DEG group comparison analysis for Bladder Cancer - Stage 1 vs. Pre Cancer

Project: Bladder Cancer Data Analysis

Purpose: Exploring molecular changes and identify potential bio-markers at different stages of bladder cancer, as well as differences in immune cells types that have infiltrated tumors.

Team 4: My team is tasked with completing two comparisons. These include comparing non invasive stage vs. precancerous tissue, and invasive stage vs non invasive stage cancer tissue

This file: Differential Gene Expression analysis using Ttest in R (group comparison) for Part I - Comparing Stage 1 (non-invasive) Cancer vs. Precancerous Tissue.

1a - Read in clinical data

```
clinData <- read.csv(file = "input/BC_ClinData_233rows.csv", header = T, stringsAsFactors = F, row.names
```

```
head(clinData)
```

```
##        SEX AGE invasiveness Intravesical.therapy systemic.chemo TMN.stage Grade
## BS017 <NA>  NA           NA                   NA             NA      <NA>    NA
## BS038 <NA>  NA           NA                   NA             NA      <NA>    NA
## BS039 <NA>  NA           NA                   NA             NA      <NA>    NA
## BS040 <NA>  NA           NA                   NA             NA      <NA>    NA
## BS041 <NA>  NA           NA                   NA             NA      <NA>    NA
## BS044 <NA>  NA           NA                   NA             NA      <NA>    NA
##       Classifier.recurrence Classifier.progression..non.muscle.invasive.
## BS017                    NA                                           NA
## BS038                    NA                                           NA
## BS039                    NA                                           NA
## BS040                    NA                                           NA
## BS041                    NA                                           NA
## BS044                    NA                                           NA
##       Classifier.progression..muscle.invasive.
## BS017                                       NA
## BS038                                       NA
## BS039                                       NA
## BS040                                       NA
## BS041                                       NA
## BS044                                       NA
##       classifier.cancer.specific.survival classifier.overall.survival
```

```
## BS017                                  NA                          NA
## BS038                                  NA                          NA
## BS039                                  NA                          NA
## BS040                                  NA                          NA
## BS041                                  NA                          NA
## BS044                                  NA                          NA
##       recurrence progression overall.survival cancer.specific.survival
## BS017         NA          NA               NA                        NA
## BS038         NA          NA               NA                        NA
## BS039         NA          NA               NA                        NA
## BS040         NA          NA               NA                        NA
## BS041         NA          NA               NA                        NA
## BS044         NA          NA               NA                        NA
##       survivalMonth           Patient     GSMid
## BS017            NA Surrounding BS017 GSM340547
## BS038            NA Surrounding BS038 GSM340548
## BS039            NA Surrounding BS039 GSM340549
## BS040            NA Surrounding BS040 GSM340550
## BS041            NA Surrounding BS041 GSM340551
## BS044            NA Surrounding BS044 GSM340552
##                  X.Sample_source_name_ch1           PrimaryBladderCancerType
## BS017 Bladder mucosae surrounding cancer Bladder mucosae surrounding cancer
## BS038 Bladder mucosae surrounding cancer Bladder mucosae surrounding cancer
## BS039 Bladder mucosae surrounding cancer Bladder mucosae surrounding cancer
## BS040 Bladder mucosae surrounding cancer Bladder mucosae surrounding cancer
## BS041 Bladder mucosae surrounding cancer Bladder mucosae surrounding cancer
## BS044 Bladder mucosae surrounding cancer Bladder mucosae surrounding cancer
```

```
knitr::kable(head(clinData), caption = "Clinical Data with rows as patients and attributes as columns")
```

Table 1: Clinical Data with rows as patients and attributes as columns

	SEX	AGE	Invasive	Intermediate	TMB	Class	Classifier	Classifier	Classifier	Classifier	Classifier	Classifier	Classifier	Classifier	Classifier	Classifier	Patient	GSMid	X.Sample_source_name_ch1	PrimaryBladderCancerType
BS017	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	Surrounding BS017	GSM340547	Bladder mucosae surrounding cancer	Bladder mucosae surrounding cancer
BS038	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	Surrounding BS038	GSM340548	Bladder mucosae surrounding cancer	Bladder mucosae surrounding cancer

	SEX	AGE	Invasive	Intravesical	TMR	Chemo	Classifier	Classifier	Classifier	Classifier	Classifier	Classifier	Classifier	Classifier	Classifier	Classifier	Primary	Bladder	Cancer
BS039	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	Surrounding BS039	GSM340546	Bladder mucosae surrounding cancer
BS040	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	Surrounding BS040	GSM340546	Bladder mucosae surrounding cancer
BS041	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	Surrounding BS041	GSM340546	Bladder mucosae surrounding cancer
BS044	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	Surrounding BS044	GSM340546	Bladder mucosae surrounding cancer

Checking the shape and overall information about the clinical dataset

```
dim(clinData)
```

```
## [1] 233  21
```

1b: Read in processed normalized gene expression data in log2 scale, includes gene annotation

```
# Read in gene expression file - features (rows), 30 patients (columns)
geneExp <- read.csv(file = "input/BC_GeneExpData_withAnno_233.tsv", sep = "\t", header = T, stringsAsFa
head(geneExp[1:5, 1:4])
```

```
##                         GSM340547 GSM340548 GSM340549 GSM340550
## ILMN_1651199|LOC643334      7.00      6.98      7.05      7.53
```

```
## ILMN_1651209|SLC35E2        8.01      7.83      7.67      6.97
## ILMN_1651210|DUSP22        10.12     10.25     10.44      9.90
## ILMN_1651217|PDCD1LG2       6.86      6.90      6.86      7.38
## ILMN_1651221|LOC642820      7.10      7.45      7.37      7.34
```

```r
knitr::kable(head(geneExp), caption = "Processed gene expression data in log2 scale with gene annotatio
```

	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]

