Experimental assessment of accuracy of genomic selection in sugarcane

M. Gouy[1,2,3], Y. Rousselle[2], D. Bastianelli[8], P. Lecomte[8], L. Bonnal[8], D. Roques[4], J-C. Efile[4], S. Rocher[4,7], J. Daugrois[6], L. Toubi[4], S. Nabeneza[9], C. Hervouet[5], H. Telismart[2], M. Denis[5], A. Thong Chane[1], J.C. Glaszmann[5], J.-Y. Hoarau[1,4]*, S. Nibouche[2] and L. Costet[2]

[1]eRcane, F-97494 Sainte-Clotilde, La Réunion, France
[2]Cirad, UMR PVBMT, F-97410 Saint-Pierre, La Réunion, France
[3]Université de la Réunion, UMR PVBMT, F-97410 Saint-Pierre, La Réunion, France
[4]Cirad, UMR AGAP, F-97170 Petit Bourg, Guadeloupe, France
[5]Cirad, UMR AGAP, F-34398 Montpellier, France
[6]Cirad, UMR BGPI, F-97170 Petit Bourg, Guadeloupe, France
[7]Université des Antilles et de la Guyane, F-97157 Pointe-à-Pitre, Guadeloupe, France
[8]Cirad, UMR SELMET, F-34398 Montpellier, France
[9]Cirad, UMR SELMET, F-97410 Saint-Pierre, La Réunion, France

eRcane — Valoriser la ressource canne
Département Hybridation-Sélection

Pvbmt cirad
Peuplement Végétaux et Bio-agresseurs en Milieu Tropical

agap cirad
Amélioration Génétique et Adaptation des Plantes

SELMET cirad
Système d'Elevage Méditerranéen et Tropicaux

UMR BGPI cirad
Biologie et Génétique des Interactions Plantes-Parasites

ANR — GENOPLANTE — Qualitropic — Département de la Réunion — Region Reunion Conseil Régional — l'Europe s'engage en Guadeloupe avec le FEDER

Genomic Selection (GS)

- Novel approach for selecting individuals in breeding programs
- Improvement of complex traits requiring long field experiments
- Predict performance of individuals (breeding or clonal value) on the basis of genome-wide fingerprinting, calibration with a 'training population'
- Exploit the whole marker information by simultaneously estimating the effect of each marker across the entire genome to predict genetic value
- GS does not rely on subset of significant markers. Unlike conventional MAS, has the ability to capture more of the genetic variation (grasp QTLs with small-effects and part of epistasis)

Objectives of our work

Explore potential of GS in the context of sugarcane

- 2 independent panels
- 4 models of prediction
- 10 quantitative traits of interest

Evaluate accuracy of GS by cross validations within or between panels

Panels and traits

- 2 independent populations

GUA

167 varieties

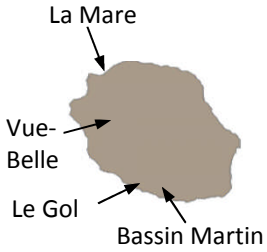
REU

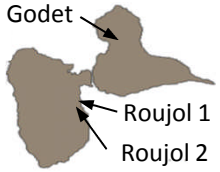
167 varieties

- 10 traits :
  - Morphological traits: stalk number (SN), stalk diameter (SD)
  - Technological trait: brix (BR), Fiber Content (FB)
  - Ligno-cellulosic traits : Acid Detergent Lignin (ADL), Acid Detergent Fiber (ADF), In Vitro Detergent Fiber Digestibility (IVDFD)
  - Disease resistance : Rust (RST), Smut (SM), feuille jaune (SCYLV)

Experimental data

Alpha-lattice
3 replications





RCBD
3 replications

Panel	Trial	Crop cycle	Morphological traits		Technological traits		Ligno-cellulosic traits			Disease traits		
			SN	SD	BR	BC	ADL	ADF	IVNDFD	RST	SCYLV	SM
REU	La Mare	2010	■	■	■	■						
	Vue-Belle	2010	■	■	■	■						
	Bassin Martin	2007	■	■	■	■	■	■	■	■	■	
		2008									■	
		2009			■	■	■	■	■		■	
	Le Gol	2007										■
		2008										■
		2009										■
GUA	Roujol-1	2005								■	■	■
		2006										■
		2007									■	■
	Roujol-2	2008	■	■	■	■	■	■	■		■	■
		2009	■	■		■	■	■	■			■
		2010	■	■	■						■	■
	Godet	2010	■	■	■							
		2011	■	■	■							

