GENOME INFORMATICS

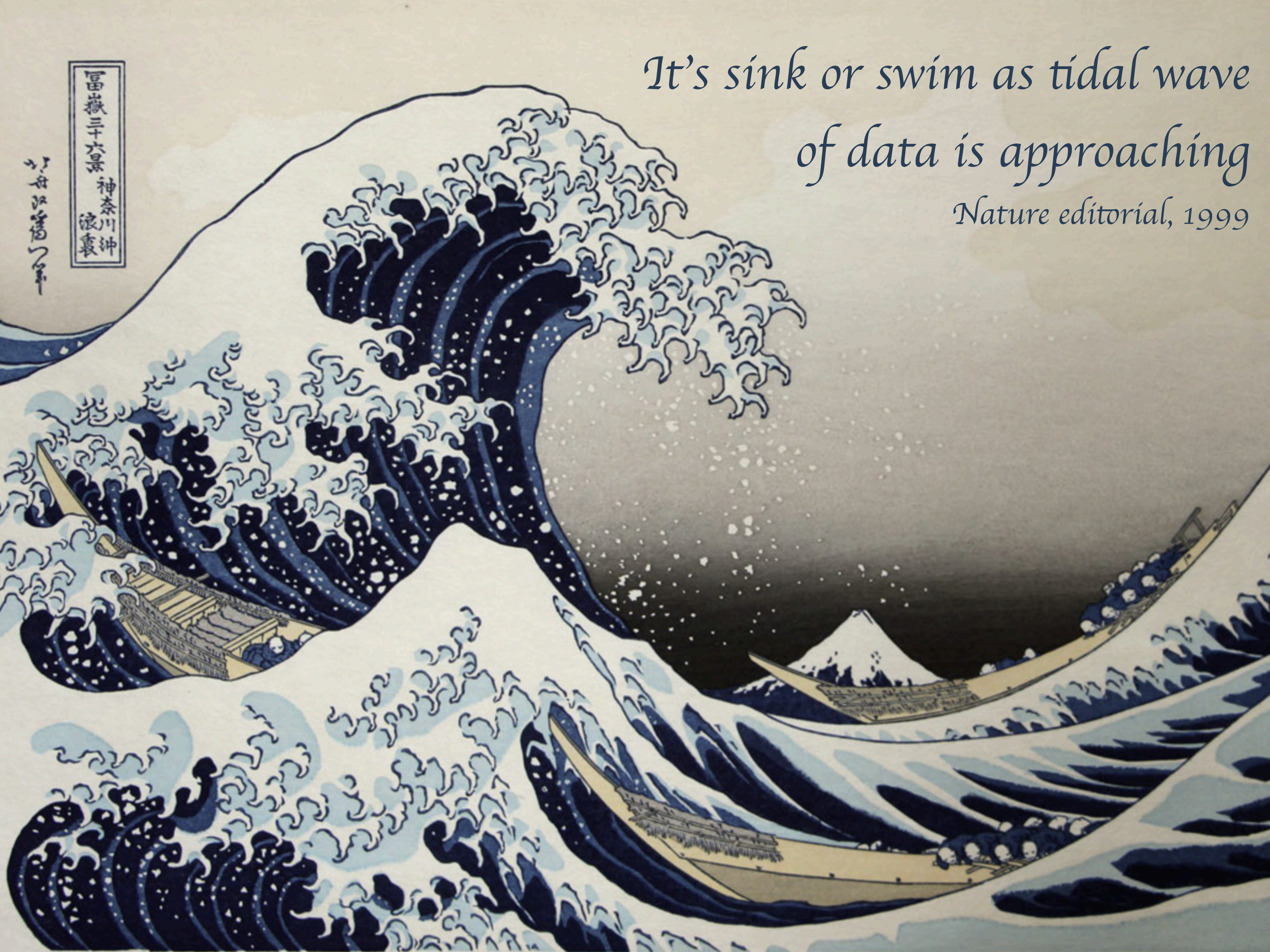
http://bioinformatics.uni-muenster.de/teaching/Current/Genome_informatics/index.hbi

Prof. Dr. Wojciech Makałowski
Institute of Bioinformatics
University of Münster, Germany

SEQUENCING TECHNOLOGY

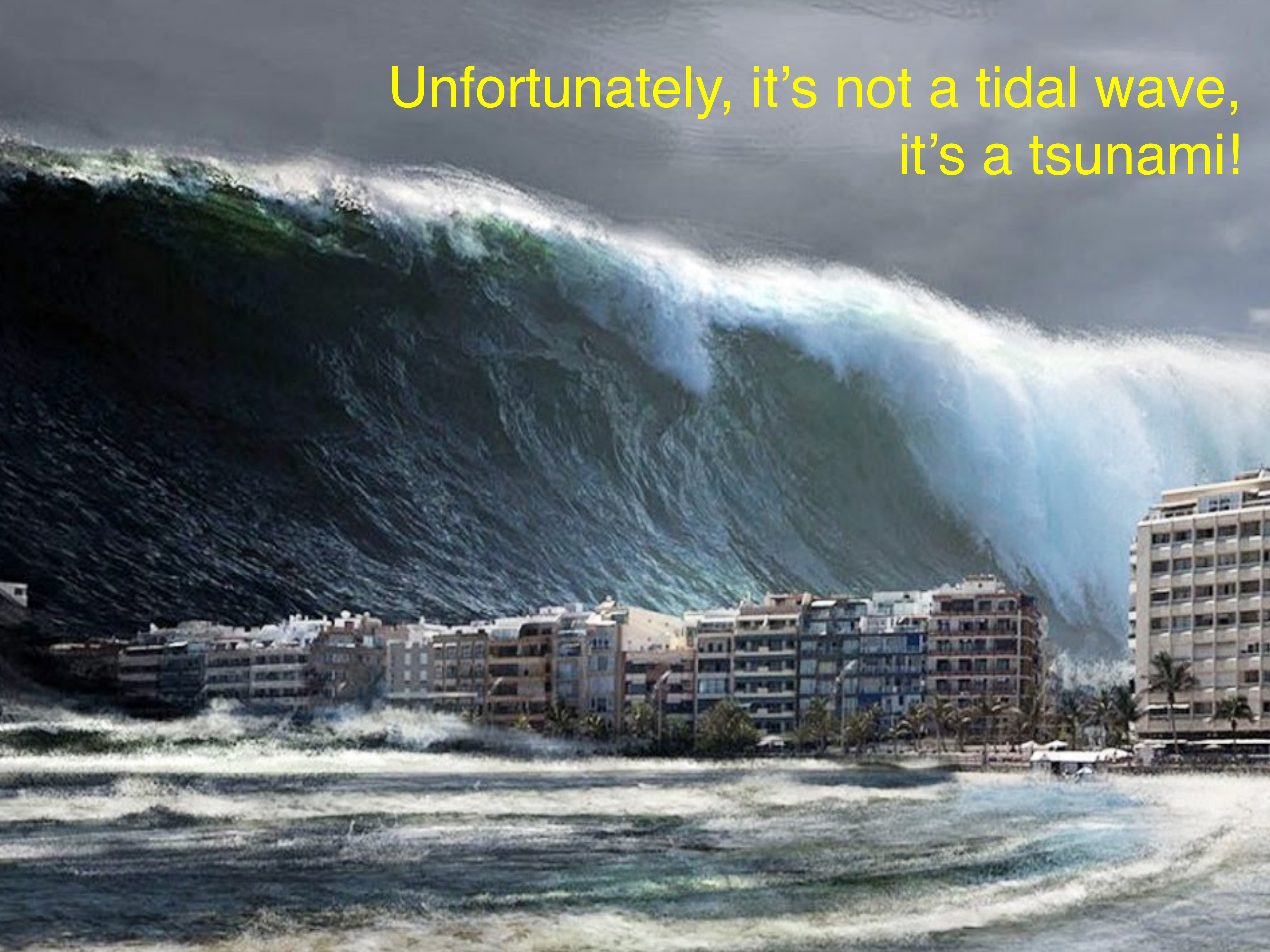
bioinformatic challenges

Prof. Dr. Wojciech Makałowski
Institute of Bioinformatics
University of Münster, Germany



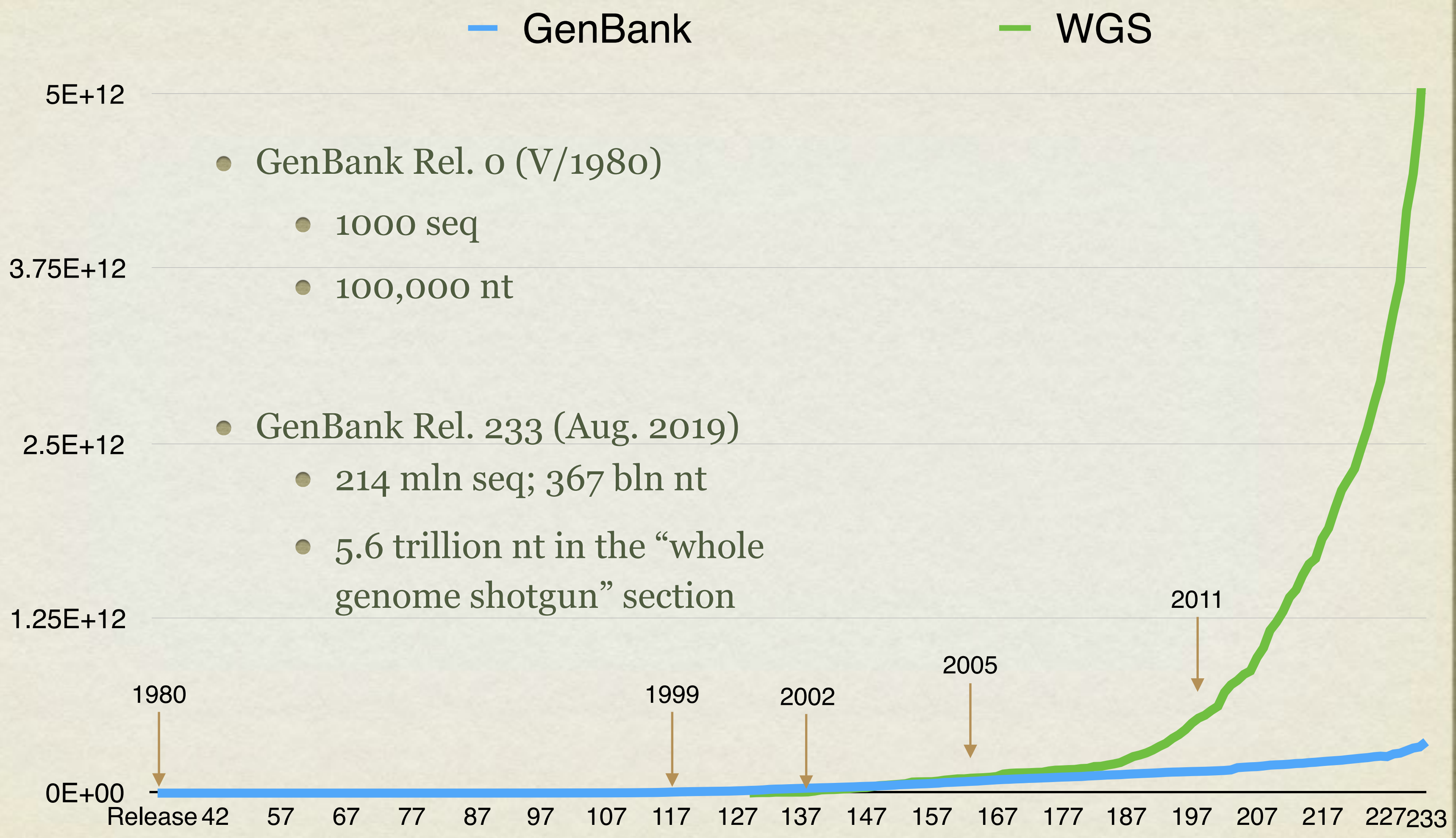
It's sink or swim as tidal wave of data is approaching

Nature editorial, 1999



Unfortunately, it’s not a tidal wave,
it’s a tsunami!

GROWTH OF BIOMEDICAL INFORMATION - GENBANK

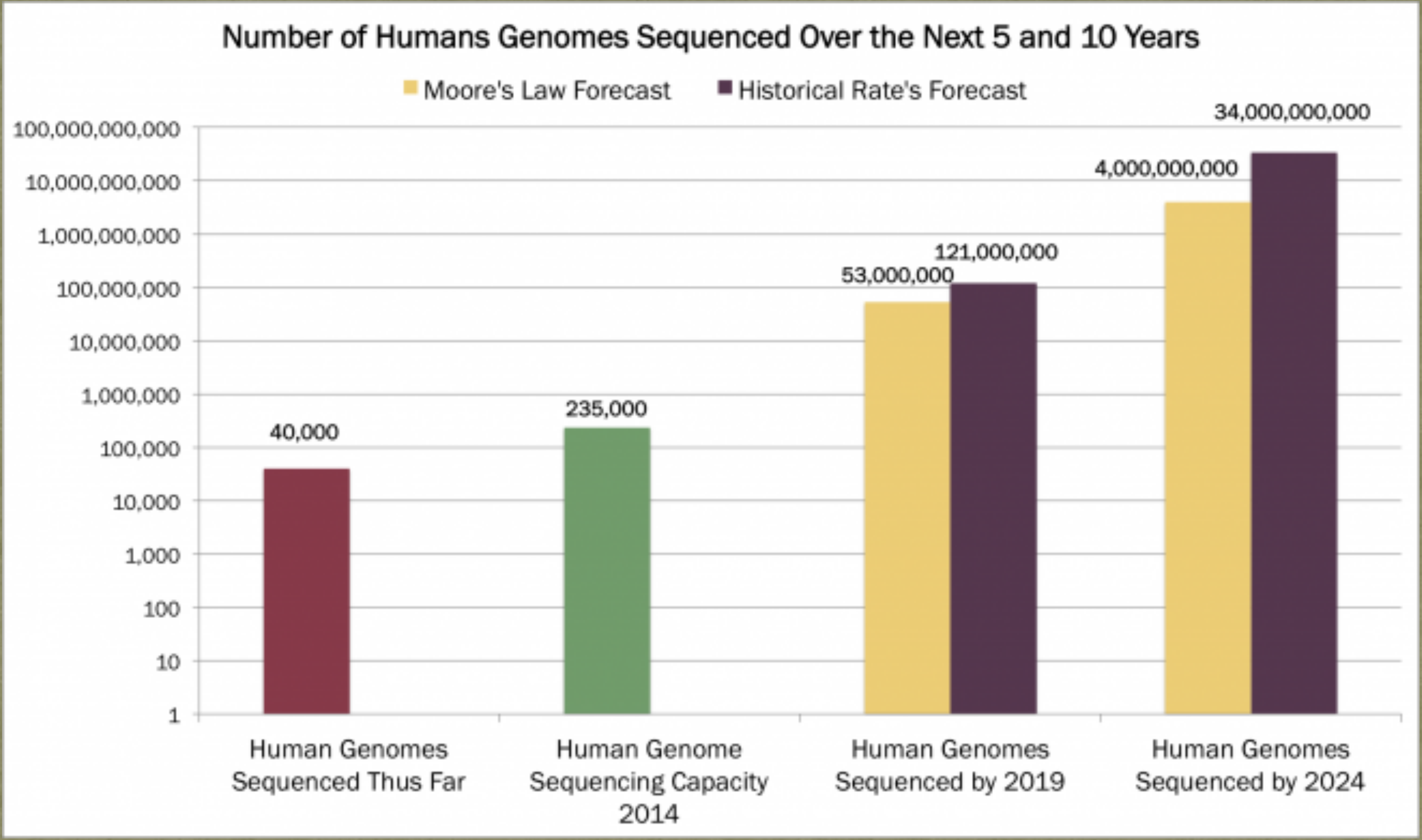


- GenBank Rel. 0 (V/1980)
  - 1000 seq
  - 100,000 nt
- GenBank Rel. 233 (Aug. 2019)
  - 214 mln seq; 367 bln nt
  - 5.6 trillion nt in the “whole genome shotgun” section

https://www.ncbi.nlm.nih.gov/genbank/statistics/

TECHNOLOGY MEETS BIOLOGY

IMPROVING TECHNOLOGY



http://ark-invest.com/genomic-revolution/declining-costs-of-genome-sequencing

GETTING SEQUENCES

```
TGCATCGATCGTAGCTAGCTAGCGCATGCTAGCTAGCTAGCTAGCTACGATGCATCG
TGCATCGATCGATGCATGCTAGCTAGCTAGCTAGCATGCTAGCTAGCTAGCTATTGG
CGCTAGCTAGCATGCATGCATGCATCGATGCATCGATTATAAGCGCGATGACGTCAG
CGCGCGCATTATGCCGCGGCATGCTGCGCACACACAGTACTATAGCATTAGTAAAAA
GGCCGCGTATATTTTACACGATAGTGCGGCGCGGCGCGTAGCTAGTGCTAGCTAGTC
TCCGGTTACACAGGTAGCTAGCTAGCTGCTAGCTAGCTGCTGCATGCATGCATTAGT
AGCTAGTGTAGCTAGCTAGCATGCTGCTAGCATGCAGCATGCATCGGGCGCGATGCT
GCTAGCGCTGCTAGCTAGCTAGCTAGCTAGGCGCTAATTATTTATTTTGGGGGGTTA
AAAAAAAAAATTTCGCTGCTTATACCCCCCCCCACATGATGATCGTTAGTAGCTACT
AGCTCTCATCGCGCGGGGGGATGCTTAGCGTGGTGTGTGTGTGGTGTGTGTGTGGTC
CTATAATTAGTGCATCGGCGCATCGATGGCTAGTCGATCGATCGATTTTATATATCT
AAAGACCCCATCTCTCTCTCTTTTCCCTTCTCTCGCTAGCGGGCGGTACGATTTACC
GGCCGCGTATATTTTACACGATAGTGCGGCGCGGCGCGTAGCTAGTGCTAGCTAGTC
AGCTCTCATCGCGCGGGGGGATGCTTAGCGTGGTGTGTGTGTGGTGTGTGTGTGGTC
TGCATCGATCGATGCATGCTAGCTAGCTAGCTAGCATGCTAGCTAGCTAGCTATTGG
CTATAATTAGTGCATCGGCGCATCGATGGCTAGTCGATCGATCGATTTTATATATCT
CGCTAGCTAGCATGCATGCATGCATCGATGCATCGATTATAAGCGCGATGACGTCAG
TCCGGTTACACAGGTAGCTAGCTAGCTGCTAGCTAGCTGCTGCATGCATGCATTAGT
```

READING ≠ UNDERSTANDING

Carmina qui quondam studio florente peregi, flebilis heu maestos cogor inire modos.

Ecce mihi lacerae dictant scribenda Camenae et ueris elegi fletibus orarigant.

Boethius, *Consolatio Philosophiae*

READING ≠ UNDERSTANDING

We shall best understand the probable course of natural selection by taking the case of a country undergoing some physical change. If the country were open were open on its borders, new forms would certainly immigrate, and this also would bla, bla bla become extinct inhabitants.

Charles Darwin - *The Origin of Species*

READING ≠ UNDERSTANDING

[redacted]erstand the probable course of by taking the case of a country undergoing some physical change. If the country were open were open on its borders, new forms would certainly immigrate, and this also would bla, bla bla become extinct inhabitants.

