



Accelerating Plant Science with CyVerse

Genomics Workflows, Data Management, and Training

Jason Williams – Lead, CyVerse – Education, Outreach, Training
Cold Spring Harbor Laboratory

williams@cshl.edu [@JasonWilliamsNY]



Slides and Tutorials



www.cyverse.org/sol2016



CyVerse Evolution



CyVerse 2016

Transforming Science
Through Data-Driven
Discovery



Vision:

Transforming science through data-driven discovery

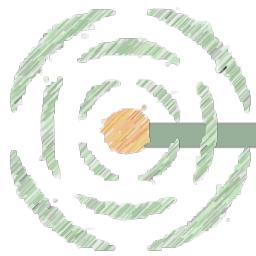
Mission:

Design, develop, deploy, and expand a national
cyberinfrastructure for life science research, and train
scientists in its use

**More than 30K users, PB of data, and hundreds of
publications, courses, and discoveries**



CyVerse Evolution



iPlant 2008
Empowering a New
Plant Biology



iPlant 2013
Cyberinfrastructure for
Life Science



CyVerse 2016
Transforming Science
Through Data-Driven
Discovery



CyVerse is Built for Data



Plant / Microbial



Animal



Biomedical



Ecological/Climate

CyVerse supports all domains of life science

CyVerse Evolution



DBI-0735191 and DBI-1265383

We are funded by the National
Science Foundation

- We are your colleagues and collaborators!
- \$100 Million in investment
- Freely available to the community
- Spur national/international collaboration
- Cite CyVerse:
CyVerse.org/acknowledge-cite-cyverse



What is Cyberinfrastructure?



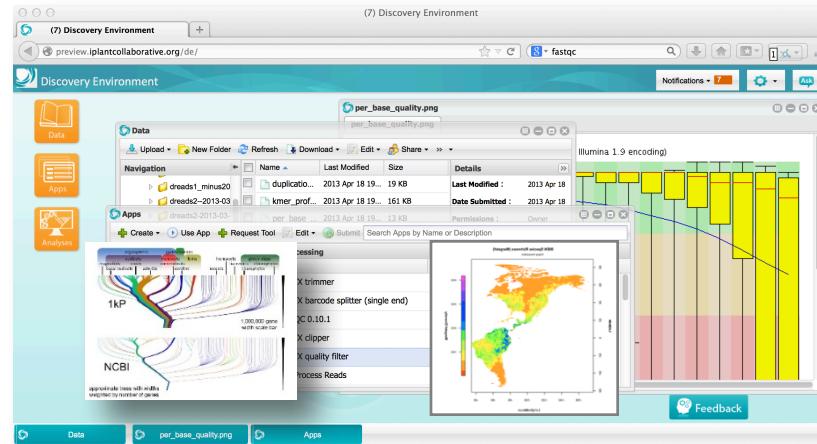
Cyberinfrastructure for Data-intensive Biology

- Data storage
- Software
- High-performance computing
- People

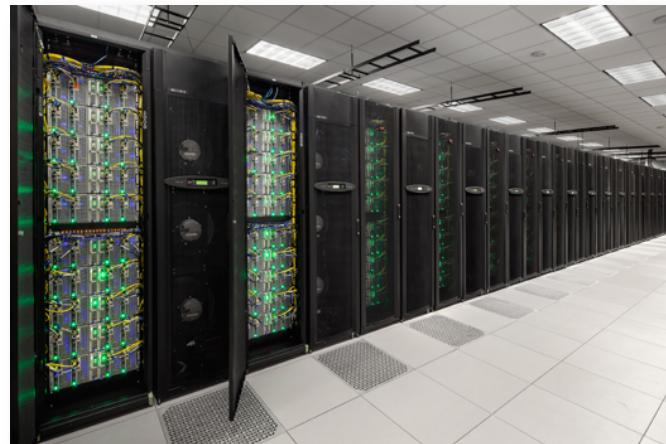
organized to solve problems of size and scope not otherwise solvable.



Cyberinfrastructure for Data-intensive Biology



Platforms, tools, datasets



Storage and compute



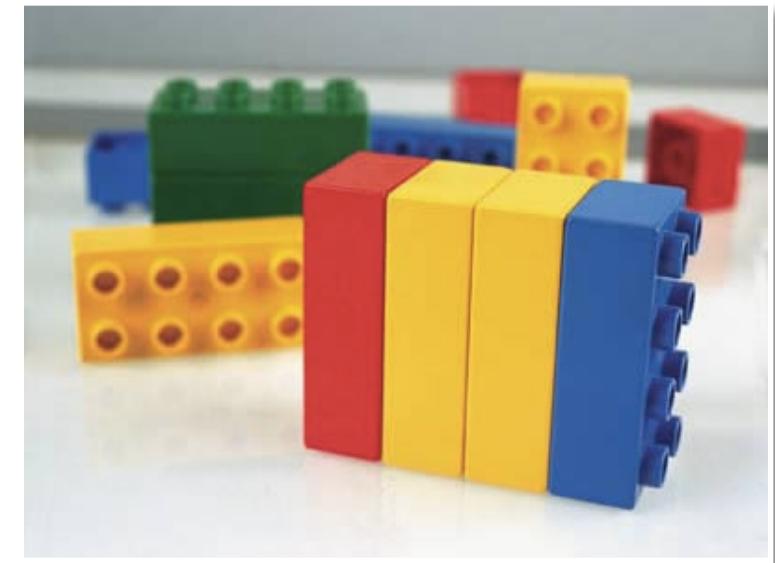
Training and support

What is available in CyVerse?



CyVerse Products

- We strive to be the **CI Lego blocks**
- Danish 'leg godt' - **'play well'**
- Also translates as '**I put together**' in Latin
- If a solution is not available you can craft your own using CyVerse CI components



CyVerse Features



Get Science Done

- Ability to access and manage data
- Software to analyze data
- Computing resources
- Skills and help to use software and interpret results

