Assess the relation between co-expression and PPI in scRNA-seq data

Wei Sun

2020-09-01

Contents

libraries and path

```
grp = "PFC_L2_3"
data.dir.github = "../../ideas/Autism/data/"

library(MASS)
library(Matrix)
library(data.table)
library(dplyr)
library(doParallel)
library(svd)

library(ggplot2)
library(ggpubr)
library(ggpointdensity)
theme_set(theme_bw())
```

read in cell information and count data

```
cell_info = fread(file.path(data.dir.github, "meta.tsv"),
                  stringsAsFactors=TRUE)
dim(cell_info)
```

```
## [1] 104559     16
```

```
cell_info[1:2,]
```

```
##                            cell     cluster    sample individual region age sex
## 1: AAACCTGGTACGCACC-1_1823_BA24 Neu-NRGN-II 1823_BA24       1823    ACC  15   M
## 2: AAACGGGCACCAGATT-1_1823_BA24        L5/6 1823_BA24       1823    ACC  15   M
##    diagnosis Capbatch Seqbatch post-mortem interval (hours)
## 1:   Control      CB8      SB3                           18
## 2:   Control      CB8      SB3                           18
##    RNA Integrity Number genes  UMIs RNA mitochondr. percent
## 1:                    7   622   774               2.4547804
## 2:                    7  6926 24042               0.4450545
##    RNA ribosomal percent
## 1:             1.4211886
## 2:             0.4284169
```

```
dat1 = readRDS(file.path(data.dir.github, sprintf("ct_mtx/%s.rds", grp)))
class(dat1)
```

```
## [1] "dgCMatrix"
## attr(,"package")
## [1] "Matrix"
```

```
dim(dat1)
```

```
## [1] 18041  8626
```

```
dat1[1:5,1:4]
```

```
## 5 x 4 sparse Matrix of class "dgCMatrix"
##          AAACCTGCACCCATTC-1_4341_BA46 AAACGGGGTCGGCATC-1_4341_BA46
## DNAJC11                            1                            3
## NADK                               .                            .
## MASP2                              .                            .
## CLCN6                              .                            .
## TNFRSF1B                           .                            .
##          AAAGATGCAGCGTCCA-1_4341_BA46 AAAGATGGTCCGAATT-1_4341_BA46
## DNAJC11                            .                            .
## NADK                               .                            .
## MASP2                              .                            .
## CLCN6                              .                            .
## TNFRSF1B                           .                            .
```

