Rice Galaxy: an open resource for plant science

Venice Juanillas[1], Alexis Dereeper[2], Nicolas Beaume[1], Gaetan Droc[3], Joshua Dizon[1], John Robert Mendoza[8], Jon Peter Perdon[8], Locedie Mansueto[1], Lindsay Triplett[7], Jillian Lang[7], Gabriel Zhou[4], Kunalan Ratharanjan[4], Beth Plale[4], Jason Haga[5], Jan E. Leach[7], Manuel Ruiz[3], Michael Thomson[1,6], Nickolai Alexandrov[1,10], Pierre Larmande[2], Tobias Kretzschmar[1,9], Ramil P. Mauleon[1]

Author affiliations

[1] International Rice Research Institute, Manila, Philippines

[2] Institut de recherche pour le développement (IRD), University of Montpellier, DIADE, IPME, Montpellier, France

[3] CIRAD, UMR AGAP, F-34398 Montpellier, France

[4] Indiana University, 107 S Indiana Ave, Bloomington, IN 47405, USA

[5] National Institute of Advanced Industrial Science and Technology, AIST Tsukuba Central 1,1-1-1 Umezono, Tsukuba, Ibaraki 305-8560 JAPAN

[6] Department of Soil and Crop Sciences, Texas A&M University, Houston, USA

[7] Department of Bioagricultural Sciences and Pest Management, Colorado State University, Fort Collins, CO 80523-1177

[8] Advanced Science and Technology Institute, Department of Science and Technology, Quezon City, Philippines

[9] Southern Cross Plant Science, Southern Cross University, Lismore, Australia

[10]Inari Agriculture Inc., 200 Sidney St, Cambridge, Massachusetts, USA

Correspondence should be addressed to R.P.M. (r.mauleon@irri.org, ORCID: 0000-0001-8512-144X)

Abstract

Background

Rice molecular genetics, breeding, genetic diversity, and allied research (such as rice-pathogen interaction) have adopted sequencing technologies and high density genotyping platforms for genome variation analysis and gene discovery. Germplasm collections representing rice diversity, improved varieties and elite breeding materials are accessible through rice gene banks for use in research and breeding, with many having genome sequences and high density genotype data available. Combining phenotypic and genotypic information on these accessions enables genome-wide association analysis, which is driving quantitative trait loci (QTL) discovery and molecular marker development. Comparative sequence analyses across QTL regions facilitate the discovery of novel alleles. Analyses involving DNA sequences and large genotyping matrices for thousands of samples, however, pose a challenge to non-computer savvy rice researchers.

Findings

We adopted the Galaxy framework to build the federated Rice Galaxy resource, with shared datasets, tools, and analysis workflows relevant to rice research. The shared datasets include high density genotypes from the 3,000 Rice Genomes project and sequences with corresponding annotations from nine published rice genomes. Rice Galaxy includes tools for designing single nucleotide polymorphism (SNP) assays, analyzing genome-wide association studies, population diversity, rice-bacterial pathogen diagnostics, and a suite of published genomic prediction methods. A prototype Rice Galaxy compliant to Open Access, Open Data, and Findable, Accessible, Interoperable, and Reproducible principles is also presented.

Conclusions

Rice Galaxy is a freely available resource that empowers the plant research community to perform state-of-the-art analyses and utilize publicly available big datasets for both fundamental and applied science.

Keywords

Rice, Breeding, workflow, genomes, high-density genotypes, reproducibility, SNP, GWAS, Galaxy project

Findings

Background

With the decreasing cost of genome sequencing, rice molecular geneticists, breeders and diversity researchers are increasingly adopting genotyping technologies as routine components in their workflows, generating large datasets of genotyping and genome sequence information. Concurrently international consortia have made re-sequencing or high density genotyping data from representative diversity collections publically available. These include, but are not limited to the medium-depth (15-20x coverage) resequencing data of the 3,010 accessions from the 3K Rice Genome (3K RG) Project (~1 – 2 million SNPs per accession) [1,2] and the 700,000 SNP Affymetrix array data for the 1,445 accessions of the High Density Rice Array (HDRA) germplasm collections [3]. The corresponding accessions are available at non-profit prices from the Genetic Resource Center (GRC) of the International Rice Research Institute (IRRI) for phenotyping, allowing subsequent Genome-Wide Association Studies (GWAS). Analysis of such datasets is a challenge to rice researchers due to (1) the fairly large data matrix and the compute-intensive algorithms that requires specialized computing infrastructure (a fairly large RAM, powerful CPU, and large disk space), and (2) the relative difficulty in using Open Source / free software tools for analysis, which are commonly provided without graphical user interface and require proper installation in a Linux operating system environment.

On the computational side, public web resources with specialized tools already exist, and are maintained at different institutions. The Rice SNP-Seek database [4,5], largely developed and hosted by IRRI, contains phenotypic, genotypic, and passport information for over 4,400 rice accessions from large scale rice diversity projects such as the 3K RG and the HDRA collections. SNP-Seek (http://snp-seek.irri.org) currently contains phenotype data for 70 different morphological and agronomic traits and stores SNPs and small indels discovered by mapping the 3K RG accessions to four published rice draft genome assemblies, collectively resulting in the discovery of ~11M new SNPs and ~0.5M new indels. While SNP-Seek focused on delivery of prior analyzed content rather than providing an analysis platform, Gigwa [6] (http://gigwa.southgreen.fr/gigwa/), hosted at the South Green portal [7] (http://www.southgreen.fr/), is a scalable and user-friendly web-based tool which provides an easy and intuitive way to explore large amounts of genotyping data from next-generation sequencing (NGS) experiments. Gigwa allows for filtering of genomic and genotyping data from NGS analyses based not only on variant features, including functional annotations, but also on genotype patterns to explore the structure of genomes in an evolutionary context for a better understanding of the ecological adaptation of organisms. Gramene [8] is a curated, open-source, integrated data resource for comparative functional genomics in crops and model plant species that, among other species, includes rice. Data and analysis tools are available as portals at the Gramene site (http://gramene.org/). In these resources mentioned, the analyses methodologies are custom-built by the respective projects.

There are other freely available web-based bioinformatics and breeding informatics software tools, optimized for plant species other than rice, including Araport (https://www.araport.org/) for Arabidopsis, Cassavabase (https://cassavabase.org/) for cassava, and The Triticeae Toolbox (T3, https://triticeaetoolbox.org/) for wheat and barley. While these tools are very useful, they are species/crop specific and custom-built for the specialized requirements of their respective communities (such as project datasets), making adoption in rice challenging for at least two reasons: (1) the need to produce curated rice datasets that work seamlessly with the software system (e.g. genome-browser ready data, curated genes, published QTLs from biparental crosses and GWAS and markers associated to traits), and (2) the need for a dedicated software development team to customize the application for rice-specific data and analyses.

The ability of software to automate repetitive analyses task is attractive for data analysts, and the public sharing of the analytical methodology (as opposed to just the raw data and the results) enhances reproducibility and is being supported by academic communities of practice such as FORCE11 (https://www.force11.org). Many research groups working with NGS data have a high demand for computing infrastructure and their complex analyses often comprise several steps using different software tools (pipeline). The deployment of these different software tools is a big challenge to small institutions without dedicated scientific computing support staff. There is no single solution to address these challenges. Our approach to help overcome them is the integration of a range of these different bioinformatics tools into the Galaxy bioinformatics system. Galaxy [9] is a web-based analysis workbench and workflow management system initiated at the Penn State University. It includes a collection of software packages which can be operated via a web browser on a public server. Galaxy is a mature community effort, supported by various high-powered institutions, is relatively easy to deploy and maintain, and thus well-suited to serve low and moderately resourced institutions such as IRRI. The graphical user interface of Galaxy means that no knowledge of code is needed, thus facilitating bioinformatics analyses by researches without computational expertise.

We built a suite of federated Galaxy resources and tools, which we collectively named **Rice Galaxy** (Figure 1). Rice Galaxy contains shared software tools and datasets tailored to the needs of rice researchers and breeders, providing computing resources through an easy-to-use interface, and allowing reproducibility and publication of analytical methodology and results.

The Rice Galaxy federated resources are hosted at:

- IRRI Galaxy @ International Rice Research Institute: http://galaxy.irri.org
- Rice Galaxy (common) Toolshed: http://52.76.88.51:8081/

DISCUSSION

1. Built-in / interoperable rice data

The Rice Galaxy system is customized to provide rice-specific genomic and genotypic data. Of primary importance is the gold-standard *japonica* variety reference genome (Nipponbare IRGSP release 1.0) [10], to which the reference gene models and most of the SNPs published have been anchored. In addition, eight medium to high quality published genomes from various sequencing projects and the respective genome annotations for each are installed as alternative genome builds and are available as drop-down menu choices in Rice Galaxy. These include four high-quality builds from *indica*-type varieties Minghui 63 and Zhengshan 97 [11], IR 8 (GenBank: MPPV00000000.1), Shuhui 498 [12], as well as an aus-type variety N 22 (GenBank: LWDA00000000.1), as well as four medium to low quality genomes, two *indica* (IR 64 , [13] and 93-11, [14]) and two aus-type rice genomes (DJ 123, [13] and Kasalath, [15]). While these references were selected to represent diversity, they further represent variety groups that display agronomically important characteristics, such as heat and drought tolerance, disease resistance, submergence tolerance, adaptation to low-phosphorus soil, wide adaptability, good grain quality, aerobic (upland) adaptation and deep roots [16-18]. Even though these genomes are highly similar to each other, they each contain unique regions (from 12.3 Mbp to 79.6 Mbp) that may harbor genes restricted to these variety-groups [5]. With the availability of several reference genomes, it becomes relatively straightforward to custom design SNP assays that are either of broad utility across varietal groups or specific to single groups.

Rice Galaxy includes genotyping data of the 3k RG (such as the 3K RG 3024 accessions x 4.8M filtered SNPs, 440K core SNPs, 1M GWAS-ready SNPs, and 2.3M indels) useful for GWAS, region-specific diversity analyses, and single locus allele mining in the shared data library.

2. Toolkits Built (and detailed discussion of each toolkit)

SNP assay design: Lift-over of SNPs from one genome to another

SNPs discovered relative to the gold-standard reference genome (Nipponbare IRGSP 1.0, [10]) are commonly used in QTL mapping (either by GWAS or biparental cross). In order to develop robust markers associated with the trait of interest, however, a SNP assay that works in the target varietal groups is needed. Consequently there is a need to "lift-over" SNPs from one genome to another (for example from Nipponbare *japonica* to an *indica* varietal group represented by IR 64). The workflow is as follows: (1) Get flanking sequences surrounding the target SNP in source genome (the main reference Nipponbare), (2) align these flanking sequences to target genome of variety of interest to verify if it hits a unique region in the target genome of similar location from the source genome, allowing some mismatches but not allowing multiple region hits, and (3) identify the flanking sequences surrounding the lifted-over SNP in the context of the target genome, for SNP assay design. The shared workflow is published in Rice Galaxy as [SNP lift-over].

