iSTOP Paper

This project accompanies the iSTOP R package, and is intended to aid reproduction of figures for a forthcoming publication detailing the scope of possible nonsense mutations capable using CRISPR mediated base editing. If you just want to search for iSTOP targetable sites in your favorite gene, you should head over the the iSTOP R package page.

Installation

To reproduce the analysis in TBD, begin by installing R (~100 MB) and RStudio (~500 MB).

Then, clone this project to your computer using the big green `Clone or download>Download ZIP` button near the top of this page. Then unzip and open the project by clicking on the `iSTOP-paper.Rproj` file. This will open RStudio with the working directory set to the project's folder.

System requirements

This analysis requires ~11 GB of disk space, at least 4 GB of memory, and a little bit of patience. While the analysis should be cross-platform, it has only been tested on macOS Sierra Version 10.12.4 with a 4 GHz processor and 32 GB of memory. Processing time on this system using 4 cores for just the Human genome is ~30 minutes to locate all iSTOP codons and targets, and an additional ~40 minutes for each RFLP annotation width (i.e. `add_RFLP(width = 150)` takes ~40 minutes).

R packages

To install all necessary R packages, run the following R commands in the RStudio console.

```
# Source the Biocoductor installation tool - installs and loads the
# BiocInstaller package which provides the biocLite function.
source("https://bioconductor.org/biocLite.R")

# Install the following required packages ~ 350 MB
BiocInstaller::biocLite(c(
  # Packages from CRAN (cran.r-project.org)
  'tidyverse',
  'gridExtra',
  'cowplot',
  'assertthat',
  'pbapply',
  'devtools',
  # Packages from GitHub (github.com) (version with genome_*_join)
  'dgrtwo/fuzzyjoin',
  # Packages from Bioconductor (bioconductor.org)
  'BSgenome',
  'Biostrings',
  'GenomicRanges',
  'IRanges'
))

# Install genomes and annotation packages ~ 3 GB
```

```
BiocInstaller::biocLite(c(
  # Arabidopsis (not available from UCSC)
  'TxDb.Athaliana.BioMart.plantsmart28', # ~ 24  MB
  'org.At.tair.db',                      # ~ 239 MB
  # Genomes
  'BSgenome.Hsapiens.UCSC.hg38',         # ~ 802 MB
  'BSgenome.Celegans.UCSC.ce11',         # ~ 25  MB
  'BSgenome.Dmelanogaster.UCSC.dm6',     # ~ 36  MB
  'BSgenome.Drerio.UCSC.danRer10',       # ~ 343 MB
  'BSgenome.Mmusculus.UCSC.mm10',        # ~ 683 MB
  'BSgenome.Rnorvegicus.UCSC.rn6',       # ~ 719 MB
  'BSgenome.Scerevisiae.UCSC.sacCer3',   # ~ 3   MB
  'BSgenome.Athaliana.TAIR.TAIR9'        # ~ 34  MB
))

# If all goes well, install the iSTOP package hosted on GitHub
BiocInstaller::biocLite('CicciaLab/iSTOP')
```

Datasets

COSMIC

Download the "COSMIC Mutation Data" (~300 MB compressed) from the Catalogue of Somatic Mutations in Cancer. This dataset requires registration with a valid email address, then in your terminal (not the R console!) you can download the file with the following commands (substitute `your_email_address` with the email used to register).

```
sftp "your_email_address"@sftp-cancer.sanger.ac.uk
# You will be prompted for the password you provided when you registered
get /files/grch38/cosmic/v80/CosmicMutantExport.tsv.gz
```

Move this file to the `data/COSMIC` directory of this project. It should already be named `CosmicMutantExport.tsv.gz`.

CGC

Download the Cancer Gene Census (CGC) dataset by clicking on the `CSV` Export button. You will need to login with the same credentials used to download the COSMIC dataset. Move this file to the `data/COSMIC` directory of this project. Rename the file `CGC.csv`.