subset cell information

```
table(colnames(dat1) %in% cell_info$cell)
```

```
##
## TRUE
## 8626
```

```
meta_cell = cell_info[match(colnames(dat1), cell_info$cell),]
dim(meta_cell)
```

```
## [1] 8626   16
```

```
meta_cell[1:2,]
```

```
##                            cell cluster    sample individual region age sex
## 1: AAACCTGCACCCATTC-1_4341_BA46    L2/3 4341_BA46       4341    PFC  13   M
## 2: AAACGGGGTCGGCATC-1_4341_BA46    L2/3 4341_BA46       4341    PFC  13   M
##    diagnosis Capbatch Seqbatch post-mortem interval (hours)
## 1:   Control      CB6      SB2                          16
## 2:   Control      CB6      SB2                          16
##    RNA Integrity Number genes  UMIs RNA mitochondr. percent
## 1:                  7.2  3967  8526               0.4691532
## 2:                  7.2  6891 23815               0.3023305
##    RNA ribosomal percent
## 1:             0.5160685
## 2:             0.4870880
```

```
names(meta_cell)[11:12] = c("PMI", "RIN")
names(meta_cell)[15:16] = c("mitoPercent", "riboPercent")
dim(meta_cell)
```

```
## [1] 8626   16
```

```
meta_cell[1:2,]
```

```
##                            cell cluster    sample individual region age sex
## 1: AAACCTGCACCCATTC-1_4341_BA46    L2/3 4341_BA46       4341    PFC  13   M
## 2: AAACGGGGTCGGCATC-1_4341_BA46    L2/3 4341_BA46       4341    PFC  13   M
##    diagnosis Capbatch Seqbatch PMI RIN genes  UMIs mitoPercent riboPercent
## 1:   Control      CB6      SB2  16 7.2  3967  8526   0.4691532   0.5160685
## 2:   Control      CB6      SB2  16 7.2  6891 23815   0.3023305   0.4870880
```

```
summary(meta_cell)
```

```
##                            cell             cluster          sample
##  AAACCTGAGCCATCGC-1_5294_BA9 :   1   L2/3       :8626   5387_BA9:1142
##  AAACCTGCAAGCGTAG-1_5419_PFC :   1   AST-FB     :   0   5278_PFC: 759
##  AAACCTGCAAGTAATG-1_5945_PFC :   1   AST-PP     :   0   5939_BA9: 733
##  AAACCTGCACATTTCT-1_5531_BA9 :   1   Endothelial:   0   5958_BA9: 542
##  AAACCTGCACCAGATT-1_5565_BA9 :   1   IN-PV      :   0   5841_BA9: 451
##  AAACCTGCACCCATTC-1_4341_BA46:   1   IN-SST     :   0   5531_BA9: 431
##  (Other)                     :8620   (Other)    :   0   (Other) :4568
##    individual   region          age        sex        diagnosis
##  Min.   :4341   ACC:   0   Min.   : 4.00   F:2138   ASD    :4925
##  1st Qu.:5387   PFC:8626   1st Qu.:12.00   M:6488   Control:3701
##  Median :5565              Median :15.00
##  Mean   :5582              Mean   :15.79
##  3rd Qu.:5939              3rd Qu.:20.00
##  Max.   :6033              Max.   :22.00
##
##     Capbatch    Seqbatch        PMI             RIN            genes
##  CB6    :3775   SB1:3087   Min.   : 3.00   Min.   :6.500   Min.   : 1634
##  CB7    :1764   SB2:5539   1st Qu.:13.00   1st Qu.:7.400   1st Qu.: 3510
##  CB1    :1663   SB3:   0   Median :21.00   Median :7.600   Median : 4680
##  CB2    :1424              Mean   :19.65   Mean   :7.834   Mean   : 4829
##  CB3    :   0              3rd Qu.:24.00   3rd Qu.:8.300   3rd Qu.: 5957
##  CB4    :   0              Max.   :42.00   Max.   :9.000   Max.   :12304
##  (Other):   0
##       UMIs          mitoPercent      riboPercent
##  Min.   :  2608   Min.   :0.0000   Min.   :0.02842
##  1st Qu.:  7854   1st Qu.:0.1886   1st Qu.:0.41331
##  Median : 12263   Median :0.4280   Median :0.58264
##  Mean   : 14371   Mean   :0.7451   Mean   :0.77375
##  3rd Qu.: 18414   3rd Qu.:0.9721   3rd Qu.:0.90364
##  Max.   :114644   Max.   :4.9753   Max.   :4.92958
##
```

```
meta_cell$Capbatch = droplevels(meta_cell$Capbatch)
meta_cell$Seqbatch = droplevels(meta_cell$Seqbatch)
table(meta_cell$Capbatch, meta_cell$Seqbatch)
```

```
##
##        SB1  SB2
##   CB1 1663    0
##   CB2 1424    0
##   CB6    0 3775
##   CB7    0 1764
```

```
summary(meta_cell$UMIs/meta_cell$genes)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.469   2.230   2.622   2.748   3.121   9.318
```

```
sort(table(paste(meta_cell$individual, meta_cell$diagnosis, sep=":")))
```

```
##
##     5978:ASD     5403:ASD 5976:Control 5408:Control 5936:Control     5144:ASD
##           65           69          106          162          193          202
## 5577:Control     5864:ASD 5879:Control 5893:Control     5419:ASD     6033:ASD
##          215          275          278          284          327          362
## 4341:Control 5538:Control     5565:ASD     5294:ASD     5945:ASD     5531:ASD
##          388          391          414          415          422          431
##     5841:ASD 5958:Control     5939:ASD     5278:ASD 5387:Control
##          451          542          733          759         1142
```

generate individual level information

```
length(unique(meta_cell$individual))
```

```
## [1] 23
```

```
meta_ind = distinct(meta_cell[,3:12])
dim(meta_ind)
```

```
## [1] 23 10
```

```
meta_ind[1:2,]
```

```
##       sample individual region age sex diagnosis Capbatch Seqbatch PMI RIN
## 1: 4341_BA46       4341    PFC  13   M   Control      CB6      SB2  16 7.2
## 2:  5144_PFC       5144    PFC   7   M       ASD      CB1      SB1   3 8.0
```

```
meta_ind$diagnosis = relevel(meta_ind$diagnosis, ref="Control")
table(meta_ind$diagnosis)
```

```
##
## Control     ASD
##      10      13
```

```
if(nrow(meta_ind) != length(unique(meta_cell$individual))){
  stop("there is non-unique information\n")
}

table(meta_ind$Seqbatch, meta_ind$Capbatch)
```

```
##
##       CB1 CB2 CB6 CB7
##   SB1   6   4   0   0
##   SB2   0   0   7   6
```

