Introduction to *AMARETTO*

***Jayendra Shinde*[1], *Celine Everaert*[2], *Shaimaa Bakr*[1], *Mohsen Nabian*[2], *Jishu Xu*[2], *Nathalie Pochet*[*2], *and Olivier Gevaert*[†1]**

[1]Stanford Center for Biomedical Informatics Research (BMIR), Department of Medicine and Biomedical Data Science, 1265 Welch Rd, Stanford, CA, USA
[2]Brigham and Women's Hospital, Harvard Medical School, Broad Institute of MIT and Harvard, Cambridge, MA, USA

[*]npochet@broadinstitute.org [†]olivier.gevaert@stanford.edu

15 February 2019

Abstract

Integrating an increasing number of available multi-omics cancer data remains one of the main challenges to improve our understanding of cancer. One of the main challenges is using multi-omics data for identifying novel cancer driver genes. We have developed an algorithm, called AMARETTO, that integrates copy number, DNA methylation and gene expression data to identify a set of driver genes by analyzing cancer samples and connects them to clusters of co-expressed genes, which we define as modules. We applied AMARETTO in a pancancer setting to identify cancer driver genes and their modules on multiple cancer sites. AMARETTO captures modules enriched in angiogenesis, cell cycle and EMT, and modules that accurately predict survival and molecular subtypes. This allows AMARETTO to identify novel cancer driver genes directing canonical cancer pathways.

Package

Report issues on https://github.com/gevaertlab/AMARETTO

Contents

1 Introduction

The package *AMARETTO* contains functions to use the statistical algorithm AMARETTO, an algorithm to identify cancer drivers by integrating a variety of omics data from cancer and normal tissue. Due to the increasing availability of multi-omics data sets, there is a need for computational methods that integrate multi-omics data set and create knowl-edge. Especially in the field of cancer research, large international projects such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) are producing large quantities of multi-omics data for each cancer site. However it remains unknown which profile is the most meaningful and how to effciently integrate different omics profiles. AMARETTO is an algorithm to unravel cancer drivers by reducing the data dimensionality into cancer modules. AMARETTO first models the effects of genomic/epigenomic data on disease specific gene expression. AMARETTO's second step involves constructing co-expressed modules to connect the cancer drivers with their downstream targets. We applied AMARETTO to several cancer sites of the TCGA project allowing to identify several cancer driver genes of interest, including novel genes in addition to known drivers of cancer. This package also includes functionality to access TCGA data directly so the user can immediately run AMARETTO on the most recent version of the data.

2 Installation Instructions

To install the AMARETTO package, the easiest way is through bioconductor:

```
--------------------------------------------------------------------
>install.packages("BiocManager")
>BiocManager::install("gevaertlab/AMARETTO")
--------------------------------------------------------------------
```

Another way to install AMARETTO is to first download the appropriate file for your platform from the Bioconductor website http://www.bioconductor.org/. For Windows, start R and select the Packages menu, then Install package from local zip file. Find and highlight the location of the zip file and click on open. For Linux/Unix, use the usual command R CMD INSTALL or install from bioconductor

The package can be installed from the GitHub repository using `devtools`:

```
--------------------------------------------------------------------
>library(devtools)
>devtools::install_github("gevaertlab/AMARETTO")
--------------------------------------------------------------------
```

Help files. Detailed information on AMARETTO package functions can be obtained in the help files. For example, to view the help file for the function AMARETTO in an R session, use ?AMARETTO.

3 Data Input

AMARETTO combines gene expression, DNA copy number and DNA methylation data into co-expressed gene expression models. Ideally, we recommend a cohort of at least 100 samples for each of these three technologies, where for most patients all data modalities have to be present.AMARETTO can be run with your own data but when interested in TCGA data, AMARETTO can also download TCGA data for you, see the next section.

3.1 Data Access

The data in this vignette is accessible at The Cancer Genome Atlas (TCGA) portal. A programmatic way of downloading data is through the firehose get tool developed by the broad institute ("https://confluence.broadinstitute.org/display/GDAC/ Download"). Firehose get provides a unified way to download data for all cancer sites and all platforms.

