Tumor cell and microenvironment transcriptome profiling in breast cancer using Single Cell RNA-seq

Akanksha Puri (BIOL), Rajveer Singh(Data), Claudia Rzucidlo(Stats), Kerwin George(Writer)

4/24/2021

Table of Contents

Background and Questions

Cancers usually have intratumoral heterogeneity with therapeutic consequences from molecular-targeted treatments used for breast cancers. The tumor microenvironment is formed by a mixture of non-carcinoma cells that infiltrate the tumor cells. Interaction of such cells with each other and with tumor cells influences tumor development and progression. Single-cell genome analysis allows for monitoring and characterization of such a microenvironment to look for transcriptome heterogeneity and heterogeneous tumor signatures such as immune cells infiltrating the tumor cells that have pleiotropic roles. This research characterizes transcriptomes of 246 single cells obtained from four breast cancer patients, and each of the patients is classified among four known subtypes of breast cancer. Through single-cell RNA-sequencing data, the single cells were divided into tumor and non tumor cells based on their genomic copy number alterations. Also, depicts that each subtype of breast cancer has distinguishing features, which include the proportion of microenvironmental non tumor cells, cellular heterogeneity of tumor cells, and gene expression signatures. The test performed here, were one-way ANOVA and chi square. One-way analysis can be defined as a statistical approach to determining any substantial differences between two or more samples. The Chi-Square Test is used to test if two categorical variables are associated.One-way ANOVA, after normally distributing our data, is to analyze whether the copy numbers vary significantly between the different tumor types. The questions based on these are; first-Is there a significant variation in the number of copies of genes in the subtypes of tumor? Second-What are the characteristics of different tumor subtypes that are shaped by tumor cells and immune cells in the microenvironment? Third- What are the heterogeneous characteristics of breast cancer transcriptomes in stromal/immune cells? Fourth- Whether the copy numbers vary significantly between the different tumor types?

Statistical Analysis

Null Hypothesis: There is no significant variation in the number of copy of genes in the subtypes.

Samples, NGS and Statistical methods

The experiment was performed with a sample size of **515 cells from 11 patients**. Four subtypes & markers: ER-positive (BC01 and BC02; luminal A), ER/HER2-positive (BC03; luminal B), HER2-positive (BC04, BC05 and BC06; HER2) and triple-negative (BC07–BC11; TNBC). NGS platforms: **Illumina HiSeq Rapid SBS & Nextera XT**. Computational & Statistical analysis: **Pearson's Correlation Analysis**, Gene Set Variation Analysis (GSVA), Over-Representation Analysis using Hypergeometric Test, Likelihood Ratio Test (LRT) and the Receiver-Operating Characteristic Test using R Package Seurat, TNBCtype program for subtypes of triple negative breast cancer cells Sequencing libraries constructed with Nextera XT and sequenced using the HiSeq2500 system.

Results

```
library(knitr) #load knitr
library(tidyverse) # load the tidyverse package
library(ggthemes) # load the package to get gg themes
library(moments) # load the package to use the skewness() function to find
skewness coefficient
library(LambertW) # load the package that allows us to use the Gaussianize()
function to automatically make the skewed, heavy-tailed data into a normal
distribution
library(ggridges) # load the package in order to order to make ridgeline
plots
library(rcompanion) # load the package
library(ggsci) # load the package for themes of ggplot

library(ggpubr) # load the package for more themes of ggplot

library(viridis)# load the package for viridis colors to fill

setwd("C:\\Users\\Kerpeck\\Desktop") # set the working directory
opts_chunk$set(echo = TRUE, message= FALSE, warning = FALSE)

copynumvar <- read_csv("CopyNumVar2.csv") # read the data into r and save it
to variable
```

