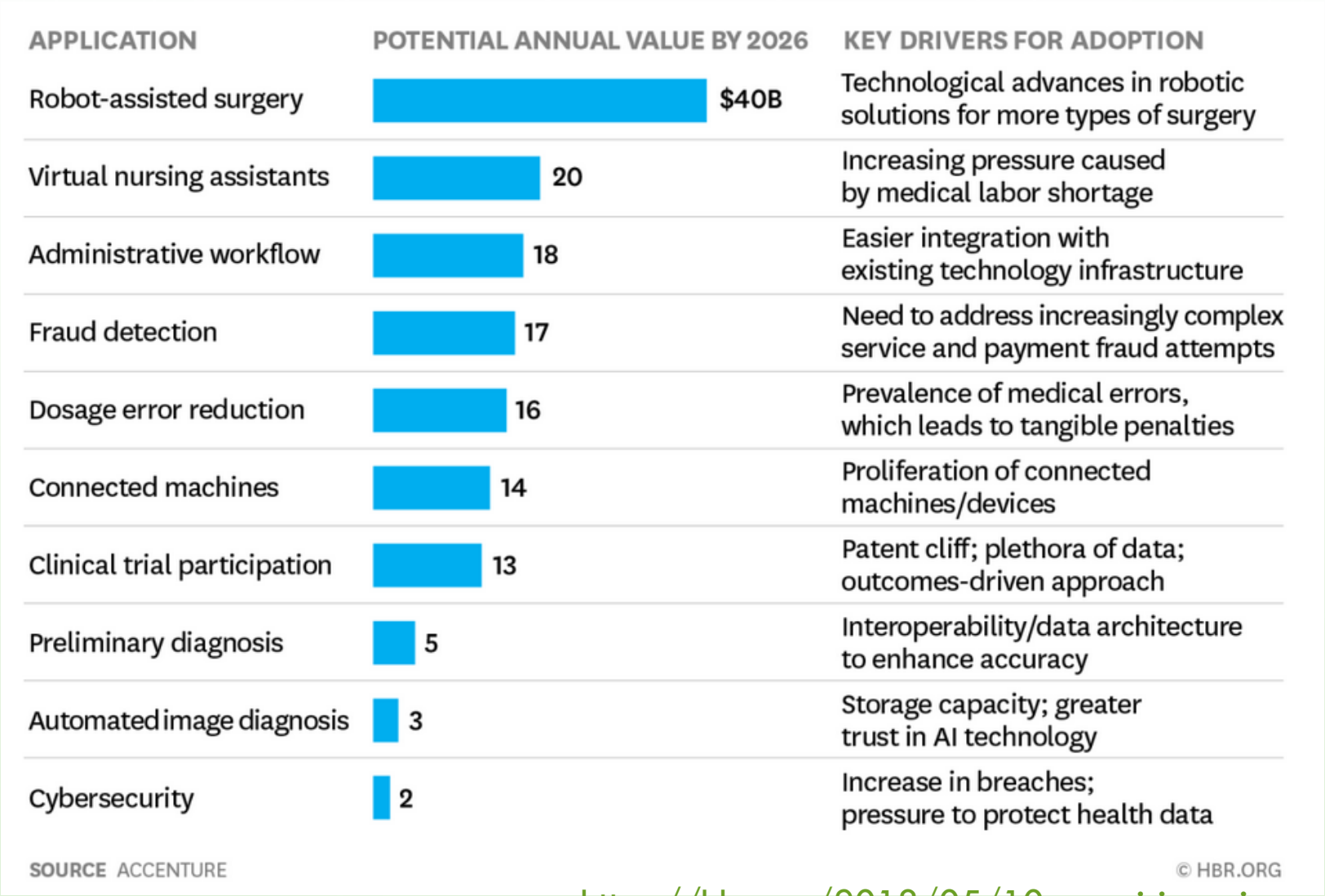
ML LAB 2024

1. INTRODUCTION

FRANCESCA M. BUFFA

HEALTH MAJOR APPLICATION FOR AI

Harvard Business Review

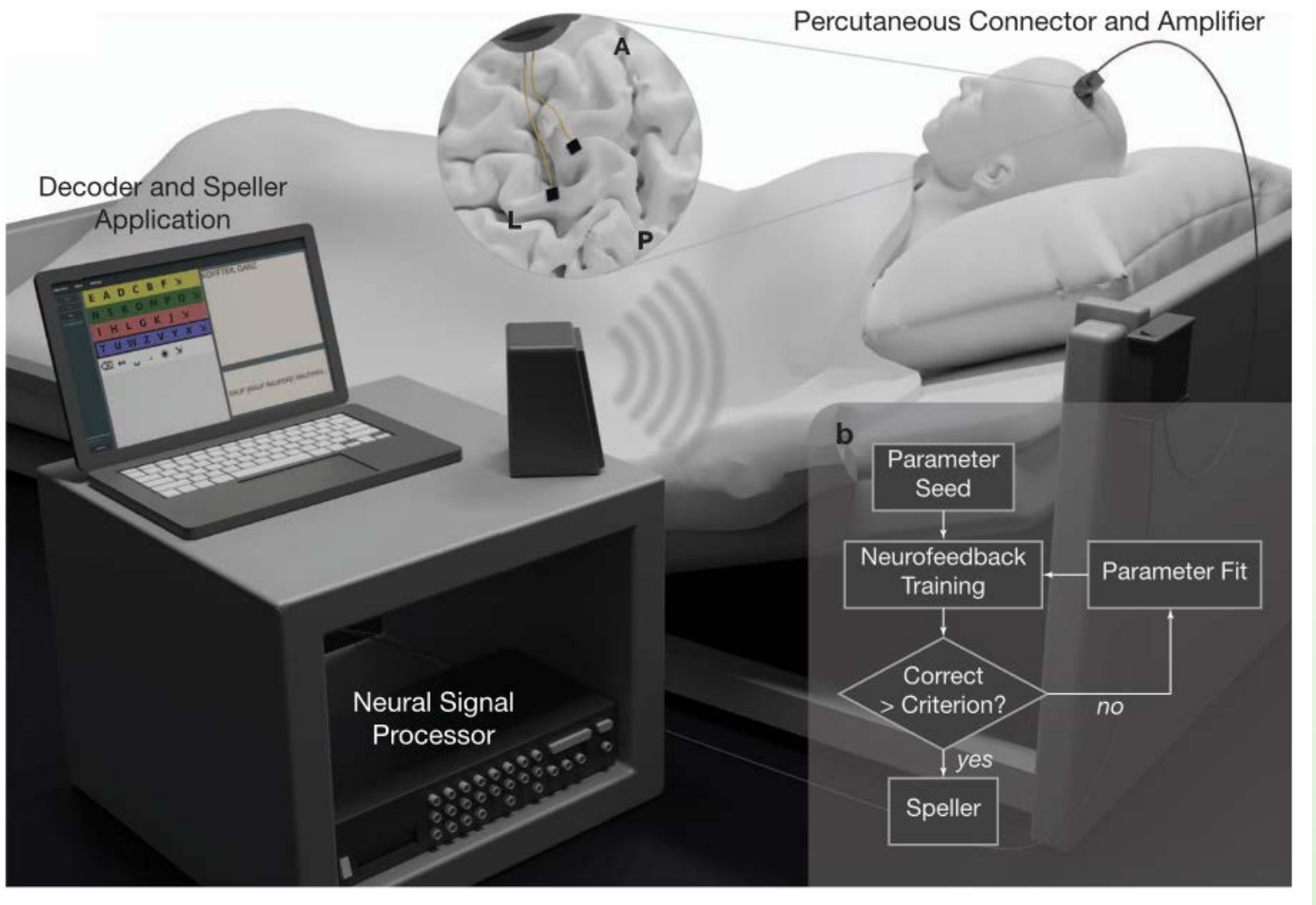


APPLICATION	POTENTIAL ANNUAL VALUE BY 2026	KEY DRIVERS FOR ADOPTION
Robot-assisted surgery	$40B	Technological advances in robotic solutions for more types of surgery
Virtual nursing assistants	20	Increasing pressure caused by medical labor shortage
Administrative workflow	18	Easier integration with existing technology infrastructure
Fraud detection	17	Need to address increasingly complex service and payment fraud attempts
Dosage error reduction	16	Prevalence of medical errors, which leads to tangible penalties
Connected machines	14	Proliferation of connected machines/devices
Clinical trial participation	13	Patent cliff; plethora of data; outcomes-driven approach
Preliminary diagnosis	5	Interoperability/data architecture to enhance accuracy
Automated image diagnosis	3	Storage capacity; greater trust in AI technology
Cybersecurity	2	Increase in breaches; pressure to protect health data

SOURCE ACCENTURE

https://hbr.org/2018/05/10-promising-ai-applications-in-health-care

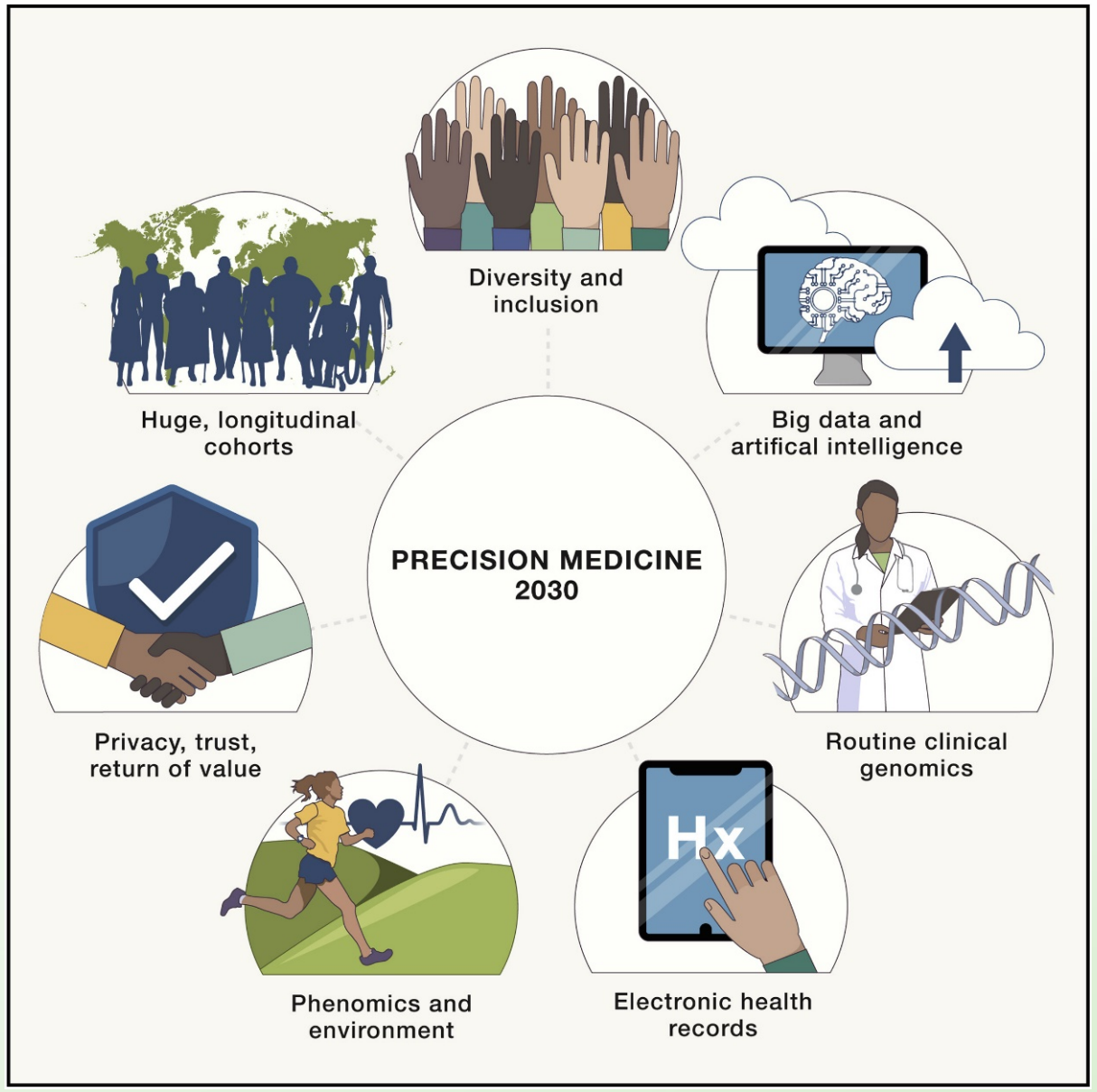
SPELLING INTERFACE USING INTRACORTICAL SIGNALS ENABLED VIA AUDITORY NEUROFEEDBACK TRAINING



https://www.nature.com/articles/s41467-022-28859-8

AI IN HEALTH

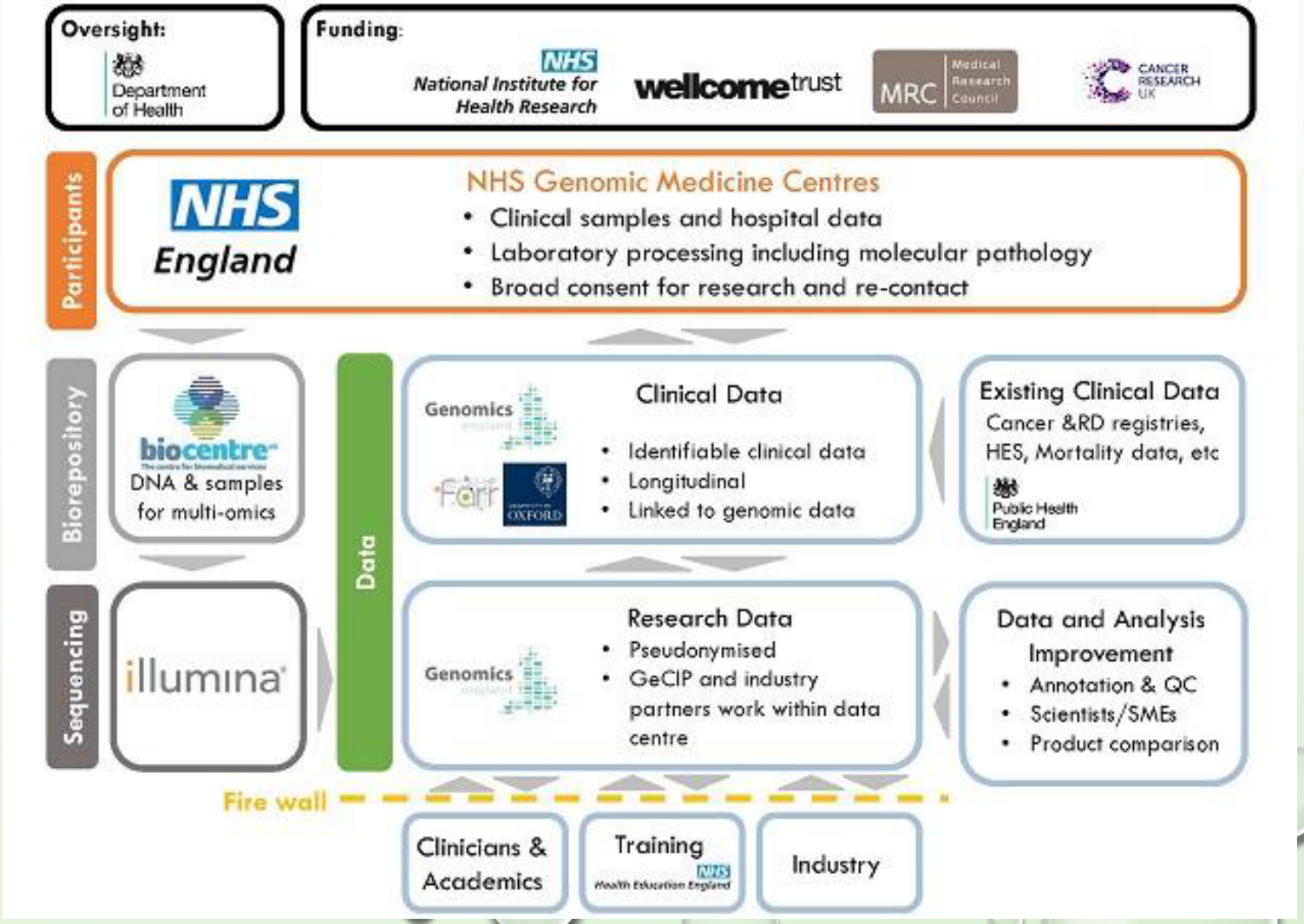
- New technologies enable to obtain and store large amount of biological and medical data.
- Both the quantity and type of data collected has changed dramatically, and is very diverse.
- To achieve impact on health we need to integrate and transform such data in information.
- This requires collaboration between disciplines: medicine, biology, genomics, engineering, ethics, law, big data, maths, stats, AI.



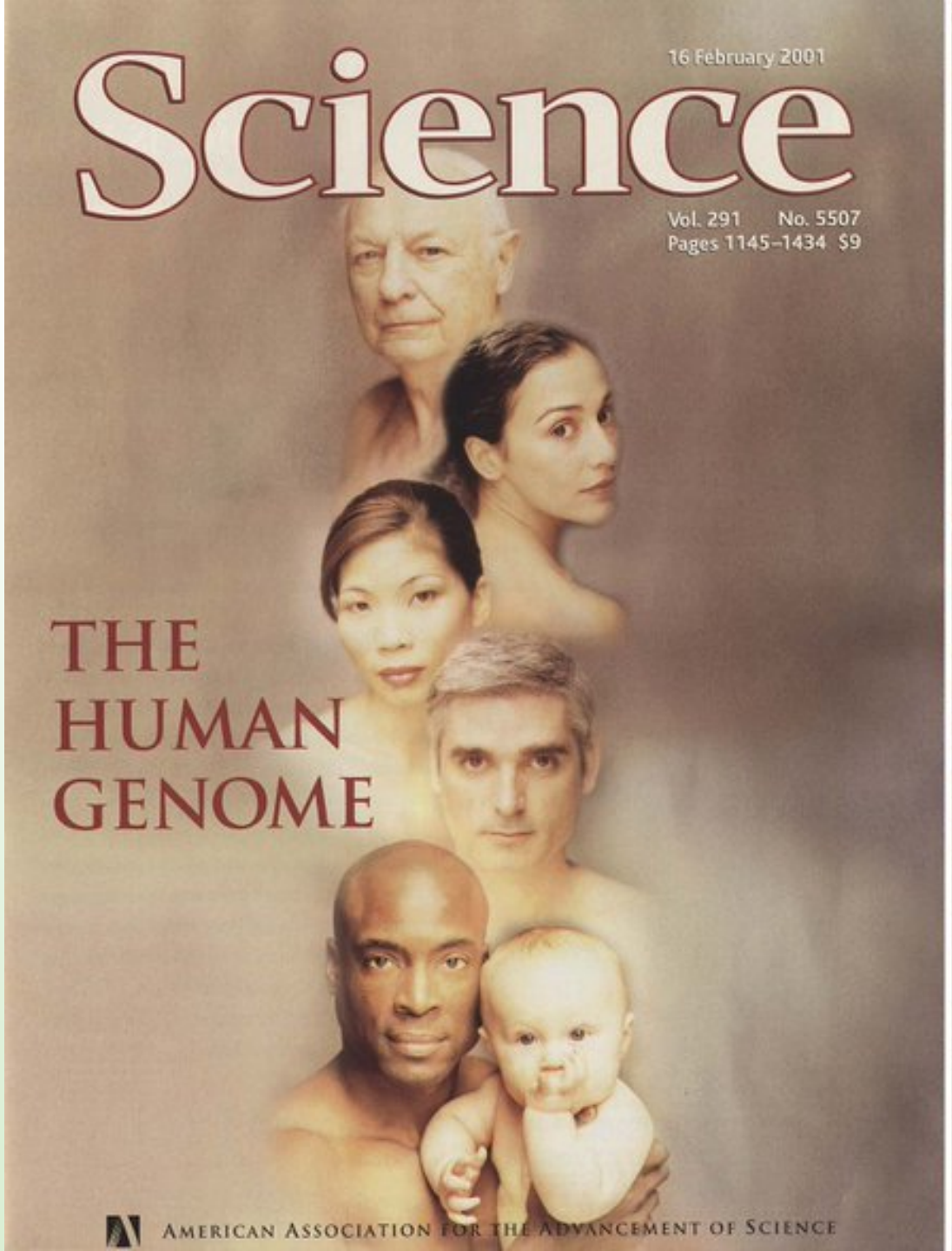
Denny & Collins, Cell, March 2021

WHAT IS BIOMEDICAL RAW DATA?

- SEQUENCING DATA
- MICROARRAY DATA
- PROTEIN STRUCTURE DATA
- MORPHOLOGICAL DATA
- IMAGES FROM MICROSCOPES
- CLINICAL IMAGING
- CLINICAL DATA
- ECOLOGICAL DATA
- BIOGEOGRAPHICAL DATA
- DEMOGRAPHIC DATA

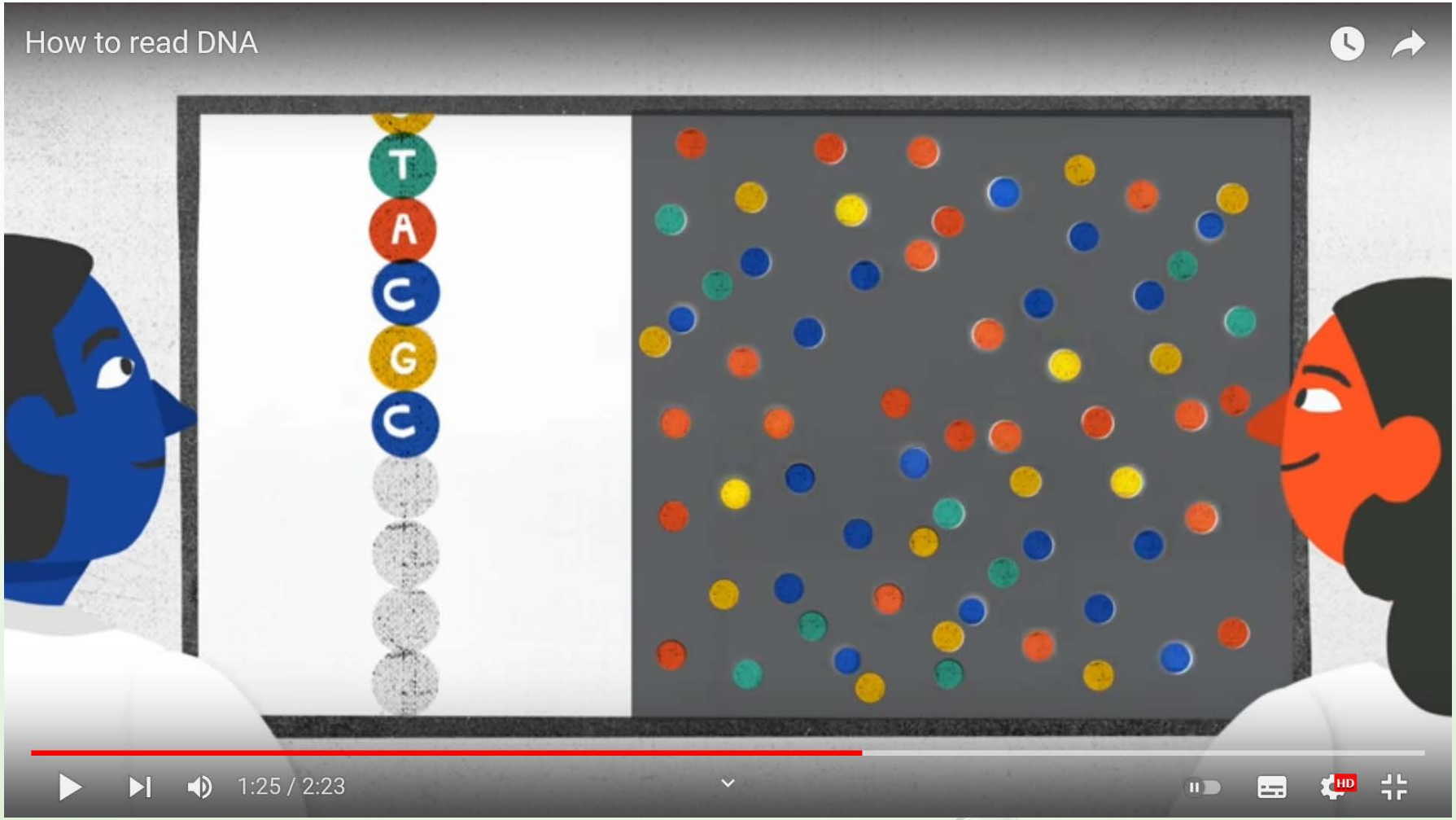


https://www.rcpath.org/profession/publications/college-bulletin/april-2021/the-100-000-genomes-project-research-and-clinical-legacy.html


Science
16 February 2001
Vol. 291 No. 5507
Pages 1145–1434 $9
THE HUMAN GENOME
AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE


15 February 2001
nature
$10.00
www.nature.com
the human genome
Nuclear fission
Five-dimensional energy landscapes
Seafloor spreading
The view from under the Arctic ice
Career prospects
Sequence creates new opportunities
naturejobs
genomics special

DNA: WE CAN READ IT AND WRITE IT



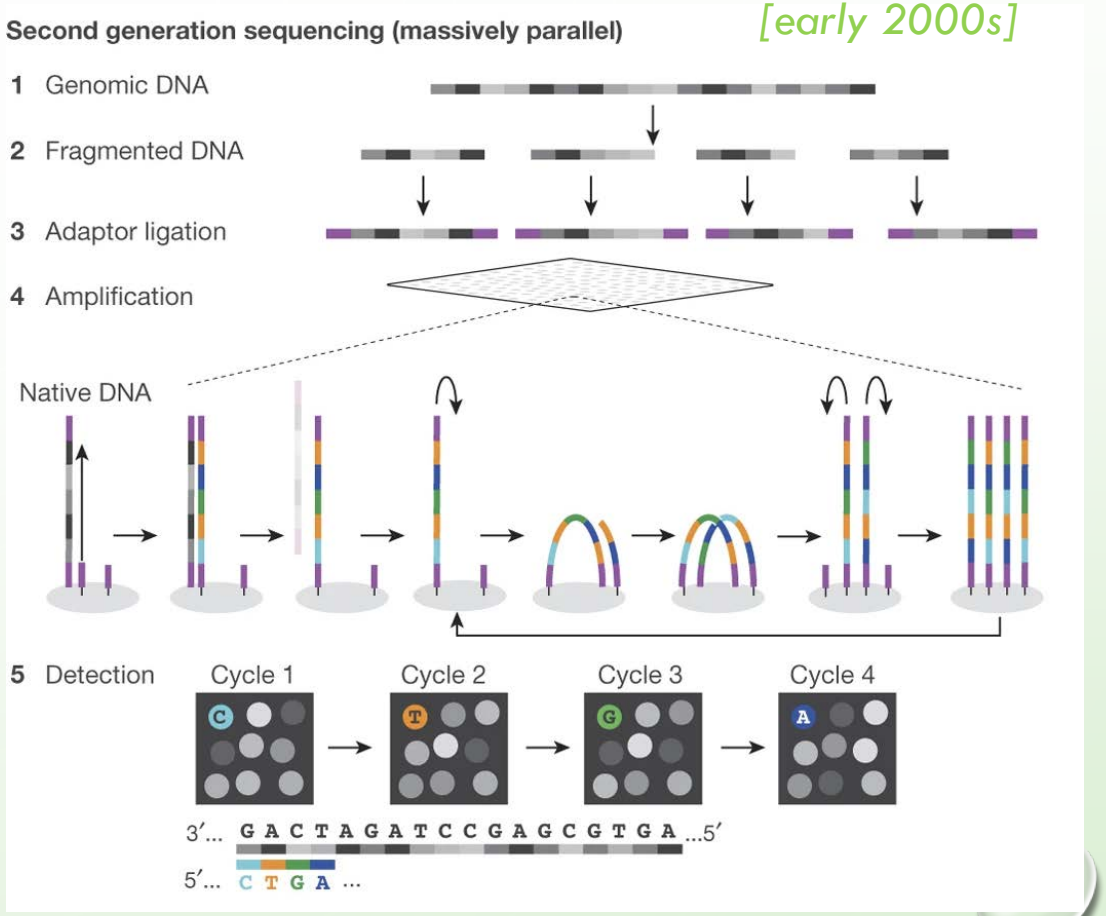
OxfordSparks

https://www.youtube.com/watch?v=bQwTa8Rf31A

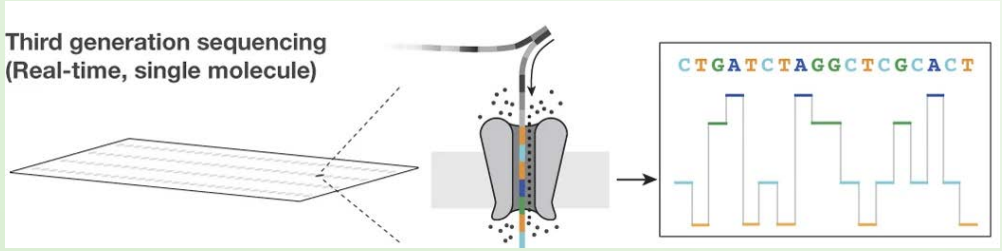
DNA SEQUENCING TECHNOLOGIES

[1977]

[early 2000s]

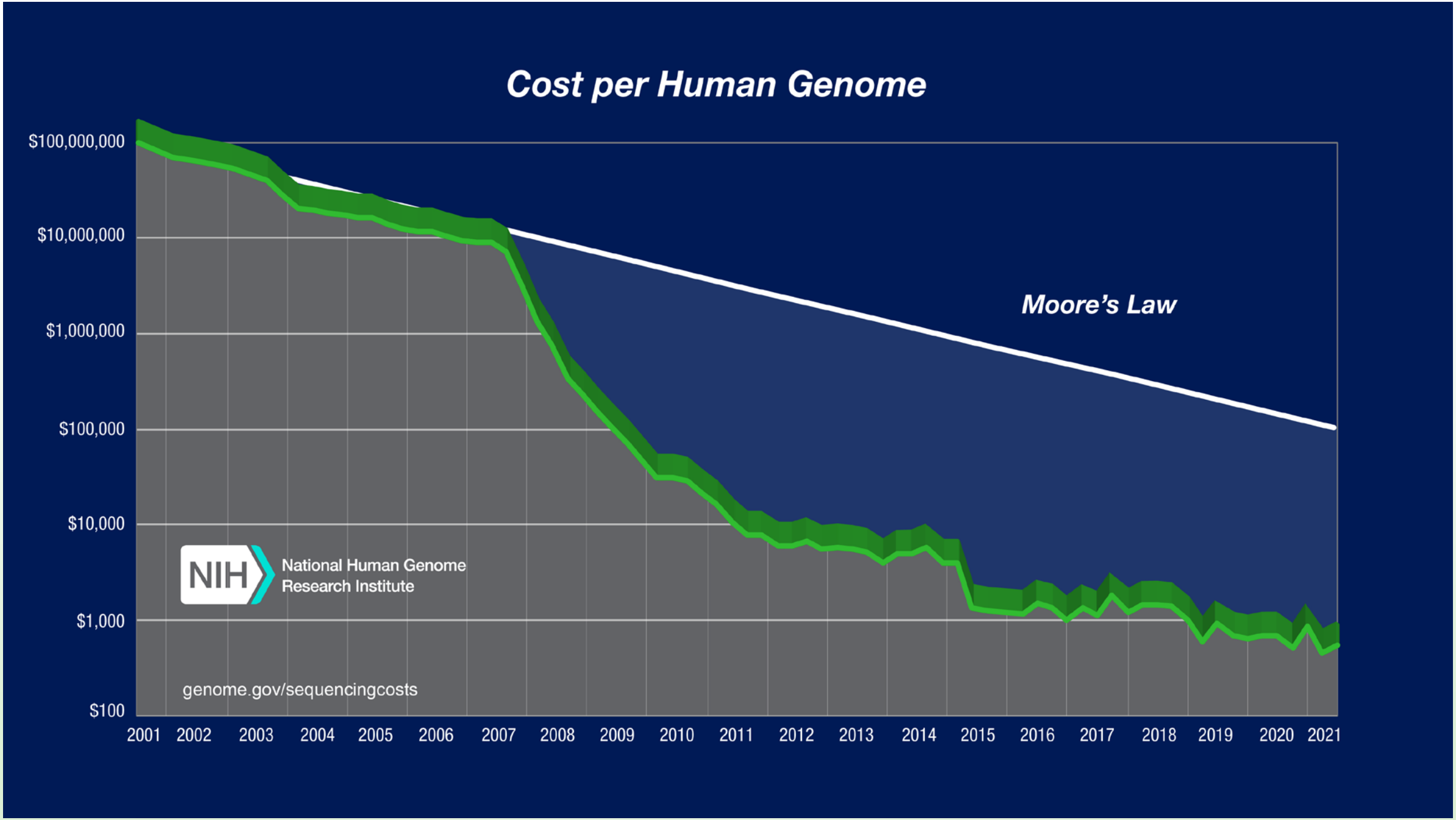


https://www.youtube.com/watch?v=ONGdehkB8jU

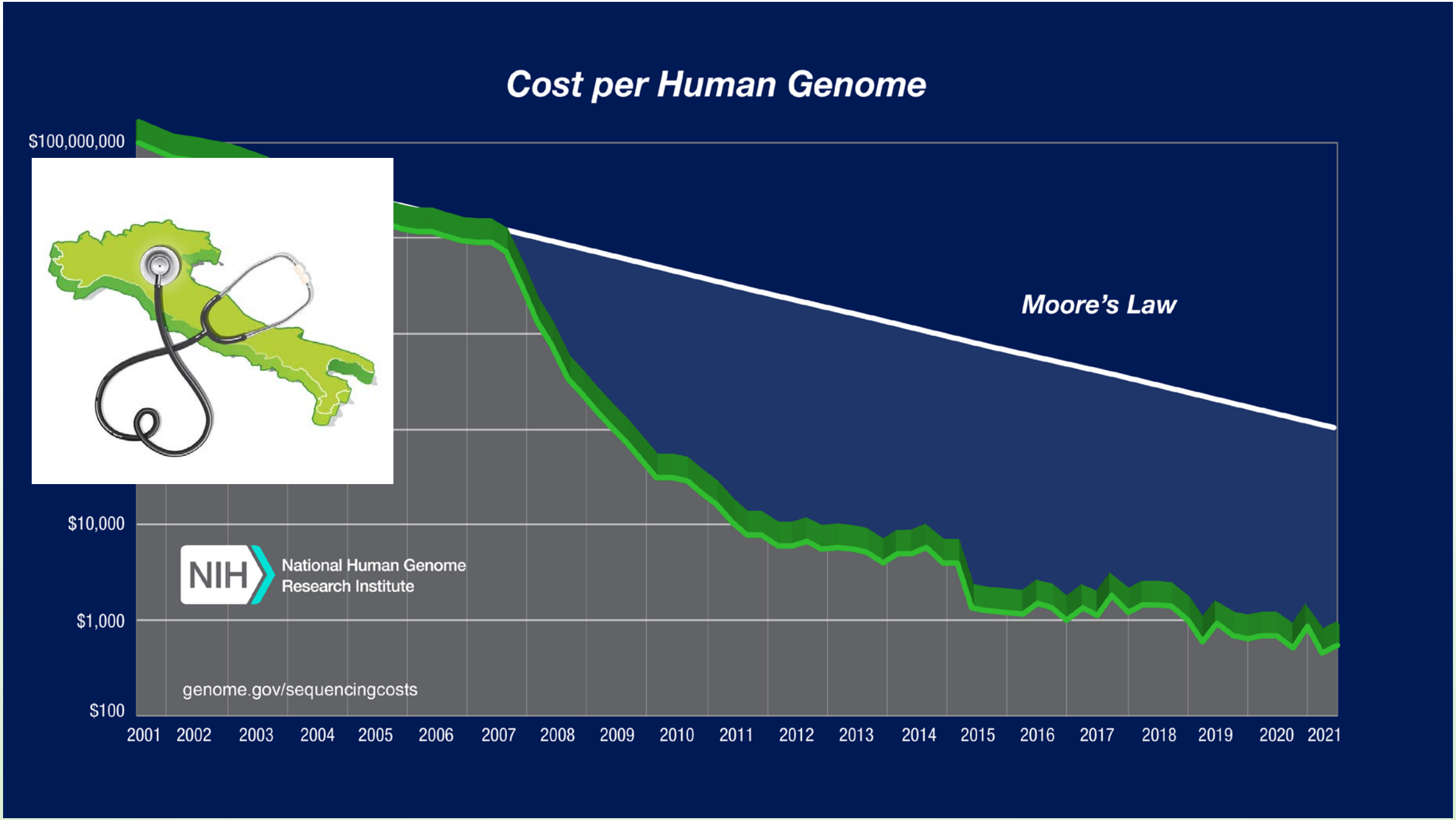


https://www.nature.com/articles/nature24286/

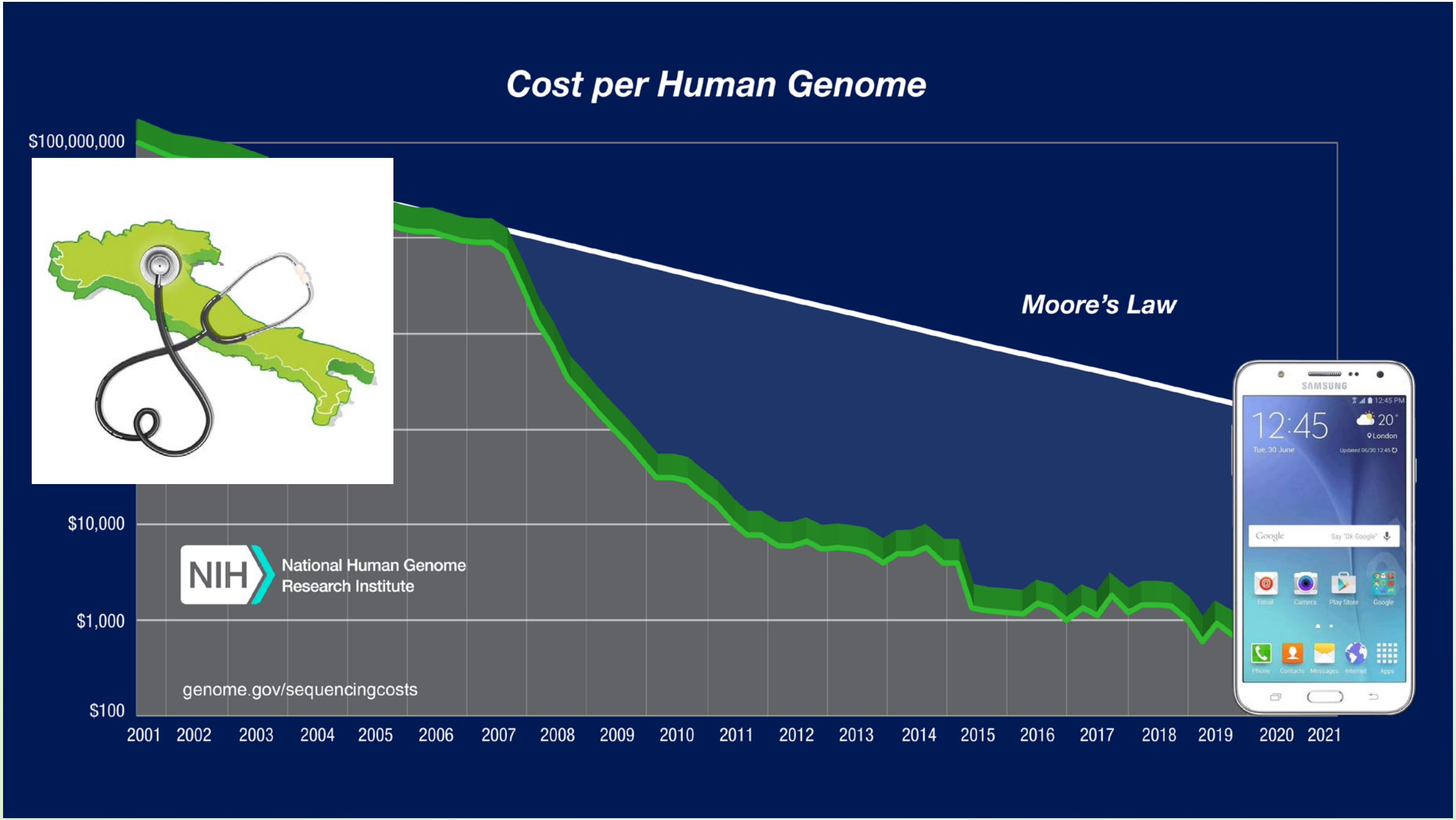
https://www.youtube.com/watch?v=v10bUR2aL5g



