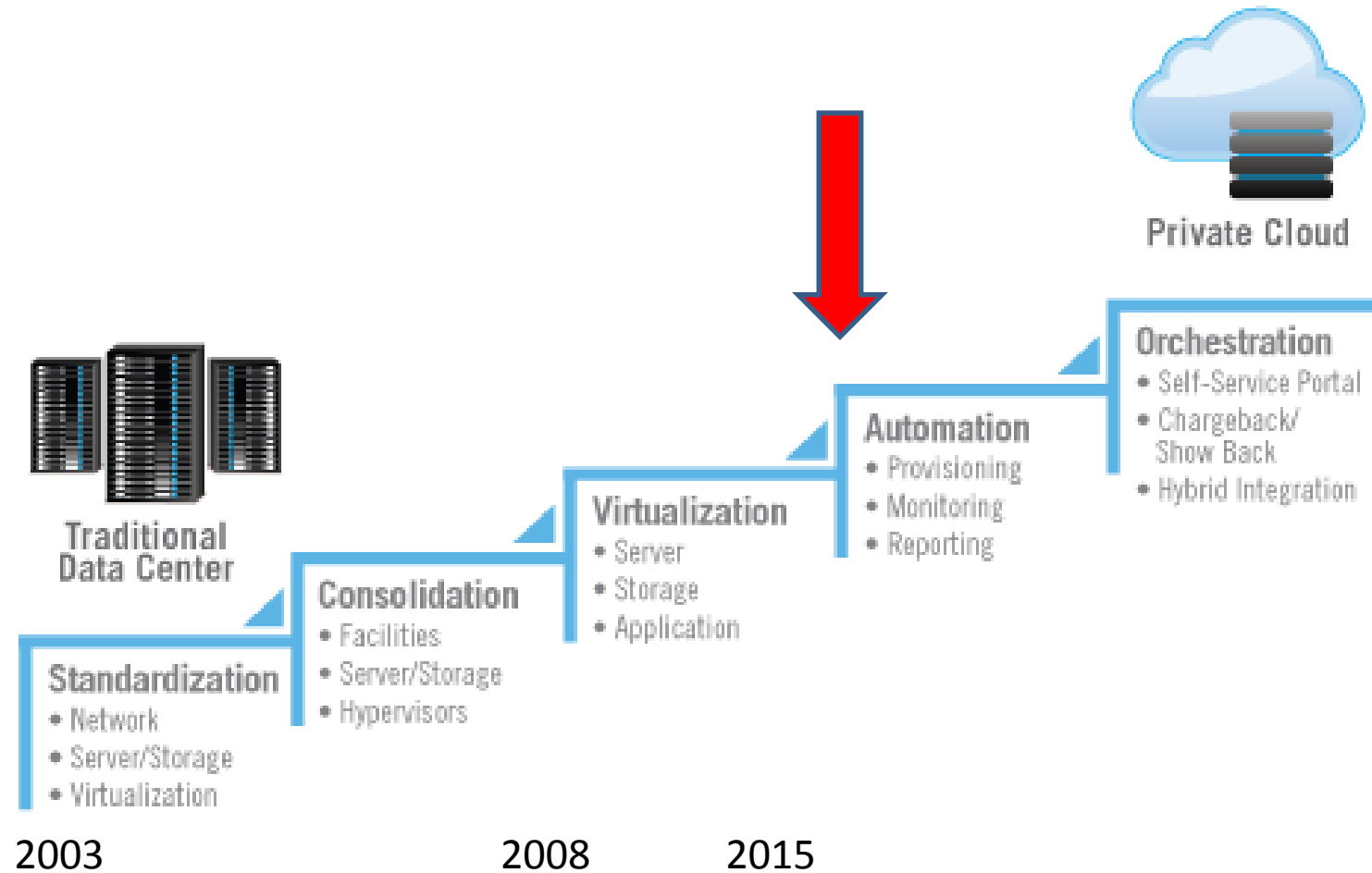


**Basic Course on Bioinformatics tools for
Next Generation Sequencing data mining**

**IT infrastructure and user interface:
The Galaxy architecture and
ARIES cluster**

Arnold Knijn
IT Sector - ISS

Data Center evolution



ISS IT infrastructure

> 130 virtual servers

50 TB



180 TB

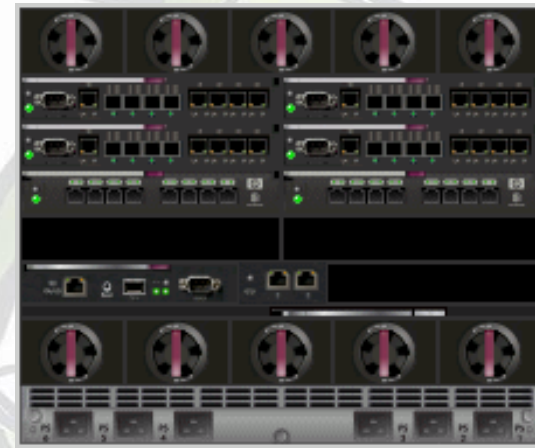


Storage



256 CPU/512 thread

1,552 GB RAM



32 x 1 GB/s

Networking

Data intensive biology *for everyone.*

Galaxy is an open, web-based platform for data intensive biomedical research. Whether on the free public server or your own instance, you can perform, reproduce, and share complete analyses.

