Intro. to Mini Project 2, Hierarchical Clustering and Regression Examples

ECE/CS 498 DS U/G

Lecture 10

Ravi K. Iyer

Dept. of Electrical and Computer Engineering

University of Illinois at Urbana Champaign

Announcements

- HW2 due on Friday, Feb 22
- MP1 presentation on Friday, Feb 22
  - 10 min per group; all group members must be present
  - Watch Piazza for information on location
- No Discussion Section on Friday (MP1 Presentations)

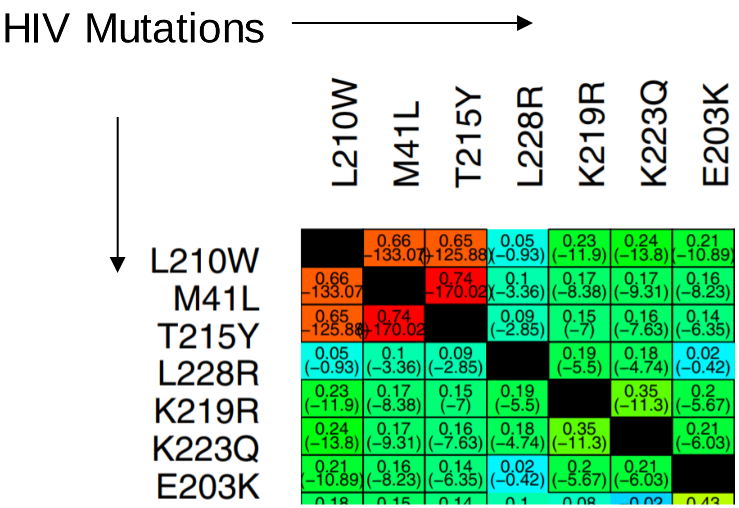
- Today:
  - Examples: Hierarchical clustering and Regressions
  - MP 2 analysis of Breast Cancer Data
  - Start on Princ. Comp Analysis (PCA)

Hierarchical Clustering Example 1

Characterization Novel HIV Drug Resistance Mutations using Clustering

Reference: Sing, Tobias, et al. "Characterization of novel HIV drug resistance mutations using clustering, multidimensional scaling and SVM-based feature ranking." *European Conference on Principles of Data Mining and Knowledge Discovery.* Springer, Berlin, Heidelberg, 2005.

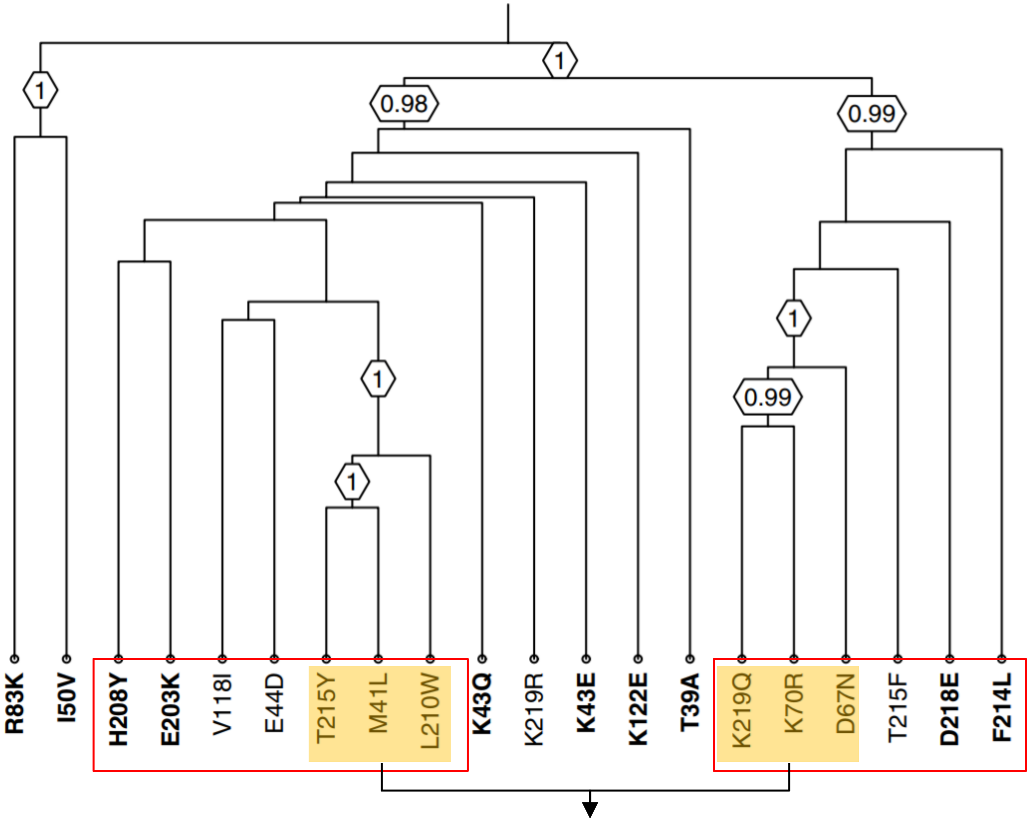
- **Objective:** By clustering new HIV mutations with HIV mutations that have known resistance mechanisms, we can infer the possible resistance mechanisms of the new mutations
- **Clustering Technique:** Agglomerative Hierarchical Clustering using Average-link
- **Distance Metric:** Matthews correlation coefficient



	L210W	M41L	T215Y	L228R	K219R	K223Q	E203K
L210W		0.66 (−133.07)	0.65 (−125.88)	0.05 (−0.93)	0.23 (−11.9)	0.24 (−13.8)	0.21 (−10.89)
M41L	0.66 (−133.07)		0.74 (−170.02)	0.1 (−3.36)	0.17 (−8.38)	0.17 (−9.31)	0.16 (−8.23)
T215Y	0.65 (−125.88)	0.74 (−170.02)		0.09 (−2.85)	0.15 (−7)	0.16 (−7.63)	0.14 (−6.35)
L228R	0.05 (−0.93)	0.1 (−3.36)	0.09 (−2.85)		0.19 (−5.5)	0.18 (−4.74)	0.02 (−0.42)
K219R	0.23 (−11.9)	0.17 (−8.38)	0.15 (−7)	0.19 (−5.5)		0.35 (−11.3)	0.2 (−5.67)
K223Q	0.24 (−13.8)	0.17 (−9.31)	0.16 (−7.63)	0.18 (−4.74)	0.35 (−11.3)		0.21 (−6.03)
E203K	0.21 (−10.89)	0.16 (−8.23)	0.14 (−6.35)	0.02 (−0.42)	0.2 (−5.67)	0.21 (−6.03)	
	0.18	0.15	0.14	0.1	0.08	−0.02	0.43

