

# *The Challenges of Molecular Genetics for Genebanks*

*"the IRG perspective"*

*Kenneth L. McNally and Ruaraidh Sackville Hamilton*

T.T. Chang Genetic Resources Center  
International Rice Research Institute  
Los Baños, Laguna, Philippines

# IRGC - the International Rice Genebank Collection

World's largest collection of rice germplasm held in trust for the world community and source countries

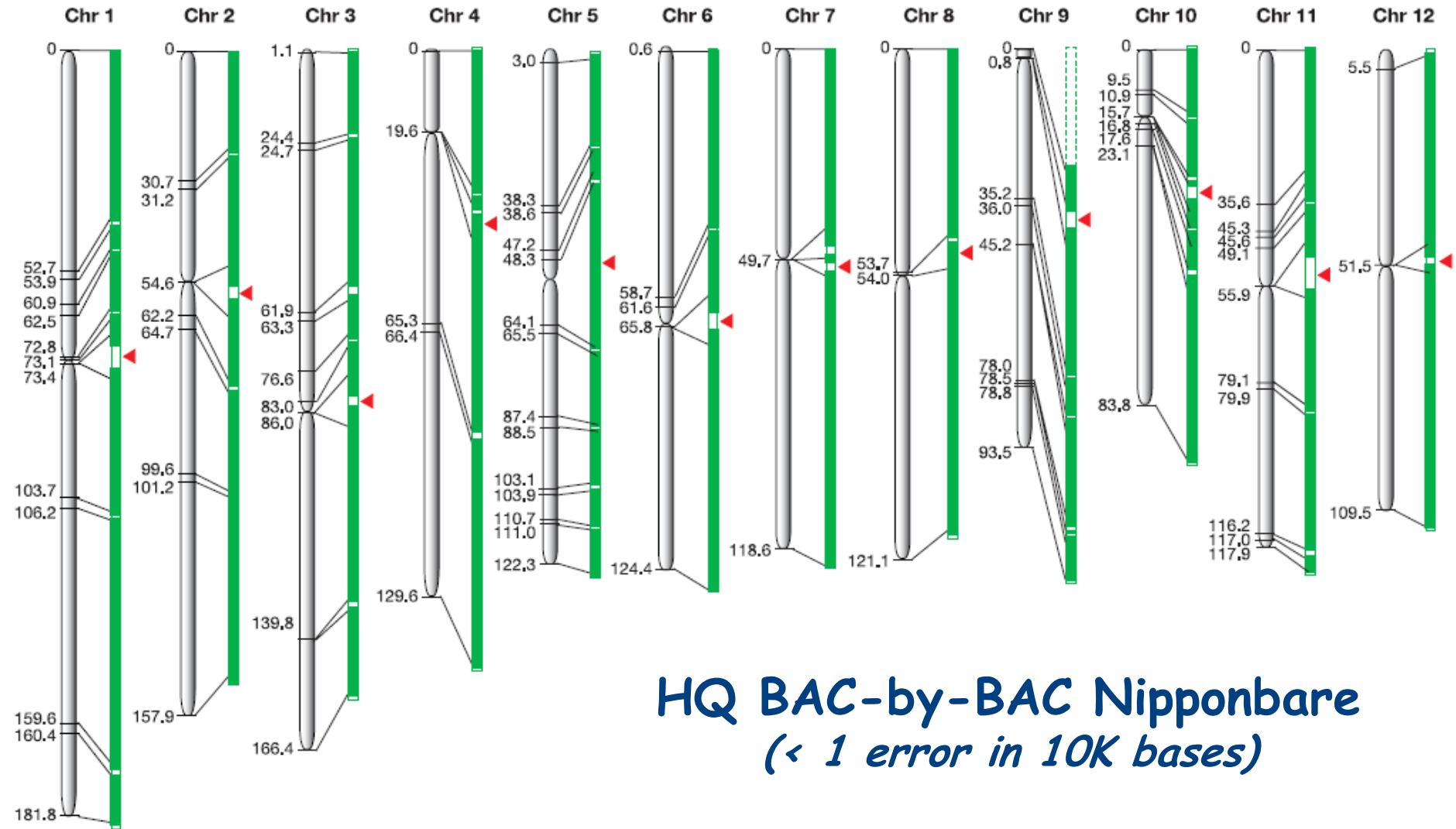


- Over 115,000 registered and incoming accessions from 117 countries
- Two cultivated species
  - Oryza sativa*
  - Oryza glaberrima*
- 22 wild species
- Relatively few accessions have donated alleles to current, high-yielding varieties
- <http://www.irri.org/GRC>

