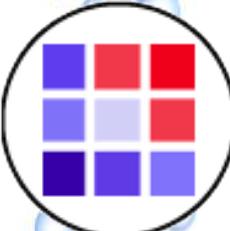


Thorin Tabor
JupyterDays Boston 2016

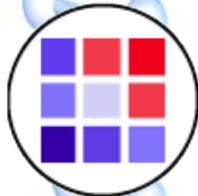


GenePattern Notebooks

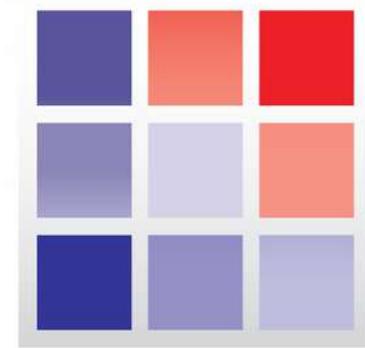
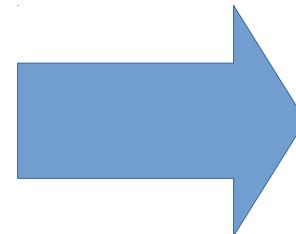
Jupyter for Bioinformatic Research

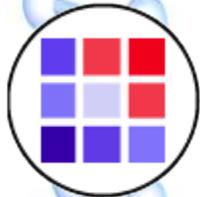
UC San Diego





From Jupyter to GenePattern





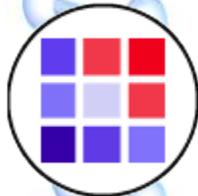
GenePattern Notebook

The screenshot displays the GenePattern Notebook Extension running within a Jupyter notebook environment. The top navigation bar includes File, Edit, View, Insert, Cell, Kernel, Help, and a Python 2 kernel selector. A toolbar below the menu bar contains various icons for file operations like Open, Save, and Print.

A central message box from GenePattern prompts users to leave feedback, with a "Leave Feedback" button. Below this, a code cell shows the execution of a job and its output URL:

```
In [9]: job1251770.get_output_files()[0].get_url()
Out[9]: u'http://genepattern.broadinstitute.org/gp/jobResults/1251770/test.cvt.txt'
```

Below the code cell is a GenePattern ConvertLineEndings tool window. It has a "Run" button and a "Drag Files Here" input field for file uploads. The input field also includes "Upload File..." and "Add Path or URL..." buttons, and a note about a 2GB file upload limit.



Two Open Source Projects





What is GenePattern?