Charles Darwin - *The Origin of Species*

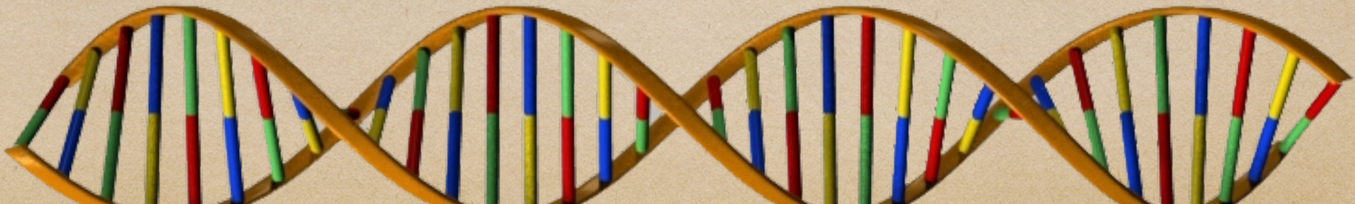
CHALLENGE: HOW FROM THIS...

```
TGCATCGATCGTAGCTAGCTAGCGCATGCTAGCTAGCTAGCTAGCTACGATGCATCG
TGCATCGATCGATGCATGCTAGCTAGCTAGCTAGCATGCTAGCTAGCTAGCTATTGG
CGCTAGCTAGCATGCATGCATGCATCGATGCATCGATTATAAGCGCGATGACGTCAG
CGCGCGCATTATGCCGCGGCATGCTGCGCACACACAGTACTATAGCATTAGTAAAAA
GGCCGCGTATATTTTACACGATAGTGCGGCGCGGCGCGTAGCTAGTGCTAGCTAGTC
TCCGGTTACACAGGTAGCTAGCTAGCTGCTAGCTAGCTGCTGCATGCATGCATTAGT
AGCTAGTGTAGCTAGCTAGCATGCTGCTAGCATGCAGCATGCATCGGGCGCGATGCT
GCTAGCGCTGCTAGCTAGCTAGCTAGCTAGGCGCTAATTATTTATTTTGGGGGGTTA
AAAAAAAAAATTTCGCTGCTTATACCCCCCCCCACATGATGATCGTTAGTAGCTACT
AGCTCTCATCGCGCGGGGGGATGCTTAGCGTGGTGTGTGTGTGTGGTGTGTGTGGTC
CTATAATTAGTGCATCGGCGCATCGATGGCTAGTCGATCGATCGATTTTATATATCT
AAAGACCCCATCTCTCTCTCTTTTCCCTTCTCTCGCTAGCGGGCGGTACGATTTACC
GGCCGCGTATATTTTACACGATAGTGCGGCGCGGCGCGTAGCTAGTGCTAGCTAGTC
AGCTCTCATCGCGCGGGGGGATGCTTAGCGTGGTGTGTGTGTGTGGTGTGTGTGGTC
TGCATCGATCGATGCATGCTAGCTAGCTAGCTAGCATGCTAGCTAGCTAGCTATTGG
CTATAATTAGTGCATCGGCGCATCGATGGCTAGTCGATCGATCGATTTTATATATCT
CGCTAGCTAGCATGCATGCATGCATCGATGCATCGATTATAAGCGCGATGACGTCAG
TCCGGTTACACAGGTAGCTAGCTAGCTGCTAGCTAGCTGCTGCATGCATGCATTAGT
```

Infer this

“The double helix is indeed a remarkable molecule. Modern man is perhaps 50,000 years old, civilization has existed for scarcely 10,000 years and the United States for only just over 200 years; but DNA and RNA have been around for at least several billion years. All that time the double helix has been there, and active, and yet we are the first creatures on Earth to become aware of its existence.”

Francis Crick (1916–2004)

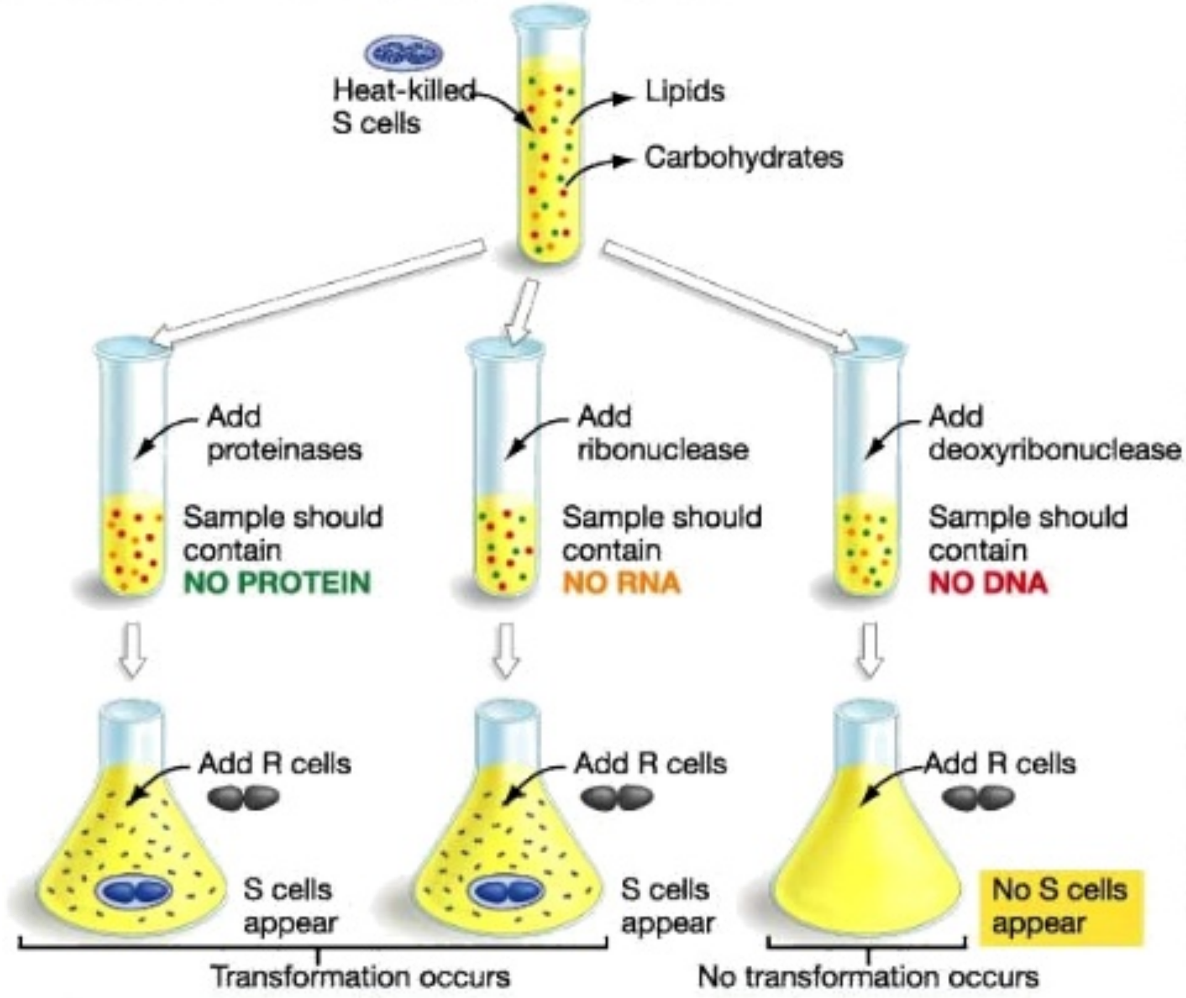


DNA story

1870 Friedrich Miescher discovers DNA

1944 Oswald Avery proves that DNA is a genetic material

DETERMINING THAT DNA IS THE HEREDITARY MATERIAL



1. Remove the lipids and carbohydrates from a solution of heat-killed S cells. Proteins, RNA, and DNA remain.

2. Subject the solution to treatments of enzymes to destroy either the proteins, RNA, or DNA.

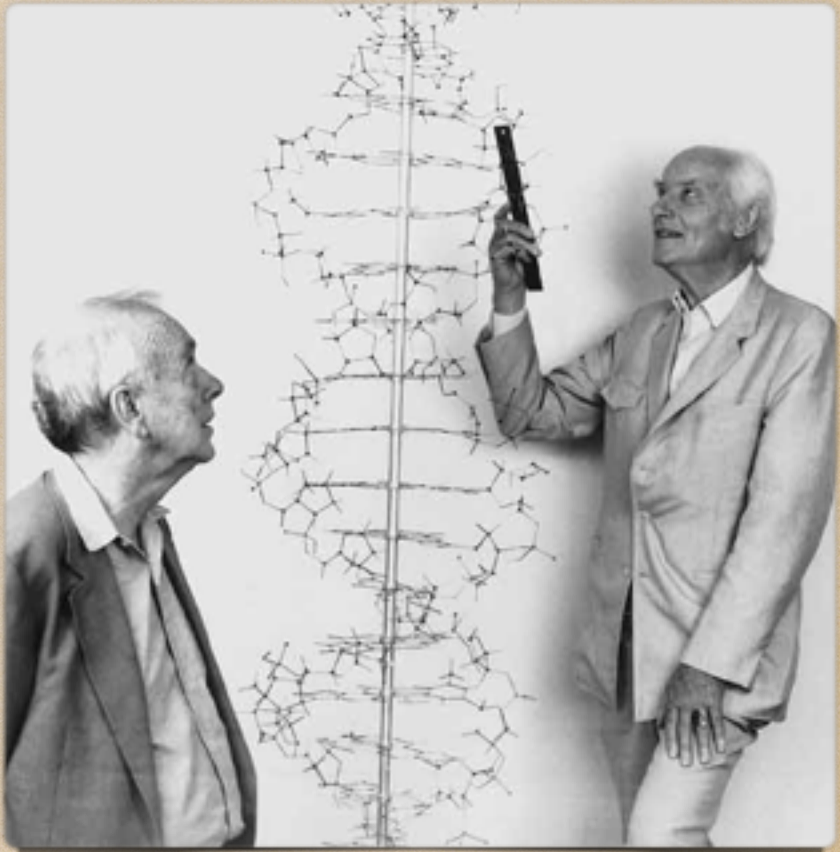
3. Add a small portion of each sample to a culture containing R cells. Observe whether transformation has occurred by testing for the presence of virulent S cells.

http://biology.kenyon.edu/courses/biol114/KH_lecture_images/How_DNA_works/FG11_02.JPG

DNA story

1953 James Watson and Francis Crick discover DNA structure

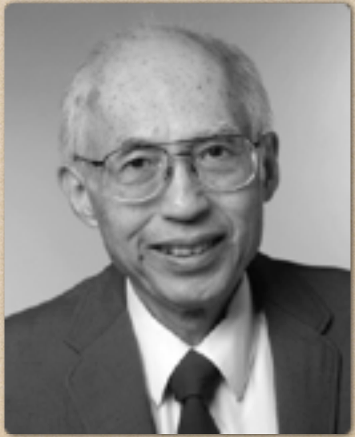
("Double Helix")



Sequencing: beginnings

1964 Robert W. Holley determines nucleotide sequences (77 nt) of the yeast Alanine tRNA
J. Biol. Chem. 240: 2122-2128

1968 Ray Wu and A. Dale Kaiser sequenced 12 bases (!) of λ phage's 5' cohesive ends of its DNA, using radioactively labeled nucleotides and polyacrylamide gel electrophoresis
J. Mol. Biol. 35: 523-537

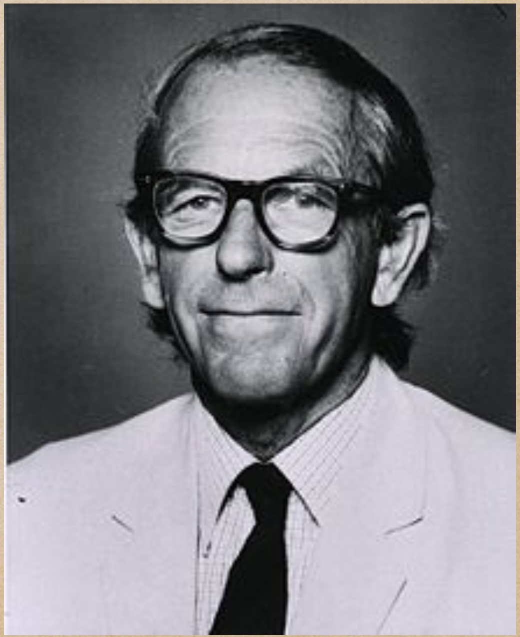


Sequencing:

1st generation sequencing

1977 - Allan Maxam and Walter Gilbert develop DNA sequencing method by chemical degradation

1977 Fred Sanger develops 2',3'-dideoxy chain termination method



Chemical degradation sequencing

(Maxam & Gilbert)

DNA extraction

↓

DNA fragmentation

↓

Strand dissociation

↓

Radioactive labeling of 5’ end

↓

Nucleotide-specific chemical reaction

↓

DNA cleavage at modified site

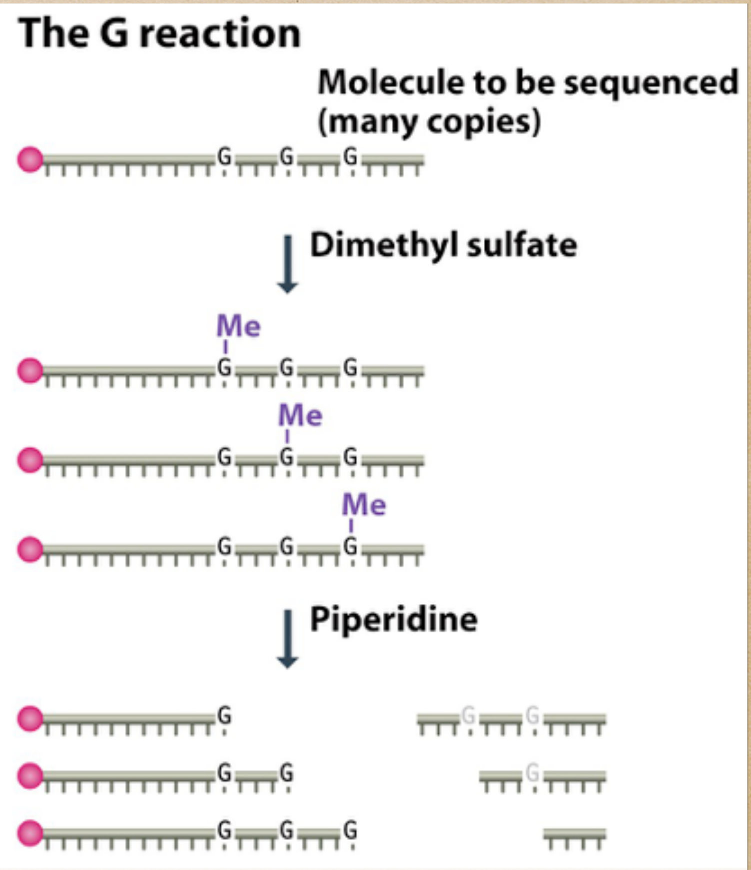
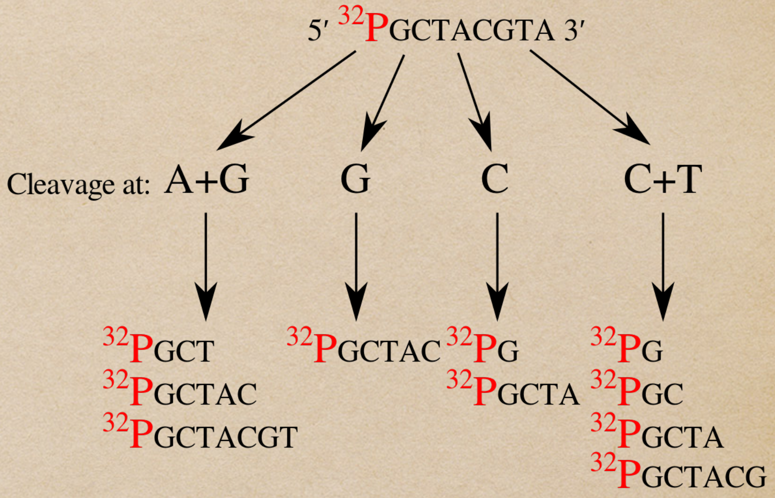


Figure 4.8 *Genomes 3* (© Garland Science 2007)

Chemical degradation sequencing

(Maxam&Gilbert)



5′ ^{32}PGCTACGTA 3′

Cleavage at: A+G G C C+T

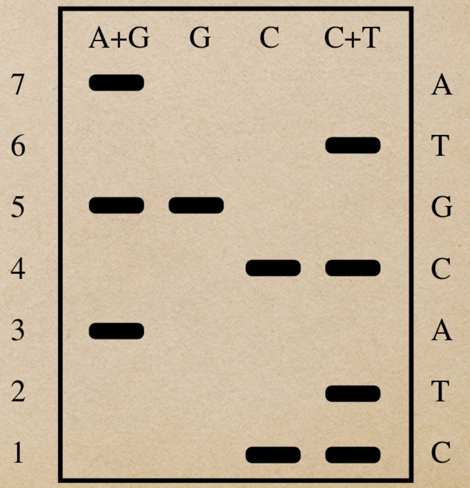
^{32}PGCT
^{32}PGCTAC
^{32}PGCTACGT

^{32}PGCTAC

^{32}PG
^{32}PGCTA

^{32}PG
^{32}PGC
^{32}PGCTA
^{32}PGCTACG

Four different reactions to detect four different nucleotides

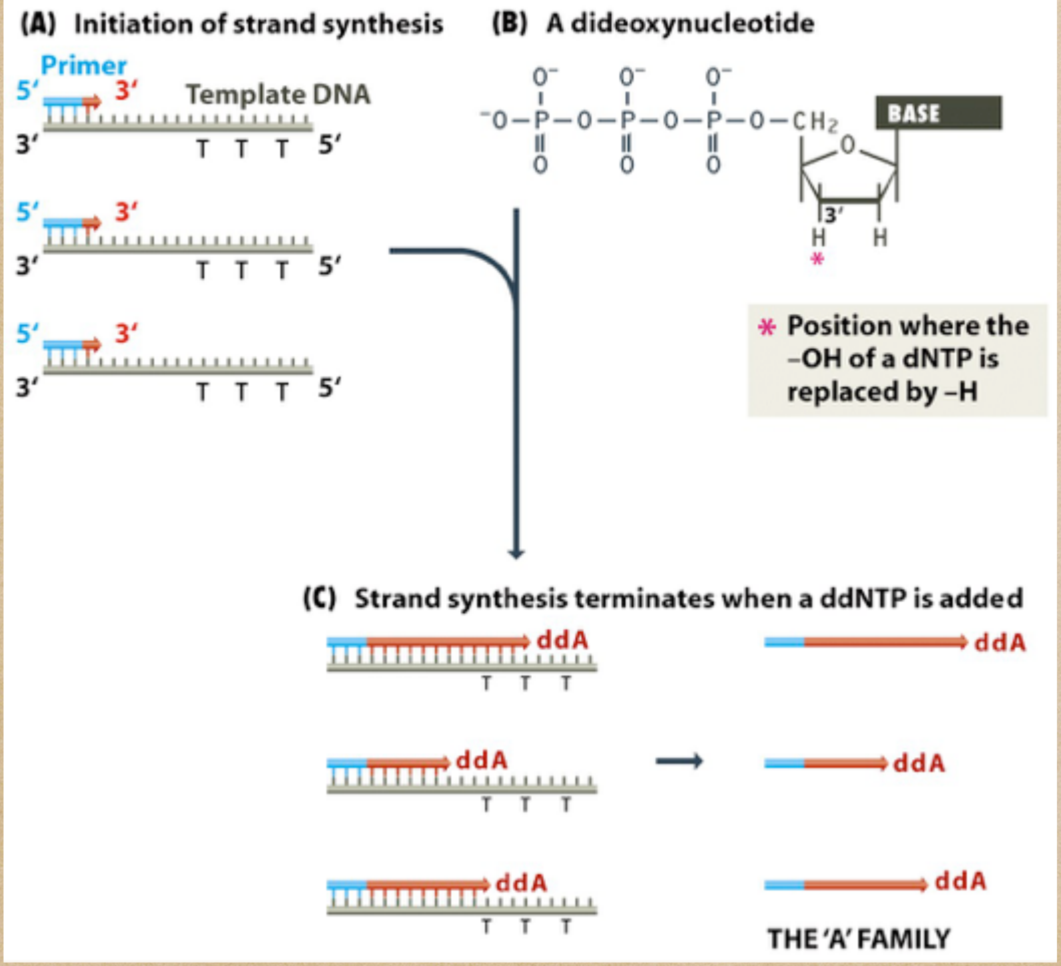


Sequencing Gel

Polyacrylamide gel electrophoresis can resolve single-stranded DNA molecules that differs in length by just one nucleotide and a sequence is read from an autoradiograph

Chain termination DNA sequencing

(Sanger)


(A) Initiation of strand synthesis
Primer
5′
3′
Template DNA
3′
T T T
5′
(B) A dideoxynucleotide
O⁻
⁻O–P–O–P–O–P–O–CH₂
O
BASE
3′
H
H
*
* Position where the –OH of a dNTP is replaced by –H
(C) Strand synthesis terminates when a ddNTP is added
ddA
T T T
THE 'A' FAMILY