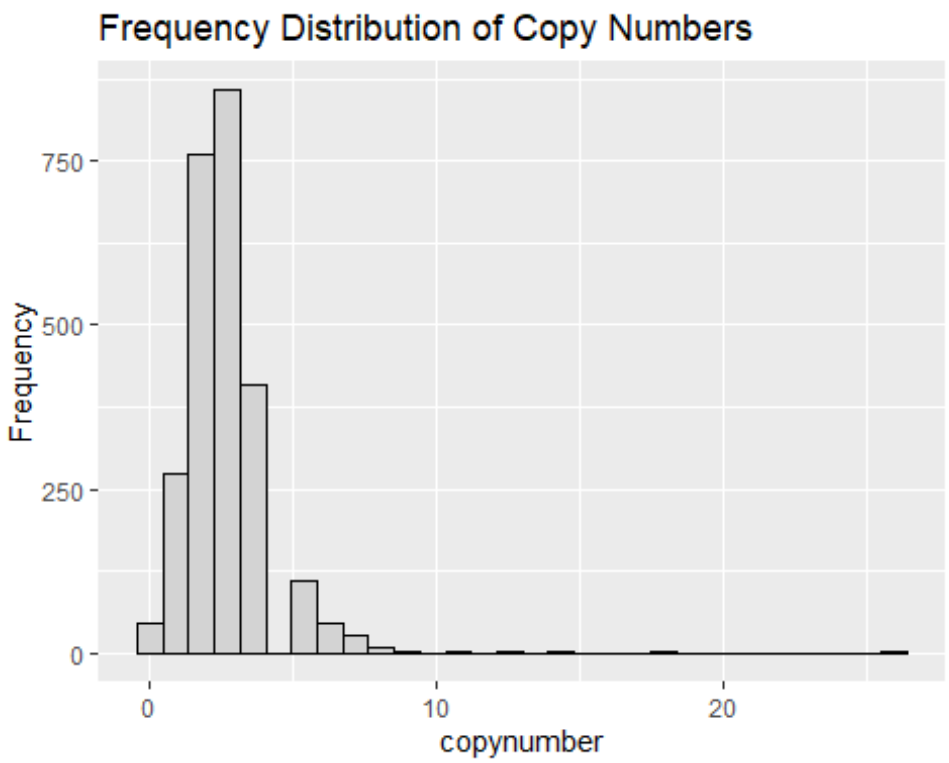
Table

```
head(copynumvar) # view the first parts of data

## # A tibble: 6 x 9
##   Tumor.Type Patient.tissue Chrom Position genes copynumber status
Mutation
##   <chr>      <chr>          <chr> <chr>    <chr>      <dbl> <chr>  <chr>
## 1 luminal A  BC01           chr1  4547400~ AKR1~          2 normal somatic
## 2 luminal A  BC01           chr1  8928725~ CCBL~          2 normal somatic
## 3 luminal A  BC01           chr1  1450757~ NBPF~          2 normal somatic
## 4 luminal A  BC01           chr1  2058632~ LOC2~          4 gain   somatic
## 5 luminal A  BC01           chr2  1172550~ GREB1          4 gain   somatic
## 6 luminal A  BC01           chr2  5928800~ LINC~          0 loss   somatic
## # ... with 1 more variable: WilcoxonRankSumTestPvalue <dbl>

# compute the skewness coefficient before transforming into normal
distribution (expect a positive value greater than 0)
skewness(copynumvar$copynumber)

## [1] 5.508167

summary(copynumvar) # provide general statistics for the data
```

```
##   Tumor.Type        Patient.tissue        Chrom              Position
##  Length:2557        Length:2557        Length:2557        Length:2557
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##     genes              copynumber        status            Mutation
##  Length:2557        Min.   : 0.000   Length:2557        Length:2557
##  Class :character   1st Qu.: 2.000   Class :character   Class :character
##  Mode  :character   Median : 3.000   Mode  :character   Mode  :character
##                     Mean   : 2.903
##                     3rd Qu.: 3.000
##                     Max.   :26.000
##  WilcoxonRankSumTestPvalue
##  Min.   :0.000e+00
##  1st Qu.:0.000e+00
##  Median :0.000e+00
##  Mean   :3.001e-03
##  3rd Qu.:6.960e-06
##  Max.   :4.980e-02
```

Histogram Before Data Transformation

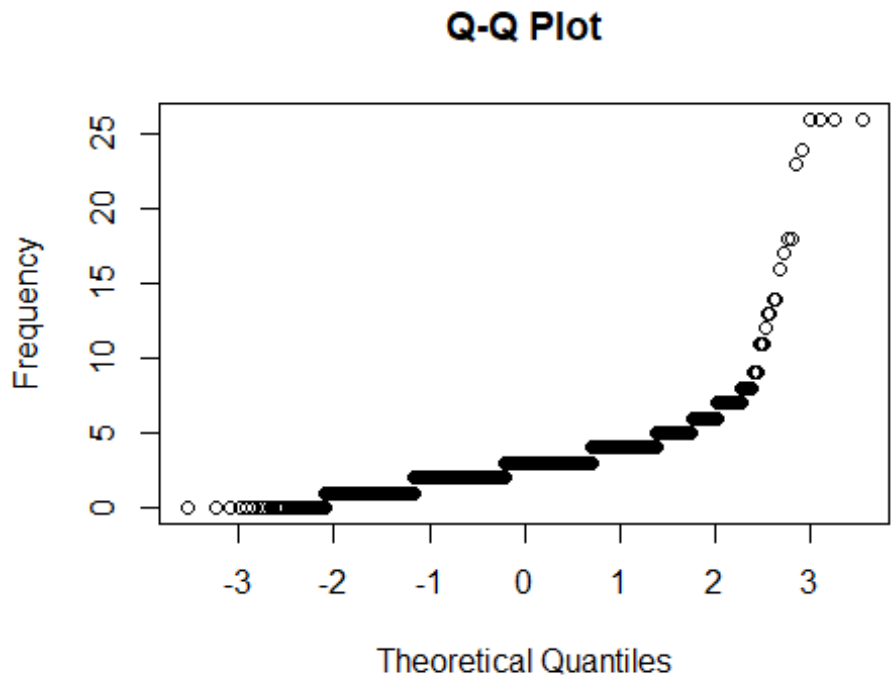
```
# Visualize distribution with a histogram before it is transformed into
normal
ggplot(copynumvar, aes(x=copynumber)) + geom_histogram(fill="light grey",
color="black") + labs(title="Frequency Distribution of Copy Numbers", y =
"Frequency")
```



Histogram displays frequency distribution of how often every copy number in the data occurs. The x-axis indicates copy number, a numerical variable, and the y-axis indicates the number of times that the values in x-axis occur within the intervals determined by values in x-axis. The longer tail on the right side of the distribution indicates an asymmetrical distribution and a positive skew (5.508167 is the skewness constant) in which the mean is greater than the median.

