

Powered by iPlant

The iPlant Collaborative as underlying infrastructure for bioinformatics projects

Plant and Animal Genome Meeting

1/16/2012

Dan Stanzione

iPlant Collaborative

Texas Advanced Computing Center

dan@tacc.utexas.edu

dan@iplantcollaborative.org



iPlant's Central Challenge

- To define what it means to build a lasting, community driven **Cyberinfrastructure** for the ***Grand Challenges*** of **Plant Science**, to get **community buy-in** of this vision, and to execute this vision.



iPlant Cyberinfrastructure

- Universal, accessible, capacious storage
- Experimental reproducibility
- Sharing and collaboration
- Information Visualization
- Rapid, iterative workflow construction
- Ability to integrate community algorithms and practices
- Vast computing power (support to use it)
- **Support for community driven science**



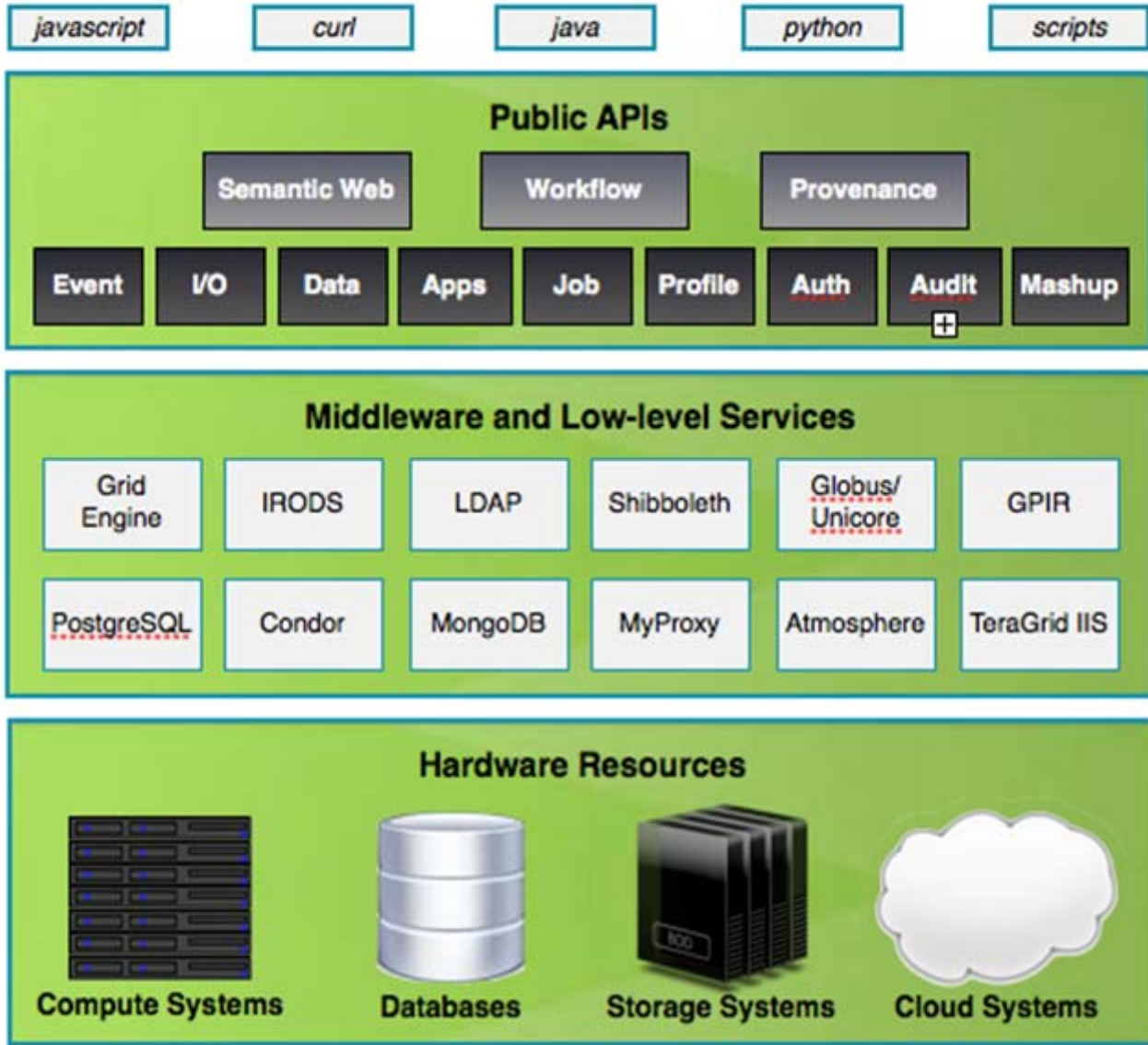


Language/OS-independent application development

Complexity is abstracted behind API layer

Reuse and integration of Open Source Middleware

National-scale physical resources



Powered by iPlant

- The iPlant CI is designed as infrastructure. This means it is a platform upon which other projects can build.
- Use of the iPlant infrastructure can take one of several forms:
 - Storage
 - Computation
 - Hosting
 - Web Services
 - Scalability



