main_r

July 10, 2021

1 Generate log2CPM for CMC data

[1]:

```r
library(tidyverse)
library(synapser)
```

```
  Attaching packages                                    tidyverse
1.3.1

  ggplot2 3.3.5      purrr   0.3.4
  tibble  3.1.2      dplyr   1.0.7
  tidyr   1.1.3      stringr 1.4.0
  readr   1.4.0      forcats 0.5.1

  Conflicts
tidyverse_conflicts()
  dplyr::filter() masks stats::filter()
  dplyr::lag()    masks stats::lag()


New synapser version detected:
    You are using synapser version 0.9.77.
    synapser version 0.10.101 is detected at http://ran.synapse.org.
    To upgrade to the latest version of synapser, please run the following
command:
    install.packages("synapser", repos="http://ran.synapse.org")



TERMS OF USE NOTICE:
  When using Synapse, remember that the terms and conditions of use require that
you:
  1) Attribute data contributors when discussing these data or results from
these data.
  2) Not discriminate, identify, or recontact individuals or groups represented
by the data.
  3) Use and contribute only data de-identified to HIPAA standards.
  4) Redistribute data only under these same terms of use.
```

[2]:

```
synLogin()
```

```
Welcome, kj.benjamin!
```

```
NULL
```

1.1 Phenotypes

[3]:

```
# Download clinical metadata
CLINICAL_ID = 'syn3354385'
clinical = data.table::fread(synGet(CLINICAL_ID, version = 4)$path)

# Download RNASeq metadata
METADATA_QC_DLPFC_ID = 'syn18358379'
metadata = data.table::fread(synGet(METADATA_QC_DLPFC_ID, version = 3)$path)

# Join clinical and RNASeq metadata
md = right_join(clinical, metadata, by = c("Individual ID" = "Individual_ID"))␣
 ↪%>%
    mutate(Dx = fct_recode(Dx, AFF_BP = "BP", AFF_BP = "AFF", Other =␣
 ↪"undetermined",
                           Control = "Control", SCZ = "SCZ")) %>%
    filter(Dx %in% c("Control", "SCZ"), Sex %in% c('XX', 'XY'))

# Compute read pair metrics and add Institution-Dx variable
md <- md %>%
  mutate(MappedRead_Pairs = Mapped_Reads/2) %>%
  mutate(`Institution-Dx` = paste0(`Institution`, "-", `Dx`)) %>%
  mutate(TotalRead_Pairs = Total_Reads/2)

# Add MDS from SNPs
mds_file = paste0('/ceph/users/jbenja13/projects/sex_sz_ria/input/commonMind/',
                  'genotypes/mds/_m/CMC_MSSM-Penn-Pitt_DLPFC_QC.mds')
mds = data.table::fread(mds_file)
colnames(mds) = gsub('C', 'snpPC', colnames(mds))

pheno_file = paste0('/ceph/users/jbenja13/projects/sex_sz_ria/input/commonMind/
 ↪',
                    'phenotypes/combine_files/_m/CMC_phenotypes_all.csv')
pheno = read.csv(pheno_file, stringsAsFactors = F)
genotypes = merge(pheno, mds, by.y='IID', by.x='Genotypes.Genotyping_Sample_ID')

genotypes = genotypes %>%
    dplyr::select("Individual_ID", starts_with("snpPC")) %>%
    rename("Individual ID"=Individual_ID)
```

```
md = md %>% left_join(genotypes, by="Individual ID") %>% distinct

md %>% dim
```

1. 858 2. 65

[4]:

```
phenotypes = md %>% select("SampleID", "Individual ID", "Institution",
                           "Reported Gender", "Dx", "Age of Death") %>%
    mutate(`Age of Death` = ifelse(`Age of Death` == "90+", "90", `Age of␣
 ↪Death`))
colnames(phenotypes) <- gsub(' ', '_', colnames(phenotypes))
phenotypes %>% data.table::fwrite("cmc_phenotypes.csv", sep=',')
phenotypes %>% head(2)
```