Now lets combine the dataset using the Patient's GSMID as the unique identifier

```r
head(clinData$GSMid) # these are the identifiers in the clinical data
```

```
## [1] "GSM340547" "GSM340548" "GSM340549" "GSM340550" "GSM340551" "GSM340552"
```

```r
head(colnames(geneExp)) # same identifiers in the gene exp data
```

```
## [1] "GSM340547" "GSM340548" "GSM340549" "GSM340550" "GSM340551" "GSM340552"
```

```r
matchingSamples = which(colnames(geneExp) %in% (clinData$GSMid)) # 233 samples
head(matchingSamples)
```

```
## [1] 1 2 3 4 5 6
```

```r
subsetGeneExp <- geneExp[, matchingSamples]
head(subsetGeneExp)
```

```
##                          GSM340547 GSM340548 GSM340549 GSM340550 GSM340551
## ILMN_1651199|LOC643334      7.00      6.98      7.05      7.53      7.21
## ILMN_1651209|SLC35E2        8.01      7.83      7.67      6.97      7.91
## ILMN_1651210|DUSP22        10.12     10.25     10.44      9.90     10.63
## ILMN_1651217|PDCD1LG2       6.86      6.90      6.86      7.38      7.04
## ILMN_1651221|LOC642820      7.10      7.45      7.37      7.34      7.27
## ILMN_1651228|RPS28         15.55     14.51     15.40     15.57     15.02
##                          GSM340552 GSM340553 GSM340554 GSM340555 GSM340556
## ILMN_1651199|LOC643334      6.96      7.16      6.99      7.04      6.98
## ILMN_1651209|SLC35E2        8.06      8.11      7.54      7.71      7.55
## ILMN_1651210|DUSP22        10.05     10.18     10.04      9.82      9.57
## ILMN_1651217|PDCD1LG2       7.13      6.99      7.23      7.01      7.07
```

```
## ILMN_1651221|LOC642820      7.09      7.23      7.09      7.26      7.02
## ILMN_1651228|RPS28         14.72     15.51     15.09     15.38     15.41
##                         GSM340557 GSM340558 GSM340559 GSM340560 GSM340561
## ILMN_1651199|LOC643334      6.89      6.97      7.06      7.13      7.20
## ILMN_1651209|SLC35E2        7.67      7.97      7.48      7.47      7.35
## ILMN_1651210|DUSP22        10.31     10.19      9.75      9.95     10.23
## ILMN_1651217|PDCD1LG2       7.12      7.00      7.09      7.14      7.14
## ILMN_1651221|LOC642820      7.27      7.09      7.03      7.11      7.34
## ILMN_1651228|RPS28         15.26     14.97     15.32     15.16     14.89
##                         GSM340562 GSM340563 GSM340564 GSM340565 GSM340566
## ILMN_1651199|LOC643334      6.79      6.95      7.08      6.91      6.97
## ILMN_1651209|SLC35E2        7.89      7.69      7.92      7.99      7.81
## ILMN_1651210|DUSP22        10.47     10.22     10.05     10.55     10.02
## ILMN_1651217|PDCD1LG2       7.04      7.38      7.10      7.22      6.99
## ILMN_1651221|LOC642820      7.13      7.44      7.14      7.33      7.38
## ILMN_1651228|RPS28         14.60     15.14     14.59     14.86     14.42
##                         GSM340567 GSM340568 GSM340569 GSM340570 GSM340571
## ILMN_1651199|LOC643334      6.88      7.02      7.11      7.09      7.00
## ILMN_1651209|SLC35E2        7.72      7.93      7.55      7.82      7.83
## ILMN_1651210|DUSP22        10.67     10.14     10.42     10.84     10.54
## ILMN_1651217|PDCD1LG2       7.13      7.17      7.10      7.31      7.64
## ILMN_1651221|LOC642820      7.20      7.15      6.98      7.20      7.20
## ILMN_1651228|RPS28         14.20     14.42     13.78     14.10     14.18
##                         GSM340572 GSM340573 GSM340574 GSM340575 GSM340576
## ILMN_1651199|LOC643334      6.94      7.02      7.00      6.83      6.95
## ILMN_1651209|SLC35E2        7.89      7.62      7.90      7.43      7.54
## ILMN_1651210|DUSP22        10.09     10.56     10.31     10.50     10.44
## ILMN_1651217|PDCD1LG2       7.03      7.42      7.12      7.09      7.28
## ILMN_1651221|LOC642820      7.10      7.13      7.23      7.11      7.17
## ILMN_1651228|RPS28         14.38     13.59     15.24     14.91     14.58
##                         GSM340577 GSM340578 GSM340579 GSM340580 GSM340581
## ILMN_1651199|LOC643334      6.92      6.99      6.92      6.94      7.03
## ILMN_1651209|SLC35E2        7.82      7.85      7.58      7.60      8.24
## ILMN_1651210|DUSP22         9.90     10.05     10.26      9.95     10.53
## ILMN_1651217|PDCD1LG2       7.18      6.96      7.15      7.09      7.03
## ILMN_1651221|LOC642820      7.12      7.14      7.10      7.13      7.08
## ILMN_1651228|RPS28         14.87     15.23     14.61     14.27     11.83
##                         GSM340582 GSM340583 GSM340584 GSM340585 GSM340586
## ILMN_1651199|LOC643334      6.91      7.06      7.17      7.15      6.89
## ILMN_1651209|SLC35E2        8.03      7.05      7.77      7.02      7.65
## ILMN_1651210|DUSP22        10.21     10.17     10.25      9.89     10.00
## ILMN_1651217|PDCD1LG2       7.13      7.06      6.91      7.14      7.02
## ILMN_1651221|LOC642820      7.23      6.99      7.29      7.05      7.22
## ILMN_1651228|RPS28         14.89     15.49     15.26     15.56     15.13
##                         GSM340587 GSM340588 GSM340589 GSM340590 GSM340591
## ILMN_1651199|LOC643334      7.04      6.88      6.97      6.89      6.92
## ILMN_1651209|SLC35E2        7.50      7.59      7.74      7.52      7.58
## ILMN_1651210|DUSP22        10.37     10.18      9.67     10.14      9.94
## ILMN_1651217|PDCD1LG2       6.98      7.09      7.02      7.24      7.12
## ILMN_1651221|LOC642820      7.11      7.16      7.15      7.14      7.20
## ILMN_1651228|RPS28         14.49     14.49     15.09     14.96     15.25
##                         GSM340592 GSM340593 GSM340594 GSM340595 GSM340596
## ILMN_1651199|LOC643334      7.03      6.91      6.84      7.01      6.81
## ILMN_1651209|SLC35E2        7.45      7.45      7.52      7.65      8.01
```

```
## ILMN_1651210|DUSP22        10.36     10.26      9.88      9.99     10.07
## ILMN_1651217|PDCD1LG2       7.44      6.98      7.16      7.05      6.96
## ILMN_1651221|LOC642820      7.31      7.27      7.03      7.21      7.22
## ILMN_1651228|RPS28         15.09     15.30     14.11     15.45     14.70
##                         GSM340597 GSM340598 GSM340599 GSM340600 GSM340601
## ILMN_1651199|LOC643334      6.91      6.83      6.96      6.89      6.87
## ILMN_1651209|SLC35E2        7.79      7.67      7.62      7.60      7.52
## ILMN_1651210|DUSP22         9.94      9.79     10.10      9.81      9.86
## ILMN_1651217|PDCD1LG2       7.03      7.20      7.00      7.28      6.88
## ILMN_1651221|LOC642820      7.18      6.94      7.10      7.03      7.17
## ILMN_1651228|RPS28         15.12     15.18     15.36     15.00     14.40
##                         GSM340602 GSM340603 GSM340604 GSM340605 GSM340606
## ILMN_1651199|LOC643334      6.89      7.13      7.07      6.97      7.12
## ILMN_1651209|SLC35E2        7.52      7.57      7.58      8.38      9.62
## ILMN_1651210|DUSP22        10.09     10.46     10.49      9.14      7.25
## ILMN_1651217|PDCD1LG2       6.95      7.15      7.33      6.92      6.89
## ILMN_1651221|LOC642820      7.16      7.07      6.97      7.38      7.95
## ILMN_1651228|RPS28         15.38     15.27     14.79     15.34     14.47
##                         GSM340607 GSM340608 GSM340609 GSM340610 GSM340611
## ILMN_1651199|LOC643334      6.88      6.94      7.06      6.97      6.95
## ILMN_1651209|SLC35E2        8.10      8.58      8.78      8.13      8.09
## ILMN_1651210|DUSP22        10.08     10.34      9.43     10.04      9.13
## ILMN_1651217|PDCD1LG2       7.11      7.08      6.92      6.91      6.99
## ILMN_1651221|LOC642820      7.07      7.23      6.99      7.16      7.30
## ILMN_1651228|RPS28         15.15     15.05     15.51     14.40     15.18
##                         GSM340612 GSM340613 GSM340614 GSM340615 GSM340616
## ILMN_1651199|LOC643334      6.94      6.99      7.07      6.92      6.91
## ILMN_1651209|SLC35E2        8.00      8.29      8.38      7.95      7.78
## ILMN_1651210|DUSP22        10.28     10.23      9.42     10.39     10.40
## ILMN_1651217|PDCD1LG2       7.32      7.10      7.00      7.10      7.08
## ILMN_1651221|LOC642820      7.15      7.17      7.34      7.12      7.37
## ILMN_1651228|RPS28         15.26     15.51     15.15     13.91     15.01
##                         GSM340617 GSM340618 GSM340619 GSM340620 GSM340621
## ILMN_1651199|LOC643334      7.16      6.90      7.02      7.08      7.21
## ILMN_1651209|SLC35E2        8.16      8.06      9.05      8.56      8.11
## ILMN_1651210|DUSP22        10.34      9.75      9.96     10.02      9.74
## ILMN_1651217|PDCD1LG2       6.92      7.00      6.96      7.09      6.96
## ILMN_1651221|LOC642820      7.09      7.16      7.19      7.12      7.14
## ILMN_1651228|RPS28         15.24     14.54     15.31     15.26     15.33
##                         GSM340622 GSM340623 GSM340624 GSM340625 GSM340626
## ILMN_1651199|LOC643334      6.95      7.02      7.05      6.76      6.91
## ILMN_1651209|SLC35E2        8.17      8.39      8.15      8.56      9.08
## ILMN_1651210|DUSP22         9.78     10.82     10.47     10.40     10.01
## ILMN_1651217|PDCD1LG2       6.90      7.15      7.04      7.11      6.99
## ILMN_1651221|LOC642820      7.12      7.10      7.15      7.42      7.30
## ILMN_1651228|RPS28         14.66     15.39     14.88     15.39     15.33
##                         GSM340627 GSM340628 GSM340629 GSM340630 GSM340631
## ILMN_1651199|LOC643334      6.99      7.04      7.07      7.01      6.87
## ILMN_1651209|SLC35E2        8.66      7.82      7.62      8.64      8.79
## ILMN_1651210|DUSP22        10.68     10.26     10.07      9.42      9.72
## ILMN_1651217|PDCD1LG2       7.03      7.25      7.03      6.94      7.10
## ILMN_1651221|LOC642820      7.38      7.40      7.19      7.05      7.21
## ILMN_1651228|RPS28         15.34     15.13     15.24     15.26     14.99
##                         GSM340632 GSM340633 GSM340634 GSM340635 GSM340636
```

