QUEEN'S UNIVERSITY BELFAST
SCHOOL OF ELECTRONICS, ELECTRICAL ENGNIEERING AND COMPUTER SCIENCE

# Data Analysis & Visualisation

**CSC3062**

**BEng (CS & SE), MEng (CS & SE), BIT & CIT**

Dr Reza Rafiee

23rd September 2019

# What is data analysis?

**Data analysis** is a process of

- Inspecting,
- Cleaning,
- Transforming,
- And modelling data

With the **goal** of

- Discovering useful information,
- Informing conclusions,
- & Supporting decision-making

# Data analysis and visualisation

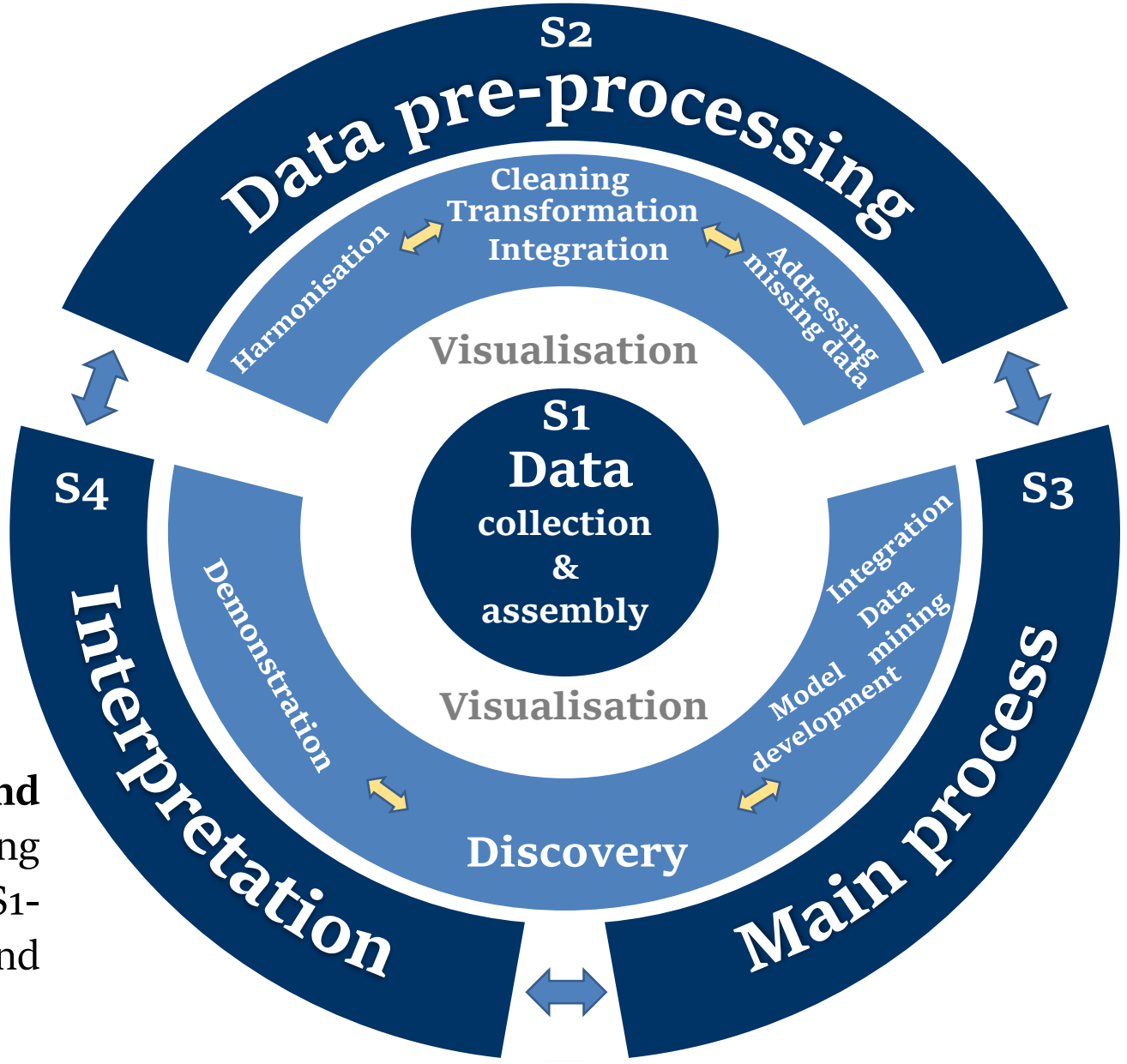



**Figure 1.1 | Data analysis and visualisation topic**, showing the four integrated stages (S1-S4), their key components and inter-relationships.

# Three main sources of data

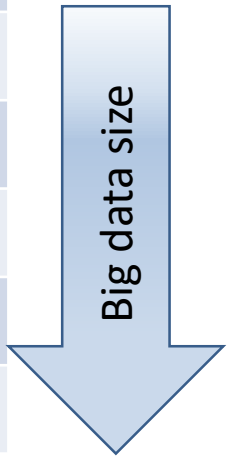

**Figure 1.2** | Three main sources of data

https://wearesocial.com/blog/2018/01/global-digital-report-2018

# Units of data & information

- **The smallest addressable unit in a computer memory is the *byte*. A byte is equal to 8 consecutive binary digits or *bits*.**

| Symbol | Prefix | SI* meaning | Binary meaning | Size difference |
|---|---|---|---|---|
| K | kilo | $10^3 = 1000^1$ | $2^{10} = 1024^1$ | 2.40% |
| M | mega | $10^6 = 1000^2$ | $2^{20} = 1024^2$ | 4.86% |
| G | giga | $10^9 = 1000^3$ | $2^{30} = 1024^3$ | 7.37% |
| T | tera | $10^{12} = 1000^4$ | $2^{40} = 1024^4$ | 9.95% |
| P | peta | $10^{15} = 1000^5$ | $2^{50} = 1024^5$ | 12.59% |
| E | exa | $10^{18} = 1000^6$ | $2^{60} = 1024^6$ | 15.29% |
| Z | zetta | $10^{21} = 1000^7$ | $2^{70} = 1024^7$ | 18.06% |
| Y | yotta | $10^{24} = 1000^8$ | $2^{80} = 1024^8$ | 20.89% |




*SI: International System of Units

# Data - image & video

Aim: detecting and extracting interest regions from an image

**a**

How?