Hierarchical Clustering Example 1

- **Dendrogram after clustering**



Mutations with known drug resistance mechanisms

Hierarchical Clustering Example 2

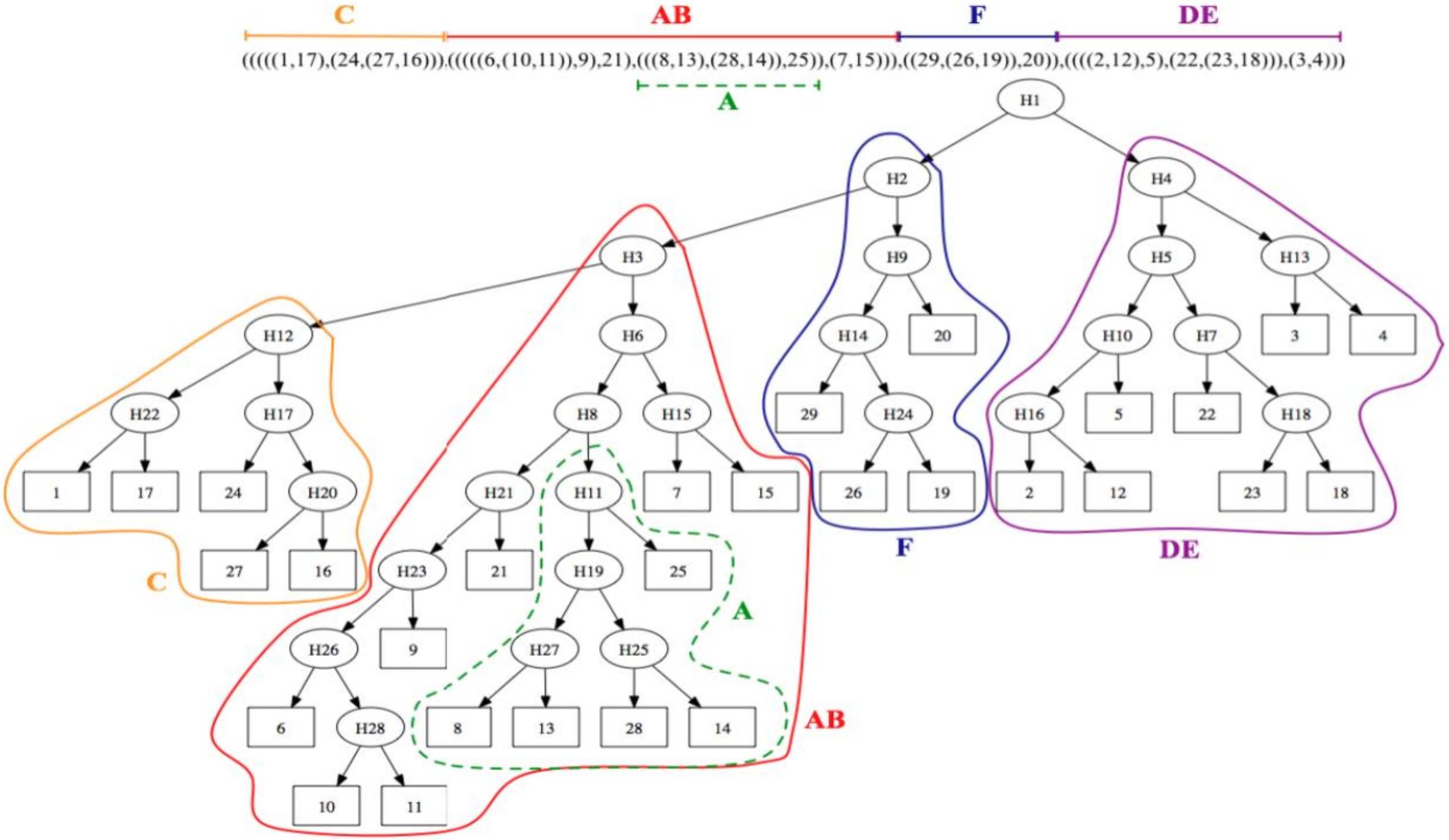
Hierarchical Clustering of Phylogenetic Trees for Evolution

Reference: Blanchette, Glenn, Richard O'Keefe, and Lubica Benuskova. "Inference of a phylogenetic tree: hierarchical clustering versus genetic algorithm." *Australasian Joint Conference on Artificial Intelligence*. Springer, Berlin, Heidelberg, 2012.

- Uses an artificial organism *Caminalcules* invented by Joseph Camin (1965) Univ. of Kansas (Camin + animalcule)
- **Objective:** To chart the evolution of the organism using hierarchical clustering on 29 species and generate a Phylogenetic Tree.
- **Clustering Technique:** Agglomerative Hierarchical Clustering using Average-link
- **Distance Metric:** Fitch/Hamming Distance (Minimum number of substitutions required to change one string into the other). Here the string corresponds to genetic sequence.

Hierarchical Clustering Example 2

- **Phylogenetic Tree**
  - Clustering matches the classifications published by Sokal (1983)
  - There is no single tree that is perfect.