3.2 Gene Expression and Copy Number Alterations

We have provided R functions that directly link with firehose get to download mRNA expression data and copy number data processed by GISTIC. Downloading TCGA data has been tested for twenty five cancer sites:

Cancer code	Cancer site
BLCA	bladder urothelial carcinoma
BRCA	breast invasive carcinoma
CESC	cervical squamous cell carcinoma and endocervical adenocarcinoma
CHOL	cholangiocarcinoma
COAD	colon adenocarcinoma
ESCA	esophageal carcinoma
GBM	glioblastoma multiforme
HNSC	head and neck squamous cell carcinoma
KIRC	kidney renal clear cell carcinoma
KIRP	kidney renal papillary cell carcinoma
LAML	acute myeloid leukemia
LGG	brain lower grade glioma
LIHC	liver hepatocellular carcinoma
LUAD	lung adenocarcinoma
LUSC	lung squamous cell carcinoma
OV	ovarian serous cystadenocarcinoma
PAAD	pancreatic adenocarcinoma
PCPG	pheochromocytoma and paraganglioma
READ	rectum adenocarcinoma
SARC	sarcoma
STAD	stomach adenocarcinoma
THCA	thyroid carcinoma
THYM	thymoma
UCEC	endometrial carcinoma
COADREAD	colon cancer + rectal cancer

We also added COADREAD as a combination of colon and rectal cancer, as reports have shown that both can be seen as a single disease. The cancer code is needed to download data from TCGA and one needs to decide on a target location to save the data locally in the TargetDirectory, e.g. the /Downloads/ folder on a mac.

```
------------------------------------------------------------------
> TargetDirectory <- "./Downloads/"
> CancerSite <- "READ"
> DataSetDirectories <- AMARETTO_Download(CancerSite,TargetDirectory)
------------------------------------------------------------------
```

We recommend to use one TargetDirectory for all cancer sites, as this will save all data in one hierarchy is convenient when revisting results later on. The directory structure that is created will also include the data version history, so it is easy to report what version of the data is used. AMARETTO_Download() downloads the data, extracts archives and provides the paths to the downloaded folder for preprocessing. AMARETTO_Download() can also be run without actually downloading the data as follows:

```
------------------------------------------------------------------
> TargetDirectory <- "./Downloads/"
> CancerSite <- "READ"
> DataSetDirectories <- AMARETTO_Download(CancerSite,TargetDirectory,FALSE)
------------------------------------------------------------------
```

This is convenient when revisiting a data set because you want to redo-downstream analysis, but not the actual down- loading. Running this way, will only set the data paths. The next step is preprocessing.

3.3 DNA Methylation Data

DNA methylation data has to be run by MethylMix which is also computationally intensive and therefore we have chosen to provide add the MethylMix output to the AMARETTO package instead of processing the raw DNA methylation data. This functionality is available in the MethylMix package

3.4 Data Preprocessing

The data preprocessing step will take care of preprocessing the gene expression and DNA copy number data. Data preprocessing is done by Preprocess CancerSite which takes the CancerSite and the data set directories as parameters:

```
------------------------------------------------------------------
> CancerSite <- 'READ'
> ProcessedData <- AMARETTO_Preprocess(CancerSite,DataSetDirectories)
------------------------------------------------------------------
```

This function preprocessed the gene expression data and the DNA copy number data. For the gene expression data, different preprocessing is done for microarray and RNA sequencing data. This involves missing value estimation using K-nearest neighbors. Also genes or patients that have more than 10% missing values are removed. Next, batch correction is done using the Combat method. For certain cancer sites, the gene expression data is split up in separate sub-data sets. This function first uses the preprocessing pipeline on each sub-data set separately

and combines the data afterwards. For the DNA copy number data, the GISTIC algorithm output data is used. All genes that are in amplifications or deletions based on GISTIC output are extracted and the segmented DNA copy number data is stored. The segmented DNA copy number data is also batch corrected using Combat.

4 Running AMARETTO

In the case that you run AMARETTO with your own data, three data matrices are needed with preprocessed gene expression, DNA copy number and DNA methylation data, where genes are in the rows and patients are in the columns. Once you have your own data in this format or using a previously downloaded TCGA data set, you can start doing analysis with AMARETTO. First, we need to initialize the algorithm by clustering the gene expression data and creating the regulator data object. This is done by the AMARETTO Initialize function and the TCGA LIHC data set:

```
----------------------------------------------------------------------
> data(ProcessedDataLIHC)
> AMARETTOinit <- AMARETTO_Initialize(MA_matrix = ProcessedDataLIHC$MA_matrix,
                                      CNV_matrix = ProcessedDataLIHC$CNV_matrix,
                                      MET_matrix = ProcessedDataLIHC$MET_matrix,
                                      NrModules = 20, VarPercentage = 50)
----------------------------------------------------------------------
```

Besides the three data sets, you need to decide how many modules to build and how much of the gene expression data is going to be used. For a first run we recommend 100 modules and to use the top 25% most varying genes. The AMARETTOinit object now contains cluster information to initialize an AMARETTO run and also stores the parameters that are required for AMARETTO. Now we can run AMARETTO as follows:

```
----------------------------------------------------------------------
> AMARETTOresults <- AMARETTO_Run(AMARETTOinit)
----------------------------------------------------------------------
```

This can take anywhere from 10 minutes up to 1 hour to build the modules for the TCGA cohorts depending on the number of genes that is modeled and the number of patients that is available. The breast cancer data set (BRCA) is the largest data set and will take the longest time to converge. AMARETTO will stop when less than 1% of the genes are reassigned to other modules. Next, one can test the AMARETTO model on the training set by calculating the Rsquare for each module using the AMARETTO EvaluateTestSet function:

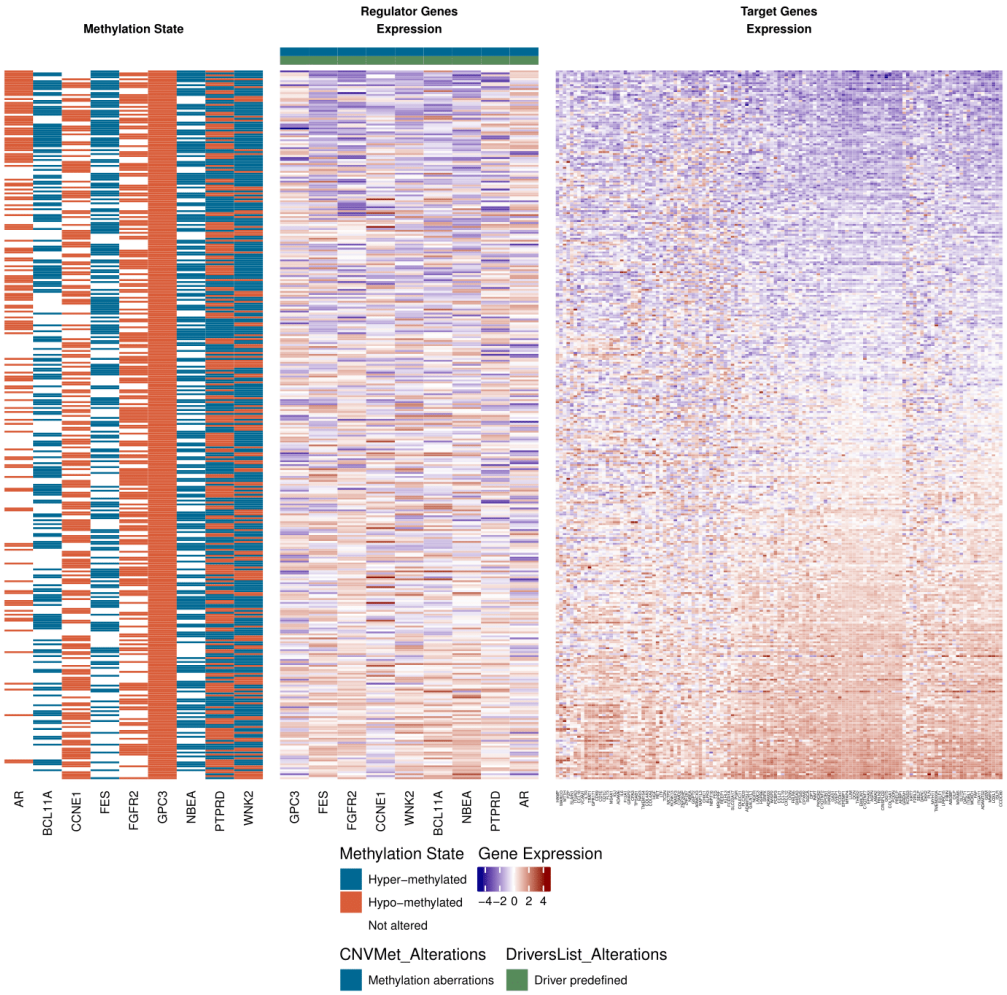
```
----------------------------------------------------------------------
> AMARETTOtestReport <- AMARETTO_EvaluateTestSet(AMARETTOresults,
                        AMARETTOinit$MA_TCGA_Var,AMARETTOinit$RegulatorData)
----------------------------------------------------------------------
```

