USE OF LOW-DEPTH GBS DATA FOR GENOMIC PREDICTION ACROSS DIFFERENT MULTI-PARENTAL POOLS AND SINGLE PLANTS IN PERENNIAL RYEGRASS (LOLIUM PERENNE L.).

The aim of the project:

Collaboration between university and industry

Develop methods for genomic prediction to enhance the ryegrass breeding procedure

Ryegrass: outcrossing crop, breed and maintained as multi-parental genetic pools.

P1 P2
F1 F2 bi-parental families Pools Ploidy = 4n

Single plants
Ploidy = 2n

Synthetic multi-parental Pools
Ploidy = ~16n

Official trials

Ryegrass genotyping

Genotyping family Pools

Genotyping by Sequencing

Allele FREQUENCY

<table>
<thead>
<tr>
<th>REF</th>
<th>ALT</th>
<th>CALL</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>3</td>
<td>0.33</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0.5</td>
</tr>
</tbody>
</table>

~1.5 M SNP markers per assay

Two issues

1. Reduce the sequencing coverage depth (CD)
   - Reduced accuracy of the alleles frequency estimation.
   - Increased number of missing values.

2. Combining Genotyping data with
   - Different sequencing depth
   - Different samples with different ploidy levels

Material:
945 F2 bi-parental family pools
Phenotypes: Heading Date (HD) and Crown Rust Resistance (CRR)
Genotypes: ~1.5M SNPs

Simulated reduction of sequencing power:
Four dataset considered with a reduction factors of: 100%, 25%, 12.5%, 6.25%

Reduced GBS coverage depth

Sequencing depth dropped from 12.7 for SD100% to 0.79 for SD6%
Missing rate increased from 13.54% of SD100% to 63.7% of SD6%

Collaboration between university and industry

Develop methods for genomic prediction to enhance the ryegrass breeding procedure

Ryegrass: outcrossing crop, breed and maintained as multi-parental genetic pools.

P1 P2
F1 F2 bi-parental families Pools Ploidy = 4n

Single plants
Ploidy = 2n

Synthetic multi-parental Pools
Ploidy = ~16n

Official trials

Ryegrass genotyping

Genotyping family Pools

Genotyping by Sequencing

Allele FREQUENCY

<table>
<thead>
<tr>
<th>REF</th>
<th>ALT</th>
<th>CALL</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>3</td>
<td>0.33</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0.5</td>
</tr>
</tbody>
</table>

~1.5 M SNP markers per assay

Two issues

1. Reduce the sequencing coverage depth (CD)
   - Reduced accuracy of the alleles frequency estimation.
   - Increased number of missing values.

2. Combining Genotyping data with
   - Different sequencing depth
   - Different samples with different ploidy levels

Material:
945 F2 bi-parental family pools
Phenotypes: Heading Date (HD) and Crown Rust Resistance (CRR)
Genotypes: ~1.5M SNPs

Simulated reduction of sequencing power:
Four dataset considered with a reduction factors of: 100%, 25%, 12.5%, 6.25%

Reduced GBS coverage depth

Sequencing depth dropped from 12.7 for SD100% to 0.79 for SD6%
Missing rate increased from 13.54% of SD100% to 63.7% of SD6%
Two Cross Validation strategy (using standard gBLUP model):
- Leave one out (LOO): K-fold cross validation with K equal to the numbers of the samples
- Leave Siblings Out (LSO): leave out one sample and all its siblings

- LOO yield higher Predictive abilities compared to LSO
- GP are possible even at very low SD

Cross validation – depth correction
Gain in Predictive ability when the dataset with corrected diagonal are used. Only the result obtained for the LSO cross validation strategy are showed.
- The bias correction improve the prediction accuracy
- It is more useful at lower depths
- It is relevant even at high depths

Cross validation – impuded datasets
Gain in Predictive ability when the imputed marker are used
- The missing data imputation improve the prediction accuracy
- The RF is the most effective method (high computation time)
- Imputation is more useful at lower depths
- It is still relevant even at high depths

Gain in Predictive ability when the dataset with corrected diagonal are used. Only the result obtained for the LSO cross validation strategy are showed.