https://www.genome.gov/about-genomics/fact-sheets/Genomic-Data-Science

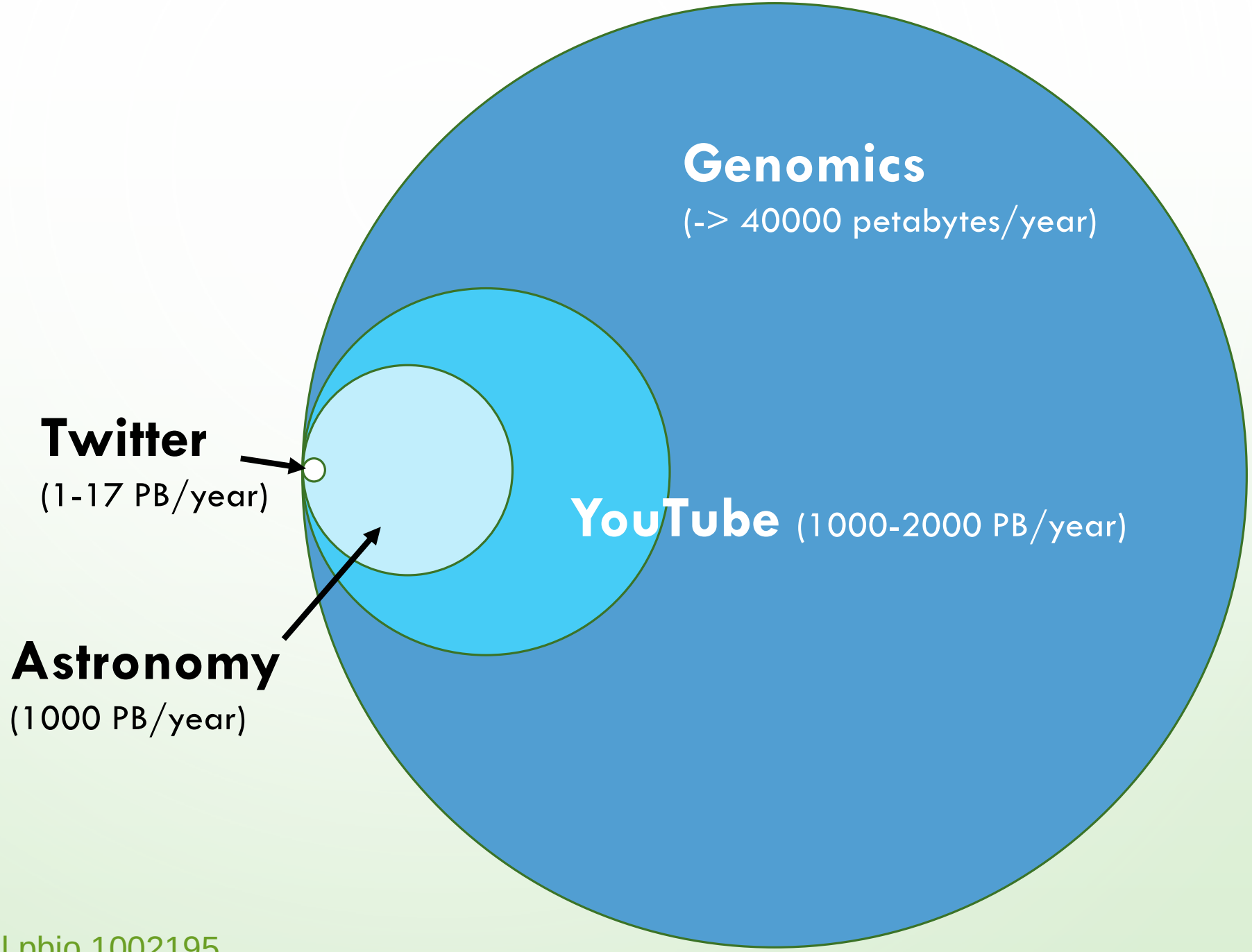


https://www.genome.gov/about-genomics/fact-sheets/Genomic-Data-Science



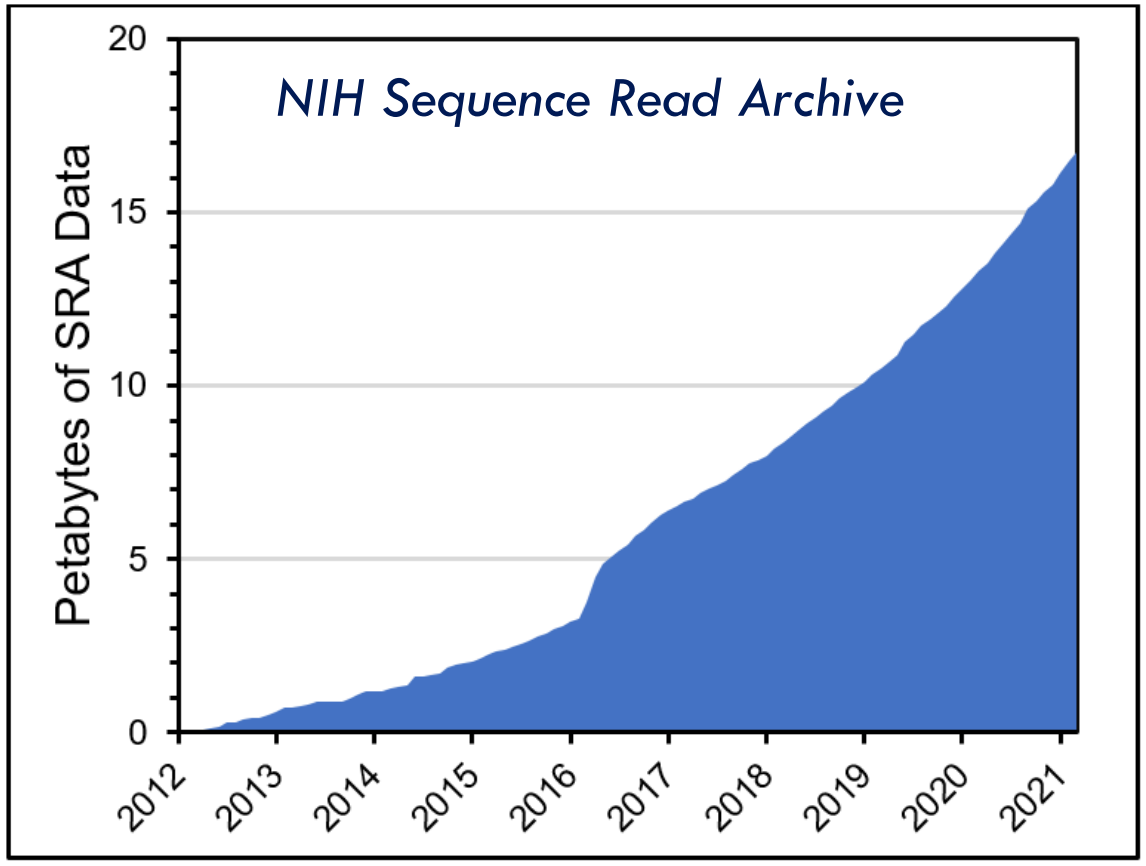
https://www.genome.gov/about-genomics/fact-sheets/Genomic-Data-Science

Big data

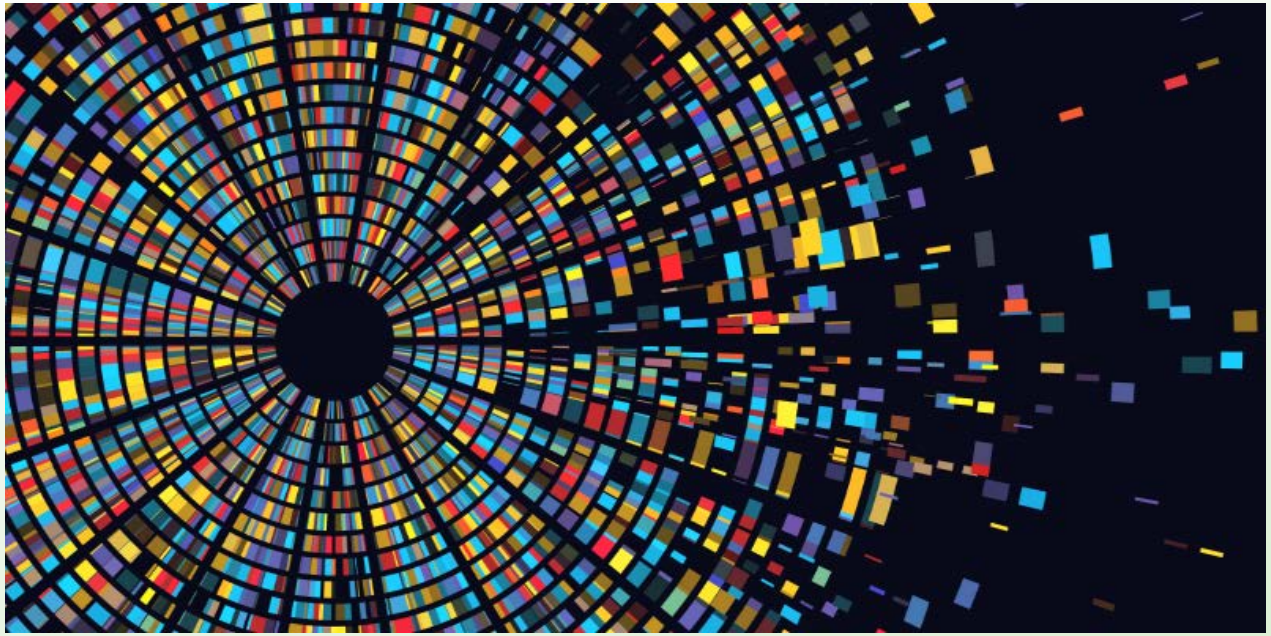


https://doi.org/10.1371/journal.pbio.1002195

TACKLING PETABYTE SCALE SEQUENCE SEARCH CHALLENGES



Dr. Susan Gregurick, NIH Associate Director for Data Science and ODSS Director, said: "We all share a common problem and a need to develop, enhance, and implement methods that streamline data access, search or findability, and ultimately data reuse."



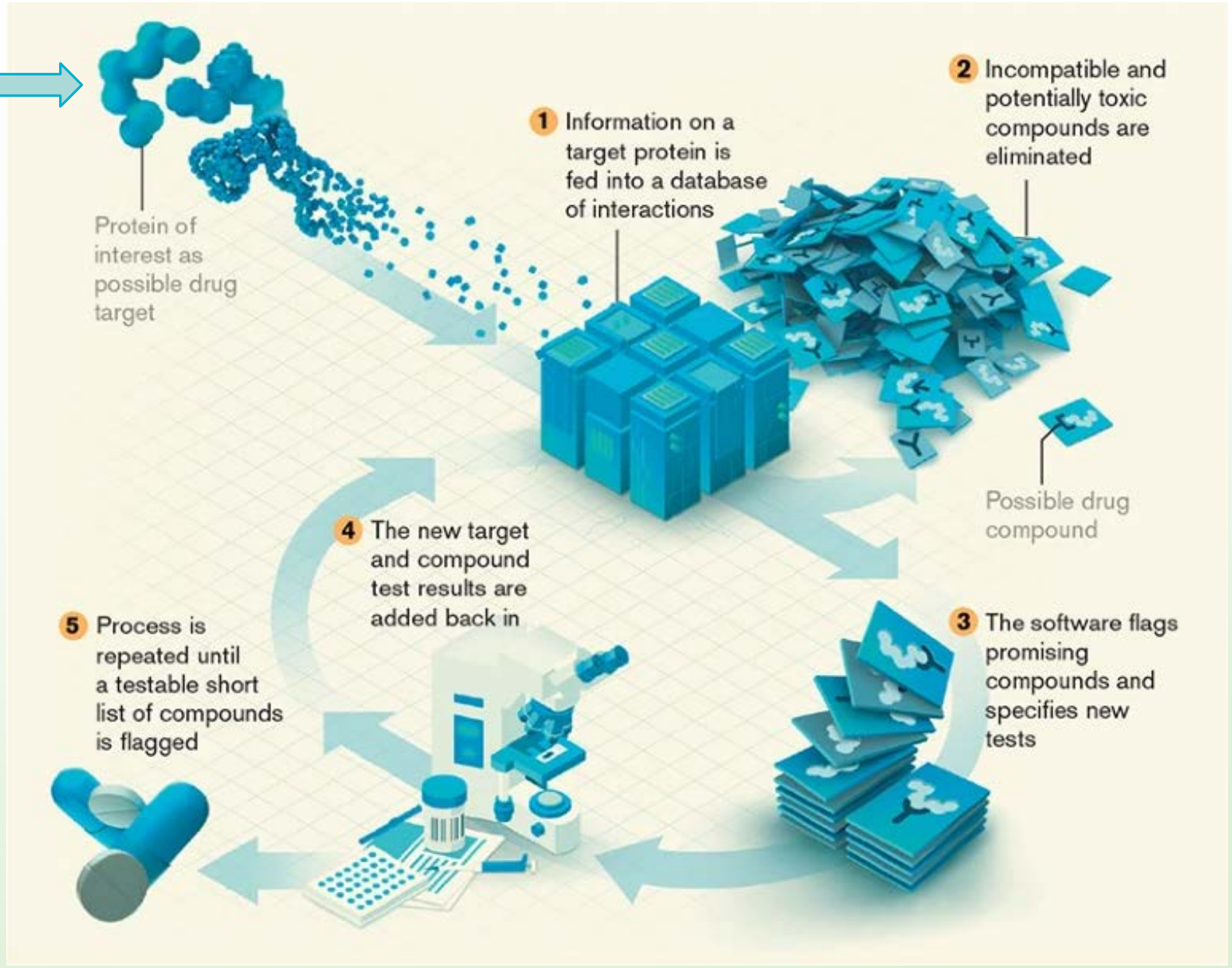
https://ncbiinsights.ncbi.nlm.nih.gov/2021/08/09/espsss-workshop/

Your genes +

Your environment & lifestyle =

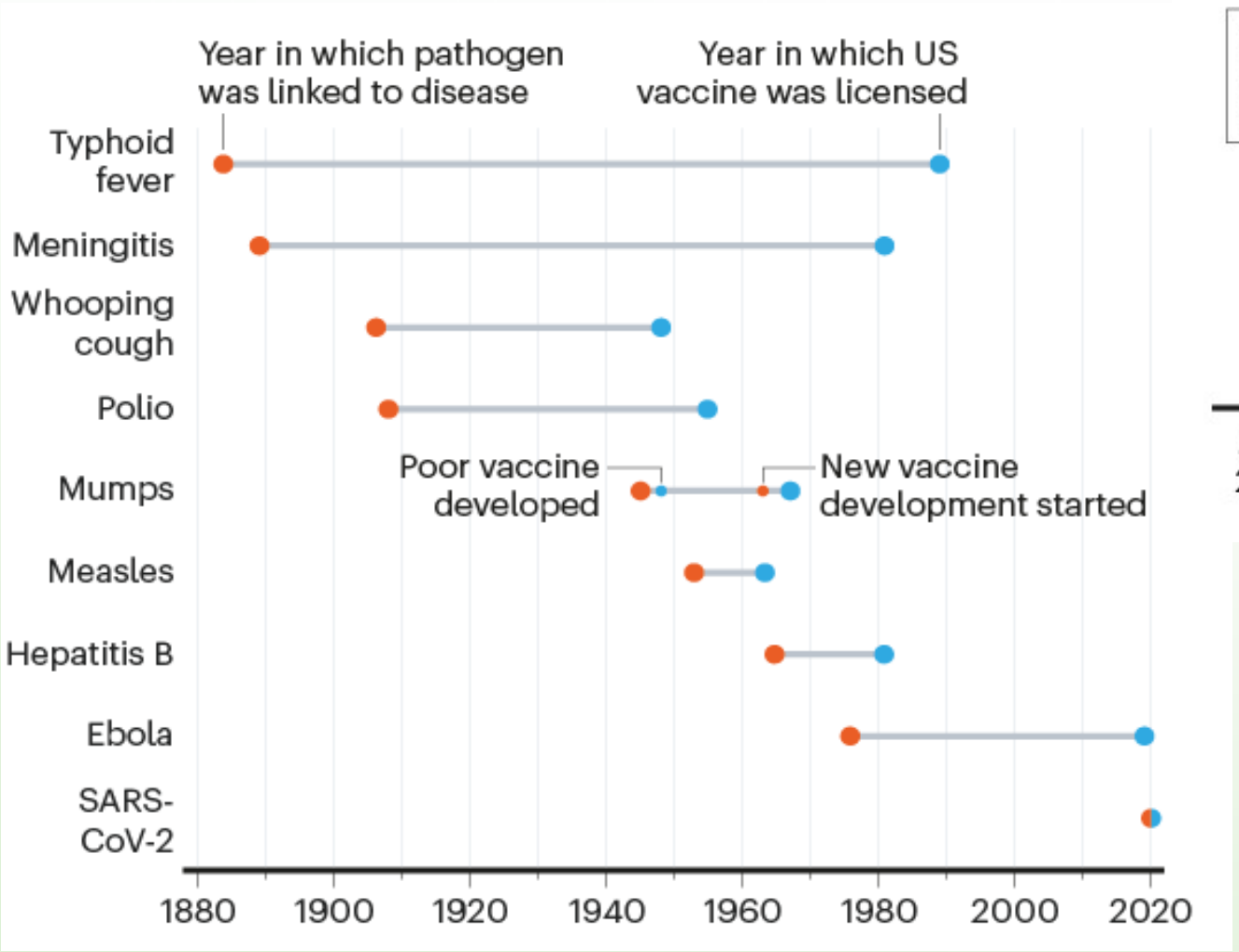
YOU!

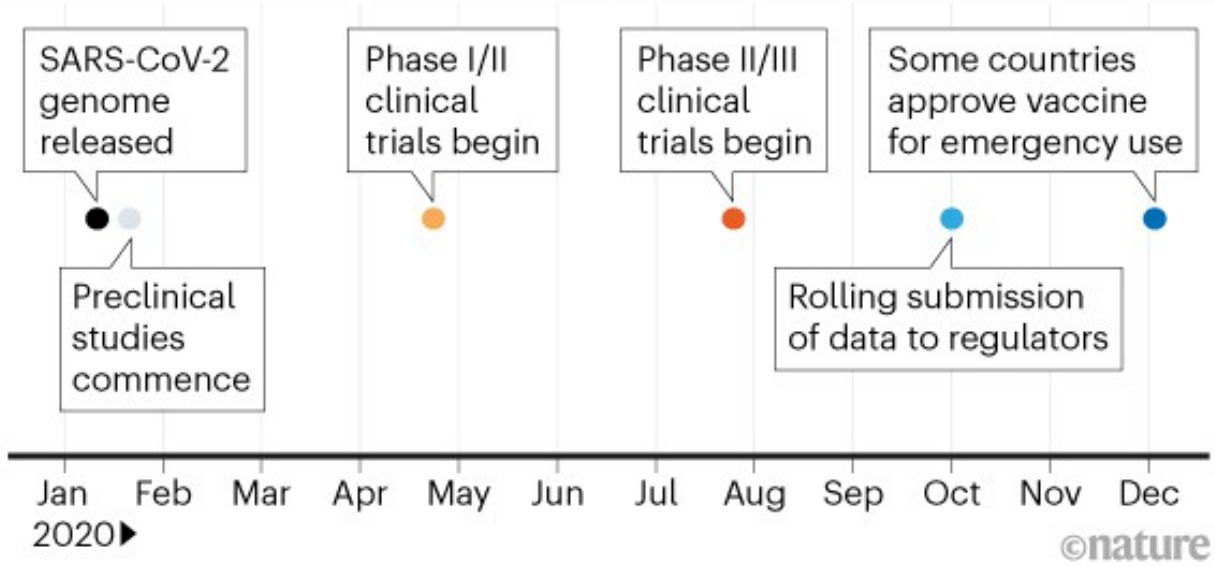
SPEEDING UP THE SEARCH FOR DRUGS WITH AI



https://www.nature.com/articles/d41586-019-03846-0

AI and vaccine development



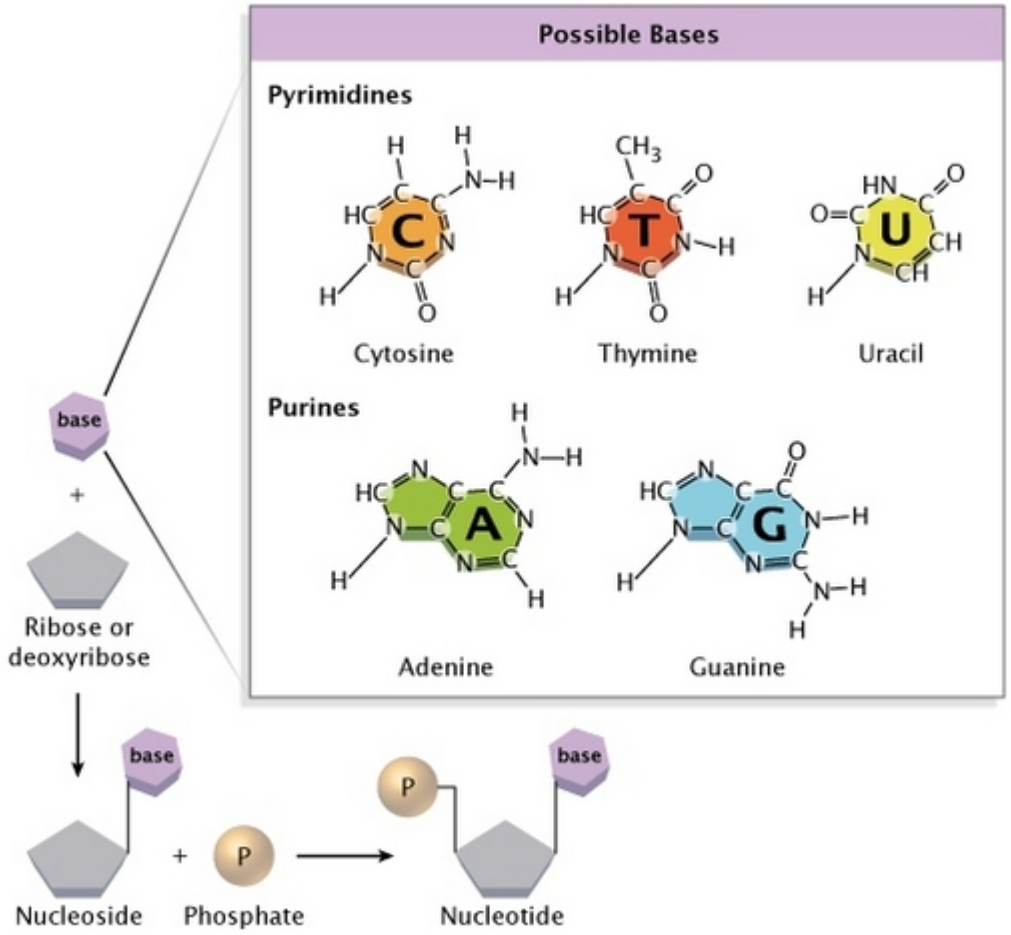


Dave Johnson, PhD in Information Physics, chief data and AI officer at Moderna: "We've seen how this digital infrastructure and how these algorithms can really help push things forward".

https://sloanreview.mit.edu/audio/ai-and-the-covid-19-vaccine-modernas-dave-johnson/

Nature News, 18 December 2020

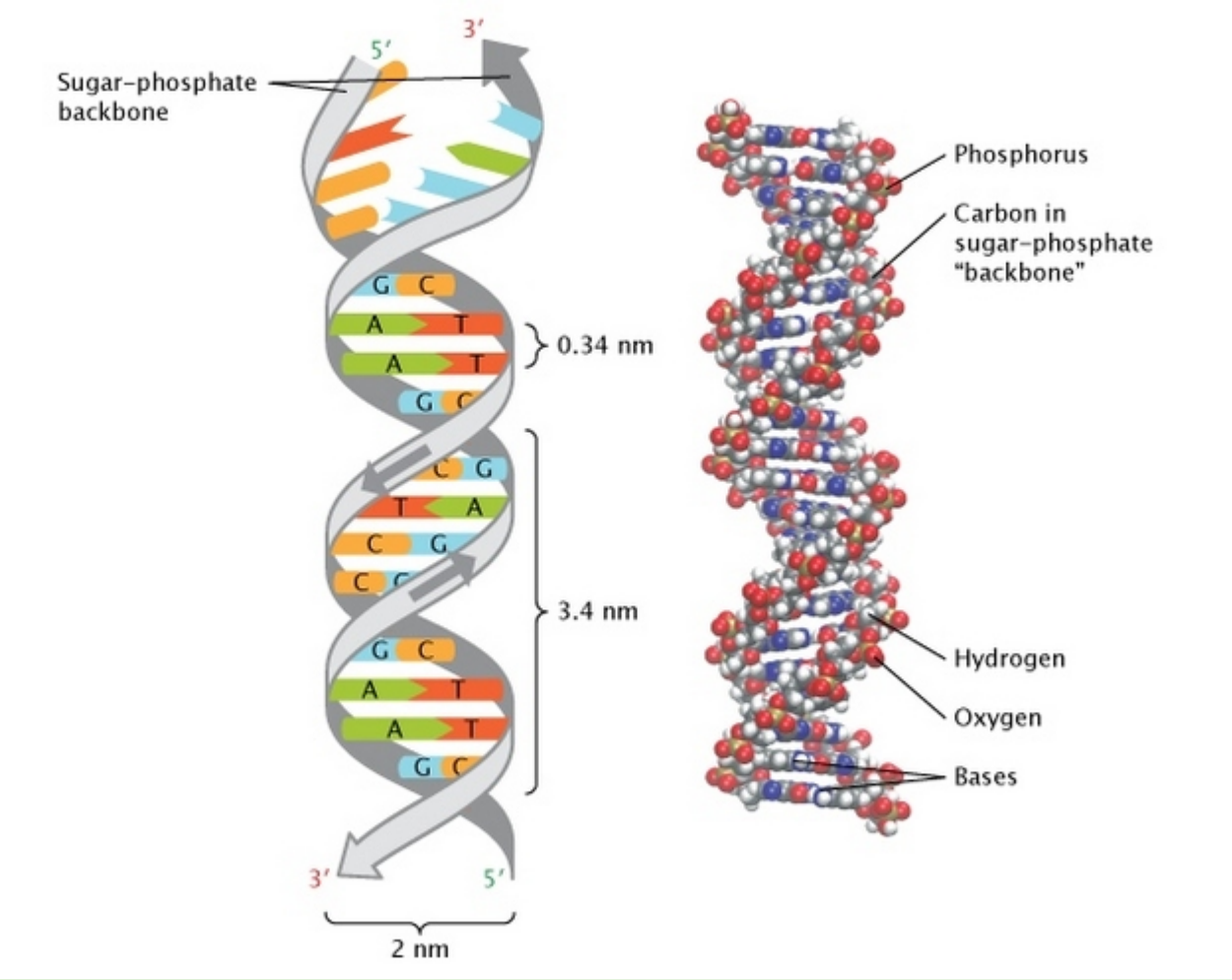
DNA/RNA STRING REPRESENTATION



- A (ADENINE)
- T (THYMINE)
- C (CYTOSINE)
- G (GUANINE)

https://www.nature.com/scitable/topicpage/discovery-of-dna-structure-and-function-watson-397/

DNA/RNA STRING REPRESENTATION



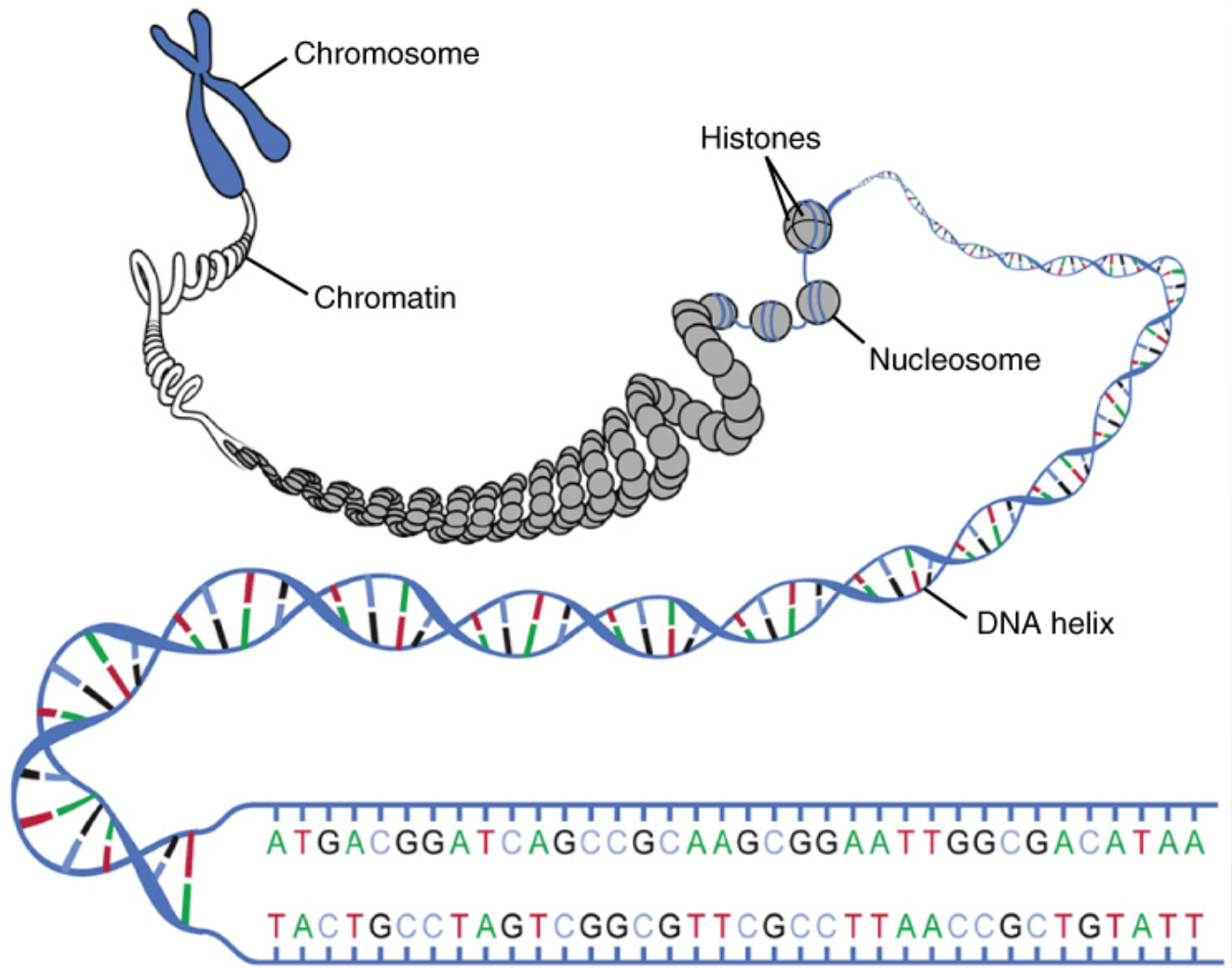
- A (ADENINE)
- T (THYMINE)
- C (CYTOSINE)
- G (GUANINE)

- ORGANIZED IN A DOUBLE HELIX OF PAIRED BASES: A-T, C-G

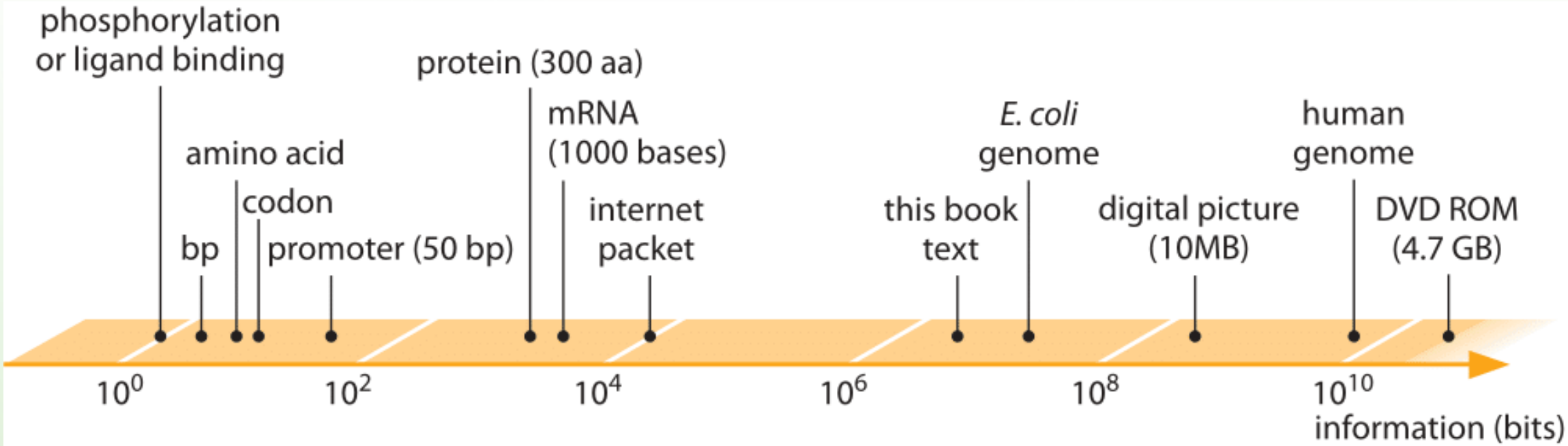
- IN RNA T -> U (URACIL)

https://www.nature.com/scitable/topicpage/discovery-of-dna-structure-and-function-watson-397/

THE DIGITAL CODE OF DNA



INFORMATION CONTENT: BIOLOGICAL ENTITIES VS STORAGE DEVICES



information, $I = \log_2$ (# possible configurations)

e.g. base pair (bp) has four possibilities $\Rightarrow I = \log_2(4) = 2$

http://book.bionumbers.org/wp-content/uploads/2014/08/500-f1-InformationAxis-1.png

DNA SEQUENCING DATA

Base Calling
(Vendor tool, non-text)

FASTQ files
Sequencing Reads

FASTQ

De facto standard for storing high-throughput sequencing data

- ☐ Text-based file containing raw + quality
- ☐ Can contain millions of entries (size ~GBs)

```
@ML-P2-14:9:000H003HG:1:11102:17290:1073 1:N:0:TCCTGAGC+GCGATCTA
TTTGGTAACAGCATGAATTATTCTAGCCACTAAAACTCTATGAACATCTTGTGAAGGTTTCAGATAGAGCCTGAAGTACACAGAGAACAATTCTTAAAAAA
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE<AEEEEEEE
```

1) Begins with @, followed by sequence identifier and optional additional information/description
2) The sequence: the base calls A, C, T, G and N
3) A separator, which is simply a plus (+) sign.
4) The base call quality scores. It represents the probability of an error in base call. These are Phred +33 encoded, using ASCII characters to represent the numerical quality scores.

DNA SEQUENCING DATA

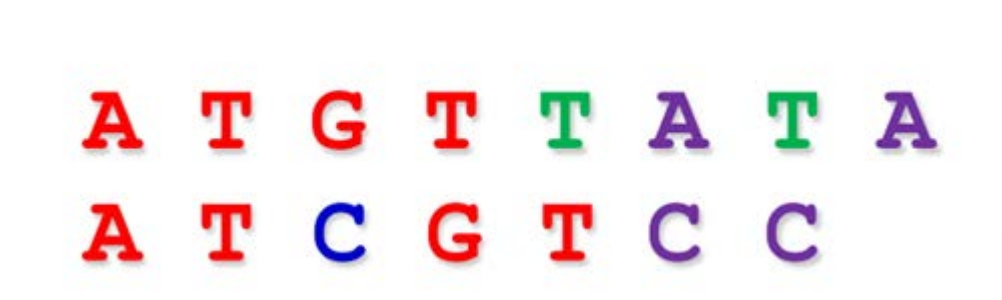
Base Calling
(Vendor tool, non-text)

FASTQ files
Sequencing Reads

Alignment or assembly

SAM/BAM files:
Aligned sequencing reads
.sam: uncompressed text file
.bam: compressed text file

ALIGNMENT: SEQUENCE MATCHING PROBLEM

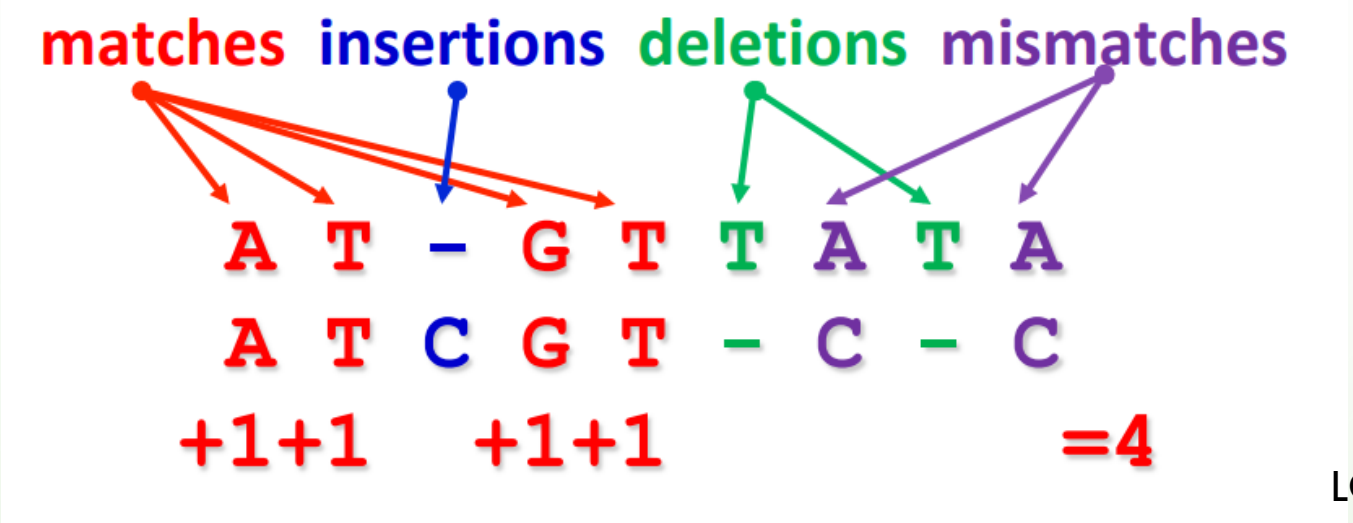


Finding regions of similarities and dissimilarities between sequences, and infer a measure of relatedness, is vital for phylogenetic research:

Comparison of genetic material across organisms allows for example to:

- Infer functional relationships
- Infer evolutionary relationships

ALIGNMENT: SEQUENCE MATCHING PROBLEM



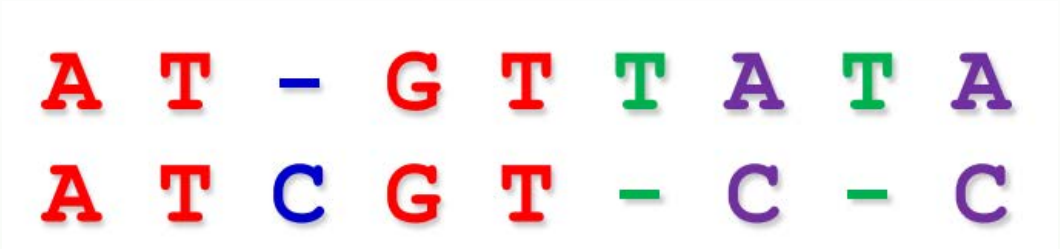
LONGEST COMMON SEQUENCE

Alignment of two sequences is a two-row matrix:

1st row: symbols of the 1st sequence (in order) interspersed by "-"

2nd row: symbols of the 2nd sequence (in order) interspersed by "-"

LONGEST COMMON SUBSEQUENCE



Matches in alignment of two sequences (ATGT) form their Common Subsequence

Longest Common Subsequence Problem:
Find a longest common subsequence of two strings.

