



# Use of Globus in GenePattern at UCSD

Ted Liefeld

Mesirov Lab, School of Medicine  
University of California, San Diego

# > 10,000 active tools for bioinformatics

**gatk** [Best-Practices](#) [Documentation](#) [Blog](#) [Forum](#) [Events](#) [Download](#)

[Back to Tool Docs Index](#)

## MuTect2

Call somatic SNPs and indels via local re-assembly of haplotypes

## HISAT2

graph-based alignment of next generation sequencing reads to a population of genomes

## DESeq2

platforms **all** downloads **top 5%** in Bioc **4.5 years** build **ok**

DOI: [10.18129/B9.bioc.DESeq2](#) [f](#) [t](#)

Differential gene expression analysis based on the negative binomial distribution

## Samtools

Samtools is a suite of programs for interacting with high-throughput sequencing data. It consists of three separate repositories:

- Samtools** Reading/writing/editing/indexing/viewing SAM/BAM/CRAM format
- BCFtools** Reading/writing BCF2/VCF/gVCF files and calling/filtering/summarising SNP and short indel sequence variants
- HTSlib** A C library for reading/writing high-throughput sequencing data

Samtools and BCFtools both use HTSlib internally, but these source packages contain their own copies of htlib so they can be built independently.

## Bowtie 2

Fast and sensitive read alignment

## Principal Component Analysis

## Classification And Regression Trees for Machine Learning

## Bowtie 2

Fast and sensitive read alignment

## 1.4. Support Vector Machines

## Picard

build **passing**

A set of command line tools (in Java) for manipulating sequencing (HTS) data and formats such as SAM

## COMBAT:

'COMBATING' BATCH EFFECTS WHEN COMBINING BATCHES OF GENE EXPRESSION MICROARRAY DATA

## Hierarchical Clustering / Dendrograms

## Cufflinks

Transcriptome assembly and differential expression analysis for RNA-Seq.

## Burrows-Wheeler Aligner

Introduction

BWA is a software package for mapping low-divergent sequences against a large reference genome, such as the human genome. It consists of three algorithms:

## flowCore: a Bioconductor package for high throughput flow cytometry

## Integrative Genomics Viewer

## NMF: Non-negative Matrix Factorization

What is HAPSEG?

HAPSEG is a probabilistic method to interpret bi-allelic marker data in cancer samples.

## GSEA

Gene Set Enrichment Analysis

## MAGeCK

Model-based Analysis of Genome-wide CRISPR-Cas9 Knockout

What is RNA-SeQC?

RNA-SeQC is a java program which

## SEURAT

R toolkit for single cell genomics

## Cytoscape

Network Data Integration, Analysis, and Visualization in a Box

## Trimmomatic: A flexible read trimming tool for Illumina NGS data

## MSigDB

Molecular Signatures Database

## FLAME

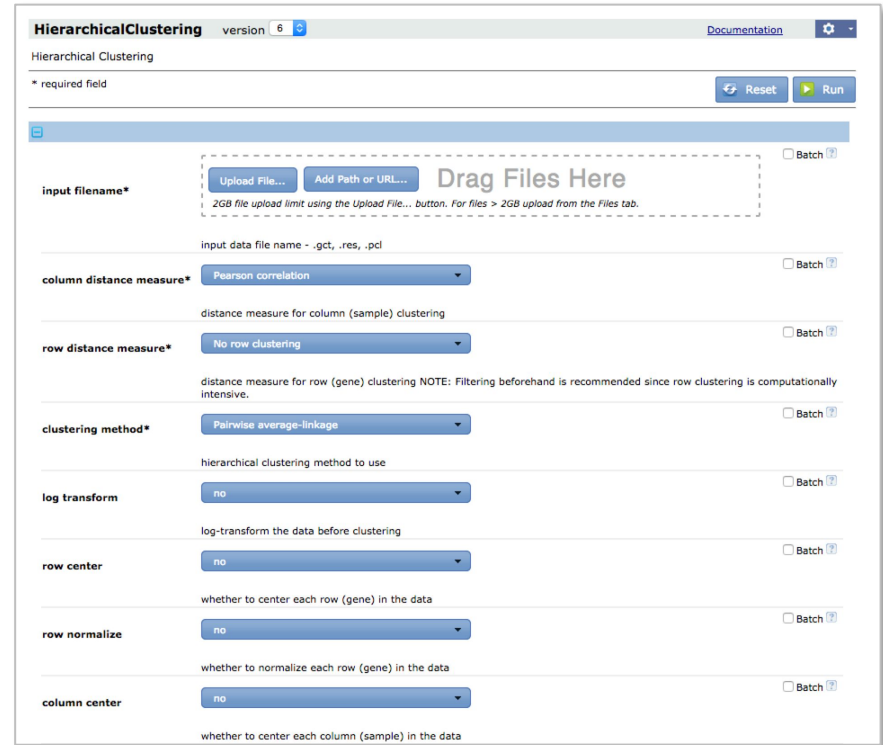
Flow analysis with Automated Multivariate Estimation

