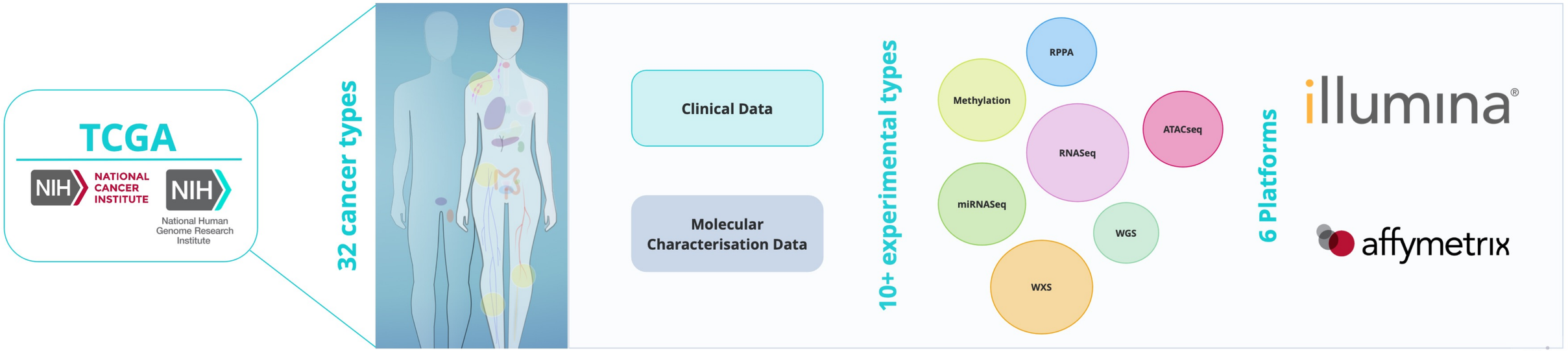
The Cancer Genome Atlas (TCGA)

Biomedical Data Repository for Cancer Data Analysis

Pawan Verma, Sun 25 Sept 2022

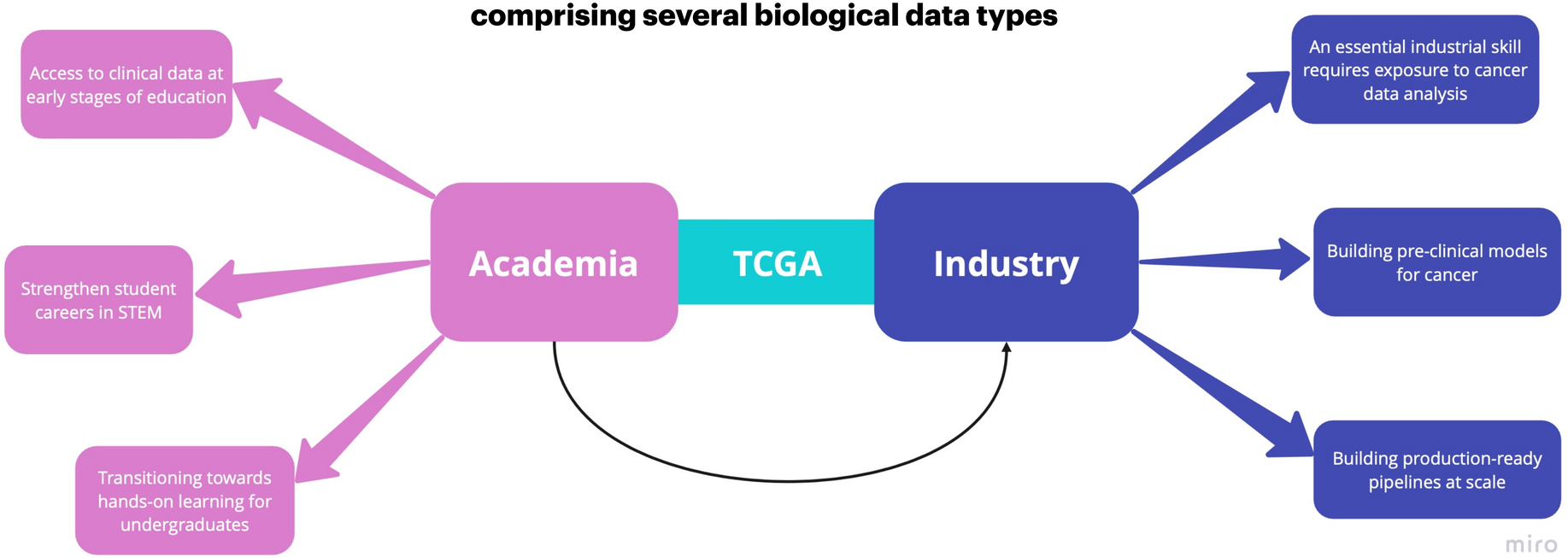
Summary of TCGA Data Repository



For more information, visit https://portal.gdc.cancer.gov/

Why TCGA?

Aside from being a one-stop shop for accessing biomedical data comprising several biological data types



How to access TCGA Data ?

Various public platforms from where TCGA datasets can be viewed, analysed and downloaded.
These platforms are created by various stakeholders in TCGA

The Cancer Imaging Archive (TCIA)

Provide access to radiological imaging data

cBioPortal for cancer genomics

Analyse, visualise and download data on large scale

Copy Number Portal

Explore copy number alterations of TCGA data

The Cancer Imaging Archive (TCIA)

Provide access to radiological imaging data

Firebrowse

Tool to explore and visualise cancer data and provide API access

Regulome Explorer

Web based tool for exploring associations between mutation and clinical data

SurvNet

A web-based tool for identifying network-based biomarkers that correlate with patient survival data

Xena

Suite of web-based tools to visualise, integrate and analyse cancer data

and more.....