**b**

**Data (only colour)**

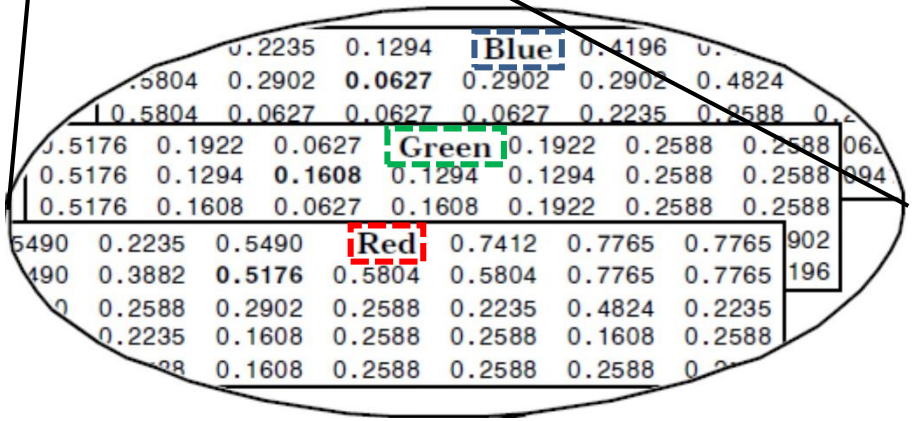



**Figure 1.3** | **a**, A colour image including an interest region (i.e., butterfly). **b**, Interest region extracted by a computer program

**What is a content based image retrieval (CBIR) system?**

https://www.sciencedirect.com/science/article/pii/S0031320313001179

# Data - image & video

**Figure 1.4** | 2D visualisation of a large image dataset of various objects using t-SNE dimensionality reduction technique

https://github.com/genekogan

# Data - image & video

- ~ 50 million creators are generating and uploading videos on YouTube™ repository.
- Almost 5 billion videos are watched on YouTube™ repository every single day.


**Figure 1.5** | Example of "number of views" in a YouTube channel (per minute - live)

# Data - image & video

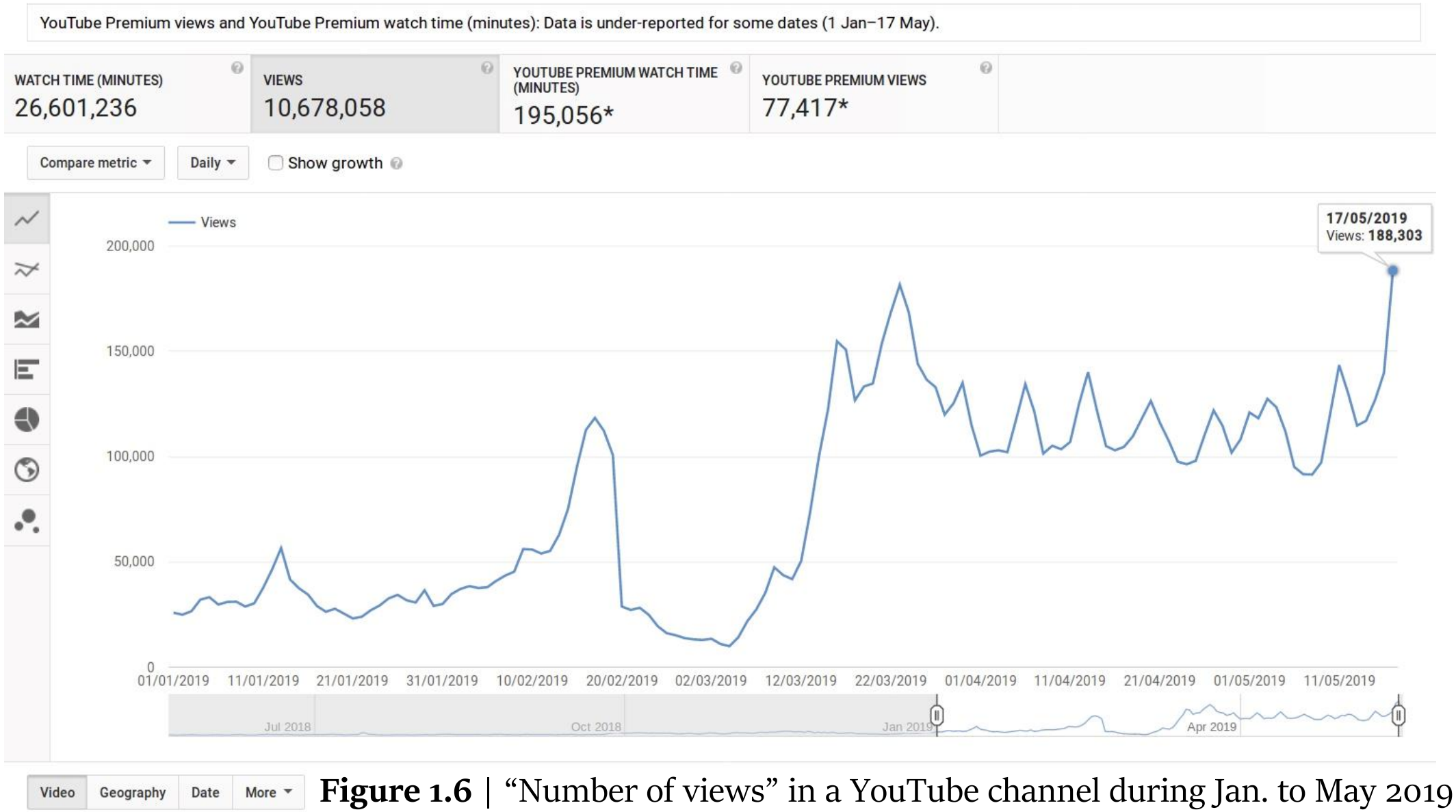



**Figure 1.6** | "Number of views" in a YouTube channel during Jan. to May 2019

# Data in medicine: clinical vs. omics

- **Clinical data** (these data are either collected during *the course of ongoing patient care* or as part of a formal *clinical trial* program.)
  - **Patient id, Age, diseases type, survival data, treatment and so on.**
- **Omics data** (these data are mainly generated using *high-throughput technologies*.)
  - **DNA sequence data (DNA base pairs: A,T,C & G)**
  - **Genomic data (gene expression)**
  - **Epigenomic data (DNA methylation)**
  - **Mutations, copy number, etc.**

