



IRRI Genotyping Service Laboratory

Galaxy: bioinformatics for rice scientists

Ramil P. Mauleon

Scientist – Bioinformatics Specialist

TT Chang Genetic Resources Center

International Rice Research Institute

ICG-8, Shenzhen, China

Presented in behalf of my co-authors from IRRI

Lead Scientists

- Kenneth L. McNally – Genebank resequencing
- Nickolai Alexandrov – rice informatics consortium
- Michael Thomson – Genotyping Service Laboratory
- Hei Leung – Program Leader

Laboratory, software team

- Venice Margaret Juanillas
- Christine Jade Dilla-Ermita



Outline

- Introduction to IRRI & it's research agenda
- Bioinformatics support to molecular rice breeding at IRRI: IRRI GSL Galaxy
- Bioinformatics support to efforts for harnessing Rice Genetic Diversity
- Future activity: International Rice Informatics Consortium

INTERNATIONAL RICE RESEARCH INSTITUTE

Los Baños, Philippines

Mission:

Reduce poverty and hunger,

Improve the health of rice farmers and consumers,

Ensure environmental sustainability

All done through research, partnerships



Home of the Rice Green Revolution
Established 1960

www.irri.org

Aims to help rice farmers improve the yield and quality of their rice by developing..

- New rice varieties
- Rice crop management techniques

IRRI

A single strategic work plan for global rice research...

Global Rice Science Partnership : GRiSP

- o Core: 3 international research centers
- o Numerous research partners
- o **NEED TO SHARE RESEARCH SOLUTIONS**



AfricaRice

IRRI



CGIAR



Many more...

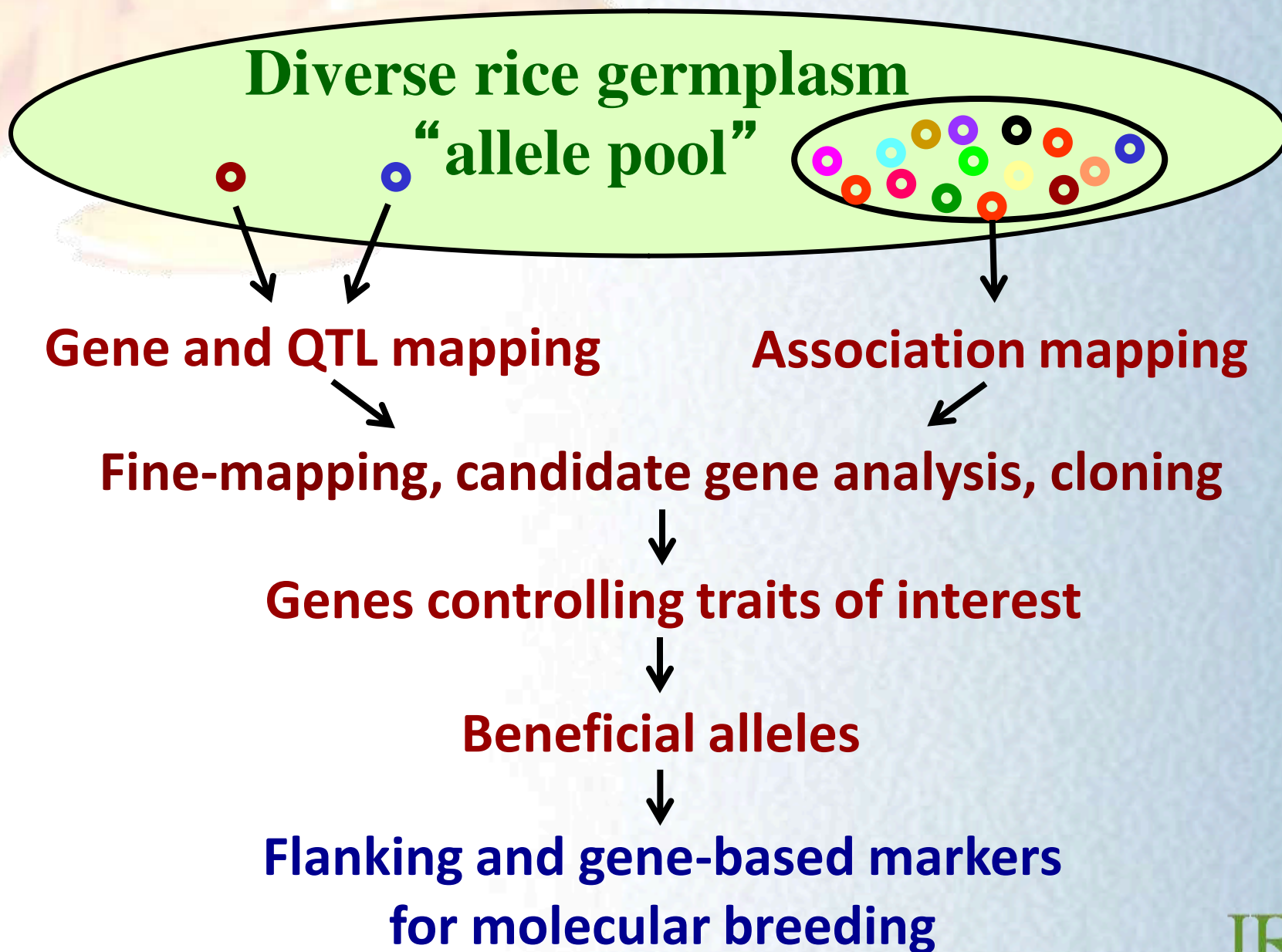
IRRI

First GRiSP Research Theme with heavy bioinformatics ...

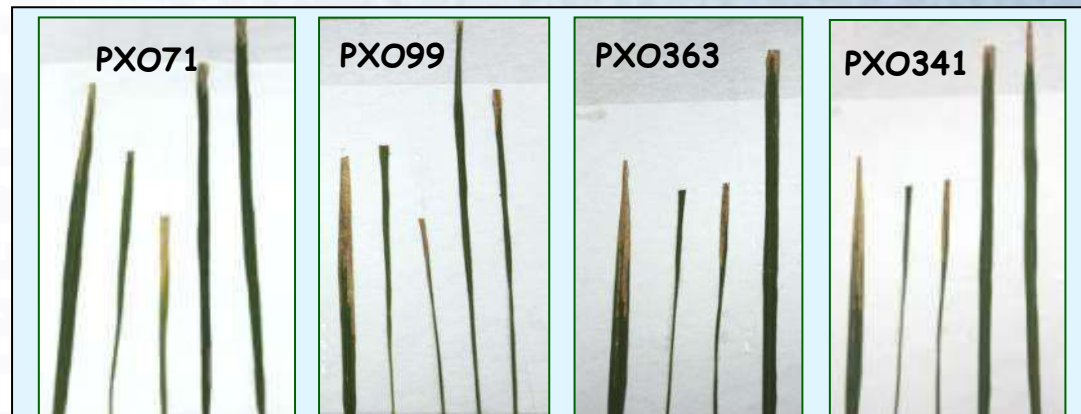
Accelerating the development, delivery, and adoption of **improved rice varieties**

