Use of the Elastic-Net algorithm for genomic selection in dairy cattle

P. Croiseau, F. Guillaume, S. Fritz and V. Ducrocq

ALIMENTATION
AGRICULTURE
ENVIRONNEMENT

EAAP : 24-27 August 2009

Genetic context :

- 54K SNP array available in dairy cattle
  - ➢ makes possible to envision the use of genomic prediction instead of the classical genetic evaluations in selection programmes (polygenic model)
- However, this genomic prediction requires :
  - ➢ to achieve a regression using all SNP together
  - ➢ to deal with a p >> n problem
  - ➢ to perform the analysis in a limited time (1 or 2 days /trait)

Variable Reduction methods :

- It is known that not all the SNP available are involved in the trait of interest
    - ➢ The idea is to select only the SNP which are involved rather than to estimate the effects of the complete set of SNP
- Here, we focus on penalized regression approaches:
    - ➢ Ridge Regression
    - ➢ Lasso
    - ➢ Elastic-Net (EN)

Penalized regression approaches :

- Ridge Regression :

$$\hat{\beta} = \arg\min\left\{\sum_{i=1}^{n}\left(Y_i - X_i\beta\right)^2 + \lambda\sum_{j}\beta_j^2\right\}$$

- Lasso :

penalty

$$\hat{\beta} = \arg\min\left\{\sum_{i=1}^{n}\left(Y_i - X_i\beta\right)^2 + \mu\sum_{j}\left|\beta_j\right|\right\}$$

- Zou & Hastie show that :
  - ➢ Ridge Regression retains all the predictors
  - ➢ Lasso retains the most significant predictors and removes the others

Elastic-Net : a combination of RR and Lasso

$$\hat{\beta} = \arg\min\left\{\sum_{i=1}^{n}\left(Y_i - X_i\beta\right)^2 + \lambda\left(\underbrace{\alpha\sum_j \beta_j^2}_{\boxed{\text{RR}}} + \underbrace{\left(1-\alpha\right)\sum_j\left|\beta_j\right|}_{\boxed{\text{LASSO}}}\right)\right\}$$

- λ : penalty intensity
- α = 0 → LASSO
- α = 1 → RR

- The method depends on 2 parameters :
  - α → [0 ; 0.1 ; 0.2 ; … ; 1]
  - λ → [0 ; 5 ; 10 ; … ; 100]

DATA

- 2 breeds studied : Montbéliarde and Holstein
  - ➢ Use of DYD (Daughter Yield Deviation) :

    Average of the daughters' performance adjusted for fixed and non genetic random effects of the daughters and for the genetic effects of the bull's mates

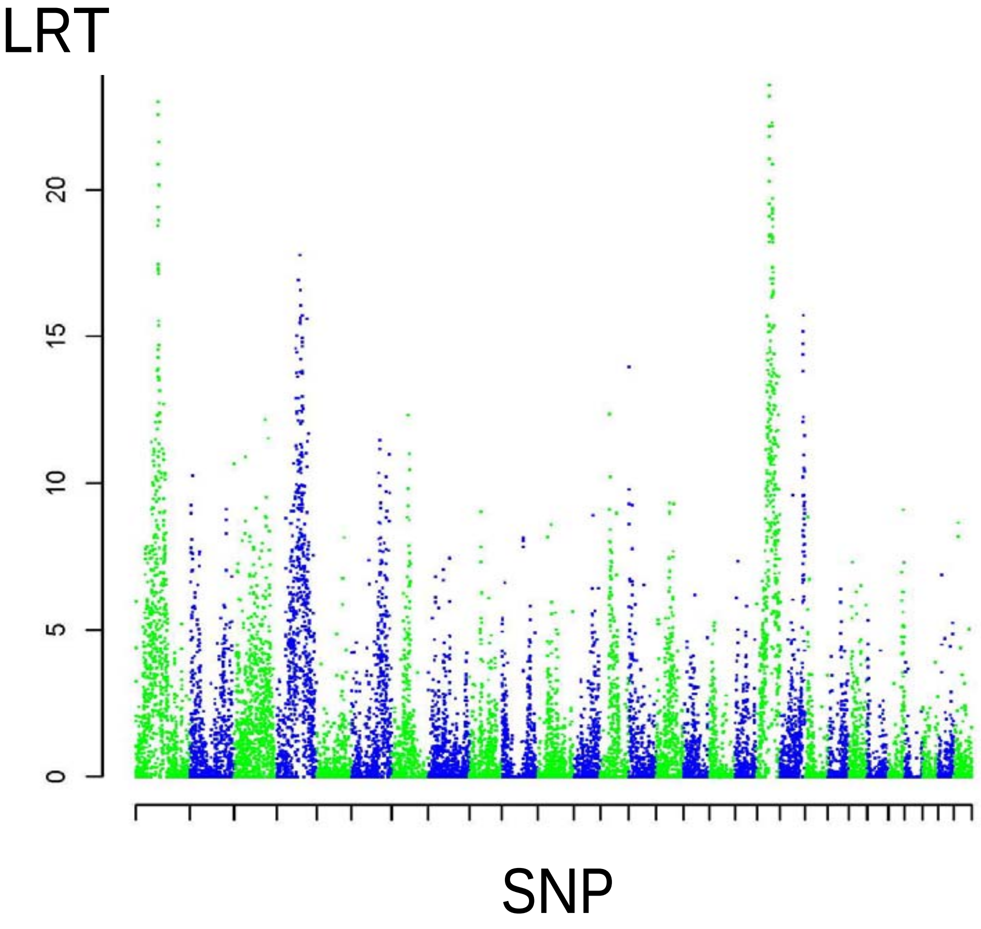
	Montbéliarde	*Holstein*
Training step	694	1827
Validation step	227	540

