Thorin Tabor
JupyterDays Boston 2016

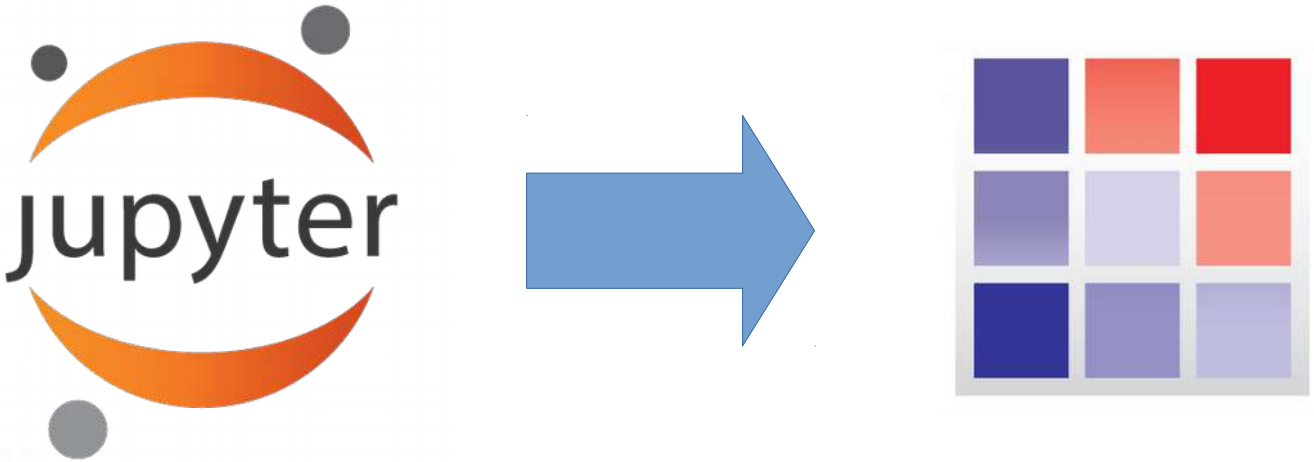
GenePattern Notebooks

Jupyter for Bioinformatic Research

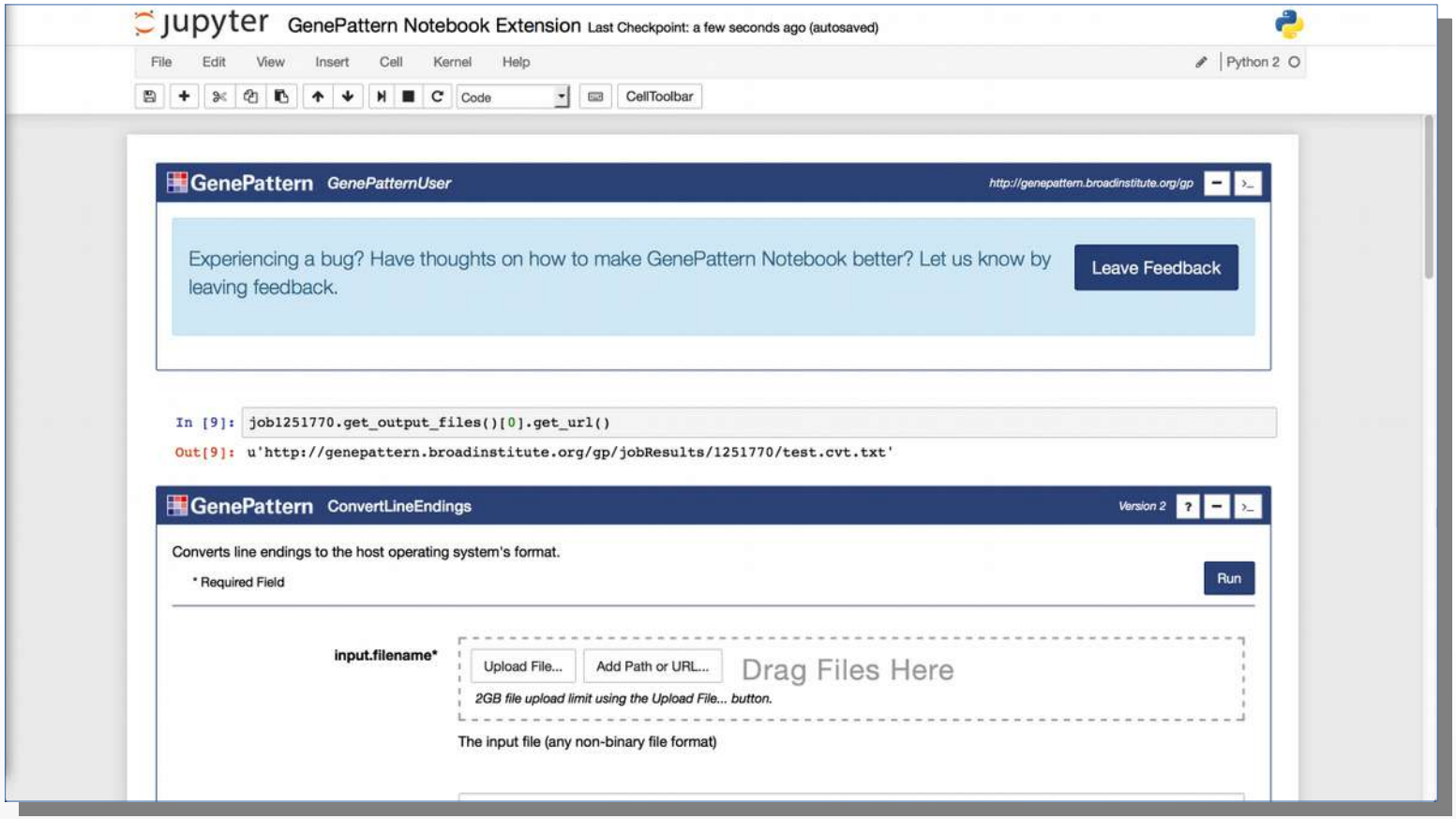
UC San Diego

BROAD INSTITUTE

From Jupyter to GenePattern



GenePattern Notebook



Two Open Source Projects

What is GenePattern?

