AI LAB 2023

1.2 – THE DATA

FRANCESCA M. BUFFA

LAB STRUCTURE

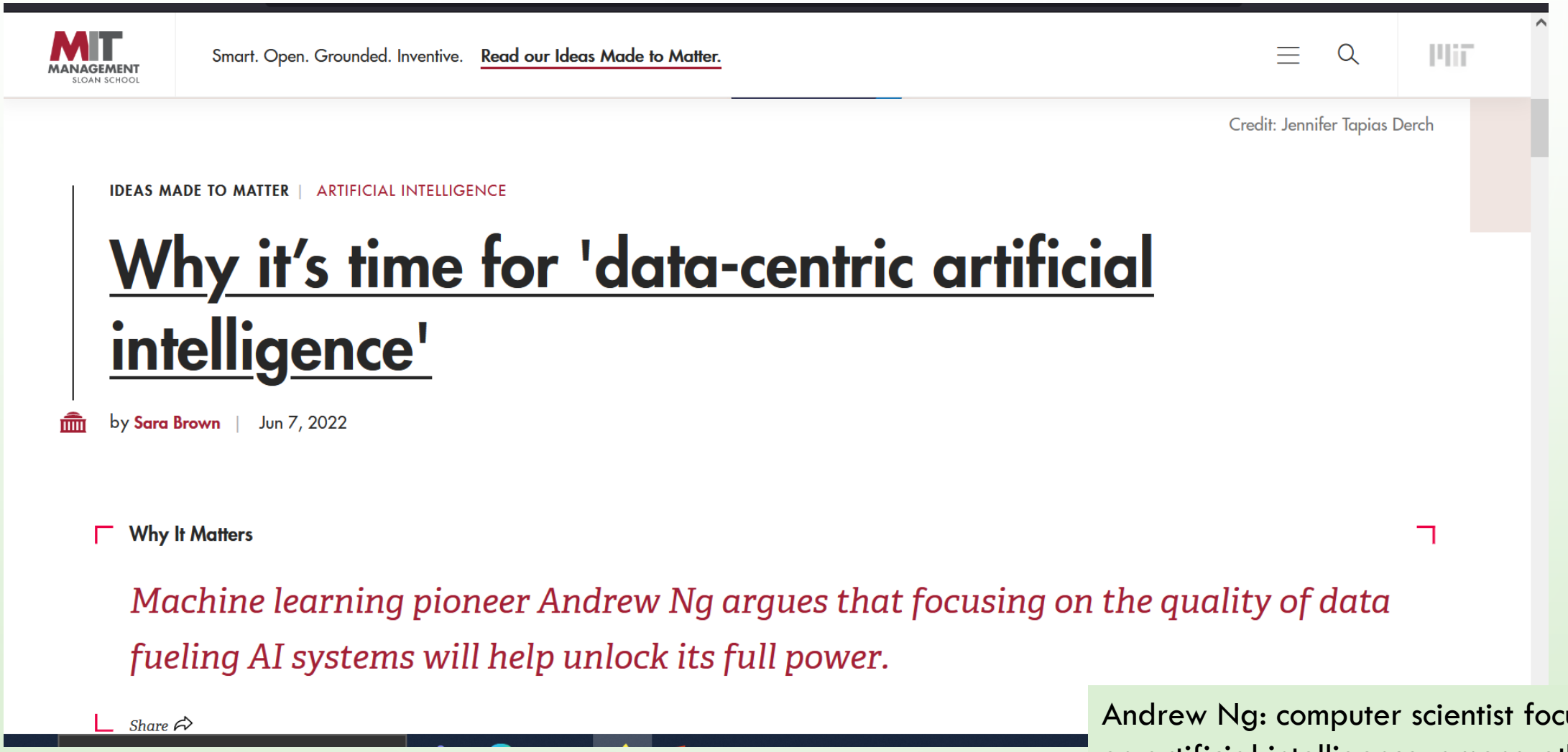
- INTRO AND BACKGROUND
- THE DATA
- THE AI-LAB CHALLENGE – PART 1
- PART 1 - SHARING AND DISCUSSION

- UNSUPERVISED LEARNING EXAMPLES
- THE AI-LAB CHALLENGE – PART 2
- PART 2 - SHARING AND DISCUSSION

- SUPERVISED LEARNING EXAMPLES
- THE AI-LAB CHALLENGE – PART 3
- LARGE PROJECTS AND DATABASES

- THE AI-LAB CHALLENGE PARTS 1-3, SHARING AND DISCUSSION
- DATA INTERPRETATION
- DISCUSS AND PREPARE WORKSHOP PRESENTATIONS

THE DATA



MIT MANAGEMENT SLOAN SCHOOL

Smart. Open. Grounded. Inventive. Read our Ideas Made to Matter.

Credit: Jennifer Tapias Derch

IDEAS MADE TO MATTER | ARTIFICIAL INTELLIGENCE

Why it's time for 'data-centric artificial intelligence'

by Sara Brown | Jun 7, 2022

Why It Matters

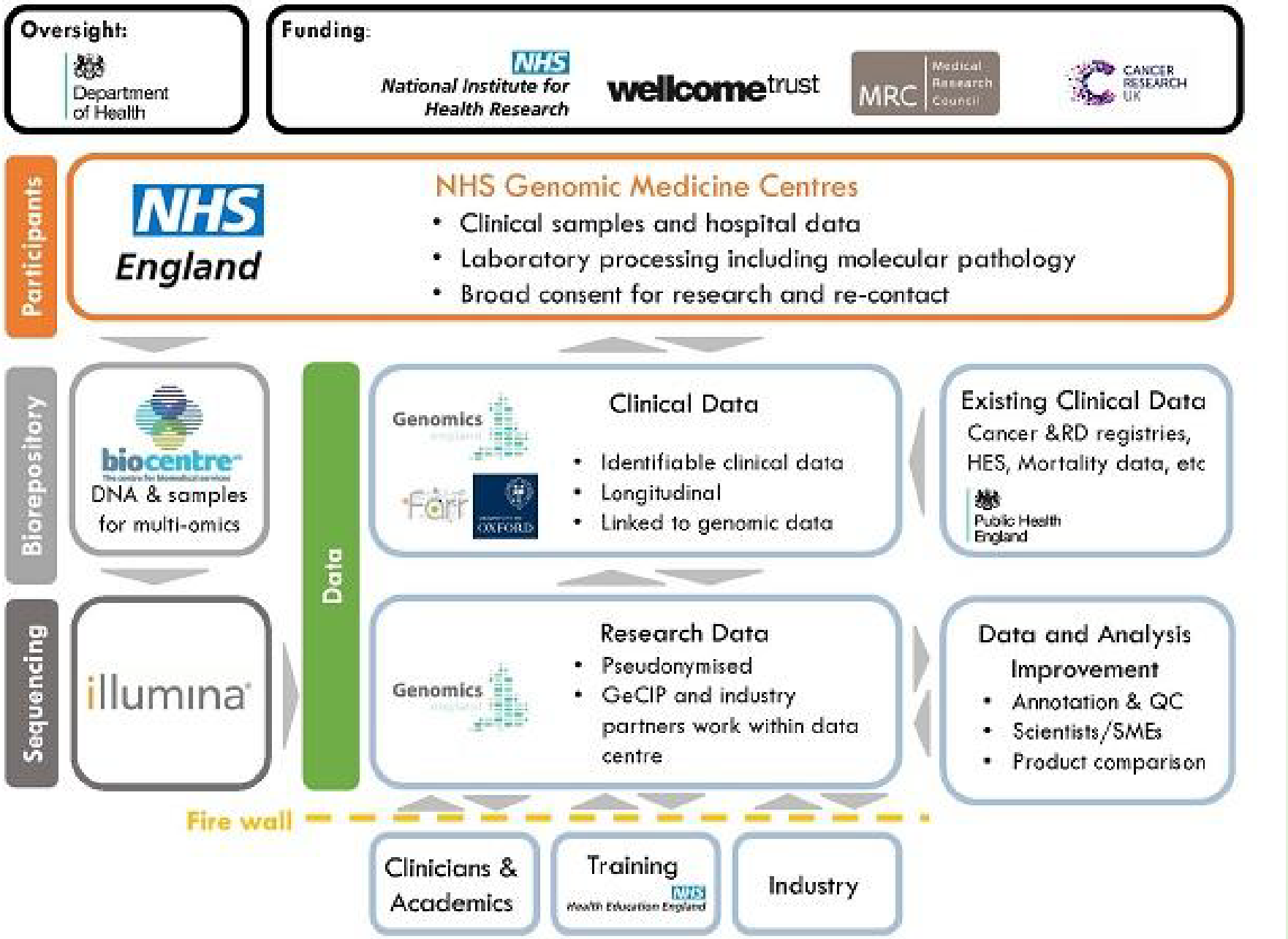
Machine learning pioneer Andrew Ng argues that focusing on the quality of data fueling AI systems will help unlock its full power.

Share

Andrew Ng: computer scientist focusing on artificial intelligence, among others cofounder and head of Google Brain.

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

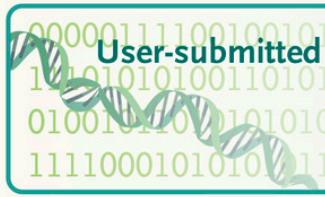
Genomic and Clinical Data Sources

The Cancer Genome Atlas (TCGA)

TARGET (Therapeutically Applicable Research to Generate Effective Treatments)

International Cancer Genome Consortium

NCI clinical trials



User-submitted studies

Genomic Data Commons (GDC)

1. Import and standardize genomic and clinical data from legacy and current NCI programs
2. Harmonize mapping of sequence data to the most current genome and transcriptome build
3. Implement state-of-the-art methods for derived data:
   - Mutation calls
   - Copy number
   - Structural variants
   - Digital gene expression

4. Maintain data security and manage authorized access
5. Provide data for browsing, download, or analysis on a colocalized computer cluster
6. Open GDC for upload of new cancer genomic data from researchers worldwide for comparison with existing data and sharing