Constellation Map: Downstream visualization and interpretation of gene set enrichment results [version 1; referees: 2 approved]

# GenePattern wraps software tools in an accessible visible format

```
> java -Djava.awt.headless=true
-Dwin=cluster.exe -Dmac=clusterMac
-Dlinux=clusterLinux
-Dlinux64=clusterLinux64 -cp
hcl.jar/legacy-gp-modules.jar/ant.jar
org.genepattern.modules.hcl.RunCluster -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
```

Standard “command-line” method for running analysis



The screenshot shows the 'HierarchicalClustering' web interface. At the top, it displays 'HierarchicalClustering version 6' and a 'Documentation' link. Below the title, there are 'Reset' and 'Run' buttons. The main area contains several configuration options, each with a dropdown menu and a 'Batch' checkbox:

- input filename\***: A dashed box for file upload with 'Upload File...', 'Add Path or URL...', and 'Drag Files Here' options. A note below states: '2GB file upload limit using the Upload File... button. For files > 2GB upload from the Files tab.'
- input data file name**: A text field with a dropdown menu showing '-.gct, .res, .pcl'.
- column distance measure\***: A dropdown menu set to 'Pearson correlation'.
- distance measure for column (sample) clustering**: A text field.
- row distance measure\***: A dropdown menu set to 'No row clustering'.
- distance measure for row (gene) clustering**: A text field with a note: 'NOTE: Filtering beforehand is recommended since row clustering is computationally intensive.'
- clustering method\***: A dropdown menu set to 'Pairwise average-linkage'.
- hierarchical clustering method to use**: A text field.
- log transform**: A dropdown menu set to 'no'.
- log-transform the data before clustering**: A text field.
- row center**: A dropdown menu set to 'no'.
- whether to center each row (gene) in the data**: A text field.
- row normalize**: A dropdown menu set to 'no'.
- whether to normalize each row (gene) in the data**: A text field.
- column center**: A dropdown menu set to 'no'.
- whether to center each column (sample) in the data**: A text field.

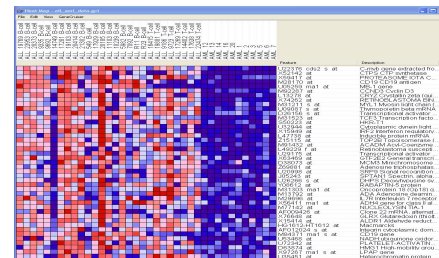
Corresponding GenePattern visual representation