Use Galaxy



Use project's free server or other public servers

Get Galaxy



Install locally or in the cloud or get Galaxy on SlipStream

Learn Galaxy



Screencasts, Galaxy 101, ...

Get Involved



Mailing lists, Tool Shed, wiki

[Search all resources](#)

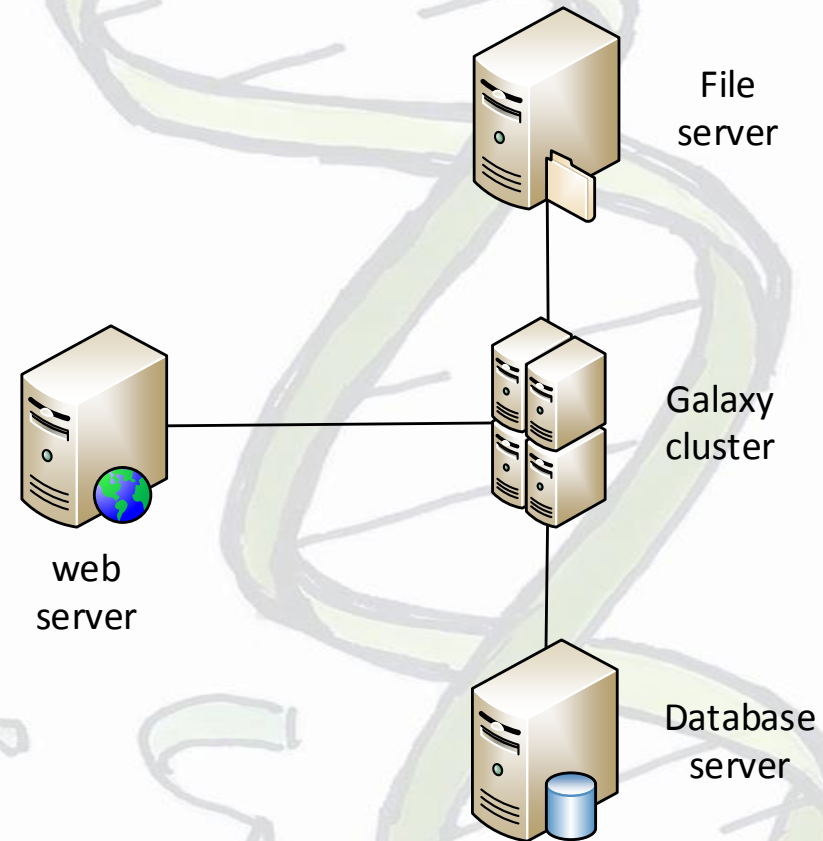
The Galaxy Team is a part of the Center for Comparative Genomics and Bioinformatics at Penn State University, and the Department of Biology at Johns Hopkins University. The Galaxy Project is supported in part by NHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Johns Hopkins University.

Which Galaxy?