Mixed models

variety BLUPs

variety BLUPs

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}_1\mathbf{b} + \mathbf{Z}_2\mathbf{c} + \mathbf{Z}_3\mathbf{cl} + \mathbf{e}$$

β : vector of fixed effects (location, crop cycle and replication)
b : vector of random incomplete block effects within each replication
c : vector of random effects of clones
cl : vector of random effects of interaction between genotypes and location or crop cycle
e : vector of residual error of the model

Summary statistics from mixed models

Traits	Panel	$\hat{\sigma}_G^{2a}$	$\hat{\sigma}_e^{2b}$	$H^{2\,c}$	$CV_g{}^{d}$	Mean ± SEM
Morphological traits						
SN (stalk/m²)	REU	79.16	99.13	0.80	23	39.21 ± 0.40
	GUA	118.48	51.14	0.90	26.1	41.66 ± 0.42
SD (mm)	REU	6.29	3.52	0.89	9.8	25.93 ± 0.10
	GUA	7.85	0.54	0.96	10.2	26.86 ± 0.07
Technological traits						
BR (%)	REU	0.98	1.29	0.83	5.4	17.94 ± 0.05
	GUA	1.63	0.58	0.88	6.2	20.49 ± 0.06
BC (%)	REU	2.04	1.59	0.89	8.3	17.52 ± 0.05
	GUA	1.81	0.88	0.71	8.1	15.98 ± 0.07
Lignocellulose traits						
ADL (%)	REU	0.32	0.37	0.84	4.9	11.77 ± 0.03
	GUA	0.68	0.092	0.87	6.9	11.88 ± 0.04
ADF (%)	REU	1.2	1.9	0.78	1.7	62.62 ± 0.06
	GUA	2.1	0.22	0.86	2.3	62.01 ± 0.07
IVNDFD (%)	REU	2.84	4.97	0.79	19.4	8.70 ± 0.13
	GUA	6.99	0.84	0.87	27.8	9.52 ± 0.15
Disease traits						
RST (score)	REU	4.12	0.51	–	–	2.95 ± 0.08
	GUA	13.17	0.48	–	–	3.19 ± 0.08
SCYLV (%)	REU	28.58	0.03	–	–	72.33 ± 1.85
	GUA	10.90	0.12	–	–	76.57 ± 1.12
SM (whip/m²)	REU	5.06	1.13	–	–	5.91 ± 0.69
	GUA	5.99	1.76	–	–	5.48 ± 0.51

GS models

- RR : Ridge-Regression (parametric) [rrBLUP]
- BL : Bayesian LASSO regression (parametric) [BLR]
- RKHS : Reproducing Kernel Hilbert Spaces regression (semi parametric) [rrBLUP]
- PLS : Partial Least Square regression (non parametric) [PLS]

GUA

167 cultivars
(origine mondiale)

