IRRI Galaxy: bioinformatics for rice scientists

Ramil P. Mauleon

Scientist – Bioinformatics Specialist

TT Chang Genetic Resources Center

International Rice Research Institute

IRRI

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

IRRI

Outline

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

IRRI

International Rice Research Institute: part of the Consultative Group on International Agricultural Research

CGIAR

IRRI

CGIAR - global partnership that unites organizations engaged in research for a food-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)

IRRI

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

IRRI

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.

cirad

Embrapa

IRRI

AfricaRice

CIAT

+++

JIRCAS

IRD

IRRI

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

IRRI

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, high-yielding varieties**
- **http://www.irri.org/GRC**

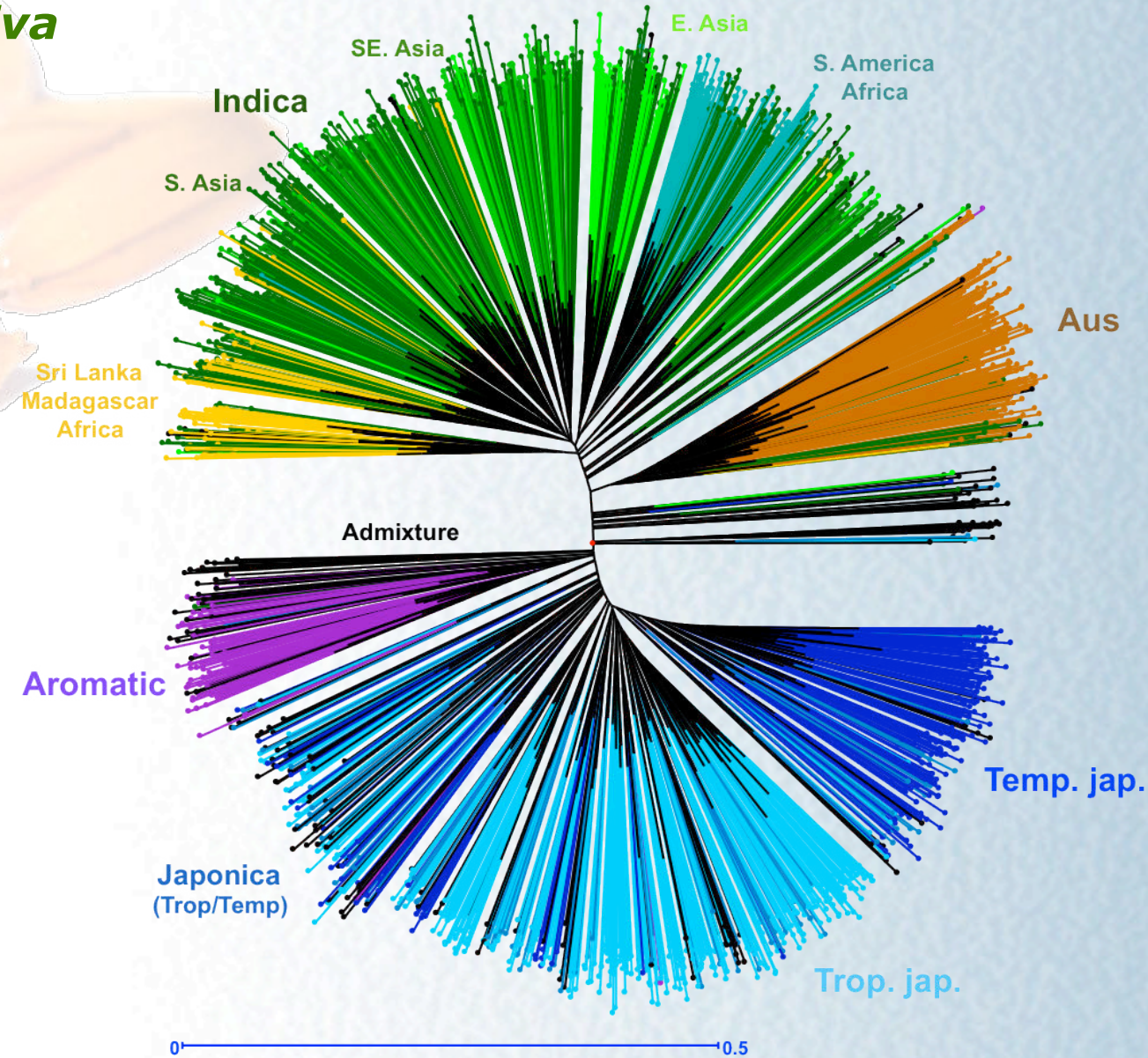
IRRI

Rice is morphologically very diverse

Structure of *O. sativa*

45 SSR Loci on 2252 lines. (DARwin5, unwtd NJ, SM coef.)

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



IRRI

cirad

AfricaRice

CORNELL

CIAT

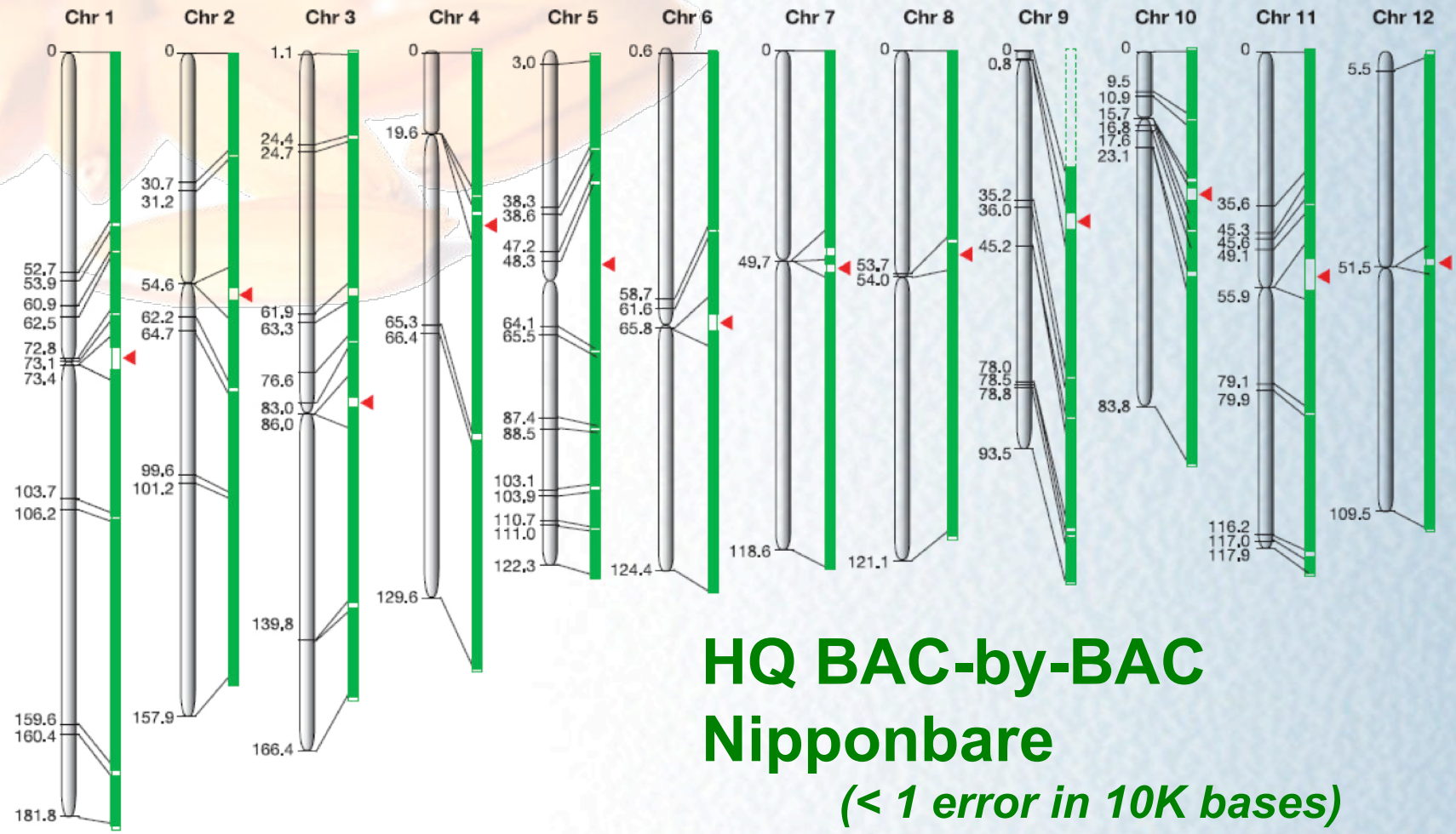
Embrapa

Generation
Cultivating Plant Diversity for the Resource-Poor

Rice exhibits deep population structure.

IRRI

A high quality reference genome is available



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.

