Histologies File QC

Jo Lynne Rokita, Krutika Gaonkar (D3B)

Contents

In this notebook we are using v18 base histology to create a base histology for v19 release. "Base histology" file has the basic clinical information manifest that is required by subtyping modules to add in OpenPBTA subtyping information.

The v18 base histologies was generated in this script: script.

CNS_region values were mis-assigned by a bug in v18 which will be fixed and QC-ed as well #14 and original issue on OpenPBTA is in 838

Load packages

```
suppressMessages(library(emo))
suppressMessages(library(tidyverse))
```

```
## Warning: package 'tidyverse' was built under R version 3.5.2

## Warning: package 'ggplot2' was built under R version 3.5.2

## Warning: package 'dplyr' was built under R version 3.5.2

## Warning: package 'stringr' was built under R version 3.5.2

## Warning: package 'forcats' was built under R version 3.5.2
```

Directories and Files

Directories

```
# Input directory
input_dir <- file.path("input")
# soft linked previous release histology
prev_hist_file <- params$prev_histology

# adapt histology
latest_hist_file <- params$latest_histology

##--- KEEP LINK to G-DRIVE --- ##

# pathology diagnosis is needed to match tumor samples
# to broad/short histology
#path_dx <- read_sheet('https://docs.google.com/spreadsheets/d/1fDXt_YODcSAWDvyI5ISBVhUCu4b5-TFCVWMOwiP
#  dplyr::select(pathology_diagnosis,broad_histology, short_histology) %>%
```

```
#  write_tsv(file.path(input_dir,"pathology_diagnosis_for_subtyping.tsv"))

# pathology free text diagnosis is needed to match to
# samples marked as "Other" in pathology_diagnosis
#path_free_text <- read_sheet('https://docs.google.com/spreadsheets/d/1fDXt_YODcSAWDvyI5ISBVhUCu4b5-TFC
#  dplyr::select(pathology_free_text_diagnosis,broad_histology, short_histology)%>%
#  write_tsv(file.path(input_dir,"pathology_free_text_diagnosis_for_subtyping.tsv"))

## ----------- ##
```

Read in old base histology

```
prev_hist <- read_tsv(prev_hist_file,
                      # NAs are being read as logical so specifying as character here
                           col_types = readr::cols(molecular_subtype = readr::col_character(),
                                                   short_histology = readr::col_character(),
                                                   integrated_diagnosis = readr::col_character(),
                           broad_histology = readr::col_character(),
                           Notes = readr::col_character()))
```

```
## Warning: 326 parsing failures.
##  row        col           expected actual                                 file
## 2743 pnoc003-dx 1/0/T/F/TRUE/FALSE   Y    '20201215-data/pbta-histologies.tsv'
## 2743 P_id       1/0/T/F/TRUE/FALSE   P-01 '20201215-data/pbta-histologies.tsv'
## 2744 P_id       1/0/T/F/TRUE/FALSE   P-01 '20201215-data/pbta-histologies.tsv'
## 2745 pnoc003-dx 1/0/T/F/TRUE/FALSE   Y    '20201215-data/pbta-histologies.tsv'
## 2745 P_id       1/0/T/F/TRUE/FALSE   P-01 '20201215-data/pbta-histologies.tsv'
## .... .......... .................. ...... .....................................
## See problems(...) for more details.
```

```
path_dx <- read_tsv(file.path(input_dir,"pathology_diagnosis_for_subtyping.tsv")) %>%
  dplyr::select(pathology_diagnosis, broad_histology, short_histology)
```

```
## Parsed with column specification:
## cols(
##   pathology_diagnosis = col_character(),
##   broad_histology = col_character(),
##   short_histology = col_character()
## )
```

```
path_free_text <- read_tsv(file.path(input_dir,"pathology_free_text_diagnosis_for_subtyping.tsv")) %>%
  dplyr::select(pathology_free_text_diagnosis, broad_histology, short_histology)
```

```
## Parsed with column specification:
## cols(
##   pathology_free_text_diagnosis = col_character(),
##   broad_histology = col_character(),
##   short_histology = col_character()
## )
```

