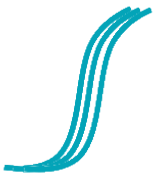
iPlant Collaborative™
Cyberinfrastructure for Life Sciences

Get Started with iPlant Toolkit

Use iPlant Resources to Support Diverse Research Programs

Get Science Done



Reproducibly

Productively

ISMB/ECCB Applied Knowledge Exchange Session

July 11th, 2015

Workshop Wiki: www.iplantc.org/akeswiki1

Account issues/information:	user.iplantcollaborative.org
iPlant support forum:	ask.iplantcollaborative.org
Support issues:	support@iplantcollaborative.org
General questions:	411@iplantcollaborative.org

iPlant staff @ this workshop

John Fonner	jfonner@tacc.utexas.edu
Jason Williams	williams@cshl.edu

Acknowledging iPlant

The iPlant Collaborative is funded by the National Science Foundation under grant #DBI - 1265383. Learn how to properly acknowledge any iPlant tools or services that you make use of: **www.iplantc.org/citeus**

iPlant's Vision *- The iPlant Collaborative enables life science researchers and educators to use and extend iPlant's foundational cyberinfrastructure to understand and ultimately predict the complexity of biological systems and their dynamic nature under various environmental conditions, thus stimulating the transformation of how the life sciences community does research.*

Introduction

The iPlant Collaborative was funded in 2008 by the National Science Foundation to develop cyberinfrastructure (software, high-performance computing, data management, and people) needed to support data-intensive biology. Since then, we have developed a number of resources for phylogenetics and genotype-to-phenotype science – resources you can use to get more science done and ask bigger questions while saving you time, effort, and funding.

What is Cyberinfrastructure?

Cyberinfrastructure (CI) is data storage, software, high-performance computing, and people – organized into systems that solve problems of size and scope that would not otherwise be solvable. Some of these components you likely have already – your personal CI. However, iPlant CI can help fill in gaps where you don't have access to resources you need, and expand your ability to address larger questions that involve bigger compute and a greater need to share data with collaborators and the community.

Where to start?

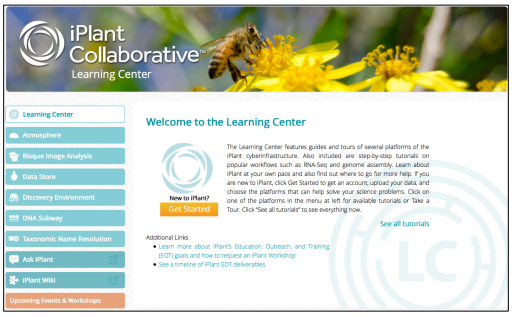
First – signup for your free iPlant account by visiting our homepage:
www.iplantcollaborative.org

Next, visit the Learning Center on the iPlant homepage:
www.iplantcollaborative.org/learning-center

At the Learning Center, you will find the latest version of this guide as well as **videos and tutorials** that guide you through the iPlant platforms and several popular science tutorials on analyses like RNA-Seq and genome assembly.

Finally, after reading this guide visit our iPlant user forum:
ask.iplantcollaborative.org

The iPlant forum is the best way to get answers from other community members and iPlant support staff on all questions technical and scientific.



Visit the iPlant Learning center to see detailed tutorials and videos on iPlant CI and popular science workflows

www.iplantcollaborative.org/learning-center

Toolkit – Item One: Strategies for Improving Bioinformatics Capabilities

Tips for addressing challenges

If you have training and/or experience in bioinformatics, than you probably already have this item in your toolkit. For the majority of biologists and many iPlant users however, bioinformatics skills are something they're just getting started with or have only recently begun to pick up. Even though these reminders are general, they give important context that can help you make the best use of iPlant.

Approach learning bioinformatics as you would other new skills in the lab or field

Most day-to-day computing and software allows us to get what we want done quickly – almost without thinking. Realistically though, bioinformatics is not always going to be straight forward, and even with large computing power some types of analyses will still take hours or days to run. When you first learn a new laboratory technique you probably don't expect every new protocol will work the first time without optimization. As you work with a new workflow or dataset it will be helpful approach things with a similar perspective.

Use "Computational Thinking" to decide when to settle and when to optimize

Setting realistic expectations does not mean that you should settle for slow software or processes. Most workflows have some bottlenecks and at least some of these may be areas where automation or gaining access to a larger pool of resources will help.

Problem	Suggestion
I have a workflow that involves large number of repetitive steps.	Build a pipeline using the iPlant Discovery Environment (p.13) or automate tasks at the command line using Atmosphere (p.19).
I have analyses that take hours or days to run, forcing me to give up a computer to these tasks. I don't know if there are faster ways to do things.	Post to the iPlant user forums to ask if your expectations are realistic or if there are software solutions that can take advantage of high-performance computing to speed runtimes. Use an Atmosphere instance to move long-running jobs off local hardware.
The sole copy of my data is on a single hard drive. I don't have a solution that allows me to share it and analyze it easily with collaborators.	Check out the capabilities of the iPlant Data Store (p.6) to share and manage data.

Additionally, computational thinking means considering what data challenges you might encounter before you start a project (i.e. how many files will be created, how will they be named, what metadata should be collected, etc.).

Tip: If you are beginning a data-intensive project for the first time, post your questions to **ask.iplantcollaborative.org** – we, and other community members would be happy to share advice and experience.

Know what type of bioinformatics skills you have and would like to have

With unlimited time, you could probably pick up a great deal of bioinformatics skills. However, most iPlant users would like to achieve competency – being able to complete routine analyses, under reasonable circumstances, in a reasonable amount of time. Some useful personas developed in Welch et.al describe relevant use cases – ones that will **also help you determine how to get the most out of iPlant**. Which classification best describes you? See ideas on this from: *Welch L, Lewitter F, Schwartz R, et al. (2014)*

iPlant platforms accommodates diverse types of users

Bioinformatics Users (Bench/Field Scientists)

- ✓ Spends 60% or less effort on bioinformatics related work – other efforts are on bench/field work and other tasks
- ✓ Vast majority of bioinformatics work done using programs with graphical user interface (not command line)
- ✓ Pain points may include lack of access to local compute resources and/or bioinformatics support

Bioinformatics Scientists

- ✓ All time is spent on computational work
- ✓ Majority of bioinformatics is done at the command line; occasionally uses GUIs
- ✓ Pain points may include unbudgeted time providing support, and finding students and staff with specialized math and stats skills

Bioinformatics Engineers (core facilities)

- ✓ Time is mostly spent on computation work with significant time dedicated to user support
- ✓ Majority of bioinformatics is done at the command line; occasionally uses and develops GUIs
- ✓ Pain points may include challenges working with collaborators, staff, and users with challenging needs or underspecified requests

iPlant has platforms, resources, and tools that address the needs of all of these types of use cases – throughout this booklet we'll point these out!

Don't put off moving to the next level

For those just beginning bioinformatics it is important to take on new challenges to grow your abilities. Getting more skills with the command line will allow you to access the greater body of bioinformatics work. Learning a programming language like R or Python will make it possible for you to ask larger questions more efficiently. For advice on these topics see the **resources section** at the end of this booklet.

Toolkit – Item Two: Data Storage that supports the Life Cycle of Data

Use the iPlant Data Store to Share and Manage Big Data

"[G]enomics technologies will enable individual laboratories to generate terabyte or even petabyte scales of data at a reasonable cost. However, the computational infrastructure that is required to maintain and process these large-scale data sets, and to integrate them with other large-scale sets, is typically beyond the reach of small laboratories and is increasingly posing challenges even for large institutes." – Schadt et. al, 2010

How the Data Store "Gets Science Done" reproducibly and productively

- Store any type of files related to your research
- An evolving "Data Commons" gives you access to important datasets

- Metadata captures information needed to ensure reproducibility
- Automatic backup and easy accessibility supports your investigation's data management plan

- IRODS technology makes high-speed transfers possible (100GB in 30 min)
- Share data instantly with collaborators and make it accessible to the world

How the Data Store Helped

"The ability to transport 2TB of data overnight using the iRODS system was particularly helpful because previously, we had been mailing hard drives which is not an optimal solution to sharing big data. Among the most helpful aspects of using iplant has been the ability to more efficiently conduct collaborative research." **J.Koltes – Iowa State**

Selected Features of the Data Store

Feature	Details	How this benefits you
Generous storage	100 GB allocations for each user – terabyte allocations available with justification (e.g. for community-sized projects or other special needs).	You can use iPlant knowing that you will have all the space you need to complete your work.
Data co-localized with compute	Data and compute reside together in the iPlant CI.	Data-intensive operations perform better when large amounts of data do not have to be moved from local to remote systems.
Automatic backup	Data are automatically backed up at two locations (Arizona and Texas).	All projects require a solid data management a plan. Backups reduce your vulnerability to risks like hard drive failures.
Fast up/download	Multiple ways to upload and download data at the command line, through standalone drag-and-drop software, or through the platforms themselves.	Transferring large datasets made relatively routine.
Easy sharing	Share large datasets with other iPlant users, or create a community folder with fine-grained access control. Share data with anyone through on-demand URLs.	You can make any data related to a publication or community effort accessible to everyone. You can also use iPlant to create a shared repository for your lab, colleagues, or community.
Metadata management	iPlant allows you to manage the metadata related to any file – templates allow you meet "minimum information" standards associated with specific data types.	Simple metadata strategies (like informative filenames) reach beyond their usefulness with larger datasets. Working within the iPlant CI, you can keep track of metadata to ensure important attributes (how was a file analyzed, how was it edited, etc.) remain tied to the data.

