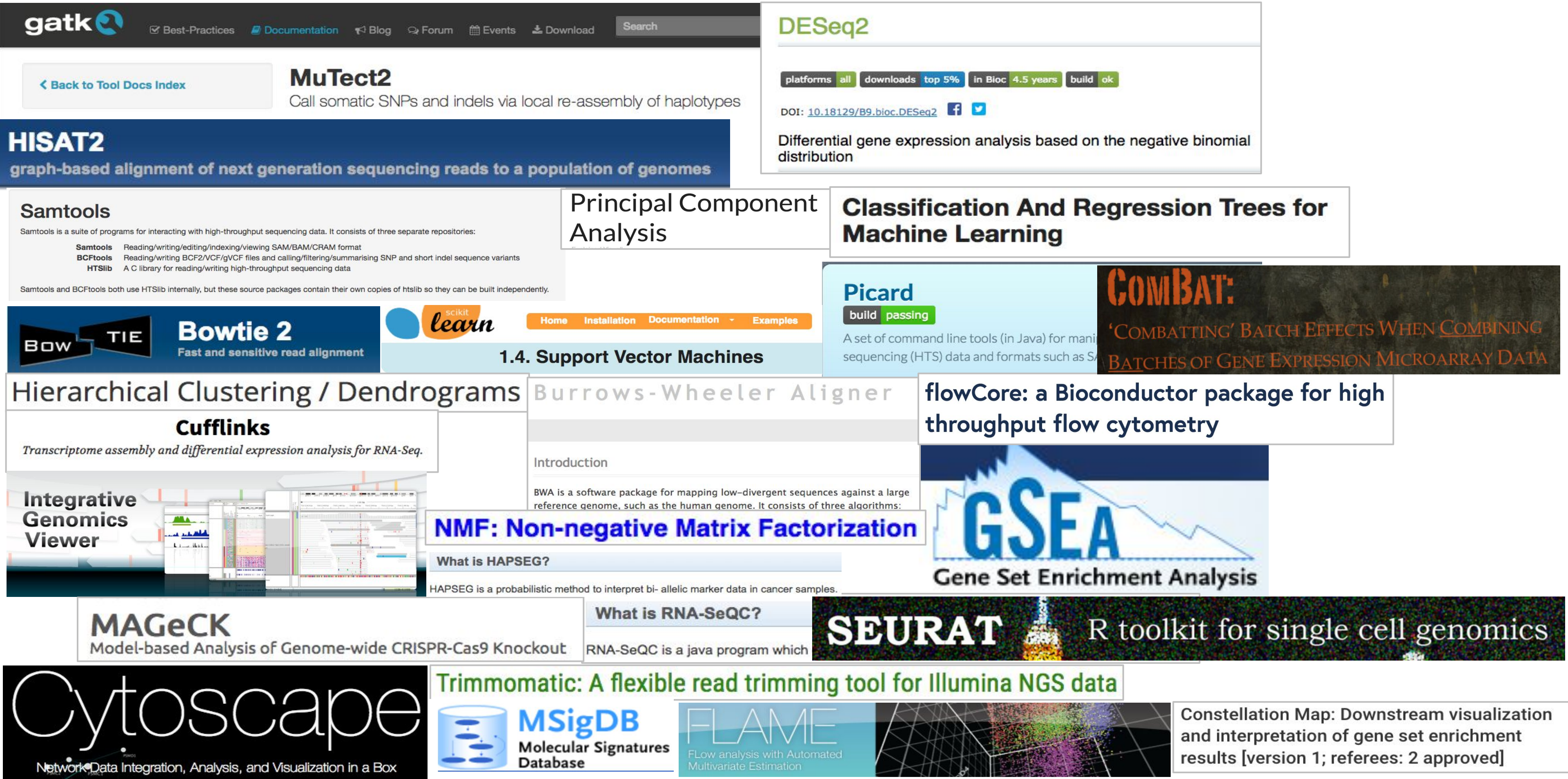
Use of Globus in GenePattern at UCSD

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 · Search

< 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

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 htslib so they can be built independently.

Principal Component Analysis

Classification And Regression Trees for Machine Learning

Picard

build passing

A set of command line tools (in Java) for mani

sequencing (HTS) data and formats such as S

COMBAT:

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

BOW TIE

Bowtie 2

Fast and sensitive read alignment

scikit learn · Home · Installation · Documentation · Examples

1.4. Support Vector Machines

Hierarchical Clustering / Dendrograms

Burrows-Wheeler Aligner

flowCore: a Bioconductor package for high throughput flow cytometry

Cufflinks

Transcriptome assembly and differential expression analysis for RNA-Seq.

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:

Integrative Genomics Viewer

NMF: Non-negative Matrix Factorization

GSEA

Gene Set Enrichment Analysis

What is HAPSEG?

