annotate_ALL_cryptic_introns

Figure 2B three_database_info_all_junctions.tsv

Requires simplified annotation of introns and exons as produced by leafcutter/leafviz/gtf2leafcutter.pl from gtf files for each annotation.

```
library(tidyr)
library(ggplot2)
library(dplyr)
library(ggrepel)
here::i_am("R/12_annotate_ALL_cryptic_introns.Rmd")
library(here)
knitr::opts_chunk$set(dev = "pdf",
                      dpi = 300,
                      echo = FALSE,
                      fig.path="plots/annotate_nonsig_cryptic_plots/")
```

```
## [1] 132587     19
```

Add annotation information from ensembl

```
## [1] 1145034      10
```

```
## `summarise()` has grouped output by 'chr', 'start', 'end', 'strand',
## 'cluster_id', 'deltapsi'. You can override using the `.groups` argument.
```

```
## # A tibble: 6 x 13
## # Groups:   chr, start, end, strand, cluster_id, deltapsi [6]
##   chr      start      end strand cluster_id   deltapsi  p.adjust transcript_ids
##   <chr>    <int>    <int> <chr>  <chr>           <dbl>     <dbl> <chr>
## 1 chr7  43648652 43650493 -      clu_35616_- -0.0135   3.89e-114 ENST0000031056~
## 2 chr7  43648652 43650612 -      clu_35616_- -0.00546  3.89e-114 ENST0000044656~
## 3 chr7  43648652 43665658 -      clu_35616_-  0.0139   3.89e-114 ENST0000043844~
## 4 chr7  43648652 43711400 -      clu_35616_-  0.000345 3.89e-114 ENST0000041579~
## 5 chr7  43648652 43729429 -      clu_35616_- -0.00985  3.89e-114 ENST0000041507~
## 6 chr7  43650712 43656033 -      clu_35616_- -0.00196  3.89e-114 ENST0000043165~
## # i 5 more variables: min_intron_number <int>, mode_intron_number <dbl>,
## #   gene <chr>, biotype <chr>, genes_in_cluster <chr>
```

```
## [1] 89495
```

```
## [1] 43092
```

