



Integrative Analysis of Circulating Tumor Cell Counts and Gene Expression Levels

Zhenya Cherkas, PhD
Janssen R&D
Springhouse, PA

PhUSE US Connect 2019 AB03

Dorethey Gorham, *Day of the Armada* Dorethey is a joyful, self-taught artist living with arthritis, general anxiety syndrome, and diabetes.

Disclaimer

- Employment by Janssen R&D

Disclaimer

- Employment by Janssen R&D
- Used R/Rstudio for analysis

Disclaimer

- Employment by Janssen R&D
- Used R/Rstudio for analysis
- Project was done to support a collaboration of Janssen Oncology, University of Pennsylvania and City of Hope

Disclaimer

- Employment by Janssen R&D
- Used R/Rstudio for analysis
- Project was done to support a collaboration of Janssen Oncology, University of Pennsylvania and City of Hope
- Focus on data science

Domain Knowledge

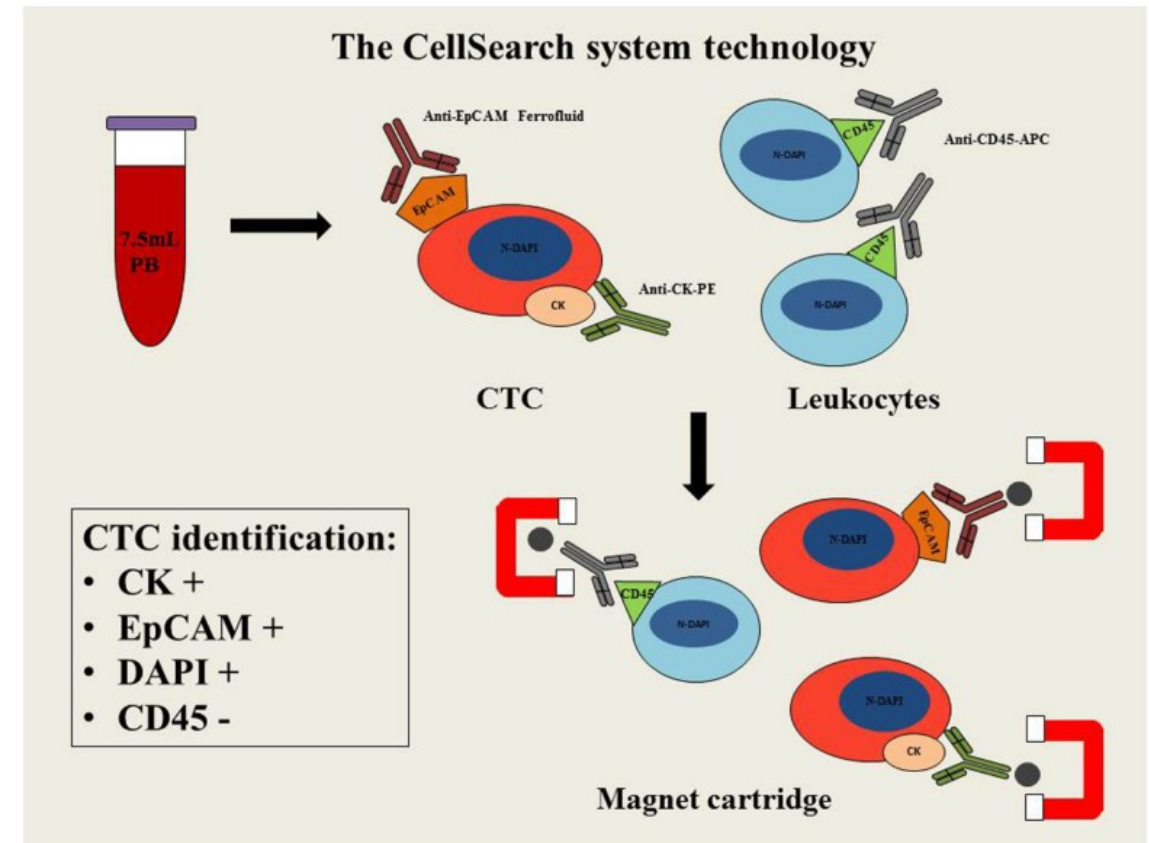
- Oncology:
 - Advanced prostate cancer
 - Mechanisms of resistance

Domain Knowledge

- Oncology:
 - Advanced prostate cancer
 - Mechanisms of resistance
- Biomarkers:
 - Circulating Tumor Cells (CTC)
 - Gene Expression Data

Domain Knowledge

- Oncology:
 - Advanced prostate cancer
 - Mechanisms of resistance
- Biomarkers:
 - **Circulating Tumor Cells (CTC)**
 - Gene Expression Data



Domain Knowledge

- Oncology:
 - Advanced prostate cancer
 - Mechanisms of resistance
- Biomarkers:
 - Circulating Tumor Cells (CTC)
 - **Gene Expression Data**
 - Genes from previous findings

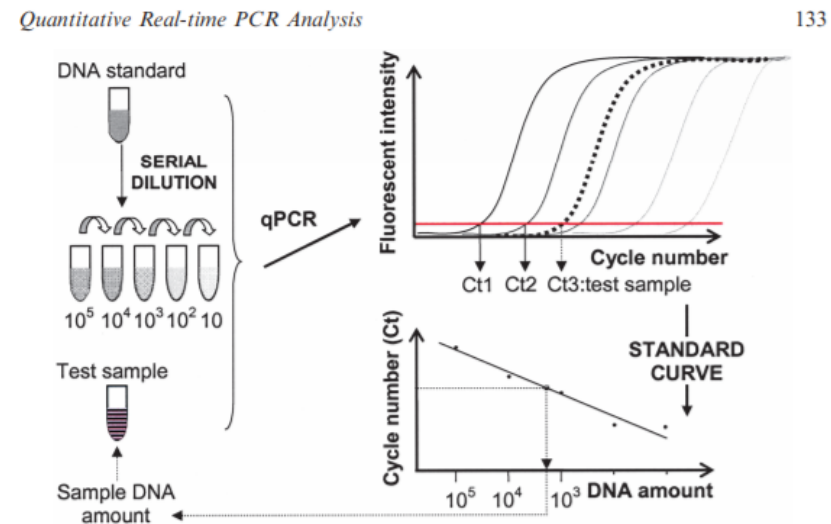


Figure 7.1 Schematic representation of the basis of real-time PCR using a quantitative standard.