```
## ILMN_1651199|LOC643334      6.93      6.98      7.08      7.10      6.99
## ILMN_1651209|SLC35E2        8.48      8.64      7.76      8.17      8.47
## ILMN_1651210|DUSP22         9.91     10.41      9.16     10.11     10.29
## ILMN_1651217|PDCD1LG2       6.78      7.06      7.13      6.66      7.03
## ILMN_1651221|LOC642820      7.16      7.16      7.12      6.92      7.33
## ILMN_1651228|RPS28         14.72     14.57     14.86     15.49     15.16
##                         GSM340637 GSM340638 GSM340639 GSM340640 GSM340641
## ILMN_1651199|LOC643334      6.86      6.90      6.85      6.91      7.16
## ILMN_1651209|SLC35E2        7.63      8.25      8.07      7.65      7.95
## ILMN_1651210|DUSP22        10.17     10.05     10.15      9.68     10.01
## ILMN_1651217|PDCD1LG2       7.28      7.09      7.23      6.88      6.88
## ILMN_1651221|LOC642820      7.27      7.16      7.23      7.25      7.05
## ILMN_1651228|RPS28         15.12     15.34     14.52     15.36     15.27
##                         GSM340642 GSM340643 GSM340644 GSM340645 GSM340646
## ILMN_1651199|LOC643334      7.06      6.91      7.15      6.83      6.96
## ILMN_1651209|SLC35E2        8.26      8.11      8.21      7.86      7.54
## ILMN_1651210|DUSP22         9.46     10.17     10.37     10.32      9.69
## ILMN_1651217|PDCD1LG2       7.19      7.04      6.87      7.16      7.00
## ILMN_1651221|LOC642820      7.22      7.24      7.14      7.15      7.05
## ILMN_1651228|RPS28         15.23     15.47     15.02     15.09     15.46
##                         GSM340647 GSM340648 GSM340649 GSM340650 GSM340651
## ILMN_1651199|LOC643334      6.87      6.76      6.67      7.04      7.06
## ILMN_1651209|SLC35E2        8.61      8.12      7.89      8.00      8.35
## ILMN_1651210|DUSP22        10.03      9.70     10.04     10.78     10.21
## ILMN_1651217|PDCD1LG2       6.94      7.00      6.95      7.19      7.06
## ILMN_1651221|LOC642820      7.18      7.07      7.05      7.23      7.09
## ILMN_1651228|RPS28         14.35     15.22     15.16     15.30     14.80
##                         GSM340652 GSM340653 GSM340654 GSM340655 GSM340656
## ILMN_1651199|LOC643334      6.85      7.05      7.06      7.09      6.99
## ILMN_1651209|SLC35E2        8.16      8.51      7.93      8.09      7.81
## ILMN_1651210|DUSP22         9.83      9.66      9.79      9.29      9.49
## ILMN_1651217|PDCD1LG2       6.91      7.02      6.92      6.96      7.17
## ILMN_1651221|LOC642820      7.11      7.18      7.23      7.13      7.28
## ILMN_1651228|RPS28         15.42     14.95     15.40     15.51     15.39
##                         GSM340657 GSM340658 GSM340659 GSM340660 GSM340661
## ILMN_1651199|LOC643334      6.87      6.97      6.95      6.98      6.93
## ILMN_1651209|SLC35E2        9.06      7.90      8.24      7.41      8.23
## ILMN_1651210|DUSP22         9.70     10.33     10.06     10.92      9.82
## ILMN_1651217|PDCD1LG2       7.15      6.96      7.11      7.05      7.16
## ILMN_1651221|LOC642820      7.20      7.47      7.22      7.19      7.05
## ILMN_1651228|RPS28         14.36     15.36     15.03     15.09     15.33
##                         GSM340662 GSM340663 GSM340664 GSM340665 GSM340666
## ILMN_1651199|LOC643334      7.05      6.88      7.18      7.12      6.80
## ILMN_1651209|SLC35E2        8.57      7.76      8.32      8.58      8.54
## ILMN_1651210|DUSP22         9.24      9.53     10.30     10.26      9.29
## ILMN_1651217|PDCD1LG2       7.05      7.14      7.07      7.13      7.07
## ILMN_1651221|LOC642820      7.03      7.22      7.33      7.05      7.22
## ILMN_1651228|RPS28         15.13     14.90     15.46     15.26     15.46
##                         GSM340667 GSM340668 GSM340669 GSM340670 GSM340671
## ILMN_1651199|LOC643334      7.02      7.01      7.04      7.01      6.99
## ILMN_1651209|SLC35E2        8.22      8.74      7.97      8.21      9.44
## ILMN_1651210|DUSP22         9.96     10.03      9.97     10.69      9.39
## ILMN_1651217|PDCD1LG2       7.02      7.12      6.99      7.13      7.00
## ILMN_1651221|LOC642820      7.16      7.07      7.43      7.26      7.19
```

```
## ILMN_1651228|RPS28         15.50     15.41     15.02     14.16     14.94
##                         GSM340672 GSM340673 GSM340674 GSM340675 GSM340676
## ILMN_1651199|LOC643334      7.02      6.86      6.90      7.05      6.88
## ILMN_1651209|SLC35E2        8.32      7.71      7.45      8.84      7.92
## ILMN_1651210|DUSP22        10.37     10.17     10.51      9.54      9.80
## ILMN_1651217|PDCD1LG2       6.95      7.04      7.27      7.02      7.19
## ILMN_1651221|LOC642820      7.08      7.00      7.08      7.28      6.98
## ILMN_1651228|RPS28         15.29     15.43     14.97     15.02     14.89
##                         GSM340677 GSM340678 GSM340679 GSM340680 GSM340681
## ILMN_1651199|LOC643334      6.93      6.92      6.99      6.82      6.97
## ILMN_1651209|SLC35E2        8.16      7.74      8.05      8.52      8.07
## ILMN_1651210|DUSP22         8.97      9.66     10.19      9.33      9.57
## ILMN_1651217|PDCD1LG2       7.27      6.86      7.13      6.94      7.09
## ILMN_1651221|LOC642820      7.24      7.28      7.40      7.24      7.16
## ILMN_1651228|RPS28         15.05     14.88     15.08     15.49     14.94
##                         GSM340682 GSM340683 GSM340684 GSM340685 GSM340686
## ILMN_1651199|LOC643334      7.02      6.88      6.91      6.96      6.94
## ILMN_1651209|SLC35E2        7.63      8.50      8.00      9.15      7.84
## ILMN_1651210|DUSP22         9.75     10.08      9.08     10.22     10.00
## ILMN_1651217|PDCD1LG2       6.88      6.84      7.04      7.08      6.81
## ILMN_1651221|LOC642820      7.35      7.15      7.04      7.18      7.28
## ILMN_1651228|RPS28         14.77     14.32     15.18     14.79     15.22
##                         GSM340687 GSM340688 GSM340689 GSM340690 GSM340691
## ILMN_1651199|LOC643334      7.00      7.08      7.08      6.99      7.04
## ILMN_1651209|SLC35E2        7.98      8.73      8.24      7.96      8.26
## ILMN_1651210|DUSP22        10.23      9.06     10.25     10.26     10.17
## ILMN_1651217|PDCD1LG2       7.77      7.03      6.99      7.09      7.14
## ILMN_1651221|LOC642820      7.15      7.21      7.05      7.12      7.22
## ILMN_1651228|RPS28         13.60     14.90     14.25     15.18     15.35
##                         GSM340692 GSM340693 GSM340694 GSM340695 GSM340696
## ILMN_1651199|LOC643334      6.79      6.97      6.77      7.06      7.07
## ILMN_1651209|SLC35E2        8.51      8.09      7.87      8.18      7.78
## ILMN_1651210|DUSP22        10.15     10.13     10.11     10.53     10.26
## ILMN_1651217|PDCD1LG2       7.04      6.77      6.87      7.02      7.39
## ILMN_1651221|LOC642820      6.98      7.12      7.23      7.14      7.28
## ILMN_1651228|RPS28         15.23     15.29     15.07     14.21     14.86
##                         GSM340697 GSM340698 GSM340699 GSM340700 GSM340701
## ILMN_1651199|LOC643334      6.92      6.95      6.94      6.84      6.83
## ILMN_1651209|SLC35E2        8.35      8.29      7.86      8.62      8.88
## ILMN_1651210|DUSP22        10.14      9.71     10.00      9.34     10.02
## ILMN_1651217|PDCD1LG2       7.20      7.12      7.18      6.85      6.87
## ILMN_1651221|LOC642820      6.89      7.22      7.17      7.13      7.07
## ILMN_1651228|RPS28         12.79     15.17     14.95     14.90     14.85
##                         GSM340702 GSM340703 GSM340704 GSM340705 GSM340706
## ILMN_1651199|LOC643334      7.03      6.98      6.95      7.35      6.82
## ILMN_1651209|SLC35E2        8.03      7.96      8.71      7.87      8.82
## ILMN_1651210|DUSP22         9.84     10.24      9.79     10.35     10.27
## ILMN_1651217|PDCD1LG2       7.00      7.06      7.08      7.15      6.90
## ILMN_1651221|LOC642820      7.34      7.12      7.28      7.25      7.19
## ILMN_1651228|RPS28         15.06     14.23     14.98     15.24     14.24
##                         GSM340707 GSM340708 GSM340709 GSM340710 GSM340711
## ILMN_1651199|LOC643334      6.92      6.86      7.10      6.97      6.99
## ILMN_1651209|SLC35E2        7.85      8.41      7.46      7.85      7.97
## ILMN_1651210|DUSP22        10.08      9.55      9.30      9.44     10.44
```

