R SQL Data Manipulation

Lavinia Carabet

```
#install.packages(c("dplyr", "dbplyr", "RSQLite", "stringr", "purrr"))
```

Load libraries

```
library(dplyr)
library(dbplyr)
library(RSQLite)
library(stringr)
library(purrr)
```

Connect to database

```
cuff_data.db <- DBI::dbConnect(RSQLite::SQLite(), './cuffData.db')
```

List database tables

```
dplyr::tbl(cuff_data.db, sql("SELECT name FROM sqlite_master WHERE type='table'"))
```

```
## # Source:   SQL [?? x 1]
## # Database: sqlite 3.37.0 [D:\Repository\Portfolio\DDL_DML\SQL_R\cuffData.db]
##    name
##    <chr>
##  1 genes
##  2 biasData
##  3 samples
##  4 TSS
##  5 TSSData
##  6 CDS
##  7 CDSData
##  8 splicingDiffData
##  9 TSSExpDiffData
## 10 CDSDiffData
## # ... with more rows
```

```
RSQLite::dbListTables(cuff_data.db)
```

```
##  [1] "CDS"                   "CDSCount"              "CDSData"
##  [4] "CDSDiffData"           "CDSExpDiffData"        "CDSFeatures"
##  [7] "CDSReplicateData"      "TSS"                   "TSSCount"
## [10] "TSSData"               "TSSExpDiffData"        "TSSFeatures"
## [13] "TSSReplicateData"      "attributes"            "biasData"
## [16] "features"              "geneCount"             "geneData"
## [19] "geneExpDiffData"       "geneFeatures"          "geneReplicateData"
## [22] "genes"                 "isoformCount"          "isoformData"
## [25] "isoformExpDiffData"    "isoformFeatures"       "isoformReplicateData"
## [28] "isoforms"              "model_transcripts"     "phenoData"
## [31] "promoterDiffData"      "replicates"            "runInfo"
## [34] "samples"               "splicingDiffData"      "sqlite_sequence"
## [37] "varModel"
```

Querying the database

With SQL syntax

```
tbl(cuff_data.db, sql("SELECT * FROM genes LIMIT 10"))
```

```
## # Source:   SQL [?? x 7]
## # Database: sqlite 3.37.0 [D:\Repository\Portfolio\DDL_DML\SQL_R\cuffData.db]
##    gene_id     class_code nearest_ref_id gene_short_name locus   length coverage
##    <chr>       <chr>      <chr>          <chr>           <chr>    <int>    <dbl>
##  1 XLOC_000001 <NA>       <NA>           DDX11L1         chr1:1~     NA       NA
##  2 XLOC_000002 <NA>       <NA>           MIR1302-2       chr1:3~     NA       NA
##  3 XLOC_000003 <NA>       <NA>           OR4F5           chr1:6~     NA       NA
##  4 XLOC_000004 <NA>       <NA>           LOC100287934    chr1:7~     NA       NA
##  5 XLOC_000005 <NA>       <NA>           LOC100287934    chr1:7~     NA       NA
##  6 XLOC_000006 <NA>       <NA>           FAM87B          chr1:8~     NA       NA
##  7 XLOC_000007 <NA>       <NA>           LINC01128       chr1:8~     NA       NA
##  8 XLOC_000008 <NA>       <NA>           LOC284600       chr1:9~     NA       NA
##  9 XLOC_000009 <NA>       <NA>           SAMD11          chr1:9~     NA       NA
## 10 XLOC_000010 <NA>       <NA>           KLHL17          chr1:9~     NA       NA
```