IRGSP 2005 Nature 436:793-800

IRRI

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

IRRI

Galaxy 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.

IRRI

GRiSP 1.2.1: Rice SNP Consortium for enabling genome-wide association studies

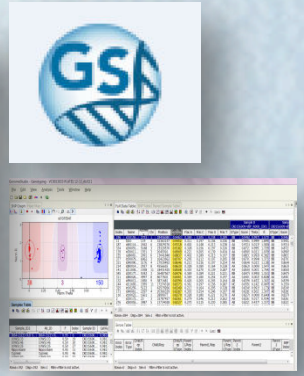
- Data from high-density genotyping using 44K, 700k Affymetrix SNP arrays and Illumina Beadstudio, Fluidigm medium density platforms
- Bioinformatics needs
  - **Genotype data management system**: SNP calling, storage, integration, retrieval, formatting for analysis
  - **Analysis**: GWAS pipelines,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

GRiSP 2.1.3 High-throughput SNP genotyping platform for breeding applications

IRRI

Our 1st Galaxy: SNP calling workflow at IRRI

BeadXpress Scan Results (384 SNPs)



GenomeStudio + Alchemy plug-in

Allele calling with ALCHEMY

IRRI

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

IRRI

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

华大基因
BGI

CAAS

+ Many Analysis Partners

NIAS	Cornell	TGAC
MIPS	Cirad	IRD
CAS	CAAS	BGI
Academia Sinica	MPI	KZI
EMBRAPA	AGI	Wageningen
CSHL	Gramene	Plant Onto
…	Uni Queensland	…

IRRI

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 toolkits/workbenches 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

IRRI

Focus of bioinformatics developments in 3k project

- Sequence/genotype data management, manipulation system
  - 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

