



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

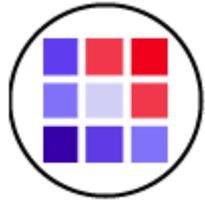
GSEA

Clustering

Class Prediction

Register for an account at:
notebook.genepattern.org

No spaces or special characters in usernames



GenePattern Overview

Tools for Bioinformatics



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

Samtools

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

Samtools	Reading/writing/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 htseq so they can be built independently.



Bowtie 2
Fast and sensitive read alignment



Home Installation Documentation Examples

1.4. Support Vector Machines

Principal Component Analysis

Picard

build passing

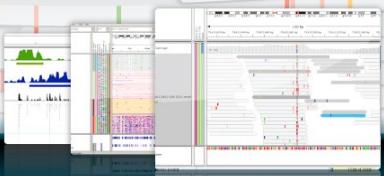
A set of command line tools (in Java) for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF.

Hierarchical Clustering / Dendrograms

Cufflinks

Transcriptome assembly and differential expression analysis for RNA-Seq.

Integrative Genomics Viewer



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:

NMF: Non-negative Matrix Factorization

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 computes a series of quality control metrics for RNA-seq data.



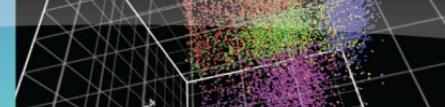
Network Data Integration, Analysis, and Visualization in a Box

Trimmomatic: A flexible read trimming tool for Illumina NGS data



MSigDB
Molecular Signatures Database

FLAME
Flow analysis with Automated Multivariate Estimation



Gene Set Enrichment Analysis

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

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 --chunkmb 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

The screenshot shows a web browser window titled "GenePattern - Bowtie.aligner" at the URL <https://genepattern.broadinstitute.org/gp/pages/index.jsf?lsid=urn:lsid:broad.mit.edu:cancer.software.genepattern.module.analysis.Bowtie.aligner>. The page displays the "Bowtie.aligner" module configuration. The top navigation bar includes links for Modules & Pipelines, Suites, Job Results, Resources, Help, and GenomeSpace. The main content area is titled "Bowtie.aligner" version 4. It provides a brief description: "Bowtie2 (v. 2.1.0) is an ultrafast and memory-efficient short read aligner." Below this, there are several input fields:

- "prebuilt bowtie index": A dropdown menu currently set to "None".
- "custom bowtie index": A section with "Upload File..." and "Add Path or URL..." buttons, and a "Drag Files Here" area.
- "input format*": A dropdown menu currently set to "FASTQ".
- "reads pair 1*": A section with "Upload File..." and "Add Path or URL..." buttons, and a "Drag Files Here" area.
- "reads pair 2": A section with "Upload File..." and "Add Path or URL..." buttons, and a "Drag Files Here" area.
- "quality value scale": A dropdown menu currently set to "Phred".
- "integer quality values": A dropdown menu currently set to "no".
- "max reads to align": An empty text input field.

Each input field has a "Batch" checkbox to its right. At the bottom right of the form are "Reset" and "Run" buttons.

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

AddNoiseToFCS	AffySExpressionFileCreator
Add noise to specified parameters in an FCS data file. Flow Cytometry	[Beta Release** Contact gp-help with any issues. Check stdout.txt and stderr.txt for errors]... Preprocess & Utilities
ApplyGatingML	ARACNE
Apply a Gating-ML file on an FCS data file (gate and/or transform list mode data) Flow Cytometry	Runs the ARACNE algorithm for reverse engineering cellular networks Pathway Analysis
AreaChange	Arff2Gct
Calculates fraction of area under the spectrum that is attributable to signal (area after noise)... Proteomics, ProteomicsSuite	Convert an .arff file into a gene pattern .gct / .cls file pair Multi-label Protein Prediction Suite (MiPPS), Preprocess ...
ATARI	AuDIT
Runs ATARI on RNAi reagent-level data. RNAi	Automated Detection of Inaccurate and Imprecise Transitions in MRM Mass Spectrometry Proteomics
BedToGtf	Beroukhim.Getz.2007.PNAS.Glioma.... pipeline
Converts BED files to GFF or GTF format Data Format Conversion	
BlastTrainTest	BlastXValidation
Sequence similarity classification using BLAST Multi-label Protein Prediction Suite (MiPPS), Prediction	Sequence similarity cross validation prediction using BLAST Multi-label Protein Prediction Suite (MiPPS), Prediction
Bowtie.aligner	Bowtie.indexer
Bowtie2 (v. 2.1.0) is an ultrafast and memory-efficient short read aligner. RNA-seq	Builds a Bowtie2 (v. 2.1.0) index from a set of DNA sequences RNA-seq
BWA.aln	BWA.bwasw
[**Beta Release** Contact gp-help with any issues. Check stdout.txt and stderr.txt for errors]... RNA-seq	[**Beta Release** Contact gp-help with any issues. Check stdout.txt and stderr.txt for errors]... RNA-seq
BWA.indexer	CaArray2ImportViewer
[**Beta Release** Contact gp-help with any issues. Check stdout.txt and stderr.txt for errors]... RNA-seq	Imports data files from CaArray 2.4.1 and creates gct or cls files Visualizer
CART	CARTXValidation
Classification and Regression Tree Prediction	Classification and Regression Tree Cross-Validation Prediction
CBS	ChIPSeq.CreateHeatmap
Segments DNA copy number data into regions of estimated equal copy number using circular binary... SNP Analysis	Generates a heatmap based on the ChIP-Seq signal extracted from a BAM file, according to the...

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

Copy Number
Divide
by Normals

GISTIC

Classification
and
Regression Trees

GISTIC

GSEA

CBS

Support
Vector
Machines

Expression
File
Creator

Variation
Filter

k-Nearest
Neighbors

**Hierarchical
Clustering**

Metagene
Projection

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

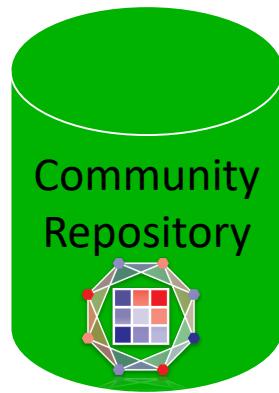
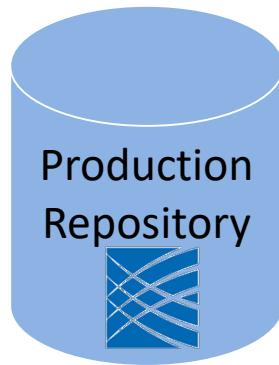
The screenshot shows the GenePattern web application interface. At the top, there's a navigation bar with tabs for 'Modules & Pipelines', 'Suites', 'Job Results', 'Resources', 'Help', and 'GenomeSpace'. Below the navigation bar, there are three main sections: 'Modules' (selected), 'Jobs', and 'Files'. A search bar says 'Search Modules & Pipelines' with a 'Browse Modules' button. Under 'Recent Modules', there are links to 'ComparativeMarkerSelectionViewer', 'FeatureSummaryViewer', 'HeatMapView', and 'HierarchicalClusteringViewer'. On the left, there's a dashed box labeled 'Drag Modules Here' for dragging and dropping modules. The central part of the screen is titled 'Browse Modules > All Modules' and contains a grid of module cards. A red box highlights the first two columns of this grid. Each card includes the module name, a brief description, and its category. Some cards have a gear icon for configuration.

Module	Description	Category
ABSOLUTE	Extracts absolute copy numbers per cancer cell from a mixed DNA population. Use this module... SNP Analysis	SNP Analysis
ABSOLUTE.review	Extracts the absolute copy number per cancer cell from a mixed DNA population. Use this module... SNP Analysis	SNP Analysis
ABSOLUTE.summarize	Summarizes the results from multiple ABSOLUTE runs so that an analyst can manually review the solutions. SNP Analysis	SNP Analysis
AddFCSEventIndex	Adds indexes to events in a Flow Cytometry Standard (FCS) data file. Flow Cytometry	Flow Cytometry
AddFCSParameter	Add parameters and their values to a FCS data file Flow Cytometry	Flow Cytometry
AddNoiseToFCS	Add noise to specified parameters in an FCS data file. Flow Cytometry	Flow Cytometry
aml.all.pipeline	ALL/AML methodology, from Golub and Slonim et al., 1999 pipeline	
ApplyGatingML	Apply a Gating-ML file on an FCS data file (gate and/or transform list mode data) Flow Cytometry	Flow Cytometry
ARACNE	Runs the ARACNE algorithm for reverse engineering cellular networks Pathway Analysis	Pathway Analysis
AreaChange	Calculates fraction of area under the spectrum that is attributable to signal (area after noise... Proteomics, ProteomicsSuite	Proteomics, ProteomicsSuite
Arff2Gct	Convert an .arff file into a gene pattern .gct / .cls file pair Multi-label Protein Prediction Suite (MiPPS), Preprocess ...	MiPPS, Preprocess
ATARIS	Runs ATARIS on RNAi reagent-level data RNAi	RNAi
AuDIT	Automated Detection of Inaccurate and Imprecise Transitions in MRM Mass Spectrometry Proteomics	Proteomics
BedToGtf	Converts BED files to GFF or GTF format Data Format Conversion	Data Format Conversion
Beroukhim.Getz.2007.PNAS.Glioma.GI! pipeline		
Birdseed	SNP genotyping algorithm that runs on the Affymetrix 500K, SNP5.0, and SNP6.0 platforms SNP Analysis	SNP Analysis
BirdseedCallRate	Computes the call rate of the Birdseed algorithm SNP Analysis	SNP Analysis
BirdseedDataPreparation	Prepare a bspn file for running Birdseed SNP Analysis	SNP Analysis
BlastTrainTest	Sequence similarity classification using BLAST Multi-label Protein Prediction Suite (MiPPS), Prediction	MiPPS, Prediction
BlastXValidation	Sequence similarity cross validation prediction using BLAST Multi-label Protein Prediction Suite (MiPPS), Prediction	MiPPS, Prediction
Bowtie.aligner	Bowtie2 (v. 2.1.0) is an ultrafast and memory-efficient short read aligner. RNA-seq	RNA-seq
Bowtie.indexer	Builds a Bowtie2 (v. 2.1.0) index from a set of DNA sequences RNA-seq	RNA-seq
BWA.aln		
BWA.bwasw		

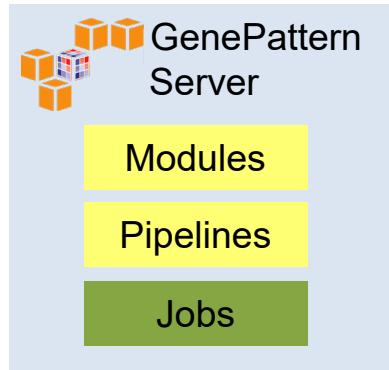
About GenePattern | Contact Us

How GenePattern works

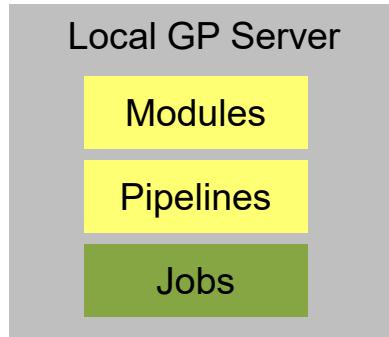
Repositories



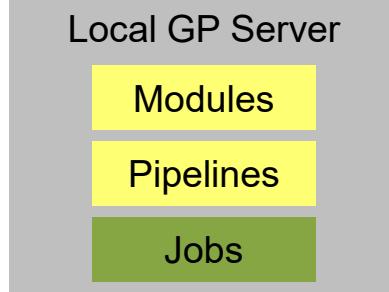
Servers



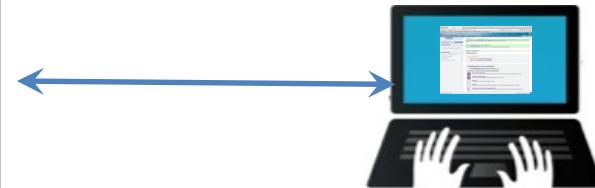
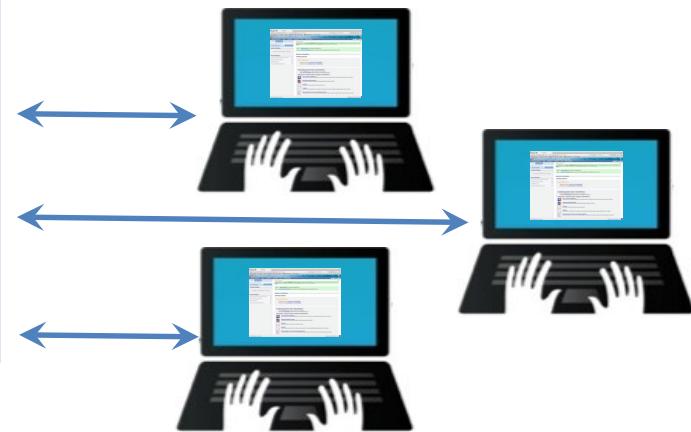
Local GP Server

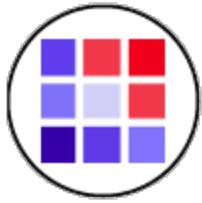


Local GP Server



GenePattern Users





GenePattern Notebook Environment

GenePattern Notebook Tutorial GenePattern Notebook Tutorial (unsaved changes)

Control Panel Logout gpdemo

File Edit View Insert Cell Kernel Widgets Help

Not Trusted Python 3.6

Markdown Tools

GenePattern Notebook Tutorial

The GenePattern Notebook Environment provides a variety of features for both basic and advanced users. This tutorial will familiarize you with some of its most important features.