- 2.1. Breeding informatics, **high-throughput marker applications**, and multi-environment testing

Allele mining for crop improvement



Major QTLs/genes for breeding



Xa genes for bacterial leaf blight

- QTLs and major genes for stress tolerance and disease resistance are known
- Flanking SSRs and gene-based STS markers have been used to transfer these major QTLs
- Move to SNP markers for Marker Assisted Backcrossing (**MABC**), Marker Assisted Selection (**MAS**), Genomic Selection (**GS**)

Challenges for IRRI scientists/breeders

- Not familiar with SNP-based genotyping
 - How do I score the alleles? (no gel image!!!)
 - Data does not fit my spreadsheet (run out of columns, rows)...
 - Cannot even view the data file using “ordinary” apps
 - Computer runs out of memory when I load the dataset...
 - Trusted analysis software crashes inexplicably...
- We need to
 - enable field/bench researchers for bioinformatics
 - Share solutions openly across GRiSP partners, with rice research community as a whole

Galaxy has features that fit our needs

Open, web-based platform for [accessible](#), [reproducible](#), and [transparent](#) computational biomedical research.

- **Accessible:** Users [w/o programming experience](#) can easily specify parameters and run tools and workflows.
- **Reproducible:** Galaxy captures info so that any user can [repeat](#) and [understand](#) a complete computational analysis.
- **Transparent:** Users [share and publish analyses](#) via the web and create interactive, web-based documents that describe a complete analysis.

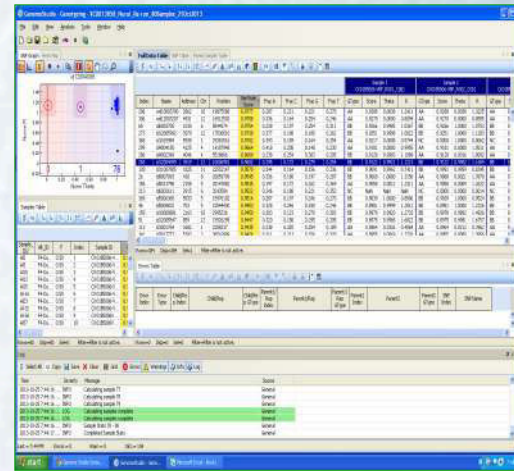
Integration of Galaxy to Genotyping Service Lab workflow

Illumina BeadXpress Genotyping



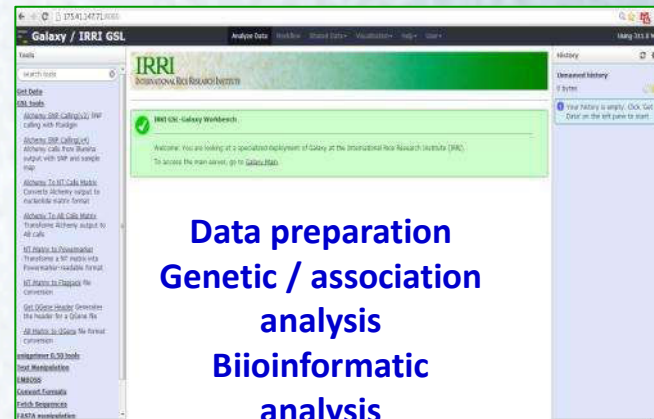
Infinium Custom 6k chip

GenomeStudio with Alchemy Plugin



Fluidigm EP1 Genotyping

Software tools on IRRI GSL-Galaxy



Data preparation
Genetic / association analysis
Biioinformatic analysis

Standard Galaxy release

Galaxy Analyze Data Workflow Shared Data Visualization Cloud Help User Using 0%

The cluster on which many NGS tools run will be down for maintenance from 4 PM, Monday, Nov. 19 until 9 AM the following day (EST5EDT, UTC-0400). Jobs running on that cluster

Tools

search tools

- Get Data
- Send Data
- ENCODE Tools
- Lift-Over
- Text Manipulation
- Convert Formats
- FASTA manipulation
- Filter and Sort
- Join, Subtract and Group
- Extract Features
- Fetch Sequences
- Fetch Alignments
- Get Genomic Scores
- Operate on Genomic Intervals
- Statistics
- Graph/Display Data
- Regional Variation
- Multiple regression
- Multivariate Analysis
- Evolution
- Motif Tools
- Multiple Alignments
- Metagenomic analyses
- Genome Diversity
- Phenotype Association
- EMBOSS

Running Your Own
Understanding how Galaxy works
An in-depth tutorial

Live Quickies

- 454 Mapping: Single End (Galactic quickie # 15)
- Uploading Data using FTP (Galactic quickie # 17)
- Managing account histories (Galactic quickie # 19)

History 1.2 MB

- Unnamed history
- 2: Filter FASTQ on data 1
- 1: human Illumina dataset

Galaxy is an open, web-based platform for data intensive biomedical research. Whether on this free public server or your own instance, you can perform, reproduce, and share complete analyses. The Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.

Galaxy build: \$Rev 8154:5dcbbdf1087\$

IRRI GALAXY (current)

Galaxy Analyze Data Workflow Shared Data Visualization Admin Help User Using 0%

Tools search tools

Get Data

GSI tools

- [Alchemy SNP Calling\(v2\)](#)
Alchemy calls from Illumina output with SNP and sample map
- [Alchemy SNP Calling\(v4\)](#)
Alchemy calls from Illumina output with SNP and sample map
- [Alchemy To NT Calls Matrix](#)
Converts Alchemy output to nucleotide matrix format
- [Alchemy To AB Calls Matrix](#)
Transforms Alchemy output to AB calls
- [NT Matrix to Powermarker](#)
Transforms a NT matrix into Powermarker-readable format
- [NT Matrix to Flapjack](#) file conversion
- [Get QGene Header](#)
Generates the header for a QGene file
- [AB Matrix to QGene file](#) format conversion

unigprimer 0.50 tools

EMBOSS

Text Manipulation

Convert Formats

Fetch Sequences

IRRI
INTERNATIONAL RICE RESEARCH INSTITUTE

IRRI GALAXY-BIOINFORMATICS WORKBENCH

Welcome! You are looking at a local instance of Galaxy at the International Rice Research Institute (IRRI).
To access the main server, go to [Galaxy main](#).

