Utilisation des méthodes de sélection génomique pour l'analyse du génome

Vincent Ducrocq

(présentation « volontaire » suggérée au titre imposé)

Objectif

• Question initiale de Jean-Michel:

« Les méthodes de sélection génomique peuvent-elles être utiles pour connaître le fonctionnement du génome »

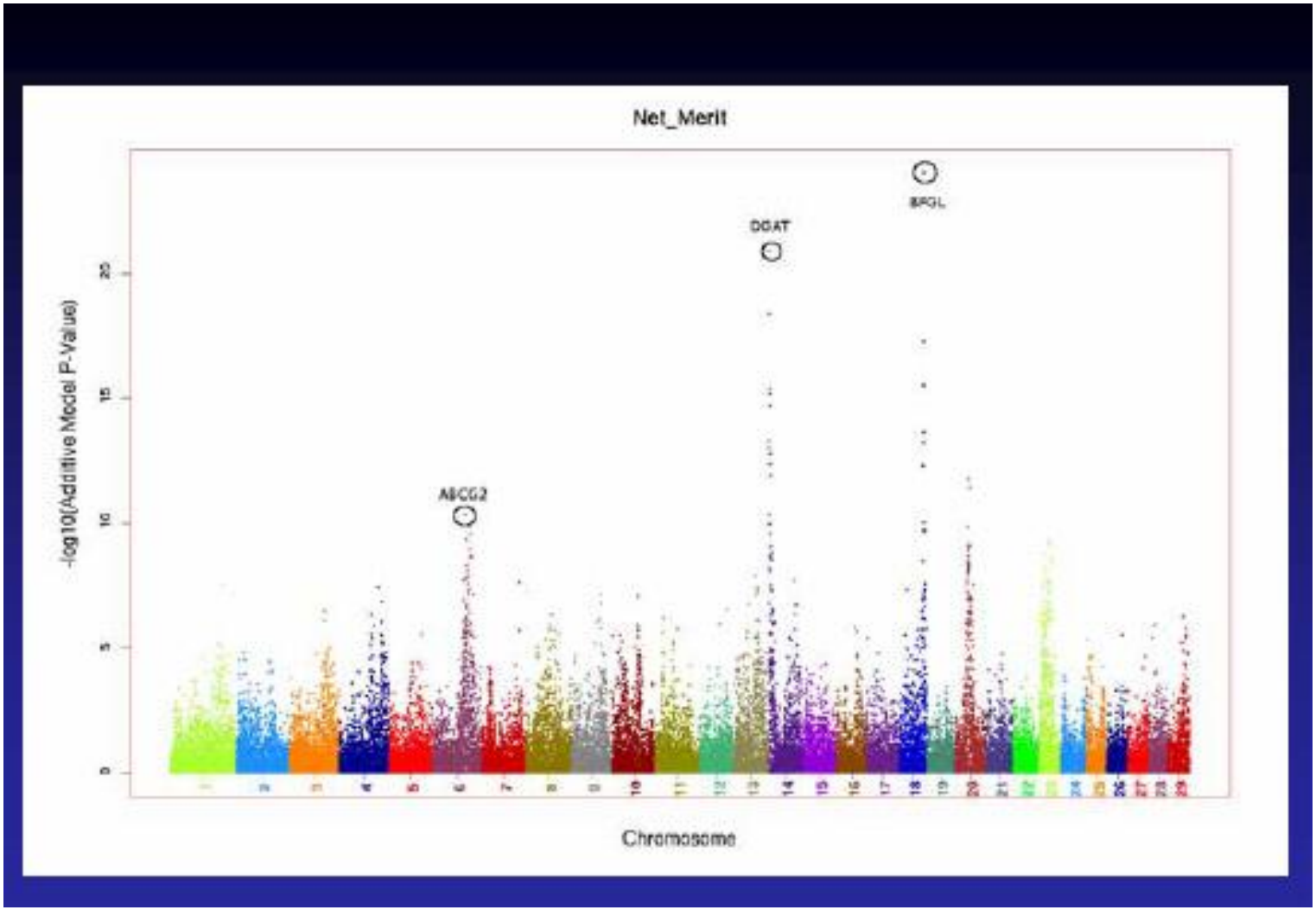
• Message WCGALP:

« l'avantage de la sélection génomique, c'est qu'on n'a pas à connaître le fonctionnement du génome »…

• Sélection génomique: = sélection de candidats dont la valeur génétique est prédite à partir du génotype de milliers de marqueurs répartis sur tout le génome

➔ estimation des effets de marqueurs

Exemple



Cole et al., 2008

Exemple (2)

• A peu près les mêmes marqueurs sur le BTA 18 ont un gros effet sur plusieurs caractères:

- Facilité de naissance et de vêlage
- Largeur du bassin, taille, Profondeur de corps
- Longévité
- « Net merit »

➔ des veaux trop gros conduisent à plus de difficultés de vêlage, ce qui a des conséquences négatives sur la longévité des vaches et l'efficacité économique globale

Les méthodes d'évaluation génomique

- Données et modèles
- BLUP génomique (modèle marqueurs ou animal, en une ou deux étapes)
- La famille Bayes (avec tous les marqueurs ou une partie d'entre eux)
- Les méthodes adaptées au problème p>>n
- (Autres approches)
- L'approche française (SAM + SG)

Données de base de la sélection génomique

• **y** = YD (Yield deviation) = données individuelles corrigées pour tous les effets fixes et aléatoires non génétiques

ou

• **y** = DYD (Daughter yield deviation) = (2 x) moyenne pour chaque taureau des YD de leurs filles corrigées pour le niveau génétique de leur mère
(avec poids associé = EDC (Equivalent Daughter Contribution)

ou

• **y** = index dérégressés (= DYD reconstituées à partir des valeurs génétiques estimées, des EDC et des parentés)

ou

• **y** = index (~« proxy » des DYD, valable que si CD très élevés)

Modèle de base

$$y_j = \mu + \sum_{i=1}^{N} x_i m_i + e_j \qquad \text{avec } x_i = 0, 1 \text{ ou } 2 \qquad \mathbf{y} = \mu\mathbf{1} + (\mathbf{Z})\mathbf{Xm} + \mathbf{e}$$

Si on suppose $m_i \sim N(0, \sigma^2_m)$ ➔ **(G)BLUP**

(Hypothèse : toute la variabilité aux QTL est expliquée par les marqueurs !)

$$\begin{bmatrix} \mathbf{1'1} & \mathbf{1'X} \\ \mathbf{X'1} & \mathbf{X'X} + \dfrac{\sigma^2_e}{\sigma^2_m}\mathbf{I} \end{bmatrix} \begin{bmatrix} \mu \\ \hat{\mathbf{m}} \end{bmatrix} = \begin{bmatrix} \mathbf{1'y} \\ \mathbf{X'y} \end{bmatrix}$$

Variance? $\sigma^2_m = \sigma^2_a \Big/ 2\sum_N p_i(1-p_i)$ (p_i = fréquence allélique)

Nouveaux animaux génotypés (sans performances)

$$GEBV = \sum_{i=1}^{N} x_i \hat{m}_i$$

Modèle équivalent

On prend en compte les fréquences alléliques en recentrant:

$w_i = x_i - 2\,p_i$ et $\mathbf{y} = \mu^{*}\mathbf{1} + \mathbf{Wm} + \mathbf{e}$

On écrit $\mathbf{a} = \mathbf{Wm}$ et $\mathbf{G} = \dfrac{\mathbf{WW}'}{2\sum_{N} p_i(1-p_i)}$

On a alors $\mathrm{Var}(\mathbf{a}) = \mathbf{G}\sigma^2_a$

$$\mathbf{y} = \mu^{*}\mathbf{1} + \mathbf{Z}(\mathbf{Wm}) + \mathbf{e} = \mu^{*}\mathbf{1} + \mathbf{Za} + \mathbf{e}$$

$$\begin{bmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{1} & \mathbf{Z}'\mathbf{Z} + \dfrac{\sigma^2_e}{\sigma^2_a}\mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \mu^{*} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}$$

