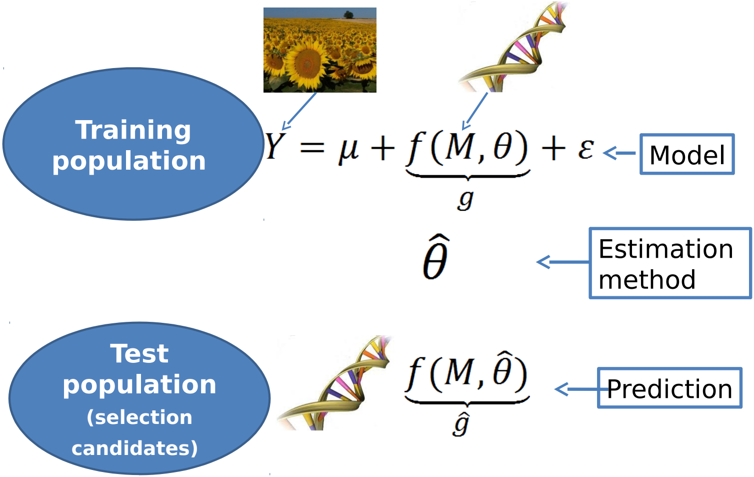
The SgenoLasso for gene mapping and genomic prediction

Charles-Elie Rabier, Céline Delmas

IMAG, Institut Montpelliérain Alexander Grothendieck
Key Initiative MUSE Data & Life Sciences
Université de Toulouse, INRAE, UR MIAT

08/25/2022

Genomic Selection (GS)



Training population

$$Y = \mu + \underbrace{f(M,\theta)}_{g} + \varepsilon$$

Model

$$\hat{\theta}$$

Estimation method

Test population (selection candidates)

$$\underbrace{f(M,\hat{\theta})}_{\hat{g}}$$

Prediction

GS motivated by Meuwissen et al (Genetics, 2001)

Selective Genotyping is highly linked to GS

Genotyping was expensive in the past

⇒ Selective Genotyping : we genotype only individuals who present extreme phenotypes Y

At a given power, a large increase of the number of individuals
leads to a decrease of the number of individuals genotyped

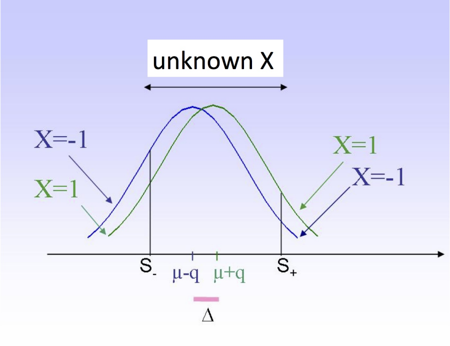
Lebowitz et al. (Theoretical and Applied Genetics, 1987)
Darvasi and Soller (Theoretical and Applied Genetics, 1992)

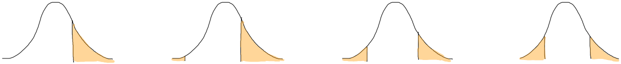
To go further in the statistical theory :

R. (Journal of Statistical Planning and Inference, 2014)

Model corresponding to selective genotyping

Probability distribution of the phenotypes *Y*





Worst scenario

Best scenario

Can we elaborate a method able to learn
a model based on extreme individuals ?

Context of our study

- The chromosome is represented by a segment $[0, T]$
- The distance on $[0, T]$ is called the genetic distance
- $X(.)$: random process representing the genome of one individual
- We consider Haldane modeling

Haldane Modeling (1919)

- no crossover interference
- $X(t)$: random variable corresponding to the genome information at t

$$X(0) \sim \frac{1}{2}(\delta_{+1} + \delta_{-1}), \quad X(t) = X(0)(-1)^{N(t)}$$

where $N(.)$ is a Poisson process with intensity 1 on $[0, T]$

- $r(t, t')$: probability of recombination between two loci

$$\begin{aligned} r(t, t') &= \mathbb{P}(X(t)X(t') = -1) = \mathbb{P}(\left|N(t) - N(t')\right| \text{ odd}) \\ &= \frac{1}{2}\left(1 - e^{-2\left|t-t'\right|}\right) = \frac{1}{2}\left(1 - \rho(t, t')\right) \end{aligned}$$

