IRRI Galaxy: bioinformatics for rice scientists

Ramil P. Mauleon Scientist – Bioinformatics Specialist TT Chang Genetic Resources Center International Rice Research Institute



Presented in behalf of my co-authors & the development team @ IRRI

Scientists/product/theme leaders

- Michael Thomson
- Kenneth L. McNally
- Hei Leung

Laboratory, software team

- Venice Margaret Juanillas
- Christine Jade Dilla-Ermita



Outline

- Overview of IRRI & it's research agenda
- Bioinformatics activities at IRRI
- IRRI Galaxy: current state, future developments



International Rice Research Institute: part of the <u>Consultative Group on</u> <u>International Agricultural Research</u>

CGIAR



CGIAR - global partnership that unites organizations engaged in research for a foodcGIAR secure future

- International Rice Research Institute (IRRI)
- Africa Rice Center
- International Center for Tropical Agriculture (CIAT)
- International Crops Research Institute for the Semi-Arid Tropics (ICRISAT)
- International Maize and Wheat Improvement Center (CIMMYT)
- International Potato Center (CIP)

- International Center for Agricultural Research in the Dry Areas (ICARDA)
- International Institute of Tropical Agriculture (IITA)
- International Livestock Research Institute (ILRI)
- International Water
 Management Institute (IWMI)



INTERNATIONAL RICE RESEARCH INSTITUTE Los Baños, Philippines

Mission:

Reduce poverty and hunger,

Improve the health of rice farmers and consumers,

Ensure environmental sustainability

Through research, partnerships



Home of the 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

Global Rice Science Partnership : GRiSP

- A single strategic and work plan for global rice research
- Streamlines current research for development activities of the CGIAR, aligns it with numerous partners, and
- Adds new activities of high priority, in areas where science is expected to make significant contributions.



6 GRiSP Research Themes (2 are rice – research, per se)

- 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 (SNP genotypes, whole genome sequencing, phenotypes)
 - **1.3.** Genes and allelic diversity conferring stress tolerance and enhanced nutrition (candidate genes)
 - 1.4. C4 rice (Converted from C3 photosynthesis)
- 2. Accelerating the development, delivery, and adoption of improved rice varieties
 - 2.1. Breeding informatics, **high-throughput marker applications**, and multi-environment testing

IRGC – the International Rice Genebank Collection

World's largest collection of rice germplasm (located at IRRI) held in trust for the world community and source countries



- 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, highyielding varieties
- http://www.irri.org/GRC



Rice is morphologically very diverse



Structure of O. sativa SE. Asia 45 SSR Loci on 2252 lines. Indica (DARwin5, unwtd NJ, SM coef.) S. Asia The color represents group

assignment for K = 9 with a minimum allele frequency of 0.65 for model-based structure analysis.



Asia

IRRI

Plant Diversity for the Resource-Poo



A high quality reference genome is available

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

chorespiez005 Nature



Research themes, Bioinformatics & Galaxy

- Leveraging the reference genome, datasets are sequencing technology-based
 - Requires bioinformatics knowledge
 - Small bioinformatics team at IRRI =
- We need to
 - enable field/bench researchers for bioinformatics
 - share bioinformatics solutions across GRiSP partners
 - share solutions with rice research community as a whole
- Galaxy bioinformatics workbench (http://galaxyproject.org/) an easy choice



Galaxy features that fit our needs

Open, web-based platform for <u>accessible</u>, <u>reproducible</u>, and <u>transparent</u> computational biomedical research.

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



GRiSP 1.2.1: Rice SNP Consortium for enabling genomewide association studies

- Data from high-density genotyping using 44K, 700k Affymetrix SNP arrays and Illumina Beadstudio, Fluidigm medium density platforms
- Bioinformatics needs
 - Genotype data management system: SNP calling, storage, integration, retrieval, formatting for analysis
 - Analysis: GWAS pipelines, genetic analysis tools (for standard & specialized populations)
 - Genome browser: integrating published datasets & visualizing

GRiSP 2.1.3 Highthroughput SNP genotyping platform for breeding applications

Our 1st Galaxy: SNP calling workflow at IRRI



BeadXpress Scan Results (384 SNPs)



GenomeStudio + Alchemy plug-in





Galaxy _ O **_ X** 🖥 Galax nalyze Data Tools History Options * IRRI GALAXY-BIONFORMATICS WORKBENCH Get Data 0 F ALCHEMY SNP calling algorithm written by Mark Wright (Wright, McCouch, et al., 2010, CORNELL UNIVERSITY) ALCHEMY tools Your history is empty. Click 'Get Data' on the left pane to start Text Manipulation EMBOSS WWFSMD? usegalaxy.org This project is supported in part by <u>NSF</u>, <u>NHGRI</u>, and <u>the Huck</u> Institutes of the Life Sciences.

