ABE - extended methods

**Eric Bryant**1, **Pierre Billon**2, and **Alberto Ciccia**2,3,\*

1 Department of Biological Sciences, Columbia University, New York, NY, USA  
2 Department of Genetics & Development, Columbia University Medical Center, New York, NY, USA  
3 Lead Contact  
\* Correspondence: [ac3685@columbia.edu](mailto:ac3685@columbia.edu)

Abstract …

# SETUP & REQUIREMENTS

## SOFTWARE

This project is available on [GitHub](https://github.com/CicciaLab/Manuscript-ABE.git). The analysis requires [R](https://cloud.r-project.org), and recommends [RStudio](https://www.rstudio.com/products/rstudio/download/#download). After installing R & RStudio, and downloading this project, open the RStudio project file named Manuscript-ABE.Rproj.

The analysis will aslo require several R packages. The following command will check for any missing required packages, and install them to R/library (this directory can be changed by modifying the project’s .Rprofile).

.Rprofile$install\_requirements()

After all packages have been installed, prepare the workspace by loading requirements with the following command.

.Rprofile$load\_requirements()

## DATASETS

Before beginning the analysis, there are several datasets that must be downloaded. The following sections provide detailed instructions for downloading all required datasets.

## Coding sequence coordinates

Coding sequence (CDS) coordinates can be downloaded from the UCSC genome browser with the following command.

data\_import$CDS(save\_as = 'data/CDS/Hsapiens-UCSC-hg38.csv')  
data\_import$CDS\_contiguous\_intervals(  
 given = 'data/CDS/Hsapiens-UCSC-hg38.csv',   
 save\_as = 'data/CDS/Hsapiens-UCSC-hg38-contiguous-intervals.csv'  
)

## Cosmic

To download the [Catalogue of Somatic Mutations in Cancer](https://cancer.sanger.ac.uk/cosmic/download) (Forbes et al., 2017), first [register with a valid email address](https://cancer.sanger.ac.uk/cosmic/register), then download the CosmicMutantExport.tsv.gz dataset into the data/COSMIC directory using the following commands:

# In Terminal  
cd data/COSMIC # Navigate to destination  
sftp "your@email.com"@sftp-cancer.sanger.ac.uk # Enter password when prompted  
get /files/grch38/cosmic/v83/CosmicMutantExport.tsv.gz

After downloading the main Cosmic dataset, we will generate three processed versions of this dataset using the following commands.

data\_import$Cosmic\_clean(  
 given\_cosmic = 'data/COSMIC/CosmicMutantExport.tsv.gz',   
 given\_cds = 'data/CDS/Hsapiens-UCSC-hg38-contiguous-intervals.csv',  
 save\_as = 'data/COSMIC/Cosmic.csv.gz'  
)  
data\_import$Cosmic\_mutation\_subset(  
 given = 'data/COSMIC/Cosmic.csv.gz',  
 class = 'Nonsense',  
 save\_as = 'data/COSMIC/Cosmic-nonsense.csv'  
)  
data\_import$Cosmic\_mutation\_subset(  
 given = 'data/COSMIC/Cosmic.csv.gz',  
 class = 'Nonstop extension',  
 save\_as = 'data/COSMIC/Cosmic-nonstop-extension.csv'  
)

## ClinVar

[Clinical Variants](https://www.ncbi.nlm.nih.gov/clinvar/) can be downloaded using the following commands:

# In Terminal  
cd data/ClinVar # Navigate to destination  
wget ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/tab\_delimited/variant\_summary.txt.gz  
wget ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/tab\_delimited/allele\_gene.txt.gz  
wget ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/tab\_delimited/README # Conatins field descriptions

## Coriell

Go to [Coriell](https://www.coriell.org/0/Sections/BrowseCatalog/Genes.aspx?PgId=5) and enter \*.\* into the search bar. Then click on Export to Excel and, when prompted, check the boxes for ID, Description, Gene and Mutations. Then click Export All Samples in Grid. Move this file to the data/Coriell directory of this project and rename the file to Catalog-Export.csv.

# ANALYSIS

#### SESSION INFORMATION

── R 3.4.2 "Short Summer" ──────────────────────────────────────────────────────

● System macOS High Sierra 10.13.1   
● Platform x86\_64-apple-darwin15.6.0 (64-bit)   
● Library /Users/ericbryant/GitServer/Manuscript-ABE/R/library

── Packages attached ───────────────────────────────────────────────────────────

✔ tidyverse 1.2.1 (CRAN 2017-11-20) ✔ purrr 0.2.4 (CRAN 2017-11-20)   
✔ ggplot2 2.2.1 (CRAN 2017-11-20) ✔ dplyr 0.7.4 (CRAN 2017-11-20)   
✔ tibble 1.3.4 (CRAN 2017-11-20) ✔ stringr 1.2.0 (CRAN 2017-11-20)   
✔ tidyr 0.7.2 (CRAN 2017-11-20) ✔ forcats 0.2.0 (CRAN 2017-11-20)   
✔ readr 1.1.1 (CRAN 2017-11-20) ✔ iSTOP 0.1.0 (Github.com/EricEdwardBryant/iSTOP@99fc1c1)