filter out genes with too many zero's

```
n.zeros = rowSums(dat1 == 0)
summary(n.zeros)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0    5334    7141    6380    8037    8619
```

```
0.6*ncol(dat1)
```

```
## [1] 5175.6
```

```
0.8*ncol(dat1)
```

```
## [1] 6900.8
```

```
table(n.zeros < 0.6*ncol(dat1))
```

```
##
## FALSE  TRUE
## 13781  4260
```

```
table(n.zeros < 0.8*ncol(dat1))
```

```
##
## FALSE  TRUE
##  9781  8260
```

```
table(n.zeros < 0.9*ncol(dat1))
```

```
##
## FALSE  TRUE
##  6335 11706
```

```
w2kp = which(n.zeros < 0.6*ncol(dat1))
dat1 = dat1[w2kp,]

dim(dat1)
```

```
## [1] 4260 8626
```

```
dat1[1:5,1:4]
```

```
## 5 x 4 sparse Matrix of class "dgCMatrix"
##        AAACCTGCACCCATTC-1_4341_BA46 AAACGGGGTCGGCATC-1_4341_BA46
## VPS13D                            2                            3
## KIF1B                             2                            5
## PRKCZ                             1                            7
## KCNAB2                            1                            5
## ENO1                              .                            2
##        AAAGATGCAGCGTCCA-1_4341_BA46 AAAGATGGTCCGAATT-1_4341_BA46
## VPS13D                            1                            1
## KIF1B                             6                            2
## PRKCZ                             1                            .
## KCNAB2                            1                            2
## ENO1                              2                            .
```

add read-depth information

```
dim(meta_cell)
```

```
## [1] 8626   16
```

```
meta_cell[1:2,]
```

```
##                           cell cluster    sample individual region age sex
## 1: AAACCTGCACCCATTC-1_4341_BA46    L2/3 4341_BA46       4341    PFC  13   M
## 2: AAACGGGGTCGGCATC-1_4341_BA46    L2/3 4341_BA46       4341    PFC  13   M
##    diagnosis Capbatch Seqbatch PMI RIN genes  UMIs mitoPercent riboPercent
```

```
## 1:   Control      CB6      SB2  16 7.2  3967  8526   0.4691532   0.5160685
## 2:   Control      CB6      SB2  16 7.2  6891 23815   0.3023305   0.4870880
```

```
table(meta_cell$cell == colnames(dat1))
```

```
##
## TRUE
## 8626
```

```
rd_cell = colSums(dat1)
summary(rd_cell)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2053    6332    9864   11606   14878   90156
```

```
meta_cell$rd = rd_cell

med_rd_cell = tapply(rd_cell, meta_cell$individual, median)
med_rd_cell
```

```
##    4341    5144    5278    5294    5387    5403    5408    5419    5531    5538
##  6863.5  8513.0  7743.0 15402.0 12012.5  8611.0 13326.5  6109.0 10884.0 11347.0
##    5565    5577    5841    5864    5879    5893    5936    5939    5945    5958
## 11039.5 14714.0 12059.0 10019.0 12753.0 10342.0 10099.0  7835.0  8773.0 13170.0
##    5976    5978    6033
##  5245.5  6441.0  8812.5
```

```
rd_ind = tapply(rd_cell, meta_cell$individual, sum)
rd_ind
```

```
##     4341     5144     5278     5294     5387     5403     5408     5419
##  3697794  1826162  6535060  6760075 17259725   753499  2230486  2173475
##     5531     5538     5565     5577     5841     5864     5879     5893
##  5208864  4609338  5064137  3410924  6117156  3044208  3716218  2992120
##     5936     5939     5945     5958     5976     5978     6033
##  2172185  6596446  3783705  7487681   635864   477067  3563892
```

```
table(names(med_rd_cell) == meta_ind$individual)
```

```
##
## TRUE
##   23
```

```
meta_ind$med_rd_cell = med_rd_cell
meta_ind$rd = rd_ind
```

