Statistical analysis: concept, practice and interpretation

Maxwell Lee

High-dimension Data Analysis Group
Laboratory of Cancer Biology and Genetics
Center for Cancer Research
National Cancer Institute

April 27, 2018

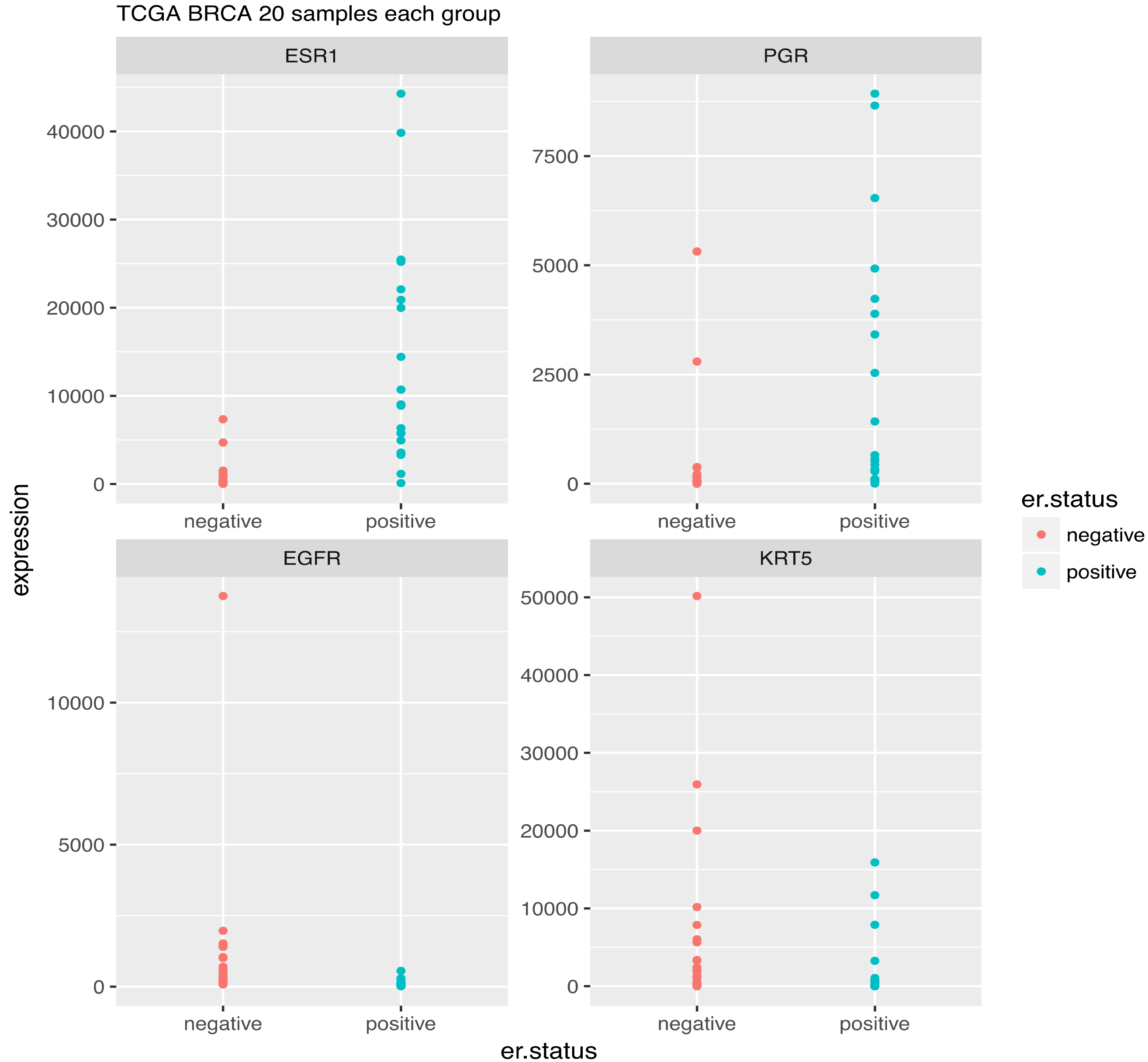
Outline of the talk

1) **Differential gene expression between two groups**
t-test, ANOVA, and linear modeling

2) **Association between two variables**
correlation, linear regression, and geometric representation

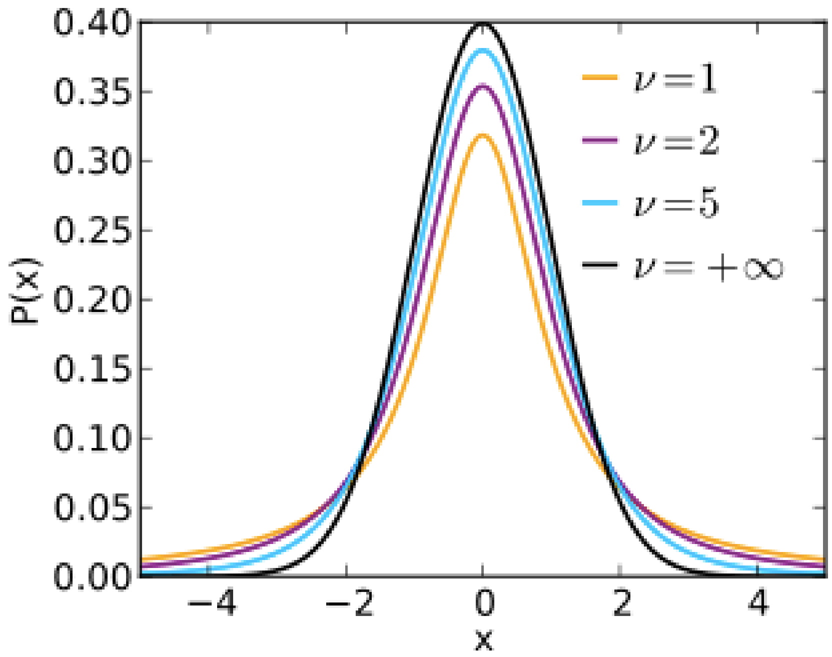
3) **Relationship between samples**
hierarchical clustering and PCA

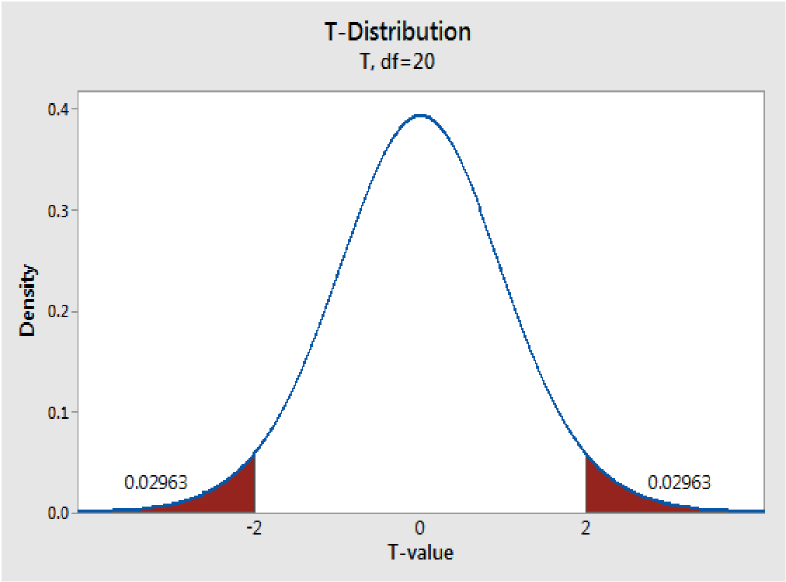
How do we know if the difference is statistically significant?



