EDA_tcga

2025-06-09

```
library(foreign)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

Reading in tcga data

```
tcga = read.csv("TCGA-READ_clinical.csv",header = T,stringsAsFactors = F)
#str(tcga)
```

Checking data structure

```
str(tcga)
```

```
## 'data.frame':    172 obs. of  167 variables:
##  $ project                                             : chr  "TCGA-READ" "TCGA-READ" "TCGA-READ" "TC
##  $ submitter_id                                        : chr  "TCGA-AF-3912" "TCGA-AG-4009" "TCGA-G5-
##  $ synchronous_malignancy                              : chr  "Not Reported" "No" "No" "No" ...
##  $ ajcc_pathologic_stage                               : chr  NA "Stage I" NA "Stage I" ...
##  $ days_to_diagnosis                                   : int  NA 0 0 0 0 0 0 0 0 0 ...
##  $ created_datetime                                    : chr  "2019-04-28T10:56:20.877264-05:00" NA N
##  $ last_known_disease_status                           : chr  "not reported" NA NA NA ...
##  $ tissue_or_organ_of_origin                           : chr  "Rectosigmoid junction" "Rectum, NOS"
##  $ age_at_diagnosis                                    : int  NA 30559 20475 24775 24411 27443 24806
##  $ primary_diagnosis                                   : chr  "Adenocarcinoma, NOS" "Adenocarcinoma,
##  $ updated_datetime                                    : chr  "2025-01-08T13:19:11.073020-06:00" "202
##  $ prior_malignancy                                    : chr  "not reported" "no" "no" "no" ...
##  $ year_of_diagnosis                                   : int  NA 2009 2004 2007 2009 2009 2008 2005 2
##  $ state                                               : chr  "released" "released" "released" "relea
##  $ prior_treatment                                     : chr  "Not Reported" "No" "No" "No" ...
##  $ diagnosis_is_primary_disease                        : logi  NA TRUE TRUE TRUE TRUE TRUE ...
##  $ days_to_last_known_disease_status                   : logi  NA NA NA NA NA NA ...
##  $ ajcc_staging_system_edition                         : chr  NA "6th" NA "5th" ...
##  $ ajcc_pathologic_t                                   : chr  NA "T2" NA "T1" ...
##  $ days_to_recurrence                                  : logi  NA NA NA NA NA NA ...
##  $ morphology                                          : chr  "8140/3" "8140/3" "8140/3" "8140/3" ..
##  $ ajcc_pathologic_n                                   : chr  NA "N0" NA "N0" ...
##  $ ajcc_pathologic_m                                   : chr  NA "M0" NA "M0" ...
##  $ residual_disease                                    : chr  NA "R0" NA "R0" ...
##  $ classification_of_tumor                             : chr  "not reported" "primary" "primary" "pr
##  $ diagnosis_id                                        : chr  "82faa96d-45c6-5943-ba0f-39df276eb4b5"
##  $ icd_10_code                                         : chr  NA "C20" "C19" "C19" ...
##  $ site_of_resection_or_biopsy                         : chr  "Rectosigmoid junction" "Rectum, NOS"
##  $ tumor_grade                                         : chr  "Not Reported" NA NA NA ...
##  $ progression_or_recurrence                           : chr  "not reported" NA NA NA ...
##  $ tumor_of_origin                                     : logi  NA NA NA NA NA NA ...
##  $ irs_stage                                           : logi  NA NA NA NA NA NA ...
##  $ iss_stage                                           : logi  NA NA NA NA NA NA ...
##  $ ann_arbor_clinical_stage                            : logi  NA NA NA NA NA NA ...
##  $ enneking_msts_stage                                 : logi  NA NA NA NA NA NA ...
##  $ inrg_stage                                          : logi  NA NA NA NA NA NA ...
##  $ enneking_msts_metastasis                            : logi  NA NA NA NA NA NA ...
##  $ esophageal_columnar_dysplasia_degree                : logi  NA NA NA NA NA NA ...
##  $ cog_liver_stage                                     : logi  NA NA NA NA NA NA ...
##  $ child_pugh_classification                           : logi  NA NA NA NA NA NA ...
##  $ metastasis_at_diagnosis_site                        : logi  NA NA NA NA NA NA ...
##  $ cog_rhabdomyosarcoma_risk_group                     : logi  NA NA NA NA NA NA ...
##  $ primary_gleason_grade                               : logi  NA NA NA NA NA NA ...
##  $ inpc_grade                                          : logi  NA NA NA NA NA NA ...
##  $ irs_group                                           : logi  NA NA NA NA NA NA ...
##  $ medulloblastoma_molecular_classification            : logi  NA NA NA NA NA NA ...
##  $ wilms_tumor_histologic_subtype                      : logi  NA NA NA NA NA NA ...
##  $ weiss_assessment_score                              : logi  NA NA NA NA NA NA ...
##  $ tumor_focality                                      : logi  NA NA NA NA NA NA ...
##  $ ann_arbor_b_symptoms                                : logi  NA NA NA NA NA NA ...
##  $ cog_renal_stage                                     : logi  NA NA NA NA NA NA ...
##  $ figo_stage                                          : logi  NA NA NA NA NA NA ...
##  $ burkitt_lymphoma_clinical_variant                   : logi  NA NA NA NA NA NA ...
##  $ days_to_best_overall_response                       : logi  NA NA NA NA NA NA ...
##  $ inss_stage                                          : logi  NA NA NA NA NA NA ...
##  $ supratentorial_localization                         : logi  NA NA NA NA NA NA ...
##  $ ishak_fibrosis_score                                : logi  NA NA NA NA NA NA ...
##  $ tumor_confined_to_organ_of_origin                   : logi  NA NA NA NA NA NA ...
##  $ gleason_grade_group                                 : logi  NA NA NA NA NA NA ...
##  $ goblet_cells_columnar_mucosa_present                : logi  NA NA NA NA NA NA ...
##  $ laterality                                          : logi  NA NA NA NA NA NA ...
##  $ ajcc_clinical_stage                                 : logi  NA NA NA NA NA NA ...
##  $ cog_neuroblastoma_risk_group                        : logi  NA NA NA NA NA NA ...
##  $ metastasis_at_diagnosis                             : logi  NA NA NA NA NA NA ...
##  $ enneking_msts_tumor_site                            : logi  NA NA NA NA NA NA ...
##  $ secondary_gleason_grade                             : logi  NA NA NA NA NA NA ...
##  $ best_overall_response                               : logi  NA NA NA NA NA NA ...
##  $ ann_arbor_pathologic_stage                          : logi  NA NA NA NA NA NA ...
##  $ ann_arbor_extranodal_involvement                    : logi  NA NA NA NA NA NA ...
##  $ method_of_diagnosis                                 : logi  NA NA NA NA NA NA ...
##  $ mitosis_karyorrhexis_index                          : logi  NA NA NA NA NA NA ...
##  $ esophageal_columnar_metaplasia_present              : logi  NA NA NA NA NA NA ...
##  $ ajcc_clinical_m                                     : logi  NA NA NA NA NA NA ...
##  $ ajcc_clinical_n                                     : logi  NA NA NA NA NA NA ...
##  $ ajcc_clinical_t                                     : logi  NA NA NA NA NA NA ...
##  $ inpc_histologic_group                               : logi  NA NA NA NA NA NA ...
##  $ masaoka_stage                                       : logi  NA NA NA NA NA NA ...
##  $ micropapillary_features                             : logi  NA NA NA NA NA NA ...
##  $ igcccg_stage                                        : logi  NA NA NA NA NA NA ...
##  $ tumor_regression_grade                              : logi  NA NA NA NA NA NA ...
##  $ first_symptom_prior_to_diagnosis                    : logi  NA NA NA NA NA NA ...
##  $ enneking_msts_grade                                 : logi  NA NA NA NA NA NA ...
##  $ gastric_esophageal_junction_involvement             : logi  NA NA NA NA NA NA ...
##  $ alcohol_days_per_week                               : logi  NA NA NA NA NA NA ...
##  $ type_of_smoke_exposure                              : logi  NA NA NA NA NA NA ...
##  $ smoking_frequency                                   : logi  NA NA NA NA NA NA ...
##  $ type_of_tobacco_used                                : logi  NA NA NA NA NA NA ...
##  $ alcohol_drinks_per_day                              : logi  NA NA NA NA NA NA ...
##  $ environmental_tobacco_smoke_exposure                : logi  NA NA NA NA NA NA ...
##  $ radon_exposure                                      : logi  NA NA NA NA NA NA ...
##  $ alcohol_intensity                                   : logi  NA NA NA NA NA NA ...
##  $ pack_years_smoked                                   : logi  NA NA NA NA NA NA ...
##  $ asbestos_exposure                                   : logi  NA NA NA NA NA NA ...
##  $ cigarettes_per_day                                  : logi  NA NA NA NA NA NA ...
##  $ tobacco_smoking_quit_year                           : logi  NA NA NA NA NA NA ...
##  $ tobacco_smoking_status                              : logi  NA NA NA NA NA NA ...
##  $ alcohol_history                                     : chr  "Not Reported" NA NA NA ...
##  $ exposure_id                                         : chr  "91667b69-dd07-5114-8c71-e941d31581de"
##  $ tobacco_smoking_onset_year                          : logi  NA NA NA NA NA NA ...
##   [list output truncated]
```

Most data types that are not numeric should be categorical. So will convert the characters to categorical

```
tcga = tcga %>%
  mutate_if(is.character, as.factor)

