



PhUSE US Connect 2019

Paper AB03

Integrative Analysis of Circulating Tumor Cell Counts and Gene Expression Levels

Yauheniya Cherkas, Janssen R&D, Springhouse, PA, USA

ABSTRACT

Prostate cancer is one of the most common cancers affecting males. With the existence of patients resistant to their current therapies, there is need to understand mechanisms of resistance that enable patients to progress to more aggressive forms such as mCRPC. Two cohorts of patients were collected through external collaboration and their samples were profiled for biomarkers associated with cancer. Measurements included CTC (circulating tumor cell) counts, and gene expression values in blood enriched for CTCs measured via qPCR (quantitative polymerase chain reaction). Various approaches were applied and compared to identify resistance genes, addressing the confounding between gene expression profiles and CTC counts.

INTRODUCTION

In the era of constant advancement in biomarker data generation and precision medicine, predictive and prognostic tests have become hallmarks of the new generation of cancer therapeutics. To take a precision medicine approach to the management of advanced prostate cancer, we sought to develop a peripheral blood circulating tumor cell (CTC)-based test to identify patients most likely to benefit from antiandrogen therapies. In this paper, we will describe a strategy taken to implement this approach and present relevant resources and methods.

STUDY DESCRIPTION

In this multi-institutional prospective cohort study, men with advanced prostate cancer were enrolled prior to starting antiandrogen therapies. At baseline, 12 weeks, and progression, peripheral blood samples were collected and profiled for circulating tumor cell (CTC) counts (Truini et. al., 2014) and gene expression analysis measuring levels for a subset of genes. It was shown previously (Scher et al., 2016) that CTC0 conversion (circulating tumor cell nonzero at baseline and 0 at 3 months) had the highest discriminatory power for overall survival. CTC0 conversion was used as an early readout for the response within this cohort. Baseline gene expression data from patients with CTC0 conversion were used for univariate analysis and to develop predictive models. The results were contrasted, and models were then compared using different performance characteristics.

DATA ANALYSIS

Typically, the steps for data analysis can be divided into three major categories:

1. Relevant Data Collection, Pre-processing and Cleaning
This step includes obtaining relevant data and context; cleaning noisy, missing and inconsistent data; combining the data from multiple sources, and data transformations
2. Exploratory Data Analysis (Tukey, 1977)
A method of looking at data that does not include modeling and inference (can be graphical and non-graphical) to inform detection of mistakes, checking of assumptions, preliminary selection of appropriate models and determining relationships between variables
3. Analysis and Modeling
Involves development of statistical models for the analysis of data to answer a research question

Analysis was executed using R: A Language and Environment for Statistical Computing (R Core Team, 2018)

DATA COLLECTION, PRE-PROCESSING AND CLEANING

In the current study, data was collected from three different sources:

1. clinical data collected as part of the study at the following timepoints:



2. gene expression data generated in the laboratory (Keer, 2008):

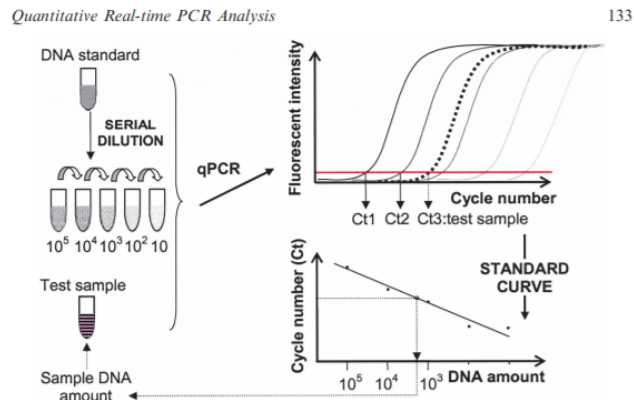
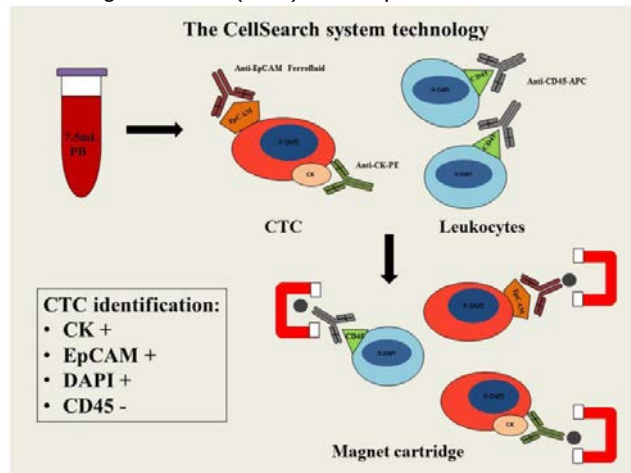