Différences entre les deux modèles

- Modèle « animal » se rapproche du BLUP classique
  en particulier pour l'interprétation de **G** (parenté « observée »)
- Si les marqueurs n'expliquent pas tous ➔ rajout d'un terme polygénique: $y_j = \mu + u_j + \sum_{i=1}^{N} x_i m_i + e_j$
- On obtient « l'index génomique » d'un animal directement
  - ➔ on n'a plus les effets individuels des marqueurs
  - ➔ totalement inutile pour l'analyse du génome…
- Intéressant si nombre de marqueurs > nombre d'animaux génotypés, beaucoup moins sinon, car $\mathbf{G}^{-1}$ est dense !
- Volonté de combiner les résultats des évaluations classiques génomiques ➔ cuisine « à la Van Raden » pas très satisfaisante

➔Modèle en une étape

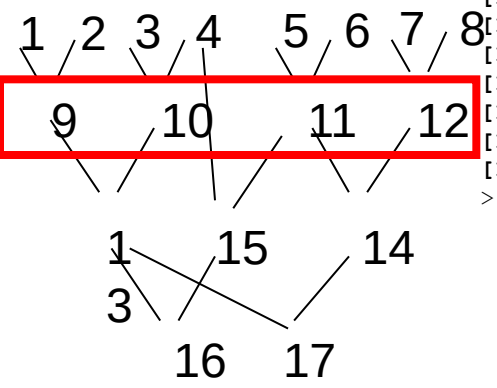
- Andrès Legarra, Misztal, Aguilar, Christensen, …
- Pour « diffuser » l'information génomique aux apparentés non génotypés:

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} \\ \mathbf{H}_{21} & \mathbf{H}_{22} \end{bmatrix} = \begin{bmatrix} \overbrace{\mathbf{A}_{11} + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\left(\mathbf{G} - \mathbf{A}_{22}\right)\mathbf{A}_{22}^{-1}\mathbf{A}_{21}}^{\textit{non génotypés}} & \underbrace{\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}}_{\textit{génotypés}} \\ \mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G} \end{bmatrix}$$

- Incroyable mais vrai: $\mathbf{H^{-1}}$ a une forme assez simple:

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}$$

Exemple



A

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]	[,13]	[,14]	[,15]	[,16]	[,17]
[1,]	1.00									0.50				0.25			0.12	0.12
[2,]		1.00								0.50				0.25			0.12	0.12
[3,]			1.00								0.50			0.25			0.12	0.12
[4,]				1.00							0.50			0.25		0.50	0.38	0.12
[5,]					1.00							0.50			0.25	0.25	0.12	0.12
[6,]						1.00						0.50			0.25	0.25	0.12	0.12
[7,]							1.00						0.50		0.25			0.12
[8,]								1.00					0.50		0.25			0.12
[9,]	0.50	0.50								1.00				0.50			0.25	0.25
[10,]			0.50	0.50							1.00			0.50		0.25	0.38	0.25
[11,]					0.50	0.50						1.00			0.50	0.50	0.25	0.25
[12,]							0.50	0.50					1.00		0.50			0.25
[13,]	0.25	0.25	0.25	0.25					0.50	0.50				1.00		0.12	0.56	0.50
[14,]					0.25	0.25	0.25	0.25			0.50	0.50			1.00	0.25	0.12	0.50
[15,]				0.50	0.25	0.25				0.25	0.50			0.12	0.25	1.00	0.56	0.19
[16,]	0.12	0.12	0.12	0.38	0.12	0.12			0.25	0.38	0.25			0.56	0.12	0.56	1.06	0.34
[17,]	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.25	0.25	0.25	0.25	0.50	0.50	0.19	0.34	1.00	