With dplyr syntax

```
genes <- tbl(cuff_data.db, "genes")

genes %>%
  select(gene_id, gene_short_name, locus) %>%
  head(10)
```

```
## # Source:   lazy query [?? x 3]
## # Database: sqlite 3.37.0 [D:\Repository\Portfolio\DDL_DML\SQL_R\cuffData.db]
##    gene_id     gene_short_name locus
##    <chr>       <chr>           <chr>
##  1 XLOC_000001 DDX11L1         chr1:11873-29370
##  2 XLOC_000002 MIR1302-2       chr1:30365-30503
##  3 XLOC_000003 OR4F5           chr1:69090-70008
##  4 XLOC_000004 LOC100287934    chr1:764864-810022
##  5 XLOC_000005 LOC100287934    chr1:764864-810022
```

```
##  6 XLOC_000006 FAM87B          chr1:817370-819834
##  7 XLOC_000007 LINC01128       chr1:827590-859446
##  8 XLOC_000008 LOC284600       chr1:911422-914782
##  9 XLOC_000009 SAMD11          chr1:925740-959309
## 10 XLOC_000010 KLHL17          chr1:960586-965897
```

```
show_query(head(genes, 10))
```

```
## <SQL>
## SELECT *
## FROM `genes`
## LIMIT 10
```

```
genes %>%
  filter(substr(locus,1,4) == "chr1") %>%
  select(gene_id, gene_short_name, locus)  %>%
  head(10)
```

```
## # Source:   lazy query [?? x 3]
## # Database: sqlite 3.37.0 [D:\Repository\Portfolio\DDL_DML\SQL_R\cuffData.db]
##    gene_id     gene_short_name locus
##    <chr>       <chr>           <chr>
##  1 XLOC_000001 DDX11L1         chr1:11873-29370
##  2 XLOC_000002 MIR1302-2       chr1:30365-30503
##  3 XLOC_000003 OR4F5           chr1:69090-70008
##  4 XLOC_000004 LOC100287934    chr1:764864-810022
##  5 XLOC_000005 LOC100287934    chr1:764864-810022
##  6 XLOC_000006 FAM87B          chr1:817370-819834
##  7 XLOC_000007 LINC01128       chr1:827590-859446
##  8 XLOC_000008 LOC284600       chr1:911422-914782
##  9 XLOC_000009 SAMD11          chr1:925740-959309
## 10 XLOC_000010 KLHL17          chr1:960586-965897
```

```
genes %>%
  filter(substr(locus,1,4) == "chr1") %>%
  select(gene_id, gene_short_name, locus) %>%
  show_query()
```

```
## <SQL>
## SELECT `gene_id`, `gene_short_name`, `locus`
## FROM `genes`
## WHERE (SUBSTR(`locus`, 1, 4) = 'chr1')
```

```
genes %>%
  collect() %>%
  group_by(substr(locus,1,str_locate(locus, ":"))) %>%
  tally()
```

```
## # A tibble: 289 x 2
##    `substr(locus, 1, str_locate(locus, ":"))`     n
##    <chr>                                      <int>
```

```
##  1 chr1:                                      3326
##  2 chr1_GL383518v1_alt:                          3
##  3 chr1_GL383519v1_alt:                          9
##  4 chr1_GL383520v2_alt:                          2
##  5 chr1_KI270706v1_random:                       5
##  6 chr1_KI270711v1_random:                       4
##  7 chr1_KI270712v1_random:                       1
##  8 chr1_KI270713v1_random:                       5
##  9 chr1_KI270714v1_random:                       3
## 10 chr1_KI270759v1_alt:                          1
## # ... with 279 more rows
```

```
genes %>%
  collect() %>%
  filter(str_starts(locus, "chr1:")) %>%
  tally()