A data.table: 2 × 6

SampleID	Individual_ID	Institution	Reported_Gender	Dx
<chr>	<chr>	<chr>	<chr>	<fct>
MSSM_RNA_PFC_155	CMC_MSSM_087	MSSM	Female	Control
MSSM_RNA_PFC_280	CMC_MSSM_226	MSSM	Female	Control

1.2 Gene expression (counts)

1.2.1 Combined counts

[5]:

```
# Download counts (DLPFC - MSSM)
COUNT_ID = 'syn17346208'
count = data.table::fread(synGet(COUNT_ID, version=2)$path) #synapser has␣
 ↪updated without backwards compatibility
count$transcript_id.s. = NULL

# Download gene lengths (DLPFC - MSSM)
genelen_CMC = data.table::fread(synGet('syn17346397', version = 2)$path) %>%
    gather(sampleID, Length, -gene_id, -`transcript_id(s)`) %>%
    group_by(gene_id) %>%
    summarise(Length = median(Length, na.rm = T)) %>%
    ungroup() %>% data.frame()

 # Download counts (DLPFC - HBCC)
COUNT_ID = 'syn17894685'
count_HBCC = data.table::fread(synGet(COUNT_ID, version = 4)$path)
count_HBCC$transcript_id.s. = NULL

# Join HBCC and MSSM counts
NEW.COUNTS = full_join(count, count_HBCC, by = c("gene_id")) %>%
    column_to_rownames(var='gene_id') %>% t %>%
    as.data.frame %>% rownames_to_column %>%
    filter(rowname %in% md$SampleID) %>%
    column_to_rownames(var="rowname") %>% t %>%
    as.data.frame
```

```
NEW.COUNTS[1:2, 1:5]
```

A data.frame: 2 × 5

	MSSM_RNA_PFC_1	MSSM_RNA_PFC_2	MSSM_RNA_P
	<dbl>	<dbl>	<dbl>
ENSG00000000003.14	124	103	160
ENSG00000000005.5	1	0	1

[6]:

```
NEW.COUNTS %>% dim
```

1. 58347 2. 858

1.2.2 CPM transformation and save

[7]:

```
edgeR::cpm(NEW.COUNTS, log=TRUE) %>% as.data.frame %>%
    rownames_to_column %>% rename("Geneid"="rowname") %>%
    data.table::fwrite("cmc_log2cpm.tsv", sep='\t')
```

1.3 Gene annotation

[8]:

```
# Get background genes
backgroundGenes = data.frame(gene_id = rownames(NEW.COUNTS)) %>%
  mutate(id = gene_id) %>%
  separate(id, c('ensembl_gene_id','position'), sep = '\\.')

# Define biomart object
mart <- biomaRt::useMart(biomart = "ENSEMBL_MART_ENSEMBL",
                host = "uswest.ensembl.org", # Ensembl Release 99 (January 2020)
                dataset = "hsapiens_gene_ensembl")
# Query biomart
Ensemble2HGNC <- biomaRt::getBM(attributes = c("ensembl_gene_id", "hgnc_symbol",
                                        "percentage_gene_gc_content",␣
 ↪"gene_biotype",
                                        "chromosome_name"),
                      filters = "ensembl_gene_id",
                      values = backgroundGenes$ensembl_gene_id,
                      mart = mart)
```

[9]:

```
backgroundGenes %>%
    inner_join(Ensemble2HGNC, by=c("ensembl_gene_id")) %>%
    select(-c(percentage_gene_gc_content, gene_biotype, position)) %>%
    data.table::fwrite("cmc_gene_annotation.tsv", sep='\t')
```