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

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

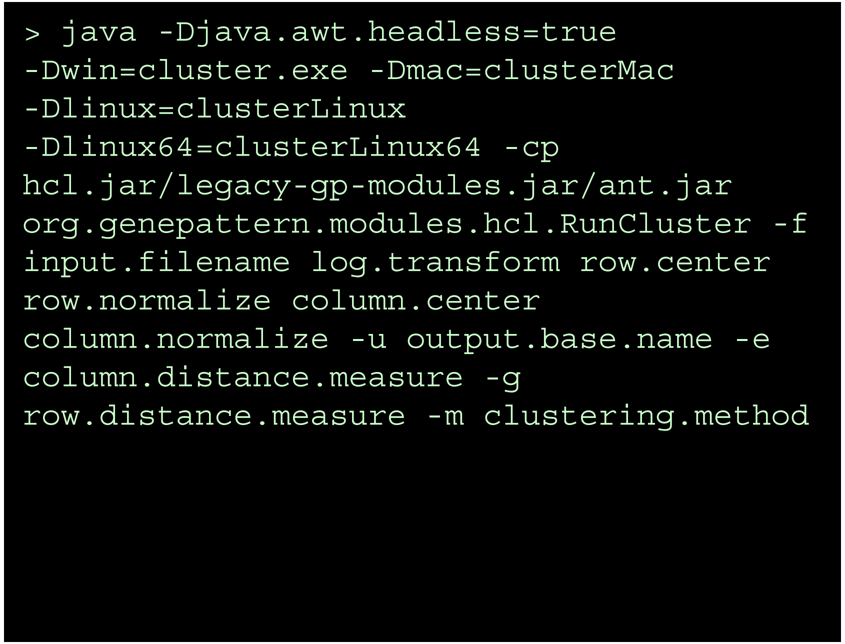
Molecular Signatures Database

FLAME

FLow analysis with Automated Multivariate Estimation

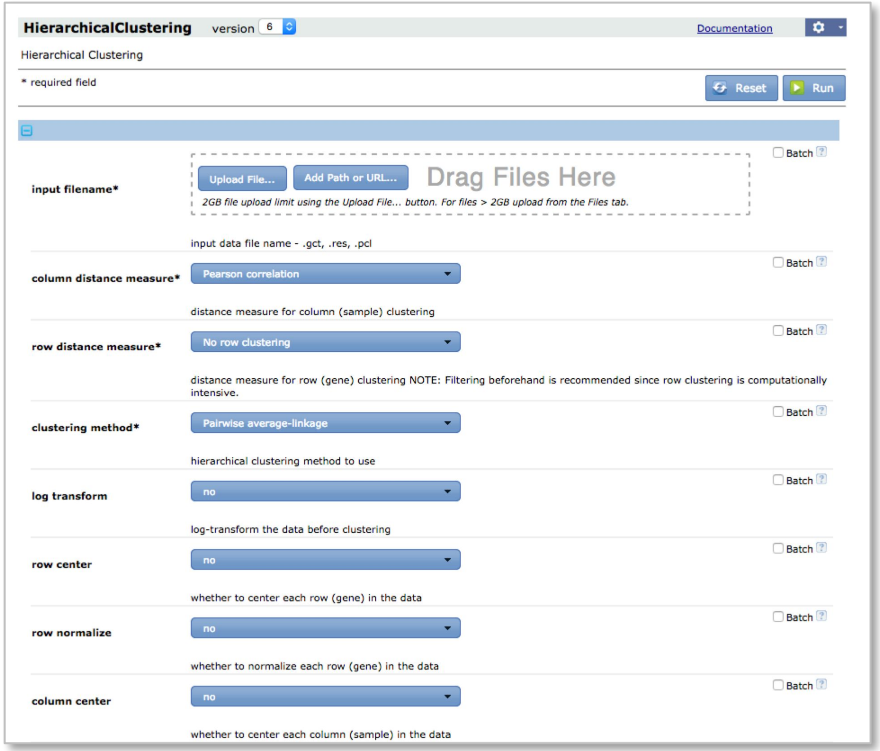
Constellation Map: Downstream visualization and interpretation of gene set enrichment results [version 1; referees: 2 approved]

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



HierarchicalClustering version 6

Hierarchical Clustering

* required field

input filename*

2GB file upload limit using the Upload File... button. For files > 2GB upload from the Files tab.

input data file name - .gct, .res, .pcl

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

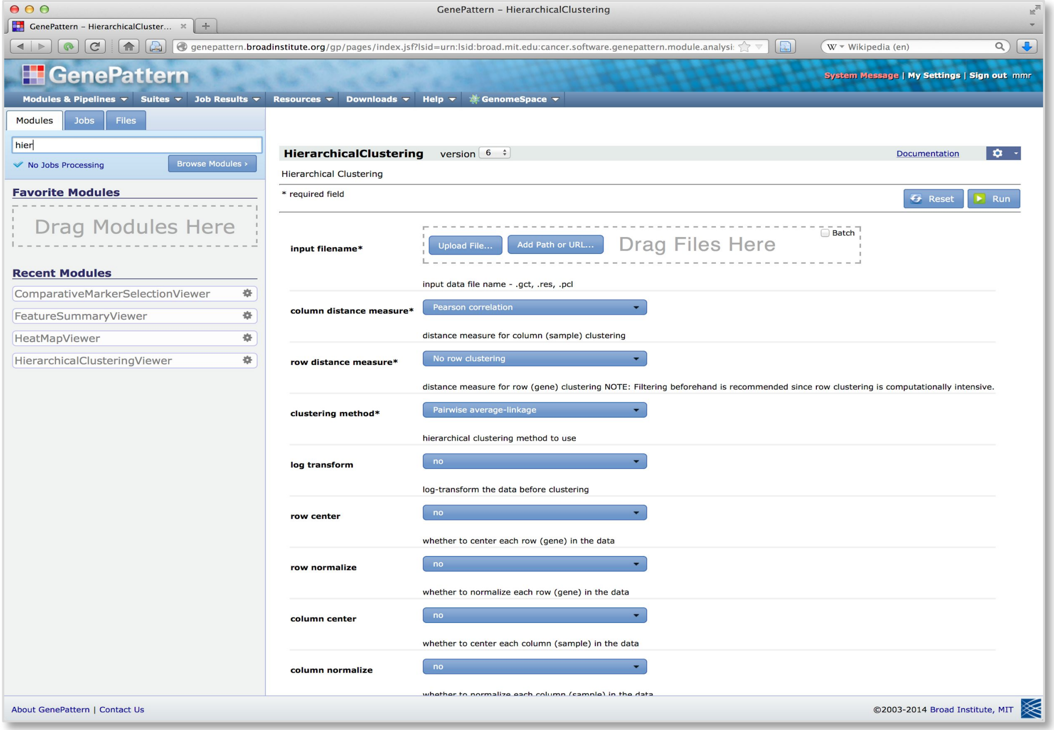
column center no

whether to center each column (sample) in the data

Corresponding GenePattern visual representation

GenePattern

Analysis interface



GenePattern – HierarchicalClustering

genepattern.broadinstitute.org/gp/pages/index.jsf?lsid=urn:lsid:broad.mit.edu:cancer.software.genepattern.module.analysi