Powered by iPlant

- Other major projects are beginning to adopt the iPlant CI as their underlying infrastructure (some completely, some in limited ways):
 - BioExtract (*web service platform*)
 - CiPRES (*computation*)
 - Gates Integrated Breeding Platform (*hosting, development*)
 - Galaxy (*storage, for now*)
 - IAIC? TAIR?



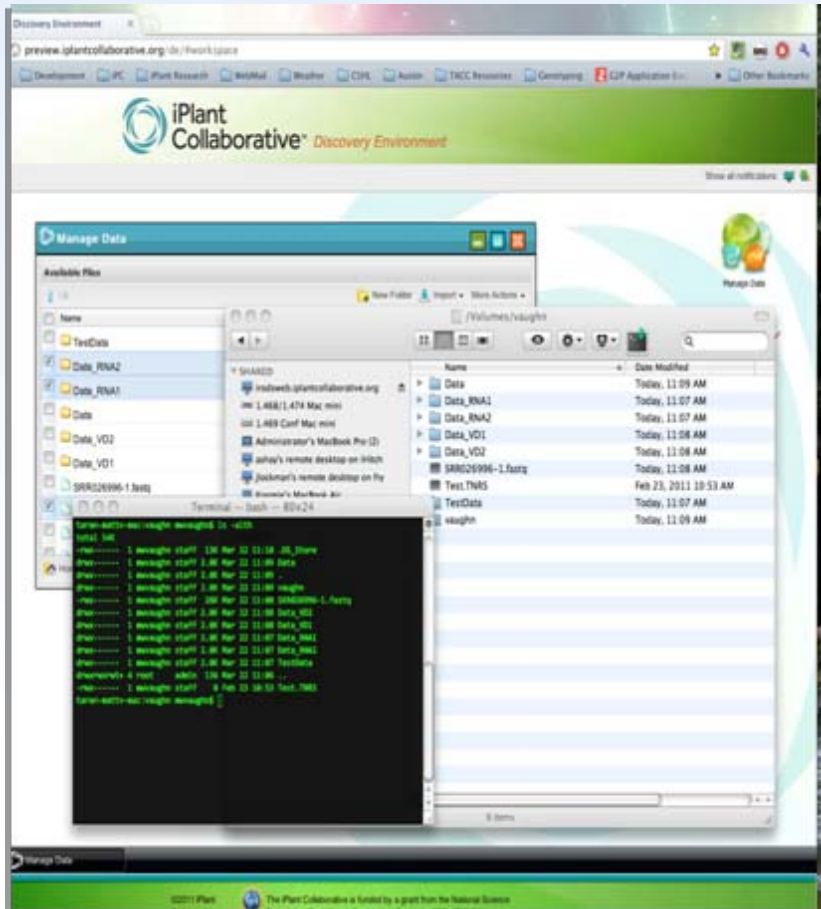
Major Ways to Access iPlant

- Atmosphere: For virtual hosting of web apps, sites, databases.
- iPlant Data Storage: All data large and small
- The Discovery Environment: Integrated Web apps.
- MyPlant: Social Networking.
- DNASubway: Annotation and more
- Standalone Apps: TNRS, TreeViewer, PhytoBisque, etc
- The API: For programmers embedding iPlant CI capabilities
- Command line for experts (thru TeraGrid/XSEDE)



iPlant Data Store

This is “Cloud Storage”... but it's not Amazon



Fast data transfers via parallel, non-TCP file transfer

- Move large (>2 GB) files with ease

Multiple, consistent access modes

- iPlant API
- iPlant web apps
- Desktop mount (FUSE/DAV)
- Java applet (iDrop)
- **Command line**

Fine-grained ACL permissions

Access and a storage allocation is automatic with your iPlant account

- Sharing made simple



iPlant Data Store

- When we say “capacious”, we mean $>100,000$ Terabytes of disk and tape



Project Atmosphere™

- API-compatible implementation of Amazon EC2/S3 interfaces
- Virtualize the execution environment for applications and services
- Up to 12 core / 48 GB instances
- Access to Cloud Storage + EBS
- Run servers, CloudBurst desktop use cases. Big data and the desktop are co-local



>60 hosted applications in Atmosphere today, including users from USDA, Forest Service, database providers, etc.

