Supplementary Information

APOBEC3B expression in breast cancer reflects cellular proliferation, while a deletion polymorphism is associated with immune activation

David W. Cescon*, Benjamin Haibe-Kains*, Tak W. Mak

* Co-first authors

Full Reproducibility of the Analysis Results

We will describe how to fully reproduce the figures and tables reported in the main manuscript. We automated the analysis pipeline so that minimal manual interaction is required to reproduce our results. To do this, one must simply:

1. Set up the software environment
2. Run the R scripts

Set up the software environment

We developed and tested our analysis pipeline using R running on linux and Mac OSX platforms.

To mimic our software environment the following R packages should be installed:

- R version 3.1.1 (2014-07-10), `x86_64-unknown-linux-gnu`
- Base packages: base, datasets, graphics, grDevices, grid, methods, parallel, splines, stats, utils
- Other packages: AER 1.2-2, AnnotationDbi 1.26.0, Biobase 2.24.0, BiocGenerics 0.10.0, BiocInstaller 1.14.2, biomaRt 2.20.0, bitops 1.0-6, car 2.0-21, cgdsr 1.1.32, cluster 1.15.3, DBI 0.3.0, devtools 1.5, foreign 0.8-61, Formula 1.1-2, gdata 2.13.3, genefu 1.13.2, GenomeInfoDb 1.0.2, ggplot2 1.0.0, Hmisc 3.14-5, iC10 1.1.2, iC10TrainingData 1.0.1, igraph 0.7.1, inSilicoDb 2.1.1, jetset 1.6.0, lattice 0.20-29, lmtest 0.9-33, lsa 0.73, MASS 7.3-34, maxLik 1.2-0, mclust 4.3, MetaGx 0.0.2, miscTools 0.6-16, mlogit 0.2-4, mRMRe 2.0.5, nnet 7.3-8, org.Hs.eg.db 2.14.0, pamr 1.55, plotrix 3.5-7, prodlim 1.4.5, RCurl 1.95-4.3, reshape2 1.4, rjson 0.2.14, RSQLite 0.11.4, sandwich 2.3-2, SnowballC 0.5, SuppDists 1.1-9.1, survcomp 1.15.1, survival 2.37-7, testthat 0.8.1, vcd 1.3-2, WriteXLS 3.5.0, xtable 1.7-4, zoo 1.7-11
- Loaded via a namespace (and not attached): acepack 1.3-3.3, amap 0.8-12, bootstrap 2014.4, colorspace 1.2-4, digest 0.6.4, evaluate 0.5.5, GSA 1.03, gtable 0.1.2, gtools 3.4.1, httr 0.5, IRanges 1.22.10, KernSmooth 2.23-13, latticeExtra 0.6-26, lava 1.2.6, memoise 0.2.1, munsell 0.4.2, plyr 1.8.1, proto 0.3-10, RColorBrewer 1.0-5, Rcpp 0.11.3, rmeta 2.16, R.methodsS3 1.6.1, R.oo 1.18.0, rpart 4.1-8, scales 0.2.4, statmod 1.4.20, stats4 3.1.1, stringr 0.6.2, survivalROC 1.0.3, tcltk 3.1.1, tools 3.1.1, whisker 0.3-2, XML 3.98-1.1

All these packages are available on CRAN[1] or Bioconductor[2]

Run the following commands in a R session to install all the required packages:

```
source("http://bioconductor.org/biocLite.R")
biocLite(c("AnnotationDbi", "Biobase",
    "BiocGenerics", "biomaRt", "bitops", "cgdsr",
    "cluster", "DBI", "Formula", "gdata",
    "genefu", "Hmisc", "igraph", "inSilicoDb",
    "jetset", "lattice", "lsa", "mclust", "mRMRe", "org.Hs.eg.db",
    "plotrix", "prodlim", "RCurl", "rjson", "R.methodsS3",
    "RSQLite", "SnowballC" "survcomp", "survival", "WriteXLS",
    "xtable", "devtools")
)
```

[1] http://cran.r-project.org
[2] http://www.bioconductor.org

The latest version of MetaGx, survcomp and gene can be installed using the following commands:

```
library(devtools)
install_github("survcomp", username="bhaibeka", ref="master")
install_github("genefu", username="bhaibeka", ref="master")
install_github("MetaGx", username="bhaibeka", ref="master")
```

Note that you may need to install Perl[3] and its module Text::CSV_XS for the WriteXLS package to write xls file; once Perl is installed in your system, use the following command to install the Text::CSV_XS module through CPAN[4]:

```
cpan Text/CSV_XS.pm
```

Lastly, follow the instructions on the CBS website to properly install the jetset package or use the following commands in R:

```
download.file(url="http://www.cbs.dtu.dk/biotools/jetset/current/jetset_1.4.0.tar.gz",
 destfile="jetset_1.4.0.tar.gz")
install.packages("jetset_1.4.0.tar.gz", repos=NULL, type="source")
```

Once the packages are installed, uncompress the archive provided as **Supplementary data** accompanying the manuscript[5]. This should create a directory on the file system containing the following files:

`apobec_.pipeline.R` Master script running all the scripts listed above to generate the analysis results.

`apobec_data_TCGA.R` Script to to downlaod and format TCGA data.

`apobec_data_METABRIC.R` Script to to downlaod and format METABRIC data.

`apobec_analysis.R` Script generating all the figures and tables reported in the manuscript.

`gsea2-2.1.0.jar` GSEA java executable; it can also be downloaded from the GSEA website[6].

`c5.all.v4.0.entrez.gmt` Definition of genesets based on Entrez Gene IDs; it can also be downloaded from the GSEA website[7].

All the files required to run the automated analysis pipeline are now in place. It is worth noting that raw gene expression and drug sensitivity data are voluminous, please ensure that at least 5GB of storage are available.

Run the R scripts

Open a terminal window and go to the `apobec` directory. You can easily run the analysis pipeline either in batch mode or in a R session. Before running the pipeline you can specify the number of CPU cores you want to allocate to the analysis (by default only 1 CPU core will be used). To do so, open the script `apobec_pipeline.R` and update line #33:

```
nbcore <- 4
```

[3] http://www.perl.org/get.html
[4] http://www.cpan.org/modules/INSTALL.html
[5] The code is also available on GitHub within the **https://github.com/bhaibeka/APOBEC3B** repository.
[6] http://www.broadinstitute.org/gsea/msigdb/download_file.jsp?filePath=/resources/software/gsea2-2.1.0.jar
[7] http://www.broadinstitute.org/gsea/msigdb/download_file.jsp?filePath=/resources/msigdb/4.0/c5.all.v4.0.entrez.gmt

to allocate four CPU cores for instance.

To run the full pipeline in batch mode, simply type the following command:

```
R CMD BATCH apobec_pipeline.R Rout &
```

The progress of the pipeline could be monitored using the following command:

```
tail -f Rout
```

To run the full analysis pipeline in an R session, simply type the following command:

```
source("apobec_pipeline.R")
```

Key messages will be displayed to monitor the progress of the analysis.

The analysis pipeline was developed so that all intermediate analysis results are saved in the directories `data` and `saveres`. Therefore, in case of interruption, the pipeline will restart where it stopped