Lecture 25.2

"Big data" in genomics experiments

Nicola Romanò
nicola.romano@ed.ac.uk

浙江大学爱丁堡大学联合学院
ZJU-UoE Institute

Learning objectives

At the end of this you should be able to:

- Define “big data” in the context of biomedical research
- Discuss the issues associated with big data
- Describe a typical pipeline of genomic sequencing

Big data – an historical perspective

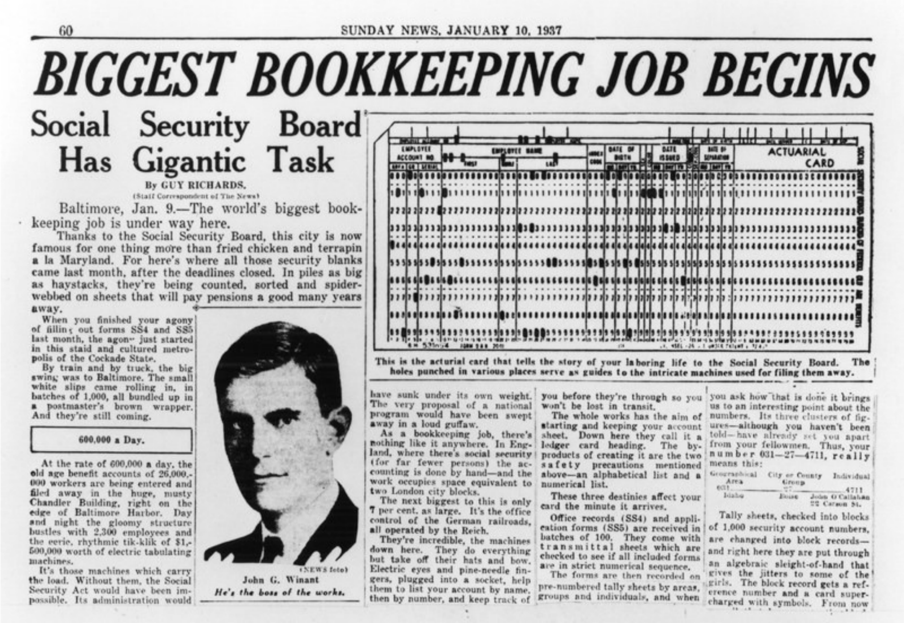
- What is “big data”?

  > *“Big data is a simple way of referring to data sets whose size grows beyond the ability of software and hardware tools to manage, capture, and process in a reasonable timeframe. Big data is where the amount of data is so massive that it becomes very difficult to control.”*
  >
  > Iannarelli and O’Shaughnessy, 2015

- Term (in its current form) popularized in 2005 by Roger Mougalas from O'Reilly Media…

- … but the problem has been around much longer.

Big data – an historical perspective



60 SUNDAY NEWS, JANUARY 10, 1937

BIGGEST BOOKKEEPING JOB BEGINS

Social Security Board Has Gigantic Task

By GUY RICHARDS.
(Staff Correspondent of The News)

Baltimore, Jan. 9.—The world's biggest bookkeeping job is under way here.

Thanks to the Social Security Board, this city is now famous for one thing more than fried chicken and terrapin a la Maryland. For here's where all those security blanks came last month, after the deadlines closed. In piles as big as haystacks, they're being counted, sorted and spider-webbed on sheets that will pay pensions a good many years away.

When you finished your agony of fillin; out forms SS4 and SS5 last month, the agon- just started in this staid and cultured metropolis of the Cockade State.

By train and by truck, the big swing was to Baltimore. The small white slips came rolling in, in batches of 1,000, all bundled up in a postmaster's brown wrapper. And they're still coming.

600,000 a Day.

At the rate of 600,000 a day, the old age benefit accounts of 26,000,000 workers are being entered and filed away in the huge, musty Chandler Building, right on the edge of Baltimore Harbor. Day and night the gloomy structure bustles with 2,300 employees and the eerie, rhythmic tik-klik of $1,500,000 worth of electric tabulating machines.

It's those machines which carry the load. Without them, the Social Security Act would have been impossible. Its administration would have sunk under its own weight. The very proposal of a national program would have been swept away in a loud guffaw.

As a bookkeeping job, there's nothing like it anywhere. In England, where there's social security (for far fewer persons) the accounting is done by hand—and the work occupies space equivalent to two London city blocks.

The next biggest to this is only 7 per cent. as large. It's the office control of the German railroads, all operated by the Reich.

They're incredible, the machines down here. They do everything but take off their hats and bow. Electric eyes and pine-needle fingers, plugged into a socket, help them to list your account by name, then by number, and keep track of you before they're through so you won't be lost in transit.

The whole works has the aim of starting and keeping your account sheet. Down here they call it a ledger card heading. The by-products of creating it are the two safety precautions mentioned above—an alphabetical list and a numerical list.

These three destinies affect your card the minute it arrives.

Office records (SS4) and application forms (SS5) are received in batches of 100. They come with transmittal sheets which are checked to see if all included forms are in strict numerical sequence.

The forms are then recorded on pre-numbered tally sheets by areas, groups and individuals, and when you ask how that is done it brings us to an interesting point about the numbers. Its three clusters of figures—although you haven't been told—have already set you apart from your fellowmen. Thus, your number 031—27—4711, really means this:

Geographical Area	City or County Group	Individual
031	27	4711
Idaho	Boise	John O'Callahan 22 Carson St.

Tally sheets, checked into blocks of 1,000 security account numbers, are changed into block records—and right here they are put through an algebraic sleight-of-hand that gives the jitters to some of the girls. The block record gets a reference number and a card super-charged with symbols. From now

This is the acturial card that tells the story of your laboring life to the Social Security Board. The holes punched in various places serve as guides to the intricate machines used for filing them away.

John G. Winant
He's the boss of the works.

Early “big data” management in the 1930s – source Wired

- In 1937 the US government enacted the Social Security Act
- Data from 26 million Americans >3 million employers
- IBM developed punch card readers specifically for the job

Big data and computing technology

- The sheer volume of data produced today requires
  - A lot of space (virtual and physical) to **store** data
  - Large computing power to **analyse** data
  - Enhanced security to **protect** data