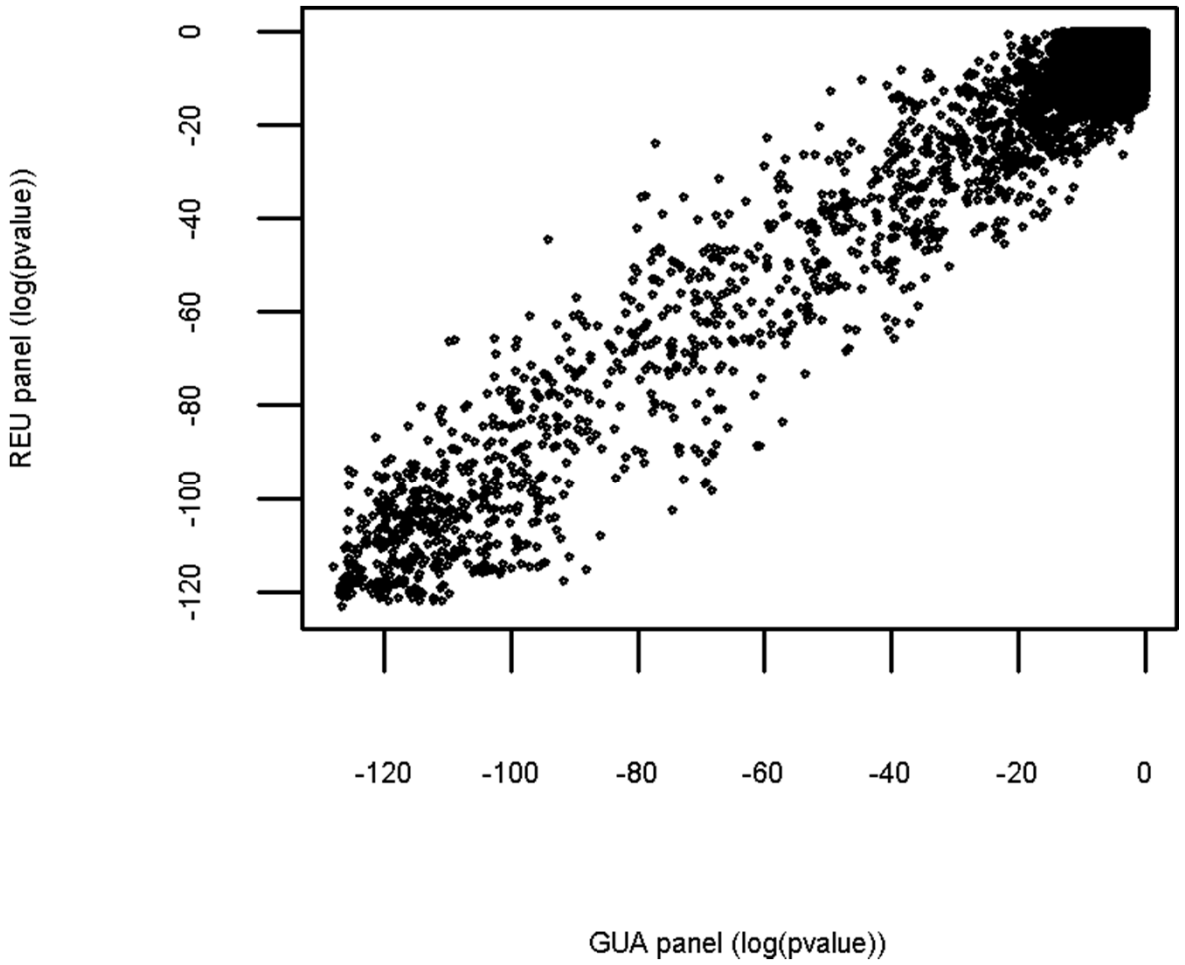
REU

167 cultivars
(origine mondiale)