Figure 7.1 Schematic representation of the basis of real-time PCR using a quantitative standard.

3. circulating tumor cell (CTC) counts profiled via CellSearch technology (Truini et. al., 2014):



Data Collection, Pre-processing and Cleaning included:

- merging the data from all the sources
- creation of relevant variables (site, time of sample processing, CTC0 response category)
- exclusion of missing data
- exclusion of gene expression data with less than 10% of non-zero values
- more standard data cleaning and transformations

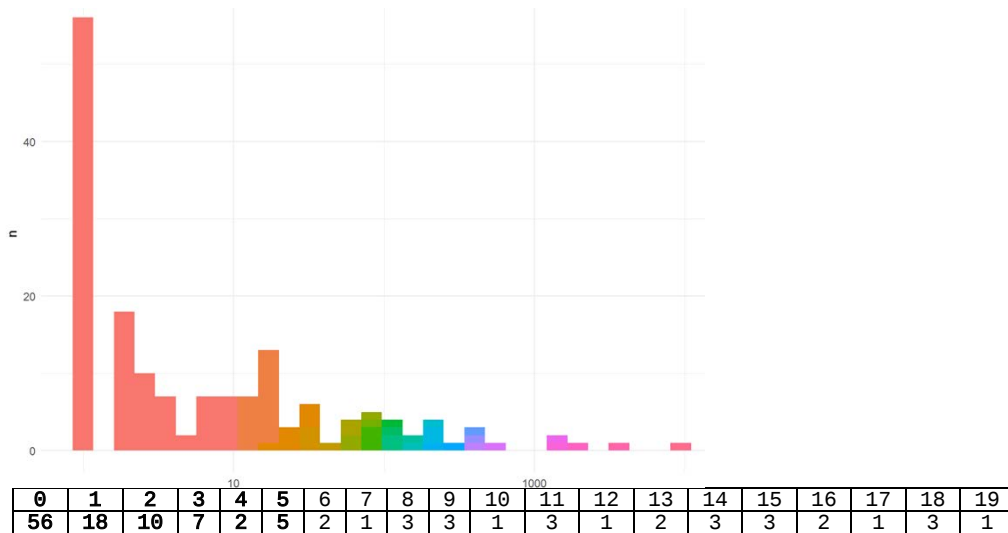
Modern set of R packages tidyverse was primarily used for data cleaning and pre-processing (Wickham, 2017).

EXPLORATORY DATA ANALYSIS

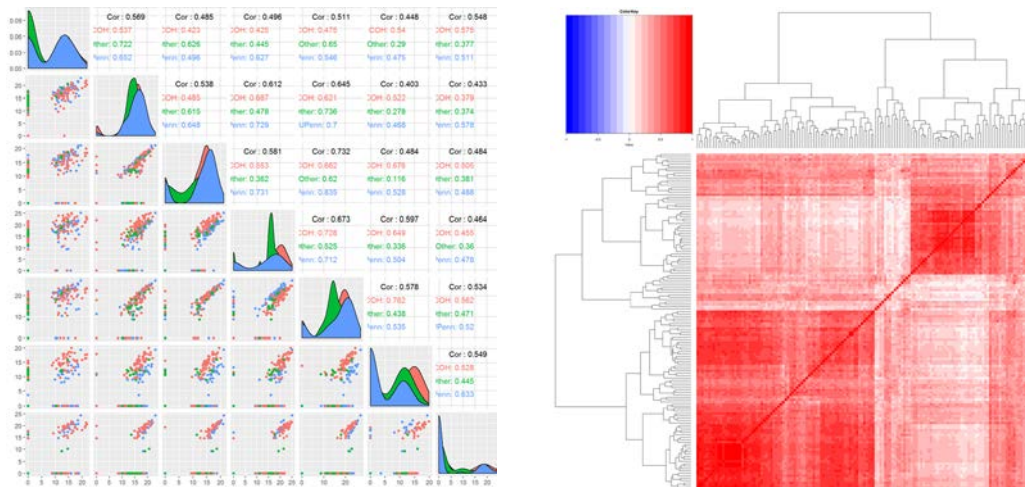
Exploratory Data Analysis (EDA) is essential to develop an understanding of data and how it is distributed; identify sources of variability and data artifacts (including technical artifacts and outliers); and address structure within the data (especially for high dimensional data).

