SRA_runVSexpr

Afrooz Razi

2024-06-03

SRA Experiment ID vs Run ID

In this document, I will look at number of samples based on experiment ID and run ID. I think we should add all the runs that are submitted under the same experiment ID. Before moving forward with this, we want to know how many experiments have 1 run and whether are the runs for an experiment were submitted on the same date.

```
library(tidyverse)
library(ggplot2)
library(data.table)


recount3_metadata<-read_tsv("/dcs04/hansen/data/recount_genotype/PCA/SRA/Recount3_metadata.tsv")
```

```
## Warning: One or more parsing issues, call `problems()` on your data frame for details, e.g.:
##   dat <- vroom(...)
##   problems(dat)

## Rows: 316447 Columns: 163
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr   (30): external_id, study, sample_acc, experiment_acc, submission_acc, ...
## dbl  (132): rail_id, paired_nominal_length, paired_nominal_stdev, spot_lengt...
## dttm   (1): run_published
##
## 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.
```

Number of Runs and Experiments:

We have 316,447 runs and 265,949 experiments

```
met<-recount3_metadata[,c(colnames(recount3_metadata)[1:6],"run_published","seq_type")]

#How many Runs do we have?
length(unique(met$external_id))
```

```
## [1] 316447
```

```
length(unique(met$experiment_acc))
```

```
## [1] 265949
```

What is this stat when separated based on sequencing type?

- Max number of runs in one experiment:
- Categorize the experiments based on number of runs:

```
pp<-met %>% group_by(experiment_acc,seq_type) %>% summarise(n=n())
```

```
## `summarise()` has grouped output by 'experiment_acc'. You can override using
## the `.groups` argument.
```

```
#Max number of runs in one experiment:
pp %>% group_by(seq_type) %>%  summarize(max(n))
```

```
## # A tibble: 2 x 2
##   seq_type `max(n)`
##   <chr>       <int>
## 1 bulk          120
## 2 smartseq       94
```

```
#Categorize the experiments based on number of runs:
pp2<-pp %>% mutate(run_num= case_when( n==1 ~ "1",
                                       n>1 & n<=3 ~ "1-3",
                                       n>3 & n<=6 ~ "3-6",
                                       n>6 & n<=15 ~ "6-15",
                                       n>15 & n<=25 ~ "15-25",
                                       n>25 ~ ">25"))

pp2_sum<-pp2 %>% group_by(seq_type,run_num) %>%   summarize(total_exper=n())
```

```
## `summarise()` has grouped output by 'seq_type'. You can override using the
## `.groups` argument.
```

```
pp2_sum$run_num<- factor(pp2_sum$run_num, levels= c("1","1-3","3-6","6-15","15-25",">25"))
pp2_sum<-pp2_sum[order(pp2_sum$run_num),]

print(pp2_sum)
```

```
## # A tibble: 11 x 3
## # Groups:   seq_type [2]
##    seq_type run_num total_exper
##    <chr>    <fct>         <int>
##  1 bulk     1            167067
##  2 smartseq 1             78403
##  3 bulk     1-3            8678
```

```
##  4 smartseq 1-3            4317
##  5 bulk     3-6            4809
##  6 smartseq 3-6            1011
##  7 bulk     6-15           1283
##  8 smartseq 6-15            204
##  9 bulk     15-25           132
## 10 bulk     >25              41
## 11 smartseq >25               4
```

Some runs for the same experiment ID were not published on the same date.

There are 211 experiments with more than 1 publishing date

```
nrow(publish_date_abnorm)
```

```
## [1] 211
```

```
publish_date<-met %>% mutate(run_published=date(run_published)) %>% group_by(experiment_acc,seq_type) %
```

```
## `summarise()` has grouped output by 'experiment_acc'. You can override using
## the `.groups` argument.
```

```
publish_date_abnorm<-publish_date %>% filter(pub_date>1)
head(publish_date_abnorm)
```

```
## # A tibble: 6 x 4
## # Groups:   experiment_acc [6]
##   experiment_acc seq_type pub_date run_num
##   <chr>          <chr>       <int>   <int>
## 1 DRX001126      bulk            2       4
## 2 DRX001127      bulk            3       5
## 3 DRX001128      bulk            3       6
## 4 DRX021425      bulk            2       8
## 5 DRX021429      bulk            2      18
## 6 DRX021430      bulk            2      21
```