Subset new file for only those sampleIDs required

v18 but we will remove BS_JXF8A2A6 for v19 #862

```
latest_hist <- read_tsv(latest_hist_file,
                        # NAs are being read as logical so specifying as character here
                             col_types = readr::cols(molecular_subtype = readr::col_character(),
                                                     short_histology = readr::col_character(),
                                                     integrated_diagnosis = readr::col_character(),
                             broad_histology = readr::col_character(),
                             Notes = readr::col_character()))

# get ids to subset
id_to_subset <- prev_hist %>%
  pull(Kids_First_Biospecimen_ID)
```

Add ids to previous release?

```
if (params$add_ids != ""){
  add_ids <- unlist(str_split(params$add_ids, ","))
  # add new ids to previous releases
  id_to_subset <- c( id_to_subset, add_ids)
  print(paste(toString(add_ids)," added"))
}
```

```
## [1] "BS_KKDTW11T, BS_X1TRW9RH, BS_46MV2DSY, BS_00FD2KMP, BS_K24D4BGK, BS_VF1R7VC2  added"
```

subset to previous ids (and new ids if provided)

```
# subset final histology
latest_hist <- latest_hist %>%
  filter(Kids_First_Biospecimen_ID  %in% id_to_subset)
```

Check 1: Assess dimensions whether new column names match the old

Check 1a: assess ids overlap in new and old

```
source("code/util/check_rows_cols.R")
check_rows(new_hist = latest_hist,old_hist = prev_hist,
          column_name = "Kids_First_Biospecimen_ID")
```

```
## [1] "ids in new and not in old in: Kids_First_Biospecimen_ID BS_00FD2KMP, BS_46MV2DSY, BS_K24D4BGK, B
## [1] "ids in old and not in new in: Kids_First_Biospecimen_ID "
```

Check 1b: assess columns overlap in new and old

```
check_cols(new_hist = latest_hist,old_hist = prev_hist)
```

```
## [1] "Columns in new that are not in old: extent_of_tumor_resection"
##  [1] "Columns in old that are not in new: parent_aliquot_id"
##  [2] "Columns in old that are not in new: cancer_predisposition"
##  [3] "Columns in old that are not in new: cell_line_composition"
##  [4] "Columns in old that are not in new: HIST1H3C_MOD_HIGH_mut_consensus"
##  [5] "Columns in old that are not in new: HIST1H3B_MOD_HIGH_mut_consensus"
##  [6] "Columns in old that are not in new: H3F3A_MOD_HIGH_mut_consensus"
##  [7] "Columns in old that are not in new: H3F3A_MOD_HIGH_mut_rna_seq_snv"
##  [8] "Columns in old that are not in new: TP53_MOD_HIGH_mut_consensus"
##  [9] "Columns in old that are not in new: TP53_MOD_HIGH_mut_rna_seq_snv"
## [10] "Columns in old that are not in new: TP53_gain_consensus_cnv"
## [11] "Columns in old that are not in new: TP53_loss_consensus_cnv"
## [12] "Columns in old that are not in new: TP53_MOD_HIGH_mut_HGVS_consensus"
## [13] "Columns in old that are not in new: TP53_MOD_HIGH_mut_HGVS_rna_seq_snv"
## [14] "Columns in old that are not in new: H3.status"
## [15] "Columns in old that are not in new: TP53alteration_status"
## [16] "Columns in old that are not in new: age_at_diagnosis_less_than_1_year"
## [17] "Columns in old that are not in new: age_at_diagnosis_more_than_1_less_than_5_years"
## [18] "Columns in old that are not in new: age_at_diagnosis_more_than_5_less_than_10_years"
## [19] "Columns in old that are not in new: age_at_diagnosis_more_than_10_years"
## [20] "Columns in old that are not in new: pnoc003-dx"
## [21] "Columns in old that are not in new: pnoc003-dx-prog-pm"
## [22] "Columns in old that are not in new: pbta-hgat-dx"
## [23] "Columns in old that are not in new: pbta-hgat-dx-prog-pm"
## [24] "Columns in old that are not in new: pbta-hgat-dx-wgs-rna"
## [25] "Columns in old that are not in new: pnoc003-dx-celn-wgs"
## [26] "Columns in old that are not in new: P_id"
## [27] "Columns in old that are not in new: ancestry"
```

