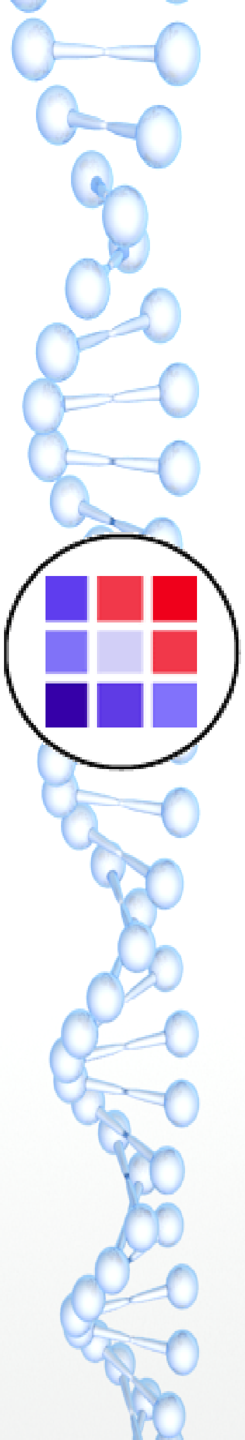
Thorin Tabor
JupyterCon 2017



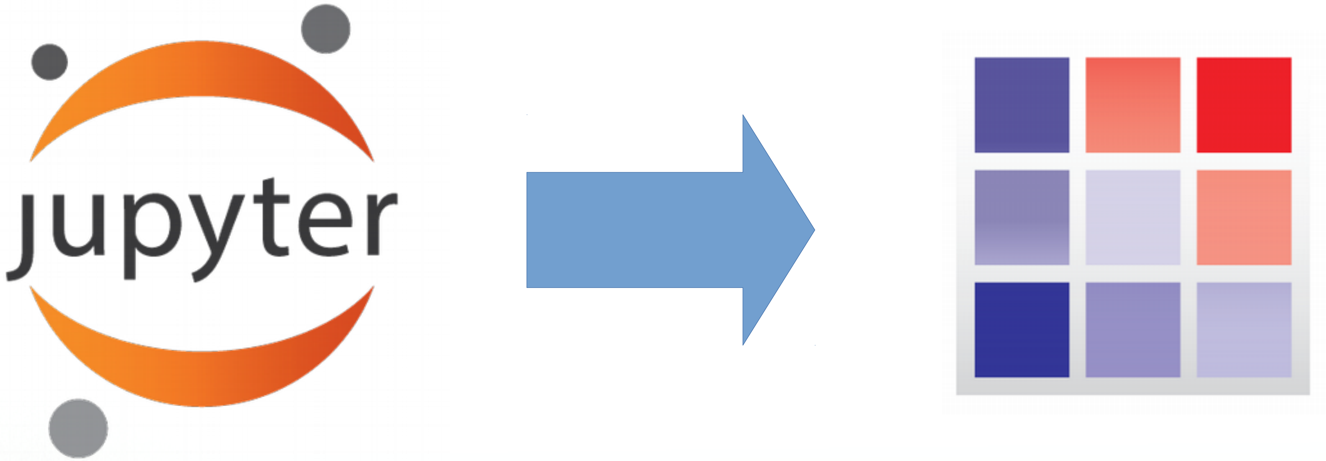
GenePattern Notebooks

Jupyter for Integrative Genomics

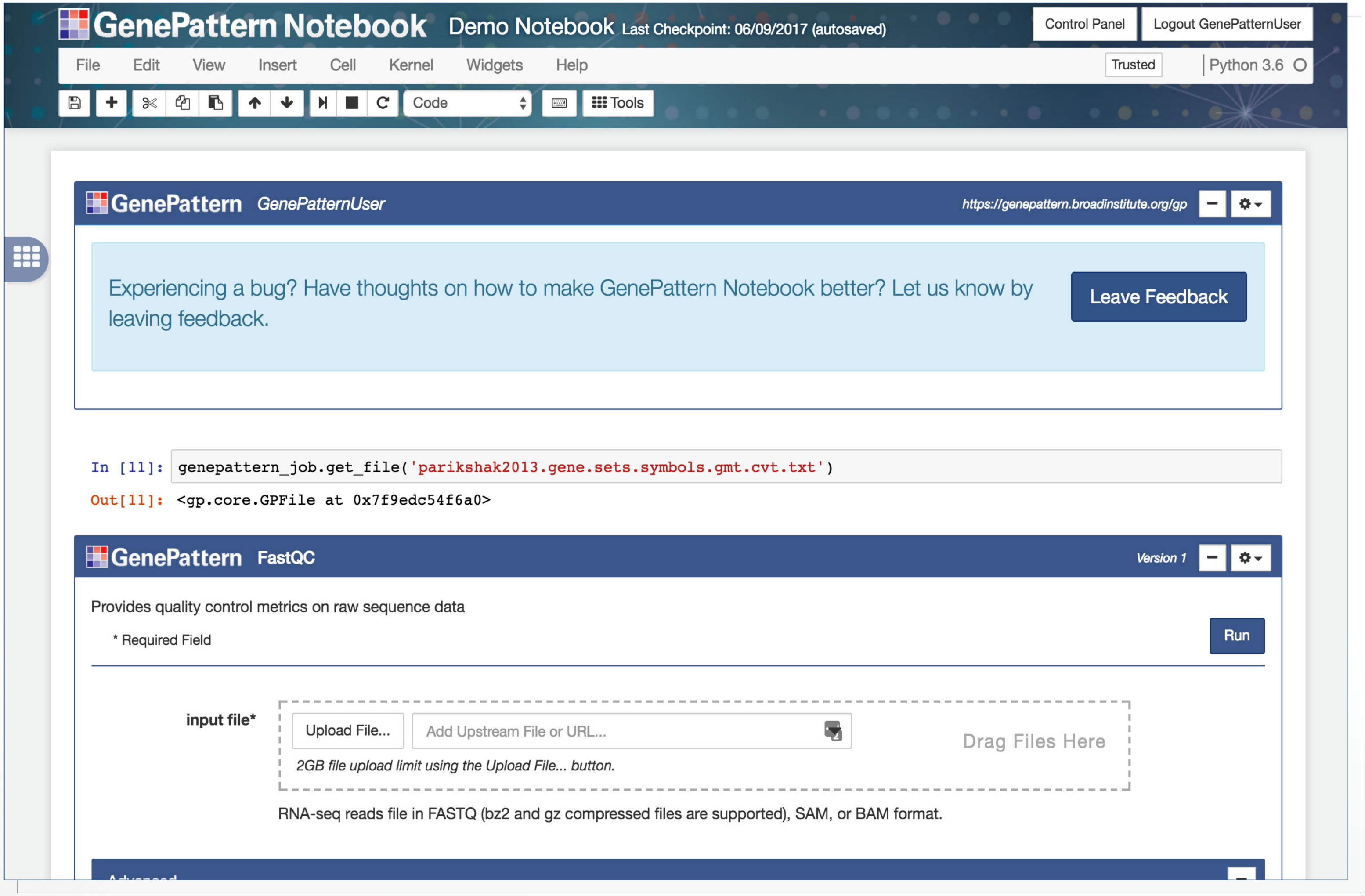
UC San Diego

BROAD INSTITUTE

From Jupyter to GenePattern



GenePattern Notebook



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

```
In [11]: genepattern_job.get_file('parikshak2013.gene.sets.symbols.gmt.cvt.txt')
Out[11]: <gp.core.GPFile at 0x7f9edc54f6a0>
```

Provides quality control metrics on raw sequence data

* Required Field

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

RNA-seq reads file in FASTQ (bz2 and gz compressed files are supported), SAM, or BAM format.

Two Open Source Projects

What is GenePattern?

Module Repository

NMF | Cuffdiff

GISTIC | GSEA

Tophat | MuTect

Hundreds of analyses and visualizations

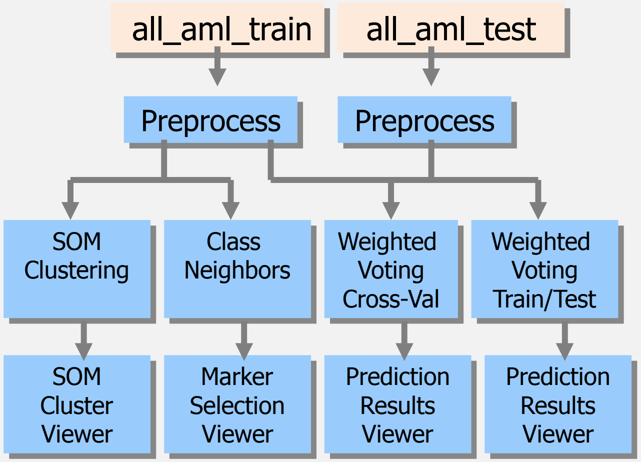
Module Integrator

Easy addition of new tools

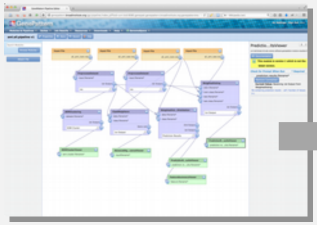
Community Repository

User-contributed modules

Pipeline Environment



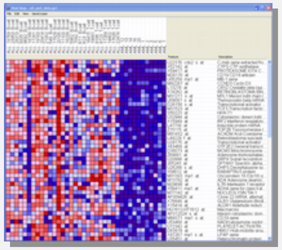
Golub and Slonim et. al 1999



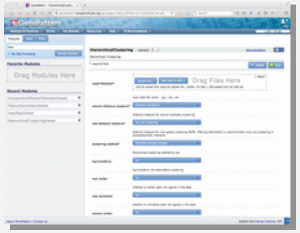
Golub.Slonim.1999.Nature.all.aml.pipeline.zip

Support for *in silico* reproducible research

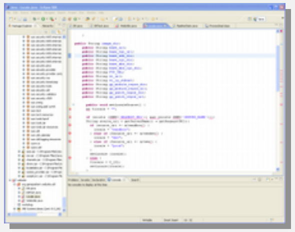
Clients



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
- ~50,000 registered users
- Public server runs ~4,000 analyses per week
- Community-contributed methods
  - CRISPR analysis
  - Bisulfite sequencing
  - Flow cytometry
  - RNAi screens

GenePattern

A Platform for Reproducible Bioinformatics

Use GenePattern

GenePattern Basics

Download

Features

Analysis Pipelines

Reproducible Research

Programming Environment

GP updates

Related Publications

genepattern.org

GenePattern

Drag Modules Here

Getting Started

Analyzing genomic data in GenePattern

genepattern.broadinstitute.org

GPARC

Upload a Module

Resources

Contact Us

Login

GParc

A repository and community where users can share and discuss their own GenePattern modules.