The Process

1. Relevant Data Collection, Pre-processing and Cleaning
2. Exploratory Data Analysis
3. Analysis and Modeling

Data Collection, Pre-processing and Cleaning

- Import
- Tidy
- Transform

Data Collection, Pre-processing and Cleaning

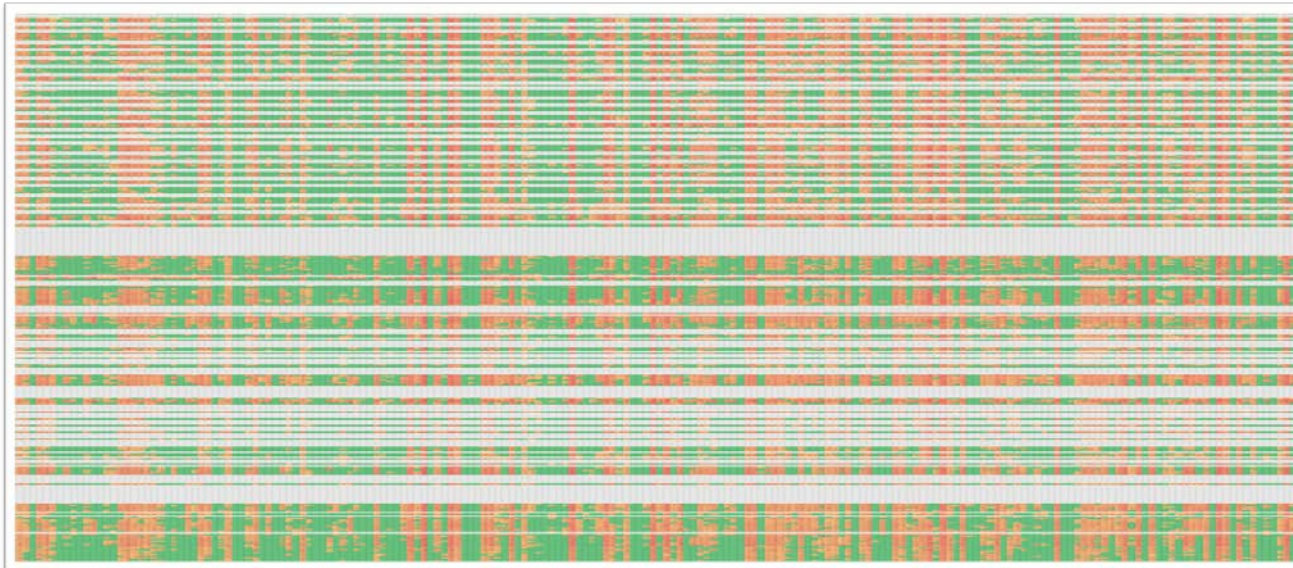
- Import
- Tidy
- Transform

R Tips and tricks:

- The Tidyverse is an opinionated collection of R packages designed for data science
- A tibble is a modern take on the `data.frame`

Data Collection, Pre-processing and Cleaning

- Import
- Tidy
- Transform
- Bird's Eye view of the expression data



R Tips and tricks:

- The Tidyverse is an opinionated collection of R packages designed for data science
- A tibble is a modern take on the data.frame

Exploratory Data Analysis

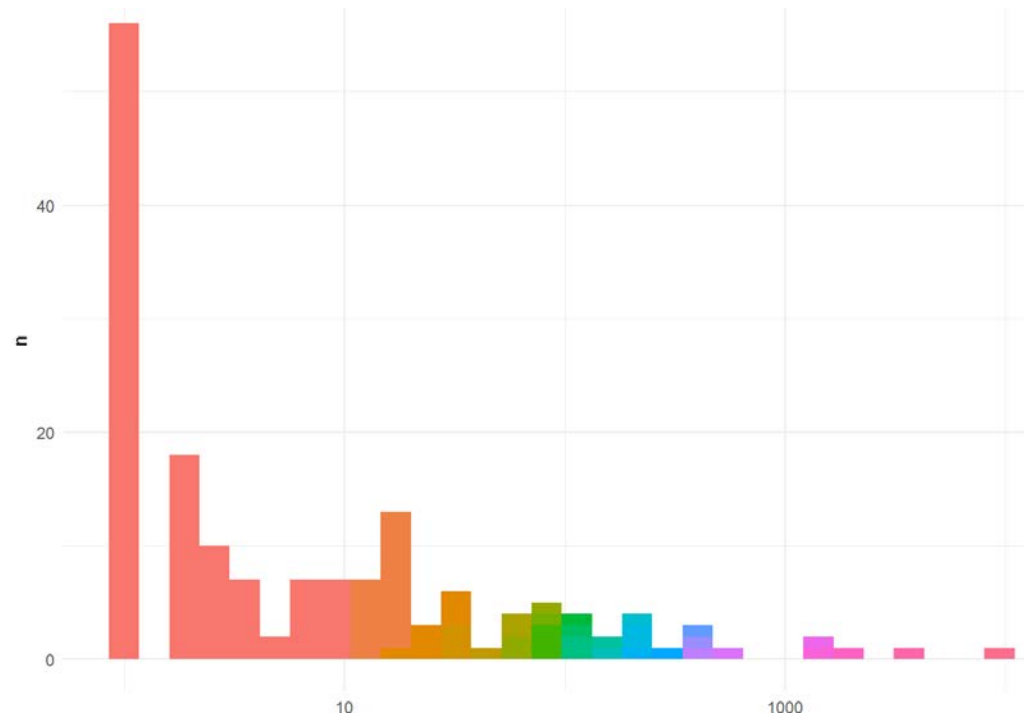
- Detection of mistakes
- Checking of assumptions
- Preliminary selection of appropriate models
- Determining relationships between variables