Module Repository

NMF | Cuffdiff | GISTIC | GSEA | Tophat | MuTect

Hundreds of analyses and visualizations

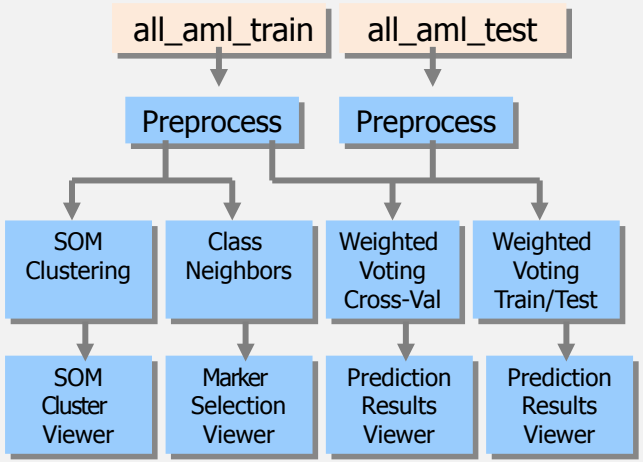
Module Integrator

Easy addition of new tools

Community Repository

User-contributed modules

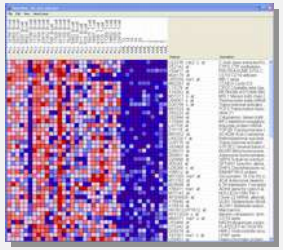
Pipeline Environment



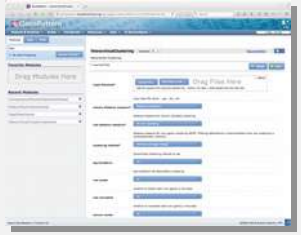
Golub and Slonim et. al 1999

Support for *in silico* reproducible research

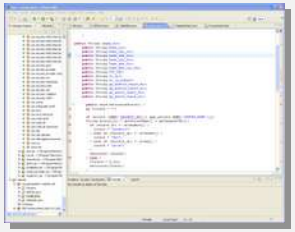
Clients



Visualizer



Web



Programming

Access for all levels of user

Platform for Reproducible Bioinformatic Research

- First public release in 2004 (similar footing to IPython)
- Open Source
- ~40,000 registered users
- Public server runs ~4,000 analyses per week
- Community-contributed methods
  - CRISPR analysis
  - Bisulfite sequencing
  - Flow cytometry
  - RNAi screens

genepattern.org

genepattern.broadinstitute.org

gparc.org

Analytical Tool Repository

Copy Number Divide by Normals	GSEA	Variation Filter	Cuffdiff
GISTIC	CBS	k-Nearest Neighbors	MutSigCV
Classification and Regression Trees	Support Vector Machines	Hierarchical Clustering	Picard Sort Sam
TopHat	Expression File Creator	Metagene Projection	RNASeQC

