Exploratory Data Analysis

STOR 565

Andrew Nobel

January, 2020

Data: Definitions from the O.E.D.

General: Facts and statistics collected together for reference or analysis.

Philosophy: Things known or assumed as facts, making the basis of reasoning or calculation.

Computing: The quantities, characters, or symbols on which operations are performed by a computer, being stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or mechanical recording media.

Datasets

Rough definition: An organized collection of information, usually numerical or categorical, obtained by measuring a common set of features across a given set of related objects.

Statistical terminology

- Objects/individuals under study are called **samples**
- Measured features are referred to as **variables**

Dataset with n samples and p variables represented by $n \times p$ Matrix

- i'th row contains measurements from i'th sample
- j'th column contains measurements of j'th variable

Example: Fisher's Iris Data

Figure: from Wikipedia

Measurements of length and width of sepals and petals of 50 Iris flowers of three different species (setosa, versicolor, and virginica).

Analyzed by R. A. Fisher "The use of multiple measurements in taxonomic problems" Annals of Eugenics, 1936

Iris Data

Data matrix with 150 samples and 4 variables.

Species	Sepal Length	Sepal Width	Petal Length	Petal Width
Setosa	5.1	3.5	1.4	0.2
Setosa	4.6	3.4	1.4	0.3
Versicolor	5.0	2.0	3.5	1.0
Virginica	7.2	3.6	6.1	2.5
...	...	...	...	...

Fisher's Iris Data: Scatterplots

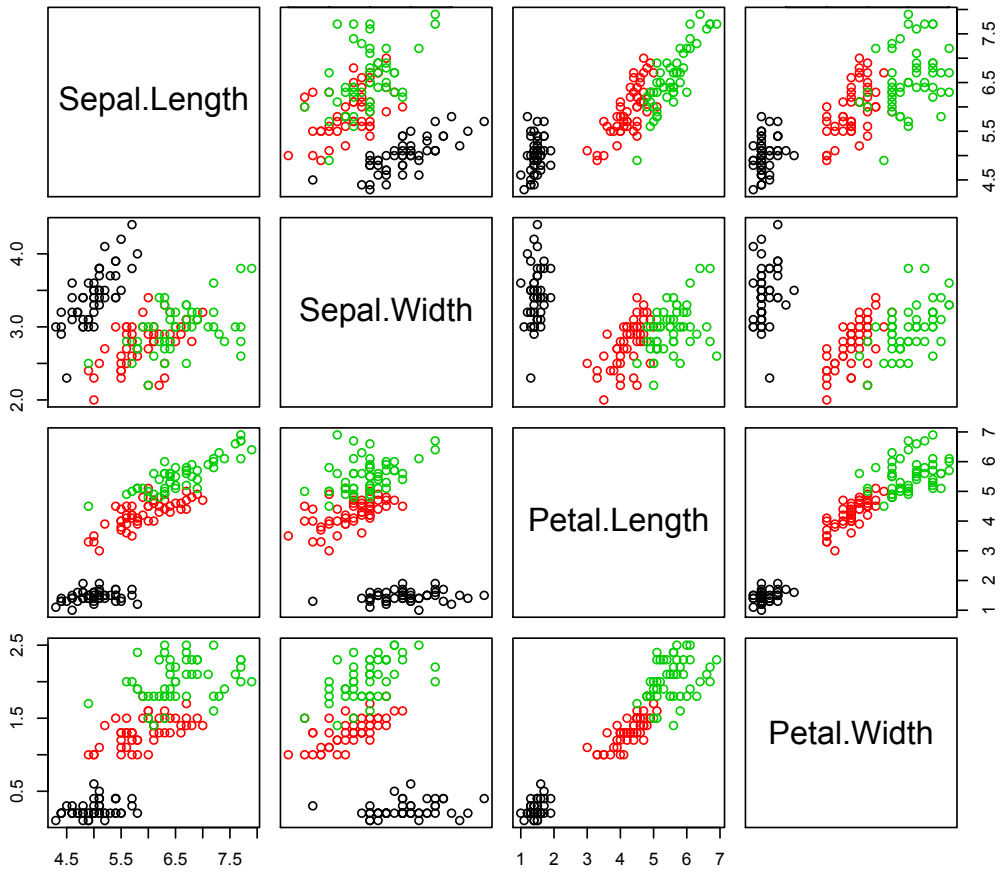


Figure: Pairwise scatterplot of Iris measurements. Colors: *Setosa*, *Virginica*, *Versicolor*.

Fisher's Iris Data: Principal Component Analysis

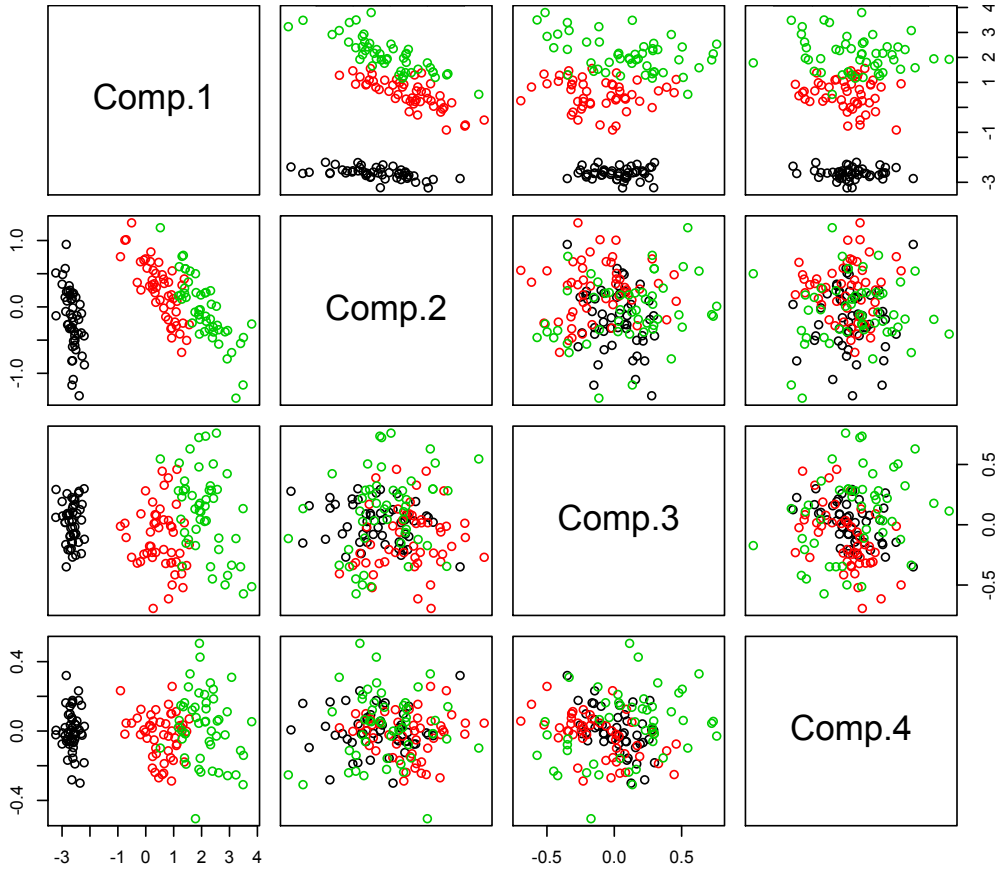
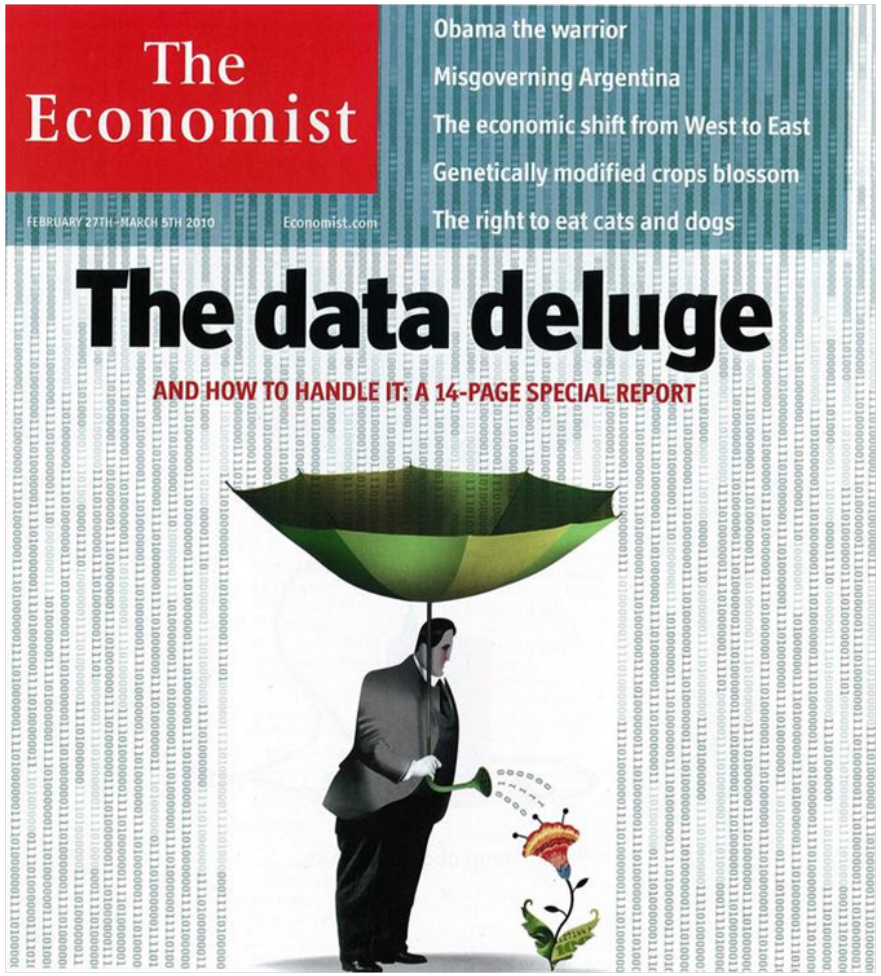


Figure: Pairwise scatterplot of principal components of Iris measurements. Colors: *Setosa*, *Virginica*, *Versicolor*.