Module Repository



Hundreds of analyses and visualizations

Module Integrator



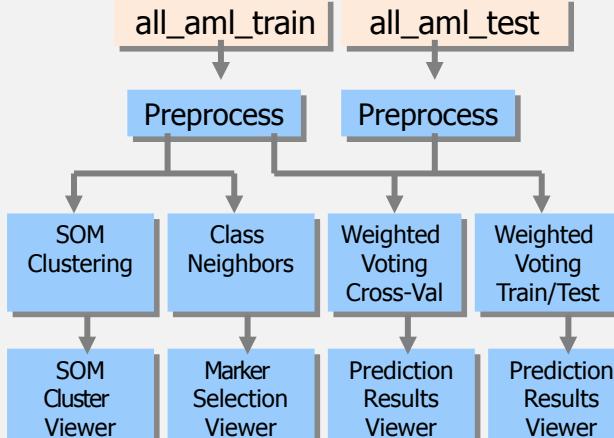
Easy addition of new tools

Community Repository

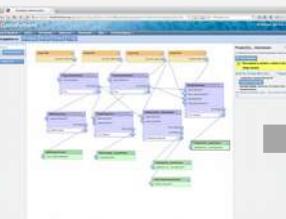


User-contributed modules

Pipeline Environment

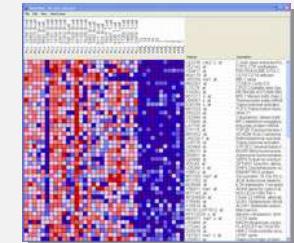


Golub and Slonim et. al 1999



Support for *in silico* reproducible research

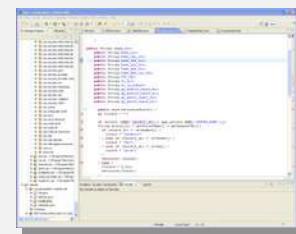
Clients



Visualizer

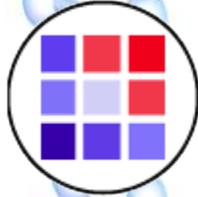


Web



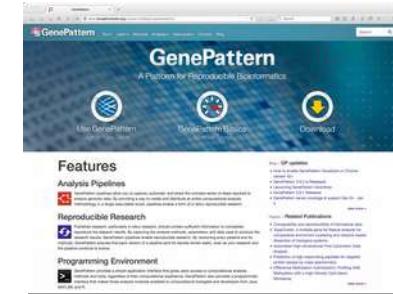
Programming

Access for all levels of user

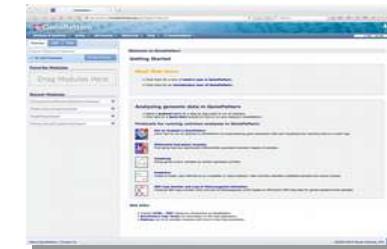


Platform for Reproducible Bioinformatic Research

- 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

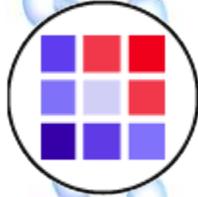
Picard Sort Sam

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

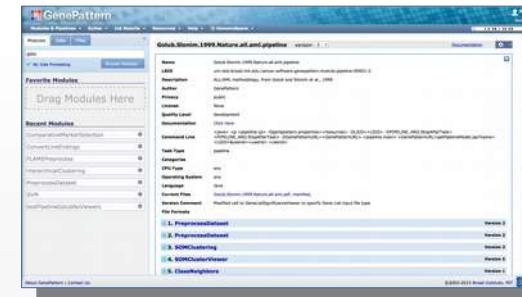
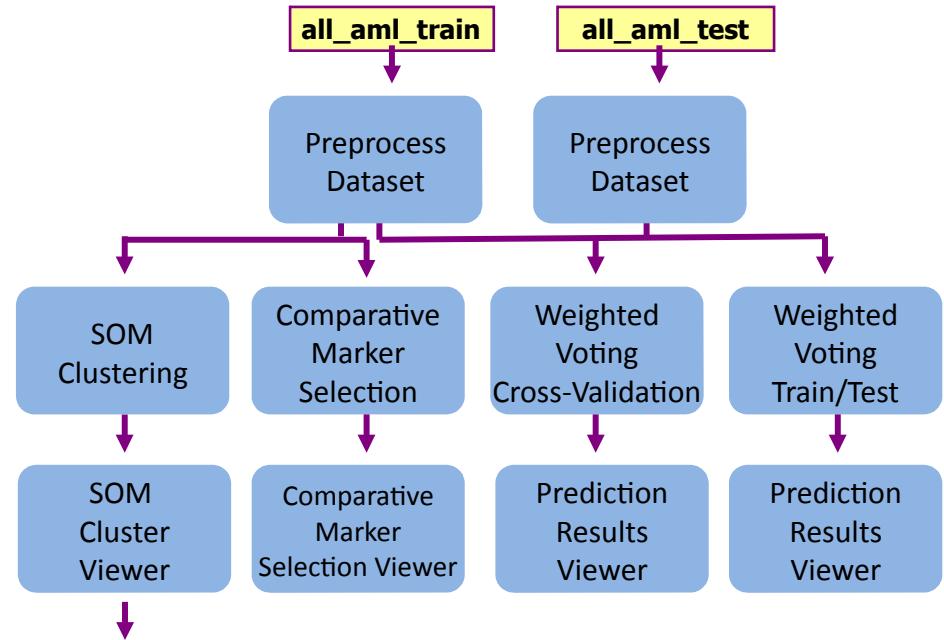
Documentation