- Input: Two strings.
- Output: A longest common subsequence of these strings.

For arbitrary number of input sequences, the problem is NP-hard.
When the number of sequences is constant, the problem is solvable in polynomial time with a number of approaches

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5425171/

SEQUENCE ALIGNMENT MAP - SAM FILE

TAB-delimited text format for storing alignment sequences against a reference

```
@HD     VN:1.0  SO:coordinate
@SQ     SN:chr20        LN:64444167
@PG     ID:TopHat       VN:2.0.14       CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-rea
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr
20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714        16      chr20   190930  3       100M    *       0       0
       CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C    BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
    AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714   HI:i:0
HWI-ST1145:74:C101DACXX:7:1114:2759:41961        16      chr20   193953  50      100M    *       0       0
       TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGGGGCATGGACGA
G    DCDDDDEDDDDDDDCDDDDDDDCCCDDDCDDDDDEEC>DFFFEJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJIHJJJJJJHHHHHFFFFFCCC
    AS:i:-16        XM:i:3  XO:i:0  XG:i:0  MD:Z:60G16T18T3 NM:i:3  NH:i:1
HWI-ST1145:74:C101DACXX:7:1204:14760:4030        16      chr20   270877  50      100M    *       0       0
       GGCTTTATTGGTAAAAAAGGAATAGCAGATTTAATCAGAAATTCCCACCTGGCCCAGCAGCACCAACCAGAAAGAAGGGAAGAAGACAGGAAAAAACCA
C    DDDDDDDDDCCDDDDDDDDDEEEEEEEFFFEFFEGHHHHFGDJJIHJJIJIJJJIIIIGGFJJIHIIIIJJJJJJIGHHFAHGFHJHFGGHFFFDD@BB
    AS:i:-11        XM:i:2  XO:i:0  XG:i:0  MD:Z:0A85G13    NM:i:2  NH:i:1
HWI-ST1145:74:C101DACXX:7:1210:11167:8699        0       chr20   271218  50      50M4700N50M     *       0
       0       GTGGCTCTTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGGTGCACTTGGGTCTCCGAAGCAGAACATCCTCAAATATGACCTCTCG
accepted_hits.sam
```

Binary Alignment/Map (BAM) file (*.bam) is a compressed binary SAM file (smaller size + faster access)

SEQUENCE ALIGNMENT MAP - SAM FILE

TAB-delimited text format for storing alignment sequences against a reference

```
@HD     VN:1.0  SO:coordinate
@SQ     SN:chr20        LN:64444167
@PG     ID:TopHat       VN:2.0.14       CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-rea
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr
20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714        16      chr20   190930  3       100M    *       0       0
       CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C    BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
    AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714   HI:i:0
HWI-ST1145:74:C101DACXX:7:1114:2759:41961        16      chr20   193953  50      100M    *       0       0
       TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGGGGCATGGACGA
G    DCDDDDEDDDDDDDCDDDDDDDCCCDDDCDDDDDEEC>DFFFEJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJIHJJJJJJHHHHHFFFFFCCC
    AS:i:-16        XM:i:3  XO:i:0  XG:i:0  MD:Z:60G16T18T3 NM:i:3  NH:i:1
HWI-ST1145:74:C101DACXX:7:1204:14760:4030        16      chr20   270877  50      100M    *       0       0
       GGCTTTATTGGTAAAAAAGGAATAGCAGATTTAATCAGAAATTCCCACCTGGCCCAGCAGCACCAACCAGAAAGAAGGGAAGAAGACAGGAAAAAACCA
C    DDDDDDDDDCCDDDDDDDDDEEEEEEEFFFEFFEGHHHHFGDJJIHJJIJIJJJIIIIGGFJJIHIIIIJJJJJJIGHHFAHGFHJHFGGHFFFDD@BB
    AS:i:-11        XM:i:2  XO:i:0  XG:i:0  MD:Z:0A85G13    NM:i:2  NH:i:1
HWI-ST1145:74:C101DACXX:7:1210:11167:8699        0       chr20   271218  50      50M4700N50M     *       0
       0       GTGGCTCTTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGGTGCACTTGGGTCTCCGAAGCAGAACATCCTCAAATATGACCTCTCG
accepted_hits.sam
```

File-level metadata. Optional.

Binary Alignment/Map (BAM) file (*.bam) is a compressed binary SAM file (smaller size + faster access)

SEQUENCE ALIGNMENT MAP - SAM FILE

TAB-delimited text format for storing alignment sequences against a reference

```
@HD     VN:1.0  SO:coordinate
@SQ     SN:chr20        LN:64444167
@PG     ID:TopHat       VN:2.0.14       CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-rea
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr
20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714        16      chr20   190930  3       100M    *       0       0
       CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C    BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
    AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714   HI:i:0
HWI-ST1145:74:C101DACXX:7:1114:2759:41961        16      chr20   193953  50      100M    *       0       0
       TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGGGGCATGGACGA
G    DCDDDDEDDDDDDDCDDDDDDDCCCDDDCDDDDDEEC>DFFFEJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJIHJJJJJJHHHHHFFFFFCCC
    AS:i:-16        XM:i:3  XO:i:0  XG:i:0  MD:Z:60G16T18T3 NM:i:3  NH:i:1
HWI-ST1145:74:C101DACXX:7:1204:14760:4030        16      chr20   270877  50      100M    *       0       0
       GGCTTTATTGGTAAAAAAGGAATAGCAGATTTAATCAGAAATTCCCACCTGGCCCAGCAGCACCAACCAGAAAGAAGGGAAGAAGACAGGAAAAAACCA
C    DDDDDDDDDCCDDDDDDDDDEEEEEEEFFFEFFEGHHHHFGDJJIHJJIJIJJJIIIIGGFJJIHIIIIJJJJJJIGHHFAHGFHJHFGGHFFFDD@BB
    AS:i:-11        XM:i:2  XO:i:0  XG:i:0  MD:Z:0A85G13    NM:i:2  NH:i:1
HWI-ST1145:74:C101DACXX:7:1210:11167:8699        0       chr20   271218  50      50M4700N50M     *       0
       0       GTGGCTCTTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGGTGCACTTGGGTCTCCGAAGCAGAACATCCTCAAATATGACCTCTCG
accepted_hits.sam
```

Reference sequence dictionary

Binary Alignment/Map (BAM) file (*.bam) is a compressed binary SAM file (smaller size + faster access)

SEQUENCE ALIGNMENT MAP - SAM FILE

TAB-delimited text format for storing alignment sequences against a reference

```
@HD     VN:1.0  SO:coordinate
@SQ     SN:chr20        LN:64444167
@PG     ID:TopHat       VN:2.0.14       CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-rea
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr
20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714        16      chr20   190930  3       100M    *       0       0
       CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C    BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
    AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714   HI:i:0
HWI-ST1145:74:C101DACXX:7:1114:2759:41961        16      chr20   193953  50      100M    *       0       0
       TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGGGGCATGGACGA
G    DCDDDDEDDDDDDDCDDDDDDDCCCDDDCDDDDDEEC>DFFFEJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJIHJJJJJJHHHHHFFFFFCCC
    AS:i:-16        XM:i:3  XO:i:0  XG:i:0  MD:Z:60G16T18T3 NM:i:3  NH:i:1
HWI-ST1145:74:C101DACXX:7:1204:14760:4030        16      chr20   270877  50      100M    *       0       0
       GGCTTTATTGGTAAAAAAGGAATAGCAGATTTAATCAGAAATTCCCACCTGGCCCAGCAGCACCAACCAGAAAGAAGGGAAGAAGACAGGAAAAAACCA
C    DDDDDDDDDCCDDDDDDDDDEEEEEEEFFFEFFEGHHHHFGDJJIHJJIJIJJJIIIIGGFJJIHIIIIJJJJJJIGHHFAHGFHJHFGGHFFFDD@BB
    AS:i:-11        XM:i:2  XO:i:0  XG:i:0  MD:Z:0A85G13    NM:i:2  NH:i:1
HWI-ST1145:74:C101DACXX:7:1210:11167:8699        0       chr20   271218  50      50M4700N50M     *       0
       0       GTGGCTCTTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGGTGCACTTGGGTCTCCGAAGCAGAACATCCTCAAATATGACCTCTCG
accepted_hits.sam
```

Program

Binary Alignment/Map (BAM) file (*.bam) is a compressed binary SAM file (smaller size + faster access)

SEQUENCE ALIGNMENT MAP - SAM FILE

TAB-delimited text format for storing alignment sequences against a reference

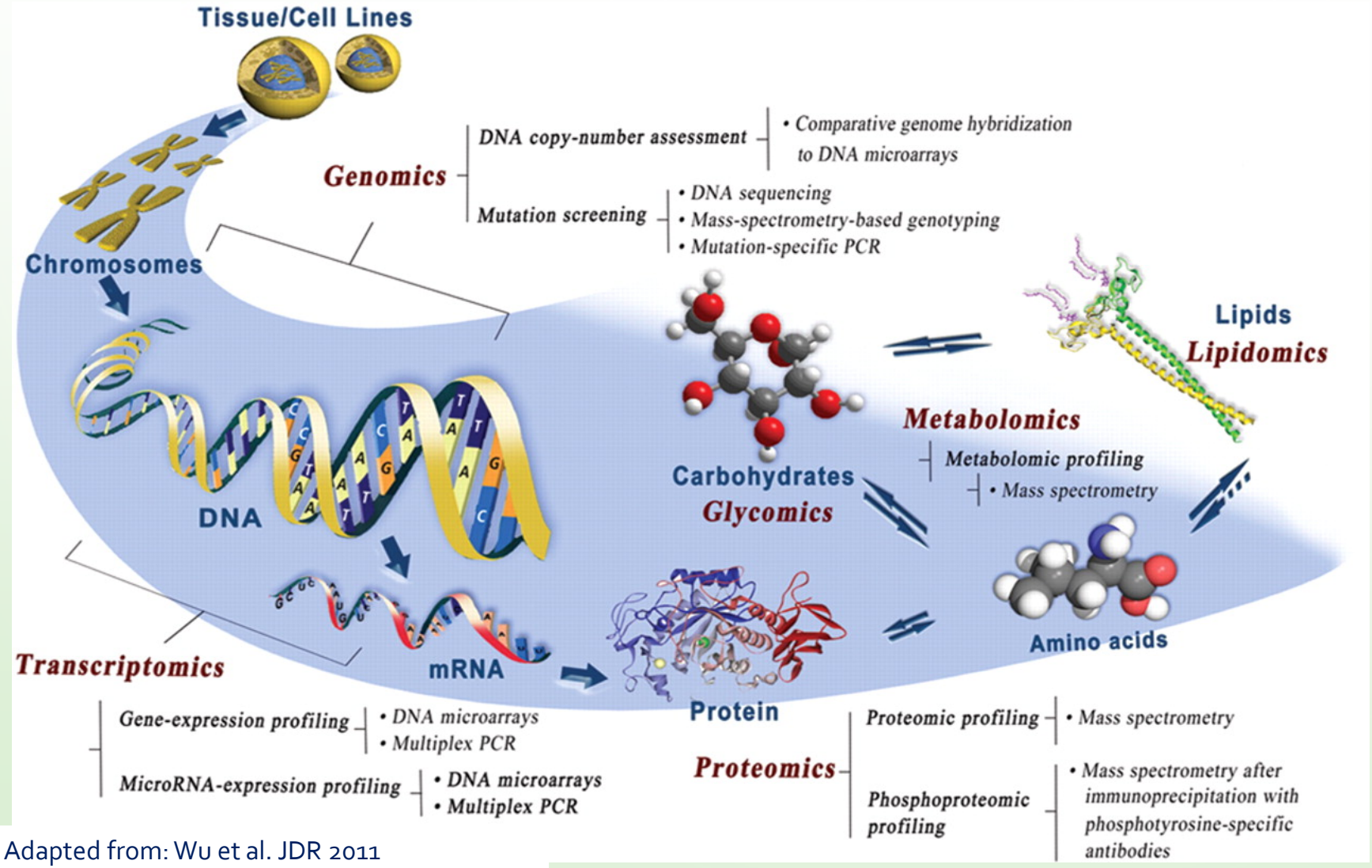
```
@HD     VN:1.0  SO:coordinate
@SQ     SN:chr20        LN:64444167
@PG     ID:TopHat       VN:2.0.14       CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-rea
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr
20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714        16      chr20   190930  3       100M    *       0       0
       CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C    BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
    AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714   HI:i:0
HWI-ST1145:74:C101DACXX:7:1114:2759:41961        16      chr20   193953  50      100M    *       0       0
       TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGGGGCATGGACGA
G    DCDDDDEDDDDDDDCDDDDDDDCCCDDDCDDDDDEEC>DFFFEJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJIHJJJJJJHHHHHFFFFFCCC
    AS:i:-16        XM:i:3  XO:i:0  XG:i:0  MD:Z:60G16T18T3 NM:i:3  NH:i:1
HWI-ST1145:74:C101DACXX:7:1204:14760:4030        16      chr20   270877  50      100M    *       0       0
       GGCTTTATTGGTAAAAAAGGAATAGCAGATTTAATCAGAAATTCCCACCTGGCCCAGCAGCACCAACCAGAAAGAAGGGAAGAAGACAGGAAAAAACCA
C    DDDDDDDDDCCDDDDDDDDDEEEEEEEFFFEFFEGHHHHFGDJJIHJJIJIJJJIIIIGGFJJIHIIIIJJJJJJIGHHFAHGFHJHFGGHFFFDD@BB
    AS:i:-11        XM:i:2  XO:i:0  XG:i:0  MD:Z:0A85G13    NM:i:2  NH:i:1
HWI-ST1145:74:C101DACXX:7:1210:11167:8699        0       chr20   271218  50      50M4700N50M     *       0
       0       GTGGCTCTTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGGTGCACTTGGGTCTCCGAAGCAGAACATCCTCAAATATGACCTCTCG
accepted_hits.sam
```

Alignment information

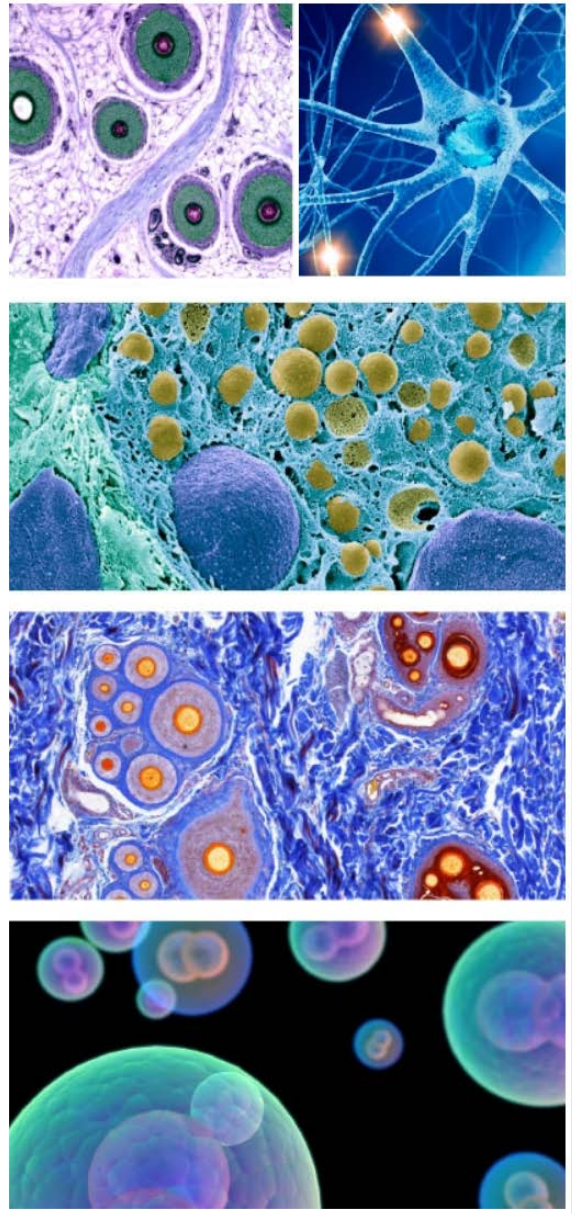
http://samtools.github.io/hts-specs/SAMv1.pdf

Binary Alignment/Map (BAM) file (*.bam) is a compressed binary SAM file (smaller size + faster access)

From genotype to phenotype

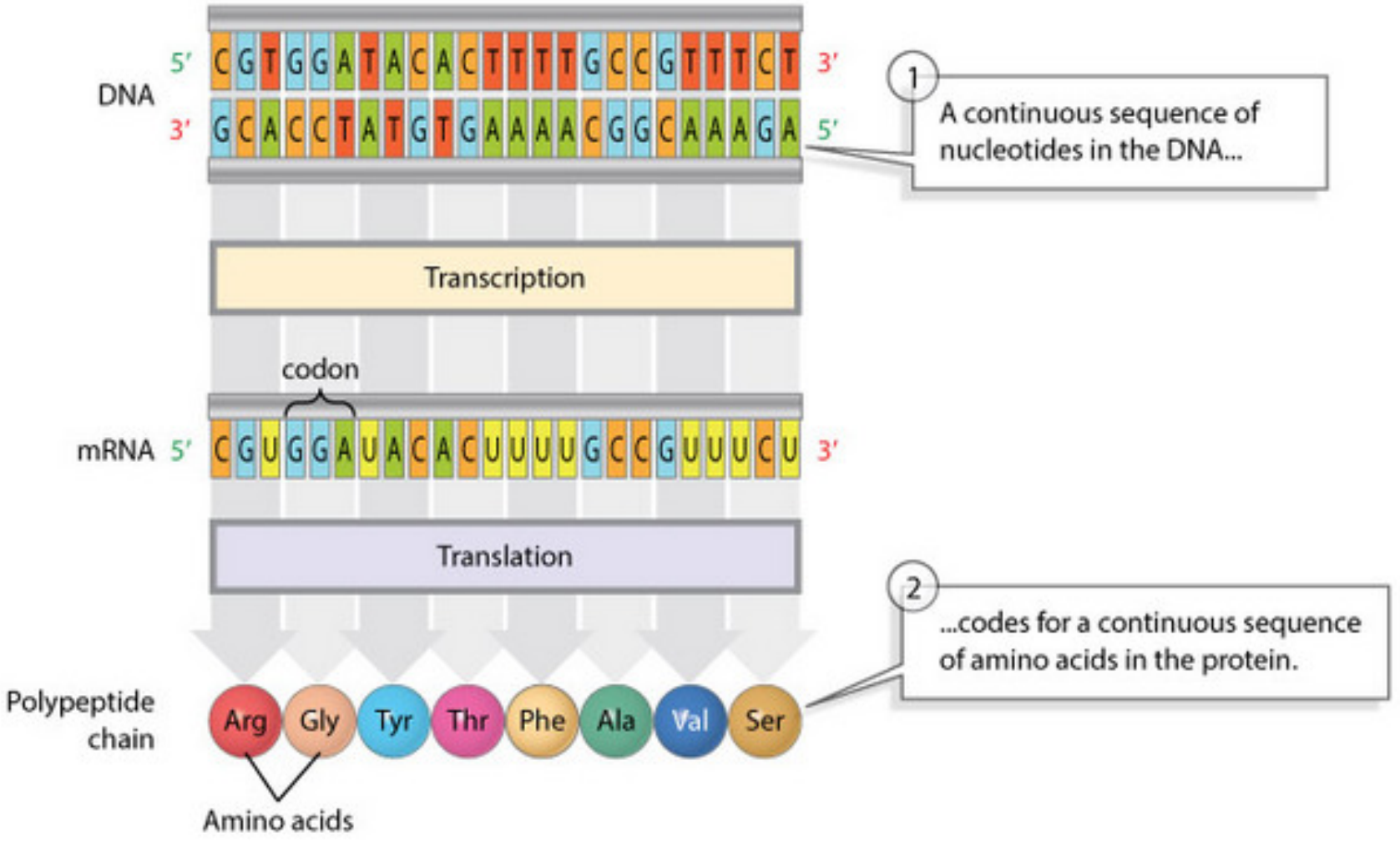


Adapted from: Wu et al. JDR 2011



Adapted from https://www.broadinstitute.org

FROM DNA TO FUNCTION



With colinearity, the number of nucleotides in the gene is proportional to the number of amino acids in the protein.

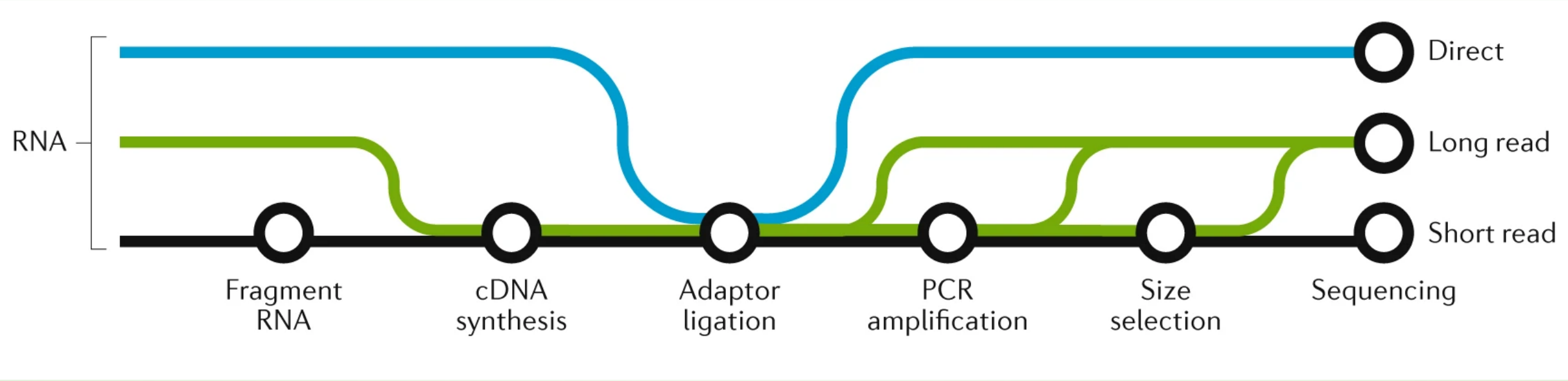
4 bases

3 letters (triplet – a codon) code for one amino acid in a protein

Codons (3 leter strings) specify 20 'words' (amino acid)

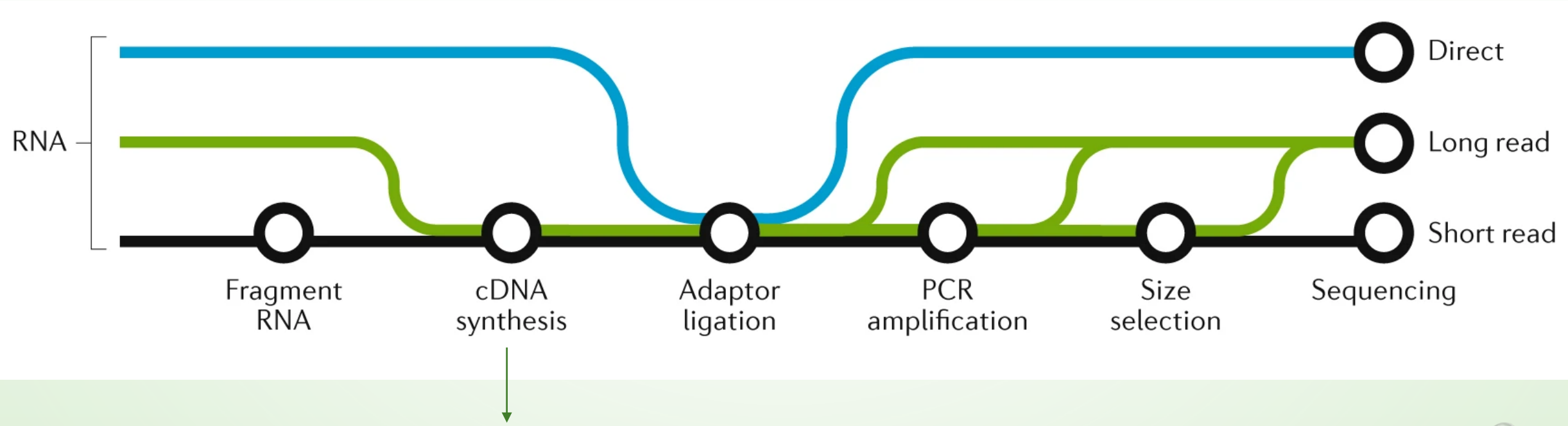
https://www.nature.com/scitable/topicpage/what-is-a-gene-colinearity-and-transcription-430/

RNA SEQUENCING



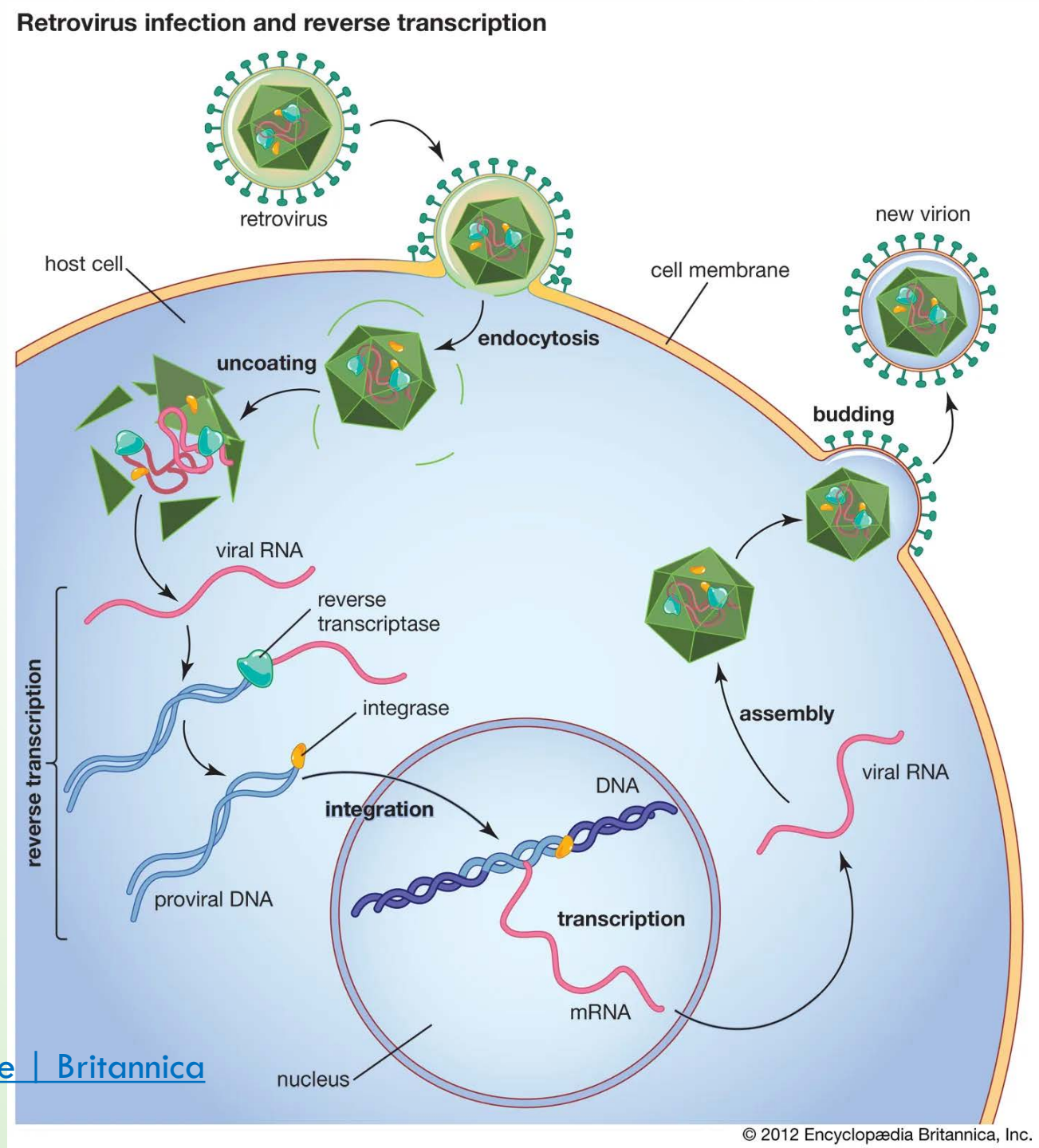
https://www.nature.com/articles/s41576-019-0150-2/figures/1

RNA SEQUENCING



Complementary DNA (cDNA) is generated via reverse transcription: mRNA is the template and DNA is the product

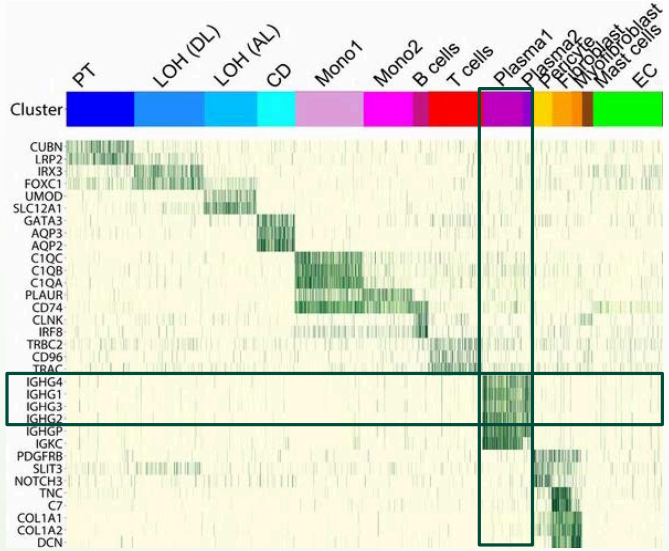
https://www.nature.com/articles/s41576-019-0150-2/figures/1



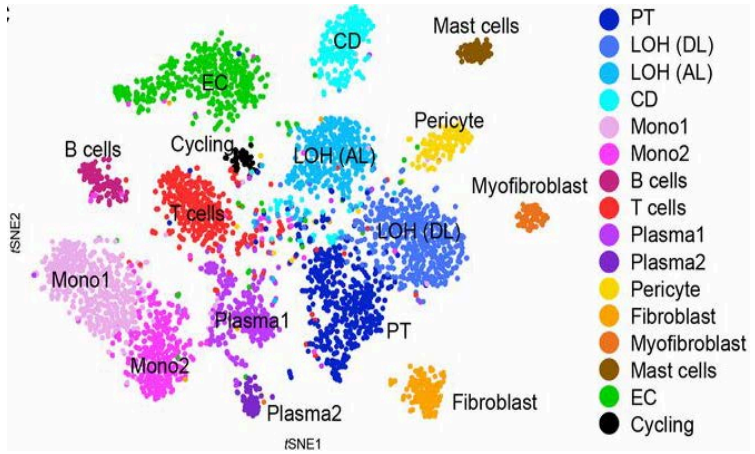
Encyclopedia Britannica
reverse transcriptase | enzyme | Britannica

We can sequence the RNA of single cells

Identifying cell-type marker genes



Discovering sample heterogeneity



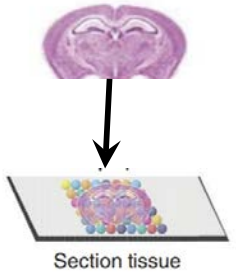
Uncovering tissue dynamics

Images from: Wu et al (2018) JASN and Farbehi et al (2019) Elife

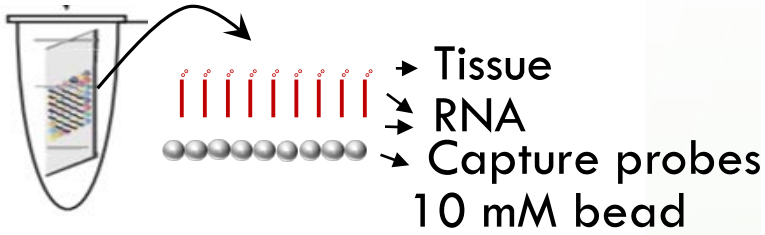
We can acquire a spatial image of single-cell expression

Spatial slide

Tissue overlay

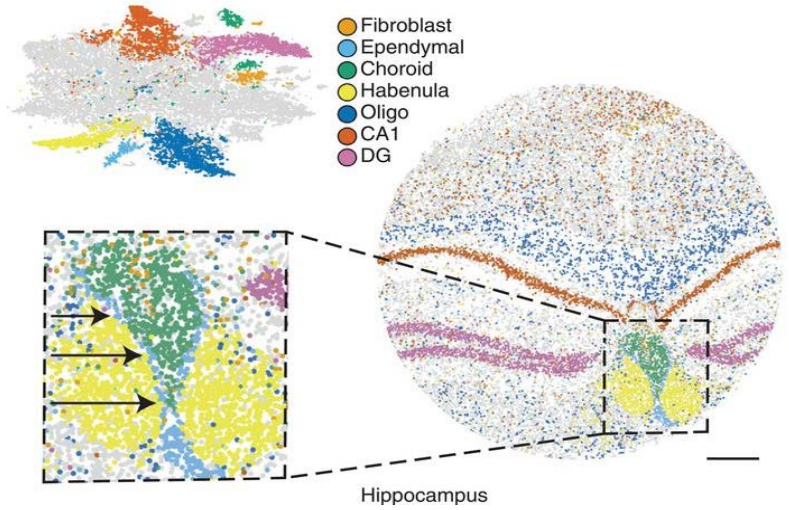


Tissue permeation

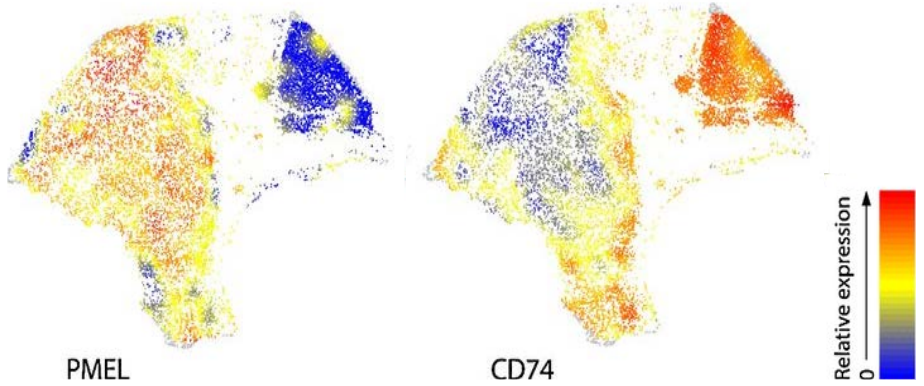


Library preparation

Spatial heatmaps of cell populations

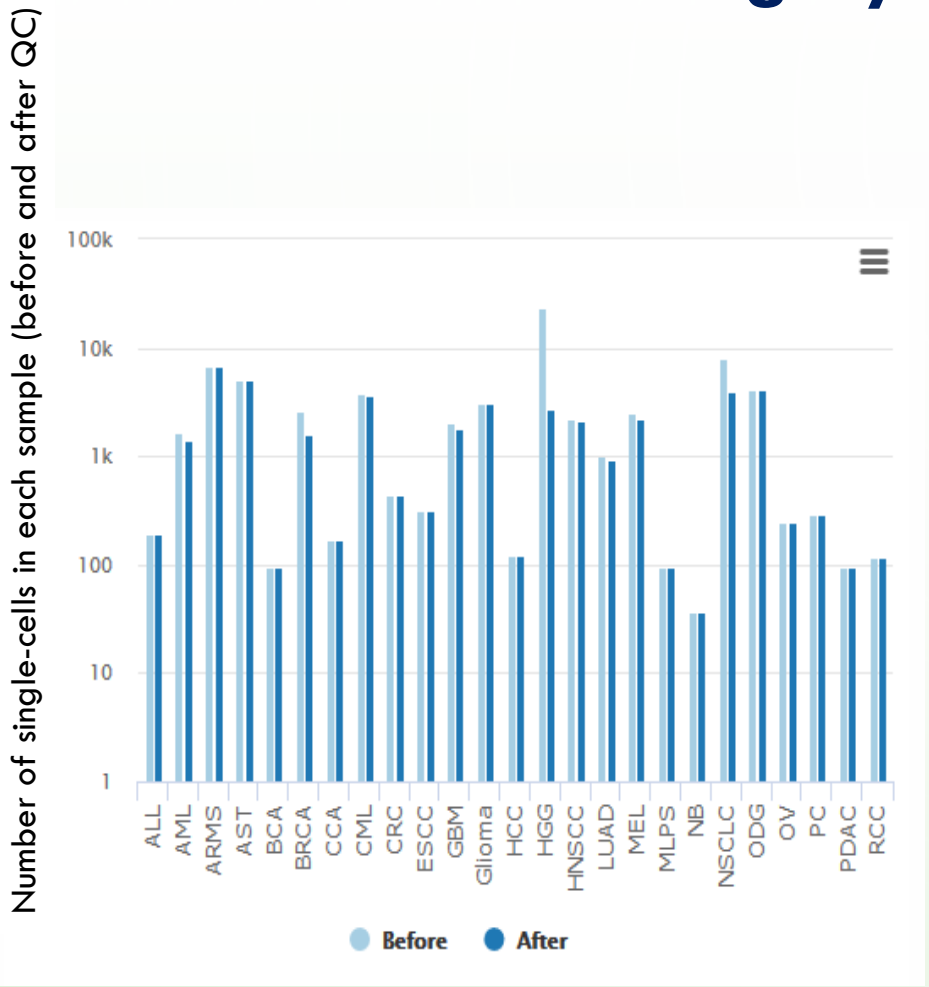


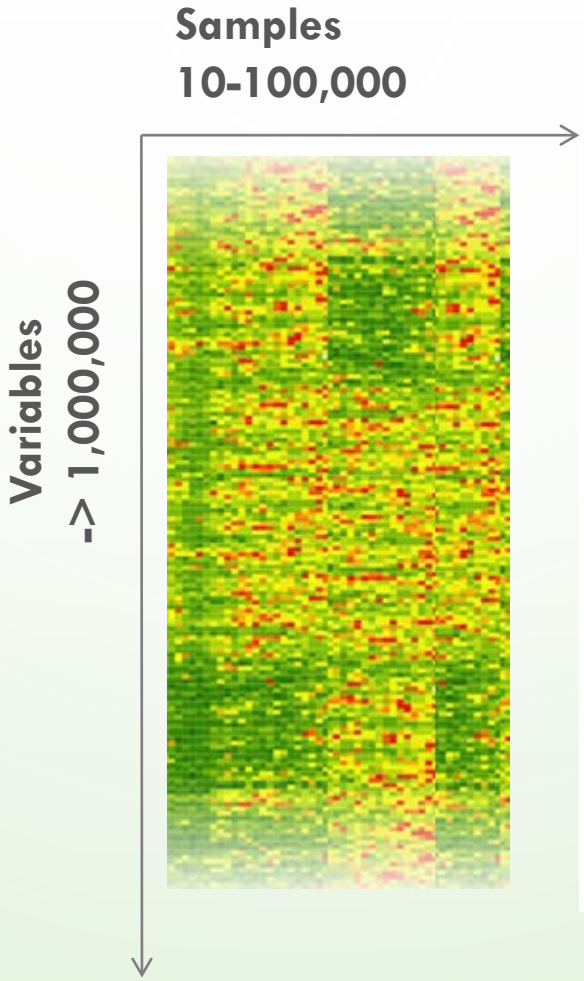
Spatial heatmaps of selected genes

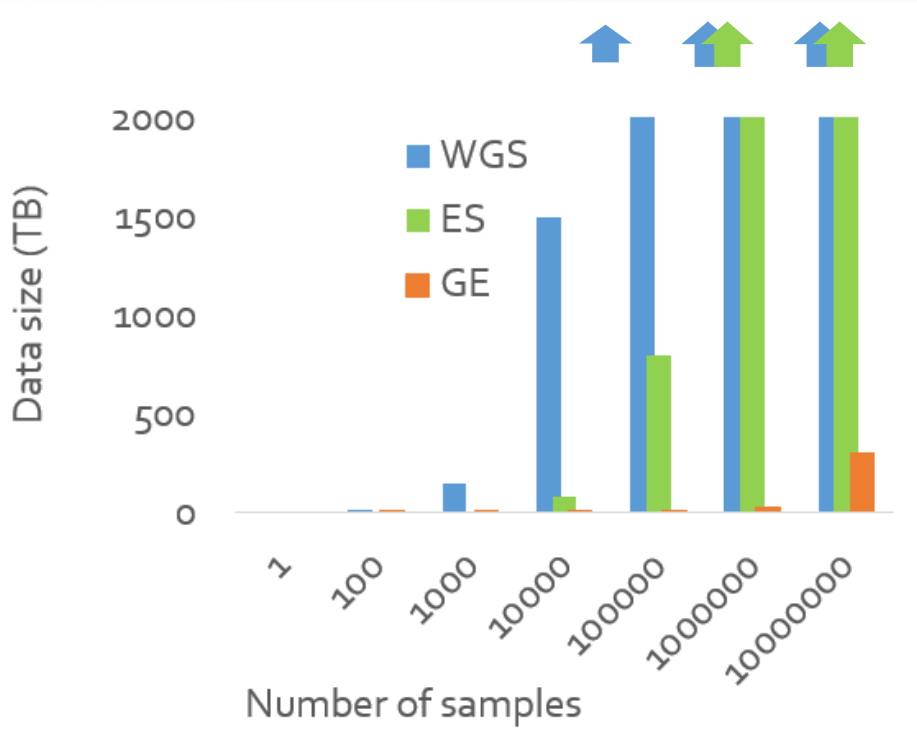


Images from: Vickovic et al (2019) Nature Methods, Thrane (2018) Cancer Res and Rodriques et al (2019) Science