The Galaxy project is supported in part by [NSF](#), [NHGRI](#), and [the Huck Institutes of the Life Sciences](#).

History [refresh] [settings]

- Deployed in the cloud (Amazon Web Services Large instance in Asia-Pacific region)
- Streamlined to contain rice-specific tools and genotyping data
- NO NGS assembly tools included

Rice genome browser installed as data source for curated SNP, genome information

Galaxy Analyze Data Workflow Shared Data Visualization Admin Help User

Tools

search tools

Get Data

- Upload File from your computer
- GBrowse local MSU7 server
- Get Gene List on MSU7 from the specified position
- Get Gene Sequences on MSU7 from the specified position

GSL tools

unigprimer 0.50 tools

EMBOSS

Text Manipulation

Convert Formats

Fetch Sequences

FASTA manipulation

Workflows

- All workflows

Rice Genome Annotation Project - MSU Rice Genome Annotation(Osa1) Release 7: 30 kbp from Chr1:10,000..40,000

Browser Select Tracks Snapshots Community Tracks Custom Tracks Preferences

Search

Landmark or Region: Chr1:10,000..40,000 Search

Export to Galaxy Cancel

Save Snapshot Load Snapshot

Examples: Chr1:80,000..120,000, Chr4:22,000,000..22,500,000, LOC_Os01g01190, LOC_Os01g40150, aldehyde dehydrogenase.

Data Source: Rice Genome Annotation Project - MSU Rice Genome Annotation(Osa1) Release 7

Scroll/Zoom: Show 30 kbp Flip

Overview

Chr1

0M 10M 20M 30M 40M

Region

0k 10k 20k 30k 40k 50k 60k 70k 80k 90k 100k 110k 120k 130k 140k 150k 160k 170k 180k 190k 20k

GC Content

% gc

Details

10 kbp

8k 9k 10k 11k 12k 13k 14k 15k 16k 17k 18k 19k 20k 21k 22k 23k 24k 25k 26k 27k 28k 29k 30k 31k 32k 33k 34k 35k 36k 37k 38k 39k 40k

Gene

LOC_Os01g01030 protein, expressed

LOC_Os01g01050 R3H domain containing protein, expressed

LOC_Os01g01019 expressed protein

LOC_Os01g01040 expressed protein

LOC_Os01g01060 40S ribosomal protein S5, putative, expressed

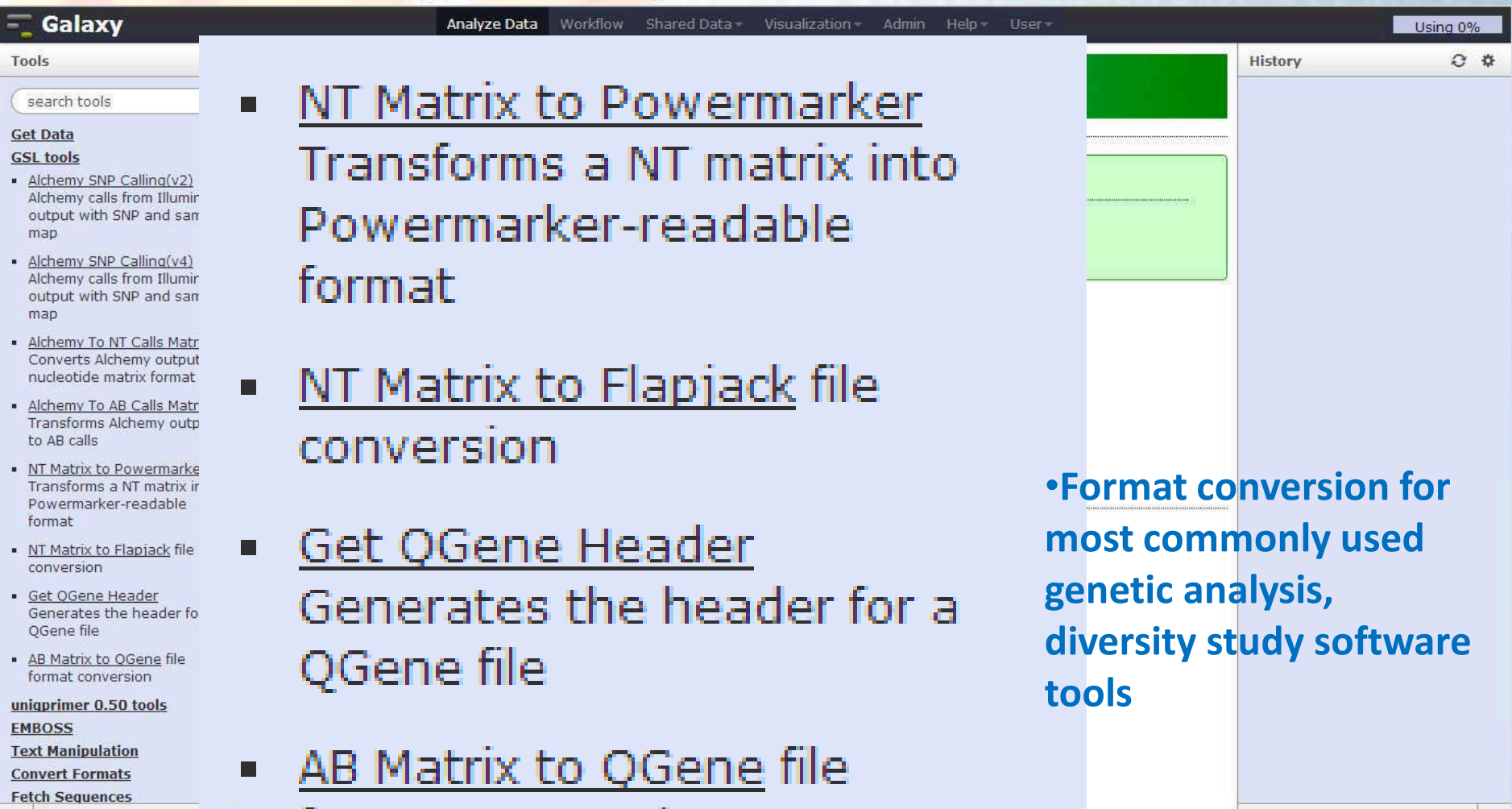
LOC_Os01g01070 expressed protein

LOC_Os01g01080 decarboxylase, putative

Select Tracks Clear highlighting

Comprehensive information on SNPs used in GSL

Data manipulation tools in GSL Galaxy



The screenshot shows the Galaxy web interface. The top navigation bar includes 'Galaxy', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Admin', 'Help', and 'User'. The left sidebar contains a 'Tools' section with a search bar and a list of tool categories: 'Get Data', 'GSL tools', 'uniqprimer 0.50 tools', 'EMBOSS', 'Text Manipulation', 'Convert Formats', and 'Fetch Sequences'. The 'GSL tools' category is expanded, showing a list of tools. A central workspace area is overlaid with a blue box containing a list of tool descriptions. The right sidebar shows a 'History' panel with a refresh icon and a gear icon. The top right corner indicates 'Using 0%'.