Quantile-Quantile Before Data Transformation

```
# Visualize with Q-Q Plot that deviates from a line as much to show it is not
normally distributed.
qqnorm(copynumvar$copynumber, ylab = "Frequency", main = "Q-Q Plot")
```



Quantile-Quantile (Q-Q) plot shows that the data of copy number doesn't approximate a straight line and thus doesn't follow a theoretical distribution such as a normal distribution. This particular plot reveals discrepancies in the upper and lower tail of the copy number distribution as seen by a wide curve towards higher quartiles. The x-axis represents the theoretical quantiles values that would contain a probability distribution into equal intervals, and the y axis represents the frequency of the copy number

Data Transformation

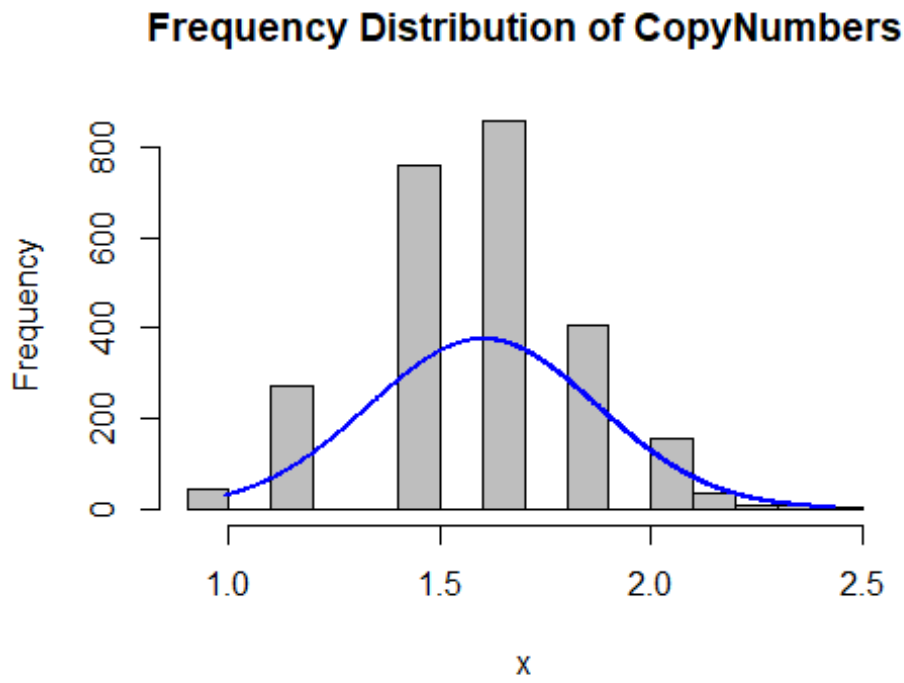
```
copynumvar$copynumber <- Gaussianize(copynumvar$copynumber) # transform the
data into normal distribution using Gaussian function


copynumvar$copynumber <- transformTukey(copynumvar$copynumber, plotit =
FALSE) # Further transform the data into normal distribution using
transformTukey function


head(copynumvar) # view the first parts of the transformed normal data
```

```
## # A tibble: 6 x 9
##   Tumor.Type Patient.tissue Chrom Position genes copynumber[,"Y1~ status
##   <chr>      <chr>          <chr> <chr>    <chr>            <dbl> <chr>
## 1 luminal A  BC01           chr1  4547400~ AKR1~            1.41  normal
```

```
## 2 luminal A  BC01           chr1  8928725~ CCBL~            1.41  normal
## 3 luminal A  BC01           chr1  1450757~ NBPF~            1.41  normal
## 4 luminal A  BC01           chr1  2058632~ LOC2~            1.89  gain
## 5 luminal A  BC01           chr2  1172550~ GREB1            1.89  gain
## 6 luminal A  BC01           chr2  5928800~ LINC~            0.990 loss
## # ... with 2 more variables: Mutation <chr>, WilcoxonRankSumTestPvalue
<dbl>

# check the skewness coefficient to see if the normal distribution worked
(expect value close to 0)
skewness(copynumvar$copynumber)

## [1] 0.01437938

summary(copynumvar) # provide general statistics for the data

##   Tumor.Type        Patient.tissue        Chrom             Position
##  Length:2557        Length:2557        Length:2557        Length:2557
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##     genes             copynumber.Y1.X      status            Mutation
##  Length:2557        Min.   :0.9896153   Length:2557        Length:2557
##  Class :character   1st Qu.:1.4076135   Class :character   Class
:character
##  Mode  :character   Median :1.6847926   Mode  :character   Mode
:character
##                     Mean   :1.6020404
##                     3rd Qu.:1.6847926
##                     Max.   :2.4346812
##  WilcoxonRankSumTestPvalue
##  Min.   :0.000e+00
##  1st Qu.:0.000e+00
##  Median :0.000e+00
##  Mean   :3.001e-03
##  3rd Qu.:6.960e-06
##  Max.   :4.980e-02
```