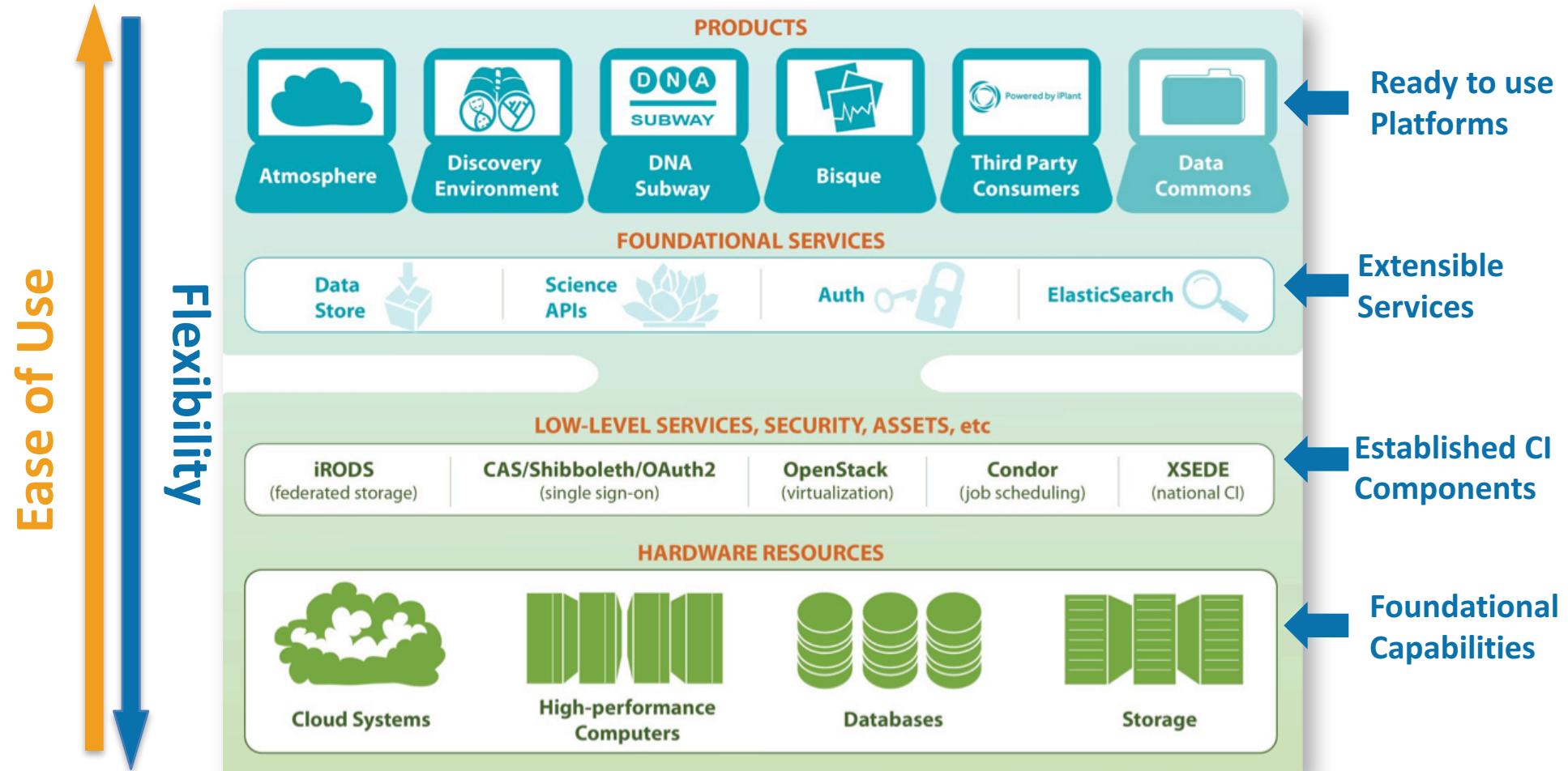
Ensure Reproducibility

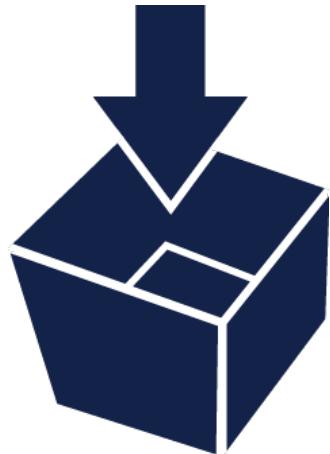
- Metadata management
- Ability to share data and workflows
- Open source sustainable tools

Increase Productivity

- High-performance and scalable computing
- Ability automate and collaborate
- Funding spent on science, not software or hardware

CyVerse product stack





Data Store

The resources you need to share and manage data with your lab, colleagues and community

- ✓ Initial 100 GB allocation – TB allocations available
- ✓ Automatic data backup
- ✓ Easy upload /download and sharing

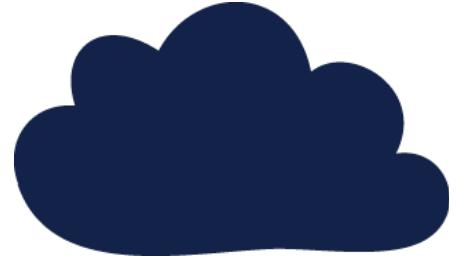




Discovery Environment

Hundreds of bioinformatics Apps in an easy-to-use interface

- ✓ A platform that can run almost any bioinformatics application
- ✓ Seamlessly integrated with data and high performance computing
- ✓ User extensible – add your own applications



Atmosphere

Cloud computing for the life sciences

- ✓ Simple: quick access to hundreds of virtual machine images
- ✓ Flexible: fully customize your software setup
- ✓ Powerful: integrated with CyVerse computing and data resources





Science APIs

Fully customize CyVerse resources

- ✓ Science-as-a-service platform
- ✓ Define your own compute, and storage resources (local and *CyVerse*)
- ✓ Build your own app store of scientific codes and workflows





DNA Subway

Educational workflows for Genomes, DNA Barcoding, RNA-Seq

- ✓ Commonly used bioinformatics tools in streamlined workflows
- ✓ Teach important concepts in biology and bioinformatics
- ✓ Inquiry-based experiments for novel discovery and publication of data





Bisque

Image analysis, management, and metadata

- ✓ Secure image storage, analysis, and data management
- ✓ Integrate existing applications or create new ones
- ✓ Custom visualization and image handling routines and APIs



How can you do science in CyVerse?





CyVerse Data Commons

Making data discoverable and reusable

Data Store architecture is built for metadata management

- Data within CyVerse maintains provenance and can be labeled using a variety of metadata templates or using custom schemas
- Direct submission of sequencing data to NCBI Sequence Read Archive
- Request DOIs and ARKs (Archival Resource Key) for permanent data identification
- Support for elastic search, data sharing, and anonymous access