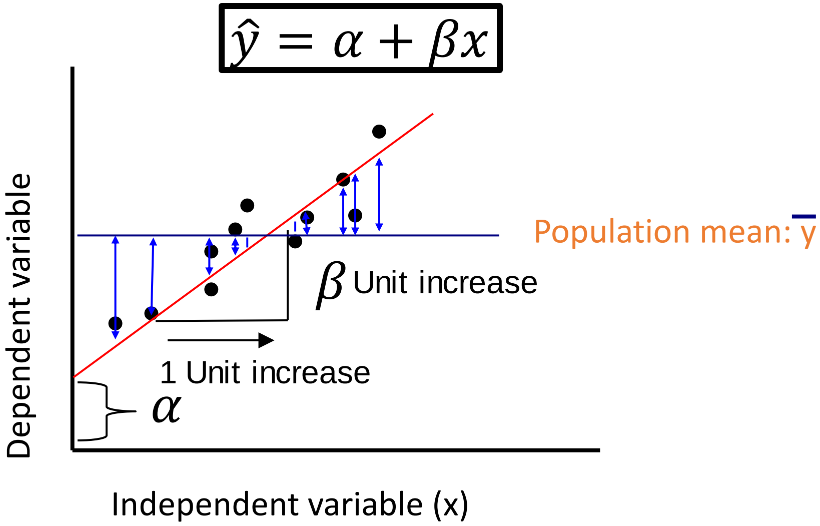


Logistic Regression Example – Donner Party

In 1846, the Donner and Reed families left Springfield, IL for California by covered wagon. In July, the Donner Party, as it became known, reached Fort Bridger, Wyoming. There its leaders decided to attempt a new and untested route to the Sacramento Valley. Having reached its full size of 87 people and 20 wagons, the party was delayed by a difficult crossing of the Wasatch Range and again in the crossing of the desert west of the Great Salt Lake. The group became stranded in the eastern Sierra Nevada mountains when the region was hit by heavy snows in late October. By the time the last survivor was rescued on April 21, 1847, 40 of the 87 members had died from famine and exposure to extreme cold.

From Ramsey, F.L. and Schafer, D.W. (2002). The Statistical Sleuth: A Course in Method of Data Analysis (2nd edition)

Simple Linear Regression



The Total Sum of Squares (SST) is equal to SSR + SSE.

- $SSR = \sum(\hat{y} - \bar{y})^2$ (measure of explained variation)
- $SSE = \sum(y - \hat{y})^2$ (measure of unexplained variation)
- $SST = \sum(y - \bar{y})^2$ (measure of total variation in y)

- A least squares regression selects the line with the lowest total sum of squared prediction errors, which is referred as Sum of Squares of Error, or SSE
- The Sum of Squares Regression (SSR) is the sum of the squared differences between the prediction for each observation and the population mean
- The proportion of total variation (SST) that is explained by the regression (SSR) is known as the Coefficient of Determination, and is often referred to as r .

$$r^2 = \frac{SSR}{SST} = \frac{1 - SSE}{SST}$$

GLM: Generalized Linear Model

l

A very general way of addressing this type of problem in regression, and the resulting models are called generalized linear models (GLMs). Logistic regression is just one example type of model.

All generalized linear models have the following three characteristics:

1. A probability distribution describing the outcome variable

2. A linear model $\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$

3. A link function that relates the linear model to the parameter of the outcome distribution

$$g(p) = \eta \quad or \quad p = g^{-1}(\eta)$$

Logistic Regression

- A logistic regression is a GLM used to Model a **Binary Categorical variable**
- Assume a Binary variable produced the outcomes and we want to model the prob. *p of success*. Logistic regression is one example. *All GLMs have the following characteristics*
- *A prob dist. Describing the outcome variable*
- *A linear model*

$\eta = \beta 0 + \beta 1X1 + \cdots + \beta nXn$

- *A link function that relates the outcome to parameters of the outcome distribution*
- $g(p) = \eta$ or $p = g{-}1(\eta)$
- A **link function** transforms the probabilities of the levels of a categorical response variable to a continuous scale that is unbounded. Once the transformation is complete, the relationship between the predictors and the response can be modeled with linear **regression**.

Logistic Regression- Link Functions

- Logistic regression is a GLM used to model a binary categorical variable using numerical and categorical predictors.

- We assume a binomial distribution produced the outcome variable and we therefore want to model *p* the *probability of success* for a given set of predictors.
-
- To finish specifying the Logistic model we just need to establish a reasonable link function that connects η to *p*. There are a variety of options but the most commonly used is the logit function.

Logit Function

$$logit(p) = \log\left(\frac{p}{1-p}\right), \text{ for } 0 \le p \le 1$$

- The logit function takes a value between 0 and 1 and maps it to a value between $-\infty$ and ∞

- Inverse logit (logistic) function

$$g^{-1}(x) = \frac{\exp(x)}{1+\exp(x)} = \frac{1}{1+\exp(-x)}$$

- The inverse logit function takes a value between $-\infty$ and ∞ and maps it to a value between 0 and 1

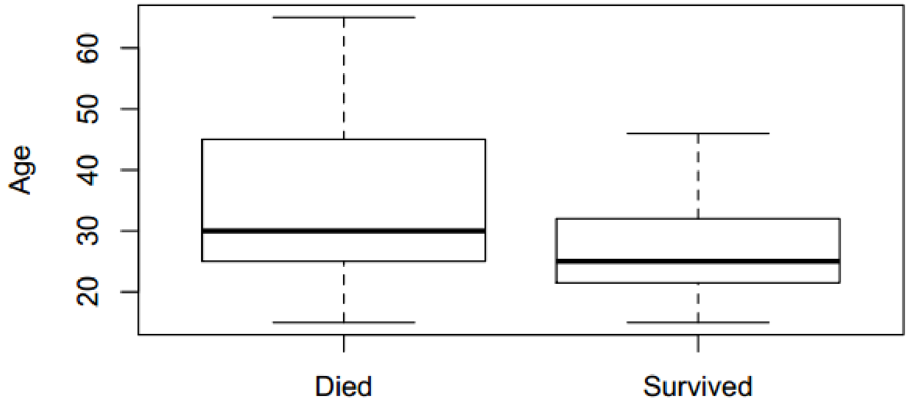
Donner Party - Data

	Age	Sex	Status
1	23.00	Male	Died
2	40.00	Female	Survived
3	40.00	Male	Survived
4	30.00	Male	Died
5	28.00	Male	Died
⋮	⋮	⋮	⋮
43	23.00	Male	Survived
44	24.00	Male	Died
45	25.00	Female	Survived