DATA

- Complete data required
  - Missing data are imputed using DualPhase (Druet et al.)

	number of SNP	
MAF	**Montbéliarde**	**Holstein**
5%	38959	41101

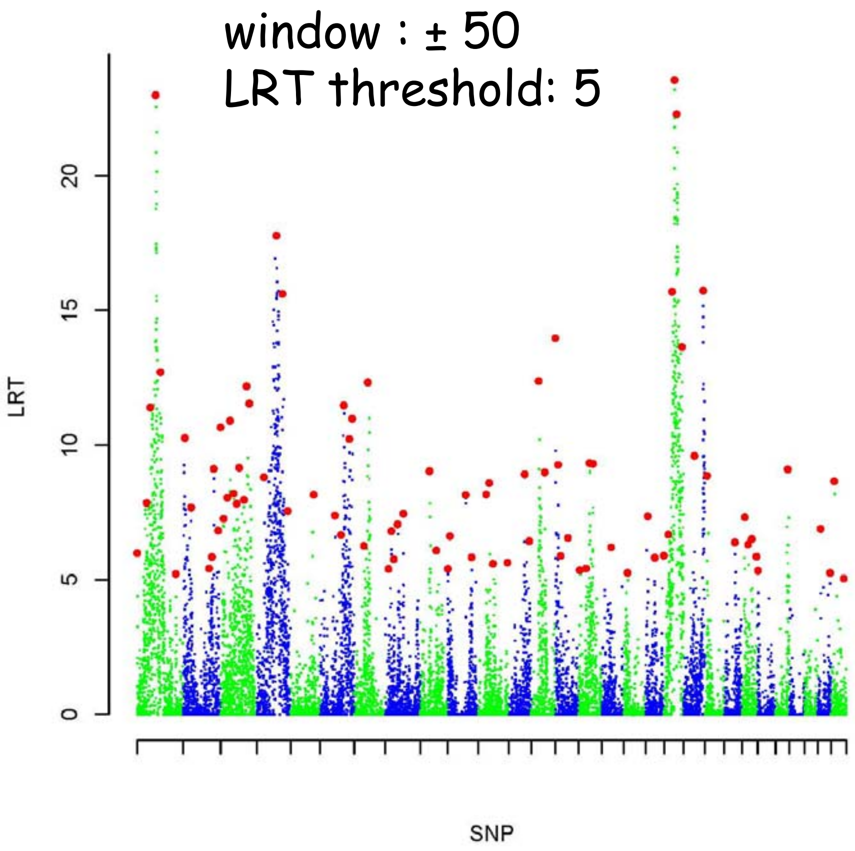
Preselection of the SNP using a LDLA approach



