Big Data in Medicine

Thomas Mohr

Institute for Analytical Chemistry - University of Vienna, Vienna.
Center of Cancer Research - Medical University of Vienna, Vienna.
ScienceConsult - DI Thomas Mohr KG, Guntramsdorf.

SS2024

First lecture

A few administrative pieces of information

This lecture will be structured into the following parts.

- What is Big Data?
- General principles for developing Big Data Projects
- The biological foundations of our analytes
- How to get data
- Data preprocessing
- Determination of differentially expressed features
- The biological context of differentially expressed features
- Network-based analysis methods
- Data integration using multi-omics
- A small amount of machine learning

I will organise an excursion into a molecular biology lab.

The big question - the final exam

The final exam will consist of two parts.

- The coding of a Big Data workflow
- A presentation that mimics actual presentations you will have
- A short question by me about the presentation.

What are Big Data?

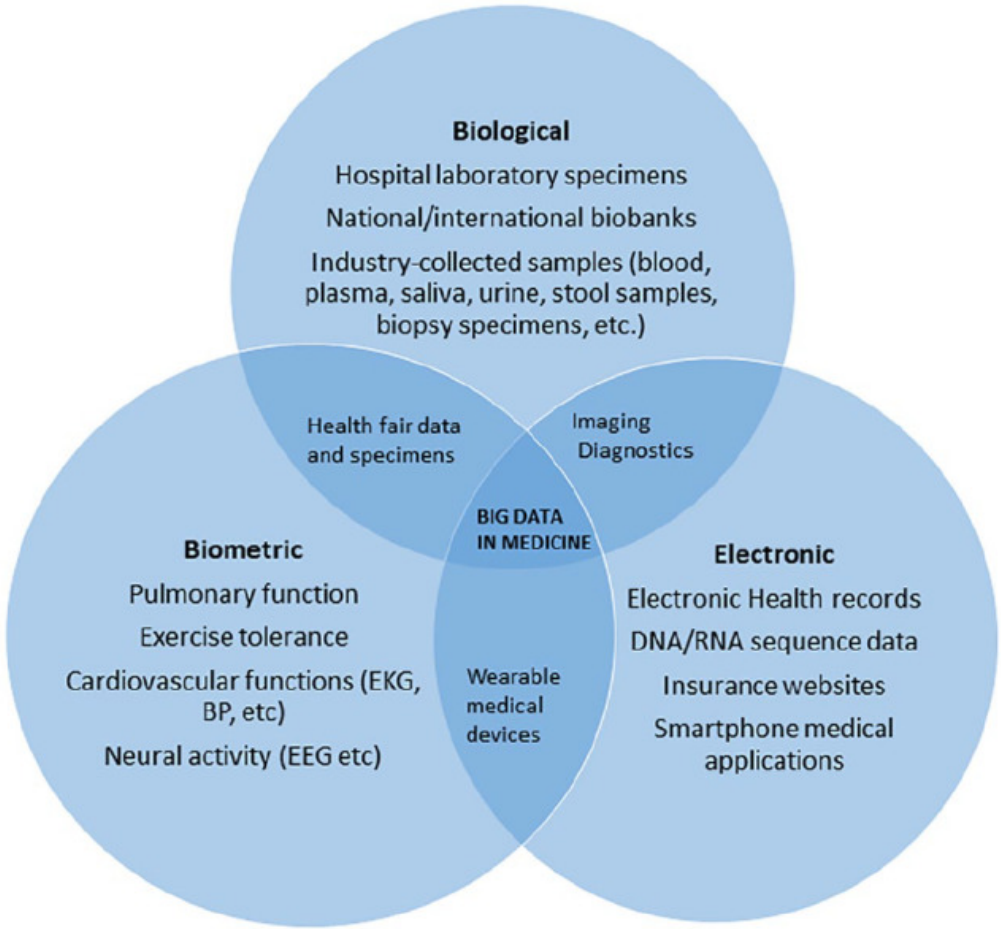
The term “Big Data” was coined in the 90ties to describe data sets too large to be processed with standard software.

These data are defined as

- high-volume,
- high-velocity
- and a high variety of information

That requires specific scientific technologies for interpretation and analysis.

How does Big Data fit into the triade of Medical Data?



Why are Big Data so important?

Example I: Deriving Gene Signatures to predict cancer outcome in Glioblastoma Multiforme

Verhaak RG, et al.: Cancer Genome Atlas Research Network. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterised by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell. 2010 Jan 19;17(1):98-110. doi: 10.1016/j.ccr.2009.12.020. PMID: 20129251; PMCID: PMC2818769.

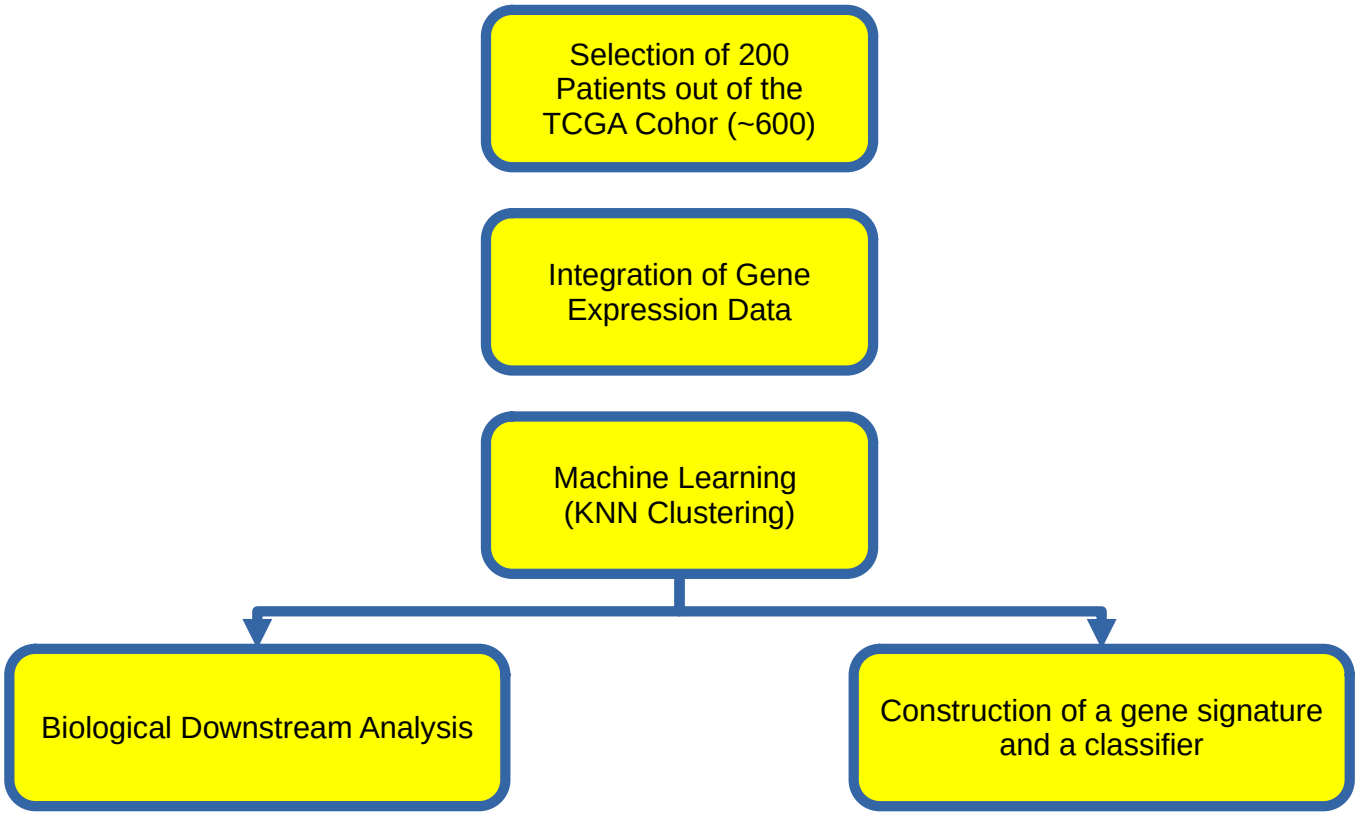
What is glioblastoma multiforme?

Glioblastoma multiforme (GBM) is adults' most common form of malignant brain cancer.

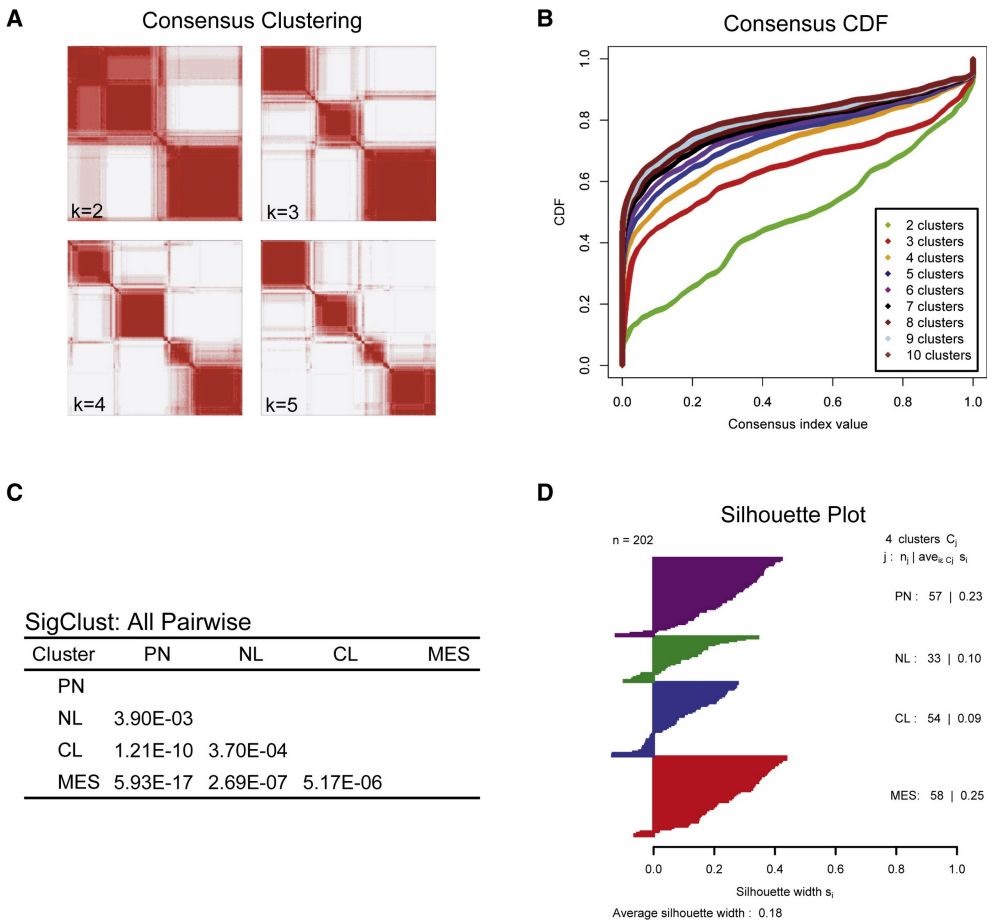
It is characterised by

- Short overall survival.
- Poor access by surgery.
- Poor access to drugs.

What did Verhaak et al. do?

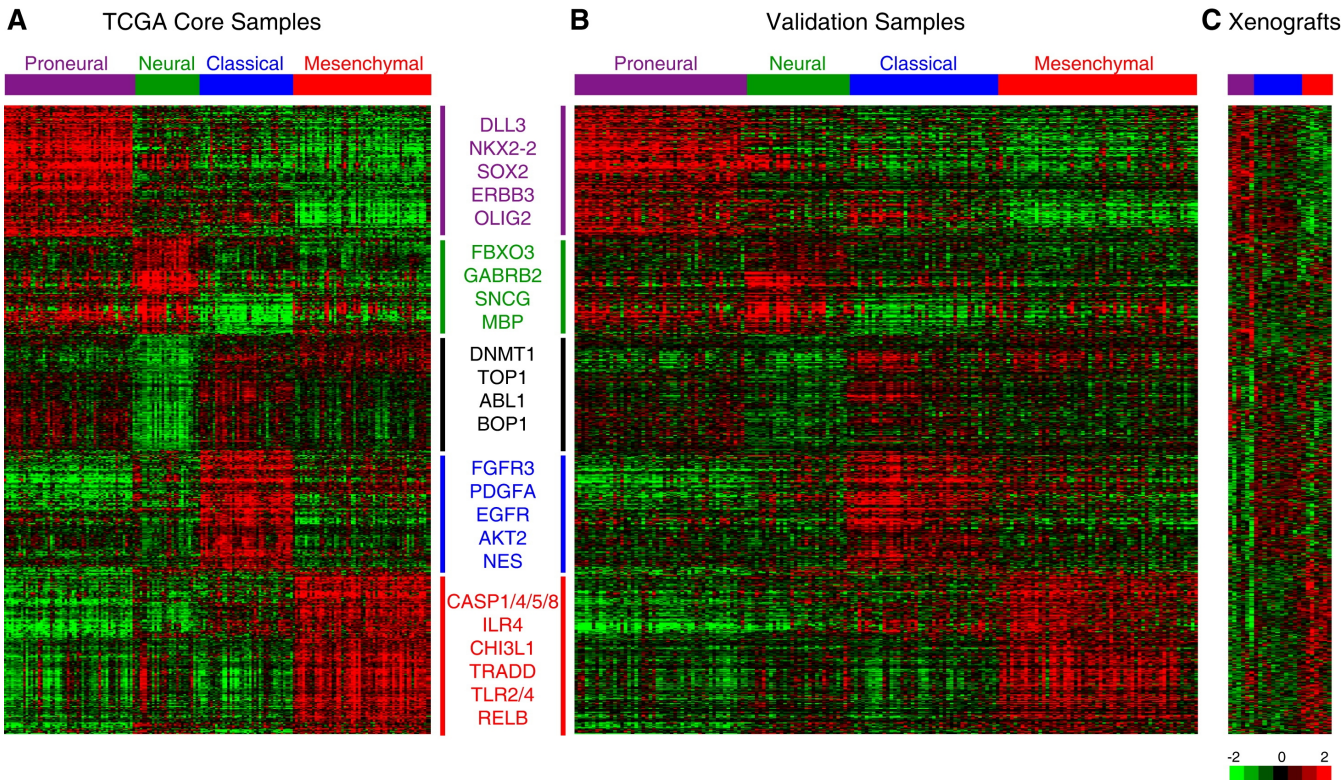


The first step - apply some machine learning to find structure in the data



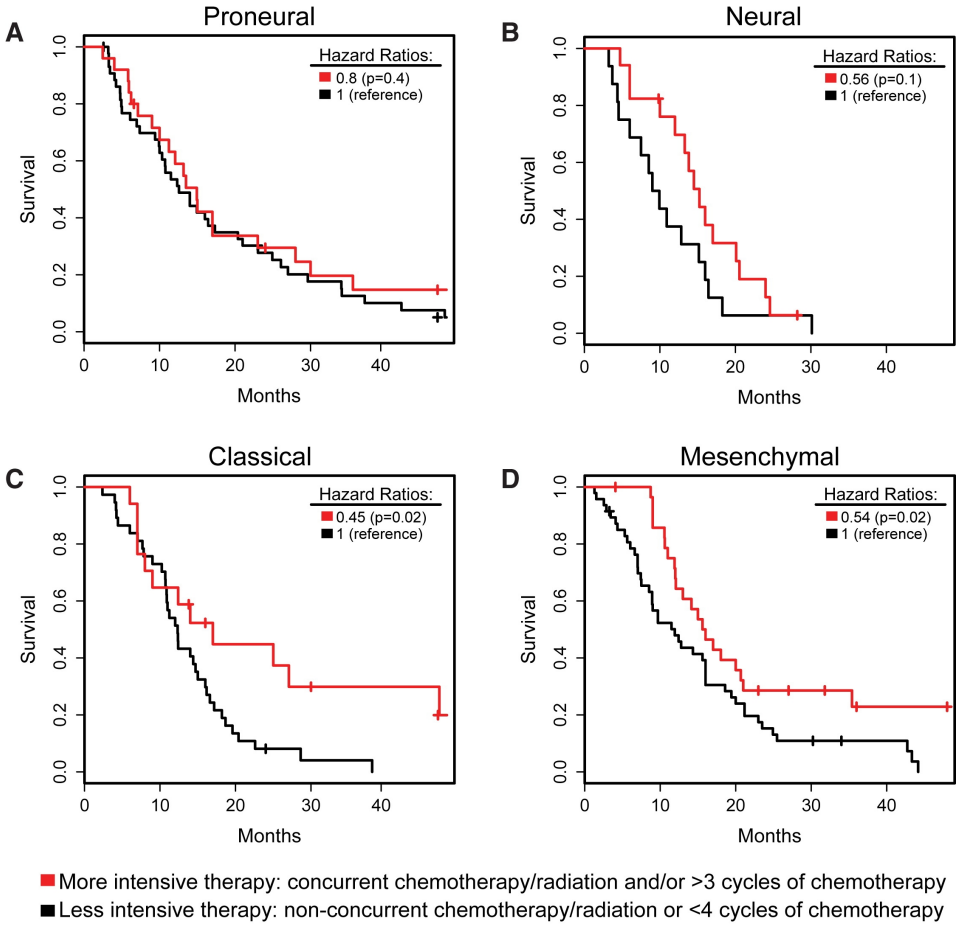
Verhaak RG, et al.: Cancer Genome Atlas Research Network. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterised by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell. 2010 Jan

The second step - validate the subtypes in an independent data set



Verhaak RG, et al.: Cancer Genome Atlas Research Network. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterised by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell. 2010 Jan

The third step - does this subtyping impact patient survival?



Verhaak RG, et al.: Cancer Genome Atlas Research Network. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterised by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell. 2010 Jan

Example II: Mutations in Melanoma

What is Melanoma?

Melanoma is a skin cancer that develops from melanocytes, the cells responsible for skin pigmentation.

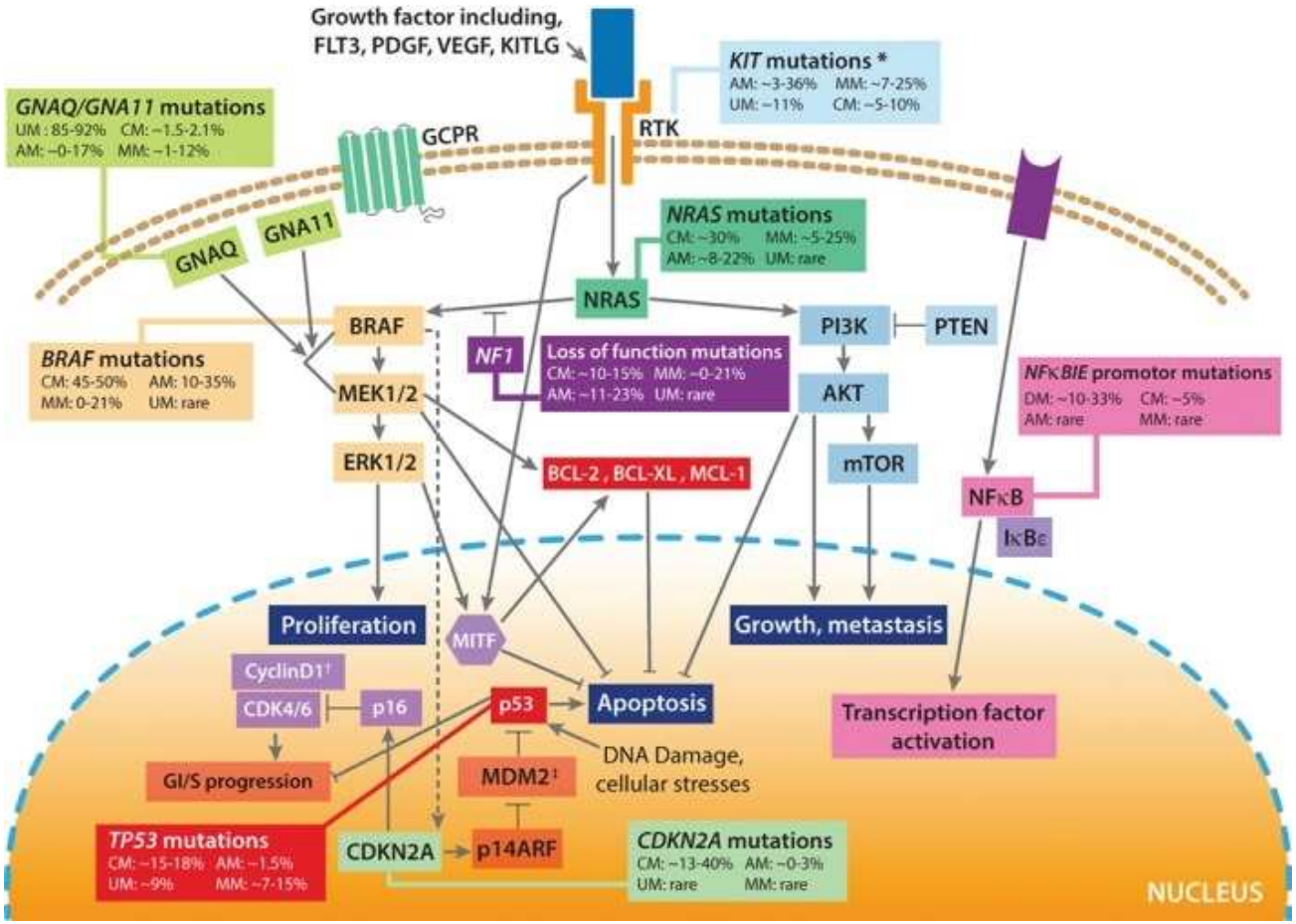
One characteristic of Melanoma is the existence of so-called *driver* mutations, which significantly drive melanoma development.

These mutations include:

- Activation of the BRAF gene
- Activation of the KRAS gene
- Loss of NF1 gene activity

What does that mean?

The impact of mutations in Melanoma



Rabbie R, Ferguson P, Molina-Aguilar C, Adams DJ, Robles-Espinoza CD. Melanoma subtypes: genomic profiles, prognostic molecular markers and therapeutic possibilities. J Pathol. 2019 Apr;247(5):539-551. doi: 10.1002/path.5213. Epub 2019 Feb 15. PMID: 30511391; PMCID: PMC6492003.

What does the mean in terms of treatment?

Mutation	Percentage (%)	Clinical features or other comments
BRAF V600	40–50	Confers susceptibility to BRAF/MEK inhibitors, more common on intermittent sun-exposed skin
NRAS	15–20	Poor prognosis, may have a higher response to immunotherapy
NF1	10–15	More common on sun-exposed skin, may have a higher response to immunotherapy
KIT	1–2	Confers susceptibility to KIT inhibitors, more common in mucosal (15–20%) and acral melanomas (15–20%)
Atypical BRAF (non-V600)	4–5	May confer susceptibility to MEK or RAF inhibitors
GNAQ/GNA11	80–90(uveal)	
TERT promoter	40–50	Poor prognosis, UV-mediated mutation
CDKN2A	25–35	Deep deletions more common than mutations
PTEN	4–8	May correlate with immune resistance, deep deletions more common than mutations

Example III - inferring potential therapeutic targets

Background information

Hepatocellular carcinoma (HCC) is

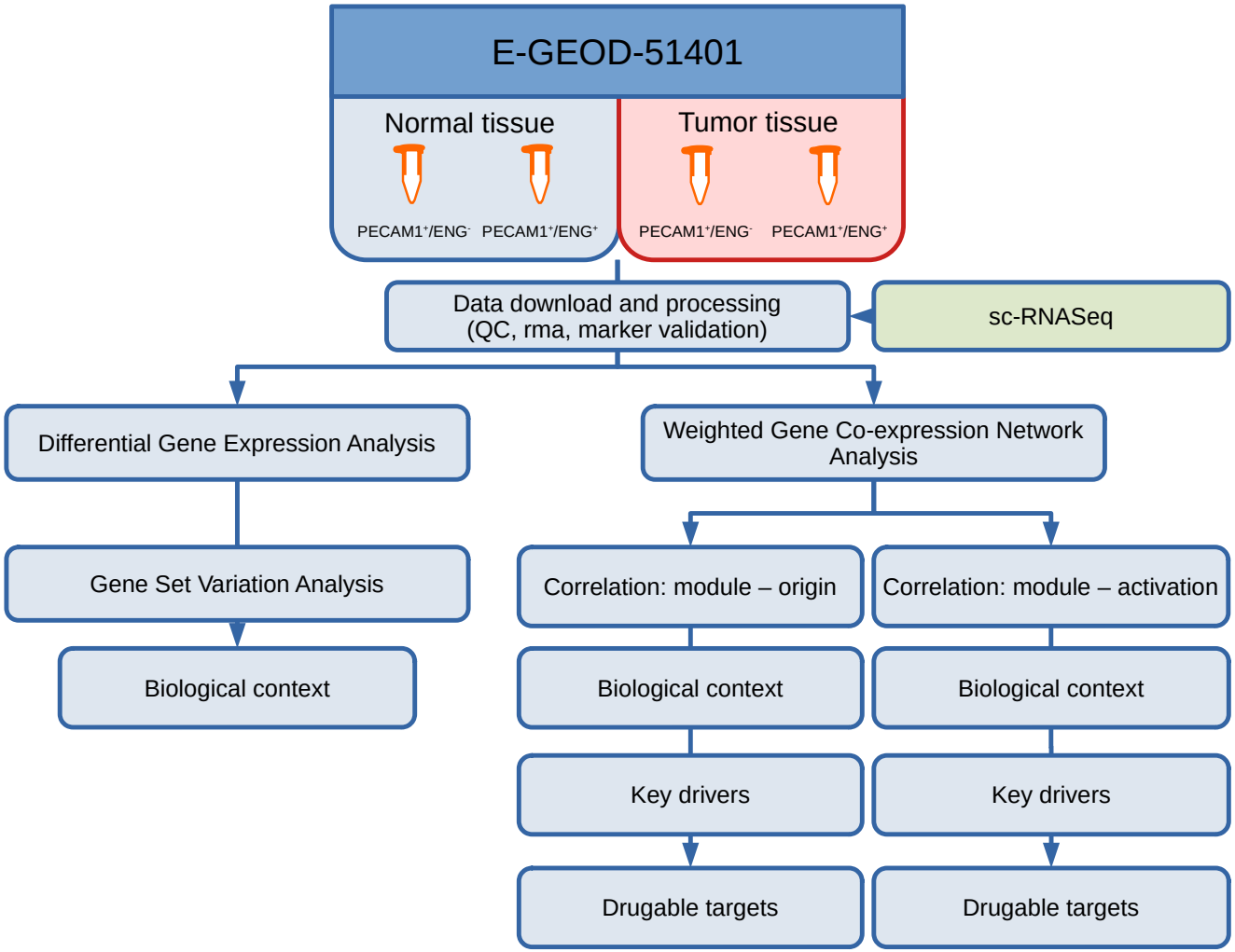
- sixth most common cancer and the third most common cause of cancer-related death,
- affecting roughly 500,000 people worldwide each year
- arises in a multistep process from preexisting cellular lesions.
- chemotherapeutic options are often limited
- prognosis after primarily curative surgery is usually poor.

It is, therefore, necessary to identify potential new chemotherapeutic targets.

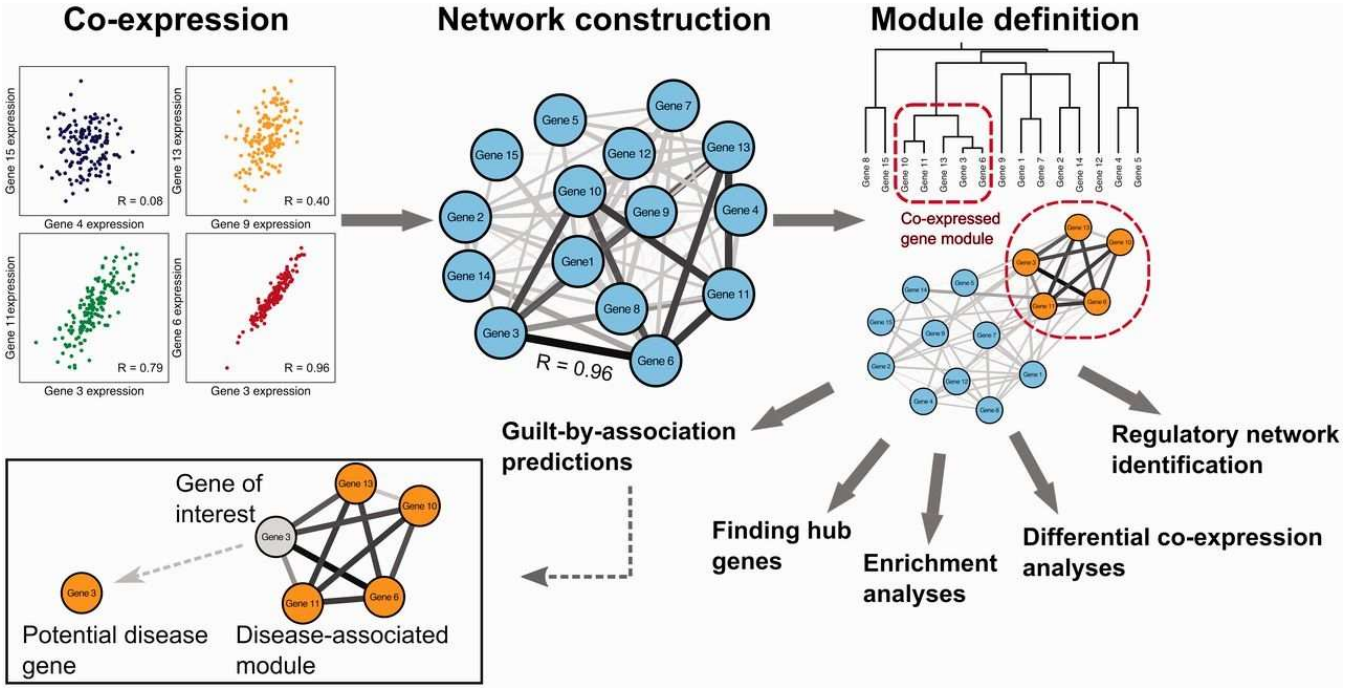
Which targets may be present in HCC?

Mohr, T.; Katz, S.; Paulitschke, V.; Aizarani, N.; Tolios, A. Systematic Analysis of the Transcriptome Profiles and Co-Expression Networks of Tumour Endothelial Cells Identifies Several Tumour-Associated Modules and Potential Therapeutic Targets in Hepatocellular Carcinoma. Cancers 2021, 13, 1768. https://doi.org/10.3390/cancers13081768

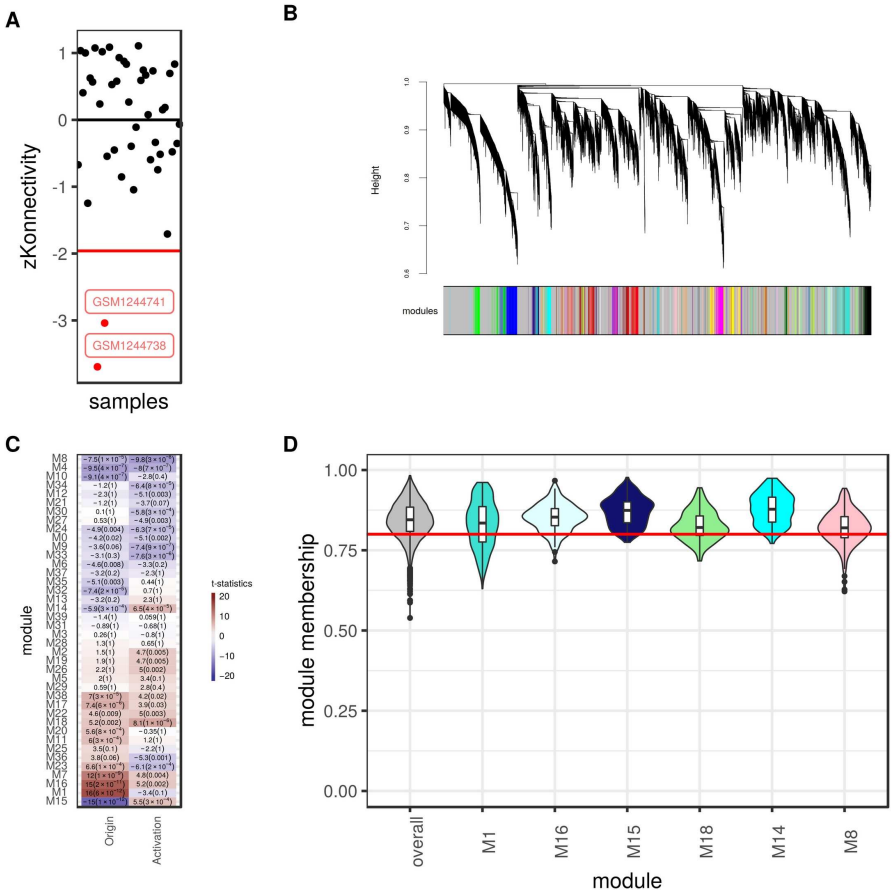
Study design



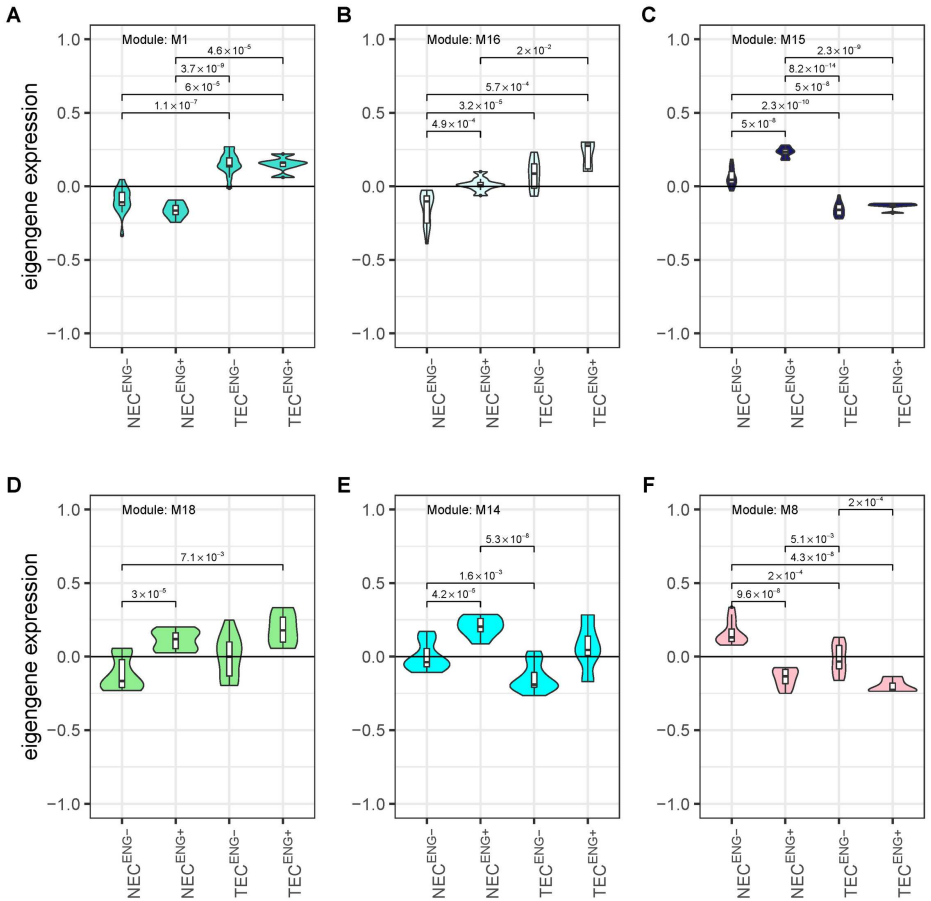
Determination of central genes



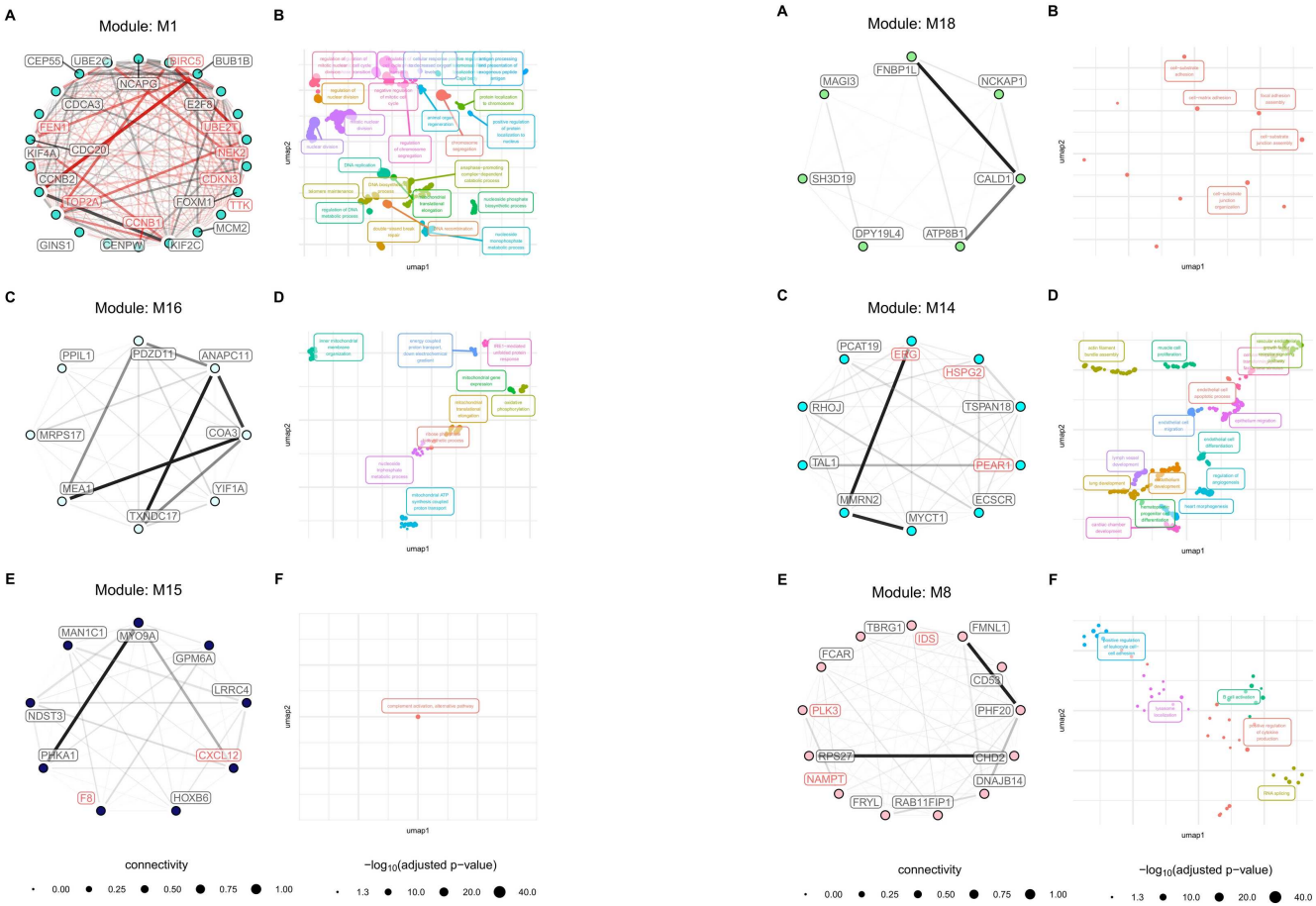
Which modules did we identify?



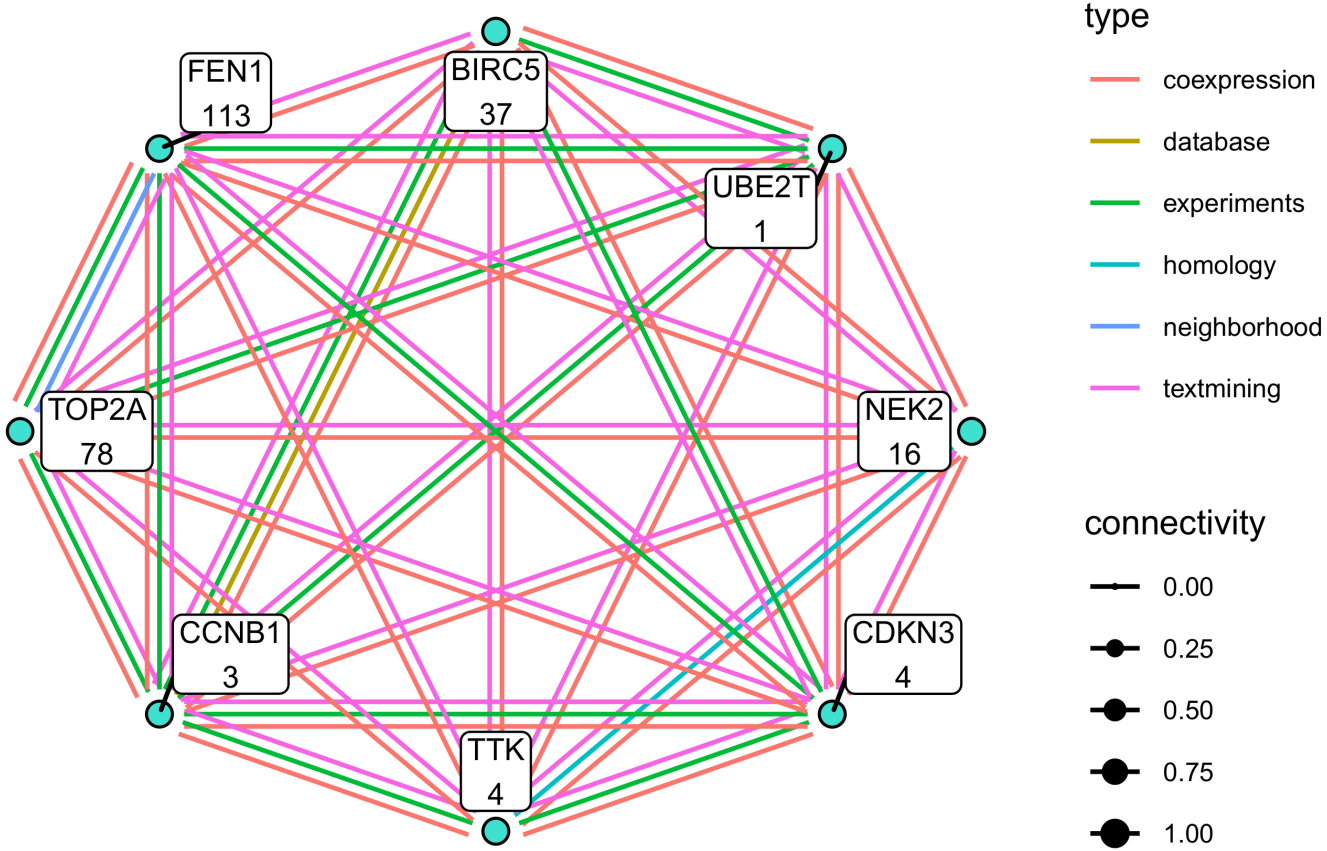
Which core genes and biological functions did these modules reveal?



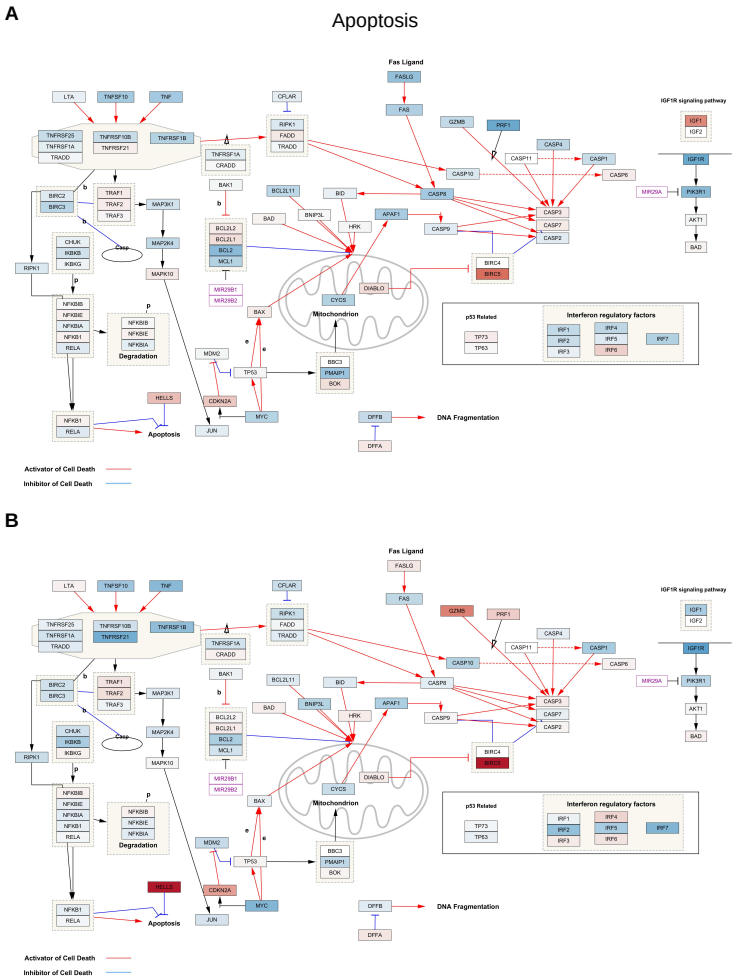
Which core genes and biological functions did these modules reveal?



Is there an underlying protein-interaction network?



How does BIRC5 fit into biological pathways?