HierarchicalClustering.pdf

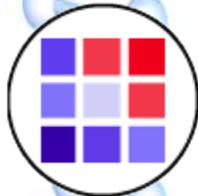
Parameter descriptions

```
-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

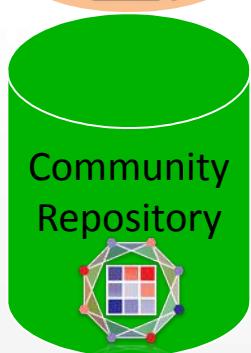
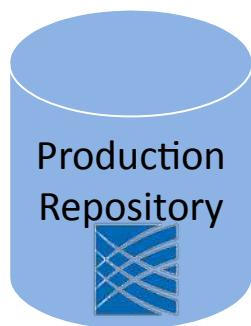


Golub.Slonim.1999.Nature.all.aml.pipeline.zip

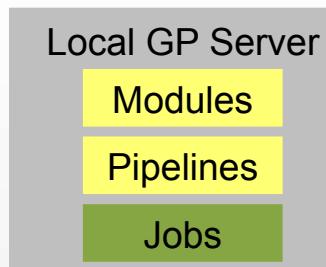
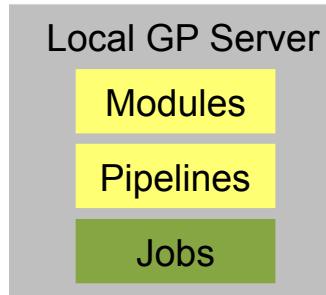
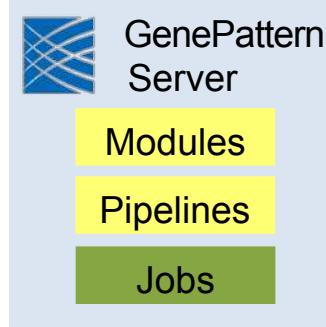


Web Server Architecture

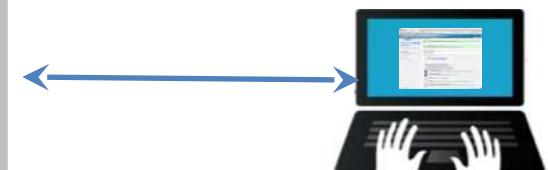
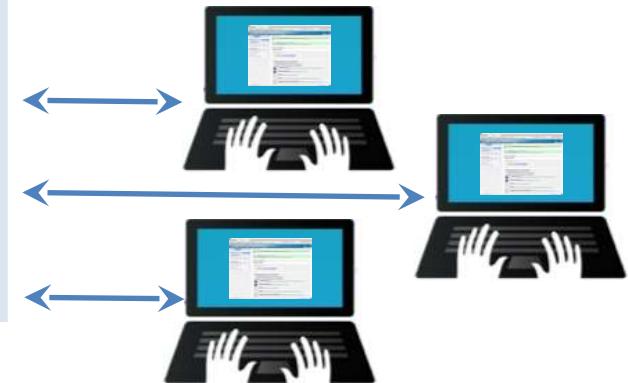
Repositories

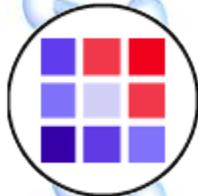


Servers



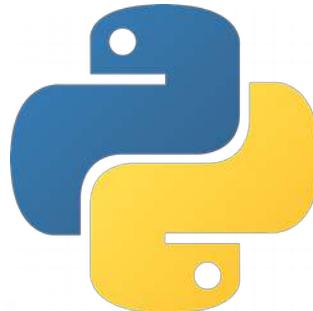
GenePattern Users



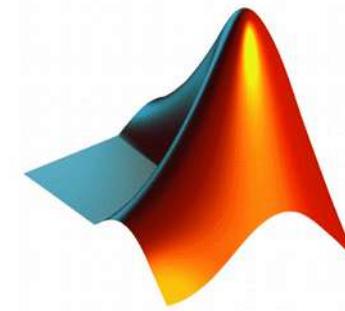


Programmatic APIs

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



REST API

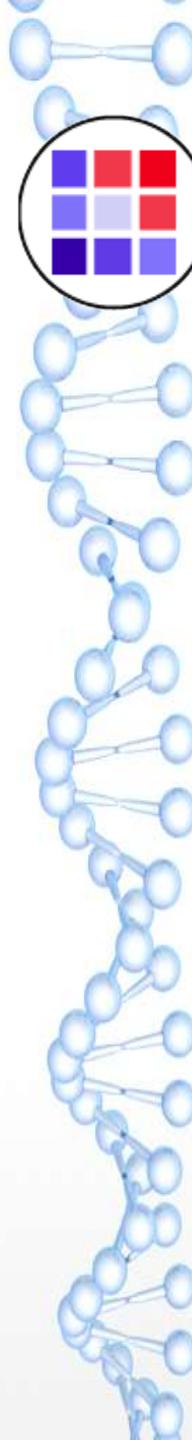




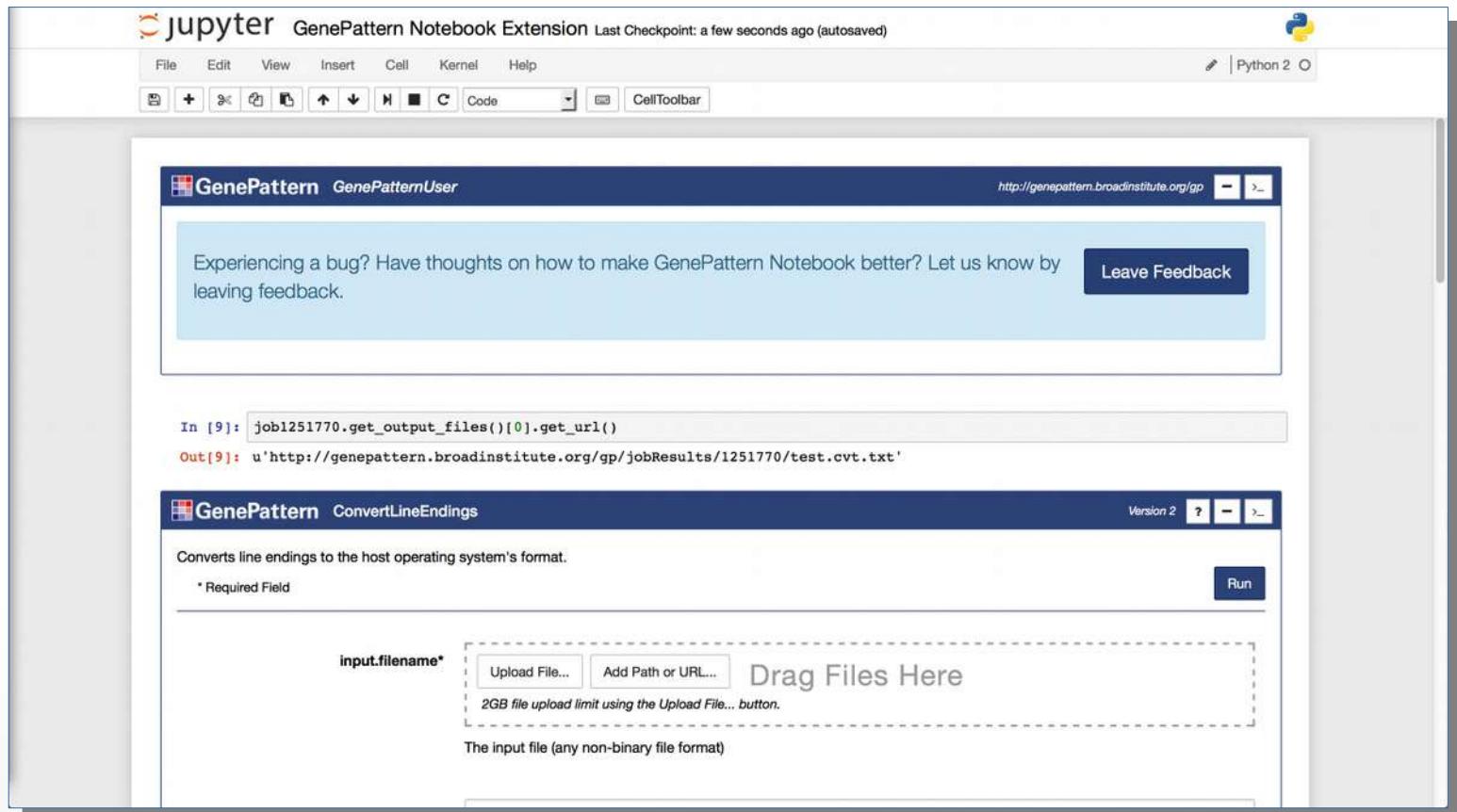
User-Friendly Interface

- Permits complex analyses without the need for a coding background

The screenshot shows the GenePattern software interface. At the top, there's a navigation bar with links for Modules & Pipelines, Suites, Job Results, Resources, Help, and GenomeSpace. A user icon labeled "test" is in the top right corner. The main area has a blue header "Welcome to GenePattern". Below it, a "Getting Started" section features "New! Web tours" with two links: "Click here for a tour of what's new in GenePattern." and "Click here for an introductory tour of GenePattern.". Another section titled "Analyzing genomic data in GenePattern" lists several protocols: "Run an Analysis in GenePattern" (described as learning how to run an analysis by preprocessing