QC file for 003 histology file

Background

QC for histology for stake holders to check through

"Not Reported" in column age_at_diagnois is converted to `NA`

RNA library

what are all the rna_libraries

```
cohort_plot_labels %>% dplyr::select(RNA_library,cohort) %>% table()
```

```
##             cohort
## RNA_library CBTTC PNOC003 PNOC008
##   poly-A       26      32       0
##   rna_exome     0       0       9
##   stranded    970       0      13
```

Any duplicates in Kids_First_Biospecimen_IDs

```
check_df<-cohort_plot_labels %>% group_by(Kids_First_Biospecimen_ID) %>% tally() %>% arrange(desc(n))
any(check_df$n>1)
```

```
## [1] FALSE
```

TP53 status experimental_strategy check

Here I'm checking - per sample_id are there are different values - if WGS/WXS alteration is matched with RNA-Seq per sample_id

```
cohort_plot_labels %>%
  dplyr::filter(TP53alteration_status %in% c("Yes","No"),sample_type=="Tumor",short_histology=="HGAT") %
  dplyr::group_by(sample_id) %>%
  dplyr::summarise(number_anno=toString(length(unique(TP53alteration_status))),
                   TP53alteration_status = toString(unique(TP53alteration_status)),
                   experimental_strategy = toString(experimental_strategy)
                   ) %>%
  dplyr::arrange(desc(number_anno))
```

```
## # A tibble: 234 x 4
##    sample_id number_anno TP53alteration_status experimental_strategy
##    <chr>     <chr>       <chr>                 <chr>
##  1 7316-1052 1           Yes                   RNA-Seq
##  2 7316-1055 1           Yes                   RNA-Seq
##  3 7316-1057 1           Yes                   RNA-Seq
##  4 7316-1059 1           Yes                   RNA-Seq
##  5 7316-1060 1           Yes                   RNA-Seq
##  6 7316-1062 1           Yes                   RNA-Seq
##  7 7316-1064 1           Yes                   RNA-Seq
##  8 7316-1068 1           Yes                   RNA-Seq
##  9 7316-1085 1           Yes                   WGS, RNA-Seq
## 10 7316-1099 1           Yes                   WGS, RNA-Seq
## # ... with 224 more rows
```

H3 status experimental_strategy check

Here I'm checking - per sample_id are there are different values - if WGS/WXS alteration is matched with RNA-Seq per sample_id

```
cohort_plot_labels %>%
  dplyr::filter(grepl("K28|wildtype|G35",H3.status),sample_type=="Tumor",short_histology=="HGAT") %>%
  dplyr::group_by(sample_id) %>%
  dplyr::summarise(number_anno=toString(length(unique(H3.status))),
                   H3.status = toString(unique(H3.status)),
                   experimental_strategy = toString(experimental_strategy)
                   ) %>%
  dplyr::arrange(desc(number_anno))
```

```
## # A tibble: 234 x 4
##    sample_id number_anno H3.status   experimental_strategy
##    <chr>     <chr>       <chr>       <chr>
##  1 7316-1052 1           H3.3 K28    RNA-Seq
##  2 7316-1055 1           H3 wildtype RNA-Seq
##  3 7316-1057 1           H3 wildtype RNA-Seq
##  4 7316-1059 1           H3 wildtype RNA-Seq
##  5 7316-1060 1           H3.3 G35    RNA-Seq
##  6 7316-1062 1           H3.3 K28    RNA-Seq
##  7 7316-1064 1           H3 wildtype RNA-Seq
##  8 7316-1068 1           H3 wildtype RNA-Seq
##  9 7316-1085 1           H3.3 K28    WGS, RNA-Seq
## 10 7316-1099 1           H3.3 G35    WGS, RNA-Seq
## # ... with 224 more rows
```

check if all Diffuse astrocytic and oligodendroglial tumor are annotated

```
all_sample_id <-cohort_plot_labels %>%
  dplyr::filter(sample_type=="Tumor",short_histology=="HGAT") %>%
```

```
  dplyr::select(sample_id,short_histology) %>%
  pull(sample_id) %>% unique()


h3_sample_id<-cohort_plot_labels %>%
  dplyr::filter(grepl("K28|wildtype|G35",H3.status),short_histology=="HGAT") %>%
  dplyr::pull(sample_id) %>% unique()

tp53_sample_id <- cohort_plot_labels %>%
  dplyr::filter(TP53alteration_status %in% c("Yes","No"),short_histology=="HGAT") %>%
  dplyr::pull(sample_id)%>% unique()

print("All sample_ids have h3 mutation status")
```

```
## [1] "All sample_ids have h3 mutation status"
```

```
all(all_sample_id %in% h3_sample_id & length(all_sample_id)==length(h3_sample_id))
```