Learn More »

Upload a module »

Browse Modules

Search

Show All

Show Most Recent Uploads

Acgh2Tab v4

Converts acgh files to a tab-delimited format usable by Genomica.

acgh

Data Format Conversion

Filter By Available Tags

View All Tags

Annotation

Bisulfite Sequencing

Clustering

ComparativeMarkerSelection

CRISPR

gparc.org

Analysis 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

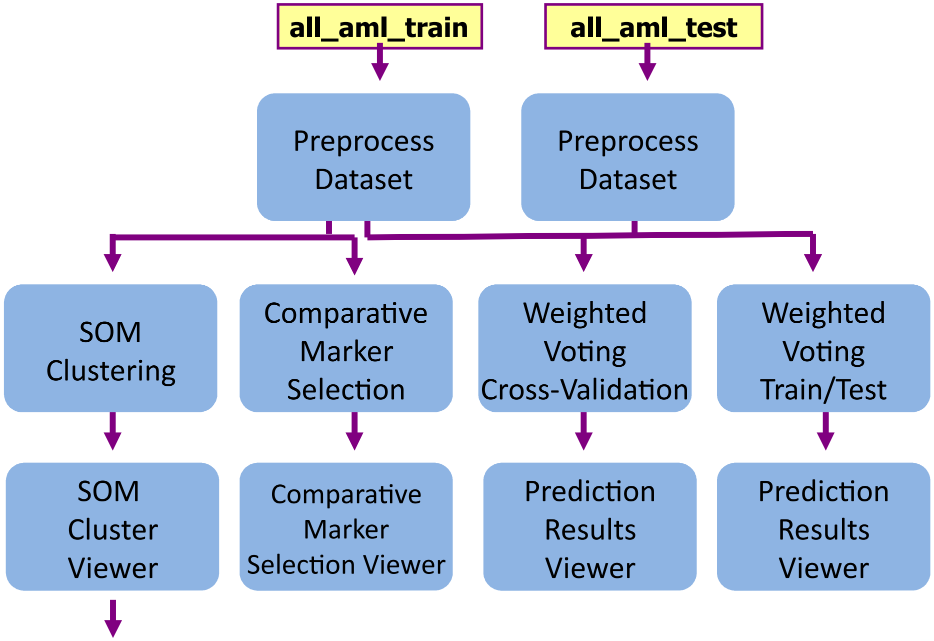
Documentation

HierarchicalClustering.pdf

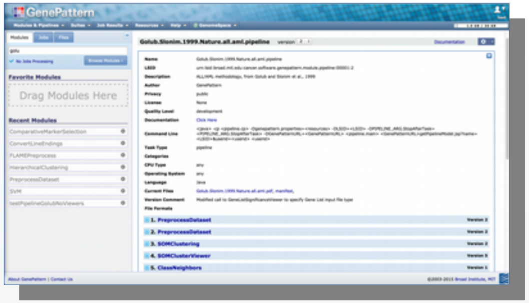
Parameter descriptions

-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



all_aml_train
all_aml_test
Preprocess Dataset
Preprocess Dataset
SOM Clustering
Comparative Marker Selection
Weighted Voting Cross-Validation
Weighted Voting Train/Test
SOM Cluster Viewer
Comparative Marker Selection Viewer
Prediction Results Viewer
Prediction Results Viewer



Golub.Slonim.1999.Nature.all.aml.pipeline version 2

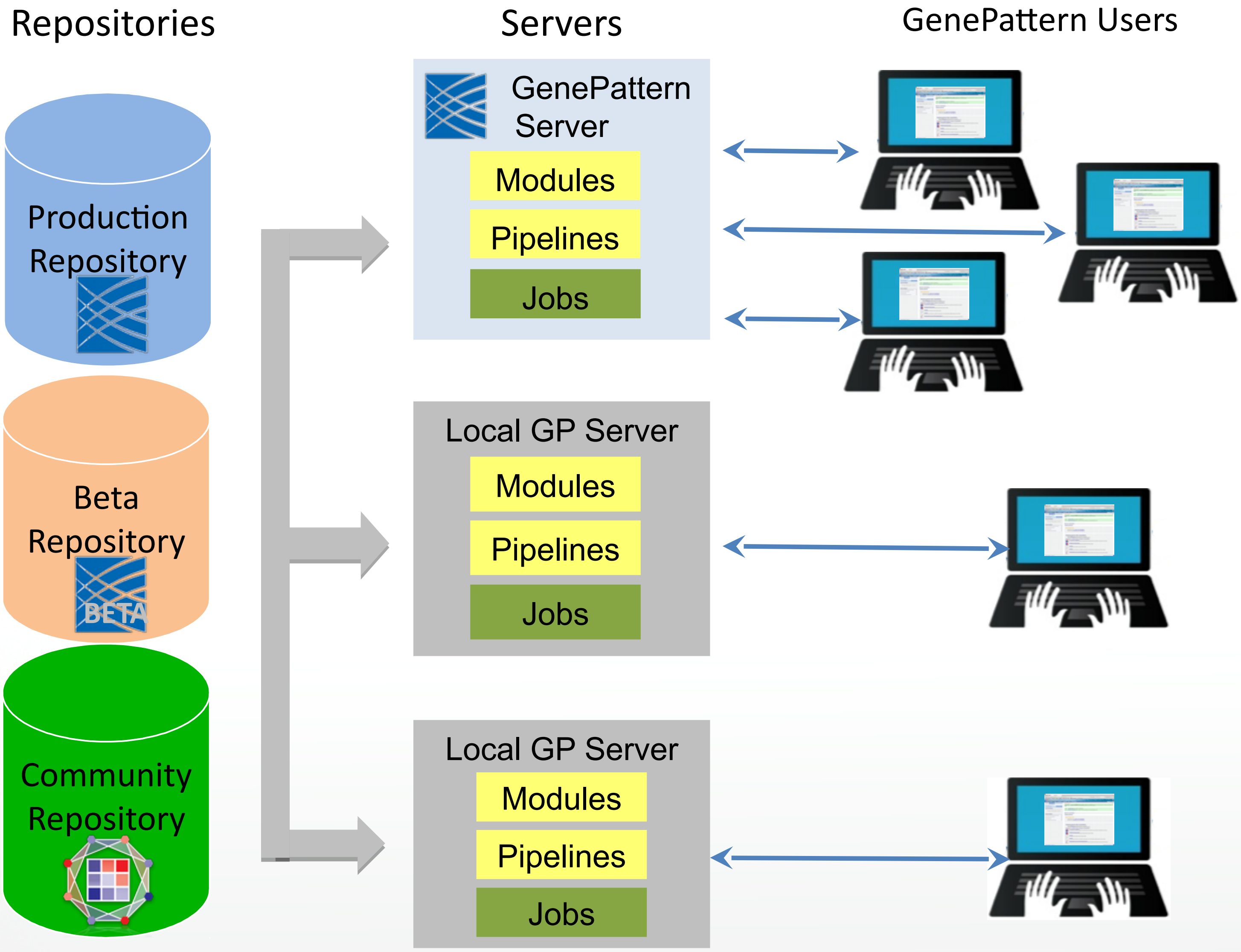
Name Golub.Slonim.1999.Nature.all.aml.pipeline
LSID urn:lsid:broad.mit.edu:cancer.software.genepattern.module.pipeline:00001:2
Description ALL/AML methodology, from Golub and Slonim et al., 1999
Author GenePattern
Privacy public
License None
Quality Level development
Documentation Click Here
Command Line [illegible]
Task Type pipeline
Categories
CPU Type any
Operating System any
Language Java
Current Files Golub.Slonim.1999.Nature.all.aml.pdf, manifest,
Version Comment Modified call to GeneListSignificanceViewer to specify Gene List input file type
File Formats

1. PreprocessDataset Version 2
2. PreprocessDataset Version 2
3. SOMClustering Version 2
4. SOMClusterViewer Version 5
5. ClassNeighbors Version 1