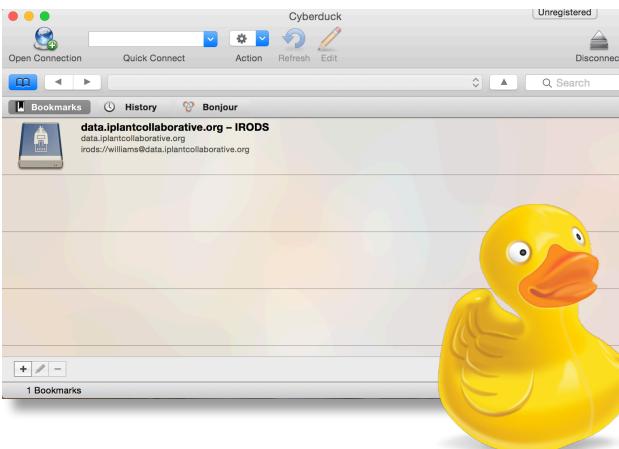




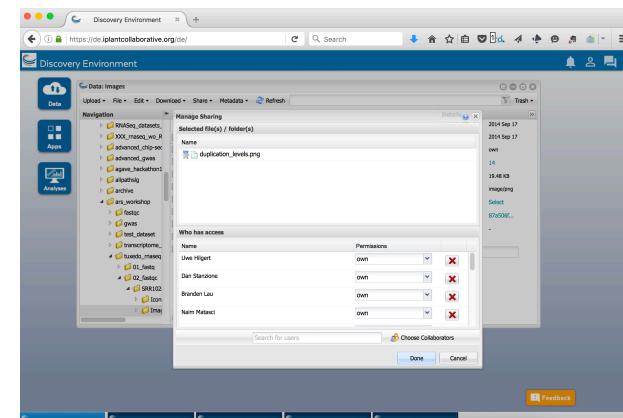
CyVerse Data Commons

Supporting the lifecycle of data

Simple upload



markup, search, share



publish

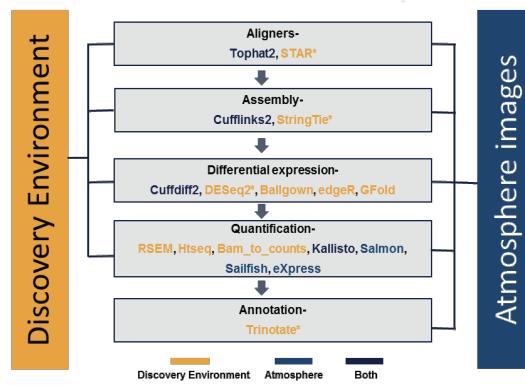
A screenshot of the CyVerse Data Commons interface. The top navigation bar includes "CyVerse Home", "Data Store", "Discovery Environment", "Atmosphere", and "About". The main content area shows a "shared" folder with various items like "LigninSystemsDB", "aangs", "common_rpo", "cshl_nps.edu", "earthenv_demo_data", "gbs", "Minus", "imicrobe", "iplant_cdp_sample_data", "iplant_kobas", "iplant_public_test", "iplant_training", "iplant_uk_training", "iplantcollaborative", "legume_wgd", "onemap_pilot", "panzea", "plant_phenotype_ontology_dar", and "publictemp". On the right, there's a "SRA" section with "Sequence Read Archive (SRA) makes biological sequence data available to the research community to enhance reproducibility and allow for new discoveries by comparing data sets. The SRA stores raw sequencing data and alignment information from major sequencing centers around the world, including the SRA Systemic, Helicos Helicosphere, Complete Genomics, and Pacific Biosciences SMRT." Below it are sections for "Getting Started", "Tools and Software", "Related Resources", and "Help Desk".