Load refseq information

```
##     chr    start      end gene_name   gene_id strand   transcript_id
## 1 chr19 58353327 58353404      A1BG gene-A1BG      - rna-NM_130786.4
## 2 chr19 58353197 58353292      A1BG gene-A1BG      - rna-NM_130786.4
## 3 chr19 58352555 58352928      A1BG gene-A1BG      - rna-NM_130786.4
## 4 chr19 58351687 58352283      A1BG gene-A1BG      - rna-NM_130786.4
## 5 chr19 58350651 58351391      A1BG gene-A1BG      - rna-NM_130786.4
## 6 chr19 58347640 58350370      A1BG gene-A1BG      - rna-NM_130786.4
##   intron_number biotype annotation
## 1             1 Unknown         NA
## 2             2 Unknown         NA
## 3             3 Unknown         NA
## 4             4 Unknown         NA
## 5             5 Unknown         NA
## 6             6 Unknown         NA
```

```
##
##                   chr1  chr1_KI270706v1_random  chr1_KI270708v1_random
##                 188151                      11                       3
##  chr1_KI270711v1_random  chr1_KI270712v1_random  chr1_KI270713v1_random
##                     25                      14                      11
##  chr1_KI270714v1_random                   chr10                   chr11
##                     16                   90080                  100795
##                  chr12                   chr13                   chr14
##                 109617                   39055                   59969
## chr14_GL000009v2_random chr14_GL000194v1_random chr14_KI270722v1_random
##                     24                      23                       1
## chr14_KI270723v1_random chr14_KI270724v1_random chr14_KI270725v1_random
##                      1                       2                       2
## chr14_KI270726v1_random                   chr15 chr15_KI270727v1_random
##                      7                   70982                      55
##                  chr16 chr16_KI270728v1_random                   chr17
##                 72812                     176                  105762
## chr17_GL000205v2_random                   chr18                   chr19
##                      2                   39671                   79265
##                   chr2                   chr20                   chr21
##                 171424                   38128                   18813
##                  chr22 chr22_KI270731v1_random chr22_KI270733v1_random
##                  36821                     178                      16
##                   chr3  chr3_GL000221v1_random                    chr4
##                 134762                      83                   84238
##  chr4_GL000008v2_random                    chr5                    chr6
##                      5                   81849                   93966
##                   chr7                    chr8                    chr9
##                  84317                   74623                   87637
##  chr9_KI270718v1_random  chr9_KI270719v1_random  chr9_KI270720v1_random
##                      2                      18                      20
##        chrUn_GL000195v1        chrUn_GL000213v1        chrUn_GL000214v1
##                     60                      43                       8
##        chrUn_GL000218v1        chrUn_GL000219v1        chrUn_GL000220v1
##                     18                      11                       7
##        chrUn_GL000224v1        chrUn_KI270442v1        chrUn_KI270741v1
##                     14                       7                      11
```

```
##        chrUn_KI270742v1        chrUn_KI270743v1        chrUn_KI270744v1
##                      19                      13                      41
##        chrUn_KI270745v1        chrUn_KI270746v1        chrUn_KI270748v1
##                      14                       7                      21
##        chrUn_KI270750v1        chrUn_KI270751v1        chrUn_KI270754v1
##                       5                       8                       8
##        chrUn_KI270755v1                    chrX                    chrY
##                       1                   55702                    7149
```

```
## `summarise()` has grouped output by 'chr', 'start', 'end', 'strand',
## 'deltapsi', 'p.adjust'. You can override using the `.groups` argument.
```

```
## [1] 11632
```

```
## # A tibble: 6 x 12
## # Groups:   chr, start, end, strand, deltapsi, p.adjust [6]
##   chr    start    end strand deltapsi     p.adjust cluster_id  transcript_ids
##   <chr>  <int>  <int> <chr>     <dbl>        <dbl> <chr>       <chr>
## 1 chr1   14829  14970 -      0.0146   0.238        clu_27295_- rna-NR_024540.1
## 2 chr1   24891  29321 -      0.000665 0.227        clu_27299_- rna-NR_024540.1
## 3 chr1  120932 165884 -      0.0302   0.0000000355 clu_27300_- rna-XR_001737579~
## 4 chr1  187287 187380 -      0.00634  0.775        clu_27297_- rna-NR_186787.1
## 5 chr1  195416 199837 -      0.00464  0.227        clu_27299_- rna-NR_186787.1
## 6 chr1  729804 729898 -      0.00343  0.242        clu_27306_- rna-NR_168328.1
## # i 4 more variables: min_intron_number <int>, mode_intron_number <dbl>,
## #   gene <chr>, genes_in_cluster <chr>
```

```
##
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16
## 4582 1778  870  598  505  448  354  296  308  252  173  186  160  139  118  122
##   17   18   19   20   21   22   23   24   25   26   27   28   29   30   31   32
##   95   68   66   50   52   64   44   31   32   31   24   19   15   11   16   17
##   33   34   35   36   37   38   39   40   41   42   43   44   46   47   48   49
##   17    6    4    6    8    4    4    3    5    3    1    5    6    2    8    2
##   50   51   52   55   56   58   59   61   62   64   65   66   67   80   82  106
##    3    2    3    1    1    4    1    1    1    1    1    1    1    1    1    1
```

```
##
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16
## 4485 1724  867  598  529  433  371  306  314  259  192  187  178  139  129  138
##   17   18   19   20   21   22   23   24   25   26   27   28   29   30   31   32
##   94   72   82   48   50   65   47   35   33   39   24   16   16   11   17   17
##   33   34   35   36   37   38   39   40   41   42   44   45   46   47   48   49
##   18    7    4    5   10    7    5    3    6    3    5    1    6    3    4    4
##   50   51   52   55   56   58   59   62   64   65   66   67   82   98  106  109
##    4    1    4    1    1    4    1    1    2    1    1    1    1    1    1    1
```

```
##         chr    start      end strand  cluster_id    deltapsi     p.adjust
## 11354 chr17 17577027 17577107      - clu_19605_-  0.16661846 6.921348e-70
## 11355 chr17 17577027 17591967      - clu_19605_- -0.01394206 6.921348e-70
##       genes_in_cluster gene_name   gene_id      transcript_id intron_number
## 11354             PEMT      PEMT gene-PEMT rna-XM_006721418.5             1
```

```
## 11355             PEMT      PEMT gene-PEMT rna-XM_024450532.2            1
##       biotype annotation
## 11354 Unknown         NA
## 11355 Unknown         NA
```

```
## [1] 31584
```

```
##     chr     start       end strand  cluster_id   deltapsi     p.adjust
## 1 chr11 111844723 111845565      -  clu_2011_-  0.1255768 1.679369e-62
## 2  chr2 238909563 238909842      + clu_32408_+  0.2121182 1.562233e-34
## 3 chr11 111844723 111846253      -  clu_2011_-  0.1117538 1.679369e-62
## 4 chr15  62570852  62589687      + clu_30508_+  0.1791942 1.952222e-34
## 5 chr11 111923754 111924326      + clu_39178_+  0.3102886 2.090005e-18
## 6  chr1 209724144 209742497      - clu_29110_- -0.1053003 2.688258e-50
##          genes_in_cluster
## 1         ALG9,AP001781.2
## 2                  TWIST2
## 3         ALG9,AP001781.2
## 4                    TLN2
## 5 HSPB2-C11orf52,C11orf52
## 6             HSD11B1-AS1
```