str(tcga)
```

```
## ’data.frame':    172 obs. of  167 variables:
##  $ project                                             : Factor w/ 1 level "TCGA-READ": 1 1 1 1 1 1
##  $ submitter_id                                        : Factor w/ 172 levels "TCGA-AF-2687",..: 9 7
##  $ synchronous_malignancy                              : Factor w/ 3 levels "No","Not Reported",..:
##  $ ajcc_pathologic_stage                               : Factor w/ 11 levels "Stage I","Stage II",..
##  $ days_to_diagnosis                                   : int  NA 0 0 0 0 0 0 0 0 0 ...
##  $ created_datetime                                    : Factor w/ 2 levels "2019-04-28T10:56:20.877
##  $ last_known_disease_status                           : Factor w/ 1 level "not reported": 1 NA NA N
##  $ tissue_or_organ_of_origin                           : Factor w/ 6 levels "Colon, NOS","Connective
##  $ age_at_diagnosis                                    : int  NA 30559 20475 24775 24411 27443 24806
##  $ primary_diagnosis                                   : Factor w/ 6 levels "Adenocarcinoma in tubul
##  $ updated_datetime                                    : Factor w/ 2 levels "2025-01-08T13:19:11.073
##  $ prior_malignancy                                    : Factor w/ 3 levels "no","not reported",..:
##  $ year_of_diagnosis                                   : int  NA 2009 2004 2007 2009 2009 2008 2005
##  $ state                                               : Factor w/ 1 level "released": 1 1 1 1 1 1 1
##  $ prior_treatment                                     : Factor w/ 3 levels "No","Not Reported",..:
##  $ diagnosis_is_primary_disease                        : logi  NA TRUE TRUE TRUE TRUE TRUE ...
##  $ days_to_last_known_disease_status                   : logi  NA NA NA NA NA NA ...
##  $ ajcc_staging_system_edition                         : Factor w/ 3 levels "5th","6th","7th": NA 2
##  $ ajcc_pathologic_t                                   : Factor w/ 6 levels "T1","T2","T3",..: NA 2
##  $ days_to_recurrence                                  : logi  NA NA NA NA NA NA ...
##  $ morphology                                          : Factor w/ 6 levels "8140/3","8211/3",..: 1
##  $ ajcc_pathologic_n                                   : Factor w/ 9 levels "N0","N1","N1a",..: NA 1
##  $ ajcc_pathologic_m                                   : Factor w/ 4 levels "M0","M1","M1a",..: NA 1
##  $ residual_disease                                    : Factor w/ 4 levels "R0","R1","R2",..: NA 1
##  $ classification_of_tumor                             : Factor w/ 2 levels "not reported",..: 1 2 2
##  $ diagnosis_id                                        : Factor w/ 172 levels "0030ab60-e0e7-58ae-849
##  $ icd_10_code                                         : Factor w/ 5 levels "C18.9","C19",..: NA 3 2
##  $ site_of_resection_or_biopsy                         : Factor w/ 4 levels "Not Reported",..: 2 3 2
##  $ tumor_grade                                         : Factor w/ 1 level "Not Reported": 1 NA NA N
##  $ progression_or_recurrence                           : Factor w/ 1 level "not reported": 1 NA NA N
##  $ tumor_of_origin                                     : logi  NA NA NA NA NA NA ...
##  $ irs_stage                                           : logi  NA NA NA NA NA NA ...
##  $ iss_stage                                           : logi  NA NA NA NA NA NA ...
##  $ ann_arbor_clinical_stage                            : logi  NA NA NA NA NA NA ...
##  $ enneking_msts_stage                                 : logi  NA NA NA NA NA NA ...
##  $ inrg_stage                                          : logi  NA NA NA NA NA NA ...
##  $ enneking_msts_metastasis                            : logi  NA NA NA NA NA NA ...
##  $ esophageal_columnar_dysplasia_degree                : logi  NA NA NA NA NA NA ...
##  $ cog_liver_stage                                     : logi  NA NA NA NA NA NA ...
##  $ child_pugh_classification                           : logi  NA NA NA NA NA NA ...
##  $ metastasis_at_diagnosis_site                        : logi  NA NA NA NA NA NA ...
##  $ cog_rhabdomyosarcoma_risk_group                     : logi  NA NA NA NA NA NA ...
##  $ primary_gleason_grade                               : logi  NA NA NA NA NA NA ...
##  $ inpc_grade                                          : logi  NA NA NA NA NA NA ...
##  $ irs_group                                           : logi  NA NA NA NA NA NA ...
##  $ medulloblastoma_molecular_classification            : logi  NA NA NA NA NA NA ...
##  $ wilms_tumor_histologic_subtype                      : logi  NA NA NA NA NA NA ...
##  $ weiss_assessment_score                              : logi  NA NA NA NA NA NA ...
##  $ tumor_focality                                      : logi  NA NA NA NA NA NA ...
##  $ ann_arbor_b_symptoms                                : logi  NA NA NA NA NA NA ...
##  $ cog_renal_stage                                     : logi  NA NA NA NA NA NA ...
##  $ figo_stage                                          : logi  NA NA NA NA NA NA ...
##  $ burkitt_lymphoma_clinical_variant                   : logi  NA NA NA NA NA NA ...
##  $ days_to_best_overall_response                       : logi  NA NA NA NA NA NA ...
##  $ inss_stage                                          : logi  NA NA NA NA NA NA ...
##  $ supratentorial_localization                         : logi  NA NA NA NA NA NA ...
##  $ ishak_fibrosis_score                                : logi  NA NA NA NA NA NA ...
##  $ tumor_confined_to_organ_of_origin                   : logi  NA NA NA NA NA NA ...
##  $ gleason_grade_group                                 : logi  NA NA NA NA NA NA ...
##  $ goblet_cells_columnar_mucosa_present                : logi  NA NA NA NA NA NA ...
##  $ laterality                                          : logi  NA NA NA NA NA NA ...
##  $ ajcc_clinical_stage                                 : logi  NA NA NA NA NA NA ...
##  $ cog_neuroblastoma_risk_group                        : logi  NA NA NA NA NA NA ...
##  $ metastasis_at_diagnosis                             : logi  NA NA NA NA NA NA ...
##  $ enneking_msts_tumor_site                            : logi  NA NA NA NA NA NA ...
##  $ secondary_gleason_grade                             : logi  NA NA NA NA NA NA ...
##  $ best_overall_response                               : logi  NA NA NA NA NA NA ...
##  $ ann_arbor_pathologic_stage                          : logi  NA NA NA NA NA NA ...
##  $ ann_arbor_extranodal_involvement                    : logi  NA NA NA NA NA NA ...
##  $ method_of_diagnosis                                 : logi  NA NA NA NA NA NA ...
##  $ mitosis_karyorrhexis_index                          : logi  NA NA NA NA NA NA ...
##  $ esophageal_columnar_metaplasia_present              : logi  NA NA NA NA NA NA ...
##  $ ajcc_clinical_m                                     : logi  NA NA NA NA NA NA ...
##  $ ajcc_clinical_n                                     : logi  NA NA NA NA NA NA ...
##  $ ajcc_clinical_t                                     : logi  NA NA NA NA NA NA ...
##  $ inpc_histologic_group                               : logi  NA NA NA NA NA NA ...
##  $ masaoka_stage                                       : logi  NA NA NA NA NA NA ...
##  $ micropapillary_features                             : logi  NA NA NA NA NA NA ...
##  $ igcccg_stage                                        : logi  NA NA NA NA NA NA ...
##  $ tumor_regression_grade                              : logi  NA NA NA NA NA NA ...
##  $ first_symptom_prior_to_diagnosis                    : logi  NA NA NA NA NA NA ...
##  $ enneking_msts_grade                                 : logi  NA NA NA NA NA NA ...
##  $ gastric_esophageal_junction_involvement             : logi  NA NA NA NA NA NA ...
##  $ alcohol_days_per_week                               : logi  NA NA NA NA NA NA ...
##  $ type_of_smoke_exposure                              : logi  NA NA NA NA NA NA ...
##  $ smoking_frequency                                   : logi  NA NA NA NA NA NA ...
##  $ type_of_tobacco_used                                : logi  NA NA NA NA NA NA ...
##  $ alcohol_drinks_per_day                              : logi  NA NA NA NA NA NA ...
##  $ environmental_tobacco_smoke_exposure                : logi  NA NA NA NA NA NA ...
##  $ radon_exposure                                      : logi  NA NA NA NA NA NA ...
##  $ alcohol_intensity                                   : logi  NA NA NA NA NA NA ...
##  $ pack_years_smoked                                   : logi  NA NA NA NA NA NA ...
##  $ asbestos_exposure                                   : logi  NA NA NA NA NA NA ...
##  $ cigarettes_per_day                                  : logi  NA NA NA NA NA NA ...
##  $ tobacco_smoking_quit_year                           : logi  NA NA NA NA NA NA ...
##  $ tobacco_smoking_status                              : logi  NA NA NA NA NA NA ...
##  $ alcohol_history                                     : Factor w/ 1 level "Not Reported": 1 NA NA N
##  $ exposure_id                                         : Factor w/ 1 level "91667b69-dd07-5114-8c71-
##  $ tobacco_smoking_onset_year                          : logi  NA NA NA NA NA NA ...
##   [list output truncated]
```

Drop extra factors from variables that are categorical

Count the NA values per sample

```
na_count = apply(tcga,2,function(x){sum(is.na(x))})
na_count_o = sort(na_count,decreasing = TRUE)
na_count_o[1:10]
```

```
##    days_to_last_known_disease_status                   days_to_recurrence
##                                  172                                  172
##                      tumor_of_origin                            irs_stage
##                                  172                                  172
##                            iss_stage             ann_arbor_clinical_stage
##                                  172                                  172
##                  enneking_msts_stage                           inrg_stage
##                                  172                                  172
##             enneking_msts_metastasis esophageal_columnar_dysplasia_degree
##                                  172                                  172
```

```
varibles_half_na = names(na_count_o)[na_count_o >= max(na_count_o)/2]
`%ni%` = Negate(`%in%`)
tcga_use = tcga[,colnames(tcga) %ni% varibles_half_na]
```

Using mice for missing data analysis among remaining variables

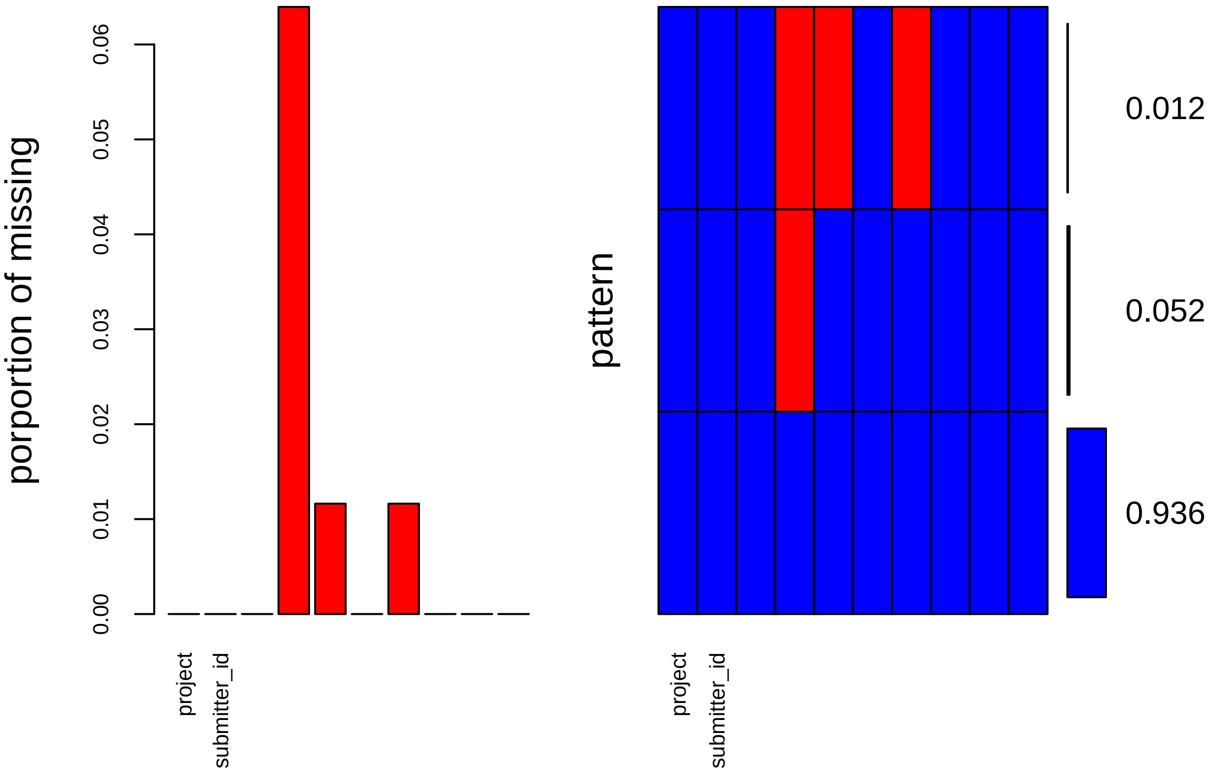
```
library(mice)
```

```
##
## Attaching package: 'mice'

## The following object is masked from 'package:stats':
##
##     filter

## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```
library(VIM)
```

```
## Loading required package: colorspace

## Loading required package: grid

## VIM is ready to use.

## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues

##
## Attaching package: 'VIM'

## The following object is masked from 'package:datasets':
##
##     sleep
```

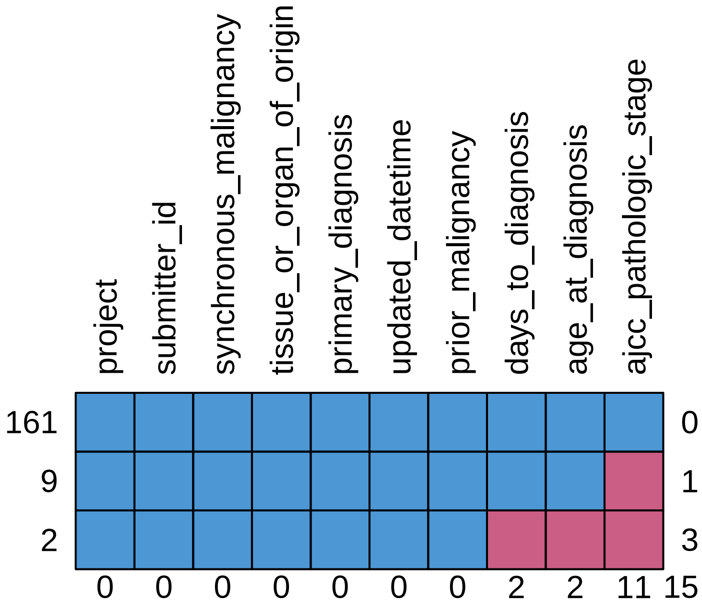
```
## red is missing and blue is not missing
missing_val_plot = aggr(tcga_use[,c(1:10)], col=c("blue","red"),
                   numbers=TRUE,sortVard=TRUE,
                     labels = names(tcga_use),cex.axis=.7,
                       gap=3,ylab=c("porportion of missing","pattern"))
```



```
#md.pattern(tcga_use,rotate.names = TRUE)
```

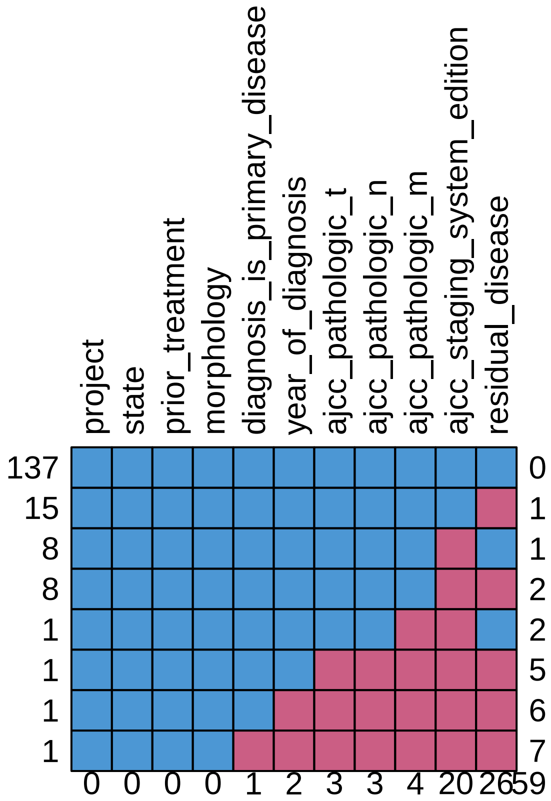
Display missing-data patterns

