1. **Q1. Is there a statistically significant difference in the expression levels of the 4 immune cell markers between treatment arms before treatment (analyze each marker separately)?**

* These are the results summarised in the table below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **IL1B** | **CX3CL1** | **TNFA** | **CCL20** |
| **Distribution** | Normal | Normal | Non normal | Non normal |
| **Test used** | Two sample T-test | Two sample T test | Wilcoxon | Wilcoxon |
| **Statistically different?** | Yes | No | No | No |

* First step was to check the distribution of the data to determine whether to use a parametric or non-parametric test.
* The data was visualised first, then a hypothesis test confirmed the distribution.
* A value for each arm and corresponding marker was generated.
* Text

  Description automatically generated
* The data was next visually appraised using histograms with the standard deviation superimposed with a red line. A QQ plot for normalization with a trend line was also generated and assessed visually. The Shapiro- wilks test was then preformed as a hypothesis test of normality.
* This is the code that was used and the variables subbed in:
* **Text

  Description automatically generated with medium confidence**

|  |  |
| --- | --- |
| **IL1B\_pre Lapatinib** | **IL1B\_pre Trastuzumab** |
|  |  |
| * Visually the data appears to relatively flat, you could argue a slight right skew as the first two bars of the histogram are above the expected probability as seen on the red line. * There is also a drop in the histogram bars where the probability of finding a value should be highest. | * The data visually appears to be right skewed as the histogram bars tail off. * There is also a drop in the histograms as we go towards the peak, i.e. the bar is less than the expected probability of finding a value of this value. |
|  |  |
| * The points follow the line of normal distribution quite well with the majority of points on the line. * We can see the slight right tailing as the points deviate from the line at the higher expression values. | * You can see how the data tails as the points drift from the line of normal distribution towards the right of the graph. * In general, the rest of the points do follow the line closely. |
| Shapiro p value: p-value = 0.07388 | Shapiro p value: p-value =0.1837 |
| Distribution: Normal distribution | |
| Variation:   * The variation of the data in the two treatment arms was analysed to determine if a student-t test or a welch t test should be used. The data was separated by treatment arm to run the comparison of variation. * **Text    Description automatically generated** | |
| Significance:   * We let the variance=TRUE as the p value was not significant, indicating to use a two sample T test. * This returned a significant p value * There is a significant difference in the expression levels of the IL1B immune cell marker between treatment arms before treatment. * We can also see there is a notable difference in the exact values of the means to support this. | |

|  |  |
| --- | --- |
| **CX3CL1 Lapatinib arm:** | **CX3CL1 Trastuzumab arm:** |
| Chart, line chart, histogram  Description automatically generated | Chart, line chart, histogram  Description automatically generated |
| * This data appears to follow the lline of. Normal distribution quite well. * The peak does slightly trend above the red line of normal distribution and the bars move quite well in a step like motion. | * The data does not appear to be normally distributed as the bar at 1.5-2.0 trends above the expected probability. * There is also a big drop at the 2.5-3.0 position before the data rises again |
|  |  |
| * The points follow the direction of the line very well but very few actually lie on the line of normality. * The variation appears to be quite small so the data could still be normally distributed. | * More of these points lie on of very close to the line of normal distribution. * You can see the peak in the data of the higher values as the points digress from the line. |
| Shapiro p value: p=0.05118 | Shapiro p value: p=0.2927 |
| Distribution: Normal Distribution | Distribution: Normal Distribution |
| * Variance was then tested to determine the parametric t test to be used in the same process as for IL1B.  * This returned a non-significant p-value thus the two groups have equal variance, so we use a two-sample t test.  * This test returned a non-significant p value. Thus, we can say there was no significant difference in the means between the treatment arms for the CX3CL1 marker. | |

|  |  |
| --- | --- |
| **TNFA Lapatinib arm:** | **TNFA Trastuzumab arm:** |
| Chart, line chart, histogram  Description automatically generated | Chart, line chart  Description automatically generated |
| * This data is right tailed or positvely skewed. * The lower values appear more than the expected normal distribution and the values tail off witth 10-15 being much lower than the expected probability. | * This data is also right tailed or positvely skewed. * The lower bars are higher than the expectecd probability and the data then drops below the expected probability and plateaus. |
|  |  |
| * You can see the tailing of the data as the points stray away from the line of normal distribution. * The middle values do follow the line of normal distribution quite well. * Looking at this you would say the data is not normally distributed. | * Again we can see the skewing of the data as the points drift from the line of normal distribution. * The middle points do relatively follow the line of normal distribution but overall you could say the data is not normally distributed. |
| Shapiro p value: p=1.122e-05 | Shapiro p value: p=4.718e-05 |
| Distribution: Non-normal distribution | Distribution: Non-normal distribution |
| Significance:   * Since the data is not normally distributed a non-parametric Wilcox test will be used again to compare the medians.  * The wilcox rank sum test returned a non-significant p value indicating that there is no significant difference between the expression levels of TNFA between treatment arms pretreatment. * The power of the p value is reduced because it is a non-parametric test but this result is not close to significant. | |