Check FANTOM CAT database

```
##     chr    start      end         gene_name           gene_id strand
## 1 chr10 69537590 69537931 CATG00000000020.1 CATG00000000020.1      -
## 2 chr10 69536724 69537457 CATG00000000020.1 CATG00000000020.1      -
## 3 chr10 69800136 69801046 CATG00000000025.1 CATG00000000025.1      -
## 4 chr10 69800136 69801046 CATG00000000025.1 CATG00000000025.1      -
## 5 chr10 69769119 69798157 CATG00000000025.1 CATG00000000025.1      -
## 6 chr10 69800136 69801046 CATG00000000025.1 CATG00000000025.1      -
##       transcript_id intron_number biotype annotation
## 1 MICT00000043247.1             1 Unknown         NA
## 2 MICT00000043247.1             2 Unknown         NA
## 3 ENCT00000057045.1             1 Unknown         NA
## 4 MICT00000043307.1             1 Unknown         NA
## 5 MICT00000043307.1             2 Unknown         NA
## 6 MICT00000043308.1             1 Unknown         NA
```

```
##
##                    chr1                   chr10                   chr11
##                  288162                  123666                  164753
##                   chr12                   chr13                   chr14
##                  173866                   60791                  100581
## chr14_GL000009v2_random                   chr15    chr15_KI270850v1_alt
##                      52                  111812                      49
##                   chr16                   chr17    chr17_KI270909v1_alt
##                  116594                  174394                      13
##                   chr18                   chr19    chr19_KI270938v1_alt
##                   52930                  143379                      10
##                    chr2                   chr20                   chr21
##                  247697                   71316                   33449
```

```
##                   chr22    chr22_KI270879v1_alt                   chr3
##                   61191                      54                 185623
##                    chr4  chr4_GL000008v2_random                   chr5
##                  127234                      24                 137231
##                    chr6                    chr7     chr7_KI270803v1_alt
##                  155684                  137108                      44
##                    chr8                    chr9                    chrM
##                  106714                  123941                      20
##        chrUn_KI270742v1                    chrX                    chrY
##                      28                   90050                    2501
```

```
## `summarise()` has grouped output by 'chr', 'start', 'end', 'strand',
## 'deltapsi', 'p.adjust'. You can override using the `.groups` argument.
```

```
## [1] 9713
```

```
## # A tibble: 6 x 12
## # Groups:   chr, start, end, strand, deltapsi, p.adjust [6]
##   chr    start    end strand deltapsi   p.adjust cluster_id  transcript_ids
##   <chr>  <int>  <int> <chr>     <dbl>      <dbl> <chr>       <chr>
## 1 chr1  743350 746695 -      -0.00140 0.278      clu_27307_- MICT00000000067.1
## 2 chr1  749381 753663 -       0.00998 0.978      clu_27305_- MICT00000000067.1
## 3 chr1  774280 778559 -       0.0364  0.00000288 clu_27309_- HBMT00000049489.1,~
## 4 chr1  774280 805799 -      -0.00621 0.00000288 clu_27309_- MICT00000000069.1
## 5 chr1  801160 805799 -       0.00195 0.00000288 clu_27309_- ENCT00000020342.1
## 6 chr1  805891 810067 -       0.00368 0.00000288 clu_27309_- FTMT20100027364.1,~
## # i 4 more variables: min_intron_number <int>, mode_intron_number <dbl>,
## #   gene <chr>, genes_in_cluster <chr>
```

```
##
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16
## 3820 1620  968  669  521  417  300  257  201  163  120   97   82   67   75   50
##   17   18   19   20   21   22   23   24   25   26   27   28   29   30   31   32
##   42   37   22   34   17   22   13   12    9   10    8    7    5    4    5    3
##   33   34   35   36   37   38   39   40   41   42   43   47   48   50   56   57
##    5    2    1    2    1    4    2    3    1    4    1    2    1    2    2    2
##   60
##    1
```

```
##
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16
## 3605 1542  996  685  538  436  328  260  229  178  130  114   93   80   89   60
##   17   18   19   20   21   22   23   24   25   26   27   28   29   30   31   32
##   48   42   25   33   26   33   16   13   11   17   10    9    6    4    5    5
##   33   34   35   36   37   38   39   40   41   42   45   47   48   49   50   56
##    5    2    2    3    2    4    4    5    1    6    1    2    1    1    3    2
##   57   60
##    2    1
```