```
md.pattern(tcga_use[,c(1:10)],rotate.names = TRUE) ## difficul tot understand
```



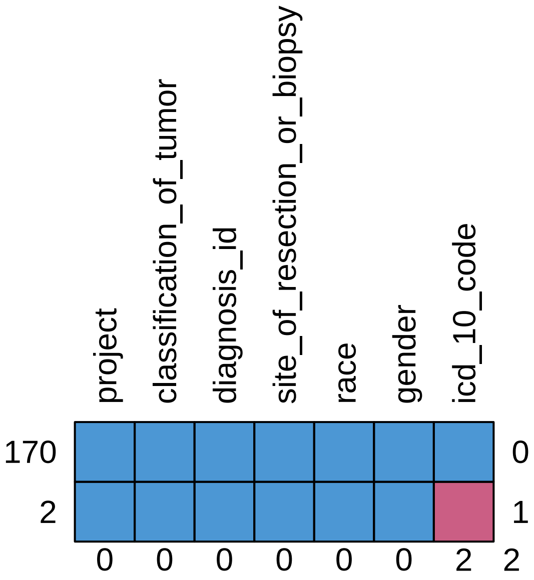
```
##     project submitter_id synchronous_malignancy tissue_or_organ_of_origin
## 161       1            1                      1                         1
## 9         1            1                      1                         1
## 2         1            1                      1                         1
##           0            0                      0                         0
##     primary_diagnosis updated_datetime prior_malignancy days_to_diagnosis
## 161                 1                1                1                 1
## 9                   1                1                1                 1
## 2                   1                1                1                 0
##                     0                0                0                 2
##     age_at_diagnosis ajcc_pathologic_stage   
## 161                1                     1  0
## 9                  1                     0  1
## 2                  0                     0  3
##                    2                    11 15
```

```
md.pattern(tcga_use[,c(1,11:20)],rotate.names = TRUE)
```



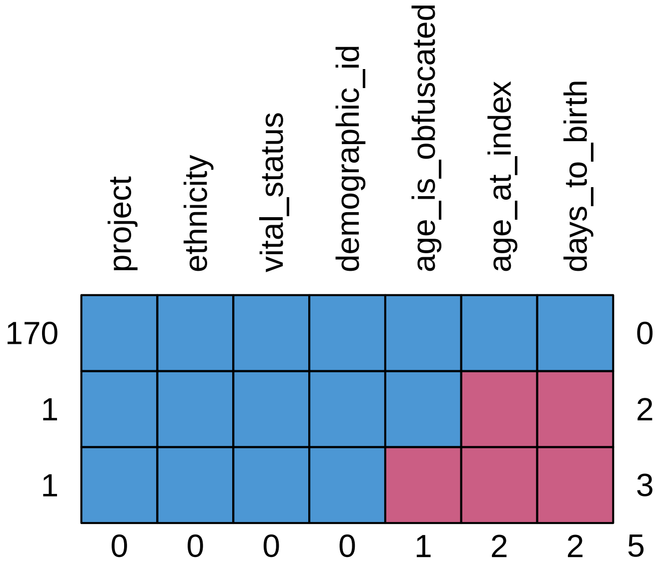
```
##     project state prior_treatment morphology diagnosis_is_primary_disease
## 137       1     1               1          1                            1
## 15        1     1               1          1                            1
## 8         1     1               1          1                            1
## 8         1     1               1          1                            1
## 1         1     1               1          1                            1
## 1         1     1               1          1                            1
## 1         1     1               1          1                            1
## 1         1     1               1          1                            0
##           0     0               0          0                            1
##     year_of_diagnosis ajcc_pathologic_t ajcc_pathologic_n ajcc_pathologic_m
## 137                 1                 1                 1                 1
## 15                  1                 1                 1                 1
## 8                   1                 1                 1                 1
## 8                   1                 1                 1                 1
## 1                   1                 1                 1                 0
## 1                   1                 0                 0                 0
## 1                   0                 0                 0                 0
## 1                   0                 0                 0                 0
##                     2                 3                 3                 4
##     ajcc_staging_system_edition residual_disease
## 137                           1                1  0
## 15                            1                0  1
## 8                             0                1  1
## 8                             0                0  2
## 1                             0                1  2
## 1                             0                0  5
## 1                             0                0  6
## 1                             0                0  7
##                              20               26 59
```

```
md.pattern(tcga_use[,c(1,21:26)],rotate.names = TRUE)
```



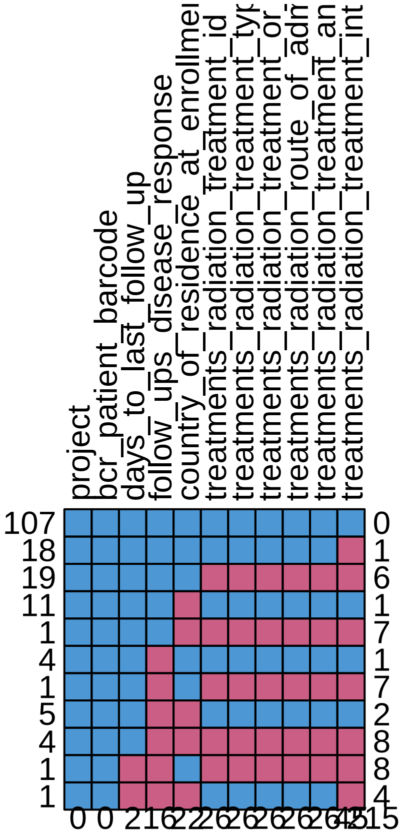
```
##     project classification_of_tumor diagnosis_id site_of_resection_or_biopsy
## 170       1                       1            1                           1
## 2         1                       1            1                           1
##           0                       0            0                           0
##     race gender icd_10_code
## 170    1      1           1 0
## 2      1      1           0 1
##        0      0           2 2
```

```
md.pattern(tcga_use[,c(1,27:32)],rotate.names = TRUE)
```



```
##     project ethnicity vital_status demographic_id age_is_obfuscated
## 170       1         1            1              1                 1
## 1         1         1            1              1                 1
## 1         1         1            1              1                 0
##           0         0            0              0                 1
##     age_at_index days_to_birth
## 170            1             1 0
## 1              0             0 2
## 1              0             0 3
##                2             2 5
```

```
md.pattern(tcga_use[,c(1,33:42)],rotate.names = TRUE)
```



```
##     project bcr_patient_barcode days_to_last_follow_up
## 107       1                   1                      1
## 18        1                   1                      1
## 19        1                   1                      1
## 11        1                   1                      1
## 1         1                   1                      1
## 4         1                   1                      1
## 1         1                   1                      1
## 5         1                   1                      1
## 4         1                   1                      1
## 1         1                   1                      0
## 1         1                   1                      0
##           0                   0                      2
##     follow_ups_disease_response country_of_residence_at_enrollment
## 107                           1                                  1
## 18                            1                                  1
## 19                            1                                  1
## 11                            1                                  0
## 1                             1                                  0
## 4                             0                                  1
## 1                             0                                  1
## 5                             0                                  0
## 4                             0                                  0
## 1                             0                                  1
## 1                             0                                  0
##                              16                                 22
##     treatments_radiation_treatment_id treatments_radiation_treatment_type
## 107                                 1                                   1
## 18                                  1                                   1
## 19                                  0                                   0
## 11                                  1                                   1
## 1                                   0                                   0
## 4                                   1                                   1
## 1                                   0                                   0
## 5                                   1                                   1
## 4                                   0                                   0
## 1                                   0                                   0
## 1                                   1                                   1
##                                    26                                  26
##     treatments_radiation_treatment_or_therapy
## 107                                         1
## 18                                          1
## 19                                          0
## 11                                          1
## 1                                           0
## 4                                           1
## 1                                           0
## 5                                           1
## 4                                           0
## 1                                           0
## 1                                           1
##                                            26
##     treatments_radiation_route_of_administration
## 107                                            1
## 18                                             1
## 19                                             0
## 11                                             1
## 1                                              0
## 4                                              1
## 1                                              0
## 5                                              1
## 4                                              0
## 1                                              0
## 1                                              1
##                                               26
##     treatments_radiation_treatment_anatomic_sites
## 107                                             1
## 18                                              1
## 19                                              0
## 11                                              1
## 1                                               0
## 4                                               1
## 1                                               0
## 5                                               1
## 4                                               0
## 1                                               0
## 1                                               1
##                                                26
##     treatments_radiation_treatment_intent_type
## 107                                          1   0
## 18                                           0   1
## 19                                           0   6
## 11                                           1   1
## 1                                            0   7
## 4                                            1   1
## 1                                            0   7
## 5                                            1   2
## 4                                            0   8
## 1                                            0   8
## 1                                            0   4
##                                             45 215
```

What could be some of the most interesting clinical variables? tissue__or__organ__of__origin - Rectum, NOS or Rectosigmoid junction primary__diagnosis - Adenocarcinoma, NOS, ajcc__pathologic__t ajcc__pathologic__m residual__disease site__of__resection__or__biopsy race gender ethnicity vital__status

Majority of the Adenocarcinoma, NOS are ones are either Rectosigmoid junction or Rectum, NOS. Compare Most of the Rectosigmoid junction and reactum NOS are T3 in ajcc__pathologic__t Most of the Rectosigmoid junction and reactum NOS are M0 in ajcc__pathologic__m

```
count_by_race_tissue_origin=tcga_use %>%
  group_by(tissue_or_organ_of_origin, race) %>%
  summarise(n= n())
```

```
## `summarise()` has grouped output by 'tissue_or_organ_of_origin'. You can
## override using the `.groups` argument.
```

```
chisq.test(count_by_race_tissue_origin$race,count_by_race_tissue_origin$tissue_or_organ_of_origin)## No
```

```
## Warning in chisq.test(count_by_race_tissue_origin$race,
## count_by_race_tissue_origin$tissue_or_organ_of_origin): Chi-squared
## approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  count_by_race_tissue_origin$race and count_by_race_tissue_origin$tissue_or_organ_of_origin
## X-squared = 8.6472, df = 15, p-value = 0.8952
```

```
count_by_gender_tissue_origin=tcga_use %>%
  group_by(tissue_or_organ_of_origin, gender) %>%
  summarise(n= n())
```

```
## `summarise()` has grouped output by 'tissue_or_organ_of_origin'. You can
## override using the `.groups` argument.
```

```
chisq.test(count_by_gender_tissue_origin$gender,count_by_gender_tissue_origin$tissue_or_organ_of_origin
```

```
## Warning in chisq.test(count_by_gender_tissue_origin$gender,
## count_by_gender_tissue_origin$tissue_or_organ_of_origin): Chi-squared
## approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  count_by_gender_tissue_origin$gender and count_by_gender_tissue_origin$tissue_or_organ_of_orig
## X-squared = 8.3333, df = 10, p-value = 0.5963
```

How is vital status affected by tumor grading or biopsy?