# GenePattern

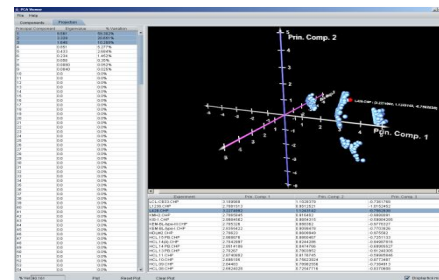
## Analysis interface

The screenshot shows the GenePattern web interface for HierarchicalClustering. The browser address bar shows the URL: [genepattern.broadinstitute.org/gp/pages/index.js?fbid=urn:isid:broad.mit.edu:cancer.software.genepattern.module.analysis](http://genepattern.broadinstitute.org/gp/pages/index.js?fbid=urn:isid:broad.mit.edu:cancer.software.genepattern.module.analysis). The interface includes a navigation menu with 'Modules & Pipelines', 'Suites', 'Job Results', 'Resources', 'Downloads', and 'Help'. A sidebar on the left lists 'Favorite Modules' (ComparativeMarkerSelectionViewer, FeatureSummaryViewer, HeatMapView, HierarchicalClusteringViewer) and 'Recent Modules'. The main panel is titled 'HierarchicalClustering version 1.6'. It features a 'Reset' button and a 'Run' button. The 'input filename\*' field has an 'Upload File...' button and a 'Drag Files Here' area. Below this, various clustering parameters are set to default values: 'column distance measure\*' is 'Pearson correlation', 'row distance measure\*' is 'No row clustering', 'clustering method\*' is 'Pairwise average-linkage', and 'log transform', 'row center', 'row normalize', 'column center', and 'column normalize' are all set to 'no'. The footer contains 'About GenePattern | Contact Us' and '©2003-2014 Broad Institute, MIT'.

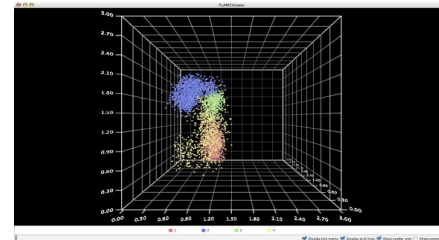
[www.genepattern.org](http://www.genepattern.org)



Gene expression heatmap



3D principal components



Cell populations

# Hundreds of genomics analysis tools

## Machine learning

- Clustering, classification, dimension reduction

## Gene expression analysis

- DESeq2, BWA, HISAT2, HT-Seq, Salmon, Kallisto, Cufflinks, etc.

## Single-cell RNA-seq analysis

- Seurat, Scanpy, STREAM, CONOS

## Cancer genomics

- GISTIC, MutSigCV, HAPSEG,

## Gene Set Enrichment Analysis

- GSEA, ssGSEA, GSEAPreranked

## Collaborative projects

- OpenCRAVAT *Karchin Lab*
- AMARETTO *Pochet Lab*
- CoGAPS *Fertig Lab*
- MutPanning *Van Allen Lab*
- NDEx *Ideker Lab*
- Next-Generation Clustered Heatmaps *Weinstein Lab (beta)*

## Other

- Proteomics, Flow Cytometry, Network Analysis, Data import and formatting utilities, etc

The screenshot displays the GenePattern Notebook interface for the Hierarchical Clustering tool. The interface is titled "GenePattern Notebook Untitled (unsaved changes)" and includes a menu bar with options like File, Edit, View, Insert, Cell, Kernel, Navigate, Widgets, and Help. The tool configuration panel is highlighted with a red border and contains the following settings:

- input filename\***: Upload File... Add File or URL...
- column distance measure\***: Pearson correlation (distance measure for column (sample) clustering)
- row distance measure\***: No row clustering (distance measure for row (gene) clustering NOTE: Filtering beforehand is recommended since row clustering is computationally intensive.)
- clustering method\***: Pairwise average-linkage (hierarchical clustering method to use)
- log transform**: no (log-transform the data before clustering)
- row center**: no (whether to center each row (gene) in the data)
- row normalize**: no (whether to normalize each row (gene) in the data)
- column center**: no (whether to center each column (sample) in the data)
- column normalize**: no (whether to normalize each column (sample) in the data)
- output base name\***: <input.filename\_basename> (base name for output files)

Annotations on the screenshot include "Upload datasets" pointing to the input filename section and "Set input parameters" pointing to the distance and clustering method sections.