All instructions for you to follow will appear in a blue panel like this one.

GenePattern Notebook Introduction Video

Below is a brief video introduction to the GenePattern Notebook Environment. This video introduces many of the basic concepts and features provided by the tool. If you would prefer a more "hands on" introduction, scroll down and follow the subsequent interactive tutorial.

To view the video, click the Play button in the middle of the video cell.

If the video is not visible, highlight the cell below and press the Run Cell (▶) button in the toolbar to see the video.

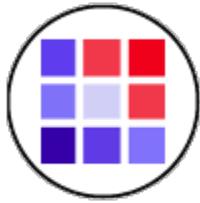
In [3]: `%HTML
<iframe width="854" height="480" src="https://www.youtube.com/embed/r5Km4UPhb1Q" frameborder="0" allowfullscreen></iframe>`

GenePattern Notebook Environment

jupyter Untitled Last Checkpoint: 2 hours ago (unsaved changes)

File Edit View Insert Cell Kernel Help

Test Dataset → Bayesian Predictor of Outcome → Probability of Relapse



Run an Analysis Notebook

2019-04-16_01 BioITWorld GenePattern Notebook
Introduction

The screenshot shows the GenePattern Notebook interface. At the top, there's a header bar with the title "GenePattern Notebook" and the date "2019-04-16_01 BioITW...". It also shows "Last Checkpoint: 2 minutes ago (unsaved changes)", "Logout gpdemo", and "Control Panel". Below the header is a toolbar with various icons for file operations like Open, Save, Insert, Run, etc., followed by "Python 3.6". The main content area has a section titled "Introduction to GenePattern Notebook" which contains instructions about preprocessing and viewing results in a heat map. A blue box highlights the instruction "Instructions are given in blue boxes, such as with the one below." Below this is a "Login" section with fields for "GenePattern Username" and "Username". A message in the center says "Log into GenePattern Server" and "You have already authenticated with the GenePattern Public Server. Would you like to automatically sign in now?".

GenePattern Notebook 2019-04-16_01 BioITW... Last Checkpoint: 2 minutes ago (unsaved changes) Logout gpdemo Control Panel

File Edit View Insert Cell Kernel Widgets Help

Not Trusted Python 3.6

Introduction to GenePattern Notebook

This document should help you understand how to run an analysis in the GenePattern Notebook environment. In it you will perform a simple preprocessing step and then view the results in a heat map.

Instructions are given in blue boxes, such as with the one below.

Instructions Sign in to GenePattern by clicking the login button or entering your username and password into the form below.

Dataset information

In this example we will preprocess a dataset of 38 samples of leukemia, 27 of subtype ALL and 11 of subtype AML. The data was created on a microarray platform, but the resulting [GCT](#) file is compatible with RNA-Seq, as well as any other data type that can be expressed with samples as columns and features as rows.

GenePattern Login

GenePattern Server

GenePattern Cloud

GenePattern Username

Username

Log into GenePattern Server

You have already authenticated with the GenePattern Public Server.
Would you like to automatically sign in now?





Jupyter Notebook

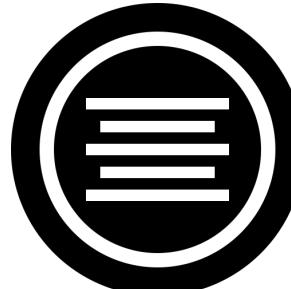
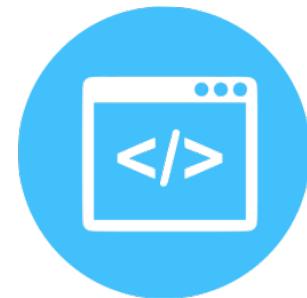
- Popular and well-supported framework for scientific computing
- Ecosystem of available extensions and resources
- Open source

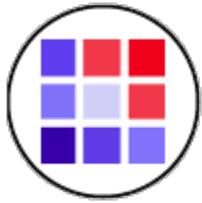




Complete Research Narrative

- Leverages the best of Jupyter and GenePattern
- Interleave text, visualization, graphics and analytical aspects





GenePattern Notebook Repository

GenePattern Notebook

Sign in

Username:

Password:

Sign In [Forgot Password?](#)

GenePattern Server Status

For more information please contact us on the [GenePattern Help Forum](#)

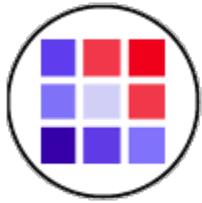
Register GenePattern Account

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

[Register a New GenePattern Account](#)

Documentation is available on the GenePattern Notebook website.

<https://notebook.genepattern.org>



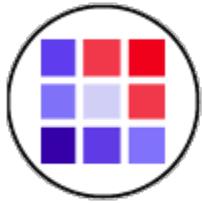
Notebook Workspace

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

The screenshot shows the GenePattern Notebook interface. At the top, there's a navigation bar with tabs for "Files" (which is selected), "Running", and "Public Notebooks". On the right side of the bar are "Control Panel" and "Logout genepattern" buttons. Below the navigation bar, a message says "Select items to perform actions on them." A file list table follows, containing two entries:

	Name	Last Modified
<input type="checkbox"/>	GenePattern Notebook Tutorial.ipynb	seconds ago
<input type="checkbox"/>	GenePattern Python Tutorial.ipynb	seconds ago

On the far right of the table, there are "Upload", "New", and a refresh icon buttons. There are also "Name" and "Last Modified" sort buttons at the top right of the table area.

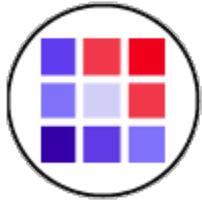


Browse Public Notebooks

- 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.

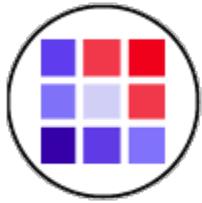
The screenshot shows the GenePattern Notebook interface. At the top, there's a navigation bar with tabs for 'Files', 'Running', and 'Notebook Library'. The 'Notebook Library' tab is active. On the left, there's a sidebar with categories like 'Public Notebooks' (with 'Featured' selected), 'All Notebooks', 'Tutorial', 'Workshop', 'Community', 'My Notebooks', and 'Shared Notebooks' (with 'Shared By Me' and 'Shared With Me'). The main area is titled 'Featured Notebooks' and contains a table with the following data:

Notebook	Authors	Updated	Quality
Classification and Prediction - RNAseq Use RNA-seq data with k-Nearest Neighbors (kNN) to build a predictor, use it to classify leukemia subtypes, and assess its accuracy in cross-validation. <small>featured</small>	GenePattern Team	2018-03-20	Release
Classification and Prediction Example of how to use k-Nearest Neighbors (kNN) to build a predictor, use it to classify leukemia subtypes, and assess its accuracy in cross-validation. <small>featured</small>	GenePattern Team	2018-03-20	Release
Differential Expression Analysis Example of using differential expression analysis to find genes that are significantly differentially expressed between classes of samples. <small>featured</small>	GenePattern Team	2018-03-19	Release
Hierarchical Clustering - RNASeq Use RNA-seq data to cluster genes and/or samples agglomeratively, based on how close they are to one another. <small>featured</small>	GenePattern Team	2018-03-19	Release



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 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

The screenshot shows the GenePattern Notebook Environment interface. The top navigation bar includes File, Edit, View, Insert, Cell, Kernel, Widgets, and Help. The status bar indicates "Not Trusted" and "Python 3.6". The main content area displays a "pattern Notebook Tutorial" with text about its features and a "pattern Notebook Introduction Video" section with a video player.

File

- New Notebook
- Open...
- Make a Copy...
- Save as...
- Rename...
- Save and Checkpoint
- Revert to Checkpoint**
- Print Preview
- Download as
- Publish to Repository
- Share with Collaborators
- Trust Notebook
- Close and Halt

pattern Notebook Tutorial

pattern Notebook Environment provides a variety of features for both basic and advanced users. This tutorial will familiarize you with some of its important features.

Actions for you to follow will appear in a blue panel like this one.

pattern Notebook Introduction Video

Brief video introduction to the GenePattern Notebook Environment. This video introduces many of the basic concepts and features provided by the environment. You may also prefer a more "hands on" introduction, scroll down and follow the subsequent interactive tutorial.

To view the video, click the Play button in the middle of the video cell.

If the video is not visible, highlight the cell below and press the Run Cell (▶) button in the toolbar to see the video.



GenePattern Cells

Authentication Cell

GenePattern Login

GenePattern Server

GenePattern Cloud

GenePattern Username

Username

GenePattern Password

Password

Log into GenePattern Register an Account

Analysis Cell

GenePattern ConvertLineEndings Version 2

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

Run

input filename* https://cloud.genepattern.org/gp/jobResults/104715/all_aml_train.preprocessed.gct

The input file (any non-binary file format)

output file* <input.filename_basename>.cvt.<input.filename_extension>

The output file

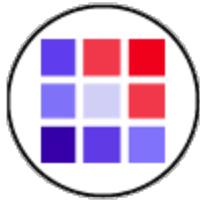
Run

Job #104716

Completed Submitted by GPDemo on 2019-03-29T18:29:45+00:00

all_aml_train.preprocessed.cvt.gct

gp_execution_log.txt



Authentication Cells

GenePattern Login

GenePattern Server
GenePattern Cloud

GenePattern Username
Username

GenePattern Password
Password

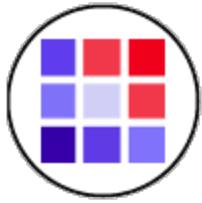
Log into GenePattern **Register an Account**

GenePattern GPDemo <https://cloud.genepattern.org/gp>

-- March 18, 2019 -- Due to recent browser security updates, cracking down on mixed content, some of our viewers are unable to display their content. We are currently only aware of one module (ConstellationMap) being impacted. Please let us know if you discover others. Sincerely, The GenePattern Team Follow the GenePattern team on Twitter, Instagram or Facebook to keep up with the latest news and events or join the conversation in our forum!

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

Leave Feedback



Analysis Cells

GenePattern CollapseDataset Version 1 Run

Collapses all probe set values for a gene into a single vector of values

dataset file* Upload File... Add File or URL...