(30 more for postdocs and grad students for training classes)



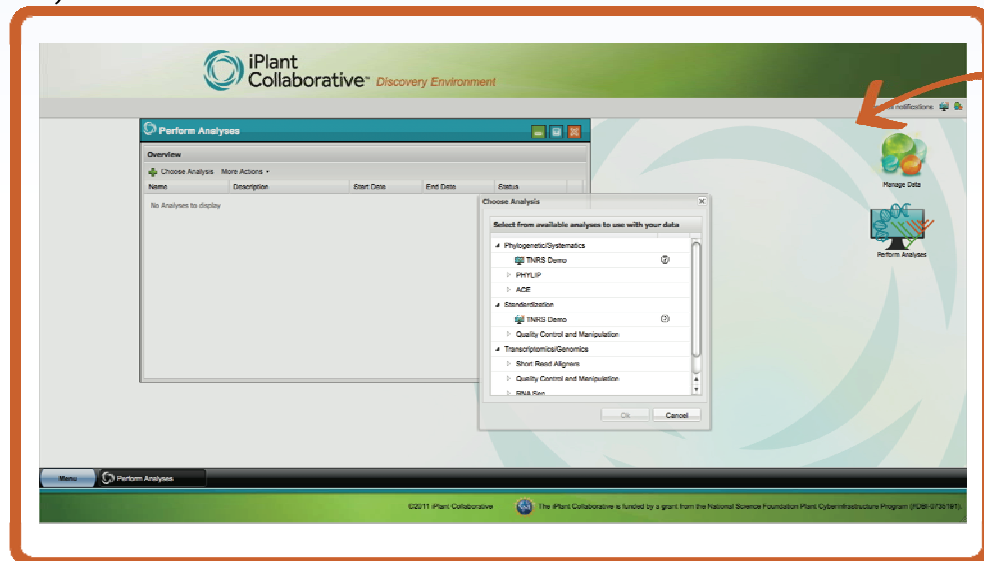
Scalable Computation

- 90,000 Compute Cores
- Up to 1TB shared memory
- Growing to ~500,000 cores by end of 2012



Discovery Environment

- A rich web client
 - Provides a consistent interface to a range of bioinformatics tools
 - Provides a portal to users not wishing to interact with lower level infrastructure
- An integrated, extensible system of applications and services
 - Provides additional intelligence above low level APIs – Provenance, Collaboration, etc.

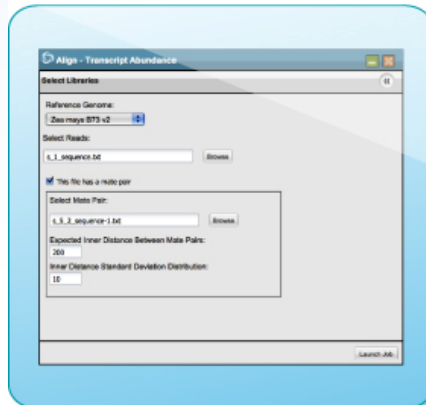


Integration of New Tools to UI without Programming

JSON

```
{
  "uuid": "jd16cf1ec-81f6-4f48-b353-bf7b77acc4f5",
  "type": "rscript",
  "request_type": "submit",
  "component": {
    "id": "c12bd559051333352e31302e3231d750adbbb3b582a",
    "location": "/usr/local/bin",
    "description": "R Statistical Package",
    "name": "R",
    "type": "executable"
  },
  "config": {
    "name": "simplesqqfwe",
    "tid": 5,
    "type": "rscript",
    "config": {
      "inputs": {
        "testinput": "some_file.csv",
        "type": "file"
      },
      "selectedOption": "nothing",
      "scriptN": "/usr/local/iplant/scripts/testingScript.R"
    }
  }
}
```

Wizard



Computing Nodes



Metadata

eFP in iPlant: in 38 minutes in September

iPlant Collaborative™ Discovery Environment

Show all notifications

Manage Data

Import ▾ More Actions ▾

- vaughn
 - 0.30
 - abyss
 - analyses
 - ath250_1
 - ath250_2
 - ath250_bundle
 - ath500_1
 - ath500_2
 - ath500_bundle
 - athSE_bundle
 - bbc1
 - cuffdiff_2
 - cuffdiff_3
 - cuff_diff_gt_1.5
 - cuffdiff qt3.34