Read in gene-gene interaction information.

```
path = "../gene_annotation/"
ppi = fread(paste0(path, "BIOGRID-ORGANISM-Homo_sapiens-3.5.185.tab3.txt.gz"))
dim(ppi)
```

```
## [1] 550660     37
```

```
names(ppi)
```

```
##  [1] "#BioGRID Interaction ID"           "Entrez Gene Interactor A"
##  [3] "Entrez Gene Interactor B"          "BioGRID ID Interactor A"
##  [5] "BioGRID ID Interactor B"           "Systematic Name Interactor A"
##  [7] "Systematic Name Interactor B"      "Official Symbol Interactor A"
```

```
##  [9] "Official Symbol Interactor B"       "Synonyms Interactor A"
## [11] "Synonyms Interactor B"              "Experimental System"
## [13] "Experimental System Type"           "Author"
## [15] "Publication Source"                 "Organism ID Interactor A"
## [17] "Organism Name Interactor A"         "Organism ID Interactor B"
## [19] "Organism Name Interactor B"         "Throughput"
## [21] "Score"                              "Modification"
## [23] "Qualifications"                     "Tags"
## [25] "Source Database"                    "SWISS-PROT Accessions Interactor A"
## [27] "TREMBL Accessions Interactor A"     "REFSEQ Accessions Interactor A"
## [29] "SWISS-PROT Accessions Interactor B" "TREMBL Accessions Interactor B"
## [31] "REFSEQ Accessions Interactor B"     "Ontology Term IDs"
## [33] "Ontology Term Names"                "Ontology Term Categories"
## [35] "Ontology Term Qualifier IDs"        "Ontology Term Qualifier Names"
## [37] "Ontology Term Types"
```

```
ppi[1:2,]
```

```
##    #BioGRID Interaction ID Entrez Gene Interactor A Entrez Gene Interactor B
## 1:                     103                    6416                    2318
## 2:                     117                   84665                      88
##    BioGRID ID Interactor A BioGRID ID Interactor B Systematic Name Interactor A
## 1:                  112315                  108607                            -
## 2:                  124185                  106603                            -
##    Systematic Name Interactor B Official Symbol Interactor A
## 1:                            -                       MAP2K4
## 2:                            -                         MYPN
##    Official Symbol Interactor B
## 1:                         FLNC
## 2:                        ACTN2
##                                        Synonyms Interactor A
## 1: JNKK|JNKK1|MAPKK4|MEK4|MKK4|PRKMK4|SAPKK-1|SAPKK1|SEK1|SERK1|SKK1
## 2:                                    CMD1DD|CMH22|MYOP|RCM4
##                      Synonyms Interactor B Experimental System
## 1: ABP-280|ABP280A|ABPA|ABPL|FLN2|MFM5|MPD4          Two-hybrid
## 2:                                 CMD1AA          Two-hybrid
##    Experimental System Type         Author Publication Source
## 1:                physical Marti A (1997)     PUBMED:9006895
## 2:                physical Bang ML (2001)    PUBMED:11309420
##    Organism ID Interactor A Organism Name Interactor A Organism ID Interactor B
## 1:                     9606              Homo sapiens                     9606
## 2:                     9606              Homo sapiens                     9606
##    Organism Name Interactor B     Throughput Score Modification Qualifications
## 1:              Homo sapiens Low Throughput     -            -              -
## 2:              Homo sapiens Low Throughput     -            -              -
##    Tags Source Database SWISS-PROT Accessions Interactor A
## 1:    -         BIOGRID                            P45985
## 2:    -         BIOGRID                            Q86TC9
##    TREMBL Accessions Interactor A      REFSEQ Accessions Interactor A
## 1:                              -              NP_003001|NP_001268364
## 2:                     A0A087WX60 NP_001243197|NP_001243196|NP_115967
##    SWISS-PROT Accessions Interactor B TREMBL Accessions Interactor B
## 1:                             Q14315                         Q59H94
## 2:                             P35609                  Q59FD9|F6THM6
##         REFSEQ Accessions Interactor B Ontology Term IDs Ontology Term Names
```

```
## 1:              NP_001120959|NP_001449                 -                   -
## 2: NP_001094|NP_001265272|NP_001265273                 -                   -
##    Ontology Term Categories Ontology Term Qualifier IDs
## 1:                        -                           -
## 2:                        -                           -
##    Ontology Term Qualifier Names Ontology Term Types
## 1:                             -                   -
## 2:                             -                   -
```

```
table(ppi$`Experimental System`, ppi$`Experimental System Type`)
```

```
##
##                                genetic physical
##   Affinity Capture-Luminescence       0     2341
##   Affinity Capture-MS                 0   266593
##   Affinity Capture-RNA                0    18451
##   Affinity Capture-Western            0    67745
##   Biochemical Activity                0    12433
##   Co-crystal Structure                0     1814
##   Co-fractionation                    0    45651
##   Co-localization                     0     3556
##   Co-purification                     0     1757
##   Dosage Growth Defect               15        0
##   Dosage Lethality                  114        0
##   Dosage Rescue                      81        0
##   Far Western                         0      838
##   FRET                                0     2015
##   Negative Genetic                 3449        0
##   PCA                                 0      999
##   Phenotypic Enhancement            216        0
##   Phenotypic Suppression            224        0
##   Positive Genetic                 2332        0
##   Protein-peptide                     0     2110
##   Protein-RNA                         0      552
##   Proximity Label-MS                  0    23222
##   Reconstituted Complex               0    34934
##   Synthetic Growth Defect           492        0
##   Synthetic Lethality              2044        0
##   Synthetic Rescue                  154        0
##   Two-hybrid                          0    56528
```

```
genes = unique(ppi$`Official Symbol Interactor A`)
length(genes)
```

```
## [1] 16086
```

```
genes = union(genes, unique(ppi$`Official Symbol Interactor B`))
length(genes)
```

```
## [1] 24074
```