Dataset file - .gct, .res

chip platform* ftp://ftp.broadinstitute.org/pub/gsea/annotations/HU6800.chip X

The chip platform

collapse mode* Maximum

Collapse mode for probe sets => 1 gene

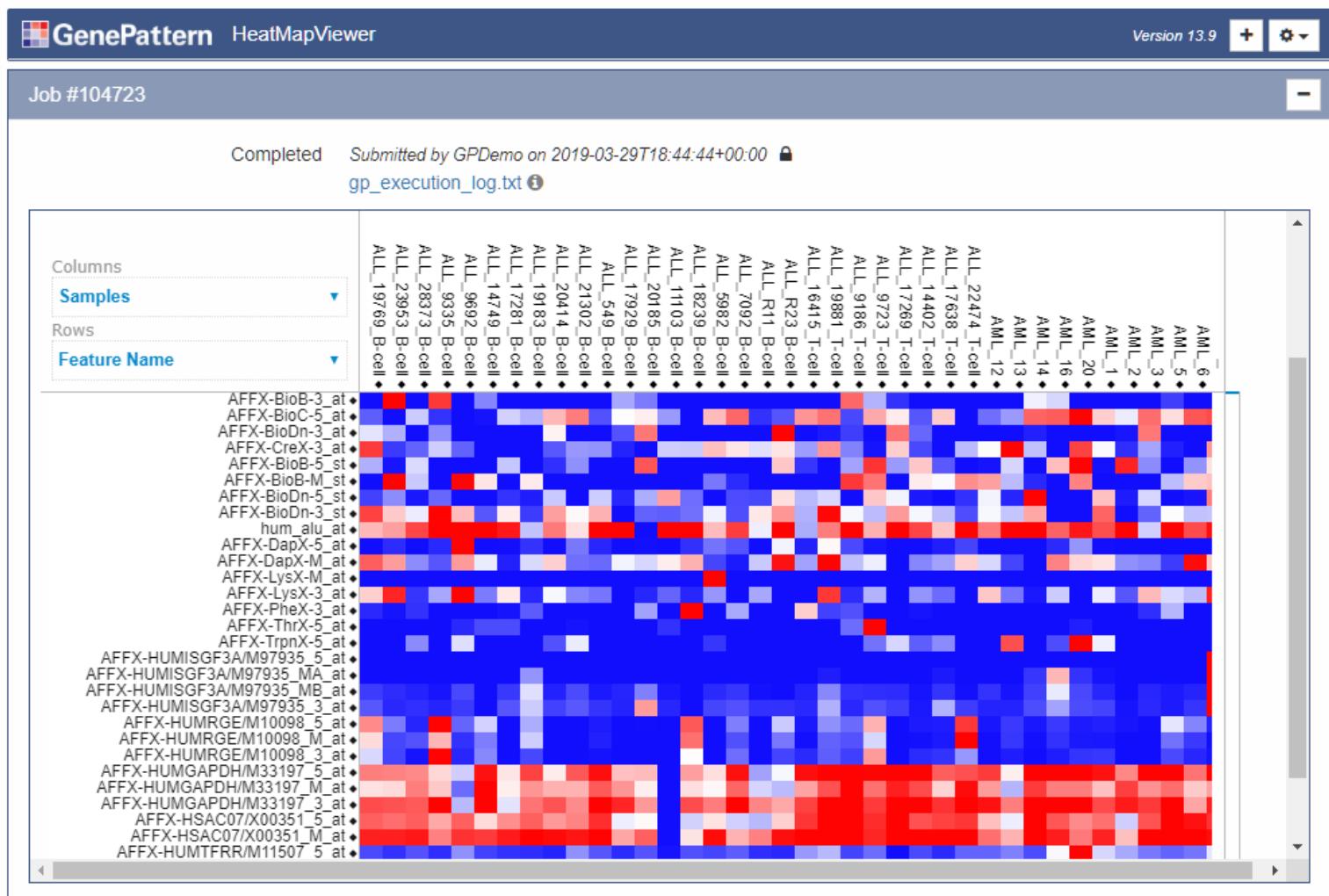
output file name* <dataset.file_basename>.collapsed

The output file name

Run



Job/Result Cells



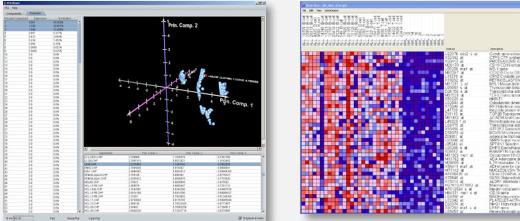
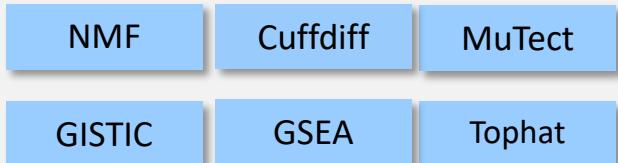


Running GenePattern Notebook

- 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.

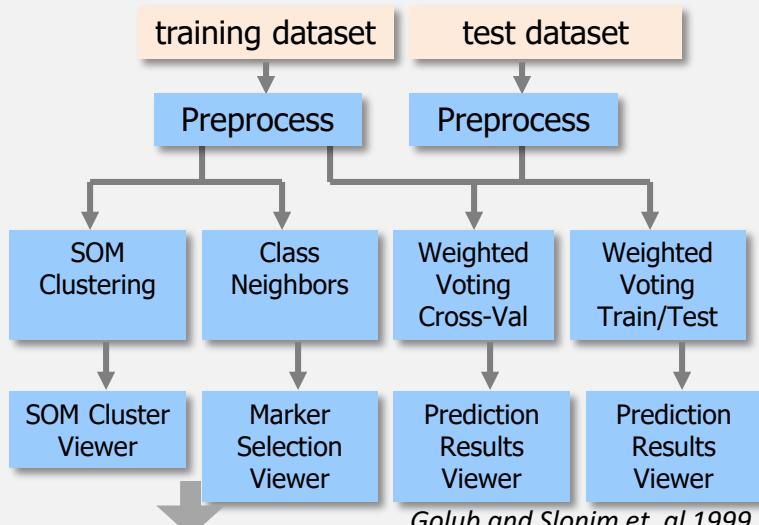
The GenePattern Ecosystem: Architecture

Module Repository

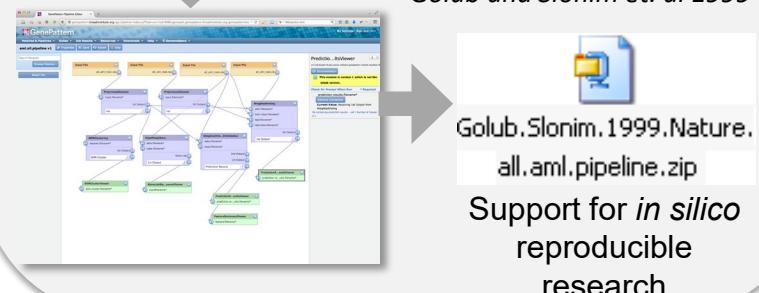


Hundreds of analysis and visualization tools

Pipeline Environment



Golub and Slonim et. al 1999

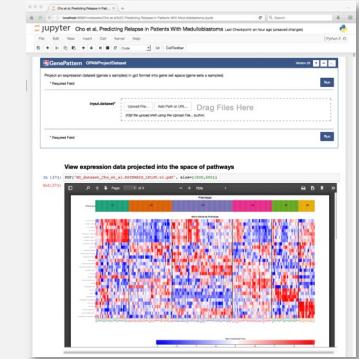


Support for *in silico* reproducible research

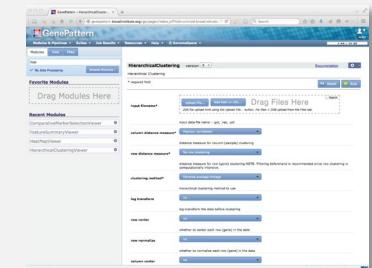
Analysis Engine



Record/replay analyses
Versioning of methods
Web service access



Notebook

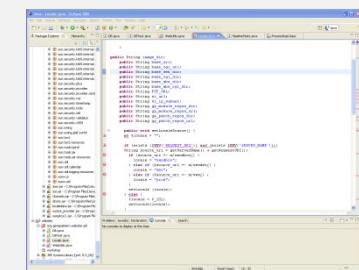


Web

Module Integrator



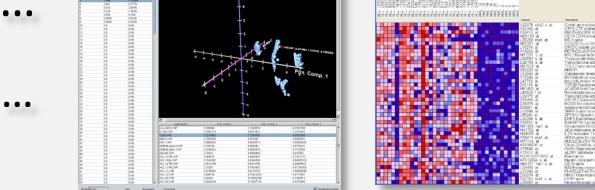
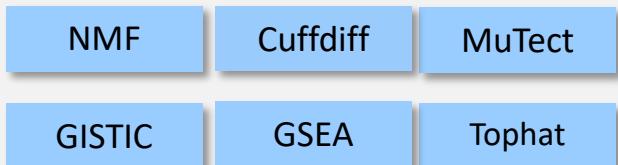
Easy addition of new tools