- NT Matrix to Powermarker
Transforms a NT matrix into Powermarker-readable format
- NT Matrix to Flapjack file conversion
- Get QGene Header
Generates the header for a QGene file
- AB Matrix to QGene file format conversion

• Format conversion for most commonly used genetic analysis, diversity study software tools

Workflows for rice data analysis already available

The screenshot displays the Galaxy web interface. The top navigation bar includes 'Galaxy', 'Analyze Data', 'Workflow' (circled in red), 'Shared Data', 'Visualization', 'Help', and 'User'. The main area is titled 'Workflow Canvas | Alchemy to powermarker'. On the left, a sidebar lists various tool categories like 'GDMS', 'SNP management tools', and 'Alchemy tools'. The central canvas shows a workflow with four steps: 'Alchemy-2' (Intensity file, SNP map file, Sample map file, out (tabular), log (tabular)), 'Alchemy2matrix' (Alchemy call file, out1 (tabular)), 'transposeTable' (Tab-delimited Table/Matrix, out1 (tabular)), and 'Matrix to Powermarker' (Matrix file, output (tabular)). The 'Matrix to Powermarker' step is expanded on the right, showing its configuration and a description: 'This tool converts a SNP matrix file to Powermarker file format.'

In place for Illumina BeadXpres, Infinium platforms, being tested on Fluidigm system...



Software Tools for SNP analysis

- SNP calling: [Alchemy](#) (Wright et al 2010)
- SNP data exploration, visualization: [Flapjack](#), [TASSEL](#)
- Genetic linkage mapping: [Mapmanager QTX](#), [R/QTL](#)
- QTL analysis: [R/QTL](#), [Qgene](#), [MPMap](#) (for multi-parent crosses)
- GWA analysis: [TASSEL](#)
- Population structure / diversity analysis : [Powermarker](#), [Structure](#)

Flapjack: visualize, manipulate SNP data

File Edit View Visualization Data Help

New Project Open Project Import Data

Data Sets

- diversity_rice32.Zhao_1536.flapj
- Trait Data
- Default View

Chromosome: 2 396 lines, 130 markers, length: 35,928,128

Line	Marker	Genotype
Aichi Asahi-GSOR301002	T C A A	G G G G C A A G A C A A G A G G T A C G T C
Ai-Chiao-Hong-GSOR301003	A C G G	A A G G A C A G G C A C G G C A A C C C T C T
NSF-TV 4-GSOR301004	T C G A G	A G A T A A A G A C G A G C C A G C A C T C T
NSF-TV 5-GSOR301005	T C G G G	A A A T A A A G G C C A A G C C A G C C C G C T
ARC 7229-GSOR301006	T C G A	A G A T A A A G A C G C G G C A G A C T C T
Arias-GSOR301007	T C G A G	A A A T A A A G A C A A G G C G G C A C G C C
Asse Y Pung-GSOR301008	T C G A G	A A A T A A A G A C C A A G C C G G C C G C C
Baber-GSOR301009	T C G G G	G G G T A A A G A C A A A G A G G C A C T T T
Baghlani Nangarhar-GSOR301010	T C A A G	G G G G G C A A G A C A A A G A G G T A C G T C
Baguamon 15-GSOR301382	A T G G G	A A G T A A G G C C A A G G C A G C A C T C T
Basmati-GSOR301011	T C G G G	A A A T A A A G G C C A A G C C G G T A C G C C
NSF-TV 13-GSOR301012	A T G G A	A A G G A A A G A C C A A G C C A G C C T T C T
Beonjo-GSOR301383	T C A A G	G G G T A A A G A C A A G C C G G T A C G T C
Bico Branco-GSOR301014	T C G G G	A A A T A A A G G C C A A G C C A G C C C G C T
Binulawan-GSOR301015	A C G G G	A G A T C A G G C C G G C A A C T T T
BJ 1-GSOR301016	A C G	A A G G C G G G C T A A G C C A G C C C T C T
Black Gora-GSOR301017	A T G G G	A A G G A C A G A C C A C G C C G C C T T C T
Blue Rose-GSOR301018	T C A A G	G A T A A A G A C A A G G C G G C A C G C C
Byakkoku Y 5006 Seln-GSOR301019	A C G G	A A G A C G G A T G A G C A A C C C T T
Caawa/Fortuna 6-103-15-GSOR301020	T C G A G	A A G T A A A G A C C A A G G C G G T A C T C T
Canella De Ferro-GSOR301021	T C G A G	A A A T A A A G A C C A A G G C G G C A C G C C
Carolina Gold-GSOR301022	T C G A G	A A A T G A G A C C A A G G C G G C C G C C
Carolina Gold-GSOR301023	T C G A G	A A A T A A A G A C C A A G G C G G C A C G C C
Carolina Gold Sel-GSOR301024	T C G A G	A A A T A A A G A C C A A G G C G G C A C G C C
NSF-TV 27-GSOR301025	T C G A G	A A A T A A A G A C C A A G G C G G C A C G C C
Champa Tong 54-GSOR301026	T T G G	A A G G A A A G A C C G A G C C A G C C T T C T

Line: _____
 Marker: _____
 Genotype: _____

Zoom:

- View genotypes graphically, with color code (nucleotide, compared to selected line ...)
- Select/deselect lines, markers from dataset,
- Filter lines by markers
- Basic statistics on dataset

4C, 9T, 140.75MB



<http://bioinf.scri.ac.uk/flapjack/index.shtml>

Flapjack Tip: Hold CTRL while clicking and dragging lines or markers to move them to new positions

What's new?

- [Version 3.25 released on 2/5/2006. This is a permanent version!! Please cite the Bioinformatics paper. Choose 'How To Cite' from Help menu to see more details.](#)
- [Version 3.09 released on 7/14/2004. Support 120 dpi settings.](#)
- [Version 3.08 released on 6/14/2004. Fixed a bug in core set selection.](#)
- [Version 3.07 released on 3/20/2004. \(1\) New SNP identification tool \(2\) Fix several minor bugs.](#)
- [Version 3.03 is available for download. \(1\) Fix a bug in batch export. \(2\) New functionality for single haplotype phase assignment. \(3\) Allow for all-missing individuals for most analyses.](#)
- [PowerMarker V3.0 was officially released on January 30th, 2004. Full documentation is included.](#)
- [The following functions are disabled in PowerMarker: \(1\) Marker selection for haplotype data, genotype data and trio data \(2\) Population structure inference based on EM algorithm \(3\) Logistic regression and least angle regression for association study. The author is stilling working on the publications. The algorithms have been implemented and tested with simulated data. Please \[contact\]\(#\) the author if you are interested in these algorithms.](#)

What's PowerMarker?