gene expression data and visualizing the resulting data as a heat map), "Differential Expression Analysis" (described as finding genes that are significantly differentially expressed between classes of samples), "Clustering" (described as grouping genes and/or samples by similar expression profiles), "Prediction" (described as creating a model, also referred to as a classifier or class predictor, that correctly classifies unlabeled samples into known classes), and "SNP Copy Number and Loss of Heterozygosity Estimation". On the left side, there's a sidebar with sections for "Favorite Modules" (with a "Drag Modules Here" placeholder) and "Recent Modules" (listing ComparativeMarkerSelection, ConvertLineEndings, FLAMEPreprocess, HierarchicalClustering, PreprocessDataset, SVM, and testPipelineGolubNoViewers). At the bottom, there are links for "About GenePattern | Contact Us" and a copyright notice "©2003-2015 Broad Institute, MIT".



GenePattern Notebook Jupyter Extension



In [9]: `job1251770.get_output_files()[0].get_url()`

Out[9]: `u'http://genepattern.broadinstitute.org/gp/jobResults/1251770/test.cvt.txt'`

GenePattern ConvertLineEndings

Converts line endings to the host operating system's format.

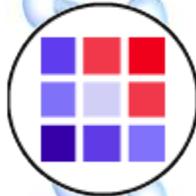
* Required Field

input.filename*

Upload File... Add Path or URL... Drag Files Here

2GB file upload limit using the Upload File... button.

The input file (any non-binary file format)



Complete Research Narrative

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





GenePattern Cells

Auth
Cell

GenePattern Login

GenePattern Server

Broad Institute

