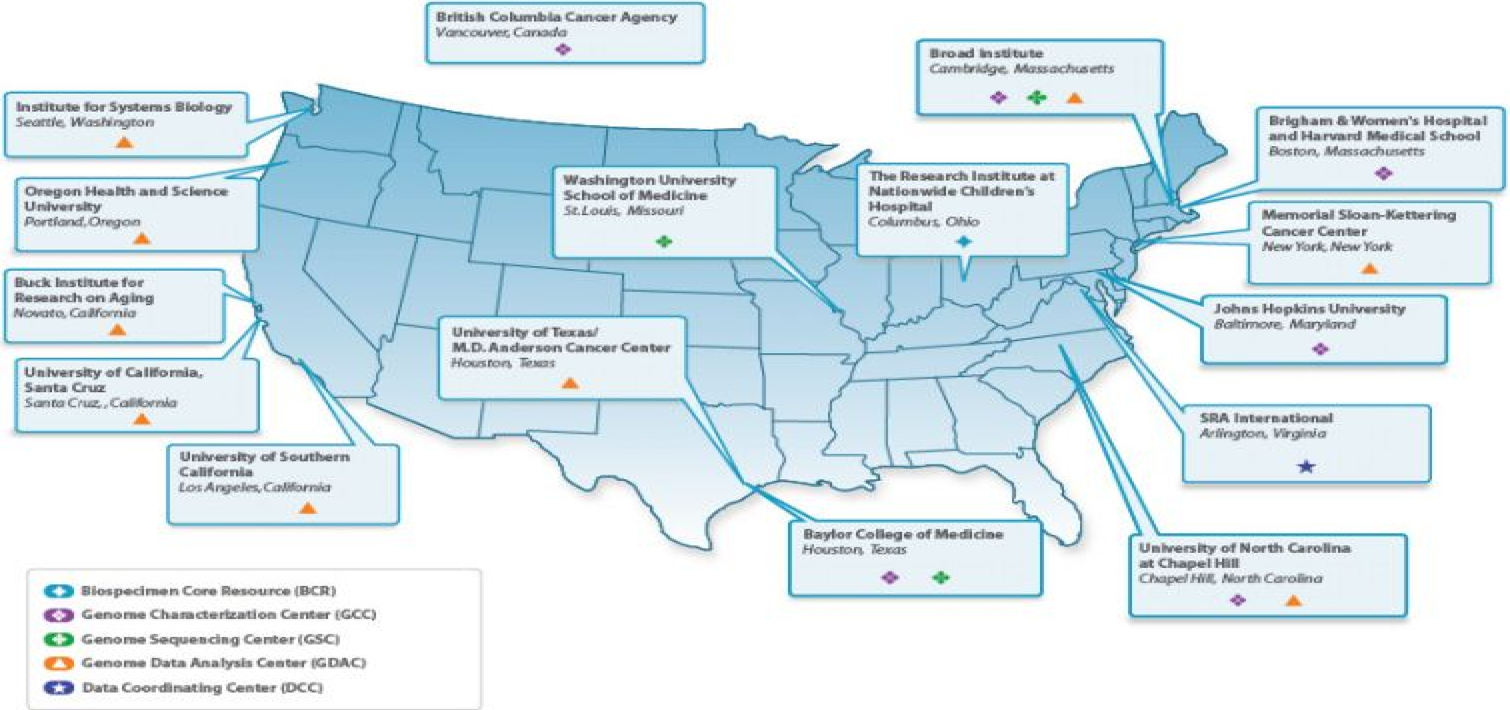
TCGA

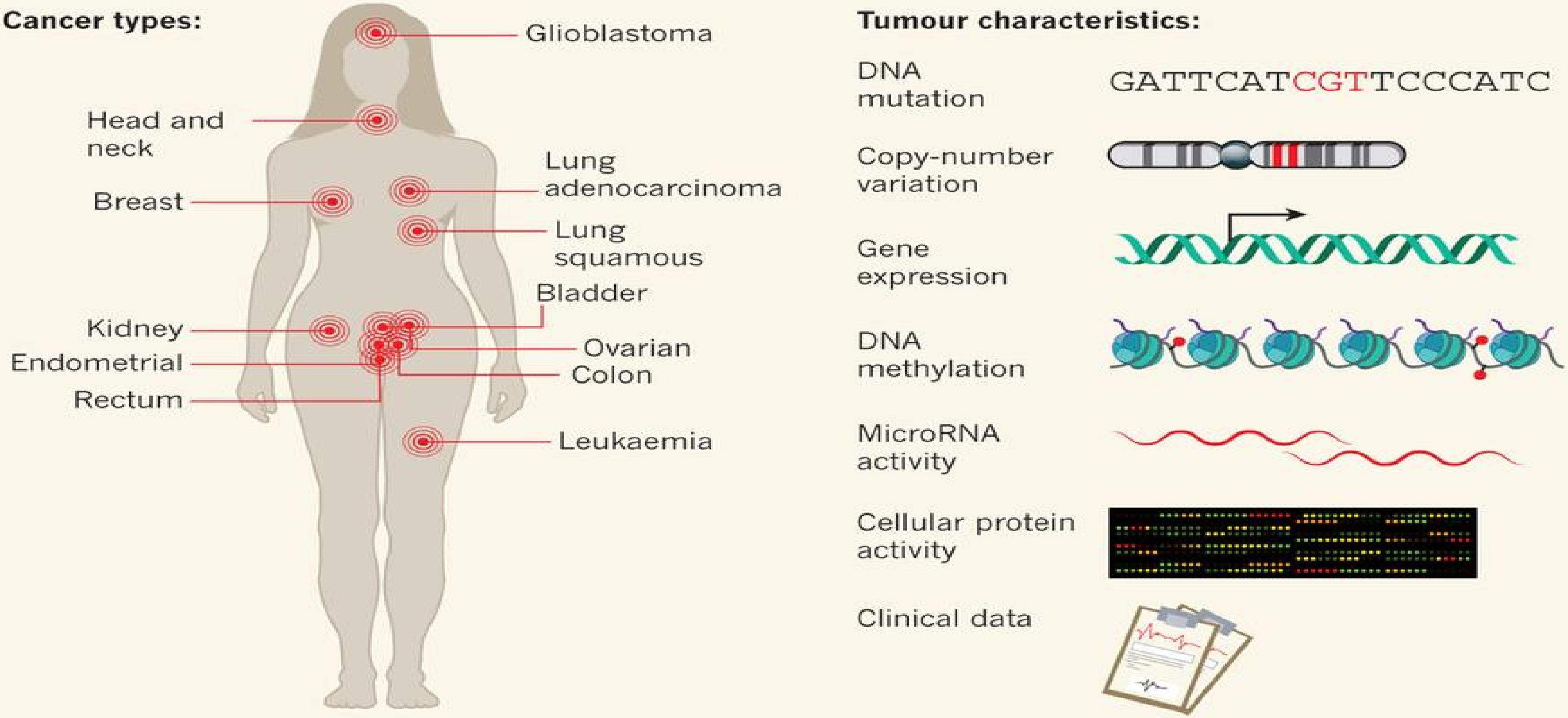
The Cancer Genome Atlas

TCGA: History and Goal

- History:
  - Started in 2005 by the **National Cancer Institute** (NCI) and the **National Human Genome Research Institute** (NHGRI) with *$110 Million* to catalogue genetic mutations responsible for cancer (2006-2009).
  - US Government dedicated ~*$500 Million for the next 5 years* (2010-2015) to characterize **20-30 Cancers**.
- Objective/Goal:
  - Comprehensive and coordinated effort to accelerate our **understanding** of the **molecular basis of cancer** through the application of genome analysis technologies, **including large-scale genome sequencing.**
  - To improve our ability to **diagnose, treat, and prevent cancer** through a better understanding of the **molecular basis of this disease**.

TCGA Map - USA


British Columbia Cancer Agency
Vancouver, Canada
Institute for Systems Biology
Seattle, Washington
Oregon Health and Science University
Portland, Oregon
Buck Institute for Research on Aging
Novato, California
University of California, Santa Cruz
Santa Cruz, , California
University of Southern California
Los Angeles, California
University of Texas/ M.D. Anderson Cancer Center
Houston, Texas
Washington University School of Medicine
St.Louis, Missouri
The Research Institute at Nationwide Children's Hospital
Columbus, Ohio
Broad Institute
Cambridge, Massachusetts
Brigham & Women's Hospital and Harvard Medical School
Boston, Massachusetts
Memorial Sloan-Kettering Cancer Center
New York, New York
Johns Hopkins University
Baltimore, Maryland
SRA International
Arlington, Virginia
Baylor College of Medicine
Houston, Texas
University of North Carolina at Chapel Hill
Chapel Hill, North Carolina
Biospecimen Core Resource (BCR)
Genome Characterization Center (GCC)
Genome Sequencing Center (GSC)
Genome Data Analysis Center (GDAC)
Data Coordinating Center (DCC)



Cancer types:

Glioblastoma

Head and neck

Breast

Lung adenocarcinoma

Lung squamous

Bladder

Kidney

Ovarian

Endometrial

Colon

Rectum

Leukaemia

Tumour characteristics:

DNA mutation

GATTCATCGTTCCCATC

Copy-number variation

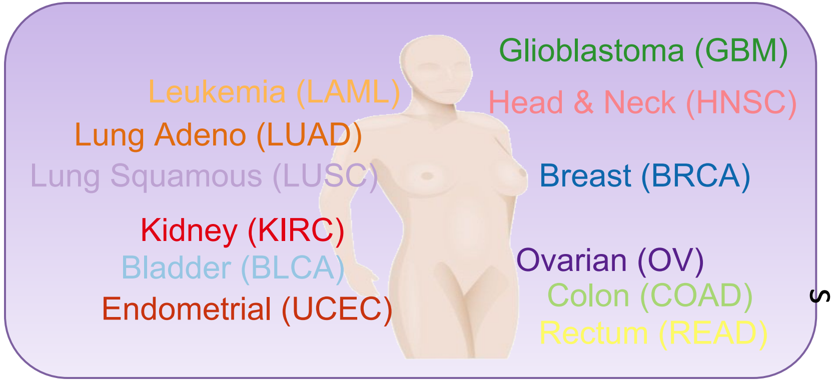
Gene expression

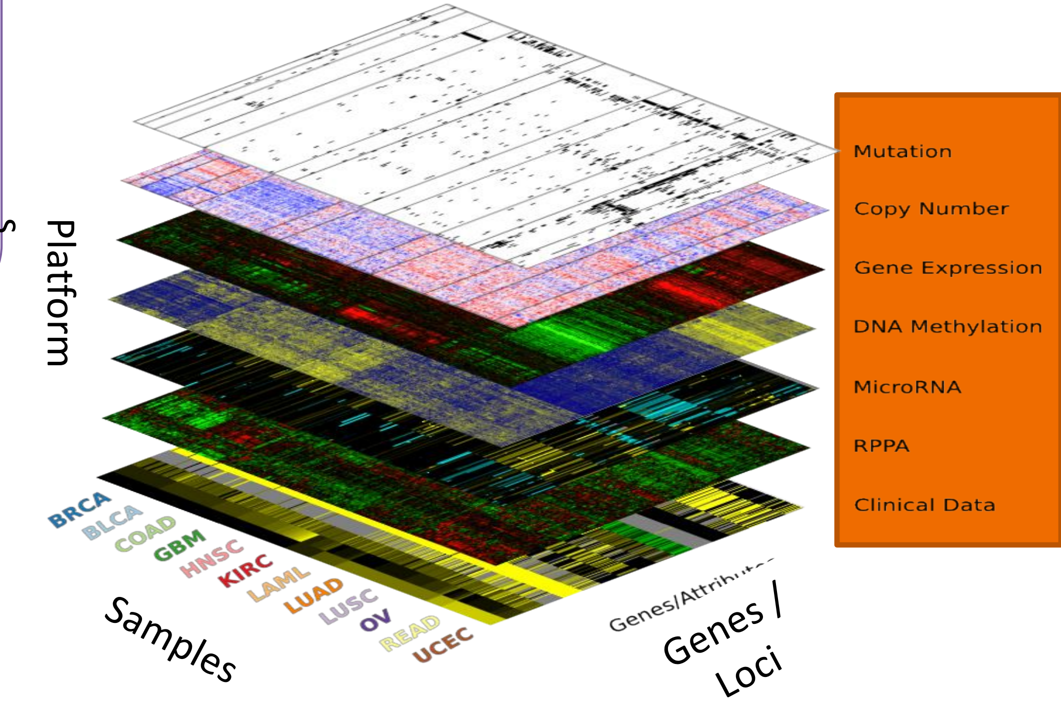
DNA methylation

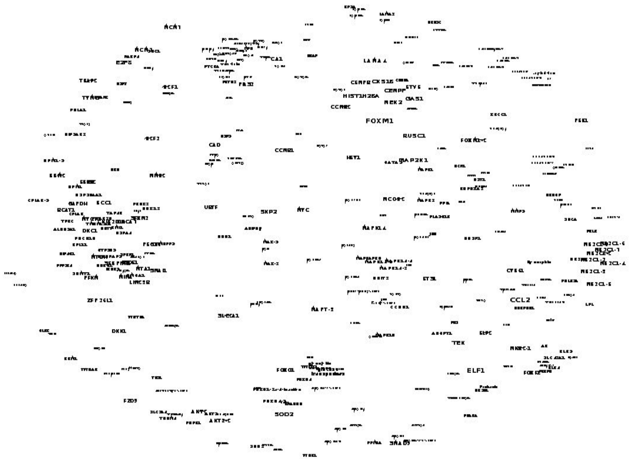
MicroRNA activity

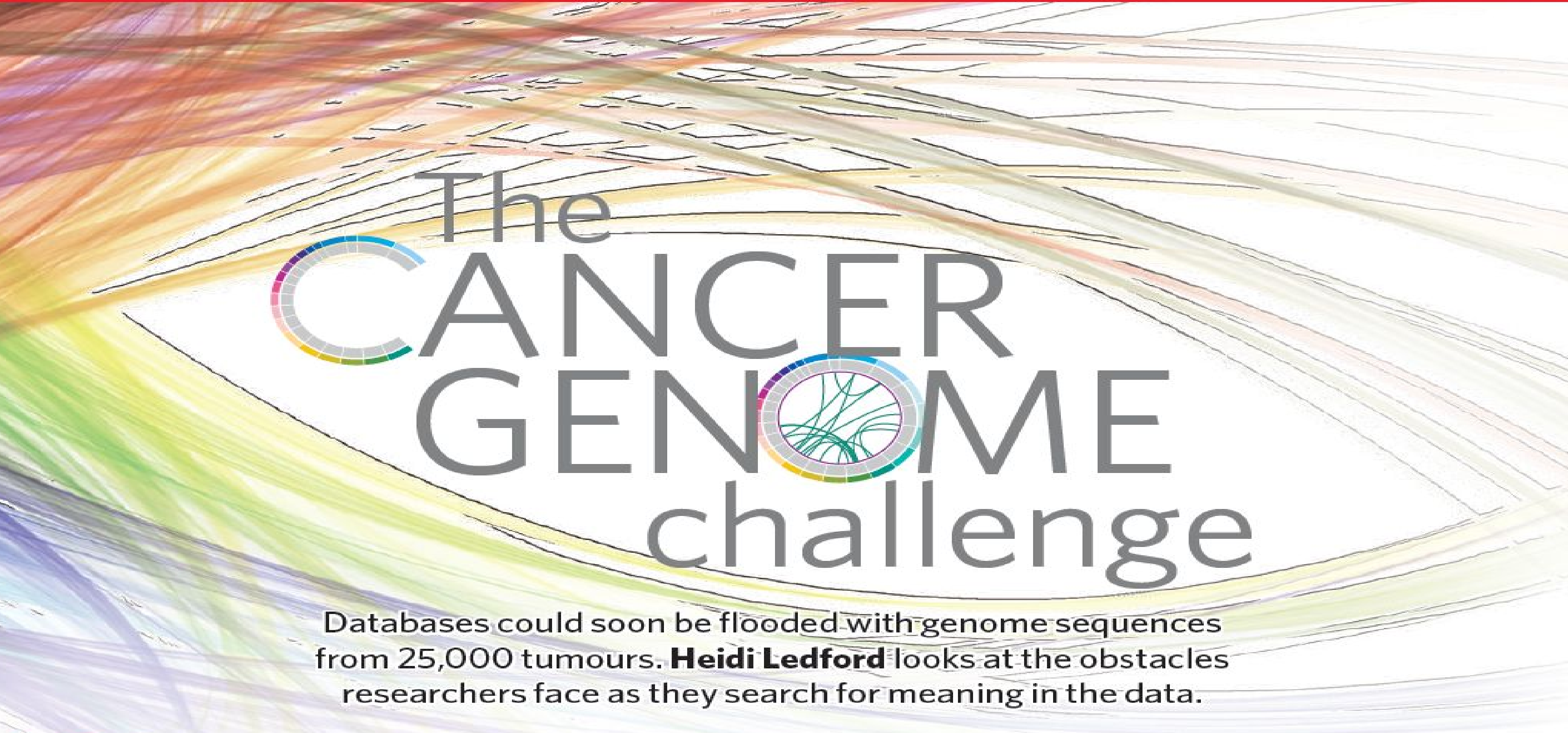
Cellular protein activity

Clinical data


Glioblastoma (GBM)
Leukemia (LAML)
Head & Neck (HNSC)
Lung Adeno (LUAD)
Lung Squamous (LUSC)
Breast (BRCA)
Kidney (KIRC)
Bladder (BLCA)
Ovarian (OV)
Colon (COAD)
Endometrial (UCEC)
Rectum (READ)


s
Platform
Mutation
Copy Number
Gene Expression
DNA Methylation
MicroRNA
RPPA
Clinical Data
BRCA
BLCA
COAD
GBM
HNSC
KIRC
LAML
LUAD
LUSC
OV
READ
UCEC
Samples
Genes/Attributes
Genes / Loci


FOXM1
CCNE1
CAD
SKP2
GAPDH
ZFP36L1
CCL2
TEK
ELF1
SOD2



The CANCER GENOME challenge

Databases could soon be flooded with genome sequences from 25,000 tumours. **Heidi Ledford** looks at the obstacles researchers face as they search for meaning in the data.

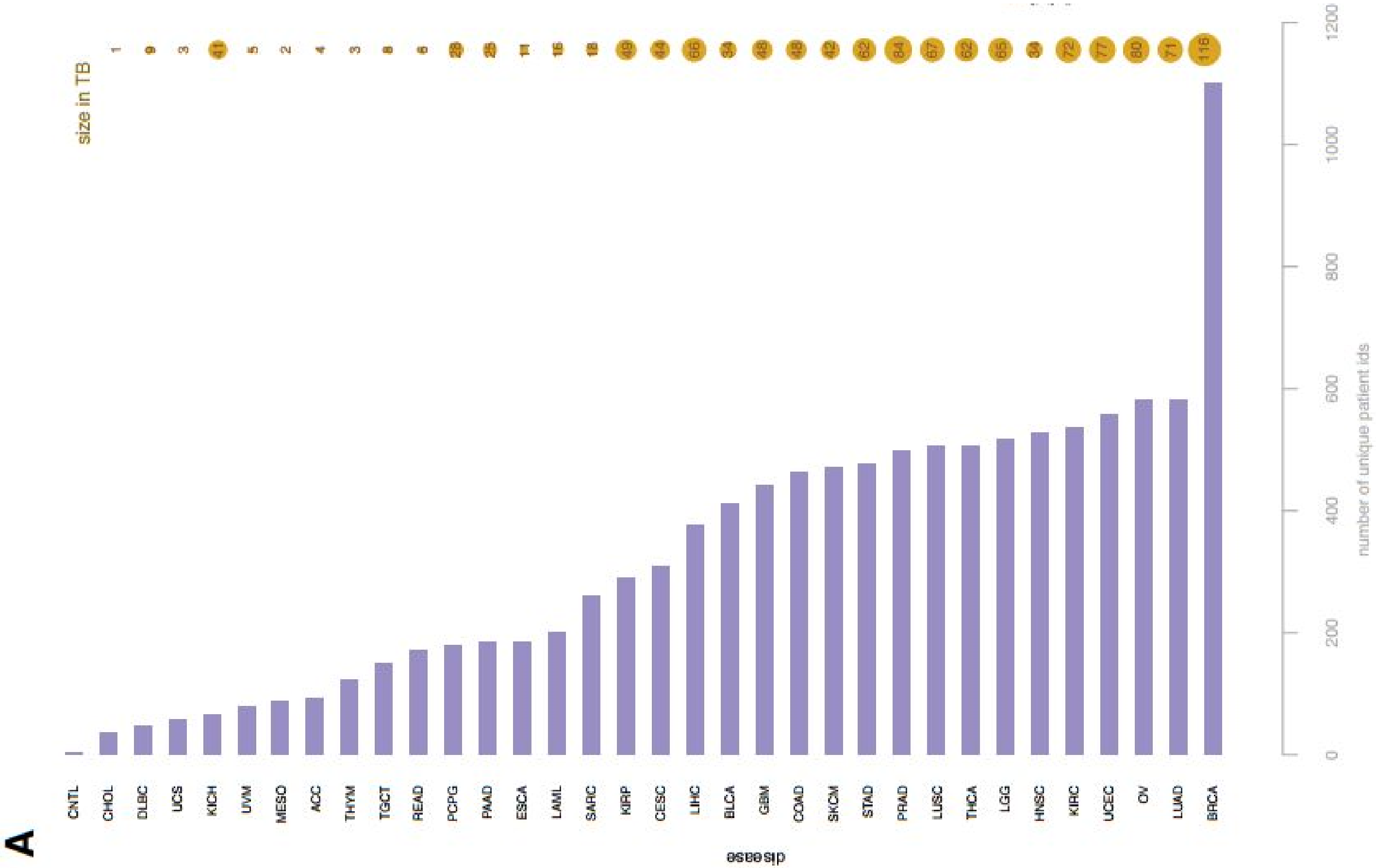
When it was first discovered, in 2006, in a study of 35 colorectal cancers[1], the mutation in the

needle pulled from a veritable haystack of cancer-associated mutations thanks to high-powered genome sequencing. In the past two

