

# Genotyping Validation of GCP reference sets



Sarah McGrath, Jean-Francois Rami,  
Jean-Christophe Glaszmann

CIRAD, avenue Agropolis, 34398 Montpellier Cedex 5, FRANCE



## What is the purpose of the validation project?

Under the SP1 umbrella, global genetic characterisation of 21 species was undertaken. An aim of this genetic characterisation was to choose a sub-sample (reference set) of the overall collection based on several criteria of representativeness and minimised structure.

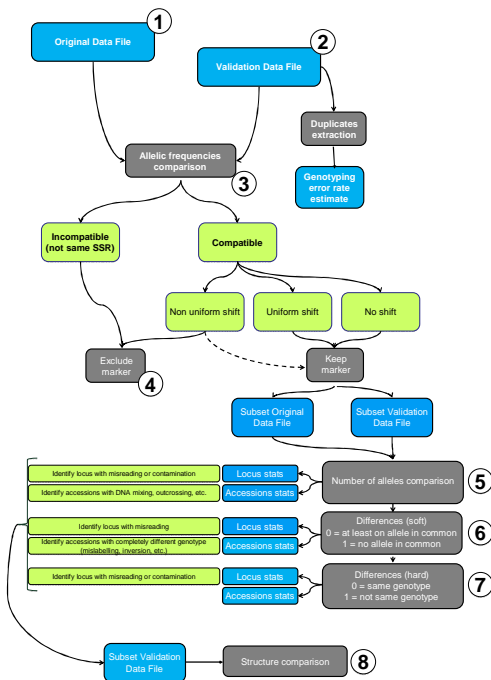
The data available on these reference sets (genotypic, phenotypic, passport data) will be made available as a public good. As such, the genotypic information on these reference sets needs to be validated.

The validation consists of re-genotyping the reference set of samples with a subset of top quality and most discriminant markers (about 20) by a single non-consortium lab (service provider).

## Status of the 21 species being validated

Species	Original Genotyping	Reference set chosen	Size of Ref. set	Nb of markers	DNA delivered	Genotyping completed
Barley	Y	Y	300	15	Y	Y
Chickpea	Y	Y	300	20	Y	Y
Coconut	Y	Y	359	20	Y	Y
Finger Millet	Y	Y	300	20	Y	Y
Groundnut	Y	Y	300	20	Y	Y
Maize	Y	Y	234	20	Y	Y
Pigeon Pea	Y	Y	300	20	Y	Y
Sorghum	Y	Y	345	20	Y	Y
Wheat	Y	Y	372	20	Y	Y
Common Bean	Y	Y	192	20	Y	N
Cowpea	Y	Y	345	20	Y	N
Lentil	Y	Y	150	10	Y	N
Musa	Y	Y	96	20	Y	N
Yam	Y	Y	342	20	Y	N
Cassava	Y	Y	250	20	N	-
Foxtail Millet	Y	Y	200	20	N	-
Pearl Millet	Y	Y	300	20	N	-
Rice	Y	N	-	20	N	-
Sweet Potato	Y	N	-	20	N	-
Potato	Y	N	-	20	N	-
Fababean	Ongoing	N	-	20	N	-

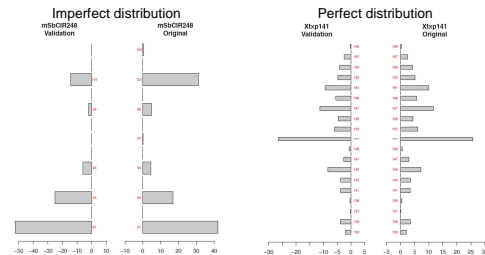
## Analysis protocol



## Results: the sorghum case

### 3 Allele frequencies comparison

All 20 markers were compatible based on frequency distributions. All markers showed uniform/quasi-uniform shift. 3 markers showed some distortion, however, the remaining 17 markers had perfect distributions.



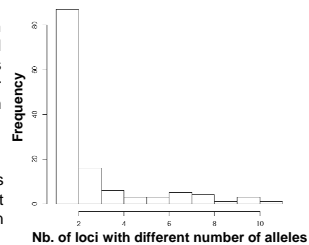
### 5 Number of alleles comparison

#### Accessions per locus

17/20 markers: less than 10% of accessions had different numbers of alleles between datasets. (The other 3 loci were excluded from further analysis).

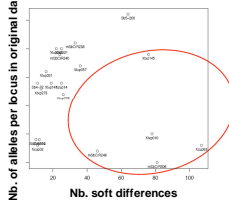
#### Loci per accession

283/300 accessions had less than 5 loci with a different number of alleles between datasets.

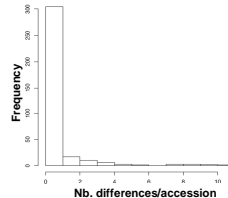


### 6 7 Number of differences between datasets

Nb. of alleles per locus in original dataset



Distribution of nb of differences/accession

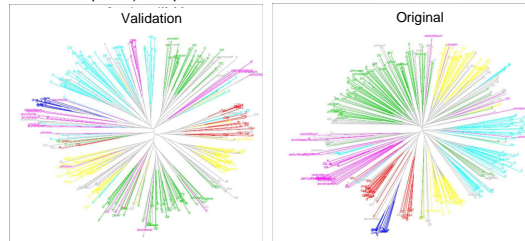


5 loci were identified as outliers (circled in red above) based on "soft" criteria

13 accessions showed 5 differences or more between both loci.

### 8 Structure

A tree for each dataset was constructed using the weighted neighbour joining method based on simple matching distances. Structure as determined with original data using the "Structure" software (K=6, membership=0.9) is represented on both datasets.



## Conclusions and outlooks

In sorghum, the validation datasets need additional data curation after these steps in order to identify the sources of differences based on comparison with the original dataset.

Validation is almost complete for 9 of the 21 species, and is expected to start for 5 further species in the coming months.

This process of validation should facilitate the identification of misinterpreted SSRs, of contaminated or mislabelled samples, validate the structure present in the reference set, and ultimately make the data publicly available.