BRCA data exploration

Urminder Singh

May 2, 2019

BRCA data exploration

This is a short lesson on how to explore the mutation data at TCGA using TCGABiolinks and maftools libraries. Using study metadata is crucial to the analysis of TCGA datasets. I have provided some functions which user can use to format TCGA metadata and easily use that in analysis. I will particulary look at BRCA data at TCGA but the methods and functions provided here could easily be extended to any cancer type.

Downloading TCGA Metadata

For all the data deposited in TCGA, there is associated clinical metadata. Clinical metadata could be critical in understanding the data from different perspectives. For example, the clinical data has information about tumor stage, gender, age etc. all of which could be helpful in data analysis from various perspectives.

TCGAbiolinks provides the function "*GDCquery_clinic*" to download the clinical metadata. The clinical metadata is separated into two categories i.e. Clinical and Biospecimen. More information on this is available at https://bioconductor.org/packages/release/bioc/vignettes/TCGAbiolinks/inst/doc/clinical.html

A simple example to use *GDCquery_clinic* function is shown below:

```
clinicalBRCA <- GDCquery_clinic(project = "TCGA-BRCA", type = "clinical")
biospecimenBRCA <- GDCquery_clinic(project = "TCGA-BRCA", type = "Biospecimen")

head(clinicalBRCA)
```

```
##   submitter_id classification_of_tumor last_known_disease_status
## 1 TCGA-3C-AAAU            not reported              not reported
## 2 TCGA-3C-AALI            not reported              not reported
## 3 TCGA-3C-AALJ            not reported              not reported
## 4 TCGA-3C-AALK            not reported              not reported
## 5 TCGA-4H-AAAK            not reported              not reported
## 6 TCGA-5L-AAT0            not reported              not reported
##                   updated_datetime               primary_diagnosis
## 1 2018-09-06T13:49:20.245333-05:00          Lobular carcinoma, NOS
## 2 2018-09-06T13:49:20.245333-05:00 Infiltrating duct carcinoma, NOS
## 3 2018-09-06T13:49:20.245333-05:00 Infiltrating duct carcinoma, NOS
## 4 2018-09-06T13:49:20.245333-05:00 Infiltrating duct carcinoma, NOS
## 5 2018-09-06T13:49:20.245333-05:00          Lobular carcinoma, NOS
## 6 2018-09-06T13:49:20.245333-05:00          Lobular carcinoma, NOS
##   tumor_stage age_at_diagnosis vital_status morphology days_to_death
## 1     stage x            20211        alive     8520/3            NA
## 2   stage iib            18538        alive     8500/3            NA
## 3   stage iib            22848        alive     8500/3            NA
## 4    stage ia            19074        alive     8500/3            NA
## 5  stage iiia            18371        alive     8520/3            NA
## 6   stage iia            15393        alive     8520/3            NA
##   days_to_last_known_disease_status created_datetime    state
## 1                                NA               NA released
```

```
## 2                                NA               NA released
## 3                                NA               NA released
## 4                                NA               NA released
## 5                                NA               NA released
## 6                                NA               NA released
##   days_to_recurrence                        diagnosis_id  tumor_grade
## 1                 NA 8cfb8afb-b915-5255-865b-a5923f47b351 not reported
## 2                 NA 8cafc022-585f-54a1-a7d4-cfa632b3991e not reported
## 3                 NA 63d85b81-8eba-5f17-8552-92babf137c00 not reported
## 4                 NA 8c90b19d-54f7-5788-a5eb-49abe239ef0b not reported
## 5                 NA 81c406cd-ad2d-552f-b543-8b2b03044686 not reported
## 6                 NA ebff6a7b-3b6c-5f71-86b0-1bdd3b78edd4 not reported
##   tissue_or_organ_of_origin days_to_birth progression_or_recurrence
## 1               Breast, NOS        -20211              not reported
## 2               Breast, NOS        -18538              not reported
## 3               Breast, NOS        -22848              not reported
## 4               Breast, NOS        -19074              not reported
## 5               Breast, NOS        -18371              not reported
## 6               Breast, NOS        -15393              not reported
##   prior_malignancy site_of_resection_or_biopsy days_to_last_follow_up
## 1     not reported                