── Packages loaded via namespace ───────────────────────────────────────────────

✔ Rcpp 0.12.13 (CRAN 2017-11-20) ✔ mnormt 1.5-5 (CRAN 2017-11-20)  
✔ parallel 3.4.2 (CRAN 2017-11-20) ✔ cli 1.0.0 (CRAN 2017-11-20)  
✔ stringi 1.1.6 (CRAN 2017-11-20) ✔ lazyeval 0.2.1 (CRAN 2017-11-20)  
✔ rprojroot 1.2 (CRAN 2017-11-20) ✔ BiocGenerics 0.22.1 (Bioc 3.5)   
✔ bindr 0.1 (CRAN 2017-11-20) ✔ pbapply 1.3-3 (CRAN 2017-11-20)  
✔ rtracklayer 1.36.6 (Bioc 3.5) ✔ SummarizedExperiment 1.6.5 (Bioc 3.5)   
✔ knitr 1.17 (CRAN 2017-11-20) ✔ IRanges 2.10.5 (Bioc 3.5)   
✔ lubridate 1.7.1 (CRAN 2017-11-20) ✔ BiocParallel 1.10.1 (Bioc 3.5)   
✔ GenomicAlignments 1.12.2 (Bioc 3.5) ✔ scales 0.5.0 (CRAN 2017-11-20)  
✔ R6 2.2.2 (CRAN 2017-11-20) ✔ zlibbioc 1.22.0 (Bioc 3.5)   
✔ reshape2 1.4.2 (CRAN 2017-11-20) ✔ haven 1.1.0 (CRAN 2017-11-20)  
✔ fuzzyjoin 0.1.3 (CRAN 2017-11-20) ✔ broom 0.4.2 (CRAN 2017-11-20)  
✔ gtable 0.2.0 (CRAN 2017-11-20) ✔ pkgconfig 2.0.1 (CRAN 2017-11-20)  
✔ rmarkdown 1.8 (CRAN 2017-11-20) ✔ XML 3.98-1.9 (CRAN 2017-11-20)  
✔ bitops 1.0-6 (CRAN 2017-11-20) ✔ psych 1.7.8 (CRAN 2017-11-20)  
✔ rstudioapi 0.7 (CRAN 2017-11-20) ✔ plyr 1.8.4 (CRAN 2017-11-20)  
✔ RCurl 1.95-4.8 (CRAN 2017-11-20) ✔ Matrix 1.2-11 (CRAN 2017-11-20)  
✔ grid 3.4.2 (CRAN 2017-11-20) ✔ htmltools 0.3.6 (CRAN 2017-11-20)  
✔ rlang 0.1.4 (CRAN 2017-11-20) ✔ colorspace 1.3-2 (CRAN 2017-11-20)  
✔ GenomeInfoDb 1.12.3 (Bioc 3.5) ✔ rvest 0.3.2 (CRAN 2017-11-20)  
✔ compiler 3.4.2 (CRAN 2017-11-20) ✔ XVector 0.16.0 (Bioc 3.5)   
✔ Biostrings 2.44.2 (Bioc 3.5) ✔ GenomicRanges 1.28.6 (Bioc 3.5)   
✔ bindrcpp 0.2 (CRAN 2017-11-20) ✔ digest 0.6.12 (CRAN 2017-11-20)  
✔ DelayedArray 0.2.7 (Bioc 3.5) ✔ glue 1.2.0 (CRAN 2017-11-20)  
✔ munsell 0.4.3 (CRAN 2017-11-20) ✔ lattice 0.20-35 (CRAN 2017-11-20)  
✔ S4Vectors 0.14.7 (Bioc 3.5) ✔ backports 1.1.1 (CRAN 2017-11-20)  
✔ matrixStats 0.52.2 (CRAN 2017-11-20) ✔ yaml 2.1.14 (CRAN 2017-11-20)  
✔ hms 0.3 (CRAN 2017-11-20) ✔ Rsamtools 1.28.0 (Bioc 3.5)   
✔ tools 3.4.2 (CRAN 2017-11-20) ✔ cellranger 1.1.0 (CRAN 2017-11-20)  
✔ foreign 0.8-69 (CRAN 2017-11-20) ✔ GenomeInfoDbData 0.99.0 (Bioc 3.5)   
✔ xml2 1.1.1 (CRAN 2017-11-20) ✔ BSgenome 1.44.2 (Bioc 3.5)   
✔ nlme 3.1-131 (CRAN 2017-11-20) ✔ stats4 3.4.2 (CRAN 2017-11-20)  
✔ evaluate 0.10.1 (CRAN 2017-11-20) ✔ assertthat 0.2.0 (CRAN 2017-11-20)  
✔ readxl 1.0.0 (CRAN 2017-11-20) ✔ modelr 0.1.1 (CRAN 2017-11-20)  
✔ crayon 1.3.4 (CRAN 2017-11-20) ✔ jsonlite 1.5 (CRAN 2017-11-20)  
✔ magrittr 1.5 (CRAN 2017-11-20) ✔ httr 1.3.1 (CRAN 2017-11-20)  
✔ Rcpp 0.12.13 (CRAN 2017-11-20) ✔ Biobase 2.36.2 (Bioc 3.5)

#### REFERENCES

Forbes, S.A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., Cole, C.G., Ward, S., Dawson, E., Ponting, L., et al. (2017). COSMIC: Somatic cancer genetics at high-resolution. Nucleic Acids Res *45*, D777–D783.