Highly multidimentional datasets

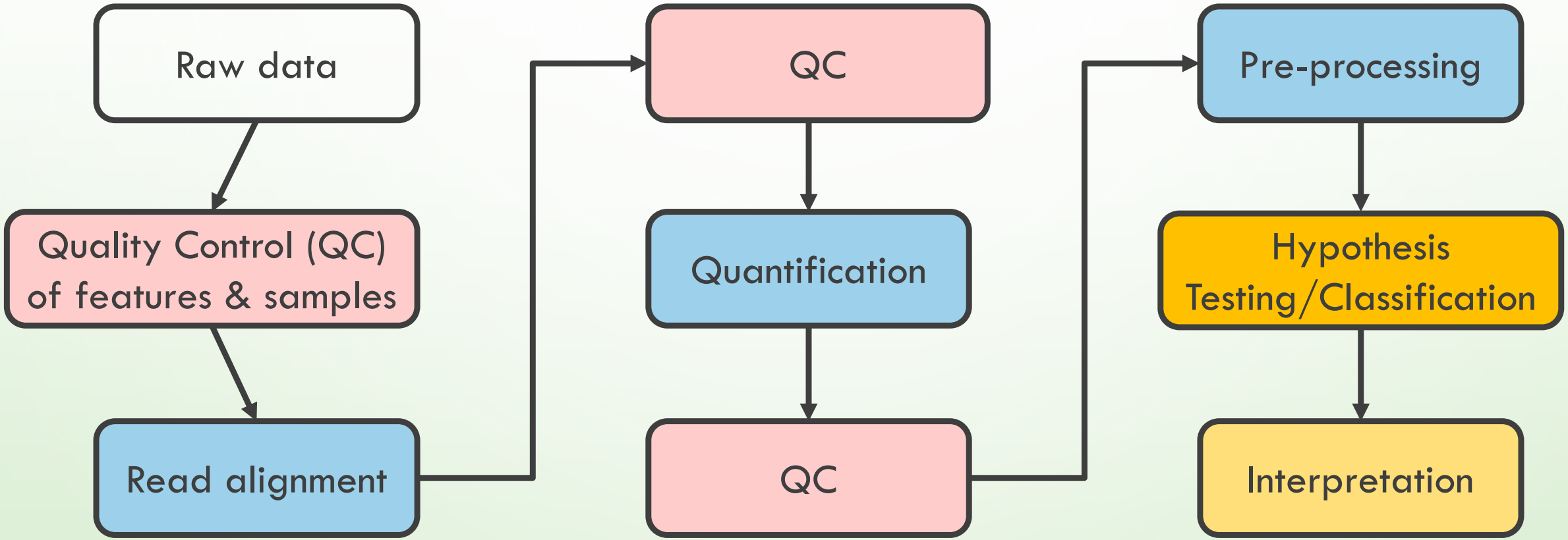




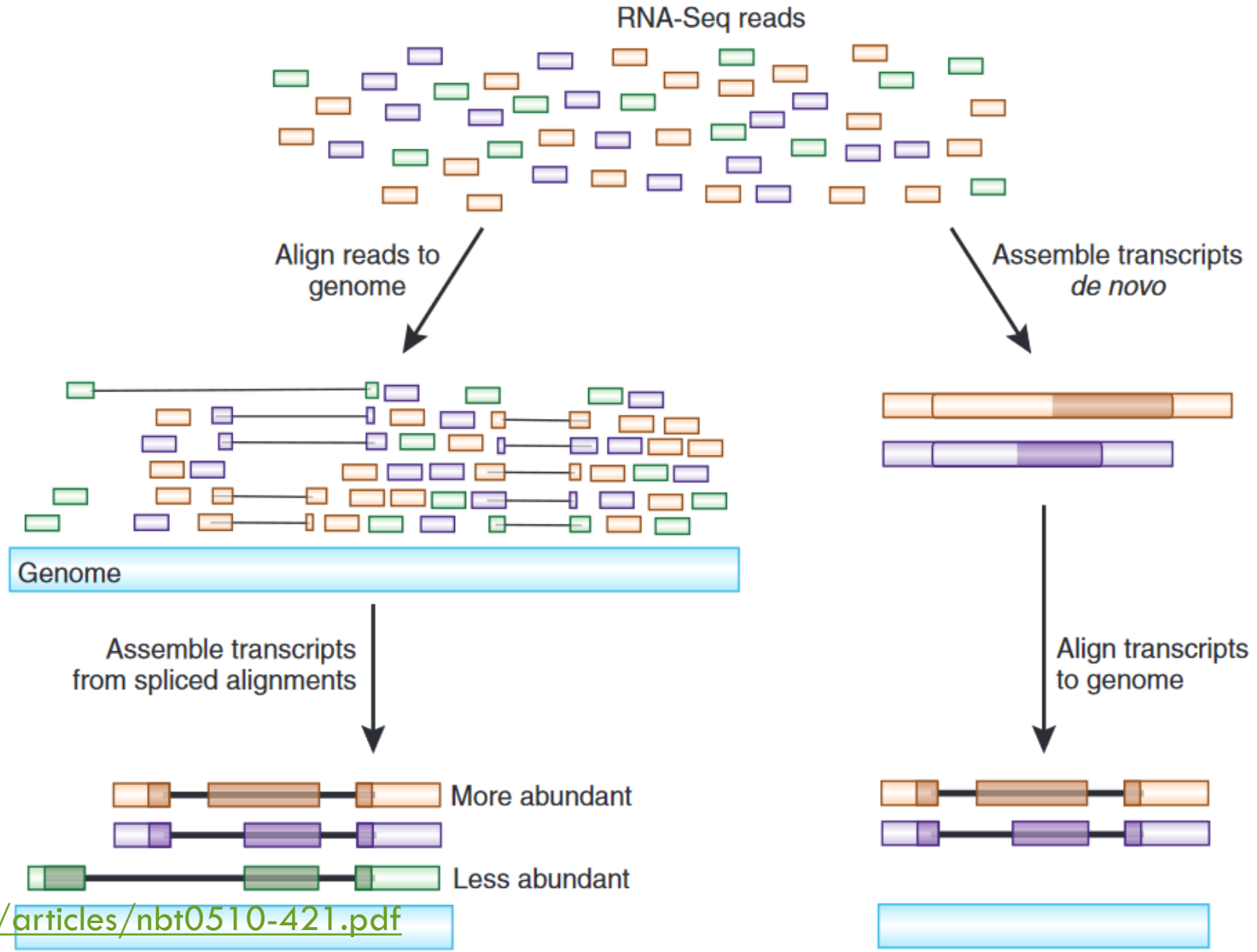


http://biocc.hrbmu.edu.cn/CancerSEA/home.jsp

RNA-SEQ: FROM LOOKING AT THE DATA TO ANALYSIS



RNA-SEQ ALIGNEMENT



Haas & Zody, 2010
https://www.nature.com/articles/nbt0510-421.pdf

NUMBERS OF GENES IN GENOMES

- Simple assumption: the whole of the genome codes for genes of interest
- If we assume that the number of amino acids in a typical protein is roughly 300 (very simplistic!)
- Then the number of bases needed to code for our typical protein is ~ 1000 (3 base pairs per amino acid)
- Genes contained in a genome estimated as genome size/1000

- For bacterial genomes this works

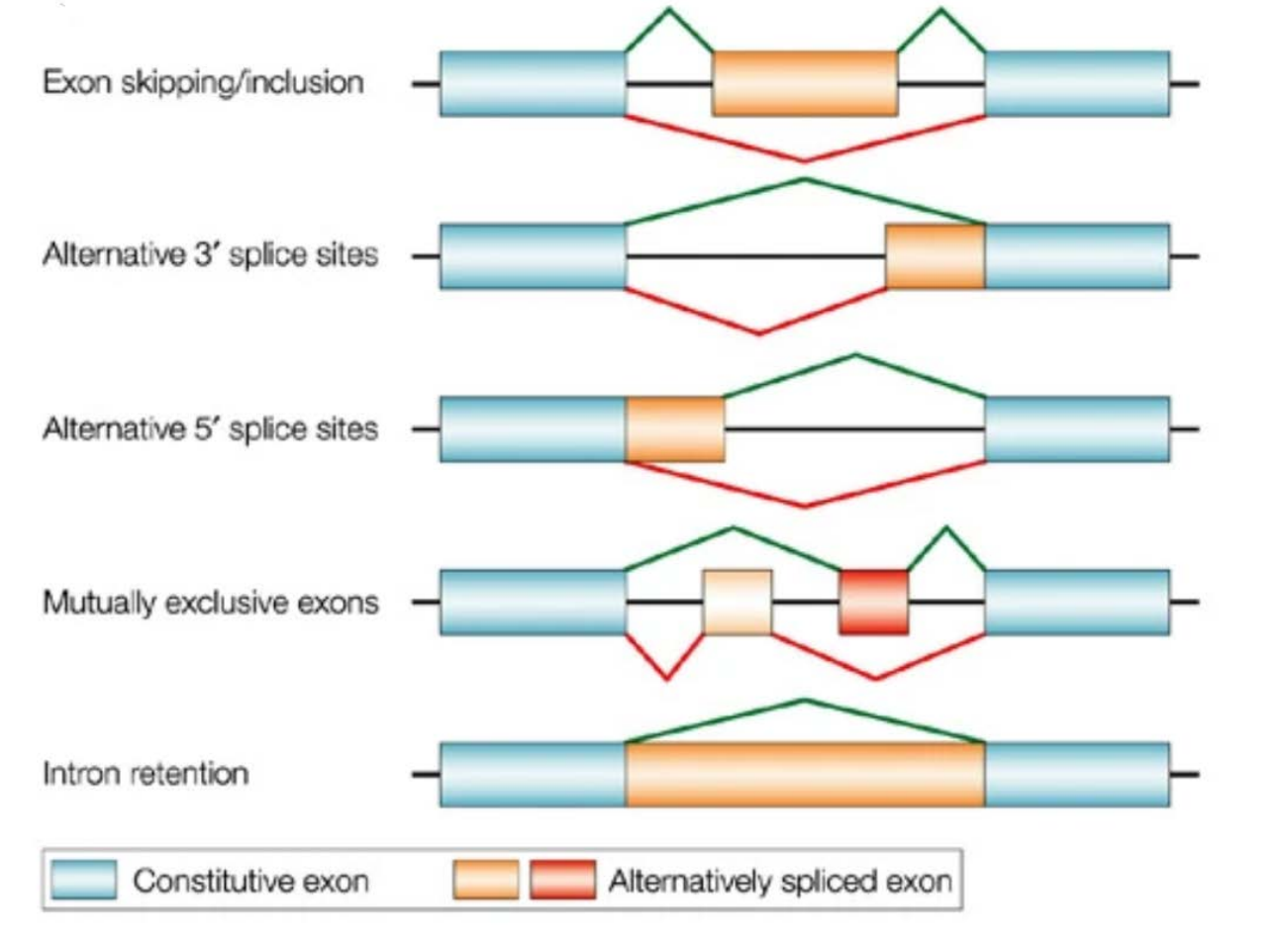
- For eukaryotic genomes this completely fails!

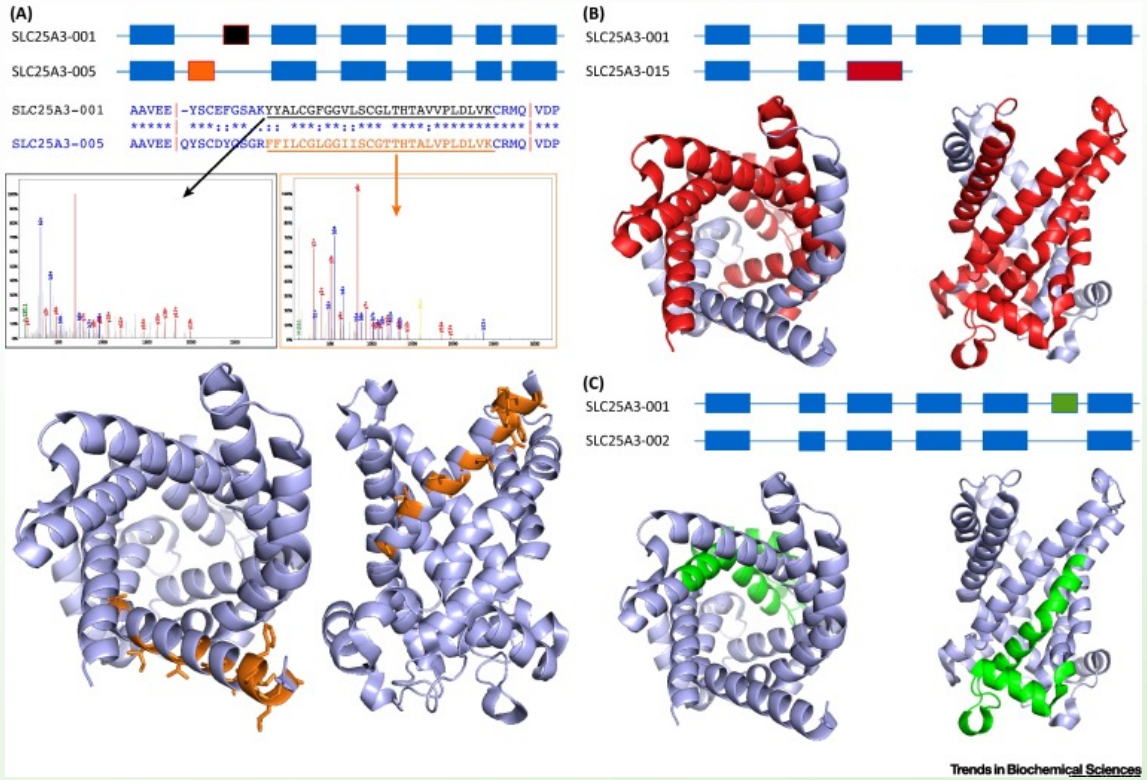
	Organism	# of protein-coding genes	# of genes naïve estimate: (genome size /1000)
viruses	HIV 1	9	10
	Influenza A virus	10-11	14
	Bacteriophage λ	66	49
	Epstein Barr virus	80	170
prokaryotes	*Buchnera sp.*	610	640
	T. maritima	1,900	1,900
	S. aureus	2,700	2,900
	V. cholerae	3,900	4,000
	B. subtilis	4,400	4,200
	E. coli	4,300	4,600
eukaryotes	*S. cerevisiae*	6,600	12,000
	C. elegans	20,000	100,000
	A. thaliana	27,000	140,000
	D. melanogaster	14,000	140,000
	F. rubripes	19,000	400,000
	Z. mays	33,000	2,300,000
	M. musculus	20,000	2,800,000
	H. sapiens	21,000	3,200,000

http://book.bionumbers.org/how-many-genes-are-in-a-genome/

HOW CAN SO FEW GENES ENCODE FOR THE COMPLEXITY OF EUKARYOTES?

From one RNA many possibilities for proteins



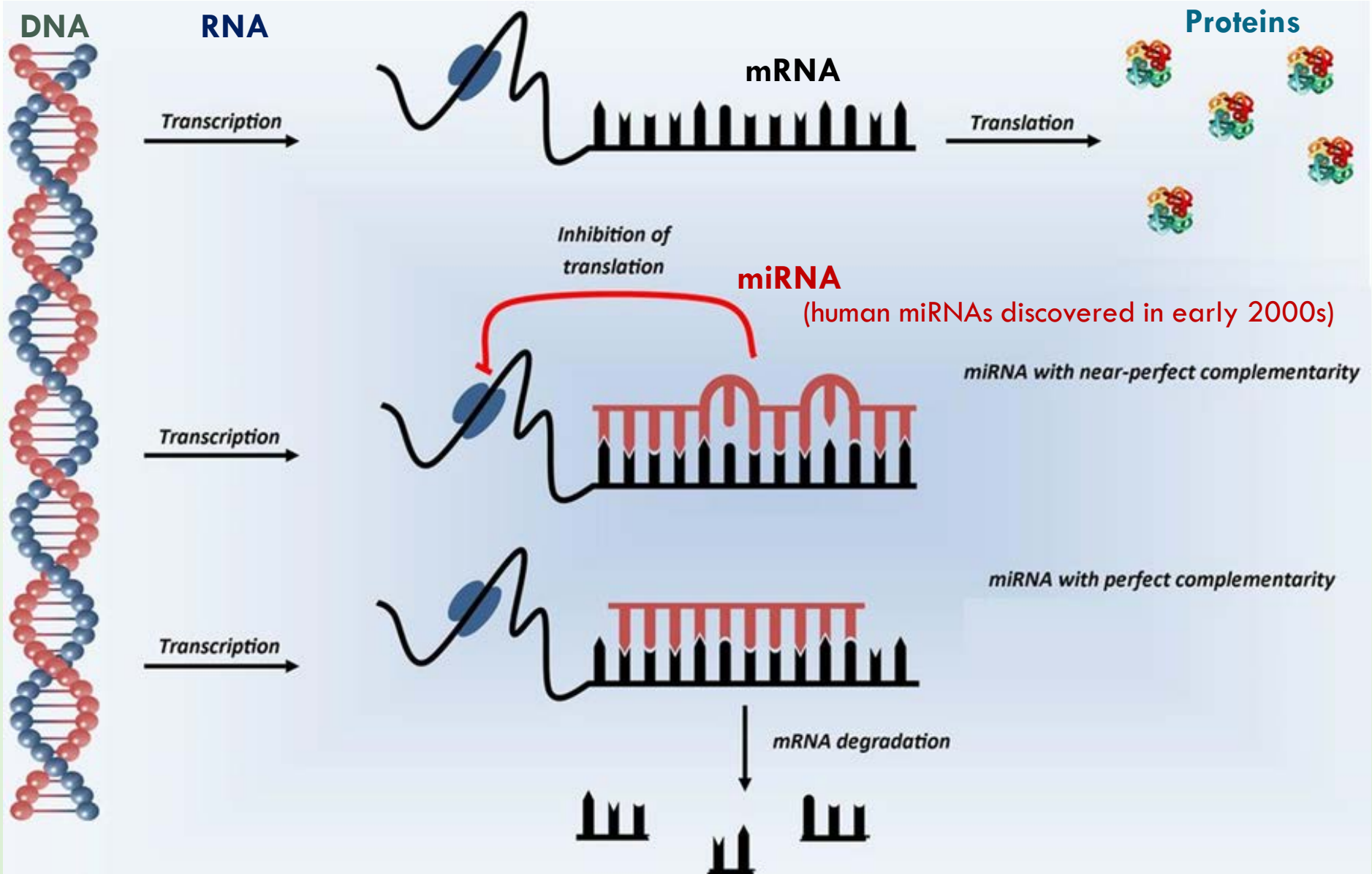


https://www.nature.com/articles/nrg775

https://www.cell.com/trends/biochemical-sciences/fulltext/S0968-0004(16)30118-9

HOW CAN SO FEW GENES ENCODE FOR THE COMPLEXITY OF EUKARYOTES?

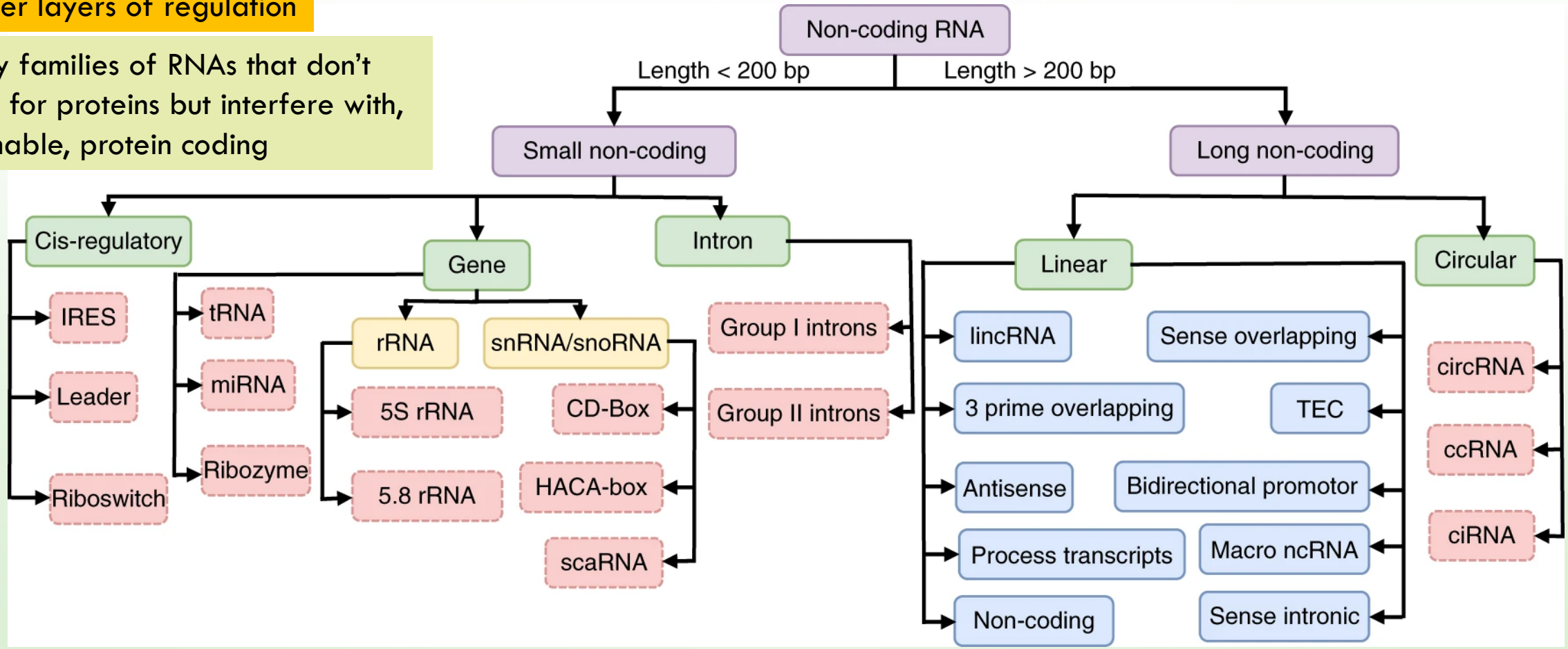
Further layers of regulation



HOW CAN SO FEW GENES ENCODE FOR THE COMPLEXITY OF EUKARYOTES?

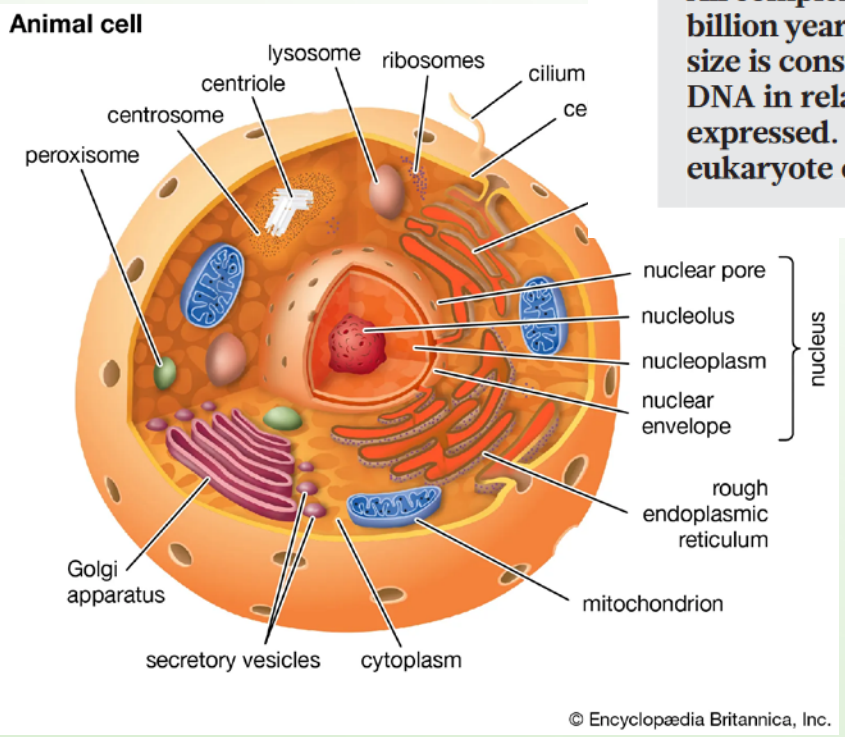
Further layers of regulation

Many families of RNAs that don't code for proteins but interfere with, or enable, protein coding



https://www.nature.com/articles/s42256-019-0051-2

HOW CAN SO FEW GENES ENCODE FOR THE COMPLEXITY OF EUKARYOTES?



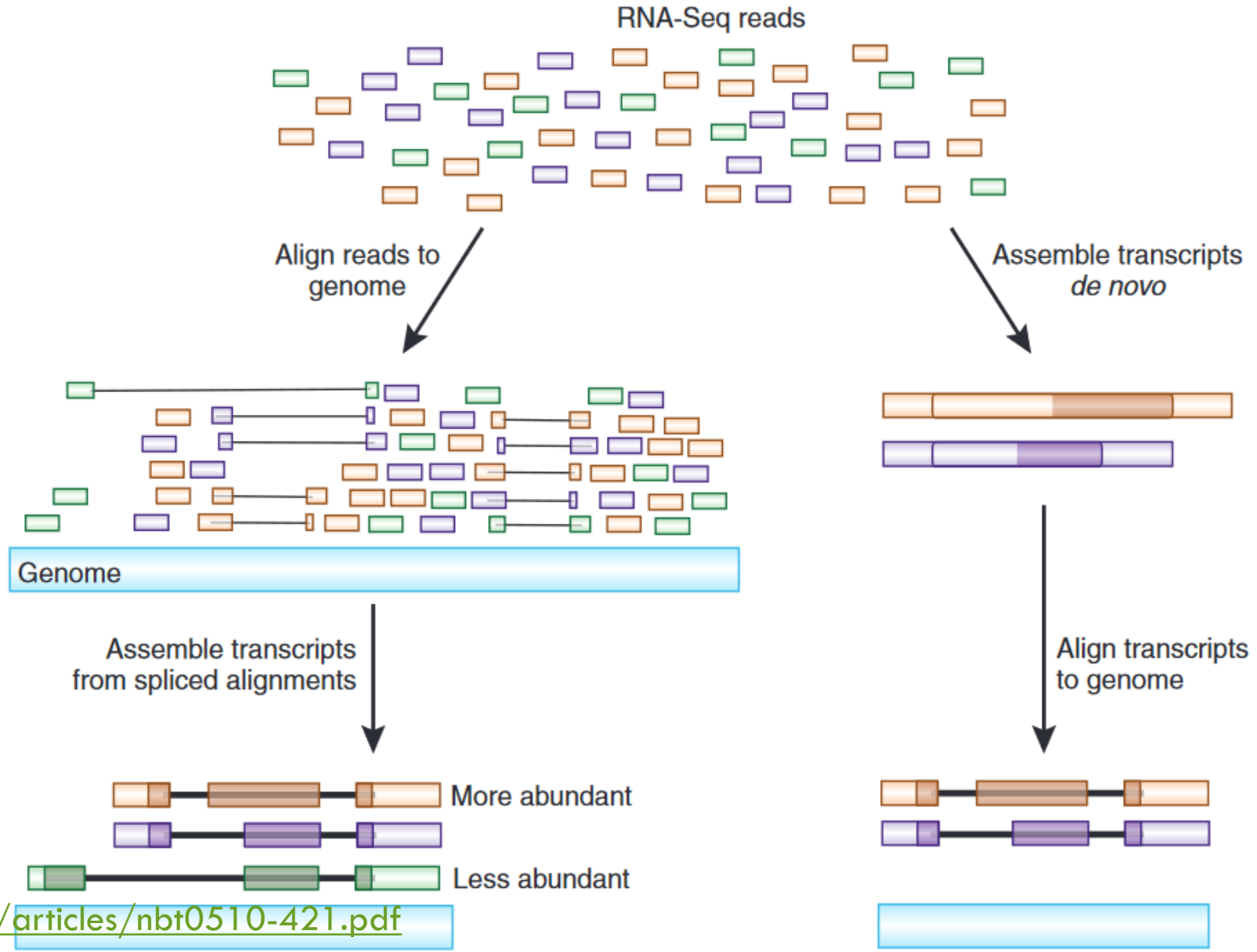
The energetics of genome complexity

Nick Lane[1] & William Martin[2]

All complex life is composed of eukaryotic (nucleated) cells. The eukaryotic cell arose from prokaryotes just once in four billion years, and otherwise prokaryotes show no tendency to evolve greater complexity. Why not? Prokaryotic genome size is constrained by bioenergetics. The endosymbiosis that gave rise to mitochondria restructured the distribution of DNA in relation to bioenergetic membranes, permitting a remarkable 200,000-fold expansion in the number of genes expressed. This vast leap in genomic capacity was strictly dependent on mitochondrial power, and prerequisite to eukaryote complexity: the key innovation en route to multicellular life.

https://www.nature.com/articles/nature09486

RNA-SEQ ALIGNEMENT

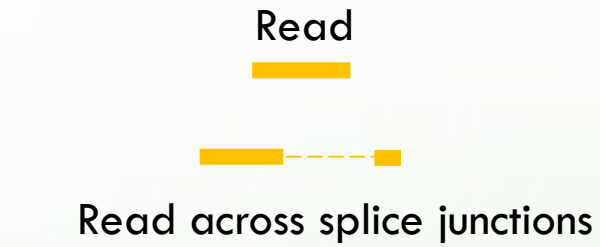


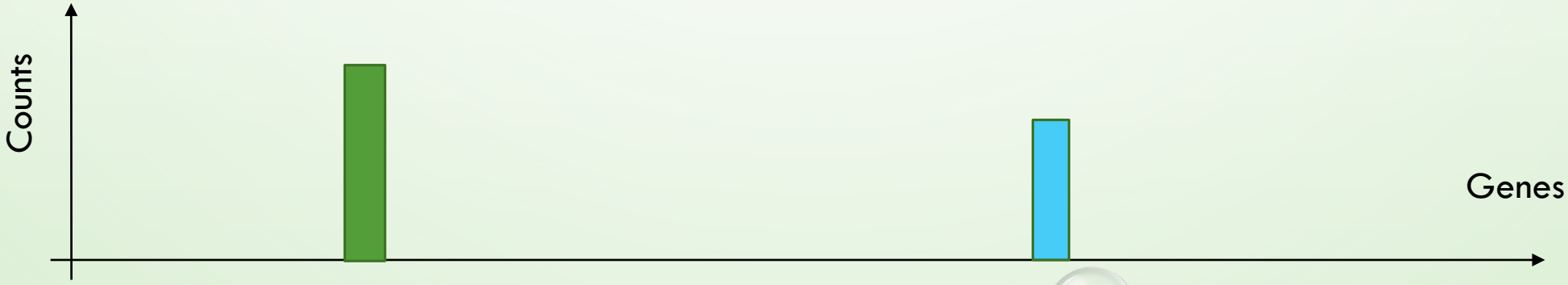
Haas & Zody, 2010
https://www.nature.com/articles/nbt0510-421.pdf

QUANTIFYING GENE EXPRESSION

Sequencing reads can be counted on any feature (e.g. exons, introns, genes)

- Gene expression
- Expression of different isoforms (challenging)

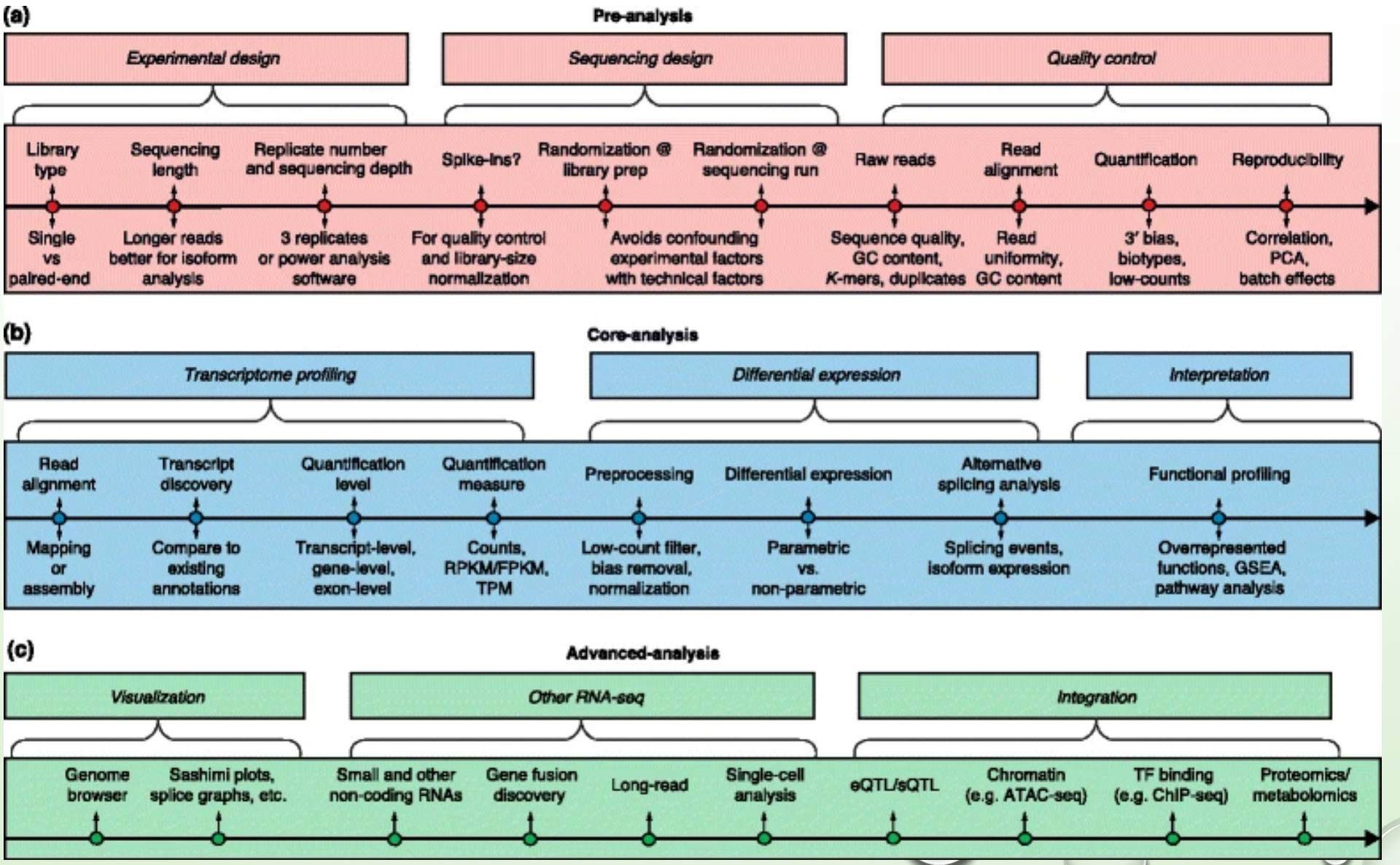




COUNTS/GENE EXPRESSION MATRIX

	Sample1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7
Gene A	345	0	23	56	76	4	3
Gene B	60	45	56	32	24	58	54
Gene C	0	0	0	0	0	0	0
Gene D	453	569	764	897	564	432	865

RNA-SEQ DATA ANALYSIS



https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8

UNSUPERVISED LEARNING APPLICATIONS

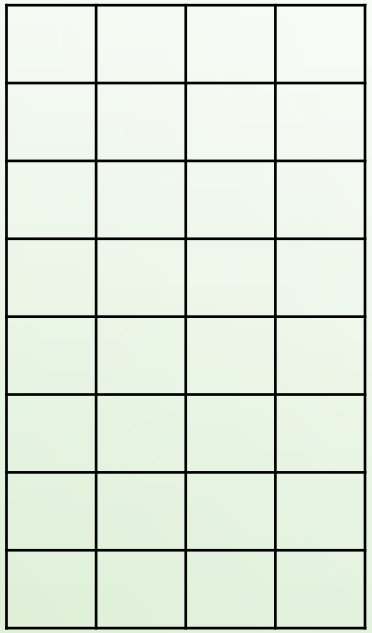
- EXPLORATORY DATA ANALYSIS
- QUALITY CONTROL
- CLASS DISCOVERY
- DIMENSIONALITY REDUCTION

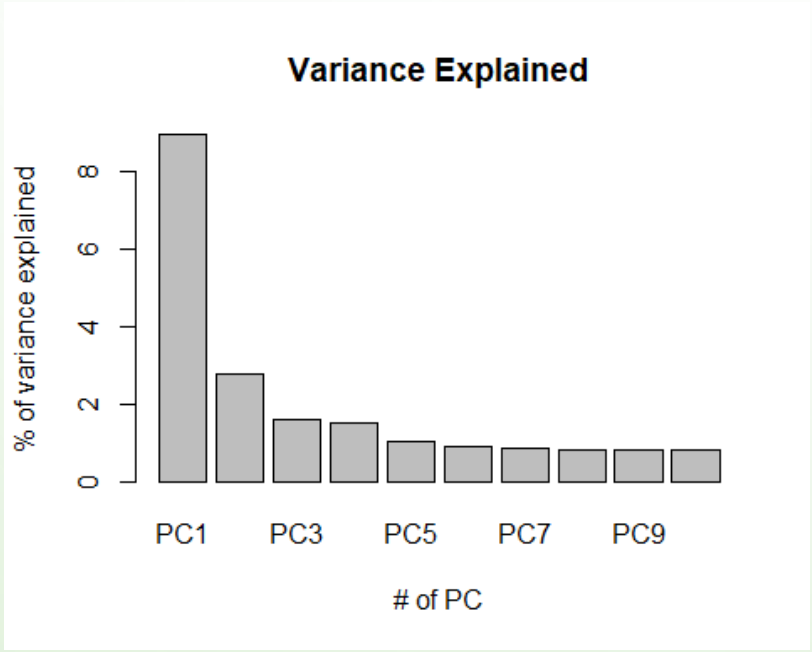
EXAMPLE: QC IN SINGLE CELL SEQUENCING EXPERIMENT

