Integrative Genomic Analysis with GenePattern

April 16, 2019

Barbara Hill Thorin Tabor

Agenda

Introduction
GenePattern and Jupyter Notebook essentials

ML and bioinformatic analysis and visualization

Break
ML and bioinformatic analysis and visualization

Notebooks for reproducible research and open science

Best practices for authoring and disseminating notebooks

Case study: single-cell RNA Seq analysis

Discussion and Q&A

GenePattern Overview and Motivation
Introduction to GenePattern Notebook
Data Prep
Differential Analysis
GSEA

Clustering
Class Prediction

Register for an account at:
notebook.genepattern.org

No spaces or special characters in usernames

GenePattern Overview

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

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

HISAT2

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

Classification And Regression Trees for Machine Learning

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

Picard

build passing

A set of command line tools (in Java) for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF.

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

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:

GSEA

Gene Set Enrichment Analysis

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.

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

MAGeCK

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

What is RNA-SeQC?

RNA-SeQC is a java program which computes a series of quality control metrics for RNA-seq data.

Cytoscape

Network Data Integration, Analysis, and Visualization in a Box

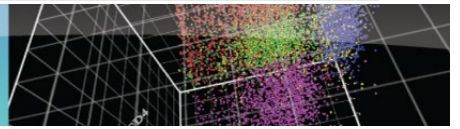
Trimmomatic: A flexible read trimming tool for Illumina NGS data

MSigDB

Molecular Signatures Database

FLAME

FLow analysis with Automated Multivariate Estimation



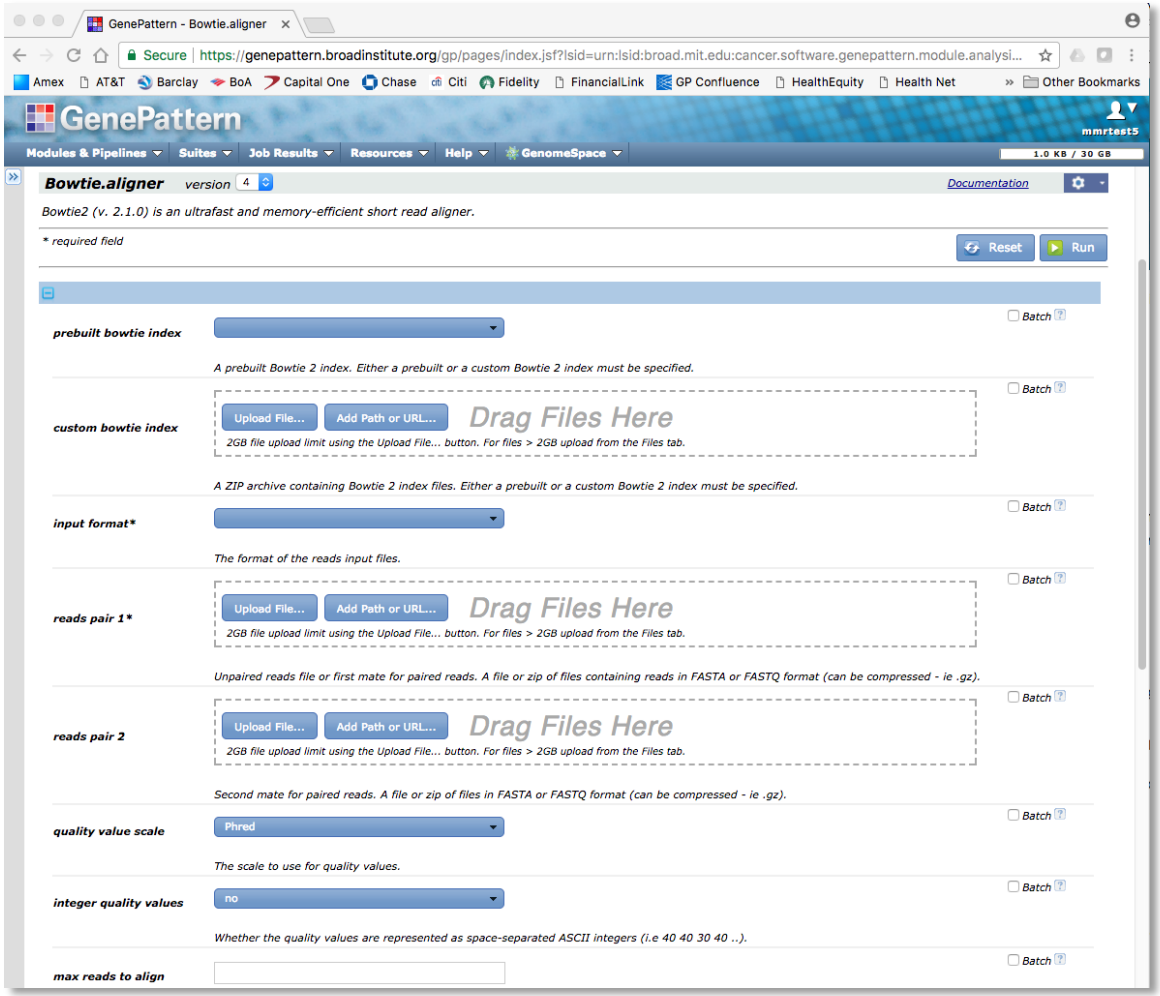
Problems with bioinformatics tool use and interoperability

- Tools are built using different languages and with different architectural assumptions.
- Each tool has its own installation and operational requirements.
- Tools require (sometimes extensive) Unix knowledge.
- Tools are not designed to communicate with each other.

```
bowtie -a --best --strata -S -m 100 -X 400 --chunkmbs 256 --fullref -p 4
Dmel.BDGP5-transcripts \ -1 SRR031714_1.fastq -2 SRR031714_2.fastq |
samtools view -F 0xC -bS - | \ samtools sort -n - ~/Desktop/untreated3-
transcriptome
```

Solution features: Wrapping tools

- "Wrap" tools in a web-based interface
- No installation/running requirements
- No programming required
- Fill out required parameters and provide input files



Solution features: Tool Repository

- Collection of hundreds of wrapped tools
- Gene expression, sequence variation, proteomics, network analysis, machine learning, flow cytometry, etc.
- Searchable by name, keyword, etc.
- Widely-used community tools, lab-developed tools, utilities
- Users can contribute their own tools

Solution features: Reproducibility

- Record and replay of all analyses
- Retain all versions of code – so results can be reproduced even if code changes
- Chain analyses into “pipelines” that can be shared and published

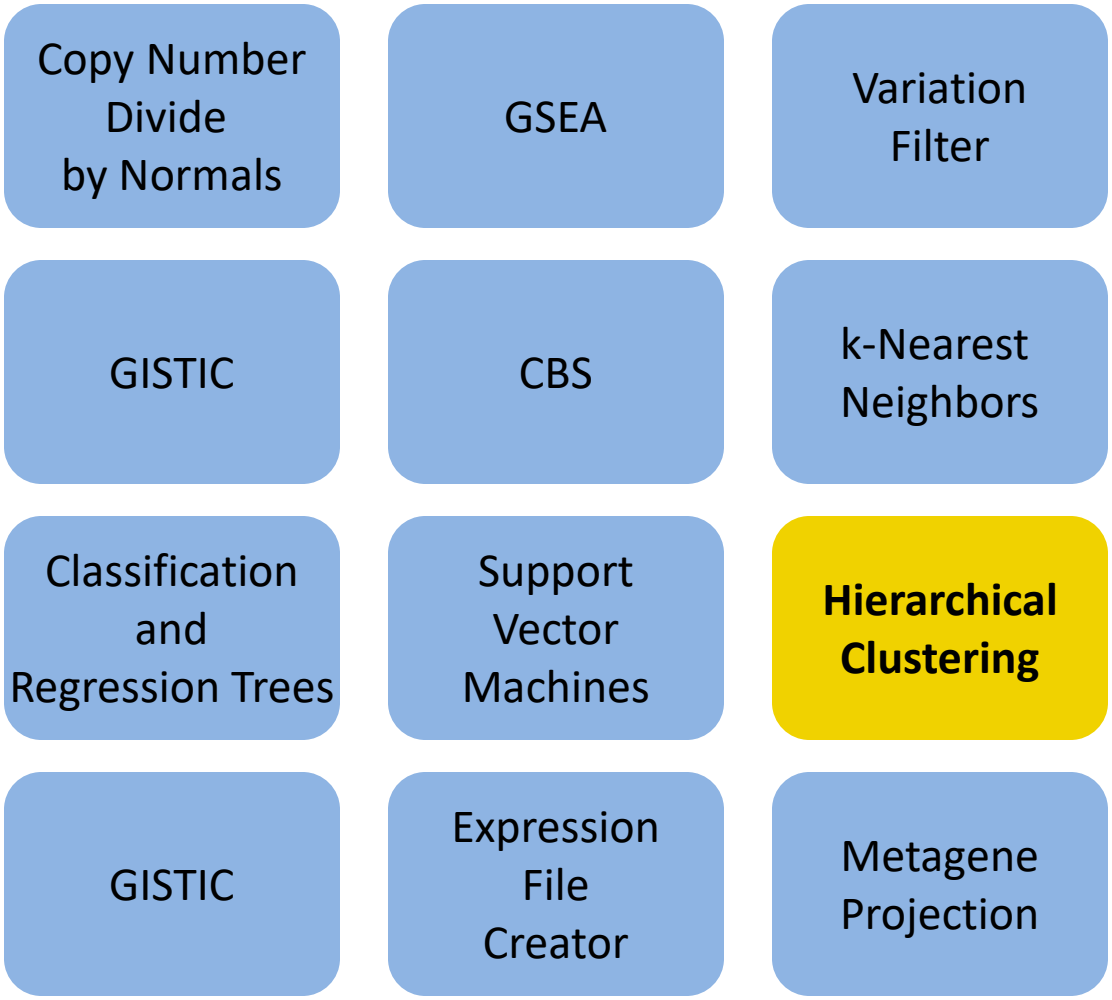
What can you do with GenePattern

Function	Description
Differential Expression	Find the genes that distinguish between two conditions (i.e., tumor vs normal, relapse vs non-relapse, etc.)
Gene Set Enrichment Analysis	Find pathways of genes that are “enriched” between two conditions
Clustering	Find subsets of a dataset (e.g., genes, samples) that are similar to one another in structure or function
Classification	Create a model that will predict the class of an unknown sample (e.g., relapse, non-relapse, etc.)
Sequence Variation analysis	Find chromosomal regions of similar copy number, call genotypes, etc.
Disease-specific analyses	E.g. genomic identification of significant targets in cancer (GISTIC), identify loss of homozygosity, etc.
Dimension Reduction	Transform your data into a lower number of dimensions to facilitate analysis

GenePattern vocabulary: Modules

Copy Number Divide by Normals	GSEA	Variation Filter
GISTIC	CBS	k-Nearest Neighbors
Classification and Regression Trees	Support Vector Machines	Hierarchical Clustering
GISTIC	Expression File Creator	Metagene Projection