AT5G45082.png

Image

32 257898_s_at

Arabidopsis eFP Browser at bar.utoronto.ca
Winter et al., 2007. PLoS One 2(8): e718

Seed/Siliques 4

Shoot Apex Inflorescence

Shoot Apex Transition

Shoot Apex Vegetative

Cotyledons

Hypocotyl

Root

2nd Internode

1st Node

Cauline Leaf

24 h Imbibed Seed

Root

Seed/Siliques 5

Seed/Siliques 4

Seed/Siliques 3 + globular

Seed/Siliques 4 + heart

Seed/Siliques 5 + torpedo

Seed/Siliques 6 - walking-stick

Seed/Siliques 7 - curled cotyledons

Seed/Siliques 8 - green cotyledons

Seed/Siliques 9 - green cotyledons

Seed/Siliques 10 - green cotyledons

Vegetative Rosette

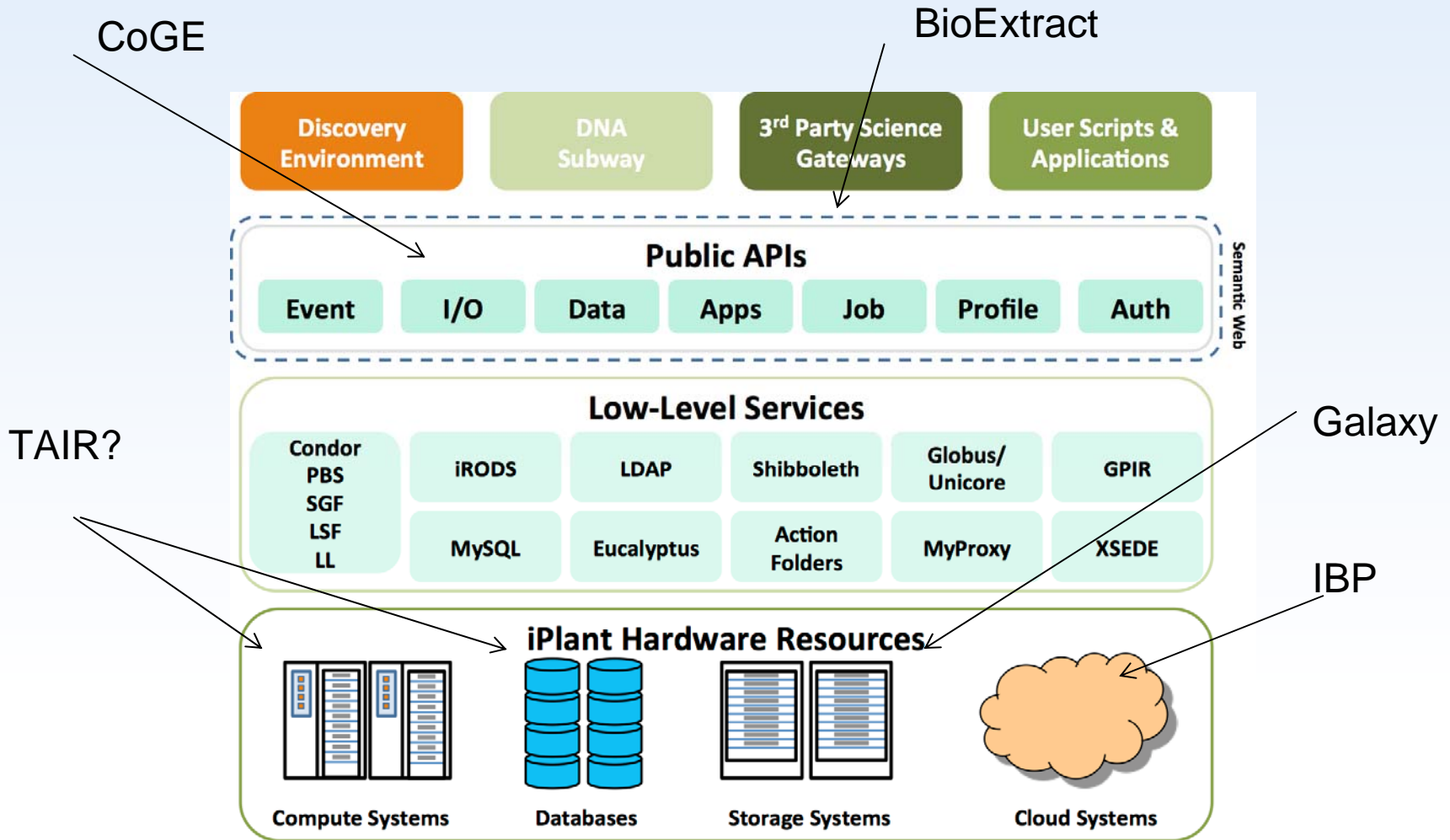
Perform Analyses

Manage Data

AT5G45082.png



Using the iPlant CI as a Foundation



Starting a New Collaboration

- If you just want to use the existing platforms, or add your own tools and workflows, just go to the web page, sign up, and read the documentation
- If you want to develop your own applications on top of the iPlant infrastructure, contact us directly:
 - dan@iplantcollaborative.org
 - support@iplantcollaborative.org
 - Best discussions are before proposals are submitted.
- Work with us to develop proposals for IAIC!



TakeAways

- iPlant is building a robust CI, and many services and tools are available now.
 - >100 tools in discovery environment, and growing fast
 - >50 Applications in Atmosphere
 - Vast Computing and Storage resources available
- But iPlant is not about the specific tools available on a given day. It's about changing the way we build the future of bioinformatics.
 - Make use of the emerging international CI
 - Standard ways to interface tools; **things that work together!**
 - Standard ways to find and use them.



Stop doing this...

```
tophat -r 160 -o top_SRR027863-65 ../../reference/hg19
SRR027863_1.fastq,SRR027864_1.fastq,SRR027865_1.fastq
SRR027863_2.fastq,SRR027864_2.fastq,SRR027865_2.fastq

tophat -r 160 -o top_SRR027866-67 ../../reference/hg19