# The GenePattern Notebook Environment

**Single-Cell RNA-seq Clustering Analysis Notebook**

Author - Clarence Mah  
Email - ckmah.ucsd.edu

This notebook analyzes a dataset of 3K Peripheral Blood Mononuclear Cells (PBMCs) from a Healthy Donor available from 10X Genomics, sequenced on the Illumina NextSeq 500. Steps are modeled after the [Seurat Guided Clustering Tutorial](#) using the [scanpy](#) library.

**Objective**

The goal of this notebook is to provide a standard single-cell RNA-seq analysis workflow for pre-processing, identifying sub-populations of cells by clustering, and exploring biomarkers to explain intra-population heterogeneity.

**Step 1: Setup Analysis**

Load a raw count matrix for a single-cell RNA-seq experiment.

**Instructions**

Provide your data file either as a URL or local file path. Select "2700 PBMCs from a Healthy Donor (example)" from the dropdown menu to load the example dataset.

Supported file formats: csv, xlsx, txt, tsv, tab, data, h5, h5ad, soft.gz, txt.gz, anndata, ztx\*

Text and Excel files (csv, txt, tsv, tab, data, xlsx): Gene and sample names are assumed to be the first column and row respectively.

NOTE\*: The 10x Genomics genomics pipeline generates gene-barcode matrices usually named `matrix.txt`, `genes.tsv`, and `barcodes.tsv`. If the `ztx` files is provided, the genes and barcodes files will automatically be imported from the same folder.

GenePattern Setup Analysis [ ]

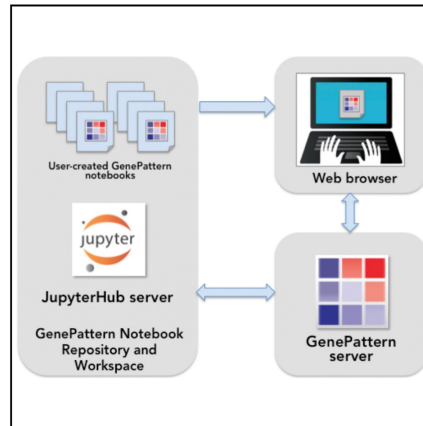
Downloaded file: matrix.txt.

Downloaded file: barcodes.tsv.

Downloaded file: genes.tsv.

- Integrates GenePattern with Jupyter Notebook
- Access hundreds of GenePattern genomic analyses from within a notebook without the need for code

[notebook.genepattern.org](https://notebook.genepattern.org)