GenePattern vocabulary: Modules



Hierarchical Clustering Files

HCL.jar
cluster.exe
ant.jar
gp-modules.jar
Jama-1.0.2.jar

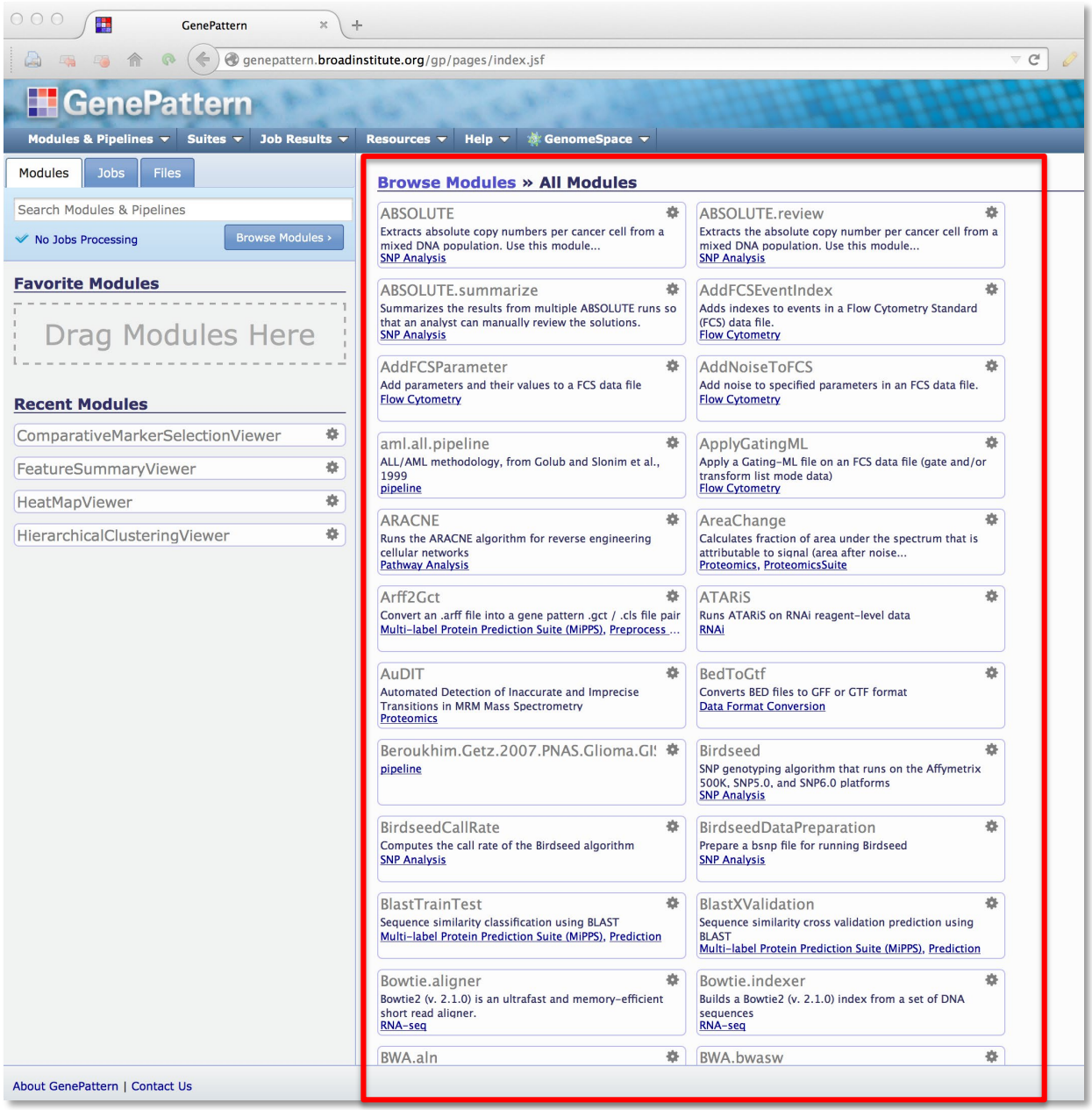
Documentation

HierarchicalClustering.pdf

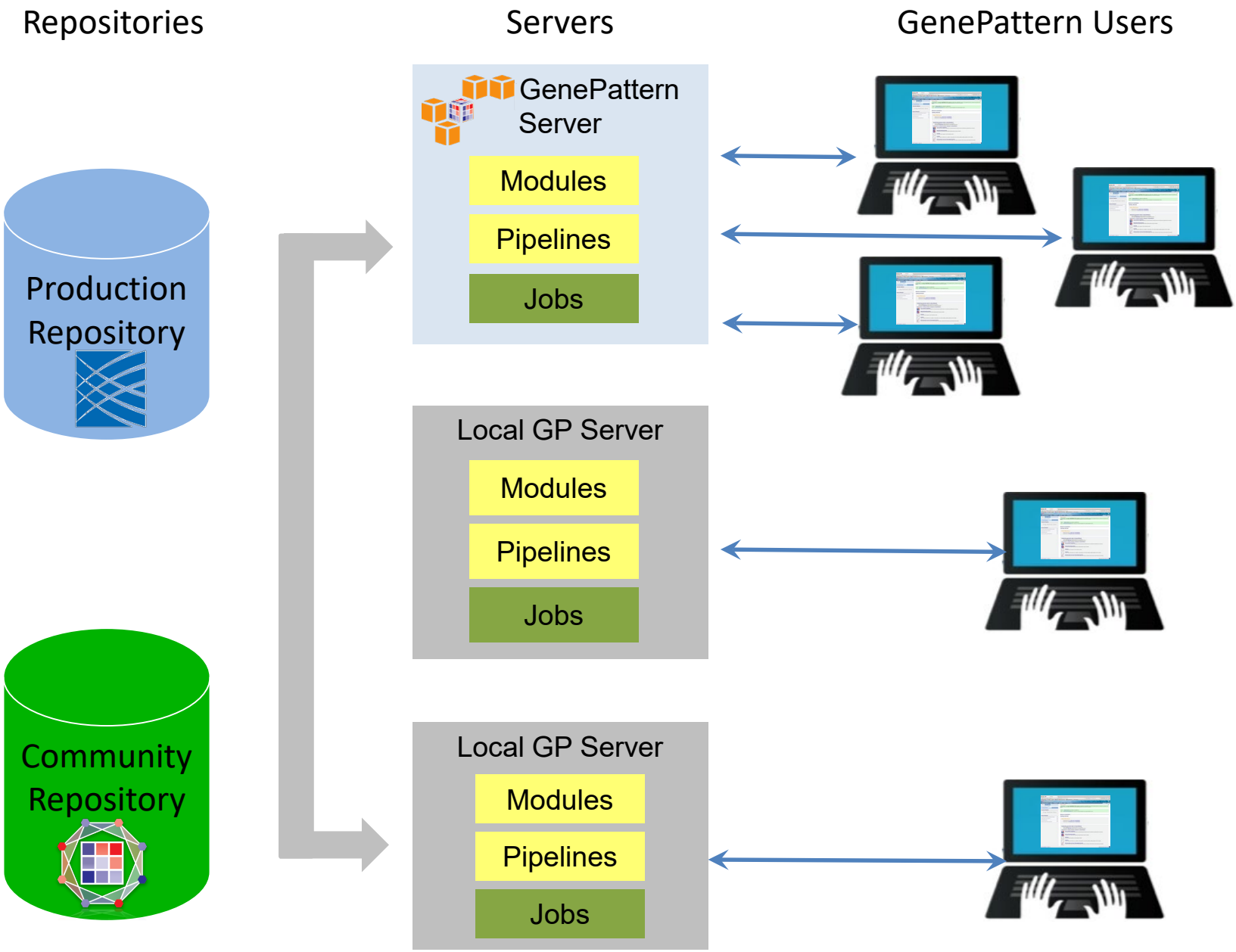
Parameter descriptions

-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>

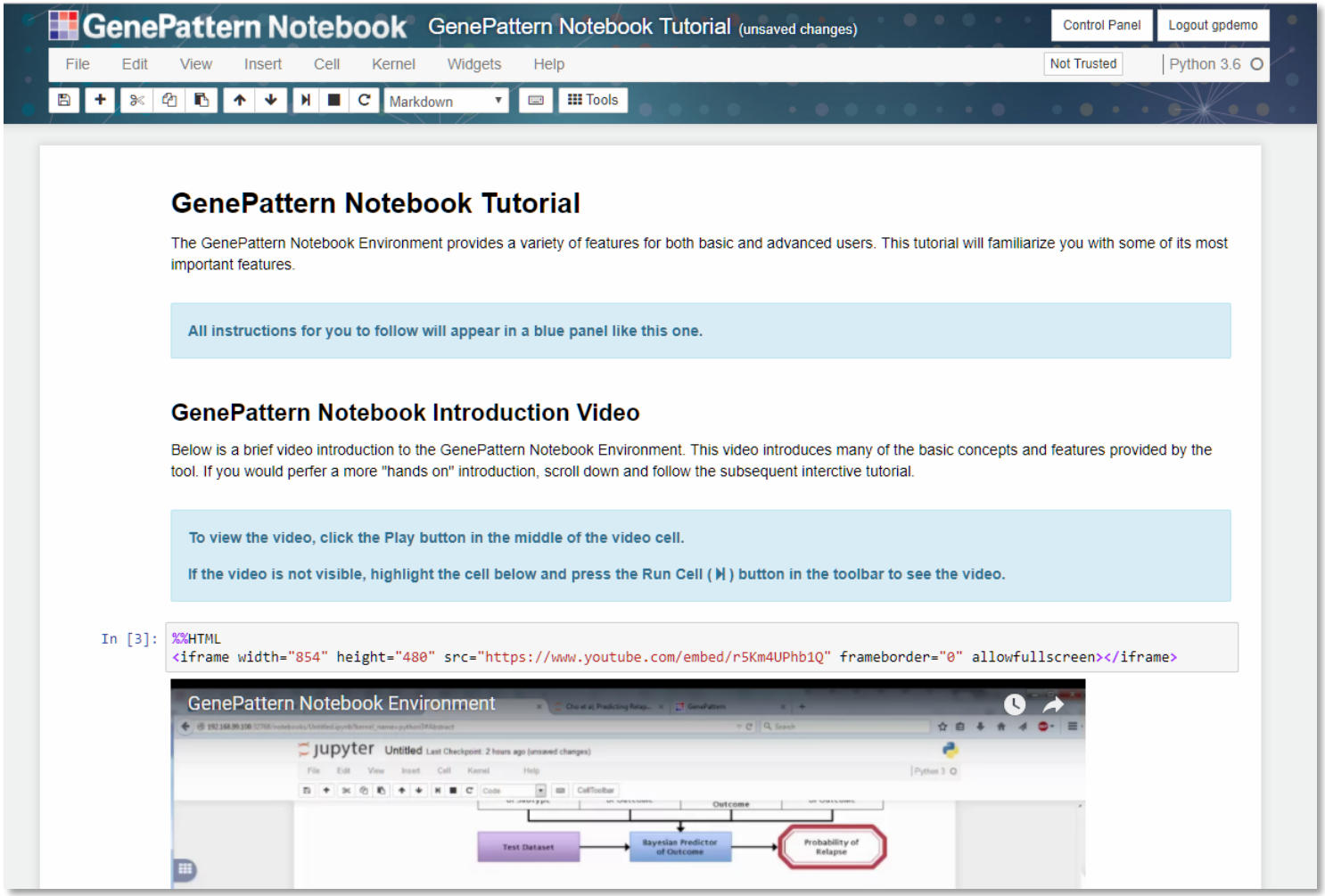
>250 GenePattern Modules, 4/2019



How GenePattern works

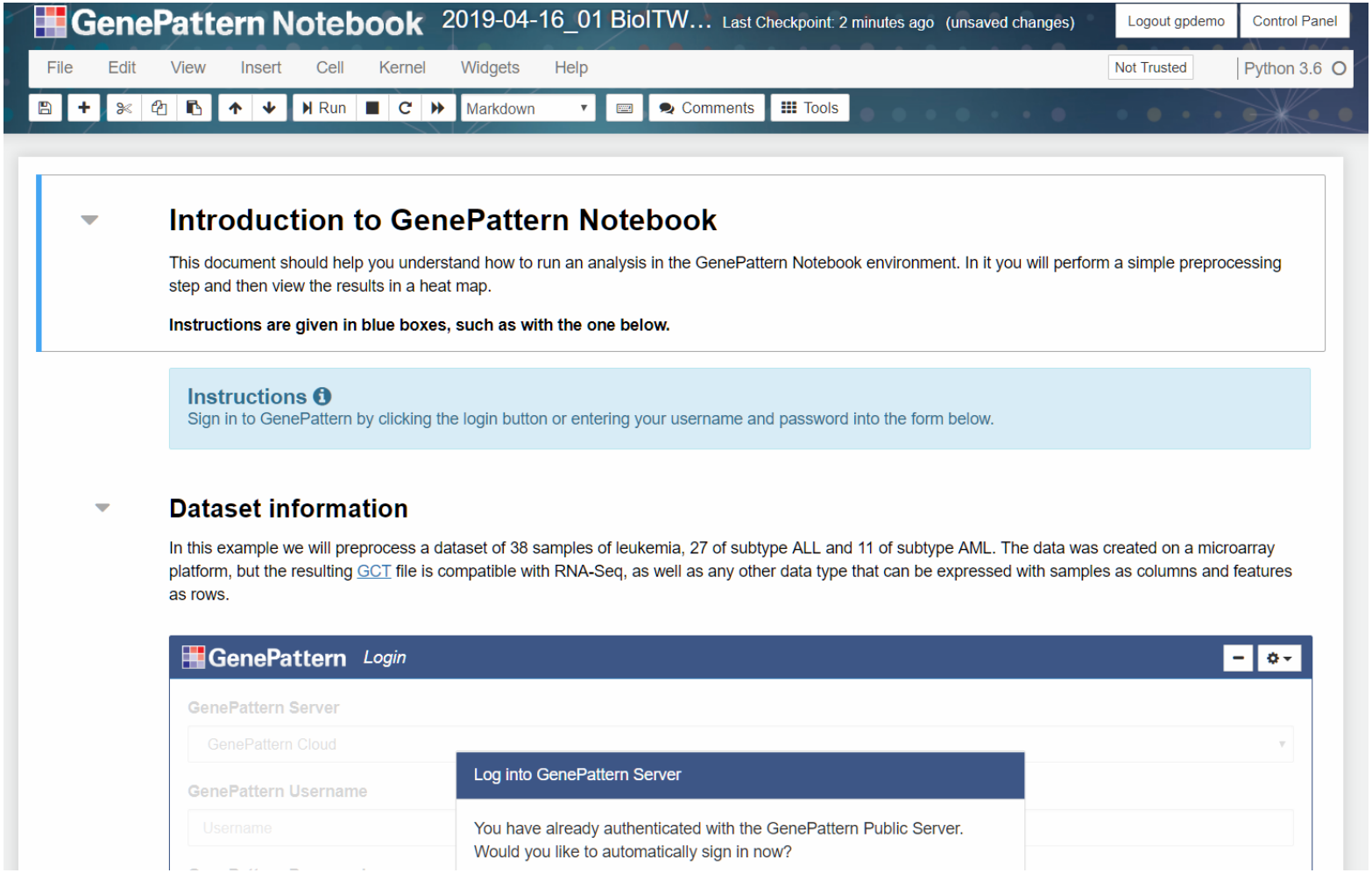


GenePattern Notebook Environment



Run an Analysis Notebook

2019-04-16_01 BioITWorld GenePattern Notebook Introduction



Jupyter Notebook

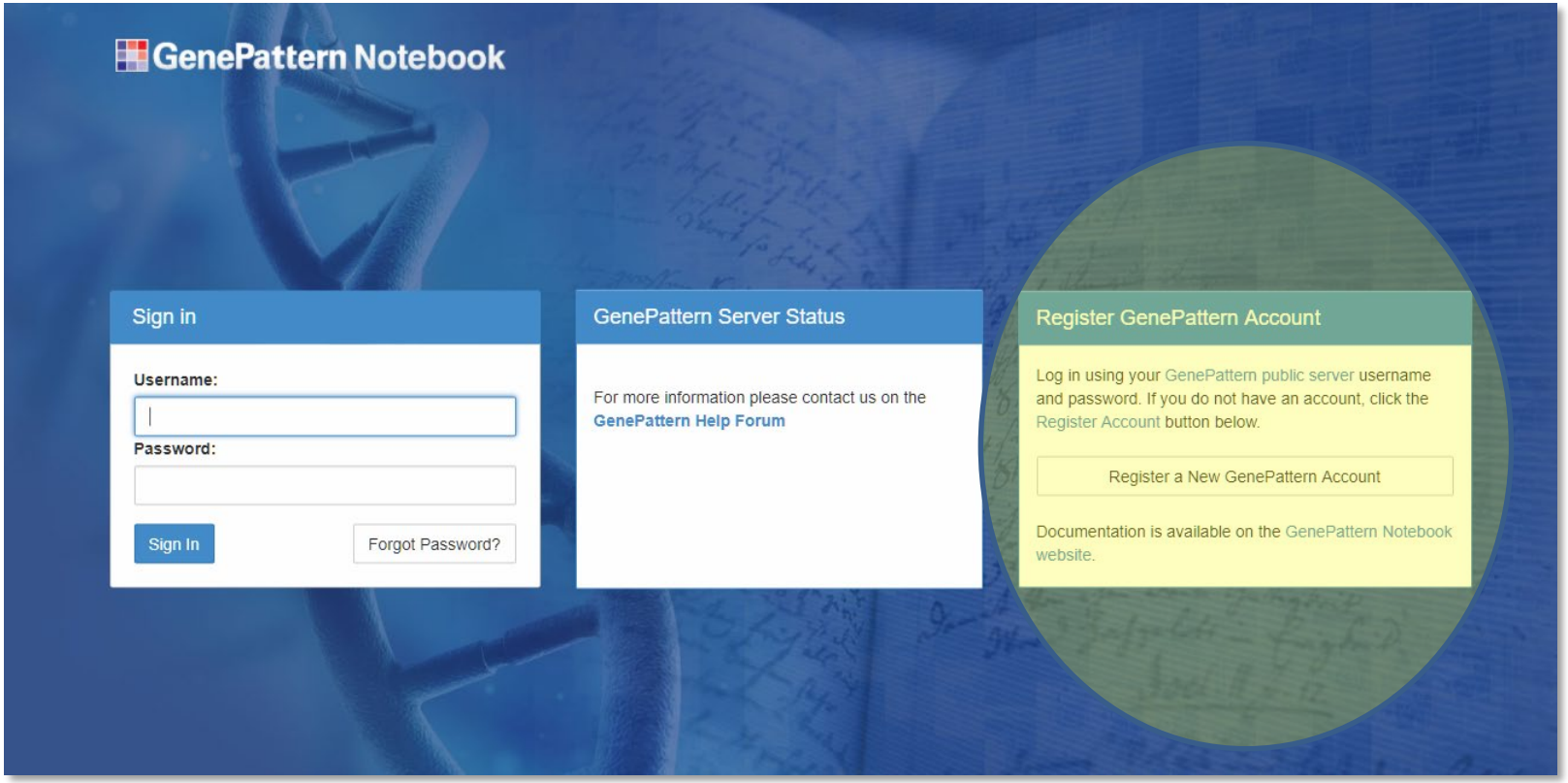
- Popular and well-supported framework for scientific computing
- Ecosystem of available extensions and resources
- Open source

Complete Research Narrative

- Leverages the best of Jupyter and GenePattern

- Interleave text, visualization, graphics and analytical aspects

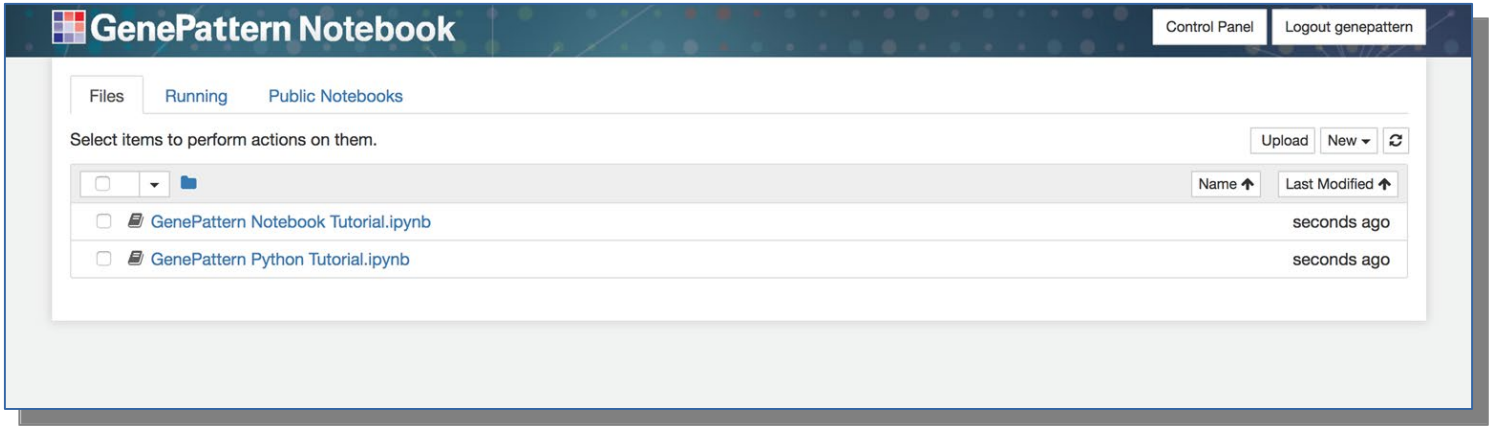
GenePattern Notebook Repository



https://notebook.genepattern.org

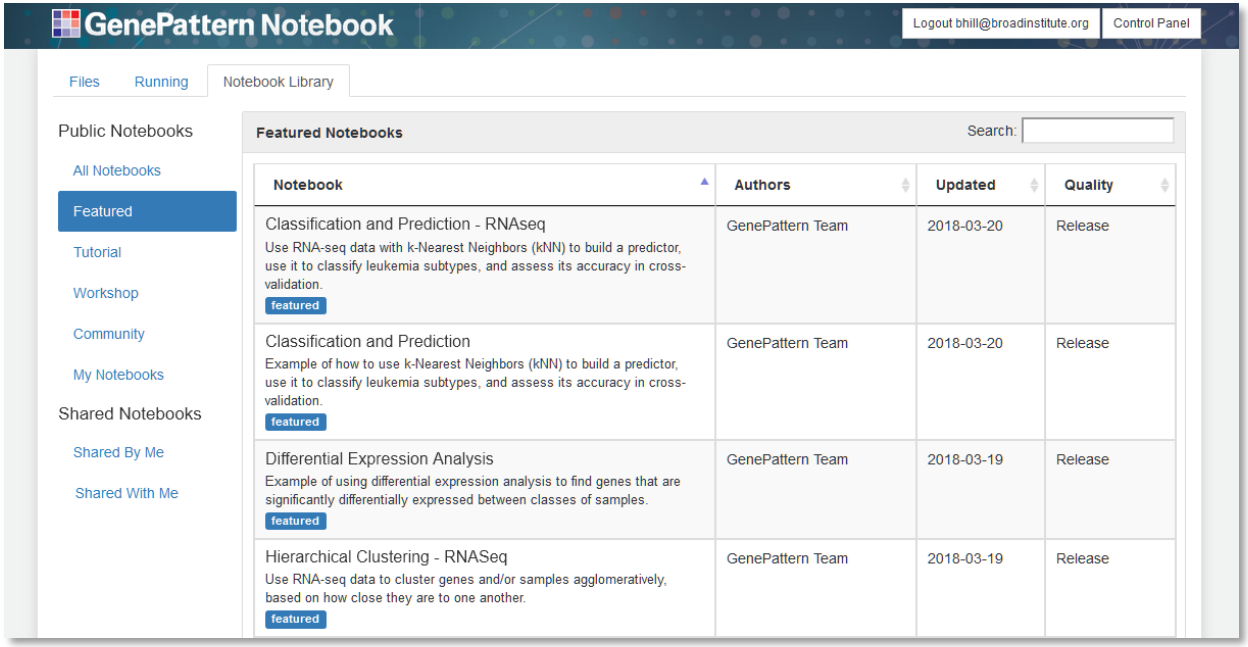
Notebook Workspace

- Lists your private notebooks and associated files.
- Copy, move, rename, delete, download and publish notebooks from here.