3k RG data access

Rice Galaxy provides access to the raw variant call format (VCF) files of each accession in the 3K RG project via connection (as data source in Rice Galaxy) to the 3,000 rice genomes at Amazon Web Services (AWS) Public Data (https://aws.amazon.com/public-datasets/3000-rice-genome/), with tools allowing region-specific download. In Rice Galaxy, tools in the [Get Data / FROM 3KRG] section allows listing of the accessions in the 3K RG and retrieval of genotype data for a selected accession of interest from the 3K RG collection. The subset genome region of interest (chromosome name – base start – base end) can be specified and extracted from the VCF of the accession of interest stored in AWS Public Datasets.

In addition, we developed an original Rice Galaxy component called Rapid Allelic Variant extractor (RAVE), which allows simultaneous extraction of genotyping data from several accessions of the internal 3K RG resource. It relies on the PLINK software [20], which efficiently builds a user-adjusted genotyping submatrix from a compressed PLINK binary biallelic genotype table (bed file + bim, fam files). Users can customize the genotyping dataset vertically by choosing a subpopulation (*indica*, *japonica*, *aromatic, aus, tropical, temperate*, etc.) or setting a list of varieties, and horizontally by restricting variations with a list of genomic regions, or a list of gene names. Additionally, users can filter the SNP positions by specifying thresholds for missing data or minor allele frequency (MAF). The extracted VCFs can be directly generated by Rice Galaxy, stored as output into the history pane of the Galaxy interface, and can be reformatted to Hapmap, a versatile file format for further analyses such as marker (SNP) design, GWAS analyses, or visualization in within a JBrowse [21] genome browser (Vcf2jbrowse component). External SNP datasets can also be imported into Rice Galaxy and merged with 3k accessions in order to compare and look at the closest genotypes using SNiPlay [22] workflow.

GWAS analysis using TASSEL

Using this feature, it is relatively easy to construct a genotyping matrix for a subset of accessions from the 3K RG and connect associated phenotypic information to perform GWAS analyses online, with outputs being decorated with various graphical enhancements. For the 3K RG accessions, the subset 1M GWAS and 440K Core SNPs that is usable for GWAS is already available as shared dataset in Rice Galaxy (Figure 2). Researchers working on the 3K RG panel can generate new phenotyping data from their respective experiments, upload the phenotype data into Rice Galaxy, and then perform GWAS using the TASSEL bioinformatics tool [23]. The GWAS Rice Galaxy workflow implementing TASSEL and Multi-Locus Mixed-Model package for association studies is shared from SNiPlay at Rice Galaxy (Figure 3). Aside from GWAS with 3K RG datasets, researcher-generated marker and phenotype data (outside of 3K RG) can also be uploaded to Rice Galaxy for GWAS analysis.

Genomic selection using Oghma genome prediction tool

Genomic selection (GS) is a promising breeding technique with potential to improve the efficiency and speed of the breeding process in rice [24]. With the intent of enabling the GS analysis process used on the 2 datasets in the Spindel et al. [24] study, (encoding data, filtering data to keep informative markers, creating a model from training set, evaluating the model and finally, performing the prediction itself), and to automate the analysis pipeline, the relevant packages (`methods`, `fpc`, `cluster`, vegan, pheatmap, pROC, randomForest, miscTools, pRF, e1076, rrBLUP, and glmnet) for the R Statistical language (https://www.r-project.org/) were installed in Rice Galaxy and the tool suite was collectively named Oghma (Operators for Genome decipHering by MAchine learning). Quality control tool (based on PLINK) and imputation tool using Beagle [25] (https://faculty.washington.edu/browning/beagle/beagle.html) were also installed. Four phenotype prediction/classifier methods (rrBLUP, random forest, SVM and lasso) were identified as relevant and deployed as tools in Rice Galaxy (Figure 4).

Figure 5 shows the overall GS analysis workflow using Oghma. Genotypes are encoded through [encode] tool. For the training set, an encoded genotype and the corresponding phenotype files are used by a classifier tool to train a model, which can be used with another encoded genotype file to predict trait values (the genomic prediction). It is important to note that (1) both genotype for training and genotype to predict must have the same markers (and thus, genotype files must have the same number of columns) to make a prediction, and (2) the "evaluation" option of the classifier tool can have any value except 1 (it is recommended to keep the default value = 0).

A big challenge when using machine learning approaches for genomic prediction is the optimization of the model based on training data, specifically setting the best parameters of the methods mentioned prior. Oghma was designed to automate the optimization of the parameter(s) of the classifiers on the fly (as opposed to manual tweaking), thus allowing users without experience of machine learning to easily optimize a model for their own data. Oghma includes some tools to evaluate prediction accuracy to allow the user to choose the most accurate method on their data by performing a cross-validation with a user-uploaded training set. Two metrics, the coefficient of determination (R^2) and the correlation between predicted and observed phenotype, and a visualization (scatterplot of predicted vs observed) have been implemented to evaluate the methods. The [computeR2] and [plotPrediction] tools are used to compute R^2 and visualize the accuracy of prediction. These tools both take the true phenotypes and the predicted outputs as inputs (take note that both predictions and phenotypes data must be in the same order), and return the computed R^2 or the scatterplot display of true phenotype vs prediction.

Oghma can be used to evaluate a classifier (Figure 6). Like the general GS workflow, genotype and phenotype are used as input for any classifier, but the "evaluation" option must be set to 1. Fold for cross-validations are designed through the [fold] tool, which take as input the encoded file. These folds are used as extra argument by the classifier tools. The chosen classifier tool produces a file, which is not a model but the prediction of the test set for each cross-validation. This output is used as input, along with the phenotypes and folds, by [evaluation], which output some performances index (R^2 and correlation). Although it does give a real indication of performances, trying to predict the training set (i.e. using the same genotype file in the pipeline described above), or at the least, showing if the classifier is not under-fitting the data.

We installed several classifiers in Oghma to allow users to test the best one(s) suited for their dataset, as our literature survey shows that no method seems to outperform the others on all genomic prediction tasks. It was noticed that Random Forest was the most accurate and the most stable classifier on Spindel dataset [24], thus we set this as default in Oghma. An original aggregation method is also implemented in Oghma, aggregating outputs of multiple classifiers to improve prediction. This tool takes as input the prediction of *n* classifiers and tries to aggregate them through weighted mean of the prediction (weight optimized by genetic algorithm) or regression (multiple type of regression have been implemented, based on decision tree, SVM and Random Forest). Limited testing shows that this approach is promising, matching Random Forest in some cases, especially with a meta-SVM, with polynomial or linear model, as aggregation method, but still needs some improvement as the accuracy remains unstable when evaluated through cross-validation (data not shown). The aggregation method can also be evaluated using the aforementioned evaluation tools.

Diversity and population structure analysis of end-user datasets

SNP datasets - such as those extracted from the 3K RG resource after a filtering by the RAVE module or custom sets directly uploaded in Rice Galaxy environment (Figure 7) can be processed for a complete exploration and large scale analysis thanks to the SNiPlay Rice Galaxy workflow (Figure 8). The workflow is available through the instance, requiring a VCF file as input. This workflow allows various analyses: (i) SNP annotation by snpEff (http://snpeff.sourceforge.net/) wrapper preconfigured for RGAP release 7.0 [26] (http://rice.plantbiology.msu.edu/) gene models (ii) variant filtration using PLINK wrapper, (iii) general statistics such as Transition-Transversion ratio, levels of heterozygosity and missing data for each variety using VCFtools, (iv) SNP density analysis, (v) diversity indices calculation in sliding windows along the genome using VCFtools (Pi, Tajima's D, FST if subpopulations provided), (vi) linkage disequilibrium, (vii) population structure by sNMF (http://membres-timc.imag.fr/Olivier.Francois/snmf/index.htm), (viii) Principal Component Analysis and Identity By State (IBS) clustering of varieties by PLINK, and (ix) SNP-based distance phylogenetic tree by FastME (http://www.atgc-montpellier.fr/fastme/). Most key steps are decorated with sophisticated visualizations using a dedicated plugin. Visualization can be displayed by clicking on the [visualization] icon.

In practice, this workflow can be processed for many applications such as the identification of possible introgression events, the identification of putative genomic regions involved in the control of qualitative trait through a FST approach, the investigation for potential duplicates in the 3K RG accessions dataset and custom datasets, or the estimate of closest varieties of new sequenced accessions, by ranking a list of varieties from the database most closely matching the given sample. It can be used also for the close inspection of genomic region of interest after a GWAS analysis, through a linkage disequilibrium focus or the haplotyping of candidate genes.

Uniqprimer module

Uniqprimer is a workflow for comparative genomics-based diagnostic primer design, developed from a pipeline used in-house at Colorado State University to develop novel species and subspecies-level diagnostic tools for bacterial plant pathogens including pathovars of *Xanthomonas translucens* [27], geographical variants of rice-associated *Xanthomonas* spp. [28-30], and the genetically diverse rice pathogen *Pseudomonas fuscovaginae* [31]. Uniqprimer is now deployed in Rice Galaxy for user-friendly diagnostic primer design from draft or complete pathogen genomes. The user inputs multiple bacterial genomes from diagnostic target species as well as non-target species (i.e. “include” and “exclude” genome files), and the tool performs comparative alignment, primer design, and primer validation to output a list of primers that are specific to the target genomes (Figure 9). The uniqprimer standalone program is written in Python and is available at the Southgreen github repository (https://github.com/SouthGreenPlatform/Uniqprimer), along with the detailed documentation for developers and end-users.

3. Rice Galaxy OA: a Prototype for Open Access

IRRI, as a member center of the Consultative Group for International Agricultural Research (CGIAR, https://www.cgiar.org/) , complies with the CGIAR policy on Open Access and Open Data (https://www.cgiar.org/how-we-work/accountability/open-access/). In collaboration with Indiana University in the United States and National Institute of Advanced Industrial Science and Technology in Japan, and carried out through grants from the National Science Foundation (NSF) in the US and the MacArthur Foundation through the Research Data Alliance (RDA - https://www.rd-alliance.org/), the team undertook a prototyping effort to bring the Rice Galaxy system to maximum compliance with the CGIAR policy.

The basis for the design to add open access to Rice Galaxy is a foundational technical idea emerging from activities occurring in the international RDA. This idea acknowledges that for open data access to be broadly realized, all meaningful data objects must have a globally unique and persistent identifier

(PID). Globally unique means the name is not shared with other objects on a global scale. An identifier is persistent when the PID itself cannot be destroyed, and when the relationship between the identifier and the data object it points to is permanent. Through an international working group in RDA, a team of researchers is advancing the notion of PID Kernel Information, which injects a tiny amount of carefully selected metadata into a PID record. This technique has the potential to stimulate an entirely new ecosystem of third party services that can process the billions of expected PIDs. The key challenge of this working group is to determine which from amongst thousands of relevant metadata are suitable to embed in the PID record.

Our design draws on earlier work by us in data provenance capture and representation [32-34] and employs a hands-off technique (*data provenance capture*) to gather information about a researcher's rice genomics analysis as the analysis is running. Through this technique, information acquired while the analysis is running is compiled and combined with pre-analysis information that is available at the beginning of the analysis workflow. Such information includes who performed the analysis, when it was performed, and under what conditions.

