. The subset of features obtained from the filter step, having a size that is largely reduced from the original input set, is fed into a wrapper step, which uses a naïve Bayesian (NB) classifier.

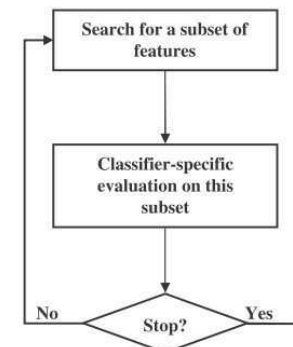


Figure 2 Outline of the wrapper selection procedure. The usefulness of a newly found feature subset is measured by a predefined classifier. If the stopping criterion (usually a threshold of prediction accuracy) is met, search stops and the subset is accepted. Otherwise, a new search is conducted.

14 ème journée "CaSciModOT"
30 juin 2011 - Tours

ALIMENTATION
AGRICULTURE
ENVIRONNEMENT



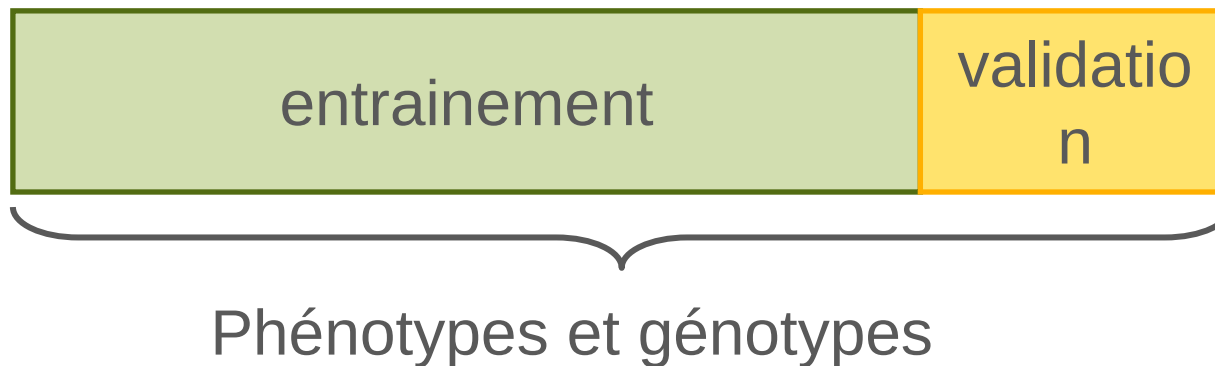
Comparer les modèles statistiques

> 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

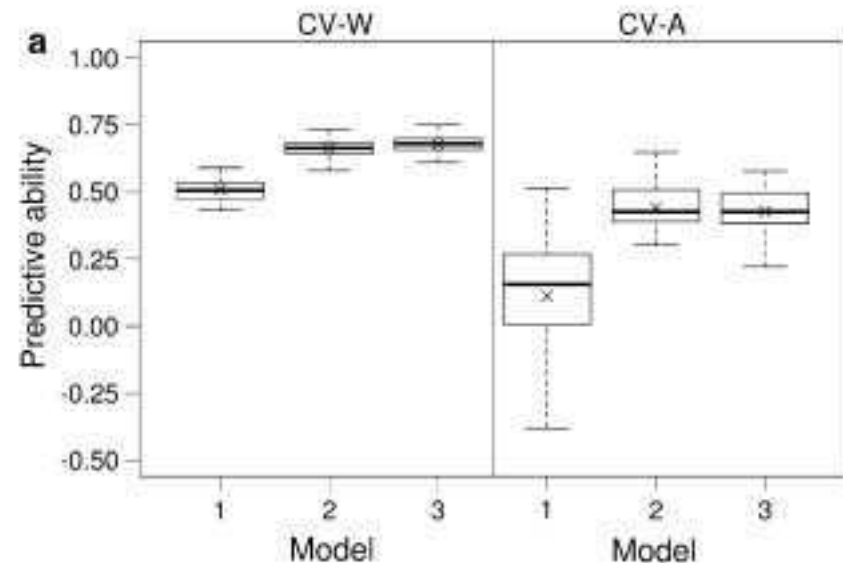
d'après T. Albrecht *et al.*,
TAG online 20/04/2011



- 1 : pedigree seulement
- 2: apparemment 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 $C\pi$, 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