```
## ILMN_1651217|PDCD1LG2       7.05      7.09      7.52      7.03      7.01
## ILMN_1651221|LOC642820      7.14      7.09      7.32      7.16      7.05
## ILMN_1651228|RPS28         14.91     14.10     15.09     15.13     14.96
##                         GSM340712 GSM340713 GSM340714 GSM340715 GSM340716
## ILMN_1651199|LOC643334      7.04      6.88      6.87      7.03      6.90
## ILMN_1651209|SLC35E2        8.16      8.93      7.58      7.84      7.69
## ILMN_1651210|DUSP22        10.51      9.63     10.68      9.68     10.06
## ILMN_1651217|PDCD1LG2       7.04      7.16      7.15      7.16      7.05
## ILMN_1651221|LOC642820      7.16      7.31      7.23      7.34      7.07
## ILMN_1651228|RPS28         15.06     15.28     14.92     15.41     15.52
##                         GSM340717 GSM340718 GSM340719 GSM340720 GSM340721
## ILMN_1651199|LOC643334      6.94      7.12      6.94      6.97      6.91
## ILMN_1651209|SLC35E2        7.95      7.88      8.03      7.22      8.47
## ILMN_1651210|DUSP22        10.32     10.38     10.16     10.57      9.27
## ILMN_1651217|PDCD1LG2       6.97      7.00      6.96      6.99      7.05
## ILMN_1651221|LOC642820      7.01      6.98      7.02      6.93      7.23
## ILMN_1651228|RPS28         15.25     15.06     14.75     15.26     14.11
##                         GSM340722 GSM340723 GSM340724 GSM340725 GSM340726
## ILMN_1651199|LOC643334      7.10      6.92      7.06      7.01      7.00
## ILMN_1651209|SLC35E2        8.07      8.12      7.97      7.98      8.01
## ILMN_1651210|DUSP22        10.44      9.70     10.22     10.03     10.32
## ILMN_1651217|PDCD1LG2       6.96      7.19      7.04      7.24      7.09
## ILMN_1651221|LOC642820      6.93      6.84      7.16      7.11      7.37
## ILMN_1651228|RPS28         14.80     15.17     15.18     15.33     15.42
##                         GSM340727 GSM340728 GSM340729 GSM340730 GSM340731
## ILMN_1651199|LOC643334      6.96      7.02      6.94      7.02      7.02
## ILMN_1651209|SLC35E2        8.00      7.52      7.74      8.27      7.19
## ILMN_1651210|DUSP22        10.13     10.57      9.78      9.84     10.46
## ILMN_1651217|PDCD1LG2       7.10      7.18      6.94      7.05      7.01
## ILMN_1651221|LOC642820      7.08      7.09      7.16      7.35      7.52
## ILMN_1651228|RPS28         15.09     14.95     15.40     15.11     15.14
##                         GSM340732 GSM340733 GSM340734 GSM340735 GSM340736
## ILMN_1651199|LOC643334      6.99      7.06      6.98      7.10      7.13
## ILMN_1651209|SLC35E2        7.80      7.50      8.09      8.00      9.10
## ILMN_1651210|DUSP22         9.98     10.07     10.26     10.12     10.44
## ILMN_1651217|PDCD1LG2       7.36      7.24      7.11      7.02      7.06
## ILMN_1651221|LOC642820      7.27      7.19      7.25      7.16      6.91
## ILMN_1651228|RPS28         14.48     15.22     15.01     14.83     14.28
##                         GSM340737 GSM340738 GSM340739 GSM340740 GSM340741
## ILMN_1651199|LOC643334      6.93      6.94      6.91      7.02      7.03
## ILMN_1651209|SLC35E2        7.83      7.97      8.31      7.75      8.86
## ILMN_1651210|DUSP22        10.18     10.10     10.43      9.97     10.02
## ILMN_1651217|PDCD1LG2       6.99      7.06      7.07      7.11      6.84
## ILMN_1651221|LOC642820      7.07      7.15      7.05      7.03      7.35
## ILMN_1651228|RPS28         15.22     15.19     15.15     15.19     15.51
##                         GSM340742 GSM340743 GSM340744 GSM340745 GSM340746
## ILMN_1651199|LOC643334      6.96      7.08      6.79      6.81      7.08
## ILMN_1651209|SLC35E2        7.65      8.16      8.03      8.42      7.35
## ILMN_1651210|DUSP22         9.44      9.83      9.85      9.59     10.57
## ILMN_1651217|PDCD1LG2       7.09      7.30      7.07      6.93      7.10
## ILMN_1651221|LOC642820      7.45      7.16      7.19      7.26      7.50
## ILMN_1651228|RPS28         15.35     15.52     15.29     14.93     15.00
##                         GSM340747 GSM340748 GSM340749 GSM340750 GSM340751
## ILMN_1651199|LOC643334      7.07      7.14      7.03      7.04      7.13
```

```
## ILMN_1651209|SLC35E2        8.26      8.15      7.68      7.72      8.12
## ILMN_1651210|DUSP22        10.59     10.45     10.22     10.14     10.93
## ILMN_1651217|PDCD1LG2       7.03      7.22      6.93      7.17      7.18
## ILMN_1651221|LOC642820      7.49      7.10      7.08      7.33      7.16
## ILMN_1651228|RPS28         15.43     15.20     15.26     15.23     15.23
##                         GSM340752 GSM340753 GSM340754 GSM340755 GSM340756
## ILMN_1651199|LOC643334      7.00      6.95      6.89      6.94      6.76
## ILMN_1651209|SLC35E2        8.37      8.29      8.23      7.87      8.56
## ILMN_1651210|DUSP22         9.82     10.12     10.15      9.51     10.78
## ILMN_1651217|PDCD1LG2       7.07      7.06      7.24      6.92      7.07
## ILMN_1651221|LOC642820      7.04      7.21      7.15      7.16      7.24
## ILMN_1651228|RPS28         15.34     15.41     15.23     15.35     14.18
##                         GSM340757 GSM340758 GSM340759 GSM340760 GSM340761
## ILMN_1651199|LOC643334      7.06      6.91      6.89      6.77      6.81
## ILMN_1651209|SLC35E2        8.20      7.62      7.43      8.05      7.73
## ILMN_1651210|DUSP22         9.46     10.68     10.29      9.22      9.94
## ILMN_1651217|PDCD1LG2       7.12      7.29      7.24      7.16      7.13
## ILMN_1651221|LOC642820      7.24      6.98      6.99      7.16      7.19
## ILMN_1651228|RPS28         14.90     15.26     15.10     15.02     14.90
##                         GSM340762 GSM340763 GSM340764 GSM340765 GSM340766
## ILMN_1651199|LOC643334      7.10      7.06      6.89      6.84      6.85
## ILMN_1651209|SLC35E2        7.71      7.65      8.06      8.39      7.63
## ILMN_1651210|DUSP22         9.68      9.96     11.09     10.05      9.97
## ILMN_1651217|PDCD1LG2       7.47      7.07      7.11      7.08      8.83
## ILMN_1651221|LOC642820      6.91      7.11      7.09      7.31      7.05
## ILMN_1651228|RPS28         15.15     14.99     15.22     15.33     14.49
##                         GSM340767 GSM340768 GSM340769 GSM340537 GSM340538
## ILMN_1651199|LOC643334      6.89      6.91      6.97      7.12      6.90
## ILMN_1651209|SLC35E2        7.41      8.93      7.41      7.54      7.53
## ILMN_1651210|DUSP22         9.66      9.88      9.96     10.29     10.34
## ILMN_1651217|PDCD1LG2       7.31      7.11      7.25      7.06      7.10
## ILMN_1651221|LOC642820      7.09      7.12      7.12      7.32      7.06
## ILMN_1651228|RPS28         15.41     15.31     15.20     15.51     15.34
##                         GSM340539 GSM340540 GSM340541 GSM340542 GSM340543
## ILMN_1651199|LOC643334      6.98      6.93      7.10      6.97      7.00
## ILMN_1651209|SLC35E2        7.75      7.69      7.45      7.52      7.55
## ILMN_1651210|DUSP22        10.06     10.30     10.02     10.90     10.33
## ILMN_1651217|PDCD1LG2       7.03      7.03      6.89      7.46      7.39
## ILMN_1651221|LOC642820      7.00      7.07      7.08      7.26      7.14
## ILMN_1651228|RPS28         15.16     15.34     15.26     15.33     15.18
##                         GSM340544 GSM340545 GSM340546
## ILMN_1651199|LOC643334      7.07      6.92      6.93
## ILMN_1651209|SLC35E2        7.62      7.72      7.66
## ILMN_1651210|DUSP22        10.56      9.98      9.91
## ILMN_1651217|PDCD1LG2       6.99      7.18      7.06
## ILMN_1651221|LOC642820      7.13      7.28      7.18
## ILMN_1651228|RPS28         15.23     15.11     15.18
```