Histogram After Data Transformation

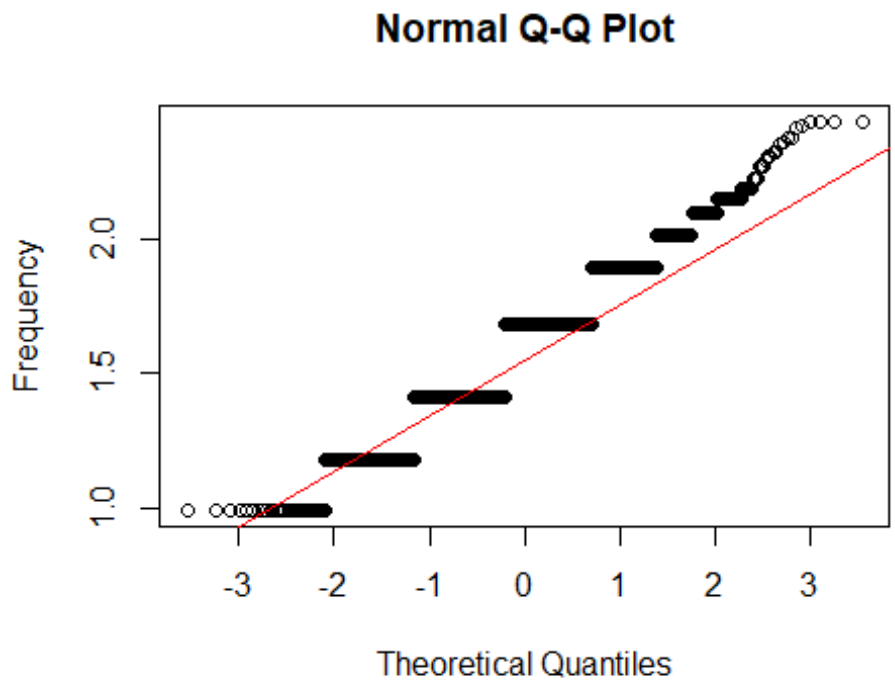
```
# Visualize the distribution with a histogram and a curve overlaying it after the data has been transformed to normal.
plotNormalHistogram(copynumvar$copynumber, main = "Frequency Distribution of CopyNumbers")
```



Histogram displays frequency distribution of how often every copy number in the data occurs. The x-axis indicates normalized values for the copy numbers and the y-axis indicates the number of times that these values in x-axis occur within the intervals determined by values in x-axis. The near symmetrical shape of the histogram with a near zero (0.0144 is the skewness constant) skew indicates an almost normal distribution, where the mean is almost equal to the median. Complete transformation to normal distribution is not achieved. The gaps in between bars indicates gaps in data where the data is not continuous.

Quantile-Quantile After Data Transformation

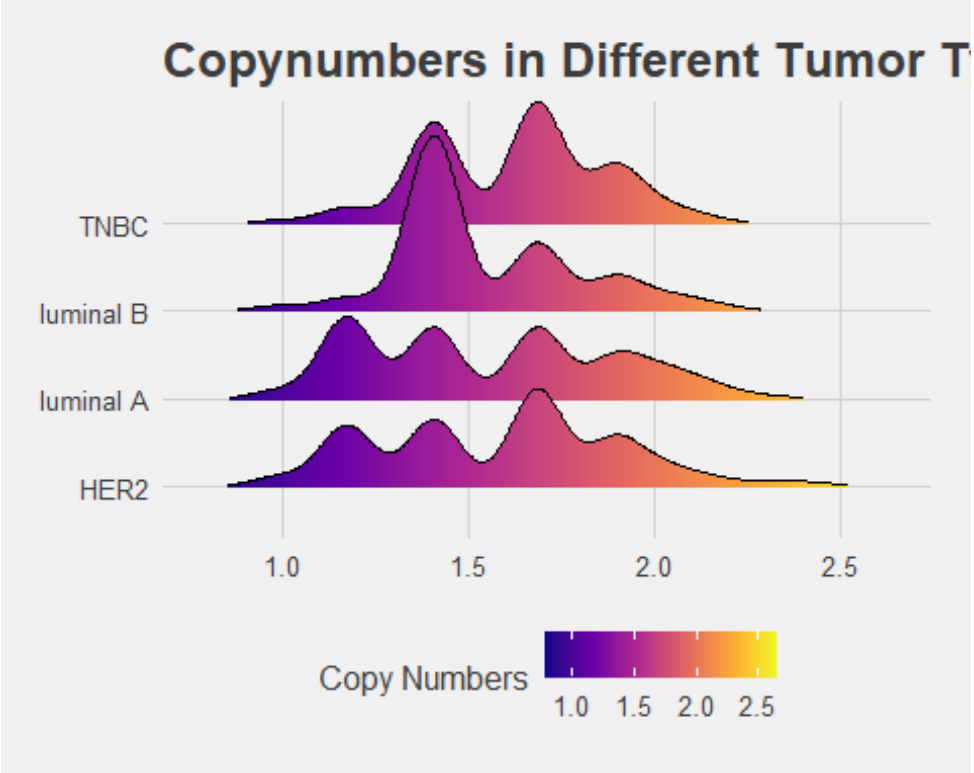
```
# On a Q-Q plot normally distributed data would appear as roughly a straight line
qqnorm(copynumvar$copynumber, ylab = "Frequency", main = "Normal Q-Q Plot")
qqline(copynumvar$copynumber, col = "red", lwd=1) # add a line to the normal quantile-quantile plot
```