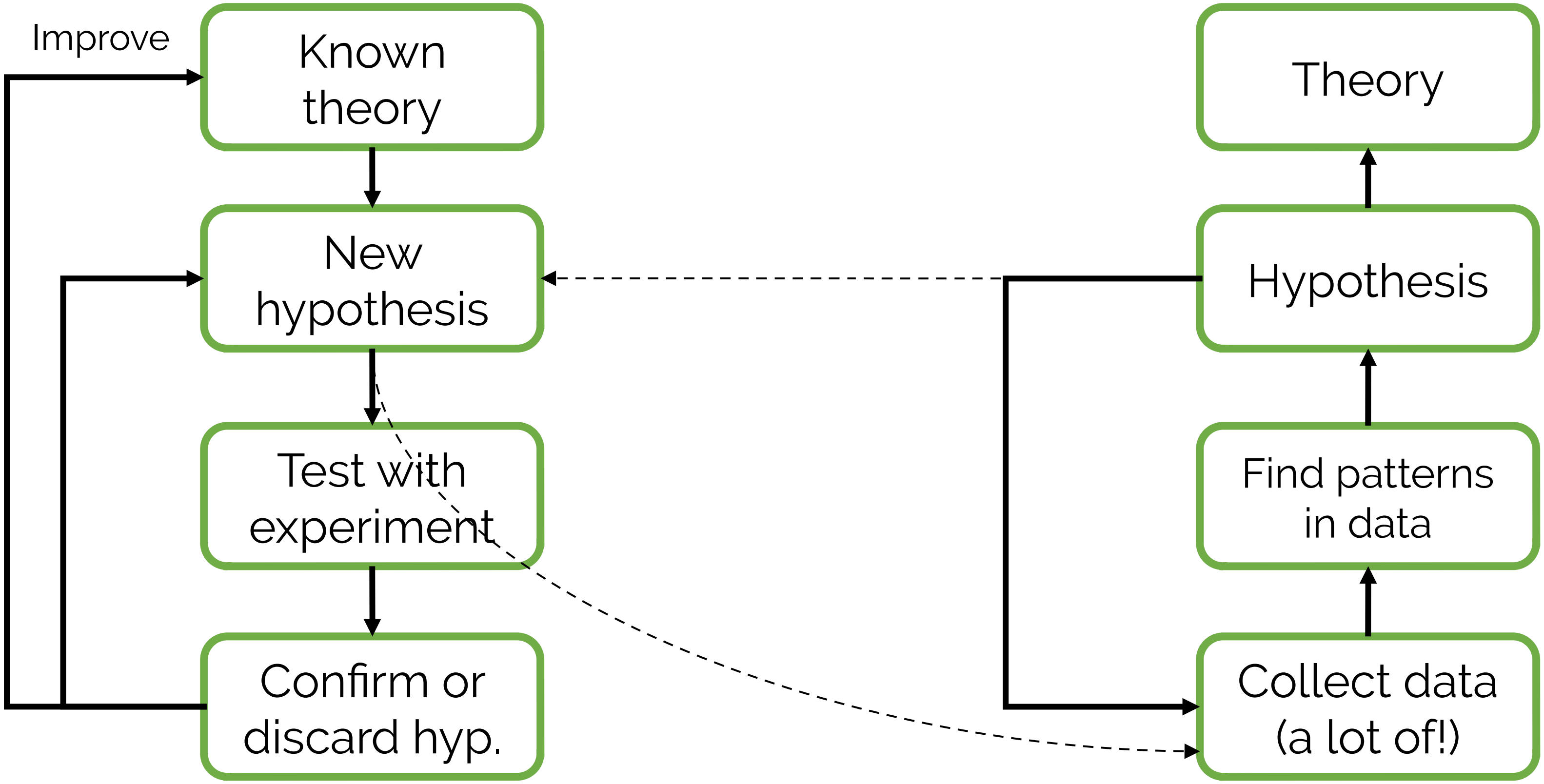
One of Google's data centres – Source Google

Big data in biology

Biology has been affected as well

From “traditional”, hypothesis-driven experimental strategies to large scale exploratory studies

Deduction vs induction



Big data in biology

Biology has been affected as well

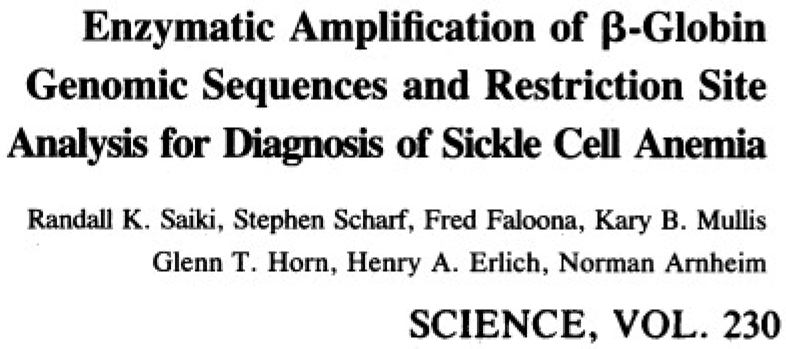
From “traditional”, hypothesis-driven experimental strategies to large scale exploratory studies

“Big data” is found in many domains of biomedical research

- Genomics
- Proteomics
- Imaging
- …

New technologies in genetics

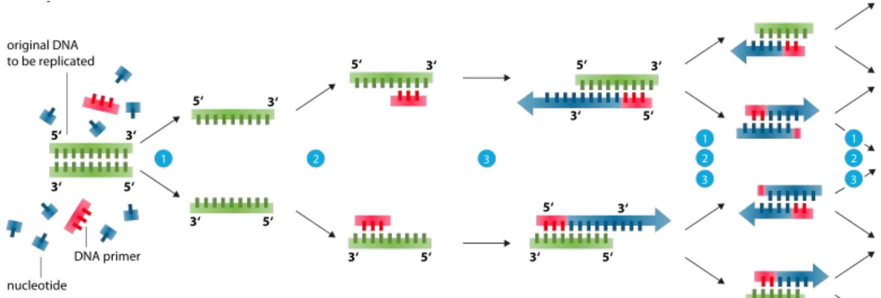
Enormous efforts to sequence the human genome contributed to a big technological push in technologies related to genetics.



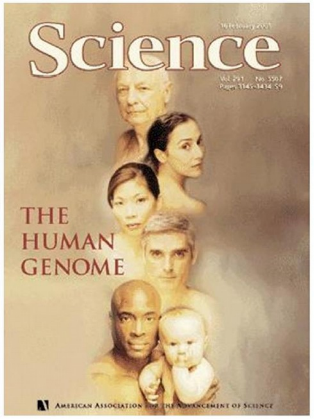
Enzymatic Amplification of β-Globin Genomic Sequences and Restriction Site Analysis for Diagnosis of Sickle Cell Anemia

Randall K. Saiki, Stephen Scharf, Fred Faloona, Kary B. Mullis
Glenn T. Horn, Henry A. Erlich, Norman Arnheim

SCIENCE, VOL. 230



PCR – source: Wikipedia



Venter, J. C. *et al.* **Science** 291, 304–1351 (2001).