Custom Modules & Pipelines

Modules

Hierarchical Clustering Files

HCL.jar
cluster.sh
ant.jar
gp-modules.jar
Jama-1.0.2.jar

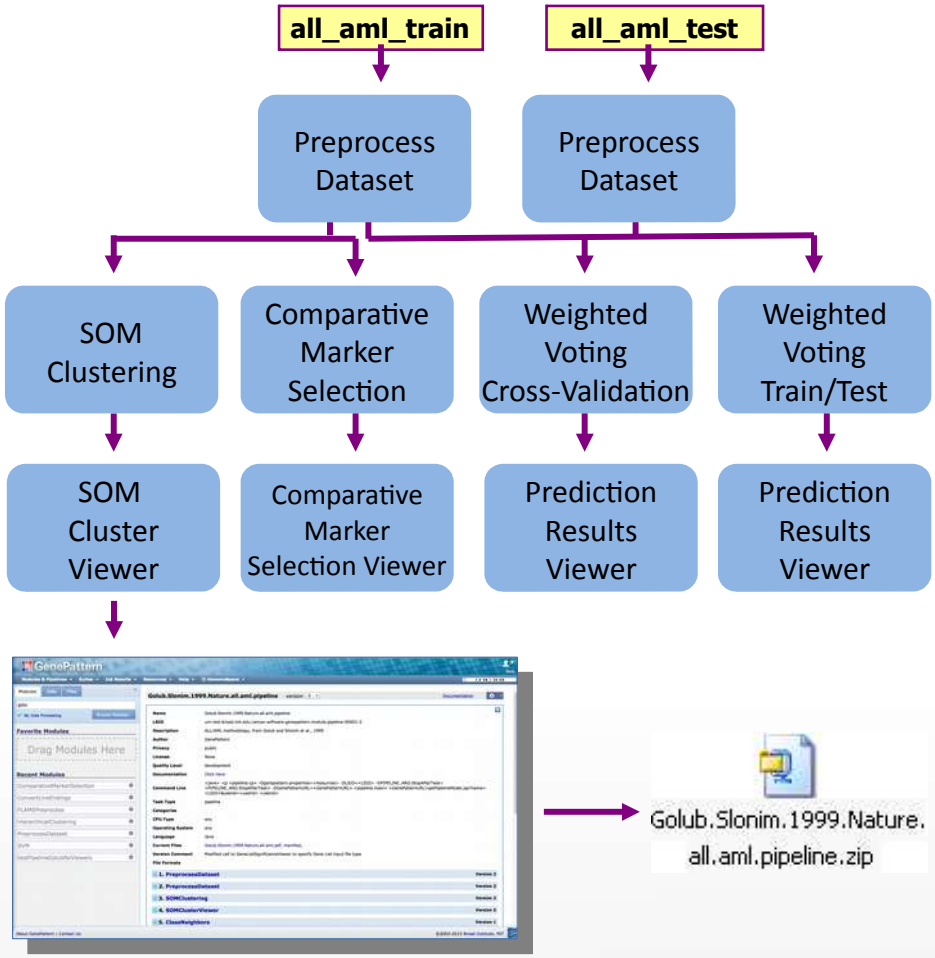
Documentation

HierarchicalClustering.pdf

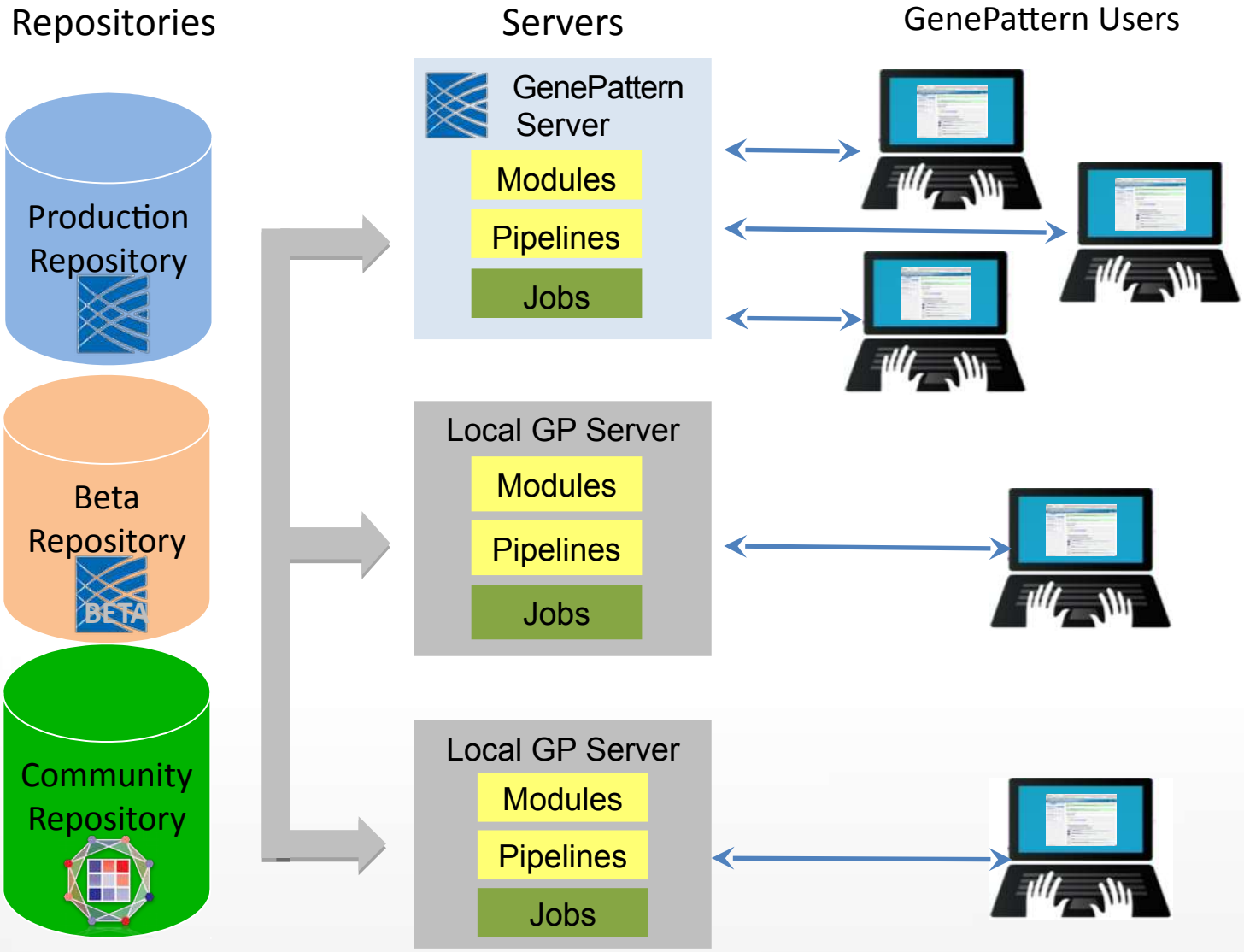
Parameter descriptions

-f <input.filename>
<log.transform>
<row.center>
<row.normalize>
<column.center>
<column.normalize>