There have been earlier approaches to capture provenance of Galaxy workflows. Geocks *et.al* [35] developed a history panel for users to facilitate reproducibility. Gaignard *et al*. [36] proposes the SHARP toolset, a semantic web (i.e. linked data) approach of harmonizing provenance collected from both the Galaxy and Taverna workflow systems. Kanwal *et al*. [37] captured the activity of a workflow (called a *provenance trace*) including the version of analysis tools run, the software parameters used, and the data objects produced at each workflow step. This work also targets increased reproducibility of past workflow instances. Missier *et al.* [38] proposes the "Golden Trail" architecture to describe and store workflow runtime provenance retrieved from Galaxy. The golden trail of provenance that is collected can be used to construct a virtual experiment view of past workflow runs. The four research contributions described further underline the need for the capture of provenance from workflow systems. They propose different but equally important uses of data provenance, that is, to facilitate the improvement of science through reproducibility and construction of virtual views of an experiment once it has completed.

Our design for Rice Galaxy Open Access (OA) shares similarities with these other techniques, however its end goal is different, which is to advance open access, hence making Rice Galaxy consistent with CGIAR's open access policy. To do this, we focus on each piece of data and information deemed valuable that emerges from workflow runs deemed to be of importance. This particular data and information must be retained and shared with others, while being subject to reasonable restrictions**.** This is a highly selective approach to provenance capture, and one that makes our work unique. We briefly outline the solution here and identify resources for those interested in pursuing the topic in more detail.

The architecture of Rice Galaxy OA (Figure 10A) utilizes the Handle system [39] and two standards emerging from the Research Data Alliance, RDA PID Type [40] and the Data Type Registry [41]. It additionally uses storage and compute resources provisioned through the NSF funded project, Pacific Rim Applications and Grid Middleware Assembly (PRAGMA).

A researcher interacts with the open access enhanced Rice Galaxy system as follows:

(1) Researcher performs an analysis in Rice Galaxy

(2) Data objects (input data, output data, information such as configuration parameters) are extracted from Rice Galaxy OA into a PRAGMA Data Repository Database (MongoDB) (Figure 10A),

(3) The data objects are assigned Persistent Identifiers, the PID Kernel Information is assigned into the PID record at this time, and a landing page created for each (Figure 10B).

(4) Data objects can be downloaded from the Data Identity server and re-loaded to the Rice Galaxy server for full faithful reproduction of the analysis

The resulting system appears to be promising and addresses a number of the recommendations from CGIAR. The Rice Galaxy OA system is a user transparent means of harvesting digital objects from applications and assigning PIDs to scientific outcomes. The architecture is modular and built with default PID information types and metadata using RDA products (Figure 10A). Although this proof-of-concept prototype successfully demonstrates the feasibility of this approach, there remains some future work. The community needs to provide feedback on which data and information products are most important to retain and make available. Additionally, not all workflow runs are important to a researcher as they could be system tests or new workflow tests. Thus, how a researcher identifies the items he/she wishes to make available to others and when, remains an important consideration for this system. For more information, points of contact to the team, the underlying software for Rice Galaxy OA, and the link to the prototype server can be found at https://github.com/Data-to-Insight-Center/RDA-PRAGMA-Data-Service/wiki/Welcome-to-PRAGMA-Data-Service-Prototype .

4. Rice Galaxy architecture discussion

We deployed IRRI Galaxy in an AWS EC2 instance (t2.large instance 2 vCPU, 4 GiB RAM) for the production server deployment in the cloud with Linux Ubuntu release 12.04.2 LTS (GNU/Linux 3.2.0-40-virtual x86_64) using Galaxy release 14 as described in the Galaxy documentation.

External data from the 3K RG Project files stored in the 3K RG AWS Simple Storage Service (S3) Public Data resource hosted at http://s3.amazonaws.com/3kricegenome/ (or s3:// 3kricegenome/) is accessed using AWS S3 Command Line Interface, a command line tool utility in AWS that provides an interface to access AWS S3 objects (CLI, https://docs.aws.amazon.com/cli/latest/reference/s3/). First, Rice Galaxy connects to the 3K RG AWS bucket using s3API and allows the objects inside the bucket to be transparent to Galaxy. VCF files (and the accompanying index files) are downloaded to Rice Galaxy using the S3 CLI with the `aws S3 cp` command, executed as:

```
      aws –profile user s3 cp
      s3://3kricegenome/REFERENCE/VCF_FILE.snp.vcf.gz* .
```

The subset region of the VCF file (chromosome:start-end) is then extracted using BCFtools (http://samtools.github.io/bcftools/) wrapped in Rice Galaxy and exported to the history pane as bgzipped, indexed BCF file, which can then be converted back to VCF using [VCFTOOLS] in Rice Galaxy.

Standard methods for tool wrapper development and deployment were followed. All tool wrapper XMLs developed specifically for Rice Galaxy are deposited and shared in a project-specific Rice Galaxy toolshed repository at http://52.76.88.51:8081/ (Figure 11) and will also be deposited in the central Galaxy toolshed (https://toolshed.g2.bx.psu.edu/). All developments and testing of Rice Galaxy and Rice Galaxy Toolshed were done in Docker containers hosted in virtual machines at the Advanced Science and Technology Institute, Department of Science and Technology of the Philippine Government (ASTI – DOST) prior to final deployment to the AWS instance.

This resource will empower the rice research community to benefit from publicly available datasets (e.g. 3K RG) and materials (seed/accessions), to enhance or even drive their own respective institutional genetic/genomic/breeding efforts. The Rice Galaxy instance (data, tools, computing resources) is free for use by all.

In addition to the integration of these tools, new Galaxy wrappers and visualization plugins are being developed for visualizing chromosomes and their information (SNP density, structural variants, translocations) either in linear or circular mode, using recent web technologies (Ideogram.js [42] , BioCircos.js [43], respectively).

Finally, a Docker container of Rice Galaxy is under development so that it can be easily shared and deployed through the Galaxy Docker flavor initiative (https://github.com/bgruening/docker-galaxy-stable).

Conclusion

Rice Galaxy is a federated Galaxy resource specialized for rice genetics, genomics, and breeding. The resource empowers the rice research community to utilize publicly available datasets (3K RG), materials (seed/accessions), and their own data, allowing complex data analyses to be performed even without investment in their own computational infrastructure and software development team. Rice research – related tools are also hosted in Rice Galaxy (i.e. Uniqprimer rice pathogen diagnostic design).

Rice Galaxy is freely accessible to all and we invite the rice research community to participate in enriching the tools hosted by the resource. It can serve as a repository for data, analyses results, and new bioinformatics tools coming from institutions that have used the publicly available rice diversity panels from 3K RG, or have developed rice genomic/genetic analyses tools that they wish to share to the community, and a computing infrastructure for small institutes without in-house computing capability.

Availability and requirements

Project name: RICE GALAXY

Project home page: https://github.com/InternationalRiceResearchInstitute/RiceGalaxy

Operating system(s): Linux Ubuntu release 12.04.2 LTS

Programming language: Python

Other requirements: R release 3.2.3 and following packages: methods, fpc, cluster, vegan, pheatmap, pROC, randomForest, miscTools, pRF, e1076, rrBLUP, glmnet ;TASSEL release 5.2.40; plink v1.90b3k; JBrowse 1.14.1; snpEff 4.3T; sNMF 1.2 (and as R package LEA); FastME 2.0

License: Rice Galaxy tools are released under GNU GPL. All software from external sources is bound by their respective licenses.

Any restrictions to use by non-academics: Rice Galaxy tools are without restriction to non-academics. All software from external sources is bound by their respective non-academic restrictions

Code availability: Tool wrappers at Rice Galaxy Toolshed (http://52.76.88.51:8081/). Rice Galaxy is available at IRRI Github (https://github.com/InternationalRiceResearchInstitute/RiceGalaxy).

Availability of supporting data

3,000 Rice Genomes Project at Gigascience database (http://gigadb.org/dataset/200001)

3K RG BAM and VCF files available from Amazon Public data and ASTI-DOST IRODs site, instructions at http://iric.irri.org/resources/3000-genomes-project .

SNP sets and morpho-agronomic characterization from 3K RG at SNP-Seek download site (http://snp-seek.irri.org/_download.zul)

Availability of supporting source code and requirements

Project name: Uniqprimer

Project home page: https://github.com/SouthGreenPlatform/Uniqprimer

Operating system(s): Linux OS

Programming Language: Python

Other requirements: MUMmer 3

License: GNU GPL

Project name: PRAGMA Data Service

Project home page: repository https://github.com/Data-to-Insight-Center/RDA-PRAGMA-Data-Service/wiki/Welcome-to-PRAGMA-Data-Service-Prototype

Operating system(s): Platform independent

License: Apache License 2.0

Declarations

Abbreviations

3K RG:3,000 Rice Genomes;HDRA: High Density Rice Array; SNP: single nucleotide polymorphism; GWAS: Genome-Wide Association Studies; RAM:random access memory; CPU: central processing unit; IRRI:International Rice Research Institute;NGS: next-generation sequencing;QTL:quantitative trait loci; IRGSP:International Rice Genome Sequencing Project; RGAP:Rice Genome Annotation Project;KASP:

Kompetitive Allele Specific PCR; VCF : variant call format ;AWS: Amazon Web Services; RAVE: Rapid Allelic Variant extractor; MAF: minor allele frequency;TASSEL: Trait Analysis by aSSociation, Evolution and Linkage; GS:genomic selection;Oghma: Operators for Genome decipHering by MAchine learning; rrBLUP:ridge regression best linear unbiased predictor; SVM:support vector machine; FST:fixation index; NSF:National Science Foundation;CGIAR: Consultative Group for International Agricultural Research; RDA: Research Data Alliance;PID:persistent identifier;OA:open access; PRAGMA:Pacific Rim Applications and Grid Middleware Assembly; EC2:elastic computing cloud; CLI: command line interface; S3: Simple Storage Service; API: Python Application Programming Interfaces; XML: eXtensible Markup Language;

Competing interests

The author(s) declare that they have no competing interests.

Funding

Components of the project are supported by the following grants: Taiwan Council of Agriculture Grant to IRRI, International Rice Informatics Consortium, and CGIAR Excellence in Breeding Platform for financial support to the Rice Galaxy main server, the USA National Science Foundation PRAGMA grant number: NSF OCI 1234983, the RDA/US-sponsored adoption program funded by the MacArthur Foundation, and the AIST ICT International Collaboration Grant.

Authors’ contributions

VJ and AD equally contributed to create Rice Galaxy. NB contributed the genomic prediction tools. AD, GD, PL, and MR contributed the RAVE and SNiPLAY tools. JD, JRM, JPP created the development Rice Galaxy cloud instances hosted at DOST-ASTI. LM created the SNP-Seek interfaces. LT, JL, JEL contributed the Uniqprimer tool. GZ, KR, BP, and JH contributed the Rice Galaxy Open Access, MT, NA, and TK contributed to funding acquisition and writing, RM coordinated the conceptualization of the project and the writing process.