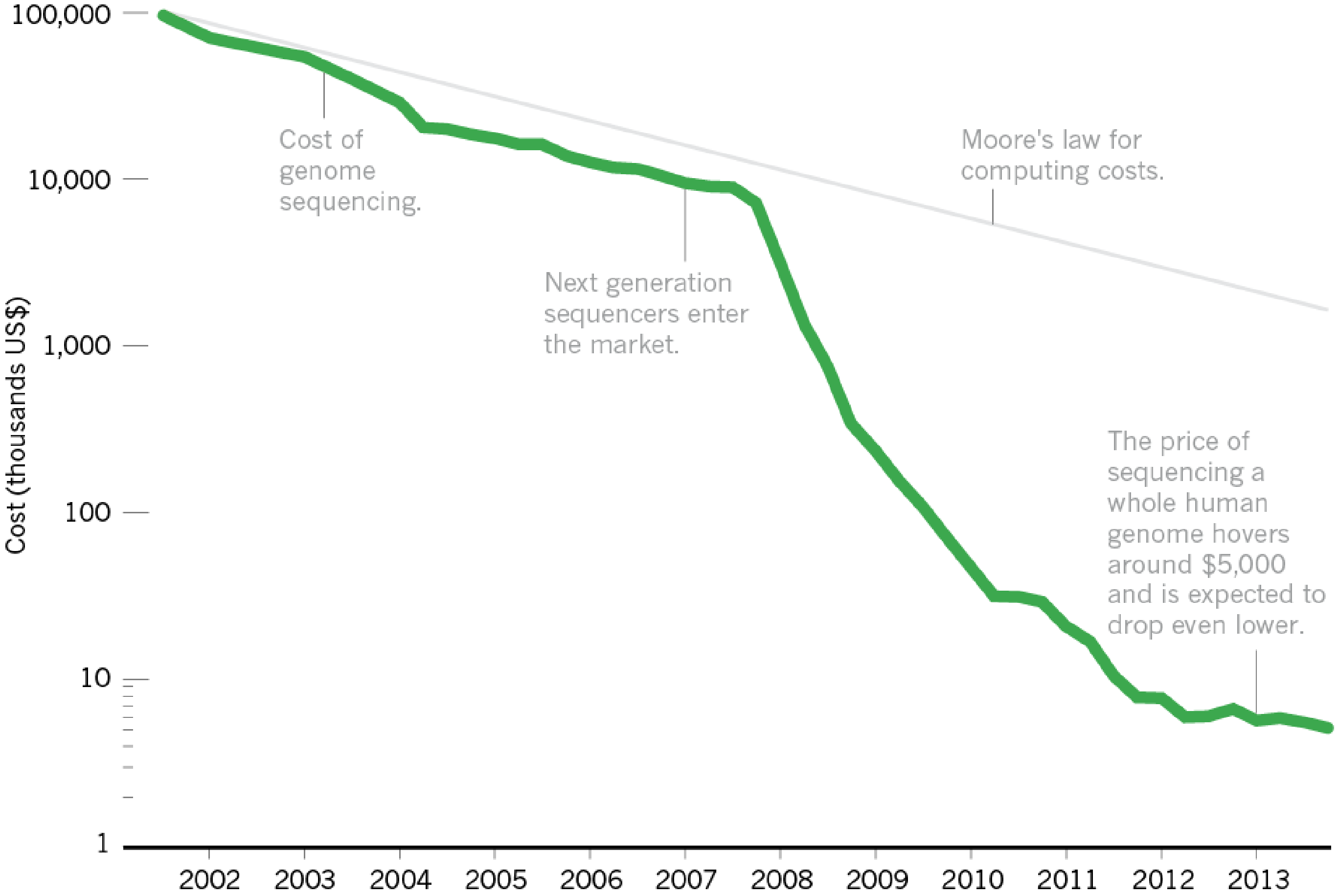
International Human Genome Sequencing Consortium, **Nature** 409, 860–921 (2001).

1985 – Kary Mullis develops PCR

2001 – The human genome is sequenced

New technologies in genetics

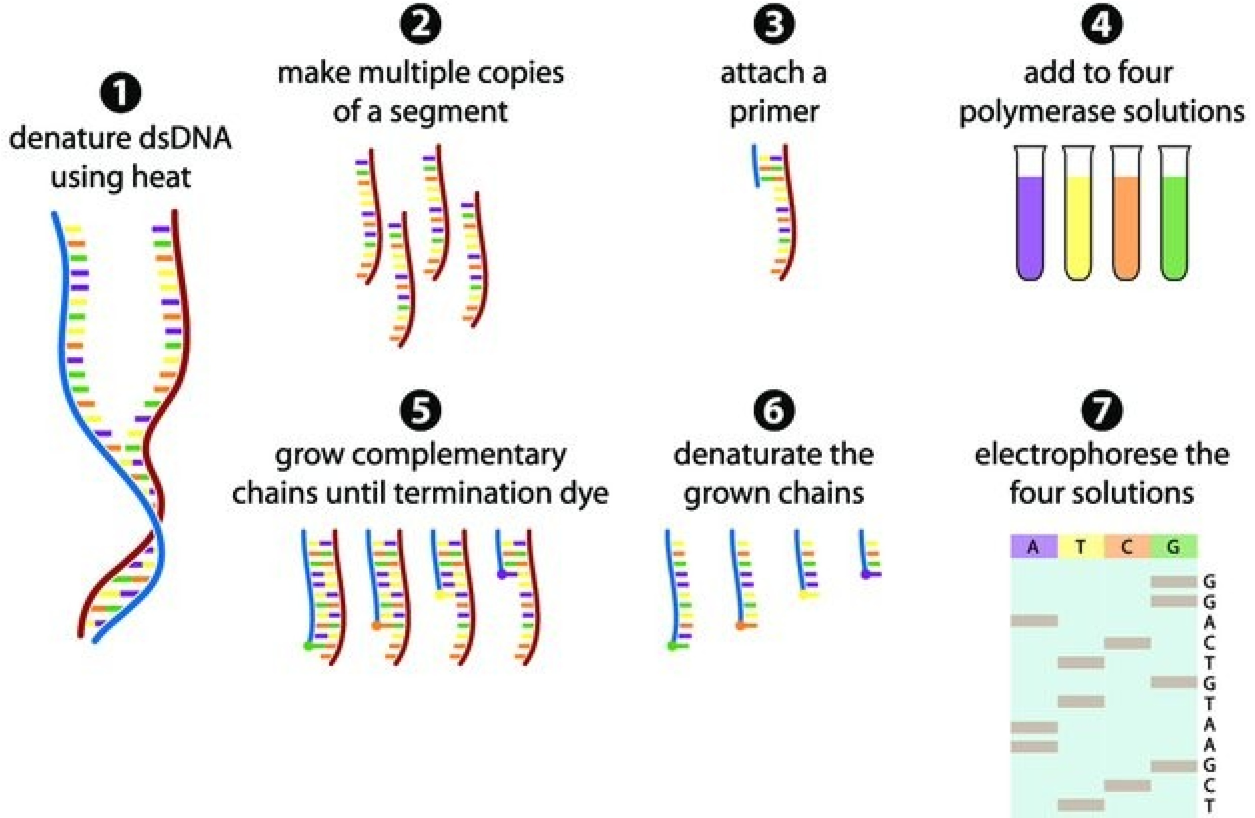
Introduction of next generation sequencing (NGS) made sequencing price drop faster then predicted. Today you can sequence a genome for <$1000.

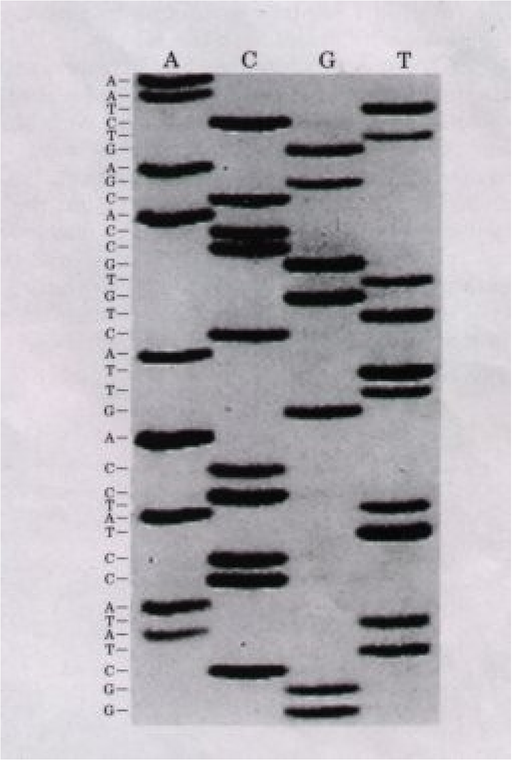


Nature News, 19/3/2014

Traditional Sanger sequencing

Based on PCR using "terminator nucleotides" (analogues that can be incorporated in DNA, but cannot have another nucleotide attached)