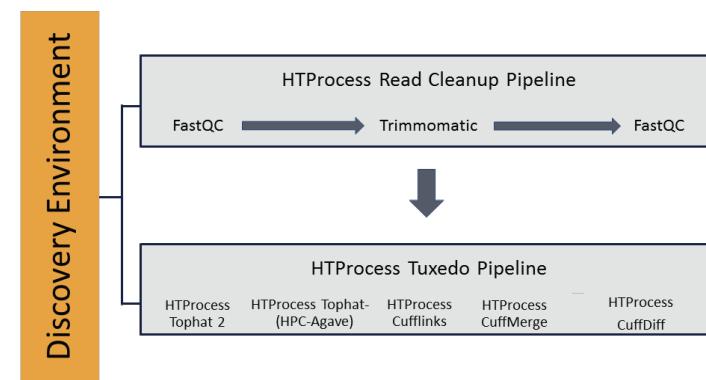
Discovery Environment

End-to-end workflows for genomics

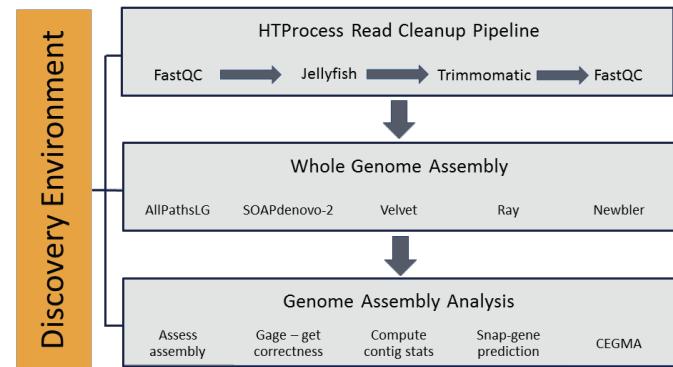
RNA-Seq 1: Differential Expression



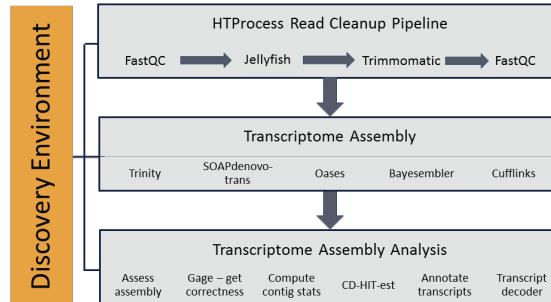
RNA-Seq 2: High-throughput



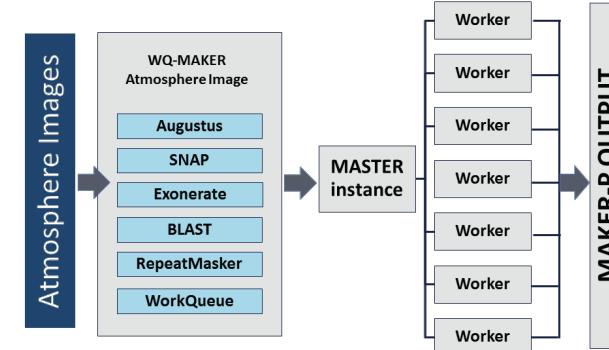
Genome Assembly



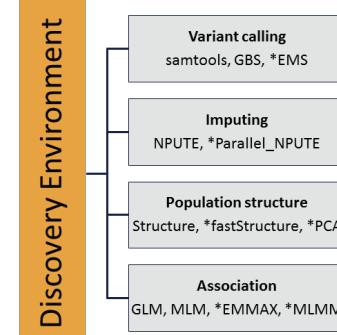
Transcriptome Assembly



WQ-Maker Annotation



Association Studies

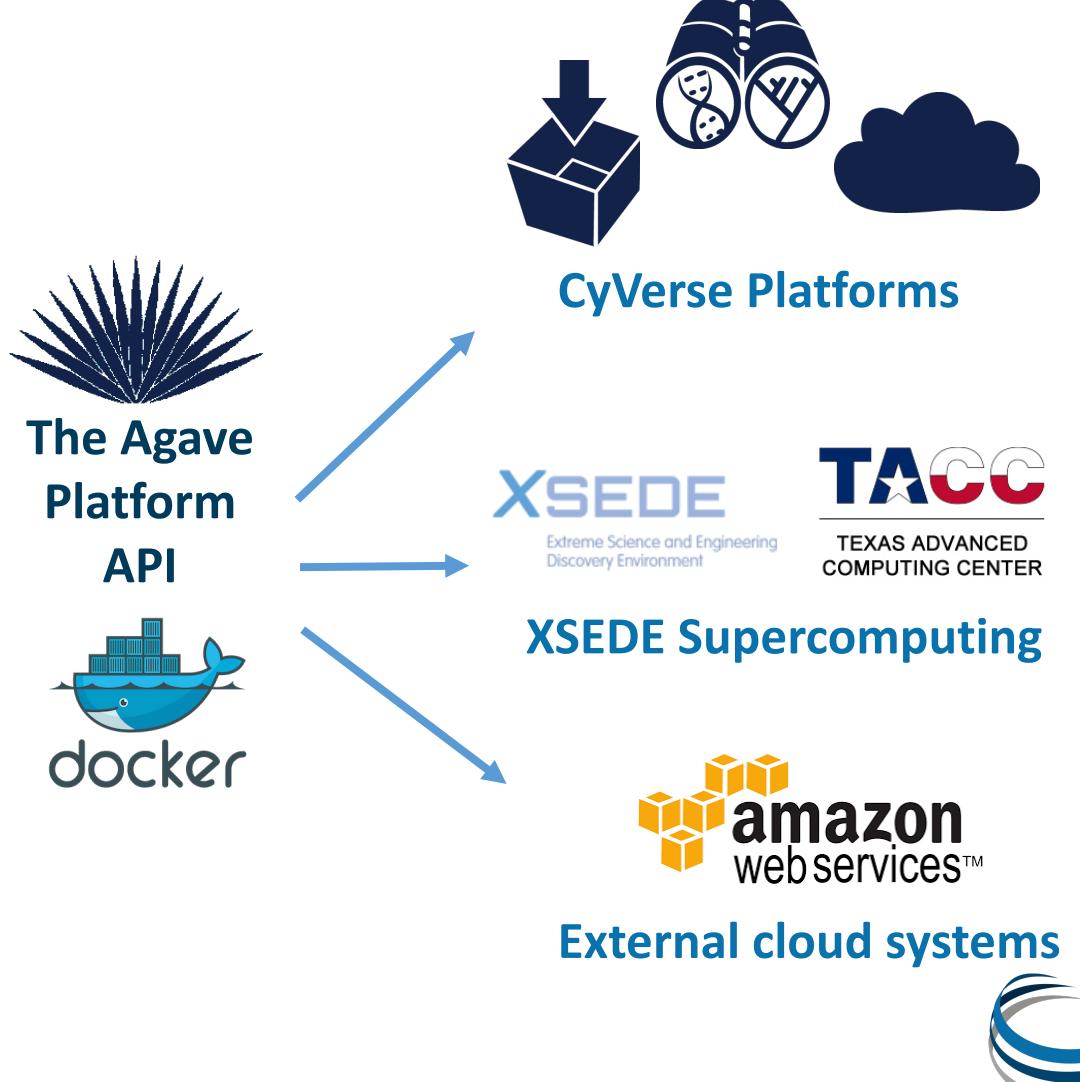




Discovery Environment

User friendly, developer friendly

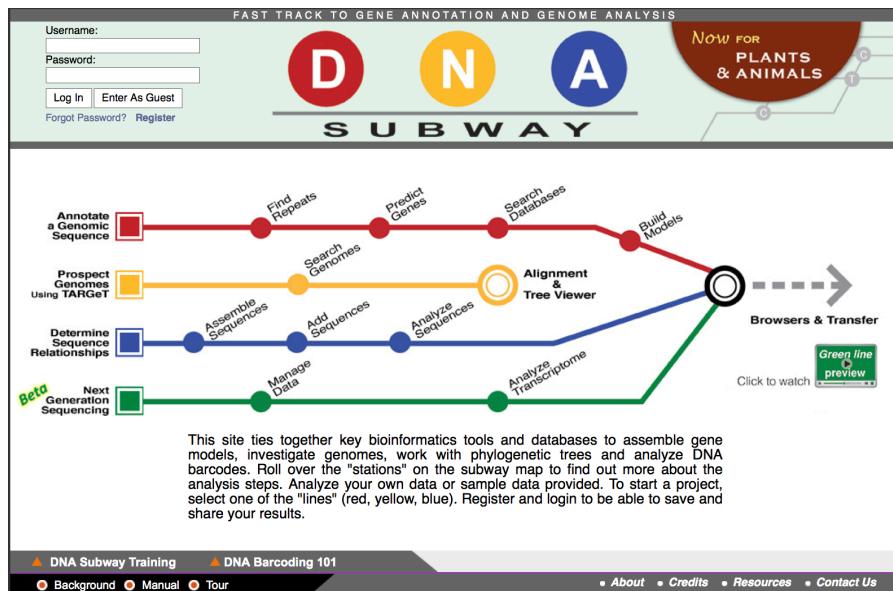
The screenshot shows the Discovery Environment web interface. On the left, there's a file browser window titled "Data: Images" showing files like "density_levels.png", "kmer_profiles.png", and "WT_hy5_volcano_plot.png". In the center, there's a "Tabular View" of a dataset named "gene_exp.diff" with columns: test_id, gene_id, gene, locus, sample_1, sample_2, status, value_1. Below it is a "density_plot.png" showing density vs log10(fpkm). To the right is a "WT_hy5_volcano_plot.png" showing -log10(p value) vs log10(fold change).



DNA SUBWAY

DNA Subway

Classroom applications



RNA-Seq for the Next Generation

- Classroom-friendly RNA-Seq workflows
- More than 900Gbs Student-faculty projects analyzed in 36 genomes
- Online educational materials



- DNA Barcoding workflow
- ~10,000 student-analyzed barcodes and 125 novel NCBI submissions
- Online educational materials

How can I use CyVerse?



Hands-on training

Workshops that reach users where they are

Tools & Services



- Two days; targeted to researchers
- Hands-on learning modules tailor to interests
- Individual consultations
- 1026 participants since 2011

Genomics in Education



- Two days; targeted to educators
- Pair bioinformatics with classroom labs
- Help for generating lesson plans
- 748 participants since 2010

Online learning materials

Documentation, tutorials, and help

Learning Center

The screenshot shows the 'WELCOME TO THE LEARNING CENTER' page. It features a sidebar with 'PRODUCTS' (Products Overview, Atmosphere, BioQue Image Analysis Environment, Data Store, Discovery Environment, DNA Subway, Learning Center, Powered by CyVerse, Science APIs) and 'LEARNING CENTER' (Learning Center, Create Account). The main content includes sections for 'ABOUT THE LEARNING CENTER' (describing its purpose), 'KEY FEATURES' (such as account creation, email management, and platform usage), and 'TUTORIALS'.

Wiki

The screenshot shows a CyVerse Wiki page titled 'Getting Started with CyVerse'. It includes a 'UPCOMING MAINTENANCE TUESDAY MARCH 28' section with details about system downtime from 7:00 AM to 5:00 PM MST. Below this, there's a summary of what will be affected (CyVerse User Portal, Discovery Environment, Atmosphere, BioQue, and Data Store) and a note about BiQue processing being unavailable. The page also contains a 'Getting Started with CyVerse Home' section and a 'Learn the basic and overviews on the website' section.

Forum

The screenshot shows the 'Ask iPlant' forum homepage. It displays several threads: 'Can I use the iPlant Data Store to get a genome into CoGe?' (3 votes, 1 answer, 63 views), 'Incomplete transfer of results from TACC back to DE' (3 votes, 1 answer, 37 views), 'How to make an App public?' (2 votes, 3 answers, 43 views), 'Problems with uploading data' (2 votes, 5 answers, 103 views), and 'Error installing R package' (2 votes, 5 answers, 57 views). The interface includes a search bar and navigation links for 'ALL' and 'UNANSWERED' posts.

- cyverse.org/learning-center
- Tutorials and quick-starts

- wiki.cyverse.org
- Documentation

- ask.iplantcollaborative.org
- Community help/Q&A





CYVERSE™

Transforming Science Through Data-driven Discovery

CyVerse Executive Team



THE UNIVERSITY
OF ARIZONA®

Parker Antin
Nirav Merchant
Eric Lyons



TEXAS ADVANCED
COMPUTING CENTER

Matt Vaughn



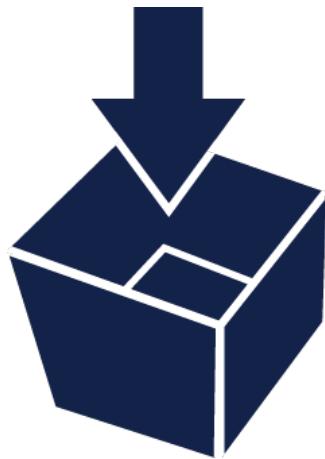
Cold
Spring
Harbor
Laboratory

Doreen Ware
Dave Micklos

CyVerse is supported by the National Science Foundation under Grant No. DBI-0735191 and DBI-1265383.