GenePattern Username

Username

GenePattern Password

Password

Log into GenePattern Register an Account

Analysis
Cell

GenePattern ConvertLineEndings

Converts line endings to the host operating system's format.

* Required Field

Run

input.filename*

Upload File... Add Path or URL... Drag Files Here

2GB file upload limit using the Upload File... button.

The input file (any non-binary file format)

output.file*

<input.filename_basename>.cvt.<input.filename_extension>

The output file

Run

* Required Field

Job
Cell

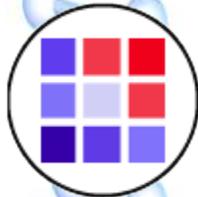
GenePattern 1251770. ConvertLineEndings

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

test.cvt.txt ⓘ

gp_execution_log.txt ⓘ

Completed



Authentication Cells

GenePattern Login

GenePattern Server
Broad Institute

GenePattern Username
Username

GenePattern Password
Password

Log into GenePattern **Register an Account**

GenePattern tabor <http://genepattern.broadinstitute.org/gp>

-- Sun 5:00 pm -- Update: The job queue is back online and accepting new jobs. For best results you should cancel any jobs which you had started before today at 5:00 pm. We can not make any guarantees about results obtained for jobs that had not yet completed before the start of the maintenance window. Thanks, The GenePattern Team -- Sat 5:00 pm -- Update: The job queue is not yet ready to accept new jobs. Please refrain from starting new jobs until further notice. We expect it to be ready during the day Sunday. Thanks, The GenePattern Team Important message: The GenePattern Server will go offline for quarterly maintenance just before 8:00 am, Saturday March 