Pick one individual with large number of cells and calculate correlations.

```
meta_cell$individual = as.character(meta_cell$individual)

dim(dat1)
```

```
## [1] 4260 8626
```

```
dat1[1:5,1:4]
```

```
## 5 x 4 sparse Matrix of class "dgCMatrix"
##        AAACCTGCACCCATTC-1_4341_BA46 AAACGGGGTCGGCATC-1_4341_BA46
## VPS13D                            2                            3
## KIF1B                             2                            5
## PRKCZ                             1                            7
## KCNAB2                            1                            5
## ENO1                              .                            2
##        AAAGATGCAGCGTCCA-1_4341_BA46 AAAGATGGTCCGAATT-1_4341_BA46
## VPS13D                            1                            1
## KIF1B                             6                            2
## PRKCZ                             1                            .
## KCNAB2                            1                            2
## ENO1                              2                            .
```

```
dim(meta_cell)
```

```
## [1] 8626   17
```

```
meta_cell[1:2,]
```

```
##                            cell cluster    sample individual region age sex
## 1: AAACCTGCACCCATTC-1_4341_BA46    L2/3 4341_BA46       4341    PFC  13   M
## 2: AAACGGGGTCGGCATC-1_4341_BA46    L2/3 4341_BA46       4341    PFC  13   M
##    diagnosis Capbatch Seqbatch PMI RIN genes  UMIs mitoPercent riboPercent
## 1:   Control      CB6      SB2  16 7.2  3967  8526   0.4691532   0.5160685
## 2:   Control      CB6      SB2  16 7.2  6891 23815   0.3023305   0.4870880
##       rd
## 1:  6707
## 2: 19279
```

```
w2kp = which(meta_cell$individual == "5278")

count_matrix = dat1[,w2kp]
rd           = meta_cell$rd[w2kp]

dim(count_matrix)
```

```
## [1] 4260  759
```

```
count_matrix[1:2,1:5]
```

```
## 2 x 5 sparse Matrix of class "dgCMatrix"
##        AAACCTGGTAGCGTGA-1_5278_PFC AAACGGGAGGACTGGT-1_5278_PFC
## VPS13D                           1                           7
## KIF1B                            4                           2
##        AAAGTAGCAGACAAGC-1_5278_PFC AACACGTAGAGCTGGT-1_5278_PFC
## VPS13D                           .                           1
## KIF1B                            .                           4
##        AACACGTTCTCGCATC-1_5278_PFC
## VPS13D                           .
## KIF1B                            3
```

```
summary(apply(count_matrix, 2, median))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##  0.0000  0.0000  1.0000  0.7286  1.0000  3.0000
```

```
summary(apply(count_matrix, 1, median))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    1.00    1.54    1.00 1042.00
```

```
table(rownames(count_matrix) %in% genes)
```

```
##
## FALSE  TRUE
##   646  3614
```

```
count_matrix = count_matrix[which(rownames(count_matrix) %in% genes),]
dim(count_matrix)
```

```
## [1] 3614  759
```

```
X = as.matrix(log(t(count_matrix + 0.5)/rd))
dim(X)
```

```
## [1]  759 3614
```

```
cr = cor(X)

ww1 = ppi$`Official Symbol Interactor A` %in% rownames(count_matrix)
ww2 = ppi$`Official Symbol Interactor B` %in% rownames(count_matrix)
table(ww1)
```

```
## ww1
##  FALSE   TRUE
## 402537 148123
```

```
table(ww2)
```

```
## ww2
##  FALSE   TRUE
## 370158 180502
```

```
col_nms = c("Official Symbol Interactor A", "Official Symbol Interactor B",
            "Experimental System", "Experimental System Type")
ppi.X = ppi[which(ww1 & ww2), ..col_nms]
names(ppi.X) = c("geneA", "geneB", "System", "Type")
dim(ppi.X)
```

```
## [1] 52837     4
```

```
ppi.X[1:2,]
```

```
##    geneA  geneB     System     Type
## 1:  XRN1  ALDOA Two-hybrid physical
## 2:   APP APPBP2 Two-hybrid physical
```

```
table(ppi.X$geneA == ppi.X$geneB)
```

```
##
## FALSE  TRUE
## 51096  1741
```

```
ppi.X = ppi.X[which(ppi.X$geneA != ppi.X$geneB),]
t1 = table(ppi.X$System)
```

```
sort(t1)
```

```
##
##                 Dosage Rescue       Synthetic Growth Defect
##                             3                           17
##        Phenotypic Suppression        Phenotypic Enhancement
##                            18                           21
##                           PCA                  Protein-RNA
##                            34                           55
##           Synthetic Lethality                  Far Western
##                            63                          118
##               Co-purification                         FRET
##                           139                          146
##               Protein-peptide          Co-crystal Structure
##                           193                          201
## Affinity Capture-Luminescence             Positive Genetic
##                           202                          235
##               Co-localization             Negative Genetic
##                           332                          391
##          Affinity Capture-RNA          Biochemical Activity
##                           652                         1620
##            Proximity Label-MS                   Two-hybrid
##                          2396                         4521
##         Reconstituted Complex             Co-fractionation
##                          4858                         6059
##      Affinity Capture-Western           Affinity Capture-MS
##                          8156                        20666
```

```
eSystem2kp = names(t1)[t1 > 50]
ppi.X = ppi.X[which(ppi.X$System %in% eSystem2kp),]
dim(ppi.X)
```

```
## [1] 51003     4
```

```
ppi.X[1:2,]
```

```
##    geneA  geneB     System     Type
## 1:  XRN1  ALDOA Two-hybrid physical
## 2:   APP APPBP2 Two-hybrid physical
```

```
table(ppi.X$System, ppi.X$Type)
```

```
##
##                                 genetic physical
##   Affinity Capture-Luminescence       0      202
##   Affinity Capture-MS                 0    20666
##   Affinity Capture-RNA                0      652
##   Affinity Capture-Western            0     8156
##   Biochemical Activity                0     1620
##   Co-crystal Structure                0      201
##   Co-fractionation                    0     6059
##   Co-localization                     0      332
##   Co-purification                     0      139
##   Far Western                         0      118
##   FRET                                0      146
##   Negative Genetic                  391        0
```

```
##   Positive Genetic                  235        0
##   Protein-peptide                     0      193
##   Protein-RNA                         0       55
##   Proximity Label-MS                  0     2396
##   Reconstituted Complex               0     4858
##   Synthetic Lethality                63        0
##   Two-hybrid                          0     4521
```

```
ppi.X$cr = rep(NA, nrow(ppi.X))

for(s1 in unique(ppi.X$System)){
  w2 = which(ppi.X$System == s1)
  wA = match(ppi.X$geneA[w2], rownames(count_matrix))
  wB = match(ppi.X$geneB[w2], rownames(count_matrix))
  ppi.X$cr[w2] = diag(cr[wA,wB])
}
dim(ppi.X)
```

```
## [1] 51003     5
```

```
ppi.X[1:2,]
```

```
##    geneA  geneB     System     Type          cr
## 1:  XRN1  ALDOA Two-hybrid physical  0.14077599
## 2:   APP APPBP2 Two-hybrid physical -0.03104515
```