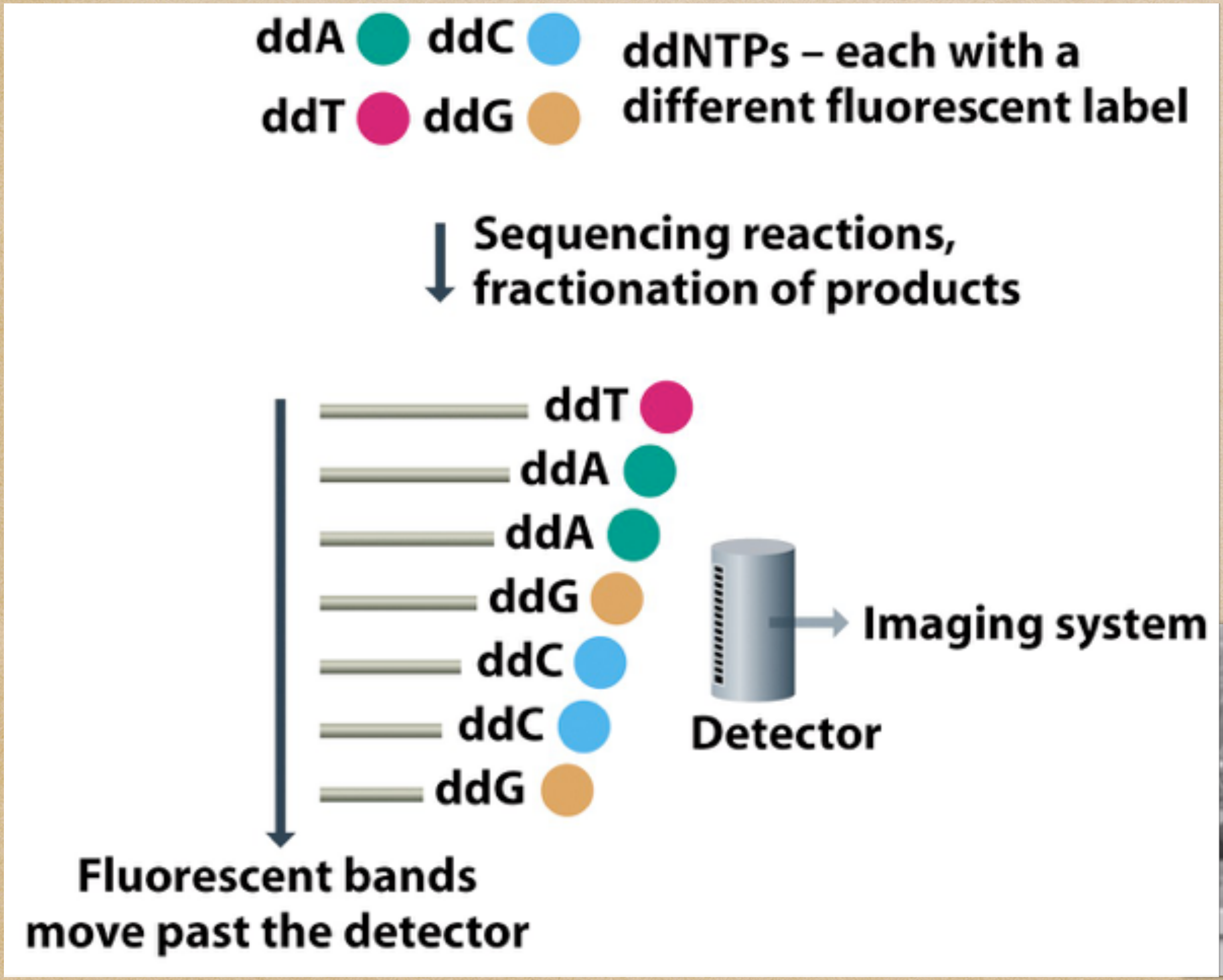
- use of DNA polymerase
- need for primers
- for each nucleotide a different analog
- similarly to M&G method separation of DNA fragments on polyacrylamide gel
- for each nucleotide a separate reaction
- sequence reading from an autoradiograph

Figure 4.2 *Genomes 3* (© Garland Science 2007)

Sequencing: maturation

- 1983 - Marvin Caruthers developed a method to construct fragments of DNA of predetermined sequence from five to about 75 base pairs long. He and Leroy Hood invented instruments that could make such fragments automatically.
- 1983 - Kary Mullis invented the polymerase chain reaction (PCR) technique
- 1987 - ABI 370; first fully automated sequencing machine by Leroy Hood
- 1995 - Craig Venter uses whole-genome shotgun sequencing technique to determine complete genome of bacterium Haemophilus influenzae
- 2005 - introduction of GS-20 sequencing machine; first in the line of “Next Generation Sequencing”, allowing hihg-throughput production

Sequencing: maturation



ddA ● ddC ●

ddT ● ddG ●

ddNTPs – each with a different fluorescent label

Sequencing reactions, fractionation of products

ddT ●

ddA ●

ddA ●

ddG ●

ddC ●

ddC ●

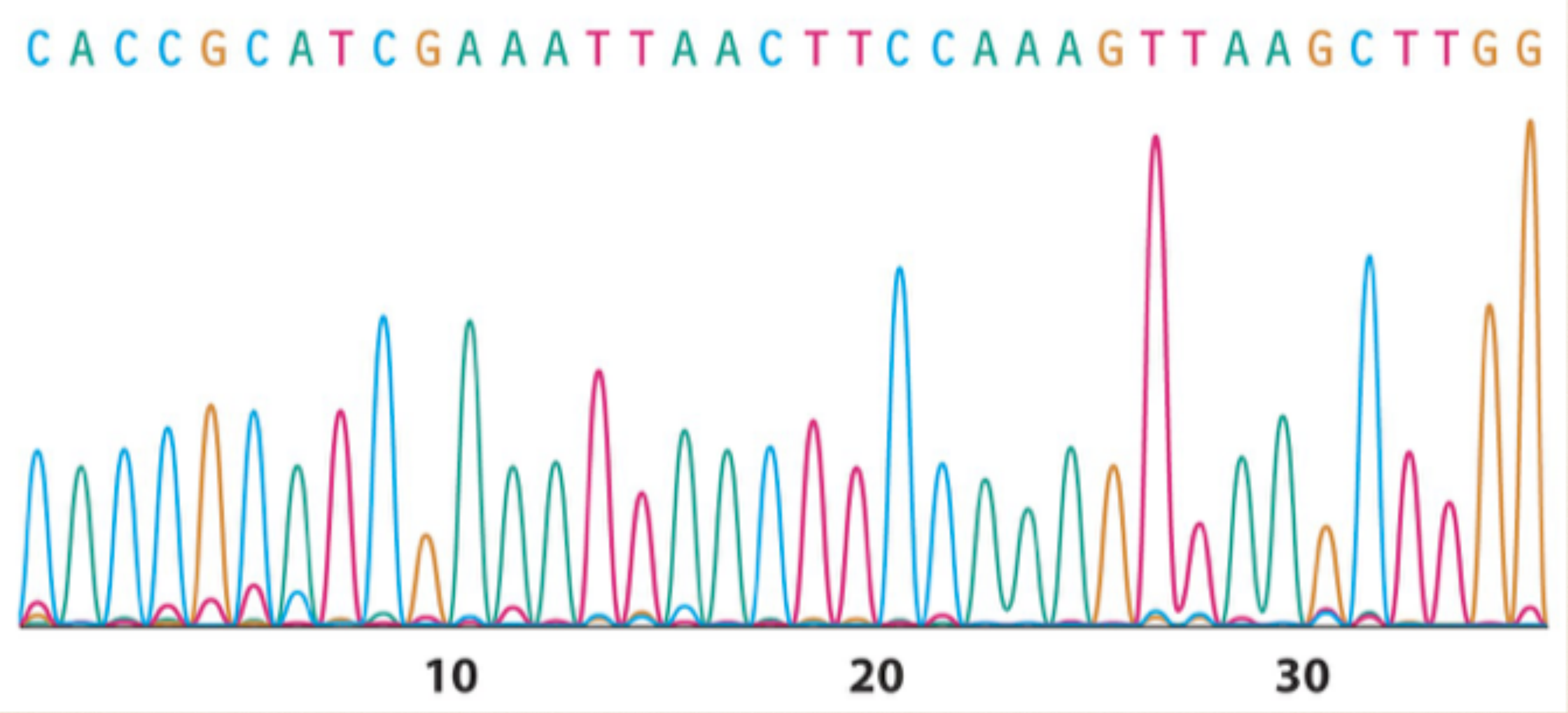
ddG ●

Detector

Imaging system

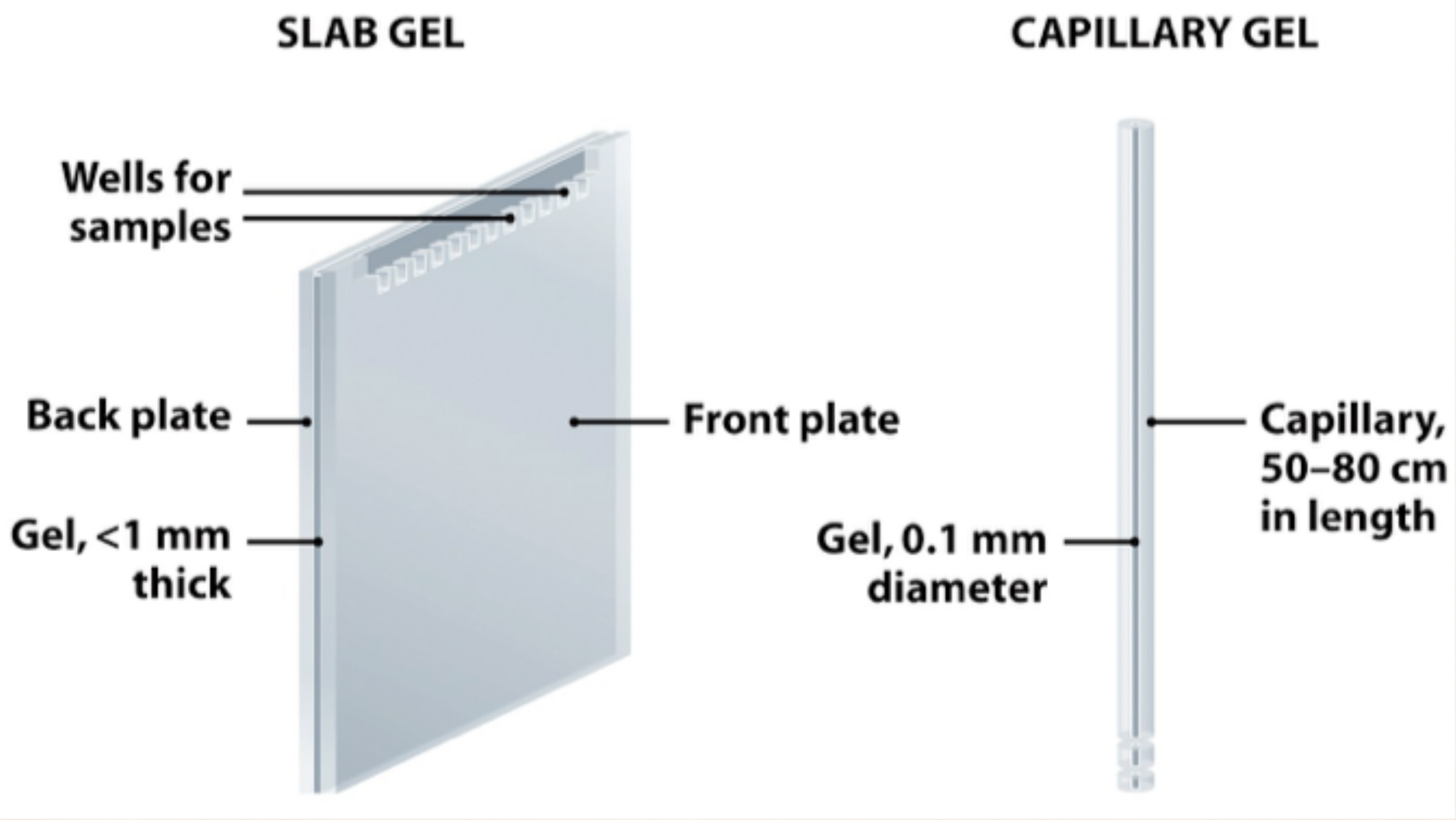
Fluorescent bands move past the detector

Sequencing: maturation



Chromatogram of a DNA sequence generated by ABI sequencing machine **(**[**https://www.dnalc.org/view/15912-Sequencing-DNA.html**](https://www.dnalc.org/view/15912-Sequencing-DNA.html) **)**

Sequencing: maturation



Sequencing: maturation

- 1983 - Marvin Caruthers developed a method to construct fragments of DNA of predetermined sequence from five to about 75 base pairs long. He and Leroy Hood invented instruments that could make such fragments automatically.
- 1983 - Kary Mullis invented the polymerase chain reaction (PCR) technique
- 1987 - ABI 370; first fully automated sequencing machine
- 1995 - Craig Venter uses whole-genome shotgun sequencing technique to determine complete genome of bacterium Haemophilus influenzae
- 2005 - introduction of GS20 sequencing machine (454 Lige Sciences); first in the line of "Next Generation Sequencing"

Next Generation Sequencing

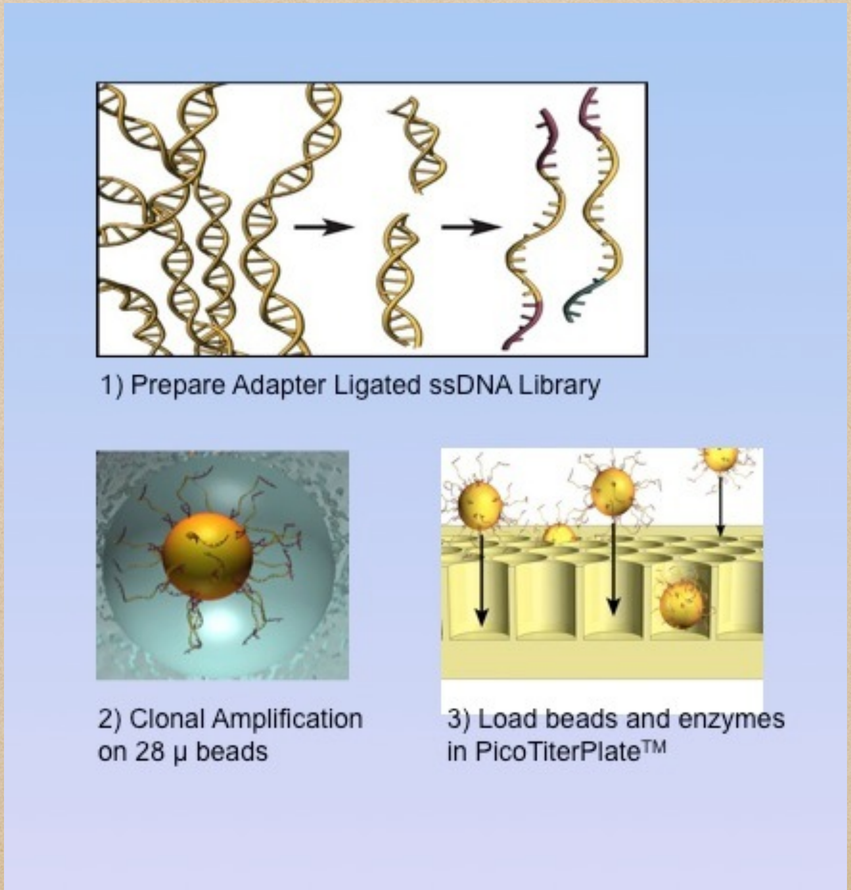
- Massive parallelization of the sequencing process
- Relatively short reads
- Different approaches from improving Sanger's technique to direct "observation" of DNA through a microscope
- Attempts to sequence single molecules without amplification step

Next Generation Sequencing

- 1 – Pyrosequencing (Roche454)
- 2 – Ion torrent (Thermo Fisher)
- 3 – Illumina

NGS – pyrosequencing

library preparation



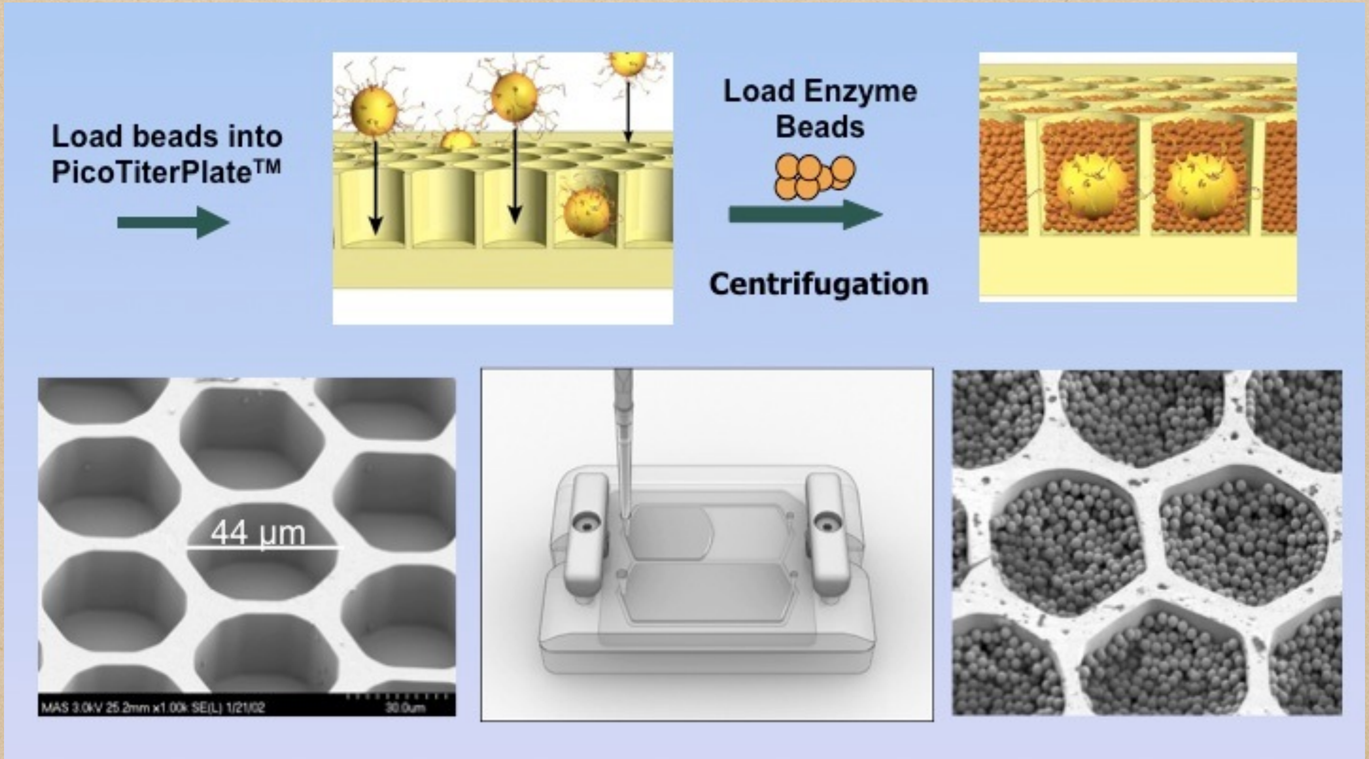
1) Prepare Adapter Ligated ssDNA Library

2) Clonal Amplification on 28 μ beads

3) Load beads and enzymes in PicoTiterPlate™

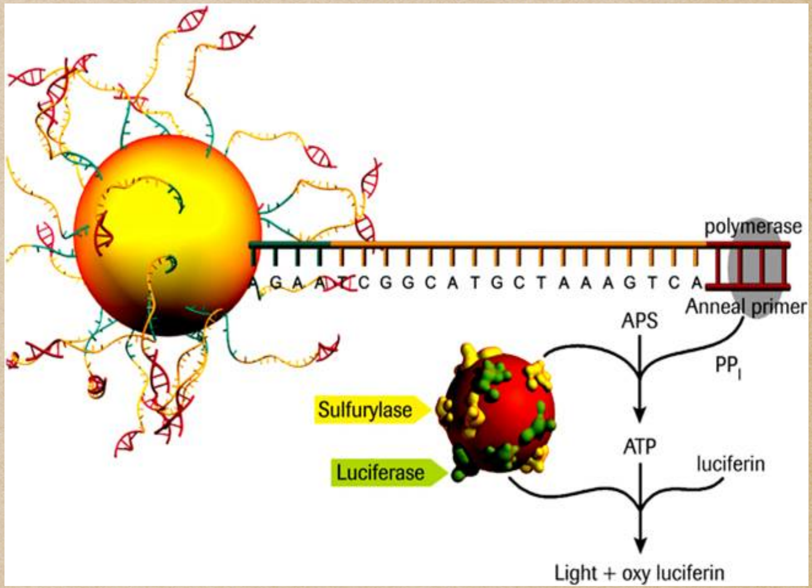
NGS - pyrosequencing

sample preparation


Load beads into PicoTiterPlate™
Load Enzyme Beads
Centrifugation
44 µm
MAS 3.0kV 25.2mm x1.00k SE(L) 1/21/02
30.0um

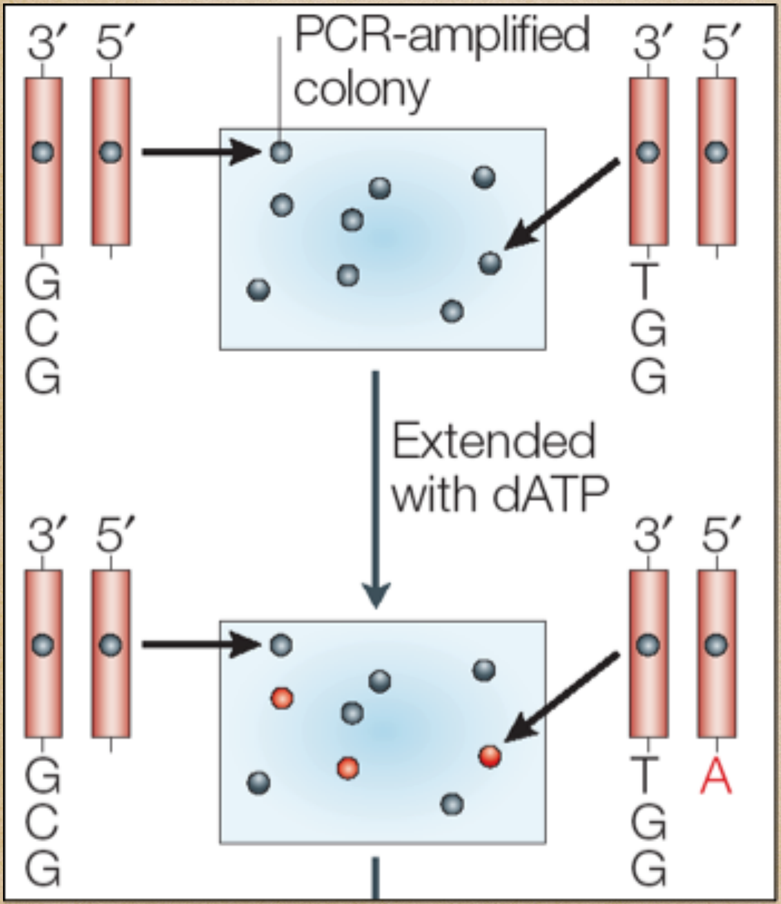
NGS - pyrosequencing

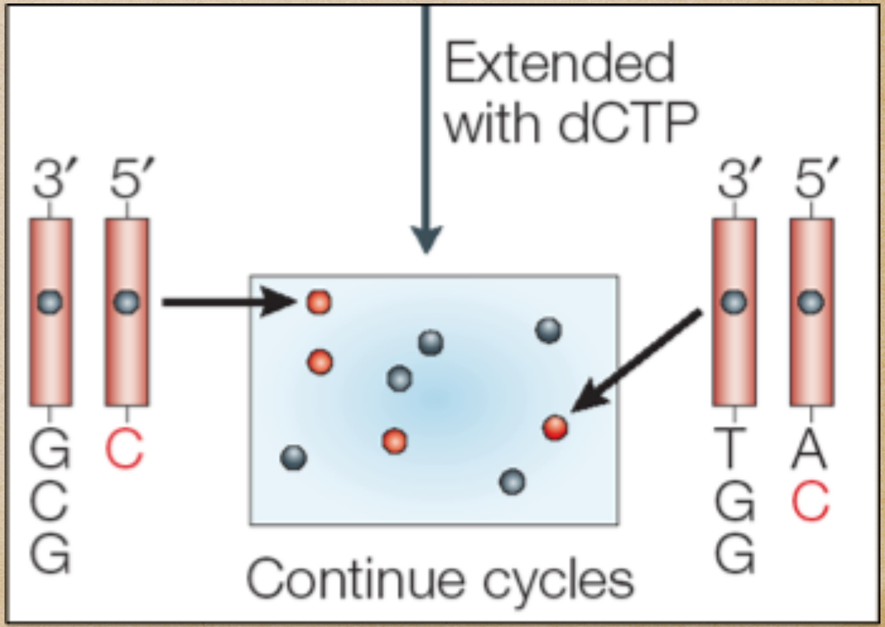
- After the emulsion PCR has been performed, the oil is removed, and the beads are put into a "picotiter" plate. Each well is just big enough to hold a single bead.
- The pyrosequencing enzymes are attached to much smaller beads, which are then added to each well.
- The plate is then repeatedly washed with the each of the four dNTPs, plus other necessary reagents, in a repeating cycle.
- The plate is coupled to a fiber optic chip. A CCD camera records the light flashes from each well.



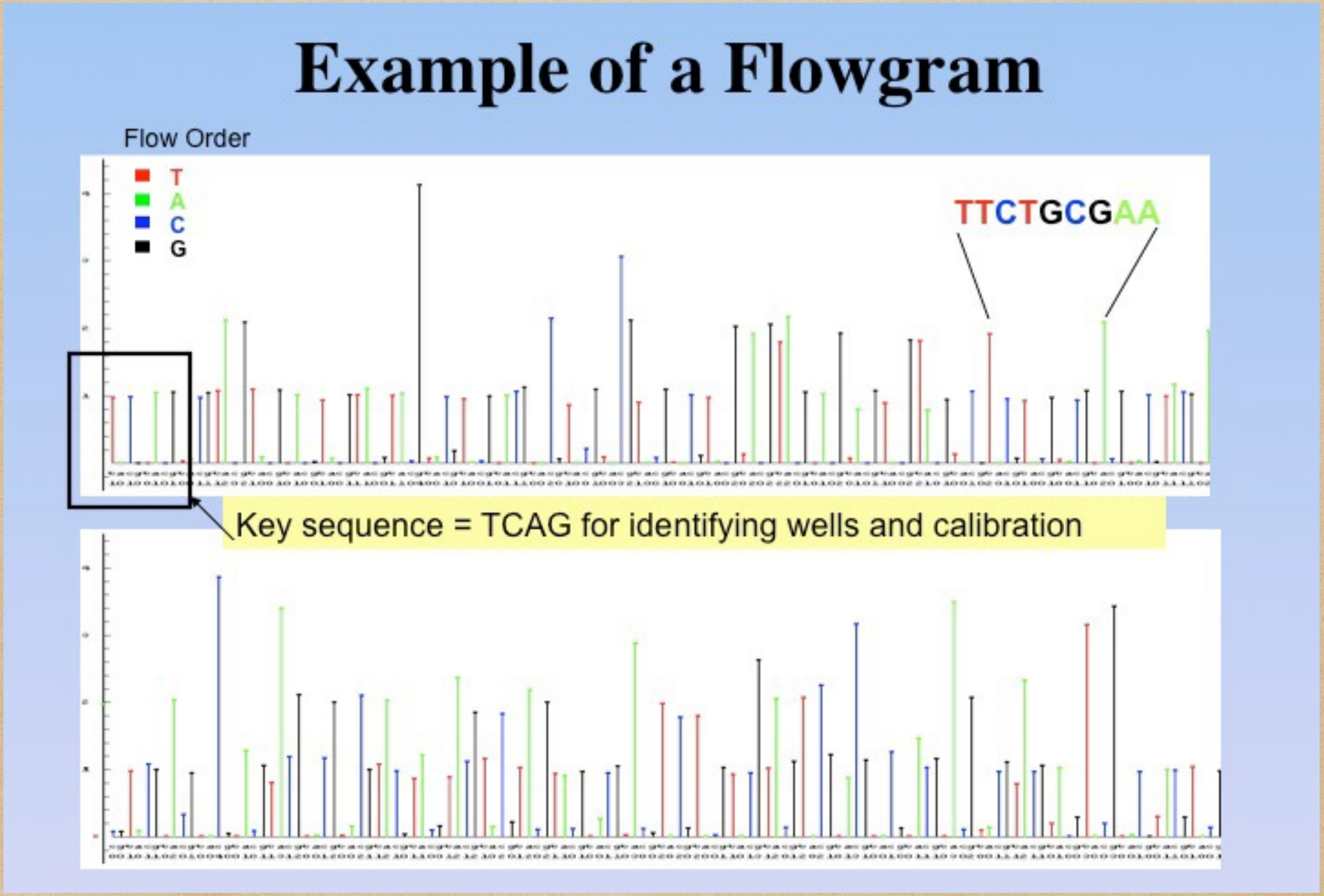
NGS - pyrosequencing

Extension with individual dNTPs gives a readout. The readout is recorded by a detector that measures position of light flashes and intensity of light flashes.





NGS - pyrosequencing



Example of a Flowgram

Key sequence = TCAG for identifying wells and calibration

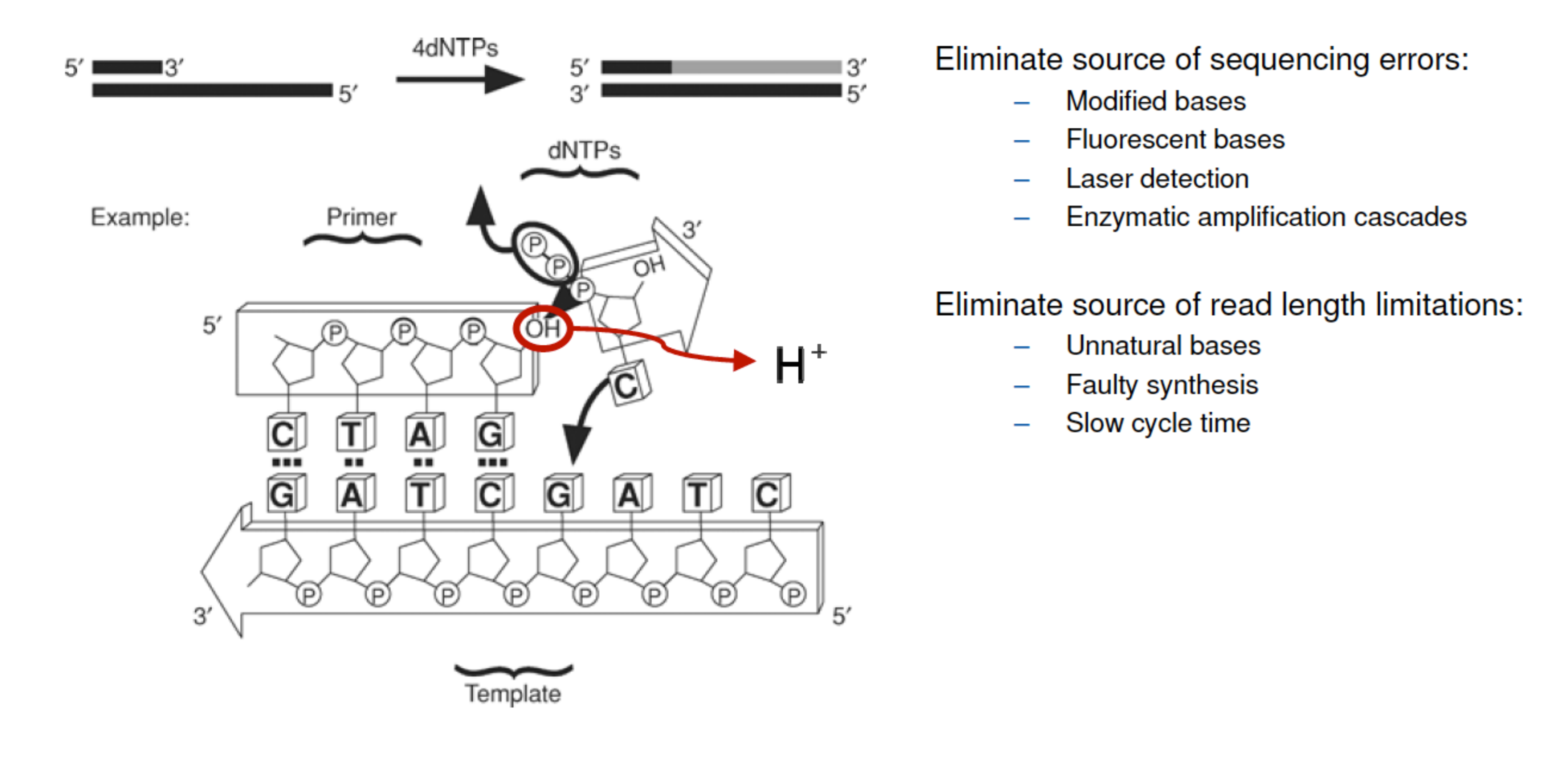
NGS -ion torrent

- Ten times faster workflow than other NGS systems
- ~2 hour sequencing runs (real-time detection of sequence extension)
- Batch sample preparation (six samples in six hours)
- Capable of six samples/day on two PGM Systems

https://www.youtube.com/watch?v=DyijNS0LWBY

NGS -ion torrent
Simple Natural Chemistry

Sequencing by synthesis



Eliminate source of sequencing errors:

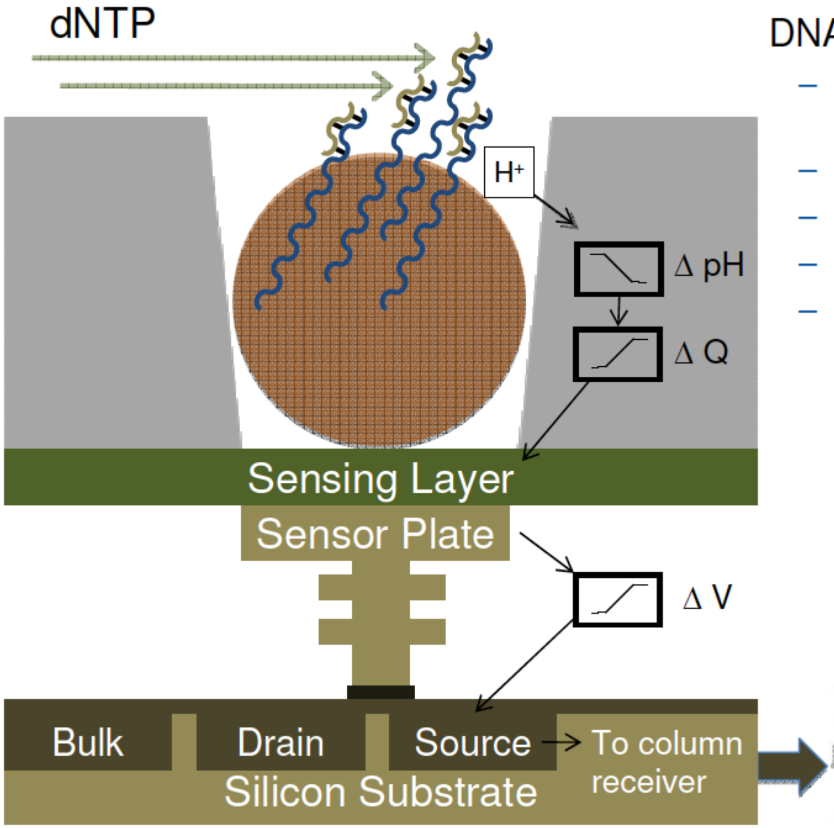
- Modified bases
- Fluorescent bases
- Laser detection
- Enzymatic amplification cascades

Eliminate source of read length limitations:

- Unnatural bases
- Faulty synthesis
- Slow cycle time

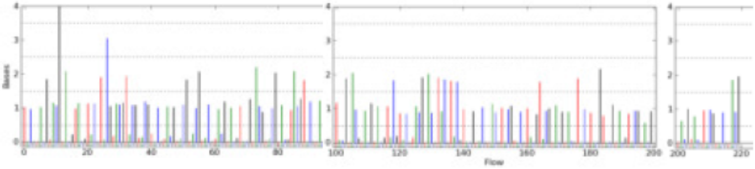
NGS -ion torrent

Fast Direct Detection



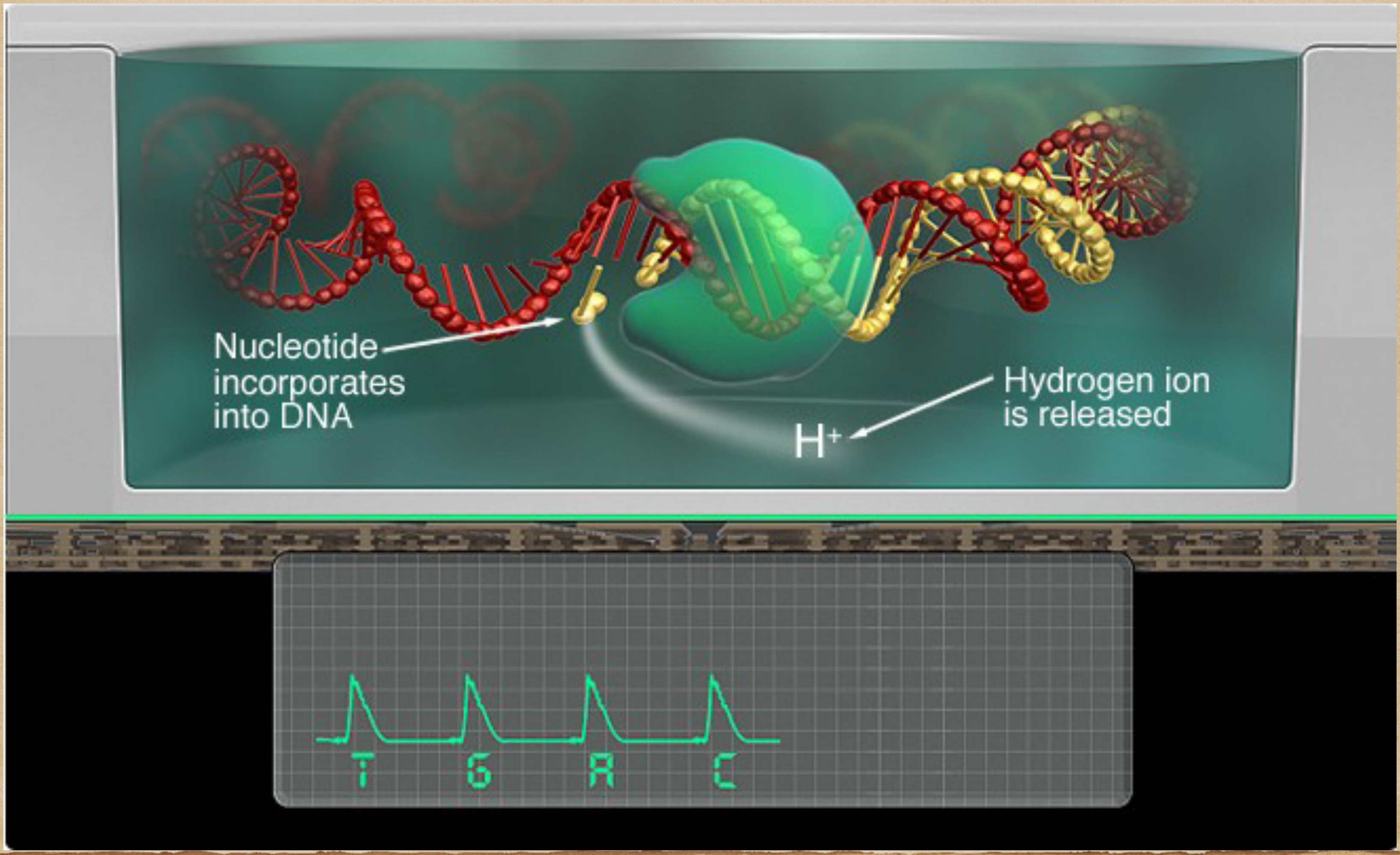
DNA → Ions → Sequence

- Nucleotides flow sequentially over Ion semiconductor chip
- One sensor per well per sequencing reaction
- Direct detection of natural DNA extension
- Millions of sequencing reactions per chip
- Fast cycle time, real time detection



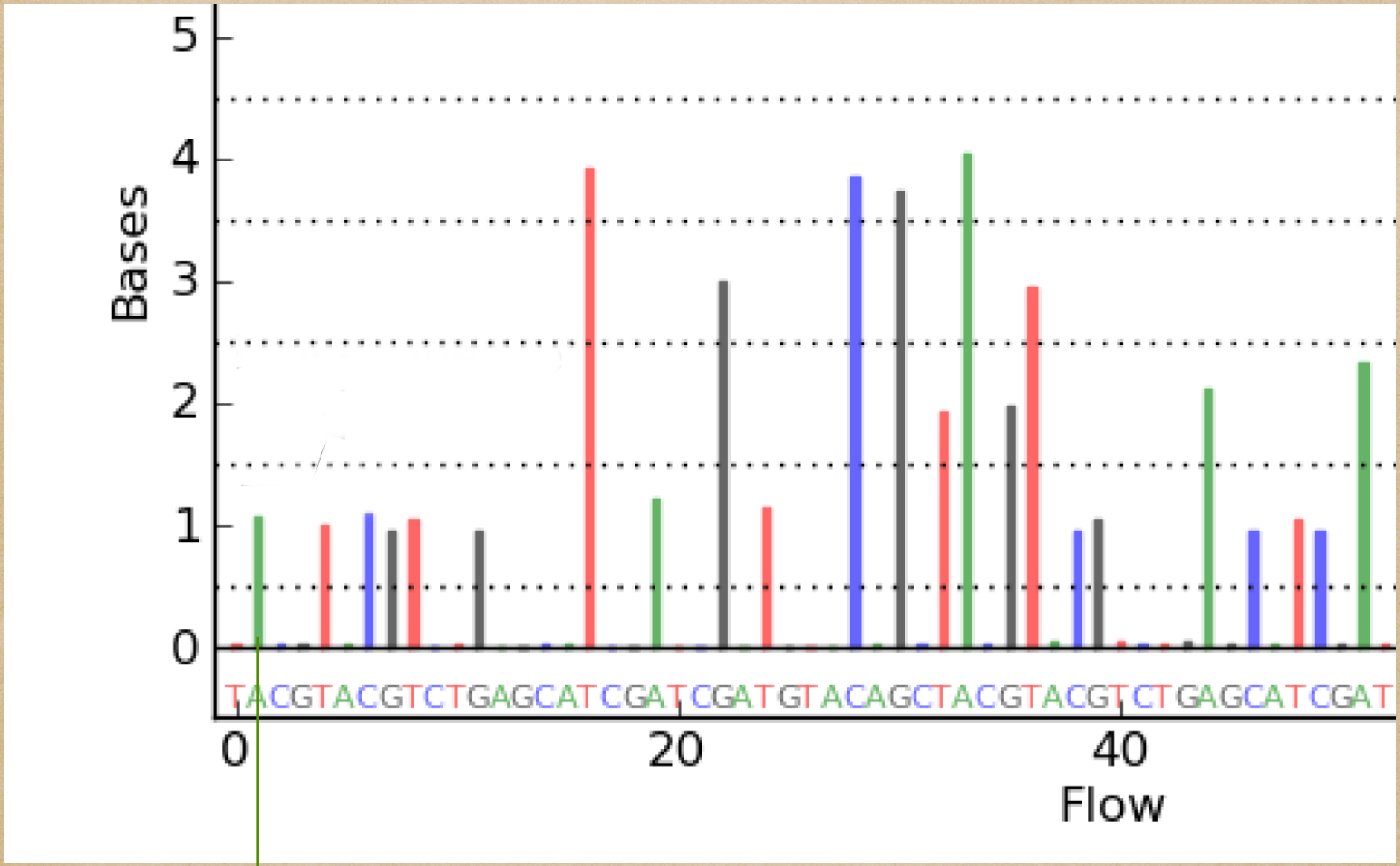
NGS -ION TORRENT

Four nucleotides flow sequentially



Base call

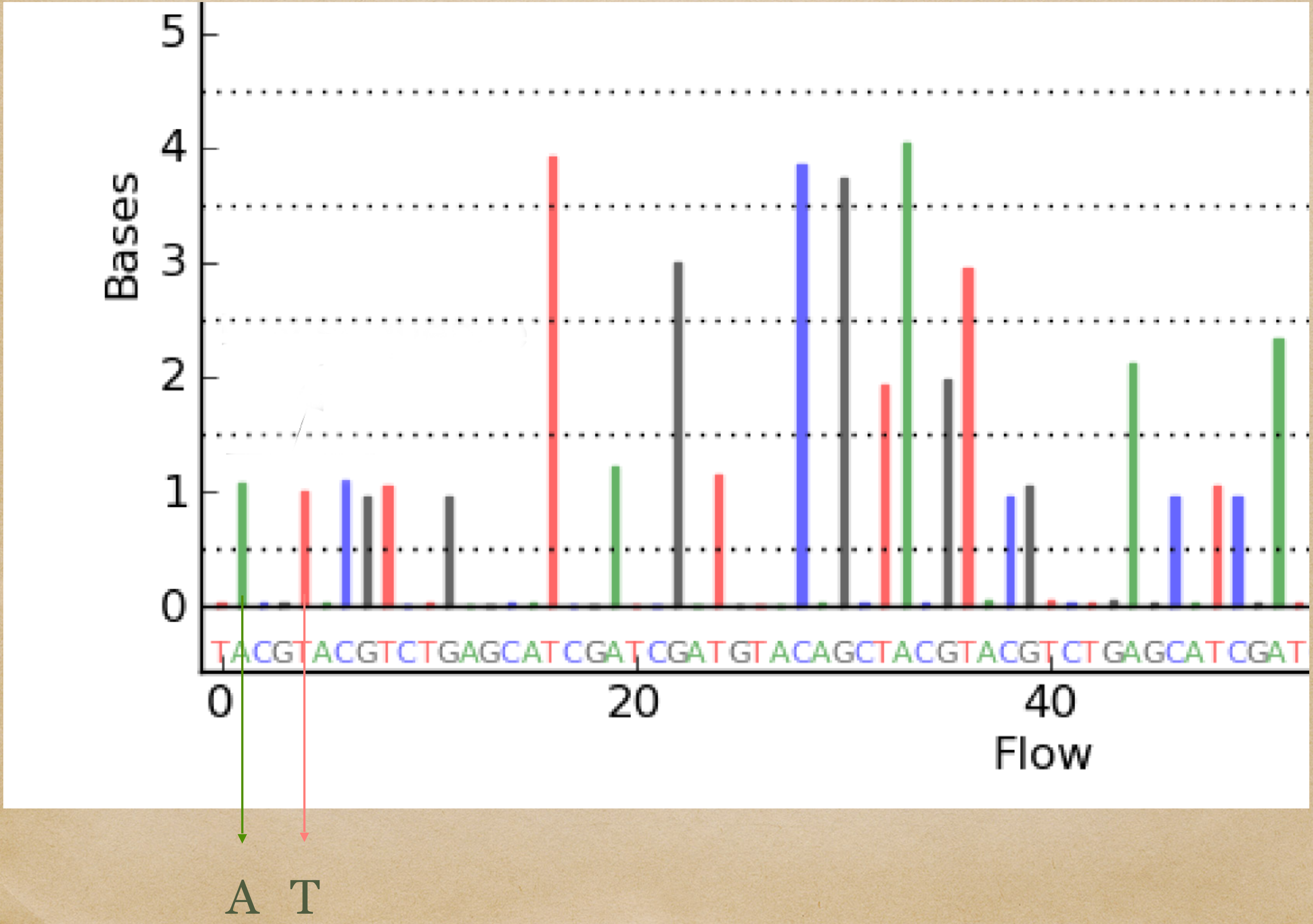
NGS -ION TORRENT



A

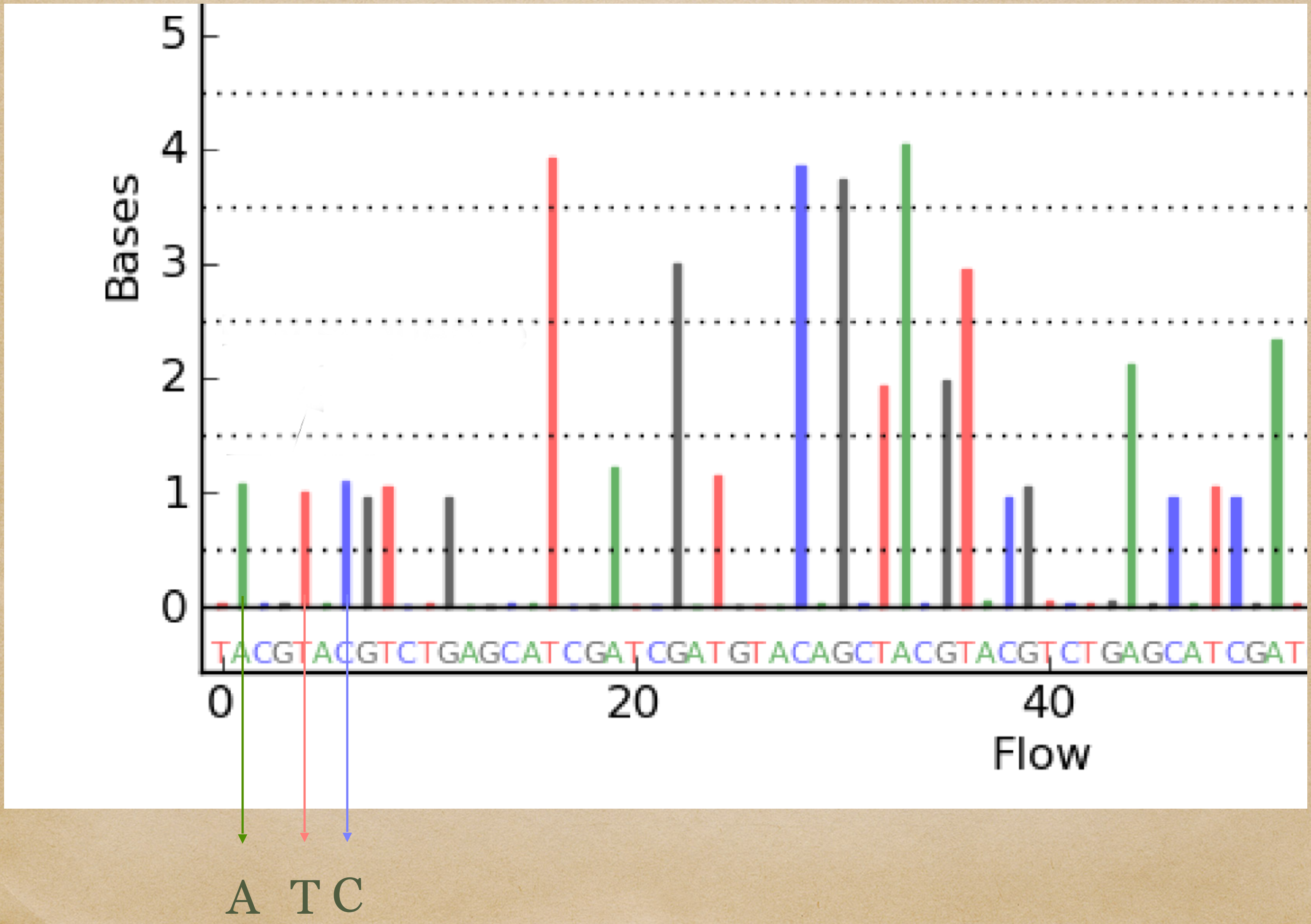
Base call

NGS -ION TORRENT



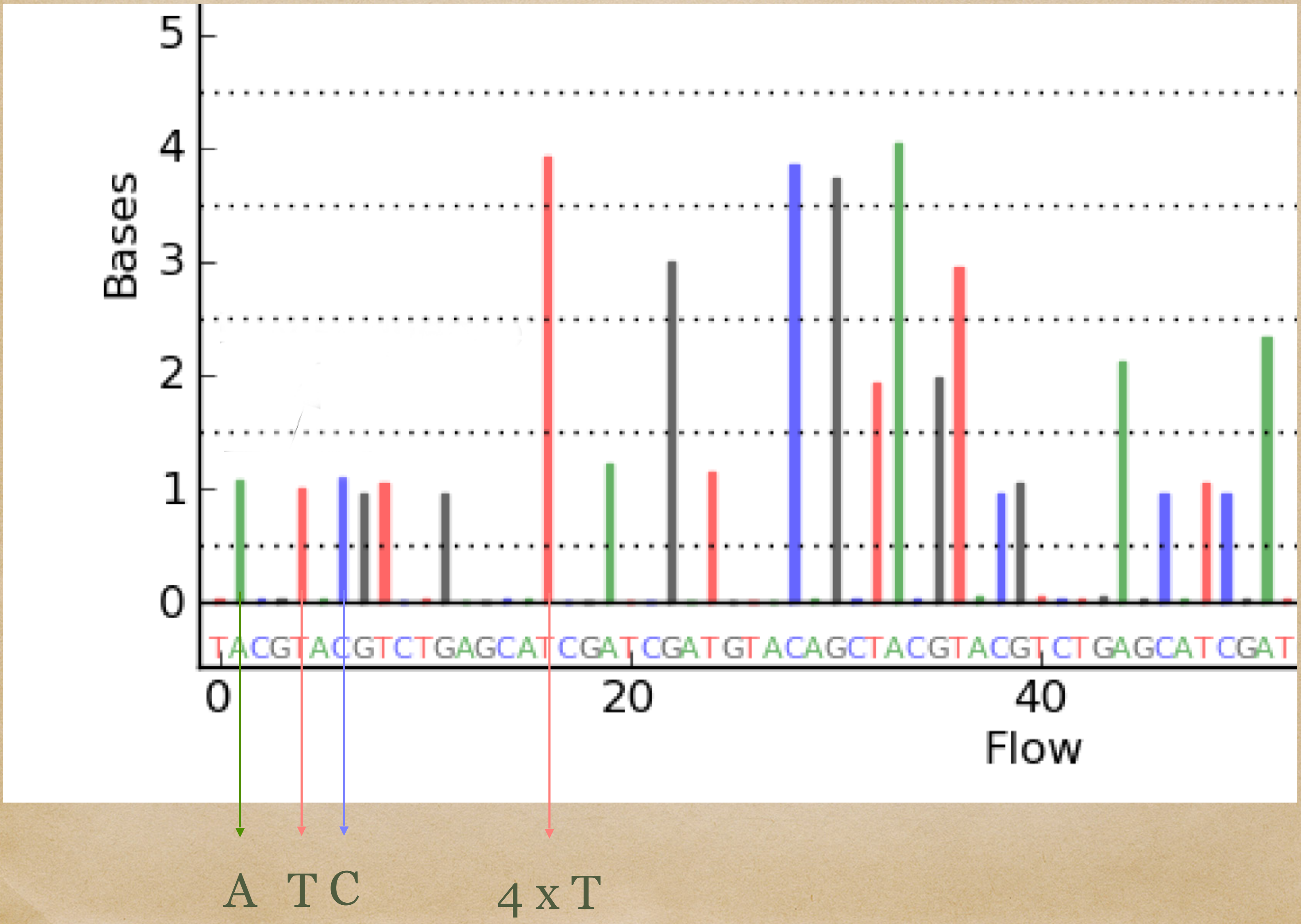
Base call

NGS -ION TORRENT



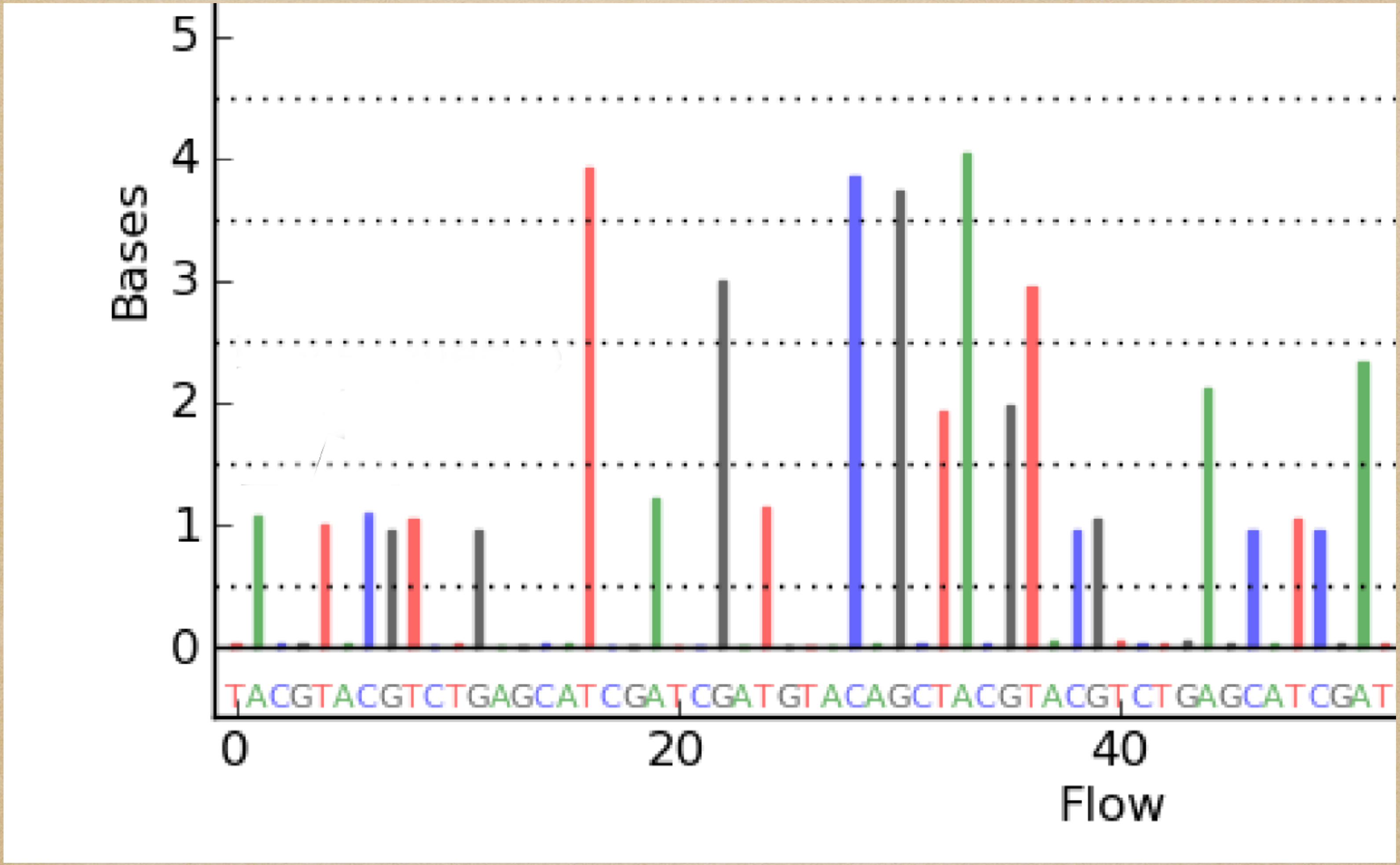
Base call

NGS -ION TORRENT



Base call

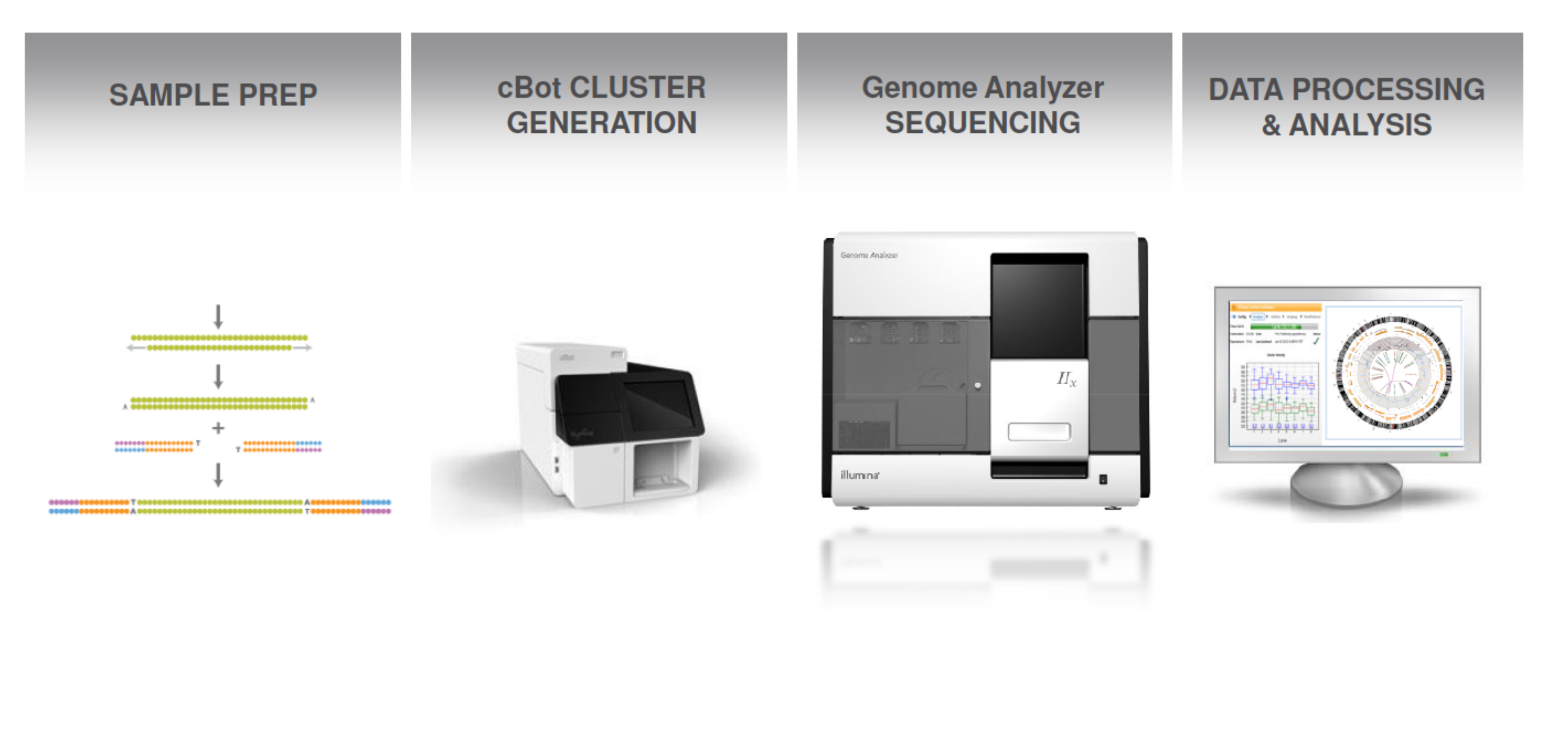
NGS -ION TORRENT



ATCGTGTTTTAGGGTCCCCGGGGTTAAAA...

NGS - Illumina

Workflow



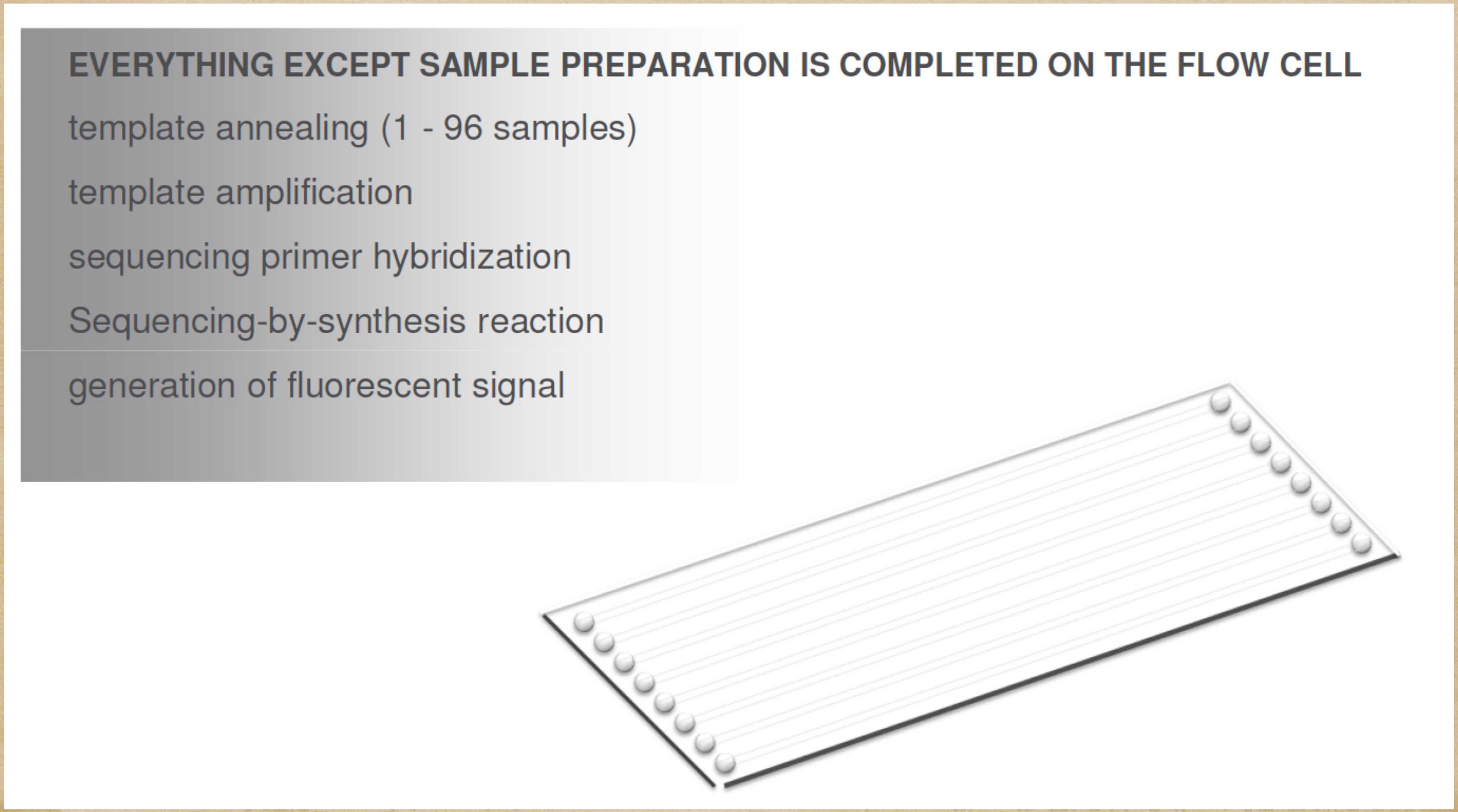
SAMPLE PREP

cBot CLUSTER GENERATION

Genome Analyzer SEQUENCING

DATA PROCESSING & ANALYSIS

NGS - Illumina
The flow cell - a core component



EVERYTHING EXCEPT SAMPLE PREPARATION IS COMPLETED ON THE FLOW CELL

template annealing (1 - 96 samples)

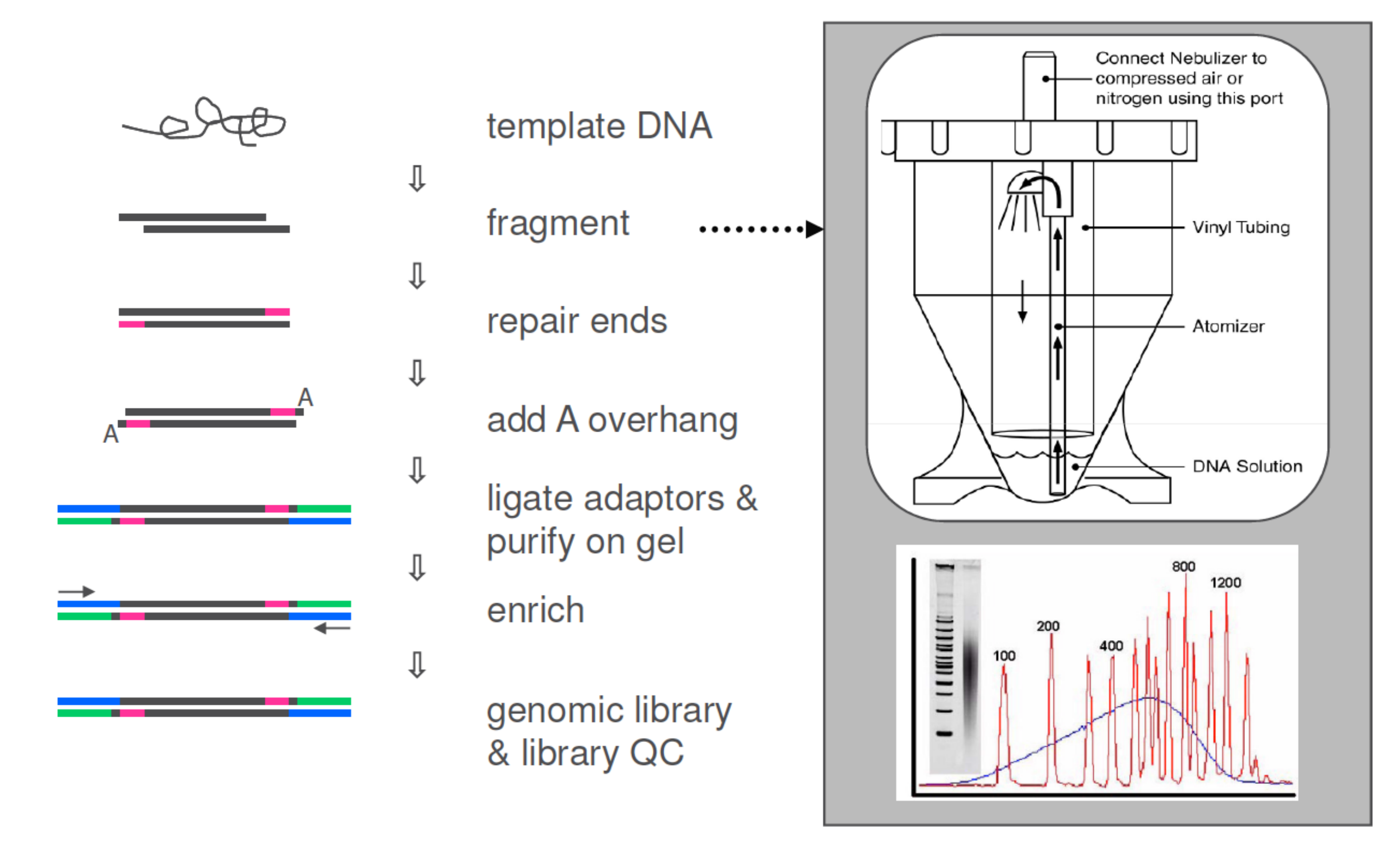
template amplification

sequencing primer hybridization

Sequencing-by-synthesis reaction

generation of fluorescent signal

NGS - Illumina
Preparation of template



template DNA

⇩

fragment

⇩

repair ends

⇩

add A overhang

⇩

ligate adaptors & purify on gel

⇩

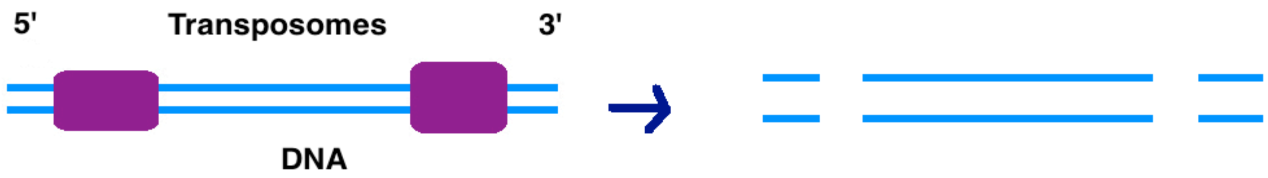
enrich

⇩

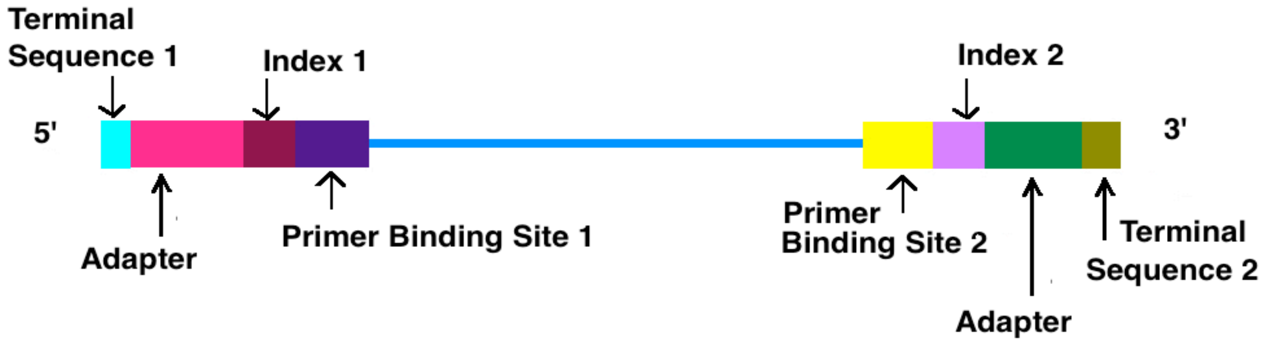
genomic library & library QC

A
A
Connect Nebulizer to compressed air or nitrogen using this port
Vinyl Tubing
Atomizer
DNA Solution
100
200
400
800
1200

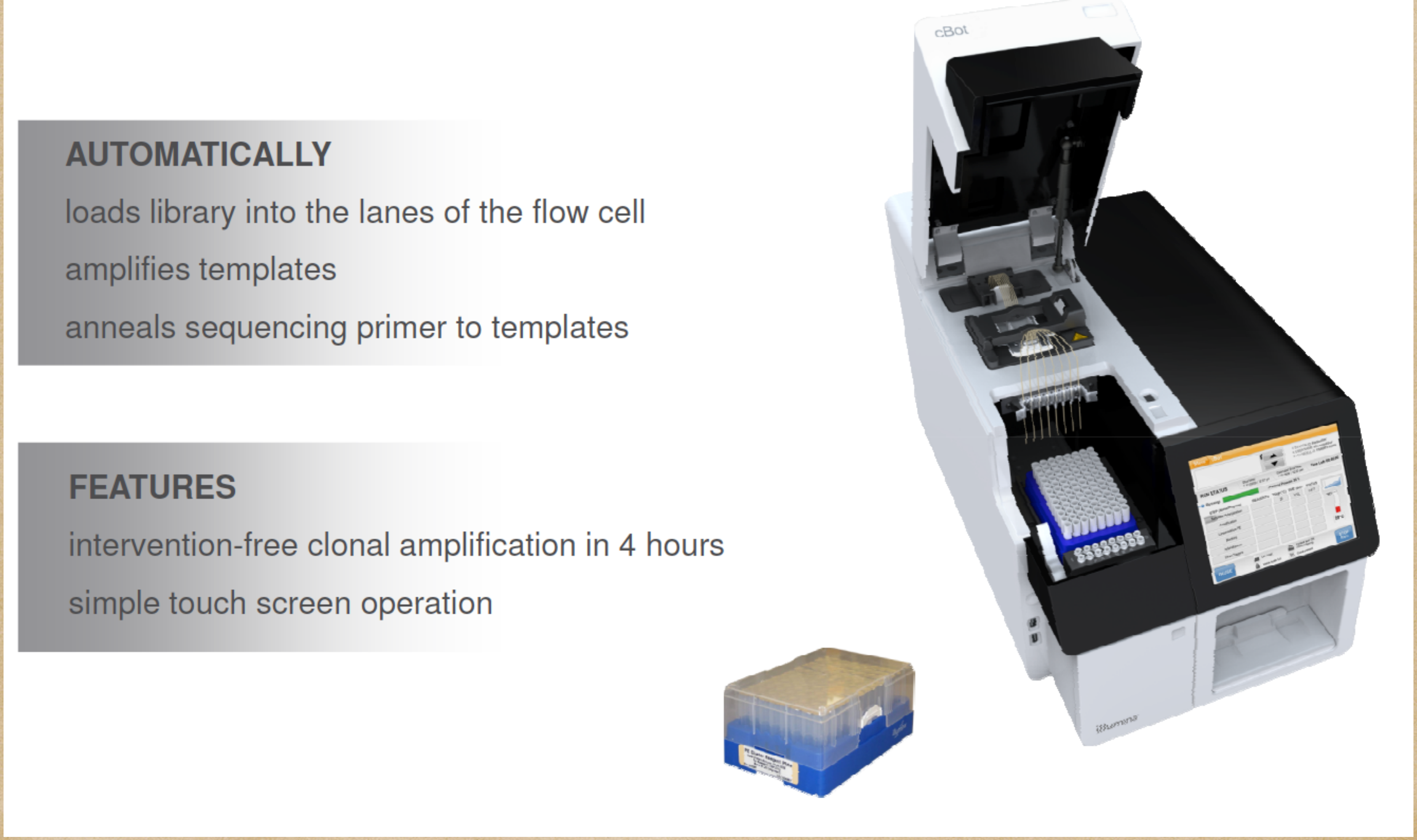
NGS - Illumina
Preparation of template



Another sheering method: transposomes – enzymes for DNA cleavage



NGS - Illumina
The flow cell is mounted on the cBot



AUTOMATICALLY

loads library into the lanes of the flow cell

amplifies templates

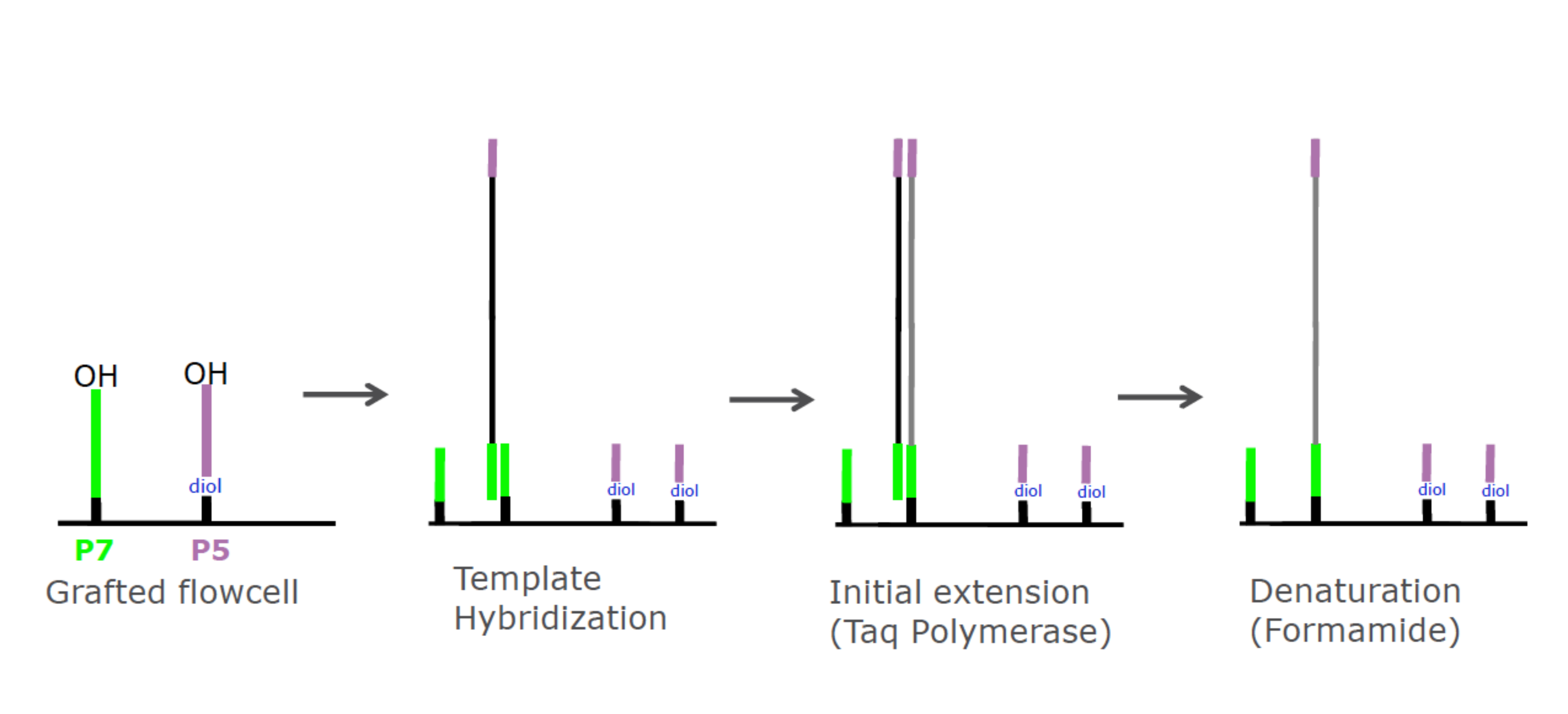
anneals sequencing primer to templates

FEATURES

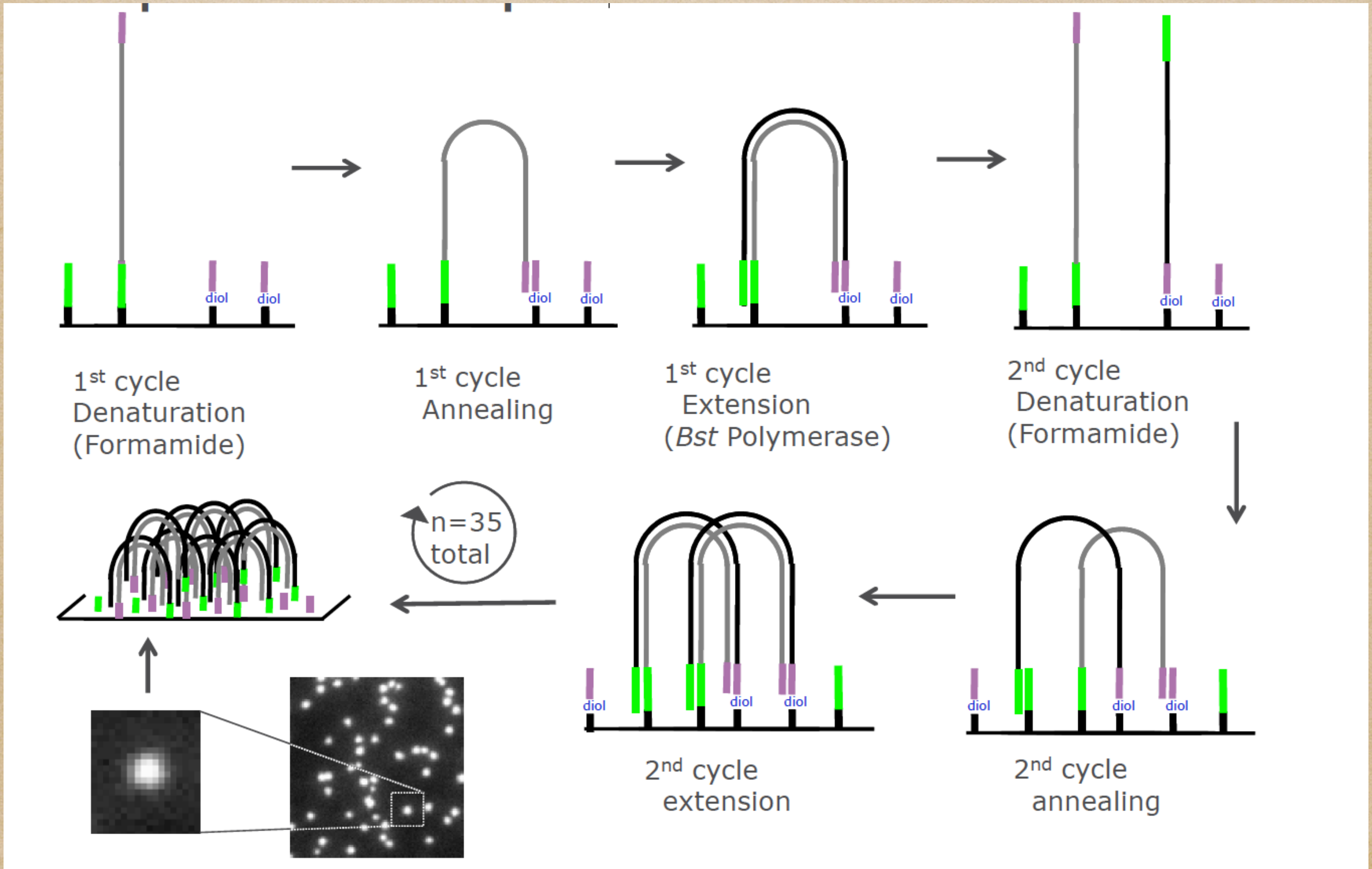
intervention-free clonal amplification in 4 hours

simple touch screen operation

NGS - Illumina
Hybridization of template

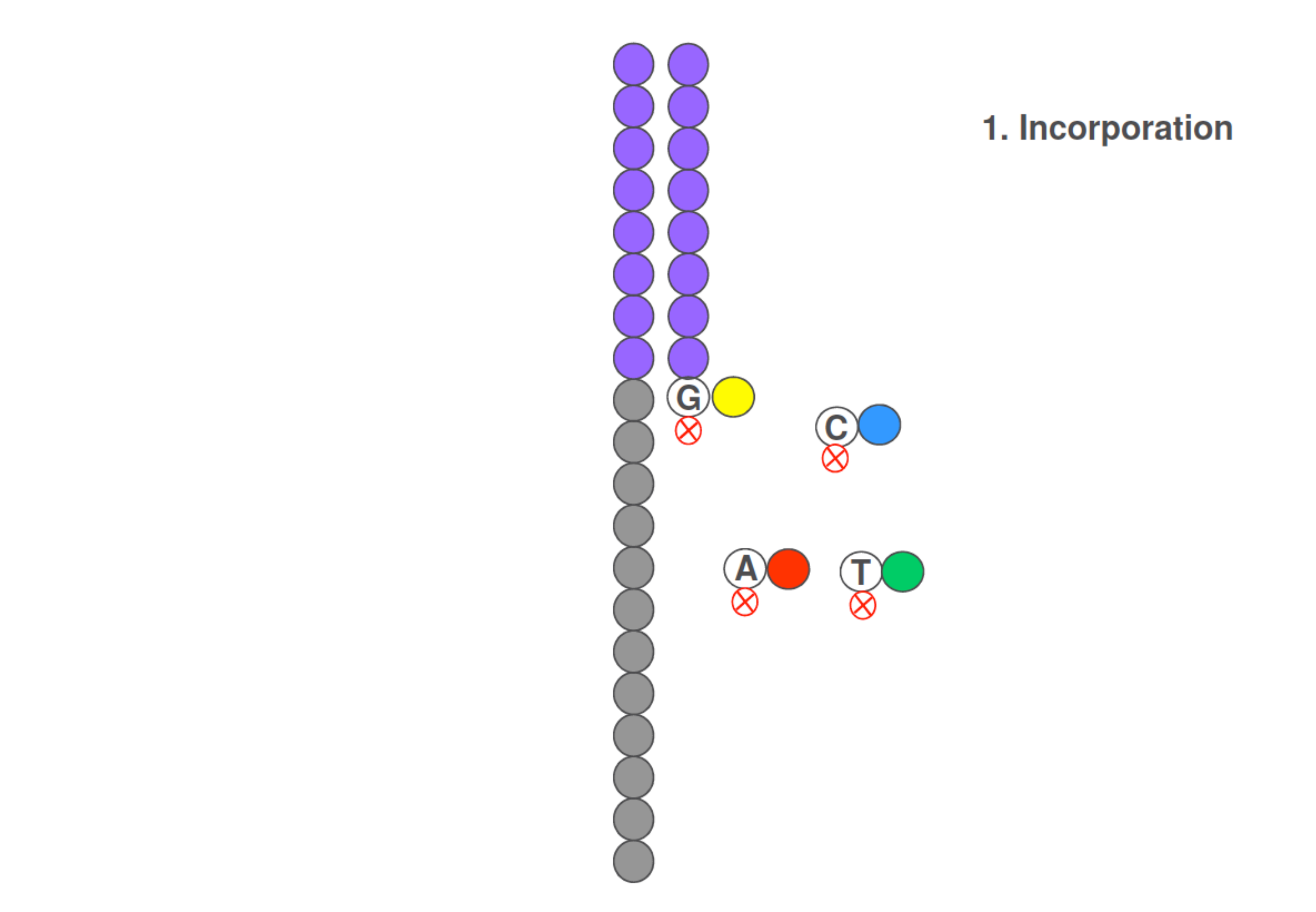


NGS - Illumina
Amplification of template



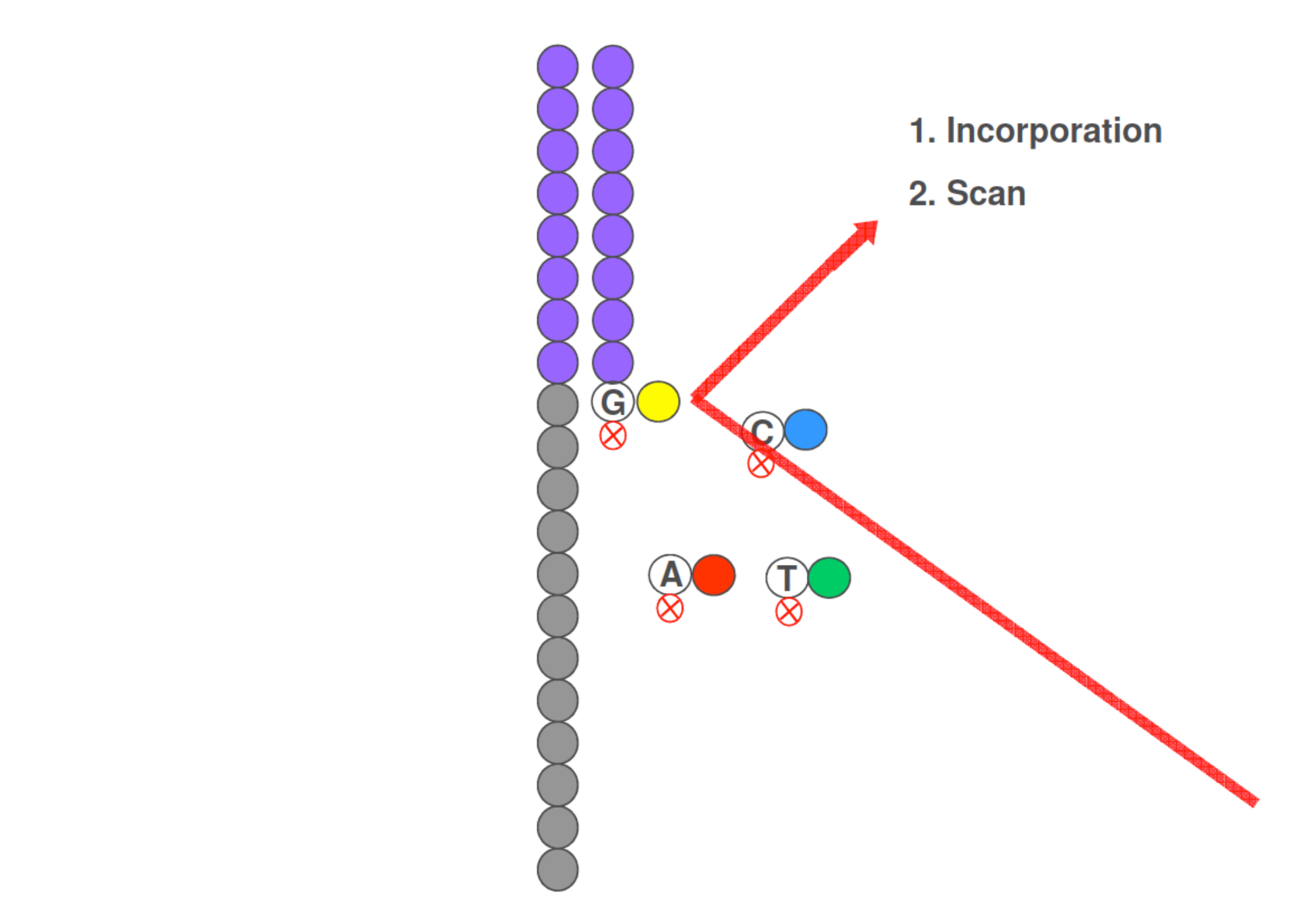
NGS - Illumina

Incorporation



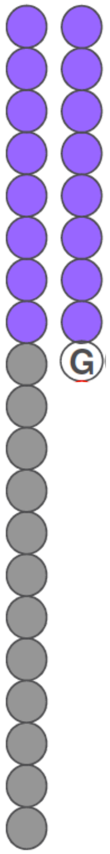
NGS - Illumina

Scanning



1. Incorporation
2. Scan

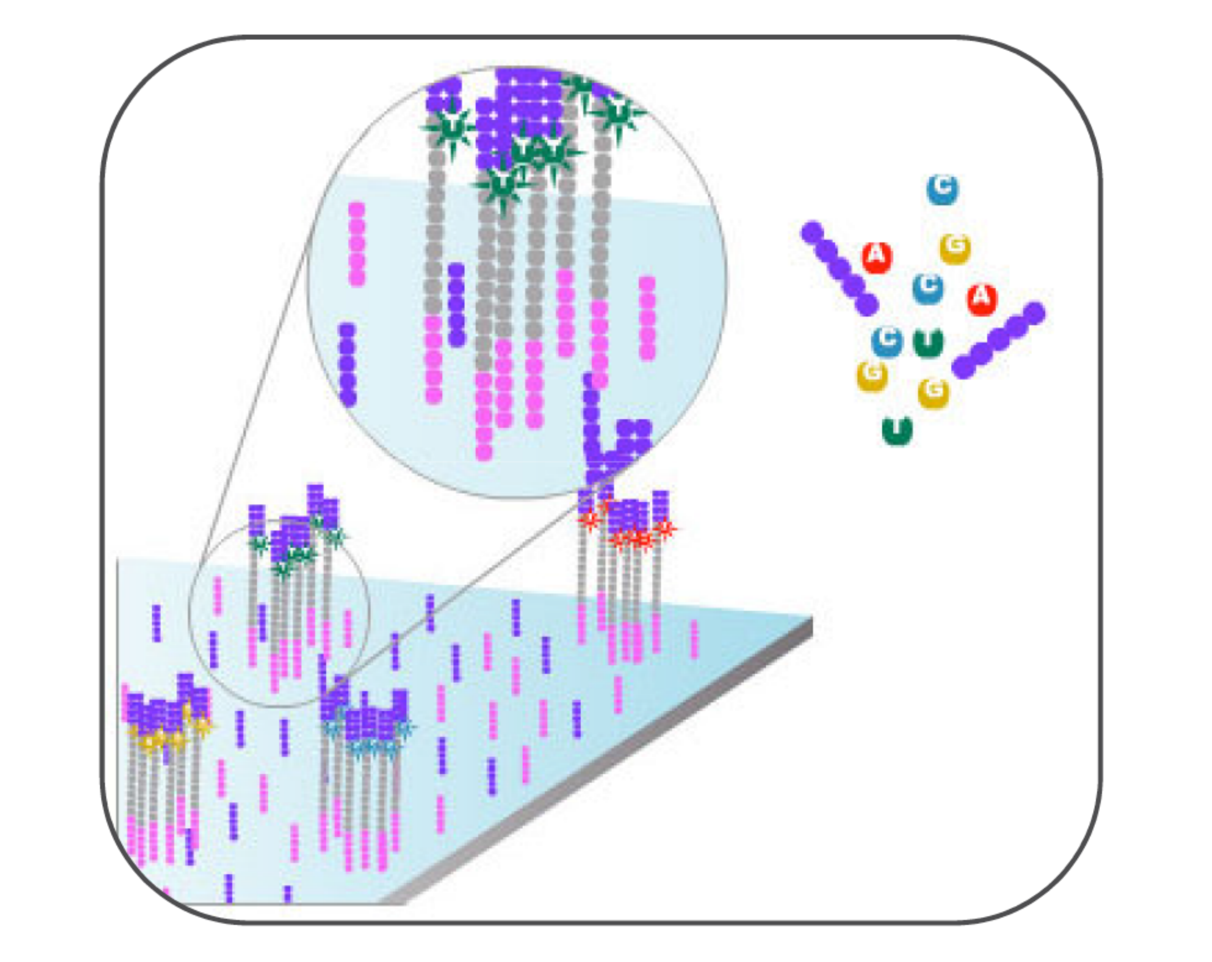
NGS - Illumina
Cleavage



1. Incorporation
2. Scan
3. Cleavage

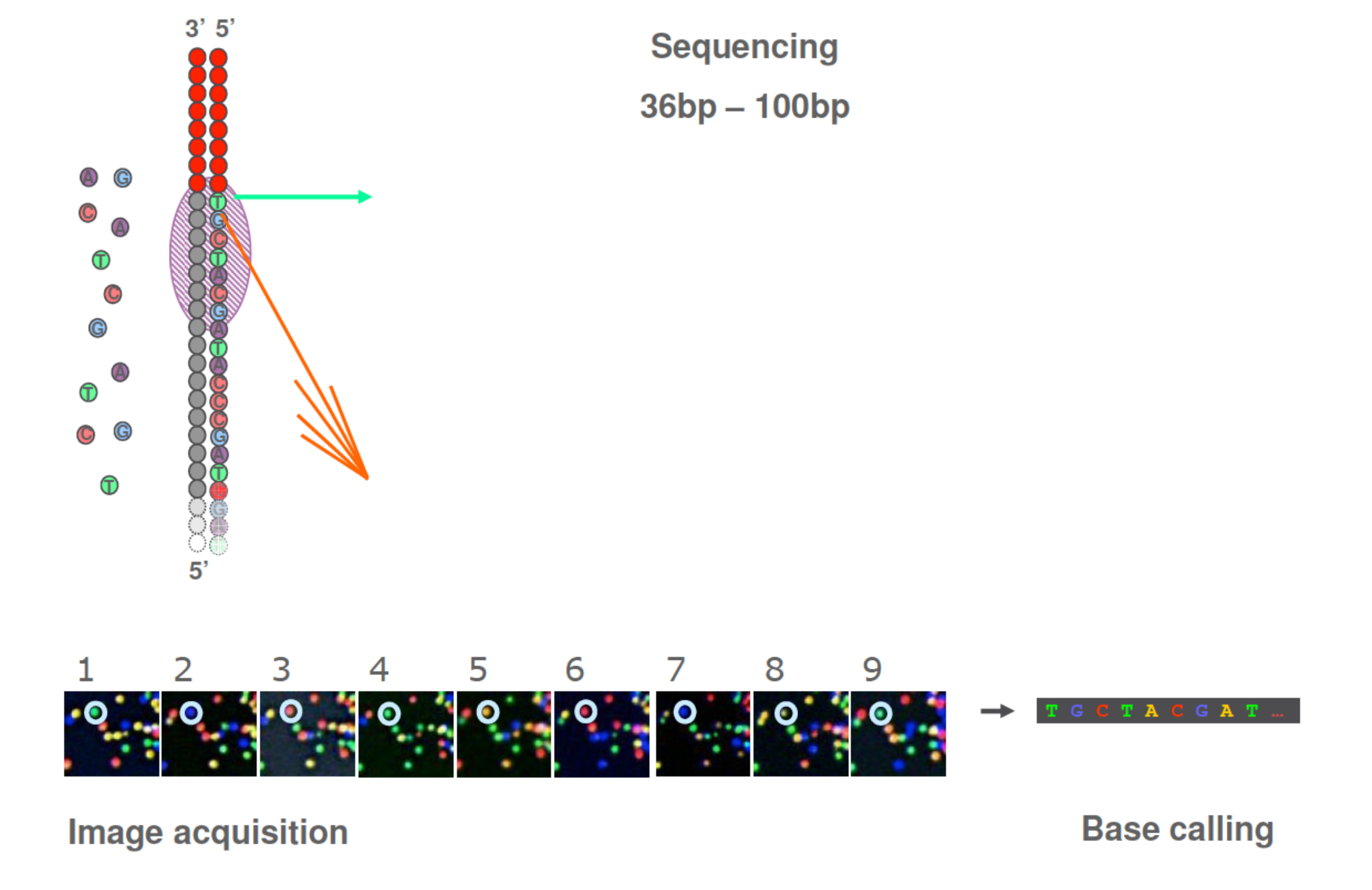
NGS - Illumina

Millions of clusters are sequenced in parallel



NGS - Illumina

A picture is taken every time a new base is added



Sequencing

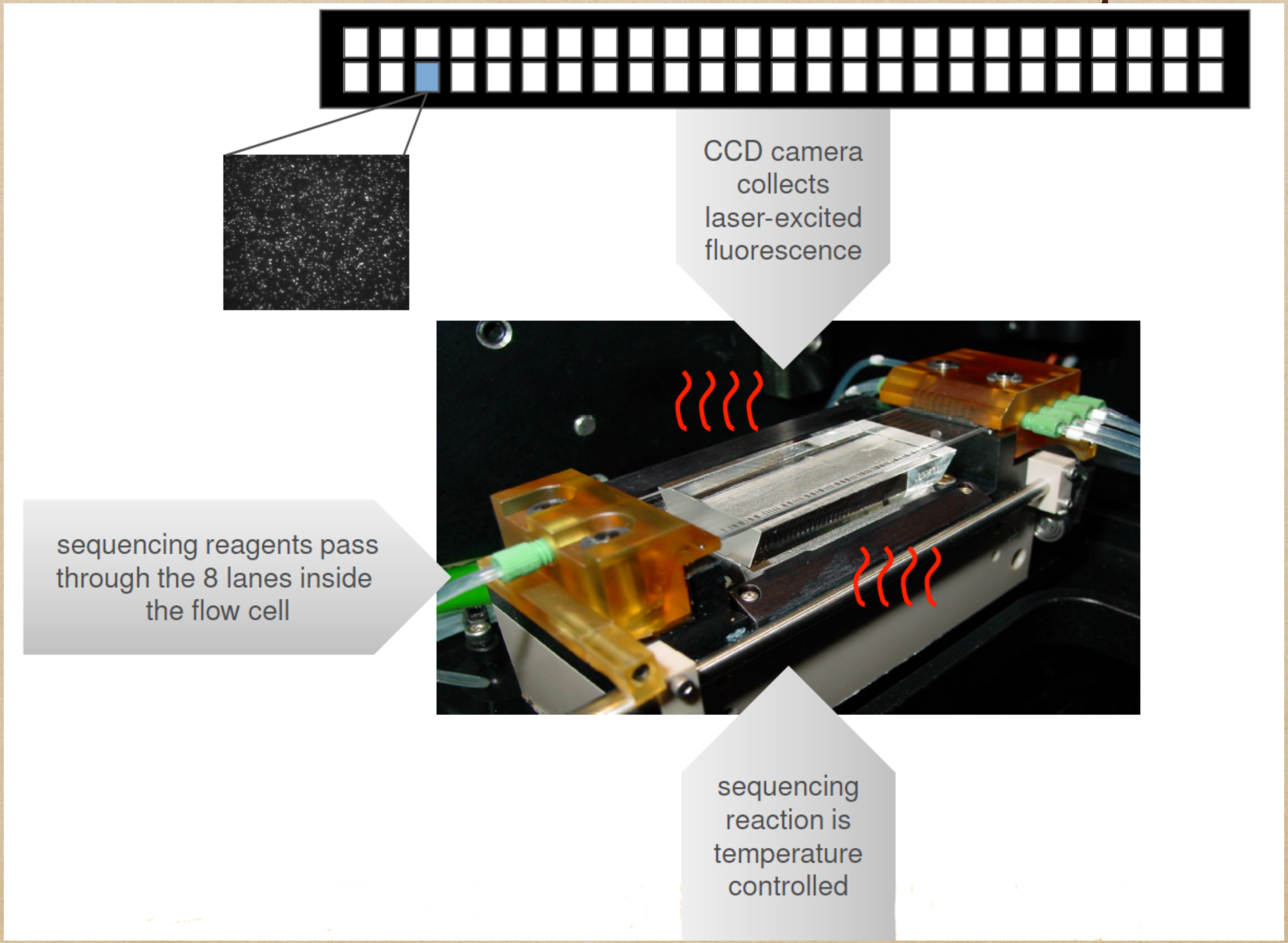
36bp – 100bp

Image acquisition

Base calling

NGS - Illumina

The flow cell is mounted on the sequencer



CCD camera collects laser-excited fluorescence
sequencing reagents pass through the 8 lanes inside the flow cell
sequencing reaction is temperature controlled

