bassez-2021 metadata

```r
globus_metadata <- read.table("/mnt/nmorais-nfs/marta/pC_myeinfobank/metadata/globus/bassez_metadata.tx
```

```r
globus_barcodes <- read.table("/mnt/nmorais-nfs/marta/pC_myeinfobank/metadata/globus/bassez_barcodes.tx
```

```r
globus_metadata[1:5,]
```

```
##          V1        V2          V3                                   V4       V5
## 1 orig.ident nCount_RNA nFeature_RNA                              Cell timepoint
## 2     BIOKEY     684.0         430 BIOKEY_13_Pre_AAACCTGCAACAACCT-1       Pre
## 3     BIOKEY    1252.0         700 BIOKEY_13_Pre_AAACCTGCAAGAAGAG-1       Pre
## 4     BIOKEY     522.0         330 BIOKEY_13_Pre_AAACCTGGTCTCCACT-1       Pre
## 5     BIOKEY    8454.0        2637 BIOKEY_13_Pre_AAACCTGTCAACGAAA-1       Pre
##          V6               V7                 V8     V9          V10
## 1 expansion           cohort       percent_mito  batch    cell_type
## 2       n/a treatment_naive 2.4853801169590644 Bassez Myeloid_cell
## 3       n/a treatment_naive  4.073482428115016 Bassez       T_cell
## 4       n/a treatment_naive 0.7662835249042145 Bassez Myeloid_cell
## 5       n/a treatment_naive  4.222853087295954 Bassez Myeloid_cell
##                         V11     V12       V13    V14    V15
## 1                 condition     sex   patient tissue  ident
## 2 Hormone_receptor_positive unknown BIOKEY_13 breast BIOKEY
## 3 Hormone_receptor_positive unknown BIOKEY_13 breast BIOKEY
## 4 Hormone_receptor_positive unknown BIOKEY_13 breast BIOKEY
## 5 Hormone_receptor_positive unknown BIOKEY_13 breast BIOKEY
```

```r
globus_barcodes[1:5,]
```

```
## [1] "BIOKEY_13_Pre_AAACCTGCAACAACCT-1" "BIOKEY_13_Pre_AAACCTGCAAGAAGAG-1"
## [3] "BIOKEY_13_Pre_AAACCTGGTCTCCACT-1" "BIOKEY_13_Pre_AAACCTGTCAACGAAA-1"
## [5] "BIOKEY_13_Pre_AAACGGGAGAGTAAGG-1"
```

```r
metadata <- data.frame(Cell_ID = globus_metadata$V4[-1],
                       Treatment_Status = globus_metadata$V7[-1],
                       Subtype = globus_metadata$V11[-1],
                       Patient = globus_metadata$V13[-1]
                       )
```

```r
cohort1 <- read.table("1872-BIOKEY_metaData_cohort1_web.csv", sep = ",")
cohort2 <- read.table("1871-BIOKEY_metaData_cohort2_web.csv", sep = ",")
```

```r
cohort1[1:5,]
```

```
##                                 V1         V2           V3         V4        V5
## 1                             Cell nCount_RNA nFeature_RNA patient_id timepoint
## 2 BIOKEY_13_Pre_AAACCTGCAACAACCT-1        684          430  BIOKEY_13       Pre
## 3 BIOKEY_13_Pre_AAACCTGCAAGAAGAG-1       1252          700  BIOKEY_13       Pre
## 4 BIOKEY_13_Pre_AAACCTGGTCTCCACT-1        522          330  BIOKEY_13       Pre
## 5 BIOKEY_13_Pre_AAACCTGTCAACGAAA-1       8454         2637  BIOKEY_13       Pre
##        V6     V7        V8             V9
```

```
## 1 expansion BC_type    cellType           cohort
## 2      n/a   HER2+ Myeloid_cell treatment_naive
## 3      n/a   HER2+       T_cell treatment_naive
## 4      n/a   HER2+          pDC treatment_naive
## 5      n/a   HER2+ Myeloid_cell treatment_naive
```

```
cohort2[1:5,]
```

```
##                                      V1         V2            V3         V4        V5
## 1                                  Cell nCount_RNA nFeature_RNA patient_id timepoint
## 2 BIOKEY_33_Pre_AAACCTGAGAGACTTA-1       3911         1665  BIOKEY_33       Pre
## 3 BIOKEY_33_Pre_AAACCTGAGTAGCGGT-1        605          491  BIOKEY_33       Pre
## 4 BIOKEY_33_Pre_AAACCTGCATGGTAGG-1        596          461  BIOKEY_33       Pre
## 5 BIOKEY_33_Pre_AAACCTGGTATAGGGC-1       2983         1615  BIOKEY_33       Pre
##           V6      V7       V8                V9
## 1 expansion BC_type    cellType            cohort
## 2         E    TNBC      T_cell neoadjuvant_chemo
## 3         E    TNBC Fibroblast neoadjuvant_chemo
## 4         E    TNBC      T_cell neoadjuvant_chemo
## 5         E    TNBC Fibroblast neoadjuvant_chemo
```

```
table(metadata$Treatment_Status)
```

```
##
## neoadjuvant_chemo    treatment_naive
##            50656             175758
```

```