# Rice Diversity



*O. sativa* Panicles

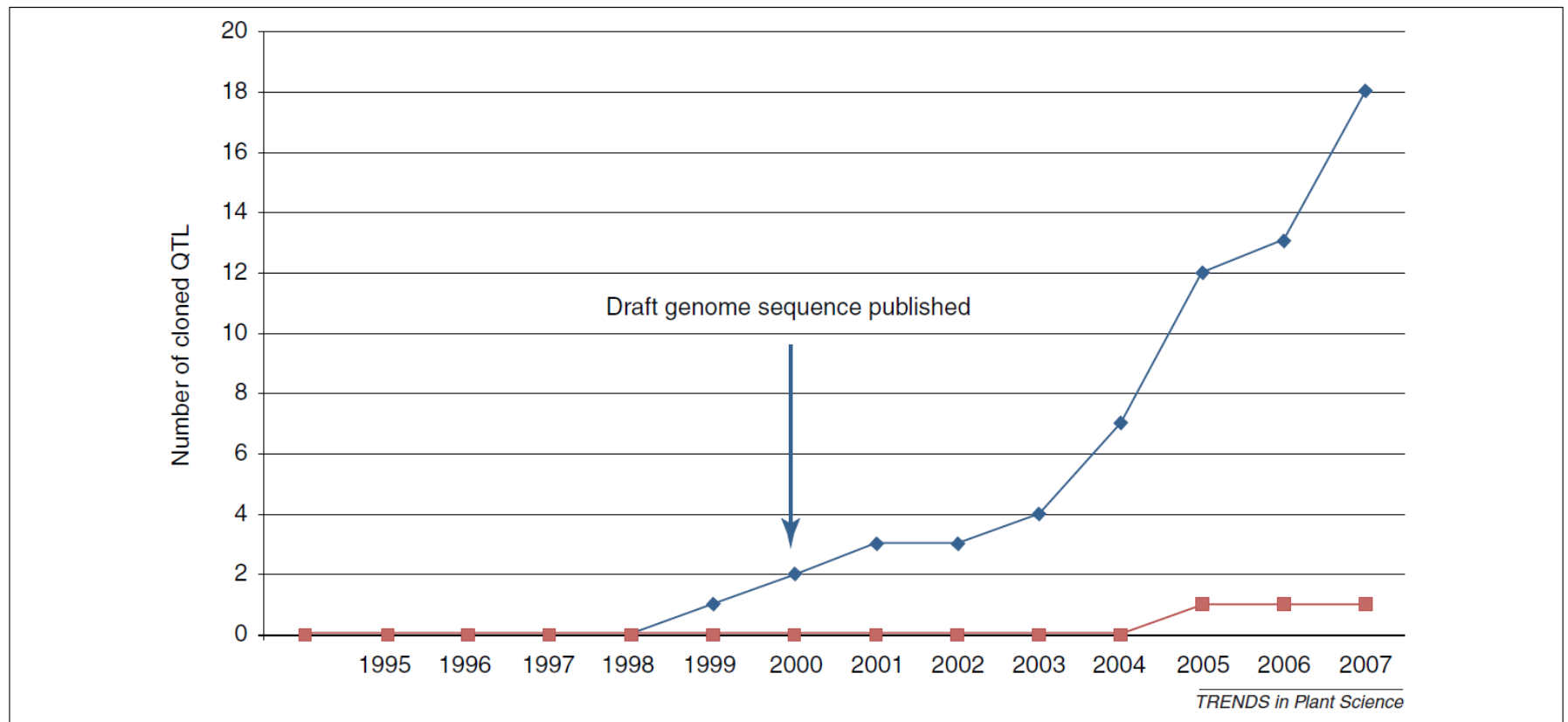


## HQ BAC-by-BAC Nipponbare (*< 1 error in 10K bases*)

**Figure 1 | Maps of the twelve rice chromosomes.** For each chromosome (Chr 1–12), the genetic map is shown on the left and the PAC/BAC contigs on the right. The position of markers flanking the PAC/BAC contigs (green) is indicated on the genetic map. Physical gaps are shown in white and the nucleolar organizer on chromosome 9 is represented with a dotted green line. Constrictions in the genetic maps and arrowheads to the right of

physical maps represent the chromosomal positions of centromeres for which rice CentO satellites are sequenced. The maps are scaled to genetic distances in centimorgans (cM) and the physical maps are depicted in relative physical lengths. Please refer to Table 2 for estimated lengths of the chromosomes.

# Rice genome sequencing already has impact but much more can be done in scale and scope



**Figure 2.** Number of QTL cloned in rice (blue) and wheat (red) since 1995. The blue arrow indicates the year in which the rice genome sequence became available and spurred the number of cloned genes and QTL (published source: NCBI and [42,61]). The Y axis represents the number of cloned QTL.

# Develop a genetic diversity platform

Single genome  
Nipponbare  
(Temp Japonica)