Data

	Male	Female
Died	20	5
Survived	10	10

Status vs Gender



Status vs Age

Donner Party

- Both Age and Gender have an effect on the survival
- One way to think about the problem – survival (or death) is a success (or failure) arising from a Bernoulli distribution
- Parameter of the Bernoulli distribution depends on Age and Gender
- Logistic Regression allows to do the same!

$$p = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 \times Age + \beta_2 \times Gender))}$$

- Let $x_{i,1}$ represent the Age, $x_{i,2}$ represent the Gender, and p_i represents the probability of survival of person i

$$p_i = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2}))}$$

Donner Party Solution

- Get the following model after training

$$p = \frac{1}{1 + \exp(-(1.63\ - 0.078 \times Age + 1.59 \times Gender))}$$

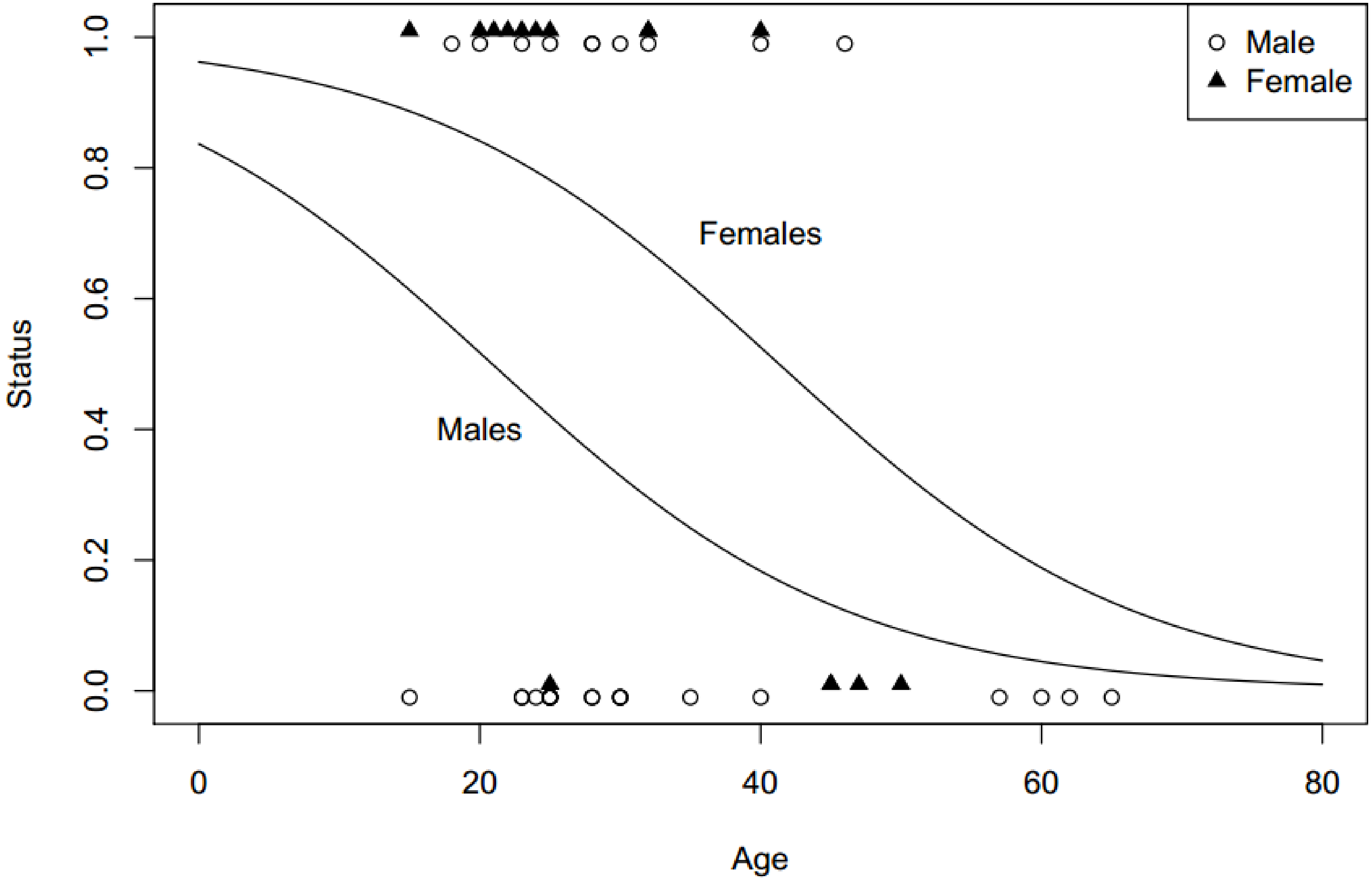
- Male model (Gender = 0)

$$p = \frac{1}{1 + \exp(-(1.63\ - 0.078 \times Age))}$$

- Female model (Gender = 1)

$$p = \frac{1}{1 + \exp(-(1.63\ - 0.078 \times Age + 1.59))}$$
$$= \frac{1}{1 + \exp(-(3.22\ - 0.078 \times Age))}$$

Donner Party: Male and Female models



Moving on!! MP 2 task and Problem Overview

Task description

Problem statement: Investigate how metformin impacts genes that might be responsible for the spreading of triple negative breast cancer

Checkpoint 1: Monday, Feb 25

- Task 0: Getting started; understanding the data

Checkpoint 2: Friday, Mar 8

- Task 1: Data pre-processing and visual inspection
- Task 2: Statistical analysis of genes with significantly altered expression values

Checkpoint 3: Wednesday, Mar 27

- Task 3: Dimensionality reduction and clustering
- Task 4: Statistical analysis of genes with significantly altered expression and interpret the results

Unsupervised Single-Cell Analysis in Triple-Negative Breast Cancer: A Case Study

ILLINOIS

MAYO CLINIC

Arjun P. Athreya, Alan Gaglio, Junmei Cairns, Krishna R. Kalari, Richard Weinshilboum, Liewei Wang, Zbigniew Kalbarczyk and Ravishankar K. Iyer