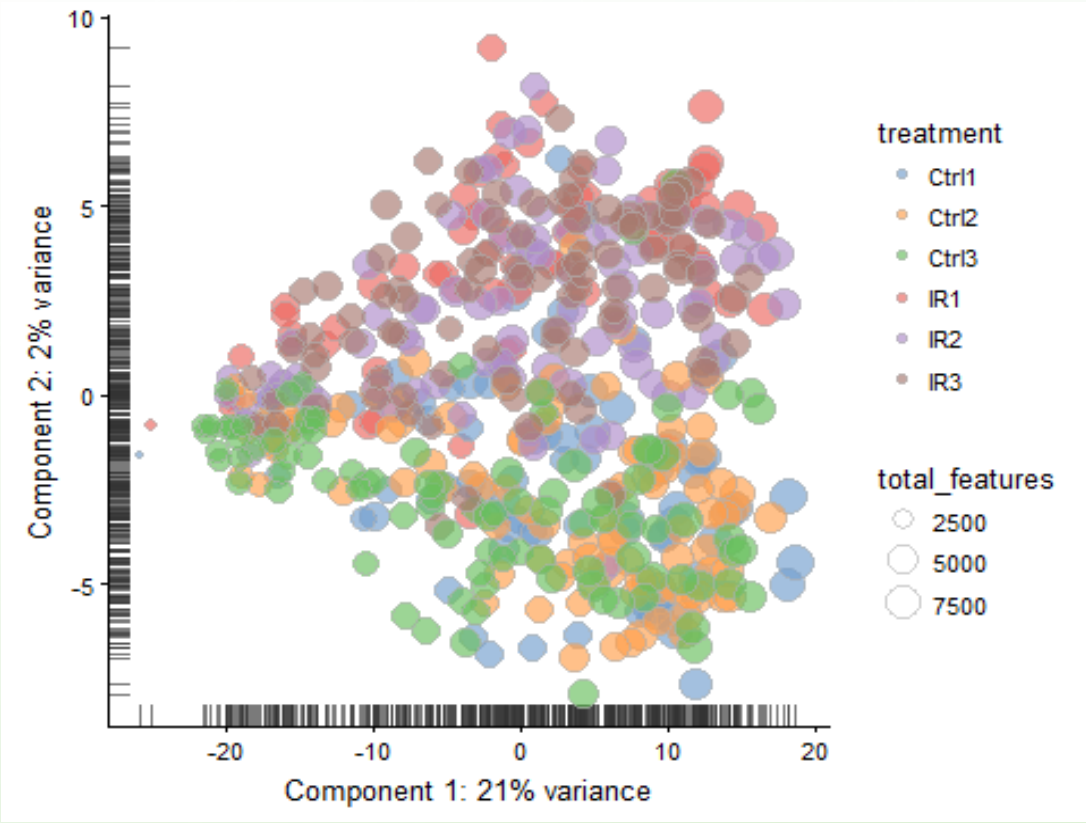
#Cells 528

 # Samples (Cntrl vs Treatment)

#Features/Genes 2637





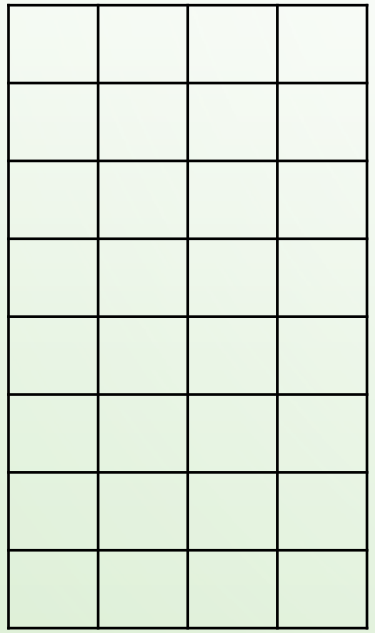


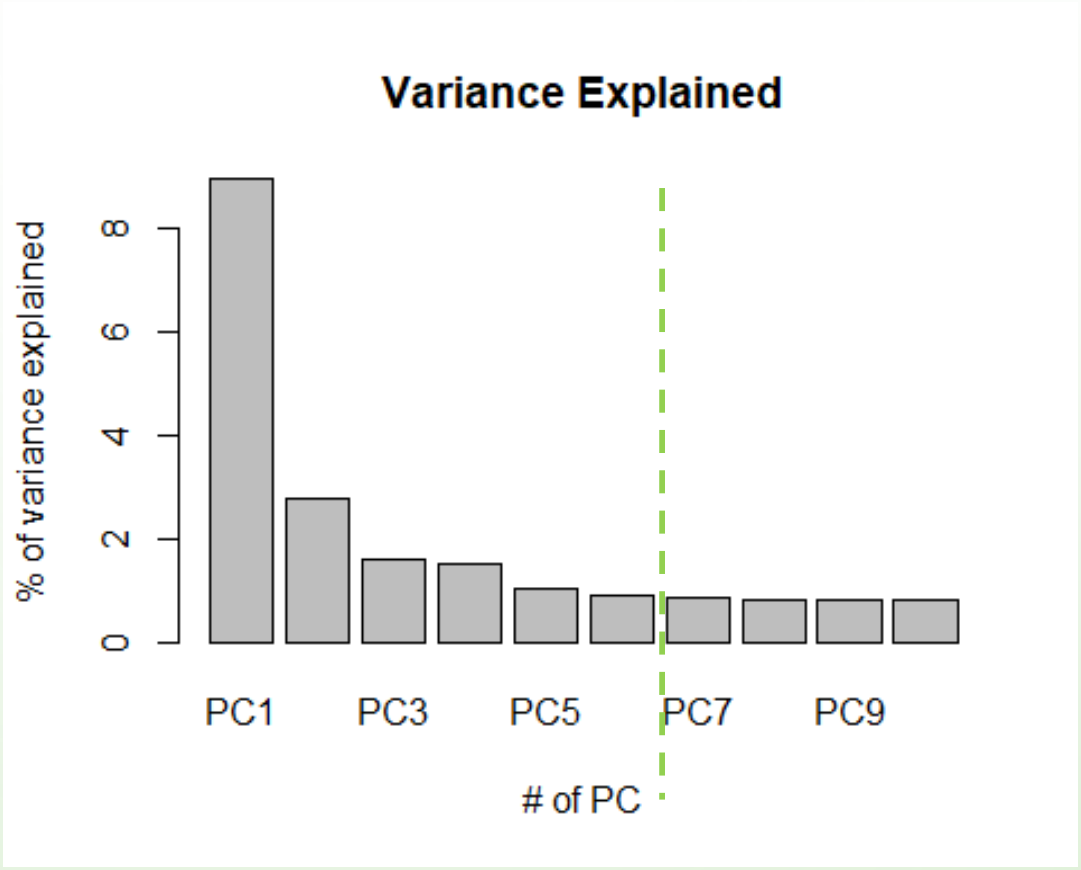
PCA DIMENSIONALITY REDUCTION

#Cells 528

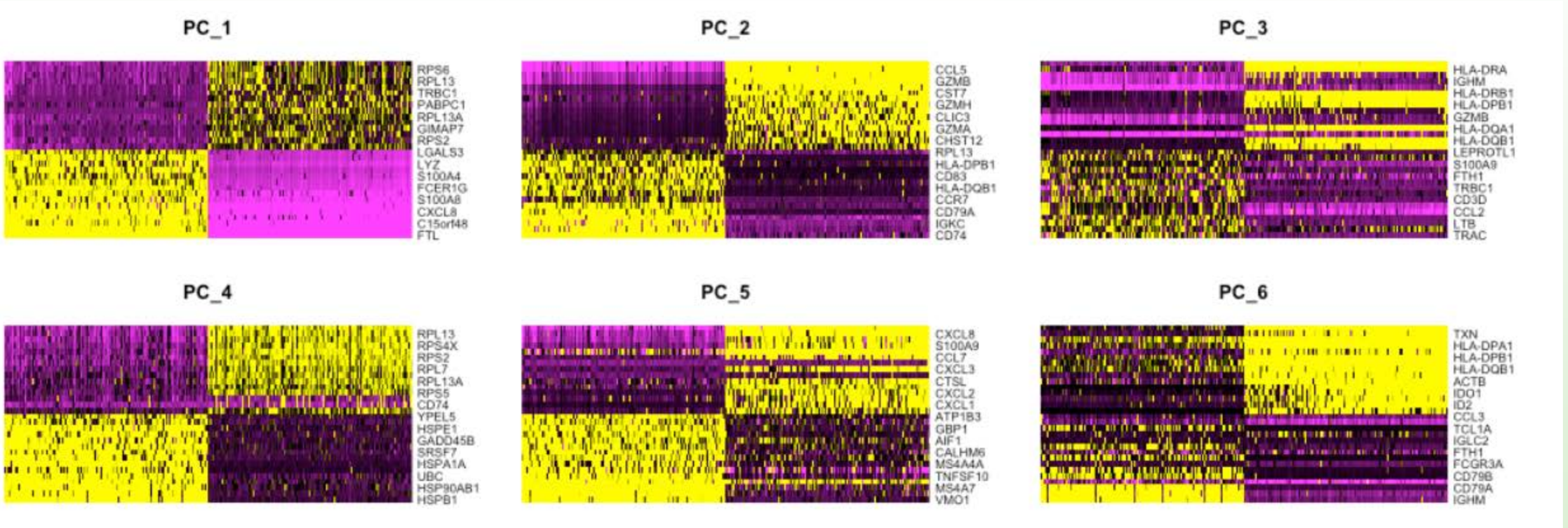
 # Samples (Cntrl vs Treatment)

#Features/Genes 2637

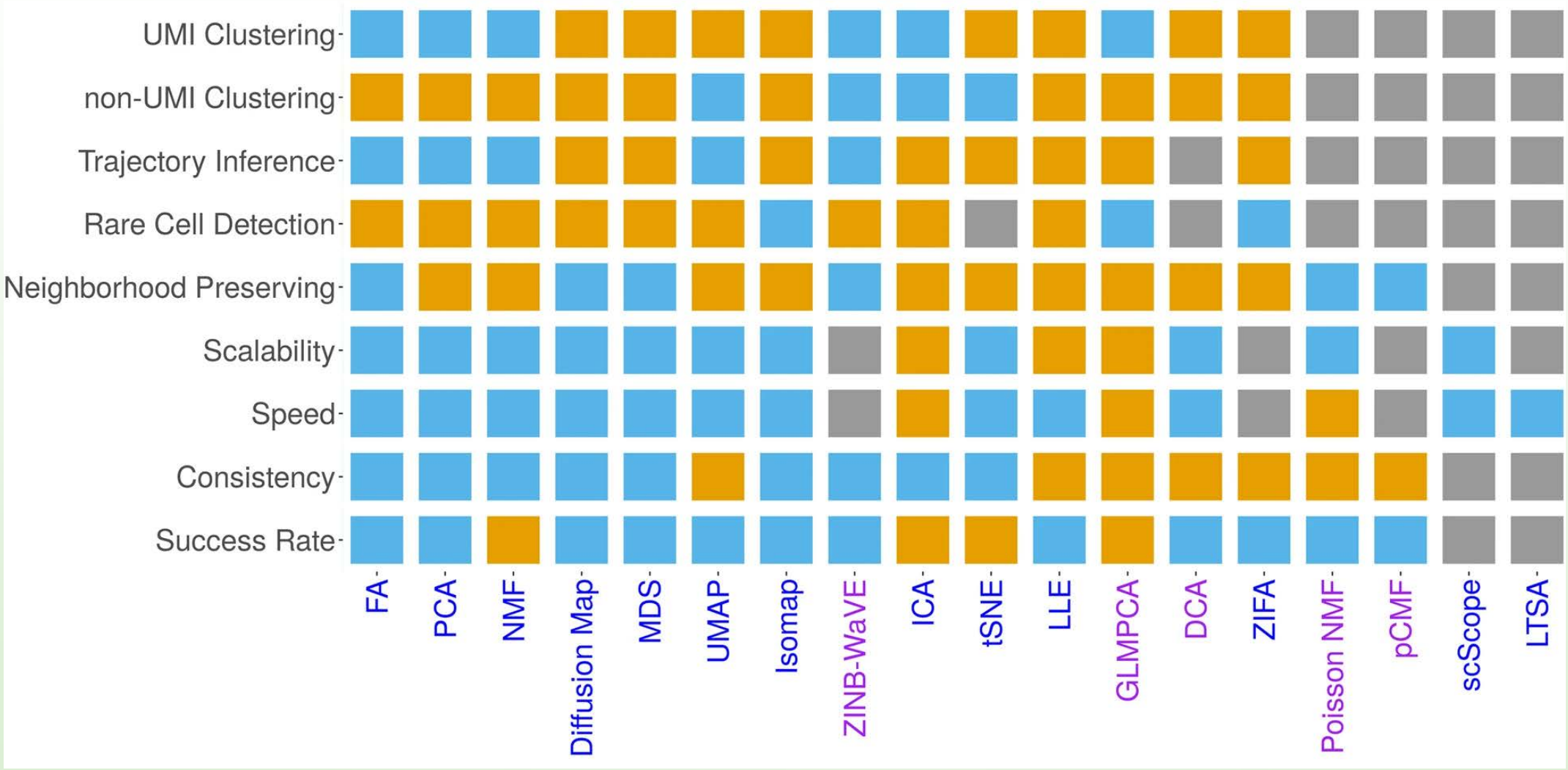




PCA DIMENSIONALITY REDUCTION



DIMENSIONALITY REDUCTION



https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1898-6

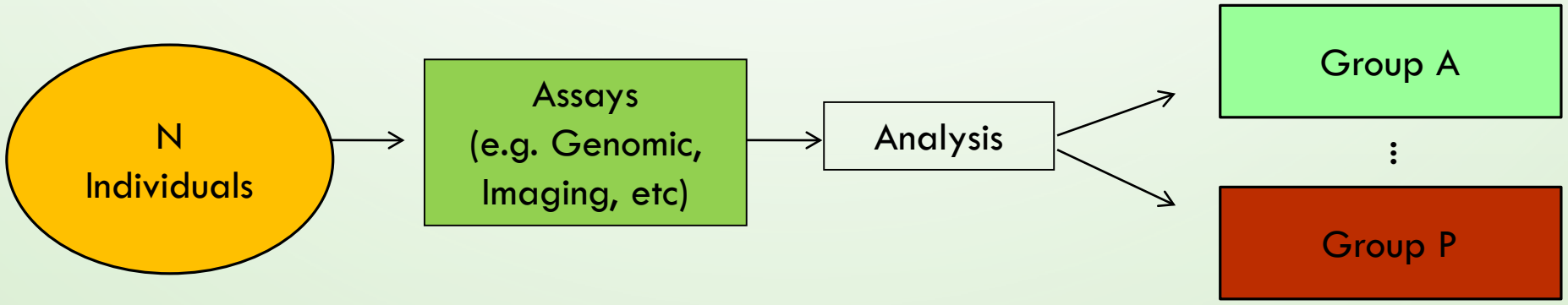
CLUSTERING APPLICATIONS

ARE THERE GROUPS OF SIMILAR DISEASES/CELLS?

"UNSUPERVISED APPROACH"

CLASS DISCOVERY

FIND GROUPS OF SIMILAR CASES OR SIMILAR FEATURES



Clustering: guilty-by-association

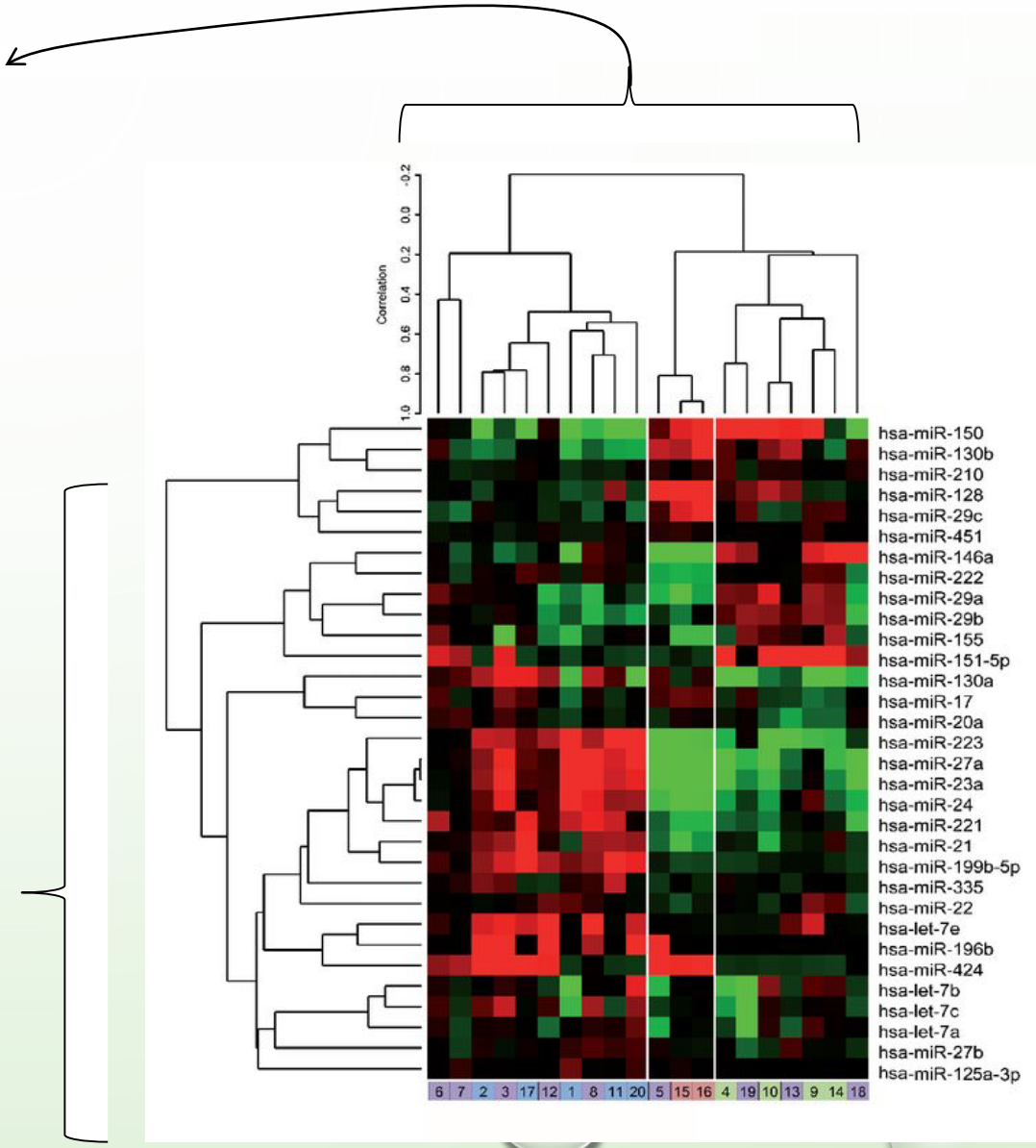
Are my samples similar?

Samples with similar genomic profile might for example have a similar prognosis or response to treatment.

Or in single cell might come from the same cell population.

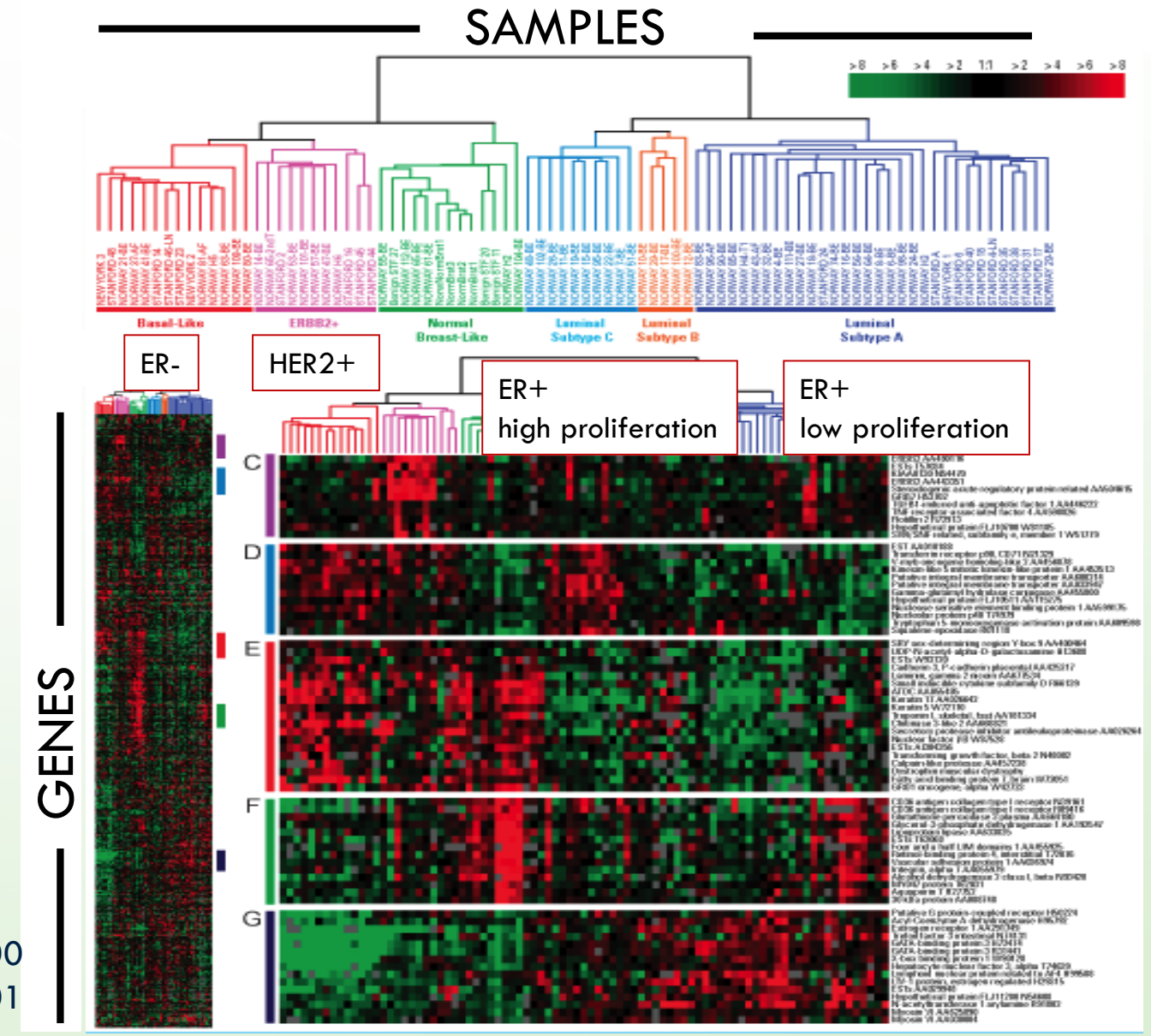
Are my genes similar?

Suppose genes A and B are grouped in the same cluster. This mean they are expressed under the same conditions. Then we can hypothesize that genes A and B are involved in similar pathways/share function.



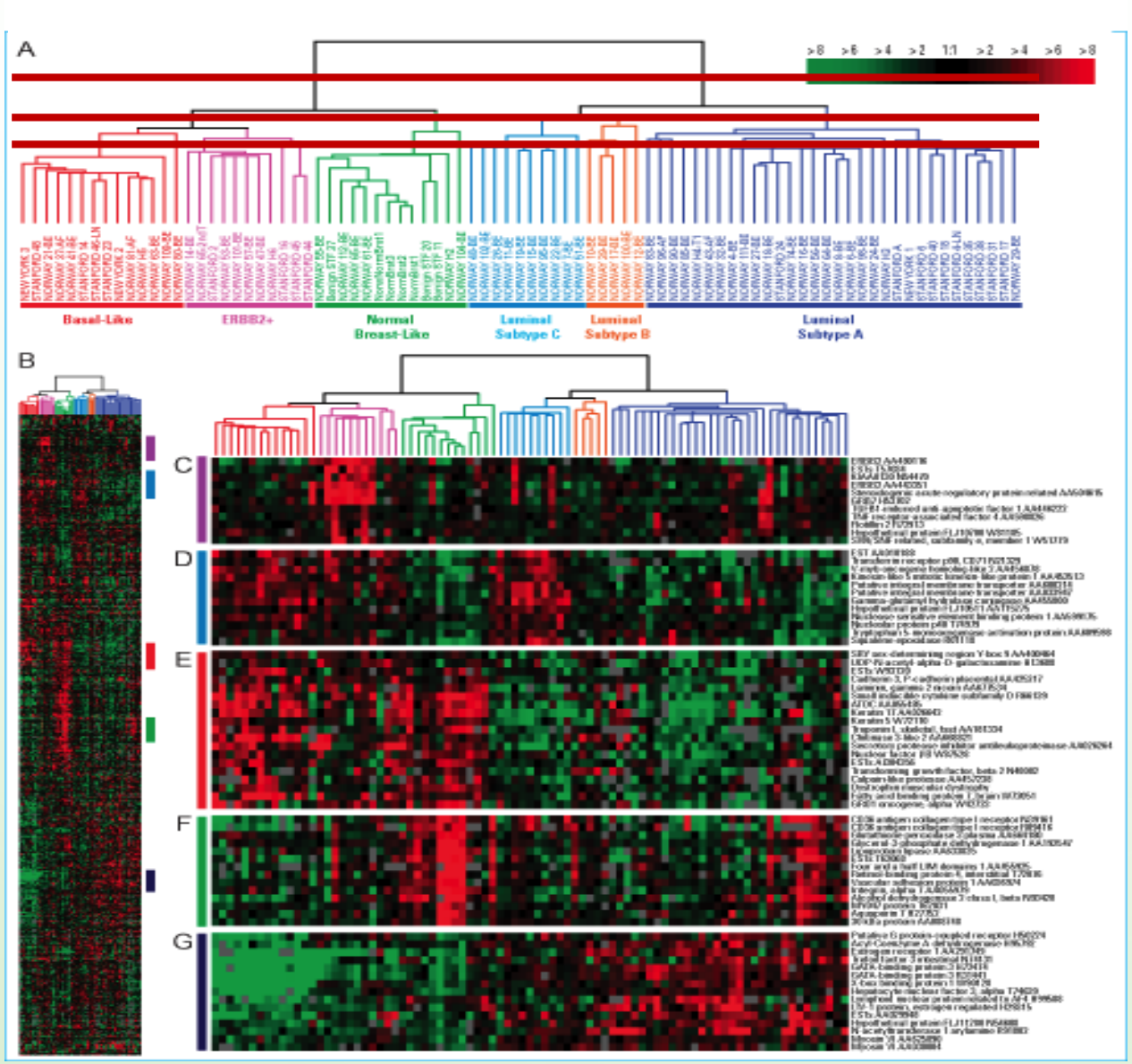
Unsupervised: Breast Cancer Subtypes (PAM50)

- Unsupervised approach: outcome not considered
- Classification based on Expression of Breast Cancer Intrinsic Genes
- Gene expression microarrays
- Hierarchical clustering used to represent distance between samples
- Groups identified matching existing clinical knowledge



Perou et al, Nature 2000
Sorlie et al, PNAS 2001

How many clusters?



Perou et al, Nature 2000
Sorlie et al, PNAS 2001

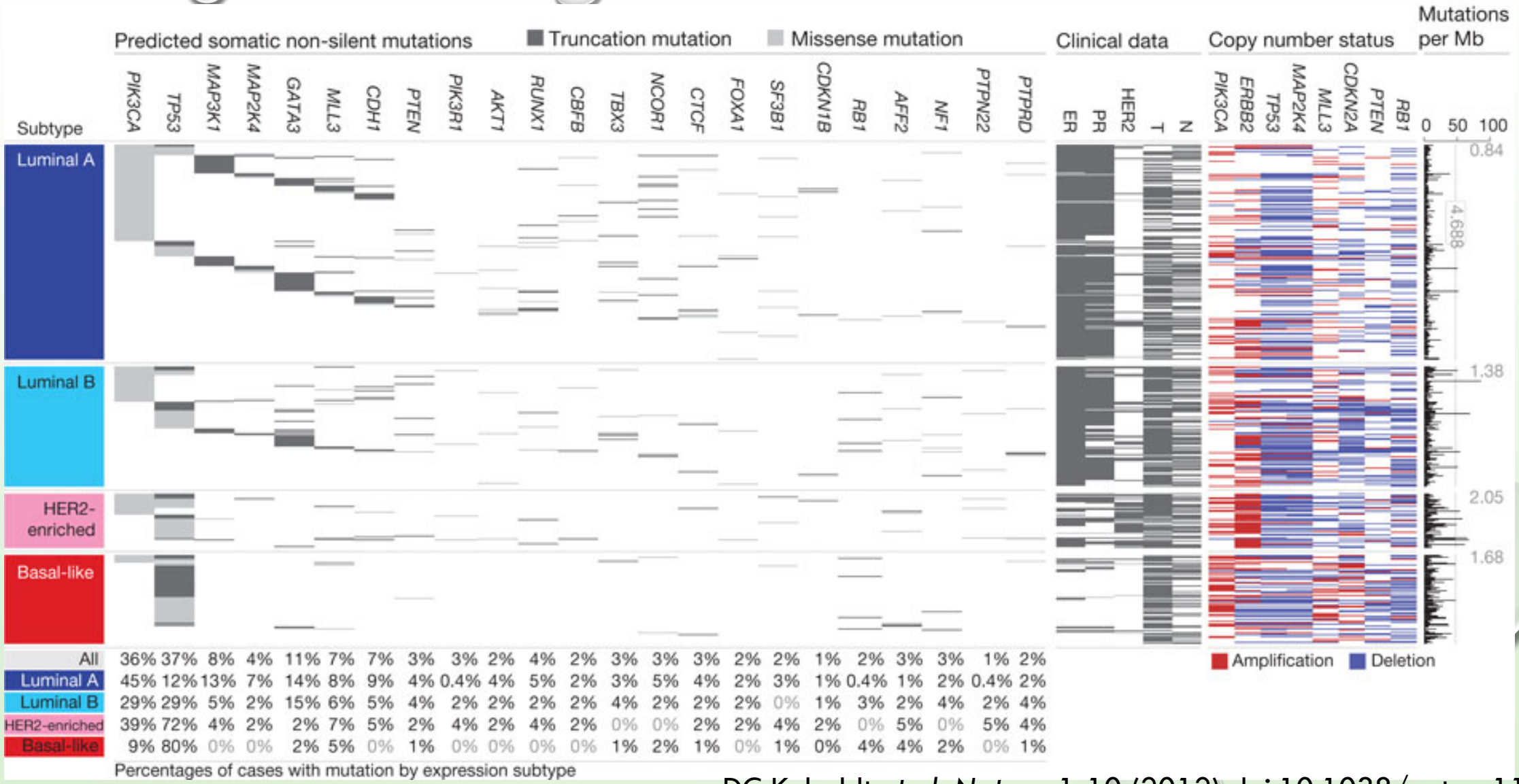
REPRODUCIBILITY

WEIGELT, B ET AL, LANCET ONCOLOGY, 2010

WHEN USING DIFFERENT METHODS TO ASSIGN PATIENTS TO EACH CLUSTER:

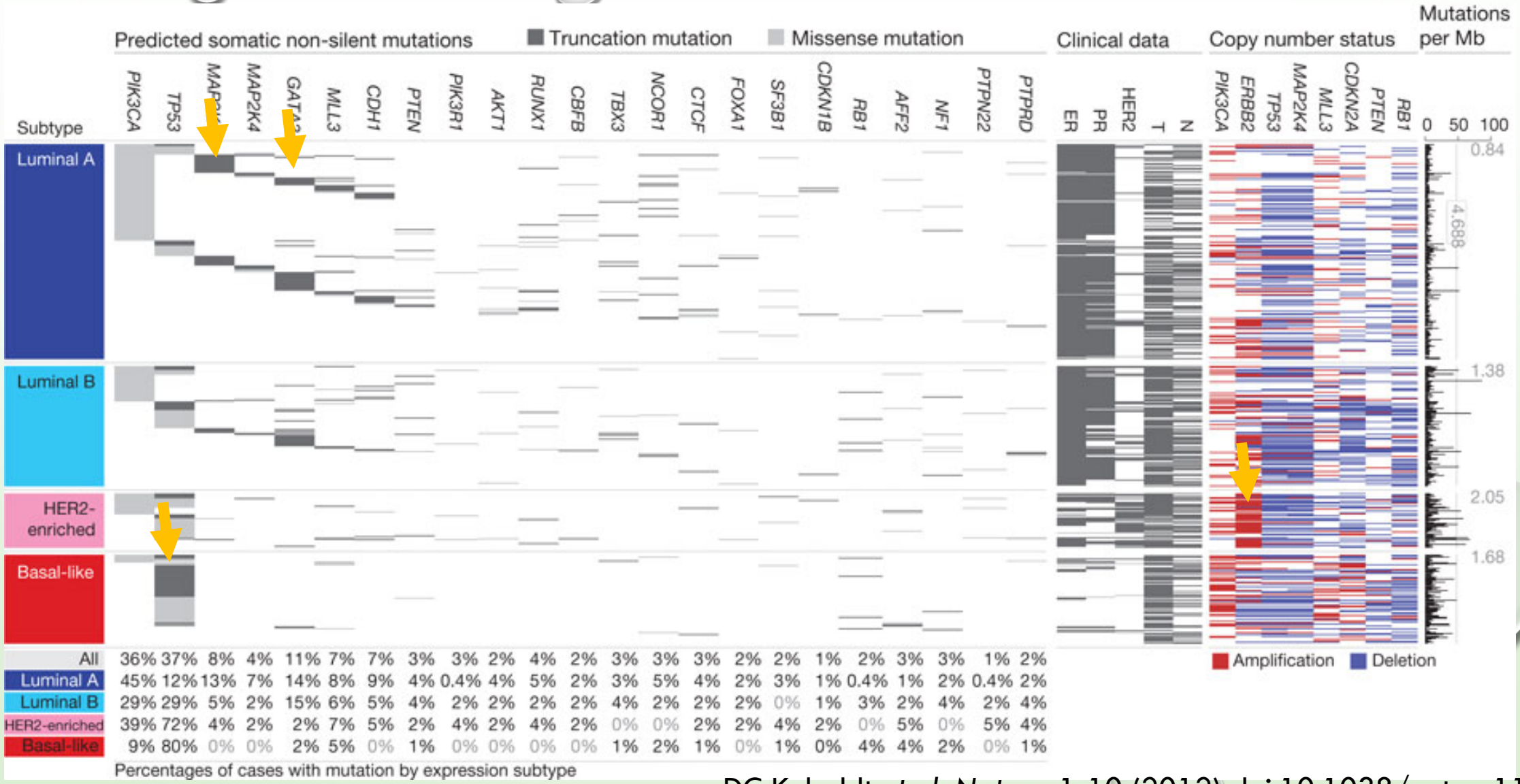
- BASAL-LIKE CANCERS WERE CONSISTENTLY CLASSIFIED.
- ASSIGNMENT OF INDIVIDUAL CASES TO LUMINAL A, LUMINAL B, HER2, AND NORMAL BREAST-LIKE SUBTYPES WAS DEPENDENT ON THE METHOD USED.
- THE SIGNIFICANCE OF ASSOCIATIONS WITH OUTCOME OF EACH MOLECULAR SUBTYPE, OTHER THAN BASAL-LIKE AND LUMINAL A, VARIED DEPENDING ON THE METHOD USED.