Breast, NOS                   4047
## 2     not reported                Breast, NOS                   4005
## 3     not reported                Breast, NOS                   1474
## 4     not reported                Breast, NOS                   1448
## 5     not reported                Breast, NOS                    348
## 6     not reported                Breast, NOS                   1477
##   cigarettes_per_day weight alcohol_history alcohol_intensity bmi
## 1                 NA     NA              NA                NA  NA
## 2                 NA     NA              NA                NA  NA
## 3                 NA     NA              NA                NA  NA
## 4                 NA     NA              NA                NA  NA
## 5                 NA     NA              NA                NA  NA
## 6                 NA     NA              NA                NA  NA
##   years_smoked                          exposure_id height gender
## 1           NA 72f0be98-dffa-5d35-88fe-f9ca774d6db0     NA female
## 2           NA 63b59ac3-ccc2-5590-bff7-673f15713369     NA female
## 3           NA 05defab8-b347-540a-8950-8a180faeb67e     NA female
## 4           NA a97db788-0772-5d32-878c-d2080d979c37     NA female
## 5           NA e80a24a3-d0fc-5067-a47d-83ac937af2f0     NA female
## 6           NA 47c9213f-b2b8-5297-a0b6-21bce2cfa3f8     NA female
##   year_of_birth                    race
## 1          1949                   white
## 2          1953 black or african american
## 3          1949 black or african american
## 4          1959 black or african american
## 5          1963                   white
## 6          1968                   white
##                          demographic_id              ethnicity
## 1 cee0a94c-1d9e-5650-a500-a6b021fe138d not hispanic or latino
## 2 583a1ee5-3175-523e-ba1f-75a30a3e1e41 not hispanic or latino
## 3 9619a908-1684-547f-a407-6adf93e15b8d not hispanic or latino
## 4 e54b1469-fffc-5291-a8e3-df2092ab5f34 not hispanic or latino
## 5 64a204f0-b380-5962-a927-84af1841b6d6 not hispanic or latino
## 6 776bb6f9-f5c8-57b6-bf4d-4014b8025a06     hispanic or latino
```

```
##   year_of_death                          treatment_id therapeutic_agents
## 1            NA 2d88df62-dc75-5c01-b249-3b914cd7380a                 NA
## 2            NA 89c2d475-7048-52d3-8dc0-1425330d35ee                 NA
## 3            NA a1ecb0cf-fa4a-58df-8c6d-c1baaad53f7e                 NA
## 4            NA 35fe07bb-5a79-5549-9a08-a851d9aa3de1                 NA
## 5            NA 68417a03-5e66-535d-ac9a-fa6ffb98b571                 NA
## 6            NA 38f1dbba-d771-539b-91e7-5b9761c0f592                 NA
##   treatment_intent_type treatment_or_therapy bcr_patient_barcode disease
## 1                    NA                   NA        TCGA-3C-AAAU    BRCA
## 2                    NA                   NA        TCGA-3C-AALI    BRCA
## 3                    NA                   NA        TCGA-3C-AALJ    BRCA
## 4                    NA                   NA        TCGA-3C-AALK    BRCA
## 5                    NA                   NA        TCGA-4H-AAAK    BRCA
## 6                    NA                   NA        TCGA-5L-AAT0    BRCA
```

```
head(biospecimenBRCA)
```

```
##   sample_type_id                 updated_datetime
## 1             01 2018-11-15T21:38:54.195821-06:00
## 2             11 2018-11-15T21:38:54.195821-06:00
## 3             01 2018-11-15T21:10:03.529893-06:00
## 4             01 2018-11-15T21:38:54.195821-06:00
## 5             10 2018-11-15T21:38:54.195821-06:00
## 6             01 2018-11-15T21:10:03.529893-06:00
##   time_between_excision_and_freezing oct_embedded tumor_code_id
## 1                                 NA         true            NA
## 2                                 NA         true            NA
## 3                                 NA           No            NA
## 4                                 NA         true            NA
## 5                                 NA        false            NA
## 6                                 NA           No            NA
##       submitter_id intermediate_dimension
## 1 TCGA-BH-A0C3-01A                     NA
## 2 TCGA-BH-A0C3-11A                     NA
## 3 TCGA-BH-A0C3-01Z                     NA
## 4 TCGA-BH-A0HQ-01A                     NA
## 5 TCGA-BH-A0HQ-10A                     NA
## 6 TCGA-BH-A0HQ-01Z                     NA
##                               sample_id is_ffpe
## 1 21ba28ff-89f6-4f02-a135-821efc4f42f8   FALSE
## 2 