>

H

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]	[,13]	[,14]	[,15]	[,16]	[,17]
[1,]	1.00		0.17	0.17	0.17	0.17	0.17	0.17	0.50	0.35	0.35	0.35	0.42	0.35	0.26	0.34	0.39
[2,]		1.00	0.17	0.17	0.17	0.17	0.17	0.17	0.50	0.35	0.35	0.35	0.42	0.35	0.26	0.34	0.39
[3,]	0.17	0.17	1.00		0.17	0.17	0.17	0.17	0.35	0.50	0.35	0.35	0.42	0.35	0.17	0.30	0.39
[4,]	0.17	0.17		1.00	0.17	0.17	0.17	0.17	0.35	0.50	0.35	0.35	0.42	0.35	0.68	0.55	0.39
[5,]	0.17	0.17	0.17	0.17	1.00		0.17	0.17	0.35	0.35	0.50	0.35	0.35	0.42	0.34	0.34	0.39
[6,]	0.17	0.17	0.17	0.17		1.00	0.17	0.17	0.35	0.35	0.50	0.35	0.35	0.42	0.34	0.34	0.39
[7,]	0.17	0.17	0.17	0.17	0.17	0.17	1.00		0.35	0.35	0.35	0.50	0.35	0.42	0.26	0.31	0.39
[8,]	0.17	0.17	0.17	0.17	0.17	0.17		1.00	0.35	0.35	0.35	0.50	0.35	0.42	0.26	0.31	0.39
[9,]	0.50	0.50	0.35	0.35	0.35	0.35	0.35	0.35	1.00	0.70	0.70	0.70	0.85	0.70	0.52	0.69	0.77
[10,]	0.35	0.35	0.50	0.50	0.35	0.35	0.35	0.35	0.70	1.00	0.70	0.70	0.85	0.70	0.60	0.72	0.77
[11,]	0.35	0.35	0.35	0.35	0.50	0.50	0.35	0.35	0.70	0.70	1.00	0.70	0.70	0.85	0.68	0.69	0.77
[12,]	0.35	0.35	0.35	0.35	0.35	0.35	0.50	0.50	0.70	0.70	0.70	1.00	0.70	0.85	0.52	0.61	0.77
[13,]	0.42	0.42	0.42	0.42	0.35	0.35	0.35	0.35	0.85	0.85	0.70	0.70	1.35	0.70	0.56	0.96	1.02
[14,]	0.35	0.35	0.35	0.35	0.42	0.42	0.42	0.42	0.70	0.70	0.85	0.85	0.70	1.35	0.60	0.65	1.02
[15,]	0.26	0.26	0.17	0.68	0.34	0.34	0.26	0.26	0.52	0.60	0.68	0.52	0.56	0.60	1.18	0.87	0.58
[16,]	0.34	0.34	0.30	0.55	0.34	0.34	0.31	0.31	0.69	0.72	0.69	0.61	0.96	0.65	0.87	1.41	0.80
[17,]	0.39	0.39	0.39	0.39	0.39	0.39	0.39	0.39	0.77	0.77	0.77	0.77	1.02	1.02	0.58	0.80	1.52