1.4 Repredicibility Information

[10]:

```
Sys.time()
proc.time()
options(width = 120)
sessioninfo::session_info()
```

```
[1] "2021-07-10 10:12:35 EDT"

   user  system elapsed
 56.632   6.197 153.661

  Session info
 setting  value
 version  R version 4.0.3 (2020-10-10)
 os       Arch Linux
 system   x86_64, linux-gnu
 ui       X11
 language (EN)
 collate  en_US.UTF-8
 ctype    en_US.UTF-8
 tz       America/New_York
 date     2021-07-10

  Packages
 package        * version  date       lib source
 AnnotationDbi    1.52.0   2020-10-27 [1] Bioconductor
 askpass          1.1      2019-01-13 [1] CRAN (R 4.0.2)
 assertthat       0.2.1    2019-03-21 [1] CRAN (R 4.0.2)
 backports        1.2.1    2020-12-09 [1] CRAN (R 4.0.2)
 base64enc        0.1-3    2015-07-28 [1] CRAN (R 4.0.2)
 Biobase          2.50.0   2020-10-27 [1] Bioconductor
 BiocFileCache    1.14.0   2020-10-27 [1] Bioconductor
 BiocGenerics     0.36.1   2021-04-16 [1] Bioconductor
 biomaRt          2.46.3   2021-02-09 [1] Bioconductor
 bit              4.0.4    2020-08-04 [1] CRAN (R 4.0.2)
 bit64            4.0.5    2020-08-30 [1] CRAN (R 4.0.2)
 blob             1.2.1    2020-01-20 [1] CRAN (R 4.0.2)
 broom            0.7.8    2021-06-24 [1] CRAN (R 4.0.3)
 cachem           1.0.5    2021-05-15 [1] CRAN (R 4.0.3)
 cellranger       1.1.0    2016-07-27 [1] CRAN (R 4.0.2)
 cli              3.0.0    2021-06-30 [1] CRAN (R 4.0.3)
 codetools        0.2-16   2018-12-24 [2] CRAN (R 4.0.3)
 colorspace       2.0-2    2021-06-24 [1] CRAN (R 4.0.3)
 crayon           1.4.1    2021-02-08 [1] CRAN (R 4.0.3)
 curl             4.3.2    2021-06-23 [1] CRAN (R 4.0.3)
 data.table       1.14.0   2021-02-21 [1] CRAN (R 4.0.3)
 DBI              1.1.1    2021-01-15 [1] CRAN (R 4.0.2)
 dbplyr           2.1.1    2021-04-06 [1] CRAN (R 4.0.3)
 digest           0.6.27   2020-10-24 [1] CRAN (R 4.0.2)
 dplyr          * 1.0.7    2021-06-18 [1] CRAN (R 4.0.3)
 edgeR            3.32.1   2021-01-14 [1] Bioconductor
 ellipsis         0.3.2    2021-04-29 [1] CRAN (R 4.0.3)
 evaluate         0.14     2019-05-28 [1] CRAN (R 4.0.2)
 fansi            0.5.0    2021-05-25 [1] CRAN (R 4.0.3)
 fastmap          1.1.0    2021-01-25 [1] CRAN (R 4.0.2)
```

```
forcats        * 0.5.1    2021-01-27 [1] CRAN (R 4.0.2)
fs               1.5.0    2020-07-31 [1] CRAN (R 4.0.2)
generics         0.1.0    2020-10-31 [1] CRAN (R 4.0.2)
ggplot2        * 3.3.5    2021-06-25 [1] CRAN (R 4.0.3)
glue             1.4.2    2020-08-27 [1] CRAN (R 4.0.2)
gtable           0.3.0    2019-03-25 [1] CRAN (R 4.0.2)
haven            2.4.1    2021-04-23 [1] CRAN (R 4.0.3)
hms              1.1.0    2021-05-17 [1] CRAN (R 4.0.3)
htmltools        0.5.1.1  2021-01-22 [1] CRAN (R 4.0.2)
httr             1.4.2    2020-07-20 [1] CRAN (R 4.0.2)