- Public server (80+)
- Own computer
- Appliance Galaxy Edition
- Data Center cluster
- Cloud standalone: Galaxy on Jetstream
- Cloud virtual cluster: Cloudman

Galaxy components

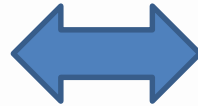
- Job runner
- File server
- Database server
- Web server



Default vs production installation

All-in-one (default)

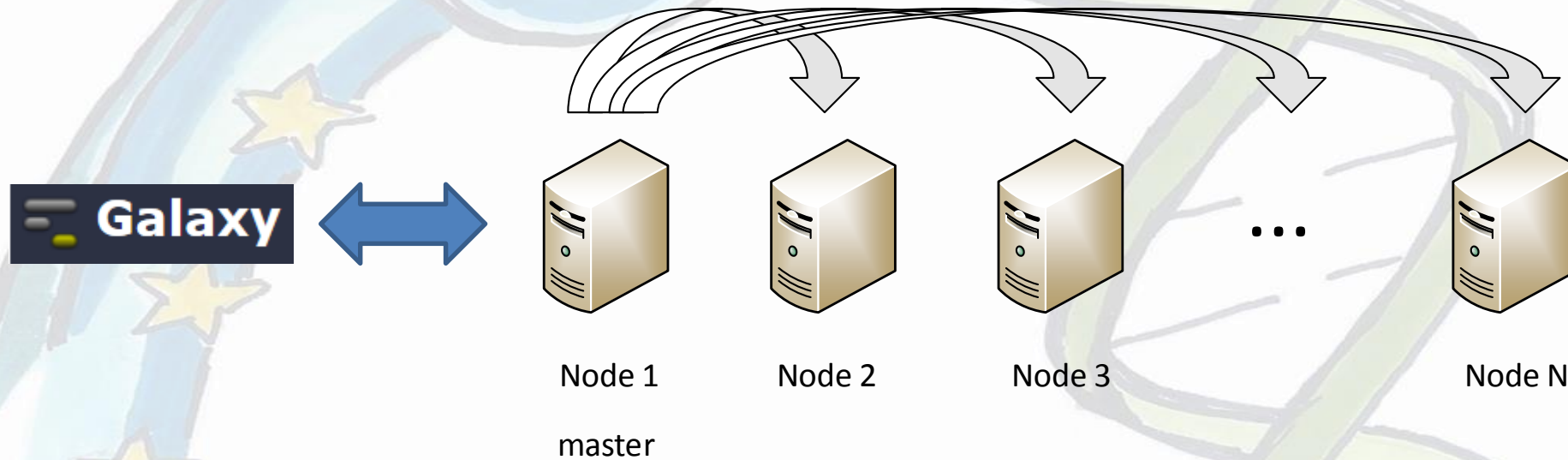
- ✗ Internal file system
- ✗ SQLite
- ✗ Built-in HTTP server for all tasks
- ✗ Local job runner
- ✗ Single process
- ✓ Simplest error-proof configuration



Production (scalable)

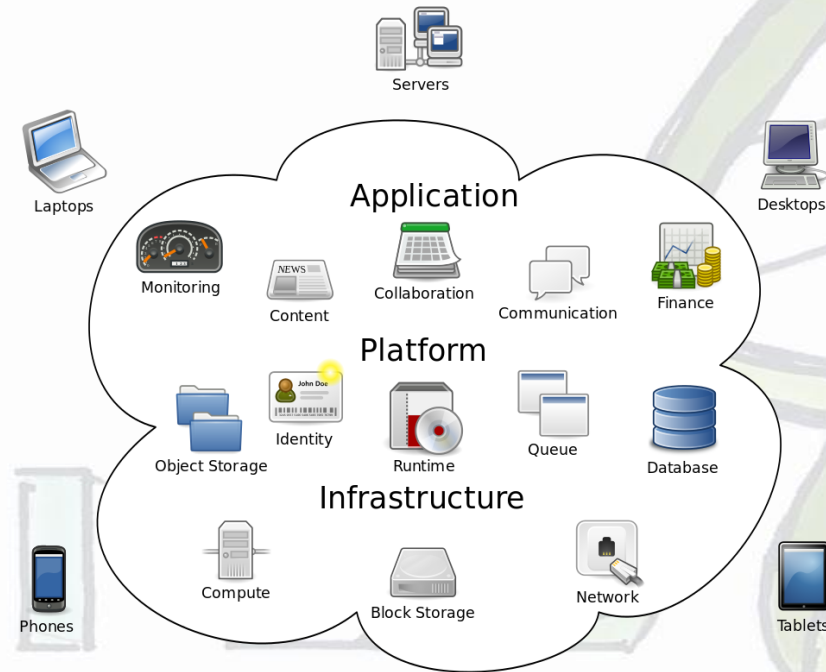
- ✓ External file server
- ✓ Real database
- ✓ Real HTTPS server for many tasks
- ✓ Cluster job runner
- ✓ Multi process
- ✗ More complex configuration

Galaxy cluster



Cloud?