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1  3326
```

```
genes %>%
  collect() %>%
  tally()
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1 37167
```

```
genes %>%
  summarize(
    n=n_distinct(gene_id)
    )
```

```
## # Source:   lazy query [?? x 1]
## # Database: sqlite 3.37.0 [D:\Repository\Portfolio\DDL_DML\SQL_R\cuffData.db]
##       n
##   <int>
## 1 37167
```

```
genes %>%
  summarize(
    n=n_distinct(gene_id)
    ) %>%
  show_query()
```

```
## <SQL>
## SELECT COUNT(DISTINCT `gene_id`) AS `n`
## FROM `genes`
```

```
geneExpDiffData <- tbl(cuff_data.db, "geneExpDiffData")
head(geneExpDiffData, 10)
```

```
## # Source:   lazy query [?? x 11]
## # Database: sqlite 3.37.0 [D:\Repository\Portfolio\DDL_DML\SQL_R\cuffData.db]
##    gene_id   sample_1 sample_2 status value_1 value_2 log2_fold_change test_stat
##    <chr>     <chr>    <chr>    <chr>    <dbl>   <dbl>            <dbl>     <dbl>
##  1 XLOC_000~ C1       C2       NOTEST  0.120   0.187            0.641      0
##  2 XLOC_000~ C1       C2       NOTEST  0       0                0          0
##  3 XLOC_000~ C1       C2       NOTEST  0       0                0          0
##  4 XLOC_000~ C1       C2       OK      0.0591  0.379            2.68       1.97
##  5 XLOC_000~ C1       C2       OK      0.572   0.452           -0.339     -0.414
##  6 XLOC_000~ C1       C2       NOTEST  0.0384  0.0173          -1.15       0
##  7 XLOC_000~ C1       C2       OK      3.83    3.92             0.0303     0.131
##  8 XLOC_000~ C1       C2       NOTEST  0       0                0          0
##  9 XLOC_000~ C1       C2       NOTEST  0.0124  0.0141           0.192      0
## 10 XLOC_000~ C1       C2       OK      3.45    3.14            -0.136     -0.535
## # ... with 3 more variables: p_value <dbl>, q_value <dbl>, significant <chr>
```

```
geneExpDiffData %>%
  collect() %>%
  tally()
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1 37167
```

```
geneData <- tbl(cuff_data.db, "geneData")
head(geneData, 10)
```

```
## # Source:   lazy query [?? x 6]
## # Database: sqlite 3.37.0 [D:\Repository\Portfolio\DDL_DML\SQL_R\cuffData.db]
##    gene_id     sample_name   fpkm conf_hi conf_lo quant_status
##    <chr>       <chr>        <dbl>   <dbl>   <dbl> <chr>
##  1 XLOC_000001 C1          0.120    0.315  0       OK
##  2 XLOC_000001 C2          0.187    0.409  0       OK
##  3 XLOC_000002 C1          0        0      0       OK
##  4 XLOC_000002 C2          0        0      0       OK
##  5 XLOC_000003 C1          0        0      0       OK
##  6 XLOC_000003 C2          0        0      0       OK
##  7 XLOC_000004 C1          0.0591   0.159  0       OK
##  8 XLOC_000004 C2          0.379    0.693  0.0648  OK
##  9 XLOC_000005 C1          0.572    0.973  0.171   OK
## 10 XLOC_000005 C2          0.452    0.856  0.0485  OK
```

```
geneData %>%
  collect() %>%
  tally()
```

```
## # A tibble: 1 x 1
```

```
##       n
##   <int>
## 1 74334
```