82e6dc7b-fe63-4fc6-af9f-68dc76c2cb88   FALSE
## 3 bb18ed04-1c02-41f5-af7a-d82980d185f3    TRUE
## 4 ef48e806-e31c-4a11-afe8-dcc232357329   FALSE
## 5 c14eb1f6-791e-491f-8b60-f9856daf77b8   FALSE
## 6 6b4f016a-8a55-49fd-9331-855a5fde317e    TRUE
##                   pathology_report_uuid                  created_datetime
## 1 3A54CF6E-AFDB-4609-A827-77D75BB376A7                              <NA>
## 2                                  <NA>                              <NA>
## 3                                  <NA> 2018-05-17T12:14:28.274820-05:00
## 4 A76A272F-675E-4E56-8761-96B71419A012                              <NA>
## 5                                  <NA>                              <NA>
## 6                                  <NA> 2018-05-17T12:12:29.643720-05:00
##   tumor_descriptor          sample_type    state current_weight
## 1               NA        Primary Tumor released             NA
## 2               NA  Solid Tissue Normal released             NA
```

```
## 3               NA        Primary Tumor released             NA
## 4               NA        Primary Tumor released             NA
## 5               NA Blood Derived Normal released             NA
## 6               NA        Primary Tumor released             NA
##   composition time_between_clamping_and_freezing shortest_dimension
## 1          NA                                NA                 NA
## 2          NA                                NA                 NA
## 3          NA                                NA                 NA
## 4          NA                                NA                 NA
## 5          NA                                NA                 NA
## 6          NA                                NA                 NA
##   tumor_code  tissue_type days_to_sample_procurement freezing_method
## 1         NA Not Reported                        NA              NA
## 2         NA Not Reported                        NA              NA
## 3         NA Not Reported                         0              NA
## 4         NA Not Reported                        NA              NA
## 5         NA Not Reported                        NA              NA
## 6         NA Not Reported                         0              NA
##
## 1
## 2
## 3
## 4 1277769600, 11, 30, 2018-09-06T13:49:20.245333-05:00, NA, 2018-09-06T13:49:20.245333-05:00, 2018-09
## 5
## 6
##   preservation_method days_to_collection initial_weight longest_dimension
## 1               <NA>               1335            160                NA
## 2               <NA>               1335            200                NA
## 3               FFPE                 NA             NA                NA
## 4               <NA>                962            110                NA
## 5               <NA>                962             NA                NA
## 6               FFPE                 NA             NA                NA
##   distance_normal_to_tumor biospecimen_anatomic_site
## 1                       NA                        NA
## 2                       NA                        NA
## 3                       NA                        NA
## 4                       NA                        NA
## 5                       NA                        NA
## 6                       NA                        NA
##   diagnosis_pathologically_confirmed distributor_reference
## 1                                 NA                    NA
## 2                                 NA                    NA
## 3                                 NA                    NA
## 4                                 NA                    NA
## 5                                 NA                    NA
## 6                                 NA                    NA
##   method_of_sample_procurement passage_count biospecimen_laterality
## 1                           NA            NA                     NA
## 2                           NA            NA                     NA
## 3                           NA            NA                     NA
## 4                           NA            NA                     NA
## 5                           NA            NA                     NA
## 6                           NA            NA                     NA
##   growth_rate catalog_reference
```

```
## 1          NA                NA
## 2          NA                NA
## 3          NA                NA
## 4          NA                NA
## 5          NA                NA
## 6          NA                NA
```