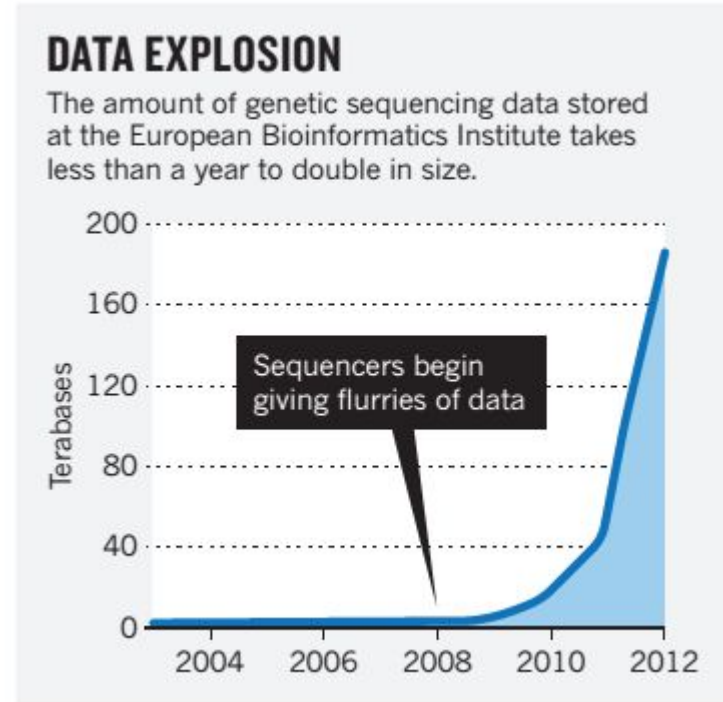
Reich et al.,  
*Cell systems*, 2017

# Genomic Dataset Growth

~2000, 100 samples, 10000 gene transcripts  
<10 MB file sizes

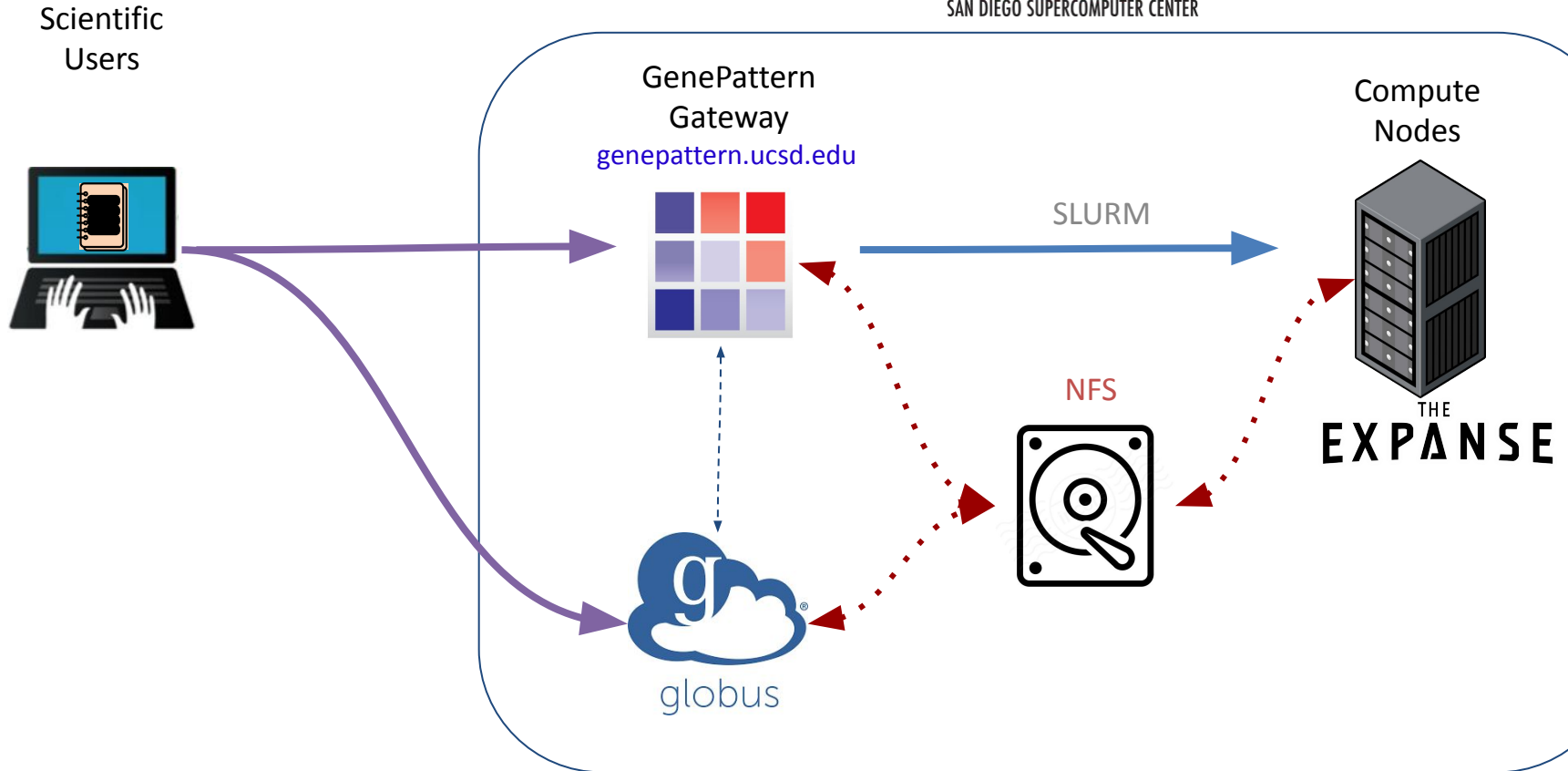
~2015, 3000 samples, 60000 transcripts,  
10 - 100MB