Acknowledgments

The authors are grateful to the following people and institutions/agencies for their support: DOST-ASTI for hosting the Rice Galaxy toolshed server, Jay Santos and Denis Diaz for assistance with AWS architecture.

References

1. 3,000 rice genomes project. The 3,000 rice genomes project. Gigascience. 2014;3:7.
2. Wang, W-S, Mauleon R, Chebotarov, D, et al. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature. 2018;557: 43–49 .doi:10.1038/s41586-018-0063-9.
3. McCouch S, Wright M, Tung C-W, Maron L, McNally K, Fitzgerald M, et al. Open Access Resources for Genome Wide Association Mapping in Rice. Nature Comm. 2016;7: 10532, doi 10.1038/ncomms10532.
4. Alexandrov N, Tai S, Wang W, Mansueto L, Palis K, Fuentes RR, et al. SNP-Seek database of SNPs derived from 3000 rice genomes. Nucleic Acids Res. 2015;63:2–6.
5. Mansueto L, Fuentes RR, Chebotarov D, Borja FN, Detras J, Abriol-Santos JM, et al. SNP-Seek II: A resource for allele mining and analysis of big genomic data in *Oryza sativa*. Curr. Plant Biol. 2016;6628:16–25.
6. Sempéré G, Philippe F, Dereeper A, Ruiz M, Sarah G, Larmande P. Gigwa-Genotype investigator for genome-wide analyses. Gigascience. 2016;5:25.
7. The South Green Collaborators. The South Green portal: a comprehensive resource for tropical and Mediterranean crop genomics. Curr Plant Biol. Elsevier. 2016;7–8: 6–9. doi:10.1016/J.CPB.2016.12.002.
8. Tello-Ruiz MK, Naithani S, Stein JC, Gupta P, Campbell M, et al. Gramene 2018: unifying comparative genomics and pathway resources for plant research. Nucleic Acids Res. 2017; PMID: 29165610. doi: 10.1093/nar/gkx1111.

9. Afgan, E, Baker D. van den Beek M, Blankenberg D, Bouvier D, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Research. 2016;44(W1): W3-W10 doi:10.1093/nar/gkw343.

10. Kawahara, T., et al. Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. Rice. 2013;6:4 .

11. Zhang J, Chen L-L, Xing F, Kudrna DA, Yao W, Copetti D, et al. Extensive sequence divergence between the reference genomes of two elite *indica* rice varieties Zhenshan 97 and Minghui 63. Proc Natl Acad Sci U S A. 2016;113: E5163–71.

12. Du, H, Yu Y, Ma Y, Gao Q, et al. Sequencing and *de novo* assembly of a near complete *indica* rice genome. Nature Communications. 2017;8 (15324). doi:10.1038/ncomms15324.

13. Schatz, M., et al. Whole genome *de novo* assemblies of three divergent strains of rice, *Oryza sativa*, document novel gene space of *aus* and *indica*. Genome Biology. 2014;15, 506.

14. Gao, Z.Y., et al. Dissecting yield-associated loci in super hybrid rice by resequencing recombinant inbred lines and improving parental genome sequences. PNAS. 2013; 110 (35), 14492-14497.

15. Sakai, H., et al. Construction of pseudomolecule sequences of the aus rice cultivar Kasalath for comparative genomics of Asian cultivated rice. DNA Research. 2014; do:10.1093/dnares/dsu006.

16. Xu K, Xu X, Fukao T, Canlas P, Maghirang-Rodriguez R, Heuer S, et al. *Sub1A* is an ethylene-response-factor-like gene that confers submergence tolerance to rice. Nature. 2006;442: 705–708.

17. Gamuyao R, Chin JH, Pariasca-Tanaka J, Pesaresi P, Catausan S, Dalid C, et al. The protein kinase *Pstol1* from traditional rice confers tolerance of phosphorus deficiency. Nature. 2012;488: 535–539.

18. Uga Y, Sugimoto K, Ogawa S, Rane J, Ishitani M, Hara N, et al. Control of root system architecture by *DEEPER ROOTING 1* increases rice yield under drought conditions. Nat Genet. 2013;45: 1097–1102.

19. Thomson MJ, Singh N, Dwiyanti MS, Wang DR, Wright MH, et al. Large-scale deployment of a rice 6 K SNP array for genetics and breeding applications. Rice. 2017;10:40 doi:10.1186/s12284-017-0181-2.

20. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, et al.PLINK: a toolset for whole-genome association and population-based linkage analysis". American Journal of Human Genetics.2007. 81: 559–75. doi:10.1086/519795.

21. Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH. JBrowse: a next-generation genome browser. Genome Res. 2009;19:1630–8.

22. Dereeper A, Homa F, Andres G, Sempere G, Sarah G, Hueber Y, et al. SNiPlay3: a web-based application for exploration and large scale analyses of genomic variations. Nucleic Acids Res. 2015;43:W295-300.

23. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics.2007; 23:2633-2635.

24. Spindel J, Begum H, Akdemir D, Virk P, Collard B, Redona E, Atlin G, Jannink JL, McCouch SR. Genomic selection and association mapping in rice (*Oryza sativa*): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. PLOS Genetics.2015; 11(2): e1004982. doi:10.1371/journal.pgen.1004982.

25. Browning BL, Browning SR. Genotype imputation with millions of reference samples. Am J Hum Genet. 2016. 98:116-126. doi:10.1016/j.ajhg.2015.11.020.

26. Rice Genome Annotation Project (RGAP) release 7. 2013; http://rice.plantbiology.msu.edu/. Accessed 3 May 2018.

27. Langlois, PA, Snelling J, Hamilton JP, Bragard C, Koebnik R, Verdier V, et al. Characterization of the *Xanthomonas translucens* complex using draft genomes, comparative genomics, phylogenetic analysis, and diagnostic LAMP assays. Phytopathology. 2017; 107: 519-527.

28. Triplett, L, Hamilton JP, Buell CR, Tisserat NA, Verdier V, Zink F, Leach JE. Genomic Analysis of *Xanthomonas oryzae* from US rice reveals substantial divergence from known *X. oryzae* pathovars. Appl. Environ. Microbiol. 2011.;77(12):3930-3937. doi:10.1128/AEM.00028-11.

29. Lang, JM, Langlois P, Nguyen MHR, Triplett LR, Purdie L, et al. Sensitive detection of *Xanthomonas oryzae pv. oryzae* and *X. oryzae pv. oryzicola* by Loop-Mediated Isothermal Amplification. Applied and Environmental Microbiology. 2014; 80:4519-4530.

30. Triplett L, Verdier V, Campillo T, Van Malderghem C, et al. Characterization of a novel clade of *Xanthomonas* isolated from rice leaves in Mali and proposal of *Xanthomonas maliensis* sp. nov. 2015;Antonie van Leeuwenhoek 107:869-81. doi: 10.1007/s10482-015-0379-5. http://link.springer.com/journal/10482/onlineFirst/page/1.

31. Ash GJ, Lang JM, Triplett LR, Stodart BJ, Verdier V, et al. Development of a genomics-based LAMP (Loop-1 mediated isothermal amplification) assay for detection of *Pseudomonas fuscovaginae* from rice. 2014.;Plant Dis 98:909-915 doi.org/10.1094/PDIS-09-13-0957-RE .

32. Yogesh LS, Plale B, Gannon D. A survey of data provenance in e-science. ACM Sigmod Record . 2005;34.3, p. 31-36.

33. Zhou Q, Ghoshal D ,Plale B. Study in Usefulness of Middleware-Only Provenance. 2014 IEEE 10th International Conference on e-Science, Sao Paulo. 2014; pp. 215-222. doi:10.1109/eScience.2014.49.

34. Suriarachchi I, Zhou Q, Plale B, Komadu. A Capture and Visualization System for Scientific Data Provenance . Journal of Open Research Software. 2015;3 p . e4,. doi:10.5334/jors.bq.

35. Goecks J, Nekrutenko A, Taylor J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome biology. 2010; 11(8):R86.

36. Gaignard A, Belhajjame K, Skaf-Molli H. Sharp: Harmonizing and bridging cross-workflow provenance. In: Blomqvist E, Hose K, Paulheim H, Lawrynowicz A, Ciravegna F, Hartig O, editors. The Semantic Web: ESWC 2017 Satellite Events. 2017.; p. 219-234, Cham. Springer International Publishing.

37. Kanwal S, Zaib Khan F, Lonie A, Sinnott RO. Investigating reproducibility and tracking provenance - a genomic workflow case study. BMC bioinformatics. 2017; 18(1):337.p. 25.

38. Missier P, Ludascher B, Dey S, Wang M, McPhillips T, et al. Golden trail: Retrieving the data history that matters from a comprehensive provenance repository. International Journal of Digital Curation. 2012; 7(1):139-150.

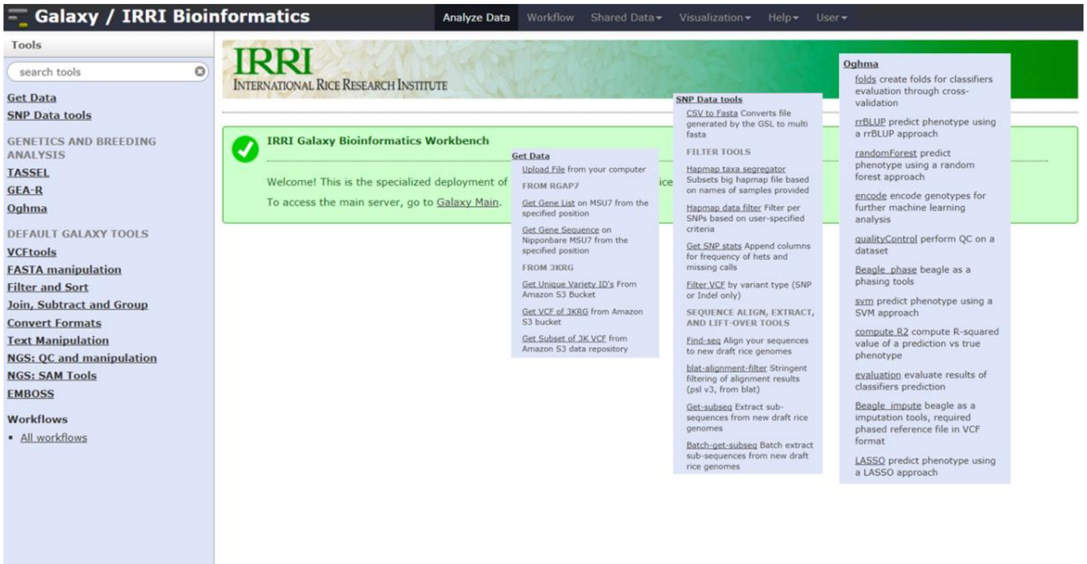
39. Kahn R, Wilensky R. A Framework for Distributed Digital Object Services. Int. J. Digit. Libr. 2006; 6, 2: 115–123. doi:10. 1007/s00799- 005- 0128- x.

40. Research Data Alliance PID Kernel Information Working Group. PID Kernel Information guiding principles. 2018; https://www.rd-alliance.org/group/ pid- kernel- information- wg/wiki/pid-kernel- information- guiding- principles.). Accessed 15 May-2018.