PowerMarker is a comprehensive set of statistical methods for genetic marker data analysis, designed especially for SSR/SNP data analysis. PowerMarker builds a [powerful user interface](#) around both new and traditional statistical methods for population genetic analysis. See [analysis](#) to check out the versatility of PowerMarker. PowerMarker is also a 2D Viewer - which was used intensively for visualizing linkage disequilibria results. [Download](#) PowerMarker now and speed up your data analysis!

<http://statgen.ncsu.edu/powermarker/>

TASSEL

-Trait Analysis by aSSociation, Evolution and Linkage

TASSEL (Buckler Laboratory, Cornell University) : a software package to evaluate traits associations, evolutionary patterns, and linkage disequilibrium.

Three areas of strength:

1. Integrates with various diversity databases (Panzea, Gramene, Sorghum , and GRIN)

2. Provides new and powerful statistical approaches to association mapping eg. General Linear Model (GLM) and Mixed Linear Model (MLM).

3. handles a wide range of indels (insertion & deletions) which is the most common type of polymorphism in maize.

www.maizegenetics.net/tassel

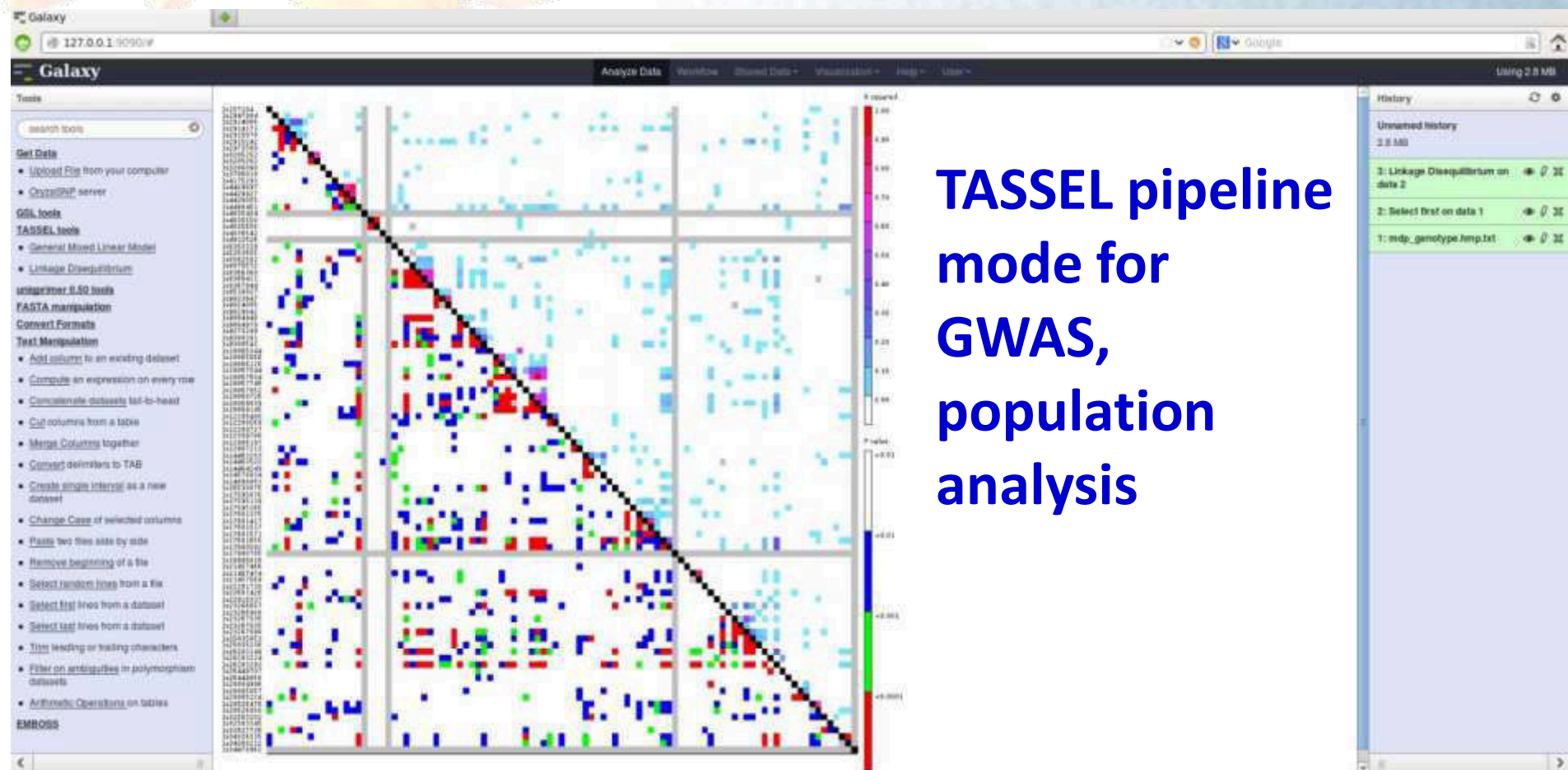
Tassel stand-alone has lots of analysis tools...

The screenshot displays the TASSEL (Trait Analysis by aSSociation, Evolution, and Linkage) 3.0.82 software interface. The main window features a menu bar (File, Tools, Help, GPC) and a toolbar with buttons for Data, Analysis, Results, Delete, and Wizard. Below the toolbar, there are buttons for GPC, Load, Export, Sites, Taxa, Traits, Impute SNPs, and Transform. The left sidebar shows a tree view of data files, including Sequence (chr1), Polymorphisms, Numerical, Matrix, Tree, Fusions, Synonymizer, and Result. The main workspace shows a list of physical positions (422620, 5773719, 11124818, 16475917, 21827016) and a list of SNPs (e.g., 6: 14430109, 7: 14448881, 8: 14626329, 9: 14818014, 0: 15149897, 1: 16484021, 2: 17795432, 3: 17937357, 4: 18045959, 5: 18260207, 6: 18372416, 7: 18461390, 8: 18617319, 9: 19245044, 0: 19354113, 1: 19384773, 2: 19688376, 3: 20585984, 4: 21020525, 5: 21794153). Below the SNP list, a list of numerical values is shown (4920, 4921, 4922, 4923, 4924, 4925, 4926, 4927, 4928, 4929, 5268).

