

Challenge Programme *GENERATION*
Subprogramme I
Genetic diversity of global genetic resources

Data analysis Workshop

held in CIHEAM,
Zaragoza, Spain
June 21-25, 2004



The data analysis workshop was organized by C de Vicente and JC Glaszmann. It took place at CIHEAM in Zaragoza, Spain. It gathered 31 participants (appendix 1). Most were scientists from the consortium, directly involved in germplasm identification and molecular analysis. One scientist from Ceraas (Senegal) in the course of a training at Agropolis joined the group, as well as two scientists from the University of Western Australia involved in chickpea research.

Initially planned to allow practise of various pieces of software to analyse a first batch of genotyping data, the workshop left room for wider discussions on Subprogramme 1, its organization and its evolution.

A first day was devoted to exposing the audience to a range of examples of association studies, selected to illustrate the type of samples that present particular interest for this approach. Examples were in maize (presented by invited speaker Letizia Kulandaivelu from Inra), sorghum, sugarcane and cocoa. The key notion there is that of linkage disequilibrium (LD), that is the association among alleles between distinct loci (see appendix 2).

The example on maize showed how anonymous molecular markers dispersed along the genome enable tracing the history of a crop, in particular the migrations and foundations that occurred with the spread of the crop in new areas. Conversely, the survey of variation in candidate genes may allow exploration of associations that have a functional basis, through the association between certain variants and particular phenotypic features. This serves to confirm the involvement of the gene in the elaboration of the trait.

The example on sugarcane showed how a population of cultivars derived from a limited genetic base through a limited number of reproduction cycles exhibits strong LD. This precludes the resolution of the genes involved when associations are observed between chromosome segments and traits, but it allows development of QTL mapping strategies on the basis of diversity studies with a reasonable number of markers scanning the genome.

The example of cocoa showed how a good understanding of the domestication of a crop enables identifying germplasm sections amenable to specific approaches. A cycle of bottleneck followed by new gene flow and recombination induces a particular population structure with strong LD at the interface between the components. This type of situation also enables mapping on the basis of global population analysis.

The example of sorghum showed how an annual inbreeder can develop a cM-scale LD during a diffuse domestication process. It also showed how a sampling of materials on the basis of anonymous markers dispersed along the genome can diminish genome-wide LD and retain essentially only those associations that are indicative of proximity in the genetic map. This helps eliminate spurious associations derived only from population structure.

Two half-day sessions were devoted to a review of the progress for each of the eleven crops represented. Each presentation involved some general information on the biology of the crop, its breeding history, its germplasm (amount and currently documented structure), and other elements worth knowing for proper germplasm management; a report of the progress on assembling datasets and a description of the germplasm contributors in the CP; a report of the progress on marker identification and a description of the genotyping network in the CP; a description of the current state of the data analysis (markers, phenotypes, cross comparisons), and; a brief analysis of constraints, difficulties and perspectives. A written progress report was provided. A number of groups requested revision of the workplan calendar to account for some difficulty of mixed origin: slow communication, difficulty of exchange of materials, new equipment awaited. The updated workplans are given in appendix 3.

Two half-day sessions were kept for participants to test various softwares, with the guidance of several advisers, including B Courtois who centralises these matters in SP4. The participants agreed to use a common file format for data exchange. The fields (mandatory ones in bold) are:

Laboratory/Institute,

Species,

Sample ID,

Germplasm ID,

Locus,

Name of internal standard (=name of the molecular weight standard for peak size estimation),

Dye,

Allele (size in bp),

Peak size,

Quality (scale from 1 to 100),

Peak height,

Volume (area under the curve),

Allele amount (2n, 3n, 4n,...., bulk).

The content of the file can be pasted in a web site and converted to various input file formats adapted to various software packages using a web tool box.

The suggested softwares are global packages such as SAS, Genstat, PowerMarker, etc., or specific softwares such as DarWin, Structure, Partition, Mstrat, etc. The input formats will include SAS files; Individual x allele matrices with various column types (1 column per ploidy level, alleles separated by a "/" or concatenated alleles); disjunctive tables (1 column per allele); fully disjunctive tables (2 columns per allele). The list is not limitative. Support is requested from SP4 for these conversions.

Two half-day discussion sessions highlighted several issues.

The composition of the initial composite set must leave room for wild materials, landraces and improved materials. Although there is no single rule for the diverse crops, a distribution of 5%W:75%L:20%I seemed appropriate to the participants.