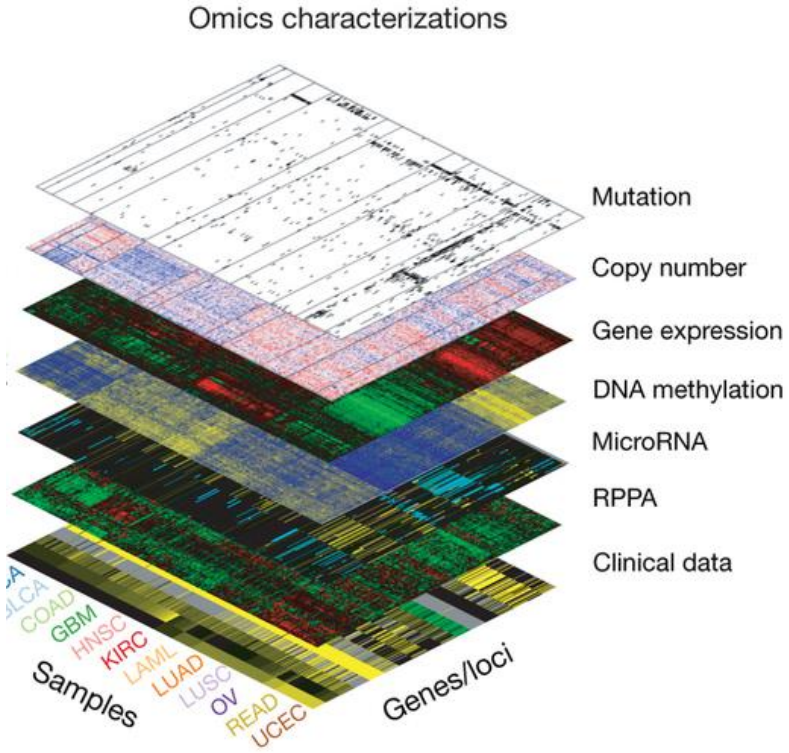
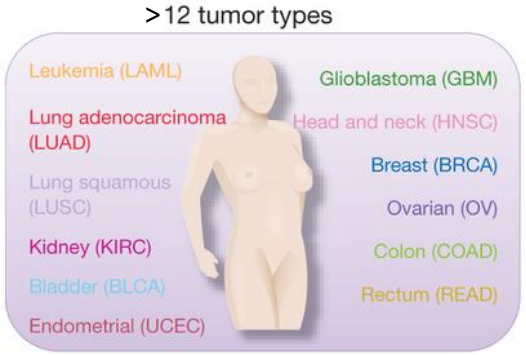



**Figure 1.7 | Data types in medicine: clinical vs. omics**. Illustration of various clinical and omics data across samples