Exploratory Data Analysis

- Cell counts

Exploratory Data Analysis

- Cell counts



0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
56	18	10	7	2	5	2	1	3	3	1	3	1	2	3	3	2	1	3	1

R Tips and tricks:

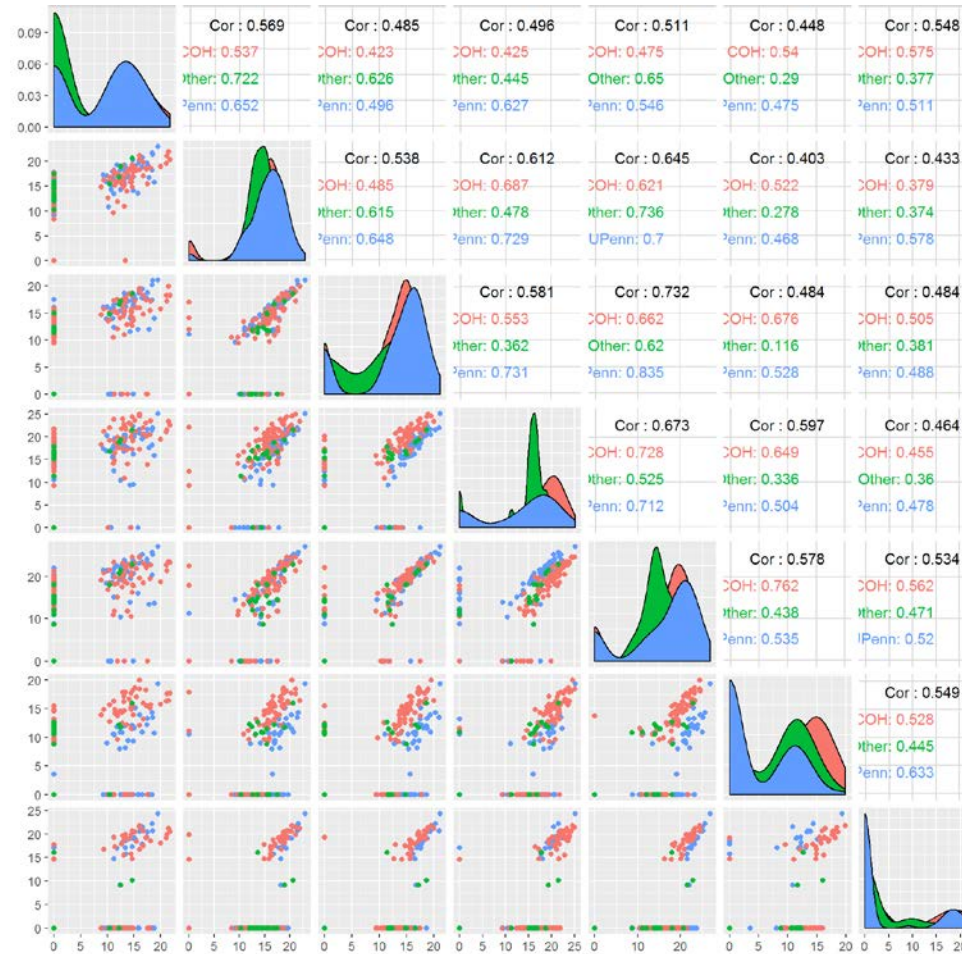
- `geom_histogram`
- `table`

Exploratory Data Analysis

- Expression Data
 - ~190 genes
 - Genes with near zero variance filtered
 - ~140 genes

Exploratory Data Analysis

- Expression Data
 - ~190 genes
 - Genes with near zero variance filtered
 - ~140 genes

