Characterizing genetic diversity and creating novel gene pools in rice for trait dissection and gene function discovery

Kenneth McNally, Hei Leung, Ramil Mauleon, Richard Bruskiewich & Ruairidh Sackville-Hamilton
International Rice Research Institute
IRRI Campus, Los Banos, Laguna
Philippines 4031
CGIAR Thematic Area 3:
Sustainable crop productivity increase for global food security

A Global Rice Science Partnership (GRiSP)

an evolving alliance of IRRI, AfricaRice & CIAT
with Cirad, IRD, JIRCAS and hundreds of research and development partners
GRiSP R&D Themes

1 Genetic Resources
2 New Varieties
3 Production Systems
4 New Products & Value Chains
5 Targeting & Policy
6 Regional Delivery

6 research themes

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

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

1.4. C4 Rice
Milestones: development of genetic diversity platform

- Single genome
- 20 varieties genome-wide SNP
- >100K lines from Gene Bank
- 2000+ lines genome-wide SNP
- Association genetics platform
- SNP haplotype-phenotype association
- QTL prediction
- Parental choices
- Pedigree/trait tracking
- Natural reverse genetics system
- Probe deep into available, useful diversity
- Selective trait evaluation
- Benchmark physical map
- Enabling -omics technology
- "e-cloning"
Public Genetic Diversity Research Platform

- Drought tolerance
- Grain quality
- Durable disease-pest resistance
- Problem soils
- C4 Rice

Current problems
- Future challenges
- Use
- Gene Function
- Conserved Germplasm Breeding Lines Specialized Genetic Stocks
- Genetic Resources and Diversity as Foundation

Phenotype-genotype association
- dissemination
- conservation

Establishing Gene-Trait Relationships

Sample 2500 diverse rice

Phenotype subsets for target traits with impact

Genotype SNP on 600K Affy arrays
http://www.ricesnp.org

Salinity tolerance

Associate SNP haplotypes with phenotypes

Predict “performance peaks” contributed by multiple SNP haplotypes
1.2: Characterizing genetic diversity and creating novel gene pools & associated informatics

- 1.2.1 Rice 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
- 1.2.1 – 1.2.4 Databases and bioinformatic tools supporting genetic diversity & gene discovery
PL 1.2.1 Rice SNP Consortium
Rice SNP Consortium for enabling Genome-wide association studies

- Developing high-density genotyping AffY arrays with >800k SNPs
- Includes newly discovered SNPs from 80+ genomes and other projects
- Genotype 2500 rice lines spanning range of diversity
  - 2000 – IRRI; 500 – NSF-AFRI
- [http://www.ricesnp.org](http://www.ricesnp.org)
- Partners include Cornell, USDA, AfricaRice, CIRAD, Bayer CropSciences, Syngenta, CIAT
- Other funding from Japan-Breeding Project
- Precise phenotyping of traits in target environments for an integrated Rice Diversity Platform
Germplasm for Deep Genotyping

Diversity (coverage), utility, trait donors, nominations

- NSF-TV (500)
- GCP genotyping set (2339)
- GCP drought (800)
- GCP Aus (300)
- Orytage/Eurigen (600)
- Other nominations of pure lines

- Madagascar (50)
- Rufipogon/nivara (100)
- MAGIC parents (16)
- ACIAR chalk (1300)
- Various donors (5)
- USDA core (1500)

Now have > 5000 SSD lines
Multiplication of SSD seed is under GRC management
• with QC by 384 plex SNPs
Resolve choices (remove duplicates) by using SNP fingerprints
Rice Genetic Diversity Platform

Home

Genetic diversity is the foundation for crop improvement. Understanding that diversity – how variation is distributed across the genome, how genetic architecture distinguishes local subpopulations, how sequence variation influences phenotype and crop performance – is essential for effective germplasm conservation and use. Available genomic and molecular tools offer unprecedented opportunities to explore and utilize the diversity of the rice gene pool.

We have recently received a grant from the USDA_AFRI Program (#2009-65300-05698) to develop a 600K SNP chip for rice and to genotype 500 diverse accessions of O. sativa and O. rufipogon.

www.ricesnp.org

- IRRI contributing to Rice SNP consortium webpage:
  - List of germplasm accessions being purified
  - 384-plex SNP chips specific for different varietal groups
  - Training workshops for SNP analysis
SNP panels for breeding and genetic diversity
Running 384-plex SNP sets at IRRI

Working with Susan McCouch (Cornell University) to develop custom 384-plex SNP sets (Illumina BeadXpress) for different subgroups:

- 384-plex for *indica* x *japonica* populations
- 384-plex for *indica* and *aus* germplasm

- 96 samples x 384 SNP markers per run
- Less than $0.10 per marker data point

- Diversity analysis
- DNA fingerprinting
- QTL mapping
- Marker-assisted breeding
- Testing genetic integrity of germplasm collection

2.3 million SNP data points genotyped at IRRI-GAMMA Lab in the past year
Future plans for SNP genotyping at IRRI

**Increase marker efficiency**
- Upgrade GAMMA Lab to increase marker capacity and reduce cost per genotype
- In-house functional SNP assays (fewer SNPs by many lines)
- Improve data quality by implementing barcoding and LIMS

**Develop functional SNPs**
- Validate & deploy SNPs for alleles needed for breeding programs

**Varietal assessment**
- Track variety usage by DNA fingerprinting of varieties in farmer’s fields for impact assessment
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 2500 lines for subsets targeted to specific traits and environments
- Phenotype these traits using standardized procedures
- Centralized database for data capture
Phenotyping network: example traits for impact

*Focus on traits affected by global climate change*

<table>
<thead>
<tr>
<th>Trait</th>
<th>Site</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yield components</td>
<td>Field</td>
</tr>
<tr>
<td>Disease resistance</td>
<td>GH + disease nursery</td>
</tr>
<tr>
<td>Salinity (vegetative and reproductive)</td>
<td>GH + Field</td>
</tr>
<tr>
<td>Drought</td>
<td>GH + Field</td>
</tr>
<tr>
<td>Heat (humid and dry)</td>
<td>Growth chamber + Field</td>
</tr>
<tr>
<td>Grain quality</td>
<td>Laboratory</td>
</tr>
<tr>
<td>Seed physiology</td>
<td>Laboratory</td>
</tr>
</tbody>
</table>

Gene for virus resistance

Root properties relevant drought tolerance
Phenotyping OryzaSNP set for traits.

