Building an open Genome Wide Association Study (GWAS) platform

Liya Wang[1,3], Doreen Ware[1,2,3]
[1]iPlant Collaborative, [2]USDA ARS
[3]Cold Spring Harbor Laboratory
Cold Spring Harbor, USA
[wangli, ware]@cshl.edu

Nirav Merchant
iPlant Collaborative
University of Arizona
Tucson, USA
nirav@iplantcollaborative.org

Carol Lushbough
University of South Dakota
Vermillion, USA
Carol.Lushbough@usd.edu

Abstract— We demonstrated how a flexible Genome Wide Association Study (GWAS) platform was built using the iPlant Collaborative Cyber-infrastructure. The platform is open for end users to add additional analysis tools. With the platform, customized GWAS workflows can be built in both iPlant Discovery Environment and BioExtract server.

Keywords— GWAS, iPlant Data Store, iPlant Foundation API, iPlant Discovery Environment, BioExtract server

I. Introduction

The iPlant Collaborative (iPlant) is a United States National Science Foundation (NSF) funded project that has created an innovative, comprehensive, and foundational cyber-infrastructure (CI) in support of plant biology research [1]. iPlant is an open-source project with application programming interfaces that allow the community to extend the infrastructure to meet their needs.

Here, we present a cloud-based open platform for GWAS built on iPlant CI and BioExtract Server [2]. The platform is parallelizable for running on high performance computing clusters (cloud-based) and customizable by an end users (open) for adding new analysis tools to extend the association analysis.

II. Technical Details

A. Data Store

The iPlant Data Store is a centralized facility to address the existing needs of the community to share and store scientifically relevant data sets and metadata. Underlying the Data Store is a federated network of Integrated Rule-Oritented Data System (iRODS) [3] servers running at University of Arizona and mirrored at the Texas Advanced Computing Center (TACC). Users can access the Data Store in multiple ways, including iPlant Discovery Environment (DE) [4] and Davis web application [5] (web interfaces), iDrop [6] (desktop application), RESTful web service interface through iPlant Foundation API [7], and FUSE interface [8] for command line tools. The provenance in the Data Store is addressed through the use of universally unique identifiers (UUID) for every file, folder, and piece of metadata. Every action taken by a user is associated with one or more UUID and logged by a centralized tracking service.

B. Foundation Api

The iPlant Foundation API (fAPI) is a hosted, Software-as-a-Service (SaaS) resource for the computational biology field. Operating as a set of RESTful web services, the fAPI bridges the gap between the HPC and web worlds that allows modern applications to interact with the underlying infrastructure. To deploy a new software package that runs on the command line, user simply needs to upload it to iPlant Data Store, and register it with a Javascript Object Notation (JSON) file through fAPI. The JSON file contains meta data describing a Graphical user interface (GUI) and computing environment.

The fAPI has been adopted by iPlant's own DE, the BioExtract Server, and Easy Terminal Alternative [9].

C. Discovery Environment

The iPlant DE is a distributed, service-oriented architecture (SOA) that exposes service endpoints in a RESTful manner and primarily communicates using JSON. The focus of the DE centers on providing a “software workbench” or “platform” for the execution and management of scientific analyses. GUIs for scientific analysis are driven by metadata “descriptions” encoded in a different JSON file other than fAPI. These “descriptions” can be authored by users through a tool integration service (TiTo), enabling the collaborative extension of the overall application's analysis capabilities. The execution of an analysis can be tailored to the needs of the computation and can either run on a local condor cluster or a remote computing resource through iPlant fAPI's Data, Authentication (Auth), Applications (Apps), and Jobs services.

D. BioExtract

The BioExtract Server is an open, Web-based system designed to aid researchers in the analysis of genomic data by providing a platform to facilitate the creation of bioinformatics workflow. There is a unified authentication mechanism between iPlant and BioExtract using the fAPI's Auth service. Therefore, any analysis tools registered through iPlant fAPI will be accessible in BioExtract with an auto-generated GUI defined by fAPI JSON file. Different from DE's pre-built workflows, BioExtract's scientific workflows are created by recording tasks performed by the user. These tasks may include execution database queries, saving query results as searchable data extracts, and executing local and web-accessible analytic tools. The series of recorded tasks can then be saved as a reproducible, sharable workflow available for subsequent execution with the original or modified input and parameter settings.

