PREDICTING THE EFFECTS OF WAVE GENES ON CANCER SEVERITY AND OUTCOMES

Ekaterina Cole, SUNY Oswego, Biomedical Informatics MS

Joshua Harkness, SUNY Oswego, Information Science BA

Seham Al-Masri, SUNY ESF, Biotechnology BS

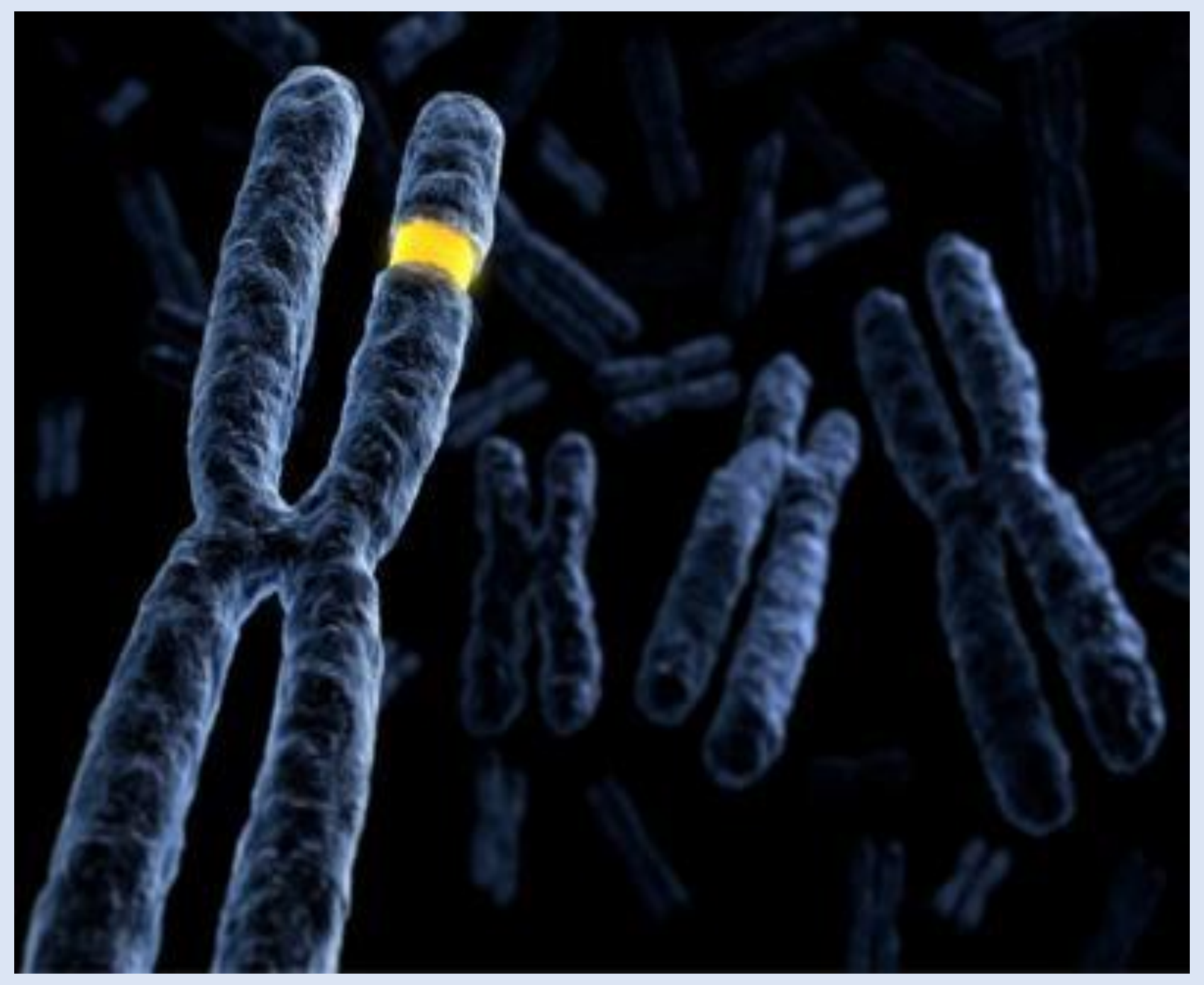
Dr. Isabelle Bichindaritz, SUNY Oswego, Director of Biomedical Informatics

Abstract:

The WAVE regulatory complex has been identified as one of the most important actors contributing to breast tumor progression. The WAVE is represented by a set of 11 genes, namely, ABI1 ABI2 ABI3 BRK1 CYFIP1 CYFIP2 NCKAP1 NCKAP1L WASF1 WASF2 WASF3, functioning in a synergistic manner. The main objective was to define the characteristics of the WAVE and determine its role in breast cancer as a phenomenon. A multi-systemic dataset, available in the public domain, was obtained, structured and analyzed. This project implied a collaboration with Dr. Kotula, an oncologist at Upstate Medical Center.