Funders now rightly view data as assets that they are underwriting and so seek the greatest pay-off for their investments. They demand that researchers and host institutions document and implement data-management and data-sharing plans that address the full life cycle of data — including what happens after a grant finishes - Lynch, 2008

How the iPlant Data Store relates to our cyberinfrastructure

The iPlant Data Store unifies all iPlant cyberinfrastructure. In practical terms this means that though the interfaces to the Data Store differ (e.g. iDrop, iCommands, Discovery Environment, APIs, etc.) they are all operating on the same system – giving you the freedom to move between platforms. Here are some solutions for common data management tasks in iPlant:

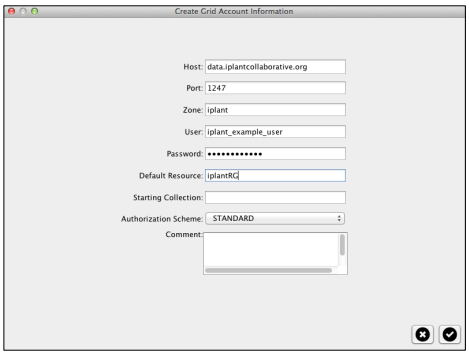
Common Tasks	Recommendation	Difficulty
Upload or Download files (including large files, large numbers of files, and/or folders)	Use the iDrop data transfer application (for Windows, Mac, Linux)	Easier
Share Data with other iPlant users (files or folders) or create Public URLs to specific datasets	Use the Discovery Environment (any web browser)	Easier
Write scripts or work in the terminal/shell to manage data	Use iCommands (Mac or Linux terminal)	Medium
Manage file metadata	Use the Discovery Environment (any web browser)	Easier

Easiest way to get your data uploaded to iPlant

iDrop is a file transfer application that will allow you to drag-and-drop files between your local computer and the iPlant Datastore. There are other options for moving data - to see a complete list of tutorials and documentation on all the above solutions see the iPlant Learning Center.

Transfer Data with iDrop

1. Download the iDrop version for your operating system (Windows, Mac, Linux) – see the "download iDrop link" at www.iplantcollaborative.org/learning-center/data-store/tour



2. Depending on your version, When iDrop starts, it may request a "pass phrase" – since this step does not allow authentication into your Data Store account a simple, memorable phrase is all that is required. **If you download the pre-configured iDrop, you will only be asked for a username and password – simply enter your iPlant Username and Password and continue this guide at step 6.**

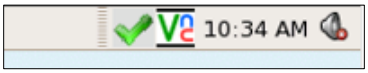
3. Click the  (Add User) icon to connect iDrop to the iPlant Data Store

4. If you are prompted to "Create Grid Account Information" enter the following (paying attention to capitalization) and entering your iPlant credentials where required:

Field	Entry
Host	data.iplantcollaborative.org
Port	1247
Zone	iplant
User	your_iplant_username
Password	your_iplant_password
Default Resource	iplantRG

All other fields can be left blank

Windows



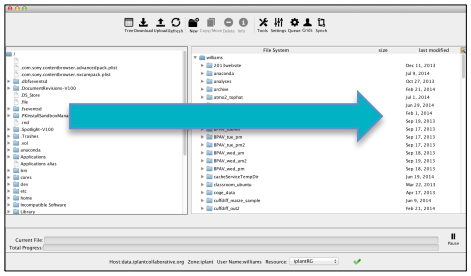
Mac/Linux

5. Select the data.iplantcollaborative.org line for your host and then lick click the  (authentication) icon to access the data store. One first setup you may wish to select the "Show GUI at Startup." option if prompted.

Tip: *When iDrop starts, it may appear in your taskbar as a* ***green check*** *(usually lower right for Windows computers, and upper left for Mac/Linux). Click on the green check and select* ***iDrop*** *from the menu to start the interface.*

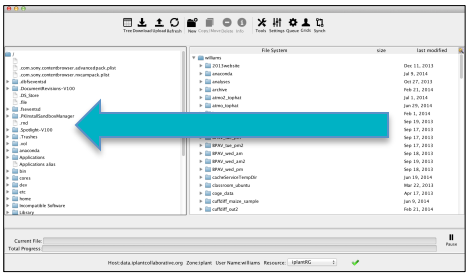
6. To upload – select the place on your local computer (left column) where the files are located:
    - **Windows** – select your hard drive (e.g. C:\) and then navigate your directory
    - **Mac/Linux** – select the files or folder you wish to move (most items will be in the /home directory

   Then drag the file(s) or folder(s) you wish to transfer into your iPlant Data Store (right column).



Drag (left to right) – upload

7. To download – select the file(s) or folder(s) in your iPlant Data Store and drag them to a location on your local computer (left column)



Drag (right to left) – download

Tip: *In the right column of iDrop (your iPlant Data Store) your home folder is the same name as your iPlant username. You will see other folders of data at this level - these are directories that either are setup by other iPlant services or directories of data shared with you. If you get lost navigating, use the* ***Tree*** *menu at the top of the application to navigate back to your home folder.*

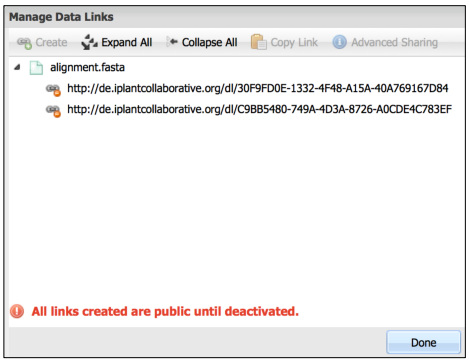
Easiest way to share Data with iPlant

The easiest way to share large datasets within iPlant is to give access to another user. Rather than copying data (which costs space and time), the iPlant Data Store allows you to search for other iPlant users and share data. You can decide what level of access you want to grant to any specific user (e.g. read-only, write access, ownership). You can also use "Data Links" to share individual files even with those who do not have an iPlant account. This is convenient for smaller files, although sensitive data should not be shared this way as anyone with the link can download the data (until you terminate the link).

Sharing with a data link in the iPlant Discovery Environment

1. Log into the iPlant Discovery Environment (click the *Launch* link from the iPlant homepage at www.iplantcollaborative.org)

2. In the **Data** console, next to the file click the ( ) icon or select a file you wish to share, then from the **Share** menu click **via Public Link** (Note, only individual files can be shared from a public link).



3. In the “Manage Data Links” window, select the file you wish to share, and click **Create**.

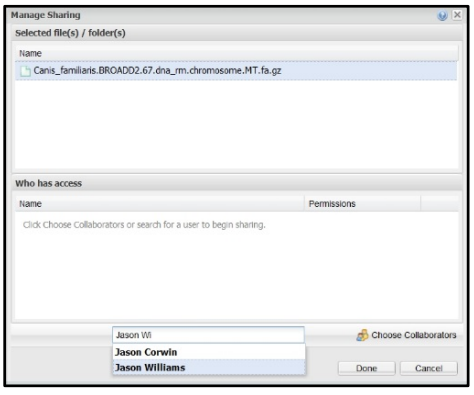
4. A new URL should appear beneath the file name. Click on this URL and the click on **Copy** in order to be presented with a window that will allow you to copy the URL to your clipboard. Anyone who you share this link with will be able to download the file.

5. Deactivate a data link by selecting the file; from the **Share** menu click **via Public Link.** Clicking the( ) icon next to the links you wish to inactivate. Once you deactivate the link, anyone with whom you shared it with will no longer be able to access that data.

Sharing with other iPlant users in the iPlant Discovery Environment

1. Click the ( ) icon or check-select a file(s), folder(s) you wish to share with another user; then either click ‘Begin Sharing’ on the right-hand Details menu or click **Sharing** and select **via Discovery Environment.**

2. In the “Manage Sharing” menu, under ‘Selected File(s)/Folder(s) the name of the files and folders you are currently sharing are displayed. Ensure the file you wish to share now is selected.

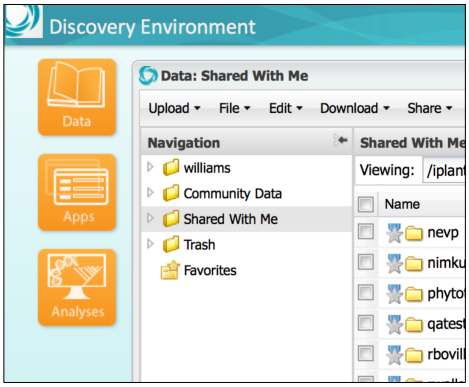


3. In the ‘search for users’ menu search for the iPlant user you wish to share with by search for their name, or iPlant username. You may also select ‘Choose from Collaborators’ which will bring up a list of people you have designated as collaborators.

4. Next, under ‘Permissions’ choose what permission you want to grant the person you are sharing this file with.

5. Once you are finished, click **Done** to begin sharing. The user will be notified that a file has been shared with them. Files shared with you appear in the 'Shared With Me' top-level folder in the **Data** console.