Data in iPlant Data Store can be accessed by BioExtract applications or workflows directly through the integrated

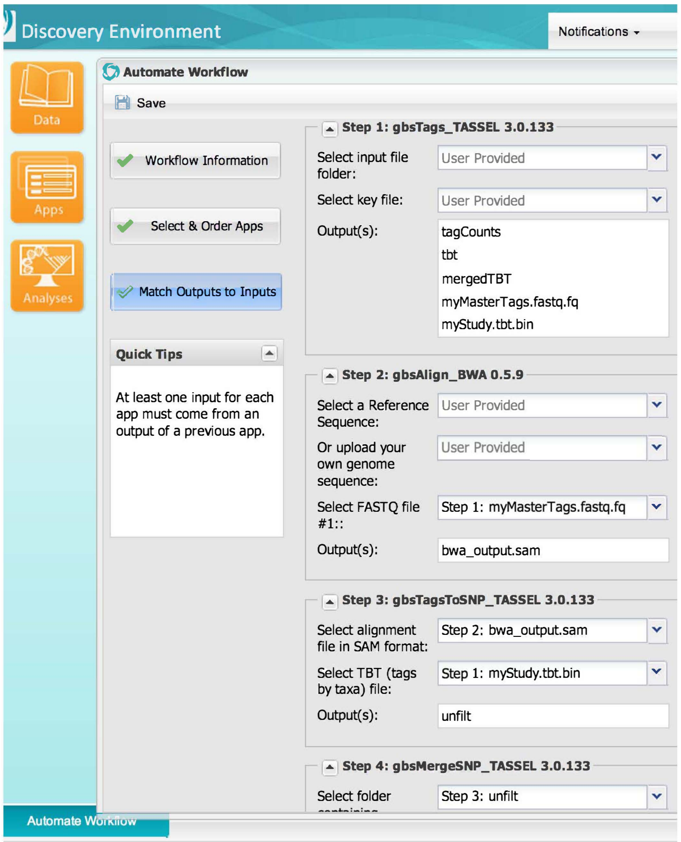
978-1-4799-0898-1/13/$31.00©2013 IEEE

iPlant fAPI IO and Data services. The analysis runs across systems at TACC and SDSC using the iPlant fAPI job services.

III. Constructing Workflow

The instructions for registering iPlant account, deploying applications, and tutorial for this workflow can be found on the public wiki site: https://pods.iplantc.org/wiki/. Once registered, user gets 100 GB initial allocation in iPlant Data Store, and the allocation can be increased by filling out the request form. User can share data with collaborators through iPlant DE and web links. The easiest way to get familiar with iPlant CI (Data, Apps, and Analysis) is through iPlant DE (https://de.iplantc.org/de). User can contact support@iplantc.org if running into any issues.

There are over 400 public applications integrated into iPlant CI. User can deploy their own applications by either requesting application being installed by iPlant staff or through fAPI by following instructions on above wiki site. Using both public and his/her private apps, user can construct automate workflow in DE with the interface shown in Figure 1. Figure 1 shows the three steps (describe, add and order apps, and map outputs/inputs among apps) in constructing the variance calling workflow – the GBS workflow. Here, GBS represents the analysis pipeline for the Genotyping-By-Sequencing protocol [10]. This pipeline takes short reads from sequencing machine and outputs variants for GWAS. The workflow can be kept private or submitted for public use.



Quick Tips

At least one input for each app must come from an output of a previous app.