Step 3 - Identifying the groups to be compared

Identifying the groups to be compared (Baseline and Comparison Grps)

In this case:

- Baseline = Bladder mucosae surrounding cancer (Precancerous)
- Comparison = Primary_BC_Superficial (Stage 1 Non Invasive)

```
# Labels (row numbers) that can identify the baseline group patients
baselineGrpLabels <-  which(clinData$PrimaryBladderCancerType == "Bladder mucosae surrounding cancer") 
head(baselineGrpLabels)
```

```
## [1] 1 2 3 4 5 6
```

```
length(baselineGrpLabels)
```

```
## [1] 58
```

```
# Use the labels (row numbers) to subset baseline patients in clinical data file
clinBase <- clinData[baselineGrpLabels, ]
head(clinBase)
```

```
##        SEX AGE invasiveness Intravesical.therapy systemic.chemo TMN.stage Grade
## BS017 <NA>  NA           NA                   NA             NA      <NA>    NA
## BS038 <NA>  NA           NA                   NA             NA      <NA>    NA
## BS039 <NA>  NA           NA                   NA             NA      <NA>    NA
## BS040 <NA>  NA           NA                   NA             NA      <NA>    NA
## BS041 <NA>  NA           NA                   NA             NA      <NA>    NA
## BS044 <NA>  NA           NA                   NA             NA      <NA>    NA
##       Classifier.recurrence Classifier.progression..non.muscle.invasive.
## BS017                    NA                                           NA
## BS038                    NA                                           NA
## BS039                    NA                                           NA
## BS040                    NA                                           NA
## BS041                    NA                                           NA
## BS044                    NA                                           NA
##       Classifier.progression..muscle.invasive.
## BS017                                       NA
## BS038                                       NA
## BS039                                       NA
## BS040                                       NA
## BS041                                       NA
## BS044                                       NA
##       classifier.cancer.specific.survival classifier.overall.survival
## BS017                                  NA                          NA
## BS038                                  NA                          NA
## BS039                                  NA                          NA
## BS040                                  NA                          NA
## BS041                                  NA                          NA
## BS044                                  NA                          NA
##       recurrence progression overall.survival cancer.specific.survival
## BS017         NA          NA               NA                        NA
## BS038         NA          NA               NA                        NA
## BS039         NA          NA               NA                        NA
## BS040         NA          NA               NA                        NA
## BS041         NA          NA               NA                        NA
```

```
## BS044         NA          NA               NA                        NA
##       survivalMonth           Patient     GSMid
## BS017            NA Surrounding BS017 GSM340547
## BS038            NA Surrounding BS038 GSM340548
## BS039            NA Surrounding BS039 GSM340549
## BS040            NA Surrounding BS040 GSM340550
## BS041            NA Surrounding BS041 GSM340551
## BS044            NA Surrounding BS044 GSM340552
##                 X.Sample_source_name_ch1           PrimaryBladderCancerType
## BS017 Bladder mucosae surrounding cancer Bladder mucosae surrounding cancer
## BS038 Bladder mucosae surrounding cancer Bladder mucosae surrounding cancer
## BS039 Bladder mucosae surrounding cancer Bladder mucosae surrounding cancer
## BS040 Bladder mucosae surrounding cancer Bladder mucosae surrounding cancer
## BS041 Bladder mucosae surrounding cancer Bladder mucosae surrounding cancer
## BS044 Bladder mucosae surrounding cancer Bladder mucosae surrounding cancer
```

```
# Labels (row numbers) that can identify the comp group patients
compGrpLabels <- which(clinData$PrimaryBladderCancerType == "Primary_BC_Superficial") #103 samples
head(compGrpLabels)
```

```
## [1] 59 60 61 62 63 64
```

```
length(compGrpLabels)
```

```
## [1] 103
```

```
# Use the labels (row numbers) to subset comp patients in clinical data file
clinComp <- clinData[compGrpLabels, ]
head(clinComp)
```

```
##       SEX AGE invasiveness Intravesical.therapy systemic.chemo TMN.stage Grade
## BT001   M  78            1                    2              1    T1N0M0     1
## BT002   F  54            1                    1              1    TaN0M0     1
## BT003   M  37            1                    1              1    TaN0M0     1
## BT004   M  72            1                    2              1    T1N0M0     1
## BT005   M  68            1                    1              1    T1N0M0     1
## BT006   M  80            1                    1              1    T1N0M0     1
##       Classifier.recurrence Classifier.progression..non.muscle.invasive.
## BT001                     2                                             1
## BT002                     1                                             1
## BT003                     2                                             1
## BT004                     1                                             1
## BT005                     2                                             2
## BT006                     1                                             2
##       Classifier.progression..muscle.invasive.
## BT001                                       NA
## BT002                                       NA
## BT003                                       NA
## BT004                                       NA
## BT005                                       NA
## BT006                                       NA
##       classifier.cancer.specific.survival classifier.overall.survival
```

```
## BT001                                NA                          NA
## BT002                                NA                          NA
## BT003                                NA                          NA
## BT004                                NA                          NA
## BT005                                NA                          NA
## BT006                                NA                          NA
##       recurrence progression overall.survival cancer.specific.survival
## BT001          2           1                2                         1
## BT002          1           1                1                         1
## BT003          1           1                1                         1
## BT004          1           1                2                         1
## BT005          1           1                2                         1
## BT006          1           1                2                         1
##       survivalMonth                     Patient     GSMid
## BT001      18.16667 Primary bladder cancer BT001 GSM340605
## BT002     136.96667 Primary bladder cancer BT002 GSM340606
## BT003     136.36667 Primary bladder cancer BT003 GSM340607
## BT004      26.90000 Primary bladder cancer BT004 GSM340608
## BT005      87.06667 Primary bladder cancer BT005 GSM340609
## BT006      15.30000 Primary bladder cancer BT006 GSM340610
##       X.Sample_source_name_ch1 PrimaryBladderCancerType
## BT001   Primary bladder cancer   Primary_BC_Superficial
## BT002   Primary bladder cancer   Primary_BC_Superficial
## BT003   Primary bladder cancer   Primary_BC_Superficial
## BT004   Primary bladder cancer   Primary_BC_Superficial
## BT005   Primary bladder cancer   Primary_BC_Superficial
## BT006   Primary bladder cancer   Primary_BC_Superficial
```

```
#### Use the clinBase and clinComp objects to subset gene expression data
geneExpBase <- subsetGeneExp[, clinBase$GSMid] # 43148 feature (rows), 58 samples columns
geneExpComp <- subsetGeneExp[, clinComp$GSMid] # 43148 feature (rows), 103 samples columns

head(geneExpBase)
```

```