So perhaps we can put this info into our bar graph

```
## [1] 21871
```

```
## < table of extent 0 >
```

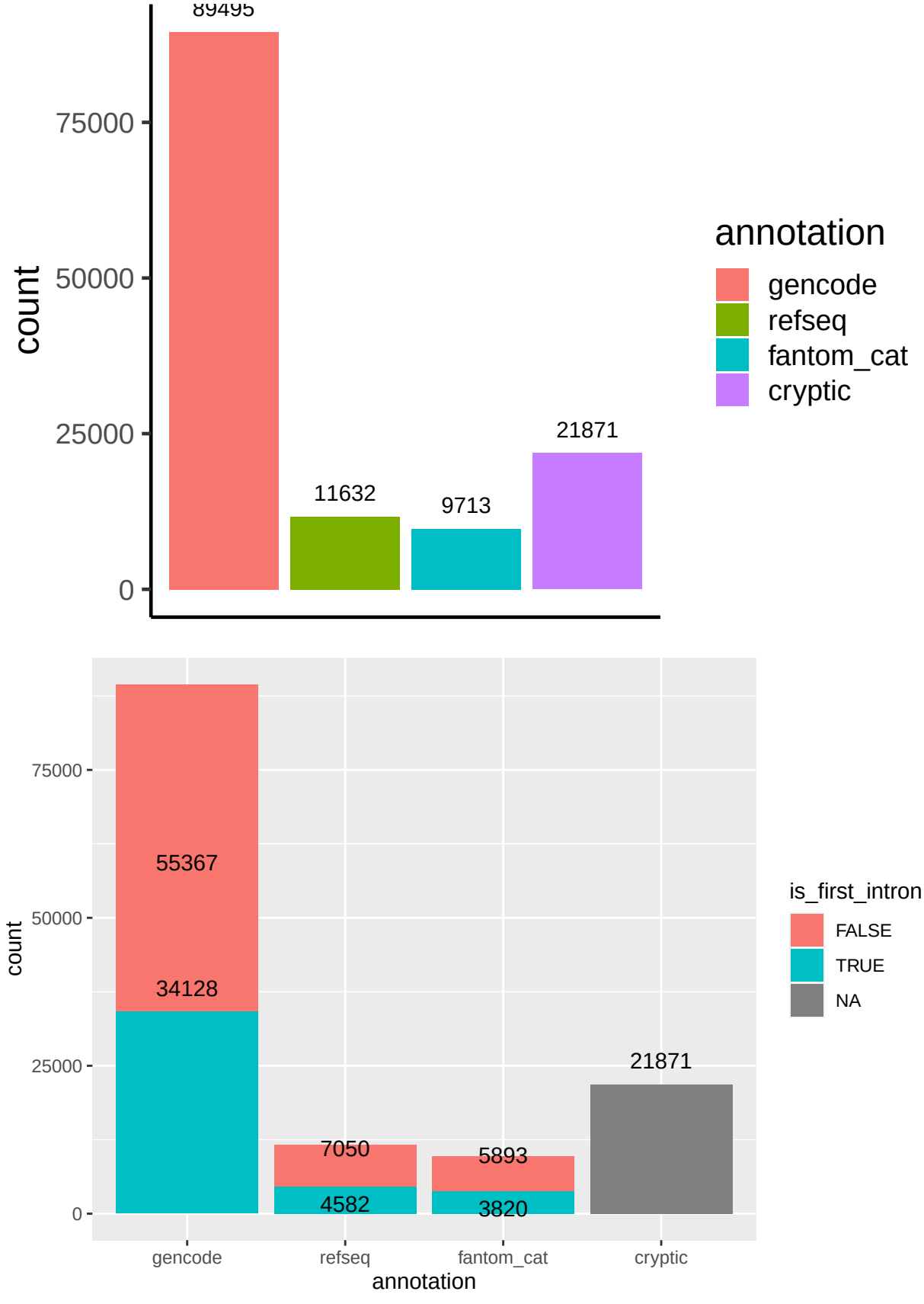
21831 introns cannot be annotated to a osition in a transcript, so are exlcuded from TSS analysis.