Getting Data into CyVerse





CyVerse Data Store

- Store any type of file related to your research
- Move files seamlessly between CyVerse platforms
- Automate file transfers
- Share files with lab members, collaborators, and communities



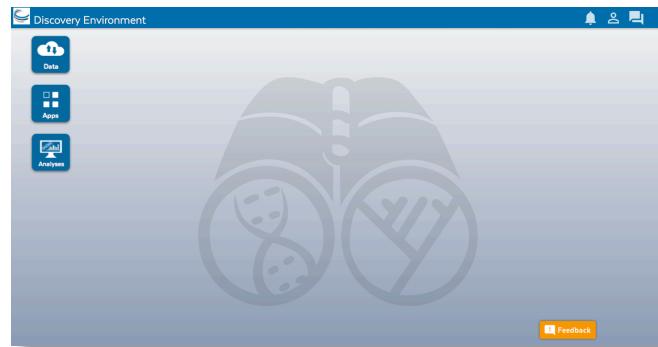
CyVerse Data Store

Multiple ways to access

Point-and-click



Cyberduck



Discovery Environment

Command line

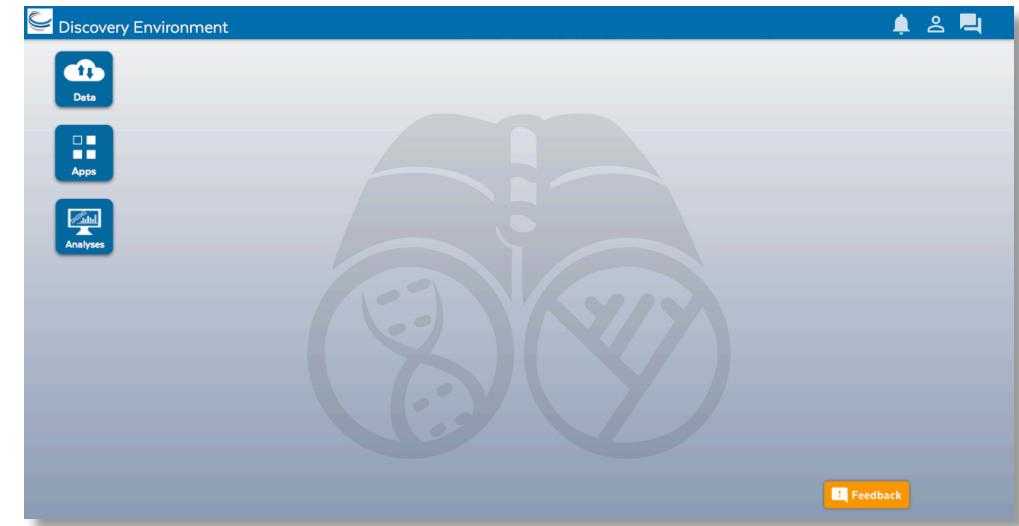
```
C- /iplant/home/williams/analyses/Soapdenovo_2.04b_analysis1-2013-09-18-13-35-36.219
C- /iplant/home/williams/analyses/Soapdenovo_2.04b_analysis1_47-2013-09-18-22-50-52.016
C- /iplant/home/williams/analyses/TASSEL_4.3.0 MLM_analysis1-2013-09-11-20-17-30.232
C- /iplant/home/williams/analyses/TASSEL_4.3.0 MLM_analysis1-2013-09-12-14-52-35.844
C- /iplant/home/williams/analyses/Test_of_New_App_analysis1-2013-10-25-14-40-49.857
```

iCommands



Discovery Environment

- Simple upload/download for small files
- Bulk upload files and folders (<10GB)
- Import from URL (no size limit)



Advantage +

Covers most upload/download sharing needs

Disadvantage -

Some size/speed limitations



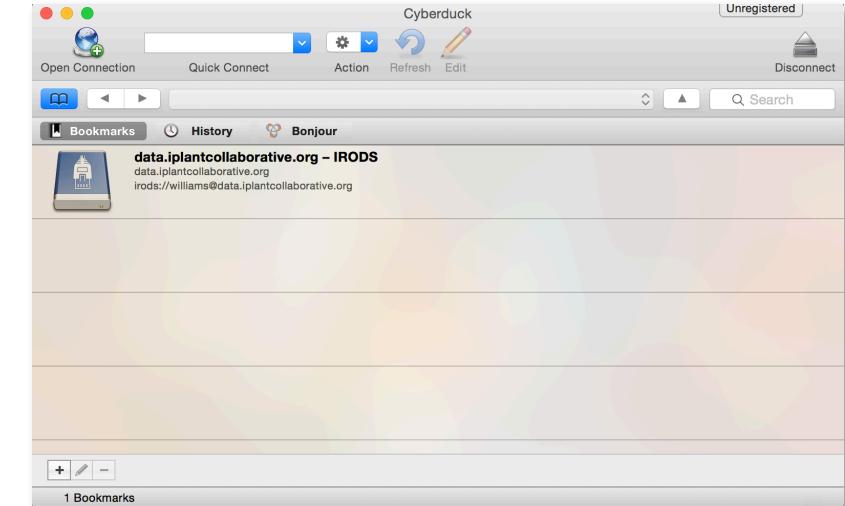


Cyberduck

- Drag and drop files and folders
- No size limit, file editing/previews
- Easy Desktop functionality

Advantage +

More like desktop file systems



Disadvantage -

No permissions/metadata control



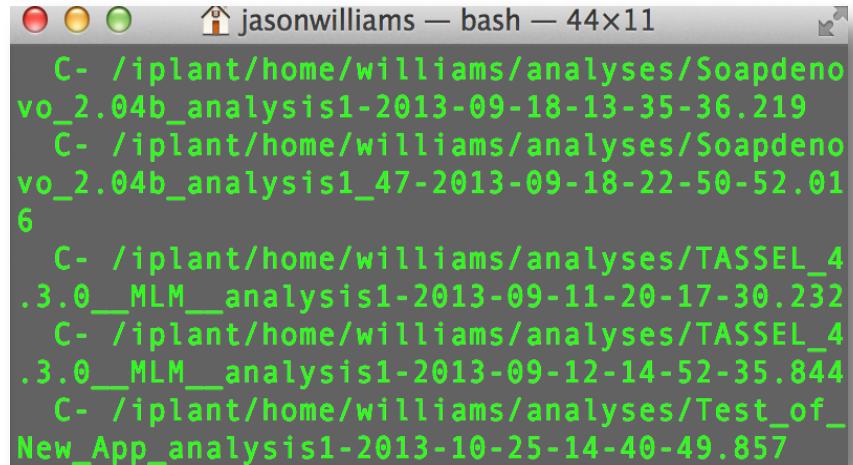


iCommands

- Full flexibility
- Ability to script and automate
- Access from terminal/server