- Cloud Storage (Google drive, iCloud, OneDrive, DropBox, ecc.)
- Cloud Computing: internet-based on-demand access to a shared pool of configurable computing resources

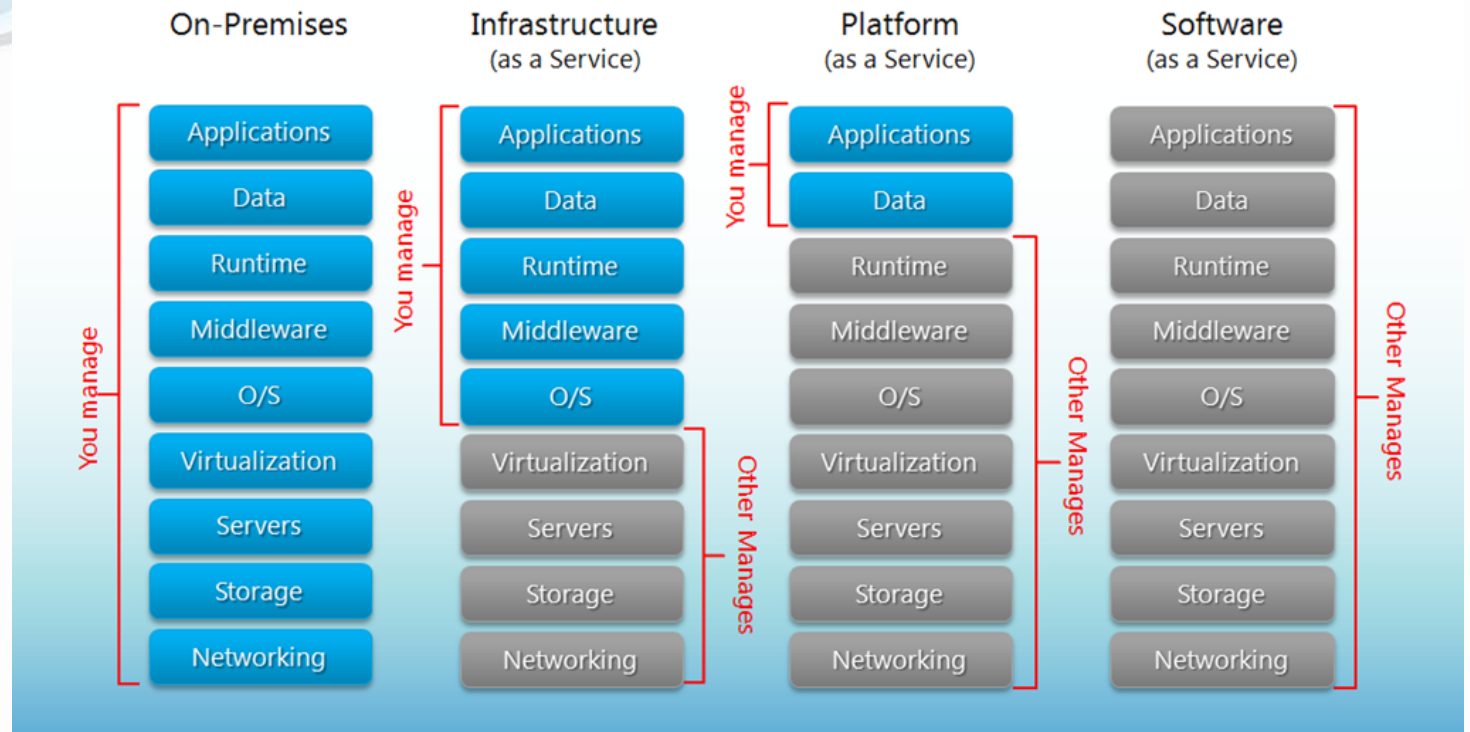


Cloud computing


Galaxy on the cloud

- Infrastructure as a Service (system administrators)
- Platform as a Service (software engineers)
- Software as a Service (users)
- Self-service
- Your wallet is the limit


Separation of Responsibilities



Galaxy on Jetstream (USA)

 Images Help Login

SEARCH TAGS

 Galaxy 16.01 Standalone

Name: Galaxy 16.01 Standalone


Created: 3/30/2016 05:08 pm CEST

Created by: admin


Description: Galaxy 16.01 Standalone - based on Ubuntu 14.04.4 LTS
This is a standalone Galaxy server that comes preconfigured with hundreds of tools and commonly used reference datasets: just launch and use.
It is necessary to launch an instance type of Large or larger.