Big Data: Cover of The Economist, 2010



Big Data: Enron Email Graph

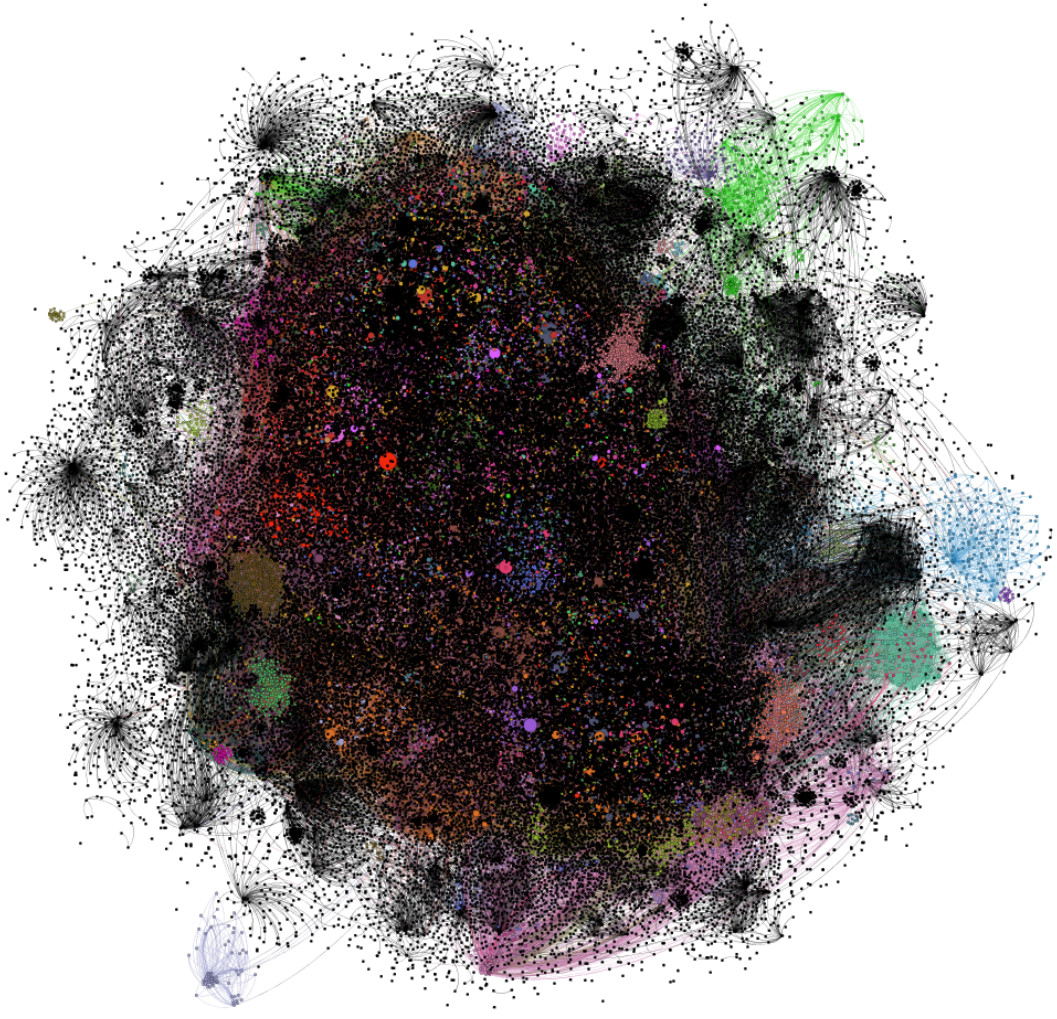
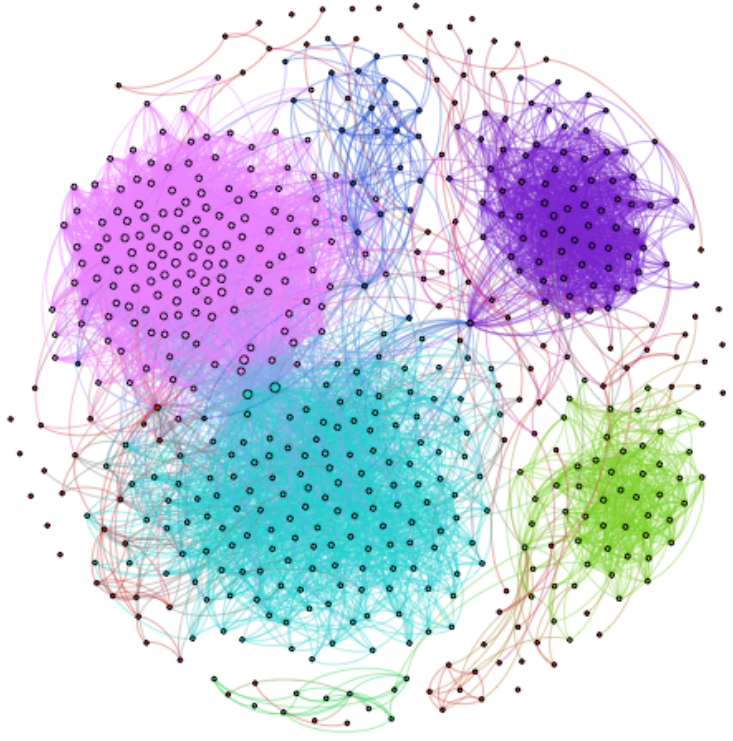


Figure: Graph of three million emails between 80,705 people. Vertices are individuals. Edges are colored according to number of emails exchanged.

Facebook Data: Community Detection



Facebook friendship network (8 underlying groups)

- **Vertices** = 561 Friends on Facebook
- **Edges** = 8375 Friendships among friends

The Cancer Genome Atlas

Multi-Institution consortium supported by the National Institutes of Health.

Goal: “[T]o accelerate our understanding of the molecular basis of cancer through the application of genome analysis technologies, including large-scale genome sequencing”

Consortium collected tissue from thousands of tumors across numerous cancer types. Data derived from high-throughput technologies measuring

- Gene expression
- Micro-RNA
- DNA copy number
- Methylation

TCGA Data Portal Overview

The Cancer Genome Atlas (TCGA) Data Portal provides a platform for researchers to search, download, and analyze data sets generated by TCGA. It contains clinical information, genomic characterization data, and high level sequence analysis of the tumor genomes.

Please note some data on the TCGA Data Portal are in controlled-access. Please visit the Access Tiers page for more information.

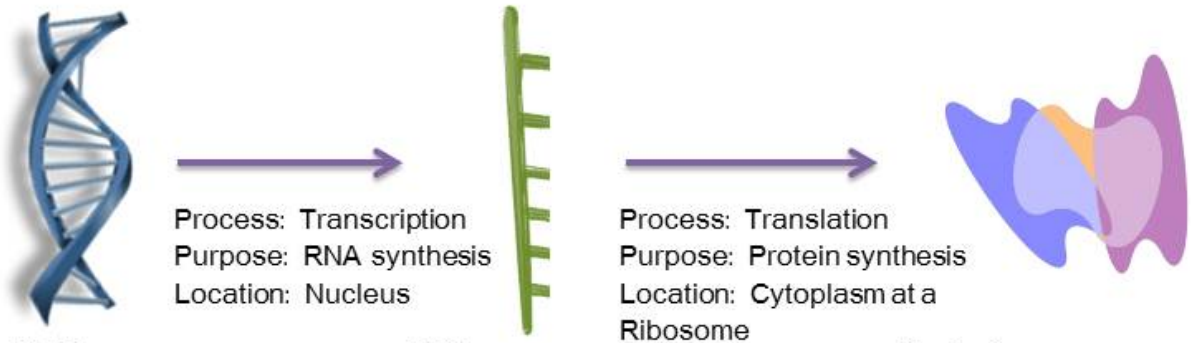
The TCGA Data Portal does not host lower levels of sequence data. NCI's Cancer Genomics Hub (CGHub) is the new secure repository for storing, cataloging, and accessing BAM files and metadata for sequencing data.

Download Data

Choose from four ways to download data

Available Cancer Types	# Cases Shipped by BCR*	# Cases with Data*	Date Last Updated (mm/dd/yy)
Acute Myeloid Leukemia [LAML]	200	200	04/29/15
Adrenocortical carcinoma [ACC]	80	80	08/27/15
Bladder Urothelial Carcinoma [BLCA]	412	412	08/27/15
Brain Lower Grade Glioma [LGG]	516	516	08/27/15
Breast invasive carcinoma [BRCA]	1100	1098	08/28/15
Cervical squamous cell carcinoma and endocervical adenocarcinoma [CESC]	308	308	08/21/15
Cholangiocarcinoma [CHOL]	36	36	08/21/15
Colon adenocarcinoma [COAD]	461	461	08/27/15

The Central Dogma



DNA contains the original codes for making the proteins that living cells need. mRNA is a copy of a gene located on the DNA molecule. mRNA will leave the nucleus of the cell and the ribosome will read its coding sequences and put the appropriate amino acids together.

Screenshot of Expression Data

Heat Maps of Numerical Data

Heat map: Means of displaying a data matrix with numerical entries

- positive entries are red
- negative entries are green
- entries close to zero are black

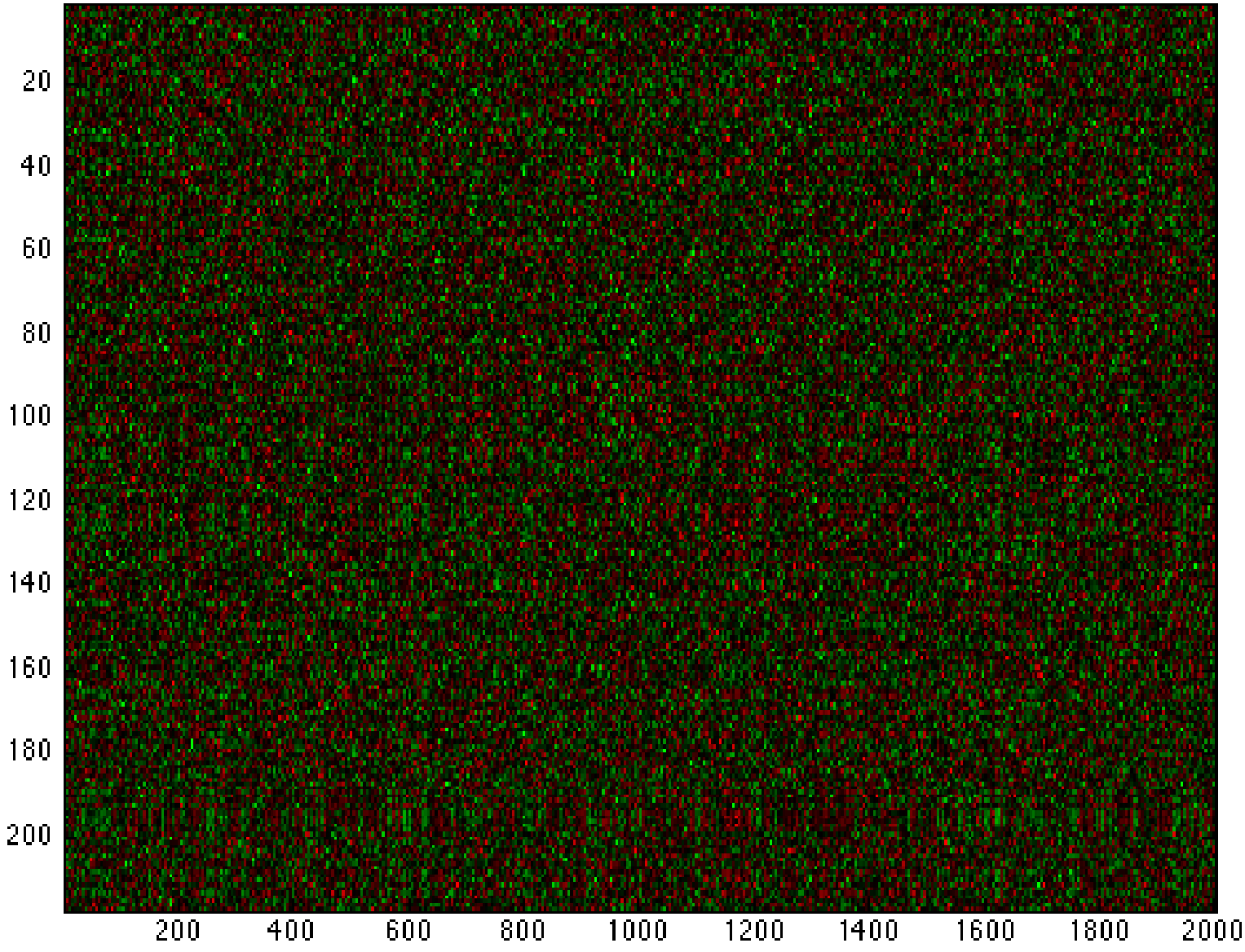
Example: Data set from The Cancer Genome Atlas (TCGA)