-u <output.base.name>
-e <column.distance.measure>
-g <row.distance.measure>
-m <clustering.method>

Pipelines

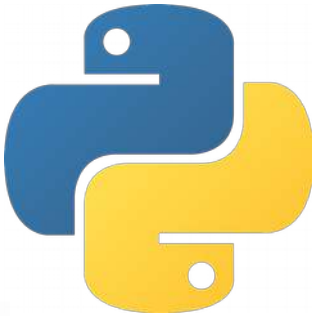


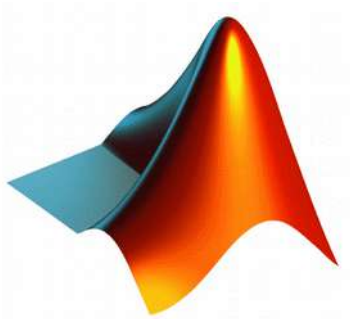
Web Server Architecture



Programmatic APIs

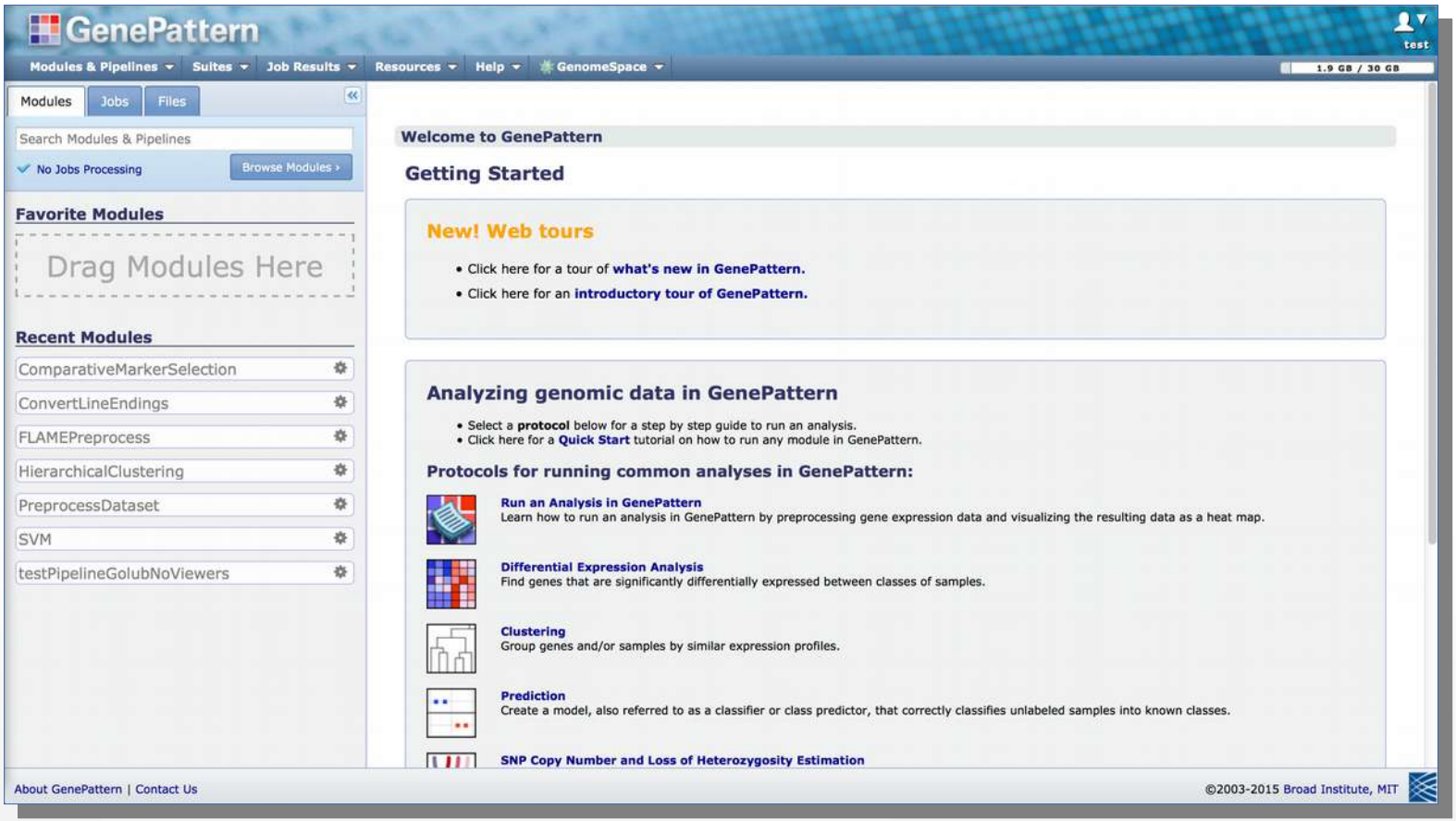
- Libraries for Python, R, MATLAB & Java
- REST API
- Used to back portals and other web applications