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

```
print("All sample_ids have tp53 mutation status")
```

```
## [1] "All sample_ids have tp53 mutation status"
```

```
all(all_sample_id %in% tp53_sample_id & length(all_sample_id)==length(tp53_sample_id))
```

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

```
all_kf_bs_id <-cohort_plot_labels %>%
  dplyr::filter(sample_type=="Tumor",short_histology=="HGAT") %>%
  pull(Kids_First_Biospecimen_ID) %>% unique()


h3_kf_bs_id<-cohort_plot_labels %>%
  dplyr::filter(grepl("K28|wildtype|G35",H3.status),short_histology=="HGAT") %>%
  dplyr::pull(Kids_First_Biospecimen_ID) %>% unique()

tp53_kf_bs_id <- cohort_plot_labels %>%
  dplyr::filter(TP53alteration_status %in% c("Yes","No"),short_histology=="HGAT") %>%
  dplyr::pull(Kids_First_Biospecimen_ID)%>% unique()

print("All Kids_First_Biospecimen_IDs have h3 mutation status")
```

```
## [1] "All Kids_First_Biospecimen_IDs have h3 mutation status"
```

```
all(all_kf_bs_id %in% h3_kf_bs_id & length(all_kf_bs_id)==length(h3_kf_bs_id))
```

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

```
print("All Kids_First_Biospecimen_IDs have tp53 mutation status")
```

```
## [1] "All Kids_First_Biospecimen_IDs have tp53 mutation status"
```

```
all(all_kf_bs_id %in% tp53_kf_bs_id & length(all_kf_bs_id)==length(tp53_kf_bs_id))
```

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

Get sample number per cohort

```
pnoc003_dx <- cohort_plot_labels %>%
  dplyr::filter(grepl("Y",.$`pnoc003-dx`)) %>%
  dplyr::select(Kids_First_Biospecimen_ID, experimental_strategy) %>%
  unique() %>%
  pull(experimental_strategy) %>%
  table() %>%
  as.list()

pnoc003_dx_prog <- cohort_plot_labels %>%
  dplyr::filter(grepl("Y",.$`pnoc003-dx-prog-pm`)) %>%
  dplyr::select(Kids_First_Biospecimen_ID, experimental_strategy) %>%
  unique() %>%
  pull(experimental_strategy) %>%
  table() %>%
  as.list()

pbta_hgat_dx <- cohort_plot_labels %>%
  dplyr::filter(grepl("Y",.$`pbta-hgat-dx`)) %>%
  dplyr::select(Kids_First_Biospecimen_ID, experimental_strategy) %>%
  unique() %>%
  pull(experimental_strategy) %>%
  table() %>%
  as.list()

pbta_hgat_dx_prog <- cohort_plot_labels %>%
  dplyr::filter(grepl("Y",.$`pbta-hgat-dx-prog-pm`)) %>%
  dplyr::select(Kids_First_Biospecimen_ID, experimental_strategy) %>%
  unique() %>%
  pull(experimental_strategy) %>%
  table() %>%
  as.list()


pbta_hgat_dx_wgs_rna <- cohort_plot_labels %>%
  dplyr::filter(grepl("Y",.$`pbta-hgat-dx-wgs-rna`)) %>%
  dplyr::select(Kids_First_Biospecimen_ID, experimental_strategy) %>%
  unique() %>%
  pull(experimental_strategy) %>%
  table() %>%
```

```
  as.list()

count <- list("pnoc003-dx" = pnoc003_dx,
     "pnoc003-dx-prog" = pnoc003_dx_prog,
     "pbta-hgat-dx"= pbta_hgat_dx,
     "pbta-hgat-dx-prog" = pbta_hgat_dx_prog,
     "pbta-hgat-dx-wgs-rna"= pbta_hgat_dx_wgs_rna)

rlist::list.save(count,
             file = file.path(root_dir,
                       "analyses",
                       "mol-clinical-annotation-files",
                       "output",
                       "annotation_files",
                       "count_experimental_strategy_per_cohort.yaml"))
```