GenePattern

System Message | My Settings | Sign out mmr

Modules & Pipelines ▾ Suites ▾ Job Results ▾ Resources ▾ Downloads ▾ Help ▾ GenomeSpace ▾

Modules | Jobs | Files

hier

✓ No Jobs Processing Browse Modules ›

Favorite Modules

Drag Modules Here

Recent Modules

- ComparativeMarkerSelectionViewer
- FeatureSummaryViewer
- HeatMapViewer
- HierarchicalClusteringViewer

HierarchicalClustering version 6 Documentation

Hierarchical Clustering

* required field Reset Run

input filename*: Upload File... | Add Path or URL... | Drag Files Here | □ Batch
input data file name - .gct, .res, .pcl

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

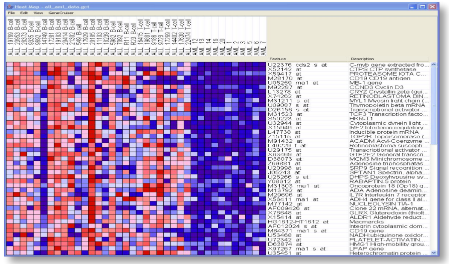
column center: no
whether to center each column (sample) in the data

column normalize: no
whether to normalize each column (sample) in the data

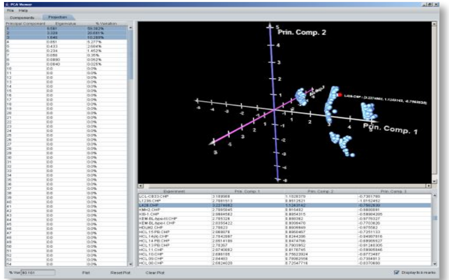
About GenePattern | Contact Us

©2003-2014 Broad Institute, MIT

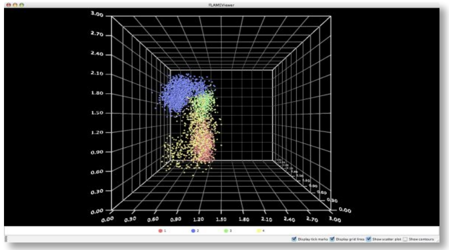
www.genepattern.org



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

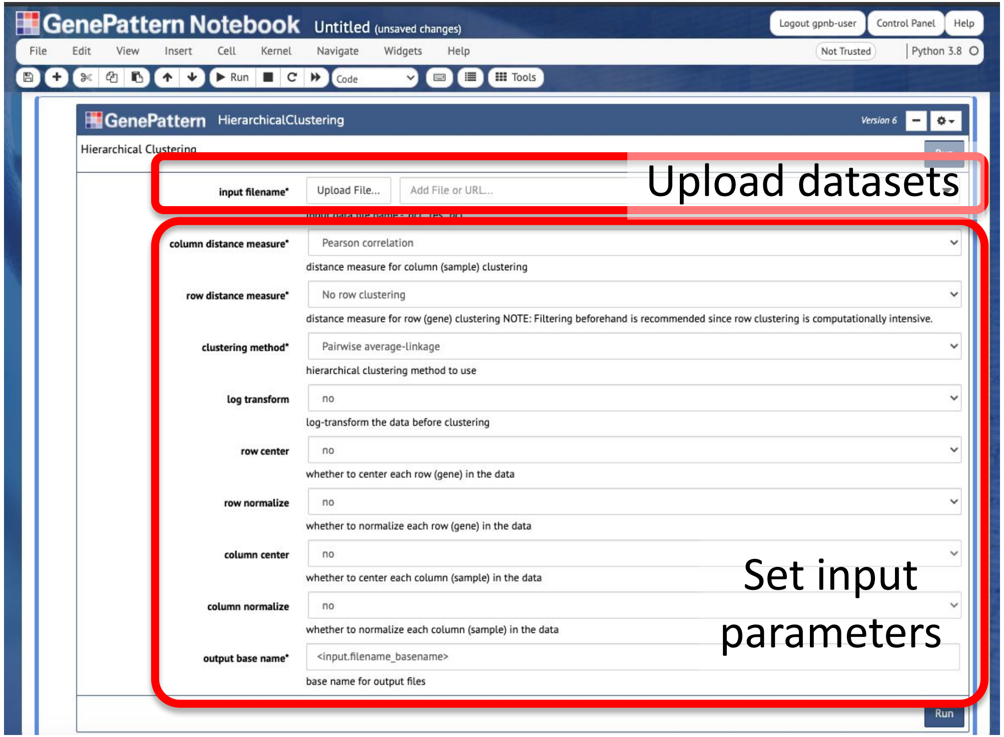
- GSEA, ssGSEA, GSEAPreranked

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



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