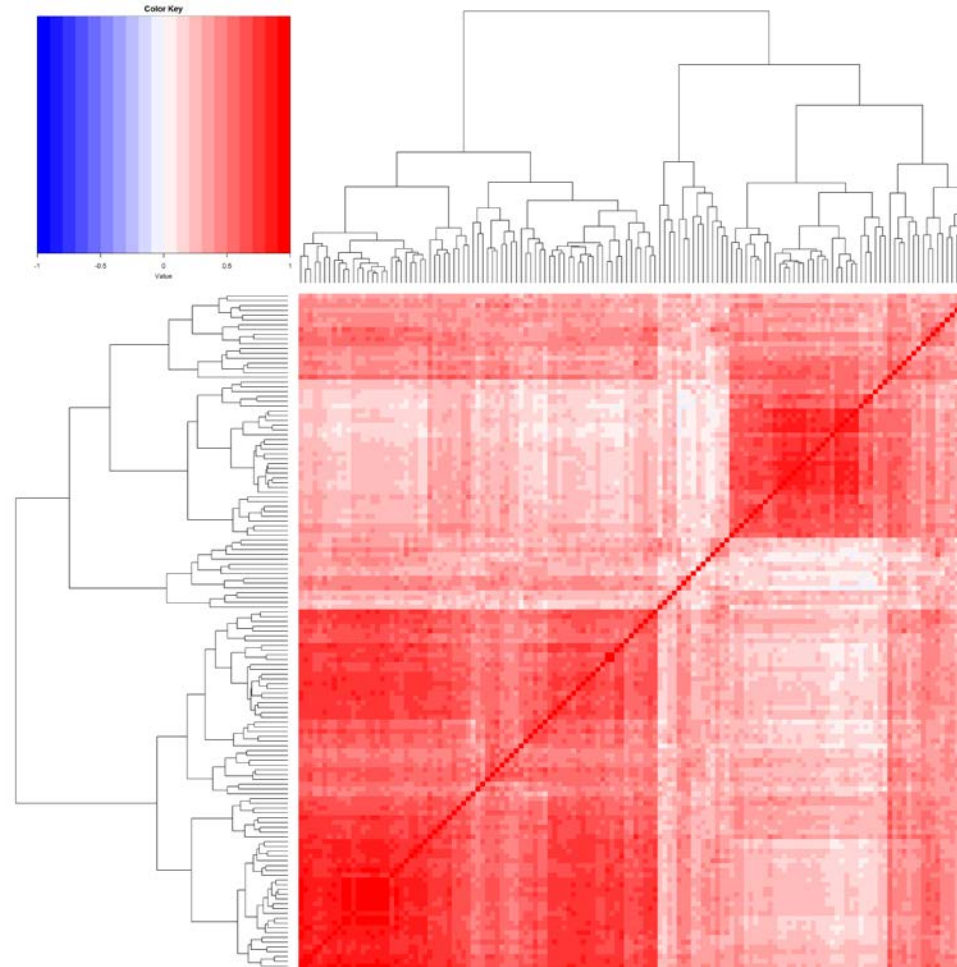


R Tips and tricks:

- ggpairs

Exploratory Data Analysis

- Expression Data
 - ~190 genes
 - Genes with near zero variance filtered
 - ~140 genes

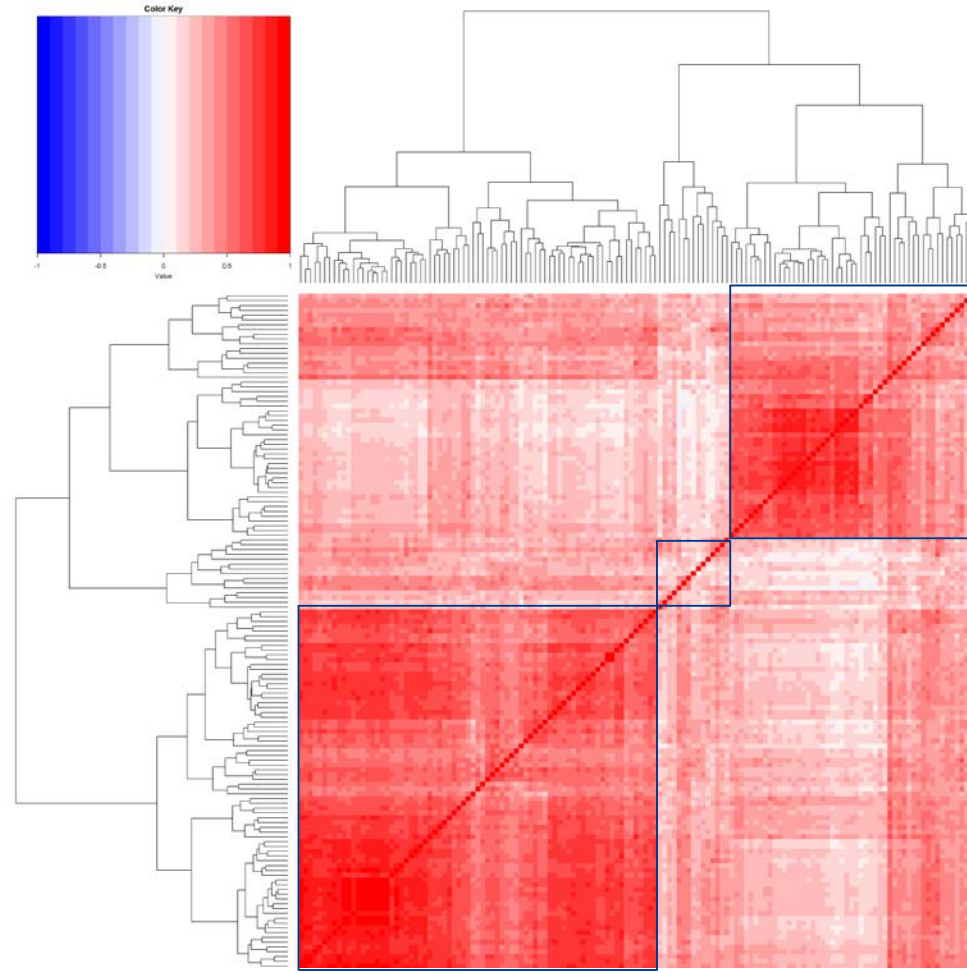


R Tips and tricks:

- `heatmap.2`

Exploratory Data Analysis

- Expression Data
 - ~190 genes
 - Genes with near zero variance filtered
 - ~140 genes