41. Research Data Alliance Data Type Registry Working Group. RDA Data Type Registries Working Group Output. 2016; doi:10.15497/ A5BCD108-ECC4-41BE-91A7-20112FF77458. Accessed 15 May 2018.

42. Dereeper A, Bocs S, Rouard M, Guignon V, Ravel S, Tranchant-Dubreuil C, et al. The coffee genome hub: a resource for coffee genomes. Nucleic Acids Res. 2015; 43:D1028-35.

43. Cui Y, Chen X, Luo H, Fan Z, Luo J, He S, et al. BioCircos.js: an interactive Circos JavaScript library for biological data visualization on web applications. Bioinformatics. 2016; 32(11):1740-2. doi:10.1093/bioinformatics/btw041.



Get Data
SNP Data tools

GENETICS AND BREEDING ANALYSIS
TASSEL
GEA-R
Oghma

DEFAULT GALAXY TOOLS
VCFtools
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Text Manipulation
NGS: QC and manipulation
NGS: SAM Tools
EMBOSS

Workflows
- All workflows

IRRI Galaxy Bioinformatics Workbench

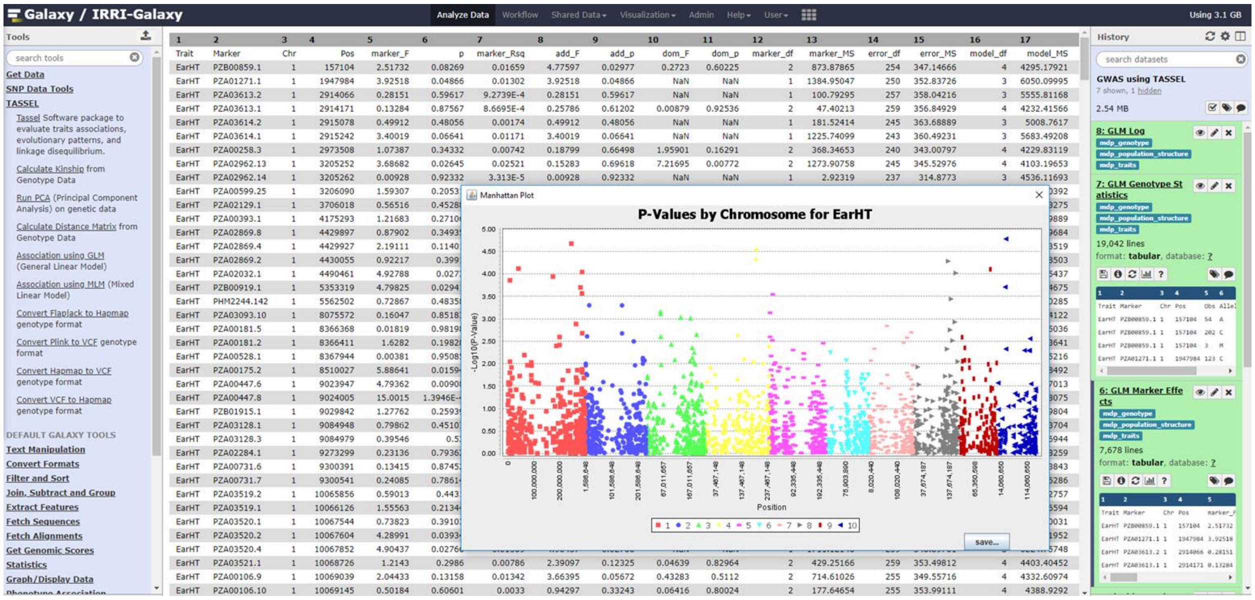
Welcome! This is the specialized deployment of … ice
To access the main server, go to Galaxy Main.

Get Data
Upload File from your computer
FROM RGAP7
Get Gene List on MSU7 from the specified position
Get Gene Sequence on Nipponbare MSU7 from the specified position
FROM 3KRG
Get Unique Variety ID's From Amazon S3 Bucket
Get VCF of 3KRG from Amazon S3 bucket
Get Subset of 3K VCF from Amazon S3 data repository

SNP Data tools
CSV to Fasta Converts file generated by the GSL to multi fasta
FILTER TOOLS
Hapmap taxa segregator Subsets big hapmap file based on names of samples provided
Hapmap data filter Filter per SNPs based on user-specified criteria
Get SNP stats Append columns for frequency of hets and missing calls
Filter VCF by variant type (SNP or Indel only)
SEQUENCE ALIGN, EXTRACT, AND LIFT-OVER TOOLS
Find-seq Align your sequences to new draft rice genomes
blat-alignment-filter Stringent filtering of alignment results (psl v3, from blat)
Get-subseq Extract sub-sequences from new draft rice genomes
Batch-get-subseq Batch extract sub-sequences from new draft rice genomes

Oghma
folds create folds for classifiers evaluation through cross-validation
rrBLUP predict phenotype using a rrBLUP approach
randomForest predict phenotype using a random forest approach
encode encode genotypes for further machine learning analysis
qualityControl perform QC on a dataset
Beagle_phase beagle as a phasing tools
svm predict phenotype using a SVM approach
compute R2 compute R-squared value of a prediction vs true phenotype
evaluation evaluate results of classifiers prediction
Beagle_impute beagle as a imputation tools, required phased reference file in VCF format
LASSO predict phenotype using a LASSO approach

Figure 1. Rice Galaxy @ IRRI with customized analyses tools for genetics, breeding, and custom data sources (i.e. 3,000 Rice Genomes project).



Get Data
SNP Data Tools
TASSEL
Tassel Software package to evaluate traits associations, evolutionary patterns, and linkage disequilibrium.
Calculate Kinship from Genotype Data
Run PCA (Principal Component Analysis) on genetic data
Calculate Distance Matrix from Genotype Data
Association using GLM (General Linear Model)
Association using MLM (Mixed Linear Model)
Convert Flapjack to Hapmap genotype format
Convert Plink to VCF genotype format
Convert Hapmap to VCF genotype format
Convert VCF to Hapmap genotype format

DEFAULT GALAXY TOOLS
Text Manipulation
Convert Formats
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Statistics
Graph/Display Data
Phenotype Association

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Trait	Marker	Chr	Pos	marker_F	p	marker_Rsq	add_F	add_p	dom_F	dom_p	marker_df	marker_MS	error_df	error_MS	model_df	model_MS
EarHT	PZB00859.1	1	157104	2.51732	0.08269	0.01659	4.77597	0.02977	0.2723	0.60225	2	873.87865	254	347.14666	4	4295.17921
EarHT	PZA01271.1	1	1947984	3.92518	0.04866	0.01302	3.92518	0.04866	NaN	NaN	1	1384.95047	250	352.83726	3	6050.09995
EarHT	PZA03613.2	1	2914066	0.28151	0.59617	9.2739E-4	0.28151	0.59617	NaN	NaN	1	100.79295	257	358.04216	3	5555.81168
EarHT	PZA03613.1	1	2914171	0.13284	0.87567	8.6695E-4	0.25786	0.61202	0.00879	0.92536	2	47.40213	259	356.84929	4	4232.41566
EarHT	PZA03614.2	1	2915078	0.49912	0.48056	0.00174	0.49912	0.48056	NaN	NaN	1	181.52414	245	363.68889	3	5008.7617
EarHT	PZA03614.1	1	2915242	3.40019	0.06641	0.01171	3.40019	0.06641	NaN	NaN	1	1225.74099	243	360.49231	3	5683.49208
EarHT	PZA00258.3	1	2973508	1.07387	0.34332	0.00742	0.18799	0.66498	1.95901	0.16291	2	368.34653	240	343.00797	4	4229.83119
EarHT	PZA02962.13	1	3205252	3.68682	0.02645	0.02521	0.15283	0.69618	7.21695	0.00772	2	1273.90758	245	345.52976	4	4103.19653
EarHT	PZA02962.14	1	3205262	0.00928	0.92332	3.313E-5	0.00928	0.92332	NaN	NaN	1	2.92319	237	314.8773	3	4536.11693
EarHT	PZA00599.25	1	3206090	1.59307	0.2053											0392
EarHT	PZA02129.1	1	3706018	0.56516	0.4528											3275
EarHT	PZA00393.1	1	4175293	1.21683	0.2710											9889
EarHT	PZA02869.8	1	4429897	0.87902	0.3493											9684
EarHT	PZA02869.4	1	4429927	2.19111	0.1140											3519
EarHT	PZA02869.2	1	4430055	0.92217	0.399											3503
EarHT	PZA02032.1	1	4490461	4.92788	0.027											5437
EarHT	PZB00919.1	1	5353319	4.79825	0.0294											4675
EarHT	PHM2244.142	1	5562502	0.72867	0.4835											0285
EarHT	PZA03093.10	1	8075572	0.16047	0.8518											4122
EarHT	PZA00181.5	1	8366368	0.01819	0.9819											5036
EarHT	PZA00181.2	1	8366411	1.6282	0.1982											3641
EarHT	PZA00528.1	1	8367944	0.00381	0.9508											5216
EarHT	PZA00175.2	1	8510027	5.88641	0.0159											3492
EarHT	PZA00447.6	1	9023947	4.79362	0.0090											7013
EarHT	PZA00447.8	1	9024005	15.0015	1.3946E-											3075
EarHT	PZB01915.1	1	9029842	1.27762	0.2593											9804
EarHT	PZA03128.1	1	9084948	0.79862	0.4510											3704
EarHT	PZA03128.3	1	9084979	0.39546	0.5											5944
EarHT	PZA02284.1	1	9273299	0.23136	0.7936											3259
EarHT	PZA00731.6	1	9300391	0.13415	0.8745											3843
EarHT	PZA00731.7	1	9300541	0.24085	0.7861											5286
EarHT	PZA03519.2	1	10065856	0.59013	0.443											2757
EarHT	PZA03519.1	1	10066126	1.55563	0.2134											5594
EarHT	PZA03520.1	1	10067544	0.73823	0.3910											0031
EarHT	PZA03520.2	1	10067604	4.28991	0.0393											1952
EarHT	PZA03520.4	1	10067852	4.90437	0.0276											5748
EarHT	PZA03521.1	1	10068726	1.2143	0.2986	0.00786	2.39097	0.12325	0.04639	0.82964	2	429.25166	259	353.49812	4	4403.40452
EarHT	PZA00106.9	1	10069039	2.04433	0.13158	0.01342	3.66395	0.05672	0.43283	0.5112	2	714.61026	255	349.55716	4	4332.60974
EarHT	PZA00106.10	1	10069145	0.50184	0.60601	0.0033	0.94297	0.33243	0.06416	0.80024	2	177.64654	255	353.99111	4	4388.9292

History
search datasets
GWAS using TASSEL
7 shown, 1 hidden
2.54 MB

8: GLM Log
mdp_genotype
mdp_population_structure
mdp_traits

7: GLM Genotype Statistics
mdp_genotype
mdp_population_structure
mdp_traits
19,042 lines
format: tabular, database: ?

```
1     2          3  4        5    6
Trait Marker     Chr Pos     Obs  Allel
EarHT PZB00859.1 1   157104  54   A
EarHT PZB00859.1 1   157104  202  C
EarHT PZB00859.1 1   157104  3    M
EarHT PZA01271.1 1   1947984 123  C
```

6: GLM Marker Effects
mdp_genotype
mdp_population_structure
mdp_traits
7,678 lines
format: tabular, database: ?

```
1     2          3   4       5
Trait Marker     Chr Pos     marker_F
EarHT PZB00859.1 1   157104  2.51732
EarHT PZA01271.1 1   1947984 3.92518
EarHT PZA03613.2 1   2914066 0.28151
EarHT PZA03613.1 1   2914171 0.13284
```

Figure 2. Genome-Wide Association Studies analysis (implemented by TASSEL software) in Rice Galaxy.