Step 1: gbsTags_TASSEL 3.0.133

Select input file folder:	User Provided
Select key file:	User Provided
Output(s):	tagCounts tbt mergedTBT myMasterTags.fastq.fq myStudy.tbt.bin

Step 2: gbsAlign_BWA 0.5.9

Select a Reference Sequence:	User Provided
Or upload your own genome sequence:	User Provided
Select FASTQ file #1::	Step 1: myMasterTags.fastq.fq
Output(s):	bwa_output.sam

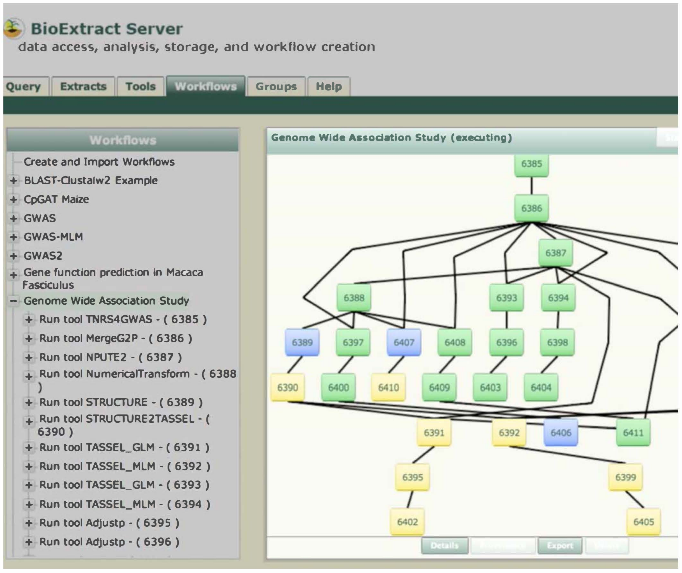
Step 3: gbsTagsToSNP_TASSEL 3.0.133

Select alignment file in SAM format:	Step 2: bwa_output.sam
Select TBT (tags by taxa) file:	Step 1: myStudy.tbt.bin
Output(s):	unfilt

Step 4: gbsMergeSNP_TASSEL 3.0.133

Select folder containing	Step 3: unfilt

Figure 1. iPlant DE interface for constructing automate workflow.



- Create and Import Workflows
- + BLAST-Clustalw2 Example
- + CpGAT Maize
- + GWAS
- + GWAS-MLM
- + GWAS2
- + Gene function prediction in Macaca Fasciculus
- − Genome Wide Association Study
  - + Run tool TNRS4GWAS - (6385)
  - + Run tool MergeG2P - (6386)
  - + Run tool NPUTE2 - (6387)
  - + Run tool NumericalTransform - (6388)
  - + Run tool STRUCTURE - (6389)
  - + Run tool STRUCTURE2TASSEL - (6390)
  - + Run tool TASSEL_GLM - (6391)
  - + Run tool TASSEL_MLM - (6392)
  - + Run tool TASSEL_GLM - (6393)
  - + Run tool TASSEL_MLM - (6394)
  - + Run tool Adjustp - (6395)
  - + Run tool Adjustp - (6396)

Figure 2. Screen shot of executing GWAS workflow on BioExtract server. The status of each application (icon) is distinguished by the color (green: completed; blue: executing; yellow: waiting for being submitted to the cluster once input data dependency is clear).

The public fAPI based apps of iPlant are also available on BioExtract server (under Tools/iPlant at http://bioextract.org). To use these apps and access iPlant Data Store, user's iPlant credential needs to be synchronized on BioExtract server by "Register iPlant Account" (http://bioextract.org/users/create-account.xhtml). Once synchronized, all iPlant's public as well as users own private apps can be used to create automate workflows. Different from DE's build-before-run workflow construction, BioExtract implemented the run-then-build approach. After logging in, user can click on "Create and Import Workflows" after clicking on the "Workflow" button on the top bar, then click on "Record Workflow". All subsequent analysis will be recorded and need to be saved as a workflow.

Figure 2 shows the executing of the GWAS workflow, which submits 29 jobs to TACC's Lonestar cluster with one click and returns analysis results from six different association models.

IV. GWAS Workflow

Figure 3 shows the general workflow for GWAS. The workflow can call markers from sequence data uploaded to iPlant Data Store by end users, or stored at NCBI Short Read Archive (SRA). Retrieving trait and marker data from the public database is under development.