5. We expect the maintenance to last the majority of the day. Thanks, The GenePattern Team -- March 7 -- New Blog Post: Older Java Applet Visualizers Blocked by Default in Updated FirefoxOlder Java Applet visualizers are no longer supported in Chrome. Please read our blog post for more information.

Experiencing a bug? Have thoughts on how to make GenePattern Notebook better? Let us know by leaving feedback.

Leave Feedback



Analysis Cells

GenePattern ExtractComparativeMarkerResults Version 4 ? - > Run

Creates a derived dataset and feature list file from the results of ComparativeMarkerSelection * Required Field

comparative.marker.filename* Drag Files Here
2GB file upload limit using the Upload File... button.
The results from ComparativeMarkerSelection - .odf

dataset.filename* Drag Files Here
2GB file upload limit using the Upload File... button.
The dataset file used to select markers - .gct, .res, Dataset

statistic
The statistic to filter features on

min Select features with statistic \geq min

max Select features with statistic \leq max

number.of.neighbors Number of neighbors to select by score in each direction

base.output.name* <comparative.marker.selection.filename_basename>.filt
The base name for the output files

* Required Field Run



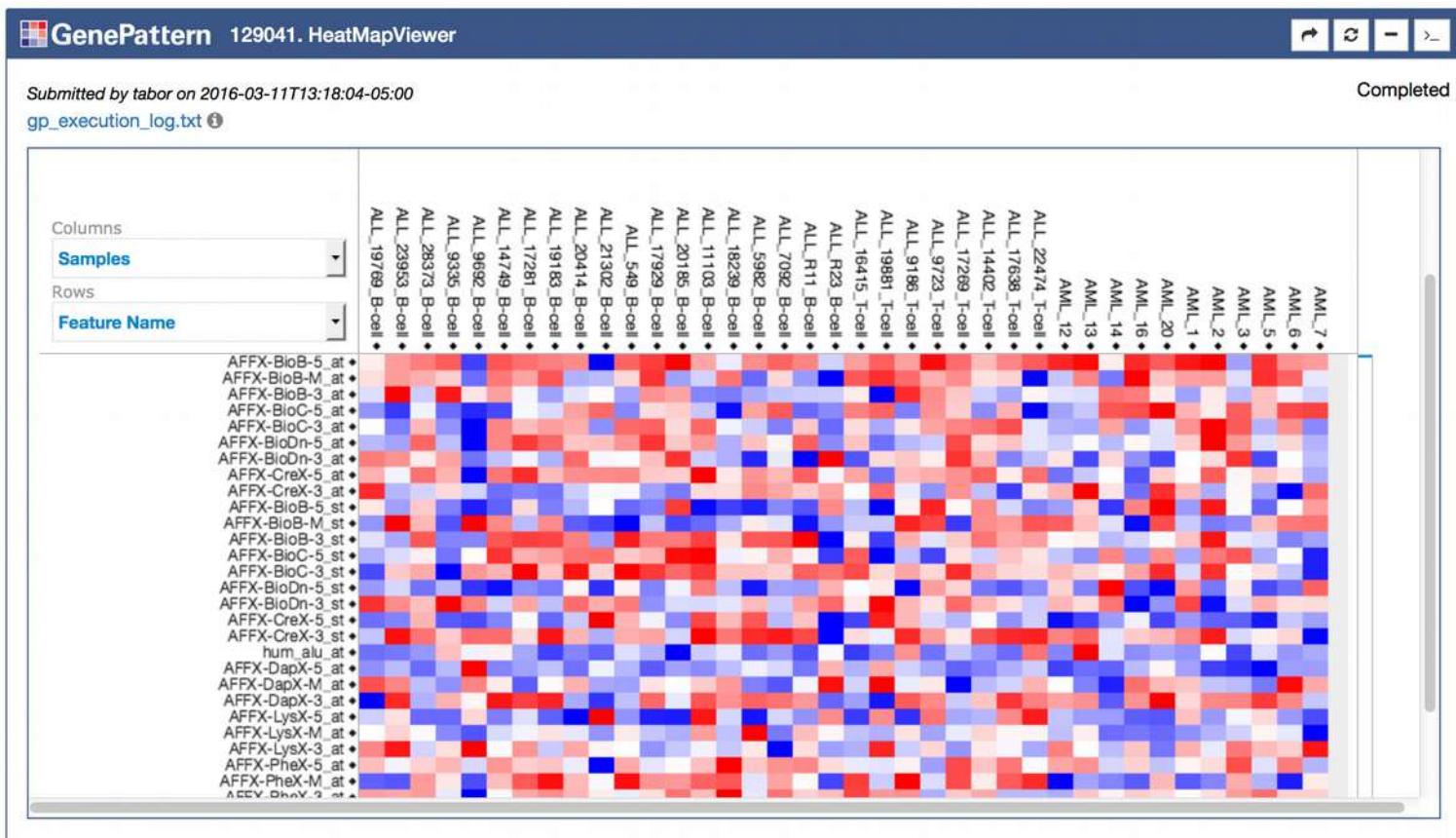
Job Cells

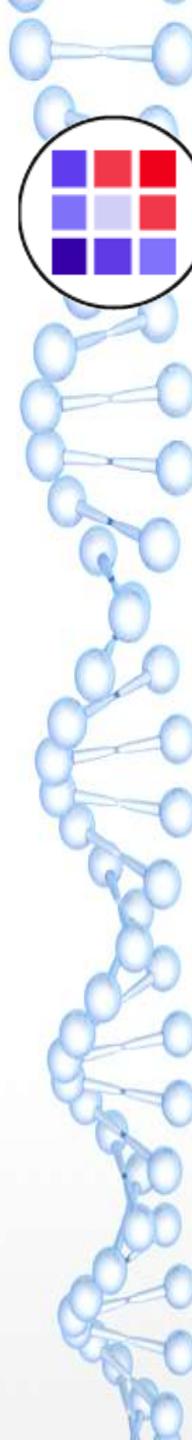
GenePattern 1251770. ConvertLineEndings

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

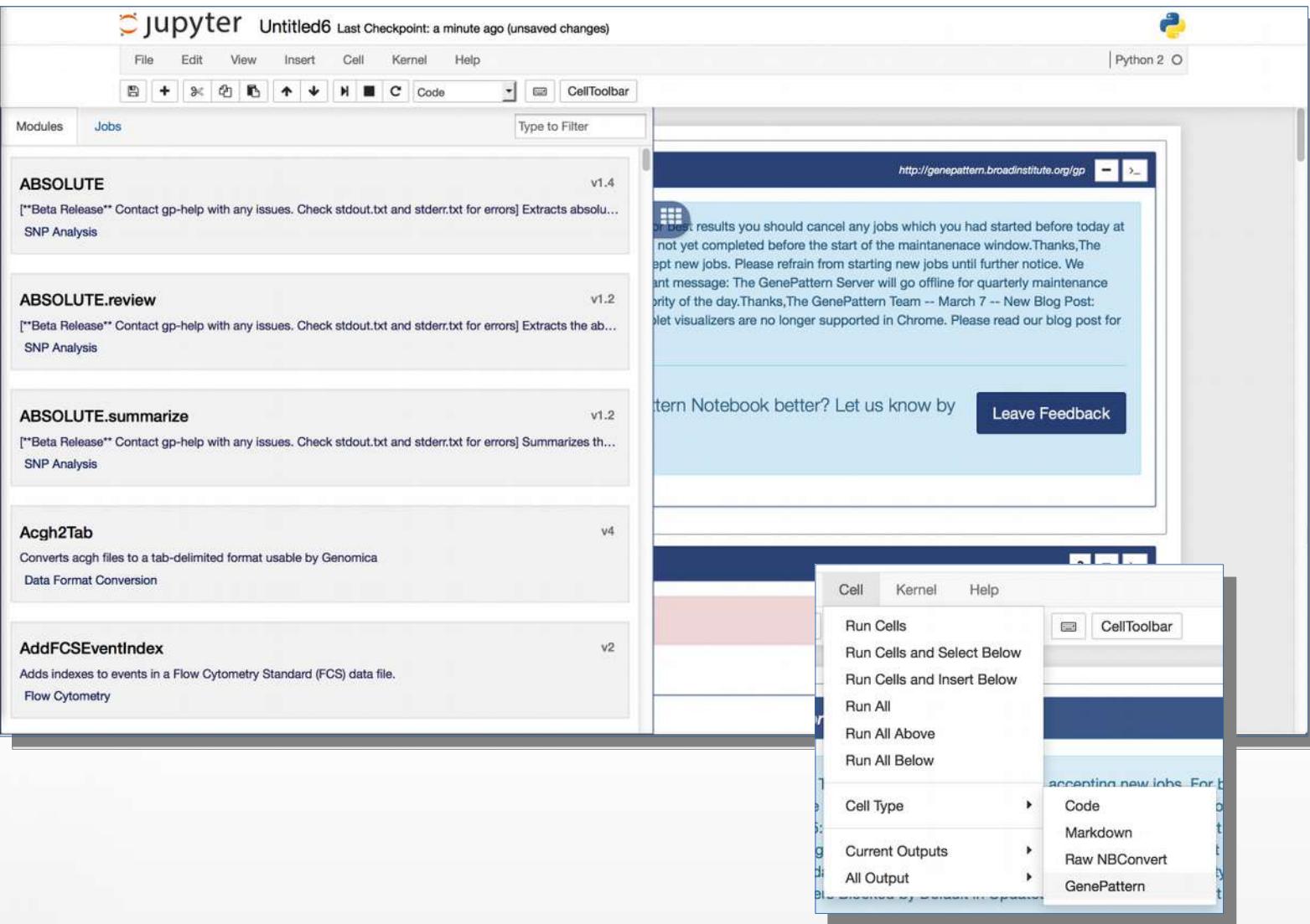
test.cvt.txt ⓘ
gp_execution_log.txt ⓘ

Completed



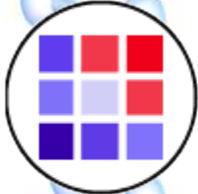


User Interface