Chalk, Grain quality, Disease (BLB, blast, SB, +), salinity, drought at reproductive stage, root traits, ... (IRRI & GCP collaborations)

Biomass, nutritional effects on immunostimulation in mice

(Jan Leach, Colorado State University)
Aus lines 2010DS
220 lines + increasing 36
3 environments
• Early vigor
• Canopy temp
• Yield
Linking root architecture with root function

IRRI lysimeter facility for high quality root phenotyping/physiology studies

Drought treatment

Dular
IR64
unplanted

Cumulative water loss (kg)

week after draining cylinders

0.5 1.5 2.5 3.5 4.5
1.2.3 Whole genome sequencing of germplasm stocks
Genetic Resources: State of knowledge on Genotype/Phenotype information

Minimal knowledge of most materials

Solution: Data on whole collection to predict performance - ecogeographic data; sequence collection

Detail required to evaluate agronomic performance

Knowledge of genotype / phenotype / agronomic value
Tapping into the unknown

- Obtain fine-scale DNA barcodes for each accession
- Use the association data between 2500 lines and trait phenotypes to select materials for specific evaluation
- Isolate novel genes and rare alleles contributing to these traits
89 Genomes sequenced by Illumina for chip design

Cornell, IRRI, USDA, DevGen, Academia Sinica, EMBRAPA, Uni Aberdeen, JBEI/JGI, NIAS, Uni Delaware

- 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)
- 1 *O. punctata* (BB genome)

- 52 genomes from Wang (Kunming Zoo Institute) & 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*
1.2.4 Specialized populations for genetic studies
MAGIC: Multi-parental Advanced Generation InterCross populations for Asia and Africa elite lines
Development of 20 RILs from OryzaSNP within and between variety groups for NAM

Thick lines ~2000
Thin lines ~400
IR64-21/Aswina F3 (1894 lines)
1.2.1 – 1.2.4: Databases and bioinformatic tools supporting genetic diversity & gene discovery
System for management & analysis of SNP & sequence datasets

Germlasm
- Natural populations
- Structured mapping populations

SNPs analysis
- Raw data output from BeadStudio or Alchemy (Wright et al 2010)

Sequence assembly/analysis
- Short sequence reads

1. Server-based data exploration, manipulation, & primary analysis tools

2. High-level analyses for genotype \(\rightarrow\) phenotype elucidation
1. Tools in place (under development)

1. Format converter

SNP data analysis & format conversion tool
Converts raw output from Illumina BeadStudio to other standard SNP formats

Or, run TASSEL using your uploaded data

Please REMOVE spaces in your data file name, it does not work, to be fixed later

Select data format to export to:
- Flapjack
- Pika
- Structure

Data file to upload: Browse

Action: Submit Run TASSEL with loaded data Reset

2. Run TASSEL with loaded data
2. Computing for the genotype $\rightarrow$ phenotype : case study

- OryzaSNP lines were shown to have introgressed genome segments of indica, japonica, or aus origin (McNally et al 2009)
- OryzaSNP lines differ in biomass-related trait (Jahn et al, in prep)
- Do the introgressed regions affect biomass-related traits?
Computing methodology

Biomass of 20 OryzaSNP lines

REGRESSION, CORRELATION analyses

Genome regions affecting biomass traits

Introgression genotype (McNally et al.)
Introgressed segments affecting biomass traits

Computing resources required for analyses of SNPs, re-sequencing, & G→P computations

Here now..

• Bioinformatics Servers @ IRRI (Super micros)
• Philippine e-Science Grid computing facilities
• Amazon EC2 High Performance Cluster

Upcoming

• NTI Taiwan computing collaboration
• DDBJ Read annotation pipeline
• ?? Future collaborations
Issues to address

• Sequence information storage
  o Efficiency (space)
  o Retrieval speed
  o Exploring GenomeData, Hierarchical Data Format - HDF5 format

• Standardization of analyses results for sharing
  o Community standards to be surveyed & adopted

• Online publication of analyses results
  o GRisP
    • Hosting
    • Open source solutions: GMOD, ??
  o RiceSNP consortium
  o Public repositories: Gramene, RAP, Genbank
Summary: G → P strategy

SNP discovery

OryzaSNP resequencing
20 rice varieties (2008-2009)

Illumina 1536-plex and BeadXpress 384-plex
(Cornell & IRRI, ongoing)

SNP genotyping

Next-gen Resequencing
>100 varieties (2010)

Affymetrix 950k
2,500 varieties (2011)

trait phenotyping

Mapping populations and diverse germplasm
(ongoing)

Sequencing the entire genebank
100,000 varieties (2012-2015)

QTL mapping, genetic diversity analysis, MAS, DNA fingerprinting

Phenotyping Network:
Large-scale, precise phenotyping and phenomics (2011-2015)

Trait association studies, allele mining for key genes and functional SNPs
Thank you

Acknowledgements to the GRiSP Theme 1 team at IRRI