WAGENINGEN UR
For quality of life

Comparing the GCP SSR data sets regarding the ability to describe population structure: some methodological aspects

(part of GCP SP4 Project G4006.17: GCP Quality Management and Data Quality Improvement, PI: Thomas Metz)

Theo van Hintum, Centre for Genetic resources, The Netherlands (CGN), WUR

To evaluate the quality of the GCP SSR fingerprints of the composite collections, the data sets available from the GCP Central Repository were analyzed.

Data downloaded

On July 9th, 2008 all datasets with one or more of the keywords 'composite', 'SSR' or 'fingerprinting' were examined and if relevant downloaded from the GCP Central Registry. The cassava (CIAT), maize (CIMMYT) and the major wheat file (CIMMYT) were restricted in their access, but 17 files with SSR data could be downloaded.

The downloaded files were checked for a number of aspects such as : 'are all objects and markers occurring in the data list decoded?', 'are the scores within a plausible range?' and 'is the number of alleles scored per object-marker lower or equal to the ploidy?' Errors were, where possible, corrected or deleted.

The barley, coconut, cowpea, fingermillet, groundnut, lentil, pigeonpea, potato, sorghum and wheat datasets came 'in one piece'. The data sets of chickpea, musa and rice had to be constructed from data sets with different markers from different institutions. The phaseolus data set was constructed from two files with different material (Andean and Meso-American).

Analysis

The data were analysed using the following descriptors:

- Population structure (PopStr): determined on the basis of all pair wise Nei-Li distances between all objects in the dataset; the standard deviation of these distances is compared to a fitted binomial distribution and the ratio is used as an indicator of structure: the higher the ratio, the more structure in the population.
- Hetrozygosity (Hetr): average number of alleles per marker in one accession (excluding missing values).
- Randomness of the distribution of missing values over markers and over accessions (RMV_{mar} and RMV_{acc}): determined by dividing the standard deviation of frequencies of missing values over markers or accessions with that of the fitted binomial distribution.
- Data Resolution (DR): the ability of the dataset to distinguish the structure of the population.
- M_{50}: the number of markers needed to achieve a DR of 0.5.

Since it appeared that sometimes a faulty assignment of bands to alleles (binning) caused low DR values, a distance measure was introduced based on the probability that two allele-scores were identical, calculated as $1- P_{mis}{}^{d}$ where P_{mis} is a parameter indicating the chance that two bands differing one nucleotide are equal and d is the size difference of the most similar alleles in a comparison.

- Level of mis-binning (P_{mis}): value resulting in the highest DR_{Pmis}.

Conclusions

- The conformity of the files to the GCP SSR template varies strongly and needs to be improved to allow interpretation.
- Many errors in the files can be identified by visual inspection or simple analysis.
- The enormous range observed in both DR and M_{50} cannot be explained by lack of population structure; low DR combined with high M_{50} values are an indication of poor data quality.
- The fact that allowing for binning-errors improved the DR values indicates that this is a major weak spot in the genotyping protocols using SSRs in the GCP.

Acknowledgement

Only datasets that could be downloaded could be analyzed. I thank those who made their data available (and wonder about the rest…)

crop	data producer(s)	#acc	#mar	%miss	PopStr	Hetr	RMV_{mar}	RMV_{acc}	DR	M_{50}	P_{mis}	DR_{Pmis}
barley	ICARDA	2676	14	4%	1.12	1.07	2.50	1.15	.13	97.8	0.8	.20
chickpea	ICARDA - ICRISAT	3024	50	5%	1.50	1.02	3.77	2.40	.39	77.1	0.9	.49
coconut	CIRAD	1014	30	4%	1.57	1.45	5.27	1.10	.64	17.0	0.7	.67
cowpea	IITA	1871	16	4%	1.11	1.03	4.71	1.59	.20	63.3	0.7	.23
fingermillet	ICRISAT	1000	20	7%	1.24	1.09	2.94	0.91	.26	56.1	0.7	.29
groundnut	ICRISAT	911	21	5%	1.64	1.54	1.89	1.47	.61	13.3	0.8	.67
lentil	ICARDA	1000	24	22%	1.67	1.20	11.69	1.65	.15	137.1	0.7	.20
musa	IITA - CIRAD	327	48	20%	2.06	1.75	5.59	1.89	.70	20.6	0.6	.72
phaseolus	CIAT	625	36	9%	2.34	1.18	2.26	1.78	.73	13.2	0.5	.74
pigeonpea	ICRISAT	1000	20	6%	1.92	1.16	2.47	1.37	.73	7.3	0.9	.85
potato	CIP	944	50	2%	1.34	1.84	2.60	0.79	.70	21.7	0.4	.68
rice	IRRI - WARDA - CIRAD - CIAT- EMBRAPA	2757	50	10%	1.79	1.08	13.12	0.84	.52	46.8	0.7	.59
sorghum	ICRISAT - CIRAD	3393	39	4%	1.44	1.04	5.40	1.29	.41	55.0	0.7	.44
wheat	CIMMYT	464	13	13%	1.33	1.10	2.96	0.68	.29	32.4	0.8	.38

WUR – Centre for Genetic Resources, the Netherlands (CGN)
P.O. Box 16, 6700AA Wageningen
Tel: +31 317 – 480913
E-mail: theo.vanhintum@wur.nl