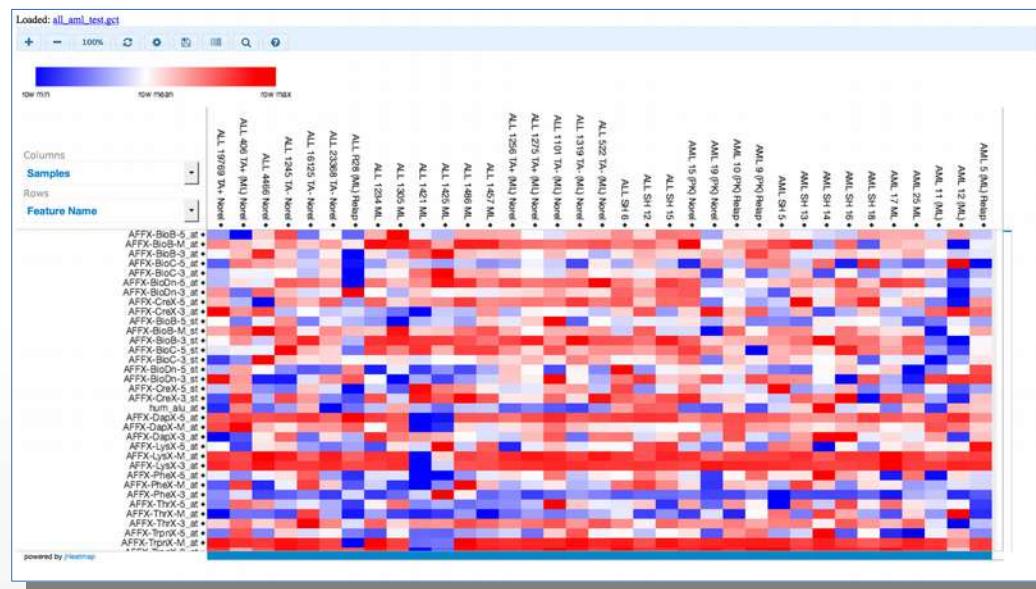
The screenshot shows a Jupyter Notebook interface with the following details:

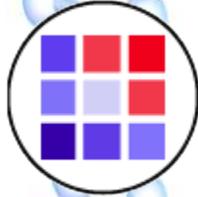
- Header:** jupyter Untitled6 Last Checkpoint: a minute ago (unsaved changes) | Python 2 O
- Toolbar:** File, Edit, View, Insert, Cell, Kernel, Help
- Cell Types:** Code, CellToolbar
- Modules:**
 - ABSOLUTE** v1.4: [Beta Release] Contact gp-help with any issues. Check stdout.txt and stderr.txt for errors] Extracts absolute values from a file.
 - ABSOLUTE.review** v1.2: [Beta Release] Contact gp-help with any issues. Check stdout.txt and stderr.txt for errors] Extracts the absolute values from a file.
 - ABSOLUTE.summarize** v1.2: [Beta Release] Contact gp-help with any issues. Check stdout.txt and stderr.txt for errors] Summarizes the absolute values.
 - Acgh2Tab** v4: Converts acgh files to a tab-delimited format usable by Genomics.
 - AddFCSEventIndex** v2: Adds indexes to events in a Flow Cytometry Standard (FCS) data file.
- Message Box:** A modal window from <http://genepattern.broadinstitute.org/gp> displays a notice about maintenance and job cancellation.
- Cell Context Menu:** A dropdown menu for the "Cell" option in the toolbar, listing options like Run Cells, Run All, and Cell Type.



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')

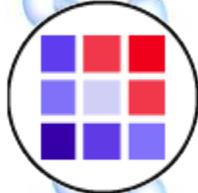
# Obtain GPTask by module name
module = gp.GPTask(gpserver, "PreprocessDataset")

# 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())

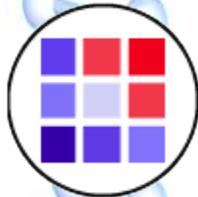
# Submit the job to the GenePattern server
job = gpserver.run_job(job_spec)
```