|  |  |
| --- | --- |
| **CCL20 Lapatinib arm:** | **CCL20 Trastuzumab arm:** |
| Chart, line chart, histogram  Description automatically generated | Chart, line chart, histogram  Description automatically generated |
| * This data is clearly not normally distributed. * The first bar in the histogram is well above the line of normal distribtuion. * The rest of the data barely appears on the histogram indicating the right skew of the data. | * This data is also clearly not normally distributed. * Similar to the Lapatinib arm the first value is well above the line of expected probability for normal distribution. * The data again tails to the right. |
| * This could be a trend in the CCL20 marker that it may normally sit at low levels and these higher values are possible outliers. * Chart, box and whisker chart    Description automatically generated with medium confidence * This box plot shows the inter quartile range is very narrow and lies very xlose to the median. * The data has quite a few outliers which are seen as the data points above the box plot. * It is possible as per the central limit therom that if we were to increase the population size these outliers wouldn’t have as much of an effect on the data and lead to normally distributed data. | |
|  |  |
| * The data points appear to follow the line of normal distribution but ideally we would like a line at a 45º angle from the origin. * We can see the tailing of the data as the points stray away from the line of normal distribution. | * Simialr to the points in the lapatinib arm majority of the points fit the line of normal distribution quite well but we would desire the line at a 45º angle from the origin. * Again we can see the tailing of the data as the points stray from the line of normal distribution. |
| Shapiro p value: p=1.122e-05 | Shapiro p value: p=4.718e-05 |
| Distribution: Non-normal distribution | Distribution: Non-normal distribution |
| * Significance: * Since the data is not normally distributed a non-parametric Wilcox test will be used again to compare the medians.  * The wilcox rank sum test returned a non-significant p value indicating that there is no significant difference between the expression levels of CCL20 between treatment arms pretreatment. * Again, the power of the p value is reduced because it is a non-parametric test, but this is well above the significance level. | |

**Q2. We also have information on “responders” and the “nonresponders”. Are estrogen receptor positive sample more likely to be non-responders?**

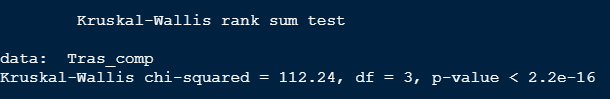
|  |
| --- |
| Count:   * I counted the data using the code below. * Text    Description automatically generated |
| Matrix:   * I then input this data into a matrix to produce a table to carry out the hypothesis test.      * This gave us this table below: * Text    Description automatically generated |
| * We can see from the mosaic plot [code: mosaicplot(ERvsResponse] that in the testing group that there are more non responders in the ER positive group than the responders group. * You can also see the reverse that there are more responders when the ER status is negative. |
| Hypothesis test:   * I used a Fishers test to test if ER status was related to treatment response. I used fishers over a chi-square test as there were values <20.  * This test returned a significant p value. Thus, we reject the null hypothesis and accept the alternative hypothesis that there is a relationship between ER status and treatment response. * The odds ratio also tells we are 3.5 times more likely to a non-responder when ER status is positive. |

**Q3. Is there a statistically significant difference between the expression levels of the 4 immune cell markers in the Trastuzumab treatment arm (pre-treatment), and where does that difference lie?**

* The data frame was edited in excel to only include data associated with Trastuzumab and attached as the dataset. This was then piped into a new data frame to only select the four variables of interest.
* A box plot was also created to visually assess if there may be a significant difference between the markers:
* Chart, box and whisker chart

  Description automatically generated
* We can see visually that there is a difference in the medians of the groups but let’s confirm this with a hypothesis test.
* We know from question 1
  + IL1B: Normally Distributed
  + CX3CL1: Normally Distributed
  + TNFA: Non normal Distribution
  + CCL20: Non normal Distribution.
* Since there is non-normally distributed data, we will use the Kruskal test. This will reduce the power of the test so our p value will not be as reliable. The power is reduced because of the ranking of data rather than the true values.
* Text

  Description automatically generated
* This test will compare the mean of the ranks.



* The P value returned is significant so we reject the null hypothesis and can say there is a significant difference in the rank of the means of at least 2 groups of the four markers.
* Next a post hoc test will tell us where this difference lies.
* Text

  Description automatically generated
* The Dunn test runs a pairwise comparison in all possible ways, in this case 6 pairs.
* This result tells us there is a significant difference between the median of all the expression markers in the pretreatment arm of Trastuzumab.
* The adjusted p value is needed because we completed several tests and it reduces the possibility of false positives as it tightens the p value cut off.
* It adjusts the p value depending on the method chosen. The idea is to make the p value more stringent as the more tests you carry out the more likely you are by chance to get a significant p value. The bh method adjusts the p value on the basis of the ranks of the p values and the number of tests carried out.
* We can see that all the p values are at varying levels of significant.
* The z scores tells us how many standard deviations away from the mean the data is. The closer to 0 the more similar the groups means are.
* We can note in group 6 that the z score shows that the means between the pair as not dissimilar, we can also see on the box plot that the medians of the variables are not too dissimilar either. However, the p values still indicated a significant difference.