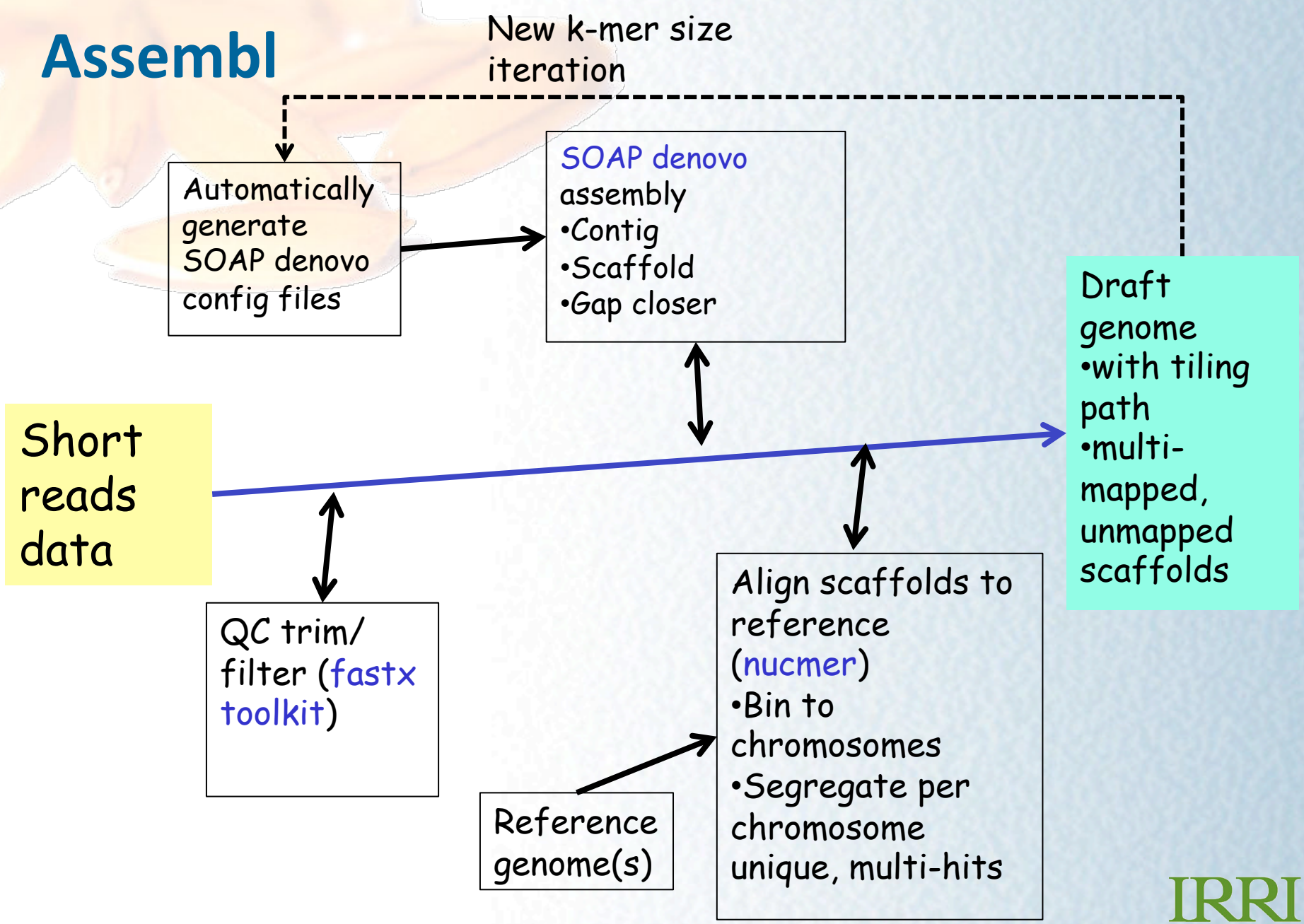
IRRI

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..

IRRI

Assembl



IRRI

Objective 1 : Sequence primary analysis (contd)

Milestone 2: SNP genotypes construction & diversity analysis: Haplotype structure & local (genome-block) 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

IRRI

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 samples with high diversity 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 high divergence between the sample and the reference sequence
- free for academic use

IRRI

PANATI technical features

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

IRRI

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)

