Tables output for manuscript

Run Jin

10/27/2021

Output Tables for OpenPBTA Manuscript

This is a Rmd files that record scripts for generating tables

```
root_dir <- rprojroot::find_root(rprojroot::has_dir(".git"))
working_dir <- file.path(root_dir, "tables")
data_dir <- file.path(root_dir, "data")

results_dir <- file.path(working_dir, "results")
if(!dir.exists(results_dir)){
  dir.create(results_dir, recursive=TRUE)
}
```

Table 1: Molecular subtypes determined for this project

```
histology_df <- readr::read_tsv(file.path(data_dir, "pbta-histologies.tsv"), guess_max =10000)
```

```
## Rows: 2840 Columns: 38

## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (33): Kids_First_Biospecimen_ID, sample_id, aliquot_id, Kids_First_Parti...
## dbl  (5): OS_days, age_last_update_days, normal_fraction, tumor_fraction, tu...

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
subtypes <- histology_df %>%
  dplyr::filter(!is.na(pathology_diagnosis) & !is.na(molecular_subtype) & !grepl("To be classified", mo
subtype_table <- subtypes %>%
  dplyr::group_by(broad_histology, molecular_subtype) %>%
  tally() %>%
  readr::write_tsv(file.path(results_dir, "Table1-molecular-subtypes.tsv"))
```

Table S1: V21 histologies table

```
openxlsx::write.xlsx(histology_df,
                     file.path(results_dir, "TableS1-histologies.xlsx"),
                     overwrite=TRUE)
```

Table S2: DNA results table

TMB

```
# read in tmb all file, select and rename columns
tmb_all <- readr::read_tsv(file.path(data_dir, "pbta-snv-mutation-tmb-all.tsv")) %>%
  dplyr::select(Tumor_Sample_Barcode, tmb) %>%
  dplyr::rename(Kids_First_Biospecimen_ID = Tumor_Sample_Barcode) %>%
  dplyr::rename(Tmb_all = tmb)
```

```
## Rows: 912 Columns: 6

## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (3): Tumor_Sample_Barcode, experimental_strategy, short_histology
## dbl (3): mutation_count, region_size, tmb

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# read in tmb coding file, select and rename columns
tmb_coding <- readr::read_tsv(file.path(data_dir, "pbta-snv-mutation-tmb-coding.tsv")) %>%
  dplyr::select(Tumor_Sample_Barcode, tmb) %>%
  dplyr::rename(Kids_First_Biospecimen_ID = Tumor_Sample_Barcode) %>%
  dplyr::rename(Tmb_coding = tmb)
```

```
## Rows: 910 Columns: 6

## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (3): Tumor_Sample_Barcode, experimental_strategy, short_histology
## dbl (3): mutation_count, region_size, tmb

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# combine files
tmb_combined <- full_join(tmb_all, tmb_coding)
```

```
## Joining, by = "Kids_First_Biospecimen_ID"
```

COSMIC mutational signatures

```
# read in the file
cosmic_mut_df <- readr::read_tsv("../analyses/mutational-signatures/results/cosmic_signatures_results.ts
  dplyr::select(Tumor_Sample_Barcode, signature, mut_per_mb) %>%
  dplyr::rename(Kids_First_Biospecimen_ID = Tumor_Sample_Barcode)
```

```
## Rows: 29977 Columns: 7
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (4): Tumor_Sample_Barcode, experimental_strategy, display_group, signature
## dbl (3): num_mutations, genome_size, mut_per_mb
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# get wide format
cosmic_mut_wide <- cosmic_mut_df %>%
  spread(signature, mut_per_mb )

# order the columns
unique_cosmic_sig <- cosmic_mut_df %>%
  pull(signature) %>% unique()
cosmic_mut_wide <- cosmic_mut_wide %>%
  dplyr::select(c(Kids_First_Biospecimen_ID, all_of(unique_cosmic_sig)))
```

Alexandrov mutational signatures

```
# read in the file
alexandrov_mut_df <- readr::read_tsv("../analyses/mutational-signatures/results/nature_signatures_resul
  dplyr::select(Tumor_Sample_Barcode, signature, mut_per_mb) %>%
  dplyr::rename(Kids_First_Biospecimen_ID = Tumor_Sample_Barcode)
```

```
## Rows: 27076 Columns: 7
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (4): Tumor_Sample_Barcode, experimental_strategy, display_group, signature
## dbl (3): num_mutations, genome_size, mut_per_mb
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# get wide format
alexandrov_mut_wide <- alexandrov_mut_df %>%
  spread(signature, mut_per_mb )
```

```
# order the columns
unique_alex_sig <- alexandrov_mut_df %>%
  pull(signature) %>% unique()
alexandrov_mut_wide <- alexandrov_mut_wide %>%
  dplyr::select(c(Kids_First_Biospecimen_ID, all_of(unique_alex_sig)))
```