Samples: Tissue from 217 breast tumors

- 95 Luminal A tumors
- 122 Basal tumors

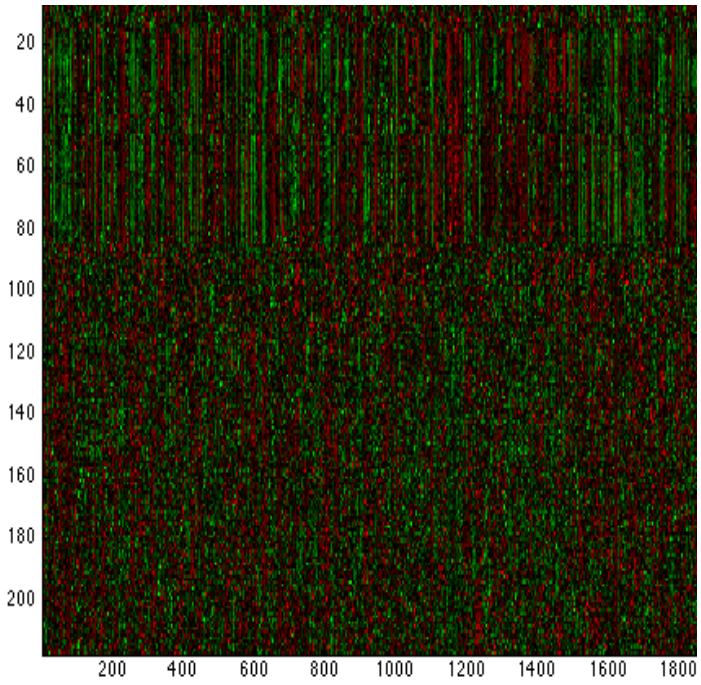
Variables: Expression levels of 2000 genes

Heat Map of TCGA Gene Expression Data



Looking ahead: Row and Column Clustering

Row Clustering



Column Clustering

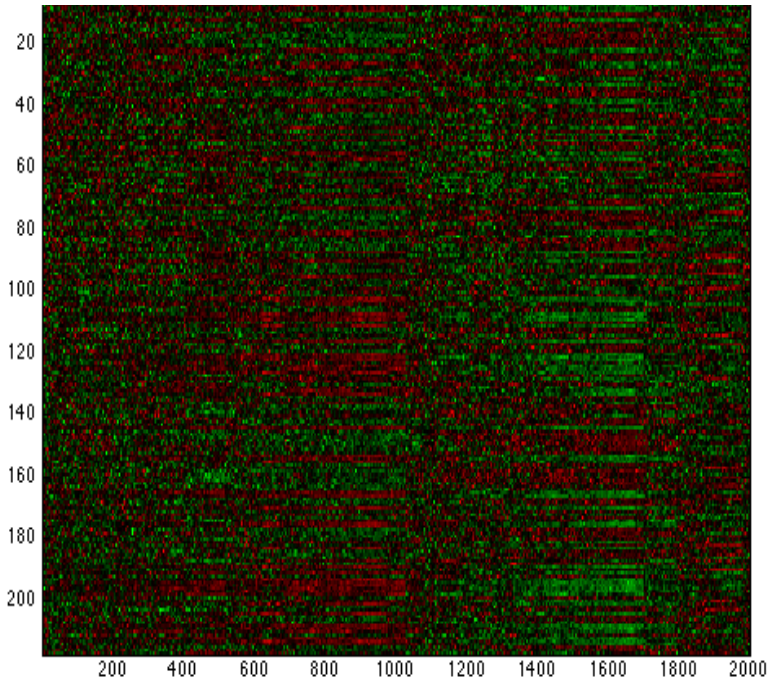
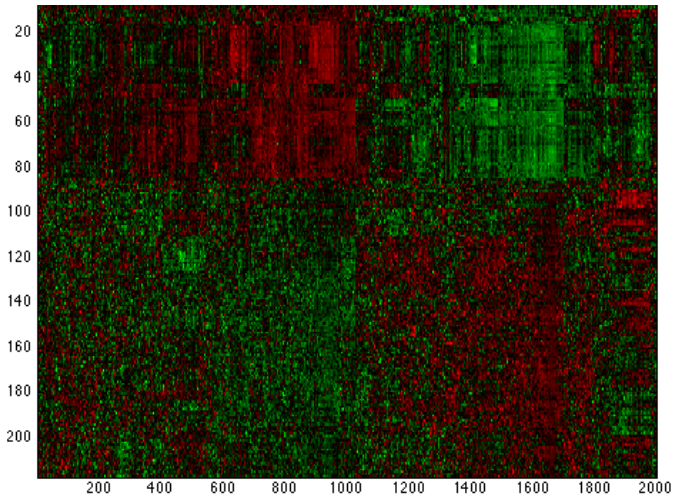


Figure: (Left): Heat map after rows are reordered according to hierarchical clustering. (Right) Heat map after columns are reordered according to hierarchical clustering.

Teaser: Co-Clustering and Biclustering

Row and Column Clustering



Biclustering

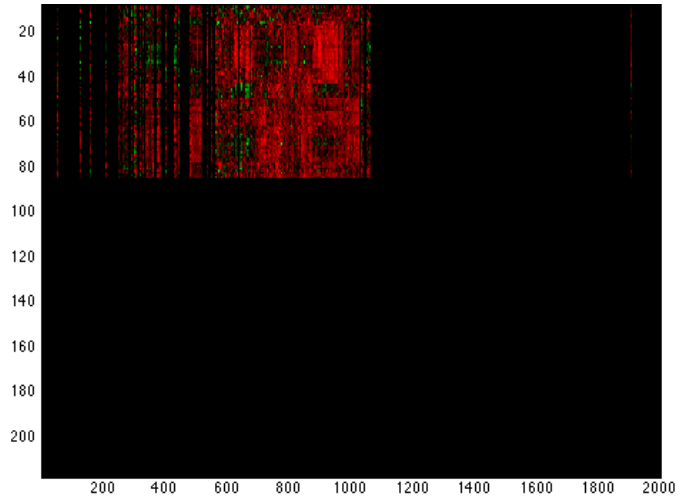
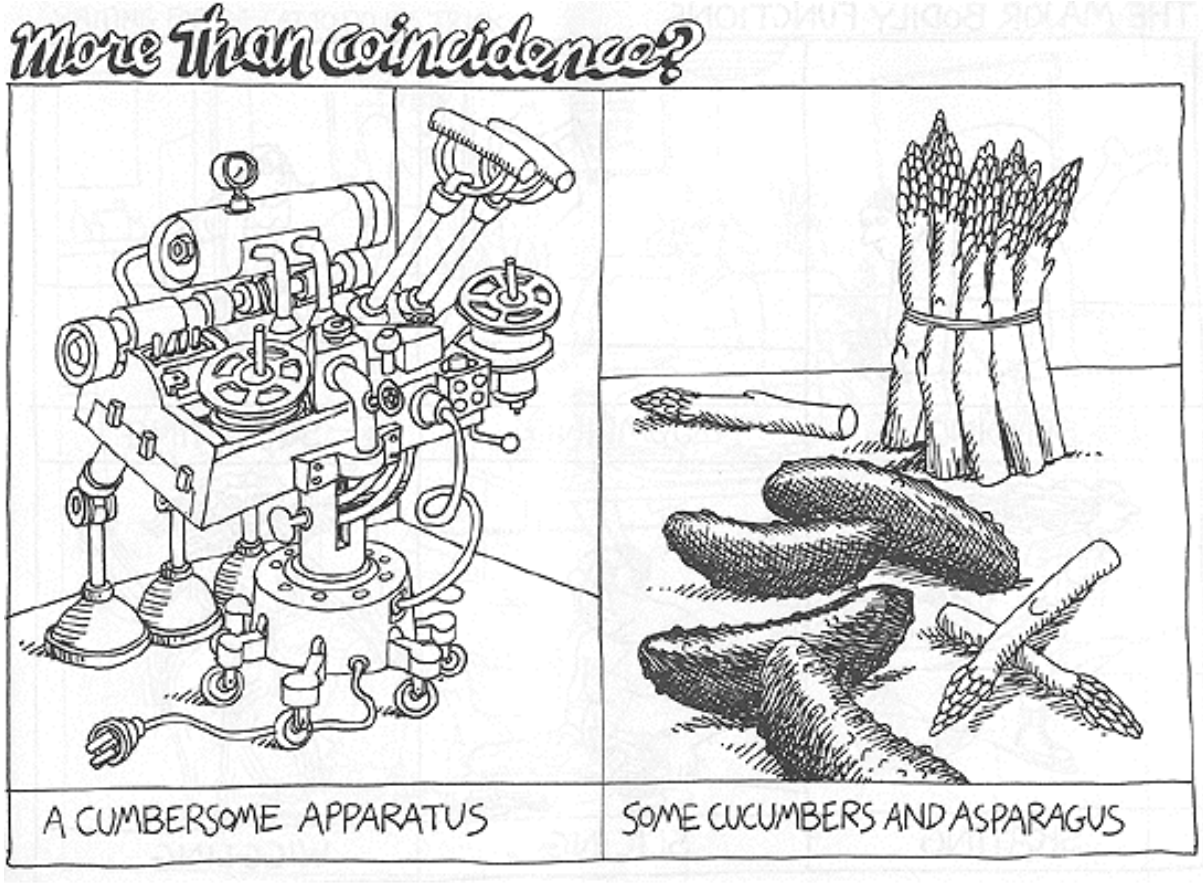


Figure: (Left): Rows and Columns are separately reordered by clustering. (Right) The first bicluster extracted from this data.

Finding Patterns in Data



Drawing by B. Kliban

Exploratory Data Analysis

First look at a data set, typically in the form of a matrix of numbers.