Advantage +

Customizability



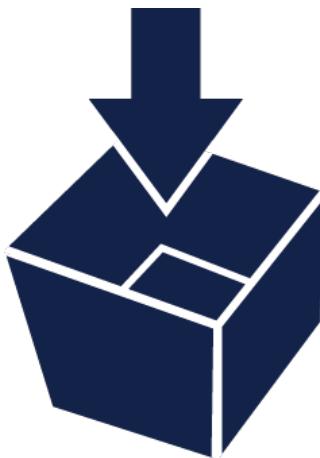
```
C- /iplant/home/williams/analyses/Soapdeno
vo_2.04b_analysis1-2013-09-18-13-35-36.219
C- /iplant/home/williams/analyses/Soapdeno
vo_2.04b_analysis1_47-2013-09-18-22-50-52.01
6
C- /iplant/home/williams/analyses/TASSEL_4
.3.0 MLM analysis1-2013-09-11-20-17-30.232
C- /iplant/home/williams/analyses/TASSEL_4
.3.0 MLM analysis1-2013-09-12-14-52-35.844
C- /iplant/home/williams/analyses/Test_of_
New_App_analysis1-2013-10-25-14-40-49.857
```

Disadvantage -

Requires some command line expertise



Cyberduck and iCommands Demo



Discovery Environment



Discovery Environment

- ✓ A platform that can run almost any bioinformatics application
- ✓ Seamlessly integrated with data and high performance computing
- ✓ User extensible – add your own applications





Discovery Environment Overview

Manage data



Data

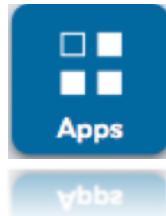
- Upload / Download files and folders
- Share files via URL (Public Links)
- Share files/folders with other users





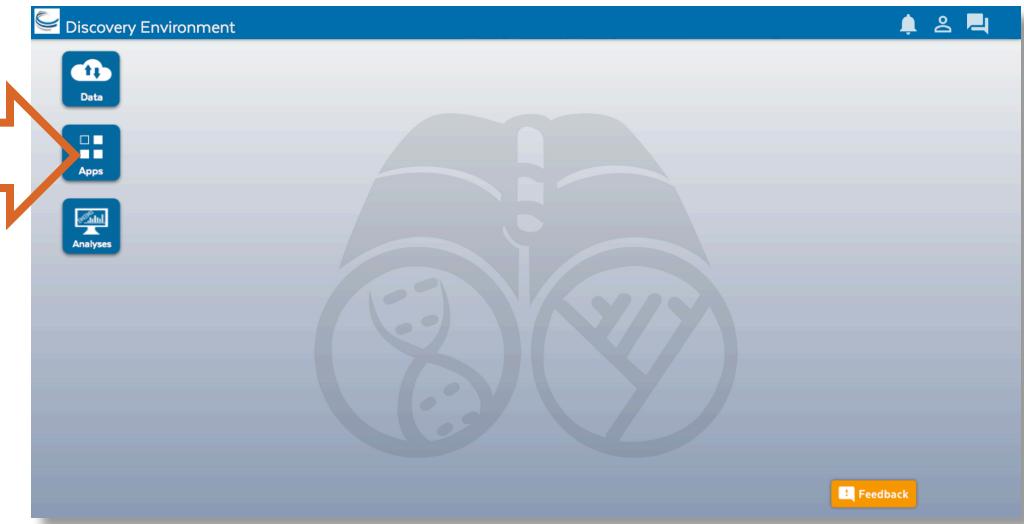
Discovery Environment Overview

Analyze data and customize Applications



Apps

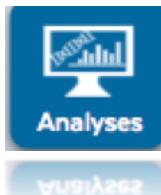
- Run hundreds of bioinformatics Apps
- Build automated workflows
- Modify Apps or integrate new ones





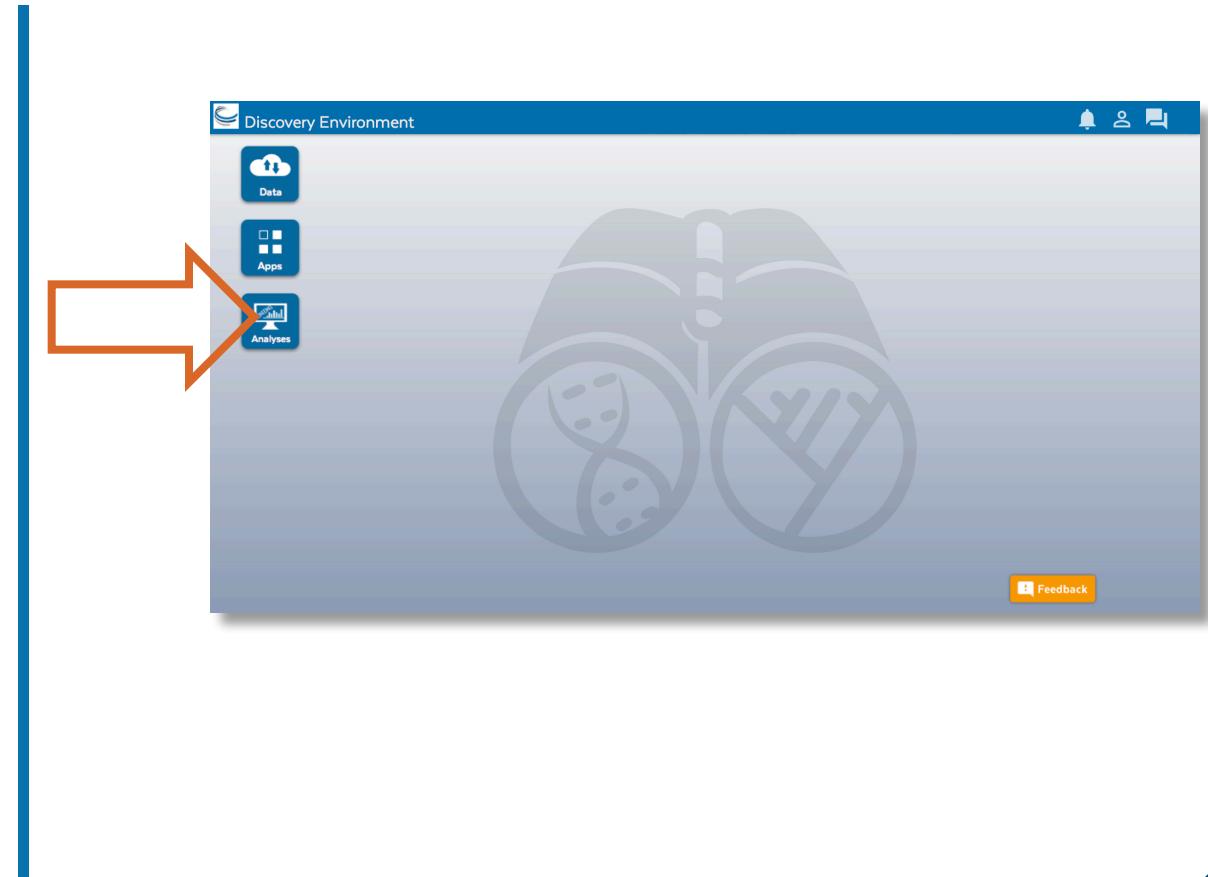
Discovery Environment Overview

View history, find results, reproduce analyses, optimize parameters

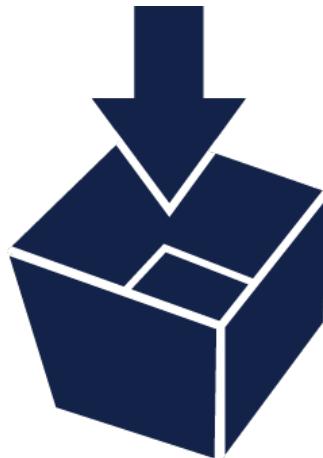


Analyses

- Monitor job status and find results
- Cancel jobs or re-launch jobs
- Detailed job history