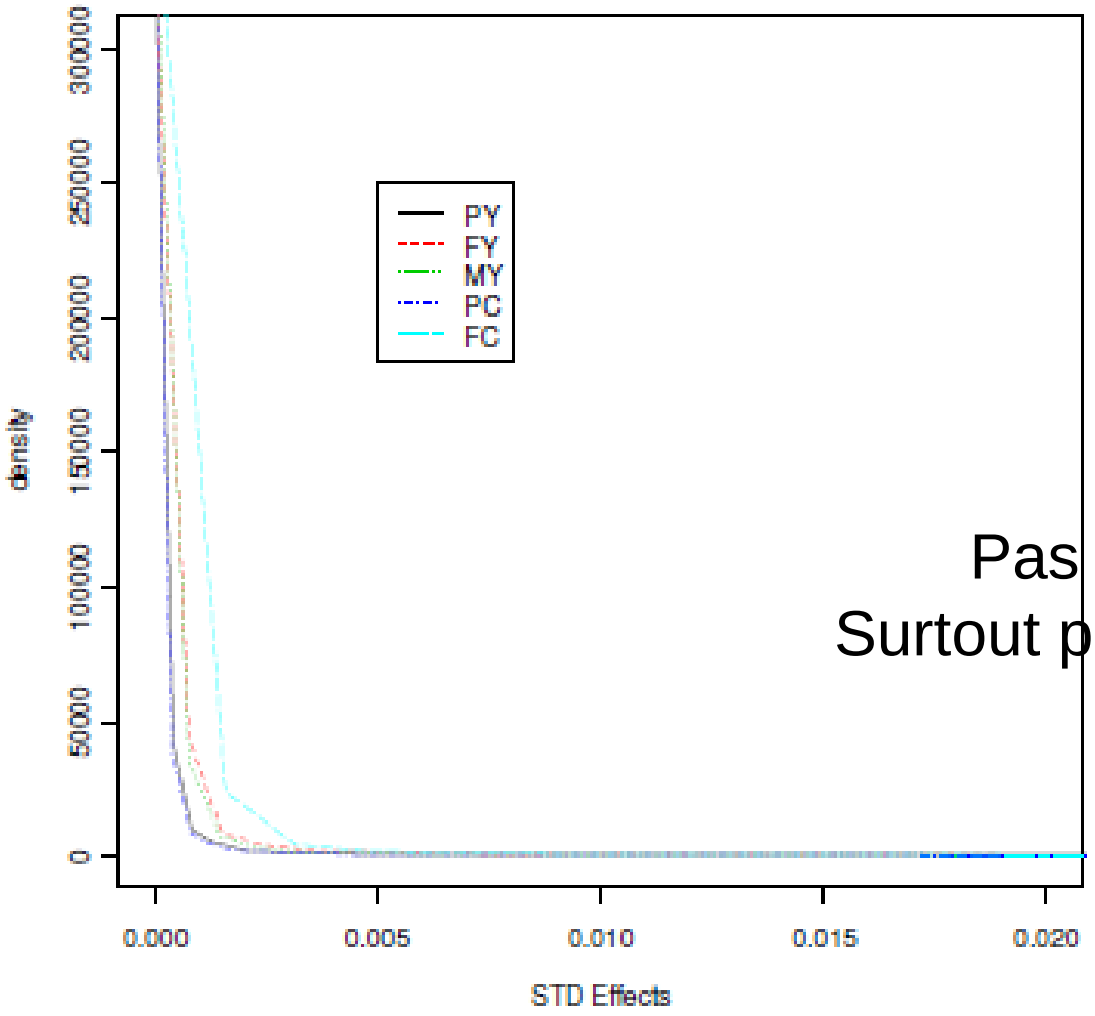
Modèle en une étape

$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{a} + \mathbf{e}$ ➔ **(G)BLUP** sur toutes les données

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \alpha\mathbf{H}^{-1} \end{bmatrix} \begin{bmatrix} \mu^* \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}$$

- Système énorme (et assez dense) mais avec des astuces, cela peut être appliqué à des fichiers de grande taille
- Trois gros inconvénients (à mon sens)
  - ne marchera plus quand le nombre d'animaux génotypés dépassera quelques dizaines de milliers
  - ne convient pas aux caractères analysés avec des modèles compliqués (multicaractères, régression aléatoire, analyse de survie)
  - conceptuellement, une puce plus dense n'apportera rien !

Distribution des effets des marqueurs



Pas franchement normal
Surtout pour le taux butyreux (FC)

Hayes et al, 2008

La famille Bayes

➢ Tous les marqueurs « contribuent »

- On suppose que m_i ~distribution t (« fat tails ») pour prendre en compte les marqueurs « à gros effet »

  De manière équivalente:

  $$m_i \sim N(0, \sigma^2_{m_i}) \text{ avec } \sigma^2_{m_i} \sim \text{chi} - \text{deux inverse}$$

  Machinerie MCMC ➔ **Bayes A**

- Problème : a posteriori de chaque variances a une distribution identique à l'a priori avec 1 degré de liberté de plus, quel que soit le nombre de performances analysés

  $$m_i \sim N(0, \sigma^2_m) \text{ avec } \sigma^2_m = \text{cste}$$ ➔ **Bayes C ou Blup Bayesien**

La famille Bayes

- (Habier)

  $m_i \sim N(0, \sigma^2_{m_i})$ avec $\sigma^2_{m_i} \sim$ chi – deux inverse

  avec paramètre du chi – deux inverse ~ Gamma (1,1) pour tous

  ➔ **Bayes D**

- (Casella et Park, Legarra…)

  $m_i \sim$ double exponentielle de paramètre τ_i ➔ **Lasso Bayésien**

- (Une alternative de Andrès Legarra)

  Estimation de τ_i par Lasso Bayésien puis BLUP Bayésien sur marqueurs avec variances hétérogènes avec $\sigma^2_{m_i} = \tau_i \, \sigma^2_m$