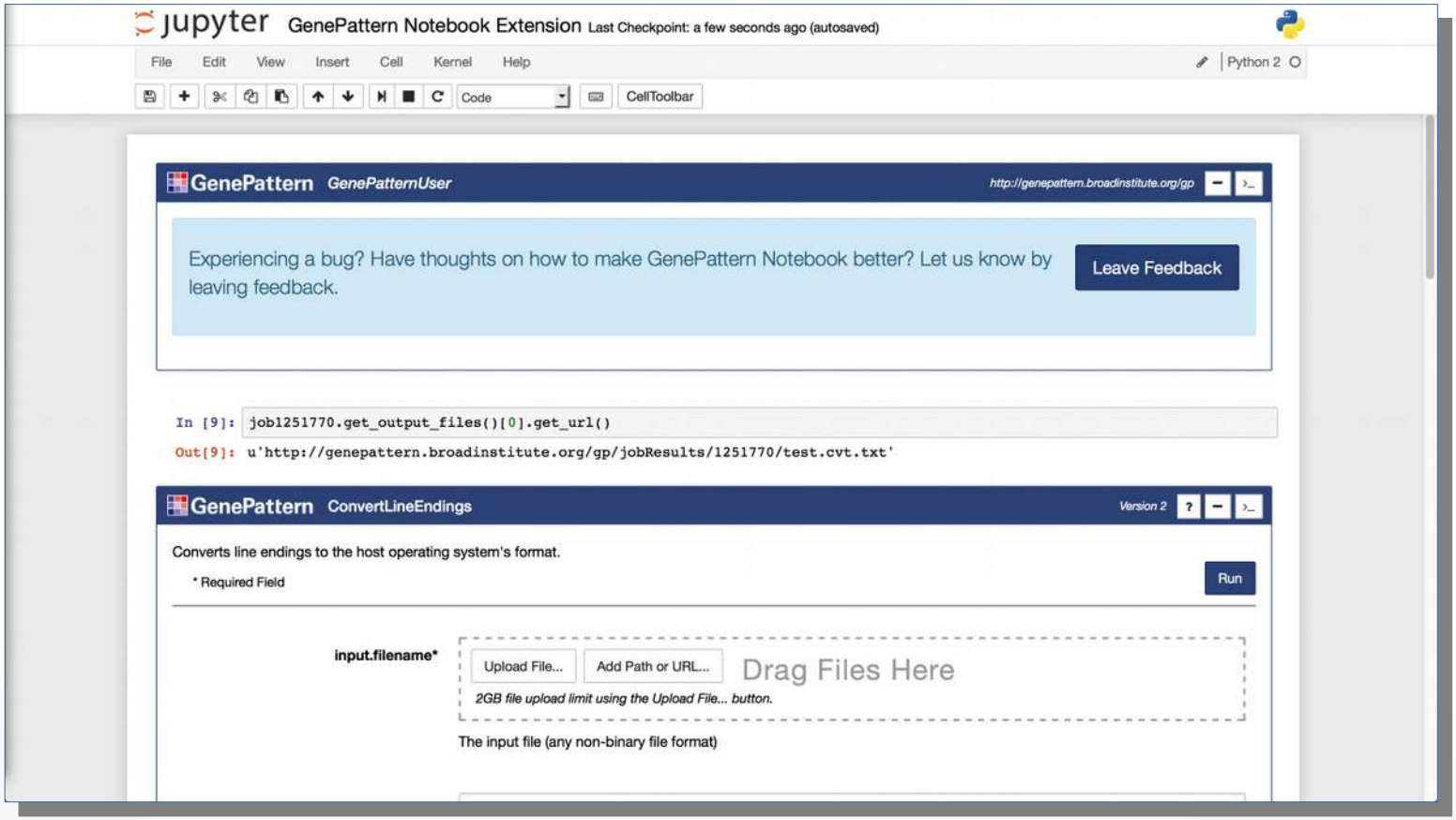


User-Friendly Interface

- Permits complex analyses without the need for a coding background

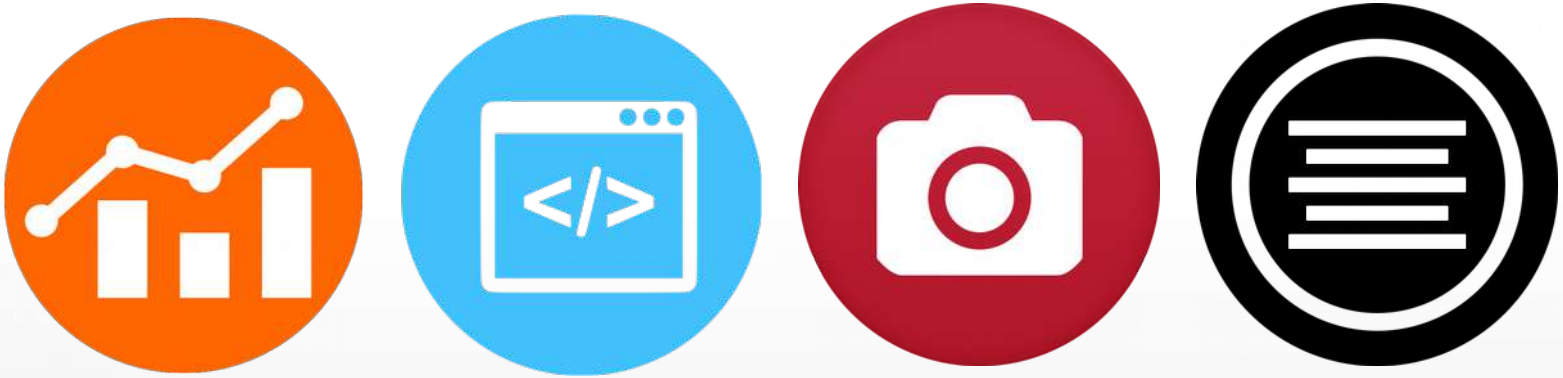


GenePattern Notebook Jupyter Extension



Complete Research Narrative

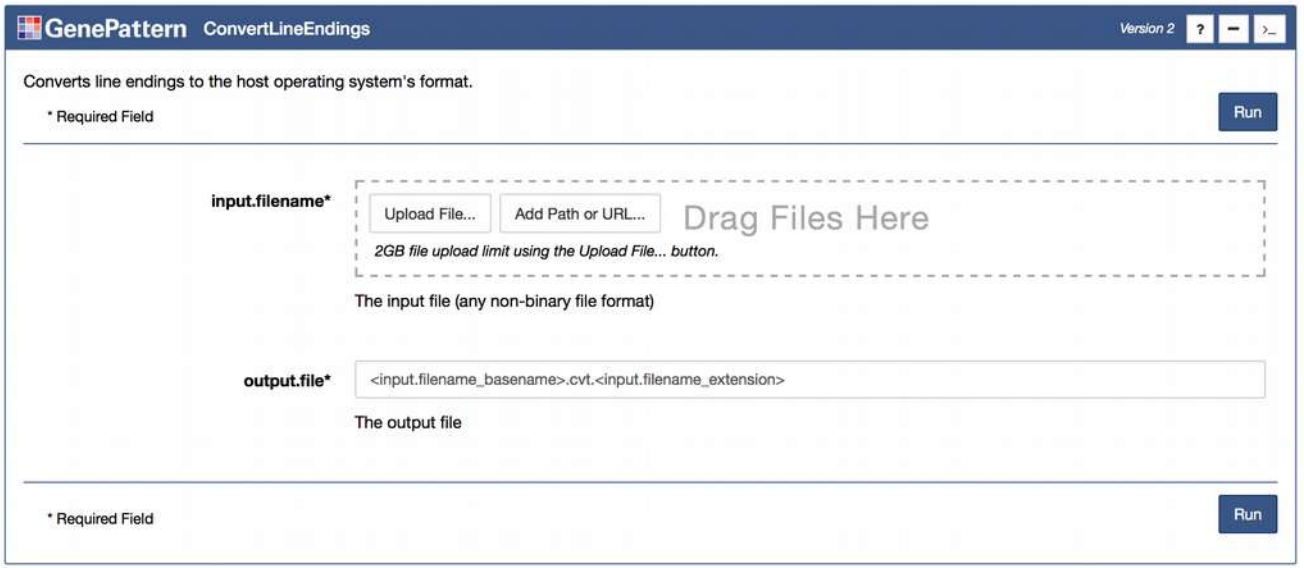
- Leverage the best of Jupyter and GenePattern
- Interleave text, visualizations, graphics and analytical aspects