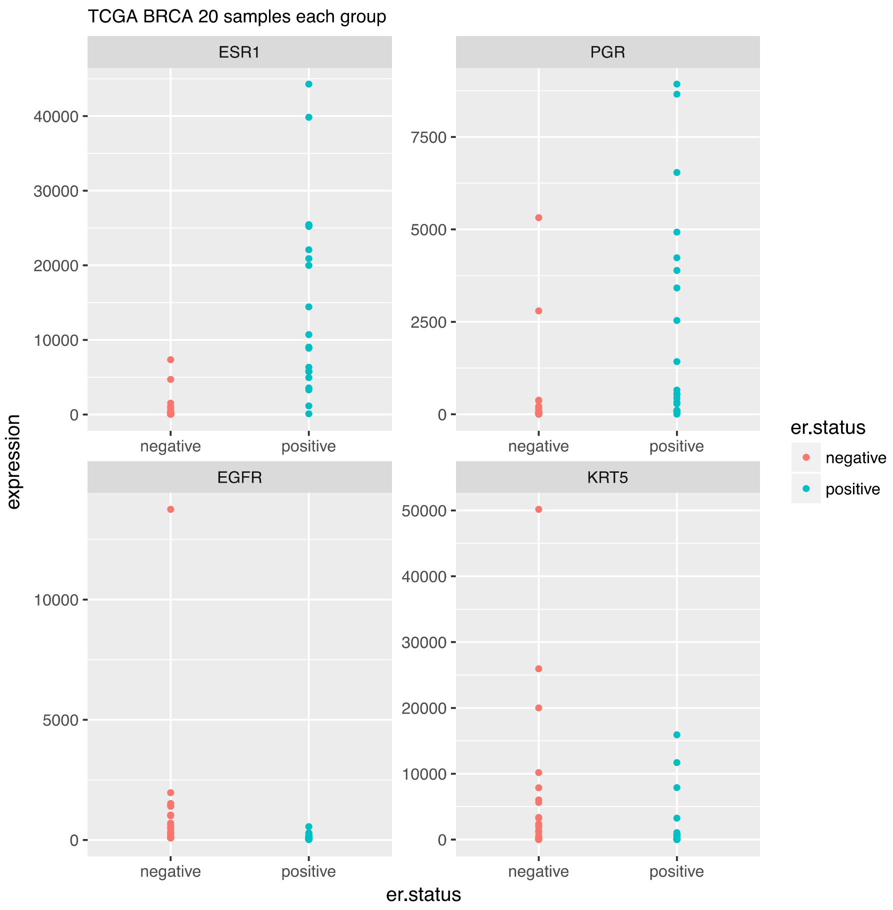
We use t-test to evaluate the difference between two groups

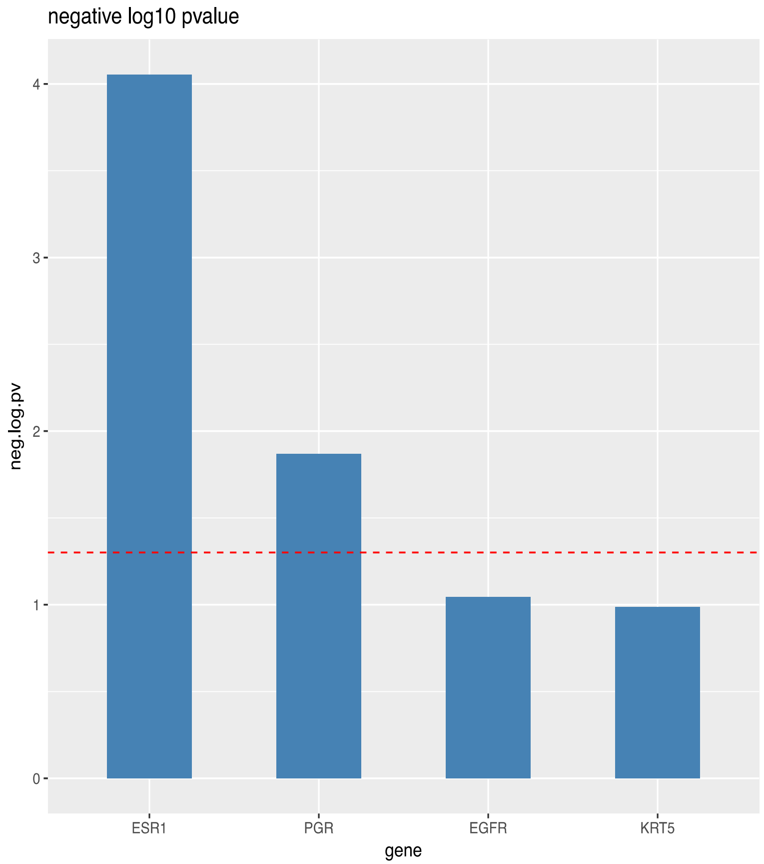
$$t = \frac{(x_1 - x_2)}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}}$$



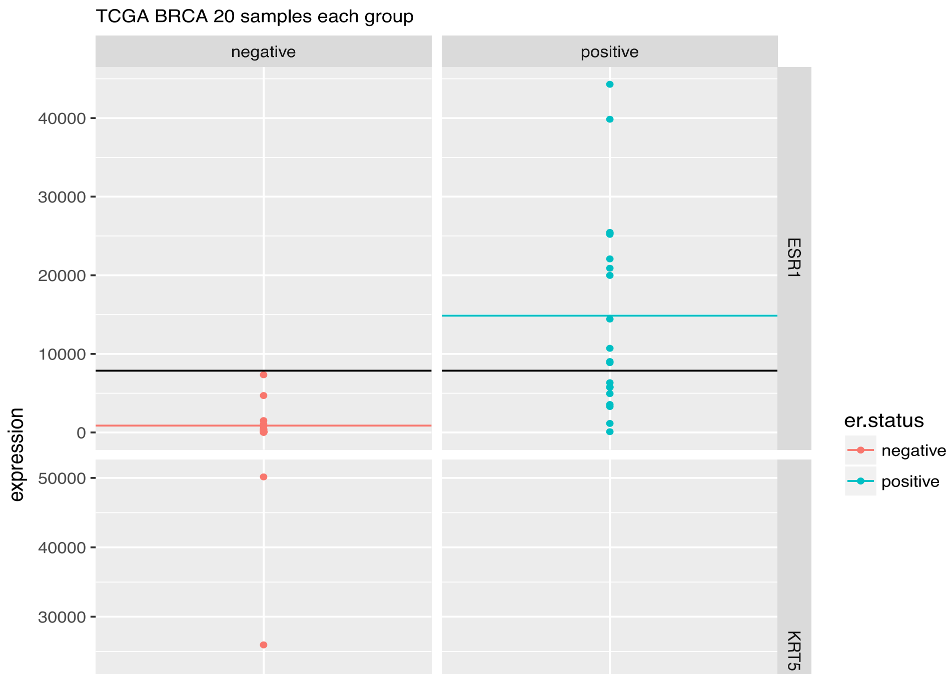


We use t-test to evaluate the difference between two groups





Analysis of variance (ANOVA)



$$SS_{total} = SS_{between} + SS_{within}$$

$$SS_{total} = \sum_{j=1}^{p}\sum_{i=1}^{n_j}(x_{ij} - \bar{x})^2$$

$$SS_{between} = \sum_{j=1}^{p} n_j(\bar{x}_j - \bar{x})^2$$

$$SS_{within} = \sum_{j=1}^{p}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2$$

Summary ANOVA

Source	Sum of Squares	Degrees of Freedom	Variance Estimate (Mean Square)	*F* Ratio
Between	SS_B	$K-1$	$MS_B = \frac{SS_B}{K-1}$	$\frac{MS_B}{MS_W}$
Within	SS_W	$N-K$	$MS_W = \frac{SS_W}{N-K}$	
Total	$SS_T = SS_B + SS_W$	$N-1$		

Conclusion of the part I

1) t statistic = $(X_1 – X_2)$ / stand error

2) t statistic provides an objective way for evaluating the statistical significance of the difference between the two groups.

3) F statistic from ANOVA can also be used to determine the statistical significance.