La famille Bayes

➢ **Une proportion π des marqueurs ne contribuent pas**

- $(1-\pi) : m_i \sim N(0, \sigma^2_{m_i})$ avec $\sigma^2_{m_i} \sim$ chi – deux inverse

  $\pi : m_i = 0$ π supposée connue

  Machinerie MCMC ➔ **Bayes B**

  Très laborieux et long…

  ➔ une approximation et un EM ➔ **Fast Bayes B**

- Une approximation plus simple

  $(1-\pi) : m_i \sim N(0, \sigma^2_m)$ ➔ Un autre **Bayes C (?)**

  $\pi : m_i \sim N(0, 0.01\sigma^2_m)$

La famille Bayes

- Comme Bayes C, mais avec $\pi : m_i = 0$

  et $\pi \sim$ uniforme(0 ,1) ➔ **Bayes Cπ**

- Comme Bayes D, mais avec $\pi : m_i = 0$

  et $\pi \sim$ uniforme(0 ,1) ➔ **Bayes Dπ**

- (Verbyla et al, 2009) Plus facile à calculer

  $m_i \sim (1-\gamma_i)\, N(0,\ 0.01\sigma^2_{m_i}) + \gamma_i N(0,\ \sigma^2_{mi})$

  avec $\sigma^2_{m_i} \sim$ chi – deux inverse

  et $\gamma_i \sim$ bernoulli ➔ **Stochastic Search variable Selection**

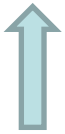
- C'est sûrement pas fini…

Résultats: corrélation avec DYD

Table 1: GEBV accuracy[a] for 115 bulls born between 1953 and 1975 depending on the Bayesian method used to estimate SNP effects, the quantitative trait and the number of training bulls born between 1995 and 2004

Trait	Training data size	BayesA	BayesB[b]	BayesCπ	BayesDπ
Milk yield	1,000	0.42	0.29	0.37	0.43
	4,000	0.44	0.42	0.41	0.45
Fat yield	1,000	0.48	0.51	0.49	0.47
	4,000	0.52	0.49	0.55	0.53
Protein yield	1,000	0.15	0.10	0.15	0.14
	4,000	0.18	0.15	0.17	0.17
Somatic cell score	1,000	0.14	0.12	0.14	0.12
	4,000	0.28	0.18	0.24	0.23

[a] standard errors: 0.08-0.09
[b] $\pi = 0.99$



Bayes A n'est pas si mal !

Résultats: nombre de SNP retenus

Table 2: Posterior mean of (1-π) multiplied by the number of SNPs used in the analyses depending on the Bayesian method used to estimate SNP effects, the quantitative trait and the number of training bulls

Trait	Training data size	BayesB[a]	BayesCπ	BayesDπ
Milk yield	1,000	404	1,180	13,982
	4,000	436	2,162	13,329
Fat yield	1,000	402	487	13,533
	4,000	441	1468	13,513
Protein yield	1,000	403	13,942	14,430
	4,000	438	6,723	13,512
Somatic cell score	1,000	398	5,057	12,962
	4,000	428	3,261	13,941

[a] $\pi = 0.99$

tout ça pour ça ?

Méthodes spécifiques au problème p>>n

- **PLS** : Partial Least Squares: régression sur les combinaisons linéaires orthogonales des SNP les mieux corrélées au phénotype à prédire
    - De nombreuses variantes existent, dont certaines intéressantes… en particulier des approches avec **sélection des SNP** en même temps (Sparse PLS)

- **LASSO**: λ = pénalisation de la norme 1 (sélection de variable)

$$\hat{m} = \arg\min \left\{ \sum_j \left(y_j - x\mathbf{m}_j \right)^2 \right\} + \lambda \sum_i |m_i|$$

= revient à mettre à 0 les effets les plus petits

- **Ridge Regression** $\hat{m} = \arg\min \left\{ \sum_j \left(y_j - \mathbf{x}\mathbf{m}_j \right)^2 \right\} + \lambda \sum_i m_i^2$

Méthodes spécifiques au problème p>>>n

- **Elastic Net** : Combine LASSO et RR

  à travers un paramètre α compris entre 0 et 1

$$\hat{m} = \arg\ \min \left\{ \sum_j \left(y_j - X\mathbf{m}_j \right)^2 \right\} + \lambda \left(\alpha \sum_i m_i^2 + (1-\alpha) \sum_i |m_i| \right) \sum_i |m_i|$$

  λ et α sont choisis de façon à donner les meilleurs résultats

  dans une population de validation

- Pour mémoire (je n'y connais rien):

  Méthodes Bayésiennes non paramétriques

  Réseaux neuronaux, …

  ➔ Boîtes noires !