IRRI

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

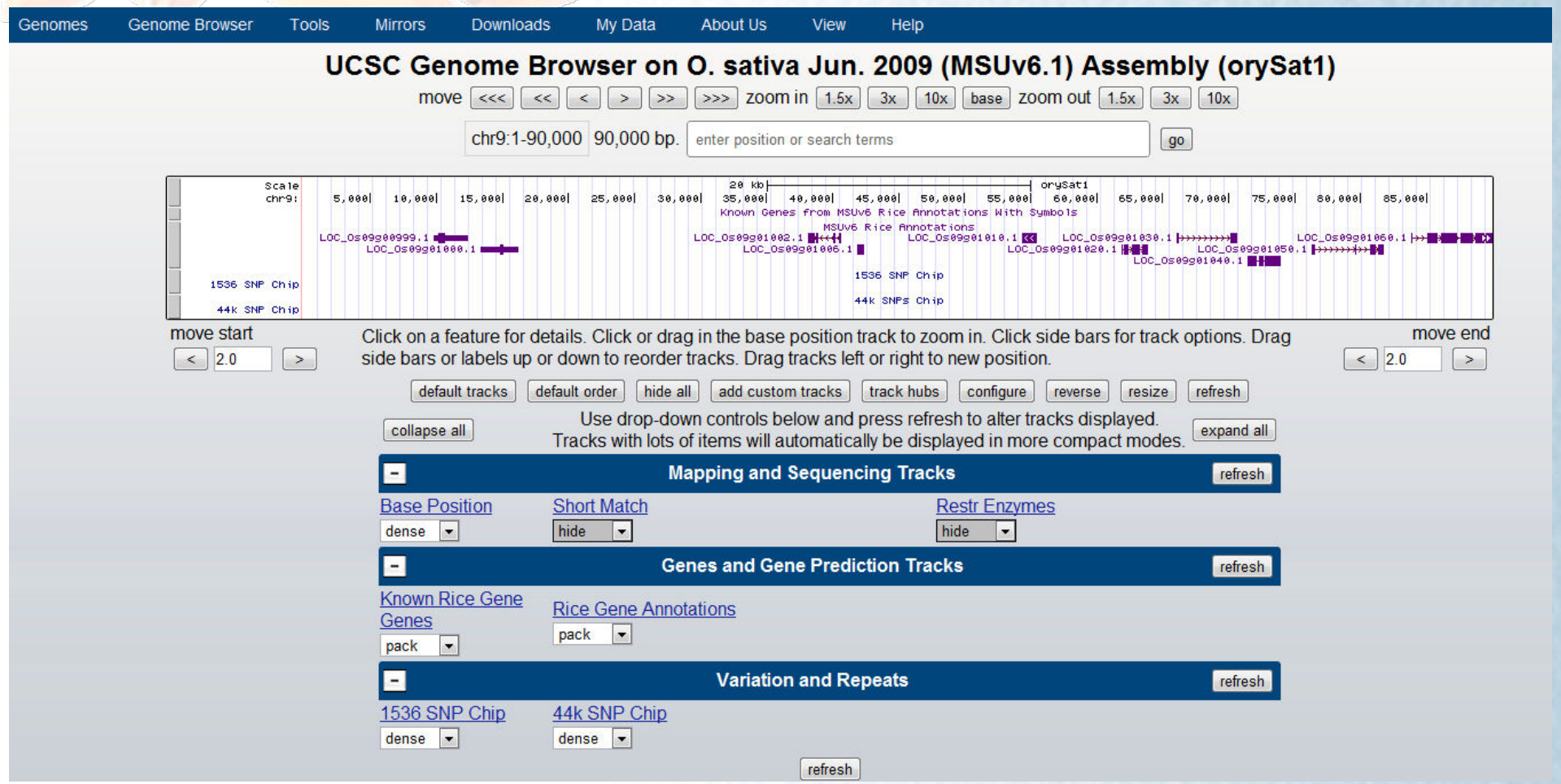
IRRI

Option 1: UCSC Genome Browser

- Good
  - 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
  - Painful installation
  - Steep learning curve (esp. for customizations)
  - Lack of comparative genome view

IRRI

UCSC Browser hosted @ CU, mirror @ IRRI



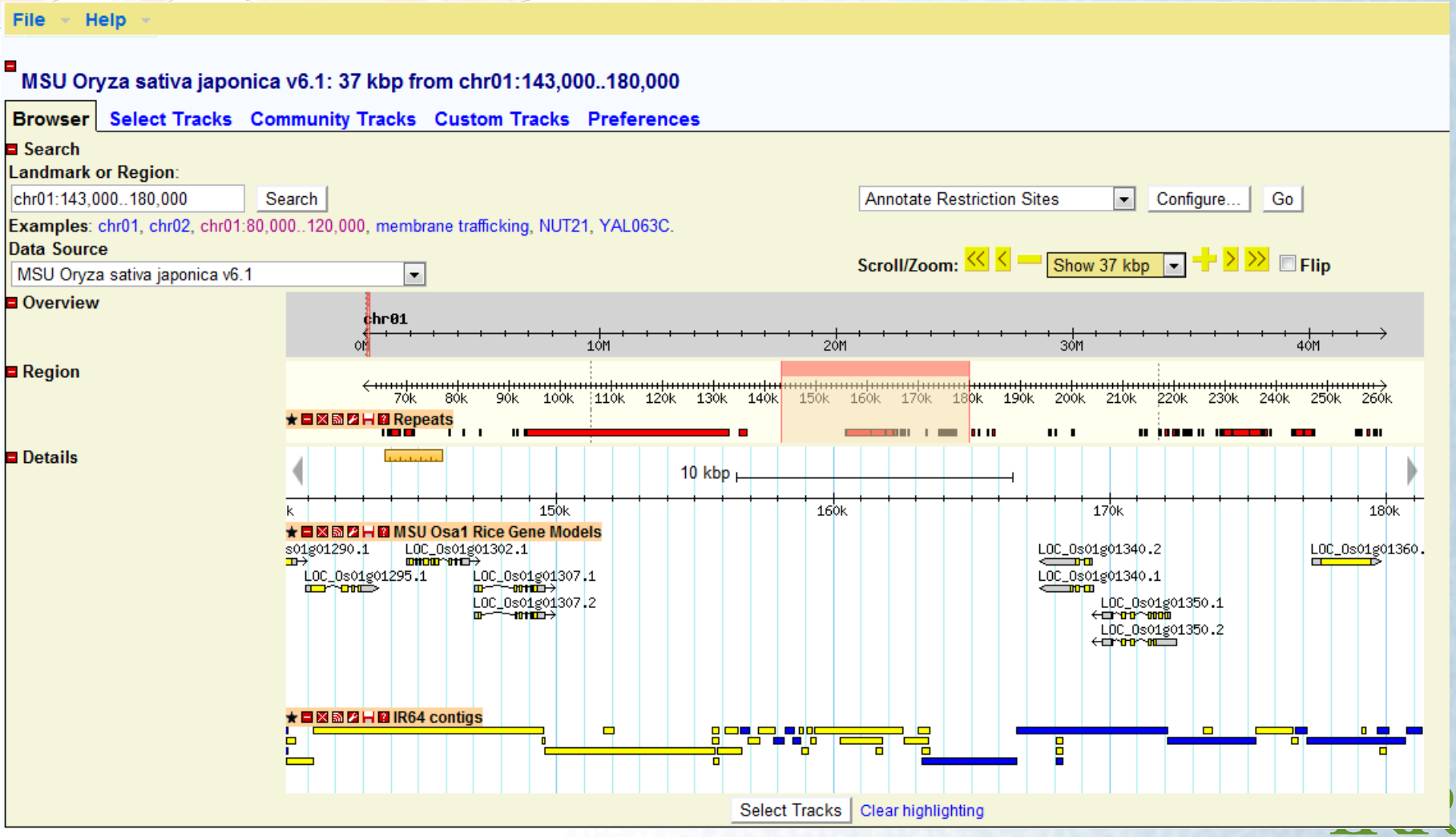
IRRI

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
  - Slow for large datasets

IRRI

GMOD Gbrowse with draft genome assembly anchored rice reference genome



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
  - Haplophyle @ CIRAD
- Build Galaxy for tools developed/adopted by project
  - Sequence/genotype management
  - Novel data analysis methods, workflows

IRRI

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

- Milestone 1 – Create an integrated phenotype database
- Milestone 2 - Association (GWAS) & genetic analysis tools
  - 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

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

IRRI

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

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.