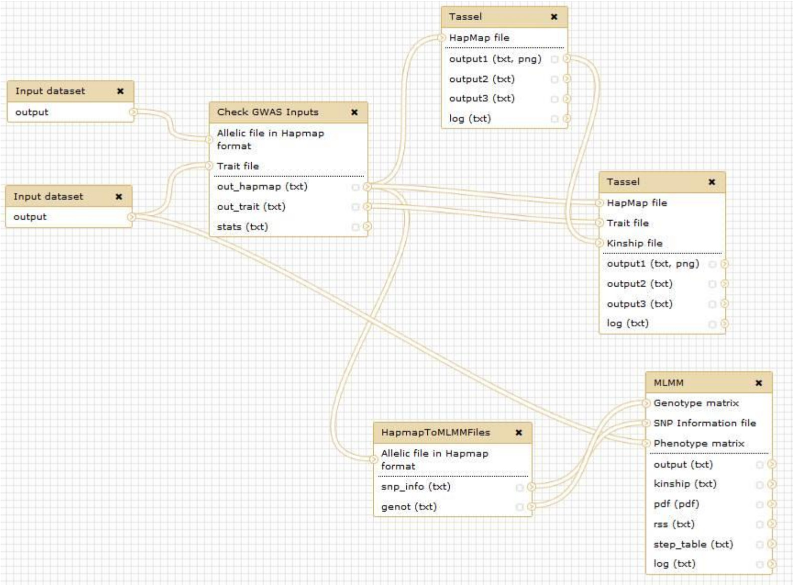
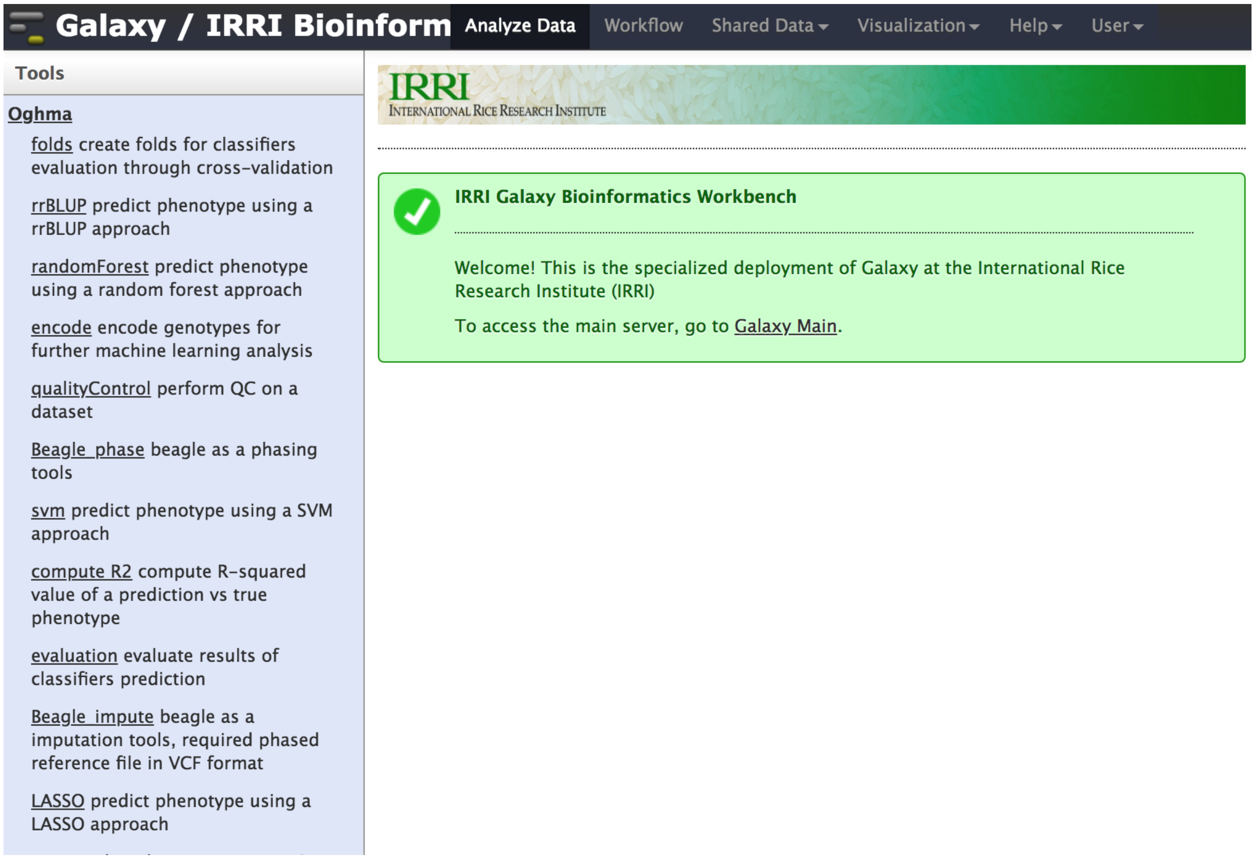


Figure 3. Genome-Wide Association Studies analysis workflow in SNiPlay as implemented in Rice Galaxy.



Tools

Oghma

folds create folds for classifiers evaluation through cross-validation

rrBLUP predict phenotype using a rrBLUP approach

randomForest predict phenotype using a random forest approach

encode encode genotypes for further machine learning analysis

qualityControl perform QC on a dataset

Beagle_phase beagle as a phasing tools

svm predict phenotype using a SVM approach

compute R2 compute R-squared value of a prediction vs true phenotype

evaluation evaluate results of classifiers prediction

Beagle_impute beagle as a imputation tools, required phased reference file in VCF format

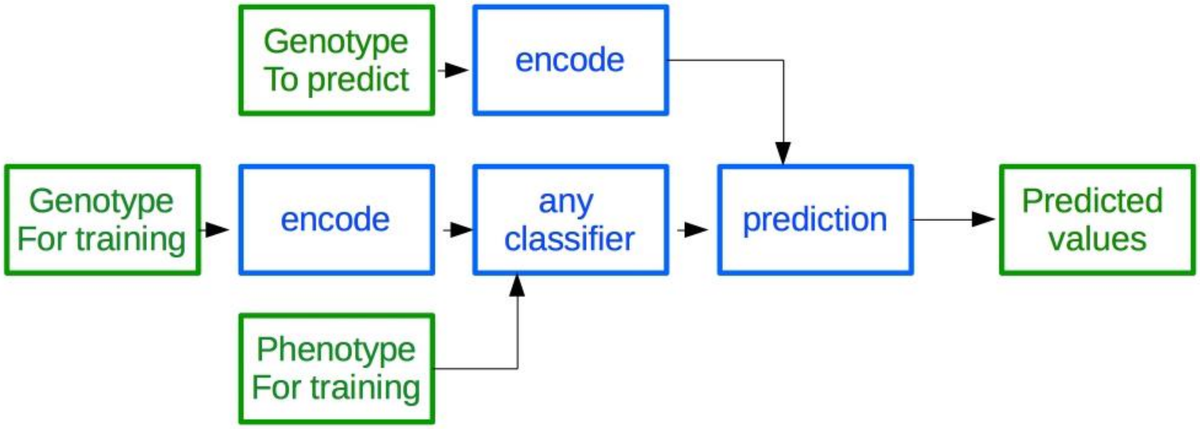
LASSO predict phenotype using a LASSO approach

IRRI Galaxy Bioinformatics Workbench

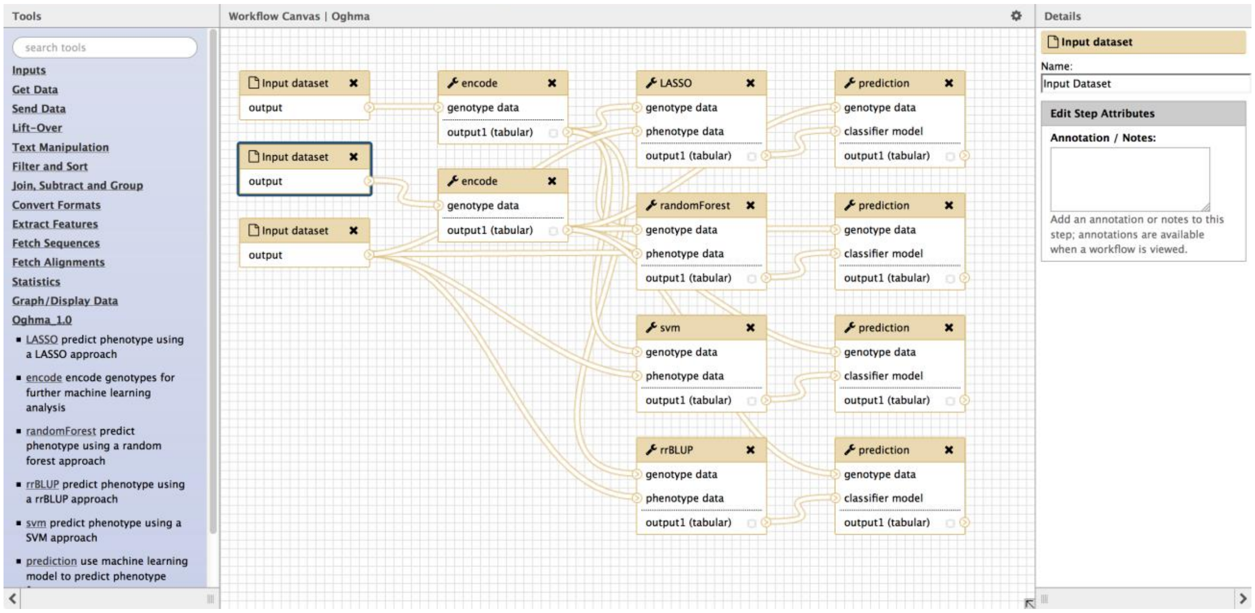
Welcome! This is the specialized deployment of Galaxy at the International Rice Research Institute (IRRI)

To access the main server, go to Galaxy Main.

Figure 4. Oghma genomic prediction and selection tools in Rice Galaxy with various classifier tools installed.



A. Overview of the Genomic Selection analyses workflow as implemented in Oghma tool suite.



Name: Input Dataset

Annotation / Notes:

Add an annotation or notes to this step; annotations are available when a workflow is viewed.