20 varieties  
genome-wide SNP  
OryzaSNP

2000+ lines  
genome-wide  
SNP  
Association  
genetics  
platform

NGS/3GS  
>10K lines  
from Gene  
Bank

2005

2008

2011

2012

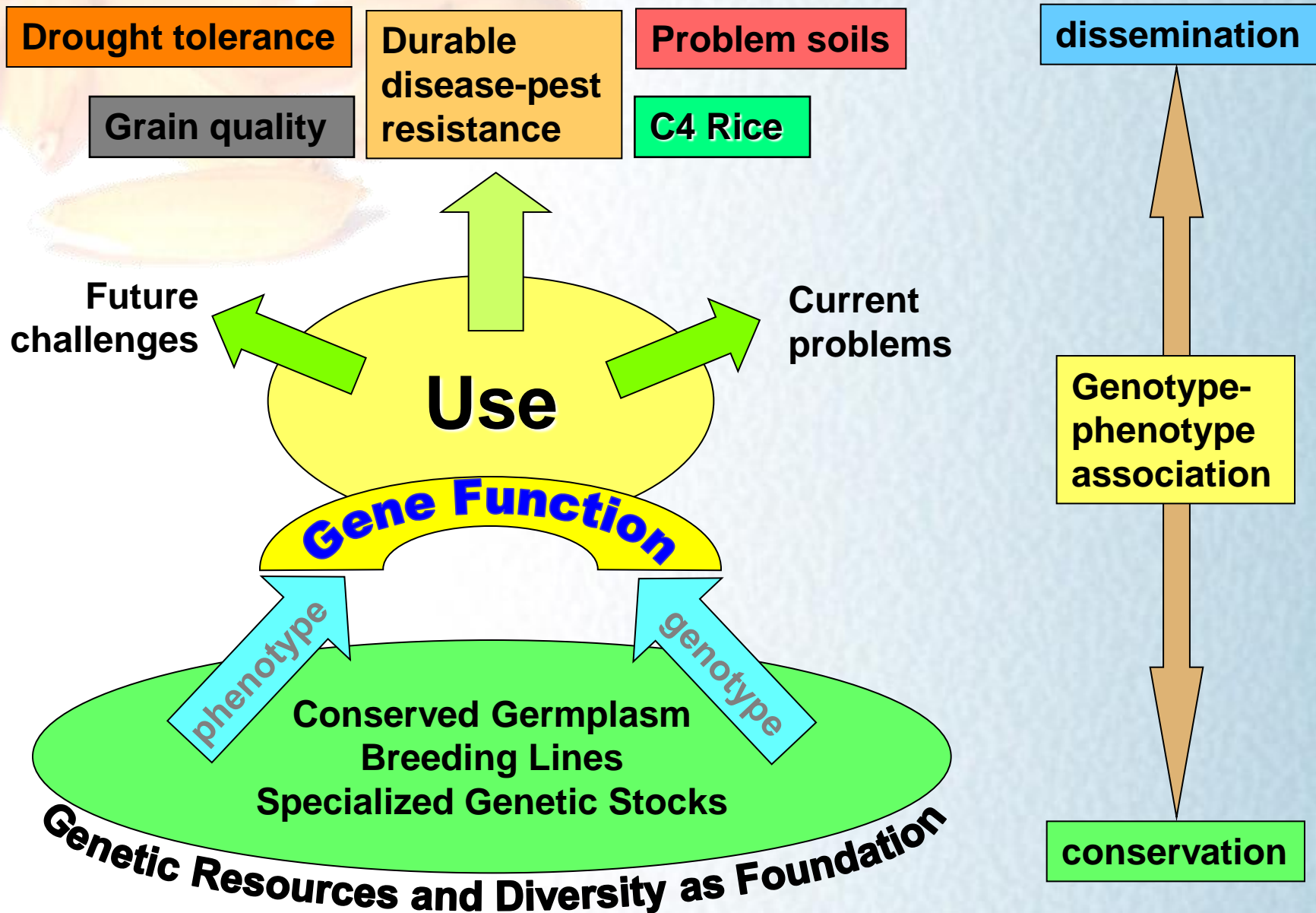
Benchmark  
physical map  
Enabling  
-omics  
technology  
“e-cloning”

Large SNP  
dataset to query  
new germplasm  
Breeding history  
Abundant markers

SNP haplotype-  
phenotype association  
QTL prediction  
Parental choices  
Pedigree/trait  
tracking

Natural reverse  
genetics system  
Probe deep into  
available, useful  
diversity  
Selective trait  
evaluation

# Public Genetic Diversity Research Platform



# Global Rice Scientific Partnership (GRiSP)

*a Research Program under the reformed CGIAR*

## *Themes and R&D Product Lines*

**IRRI**



AfricaRice



**cirad**



+++

**Embrapa**

1. **Harnessing genetic diversity to chart new productivity, quality, and health horizons**
2. **Accelerating the development, delivery, and adoption of improved rice varieties**
3. **Ecological and sustainable management of rice-based production systems**
4. **Extracting more value from rice harvests through improved quality, processing, market systems and new products**
5. **Technology evaluation, targeting and policy options for enhanced impact**
6. **Supporting the growth of the global rice sector**