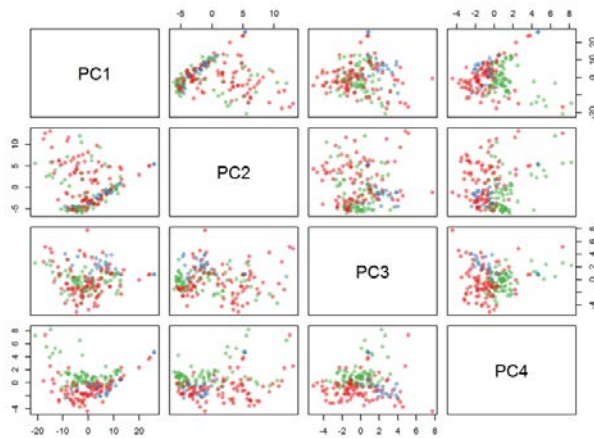
For the three sources of data the following techniques were applied:

1. CTC counts were summarized using histogram and table (R commands geom_histogram and table) and showed inflation of zero counts:



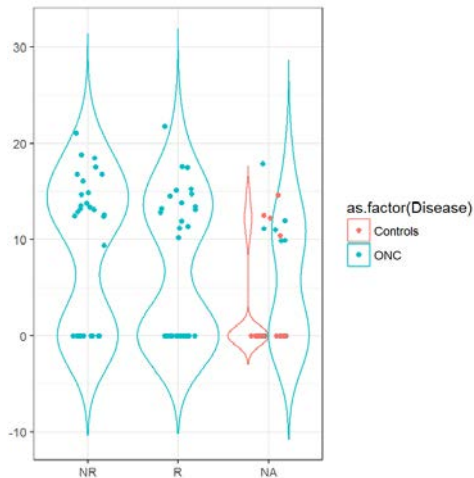
- Gene expression data included initially about 190 genes. Genes with near zero variance were filtered, and the final set included about 140 genes. Unsupervised hierarchical clustering was applied to the latter set of genes together with principal component analysis, and pairs of genes were also plotted for more in-depth look at individual genes (R commands `heatmap.2`, `princomp`, `ggpairs`). The results showed highly correlated structure present within the data, variability explained by site and response to the treatment, as well as presence of clusters of genes:



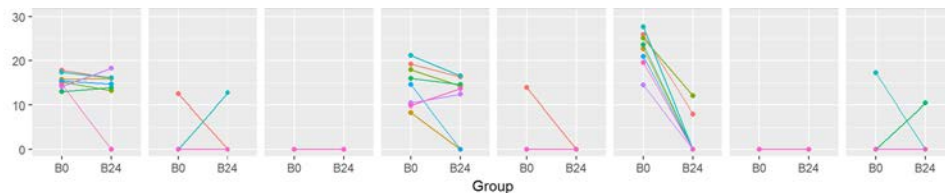


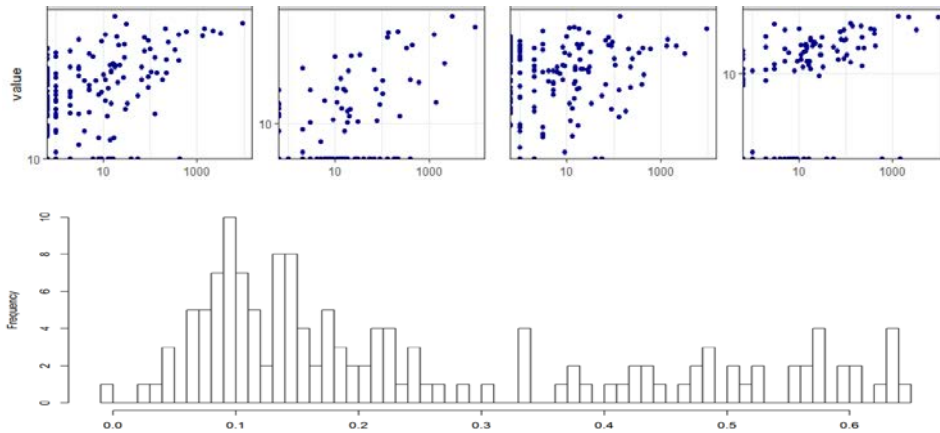
3. Clinical data exploratory analyses included summary tables and violin plots (Hintze, 1997) which are similar to boxplots with a rotated kernel density plot on each side (R commands `table`, `geom_violin`). The results showed sizes of the response groups and their relation to the gene expression data as follows:

Response	
NR	R
19	24



4. Additional evaluations included technical effect of the time of sample processing and relationships between CTC counts and gene expression data (R commands `geom_point`, `geom_line`, `hist`). The results showed differences between samples profiled at 0 and 24 hours as well as strong correlations between cell counts and some genes (histogram of Spearman correlation values):



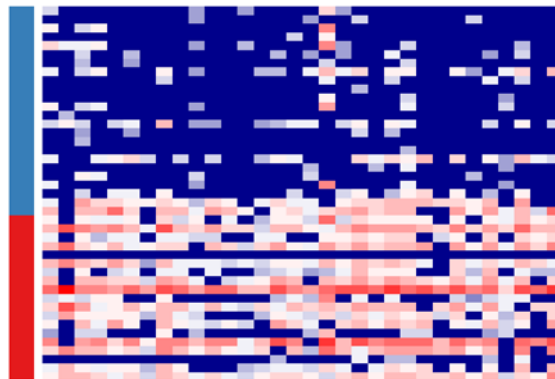


ANALYSIS AND MODELING

Equipped with the knowledge of data structure, distribution and important factors the following statistical modeling techniques were applied to the data:

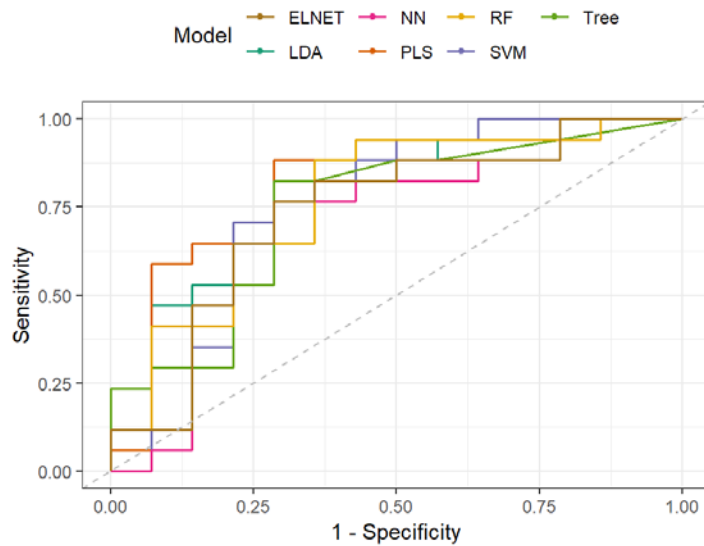
1. Univariate analysis: four generalized (binomial/logit) linear models of the response were fit for each gene
 - i. Model only including gene expression data
 - ii. Model including gene expression data and cell count
 - iii. Model including gene expression data, cell count and site
 - iv. Model including gene expression data, cell count, site and the interaction of gene expression and cell count