> ***Tip:*** *You can manage your list of collaborators from the menu that appears under your username in the Discovery Environment*



Permissions and privileges for sharing data for sharing are explained in the following chart

Permission Level	Read	Download/ Save	Metadata	Rename	Move	Delete
Read	**X**	**X**	View			
Write	**X**	**X**	Add/Edit			
Own	**X**	**X**	Add/Edit	**X**	**X**	**X**

How Different Scientists Might Make Use of the Data Store

Bioinformatics Users (Bench/Field Scientists)	• Uploads all fastq files for an RNA-Seq experiment for analysis in the Discovery Environment • Sharing all the analyses related to thesis work with an advisor
Bioinformaticians	• Use a metadata template for assembled genomes students and collaborators will place in a shared folder • Uses public links in the supplemental materials of publications to share data
Bioinformatics Engineers (Core Facilities)	• Developed a script to automate transfer of data to core users • Uses a shared folder to make large datasets accessible

Toolkit – Item Three: Web-based Graphical Bioinformatics Platform

iPlant Discovery Environment – An Extensible Bioinformatics workbench

"Over the past decade the volume of bioinformatics publications has grown tremendously. Within the scientific community, there have been concerns about disappearing databases, lack of interoperability, incomplete disclosure, and general quality and integrity issues." - Tan et.al , 2010

How Discovery Environment "Gets Science Done" reproducibly and productively

- Use hundreds of bioinformatics Apps without command line
- Add your own applications – an extensible, scalable platform

- Create and publish Apps and workflows so anyone can use them
- Detailed analysis history– "avoid forensic bioinformatics"

- High-performance computing – not dependent on your hardware
- Manage a secure and data repository and share data easily

How the Discovery Environment Helped

"The apps in the discovery environment are quite useful [and] save me from having to install scripts and command line applications on lab computers…the computing power available at iPlant, saves me a lot of time having to wait on a process to finish. And I mean a lot of time! A normal operation that would take my personal laptop a day to perform (no kidding), would take my lab computer roughly 7 hours, but takes iPlant no more than an hour, and usually much less." **A.Nelson – University of Arizona**

Selected Features of the Discovery Environment

Feature	Details	Benefits
Simple Web Interface	The Discovery Environment is a Graphical User Interface for bioinformatics application and is accessible through any web browser.	You can access bioinformatics applications that normally would only run on the command line. Point-and-click access to more than 450 installed applications.
Processing Power	The Discovery Environment is hosted on a powerful compute cluster and is integrated with iPlant APIs to access even larger HPC resources through XSEDE.	You are freed from the limitation of local resources. HPC applications that might be challenging for you to deploy on your own are readily accessible through the interface.
Data Management	Users can up/download data, share data with access control, manipulate metadata and access visualizations of data.	Easy desktop-style management allows you handle most routine tasks for data transfer
Workflow Creation	A visual workflow editor allows users to create, edit, and share simple linear workflows.	Workflows allow you to automate commonly used analysis pipelines to save time, work reproducibly, and reduce the chance of mistakes.
Analysis History	The Discovery Environment supports detailed analyses histories. Automatically assigned metadata attributes are assigned to files processed within the system.	It is easy to keep track of where data came from, and how it was used. Access to prior versions of applications allows you to reproduce prior results.
Tool Integration	The Discovery Environment is a platform that is user extensible. Users can integrate new applications and design custom interfaces.	If a required tool is not present in the Discovery Environment, you can integrate it or request support to help deploy a specific application. You can share applications publically, or selectively restrict access.

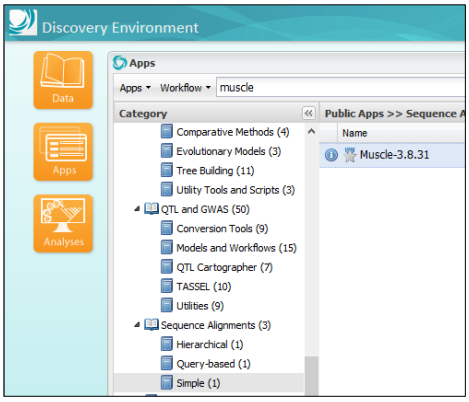
How the Discovery Environment relates to our cyberinfrastructure

The Discovery Environment (DE) was designed to be an all-purpose bioinformatics workbench – tailored to the needs of biologists who have data to analyze but who may not have command line expertise. Most of the popular bioinformatics analyses (e.g. genome assembly, RNA-Seq, phylogeny methods, GWAS, etc.) can be done from start to finish within the DE.

Example analysis in the Discovery Environment

This example demo covers the basic steps of using a bioinformatics application (in this case using MUSCLE to generate a multiple sequence alignment) within the DE. All the other DE applications have similar interfaces and work in a similar way. For a list of all the current applications see:www.iplantcollaborative.org/apps1

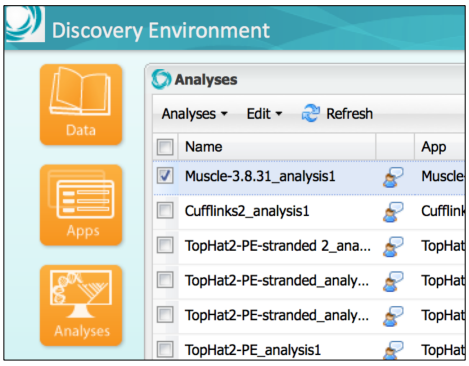
1. Login to the iPlant Discovery Environment (click the *Launch* link from the iPlant homepage at www.iplantcollaborative.org)



2. Click **Apps** from the DE workspace and select the aligner **MUSCLE-3.8.31** (Location: *Public Applications> Sequence Alignments>Simple)*. Click on the actual app name.

3. Under "Analysis Name" leave the defaults or make any desired notes.

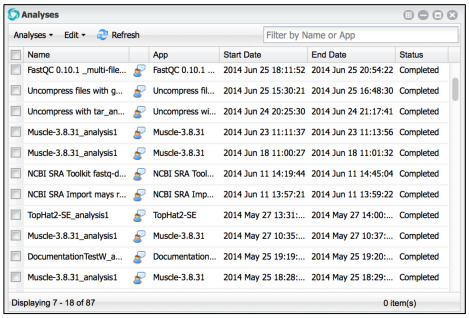
4. Under "Select Input data" click **Browse**, then navigate to and select the **DE_sample_plants.fas** file. (Location: *Community Data > iplant_training> de_walkthough)*. Then click **OK.**



5. Under "Sequence Type", select **DNA**, and then click **Launch Analysis.** You will receive a notification and may close the Apps window.

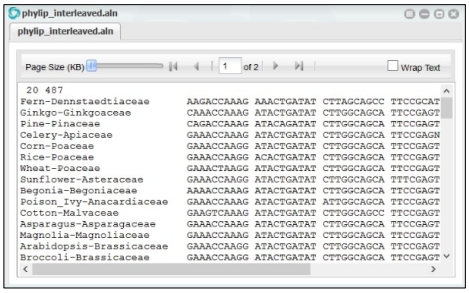
6. Click on **Analyses** from the DE workspace and monitor the status of your submitted job (You may have to click refresh to view updated status).

7. In the **Analysis** console, once your status appears as 'Completed,' click on the name of your analysis. (You could also navigate to your expected output folder from the Data console, by default: *your_iplant_username > analyses*).



8. You should have a folder (named according to your job title) with the following outputs: "logs" - (a folder of log files), "clustalw.aln", "fasta.aln", "phylip_interleaved.aln", and "phylip_sequential.aln".

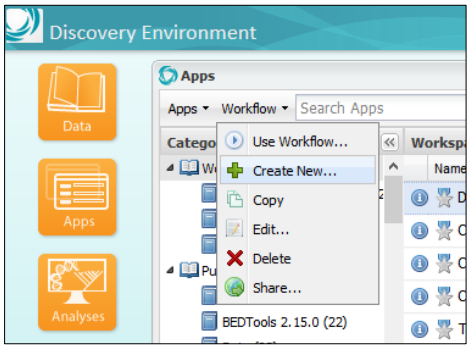
9. Click the **phylip_interleaved.aln** file to view your aligned sequences. You can download these outputs or use them in further analyses.



How to create a workflow in the Discovery Environment

Simple workflows are one way to automate analyses. Building an automated workflow is not a trivial task, and complex command line workflows offer the greatest flexibility to customize. Working with the Discovery Environment's visual workflow creator allows you to create and share workflows that are useful for a number of objective and can save significant time and effort. This example workflow uses two Apps (MUSCLE, RAxML) that will construct a multiple alignment and produce a phylogenetic tree.

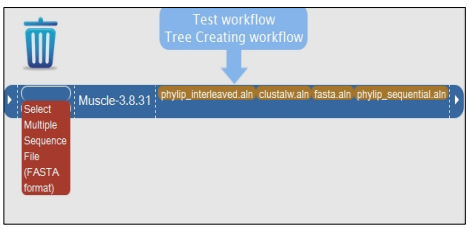
1. Click on **Apps** from the DE workspace and then click on the **Workflow** button. Select **Create New**.



2. In the visual workflow view click **Switch View** enter a name and description for your workflow.

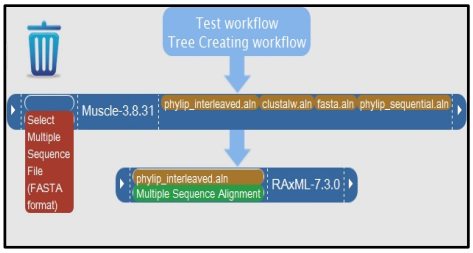
3. From the catalog of Apps (categories) drag **MUSCLE – 3.8.31**. (Location: *Public Applications > Sequence Alignments > Simple*) into the workspace.

4. Click the right-most arrow to view the list of possible outputs for **MUSCLE – 3.8.31**.



5. From the catalog of Apps (categories) drag **RAxML – 7.3.0**. (Location: *Public Applications > Phylogenetics > Tree Building*) into the workspace.

6. Drag the **phylip_interleaved.aln** output from the workflow into the empty 'Multiple Sequence Alignment' input option for **RAxML – 7.3.0** app. The green color change indicates the input option is satisfied. Notice the input option for the **MUSCLE** app is blank because user provided input is expected.

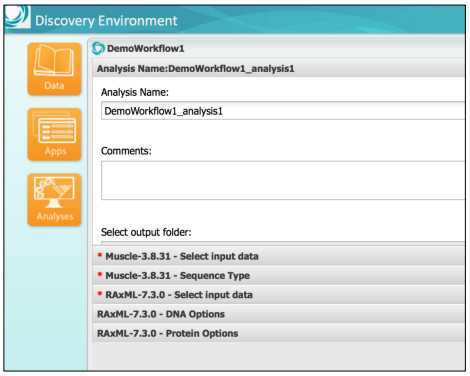


7. Click **Save** to save the workflow to your personal Apps workspace. Close all windows.

8. Click on **Apps** from the DE workspace**.** Under *Workspace > Applications under development* you should see the workflow you created. Select the workflow to run it.

9. Under "Analysis Name" leave the defaults or make any desired adjustments/entries for the analysis name, description, or output folder.