A normal quantile-quantile (q-q) plot shows data of copy number on an approximate line, indicating the data now follows an approximate normal distribution. The x-axis represents the theoretical quantiles values that would contain a probability distribution into equal intervals, and the y axis represents the frequency of the copy number that has been normalized by gaussian and transformTukey. The upper tail in this qq plot is not conspicuously longer than the bottom tail in the fashion that it was in the qq plot of the positively skewed data before it was transformed to normal.

Ridgeline

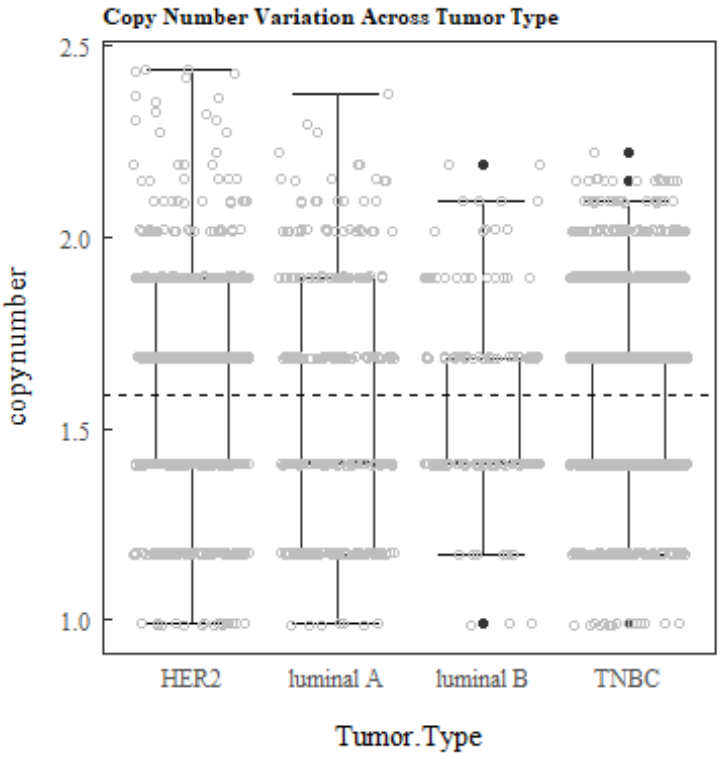
```
# make a ridgeline plot visualizing changes density distribution of over
increasing copy numbers
ggplot(copynumvar, aes(x = copynumber, y = Tumor.Type, fill = stat(x))) +
geom_density_ridges_gradient(scale = 2, rel_min_height = 0.01) +
scale_fill_viridis_c(name = "Copy Numbers", option = "C") +
labs(title = 'Copynumbers in Different Tumor Types') +
theme_fivethirtyeight()
```



Ridgeline plots display densities for each tumor type distribution of the numeric variable copy number using a density function, which tells us the probability (relative likelihood) of obtaining a specific copy number in the range of values. The x-axis shows the copy number values in our data with the scale and range determined by the normalization of the copy number variable; the y-axis shows the different tumor types. The relative color spectrum is set to the numeric variable, as shown by the legend, and filled in for each density plot. As shown by each density plot, there are lighter colors near the higher range of copy numbers closer to 2-2.5, the interval which sees a decrease in density for each tumor type and in which the probability of finding those copy numbers is relatively low.

Box Plot

```r
# manually modifiy the theme of the plot
theme_USGS_box <- function(base_family = "serif", ...){
theme_bw(base_family = base_family, ...) +
theme(
panel.grid = element_blank(),
plot.title = element_text(size = 8, face = "bold"),
axis.ticks.length = unit(-0.05, "in"),
axis.text.y = element_text(margin=unit(c(0.3,0.3,0.3,0.3), "cm")),
axis.text.x = element_text(margin=unit(c(0.3,0.3,0.3,0.3), "cm")),
axis.ticks.x = element_blank(),
aspect.ratio = 1,
legend.background = element_rect(color = "black", fill = "white")
)
}

# make a box plot to compare the distribution of a numerical variable between
several groups of a categorical variable. In this case, we want to visualize
the distribution of copy numbers between the different types of breast
tumors.
ggplot(copynumvar, aes(x=Tumor.Type, y=copynumber)) + stat_boxplot(geom
='errorbar',
width = 0.6) + geom_boxplot(width=0.5) + theme_USGS_box() +
geom_jitter(shape=1,
color="grey") + geom_hline(linetype = "dashed", yintercept = 1.58532) +
labs(title =
"Copy Number Variation Across Tumor Type")
```