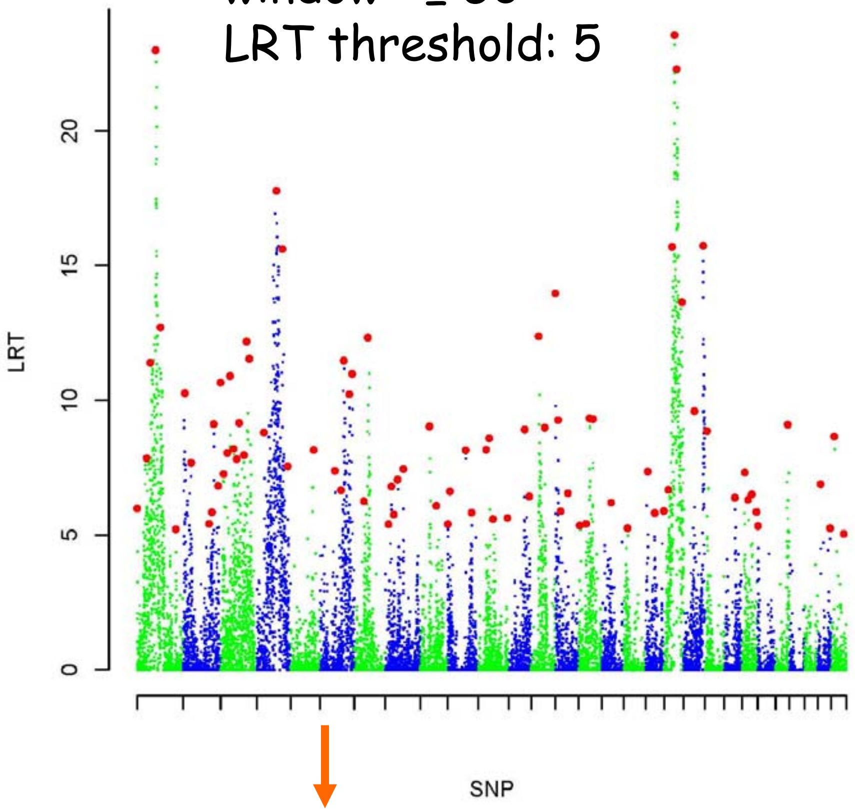
Different definitions of a LRT peak:

- Size of the window:
  - ✓ ±25, ±50, ±100 and ±200
- LRT threshold:
  - ✓ 0, 1, 3, 5 and 9

Preselection of the SNP using a LDLA approach



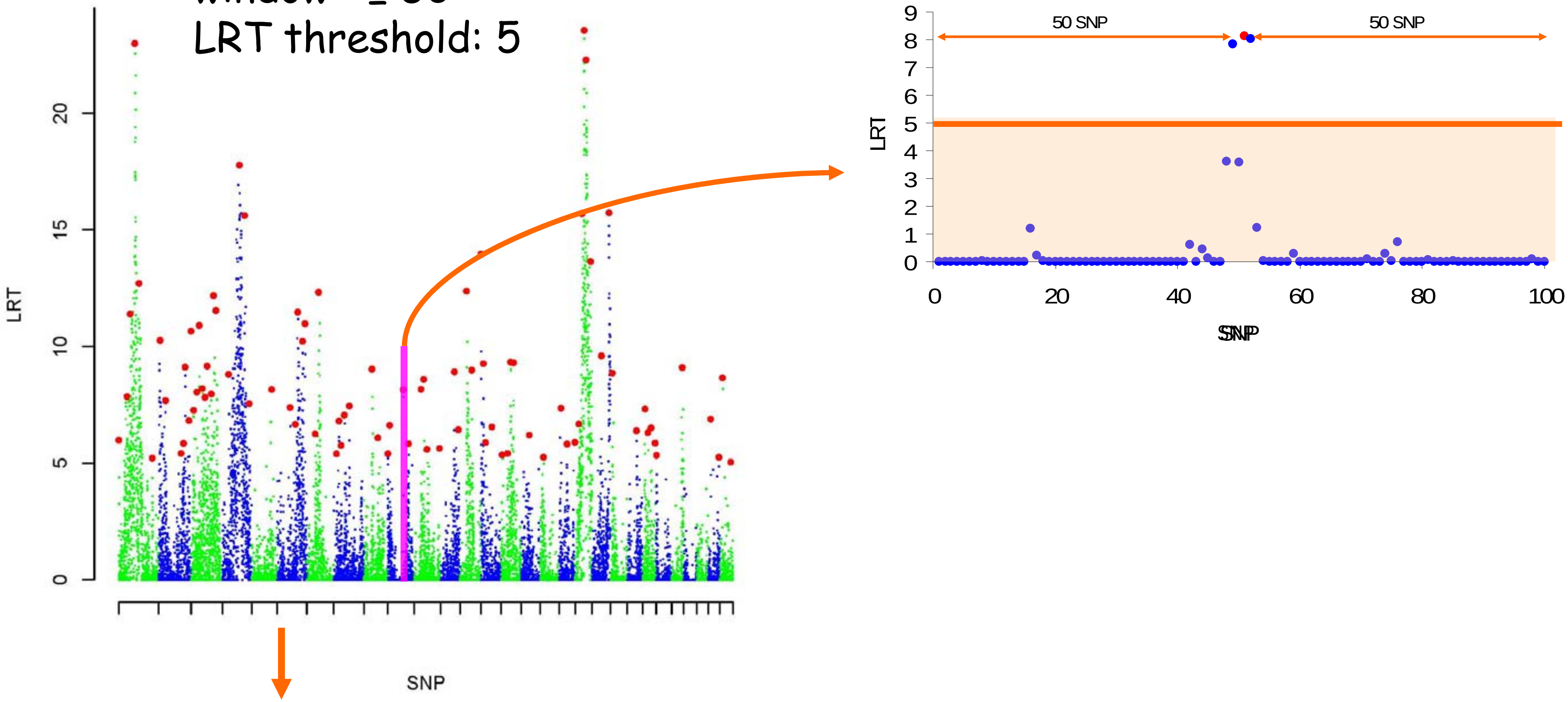
Preselection of the SNP using a LDLA approach