GenePattern Cells

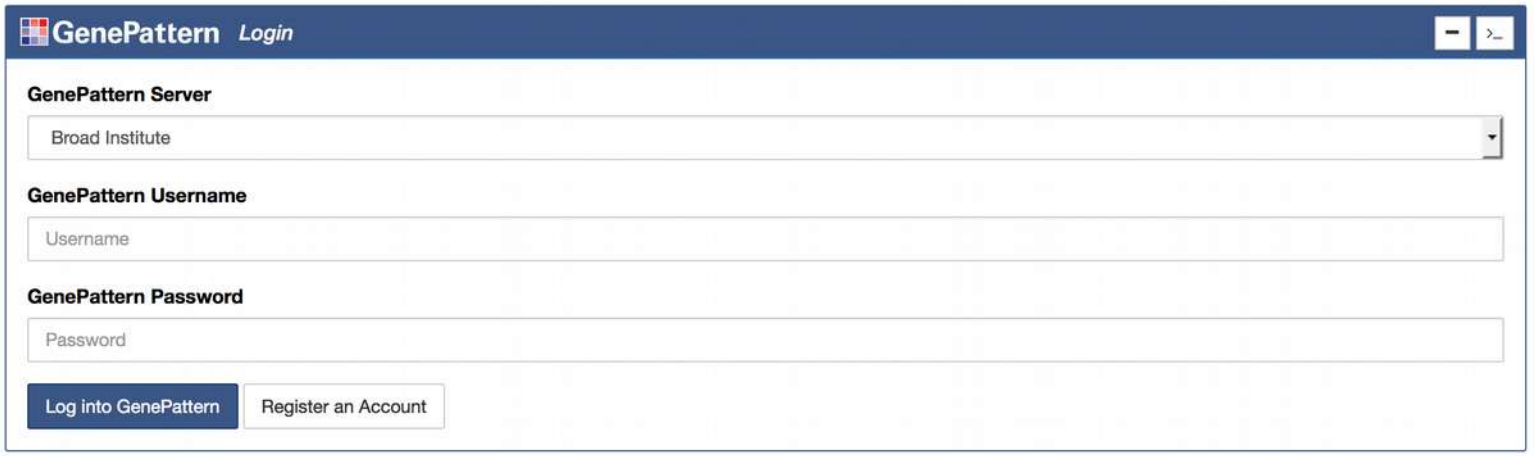
Auth Cell

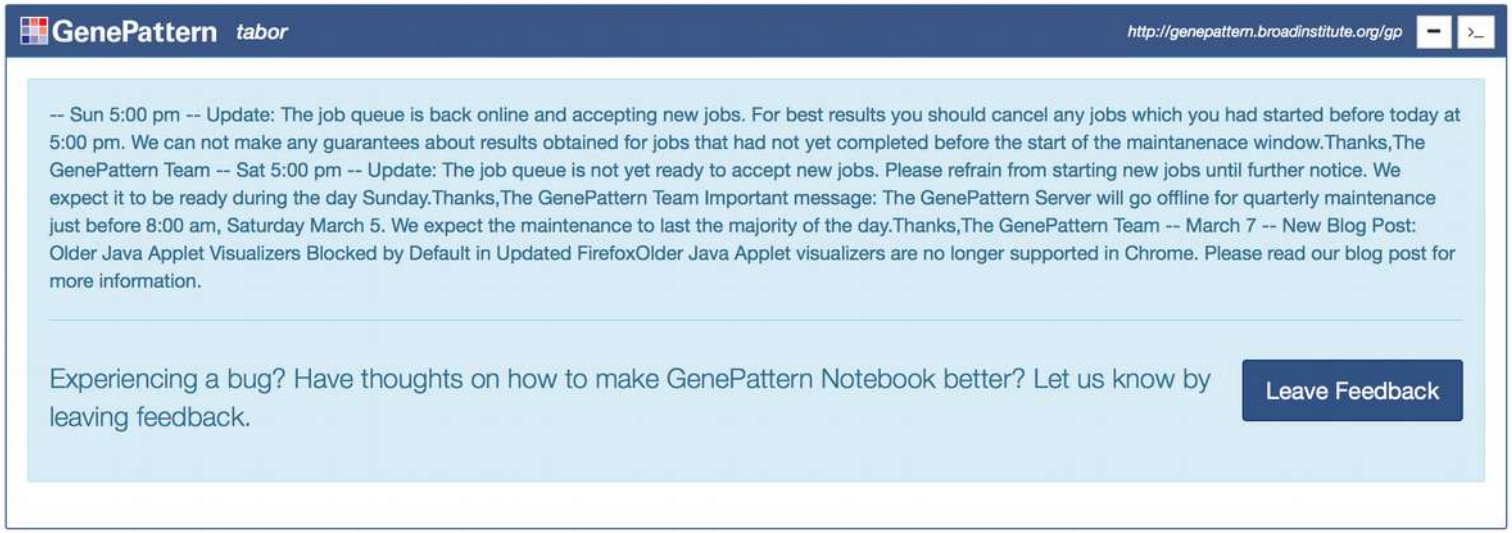
Analysis Cell



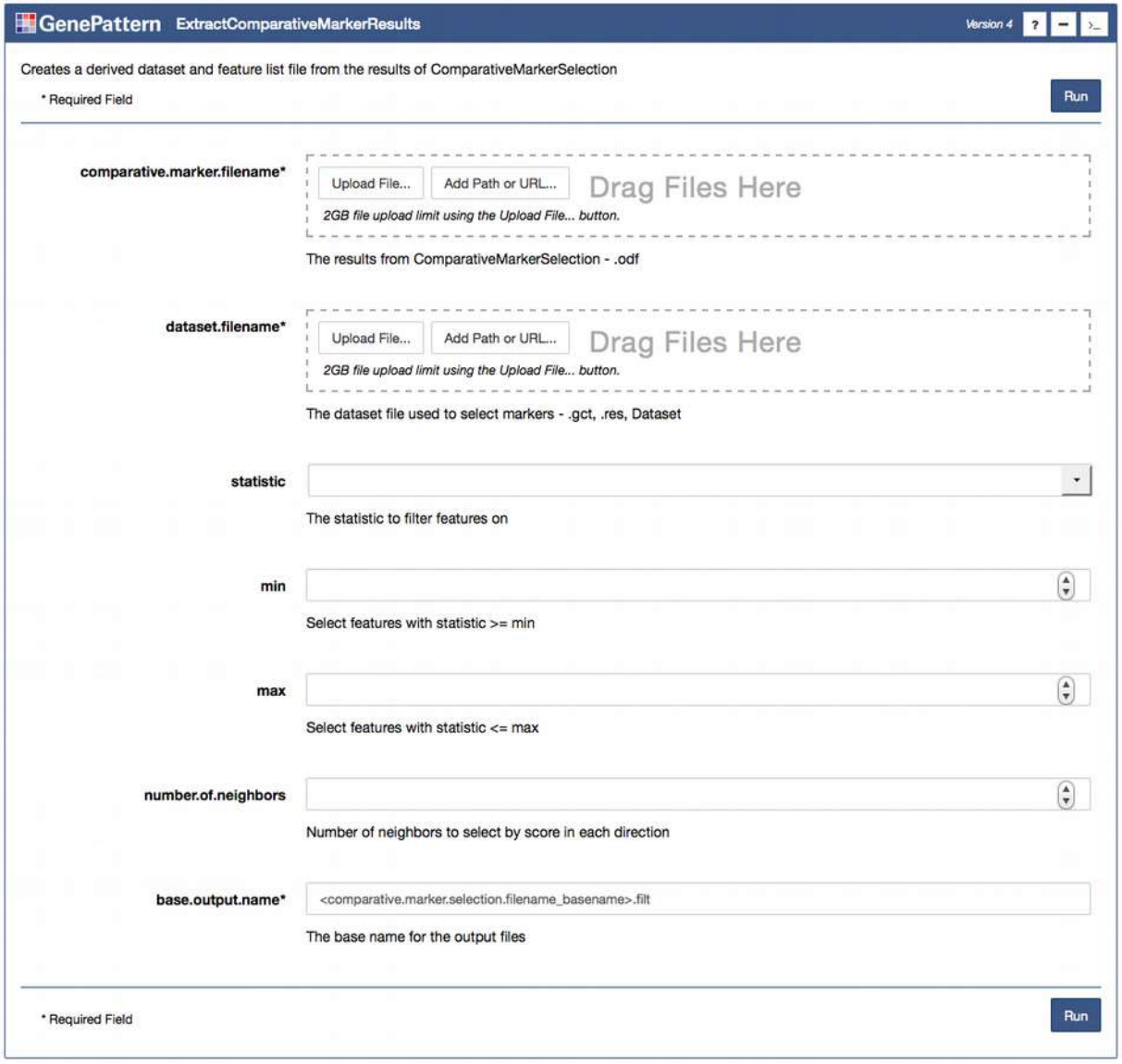
Job Cell

Authentication Cells





Analysis Cells



Job Cells

GenePattern 1251770. ConvertLineEndings