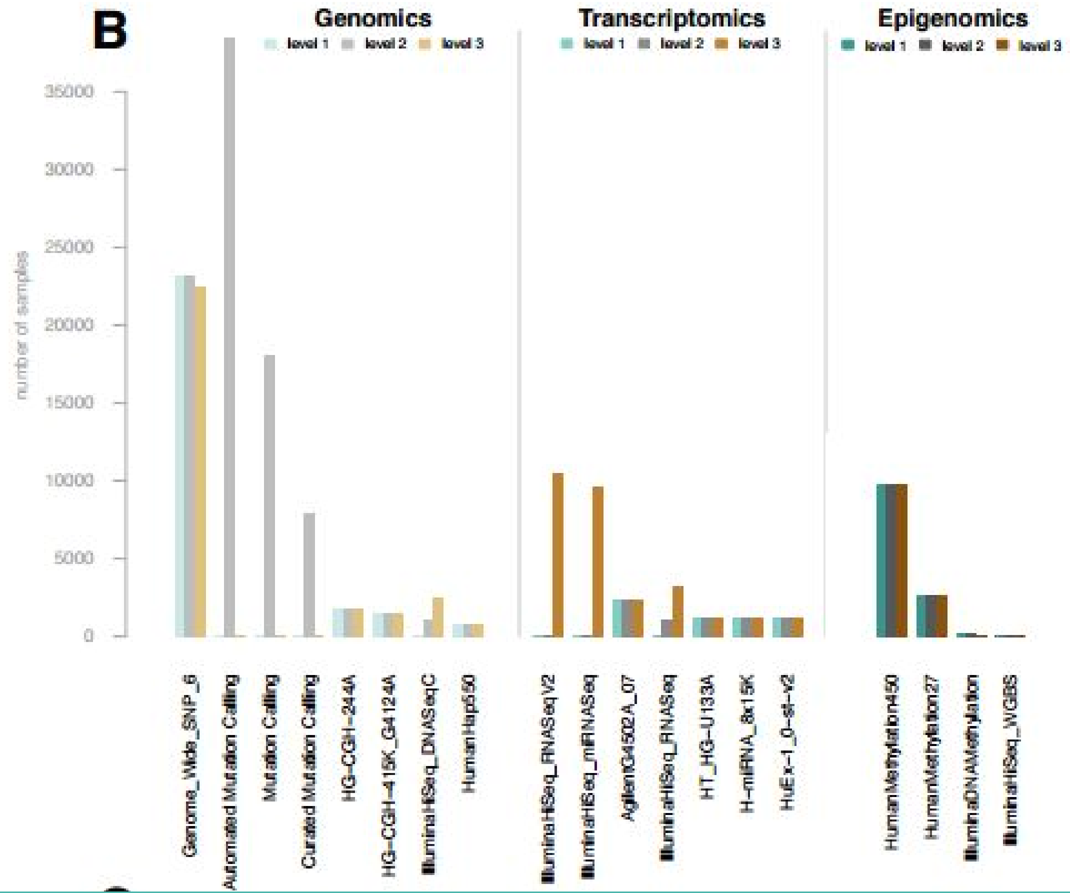
GENOMES AT A GLANCE

Circos plots can give a snapshot of the mutations within

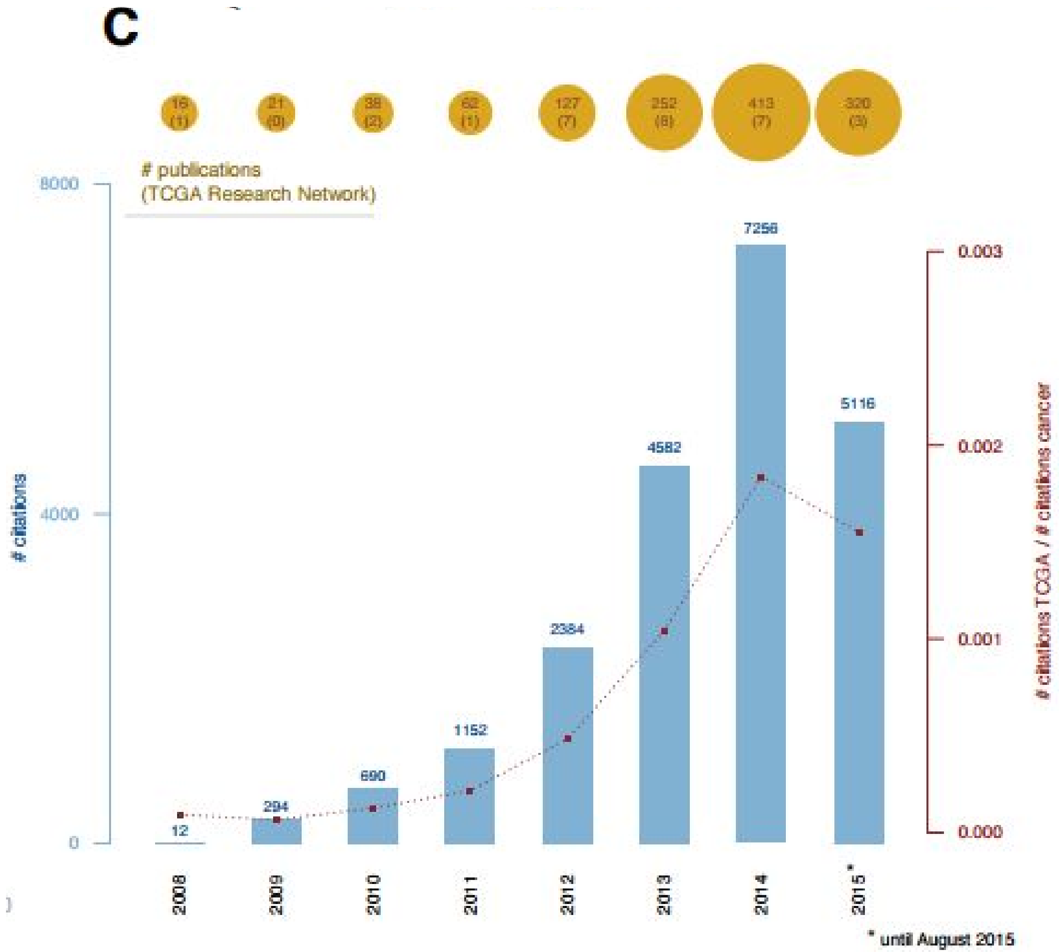
Number of TCGA Samples



TCGA Molecular Data



TCGA Impact


C
16
(1)
21
(0)
38
(2)
62
(1)
127
(7)
252
(8)
413
(7)
320
(3)
# publications
(TCGA Research Network)
8000
4000
0
# citations
12
294
690
1152
2384
4582
7256
5116
0.003
0.002
0.001
0.000
# citations TCGA / # citations cancer
2008
2009
2010
2011
2012
2013
2014
2015*
* until August 2015

Problems

i) Difficulty finding the desired information

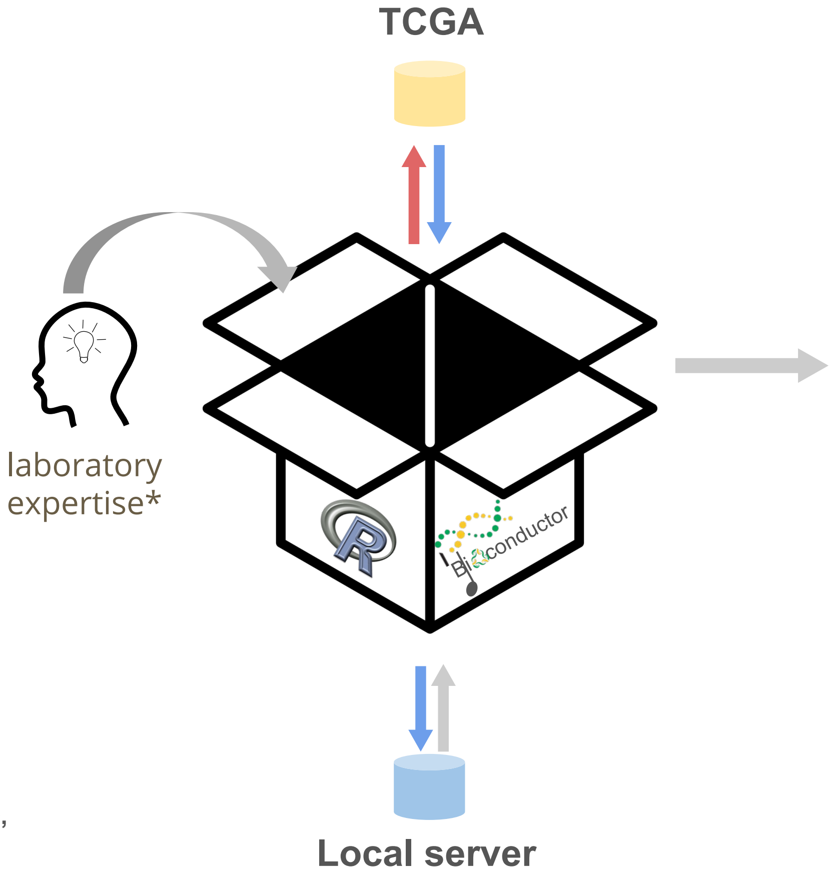
  i.i) If we want an old version of the data the only solution is via the http

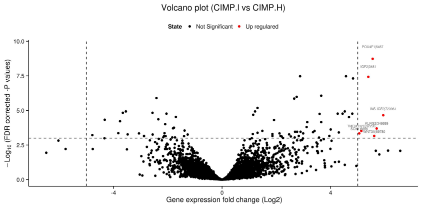
ii) How the data is organized in the portal

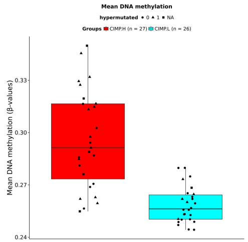
iii) Download complexity of the site

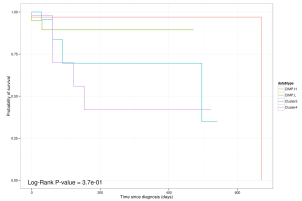
Solution

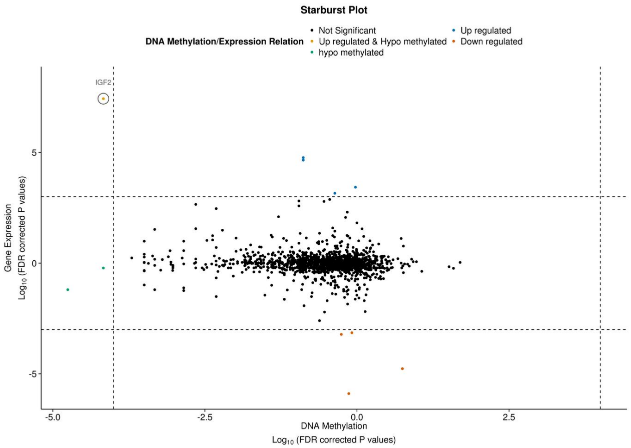
Search
Retrieve
Analysis
Local Database
Public Database