summarize the correlations.

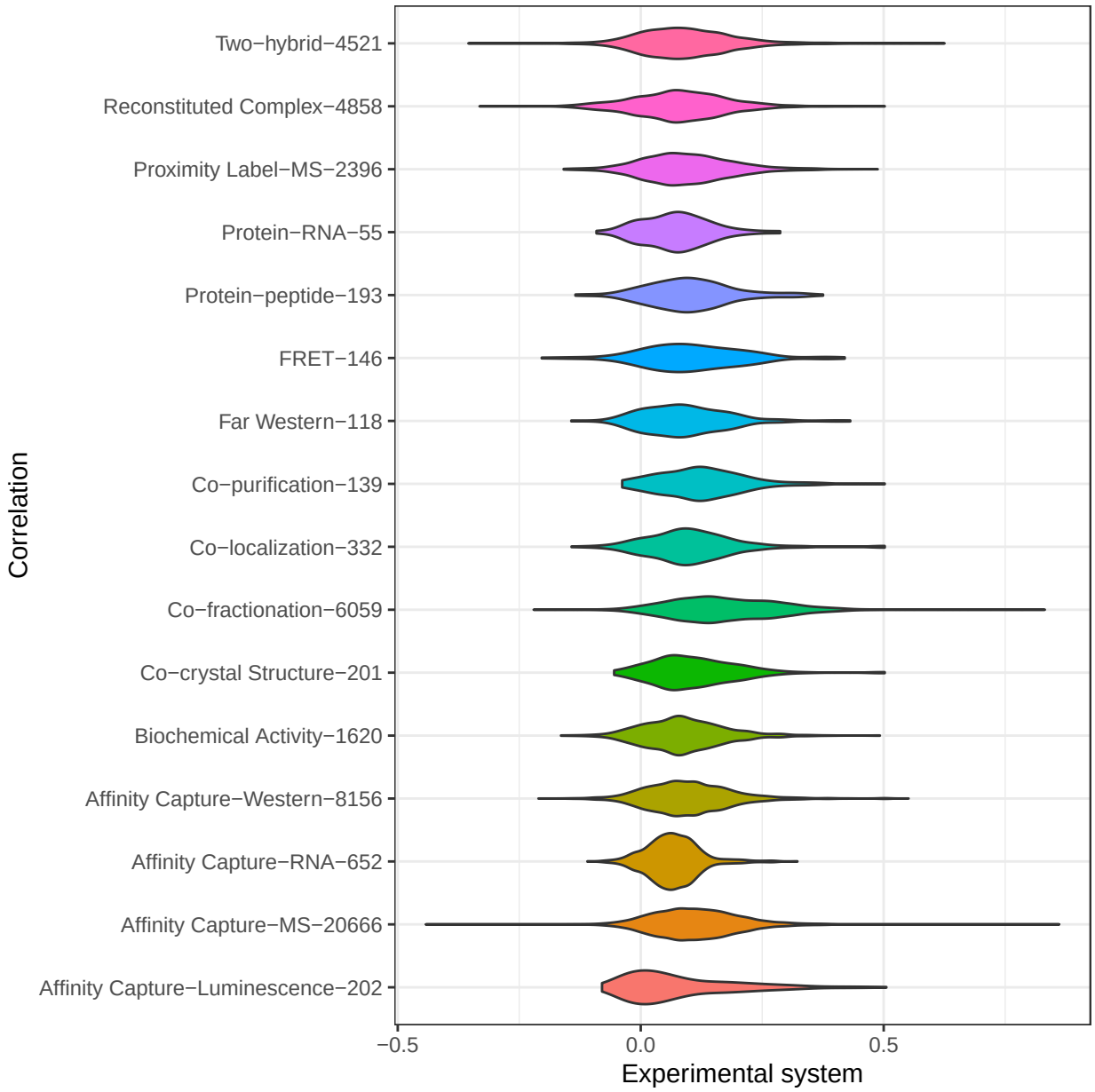
```
tapply(ppi.X$cr, ppi.X$System, summary)
```

```
## $`Affinity Capture-Luminescence`
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -0.079591 -0.005304  0.042381  0.077433  0.144666  0.505487
##
## $`Affinity Capture-MS`
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.44180  0.05071  0.10804  0.11317  0.16798  0.86067
##
## $`Affinity Capture-RNA`
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.10980  0.03220  0.06384  0.06627  0.09787  0.32246
##
## $`Affinity Capture-Western`
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.21056  0.03842  0.09147  0.09813  0.15039  0.55120
##
## $`Biochemical Activity`
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.16413  0.03002  0.08219  0.08826  0.13548  0.49205
##
## $`Co-crystal Structure`
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.05468  0.05139  0.09553  0.10830  0.15909  0.50212
##
## $`Co-fractionation`
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.22033  0.09302  0.16071  0.17247  0.24801  0.83170
```

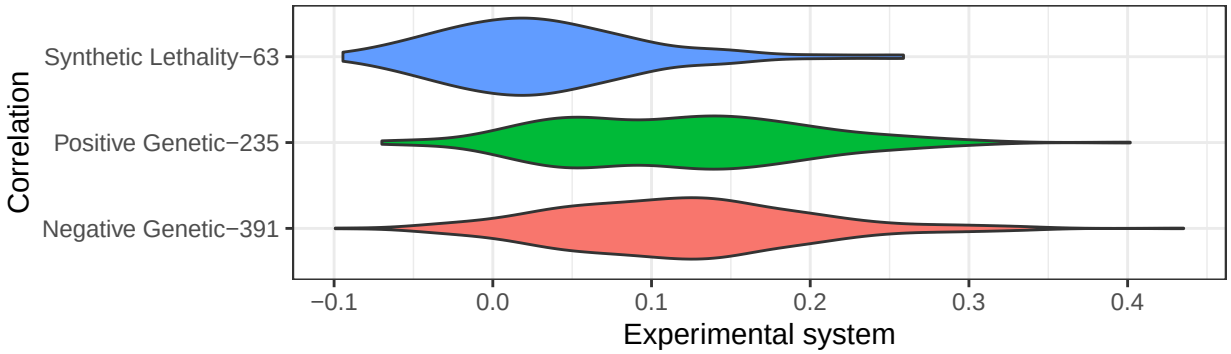
```
##
## $`Co-localization`
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.14240  0.04417  0.09786  0.10288  0.15079  0.50212
##
## $`Co-purification`
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.03810  0.05873  0.11774  0.12458  0.17360  0.50212
##
## $`Far Western`
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.14289  0.02403  0.08109  0.09055  0.14815  0.43123
##
## $FRET
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.20375  0.03615  0.09923  0.10580  0.17114  0.41982