```
vital_status_by_ajcc_t = tcga_use %>%
  group_by(vital_status,ajcc_pathologic_t) %>%
  summarise(n = n())
```

```
## 'summarise()' has grouped output by 'vital_status'. You can override using the
## '.groups' argument.
```

```
chisq.test(vital_status_by_ajcc_t$vital_status,vital_status_by_ajcc_t$ajcc_pathologic_t)## No asso
```

```
## Warning in chisq.test(vital_status_by_ajcc_t$vital_status,
## vital_status_by_ajcc_t$ajcc_pathologic_t): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  vital_status_by_ajcc_t$vital_status and vital_status_by_ajcc_t$ajcc_pathologic_t
## X-squared = 2, df = 5, p-value = 0.8491
```

```
vital_status_by_ajcc_m = tcga_use %>%
  group_by(vital_status,ajcc_pathologic_m) %>%
  summarise(n = n())
```

```
## 'summarise()' has grouped output by 'vital_status'. You can override using the
## '.groups' argument.
```

```
chisq.test(vital_status_by_ajcc_m$vital_status,vital_status_by_ajcc_m$ajcc_pathologic_m)## No asso
```

```
## Warning in chisq.test(vital_status_by_ajcc_m$vital_status,
## vital_status_by_ajcc_m$ajcc_pathologic_m): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  vital_status_by_ajcc_m$vital_status and vital_status_by_ajcc_m$ajcc_pathologic_m
## X-squared = 0.875, df = 3, p-value = 0.8315
```

```
vital_status_by_site_rejection_biopsy = tcga_use %>%
  group_by(vital_status,site_of_resection_or_biopsy) %>%
  summarise(n = n())
```

```
## 'summarise()' has grouped output by 'vital_status'. You can override using the
## '.groups' argument.
```

```r
chisq.test(vital_status_by_site_rejection_biopsy$vital_status,vital_status_by_site_rejection_biopsy$sit
```

```
## Warning in chisq.test(vital_status_by_site_rejection_biopsy$vital_status, :
## Chi-squared approximation may be incorrect

##
##  Pearson's Chi-squared test
##
## data:  vital_status_by_site_rejection_biopsy$vital_status and vital_status_by_site_rejection_biopsy$s
## X-squared = 1.6667, df = 6, p-value = 0.9477
```

```r
vital_status_by_primary_diagnosis = tcga_use %>%
  group_by(vital_status,primary_diagnosis) %>%
  summarise(n = n())
```

```
## `summarise()` has grouped output by 'vital_status'. You can override using the
## `.groups` argument.
```

```r
chisq.test(vital_status_by_primary_diagnosis$vital_status,vital_status_by_primary_diagnosis$primary_dia
```

```
## Warning in chisq.test(vital_status_by_primary_diagnosis$vital_status,
## vital_status_by_primary_diagnosis$primary_diagnosis): Chi-squared approximation
## may be incorrect

##
##  Pearson's Chi-squared test
##
## data:  vital_status_by_primary_diagnosis$vital_status and vital_status_by_primary_diagnosis$primary_
## X-squared = 7.65, df = 10, p-value = 0.663
```

Does missingness of important variables associate with outcome?

```r
missing_ajcc_t <- as.factor(is.na(tcga_use$ajcc_pathologic_t))

table(tcga_use$vital_status,missing_ajcc_t)
```

```
##                 missing_ajcc_t
##                  FALSE TRUE
##   Alive            142    0
##   Dead              27    1
##   Not Reported       0    2
```

```r
fisher.test(tcga_use$vital_status,missing_ajcc_t)## Significant
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tcga_use$vital_status and missing_ajcc_t
## p-value = 3.36e-05
## alternative hypothesis: two.sided
```

```
missing_ajcc_m <- as.factor(is.na(tcga_use$ajcc_pathologic_m))

table(tcga_use$vital_status,missing_ajcc_m)
```

```
##               missing_ajcc_m
##                FALSE TRUE
##   Alive          141    1
##   Dead            27    1
##   Not Reported     0    2
```

```
fisher.test(tcga_use$vital_status,missing_ajcc_m)## Significant
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tcga_use$vital_status and missing_ajcc_m
## p-value = 0.0001237
## alternative hypothesis: two.sided
```

```
missing_site_resection_biopsy <- as.factor(is.na(tcga_use$site_of_resection_or_biopsy))

table(tcga_use$vital_status,missing_site_resection_biopsy)
```

```
##               missing_site_resection_biopsy
##                FALSE
##   Alive          142
##   Dead            28
##   Not Reported     2
```

```
#fisher.test(tcga_use$vital_status,missing_site_resection_biopsy)## one level in missing_site_resection
```

Will adjust for missing_ajcc_t, missing_ajcc_m, gender, age, race and ethnicity while DE for gene exp and vital_status.

Reading in gene expression data

```
gene_exp = read.table(file="TCGA-READ_RNASeq_count_Data.txt",header = T,sep="\t",stringsAsFactors = F)

sample_ids = colnames(gene_exp)
```

Cleaning_up_the_sample_IDs

```
sample_ids_edited=str_replace_all(sample_ids,"\\.","_")
colnames(gene_exp) = sample_ids_edited

tcga_use$bcr_patient_barcode_mod = str_replace_all(tcga_use$bcr_patient_barcode,"-","_")
```

Now, going to take the order of the samples of the gene exp matrix and order the metadata in the same order

```
gene_exp_ids = colnames(gene_exp)

gene_exp_ids = as.data.frame(gene_exp_ids)

library(stringr)
gene_exp_ids <- str_split_fixed(gene_exp_ids$gene_exp_ids, '_',7)
gene_exp_ids = as.data.frame(gene_exp_ids)
gene_exp_ids$New_RNA_id=paste(gene_exp_ids$V1,gene_exp_ids$V2,gene_exp_ids$V3,sep="_")

length(intersect(gene_exp_ids$New_RNA_id,tcga_use$bcr_patient_barcode_mod)) ## 167
```

```
## [1] 167
```

```
gene_exp_ids$Original_ID = colnames(gene_exp)


colnames(gene_exp_ids)[1:7]=c("Project","TSS","Participant","Sample_Vial","Portion_Analyte","Plate","Cer
```

Going ot restrict the analysis to only tumor samples of vial A by filtering on the 4th field of the name whereby 01-09 for tumor and 10-19 are normal

```
table(gene_exp_ids$Sample)
```

```
##
## 01A 01B 01C 02A 11A
## 163   2   1   1  10
```

```
samples_use = subset(gene_exp_ids,gene_exp_ids$Sample_Vial=="01A")
```

Now going to use the order of the samples_use df to order the samples and extract the gene expression matrix for corresponding_Samples

```
samples_use_clinical = left_join(samples_use,tcga_use,by=c("New_RNA_id"="bcr_patient_barcode_mod"))

gene_exp_use = gene_exp[,samples_use_clinical$Original_ID]
```

Some variables have to still be converted to factor for the model

```
samples_use_clinical$Portion_Analyte = as.factor(samples_use_clinical$Portion_Analyte)

samples_use_clinical$age_diag_c = scale(samples_use_clinical$age_at_diagnosis,center = TRUE)
```

Now checking to see which of the adjusting variables have a lot of NA. Will drop it.

```
NA_count = apply(samples_use_clinical,2,function(x){sum(is.na(x))})
NA_count
```

```
##                                       Project
##                                             0
##                                            TSS
##                                              0
##                                    Participant
##                                              0
##                                    Sample_Vial
##                                              0
##                                Portion_Analyte
##                                              0
##                                          Plate
##                                              0
##                                         Center
##                                              0
##                                     New_RNA_id
##                                              0
##                                    Original_ID
##                                              0
##                                        project
##                                              0
##                                   submitter_id
##                                              0
##                         synchronous_malignancy
##                                              0
##                          ajcc_pathologic_stage
##                                             10
##                              days_to_diagnosis
##                                              1
##                      tissue_or_organ_of_origin
##                                              0
##                               age_at_diagnosis
##                                              1
##                              primary_diagnosis
##                                              0
##                               updated_datetime
##                                              0
##                               prior_malignancy
##                                              0
##                              year_of_diagnosis
##                                              1
##                                          state
##                                              0
##                                prior_treatment
##                                              0
##                   diagnosis_is_primary_disease
##                                              0
##                    ajcc_staging_system_edition
##                                             17
##                              ajcc_pathologic_t
##                                              2
##                                     morphology
##                                              0
##                              ajcc_pathologic_n
##                                              2
##                              ajcc_pathologic_m
##                                              3
##                              residual_disease 
##                                            22 
##                       classification_of_tumor 
##                                             0 
##                                  diagnosis_id 
##                                             0 
##                                   icd_10_code 
##                                             1 
##                   site_of_resection_or_biopsy 
##                                             0 
##                                          race 
##                                             0 
##                                        gender 
##                                             0 
##                                     ethnicity 
##                                             0 
##                                  vital_status 
##                                             0 
##                                  age_at_index 
##                                             1 
##                                 days_to_birth 
##                                             1 
##                                demographic_id 
##                                             0 
##                             age_is_obfuscated 
##                                             0 
##            country_of_residence_at_enrollment 
##                                            18 
##                        days_to_last_follow_up 
##                                             1 
##                   follow_ups_disease_response 
##                                            14 
##             treatments_radiation_treatment_id 
##                                            24 
##           treatments_radiation_treatment_type 
##                                            24 
##     treatments_radiation_treatment_or_therapy 
##                                            24 
##    treatments_radiation_treatment_intent_type 
##                                            42 
##  treatments_radiation_route_of_administration 
##                                            24 
## treatments_radiation_treatment_anatomic_sites 
##                                            24 
##                           bcr_patient_barcode 
##                                             0 
##                                    age_diag_c 
##                                             1
```

Will adjust for missing_ajcc_t, missing_ajcc_m, gender, age, race and ethnicity, as well as portion and analyte while DE for vital_status

```
samples_use_clinical_final = samples_use_clinical[,c("Original_ID",
  "vital_status","ajcc_pathologic_t","ajcc_pathologic_m","gender","age_diag_c","race","ethnicity","Port
samples_use_clinical_final = samples_use_clinical_final[complete.cases(samples_use_clinical_final),]

gene_exp_use = gene_exp[,samples_use_clinical_final$Original_ID]
```

Checking that the order of the samples in gene exp matrix and metdata are the same

```
table(samples_use_clinical_final$Original_ID == colnames(gene_exp_use))
```

```
##
## TRUE
##  160
```

Now carrying out DE analysis using DESeq2

```
library(DESeq2)
```

```
## Loading required package: S4Vectors
```

```
## Loading required package: stats4
```

```
## Loading required package: BiocGenerics
```

```
## Loading required package: generics
```

```
##
## Attaching package: 'generics'
```

```
## The following object is masked from 'package:lubridate':
##
##     as.difftime
```

```
## The following object is masked from 'package:dplyr':
##
##     explain
```

```
## The following objects are masked from 'package:base':
##
##     as.difftime, as.factor, as.ordered, intersect, is.element, setdiff,
##     setequal, union
```

```
##
## Attaching package: 'BiocGenerics'
```

```
## The following objects are masked from 'package:mice':
##
##     cbind, rbind
## The following object is masked from 'package:dplyr':
##
##     combine

## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs

## The following objects are masked from 'package:base':
##
##     anyDuplicated, aperm, append, as.data.frame, basename, cbind,
##     colnames, dirname, do.call, duplicated, eval, evalq, Filter, Find,
##     get, grep, grepl, is.unsorted, lapply, Map, mapply, match, mget,
##     order, paste, pmax, pmax.int, pmin, pmin.int, Position, rank,
##     rbind, Reduce, rownames, sapply, saveRDS, table, tapply, unique,
##     unsplit, which.max, which.min

##
## Attaching package: 'S4Vectors'

## The following objects are masked from 'package:lubridate':
##
##     second, second<-

## The following objects are masked from 'package:dplyr':
##
##     first, rename

## The following object is masked from 'package:tidyr':
##
##     expand

## The following object is masked from 'package:utils':
##
##     findMatches

## The following objects are masked from 'package:base':
##
##     expand.grid, I, unname

## Loading required package: IRanges

##
## Attaching package: 'IRanges'

## The following object is masked from 'package:lubridate':
##
##     %within%

## The following objects are masked from 'package:dplyr':
##
##     collapse, desc, slice
## The following object is masked from ’package:purrr’:
##
##     reduce
```

```
## Loading required package: GenomicRanges
```

```
## Loading required package: GenomeInfoDb
```

```
## Loading required package: SummarizedExperiment
```

```
## Loading required package: MatrixGenerics
```

```
## Loading required package: matrixStats
```

```
##
## Attaching package: ’matrixStats’
```

```
## The following object is masked from ’package:dplyr’:
##
##     count
```

```
##
## Attaching package: ’MatrixGenerics’
```

```
## The following objects are masked from ’package:matrixStats’:
##
##     colAlls, colAnyNAs, colAnys, colAvgsPerRowSet, colCollapse,
##     colCounts, colCummaxs, colCummins, colCumprods, colCumsums,
##     colDiffs, colIQRDiffs, colIQRs, colLogSumExps, colMadDiffs,
##     colMads, colMaxs, colMeans2, colMedians, colMins, colOrderStats,
##     colProds, colQuantiles, colRanges, colRanks, colSdDiffs, colSds,
##     colSums2, colTabulates, colVarDiffs, colVars, colWeightedMads,
##     colWeightedMeans, colWeightedMedians, colWeightedSds,
##     colWeightedVars, rowAlls, rowAnyNAs, rowAnys, rowAvgsPerColSet,
##     rowCollapse, rowCounts, rowCummaxs, rowCummins, rowCumprods,
##     rowCumsums, rowDiffs, rowIQRDiffs, rowIQRs, rowLogSumExps,
##     rowMadDiffs, rowMads, rowMaxs, rowMeans2, rowMedians, rowMins,
##     rowOrderStats, rowProds, rowQuantiles, rowRanges, rowRanks,
##     rowSdDiffs, rowSds, rowSums2, rowTabulates, rowVarDiffs, rowVars,
##     rowWeightedMads, rowWeightedMeans, rowWeightedMedians,
##     rowWeightedSds, rowWeightedVars
```