The clinical metadata is saved in clinical and Biospecimen categories to represent two different type of data associated with each patient. This also removes a lot redundancy from the data had this been a single table. This structure is also reflected in the downloaded metadata and this makes it really hard to use the metadata. For example, in the *biospecimenBRCA* table, the column *portions* contain a dataframe for each row. This is beacuse a multiple portions could be mapped to a single *submitter_id* (for more information on metadata see: https://docs.cancergenomicscloud.org/docs/tcga-metadata). This is true for other columns *portions.analytes*, *portions.analytes.aliquots"* as well.

I have written a function *getTCGAMetadata* to format the clinical metadata into one table. This makes the metadata simple and easy to use. Example usage is shown below:

```
#function to expand columns of TCGA metadata from list to df. Used in getTCGAMetadata function
expand<-function(df,colName){
  res<-data.frame()
  #for each row
  for(i in 1: dim(df)[1]){
    thisRow<-df[i, ! (colnames(df) %in% c(colName))]
    tempdf<-as.data.frame(df[i, c(colName)])
    #if list is empty skip that row
    if(dim(tempdf)[1]<1){
      next
    }
    #change colnames so they are unique
    colnames(tempdf)<-paste(paste(colName,".",sep = ""),colnames(tempdf),sep = "")
    #print(paste(i,colnames(tempdf)))
    newRow<-cbind(thisRow,tempdf,row.names = NULL)
    res<-bind_rows(res,newRow)

  }
  #print(res)
  return(res)
}

#function to download and combine TCGA clinical and Biospecimen metadata given a project name
#example usage: getTCGAMetadata("TCGA-BRCA")
getTCGAMetadata<-function(projName){
  print(paste("Downloading",projName))
  clinicalBRCA <- GDCquery_clinic(project = projName, type = "clinical")
  biospecimenBRCA <- GDCquery_clinic(project = projName, type = "Biospecimen")

  #rename all cols from clinical table with suffix clinical
  colnames(clinicalBRCA)<- paste0("clinical.",colnames(clinicalBRCA))

  #expand biospecimen data in the order portions, portions.analytes, portions.analytes.aliquots
  toUnpack<-c("portions", "portions.analytes", "portions.analytes.aliquots")
  for(s in toUnpack){
    biospecimenBRCA<-expand(biospecimenBRCA,s)
  }
```

```
  #add patient barcode to biospecimen data
  biospecimenBRCA<- biospecimenBRCA %>% mutate(clinical.bcr_patient_barcode=substr(submitter_id,1,nchar
  #join clinical and biospecimen
  finalJoined<-join(clinicalBRCA,biospecimenBRCA,by="clinical.bcr_patient_barcode")
  return(finalJoined)
}

#a list of useful (suggested) columns to retain from the TCGA metadata (to reduce the dimensions)
colsToKeep<-c("clinical.submitter_id",
              "clinical.classification_of_tumor",
              "clinical.primary_diagnosis",
              "clinical.tumor_stage",
              "clinical.age_at_diagnosis",
              "clinical.vital_status",
              "clinical.days_to_death",
              "clinical.tissue_or_organ_of_origin",
              "clinical.days_to_birth",
              "clinical.site_of_resection_or_biopsy",
              "clinical.days_to_last_follow_up",
              "clinical.cigarettes_per_day",
              "clinical.weight",
              "clinical.alcohol_history",
              "clinical.bmi",
              "clinical.years_smoked",
              "clinical.height",
              "clinical.gender",
              "clinical.year_of_birth",
              "clinical.race",
              "clinical.ethnicity",
              "clinical.year_of_death",
              "clinical.bcr_patient_barcode",
              "clinical.disease",
              "submitter_id",
              "sample_type",
              "tissue_type",
              "portions.submitter_id",
              "portions.analytes.analyte_type",
              "portions.analytes.submitter_id",
              "portions.analytes.analyte_type_id",
              "portions.analytes.aliquots.analyte_type",
              "portions.analytes.aliquots.submitter_id")


#download BRCA metadata
brcaMetadata<-getTCGAMetadata("TCGA-BRCA")
```

```
## [1] "Downloading TCGA-BRCA"
```

```
#only keep useful columns
brcaMetadata<-brcaMetadata[,colsToKeep]

head(brcaMetadata)
```

```
##   clinical.submitter_id clinical.classification_of_tumor
## 1          TCGA-3C-AAAU                      not reported
```

```
## 2          TCGA-3C-AAAU                    not reported
## 3          TCGA-3C-AAAU                    not reported
## 4          TCGA-3C-AAAU                    not reported
## 5          TCGA-3C-AAAU                    not reported
## 6          TCGA-3C-AAAU                    not reported
##   clinical.primary_diagnosis clinical.tumor_stage
## 1    Lobular carcinoma, NOS             stage x
## 2    Lobular carcinoma, NOS             stage x
## 3    Lobular carcinoma, NOS             stage x
## 4    Lobular carcinoma, NOS             stage x
## 5    Lobular carcinoma, NOS             stage x
## 6    Lobular carcinoma, NOS             stage x
##   clinical.age_at_diagnosis clinical.vital_status clinical.days_to_death
## 1                     20211                alive                     NA
## 2                     20211                alive                     NA
## 3                     20211                alive                     NA
## 4                     20211                alive                     NA
## 5                     20211                alive                     NA
## 6                     20211                