Discovery Environment Demo





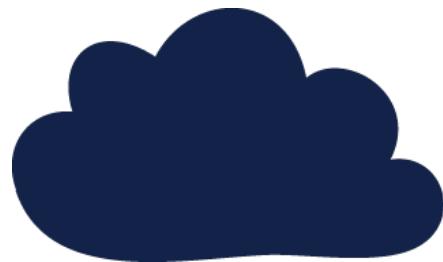
Discovery Environment

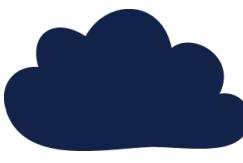
Demo analysis – sequence alignment using MUSCLE

Task: Take unaligned DNA sequences in FASTA format and create a multiple alignment

- ✓ View sample data in iPlant Data Store
- ✓ Launch a job using the MUSCLE sequence alignment app
- ✓ Monitor the job progress and view results

Atmosphere





Atmosphere

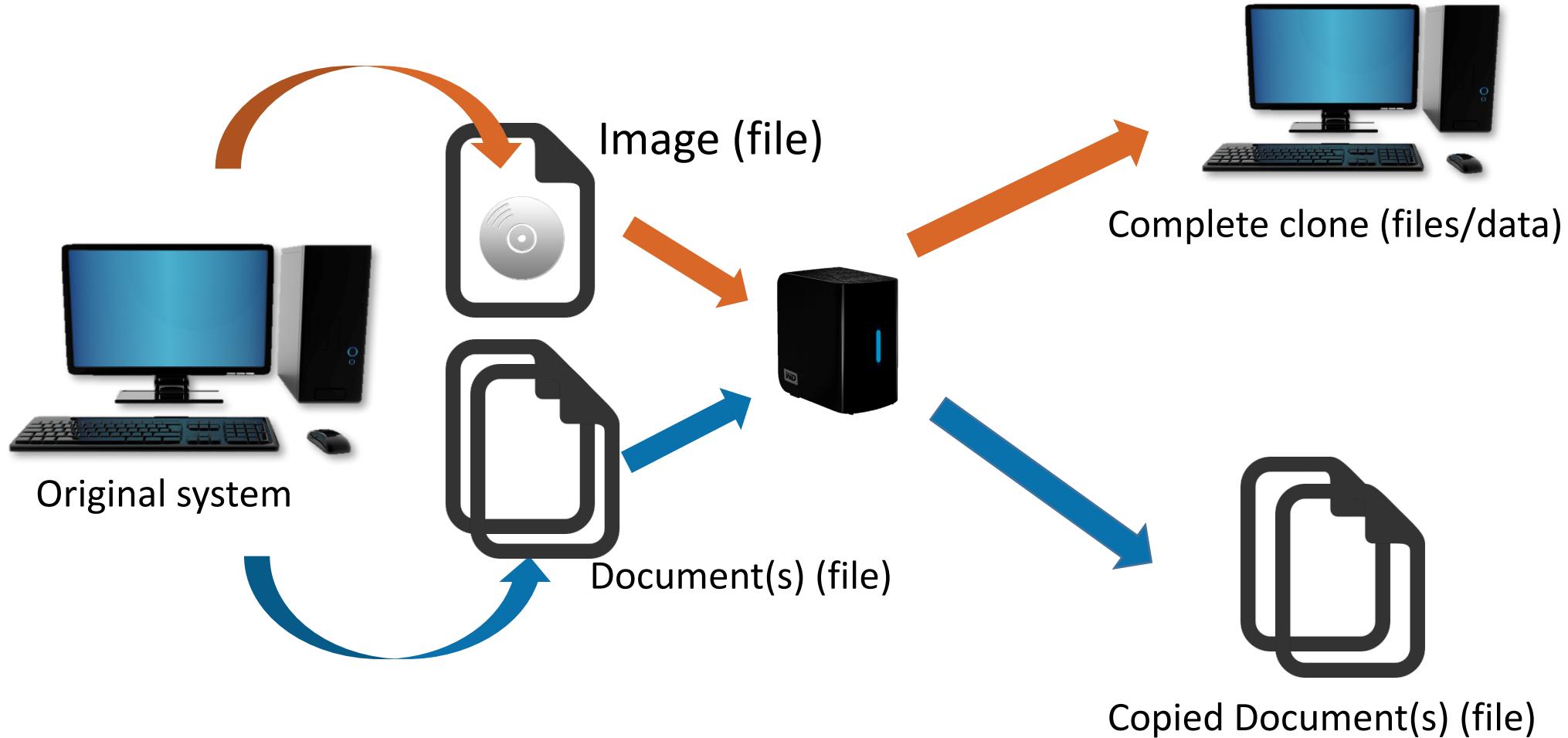
- ✓ Simple: 1-click access to hundreds of virtual machine images
- ✓ Flexible: Fully customize your software setup
- ✓ Powerful: Integrated with CyVerse computing and data resources





What is Cloud Computing?