##                       GSM340547 GSM340548 GSM340549 GSM340550 GSM340551
## ILMN_1651199|LOC643334      7.00      6.98      7.05      7.53      7.21
## ILMN_1651209|SLC35E2        8.01      7.83      7.67      6.97      7.91
## ILMN_1651210|DUSP22        10.12     10.25     10.44      9.90     10.63
## ILMN_1651217|PDCD1LG2       6.86      6.90      6.86      7.38      7.04
## ILMN_1651221|LOC642820      7.10      7.45      7.37      7.34      7.27
## ILMN_1651228|RPS28         15.55     14.51     15.40     15.57     15.02
##                       GSM340552 GSM340553 GSM340554 GSM340555 GSM340556
## ILMN_1651199|LOC643334      6.96      7.16      6.99      7.04      6.98
## ILMN_1651209|SLC35E2        8.06      8.11      7.54      7.71      7.55
## ILMN_1651210|DUSP22        10.05     10.18     10.04      9.82      9.57
## ILMN_1651217|PDCD1LG2       7.13      6.99      7.23      7.01      7.07
## ILMN_1651221|LOC642820      7.09      7.23      7.09      7.26      7.02
## ILMN_1651228|RPS28         14.72     15.51     15.09     15.38     15.41
##                       GSM340557 GSM340558 GSM340559 GSM340560 GSM340561
## ILMN_1651199|LOC643334      6.89      6.97      7.06      7.13      7.20
## ILMN_1651209|SLC35E2        7.67      7.97      7.48      7.47      7.35
## ILMN_1651210|DUSP22        10.31     10.19      9.75      9.95     10.23
## ILMN_1651217|PDCD1LG2       7.12      7.00      7.09      7.14      7.14
## ILMN_1651221|LOC642820      7.27      7.09      7.03      7.11      7.34
```

```
## ILMN_1651228|RPS28         15.26     14.97     15.32     15.16     14.89
##                         GSM340562 GSM340563 GSM340564 GSM340565 GSM340566
## ILMN_1651199|LOC643334      6.79      6.95      7.08      6.91      6.97
## ILMN_1651209|SLC35E2        7.89      7.69      7.92      7.99      7.81
## ILMN_1651210|DUSP22        10.47     10.22     10.05     10.55     10.02
## ILMN_1651217|PDCD1LG2       7.04      7.38      7.10      7.22      6.99
## ILMN_1651221|LOC642820      7.13      7.44      7.14      7.33      7.38
## ILMN_1651228|RPS28         14.60     15.14     14.59     14.86     14.42
##                         GSM340567 GSM340568 GSM340569 GSM340570 GSM340571
## ILMN_1651199|LOC643334      6.88      7.02      7.11      7.09      7.00
## ILMN_1651209|SLC35E2        7.72      7.93      7.55      7.82      7.83
## ILMN_1651210|DUSP22        10.67     10.14     10.42     10.84     10.54
## ILMN_1651217|PDCD1LG2       7.13      7.17      7.10      7.31      7.64
## ILMN_1651221|LOC642820      7.20      7.15      6.98      7.20      7.20
## ILMN_1651228|RPS28         14.20     14.42     13.78     14.10     14.18
##                         GSM340572 GSM340573 GSM340574 GSM340575 GSM340576
## ILMN_1651199|LOC643334      6.94      7.02      7.00      6.83      6.95
## ILMN_1651209|SLC35E2        7.89      7.62      7.90      7.43      7.54
## ILMN_1651210|DUSP22        10.09     10.56     10.31     10.50     10.44
## ILMN_1651217|PDCD1LG2       7.03      7.42      7.12      7.09      7.28
## ILMN_1651221|LOC642820      7.10      7.13      7.23      7.11      7.17
## ILMN_1651228|RPS28         14.38     13.59     15.24     14.91     14.58
##                         GSM340577 GSM340578 GSM340579 GSM340580 GSM340581
## ILMN_1651199|LOC643334      6.92      6.99      6.92      6.94      7.03
## ILMN_1651209|SLC35E2        7.82      7.85      7.58      7.60      8.24
## ILMN_1651210|DUSP22         9.90     10.05     10.26      9.95     10.53
## ILMN_1651217|PDCD1LG2       7.18      6.96      7.15      7.09      7.03
## ILMN_1651221|LOC642820      7.12      7.14      7.10      7.13      7.08
## ILMN_1651228|RPS28         14.87     15.23     14.61     14.27     11.83
##                         GSM340582 GSM340583 GSM340584 GSM340585 GSM340586
## ILMN_1651199|LOC643334      6.91      7.06      7.17      7.15      6.89
## ILMN_1651209|SLC35E2        8.03      7.05      7.77      7.02      7.65
## ILMN_1651210|DUSP22        10.21     10.17     10.25      9.89     10.00
## ILMN_1651217|PDCD1LG2       7.13      7.06      6.91      7.14      7.02
## ILMN_1651221|LOC642820      7.23      6.99      7.29      7.05      7.22
## ILMN_1651228|RPS28         14.89     15.49     15.26     15.56     15.13
##                         GSM340587 GSM340588 GSM340589 GSM340590 GSM340591
## ILMN_1651199|LOC643334      7.04      6.88      6.97      6.89      6.92
## ILMN_1651209|SLC35E2        7.50      7.59      7.74      7.52      7.58
## ILMN_1651210|DUSP22        10.37     10.18      9.67     10.14      9.94
## ILMN_1651217|PDCD1LG2       6.98      7.09      7.02      7.24      7.12
## ILMN_1651221|LOC642820      7.11      7.16      7.15      7.14      7.20
## ILMN_1651228|RPS28         14.49     14.49     15.09     14.96     15.25
##                         GSM340592 GSM340593 GSM340594 GSM340595 GSM340596
## ILMN_1651199|LOC643334      7.03      6.91      6.84      7.01      6.81
## ILMN_1651209|SLC35E2        7.45      7.45      7.52      7.65      8.01
## ILMN_1651210|DUSP22        10.36     10.26      9.88      9.99     10.07
## ILMN_1651217|PDCD1LG2       7.44      6.98      7.16      7.05      6.96
## ILMN_1651221|LOC642820      7.31      7.27      7.03      7.21      7.22
## ILMN_1651228|RPS28         15.09     15.30     14.11     15.45     14.70
##                         GSM340597 GSM340598 GSM340599 GSM340600 GSM340601
## ILMN_1651199|LOC643334      6.91      6.83      6.96      6.89      6.87
## ILMN_1651209|SLC35E2        7.79      7.67      7.62      7.60      7.52
## ILMN_1651210|DUSP22         9.94      9.79     10.10      9.81      9.86
```

```
## ILMN_1651217|PDCD1LG2       7.03      7.20      7.00      7.28      6.88
## ILMN_1651221|LOC642820      7.18      6.94      7.10      7.03      7.17
## ILMN_1651228|RPS28         15.12     15.18     15.36     15.00     14.40
##                         GSM340602 GSM340603 GSM340604
## ILMN_1651199|LOC643334      6.89      7.13      7.07
## ILMN_1651209|SLC35E2        7.52      7.57      7.58
## ILMN_1651210|DUSP22        10.09     10.46     10.49
## ILMN_1651217|PDCD1LG2       6.95      7.15      7.33
## ILMN_1651221|LOC642820      7.16      7.07      6.97
## ILMN_1651228|RPS28         15.38     15.27     14.79
```

```
head(geneExpComp)
```

```
##                         GSM340605 GSM340606 GSM340607 GSM340608 GSM340609
## ILMN_1651199|LOC643334      6.97      7.12      6.88      6.94      7.06
## ILMN_1651209|SLC35E2        8.38      9.62      8.10      8.58      8.78
## ILMN_1651210|DUSP22         9.14      7.25     10.08     10.34      9.43
## ILMN_1651217|PDCD1LG2       6.92      6.89      7.11      7.08      6.92
## ILMN_1651221|LOC642820      7.38      7.95      7.07      7.23      6.99
## ILMN_1651228|RPS28         15.34     14.47     15.15     15.05     15.51
##                         GSM340610 GSM340611 GSM340615 GSM340616 GSM340617
## ILMN_1651199|LOC643334      6.97      6.95      6.92      6.91      7.16
## ILMN_1651209|SLC35E2        8.13      8.09      7.95      7.78      8.16
## ILMN_1651210|DUSP22        10.04      9.13     10.39     10.40     10.34
## ILMN_1651217|PDCD1LG2       6.91      6.99      7.10      7.08      6.92
## ILMN_1651221|LOC642820      7.16      7.30      7.12      7.37      7.09
## ILMN_1651228|RPS28         14.40     15.18     13.91     15.01     15.24
##                         GSM340619 GSM340621 GSM340622 GSM340624 GSM340625
## ILMN_1651199|LOC643334      7.02      7.21      6.95      7.05      6.76
## ILMN_1651209|SLC35E2        9.05      8.11      8.17      8.15      8.56
## ILMN_1651210|DUSP22         9.96      9.74      9.78     10.47     10.40
## ILMN_1651217|PDCD1LG2       6.96      6.96      6.90      7.04      7.11
## ILMN_1651221|LOC642820      7.19      7.14      7.12      7.15      7.42
## ILMN_1651228|RPS28         15.31     15.33     14.66     14.88     15.39
##                         GSM340626 GSM340627 GSM340629 GSM340631 GSM340632
## ILMN_1651199|LOC643334      6.91      6.99      7.07      6.87      6.93
## ILMN_1651209|SLC35E2        9.08      8.66      7.62      8.79      8.48
## ILMN_1651210|DUSP22        10.01     10.68     10.07      9.72      9.91
## ILMN_1651217|PDCD1LG2       6.99      7.03      7.03      7.10      6.78
## ILMN_1651221|LOC642820      7.30      7.38      7.19      7.21      7.16
## ILMN_1651228|RPS28         15.33     15.34     15.24     14.99     14.72
##                         GSM340635 GSM340637 GSM340638 GSM340639 GSM340640
## ILMN_1651199|LOC643334      7.10      6.86      6.90      6.85      6.91
## ILMN_1651209|SLC35E2        8.17      7.63      8.25      8.07      7.65
## ILMN_1651210|DUSP22        10.11     10.17     10.05     10.15      9.68
## ILMN_1651217|PDCD1LG2       6.66      7.28      7.09      7.23      6.88
## ILMN_1651221|LOC642820      6.92      7.27      7.16      7.23      7.25
## ILMN_1651228|RPS28         15.49     15.12     15.34     14.52     15.36
##                         GSM340641 GSM340642 GSM340643 GSM340644 GSM340645
## ILMN_1651199|LOC643334      7.16      7.06      6.91      7.15      6.83
## ILMN_1651209|SLC35E2        7.95      8.26      8.11      8.21      7.86
## ILMN_1651210|DUSP22        10.01      9.46     10.17     10.37     10.32
## ILMN_1651217|PDCD1LG2       6.88      7.19      7.04      6.87      7.16
## ILMN_1651221|LOC642820      7.05      7.22      7.24      7.14      7.15
```