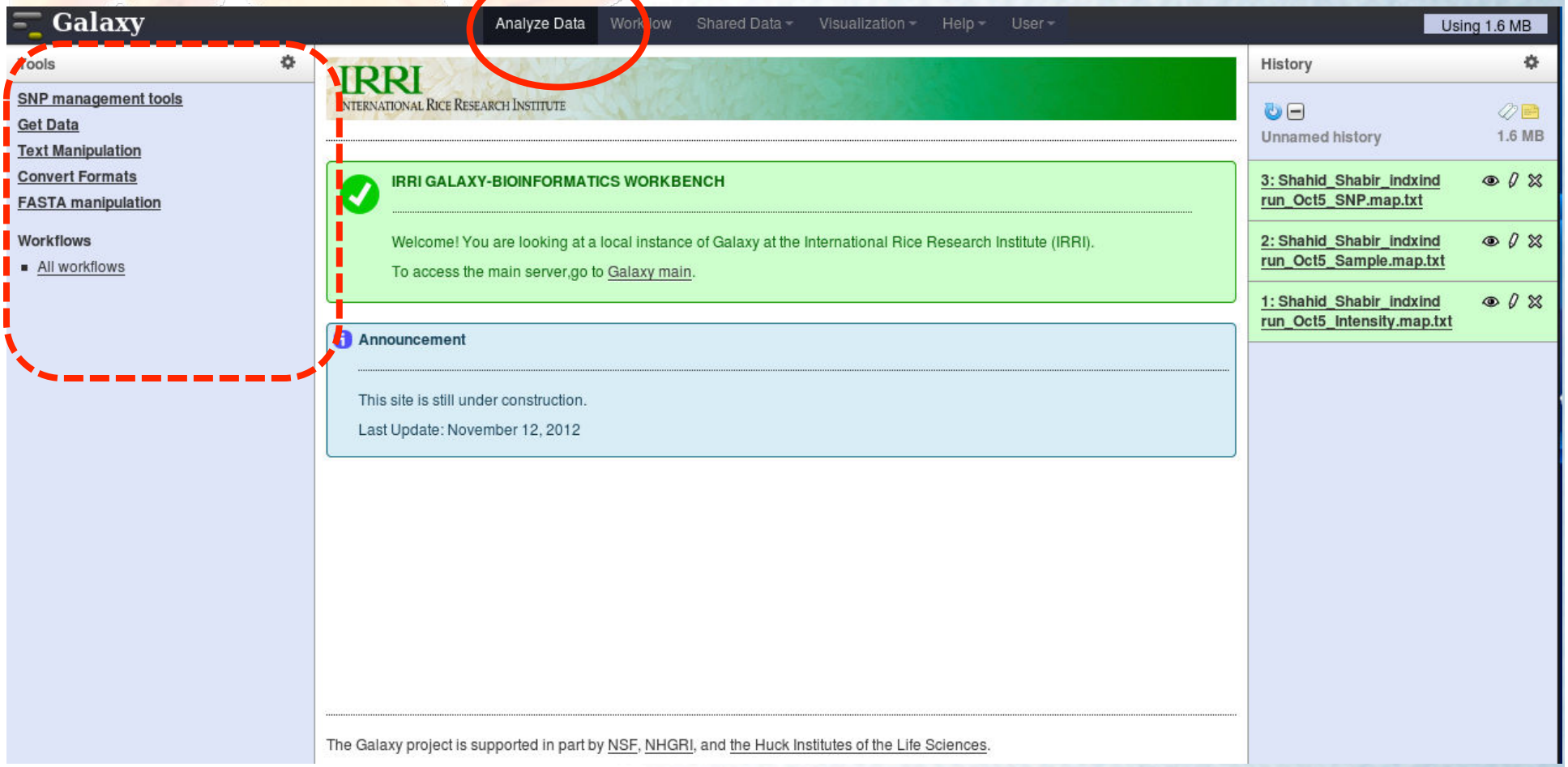
Galaxy build: **$Rev 8154:5dcbbdfe1087$**

History

Unnamed history — 1.2 MB

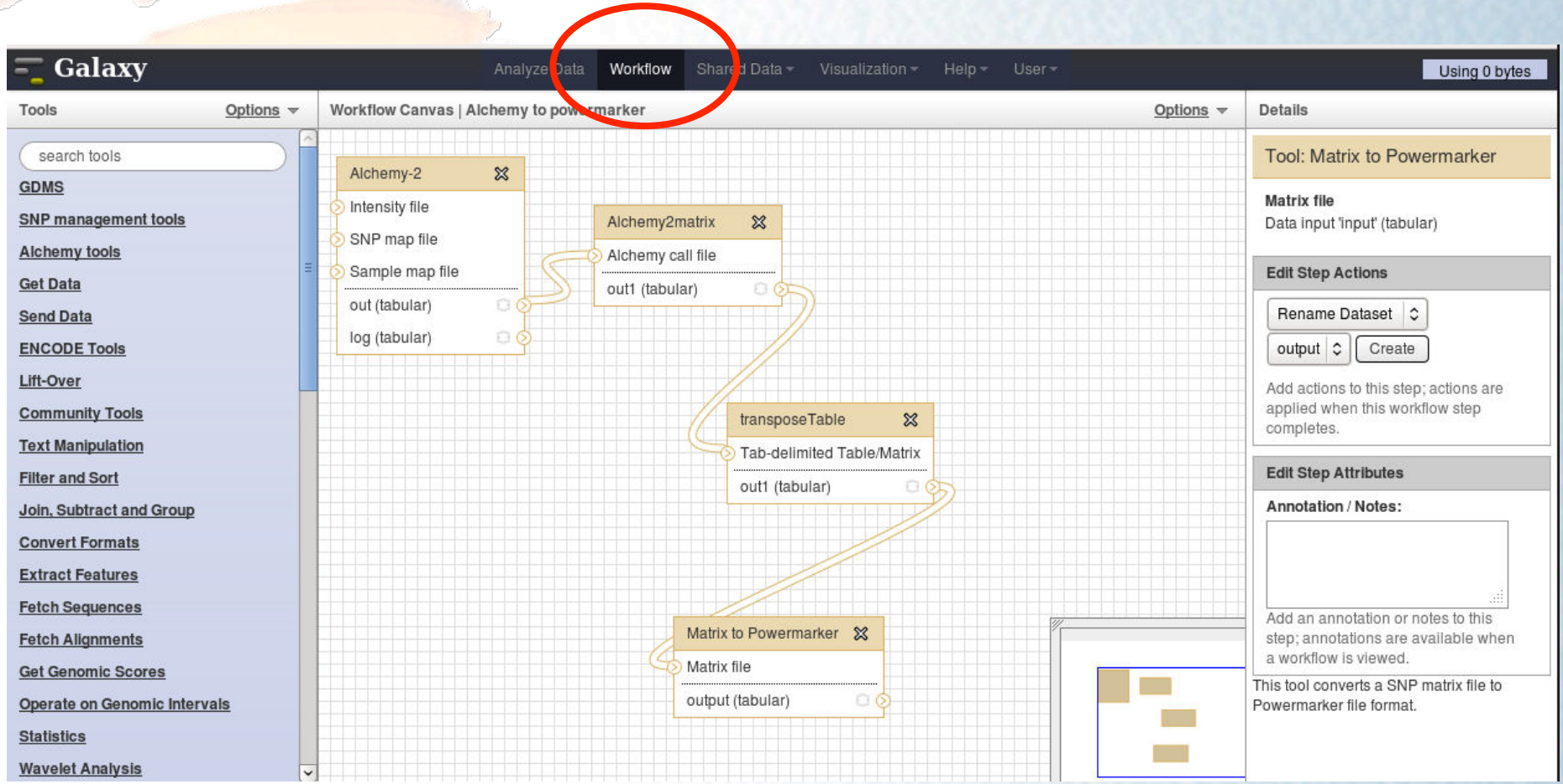
- 2: Filter FASTQ on data 1
- 1: human Illumina dataset