```
## Loading required package: Biobase
```

```
## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     ’browseVignettes()’. To cite Bioconductor, see
##     ’citation("Biobase")’, and for packages ’citation("pkgname")’.
```

```
##
## Attaching package: ’Biobase’
## The following object is masked from 'package:MatrixGenerics':
##
##     rowMedians

## The following objects are masked from 'package:matrixStats':
##
##     anyMissing, rowMedians
```

```
library(edgeR)
```

```
## Loading required package: limma

##
## Attaching package: 'limma'

## The following object is masked from 'package:DESeq2':
##
##     plotMA

## The following object is masked from 'package:BiocGenerics':
##
##     plotMA
```

```
dds = DESeqDataSetFromMatrix(gene_exp_use,colData = samples_use_clinical_final,
                             design = ~ vital_status + ajcc_pathologic_t + ajcc_pathologic_m + gender +
```

```
## factor levels were dropped which had no samples

##   Note: levels of factors in the design contain characters other than
##   letters, numbers, '_' and '.'. It is recommended (but not required) to use
##   only letters, numbers, and delimiters '_' or '.', as these are safe characters
##   for column names in R. [This is a message, not a warning or an error]
```

```
keep = rowSums( cpm(dds) >2) >= 5 ## dds >2 in at least 5 of sequenced sampels

table(keep)
```

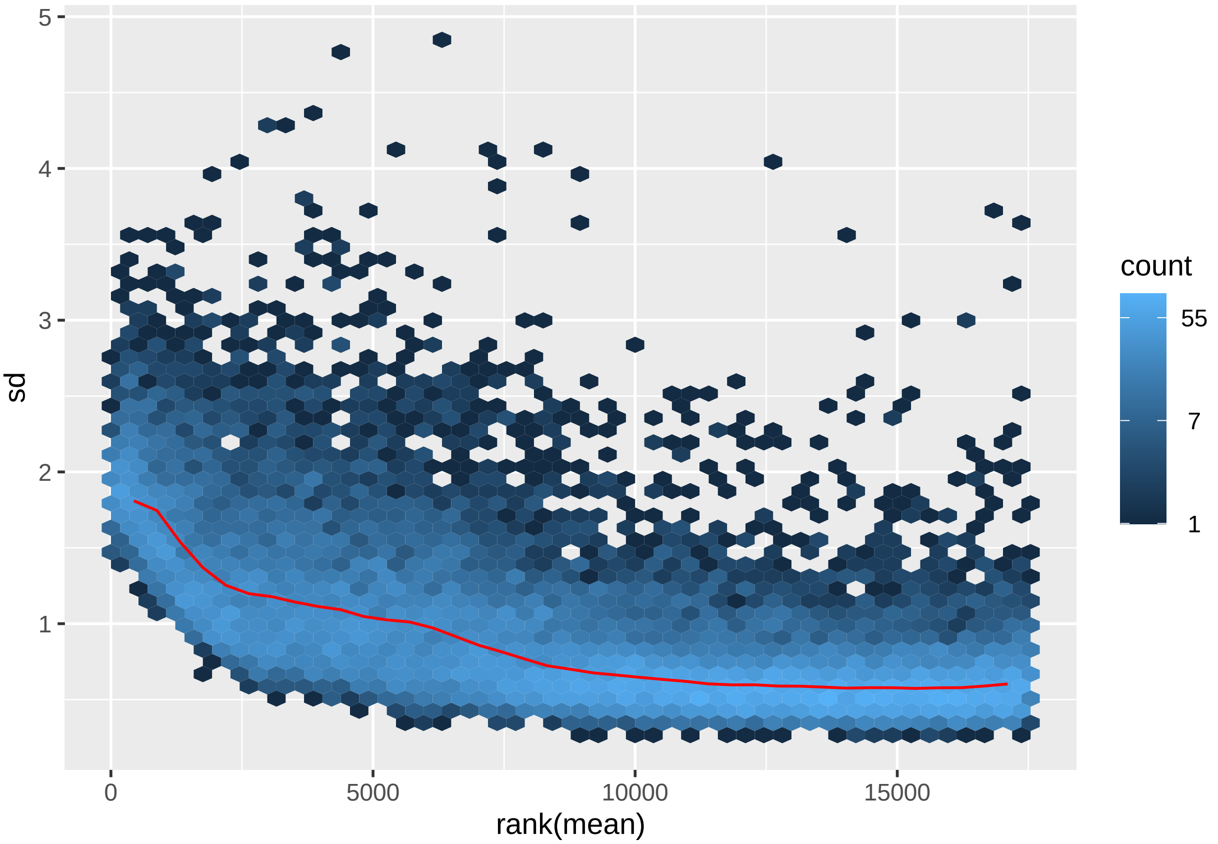
```
## keep
## FALSE  TRUE
## 43116 17544
```

```
dds = dds[keep,]
```

```
ntd <- normTransform(dds)
```

```
##   Note: levels of factors in the design contain characters other than
##   letters, numbers, '_' and '.'. It is recommended (but not required) to use
##   only letters, numbers, and delimiters '_' or '.', as these are safe characters
##   for column names in R. [This is a message, not a warning or an error]
```

```r
library("vsn")
meanSdPlot(assay(ntd))
```



```r
vsd <- vst(dds, blind = FALSE)
```

```
##   Note: levels of factors in the design contain characters other than
##   letters, numbers, '_' and '.'. It is recommended (but not required) to use
##   only letters, numbers, and delimiters '_' or '.', as these are safe characters
##   for column names in R. [This is a message, not a warning or an error]
```

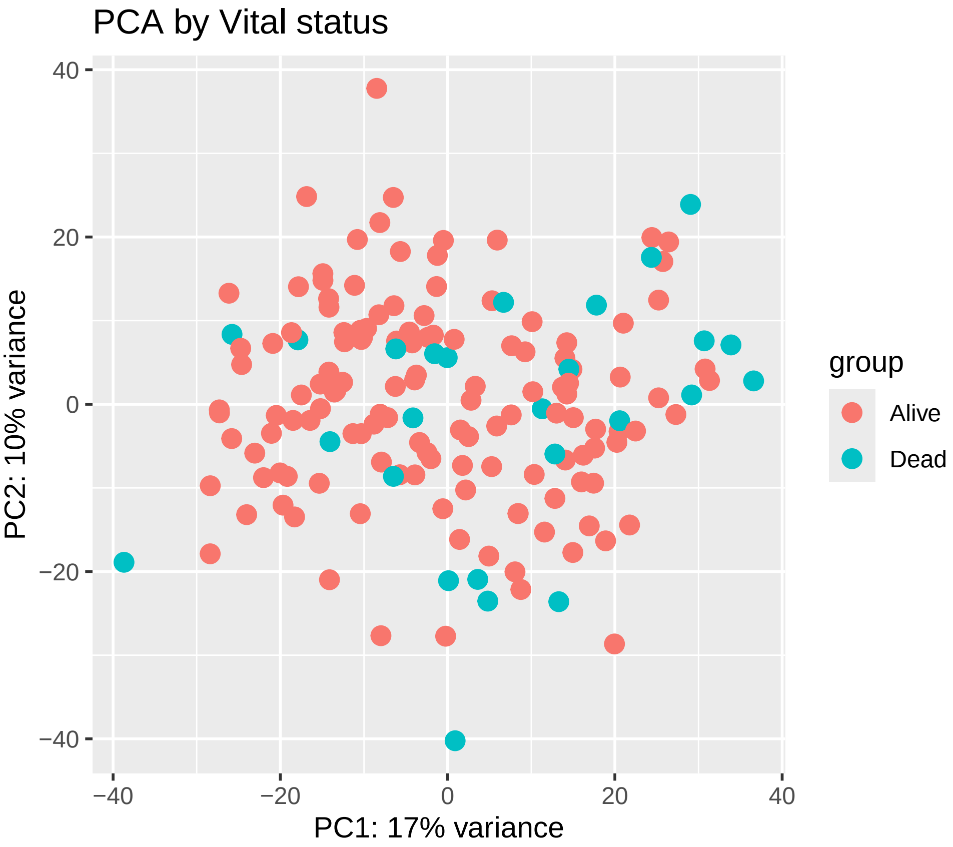
```r
dds <- estimateSizeFactors(dds)
```

```
##   Note: levels of factors in the design contain characters other than
##   letters, numbers, '_' and '.'. It is recommended (but not required) to use
##   only letters, numbers, and delimiters '_' or '.', as these are safe characters
##   for column names in R. [This is a message, not a warning or an error]
```

Now estimating the size factor and generating the PCA plots

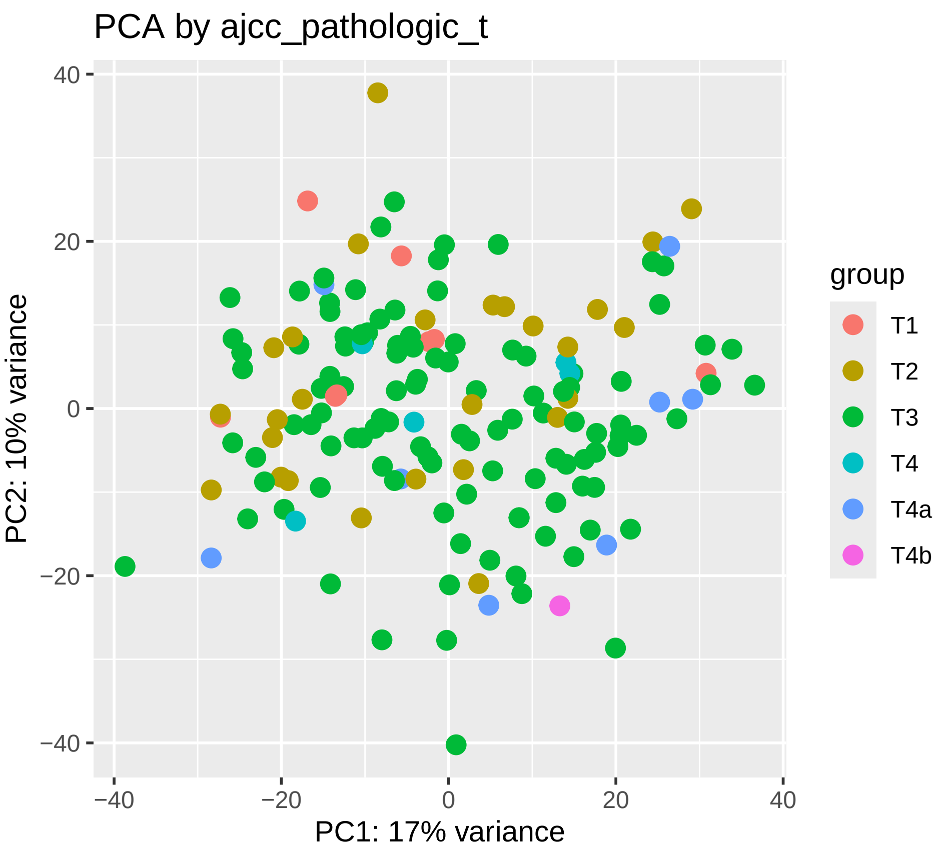
```r
plotPCA(vsd, intgroup = c("vital_status")) +ggtitle("PCA by Vital status")
```

```
## using ntop=500 top features by variance
```



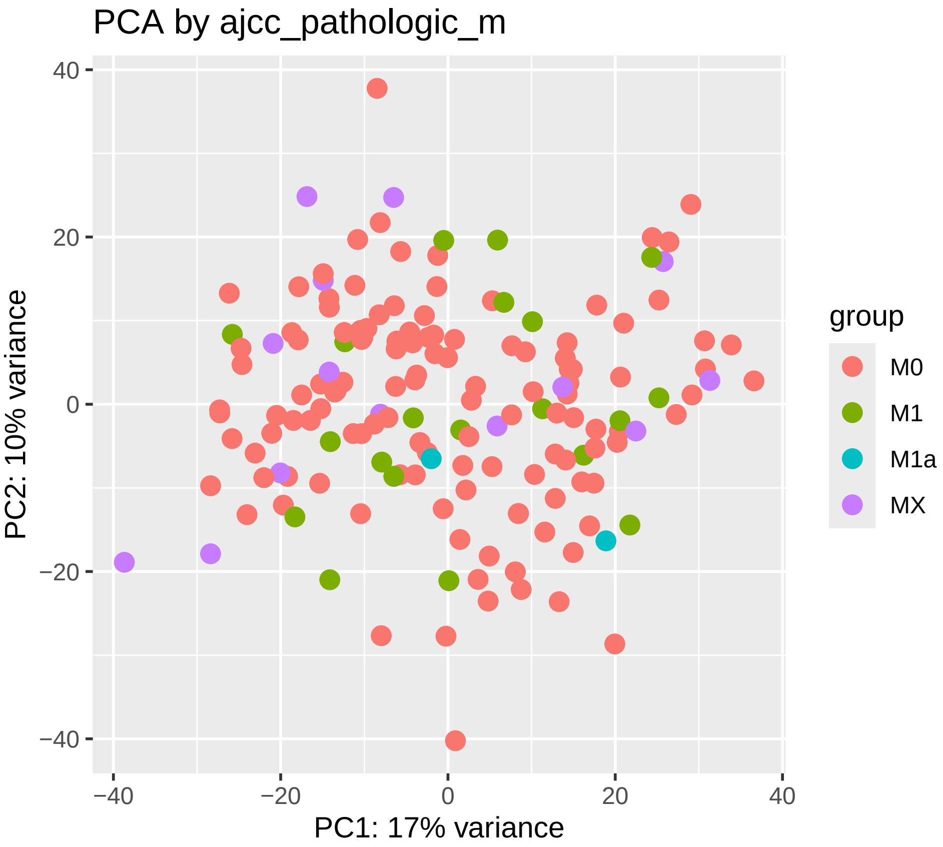
```
plotPCA(vsd, intgroup = c("ajcc_pathologic_t")) +ggtitle("PCA by ajcc_pathologic_t")
```