Allele calling with ALCHEMY



Why ALCHEMY SNP calling

- GenomeStudio's genotype calling algorithm is designed for human applications
 - does not consider inbred samples or population deficient in heterozygotes
- Alchemy : Open source, developed at Cornell University by Mark "Koni" Wright et al. (2010)
 - addresses the poor performance of the vendor's software on inbred sample sets
 - ability to estimate and incorporate inbreeding information on a per sample basis
 - written in C ; compiles neatly under the GNU/Linux environment



GRISP 1.2.3: The Rice 3,000 Genomes Project: Sequencing for Crop Improvement

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

IRRI INTERNATIONAL RICE RESEARCH INSTITUTE





+ Many Analysis Partners

NIAS	Cornell
MIPS	Cirad
CAS	CAAS
Academia Sinica	MPI
EMBRAPA	AGI
CSHL	Gramene
•••	Uni Queensland

TGAC IRD BGI KZI Wageningen Plant Onto

...



Bioinformatics challenges of the project...

- Efficient database system that allows the integration of the genebank information with phenotypic, breeding, genomic, and IPR data for enhanced utilization
- Development of <u>toolkits/workbenches</u> to enable gene/genotype->phenotype predictions by research scientists and rice breeders
- Make these databases, tools, & analyses results available (& updated) along with the rice gene bank



Focus of bioinformatics developments in 3k project

- Sequence/genotype data management, manipulation system
 - o include primary data visualization (SNPs, genome)
- Data analysis workbench
 - Analysis tools, w/ workflow management
 - Results visualization (haplotypes, population structures, GWAS results)
 - Highly efficient sequence/analysis results data storage model & phenotype database

Objective 1 : Sequence primary analysis

- Milestone 1: Construction of new variety group reference genomes for the representative clades
 - Quick draft genomes: SOAP de novo –based assembly (Assembl, V.J. Ulat - IRRI)
 - Velvet fails with our dataset (legitimate out-of-memory error, likely due to repeats)
 - New strategies (adapt/optimize/create algorithms) for high-quality assembly of new references, thru collaborations with partners mentioned before..





Objective 1 : Sequence primary analysis (contd)

Milestone 2: SNP genotypes construction & diversity analysis: Haplotype structure & local (genomeblock) diversity analysis

- Main problem:
 - Number of samples (3,042 varieties) overwhelms existing software & computers (for SNP discovery, a big problem)
- One Proposed Solution : PANATI
 - Koni Wright PhD thesis, Cornell University Very fast SNP discovery and genotype calling using SW alignment



PANATI (http://panati.sourceforge.net)

- No hard limits on the number of mismatches and in/dels imposed by the algorithm
- Designed for and best suited for analysis of population <u>samples with high diversity</u> or for the use of a divergent proxy reference sequence for species which have no adequate reference of their own
- Fast execution even when there is <u>high divergence</u> between the sample and the reference sequence
- free for academic use



PANATI technical features

- Read lengths of any size
 - Input can be mixes of different read lengths and singleend or paired-end formats
- Flexible trade-offs between speed and memory usage
- <u>Multithreaded parallel execution</u> of mapping and alignment scaling in linear performance up to 64 CPUs (higher has not been tested)
- Ability to <u>read compressed FASTQ files</u> in bzip2 or gzip formats directly
 - will automatically use pbzip2 for parallel decompression of pbzip2 compressed files if the program is available

Objective 1 : Sequence primary analysis (contd)

- Milestone 3: Annotation of constructed variety reference genomes, genotypes/haplotypes of the 10k genomes, & diversity analyses results
 - Intersection of results from various annotation pipelines
 - RAP pipeline(NIAS , T. Itoh et al)
 - PASA (TIGR)
 - Gramene evidence-based method
 - Maker (GMOD)

Objective 2 : Build database & visualization tools for the genomes / genotypes / haplotype/diversity analysis results

Milestone 1. Building the project genome browser; some issues:

- Multiple reference genomes to display & call SNPs from
 - Per reference view, several at a time
 - Super ("pan") genome view
- So many varieties to display
 - Pick & show subsets? Global Display?
 - Regional/global genome comparisons between varieties



Option 1: UCSC Genome Browser

- Good
 - o Fast even for large datasets
 - Funded, with large community support base
 - Nice integration with Galaxy
 - Pick & choose varieties in Galaxy → UCSC gbrowser visualization
- Not so good
 - o Painful installation
 - Steep learning curve (esp. for customizations)
 - Lack of comparative genome view



UCSC Browser hosted @ CU, mirror @ IRRI

Genomes	Genome Browser	Tools	Mirrors	Downloads	My Data	About Us	View	Help			
		U	CSC Ge	nome Bro	owser on	O. sativ	a Jun.	2009 (1	ASUv6.1) A	ssembly (oryS	at1)
			mov	/e <<< <<	< > >>	>>> zoom	n in 1.5x	3x 10x	base zoom out	1.5x 3x 10x	το Το τα καλού σ
				chr9:1-90,0	00 90,000 bp	. enter position	or search t	erms		go	
		2000-000 800-000-000-000-000-000-000-000-	000 10,000	15,000 20,000	9 25,000 30,	Known Gene	s from MSUV MSUV6	Rice Annotatio	ions With Symbols ns		
	1536_SNP	Chip	0509g00999.1 🏚 LOC_0509g010			LOC_0509901002 LOC_05	09g01006,1	LOC_OS09 536 SNP Chip 4k SNPs Chip	9901010.1 K LOC_03 LOC_0509901020	899901030,1]}→>>>>>>3 0,1]}=3 LOC_OS9999105 LOC_OS9990105	LOC_OS®9901066,1
	move start			eature for deta r labels up or c						s for track options. Drag	move end
			defau collapse			own controls be	elow and	oress refresh	configure reverse to alter tracks dis yed in more compa	splayed.	
			-	72 I		lapping and			ed in more compa	refresh	
			Base Po dense	-	ide <mark>→</mark>			Re: hid	<u>str Enzymes</u> le ▼		
			-		G	enes and Ge	ne Predio	tion Tracks	;	refresh	
			Known R Genes pack	_ [tice Gene Anno back 💌	<u>otations</u>					
			-			Variatio	n and Re	peats		refresh	
			1536 SN dense	A.	<u>4k SNP Chip</u> dense ▼						
							refresh				
											IDDI



Option 2: GMOD Gbrowse

- Good
 - "Comfort zone" genome browser installation, customization
 - Simple DB schema (basic install)
 - Funded, with large community support base
 - Comparative genome view supported
 - Integrates with Galaxy (similar to UCSC Gbrowser)
- Not so good
 - o Slow for large datasets



GMOD Gbrowse with draft genome assembly anchored rice reference genome

File 👻 Help

MSU Oryza sativa japonica	v6.1: 37 kbp from chr01:143	,000180,000			
Browser Select Tracks Cor	mmunity Tracks Custom Track	s Preferences			
Search Landmark or Region:					
	earch		Annotate Restriction	n Sites 🔹 Configure	Go
Examples: chr01, chr02, chr01:80,00 Data Source	00120,000, membrane trafficking, NU	JT21, YAL063C.	Sarall/Zaami 🛠 🔇		
MSU Oryza sativa japonica v6.1			Scroll/20011:		🗠 🗆 ЕПР
Overview	chr91				-+ + <u> </u> + + → 40M
■ Region	70k 80k 90k 10		150k 160k 170k 18 <mark>0k 1</mark>	90k 200k 210k 220k 230k	240k 250k 260k
Details	k 15 ★ ■ ⊠ ⊠ □ H ≅ MSU Osa1 Rice Gene M solgol290.1 L0C_0solgol302.1 → □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □	lodels 307.1 307.2		1 170k L0C_0so1g01340.2 L0C_0so1g01340.1 CDC_0so1g01350.1 ←□ropromon L0C_0s01g01350.2 ←□ropromon	180k
		Sele	t Tracks Clear highlighting		

Objective 2 : Build database & visualization tools for the genomes / genotypes / haplotype/diversity analysis results (contd)

Milestone 2: Build data analysis application tools coupled to the sequence database