- Visualization
- Identifying patterns or regularities of interest

Preliminaries

- Identifying and addressing outliers and extreme values
- Identifying and filling in missing values
- Normalization: removing systematic differences between samples
- Transforming data values using logarithm or other functions
- Checking distributional assumptions

Overview: Univariate Data Analysis

Univariate Sample: Sample $x = x_1, \ldots, x_n$ with $x_i \in \mathbb{R}$

- Sample mean $m(x) = \overline{x} = n^{-1} \sum_{i=1}^{n} x_i$
- Sample variance $s^2(x) = n^{-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$
- Sample standard deviation $s(x)$
- Coefficient of variation $\mathrm{CV}(x) = s(x) \,/\, \overline{x}$ (when x_i positive)

Standardized sample: Replace x_i with $\tilde{x}_i = (x_i - \overline{x})/s(x)$

- Standardization ensures $m(\tilde{x}) = 0$ and $s(\tilde{x}) = 1$

Univariate Data, continued

Rank based statistics

- Order statistics $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$
- αth percentile $= x_{(r)}$, where r is the integer closest to $n\,(\alpha/100) + 1/2$
- Special cases: first (25%), median (50%), and third (75%) quartiles

Definition: Empirical cumulative distribution function (CDF) of x

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(x_i \leq t) \qquad -\infty < t < \infty$$

- "staircase shape" with jumps of size $1/n$ at each data point
- can recover dataset x from $F_n(t)$ apart from order

Univariate Data, visualization

Histogram of $\{x_1, \ldots, x_n\}$, or appropriate density estimate

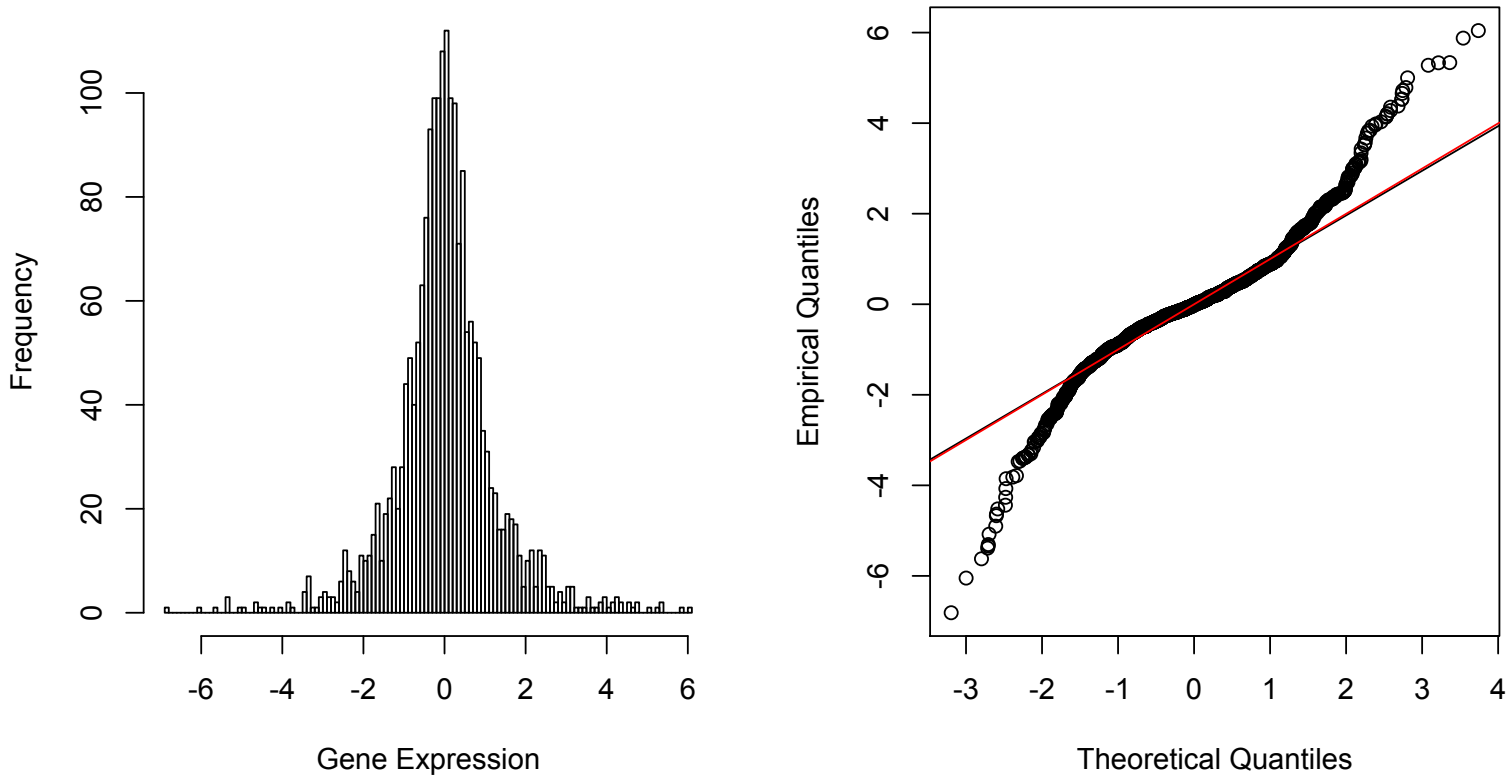
Testing Normality: Normal QQ plot of dataset $x_1, \ldots, x_n$.

- x-axis is theoretical quantiles of standard normal CDF $\Phi(x)$
- Reference line shows x versus $y = \Phi^{-1}(\Phi(x)) = x$
- Empirical plot shows x_i versus $y_i = F_n^{-1}(\Phi(x_i))$

Take-away

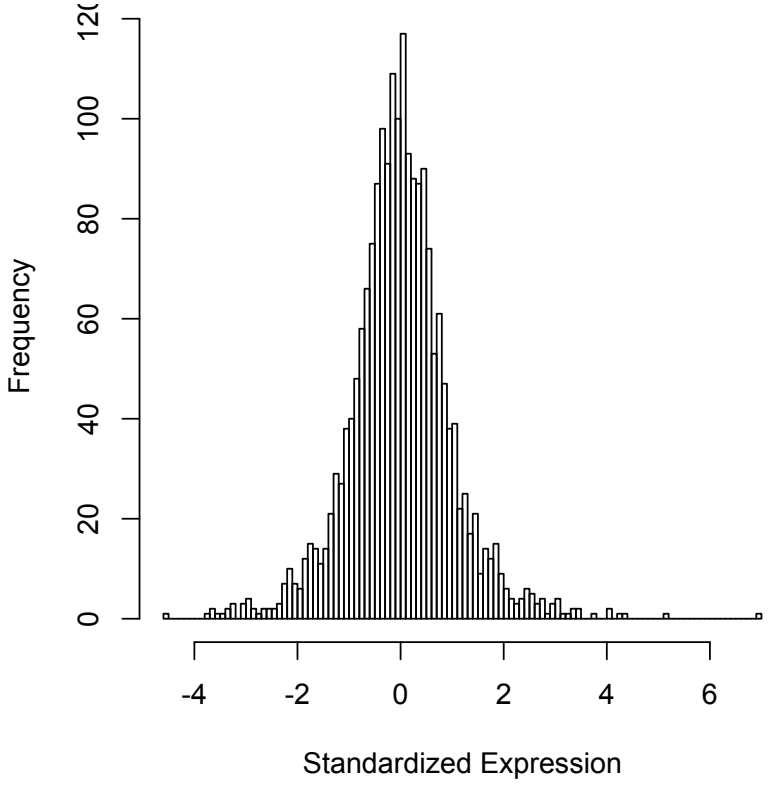
- If QQ plot steeper than reference line then empirical distribution has heavier tails (more dispersed) than normal.
- If QQ plot flatter than reference line then empirical distribution has lighter tails (less dispersed) than normal.

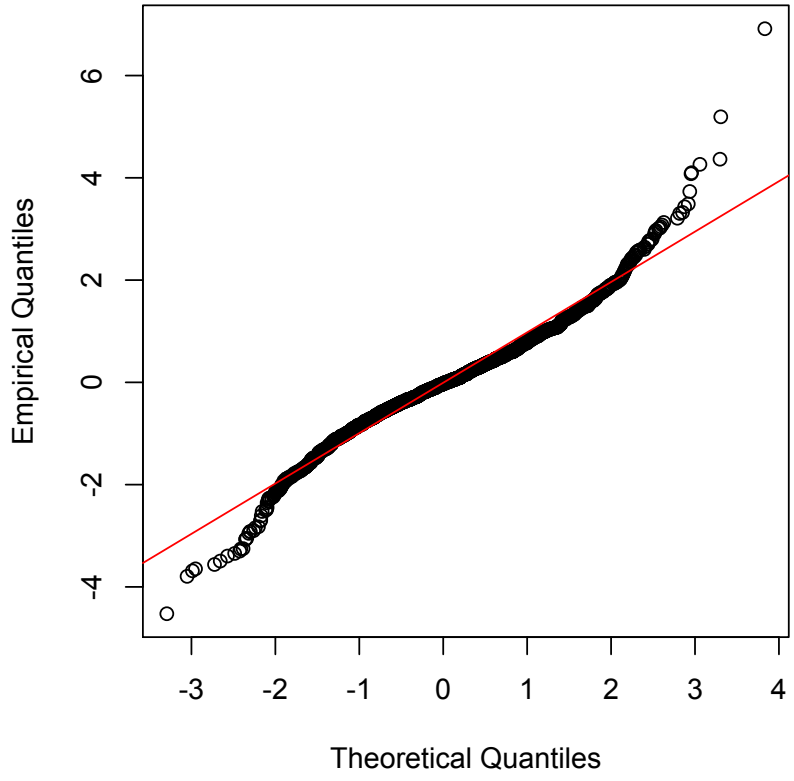
Histogram and QQ-plot, First TCGA Sample



Note: Each figure based on 2000 expression measurements in first row of the data matrix.

Histogram and QQ-plot, First Sample after Standardization





Note: Each figure based on standardized values of 2000 expression measurements in first row of the data matrix.

Bivariate Data

Bivariate Sample: $(x, y) = (x_1, y_1), \ldots, (x_n, y_n)$ with $(x_i, y_i) \in \mathbb{R}^2$

- Univariate statistics $m(x), s^2(x)$ and $m(y), s^2(y)$.