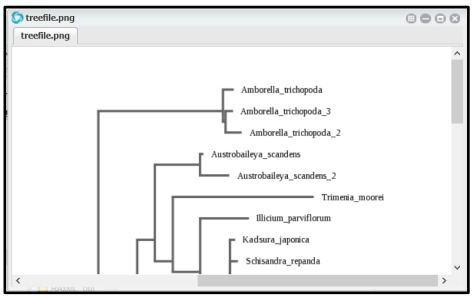
10. Complete mandatory options (red asterisk) as follows:



    - Muscle-3.8.31 – Select input data: **Community Data > iplant_training > intro_phylogenetics > 01_input_data** and select the file **atpB.fa**.
    - Muscle-3.8.31 – Sequence type
        - **DNA**
    - RAxML-7.3.0 Select input data
        - Choose multiple alignment is not an option here because previous app provides this input
        - Sequence type should be: **DNA**

11. Click **Launch Analysis**.

12. Click **Analyses** from the DE workspace and monitor the status of your job. When it is complete, select your job and click **View Output(s)** to navigate to the job output.



Tip: *Not all Apps are designed for use in a workflow. If your workflow does not have specified inputs/outputs (e.g. the arrows on the workflow editors don't expand) you may need to modify the App. See the Discovery Environment documentation on the Learning Center*

How Different Scientists Might Use the Discovery Environment

Bioinformatics Users (Bench/Field Scientists)	• Uses the DE for all data uploads and sharing • Outputs the lab's workflow results to a shared folder
Bioinformaticians	• Install HPC applications for high-memory nodes here so anyone can use them • Create custom applications with parameters selectively hidden or exposed
Bioinformatics Engineers (Core Facilities)	• Developed a workflow for sequence read QC and filtering to support local sequencing center • Teach about genome assembly to first-time learners using demo datasets

Toolkit – Item Four: On-Demand Computing

Atmosphere Cloud Computing – The largest, easiest to use open cloud for life science

"Unlike wet-laboratory experiments, where reviewers use their best judgment to consider whether the experimental disclosure is sufficient for reproduction by any personal skilled in the art, and where experiments require great time and expense to be repeated during the review process, the rigour of the review process for bioinformatics investigations can reach a level where the author's in silico experiments can be selectively or completely reproduced, depending on the computational power and resources available to the reviewer. In this way, the veracity of the claims can be tested and any queries be raised before the paper can be approved for publication. Moreover, any doubtable claims can be refuted or rebutted before publication. Any software coding errors can be detected earlier, and any database errors fixed before public release." - Tan et.al, 2010

How Atmosphere "Gets Science Done" reproducibly and productively

- Work in an on-demand Linux environment (most bioinformatics)
- Collaborate with students and colleagues on the same instance

- Make data, workflows, and analyses available in a public image
- Access previous software version and images

- Large CPU/Memory instances to run intensive applications
- Move your analyses from your laptop to the cloud

How Atmosphere Helped

"A few years ago you helped me use Atmosphere as part of an undergraduate class. It worked very well and was the first time I've ever seen students really master the Unix environment. Another "big win" was the students especially loved being able to log into their VMs from home or school. This convenience and flexibility was a big reason they continued using the VMs throughout the semester even though it wasn't a strict requirement." **A.Loraine– UNC Charlotte**

Selected Features of Atmosphere

Feature	Details	Benefits
Simple Interface	Selecting, configuring, and launching an instance can be done in as little as two mouse-clicks.	There is far less overhead to using Atmosphere compared with large-commercial solutions that can be challenging to configure.
Image "App Store"	More then 200 community-contributed images of operating systems and software configured for a variety of life science applications.	You don't have to worry about installing software or finding obscure dependencies. Most images are ready-to-use.
Integrated with iPlant Authentication and Data	iPlant credentials are used to access the virtual machine instances and for SUDO functions. Atmosphere resources are co-local with the Data Store.	Instances and data are secure. You can control and grant access to instances and configure them with full administrative rights. Data transfers are quick and efficient.
Configurable Hardware Resources	Instances can be configured with specific hardware resources (e.g. RAM, disk space, and processor count).	You can access very powerful machines without the expense or setup time for alternative commodity-hardware solutions.
On-Demand Imaging	On request, an instance can be imaged and available in the image catalogue a private or public image.	You can use imaging as a way to "save your work" by preserving the entire configuration of your virtual machine, and you can use this as a way to publish software, data, and workflows.

How Atmosphere relates to our cyberinfrastructure

Atmosphere is one of the most versatile components of the iPlant CI. Anything that you would normally be able to do with your local laptop/desktop, you can do on a virtual machine in the Atmosphere cloud. The advantage of using Atmosphere is that you can get access to greater resources (currently up to 16 CPU, 128GB RAM machines). Additionally, those resources are co-localized with the iPlant Data Store so that moving to and from your instance is very easy to do.

When to use Atmosphere? When to use the Discovery Environment

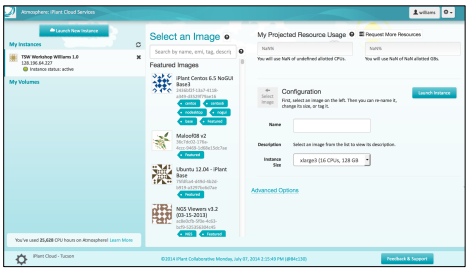
One of the most commonly asked questions is when to use which of these resources, the following recommendations help to explain advantages and limitations of each platform. The recommendations below will not cover every use case so post questions to ask.iplantcollaborative.org if you are unsure.

I Want to…	Recommendation	Why
Analyze data but I don't know Linux/command line	Discovery Environment (DE)	While some Atmosphere instances have easy-to-use Linux desktops, the DE is probably a better place to start
Use a bioinformatics application that has a GUI - graphical user interface – (e.g. CLC workbench)	Atmosphere	Software that does not have a command line interface is not suitable for the DE. You can install these software on an Atmosphere instance (assuming Linux compatibility – though in principle other OSs can be supported)
I need a great deal of memory, power, space, etc.	It depends on your application	Atmosphere makes it clear what hardware configuration you have. In the DE, resources are managed dynamically. In some cases an individual Atmosphere instance may have more power than what you would access in the DE. However, HPC Apps in the DE can be much more powerful that what Atmosphere provides. If you are unsure what you need – post to ask.iplantcollaborative.org
I want to use XXX software	It depends on your application	A list of applications installed in the DE can be found at: www.iplantcollaborative.org/apps1 You can search the list of Atmosphere images for your software. Remember, in both Atmosphere and the DE you can install your own programs.

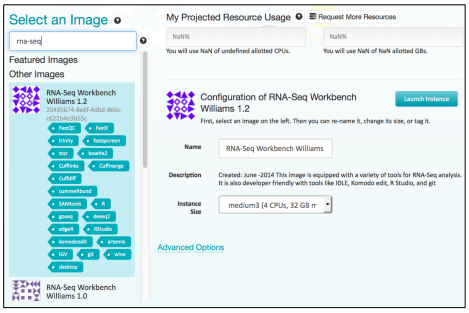
How to launch and connect to an Atmosphere instance

Creating an Atmosphere instance is like buying a new computer, you will have to select what you want and then customize it to suit your needs. Also like a new computer, your Atmosphere instance will generally come only with the listed software installed. You will have to connect that instance to your iPlant Data Store to import files. This guide will not cover all the use cases and features of Atmosphere (e.g. managing your allocation, requesting more resources, Imaging, and creating and mounting EBS volumes) See the Atmosphere page on the iPlant Learning Center.

> ***Tip**: To use Atmosphere, you must have an email address from an academic/governmental institution and request access to Atmosphere through the user portal. To request access, login to user.iplantcollaborative.org and check to see if Atmosphere is listed under 'My Services.' If it is not, scroll down and click the "Request Access" button next to Atmosphere to complete a request form.*



1. Login to Atmosphere (click the *Launch* link from the iPlant homepage at www.iplantcollaborative.org)
2. Click **Launch New Instance** either on the navigation panel (left) **or** on the home screen.
3. Under **'Select an Image'** select any image to use to (Note: Some images support a GUI Desktop and some are only acceesible through the shell – check the description and/or tags)



4. If desired, give your instance a name. Set the instance size to whatever if appropriate for your use (Note: you can not launch a size that would exceed your resource allocation or the actual available resources at any given moment)
5. Click **Launch Instance**. Your instance should be ready in 10-20 minutes.

Connecting to your Instance

Tip: Your Instance status must be '**active**' in order for you to connect.

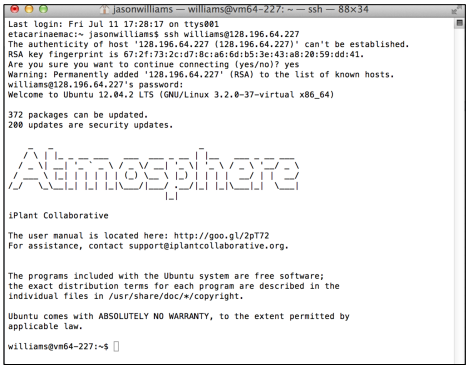
Connect via SSH

1. Locate your Instance IP address (beneath your instance name)

2. Open a terminal (Mac/Linux) and connect: **your_iplant_username@atmosphere.ip.address**

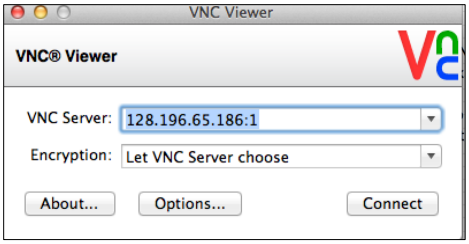
3. You will be asked to save and RSA key to the list of known hosts, enter 'yes'

4. When prompted, enter your iPlant password.



Connect via VNC

Tip: You may also try the VNC Tab within the Atmosphere interface itself – this requires you have Java installed and properly enabled.



1. Download VNC viewer (http://www.realvnc.com/download/viewer/)

2. Locate your Instance IP address (beneath your instance name)

3. Enter your IP address + ":1" in the 'VNC Server' field (e.g. 161.803.39.887:1) and click connect.

4. When connecting for the first time to an instance, you will be prompted to save a signature. Select **yes** and continue.