The results were evaluated adjusting for false discovery rate and the top set of 32 genes plotted below included genes primarily overexpressed in non-responders:

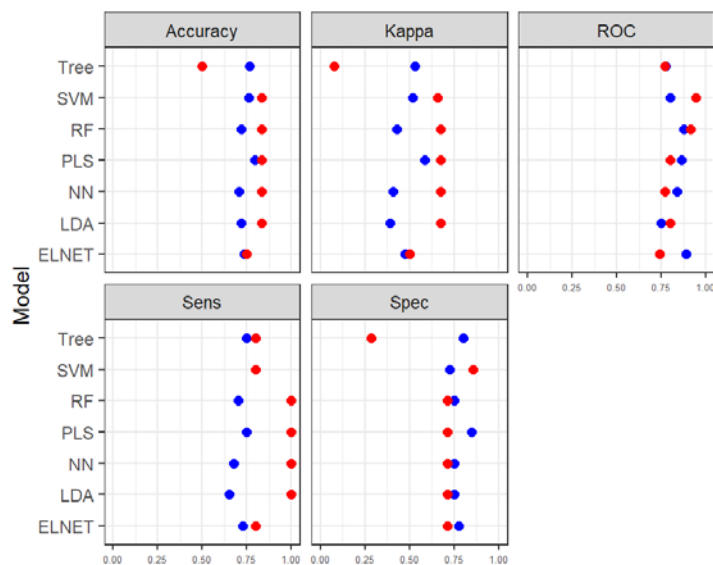


2. Multivariate predictive modeling: dataset was first randomly split into training and test sets. Baseline gene expression data for the training set was used to develop a predictive model for response, starting with a panel of 141 expressed genes/isoforms including those associated with prostate cancer. Seven different types of classification models (Hastie, et.al., 2009) were applied to the test dataset with respective tuning parameters selected in a cross-validation loop:
 - i. Linear Discriminant Analysis (LDA) (Venables et.al., 2002)
Linear Discriminant Analysis (LDA) is a linear transformation technique commonly used for dimensionality reduction. LDA computes the directions ("linear discriminants") that will represent the axes that maximize the separation between multiple classes
 - ii. Partial Least Squares (PLS) (Barker et.al., 2003)
The Partial Least Squares (PLS) method fits linear models based on linear combinations of the explanatory variables simultaneously maximizing the covariance between the variables and the response or responses
 - iii. Support Vector Machines (SVMs) (Scholkopf et.al., 2002)
Support vector machines (SVMs) are a method of classification in high-dimensional spaces that find separating hyperplanes between groups of data
 - iv. Neural Networks (Ripley, 1996)
Neural Networks represent models based on how biological neural networks in the human brain process information. It works as an adaptive system that changes its structure during a learning phase and able to model complex relationships between inputs and outputs or to find patterns in data by building layers that takes an input and computes an output for the next layer.
 - v. Recursive Partitioning (Breiman et.al., 1984)
Recursive partitioning creates a decision tree that aims to correctly classify members of the population by splitting it into sub-populations based on independent variables.
 - vi. Random Forest (Breiman et.al., 2001)
Random forests are an ensemble method that combines several individual classification trees in the following way: From the original sample several bootstrap samples are drawn, and an unpruned classification tree is fit to each bootstrap sample. Each tree in the ensemble is built based on the principle of recursive partitioning, where the feature space is recursively split into regions containing observations with similar response values. The variable selection for each split in the classification tree is conducted only from a small random subset of predictor variables, so that the overfitting is avoided. From the complete forest the status of the response variable is predicted as an average or majority vote of the predictions of all trees.
 - vii. Elastic Net (Zou et.al., 2005)
Elastic net regression is a regularized regression method that optimizes the linear combination of the objective function that measures the coherence between the model and the data and the two penalties: L1 or the sum of absolute values of all coefficients, and L2 or the sum of squares of all coefficients.

