GS rust

Salvador Osuna-Caballero

6/12/2021

Table of Contents

[Introduction 1](#_Toc95298395)

[Scientific questions 1](#_Toc95298396)

[Material and Methods 2](#_Toc95298397)

[Plant Material 2](#_Toc95298398)

[Phenotyping and statistical analysis 2](#_Toc95298399)

[Genotyping and data filtering 3](#_Toc95298400)

[Genomic regression models and data configurations 3](#_Toc95298401)

[Results 4](#_Toc95298402)

[Phenotypic variation and genotype × environment interaction 4](#_Toc95298403)

[Genome-enabled modeling 5](#_Toc95298404)

# Abstract

Genomic selection (GS) is a breeding tool, which is rapidly gaining popularity for plant breeding, particularly for traits that are difficult to measure. One such trait is rust resistance in pea (Pisum sativum L.), which is difficult to assay because it is strongly influenced by the environment. Here we report a study of the efficacy of GS for predicting rust resistance in pea, as represented by data collecting from field and controlled conditions and using DArT-Seq markers acquired through genotyping-by-sequencing. The effects on prediction accuracy of different GS models including the genomic correlation approach were compared using cross-validation. Additionally, the inclusion of marker × environment interactions in a genomic best linear unbiased prediction (GBLUP) model was evaluated. Finally, different ways of combining trait data from two environment using single traits or multi-trait index (MTI) analyses were compared to results obtained using a mean value. The best prediction accuracy achieved for MTI was 0.56, obtained using GBLUP analysis with a mean value for ASC. GBLUP and Bayesian Reproducing kernel Hilbert spaces regression (RKHS) performed slightly better than the other models trialed, whereas different missing data thresholds made minimal differences to prediction accuracy. The prediction accuracies of individual, randomly selected, testing/training partitions were highly variable, highlighting the effect that the choice of training population has on prediction accuracy. The inclusion of marker × environment interactions did not increase the prediction accuracy for lines which had not been phenotyped but did improve the results of prediction across environments. GS is potentially useful for pea breeding programs pursuing ascochyta blight resistance, both for predicting breeding values for lines that have not been phenotyped, and for providing enhanced estimated breeding values for lines for which trait data is available.

# Introduction

## Scientific questions

1. Is a multi-trait index approach a better predictor for rust disease than a single trait in controlled conditions?
2. In case the multi-trait index has a better predictive ability than single trait, either in controlled condition or field, is the multi-trait index a better predictor than single-trait for rust disease in field?

* Here we propose two scenarios:
  + the predictive ability over field (new environment) testing new lines [CV2].
  + the predictive ability over field (new environment) testing the same lines used for training in controlled condition (old environment) [CV1].

1. Is the predictive ability, in G-BLUP model case, affected by the marker per environment effect?

# Material and Methods

## Plant Material

The pea panes has 320 genotypes of *Pissum* spp. including wild relatives, landraces, cultivars, breeding lines and unknown material from all continents. It also has a representative collection of *Pissum* genre, including the three main species (*P. sativum*, *P. fulvum* and *P. abyssinucum*) and some interesting *P. sativum* subspecies (arvense, elatius, *transcaucasicum*, *choresmicum* and cinereum).

## Phenotyping and statistical analysis

The pea panel has been evaluated in controlled and field conditions for rust disease. The trait evaluated in field was disease severity (DS), as a subjective estimation of the damage caused by rust (% rust damage covering the whole plant). In controlled conditions the traits evaluated were:

* IF. Infection frequency as number of pustules per cm2 of leaf, counted every day from day 7 after inoculation to day 14, when the life cycle of rust ended.
* AUDPC. The IF measured every day allow to calculate the area under the disease progress curve.
* IT. The infection type according to (Stackman, Stewart, and Loegering 1962)
* DS. Percentage of pustules over the plant.

In controlled conditions the 320 accessions were grown with two replicates and artificially inoculated with rust spores at day ~12 post sown and two different inoculations were done, so 4 replicates per accession are considered. In field, every experimental unit (10 seeds of the same accession in a single row) were sown three times in a randomized complete block design (RCBD) with Cartouche cv. as control.

Rust traits underwent an analysis of variance (ANOVA) aimed to test the variation between environments (CC or field) and among genotypes and the genotype × environment (GE) interaction. The extent of GE environments was estimated by the genetic correlation () for genotype rust responses across environments calculated according to Burdon et al. 1977. For each environment, components of variance relative to variation among genotypes () and experimental error () were estimated by a restricted maximum likelihood (REML) method, to compute best linear unbiased prediction (BLUP) genotype values according to DeLacy et al. (1996) that were used as phenotyping data for compute a multi-trait index based on factor analysis and ideotype-design proposed by Rocha et al. (2018) and for subsequent genomic prediction assessments. The extent of genetic variation in each trait and environment were expressed as the genetic coefficient of variation , where m = trait mean value. The broad-sense heritability on an entry mean basis in each trait and environment were estimated as reported in Nizam Uddin et al. (1994).