Tip: Once you connect to an Atmosphere instance, use iDrop (as described earlier in this guide) to transfer data to and from your new instance. See the iPlant online Learning

Note: Not all Atmosphere images are VNC enabled – check the image description to ensure it has Desktop/VNC support. Email support@iplantcollaborative.org if you have questions

Note: Once you have finished using Atmosphere, you should terminate the instance (this function appears in the Atmosphere home page under the 'Instance Details' for a particular instance. You can also terminate by click the **X** next to the instance name under 'My Instances.' Once you terminate an instance, all data will be lost – only terminate when you have saved your work elsewhere (e.g. to the iPlant Data Store).

How Different Scientists Might Use Atmosphere

Bioinformatics Users (Bench/Field Scientists)	• Learn how to use the shell and how to work with Linux • Master R to develop plots publication
Bioinformaticians	• Take advantage of root/SUDO access to fully customize a powerful machine • Developed a software suite with numerous R and Python library dependencies – update it regularly by making a new image.
Bioinformatics Engineers (Core Facilities)	• Link several Atmosphere instances with Apache Hadoop • Work with iPlant support to import existing Amazon images

Toolkit – Item Five: Strategies for Getting Help

People are a part of cyberinfrastructure - everyone will need help at some point

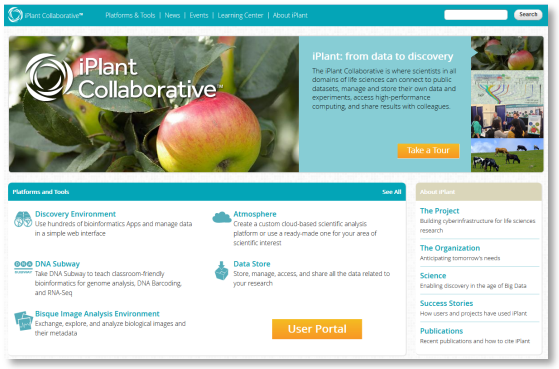
Hopefully, this guide has pointed out some ways to make use of iPlant as a part of your bioinformatics toolkit. Here are some additional resources that can get what you need.

Understand the iPlant Support System

Type of Request	Examples	Support Mechanism
Technical	Report a problem: iPlant tool, service or API is offline or not working.	Email Support
	Want your computational service to use some component of iPlant's infrastructure?	Submit Powered by iPlant request form
	A tool isn't working with the parameters I set.	Use iPlant forum "Ask iPlant"
Scientific	What's the best analysis for testing my hypothesis and what are the best tools to use?	
Technical and Scientific	Scale an algorithm Manage large-scale data Create a scalable workflow	Submit Extended Collaborative Support request form
	For groups and organizations: Develop new solutions to achieve your scientific goals.	Email Community Project Support

To get access to any of these forms visit the iPlant homepage or email support@iplantcollaborative.org

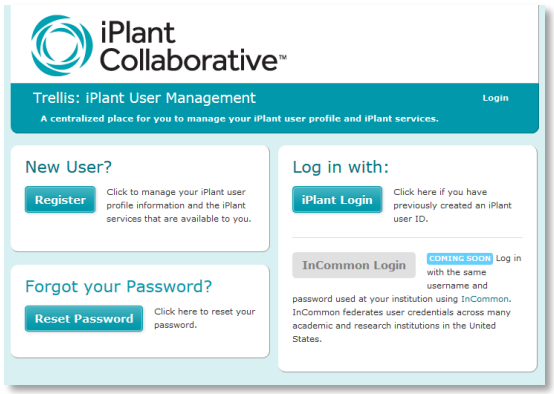
Resources you should know about



iPlant Collaborative Homepage

www.iplantcollaborative.org

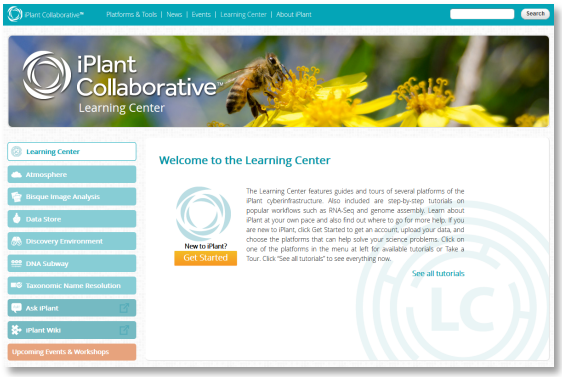
From the homepage you can learn about iPlant's mission, vision, and objectives (long-term or by quarter). You can also directly access our tools and platforms, learn about how to use iPlant platforms, and see how other persons, projects, and organizations make use of iPlant cyberinfrastructure.



iPlant User Portal

user.iplantcollaborative.org

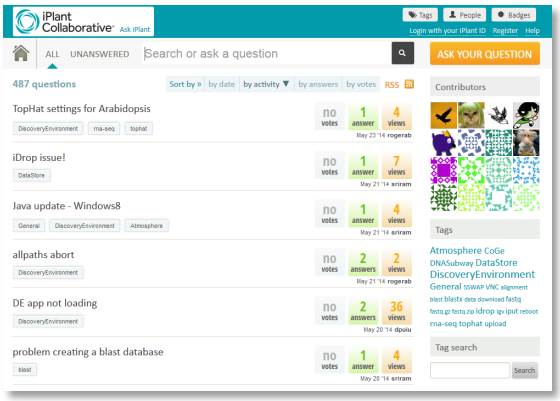
Create an iPlant account and also see what iPlant services you have access to. Once logged in, you can request access to services like **Atmosphere** from the user dashboard. From the user portal you can also reset your password.



iPlant Learning Center

www.iplantcollaborative.org/learning-center

The Learning Center features tours of the major iPlant platforms and tools as well as a growing list of science tutorials. If it is your first time using iPlant try the "Get Started" guide at **www.iplantc.org/getstarted**



ask.iPlant

ask.iplantcollaborative.org

This forum is a place for you to post any question about iPlant, from technical issues to questions and advice on a specific science objective. Any iPlant community member can help answer your questions, and the forum is monitored by iPlant support to ensure rapid responses.

How to Acknowledge iPlant

Please cite or acknowledge iPlant in any research that uses iPlant resources or extends the cyberinfrastructure. This may take the form of a citation, an acknowledgement, or both, as appropriate.

Acknowledging iPlant

The suggested format to acknowledge iPlant in a paper, a poster, or a presentation is:

This material is based upon work supported by the National Science Foundation under Grant Numbers #DBI-0735191 and #DBI-1265383. URL: www.iplantcollaborative.org

If you wish to additionally acknowledge an individual who assisted you from iPlant, the suggested format is:

We thank [consultant's name(s)] for [his/her/their] assistance with [describe the tasks accomplished], which was made possible through iPlant's Extended Collaborative Support program.

PIs should include a bibliography of articles or other manuscripts (published, accepted, submitted, or in preparation) that benefitted from use of iPlant resources as part of their annual Progress Report and Final Reports to their funding agencies.

Citing iPlant

If you would like to cite an iPlant publication, please cite this paper:

Goff, Stephen A. et al., "The iPlant Collaborative: Cyberinfrastructure for Plant Biology," Frontiers in Plant Science 2 (2011), doi: 10.3389/fpls.2011.00034.

For more information go to www.iplantcollaborative.org/citeus

Funding

The iPlant Collaborative is funded by the National Science Foundation under grant #DBI - 1265383.

References

Lynch C. Big data: How do your data grow? Nature. 2008;455(7209):28-9.

Pavelin K, Cham JA, de Matos P, Brooksbank C, Cameron G, Steinbeck C. Bioinformatics meets user-centred design: A perspective. PLoS Computational Biology. 2012;8(7).

Schadt EE, Linderman MD, Sorenson J, Lee L, Nolan GP. Computational solutions to large-scale data management and analysis. Nature Reviews Genetics. 2010;11(9):647-57.

Welch L, Lewitter F, Schwartz R, Brooksbank C, Radivojac P, Gaeta B, et al. Bioinformatics curriculum guidelines: Toward a definition of core competencies. PLoS Computational Biology. 2014;10(3).

Ranganathan S, Schönbach C, Nakai K, Tan TW. Challenges of the next decade for the asia pacific region: 2010 international conference in bioinformatics (InCoB 2010). BMC Genomics. 2010;11(SUPPL. 4).

Tools and Services Workshop: Additional Exercises

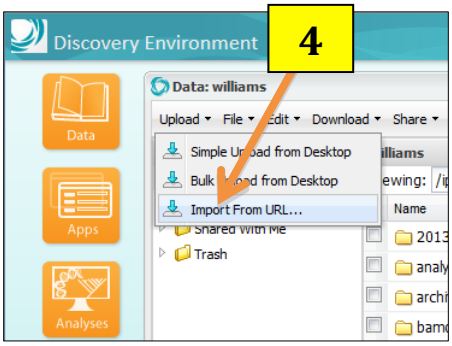
Data Store Exercises

Detailed Notes on the Wiki @: www.iplantc.org/ds1

Import a file into the DE from a URL

1. Follow your instructors' direction to choose an ftp link for import. You can select any link you like - here's one from Ensembl:

ftp://ftp.ensembl.org/pub/release-67/fasta/canis_familiaris/dna/
Canis_familiaris.BROADD2.67.dna_rm.chromosome.MT.fa.gz

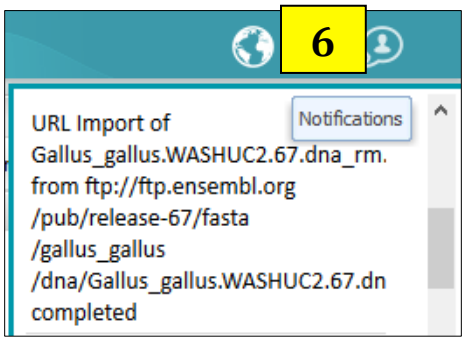


2. Click **Data** from the DE workspace.

3. Select your home directory from the directory tree (e.g. *your_iplant_username)* .

4. Click on the **Import from URL** icon from the **Upload** drop-down menu.

5. Paste the URL in the space provided using a keyboard shortcut (i.e Ctrl+V, Command+V). Delete any unnecessary spaces before, after, or within the URL, then click **Import from URL**.