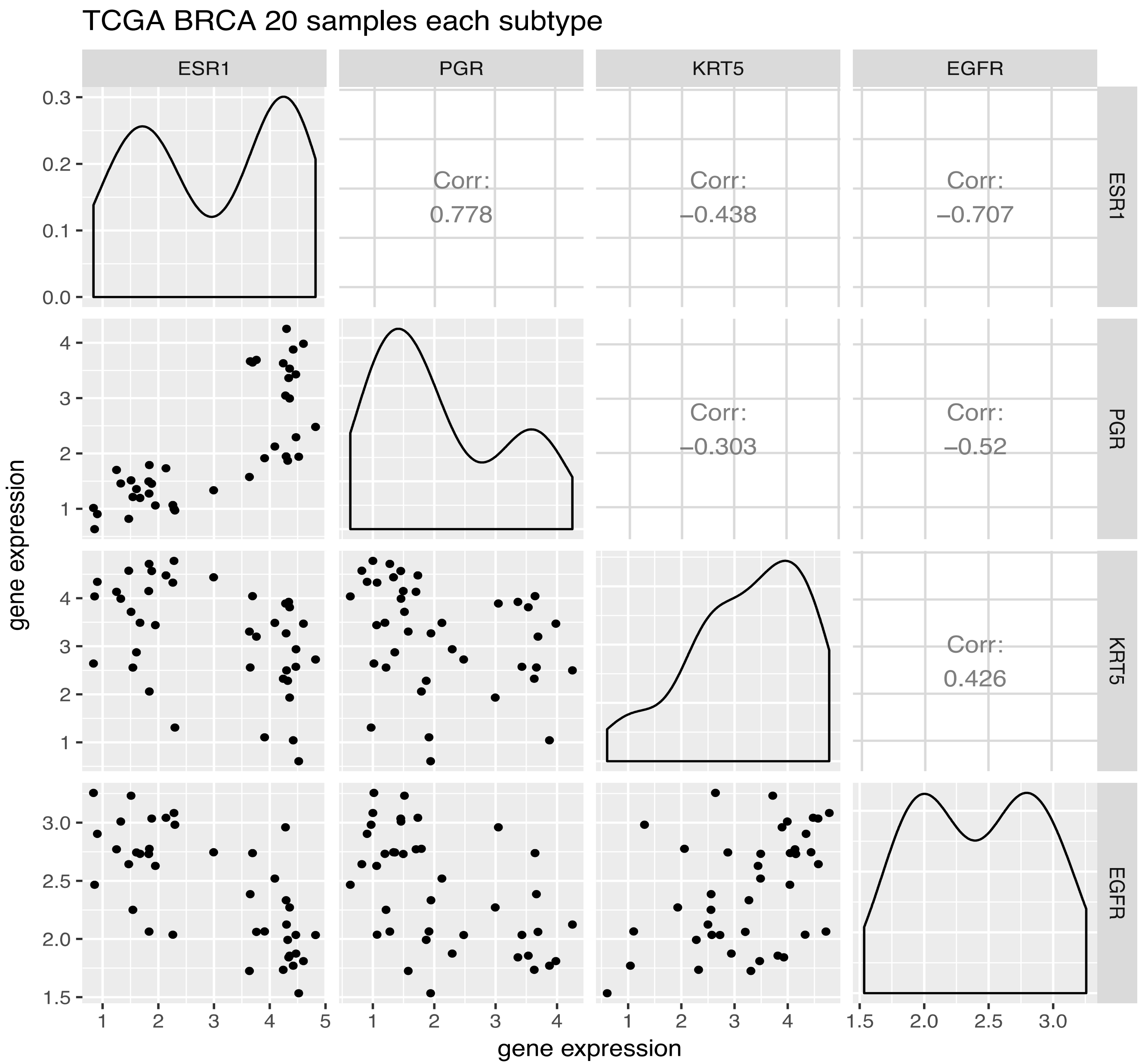
Outline of the talk

1) **Differential gene expression between two groups**
t-test, ANOVA, and linear modeling

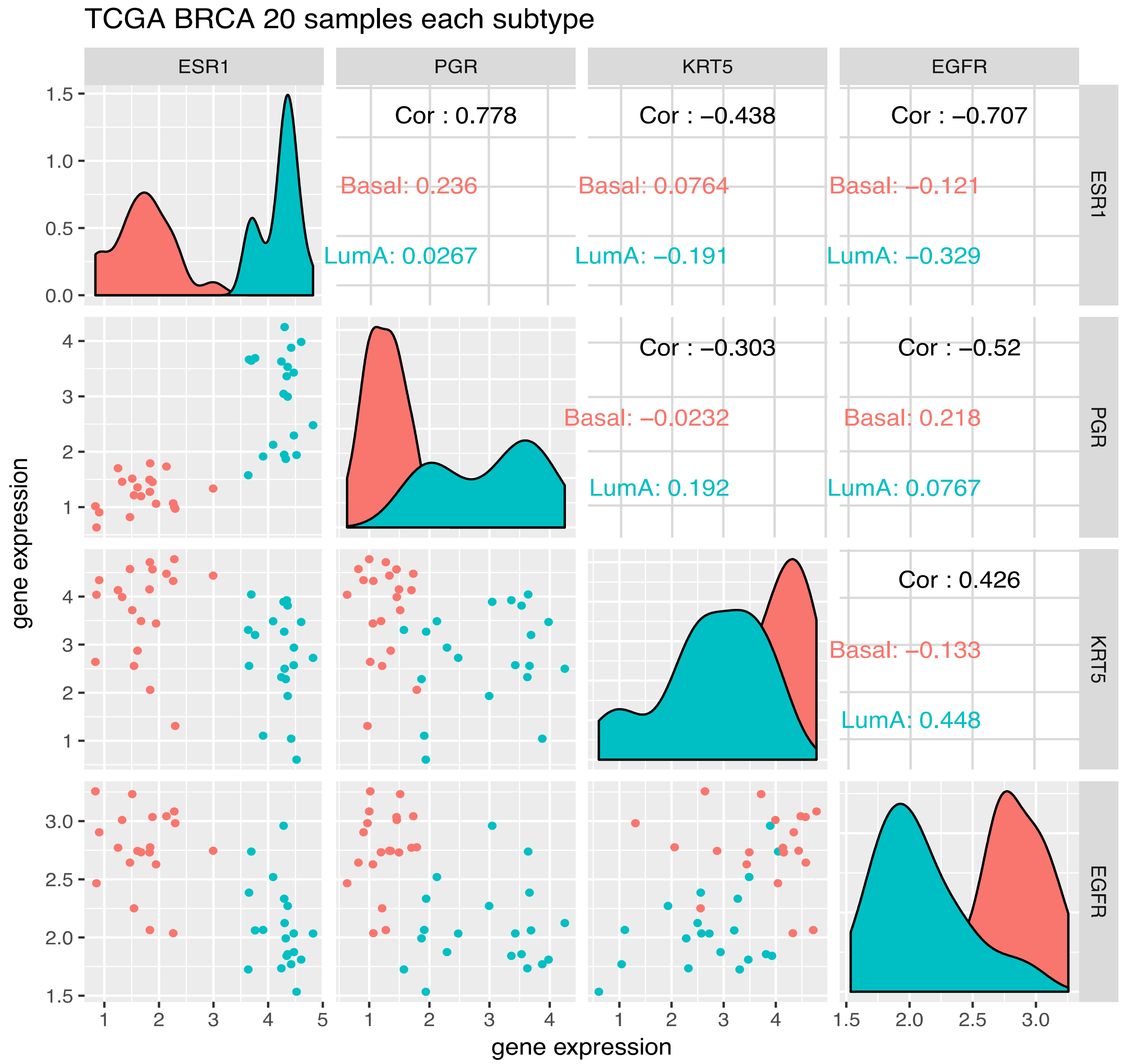
2) **Association between two variables**
correlation, linear regression, and geometric representation