# Product lines for theme 1: *Harnessing genetic diversity to chart new productivity, quality, and health horizons*

1.1. *Ex situ* conservation and dissemination of rice germplasm

1.2. Characterizing genetic diversity and creating novel gene pools

1.2.1 SNP Consortium for high density genotypes

1.2.2 Global phenotyping network for key traits

1.2.3 Whole genome sequencing of genebank stocks

1.2.4 *Specialized populations for genetic studies*

GWAS

1.3. Genes and allelic diversity conferring stress tolerance and enhanced nutrition

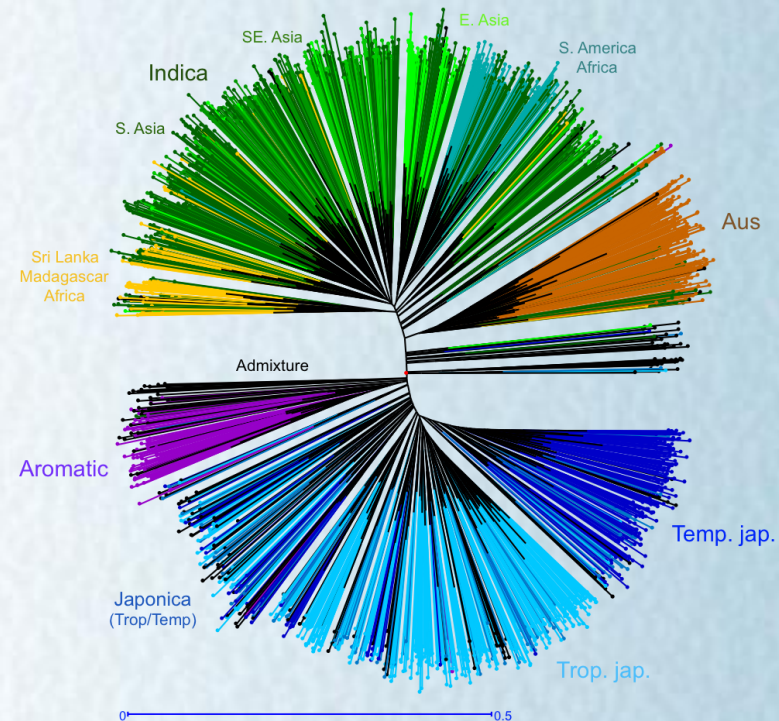
1.4. Converting rice from C3 to C4 architecture and metabolism

# Rice SNP Consortium for enabling genome-wide association studies (GWAS)

- Developed high-density genotyping Affy arrays with 1 M SNPs that include newly discovered SNPs from >150 genomes and from other projects
- Genotyping 2000 purified genetic stocks spanning range of diversity

<http://www.ricesnp.org>

- Partners include Cornell, USDA, CIAT, AfricaRice, CIRAD, Bayer CropSciences, Syngenta
- Genotyping ongoing with completion by June 2011
- *Precise phenotyping of traits in target environments for an integrated Rice Diversity Platform for GWAS*



# Germplasm for Genotyping/WGS

Diversity (coverage), utility, trait donors, nominations

- Cornell/USDA (500)
- GCP genotyping set (2339)
- GCP drought (800)
- GCP Aus (300)
- Orytage/Eurigen (600)
- **IRRI Core (13,000)**
- Madagascar (50)
- *O. rufipogon/nivara* (100)
- MAGIC parents (16)
- ACIAR chalk (1300)
- Various donors (100s)
- *O. glaberrima* (300)
- *Nominations from GRiSP (100s)*

*Now have >4100 SSD genetic stocks, preparing 8,000+ more  
(treated as new accessions, linked to original as derivatives)*

*2000 for SNPing on 1M feature Affy arrays*

*Rest are in line for sequencing by Illumina NGS with BGI*

# Multiplication of 1200 SSD lines





## PL 1.2.2 Global phenotyping network for key traits

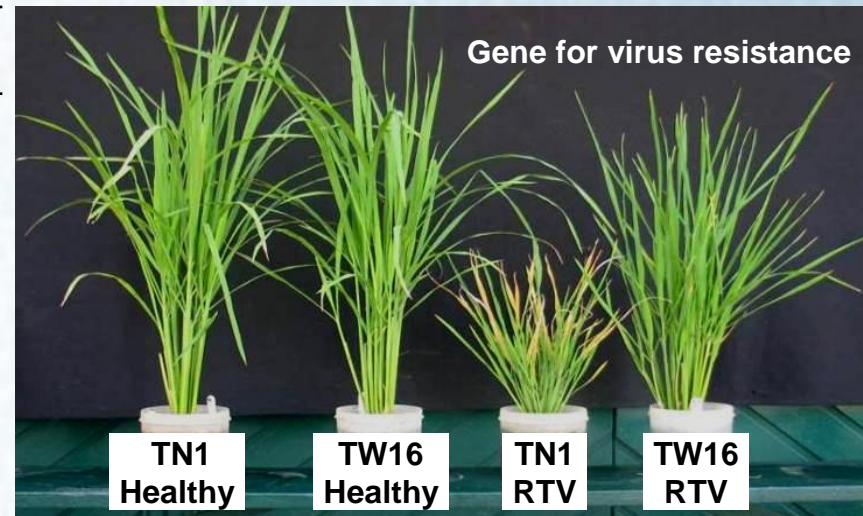
# Phenotyping consortium for traits with impact

- Build consortium of partners with expertise in particular traits
- Rely on existing networks and sites as much as possible
- Identify and prioritize traits where impact is needed
- Sample from the Rice SNP set of 2000 lines for subsets targeted to specific traits and environments
- Phenotype these traits using standardized procedures
- Capture meta-data about experimental design,
  - **Method ontology, Experimental ontology**
- Use **controlled vocabulary** and **trait ontology** for data
- Centralized database for data capture (IRIS)
  - Pedigrees, germplasm stocks, phenotype studies, SNP data
  - Updated schema (*Chiangzhi Liang, IRRI*)
  - Full use of **PO, EO, MO, TO, GO, Crop Ontology (GCP)**