Programming
Access for all levels of user

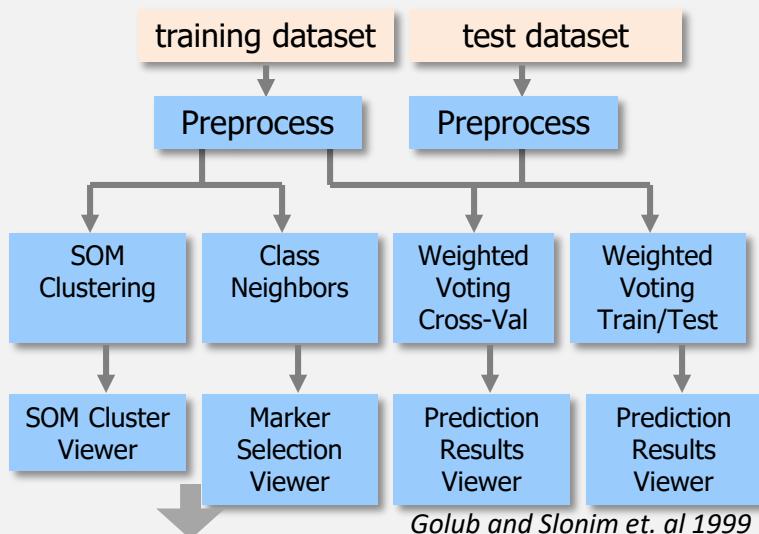
The GenePattern Ecosystem: Architecture

Tool Library

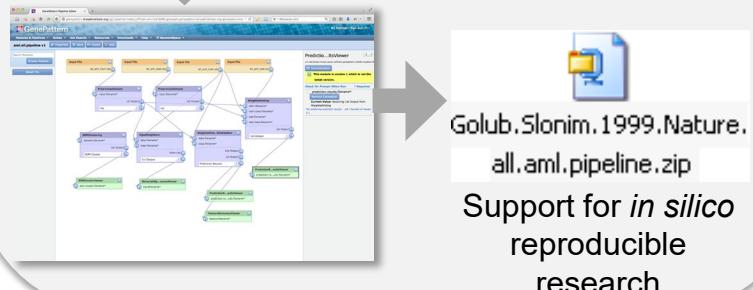


Hundreds of analysis and visualization tools

Workflows



Golub and Slonim et. al 1999

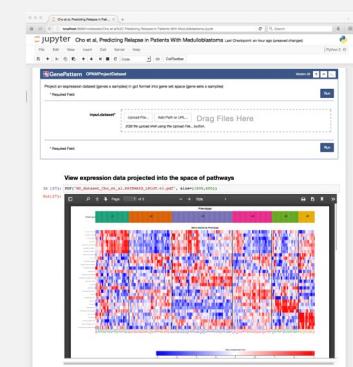


Analysis Engine

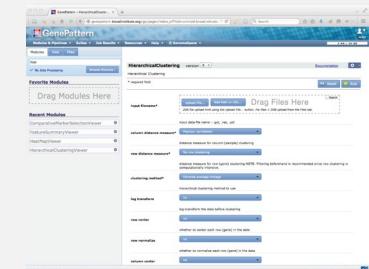


Record/replay analyses
Versioning of methods
Web service access

Access Points

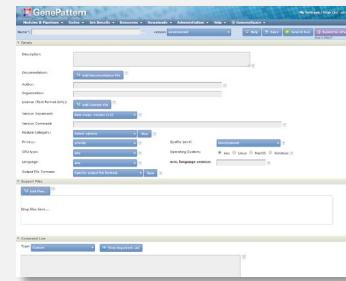


Notebook

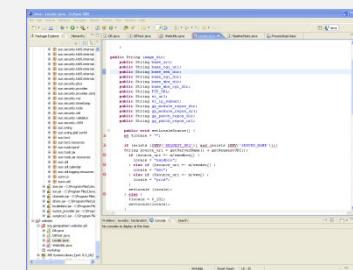


Web

Create your own tool

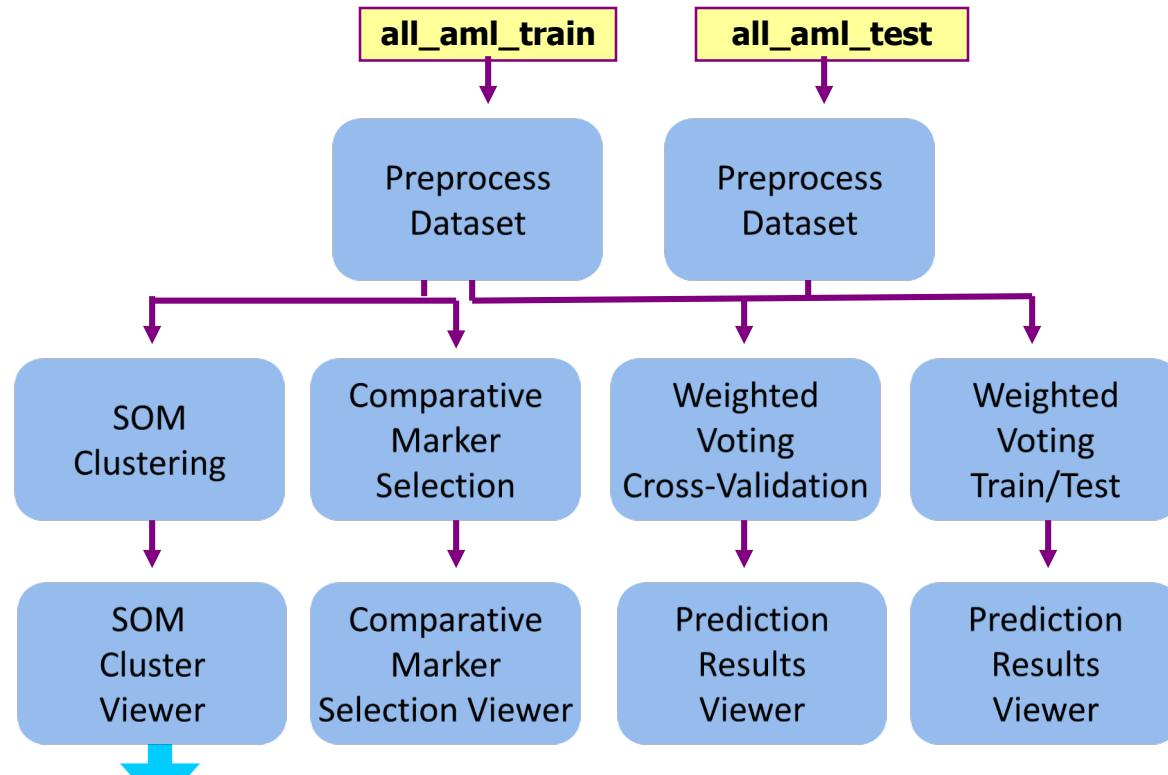


Easy addition of new tools



Programming
Access for all levels of user

GenePattern vocabulary: Pipelines



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

Pipelines in GenePattern

The screenshot shows the GenePattern web interface on a Mac OS X system. The main content area displays a list of pipelines, each represented by a card with a title, a brief description, and a gear icon for configuration. A red box highlights the first 15 pipelines in the list.

Browse Modules > pipeline

- aml.all.pipeline
ALL/AML methodology, from Golub and Slonim et al., 1999 pipeline
- CBSWrapperPipeline
A one step pipeline that runs CBS pipeline
- CopyNumberInferencePipeline.Part2of...
A pipeline that runs CopyNumberInferencePipeline.Part2of2 – Part of the pipeline
- FLAMEContourViewer.Pipeline
Pipeline which runs the FLAMECounterDataGenerator and the FLAMEViewer pipeline
- Golub.Slonim.1999.Nature.all.aml.pipeline
ALL/AML methodology, from Golub and Slonim et al., 1999 pipeline
- IlluminaDASLPipeline
creates a GenePattern gct file from raw Illumina scan data pipeline
- ImmPort_FLOCK_Individual_FCS
ImmPort FLOCK and Individual FCS pipeline pipeline
- job212786
describe it here pipeline
- job437446
describe it here pipeline
- MGED_Reich
test pipeline pipeline
- PWRGPTTestAuto_InheritType_Vis
Automated pipeline with file input as stored path (ie saved with the pipeline) and text inputs.... pipeline
- Rot13Madness
- Beroukhim.Getz.2007.PNAS.Glioma.GI! pipeline
- CopyNumberInferencePipeline.Part2of...
Second half of Pipeline for processing SNP 6 data pipeline
- CufflinksCuffmergePipeline
Beta Release Contact gp-help with any issues. Check stdout.txt and stderr.txt for errors|Creates... pipeline
- GetDataSetInSilico
downloads a compressed .tgz file from the Insilico servers and extract it pipeline
- Golub_Slonim
ALL/AML methodology, from Golub and Slonim et al., 1999 pipeline
- ImmPort_FLOCK_CrossSample
ImmPort FLOCK and CrossSample pipeline pipeline
- job212108
describe it here pipeline
- job298686
describe it here pipeline
- Lu.Getz.Miska.Nature.June.2005.mous...
Normal/tumor classifier and kNN prediction of mouse lung samples LuGetzMiska.Nature.2005.Suite, pipeline
- ParallelICBS
Runs CBS algorithm on multiple samples in parallel pipeline
- RNaseQC_CEGS
pipeline
- ScripturePipeline

Resources **Help** **GenomeSpace**

Modules & Pipelines Suites Job Results

Modules Jobs Files

Search Modules & Pipelines

No Jobs Processing Browse Modules >

Favorite Modules

Drag Modules Here

Recent Modules

- ComparativeMarkerSelectionViewer
- FeatureSummaryViewer
- HeatMapView
- HierarchicalClusteringViewer

2 MB

you through the new features. See the release notes for more

ta as a heat map.

wn classes.

jet/normal samples.

About GenePattern | Contact Us ©2003-2014 Broad Institute, MIT

Community Activity

- 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

- GParc: GenePattern community repository

- ~100 community-contributed methods
- CRISPR analysis
- Bisulfite sequencing
- Flow cytometry
- RNAi screens

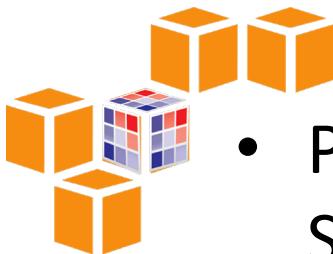
www.genepattern.org

The screenshot shows the GenePattern homepage with a dark blue header featuring the logo and navigation links: Run, Learn, Modules, Analytics, Resources, Contact, Help, and Search. Below the header, there's a banner with three circular icons: 'Use GenePattern', 'GenePattern Basics', and 'Community'. A sidebar on the right contains sections for 'Features' (describing a user-friendly interface for genomics tools), 'New: GenePattern Notebooks' (mentioning a notebook environment), and 'Analysis Pipelines' (describing how GenePattern integrates with other tools). At the bottom, there's a news section titled 'Bugs > GP Updates' with a list of recent changes.

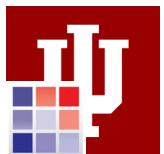
www.gparc.org

The screenshot shows the GParc homepage with a dark blue header featuring the logo and navigation links: Upload a Module, Resources, Contact Us, and Login. Below the header, there's a large graphic of a network graph with colored nodes and connecting lines. A call-to-action button says 'Upload a module'. A 'Browse Modules' section includes a search bar, a 'Search Modules...' input field, and buttons for 'Search', 'Show All', and 'Show Most Recent Uploads'. To the right, there's a 'Filter By Available Tags' sidebar with several buttons: 'View All Tags', 'Annotation', 'Bisulfite Sequencing', 'BisulfiteConversion', 'Clustering', 'ConceptualMarkerSelection', 'ConceptualMarkerConversion', and 'CRISPR'. Below the search section, there's a card for the 'Acgh2Tab v4' module, which converts acgh files to tab-delimited format usable by Genomics.

Availability



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

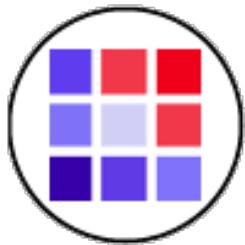


- Public server at Indiana U, backed by Carbonate HPC cluster (gp.indiana.edu)



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

	A	B	C	D	E	F	G	H	I
Always #1.2	#1.2								
Col 1: Row identifiers. Typically gene ids.	60483	30							
Col 2: Optional description	Name	Description	A7-A13F-primary	A7-A13F-normal	A7-A13G-primary	A7-A13G-normal	AC-A23H-primary	AC-A23H-normal	AC-A2FB-primary
Col 3+: Sample names, must be unique.	ENSG000000	ENSG000000	2904	3276	1009	4611	614	7362	4546
Each column contains expression or count values from one sample.	ENSG000000	ENSG000000	2	2159	65	869	0	234	10
	ENSG000000	ENSG000000	6070	1295	547	1896	4011	1813	1663
	ENSG000000	ENSG000000	1263	1178	3914	1262	4928	1684	2451
	ENSG000000	ENSG000000	1362	277	799	256	954	390	825
	ENSG000000	ENSG000000	330	480	289	555	239	471	905
	ENSG000000	ENSG000000	1090	9090	1790	9020	3629	13575	3762
	ENSG000000	ENSG000000	3897	2651	1391	3555	5525	4534	3600
	ENSG000000	ENSG000000	2248	2249	1493	6008	2352	2856	3482
	ENSG000000	ENSG000000	2735	1934	2642	1975	4098	2814	2922
	ENSG000000	ENSG000000	284	636	799	631	907	774	925
	ENSG000000	ENSG000000	597	3294	5708	1877	2589	5229	5003
	ENSG000000	ENSG000000	4201	2010	986	1911	4783	2391	2921
	ENSG000000	ENSG000000	4471	1864	1355	3751	1669	1636	2813
	ENSG000000	ENSG000000	10532	2118	2470	1666	6469	1877	8759
	ENSG000000	ENSG000000	1	30	33	38	28	80	22
	ENSG000000	ENSG000000	5099	3305	2760	5474	7075	4630	5803
	ENSG000000	ENSG000000	481	107	52	105	1163	146	370

- Tab Delimited
- Open in any spreadsheet or text editor

Sample labels: CLS file

- 30 samples

- 2 Classes

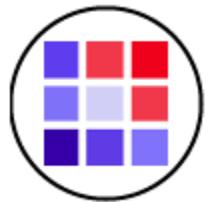
- Numeric values, start with 0

- Tab or space delimited

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

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

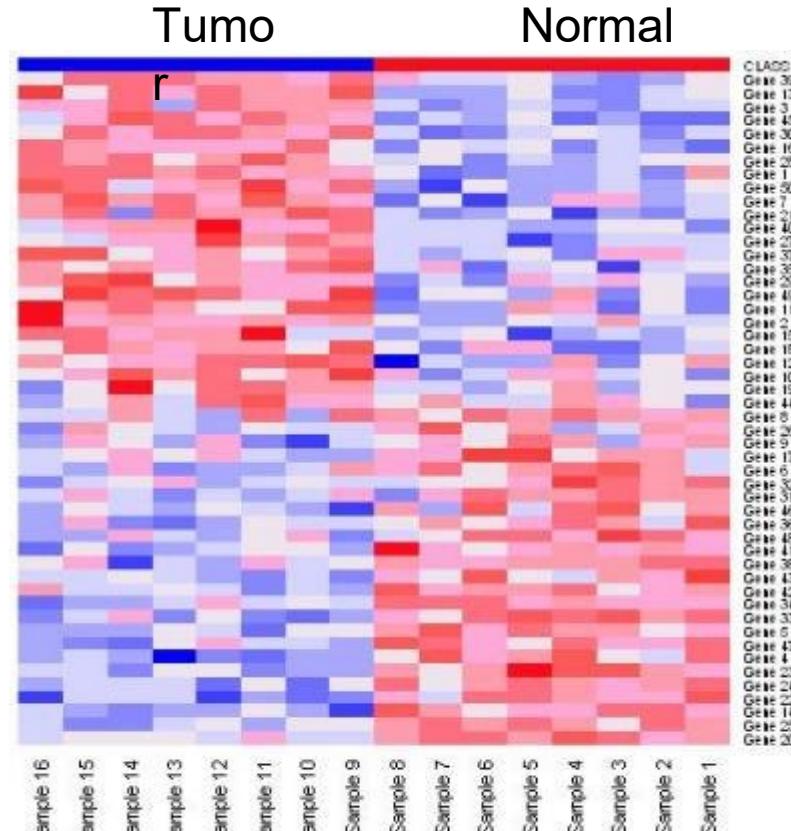
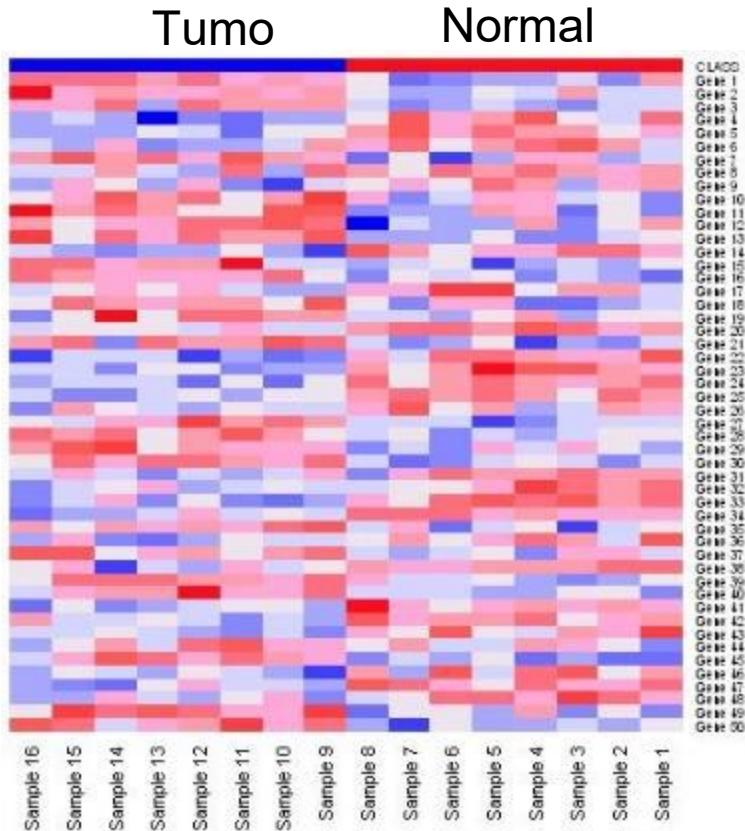