©2003-2015 Broad Institute, MIT

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

Web Server Architecture

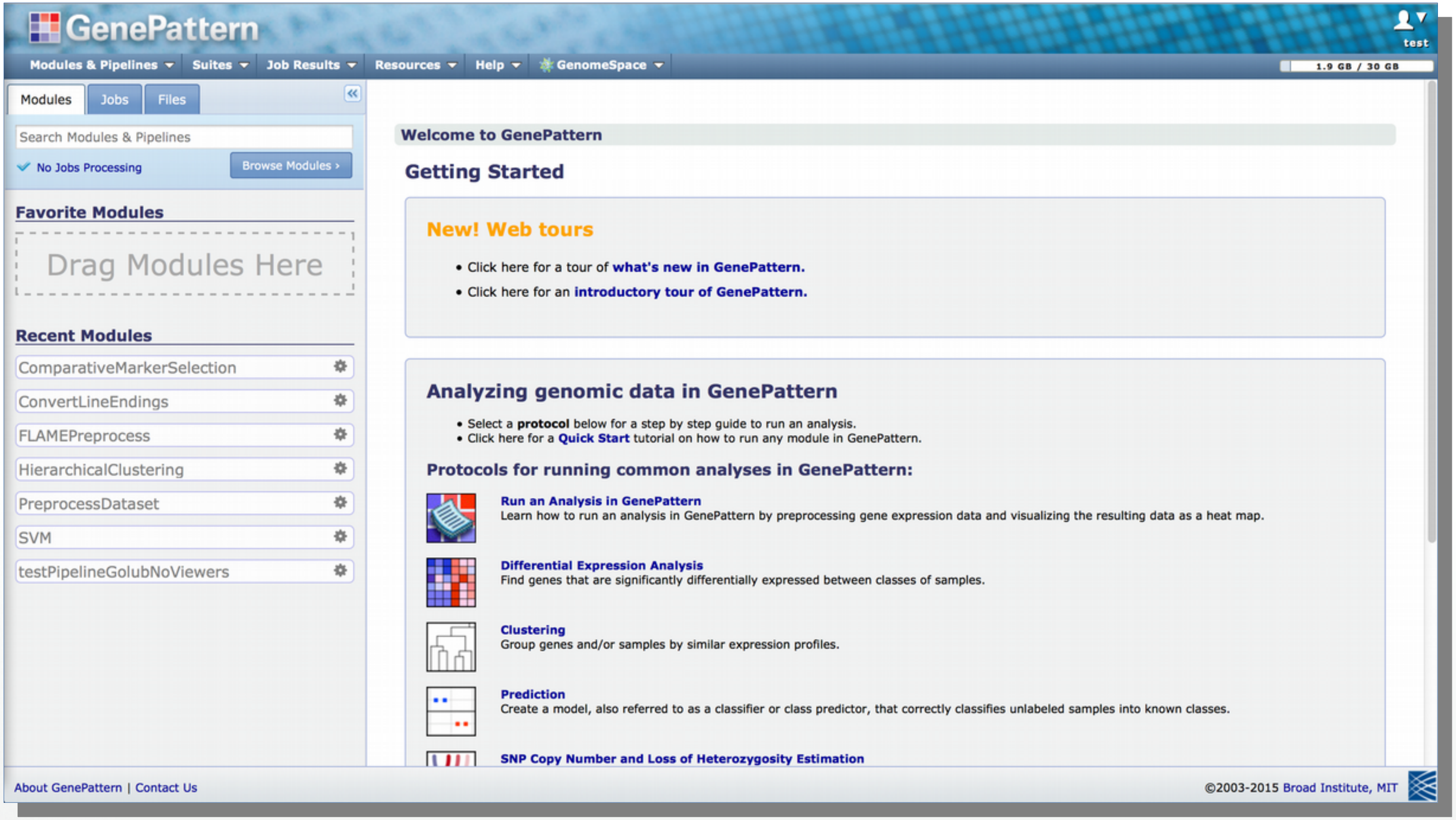


Programmatic APIs

- 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



Welcome to GenePattern

Getting Started

New! Web tours

- Click here for a tour of **what's new in GenePattern.**
- Click here for an **introductory tour of GenePattern.**

Analyzing genomic data in GenePattern

- Select a **protocol** below for a step by step guide to run an analysis.
- Click here for a **Quick Start** tutorial on how to run any module in GenePattern.

Protocols for running common analyses in GenePattern:

Run an Analysis in GenePattern
Learn how to run an analysis in GenePattern by preprocessing gene expression data and visualizing the resulting data as a heat map.

Differential Expression Analysis
Find genes that are significantly differentially expressed between classes of samples.

Clustering
Group genes and/or samples by similar expression profiles.

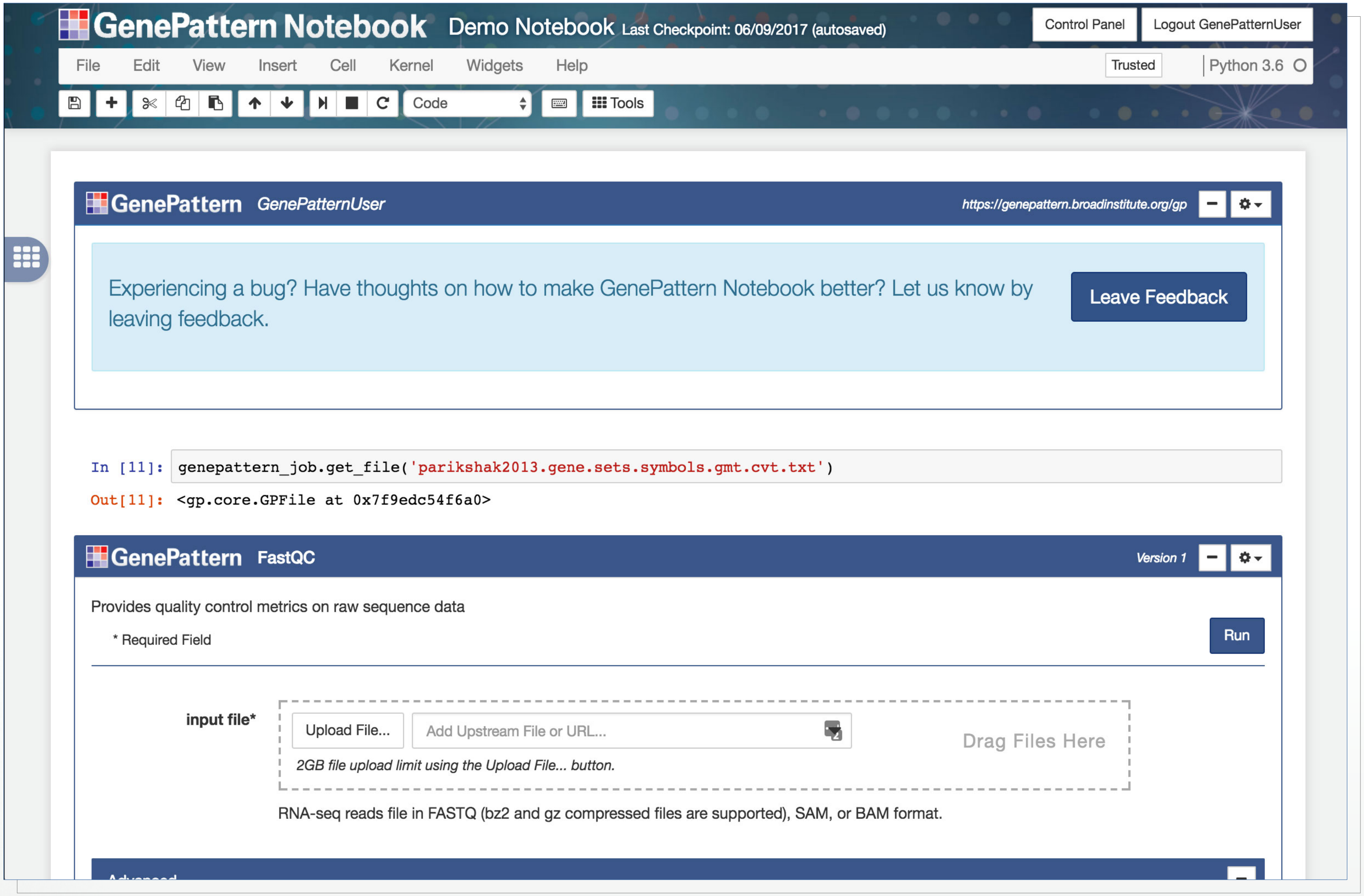
Prediction
Create a model, also referred to as a classifier or class predictor, that correctly classifies unlabeled samples into known classes.

SNP Copy Number and Loss of Heterozygosity Estimation

About GenePattern | Contact Us

©2003-2015 Broad Institute, MIT

GenePattern Notebook
Jupyter Extensions



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

```
In [11]: genepattern_job.get_file('parikshak2013.gene.sets.symbols.gmt.cvt.txt')
Out[11]: <gp.core.GPFile at 0x7f9edc54f6a0>
```

Provides quality control metrics on raw sequence data

* Required Field

input file*

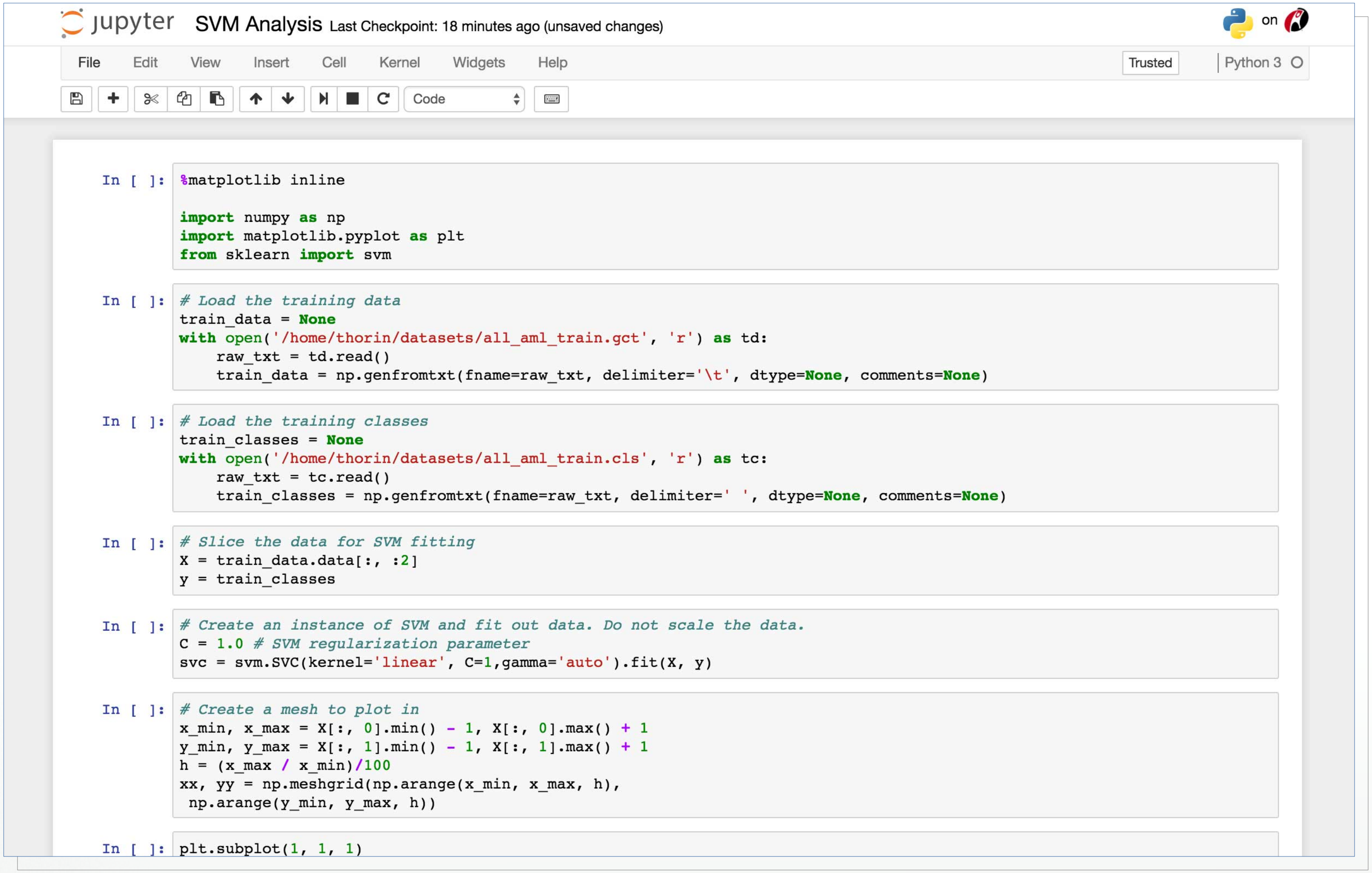
2GB file upload limit using the Upload File... button.