Browse Public Notebooks

- A variety of public notebooks are available in the GenePattern Notebook Library.
- Anyone can make a copy of these notebooks to read, run and reproduce.



Understanding Notebooks

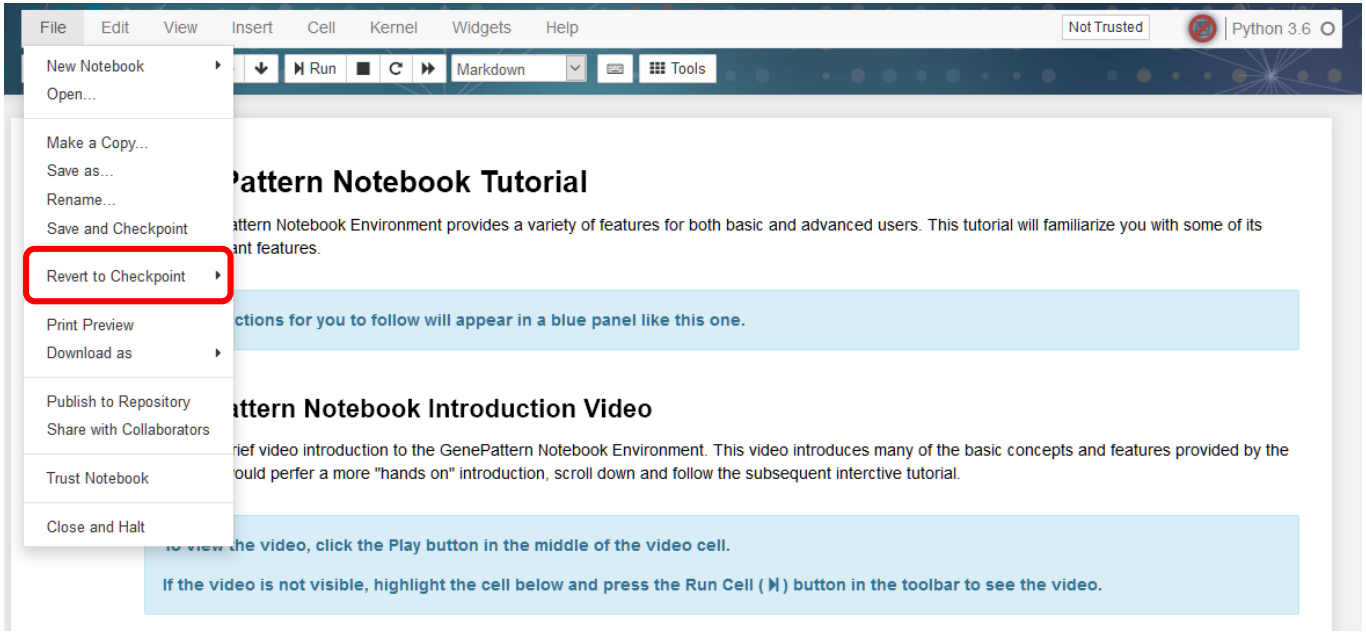
- Notebooks encapsulate a workflow, including analysis, documentation and other considerations, so that it can be easily reproduced.

- To achieve this, all notebooks are backed by a “kernel,” which is a complete contained computational environment.

- The kernel provides programmatic capabilities for users who want to code, and also allows for for interactive widgets for users who don’t want to code.

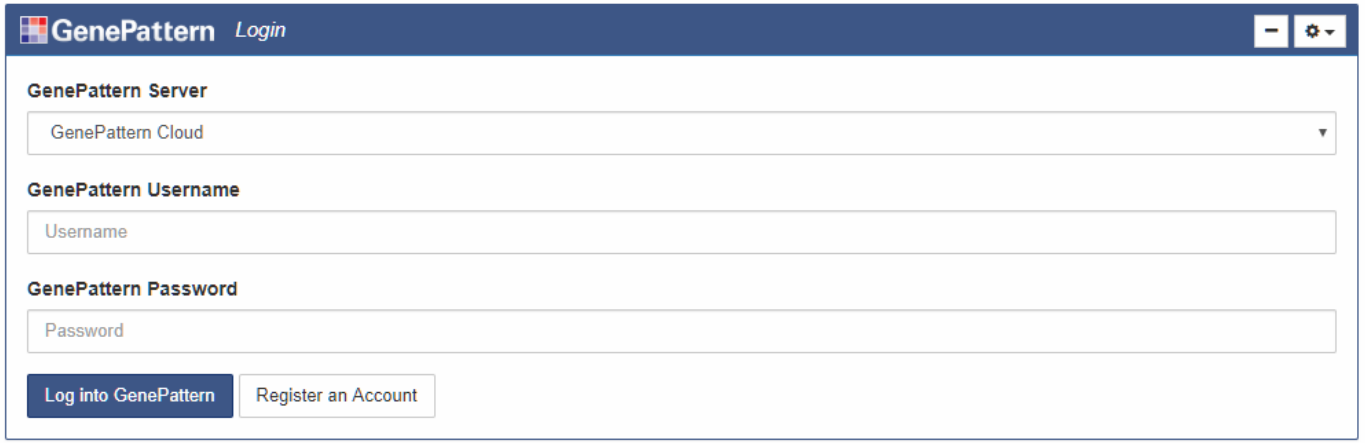
Name, Save & Checkpoint Notebooks

- Name or rename notebooks
- Save or revert to a checkpoint
- Make a duplicate notebook