- Some existing tools (input from collaborating institutes)
 - EU- transPLANT project: computational infrastructures for plant genomics
 - o Haplophyle @ CIRAD
- Build Galaxy for tools developed/adopted by project
 - Sequence/genotype management
 - Novel data analysis methods, workflows

Objective 3 : Genotype - > Phenotype analysis/ breeders' toolkit

- Milestone 1 Create an integrated phenotype database
- Milestone 2 Association (GWAS) & genetic analysis tools
 - o TASSEL , java web start in IRRI GALAXY
 - R packages integrated into IRRI GALAXY
 - R-GENETICS
 - GAPIT Buckler, et al., Cornell University
- Milestone 3 The breeders' toolkit
 - Major project.. Putting all these tools together in a target user-friendly package
 - Breeder's use cases captured as workflows

Is GALAXY up to this task??

Will breeders use it??



IRRI Galaxy: Current status

- Deployed in the cloud (Amazon Web Services Large instance – Singapore region)
- Streamlined to contain rice-specific tools and genotyping data
- NO NGS assembly tools in public site



Standard Galaxy release 🗌 Galaxy Analyze Data Workflow Shared Data - Visualization -Cloud-Help -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 Đ. Tools History search tools C 00 02 📄 **Running Your Own** Unnamed history 1.2 MB Get Data **Understanding how Galaxy** Send Data 2: Filter FASTQ on data 1 @ 0 🗱 **ENCODE Tools** works 1: human Illumina dataset 👁 🖉 💥 Lift-Over An in-depth tutorial **Text Manipulation Convert Formats FASTA** manipulation Filter and Sort Join, Subtract and Group Extract Features Live Quickies Fetch Sequences **Fetch Alignments** Get Genomic Scores **Uploading Data** 454 Mapping: Managing using FTP Single End account histories Operate on Genomic Intervals Galactic quickie # 15 alactic quickle # 17 actic quickie # 19 Statistics Graph/Display Data **Regional Variation** 111 Multiple regression **Multivariate Analysis** Galaxy is an open, web-based platform for data intensive biomedical research. Evolution 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 **Motif Tools** State, and the Biology and Mathematics and Computer Science departments at **Multiple Alignments** Emory University. The Galaxy Project is supported in part by NSF, NHGRI, The Huck Metagenomic analyses Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University. Genome Diversity Phenotype Association Galaxy build: \$Rev 8154:5dcbbdfe1087\$ EMBOSS

IRRI GALAXY (current)

- Galaxy	Analyze Data Work low Shared Data - Visualization - Help - User -	Us	ing 1.6 MB
rools 🌣	IRRI	History	¢
<u>SNP management tools</u> <u>Get Data</u> Text Manipulation	NTERNATIONAL RICE RESEARCH INSTITUTE	Unnamed history	2 1.6 MB
Convert Formats FASTA manipulation		3: Shahid_Shabir_indxind run_Oct5_SNP.map.txt	• / %
Workflows All workflows	Welcome! You are looking at a local instance of Galaxy at the International Rice Research Institute (IRRI). To access the main server.go to <u>Galaxy main</u> .	2: Shahid_Shabir_indxind run_Oct5_Sample.map.txt	• / ×
	Announcement	1: Shahid_Shabir_indxind run_Oct5_Intensity.map.txt	• / ×
	This site is still under construction. Last Update: November 12, 2012		
	The survey projectie supported in party (then, and the mark instructed of the time detended.		

IRRI

Workflows for rice data analysis already available

- Galaxy	Analyze Data Workflow Shared Data - Visualization - Help - User -	Using 0 bytes
Tools Options ~	Workflow Canvas Alchemy to powermarker Options 🔻	Details
search tools	Alchemy-2 💥	Tool: Matrix to Powermarker
GDMS SNP management tools Alchemy tools	Intensity file Alchemy2matrix X SNP map file Alchemy2matrix X	Matrix file Data input 'input' (tabular)
Get Data	Sample map file	Edit Step Actions
Send Data ENCODE Tools	out (tabular)	Rename Dataset 🗢 output 🗢 Create
Lift-Over Community Tools	transposeTable 🔀	Add actions to this step; actions are applied when this workflow step completes.
Text Manipulation Filter and Sort	Tab-delimited Table/Matrix out1 (tabular)	Edit Step Attributes
Join, Subtract and Group Convert Formats		Annotation/ Notes.
Extract Features Fetch Sequences		Add an annotation or notes to this
Fetch Alignments Get Genomic Scores	Matrix to Powermarker 🕱	step; annotations are available when a workflow is viewed.
Operate on Genomic Intervals Statistics	output (tabular)	This tool converts a SNP matrix file to Powermarker file format.
Wavelet Analysis		