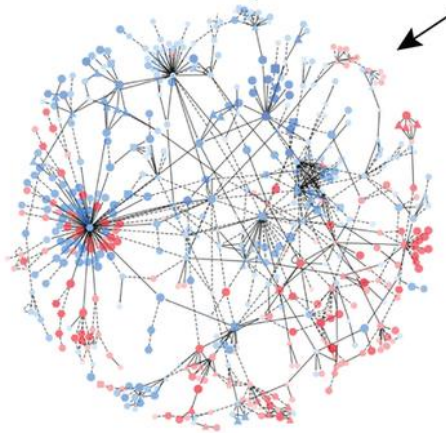
# Data in medicine: portal/repositories

## The Cancer Genome Atlas (**TCGA**) Network & the International Cancer Genome Consortium (**ICGC**)

Generating large amount of cancer genomic data from multiple technical platforms

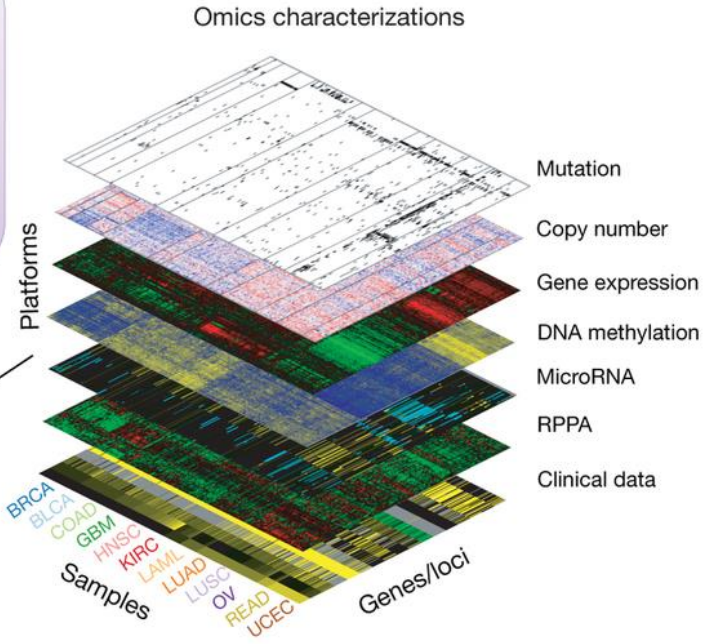


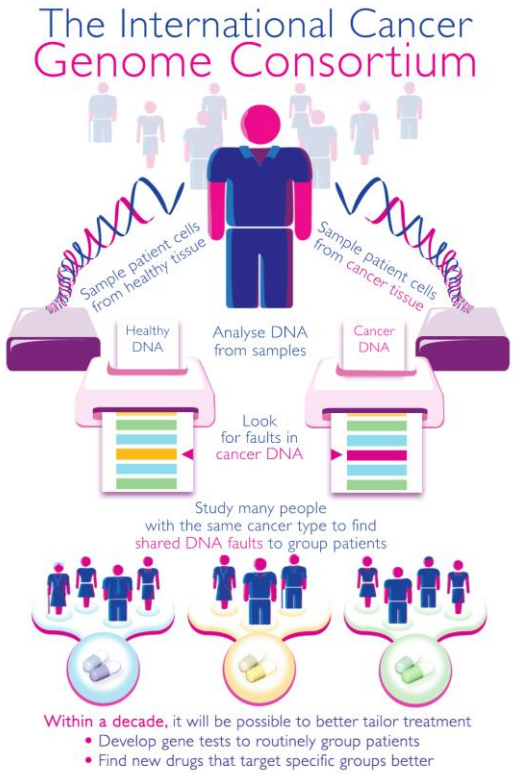

Thematic pathways

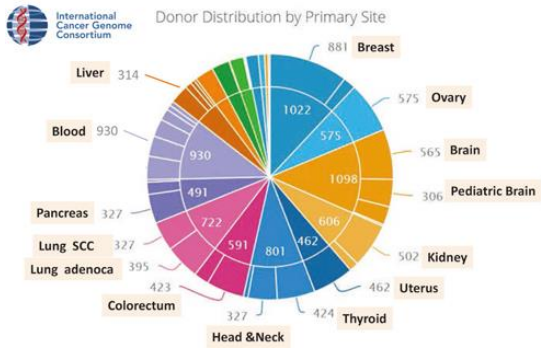


2005

https://portal.gdc.cancer.gov/




2008

https://dcc.icgc.org/

Aim: better diagnosis, treatment and prevention

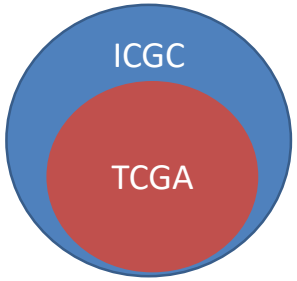



**Figure 1.8 | Data portal and repositories in cancer. a**, TCGA network. **b**, ICGC consortium.

# Data in medicine: DNA sequence

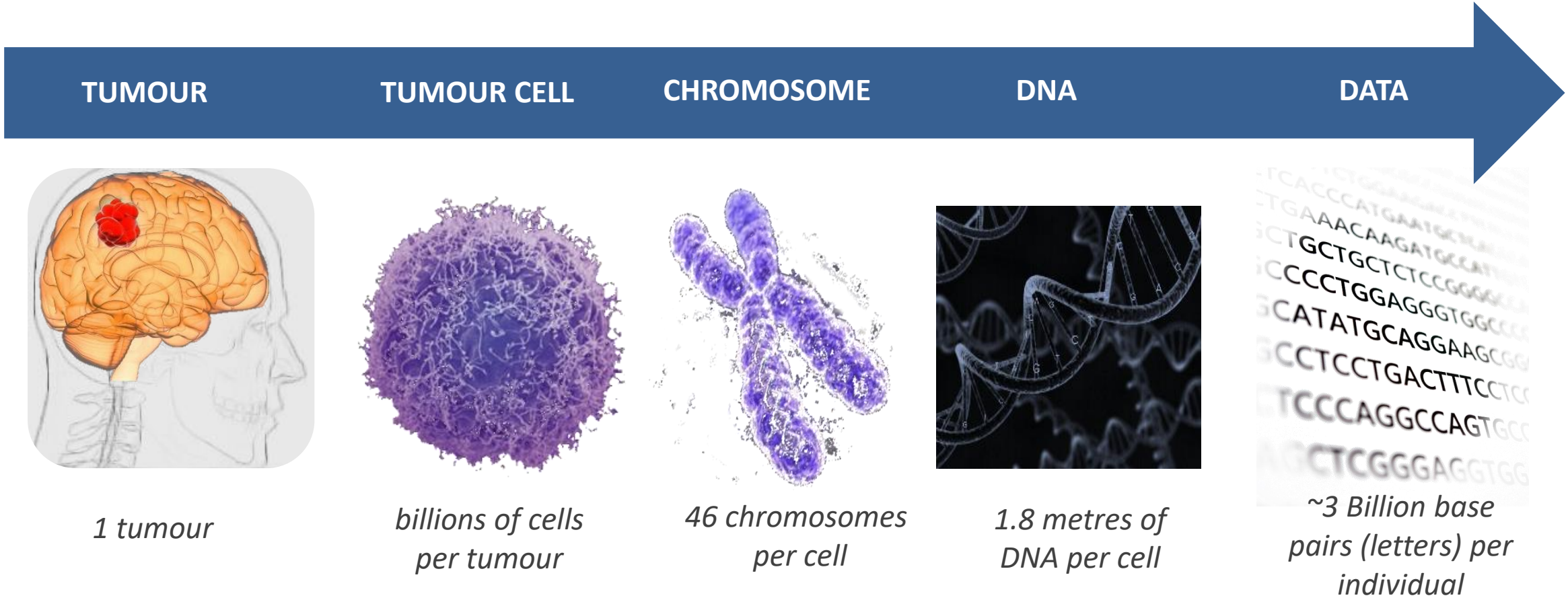


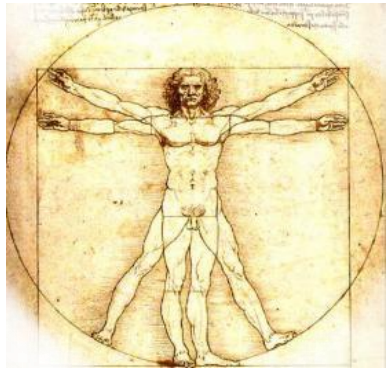

**Figure 1.9 | From cancer (tumour) to data.** Extracting DNA sequences from a brain tumour (glioblastoma).

For more details about this graphical abstract, watch the following video:
https://www.youtube.com/watch?v=XBRiwpNAleM

# Data in medicine: DNA sequence

**a**

**The Human Genome Project**
**1990-2003**



**20 x Centres**
**1000s of Scientists**
**13 Years**
**Cost = $2.7 Billion**

**...**
**... CCCTATATAGGGCCATATATA ...**
**... GCATTAACCAGATACACAGTAA ...**
**...**

**b**

**Sequencing an individual brain tumour in 2015**


**WNT subgroup**

**1-5 Scientists**
**1-5 Weeks**
**Cost = $1,245**

**...TAAATCATGCCGTAC TGAAATC ...**

**Figure 1.10** |**a**, Human genome project. **b**, Sequencing a childhood brain tumour (medulloblastoma) using recent technologies.