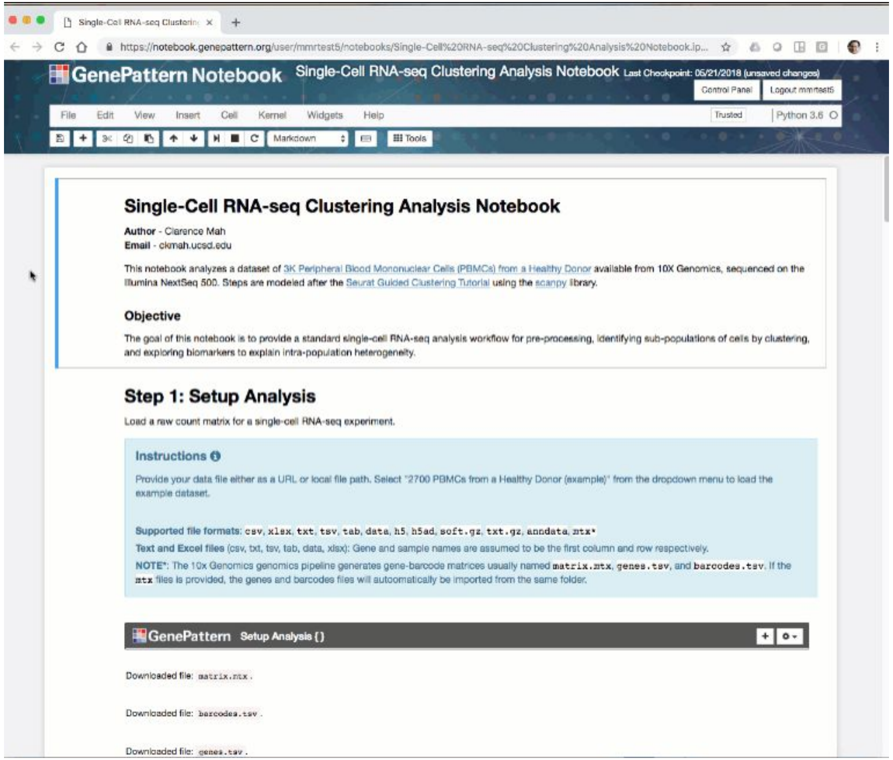
column normalize: no
whether to normalize each column (sample) in the data

output base name*: <input.filename_basename>
base name for output files

Upload datasets

Set input parameters

The GenePattern Notebook Environment


Single-Cell RNA-seq Clusterin ×
https://notebook.genepattern.org/user/mmrtest5/notebooks/Single-Cell%20RNA-seq%20Clustering%20Analysis%20Notebook.ip...
GenePattern Notebook
Control Panel
Logout mmrtest5
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3.6
Markdown
Tools

Single-Cell RNA-seq Clustering Analysis Notebook Last Checkpoint: 05/21/2018 (unsaved changes)

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`, `mtx*`

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.mtx`, `genes.tsv`, and `barcodes.tsv`. If the `mtx` files is provided, the genes and barcodes files will autoomatically be imported from the same folder.

GenePattern Setup Analysis {}

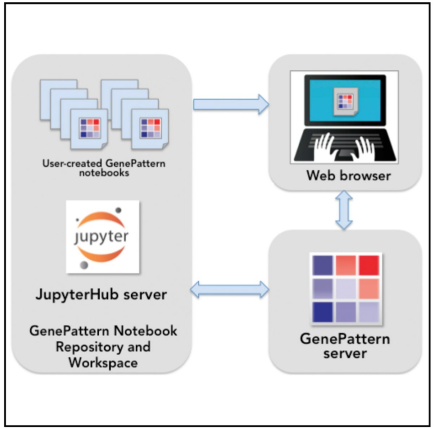
Downloaded file: `matrix.mtx` .

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


User-created GenePattern notebooks
jupyter
JupyterHub server
GenePattern Notebook Repository and Workspace
Web browser
GenePattern server