*** PMID: 24885402, 24120142, 23717510, 22684628, 22479200, 22187159, 22120008, 21659424, 20399149**

TCGAbiolinks

An R/Bioconductor package for integrative analysis with TCGA data

Preparing for tutorial

- /dados/uruguay/aluno/tcgabiolinks
- /dados/uruguay/aluno/elmer

Aim

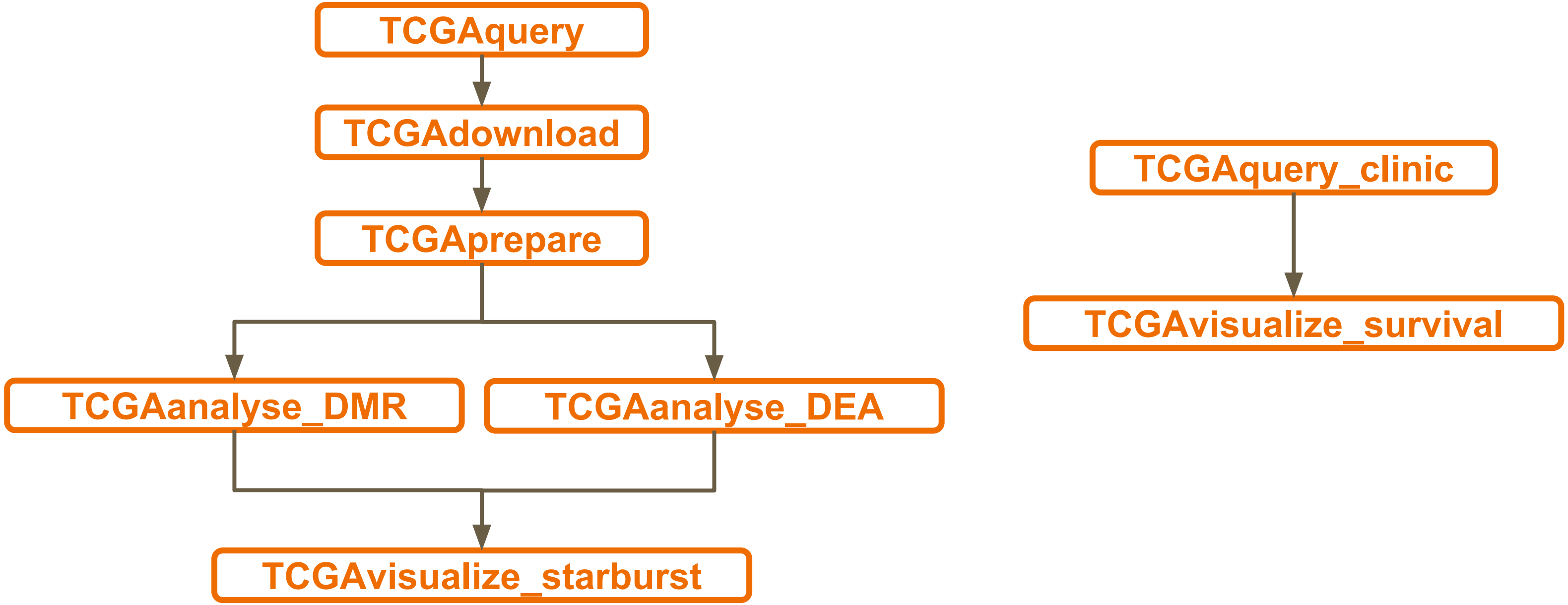
i) facilitate the TCGA open-access data retrieval

ii) prepare the data using the appropriate pre-processing strategies

iii) provide the means to carry out different standard analyses

iv) allow the user to download a specific version of the data and thus to easily reproduce earlier research results.

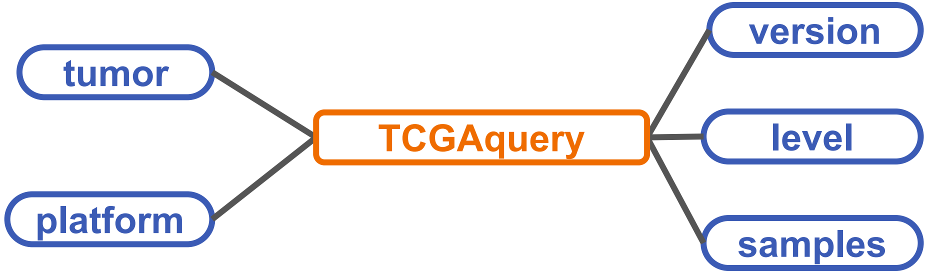
Pipeline



CODE

Final results

Searching for data



tumor

platform

TCGAquery

version

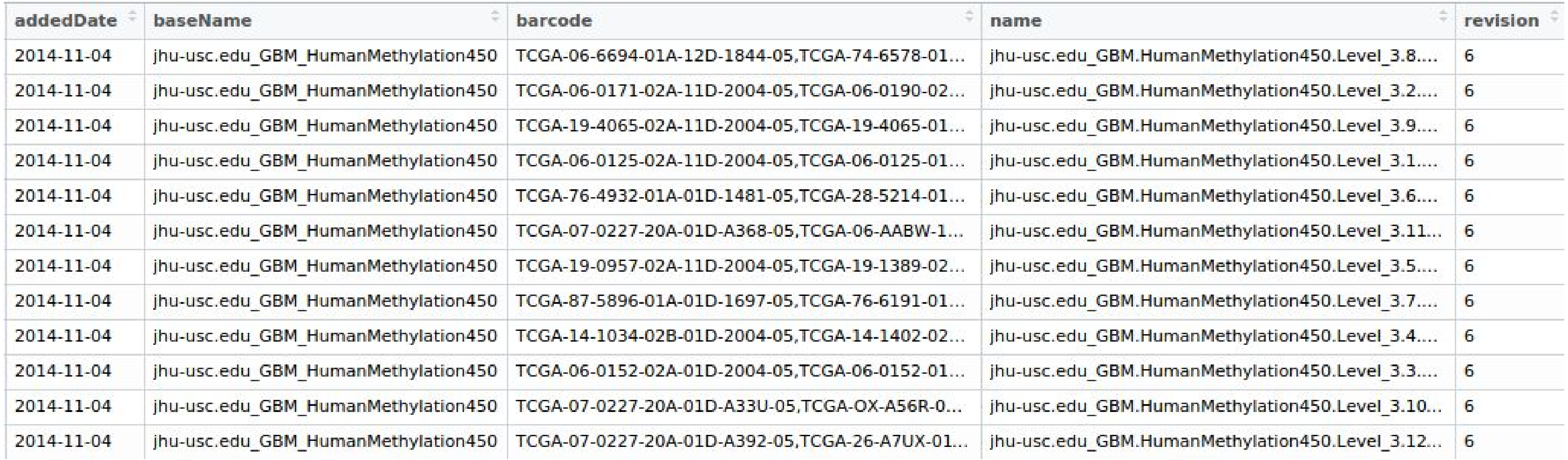
level

samples

```
query <- TCGAquery ( tumor = "GBM",                                    # Vector of tumors
                     platform = "Humanmethylation450",                 # Vector of platforms
                     level = 3,                                        # 1, 2, 3
                     version = list(c("HumanMethylation450","GBM",5)), # List of triple (tumor, platform, version)
                     samples = c("TCGA-06-6694-01A-12D-1844-05",       # Vector of barcodes
                                 "TCGA-06-0171-02A-11D-2004-05")
)
```

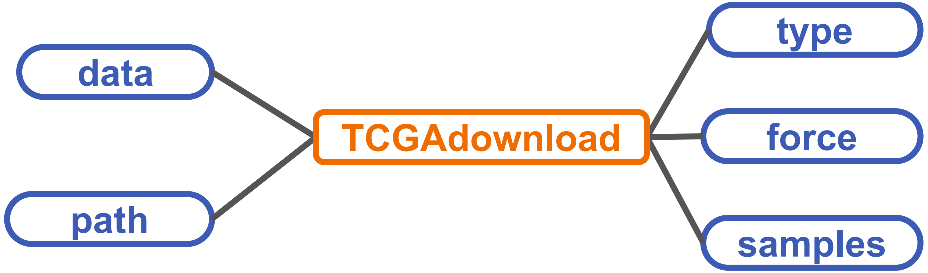
Searching for data - Example

```
query <- TCGAquery ( tumor = "GBM", platform = "Humanmethylation450", level = 3)
```



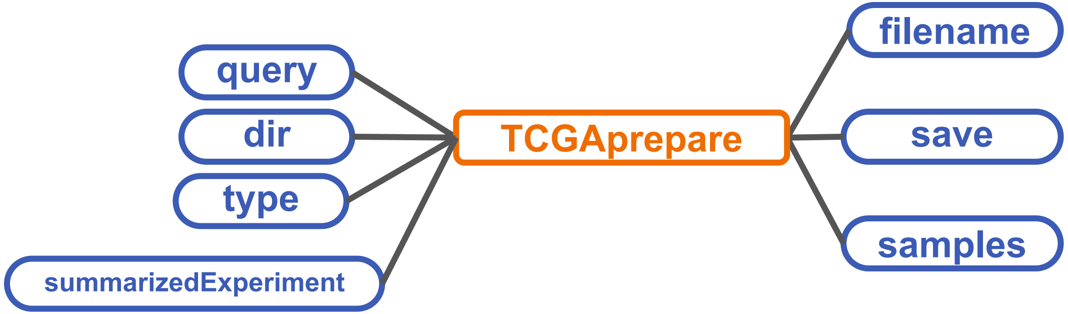
addedDate	baseName	barcode	name	revision
2014-11-04	jhu-usc.edu_GBM_HumanMethylation450	TCGA-06-6694-01A-12D-1844-05,TCGA-74-6578-01...	jhu-usc.edu_GBM.HumanMethylation450.Level_3.8....	6
2014-11-04	jhu-usc.edu_GBM_HumanMethylation450	TCGA-06-0171-02A-11D-2004-05,TCGA-06-0190-02...	jhu-usc.edu_GBM.HumanMethylation450.Level_3.2....	6
2014-11-04	jhu-usc.edu_GBM_HumanMethylation450	TCGA-19-4065-02A-11D-2004-05,TCGA-19-4065-01...	jhu-usc.edu_GBM.HumanMethylation450.Level_3.9....	6
2014-11-04	jhu-usc.edu_GBM_HumanMethylation450	TCGA-06-0125-02A-11D-2004-05,TCGA-06-0125-01...	jhu-usc.edu_GBM.HumanMethylation450.Level_3.1....	6
2014-11-04	jhu-usc.edu_GBM_HumanMethylation450	TCGA-76-4932-01A-01D-1481-05,TCGA-28-5214-01...	jhu-usc.edu_GBM.HumanMethylation450.Level_3.6....	6
2014-11-04	jhu-usc.edu_GBM_HumanMethylation450	TCGA-07-0227-20A-01D-A368-05,TCGA-06-AABW-1...	jhu-usc.edu_GBM.HumanMethylation450.Level_3.11...	6
2014-11-04	jhu-usc.edu_GBM_HumanMethylation450	TCGA-19-0957-02A-11D-2004-05,TCGA-19-1389-02...	jhu-usc.edu_GBM.HumanMethylation450.Level_3.5....	6
2014-11-04	jhu-usc.edu_GBM_HumanMethylation450	TCGA-87-5896-01A-01D-1697-05,TCGA-76-6191-01...	jhu-usc.edu_GBM.HumanMethylation450.Level_3.7....	6
2014-11-04	jhu-usc.edu_GBM_HumanMethylation450	TCGA-14-1034-02B-01D-2004-05,TCGA-14-1402-02...	jhu-usc.edu_GBM.HumanMethylation450.Level_3.4....	6
2014-11-04	jhu-usc.edu_GBM_HumanMethylation450	TCGA-06-0152-02A-01D-2004-05,TCGA-06-0152-01...	jhu-usc.edu_GBM.HumanMethylation450.Level_3.3....	6
2014-11-04	jhu-usc.edu_GBM_HumanMethylation450	TCGA-07-0227-20A-01D-A33U-05,TCGA-OX-A56R-0...	jhu-usc.edu_GBM.HumanMethylation450.Level_3.10...	6
2014-11-04	jhu-usc.edu_GBM_HumanMethylation450	TCGA-07-0227-20A-01D-A392-05,TCGA-26-A7UX-01...	jhu-usc.edu_GBM.HumanMethylation450.Level_3.12...	6