3) **Relationship between samples**
hierarchical clustering and PCA

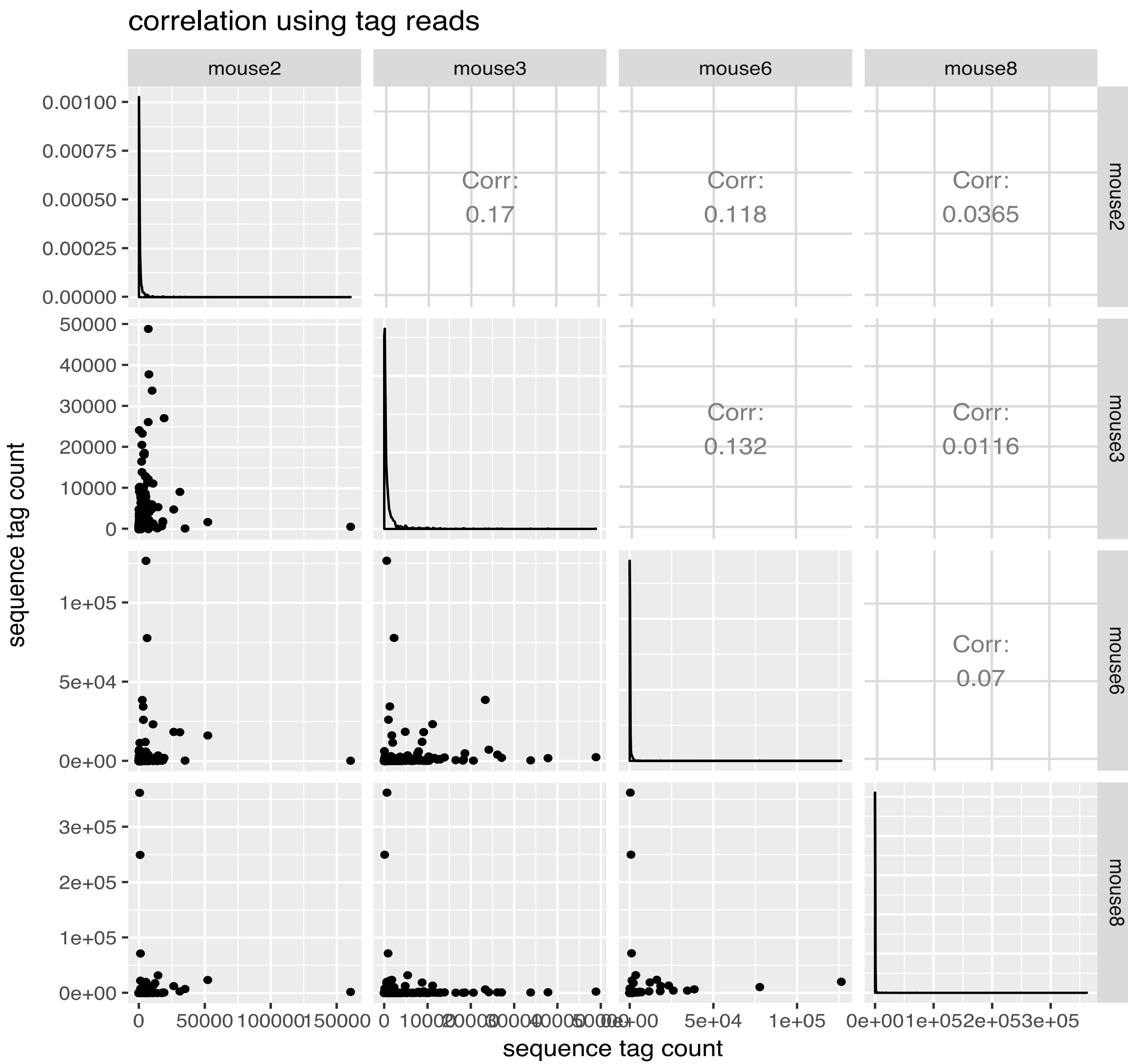
Correlation between variables



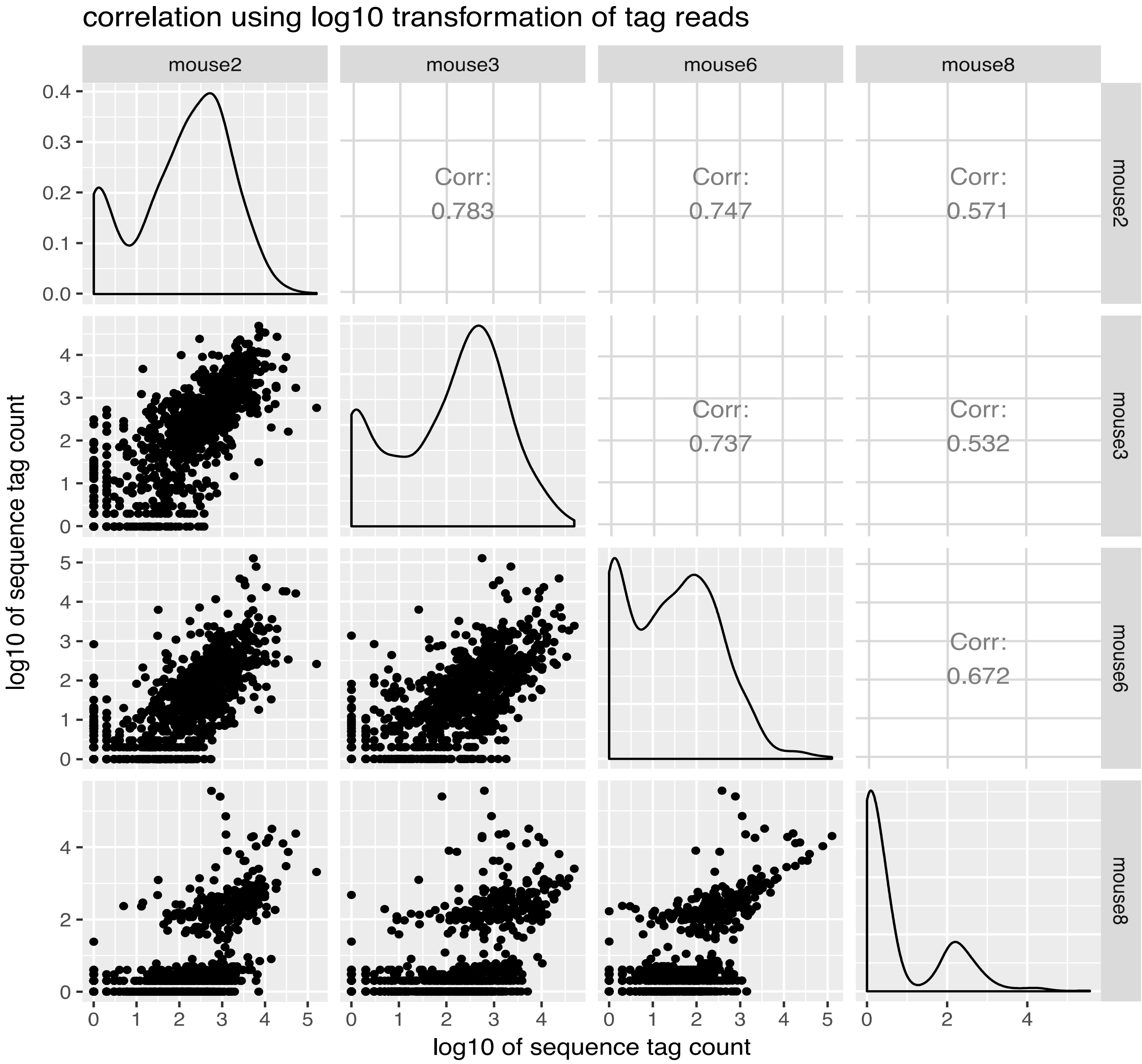
Be aware of different correlations across multiple levels