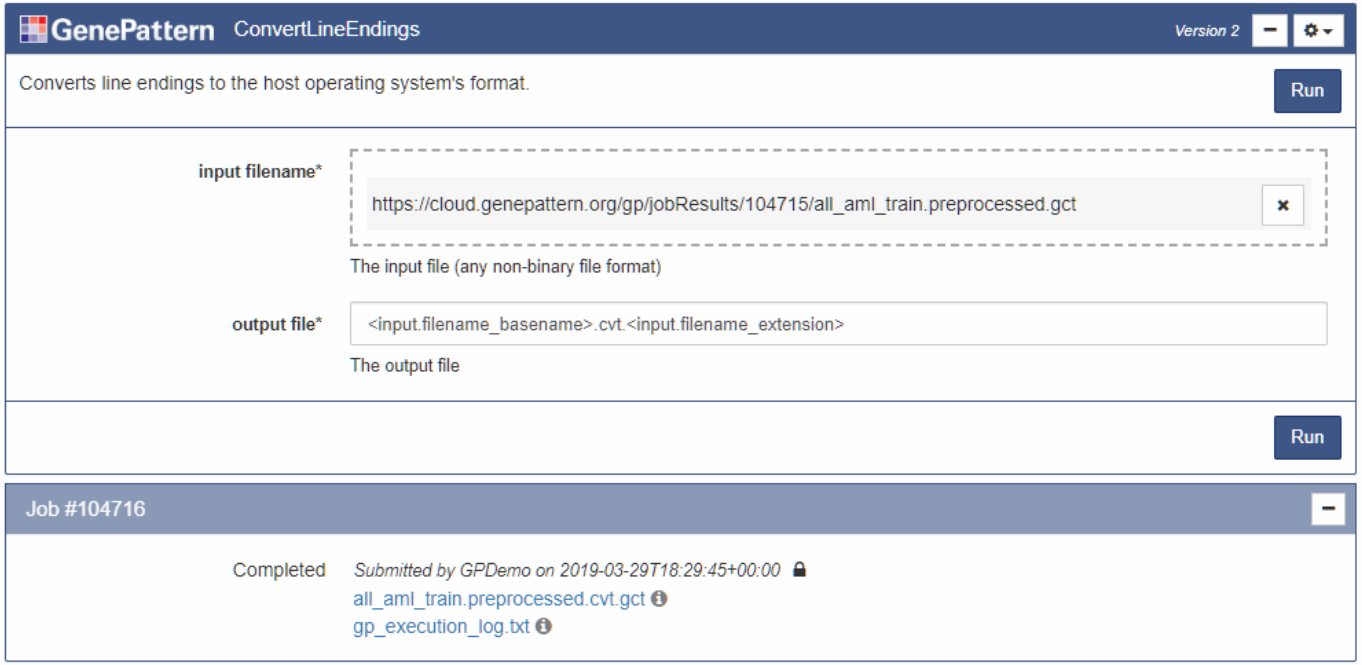


GenePattern Cells

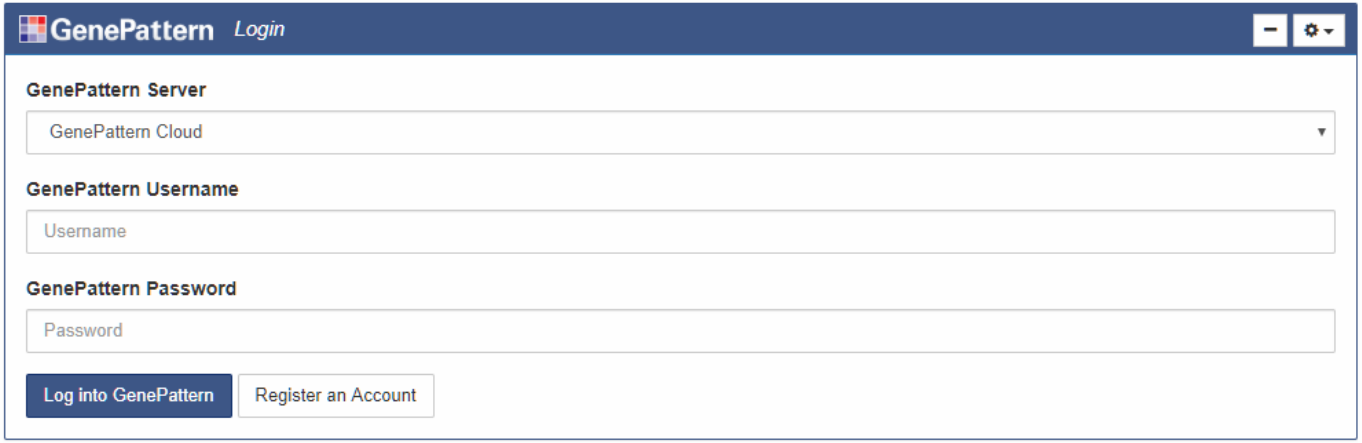
Authentication Cell

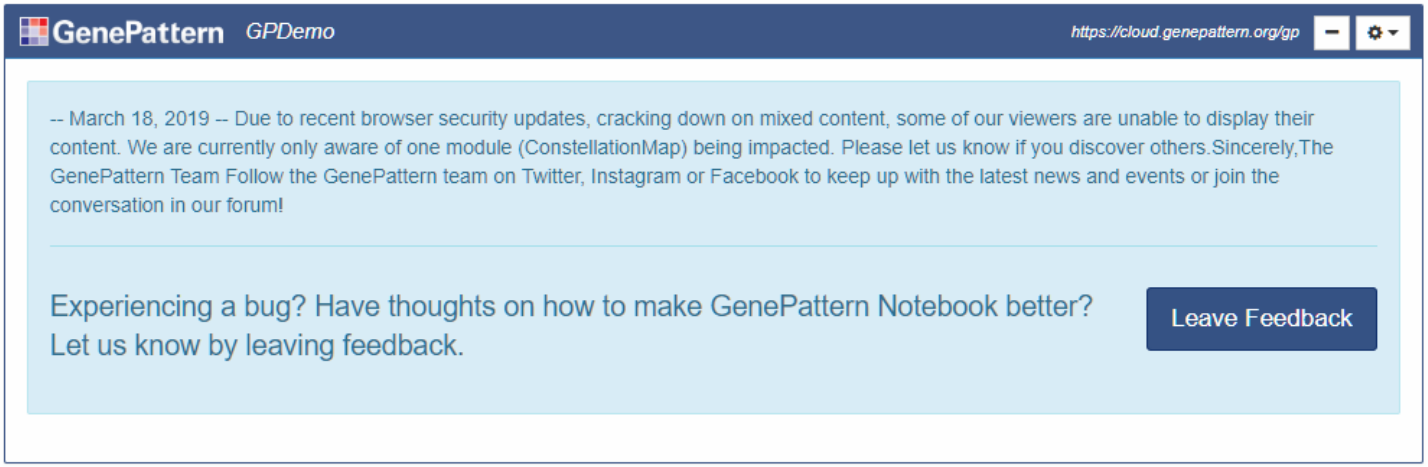


Analysis Cell

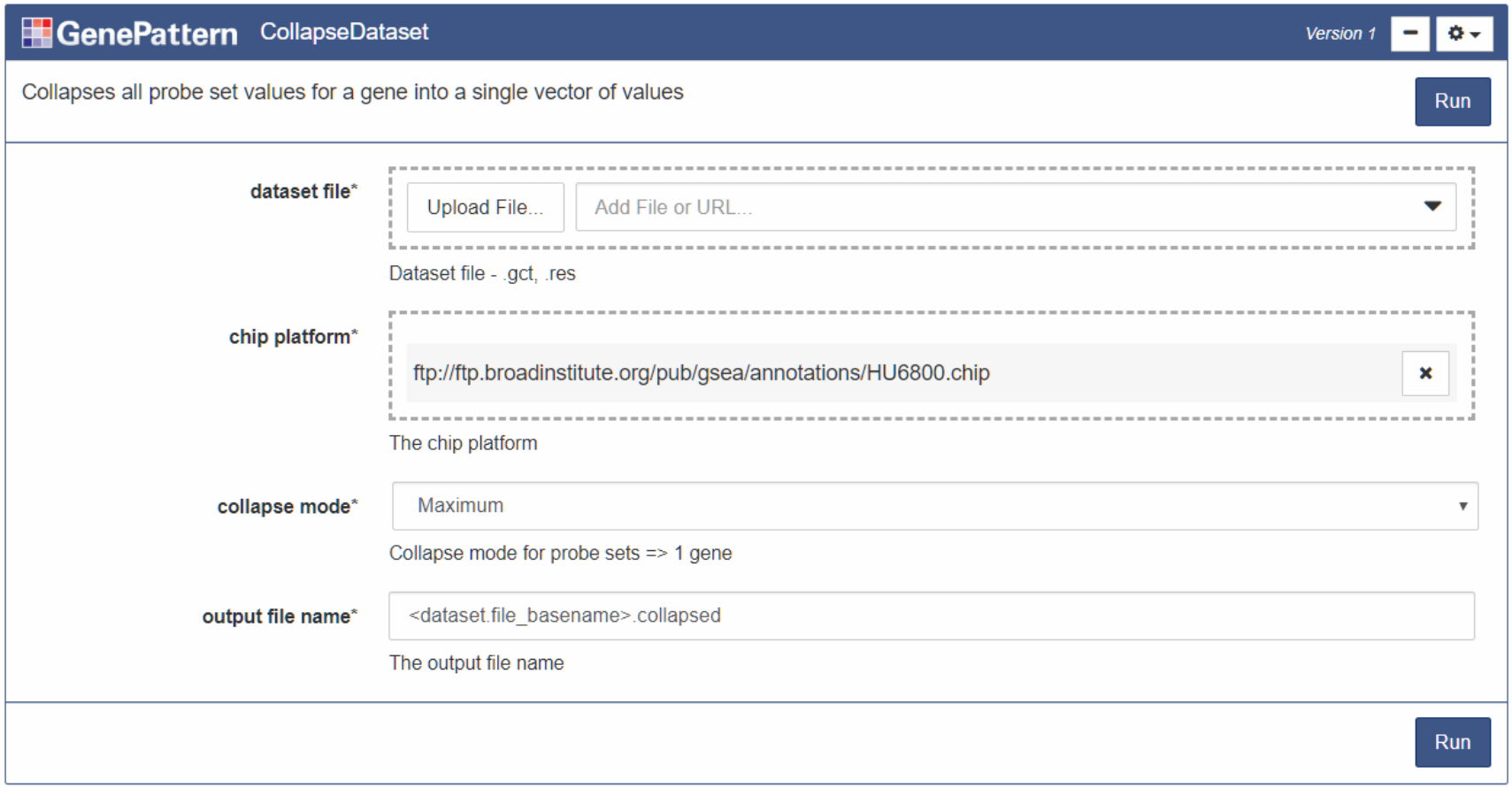


Authentication Cells

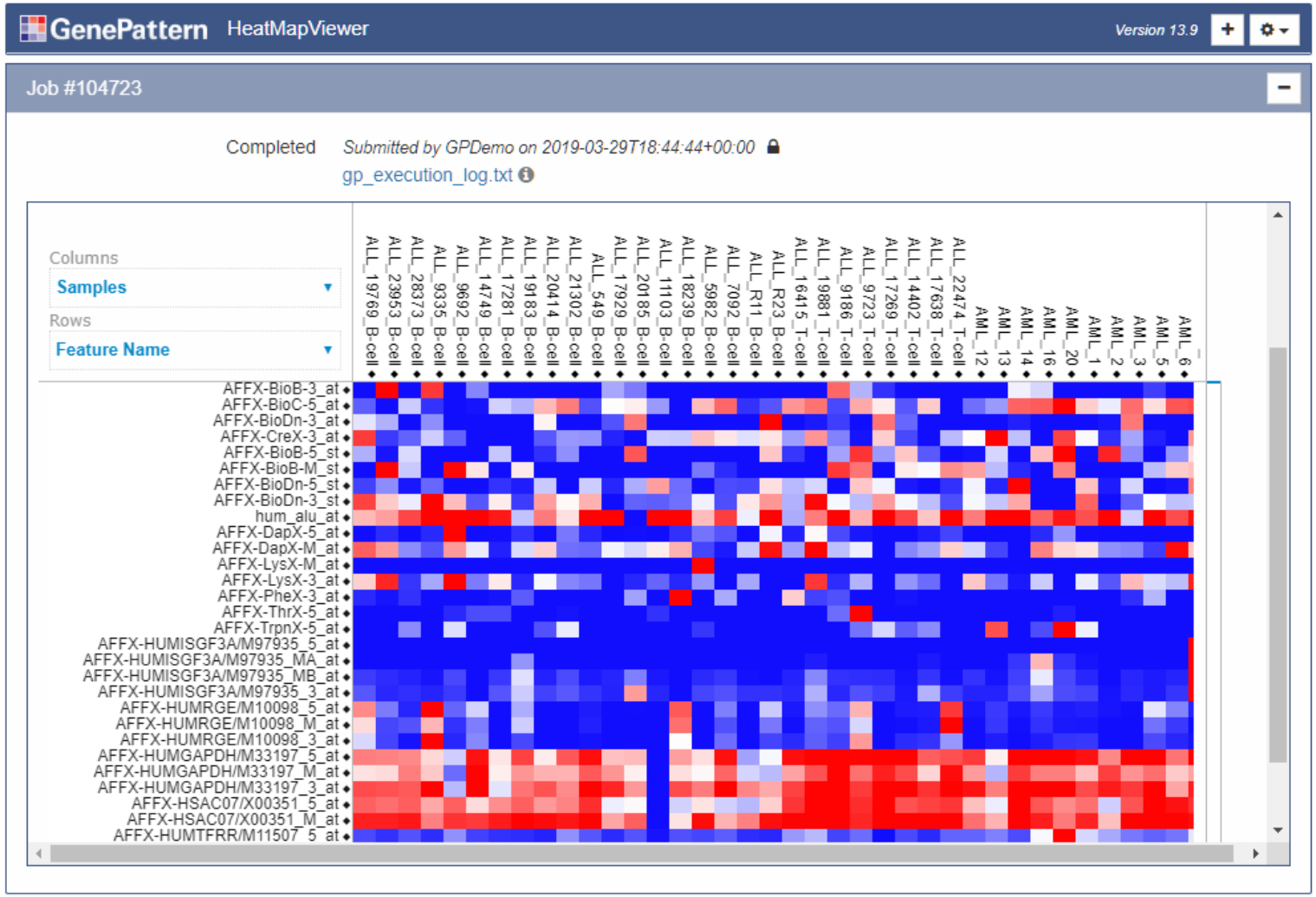




Analysis Cells



Job/Result Cells



Running GenePattern Notebook

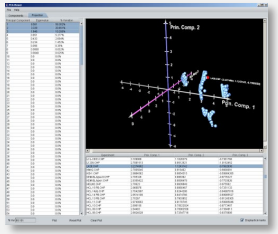
- Run using the GenePattern Notebook Repository.
- Install on your own computer by installing through the `pip` or `conda` package managers.
- A GenePattern Notebook Docker image is available.

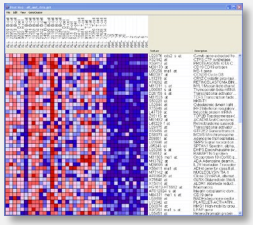
The GenePattern Ecosystem: Architecture

Module Repository

NMF | Cuffdiff | MuTect | ...

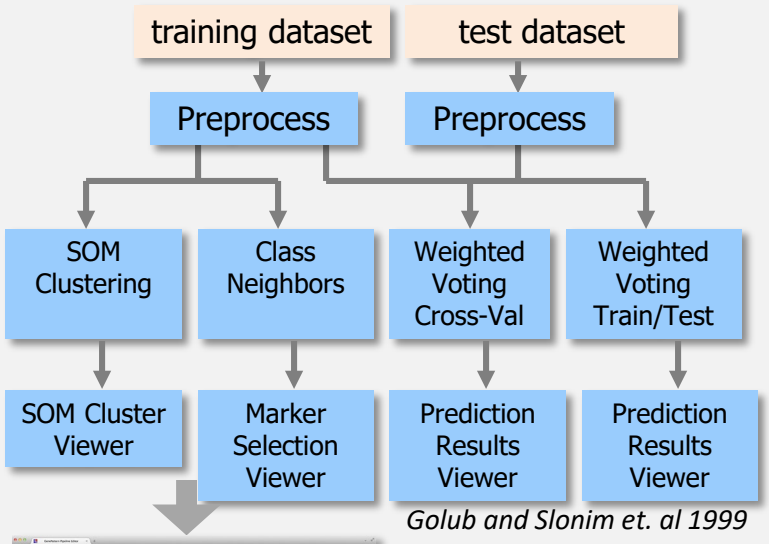
GISTIC | GSEA | Tophat | ...



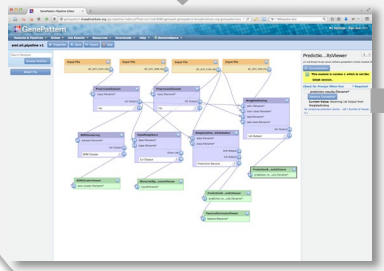


Hundreds of analysis and visualization tools

Pipeline Environment



Golub and Slonim et. al 1999



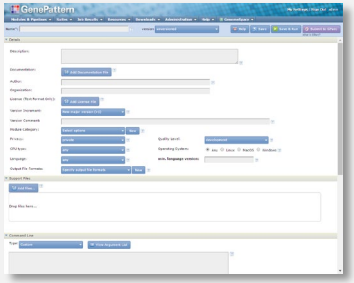
Golub.Slonim.1999.Nature.all.aml.pipeline.zip

Support for *in silico* reproducible research

Analysis Engine

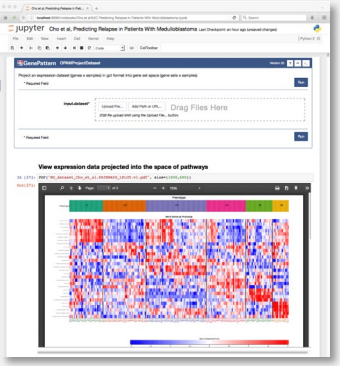
Record/replay analyses
Versioning of methods
Web service access

Module Integrator

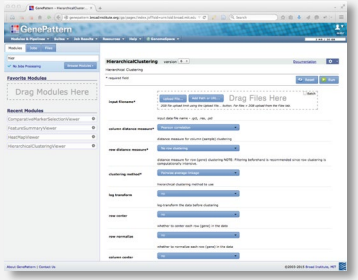


Easy addition of new tools

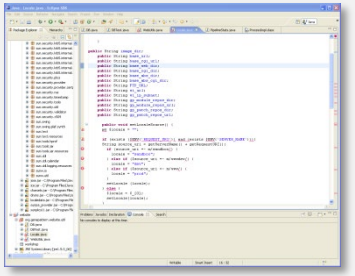
Clients



Notebook



Web



Programming

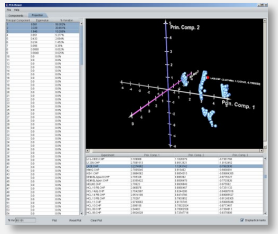
Access for all levels of user

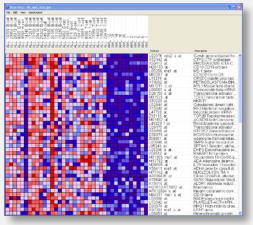
The GenePattern Ecosystem: Architecture

Tool Library

NMF | Cuffdiff | MuTect | ...

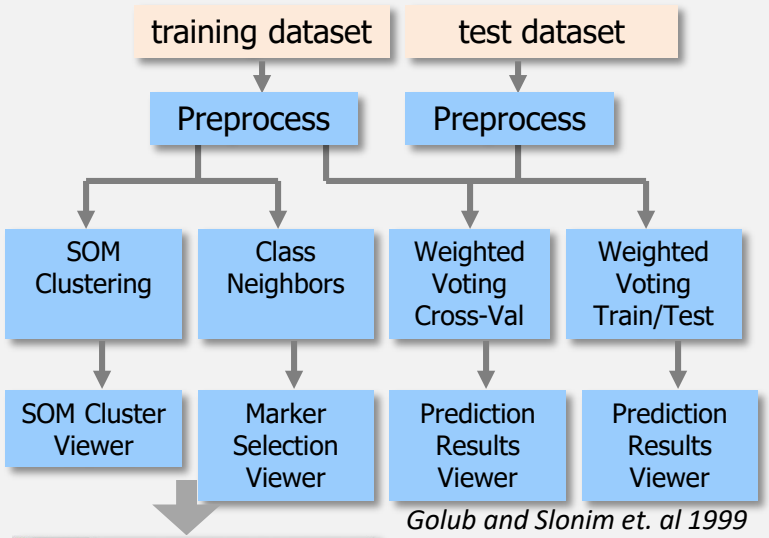
GISTIC | GSEA | Tophat | ...



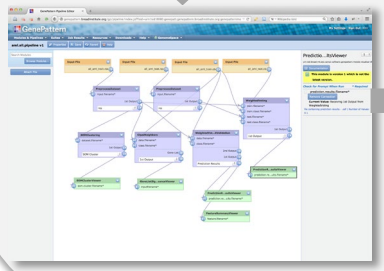


Hundreds of analysis and visualization tools

Workflows



Golub and Slonim et. al 1999



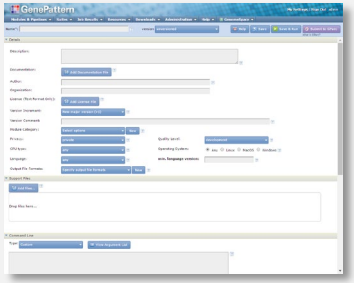
Golub.Slonim.1999.Nature.all.aml.pipeline.zip

Support for *in silico* reproducible research

Analysis Engine

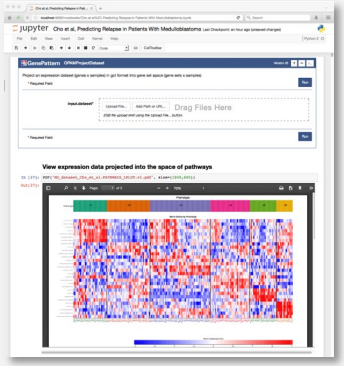
Record/replay analyses
Versioning of methods
Web service access