metadata$Treatment_Status[metadata$Treatment_Status == "treatment_naive"] <- "Naive"
metadata$Treatment_Status[metadata$Treatment_Status == "neoadjuvant_chemo"] <- "Treated"
table(metadata$Treatment_Status)
```

```
##
##   Naive Treated
##  175758   50656
```

```
table(metadata$Subtype)
```

```
##
## Hormone_receptor_positive                    TNBC
##                    107701                  118713
```

```
table(cohort1$V7)
```

```
##
## BC_type      ER+    HER2+     TNBC
##       1    80408    13414    82120
```

```
table(cohort2$V7)
```

```
##
## BC_type      ER+    HER2+     TNBC
##       1    10679     3293    36721
```

```
subtypes <- c(cohort1$V7[-1], cohort2$V7[-1])
table(subtypes)
```

```
## subtypes
##    ER+  HER2+   TNBC
##  91087  16707 118841
```

```r
barcodes <- c(cohort1$V1[-1], cohort2$V1[-1])
barcodes[1:5]
```

```
## [1] "BIOKEY_13_Pre_AAACCTGCAACAACCT-1" "BIOKEY_13_Pre_AAACCTGCAAGAAGAG-1"
## [3] "BIOKEY_13_Pre_AAACCTGGTCTCCACT-1" "BIOKEY_13_Pre_AAACCTGTCAACGAAA-1"
## [5] "BIOKEY_13_Pre_AAACGGGAGAGTAAGG-1"
```

```r
subtypes <- subtypes[order(barcodes)]
barcodes <- barcodes[order(barcodes)]
barcodes[1:5]
```

```
## [1] "BIOKEY_1_On_AAACCTGGTAGCAAAT-1" "BIOKEY_1_On_AAACCTGGTCATACTG-1"
## [3] "BIOKEY_1_On_AAACCTGGTCTAGGTT-1" "BIOKEY_1_On_AAACCTGGTTGGACCC-1"
## [5] "BIOKEY_1_On_AAACCTGTCACGATGT-1"
```

```r
metadata <- metadata[order(metadata$Cell_ID),]

subtypes <- subtypes[barcodes %in% metadata$Cell_ID]
barcodes <- barcodes[barcodes %in% metadata$Cell_ID]

identical(barcodes,metadata$Cell_ID)
```

```
## [1] TRUE
```

```r
metadata$Subtype <- subtypes
table(metadata$Subtype)
```

```
##
##    ER+  HER2+   TNBC
##  91012  16689 118713
```

```r
table(metadata$Patient, metadata$Subtype)
```

```
##
##               ER+ HER2+  TNBC
##    BIOKEY_1     0     0  9785
##    BIOKEY_10    0     0 10769
##    BIOKEY_11    0     0  4638
##    BIOKEY_12 10077     0     0
##    BIOKEY_13    0  4241     0
##    BIOKEY_14    0     0  4244
##    BIOKEY_15    0     0  8381
##    BIOKEY_16    0     0  9941
##    BIOKEY_17  6391     0     0
##    BIOKEY_18  5222     0     0
##    BIOKEY_19    0     0  7406
##    BIOKEY_2     0     0  6146
##    BIOKEY_20  3304     0     0
##    BIOKEY_21  6195     0     0
##    BIOKEY_22  2923     0     0
##    BIOKEY_23    0  4873     0
##    BIOKEY_24  4420     0     0
##    BIOKEY_25    0     0   777
##    BIOKEY_26    0     0  7535
##    BIOKEY_27  4509     0     0
##    BIOKEY_28    0  4282     0
##    BIOKEY_29  1375     0     0
```

```
##    BIOKEY_3  6128    0     0
##    BIOKEY_30 6820    0     0
##    BIOKEY_31    0    0  6786
##    BIOKEY_32 1374    0     0
##    BIOKEY_33    0    0  6237
##    BIOKEY_34    0    0  3367
##    BIOKEY_35    0    0 12002
##    BIOKEY_36    0    0  7213
##    BIOKEY_37 3953    0     0
##    BIOKEY_38    0 3293     0
##    BIOKEY_39    0    0  2326
##    BIOKEY_4  8256    0     0
##    BIOKEY_40 2712    0     0
##    BIOKEY_41    0    0  5549
##    BIOKEY_42 2630    0     0
##    BIOKEY_5  4013    0     0
##    BIOKEY_6  7810    0     0
##    BIOKEY_7  2900    0     0
##    BIOKEY_8     0    0  1535
##    BIOKEY_9     0    0  4076
```

```r
table(cohort1$V4[-1], cohort1$V7[-1])
```

```
##
##               ER+ HER2+  TNBC
##    BIOKEY_1     0    0  9789
##    BIOKEY_10    0    0 10796
##    BIOKEY_11    0    0  4656
##    BIOKEY_12 10078    0     0
##    BIOKEY_13    0 4241     0
##    BIOKEY_14    0    0  4247
##    BIOKEY_15    0    0  8393
##    BIOKEY_16    0    0  9945
##    BIOKEY_17 6391    0     0
##    BIOKEY_18 5222    0     0
##    BIOKEY_19    0    0  7410
##    BIOKEY_2     0    0  6147
##    BIOKEY_20 3306    0     0
##    BIOKEY_21 6209    0     0
##    BIOKEY_22 2925    0     0
##    BIOKEY_23    0 4878     0
##    BIOKEY_24 4423    0     0
##    BIOKEY_25    0    0   777
##    BIOKEY_26    0    0  7559
##    BIOKEY_27 4510    0     0
##    BIOKEY_28    0 4295     0
##    BIOKEY_29 1375    0     0
##    BIOKEY_3  6133    0     0
##    BIOKEY_30 6821    0     0
##    BIOKEY_31    0    0  6788
##    BIOKEY_4  8261    0     0
##    BIOKEY_5  4014    0     0
##    BIOKEY_6  7810    0     0
##    BIOKEY_7  2930    0     0
##    BIOKEY_8     0    0  1537
```