1
denature dsDNA using heat
2
make multiple copies of a segment
3
attach a primer
4
add to four polymerase solutions
5
grow complementary chains until termination dye
6
denaturate the grown chains
7
electrophorese the four solutions
A
T
C
G
G
G
A
C
T
G
T
A
A
G
C
T



Gauthier, 2007

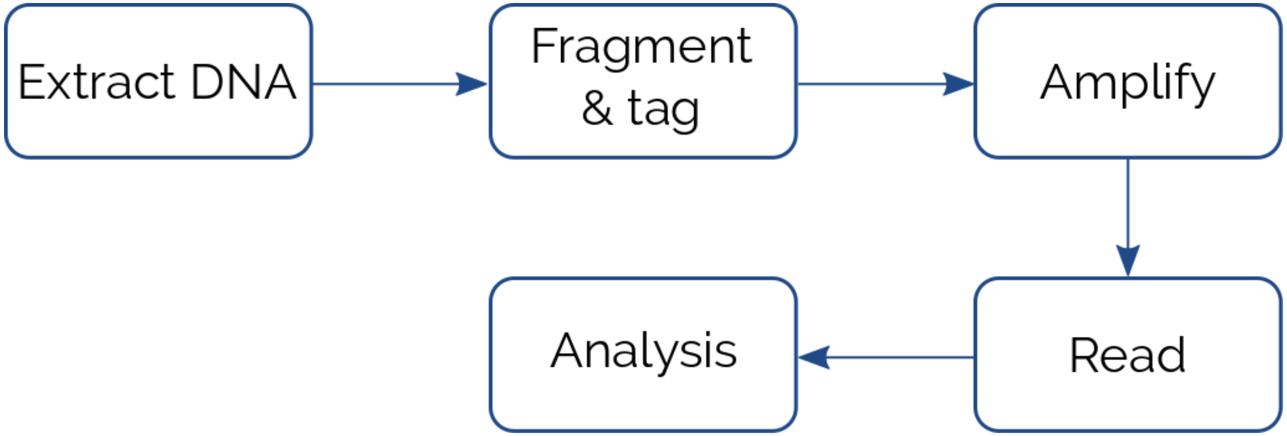
Next generation sequencing

Also called “massive parallel sequencing”. Improved speed and efficiency, reduced costs.

In most cases based on using microfluidics platform

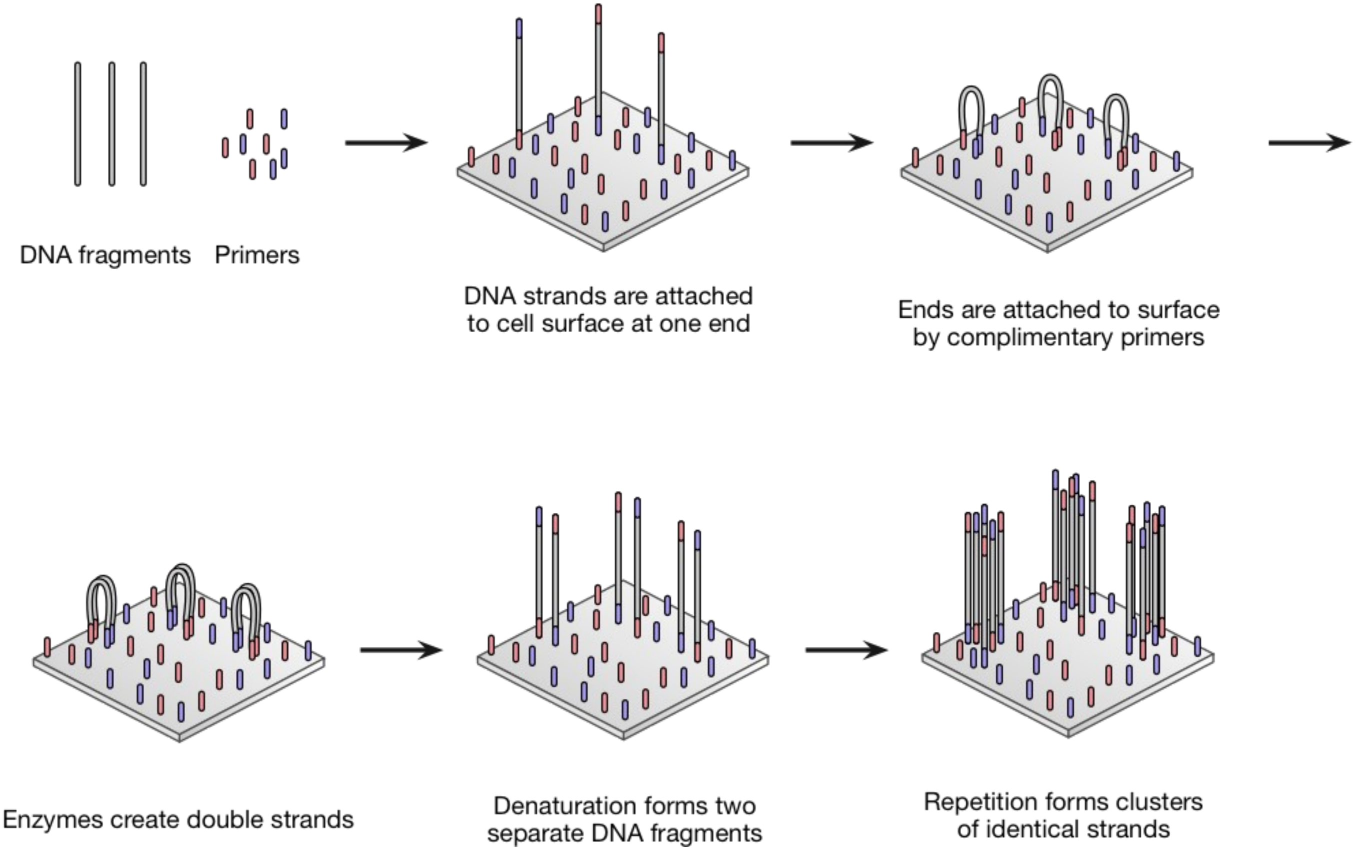
Generally reads small fragments of DNA or RNA (150-300 bp)
Each fragment is called a “read”

State-of-the art machines can sequence up to >30M reads/hour



Illumina sequencing

In Illumina sequencing amplification is obtained by “bridge method”



DNA fragments

Primers

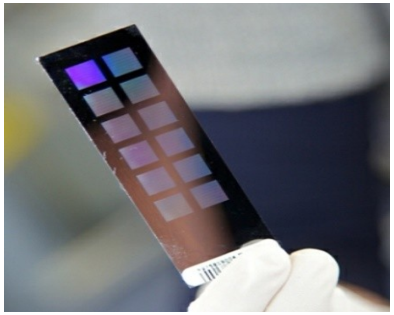
DNA strands are attached to cell surface at one end