Univ. of Illinois at Urbana-Champaign
Mayo Clinic

Breast Cancer Precision Medicine

ER/PgR	HER2	TNBC
Endocrine Rx (SERMs and AIs)	Trastuzamab (Herceptin)	Chemotherapy (taxanes-anthracyclines)
Targeted Rx	Targeted Rx	Untargeted Rx

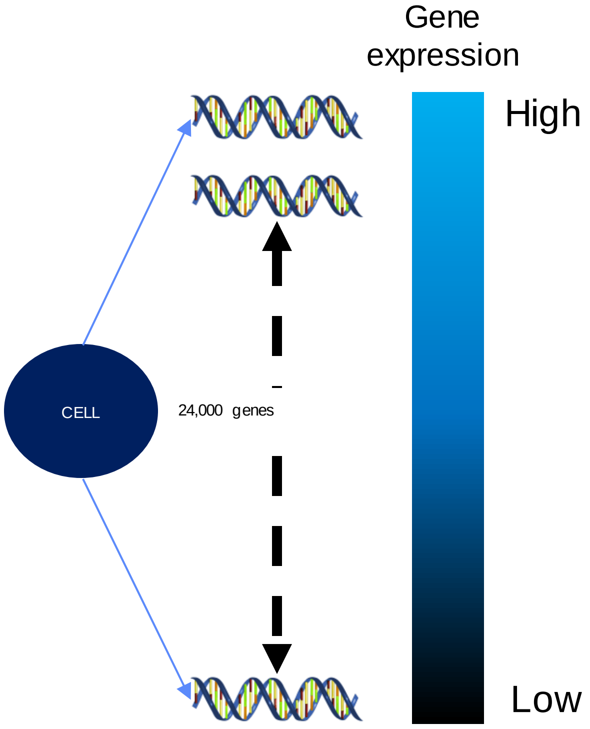
No targeted treatments yet for TNBC

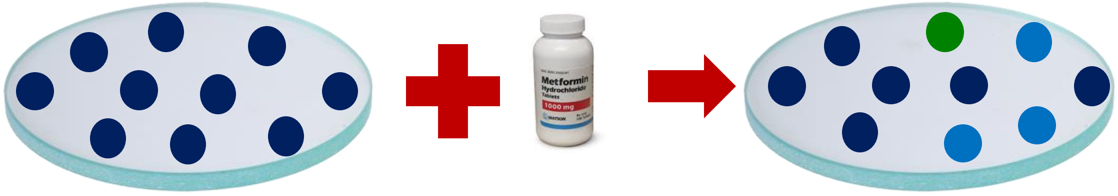
Population studies in 2009: Diabetic patients taking metformin have decreased incidence of cancer, including breast cancer

Question: What are the molecular mechanisms of metformin in TNBC?

Data: Breast Cancer Single-Cells

Single-cell RNAseq





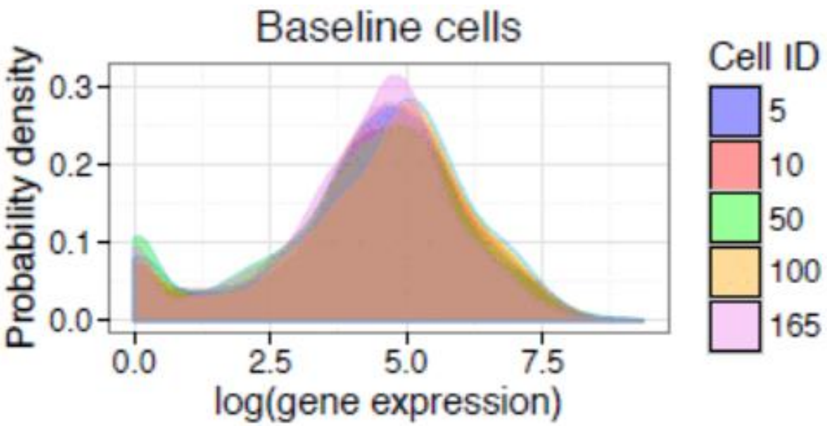
200 Baseline cells

200 Metformin-treated cells

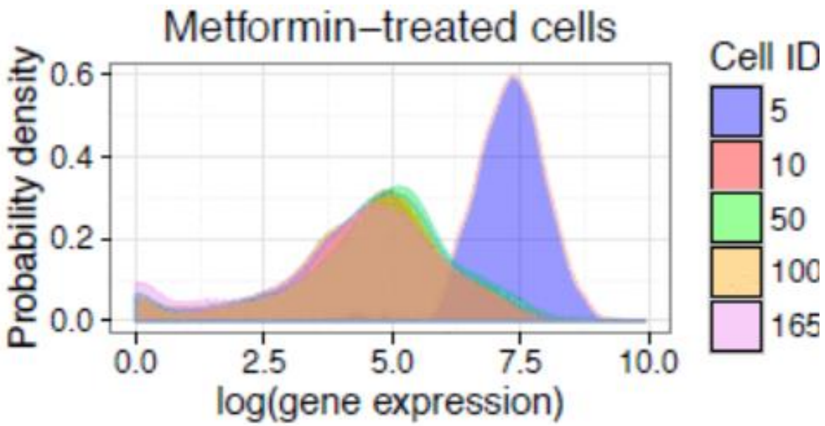
Coverage: 1M gene count per cell

Deploy unsupervised learning to quantify how metformin affects the gene expression using single-cell resolution

Preliminary Analysis: Studying Probability Density Functions



(b) Baseline cells



(c) Metformin-treated cells

Insights:

1) Data's distribution is a **mixture of Gaussians**
2) Some metformin-treated cells have components of the distribution **phase-shifted**

Problem:

Identify cluster of cells which have similar distributions

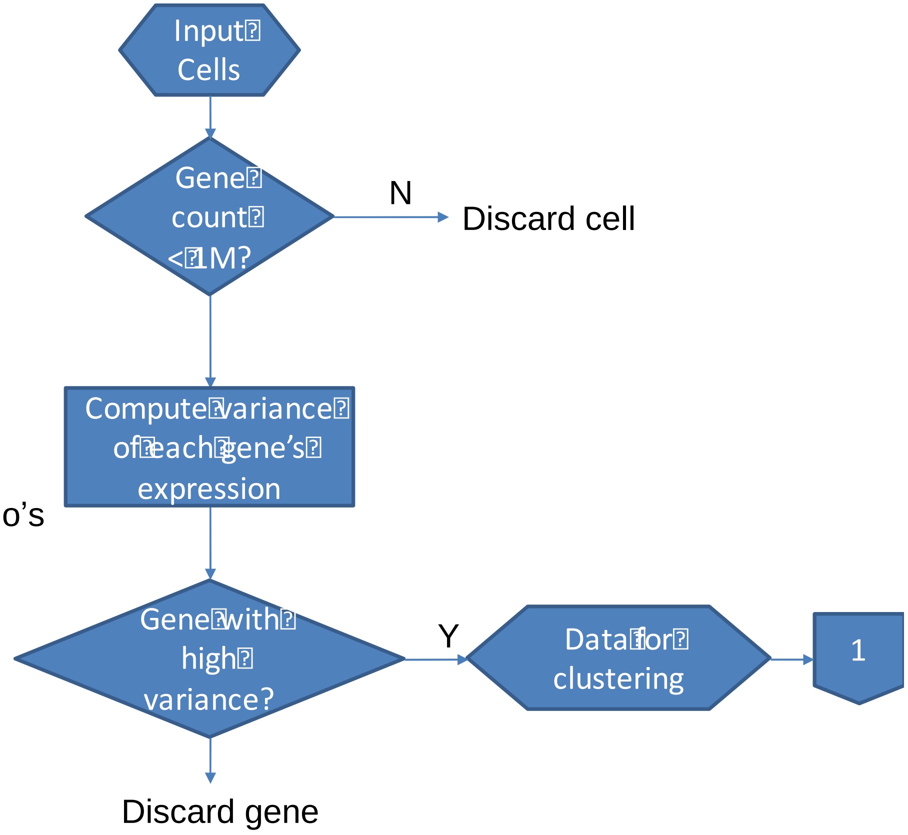
MiMoSA

Mixture-Model based Single-cell Analysis

PHASE 1

Removing cells with low gene counts

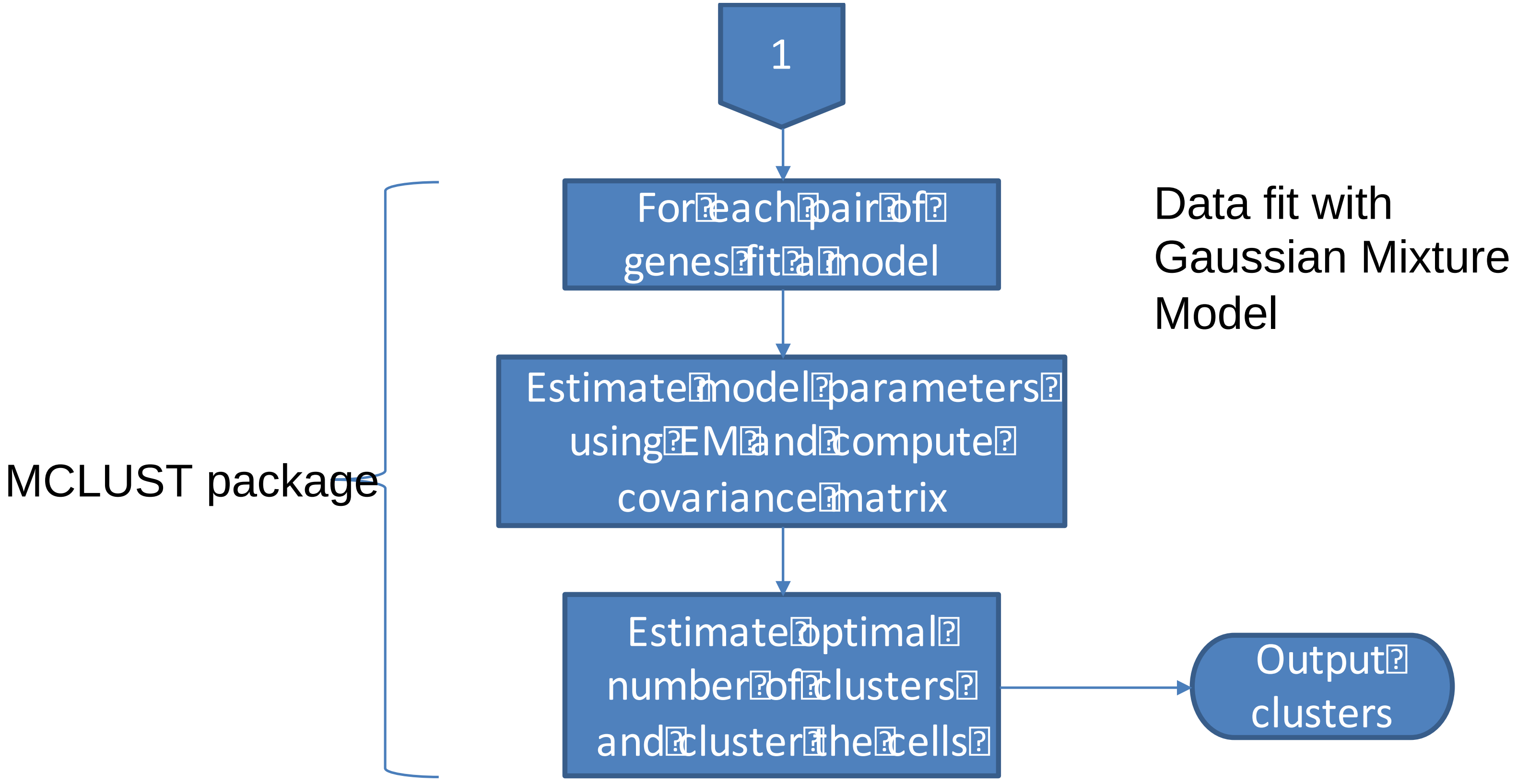
Accounting for genes who’s expression is drastically affected by Metformin