GDC Use Cases

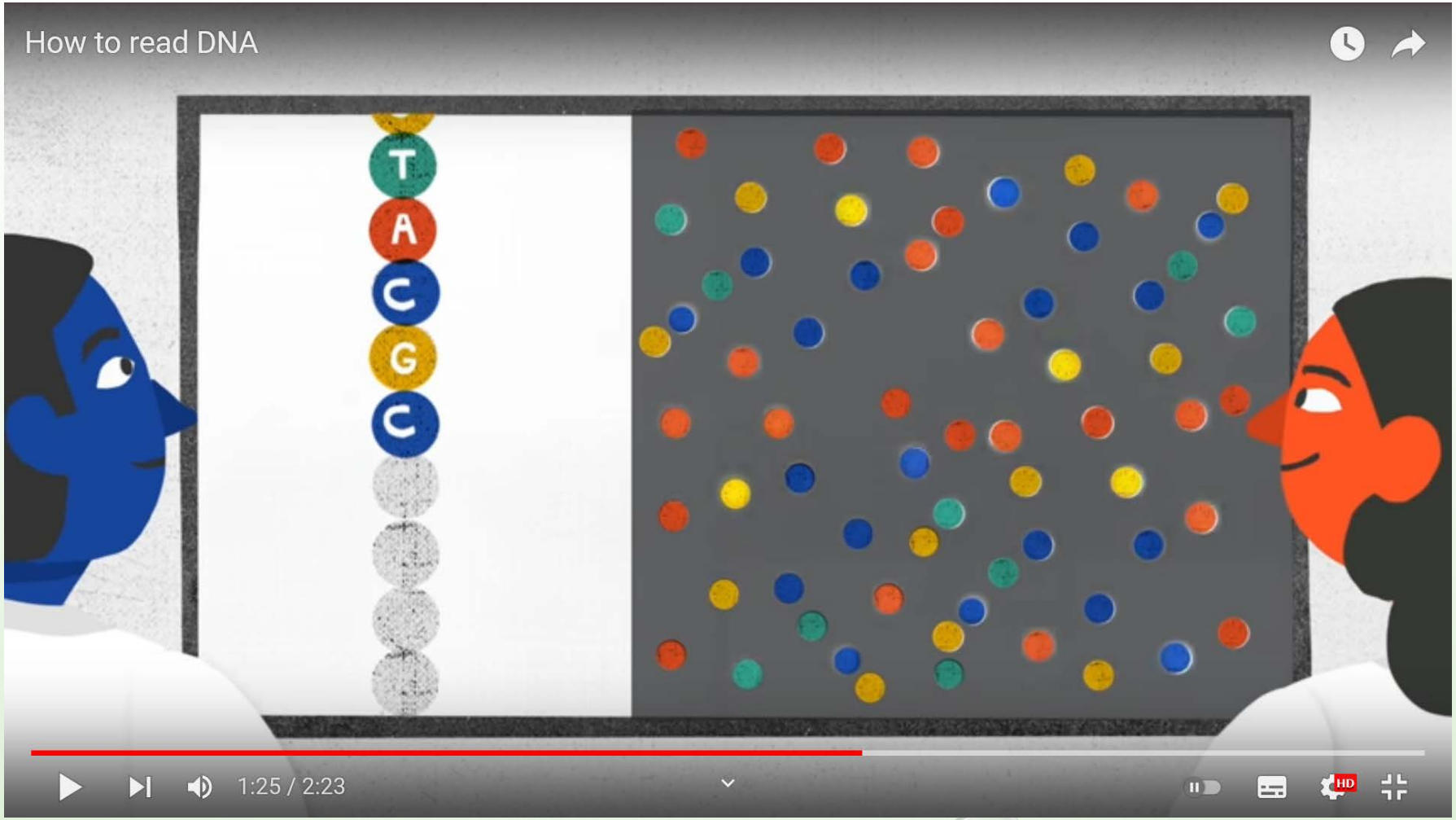
Identify low-frequency cancer drivers

Define genomic determinants of treatment response

Compose clinical trial cohorts sharing targetable genetic lesions

https://www.nejm.org/doi/full/10.1056/nejmp1607591

DNA: WE CAN READ IT AND WRITE IT



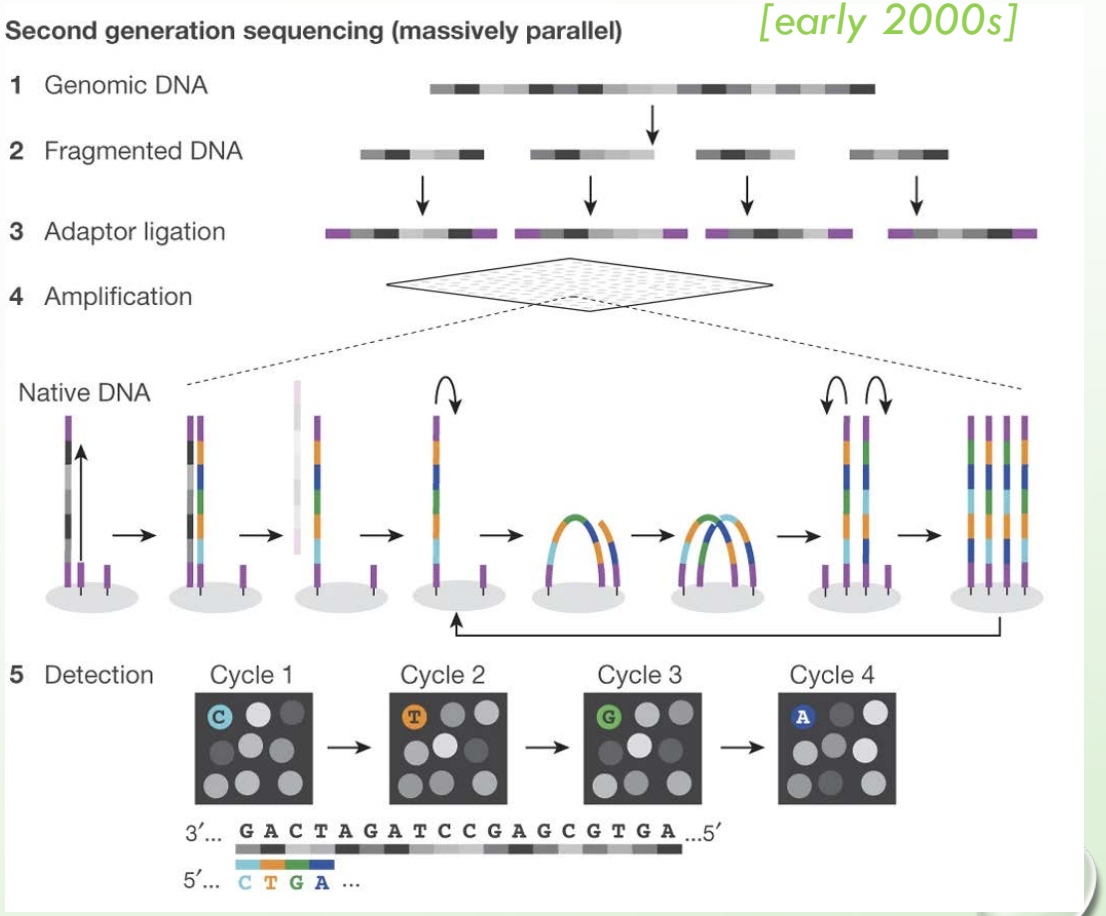
OxfordSparks

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

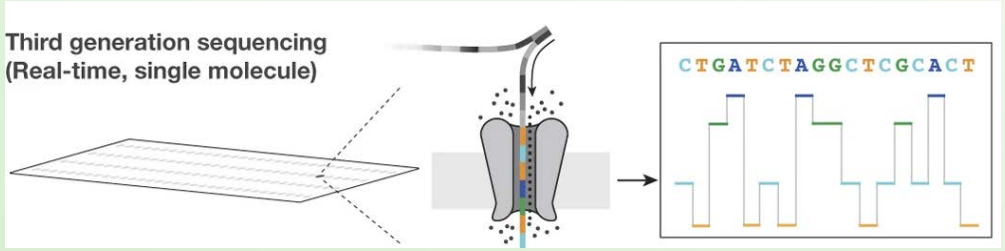
DNA SEQUENCING TECHNOLOGIES

[1977]

[early 2000s]



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

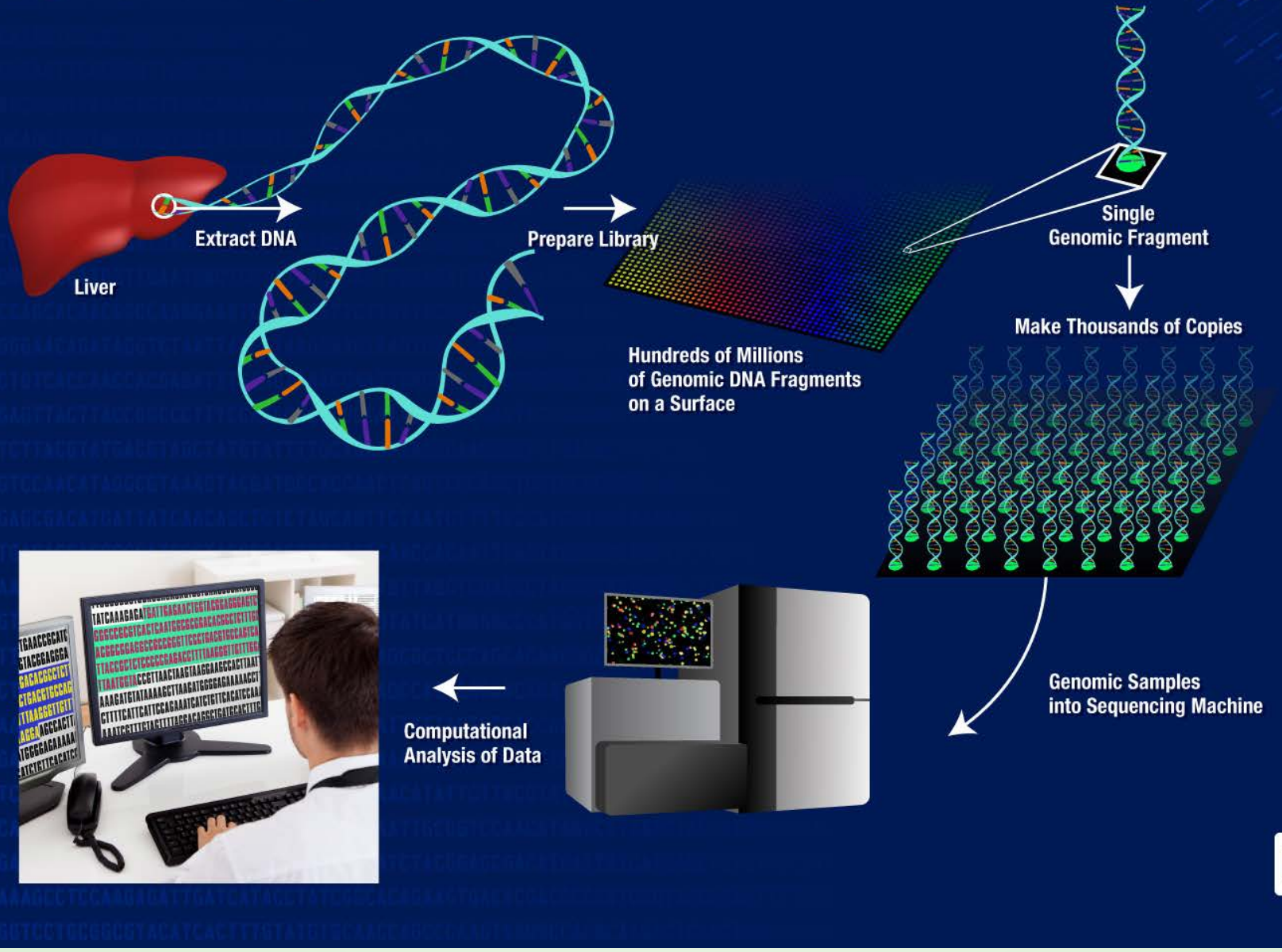


https://www.nature.com/articles/nature24286/

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

Dna Sequencing

NHGRI FACT SHEETS
genome.gov



NIH National Human Genome Research Institute

https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Fact-Sheet

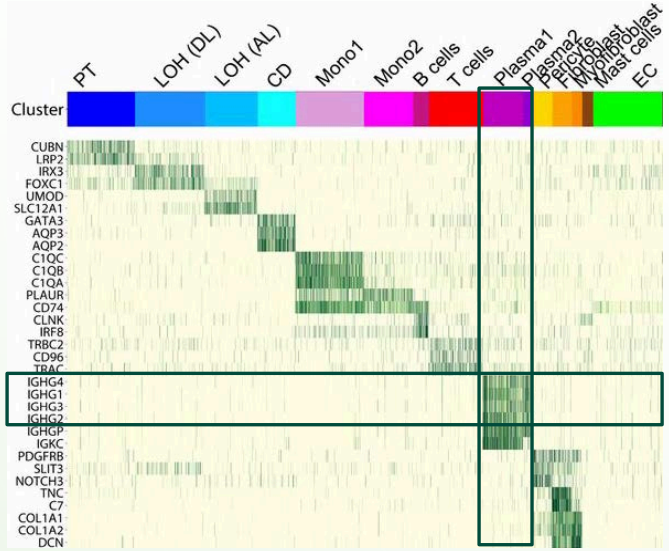
DNA/RNA STRING REPRESENTATION

```
Sequence: GTTCTCTTCTTCCCTAGCGGATAACAGATTCGGCCCACTGTCCCCTTCAG

[1] + Sequence Length: 50

[2] + Nucleotide Frequency: {'A': 8, 'C': 18, 'G': 9, 'T': 15}

[3] + DNA/RNA Transcription: GUUCUCUUCUUCCCUAGCGGAUAACAGAUUCGGCCCACUGUCCCCUUCAG

[4] + DNA String + Complement + Reverse Complement:
5' GTTCTCTTCTTCCCTAGCGGATAACAGATTCGGCCCACTGTCCCCTTCAG 3'
   ||||||||||||||||||||||||||||||||||||||||||||||||||
3' CAAGAGAAGAAGGGATCGCCTATTGTCTAAGCCGGGTGACAGGGGAAGTC 5' [Complement]
5' CTGAAGGGGACAGTGGGCCGAATCTGTTATCCGCTAGGGAAGAAGAGAAC 3' [Rev. Complement]
```