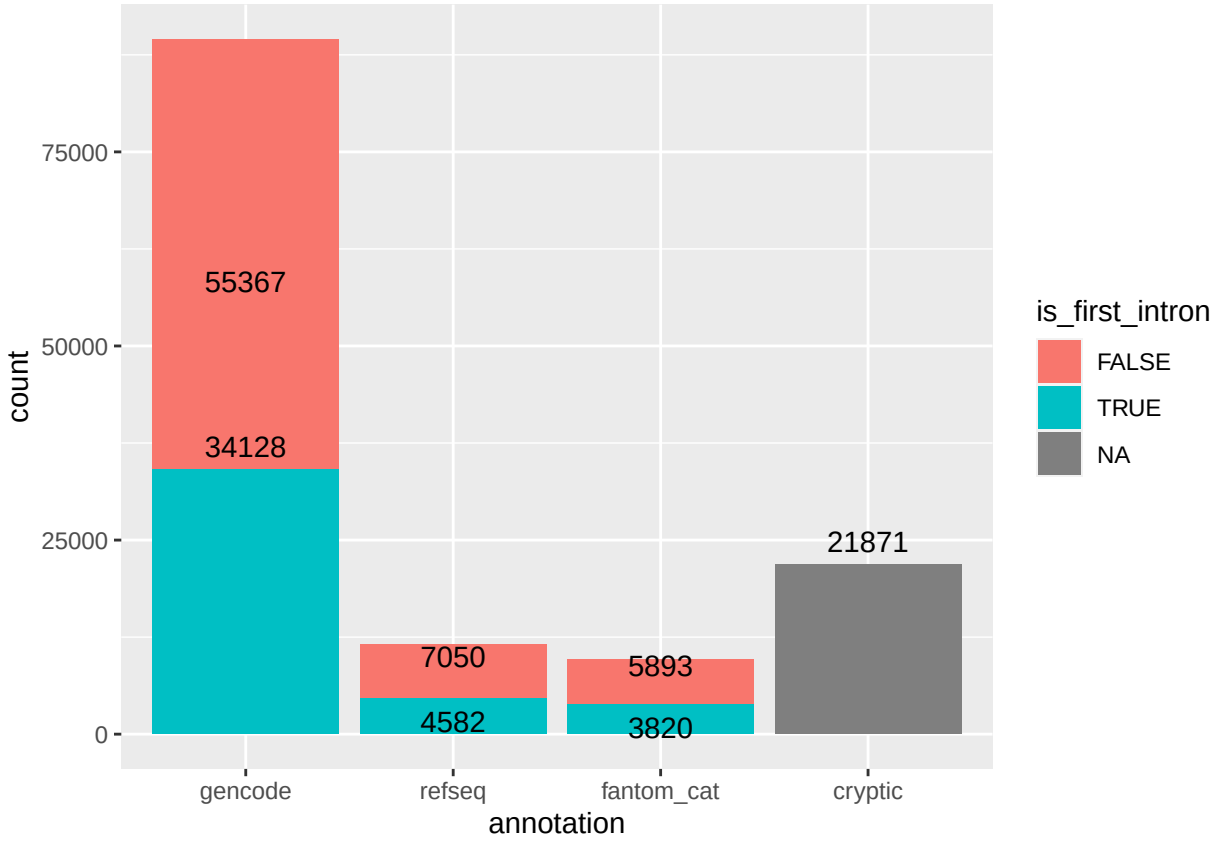
```
## [1] "chr"              "start"            "end"              "strand"
## [5] "cluster_id"       "deltapsi"         "p.adjust"         "genes_in_cluster"
```

```
## # A tibble: 6 x 14
## # Groups:   chr, start, end, strand, cluster_id, deltapsi [6]
##   annotation chr      start      end strand cluster_id   deltapsi  p.adjust
##   <chr>      <chr>    <int>    <int> <chr>  <chr>           <dbl>     <dbl>
## 1 gencode    chr7  43648652 43650493 -      clu_35616_- -0.0135   3.89e-114
## 2 gencode    chr7  43648652 43650612 -      clu_35616_- -0.00546  3.89e-114
## 3 gencode    chr7  43648652 43665658 -      clu_35616_-  0.0139   3.89e-114
## 4 gencode    chr7  43648652 43711400 -      clu_35616_-  0.000345 3.89e-114
## 5 gencode    chr7  43648652 43729429 -      clu_35616_- -0.00985  3.89e-114
## 6 gencode    chr7  43650712 43656033 -      clu_35616_- -0.00196  3.89e-114
## # i 6 more variables: transcript_ids <chr>, min_intron_number <int>,
## #   mode_intron_number <dbl>, gene <chr>, biotype <chr>, genes_in_cluster <chr>
```

```
## [1] 132711
```