Link between gene expression clusters and genomic features



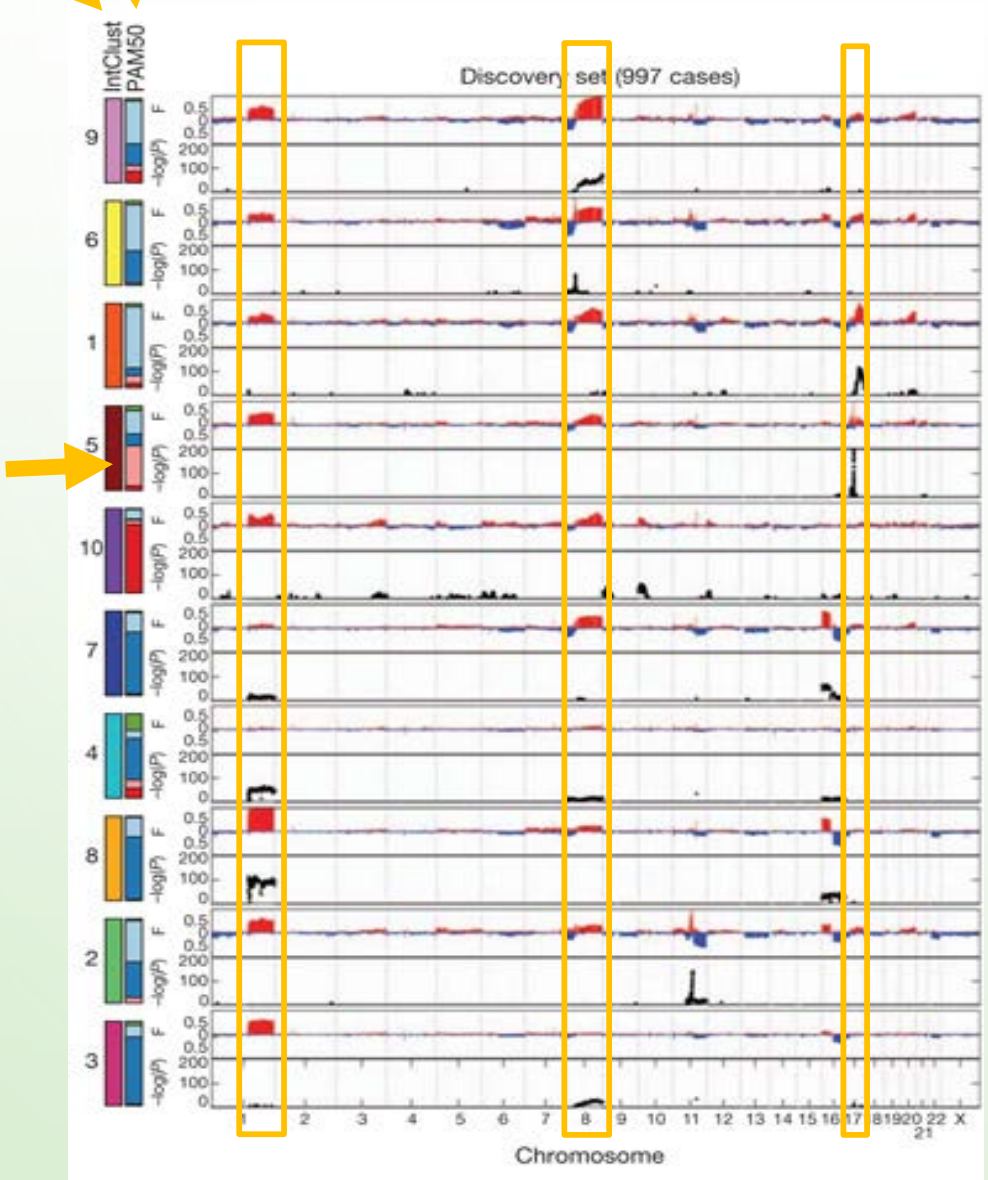
DC Koboldt *et al. Nature*, 1-10 (2012) doi:10.1038/nature11412

Link between gene expression clusters and genomic features



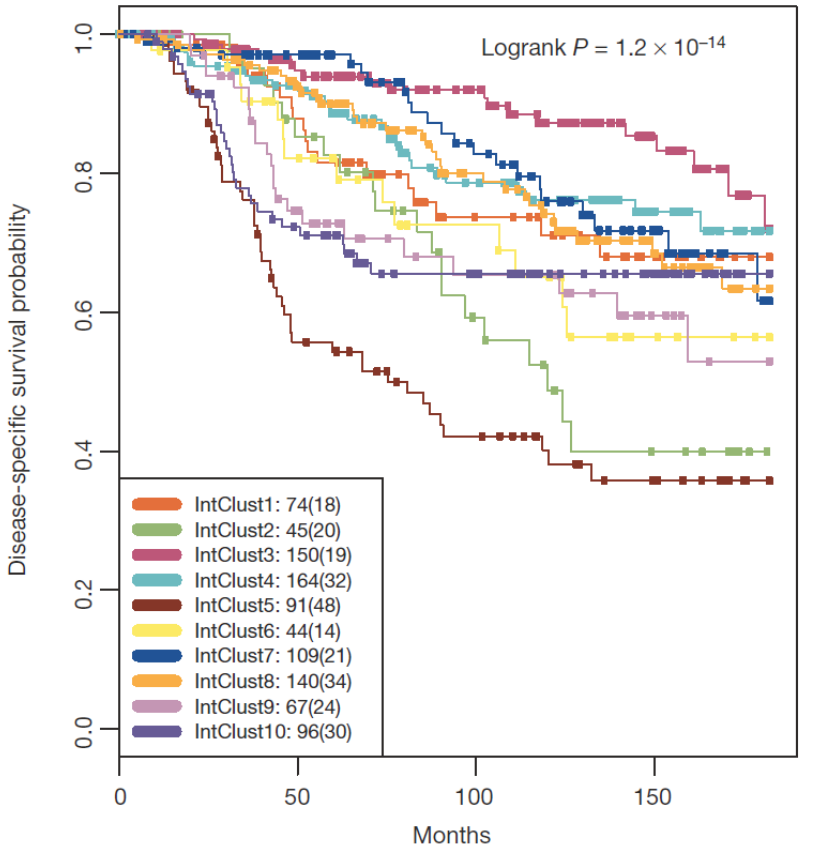
DC Koboldt *et al. Nature*, 1-10 (2012) doi:10.1038/nature11412

Clustering on Copy Number Changes and transcriptional landscape of thousands of tumours (*IntClust*)



The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups

Christina Curtis[1,2]*, Sohrab P. Shah[3,4]*, Suet-Feung Chin[1,2]*, Gulisa Turashvili[3,4]*, Oscar M. Rueda[1,2], Mark J. Dunning[2], Doug Speed[2,5]†, Andy G. Lynch[1,2], Shamith Samarajiwa[1,2], Yinyin Yuan[1,2], Stefan Gräf[1,2], Gavin Ha[3], Gholamreza Haffari[3], Ali Bashashati[3], Roslin Russell[2], Steven McKinney[3,4], METABRIC Group‡, Anita Langerød[6], Andrew Green[7], Elena Provenzano[8], Gordon Wishart[8], Sarah Pinder[9], Peter Watson[3,4,10], Florian Markowetz[1,2], Leigh Murphy[10], Ian Ellis[7], Arnie Purushotham[9,11], Anne-Lise Børresen-Dale[6,12], James D. Brenton[2,13], Simon Tavaré[1,2,5,14], Carlos Caldas[1,2,8,13] & Samuel Aparicio[3,4]

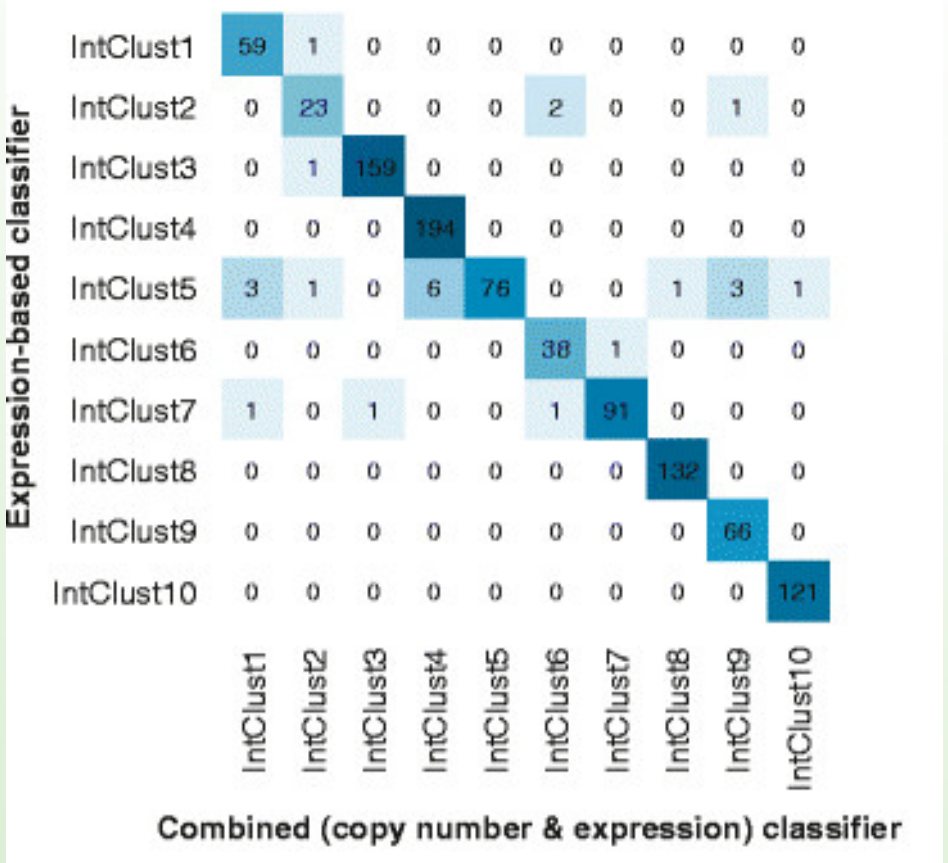


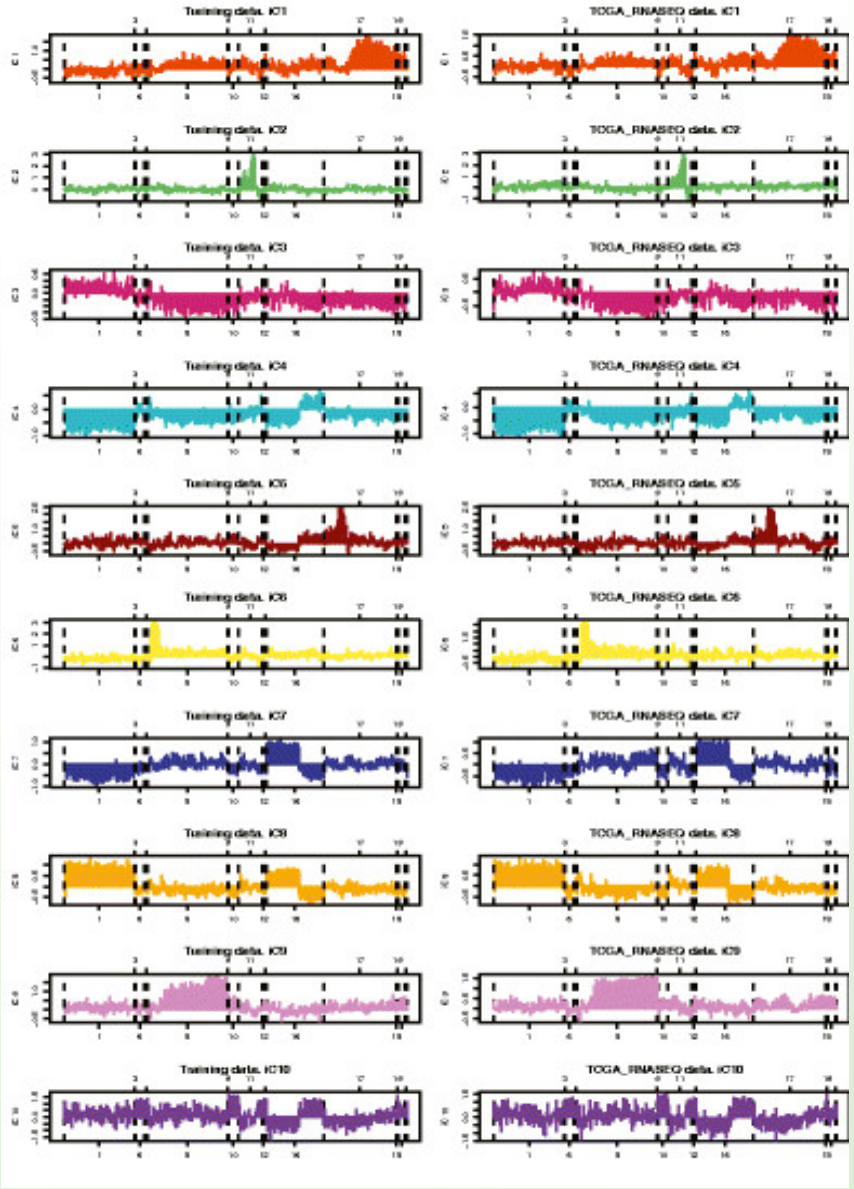
Gene expression-based approach for classifying breast tumors into the ten IntClust subtypes

Genome-driven integrated classification of breast cancer validated in over 7,500 samples

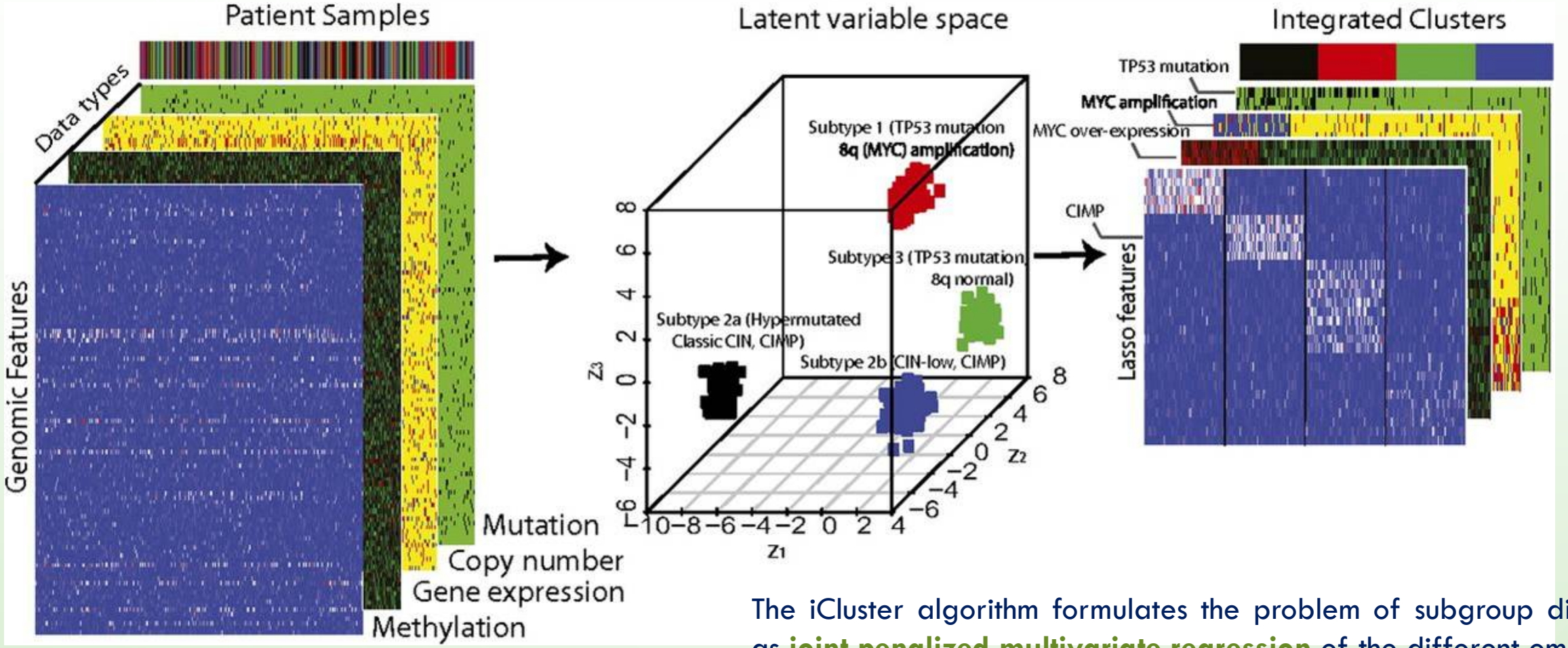
H Raza Ali, Oscar M Rueda, Suet-Feung Chin, Christina Curtis, Mark J Dunning, Samuel AJR Aparicio and Carlos Caldas

Genome Biology 2014 **15**:431 | **DOI:** 10.1186/s13059-014-0431-1





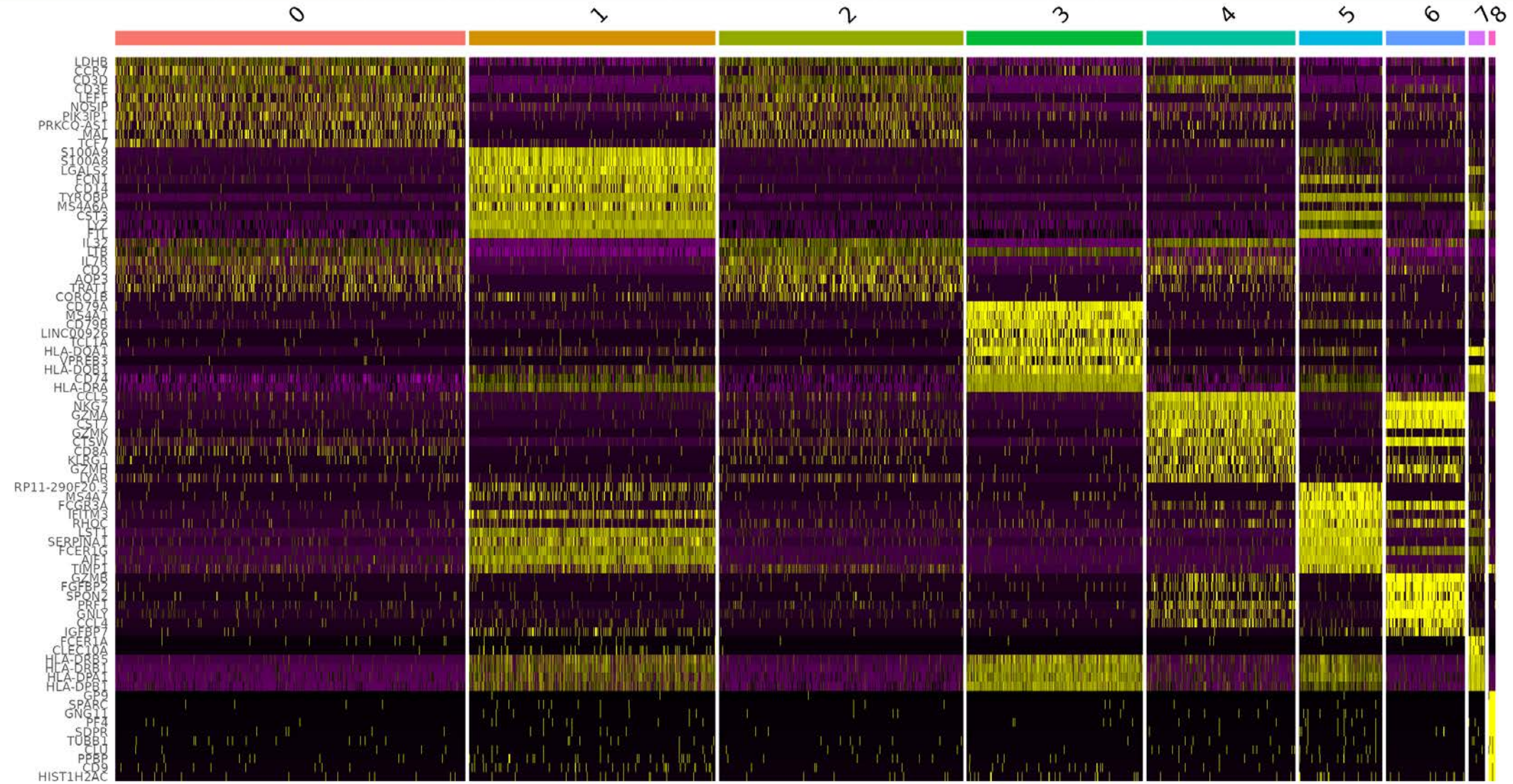
EXAMPLE OF INTEGRATED CLUSTERING METHOD: ICLUSTER



https://sites.google.com/site/ronglais/icluster

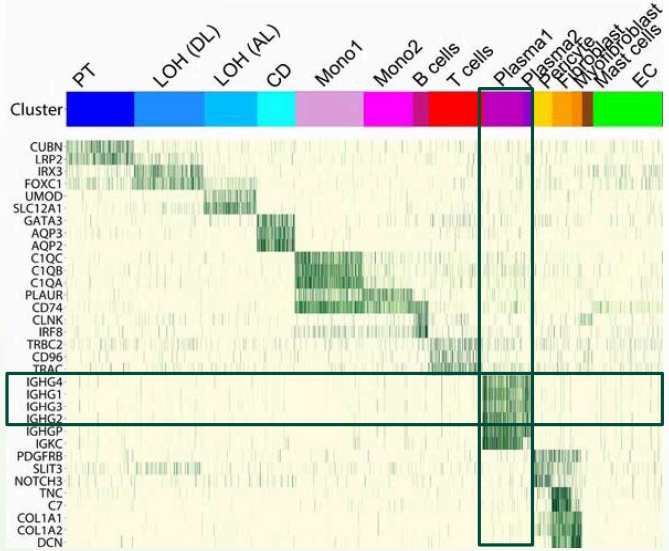
The iCluster algorithm formulates the problem of subgroup discovery as **joint penalized multivariate regression** of the different omics data types with **reference to a set of common latent variables**, which represent the underlying tumor subtypes Gaussian joint latent variable model.

CLUSTERING SINGLE CELLS RNASEQ

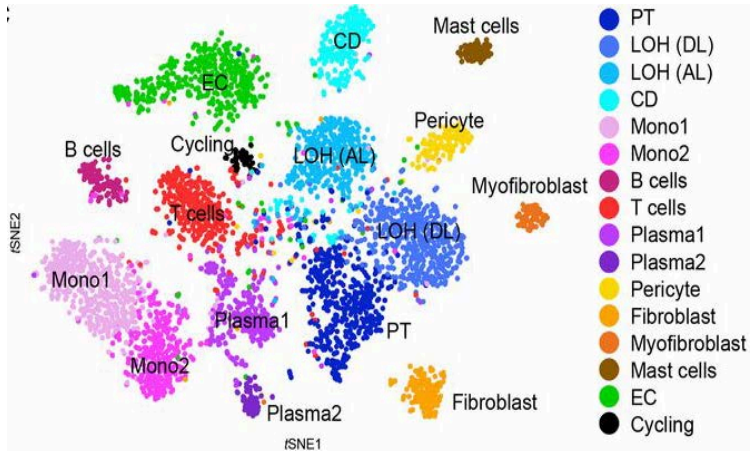


DISCOVERY OF DIFFERENT CELL PHENOTYPES

Identifying cell-type marker genes



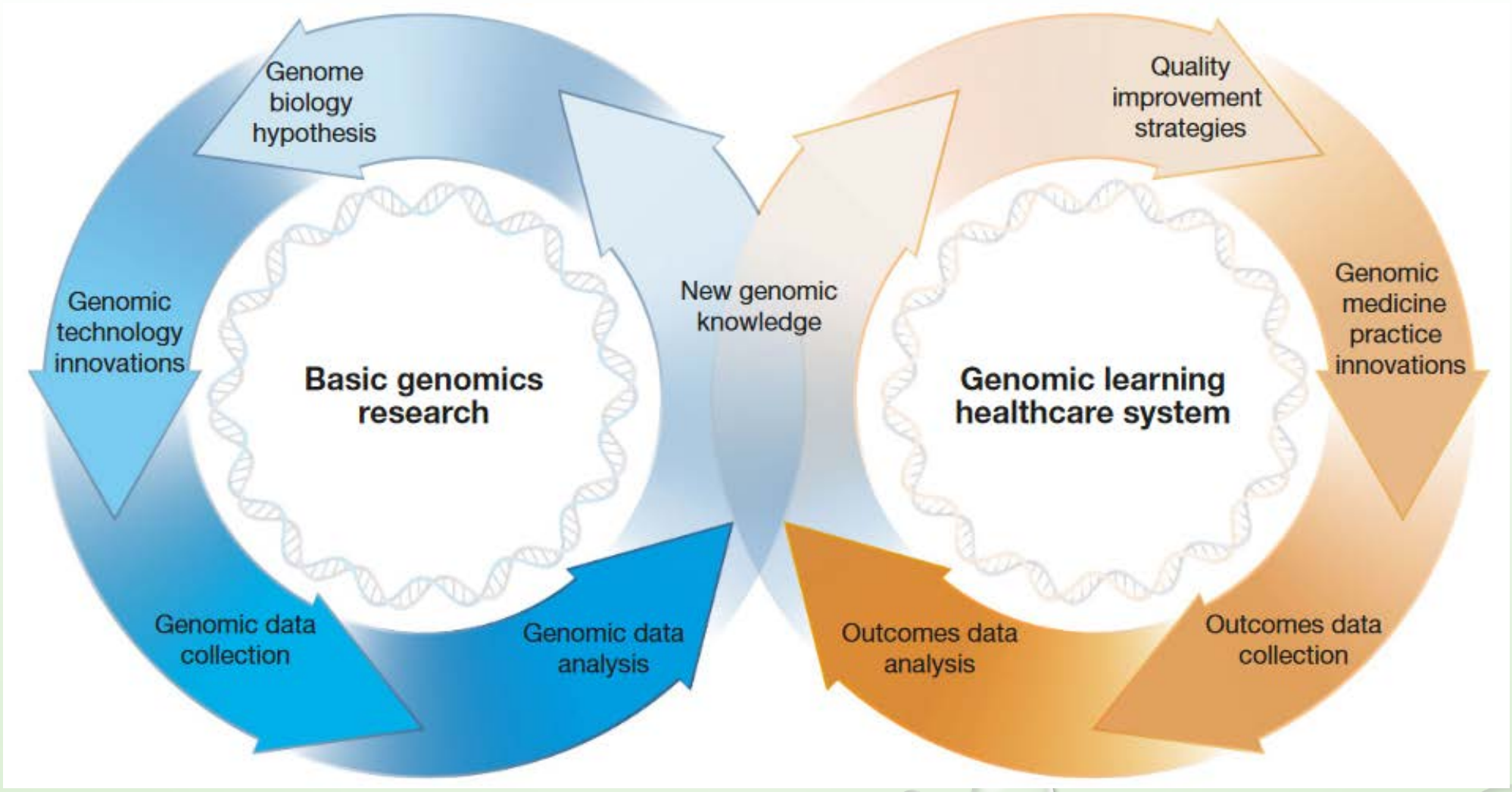
Discovering sample heterogeneity



Uncovering tissue dynamics

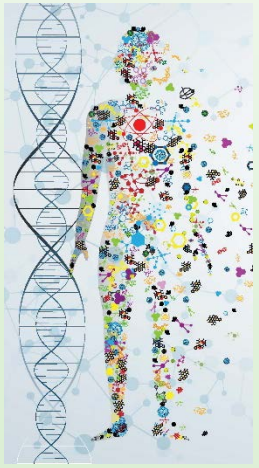
Images from: Wu et al (2018) JASN and Farbehi et al (2019) Elife

VIRTUOUS CYCLES IN HUMAN GENOMICS RESEARCH AND CLINICAL CARE



https://www.nature.com/articles/s41586-020-2817-4

Precision medicine paradigm

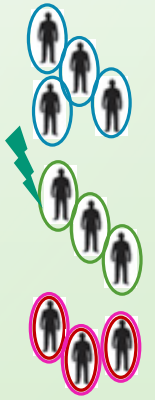


No benefit
High toxicity

High benefit
High toxicity

No benefit
Low toxicity

High benefit
Low toxicity



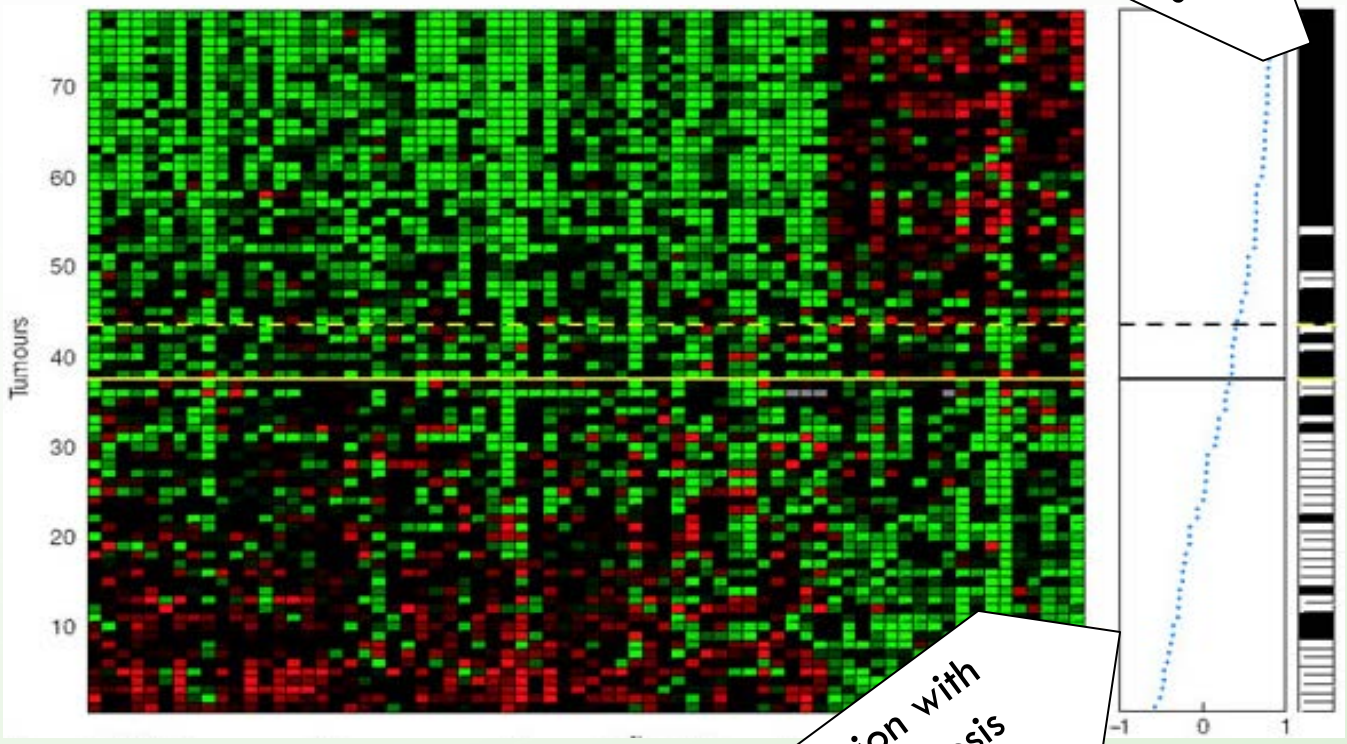
Biomarkers *Machine Learning* *Prediction* *Treatment*

A 70-GENE SIGNATURE FOR RISK OF METASTASIS

Sporadic breast tumours
Patient < 55 years
Tumour size < 5cm
Lymph node negative (LN0)

No distant metastases
> 5 years

Distant metastases
< 5 years



Van 't Veer et al, Nature, 2002

A 70-GENE SIGNATURE FOR RISK OF METASTASIS

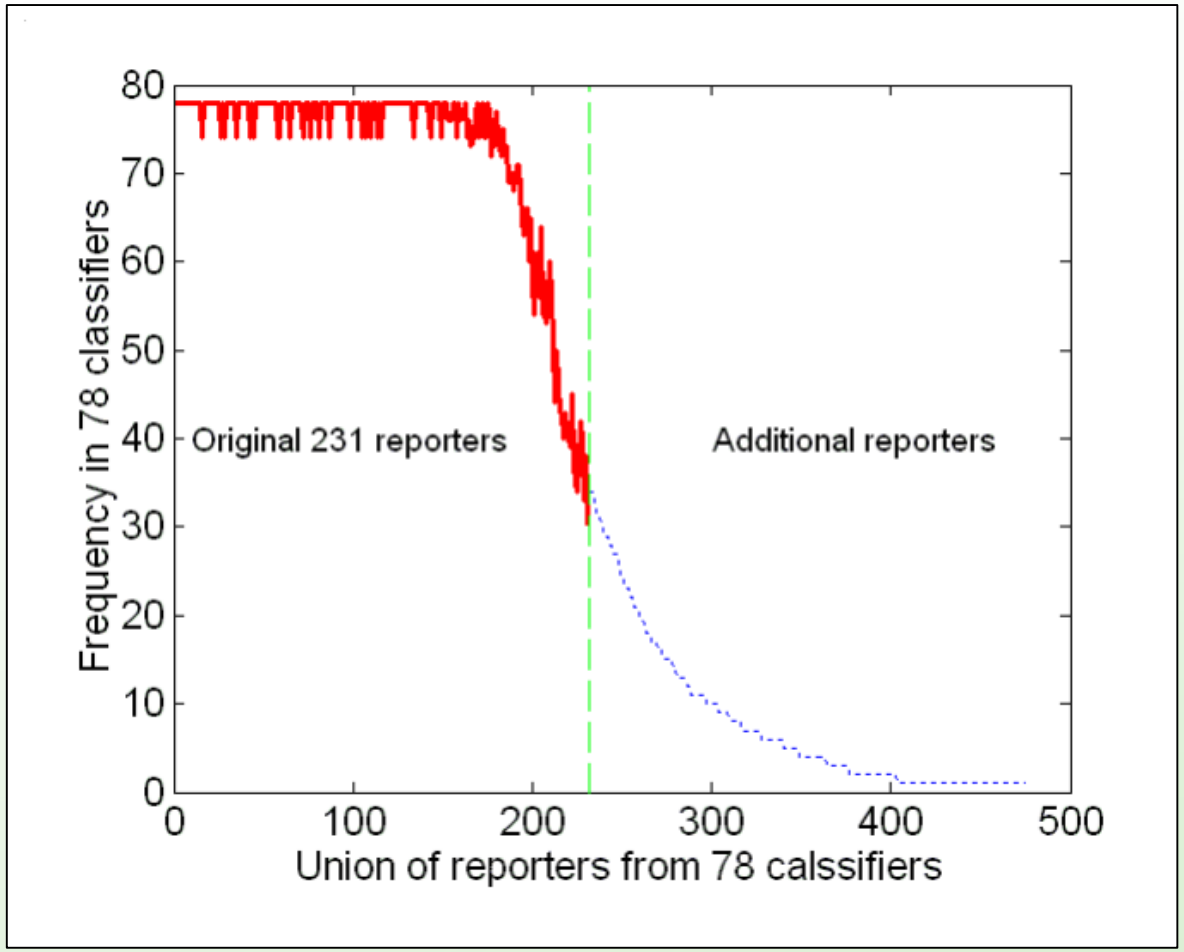
Correlation between the prognostic category (metastasis vs. no-metastasis) and the logarithmic expression ratio across all 78 samples for each individual gene in 5,000 significantly expressed genes.

Permutation to evaluate significance.

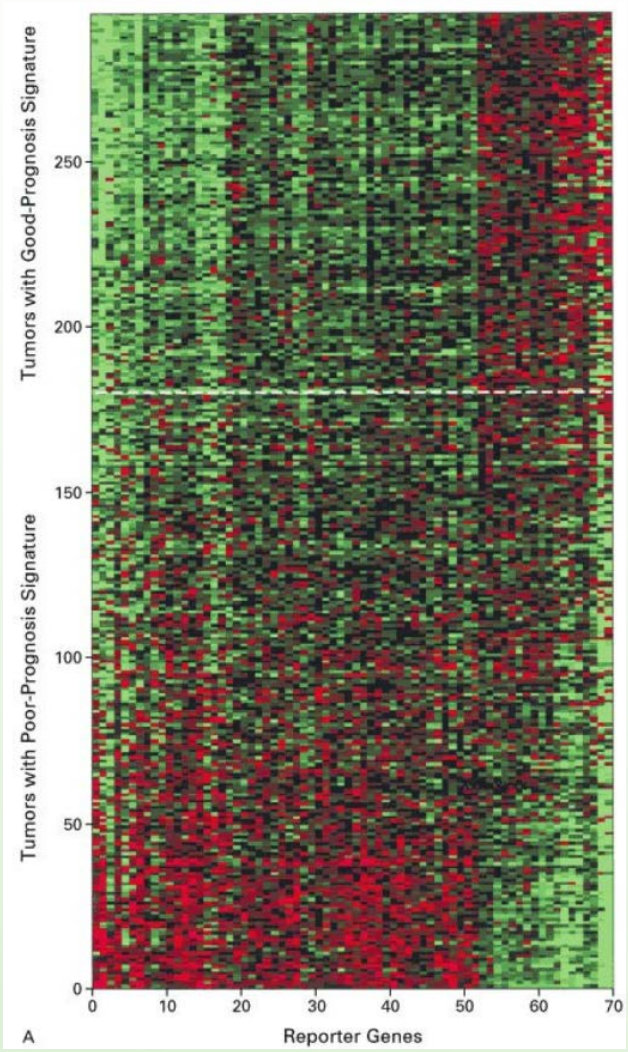
231 genes corr coeff > 0.3 or < – 0.3 selected (p= 0.3%)

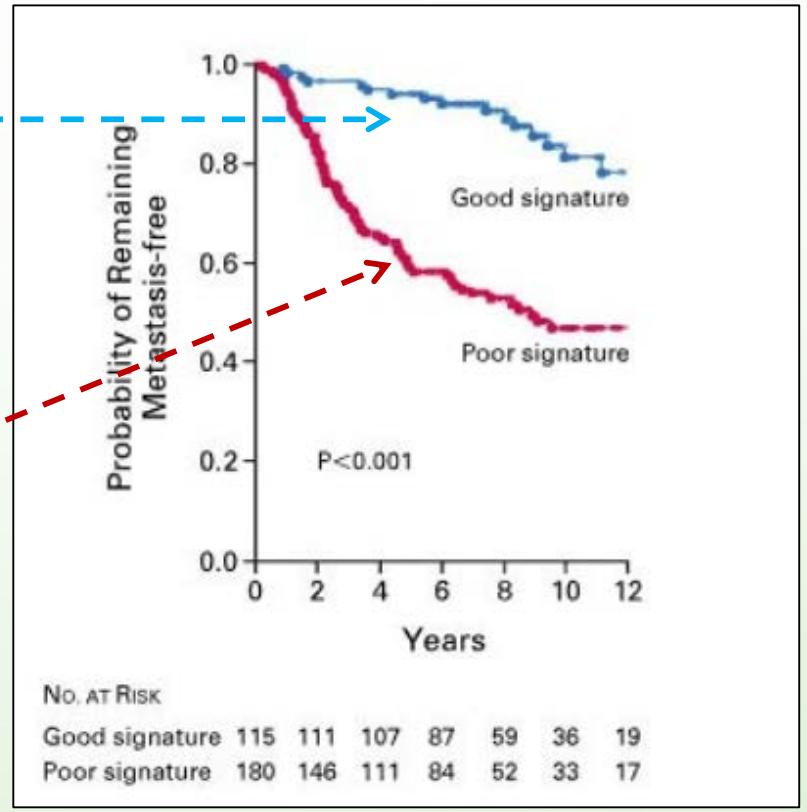
Leave-one-out: (1) leave one sample out, (2) define reporters based on the remaining 77 samples among the set of ~5000 significant genes, (3) use the reporters to predict the outcome of the one sample that was left out in step (1), (4) repeat steps (1)-(3) exhaustively for all 78 samples.