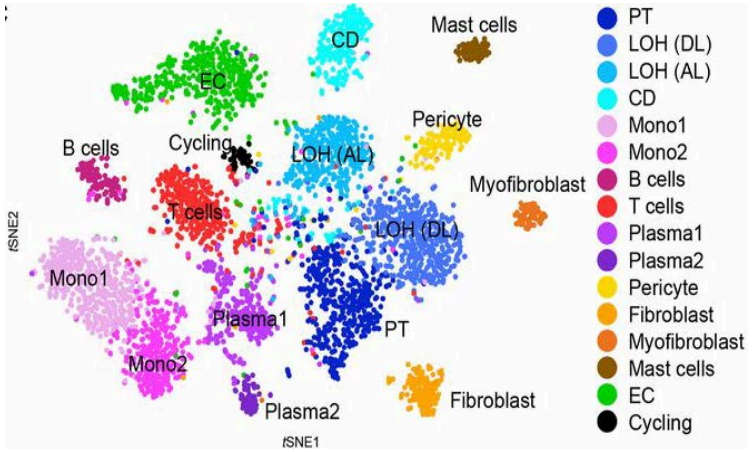
https://python.plainenglish.io/bioinformatics-in-python-dna-toolkit-part-2-transcription-and-reverse-complement-1ed194e0d06e

We can sequence the DNA and RNA of single cells

Identifying cell-type marker genes



Discovering sample heterogeneity



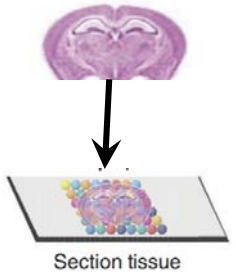
Uncovering tissue dynamics

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

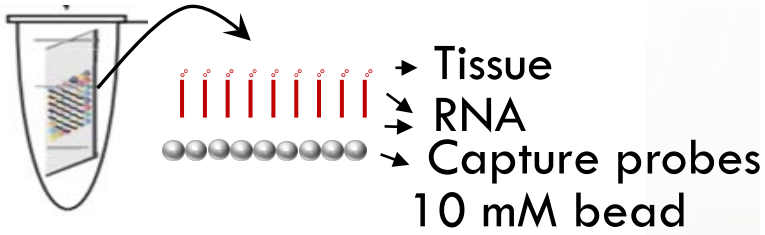
We can acquire a spatial image of single-cell expression

