mrr_project_bio

2025-11-10

Classification of Cancer outcome using Genetic and Clinical data

Introduction

This project investigates the effect of genetic and clinical variables on the survival outcome of breast cancer patients. The data contains records of more than 1000 breast cancer patients from several research institutions. Clinical data contains patient-related and tumor-related information. Additionally mRNA gene expression data is available for each patient. The gene expression data has been processed to include only the top 5000 most variable genes on the transformed scale (log2(counts + 1)).

We consider that the main outcome variable of interest is **vital_status**, which is defined in clinical data.

Load data

```
load("mrr_bio.Rdata")

# genetic data: n x p matrix
dim(GeneX) # 1231 x 5000
```

```
## [1] 1231 5000
```

```
str(GeneX)
```

```
##  num [1:1231, 1:5000] 0 0 10.92 3.17 5 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:1231] "EW-A2FS-01A-11R-A17B" "OL-A6VR-01A-32R-A33J" "E9-A226-01A-21R-A157" "A8-A08H-0
##   ..$ : chr [1:5000] "CLEC3A" "SCGB2A2" "CPB1" "GSTM1" ...
```

```
colnames(GeneX)[1:10] # first 10 names of genes
```

```
##  [1] "CLEC3A"    "SCGB2A2"   "CPB1"      "GSTM1"     "TFF1"      "SCGB1D2"
##  [7] "KCNJ3"     "MUCL1"     "LINC00993" "ANKRD30A"
```

```
# clinical data
dim(clinical_data) # 1231 x 24
```

```
## Loading required namespace: S4Vectors
```

```
## NULL
```

```
head(clinical_data)
```

```
## DataFrame with 6 rows and 24 columns
##                      initial_weight tissue_type  laterality
##                           <numeric> <character> <character>
## EW-A2FS-01A-11R-A17B            110       Tumor        Left
## OL-A6VR-01A-32R-A33J            380       Tumor       Right
## E9-A226-01A-21R-A157            510       Tumor        Left
## A8-A08H-01A-21R-A00Z            250       Tumor        Left
## D8-A27H-01A-11R-A16F            170       Tumor       Right
## D8-A3Z6-01A-11R-A239             60       Tumor        Left
##                      tissue_or_organ_of_origin age_at_diagnosis
##                                    <character>        <integer>
## EW-A2FS-01A-11R-A17B               Breast, NOS            15302
## OL-A6VR-01A-32R-A33J               Breast, NOS            17702
## E9-A226-01A-21R-A157               Breast, NOS            16511
## A8-A08H-01A-21R-A00Z               Breast, NOS            24137
## D8-A27H-01A-11R-A16F               Breast, NOS            26588
## D8-A3Z6-01A-11R-A239               Breast, NOS            20629
##                           primary_diagnosis prior_treatment
##                                 <character>     <character>
## EW-A2FS-01A-11R-A17B Infiltrating duct ca..              No
## OL-A6VR-01A-32R-A33J Infiltrating duct ca..              No
## E9-A226-01A-21R-A157 Infiltrating duct ca..              No
## A8-A08H-01A-21R-A00Z Infiltrating duct ca..              No
## D8-A27H-01A-11R-A16F Infiltrating duct ca..              No
## D8-A3Z6-01A-11R-A239 Lobular carcinoma, NOS              No
##                      diagnosis_is_primary_disease ajcc_pathologic_t  morphology
##                                         <logical>       <character> <character>
## EW-A2FS-01A-11R-A17B                         TRUE                T2      8500/3
## OL-A6VR-01A-32R-A33J                         TRUE               T1b      8500/3
## E9-A226-01A-21R-A157                         TRUE                T2      8500/3
## A8-A08H-01A-21R-A00Z                         TRUE                T2      8500/3
## D8-A27H-01A-11R-A16F                         TRUE                T2      8500/3
## D8-A3Z6-01A-11R-A239                         TRUE                T3      8520/3
##                      classification_of_tumor   sites_of_involvement
##                                  <character>                 <list>
## EW-A2FS-01A-11R-A17B                 primary Breast, Left Upper O..
## OL-A6VR-01A-32R-A33J                 primary            Breast, NOS
## E9-A226-01A-21R-A157                 primary Breast, Left Lower I..
## A8-A08H-01A-21R-A00Z                 primary Breast, Left Upper I..
## D8-A27H-01A-11R-A16F                 primary Breast, Right Upper ..
## D8-A3Z6-01A-11R-A239                 primary            Breast, NOS
##                      days_to_last_follow_up follow_ups_disease_response
##                                   <integer>                 <character>
## EW-A2FS-01A-11R-A17B                  1604               TF-Tumor Free
## OL-A6VR-01A-32R-A33J                  1220               TF-Tumor Free
## E9-A226-01A-21R-A157                  1048               WT-With Tumor
## A8-A08H-01A-21R-A00Z                     0               TF-Tumor Free
## D8-A27H-01A-11R-A16F                   397               TF-Tumor Free
## D8-A3Z6-01A-11R-A239                   563               TF-Tumor Free
##                                      race      gender             ethnicity
##                                 <character> <character>           <character>
```

```
## EW-A2FS-01A-11R-A17B black or african ame..      female not hispanic or latino
## OL-A6VR-01A-32R-A33J black or african ame..      female not hispanic or latino
## E9-A226-01A-21R-A157                 white      female not hispanic or latino
## A8-A08H-01A-21R-A00Z          not reported      female           not reported
## D8-A27H-01A-11R-A16F                 white      female not hispanic or latino
## D8-A3Z6-01A-11R-A239                 white      female not hispanic or latino
##                      vital_status age_at_index days_to_birth age_is_obfuscated
##                       <character>    <integer>     <integer>         <logical>
## EW-A2FS-01A-11R-A17B        Alive           41        -15302             FALSE
## OL-A6VR-01A-32R-A33J        Alive           48        -17702             FALSE
## E9-A226-01A-21R-A157         Dead           45        -16511             FALSE
## A8-A08H-01A-21R-A00Z        Alive           66        -24137             FALSE
## D8-A27H-01A-11R-A16F        Alive           72        -26588             FALSE
## D8-A3Z6-01A-11R-A239        Alive           56        -20629             FALSE
##                      bcr_patient_barcode primary_site
##                                <character>       <list>
## EW-A2FS-01A-11R-A17B    TCGA-EW-A2FS-01A       Breast
## OL-A6VR-01A-32R-A33J    TCGA-OL-A6VR-01A       Breast
## E9-A226-01A-21R-A157    TCGA-E9-A226-01A       Breast
## A8-A08H-01A-21R-A00Z    TCGA-A8-A08H-01A       Breast
## D8-A27H-01A-11R-A16F    TCGA-D8-A27H-01A       Breast
## D8-A3Z6-01A-11R-A239    TCGA-D8-A3Z6-01A       Breast
##                                                                        disease_type
##                                                                              <list>
## EW-A2FS-01A-11R-A17B Adenomas and Adenoca..,Adnexal and Skin App..,Fibroepithelial Neop..,...
## OL-A6VR-01A-32R-A33J Adenomas and Adenoca..,Adnexal and Skin App..,Fibroepithelial Neop..,...
## E9-A226-01A-21R-A157 Adenomas and Adenoca..,Adnexal and Skin App..,Fibroepithelial Neop..,...
## A8-A08H-01A-21R-A00Z Adenomas and Adenoca..,Adnexal and Skin App..,Fibroepithelial Neop..,...
## D8-A27H-01A-11R-A16F Adenomas and Adenoca..,Adnexal and Skin App..,Fibroepithelial Neop..,...
## D8-A3Z6-01A-11R-A239 Adenomas and Adenoca..,Adnexal and Skin App..,Fibroepithelial Neop..,...
```

```
colnames(clinical_data) # variables
```

```
##  [1] "initial_weight"              "tissue_type"
##  [3] "laterality"                  "tissue_or_organ_of_origin"
##  [5] "age_at_diagnosis"            "primary_diagnosis"
##  [7] "prior_treatment"             "diagnosis_is_primary_disease"
##  [9] "ajcc_pathologic_t"           "morphology"
## [11] "classification_of_tumor"     "sites_of_involvement"
## [13] "days_to_last_follow_up"      "follow_ups_disease_response"
## [15] "race"                        "gender"
## [17] "ethnicity"                   "vital_status"
## [19] "age_at_index"                "days_to_birth"
## [21] "age_is_obfuscated"           "bcr_patient_barcode"
## [23] "primary_site"                "disease_type"
```

```
# outcome variable
y = clinical_data$vital_status
table(y)
```

```
## y
## Alive  Dead
##  1029   201
```