SRR027866_1.fastq,SRR027867_1.fastq SRR027866_2.fastq,SRR027867_2.fastq

cufflinks -o cuff_SRR027863-65 top_SRR027863-65/accepted_hits.bam
cufflinks -o cuff_SRR027866-67 top_SRR027866-67/accepted_hits.bam

cuffmerge -s ../../reference/hg19.fa assemblies.txt

cuffdiff merged_asm/merged.gtf top_SRR027863-65/accepted_hits.bam
top_SRR027866-67/accepted_hits.bam
```

Start doing this...

The screenshot displays the iPlant Collaborative Discovery Environment interface. The top header features the iPlant Collaborative logo and the text "Discovery Environment". On the right, there are user links: "dan", "Help", and "Notifications".

On the left sidebar, there are three main icons: "Data", "Analyses", and "Apps". The "Data" panel is active, showing a tree structure under the "dan" user. The "Apps" panel is also active, showing a list of categories and a table of applications.

Data Panel:

- dan
 - analyses
 - CuffDiff
 - Cufflinks-HY5
 - Cufflinks-WT
 - fastq-dump-HY5
 - FPKM-ColumnExt
 - Import_SRX02958
 - Import_SRX02958
 - job1-Dan
 - TopHat-HY5
 - Tophat-WT
 - WT-FastQ-Dump

Apps Panel:

Categories:

- Utility Tools and Scripts (9)
- NGS (24)
 - Aligners (6)
 - QC and Processing (5)
 - Assembly and Annotation (6)
 - RNAseq Analysis (3)
 - ChIPseq Analysis (1)
 - Utilities (2)
 - Variant Identification (1)
 - QTL and GWAS (10)
 - Data Sources (2)
- Phylogenetics (20)
 - Tree Building (11)**
 - Comparative Methods (4)
 - Evolutionary Models (2)
 - Community Ecology (1)
 - Utility Tools and Scripts (2)
 - Functional Analysis (5)

Tree Building Table:

Name	Integrated by	Published on	Rating
PhyML	Eric Lyons		★★★★★
RAXML - Proteins	Naim Matasci		★★★★★
RAXML - Nucleotides	Naim Matasci		★★★★★
Ninja	Naim Matasci		★★★★★
PROTNJ	Sheldon McKay		★★★★★
DNAPARS	Sheldon McKay		★★★★★
FastTree2	Sheldon McKay		★★★★★
DNANJ	Sheldon McKay		★★★★★
PROML	Sheldon McKay		★★★★★
DNAML	Sheldon McKay		★★★★★
PROTPARS	Sheldon McKay		★★★★★

©2011 iPlant Collaborative