Model

- K genetic markers on $[0, T]$ located at

$$t_1 = 0 < t_2 < ... < t_K = T$$

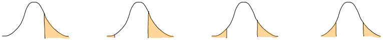
- m QTLs (i.e. Quantitative Trait Loci) located at

$$0 \leq t_1^\star < t_2^\star < \ldots < t_m^\star \leq T$$

- Assuming a linear model for the phenotype Y

$$Y = \mu + \sum_{s=1}^{m} X(t_s^\star) q_s + \sigma\varepsilon \qquad \text{with} \quad \varepsilon \sim N(0,1)$$

- Genome information X(.) available :
  - only at genetic markers $t_1, \ldots, t_K$
  - only if Y is extreme (i.e. $Y > S_+$ or $Y < S_-$)



$\Rightarrow$ Dependency between the alleles at the markers and the extreme phenotypes Y

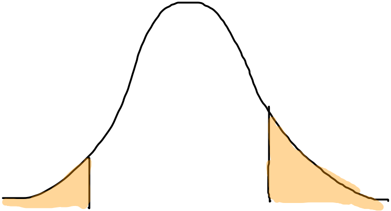
One observation

$\overline{X}(t)$ is the random variable such as

$$\overline{X}(t) = \begin{cases} X(t) & \text{if } \; Y \notin [S_- \; , \; S_+] \\ 0 & \text{otherwise} \end{cases}$$

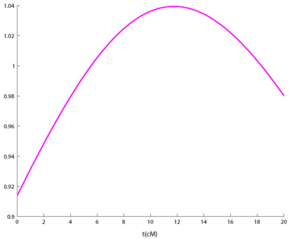
then, under our selective genotyping framework, one observation is

$$\left(Y, \; \overline{X}(t_1), \; \overline{X}(t_2), \; \ldots, \; \overline{X}(t_K) \right) .$$



The Interval Mapping of Lander and Botstein (1989)

- It assumes a maximum of $m = 1$ QTL
- $\Lambda_n(t)$: Likelihood Ratio Test at a given location $t \in [0, T]$, for testing $q_1 = 0$ vs $q_1 \neq 0$
- $\Lambda_n(.)$: Likelihood Ratio Test process on $[0, T]$
- $\sup_{t \in [0,T]} \Lambda_n(t)$: Likelihood Ratio Test of H_0 “no QTL on $[0, T]$” vs H_1 “there exists one QTL at $t_1^\star$”, i.e. LRT on the whole interval
- $\arg\sup \Lambda_n(.)$: natural estimator for the QTL location



One path of the process $\Lambda_n(.)$ ($T = 20$cM, $K = 2$)

The true probability distribution when $m = 1$

When only one QTL lies on the genome (i.e. $m = 1$) at $t = t_1^\star$:

$$L_{t_1^\star}(q_1,\ \mu,\ \sigma) = \Big[\Big\{p(t_1^\star) f_{(\mu+q_1,\sigma)}(Y) + (1 - p(t_1^\star)) f_{(\mu-q_1,\sigma)}(Y)\Big\}\, 1_{Y \notin [S_-, S_+]} + \left\{\frac{1}{2} f_{(\mu+q_1,\sigma)}(Y) + \frac{1}{2} f_{(\mu-q_1,\sigma)}(Y)\right\} 1_{Y \in [S_-, S_+]}\Big]\ g(.)$$

where

- $p(t_1^\star) = P[X(t_1^\star) = 1 | X(t_1), \cdots, X(t_K)]$
- $f_{(m,\sigma)}$ is the Gaussian density with parameters (m, σ)
- $g(.)$ is a function which does not depend on parameters q_1, μ and σ

Score statistic and LRT statistic

- $\theta^1 = (q_1,\ \mu,\ \sigma)$ parameter of the model at t fixed
- $\theta_0^1 = (0,\ \mu,\ \sigma)$ stands for H_0

Score statistic at t :

$$S_n(t) = \frac{\frac{\partial l_t^n}{\partial q_1} \, |_{\theta_0^1}}{\sqrt{\mathrm{Var}\left(\frac{\partial l_t^n}{\partial q_1} \, |_{\theta_0^1}\right)}} \ ,$$

with $l_t^n(\theta^1)$ log likelihood at t, associated to n observations.