```
##    BIOKEY_9      0     0  4076
```

```
metadata$Dissociation <- "Mechanical,Enzymatic"
metadata$Protocol <- "10X_chromium_5'"
metadata$Technology <- "10X_Genomics"
metadata$Sample_Type <- "Fresh"
metadata$CellOrNucleus <- "Cell"
metadata$Sequencing_Machine <- "Illumina_NextSeq/NovaSeq6000"
metadata$Preprocessing <- "CellRanger"
metadata$Reference_Genome <- "GRCh38"
```

```
metadata$Treatment_Type <- metadata$Treatment_Status
metadata$Treatment_Type[metadata$Treatment_Status == "Treated"] <- "Pembrolizumab"
metadata$Treatment_Type[metadata$Treatment_Status == "Naive"] <- "None"
metadata$Treatment_Type[metadata$Treatment_Status == "Treated" &
                        metadata$Subtype == "HER2+" ] <- "Pembrolizumab, anti-HER2"
table(metadata$Treatment_Type, metadata$Treatment_Status)
```

```
##
##                          Naive Treated
##    None                 175758       0
##    Pembrolizumab             0   47363
##    Pembrolizumab, anti-HER2  0    3293
```

```
table(metadata$Treatment_Type, metadata$Subtype)
```

```
##
##                           ER+ HER2+  TNBC
##    None                 80343 13396 82019
##    Pembrolizumab        10669     0 36694
##    Pembrolizumab, anti-HER2  0  3293     0
```

```
rownames(metadata) <- NULL
```

```
metadata$Age_Category <- metadata$Patient
metadata$Age_Category[metadata$Patient == "BIOKEY_1"] <- "30-40"
metadata$Age_Category[metadata$Patient == "BIOKEY_10"] <- "41-50"
metadata$Age_Category[metadata$Patient == "BIOKEY_11"] <- "51-60"
metadata$Age_Category[metadata$Patient == "BIOKEY_12"] <- "61-70"
metadata$Age_Category[metadata$Patient == "BIOKEY_13"] <- "41-50"
metadata$Age_Category[metadata$Patient == "BIOKEY_14"] <- "81-90"
metadata$Age_Category[metadata$Patient == "BIOKEY_15"] <- "41-50"
metadata$Age_Category[metadata$Patient == "BIOKEY_16"] <- "51-60"
metadata$Age_Category[metadata$Patient == "BIOKEY_17"] <- "41-50"
metadata$Age_Category[metadata$Patient == "BIOKEY_18"] <- "71-80"
metadata$Age_Category[metadata$Patient == "BIOKEY_19"] <- "61-70"
metadata$Age_Category[metadata$Patient == "BIOKEY_2"] <- "71-80"
metadata$Age_Category[metadata$Patient == "BIOKEY_20"] <- "51-60"
metadata$Age_Category[metadata$Patient == "BIOKEY_21"] <- "41-50"
metadata$Age_Category[metadata$Patient == "BIOKEY_22"] <- "51-60"
metadata$Age_Category[metadata$Patient == "BIOKEY_23"] <- "71-80"
metadata$Age_Category[metadata$Patient == "BIOKEY_24"] <- "41-50"
metadata$Age_Category[metadata$Patient == "BIOKEY_25"] <- "61-70"
metadata$Age_Category[metadata$Patient == "BIOKEY_26"] <- "71-80"
metadata$Age_Category[metadata$Patient == "BIOKEY_27"] <- "61-70"
metadata$Age_Category[metadata$Patient == "BIOKEY_28"] <- "61-70"
```

```r
metadata$Age_Category[metadata$Patient == "BIOKEY_29"] <- "51-60"
metadata$Age_Category[metadata$Patient == "BIOKEY_3"] <- "41-50"
metadata$Age_Category[metadata$Patient == "BIOKEY_30"] <- "41-50"
metadata$Age_Category[metadata$Patient == "BIOKEY_31"] <- "61-70"
metadata$Age_Category[metadata$Patient == "BIOKEY_32"] <- "61-70"
metadata$Age_Category[metadata$Patient == "BIOKEY_33"] <- "30-40"
metadata$Age_Category[metadata$Patient == "BIOKEY_34"] <- "51-60"
metadata$Age_Category[metadata$Patient == "BIOKEY_35"] <- "30-40"
metadata$Age_Category[metadata$Patient == "BIOKEY_36"] <- "71-80"
metadata$Age_Category[metadata$Patient == "BIOKEY_37"] <- "51-60"
metadata$Age_Category[metadata$Patient == "BIOKEY_38"] <- "51-60"
metadata$Age_Category[metadata$Patient == "BIOKEY_39"] <- "30-40"
metadata$Age_Category[metadata$Patient == "BIOKEY_4"] <- "41-50"
metadata$Age_Category[metadata$Patient == "BIOKEY_40"] <- "51-60"
metadata$Age_Category[metadata$Patient == "BIOKEY_41"] <- "41-50"
metadata$Age_Category[metadata$Patient == "BIOKEY_42"] <- "61-70"
metadata$Age_Category[metadata$Patient == "BIOKEY_5"] <- "61-70"
metadata$Age_Category[metadata$Patient == "BIOKEY_6"] <- "51-60"
metadata$Age_Category[metadata$Patient == "BIOKEY_7"] <- "41-50"
metadata$Age_Category[metadata$Patient == "BIOKEY_8"] <- "71-80"