# Phenotyping network: example traits for impact

*Focus on traits affected by global climate change*

Trait	Site
Yield components	Field
Disease resistance	GH + disease nursery
Salinity (vegetative and reproductive)	GH + Field
Drought	GH + Field
Heat (humid and dry)	Growth chamber + Field
Grain quality	Laboratory
Seed physiology	Laboratory





Aus lines 2010DS  
250 purified genetic stocks  
from GB accessions

3 environments

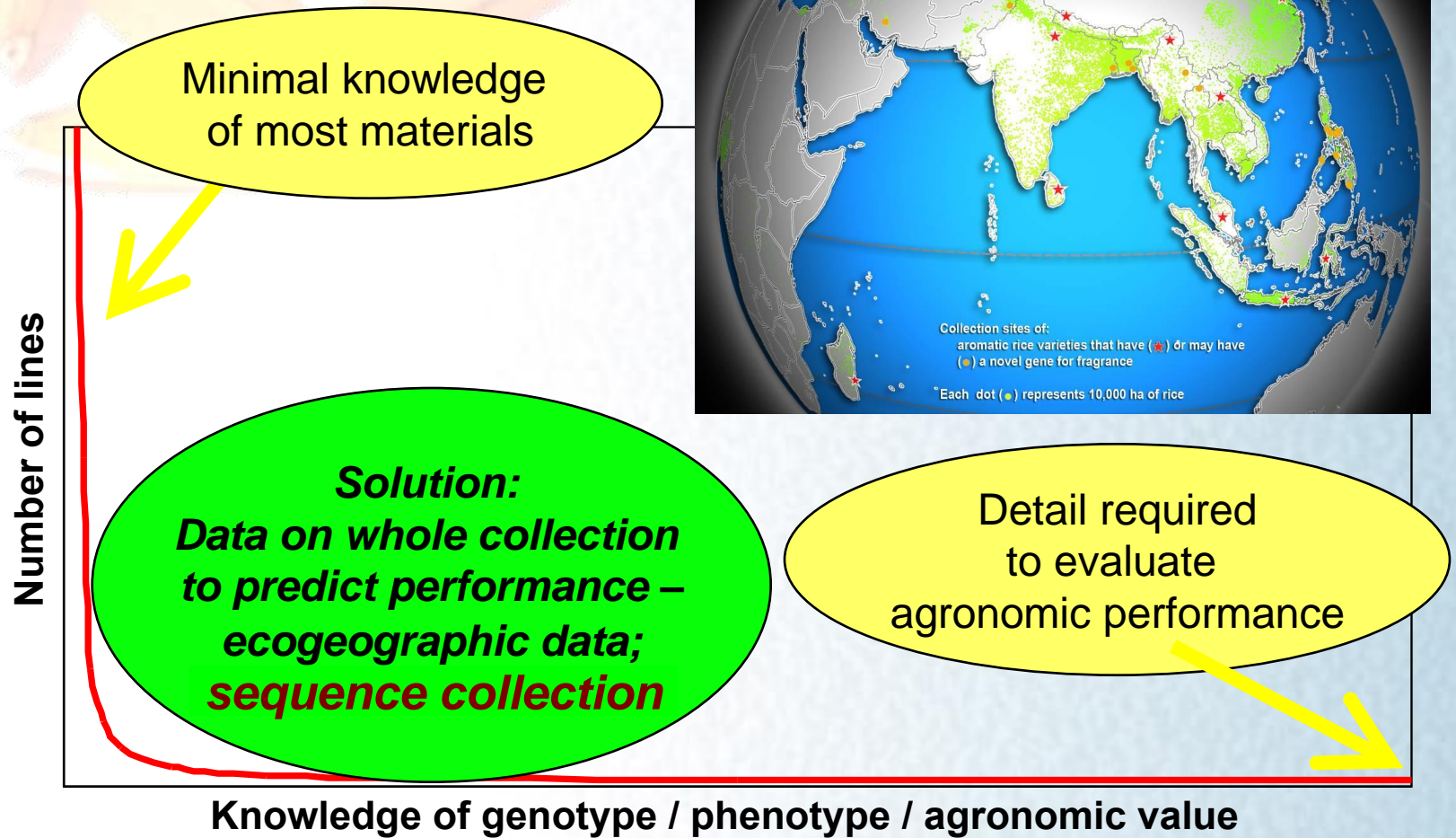
- Early vigor
- Canopy temp
- Yield



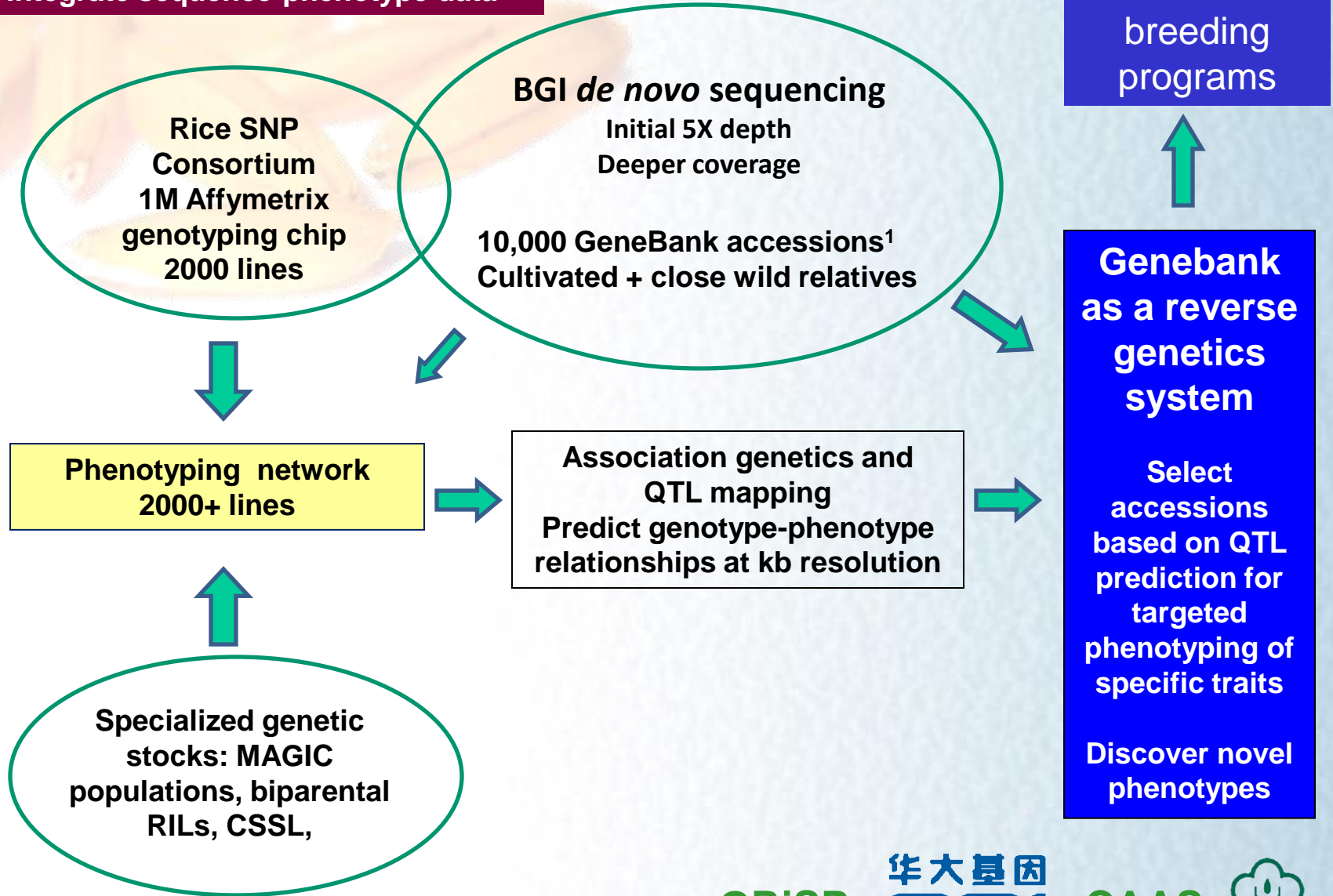


### **1.2.3** Whole genome sequencing of germplasm stocks

# Genetic Resources: Genotype/Phenotype information



**Bioinformatics and database to  
Integrate sequence-phenotype data**

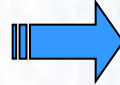


<sup>1</sup> Include publicly accessible germplasm from IRRI, CIRAD, AfricaRice, CIAT and regional collections

# Tapping into the unknown

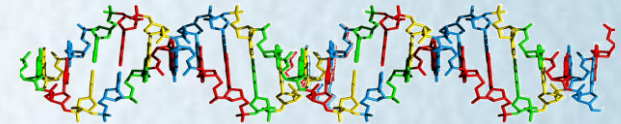


IRG Traditional Germplasm  
100,000 cultivated accessions



**Iterative  
sampling**

*Start @  
5X depth*