The performance of all seven models is summarized below via ROC curve:



Additionally, below are the five additional characteristics of model performance (test set value in red and cross-validation in blue): accuracy, kappa, ROC, Sensitivity and Specificity. Random forest has the highest average ROC between test and cross-validated estimates:



CONCLUSION

Baseline gene expression data from patients with CTC0 conversion were used for univariate analysis and to develop seven predictive models. The results were then compared using different performance characteristics. Of the seven predictive models tested (linear discriminant analysis, partial least squares, support vector machines, neural networks, recursive partitioning, random forests and elastic net), random forest yielded the best performance, with respective cross-validation and test set sensitivities of 0.7 and 1, specificities 0.75 and 0.71, AUC 0.88 and 0.91. Top genes identified include those previously associated with disease. A prospective study with a larger number of patients will be required to further validate our findings. Ultimately, this test has the potential to select the patients most likely to benefit from antiandrogen treatment.

REFERENCES

1. Truini A., Alama A., Dal Bello MG., Coco S., Vanni I., Rijavec E., Genova C., Barletta G., Biello F., Grossi F. (2014). Clinical Applications of Circulating Tumor Cells in Lung Cancer Patients by CellSearch System. *Frontiers in oncology*. 4. 242. 10.3389/fonc.2014.00242.
2. Scher, H.I., et al., Trial Design and Objectives for Castration-Resistant Prostate Cancer: Updated Recommendations From the Prostate Cancer Clinical Trials Working Group 3. *J Clin Oncol*, 2016. 34(12): p. 1402-18.
3. Tukey, J. W. (1977). *Exploratory data analysis*. Reading, Mass: Addison-Wesley Pub. Co.
4. 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/>.
5. Keer, J. C. (2008) Quantitative Real-time PCR Analysis, In *Essentials of Nucleic Acid Analysis: A Robust Approach* (Keer, J. T., and Birch, L., Eds.) pp 132–166, Chapter 7, RSC Publishing, Cambridge
6. Hadley Wickham (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. <https://CRAN.R-project.org/package=tidyverse>
7. Hintze J. L., Nelson R. D. (1997) Violin Plots: A Box Plot-Density Trace Synergism. *Statistical Computing and Graphic Pages* 181-184
8. Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction*.
9. Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. New York: Springer.
10. Venables, W. N. and Ripley, B. D. (2002) *Modern Applied Statistics with S*. Fourth edition. Springer
11. Barker M, Rayens W (2003). Partial Least Squares for Discrimination." *Journal of Chemometrics*, 17(3), 166–173
12. Scholkopf B. and Smola A.J. 2002. *Learning with Kernels*. MIT Press
13. Ripley, B. D. (1996) *Pattern Recognition and Neural Networks*. Cambridge
14. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees*. New York: Chapman and Hall; 1984
15. Breiman L. Random forests. *Machine Learning*. 2001a;45(1):5–32
16. Zou H, Hastie T. 2005. Regularization and variable selection via the elastic net
17. Ondra, T., Dmitrienko, A., Friede, T., Graf, A., Miller, F., Stallard, N., Posch, M. (2016). Methods for identification and confirmation of targeted subgroups in clinical trials: a systematic review. *Journal of biopharmaceutical statistics*, 26, 99-119.

TEAM ACKNOWLEDGMENTS

Michael J. LaRiviere, MD,¹ Naomi B. Haas, MD,² Yauheniya Cherkas, PhD,³ Karl Nielsen, MSc,³ Brad Foulk, MS,³ Jaymala Patel, MS,³ Denis Smirnov, PhD,³ David J. Vaughn, MD,² Ravi Amaravadi, MD,² Samantha L. Savitch, BA,² Krishna Majmundar, BS,² Thomas H. Buckingham, BA, BS,² Stephanie S. Yee, MS,² Jeremy Jones, PhD,⁴ Sumanta K. Pal, MD,⁴ Erica L. Carpenter, MBA, PhD^{2,5}

¹Department of Radiation Oncology, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania.

²Division of Hematology/Oncology, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania.

³Janssen, Pharmaceutical Companies of Johnson and Johnson, Spring House, Pennsylvania.

⁴Department of Medical Oncology and Therapeutics Research, City of Hope, Duarte, California.