metadata$Age_Category[metadata$Patient == "BIOKEY_9"] <- "41-50"
```

```r
table(metadata$Age_Category, metadata$Patient)
```

```
##
##          BIOKEY_1 BIOKEY_10 BIOKEY_11 BIOKEY_12 BIOKEY_13 BIOKEY_14 BIOKEY_15
##   30-40      9785         0         0         0         0         0         0
##   41-50         0     10769         0         0      4241         0      8381
##   51-60         0         0      4638         0         0         0         0
##   61-70         0         0         0     10077         0         0         0
##   71-80         0         0         0         0         0         0         0
##   81-90         0         0         0         0         0      4244         0
##
##          BIOKEY_16 BIOKEY_17 BIOKEY_18 BIOKEY_19 BIOKEY_2 BIOKEY_20 BIOKEY_21
##   30-40          0         0         0         0        0         0         0
##   41-50          0      6391         0         0        0         0      6195
##   51-60       9941         0         0         0        0      3304         0
##   61-70          0         0         0      7406        0         0         0
##   71-80          0         0      5222         0     6146         0         0
##   81-90          0         0         0         0        0         0         0
##
##          BIOKEY_22 BIOKEY_23 BIOKEY_24 BIOKEY_25 BIOKEY_26 BIOKEY_27 BIOKEY_28
##   30-40          0         0         0         0         0         0         0
##   41-50          0         0      4420         0         0         0         0
##   51-60       2923         0         0         0         0         0         0
##   61-70          0         0         0       777         0      4509      4282
##   71-80          0      4873         0         0      7535         0         0
##   81-90          0         0         0         0         0         0         0
##
##          BIOKEY_29 BIOKEY_3 BIOKEY_30 BIOKEY_31 BIOKEY_32 BIOKEY_33 BIOKEY_34
##   30-40          0        0         0         0         0      6237         0
##   41-50          0     6128      6820         0         0         0         0
##   51-60       1375        0         0         0         0         0      3367
##   61-70          0        0         0      6786      1374         0         0
```

6

```
##    71-80         0         0         0         0         0         0         0
##    81-90         0         0         0         0         0         0         0
##
##           BIOKEY_35 BIOKEY_36 BIOKEY_37 BIOKEY_38 BIOKEY_39 BIOKEY_4 BIOKEY_40
##    30-40     12002         0         0         0      2326        0         0
##    41-50         0         0         0         0         0     8256         0
##    51-60         0         0      3953      3293         0        0      2712
##    61-70         0         0         0         0         0        0         0
##    71-80         0      7213         0         0         0        0         0
##    81-90         0         0         0         0         0        0         0
##
##           BIOKEY_41 BIOKEY_42 BIOKEY_5 BIOKEY_6 BIOKEY_7 BIOKEY_8 BIOKEY_9
##    30-40         0         0        0        0        0        0        0
##    41-50      5549         0        0        0     2900        0     4076
##    51-60         0         0        0     7810        0        0        0
##    61-70         0      2630     4013        0        0        0        0
##    71-80         0         0        0        0        0     1535        0
##    81-90         0         0        0        0        0        0        0
```

```r
metadata$Menopause <- metadata$Patient
metadata$Menopause[metadata$Patient == "BIOKEY_1"] <- "Pre"
metadata$Menopause[metadata$Patient == "BIOKEY_10"] <- "Pre"
metadata$Menopause[metadata$Patient == "BIOKEY_11"] <- "Pre"
metadata$Menopause[metadata$Patient == "BIOKEY_12"] <- "Post"
metadata$Menopause[metadata$Patient == "BIOKEY_13"] <- "Pre"
metadata$Menopause[metadata$Patient == "BIOKEY_14"] <- "Post"
metadata$Menopause[metadata$Patient == "BIOKEY_15"] <- "Pre"
metadata$Menopause[metadata$Patient == "BIOKEY_16"] <- "Post"