**Apply low-cost sequencing  
by next generation and 3<sup>rd</sup>  
generation technologies**

- Working with BGI-Shenzhen to sequence 10,000 genomes by mid 2012
- First 3,000 to be completed by September 2011
- Use the association data between 2000 lines and trait phenotypes to select materials for specific evaluation
- Isolate novel genes and rare alleles contributing to these traits

华大基因  
**BGI**

# 103 Genomes by Illumina NGS for 1M SNP Affy chip

*Cornell, IRRI, USDA, DevGen, Academia Sinica, EMBRAPA, Uni Aberdeen, JBEI/JGI, NIAS, Uni Delaware, Arizona Genome Institute, AfricaRice ...*

- 15 indica
- 6 indica/admixed (unique type in some analyses)
- 12 aus
- 17 temperate japonica
- 7 aromatic
- 16 tropical japonica
- 14 *O. rufipogon* and *nivara* (AA genome)
- 1 *O. meridionalis* (AA genome)
- 7 *O. glaberrima* (African cultivated, AA genome)
- 7 *O. barthii* (AA genome)
- 1 *O. punctata* (BB genome, outgroup)

## ***52 genomes from W Wang (Kunming Zoo Institute) & FY Hu (YAAS)***

- 5 indica, 4 aus, 2 deep-water, 6 aromatic, 5 trop. japonica, 4 temp. japonica,
- 25 *O. rufipogon* and *nivara*, 1 *O. longistaminata*