Two analysis windows are open:

- Linkage Disequilibrium**: A heatmap showing the correlation between SNPs. The x and y axes are labeled with SNP IDs (e.g., Working10, Working11, Working12, Working13, Working14, Working15, Working16, Working17, Working18). The color scale ranges from 0.00 (white) to 1.00 (black), with a legend for Upper P-values (0.01 to 0.001) and Lower P-values (0.01 to 0.001).
- Manhattan Plot**: A plot titled "P-Values by Chromosome for AC". The y-axis is labeled "-Log10(P-value)" and ranges from 0.0 to 9.5. The x-axis is labeled "Position" and ranges from 0 to 350,000,000. The plot shows a dense cloud of points colored by chromosome (1-12), with a significant peak at approximately 250,000,000 on chromosome 8.

TASSEL analysis tools are being incorporated into Galaxy ...



**TASSEL pipeline
mode for
GWAS,
population
analysis**

IRRI Galaxy Toolshed (“APPS STORE”) is under development

The screenshot displays the Galaxy Toolshed interface. At the top, the navigation bar includes 'Galaxy Tool Shed', 'Repositories', 'Help', and 'User'. A sidebar on the left contains navigation links such as 'Search for valid tools', 'Search for workflows', 'All Repositories', 'My Repositories and Tools', and 'Available Actions'. The main content area shows a repository page for 'file_conversion_tools'. It includes a search bar, a 'Clone this repository' section with a URL, and fields for Name, Synopsis, Revision, Owner, and Times downloaded. Below this is a table of valid tools with columns for name, description, version, and requirements.

6 valid tools on Nov 18, 2012

Search

- Search for valid tools
- Search for workflows

All Repositories

- Browse by category

My Repositories and Tools

- Repositories I own
- My writable repositories
- My invalid tools

Available Actions

- Create new repository

Repositories Help User

Repository Actions

file_conversion_tools

Clone this repository:
hg clone http://mauleon@localhost:8001/toolshed/repos/sample6/file_conversion_tools

Name:
file_conversion_tools

Synopsis:
matrix to powermarker X matrix to qgene

Revision:
[1:bf50914d2d07](#)

Owner:
sample6

Times downloaded:
0

Preview tools and inspect metadata by tool version

Valid tools - click the name to preview the tool and use the pop-up menu to inspect all metadata

name	description	version	requirements
Matrix to QGene	file format conversion	1.0.0	none
Matrix to Powermarker	file conversion	1.0.0	none
Alchemy to Matrix	file converter	1.0.0	none

Categories



Genotyping data management

IRRI GSL manages data of customers ...

- Customer declares as private – retained in GSL Galaxy account of customer
- Customer declares data as public – loaded into Genotyping Data Management System; shared with research community



GENOTYPIC DATA MANAGEMENT

- Project planning and queries >
- Germplasm management >
- Field trial management >
- Genotypic data management** v
- GDMS >
- Data analysis >
- Breeding decision >

Genotypic Data Management System

Category Data management

Usability Easy to use



Version Beta 1

Developers Trushar Shah (ICRISAT)

Genotypic Data Management System (GDMS) enables users to store, search and retrieve molecular marker information, genotyping data and genetic maps.



This tool is in development

Login or register to post comments: [Genotypic data management](#)

Welcome

Username: Password:

The **Genotyping Data Management System** aims to provide a comprehensive public repository for genotype, linkage map and QTL data from crop species relevant to the GCP's aim to boost crop productivity and resilience for smallholders in drought-prone environments.

This system is developed in Java and the database is MySQL. The initial release records details of current genotype datasets generated for GCP mandate crops along with details of molecular markers and related metadata. The Retrieve tab on banner is a good starting point to browse or query the database contents. The datasets available for each crop species can be queried. Access to datasets requires user login.

Data may be currently exported to the following formats: 2x2 matrix and Flapjack formats. Data submission is through templates; upload templates are available for genotype, QTL and map data (type of markers - SSR, SNP and DArT).

[Contact](#)

Login Successful!

Now you can [Upload](#) , [Retrieve](#) or [Delete](#) Data

Data Uploading

(Data can be uploaded using provided templates.
To upload, select button, browse & upload template containing data.)

Please upload Marker Information before uploading Genotyping Data

Marker Information

Genotyping Data

Maps/QTLs

SSR Marker

[SSR Marker Sample Template](#)

SNP Marker

[SNP Marker Sample Template](#)

CIRS Marker

[CIRS Marker Sample Template](#)

CAP Marker

[CAP Marker Sample Template](#)

Choose File No file chosen

Submit

Genotyping Data Retrieval

Genotyping Matrix

Polymorphic Markers

Map/QTL Data

Retrieve using :

GIDs

GermplasmNames

Markers

Dataset

Genotyping Data Retrieval

Genotyping Matrix

Polymorphic Markers

Map/QTL Data

Retrieve using :

GIDs

GermplasmNames

Markers

Dataset

Select the Dataset :

Choose Data Export Format You Would Like to View

Genotyping X Marker Matrix

Flapjack

Please select the map :

Next

Flapjack data file

Flapjack Map file

Run Flapjack

Genotyping Data Retrieval

Genotyping Matrix

Polymorphic Markers

Map/QTL Data

Retrieve using :

GIDs

GermplasmNames

Markers

Dataset

Select the Dataset :

Choose Data Export Format You Would Like to View

Genotyping X Marker Matrix

Flapjack

Next

2016 Germplasm ID(s) **384** Marker(s)

Data Export Formats

Genotyping X Marker Matrix

Back

Genotype data matrix ...