Tags: community-contributed Featured Ubuntu

Versions:




1.0-latest
3/30/2016 08:38 pm CEST by admin



Login to Launch

16/06/2016

IT infrastructure and user interface: The Galaxy architecture and ARIES cluster

 Istituto Superiore di Sanità

Galaxy on Amazon

Galaxy Cloud Launch

Easily launch your own cloud servers for use with [Galaxy](#) and [CloudMan](#). See [this page](#) for detailed instructions on how to get started.

Release Candidate 1 for Galaxy on the Cloud with Galaxy 16.04 is available. If you would like to use/test it, choose *Testing: Galaxy 16.04* flavor under the Advanced options dropdown.

Cloud

Choose from the available clouds. The credentials you provide below must match (ie, exist on) the chosen cloud.

Access key

Your cloud account API access key. For the Amazon cloud, available from the [security credentials page](#).

Secret key

Your cloud account API secret key. For the Amazon cloud, also available from the [security credentials page](#).

Cluster name or

Name of your cluster used for identification and restarting. If creating a new cluster, type any name you like.

Password

Your choice of password, for the CloudMan web interface and accessing the server via ssh.

Instance type

Type (ie, virtual hardware configuration) of the server to start.

Cluster type ☒ Cluster with Galaxy
☐ Cluster only
☐ Do not set cluster type now

The cluster type determines the initial startup template used by CloudMan. See [this page](#) for details on cluster types.

Storage type ☒ Persistent volume storage
☐ Transient instance storage

The type of storage to use for the main file system. See [this page](#) for more details on storage types.

Storage size

The size of the storage (in GB; number only). The default is 10.

↓ Show advanced startup options

Galaxy user interface

Galaxy / ARIES - ISS Analyze Data Workflow Shared Data Visualization Admin Help User Using 1.6 GB

Tools

search tools

--- COMMON TOOLS ---

- [Get Data](#)
- [Send Data](#)
- [Lift-Over](#)
- [Text Manipulation](#)
- [Filter and Sort](#)
- [Join, Subtract and Group](#)
- [Convert Formats](#)
- [Extract Features](#)
- [Fetch Sequences](#)
- [Fetch Alignments](#)
- [Statistics](#)
- [Graph/Display Data](#)
- [GraPhlAn](#)

---HREVAP TOOLS---

[HReVAP](#)

---NGS TOOLS---

- [NGS: Assembly](#)
- [NCBI Blast](#)
- [Manipulation](#)
- [kSNP3](#)
- [Gene Annotation](#)
- [FASTA manipulation](#)
- [NGS: Mapping](#)
- [NGS: SAM Tools](#)
- [NGS: QC and manipulation](#)

Istituto Superiore di Sanita'

ARIES - Advanced Research Infrastructure for Experimentation in Genomics - Galaxy Instance at ISS

Tweets by @ARIES_GENOMICS

Aries Group Retweeted

TGAC
@GenomeAnalysis

5 reasons #computing isn't scary & why learning to #code is one of the most useful skills bit.ly/1q5oEJF

Five reasons why computing isn't as scary as you think

07 May

Embed View on Twitter

Please read our [disclaimer](#) before using ARIES.

Galaxy is an open, web-based platform for data intensive biomedical research. The Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.

History

search datasets

Unnamed history
287 shown, 665 deleted, 7 hidden
2.77 GB

- [955: BAM-to-SAM on data 147: converted SAM](#)
- [948: Cut on data 947](#)
- [947: Select on data 942](#)
- [945: Cut on data 942](#)
- [944: Cut on data 942](#)
- [943: Cut on data 942](#)
- [942: Galaxy1128-\[BAM-to-SAM on data 1124 converted SAM\].sam](#)
- [941: comment on data 194, data 193, and others: heatmap_plain.png](#)
- [940: comment on data 194, data 193, and others: heatmap_percentage.png](#)


Galaxy user interface

The screenshot displays the Galaxy web interface with several key components highlighted by red boxes and labels:




- MENU:** Located at the top center, it points to the navigation bar containing links like 'Analyze Data', 'Workflow', 'Admin', 'Help', 'User', and a grid icon.
- TOOLS:** Located on the left side, it points to the 'Tools' panel which lists various bioinformatics tools such as 'GFF3', 'search for patterns in DNA using PatMaN', 'Parse Primer search', 'EMBOSS primer search output to tabular', 'GVF Feature ID Extractor', 'LOCI ASSEMBLYING-TERMOTYPING TOOLS', 'Allele Cluster', 'LociExtractor', 'T-Melt Calculator', 'Bin Assigner', 'CLUSTERING TOOLS', 'Concatenate', 'TermoType Clustering', 'NGS: Assembly', and 'NCBI Blast'.
- WORKING AREA:** Located in the center, it points to the main workspace where the 'Bin Assigner' tool is currently running. The tool's interface shows a 'Melting Temperatures Tab' with a file selection dropdown (showing '955: BAM-to-SAM on data 147: converted SAM') and a 'Column to consider to Assign Bin' field. Below the input fields is an 'Execute' button and a description: 'Loads the file with the temperatures; matches the temperatures with reference bins and assigns the corresponding alleles'.
- DATA:** Located on the right side, it points to the 'History' panel which lists previous jobs. Each entry includes a job ID (e.g., 455, 454, 452, 451, 450, 449, 448, 447, 444, 440), a description (e.g., 'Escherichia coli classification of data 454', 'Tm 15 strains_geniTolti.txt'), and icons for viewing, editing, and deleting the job.

Two blue curved arrows originate from the 'TOOLS' and 'DATA' panels and point towards the 'WORKING AREA', indicating the flow of data and tool execution within the Galaxy environment.

Intuitive and self-documenting

 **Bin Assigner** Home-made tool for bin assignment (Galaxy Tool Version 1.0.0) ▼ Options

Melting Temperatures Tab


 953: Tm_77_new.txt ▼

Column to consider to Assign Bin

7

✓ Execute

Loads the file with the temperatures; matches the temperatures with reference bins and assigns the corresponding alleles

Citations  Show BibTeX

Michelacci, Valeria and Orsini, Massimiliano and Knijn, Arnold and Delannoy, Sabine and Fach, Patrick and Caprioli, Alfredo and Morabito, Stefano (2016). Development of a High Resolution Virulence Allelic Profiling (HReVAP) Approach Based on the Accessory Genome of Escherichia coli to Characterize Shiga-Toxin Producing E. coli (STEC). In *Frontiers in Microbiology*, 7. [[doi:10.3389/fmicb.2016.00202](https://doi.org/10.3389/fmicb.2016.00202)][[Link](#)]

UI vs Command-Line

Tool: Bin Assigner

Name:	BinAssigner Log File
Created:	Fri Feb 13 07:43:59 2015 (UTC)
Filesize:	877 bytes
Dbkey:	?
Format:	txt
Galaxy Tool ID:	binassigner
Galaxy Tool Version:	1.0.0
Tool Version:	
Tool Standard Output:	<u>stdout</u>
Tool Standard Error:	<u>stderr</u>
Tool Exit Code:	0
API ID:	e9fb797960230e8a
History ID:	f597429621d6eb2b
UUID:	dc1676ef-87b7-48bf-a24e-4359f57cf2fa
Full Path:	/home/galaxy/galaxy-dist/database/files/001/dataset_1439.dat
Job Command-Line	python /home/galaxy/galaxy-dist/tools/Hrevap/BinAssigner.py -t /home/galaxy/galaxy-dist/database/files/001/dataset_1433.dat -o /home/galaxy/galaxy-dist/database/files/001/dataset_1438.dat -c 7 > /home/galaxy/galaxy-dist/database/files/001/dataset_1439.dat
Job Runtime (Wall Clock)	1 seconds
Cores Allocated	1
Job Start Time	2015-02-13 08:44:00
Job End Time	2015-02-13 08:44:01

Input Parameter	Value	Note for rerun
Melting Temperatures Tab	176: ThermoTyping Summary File	
Column to consider to Assign Bin	7	