```
## using ntop=500 top features by variance
```



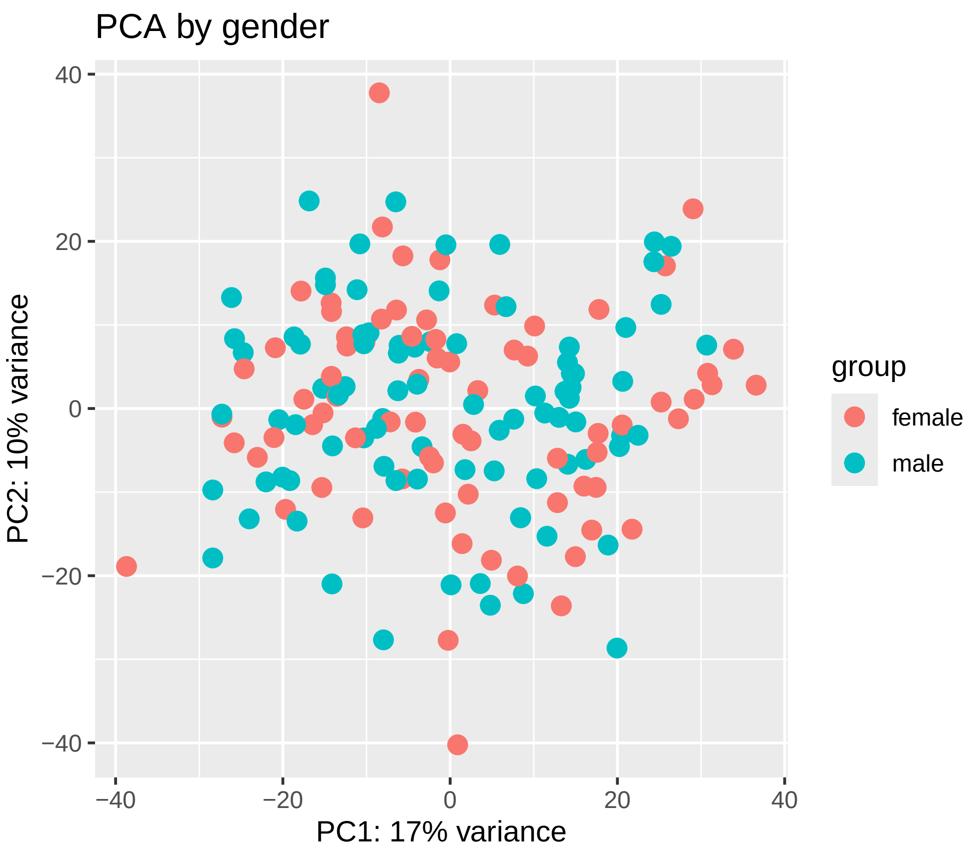
```
plotPCA(vsd, intgroup = c("ajcc_pathologic_m")) +ggtitle("PCA by ajcc_pathologic_m")
```

```
## using ntop=500 top features by variance
```



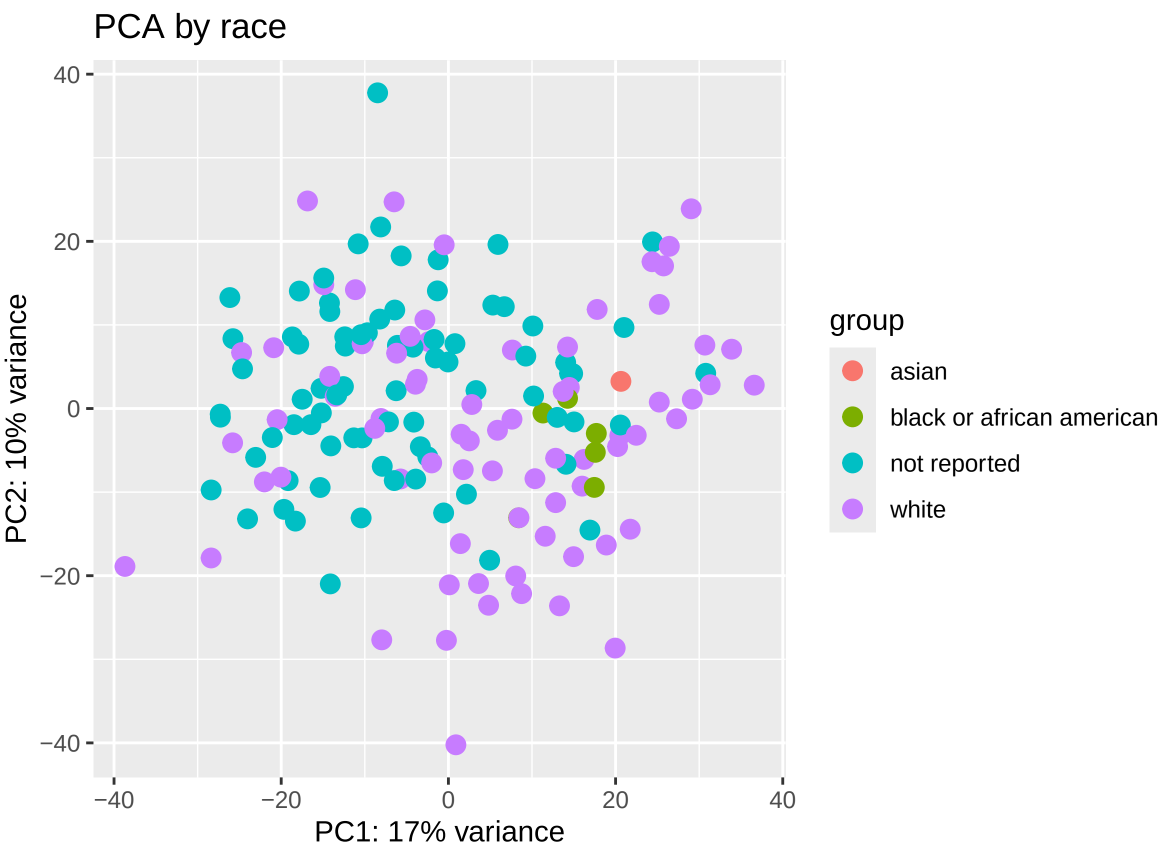
```
plotPCA(vsd, intgroup = c("gender")) +ggtitle("PCA by gender")
```

```
## using ntop=500 top features by variance
```



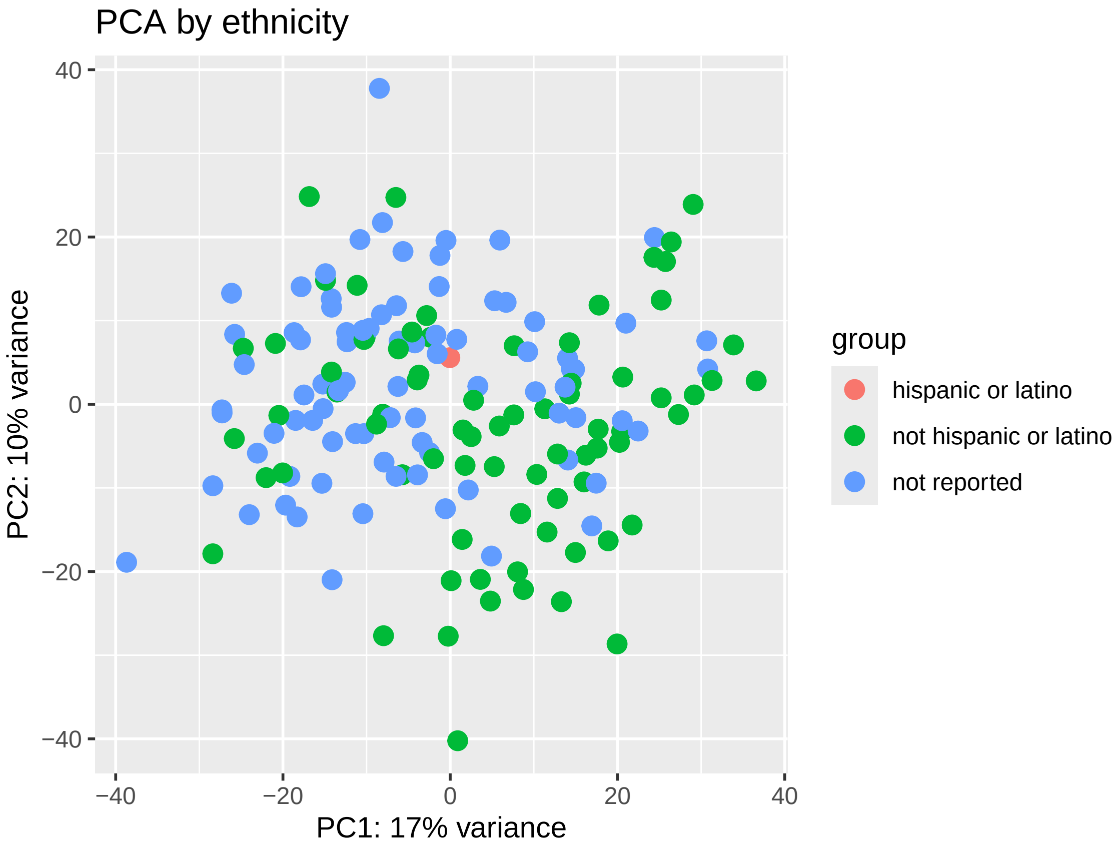
```
plotPCA(vsd, intgroup = c("race")) +ggtitle("PCA by race")
```

```
## using ntop=500 top features by variance
```



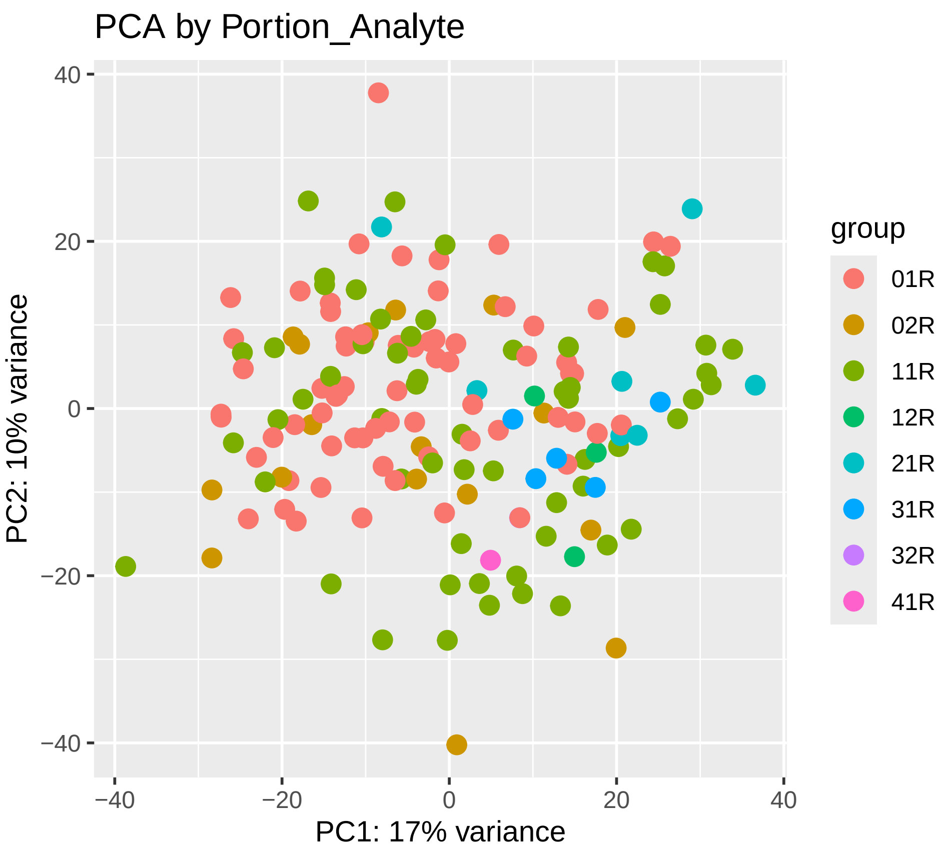
```
plotPCA(vsd, intgroup = c("ethnicity")) +ggtitle("PCA by ethnicity")
```