We will use Sorghum Association Panel re-sequencing data [11] as an example to walk through the workflow. The sequencing data is deposited into NCBI SRA with accession number SRR636574 and SRR636575. The workflow starts in iPlant DE by importing data into iPlant Data Store as Fastq files with applications "NCBI SRA import" and "NCBI SRA Toolkit fastq-dump". The Fastq files are further organized for feeding into the GBS workflow built in DE for calling variants with accession ids. In this case, trait data are extracted from a previous study with accession names [12]. Before merging trait data with variant data, accession names in trait data are converted to accession ids with an application named "TNRS4GWAS" for naming resolution.

It is critical to impute missing marker data before the association mapping. An imputation tool, NPUTE [13], is used to fill missing marker data using nearest neighbor algorithm. Then the marker data is filtered by minor allele frequency and converted for downstream analysis. To deal with the confounding factor of population structure in GWAS, a popular application, STRUCTURE [14], is integrated. Similarly, various kinship estimation methods are added for the similar purpose for feeding into various association models.

For association mapping, we have streamlined three popular tools, including TASSEL [15-16], EMMAX [17], and MLMM [18]. Each of these tools takes marker and trait data in different formats; therefore conversion is usually needed for automating the analysis workflow. For TASSEL, different models can be tested by including population structure, kinship or not. The association between marker and trait is quantified with p values. The p values need to be corrected or adjusted for multiple comparisons. A XYPlot function is added to plot the so-called "Manhattan Plot" for visualization. Further development of the visualization function will allow interactive selection of significant p values for extracting nearby genes and their associated pathways for further analysis. For high throughput phenotyping, image analysis tools, such as HYPOTrace [19], will be added and scaled to support extracting traits from imaging based phenotyping protocols.

NPUTE determines optimal window size for imputation by testing known markers with various sizes of windows. Similarly, STRUCTURE picks optimal number of clusters by testing various assumptions of populations. These computations are time consuming but without data dependency between each other. Thus, both NPUTE and STRUCTURE are integrated with fAPI and parallelized through TACC's parametric launcher module.

All applications are integrated in DE for running on either a Condor cluster (Variant calling pipeline) or TACC Lonestar cluster (all GWAS applications) through fAPI. A GWAS workflow is also automated on BioExtract Server, which allows testing various models in a parallel fashion once the data dependency is clear. The construction of automated workflow in BioExtract Server using fAPI based apps demonstrates that iPlant CI makes building innovative solutions with ease, without having to worry about foundational infrastructure.

V. Concrete Use Case

Using Sorghum Association Panel as an example, 310 accessions are remained after merging marker data with trait data (height). The output p values from TASSEL (top 3). EMMAX and MLMM (all after Bonferroni correction) are shown in Figure 4. The dashed line on the right highlights the location of a dwarf gene, Dw1/SbHT9.1. The bottom plot shows one stepwise regression result (MLMM model) after taking the most significant SNP (indicated by the dashed line on the right) as the co-factor to the mixed model. MLMM is designed to account for loci of larger effect, and interestingly, and adding the cofactor highlights another SNP (55424715) that is not significant in a single point mixed model analysis. The SNP also falls on the coding region of a predicted transcript FGENESH00000020814. The nonsynonymous mutation (C/T) on this loci causes the modest threonine (ACC) to isoleucine (ATC or ATT) conversion.

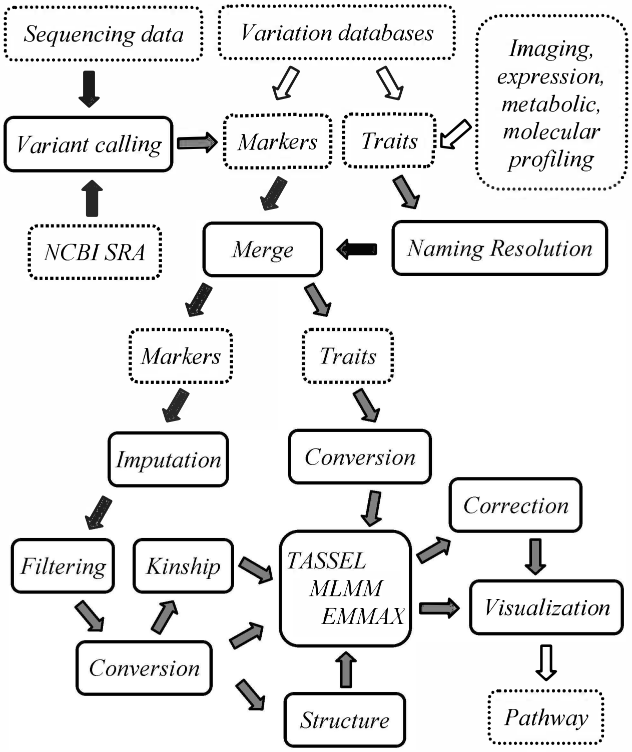


Figure 3. General GWAS workflow (data in a dashed box; applications in a solid box; open arrow for functionality to be built; close arrow for implemented functionality).