Home-made tools

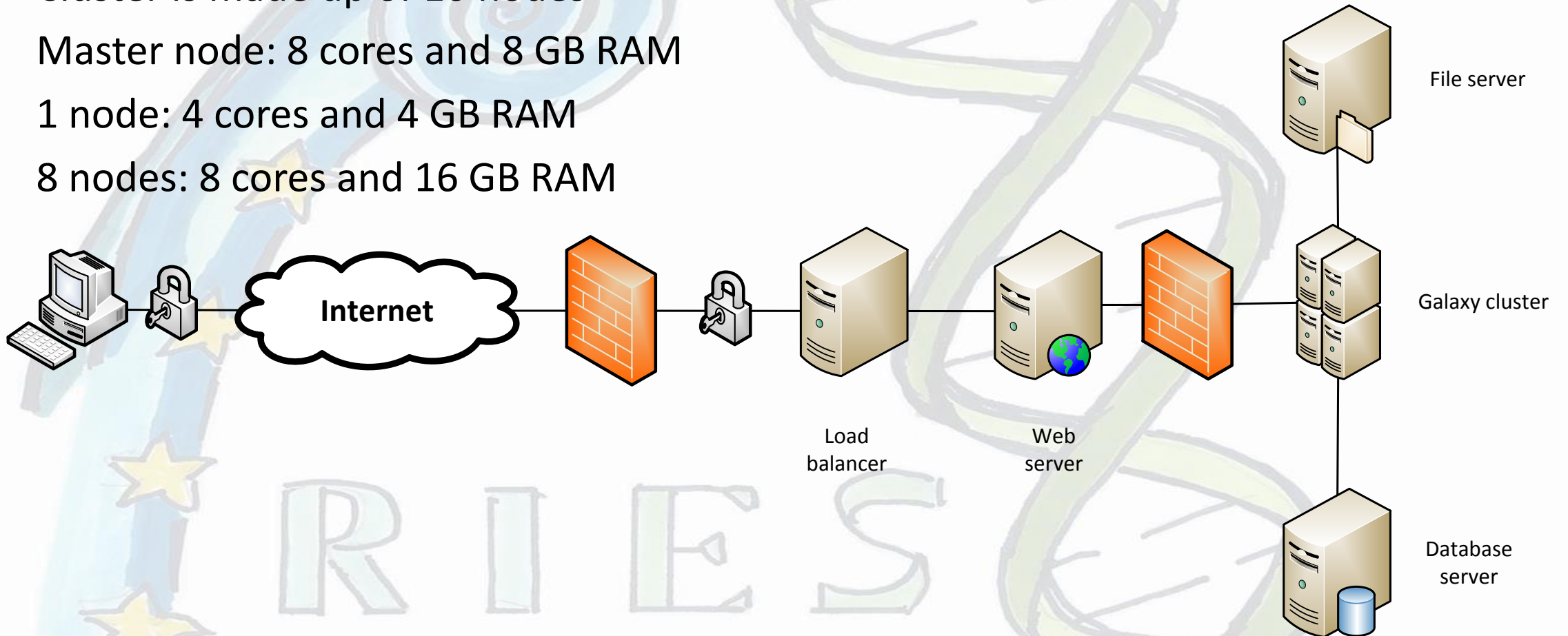
```
<tool id="binassigner" name="Bin Assigner">
  <description>Home-made tool for bin assignment</description>
  <requirements>
    <requirement></requirement>
  </requirements>
  <command interpreter="python">
    BinAssigner.py -t $tmstab -o $output -c $columnstab > $logfile
  </command>
  <inputs>
    <param name="tmstab" type="data" format="tabular" label="Melting Temperatures Tab"/>
    <param name="columnstab" type="text" format="integer" label="Column to consider to Assign Bin" />
  </inputs>
  <outputs>
    <data format="tabular" name="output" label="Allele Table"/>
    <data format="txt" name="logfile" label="BinAssigner Log File" />
  </outputs>
  <help>
    Loads the file with the temperatures; matches the temperatures with reference bins and assigns
    the corresponding alleles
  </help>
  <citations>
    <citation type="doi">10.3389/fmicb.2016.00202</citation>
  </citations>
</tool>
```

Galaxy in ISS: ARIES

- First standalone installation in April 2014
- Reinstallation in June 2014
- Installation of the cluster in March 2015
- Alpha test NGS Course last year in June 2015
- Presentation of ARIES in ISS in June 2015
- ARIES opened for ISS users in July 2015
- ARIES went public in September 2015

ARIES cluster

- Cluster is made up of 10 nodes
- Master node: 8 cores and 8 GB RAM
- 1 node: 4 cores and 4 GB RAM
- 8 nodes: 8 cores and 16 GB RAM