Exploratory data analysis: example

```
# summary statistics of first 10 genes
apply(GeneX[,1:10],2, summary)
```

```
##            CLEC3A  SCGB2A2      CPB1     GSTM1      TFF1   SCGB1D2     KCNJ3
## Min.     0.000000  0.00000  0.000000  0.000000  0.000000  0.000000  0.000000
## 1st Qu.  1.000000  7.48779  4.523562  0.000000  6.679463  5.614710  2.807355
## Median   4.523562 11.70434  7.686501  5.491853 10.796040  9.273796  5.930737
## Mean     5.925758 11.12788  8.313829  5.357693  9.712977  9.026789  6.687515
## 3rd Qu. 10.175539 15.06285 11.151939 10.344845 13.207240 12.274950 10.321917
## Max.    19.455835 22.36921 22.745581 15.874141 18.927850 21.479643 16.672715
##             MUCL1 LINC00993  ANKRD30A
## Min.     0.000000  0.000000  0.000000
## 1st Qu.  5.894767  5.643856  5.554589
## Median   8.945444 10.512740 10.338736
## Mean     9.151482  8.864887  8.650306
## 3rd Qu. 12.404866 12.230770 11.990273
## Max.    20.547814 15.914922 16.758184
```

```
# clinical variables: gender
unique(clinical_data$gender) # unique values
```

```
## [1] "female" "male"   NA
```

```
table(clinical_data$gender) # frequency
```

```
##
## female   male
##   1217     13
```