metadata$Menopause[metadata$Patient == "BIOKEY_17"] <- "Pre"
metadata$Menopause[metadata$Patient == "BIOKEY_18"] <- "Post"
metadata$Menopause[metadata$Patient == "BIOKEY_19"] <- "Post"
metadata$Menopause[metadata$Patient == "BIOKEY_2"] <- "Pre"
metadata$Menopause[metadata$Patient == "BIOKEY_20"] <- "Post"
metadata$Menopause[metadata$Patient == "BIOKEY_21"] <- "Post"
metadata$Menopause[metadata$Patient == "BIOKEY_22"] <- "Post"
metadata$Menopause[metadata$Patient == "BIOKEY_23"] <- "Pre"
metadata$Menopause[metadata$Patient == "BIOKEY_24"] <- "Pre"
metadata$Menopause[metadata$Patient == "BIOKEY_25"] <- "Post"
metadata$Menopause[metadata$Patient == "BIOKEY_26"] <- "Post"
metadata$Menopause[metadata$Patient == "BIOKEY_27"] <- "Pre"
metadata$Menopause[metadata$Patient == "BIOKEY_28"] <- "Post"
metadata$Menopause[metadata$Patient == "BIOKEY_29"] <- "Pre"
metadata$Menopause[metadata$Patient == "BIOKEY_3"] <- "Pre"
metadata$Menopause[metadata$Patient == "BIOKEY_30"] <- "Post"
metadata$Menopause[metadata$Patient == "BIOKEY_31"] <- "Post"
metadata$Menopause[metadata$Patient == "BIOKEY_32"] <- "Pre"
metadata$Menopause[metadata$Patient == "BIOKEY_33"] <- "Pre"
metadata$Menopause[metadata$Patient == "BIOKEY_34"] <- "Pre"
metadata$Menopause[metadata$Patient == "BIOKEY_35"] <- "Post"
metadata$Menopause[metadata$Patient == "BIOKEY_36"] <- "Post"
metadata$Menopause[metadata$Patient == "BIOKEY_37"] <- "Post"
metadata$Menopause[metadata$Patient == "BIOKEY_38"] <- "Post"
metadata$Menopause[metadata$Patient == "BIOKEY_39"] <- "Pre"
metadata$Menopause[metadata$Patient == "BIOKEY_4"] <- "Post"
```

```
metadata$Menopause[metadata$Patient == "BIOKEY_40"] <- "Post"
metadata$Menopause[metadata$Patient == "BIOKEY_41"] <- "Post"
metadata$Menopause[metadata$Patient == "BIOKEY_42"] <- "Pre"
metadata$Menopause[metadata$Patient == "BIOKEY_5"] <- "Post"
metadata$Menopause[metadata$Patient == "BIOKEY_6"] <- "Post"
metadata$Menopause[metadata$Patient == "BIOKEY_7"] <- "Pre"
metadata$Menopause[metadata$Patient == "BIOKEY_8"] <- "Pre"
metadata$Menopause[metadata$Patient == "BIOKEY_9"] <- "Post"
```

```
metadata$Cancer_Type <- metadata$Patient
metadata$Cancer_Type[metadata$Patient == "BIOKEY_1"] <- "IBC-NST"
metadata$Cancer_Type[metadata$Patient == "BIOKEY_10"] <- "IBC-NST"
metadata$Cancer_Type[metadata$Patient == "BIOKEY_11"] <- "ST"
metadata$Cancer_Type[metadata$Patient == "BIOKEY_12"] <- "IBC-NST"
metadata$Cancer_Type[metadata$Patient == "BIOKEY_13"] <- "IBC-NST"
metadata$Cancer_Type[metadata$Patient == "BIOKEY_14"] <- "IBC-NST"
metadata$Cancer_Type[metadata$Patient == "BIOKEY_15"] <- "IBC-NST"
metadata$Cancer_Type[metadata$Patient == "BIOKEY_16"] <- "IBC-NST"
metadata$Cancer_Type[metadata$Patient == "BIOKEY_17"] <- "IBC-NST"
metadata$Cancer_Type[metadata$Patient == "BIOKEY_18"] <- "ST"
metadata$Cancer_Type[metadata$Patient == "BIOKEY_19"] <- "ST"
metadata$Cancer_Type[metadata$Patient == "BIOKEY_2"] <- "IBC-NST"
metadata$Cancer_Type[metadata$Patient == "BIOKEY_20"] <- "IBC-NST"
metadata$Cancer_Type[metadata$Patient == "BIOKEY_21"] <- "IBC-NST"
metadata$Cancer_Type[metadata$Patient == "BIOKEY_22"] <- "IBC-NST"
metadata$Cancer_Type[metadata$Patient == "BIOKEY_23"] <- "IBC-NST"
metadata$Cancer_Type[metadata$Patient == "BIOKEY_24"] <- "IBC-NST"
metadata$Cancer_Type[metadata$Patient == "BIOKEY_25"] <- "IBC-NST"
metadata$Cancer_Type[metadata$Patient == "BIOKEY_26"] <- "IBC-NST"
metadata$Cancer_Type[metadata$Patient == "BIOKEY_27"] <- "IBC-NST"
metadata$Cancer_Type[metadata$Patient == "BIOKEY_28"] <- "IBC-NST"
metadata$Cancer_Type[metadata$Patient == "BIOKEY_29"] <- "IBC-NST"
metadata$Cancer_Type[metadata$Patient == "BIOKEY_3"] <- "IBC-NST"
metadata$Cancer_Type[metadata$Patient == "BIOKEY_30"] <- "ST"
metadata$Cancer_Type[metadata$Patient == "BIOKEY_31"] <- "IBC-NST"
metadata$Cancer_Type[metadata$Patient == "BIOKEY_32"] <- "IBC-NST"
metadata$Cancer_Type[metadata$Patient == "BIOKEY_33"] <- "IBC-NST"
metadata$Cancer_Type[metadata$Patient == "BIOKEY_34"] <- "IBC-NST"
metadata$Cancer_Type[metadata$Patient == "BIOKEY_35"] <- "IBC-NST"
metadata$Cancer_Type[metadata$Patient == "BIOKEY_36"] <- "IBC-NST"