Correlation is sensitive to data scale: impact of log transformation

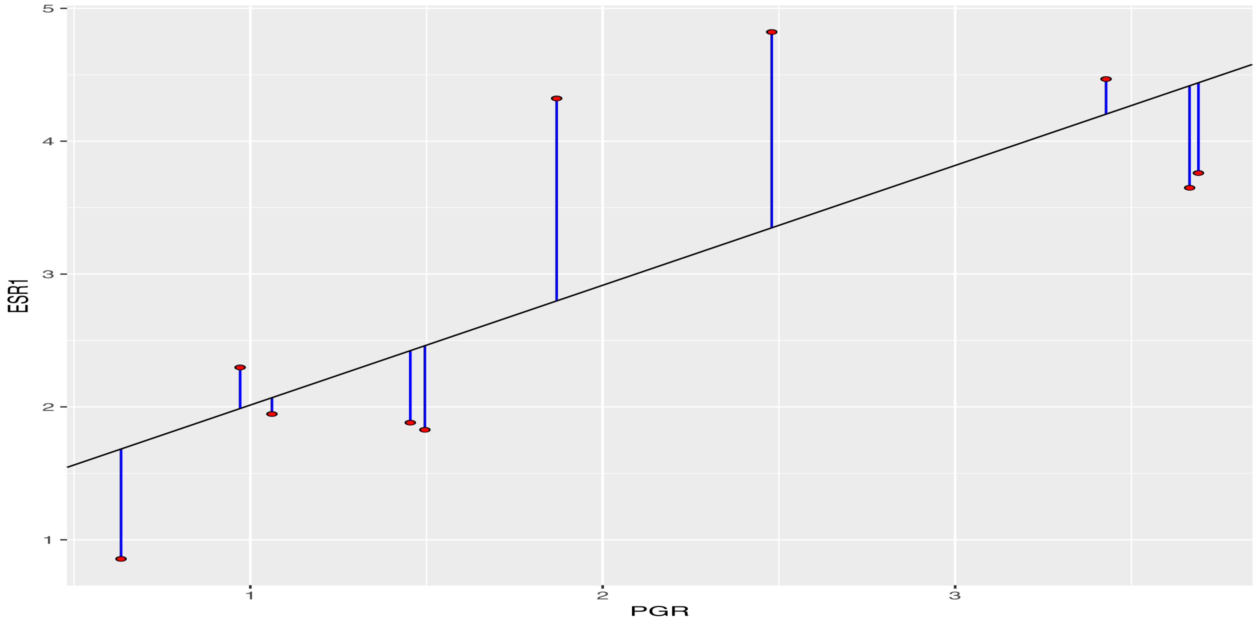


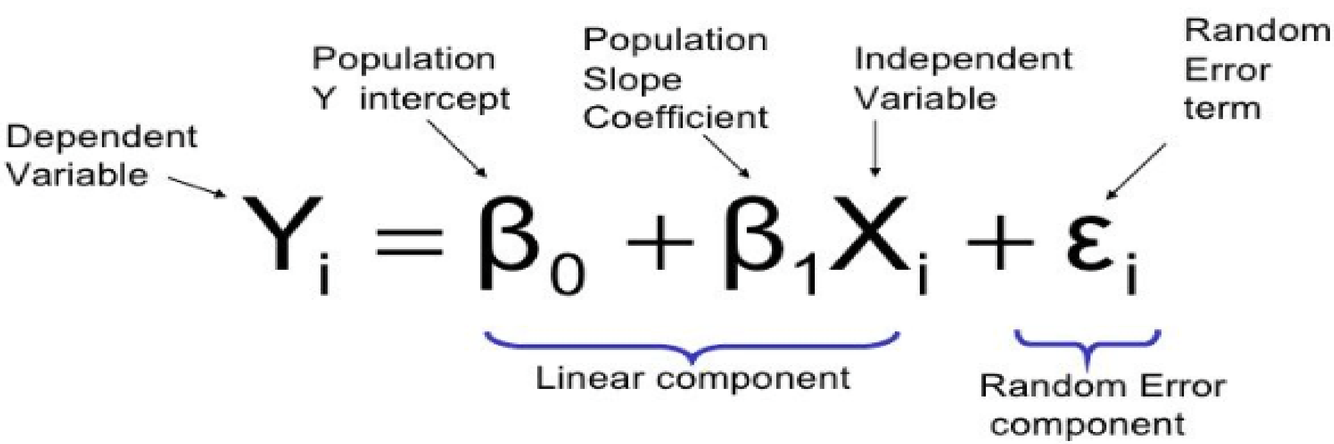
Correlation is sensitive to data scale: impact of log transformation



Simple linear regression model

$$Y = \beta_0 + \beta_1 X + e$$





Least squares solution

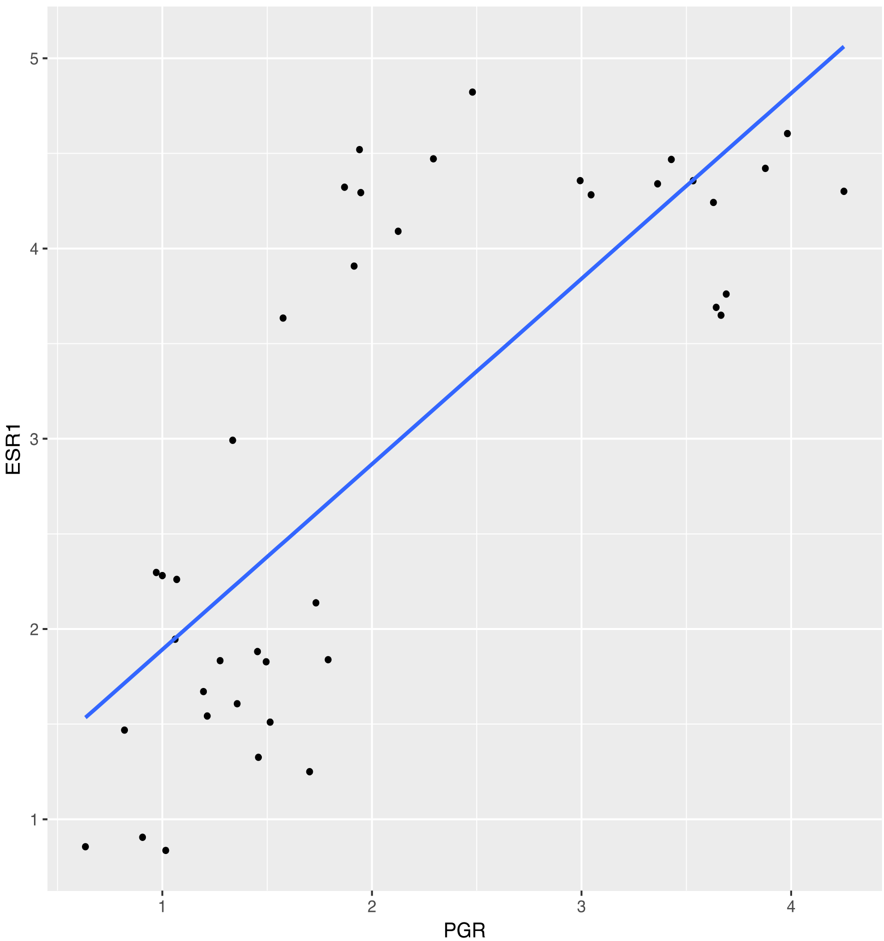
$$RSS = e_1^2 + e_2^2 + \ldots + e_n^2$$

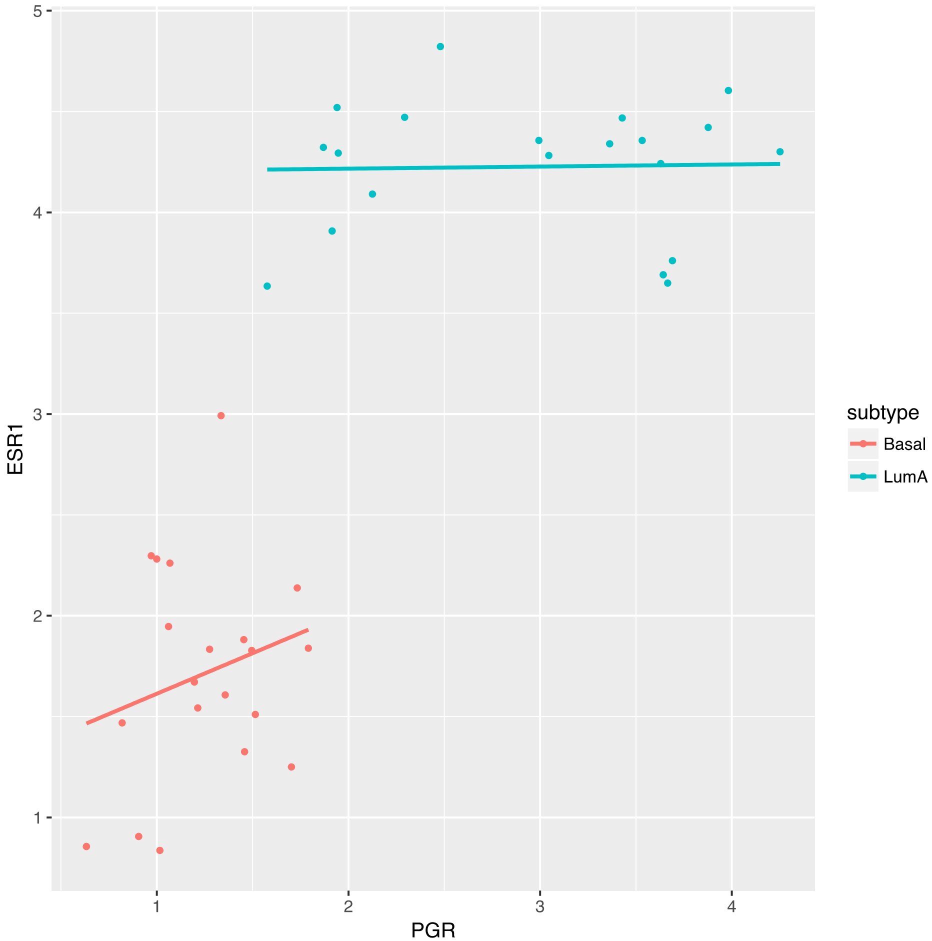
The least squares approach chooses β_0 and β_1 to minimize the RSS.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \beta_1 = r\, S_y/S_x$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