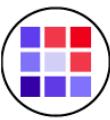
Differential Expression Analysis

2/5

Marker selection

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

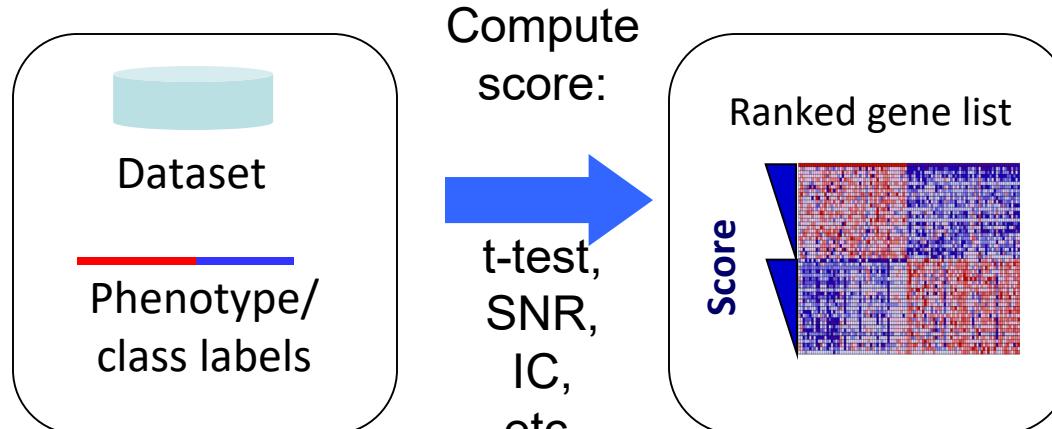




Gene Marker Selection

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.

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

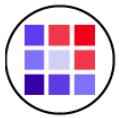
Information coefficient (IC)

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

$$IC(t, s_k)$$

$$IC(X, Y) = sign(\rho(X, Y)) \sqrt{1 - e^{-2MI(X, Y)}}$$

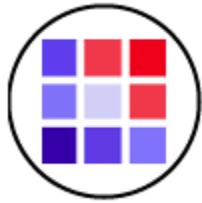
$$MI(X, Y) = \iint_{H(t)} p(x, y) \log \left(\frac{H(s_k)(x, y)}{p(x)p(y)} \right) dx dy$$



Differential Analysis Exercise

Open notebook:

BioITWorld Differential Analysis



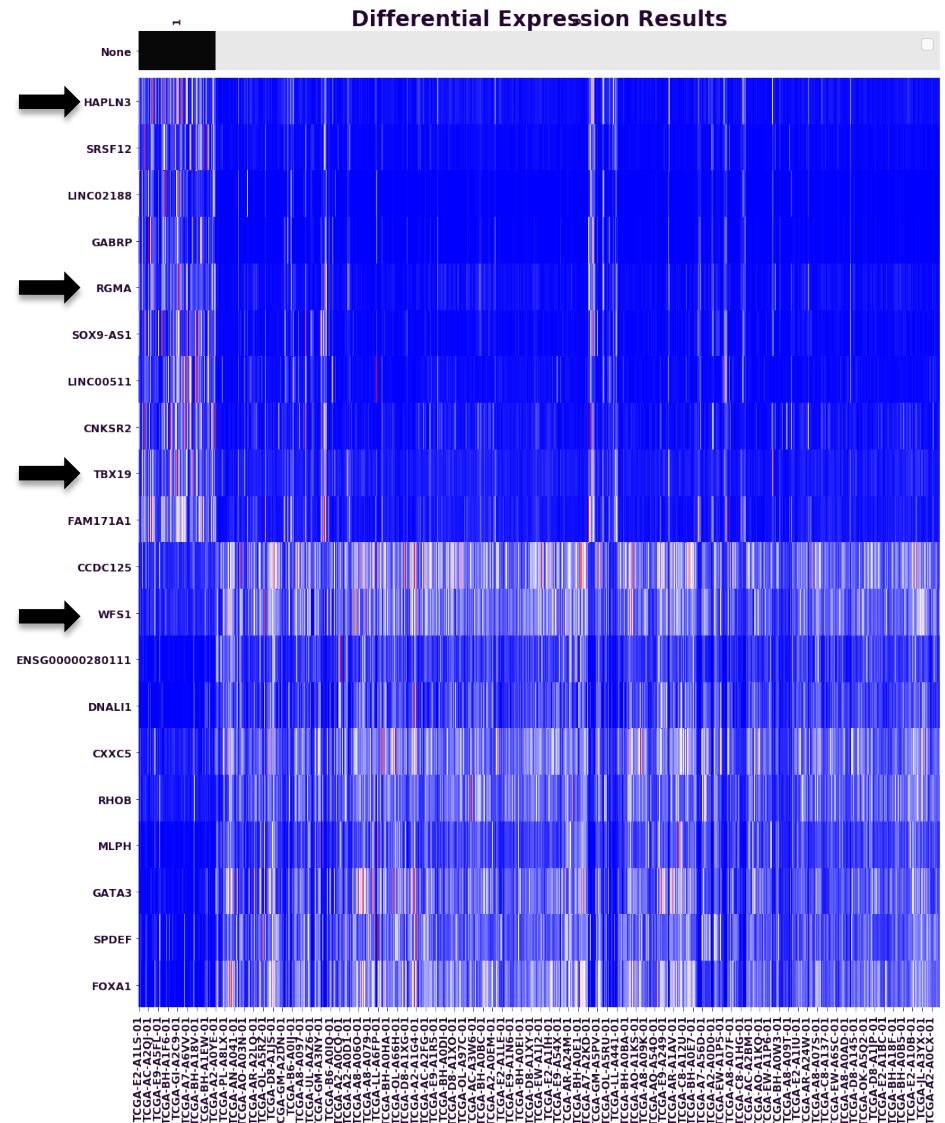
Gene Set Enrichment Analysis (GSEA)



GSEA in a Nutshell

2

- 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”)

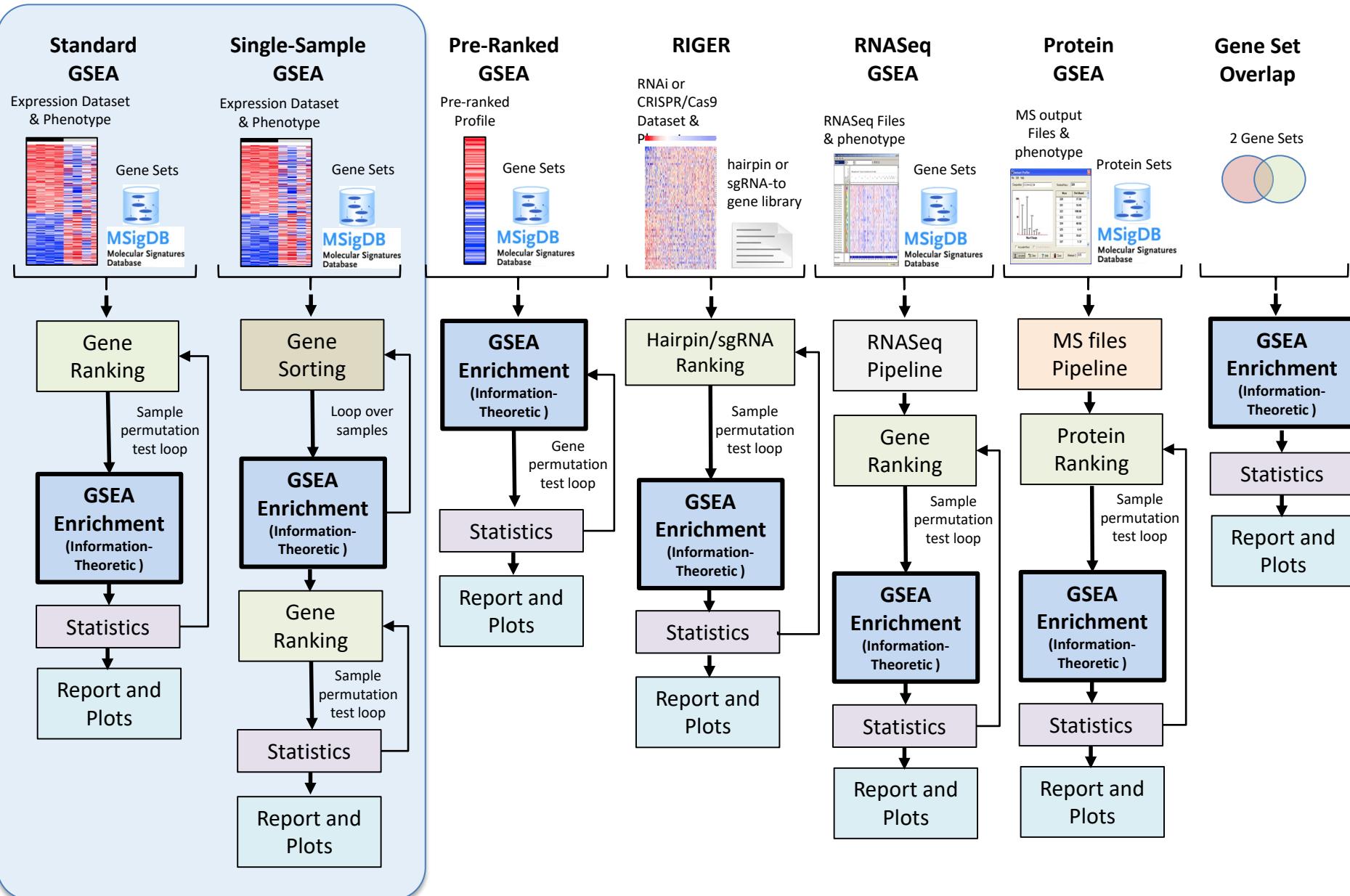


Single-Sample GSEA (ssGSEA)

- 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

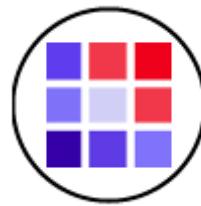
More details in : <http://software.broadinstitute.org/gsea> or msigdb.org





ssGSEA Exercise

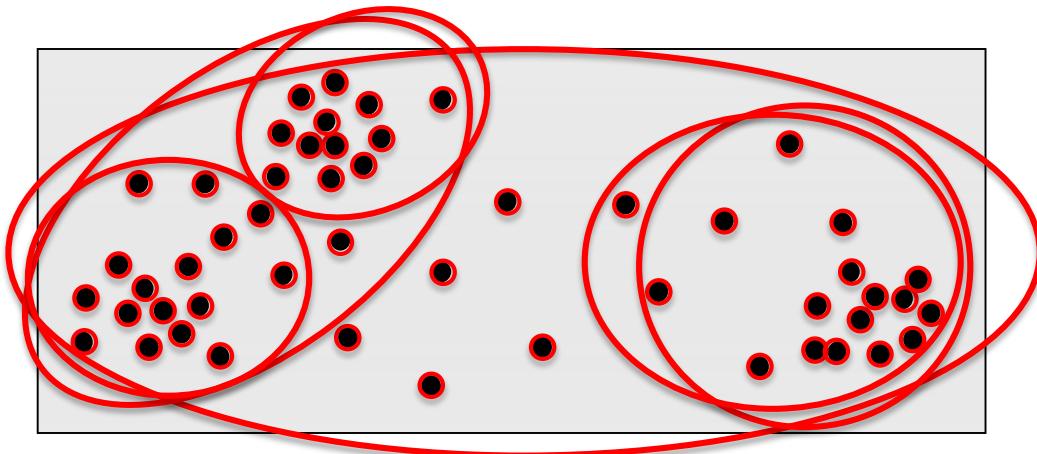
2019-04-16_04 BioITWorld ssGSEA



Clustering

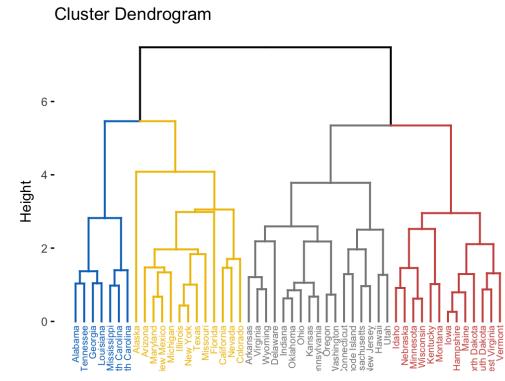
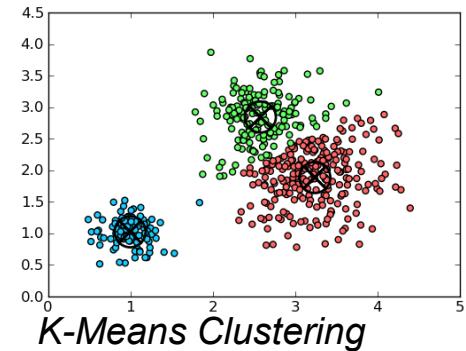
Clustering/Class Discovery

- **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

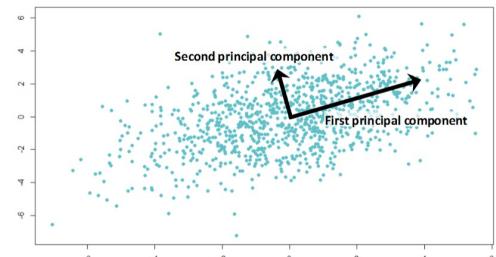


Clustering in GenePattern

- Representative based:
Find representatives/centroids of the dataset
 - K-means
 - Self Organizing Maps (SOM)
- Bottom-up (Agglomerative)
Create an ordering of the data by closeness
 - 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)



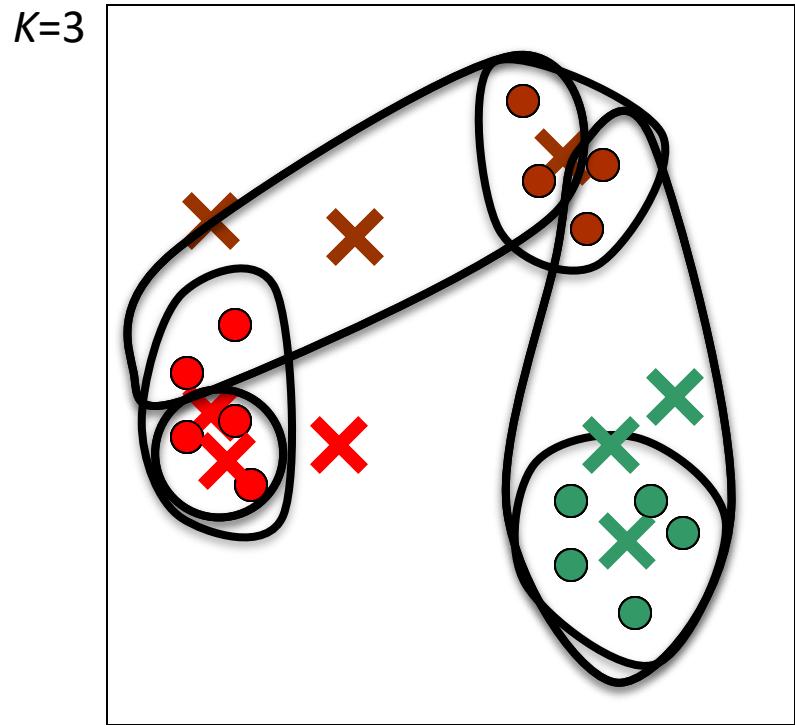
Hierarchical Clustering



Principal Components Analysis

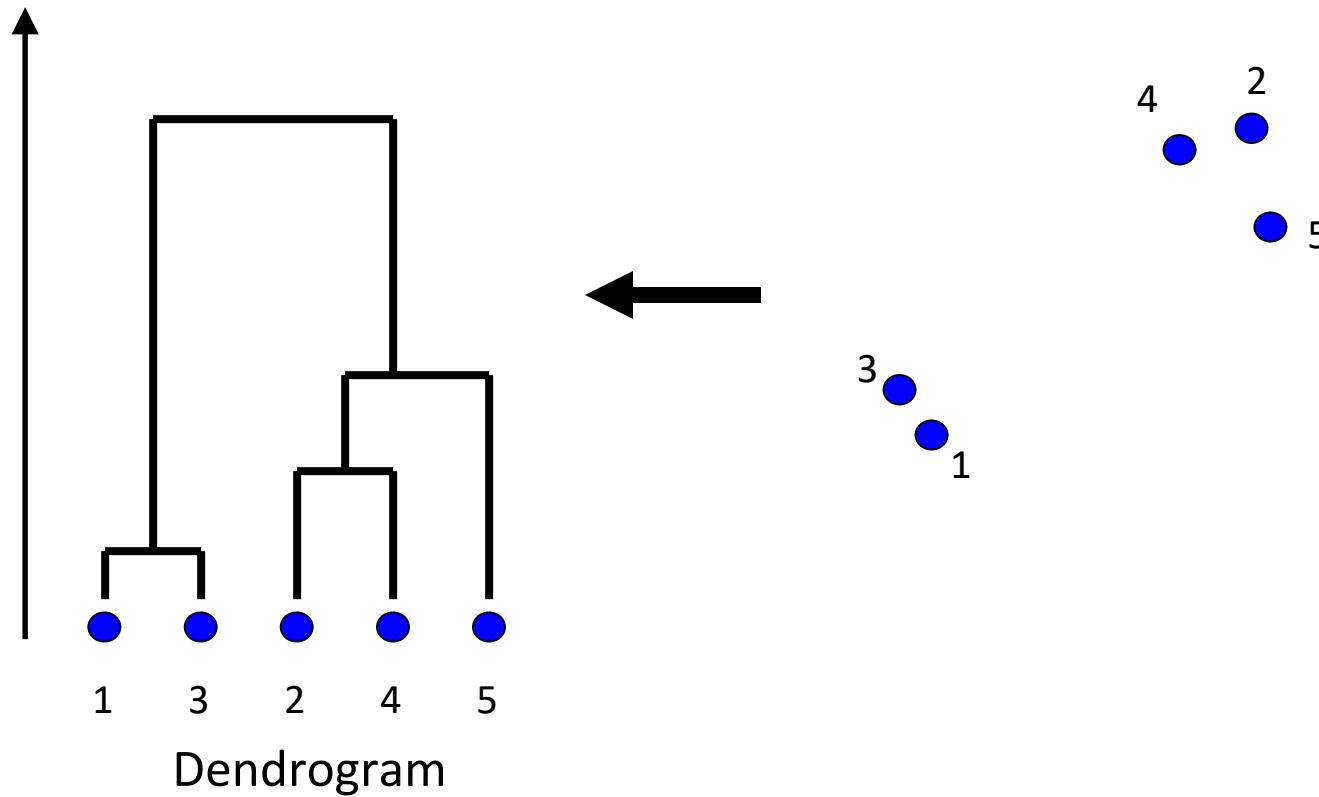
K-means Clustering

- 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



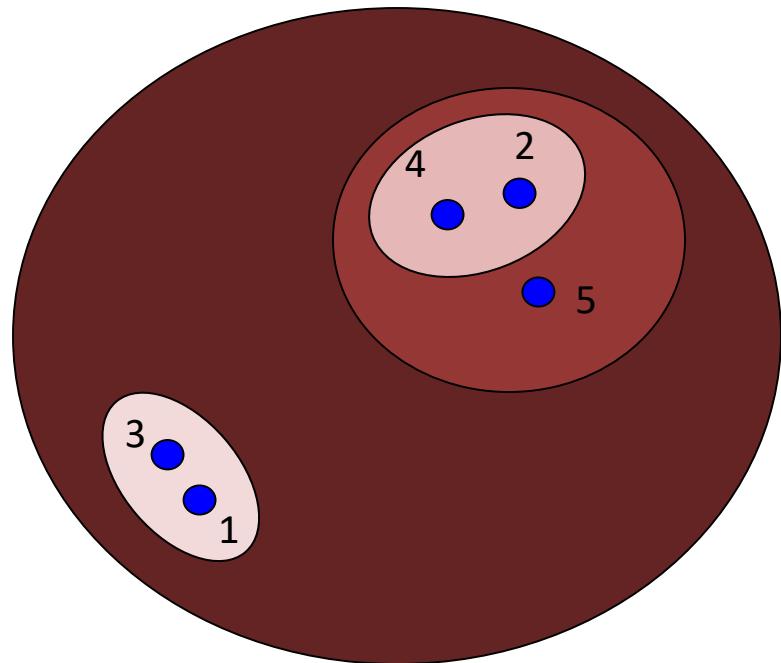
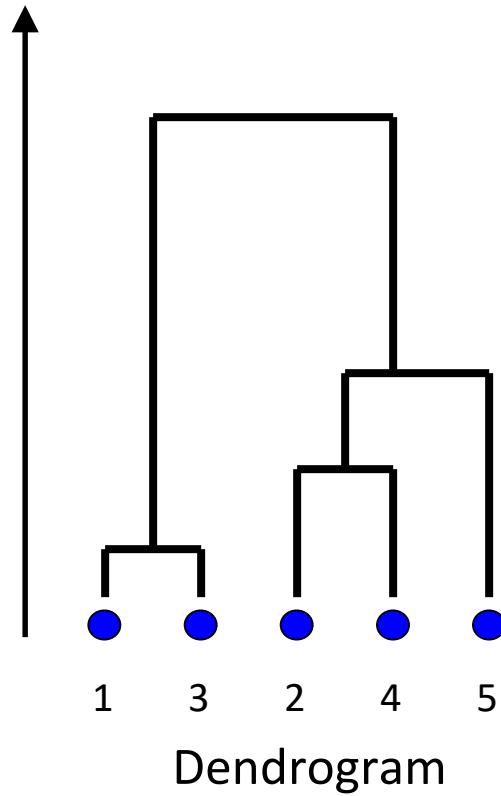
Hierarchical Clustering

Distance between joined clusters



Hierarchical Clustering

Distance between joined clusters

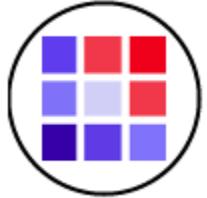


Linkage is the method for linking clusters based on the **distance** between them.

Average Linkage: average distance between all pairs

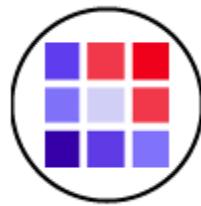
Complete Linkage: farthest distance between all pairs

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.



Recognizing differences

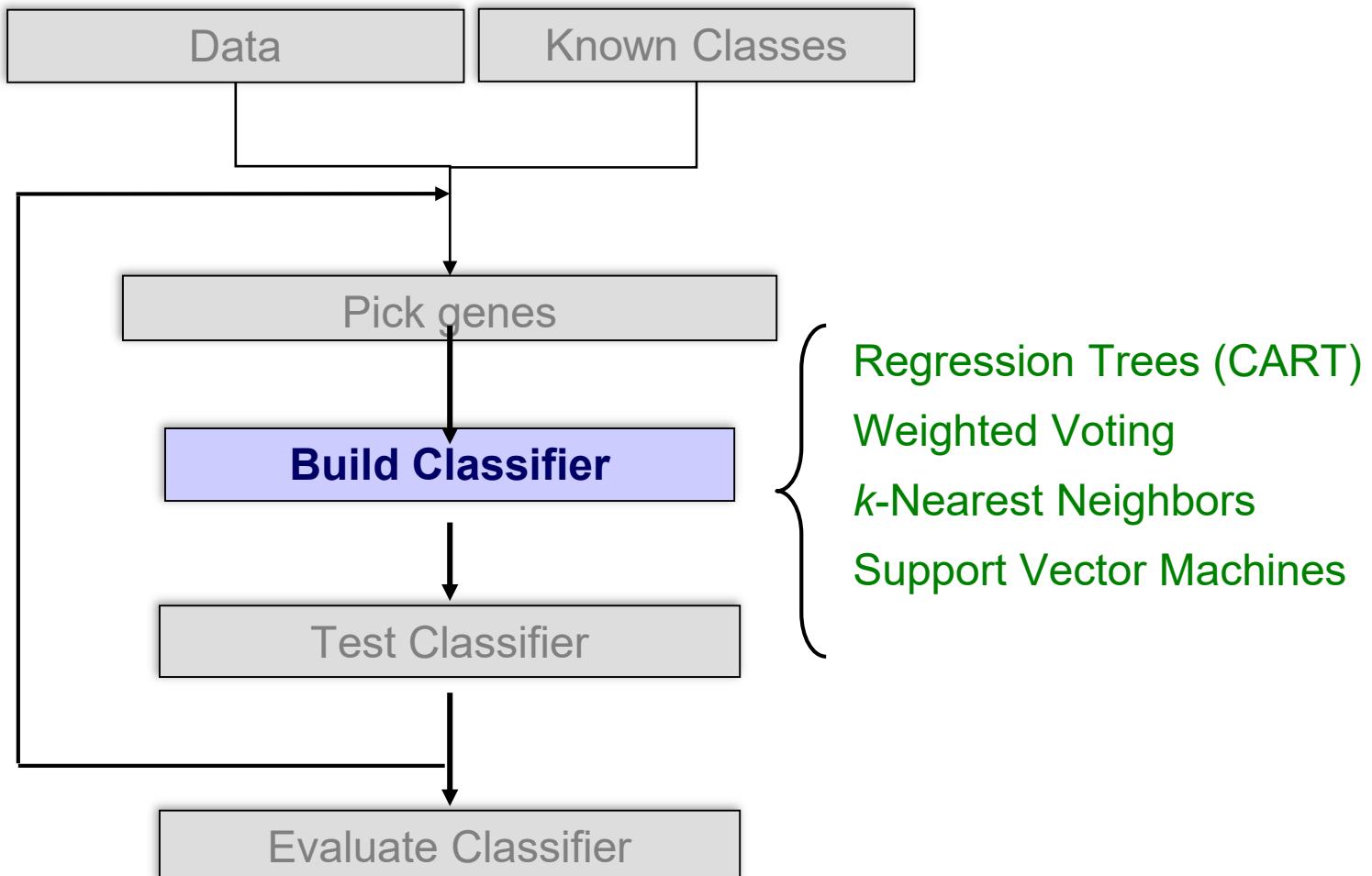


What Is a Classifier

- 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

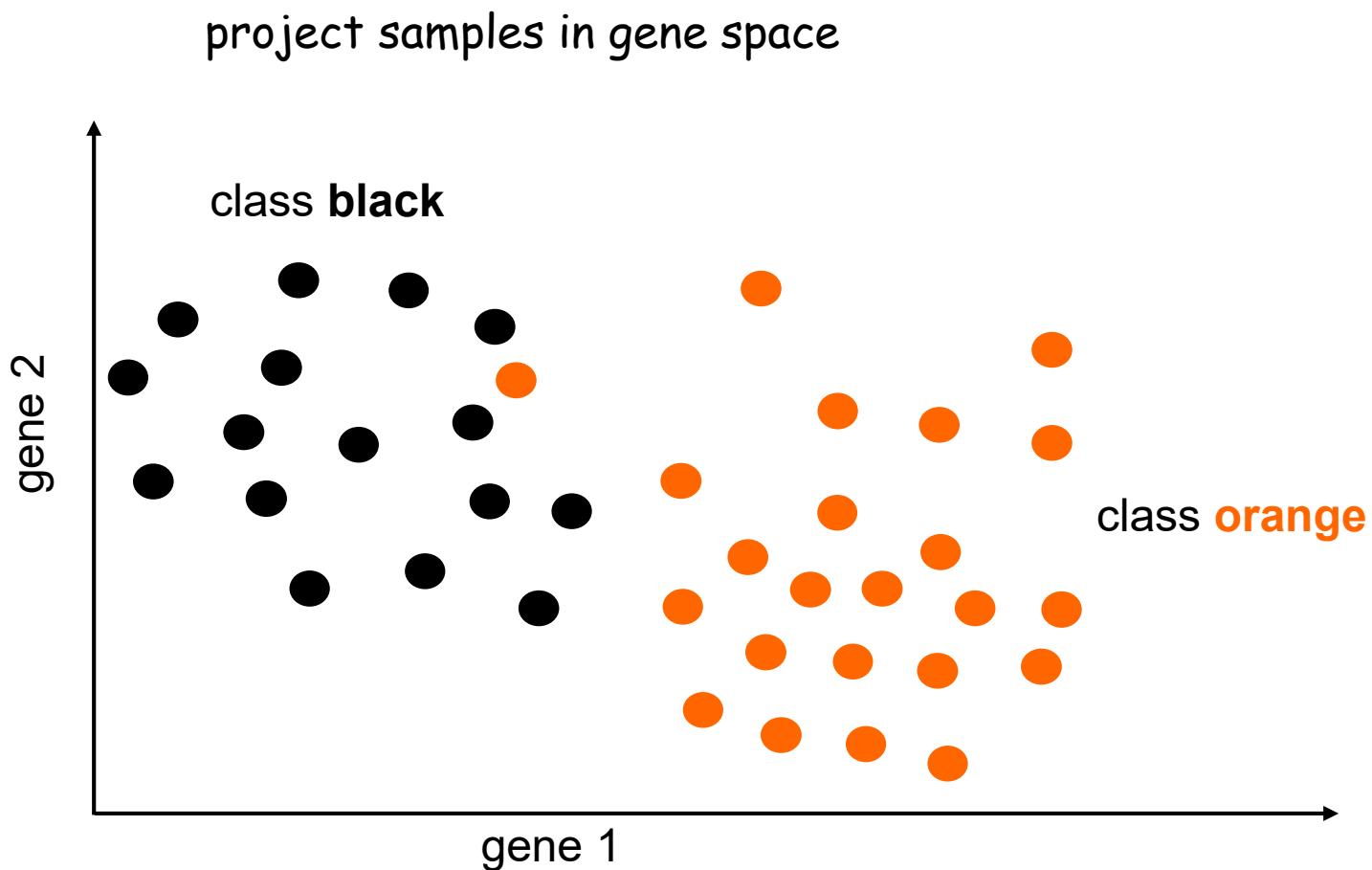
Important issues:

- 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

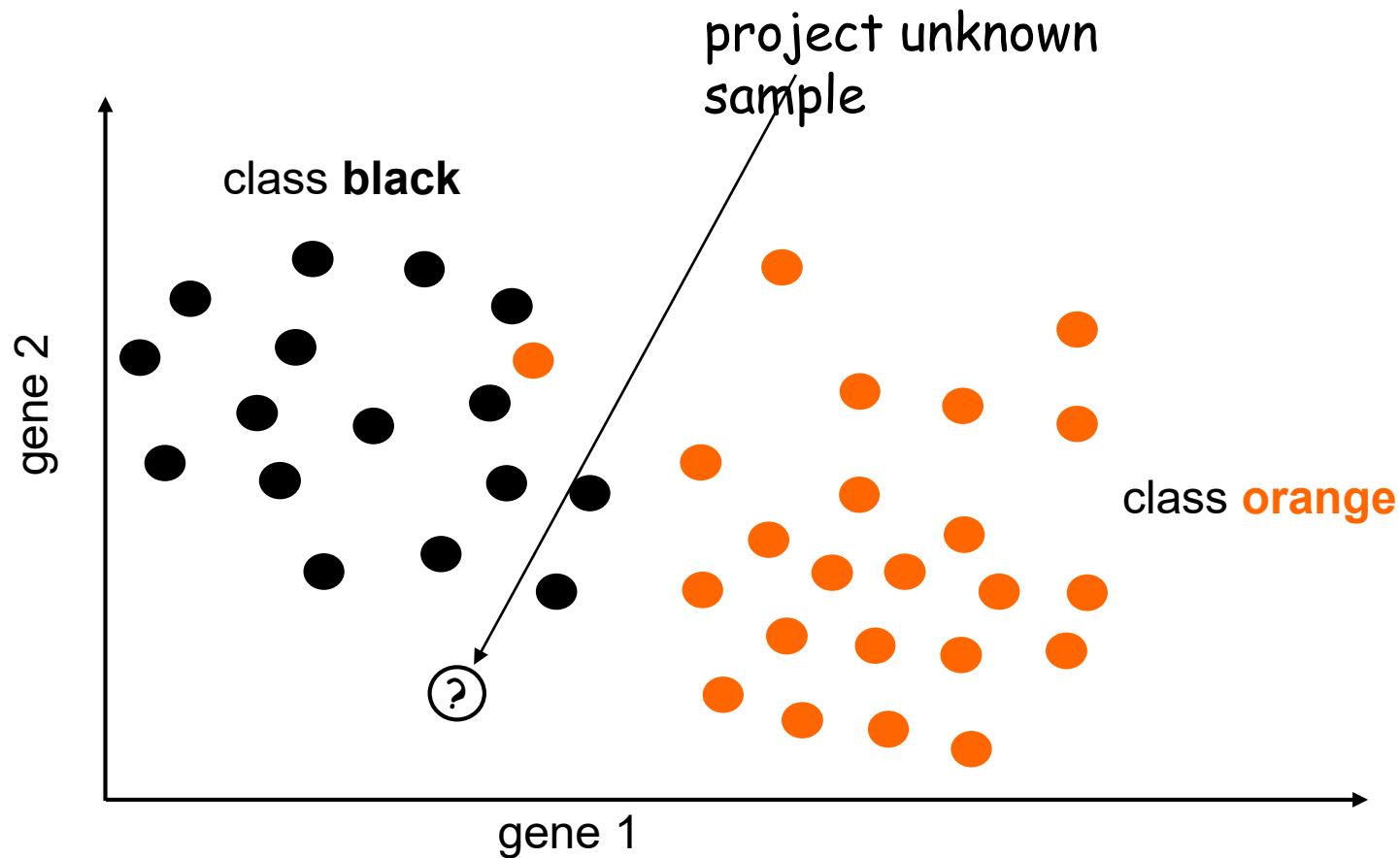
k Nearest Neighbors (kNN) Classifier

Example: k=5, 2 genes, 2 classes



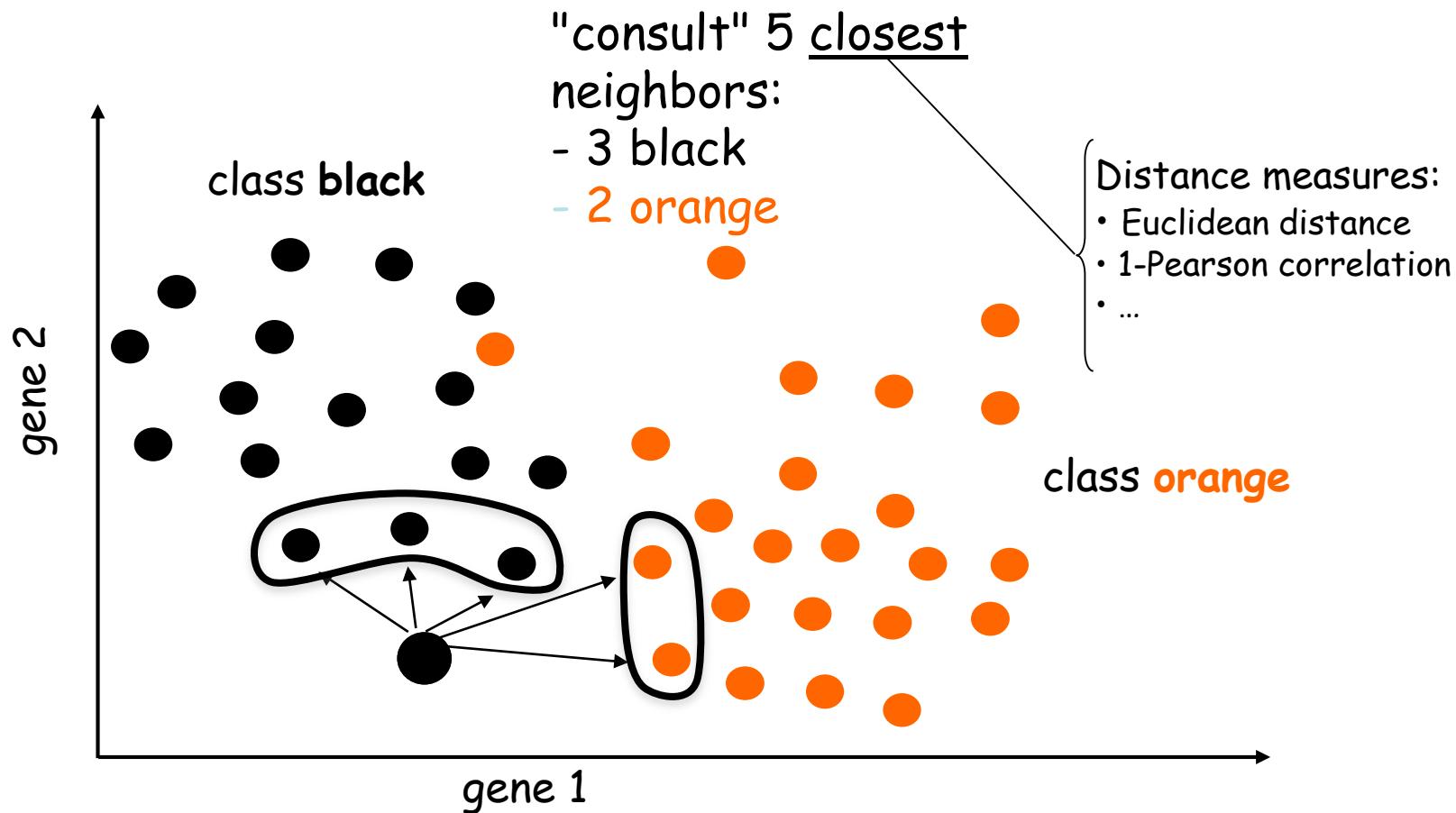
kNN Classifier

Example: k=5, 2 genes, 2 classes

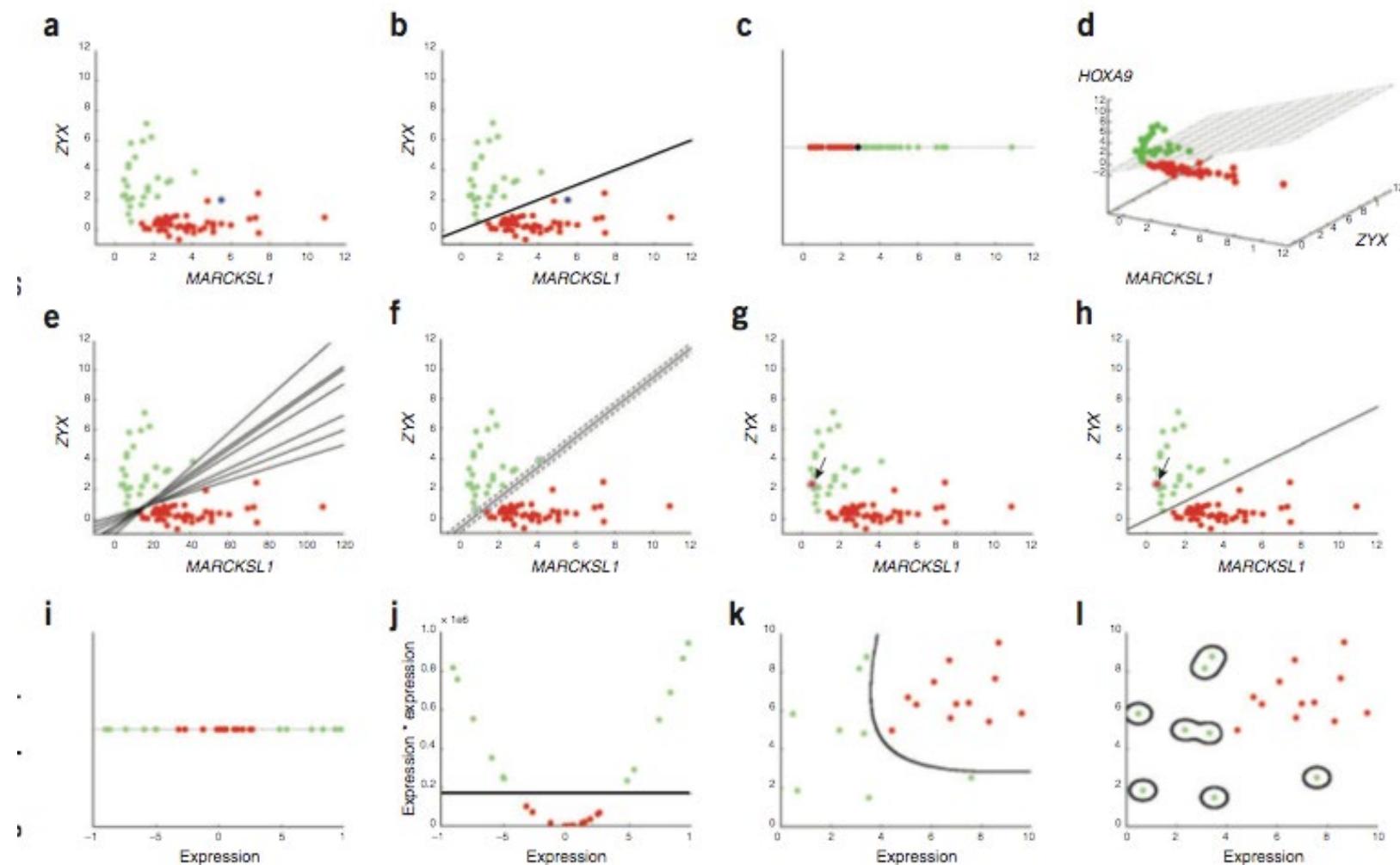


kNN Classifier

Example: K=5, 2 genes, 2 classes



Support Vector Machine (SVM)

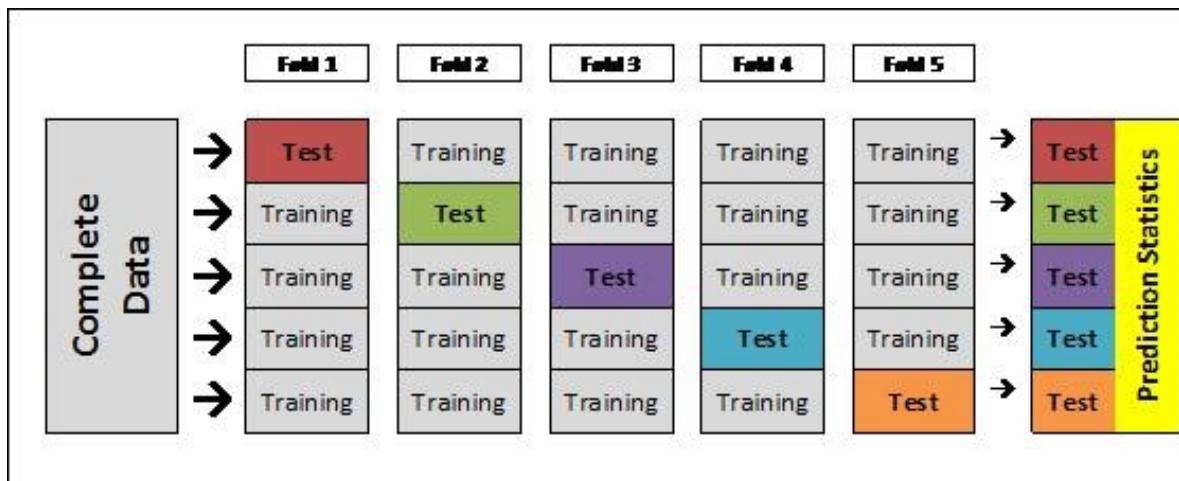


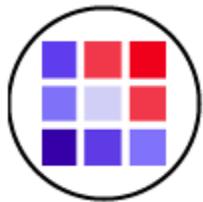
Noble, Nat Biotech 2006

Testing the Classifier

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

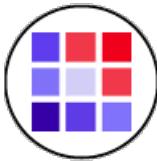
- Evaluation on independent test set
 - Build the classifier on the train set.
 - Assess prediction performance on test set.
 - What if we don't have an independent test set?
- Cross Validation (XV):
 - Split the dataset into n folds (e.g., 10 folds of 10 samples each).
 - For each fold (e.g., for each 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.



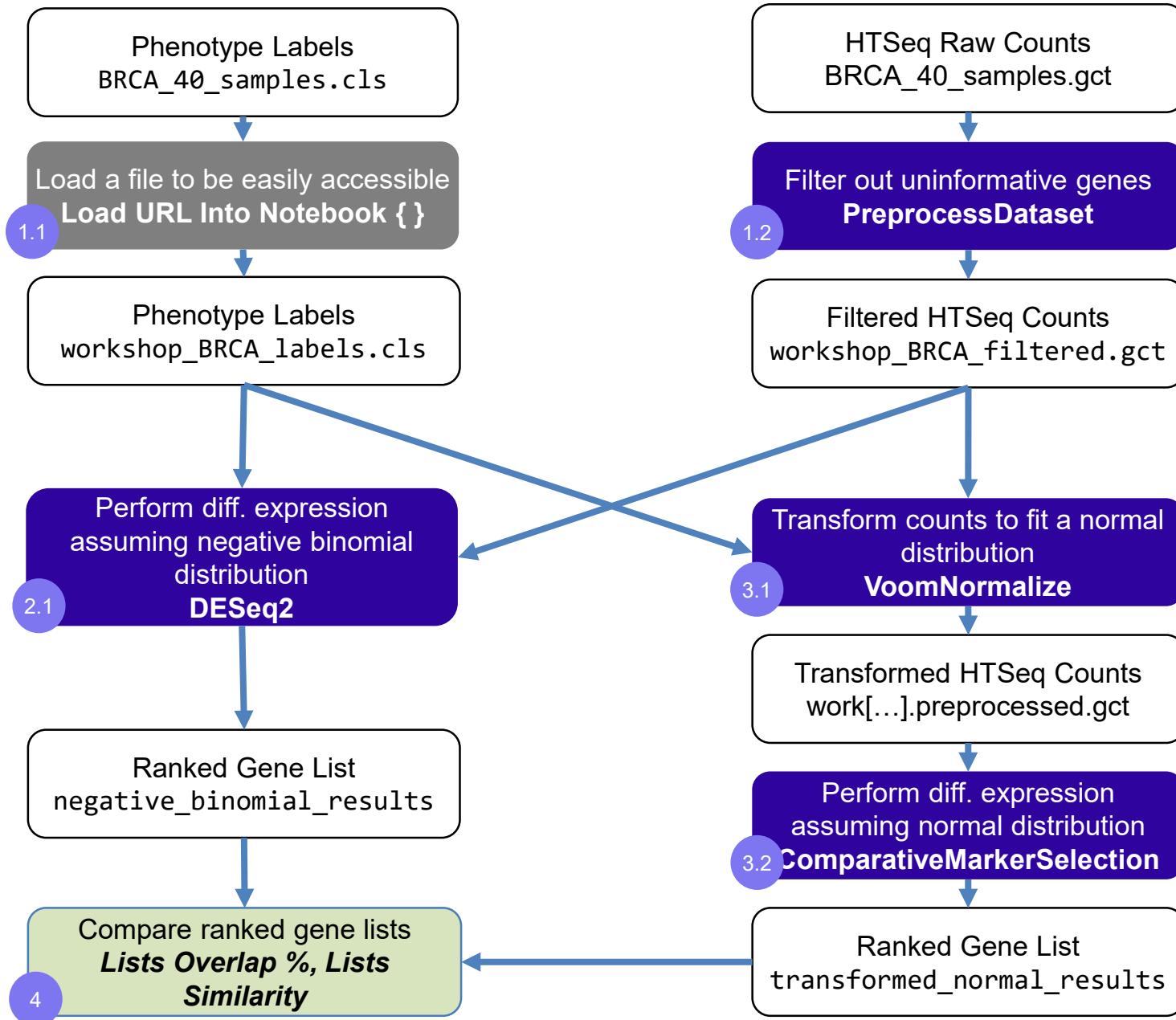


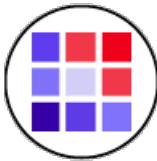
Classification and Prediction Exercise

BioITWorld Classification and Prediction



Analyzing HTSeq Data Using GenePattern





Analyzing HTSeq Data Using GenePattern