```
## using ntop=500 top features by variance
```



```
plotPCA(vsd, intgroup = c("Portion_Analyte")) +ggtitle("PCA by Portion_Analyte")
```

```
## using ntop=500 top features by variance
```



```
design(dds) = ~ vital_status + ajcc_pathologic_t + ajcc_pathologic_m + gender + age_diag_c + race + eth
```

```
##   Note: levels of factors in the design contain characters other than
##   letters, numbers, '_' and '.'. It is recommended (but not required) to use
##   only letters, numbers, and delimiters '_' or '.', as these are safe characters
##   for column names in R. [This is a message, not a warning or an error]
```

```
dds <- DESeq(dds)
```

```
## using pre-existing size factors

## estimating dispersions

## gene-wise dispersion estimates

## mean-dispersion relationship

##   Note: levels of factors in the design contain characters other than
##   letters, numbers, '_' and '.'. It is recommended (but not required) to use
##   only letters, numbers, and delimiters '_' or '.', as these are safe characters
##   for column names in R. [This is a message, not a warning or an error]

## final dispersion estimates
## fitting model and testing
```

```
## 482 rows did not converge in beta, labelled in mcols(object)$betaConv. Use larger maxit argument with
```

```r
#plotMDS(dds)
```

Comparing gene expression pattern by vital_status

```r
res_by_vital_status <- results(dds, contrast=c("vital_status","Dead","Alive"),pAdjustMethod = "BH",forma
res_by_vital_status_df=as.data.frame(res_by_vital_status)
res_by_vital_status_df$Ensembl=rownames(res_by_vital_status_df)
summary(res_by_vital_status)
```

```
##
## out of 17544 with nonzero total read count
## adjusted p-value < 0.1
## LFC > 0 (up)       : 117, 0.67%
## LFC < 0 (down)     : 39, 0.22%
## outliers [1]       : 0, 0%
## low counts [2]     : 0, 0%
## (mean count < 5)
## [1] see ’cooksCutoff’ argument of ?results
## [2] see ’independentFiltering’ argument of ?results
```

```r
res_by_vital_status_df=res_by_vital_status_df[!is.na(res_by_vital_status_df$padj),]
```

Now annotating the genes

```r
library(stringr)
res_by_vital_status_df[c('Ensembl', 'Dot')] <- str_split_fixed(res_by_vital_status_df$Ensembl, '\\.', 2

library(org.Hs.eg.db)
```

```
## Loading required package: AnnotationDbi
```

```
##
## Attaching package: ’AnnotationDbi’
```

```
## The following object is masked from ’package:dplyr’:
##
##     select
```

```
##
```

```r
res_by_vital_status_df$Entrez <- mapIds(org.Hs.eg.db, res_by_vital_status_df$Ensembl,keytype="ENSEMBL",
```

```
## ’select()’ returned 1:many mapping between keys and columns
```

```
res_by_vital_status_df$Symbol <- mapIds(org.Hs.eg.db, res_by_vital_status_df$Entrez,keytype="ENTREZID",
```

```
## 'select()' returned 1:1 mapping between keys and columns
```

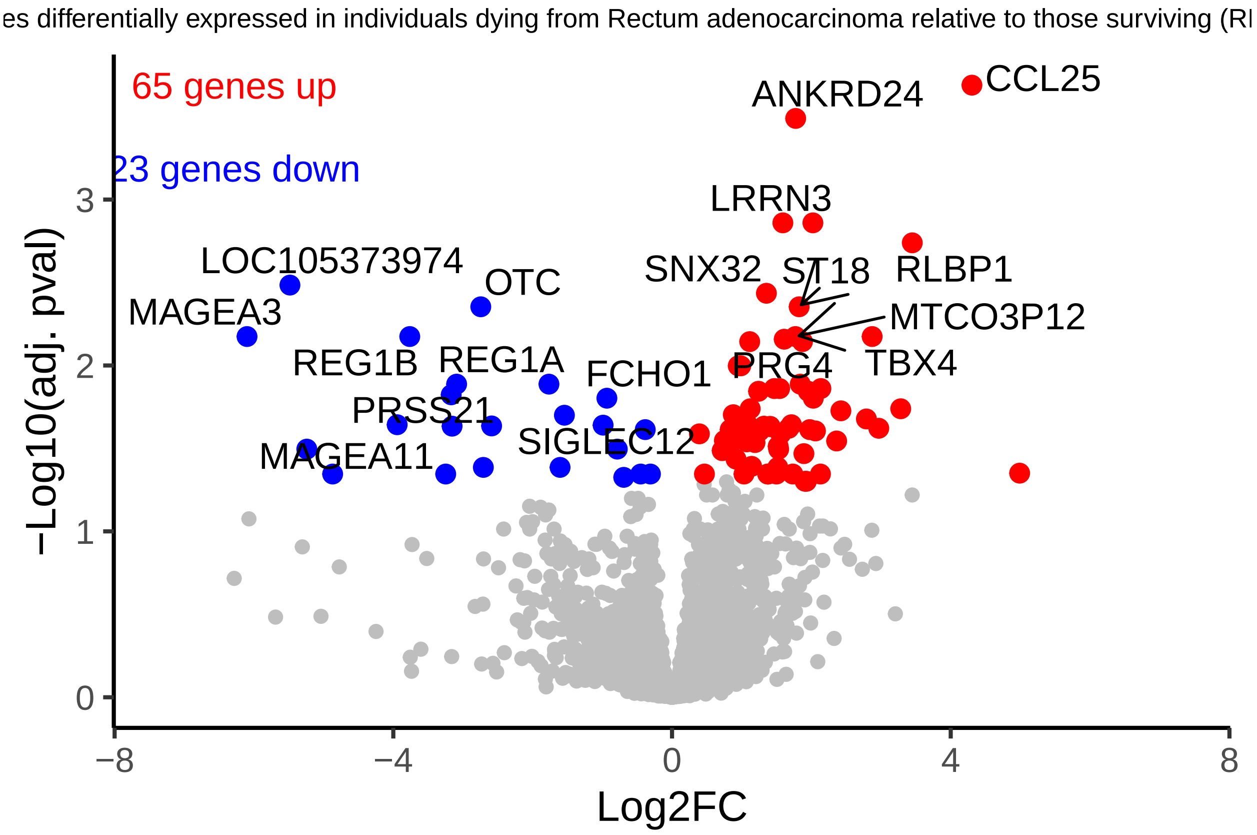
```
res_by_vital_status_df$Genename <- mapIds(org.Hs.eg.db, res_by_vital_status_df$Entrez,keytype="ENTREZID
```

```
## 'select()' returned 1:1 mapping between keys and columns
```

```
res_by_vital_status_df$Entrez=as.character(res_by_vital_status_df$Entrez)
res_by_vital_status_df$Symbol=as.character(res_by_vital_status_df$Symbol)
res_by_vital_status_df$Genename=as.character(res_by_vital_status_df$Genename)
```

Now generating a summary volcano plot

```
library(ggrepel)
res_by_vital_status_df$Log10_p_val = -log10(res_by_vital_status_df$padj)

upreg = subset(res_by_vital_status_df,res_by_vital_status_df$log2FoldChange > 0 & res_by_vital_status_d

downreg = subset(res_by_vital_status_df,res_by_vital_status_df$log2FoldChange < 0 & res_by_vital_status_

upreg_o = upreg[order(- upreg$Log10_p_val),]
downreg_o = downreg[order(- downreg$Log10_p_val),]

up_10=upreg_o[1:10,]
dn_10=downreg_o[1:10,]

up_dn_10 = rbind(up_10,dn_10)

up_dn_10_no_NULL = subset(up_dn_10,up_dn_10$Symbol!="NULL")

max_abs_val = max(abs(res_by_vital_status_df$log2FoldChange))

p=ggplot(res_by_vital_status_df, aes(log2FoldChange, Log10_p_val)) +
  theme_classic(base_size = 16)+
  geom_point(data=res_by_vital_status_df, aes(x=log2FoldChange, y=Log10_p_val), colour="grey", size=2)
p1 <- p +  geom_point(data = upreg_o, aes(x=log2FoldChange, y=Log10_p_val) ,size=3,color="red1")
p2 <- p1 +  geom_point(data = downreg_o, aes(x=log2FoldChange, y=Log10_p_val) ,size=3,color="blue1")
p3=p2+ggtitle(paste("Volcano plot for genes differentially expressed in individuals dying from Rectum ad
p3
```



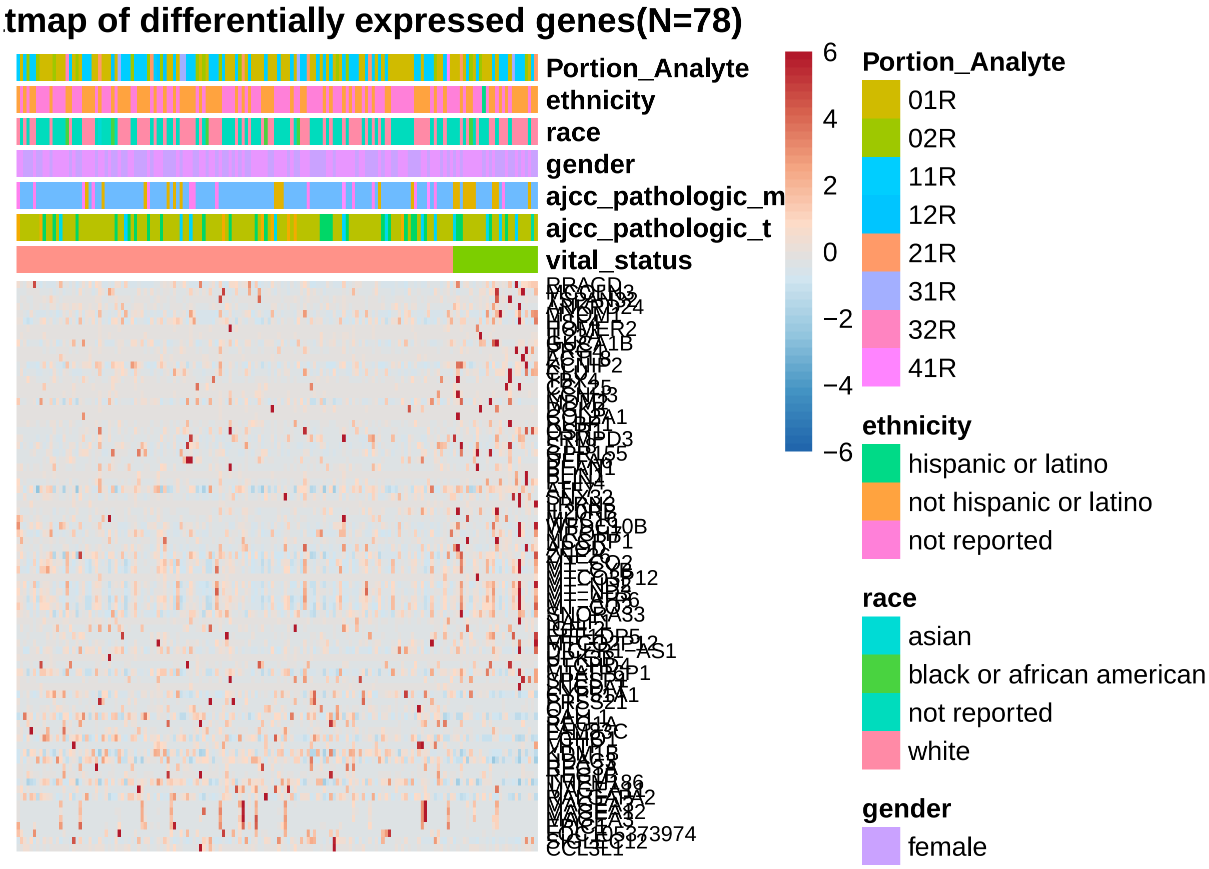
Generating aranging the genes in order to generate a Z score heatmap to show that expression data supports logFC data for the DE genes

```
cpm_ob = cpm(dds)
deg_order = rbind(upreg,downreg)
deg_no_null = subset(deg_order,deg_order$Symbol!="NULL")


cpm_diff_exp_genes = cpm_ob[rownames(deg_no_null),]
rownames(cpm_diff_exp_genes) = deg_no_null$Symbol

cpm_diff_exp_genes_dead = cpm_diff_exp_genes[,samples_use_clinical_final$Original_ID[samples_use_clinica
cpm_diff_exp_genes_alive = cpm_diff_exp_genes[,samples_use_clinical_final$Original_ID[samples_use_clini
cpm_diff_exp_genes_alive_dead=cbind(cpm_diff_exp_genes_alive,cpm_diff_exp_genes_dead)


alive_samp=subset(samples_use_clinical_final,samples_use_clinical_final$vital_status=="Alive")
dead_samp=subset(samples_use_clinical_final,samples_use_clinical_final$vital_status=="Dead")
alive_Dead_samp = rbind(alive_samp,dead_samp)
rownames(alive_Dead_samp)=alive_Dead_samp$Original_ID

length(intersect(colnames(cpm_diff_exp_genes_alive_dead),rownames(alive_Dead_samp)))
```

```
## [1] 160
```

Actual generation of the Z score heatmap for differentially expressed genes

```r
#stage_race_eth=stage_1_4_samp[,c("tumor_stage","race","ethnicity")]
#rownames(stage_race_eth) = rownames(stage_1_4_samp)
#colnames(stage_race_eth)=c("Tumor Stage","Race","Ethnicity")


library(pheatmap)
library(RColorBrewer)
pheatmap_object=pheatmap(cpm_diff_exp_genes_alive_dead,scale = "row",show_rownames = TRUE,cluster_rows =
                         col = colorRampPalette(rev(brewer.pal(8, "RdBu")))(50),breaks = seq(-6, 6, leng
```



Enrichment analysis using GO or Reactome database to find what functions are the DE genes involved in.