CNS mutational signatures

```
cns_mut_list <- readRDS("../analyses/mutational-signatures/results/fitted_cns_signature_exposures.RDS")
cns_mean <- cns_mut_list[["mean"]] %>%
  as.data.frame() %>%
  tibble::rownames_to_column("Kids_First_Biospecimen_ID")
```

Chromothripsis regions per sample

```
chromothripsis_region_df <- readr::read_tsv("../analyses/chromothripsis/results/chromothripsis_summary_
```

```
## Rows: 777 Columns: 6
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (2): Kids_First_Biospecimen_ID, any_regions
## dbl (3): count_regions_any_conf, count_regions_high_conf, count_regions_low_...
## lgl (1): any_regions_logical
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

combine S2 table

```
list_s2_table <- list(tmb = tmb_combined,
                      cosmic_mut_sigs = cosmic_mut_wide,
                      alexandrov_mut_sigs = alexandrov_mut_wide,
                      cns_denovo_mut_sigs = cns_mean,
                      chromothripsis_events = chromothripsis_region_df
                      )
openxlsx::write.xlsx(list_s2_table,
                      file.path(results_dir, "TableS2-DNA-results-table.xlsx"),
                      overwrite=TRUE)
```

Table S3: RNA results table

read in and process files

```r
# get tp53 scores
tp53_scores <- readr::read_tsv("../analyses/tp53_nf1_score/results/tp53_altered_status.tsv")
```

```
## Rows: 1166 Columns: 16
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (8): sample_id, Kids_First_Biospecimen_ID_DNA, Kids_First_Biospecimen_ID...
## dbl (8): tp53_score, SNV_indel_counts, CNV_loss_counts, SV_counts, Fusion_co...
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
# get extend scores file
telomerase_scores_polya_count <- readr::read_tsv("../analyses/telomerase-activity-prediction/results/Tel
  dplyr::select(SampleID, NormEXTENDScores) %>%
  dplyr::rename(Kids_First_Biospecimen_ID_RNA = SampleID,
                NormEXTENDScores_counts = NormEXTENDScores)
```

```
## Rows: 58 Columns: 3
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (1): SampleID
## dbl (2): RawEXTENDScores, NormEXTENDScores
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
telomerase_scores_polya_fpkm <-
  readr::read_tsv("../analyses/telomerase-activity-prediction/results/TelomeraseScores_PTBAPolya_FPKM.ts
  dplyr::select(SampleID, NormEXTENDScores) %>%
  dplyr::rename(Kids_First_Biospecimen_ID_RNA = SampleID,
                NormEXTENDScores_fpkm = NormEXTENDScores)
```

```
## Rows: 58 Columns: 3
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (1): SampleID
## dbl (2): RawEXTENDScores, NormEXTENDScores
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
telomerase_scores_polya_combined <- full_join(telomerase_scores_polya_count,
                                              telomerase_scores_polya_fpkm)
```

```
## Joining, by = "Kids_First_Biospecimen_ID_RNA"
```

```
telomerase_scores_stranded_count <- readr::read_tsv("../analyses/telomerase-activity-prediction/results/
  dplyr::select(SampleID, NormEXTENDScores) %>%
  dplyr::rename(Kids_First_Biospecimen_ID_RNA = SampleID,
                NormEXTENDScores_counts = NormEXTENDScores)
```

```
## Rows: 977 Columns: 3
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (1): SampleID
## dbl (2): RawEXTENDScores, NormEXTENDScores
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
telomerase_scores_stranded_fpkm <- readr::read_tsv("../analyses/telomerase-activity-prediction/results/T
  dplyr::select(SampleID, NormEXTENDScores) %>%
  dplyr::rename(Kids_First_Biospecimen_ID_RNA = SampleID,
                NormEXTENDScores_fpkm = NormEXTENDScores)
```

```
## Rows: 977 Columns: 3
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (1): SampleID
## dbl (2): RawEXTENDScores, NormEXTENDScores
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
telomerase_scores_stranded_combined <- full_join(telomerase_scores_stranded_count,
                                                 telomerase_scores_stranded_fpkm)
```

```
## Joining, by = "Kids_First_Biospecimen_ID_RNA"
```

```
telomerase_scores_combined <- bind_rows(telomerase_scores_polya_combined,
                                        telomerase_scores_stranded_combined)
```

combine and output file

```
list_s3_table <- list(tp53_scores = tp53_scores,
                      telomerase_scores = telomerase_scores_combined
                      )
openxlsx::write.xlsx(list_s3_table,
                     file.path(results_dir, "TableS3-RNA-results-table.xlsx"),
                     overwrite=TRUE)
```