Empirical prediction accuracy of genomic selection between experimental designs and generations in oil palm

D. Cros

david.cros@cirad.fr

1/ Selection candidates
2/ Progeny tests in A x B hybrid trials to estimate GCAs
3/ Selection of individuals with highest GCAs within each parental group
4/ Selfings and intra-group crossings to produce next generation (used to start following breeding cycle and to produce commercial hybrid seeds)

2 heterotic groups, A (mostly Asia) and B (Africa)
→ hybrid vigor for bunch production in A x B crosses
→ conventional oil palm breeding for yield = reciprocal recurrent selection (RRS):

Conventional RRS

Annual response to selection

\[
\text{Annual response to selection} = \frac{i \times r_{a,a} \times \sigma_a}{L}
\]

Towards a new breeding scheme?

Genomic RRS

1/ Training set: phenotyped & genotyped ("dense" coverage of the genome)
2/ Analyze considering all markers jointly, e.g. using a matrix of realized relationship \(G\) computed from markers ("GBLUP")
3/ Prediction of the breeding value of the selection candidates from their genotype → Genomic selection

Previous simulations and empirical studies showed the high potential of GS to improve quantitative traits in oil palm and pointed out the need to obtain:

– empirical GS accuracy between experimental designs?
– empirical GS accuracy in the generation(s) after training?

Also, the potential of genotyping-by-sequencing has not been investigated for GS in oil palm yet
Available populations = 2 experiments of progeny-tests in North Sumatra:

Experiment 1:
146 A x 156 B = TRAINING SETS

Experiment 2:
68 A x 43 B = VALIDATION SETS

Measure of GS prediction accuracy:
Validation sets =
- For group A:
  - ~60 individuals of the same generation as training individuals (eg sibs, ...)
- For group B:
  - ~20 individuals of the same generation as training individuals
  - ~20 progenies of some training individuals

Example: partial view of group B pedigree

Control method = PBLUP, ie predicting value of validation individuals with pedigree instead of SNPs, to check if GBLUP capture between family differences AND within family differences

Molecular data = SNP obtained by Genotyping-By-Sequencing
- production of raw sequence data:
- pipeline for sequence data analysis / SNP calling:
  - polymorphic SNPs / parental group:
    - group A: 3844
    - group B: 6441
  (minimum depth per datapoint = 5, max missing data / SNP = 75%)
- imputation of missing data: Beagle 4.0, using pedigrees or not

Measure of GS prediction accuracy:
1/ Estimation of « reference » GCAs for validation sets with the conventional method, ie using the phenotype of their hybrids
  → can GS predict these reference GCAs using only the genotypes?
2/ Genotyping of training and validation sets
3/ Applying GS model with training sets data (hybrids phenotypes + genotypes of their parents) and the genotype of validation parents
  → GEBVs for validation sets
4/ Prediction accuracy = correlation (GEBV, reference GCAs) in validation sets

Traits: oil yield components (measured on hybrid individuals):
- Average bunch weight (ABW), Bunch number (BN)
- Fruits to bunch ratio (FB), Pulp to fruits ratio (PF), Oil to pulp ratio (OP)
- Total bunch production (FFB), Extraction rate (OER)

SNP density / quality:
Different SNP density were tested, using:
- random SNP sampling
- SNP sampling optimizing quality parameters (% missing data, MAF, distribution along genome)

Example of SNP filtering to reach an even distribution along the genome:
**SNP density / quality:**

Different SNP density were tested, using:
- random SNP sampling
- SNP sampling optimizing quality parameters (% missing data, MAF, distribution along genome)

![Graphs showing SNP density and quality](image)

**Results**

**Prediction accuracy per trait using all SNPs:**

- **Group A**
  - GBLUP better than PBLUP for BN and FFB, but low accuracy

![Graph showing prediction accuracy](image)

**Using pedigrees to impute missing SNP data increases GS accuracy:**

Example in group B:

- Variability in accuracy can be high for a given SNP density, depending on the actual random sample of SNPs
- With random SNP sampling, using all SNP gives best accuracies

![Graphs showing imputation results](image)

**Subsetting SNPs on quality can increase GS accuracy:**

- GBLUP > PBLUP, but using 300-400 SNP with low % missing data and high MAF strongly increases prediction accuracy

![Graphs showing subsetting results](image)
Subsetting SNPs on quality can increase GS accuracy:

→ using 300-400 SNP with high MAF strongly increases prediction accuracy and makes GBLUP > PBLUP
→ best method for SNP filtering differs among traits

Prediction accuracy per trait using all SNPs:

→ GBLUP better than PBLUP for several traits
→ high PBLUP accuracy in some traits makes difficult the evaluation of GBLUP

Subsetting SNPs on quality can increase GS accuracy:

→ using 400 SNP with low % missing data, high MAF (selected on an index) and even distribution along genome strongly increases prediction accuracy
→ best method for SNP filtering differ among parental groups
→ using an index to filter SNPs on several criteria can increase accuracy

Conclusions

→ Other studies are required to take advantage of the full potential of GS:
  - What would be the effect of larger training sets on accuracy (by genotyping hybrid individuals in addition to their parents, and / or combining data from several generations of progeny tested individuals)?
  - Can we define a more robust method to subset SNPs ?
  - A measure of GS accuracy in large full sib families is necessary: using existing breeding material, PBLUP sometimes gives high accuracy → difficult to tell the ability of GS to select within families
  - SNP array versus GBS SNPs ?

GBS is a relevant method to produce the molecular data necessary for GS in oil palm

Best method to subset SNPs differ between traits and populations (and generation)
Thanks for your attention!