Introduction and Stats:

Breast cancer is one of the most omnipresent diseases in today's US Healthcare system. According to the American Cancer Society statistics (2017), in **2017** alone, an estimated **40,610** women will **die** from breast cancer while approximately **316,120** women will be **diagnosed**; a fourth of which will be non-invasive. In addition, breast cancer is reported to be the second leading cause of death for women behind lung cancer. The chance a woman will die from breast cancer is **1 in 37**. These data emphasize the importance of most profound understanding of factors that trigger breast cancer and contribute to its development.



Project Milestones and Methods:

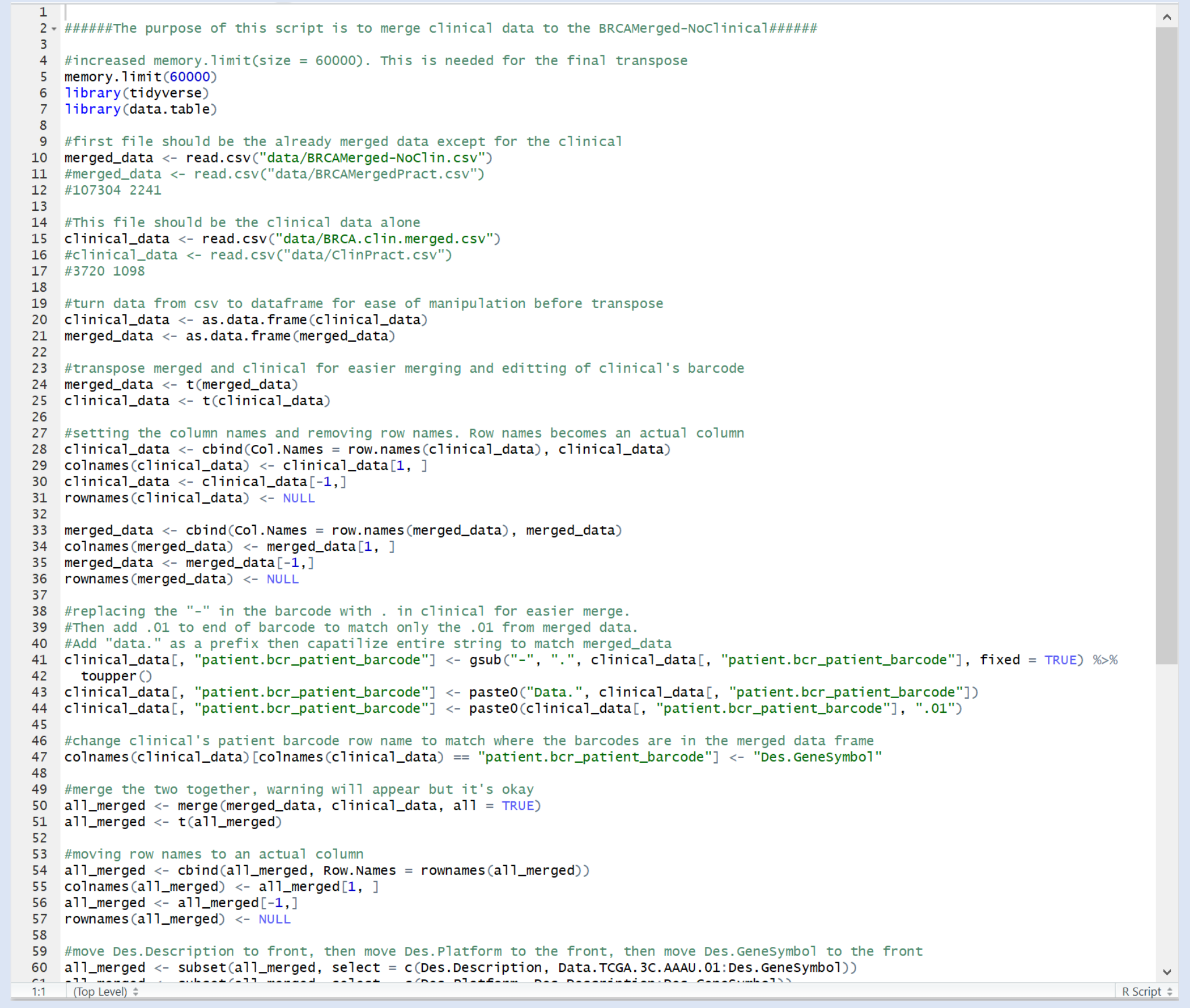
The project encompassed examining the professional domain of breast cancer and identifying the existing research. Meta-analysis of literature was expected to provide basis for data understanding.