1499 DArT

0.05<fréquence<0.95
missing data < 10%

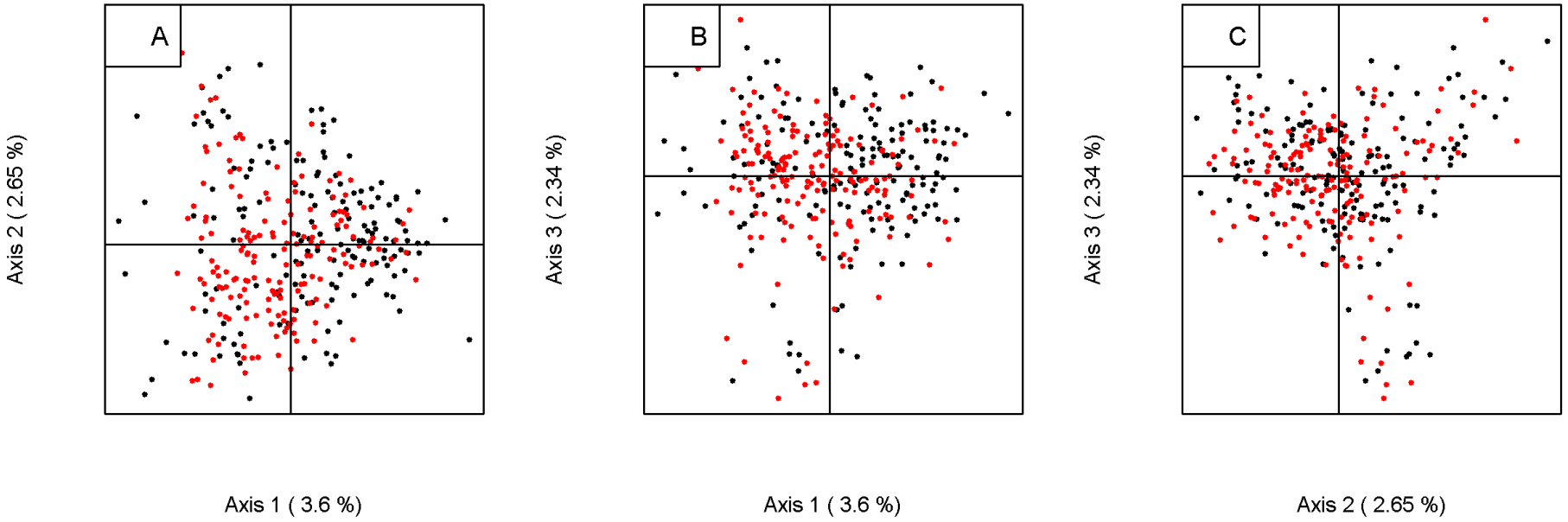
LD pattern in panels



1499 DArT

1 122 751 Tests Exacts de Fischer (pairwise marker associations)

Genetic divrsity(PCA analses)



GUA : points rouges

REU : points noirs

No disjunction of the two panels (similar organization of the genetic diversity)

Accuracy of GS

accuracy of genomic selection prediction =
correlation between genetic values (GV) predicted by GS models and observed genetic values (BLUPs)

Validation for each GS methods :