Submitted by tabor on 2016-03-03T12:09:39-05:00

Completed

test.cvt.txt

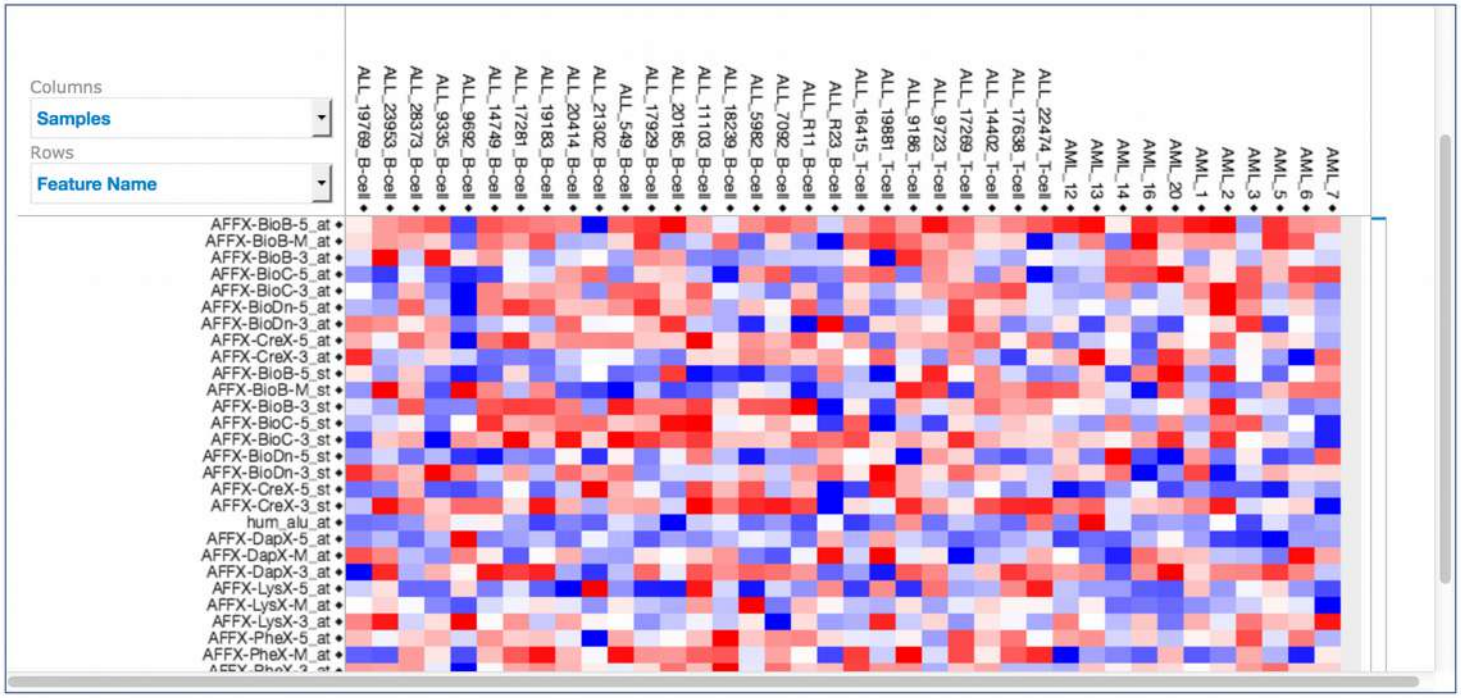
gp_execution_log.txt

GenePattern 129041. HeatMapViewer

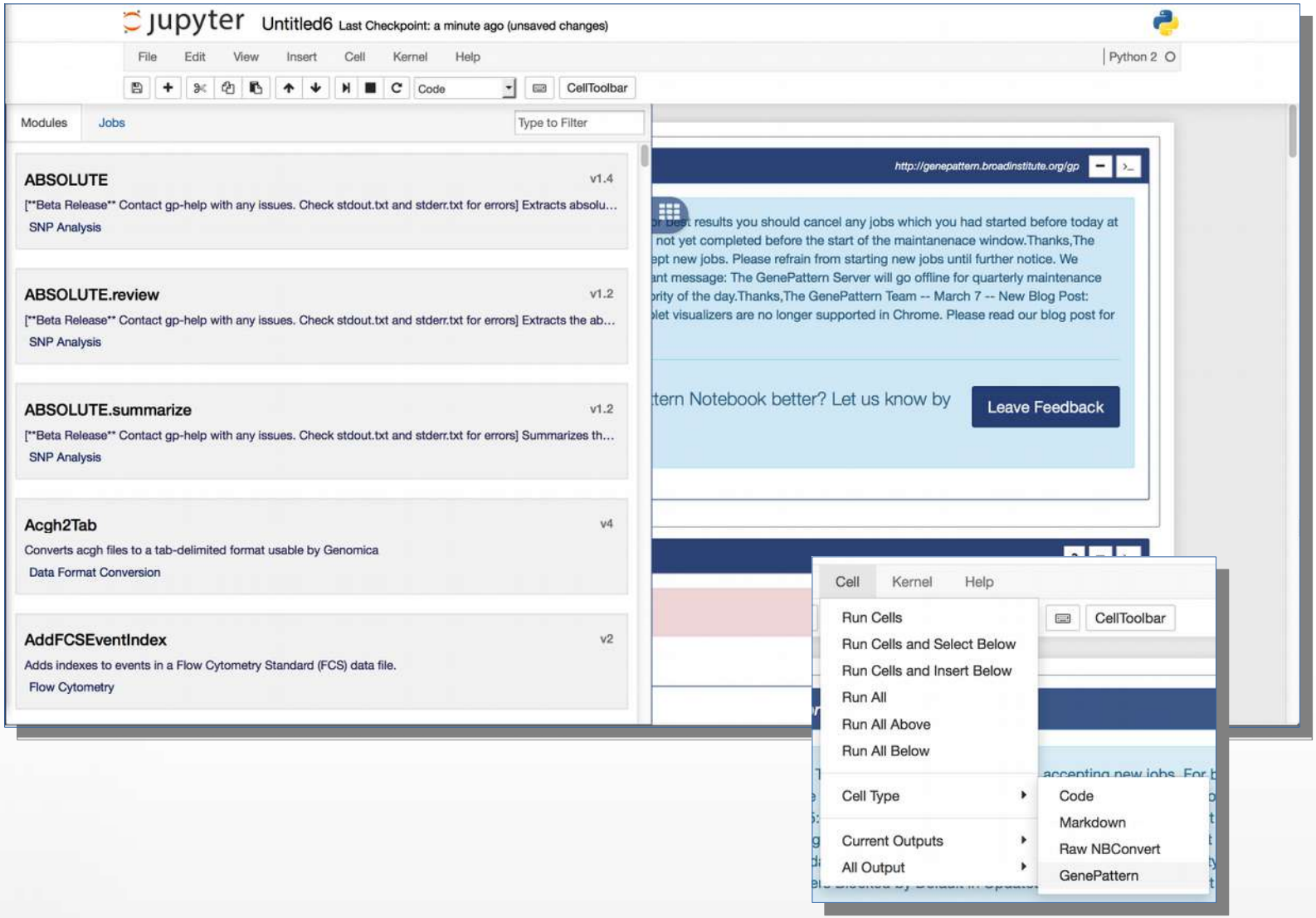
Submitted by tabor on 2016-03-11T13:18:04-05:00