# What are the challenges these efforts present?

- Maintaining the link between the original accession and its purified genetic stock
- Having an efficient database system that allows the integration of the genebank information with phenotypic, breeding, genomic, and IPR data for enhanced utilization

# Genotype-phenotype relationships are about more than just accessions

- **Other key genetic resources**
  - Specific seed stock of an accession
  - Generations of purification
  - The plant sampled for DNA extraction
  - The leaf extract, DNA extract
  - Crosses made for gene discovery
  - Selections for NILs, RILs, etc
- **Attach data to the sample studied**
  - Not to the parental accession
- **Need to track all relationships**
  - To document potential drift, selection etc from accession to sequence
  - To ensure germplasm match between genetic and phenotypic data
  - To link data correctly back to accession

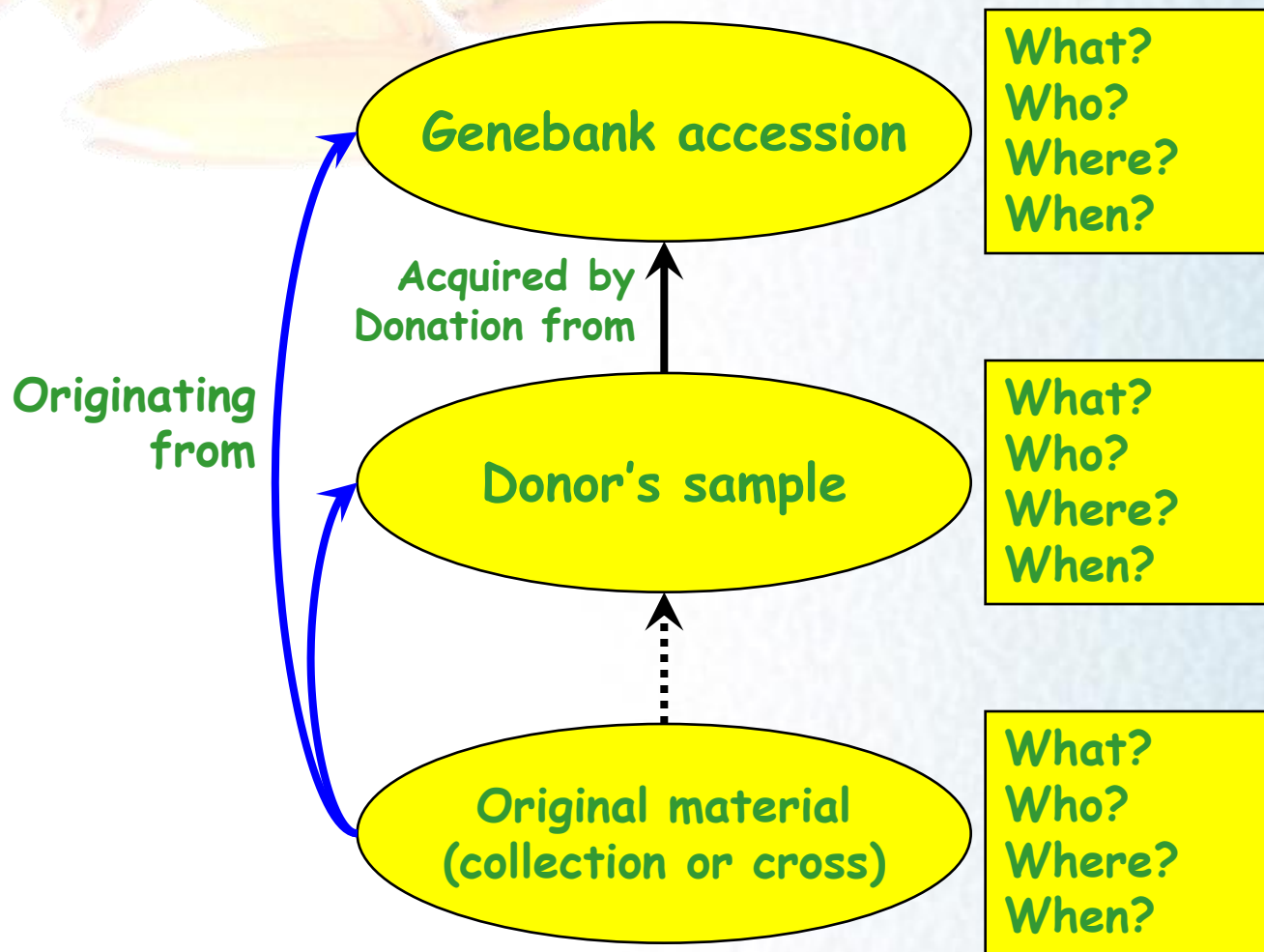
# Genebank curator's view: Multi-Crop Passport Descriptors

One record in MCPD documents 3 distinct germplasm samples

<b>Accession</b>	<b>Donor</b>	<b>Origin</b>
<b>What?</b> <b>Who?</b> <b>Where?</b> <b>When?</b> (How=acquired from donor)	<b>What?</b> <b>Who?</b> <b>Where?</b> (When excluded) (How=acquired directly or indirectly from origin)	<b>What?</b> <b>Who?</b> <b>Where?</b> <b>When?</b> <b>How?</b>