metadata$Cancer_Type[metadata$Patient == "BIOKEY_37"] <- "IBC-NST"
metadata$Cancer_Type[metadata$Patient == "BIOKEY_38"] <- "IBC-NST"
metadata$Cancer_Type[metadata$Patient == "BIOKEY_39"] <- "IBC-NST"
metadata$Cancer_Type[metadata$Patient == "BIOKEY_4"] <- "IBC-NST"
metadata$Cancer_Type[metadata$Patient == "BIOKEY_40"] <- "IBC-NST"
metadata$Cancer_Type[metadata$Patient == "BIOKEY_41"] <- "IBC-NST"
metadata$Cancer_Type[metadata$Patient == "BIOKEY_42"] <- "IBC-NST"
metadata$Cancer_Type[metadata$Patient == "BIOKEY_5"] <- "IBC-NST"
metadata$Cancer_Type[metadata$Patient == "BIOKEY_6"] <- "IBC-NST"
metadata$Cancer_Type[metadata$Patient == "BIOKEY_7"] <- "IBC-NST"
metadata$Cancer_Type[metadata$Patient == "BIOKEY_8"] <- "IBC-NST"
metadata$Cancer_Type[metadata$Patient == "BIOKEY_9"] <- "IBC-NST"
```

```
table(metadata$Menopause, metadata$Patient)
```

```
## 
##        BIOKEY_1 BIOKEY_10 BIOKEY_11 BIOKEY_12 BIOKEY_13 BIOKEY_14 BIOKEY_15
##   Post        0         0         0     10077         0      4244         0
##   Pre      9785     10769      4638         0      4241         0      8381
## 
##        BIOKEY_16 BIOKEY_17 BIOKEY_18 BIOKEY_19 BIOKEY_2 BIOKEY_20 BIOKEY_21
##   Post      9941         0      5222      7406        0      3304      6195
##   Pre          0      6391         0         0     6146         0         0
## 
##        BIOKEY_22 BIOKEY_23 BIOKEY_24 BIOKEY_25 BIOKEY_26 BIOKEY_27 BIOKEY_28
##   Post      2923         0         0       777      7535         0      4282
##   Pre          0      4873      4420         0         0      4509         0
## 
##        BIOKEY_29 BIOKEY_3 BIOKEY_30 BIOKEY_31 BIOKEY_32 BIOKEY_33 BIOKEY_34
##   Post         0        0      6820      6786         0         0         0
##   Pre       1375     6128         0         0      1374      6237      3367
## 
##        BIOKEY_35 BIOKEY_36 BIOKEY_37 BIOKEY_38 BIOKEY_39 BIOKEY_4 BIOKEY_40
##   Post     12002      7213      3953      3293         0     8256      2712
##   Pre          0         0         0         0      2326        0         0
## 
##        BIOKEY_41 BIOKEY_42 BIOKEY_5 BIOKEY_6 BIOKEY_7 BIOKEY_8 BIOKEY_9
##   Post      5549         0     4013     7810        0        0     4076
##   Pre          0      2630        0        0     2900     1535        0
```

```
metadata$Grade <- metadata$Patient
metadata$Grade[metadata$Patient == "BIOKEY_1"] <- "3"
metadata$Grade[metadata$Patient == "BIOKEY_10"] <- "3"
metadata$Grade[metadata$Patient == "BIOKEY_11"] <- "3"
metadata$Grade[metadata$Patient == "BIOKEY_12"] <- "3"
metadata$Grade[metadata$Patient == "BIOKEY_13"] <- "3"
metadata$Grade[metadata$Patient == "BIOKEY_14"] <- "3"
metadata$Grade[metadata$Patient == "BIOKEY_15"] <- "3"
metadata$Grade[metadata$Patient == "BIOKEY_16"] <- "3"
metadata$Grade[metadata$Patient == "BIOKEY_17"] <- "3"
metadata$Grade[metadata$Patient == "BIOKEY_18"] <- "2"
metadata$Grade[metadata$Patient == "BIOKEY_19"] <- "2"
metadata$Grade[metadata$Patient == "BIOKEY_2"] <- "3"
metadata$Grade[metadata$Patient == "BIOKEY_20"] <- "3"
metadata$Grade[metadata$Patient == "BIOKEY_21"] <- "3"
metadata$Grade[metadata$Patient == "BIOKEY_22"] <- "3"
metadata$Grade[metadata$Patient == "BIOKEY_23"] <- "3"
metadata$Grade[metadata$Patient == "BIOKEY_24"] <- "3"
metadata$Grade[metadata$Patient == "BIOKEY_25"] <- "3"
metadata$Grade[metadata$Patient == "BIOKEY_26"] <- "3"
metadata$Grade[metadata$Patient == "BIOKEY_27"] <- "3"
metadata$Grade[metadata$Patient == "BIOKEY_28"] <- "3"
metadata$Grade[metadata$Patient == "BIOKEY_29"] <- "3"
metadata$Grade[metadata$Patient == "BIOKEY_3"] <- "3"
metadata$Grade[metadata$Patient == "BIOKEY_30"] <- "2"
metadata$Grade[metadata$Patient == "BIOKEY_31"] <- "3"
metadata$Grade[metadata$Patient == "BIOKEY_32"] <- "3"
```

```r
metadata$Grade[metadata$Patient == "BIOKEY_33"] <- "3"
metadata$Grade[metadata$Patient == "BIOKEY_34"] <- "3"
metadata$Grade[metadata$Patient == "BIOKEY_35"] <- "2"
metadata$Grade[metadata$Patient == "BIOKEY_36"] <- "3"
metadata$Grade[metadata$Patient == "BIOKEY_37"] <- "1"