MiMoSA

Mixture-Model based Single-cell Analysis

PHASE 2



1

For each pair of genes fit a model

Estimate model parameters using EM and compute covariance matrix

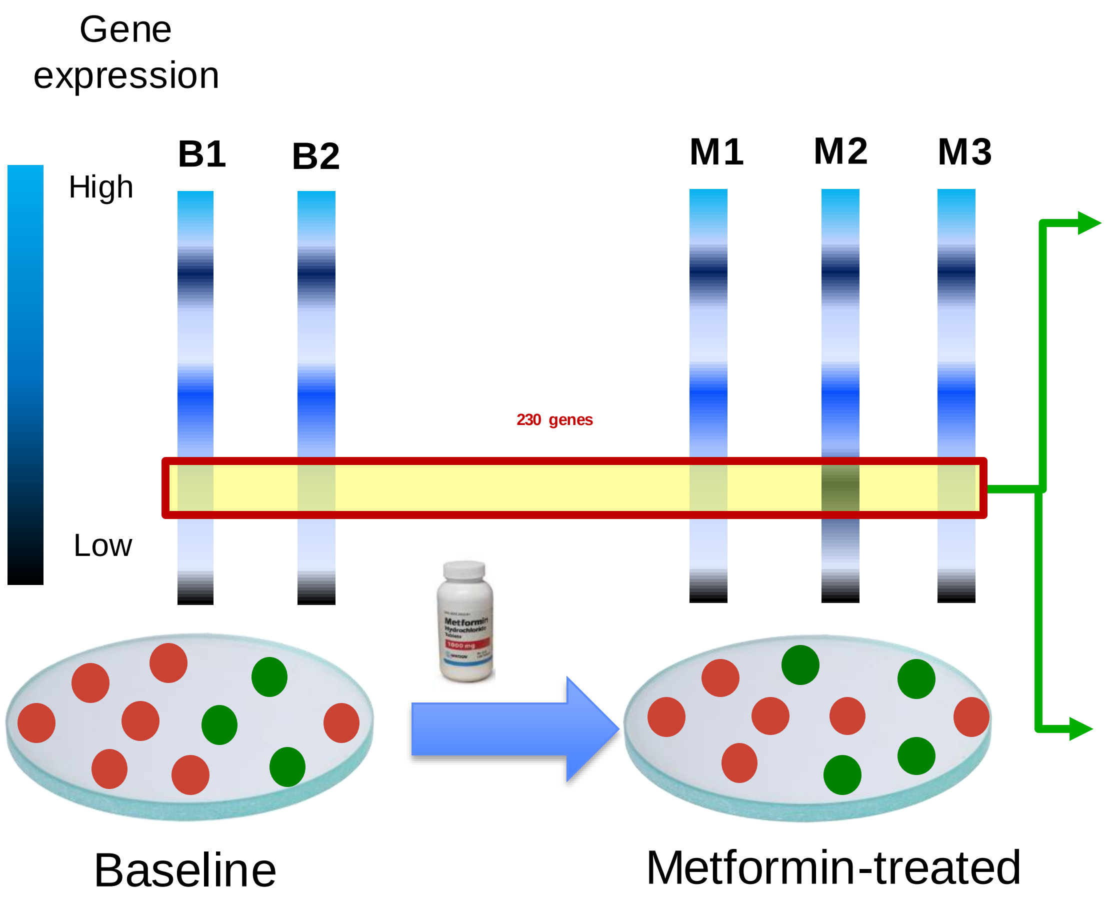
Estimate optimal number of clusters and cluster the cells

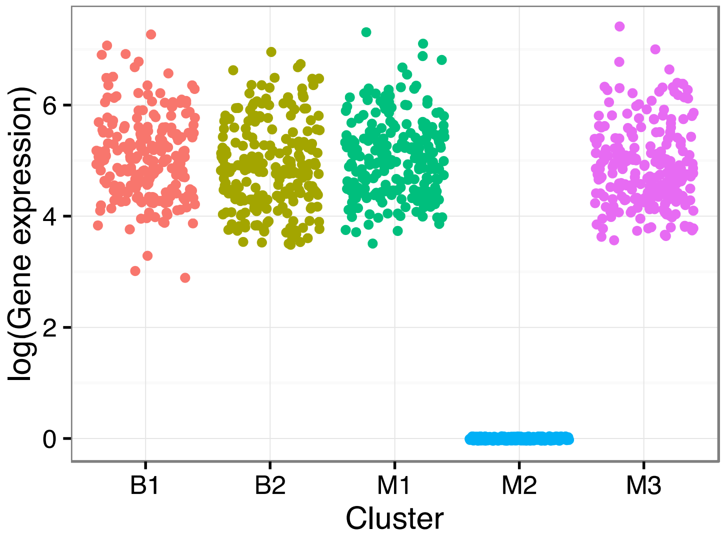
Output clusters

MCLUST package

Data fit with Gaussian Mixture Model

MiMoSA reveals cell clusters





Signaling pathway: **CDC42** (impact unknown)

Canonical pathway: (impact known)
Mitochondrial dysfunction: p-value: **1.03E-26**
Oxidative phosphorylation: p-value: **1.52E-25**
m-TOR signaling : p-value: **1.62E-07**

Machine Learning to Biology: Validation of Analytics

- In addition to significant pathways,
  - 70% of 230 genes were implicated in known metformin response pathways
  - **Validation without experiments**

- In the remaining 30% of the genes which are less studied,
  - ***CDC42*** – also down-regulated in population studies in 2009
  - Functional implication of *CDC42* knockdown from metformin not known
  - **Actionable Intelligence for Lab Experimentation: *Test CDC42's impact on anticancer properties***