```r
library(clusterProfiler)
```

```
##

## clusterProfiler v4.16.0 Learn more at https://yulab-smu.top/contribution-knowledge-mining/
##
## Please cite:
##
## S Xu, E Hu, Y Cai, Z Xie, X Luo, L Zhan, W Tang, Q Wang, B Liu, R Wang,
## W Xie, T Wu, L Xie, G Yu. Using clusterProfiler to characterize
## multiomics data. Nature Protocols. 2024, 19(11):3292-3320

##
## Attaching package: 'clusterProfiler'
## The following object is masked from 'package:AnnotationDbi':
##
##     select

## The following object is masked from 'package:IRanges':
##
##     slice

## The following object is masked from 'package:S4Vectors':
##
##     rename

## The following object is masked from 'package:purrr':
##
##     simplify

## The following object is masked from 'package:stats':
##
##     filter
```

```r
library(msigdbr)
library(org.Mm.eg.db)
```

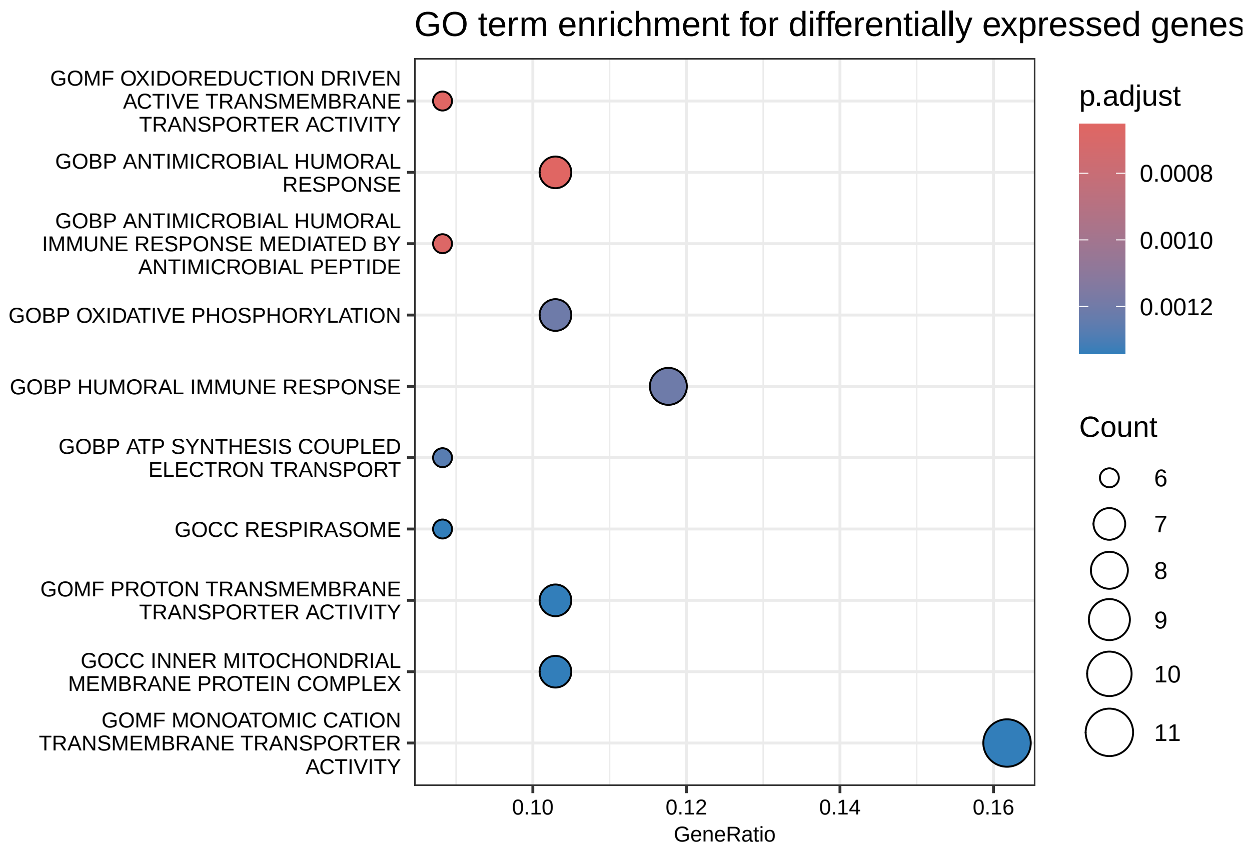
```
##
```

```r
library(magrittr)
```

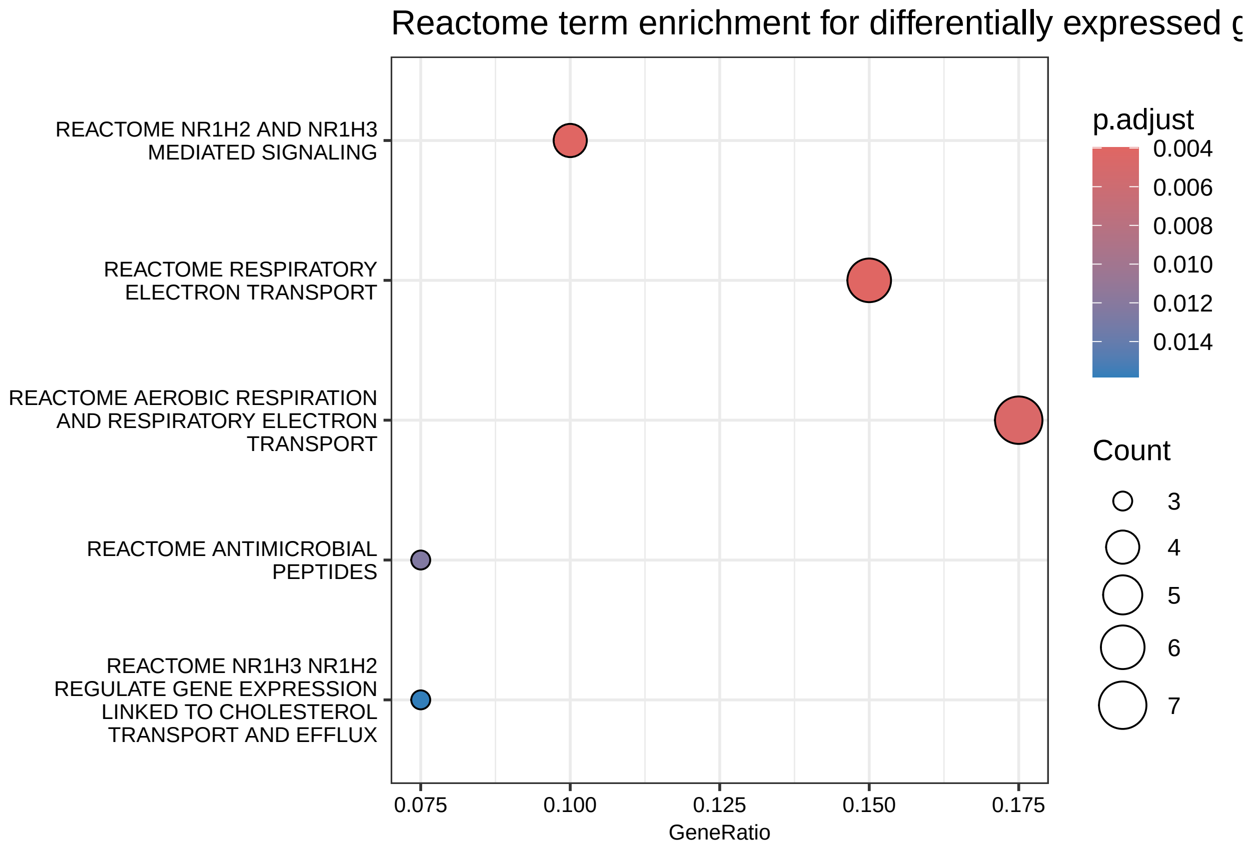
```
##
## Attaching package: 'magrittr'

## The following object is masked from 'package:GenomicRanges':
##
##     subtract

## The following object is masked from 'package:purrr':
##
##     set_names

## The following object is masked from 'package:tidyr':
##
##     extract
```

```r
mm_msigdb_df <- msigdbr(species = "Homo sapiens")

mm_GO_df <- mm_msigdb_df %>%
  dplyr::filter(
    gs_collection == "C5", # This is to filter only to the C2 curated gene sets
    gs_subcollection %in% c("GO:BP","GO:CC","GO:MF") # This is because we only want KEGG pathways
  )
```

```
mm_Reactome_df <- mm_msigdb_df %>%
  dplyr::filter(
    gs_subcollection == "CP:REACTOME" # This is to filter only to the C2 curated gene sets
     # This is because we only want KEGG pathways
  )


GO_ora_results <- enricher(
  gene = deg_order$Ensembl, # A vector of your genes of interest
  pvalueCutoff = 0.05, # Can choose a FDR cutoff
  pAdjustMethod = "BH",
  universe = res_by_vital_status_df$Ensembl,# Method to be used for multiple testing correction
  # The pathway information should be a data frame with a term name or
  # identifier and the gene identifiers
  TERM2GENE = dplyr::select(
    mm_GO_df,
    gs_name,
    ensembl_gene
  )
)


enrich_plot_GO <- enrichplot::dotplot(GO_ora_results, showCategory=10,font.size=8,title="GO term enrich
enrich_plot_GO
```



```
Reactome_ora_results <- enricher(
  gene = deg_order$Ensembl, # A vector of your genes of interest
  pvalueCutoff = 0.05, # Can choose a FDR cutoff
  pAdjustMethod = "BH",
  universe = res_by_vital_status_df$Ensembl,# Method to be used for multiple testing correction
  # The pathway information should be a data frame with a term name or
  # identifier and the gene identifiers
  TERM2GENE = dplyr::select(
    mm_Reactome_df,
    gs_name,
    ensembl_gene
  )
)


enrich_plot_Reactome <- enrichplot::dotplot(Reactome_ora_results, showCategory=10,font.size=8,title="Rea
enrich_plot_Reactome
```



Based on this analysis, the genes differentially expressed between those that died from Rectum adenocarcinoma vs those that did not priarily enrich for various transmembrane transporter genes as well immune response genes such as antimicrobial humoral response with additional involvement of ETC related genes.

In terms of pathways, NR1H2 and NR1H3 are Nuclear receptors that exhibits a ligand-dependent transcriptional activation activity and regulates cholesterol uptake through MYLIP-dependent ubiquitination of LDLR, VLDLR and LRP8; DLDLR and LRP8. These genes are also key regulators of macrophage function, controlling transcriptional programs involved in lipid homeostasis and inflammation.

Of note, macrophages control inflammation in the rectum via polarisation of M0 macrophages to M1 (pro-inflmmatory) and M2 (anti-inflammatory). They can be either pro-inflammatory, potentially inhibiting tumor growth, or anti-inflammatory, promoting tumor development and metastasis.Hence macrophage function via regulation of NR1H2 and NR1H3 receptor activity could be important to determining surival in Rectum adenocarcinoma.