Downloading the data



```
TCGAdownload(data = query,                                   # TCGAquery result
             path = ".",                                     # Path to save files
             samples = c("TCGA-06-6694-01A-12D-1844-05",     # Vector of barcodes to download file
                         "TCGA-06-0171-02A-11D-2004-05"),
             force = FALSE)                                  # If already downloaded download it again?
```

Reading the data

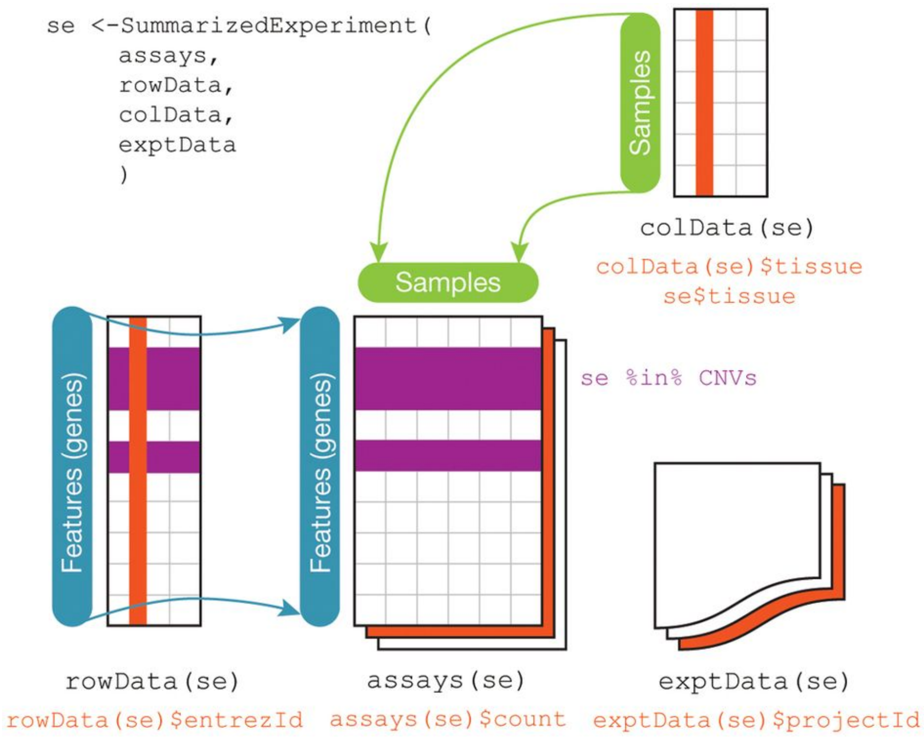


```
TCGAprepare(query = query,
            dir = ".",
            samples = c("TCGA-06-6694-01A-12D-1844-05",
                        "TCGA-06-0171-02A-11D-2004-05"),
           save = TRUE,
           filename = "name.rda",
          summarizedExperiment = TRUE )
```

```
# TCGAquery result
# Path where files were saved
# Vector of barcodes to read file

# Save prepared object?
# Name of the file with the prepared object
# If FALSE output is a data.frame
```

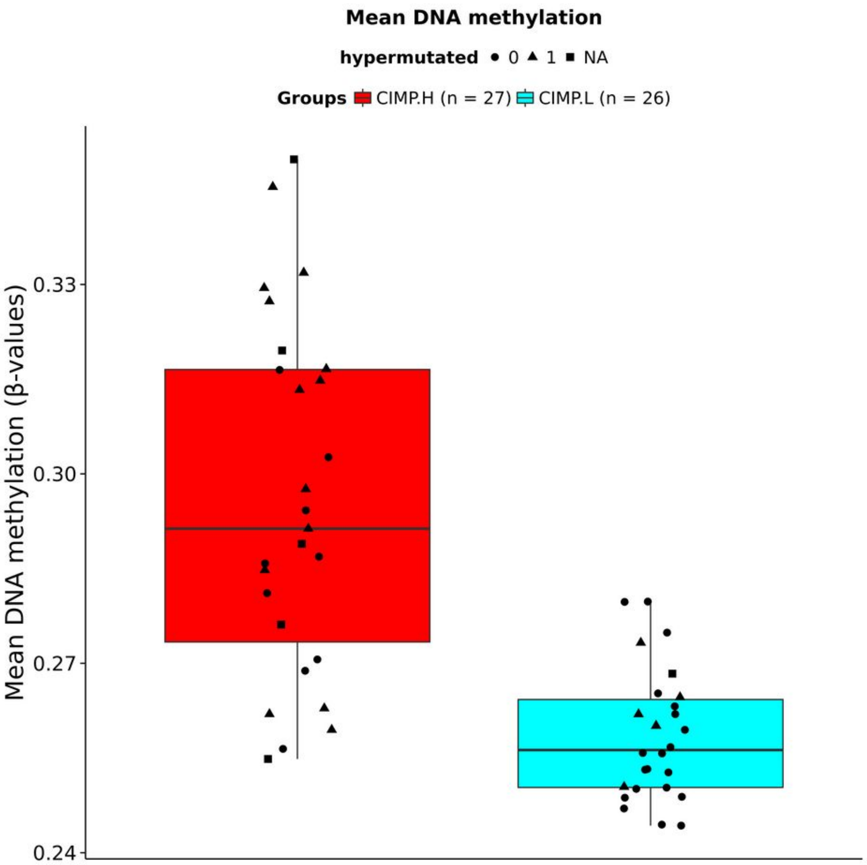
SummarizedExperiment



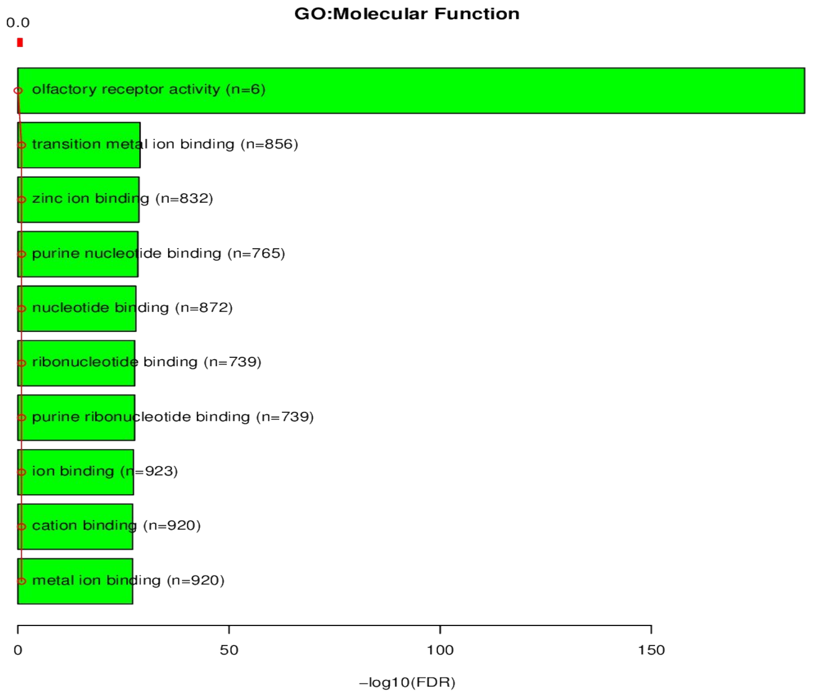
```
se <-SummarizedExperiment(
        assays,
        rowData,
        colData,
        exptData
        )
```

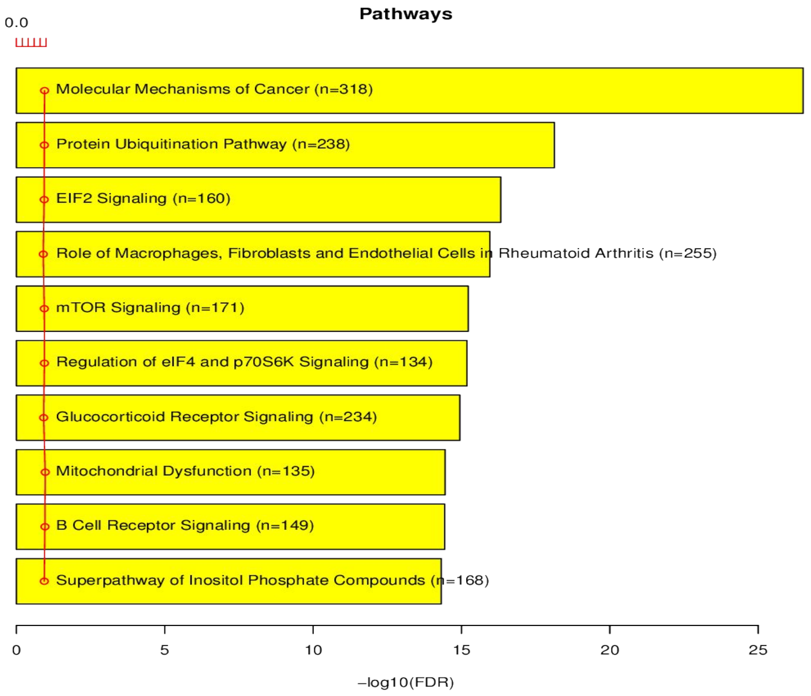
Huber, Wolfgang, et al. "Orchestrating high-throughput genomic analysis with Bioconductor." *Nature methods* 12.2 (2015): 115-121.

