ccledb

Vincent J. Carey, stvjc at channing.harvard.edu based on work of Phil Chapman at CRUK

July 2015

Contents

1 Introduction

A large indexed SQLite database has been created by Phil Chapman to represent CCLE, Achilles, and other integrative data sources relevant to cancer biology.

This document describes some approaches to user interface design. We indicate how to

- query the database directly
- provide R-level support for substantively useful self-description
- carry out some exemplary 'analyses'

For this code to work you need to have the environment variable `CCLEDB_PATH` defined to give the path to the SQLite file.

2 An R6 interface

I believe that an object that is somewhat fleshed out relative to the database view provided by dplyr will come in handy. Therefore I defined a reference class and have lightly populated it with some identifier vectors.

```
ccle = ccledb$new(.ccleSrc)
```

```
## creating guide vectors (need ~10 seconds...)
## done.
```

```
ccle
```

```
## ccledb instance. Components:
##   cell_lines: (1046) 1321N1_CENTRAL_NERVOUS_SYSTEM 143B_BONE 22RV1_PROS ...
##   cngenes: (21217) A1BG A1CF A2BP1 A2LD1 A2M A2ML1 A4GALT A4GNT AAA1  ...
##   compounds: (24) 17-AAG AEW541 AZD0530 AZD6244 Erlotinib Irinotecan ...
##   hcgenes: (1667) 36403 37134 37499 38595 AAK1 AATK ABCA3 ABCC3 ABCC ...
##   organs: (24) autonomic_ganglia biliary_tract bone breast centra ...
## ---
## use $src for reference to dplyr src_sqlite:
## src:  sqlite 3.8.6 [/Users/stvjc/Research/CCLE_CHAPMAN/CellLineData.db]
## tbls: achilles_v243, ccle_cell_line_info, ccle_copynumber_tall,
```

```
##   ccle_drug_data, ccle_exprs, ccle_exprs_tall, ccle_hybrid_capture,
##   ccle_id_mapping, ccle_oncomap, ccle_rnaseq, ccle_screening_ic50_wide,
##   cosmicclp_mutations, gdsc_cell_line_info, gdsc_copynumber_wide,
##   gdsc_genetic_tall, gdsc_mutation_wide, gdsc_screening,
##   gdsc_screening_ic50, gdsc_screening_ic50_wide
```

```
"BRAF" %in% ccle$cngenes  # copy number gene list
```

```
## [1] TRUE
```

I have not done much with this yet but I think this will help with subsetting and shiny query support.

3 dplyr-based interaction

We can get a feel for the contents with some simple commands. `.ccleSrc` is a globally defined dplyr `src`.

```
library(ccledb)
.ccleSrc
```

```
## src:  sqlite 3.8.6 [/Users/stvjc/Research/CCLE_CHAPMAN/CellLineData.db]
## tbls: achilles_v243, ccle_cell_line_info, ccle_copynumber_tall,
##   ccle_drug_data, ccle_exprs, ccle_exprs_tall, ccle_hybrid_capture,
##   ccle_id_mapping, ccle_oncomap, ccle_rnaseq, ccle_screening_ic50_wide,
##   cosmicclp_mutations, gdsc_cell_line_info, gdsc_copynumber_wide,
##   gdsc_genetic_tall, gdsc_mutation_wide, gdsc_screening,
##   gdsc_screening_ic50, gdsc_screening_ic50_wide
```

```
.ccleSrc %>% tbl("ccle_cell_line_info") %>% head()
```

```
##   row_names                     CCLE_name Cell_line_primary_name
## 1         1 1321N1_CENTRAL_NERVOUS_SYSTEM                 1321N1
## 2         2                    143B_BONE                   143B
## 3         3                22RV1_PROSTATE                 22Rv1
## 4         4               2313287_STOMACH               23132/87
## 5         5          253JBV_URINARY_TRACT                253J-BV
## 6         6            253J_URINARY_TRACT                  253J
##   Cell_line_aliases Gender           Site_Primary    Histology
## 1                         M central_nervous_system       glioma
## 2                         F                   bone osteosarcoma
## 3                         M               prostate    carcinoma
## 4                         M                stomach    carcinoma
## 5                         U          urinary_tract    carcinoma
## 6                                    urinary_tract    carcinoma
##                 Hist_Subtype1
## 1                 astrocytoma
## 2                          NS
## 3                          NS
## 4              adenocarcinoma
## 5 transitional_cell_carcinoma
## 6 transitional_cell_carcinoma
##                                                                                                  Notes
## 1                                Identical lines: U-118 MG, U-138 MG and 1321N1 share high SNP identity
## 2 Identical lines: HTK-, HOS and 143B share high SNP identity and are very likely to be osteosarcoma
## 3
## 4
```

```
## 5                                        Identical lines: 253J and 253J-BV share high SNP identity
## 6                                        Identical lines: 253J and 253J-BV share high SNP identity
##   Source                              Expression_arrays
## 1  ECACC  NIECE_p_NCLE_RNA3_HG-U133_Plus_2_B06_296024
## 2   ATCC  MAKER_p_NCLE_RNA7_HG-U133_Plus_2_F09_454702
## 3   ATCC  NIECE_p_NCLE_RNA3_HG-U133_Plus_2_F06_296120
## 4   DSMZ  WATCH_p_NCLE_RNA8_HG-U133_Plus_2_E11_474718
## 5   KCLB CRAZY_p_NCLE_RNA10_HG-U133_Plus_2_A05_569490
## 6   KCLB CRAZY_p_NCLE_RNA10_HG-U133_Plus_2_A03_569510
##                                        SNP_arrays Oncomap
## 1 HONEY_p_NCLE_DNAAffy3_S_GenomeWideSNP_6_E09_293392     yes
## 2   BOWER_p_NCLE_DNAAffy8_GenomeWideSNP_6_D02_464552     yes
## 3      LIMPS_p_NCLE_DNA2N_GenomeWideSNP_6_C09_246674     yes
## 4   CHARY_p_NCLE_DNAAffy9_GenomeWideSNP_6_D06_490336     yes
## 5                                                        yes
## 6                                                        yes
##   Hybrid_Capture_Sequencing
## 1
## 2
## 3                       yes
## 4                       yes
## 5                       yes
## 6                       yes
```