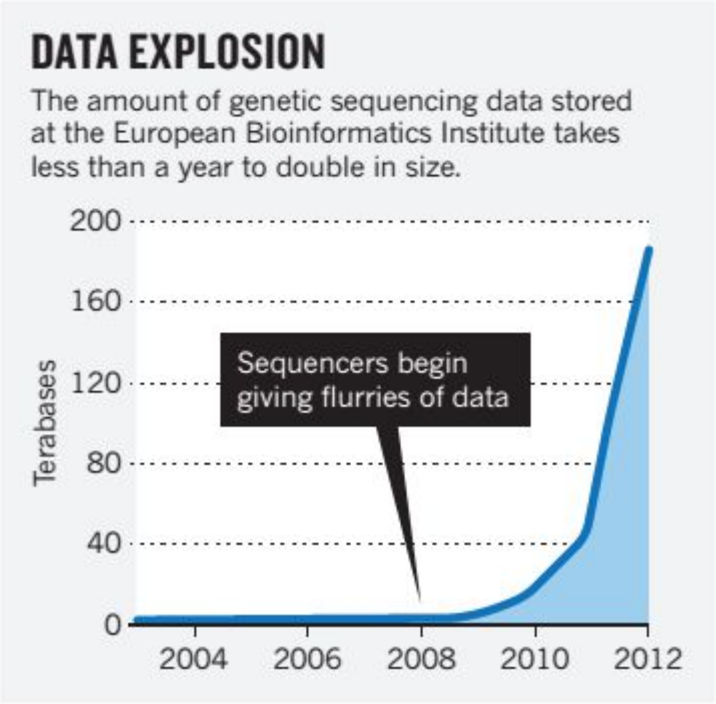
Reich et al., *Cell systems*, 2017

Genomic Dataset Growth

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

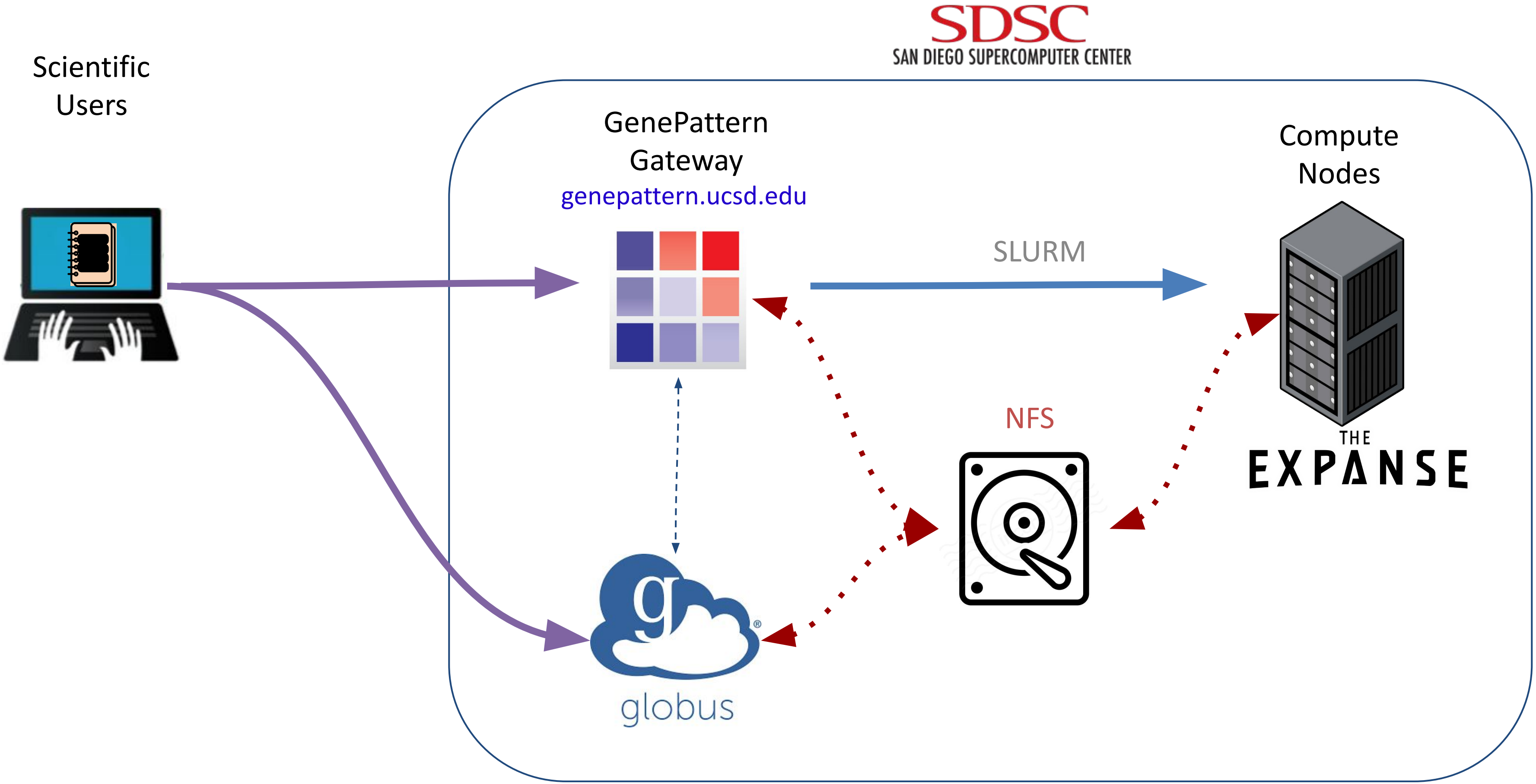
~2015, 3000 samples, 60000 transcripts,
  10 - 100MB