Linear regression is sensitive to data at multiple levels





Estimate	Std. Error	t value	Pr(>|t|)
9.747e-01	1.279e-01	7.623e+00	3.587e-09

subtype	Estimate	Std. Error	t value	Pr(>|t|)
Basal	0.40173	0.39035	1.0291	0.317
LumA	0.01045	0.09211	0.1134	0.911

Linear model: linear regression and ANOVA

$$Y = \beta_0 + \beta_1 X + e$$

Y	X	Type
continuous variable	continuous variable	linear regression
continuous variable	categorical variable	ANOVA

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$$

Y	X	Type
continuous variable	X_1 is continuous X_2 is categorical	ANCOVA

lm(ESR1 ~ PGR * subtype)

Multiple linear regression model

$$\begin{bmatrix} y_1 & x_{11} & x_{12} & \dots & x_{1p} \\ y_2 & x_{21} & x_{22} & \dots & x_{2p} \\ . & & & & . \\ . & & & & . \\ y_n & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

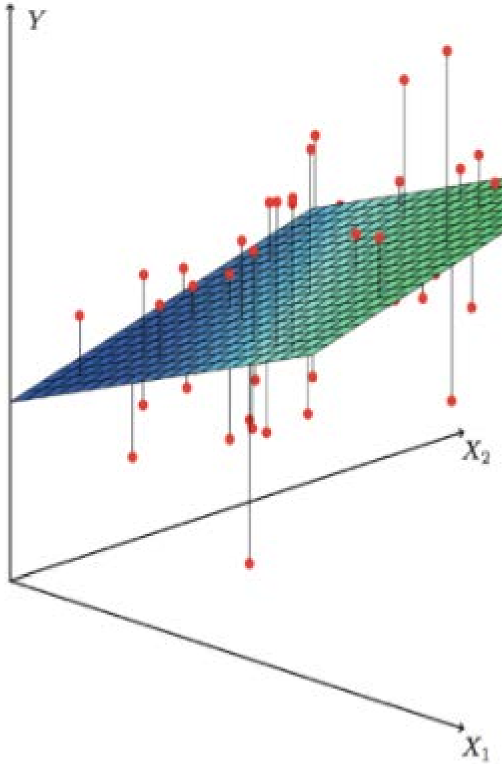
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \dots + \beta_p x_p + \varepsilon$$

$$y = X\beta + \varepsilon$$

$$RSS = (y - X\beta)^T (y - X\beta)$$

$$\beta = (X^T X)^{-1} X^T y$$

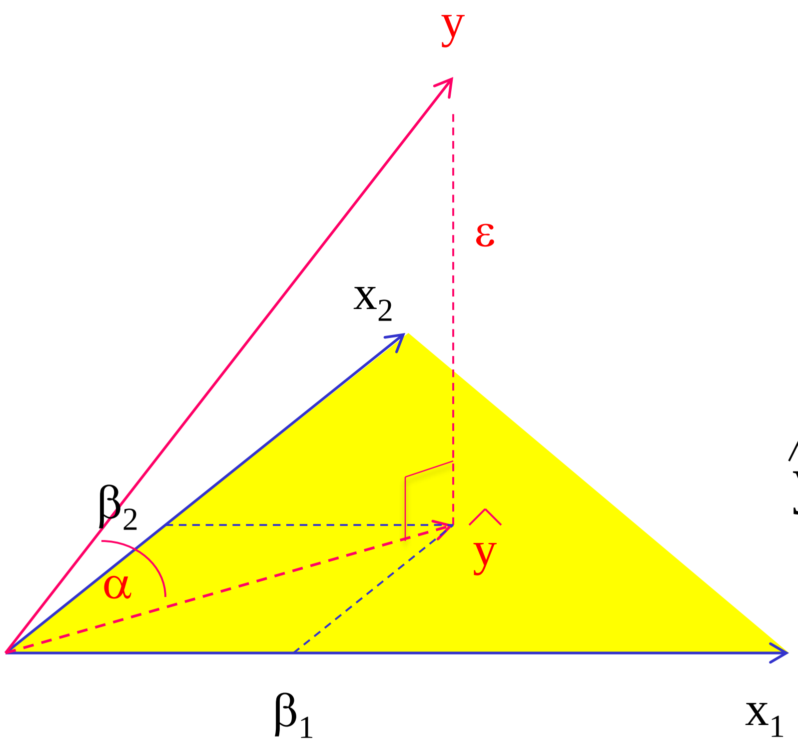
Multiple linear regression model: traditional representation



$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Multiple linear regression model: geometric representation

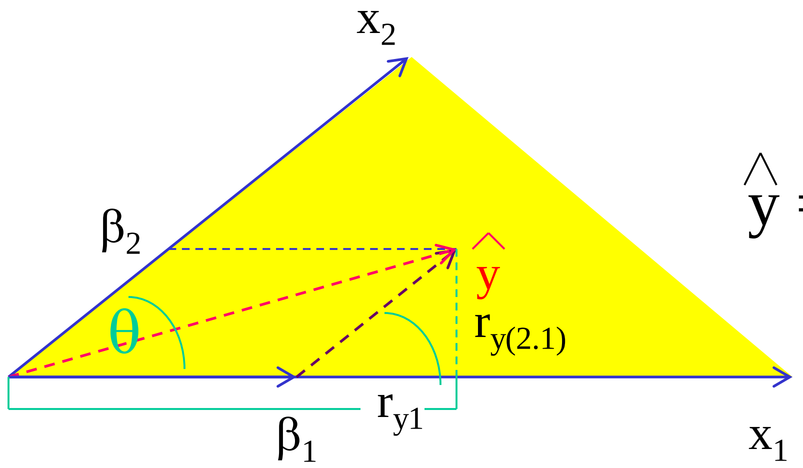


$$\hat{y} = \beta_1 x_1 + \beta_2 x_2$$

$$R^2 = 1 - (RSS/SST)$$

$$R = \cos(\alpha)$$

Multiple linear regression model: geometric representation



$$\hat{y} = \beta_1 x_1 + \beta_2 x_2$$

$$R^2 = r^2_{y1} + r^2_{y(2.1)}$$

$$r_{y(2.1)} = \beta_2 \sin(\theta)$$

Generalized linear model: linear regression and classification

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_p x_p + \varepsilon$$

Linear Regression and ANOVA

Y	X	Type
continuous variable	continuous variable	linear regression
continuous variable	categorical variable	ANOVA

Classification

Y	X	Type
categorical variable	continuous or categorical	classification

Conclusion of the part II

1) We can use correlation to evaluate association between two variables. The correlation is sensitive to data consisting of heterogeneous groups and data transformation.

2) We can also use regression model to evaluate association between two variables. Similarly, regression analysis is sensitive to data scale and transformation.

3) The linear model is a powerful approach. It can be used for regression, ANOVA, and classification.

4) Geometric representation can provide insights into the understanding of linear regression analysis.

Outline of the talk

1) **Differential gene expression between two groups**
t-test, ANOVA, and linear modeling