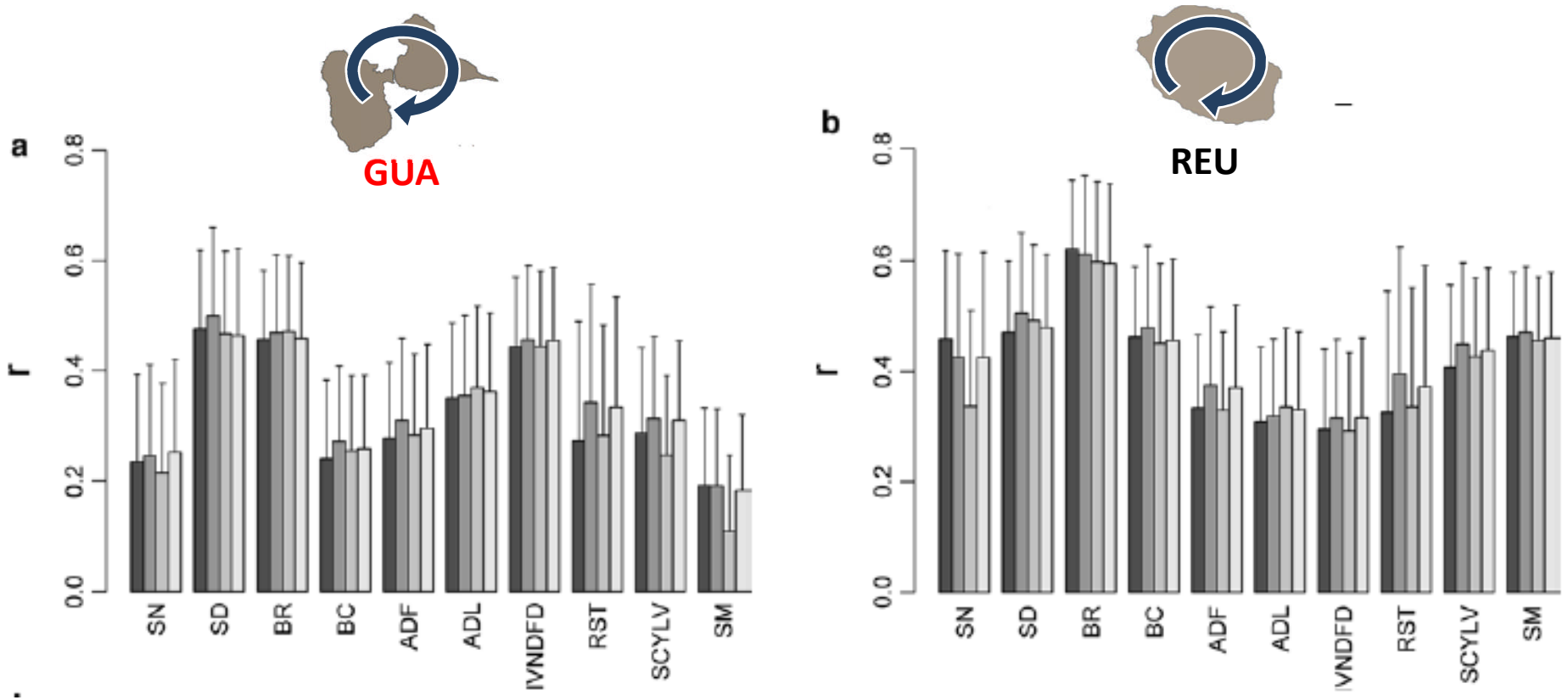
- **within panel**
- **between panels**

Cross-validation within panels

- For each data set, the 167 accessions were randomly split into five subsets
  - 4 used as the training set
  - 1 for prediction
- Random sampling of the training and validation sets repeated 500 times.
- Standard deviation (5^{th} – 95^{th} percentiles of GS predictions)

Validation within panel

Fig. 3 Median correlations (Pearson's coefficient) between observed genetic values (GV) and predicted genetic values (PGV) in a fivefold within-panel cross validation. Four genomic selection methods were compared. From darkest to lightest: bayesian LASSO, reproducing kernel Hilbert space, partial least square regression, and ridge regression. Ten traits were predicted: stalk diameter (SD), stalk number (SN), brix of the juice (BR), bagasse content (BC), acid detergent lignin as a percentage of neutral detergent fiber (ADL), acid detergent fiber as a percentage of neutral detergent fiber (ADF), in vitro neutral detergent fiber digestibility of the bagasse (IVNDFD), rust resistance (RST), yellow leaf disease resistance (SCYLV), and smut resistance (SM). **a** Cross validation within the GUA panel. **b** Cross validation within the REU panel. For RST, we focused on accessions which do not carry the major resistance gene *Bru*1. *Vertical lines* over the bars represent absolute value of the standard deviations