```
## ILMN_1651228|RPS28         15.27     15.23     15.47     15.02     15.09
##                         GSM340646 GSM340647 GSM340649 GSM340650 GSM340651
## ILMN_1651199|LOC643334      6.96      6.87      6.67      7.04      7.06
## ILMN_1651209|SLC35E2        7.54      8.61      7.89      8.00      8.35
## ILMN_1651210|DUSP22         9.69     10.03     10.04     10.78     10.21
## ILMN_1651217|PDCD1LG2       7.00      6.94      6.95      7.19      7.06
## ILMN_1651221|LOC642820      7.05      7.18      7.05      7.23      7.09
## ILMN_1651228|RPS28         15.46     14.35     15.16     15.30     14.80
##                         GSM340652 GSM340655 GSM340656 GSM340657 GSM340658
## ILMN_1651199|LOC643334      6.85      7.09      6.99      6.87      6.97
## ILMN_1651209|SLC35E2        8.16      8.09      7.81      9.06      7.90
## ILMN_1651210|DUSP22         9.83      9.29      9.49      9.70     10.33
## ILMN_1651217|PDCD1LG2       6.91      6.96      7.17      7.15      6.96
## ILMN_1651221|LOC642820      7.11      7.13      7.28      7.20      7.47
## ILMN_1651228|RPS28         15.42     15.51     15.39     14.36     15.36
##                         GSM340659 GSM340661 GSM340662 GSM340663 GSM340664
## ILMN_1651199|LOC643334      6.95      6.93      7.05      6.88      7.18
## ILMN_1651209|SLC35E2        8.24      8.23      8.57      7.76      8.32
## ILMN_1651210|DUSP22        10.06      9.82      9.24      9.53     10.30
## ILMN_1651217|PDCD1LG2       7.11      7.16      7.05      7.14      7.07
## ILMN_1651221|LOC642820      7.22      7.05      7.03      7.22      7.33
## ILMN_1651228|RPS28         15.03     15.33     15.13     14.90     15.46
##                         GSM340666 GSM340667 GSM340668 GSM340669 GSM340670
## ILMN_1651199|LOC643334      6.80      7.02      7.01      7.04      7.01
## ILMN_1651209|SLC35E2        8.54      8.22      8.74      7.97      8.21
## ILMN_1651210|DUSP22         9.29      9.96     10.03      9.97     10.69
## ILMN_1651217|PDCD1LG2       7.07      7.02      7.12      6.99      7.13
## ILMN_1651221|LOC642820      7.22      7.16      7.07      7.43      7.26
## ILMN_1651228|RPS28         15.46     15.50     15.41     15.02     14.16
##                         GSM340671 GSM340675 GSM340676 GSM340679 GSM340680
## ILMN_1651199|LOC643334      6.99      7.05      6.88      6.99      6.82
## ILMN_1651209|SLC35E2        9.44      8.84      7.92      8.05      8.52
## ILMN_1651210|DUSP22         9.39      9.54      9.80     10.19      9.33
## ILMN_1651217|PDCD1LG2       7.00      7.02      7.19      7.13      6.94
## ILMN_1651221|LOC642820      7.19      7.28      6.98      7.40      7.24
## ILMN_1651228|RPS28         14.94     15.02     14.89     15.08     15.49
##                         GSM340681 GSM340682 GSM340683 GSM340685 GSM340686
## ILMN_1651199|LOC643334      6.97      7.02      6.88      6.96      6.94
## ILMN_1651209|SLC35E2        8.07      7.63      8.50      9.15      7.84
## ILMN_1651210|DUSP22         9.57      9.75     10.08     10.22     10.00
## ILMN_1651217|PDCD1LG2       7.09      6.88      6.84      7.08      6.81
## ILMN_1651221|LOC642820      7.16      7.35      7.15      7.18      7.28
## ILMN_1651228|RPS28         14.94     14.77     14.32     14.79     15.22
##                         GSM340687 GSM340688 GSM340689 GSM340690 GSM340691
## ILMN_1651199|LOC643334      7.00      7.08      7.08      6.99      7.04
## ILMN_1651209|SLC35E2        7.98      8.73      8.24      7.96      8.26
## ILMN_1651210|DUSP22        10.23      9.06     10.25     10.26     10.17
## ILMN_1651217|PDCD1LG2       7.77      7.03      6.99      7.09      7.14
## ILMN_1651221|LOC642820      7.15      7.21      7.05      7.12      7.22
## ILMN_1651228|RPS28         13.60     14.90     14.25     15.18     15.35
##                         GSM340692 GSM340695 GSM340698 GSM340699 GSM340703
## ILMN_1651199|LOC643334      6.79      7.06      6.95      6.94      6.98
## ILMN_1651209|SLC35E2        8.51      8.18      8.29      7.86      7.96
## ILMN_1651210|DUSP22        10.15     10.53      9.71     10.00     10.24
```

```
## ILMN_1651217|PDCD1LG2       7.04      7.02      7.12      7.18      7.06
## ILMN_1651221|LOC642820      6.98      7.14      7.22      7.17      7.12
## ILMN_1651228|RPS28         15.23     14.21     15.17     14.95     14.23
##                         GSM340707 GSM340708 GSM340710 GSM340711 GSM340713
## ILMN_1651199|LOC643334      6.92      6.86      6.97      6.99      6.88
## ILMN_1651209|SLC35E2        7.85      8.41      7.85      7.97      8.93
## ILMN_1651210|DUSP22        10.08      9.55      9.44     10.44      9.63
## ILMN_1651217|PDCD1LG2       7.05      7.09      7.03      7.01      7.16
## ILMN_1651221|LOC642820      7.14      7.09      7.16      7.05      7.31
## ILMN_1651228|RPS28         14.91     14.10     15.13     14.96     15.28
##                         GSM340716 GSM340717 GSM340719 GSM340722 GSM340724
## ILMN_1651199|LOC643334      6.90      6.94      6.94      7.10      7.06
## ILMN_1651209|SLC35E2        7.69      7.95      8.03      8.07      7.97
## ILMN_1651210|DUSP22        10.06     10.32     10.16     10.44     10.22
## ILMN_1651217|PDCD1LG2       7.05      6.97      6.96      6.96      7.04
## ILMN_1651221|LOC642820      7.07      7.01      7.02      6.93      7.16
## ILMN_1651228|RPS28         15.52     15.25     14.75     14.80     15.18
##                         GSM340726 GSM340730 GSM340734 GSM340735 GSM340736
## ILMN_1651199|LOC643334      7.00      7.02      6.98      7.10      7.13
## ILMN_1651209|SLC35E2        8.01      8.27      8.09      8.00      9.10
## ILMN_1651210|DUSP22        10.32      9.84     10.26     10.12     10.44
## ILMN_1651217|PDCD1LG2       7.09      7.05      7.11      7.02      7.06
## ILMN_1651221|LOC642820      7.37      7.35      7.25      7.16      6.91
## ILMN_1651228|RPS28         15.42     15.11     15.01     14.83     14.28
##                         GSM340737 GSM340738 GSM340739 GSM340740 GSM340741
## ILMN_1651199|LOC643334      6.93      6.94      6.91      7.02      7.03
## ILMN_1651209|SLC35E2        7.83      7.97      8.31      7.75      8.86
## ILMN_1651210|DUSP22        10.18     10.10     10.43      9.97     10.02
## ILMN_1651217|PDCD1LG2       6.99      7.06      7.07      7.11      6.84
## ILMN_1651221|LOC642820      7.07      7.15      7.05      7.03      7.35
## ILMN_1651228|RPS28         15.22     15.19     15.15     15.19     15.51
##                         GSM340742 GSM340743 GSM340744 GSM340748 GSM340749
## ILMN_1651199|LOC643334      6.96      7.08      6.79      7.14      7.03
## ILMN_1651209|SLC35E2        7.65      8.16      8.03      8.15      7.68
## ILMN_1651210|DUSP22         9.44      9.83      9.85     10.45     10.22
## ILMN_1651217|PDCD1LG2       7.09      7.30      7.07      7.22      6.93
## ILMN_1651221|LOC642820      7.45      7.16      7.19      7.10      7.08
## ILMN_1651228|RPS28         15.35     15.52     15.29     15.20     15.26
##                         GSM340750 GSM340751 GSM340752 GSM340753 GSM340754
## ILMN_1651199|LOC643334      7.04      7.13      7.00      6.95      6.89
## ILMN_1651209|SLC35E2        7.72      8.12      8.37      8.29      8.23
## ILMN_1651210|DUSP22        10.14     10.93      9.82     10.12     10.15
## ILMN_1651217|PDCD1LG2       7.17      7.18      7.07      7.06      7.24
## ILMN_1651221|LOC642820      7.33      7.16      7.04      7.21      7.15
## ILMN_1651228|RPS28         15.23     15.23     15.34     15.41     15.23
##                         GSM340755 GSM340756 GSM340768
## ILMN_1651199|LOC643334      6.94      6.76      6.91
## ILMN_1651209|SLC35E2        7.87      8.56      8.93
## ILMN_1651210|DUSP22         9.51     10.78      9.88
## ILMN_1651217|PDCD1LG2       6.92      7.07      7.11
## ILMN_1651221|LOC642820      7.16      7.24      7.12
## ILMN_1651228|RPS28         15.35     14.18     15.31
```