IRRI GALAXY (current)



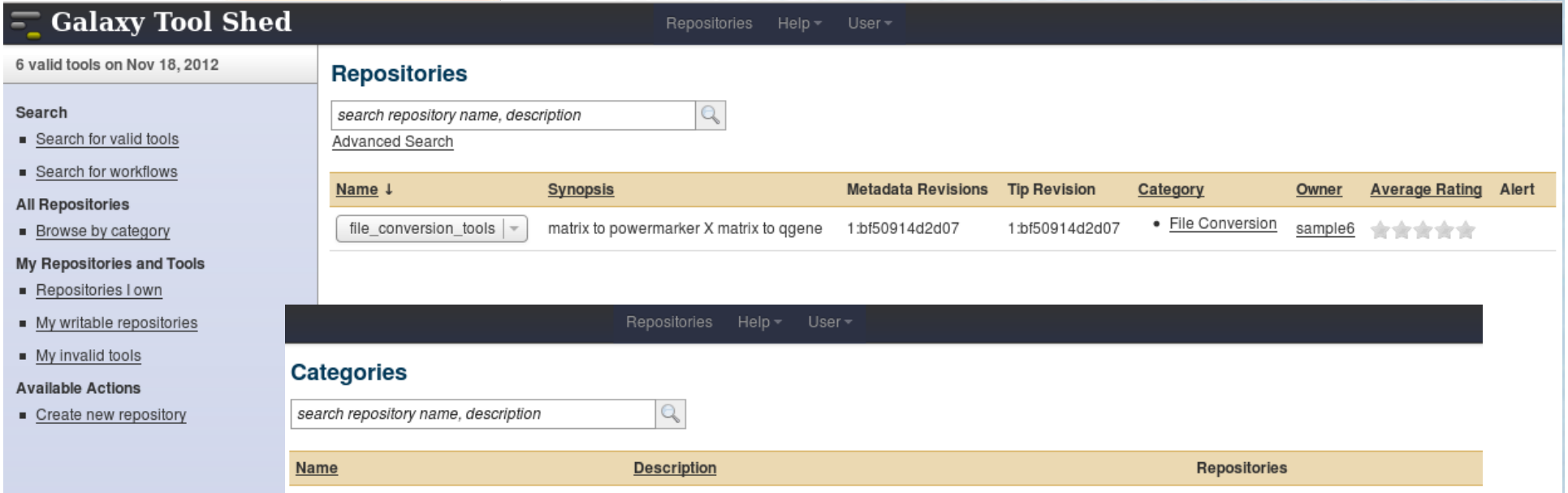
IRRI

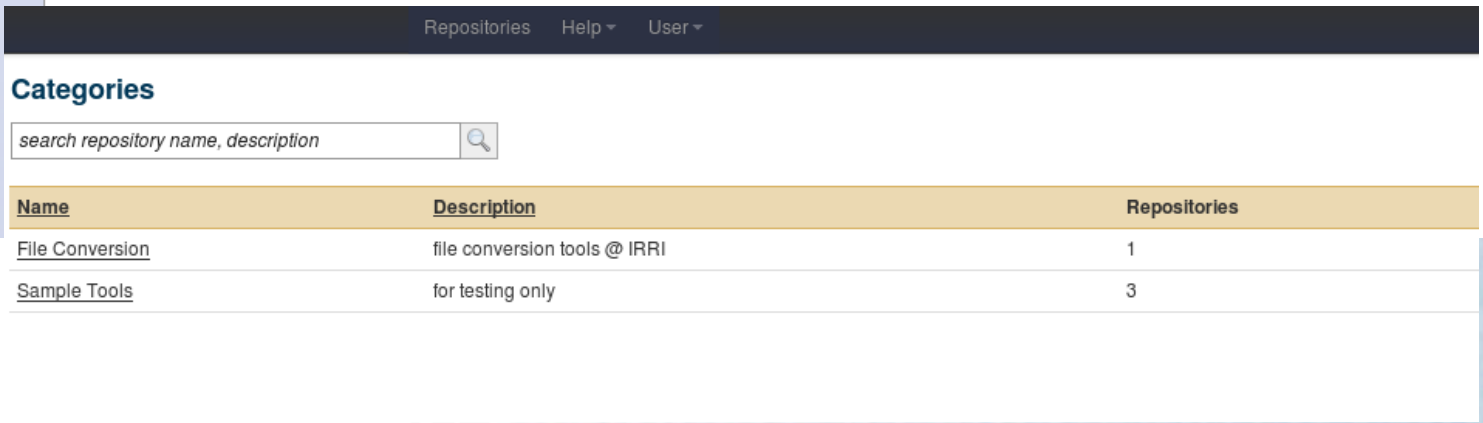
Workflows for rice data analysis already available



