From Jupyter to GenePattern
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

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**
- Preprocess
- SOM Clustering
- Class Neighbors
- Weighted Voting Cross-Val
- SOM Cluster Viewer
- Marker Selection Viewer
- Prediction Results Viewer
- Weighted Voting Train/Test
- Prediction Results Viewer

Support for in silico reproducible research

**Clients**

- Visualizer
- Web
- Programming

**Access for all levels of user**

Golub and Slonim et. al 1999

GISTIC
Tophat
GSEA
NMF
MuTect
Cuffdiff
Cuffdiff

Hundreds of analyses and visualizations

Easy addition of new tools

User-contributed modules

Support for in silico reproducible research

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

[Links: genepattern.broadinstitute.org, genepattern.org, gparc.org]
### Analysis Tool Repository

<table>
<thead>
<tr>
<th>Copy Number Divide by Normals</th>
<th>GSEA</th>
<th>Variation Filter</th>
<th>Cuffdiff</th>
</tr>
</thead>
<tbody>
<tr>
<td>GISTIC</td>
<td>CBS</td>
<td>k-Nearest Neighbors</td>
<td>MutSigCV</td>
</tr>
<tr>
<td>Classification and Regression Trees</td>
<td>Support Vector Machines</td>
<td>Hierarchical Clustering</td>
<td>Picard Sort Sam</td>
</tr>
<tr>
<td>TopHat</td>
<td>Expression File Creator</td>
<td>Metagene Projection</td>
<td>RNASeQC</td>
</tr>
</tbody>
</table>
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**
  - Preprocess Dataset
  - SOM Clustering
  - SOM Cluster Viewer
- **all_aml_test**
  - Preprocess Dataset
  - Comparative Marker Selection
  - Weighted Voting Cross-Validation
  - Weighted Voting Train/Test
  - Comparative Marker Selection Viewer
  - Prediction Results Viewer
  - Prediction Results Viewer

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 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.cv.txt')`
Out[11]: `<gp.core.GPFile at 0x7f9edc54f6a0>`

GenePattern FastQC
Provides quality control metrics on raw sequence data
* Required Field

input file*  
[Upload File]  [Add Upstream File or URL]  [Drag Files Here]

20GB 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

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

# 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)

# 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)

# Slice the data for SVM fitting
X = train_data.data[:, i2]
y = train_classes

# 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)

# 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))

plt.subplot(1, 1, 1)
```
SVM Example #2

[Image of GenePattern SVM interface]

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 cis filename
Upload File... Add Upstream File or URL... Drag Files Here
2GB file upload limit using the Upload File... button.

The training class file - .cis

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.cdf

The name of the output file for prediction results

model output file
<train.data.filename_basename>.model
```
GenePattern Cells

Authentication Cell

Analysis Cell

Job Cell
Authentication Cells

GenePattern Login

GenePattern Server
Broad Institute

GenePattern Username
Username

GenePattern Password
Password

Log into GenePattern  Register an Account

GenePattern tabor

-- 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 maintainence window.Thanks,The GenePattern Team -- Set 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

GenePattern Extract/ComparativeMarkerResults

* 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 >= 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_base_name>.fill

The base name for the output files

* Required Field
Python Function GUI

- Turn any Python function into an interactive user interface
Rich Text Markdown Editor

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

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

- Complete programmatic access
- Automatic integration with GenePattern cell data

```python
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

http://genepattern-notebook.org
Publish & Share Notebooks

- 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

- matplotlib
- pandas
- IPython
- Anaconda
- R
- JupyterHub
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
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