VI. Outlook

In the near future, the GWAS platform will be enhanced in three areas. First, the ongoing efforts on iPlant Data Commons will provide cloud storage of community re-sequencing data that are directly accessible to the platform. Second, more popular tools will be scaled and integrated into the platform by iPlant staff or community members. Third, a visualization platform will be built to verify association results, and allow exploring nearby genes and related pathways stored in iPlant Data commons.

Comparing with other GWAS platform, e.g., easyGWAS [20], the iPlant CI based GWAS platform is the first one that is open for the community to contribute new tools, and the first one that provides large scale computation support, the first one that supports seamless variant calling and imputing support, and the first one that leverages big data management and analysis for GWAS.

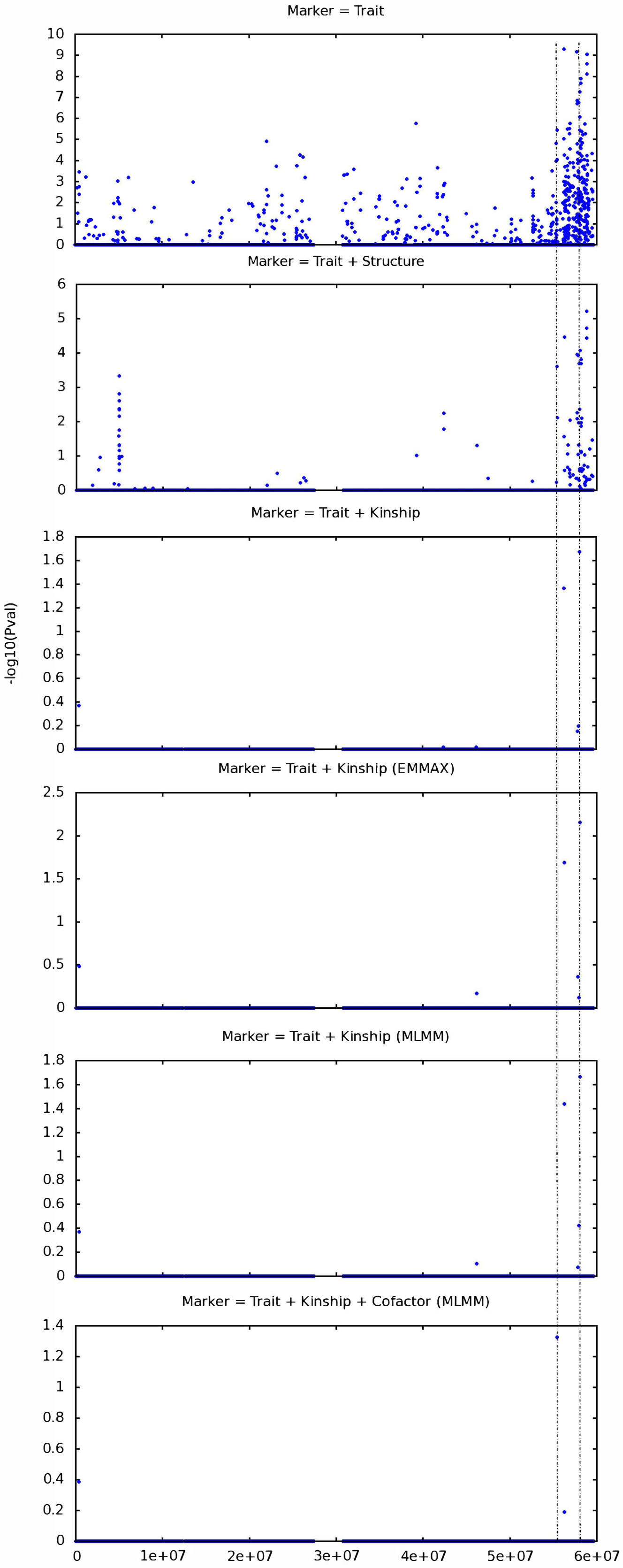


Figure 4. Output of various models with the dashed line highlighting the location of dw1/SbHT9.1 gene (right) and a predicted gene (left).