# Data pre-processing

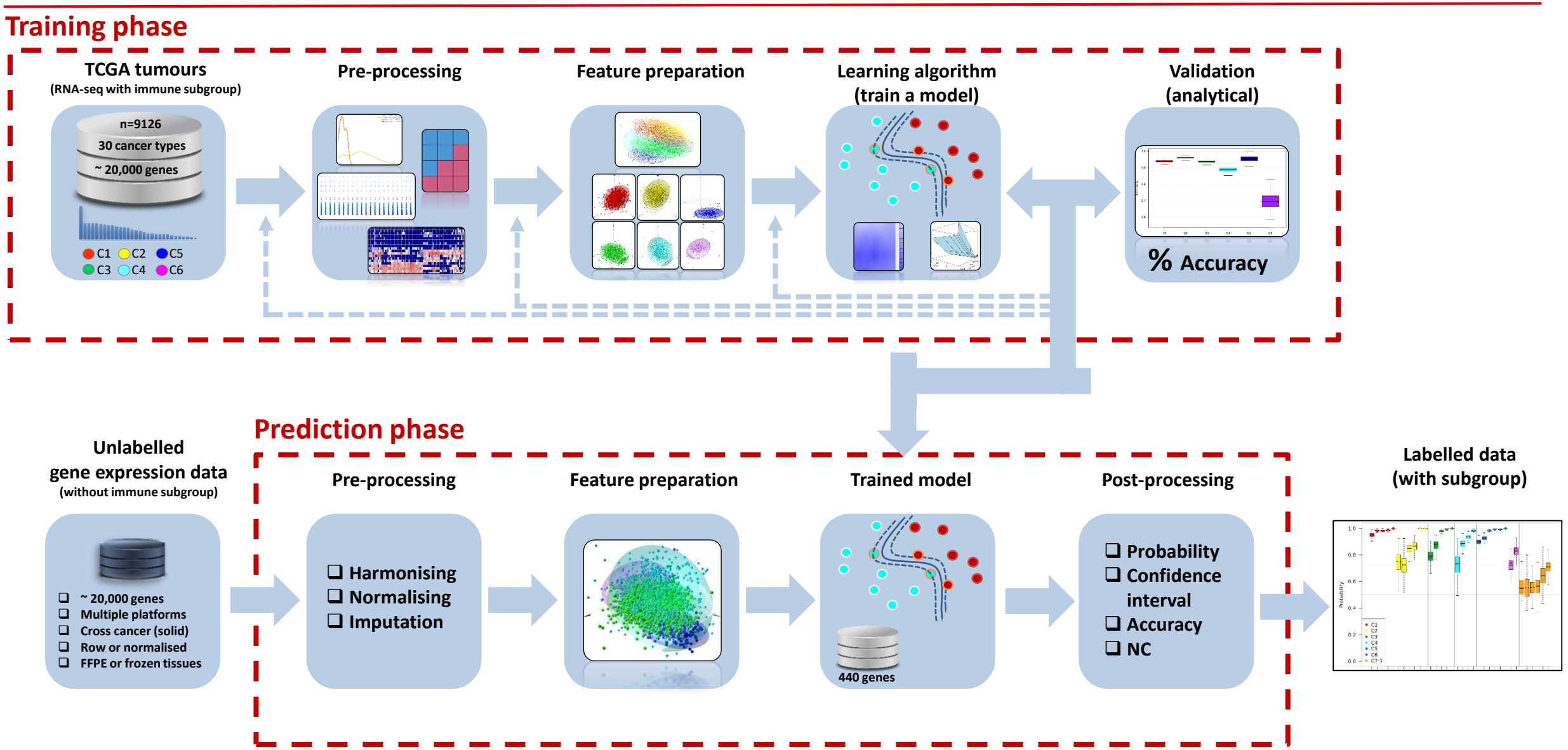



**Figure 1.11 | Data pre-processing stage. a**, Training phase. **b**, Prediction phase.

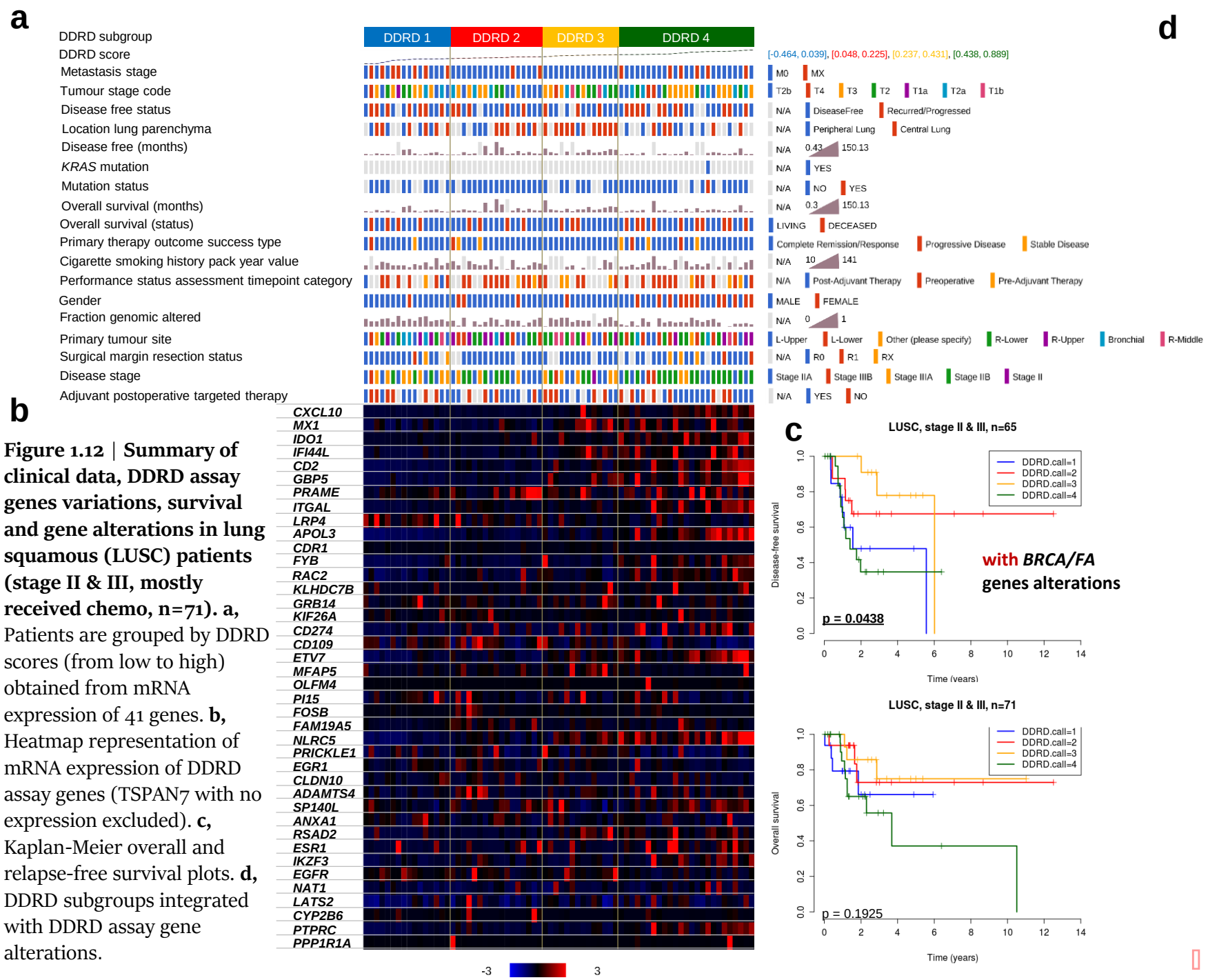


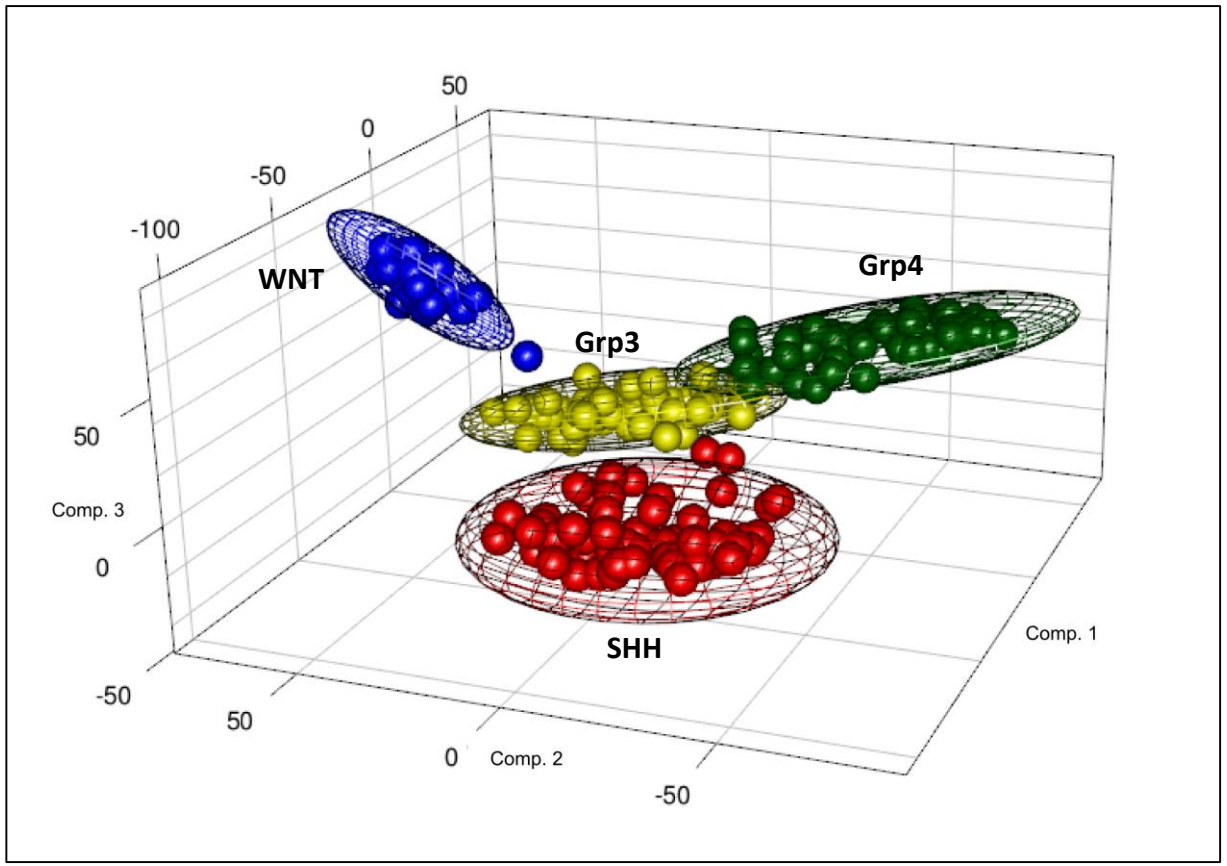

**Figure 1.12 | Summary of clinical data, DDRD assay genes variations, survival and gene alterations in lung squamous (LUSC) patients (stage II & III, mostly received chemo, n=71). a,** Patients are grouped by DDRD scores (from low to high) obtained from mRNA expression of 41 genes. **b,** Heatmap representation of mRNA expression of DDRD assay genes (TSPAN7 with no expression excluded). **c,** Kaplan-Meier overall and relapse-free survival plots. **d,** DDRD subgroups integrated with DDRD assay gene alterations.

QUEEN'S UNIVERSITY BELFAST

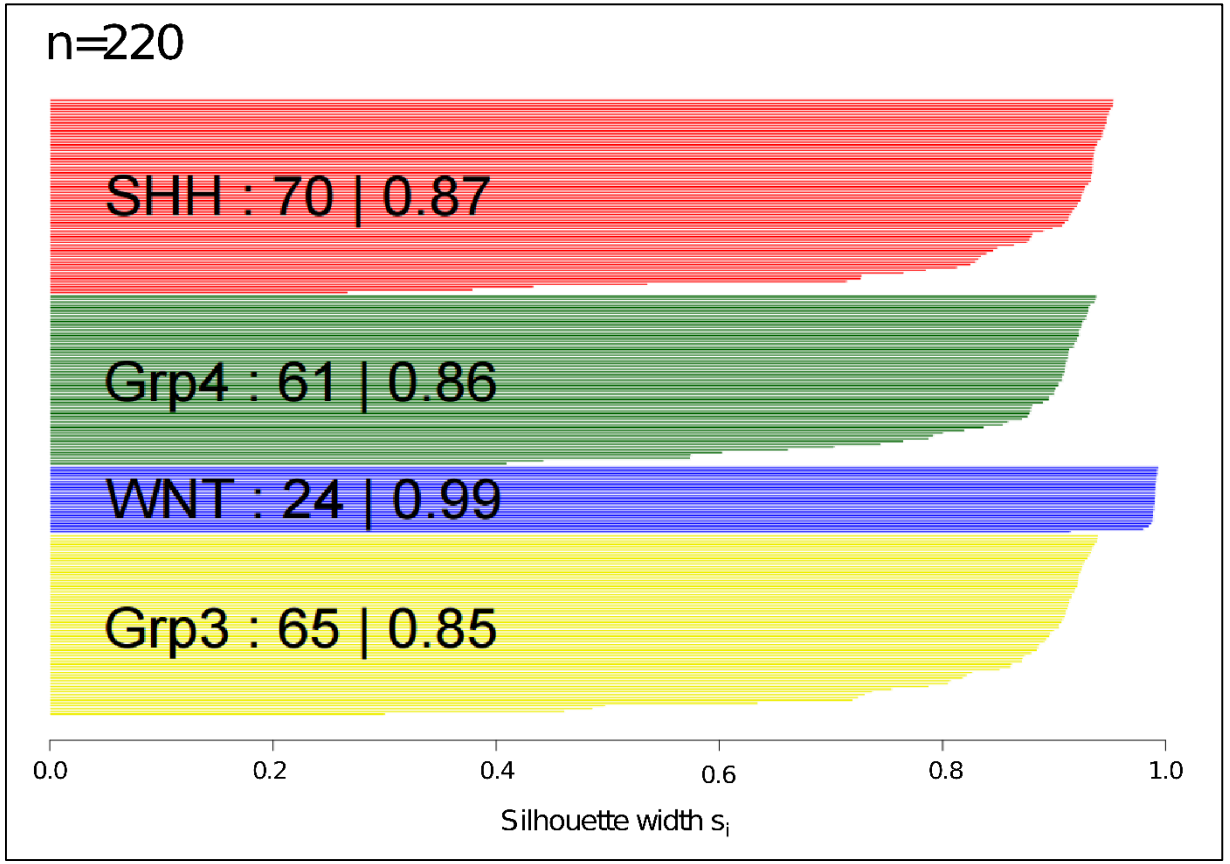
SCHOOL OF ELECTRONICS, ELECTRICAL ENGNIEERING AND COMPUTER SCIENCE

# Visualising high-dimensional data

Unsupervised consensus clustering of DNA Methylation (CpG loci) of 220 medulloblastoma (MB) samples