Between 20 and 500 LRT peaks identified

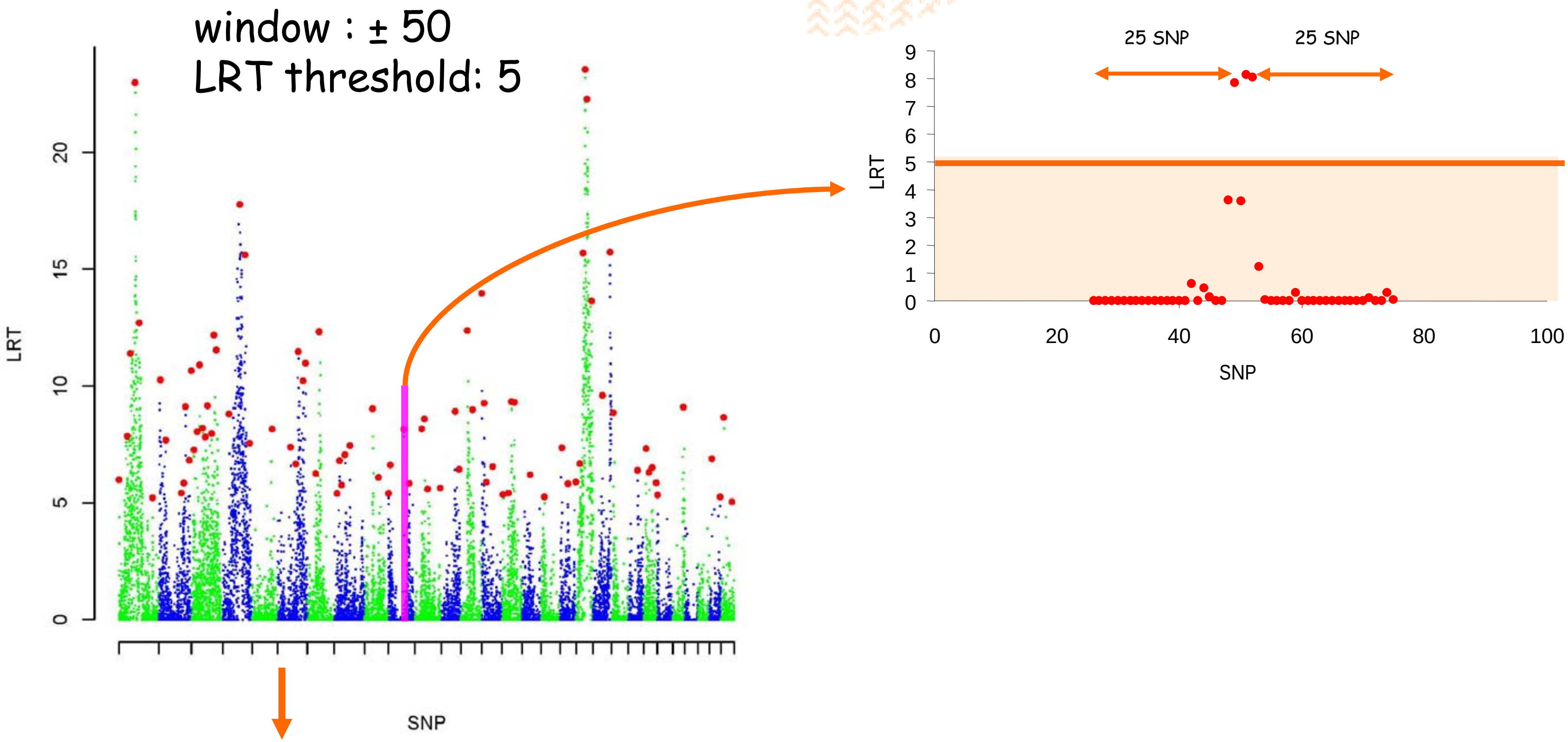
Preselection of the SNP using a LDLA approach