The issue of heterogeneous accessions was considered: when this is possible (not too late) and efficient (suitable multiplication rate), it is advisable to extract DNA from a single plant per accession and to self it (inbreeders) and use the seeds as a foundation. For outbreeders not clonally propagated such as maize, a range of methods is possible; Cimmyt and Inra have acquired considerable experience in the handling of bulks to evaluate allele frequencies. However, there is doubt as to the capacity of accessions with within accession diversity to reveal functional associations (statistical power).

The application of phenotyping with the view to association studies requires extraction of manageable samples for field/growth chamber/greenhouse experiments. The mode of sampling to extract (from the composite set) a reference sample that would be preferentially used for association studies was considered to have to include:

- representatives of the main components of the diversity ("Representation" accessions), to cover the range of allelic diversity
- a large portion of accessions that would cover in a continuous manner the global range of genotypic diversity ("Star"-like sample), ideal for species-wide association studies
- those sectors of the diversity that seem derived from recombination between two distinct components ("Interface" sections), ideal for LD mapping
- some components with large continuous variation ("Compartments"), ideal for subspecific association studies.

The value of using as many as possible common accessions across experiments for subsequent integration of information (and cross-comparisons) is recognized. The users should be encouraged to use as much as possible of this reference sample but should also be allowed to include their preferred checks and to exclude those materials that are not adapted to the experimentation environment.

The combination of all these criteria requires development of simple, easy to use softwares for elaborating the set of materials for any new experiment.

A half day was devoted to various business, communication and consultation of the participants regarding priorities for commissioned research.

Lessons learnt and conclusions

SP1 year 1 workplan was largely determined by a share of genotyping commitments between partners that was organized during the last few months of 2003. Eleven crops were selected at the Wageningen meeting in August 2003, on the basis of the declared availability of markers. Depending on the forces available, numbers of accessions were put forward. Cluster 1 has thus had a constrained framework: identifying a composite set of a given size irrespective of the diversity of collections and institutional dynamics for the various crops; in some cases each partner promoted its own sample whereas in other cases a unique rationale could be applied. The other constraint was the necessity to quickly start genotyping, leaving little (not enough) time to implement the sampling strategy.

The workshop held at the Plant and Animal Genome meeting at San Diego in January allowed timely identification of markers and planning of marker development. The workshop in Zaragoza came late on the one hand, because it was the first opportunity for a collective discussion on the global rationale and on particular issues shared among crops. On the other hand, it came early because few data were yet available, the first efforts having been concentrated on preparing DNA samples; the final goal of identifying balanced reference samples in view of association studies was still far from concrete. Altogether it was very useful for starting a community in good phase within SP1.

A number of groups (per crop) revised their calendar and now plan completion of the work from 0 to 5 months later than expected. In some cases the scientists have been overoptimistic by considering that an equipment just needs to be present to work full throughput. But most cases of delay were due to a slow start in communication of information and in exchanges of materials. Within this type of partnership, the slowest partner determines global speed. In some cases the speed may probably be improved by more institutional commitment: some partners are very large and it takes time to have the CP recognized a highest priority; some issues related to material exchange rules may be out of the institutional control. The work in cluster 4 will hopefully promote fluidity within the community. The most immediate priority is to have some exchanges of information occur, that condition the feasibility of cluster 1 tasks: it is not acceptable to have materials selected for genotyping and left without any bit of information with the cluster 1 crop coordinator. Measures can be taken at the scientists' level, some institutional support may become necessary from the steering committee in case of failure.

The group has progressed in establishing the basis for future action in the coming months: data forming, data analysis, sampling a reference set of accessions to be promoted for phenotyping, training NARS scientists in data analysis applied to germplasm management.

Training actions will include the involvement of NARS scientists in the final elaboration and documentation of a marker kit for each crop, that will be the basis for future decentralized molecular characterization of additional germplasm, through visits (several months stays) in crop-coordinating laboratories.

Rendez-vous has been taken for a conclusive workshop when all genotyping data have been acquired, with the aim of completing analysis (one scientist per crop) and training NARS scientists (two per crop), by mid 2005.

The organisers and the participants express their gratitude to the CIHEAM for their very kind hospitality and efficient organisational support for this workshop.

Appendix 1.