CDS

Download CDS coordinates for each genome (~120 MB). Back in the RStudio console, run the following commands.

```
library(tidyverse)
library(iSTOP)

CDS_Athaliana_BioMart_plantsmart28() %>%
  write_csv('data/CDS/Athaliana-plantsmart28.csv')
CDS_Celegans_UCSC_ce11() %>%
  write_csv('data/CDS/Celegans-ce11.csv')
CDS_Dmelanogaster_UCSC_dm6() %>%
```

```
  write_csv('data/CDS/Dmelanogaster-dm6.csv')
CDS_Drerio_UCSC_danRer10() %>%
  write_csv('data/CDS/Drerio-danRer10.csv')
CDS_Hsapiens_UCSC_hg38() %>%
  write_csv('data/CDS/Hsapiens-hg38.csv')
CDS_Mmusculus_UCSC_mm10() %>%
  write_csv('data/CDS/Mmusculus-mm10.csv')
CDS_Rnorvegicus_UCSC_rn6() %>%
  write_csv('data/CDS/Rnorvegicus-rn6.csv')
CDS_Scerevisiae_UCSC_sacCer3() %>%
  write_csv('data/CDS/Scerevisiae-sacCer3.csv')
```

Comprehensive search for iSTOP targetable sites

Load all required packages and functions with the following commands.

```
library(tidyverse)
library(stringr)
library(iSTOP)

# Source all R functions defined in this project
list.files('R/functions', '[.]R', full.names = T) %>% walk(source)
```

Given CDS coordinates and genomes, search for all iSTOP sites with the following commands. Raw results will be saved to the `data/iSTOP` directory, and a compacted version with RFLP annotations will be saved to the `data/iSTOP-compact` directory. To dramatically reduce computation time, comment out the `add_RFLP` lines. The `add_RFLP` lines for the Human dataset are only required to reproduce Figure 3F and Supplementary Figure 3A.

```
# Adjust the number of cores for parallel computation to suit your computer
# Assume that each core will require ~2.5 GB of Memory
# Set to 1 or 2 if you are unsure. Only 1 core is supported on Windows
cores = 1

# Only the human datasets are necessary to reproduce Figure 3
read_csv('data/CDS/Hsapiens-hg38.csv', col_types = 'cciccii') %>%
  locate_codons(BSgenome.Hsapiens.UCSC.hg38::Hsapiens, cores = cores) %>%
  locate_iSTOP(BSgenome.Hsapiens.UCSC.hg38::Hsapiens) %>%
  write_csv('data/iSTOP/Hsapiens-hg38.csv') %>%       # ~ 1 GB
  compact_iSTOP %>%
  add_RFLP(width = 150, cores = cores) %>%
  add_RFLP(width = 100, cores = cores) %>%
  add_RFLP(width = 50,  cores = cores) %>%
  write_csv('data/iSTOP-compact/Hsapiens-hg38.csv')

# The remaining species datasets are necessary to reproduce Figure 4
# This script will run the above command for the non-human species
source('R/scripts/iSTOP-non-human-species.R')
```

Once complete, compute summaries by codon and untargetable datasets with the following command.

```
# Writes two files each to
# `data/iSTOP-by-codon` and `data/iSTOP-untargetable`
list.files('data/iSTOP', '[.]csv', full.names = T) %>%
```

```
  pbapply::pblapply(summarize_by_codon, cl = cores)

# Summarize targetability for all species on codon, ORF and gene levels
source('R/scripts/summarize-by-codon-ORF-gene.R')
```

Analysis of COSMIC nonsense mutations

The raw COSMIC dataset can be cleaned and summarized by sourcing the `R/Clean-COSMIC.R` script. This will add three datasets to the `data/COSMIC` directory.

1. `COSMIC-iSTOP.csv` - (~360 MB) All frameshift and substitution mutations for GRCh38, with aggregated cancer types, and annotated as to whether or not the mutation corresponds to an iSTOP targetable coordinate.
2. `COSMIC-nonsense.csv` - (~7 MB) Only nonsense mutations from `COSMIC-iSTOP.csv`
3. `COSMIC-summary-by-cancer.csv` - Summary by cancer type that details the frequency of nonsense, and targetability with iSTOP
4. `COSMIC-summary-by-gene.csv` - Summary by gene that includes test results for frequent stoppers (likely tumor suppressors). Note that test results for "All cancers" are simply the smallest observed p-value across all cancer subtypes for a given gene.

```
source('Clean-COSMIC.R')
```