LRT statistic at t :

$$\Lambda_n(t) = 2\left\{ l_t^n(\widehat{\theta_1}) - l_t^n(\widehat{\theta}_{1\,|H_0}) \right\} \quad ,$$

with $\widehat{\theta_1}$ MLE, and $\widehat{\theta}_{1\,|H_0}$ MLE under H_0.

- known $t_1^\star \Rightarrow$ regular model
- unknown $t_1^\star \Rightarrow$ irregular model (under H_0, the Fisher Information Matrix relative to t is equal to zero)

Hypothesis studied and extra notations

We will study the asymptotic properties of $S_n(.)$ and $\Lambda_n(.)$ under the following hypothesis :

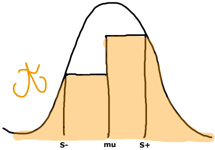
$H_{at^\star}$: “there are m QTL located at $t_1^\star$, ..., $t_m^\star$ with effects $q_1 = a_1/\sqrt{n}, \ldots, q_m = a_m/\sqrt{n}$ where $a_1 \neq 0, \ldots, a_m \neq 0$” .

A few extra notations :

- $\mathbb{T}_K := \{t_1, ..., t_K\}$
- $t^\ell := \sup \{t_k \in \mathbb{T}_K : t_k < t\}$
- $t^r := \inf \{t_k \in \mathbb{T}_K : t < t_k\}$

In other words, t belongs to the “Marker interval” (t^ℓ, t^r)

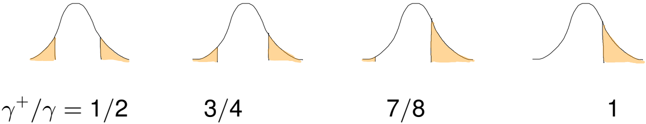
A key factor linked to selection intensity



About the key factor linked to selection intensity

$$
\begin{aligned}
\mathcal{A} &:= \sigma^2 \; \{\gamma \; + \; z_{\gamma_+} \; \varphi(z_{\gamma_+}) \; - \; z_{1-\gamma_-} \; \varphi(z_{1-\gamma_-})\} \\
\gamma &:= \mathbb{P}_{\mathcal{H}_0}\left(Y \notin [S_-, \; S_+]\right) \\
\gamma_+ &:= \mathbb{P}_{\mathcal{H}_0}\left(Y > S_+\right) \\
\gamma_- &:= \mathbb{P}_{\mathcal{H}_0}\left(Y < S_-\right)
\end{aligned}
$$

where $\varphi(x)$ and z_α denote respectively the density of a standard normal distribution taken at the point x, and the quantile of order $1 - \alpha$ of a standard normal distribution.



A non linear interpolation on the “Marker interval” (t^ℓ, t^r)

Theorem (R. & Delmas, Statistics 2021)

$$S_n(.) \Rightarrow Z(.) \quad , \quad \Lambda_n(.) \overset{F.d.}{\rightarrow} Z^2(.) \quad , \quad \sup \Lambda_n(.) \xrightarrow{\mathcal{L}} \sup Z^2(.) \qquad \textit{where}$$

- *$Z(.)$ is the non linear interpolated process such as*

$$\forall t \in [0,T]\backslash\mathbb{T}_K \quad Z(t) = \frac{\alpha(t)\ Z(t^\ell)\ +\ \beta(t)\ Z(t^r)}{\sqrt{\alpha^2(t)+\beta^2(t)+2\alpha(t)\beta(t)\rho(t^\ell,t^r)}}$$

$$\textit{with} \quad Cov\{Z(t_k),Z(t_{k'})\} = \rho(t_k,t_{k'}) \quad \forall (t_k,t_{k'}) \in \mathbb{T}_K \times \mathbb{T}_K$$

- *$Z(.)$ is a Gaussian process with unit variance and with expectation :*

$$\textit{under } H_{at^\star} \ : \ m_{t^\star}(t^\ell) = \sum_{s=1}^{m} a_s \sqrt{\mathcal{A}}\ \rho(t^\ell, t_s^\star)/\sigma^2 \quad , \quad m_{t^\star}(t^r) = \sum_{s=1}^{m} a_s \sqrt{\mathcal{A}}\ \rho(t_s^\star, t^r)/\sigma^2$$

$$\forall t \in [0,T]\backslash\mathbb{T}_K \quad m_{t^\star}(t) = \frac{\alpha(t)\ m_{t^\star}(t^\ell)\ +\ \beta(t)\ m_{t^\star}(t^r)}{\sqrt{\alpha^2(t)+\beta^2(t)+2\alpha(t)\beta(t)\rho(t^\ell,t^r)}}$$