RNA-seq reads file in FASTQ (bz2 and gz compressed files are supported), SAM, or BAM format.

Complete Research Narrative

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

SVM Example #1



```
In [ ]: %matplotlib inline

        import numpy as np
        import matplotlib.pyplot as plt
        from sklearn import svm
```

```
In [ ]: # Load the training data
        train_data = None
        with open('/home/thorin/datasets/all_aml_train.gct', 'r') as td:
            raw_txt = td.read()
            train_data = np.genfromtxt(fname=raw_txt, delimiter='\t', dtype=None, comments=None)
```

```
In [ ]: # Load the training classes
        train_classes = None
        with open('/home/thorin/datasets/all_aml_train.cls', 'r') as tc:
            raw_txt = tc.read()
            train_classes = np.genfromtxt(fname=raw_txt, delimiter=' ', dtype=None, comments=None)
```

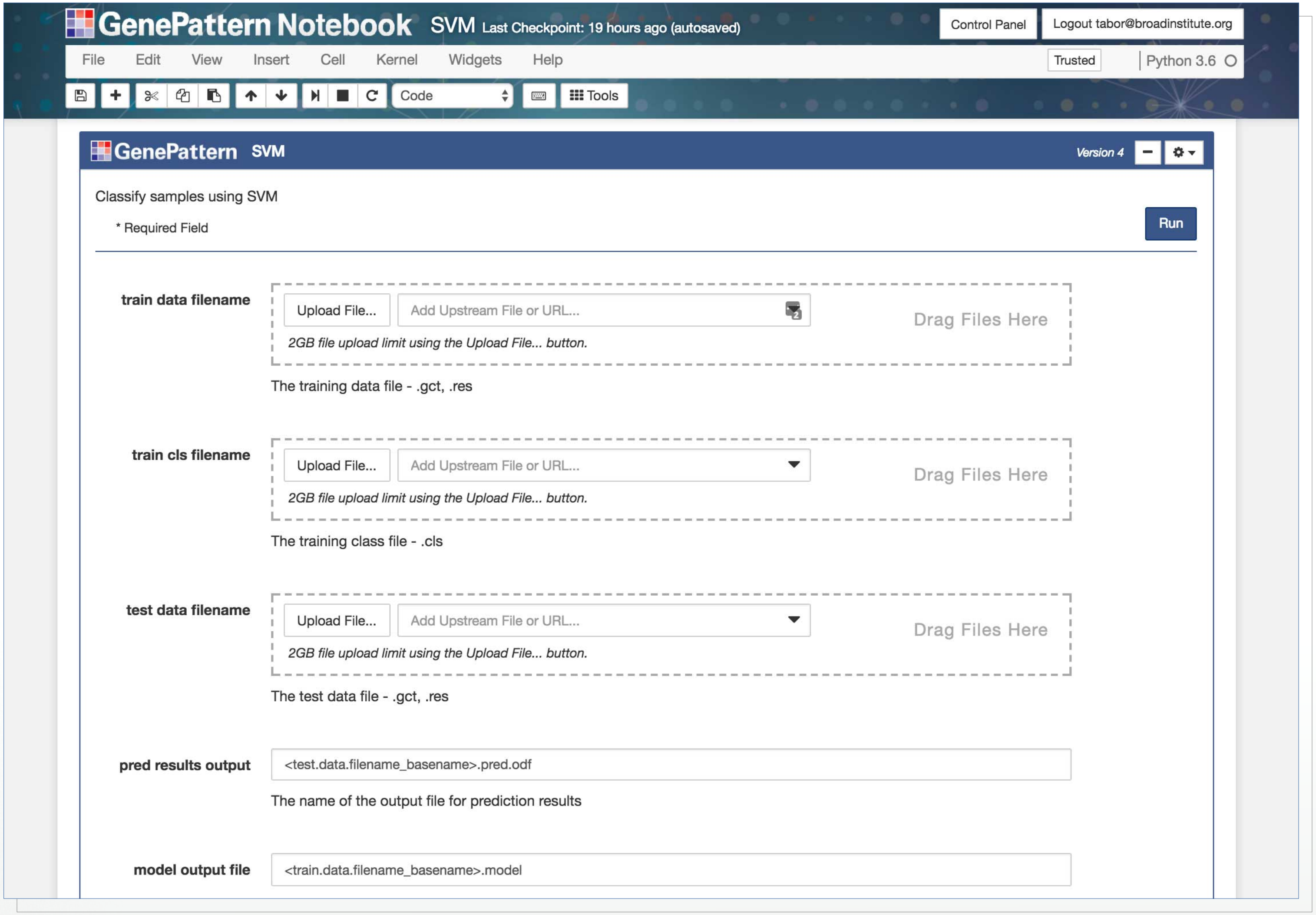
```
In [ ]: # Slice the data for SVM fitting
        X = train_data.data[:, :2]
        y = train_classes
```

```
In [ ]: # Create an instance of SVM and fit out data. Do not scale the data.
        C = 1.0 # SVM regularization parameter
        svc = svm.SVC(kernel='linear', C=1,gamma='auto').fit(X, y)
```

```
In [ ]: # Create a mesh to plot in
        x_min, x_max = X[:, 0].min() - 1, X[:, 0].max() + 1
        y_min, y_max = X[:, 1].min() - 1, X[:, 1].max() + 1
        h = (x_max / x_min)/100
        xx, yy = np.meshgrid(np.arange(x_min, x_max, h),
         np.arange(y_min, y_max, h))
```

```
In [ ]: plt.subplot(1, 1, 1)
```

SVM Example #2



Classify samples using SVM

* Required Field

train data filename

Upload File... | Add Upstream File or URL... | Drag Files Here

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

The training data file - .gct, .res

train cls filename

Upload File... | Add Upstream File or URL... | Drag Files Here

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

The training class file - .cls

test data filename

Upload File... | Add Upstream File or URL... | Drag Files Here

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

The test data file - .gct, .res

pred results output

<test.data.filename_basename>.pred.odf

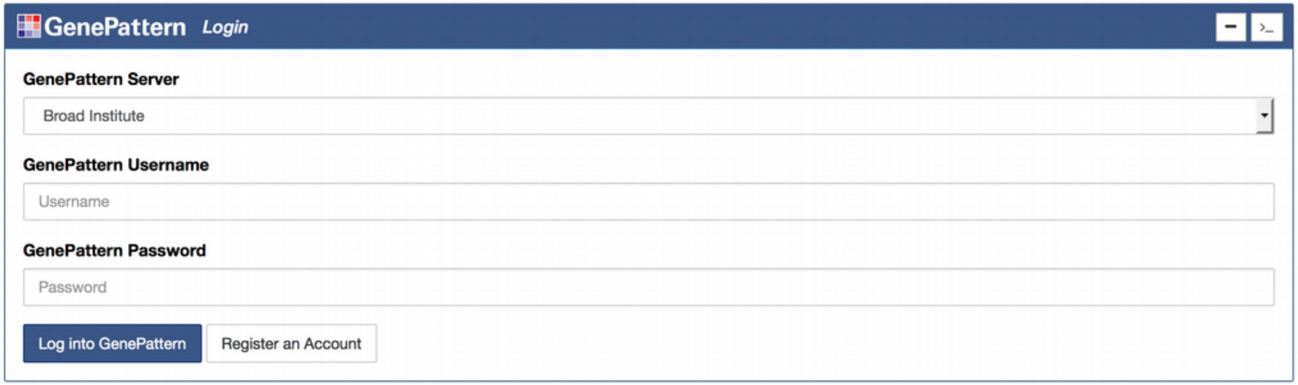
The name of the output file for prediction results

model output file

<train.data.filename_basename>.model

GenePattern Cells

Auth Cell


GenePattern Login

GenePattern Server

Broad Institute

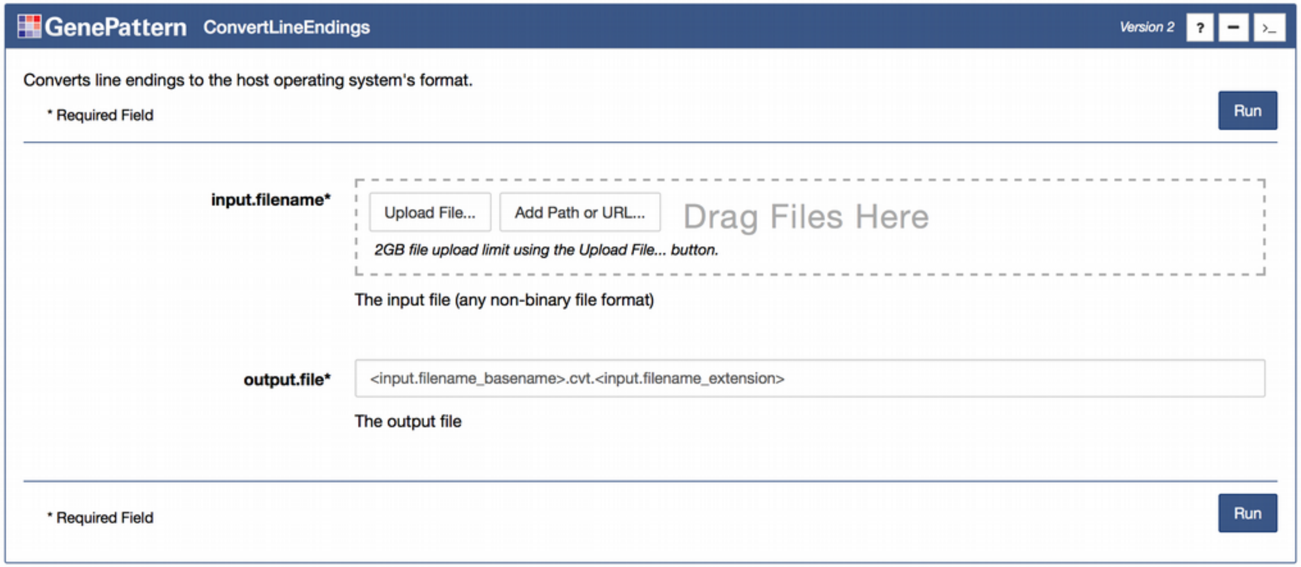
GenePattern Username

Username

GenePattern Password

Password

Analysis Cell


GenePattern ConvertLineEndings

Version 2

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

* Required Field

input.filename*

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

The input file (any non-binary file format)