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

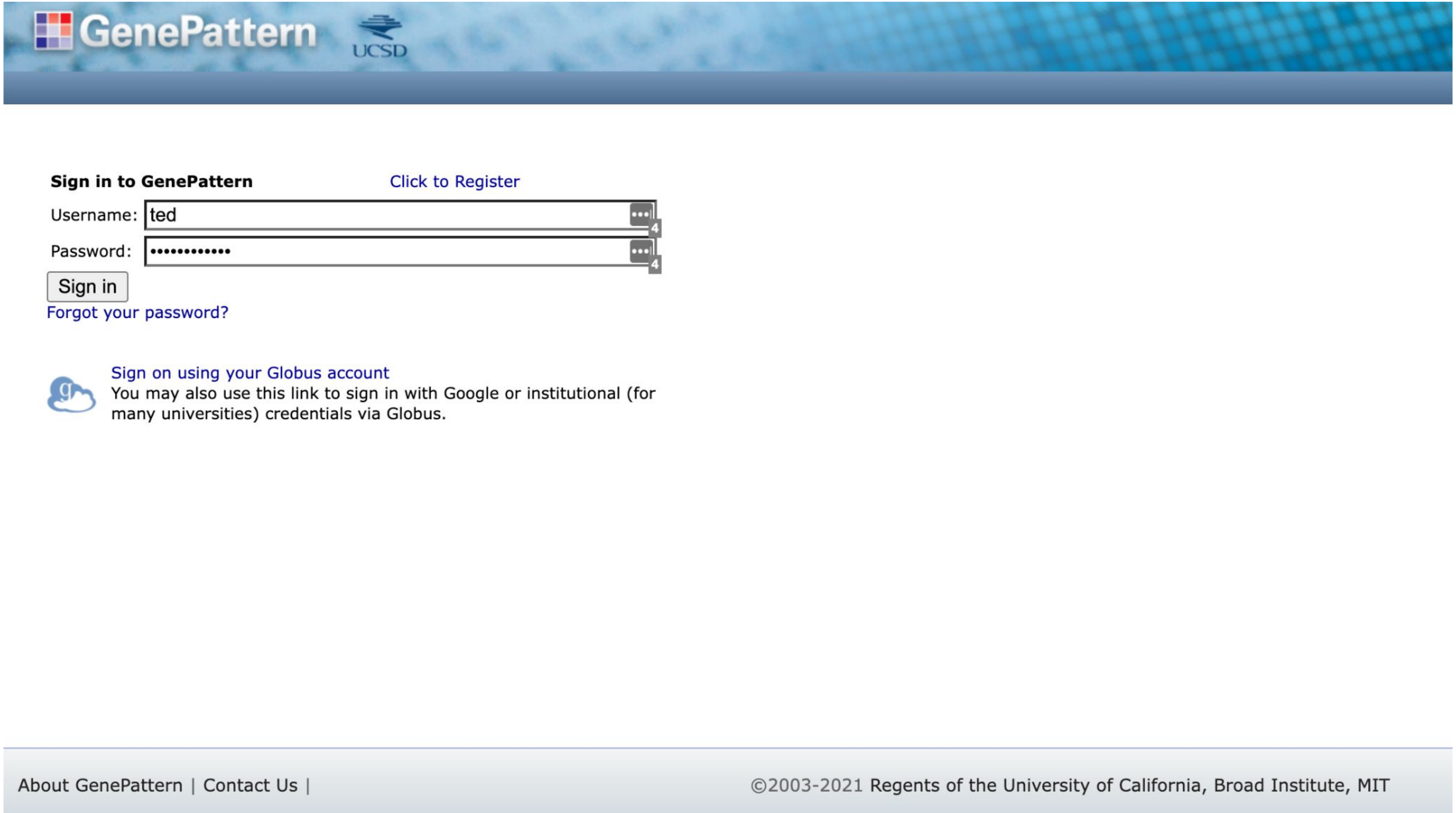


Kashish et al, 2018

High-level GenePattern Architecture



GenePattern and Globus



Sign in to GenePattern Click to Register

Username: ted

Password: •••••••••••••

Sign in

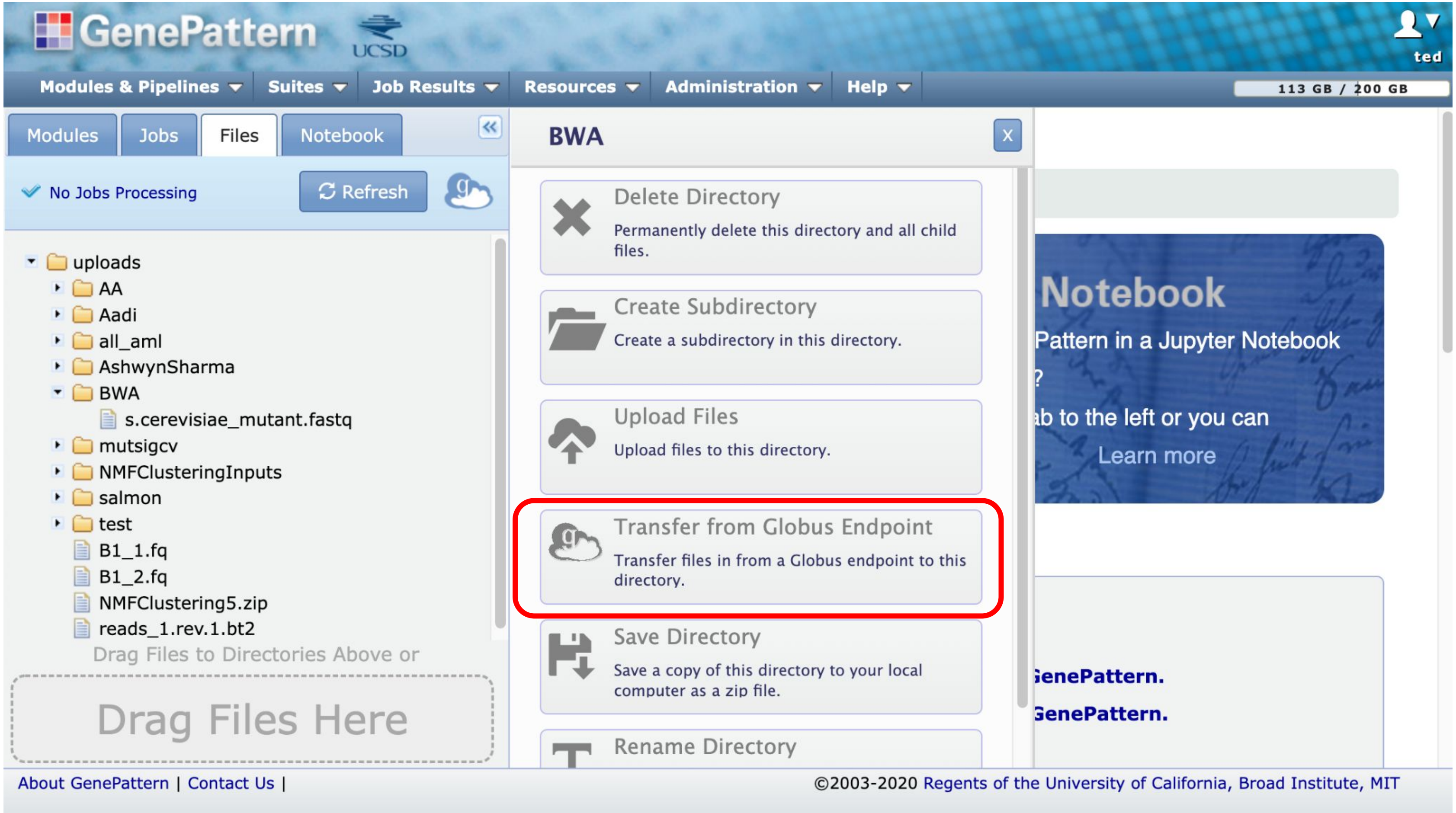
Forgot your password?

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.

About GenePattern | Contact Us |

GenePattern and Globus



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.