Intuition on asymptotic theory

At a marker t_k, the score statistic can be decomposed in the following way :

$$S_n(t_k) = \sum_{j=1}^{n} \sum_{s=1}^{m} \frac{q_s \, \overline{X}_j(t_s^\star) \, \overline{X}_j(t_k)}{\sqrt{n \, \mathcal{A}}} + \sum_{j=1}^{n} \frac{\sigma \varepsilon_j \, \overline{X}_j(t_k)}{\sqrt{n \, \mathcal{A}}}$$

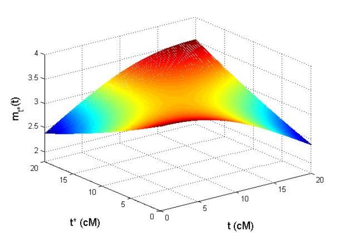
Then, according to a technical proof , we have the relationship

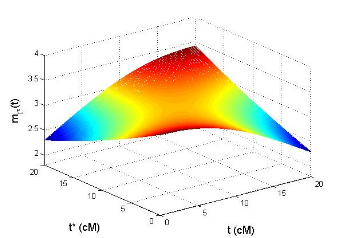
$$\sum_{j=1}^{n} \frac{\sigma \varepsilon_j \, \overline{X}_j(t_k)}{\sqrt{n \, \mathcal{A}}} \xrightarrow{\mathcal{L}} \mathcal{N}\left[\Omega, \ 1\right]$$

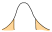
where Ω is a function of $a_1, \ldots, a_m, t_1^\star, \ldots, t_m^\star, t_k, S_-$ and S_+.

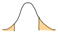
The correlation between ε and $\overline{X}(t_k)$ plays a role in the asymptotic theory

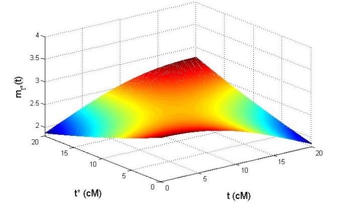
Mean function under selective genotyping ($K = 2$ markers, $T = 20$cM, $m = 1$ QTL)







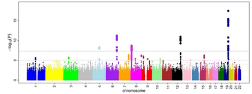




Introducing the SgenoLasso

1) we discretize the process at marker locations

$$\vec{S}_n = \vec{m}_{t^\star} \; + \; \vec{\varepsilon} \; + \; o_P(1)$$



where $\vec{S}_n = (S_n(t_1)\ ,\ S_n(t_2)\ ,\ \dots\ ,\ S_n(t_K))'$

$$\vec{m}_{t^\star} = (m_{t^\star}(t_1)\ ,\ m_{t^\star}(t_2)\ ,\ \dots, m_{t^\star}(t_K))'$$

$$\vec{\varepsilon} \sim N(0, \Sigma) \text{ with } \Sigma_{kk'} = \operatorname{Cov}\left(Z(t_k), Z(t_{k'})\right)$$

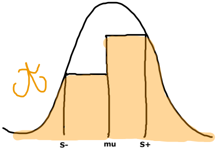
2) we decorrelate the process

Let $\mathbb{T}_K^\star := \{t_1^\star, \dots, t_m^\star\}$ and $\Sigma := BB'$, we have

$$B^{-1}\vec{S}_n = B'\Delta \; + \; B^{-1}\vec{\varepsilon} \; + \; o_P(1)$$

where $\Delta := (\Delta_1, \dots, \Delta_K)'$

$$\text{and} \quad \Delta_k = \begin{cases} 0 & \text{if} \quad t_k \notin \mathbb{T}_K^\star \\ \frac{a_s}{\sigma}\ \frac{\sqrt{\mathcal{A}}}{\sigma} & \text{if } t_k \in \mathbb{T}_K^\star \text{ with } s \mid t_s^\star = t_k \end{cases}$$



Introducing the SgenoLasso

In fact, non null Δ_k are unknown
⇒ L1 penalized regression Lasso (Tibshirani, 1996)

$$\hat{\Delta}_{\text{SgenoLasso}}(\lambda, \alpha) = \arg\min_{\Delta} \left(\left\| B^{-1}\vec{S}_n - B'\Delta \right\|_2^2 + \lambda \left\| \Delta \right\|_1 \right)$$

SgenoLasso presents all the properties of the classical Lasso !