metadata$Grade[metadata$Patient == "BIOKEY_38"] <- "3"
metadata$Grade[metadata$Patient == "BIOKEY_39"] <- "3"
metadata$Grade[metadata$Patient == "BIOKEY_4"] <- "3"
metadata$Grade[metadata$Patient == "BIOKEY_40"] <- "2"
metadata$Grade[metadata$Patient == "BIOKEY_41"] <- "3"
metadata$Grade[metadata$Patient == "BIOKEY_42"] <- "3"
metadata$Grade[metadata$Patient == "BIOKEY_5"] <- "3"
metadata$Grade[metadata$Patient == "BIOKEY_6"] <- "2"
metadata$Grade[metadata$Patient == "BIOKEY_7"] <- "3"
metadata$Grade[metadata$Patient == "BIOKEY_8"] <- "1"
metadata$Grade[metadata$Patient == "BIOKEY_9"] <- "3"
```

```r
metadata$Primary_Tumour <- metadata$Patient
metadata$Primary_Tumour[metadata$Patient == "BIOKEY_1"] <- "pT2"
metadata$Primary_Tumour[metadata$Patient == "BIOKEY_10"] <- "pT2"
metadata$Primary_Tumour[metadata$Patient == "BIOKEY_11"] <- "pT1"
metadata$Primary_Tumour[metadata$Patient == "BIOKEY_12"] <- "pT2"
metadata$Primary_Tumour[metadata$Patient == "BIOKEY_13"] <- "pT1"
metadata$Primary_Tumour[metadata$Patient == "BIOKEY_14"] <- "pT1"
metadata$Primary_Tumour[metadata$Patient == "BIOKEY_15"] <- "pT1"
metadata$Primary_Tumour[metadata$Patient == "BIOKEY_16"] <- "pT3"
metadata$Primary_Tumour[metadata$Patient == "BIOKEY_17"] <- "pT2"
metadata$Primary_Tumour[metadata$Patient == "BIOKEY_18"] <- "pT3"
metadata$Primary_Tumour[metadata$Patient == "BIOKEY_19"] <- "pT2"
metadata$Primary_Tumour[metadata$Patient == "BIOKEY_2"] <- "pT2"
metadata$Primary_Tumour[metadata$Patient == "BIOKEY_20"] <- "pT2"
metadata$Primary_Tumour[metadata$Patient == "BIOKEY_21"] <- "pT3"
metadata$Primary_Tumour[metadata$Patient == "BIOKEY_22"] <- "pT1"
metadata$Primary_Tumour[metadata$Patient == "BIOKEY_23"] <- "pT2"
metadata$Primary_Tumour[metadata$Patient == "BIOKEY_24"] <- "pT2"
metadata$Primary_Tumour[metadata$Patient == "BIOKEY_25"] <- "pT2"
metadata$Primary_Tumour[metadata$Patient == "BIOKEY_26"] <- "pT3"
metadata$Primary_Tumour[metadata$Patient == "BIOKEY_27"] <- "pT2"
metadata$Primary_Tumour[metadata$Patient == "BIOKEY_28"] <- "pT3"
metadata$Primary_Tumour[metadata$Patient == "BIOKEY_29"] <- "pT2"
metadata$Primary_Tumour[metadata$Patient == "BIOKEY_3"] <- "pT2"
metadata$Primary_Tumour[metadata$Patient == "BIOKEY_30"] <- "pT2"
metadata$Primary_Tumour[metadata$Patient == "BIOKEY_31"] <- "pT1"
metadata$Primary_Tumour[metadata$Patient == "BIOKEY_32"] <- "ypT2"
metadata$Primary_Tumour[metadata$Patient == "BIOKEY_33"] <- "ypT1"
metadata$Primary_Tumour[metadata$Patient == "BIOKEY_34"] <- "ypT1"
metadata$Primary_Tumour[metadata$Patient == "BIOKEY_35"] <- "ypT2"
metadata$Primary_Tumour[metadata$Patient == "BIOKEY_36"] <- "ypT1"
metadata$Primary_Tumour[metadata$Patient == "BIOKEY_37"] <- "ypT3"
metadata$Primary_Tumour[metadata$Patient == "BIOKEY_38"] <- "ypT0"
metadata$Primary_Tumour[metadata$Patient == "BIOKEY_39"] <- "ypT2"
metadata$Primary_Tumour[metadata$Patient == "BIOKEY_4"] <- "pT2"
metadata$Primary_Tumour[metadata$Patient == "BIOKEY_40"] <- "ypT2"
metadata$Primary_Tumour[metadata$Patient == "BIOKEY_41"] <- "ypT1"
```

```r
metadata$Primary_Tumour[metadata$Patient == "BIOKEY_42"] <- "ypT2"
metadata$Primary_Tumour[metadata$Patient == "BIOKEY_5"] <- "pT2"
metadata$Primary_Tumour[metadata$Patient == "BIOKEY_6"] <- "pT3"
metadata$Primary_Tumour[metadata$Patient == "BIOKEY_7"] <- "pT2"
metadata$Primary_Tumour[metadata$Patient == "BIOKEY_8"] <- "rpT2"
metadata$Primary_Tumour[metadata$Patient == "BIOKEY_9"] <- "pT1"

metadata$Nodes <- metadata$Patient
metadata$Nodes[metadata$Patient == "BIOKEY_1"] <- "pN1"
metadata$Nodes[metadata$Patient == "BIOKEY_10"] <- "pN0"
metadata$Nodes[metadata$Patient == "BIOKEY_11"] <- "pN0"
metadata$Nodes[metadata$Patient == "BIOKEY_12"] <- "pN0"
metadata$Nodes[metadata$Patient == "BIOKEY_13"] <- "pN0"
metadata$Nodes[metadata$Patient == "BIOKEY_14"] <- "pN0"
metadata$Nodes[metadata$Patient == "BIOKEY_15"] <- "pN0"
metadata$Nodes[metadata$Patient == "BIOKEY_16"] <- "pN0"
metadata$Nodes[metadata$Patient == "BIOKEY_17"] <- "pN1"
metadata$Nodes[metadata$Patient == "BIOKEY_18"] <- "pN3"
metadata$Nodes[metadata$Patient == "BIOKEY_19"] <- "pN1"