IRRI

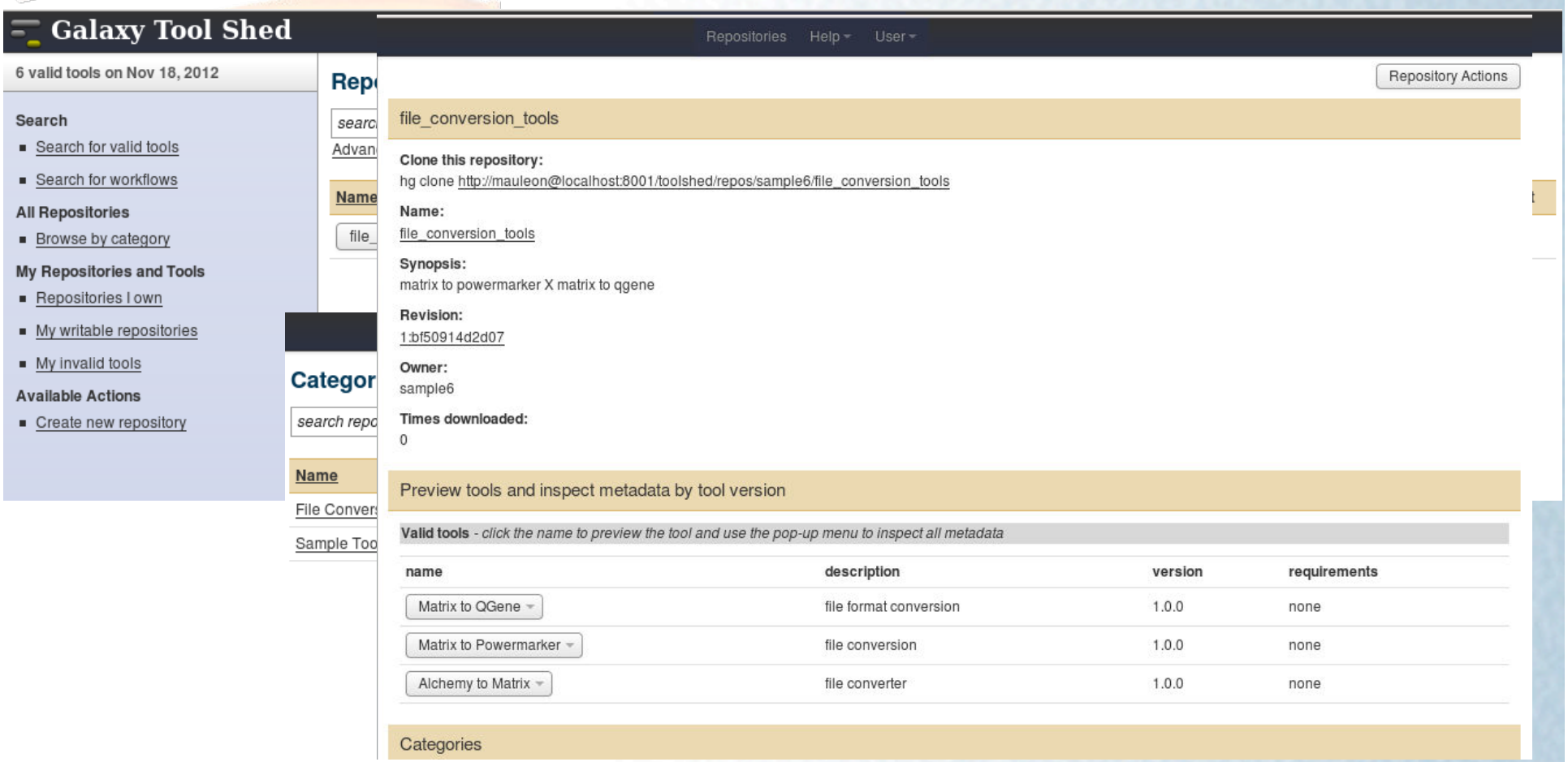
IRRI Galaxy Toolshed is under development (1)





IRRI

IRRI Galaxy Toolshed is under development (2)