4 Gene annotation

Expression data are provided at probe-set level, with symbols.

```
.ccleSrc %>% tbl("ccle_exprs_tall") %>% head()
```

```
##   row_names         Name  Description        Tumor_Sample_Barcode   Signal
## 1         1 100009676_at LOC100009676 LN18_CENTRAL_NERVOUS_SYSTEM 5.987545
## 2         2     10000_at         AKT3 LN18_CENTRAL_NERVOUS_SYSTEM 6.230233
## 3         3     10001_at         MED6 LN18_CENTRAL_NERVOUS_SYSTEM 9.363550
## 4         4     10002_at        NR2E3 LN18_CENTRAL_NERVOUS_SYSTEM 3.803069
## 5         5     10003_at      NAALAD2 LN18_CENTRAL_NERVOUS_SYSTEM 3.586430
## 6         6 100048912_at   CDKN2B-AS1 LN18_CENTRAL_NERVOUS_SYSTEM 3.824073
```

5 Compound annotation

```
.ccleSrc %>% tbl("ccle_drug_data") %>% head()
```

```
##   row_names                    CCLE_Cell_Line_Name Primary_Cell_Line_Name
## 1         1          1321N1_CENTRAL_NERVOUS_SYSTEM                1321N1
## 2         2                       22RV1_PROSTATE                  22Rv1
## 3         3          42MGBA_CENTRAL_NERVOUS_SYSTEM               42-MG-BA
## 4         4                     5637_URINARY_TRACT                  5637
## 5         5                     639V_URINARY_TRACT                 639-V
## 6         6 697_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE                   697
##   Compound Target                           Doses__uM_
## 1   17-AAG  HSP90 .0025,.0080,.025,.080,.25,.80,2.53,8
```

```
## 2   17-AAG  HSP90 .0025,.0080,.025,.080,.25,.80,2.53,8
## 3   17-AAG  HSP90 .0025,.0080,.025,.080,.25,.80,2.53,8
## 4   17-AAG  HSP90 .0025,.0080,.025,.080,.25,.80,2.53,8
## 5   17-AAG  HSP90 .0025,.0080,.025,.080,.25,.80,2.53,8
## 6   17-AAG  HSP90 .0025,.0080,.025,.080,.25,.80,2.53,8
##               Activity_Data__median_
## 1 -58,-7.5,-1.7,7.04,-58,-70,-73,-73
## 2 -5.9,-14,-.068,-15,-38,-71,-74,-80
## 3   -4.2,-62,-19,-72,-69,-84,-78,-91
## 4 3.24,-2.6,9.10,-35,-87,-91,-91,-92
## 5   -10,-14,-1.2,-23,-74,-87,-90,-91
## 6 -13,-1.1,17.2,7.23,-24,-77,-91,-92
##                                Activity_SD Num_Data FitType  EC50__uM_
## 1    43.3,11.5,3.65,2.85,.28,.18,3.31,.64        8 Sigmoid 0.19367672
## 2   .35,8.56,.53,21.6,1.08,1.37,3.71,2.17        8 Sigmoid 0.26721454
## 3 19.1,42.8,13.5,6.12,7.84,16.9,5.98,7.26        8 Sigmoid 0.05208059
## 4  5.69,14.8,29.2,30.4,5.11,.39,1.98,.041        8 Sigmoid 0.06090715
## 5  6.48,.16,2.77,2.31,4.07,1.69,.035,.036        8 Sigmoid 0.14792989
## 6   11.4,20.7,2.72,12.3,25.4,.66,.83,1.75        8 Sigmoid 0.37833494
##    IC50__uM_      Amax ActArea
## 1 0.22807844 -72.12381  3.0302
## 2 0.32970169 -76.30148  3.0622
## 3 0.05303809 -80.37766  5.0587
## 4 0.07082279 -91.65148  3.5350
## 5 0.15009449 -89.63907  3.7820
## 6 0.42257124 -91.79781  3.6605
```