~2022, 320000 samples, ~60000 transcripts,  
40 - 100 **GB**



# High-level GenePattern Architecture

**SDSC**  
SAN DIEGO SUPERCOMPUTER CENTER





# GenePattern and Globus



**Sign in to GenePattern**

[Click to Register](#)

Username:

Password:

[Forgot your password?](#)



[Sign on using your Globus account](#)

You may also use this link to sign in with Google or institutional (for many universities) credentials via Globus.

# GenePattern and Globus

The screenshot displays the GenePattern web interface. At the top, the GenePattern logo and UCSD branding are visible. The navigation bar includes 'Modules & Pipelines', 'Suites', 'Job Results', 'Resources', 'Administration', and 'Help'. A user profile icon for 'ted' and a storage indicator '113 GB / 200 GB' are also present. The main interface is divided into a left sidebar and a central content area. The sidebar shows a file tree under the 'BWA' directory, with files like 's.cerevisiae\_mutant.fastq', 'B1\_1.fq', 'B1\_2.fq', and 'reads\_1.rev.1.bt2'. A 'Drag Files Here' box is at the bottom of the sidebar. The central content area, titled 'BWA', lists several actions: 'Delete Directory', 'Create Subdirectory', 'Upload Files', 'Transfer from Globus Endpoint' (highlighted with a red box), 'Save Directory', and 'Rename Directory'. A 'Notebook' panel on the right contains text about using GenePattern in a Jupyter Notebook.

GenePattern UCSD

Modules & Pipelines Suites Job Results Resources Administration Help 113 GB / 200 GB ted

Modules Jobs Files Notebook

✓ No Jobs Processing Refresh

- uploads
  - AA
  - Aadi
  - all\_aml
  - AshwynSharma
  - BWA
    - s.cerevisiae\_mutant.fastq
  - mutsigcv
  - NMFClusteringInputs
  - salmon
  - test
    - B1\_1.fq
    - B1\_2.fq
    - NMFClustering5.zip
    - reads\_1.rev.1.bt2

Drag Files to Directories Above or

Drag Files Here

### BWA

- Delete Directory  
Permanently delete this directory and all child files.
- Create Subdirectory  
Create a subdirectory in this directory.
- Upload Files  
Upload files to this directory.
- Transfer from Globus Endpoint**  
Transfer files in from a Globus endpoint to this directory.
- Save Directory  
Save a copy of this directory to your local computer as a zip file.
- Rename Directory

## Notebook

GenePattern in a Jupyter Notebook

ab to the left or you can

Learn more

GenePattern.  
GenePattern.

# GenePattern and Globus

The screenshot shows the GenePattern File Manager interface. On the left is a vertical navigation sidebar with icons for FILE MANAGER, BOOKMARKS, ACTIVITY, COLLECTIONS, GROUPS, CONSOLE, FLOWS, ACCOUNT, and LOGOUT. The main window is titled 'File Manager' and shows the current collection as 'TedsUCSDLaptop' and the path as '/Users/liefeld/Desktop/GlobusEndpoint/GenePatternLocal/jliefeld@ucsd.edu'. A 'Bookmark' dropdown is visible. Below the path, there are navigation icons (back, forward, refresh) and a 'view' button. A 'Label' field contains 'To GenePattern' with a note '(optional) Give this task a descriptive label' and 'Submit' and 'Cancel' buttons. A table lists files and folders:

NAME	LAST MODIFIED	SIZE
<input checked="" type="checkbox"/> all_aml_test_copy.gct	8/23/2017, 11:52...	1.90 MB
<input type="checkbox"/> globus	8/9/2021, 10:47...	-

On the right side, a context menu is open, showing options: Share, Transfer or Sync to..., New Folder, Rename, Delete Selected, Download, Open, Upload, and Get Link. The user's name 'ted' is visible in the top right corner.