Completed

gp_execution_log.txt

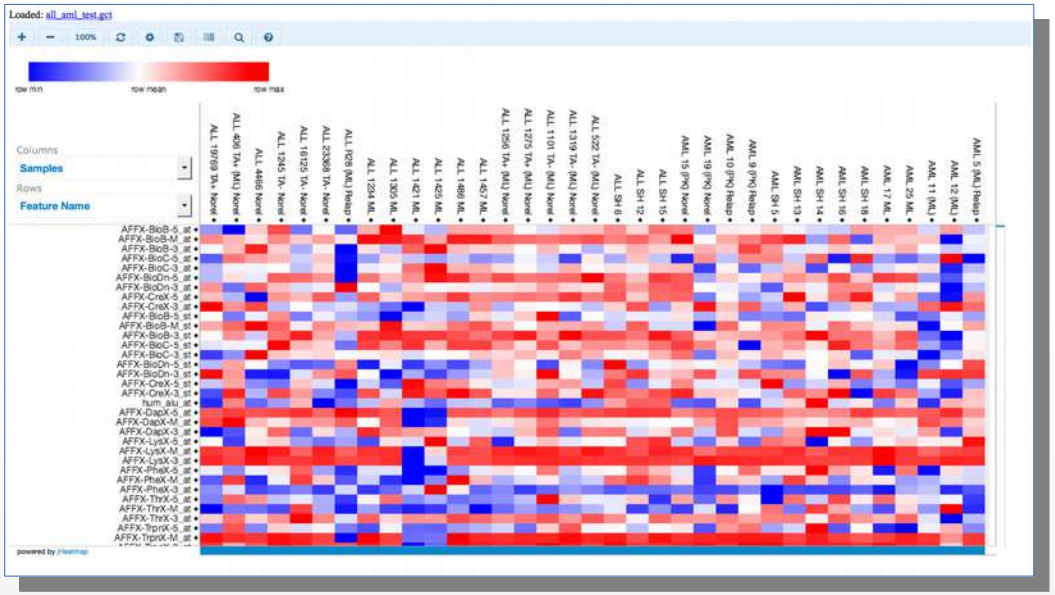


User Interface



Jupyter Widgets

- Interactive widgets use the Jupyter widget framework (ipywidgets, traitlets)
- Graphical representation of Python code
- Not limited by GenePattern analyses



GenePattern Python Library

- Complete programmatic access
- Automatic integration with GenePattern cell data

```
import gp

# Create a GenePattern server proxy instance
gpserver = gp.GPServer('http://localhost:8080/gp','myusername', 'mypassword')

# Obtain GPTask by module name
module = gp.GPTask(gpserver, "PreprocessDataset")

# Load module parameter data
module.param_load()

# Create a job specification
job_spec = module.make_job_spec()

# Upload a file to the server
uploaded_file = gpserver.upload_file("file_name", "/path/to/the/file/file_name")
job_spec.set_parameter("input.filename", uploaded_file.get_url())

# Submit the job to the GenePattern server
job = gpserver.run_job(job_spec)
```

GenePattern Magics

- New Jupyter line magics
- Shortcuts for programmatically creating new GenePattern cells

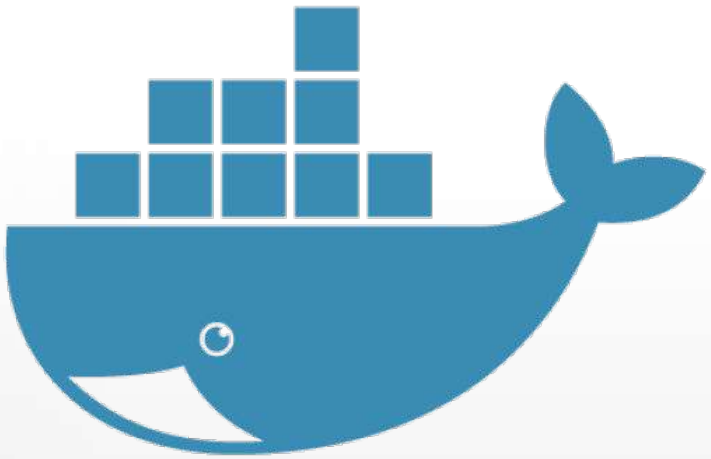
```
# Jupyter magic to obtain reference to Task
task = %get_task $gpserver ConvertLineEndings

# Jupyter magic to obtain reference to Job
job = %get_job $gpserver 1170434

# Turn task object into task widget
GPTaskWidget(task)

# Turn job object into job widget
GPJobWidget(job)
```

JupyterHub

- JupyterHub integration via GenePattern OAuth2 authenticator
- GenePattern Notebook Docker images on DockerHub

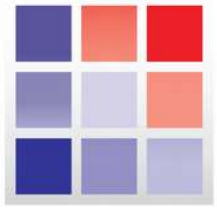
Installing the Extension

- PyPI
  - pip install genepattern-notebook
- DockerHub
  - docker pull genepattern/genepattern-notebook



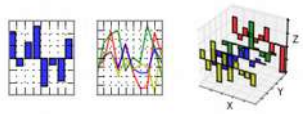
Jupyter Ecosystem

matplotlib



pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



IP[y]:
IPython

Anaconda

jupyterhub

Acknowledgments

GenePattern Team

Peter Carr
David Eby
Barbara Hill
Marc-Danie Nazaire
Michael Reich
Thorin Tabor
Helga Thorvaldsdottir

PI

Jill Mesirov

GenePattern Notebook
funded by the National Cancer Institute's Informatics Technology for Cancer Research

GenePattern Server
funded by the National Institute of General Medical Sciences

Resources

GenePattern
genepattern.org

Public GenePattern server
genepattern.broadinstitute.org

Indiana University GenePattern server
gp.indiana.edu

GenePattern Archive (GPArc)
gparc.org

GenePattern Notebook
genepattern.org/genepattern-notebooks

GenePattern Twitter
@genepattern

GenePattern GitHub
github.com/genepattern

GenePattern DockerHub
hub.docker.com/r/genepattern