List of participants

Name	Institution	Address	Email
Claire Billot	CIRAD	Avenue Agropolis, TA 40/03, 34398 Montpellier Cedex 5France	claire.billot@cirad.fr
Daniel Foncka	CERAAS	France	danielfoncka@yahoo.fr
Jean-Francois Rami	CIRAD	France	jean-francois.rami@cirad.fr
Brigitte Courtois	CIRAD	Avenue Agropolis - TA 40/03 34398Montpellier, France	brigitte.courtois@cirad.fr
Jean-Christophe Glaszmann	CIRAD	Avenue Agropolis, TA 40/03, 34398 Montpellier Cedex 5France	glaszmann@cirad.fr
Salvatore Ceccarelli	ICARDA	P.O Box 5466 Aleppo	ceccarrelli@cgiar.org
Maarten van Ginkel	CIMMYT	Apdo. Postal 6-641, 06600, Mexico D.F, Mexico	M.Van-Ginkel@cgiar.org
Ruaraidh Sackville Hamilton	IRRI	DAPO Box 7777, Metro Manila, Philippines	r.hamilton@cgiar.org
Suketoshi Taba	CIMMYT	Apdo. Postal 6-641, 06600, Mexico D.F, Mexico	s.taba@cgiar.org
Matthew Blair	CIAT	A.A. 6713Cali, Colombia	m.blair@cgiar.org
Isabelle Hippolyte	CIRAD	73 rue Jean-François Breton - TA 40/16, 34398 Montpellier Cedex 5 France	Isabelle.hippolyte@cirad.fr
Jens Berger	University of Western	Australia	Jens.berger@csiro.au
Fucheng Shan	University of Western	Australia	fshan@agric.uwa.edu.au
Violeta Bartolome	IRRI	DAPO Box 7777, Metro Manila, Philippines	irri@cgiar.org
Marcelino Pérez de la Vega	Universidad de León	Área de Genética, Fac. de CC. Biológicas y ambientales, Campus de Vegazana 24071 León, España	degmpv@unileon.es
Carmen de Vicente	IPGRI	A.A. 6713Cali, Colombia	c.devicente@cgiar.org
Pat Heslop - Harrison	University of Leicester	Leicester	pjh4@leicester.ac.uk
Marilyn Warburton	CIMMYT	Apdo. Postal 6-641, 06600, Mexico D.F, Mexico	m.warburton@cgiar.org
Maria José Peloso	EMBRAPA	Caixa Postal 179Rodovia Goiânia - Nova Veneza Km 1275.375.000 Santo Antonio de Goiás - Goiás - Brazil	mjpeloso@cnpaf.embrapa.br
Ismahane Elouafi	ICARDA	PO Box 5466 Aleppo, Syria	i.elouafi@cgiar.org
Pooran Gaur	ICRISAT	Patancheru 502 324, Andhra Pradesh, India	p.gaur@cgiar.org
H. Upadhyaya	ICRISAT	Patancheru 502 324, Andhra Pradesh, India	H.Upadhyaya@cgiar.org
Rolf Folkertsma	ICRISAT	ILRI Nairobi	r.folkertsma@cgiar.org

Kenneth McNally	IRRI	DAPO Box 7777, Metro Manila, Philippines	k.mcnally@cgiar.org
Marc Ghislain	CIP	Apdo. Postal 1558, Lima 12, Peru	m.ghislain@cgiar.org
Morag Ferguson	IITA	Nairobi	m.ferguson@cgiar.org
Mahalakshmi Visvanathan	IITA	PMB 5320, Ibadan, Nigeria	v.mahalakshmi@cgiar.org
Rodomiro Ortiz	IITA	Namulonge, PO Box 7878 Kampala, Uganda	R.Ortiz@cgiar.org

Appendix 2.

Analytical framework for the use of molecular markers for mining useful allelic diversity

Genetic diversity is the substrate for breeding. The creation of new crop varieties requires exploring the value of various materials generated by brassage among germplasm identified as potential contributor of traits of interest. Genetics has enabled researchers to consider the basis that undermines phenotypic variation and explore gene reassortments made possible by meiotic recombination. The advent of molecular markers, molecular maps of crop genomes and functional genomics enables more specific actions. Genomics-based germplasm science addresses a range of questions on the source of improvement: where does it lie in the germplasm available, where on the genome, at which particular loci, in which particular allele and at which particular bit of sequence? Germplasm scientists thus concentrate their attention on proposing global approaches that can best serve these issues.

The genetic diversity of crops can be evaluated with a number of methods, including the so-called molecular markers. These have various advantages that have made them more and more widely used since the first studies with isozymes; now a range of technologies enable revelation of DNA markers in virtually unlimited numbers. Some markers are anonymous in that they just reveal sequence variation in randomly selected loci in the genome. These can be surveyed in germplasm samples and can be located on genetic maps by segregation studies in controlled progenies. Others can be targeted, in that they reveal variation within genes whose activity may be involved in the elaboration of particular traits. These are candidate genes that result from functional studies such as mutants analysis, physiological analyses or gene expression surveys.