IRRI

IRRI Galaxy Toolshed is under development (1)

- Galaxy Tool Shed		Repositories Help -	User -					
6 valid tools on Nov 18, 2012	Repositories							
Search Search for valid tools Search for workflows	search repository name, desc Advanced Search	ription						
Search for workflows	<u>Name</u> ↓	<u>Synopsis</u>	Metadata Revisions	Tip Revision	<u>Category</u>	<u>Owner</u>	Average Rating	Alert
Browse by category	file_conversion_tools -	matrix to powermarker X matrix to qgene	1:bf50914d2d07	1:bf50914d2d07	File Conversion	sample6	****	
My Repositories and Tools Repositories I own								
 My writable repositories 		Repositories Help - Use						
My invalid tools	ategories							
Available Actions	earch repository name, descriptior							
N	ame	Description			Repositories	1		
Fi	ile Conversion	file conversion tools @ IRRI			1			
S	ample Tools	for testing only			3			



IRRI Galaxy Toolshed is under development (2)

Galaxy Tool Shed

Search

Search for valid tools

6 valid tools on Nov 18, 2012

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

d		Repositories Help + User +		
Rep				Repository Actions
searc	file_conversion_tools			
Advan Name file_ Categor search repo	Clone this repository: hg clone http://mauleon@localhost:8001/to Name: file_conversion_tools Synopsis: matrix to powermarker X matrix to qgene Revision: 1:bf50914d2d07 Owner: sample6 Times downloaded: 0	oolshed/repos/sample6/file_conversion_tools		
Name File Convers Sample Too	Preview tools and inspect metadata Valid tools - click the name to preview the	a by tool version 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



Share data, import into current analysis (upon publication of studies..)

Data Libraries	
search dataset name, info, messag Advanced Search	ge, dbl 🔍
<u>Data library name</u> ↓	Data library description
1000 Genomes	Data from the 1000 Genomes Project FTP site
AC-exome	
Bushman	Data for Nature Letter "Complete Khoisan and Bantu genomes from southern Africa"
ChIP-Seq Mouse Example	Data used in examples that demonstrate analysis of ChIP-Seq data
<u>Chobi</u>	
Codon Usage Frequencies	
Coleman	IonPGM
Erythroid Epigenetic Landscape	Dynamics of the epigenetic landscape during erythroid differentiation after GATA1 restoration
<u>Evolutionary Trajectories in a</u> Phage	Experimental evolution (Illumina)
GATK	
GCAT	Consortium
Genome Diversity	Nucleotide polymorphisms for several threatened species



Solving the data mining issue for large data/results sets

- BIO HDF5 technology (Hierarchical Data Format) http://www.hdfgroup.org/projects/biohdf/
- Bottom line:
 - very fast data mining of alignments (SAM/BAM), sequences when the data model/file organization & tools (C APIs & libraries) are used
 - o Pilot ongoing now for 2,000 samples genotype data



HDF vs BAM Performance



- Avg. 8x import improvement
- Avg. 4x export improvement
- Improved compression
- Improved organization
- Consistent scaling

geospíza

from www.hdfgroup.org/pubs/presentations/BIOHDF-BOF-SC09-final.pdf

Page 21

Copyright 2009 Geospiza, Inc. All Rights Reserved.

Projects in IRRI Galaxy bioinformatics workbench

- SNP data pre-processing & calling (Alchemy, PANATI M. Wright)
- Data format manipulation for downstream analysis tools
- Population analysis tools
 - o Structure (Pritchard et al.)
 - o Ade4 R package (Chessel et al.) for Analysis of Molecular Variance
- Downstream sequence analysis tools e.g. unique primer design (Triplett et al, Colorado State University, in prep)
- Interfaces for SNPs data management & analysis
 - o GWAS: TASSEL (Bradbury et al.), GAPIT
 - o GBS analysis pipeline
- Pick & choose data to visualize: Varieties → Genome browser



Summary



¹ Including publicly accessible germplasm from IRRI, CIRAD, AfricaRice , CIAT and regional collections

IRR

THANKS FROM OUR CUSTOMERS ⁽²⁾