Create your own tool

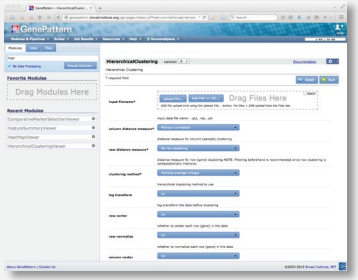


Easy addition of new tools

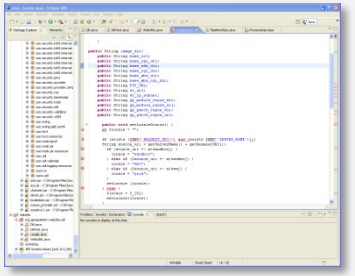
Access Points



Notebook



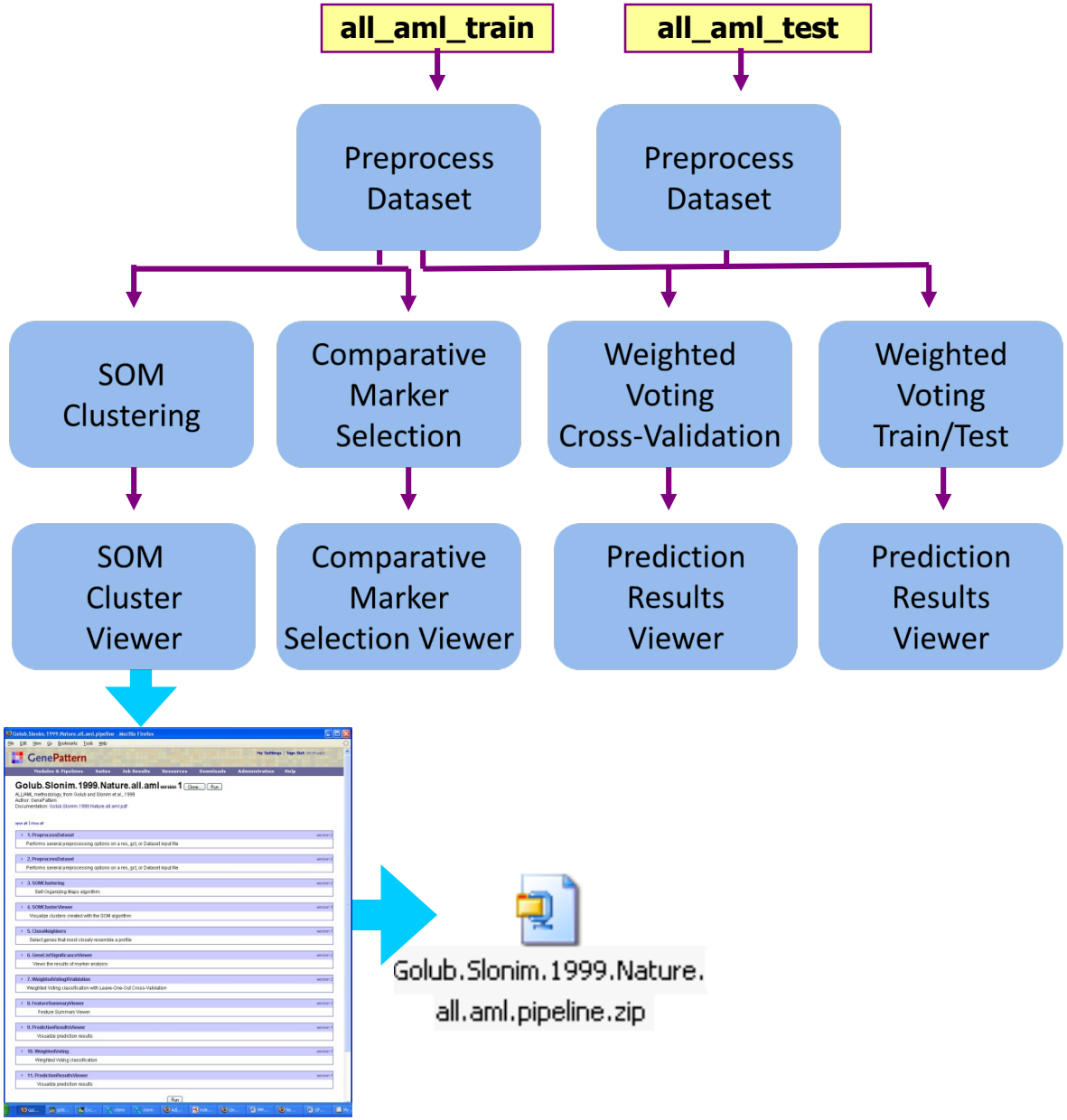
Web



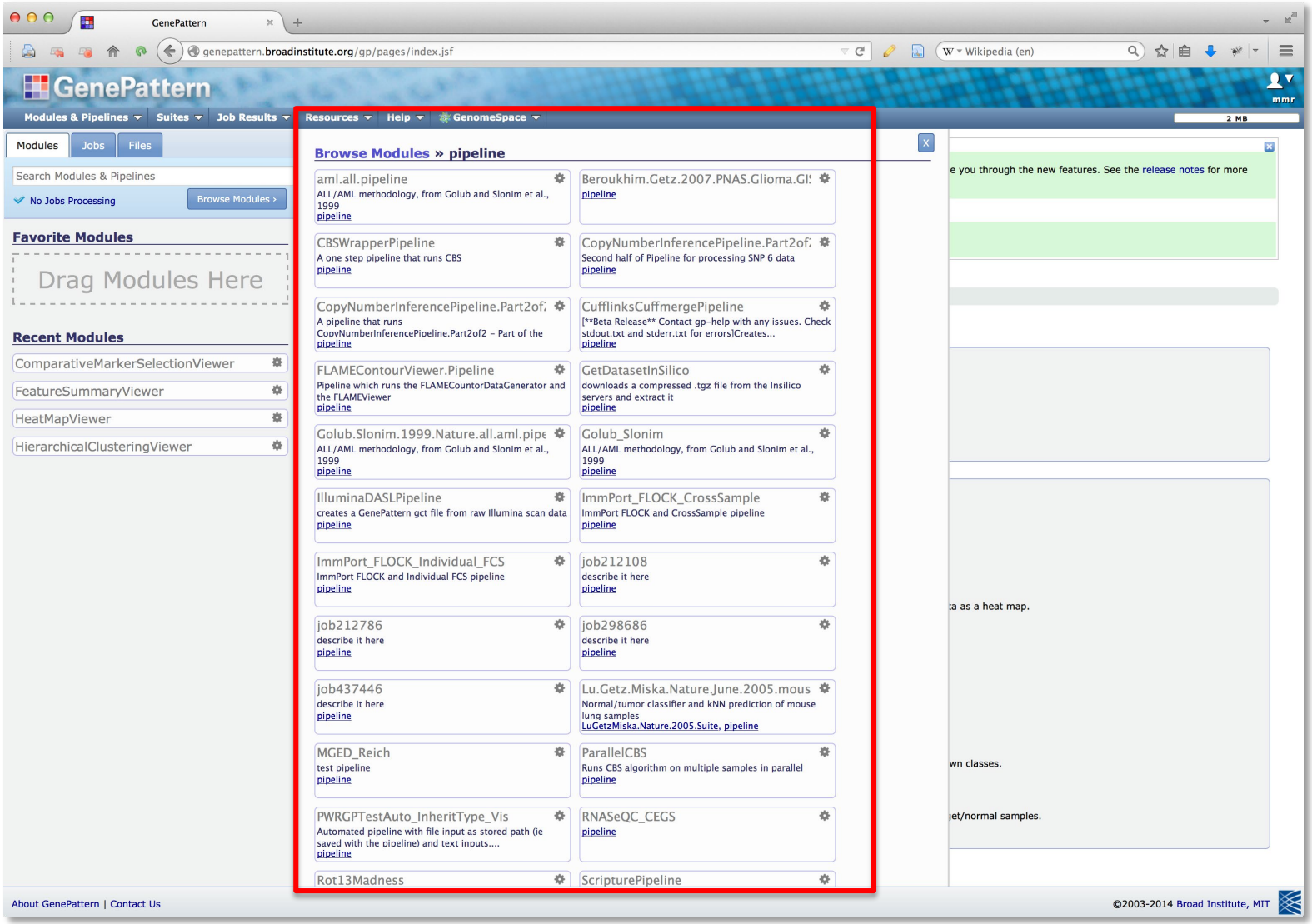
Programming

Access for all levels of user

GenePattern vocabulary: Pipelines

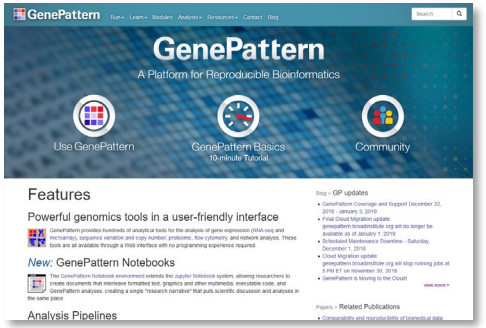


Pipelines in GenePattern



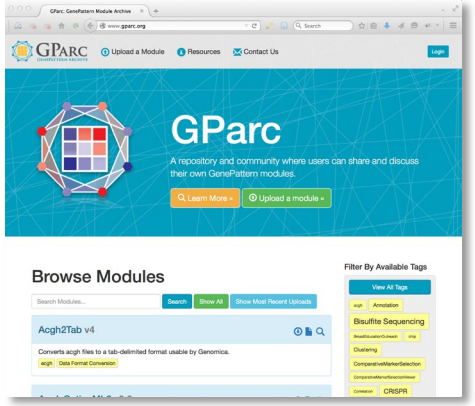
Community Activity

www.genepattern.org



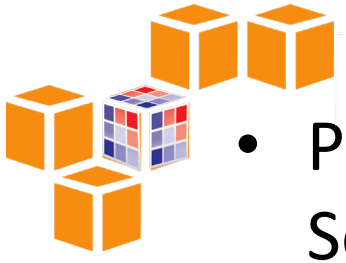
- Current version: 3.9.11 rc.4 b210 (2/2019)
- >50,000 registered users
- Open source, BSD-style license
- New Public server runs ~2,500 analyses/week

www.gparc.org

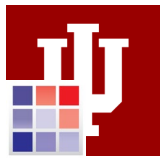


- GParc: GenePattern community repository
  - ~100 community-contributed methods
  - CRISPR analysis
  - Bisulfite sequencing
  - Flow cytometry
  - RNAi screens

Availability



- Public server running on Amazon Web Services (cloud.genepattern.org)



- Public server at Indiana U, backed by Carbonate HPC cluster (gp.indiana.edu)

- Downloadable server (www.genepattern.org)

- Amazon Machine Image (AMI)

Data Preparation

GenePattern file formats

- Each module specifies its input and output file formats
- Modules may use common GenePattern file formats or their own unique formats
- Most modules use the file formats described on the GenePattern website
- There are many GenePattern modules that are specifically meant for importing data from other systems

Example modules for data formatting

Data type	Input Format	Module
RNA-Seq	bed fpkm tracking Read group tracking Cufflinks Expr bed	BedToGtf Fpkm_trackingToGct Read_group_trackingToGct ExprToGct BedToGtf
DNA-seq	bam sam sam/bam	Picard.BamToSam Picard.SamToBam Picard.SamToFastQ
Gene expression chips	.cel	ExpressionFileCreator
SNP chips	.cel	SNPFileCreator
MAGE	Mage-ML Mage-tab	MageMLImportViewer MageTabImportViewer

- To find, search modules for “to<format>” or “<format>To” or 2<format> or <format>2
- There are many more modules, for MS, FCS and other data types

GCT file format

Always #1.2

Number of rows (ie gene ids)

Number of samples

- Tab Delimited
- Open in any spreadsheet or text editor

Col 1: Row identifiers. Typically gene ids.

Col 2: Optional description

Col 3+: Sample names, must be unique.

Each column contains expression or count values from one sample.

	A	B	C	D	E	F	G	H	I
1	#1.2								
2	60483	30							
3	Name	Description	A7-A13F-primar	A7-A13F-normal	A7-A13G-primary	A7-A13G-normal	AC-A23H-primary	AC-A23H-normal	AC-A2FB-primary
4	ENSG000000	ENSG000000	2904	3276	1009	4611	614	7362	4546
5	ENSG000000	ENSG000000	2	2159	65	869	0	234	10
6	ENSG000000	ENSG000000	6070	1295	547	1896	4011	1813	1663
7	ENSG000000	ENSG000000	1263	1178	3914	1262	4928	1684	2451
8	ENSG000000	ENSG000000	1362	277	799	256	954	390	825
9	ENSG000000	ENSG000000	330	480	289	555	239	471	905
10	ENSG000000	ENSG000000	1090	9090	1790	9020	3629	13575	3762
11	ENSG000000	ENSG000000	3897	2651	1391	3555	5525	4534	3600
12	ENSG000000	ENSG000000	2248	2249	1493	6008	2352	2856	3482
13	ENSG000000	ENSG000000	2735	1934	2642	1975	4098	2814	2922
14	ENSG000000	ENSG000000	284	636	799	631	907	774	925
15	ENSG000000	ENSG000000	597	3294	5708	1877	2589	5229	5003
16	ENSG000000	ENSG000000	4201	2010	986	1911	4783	2391	2921
17	ENSG000000	ENSG000000	4471	1864	1355	3751	1669	1636	2813
18	ENSG000000	ENSG000000	10532	2118	2470	1666	6469	1877	8759
19	ENSG000000	ENSG000000	1	30	33	38	28	80	22
20	ENSG000000	ENSG000000	5099	3305	2760	5474	7075	4630	5803
21	ENSG000000	ENSG000000	481	107	52	105	1163	146	370

Sample labels: CLS file

- 30 samples
- 2 Classes
- Numeric values, start with 0
- Tab or space delimited

```
30      2       1
#primary tumor  matched normal
0       1       0       1       0       1       0       1       0       1
```

	A	B	C	D	E
1	#1.2				
2	60483	30			
3	Name	Description	A7-A13F-primary	A7-A13F-normal	A7-A13G-primary
4	ENSG00000000003.13	ENSG00000000003.13	2904	3276	1009
5	ENSG00000000005.5	ENSG00000000005.5	2	2159	65

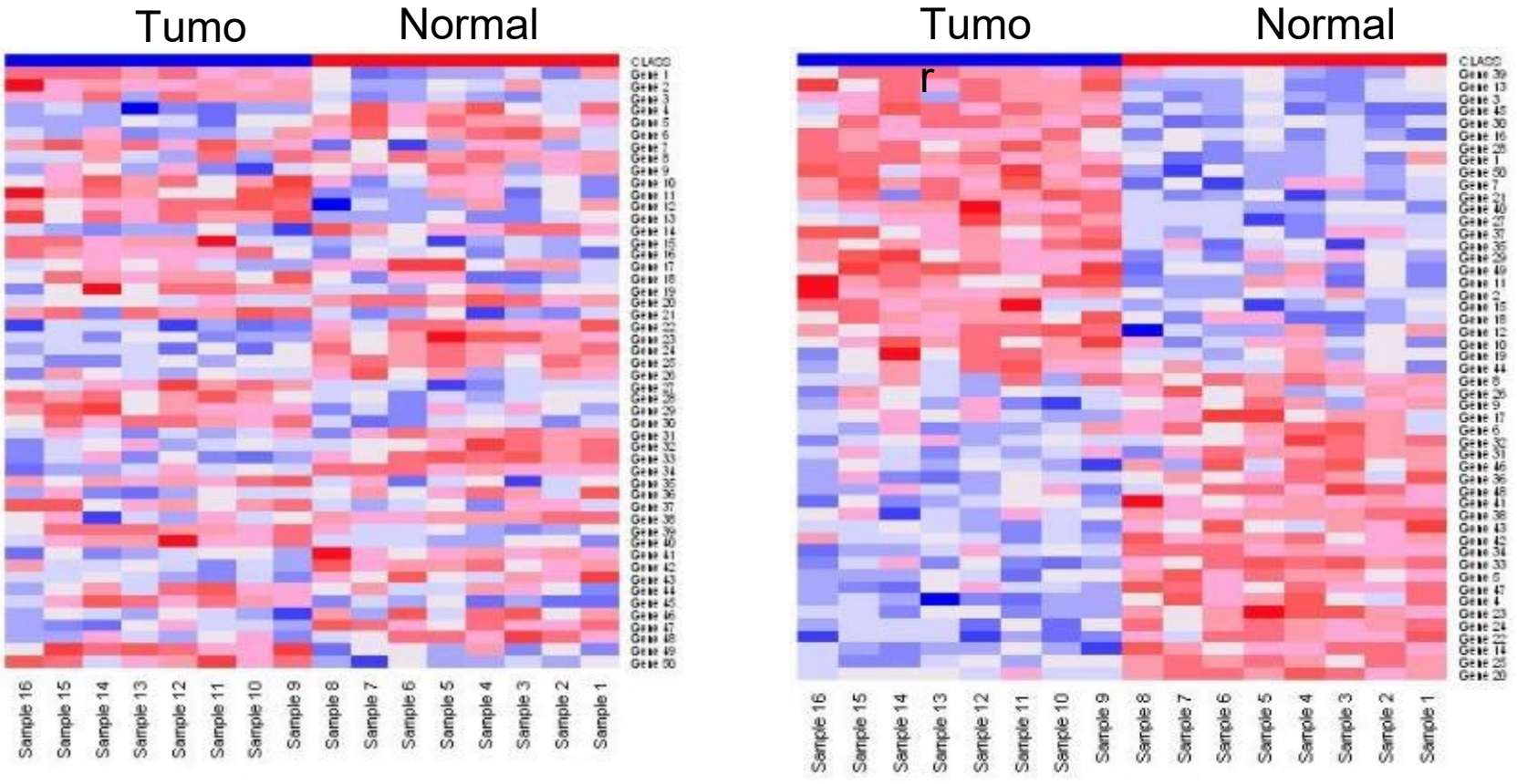
ClsFileCreator module can be used to generate a CLS file from a gct file.

Differential Gene Expression

Differential Expression Analysis

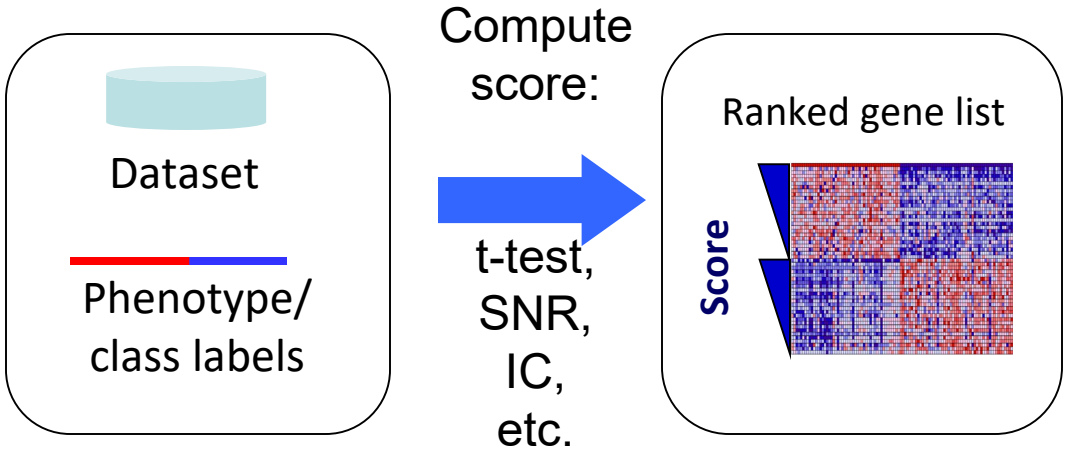
Marker selection

Given phenotypically distinct classes, find “markers” that distinguish these classes from one another



Gene Marker Selection

Compute score for each gene



μ = class mean
σ = std deviation
n = # of samples

t-test

Hypothesis testing method:
It is the **difference between the mean expression** of class A and class B **divided by the variability of expression**.

$$\frac{\mu_A - \mu_B}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}}$$

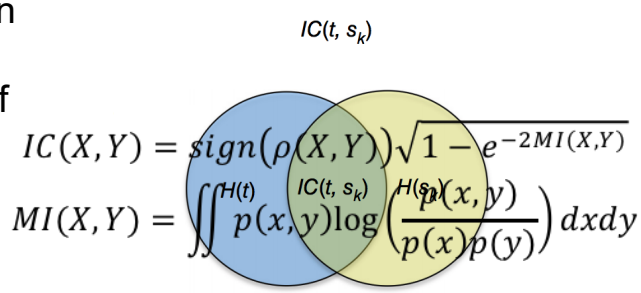
Signal-to-Noise Ratio (SNR)

Similar to the t-test but **takes the standard deviation of the two distributions into account** which is more representative of the differences between classes when there may be differences between the SD of class A and the SD of class B.

$$\frac{\mu_A - \mu_B}{\sigma_A + \sigma_B}$$

Information coefficient (IC)

This test takes **the amount of shared information** between the two classes.