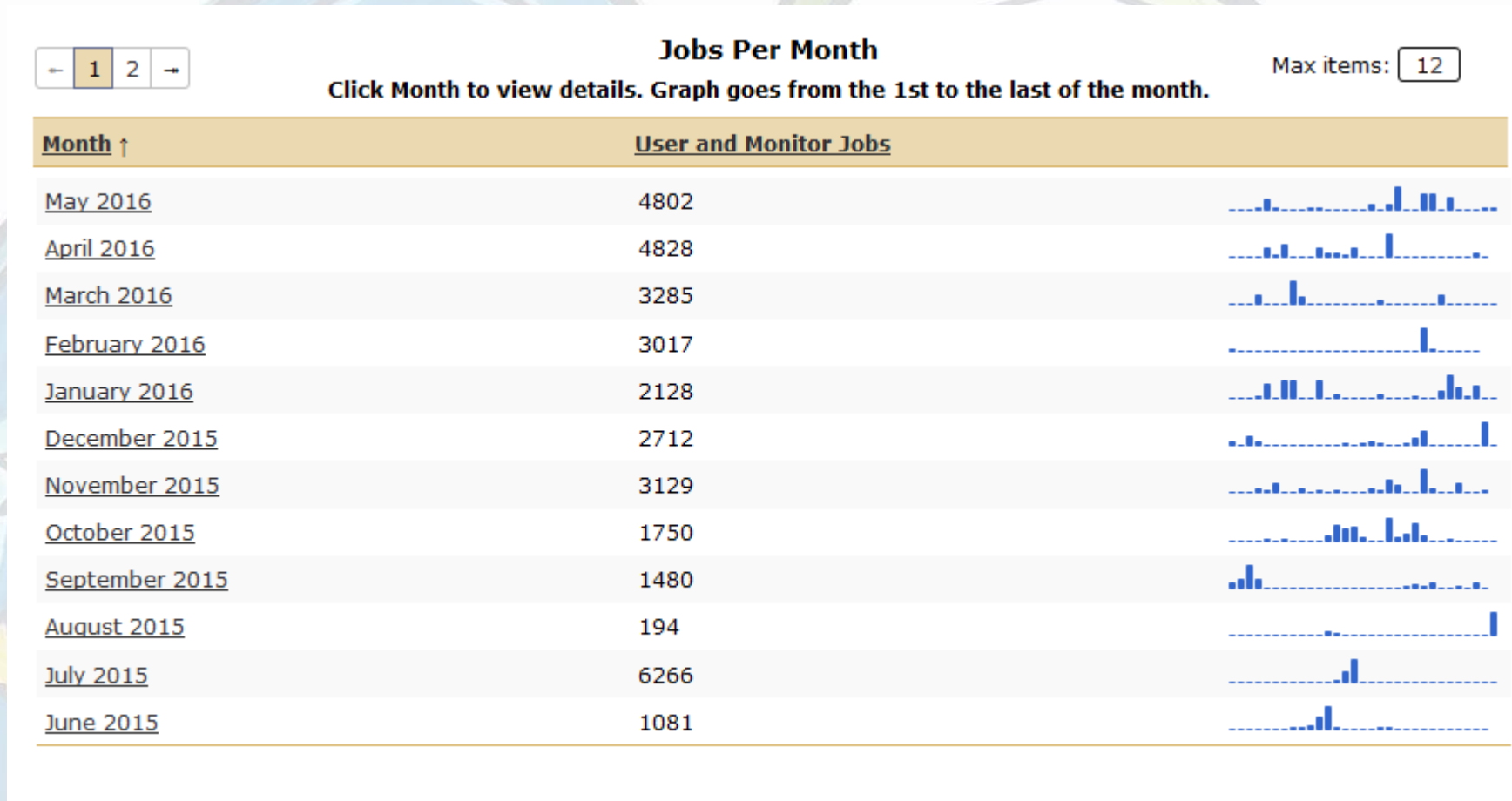
ARIES users

Currently 40 users:

- 4 Belgium
- 3 Finland
- 1 France
- 1 Ireland
- 22 Italy (16 ISS)
- 1 Luxembourg
- 5 The Netherlands
- 1 Poland
- 2 Spain



ARIES use



Questions?

- aries@iss.it
- Follow us: @ARIES_GENOMICS

ARIES