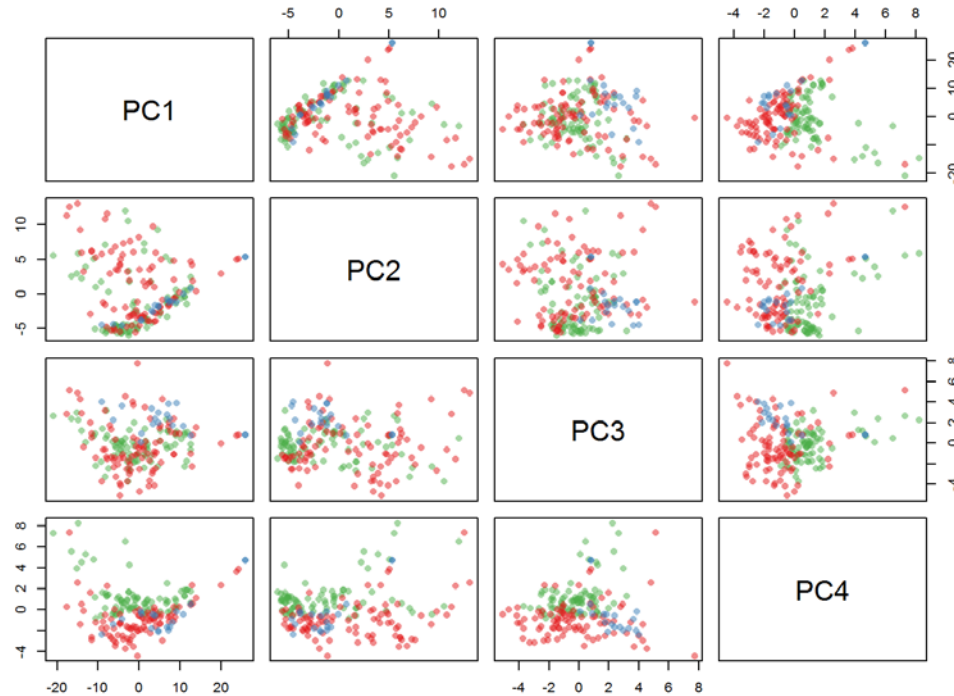


R Tips and tricks:

- `heatmap.2`

Exploratory Data Analysis

- Expression Data
 - ~190 genes
 - Genes with near zero variance filtered
 - ~140 genes

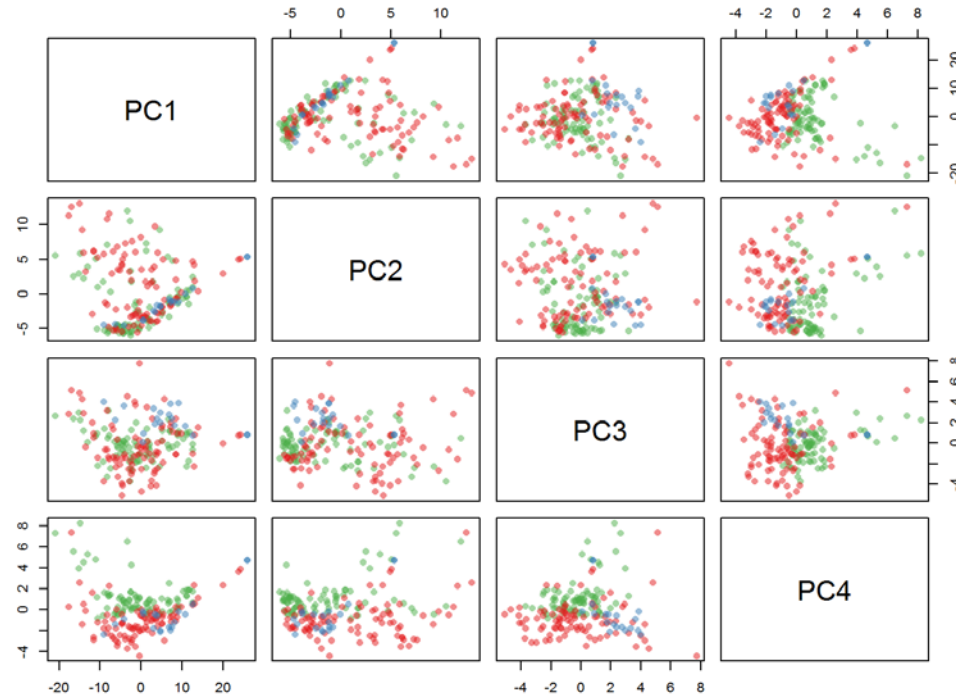


R Tips and tricks:

- `princomp`

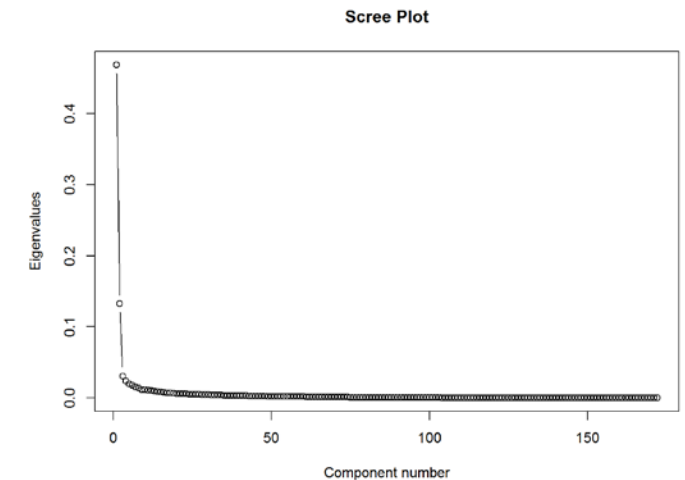
Exploratory Data Analysis

- Expression Data
 - ~190 genes
 - Genes with near zero variance filtered
 - ~140 genes



R Tips and tricks:

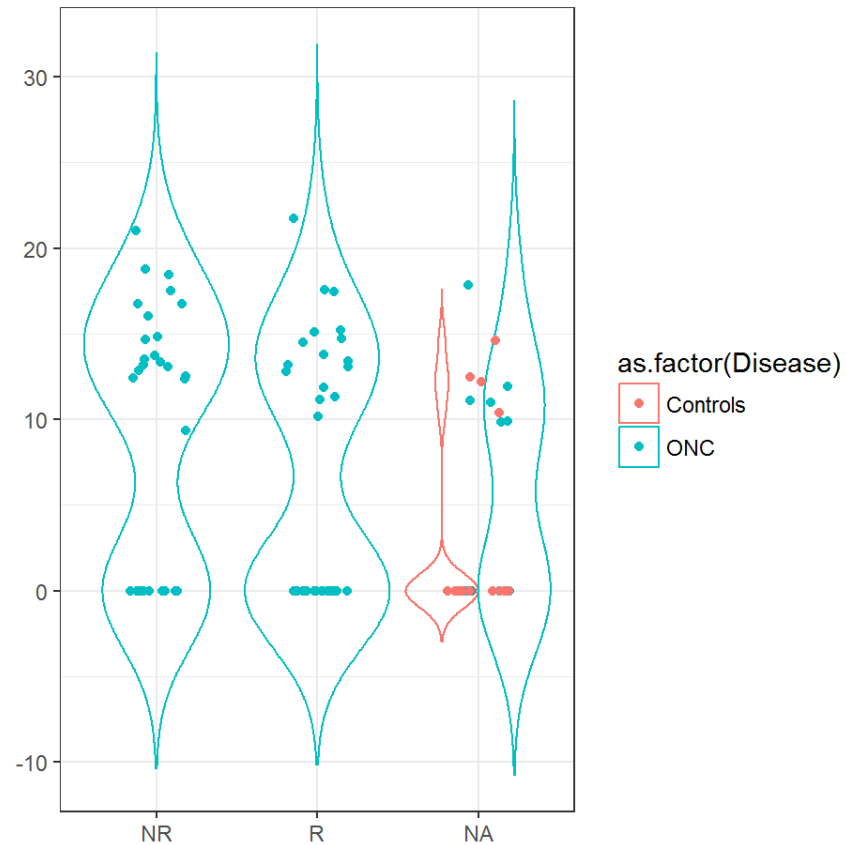
- princomp



Exploratory Data Analysis

- Response Data

Response	
NR	R
19	24



R Tips and tricks:

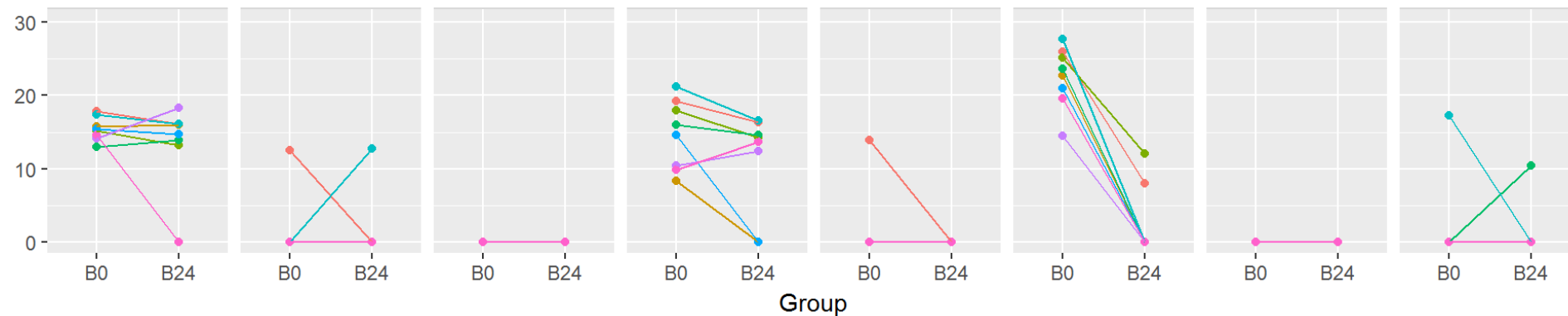
- `table`
- `geom_violin`

Exploratory Data Analysis

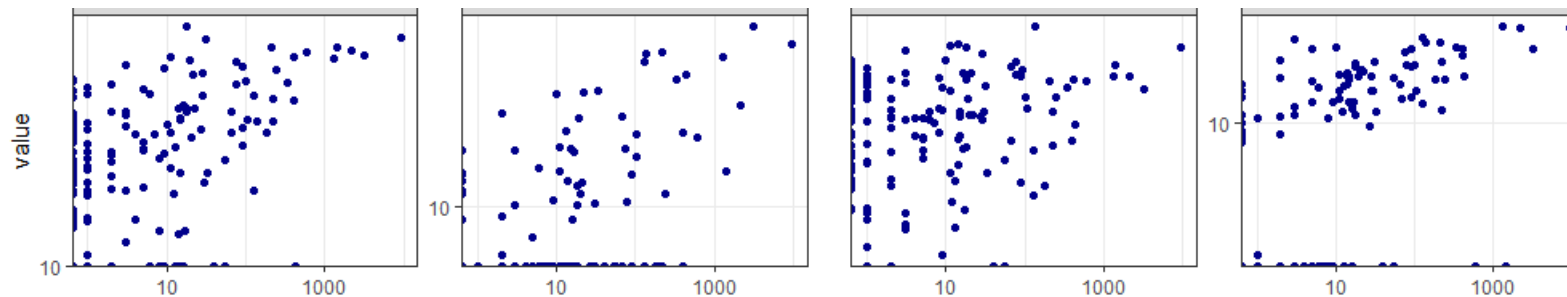
- Technical Effects
 - Delay in 24 hour sample processing

R Tips and tricks:

- `geom_point`
- `geom_line`



- Cell Counts



Analysis and Modeling

- Univariate
- Multivariate/Predictive

Analysis and Modeling

- Univariate
 - Fitting GLM (binomial/logit) for each of ~140 genes
 - $\text{Response} = \text{GeneExpr} + \text{Study} + \text{CTCs\#} + \text{GeneExpr} * \text{CTCs\#}$
 - $\text{Response} = \text{GeneExpr} + \text{Study} + \text{CTCs\#}$
 - $\text{Response} = \text{GeneExpr} + \text{CTCs\#}$
 - $\text{Response} = \text{GeneExpr}$
- Multivariate/Predictive

R Tips and tricks:

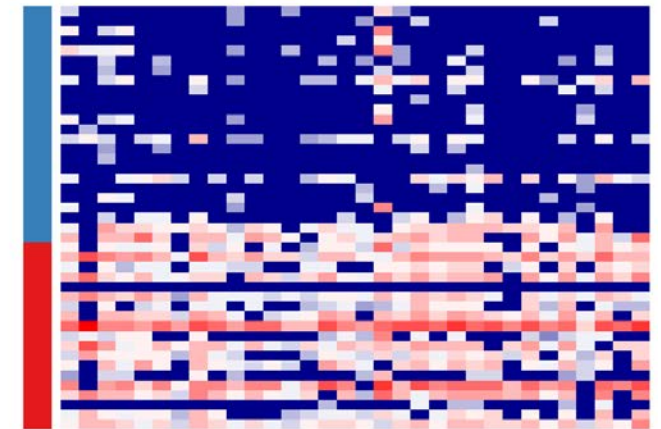
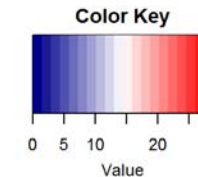
- `glm`

Analysis and Modeling

- Univariate
 - Fitting GLM (binomial/logit) for each of ~140 genes
 - $\text{Response} = \text{GeneExpr} + \text{Study} + \text{CTCs\#} + \text{GeneExpr} * \text{CTCs\#}$
 - $\text{Response} = \text{GeneExpr} + \text{Study} + \text{CTCs\#}$
 - $\text{Response} = \text{GeneExpr} + \text{CTCs\#}$
 - $\text{Response} = \text{GeneExpr}$
- Multivariate/Predictive

R Tips and tricks:

- `glm`
- `heatmap.2`



Analysis and Modeling

- Univariate
- Multivariate/Predictive

R Tips and tricks:

- caret package

Analysis and Modeling

- Univariate
- Multivariate/Predictive
 - Create data partition into test and training sets

R Tips and tricks:

- caret package

Analysis and Modeling

- Univariate
- Multivariate/Predictive
 - Create data partition into test and training sets
 - Choose methods:
 - Linear Discriminant Analysis
 - Partial Least Squares
 - Support Vector Machines
 - Neural Networks
 - Recursive Partitioning
 - Random Forests
 - Elastic Net

R Tips and tricks:

- caret package

Analysis and Modeling

- Univariate
- Multivariate/Predictive
 - Create data partition into test and training sets
 - Choose methods:
 - Linear Discriminant Analysis
 - Partial Least Squares
 - Support Vector Machines
 - Neural Networks
 - Recursive Partitioning
 - Random Forests
 - Elastic Net
 - Choose tuning parameters for methods

R Tips and tricks:

- caret package
- recipes package

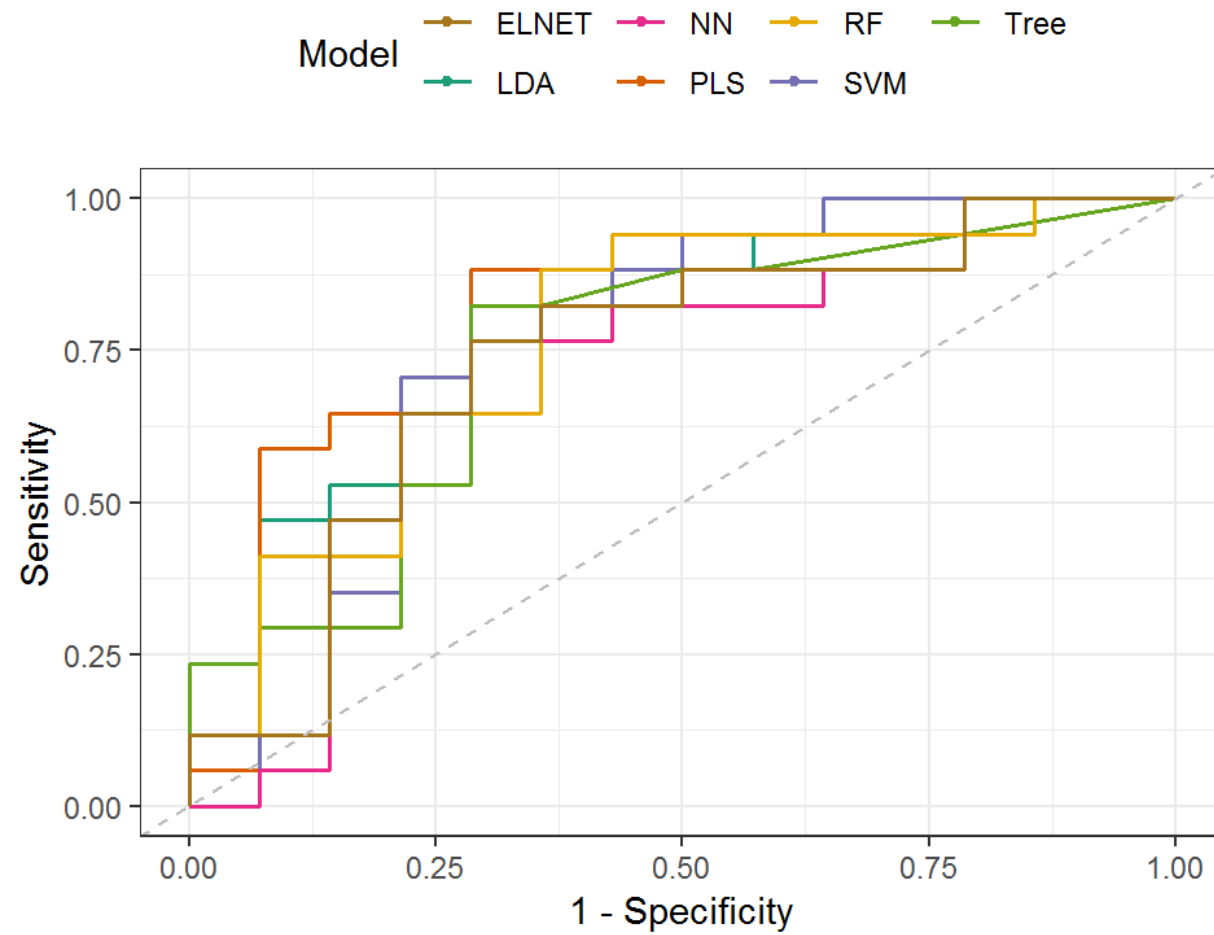
Analysis and Modeling

- Univariate
- Multivariate/Predictive
 - Create data partition into test and training sets
 - Choose methods:
 - Linear Discriminant Analysis
 - Partial Least Squares
 - Support Vector Machines
 - Neural Networks
 - Recursive Partitioning
 - Random Forests
 - Elastic Net
 - Choose tuning parameters for methods
 - Execute and evaluate

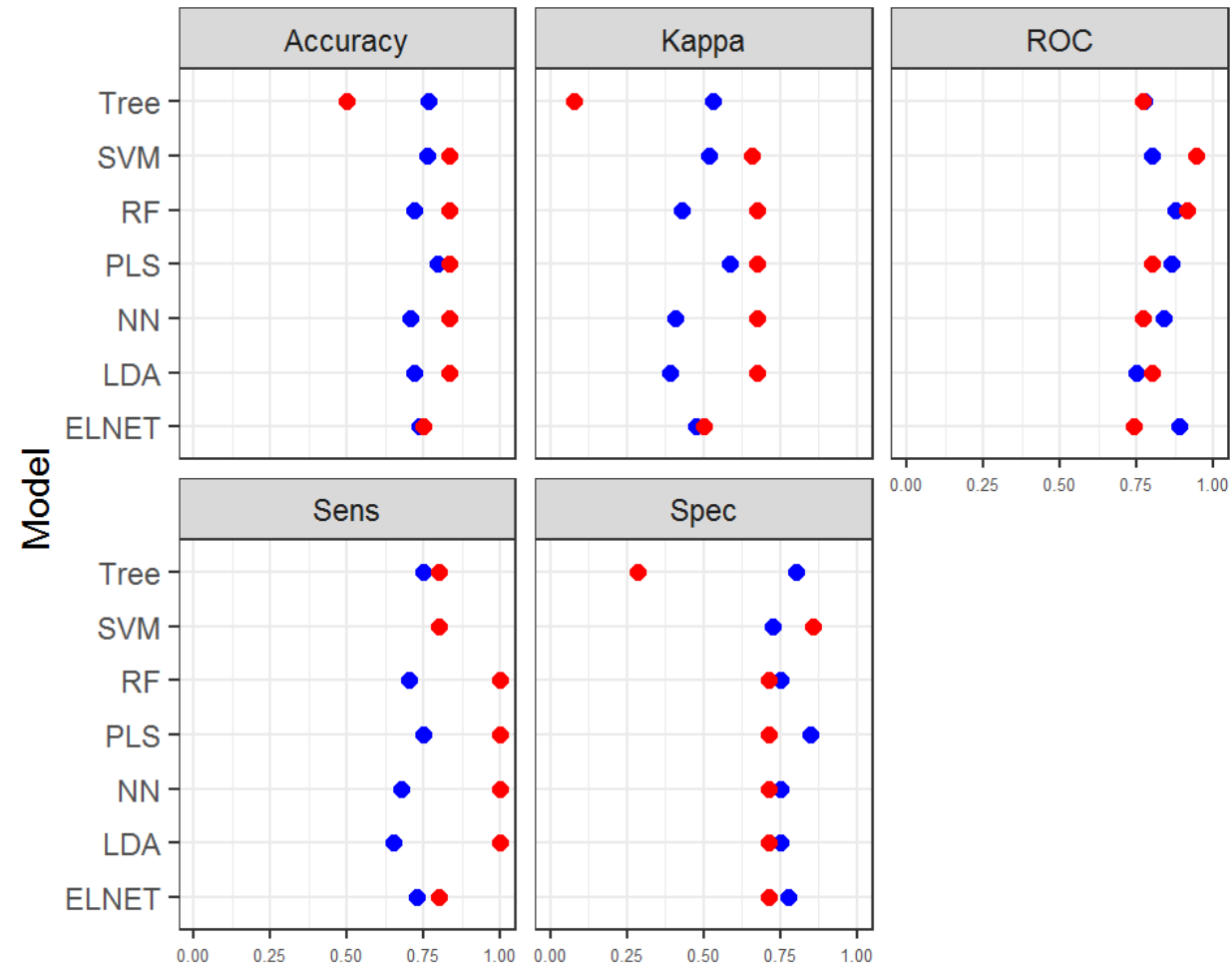
R Tips and tricks:

- caret package
- plotROC package

Results



Results



References

- R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Tukey, J. W. (1977). Exploratory data analysis. Reading, Mass: Addison-Wesley Pub. Co.
- Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). The elements of statistical learning: Data mining, inference, and prediction.
- Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. New York: Springer.