##
## $`Negative Genetic`
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.09930  0.06439  0.11532  0.11811  0.16519  0.43533
##
## $`Positive Genetic`
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.06991  0.05232  0.11611  0.11652  0.16757  0.40167
##
## $`Protein-peptide`
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.13439  0.04292  0.09891  0.10167  0.15041  0.37524
##
## $`Protein-RNA`
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.09100  0.02106  0.07304  0.07277  0.11192  0.28698
##
## $`Proximity Label-MS`
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.15857  0.03201  0.08868  0.09459  0.14975  0.48769
##
## $`Reconstituted Complex`
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.33172  0.01802  0.08005  0.08057  0.14162  0.50212
##
## $`Synthetic Lethality`
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.09443 -0.01389  0.02003  0.03157  0.06460  0.25878
##
## $`Two-hybrid`
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.35468  0.02598  0.08431  0.09184  0.14897  0.62521
```

```
t1 = table(ppi.X$System)
t1 = data.frame(System=names(t1), freq=as.numeric(t1))
ppi.X$System_count = paste(ppi.X$System,
                           t1$freq[match(ppi.X$System, t1$System)], sep="-")
p1 = ggplot(subset(ppi.X, Type %in% c("physical")),
```

```
            aes(x=System_count, y=cr, fill=System_count)) +
  geom_violin()  + coord_flip() + xlab("Correlation") +
  ylab("Experimental system") +
  theme(legend.position="none")
p1
```




```
p2 = ggplot(subset(ppi.X, Type %in% c("genetic")),
            aes(x=System_count, y=cr, fill=System_count)) +
  geom_violin()  + coord_flip() + xlab("Correlation") +
  ylab("Experimental system") +
  theme(legend.position="none")
p2
```