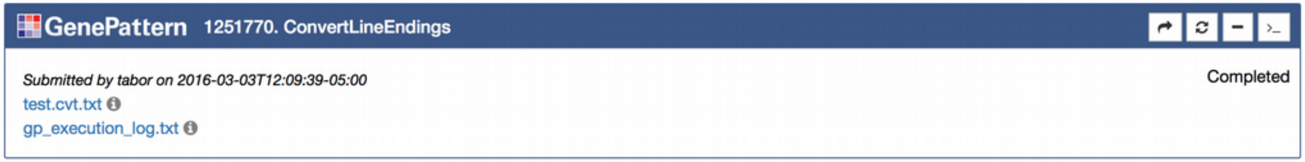
output.file*

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

The output file

* Required Field

Job Cell


GenePattern 1251770. ConvertLineEndings

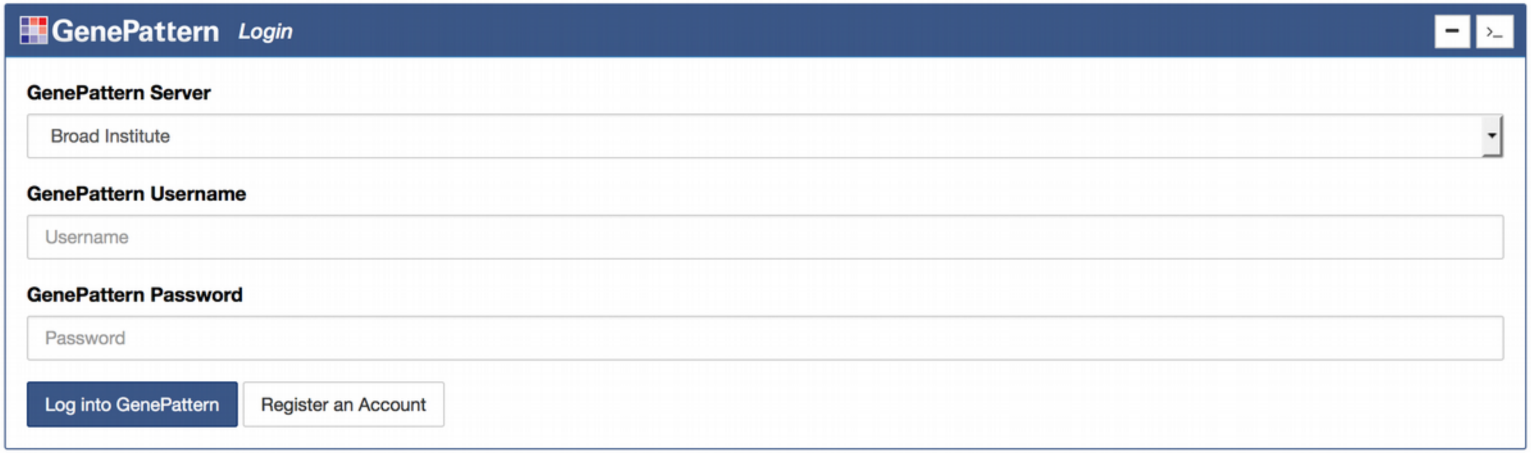
Submitted by tabor on 2016-03-03T12:09:39-05:00

Completed

test.cvt.txt

gp_execution_log.txt

Authentication Cells



GenePattern Server

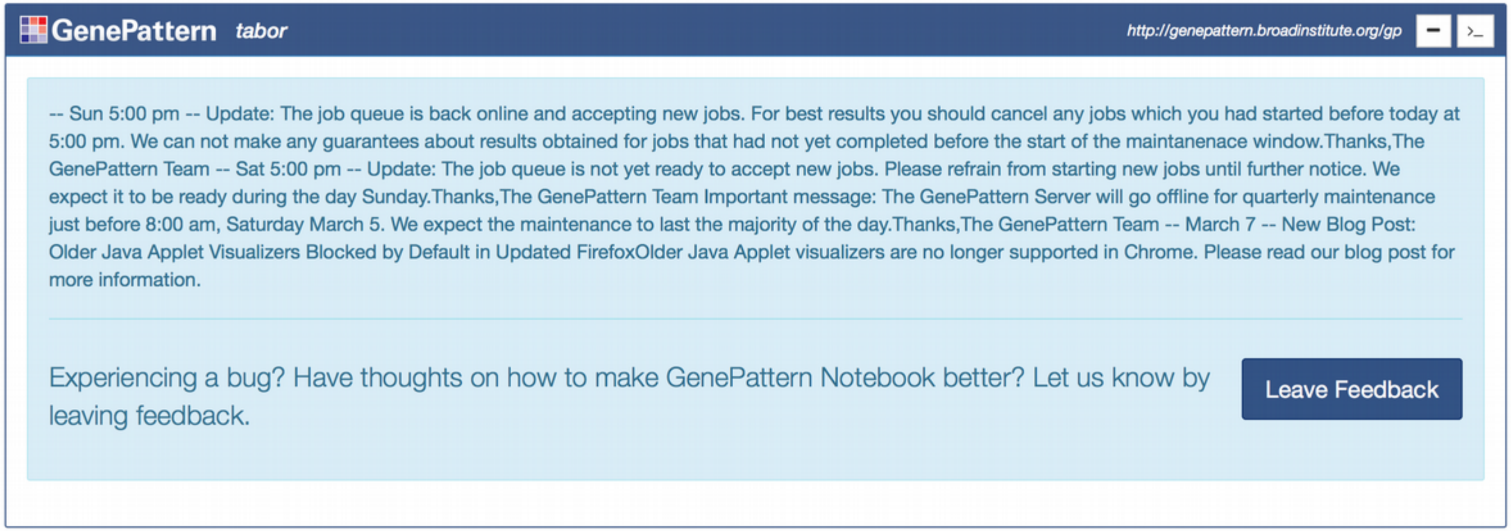
Broad Institute

GenePattern Username

Username

GenePattern Password

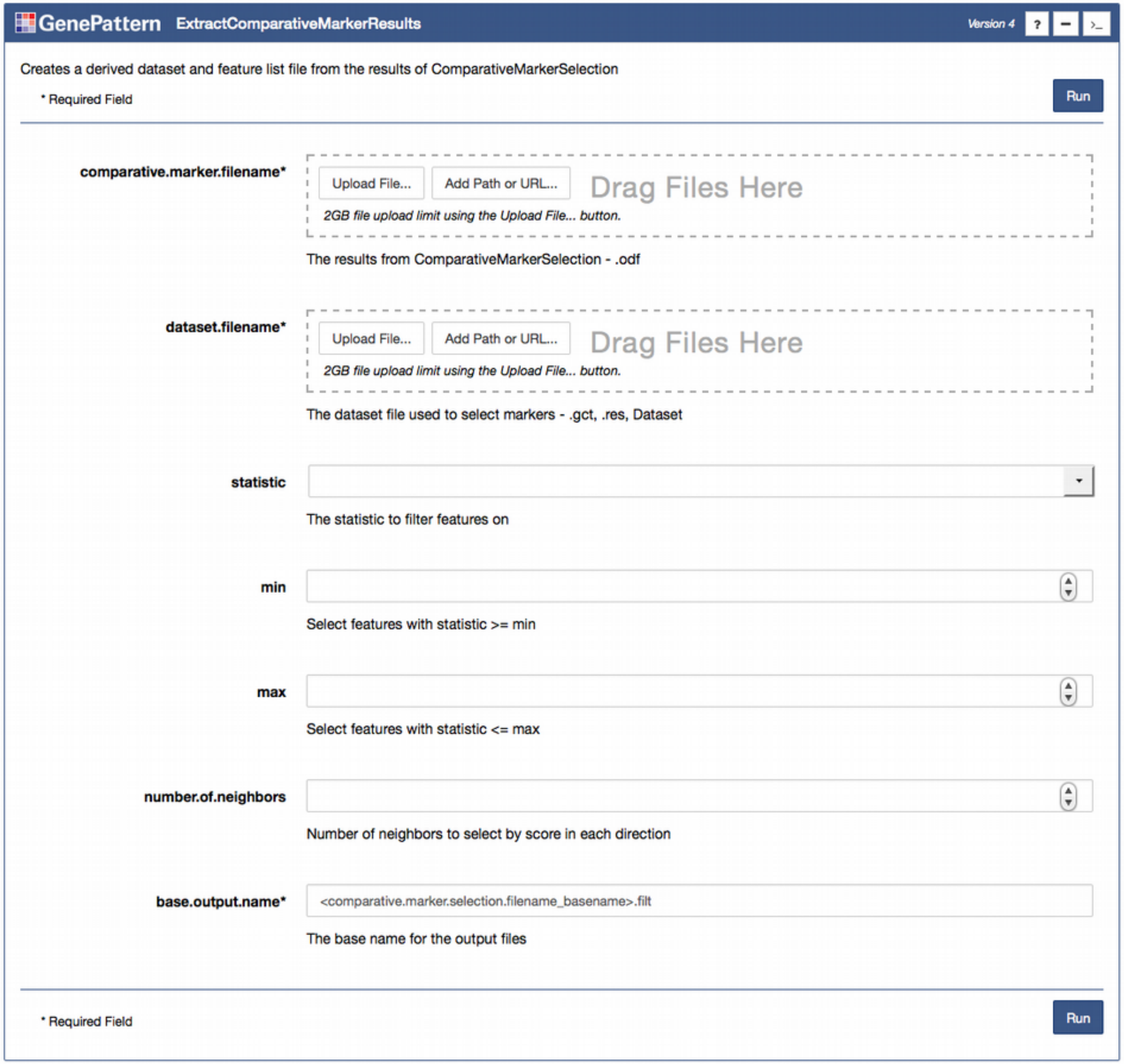
Password



-- 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 maintanenace 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.