Spatial slide

Tissue overlay

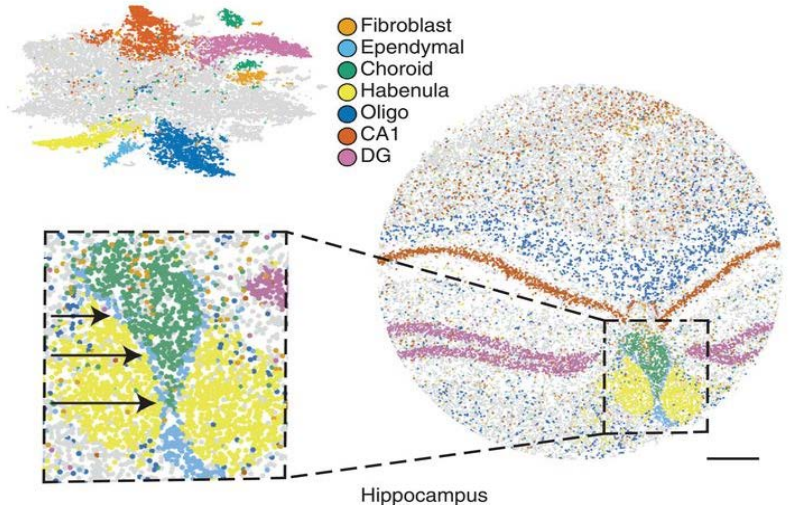


Tissue permeation

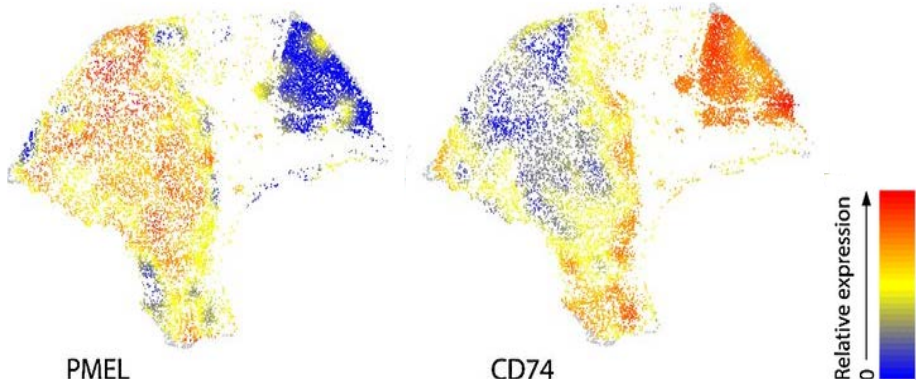


Library preparation

Spatial heatmaps of cell populations

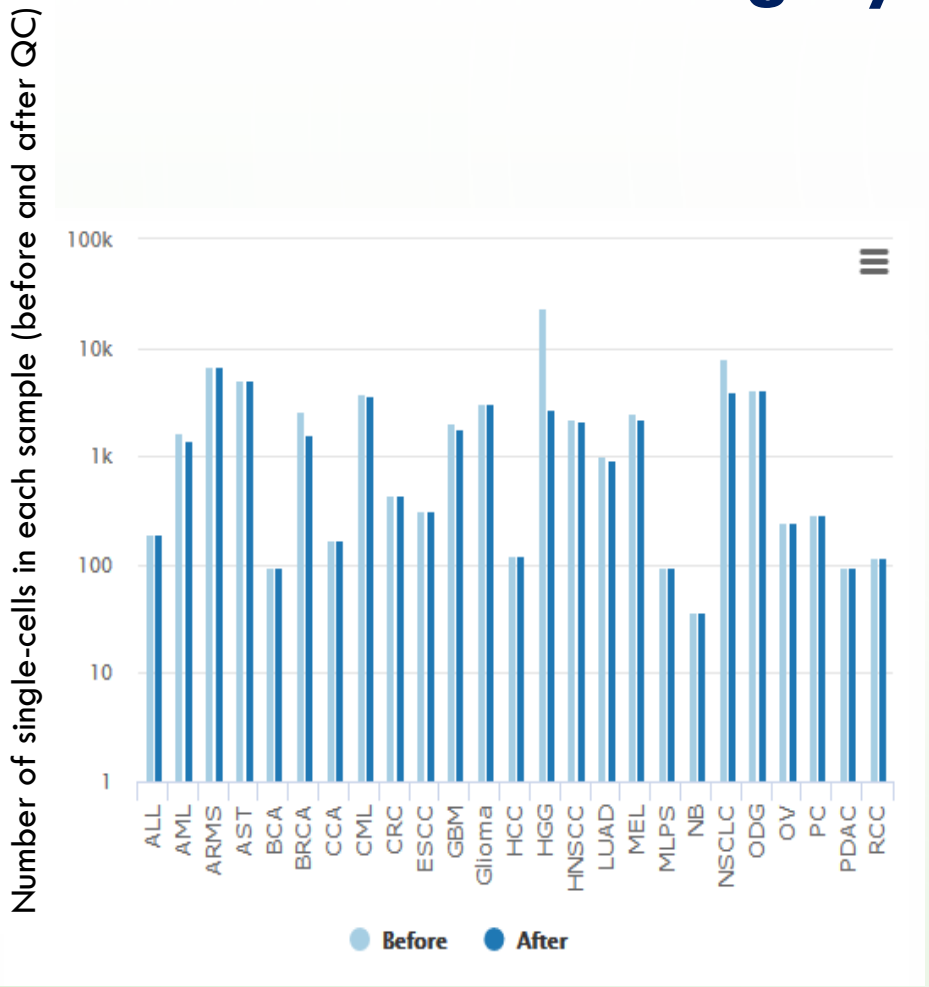


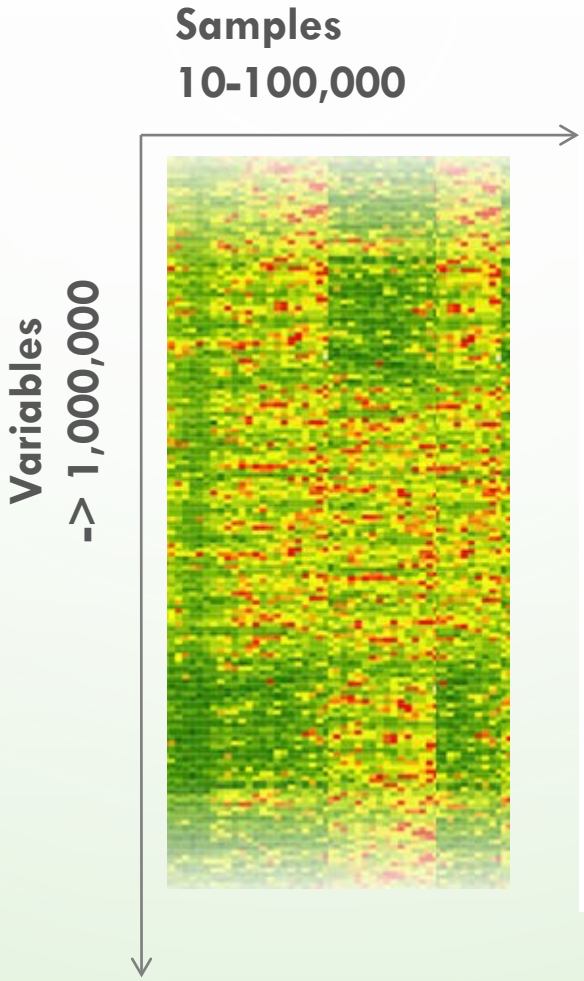
Spatial heatmaps of selected genes

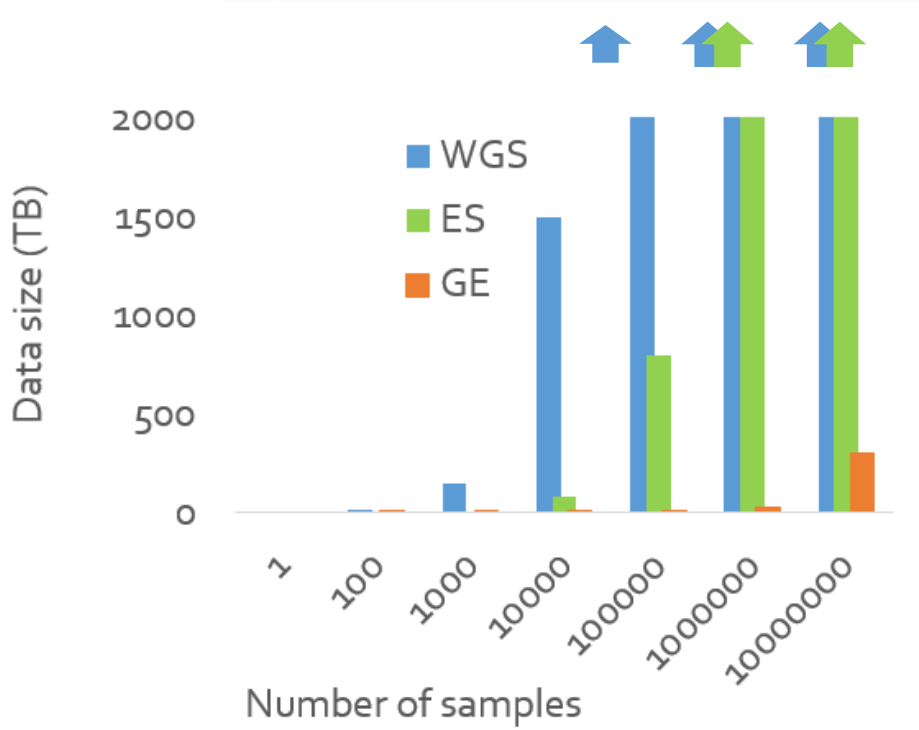


Images from: Vickovic et al (2019) Nature Methods, Thrane (2018) Cancer Res and Rodriques et al (2019) Science

Highly multidimentional datasets







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

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

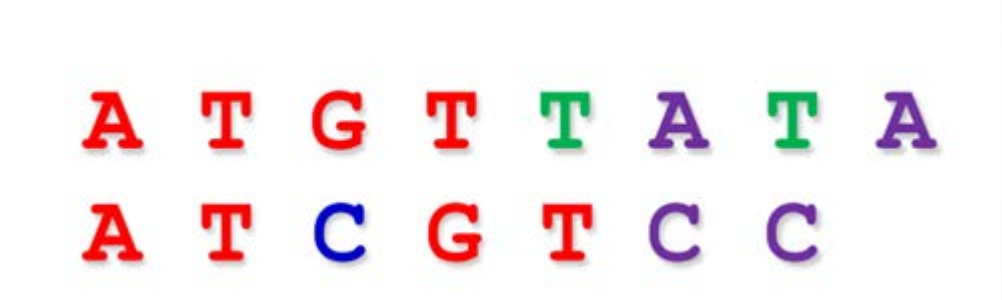
Base Calling
(Vendor tool, non-text)

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

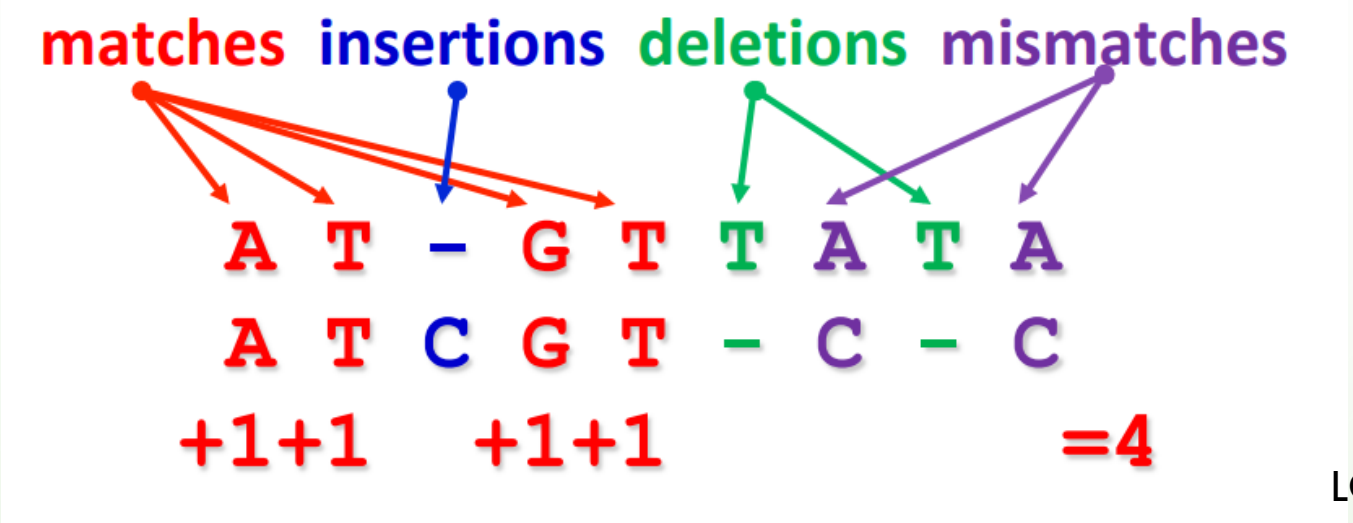


Finding regions of similarities and dissimilarities between sequences, and infer a measure of relatedness, is vital for phylogenetic research:

Comparison of genetic material across organisms allows for example to:

- Infer functional relationships
- Infer evolutionary relationships

ALIGNMENT: SEQUENCE MATCHING PROBLEM



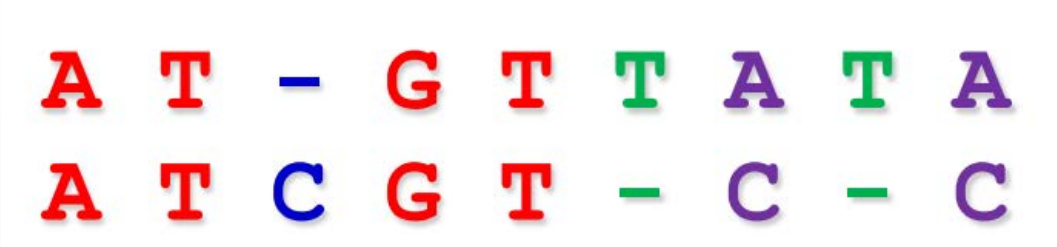
LONGEST COMMON SEQUENCE

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

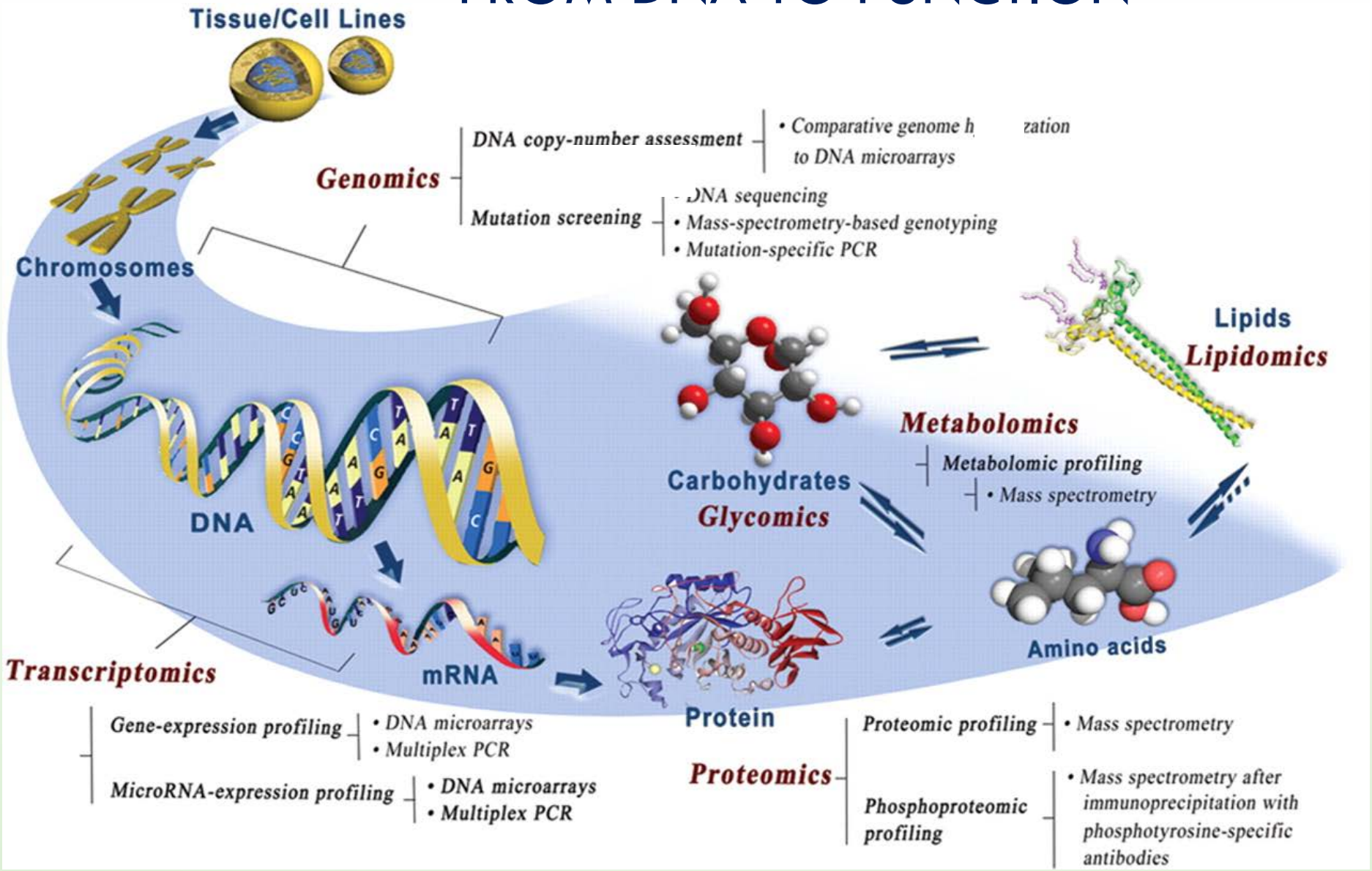
```
@HD     VN:1.0  SO:coordinate
@SQ     SN:chr20        LN:64444167
@PG     ID:TopHat       VN:2.0.14       CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-rea
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr
20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq
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

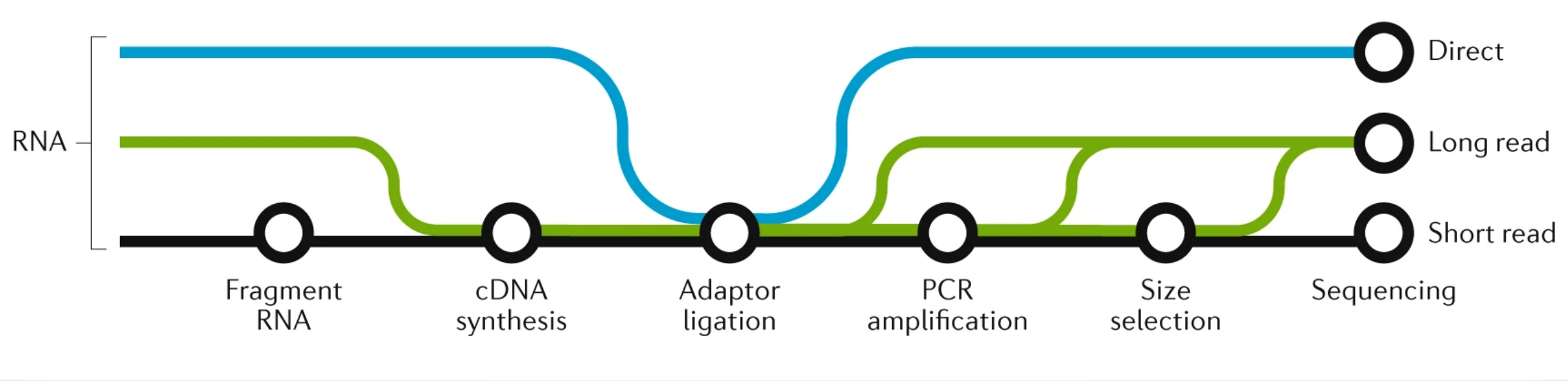
Binary Alignment/Map (BAM) file (*.bam) is a compressed binary SAM file (smaller size + faster access)

FROM DNA TO FUNCTION