Ends are attached to surface by complimentary primers

Enzymes create double strands

Denaturation forms two separate DNA fragments

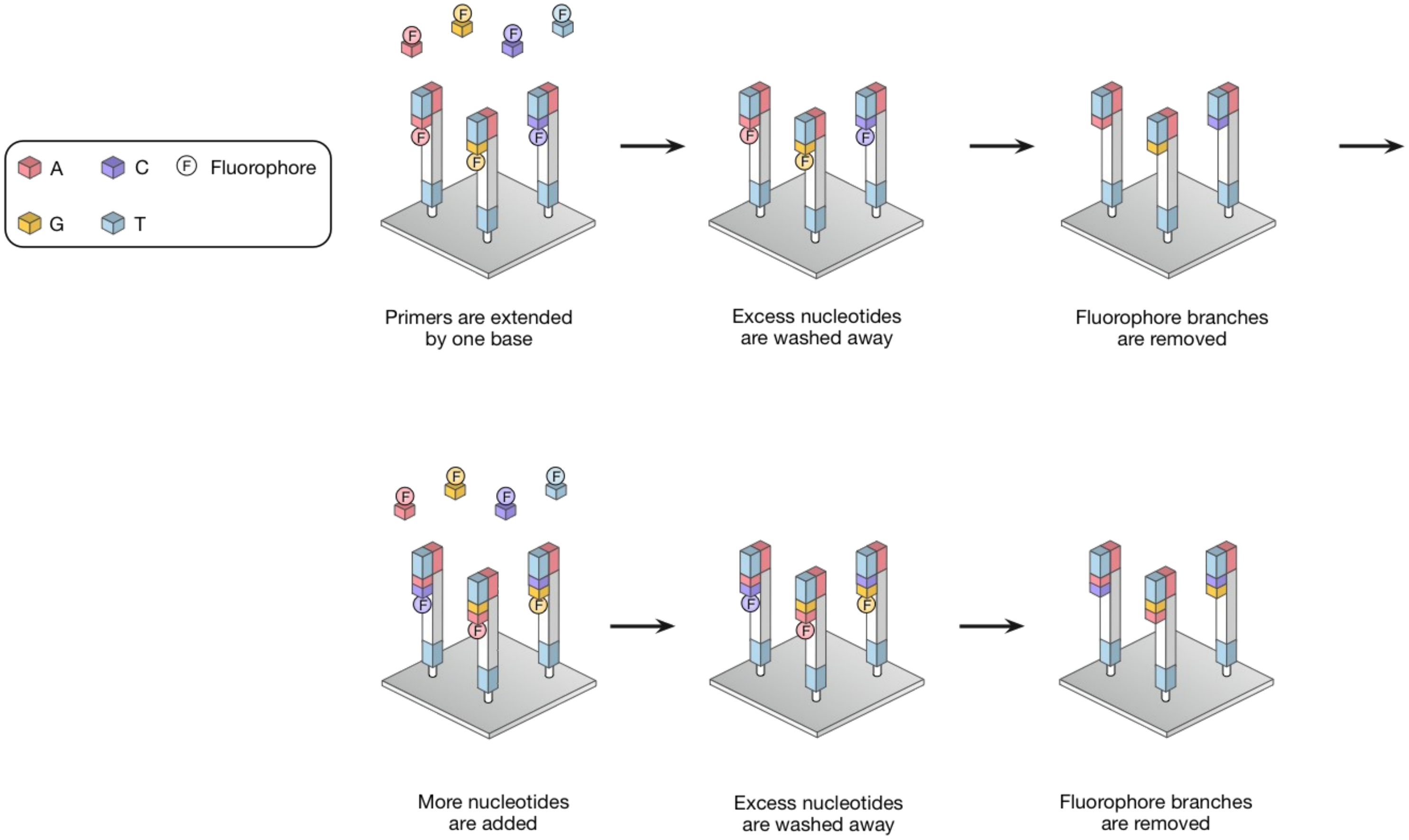
Repetition forms clusters of identical strands



Source: atdbio

Illumina sequencing

Sequencing by synthesis



Primers are extended by one base

Excess nucleotides are washed away

Fluorophore branches are removed

More nucleotides are added

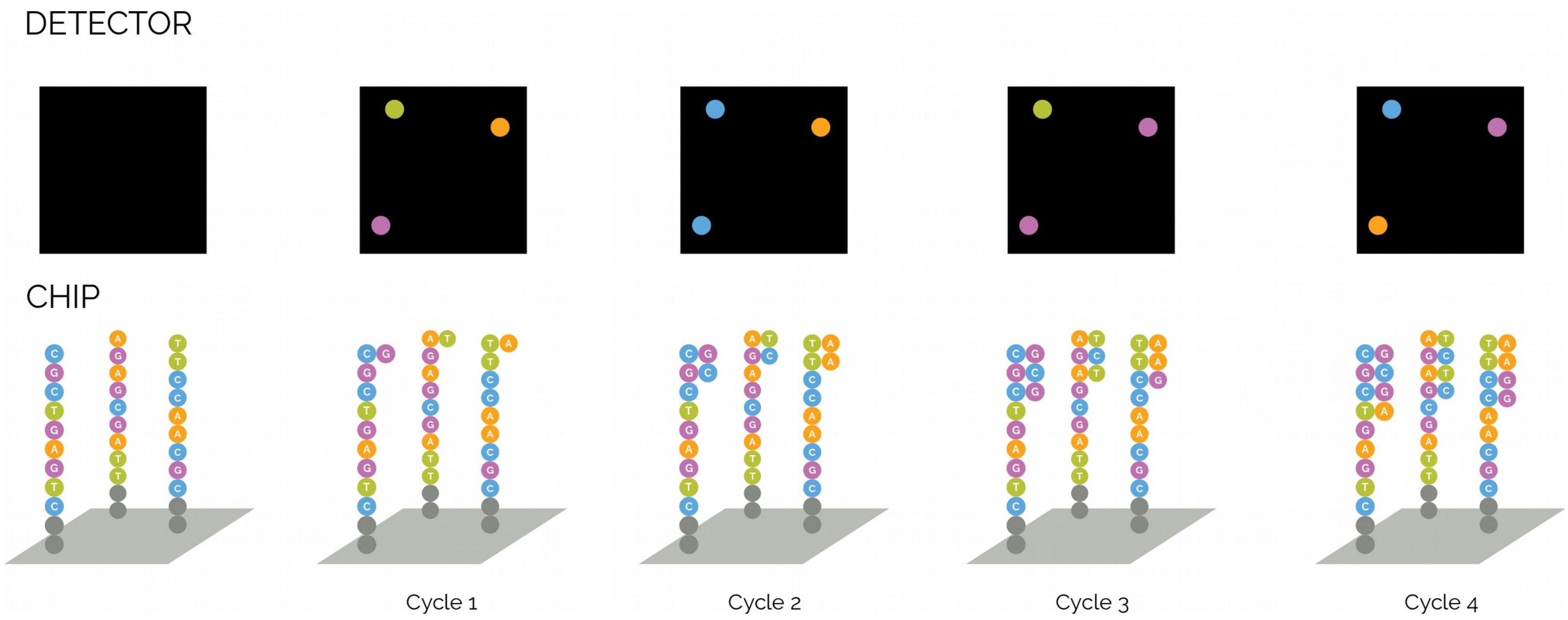
Excess nucleotides are washed away

Fluorophore branches are removed

Source: atdbio

Illumina sequencing

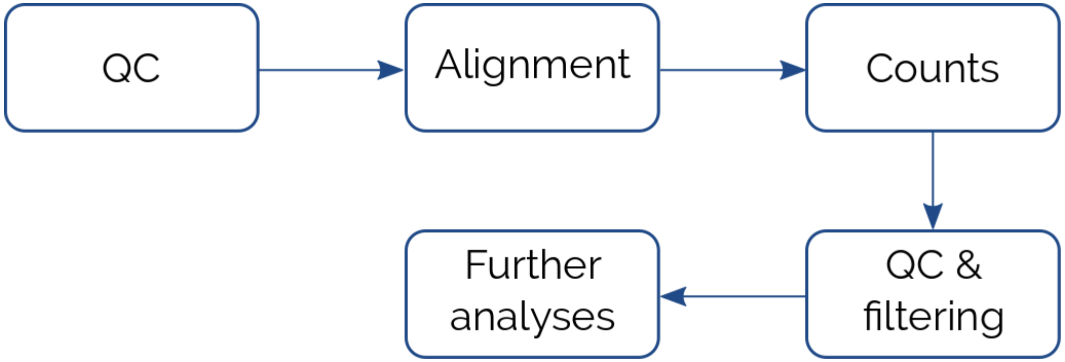
At each step a picture with the spots corresponding to each cluster is taken



Bioinformatics analysis

The next step is to analyse the data

Data needs to be aligned to the genome and counted



QC

Alignment

Counts

QC & filtering

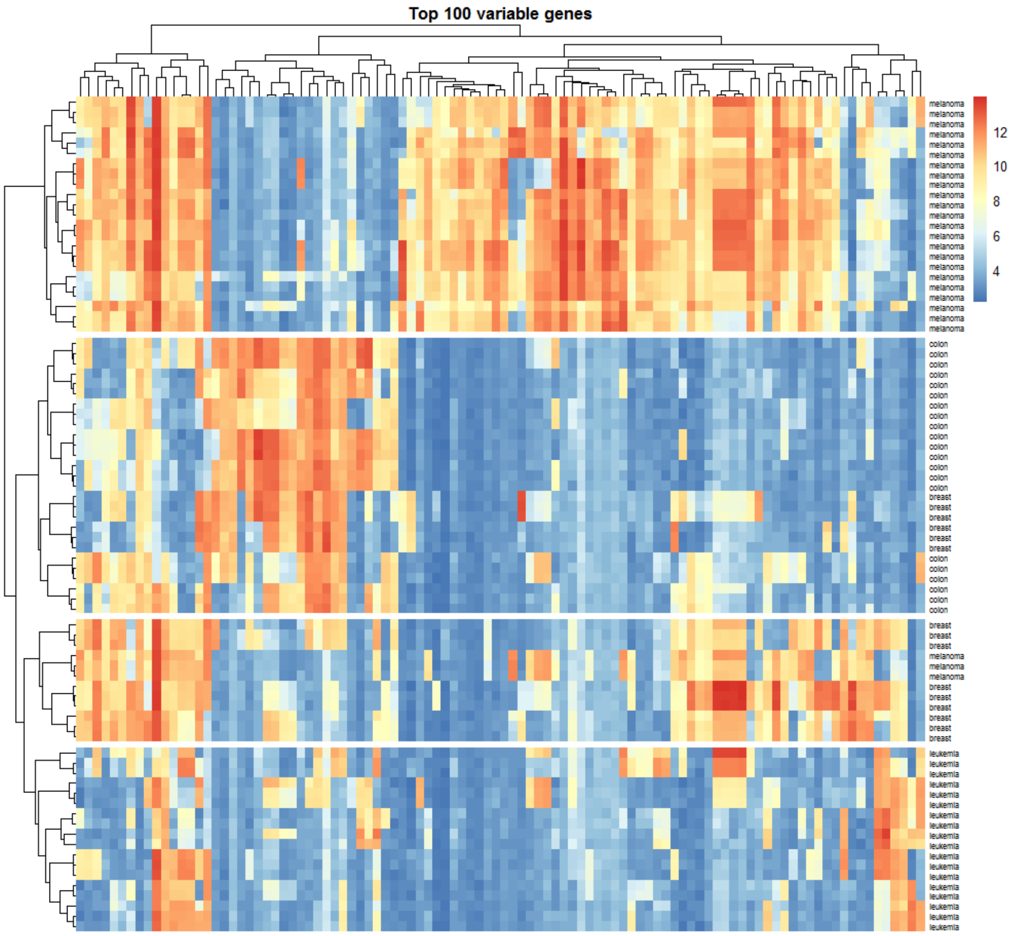
Further analyses

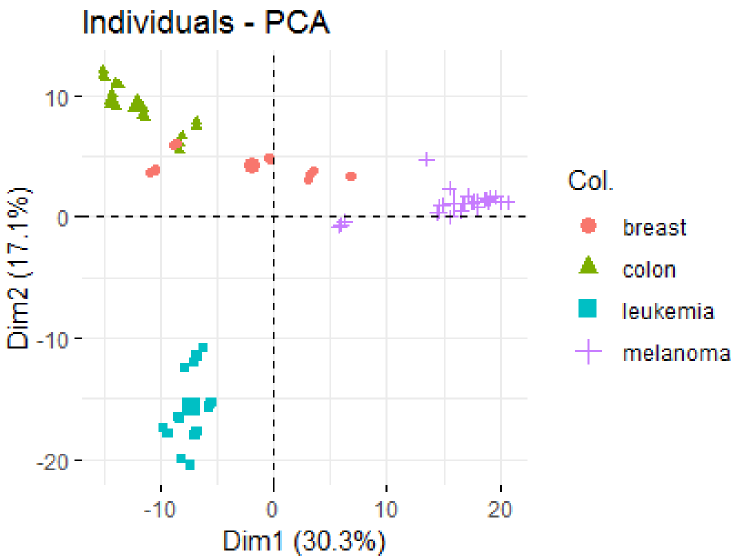
	S1	S2	S3	S4
G1	120	0	30	34
G2	85	22	15	0
G3	0	0	2	88
G4	150	230	223	743
…	…	…	…	…
G20000	45	23	55	11

Bioinformatics analysis

Visualization techniques

Heatmaps and clustering, PCA - see Workshop 7!