Results

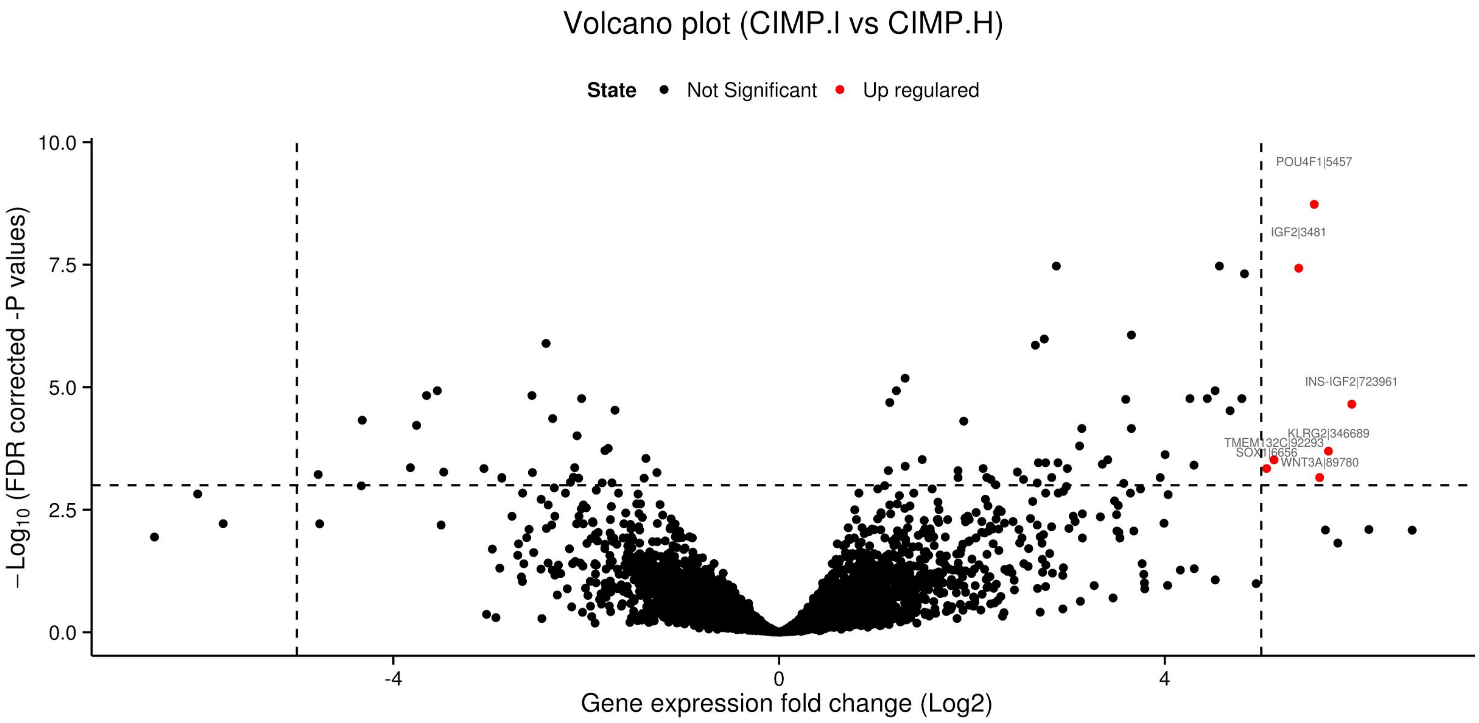

Mean DNA methylation
hypermutated ● 0 ▲ 1 ■ NA
Groups 🟥 CIMP.H (n = 27) 🟦 CIMP.L (n = 26)
Mean DNA methylation (β-values)
0.33
0.30
0.27
0.24

Results

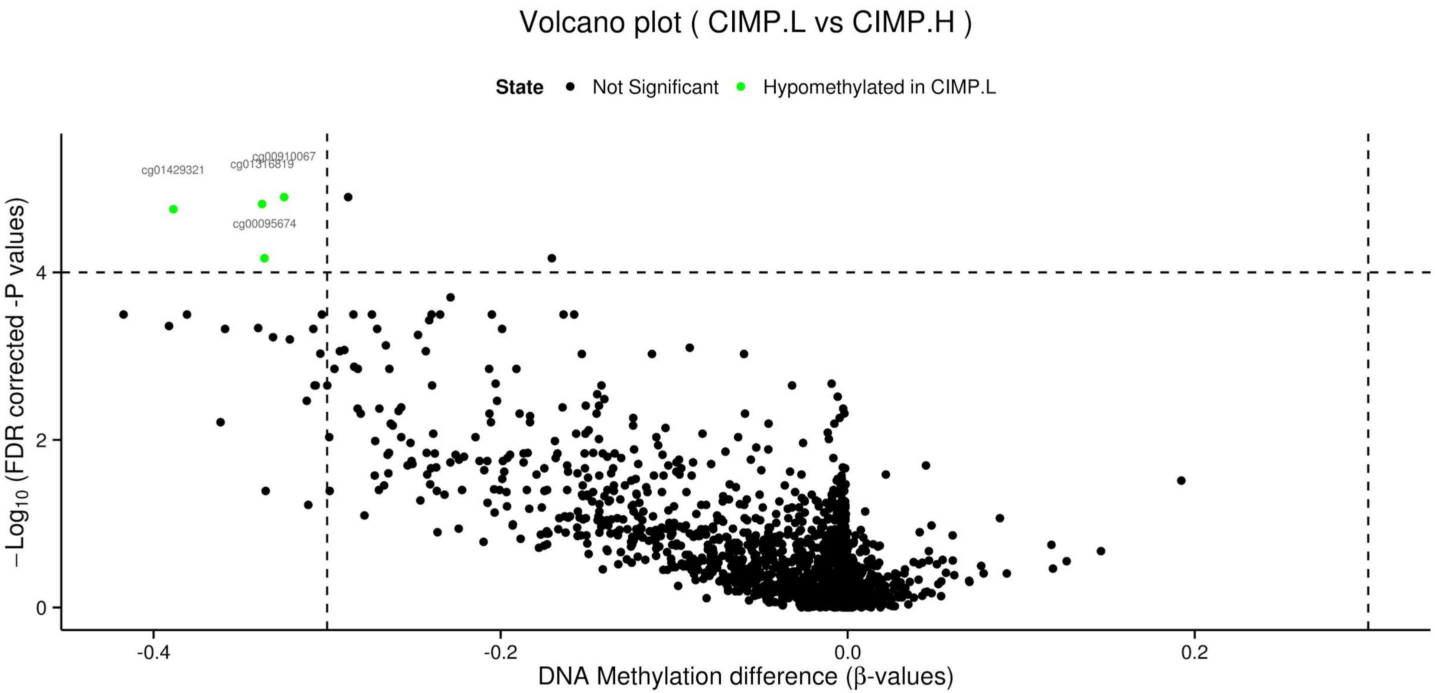

GO:Molecular Function
0.0
olfactory receptor activity (n=6)
transition metal ion binding (n=856)
zinc ion binding (n=832)
purine nucleotide binding (n=765)
nucleotide binding (n=872)
ribonucleotide binding (n=739)
purine ribonucleotide binding (n=739)
ion binding (n=923)
cation binding (n=920)
metal ion binding (n=920)
0
50
100
150
−log10(FDR)


Pathways
0.0
Molecular Mechanisms of Cancer (n=318)
Protein Ubiquitination Pathway (n=238)
EIF2 Signaling (n=160)
Role of Macrophages, Fibroblasts and Endothelial Cells in Rheumatoid Arthritis (n=255)
mTOR Signaling (n=171)
Regulation of eIF4 and p70S6K Signaling (n=134)
Glucocorticoid Receptor Signaling (n=234)
Mitochondrial Dysfunction (n=135)
B Cell Receptor Signaling (n=149)
Superpathway of Inositol Phosphate Compounds (n=168)
0
5
10
15
20
25
−log10(FDR)

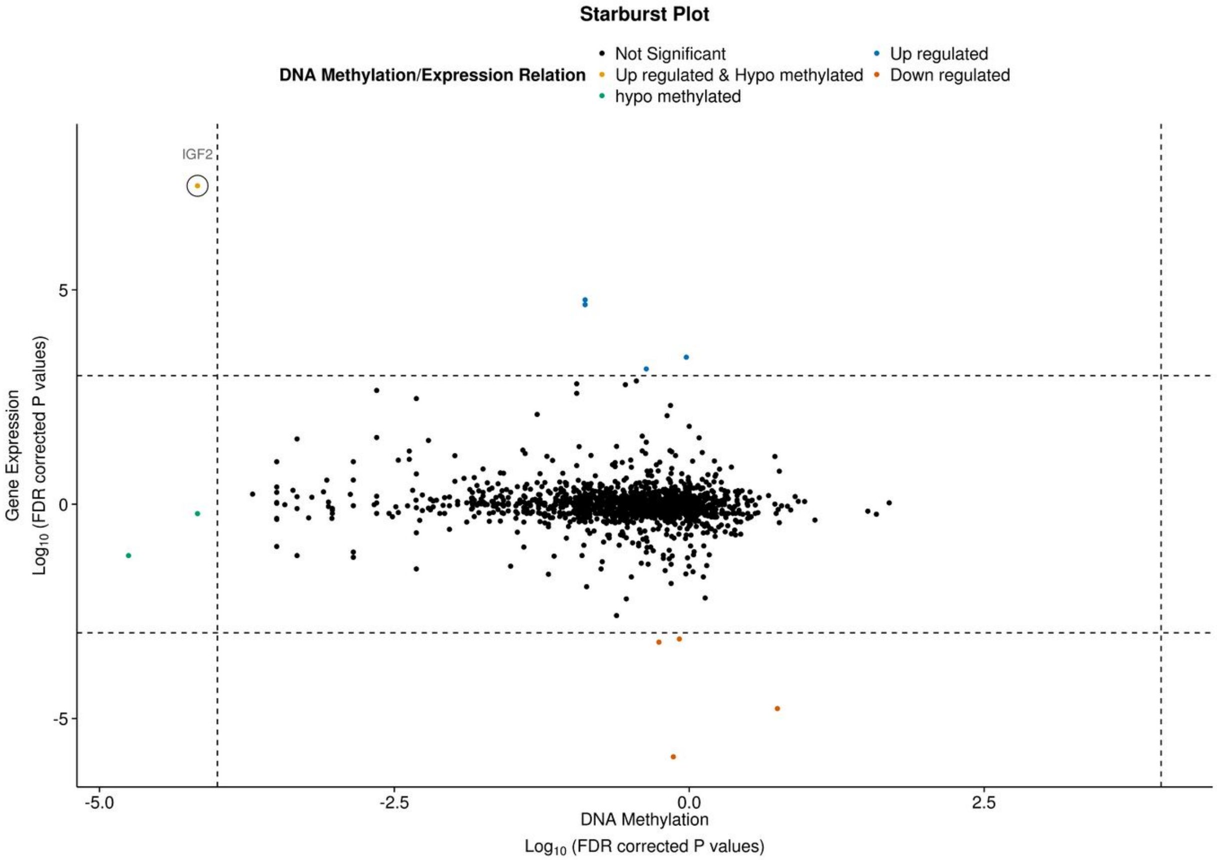
Results


Volcano plot (CIMP.l vs CIMP.H)
State
● Not Significant
● Up regulared
10.0
7.5
5.0
2.5
0.0
$-\text{Log}_{10}$ (FDR corrected -P values)
-4
0
4
Gene expression fold change (Log2)
POU4F1|5457
IGF2|3481
INS-IGF2|723961
KLRG2|346689
TMEM132C|92293
SOX1|6656
WNT3A|89780