GRM bias due to low depth sequencing
- Low coverage depth make it difficult to distinguish heterozygote from homozygote
- Inflated values on the diagonal
- Overestimated genetic variances

Cross validation

Reduced accuracy of the allel frequency estimation
- Gain in Predictive ability when the dataset with corrected diagonal are used. Only the result obtained for the LSO cross validation strategy are showed.

Each sample variance :
Genomic variance:
\[ \sigma^2_G = \frac{1}{n} \sum_{i=1}^{n} p_i (1 - p_i) \]
\( n \) = ploidy level of the sample

Bias variance:
\[ \sigma^2_B = \left( \frac{1}{n} \right) \sum_{i=1}^{n} p_i (1 - p_i) CD \]
\( CD \) = number of reeds for each SNP

Diagonal shrinkage:
\[ \sigma^2_C = \frac{\sigma^2_B}{\sigma^2_B + \sigma^2_G} \]

Random Forest (RF):
- Markers are classified by using regression trees.
- All available data are used to predict the missing value of each marker
- Both kNN and RF outperform the standard mean imputation (MI)
- The Random forest outperform the kNN

Gain in Predictive ability when the imputed marker are used
- The missing data imputation improve the prediction accuracy
- The RF is the most effective method (high computation time)
- Imputation is more useful at lower depths
- It is still relevant even at high depths

Increased number of missing values
Standard imputation methods cannot be applied:
- Current draft assembly is only assembled to the scaffold level (no pseudo-molecules)
- Difficult to build haplotype for Families pools

Random Forest (RF):
- Markers are classified by using regression trees.
- All available data are used to predict the missing value of each marker
- Both kNN and RF outperform the standard mean imputation (MI)
- The Random forest outperform the kNN

Gain in Predictive ability when the imputed marker are used
- The missing data imputation improve the prediction accuracy
- The RF is the most effective method (high computation time)
- Imputation is more useful at lower depths
- It is still relevant even at high depths

Gain in Predictive ability when the imputed marker are used
- The missing data imputation improve the prediction accuracy
- The RF is the most effective method (high computation time)
- Imputation is more useful at lower depths
- It is still relevant even at high depths
Cross validation - corrected diagonal and imputed datasets

Gain in Predictive ability when both imputed markers and a correction for the coverage depth are used.

- It is possible to combine diagonal correction and imputations
- The results are generally higher than the ones obtained by only using imputation or diagonal correction
- Low sequencing depth can be used and good PA can still be achieved
- Imputation and biases correction can greatly improve the accuracy, especially at low SD

Real datasets

Material:
- A total of 3289 samples considered
- Different plant materials:
  - Training set: 946 bi-parental F2 family pools (ploidy=4n)
  - Validation set: 846 F2 bi-parental family pools (ploidy=4n)
  - SP: 1225 single plants (ploidy=2n)
  - SYNs: 127 multi-parental synthetic family pools (ploidy=16n)

Gain in Predictive ability when both imputed markers and the corrected diagonal are used.

Cross validation

Predictive abilities have been computed using standard gBLUP model. Leave Set Out (LSO): leave out one set of plants and using all the other to predict it

It is possible to predict one set from the others

Differences on the predictions depend mostly on:
- Its relationship with the other sets
- The plant material of the sets

Cross Validation

Predictive abilities have been computed using standard gBLUP model. Leave Set Out (LSO): leave out one set of plants and using all the other to predict it

It is possible to predict one set from the others

Differences on the predictions depend mostly on:
- Its relationship with the other sets
- The plant material of the sets

Real datasets

Material:
- A total of 3289 samples considered
- Different plant materials
- Genotyped with ~1.5M SNPs in three different assays with different multiplexing parameters

Cross Validation

Gain in Predictive ability when both imputed markers and the corrected diagonal are used

The diagonal bias correction and the imputation of missing values increased the accuracy of the genomic predictions also across sets
Conclusions

- GBS is an efficient way to genotype different ryegrass breeding materials.
- Low coverage depth genotyping can still yield good PA for the analyzed traits.
- The biases at low sequencing depth can be corrected.
- The missing value can be imputed.
- It is possible to combine data from different genotyping assay and of different ploidy.
- It is possible to predict across different breeding materials.
- New perspective for reshaping the breeding schemes.

Acknowledgments