Differential Analysis Exercise

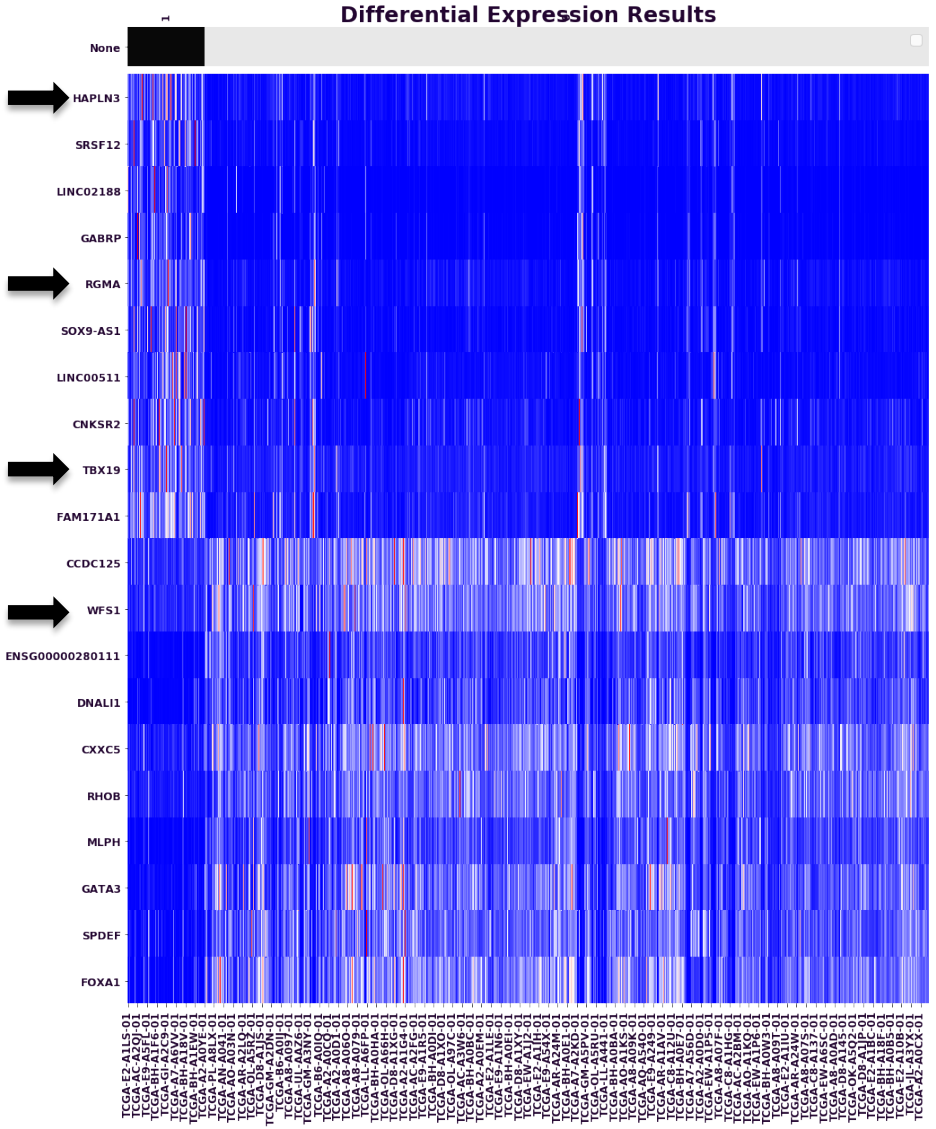
Open notebook:

BioITWorld Differential Analysis

Gene Set Enrichment Analysis (GSEA)

GSEA in a Nutshell

- GSEA answers the questions "**what are the gene sets that are significantly enriched** in my data?" and "**How enriched is my set of interest**?"
  - E.g., are the genes mentioned in a paper enriched as a set in a particular phenotype?



GSEA Outline

- Take gene expression data from two different groups and **rank all genes according to the differential expression** across the groups.
- Take a predefined **group of genes and determine whether they are differentially expressed as a set** => called enrichment.
- **Randomly swap** the gene-set or phenotype **labels** of the data and **repeat the test many times** as a gauge of **significance** of the enrichment analysis **results**.

What Is a Gene Set?

- A gene set is any group of genes, e.g.:
  - A pathway
  - A network
  - A list of over/under expressed genes
  - A group of genes from the same chromosome
- Order of genes in set is irrelevant
- Relationships between genes are not recorded
- *Wouldn't it be great if someone had created relevant collections of gene sets?*

MSigDB Gene Set Databases

http://www.msigdb.org

H	**hallmark gene sets** are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.
C1	**positional gene sets** for each human chromosome and cytogenetic band.
C2	**curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.
C3	**motif gene sets** based on conserved cis-regulatory motifs from a comparative analysis of the human, mouse, rat, and dog genomes.
C4	**computational gene sets** defined by mining large collections of cancer-oriented microarray data.
C5	**GO gene sets** consist of genes annotated by the same GO terms.
C6	**oncogenic signatures** defined directly from microarray gene expression data from cancer gene perturbations.
C7	**immunologic signatures** defined directly from microarray gene expression data from immunologic studies.

What if I only have few samples?

Or if I want to project my data onto the pathway space?

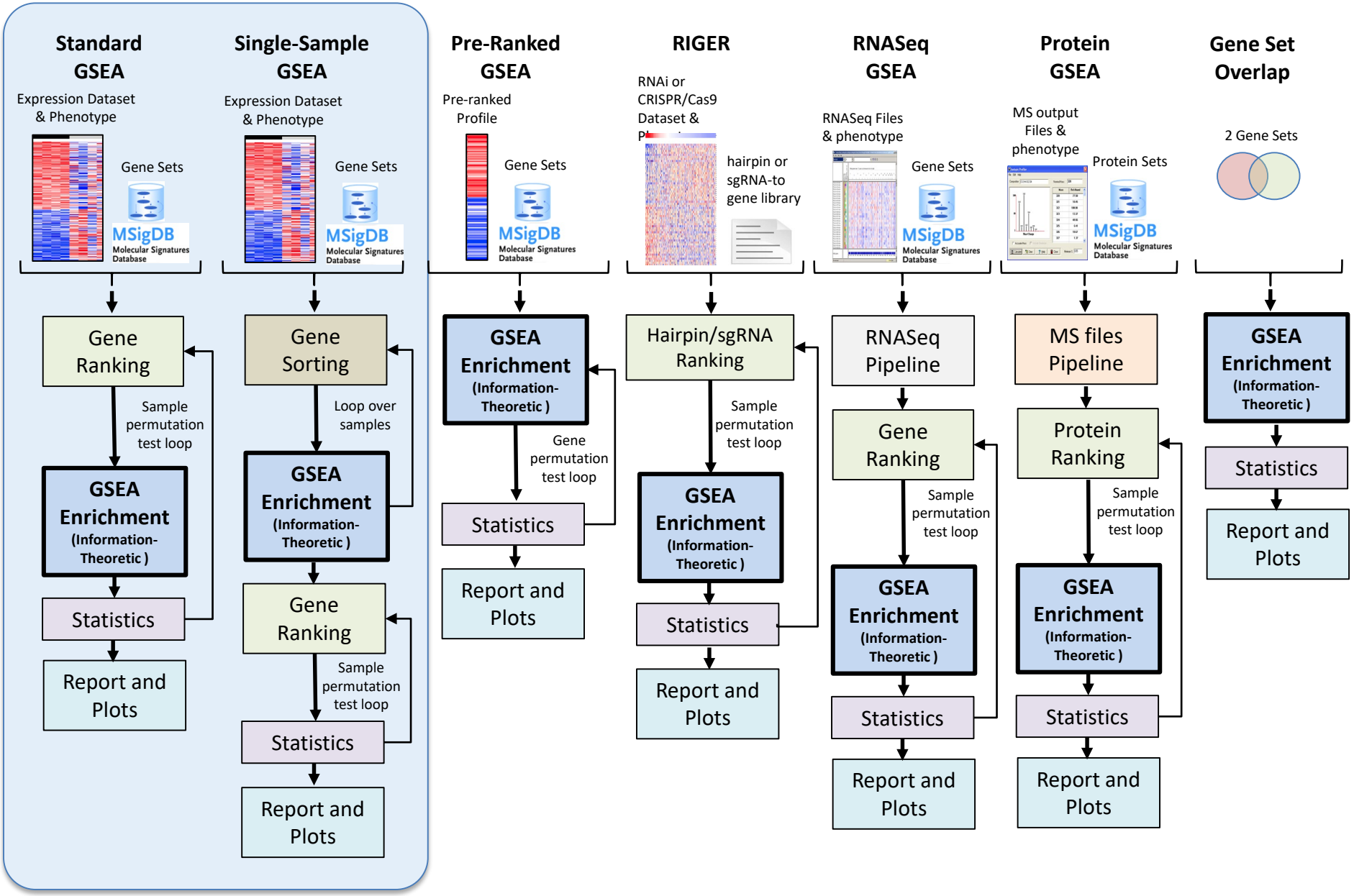
Single Sample GSEA (“projection”)

Single-Sample GSEA (ssGSEA)

- Projects gene expression onto pathways
- Dataset becomes pathways x samples
- Works on one sample at a time
- Any gene set can be used as a pathway
- Does not require predefined classes

Different Modalities of GSEA

More details in : http://software.broadinstitute.org/gsea or msigdb.org



ssGSEA Exercise

2019-04-16_04 BioITWorld ssGSEA

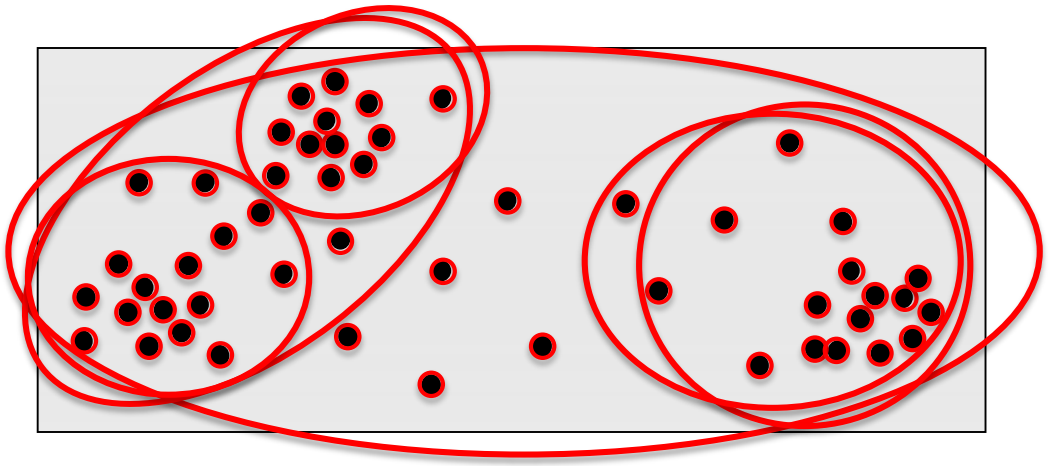
Clustering

Clustering/Class Discovery

- Aim: Partition data (e.g. genes or samples) into sub-groups (clusters), such that points of the same cluster are “more similar”.

- Example:
  How many clusters?

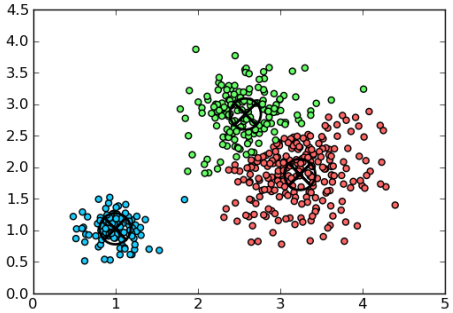
- One has to choose:
  - Clustering method
  - Similarity/distance measure
  - Evaluate clusters



Clustering in GenePattern

- Representative based:

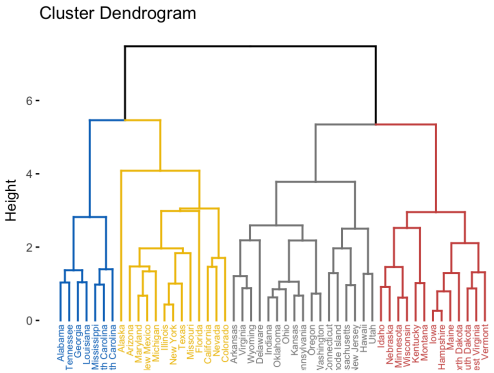
  Find representatives/centroids of the dataset
  - *K*-means
  - Self Organizing Maps (SOM)



K-Means Clustering

- Bottom-up (Agglomerative)

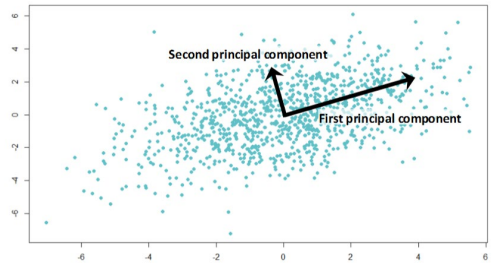
  Create an ordering of the data by closeness
  - Hierarchical clustering



Hierarchical Clustering

- Clustering-like:

  Reduce the data to a smaller number of dimensions containing the majority of the information content
  - NMF (Non-Negative Matrix Factorization)
  - PCA (Principal Components Analysis)



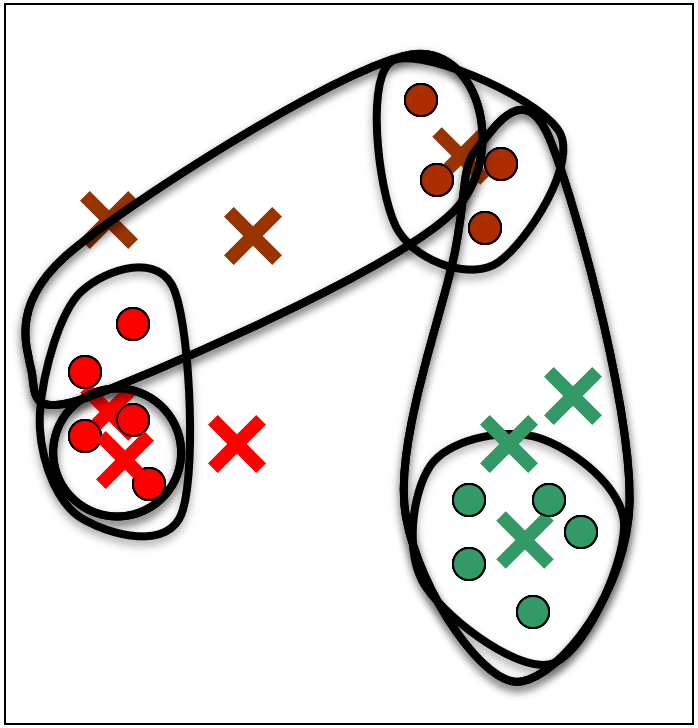
Principal Components Analysis

K-means Clustering

- Initialize centroids at random positions
- Iterate:
  - Assign each data point to its closest centroid
  - Move centroids to center of assigned points
- Stop when converged