B. Rice Galaxy workflow for genome prediction using Oghma tool suite.

Figure 5: Genomic Selection analyses workflow as implemented by Oghma tool suite.

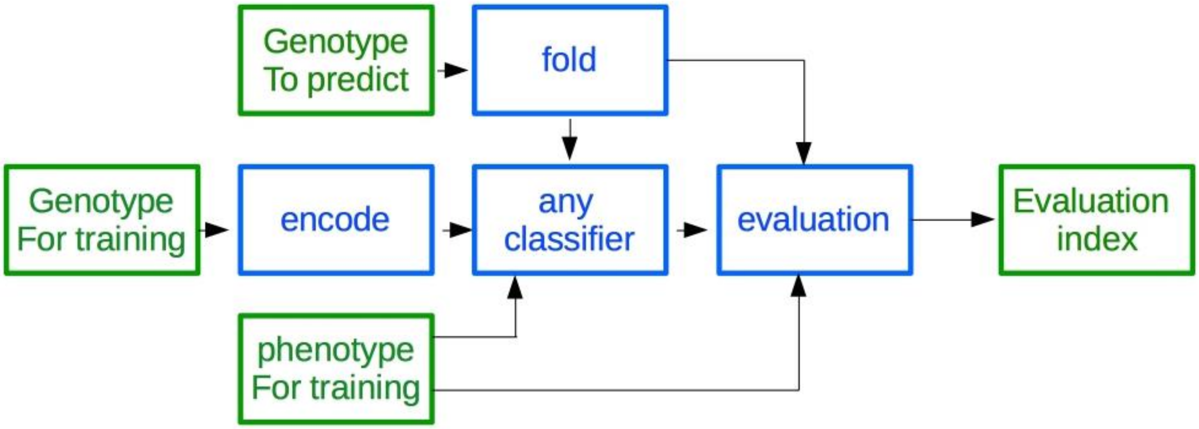


Figure 6. Workflow for classifier evaluation in the genome prediction tool suite implemented by Oghma.

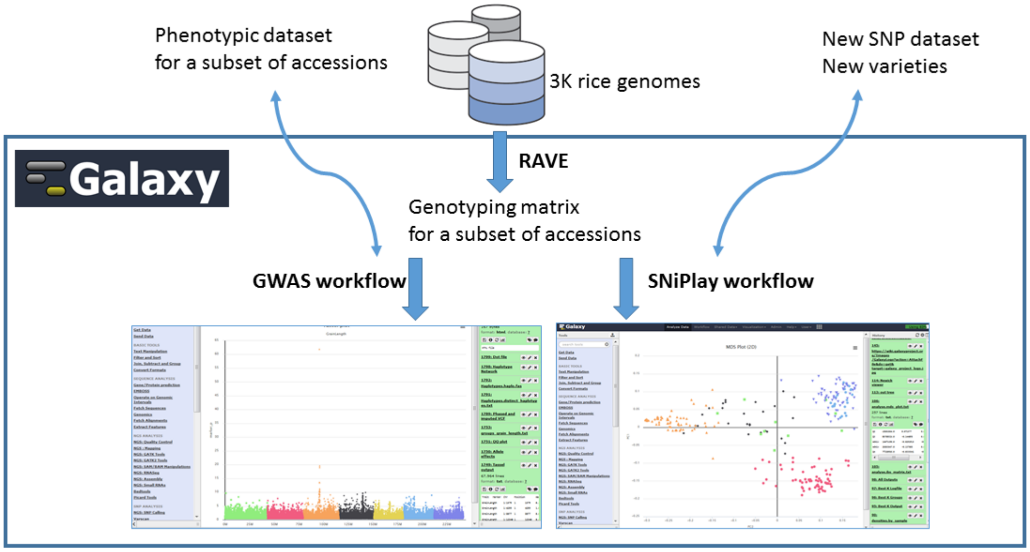


Figure 7. Overview schematic showing the integration of the 3K Rice Genomes project genotyping database and rapid extraction of subset SNPs by RAVE module for use by analyses workflows installed in Rice Galaxy.

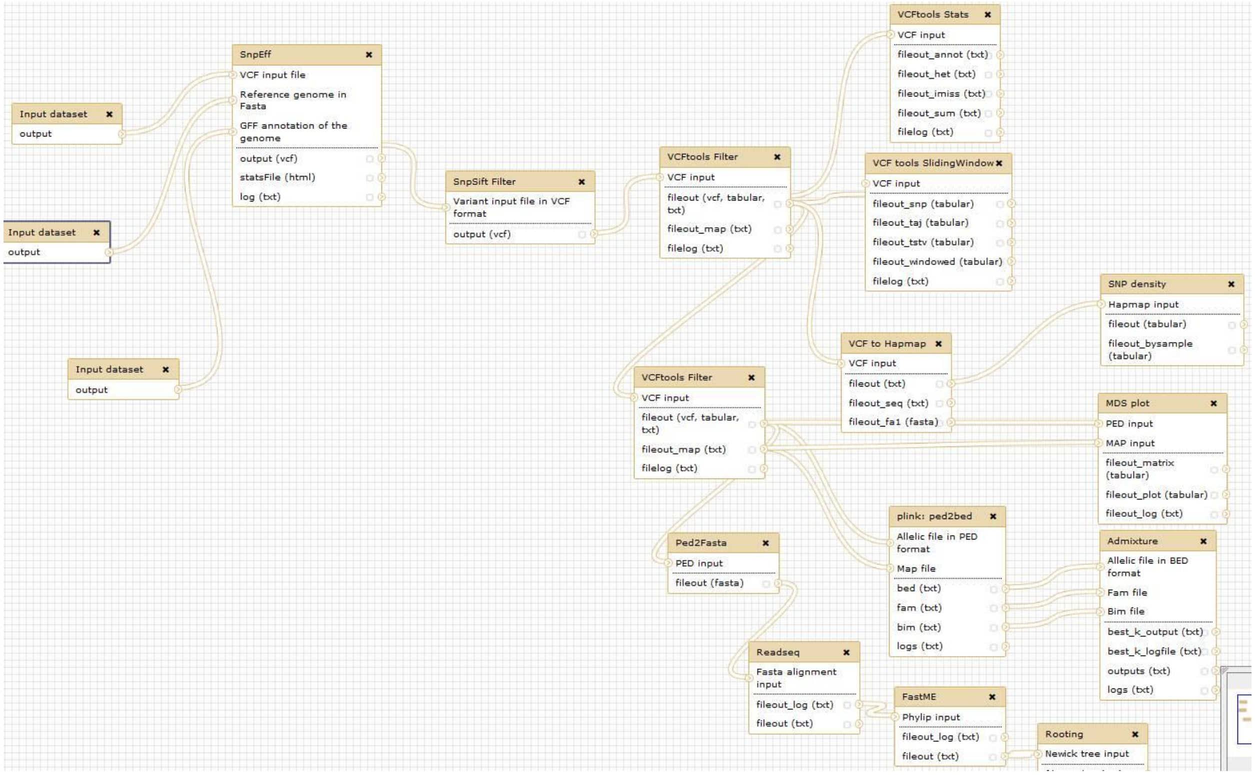
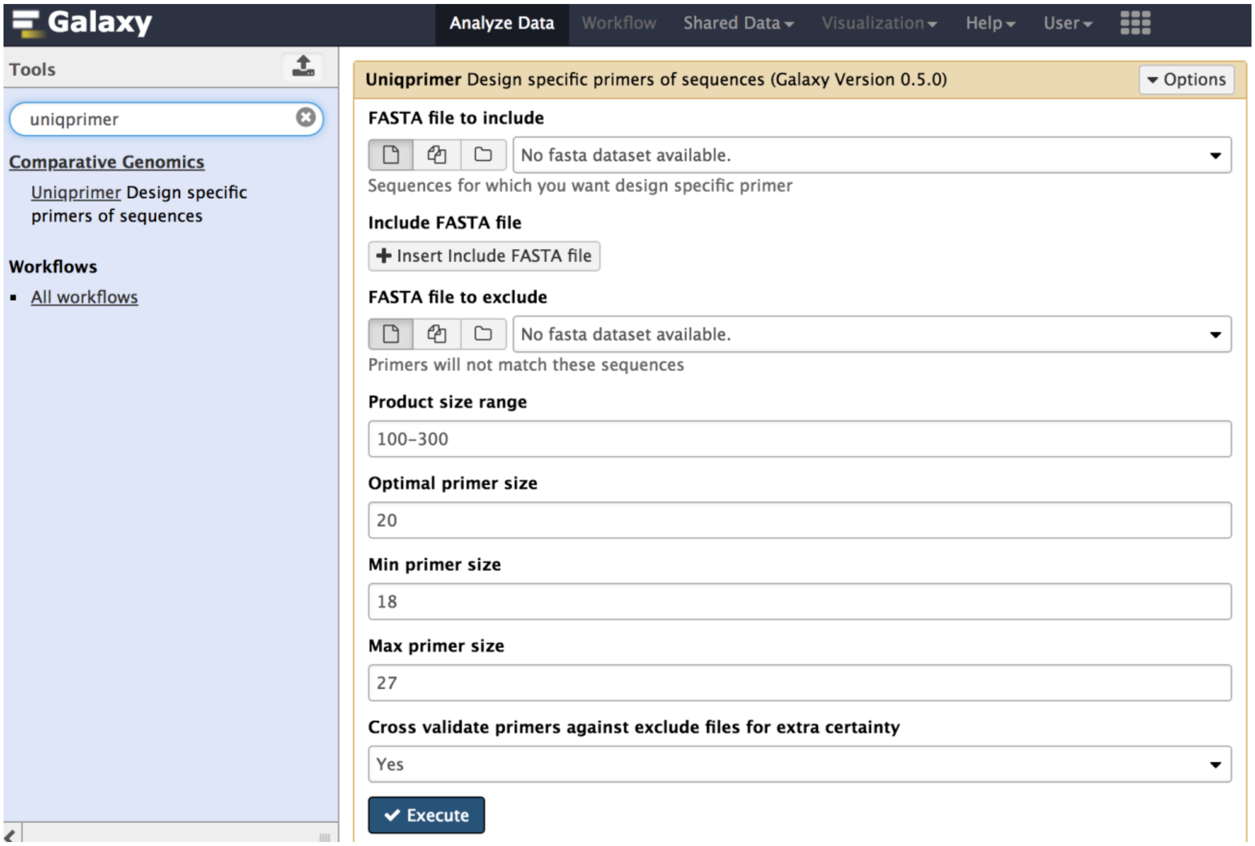


Figure 8. Rice Galaxy SNiPlay workflow for diversity and population structure analyses using various software tools.



uniqprimer

Uniqprimer Design specific primers of sequences

Uniqprimer Design specific primers of sequences (Galaxy Version 0.5.0)

FASTA file to include

No fasta dataset available.

Sequences for which you want design specific primer

Include FASTA file

FASTA file to exclude

No fasta dataset available.