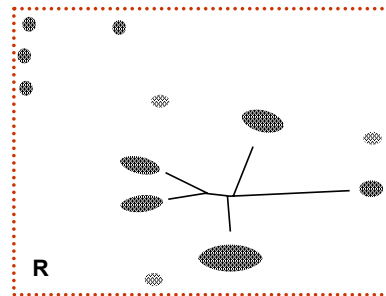
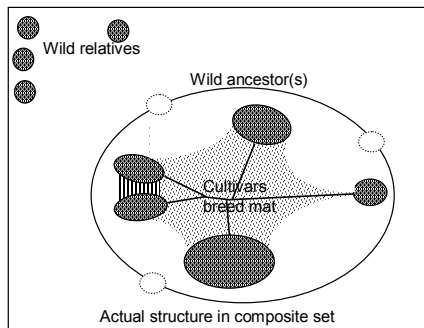
The relation between anonymous molecular diversity and variation for phenotypic traits at the whole germplasm level generally pertains to historical elements in the species evolution that fixed parallel, albeit functionally unrelated, differentiation for both types of traits. The origin of the former typically involves (insofar they are "neutral markers") phenomena such as mutation, drift, migration and all sorts of founder effects, whereas the forces that act on the latter include natural selection, to which human selection adds in the case of crops. Within this scheme, molecular markers are very useful in monitoring and sampling genetic diversity, including for phenotypic traits, for they identify and give access to all components of the germplasm structure. These components may have undergone diverse selection pressures and thus developed distinct sources of allelic variation for genes involved in those phenotypic traits. The property that is taken advantage of here is a range of statistical associations due to population heterogeneity, that involve markers/genes distributed over the whole genome and in most cases without any functional relationship in terms of physiological elaboration of the trait.

Another type of statistical association without functional relationship can be expected in certain conditions. When a particular mutation occurs, the new variant thus created will occur in a particular genotypic context and it will be associated with the other features of this genotype. After recombinations involving this genotype and its progeny have occurred, the association is maintained only within the vicinity on the chromosome around the new variant and it will decay when the genetic distance increases. This thus yields linkage disequilibrium limited to the region that bears the new variant. With the new high throughput molecular typing methods, it is possible to reach a genome coverage intensity that is compatible with LD-based localization of variants of interest. Further to the birth of the variant it is possible that LD emerge when the population meets new situations such as bottlenecks (the population restricts to a point where the variant is present in only a small number of genotypes) or admixture (the variant is present in one type of germplasm that is put in contact with another type highly differentiated). This yields LD along the whole genome. Again, if recombination occurs, the level of LD will diminish with genetic distance and the variant will be associated only with markers/genes closely linked to it. There also a fine genome coverage with markers enables localization of target variants on the basis of LD. Of course the ideal situation is when candidate genes exist and when one such gene is actually functionally involved in the trait. A statistical association between the presence of a particular allele and the variation of the target trait will depict this functional role. However, for the study to be conclusive, the contingent confusing effect of LD around the candidate gene, which might leave room to other interpretations than the responsibility of the very candidate gene, should be as low as possible, and the population under study should therefore be as little structured as possible.

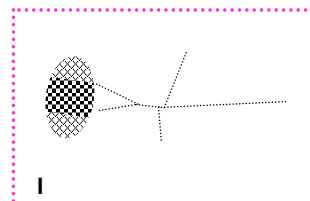
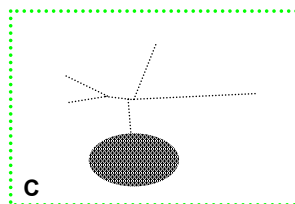
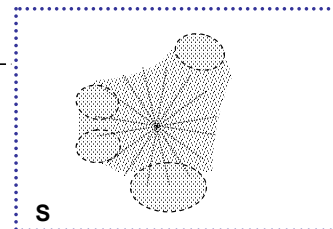
Appendix 4.

Representation of various rationales for sampling germplasm with the view to revealing associations between traits and molecular variation

Given a documented structure of the genetic resources, made of distant wild relatives, direct wild ancestors, various components within the cultivated pool, including traditional and improved materials, various approaches can be followed in order to build a reference sample that will be used to reveal associations between genotype and phenotype. They correspond to the use of two distinct sources of association: linkage or functional involvement (L or F). They will focus on subsamples with different geographic distributions (global vs local), different distributions of genotypic diversity (discontinuous vs continuous), and different ranges of adaptation that made them more or less appropriate for comparative phenotyping in a single evaluation environment.



Sampling rationale	Linkage Function	distribution	adaptation to a single environment for phenotyping	
Representation	F	global	discontinuous	???
Star	L, F	global	continuous	??
Interface	L	local	continuous	
Compartment	L, F	partial	continuous	?



How do we make use of marker data to build a reference sample?