Box plots represent distribution and the spread of the numerical variable copy numbers around the mean between the different tumor types HER2, luminal A, luminal B, and TNBC. The points on the plot represent each observation and the jitter on the point adds a small amount of random variation to the location of each point. The x-axis assigns each boxplot for the independent variable of tumor type and the y-axis represents the dependent variable copy number which has been transformed to normal with the gaussian function and transformTukey function. The dotted horizontal line crossing each boxplot represents the overall mean of the copy numbers. The boxes represent the interquartile range. For luminal B and TNBC tumors, the first quartile and third quartile, respectively, are equal to the median value making the median line appear on the edge of the boxes.

```
# run the one-way anova test
copynum4 <- lm(data = copynumvar, copynumber ~ Tumor.Type)


# The ANOVA test uses the linear model function lm() which takes in data
argument and a formula with relationship detonated using ~ (specifies between
the explanatory categorical variable of parasite, and the numerical variable
growth.rate). To make inferences from this linear , we use anova() and
summary() functions.


anova(copynum4) # produce a table of analysis of deviance to test whether
there is significance and whether the null hypothesis is correct. Provides us
```

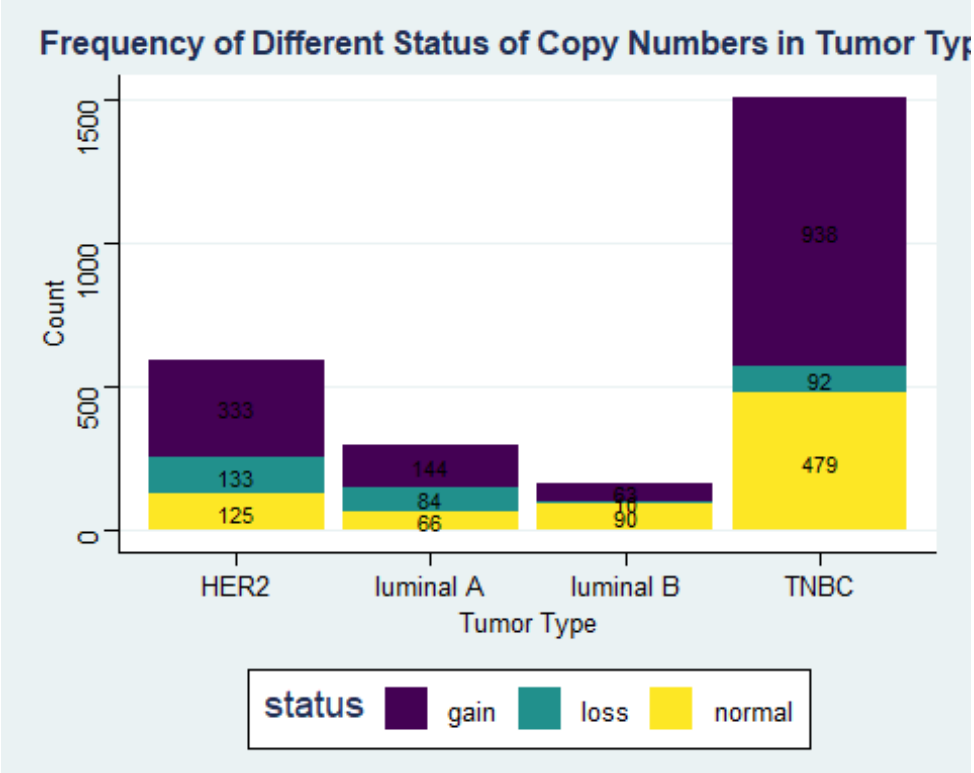
```
with the p-value.


## Analysis of Variance Table
##
## Response: copynumber
##              Df  Sum Sq Mean Sq F value   Pr(>F)
## Tumor.Type    3   2.137 0.71235  9.7451 2.16e-06 ***
## Residuals  2553 186.618 0.07310
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(copynum4) # In the estimate column there are numbers next to each
tumor type which are contrasts, differences between copy numbers of that
tumor type and the selected tumor type (HER2). As indicated by the negative
numbers, there is decreased mean copy number from HER2 to the luminal A and
luminal B tumors with luminal b having the greatest negative contrast.Tumor
TNBC has a positive contrast to HER2 so it has increased mean copy number. We
can get the means of each tumor type if we do the following by subtracting
the HER2 mean which is 1.58532 separately by each negative or positive number
in the estimate column next to each tumor



##
## Call:
## lm(formula = copynumber ~ Tumor.Type, data = copynumvar)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.63463 -0.21663  0.06055  0.13964  0.84936
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.58532    0.01112 142.547  < 2e-16 ***
## Tumor.Typeluminal A -0.03209    0.01930  -1.663  0.09646 .
## Tumor.Typeluminal B -0.04017    0.02392  -1.679  0.09320 .
## Tumor.TypeTNBC       0.03892    0.01312   2.967  0.00304 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2704 on 2553 degrees of freedom
## Multiple R-squared:  0.01132,    Adjusted R-squared:  0.01016
## F-statistic: 9.745 on 3 and 2553 DF,  p-value: 2.16e-06
```