# GenePattern and Globus

The screenshot displays the GenePattern web interface. At the top, the GenePattern logo and UCSD branding are visible. A navigation bar includes 'Modules & Pipelines', 'Suites', 'Job Results', 'Resources', 'Administration', and 'Help'. A user profile icon for 'ted' is in the top right, along with a storage indicator '113 GB / 200 GB'. Below the navigation, there are tabs for 'Modules', 'Jobs', 'Files', and 'Notebook'. A status bar shows 'No Jobs Processing' and a 'Refresh' button. The main content area shows a 'Welcome to GenePattern' message and a file browser view with an 'uploads' folder. A modal window titled 'GenePattern-Globus Transfers' is open, showing a successful transfer of 'all\_aml\_test\_copy.gct' with a 'Complete!' button and a 'Cancel' button. Below the modal are buttons for 'Clear completed', 'View on Globus', and 'Close'. At the bottom, there is a 'Drag Files Here' area and a 'Web tours' section with two links: 'Click here for a tour of what's new in GenePattern.' and 'Click here for an introductory tour of GenePattern.'

GenePattern UCSD

Modules & Pipelines Suites Job Results Resources Administration Help 113 GB / 200 GB

Modules Jobs Files Notebook

✓ No Jobs Processing Refresh

Welcome to GenePattern

GenePattern-Globus Transfers

all\_aml\_test\_copy.gct Complete! Cancel

Clear completed View on Globus Close

Drag Files to Directories Above or

Drag Files Here

Web tours

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

About GenePattern | Contact Us | ©2003-2020 Regents of the University of California, Broad Institute, MIT

# GenePattern and Globus

The screenshot displays the GenePattern web interface. At the top, the GenePattern logo and UCSD branding are visible. The navigation bar includes 'Modules & Pipelines', 'Suites', 'Job Results', 'Resources', 'Administration', and 'Help'. A user profile 'ted' is shown in the top right corner, along with a storage indicator '113 GB / 200 GB'. The main interface is divided into a left sidebar and a right main area. The sidebar shows a file tree under 'uploads', with 'all\_aml\_test\_copy.gct' highlighted in a red box. Below the file tree is a 'Drag Files Here' area. The main area shows the configuration for the 'NMFClustering' module, version 6. A green banner indicates 'This is a beta version of the module.' The module description is 'Non-negative Matrix Factorization (NMF) Consensus Clustering'. A red box highlights the 'dataset filename\*' field, which contains the URL 'https://genepattern.ucsd.edu/gp/users/ted/BWA/all\_aml\_test\_copy.gct'. Other fields include 'k initial\*' with the value '2' and 'k final\*' with the value '5'. A 'Refresh' button and a 'No Jobs Processing' status are also visible in the sidebar.

GenePattern UCSD

Modules & Pipelines Suites Job Results Resources Administration Help 113 GB / 200 GB ted

Modules Jobs Files Notebook

✓ No Jobs Processing Refresh

- uploads
  - AA
  - Aadi
  - all\_aml
  - AshwynSharma
  - BWA
    - all\_aml\_test\_copy.gct**
    - s.cerevisiae\_mutant.rastq
  - mutsigcv
  - NMFClusteringInputs
  - salmon
  - test
    - B1\_1.fq
    - B1\_2.fq
    - NMFClustering5.zip

Drag Files to Directories Above or

Drag Files Here

**NMFClustering** version 6

*This is a beta version of the module.*

Non-negative Matrix Factorization (NMF) Consensus Clustering

\* required field

dataset filename\* **https://genepattern.ucsd.edu/gp/users/ted/BWA/all\_aml\_test\_copy.gct**

k initial\* 2

k final\* 5

# Acknowledgements

---

## Mesirov Lab UCSD

Anthony Castanza

David Eby

Edwin Huang

Forrest Kim

Michael Reich

Thorin Tabor

Helga Thorvaldsdottir

Alexander Wenzel

## SDSC

Mahidhar Tatineni

Richard Wagner

## Globus

Brigitte Raumann



[www.genepattern.org](http://www.genepattern.org)