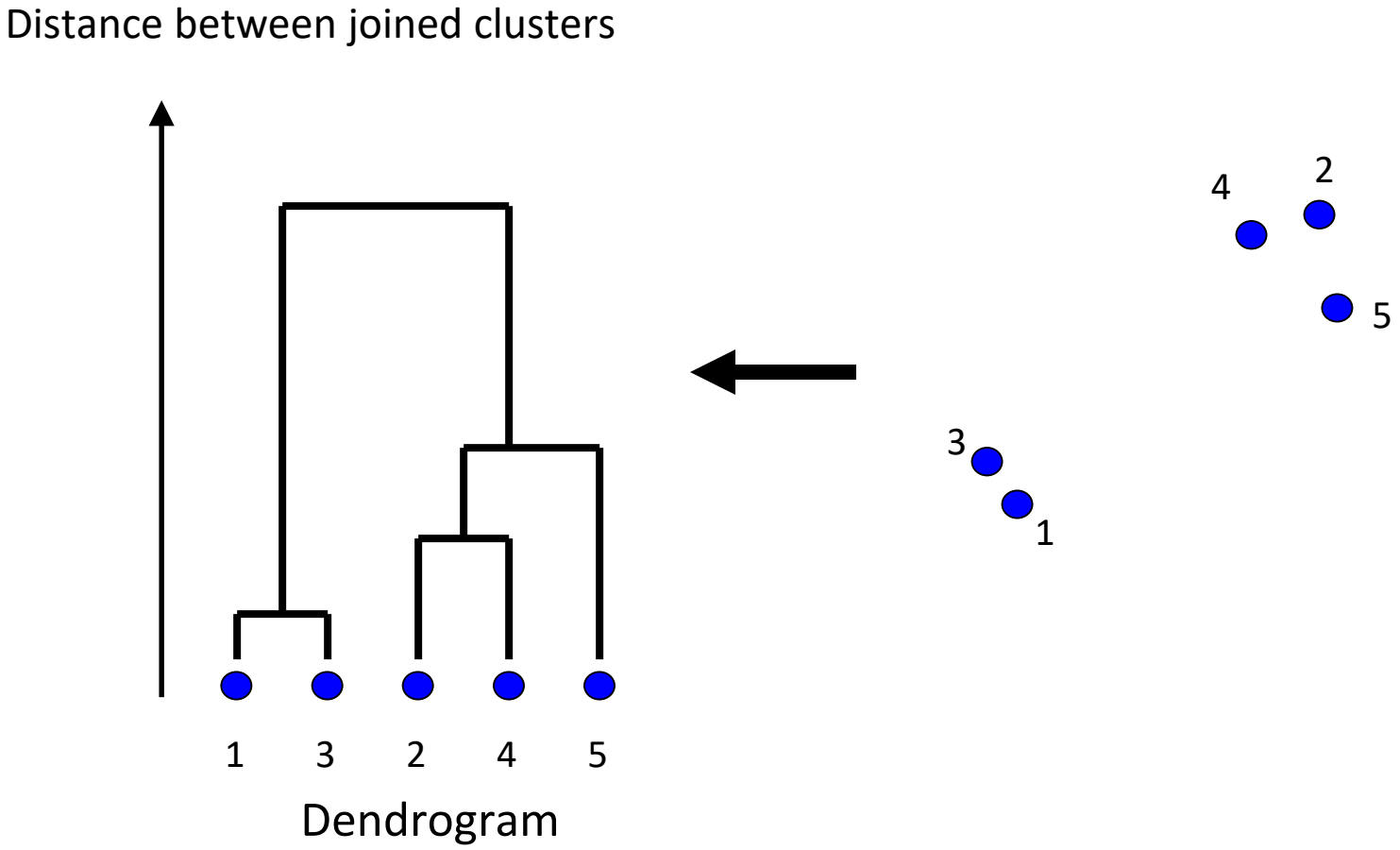
- Guaranteed to reach a local minimum

K=3



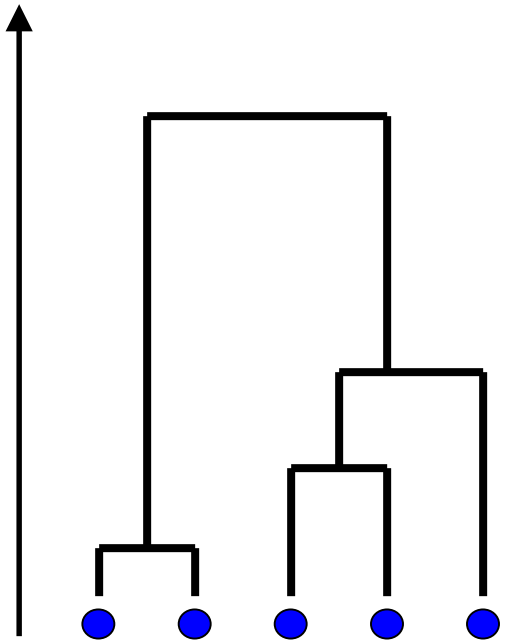
Iteration = 2

Hierarchical Clustering

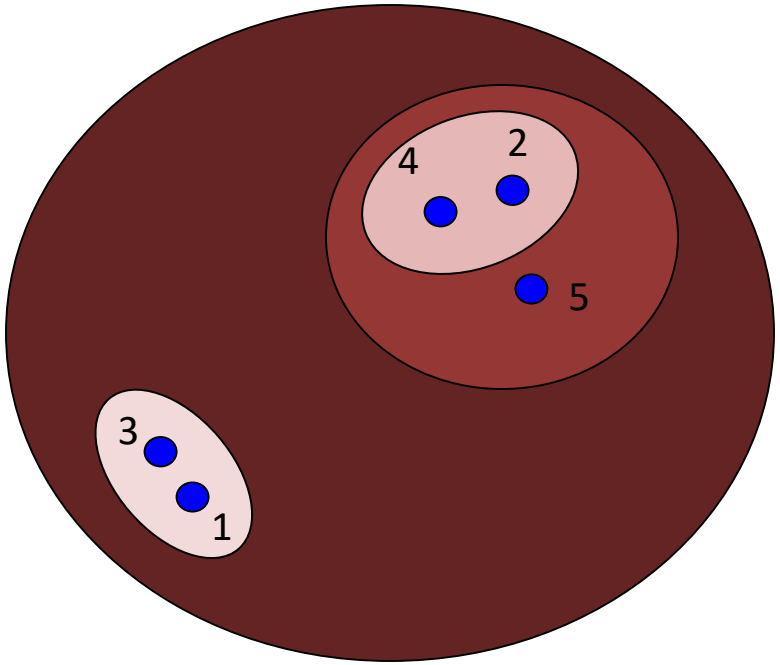


Hierarchical Clustering

Distance between joined clusters



Dendrogram



Linkage is the method for linking clusters based on the distance between them.

Average Linkage: average distance between all pairs

Complete Linkage: farthest distance between all pairs

Single Linkage: closest distance between all pairs

Clustering Exercise

BioITWorld Clustering

Classification / Prediction

Classification

"Supervised Learning"

Use a "training set" of examples to create a model that is able to predict, given an unknown sample, which of two or more classes that sample belongs to.

?

TUMOUR CLASSIFIC

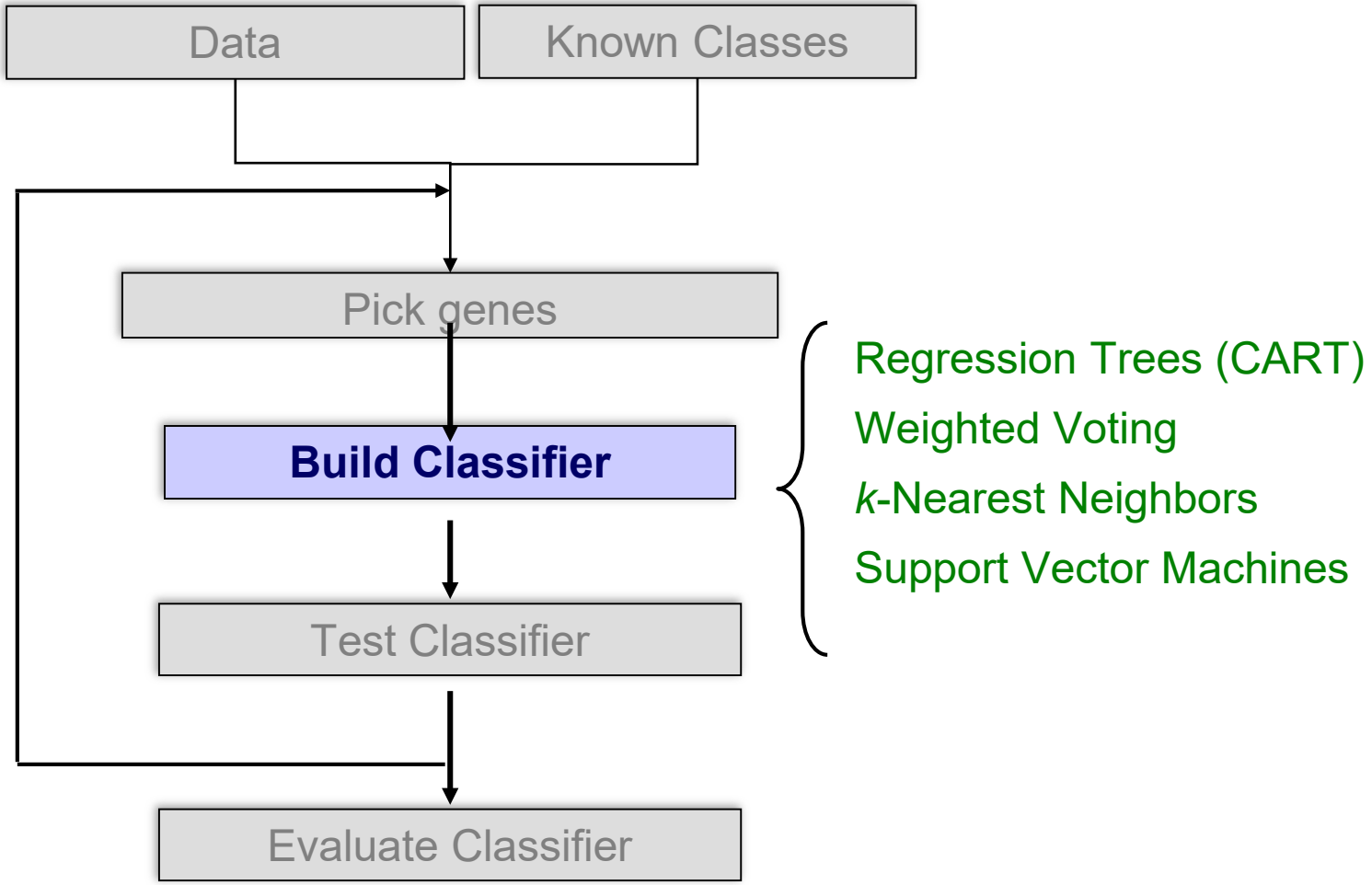
Recognizing differences

What Is a Classifier

- A **predictive** rule that uses a set of **inputs** (genes) to predict the values of the **output** (phenotype).
- Known examples (train data) are used to build the predictive rule.
- Goal:
  - Achieve high predictive power.
  - Avoid over-fitting: i.e. classifier memorizes the training data and is not generalizable to other test data

Classification

Computational methodology



Classifiers

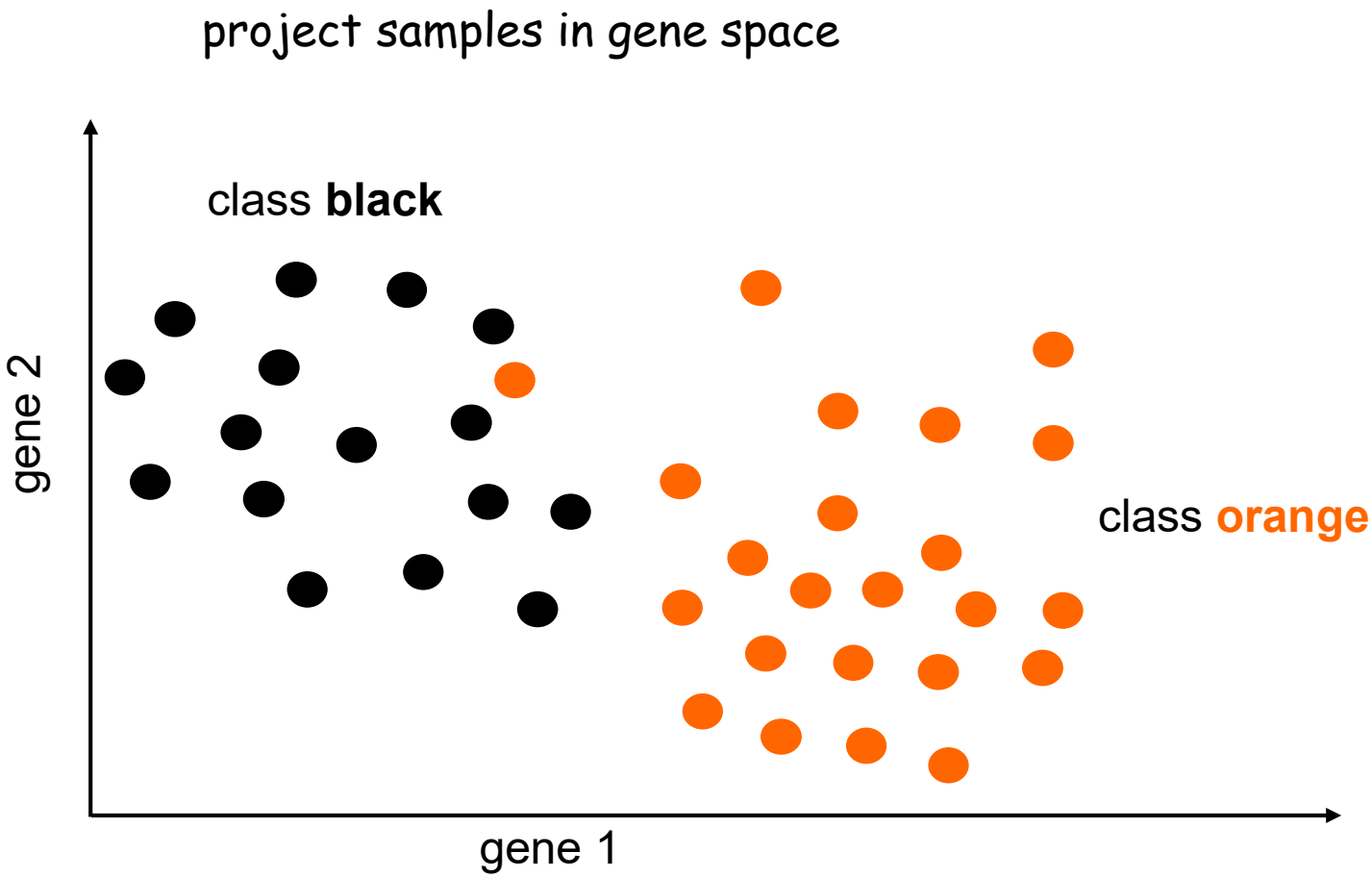
Important issues:

- Few cases, many variables (genes)
- redundancy: many highly correlated genes.
- noise: measurements are very imprecise.
- feature selection: reducing the # of genes is a necessity.