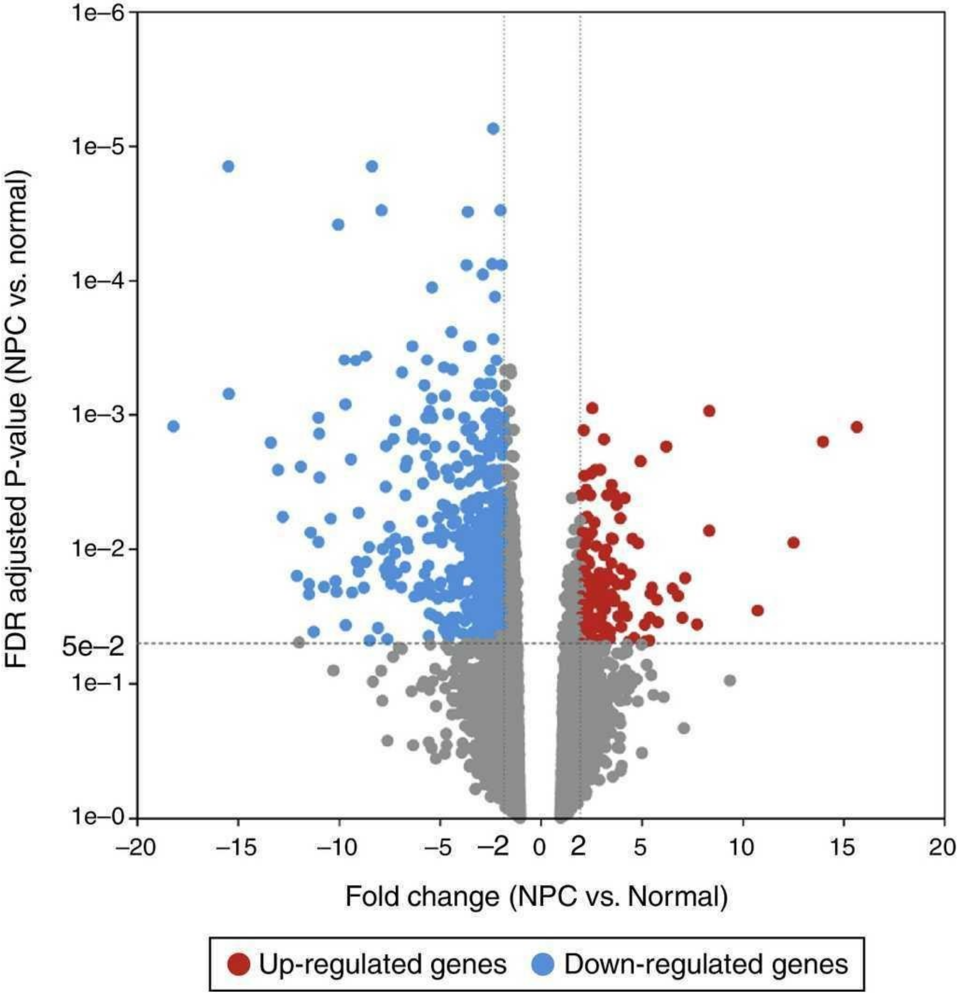




Bioinformatics analysis

Differential expression analysis

- Finds genes under-/over-expressed between two samples
- Considers fold change and adjusted p-value
- Useful to find targets for further analysis