IRRI

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

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Cloud | Help | User

Data Libraries

search dataset name, info, message, dbi

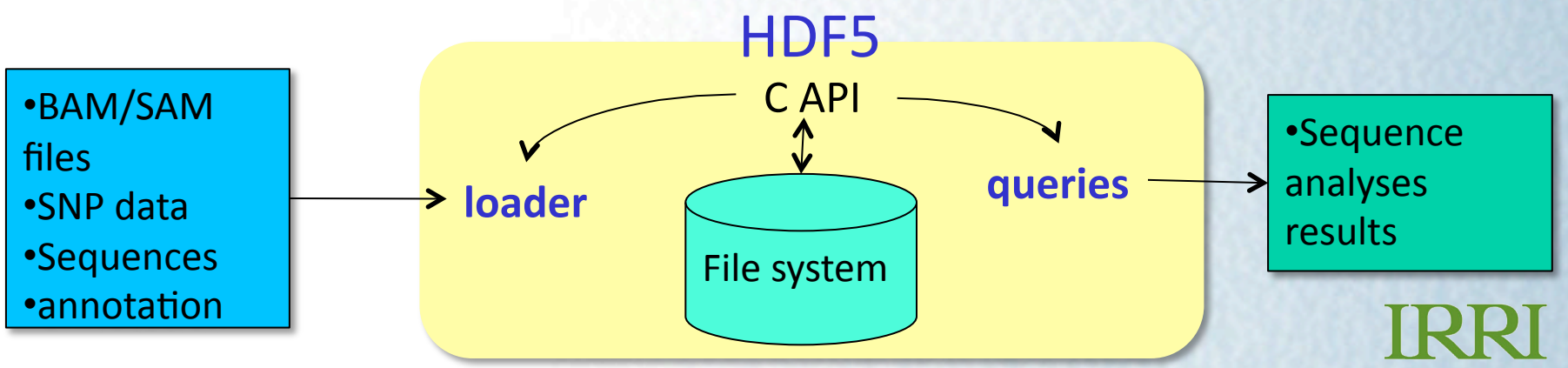
Advanced Search

Data library name ↓	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
Chobi	
Codon Usage Frequencies	
Coleman	IonPGM
Erythroid Epigenetic Landscape	Dynamics of the epigenetic landscape during erythroid differentiation after GATA1 restoration
Evolutionary Trajectories in a Phage	Experimental evolution (Illumina)
GATK	
GCAT	Consortium
Genome Diversity	Nucleotide polymorphisms for several threatened species

IRRI

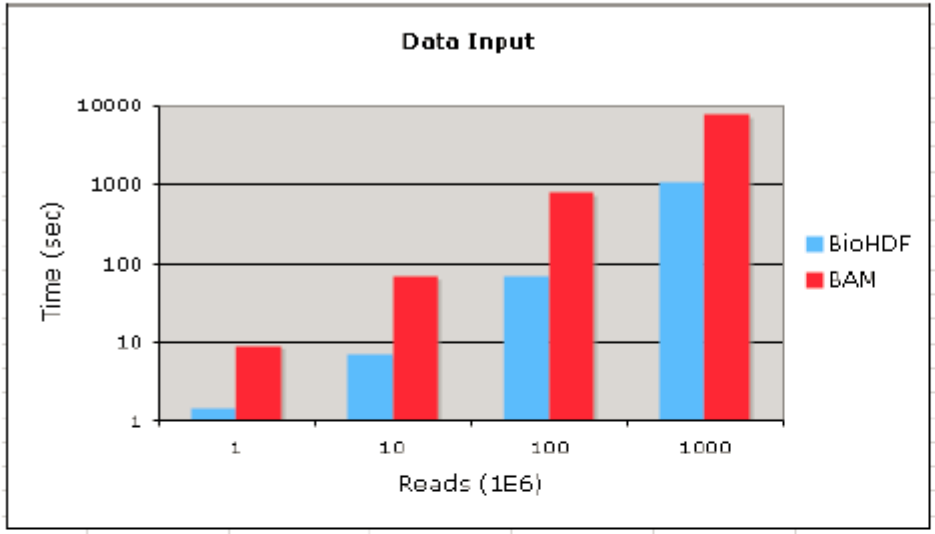
Solving the data mining issue for large data/results sets

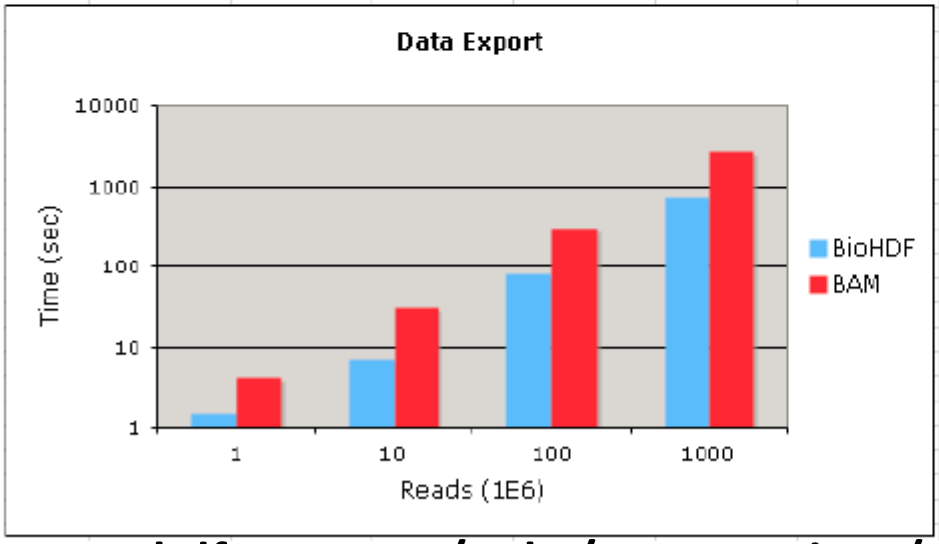
- BIO HDF5 technology (Hierarchical Data Format) - http://www.hdfgroup.org/projects/biohdf/
- Bottom line:
  - very fast data mining of alignments (SAM/BAM), sequences when the data model/file organization & tools (C APIs & libraries) are used
  - Pilot ongoing now for 2,000 samples genotype data



IRRI

HDF vs BAM Performance





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

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

geospiza

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
  - Structure (Pritchard et al.)
  - 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
  - GWAS: TASSEL (Bradbury et al.), GAPIT
  - GBS analysis pipeline
- Pick & choose data to visualize: Varieties → Genome browser

IRRI

Summary

Bioinformatics and database to Integrate sequence-phenotype data

Rice SNP Consortium
700k Affymetrix genotyping chip
2000 lines

BGI *de novo* and re-sequencing
Initial 5-10X coverage

10,000 GeneBank accessions[1]
Cultivated + close wild relatives

Phenotyping network
2000+ lines

Association genetics and QTL mapping
Predict genotype-phenotype relationships at kb resolution

Specialized genetic stocks: MAGIC populations, biparental RILs, CSSL,

Use in breeding programs

Genebank as a reverse genetics system

- Select accessions based on QTL prediction for targeted phenotyping of specific traits
- Discover novel phenotypes

IRRI & GRiSP, 华大基因 BGI CAAS

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

IRRI


THANKS FROM OUR CUSTOMERS ☺