Step 4: Sanity check

- See if filtering of clinical data in R matches filtering of clinical data in excel
- See if sample ids in clinical data match sample ids in gene exp data (if they don't match it means your step 1 and/or 2 is wrong)
- Verify you see correct number of samples in baseline and comp groups
- Export the column names from gene expression data to see if it contains only probe/gene names and no other garbage

```
#See if sample ids in clinical data match sample ids in gene exp data
clinBase$GSMid == colnames(geneExpBase)
```

```
##  [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [16] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [31] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [46] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

```
clinComp$GSMid == colnames(geneExpComp)
```

```
##   [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [16] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [31] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [46] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [61] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [76] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [91] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

```
### Checking if the samples IDs baseline and comp groups are the same ---- can put these in an excel to
write.table(clinBase$GSMid , file = "sanity/Tazeen_ClinBaseIDs.tsv", sep="\t", quote = F  )
write.table(clinComp$GSMid , file = "sanity/Tazeen_ClinCompIDs.tsv", sep="\t", quote = F  )
write.table(colnames(geneExpBase) ,
            file = "sanity/Tazeen_GeneExpBaseIDs.tsv",
            sep="\t", quote = F  )
write.table(colnames(geneExpComp) ,
            file = "sanity/Tazeen_GeneExpCompIDs.tsv",
            sep="\t", quote = F  )

#Export the features from gene expression data
#Open this file and check that it contains only probe/gene names and no other garbage
write.table(rownames(geneExp),file = "Tazeen_FeatureIDs.tsv", sep="\t", quote = F  )
```

Step 5: Preparing data for T-test

- Molecular data must have features (genes in this case) as rows, and samples as columns.
- Transpose data (if needed) to obtain this
- Objects must be data frame
- Numeric data only

```
### Checking to make sure data is a numeric data frame
knitr::kable(head(geneExpBase[1:5,1:4]))
```

	GSM340547	GSM340548	GSM340549	GSM340550
ILMN_1651199|LOC643334	7.00	6.98	7.05	7.53
ILMN_1651209|SLC35E2	8.01	7.83	7.67	6.97
ILMN_1651210|DUSP22	10.12	10.25	10.44	9.90
ILMN_1651217|PDCD1LG2	6.86	6.90	6.86	7.38
ILMN_1651221|LOC642820	7.10	7.45	7.37	7.34

```
knitr::kable(head(geneExpComp[1:5,1:4]))
```

	GSM340605	GSM340606	GSM340607	GSM340608
ILMN_1651199|LOC643334	6.97	7.12	6.88	6.94
ILMN_1651209|SLC35E2	8.38	9.62	8.10	8.58
ILMN_1651210|DUSP22	9.14	7.25	10.08	10.34
ILMN_1651217|PDCD1LG2	6.92	6.89	7.11	7.08
ILMN_1651221|LOC642820	7.38	7.95	7.07	7.23

```
source("fnTTest.R")

#### Call T-test function
results1 = fnTTest(baseGroup = geneExpBase,
                   compGroup = geneExpComp,
                   testName = "Tazeen_Team4_Step1_TTest_",
                   baseGroupName = "Precancer",
                   compGroupName =  "Stage1NonInvasive",
                   folderName = "output")
```

Function for T-test

Final Step - Sub-set top differentially expressed genes

```
#Read in the T-Test results file

ttestResults <- read.csv(file = "output/Tazeen_Team4_Step1_TTest__Stage1NonInvasive_(Comp).vs._Precancer

#check to make sure p-value column is imported as numeric
#sort by p-value (just in case the results are not sorted by p-value)

ttestResultsSorted <- dplyr::arrange(ttestResults, Pvalue)

#find rows with p-value < 0.05
whichSig <- which(ttestResultsSorted$Pvalue <= 0.05)

#Short list sig results
ttestResultsSig <- ttestResultsSorted[whichSig, ] #18395 rows

### Export short listed results
```

```r
write.table(x = ttestResultsSig,
            file = "output/Tazeen_Stage1_Precancerous_Ttest_Shortlisted.csv",
            quote = F, sep = ",")

##### First column is a list of features in thsi format : ProbeID|GeneName.
#### Use string split strsplit() function to extract gene names
funcSplit <- function(featureX) {
  f1 <- unlist(strsplit(x = featureX, split = "|", fixed = TRUE))
  f2 <- f1[2]
  return(f2)
}

# Use apply() function to run the split on every row, its faster version of a loop
geneNames1 <- apply(X = as.matrix(ttestResultsSig$Feature),
                    MARGIN = 1, FUN = funcSplit)

head(geneNames1)
```

```
## [1] "MFAP4"   "CFD"     "COL16A1" "DCN"     "ACTG2"   "AEBP1"
```

```r
#print length of short listed gene names
length(geneNames1)
```

```
## [1] 18395
```

```r
### Export list of gene names
write.table(x = geneNames1,
            file = "output/Tazeen_Stage1_Precancerous_SigDiffExpressedGenes.csv",
            quote = F, sep = ",")
```