	fd10	fd12	fd13	fd17	fd6	fd7	fd8	fd9	id10007384	id1004256	id1018329	id11000133	id11008929	id1						
-40000	CK00QJ008-RiceSNP-Plate5_R001_C001	T/T	C/C	T/T	N/N	N/N	G/G	A/A	G/G	A/A	G/G	C/C	A/A	G/G	C/C	G/G	C/C	T/T	A/A	C/C
-40001	CK00QJ008-RiceSNP-Plate5_R002_C001	T/T	C/C	T/T	N/N	N/N	G/G	A/A	G/G	A/A	G/G	C/C	C/C	G/G	C/C	G/G	C/C	T/T	A/A	C/C
-40002	CK00QJ008-RiceSNP-Plate5_R003_C001	T/T	T/T	T/T	N/N	N/N	G/G	C/C	A/A	T/T	A/A	T/T	A/A	T/T	C/C	A/A	T/T	A/A	G/G	T/T
-40003	CK00QJ008-RiceSNP-Plate5_R004_C001	T/T	C/C	T/T	N/N	N/N	G/G	A/A	A/A	A/A	G/G	C/C	A/A	G/G	C/C	G/G	C/C	T/T	A/A	C/C
-40004	CK00QJ008-RiceSNP-Plate5_R005_C001	T/T	C/C	T/T	N/N	N/N	T/T	A/A	G/G	A/A	G/G	C/C	C/C	G/G	C/C	G/G	C/C	T/T	A/A	T/T
-40005	CK00QJ008-RiceSNP-Plate5_R006_C001	T/T	C/C	T/T	N/N	N/N	G/G	A/A	G/G	A/A	G/G	C/C	C/C	G/G	C/C	G/G	T/T	T/T	A/A	C/C
-40006	CK00QJ008-RiceSNP-Plate5_R007_C001	T/T	T/T	T/T	N/N	N/N	G/G	A/A	A/A	T/T	A/A	T/T	A/A	T/T	T/T	A/A	T/T	A/A	A/A	T/T
-40007	CK00QJ008-RiceSNP-Plate5_R008_C001	T/T	T/T	T/T	N/N	N/N	T/T	A/A	A/A	T/T	A/A	T/T	A/A	T/T	C/C	A/A	T/T	A/A	G/G	T/T
-40008	CK00QJ008-RiceSNP-Plate5_R001_C002	T/T	C/C	T/T	N/N	N/N	G/G	C/C	A/A	A/A	G/G	C/C	C/C	G/G	C/C	G/G	C/C	A/A	A/A	C/C
-40009	CK00QJ008-RiceSNP-Plate5_R002_C002	T/T	T/T	T/T	N/N	N/N	G/G	C/C	A/A	T/T	A/A	T/T	A/A	T/T	C/C	A/A	T/T	A/A	G/G	T/T
-40010	CK00QJ008-RiceSNP-Plate5_R003_C002	T/T	T/T	T/T	N/N	N/N	G/G	C/C	A/A	T/T	A/A	T/T	A/A	T/T	C/C	A/A	T/T	A/A	G/G	T/T
-40011	CK00QJ008-RiceSNP-Plate5_R004_C002	T/T	T/T	T/T	N/N	N/N	G/G	C/C	A/A	T/T	A/A	T/T	A/A	T/T	C/C	A/A	T/T	A/A	G/G	T/T
-40012	CK00QJ008-RiceSNP-Plate5_R005_C002	T/T	T/T	T/T	N/N	N/N	T/T	A/A	A/A	T/T	A/A	T/T	A/A	G/G	T/T	A/A	T/T	A/A	A/A	T/T
-40013	CK00QJ008-RiceSNP-Plate5_R006_C002	T/T	T/T	T/T	N/N	N/N	T/T	A/A	A/A	T/T	A/A	T/T	A/A	T/T	T/T	A/A	T/T	A/A	A/A	T/T
-40014	CK00QJ008-RiceSNP-Plate5_R007_C002	T/T	T/T	T/T	N/N	N/N	T/T	A/A	A/A	T/T	A/A	T/T	A/A	T/T	T/T	A/A	T/T	A/A	A/A	T/T
-40015	CK00QJ008-RiceSNP-Plate5_R008_C002	T/T	T/T	T/T	N/N	N/N	T/T	A/A	A/A	T/T	A/A	T/T	A/A	T/T	T/T	A/A	T/T	A/A	A/A	T/T
-40016	CK00QJ008-RiceSNP-Plate5_R001_C003	T/T	T/T	T/T	N/N	N/N	G/G	A/A	G/G	A/A	G/G	C/C	C/C	G/G	C/C	G/G	C/C	T/T	A/A	C/C
-40017	CK00QJ008-RiceSNP-Plate5_R002_C003	T/T	C/C	T/T	N/N	N/N	G/G	A/A	G/G	A/A	G/G	C/C	C/C	G/G	C/C	G/G	C/C	T/T	A/A	C/C
-40018	CK00QJ008-RiceSNP-Plate5_R003_C003	T/T	C/C	T/T	N/N	N/N	G/G	A/A	A/A	A/A	G/G	C/C	C/C	G/G	C/C	G/G	C/C	T/T	A/A	C/C
-40019	CK00QJ008-RiceSNP-Plate5_R004_C003	T/T	C/C	T/T	N/N	N/N	G/G	A/A	G/G	A/A	G/G	C/C	A/A	G/G	C/C	G/G	C/C	T/T	A/A	C/C
-40020	CK00QJ008-RiceSNP-Plate5_R005_C003	T/T	C/C	T/T	N/N	N/N	G/G	A/A	G/G	A/A	G/G	C/C	C/C	G/G	C/C	G/G	C/C	T/T	A/A	C/C

Genotyping Data Retrieval

Genotyping Matrix

Polymorphic Markers

Map/QTL Data

Fingerprinting Data

Mapping Population

Please select the Lines :

- select -

CK001BB042-VBP_R001_C001

CK001BB042-VBP_R001_C002

CK001BB042-VBP_R001_C003

CK001BB042-VBP_R001_C004

and

CK00YR020-VBP_R001_C001

CK00YR020-VBP_R002_C001

CK00YR020-VBP_R003_C001

CK00YR020-VBP_R004_C001

CK00YR020-VBP_R005_C001

Submit

'127' markers are polymorphic between 'CK001BB042-VBP_R001_C001' & 'CK00YR020-VBP_R001_C001'

Marker	Marker	Marker	Marker	Marker
dd3000535	fd13	fd8	fd9	id10001250
id10001624	id10003555	id1003559	id1008267	id1008821
id1016436	id1021920	id11003281	id11003924	id11004215
id11004240	id11004398	id11004812	id11005646	id11007840
id11009201	id11009456	id11010245	id11010335	id12001224
id12001321	id12001996	id12002728	id12003803	id12004047
id12005205	id12008557	id12010050	id2000096	id2001102
id2001565	id2001831	id2002229	id2006486	id2006621
id2007526	id2009319	id2010102	id2010969	id2014034
id2016108	id3000111	id3001992	id3004190	id3004633
id3005145	id3005216	id3005817	id3006808	id3007932
id3008957	id3009433	id3010106	id3011048	id3013669
id3015399	id3017762	id3017899	id4002852	id4002913
id4003259	id4003973	id4004493	id4005120	id4007105
id4009024	id4010238	id4010621	id4011016	id4011774
id5000015	id5000759	id5003785	id5004668	id5007536
id5007714	id5010661	id5010992	id5012152	id5013231
id5014338	id5014589	id6000402	id6003000	id6004481
id6005661	id6011429	id6012426	id6013529	id6016941
id7000871	id7001478	id7001628	id7002260	id7002427
id7002784	id7002859	id7004429	id7004442	id8001908
id8002314	id8002662	id8004838	id8007472	id9002014
id9002497	id9003188	id9003471	id9004727	id9005086
id9006377	id9006988	id9007356	id9007763	ud1000711
ud2002015	ud4000438	ud8001072	wd11001469	wd2000409
wd5000542	wd8003200			

Second GRiSP Research Theme with heavy bioinformatics ...