Joins

With dplyr

```
genes %>%
  inner_join(geneExpDiffData, by="gene_id") %>%
  select(-c(class_code, nearest_ref_id, length, coverage)) %>%
  filter(status == "OK" & significant == "yes") %>%
  head(10)
```

```
## # Source:   lazy query [?? x 13]
## # Database: sqlite 3.37.0 [D:\Repository\Portfolio\DDL_DML\SQL_R\cuffData.db]
##    gene_id     gene_short_name locus    sample_1 sample_2 status value_1 value_2
##    <chr>       <chr>           <chr>    <chr>    <chr>    <chr>    <dbl>   <dbl>
##  1 XLOC_000437 RP4-533D7.5     chr1:46~ C1       C2       OK       0.269   0
##  2 XLOC_002533 <NA>            chr1:16~ C1       C2       OK       0.552   0
##  3 XLOC_002557 LOC101928484    chr1:16~ C1       C2       OK       0       0.205
##  4 XLOC_002620 <NA>            chr1:17~ C1       C2       OK       0       0.491
##  5 XLOC_002716 <NA>            chr1:20~ C1       C2       OK       0.339   0
##  6 XLOC_003017 <NA>            chr1:11~ C1       C2       OK       0       0.287
##  7 XLOC_003150 <NA>            chr1:11~ C1       C2       OK       0       0.223
##  8 XLOC_003196 <NA>            chr1:15~ C1       C2       OK       0       0.389
##  9 XLOC_003197 <NA>            chr1:15~ C1       C2       OK       0       0.548
## 10 XLOC_003248 <NA>            chr1:19~ C1       C2       OK       0       0.256
## # ... with 5 more variables: log2_fold_change <dbl>, test_stat <dbl>,
## #   p_value <dbl>, q_value <dbl>, significant <chr>
```

```
genes %>%
  inner_join(geneExpDiffData, by="gene_id") %>%
  select(-c(class_code, nearest_ref_id, length, coverage)) %>%
  filter(status == "OK" & significant == "yes") %>%
  show_query()
```

```
## <SQL>
## SELECT *
## FROM (SELECT `gene_id`, `gene_short_name`, `locus`, `sample_1`, `sample_2`, `status`, `value_1`, `val
## FROM (SELECT `LHS`.`gene_id` AS `gene_id`, `class_code`, `nearest_ref_id`, `gene_short_name`, `locus`
## FROM `genes` AS `LHS`
## INNER JOIN `geneExpDiffData` AS `RHS`
## ON (`LHS`.`gene_id` = `RHS`.`gene_id`)
## ))
## WHERE (`status` = 'OK' AND `significant` = 'yes')
```

```
genes %>%
  inner_join(geneExpDiffData, by="gene_id") %>%
  select(-c(class_code, nearest_ref_id, length, coverage)) %>%
  filter(status == "OK" & significant == "yes") %>%
  tally()
```

```
## # Source:   lazy query [?? x 1]
## # Database: sqlite 3.37.0 [D:\Repository\Portfolio\DDL_DML\SQL_R\cuffData.db]
##       n
##   <int>
## 1 10116
```