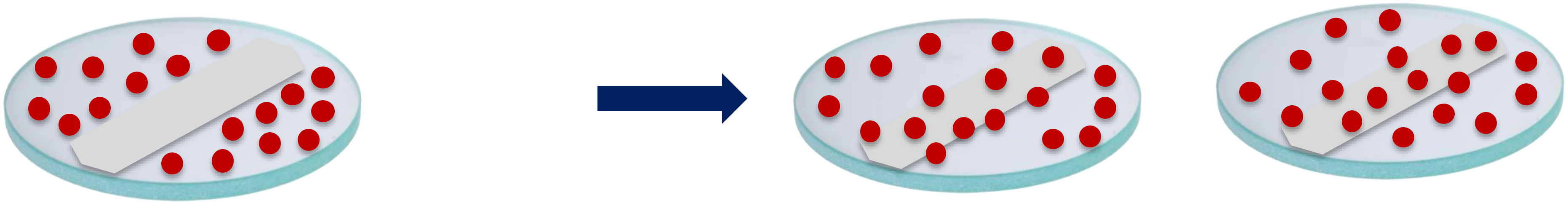
Lab Results

Down regulation of CDC42 through Metformin inhibits cell-migration

Control experiment:

48 hours

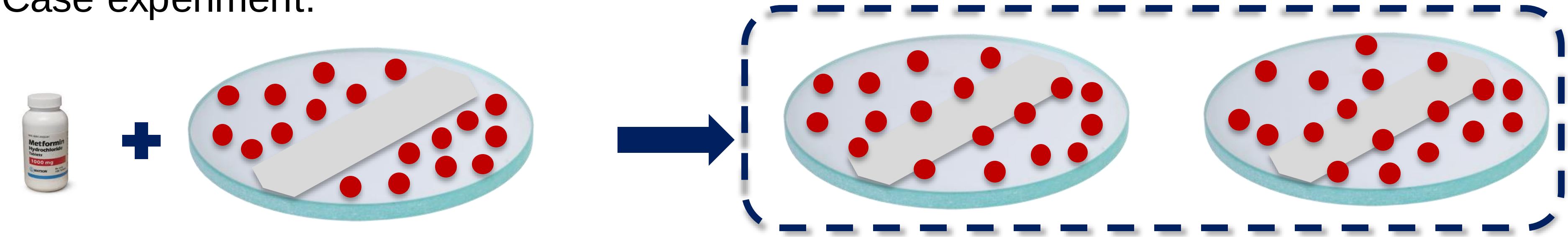
72 hours



Case experiment:

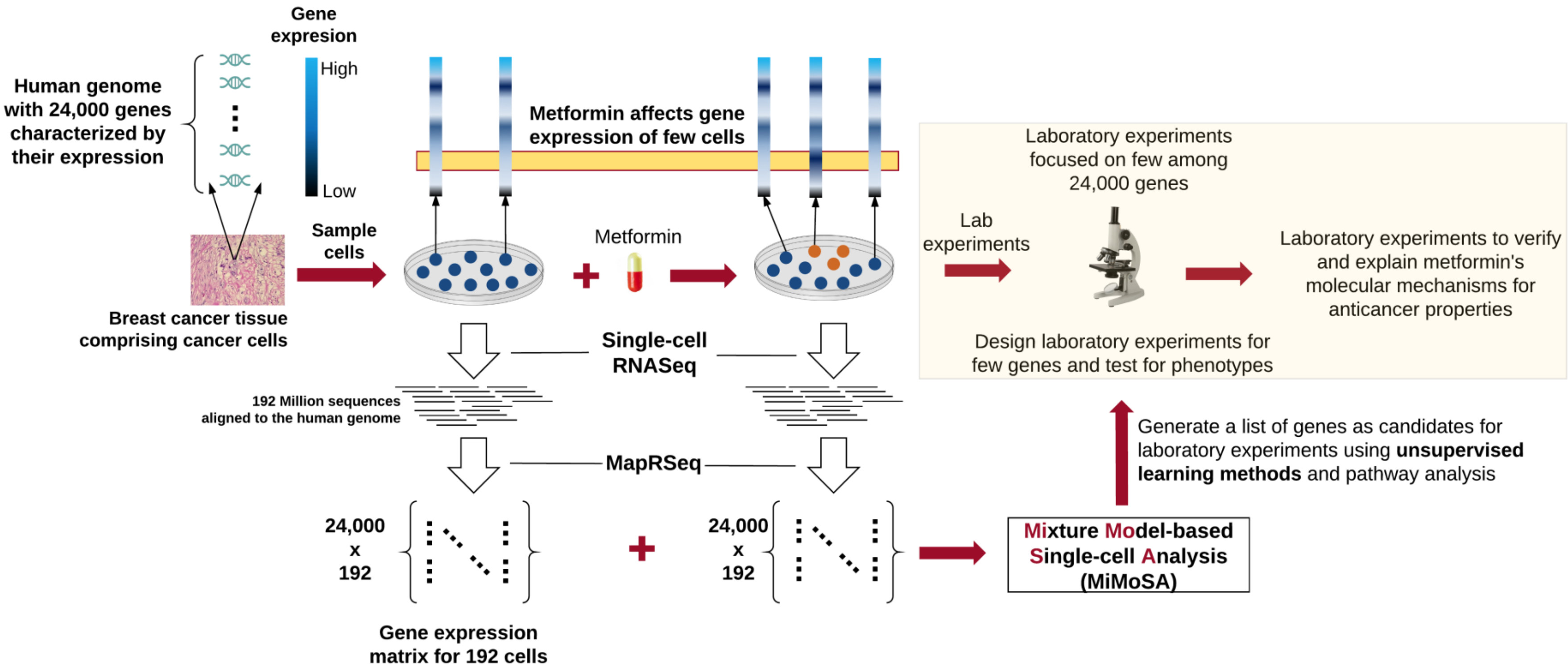
48 hours

72 hours



CELLS ARE STILL WELL-SEPARATED

Summary of Work – From Data to Actionable Intelligence


Gene expresion
High
Low
Human genome with 24,000 genes characterized by their expression
Metformin affects gene expression of few cells
Sample cells
Metformin
Breast cancer tissue comprising cancer cells
Single-cell RNASeq
192 Million sequences aligned to the human genome
MapRSeq
24,000 x 192
24,000 x 192
Gene expression matrix for 192 cells
Mixture Model-based Single-cell Analysis (MiMoSA)
Generate a list of genes as candidates for laboratory experiments using unsupervised learning methods and pathway analysis
Laboratory experiments focused on few among 24,000 genes
Lab experiments
Laboratory experiments to verify and explain metformin's molecular mechanisms for anticancer properties
Design laboratory experiments for few genes and test for phenotypes