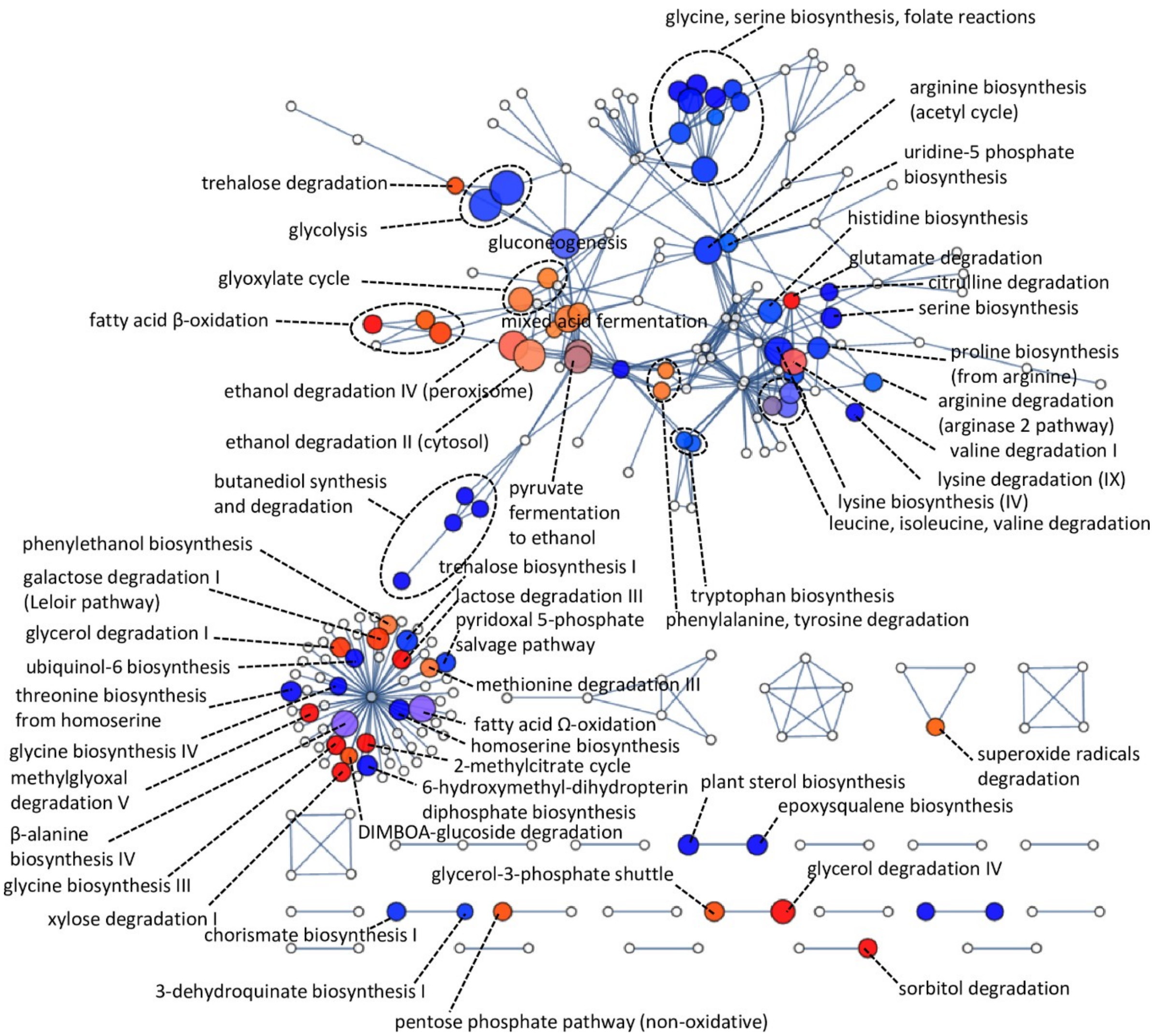
Jiang, 2016

Bioinformatics analysis

Pathway analysis

- Finds molecular pathways enriched in the sample

- Genes can be mapped to function e.g. through GO annotations



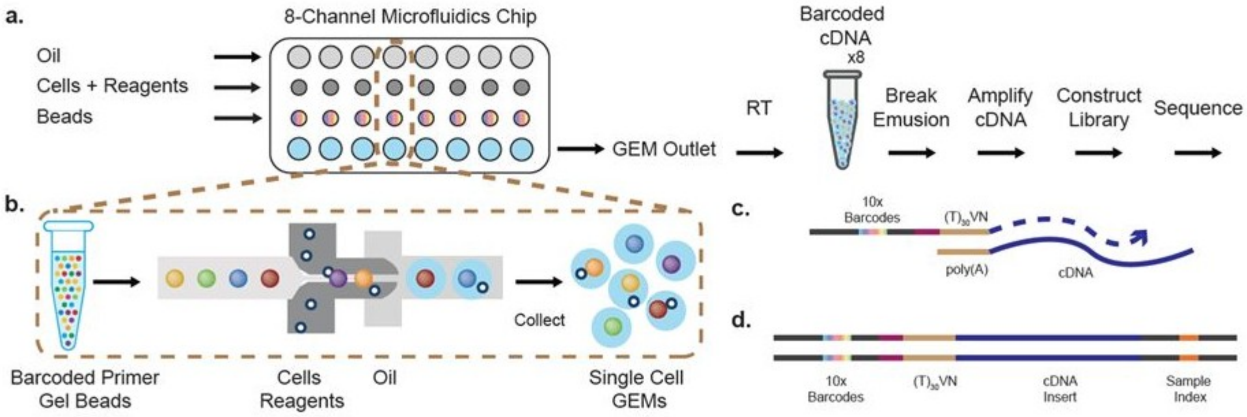
glycine, serine biosynthesis, folate reactions
arginine biosynthesis (acetyl cycle)
uridine-5 phosphate biosynthesis
histidine biosynthesis
glutamate degradation
citrulline degradation
serine biosynthesis
proline biosynthesis (from arginine)
arginine degradation (arginase 2 pathway)
valine degradation I
lysine degradation (IX)
lysine biosynthesis (IV)
leucine, isoleucine, valine degradation
trehalose degradation
glycolysis
gluconeogenesis
glyoxylate cycle
fatty acid β-oxidation
mixed acid fermentation
ethanol degradation IV (peroxisome)
ethanol degradation II (cytosol)
butanediol synthesis and degradation
pyruvate fermentation to ethanol
trehalose biosynthesis I
lactose degradation III
pyridoxal 5-phosphate salvage pathway
tryptophan biosynthesis
phenylalanine, tyrosine degradation
phenylethanol biosynthesis
galactose degradation I (Leloir pathway)
glycerol degradation I
ubiquinol-6 biosynthesis
threonine biosynthesis from homoserine
glycine biosynthesis IV
methylglyoxal degradation V
β-alanine biosynthesis IV
glycine biosynthesis III
xylose degradation I
methionine degradation III
fatty acid Ω-oxidation
homoserine biosynthesis
2-methylcitrate cycle
6-hydroxymethyl-dihydropterin diphosphate biosynthesis
DIMBOA-glucoside degradation
superoxide radicals degradation
plant sterol biosynthesis
epoxysqualene biosynthesis
glycerol-3-phosphate shuttle
glycerol degradation IV
chorismate biosynthesis I
3-dehydroquinate biosynthesis I
pentose phosphate pathway (non-oxidative)
sorbitol degradation

Schabort, 2016

From bulk to single cell sequencing

Recent advances in NGS and microfluidics technology have allowed sequencing of the transcriptome of single cells.

10X sequencing



Generally 10^3-10^5 cells sequenced – 10^4-10^5 reads/cell

Other methods (e.g. SmartSeq 2): ~10^6 reads/cell

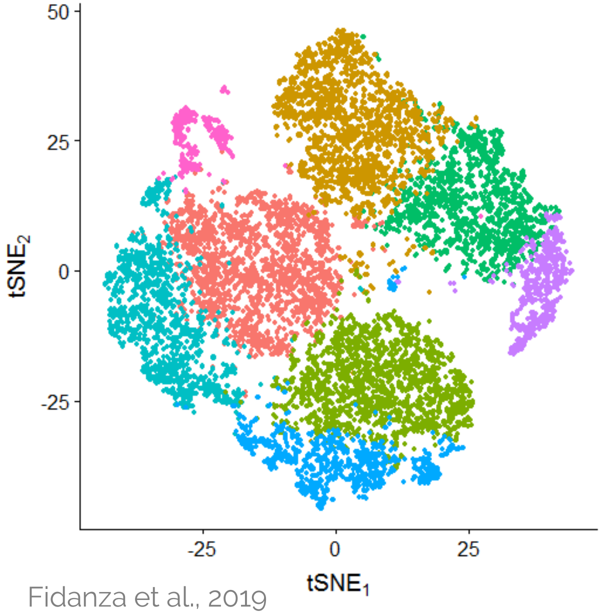
scRNAseq analysis

	Cell 1	Cell 2	Cell 3	...	Cell 12000
G 1	19	0	150	...	0
G 2	0	0	11	...	153
G 3	0	0	22	...	80
...	...	...	...	...	...
G 13000	15	81	0	...	22

Dimension reduction
(e.g. PCA, t-SNE, UMAP...)

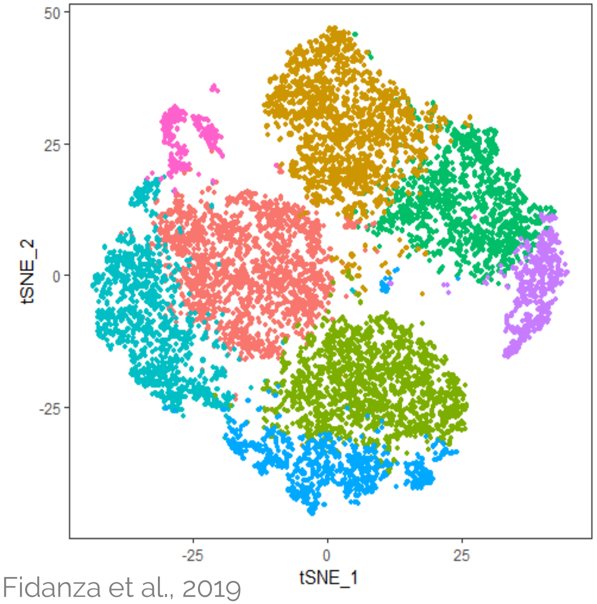
	Dim 1	Dim 2
C 1	2.34	-0.35
C 2	3.15	1.87
...	...	...
C 12000	-2.43	-0.88

Clustering

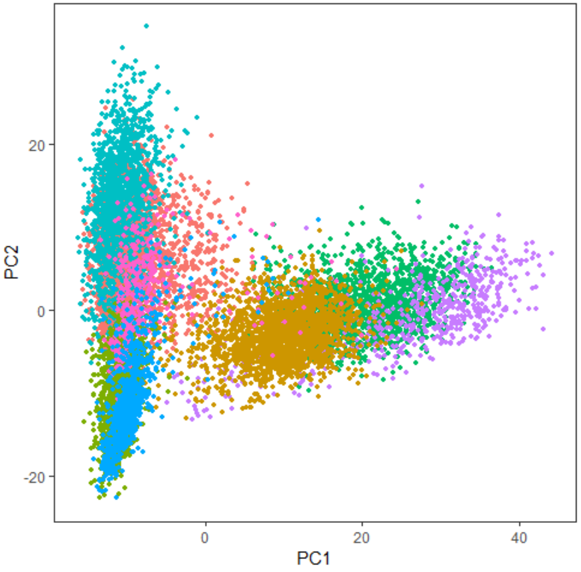


Fidanza et al., 2019

t-SNE dimension reduction



Fidanza et al., 2019



t-Distributed Stochastic Neighbour Embedding

Developed in 2008 by developed by Laurens van der Maaten and Geoffrey Hinton

Particularly suited to high dimensional data, such as that of scRNAseq experiments.

Contrary to PCA is not a linear projection, so it can capture non-linear relationships.

Newer, non-linear methods are being constantly developed, such as UMAP (Uniform Manifold Approximation and Projection for dimension reduction).

What are the issues?

An average file from a RNAseq experiment with ~50M reads is about 20 Gb

This has issues regarding

- Computing power
- Storage space
- Costs
- Data management plans are now required in most grants applications
- Bioinformatics skills are required (although pre-existing analysis pipelines make things a lot easier than they used to be)

Not only sequencing…

Other sources of big data exist including

- Proteomics (and single-cell proteomics) experiments where MS is used to characterise the proteome

- 4D imaging experiments, high content screening, super-resolution methods

- Many more will be coming!

Conclusion

New technologies are changing the way we do biology

Large amount of data are now easy and cheap to generate…

…at the cost of not being able to understand what those data mean!

New skills and technologies will be needed ever more in the future!

- Data analysis
- Machine learning
- Data visualization
- …