With purrr

```
library(purrr)

diffExp <- reduce(
  list(genes, geneExpDiffData),
  right_join) %>%
  select(-c(class_code, nearest_ref_id, length, coverage)) %>%
  filter(status == "OK" & significant == "yes")
```

```
## Joining, by = "gene_id"
```

```
diffExp %>% head(10)
```

```
## # Source:   lazy query [?? x 13]
## # Database: sqlite 3.37.0 [D:\Repository\Portfolio\DDL_DML\SQL_R\cuffData.db]
##    gene_id     gene_short_name locus    sample_1 sample_2 status value_1 value_2
##    <chr>       <chr>           <chr>    <chr>    <chr>    <chr>    <dbl>   <dbl>
##  1 XLOC_000437 RP4-533D7.5     chr1:46~ C1       C2       OK       0.269   0
##  2 XLOC_002533 <NA>            chr1:16~ C1       C2       OK       0.552   0
##  3 XLOC_002557 LOC101928484    chr1:16~ C1       C2       OK       0       0.205
##  4 XLOC_002620 <NA>            chr1:17~ C1       C2       OK       0       0.491
##  5 XLOC_002716 <NA>            chr1:20~ C1       C2       OK       0.339   0
##  6 XLOC_003017 <NA>            chr1:11~ C1       C2       OK       0       0.287
##  7 XLOC_003150 <NA>            chr1:11~ C1       C2       OK       0       0.223
##  8 XLOC_003196 <NA>            chr1:15~ C1       C2       OK       0       0.389
##  9 XLOC_003197 <NA>            chr1:15~ C1       C2       OK       0       0.548
## 10 XLOC_003248 <NA>            chr1:19~ C1       C2       OK       0       0.256
## # ... with 5 more variables: log2_fold_change <dbl>, test_stat <dbl>,
## #   p_value <dbl>, q_value <dbl>, significant <chr>
```

```
diffExp %>% tally()
```

```
## # Source:   lazy query [?? x 1]
## # Database: sqlite 3.37.0 [D:\Repository\Portfolio\DDL_DML\SQL_R\cuffData.db]
##       n
##   <int>
## 1 10116
```

```
library(purrr)

diffExp <- reduce(
  list(genes, geneExpDiffData),
  right_join) %>%
  select(-c(class_code, nearest_ref_id, length, coverage)) %>%
  filter(status == "OK" & significant == "yes" & !is.na(test_stat)) %>%
  arrange(log2_fold_change, q_value)
```

```
## Joining, by = "gene_id"
```

```
diffExp <- reduce(
  list(genes, geneExpDiffData),
  right_join) %>%
  select(-c(class_code, nearest_ref_id, length, coverage)) %>%
  filter(status == "OK" & significant == "yes" & !is.na(test_stat)) %>%
  arrange(log2_fold_change, q_value) %>%
  show_query()
```

```
## Joining, by = "gene_id"
```

```
## <SQL>
## SELECT *
## FROM (SELECT `gene_id`, `gene_short_name`, `locus`, `sample_1`, `sample_2`, `status`, `value_1`, `val
## FROM (SELECT `RHS`.`gene_id` AS `gene_id`, `class_code`, `nearest_ref_id`, `gene_short_name`, `locus`
## FROM `geneExpDiffData` AS `RHS`
## LEFT JOIN `genes` AS `LHS`
## ON (`LHS`.`gene_id` = `RHS`.`gene_id`)
## ))
## WHERE (`status` = 'OK' AND `significant` = 'yes' AND NOT(((`test_stat`) IS NULL)))
## ORDER BY `log2_fold_change`, `q_value`
```

```
diffExp %>% head(10)
```

```
## # Source:     lazy query [?? x 13]
## # Database:   sqlite 3.37.0 [D:\Repository\Portfolio\DDL_DML\SQL_R\cuffData.db]
## # Ordered by: log2_fold_change, q_value
##    gene_id     gene_short_name locus    sample_1 sample_2 status value_1 value_2
##    <chr>       <chr>           <chr>    <chr>    <chr>    <chr>    <dbl>   <dbl>
##  1 XLOC_023782 PFKFB4          chr3:48~ C1       C2       OK     6.76e+0 1.06e-1
##  2 XLOC_033415 NUDT18          chr8:22~ C1       C2       OK     4.77e+0 9.42e-2
##  3 XLOC_008085 <NA>            chr12:1~ C1       C2       OK     5.93e-1 1.41e-2
##  4 XLOC_027328 GPRIN1          chr5:17~ C1       C2       OK     2.65e+0 6.47e-2
##  5 XLOC_014414 GNGT2           chr17:4~ C1       C2       OK     3.45e-1 9.11e-3
##  6 XLOC_013802 SLC25A10        chr17:8~ C1       C2       OK     3.22e+1 8.66e-1
##  7 XLOC_014445 LOC101927337    chr17:5~ C1       C2       OK     2.66e+3 7.47e+1
##  8 XLOC_006681 GPR162          chr12:6~ C1       C2       OK     6.27e-1 1.77e-2
##  9 XLOC_001530 C1orf233        chr1:15~ C1       C2       OK     5.37e+0 1.66e-1
## 10 XLOC_028097 C6orf223        chr6:43~ C1       C2       OK     2.82e+0 9.58e-2
## # ... with 5 more variables: log2_fold_change <dbl>, test_stat <dbl>,
## #   p_value <dbl>, q_value <dbl>, significant <chr>
```

```
diffExp %>% tally()
```

```
## # Source:   lazy query [?? x 1]
## # Database: sqlite 3.37.0 [D:\Repository\Portfolio\DDL_DML\SQL_R\cuffData.db]
##       n
##   <int>
## 1 10034
```