Select top 70



70-GENE SIGNATURE IS PREDICTIVE OF SURVIVAL AND RISK OF METASTASES [VAN DE VIJVER ET AL, N ENGL J MED, 2002]





Performance of gene expression signatures in published studies

Study reference	Cancer type	Clinical endpoint	Sample size	Number of events (%)	Number of channels (type)	Number of genes after filtration*
2	Non-Hodgkin lymphoma	Survival	240	138 (58%)	2 (Lymphochip)	6693
3	Acute lymphocytic leukaemia	Relapse-free survival	233	32 (14%)	1 (Affymetrix)	12 236
4	Breast cancer	5-year metastasis-free survival	97	46 (47%)	2 (Agilent)	4948
5	Lung adenocarcinoma	Survival	86	24 (28%)	1 (Affymetrix)	6532
6,7	Lung adenocarcinoma	4-year survival	62†	31 (50%)	1 (Affymetrix)	5403
8	Medulloblastoma	Survival	60	21 (35%)	1 (Affymetrix)	6778
9	Hepatocellular carcinoma	1-year recurrence-free survival	60	20 (33%)	1 (Affymetrix)	4861

*For the data of van 't Veer and colleagues,[4] the same filter was used as in the original publication. For other studies, genes with little variation in expression were excluded. †Only patients with clinical follow-up of at least 4 years after surgical resection were analysed.[7]

Table: Description of eligible studies ordered by sample size

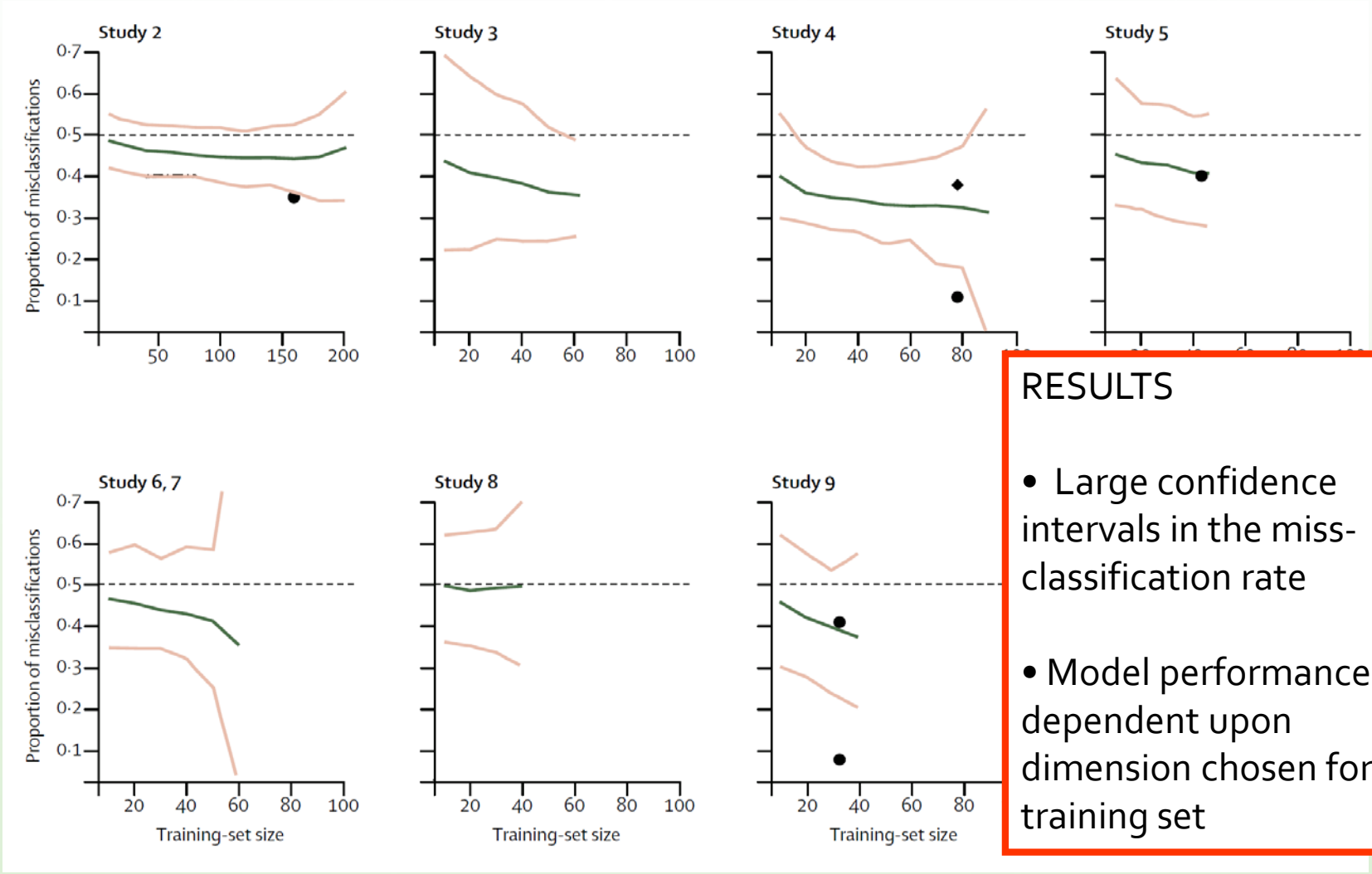
Michiels et al, Lancet 2005

Resampling strategy:

1) Sample a set of patients for training at random, and leave the rest for validation
2) Identify a gene expression signature (GES) for the clinical endpoint in the training set
3) Predict for the patients in validation set and estimate the proportion of misclassifications
4) Iterate on multiple random sets to study stability and performance of the GES

Performance of gene expression signatures in published studies

Misclassification rate from 500 random training-validation sets vs. training-set size (mean and 95% CIs)

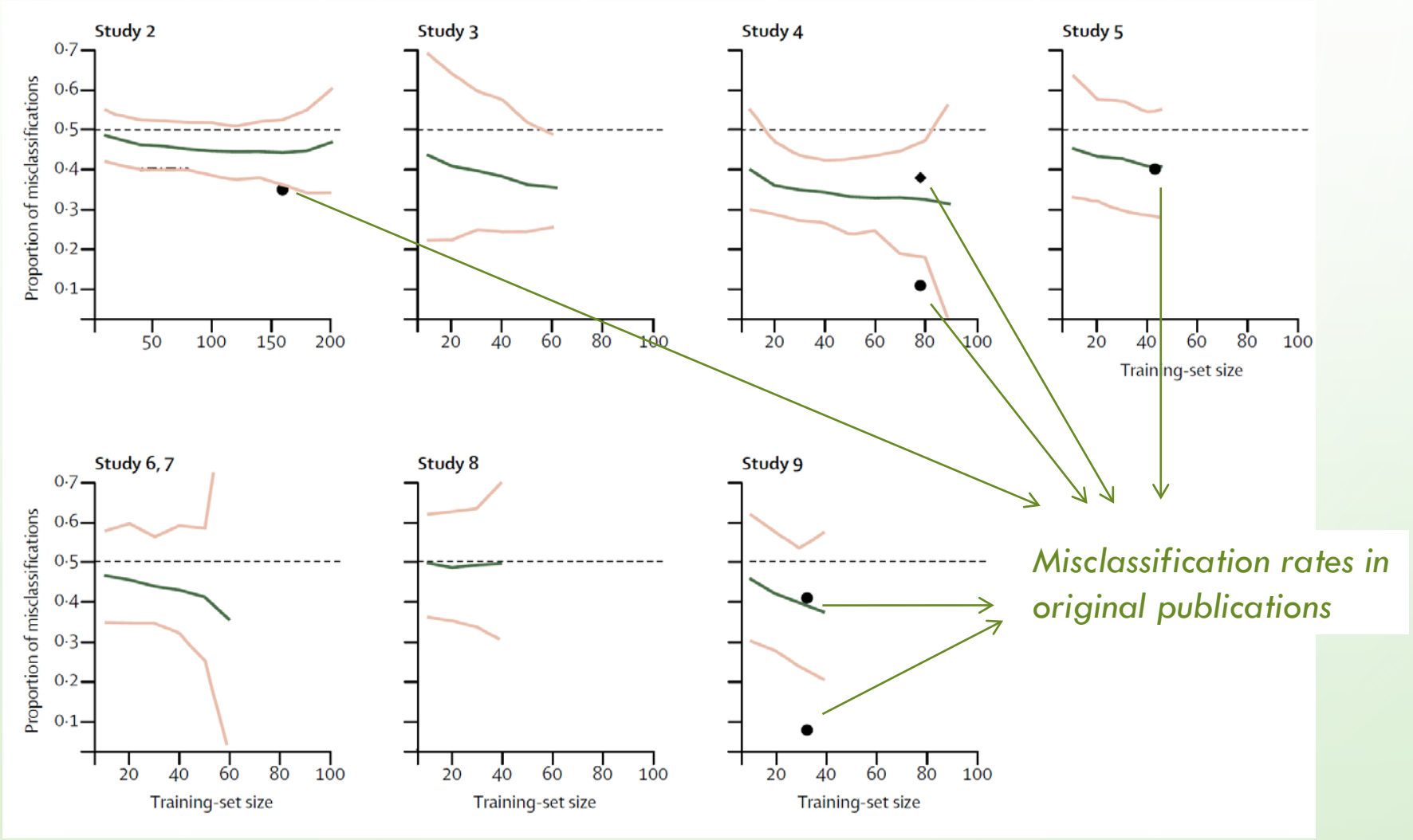


RESULTS

- Large confidence intervals in the miss-classification rate
- Model performance dependent upon dimension chosen for the training set

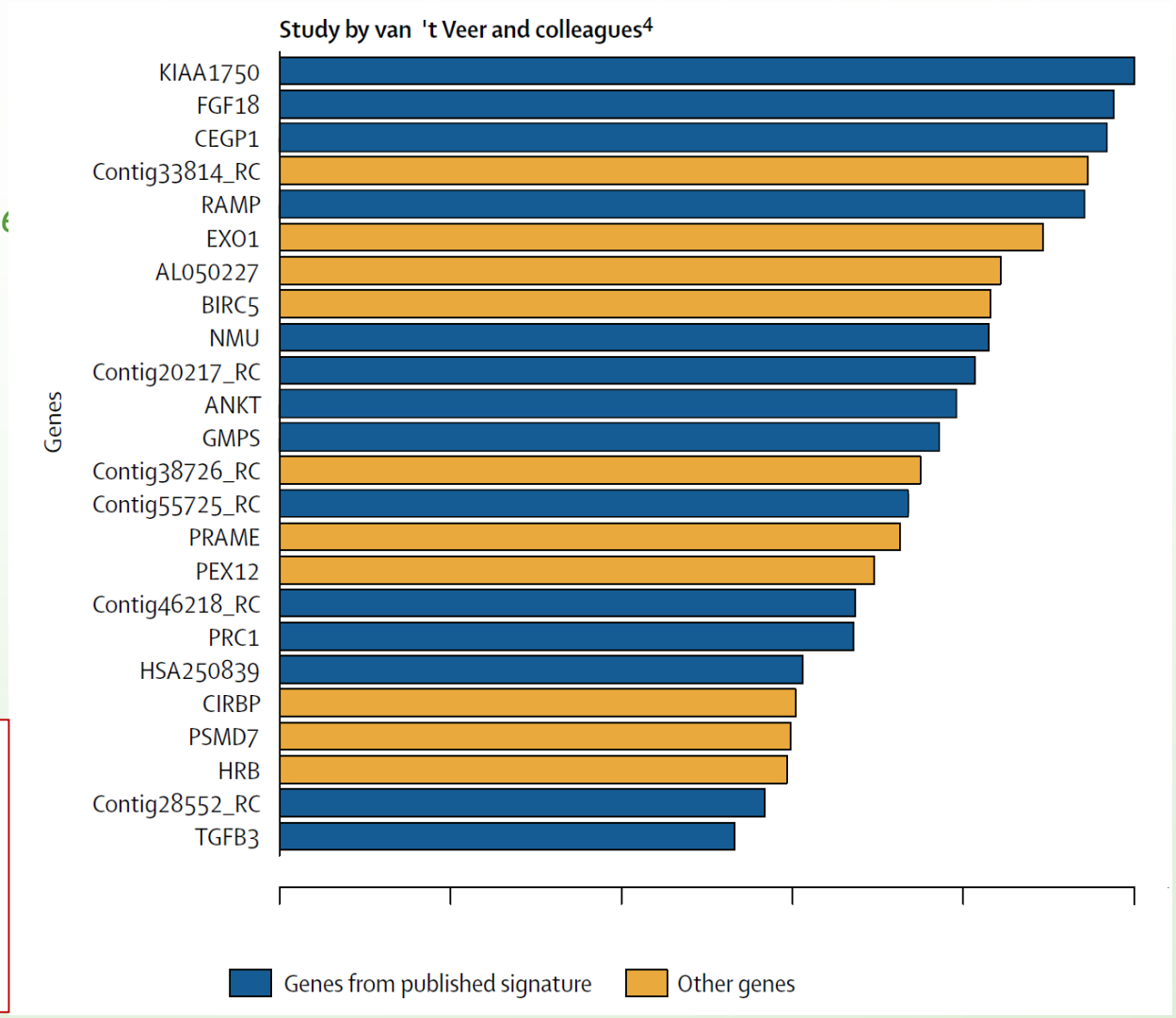
Performance of gene expression signatures in published studies

Misclassification rate from 500 random training-validation sets vs. training-set size (mean and 95% CIs)



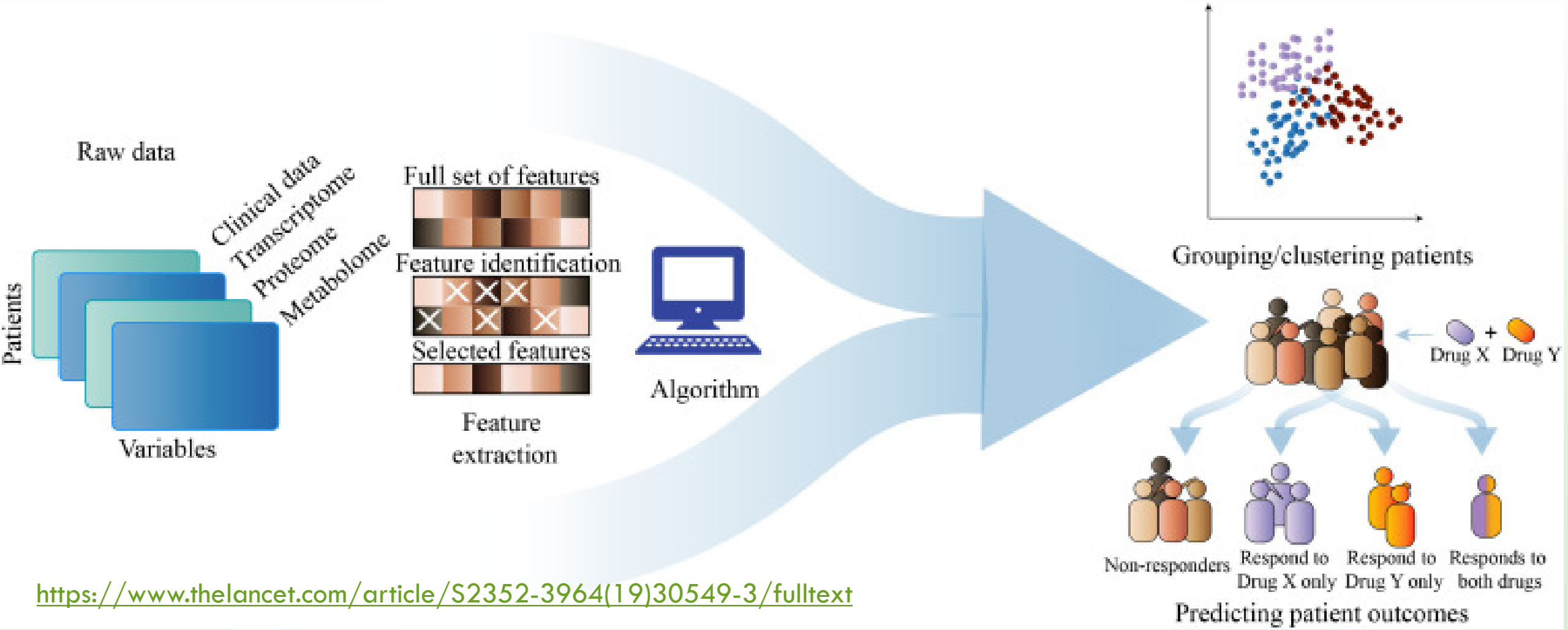
LARGE VARIABILITY IN SELECTED GENES

- *Several possible models with similar correlation with outcome*
- *Very little overlap in gene content between models*



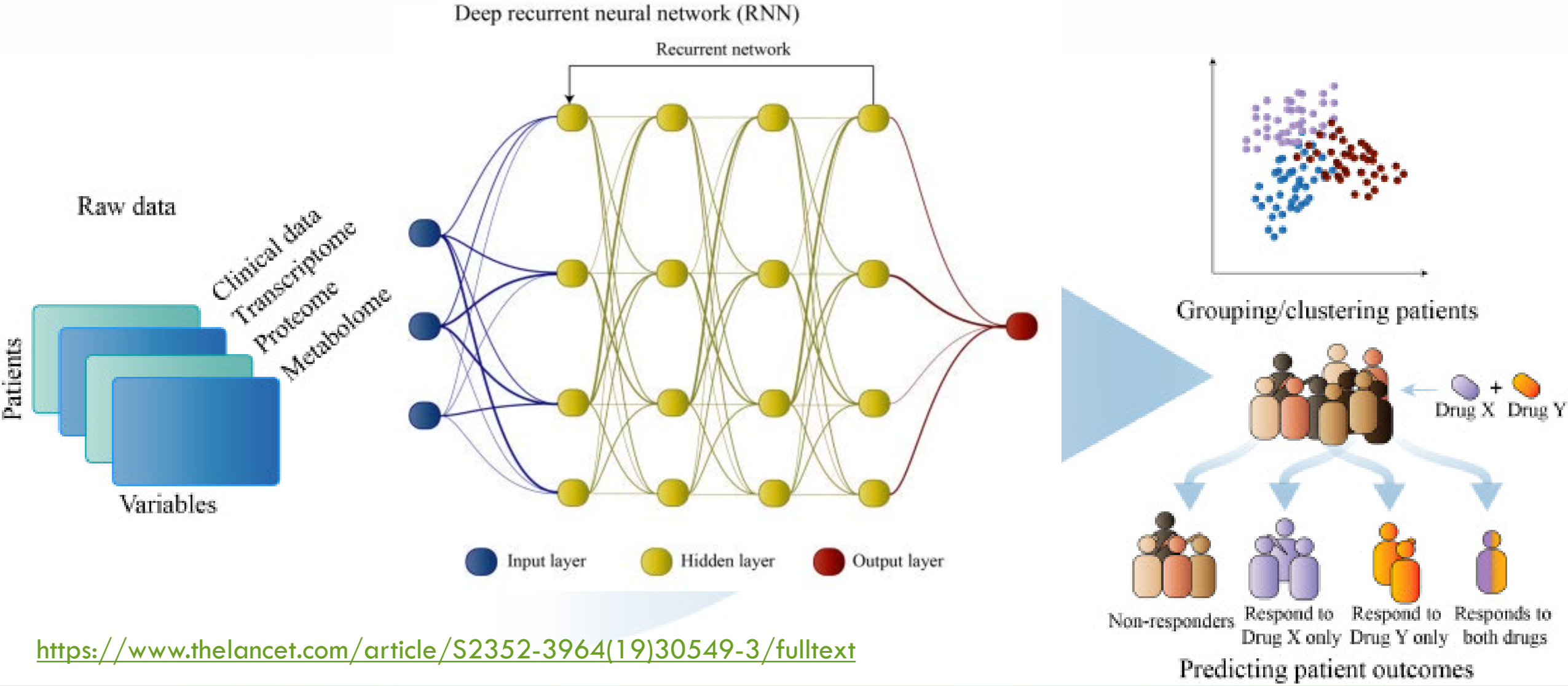
24 genes within the common set were selected in at least 50% of the random simulations

(DEEP) LEARNING TO IDENTIFY PREDICTIVE BIOMARKERS



https://www.thelancet.com/article/S2352-3964(19)30549-3/fulltext

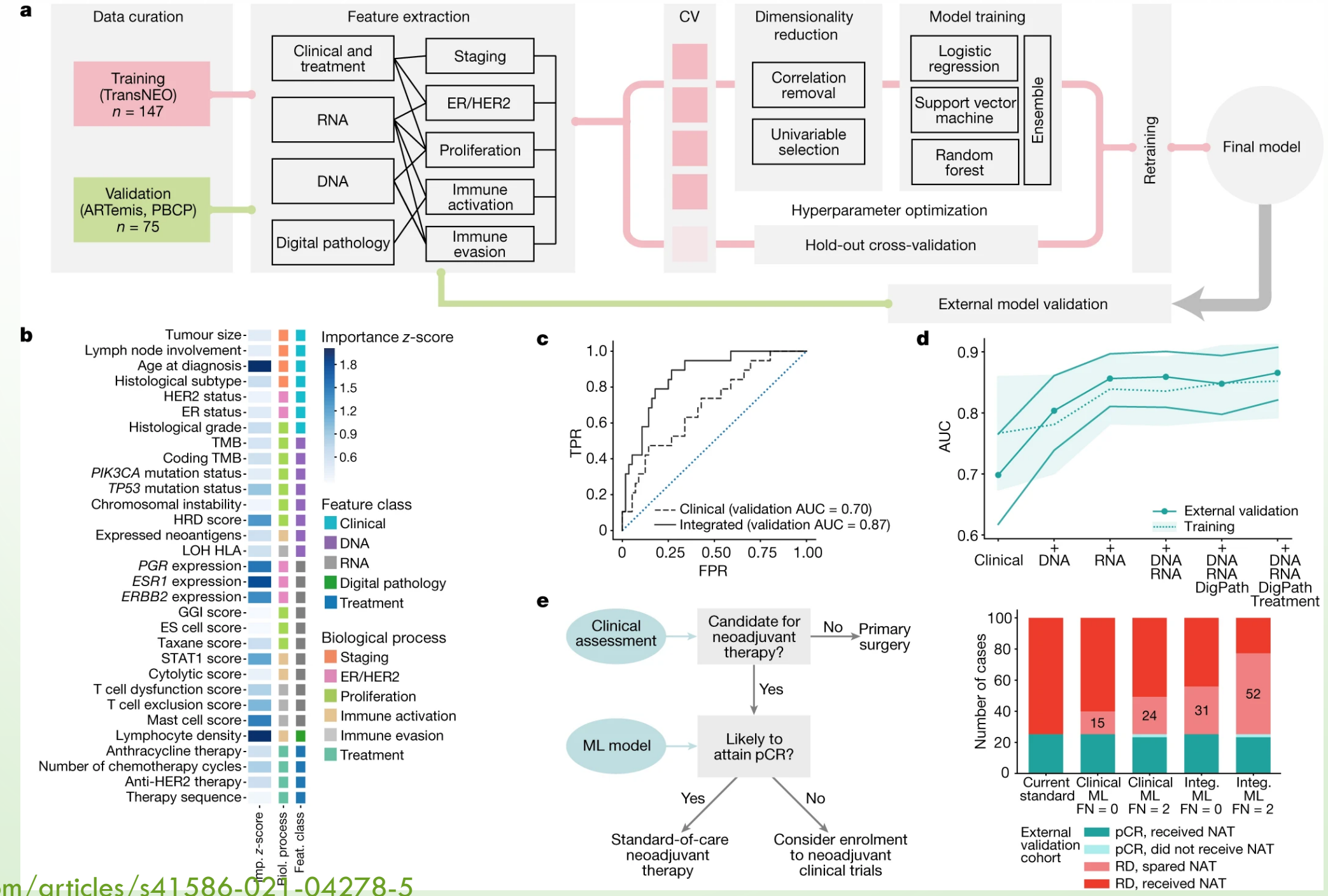
(DEEP) LEARNING TO IDENTIFY PREDICTIVE BIOMARKERS



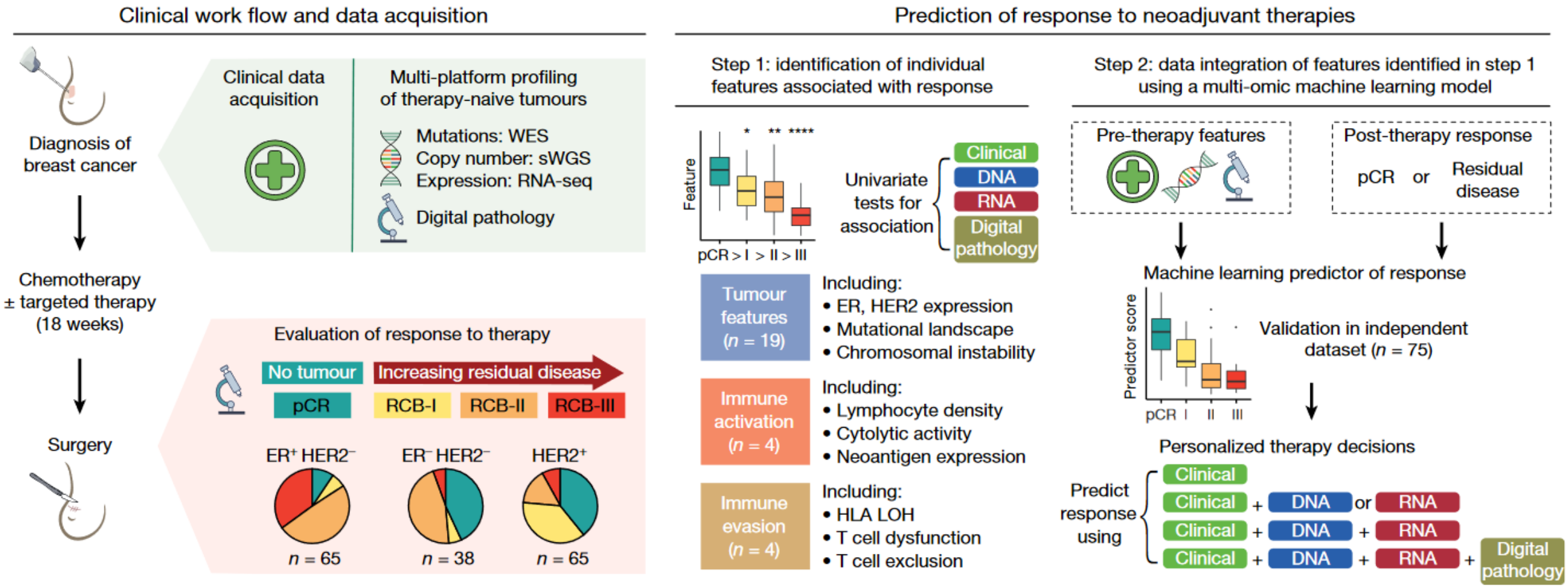
https://www.thelancet.com/article/S2352-3964(19)30549-3/fulltext

Multi-omic machine learning predictor of breast cancer therapy response

Nature, 623–629 (2022)



https://www.nature.com/articles/s41586-021-04278-5

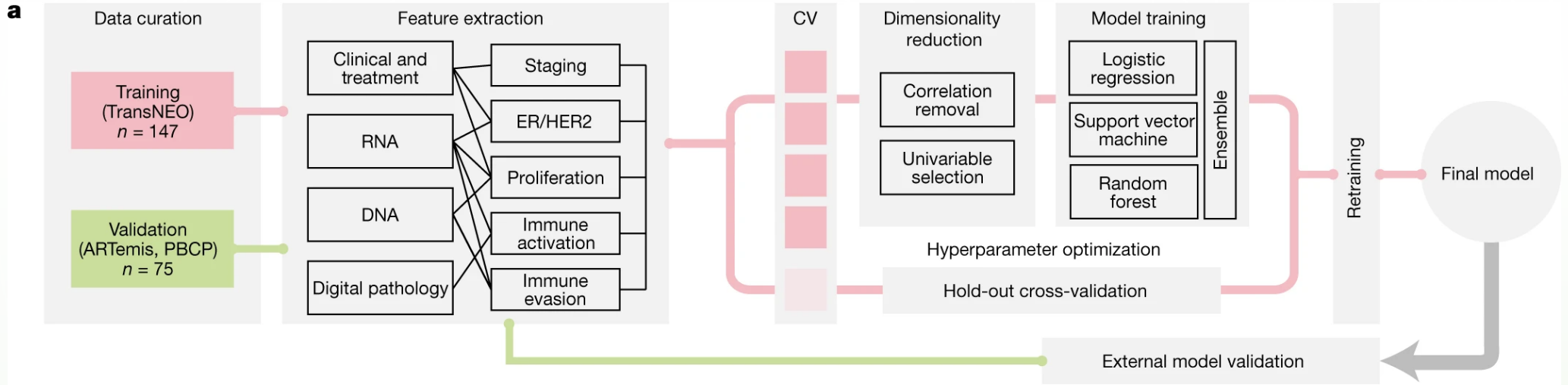


Methods

- clinical, digital pathology, genomic and transcriptomic profiles of pre-treatment biopsies of breast tumours from 168 patients treated with chemotherapy with or without HER2 (encoded by *ERBB2*)-targeted therapy before surgery.
- Pathology end points: complete response or residual disease
- Multi-omic features in these diagnostic biopsies

Multi-omic machine learning predictor of breast cancer therapy response

Nature, 623–629 (2022)



Methods

- Six pCR prediction models including different feature combinations were derived using: (1) clinical features only, and adding (2) DNA, (3) RNA, (4) DNA and RNA, (5) DNA, RNA and digital pathology, and (6) DNA, RNA, digital pathology and treatment.

- The models were based on a multi-step predictor pipeline. Features were first filtered by univariable selection and collinearity reduction, and then fed into an unweighted ensemble classifier.

- Each ensemble consisted of three algorithms acting in parallel: logistic regression with elastic net regularization, a support vector machine and a random forest. The three algorithm scores were then averaged to form the predictor.

- A fivefold cross-validation scheme was used to optimize model hyperparameters

https://www.nature.com/articles/s41586-021-04278-5

To maximise the robustness of the predictions, the models contain two levels of averaging. Firstly, predictions are obtained by averaging three classifier pipelines, as follows:

$$Prob(attaining\ pCR) = \frac{1}{3} \times (Pipeline_{LR}^{HER2+} + Pipeline_{SVC}^{HER2+} + Pipeline_{RF}^{HER2+}),$$

where

$$Pipeline_{Classifier} = \{Coll.Reduction\ (0.8) \Longrightarrow Univ.selection\ \Longrightarrow Classifier\},$$

with the corresponding hyperparameters listed in Tables 1 and 2. In particular, model hyperparameters that were *fixed a priori* are listed in Table 1, while model hyperparameters that were *optimised* for each of the classifiers using a 5-fold cross-validation setup are listed in Table 3. Once all hyperparameters are set, the model is re-trained on the entire training cohort, and subsequently frozen.

Secondly, to account for possible biases in the optimisation due to the particular cross validation splitting used, we repeated the process explained above 5 times, with 5 different cross-validation splitting seeds (integers from 1 to 5). As a result, we obtain 5 alternative optimised models. The final predictions are the average of the 5.

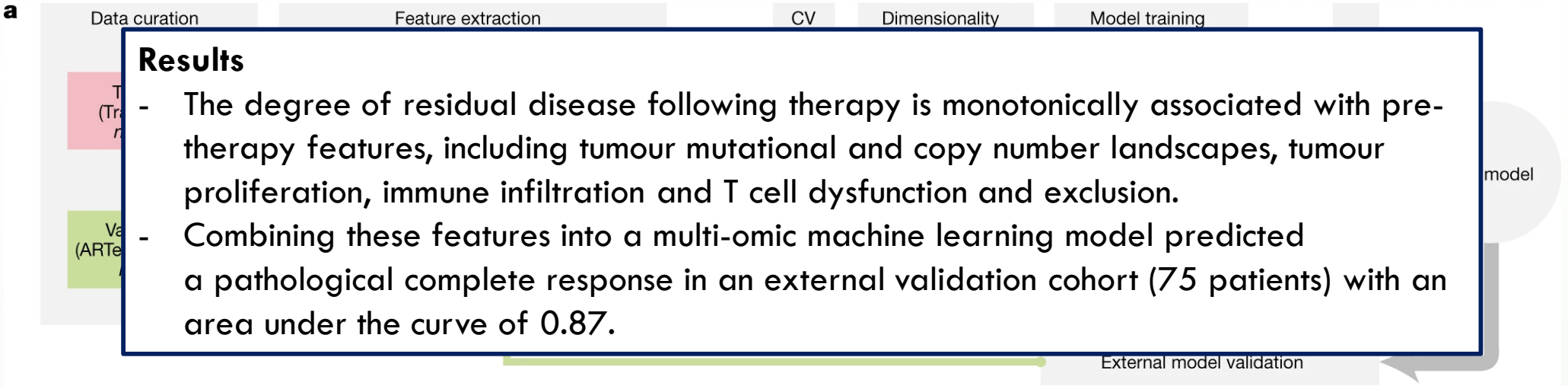
From Suppl Methods:
https://static-content.springer.com/esm/art%3A10.1038%2Fs41586-021-04278-5/MediaObjects/41586_2021_4278_MOESM1_ESM.pdf
Code used:
https://github.com/cclab-brca/neoadjuvant-therapy-response-predictor/blob/master/R/06%20-%20ML%20predictor.R

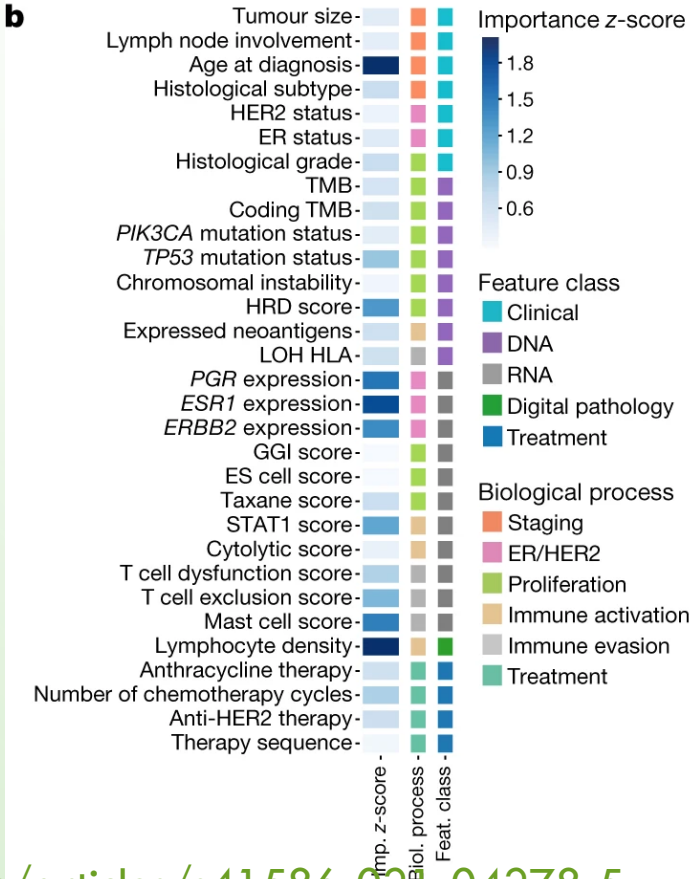
Multi-omic machine learning predictor of breast cancer therapy response

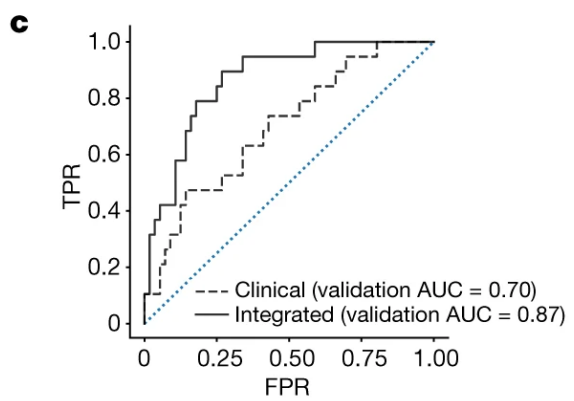
Nature, 623–629 (2022)

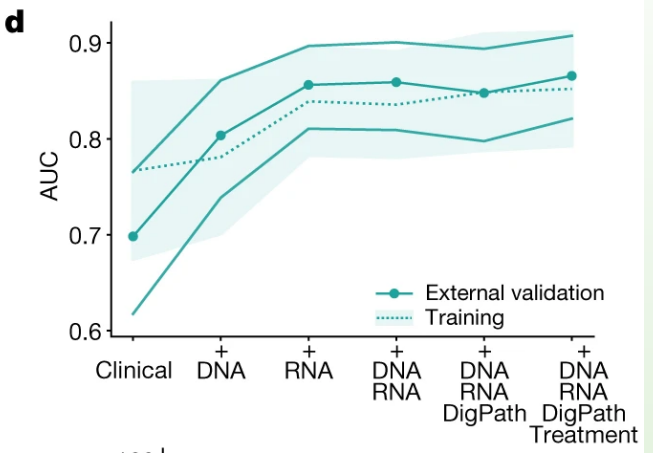
Results

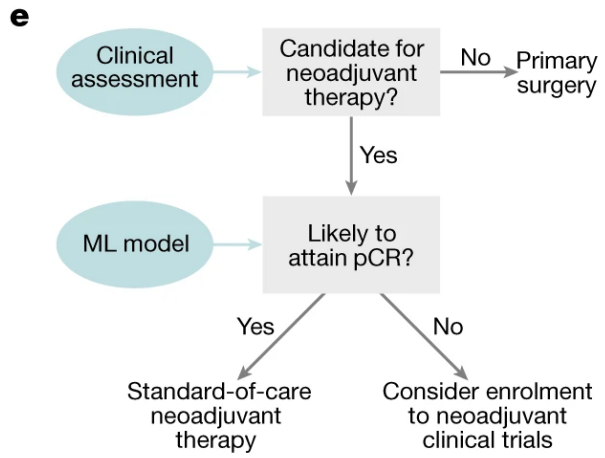
- The degree of residual disease following therapy is monotonically associated with pre-therapy features, including tumour mutational and copy number landscapes, tumour proliferation, immune infiltration and T cell dysfunction and exclusion.
- Combining these features into a multi-omic machine learning model predicted a pathological complete response in an external validation cohort (75 patients) with an area under the curve of 0.87.

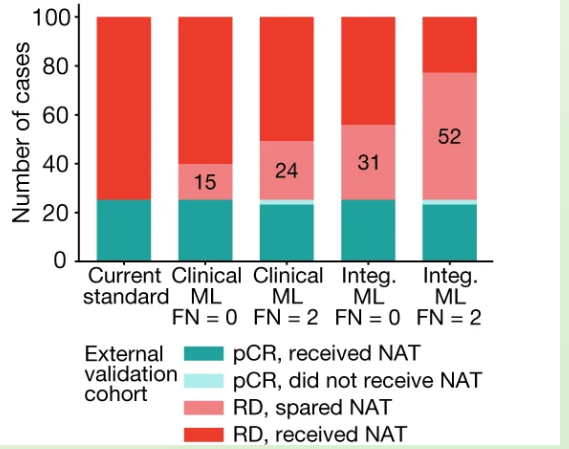




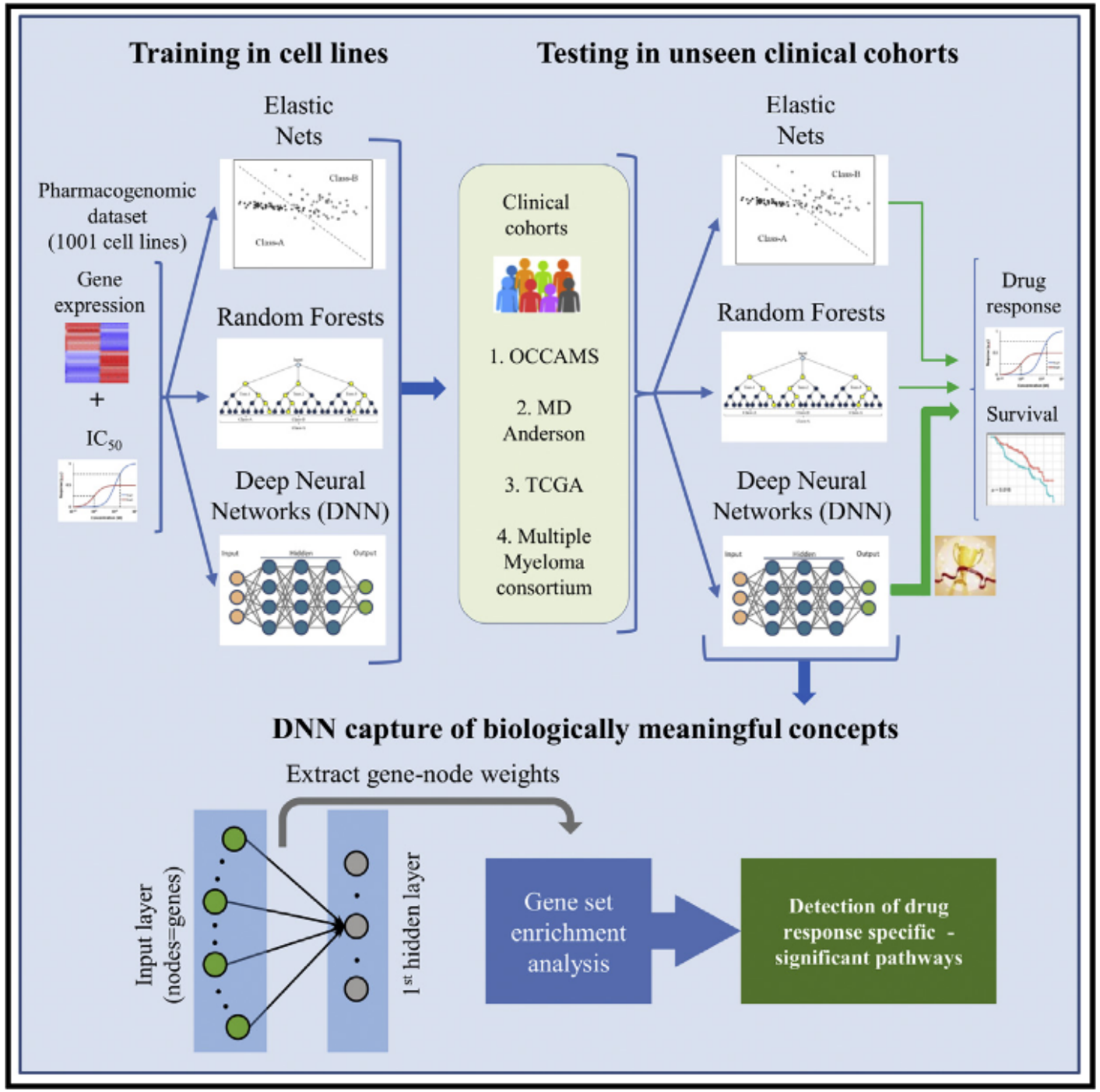








https://www.nature.com/articles/s41586-021-04278-5



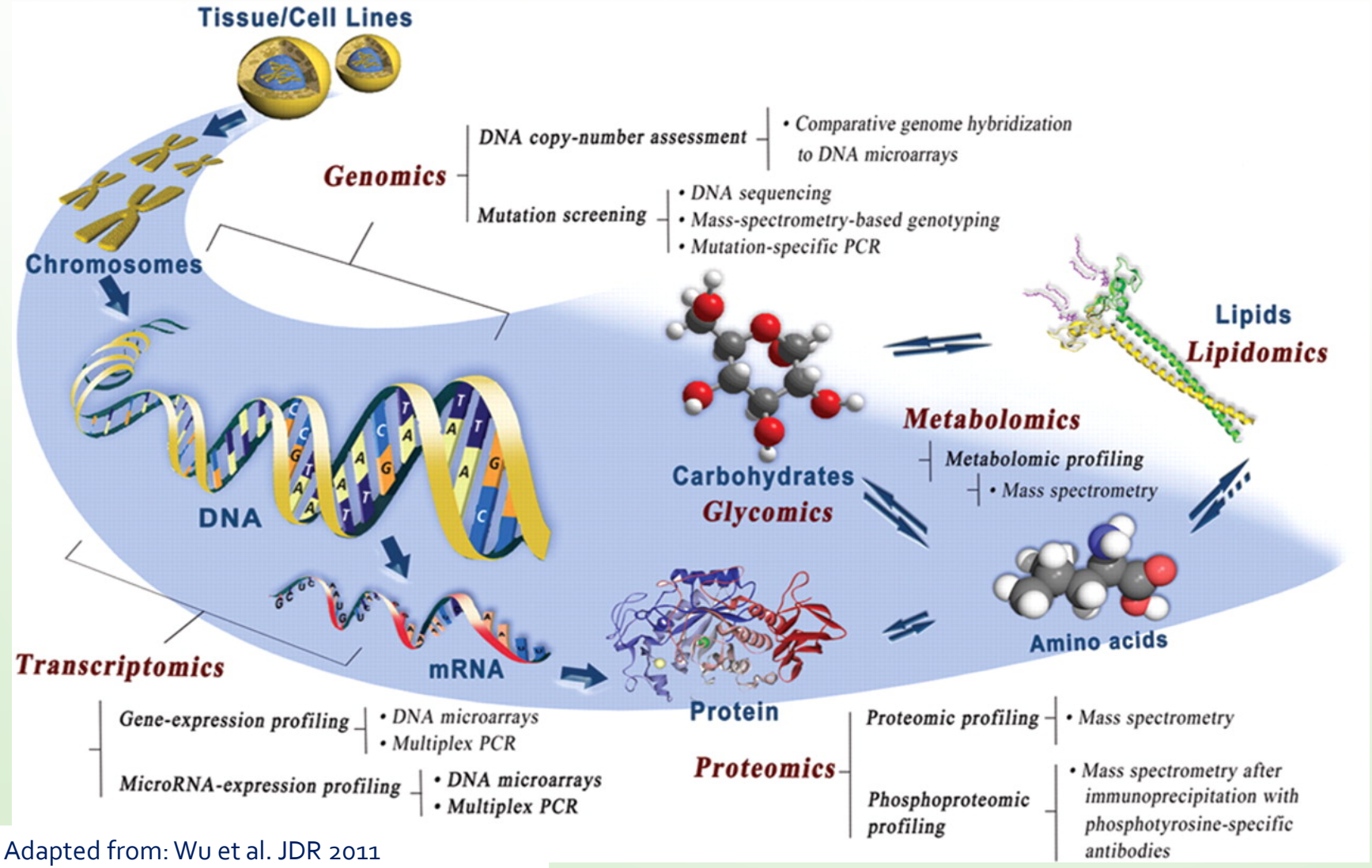
A Deep Learning Framework for Predicting Response to Therapy in Cancer

https://doi.org/10.1016/j.celrep.2019.11.017

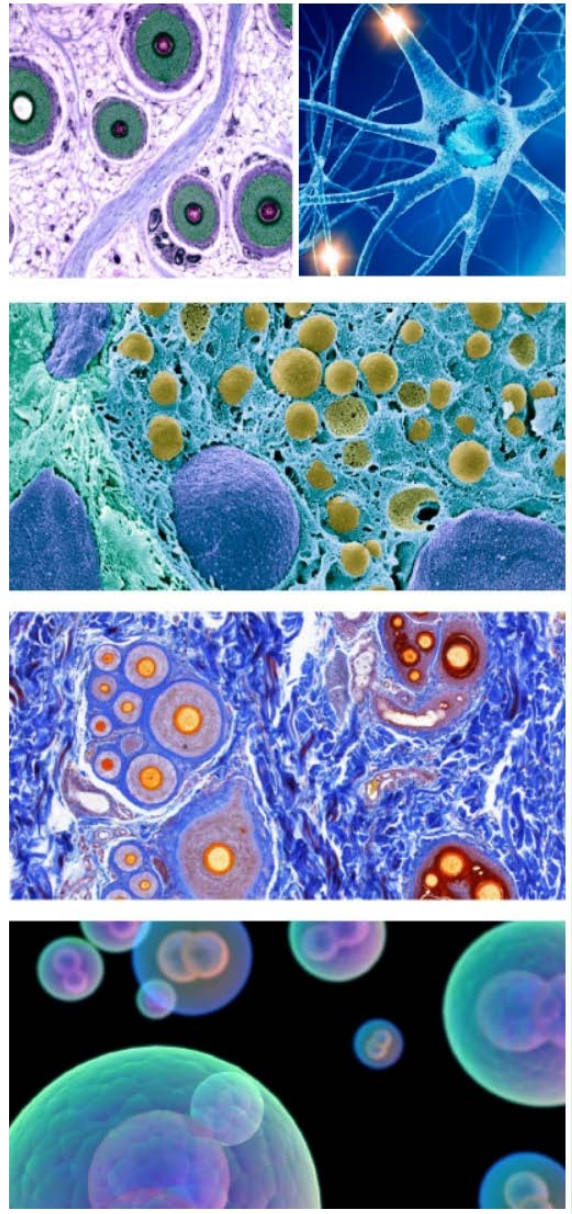
Highlights

- A machine learning (ML) workflow is designed to predict drug response in cancer patients
- Deep neural networks (DNNs) surpass current ML algorithms in drug response prediction
- DNNs predict drug response and survival in various large clinical cohorts
- DNNs capture intricate biological interactions linked to specific drug response pathways