Plot Annotation status





Check for cross-gene clusters

```
##  [1] "COA1"                    "PC"
##  [3] "CA5BP1,CA5B"             "PPARG"
##  [5] "MME"                     "CITED1,AL133500.1,HDAC8"
##  [7] "BANK1"                   "XPNPEP1"
##  [9] "BLOC1S1,AC009779.3,RDH5" "PEMT"
## [11] "C19orf12"                "CD44"
## [13] "RTN4"                    "PTK2B"
## [15] "FAR2"                    "LYRM4"
## [17] "ALG9,AP001781.2"         "PEX19,AL139011.2"
## [19] "NCALD"                   "LPIN1"
```

```
##    Mode   FALSE    TRUE
## logical  122320   10391
```

```
##    Mode   FALSE    TRUE
## logical  132611     100
```

Save

```
## # A tibble: 6 x 15
## # Groups:   chr, start, end, strand, cluster_id, deltapsi [6]
##   annotation chr      start      end strand cluster_id   deltapsi  p.adjust
```

```
##   <fct>      <chr>    <int>    <int> <chr>  <chr>           <dbl>     <dbl>
## 1 gencode    chr7  43648652 43650493 -      clu_35616_- -0.0135   3.89e-114
## 2 gencode    chr7  43648652 43650612 -      clu_35616_- -0.00546  3.89e-114
## 3 gencode    chr7  43648652 43665658 -      clu_35616_-  0.0139   3.89e-114
## 4 gencode    chr7  43648652 43711400 -      clu_35616_-  0.000345 3.89e-114
## 5 gencode    chr7  43648652 43729429 -      clu_35616_- -0.00985  3.89e-114
## 6 gencode    chr7  43650712 43656033 -      clu_35616_- -0.00196  3.89e-114
## # i 7 more variables: transcript_ids <chr>, min_intron_number <int>,
## #   mode_intron_number <dbl>, gene <chr>, biotype <chr>,
## #   genes_in_cluster <chr>, is_first_intron <lgl>
```