Primers will not match these sequences

Product size range

100-300

Optimal primer size

20

Min primer size

18

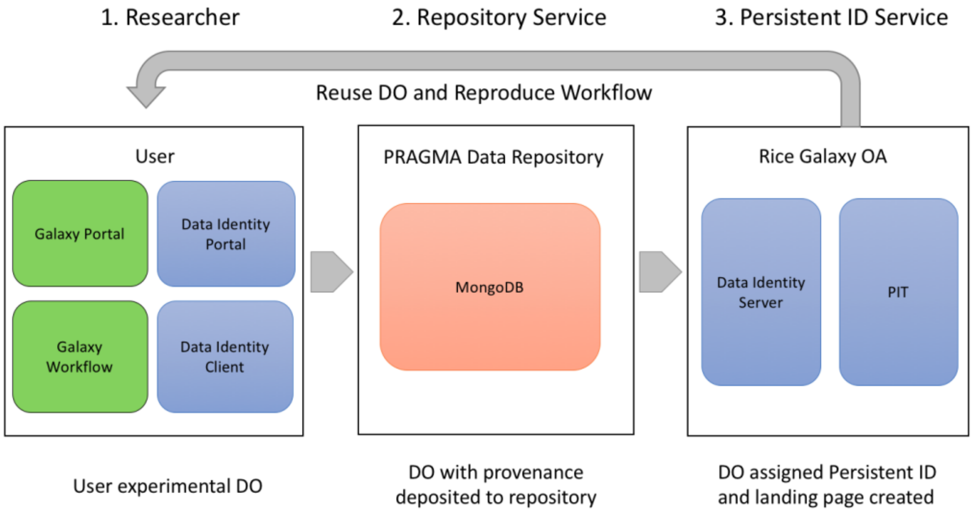
Max primer size

27

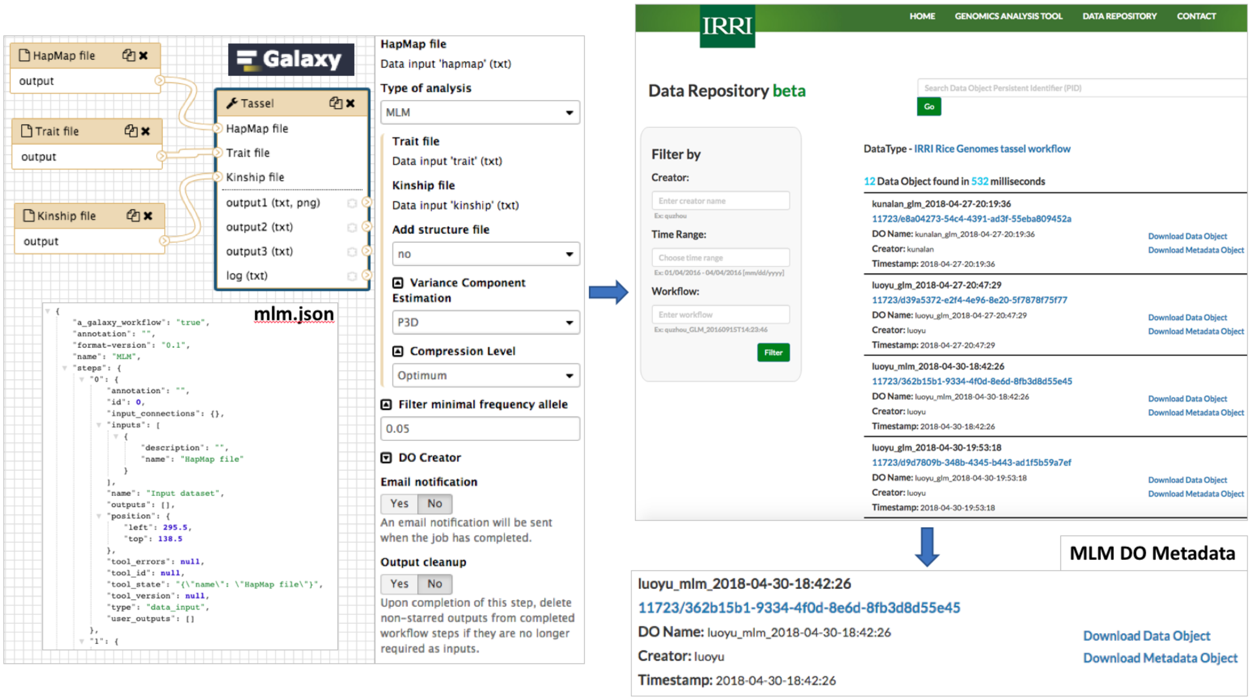
Cross validate primers against exclude files for extra certainty

Yes

Figure 9. Uniqprimer comparative genomics-based diagnostic primer design tool for microbial pathogen detection installed in Rice Galaxy.



A. The underlying software infrastructure for the components of Rice Galaxy Open Access.



HapMap file
Data input 'hapmap' (txt)

Type of analysis
MLM

Trait file
Data input 'trait' (txt)

Kinship file
Data input 'kinship' (txt)

Add structure file
no

⊟ **Variance Component Estimation**
P3D

⊟ **Compression Level**
Optimum

⊟ **Filter minimal frequency allele**
0.05

⊟ **DO Creator**

Email notification
Yes No
An email notification will be sent when the job has completed.

Output cleanup
Yes No
Upon completion of this step, delete non-starred outputs from completed workflow steps if they are no longer required as inputs.

```
{
    "a_galaxy_workflow": "true",
    "annotation": "",
    "format-version": "0.1",
    "name": "MLM",
    "steps": {
        "0": {
            "annotation": "",
            "id": 0,
            "input_connections": {},
            "inputs": [
                {
                    "description": "",
                    "name": "HapMap file"
                }
            ],
            "name": "Input dataset",
            "outputs": [],
            "position": {
                "left": 295.5,
                "top": 138.5
            },
            "tool_errors": null,
            "tool_id": null,
            "tool_state": "{\"name\": \"HapMap file\"}",
            "tool_version": null,
            "type": "data_input",
            "user_outputs": []
        },
        "1": {
```

Data Repository beta

Search Data Object Persistent Identifier (PID)
Go

Filter by
Creator: Enter creator name (Ex: quzhou)
Time Range: Choose time range (Ex: 01/04/2016 - 04/04/2016 [mm/dd/yyyy])
Workflow: Enter workflow (Ex: quzhou_GLM_20160915T14:23:46)
Filter

DataType - IRRI Rice Genomes tassel workflow

12 Data Object found in 532 milliseconds

kunalan_glm_2018-04-27-20:19:36
11723/e8a04273-54c4-4391-ad3f-55eba809452a
DO Name: kunalan_glm_2018-04-27-20:19:36
Creator: kunalan
Timestamp: 2018-04-27-20:19:36
Download Data Object
Download Metadata Object

luoyu_glm_2018-04-27-20:47:29
11723/d39a5372-e2f4-4e96-8e20-5f7878f75f77
DO Name: luoyu_glm_2018-04-27-20:47:29
Creator: luoyu
Timestamp: 2018-04-27-20:47:29
Download Data Object
Download Metadata Object

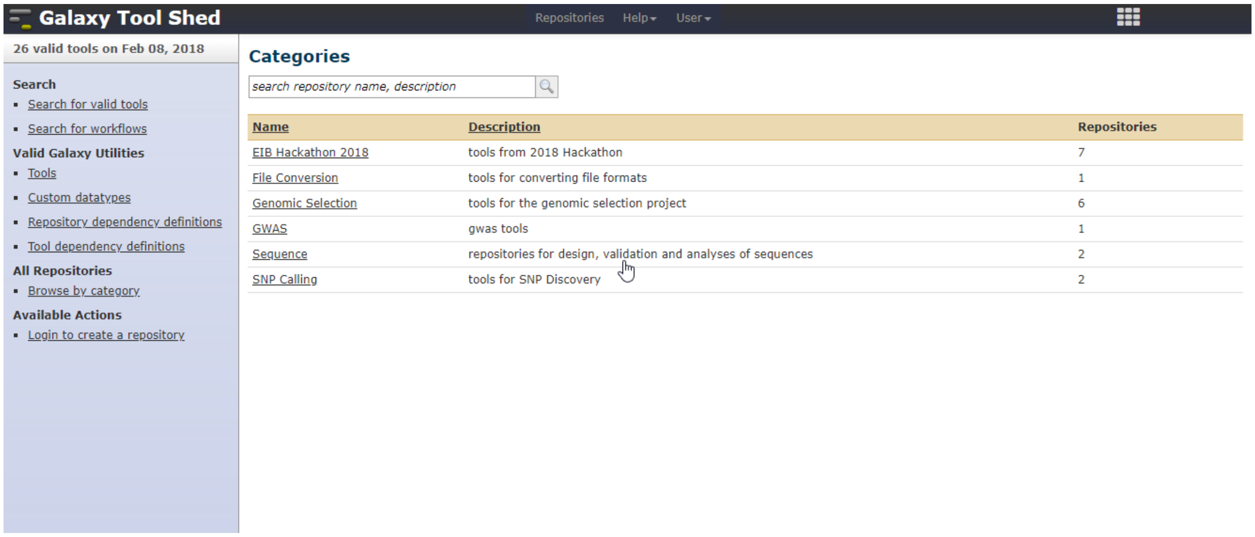
luoyu_mlm_2018-04-30-18:42:26
11723/362b15b1-9334-4f0d-8e6d-8fb3d8d55e45
DO Name: luoyu_mlm_2018-04-30-18:42:26
Creator: luoyu
Timestamp: 2018-04-30-18:42:26
Download Data Object
Download Metadata Object

luoyu_glm_2018-04-30-19:53:18
11723/d9d7809b-348b-4345-b443-ad1f5b59a7ef
DO Name: luoyu_glm_2018-04-30-19:53:18
Creator: luoyu
Timestamp: 2018-04-30-19:53:18
Download Data Object
Download Metadata Object

luoyu_mlm_2018-04-30-18:42:26
11723/362b15b1-9334-4f0d-8e6d-8fb3d8d55e45
DO Name: luoyu_mlm_2018-04-30-18:42:26
Creator: luoyu
Timestamp: 2018-04-30-18:42:26
Download Data Object
Download Metadata Object

B. Digital Object flow in Rice Galaxy Open Access. A Galaxy analysis workflow (exported as JSON file) is deposited to the DO repository, and the data identity server publishes the deposited DO + meta-data for discoverability.

Figure 10. The components (A) and the flow of Digital Objects from upload to discoverability (B) in the prototype Rice Galaxy Open Access.



26 valid tools on Feb 08, 2018

Search
- Search for valid tools
- Search for workflows

Valid Galaxy Utilities
- Tools
- Custom datatypes
- Repository dependency definitions
- Tool dependency definitions

All Repositories
- Browse by category

Available Actions
- Login to create a repository

Categories

search repository name, description

Name	Description	Repositories
EIB Hackathon 2018	tools from 2018 Hackathon	7
File Conversion	tools for converting file formats	1
Genomic Selection	tools for the genomic selection project	6
GWAS	gwas tools	1
Sequence	repositories for design, validation and analyses of sequences	2
SNP Calling	tools for SNP Discovery	2

Figure 11. Rice Galaxy Toolshed with the various available tools.