- Sample covariance of x and y

$$s(x, y) \;=\; n^{-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) \;=\; n^{-1} \sum_{i=1}^{n} x_i \, y_i - \overline{x} \, \overline{y}$$

- Sample correlation of x and y

$$r(x, y) \;=\; \frac{s(x, y)}{s(x) \, s(y)} \;\in\; [-1, 1]$$

Bivariate Data: Scatter Plots and Regression Lines

Scatter-Plot: Plot of $\{(x_i, y_i) : 1 \leq i \leq n\} \subseteq \mathbb{R}^2$

Sample regression line of y on x: This is the line $\ell^*(x)$ minimizing

$$\text{MSE}(\ell) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \ell(x_i))^2$$

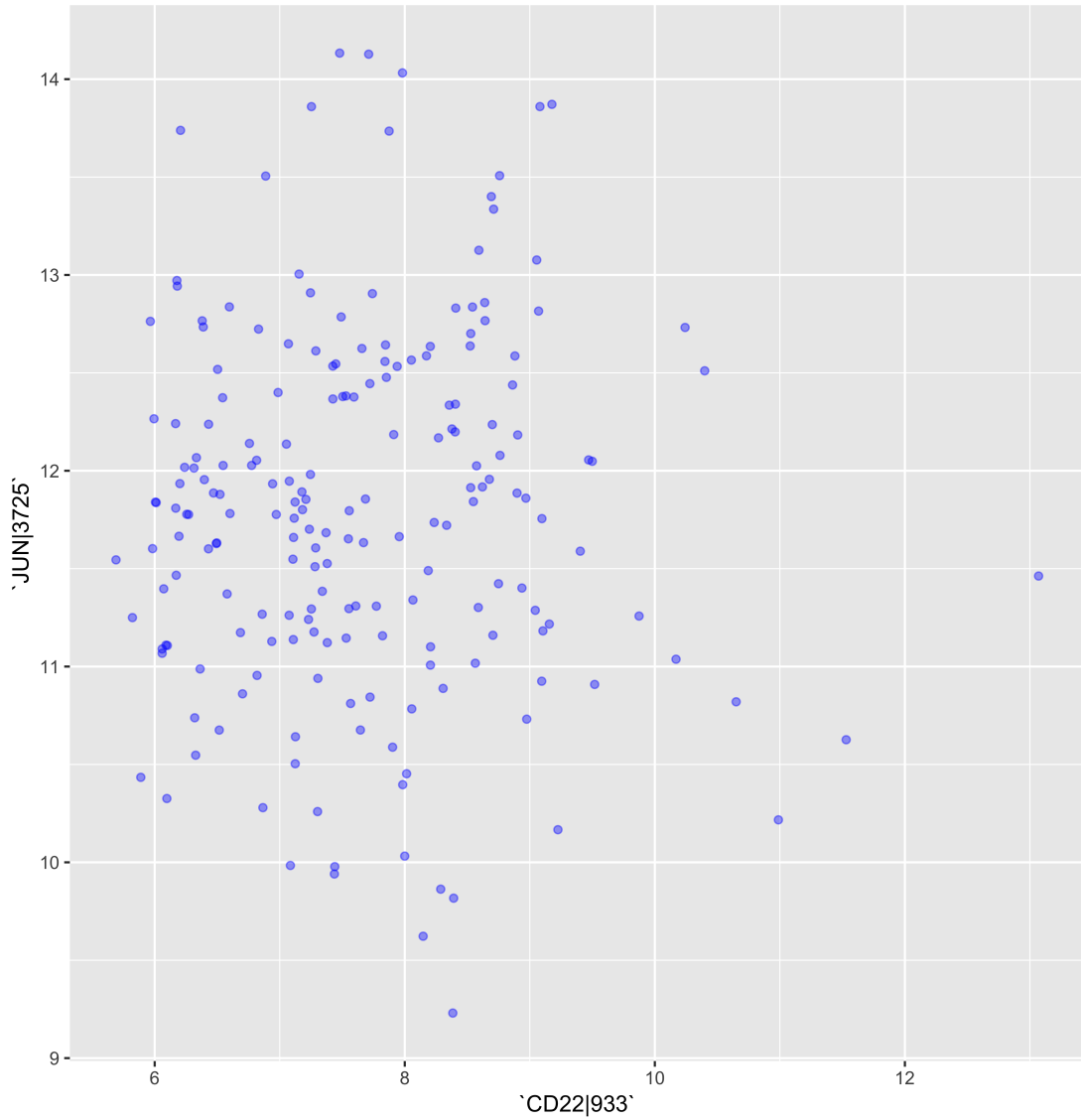
over all linear functions $\ell(x) = ax + b$.

Fact: Sample regression line ℓ^* of y on x is given by

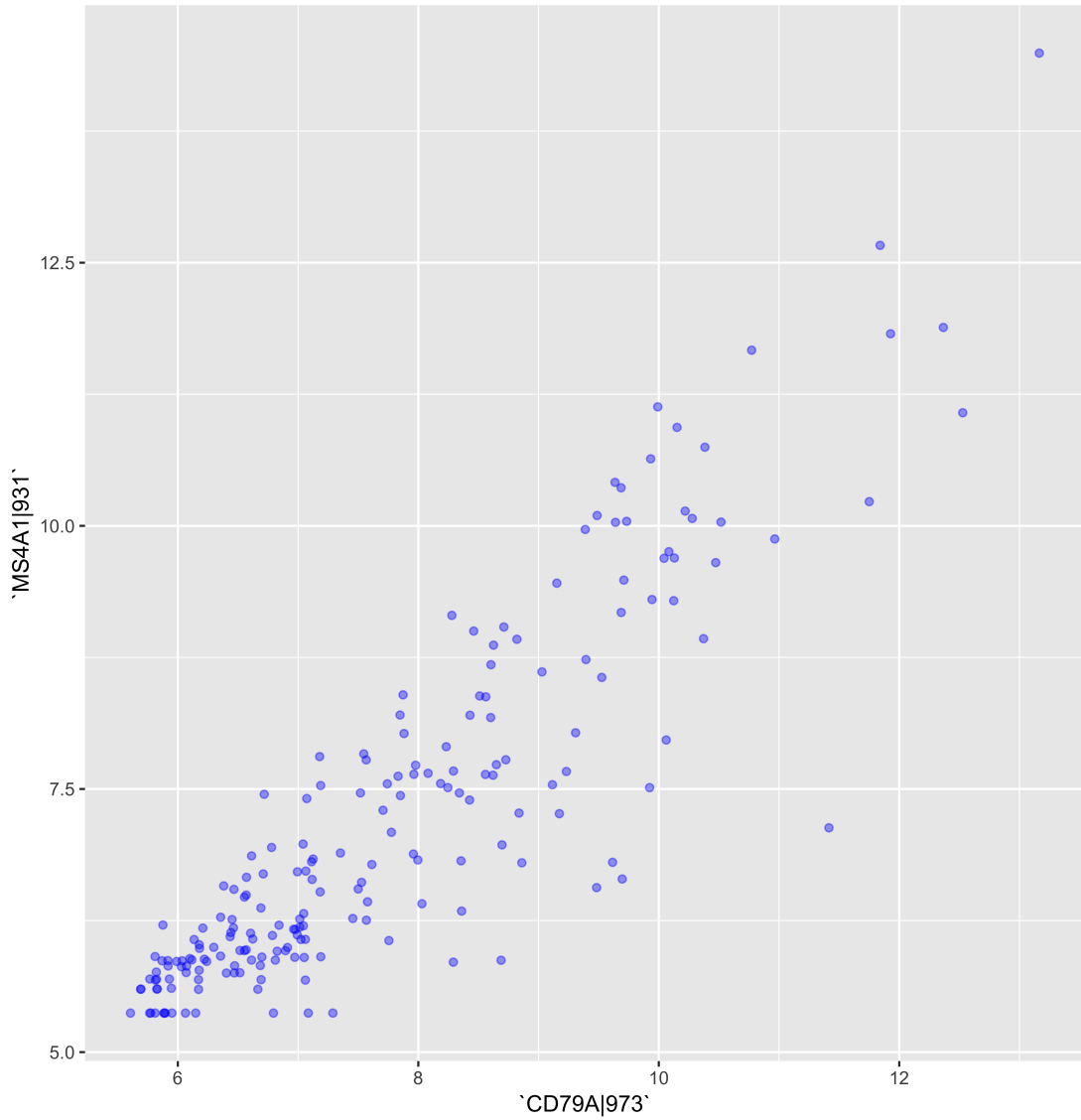
$$\ell^*(x) = m(y) + \frac{s(x,y)}{s^2(x)}[x - m(x)]$$

and satisfies $\text{MSE}(\ell^*) = s^2(y)[1 - r^2(x,y)]$.