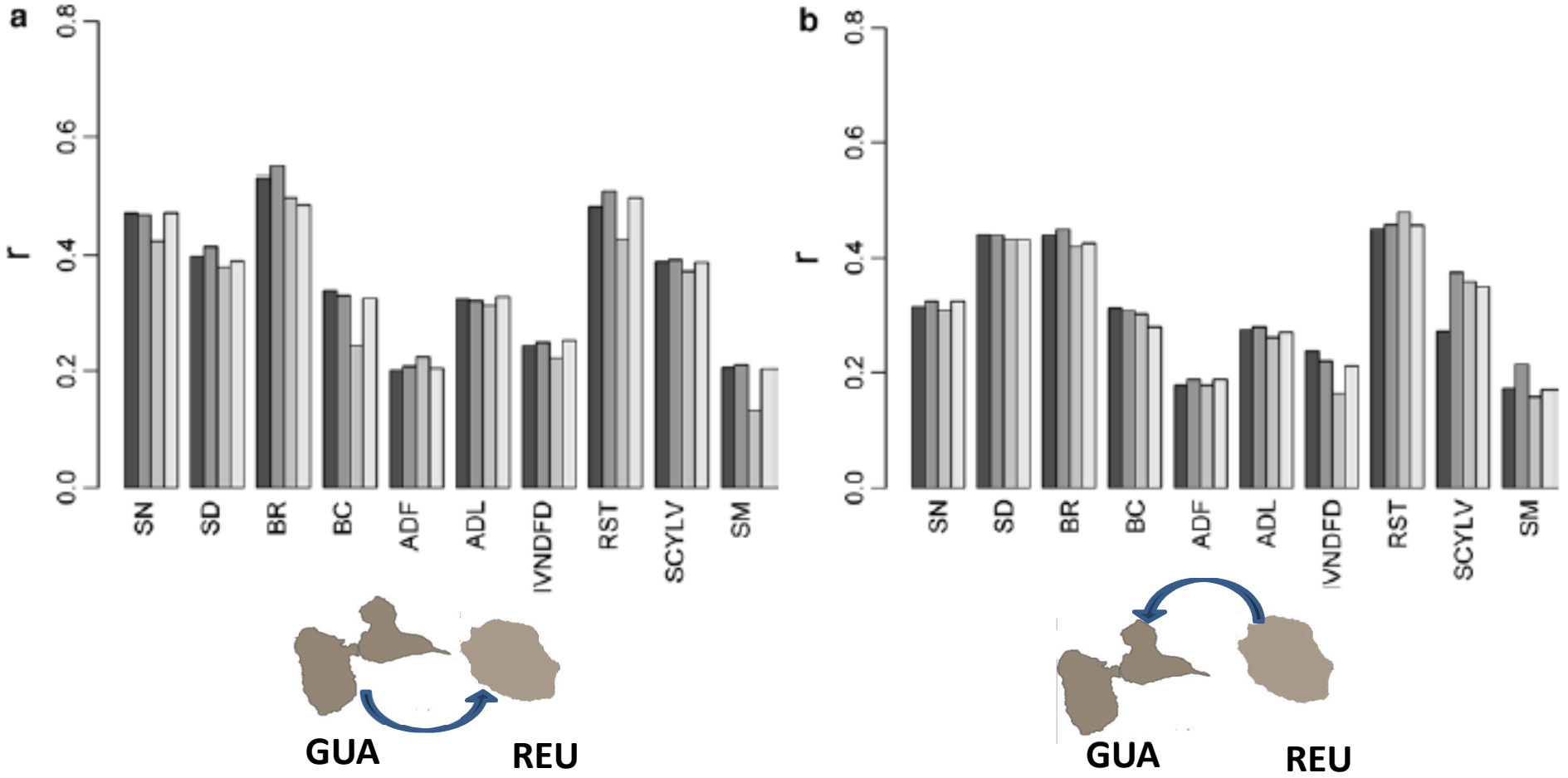
Cross validation between panels

Interchanging training and validation panels:

- Guadeloupe panel for training to predict Reunion panel
- Reunion panel for training to predict Guadeloupe panel

Cross validation between panels

Fig. 4 Correlation coefficients (Pearson's coefficient) between observed genetic values (GV) and predicted genetic values (PGV) obtained using cross validation between two independent panels. Four genomic selection methods were compared. From darkest to lightest: bayesian LASSO, reproducing kernel Hilbert space, partial least square regression and ridge regression. Ten traits were predicted: stalk diameter (SD), stalk number (SN), brix of the juice (BR), bagasse content (BC), acid detergent lignin as a percentage of neutral detergent fiber (ADL), acid detergent fiber as a percentage of neutral detergent fiber (ADF), in vitro neutral detergent fiber digestibility of the bagasse (IVNDFD), rust resistance (RST), yellow leaf disease resistance (SCYLV), and smut resistance (SM). **a** The GUA panel was used as training population to predict the REU panel. **b**



Summary

- Surprising encouraging predictions levels even with relatively low marker density
  (ranges of accuracies are similar to several published GS results)

- Four models currently available in GS literature :

  - No major difference in prediction accuracy between the methods (large number of small QTL ? weak QTL-marker associations ?)
  - Large differences between traits (0.11 to 0.62)

Factors that could have a positive influence :

- Traits assessed repeatedly across times and sites within each region
- Sound variety BLUPs : large entry-means basis (H^2 values)
- Buffering of inter-annual variations /crop-cycles and of sites differences

- Persistence of LD across panels (similar genetic diversity pattern)

Next steps

- Larger training populations
- Larger number of markers (SNP technology) anchored to a reference sequence
- Predictions based on haplotypes (to be defined with a reference sequence)
- Pedigree taken into account

Cross validation inter-panels