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()))

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"))
}
```

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 "
## [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 overlap in new and old \u2705"
```

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] "Glial-neuronal tumor NOS counts changed"
## [2] "Low-grade glioma/astrocytoma (WHO grade I/II) counts changed"
## [3] "Metastatic secondary tumors;Neuroblastoma counts changed"
## [4] "Neuroblastoma counts changed"
## [5] "Schwannoma counts changed"
## [1] "Levels overlap in new and old \u2705"
```

```
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 Low-grade glioma/astrocytoma (WHO grade I/II), High-grad
## [1] "Levels differ in  pathology_free_text_diagnosis because change in BS_16FT8V4B, BS_17AXPP1Y, BS_1
```

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: 4 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>                Brain
## 4 <NA>                          <NA>                Adjacent Brain
```

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

```
## # A tibble: 4 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>                Brain
## 4 <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"
## [1] "Levels differ in  experimental_strategy because change in BS_7KR13R3P, BS_WHZT48VG"
```

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 overlap in new and old \u2705"
```

Check 3c Tumor Descriptor

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

```
## [1] "Initial CNS Tumor counts changed" "Progressive counts changed"
## [1] "Levels overlap in new and old \u2705"
```

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 overlap in new and old \u2705"
```

Check 3f RNA library

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

```
## [1] "stranded counts changed"
## [1] "Levels overlap in new and old \u2705"
```

Check 3g: Cohort

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

```
## [1] "CBTN counts changed"
## [1] "Levels overlap in new and old \u2705"
```

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 7316-3217-T-A12398.WGS, A16404-N.WXS, A14447-N.WXS, 731
## [1] "Levels differ in  sample_id because change in BS_0ATJ22QA, BS_0DVXQN0X, BS_0N50PRC8, BS_0NC1NQ0
```

```
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 1030650, 1030648, 1030626, 1030634, 1030633, 1030646, 1
## [1] "Levels differ in  aliquot_id because change in BS_1HQ76V6D, BS_3BDAG9YN, BS_4DY83R02, BS_7F07M7
```

Check 3i: Sequencing Center

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

```
## [1] "NantOmics counts changed"
## [1] "Levels differ in  seq_center because change in BS_0DVXQN0X, BS_0N50PRC8, BS_0ZA67BBC, BS_1CQ01R
```

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 L. Pons Anterior, L. Lateral Pons, R. Posterior Pons; A
## [1] "Levels differ in  primary_site because of change in BS_1Q524P3B, BS_22VCR7DF, BS_5968GBGT, BS_A
```

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"), s
    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: 138 x 4
##    Kids_First_Biospec~ CNS_region.v18 CNS_region.v19 primary_site
##    <chr>               <chr>          <chr>          <chr>
##  1 BS_0C7VZC0A         Midline        Mixed          Basal Ganglia;Optic Pathwa~
##  2 BS_0XEG6SNV         Hemispheric    Mixed          Parietal Lobe;Ventricles
##  3 BS_0ZR4XA69         Ventricles     Mixed          Skull;Temporal Lobe
##  4 BS_1607397Q         Ventricles     Mixed          Skull;Temporal Lobe
##  5 BS_18NCV5QZ         Ventricles     Other          Meninges/Dura;Skull
##  6 BS_19EJ85F8         Midline        Mixed          Brain Stem- Pons;Cerebellu~
##  7 BS_1A6MQ9ZA         Hemispheric    Mixed          Frontal Lobe;Suprasellar/H~
##  8 BS_1D6PZNKN         Midline        Mixed          Brain Stem-Medulla;Brain S~
##  9 BS_23QW0BBA         Midline        Mixed          Brain Stem-Medulla;Brain S~
## 10 BS_2EN3X6HB         Midline        Mixed          Brain Stem- Midbrain/Tectu~
## # ... with 128 more rows
```

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

```
## .
## Mixed Other
##   132     6
```

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] "Embryonal tumor counts changed"
## [2] "Low-grade astrocytic tumor counts changed"
## [3] "Metastatic tumors counts changed"
## [4] "Neuronal and mixed neuronal-glial tumor counts changed"
## [5] "Tumor of cranial and paraspinal nerves counts changed"
## [1] "Levels overlap in new and old \u2705"
```

Check short_histology

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

```
## [1] "Glial-neuronal tumor NOS counts changed"
## [2] "LGAT counts changed"
## [3] "Metastases counts changed"
## [4] "Neuroblastoma counts changed"
## [5] "Schwannoma counts changed"
## [1] "Levels overlap in new and old \u2705"
```

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"))
}
```

```
## [1] "BS_JXF8A2A6  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
```