PCA visualisation of groups identified using a **consensus NMF clustering**




Silhouette plot illustrates robustness of each group (number and average silhouette width are shown)

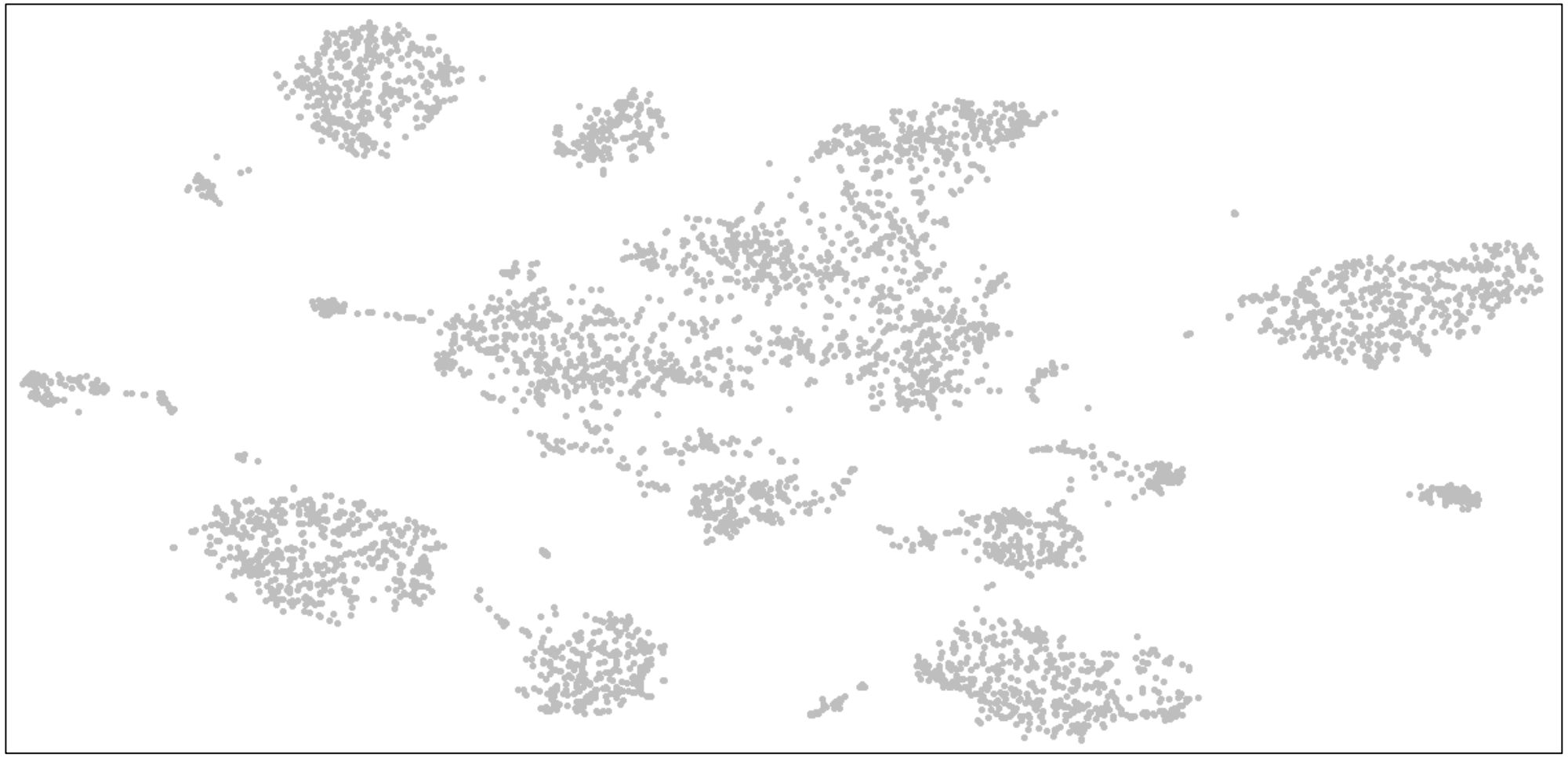
PCA: Principal component analysis

NMF: Non-negative matrix factorisation

G. Rafiee, et al., Minimal Methylation Classifier (MIMIC), *Scientific Report*, 2017

# Data summarisation & visualisation

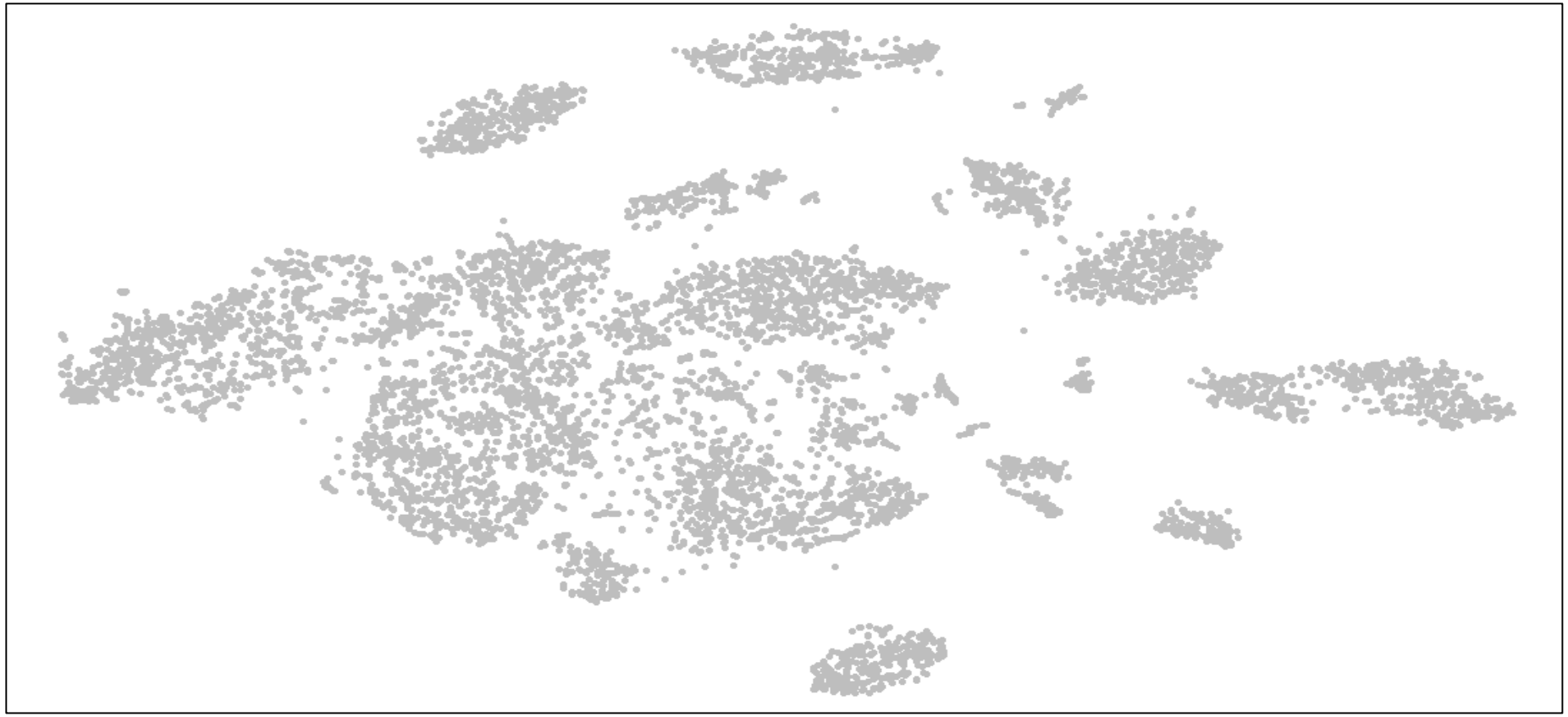
## Unsupervised clustering of 9126 solid tumours (16,335 genes) – transcriptomic map of tumour



2D visualisation of a large dataset of 30 solid tumours using *t-SNE* dimensionality reduction technique

# Data summarisation & visualisation

## Unsupervised clustering of 7302 solid tumours (440 genes)



2D visualisation of a large dataset of 30 solid tumours using *t-SNE* dimensionality reduction technique