GenePattern Magics

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

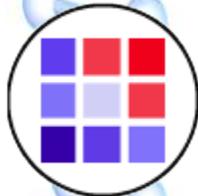
```
# Jupyter magic to obtain reference to Task  
task = %get_task $gpserver ConvertLineEndings  
  
# Jupyter magic to obtain reference to Job  
job = %get_job $gpserver 1170434  
  
# Turn task object into task widget  
GPTaskWidget(task)  
  
# Turn job object into job widget  
GPJobWidget(job)
```



JupyterHub

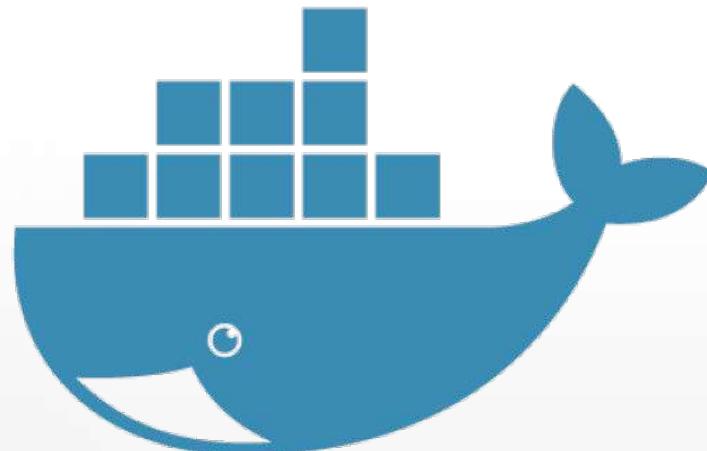
- JupyterHub integration via GenePattern OAuth2 authenticator
- GenePattern Notebook Docker images on DockerHub





Installing the Extension

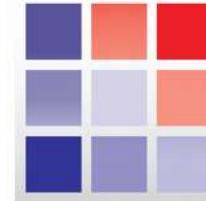
- PyPI
 - `pip install genepattern-notebook`
- DockerHub
 - `docker pull genepattern/genepattern-notebook`





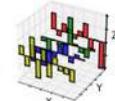
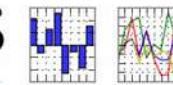
Jupyter Ecosystem

matplotlib



pandas

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



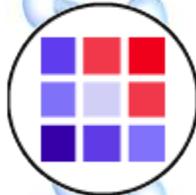
IP[y]:
IPython



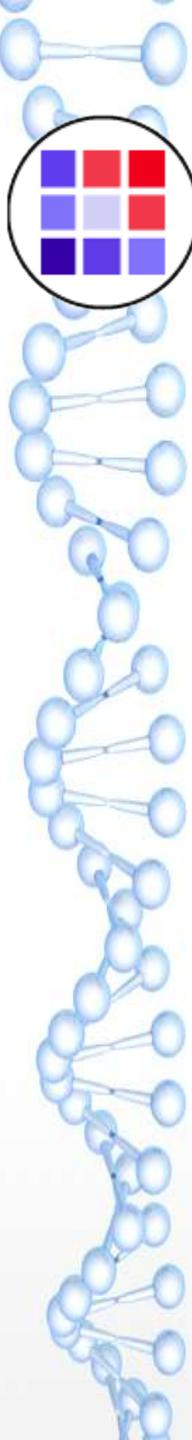

Anaconda







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](https://twitter.com/genepattern)

GenePattern GitHub
github.com/genepattern

GenePattern DockerHub
hub.docker.com/r/genepattern