Analysis Cells



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

* Required Field

comparative.marker.filename*

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

The results from ComparativeMarkerSelection - .odf

dataset.filename*

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 >= min

max

Select features with statistic <= max

number.of.neighbors

Number of neighbors to select by score in each direction

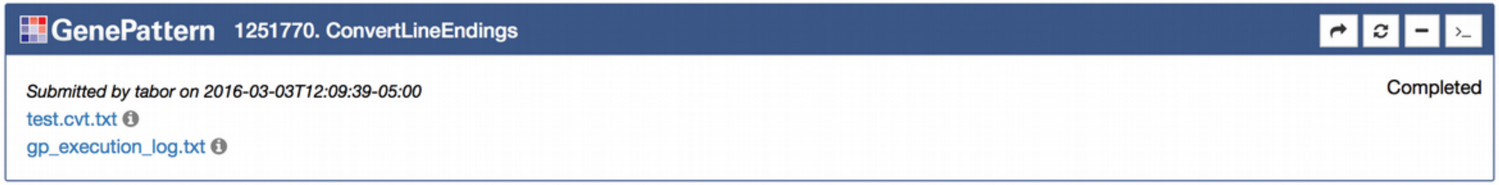
base.output.name*

<comparative.marker.selection.filename_basename>.filt

The base name for the output files

* Required Field

Job Cells



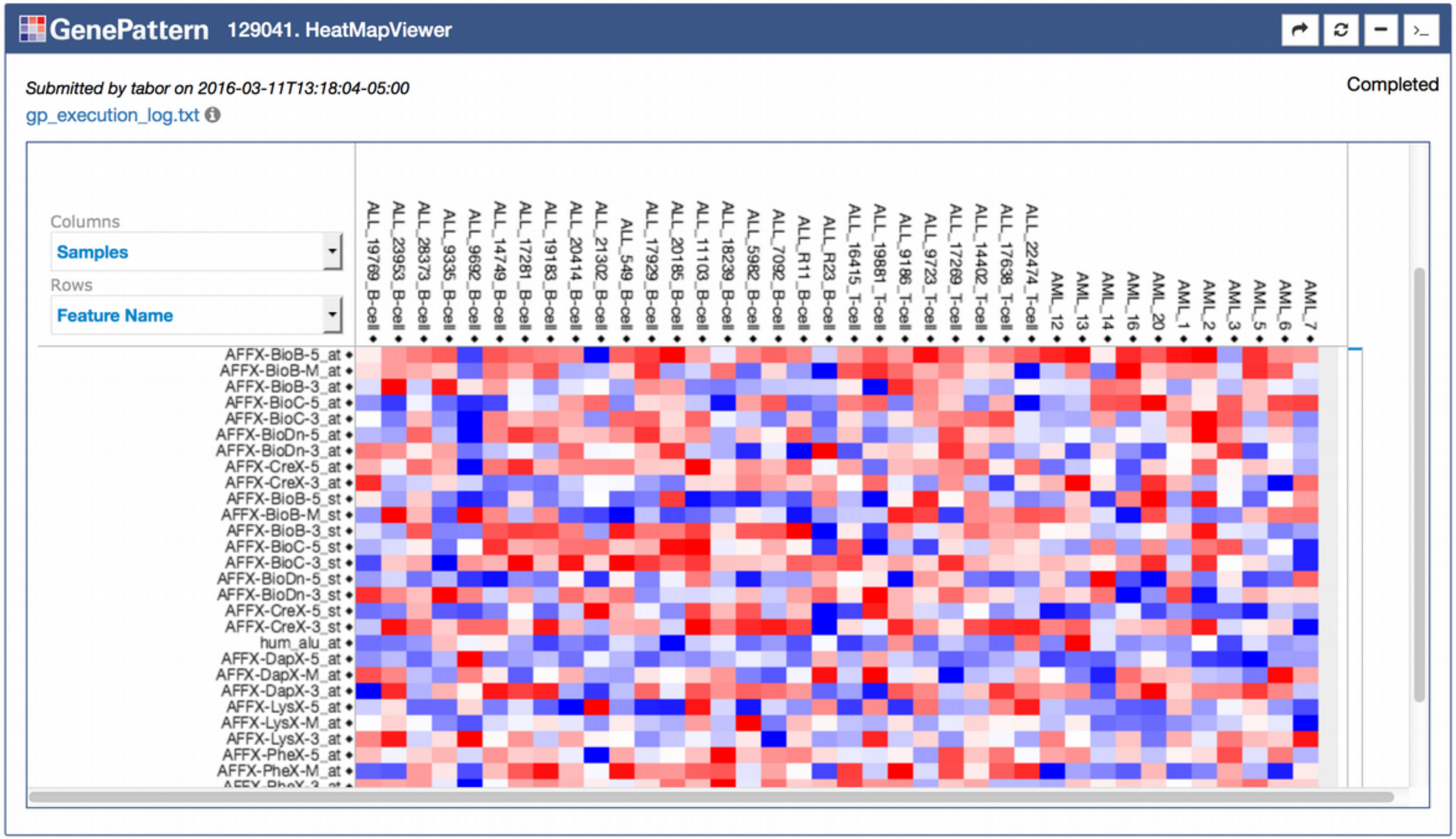
1251770. ConvertLineEndings

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

Completed

test.cvt.txt

gp_execution_log.txt



129041. HeatMapViewer

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

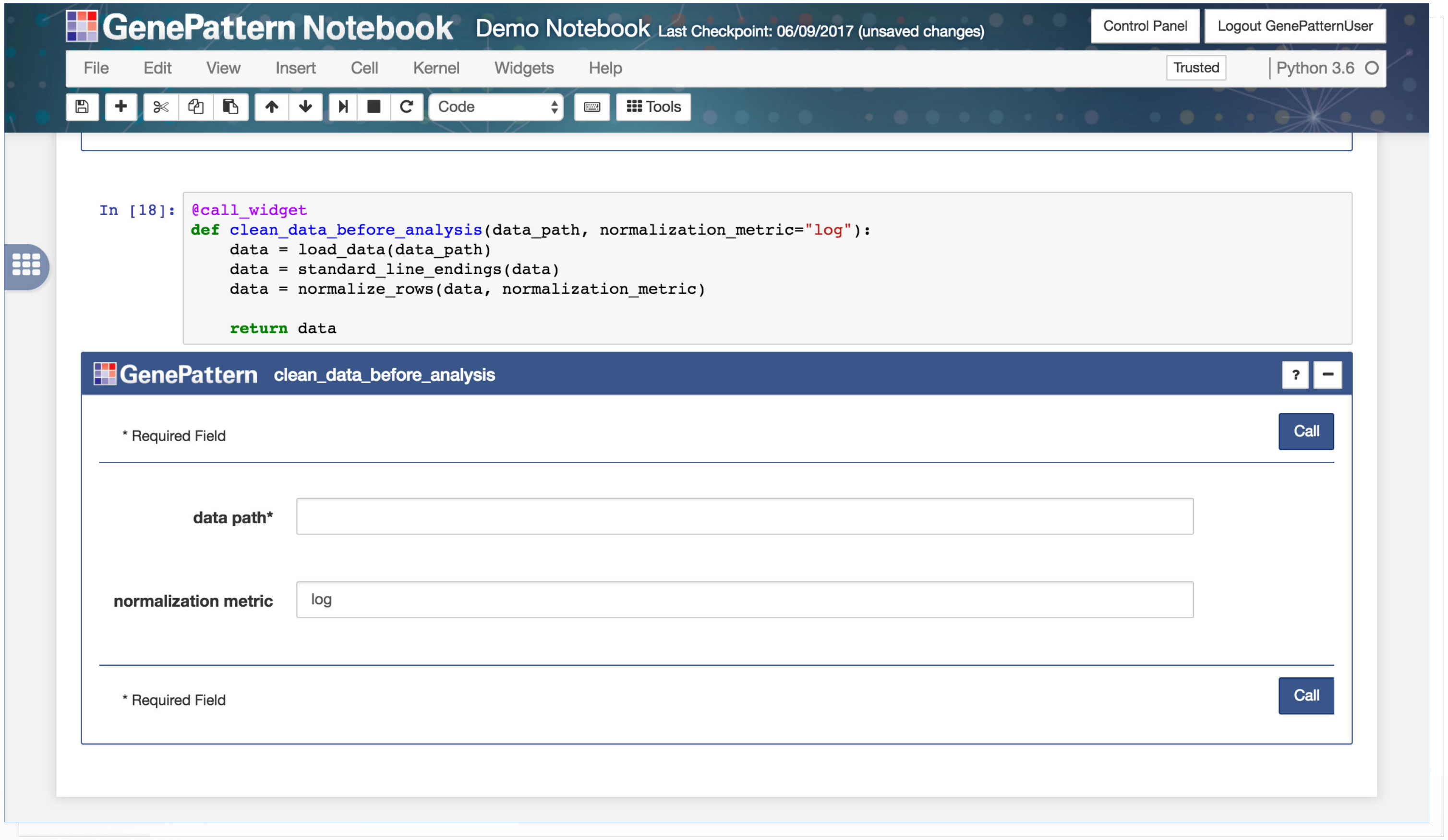
Completed

gp_execution_log.txt

Columns: Samples

Rows: Feature Name

Python Function GUI



```
In [18]: @call_widget
         def clean_data_before_analysis(data_path, normalization_metric="log"):
             data = load_data(data_path)
             data = standard_line_endings(data)
             data = normalize_rows(data, normalization_metric)

             return data
```

GenePattern clean_data_before_analysis

* Required Field

Call

data path*:

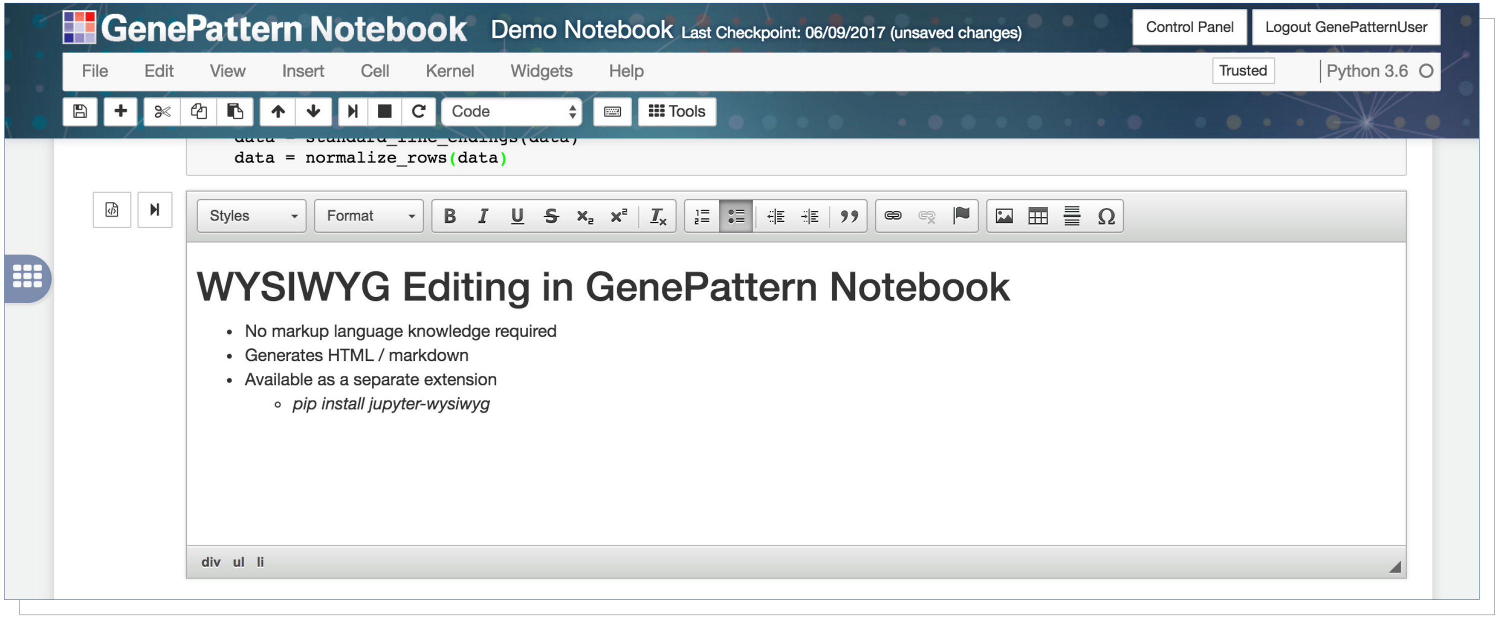
normalization metric: log

* Required Field

Call

- Turn any Python function into an interactive user interface

Rich Text Markdown Editor



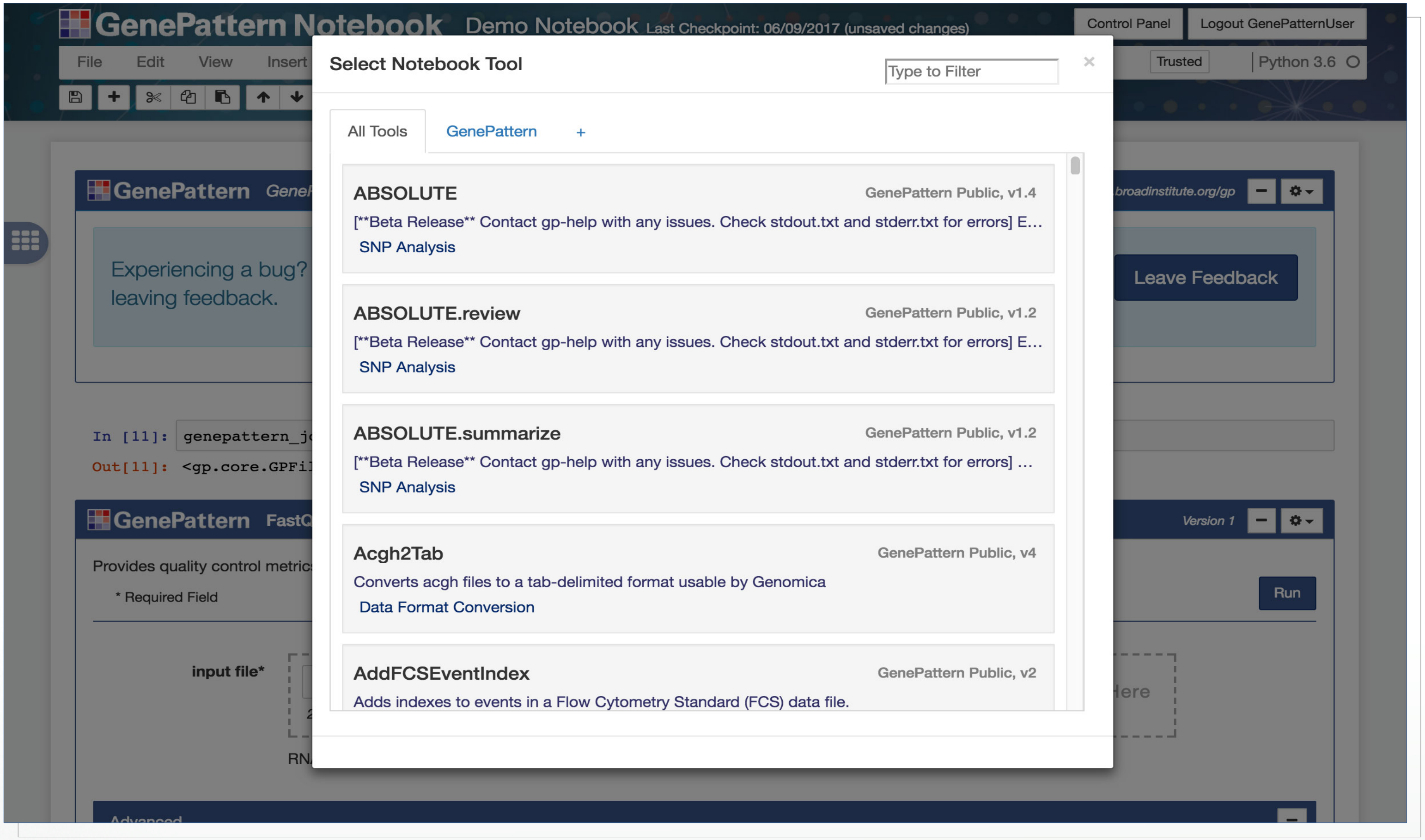
```
data = standard_line_endings(data)
data = normalize_rows(data)
```

WYSIWYG Editing in GenePattern Notebook

- No markup language knowledge required
- Generates HTML / markdown
- Available as a separate extension
  - *pip install jupyter-wysiwyg*

- No markup knowledge required
- Generates HTML / markdown
- Available as a separate extension

Notebook Tool Manager



```
In [11]: genepattern_j
Out[11]: <gp.core.GPFil
```

Provides quality control metric

* Required Field

Select Notebook Tool

Type to Filter

All Tools | GenePattern | +

ABSOLUTE — GenePattern Public, v1.4
[**Beta Release** Contact gp-help with any issues. Check stdout.txt and stderr.txt for errors] E…
SNP Analysis

ABSOLUTE.review — GenePattern Public, v1.2
[**Beta Release** Contact gp-help with any issues. Check stdout.txt and stderr.txt for errors] E…
SNP Analysis

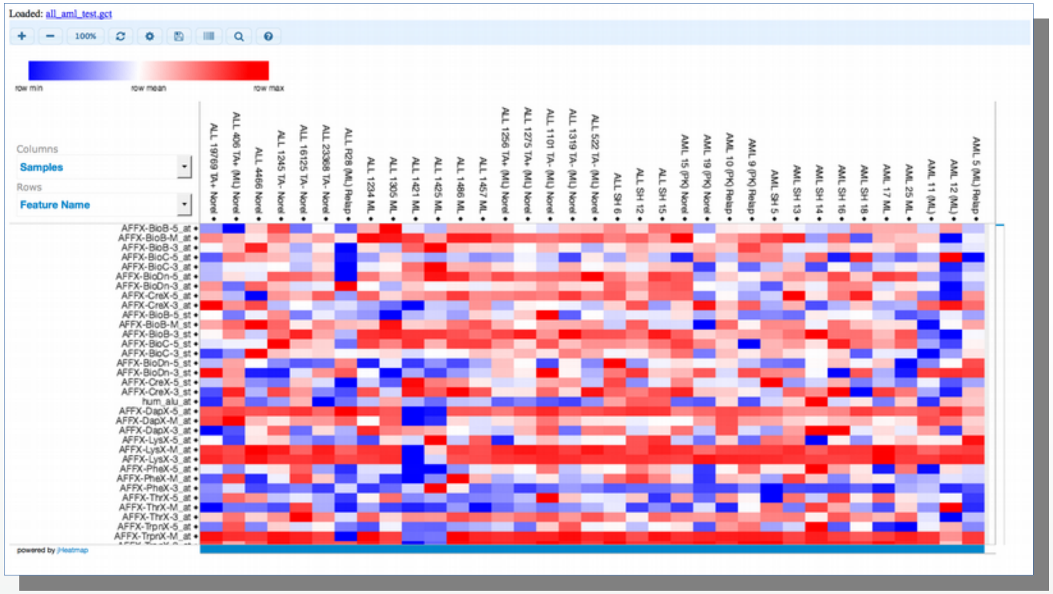
ABSOLUTE.summarize — GenePattern Public, v1.2
[**Beta Release** Contact gp-help with any issues. Check stdout.txt and stderr.txt for errors] …
SNP Analysis

Acgh2Tab — GenePattern Public, v4
Converts acgh files to a tab-delimited format usable by Genomica
Data Format Conversion

AddFCSEventIndex — GenePattern Public, v2
Adds indexes to events in a Flow Cytometry Standard (FCS) data file.

Behind the Scenes

- Interactive widgets use the Jupyter widget framework (ipywidgets, traitlets)
- Can use Python variables as input
- Not limited by GenePattern analyses



Loaded: all_aml_test.gct

Columns: Samples

Rows: Feature Name

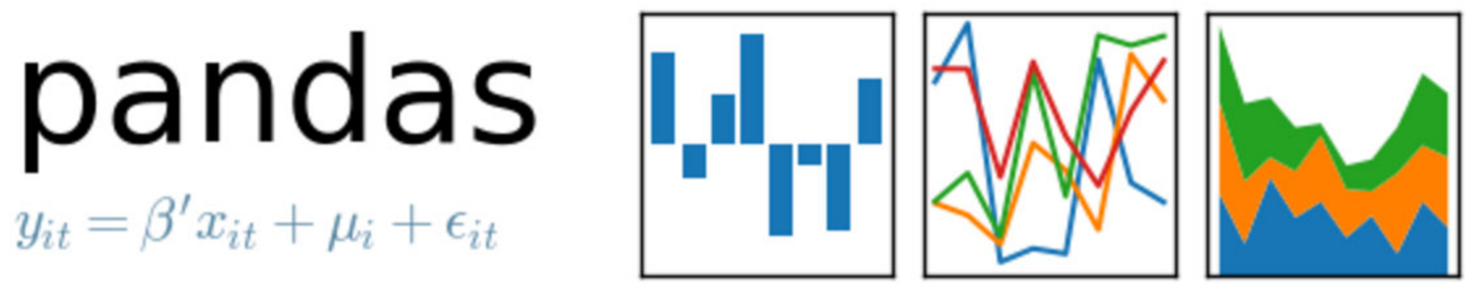
GenePattern Python Library

- Complete programmatic access
- Automatic integration with GenePattern cell data