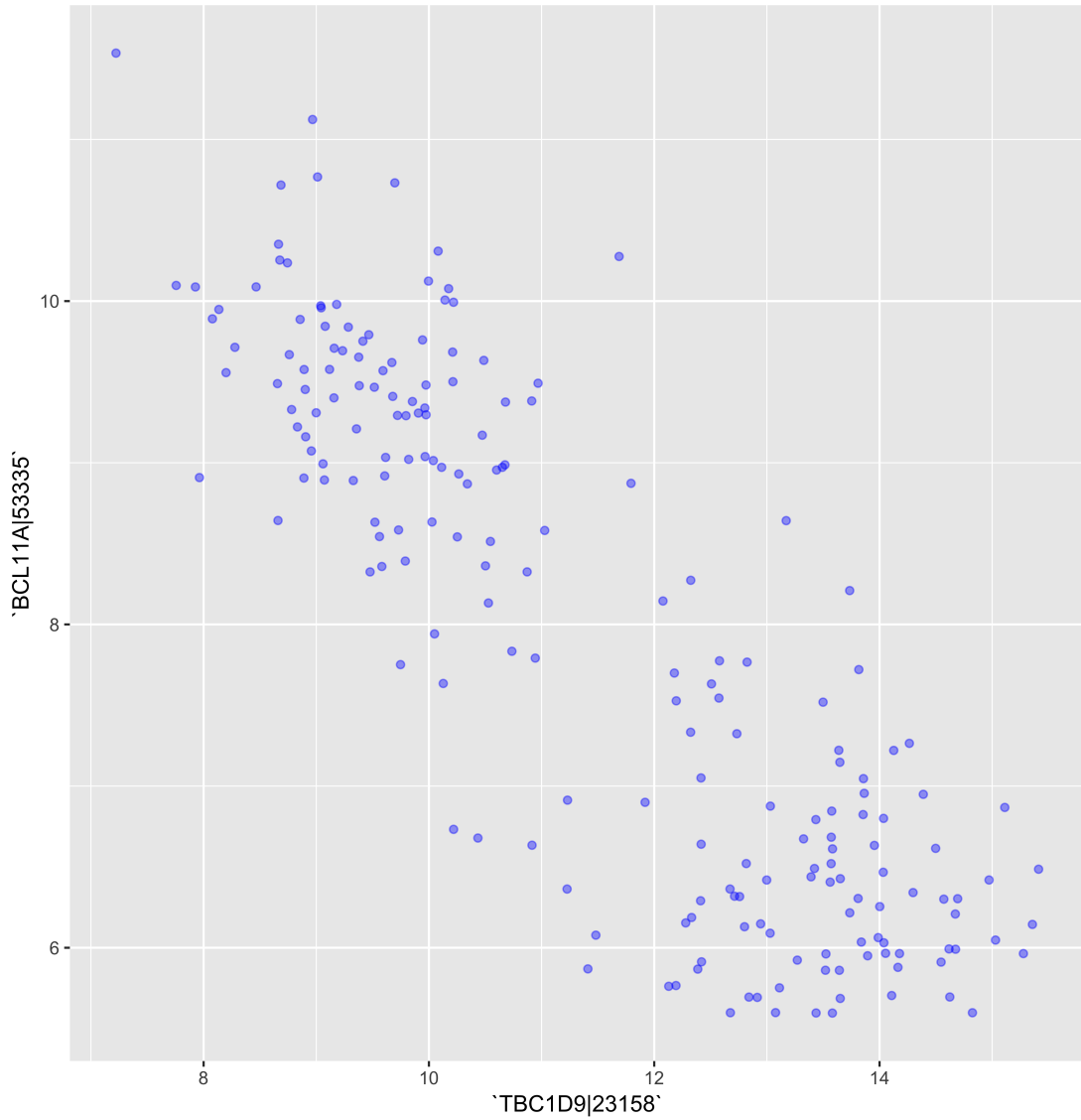
Two Uncorrelated Genes



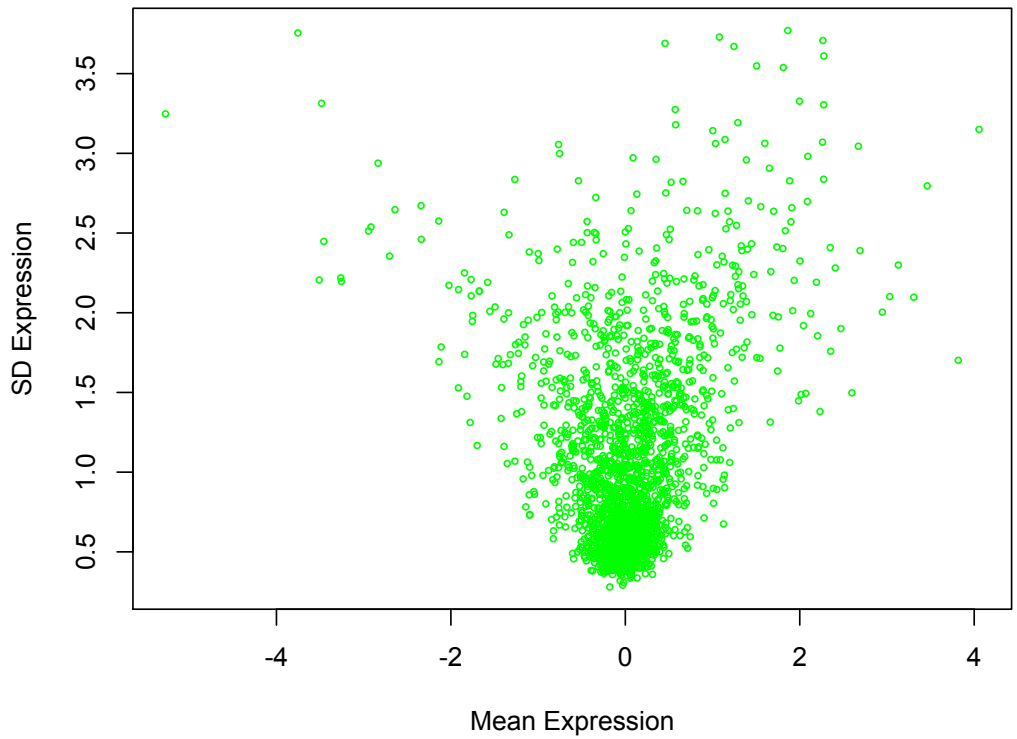
Two Positively Correlated Genes



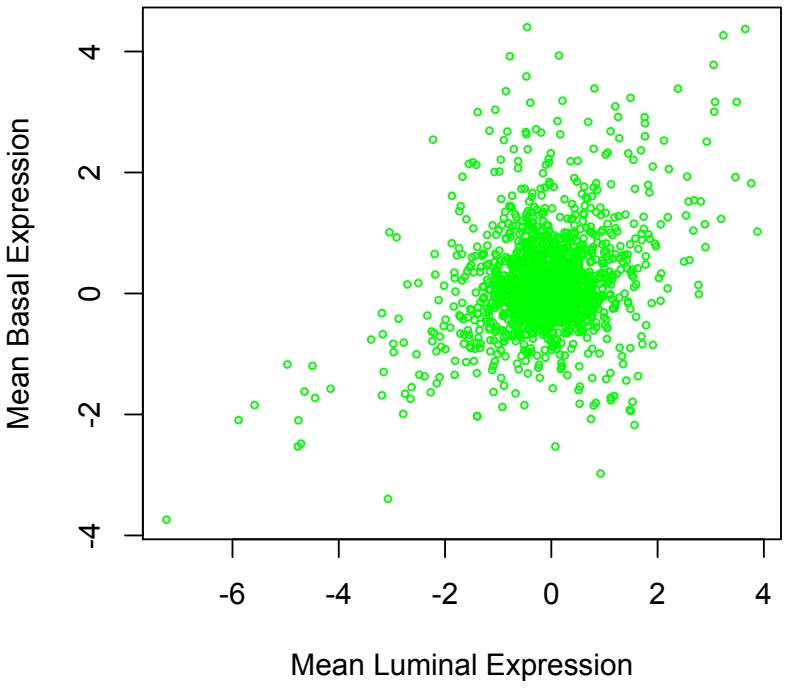
Two Negatively Correlated Genes

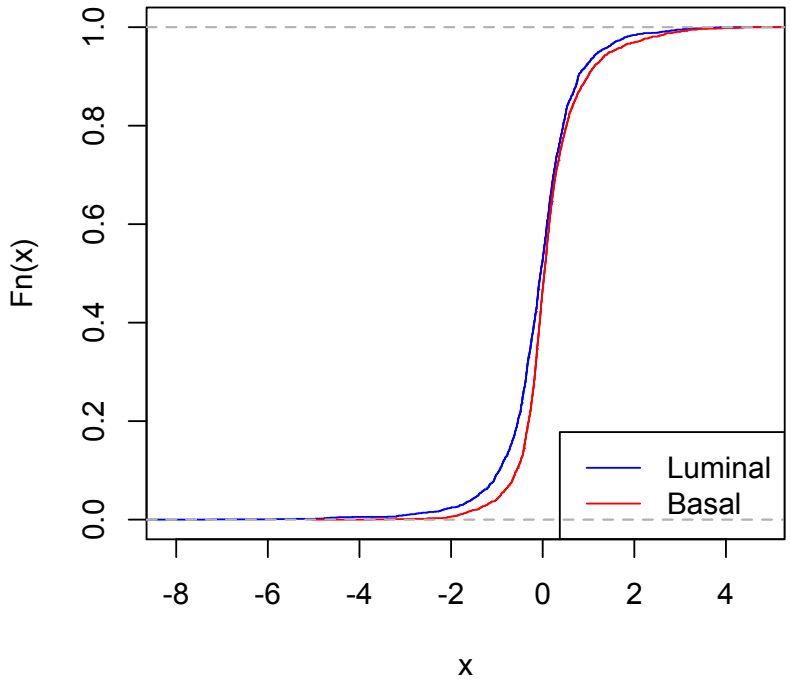


Comparison of Mean and SD of Expression for Genes



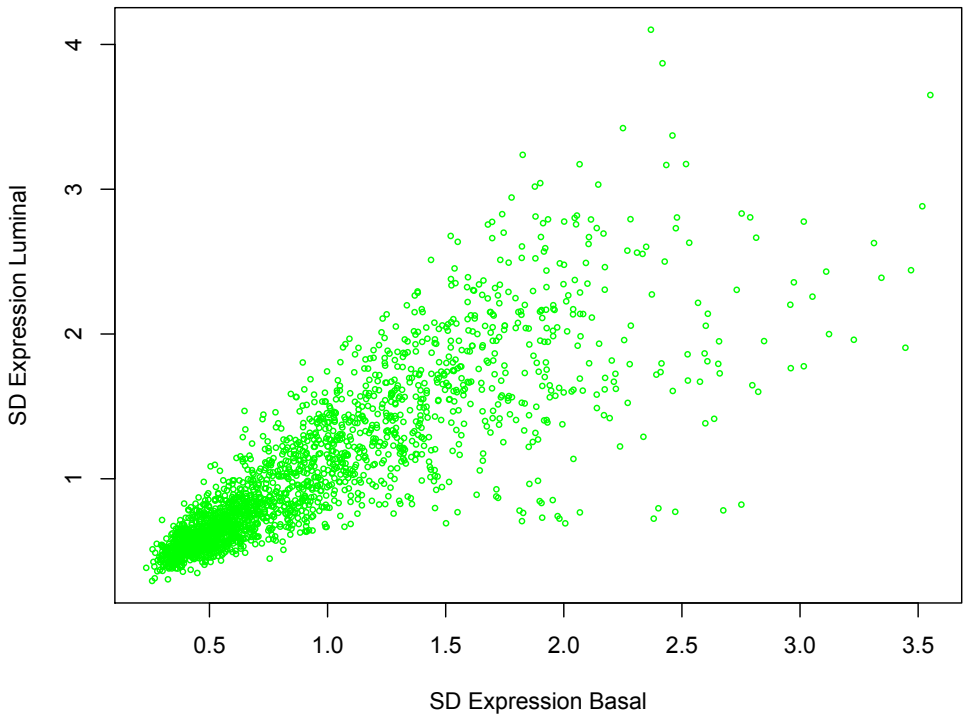
Comparison of Mean Expression of Across Subtypes





- Correlation: $r(x, y) = 0.30$; fraction of variance explained $r^2(x, y) = .090$

Comparison of SD Expression Across Subtypes



- Correlation: $r(x, y) = 0.84$; fraction of variance explained $r^2(x, y) = .71$.