Its β-min condition :

$$\min_{s | t_s^\star \in \mathbb{T}_K} \frac{|a_s| \sqrt{\mathcal{A}}}{\sigma^2 \sqrt{K}} >> \Phi^{-2} \sqrt{\frac{m \log(K)}{K}}$$

Its irrepresentable condition :

$$\left\| \Sigma^{(.,\star)} (\Sigma^{(\star,\star)})^{-1} \text{Sign}(a_1, \ldots, a_m) \right\|_\infty \leq C < 1$$

where $\|x\|_\infty = \max_j |x_j|$, $\text{Sign}(a_1, \ldots, a_m) = (\text{Sign}(a_1), \ldots, \text{Sign}(a_m))^\top$

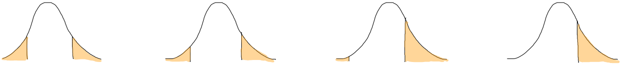
β-min condition + irrep cond ⇒ consistent variable selection

Applications to association studies (Simulated data)

$(n = 500, \gamma = 20\%)$ or $(n = 333, \gamma = 30\%)$
$K =$10,000 markers on $[0, 10M]$ / 1,000 markers on $[0, 1M]$
$m = 16$ QTLs located only on $[0, 1M]$

L1 ratio $\sum_{i=1}^{1000} |\hat{\Delta}_i| / \sum_{i=1}^{10000} |\hat{\Delta}_i|$

γ	γ^+/γ	SgenoLasso	Lasso	Group Lasso	EN	RALasso
0.2	1/2	94.19%	91.69%	97.46%	97.44%	98.09%
	3/4	91.52%	84.75%	95.88%	96.02%	95.08%
	7/8	92.38%	75.46%	94.67%	95.23%	89.33%
	1	85.03%	21.14%	21.86%	27.37%	44.93%
0.3	1/2	91.62%	83.45%	92.87%	93.67%	95.36%
	3/4	90.88%	76.18%	89.59%	91.10%	91.13%
	7/8	86.22%	65.03%	78.00%	82.84%	80.32%
	1	78.00%	20.92%	20.82%	24.92%	48.25%



$\gamma^+/\gamma =$1/2 3/4 7/8 1

The SgenoLasso has several cousins

SgenoLasso is built on the L1 penalty of Lasso (Tibshirani, 1996)

$$\hat{\Delta}_{\text{SgenoLasso}}(\lambda,\alpha) = \arg\min_{\Delta}\left(\left\|B^{-1}\vec{S}_n - B'\Delta\right\|_2^2 + \lambda\left\|\Delta\right\|_1\right)$$

SgenoElasticNet is built on the mixture of L1 and L2 penalties of Elastic Net (Zou and Hastie, 2005)

$$\hat{\Delta}_{\text{SgenoEN}}(\lambda,\alpha) = \arg\min_{\Delta}\left(\left\|B^{-1}\vec{S}_n - B'\Delta\right\|_2^2 + \frac{1-\alpha}{2}\left\|\Delta\right\|_2^2 + \alpha\left\|\Delta\right\|_1\right)$$

SgenoGroupLasso is built on the Group Lasso penalty (Yuan and Lin, 2006)

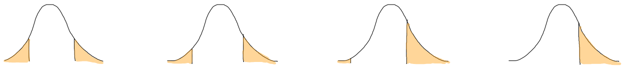
$$\hat{\Delta}_{\text{SgenoGroupLasso}}(\lambda) = \arg\min_{\Delta}\left(\left\|B^{-1}\vec{S}_n - B'\Delta\right\|_2^2 + \lambda\sum_{i=1}^{\text{nbGroup}}\sqrt{L_i}\left\|\vec{\Delta}_i\right\|_2\right)$$

The SgenoLasso has several cousins

10,000 markers on [0, 10M] / 1,000 markers on [0, 1M]
16 QTLs located only on [0, 1M]

L1 ratio $\sum_{i=1}^{1000} |\hat{\Delta}_i| / \sum_{i=1}^{10000} |\hat{\Delta}_i|$