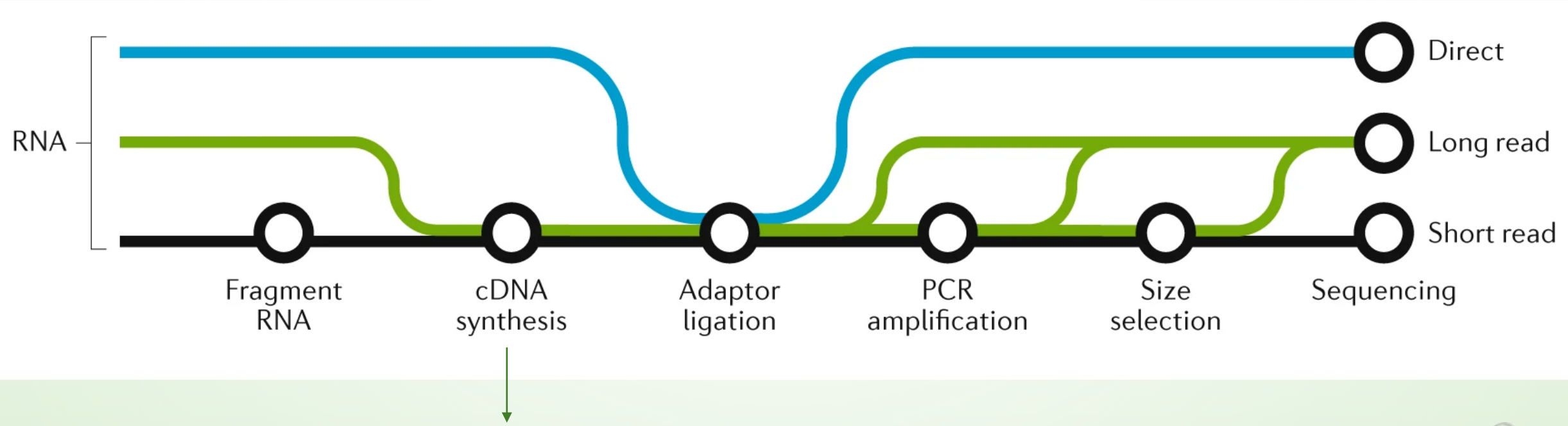
Adapted from: Wu et al. JDR 2011

RNA SEQUENCING



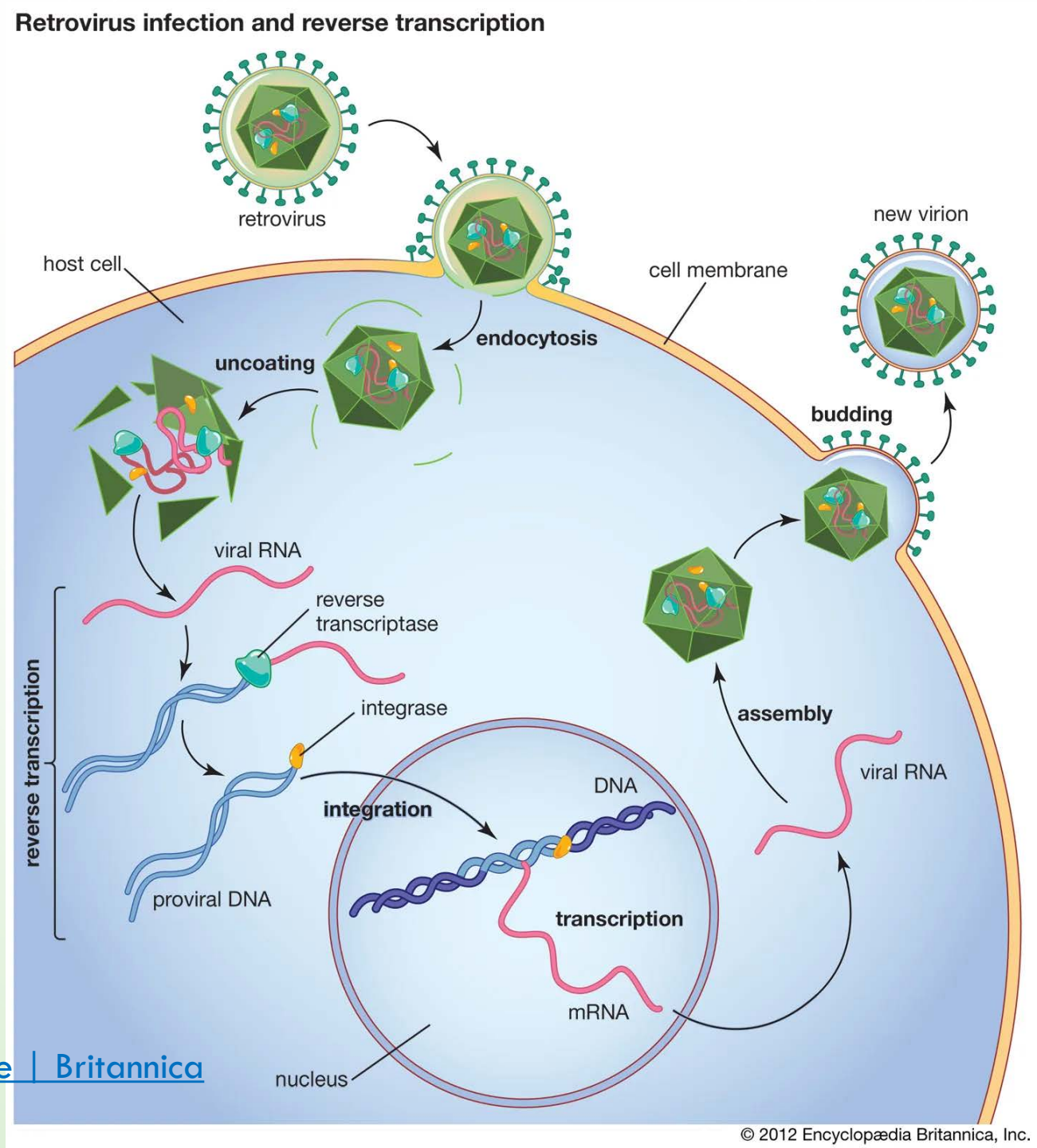
https://www.nature.com/articles/s41576-019-0150-2/figures/1

RNA SEQUENCING



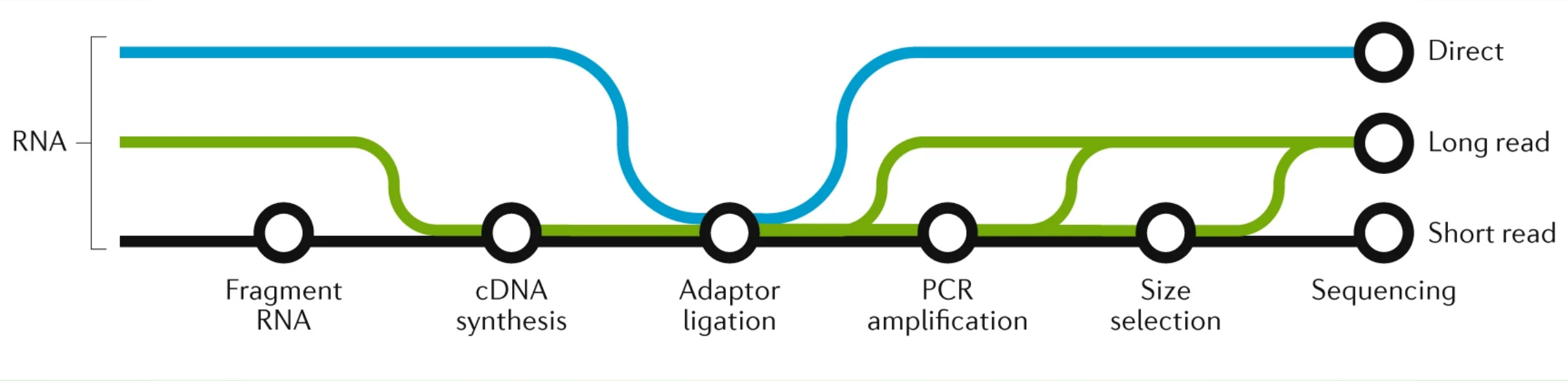
Complementary DNA (cDNA) is generated via reverse transcription: mRNA is the template and DNA is the product

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



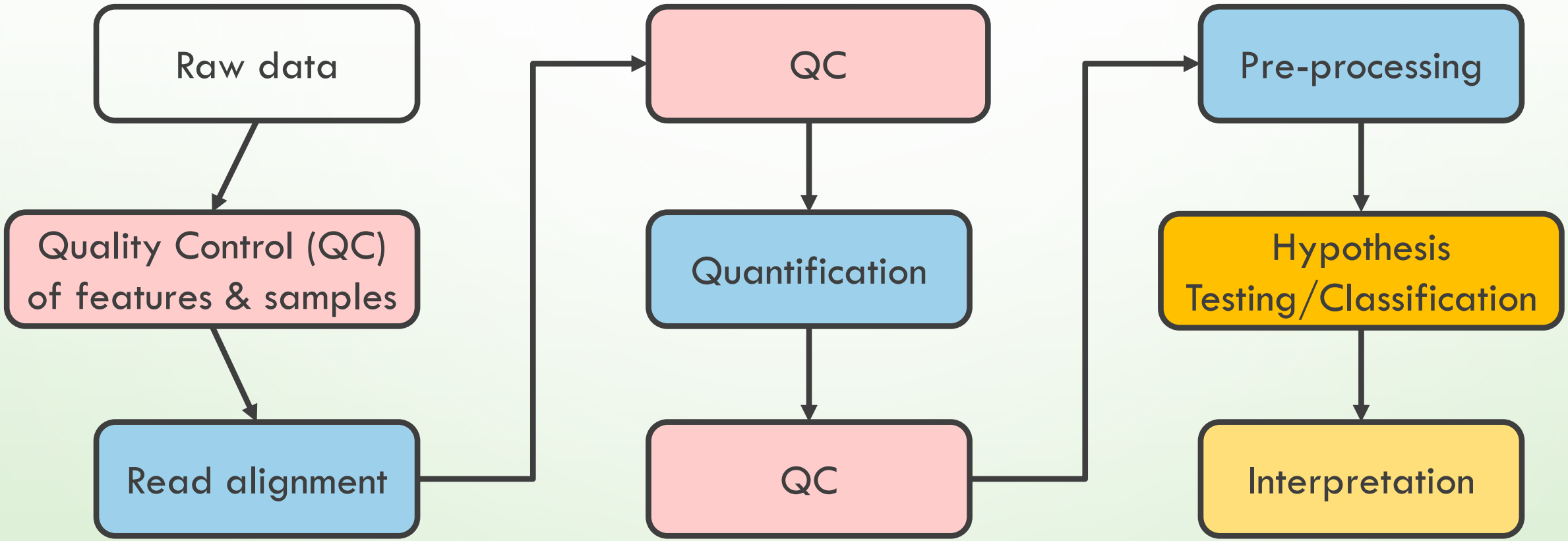
Encyclopedia Britannica
reverse transcriptase | enzyme | Britannica

RNA SEQUENCING

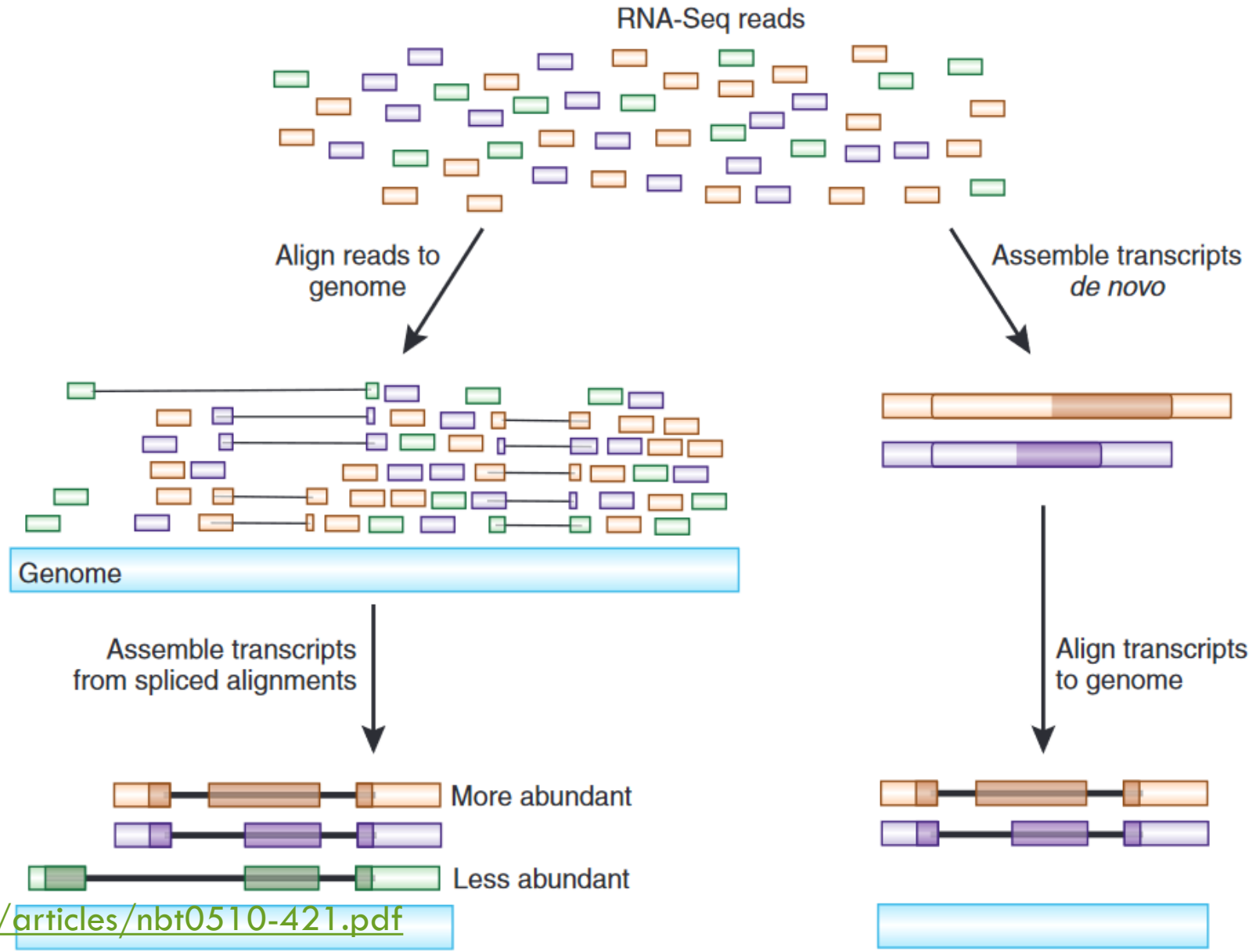


https://www.nature.com/articles/s41576-019-0150-2/figures/1

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

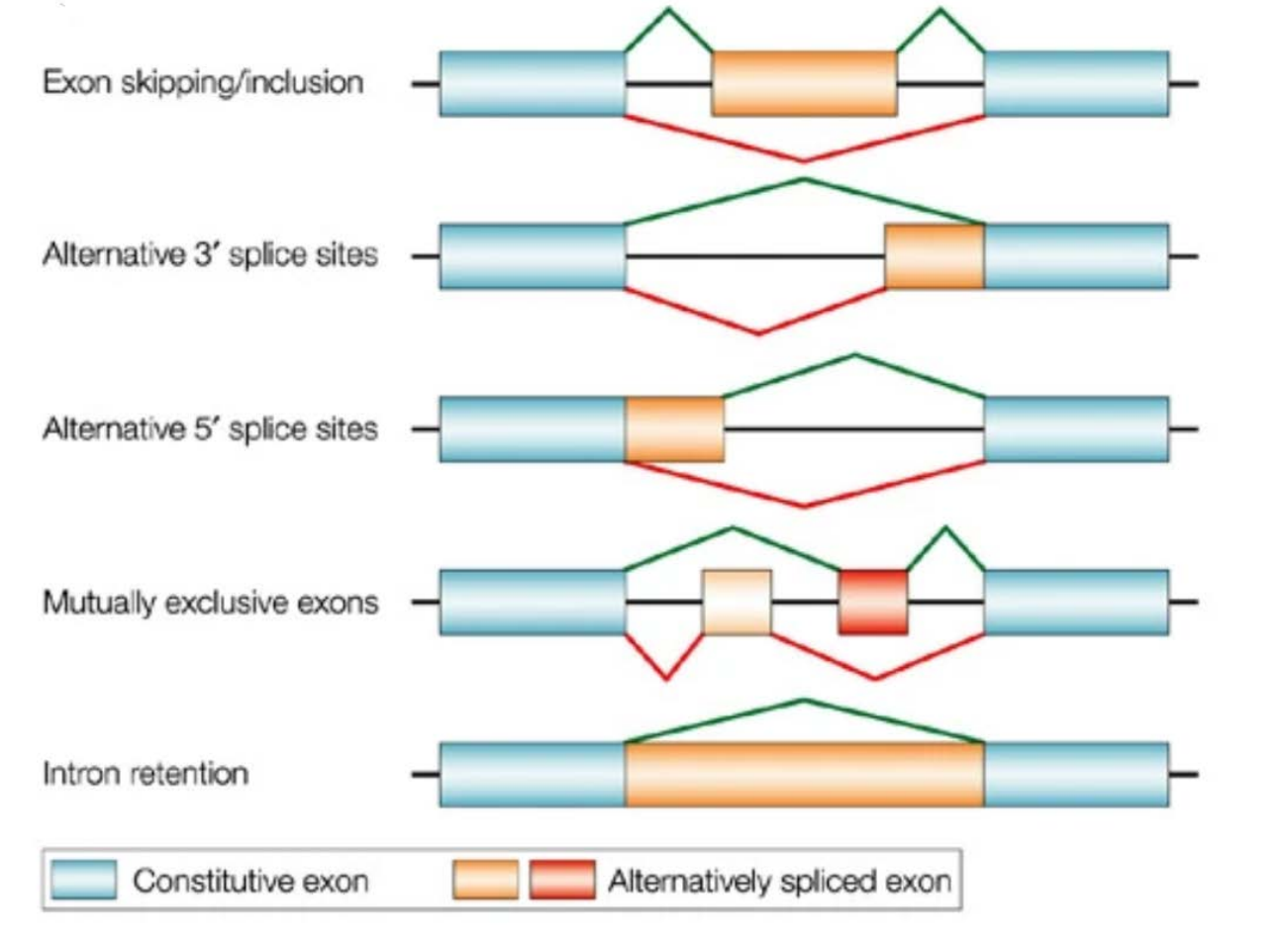