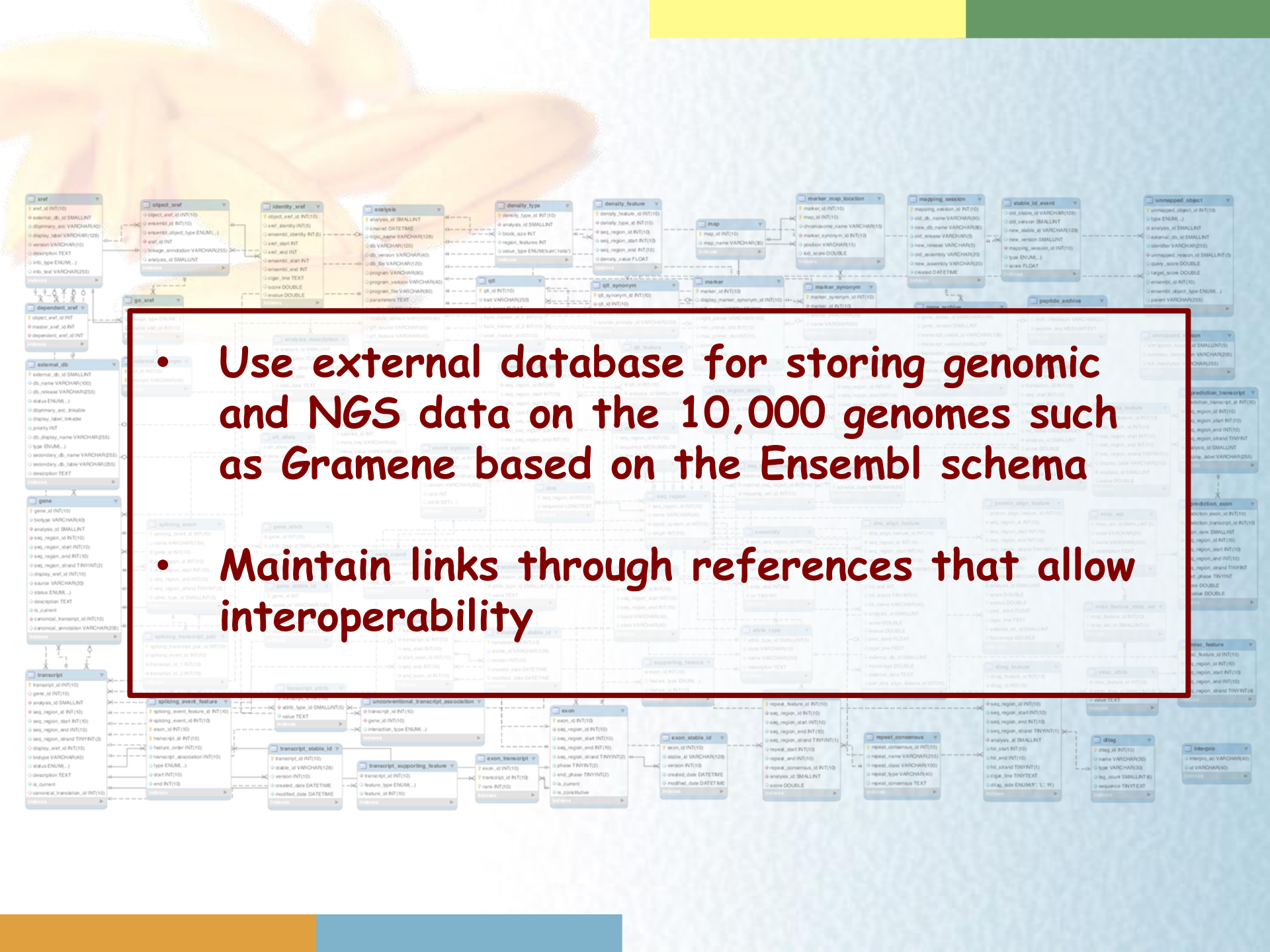
# Need to generalise beyond MCPD: 1 record per germplasm sample, of any kind



# How are we handling this?

- Migration of the Genebank from IRGCis (Oracle-based, antiquated) to the International Rice Information System (<http://iris.irri.org>)
- The genealogy management system schema allows independent tracking of individual samples (seed, leaf, DNA) under an accession “management” domain or as derivatives linked to the original accession

- 
- Extension of the data management system of IRIS so that SNP and other genotyping information can be stored.
  - A molecular data management system with binary large objects (BLOBs) for handling very large matrices
    - 2000 lines x 1M SNPs
    - Separately by row and by column
    - External references for genomic data on SNP loci

- 
- Use external database for storing genomic and NGS data on the 10,000 genomes such as Gramene based on the Ensembl schema
  - Maintain links through references that allow interoperability

# Conclusions

- Advances in genotyping and sequencing have immensely accelerated the amount of data being generated
- Curation is needed in the context of the genetic resources used for discovery
- Integration of the genotype/phenotype data in a system along with passports and pedigrees will add significant value
- All of which will lead to enhanced utilization of conserved germplasm
- The key issue is that the sample (not the "accession") is the main entity for curation.