window : ± 50
LRT threshold: 5



Between 20 and 500 LRT peaks identified

Preselection of the SNP using a LDLA approach

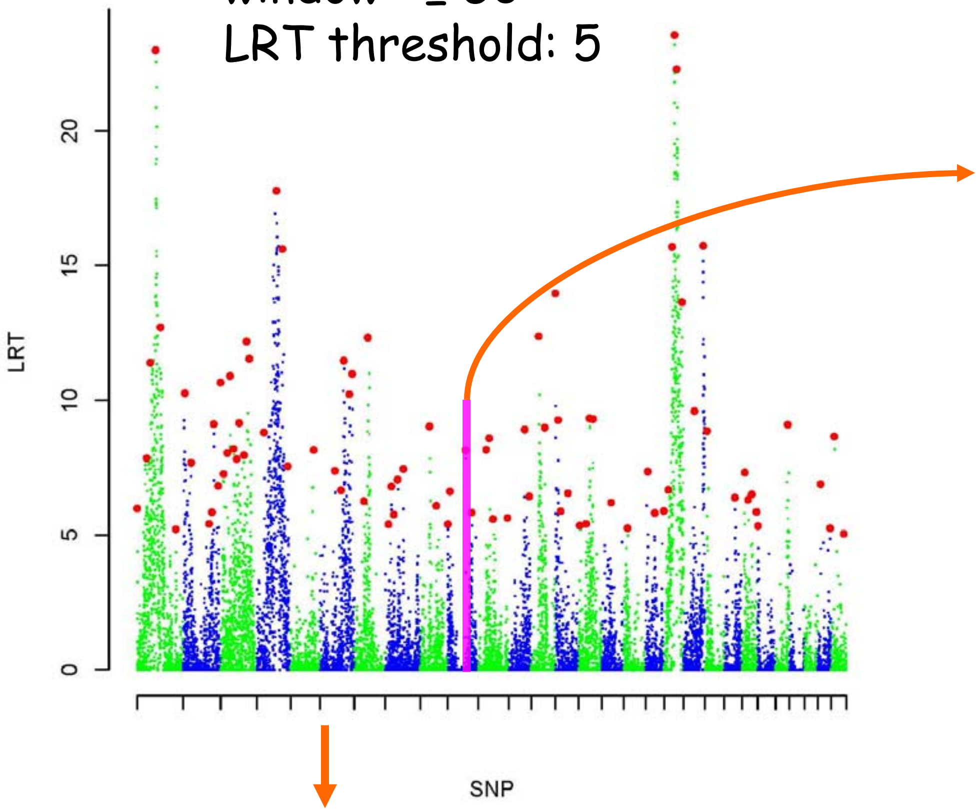


window : ± 50
LRT threshold: 5

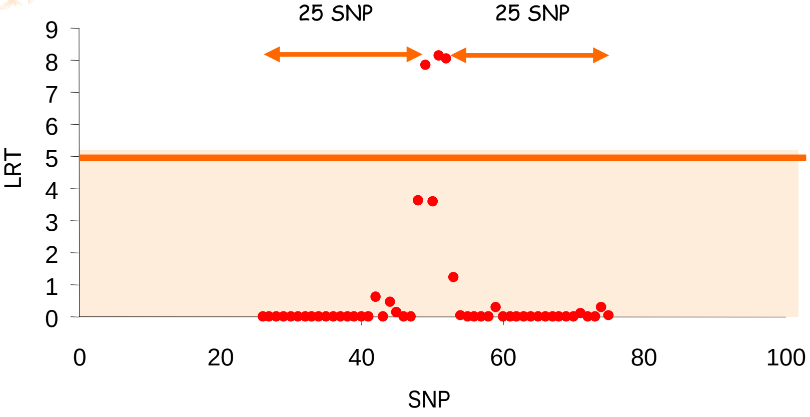
Between 20 and 500 LRT peaks identified

Preselection of the SNP using a LDLA approach

window : ± 50
LRT threshold: 5



Between 20 and 500 LRT peaks identified



Each LRT peak included in the EN model with a window of ±25 SNP

Analysis

- Measure of the quality of the predicted DYD :
  - Correlation between :

    DYD observed in 2008 on the animals of the validation set and DYD estimated on these animals using the coefficients calculated using only the animals from the training set
- We test all combinations of :
  - Preselection criteria of SNP (window size and LRT threshold)
  - Elastic-Net parameters (**α, λ**)

Results (of best combination for each trait)

Montbéliarde breed :

38959 SNP →

	Milk		Fat kg		Protein kg		Fat %		Protein %	
Elastic-Net without preselection	0.362	0.13	0.374	0.09	0.404	0.14	0.508	0.11	0.364	0.03
Elastic-Net with preselection	**0.493**		**0.469**		**0.549**		**0.614**		**0.392**	
BLUP polygenic										
french MAS approach										

Holstein breed :

41101 SNP →

	Milk		Fat kg		Protein kg		Fat %		Protein %	
Elastic-Net without preselection	0.400	0.04	0.436	0.07	0.280	0.09	**0.741**	0.01	0.591	0.04
Elastic-Net with preselection	**0.443**		**0.503**		**0.373**		0.734		**0.635**	
BLUP polygenic										
french MAS approach										