- For eukaryotic genomes this completely fails!

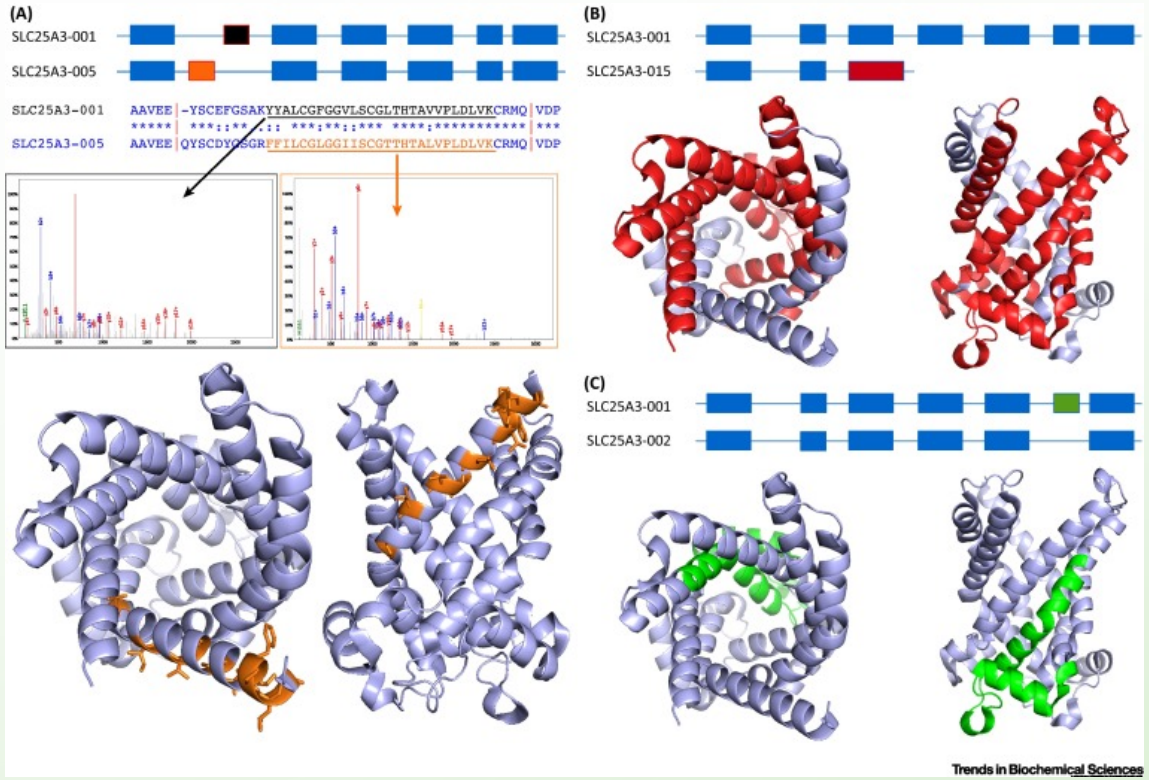
	Organism	# of protein-coding genes	# of genes naïve estimate: (genome size /1000)
viruses	HIV 1	9	10
	Influenza A virus	10-11	14
	Bacteriophage λ	66	49
	Epstein Barr virus	80	170
prokaryotes	*Buchnera sp.*	610	640
	T. maritima	1,900	1,900
	S. aureus	2,700	2,900
	V. cholerae	3,900	4,000
	B. subtilis	4,400	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



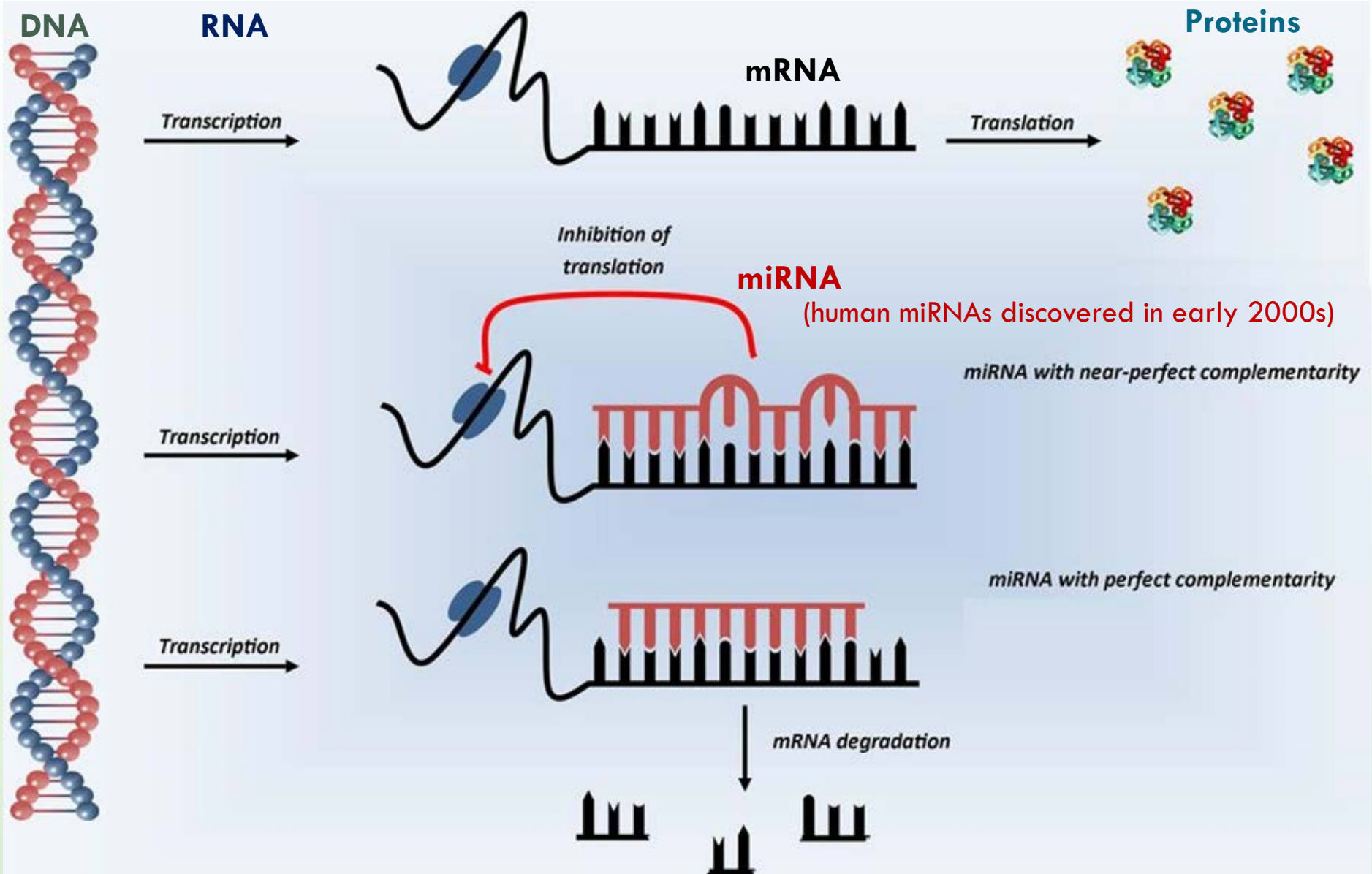


https://www.nature.com/articles/nrg775

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

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

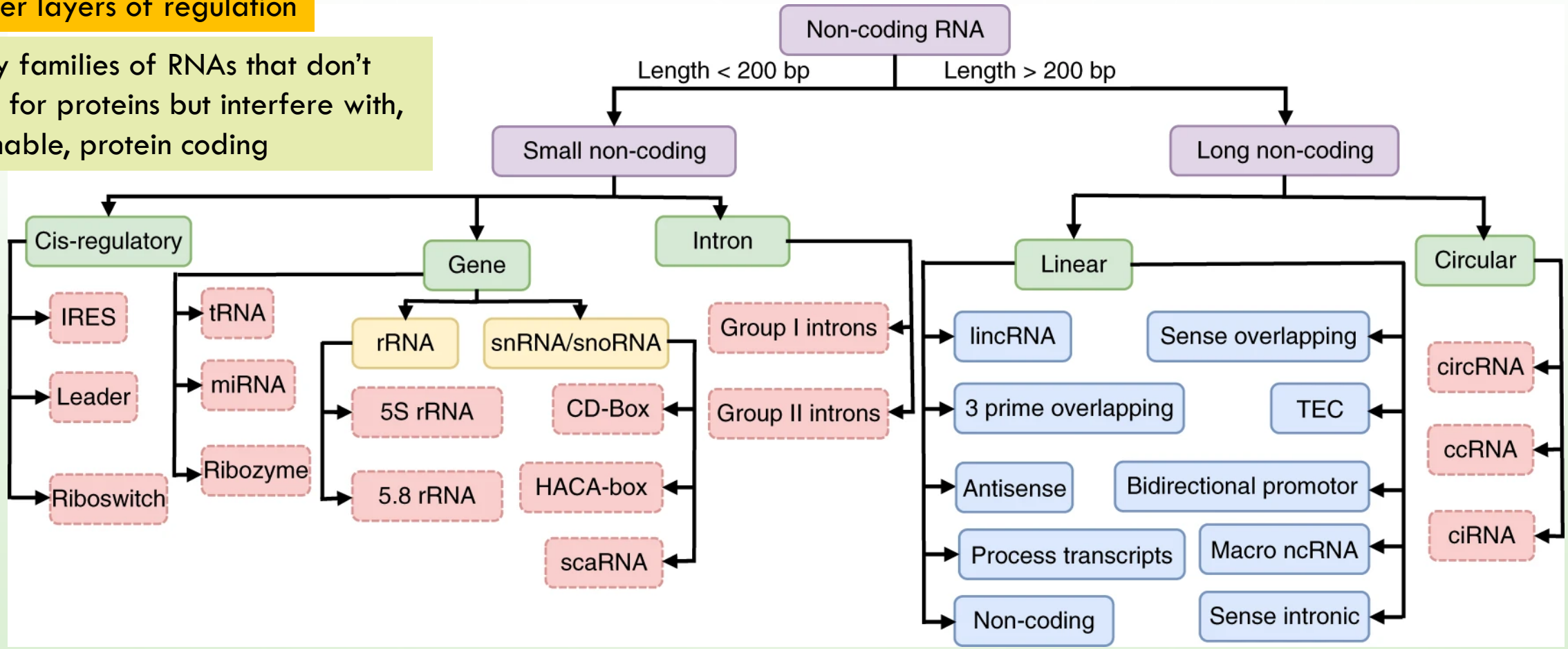
Further layers of regulation



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