alive                     NA
##   clinical.tissue_or_organ_of_origin clinical.days_to_birth
## 1                      Breast, NOS                -20211
## 2                      Breast, NOS                -20211
## 3                      Breast, NOS                -20211
## 4                      Breast, NOS                -20211
## 5                      Breast, NOS                -20211
## 6                      Breast, NOS                -20211
##   clinical.site_of_resection_or_biopsy clinical.days_to_last_follow_up
## 1                       Breast, NOS                            4047
## 2                       Breast, NOS                            4047
## 3                       Breast, NOS                            4047
## 4                       Breast, NOS                            4047
## 5                       Breast, NOS                            4047
## 6                       Breast, NOS                            4047
##   clinical.cigarettes_per_day clinical.weight clinical.alcohol_history
## 1                          NA             NA                       NA
## 2                          NA             NA                       NA
## 3                          NA             NA                       NA
## 4                          NA             NA                       NA
## 5                          NA             NA                       NA
## 6                          NA             NA                       NA
##   clinical.bmi clinical.years_smoked clinical.height clinical.gender
## 1           NA                   NA             NA          female
## 2           NA                   NA             NA          female
## 3           NA                   NA             NA          female
## 4           NA                   NA             NA          female
## 5           NA                   NA             NA          female
## 6           NA                   NA             NA          female
##   clinical.year_of_birth clinical.race     clinical.ethnicity
## 1                   1949         white not hispanic or latino
## 2                   1949         white not hispanic or latino
## 3                   1949         white not hispanic or latino
## 4                   1949         white not hispanic or latino
## 5                   1949         white not hispanic or latino
## 6                   1949         white not hispanic or latino
```

```
##   clinical.year_of_death clinical.bcr_patient_barcode clinical.disease
## 1                     NA                 TCGA-3C-AAAU             BRCA
## 2                     NA                 TCGA-3C-AAAU             BRCA
## 3                     NA                 TCGA-3C-AAAU             BRCA
## 4                     NA                 TCGA-3C-AAAU             BRCA
## 5                     NA                 TCGA-3C-AAAU             BRCA
## 6                     NA                 TCGA-3C-AAAU             BRCA
##       submitter_id          sample_type  tissue_type portions.submitter_id
## 1 TCGA-3C-AAAU-10A Blood Derived Normal Not Reported   TCGA-3C-AAAU-10A-01
## 2 TCGA-3C-AAAU-10A Blood Derived Normal Not Reported   TCGA-3C-AAAU-10A-01
## 3 TCGA-3C-AAAU-01A        Primary Tumor Not Reported   TCGA-3C-AAAU-01A-11
## 4 TCGA-3C-AAAU-01A        Primary Tumor Not Reported   TCGA-3C-AAAU-01A-11
## 5 TCGA-3C-AAAU-01A        Primary Tumor Not Reported   TCGA-3C-AAAU-01A-11
## 6 TCGA-3C-AAAU-01A        Primary Tumor Not Reported   TCGA-3C-AAAU-01A-11
##   portions.analytes.analyte_type portions.analytes.submitter_id
## 1                            DNA           TCGA-3C-AAAU-10A-01D
## 2                            DNA           TCGA-3C-AAAU-10A-01D
## 3                            DNA           TCGA-3C-AAAU-01A-11D
## 4                            DNA           TCGA-3C-AAAU-01A-11D
## 5                            DNA           TCGA-3C-AAAU-01A-11D
## 6                            RNA           TCGA-3C-AAAU-01A-11R
##   portions.analytes.analyte_type_id
## 1                                 D
## 2                                 D
## 3                                 D
## 4                                 D
## 5                                 D
## 6                                 R
##   portions.analytes.aliquots.analyte_type
## 1                                    <NA>
## 2                                    <NA>
## 3                                    <NA>
## 4                                    <NA>
## 5                                    <NA>
## 6                                    <NA>
##   portions.analytes.aliquots.submitter_id
## 1            TCGA-3C-AAAU-10A-01D-A41F-09
## 2            TCGA-3C-AAAU-10A-01D-A41E-01
## 3            TCGA-3C-AAAU-01A-11D-A41E-01
## 4            TCGA-3C-AAAU-01A-11D-A41F-09
## 5            TCGA-3C-AAAU-01A-11D-A41Q-05
## 6            TCGA-3C-AAAU-01A-11R-A41B-07
```

Similarly to download and combine metadata from multiple TCGA project one can use:

```
#download metadata of following projects into a single dataframe
tcgaProjList<-c("TCGA-BLCA","TCGA-HNSC","TCGA-ESCA","TCGA-PRAD")
#mdList will have all metadata for tcgaProjList
mdListDF<-data.frame()
for(s in tcgaProjList){
  #mdList<-c(mdList,getjoinedBiospcClinc(s))
  if(dim(mdListDF)[1]<1){
    mdListDF<-getjoinedBiospcClinc(s)
  }else{
    print("joining")
```

```
    temp<-getjoinedBiospcClinc(s)
    mdListDF<-bind_rows(mdListDF,temp)
  }
}
```

Now we can use the downloaded metadata in subsequent analysis.

Downloading Mutation Data