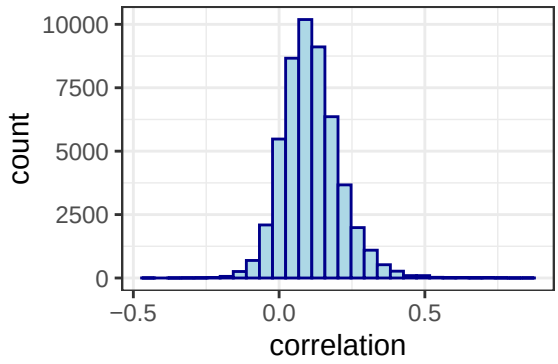



```
s2rm = c("Synthetic Lethality", "Affinity Capture-Luminescence")
ppi.X = ppi.X[which(! ppi.X$System %in% s2rm),]
table(ppi.X$System)
```

```
##
##      Affinity Capture-MS     Affinity Capture-RNA Affinity Capture-Western
##                    20666                      652                     8156
##     Biochemical Activity     Co-crystal Structure         Co-fractionation
##                     1620                      201                     6059
##          Co-localization          Co-purification              Far Western
##                      332                      139                      118
##                     FRET         Negative Genetic         Positive Genetic
##                      146                      391                      235
##          Protein-peptide              Protein-RNA       Proximity Label-MS
##                      193                       55                     2396
##    Reconstituted Complex              Two-hybrid
##                     4858                     4521
```

```
ggplot(ppi.X, aes(x=cr)) +
  geom_histogram(color="darkblue", fill="lightblue") +
  xlab("correlation")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
gc()
```

```
##            used  (Mb) gc trigger   (Mb) limit (Mb)  max used   (Mb)
## Ncells  2407312 128.6    3546515  189.5         NA   3546515  189.5
## Vcells 79780869 608.7  390212708 2977.1      32768 597689445 4560.1
```

```
sessionInfo()
```

```
## R version 3.6.2 (2019-12-12)
```

```
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS Catalina 10.15.6
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] parallel  stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
##  [1] ggpointdensity_0.1.0 ggpubr_0.2.5         magrittr_1.5
##  [4] ggplot2_3.3.1        svd_0.5              doParallel_1.0.15
##  [7] iterators_1.0.12     foreach_1.4.7        dplyr_0.8.4
## [10] data.table_1.12.8    Matrix_1.2-18        MASS_7.3-51.5
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.3        pillar_1.4.3      compiler_3.6.2    R.methodsS3_1.8.0
##  [5] R.utils_2.9.2     tools_3.6.2       digest_0.6.23     evaluate_0.14
##  [9] lifecycle_0.2.0   tibble_3.0.1      gtable_0.3.0      lattice_0.20-38
## [13] pkgconfig_2.0.3   rlang_0.4.6       yaml_2.2.1        xfun_0.12
## [17] withr_2.1.2       stringr_1.4.0     knitr_1.28        vctrs_0.3.0
## [21] grid_3.6.2        tidyselect_1.0.0  glue_1.3.1        R6_2.4.1
## [25] rmarkdown_2.1     farver_2.0.3      purrr_0.3.3       scales_1.1.0
## [29] codetools_0.2-16  ellipsis_0.3.0    htmltools_0.4.0   assertthat_0.2.1
## [33] colorspace_1.4-1  ggsignif_0.6.0    labeling_0.3      stringi_1.4.5
## [37] munsell_0.5.0     crayon_1.3.4      R.oo_1.23.0
```