Stacked Bar Graph

```
# To single out the groups and then get the counts for each status in each
tumor type
copynumvarChi <- copynumvar %>% group_by(Tumor.Type, status) %>% count()
copynumvarChi # give an output of the grouped table with counts
```

```
## # A tibble: 12 x 3
## # Groups:   Tumor.Type, status [12]
##    Tumor.Type status     n
##    <chr>      <chr>  <int>
##  1 HER2       gain     333
##  2 HER2       loss     133
##  3 HER2       normal   125
##  4 luminal A  gain     144
##  5 luminal A  loss      84
##  6 luminal A  normal    66
##  7 luminal B  gain      63
##  8 luminal B  loss      10
##  9 luminal B  normal    90
## 10 TNBC       gain     938
## 11 TNBC       loss      92
## 12 TNBC       normal   479
```

```
#To visualize data in graphical form, use a bar graph to depict the
counts/frequencies of the two categorical variables. The bars are colored
with the different statuses Note that stat = 'identity'
#in geom_bar() allows for the heights of the bars to represent values in the
data for the y variable. The position = 'dodge' allows us to dodge overlap by
adjusting position side by side.

ggplot(copynumvarChi, aes(x = Tumor.Type, y = n, fill = status)) +
geom_bar(stat =
'identity', position = 'stack') + ylab("count") + ylab("Count") + xlab("Tumor
Type") +
ggtitle("Frequency of Different Status of Copy Numbers in Tumor Types") +
geom_text(aes(label = n), size = 3, position = position_stack(vjust = 0.5)) +
theme_stata() +
theme(plot.title = element_text(size=12, face="bold")) +
scale_fill_viridis(discrete = T)
```



Bar plot shows the relationship between a numeric variable of counts (n) for the status and the two categorical variables, tumor type and status. Each bar of the tumor type is stacked into subdivided bars that are filled in with different colors representing different status. Together, this visualizes the association between the two categorical variables, tumor type and status. The x-axis represents the tumor type categorical variable, and the y axis represents the numerical variable of counts. In general, gain status has the most counts, while loss status has the least for each tumor type. The tumor type TNBC has the most counts of gain (n=938) but also most counts of normal as well (n=479). Luminal B has the least counts of both gain (n=63) and loss status (n=10). These trends mainly have to do with the differences in mutation rate and mutation occurrences that cause differences in number of observations of copy number variation for each tumor type.

```
copynumvar.tab <- xtabs(n ~ status + Tumor.Type, data = copynumvarChi)
#To make a matrix (contingency table) from a data frame, use the function
xtabs() which takes in a formula and the data frame as arguments. In this
case this function takes in the formula, n ~ status + Tumor.type, that
cross-tabulates the counts by the status and Tumor.Type variables.


copynumvar.tab # output the contingency table

##          Tumor.Type
## status    HER2 luminal A luminal B TNBC
##   gain     333       144        63  938
```

```
##   loss    133        84        10   92
##   normal  125        66        90  479
```

```
copynumvarchi <- chisq.test(copynumvar.tab) # Perform the Chi Square test with the function chiseq.test() which takes in the matrix. Save results to a variable
copynumvarchi
```

```
##
##  Pearson's Chi-squared test
##
## data:  copynumvar.tab
## X-squared = 240.2, df = 6, p-value < 2.2e-16
```

```
#We can single out names of objects in the chi square test and place it after dollar to access that.
```

```
copynumvarchi$expected # This is the expected counts
```

```
##         Tumor.Type
## status        HER2 luminal A luminal B     TNBC
##   gain   341.61048 169.93821  94.21744 872.2339
##   loss    73.73054  36.67814  20.33516 188.2562
##   normal 175.65898  87.38365  48.44740 448.5100
```

```
copynumvarchi$observed #This is the observed counts
```

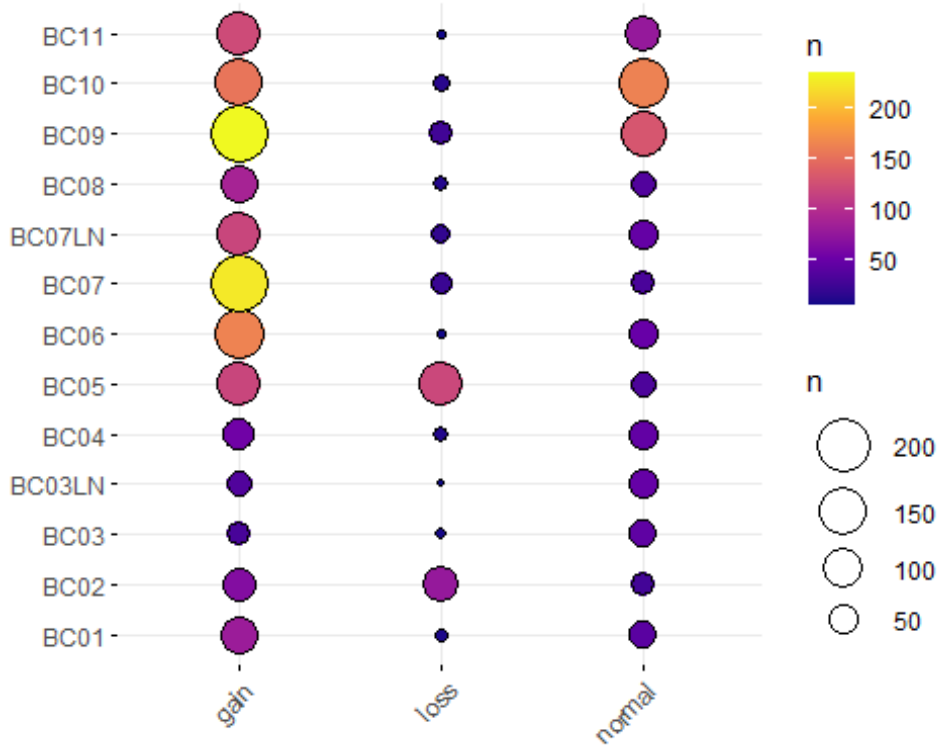
```
##         Tumor.Type
## status   HER2 luminal A luminal B TNBC
##   gain    333       144        63  938
##   loss    133        84        10   92
##   normal  125        66        90  479
```

```
copynumvarchi$p.value #This gives us the p-value.
```

```
## [1] 5.086606e-49
```