		SgenoLasso	SgenoGroupLasso	SgenoEN
γ	γ^+/γ	L1 ratio	L1 ratio	L1 ratio
0.2	$1/2$	94.19%	98.33%	96.03%
	$3/4$	91.52%	95.38%	92.59%
	$7/8$	92.38%	96.83%	93.19%
	1	85.03%	90.53%	84.93%
0.3	$1/2$	91.62%	92.35%	86.53%
	$3/4$	90.88%	94.84%	91.84%
	$7/8$	86.22%	89.96%	86.68%
	1	78.00%	82.61%	77.23%

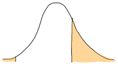


$\gamma^+/\gamma = 1/2$ $3/4$ $7/8$ 1

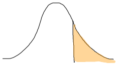
The predictive ability of the SgenoLasso (simulated data, K=10,000 markers)

Accuracy criterion Cor($\hat{y}, y$)

γ	γ^+/γ	SgenoLasso	Lasso	Group Lasso	EN	RaLasso
0.1	1	30.97%	6.49%	3.17%	4.38%	10.43%
	7/8	31.25%	30.55%	29.87%	29.74%	28.78%
0.2	1	27.88%	7.12%	4.05%	5.41%	11.08%
	7/8	28.26%	27.98%	27.86%	28.09%	26.28%
0.3	1	26.79%	9.02%	6.89%	7.48%	11.96%
	7/8	28.13%	27.85%	26.59%	28.25%	26.05%



$\gamma^+/\gamma = 7/8$



$\gamma^+/\gamma = 1$

Our answer to Brandariz and Bernardo (Crop Science, 2018) :
no need to keep the worst individuals in the breeding programs

Rice real data

Data from Spindel et al. (Plos Genetics, 2015) and from Begum et al. (Plos One, 2015)

- Trait of interest : flowering date during the dry season 2012
- K =13,101 markers, randomly chosen by the authors from their 73,147 collected markers
- $n = 312$ in total (i.e. under complete genotyping)
- only 93 extreme individuals when $\gamma = 0.3$
- we performed a symmetrical selective genotyping (i.e. $\gamma^+/\gamma = 1/2$)

Rice real data

γ	Method	Selected QTLs
1	Begum et al.	S3-1125848, S3-1165376, S3-1221494, S3-1269941, S3-1394477
0.3	SgenoLasso	4 QTLs matching those of Begum et al. (2015)
0.3	SgenoEN	5 QTLs matching those of Begum et al. (2015)
0.3	SgenoGroupLasso	5 QTLs matching those of Begum et al. (2015)
0.3	Lasso	2 QTLs matching those of Begum et al. (2015)
0.3	EN	5 QTLs matching those of Begum et al. (2015)
0.3	Group Lasso	3 QTLs matching those of Begum et al. (2015)

Thank you for listening

A few references :

- S.P. Brandariz and R. Bernardo. *Maintaining the Accuracy of Genomewide Predictions when Selection Has Occurred in the Training Population*, Crop Science (2018)
- D. Darvasi, M. Soller, *Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus*, Theor. Appl. Genet. (1992).
- J. Fan, Q. Li, Y. Wang. *Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions*, Journal of the Royal Statistical Society : Series B (Statistical Methodology) (2017)
- E.S. Lander and D. Botstein. *Mapping mendelian factors underlying quantitative traits using RFLP linkage maps*, Genetics (1989)
- R.J. Lebowitz, M. Soller, J.S. Beckmann. *Trait-based analyses for the detection of linkage between marker loci and quantitative trait loci in crosses between inbred lines*, Theor. Appl. Genet. (1987).
- C.E. Rabier. *On statistical inference for selective genotyping*, J. Stat. Plan. Infer. (2014)
- C.E. Rabier. *On stochastic processes for Quantitative Trait Locus mapping under selective genotyping*, Statistics (2015)
- C.E. Rabier and C. Delmas. *The SgenoLasso and its cousins for selective genotyping and extreme sampling*, Statistics (2021)
- R. Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society B (1996).
- M. Yuan, Y. Lin, *Model selection and estimation in regression with grouped variables*, Journal of the Royal Statistical Society Series B (2006).
- Y. Zhao, M. Gowda, F.H. Longin, T. Würschum, N. Ranc, J.C. Reif, *Impact of selective genotyping in the training population on accuracy and bias of genomic selection*, Theoretical and Applied Genetics (2012).
- H. Zou, T. Hastie, *Regularization and variable selection via the elastic net*, Journal of the Royal Statistical Society : Series B (Statistical Methodology), (2005).