Results

Montbéliarde breed :

	Milk	*Fat kg*	*Protein kg*	*Fat %*	*Protein %*
Elastic-Net without preselection	0.362	0.374	0.404	0.508	0.364
Elastic-Net with preselection	**0.493**	**0.469**	**0.549**	**0.614**	0.392
BLUP polygenic *	0.273	0.355	0.276	0.372	0.214
french MAS approach *	0.420	0.438	0.383	0.579	**0.543**

Holstein breed :

	Milk	*Fat kg*	*Protein kg*	*Fat %*	*Protein %*
Elastic-Net without preselection	0.400	0.436	0.280	0.741	0.591
Elastic-Net with preselection	0.443	0.503	0.373	0.734	0.635
BLUP polygenic *	0.423	0.317	0.330	0.449	0.390
french MAS approach *	**0.520**	**0.532**	**0.459**	**0.755**	**0.673**

* F. Guillaume et al.

Elastic-Net parameters

- Best results were obtained using :
  - a window size of ±50 SNP and a LRT threshold of 5 for the preselection
- Most of the time, the best set of parameters for the Elastic Net is with :
  - λ = 4 - 15 (relatively strong intensity of penalization)
  - α = 0.1 - 0.3 , close to a full Lasso approach
- For both breeds, around 500 SNP with non-null effect
- Worst case situation (with 500 LRT peaks, i.e. with 25000 SNP) tested in less than one day

Conclusion

- The Elastic-Net approach shows:
  - ➢ Interesting results in Montbéliarde breed
  - ➢ Need further investigation for the Holstein breed
- Other improvements need to be investigated
  - ➢ Haplotype vs Allelic coding
  - ➢ Addition of familial information (as in the French MAS approach)
  - ➢ Comparison to other approaches such as Bayes A, B, …

  is under way…