IRanges          2.24.1   2020-12-12 [1] Bioconductor
IRdisplay        1.0      2021-01-20 [1] CRAN (R 4.0.2)
IRkernel         1.2      2021-05-11 [1] CRAN (R 4.0.3)
jsonlite         1.7.2    2020-12-09 [1] CRAN (R 4.0.2)
lattice          0.20-41  2020-04-02 [2] CRAN (R 4.0.3)
lifecycle        1.0.0    2021-02-15 [1] CRAN (R 4.0.3)
limma            3.46.0   2020-10-27 [1] Bioconductor
locfit           1.5-9.4  2020-03-25 [1] CRAN (R 4.0.2)
lubridate        1.7.10   2021-02-26 [1] CRAN (R 4.0.3)
magrittr         2.0.1    2020-11-17 [1] CRAN (R 4.0.2)
memoise          2.0.0    2021-01-26 [1] CRAN (R 4.0.2)
modelr           0.1.8    2020-05-19 [1] CRAN (R 4.0.2)
munsell          0.5.0    2018-06-12 [1] CRAN (R 4.0.2)
openssl          1.4.4    2021-04-30 [1] CRAN (R 4.0.3)
pack             0.1-1    2021-02-23 [1] local
pbdZMQ           0.3-5    2021-02-10 [1] CRAN (R 4.0.3)
pillar           1.6.1    2021-05-16 [1] CRAN (R 4.0.3)
pkgconfig        2.0.3    2019-09-22 [1] CRAN (R 4.0.2)
prettyunits      1.1.1    2020-01-24 [1] CRAN (R 4.0.2)
progress         1.2.2    2019-05-16 [1] CRAN (R 4.0.2)
purrr          * 0.3.4    2020-04-17 [1] CRAN (R 4.0.2)
PythonEmbedInR   0.6.76   2021-02-23 [1] local
R6               2.5.0    2020-10-28 [1] CRAN (R 4.0.2)
rappdirs         0.3.3    2021-01-31 [1] CRAN (R 4.0.2)
Rcpp             1.0.7    2021-07-07 [1] CRAN (R 4.0.3)
readr          * 1.4.0    2020-10-05 [1] CRAN (R 4.0.2)
readxl           1.3.1    2019-03-13 [1] CRAN (R 4.0.2)
repr             1.1.3    2021-01-21 [1] CRAN (R 4.0.2)
reprex           2.0.0    2021-04-02 [1] CRAN (R 4.0.3)
rlang            0.4.11   2021-04-30 [1] CRAN (R 4.0.3)
RSQLite          2.2.7    2021-04-22 [1] CRAN (R 4.0.3)
rstudioapi       0.13     2020-11-12 [1] CRAN (R 4.0.2)
rvest            1.0.0    2021-03-09 [1] CRAN (R 4.0.3)
S4Vectors        0.28.1   2020-12-09 [1] Bioconductor
scales           1.1.1    2020-05-11 [1] CRAN (R 4.0.2)
sessioninfo      1.1.1    2018-11-05 [1] CRAN (R 4.0.2)
stringi          1.6.2    2021-05-17 [1] CRAN (R 4.0.3)
stringr        * 1.4.0    2019-02-10 [1] CRAN (R 4.0.2)
```

```
 synapser      * 0.9.77   2021-02-23 [1] local
 tibble        * 3.1.2    2021-05-16 [1] CRAN (R 4.0.3)
 tidyr         * 1.1.3    2021-03-03 [1] CRAN (R 4.0.3)
 tidyselect      1.1.1    2021-04-30 [1] CRAN (R 4.0.3)
 tidyverse     * 1.3.1    2021-04-15 [1] CRAN (R 4.0.3)
 utf8            1.2.1    2021-03-12 [1] CRAN (R 4.0.3)
 uuid            0.1-4    2020-02-26 [1] CRAN (R 4.0.2)
 vctrs           0.3.8    2021-04-29 [1] CRAN (R 4.0.3)
 withr           2.4.2    2021-04-18 [1] CRAN (R 4.0.3)
 XML             3.99-0.6 2021-03-16 [1] CRAN (R 4.0.3)
 xml2            1.3.2    2020-04-23 [1] CRAN (R 4.0.2)

[1] /home/jbenja13/R/x86_64-pc-linux-gnu-library/4.0
[2] /usr/lib/R/library
```

[ ]: