TRAINING: A practical approach for analysis and interpretation of NGS cancer data

Teachers: Antonio Colaprico and Catharina Olsen
Academic supervisor: Gianluca Bontempi

Please feel free to contact me by skype if necessary before the workshop
skype account: antonio.colaprico

TCGAbiolinks module introduces practical, real-world bioinformatics applications, going beyond the TCGAbiolinksGUI graphical interface using of TCGAbiolinks R bioconductor package published in http://nar.oxfordjournals.org/content/early/2015/12/23/nar.gkv1507.full
The module is focused on the theme of data analysis: when a large-scale experiment is performed, and bioinformatics analysis is required, how is it done? The module is organised into three units.

In the "*TCGA Data Access and Integration*" unit, students will learn how to distinguish databases and integrate data from different datatypes from The Cancer Genome Atlas (TCGA), including microRNA expression, gene expression, copy number, mutation, methylation, protein expression clinical's data.

In the "*Genomics and NGS*" unit, students will learn practical analysis of microarray and next-generation sequencing (NGS) data. Students will learn how to map sequencing data to genomes in a variety of problem settings, how to analyse differential expression studies from whole-genome experiments, and more.
In this section, the student will be learning how to analyse RNA-seq count data, using TCGAbiolinksGUI. This will include reading the data into shiny, quality control and performing differential expression analysis and gene set testing, with a focus on the edgeR analysis workflow.

The last, unit is "*Integrative Analysis*". Students will learn how bringing disparate data types together can add enormous power to analyses. Cases will include pathway analysis of expression data, and analysis of dna methylation.

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Read me first

- You will be mainly working with a graphical interface of TCGAbiolinks within Rstudio. A short introduction will be given at the beginning of the workshop. Practical experience will be obtained by applying the given information on different given cancer data sets.

- For this workshop you require:

➔ This manual
➔ Look previous presentation of TCGAbiolinks
https://www.youtube.com/watch?v=eP9C3kKA8eo
➔ Read TCGAbiolinks paper:

http://nar.oxfordjournals.org/content/early/2015/12/23/nar.gkv1507.full

https://www.researchgate.net/publication/287996967_TCGAbiolinks_An_RBioconductor_package_for_integrative_analysis_of_TCGA_data

➔ TCGAbiolinksGUI vignette / Tutorial (`TCGAbiolinks GUI_vignette.pdf`attached)
➔ In order to know the web address of the application please connect to **http://litpc45.ulb.ac.be**

You can ask questions during the workshop.

Citation

Please cite TCGAbiolinks package:

- “TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data.” Nucleic acids research (2015): gkv1507. (Colaprico, Antonio and Silva, Tiago C. and Olsen, Catharina and Garofano, Luciano and Cava, Claudia and Garolini, Davide and Sabedot, Thais S. and Malta, Tathiane M. and Pagnotta, Stefano M. and Castiglioni, Isabella and Ceccarelli, Michele and Bontempi, Gianluca and Noushmehr, Houtan 2016)

Related publications to this package:

- “TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages”. F1000Research 10.12688/f1000research.8923.1 (Silva, TC and Colaprico, A and Olsen, C and D’Angelo, F and Bontempi, G and Ceccarelli, M and Noushmehr, H 2016)

Also, if you have used ELMER analysis please cite:

- Yao, L., Shen, H., Laird, P. W., Farnham, P. J., & Berman, B. P. “Inferring regulatory element landscapes and transcription factor networks from cancer methylomes.” Genome Biol 16 (2015): 105.
- Yao, Lijing, Benjamin P. Berman, and Peggy J. Farnham. “Demystifying the secret mission of enhancers: linking distal regulatory elements to target genes.” Critical reviews in biochemistry and molecular biology 50.6 (2015): 550-573.

Content

How to install TCGAbiolinks

```
source("https://bioconductor.org/biocLite.R")
biocLite("TCGAbiolinks")
```

How to install TCGAbiolinks last / beta version (only expert)

```
devtools::install_github(repo="BioinformaticsFMRP/TCGAbiolinks")
```

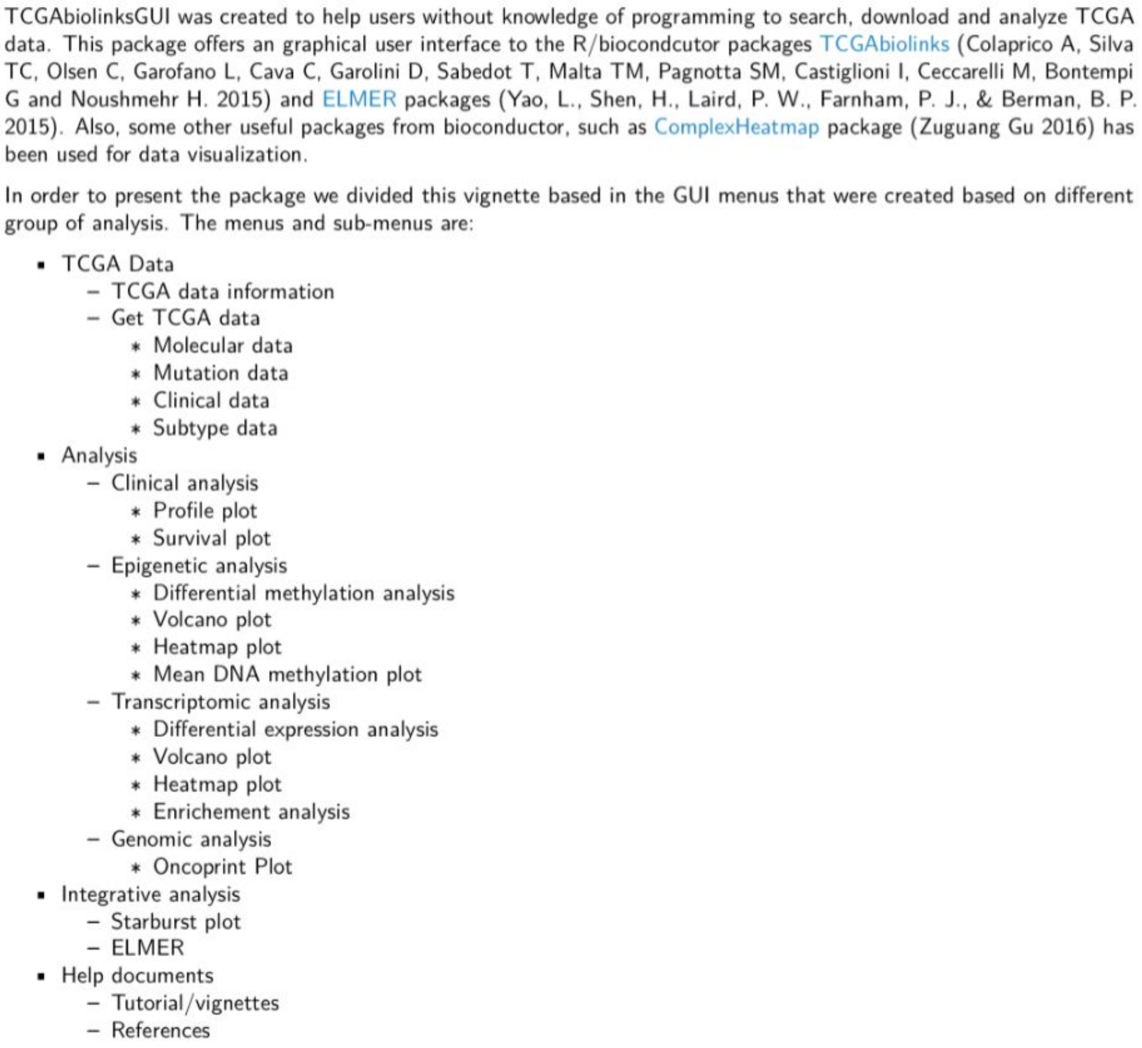
How to install TCGAbiolinksGUI

```
devtools::install_github(repo="BioinformaticsFMRP/TCGAbiolinksGUI")
```

2. Introduction to TCGA's data and TCGAbiolinks 30 min

2.1 Presentation: https://www.youtube.com/watch?v=eP9C3kKA8eo

2.2 Overview of TCGAbiolinksGUI 15min



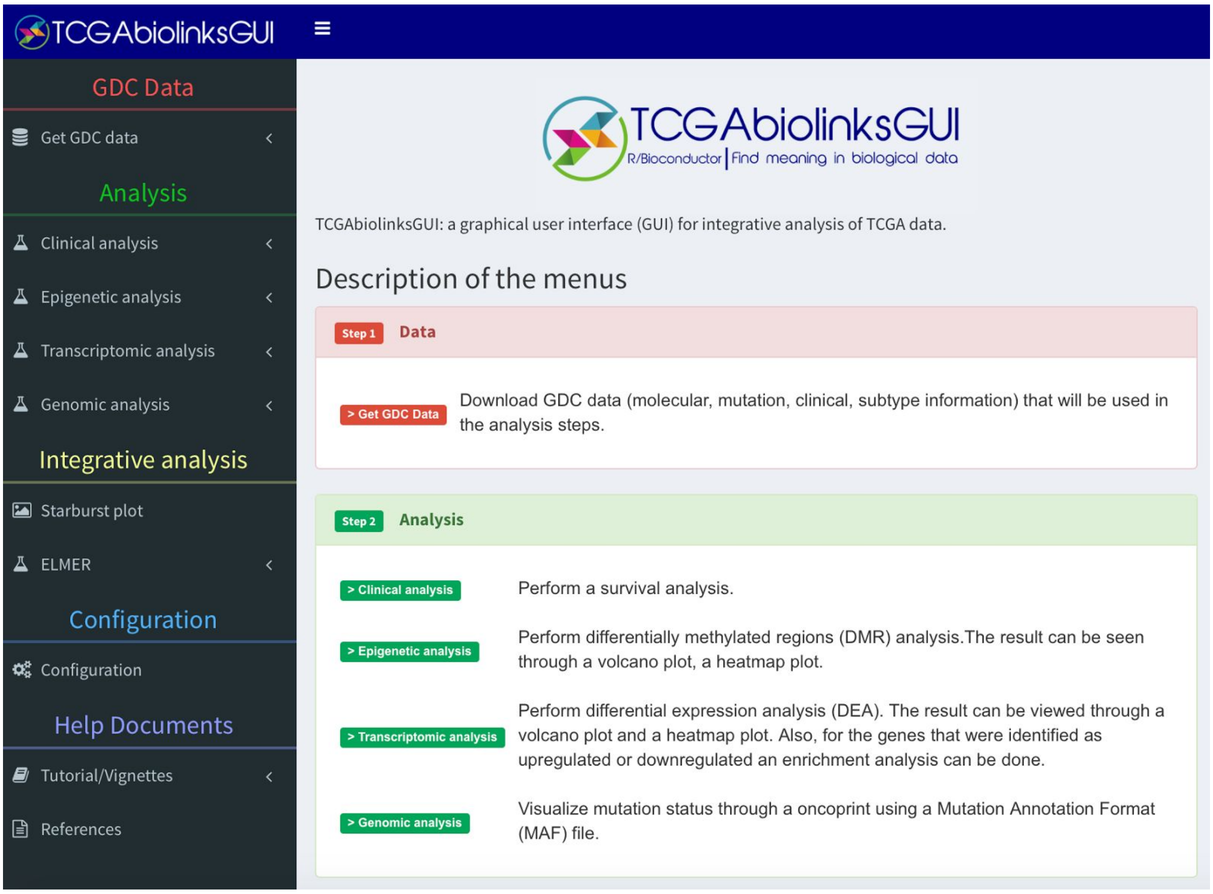
TCGAbiolinksGUI was created to help users without knowledge of programming to search, download and analyze TCGA data. This package offers an graphical user interface to the R/biocondcutor packages TCGAbiolinks (Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, Sabedot T, Malta TM, Pagnotta SM, Castiglioni I, Ceccarelli M, Bontempi G and Noushmehr H. 2015) and ELMER packages (Yao, L., Shen, H., Laird, P. W., Farnham, P. J., & Berman, B. P. 2015). Also, some other useful packages from bioconductor, such as ComplexHeatmap package (Zuguang Gu 2016) has been used for data visualization.

In order to present the package we divided this vignette based in the GUI menus that were created based on different group of analysis. The menus and sub-menus are:

- TCGA Data
  - TCGA data information
  - Get TCGA data
    - Molecular data
    - Mutation data
    - Clinical data
    - Subtype data
- Analysis
  - Clinical analysis
    - Profile plot
    - Survival plot
  - Epigenetic analysis
    - Differential methylation analysis
    - Volcano plot
    - Heatmap plot
    - Mean DNA methylation plot
  - Transcriptomic analysis
    - Differential expression analysis
    - Volcano plot
    - Heatmap plot
    - Enrichement analysis
  - Genomic analysis
    - Oncoprint Plot
- Integrative analysis
  - Starburst plot
  - ELMER
- Help documents
  - Tutorial/vignettes
  - References

1. TCGAbiolinksGUI interface

After loading TCGAbiolinksGUI you will see a web-page showing:



TCGAbiolinksGUI: a graphical user interface (GUI) for integrative analysis of TCGA data.

Description of the menus

Step 1 Data

> Get GDC Data — Download GDC data (molecular, mutation, clinical, subtype information) that will be used in the analysis steps.

Step 2 Analysis

> Clinical analysis — Perform a survival analysis.

> Epigenetic analysis — Perform differentially methylated regions (DMR) analysis.The result can be seen through a volcano plot, a heatmap plot.

> Transcriptomic analysis — Perform differential expression analysis (DEA). The result can be viewed through a volcano plot and a heatmap plot. Also, for the genes that were identified as upregulated or downregulated an enrichment analysis can be done.

> Genomic analysis — Visualize mutation status through a oncoprint using a Mutation Annotation Format (MAF) file.

QUESTIONS that will be addressed

Q1. Which data can be downloaded from TCGAbiolinksGUI?

Q2. Which analysis can be performed by TCGAbiolinksGUI?

Q3. How to run a survival analysis with Kaplan-Meier plot?

Q4. How to produce an oncoprint plot ?

Q5. How to run a DEA differentially expression analysis using gene expression's data?

Q6. How to produce a volcano plot using differentially expressed genes?

Q7. How to produce heatmap plot using differentially expressed genes?

Q8. How to perform an enrichment analysis?

Q9. How to plot a pathway graph enriched by a list of genes?

Q10. How to produce an oncoprint plot?

Q11. How to run a DMR differentially methylation region analysis?

Q12. How to produce a mean DNA methylation plot?

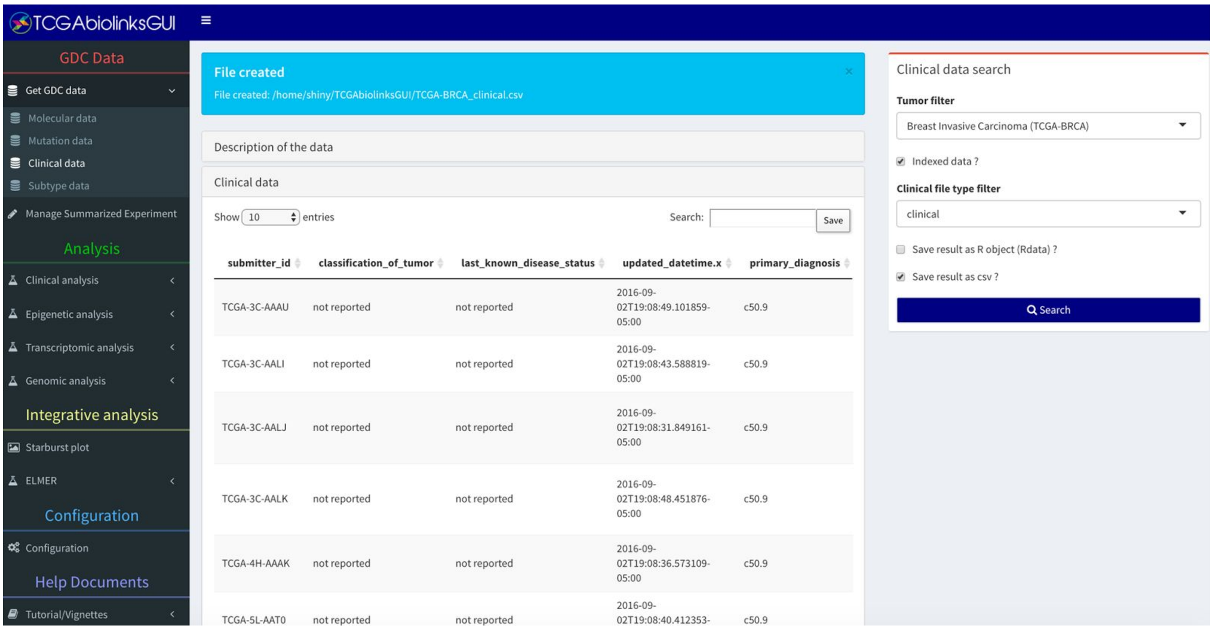
Q13. How to perform an integrative analysis using gene expression data and methylation data and showing the results in a starburst plot?

DATA that will be used can be found in folder ./Data_Workshop

- D1_TCGA_BRCA_clinical.csv
- D2_TCGA_CHOL_maf.csv
- D3_short_GDC_TCGA_BRCA_Illumina HiSeq.rda
- D4_GDC_TCGA_BRCA_Illumina Human Methylation 27.rda

How to generate DATA that will be used?

D1_TCGA_BRCA_clinical.csv



File created

File created: /home/shiny/TCGAbiolinksGUI/TCGA-BRCA_clinical.csv

Show 10 entries

Search:

submitter_id	classification_of_tumor	last_known_disease_status	updated_datetime.x	primary_diagnosis
TCGA-3C-AAAU	not reported	not reported	2016-09-02T19:08:49.101859-05:00	c50.9
TCGA-3C-AALI	not reported	not reported	2016-09-02T19:08:43.588819-05:00	c50.9
TCGA-3C-AALJ	not reported	not reported	2016-09-02T19:08:31.849161-05:00	c50.9
TCGA-3C-AALK	not reported	not reported	2016-09-02T19:08:48.451876-05:00	c50.9
TCGA-4H-AAAK	not reported	not reported	2016-09-02T19:08:36.573109-05:00	c50.9
TCGA-5L-AAT0	not reported	not reported	2016-09-02T19:08:40.412353-	c50.9

Clinical data search

Tumor filter: Breast Invasive Carcinoma (TCGA-BRCA)

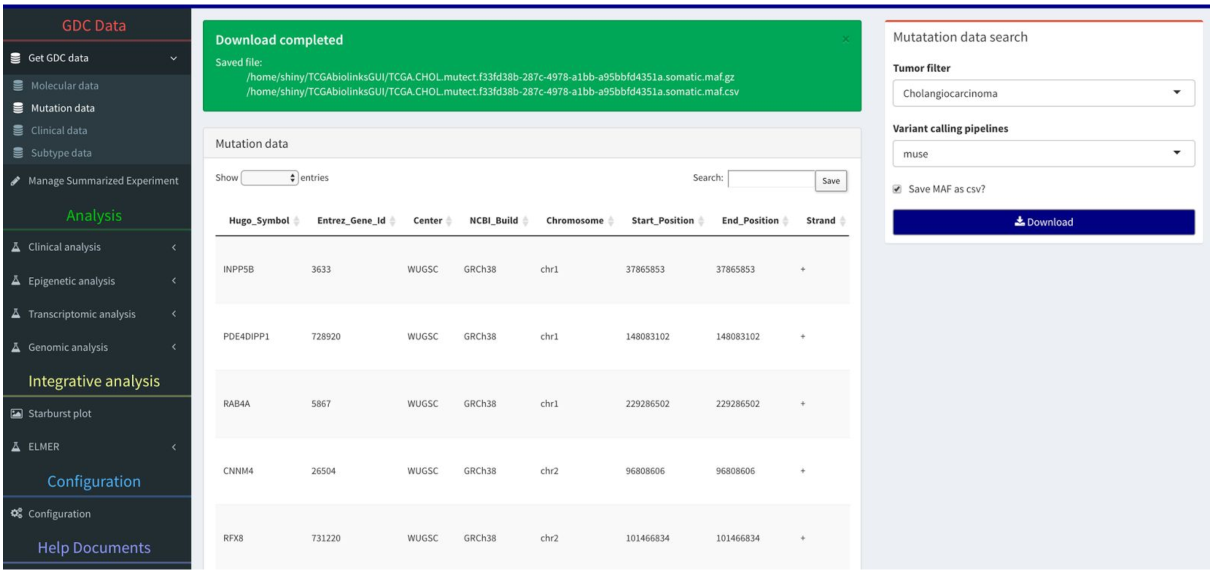
☑ Indexed data ?

Clinical file type filter: clinical

□ Save result as R object (Rdata) ?

☑ Save result as csv ?

D2_TCGA_CHOL_maf.csv



Download completed

Saved file:

```
/home/shiny/TCGAbiolinksGUI/TCGA.CHOL.mutect.f33fd38b-287c-4978-a1bb-a95bbfd4351a.somatic.maf.gz
/home/shiny/TCGAbiolinksGUI/TCGA.CHOL.mutect.f33fd38b-287c-4978-a1bb-a95bbfd4351a.somatic.maf.csv
```

Mutation data

Show entries Search: Save

Hugo_Symbol	Entrez_Gene_Id	Center	NCBI_Build	Chromosome	Start_Position	End_Position	Strand
INPP5B	3633	WUGSC	GRCh38	chr1	37865853	37865853	+
PDE4DIPP1	728920	WUGSC	GRCh38	chr1	148083102	148083102	+
RAB4A	5867	WUGSC	GRCh38	chr1	229286502	229286502	+
CNNM4	26504	WUGSC	GRCh38	chr2	96808606	96808606	+
RFX8	731220	WUGSC	GRCh38	chr2	101466834	101466834	+

Mutatation data search

Tumor filter: Cholangiocarcinoma

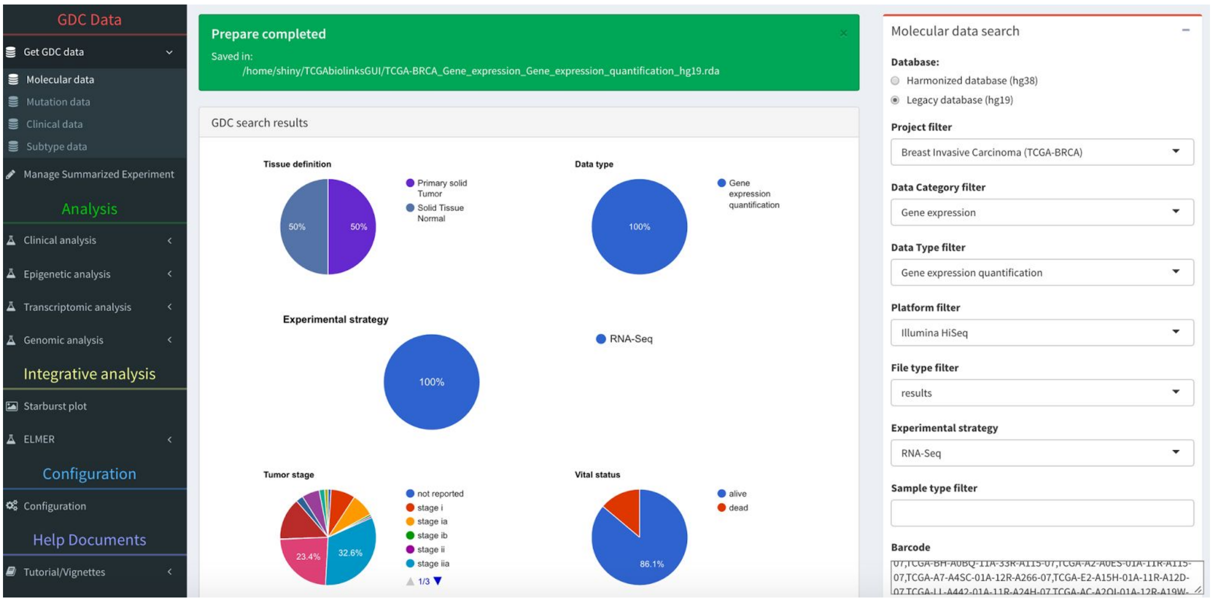
Variant calling pipelines: muse

☑ Save MAF as csv?

D3_short_GDC_TCGA_BRCA_Illumina HiSeq.rda

Barcode:

TCGA-E2-A105-01A-11R-A10J-07,TCGA-E9-A1R7-11A-42R-A14M-07,TCGA-AN-A0XU-01A-11R-A109-07,TCGA-B6-A0IN-01A-11R-A034-07,TCGA-BH-A0BQ-11A-33R-A115-07,TCGA-A2-A0ES-01A-11R-A115-07,TCGA-A7-A4SC-01A-12R-A266-07,TCGA-E2-A15H-01A-11R-A12D-07,TCGA-LL-A442-01A-11R-A24H-07,TCGA-AC-A2QI-01A-12R-A19W-07,TCGA-BH-A0HW-01A-11R-A034-07,TCGA-E2-A1BC-11A-32R-A12P-07,TCGA-E9-A1RI-11A-41R-A169-07,TCGA-E9-A1ND-11A-43R-A144-07,TCGA-BH-A0BC-11A-22R-A089-07,TCGA-AC-A23H-11A-12R-A157-07,TCGA-BH-A18P-11A-43R-A12D-07,TCGA-BH-A18K-11A-13R-A12D-07,TCGA-BH-A1FG-11B-12R-A13Q-07,TCGA-AC-A3OD-01A-11R-A21T-07



Prepare completed

Saved in:
/home/shiny/TCGAbiolinksGUI/TCGA-BRCA_Gene_expression_Gene_expression_quantification_hg19.rda

Molecular data search

Database:
○ Harmonized database (hg38)
● Legacy database (hg19)

Project filter: Breast Invasive Carcinoma (TCGA-BRCA)

Data Category filter: Gene expression

Data Type filter: Gene expression quantification

Platform filter: Illumina HiSeq

File type filter: results

Experimental strategy: RNA-Seq

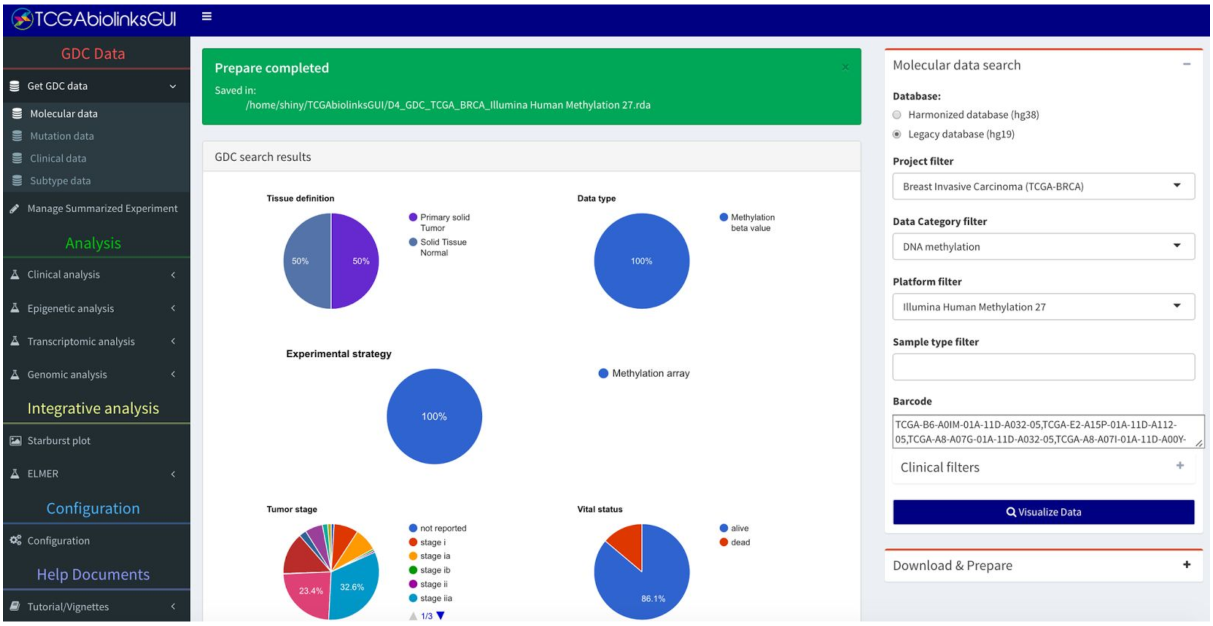
Sample type filter:

Barcode:
07,TCGA-BH-A0BQ-11A-33R-A115-07,TCGA-A2-A0ES-01A-11R-A115-
07,TCGA-A7-A4SC-01A-12R-A266-07,TCGA-E2-A15H-01A-11R-A12D-
07,TCGA-LL-A442-01A-11R-A24H-07,TCGA-AC-A2QI-01A-12R-A19W-

D4_GDC_TCGA_BRCA_Illumina Human Methylation 27.rda

Barcode:

TCGA-B6-A0IM-01A-11D-A032-05,TCGA-E2-A15P-01A-11D-A112-05,TCGA-A8-A07G-01A-11D-A032-05,TCGA-A8-A07I-01A-11D-A00Y-05,TCGA-A8-A096-01A-11D-A00Y-05,TCGA-A2-A0CU-01A-12D-A032-05,TCGA-BH-A0DL-01A-11D-A112-05,TCGA-A8-A09A-01A-11D-A00Y-05,TCGA-E2-A15O-01A-11D-A112-05,TCGA-C8-A12K-01A-21D-A112-05,TCGA-BH-A0B7-11A-34D-A112-05,TCGA-BH-A18Q-11A-34D-A12E-05,TCGA-BH-A0DL-11A-13D-A112-05,TCGA-BH-A18S-11A-43D-A12E-05,TCGA-BH-A18F-11A-22D-A12E-05,TCGA-E2-A15L-11A-31D-A12E-05,TCGA-BH-A0BO-11A-11D-A12E-05,TCGA-BH-A18L-11A-42D-A12E-05,TCGA-BH-A0BL-11A-12D-A112-05,TCGA-BH-A18J-11A-31D-A12E-05



Prepare completed

Saved in:
/home/shiny/TCGAbiolinksGUI/D4_GDC_TCGA_BRCA_Illumina Human Methylation 27.rda

Molecular data search

Database:
○ Harmonized database (hg38)
● Legacy database (hg19)

Project filter: Breast Invasive Carcinoma (TCGA-BRCA)

Data Category filter: DNA methylation

Platform filter: Illumina Human Methylation 27

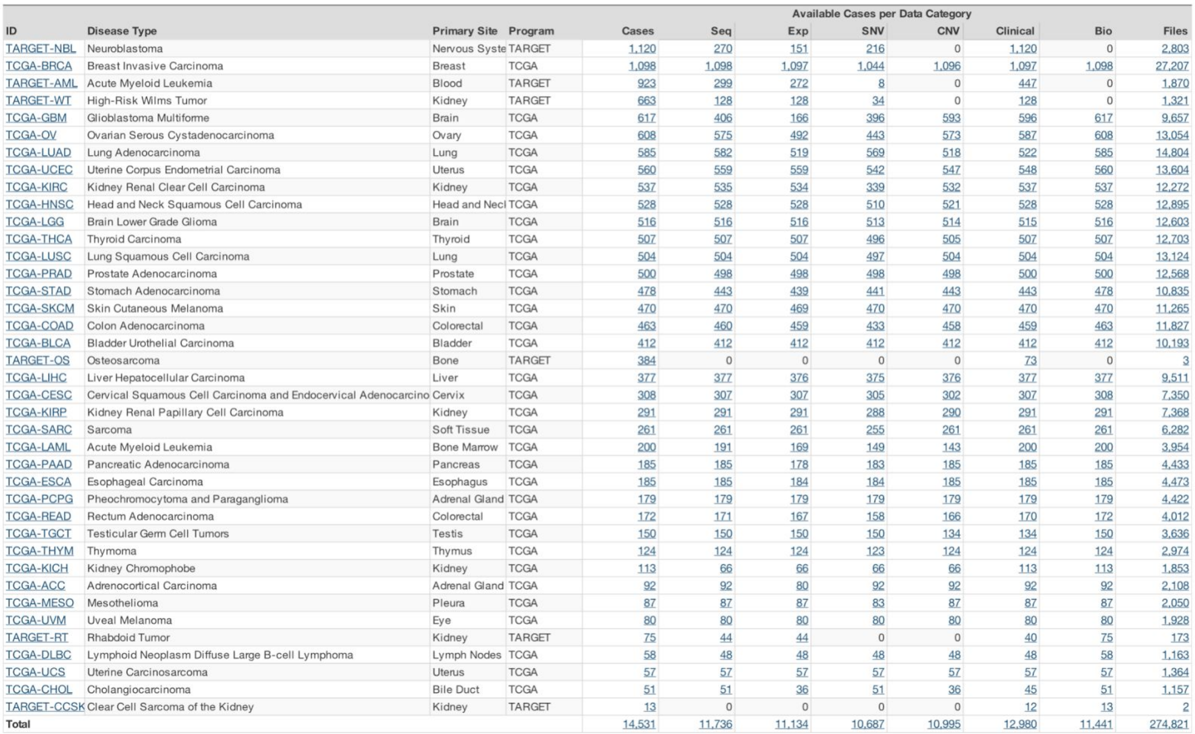
Sample type filter:

Barcode: TCGA-B6-A0IM-01A-11D-A032-05,TCGA-E2-A15P-01A-11D-A112-05,TCGA-A8-A07G-01A-11D-A032-05,TCGA-A8-A07I-01A-11D-A00Y-

1. GDC data introduction and overview

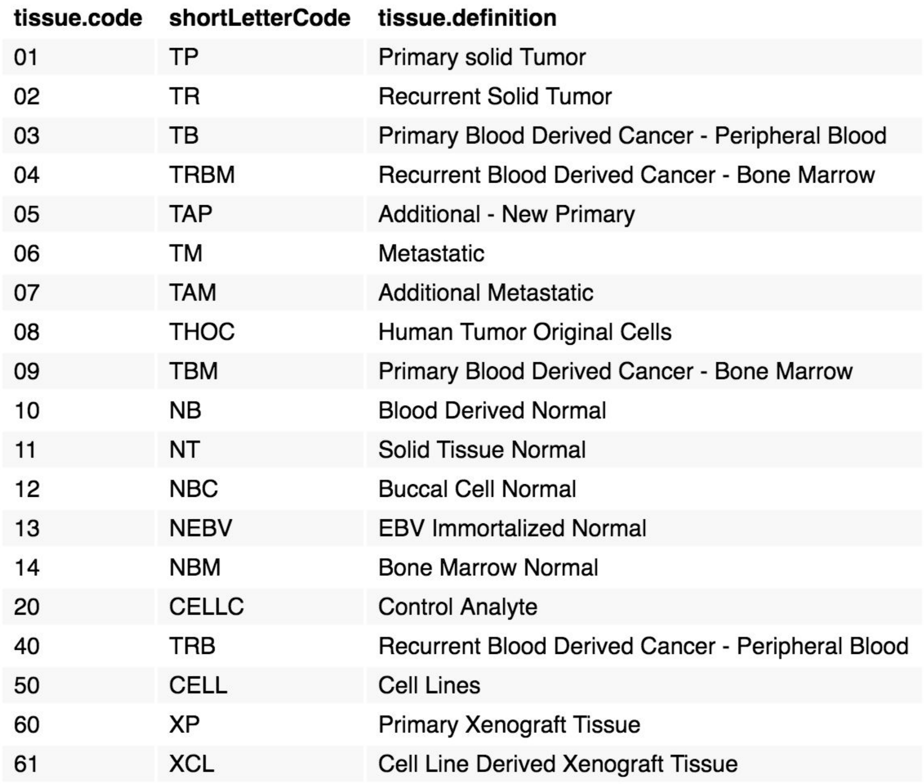
1.1. (A) Cancer Projects

From https://gdc-portal.nci.nih.gov/projects/t



					Available Cases per Data Category						
ID	Disease Type	Primary Site	Program	Cases	Seq	Exp	SNV	CNV	Clinical	Bio	Files
TARGET-NBL	Neuroblastoma	Nervous Syste	TARGET	1,120	270	151	216	0	1,120	0	2,803
TCGA-BRCA	Breast Invasive Carcinoma	Breast	TCGA	1,098	1,098	1,097	1,044	1,096	1,097	1,098	27,207
TARGET-AML	Acute Myeloid Leukemia	Blood	TARGET	923	299	272	8	0	447	0	1,870
TARGET-WT	High-Risk Wilms Tumor	Kidney	TARGET	663	128	128	34	0	128	0	1,321
TCGA-GBM	Glioblastoma Multiforme	Brain	TCGA	617	406	166	396	593	596	617	9,657
TCGA-OV	Ovarian Serous Cystadenocarcinoma	Ovary	TCGA	608	575	492	443	573	587	608	13,054
TCGA-LUAD	Lung Adenocarcinoma	Lung	TCGA	585	582	519	569	518	522	585	14,804
TCGA-UCEC	Uterine Corpus Endometrial Carcinoma	Uterus	TCGA	560	559	559	542	547	548	560	13,604
TCGA-KIRC	Kidney Renal Clear Cell Carcinoma	Kidney	TCGA	537	535	534	339	532	537	537	12,272
TCGA-HNSC	Head and Neck Squamous Cell Carcinoma	Head and Necl	TCGA	528	528	528	510	521	528	528	12,895
TCGA-LGG	Brain Lower Grade Glioma	Brain	TCGA	516	516	516	513	514	515	516	12,603
TCGA-THCA	Thyroid Carcinoma	Thyroid	TCGA	507	507	507	496	505	507	507	12,703
TCGA-LUSC	Lung Squamous Cell Carcinoma	Lung	TCGA	504	504	504	497	504	504	504	13,124
TCGA-PRAD	Prostate Adenocarcinoma	Prostate	TCGA	500	498	498	498	498	500	500	12,568
TCGA-STAD	Stomach Adenocarcinoma	Stomach	TCGA	478	443	439	441	443	443	478	10,835
TCGA-SKCM	Skin Cutaneous Melanoma	Skin	TCGA	470	470	469	470	470	470	470	11,265
TCGA-COAD	Colon Adenocarcinoma	Colorectal	TCGA	463	460	459	433	458	459	463	11,827
TCGA-BLCA	Bladder Urothelial Carcinoma	Bladder	TCGA	412	412	412	412	412	412	412	10,193
TARGET-OS	Osteosarcoma	Bone	TARGET	384	0	0	0	0	73	0	3
TCGA-LIHC	Liver Hepatocellular Carcinoma	Liver	TCGA	377	377	376	375	376	377	377	9,511
TCGA-CESC	Cervical Squamous Cell Carcinoma and Endocervical Adenocarcino	Cervix	TCGA	308	307	307	305	302	307	308	7,350
TCGA-KIRP	Kidney Renal Papillary Cell Carcinoma	Kidney	TCGA	291	291	291	288	290	291	291	7,368
TCGA-SARC	Sarcoma	Soft Tissue	TCGA	261	261	261	255	261	261	261	6,282
TCGA-LAML	Acute Myeloid Leukemia	Bone Marrow	TCGA	200	191	169	149	143	200	200	3,954
TCGA-PAAD	Pancreatic Adenocarcinoma	Pancreas	TCGA	185	185	178	183	185	185	185	4,433
TCGA-ESCA	Esophageal Carcinoma	Esophagus	TCGA	185	185	184	184	185	185	185	4,473
TCGA-PCPG	Pheochromocytoma and Paraganglioma	Adrenal Gland	TCGA	179	179	179	179	179	179	179	4,422
TCGA-READ	Rectum Adenocarcinoma	Colorectal	TCGA	172	171	167	158	166	170	172	4,012
TCGA-TGCT	Testicular Germ Cell Tumors	Testis	TCGA	150	150	150	150	134	134	150	3,636
TCGA-THYM	Thymoma	Thymus	TCGA	124	124	124	123	124	124	124	2,974
TCGA-KICH	Kidney Chromophobe	Kidney	TCGA	113	66	66	66	66	113	113	1,853
TCGA-ACC	Adrenocortical Carcinoma	Adrenal Gland	TCGA	92	92	80	92	92	92	92	2,108
TCGA-MESO	Mesothelioma	Pleura	TCGA	87	87	87	83	87	87	87	2,050
TCGA-UVM	Uveal Melanoma	Eye	TCGA	80	80	80	80	80	80	80	1,928
TARGET-RT	Rhabdoid Tumor	Kidney	TARGET	75	44	44	0	0	40	75	173
TCGA-DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	Lymph Nodes	TCGA	58	48	48	48	48	48	58	1,163
TCGA-UCS	Uterine Carcinosarcoma	Uterus	TCGA	57	57	57	57	57	57	57	1,364
TCGA-CHOL	Cholangiocarcinoma	Bile Duct	TCGA	51	51	36	51	36	45	51	1,157
TARGET-CCSK	Clear Cell Sarcoma of the Kidney	Kidney	TARGET	13	0	0	0	0	12	13	2
Total				14,531	11,736	11,134	10,687	10,995	12,980	11,441	274,821

1.2. (B) Sample types



tissue.code	shortLetterCode	tissue.definition
01	TP	Primary solid Tumor
02	TR	Recurrent Solid Tumor
03	TB	Primary Blood Derived Cancer - Peripheral Blood
04	TRBM	Recurrent Blood Derived Cancer - Bone Marrow
05	TAP	Additional - New Primary
06	TM	Metastatic
07	TAM	Additional Metastatic
08	THOC	Human Tumor Original Cells
09	TBM	Primary Blood Derived Cancer - Bone Marrow
10	NB	Blood Derived Normal
11	NT	Solid Tissue Normal
12	NBC	Buccal Cell Normal
13	NEBV	EBV Immortalized Normal
14	NBM	Bone Marrow Normal
20	CELLC	Control Analyte
40	TRB	Recurrent Blood Derived Cancer - Peripheral Blood
50	CELL	Cell Lines
60	XP	Primary Xenograft Tissue
61	XCL	Cell Line Derived Xenograft Tissue

1.3. (C) Molecular data

Molecular data can be grouped in gene expression, copy number, methylation, microRNA.

1.4. (D) Mutation data

The GDC provides access to DNA sequence data and generates associated Variant Calling Format (VCF) and Mutation Annotation Format (MAF) files that identify somatic mutations such as point mutations, missense mutations, nonsense mutations, and insertions and deletions (indels) of nucleotides in the DNA.

1.5. (E) Clinical data

GDC clinical data represent many categories of information, including vital status at time of report, disease-specific diagnostic information, and initial treatment regimens. Some but not all disease studies have additional clinical follow up information for some or all participants.

1.6. (F) Subtype data

The Cancer Genome Atlas (TCGA) Research Network has reported integrated genome-wide studies of various diseases. We have added some of the subtypes defined by these report in our package. The ACC(Cancer Genome Atlas Research Network and others 2016), BRCA (Cancer Genome Atlas Research Network and others 2012c), COAD (Cancer Genome Atlas Research Network and others 2012b), GBM (Ceccarelli, Michele and Barthel, Floris P and Malta, Tathiane M and Sabedot, Thais S and Salama, Sofie R and Murray, Bradley A and Morozova, Olena and Newton, Yulia and Radenbaugh, Amie and Pagnotta, Stefano M and others 2016), HNSC (Cancer Genome Atlas Research Network and others 2015a), KICH (Davis, Caleb F and Ricketts, Christopher J and Wang, Min and Yang, Lixing and Cherniack, Andrew D and Shen, Hui and Buhay, Christian and Kang, Hyojin and Kim, Sang Cheol and Fahey, Catherine C and others 2014), KIRC(Cancer Genome Atlas Research Network and others 2013a), KIRP (Linehan, W Marston and Spellman, Paul T and Ricketts, Christopher J and Creighton, Chad J and Fei, Suzanne S and Davis, Caleb and Wheeler, David A and Murray, Bradley A and Schmidt, Laura and Vocke, Cathy D and others 2016), LGG (Ceccarelli, Michele and Barthel, Floris P and Malta, Tathiane M and Sabedot, Thais S and Salama, Sofie R and Murray, Bradley A and Morozova, Olena and Newton, Yulia and Radenbaugh, Amie and Pagnotta, Stefano M and others 2016), LUAD (Cancer Genome Atlas Research Network and others 2014b), LUSC(Cancer Genome Atlas Research Network and others 2012a), PRAD(Cancer Genome Atlas Research Network and others 2015c), READ (Cancer Genome Atlas Research Network and others 2012b), SKCM (Cancer Genome Atlas Research Network and others 2015b), STAD (Cancer Genome Atlas Research Network and others 2014a), THCA (Cancer Genome Atlas Research Network and others 2014c), UCEC (Cancer Genome Atlas Research Network and others 2013b) tumors have data added.

ULB UNIVERSITÉ LIBRE DE BRUXELLES

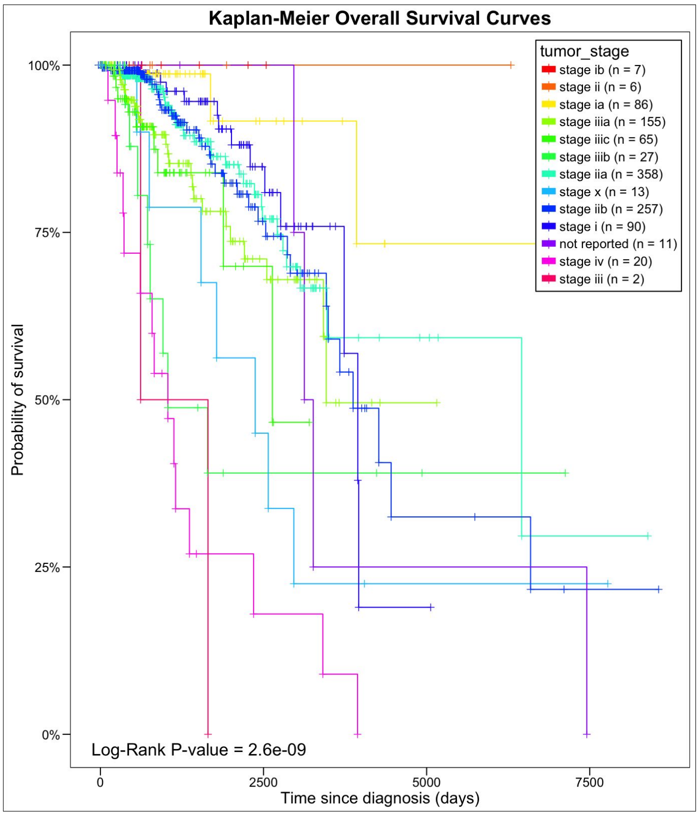
2. Clinical Analysis

2.1. (A) Survival Plot

The Kaplan–Meier estimator also known as the product limit estimator, is a non-parametric statistic used to estimate the survival function from lifetime data. In medical research, it is often used to measure the fraction of patients living for a certain amount of time after treatment.

Data selected file (csv or rda) → D1_TCGA_BRCA_clinical.csv

Group column → tumor stage



We can observe that samples with stages I or II have higher probability to survive compared to samples with stages III or IV.

ULB
UNIVERSITÉ
LIBRE
DE BRUXELLES

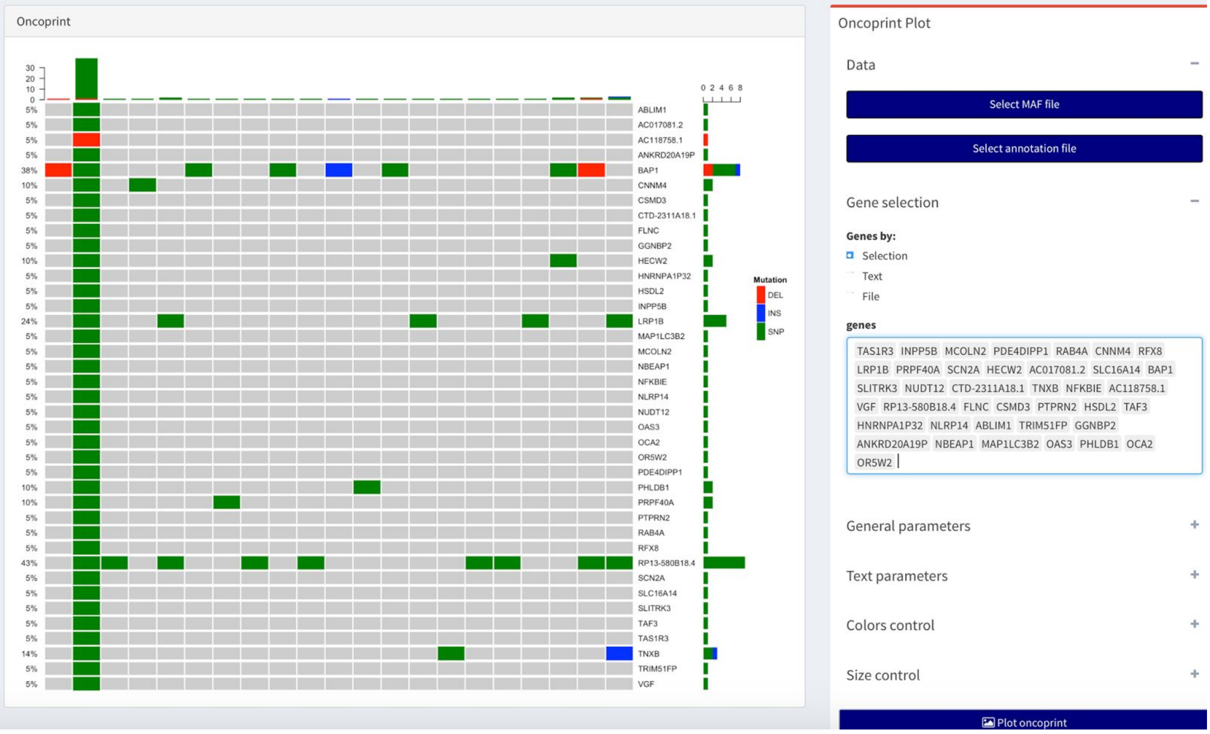
3. Genomic Analysis

3.1. (A) Oncoprint Plot

OncoPrint is a way to visualize multiple genomic alteration events by heatmap.

Select MAF file → **D2_TCGA_CHOL_maf.csv**

Genes by selection → Select the genes as shown in the following screenshot



Genes by:

● Selection
○ Text
○ File

genes

TAS1R3 INPP5B MCOLN2 PDE4DIPP1 RAB4A CNNM4 RFX8 LRP1B PRPF40A SCN2A HECW2 AC017081.2 SLC16A14 BAP1 SLITRK3 NUDT12 CTD-2311A18.1 TNXB NFKBIE AC118758.1 VGF RP13-580B18.4 FLNC CSMD3 PTPRN2 HSDL2 TAF3 HNRNPA1P32 NLRP14 ABLIM1 TRIM51FP GGNBP2 ANKRD20A19P NBEAP1 MAP1LC3B2 OAS3 PHLDB1 OCA2 OR5W2

4. Transcriptomics Analysis

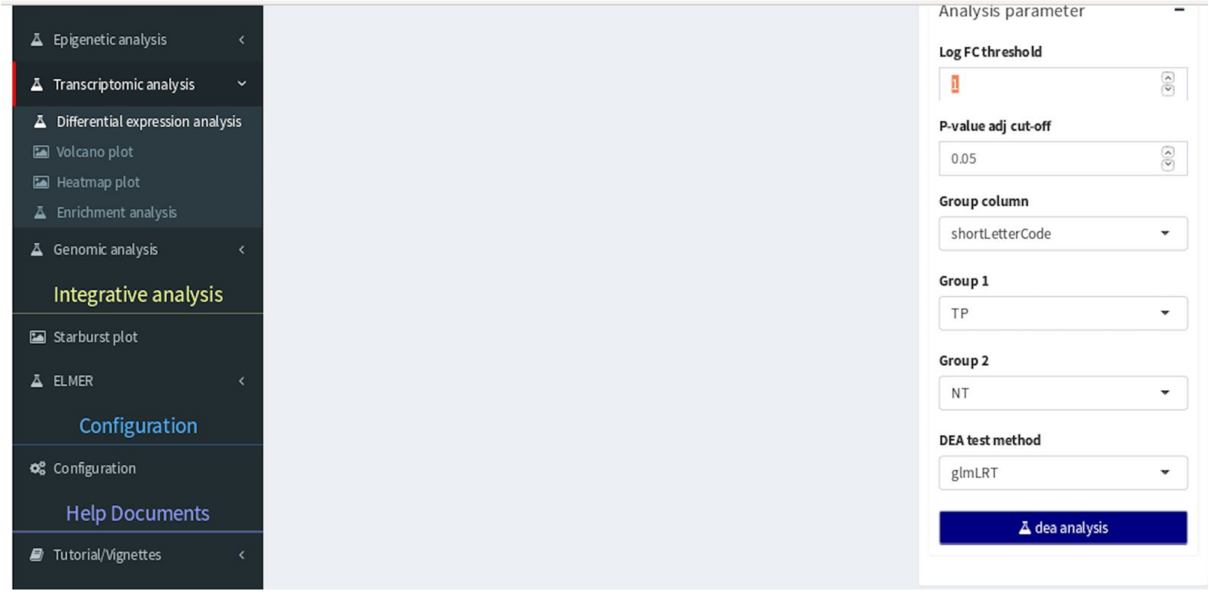
4.1. (A) Differential expression analysis

DEA is a way to find differentially expressed genes between normal and tumor samples

Select SummarizedExperiment Data file → `D3_short_GDC_TCGA_BRCA_Illumina HiSeq.rda`
Pre-analysis options → Select both normalization of genes and quantile filter of genes
Threshold selected as mean for filtering"--> 0.25
Analysis parameter → LogFC threshold = 1, P-value adj cut-off = 0.05
Group column →shortLetterCode Group 1 --> TP
Group2 → NT
DEAtest method → glmLRT

After selecting all above parameters you can click on '*dea analysis*'.

Panel A



Analysis parameter

Log FC threshold: 1

P-value adj cut-off: 0.05

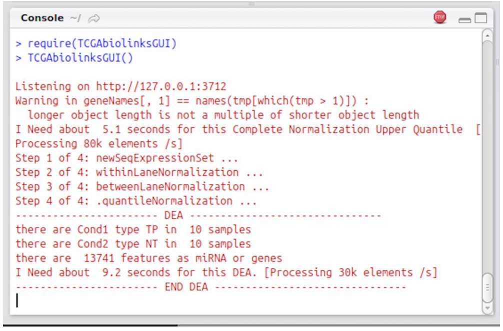
Group column: shortLetterCode

Group 1: TP

Group 2: NT

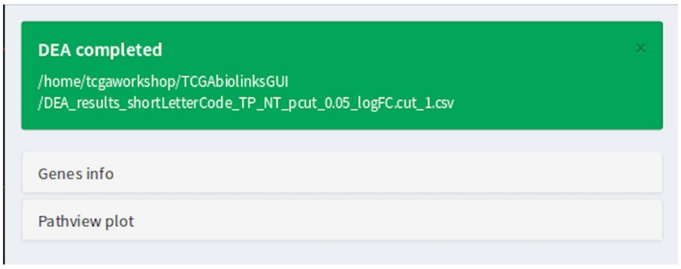
DEA test method: glmLRT

Panel B **Panel C**



```
> require(TCGAbiolinksGUI)
> TCGAbiolinksGUI()

Listening on http://127.0.0.1:3712
Warning in geneNames[, 1] == names(tmp[which(tmp > 1)]) :
  longer object length is not a multiple of shorter object length
I Need about  5.1 seconds for this Complete Normalization Upper Quantile  [
Processing 80k elements /s]
Step 1 of 4: newSeqExpressionSet ...
Step 2 of 4: withinLaneNormalization ...
Step 3 of 4: betweenLaneNormalization ...
Step 4 of 4: .quantileNormalization ...
---------------------- DEA ------------------------------
there are Cond1 type TP in  10 samples
there are Cond2 type NT in  10 samples
there are  13741 features as miRNA or genes
I Need about  9.2 seconds for this DEA. [Processing 30k elements /s]
---------------------- END DEA -------------------------------
```



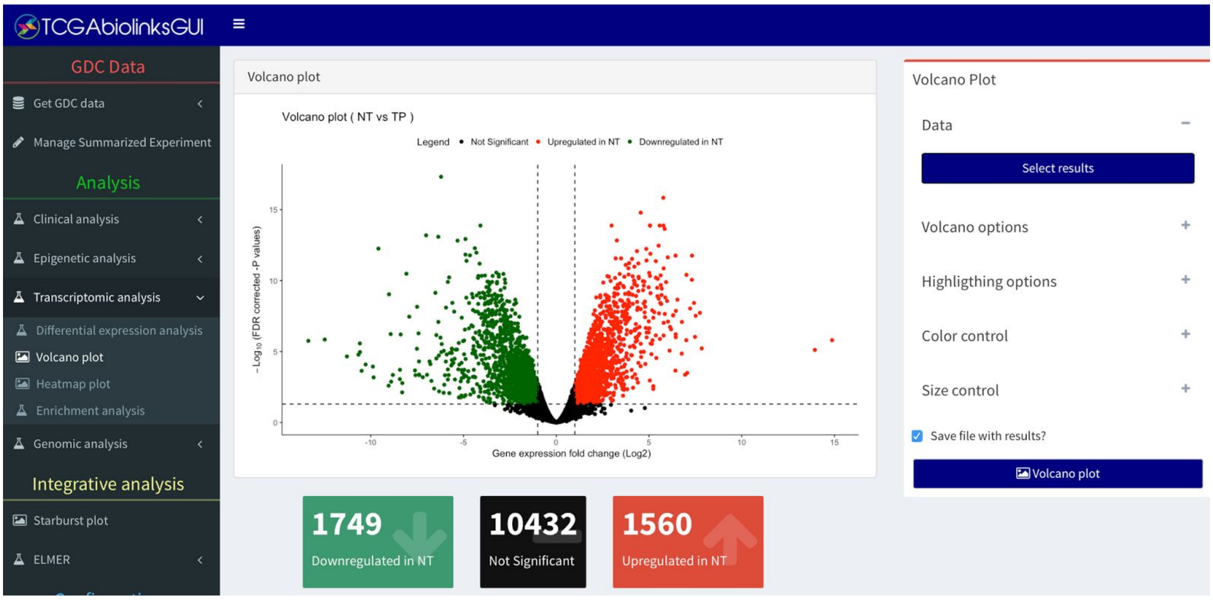
DEA completed
/home/tcgaworkshop/TCGAbiolinksGUI
/DEA_results_shortLetterCode_TP_NT_pcut_0.05_logFC.cut_1.csv

4.2. (B) Volcano plot

In statistics, a volcano plot is a type of scatter-plot that is used to quickly identify changes in large data sets composed of replicate data[1]. It plots significance versus fold-change on the y and x axes, respectively.

Select results file → `DEA_results_shortLetterCode_TP_NT_pcut_0.05_logFC.cut_1.csv`
Save file with results → Yes

After selecting all above parameters you can click on '***volcano plot***'.

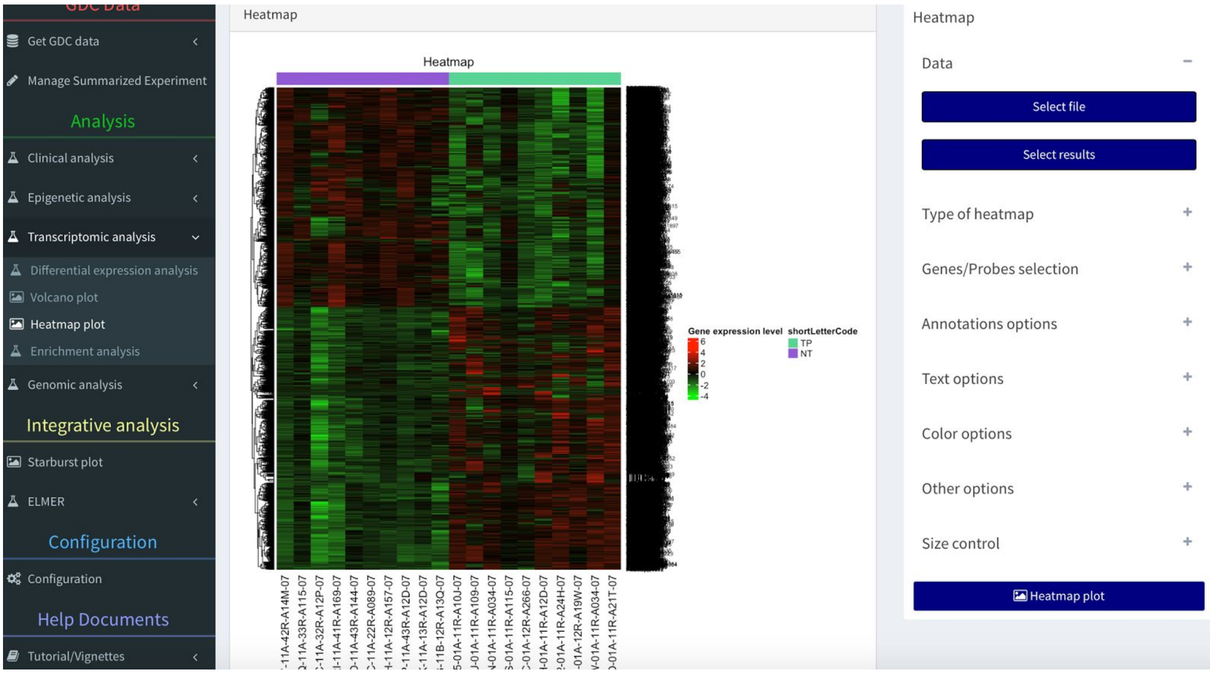


The volcano plot shows 3309 DEGs with |logFC| >=1 and FDR < 0.05.

ULB UNIVERSITÉ LIBRE DE BRUXELLES

4.3. (C) Heatmap plot

Select file → D3_short_GDC_TCGA_BRCA_Illumina HiSeq.rda
Select results → Select file output from section 3.2 Volcano plot, the same as previous step→ `DEA_results_shortLetterCode_TP_NT_pcut_0.05_logFC.cut_1.csv`
Annotation options:
Column annotations →shortLetterCode
Sort by columns →Yes
Color options / Set colors → Yes
Other options:
Scale data → row
Take the log2(matrix +1) → Yes
Cluster rows → Yes
Cluster columns → Yes
Show rownames → Yes
Show colnames → Yes
Plot height px → 600

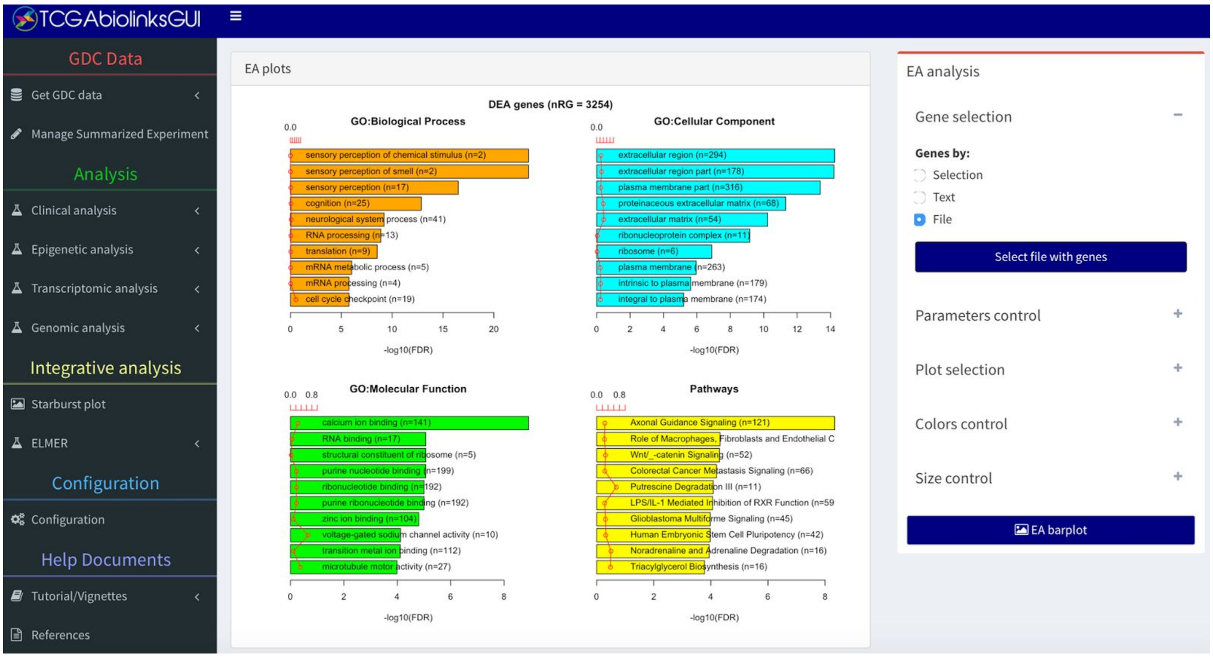


4.4. (D) Enrichment analysis

Gene Selection → Genes by file → Select file with genes → `DEA_results_shortLetterCode_TP_NT_pcut_0.05_logFC.cut_1_filtered.csv`

Plot height px → 600

After selecting all above parameters you can click on '***EA barplot***'.



TCGAbiolinksGUI
GDC Data
Get GDC data
Manage Summarized Experiment
Analysis
Clinical analysis
Epigenetic analysis
Transcriptomic analysis
Genomic analysis
Integrative analysis
Starburst plot
ELMER
Configuration
Configuration
Help Documents
Tutorial/Vignettes
References
EA plots
DEA genes (nRG = 3254)
GO:Biological Process
0.0
sensory perception of chemical stimulus (n=2)
sensory perception of smell (n=2)
sensory perception (n=17)
cognition (n=25)
neurological system process (n=41)
RNA processing (n=13)
translation (n=9)
mRNA metabolic process (n=5)
mRNA processing (n=4)
cell cycle checkpoint (n=19)
0
5
10
15
20
-log10(FDR)
GO:Cellular Component
0.0
extracellular region (n=294)
extracellular region part (n=178)
plasma membrane part (n=316)
proteinaceous extracellular matrix (n=68)
extracellular matrix (n=54)
ribonucleoprotein complex (n=11)
ribosome (n=6)
plasma membrane (n=263)
intrinsic to plasma membrane (n=179)
integral to plasma membrane (n=174)
0
2
4
6
8
10
12
14
-log10(FDR)
GO:Molecular Function
0.0
0.8
calcium ion binding (n=141)
RNA binding (n=17)
structural constituent of ribosome (n=5)
purine nucleotide binding (n=199)
ribonucleotide binding (n=192)
purine ribonucleotide binding (n=192)
zinc ion binding (n=104)
voltage-gated sodium channel activity (n=10)
transition metal ion binding (n=112)
microtubule motor activity (n=27)
0
2
4
6
8
-log10(FDR)
Pathways
0.0
0.8
Axonal Guidance Signaling (n=121)
Role of Macrophages, Fibroblasts and Endothelial C
Wnt/_-catenin Signaling (n=52)
Colorectal Cancer Metastasis Signaling (n=66)
Putrescine Degradation III (n=11)
LPS/IL-1 Mediated Inhibition of RXR Function (n=59
Glioblastoma Multiforme Signaling (n=45)
Human Embryonic Stem Cell Pluripotency (n=42)
Noradrenaline and Adrenaline Degradation (n=16)
Triacylglycerol Biosynthesis (n=16)
0
2
4
6
8
-log10(FDR)
EA analysis
Gene selection
Select file with genes
Parameters control
Plot selection
Colors control
Size control
EA barplot

Genes by:
○ Selection
○ Text
● File

The figure shows canonical pathways significantly overrepresented (enriched) by the DEGs (differentially expressed genes). The most statistically significant canonical pathways identified in DEGs list are listed according to their p value corrected FDR (-Log) (colored bars) and the ratio of list genes found in each pathway over the total number of genes in that pathway (Ratio, red line).

4.5. (E) Pathview plot

Analysis → Differential expression analysis → Pathway graph

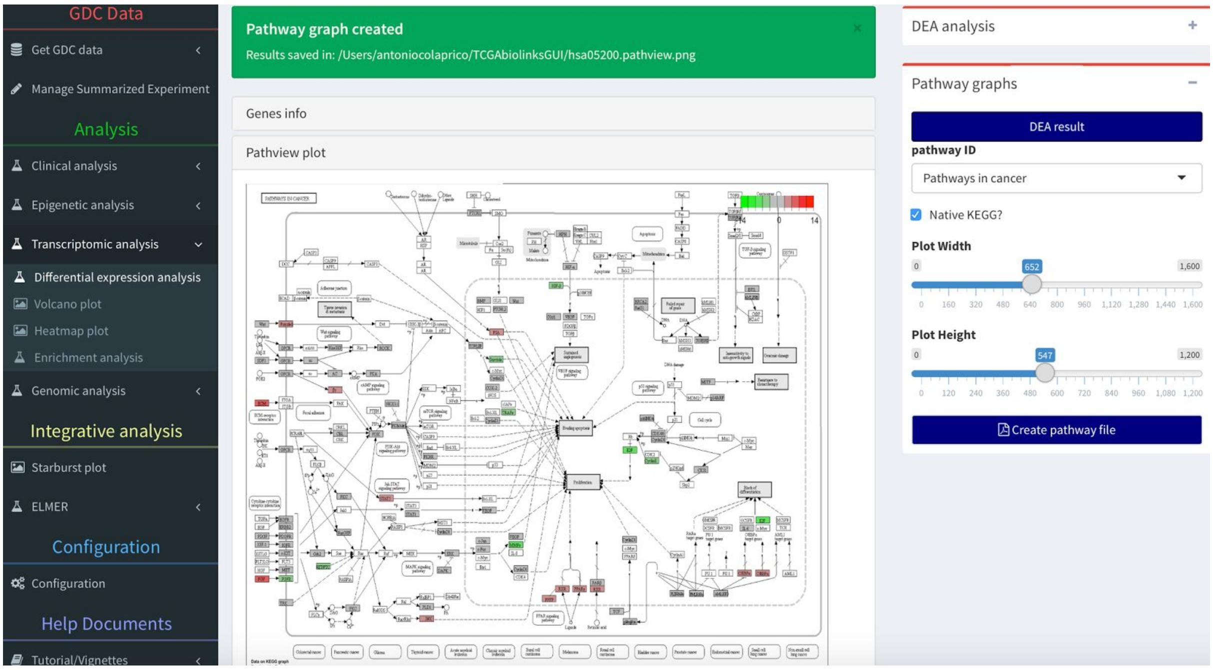
DEA results → `DEA_results_shortLetterCode_TP_NT_pcut_0.05_logFC .cut_1_filtered.csv`

Pathway ID → Pathways in cancer

Plot width px → 650

Plot height px → 550

After selecting all above parameters you can click on '***Create pathway file***'.



Pathway graph created

Results saved in: /Users/antoniocolaprico/TCGAbiolinksGUI/hsa05200.pathview.png

pathway ID: Pathways in cancer

☑ Native KEGG?

Plot Width: 652

Plot Height: 547

5. Epigenetic Analysis

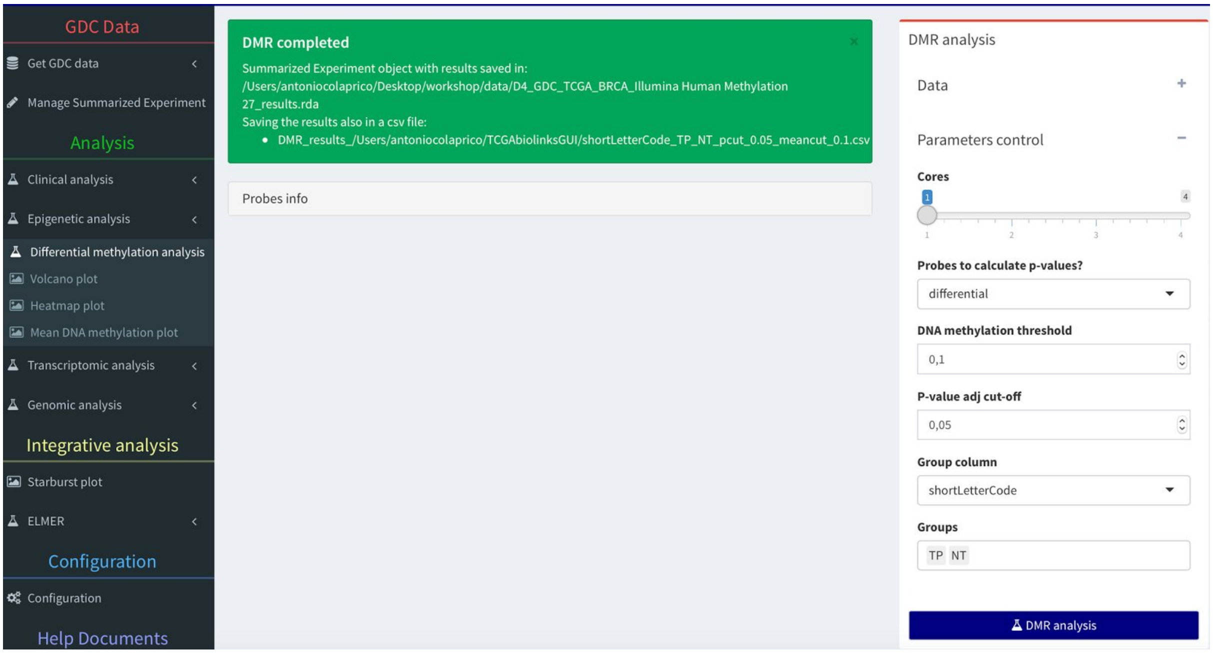
5.1. (A) Differential methylation analysis

We will search for differentially methylated CpG sites.
In order to find these regions we use the beta-values (methylation values ranging from 0.0 to 1.0) to compare two groups.

Select SummarizedExperiment Data file → `D4_GDC_TCGA_BRCA_Illumina Human Methylation 27.rda`

DNA methylation threshold → 0.1
P-value adj cut-off → 0.05
Group column →shortLetterCode
Groups --> TP, NT

After selecting all above parameters you can click on ‘***DMR analysis***’.



DMR completed

Summarized Experiment object with results saved in:
/Users/antoniocolaprico/Desktop/workshop/data/D4_GDC_TCGA_BRCA_Illumina Human Methylation 27_results.rda
Saving the results also in a csv file:

- DMR_results_/Users/antoniocolaprico/TCGAbiolinksGUI/shortLetterCode_TP_NT_pcut_0.05_meancut_0.1.csv

Cores: 1

Probes to calculate p-values?: differential

DNA methylation threshold: 0,1

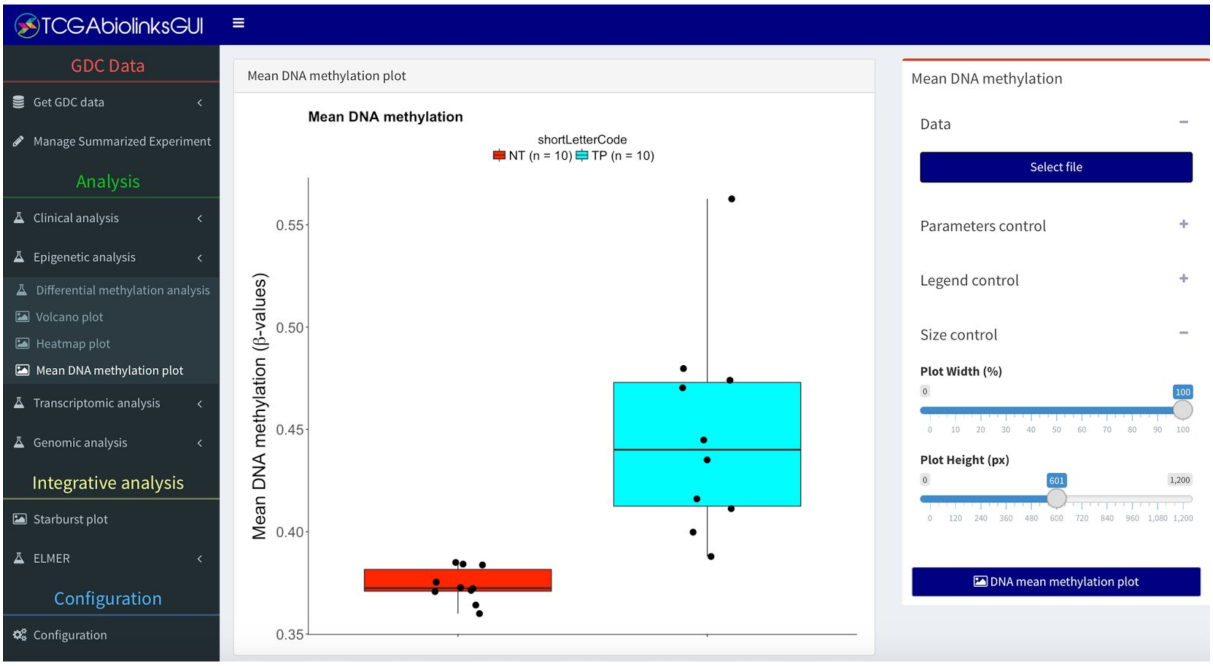
P-value adj cut-off: 0,05

Group column: shortLetterCode

Groups: TP NT

5.2. (B) Mean DNA methylation

Analysis → Epigenetic analysis → Mean DNA methylation plot
Data → Select file → `D4_GDC_TCGA_BRCA_Illumina Human Methylatio n 27_results.rda`
Parameters control → Group column → shortLetterCode
Plot width → 100%
Plot height px → 600
After selecting all above parameters you can click on ‘***DNA mean methylation plot***’.



Plot Width (%): 100

Plot Height (px): 601

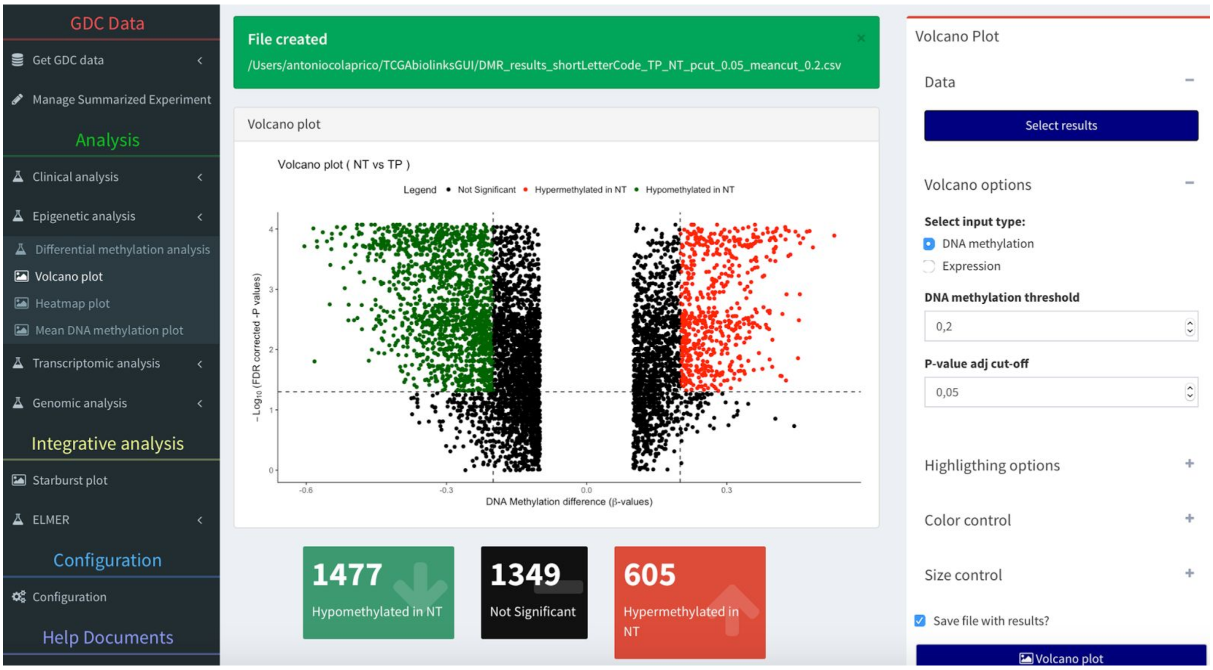
The figure shows 10 normal samples (NT) and 10 cancer samples (TP).

5.3. (C) Volcano plot

In statistics, a volcano plot is a type of scatter-plot that is used to quickly identify changes in large data sets composed of replicate data.[1] It plots significance versus fold-change on the y and x axes, respectively.

Select results file → `DMR_results_shortLetterCode_TP_NT_pcut_0.05_meancut_0.2.csv`
DNA methylation threshold → 0.2
P-value adj cut-off → 0.05
Save file with results → Yes

After selecting all above parameters you can click on '***volcano plot***'.



File created
/Users/antoniocolaprico/TCGAbiolinksGUI/DMR_results_shortLetterCode_TP_NT_pcut_0.05_meancut_0.2.csv

Select input type:
● DNA methylation
○ Expression

DNA methylation threshold: 0,2

P-value adj cut-off: 0,05

☑ Save file with results?

1477 Hypomethylated in NT

1349 Not Significant

605 Hypermethylated in NT

The volcano plot shows 2082 DMRs with |DNA methylation threshold| > 0.2 and FDR < 0.05.

6. Integrative Analysis

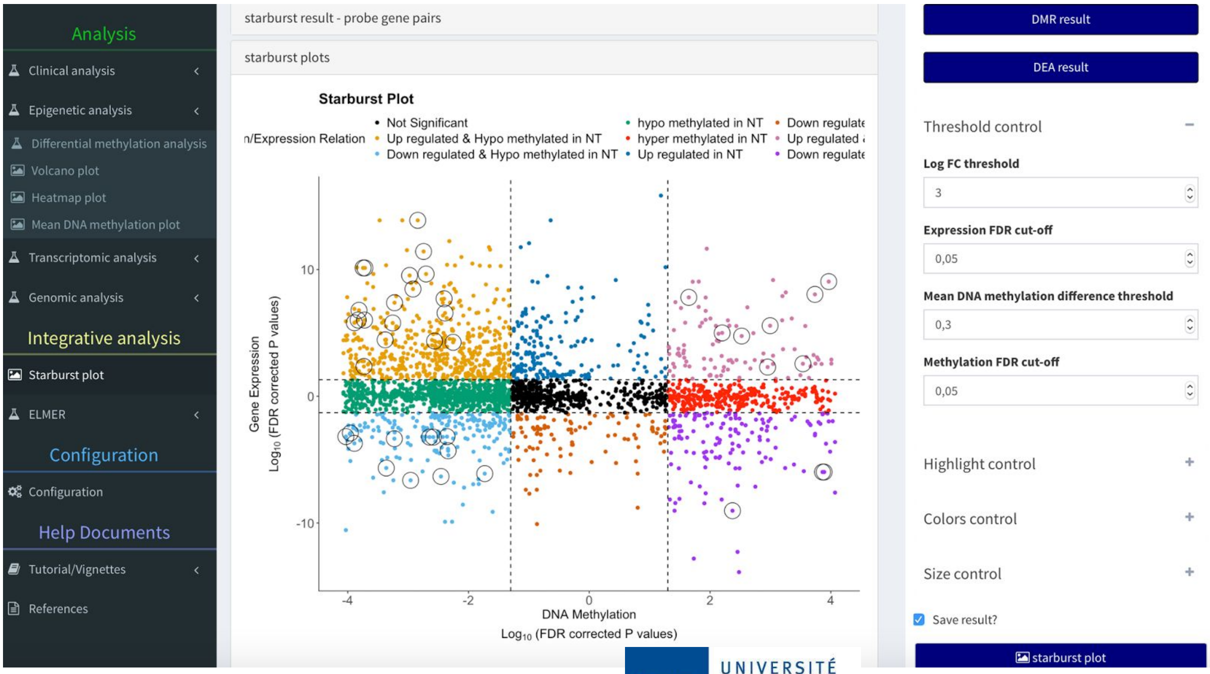
6.1. (A) Starburst plot

The starburst plot is proposed to combine information from two volcano plots, and is applied for a study of DNA methylation and gene expression. It first introduced in 2010 (Noushmehr, H., Weisenberger, D.J., Diefes, K., Phillips, H.S., Pujara, K., Berman, B.P., Pan, F., Pelloski, C.E., Sulman, E.P., Bhat, K.P. et al. 2010).
The function creates Starburst plot for comparison of DNA methylation and gene expression. The log10 (FDR-corrected P value) for DNA methylation is plotted in the x axis, and for gene expression in the y axis, for each gene. The black dashed line shows the FDR-adjusted P value of 0.01.

DMR result → `DMR_results_shortLetterCode_TP_NT_pcut_0.05_meancut_0.1.csv`
DEA result → `DEA_results_shortLetterCode_TP_NT_pcut_0.05_logFC.cut_1.csv`

LogFC threshold → 3
Expression FDR cut.off → 0.05
Mean DNA methylation difference threshold → 0.3
Methylation FDR cut-off → 0.05
Save result → Yes

After selecting all above parameters you can click on '***starburst plot***'.



Log FC threshold: 3
Expression FDR cut-off: 0,05
Mean DNA methylation difference threshold: 0,3
Methylation FDR cut-off: 0,05
☑ Save result?