## Genotyping and data filtering

Pea core collection was genotyped with the DArTSeq approach by DiversityArray Ltd (Australia). For this, third composed leaves from 20 two weeks-old seedlings of each accessions grown under controlled condition were harvested, pooled together, flash frozen in liquid nitrogen and lyophilized. Then, DNA was extracted following to the method stipulated by Diversity Arrays P/L, Canberra, Australia and adjusted at 20 ng/µl prior to DArT marker analysis using the high density Pea DArTseq 1.0 array (50,000 markers) adapted for wild *Pisum spp.* accessions. Complexity reduction with the *Pst*I,-*Mse*I restriction enzymes, library construction, amplification and Illumina sequencing were performed by Diversity Arrays Technology Pty Ltd (Canberra, Australia) as described in Barilli et al. 2015. DArTSeq sequence analysis retrieve two set of markers, SNPs and presence–absence sequence variants (Silico-DArT), collectively referred to as DArT-Seq markers.

Data cleaning was then performed for both type of DArT markers to remove low quality and non-polymorphic markers as described before. Accordingly, DArT markers with > 20% missing data, minor allele frequency (MAF) < 5% and heterozygosity > 0.1% were removed from the analysis. Missing data were imputed with SVD method following Nazzicari et al. 2016 recommendations.

## Genomic regression models and data configurations

Genome-enabled predictions were based on two data set markers: SNP markers and DArT-Seq markers. We envisaged two genomic prediction models that tended to stand out for predictive ability in previous model comparisons for legume species (Annicchiarico et al. 2017b) or other crops (Soriano Viana et al. 2016), namely, Ridge regression BLUP (rrBLUP) and Bayesian Lasso (BL).

The rrBLUP model (Meuwissen et al. 2001), which assumes the effects of all loci to have a common variance, is well suited for traits that are influenced by a large number of minor genes. In its linear mixed additive model, each marker is assigned an effect as a solution of the equation:

where is the vector of observed phenotypes, is the mean of , is the genotype matrix (i.e., {0, 1, 2} for biallelic SNP markers of genotypes or {0, 1} for absence/presence sequence variants DArT-Seq markers), is the vector of marker effects, and is the vector of residuals. Solving with the standard ridge-regression method, the solution is:  
where is the ridge parameter, representing the ratio between residual and markers variance (Searle et al. 2009). Given the vector of effects, it is then possible to predict phenotypes and estimate genetic breeding values. We estimated by a REML method implemented by a spectral decomposition algorithm (Kang et al. 2008). Bayesian models assume relatively few markers with large effects, and allow markers to have different effects and variances (Wang et al. 2018). They assign prior densities to markers effects, thereby inducing different types of shrinkage. The solution is obtained by sampling from the resulting posterior density through a Gibbs sampling approach (Casella and George 1992). Among these models, we selected BL as described by Park and Casella (2008).  
The ability of genome-enabled models to predict breeding values for rust traits on pea panel was assessed using the R package GROAN (Nazzicari and Biscarini 2017). Predictive ability () was estimated as Pearson’s correlation between observed and predicted phenotypes following three CV strategies:

1. The first CV procedure, referred to a single trait and intra-environment cross-validation, was performed testing every trait per environment with their own and considering the multi-trait index as a single trait to compare their predictive ability.
2. The second CV procedure, referred to a single trait and cross-environment validation, was performed by predicting the breeding value of the untrained environment using a model trained on the remaining one, testing the same lines [CV1]
3. The third CV procedure, also referred to a single trait and inter environment cross-validation, but using new lines for predict them in a new environment not included in the training [CV2].