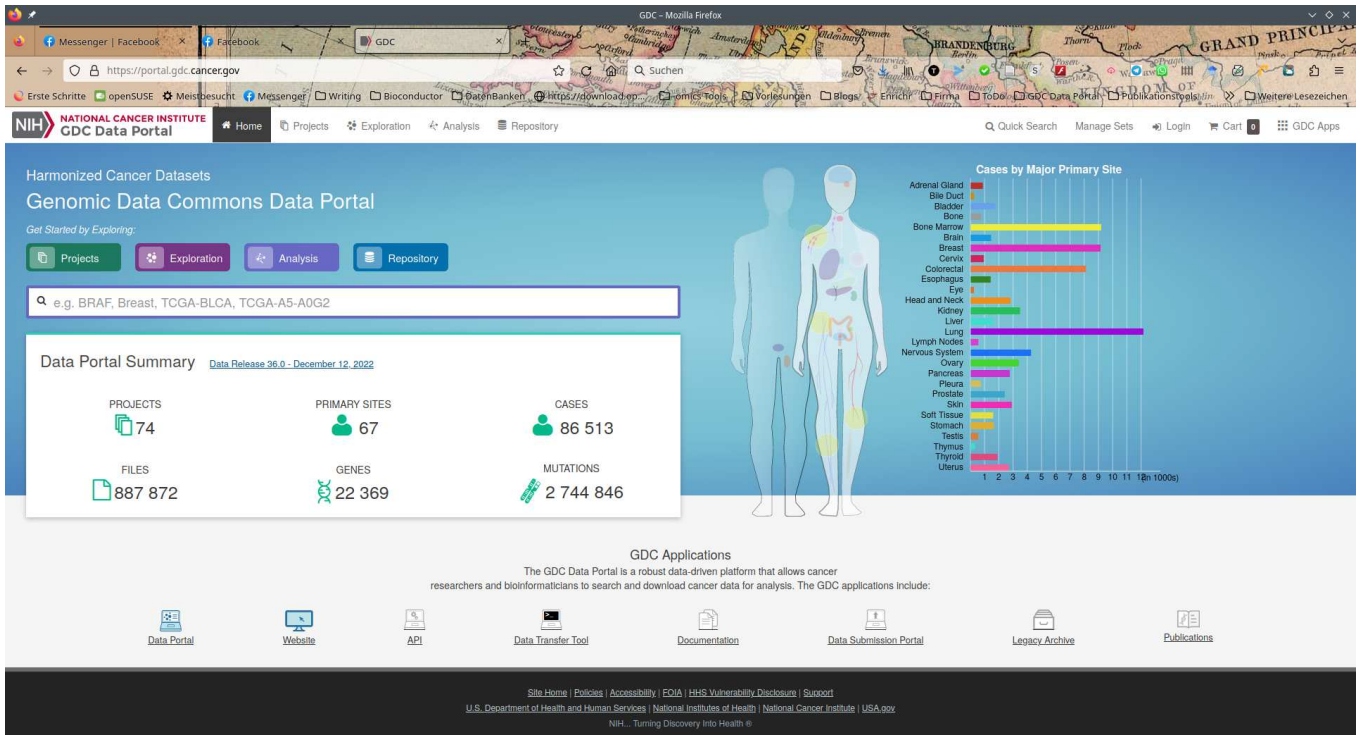
Conclusion

Big Data allow insight into biological processes in an unprecedented manner. They allow us to derive and combine information from various sources to detect structure in data and derive therapeutically and diagnostically relevant information.

Where do I access Big Data?

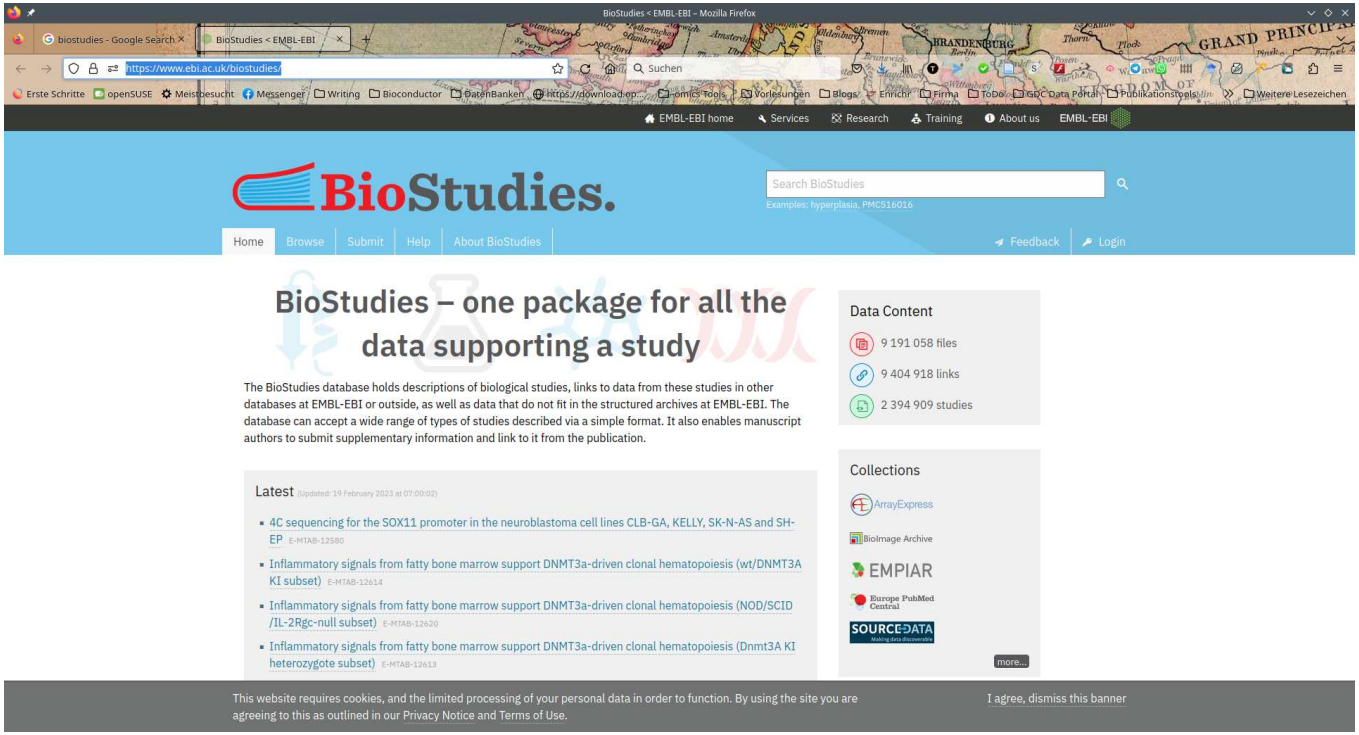
Example: Cancer

Let us have a look at some examples - the GDC portal:



Task: Explore the GDC portal (15 to 30 mins) https://portal.gdc.cancer.gov/

Example: General -omics data - the Biostudies portal



Task: Explore the Biostudies portal (15 to 30 mins) https://www.ebi.ac.uk/biostudies/

Where do I get the tools to work with Big Data?

Main tools to use in Big Data Science

- R and Bioconductor(https://bioconductor.org/): Bioconductor is a collection of tools focused on -omics data analysis.
- Python in combination with Anaconda (https://www.anaconda.com/): Anaconda is a collection of scientific libraries for python and R.