Empirical Covariance Matrices

Given: $n \times p$ data matrix $\mathbf{X}$ with

- rows/samples $\mathbf{x}_{i\cdot} = (x_{i1}, \ldots, x_{ip}) \in \mathbb{R}^p$ for $i = 1, \ldots, n$
- cols/variables $\mathbf{x}_{\cdot k} = (x_{1k}, \ldots, x_{nk}) \in \mathbb{R}^n$ for $k = 1, \ldots, p$

Covariance Matrices for Samples and Variables (symmetric)

$$(\Sigma_{\mathsf{s}})_{ij} \;=\; s(\mathbf{x}_{i\cdot}, \mathbf{x}_{j\cdot}) \;=\; \frac{1}{p}\sum_{r=1}^{p} x_{ir}x_{jr} \;-\; \overline{\mathbf{x}}_{i\cdot}\,\overline{\mathbf{x}}_{j\cdot} \quad 1 \le i, j \le n$$

$$(\Sigma_{\mathsf{v}})_{kl} \;=\; s(\mathbf{x}_{\cdot k}, \mathbf{x}_{\cdot l}) \;=\; \frac{1}{n}\sum_{r=1}^{n} x_{rk}x_{rl} \;-\; \overline{\mathbf{x}}_{\cdot k}\,\overline{\mathbf{x}}_{\cdot l} \quad 1 \le k, l \le p$$

Note: Sample Covariance Σ_{s} is $n \times n$, Variable Covariance Σ_{v} is $p \times p$

Empirical Correlation Matrices

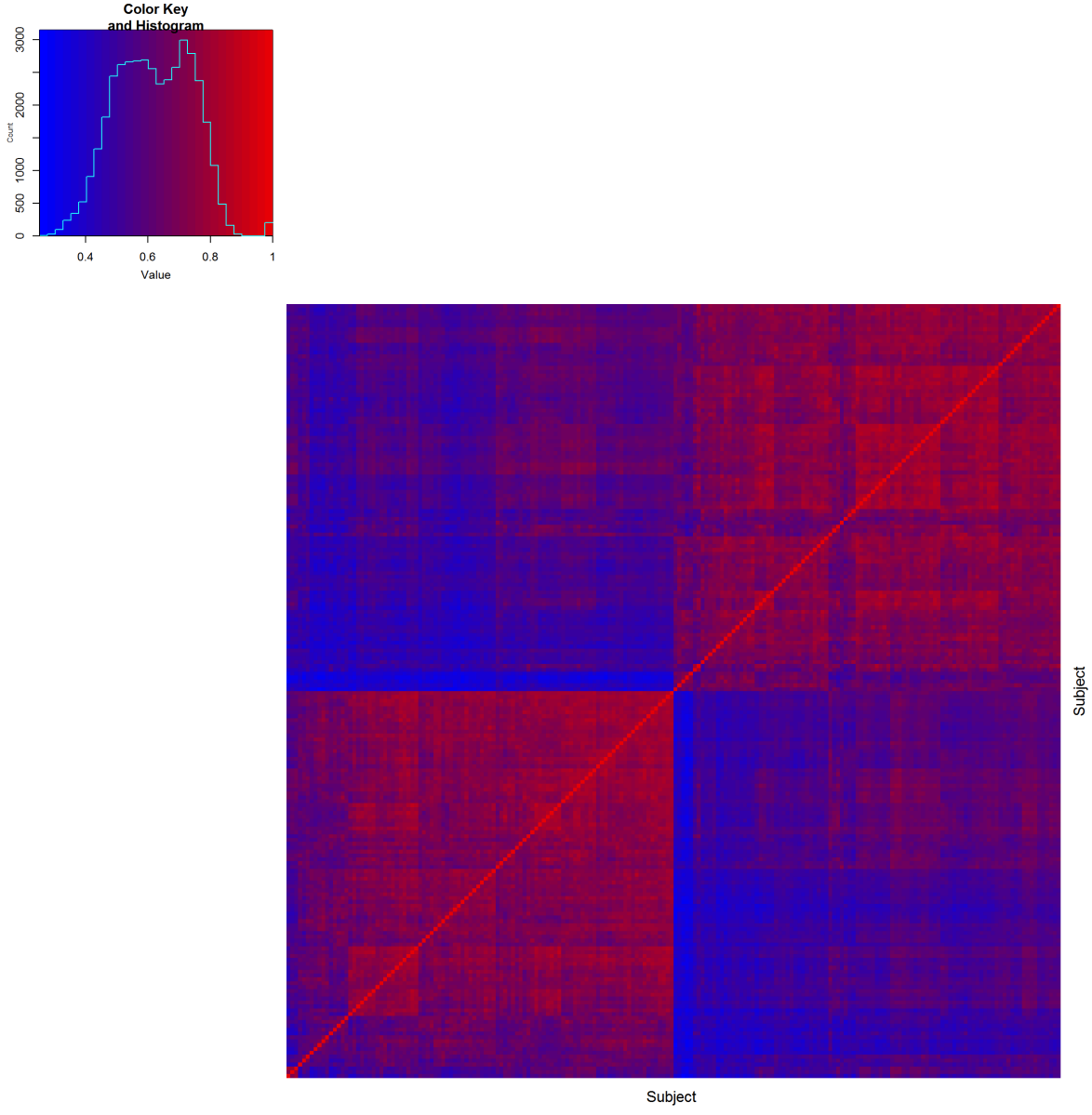
Correlation Matrices for Samples and Variables (symmetric)

$$(R_{\mathsf{s}})_{ij} = \frac{s(\mathbf{x}_{i\cdot}, \mathbf{x}_{j\cdot})}{s(\mathbf{x}_{i\cdot})\, s(\mathbf{x}_{j\cdot})} \quad 1 \leq i, j \leq n$$

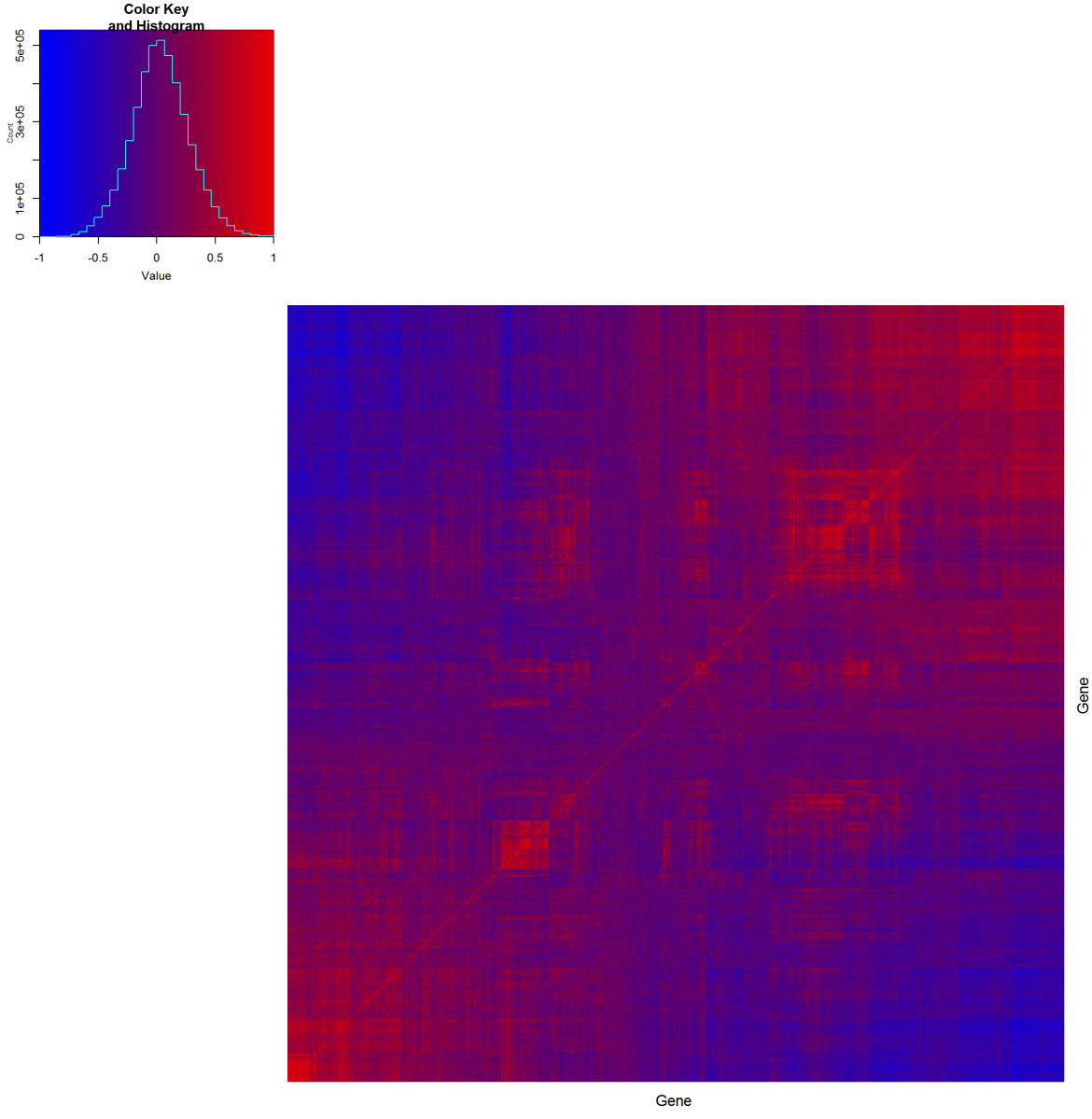
$$(R_{\mathsf{v}})_{kl} = \frac{s(\mathbf{x}_{\cdot k}, \mathbf{x}_{\cdot l})}{s(\mathbf{x}_{\cdot i})\, s(\mathbf{x}_{\cdot j})} \quad 1 \leq k, l \leq p$$

Note: Sample Covariance Σ_{s} is $n \times n$, Variable Covariance Σ_{v} is $p \times p$

Heatmap: Correlation Matrix of Samples ($n \times n$)



Heatmap: Correlation Matrix of Genes ($p \times p$)



Two Sample t-Statistics: Student

Given: Samples $x = x_1, \ldots, x_n$ and $y = y_1, \ldots, y_m$ from populations with the *same* variance. Student t-statistic

$$T_s(x, y) \;=\; \frac{\overline{x} - \overline{y}}{\sqrt{s^2_{\text{pool}}(x, y)\left(n^{-1} + m^{-1}\right)}}$$

with pooled variance estimate

$$s^2_{\text{pool}}(x, y) \;=\; \frac{1}{n + m - 2}\left(\sum_{i=1}^{n}(x_i - \overline{x})^2 + \sum_{j=1}^{m}(y_j - \overline{y})^2\right)$$