Reproducing Figures

Each script will write figures as PDF files to the `figures` directory of the project. Final figures for the paper were edited in Inkscape to reduce the size of the files and improve readability of figure legends and axis labels.

```
library(tidyverse)
library(iSTOP)
library(stringr)
library(gtable)
library(gridExtra)
library(cowplot)

source('R/figures/Figure-1.R')
source('R/figures/Figure-3.R')
source('R/figures/Figure-4.R')
source('R/figures/Figure-5.R')
source('R/figures/Supp-Figure-3.R')
source('R/figures/Supp-Figure-5.R')
```

Figure 5E demonstrates the utility of the iSTOP package for visualizing iSTOP targetable sites. This figure can be reproduced for any gene using the following commands.

```
library(tidyverse)
library(iSTOP)

my_gene = 'ATM'

COSMIC <- read_csv('data/COSMIC/COSMIC-nonsense.csv')
CDS    <- read_csv('data/CDS/Hsapiens-hg38.csv')
```

```
iSTOP <-
  filter(CDS, gene == my_gene) %>%
  locate_codons(BSgenome.Hsapiens.UCSC.hg38::Hsapiens) %>%
  locate_iSTOP(BSgenome.Hsapiens.UCSC.hg38::Hsapiens) %>%
  add_RFLP(width = 50)

Fig5E <- plot_spliced_isoforms(
  gene   = my_gene,
  coords = filter(CDS, tx %in% iSTOP$tx),
  colors = c('red', 'black', 'blue', 'darkgreen'),
  `Nonsense in cancer`    = COSMIC,
  `CAA, CAG, CGA, TGG`    = iSTOP,
  `iSTOP targetable`      = filter(iSTOP, match_any),
  `Verifiable with RFLP` = filter(iSTOP, match_any & has(RFLP_50))
)

Fig5E
ggsave('figures/Figure-5E.pdf', Fig5E, width = 18, height = 3)
```

```
# Clean up workspace - Leave Figures and Figure data
rm(list = setdiff(ls(), ls(pattern = '^Fig|COSMIC|CDS|iSTOP')))
```

