Create project metadata for nf-core/RNA-seq Analysis

Antonia Chroni achroni@stjude.org for SJCRH DNB_BINF_Core

Contents

```
## The following object is masked _by_ .GlobalEnv:
##
##     root_dir
```

PI: NA
Project: epigenomic-profiling-analysis
Task: analysis
Project Lead(s): NA
Department: Developmental Neurobiology

DNB Bioinformatics Core Analysis Team:

> **Lead Analyst(s): Antonia Chroni, PhD**
> Group Lead: Cody A. Ramirez, PhD
> **Contact E-mail:** achroni@stjude.org
> **DNB Bioinformatics Core Pipeline:** nf-core-RNA-seq-analysis

Date started: 09-03-2024
Date completed: complete
Report generated: 11:53:33 CDT 09/03/2024

Reviewed by: ________________________________ Date: __________________

1 Information about this notebook

This notebook creates the metadata for the project. The output file generated here can be used as an input file to run nf-core/RNA-seq pipeline.

1.1 What to include

The required columns for the samplesheet are “sample”, “fastq_1”, “fastq_2”, and “strandedness”. Additional columns are allowed, and we typically use the same sample sheet for downstream analyses. So adding columns for “line”, “group”, and “SJID” will be helpful. Each sample was run across two lanes, so there are two sets of FASTQs for each. These files will just be concatenated by the pipeline so long as the same sample name is provided. All required info can be inferred from the file names. Strandedness for all of our samples is always “reverse”. As an example, here’s what the filename of one sample file: SAMPLE1_C1-CELLLINE1_EXPERA24h_1_S21_L001_R1_001.fastq.gz

2 Set up

```r
suppressPackageStartupMessages({
  library(tidyverse)
  })
```

3 Directories and paths to file Inputs/Outputs

```r
attach(params)
```

```
## The following object is masked _by_ .GlobalEnv:
##
##     root_dir
```

```r
analysis_dir <- file.path(root_dir, "analyses", "create-project-metadata")

results_dir <- file.path(analysis_dir, "results")
if (!dir.exists(results_dir)) {
  dir.create(results_dir)}
```

4 Read data files

```r
# Read list of data files
files_list <- list()
files_list <- c(dir(path = data_dir,  pattern = ".fastq.gz", full.names = TRUE, recursive = TRUE))
files_df <- data.frame(file_path = unlist(files_list))

# Create list of sample names
sample_name_list <- list()
sample_name_list <- c(str_split_fixed(files_list, "/", 10)[,10])
sample_name_df <- data.frame(sample_name_drop = unlist(sample_name_list)) %>%
  mutate(sample_name_drop = str_replace(sample_name_drop, '_001.fastq.gz', ''))
```

5 Create df with `project_metadata`

```r
# Create df with files
df <- cbind(files_df, sample_name_df) %>%

  # add col `SJID`:
  mutate(SJID = str_split(sample_name_drop, "_", simplify = T)[, 1],
         # add col `unique_id`: LTC6_BPK30u72h_1_S21_L002
         unique_id = str_split(sample_name_drop, "-", simplify = T)[, 2],

         # add col `fastq`
         fastq = case_when(grepl("R1", sample_name_drop) ~ "fastq_1",
                           grepl("R2", sample_name_drop) ~ "fastq_2")) %>%

  # add col `line`: LTC6
  separate(unique_id, c('line', 'group', 'drop1', 'drop2', 'drop3')) %>%

  # add col `sample`: LTC6_BPK25_72h_1
  unite("sample", line:drop1, remove = FALSE) %>%

  # add col `unique_id`: LTC6_BPK30u72h_1_S21_L002
  unite("unique_id", line:drop3, remove = FALSE) %>%

  # group: BPK25
  mutate(group = str_sub(group, 1, 5)) %>%

  # add col `strandedness`: Strandedness for all of our samples is always "reverse".
  add_column(strandedness = "reverse") %>%

  # remove columns not needed
  select(-c(sample_name_drop, drop1, drop2, drop3)) %>%

  # add col `fastq_1`: `R1_001.fastq.gz` and col `fastq_2`: `R2_001.fastq.gz`
  pivot_wider(names_from = "fastq", values_from = "file_path") %>%

  # remove columns not needed
  select(-c(unique_id))
```

```
## Warning: Expected 5 pieces. Additional pieces discarded in 96 rows [1, 2, 3, 4,
## 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].
```

```
# head(df)
```

6 Save output file

```
write_tsv(df, file = paste0(results_dir, "/", "project-metadata", ".tsv"))
```

7 Session Info

```
## R version 4.4.0 (2024-04-24)
## Platform: x86_64-pc-linux-gnu
## Running under: Red Hat Enterprise Linux 8.4 (Ootpa)
##
## Matrix products: default
## BLAS:   /research/rgs01/applications/hpcf/authorized_apps/rhel8_apps/lapack/3.10.1/install/lib64/lib
## LAPACK: /research/rgs01/applications/hpcf/authorized_apps/rhel8_apps/lapack/3.10.1/install/lib64/lib
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## time zone: America/Chicago
## tzcode source: system (glibc)
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] lubridate_1.9.3 forcats_1.0.0   stringr_1.5.1   dplyr_1.1.4
##  [5] purrr_1.0.2     readr_2.1.5     tidyr_1.3.1     tibble_3.2.1
##  [9] ggplot2_3.5.1   tidyverse_2.0.0 yaml_2.3.10
##
## loaded via a namespace (and not attached):
##  [1] sass_0.4.9        utf8_1.2.4        generics_0.1.3    stringi_1.8.4
##  [5] hms_1.1.3         digest_0.6.37     magrittr_2.0.3    evaluate_0.24.0
##  [9] grid_4.4.0        timechange_0.3.0  fastmap_1.2.0     jsonlite_1.8.8
## [13] fansi_1.0.6       scales_1.3.0      jquerylib_0.1.4   cli_3.6.3
## [17] rlang_1.1.4       crayon_1.5.3      bit64_4.0.5       munsell_0.5.1
## [21] withr_3.0.1       cachem_1.1.0      tools_4.4.0       parallel_4.4.0
## [25] tzdb_0.4.0        colorspace_2.1-1  vctrs_0.6.5       R6_2.5.1
## [29] mime_0.12         lifecycle_1.0.4   bit_4.0.5         vroom_1.6.5
## [33] pkgconfig_2.0.3   pillar_1.9.0      bslib_0.8.0       gtable_0.3.5
## [37] glue_1.7.0        xfun_0.47         tidyselect_1.2.1  knitr_1.48
## [41] htmltools_0.5.8.1 rmarkdown_2.28    compiler_4.4.0
```