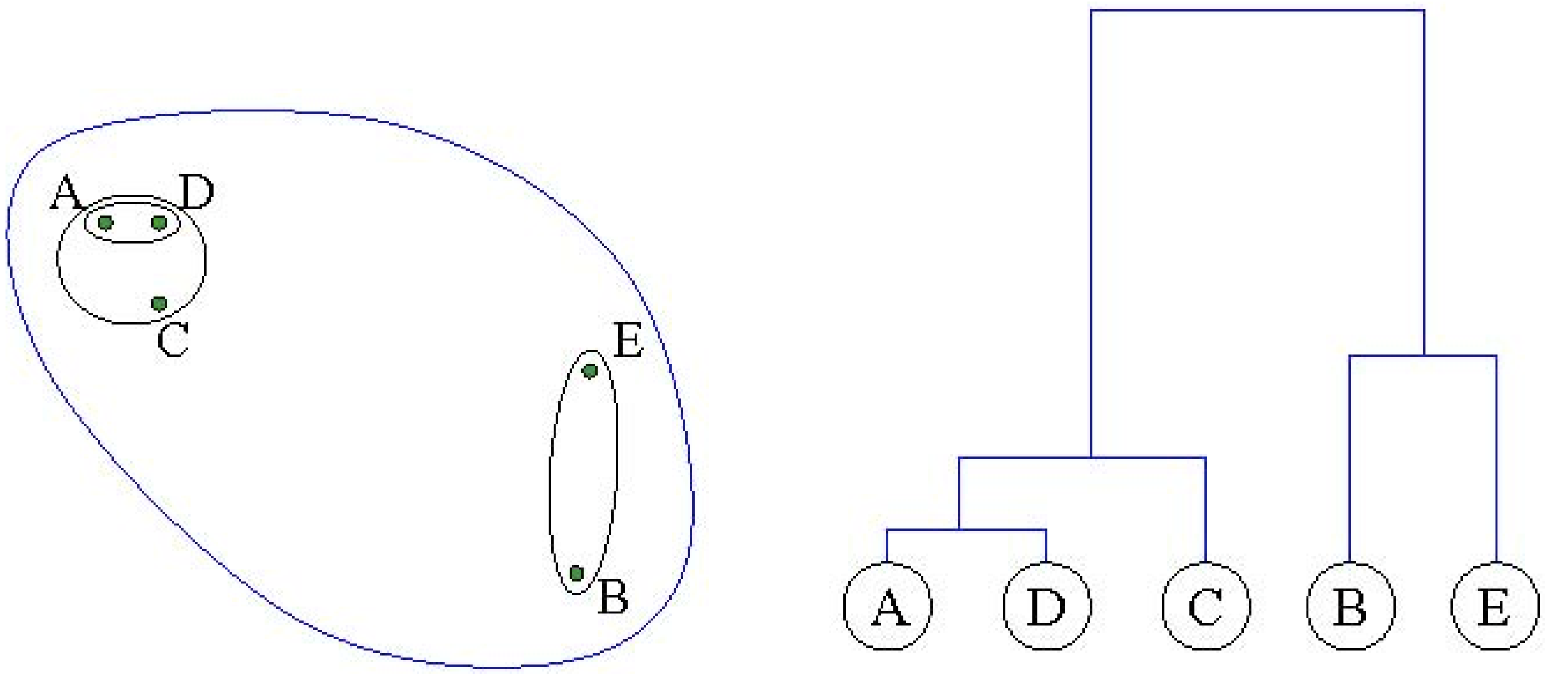
2) **Association between two variables**
correlation, linear regression, and geometric representation

3) **Relationship between samples**
hierarchical clustering and PCA

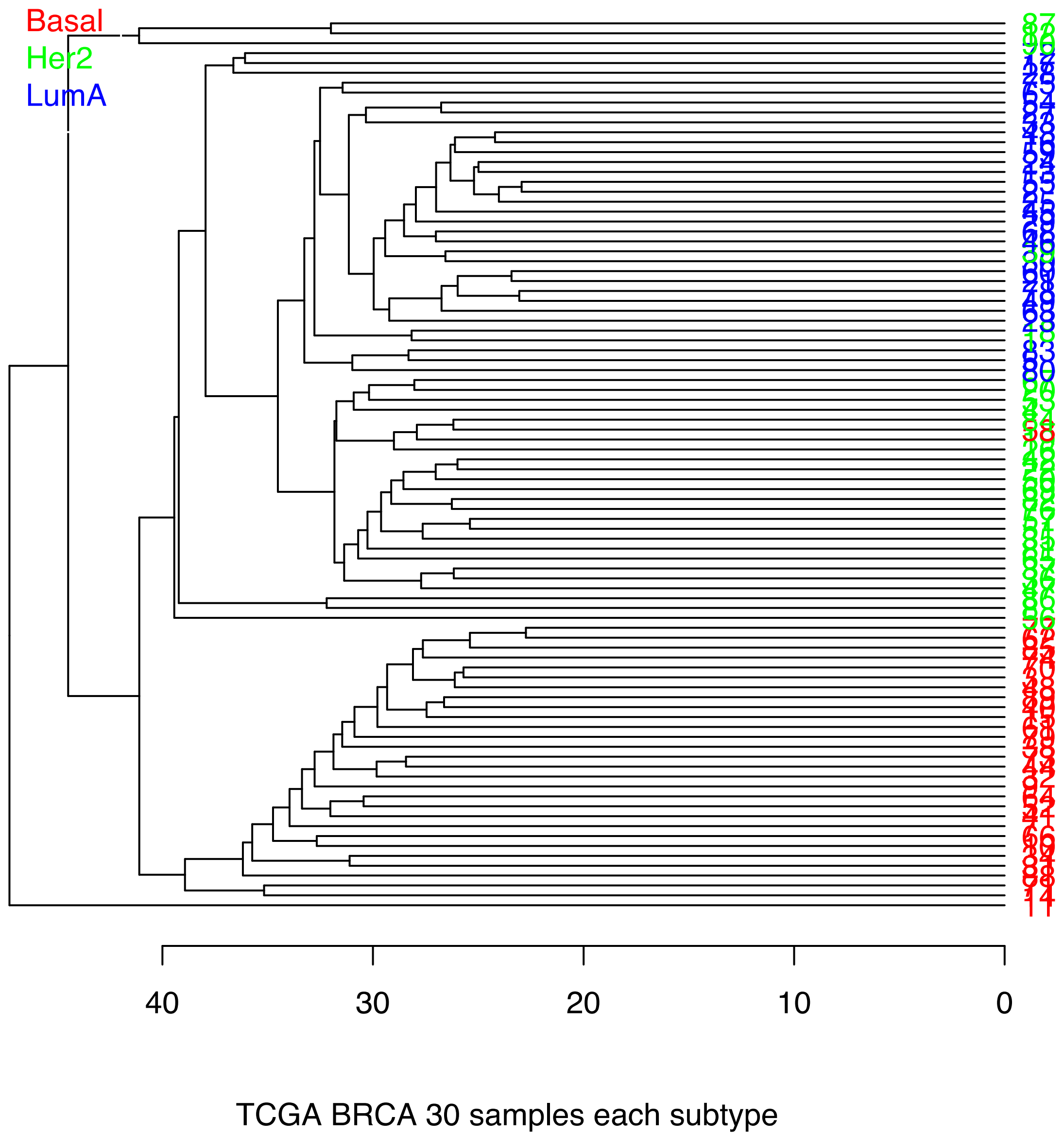
Supervised and unsupervised statistical learning

- supervised
  - continuous variable
    - Regression model
      linear regression, polynomial regression, glm
  - categorical variable
    - Classification
      glm, LDA, KNN
- unsupervised
  - Clustering analysis
    hierarchical clustering, k-means
  - Dimension reduction
    PCA, MDS, t-SNE