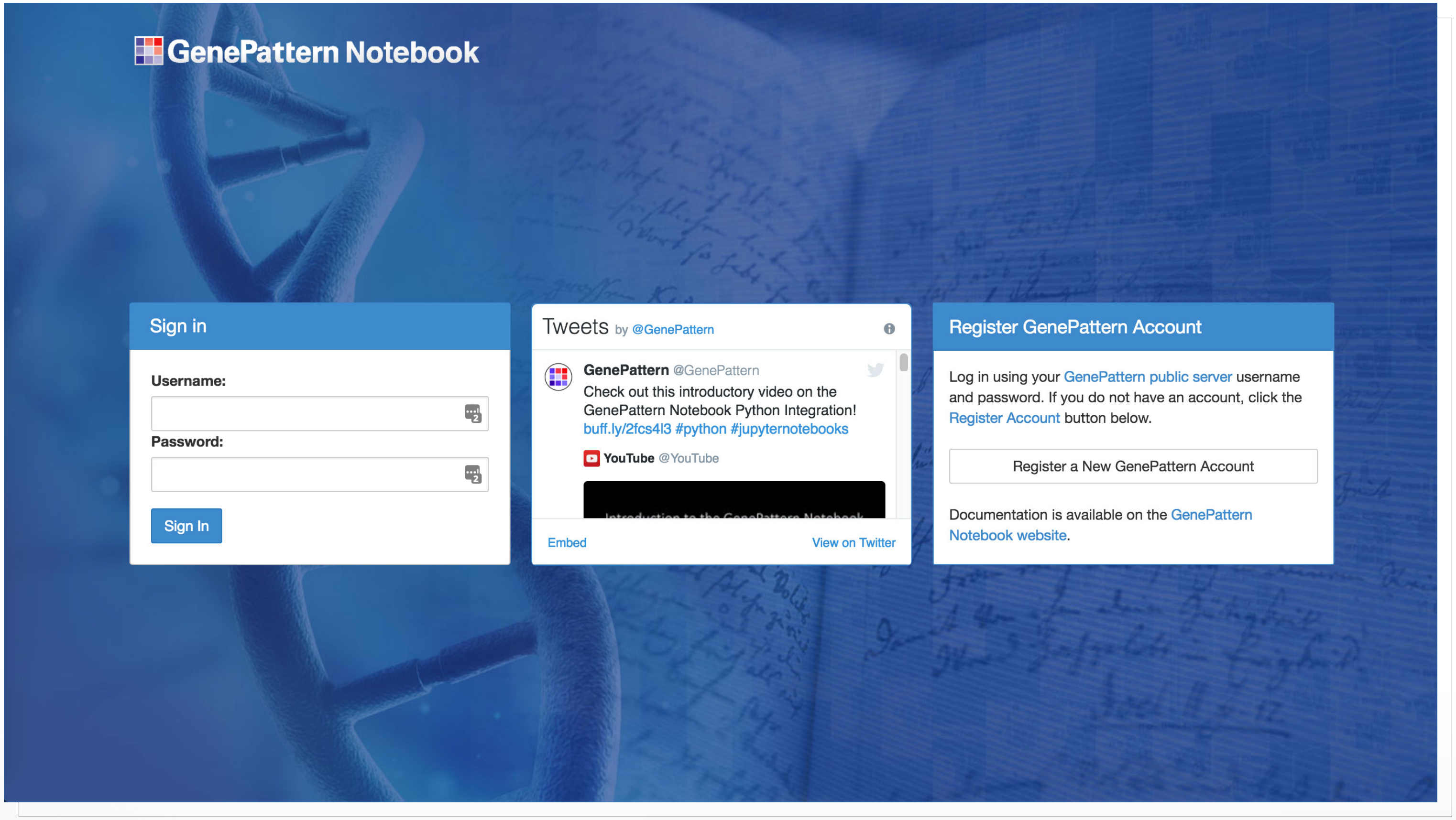
```
import gp

# 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 Data Tools

- Easily import common bioinformatic data formats as pandas DataFrames
- Work with GenePattern files using popular Python libraries



GenePattern
Notebook Repository



GenePattern Notebook

Sign in

Username:

Password:

Tweets by @GenePattern

GenePattern @GenePattern

Check out this introductory video on the GenePattern Notebook Python Integration! buff.ly/2fcs4l3 #python #jupyternotebooks

YouTube @YouTube

Introduction to the GenePattern Notebook

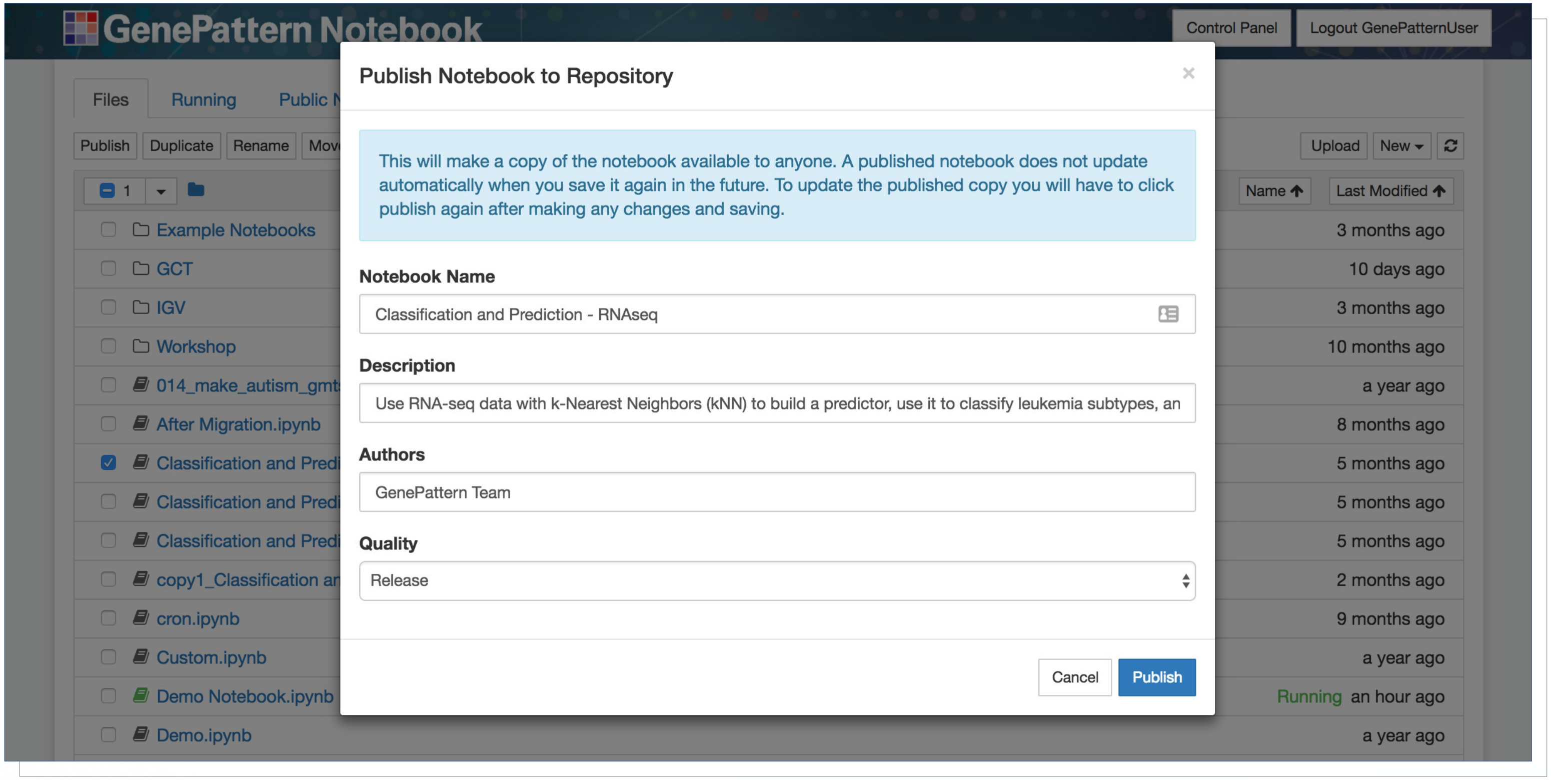
Register GenePattern Account

Log in using your GenePattern public server username and password. If you do not have an account, click the Register Account button below.

Documentation is available on the GenePattern Notebook website.

http://genepattern-notebook.org

Publish & Share Notebooks



☐	Example Notebooks	3 months ago
☐	GCT	10 days ago
☐	IGV	3 months ago
☐	Workshop	10 months ago
☐	014_make_autism_gmt	a year ago
☐	After Migration.ipynb	8 months ago
☑	Classification and Predi	5 months ago
☐	Classification and Predi	5 months ago
☐	Classification and Predi	5 months ago
☐	copy1_Classification an	2 months ago
☐	cron.ipynb	9 months ago
☐	Custom.ipynb	a year ago
☐	Demo Notebook.ipynb	Running an hour ago
☐	Demo.ipynb	a year ago

Publish Notebook to Repository ×

This will make a copy of the notebook available to anyone. A published notebook does not update automatically when you save it again in the future. To update the published copy you will have to click publish again after making any changes and saving.

Notebook Name
Classification and Prediction - RNAseq

Description
Use RNA-seq data with k-Nearest Neighbors (kNN) to build a predictor, use it to classify leukemia subtypes, an

Authors
GenePattern Team

Quality
Release

Cancel | Publish

- Publish notebooks to the GenePattern Notebook Repository.
- Browse available notebooks.

Installing the Extension

- PyPI

  -
    ```
    pip install genepattern-notebook
    ```
- Anaconda Cloud

  -
    ```
    conda install -c genepattern genepattern-notebook
    ```
- DockerHub

  -
    ```
    docker pull genepattern/genepattern-notebook
    ```

Jupyter Ecosystem

Acknowledgments

GenePattern Team

Peter Carr
David Eby
Barbara Hill
Edwin Juarez
Ted Liefeld
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

NIGMS

National Institutes of Health

National Cancer Institute

Resources

GenePattern Notebook
genepattern-notebook.org

GenePattern
genepattern.org

Public GenePattern server
genepattern.broadinstitute.org

Indiana University GenePattern server
gp.indiana.edu

GenePattern Archive (GPArc)
gparc.org

GenePattern Twitter
@genepattern

GenePattern GitHub
github.com/genepattern

GenePattern DockerHub
hub.docker.com/r/genepattern