Check 2: Assess levels of histology columns

Check 2a: path_dx and path_free_text_dx is used to match later so should have the same values in new histology

```
check_values(new_hist = latest_hist,old_hist = prev_hist,
          column_name = "pathology_diagnosis",output_dir = params$output)
```

```
## [1] "Different values found in new histology "
## [1] "Levels differ in  pathology_diagnosis because change in BS_00FD2KMP, BS_0AK4F99X, BS_0ATJ22QA, I
```

```
check_values(new_hist = latest_hist,old_hist = prev_hist,
          column_name = "pathology_free_text_diagnosis",output_dir = params$output)
```

```
## [1] "Different values found in new histology Glioblastoma, Diffuse Astrocytoma, Anaplastic Astrocytor
## [1] "Levels differ in  pathology_free_text_diagnosis because change in BS_00FD2KMP, BS_00QJE3E5, BS_
```

Check 2b: Normals, these should not have path_dx, int_dx,molecular_subtype, broad/short_hist

```
latest_hist_normals <- latest_hist %>%
  filter(sample_type=="Normal")
prev_hist_normals <- prev_hist %>%
  filter(sample_type=="Normal")

key_column_name = c("pathology_free_text_diagnosis","pathology_diagnosis","primary_site")

distinct(prev_hist_normals[,key_column_name])
```

```
## # A tibble: 5 x 3
##   pathology_free_text_diagnosis pathology_diagnosis primary_site
##   <chr>                         <chr>               <chr>
## 1 na                            <NA>                Peripheral Whole Blood
## 2 na                            <NA>                <NA>
## 3 <NA>                          <NA>                Peripheral Whole Blood
## 4 na                            <NA>                Brain
## 5 <NA>                          <NA>                Adjacent Brain
```

```
distinct(latest_hist_normals[,key_column_name])
```

```
## # A tibble: 5 x 3
##   pathology_free_text_dia~ pathology_diagnosis                primary_site
##   <chr>                    <chr>                              <chr>
## 1 <NA>                     <NA>                               Peripheral Whole~
## 2 <NA>                     <NA>                               <NA>
## 3 <NA>                     <NA>                               Brain
## 4 <NA>                     High-grade glioma/astrocytoma (WHO~ Peripheral Whole~
## 5 <NA>                     <NA>                               Adjacent Brain
```

Check3 tables per column changes

Check 3a Experimental strategy

```
check_values(new_hist = latest_hist,old_hist = prev_hist,
          column_name = "experimental_strategy",output_dir = params$output)
```

```
## [1] "Panel counts changed"   "RNA-Seq counts changed" "WGS counts changed"
## [4] "WXS counts changed"
## [1] "Levels differ in  experimental_strategy because change in BS_00FD2KMP, BS_46MV2DSY, BS_7KR13R3P
```

Check 3b Sample Type

```
check_values(new_hist = latest_hist,old_hist = prev_hist,
          column_name = "sample_type",output_dir = params$output)
```

```
## [1] "Normal counts changed" "Tumor counts changed"
## [1] "Levels differ in  sample_type because change in BS_00FD2KMP, BS_46MV2DSY, BS_K24D4BGK, BS_KKDTW
```

Check 3c Tumor Descriptor

```
check_values(new_hist = latest_hist,old_hist = prev_hist,
          column_name = "tumor_descriptor",output_dir = params$output)
```

```
## [1] "Different values found in new histology "
## [1] "Levels differ in  tumor_descriptor because change in BS_00FD2KMP, BS_18RH1034, BS_2853394H, BS_
```

Check 3d Composition

```
# update composition with to match new terms to previous composition terms


check_values(new_hist = latest_hist,old_hist = prev_hist,
          column_name = "composition",params$output)
```

```
## [1] "Peripheral Whole Blood counts changed"
## [2] "Solid Tissue counts changed"
## [1] "Levels differ in  composition because change in BS_00FD2KMP, BS_46MV2DSY, BS_K24D4BGK, BS_KKDTW
```

Check 3f RNA library

```
check_values(new_hist = latest_hist,old_hist = prev_hist,
          column_name = "RNA_library",output_dir = params$output)
```

```
## [1] "poly-A counts changed"    "rna_exome counts changed"
## [3] "stranded counts changed"
## [1] "Levels differ in  RNA_library because change in BS_00FD2KMP, BS_2853394H, BS_3ZPJAK9A, BS_46MV2
```

Check 3g: Cohort

```
check_values(new_hist = latest_hist,old_hist = prev_hist,
          column_name = "cohort",output_dir = params$output)
```

```
## [1] "CBTN counts changed"    "PNOC003 counts changed" "PNOC008 counts changed"
## [1] "Levels differ in  cohort because change in BS_00FD2KMP, BS_18RH1034, BS_2853394H, BS_2B6ZEXAP, 
```