Nominal p-values: $T_s(x, y) \sim t_{n+m-2}$

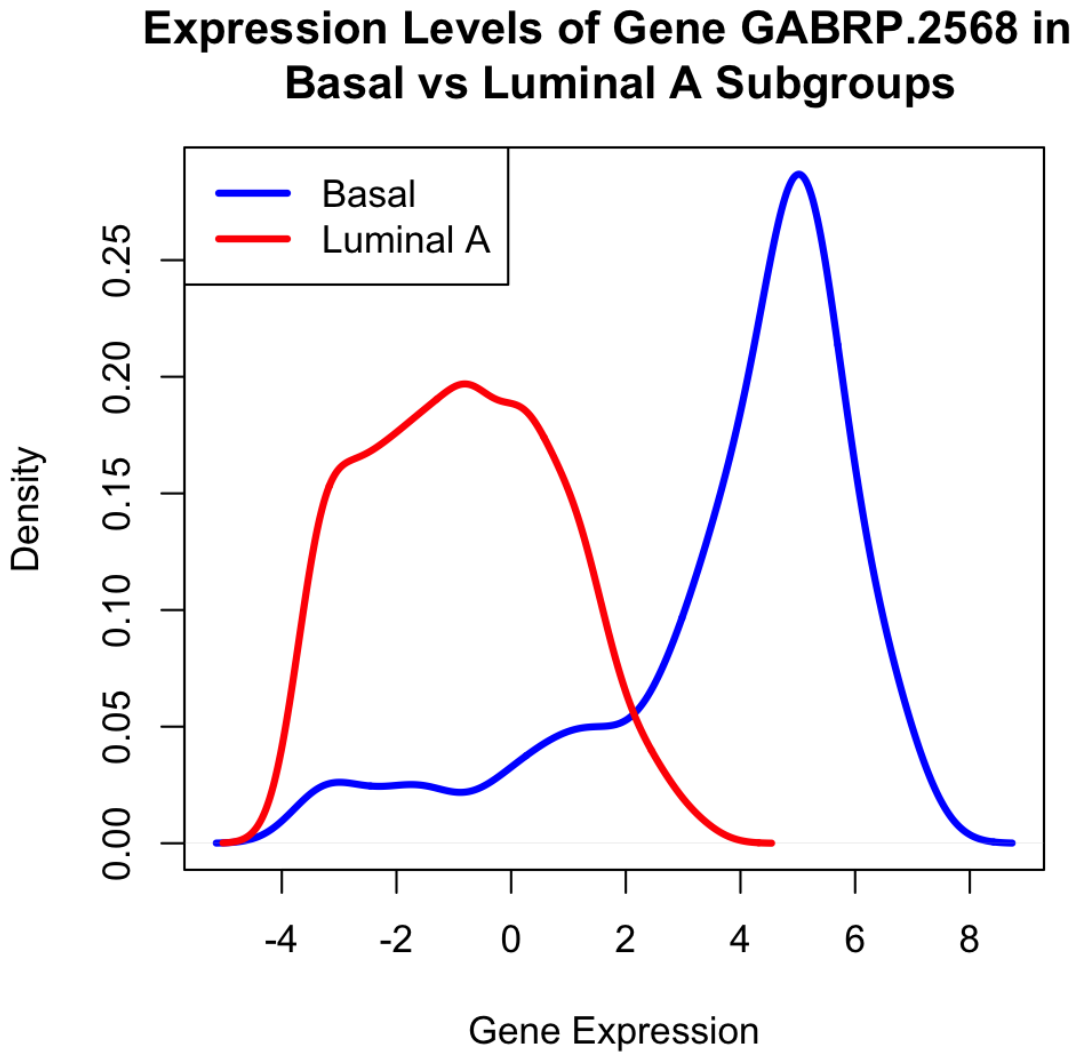
Two Sample t-Statistics: Welch

Given: Samples $x = x_1, \ldots, x_n$ and $y = y_1, \ldots, y_m$ from populations with *different* variances. Welch t-statistic

$$T_w(x, y) \;=\; \frac{\overline{x} - \overline{y}}{\sqrt{(n-1)^{-1} s^2(x) + (m-1)^{-1} s^2(y)}}$$

Nominal p-values: $T_s(x, y) \sim t_\nu$ where ν estimated from data

Example of Differential Gene Expression



Differential Expression Analysis

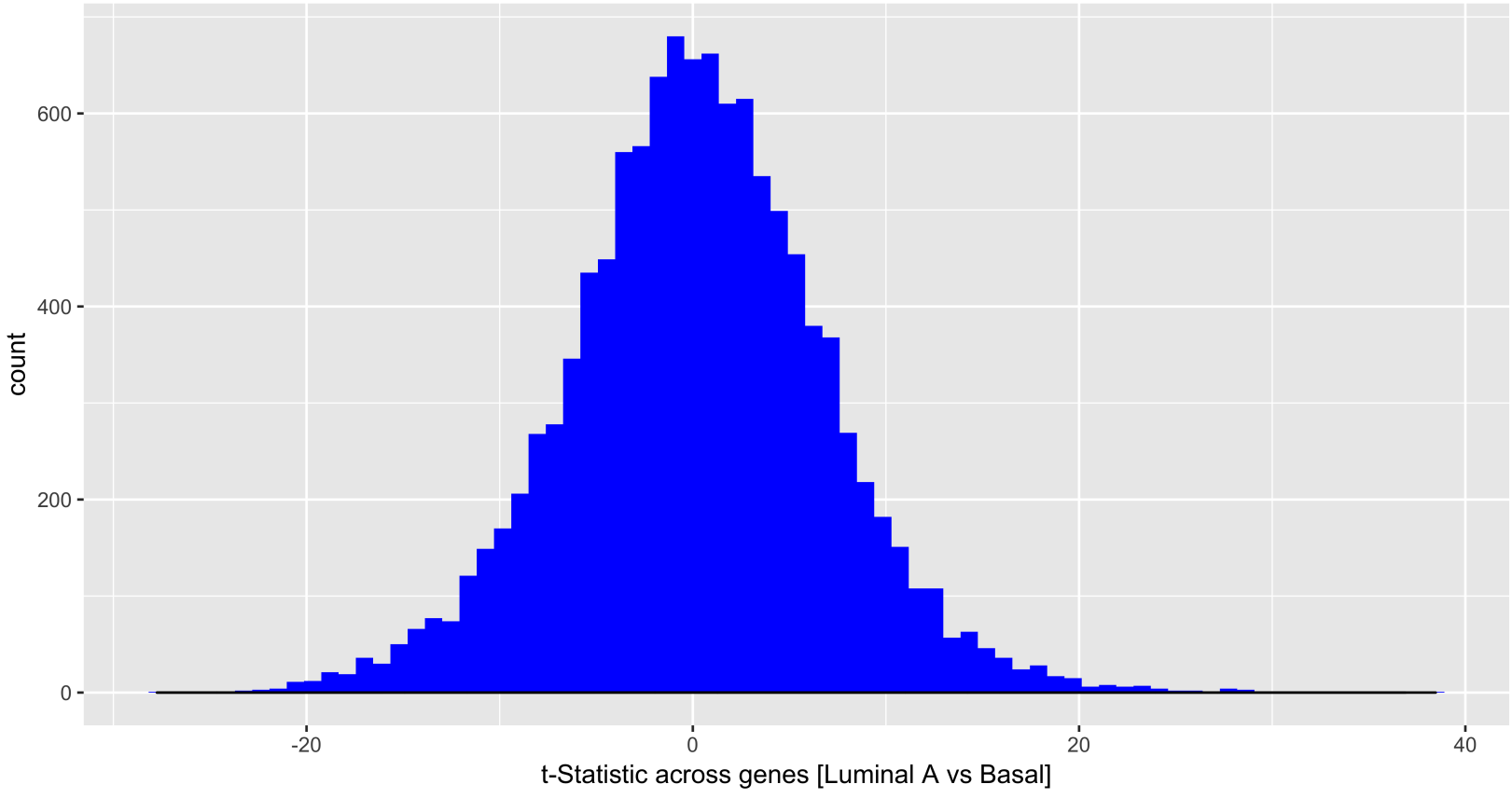
Step 1: For each gene

- Compute a two-sample t-statistic to assess whether it is differentially expressed in two cancer subtypes
- Use the t-statistic to assign a p-value assessing the evidence for differential expression

Step 2: Use a multiple testing procedure to identify genes that are differentially expressed. Procedures include

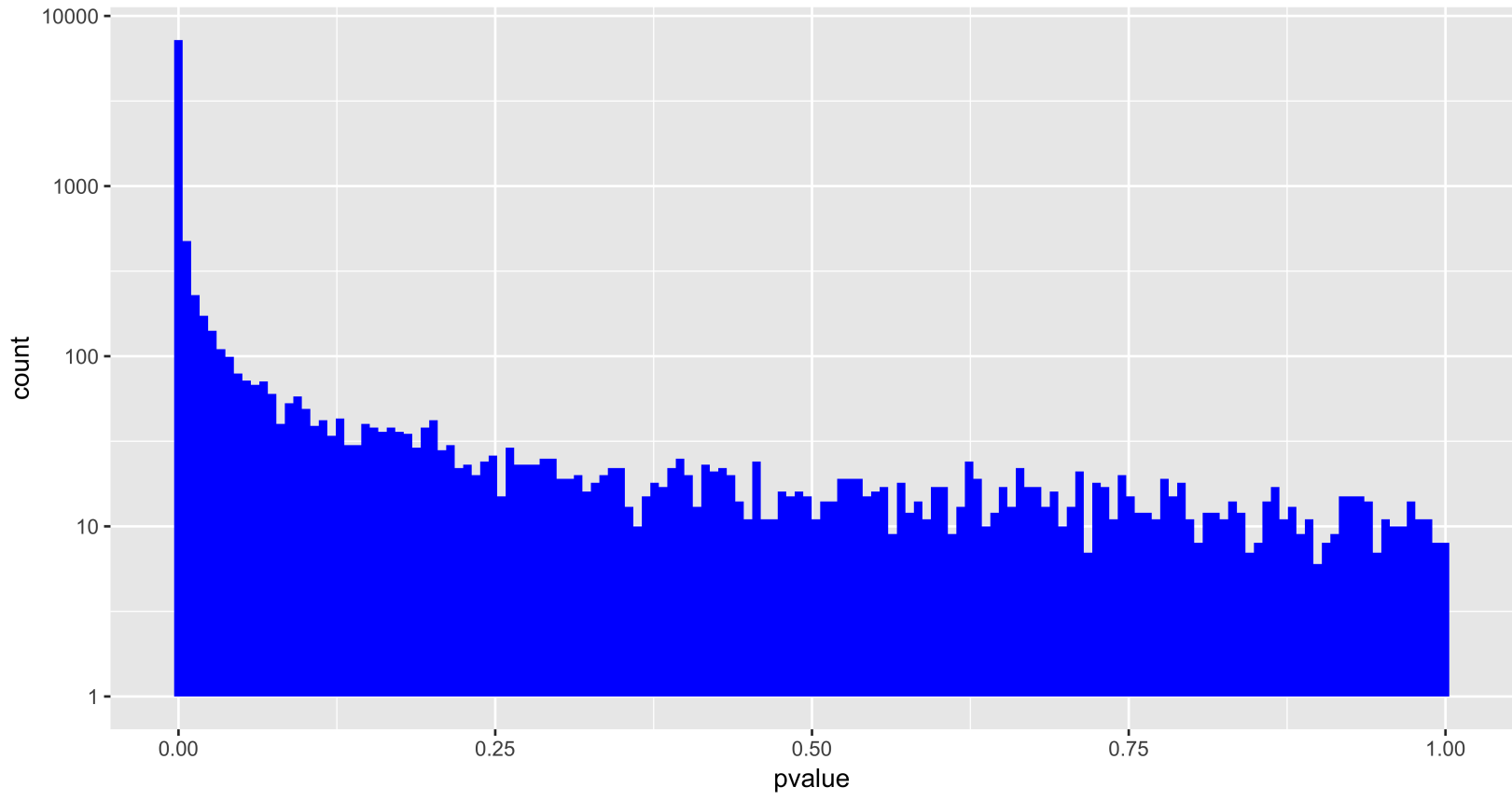
- Bonferroni correction (controls family-wise error rate)
- Benjamini-Hochberg (controls false discover rate)

t-Statistics for Differential Expression



$|t|$-cut-offs ($\alpha = .05$): 1.97 (no correction), 2.976 (FDR), 4.723 (FWER)

P-values for Differential Expression



p-cut-offs ($\alpha = .05$): $.05$ (no correction), $.00342$ (FDR), $4.38e-6$ (FWER)