This function will use the training data to calculate the performance for predicting genes expression values based on the selected regulators. Of course, it is more interesting to use an independent test set. In this case only a gene expression data set is needed, for example from the GEO database. This will allow to check how well the AMARETTO modules are generalizing to new data. Take care that the an independent data set needs to be centered and scaled to unit variance. The AMARETTOtestReport will also give information of how many regulators and cluster members are actually present. The Rsquare performance has

to be interpreted in this context as if many regulators are absent in the test data set due to platform limitations, the performance will be adversely affected. Finally, modules can be visualized using the AMARETTO VisualizeModule function:

```
-------------------------------------------------------------------
> AMARETTO_VisualizeModule(AMARETTOinit,AMARETTOresults,
                    ProcessedDataLIHC$CNV_matrix,ProcessedDataLIHC$MET_matrix,ModuleNr=1)
-------------------------------------------------------------------
```

Additionaly, to a standard version of the heatmap, one can add sample annotations to interogate biological phenotypes.



5 HTML Report of AMARETTO

To retrieve heatmaps for all of the modules and additional tables with gene set enrichment data one can run a HTML report.

```
-----------------------------------------------------------------
> gmt_file<-system.file("templates/H.C2CP.genesets.gmt",package="AMARETTO")
> AMARETTO_HTMLreport(AMARETTOinit,AMARETTOresults,
                      ProcessedDataLIHC$CNV_matrix,ProcessedDataLIHC$MET_matrix,
                      hyper_geo_test_bool = TRUE,hyper_geo_reference = gmt_file, MSIGDB=TRUE)
-----------------------------------------------------------------
```

6 References

1. Champion, M. et al. Module Analysis Captures Pancancer Genetically and Epigenetically Deregulated Cancer Driver Genes for Smoking and Antiviral Response. EBioMedicine 27, 156–166 (2018).
2. Gevaert, O., Villalobos, V., Sikic, B. I. & Plevritis, S. K. Identification of ovarian cancer driver genes by using module network integration of multi-omics data. Interface Focus 3, 20130013–20130013 (2013).
3. Gevaert, O. MethylMix: an R package for identifying DNA methylation-driven genes. Bioinformatics 31, 1839–1841 (2015).

Session info

```
## R version 3.5.2 (2018-12-20)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS High Sierra 10.13.6
##
## Matrix products: default
## BLAS: /Library/Frameworks/R.framework/Versions/3.5/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.5/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] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] BiocStyle_2.10.0
##
## loaded via a namespace (and not attached):
##  [1] BiocManager_1.30.4 compiler_3.5.2     bookdown_0.9
##  [4] magrittr_1.5       tools_3.5.2        htmltools_0.3.6
##  [7] yaml_2.2.0         Rcpp_1.0.0         stringi_1.2.4
## [10] rmarkdown_1.11     knitr_1.21         stringr_1.4.0
## [13] xfun_0.4           digest_0.6.18      evaluate_0.13
```