MMRF-CoMMpass data integration and analysis for identifying Prognostic Markers

Marzia Settino[1][0000−0001−5558−9033], Mariamena Arbitrio[2][0000−0001−8387−3664], Francesca Scionti[3][0000−0002−2210−3356], Daniele Caracciolo[3][0000−0001−7870−7565], Maria Teresa Di Martino[3][0000−0002−8205−2706], Pierosandro Tagliaferri[3][0000−0002−1535−2477], Pierfrancesco Tassone[3][0000−0002−8298−6787], and Mario Cannataro[1][0000−0003−1502−2387]

[1] Data Analytics Research Center
Department of Medical and Surgical Sciences
Magna Graecia University,
Catanzaro, Italy.
marzia.settino@studenti.unicz.it,
cannataro@unicz.it

[2] CNR-Institute of Neurological Sciences
UOS of Pharmacology,
Catanzaro, Italy
mariamena.arbitrio@cnr.it

[3] Department of Experimental and Clinical Medicine,
Medical Oncology Unit, Mater Domini Hospital, Magna Graecia University,
Catanzaro, Italy
daniele.caracciolo1@studenti.unicz.it,
{scionti.teresadm,tagliaferri,tassone}@unicz.it

Abstract. Multiple Myeloma (MM) is the second most frequent haematological malignancy in the world although the related pathogenesis remains unclear. The study of how gene expression profiling (GEP) is correlated with patients' survival could be important for understanding the initiation and progression of MM.

In order to aid researchers in identifying new prognostic RNA biomarkers as targets for functional cell-based studies, the use of appropriate bioinformatic tools for integrative analysis is required.

The main contribution of this paper is the development of a set of functionalities, extending TCGAbiolinks package, for downloading and analysing Multiple Myeloma Research Foundation (MMRF) CoMMpass study data available at the NCI's Genomic Data Commons (GDC) Data Portal. In this context, we present further a workflow based on the use of this new functionalities that allows to i) download data; ii) perform and plot the Array Array Intensity correlation matrix; ii) correlate gene expression and Survival Analysis to obtain a Kaplan–Meier survival plot.

Keywords: TCGABiolinks· MMRF-CoMMpass · Multiple Myeloma · TCGA · R · Integrative Data Analysis

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50420-5_42

1 Introduction

Multiple myeloma (MM) is a cancer of plasma cell and it is the second most common blood cancer. Myeloma is a heterogeneous disease with great genetic and epigenetic complexity. Therefore, the identification of patient subgroups defined by molecular profiling and clinical features remains a critical need for a better understanding of disease mechanism, drug response and patient relapse. In this context, the Multiple Myeloma Research Foundation (MMRF-CoMMpass) Study represents the largest genomic data set and the most widely published studies in multiple myeloma.

Transcriptomic studies have largely contributed to reveal multiple myeloma features, distinguishing multiple myeloma subgroups with different clinical and biological patterns. Based on the hypothesis that myeloma invasion would induce changes in gene expression profiles, gene expression profile (GEP) studies constitute a reliable prognostic tool [11, 3].

Various studies have identified gene expression signatures capable of predicting event-free survival and overall survival (OS) in multiple myeloma[6, 1]. In order to aid researchers in identifying new prognostic RNA biomarkers as well as targets for functional cell-based studies, the use of appropriate bioinformatic tools for integrative analysis can offer new opportunities. Among these tools a promising approach is the use of TCGABiolinks package[9, 10, 2].The main contribution of this work is to provide the researchers with a new set of functions extending TCGAbiolinks package that allows to MMRF-CoMMpass database to be investigated. Moreover, a simple workflow for searching, downloading and analyzing RNA-Seq gene level expression dataset from the MMRF-CoMMpass Studies will be described. The same workflow could be in general extended to other MMRF-CoMMpass datasets.

2 Background

Gene expression data from multiple myeloma patients can be retrieved from MMRF-CoMMpass[4] and Gene Expression Omnibus (GEO)[5]. GEO is an international public repository that archives and freely distributes high-throughput gene expression and other functional genomics datasets. The National Cancer Institute (NCI) Genomic Data Commons (GDC) [5] provides the cancer research community with a rich resource for sharing and accessing data across numerous cancer studies and projects for promoting precision medicine in oncology.

The NCI Genomic Data Commons data are made available through the GDC Data Portal[6], a platform for efficiently querying and downloading high quality and complete data. The GDC platform includes data from The Cancer Genome Atlas (TCGA), Therapeutically Applicable Research to Generate Effective Treatments (TARGET) and further studies[7].

[4] https://themmrf.org/we-are-curing-multiple-myeloma/mmrf-commpass-study/

[5] https://www.ncbi.nlm.nih.gov/geo/

[6] https://portal.gdc.cancer.gov/

[7] https://portal.gdc.cancer.gov/projects

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50420-5_42

Recently, many studies are contributing with additional datasets to GDC platform, including the MMRF CoMMpass Study among others [7]. One of the major goals of the GDC is to provide a centralized repository for accessing data from large-scale NCI programs, however it does not make available a comprehensive toolkit for data analyses and interpretation. To fulfil this need, the R/Bioconductor package TCGAbiolinks was developed to allow users to query, download and perform integrative analyses of GDC data [9, 10, 2]. TCGAbiolinks combines methods from computer science and statistics and it includes methods for visualization of results in order to easily perform a complete analysis.

The Cancer Genome Atlas (TCGA) The Cancer Genome Atlas (TCGA) contains data on 33 different cancer types from 11,328 patients and it is the world's largest and richest collection of genomic data. TCGA contains molecular data from multiple types of analysis such as DNA sequencing, RNA sequencing, Copy number, Array-based expression and others. In addition to molecular data, TCGA has well catalogued metadata for each sample such as clinical and sample information.

NCI's Genomic Data Commons (GDC) Data Portal The National Cancer Institute (NCI) Genomic Data Commons (GDC) is a publicly available database that promotes the sharing of genomic and clinical data among researchers and facilitates precision medicine in oncology. At a high level, data in GDC are organized by project (e.g. TCGA, TARGET, MMRF-CoMMpass). Each of these projects contains a variety of molecular data types, including genomics, epigenomics, proteomics, imaging, clinical and others.

Multiple Myeloma Research Foundation (MMRF) CoMMpass The MMRF-CoMMpass Study is a collaborative research effort with the goal of mapping the genomic profile of patients with newly diagnosed active multiple myeloma to clinical outcomes to develop a more complete understanding of patient responses to treatments. MMRF-CoMMpass Study identified many genomic alterations that were not previously found in multiple myeloma as well as providing a prognostic stratification of patients leading to advances in cancer care [8]. Recently the MMRF announced new discoveries into defining myeloma subtypes, identifying novel therapeutic targets for drug discovery and more accurately predicting high-risk disease[8]. The NCI announced in 2016 a collaboration with MMRF to incorporate genomic and clinical data about myeloma into the NCI Genomic Data Commons (GDC) platform.

[8] https://themmrf.org/2018/12/the-mmrf-commpass-study-drives-new-discoveries-in-multiple-myeloma/

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50420-5_42

3 Workflow for downloading and analysing MMRF-CoMMpass Data

TCGAbiolinks is a R/Bioconductor package that combines methods from computer science and statistics to address challenges with data mining and analysis of cancer genomics data stored at GDC Data Portal. More specifically, a guided workflow [10] allows users to query, download, and perform integrative analyses of GDC data. The package provides several methods for analysis (e.g. differential expression analysis, differentially methylated regions, etc.) and methods for visualization (e.g. survival plots, volcano plots and starburst plots, etc.).
TCGAbiolinks was initially conceived to interact with TCGA data through the GDC Data Portal but it can be in principle extended to other GDC datasets if the functions to handle their differences in formats and data availability are properly handled [9]. The GDC API Application Programming Interface (API) provides developers with a programmatic access to GDC functionality.

TCGAbiolinks consists of several functions but in this work we will describe only the main functions used in the workflow described in the Section 3. More specifically:

- **GDCquery** uses GDC API for searching GDC data;
- **GDCprepare** allows to read downloaded data and prepare them into an R object;
- **GDCquery_clinic** allows to download all clinical information related to a specified project in GDC;
- **TCGAanalyze_Preprocessing** performs an Array Array Intensity correlation (AAIC). It defines a square symmetric matrix of spearman correlation among samples;
- **TCGAanalyze_SurvivalKM** performs an univariate Kaplan-Meier (KM) survival analysis (SA) using complete follow up taking one gene a time from a gene list.

The *SummarizedExperiment* [4] object is the default data structure used in TCGAbiolinks for combining genomic data and clinical information. A *SummarizedExperiment* object contains sample information, molecular data and genomic ranges (i.e.gene information). MMRF-CoMMpass presents some differences in formats and data respect to TCGA dataset. For example, the sample ID format in MMRF-CoMMpass is "study-patient-visit-source" (e.g."MMRF-1234-1-BM" means patient 1234, first visit, from bone marrow). Moreover, some fileds in MMRF-CoMMpass *SummarizedExperiment* are lacking or they are named differently respect to TCGA dataset format. In order to fill this gap and to make MMRF-CoMMpass dataset suitable to be handled by previous functions we introduced the following customized functions:

- **MMRF_prepare** adds the sample type information to *SummarizedExperiment* object from *GDCprepare*;
- **MMRF_prepare_clinical** renames the data frame field "submitter_id" of clinical information from *GDCquery_clinic* as the field name found in TCGA dataset (i.e. bcr_patient_barcode);

- **MMRF_prepare_SurvivalKM** makes the MMRF-CoMMpass sample ID format in Gene Expression matrix (dataGE) from *GDCprepare* suitable for using in *TCGAanalyze_SurvivalKM* function.

The following workflow describes the steps for downloading, processing and analyzing MMRF-CoMMpass RNA-Seq gene expression using TCGABiolinks jointly with the new functions before reported.

Search the MMRF-CoMMpass data: GDCquery uses GDC API to search the data for a given project and data category as well as other filters. A valid data category for MMRF-CoMMpass project can be found using *getProjectSummary* function. The results are shown in Table 1.

Table 1. Results of *getProjectSummary* in the case of the MMRF-CoMMpass project.

Data category	File count	Case count
Simple Nucleotide Variation	10918	959
Sequencing Reads	6577	995
Transcriptome Profiling	4295	787

The following listing illustrates the use of GDCquery for searching gene expression level dataset (HTSeq - FPKM) using the ”Trascriptome Profiling” category in the list obtained from *getProjectSummary*.

For simplification purposes just a filtered by barcode subset is downloaded.

```
query.mm.fpkm <- GDCquery(project = "MMRF-COMMPASS",
data.category = "Transcriptome Profiling",
data.type = "Gene Expression Quantification",
workflow.type="HTSeq - FPKM",
barcode = c("MMRF_2473","MMRF_2111","MMRF_2270",
            "MMRF_2238","MMRF_1080","MMRF_2253",
            "MMRF_2119","MMRF_2468", "MMRF_1201",
            "MMRF_2821","MMRF_1957","MMRF_1678"))
```

Listing 1.1. GDCquery function for searching gene expression data in MMRF-CoMMpass. The datset is filtered by barcode.

Download and prepare the MMRF-COMMPASS data: The *GDCdownload* function allows to download and save the data in a local folder to be used in *GDCprepare* function that transforms the downloaded data into a *SummarizedExperiment*. The clinical data (e.g. tumor stage, days to last follow up, treatments) can be obtained using the *GDCquery_clinical* function specifying as input project ”MMRF-COMMPASS”). At this point, *MMRF_prepare* and *MMRF_prepare_clinical* functions allow to make the output of the previous functions suitable for being handled by TCGABiolinks functions.

Analyse MMRF-COMMPASS data: Once the data were downloaded and they are prepared, outliers could be discovered through the use of the function *TCGAanalyze_Preprocessing* which performs an Array Array Intensity correlation (AAIC). The plot in Fig.1 shows an example of heat map of AAIC for MMRF-CoMMpass gene expression data.

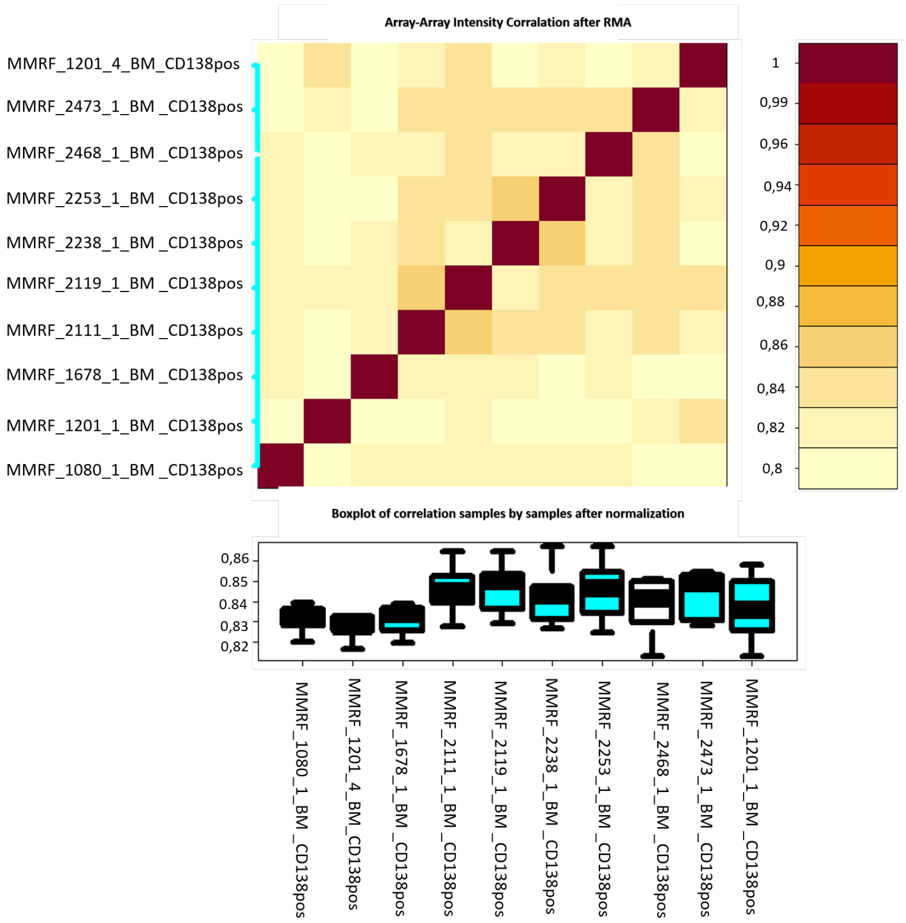

Fig. 1. Example of heat map of the array-array Spearman/Pearson rank correlation coefficients. This plot is useful for detecting outliers.

We used *MMRF_prepare_SurvivalKM* for preparing dataGE from *GDCprepare* to be handled by *TCGAanalyze_SurvivalKM* function.

Finally, we performed a KaplanMeier univariate survival analysis (KM-SA) using *TCGAanalyze_SurvivalKM* function. The resulting plot allows to correlate visually gene expression and Survival Analysis. Two thresholds are defined for each gene expression according its level of mean expression in cancer samples. In this example we used the threshold of intensity of gene expression to divide the samples in 2 groups (High, Low). The Fig. 2 shows the correlation between survival and the most high/low expressed gene.

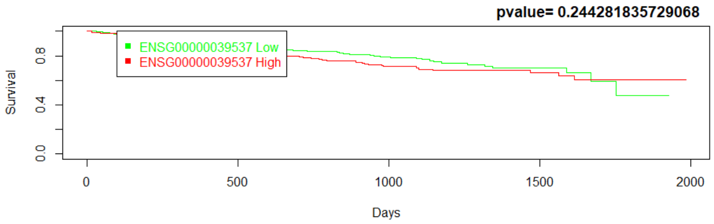

Fig. 2. GDCquery_clinic, MMRF_prepare_clinical, MMRF_prepare_SurvivalKM and TCGAanalyze_SurvivalKM jointly allow to perform an univariate KaplanMeier (KM) survival analysis (SA).

4 Conclusion and Future work

The MMRF-CoMMpass has proven itself to be a leader in scientific innovation as well as in data sharing when it decided to incorporate their data into the GDC platform. The use of appropriate bioinformatic tools for integrative analysis of MMRF-CoMMpass data can offer great opportunities. In order to take this chance, the TCGAbiolinks package represents a useful tool for data integration and analysis of cancer data. For example, TCGAbiolinks offers the possibility to integrate gene expression data from external sources (e.g GEO) obtaining a merged result that can be used for further analysis such as differential expression analysis. The main contribution of this paper is the extension of TCGABiolinks package with new functions to handle MMRF-CoMMpass data available at the NCI's Genomic Data Commons (GDC) Data Portal.This will allow to MM researchers to better exploit MMRF-CoMMpass data. As future work we plan to make available these new functions as package through a public repository and to extend them to allow further analysis of MMRF-CoMMpass data.

References

1. Chng, W.J., Chung, T.H., Kumar, S., Usmani, S., Munshi, N., Avet-Loiseau, H., Goldschmidt, H., Durie, B., Sonneveld, P.: Gene signature combinations improve prognostic stratification of multiple myeloma patients. Leukemia **30**(5), 1071–1078 (05 2016)

2. Colaprico, A., Silva, T.C., Olsen, C., Garofano, L., Cava, C., Garolini, D., Sabedot, T.S., Malta, T.M., Pagnotta, S.M., Castiglioni, I., Ceccarelli, M., Bontempi, G., Noushmehr, H.: Tcgabiolinks: an r/bioconductor package for integrative analysis of tcga data. Nucleic Acids Res **44**(8) (May 2016). https://doi.org/10.1093/nar/gkv1507, https://www.ncbi.nlm.nih.gov/pubmed/26704973
3. Gooding, S., Olechnowicz, S.W.Z., Morris, E.V., Armitage, A.E., Arezes, J., Frost, J., Repapi, E., Edwards, J.R., Ashley, N., Waugh, C., Gray, N., Martinez-Hackert, E., Lim, P.J., Pasricha, S.R., Knowles, H., Mead, A.J., Ramasamy, K., Drakesmith, H., Edwards, C.M.: Transcriptomic profiling of the myeloma bone-lining niche reveals BMP signalling inhibition to improve bone disease. Nat Commun **10**(1), 4533 (10 2019)
4. Huber, W., Carey, V.J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B.S., Bravo, H.C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K.D., Irizarry, R.A., Lawrence, M., Love, M.I., MacDonald, J., Obenchain, V., Ole?, A.K., Pag?s, H., Reyes, A., Shannon, P., Smyth, G.K., Tenenbaum, D., Waldron, L., Morgan, M.: Orchestrating high-throughput genomic analysis with Bioconductor. Nat. Methods **12**(2), 115–121 (Feb 2015)
5. Jensen, M.A., Ferretti, V., Grossman, R.L., Staudt, L.M.: The nci genomic data commons as an engine for precision medicine. Blood **130**(4), 453–459 (07 2017). https://doi.org/10.1182/blood-2017-03-735654, https://www.ncbi.nlm.nih.gov/pubmed/28600341
6. Kuiper, R., Broyl, A., de Knegt, Y., van Vliet, M.H., van Beers, E.H., van der Holt, B., el Jarari, L., Mulligan, G., Gregory, W., Morgan, G., Goldschmidt, H., Lokhorst, H.M., van Duin, M., Sonneveld, P.: A gene expression signature for high-risk multiple myeloma. Leukemia **26**(11), 2406–2413 (Nov 2012)
7. Lee, J.S., Kibbe, W.A., Grossman, R.L.: Data Harmonization for a Molecularly Driven Health System. Cell **174**(5), 1045–1048 (08 2018)
8. Liu, Y., Yu, H., Yoo, S., Lee, E., Lagan?, A., Parekh, S., Schadt, E.E., Wang, L., Zhu, J.: A Network Analysis of Multiple Myeloma Related Gene Signatures. Cancers (Basel) **11**(10) (Sep 2019)
9. Mounir, M., Lucchetta, M., Silva, T.C., Olsen, C., Bontempi, G., Chen, X., Noushmehr, H., Colaprico, A., Papaleo, E.: New functionalities in the tcgabiolinks package for the study and integration of cancer data from gdc and gtex. PLoS Comput Biol **15**(3) (03 2019). https://doi.org/10.1371/journal.pcbi.1006701, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6420023/
10. Silva, T.C., Colaprico, A., Olsen, C., D'Angelo, F., Bontempi, G., Ceccarelli, M., Noushmehr, H.: TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages. F1000Res **5**, 1542 (2016)
11. Szalat, R., Avet-Loiseau, H., Munshi, N.C.: Gene Expression Profiles in Myeloma: Ready for the Real World? Clin. Cancer Res. **22**(22), 5434–5442 (Nov 2016)