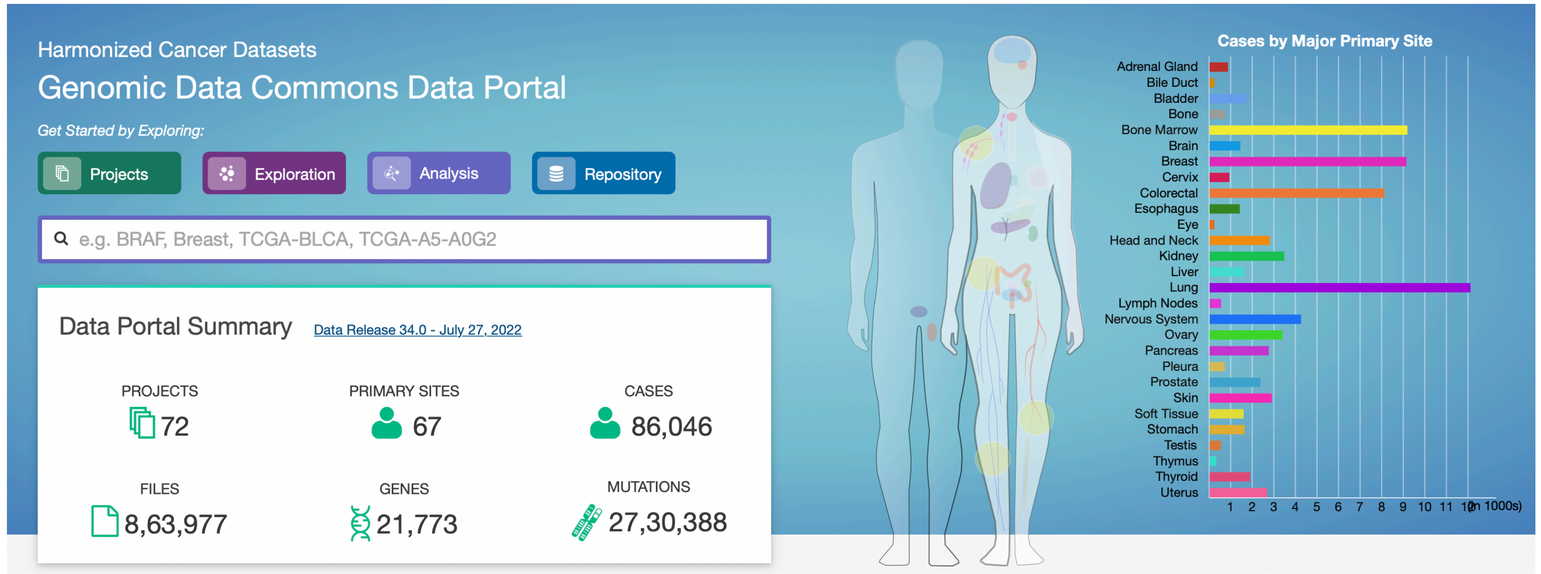
Key Pointers

1. Web based tools and GUIs are useful for performing quick investigation of the data landscape for a particular cancer type, such as number of samples, demographic data, and nature the processed data.
2. Real data analysis is performed by accessing data from a central repository which forms the primary data source for above tools.
3. Be advised that **not all** the above sources undergo frequent updation and may be outdated.

For more information, visit https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/using-tcga/tools

Accessing TCGA data from GDC

Genomic Data Commons Portal for harmonised data access



Tools for communicating with the GDC API

Many third-party tools can be used for communicating with the GDC API and for preparing and visualizing API calls.

Examples of tools for communicating with the GDC API:

Tool	Type
Curl	Command line tool
HTTPie	Command line tool
Postman REST Client	App for Google Chrome and OS X
DHC REST Client	Google Chrome extension
Google Chrome	Google Chrome web browser

Examples of tools that can help build GDC API calls:

Tool	Description
JSONLint	Validate JSON
JSON Formatter	Format, validate, and convert JSON to other formats
Percent-(URL)-encoding tool	Tool for percent-encoding strings

Key Pointers:

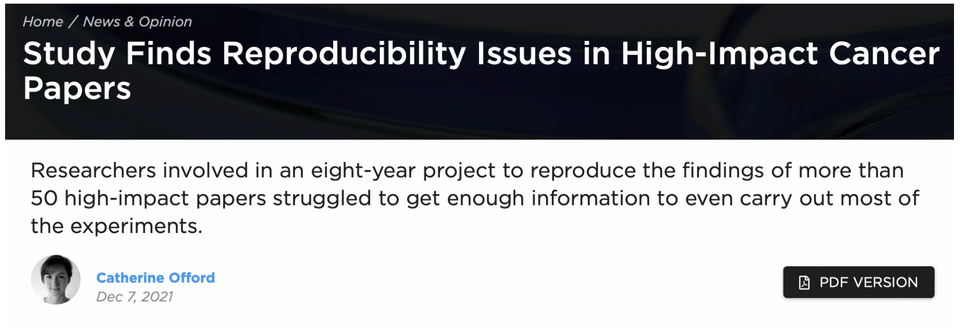
1. Users can **browse** cancer datasets directly from the interactive GDC Portal
2. It can be a bit challenging to navigate for beginners therefore you must use this portal mainly for gaining familiarity with various cancer data types.
3. For programmatic users, GDC provides a REST API facility for search and download.
4. For performing data analysis in R, there are specialised R packages available for downloading and analysing the data.

For more information, visit https://portal.gdc.cancer.gov/

Best practises for early researchers

Especially keen to learn about biomarker discovery

- A lot of time is spent in understanding biological mechanisms in cancer research.
- Researchers hardly focus on enabling clinical translation of their findings.
- Biomarkers based models are hardly generalisable and lack business value.
- This is mainly due to lack of knowledge and training to access the right datasets.
- Academia generally focussed on winning the “funding race”
- A lot of cancer research papers are not reproducible



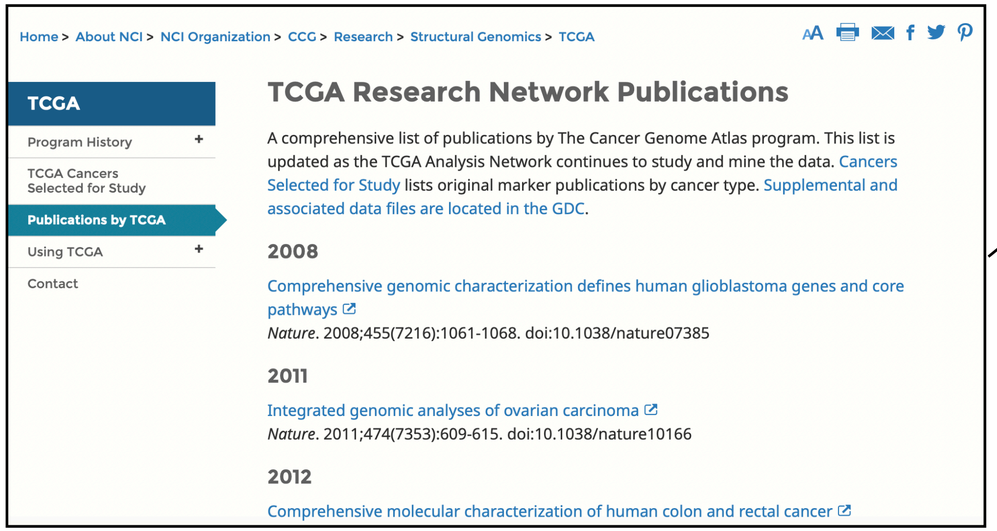
> Of course, the validation attempts may have failed because of technical differences or difficulties, despite efforts to ensure that this was not the case. Additional models were also used in the validation, because to drive a drug-development programme it is essential that findings are sufficiently robust and applicable beyond the one narrow experimental model that may have been enough for publication. To address these concerns, when findings could not be reproduced, an attempt was made to contact the original authors, discuss the discrepant findings, exchange reagents and repeat experiments under the authors' direction, occasionally even in the laboratory of the original investigator. These investigators were all competent, well-meaning scientists who truly wanted to make advances in cancer research.

Reference : https://www.nature.com/articles/483531a

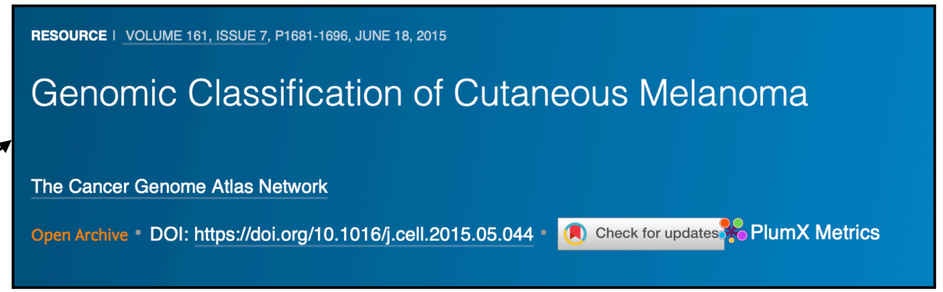
So before you dive into the data

only to generate pretty plots and AUC curves!!!!

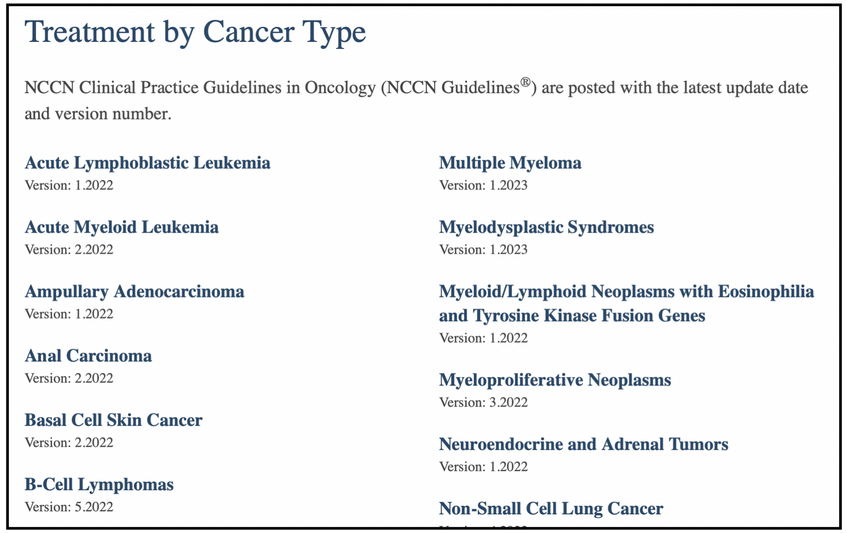
Best practise is to read **marker papers** for cancers of your interest



For more information, visit: TCGA Publications



Nearly all cancer types on TCGA provide marker papers which shed light on the nature of samples and molecular subtypes and key driver mutations



It is also vital to understand the clinical importance of known biomarkers and mutations that drive cancer. NCCN guidelines for all cancers provide information on targeted drug therapies available for different mutations at various stages of treatment

Case Study

Discovering potential prognostic biomarkers in lung adenocarcinoma patients on TCGA

Step 1: Setting up TCGABiolinks

Version > 4.0.0

Version 2022.07.2 Build 576

```
if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install("TCGAbiolinks")
```

⚠ **WARNING** : TCGABiolinks can take several minutes to download due to large number of dependencies, depending on the internet speed

TCGAbiolinks

DOI: 10.18129/B9.bioc.TCGAbiolinks

TCGAbiolinks: An R/Bioconductor package for integrative analysis with GDC data

Bioconductor version: Release (3.15)

The aim of TCGAbiolinks is : i) facilitate the GDC open-access data retrieval, ii) prepare the data using the appropriate pre-processing strategies, iii) provide the means to carry out different standard analyses and iv) to easily reproduce earlier research results. In more detail, the package provides multiple methods for analysis (e.g., differential expression analysis, identifying differentially methylated regions) and methods for visualization (e.g., survival plots, volcano plots, starburst plots) in order to easily develop complete analysis pipelines.

Author: Antonio Colaprico, Tiago Chedraoui Silva, Catharina Olsen, Luciano Garofano, Davide Garolini, Claudia Cava, Thais Sabedot, Tathiane Malta, Stefano M. Pagnotta, Isabella Castiglioni, Michele Ceccarelli, Gianluca Bontempi, Houtan Noushmehr

Maintainer: Tiago Chedraoui Silva <tiagochst at gmail.com>, Antonio Colaprico <axc1833 at med.miami.edu>

Citation (from within R, enter `citation("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, Noushmehr H (2015). " TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data." *Nucleic Acids Research*. doi: 10.1093/nar/gkv1507, http://doi.org/10.1093/nar/gkv1507.

Silva, C T, Colaprico, Antonio, Olsen, Catharina, D'Angelo, Fulvio, Bontempi, Gianluca, Ceccarelli, Michele, Noushmehr, Houtan (2016). "TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages." *F1000Research*, **5**.

Mounir, Mohamed, Lucchetta, Marta, Silva, C T, Olsen, Catharina, Bontempi, Gianluca, Chen, Xi, Noushmehr, Houtan, Colaprico, Antonio, Papaleo, Elena (2019). "New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEx." *PLoS computational biology*, **15**(3), e1006701.

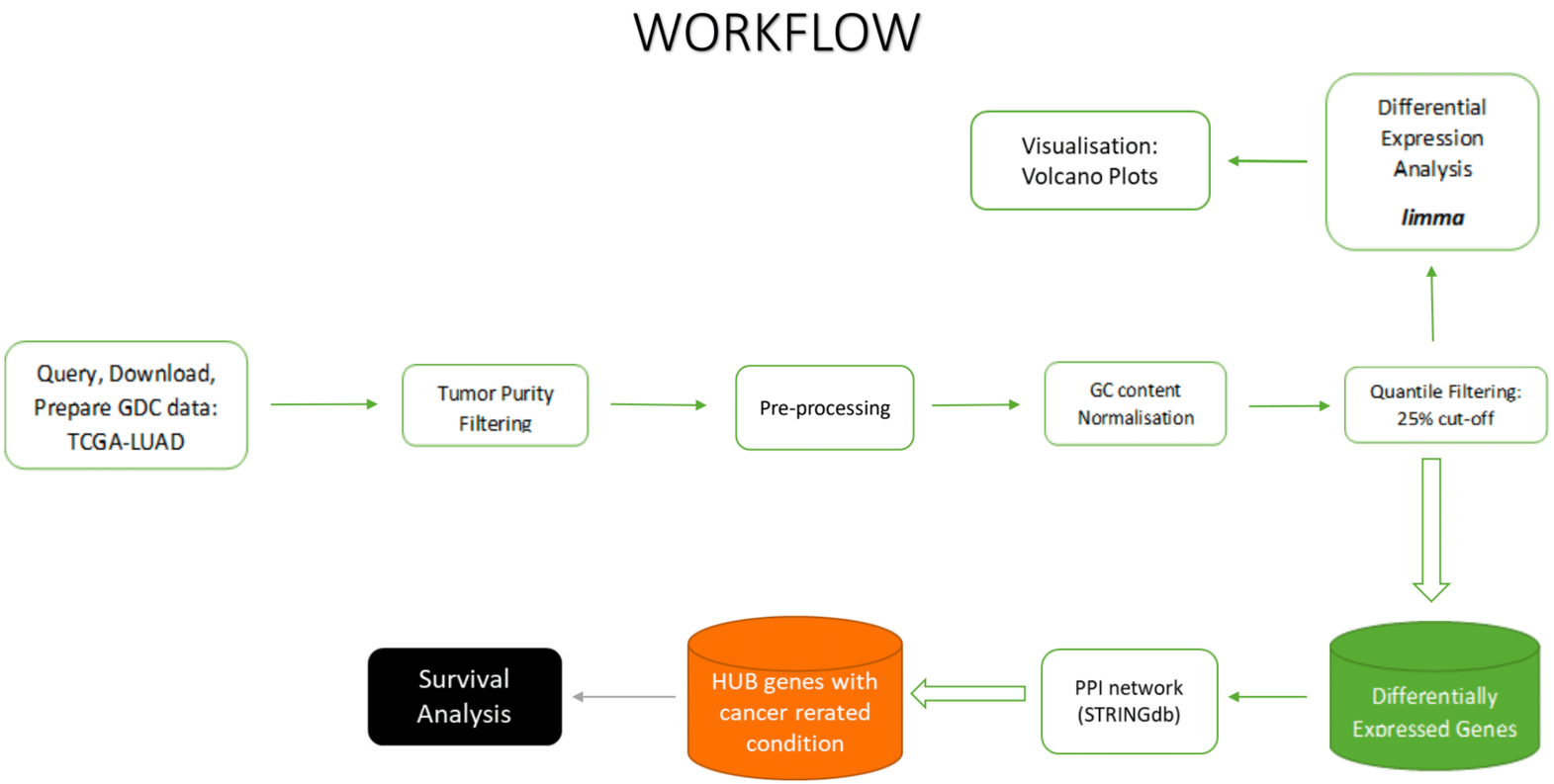
Installation

To install this package, start R (version "4.2") and enter:

```
if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install("TCGAbiolinks")
```

Overall workflow design

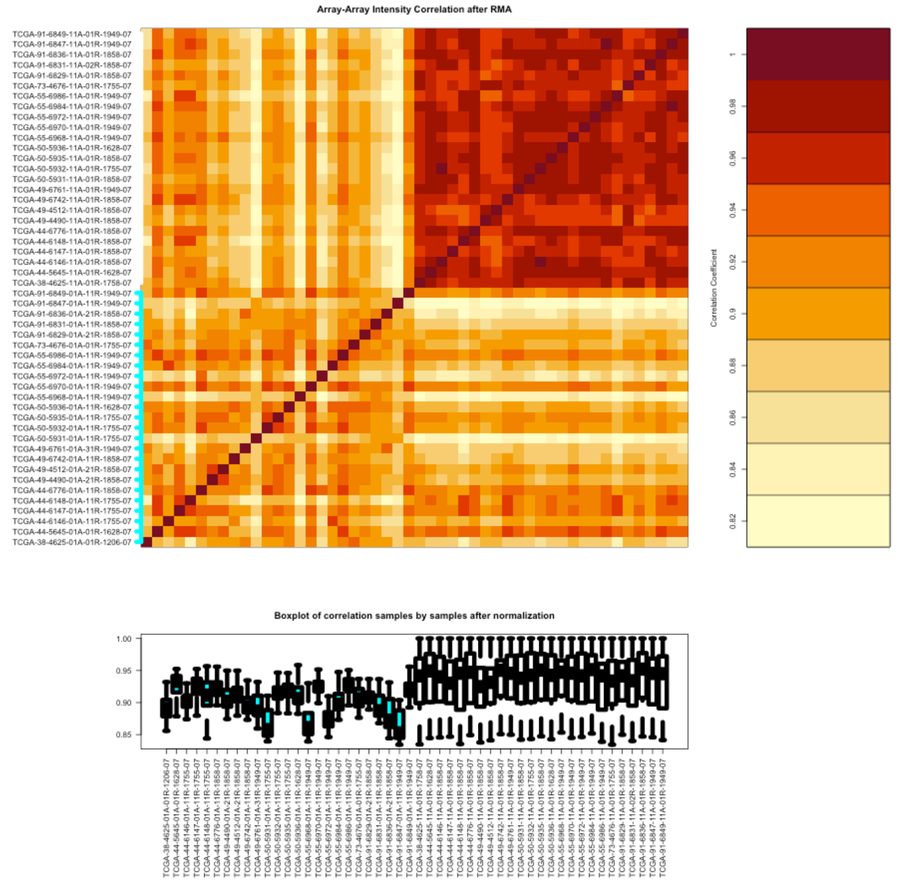


Data Download and Preprocessing

574 patient samples for TCGA-LUAD

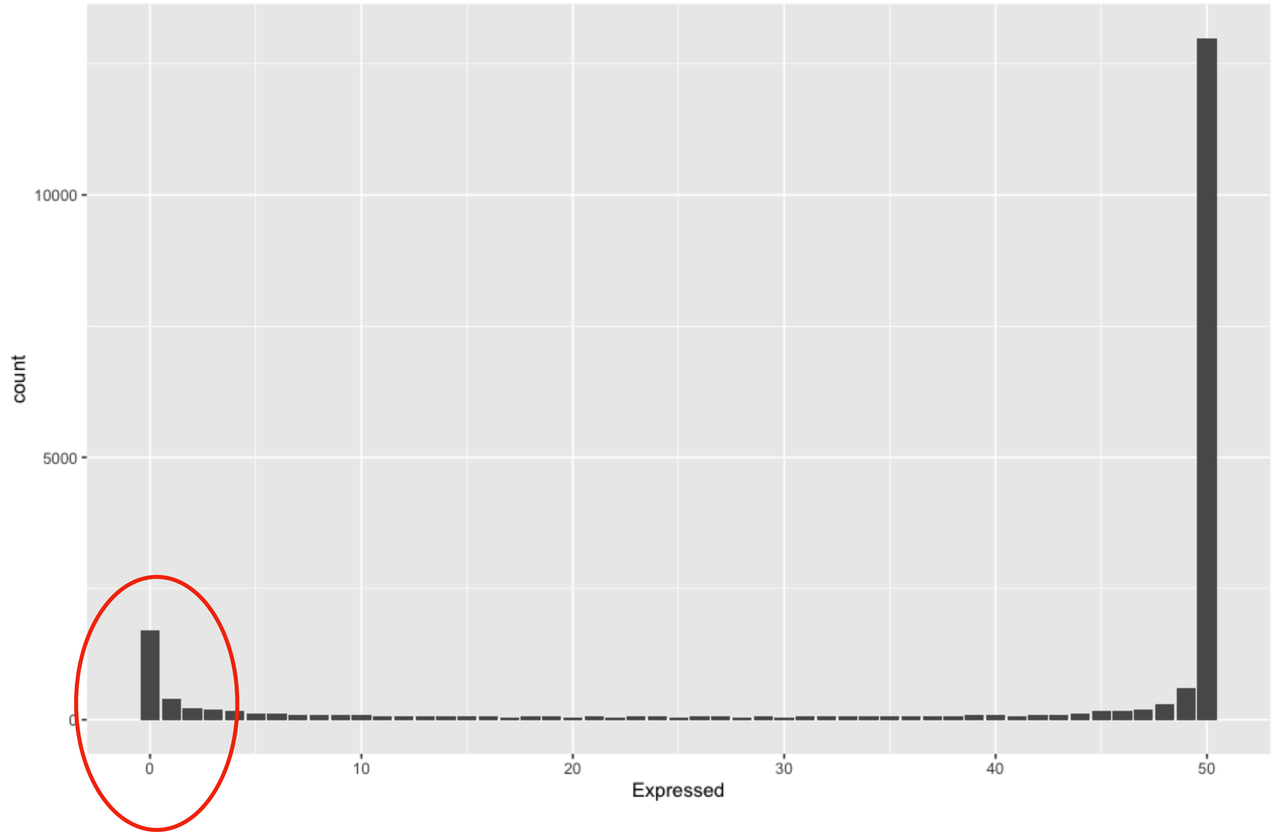
116 paired samples found (58 Tumor + 58 Normal)

25 samples containing over 60% tumor content (25 Normal + 25 Tumor)