Balloon Plot

```
copynumvarChi2 <- copynumvar %>% group_by(Patient.tissue, status) %>% count()
# To single out the groups and then get the counts for each status in each
Patient tissue

# Make a balloon plot to visualize association between status and patient
tissue relative to count (n)

ggballoonplot(copynumvarChi2, x = "status" , y = "Patient.tissue", size =
"n",
fill = "n") +
scale_fill_viridis_c(option = "C")
```



Balloon plot visualizes multivariate categorical data, of which patient tissue and status type being the categorical variables, with balloons as graphical form whose sizes and colors correspond to the counts. The x-axis represents one categorical variable, status, while the y axis represents the other categorical variable, patient tissue. The legend shows the color spectrum of the numerical variable, n, which is the counts of status. As count gets higher, the balloons get bigger in size and lighter in color. In general the gain status has balloons with greater size and lighter color, especially the balloons from tissue BC05 - BC11. The patient tissues BC06 and BC07 for the gain status, have yellow balloon color and are the largest in size (n >200). The loss status has more corresponding patient tissues with low balloon size and dark color (n<50) compared to other status types.

```
#To make a matrix (contingency table) from a data frame, use the function
xtabs() which takes in a formula and the data frame as arguments. In this
case this function takes in the formula, n ~ status + Patient.tissue,that
cross-tabulates the counts by the status and Patient.tissue variables.


copynumvar2.tab <- xtabs(n ~ status + Patient.tissue, data = copynumvarChi2)
copynumvar2.tab # output the contingency table
```

```
##         Patient.tissue
## status   BC01 BC02 BC03 BC03LN BC04 BC05 BC06 BC07 BC07LN BC08 BC09 BC10
BC11
##   gain     80   64   29     34   53  117  163  225    117   88  234  153
121
##   loss      8   76    6      4   10  118    5   21     17   11   25   13
5
##   normal   40   26   43     47   45   32   48   30     46   35  130  162
76
```

```
# Perform the Chi Sq test with the function chiseq.test() which takes in the
matrix. Save results to a variable copynumvarchi2.

copynumvarchi2 <- chisq.test(copynumvar2.tab)
```

```
## Pearson's Chi-squared test

## data:  copynumvar2.tab

## X-squared = 695.22, df = 24, p-value < 2.2e-16
```

```
copynumvarchi2$expected # This is the expected counts
```

```
##         Patient.tissue
## status       BC01     BC02      BC03   BC03LN     BC04      BC05      BC06
##   gain   73.98670 95.95151 45.085647 49.13180 62.42628 154.33164 124.85256
##   loss   15.96871 20.70943  9.730935 10.60422 13.47360  33.30974  26.94720
##   normal 38.04458 49.33907 23.183418 25.26398 32.10012  79.35862  64.20023
##         Patient.tissue
## status        BC07   BC07LN     BC08      BC09      BC10      BC11
##   gain   159.53383 104.0438 77.45483 224.85022 189.59093 116.76027
##   loss    34.43254  22.4560 16.71725  48.52992  40.91983  25.20063
##   normal  82.03363  53.5002 39.82792 115.61987  97.48925  60.03911
```

```
copynumvarchi2$observed #This is the observed counts
```

```
##         Patient.tissue
## status   BC01 BC02 BC03 BC03LN BC04 BC05 BC06 BC07 BC07LN BC08 BC09 BC10
```

```
BC11
##   gain     80   64   29     34   53  117  163  225    117   88  234  153
121
##   loss      8   76    6      4   10  118    5   21     17   11   25   13
5
##   normal   40   26   43     47   45   32   48   30     46   35  130  162
76

copynumvarchi2$p.value #This gives us the p-value.

## [1] 2.511295e-131
```

Conclusion

In the final analysis, the single-cell transcriptome profile characterizes the clinically important sub populations, the single-cell expression profile characterizes the important immune cell characteristics, and the single-cell analysis shows the unique carcinoma characteristics and diverse micro-environmental populations shared by different patient tumors. One-way ANOVA was used to analyze whether the copy numbers vary significantly between the different tumor types. After performing the one-way ANOVA test, the p value observed was less 2.16e^-06, we therefore concluded that there was variation in copy number.

Citation

Chung, Woosung et al (2017). “Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer.” Nature communications vol. 8 15081. 5 May. 2017, doi:10.1038/ncomms15081