# Data summarisation & visualisation

## Unsupervised clustering of 7302 solid tumours (440 genes)

a

LUAD
GBM
OV
LUSC
PRAD
BLCA
TGCT
ESCA
PAAD
SARC
BRCA
MESO
COAD
STAD
CHOL
KIRP
UCEC
READ
KIRC
HNSC
LIHC
SKCM
CESC
THCA
KICH
UCS
ACC
PCPG
LGG
UVM

b

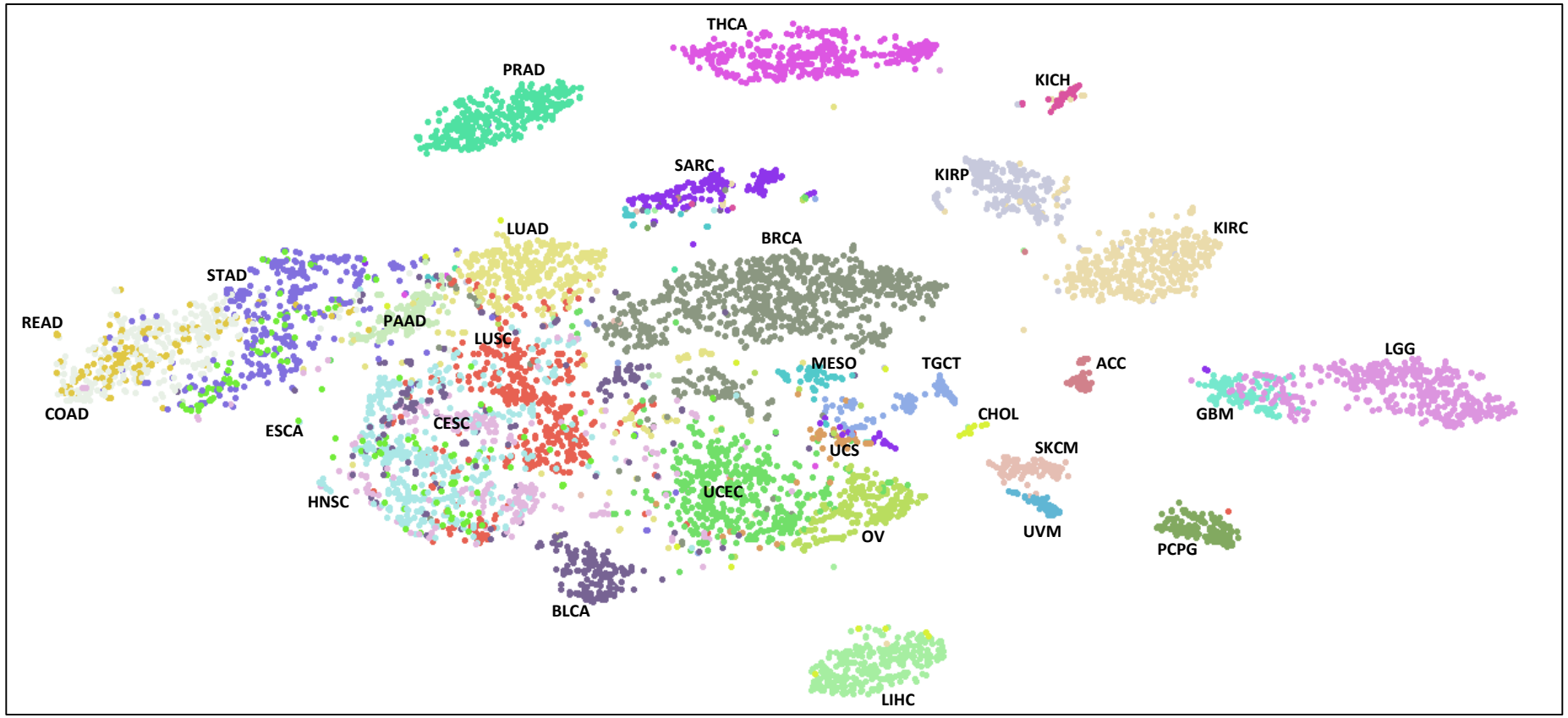



**Figure 1.16 | Establishing the gene expression-based immune solid tumours reference cohort. a**, Overview of the 30 non-hematologic/solid tumour cohorts. **b**, Unsupervised clustering of reference cohort samples (n=7,302) using t-SNE dimensionality reduction technique. Individual samples are colour-coded in the respective class colour (n=30) and labelled with the class abbreviation.