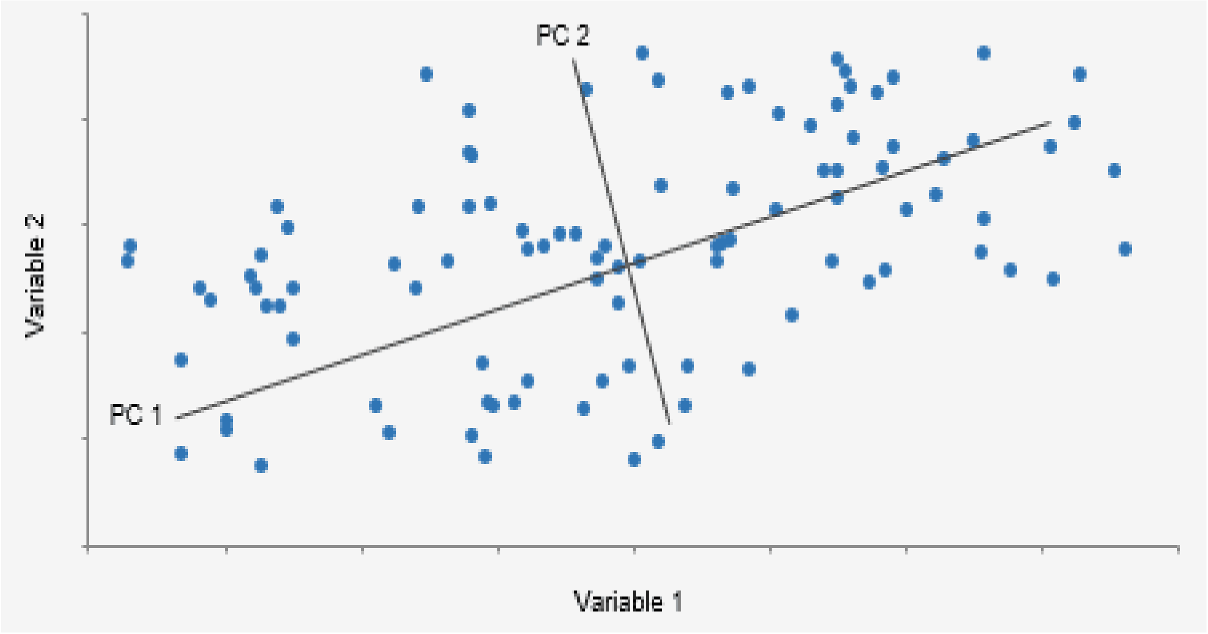
Hierarchical clustering analysis



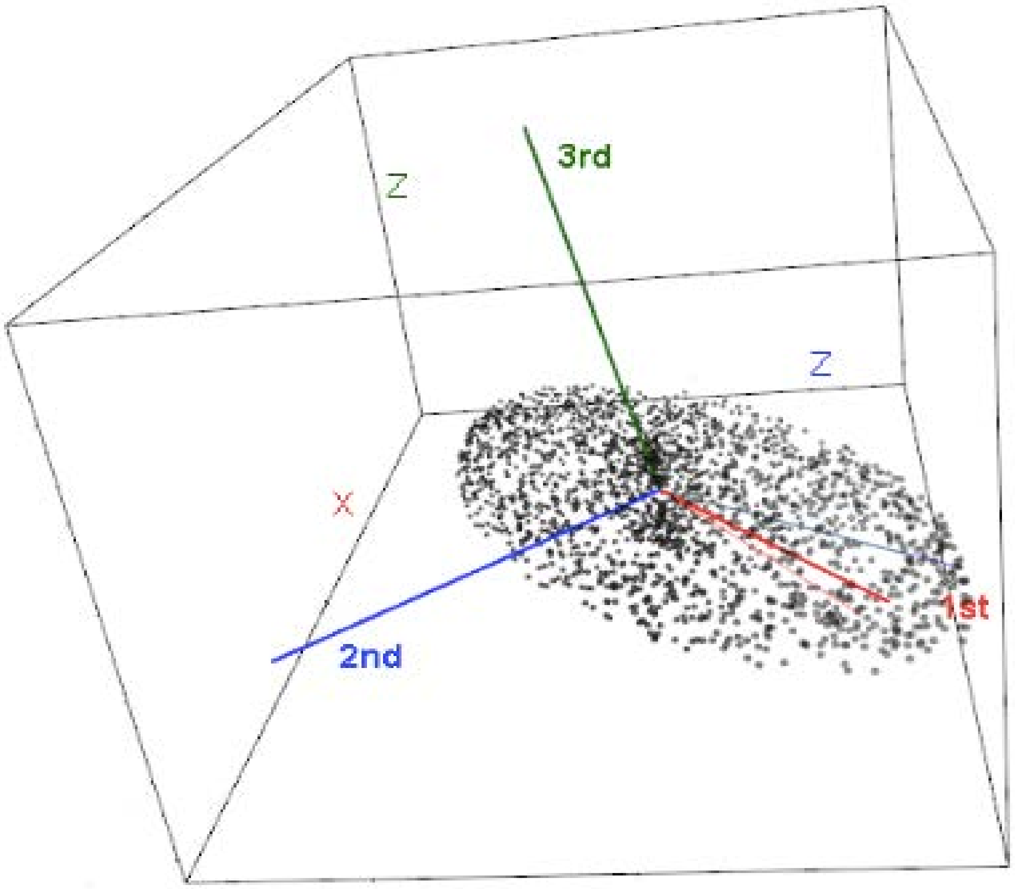
Hierarchical clustering analysis of TCGA samples



Principal component analysis (PCA)



Principal component analysis (PCA)



Algorithm of PCA

$$z_1 = Xu_1$$
$$z_2 = Xu_2$$
$$z_3 = Xu_3$$

$$Z = XU$$

$$\mathrm{var}(Z) = (XU)^T XU$$

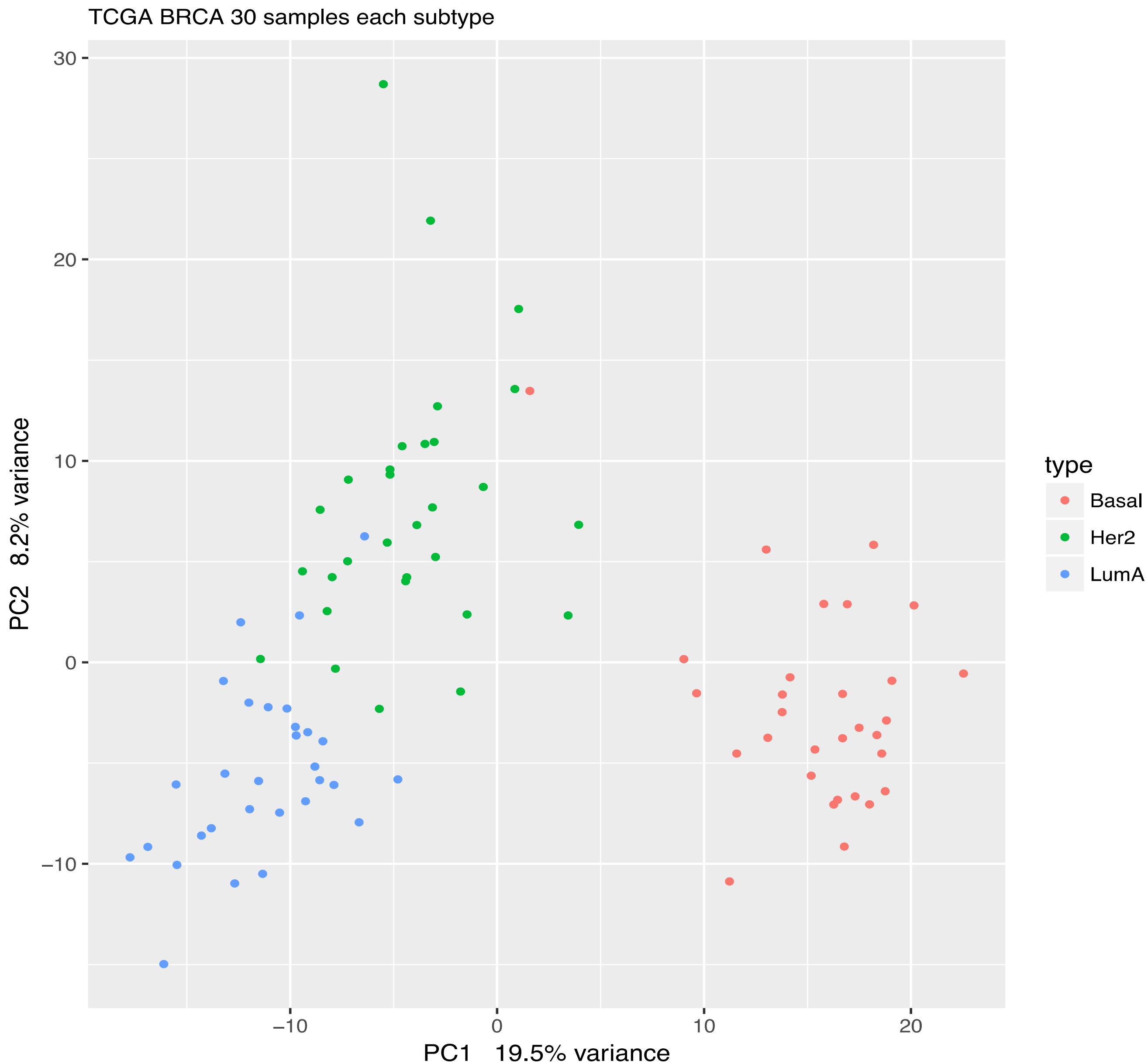
$$\mathrm{var}(Z) = U^T X^T XU = U^T RU$$

Choose U to maximize $U^T RU$
subject to $U^T U = I$

$$RU = \lambda U$$

U is the eigenvector and λ is eigenvalue

PCA analysis of TCGA samples



Conclusion of the part III

1) We can use hierarchical clustering to evaluate relationship among samples.

2) PCA involves the rotation of the coordinates so that PC1 captures the direction where samples have the largest variance, followed by PC2, PC3, and so on.

3) Each PC is a linear combination of the original variables. Ideally, the first a few PC components should capture most of the variance in the samples.