On the whole, we assessed 12-model configurations represented by combinations of two genomic prediction models (rrBLUP or BL), three CV procedures and two markers data set (DArT-seq or SNP). The accuracy () of genome-enabled models was estimated from and the square root of the broad-sense heritability on an entry mean basis in the validation environment () according to Lorenz et al. (2011) as

# Results

## Phenotypic variation and genotype × environment interaction

**Table 1.** Mean value, broad-sense heritability () on an entry mean basis and genetic coefficient of variation () for white pea rust disease of 320 accessions in two grown conditions (environments).

a Genetic variance always different from zero at P < 0.001

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Trait | Environment | Mean ± S.E. | ± S.E. | (%)a |
| IF | CC | 50.25 ± 1.19 | 0.76 ± 0.02 | 47.2 |
| AUDPC | CC | 179.84 ±4 .18 | 0.76 ±0.02 | 48.5 |
| IT | CC | 3.79 ± 0.01 | 0.67 ± 0.03 | 36.1 |
| DS | CC | 20.01 ± 0.29 | 0.86 ± 0.01 | 63.3 |
| DS | Field | 12.82 ± 0.41 | 0.67 ± 0.04 | 41.5 |

Table 2. Phenotypic () and genetic correlation () for rust pea disease traits of across both environments with traits combinations.

\*\*: different (P < 0.01) from zero  
a Significant (P < 0.01) genotype × environment interaction for all pairs of traits

|  |  |  |  |
| --- | --- | --- | --- |
| Pairs of environments | Pair of traits |  | a |
| CC vs Field | AUDPC vs DS | 0.22\*\* | 0.31 |
| CC vs Field | IF vs DS | 0.25\*\* | 0.35 |
| CC vs Field | IT vs DS | 0.23\*\* | 0.34 |
| CC vs Field | DS vs DS | 0.26\*\* | 0.34 |
| CC vs Field | Index vs DS | 0.30\*\* | 0.42 |

## Genome-enabled modeling

**Table 3**. Intra-environment predictive ability for rust pea disease in two environment with their traits evaluated using Ridge regression BLUP (rrBLUP) or Bayesian Lasso (BL) model training using a SNP or DArT-Seq marker data set a

a Using 50 repetitions of 10-fold stratified cross validations per individual analysis.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | CC |  |  |  |  | Field |  |
| Method | Model | AUDPC | IF | IT | DS | Index | DS | Mean |
| SNP-BLUP | rrBLUP | 0.602 | 0.572 | 0.579 | 0.601 | 0.627 | 0.615 | 0.599 |
| SNP-BLUP | BL | 0.602 | 0.569 | 0.571 | 0.604 | 0.630 | 0.611 | 0.598 |
| G-BLUP | RKHS | 0.885 | 0.858 | 0.921 | 0.889 | 0.912 | 0.882 | 0.891 |

**Table 4**. Cross-environment predictive ability and predictive accuracy for rust pea disease across different traits and two environments using Ridge regression BLUP (rrBLUP) or Bayesian Lasso (BL) model training with DArT-Seq marker data set a

a For model training with DArT-seq marker data set using 50 repetitions of 10-fold stratified cross validations per individual analysis

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | P. ability |  | P. accuracy |  |
| Training set (CC) | Validation set (Field) | rrBLUP | BL | rrBLUP | BL |
| AUDPC | DS | 0.222 | 0.221 | 0.255 |  |
| IF | DS | 0.257 | 0.256 | 0.295 |  |
| IT | DS | 0.262 | 0.265 | 0.320 |  |
| DS | DS | 0.263 | 0.261 | 0.284 |  |
| Index | DS | 0.311 | 0.313 | 0.356 |  |

**Table 5**. Cross-environment predictive ability and predictive accuracy for rust pea disease across different traits and two environments using Ridge regression BLUP (rrBLUP) or Bayesian Lasso (BL) model training with SNP marker data set a

a For model training with SNP marker data set using 50 repetitions of 10-fold stratified cross validations per individual analysis

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | P. ability |  | P. accuracy |  |
| Training set (CC) | Validation set (Field) | rrBLUP | BL | rrBLUP | BL |
| AUDPC | DS | 0.218 | 0.216 | 0.250 | 0.248 |
| IF | DS | 0.252 | 0.248 | 0.289 | 0.284 |
| IT | DS | 0.253 | 0.255 | 0.309 | 0.312 |
| DS | DS | 0.260 | 0.254 | 0.280 | 0.274 |
| Index | DS | 0.307 | 0.304 | 0.352 | 0.349 |

**Table 6**. Cross-environment predictive ability and predictive accuracy for rust pea disease across different traits and two environments using Ridge regression BLUP (rrBLUP) model training with DArT marker data set a [CV2]

a For model training with SNP marker data set using 50 repetitions of 10-fold stratified cross validations per individual analysis

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | P. ability |  | P. accuracy |  |
| Training set (CC) | Validation set (Field) | rrBLUP | BL | rrBLUP | BL |
| AUDPC | DS | 0.212 |  |  |  |
| IF | DS | 0.231 |  |  |  |
| IT | DS | 0.231 |  |  |  |
| DS | DS | 0.253 |  |  |  |
| Index | DS | 0.295 |  |  |  |