Quality Control

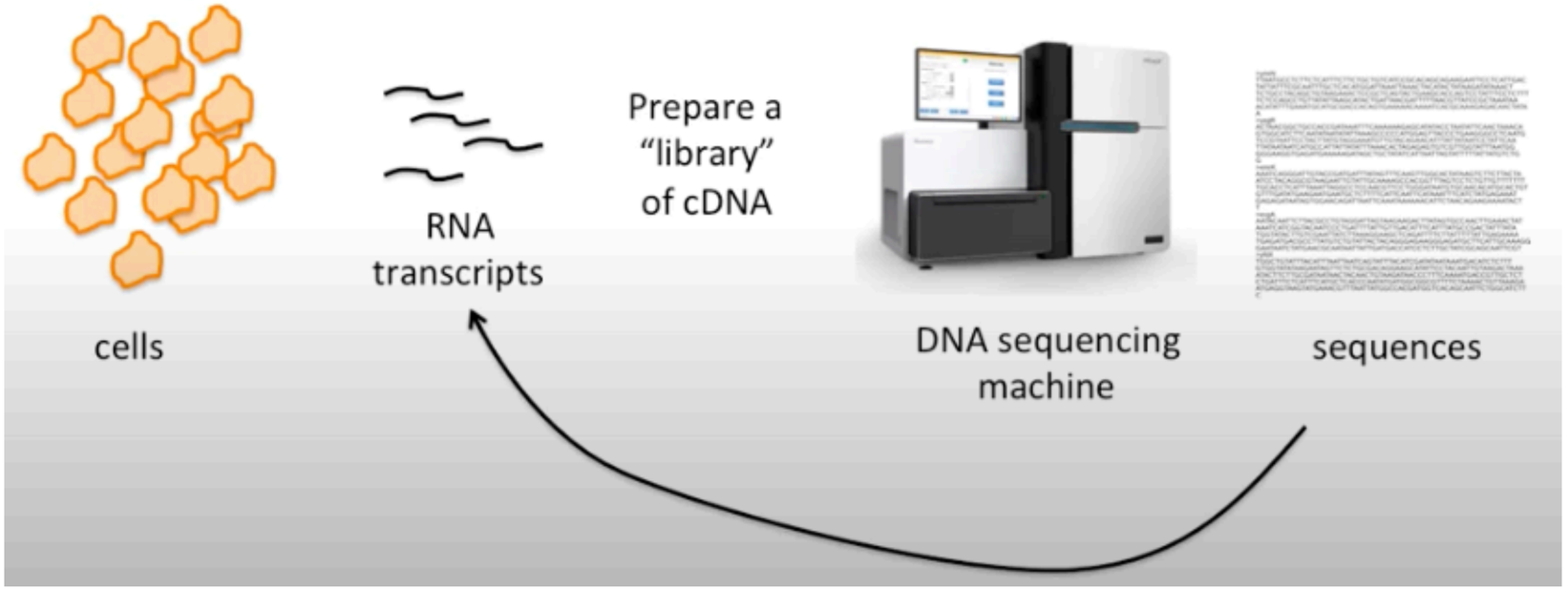
Step 1 : Removing genes which are not expressed in 5 or more samples



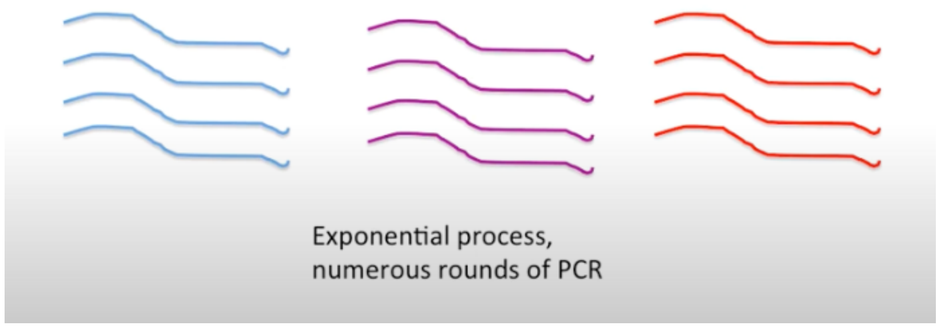
Genes that are not expressed in any samples or very few samples

Not a necessary step, but can reduce the size of the Data you use for downstream analysis and speed up computation