From genotype to phenotype

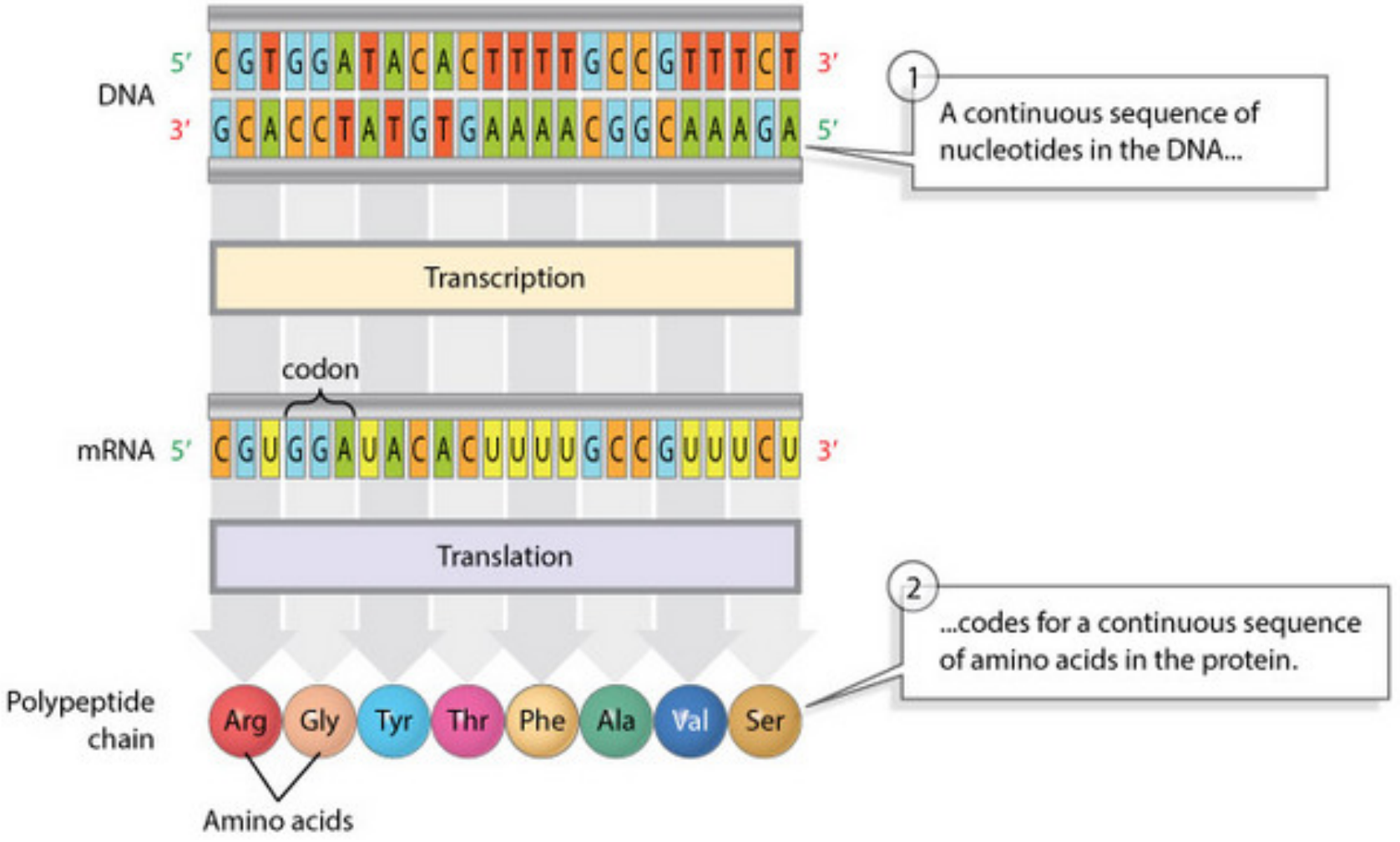


Adapted from: Wu et al. JDR 2011



Adapted from https://www.broadinstitute.org

FROM DNA TO FUNCTION



With colinearity, the number of nucleotides in the gene is proportional to the number of amino acids in the protein.

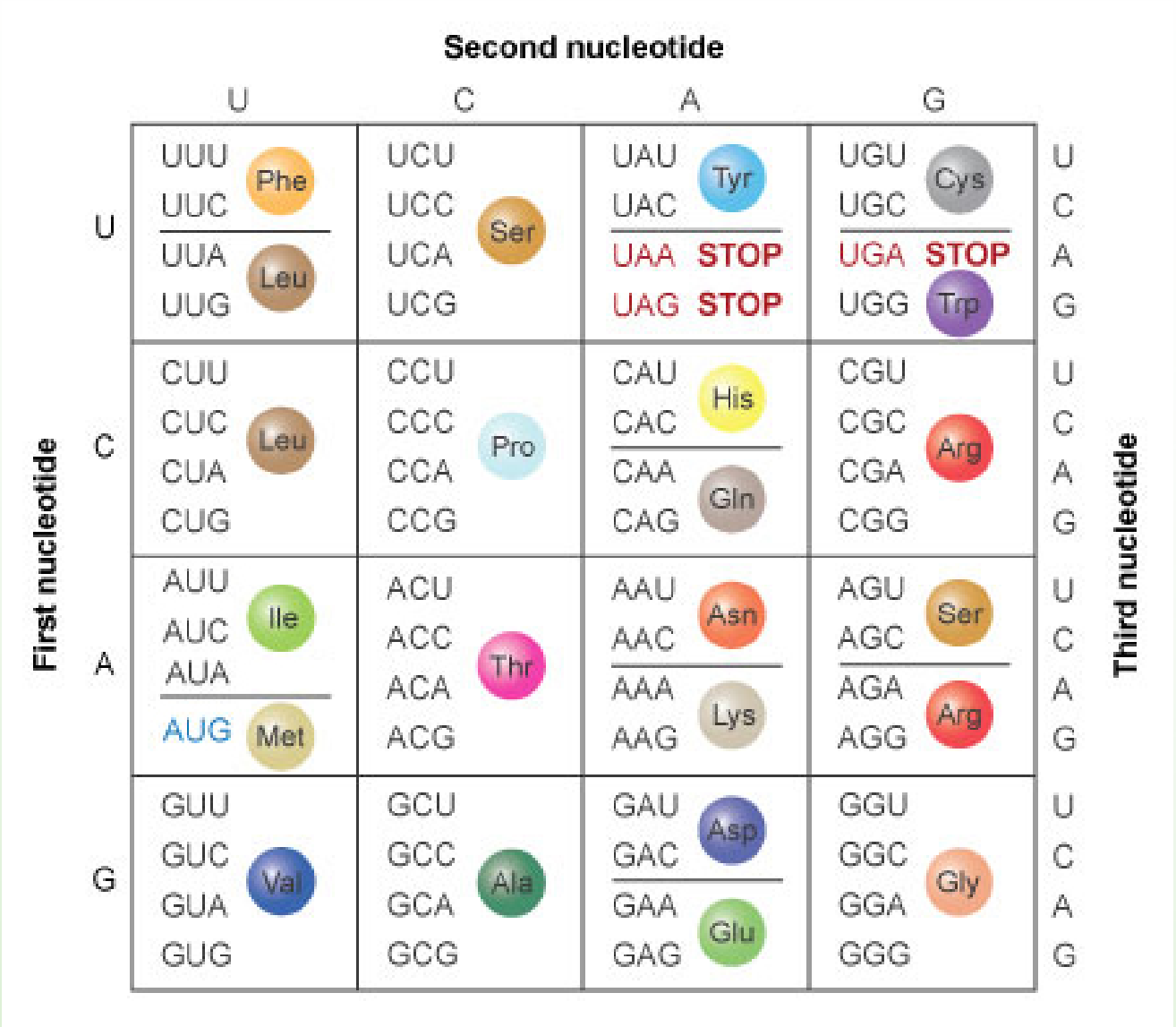
4 bases

3 letters (triplet – a codon) code for one amino acid in a protein

Codons (3 leter strings) specify 20 'words' (amino acid)

https://www.nature.com/scitable/topicpage/what-is-a-gene-colinearity-and-transcription-430/

$4^3 = 64$ possible ways (order important, repetition allowed) to code for 20 amino acids and stop codons



- Alanine
- Arginine
- Asparagine
- Aspartic Acid
- Cysteine
- Glutamic acid
- Glutamine
- Glycine
- Histidine
- Isoleucine
- Leucine
- Lysine
- Methionine
- Phenylalanine
- Proline
- Serine
- Threonine
- Tryptophan
- Tyrosine
- Valine

https://www.nature.com/scitable/topicpage/translation-dna-to-mrna-to-protein-393/

A SINGLE BASE CHANGE CAN CREATE DEVASTATING GENETIC DISORDERS

e.g. Sickle-Cell Anemia

In sickle-cell anemia, the gene for the beta chain of the hemoglobin protein (the oxygen-carrying protein that makes blood red) is mutated.

Beta hemoglobin (beta globin) is a single chain of 147 amino acids.

One single-base mutation causes the sixth amino acid in the chain to be valine, rather than glutamic acid. This changes the red blood cell shape and ability to carry oxygen, with huge consequences for the individual.

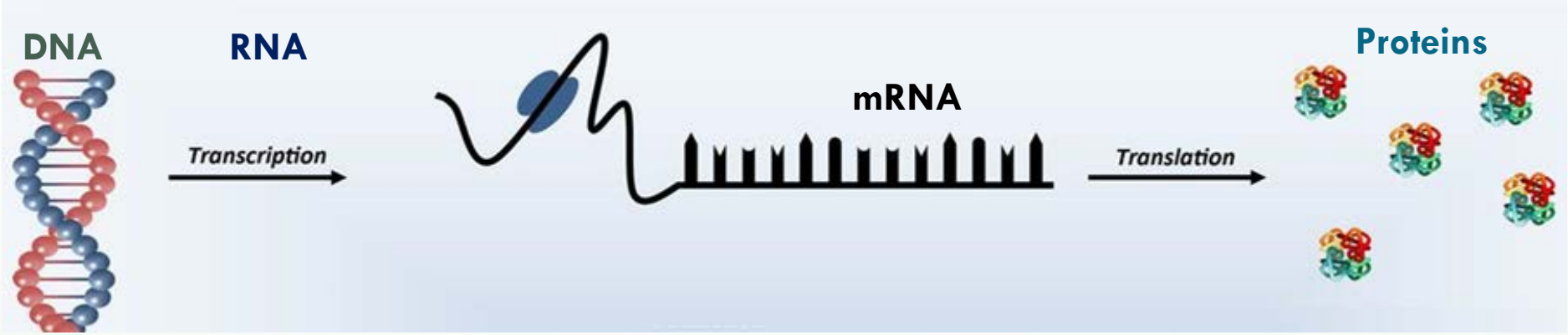
Wild-Type

ATG	GTG	CAC	CTG	ACT	CCT	GAG	GAG	AAG	TCT	GCC	GTT	ACT
Start	Val	His	Leu	Thr	Pro	Glu	Glu	Lys	Ser	Ala	Val	Thr

Mutant

ATG	GTG	CAC	CTG	ACT	CCT	GTG	GAG	AAG	TCT	GCC	GTT	ACT
Start	Val	His	Leu	Thr	Pro	Val	Glu	Lys	Ser	Ala	Val	Thr

DNA, RNA AND PROTEINS

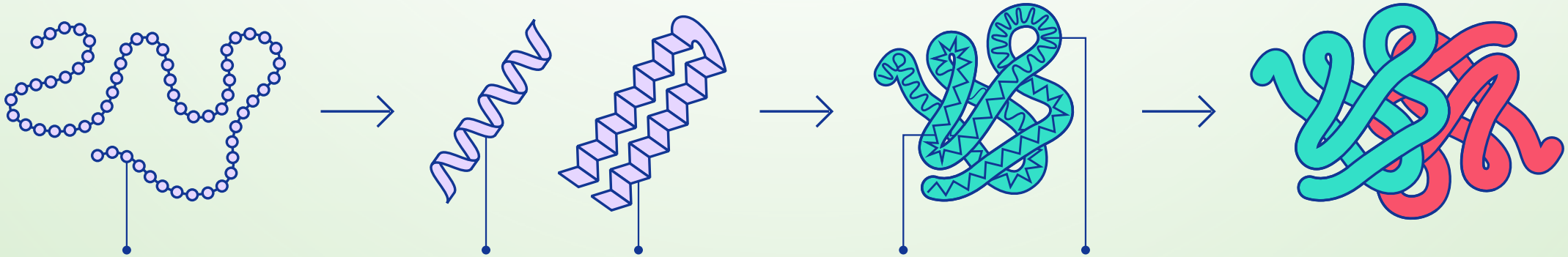


Every protein is made up of a sequence of amino acids bonded together

These amino acids interact locally to form shapes like helices and sheets

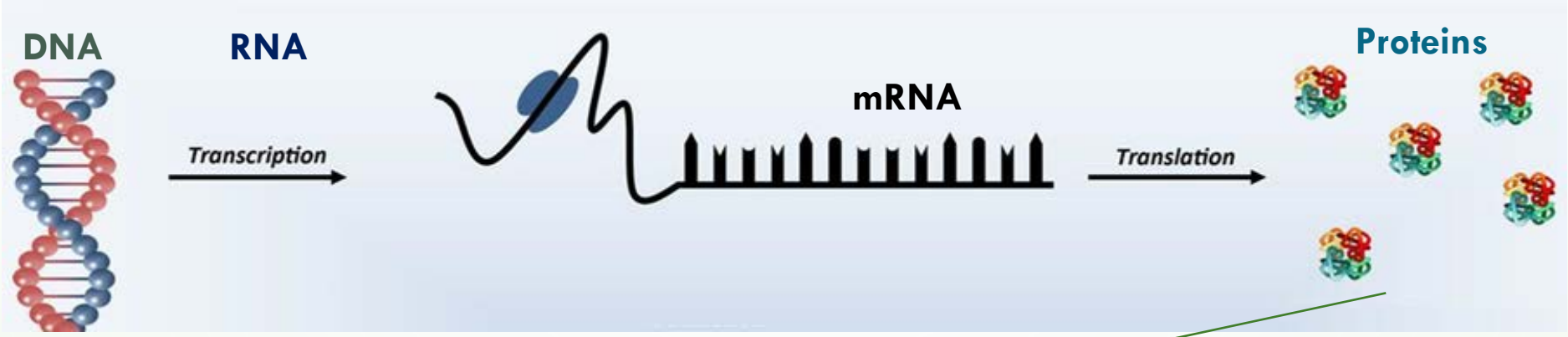
These shapes fold up on larger scales to form the full three-dimensional protein structure

Proteins can interact with other proteins, performing functions such as signalling and transcribing DNA



https://deepmind.com/blog/article/AlphaFold-Using-AI-for-scientific-discovery

DNA, RNA AND PROTEINS

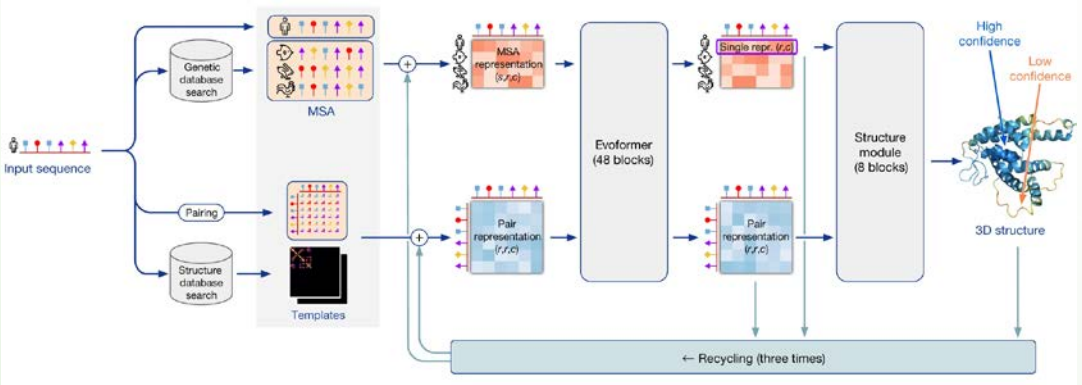


Protein 3D structure

Protein function

https://deepmind.com/blog/article/AlphaFold-Using-AI-for-scientific-discovery

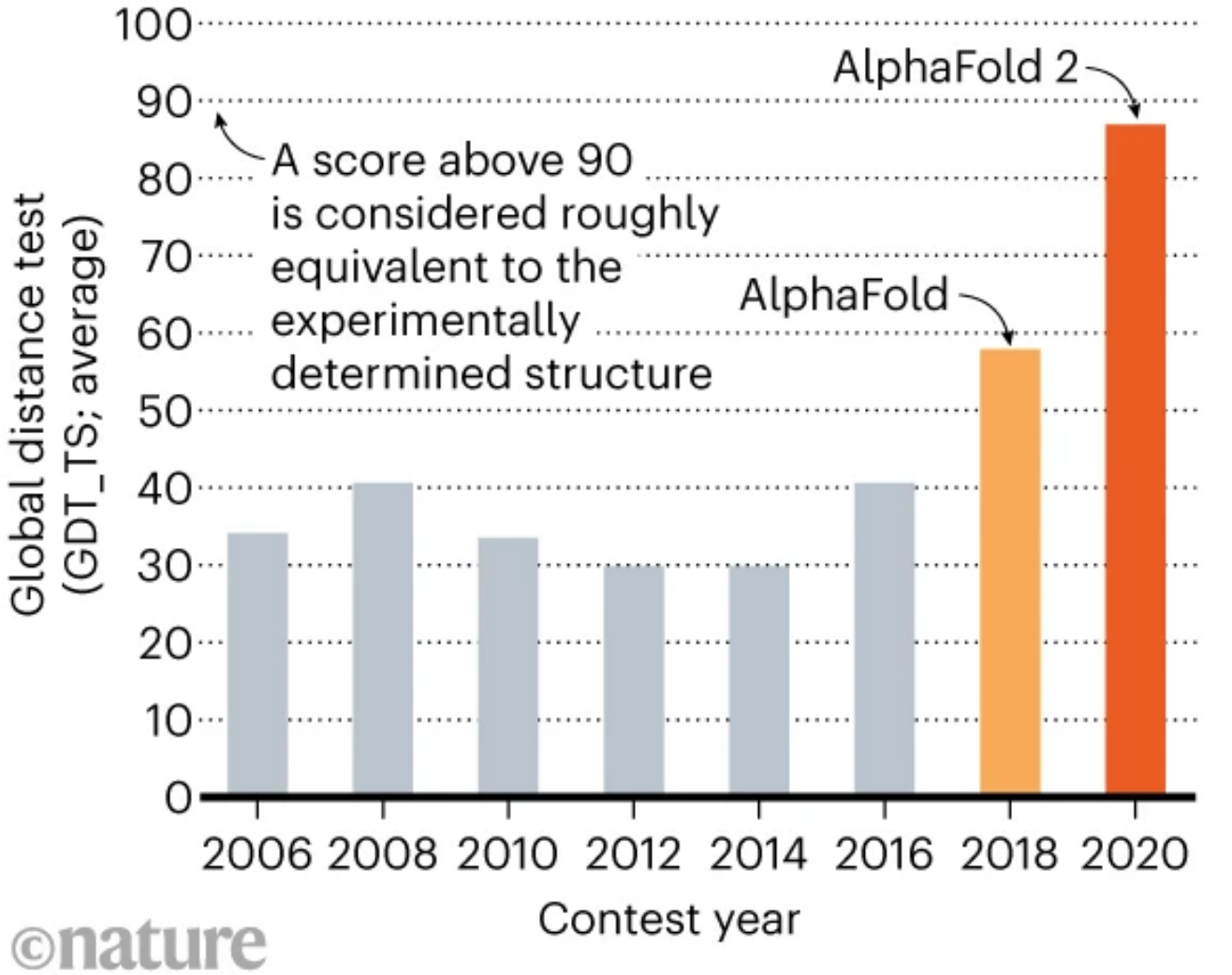
ALPHAFOLD: USING AI TO PREDICT PROTEIN STRUCTURE



AlphaFold
Protein Structure Database

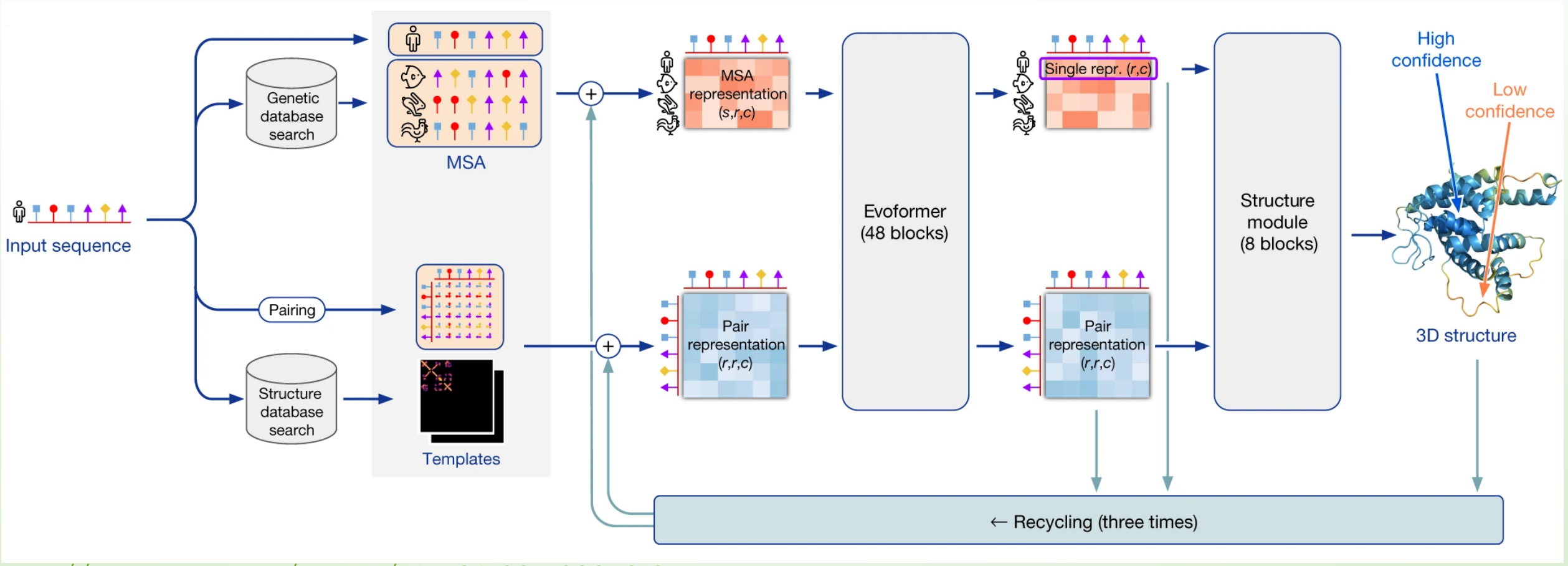
Developed by DeepMind and EMBL-EBI

https://alphafold.ebi.ac.uk/



USING AI TO PREDICT PROTEIN STRUCTURE

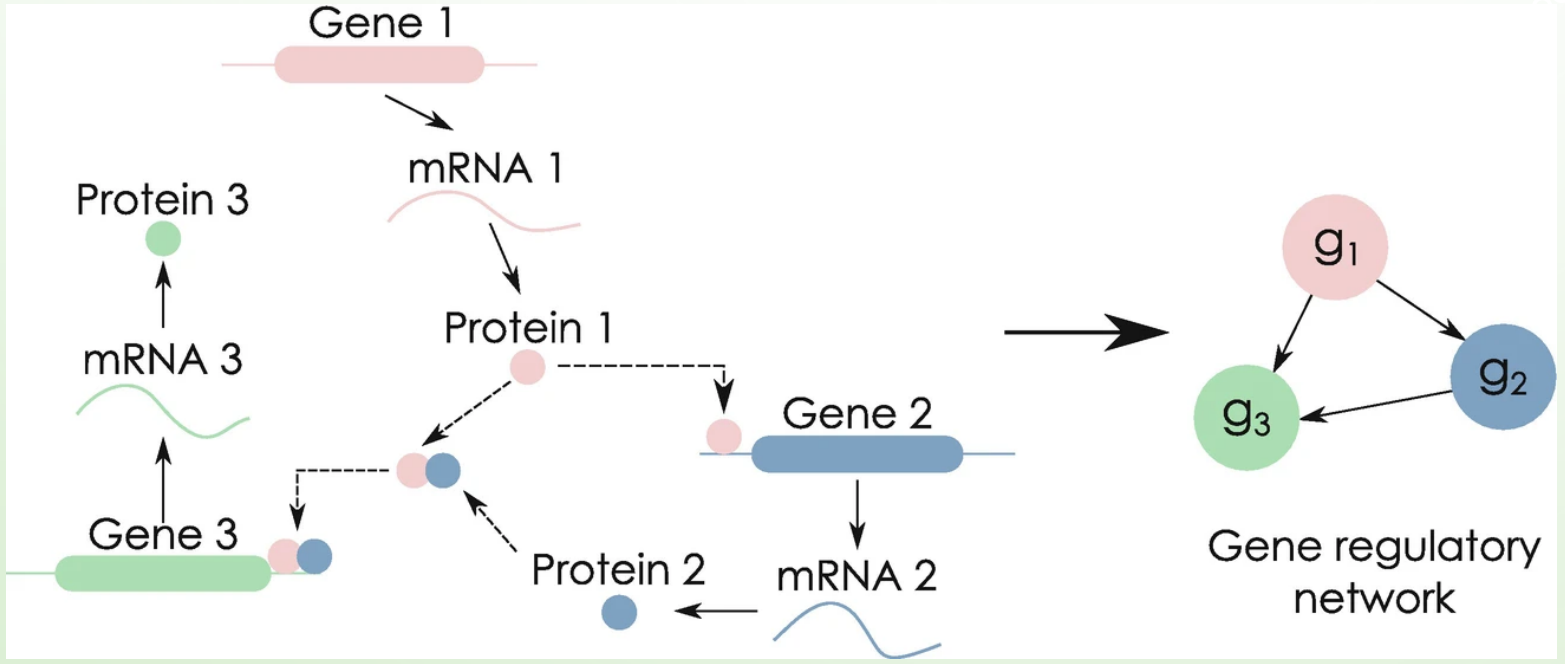
Can we determine a protein's 3D shape from its amino-acid sequence?

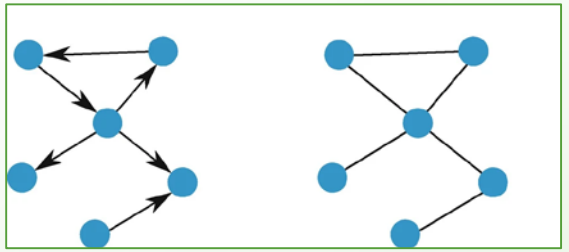


https://www.nature.com/articles/s41586-021-03819-2

REGULATION OF GENE EXPRESSION IN EUKARYOTES

At any given time, a complex set of interactions between genes, RNA molecules and proteins determine which genes are activated, and the amount of protein or RNA product.

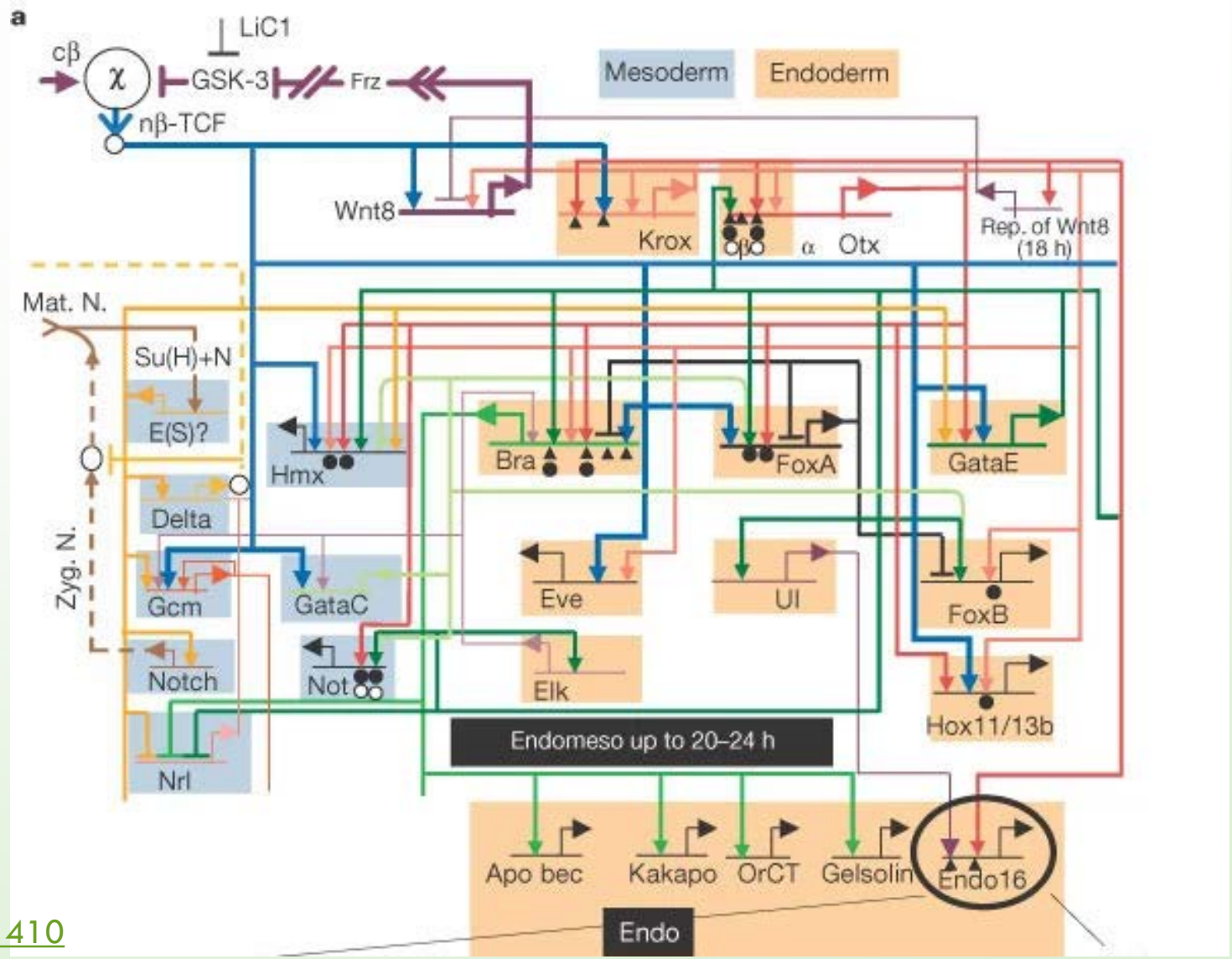




https://link.springer.com/protocol/10.1007/978-1-4939-8882-2_1

CODING NETWORK OF GENES

Gene regulatory networks that specify the behaviour of the genes can be represented computationally using methods from STEM

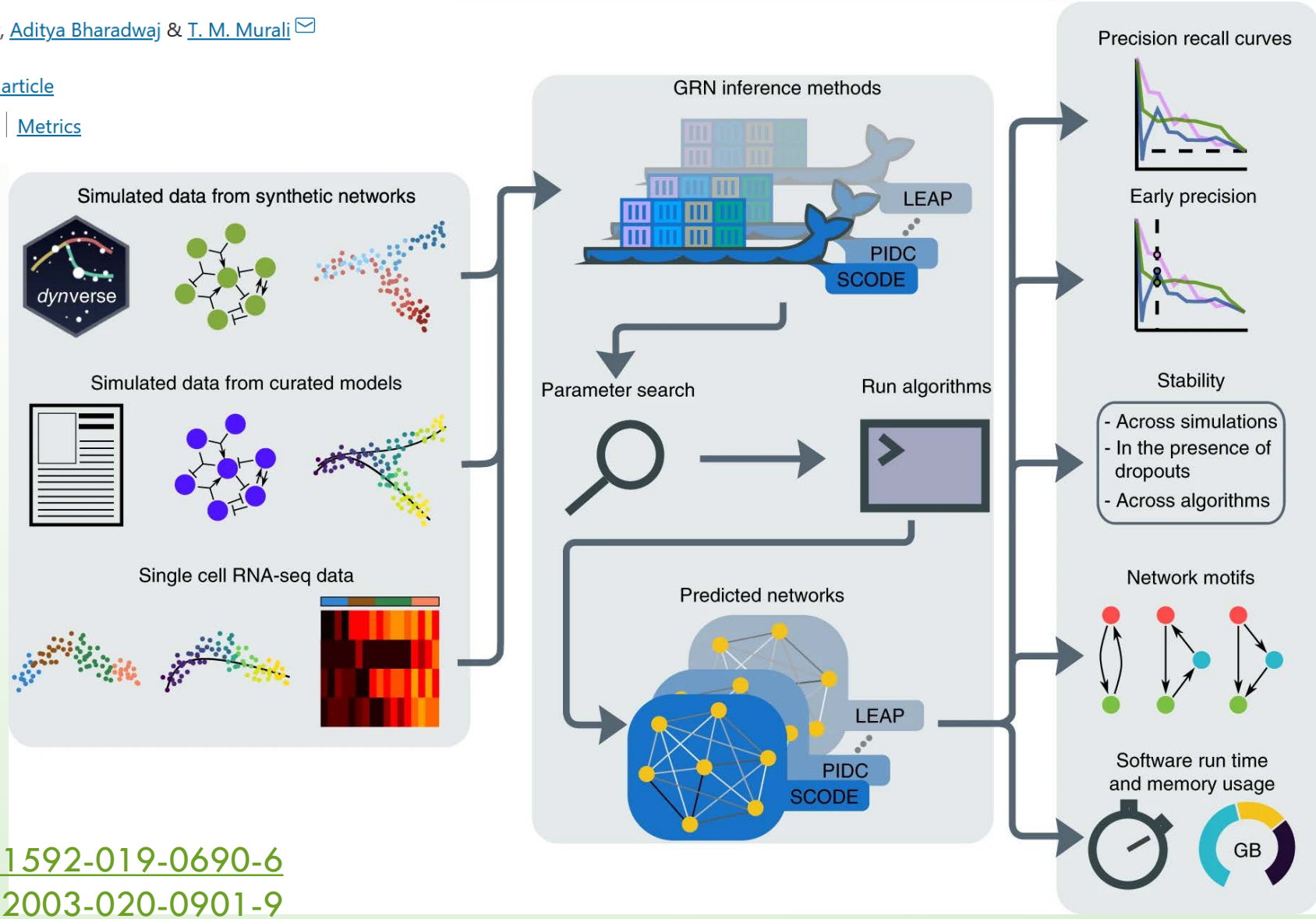


Hood and Galas, https://www.nature.com/articles/nature01410

Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data

Aditya Pratapa, Amogh P. Jalihal, Jeffrey N. Law, Aditya Bharadwaj & T. M. Murali

Nature Methods **17**, 147–154 (2020) | Cite this article



https://www.nature.com/articles/s41592-019-0690-6
https://www.nature.com/articles/s42003-020-0901-9