Exemple réel (données Holstein France)

		Milk	Protein	Fat	Protein %	Fat %	Conception rate
	pedigree-based BLUP	0.38	0.44	0.40	0.47	0.44	0.29
54K	GBLUP	0.56	0.55	0.59	0.73	0.72	**0.35**
	PLS	0.53	0.55	0.58	0.71	0.70	0.33
	Elastic-Net	**0.57**	**0.57**	**0.63**	**0.75**	**0.80**	0.34
Pre selection	GBLUP	0.56	0.54	0.59	**0.73**	0.74	**0.33**
	PLS	0.53	0.55	0.58	0.71	0.70	**0.33**
	Elastic-Net	**0.57**	**0.57**	**0.63**	**0.73**	**0.79**	**0.33**

La SAMG Française (SAM +SG)

- **Mise en place officielle: Juin 2010**
- **Etape 1** : Cartographie LDLA. Choix des principaux QTL (~70).

  Definition des haplotypes (~5)

  Pour ces QTL, pourcentage de variance génétique retenue

  proportionel aux variances aux QTL de l'analyse uni-QTL
- **Etape 2** : Choix des SNP par Elastic Net (EN) en forçant le

  nombre des retenu à <1200 SNP

  Les SNP retenus sont regroupés en QTL = haplotype si ils sont dans le même mégabase

  part de variance identique pour tous

La SAMG Française (SAM +SG)

- **Etape 3** : combinaison des 2 étapes précédentes
  - gros QTL 15-20% + EN 35-40%
- **Etape 4** : amélioration des regroupements des SNP obtenus par l'EN, en prenant 1-2 SNP voisins (➔3 à 4 SNP)

  variance globale entre 45 et 50% Mo et No ; entre 50 et 65% en Ho)
- **Etape 5** : Evaluation Assistée par Marqueurs

  Fernando et Grossman, 1989 $a_i = u_i + \sum_{n\_QTL} (h_{i1} + h_{i2})$

  200 à 700 QTL par caractère, avec 20 à 30 haplotypes chacun. Inclusion d'un effet polygénique qui représente 35 à 50% de la variabilité génétique additive, Calculs simples, avec toutes les données

Exemple réel (données Holstein France)

		Milk	Protein	Fat	Protein %	Fat %	Conception rate
	pedigree-based BLUP	0.38	0.44	0.40	0.47	0.44	0.29
54K	GBLUP	0.56	0.55	0.59	0.73	0.72	**0.35**
	PLS	0.53	0.55	0.58	0.71	0.70	0.33
	Elastic-Net	**0.57**	**0.57**	**0.63**	**0.75**	**0.80**	0.34
Pre selection	GBLUP	0.56	0.54	0.59	**0.73**	0.74	**0.33**
	PLS	0.53	0.55	0.58	0.71	0.70	**0.33**
	Elastic-Net	**0.57**	**0.57**	**0.63**	**0.73**	**0.79**	**0.33**
	MAS+EN*	**0.60**	**0.57**	**0.66**	**0.73**	**0.81**	**0.39**
	mean gain of MAS-EN	0.03	0.00	0.03	0.00	0.02	0.06

Conclusion

• De nombreuses méthodes d'évaluation génomique

• Des résultats différents quand les populations génotypées sont petites, beaucoup moins pour les grandes…

• Pour l'analyse du génome: pas très discriminant, surtout quand on se rapproche d'un modèle infinitésimal…

• Même pour les méthodes de sélection de SNP

• Les puces HD ne devraient pas trop changer la donne …

• Intérêt toujours d'actualité pour la détection de QTL…

• Une très grande masse de génotypes disponible !

Utilisation des méthodes de sélection génomique pour l'analyse du génome

Vincent Ducrocq

(présentation volontaire suggérée au titre imposé)