Check 3h: Sample and aliquot IDs - any changes?

```
check_values(new_hist = latest_hist,old_hist = prev_hist,
          column_name = "sample_id",output_dir = params$output)
```

```
## [1] "Different values found in new histology 7316-8716, 7316-4994-T-A14565.WGS, 7316-3217-T-A12398.W
## [1] "Levels differ in  sample_id because change in BS_00FD2KMP, BS_0AK4F99X, BS_0ATJ22QA, BS_0DVXQNO
```

```
check_values(new_hist = latest_hist,old_hist = prev_hist,
          column_name = "aliquot_id", output_dir = params$output)
```

```
## [1] "Different values found in new histology 1040291_RNA_T, 1030650, 1030648, 1062685_DNA_N, 1030626
## [1] "Levels differ in  aliquot_id because change in BS_00FD2KMP, BS_1HQ76V6D, BS_3BDAG9YN, BS_46MV2D
```

Check 3i: Sequencing Center

```
check_values(new_hist = latest_hist,old_hist = prev_hist,
          column_name = "seq_center",output_dir = params$output)
```

```
## [1] "BGI counts changed"
## [2] "BGI@CHOP Genome Center counts changed"
## [3] "CHOP DGD counts changed"
## [4] "NantOmics counts changed"
## [1] "Levels differ in  seq_center because change in BS_00FD2KMP, BS_0DVXQN0X, BS_0N50PRC8, BS_0ZA67B
```

Check 3f: primary_site

```
check_values(new_hist = latest_hist,old_hist = prev_hist,
          column_name = "primary_site",output_dir = params$output)
```

```
## [1] "Different values found in new histology right anterior temporal lobe, Right Temporal Lobe"
## [1] "Levels differ in  primary_site because change in BS_00FD2KMP, BS_46MV2DSY, BS_9CA93S6D, BS_K24D
```

Update CNS_region

json file was generated from the CNS_region updates ticket 838.

Match CNS_region matching primary_site and update

```
source("code/util/primary_site_matched_CNS_region.R")
latest_hist <- get_CNS_region(histology=latest_hist,
                  CNS_match_json=file.path(input_dir,"CNS_primary_site_match.json"))
```

Which samples had different CNS_region in v18?

```
diff_cns <-latest_hist %>%
  left_join(prev_hist[,c("Kids_First_Biospecimen_ID","CNS_region")],by=c("Kids_First_Biospecimen_ID"), 
    dplyr::select(Kids_First_Biospecimen_ID,CNS_region.v18,CNS_region.v19,primary_site) %>%
  dplyr::filter(CNS_region.v18 != CNS_region.v19)


diff_cns
```

```
## # A tibble: 252 x 4
##    Kids_First_Biospec~ CNS_region.v18 CNS_region.v19 primary_site
##    <chr>               <chr>          <chr>          <chr>
##  1 BS_00TRPEQX         Other          Mixed          Cerebellum/Posterior Fossa~
##  2 BS_02NZT8CE         Other          Mixed          Optic Pathway;Temporal Lobe
##  3 BS_042DVDQM         Other          Mixed          Optic Pathway;Suprasellar/~
##  4 BS_05S9WJW6         Other          Mixed          Cerebellum/Posterior Fossa~
##  5 BS_0C7VZC0A         Other          Mixed          Basal Ganglia;Optic Pathwa~
##  6 BS_0XEG6SNV         Other          Mixed          Parietal Lobe;Ventricles
##  7 BS_0ZR4XA69         Other          Mixed          Skull;Temporal Lobe
##  8 BS_10V9SAG8         Other          Mixed          Cerebellum/Posterior Fossa~
##  9 BS_1607397Q         Other          Mixed          Skull;Temporal Lobe
## 10 BS_16FT8V4B         Other          Mixed          Cerebellum/Posterior Fossa~
## # ... with 242 more rows
```

```
diff_cns$CNS_region.v19 %>% table()
```

```
## .
## Hemispheric       Mixed
##           2         250
```

135 samples were incorrectly assigned CNS_region in v18. 129 on these should be 'Mixed' and 6 'Other', fixed with an updated `cns_region_check()` function in this notebook.