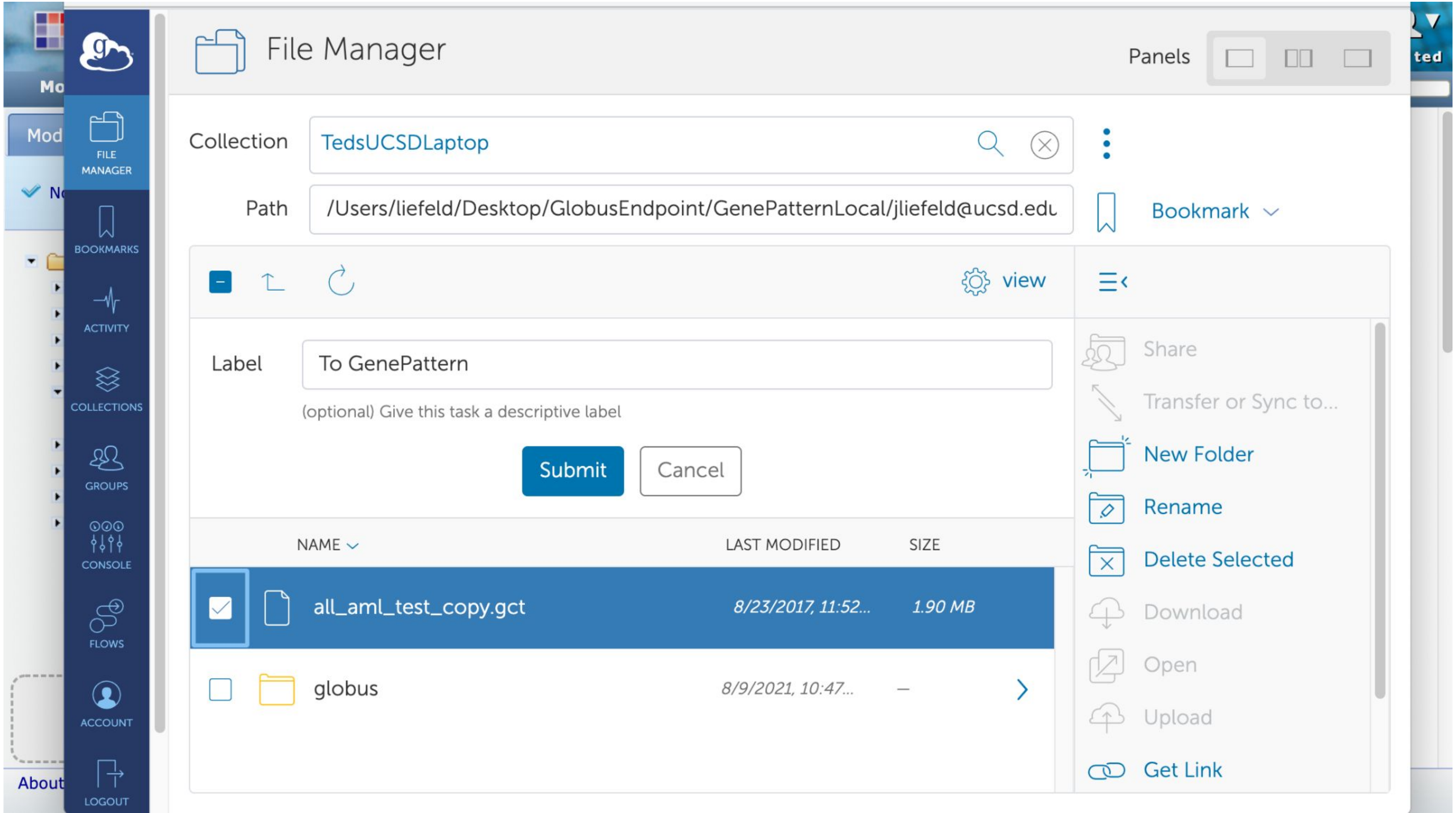
Save Directory
Save a copy of this directory to your local computer as a zip file.

Rename Directory

About GenePattern | Contact Us |

©2003-2020 Regents of the University of California, Broad Institute, MIT

GenePattern and Globus



Collection: TedsUCSDLaptop

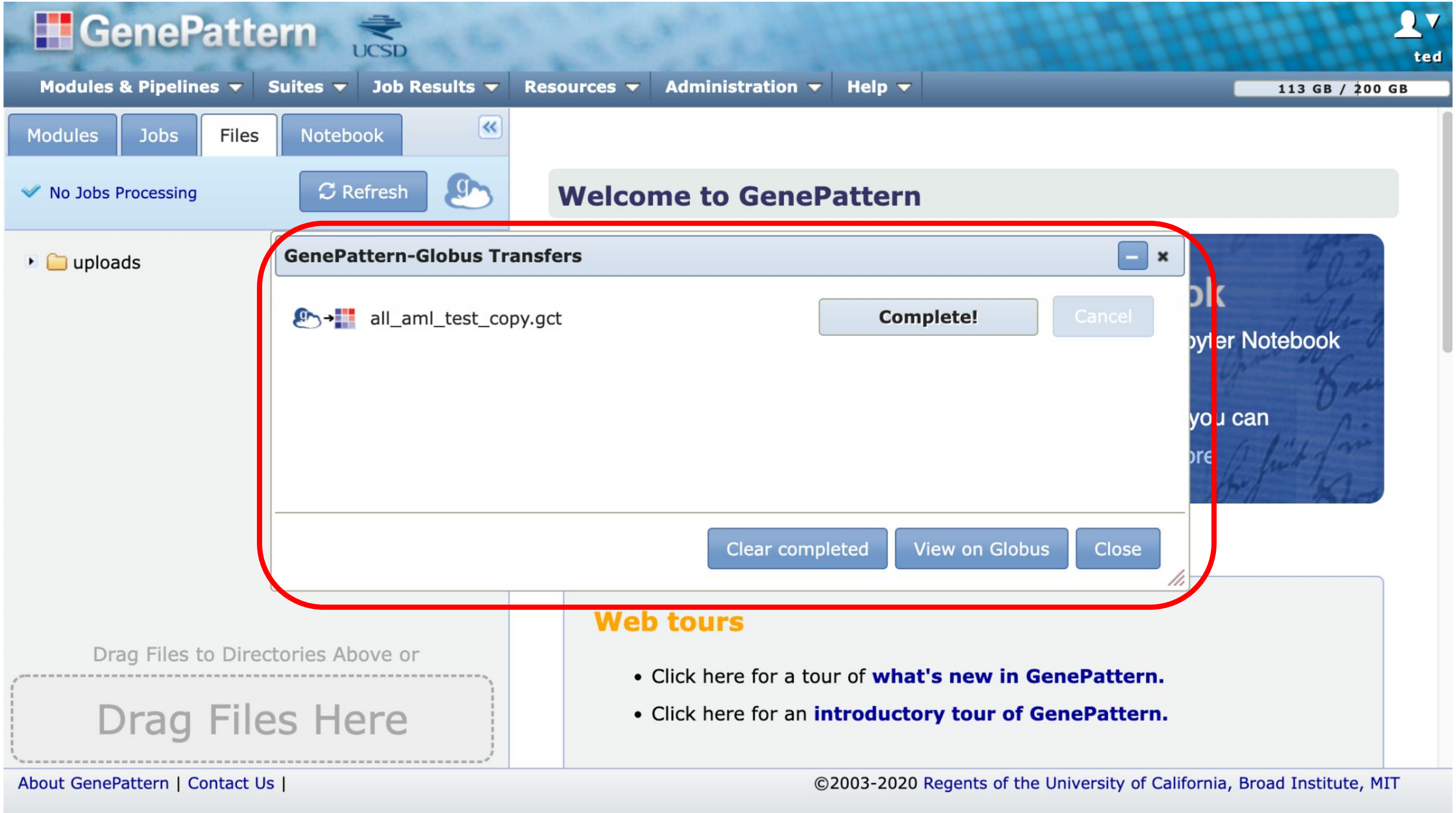
Path: /Users/liefeld/Desktop/GlobusEndpoint/GenePatternLocal/jliefeld@ucsd.edu