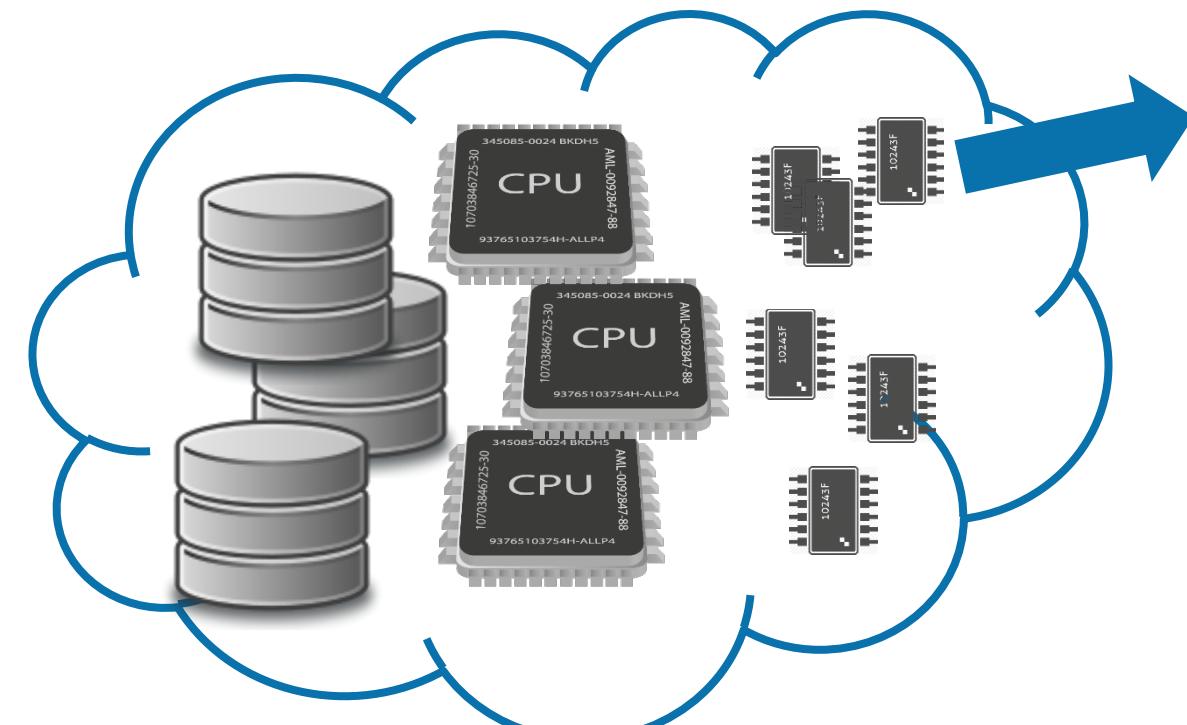
Important concepts: Image



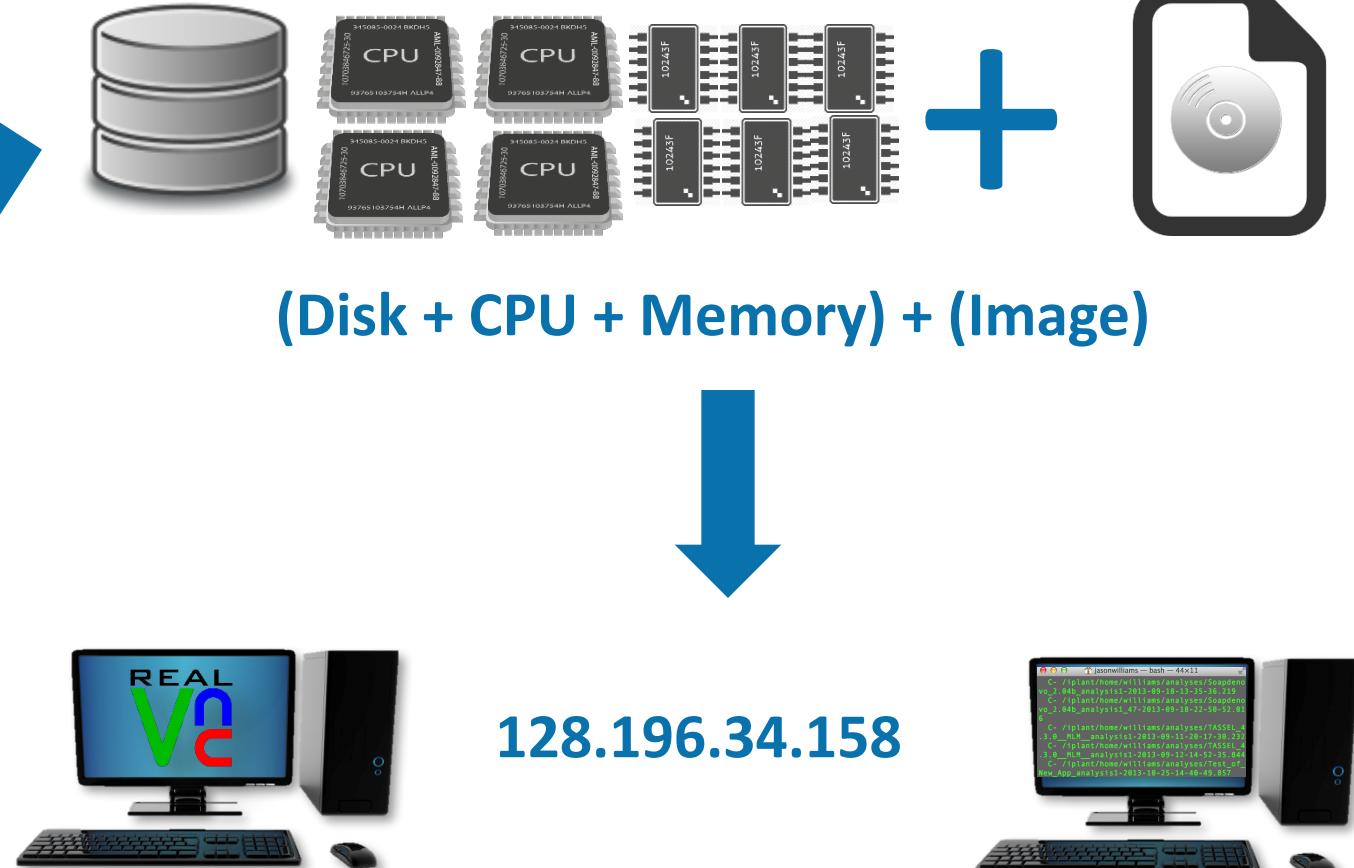


What is Cloud Computing?

Important concepts: Instance



CyVerse Cloud



Atmosphere Instance
(virtual machine)





Atmosphere Overview

Largest, easiest to use cloud for Life Sciences

The screenshot shows the CYVERSE Atmosphere Overview interface. At the top, there is a navigation bar with the CYVERSE logo (BETA), Dashboard, Projects, Images, Providers, and Help. Below the navigation bar, the title "Getting Started" is displayed. The page is divided into three main sections: "Launch New Instance" (with a rocket icon), "Browse Help Resources" (with a question mark in a cloud icon), and "Change Your Settings" (with a gear and magnifying glass icon). Each section contains descriptive text and a link.

- Launch New Instance**
Browse Atmosphere's list of available images and select one to launch a new instance.
- Browse Help Resources**
View a video tutorial, read the how-to guides, or email the Atmosphere support team.
- Change Your Settings**
Modify your account settings, view your resource quota, or request more resources.

- Choose an existing image or customize
- Instances up to 16-Core / 128 GB RAM
- Access via shell or VNC
- Share your image with selected users, or make them public





Atmosphere

Cloud computing for life sciences: sample use cases

- Run the software and data that are monopolizing your laptop/desktop
- Use desktop enabled images to run visually oriented programs (GUI)
- SUDO access – manage complex dependencies
- Uniform computing setups for your lab, collaborators, and students
- Make your own software available to a larger user community



Atmosphere Demo





Atmosphere

Cloud computing for life sciences

Windows



VNC Viewer



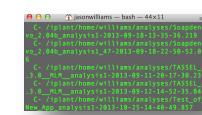
PuTTY

VNC Viewer:
PuTTY:

Mac



VNC Viewer

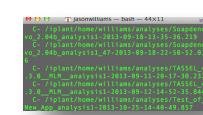


Shell/terminal

Linux



VNC Viewer



Shell/terminal

www.realvnc.com/download/viewer
www.putty.org



Where to go from here:



The screenshot shows the CyVerse homepage. At the top, there's a navigation bar with links for About, Contact, Status, Help, Sign Up, and a search bar. Below the header, a banner reads "TRANSFORMING SCIENCE through data-driven discovery". A sidebar on the left says "THE IPLANT COLLABORATIVE HAS BECOME CYVERSE" and provides a link to read more. The main content area features a circular network diagram with nodes like Discovery Environment, Bisque Image Analysis, Powered by CyVerse, Atmosphere, Data Store, Learning Center, DNA Subway, Science APIs, and a central node for Publications. An orange rectangle highlights the "Learning Center" node. At the bottom, there are sections for "UPCOMING EVENTS" and "PUBLICATIONS".

Learning Center

- Get Started Guide
- Tutorials and Videos
- Documentation

Upcoming Events

- Workshops
- Webinars



Slides and Tutorials



www.cyverse.org/sol2016