Harnessing genetic diversity to chart new productivity, quality, and health horizons

- 1.2.** Characterizing genetic diversity and creating novel gene pools (**SNP genotypes, whole genome sequencing, phenotypes**)
- 1.3.** Genes and allelic diversity conferring stress tolerance and enhanced nutrition (**candidate genes**)

IRGC – the International Rice Genebank Collection @ IRRI

World's largest collection of rice germplasm held in trust for the world community and source countries (www.irri.org/GRC)



- Over 117,000 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

Rice exhibits deep population structure.

Phylogenetic tree for
200K SNPs on 3,000 lines
McNally et al., 2013
unpublished

Unpublished data removed

The Rice 3,000 Genomes Project: Sequencing for Crop Improvement

Kenneth McNally, Nickolai Alexandrov, Ramil Mauleon, Chengzhi Liang, Ruairaidh Sackville Hamilton,
Zhikang Li, Ren Wang, Hongliang Chen, Gengyun Zhang, Hongsheng Liang,
Hei Leung, Achim Dobermann, Robert Zeigler



+ Many Analysis Partners

-
-
-

Bioinformatics challenges of the project...

- Primary data analysis: SNP calls, reference genome refinement, phylogenetic analysis, genotype → phenotype association, etc...
- Efficient database system that allows the integration of the genebank information with phenotypic, breeding, genomic, and IPR data
- Development of toolkits/workbenches for use by research scientists and rice breeders
- Make these databases, tools, & analyses results publicly accessible (& constantly updated)

More analysis ...

Can we assemble new references?

Find important SNPs (merging with current GWAS/QTL results)

- in CDSs

- in promoters and other regulatory motifs

Reconstruct large deletions/insertions/inversions in genome

Find correlated SNPs

Focus on known genes associated with traits

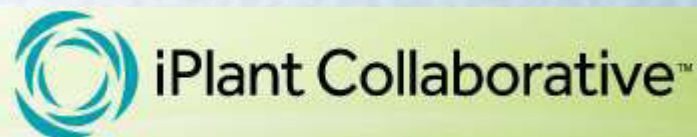
Find conserved genome regions selected by breeders

- Need for speed

- Need for collaborations

IRIC: International Rice Informatics Consortium

PAG 2013 : first introduction of the initiative to the scientific community



IRIC Portal is a central point of IRIC



NIAS
MIPS
CAS
Academia Sinica
EMBRAPA
CSHL

Cornell
Cirad
CAAS
MPI
AGI
Gramene

TGAC
IRD
KZI
Wageningen UR
Plant Onto
Uni Queensland

Initial Contact Organizations

GRiSP Centers (4)

IRRI
CIAT
IRD
Cirad

Breeding Companies (7)

Bayer CropSciences
Biogemma
Mahyco
Mars Food Global
Pioneer
RiceTec
Syngenta

Foundations (5)

Gates Foundation
NSF, U.S.A.
Sloan Foundation
USAID, U.S.A.
USDA-CREES

Universities (14)

Arizona Genomics Institute
Cornell University
Federal University of Pelotas, Brazil
Huazhong AU, China
Katsetsart University, Thailand
Kyung Hee University, Korea
Louisiana State University
Michigan State University
Oregon State University
Perpignan University, France
UC-Riverside
University of Delaware
University of Queensland, Australia
Wageningen UR, Netherlands

Others (5)

GigaScience Journal
GigaDB.org
iPlant Collaborative, U.S.A.
Gramene
Plant & Trait Ontology

Institutions (16)

Academia Sinica, Taiwan
CAAS, Beijing & Shenzhen
CAS, Beijing
Cold Spring Harbor Laboratory
EMBL-EBI, U.K.
EMBRAPA, Brazil
ICAR, India
INRA, France
Kunming Zoo Institute, China
MIPS, Germany
MPI-Tuebingen, Germany
NCGR-CAS, SIBS, Shanghai
NIAS, Japan
The Genome Analysis Centre, U.K.
USDA-Research
NCGR, Sante Fe , NM

Tech Companies (3)

BGI-Shenzhen
Affymetrix
Pacific Biosciences

Existing rice portals

MSU Rice Genome Annotation Project <http://rice.plantbiology.msu.edu/>



Rice Genome Annotation Project
Funded by the NSF



RAP-DB <http://rapdb.dna.affrc.go.jp/>



BGI-RIS Rice Information System
<http://rice.genomics.org.cn/rice/index>



to 2010 Update

RIS^e

PlantGDB
<http://www.plantgdb.org/OsGDB/>

PlantGDB



...resources for comparative
plant genomics

Gramene
<http://www.gramene.org/>



IRIC portal content

- Sequences and analysis of 3,000 genomes*
 - SNPs
 - assemblies
 - phylogenetic trees
 - genes associated with traits
 - regulatory motifs
 - most significant variations
- Other available rice genome sequences (~2,000 rice entries in SRA)
- Sequences of rice microorganisms
- Sequences of other grasses (e.g. for C4 project)
- Genotyping results from GBS, 44K and 700K affy chips
- Phenotypic data
- Gene expression data
- Gene functions and networks
- Analysis tools
- Linked to rice seeds database
- Linked to other IRRI databases and portals

***Total amount of genotyping data: $\sim 3K * 20Mio = 60Bio$**

IRIC portal development team @ IRRI

Rolando Jay Santos

Victor Jun Ulat

Frances Nikki Borja

Venice Margarete Juanillas

Jeffrey Detras

Roven Rommel Fuentes

Ramil Mauleon

Kenneth McNally

Nickolai Alexandrov

Technological advices

Marco van den Berg

Visionary input

Achim Dobermann

Management team

Hei Leung

Ruaraidh Sackville Hamilton

Kenneth McNally

Ramil Mauleon

Nickolai Alexandrov

We need you!!!

- Now Hiring: Two (2) post-doctoral positions for computational biology / bioinformatics based at IRRI
- <http://irri.org>

Join Us - The IRRI Job Board

Position Title	Apply by	Location	Job type	Specialization/Skills	
Postdoctoral Fellow - Computational Biology (PDF-2013-11-KM)	10/31/2013	Philippines	Fixed-term (Renewable), National	PhD in Genetics, Bioinformatics, Statistics or closely related disciplines.	Apply now
Postdoctoral Fellow - Bioinformatics (PDF-2013-10-NA)	10/27/2013	Philippines	Fixed-term (Renewable), International	PhD in Bioinformatics, Computer Science, (Bio) Statistics	Apply now