Check annotation columns

[1] “H3.status”
[2] “integrated_diagnosis”
[3] “cohort”
[4] “TP53alteration_status”
[5] “OS_status”
[6] “age_at_diagnosis_less_than_1_year”
[7] “age_at_diagnosis_more_than_1_less_than_5_years” [8] “age_at_diagnosis_more_than_5_less_than_10_years”
[9] “age_at_diagnosis_more_than_10_years”
[10] “reported_gender”
[11] “CNS_region”
[12] “autopsy_tumor_location”

```
annotation_cols <- unlist(str_split("H3.status,integrated_diagnosis,cohort,TP53alteration_status,OS_sta

get_count<-function(x){
cohort_plot_labels %>%
    dplyr::filter(short_histology=="HGAT",sample_type=="Tumor") %>%
    dplyr::group_by(!!as.name(x)) %>%
    tally() %>%
    arrange(desc(n))
}

lapply(annotation_cols,function(x) get_count(x))
```

```
## [[1]]
## # A tibble: 4 x 2
##   H3.status       n
##   <chr>       <int>
## 1 H3.3 K28      212
## 2 H3 wildtype   181
## 3 H3.3 G35       20
## 4 H3.1 K28       18
##
```

```
## [[2]]
## # A tibble: 2 x 2
##   integrated_diagnosis       n
##   <chr>                  <int>
## 1 Diffuse midline glioma   230
## 2 High-grade glioma        201
##
## [[3]]
## # A tibble: 3 x 2
##   cohort      n
##   <chr>   <int>
## 1 CBTTC     254
## 2 PNOC003   131
## 3 PNOC008    46
##
## [[4]]
## # A tibble: 2 x 2
##   TP53alteration_status     n
##   <chr>                 <int>
## 1 Yes                     274
## 2 No                      157
##
## [[5]]
## # A tibble: 3 x 2
##   OS_status     n
##   <chr>     <int>
## 1 DECEASED    319
## 2 LIVING       63
## 3 <NA>         49
##
## [[6]]
## # A tibble: 3 x 2
##   age_at_diagnosis_less_than_1_year     n
##   <chr>                             <int>
## 1 No                                  414
## 2 Yes                                  16
## 3 Not_Reported                          1
##
## [[7]]
## # A tibble: 3 x 2
##   age_at_diagnosis_more_than_1_less_than_5_years     n
##   <chr>                                          <int>
## 1 No                                               366
## 2 Yes                                               64
## 3 Not_Reported                                       1
##
## [[8]]
## # A tibble: 3 x 2
##   age_at_diagnosis_more_than_5_less_than_10_years     n
##   <chr>                                             <int>
## 1 No                                                  233
## 2 Yes                                                 197
## 3 Not_Reported                                          1
##
```

```
## [[9]]
## # A tibble: 3 x 2
##   age_at_diagnosis_more_than_10_years     n
##   <chr>                                 <int>
## 1 No                                      277
## 2 Yes                                     153
## 3 Not_Reported                              1
##
## [[10]]
## # A tibble: 3 x 2
##   reported_gender     n
##   <chr>           <int>
## 1 Male              221
## 2 Female            209
## 3 Not Available       1
##
## [[11]]
## # A tibble: 7 x 2
##   CNS_region          n
##   <chr>           <int>
## 1 Midline           219
## 2 Hemispheric       123
## 3 Other              61
## 4 Posterior fossa    21
## 5 <NA>                3
## 6 Spine               2
## 7 Ventricles          2
##
## [[12]]
## # A tibble: 76 x 2
##    primary_site                          n
##    <chr>                               <int>
##  1 Pons/Brainstem                      118
##  2 Thalamus                             43
##  3 Temporal Lobe                        35
##  4 Frontal Lobe                         31
##  5 Brain Stem- Pons                     18
##  6 Cerebellum/Posterior Fossa           18
##  7 Parietal Lobe                        17
##  8 Basal Ganglia;Thalamus                8
##  9 Parietal Lobe;Temporal Lobe           7
## 10 Temporal Lobe;Thalamus;Ventricles     6
## # ... with 66 more rows
```

‘‘‘