Label: To GenePattern

(optional) Give this task a descriptive label

	NAME	LAST MODIFIED	SIZE
☑	all_aml_test_copy.gct	8/23/2017, 11:52...	1.90 MB
□	globus	8/9/2021, 10:47...	–

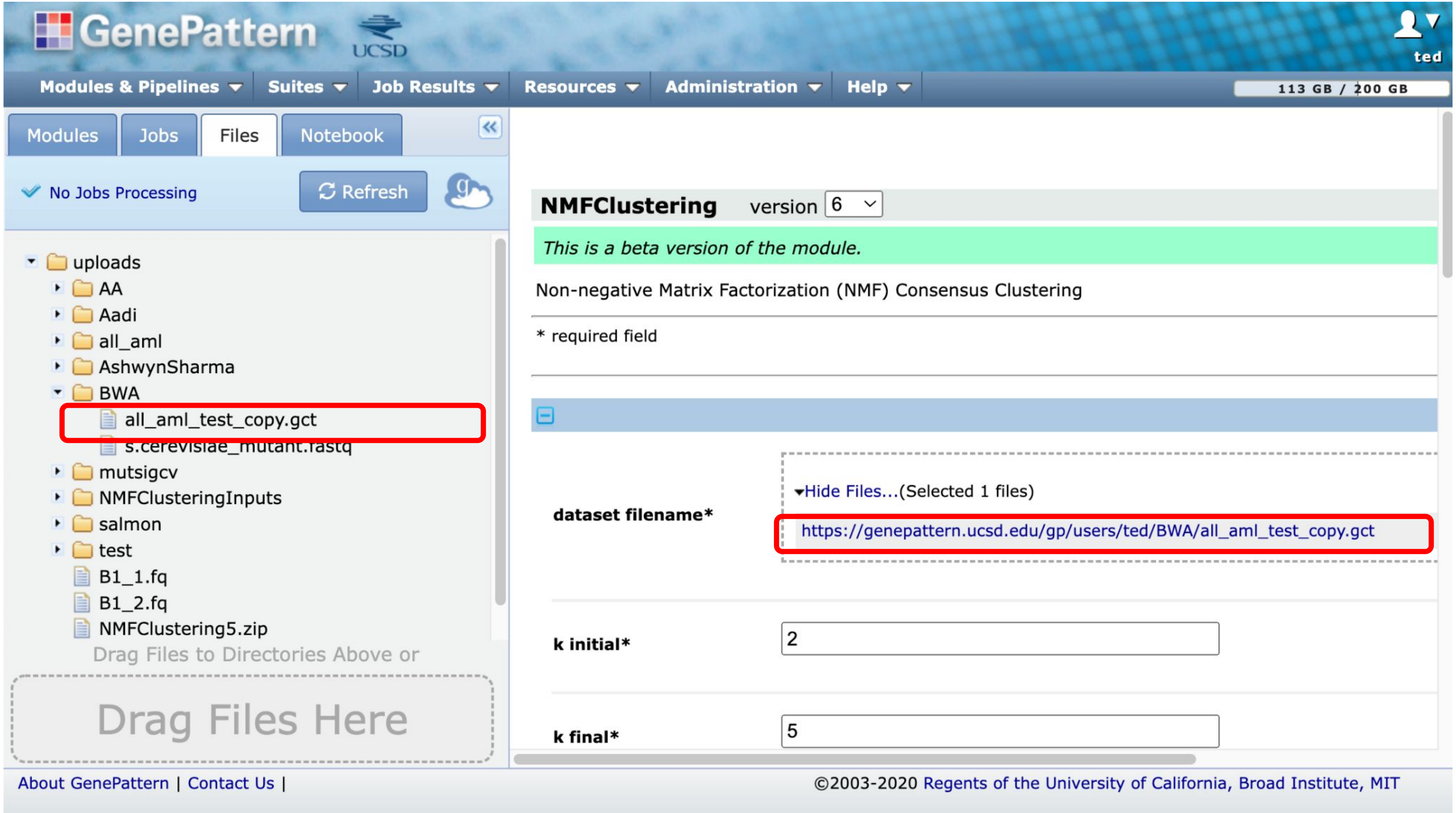
GenePattern and Globus



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

©2003-2020 Regents of the University of California, Broad Institute, MIT

GenePattern and Globus



- uploads
  - AA
  - Aadi
  - all_aml
  - AshwynSharma
  - BWA
    - all_aml_test_copy.gct
    - s.cerevisiae_mutant.fastq
  - 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* Hide Files...(Selected 1 files)
https://genepattern.ucsd.edu/gp/users/ted/BWA/all_aml_test_copy.gct

k initial* 2

k final* 5

About GenePattern | Contact Us |

©2003-2020 Regents of the University of California, Broad Institute, MIT

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

NATIONAL CANCER INSTITUTE®

XSEDE
Extreme Science and Engineering
Discovery Environment

www.genepattern.org