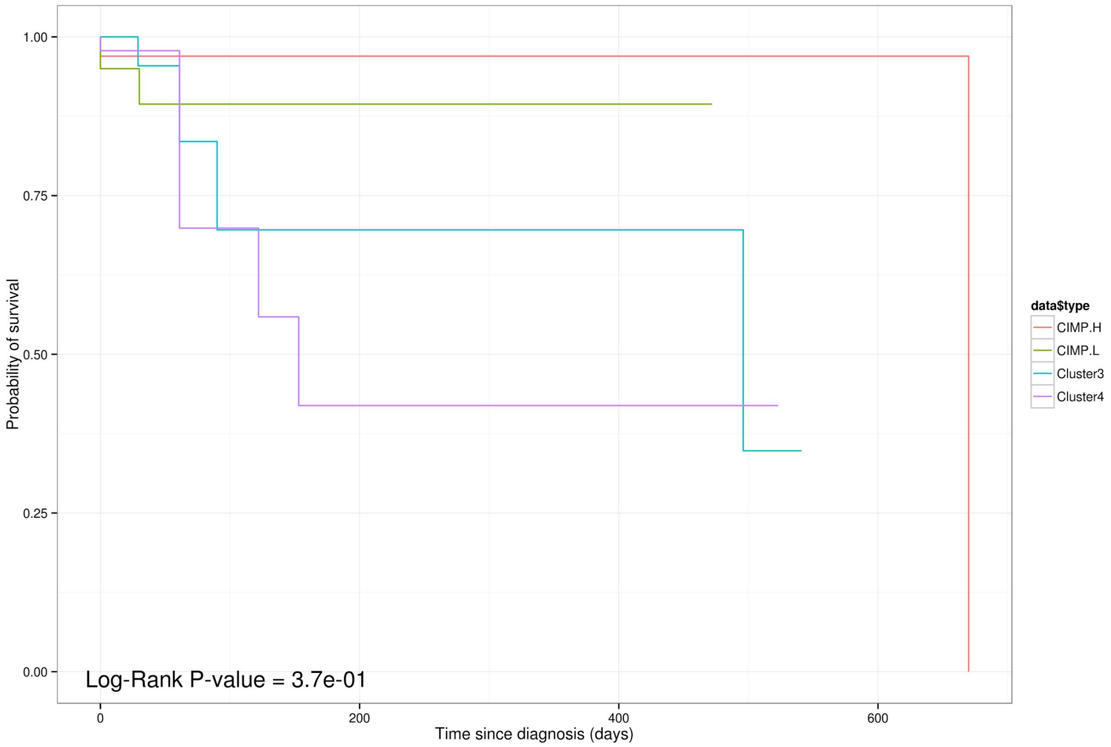
Results


Volcano plot (CIMP.L vs CIMP.H)
State
● Not Significant
● Hypomethylated in CIMP.L
cg01429321
cg00910067
cg01316819
cg00095674
−Log$_{10}$ (FDR corrected -P values)
0
2
4
-0.4
-0.2
0.0
0.2
DNA Methylation difference (β-values)

Starburst


Starburst Plot
DNA Methylation/Expression Relation
● Not Significant
● Up regulated & Hypo methylated
● hypo methylated
● Up regulated
● Down regulated
IGF2
5
0
-5
Gene Expression
Log_{10} (FDR corrected P values)
-5.0
-2.5
0.0
2.5
DNA Methylation
Log_{10} (FDR corrected P values)

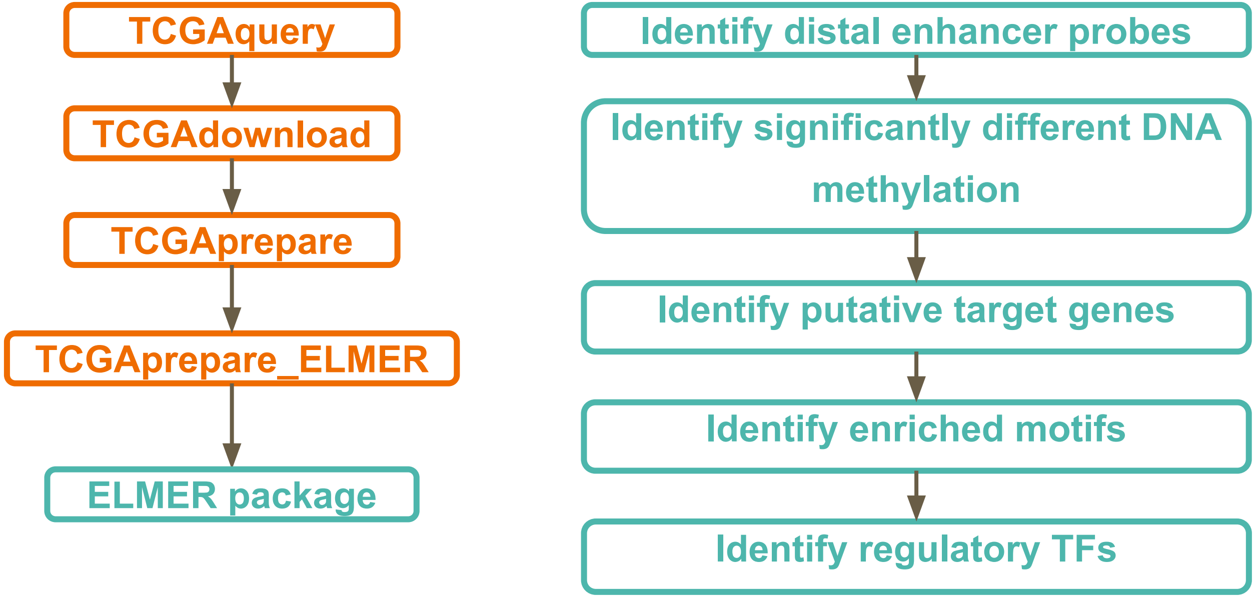
Results


1.00
0.75
0.50
0.25
0.00
Probability of survival
0
200
400
600
Time since diagnosis (days)
data$type
CIMP.H
CIMP.L
Cluster3
Cluster4
Log-Rank P-value = 3.7e-01

ELMER package

ELMER is designed to use DNA methylation and gene expression from a large number of samples to infere regulatory element landscape and transcription factor network in primary tissue.

Pipeline - TCGAbiolinks + ELMER

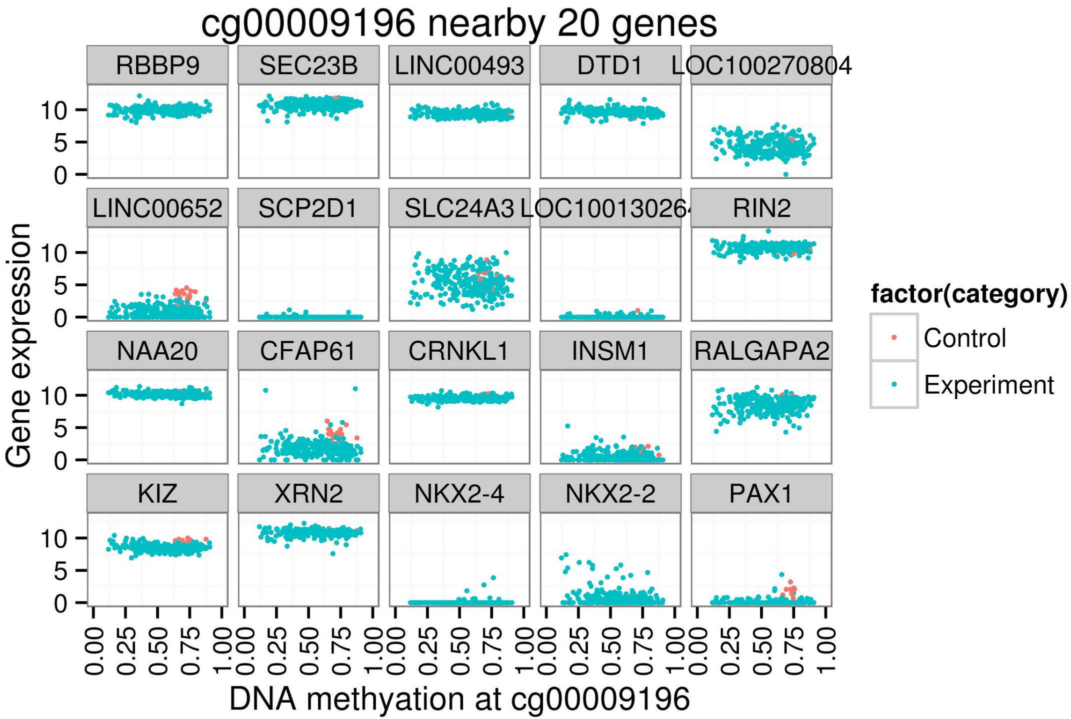

TCGAquery
TCGAdownload
TCGAprepare
TCGAprepare_ELMER
ELMER package
Identify distal enhancer probes
Identify significantly different DNA methylation
Identify putative target genes
Identify enriched motifs
Identify regulatory TFs

CODE

Final results

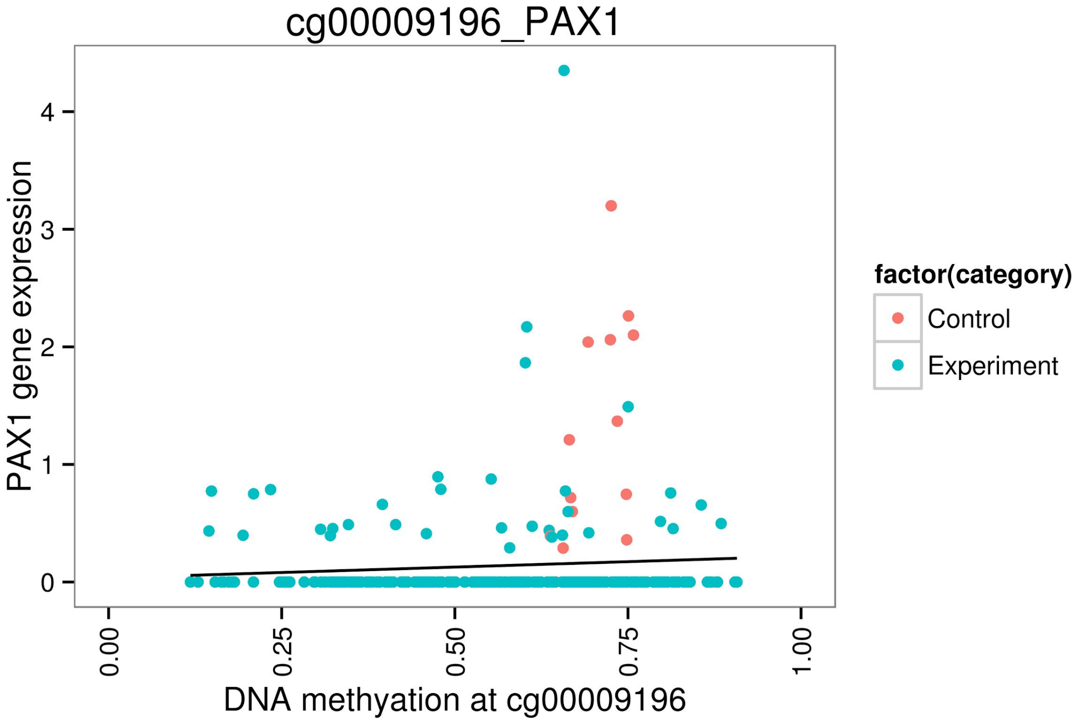
Results

Generate scatter plots for one probes’ nearby 20 gene expression vs DNA methylation at this probe.