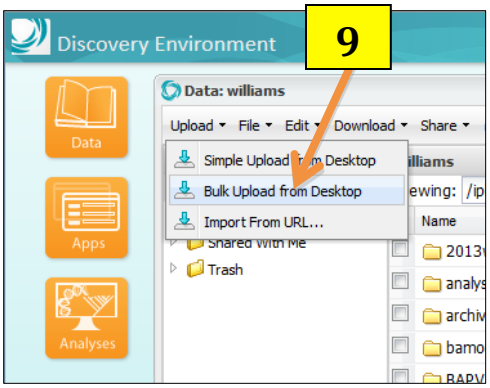


6. Click on **Notifications** in the DE workspace to monitor your notifications for the message that the upload is completed.

7. Click on **Data** from the DE workspace and check your home folder to confirm that you see the imported file.

Import a "large" file using iDrop lite in the DE

8. From the **Data** console, ensure you are somewhere in your home directory (where you have permission to write)



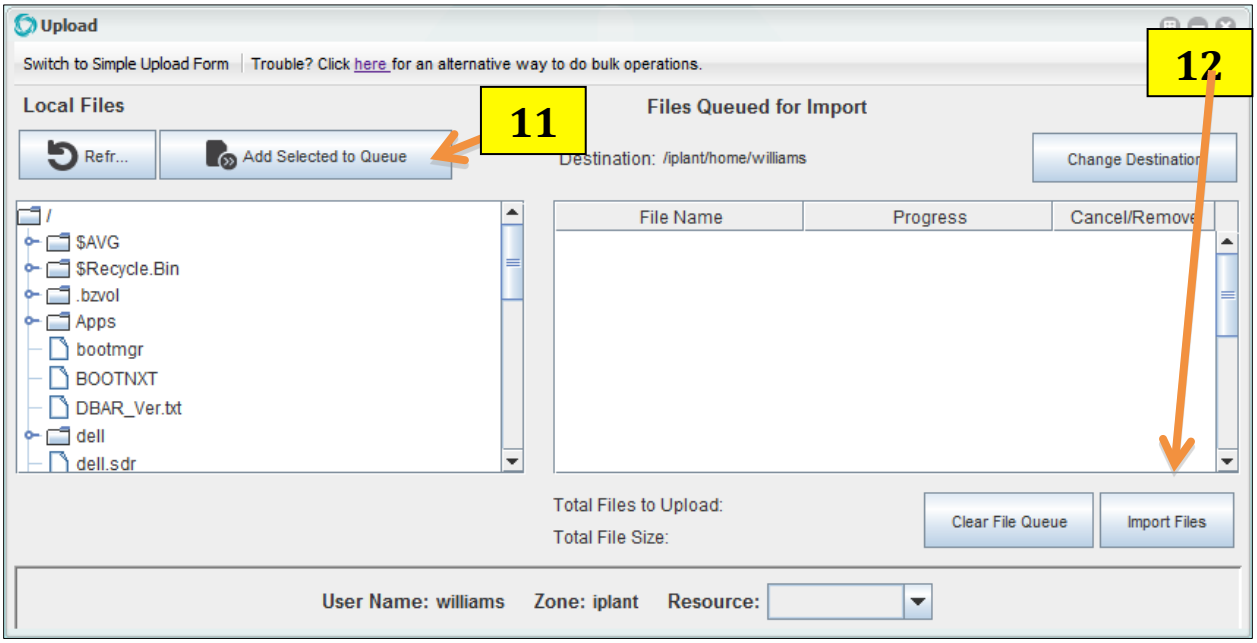
9. From the **Upload** menu select **Bulk Upload from Desktop**.

   Note: If prompted, give permission for Java (iDrop lite) to run. Selecting "Always trust content from this publisher" is recommended.

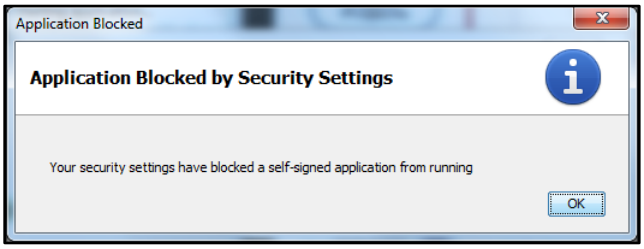
10. Under the **'Local Files'** column select a file to upload from your local file system. (*For this demo, a small file (10-20MB) is best. The process for large (3-4GB) files is exactly the same. You can use the file we downloaded in the DE demo if you choose.*)

11. Click **Add Selected to Queue** icon to add the file to your import list (*you can upload multiple files if you wish*).

12. Click **Import Files** to start the upload. When the import is complete you will get a pop-up notification. You may then close iDrop lite.



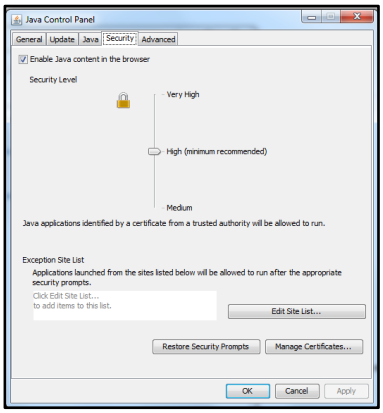
Notes on Java Compatibility



As iDrop Lite requires JAVA you must have Java installed (www.JAVA.com) and you may need to adjust your security settings.

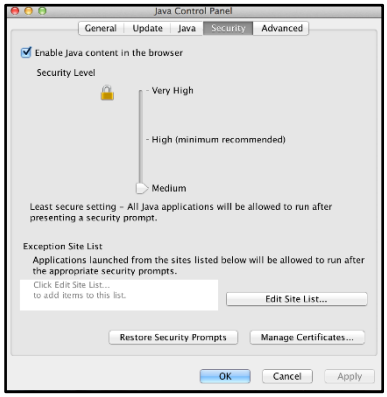
Windows:

Go to the Windows Control Panel: In the "Programs" section there will be the Java Control Panel (You can also use the windows search to find the Java Control Panel). In the Java security tab, lower the setting to medium and then apply (this may be readjusted later). Launch iDrop Lite again. (You may have to close your web browser to fully apply this setting)



Mac OS:

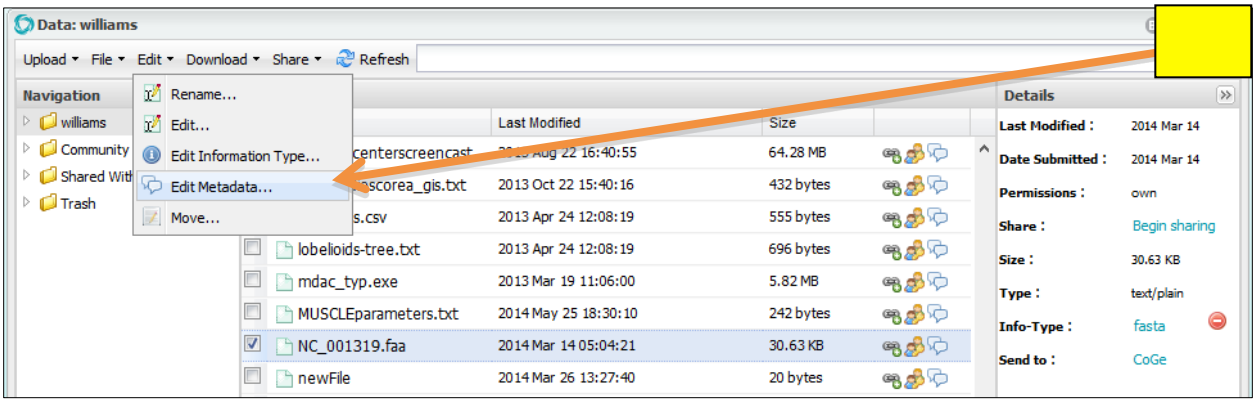
Go to the System Preferences Panel: There will be the Java control panel. In the Java security tab, lower the setting to medium and then apply (this may be readjusted later). Launch iDrop Lite again. (You may have to close your web browser to fully apply this setting)



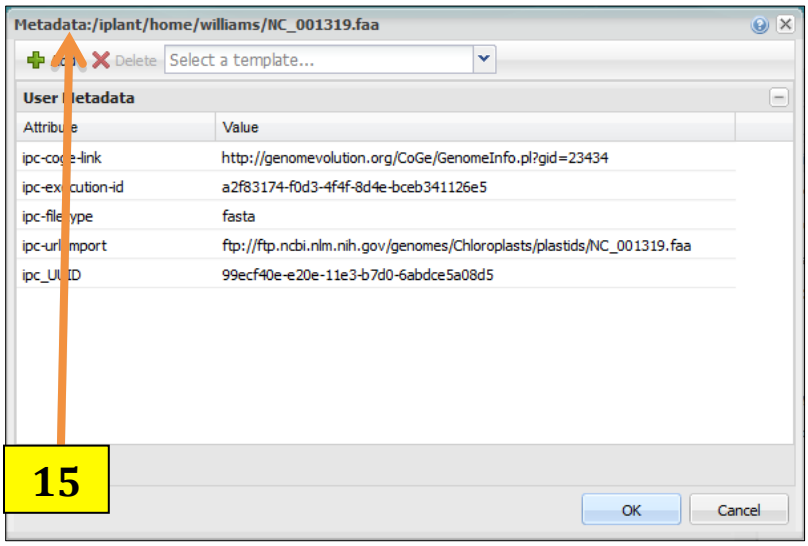
Managing and Adding Metadata

13. In the **Data** console, locate the file you imported from a URL in task 1.

14. Select the file and click the ( ) icon or click **Edit** to open a drop-down menu; click **Metadata**.



15. Click on the **Add** button and then enter information for the **Attribute** and **Value** categories, then click **OK**. You can repeat steps 14-15 for file from task 2 as well.