Avoid over-fitting

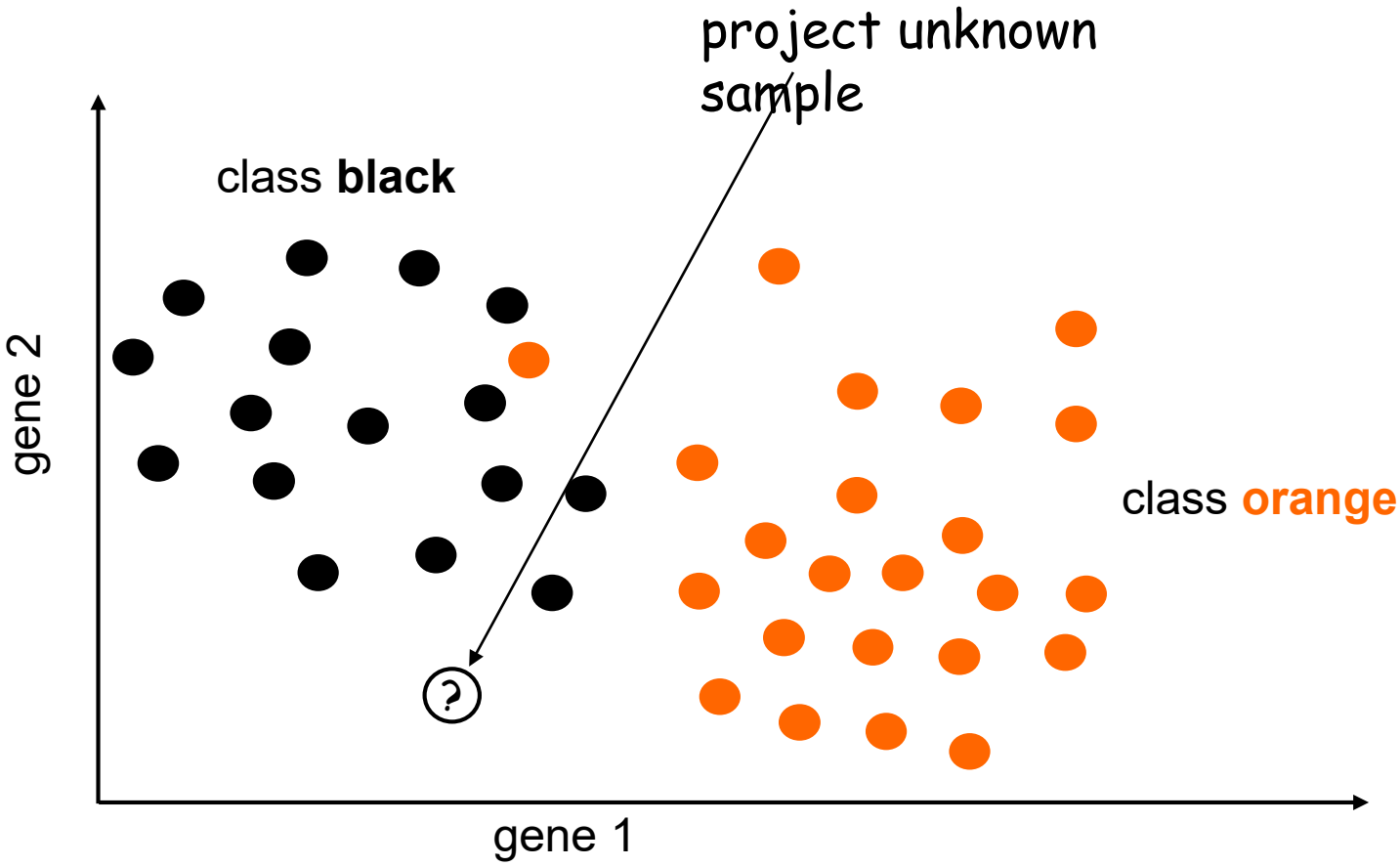
k Nearest Neighbors (kNN) Classifier

Example: k=5, 2 genes, 2 classes



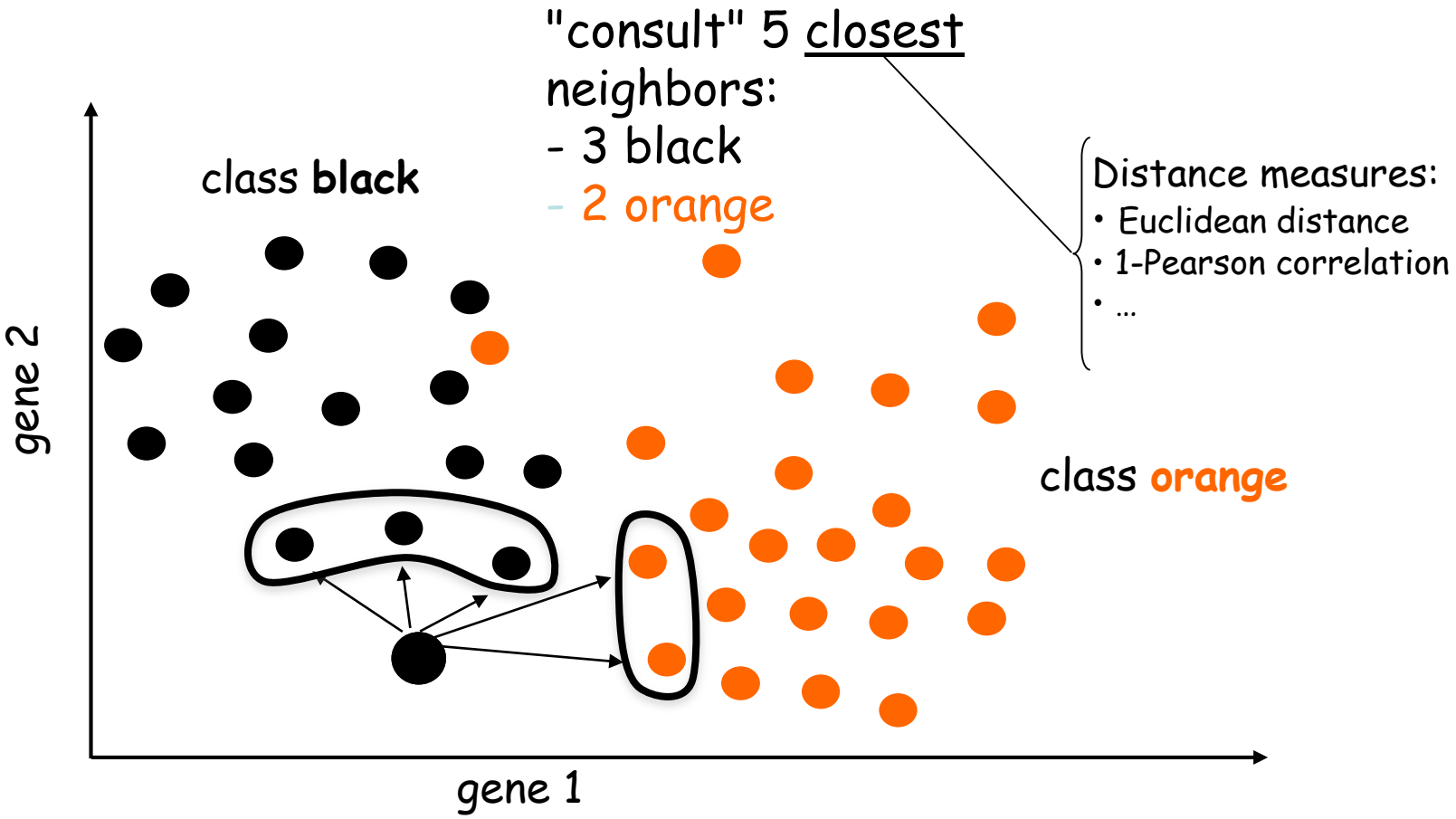
kNN Classifier

Example: k=5, 2 genes, 2 classes

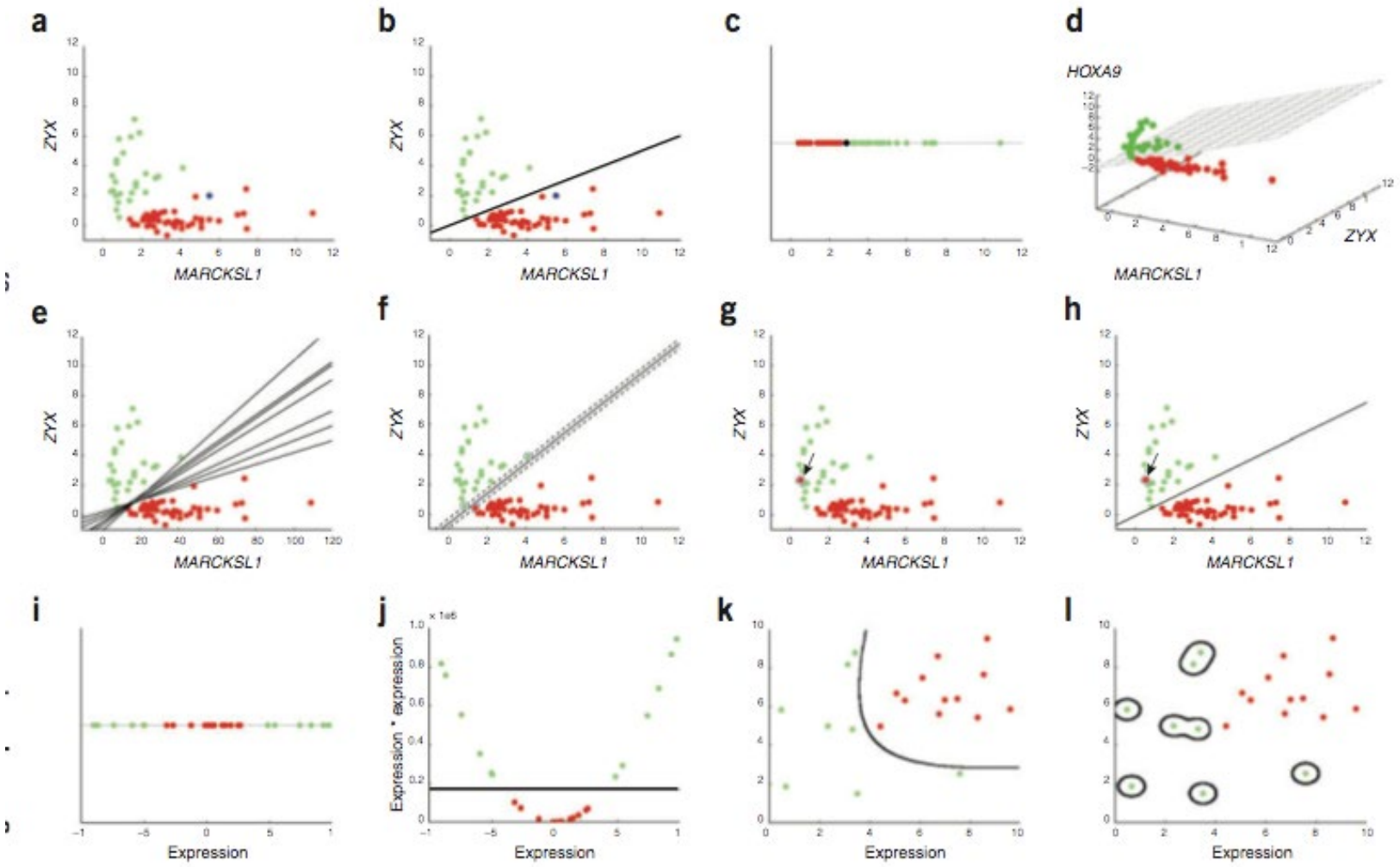


kNN Classifier

Example: K=5, 2 genes, 2 classes



Support Vector Machine (SVM)



Noble, Nat Biotech 2006

Testing the Classifier

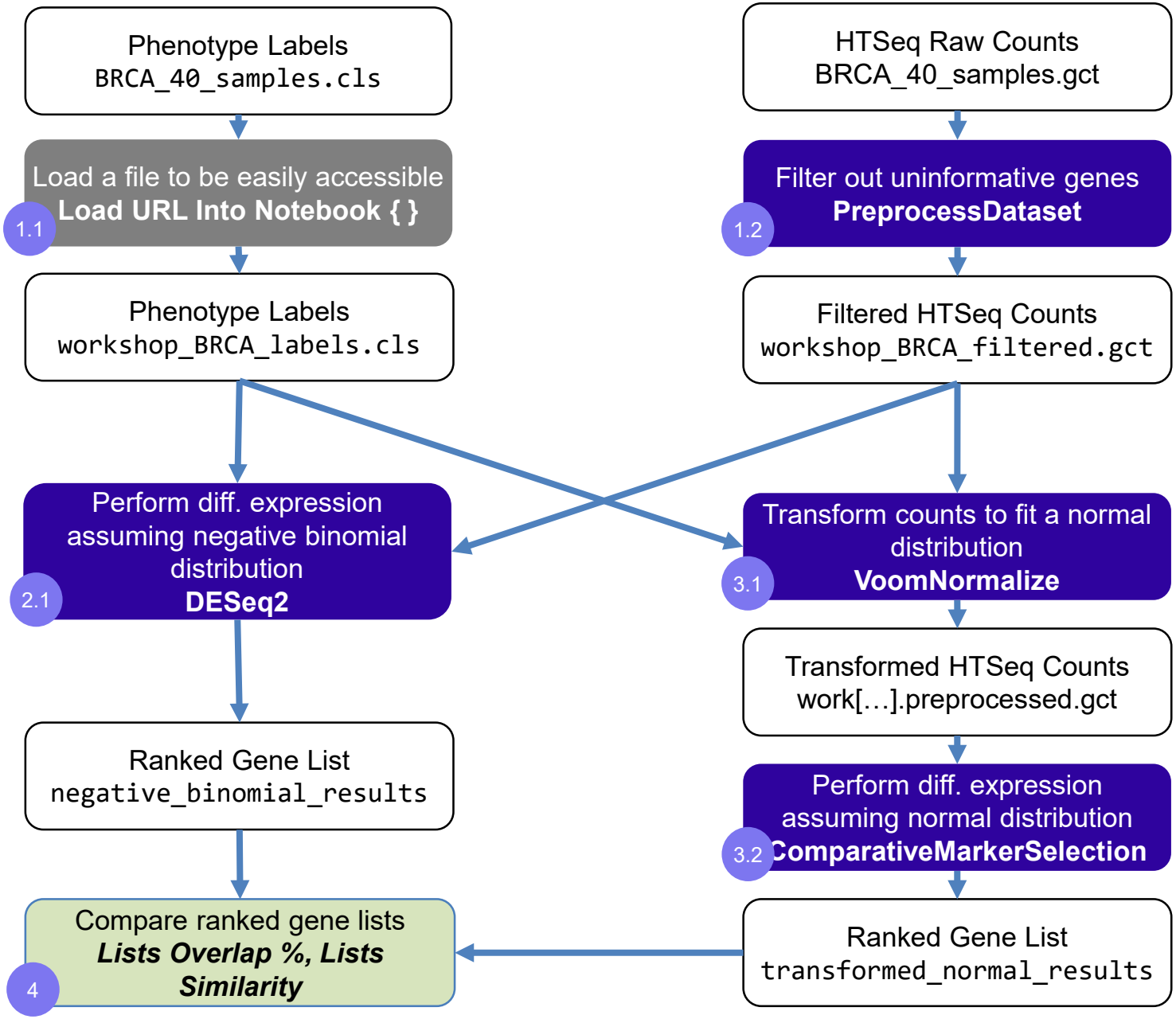
$$\text{error rate} = \frac{\text{\# of cases correctly classified}}{\text{total \# of cases}}$$

- Evaluation on independent test set
  - Build the classifier on the train set.
  - Assess prediction performance on test set.
  - What if we don't have an independent test set?
- Cross Validation (XV):
  - Split the dataset into n folds (e.g., 10 folds of 10 samples each).
  - For each fold (e.g., for each group of 10 samples),
    - train (i.e., build model) on n-1 folds (e.g., on 90 samples),
    - test (i.e., predict) on left-out fold (e.g., on remaining 10 samples).
  - Combine test results.

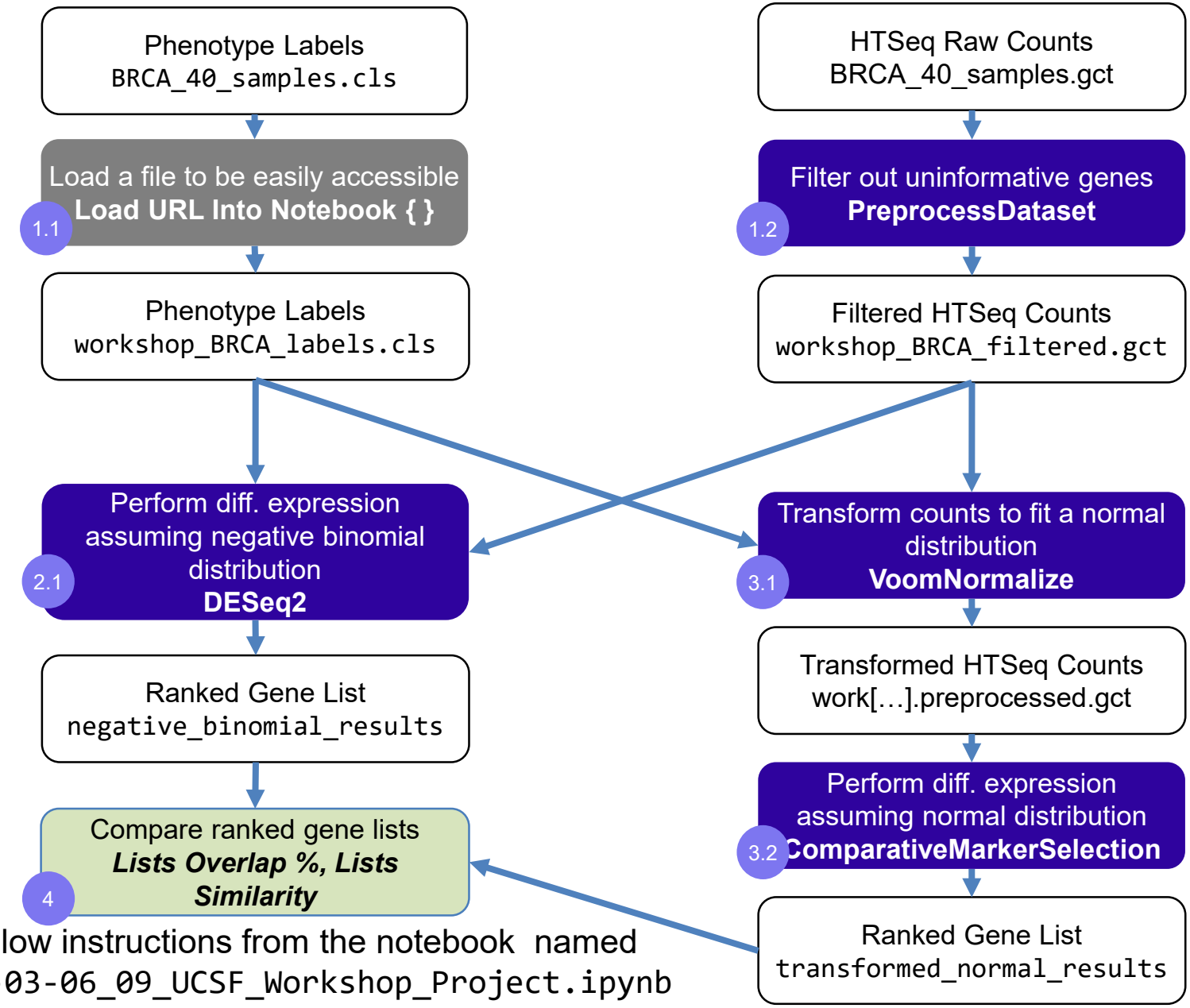
Classification and Prediction Exercise

BioITWorld Classification and Prediction

Analyzing HTSeq Data Using GenePattern



Analyzing HTSeq Data Using GenePattern



Follow instructions from the notebook named
`2019-03-06_09_UCSF_Workshop_Project.ipynb`