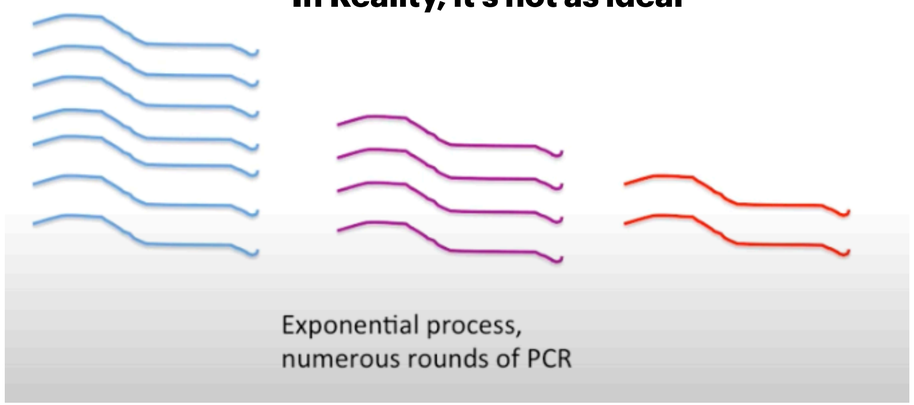
Step 2: GC Content Normalisation



An ideal scenario

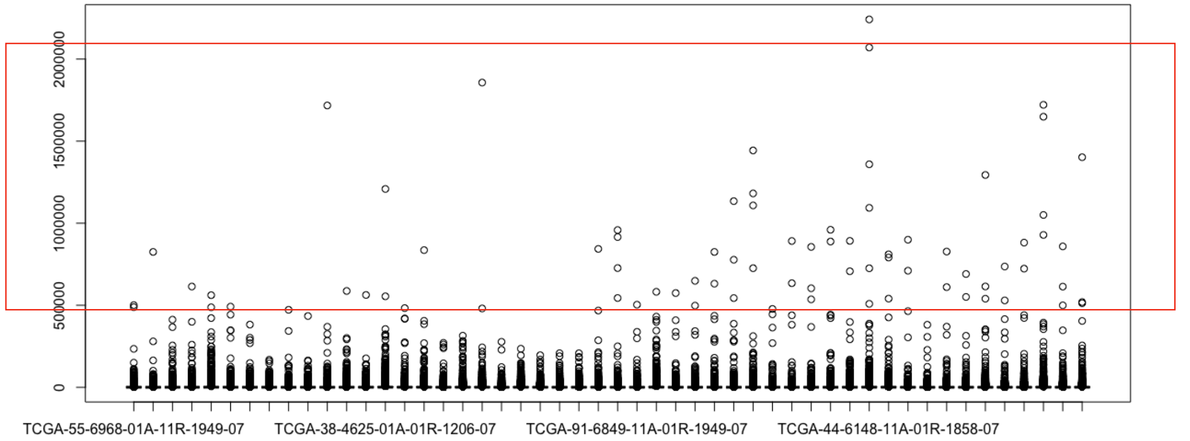


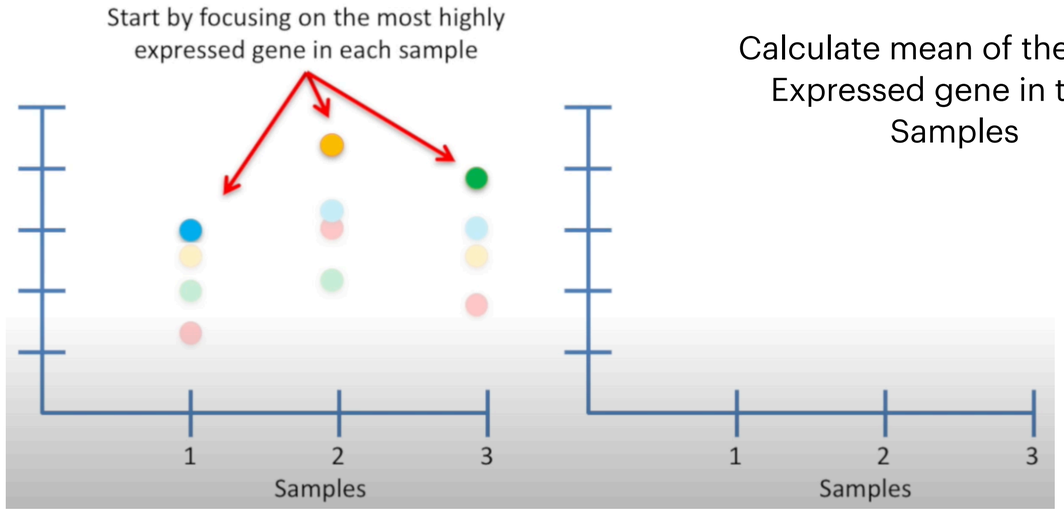
In Reality, it's not as ideal



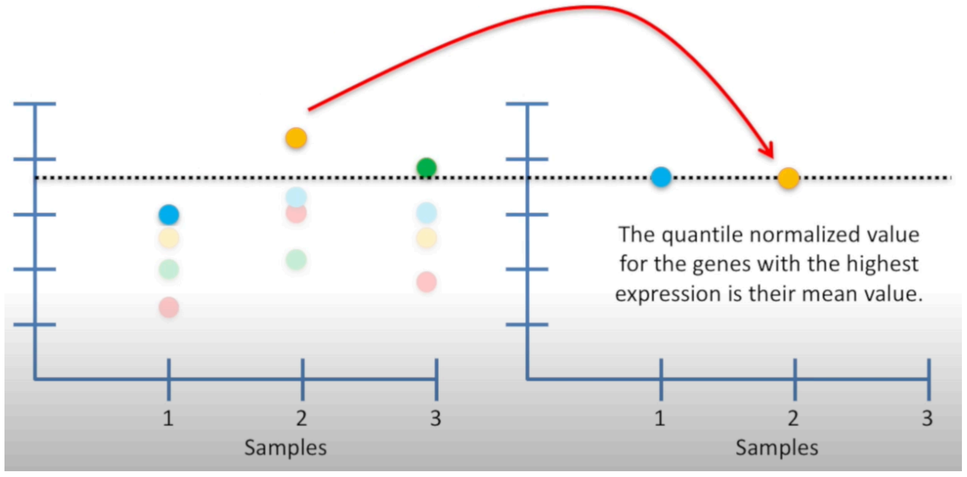
Step 3: Quantile Normalisation

Such large counts largely Influence important signals



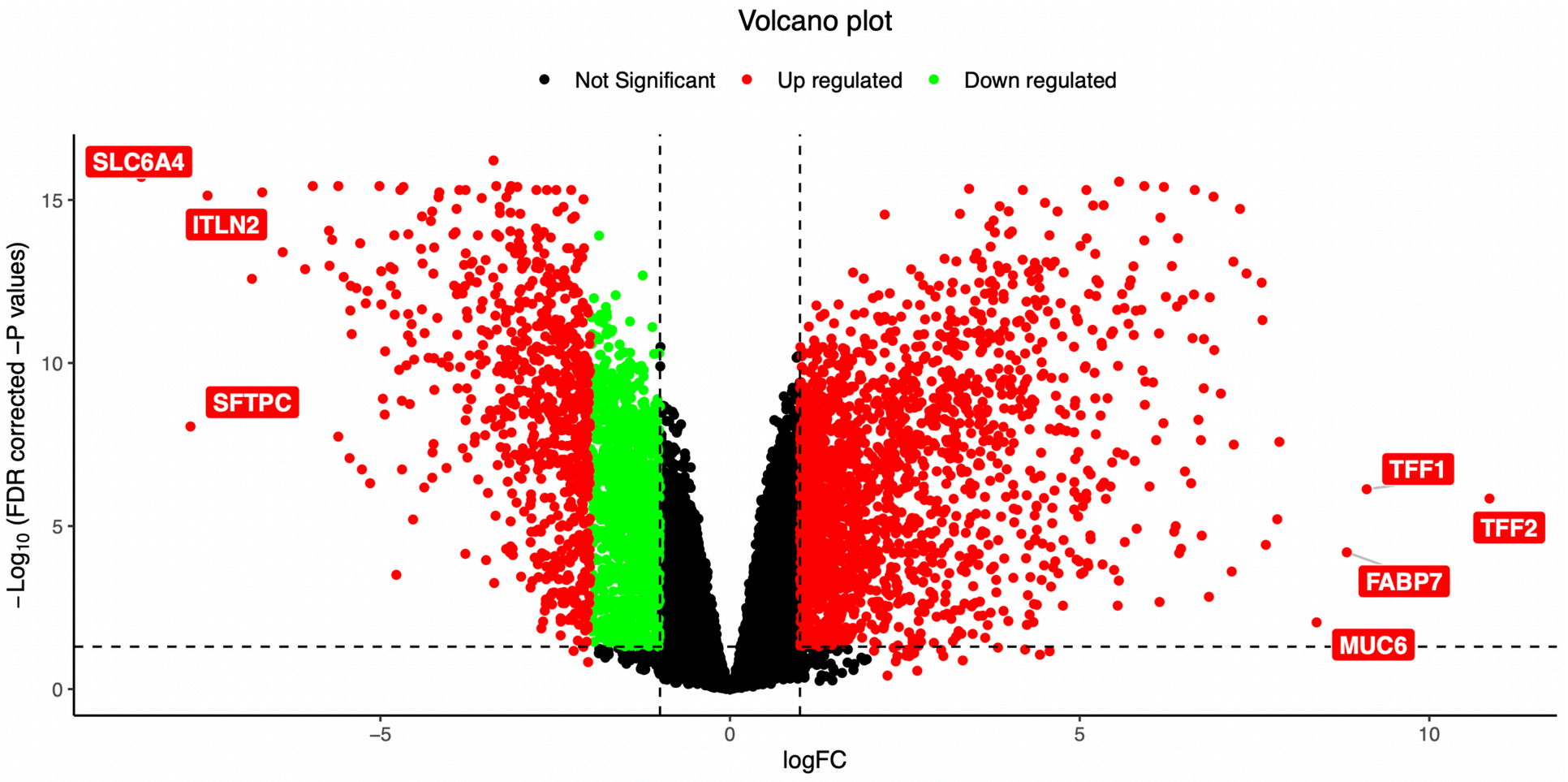


Calculate mean of the highly Expressed gene in those Samples

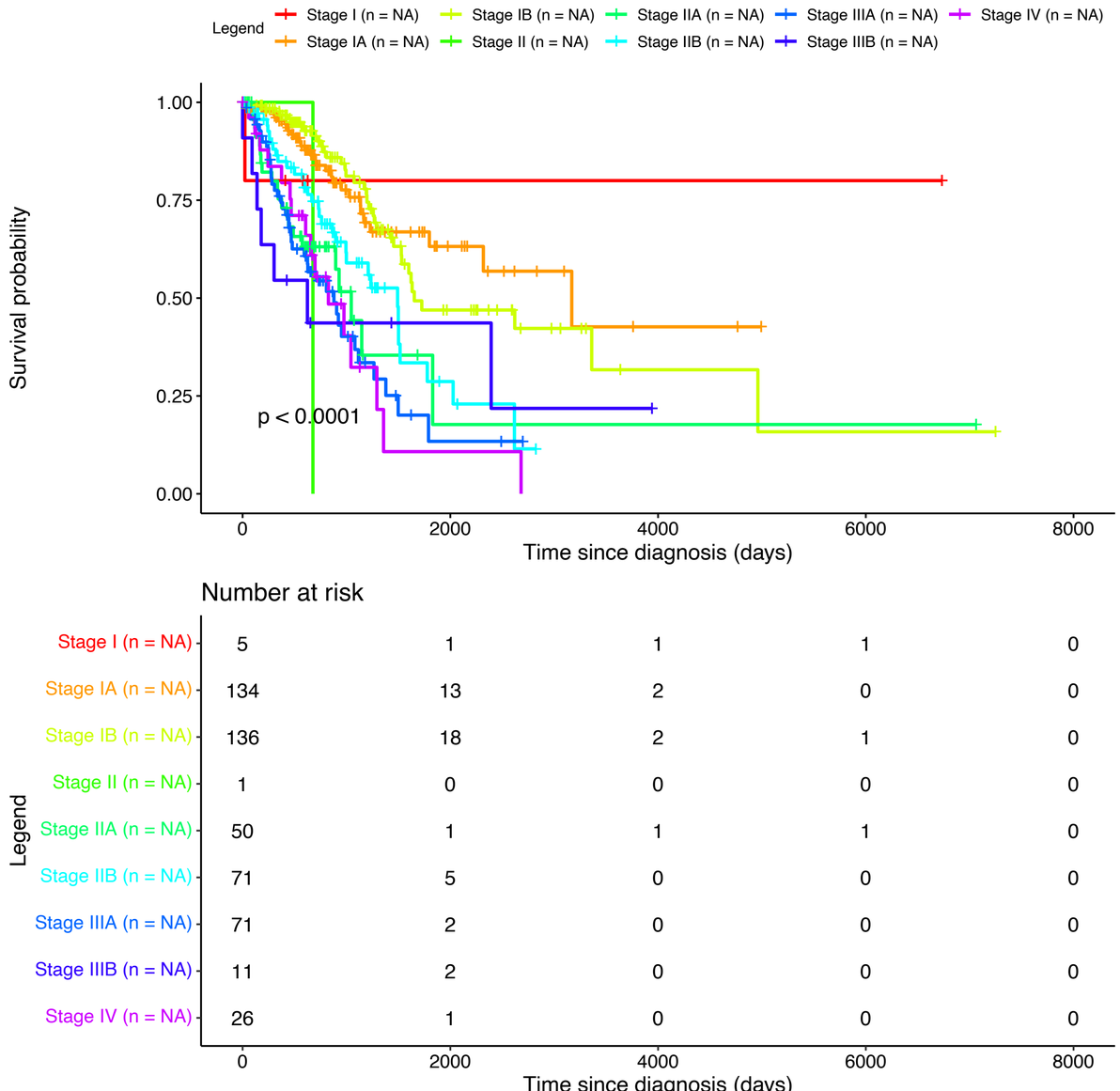


Differential Gene Expression Analysis

Limma



Survival Analysis



Take Home Assignment

- Using the list of differentially expressed genes, filter out genes that do not show any differential expression
- Apply an adj.P.value threshold of 0.05 and logFC threshold of |2|
- Find co-expression modules using the list of significant genes. HINT: See workflow diagram
- Pick a module which has genes that are mostly associated with Lung Adenocarcinoma progression. HINT: Use DisGeNET to locate genes that are associated with the above cancer type.
- Perform a survival analysis using TCGABiolinks package or any other package of your choice for the genes of interest from the above analysis. For TCGABiolinks use TCGA_SurvivalKM() function to perform survival analysis using a gene-list.

NOTE: Github link for the code will be provided after the lecture

Thank You

For more queries, contact: pawan12394@gmail.com