Using the DE to Examine Differential Expression with an RNA-Seq Dataset

Detailed Notes on the Wiki @: www.iplantc.org/rs1

Task 1: Align read data to the Arabidopsis genome using TopHat

1. Click on **Apps** from the DE workspace and select **TopHat2 – SE**. (Location: *Public Applications > NGS > Transcriptome Profiling >Tuxedo RNA-Seq 2*.)

2. Under "Analysis Name" leave the defaults or make any desired adjustments/entries for the analysis name, description, or output folder.

3. Under **Input data** for **FASTQ file(s)** use the "Add" button to browse and select each of four FASTQ files located under *Community Data > iplant_training > intro_rna-seq and select the files in the folder 01_ input_data*. Alternatively you can click and drag these files from the data folder into the Input data window

4. Under **Reference Genome (Mandatory)** for 'Select a reference genome from the list' select **Arabidopsis thaliana [mouse-ear cress] (Ensembl 14)** (*Note: This is equivalent to the TAIR 10 release)*

5. Under **Reference Annotations** select **Arabidopsis thaliana [mouse-ear cress] (Ensembl 14)**; click **Launch Analysis.**

6. Click on **Analyses** from the DE workspace to monitor the status of your job. You will also receive notifications.

7. When your job is completed click the job name in the **Analysis** console or navigate to the output in our **Data** directory. In the **tophat_out** folder created you should verify you have created 5 folders; a 'bam' folder and one folder for each of the wild type/hy5 reads.

Task 2: Assemble transcripts using Cufflinks

8. Click on **Apps** from the DE workspace and select **Cufflinks2**. (Location: *Public Applications > NGS > Transcriptome Profiling >Tuxedo RNA-Seq 2.)*

9. Under "Analysis Name" leave the defaults or make any desired adjustments/entries for the analysis name, description, or output folder.

10. For **Input Data** section for **SAM/BAM file(s)** use the "Add" button to add the bam files created by TopHat in the previous analysis. In the **'bam'** folder (See step 7) add all four **.bam** files (*hy5_rep1.bam*, *hy5_rep2.bam*, *WT_rep1.bam*, *WT_rep2.bam*). These files are also available in the Community Data folder (*Community Data > iplant_training > intro_rna-*

seq > *02_tophat* > *bam*). For convenience, a batch of TopHat bam files can be analyzed together but these files can also be processed concurrently in independent Cufflinks runs.

11. Under **Reference Annotations** under 'Select Reference Genome Annotation' we select the same genome build that we used in the TopHat assembly: **Arabidopsis thaliana [mouse-ear cress] (Ensembl 14)**; click **Launch Analysis**.

12. Click on **Analyses** from the DE workspace to monitor the status of your job. You will also receive notifications. When your job is completed click the job name in the **Analysis** window or navigate to the output in our **Data** directory. In the cufflinks output folder that is created you should find folders for each replicate as well as a folder called **gtf**.

13. In the other folders created by Cufflinks (e.g. hy5_rep1) you should find GTF and FPKM files. Click on **transcripts.gtf** to view annotated transcripts with their release annotations.

14. In the same folder (hy5_rep1) click on the **genes.fpkm_tracking** file to preview coverage expressed in fragments per kilobase of exon per million mapped reads.

Task 3: Merge all assembled transcripts into a single transcriptome annotation file with Cuffmerge

15. Click on **Apps** from the DE workspace and select **Cuffmerge2**. (Location: *Public Applications > NGS > Transcriptome Profiling > Tuxedo RNA-Seq 2.)*

16. Under "Analysis Name" leave the defaults or make any desired adjustments/entries for the analysis name, description, or output folder.

17. For **Input Data** under **GTF files to merge** use the four files in the **gtf** folder created by the cufflinks analyses (see step 12), (e.g. *hy5_rep1_transcripts.gtf*, *hy5_rep2_transcripts.gtf*, *WT_rep1_transcripts.gtf*, *WT_rep2_transcripts.gtf*). These files are also available in the Community Data folder (*Community Data* > *iplant_training* > *intro_rna-seq* > *03_cufflinks* > *gtf*).

18. Under **Reference Data** under 'Select Reference Genome Annotation' we select the same genome build that we used in the TopHat assembly: **Arabidopsis thaliana [mouse-ear cress] (Ensembl 14)**

19. Click **Launch Analysis**. Name your job (e.g. **Cuffmerge**), add a description if desired, and click **OK**.

20. Click on **Analyses** from the DE workspace to monitor the status of your job. You will also receive notifications. When your job is completed click the job name in the **Analysis**

console or navigate to the output in our **Data** directory. In the folder that is created you should find 7 files.

Task 4: Compare expression using CuffDiff

21. Click on **Apps** from the DE workspace and select **CuffDiff2**. (Location: *Public Applications* > *NGS* > *Transcriptome Profiling* > *Tuxedo RNA-Seq 2.*)

22. Under "Analysis Name" leave the defaults or make any desired adjustments/entries for the analysis name, description, or output folder.

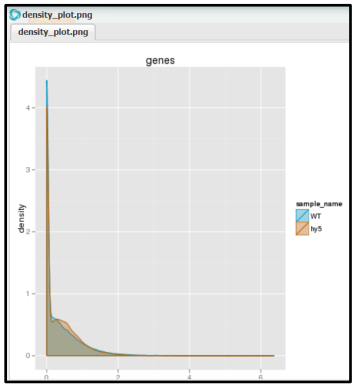
23. Under **Input Data** for **Sample 1** enter **WT** for 'Sample 1 Name.' Then add (or drag) bam files from the two wild type replicates (*WT_rep1.bam*, *WT_rep2.bam*) in **bam** from the Tophat run (Task 1 Step 7). (*WT_rep1.bam*, *WT_rep2.bam*). These files are also available in the Community Data folder (*Community Data* > *iplant_training* > *intro_rna-seq* > *02_tophat* > *bam*).

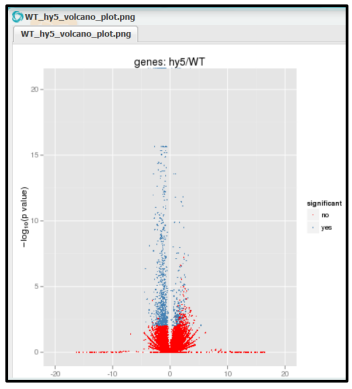
24. Under **Input Data** for **Sample 2** enter **hy5** for 'Sample 2 Name.' Then add (or drag) bam files from the two HY5 replicates(*hy5_rep1.bam*, *hy5_rep2.bam*) in **bam** from the Tophat run (see Task 1 Step 7). These files are also available in the Community Data folder (*Community Data* > *iplant_training* > *intro_rna-seq* > *02_tophat* > *bam*).

13. Under **Reference Annotations** under 'Custom annotation File' browse for the **merged_with_ref_ids.gtf** created from **cuffmerge** (See Task 3 step 20). This file is also available in the Community Data folder (*Community Data* > *iplant_training* > *intro_rna-seq* > *04_cuffmerge* > *cuffmerge_out*); click **Launch Analysis**.
14. Click on **Analyses** from the DE workspace to monitor the status of your job. You will also receive notifications. When the job is completed, click on the job name to navigate to the job output.

*In the **cuffdiff_out** folder you will see a number of outputs which are described in documentation (www.iplantcollaborative/rs1) including **gene_exp.diff** which compares expression between the two samples (**WT** and **hy5**). The **graphs** folder contains a few automatically generated plots using **cummeRbund**, part of the cufflinks RNA-Seq workflow.*

Example results and plots





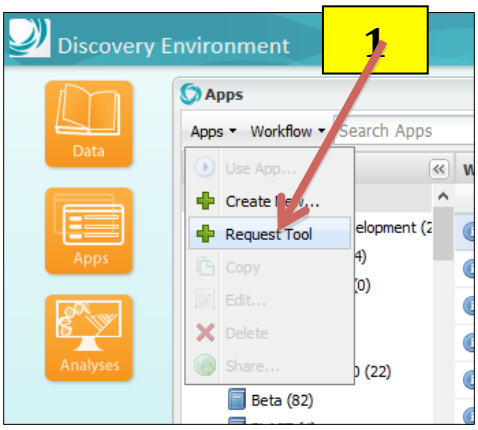
iPlant Tool Integration within the DE

Detailed Notes on the Wiki @: www.iplantc.org/ti1

Note: In order to integrate your own tools within the DE, you will need to have support (support@iplantcollabortive.org) install the tool. It will also be helpful to have tested your tool(s) and dependencies; atmosphere is a good resource for this!

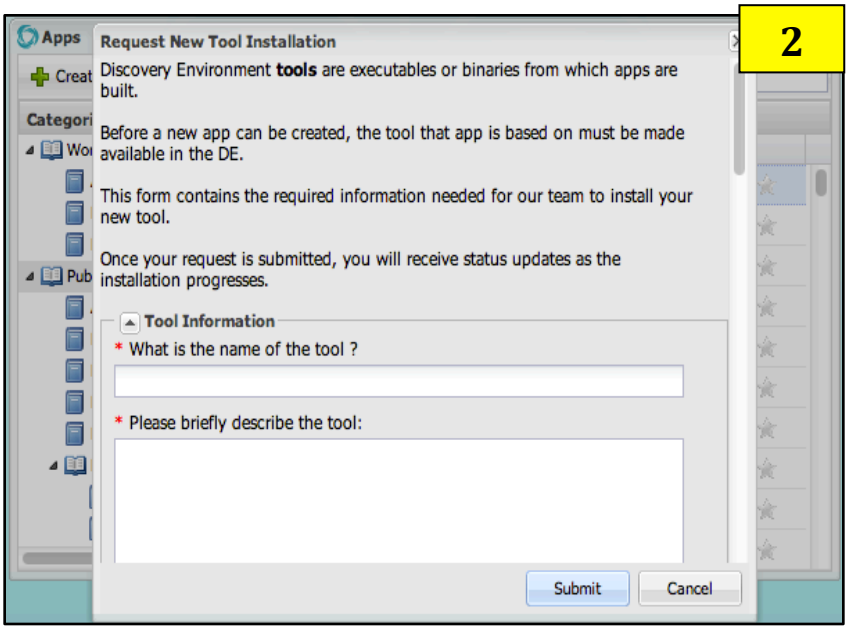
Task 0 (pre-requisite for custom installations): Deploy your app on condor

1. Click on **Apps** from the DE workspace, then select the **Apps** menu and select **Request Tool**.



2. Confirm the information provided under 'Tool Information' and 'Other Information' tab fill out the form and upload the source code and test files.