ACKNOWLEDGEMENT

This work is supported by the National Science Foundation Plant Cyber-infrastructure Program (#DBI-0735191) and USDA-ARS.

REFERENCES

[1] Goff, Stephen A.; et al. "The iPlant Collaborative: Cyberinfrastructure for Plant Biology". Frontiers in Plant Science (2011). Doi: 10.3389/fpls.2011.00034

[2] Lushbough, C., Jennewein, D., Brendel, V. “The BioExtract Server: a web-based bioinformatics workflow platform”, Nucl. Acids Res, vol. 39, iss. W528-W532 (2011).

[3] Rajasekar A., et al, "iRODS Primer: integrated Rule-Oriented Data Systems", Morgan-Claypool Publishers, San Rafael, CA, (2010).

[4] Andrew Lenards, Nirav Merchant, and Dan Stanzione. “Building an environment to facilitate discoveries for plant sciences”. In GCE ACM, New York, USA, 51-58 (2011). Doi: 10.1145/2110486.2110494

[5] “Davis: A Generic Interface for SRB and iRODS”, www.dhpc.adelaide.edu.au/reports/197/dhpc-197.pdf.

[6] Conway, M. iRODS iDrop. https://code.renci.org/gf/project/iRODSidrop/.

[7] Dooley, R. et al. “Software-as-a-Service: The iPlant Foundation API”

[8] Filesystem in Userspace. http://fuse.sourceforge.net.

[9] Easy Terminal Alternative. http://etapub.cgrb.oregonstate.edu/about.php.

[10] Elshire, R. et al. “A robust, simple Genotyping-by-Sequencing (GBS) approach for high density species”. PLoS ONE 6(5): e19379 (2011_. doi: 10.1371/journal.pone.0019379.

[11] Morris, G. et al. “Population genomic and genome-wide association studies of agroclimatic traits in sorghum”. PNAS 110(2) 453-458 (2012).

[12] Brown, P. et al. “Efficient mapping of plant height quantitative trait loci in a sorghum association population with introgressed dwarfing genes”. Genetics. 180(1): 629–637 (2008). doi: 10.1534/genetics.108.092239

[13] Roberts, A. et al. “Inferring missing genotypes in large SNP panels using fast nearest-neighbor searches over sliding windows”. Bioinformatics, 23 (13): i401-i407 (2007). doi: 10.1093/bioinformatics/btm220

[14] Pritchard J. et al. “Inference of population structure using multilocus genotype data”. Genetics. 155(2): 945-59 (2000).

[15] Bradbury, P. et al. “TASSEL: software for association mapping of complex traits in diverse samples”. Bioinformatics 23 (19): 2633-2635(2007). doi: 10.1093/bioinformatics/btm308

[16] Zhang, Z. et al. “Mixed linear model approach adapted for genome-wide association studies”. Nature Genetics 42(4): 355-360 (2010).

[17] Kang, H. et al. “Variance component model to account for sample structure in genome-wide association studies”. Nature Genetics. 42(4): 348-354 (2010).

[18] Segura V. et al. “An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations”. Nature Genetics. 44(7): 825-830 (2012). doi: 10.1038/ng.2314.

[19] Wang L. et al. “HYPOTrace: image analysis software for measuring hypocotyl growth and shape demonstrated on Arabidopsis seedlings undergoing photomorphogenesis”. Plant Physiology 149(4):1632-7 (2009). doi: 10.1104/pp.108.134072.

[20] Grimm, Dominik, et al. "easyGWAS: An integrated interspecies platform for performing genome-wide association studies." *arXiv preprint arXiv:1212.4788*(2012).