Using *R* to Empower a New Plant Biology

Naim Matasci[1,*], Matthew Vaughn[2], Edwin Skidmore[1] and Nirav Merchant[1,3]

[1] The iPlant Collaborative, BIO5 Institute, University of Arizona - [2] The iPlant Collaborative, Texas Advanced Computing Center, University of Texas at Austin - [3] Biotechnology Computing Facility, BIO5 Institute, University of Arizona - * nmatasci@iplantcollaborative.org

ABSTRACT

www.iplantcollaborative.org

The iPlant Collaborative is a community-driven U.S. National Science Foundation program to build a cyberinfrastructure for the plant sciences that will enable new conceptual advances through integrative, computational thinking. The complexity and sheer magnitude of problems like understanding plants' adaptation to climate change or increasing crop yields makes it almost impossible for individual researchers or local research groups to tackle such problems alone. The cyberinfrastructure provided by iPlant gives researchers access to High Performance Computing resources and allows them to collaborate, share and integrate data and algorithms regardless of their size or complexity.

As part of this cyberinfrastructure, iPlant is dedicated to facilitating the development, execution and distribution of *R* based tools.

Developers can work on constructing new tools in Atmosphere, iPlant's cloud computing environment. iPlant's unified data storage solution provides a single point of access to large datasets and community resources, whereas users and developers can execute computationally intensive scripts on HPC resources through iPlant's foundational API. Finally, a web based Discovery Environment provides a simple way to create graphical front-ends for *R* based applications, thus making them available to a wider user community.

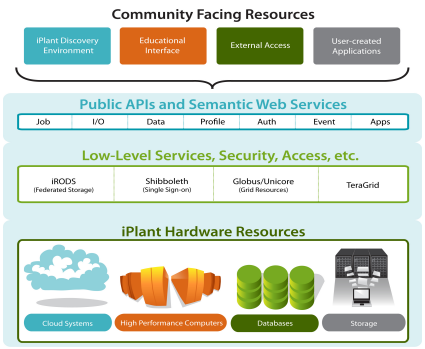

Figure 1: Overview of iPlant's cyberinfrastructure.

ATMOSPHERE

www.iplantcollaborative.org/discover/atmosphere

Atmosphere is a cloud service that gives users access to highly configurable, customized computational resources. Through a web interface users can launch their own private virtual machine (VM), which can be either a basic Linux installation or one of the VMs preconfigured for common tasks and analyses. VMs range from low-memory single CPU machines to 4 CPUs with 32 gigabytes of memory that can support analysis of large data. Atmosphere is also a gateway to access iPlant's core infrastructure resources such as high performance computing (HPC), grid computing environment and large data storage system. VMs can be kept private or made public so that users can develop tools and analytical workflows with their collaborators and then share them with the rest of the community. These VMs can also be used for software development, including VMs that mirror the HPC and grid environments. iPlant also provides a VM specifically designed with *R* users and developers in mind. *R* comes preinstalled on this machine, with additional packages and software development tools, but users are free to modify it to fit their needs.

DATA STORE

www.iplantcollaborative.org/discover

The iPlant Collaborative offers free, secure, cloud based storage space to its users. The storage system is designed to handle any kind of size, up to the terabyte range and can be accessed through a variety of interfaces: from web services to mountable filesystems to high speed command line transfers. Shared reference information, such as genome sequences can thus be accessed by the different components of iPlant's cyberinfrastructure and by the community. The fine-grained access control ensures that important data can be kept private or only be visible to selected people, permitting collaborative work on large unpublished datasets without the risks and costs involved in transferring and replicating such large datasets.

FOUNDATIONAL API

www.iplantcollaborative.org/discover

Advanced users and developers can directly interact with iPlant's physical infrastructure through an Application Programming Interface (API). iPlant's foundational API offer a RESTful web service that can be accessed by *R* scripts to import, export or convert data files, list the available applications and launch and manage remote jobs running on iPlant's HPC resources or on Atmosphere virtual machines. It supports http simple and token-based authentication, meaning that access to private resources can be directly granted or revoked by the owner. The foundational API is being currently expanded, adding more functionalities such as profile discovery and auditing. Moreover, iPlant is also providing a virtual application layer that emulates local applications while running them on HPC resources.

DISCOVERY ENVIRONMENT

www.iplantcollaborative.org/discover/discovery-environment

The Discovery Environment (DE) is a web interface that provides access to the computing, data storage, and analysis applications available through iPlant. It is geared towards users that are unfamiliar with command line tools and scripting languages and allows the integration of multiple tools and workflows under a uniform graphical interface. Developers can bring their own software and integrate it into the DE in minutes. This way, *R* scripts can be provided as standalone applications that can be shared with collaborators or made available to the entire community, so that researchers can immediately start to use them and integrate them into their analytical workflows.

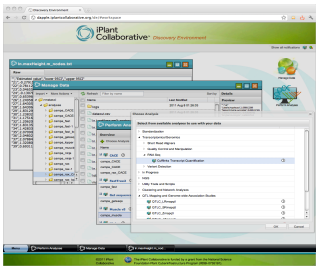

Figure 2: The Discovery Environment. Some of the analyses are powered by *R* scripts.

The iPlant Collaborative is funded by a grant from the National Science Foundation Plant Cyberinfrastructure Program (#DBI-0735191).