The breast cancer data, available in public domain (TCGA), was obtained and merged with the help of a specifically designed R script. The dataset, including both clinical and genomic data, was analyzed and the full data dictionary was completed. The WAVE complex was extracted from the dataset and combined with a set of variables, descriptive of

- **Cancer stage;**
- **Occurrence of relapse;**
- **Presence of metastasis.**

The R Script:

The R script written merges clinical data with genomic data. The datasets were merged based on a shared barcode. Editions were made to the clinical dataset for the barcodes to match exactly, due to minor differences. This ensured that the proper barcodes were given additional variables from clinical, without any repeated values.



```r

######The purpose of this script is to merge clinical data to the BRCAMerged-NoClinical######

#increased memory.limit(size = 60000). This is needed for the final transpose
memory.limit(60000)
library(tidyverse)
library(data.table)

#first file should be the already merged data except for the clinical
merged_data <- read.csv("data/BRCAMerged-NoClin.csv")
#merged_data <- read.csv("data/BRCAMergedPract.csv")
#107304 2241

#This file should be the clinical data alone
clinical_data <- read.csv("data/BRCA.clin.merged.csv")
#clinical_data <- read.csv("data/ClinPract.csv")
#3720 1098

#turn data from csv to dataframe for ease of manipulation before transpose
clinical_data <- as.data.frame(clinical_data)
merged_data <- as.data.frame(merged_data)

#transpose merged and clinical for easier merging and editting of clinical's barcode
merged_data <- t(merged_data)
clinical_data <- t(clinical_data)

#setting the column names and removing row names. Row names becomes an actual column
clinical_data <- cbind(Col.Names = row.names(clinical_data), clinical_data)
colnames(clinical_data) <- clinical_data[1, ]
clinical_data <- clinical_data[-1,]
rownames(clinical_data) <- NULL

merged_data <- cbind(Col.Names = row.names(merged_data), merged_data)
colnames(merged_data) <- merged_data[1, ]
merged_data <- merged_data[-1,]
rownames(merged_data) <- NULL

#replacing the "-" in the barcode with . in clinical for easier merge.
#Then add .01 to end of barcode to match only the .01 from merged data.
#Add "data." as a prefix then capatilize entire string to match merged_data
clinical_data[, "patient.bcr_patient_barcode"] <- gsub("-", ".", clinical_data[, "patient.bcr_patient_barcode"], fixed = TRUE) %>%
  toupper()
clinical_data[, "patient.bcr_patient_barcode"] <- paste0("Data.", clinical_data[, "patient.bcr_patient_barcode"])
clinical_data[, "patient.bcr_patient_barcode"] <- paste0(clinical_data[, "patient.bcr_patient_barcode"], ".01")

#change clinical's patient barcode row name to match where the barcodes are in the merged data frame
colnames(clinical_data)[colnames(clinical_data) == "patient.bcr_patient_barcode"] <- "Des.GeneSymbol"

#merge the two together, warning will appear but it's okay
all_merged <- merge(merged_data, clinical_data, all = TRUE)
all_merged <- t(all_merged)

#moving row names to an actual column
all_merged <- cbind(all_merged, Row.Names = rownames(all_merged))
colnames(all_merged) <- all_merged[1, ]
all_merged <- all_merged[-1,]
rownames(all_merged) <- NULL

#move Des.Description to front, then move Des.Platform to the front, then move Des.GeneSymbol to the front
all_merged <- subset(all_merged, select = c(Des.Description, Data.TCGA.3C.AAAU.01:Des.GeneSymbol))
all_merged <- subset(all_merged, select = c(Des.Platform, Des.Description:Des.GeneSymbol))
```

An outer join was used to ensure no variables were lost while still merging identical barcodes together. The merge introduced an additional 3719 variables.

The WAVE characteristics were summarized using a range of statistical tools, including Excel, R, and SPSS. The associations analysis, along with dataset enrichment through cBioPortal was performed.

Project Achievements:

- A merged dataset of clinical + genes data, with a designed R script, and full data dictionary for the dataset were created;
- WAVE complex extracted and variables, definitive of main set characteristics were identified;
- The characteristics of the WAVE were summarized and the associations have been examine:

  the highest values of gene expressions are associated with the higher cancer stages (X, IIIB), and the lower values, are associated with the stages II and III.

  genes of BRK, CYFIP and NCKAP families appear to be among the most frequently expressed genes for cases with both relapse occurrence and higher stages;

  significant differences in WAVE gene expression levels have been observed in connection with aggregated patients' replapse status;

  cBioPortal Enrichment analysis has shown that Hormonal genes of BRCA and HER families have been found to be correlated with the WAVE sequence in terms of genetic alterations.

Name of the Dataset Feature	Characteristics
Size of the Overall Dataset Analyzed (Clinical+Genes)	111025 rows x 2242 columns (1.5GB)
Complete Number of Genes	21661 gene (5 metrics per each)
WAVE genes	11 genes extracted
Number of patients	2238 patients with unique TCGA barcodes
Main Patients' Characteristics	Patient Barcode, Vital Status, New tumor event; Days to death, Tumor status; Cancer Stage; Drug/ therapy info; follow-ups.
Characteristics, Vital to Data Analysis	Patient's vital status; cancer stage; relapse status; metastasis status; patient's new tumor events and follow-up condition statuses.
Origin of the Data	NIH The Cancer Genome Atlas (TCGA)