metadata$Nodes[metadata$Patient == "BIOKEY_2"] <- "pN0"
metadata$Nodes[metadata$Patient == "BIOKEY_20"] <- "pN1"
metadata$Nodes[metadata$Patient == "BIOKEY_21"] <- "pN1"
metadata$Nodes[metadata$Patient == "BIOKEY_22"] <- "pN0"
metadata$Nodes[metadata$Patient == "BIOKEY_23"] <- "pN0"
metadata$Nodes[metadata$Patient == "BIOKEY_24"] <- "pN1"
metadata$Nodes[metadata$Patient == "BIOKEY_25"] <- "pN0"
metadata$Nodes[metadata$Patient == "BIOKEY_26"] <- "pN0"
metadata$Nodes[metadata$Patient == "BIOKEY_27"] <- "pN1"
metadata$Nodes[metadata$Patient == "BIOKEY_28"] <- "pN0"
metadata$Nodes[metadata$Patient == "BIOKEY_29"] <- "pN0"
metadata$Nodes[metadata$Patient == "BIOKEY_3"] <- "pN1"
metadata$Nodes[metadata$Patient == "BIOKEY_30"] <- "pN1"
metadata$Nodes[metadata$Patient == "BIOKEY_31"] <- "pN0"
metadata$Nodes[metadata$Patient == "BIOKEY_32"] <- "pN3"
metadata$Nodes[metadata$Patient == "BIOKEY_33"] <- "pN0"
metadata$Nodes[metadata$Patient == "BIOKEY_34"] <- "pN0"
metadata$Nodes[metadata$Patient == "BIOKEY_35"] <- "pN0"
metadata$Nodes[metadata$Patient == "BIOKEY_36"] <- "pN2"
metadata$Nodes[metadata$Patient == "BIOKEY_37"] <- "pN3"
metadata$Nodes[metadata$Patient == "BIOKEY_38"] <- "pN0"
metadata$Nodes[metadata$Patient == "BIOKEY_39"] <- "pN1"
metadata$Nodes[metadata$Patient == "BIOKEY_4"] <- "pN0"
metadata$Nodes[metadata$Patient == "BIOKEY_40"] <- "pN1"
metadata$Nodes[metadata$Patient == "BIOKEY_41"] <- "pN1"
metadata$Nodes[metadata$Patient == "BIOKEY_42"] <- "pN1"
metadata$Nodes[metadata$Patient == "BIOKEY_5"] <- "pN1"
metadata$Nodes[metadata$Patient == "BIOKEY_6"] <- "pN1"
metadata$Nodes[metadata$Patient == "BIOKEY_7"] <- "pN1"
metadata$Nodes[metadata$Patient == "BIOKEY_8"] <- "pNx"
metadata$Nodes[metadata$Patient == "BIOKEY_9"] <- "pN0"

metadata$Gender <- NA
metadata$Age <- NA
metadata$Race <- NA
```

```r
metadata$Menopause <- NA
metadata$Parity <- NA
metadata$Stage <- NA
metadata$Tissue <- "Tumour"

metadata[1:5,]
```

```
##                        Cell_ID Treatment_Status Subtype  Patient
## 1 BIOKEY_1_On_AAACCTGGTAGCAAAT-1            Naive    TNBC BIOKEY_1
## 2 BIOKEY_1_On_AAACCTGGTCATACTG-1            Naive    TNBC BIOKEY_1
## 3 BIOKEY_1_On_AAACCTGGTCTAGGTT-1            Naive    TNBC BIOKEY_1
## 4 BIOKEY_1_On_AAACCTGGTTGGACCC-1            Naive    TNBC BIOKEY_1
## 5 BIOKEY_1_On_AAACCTGTCACGATGT-1            Naive    TNBC BIOKEY_1
##             Dissociation         Protocol   Technology Sample_Type CellOrNucleus
## 1 Mechanical,Enzymatic 10X_chromium_5' 10X_Genomics       Fresh          Cell
## 2 Mechanical,Enzymatic 10X_chromium_5' 10X_Genomics       Fresh          Cell
## 3 Mechanical,Enzymatic 10X_chromium_5' 10X_Genomics       Fresh          Cell
## 4 Mechanical,Enzymatic 10X_chromium_5' 10X_Genomics       Fresh          Cell
## 5 Mechanical,Enzymatic 10X_chromium_5' 10X_Genomics       Fresh          Cell
##             Sequencing_Machine Preprocessing Reference_Genome Treatment_Type
## 1 Illumina_NextSeq/NovaSeq6000     CellRanger           GRCh38           None
## 2 Illumina_NextSeq/NovaSeq6000     CellRanger           GRCh38           None
## 3 Illumina_NextSeq/NovaSeq6000     CellRanger           GRCh38           None
## 4 Illumina_NextSeq/NovaSeq6000     CellRanger           GRCh38           None
## 5 Illumina_NextSeq/NovaSeq6000     CellRanger           GRCh38           None
##   Age_Category Menopause Cancer_Type Grade Primary_Tumour Nodes Gender Age Race
## 1        30-40        NA     IBC-NST     3            pT2   pN1     NA  NA   NA
## 2        30-40        NA     IBC-NST     3            pT2   pN1     NA  NA   NA
## 3        30-40        NA     IBC-NST     3            pT2   pN1     NA  NA   NA
## 4        30-40        NA     IBC-NST     3            pT2   pN1     NA  NA   NA
## 5        30-40        NA     IBC-NST     3            pT2   pN1     NA  NA   NA
##   Parity Stage Tissue
## 1     NA    NA Tumour
## 2     NA    NA Tumour
## 3     NA    NA Tumour
## 4     NA    NA Tumour
## 5     NA    NA Tumour
```

```r
#saveRDS(metadata,
#        "/mnt/nmorais-nfs/marta/pC_myeinfobank/metadata/bassez-2021/bassez-2021-clinical-metadata.rds")
```