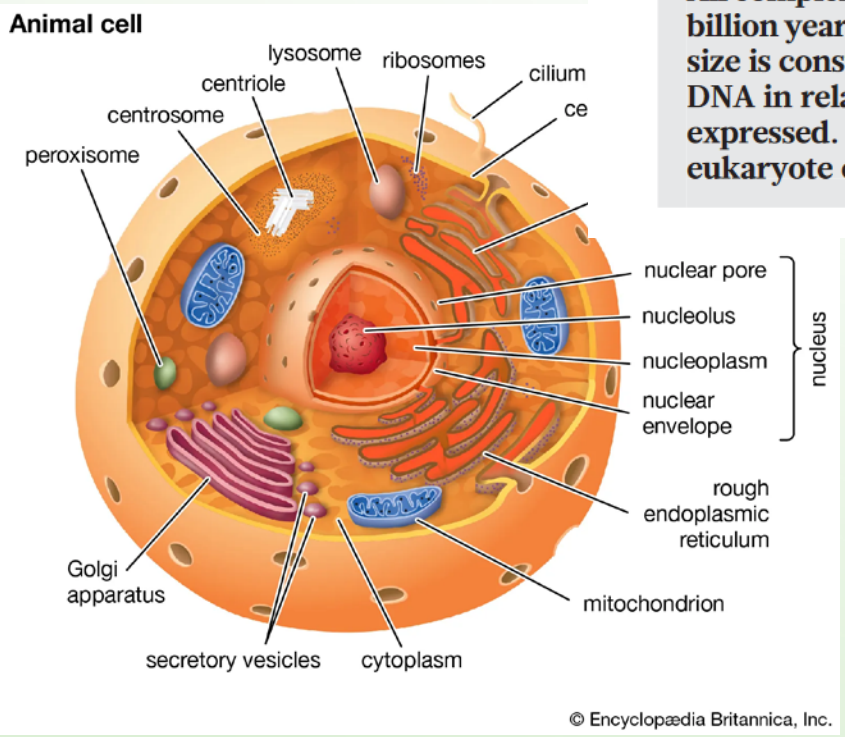
Further layers of regulation

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?



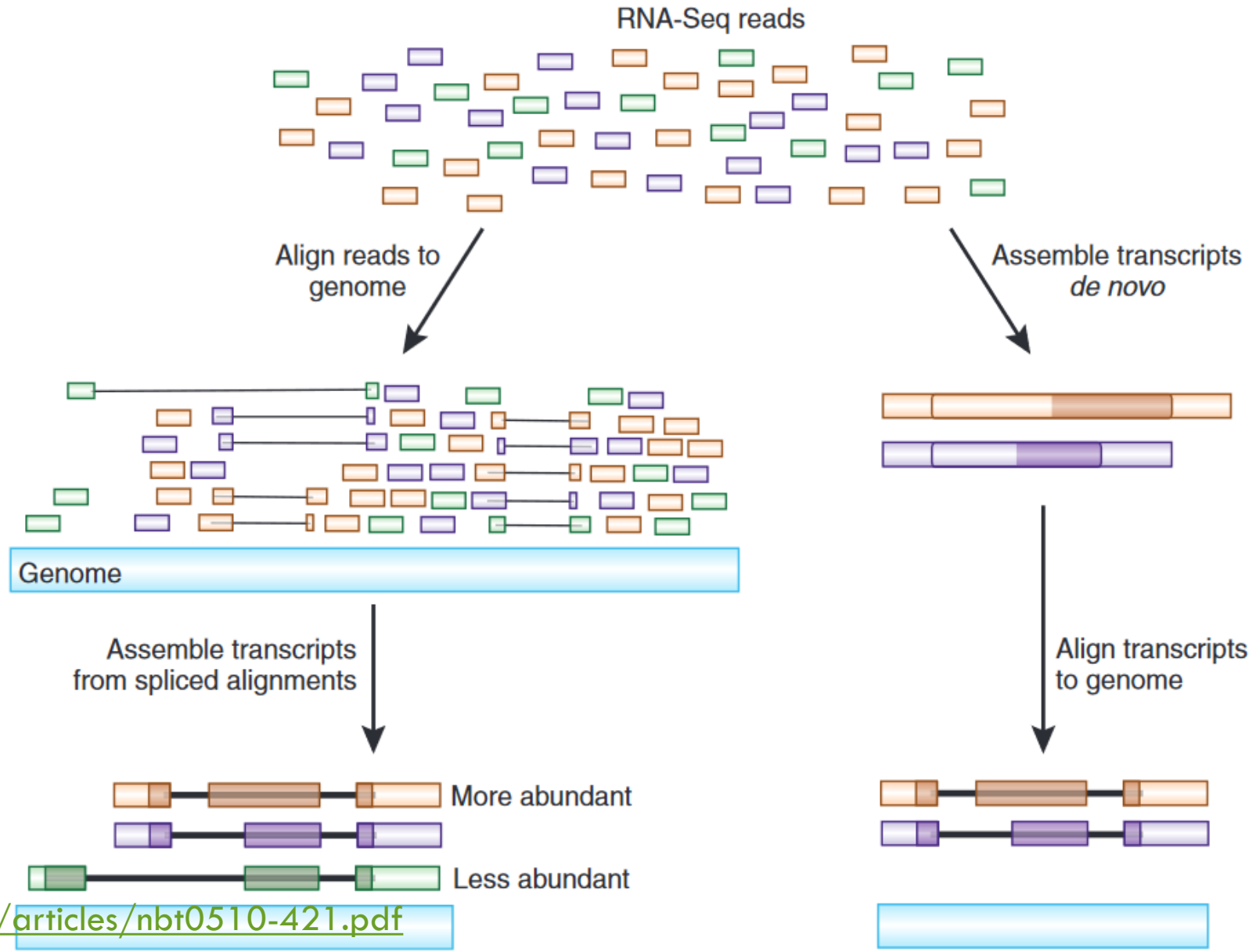
The energetics of genome complexity

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

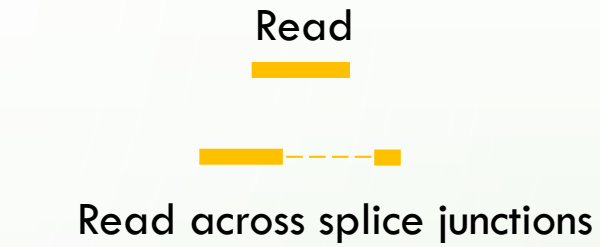


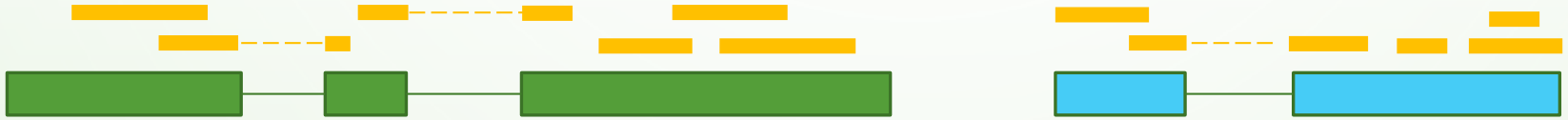
Haas & Zody, 2010
https://www.nature.com/articles/nbt0510-421.pdf

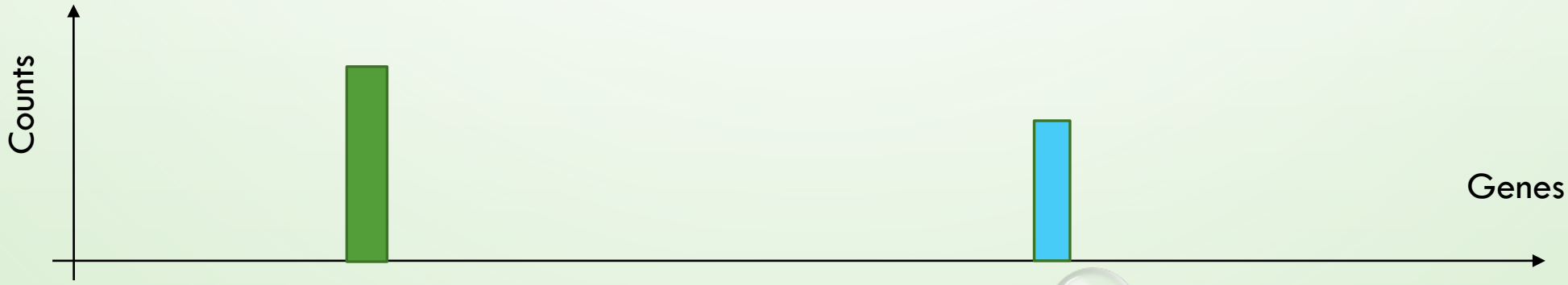
QUANTIFYING GENE EXPRESSION

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

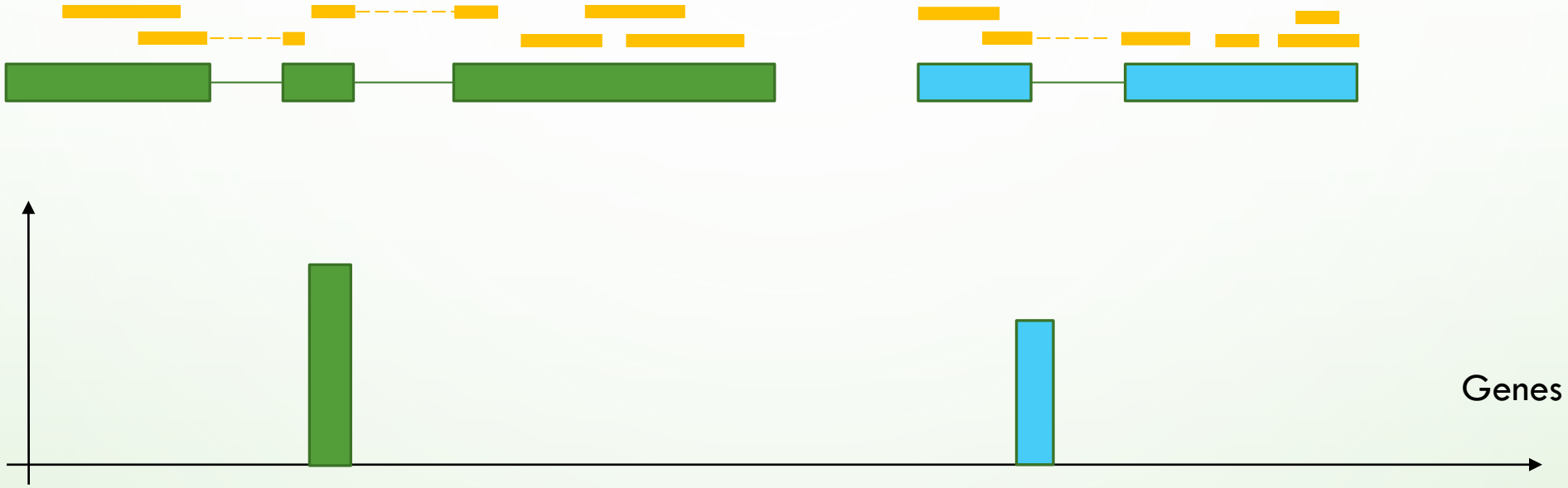
- Gene expression
- Expression of different isoforms (challenging)





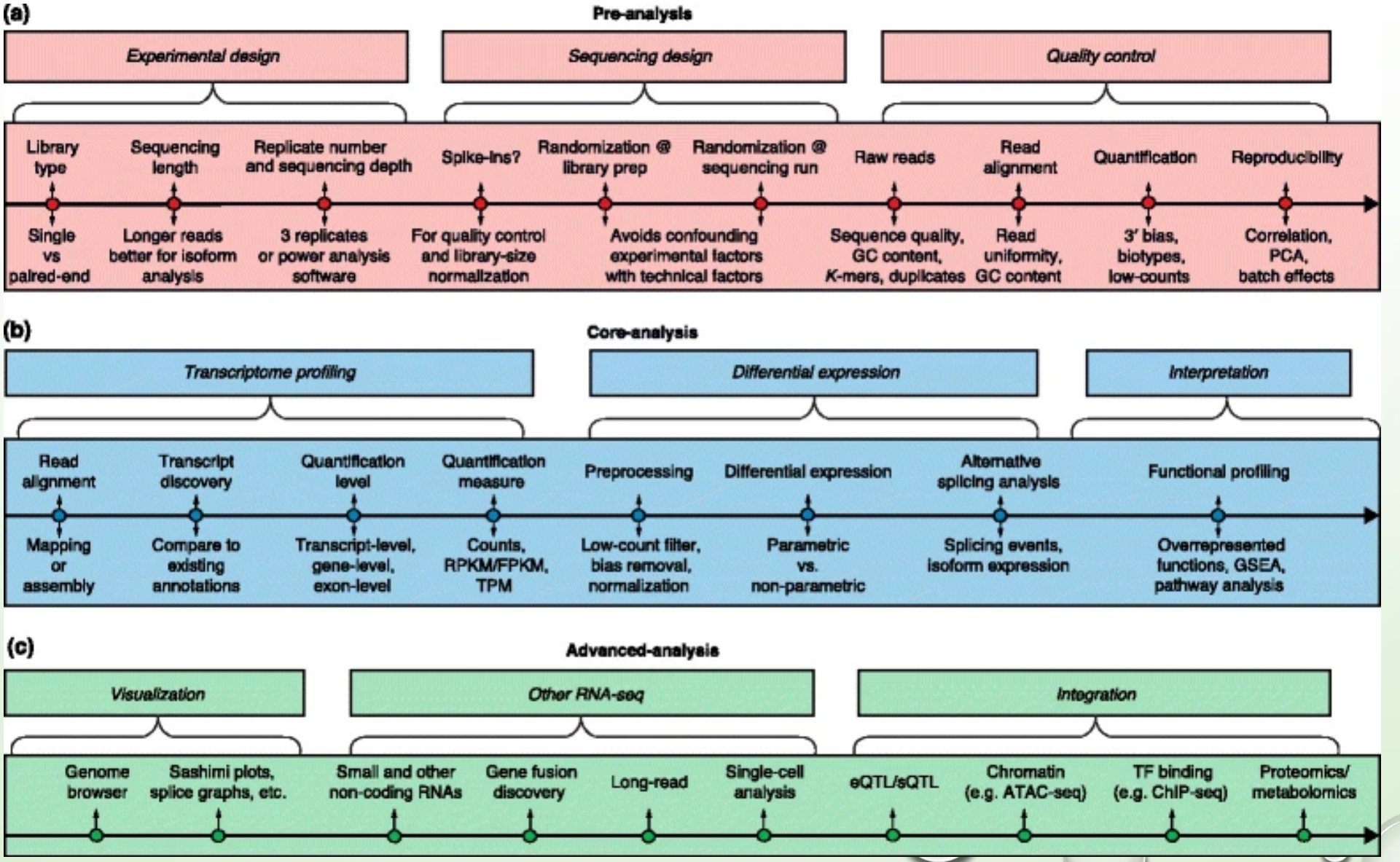


COUNTS/GENE EXPRESSION MATRIX



	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