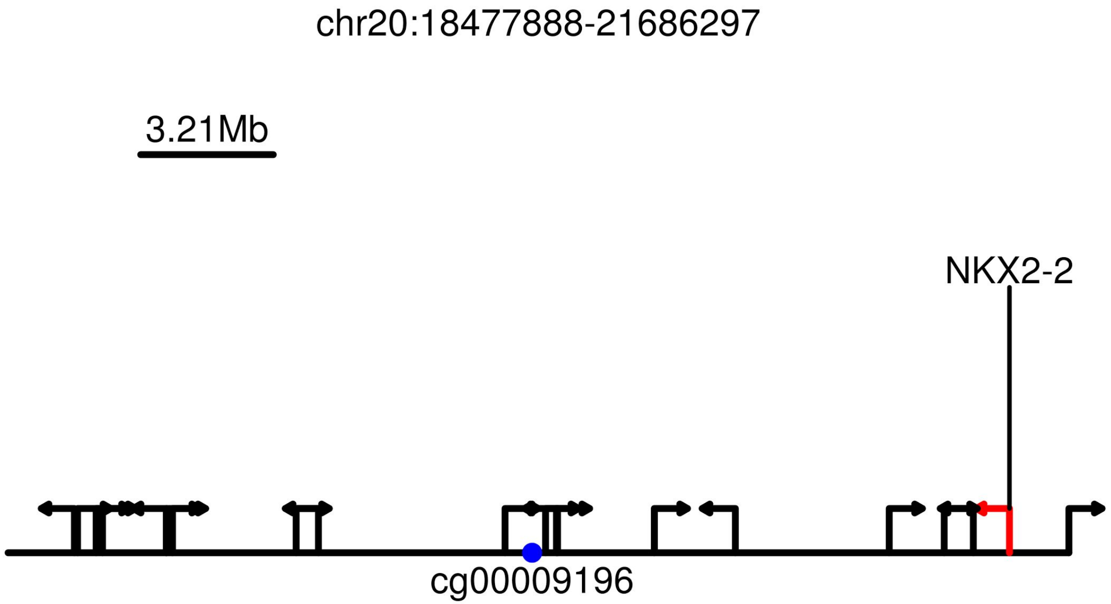
Results

You can also focus on one probe-gene pair. The entrez gene ID for PAIX is 5075.



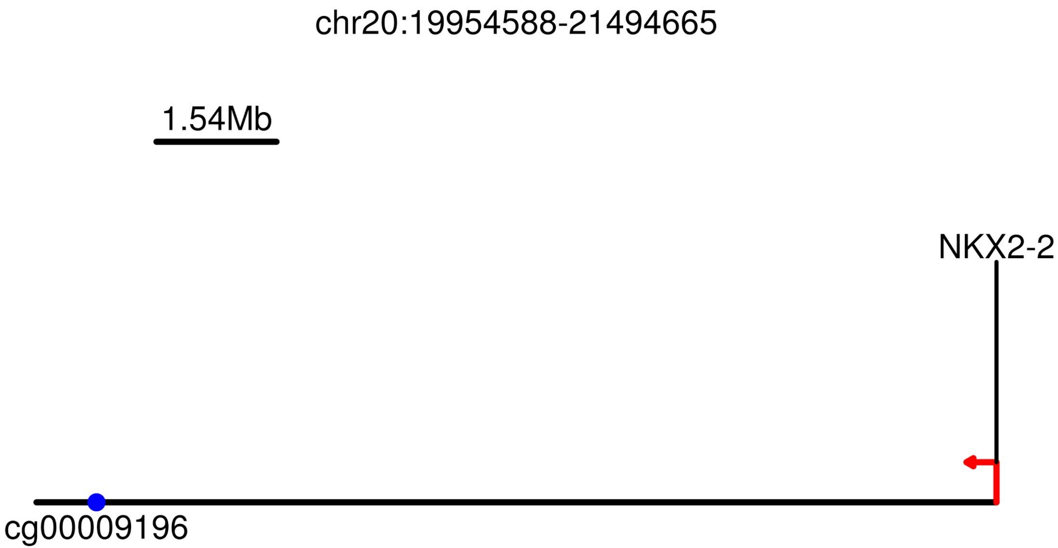
Results

You can generate schematic plot for one probe with 20 nearby genes and label the gene significantly linked with the probe in red.



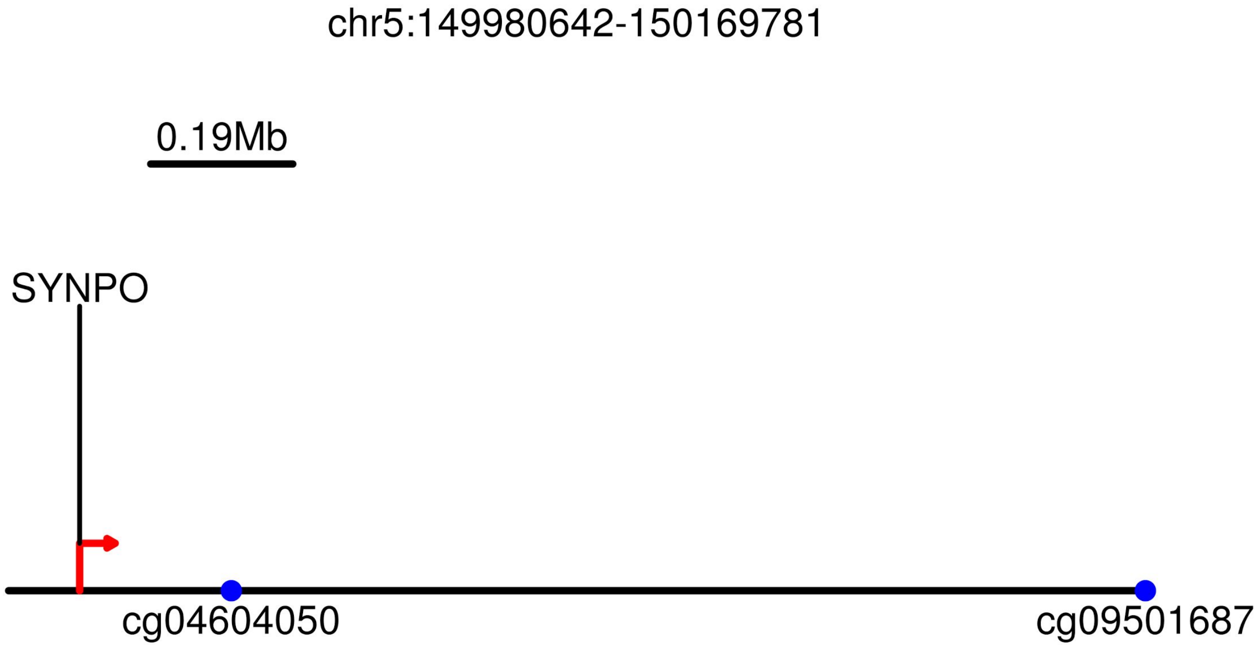
Results

Generate schematic plot for one gene with the probes which the gene is significantly linked to.



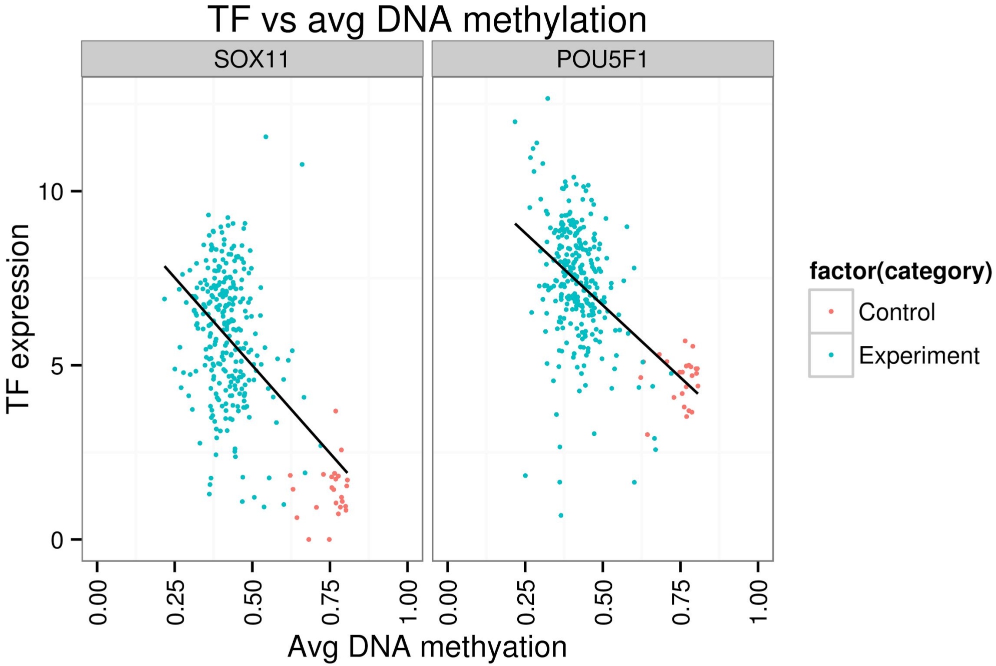
Results

Sometimes there is more than one probe nearby.



Results

We can generate a scatter plot for TF expression vs average DNA methylation of the sites with certain motif.



References

- TCGAbiolinks
- ELMER

Troubleshoot

- TCGAdownload is not working or it crashed.
    - TCGA data portal might be down or the access limit was reached and your IP was blocked. Try again in some minutes
    - If it is not working the following file has all the objects you need for the course, just upload this to your rstudio.