Update broad_histology and short_histology

Match by pathology_diagnosis and pathology_free_text_diagnosis (Other)

By path_free_text for "Other" diagnosed

Only samples with 'Other' in pathology_diagnosis will be need to be matched by path_free_text

```
latest_hist<- dplyr::select(latest_hist,c(-broad_histology,-short_histology))
latest_hist_other <- latest_hist  %>%
  dplyr::filter(pathology_diagnosis == "Other") %>%
  left_join(path_free_text,by="pathology_free_text_diagnosis")
```

By path_dx for all tumors other than "Other"

Remove samples with 'Other' in pathology_diagnosis that was already matched above

```
latest_hist <- latest_hist  %>%
  dplyr::filter(pathology_diagnosis != "Other"|
                  # add Normals
                  sample_type=="Normal") %>%
  left_join(path_dx,by="pathology_diagnosis") %>%
  bind_rows(latest_hist_other)
```

Check broad_histology

```
check_values(new_hist = latest_hist,old_hist = prev_hist,
          column_name = "broad_histology",output_dir = params$output)
```

```
## [1] "Different values found in new histology Lymphoma, Metastatic tumors, Other astrocytic tumor, Nor
## [1] "Levels differ in  broad_histology because change in BS_00FD2KMP, BS_0YVR8Q4E, BS_0ZR4XA69, BS_1(
```

Check short_histology

```
check_values(new_hist = latest_hist,old_hist = prev_hist,
          column_name = "short_histology",output_dir = params$output)
```

```
## [1] "Different values found in new histology Sarcoma, Embryonal tumor, Langerhans cell histiocytosis
## [1] "Levels differ in  short_histology because change in BS_00FD2KMP, BS_02ZSVZCB, BS_09R7GDA7, BS_0S
```

Remove ids from previous release?

```
if (params$remove_ids != ""){
  remove_ids <- unlist(str_split(params$remove_ids, ","))
  # remove ids from previous releases
  latest_hist <- latest_hist %>%
    filter(!Kids_First_Biospecimen_ID %in% remove_ids)
  print(paste(toString(remove_ids)," removed"))
}
```

Write new file

```
write.table(latest_hist, file.path(params$output,"pbta-histologies-base.tsv")
            , sep = "\t", quote = F, col.names = T, row.names = F)
```

```
sessionInfo()
```

```
## R version 3.5.1 (2018-07-02)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS Sierra 10.12.6
##
## Matrix products: default
## BLAS: /Library/Frameworks/R.framework/Versions/3.5/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.5/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
```

```
## other attached packages:
##  [1] forcats_0.5.0   stringr_1.4.0   dplyr_0.8.5     purrr_0.3.4
##  [5] readr_1.3.1     tidyr_1.0.2     tibble_3.0.0    ggplot2_3.3.0
##  [9] tidyverse_1.3.0 emo_0.0.0.9000
##
## loaded via a namespace (and not attached):
##  [1] tidyselect_1.0.0  xfun_0.19         haven_2.2.0       lattice_0.20-41
##  [5] colorspace_1.4-1  vctrs_0.2.4       generics_0.0.2    htmltools_0.5.1.1
##  [9] yaml_2.2.1        utf8_1.1.4        rlang_0.4.6       pillar_1.4.3
## [13] withr_2.2.0       glue_1.4.0        DBI_1.1.0         dbplyr_1.4.2
## [17] modelr_0.1.6      readxl_1.3.1      lifecycle_0.2.0   munsell_0.5.0
## [21] gtable_0.3.0      cellranger_1.1.0  rvest_0.3.5       evaluate_0.14
## [25] knitr_1.30        fansi_0.4.1       broom_0.5.5       Rcpp_1.0.4
## [29] backports_1.1.6   scales_1.1.0      jsonlite_1.6.1    fs_1.3.1
## [33] hms_0.5.3         digest_0.6.25     stringi_1.4.6     grid_3.5.1
## [37] cli_2.0.2         tools_3.5.1       magrittr_1.5      crayon_1.3.4
## [41] pkgconfig_2.0.3   ellipsis_0.3.0    xml2_1.3.2        reprex_0.3.0
## [45] lubridate_1.7.8   rstudioapi_0.11   assertthat_0.2.1  rmarkdown_2.3
## [49] httr_1.4.2        R6_2.4.1          nlme_3.1-137      compiler_3.5.1
```