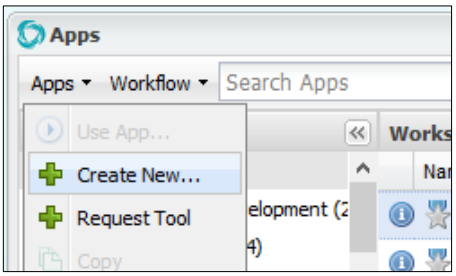
   You will need to provide several items including
   - Tool description and version
   - Tool source/binary
   - URL for tool documentation
   - Test data



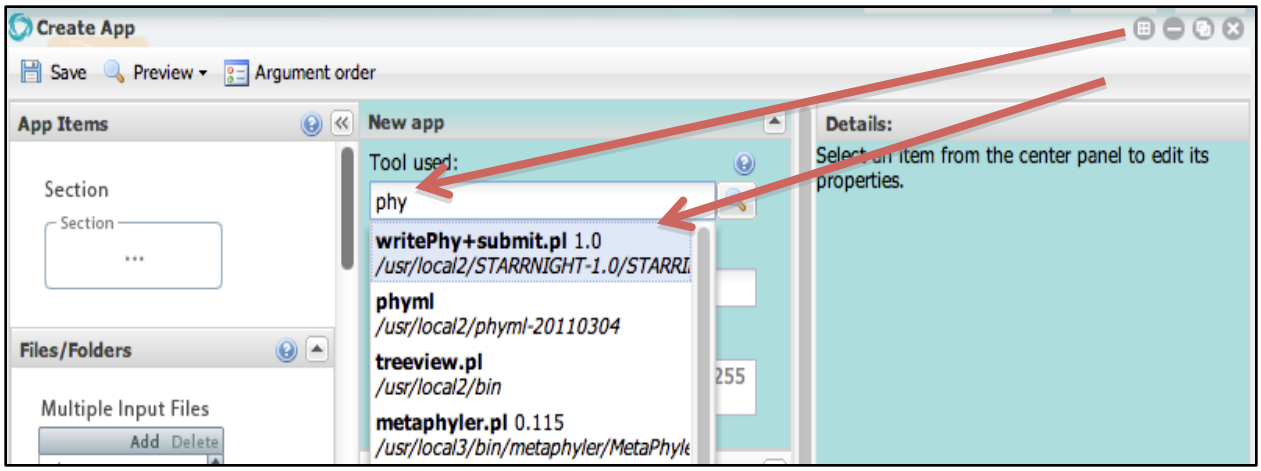
3. Once you are done filling out the form, click **Submit**. You can proceed to task 1, or if you don't have all the requirements met now for the installation, you can proceed with the description of your tool in task 1.

Task 1: Describe your app

4. Click on **Apps** from the DE workspace. Under **Create** select 'New App.' *This opens the Tool Integration console.*



5. Under 'Tool Used' click the search icon and search for **phyml**. Click **OK. You may also search for your tool by typing the name into the field.**

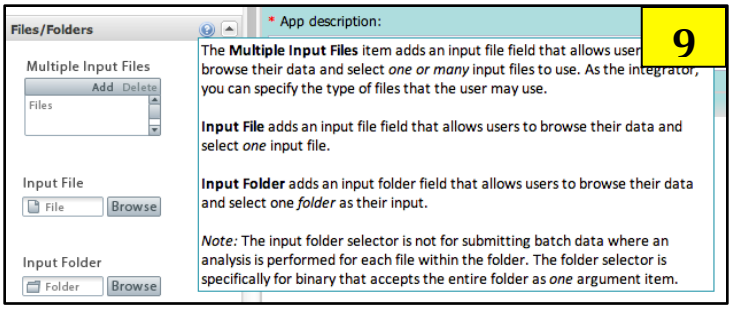


6. Enter a name and description for your app under the 'App Name' and 'App description' fields respectively. *Note that these fields are marked as required.*

Task 2: Configure arguments for your app

7. Select the header titled 'Section 1'. This will update the contents of the 'Details' panel on the right.
Selecting the header also causes the section to collapse, click it one more time to expand it.

8. In the 'Details' panel under 'Section name', change the name to "**Select Inputs**" and hit ENTER. *You should see the section's header, and the 'Details' panel header update with the new name. The left side of the editor contains draggable items, organized into groups, which can be added to your app. Each group has a contextual help icon that describes the items within the group.*



9. Select the contextual help icon in the 'Files/Folders' group on left side of the editor.

   This pops up a description of all of the items within the 'Files/Folders' group.

10. Click and drag the 'Input File' item into the 'Select Inputs' section and drop it.

11. As you drag the item, you will receive a visual indicator which lets you know if you can successfully drop the item (green circle with a white checkmark, otherwise a red circle with a slash).

   When a new argument or section is added to the app, it will be automatically selected, which causes the 'Details' panel to update with the selected item's corresponding details panel.

12. In the 'Details' panel under 'File Selector label' change the name to "**Select PHYLIP interleaved MSA file**". *(MSA: multiple sequence alignment)*

   Again, notice that the center panel updates with the change (after you hit ENTER or click another field).

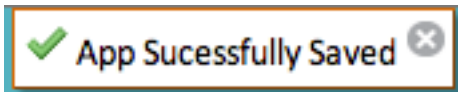
13. Check the 'Make this field required.' checkbox.

   This will prepend a red "" to the argument's label to indicate that the argument requires user input. If a section contains any argument which is marked as required, it will also be prepended with an "*".*

14. For 'Argument option' enter "**-I**".
   The change will be reflected in the "Command line view" at the bottom of the editor.

15. Under 'Tool tip text' enter "**Please select one PHYLIP interleaved multiple sequence alignmen**t". *You should see a contextual help icon appear on your argument in the center panel. Hover over it, and you should see the tool tip you just entered. The tool tip text also supports simple html markup, as you can see in the image with the bolded '**PHYLIP**'.*

16. Under 'Type of information contained in this file' select **Multiple Sequence Alignment**.
   Click 'Save' in the top left of the panel.
   If successful, you should see a banner with a green checkmark displayed at the top of the browser window that states "App Successfully Saved".



17. Next, add a new section by dragging the 'Section' icon from the left panel to the center.
   As you add new sections by dragging them to your app, all sections of the app are automatically collapsed.

18. Under 'Section name' in the 'Details' panel, name the section "**Describe sequence type**".

19. Add a list selector argument to the "Describe sequence type" section by dragging the 'List' icon into the "Describe sequence type" section. *The 'List' icon may be found within the 'Lists' group on the left side of the screen.*

20. Under 'List label' enter "**Select the sequence type contained in your file**".

21. Click the 'Edit list' button in the 'Details' panel. *This will open a popup titled 'Edit list' for editing your list.*

22. Click the 'Add' button in the 'Edit list' popup. This will insert a new row.

23. Double click on the new row to begin editing. Then enter the following values:

    a. Under 'Display' enter "**Nucleotides**"
    b. Under 'Argument' enter "**-d**"
    c. Under 'Value' enter "**nt**"

    'Display' is what users will see as options, 'Argument' and 'Value' are the command line parameters the program is expecting.

24. When done, click the 'Save' button below the row to save your changes.

25. Add another row and enter the following values:
    a. Under 'Display' enter "**Amino Acids**"
    b. Under 'Argument' enter "**-d**"
    c. Under 'Value' enter "**aa**"

26. When finished editing your list, click 'Done' on the 'Edit list' popup.

27. Select a default item from your list by clicking the down arrow under 'Default item to display' in the 'Details' panel and selecting **'Nucleotides**'. *You may also select the default item from the argument in the center panel.* Click 'Save'.

Task 3: Preview how your app will appear in the DE and Order Commands

28. Click **Preview** and select 'Preview App'.

    The App preview will be nearly identical to what is displayed in the center panel of the editor. You should also notice that there is a 'Launch Analysis' button. Clicking this button inside the app preview window does not launch the app. It is provided as a means to test any validations you've designed into your app. If there are any errors in the app, they should be displayed after clicking 'Launch Analysis'.

In the app we've designed there is one validation that will be flagged if not met, the file selector was marked as required. If there is no file selected, the app should produce an error. Furthermore, if a section contains a number of arguments with validation errors, they will be summarized in the section title. Hovering over the red exclamation point will provide a tool tip containing the validation errors. This is demonstrated in the following image:

29. Click **Argument line order**

30. Drag "**-i (Input)"** from 'Unordered Argument to the first position in the 'Order | Argument' column.

31. Drag **"Select the sequence type contained in your file"** from 'Unordered Argument to the second position in the 'Order | Argument' column.

32. Click **Done**. *Notice the command line preview is now filled in under 'Command line preview.'*

33. Click **Save** at the top left of the panel.

34. You should get a success message.

Task 4: Test your app

35. In your workspace, run your new PhyML on the file contained in this path: **Community Data > iplant_training > tool_integration > phylip_interleaved.aln**. *You should be able to create a newick string containing the data for your phylogenetic tree as output. e.g ((((((((Tomato-Solanaceae:0.04699055,Celery-Apiaceae:0.03695472)…"*