Session Information

This analysis was successfully performed with the following system, and package versions:

```
## Session info -------------------------------------------------------------------------------
```

```
##  setting  value
##  version  R version 3.3.3 (2017-03-06)
##  system   x86_64, darwin13.4.0
##  ui       RStudio (1.0.136)
##  language (EN)
##  collate  en_US.UTF-8
##  tz       America/New_York
##  date     2017-04-07
```

```
## Packages -----------------------------------------------------------------------------------
```

```
##  package              * version    date       source
##  assertthat             0.1        2013-12-06 CRAN (R 3.3.0)
##  backports              1.0.5      2017-01-18 CRAN (R 3.3.2)
##  Biobase                2.34.0     2016-10-18 Bioconductor
##  BiocGenerics           0.20.0     2016-10-18 Bioconductor
##  BiocInstaller          1.24.0     2016-10-18 Bioconductor
##  BiocParallel           1.8.1      2016-10-30 Bioconductor
##  Biostrings             2.42.1     2016-12-01 Bioconductor
##  bitops                 1.0-6      2013-08-17 CRAN (R 3.3.0)
##  broom                  0.4.2      2017-02-13 CRAN (R 3.3.2)
##  BSgenome               1.42.0     2016-10-18 Bioconductor
##  colorspace             1.3-2      2016-12-14 CRAN (R 3.3.2)
##  cowplot              * 0.7.0      2016-10-28 CRAN (R 3.3.0)
##  DBI                    0.6-1      2017-04-01 CRAN (R 3.3.2)
##  devtools               1.12.0     2016-06-24 CRAN (R 3.3.0)
```

```
##  digest                0.6.12     2017-01-27 CRAN (R 3.3.2)
##  dplyr               * 0.5.0      2016-06-24 CRAN (R 3.3.0)
##  evaluate              0.10       2016-10-11 CRAN (R 3.3.0)
##  forcats               0.2.0      2017-01-23 CRAN (R 3.3.2)
##  foreign               0.8-67     2016-09-13 CRAN (R 3.3.3)
##  fuzzyjoin             0.1.2.9000 2017-04-05 Github (dgrtwo/fuzzyjoin@2f30724)
##  GenomeInfoDb          1.10.3     2017-02-07 Bioconductor
##  GenomicAlignments     1.10.1     2017-03-18 Bioconductor
##  GenomicRanges         1.26.4     2017-03-18 Bioconductor
##  ggplot2             * 2.2.1      2016-12-30 CRAN (R 3.3.2)
##  gridExtra           * 2.2.1      2016-02-29 CRAN (R 3.3.0)
##  gtable              * 0.2.0      2016-02-26 CRAN (R 3.3.0)
##  haven                 1.0.0      2016-09-23 CRAN (R 3.3.0)
##  hms                   0.3        2016-11-22 CRAN (R 3.3.2)
##  htmltools             0.3.5      2016-03-21 CRAN (R 3.3.0)
##  httr                  1.2.1      2016-07-03 CRAN (R 3.3.0)
##  IRanges               2.8.2      2017-03-18 Bioconductor
##  iSTOP               * 0.1.0      2017-04-05 Github (ericedwardbryant/iSTOP@199f2c7)
##  jsonlite              1.3        2017-02-28 CRAN (R 3.3.2)
##  knitr                 1.15.1     2016-11-22 CRAN (R 3.3.2)
##  lattice               0.20-35    2017-03-25 CRAN (R 3.3.2)
##  lazyeval              0.2.0      2016-06-12 CRAN (R 3.3.0)
##  lubridate             1.6.0      2016-09-13 CRAN (R 3.3.0)
##  magrittr              1.5        2014-11-22 CRAN (R 3.3.0)
##  Matrix                1.2-8      2017-01-20 CRAN (R 3.3.3)
##  memoise               1.0.0      2016-01-29 CRAN (R 3.3.0)
##  mnormt                1.5-5      2016-10-15 CRAN (R 3.3.0)
##  modelr                0.1.0      2016-08-31 CRAN (R 3.3.0)
##  munsell               0.4.3      2016-02-13 CRAN (R 3.3.0)
##  nlme                  3.1-131    2017-02-06 CRAN (R 3.3.3)
##  pbapply               1.3-2      2017-03-01 CRAN (R 3.3.2)
##  plyr                  1.8.4      2016-06-08 CRAN (R 3.3.0)
##  psych                 1.7.3.21   2017-03-22 CRAN (R 3.3.2)
##  purrr               * 0.2.2      2016-06-18 CRAN (R 3.3.0)
##  R6                    2.2.0      2016-10-05 CRAN (R 3.3.0)
##  Rcpp                  0.12.10    2017-03-19 CRAN (R 3.3.2)
##  RCurl                 1.95-4.8   2016-03-01 CRAN (R 3.3.0)
##  readr               * 1.1.0      2017-03-22 CRAN (R 3.3.2)
##  readxl                0.1.1      2016-03-28 CRAN (R 3.3.0)
##  reshape2              1.4.2      2016-10-22 CRAN (R 3.3.0)
##  rmarkdown             1.4        2017-03-24 CRAN (R 3.3.3)
##  rprojroot             1.2        2017-01-16 CRAN (R 3.3.2)
##  Rsamtools             1.26.1     2016-10-22 Bioconductor
##  rstudioapi            0.6        2016-06-27 CRAN (R 3.3.0)
##  rtracklayer           1.34.2     2017-02-19 Bioconductor
##  rvest                 0.3.2      2016-06-17 CRAN (R 3.3.0)
##  S4Vectors             0.12.2     2017-03-18 Bioconductor
##  scales                0.4.1      2016-11-09 CRAN (R 3.3.2)
##  stringi               1.1.3      2017-03-21 CRAN (R 3.3.2)
##  stringr             * 1.2.0      2017-02-18 CRAN (R 3.3.2)
##  SummarizedExperiment  1.4.0      2016-10-18 Bioconductor
##  tibble              * 1.3.0      2017-04-01 CRAN (R 3.3.2)
##  tidyr               * 0.6.1      2017-01-10 CRAN (R 3.3.2)
##  tidyverse           * 1.1.1      2017-01-27 CRAN (R 3.3.2)
```

```
##  withr                1.0.2      2016-06-20 CRAN (R 3.3.0)
##  XML                  3.98-1.6   2017-03-30 CRAN (R 3.3.2)
##  xml2                 1.1.1      2017-01-24 CRAN (R 3.3.2)
##  XVector              0.14.1     2017-03-18 Bioconductor
##  yaml                 2.1.14     2016-11-12 CRAN (R 3.3.2)
##  zlibbioc             1.20.0     2016-10-18 Bioconductor
```