Third Generation Sequencing

1 – Pacific Bioscience (PacBio)

2 – MinIon (Oxford NanoTechnologies)

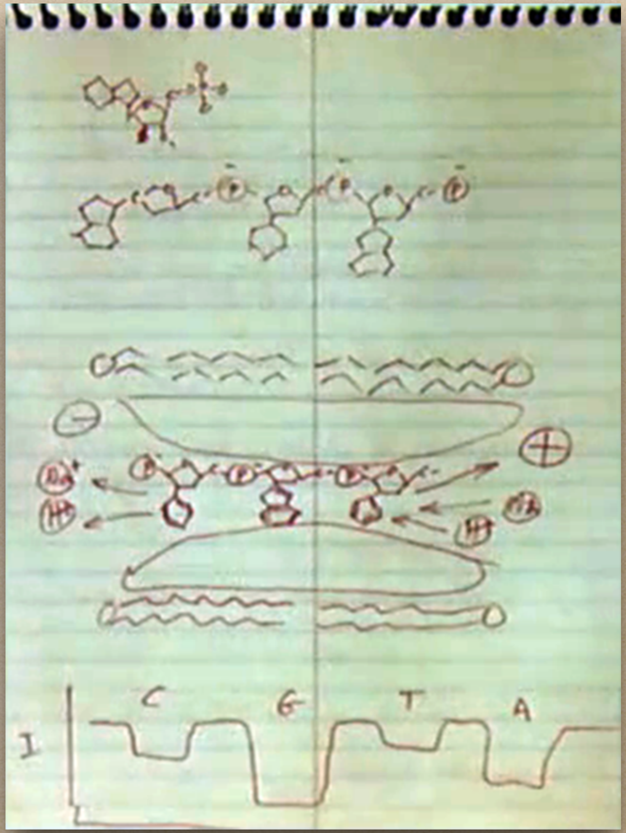
PacBio

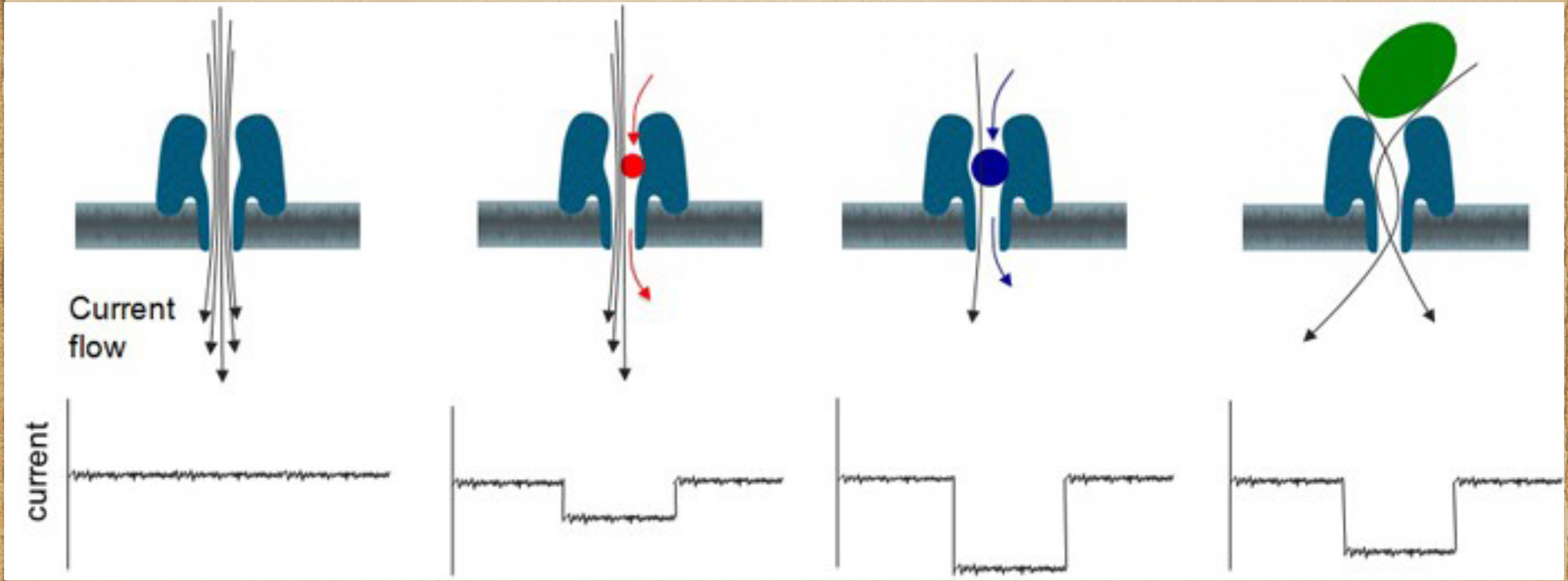
https://www.youtube.com/watch?v=_B_cUZ8hSYU

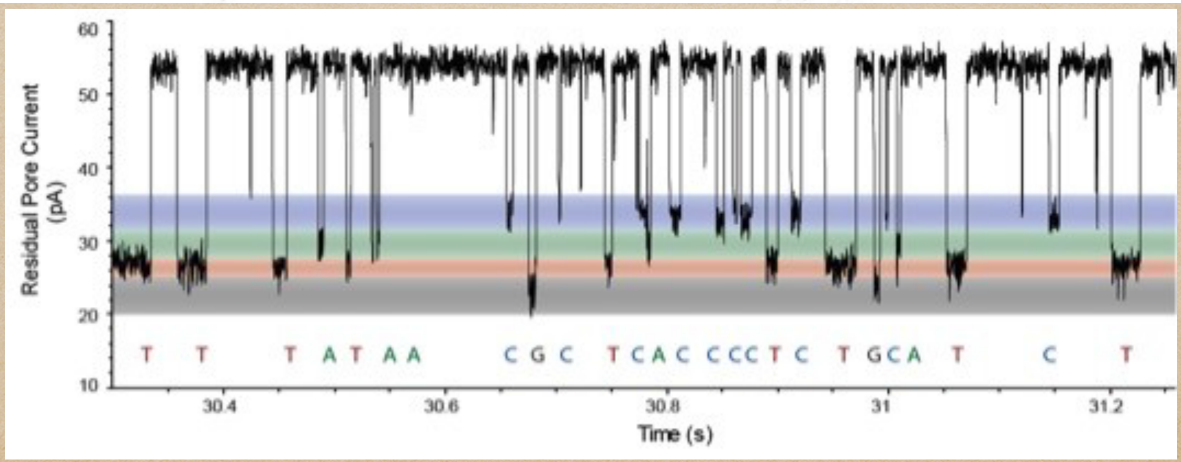
MinIon

MinIon: Sequencing using nanopores

- Nanopores as polymer sensors.
- The idea emerged in early 1990s.
- Fundamental work done by David Deamer and Daniel Branton in collaboration with John Kasianowicz. (PNAS 1996 146:13770-13773)
- Biologicaly relevant experiments – since 2010.



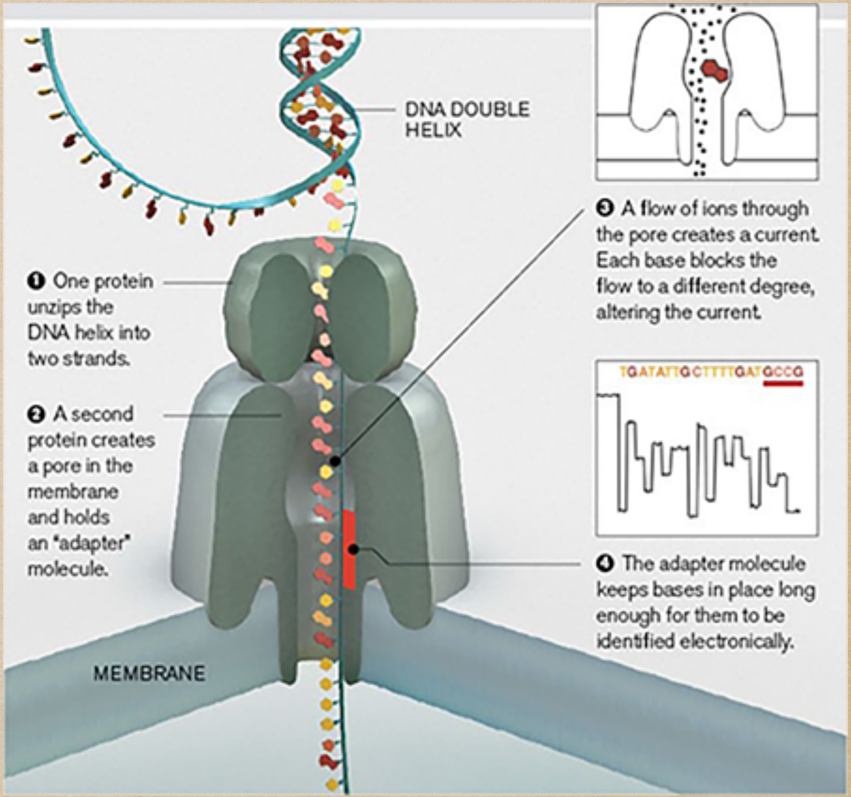

Current
flow
current


Residual Pore Current
(pA)
60
50
40
30
20
10
T T T A T A A C G C T C A C C C C T C T G C A T C T
30.4
30.6
30.8
31
31.2
Time (s)

MinION basics

https://nanoporetech.com/science-technology/introduction-to-nanopore-sensing/introduction-to-nanopore-sensing

- Synthetic membrane
- Nanopore (2) is created by modified protein pores: α-hemolysin, CsgG from E.coli
- Non-destructive motor protein (1) (actually serves as a break)



❶ One protein unzips the DNA helix into two strands.

❷ A second protein creates a pore in the membrane and holds an "adapter" molecule.

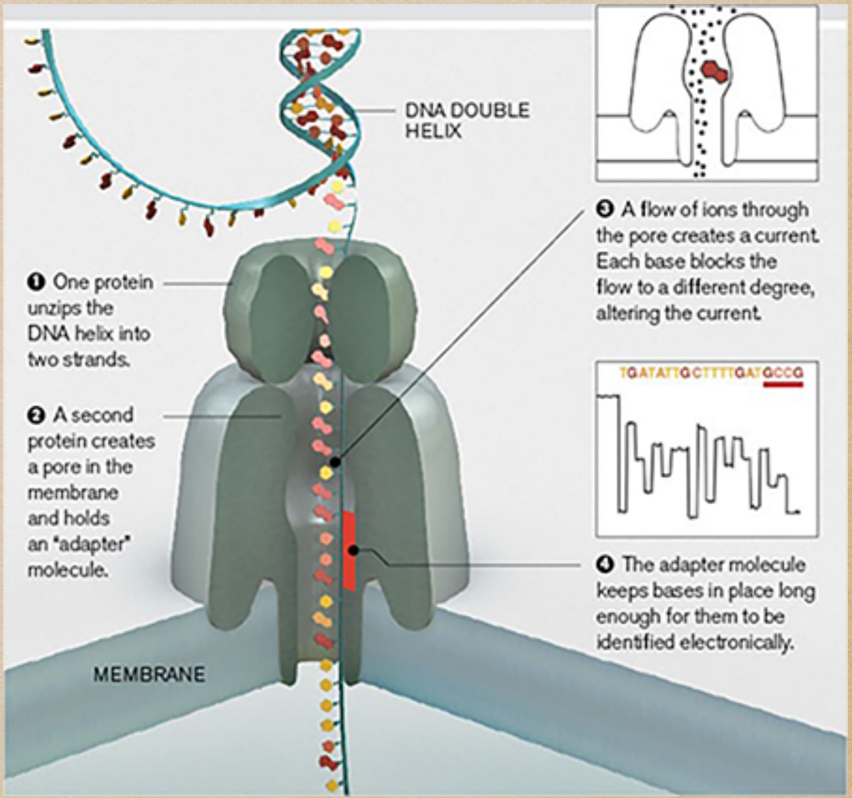
❸ A flow of ions through the pore creates a current. Each base blocks the flow to a different degree, altering the current.

❹ The adapter molecule keeps bases in place long enough for them to be identified electronically.

MinION basics

https://nanoporetech.com/science-technology/introduction-to-nanopore-sensing/introduction-to-nanopore-sensing

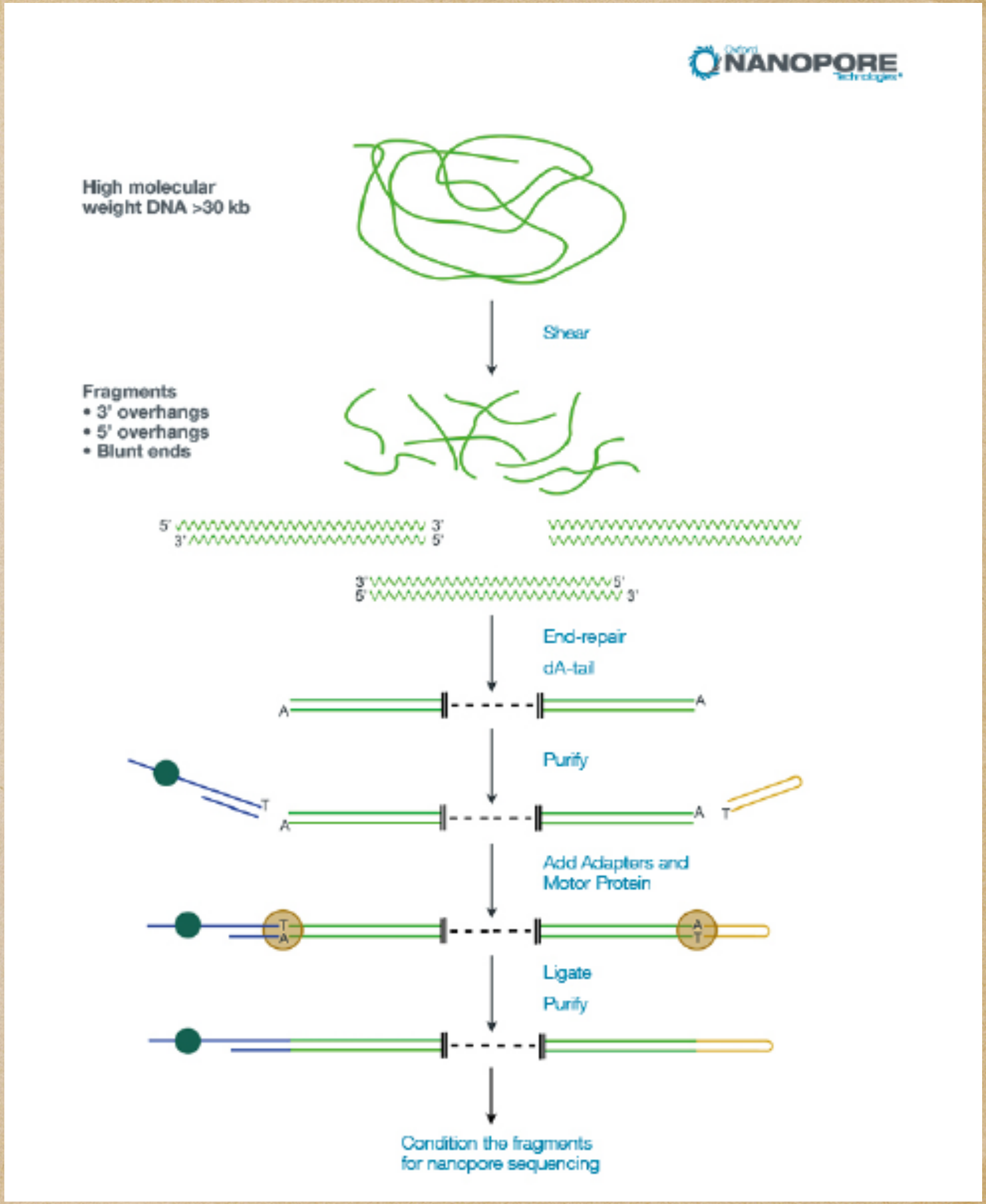
- 512 channels (pores) per flow cell. Usually about 90% are working.
- Read length: over a million of bp
- Read speed: 8 bases to 20 bases/sec
- Run time: max 48 hours
- Error rate = 5-10 %
- Sequence yield per flow cell: 15 Gb
- Complex algorithm for base calling using neural network approach


DNA DOUBLE HELIX
❶ One protein unzips the DNA helix into two strands.
❷ A second protein creates a pore in the membrane and holds an "adapter" molecule.
❸ A flow of ions through the pore creates a current. Each base blocks the flow to a different degree, altering the current.
❹ The adapter molecule keeps bases in place long enough for them to be identified electronically.
TGATATTGCTTTTGATGCCG
MEMBRANE

Easy, standard template preparation

Time of library preparation:
1D - about ten minutes
2D - up to two hours

Cost of a single run:
reagents $200
flow cell $1000


Oxford NANOPORE Technologies
High molecular weight DNA >30 kb
Shear
Fragments
• 3' overhangs
• 5' overhangs
• Blunt ends
5'
3'
3'
5'
3'
5'
5'
3'
End-repair
dA-tail
A
A
Purify
T
A
A
T
Add Adapters and Motor Protein
T
A
A
T
Ligate
Purify
Condition the fragments for nanopore sequencing

MinION dataflow

MinION – the device

Nanopore sensing is carried out on the sensor chip, contained in the flow cell inside the MinION device. Data is processed by an Application-Specific Integrated Circuit (ASIC) also in the flow cell and processed in real time by the MinKNOW software

MinKNOW – the software

MinKNOW is the software that controls the MinION. It carries out several core data tasks and can be used to change experimental workflows or parameters. MinKNOW runs on the user's computer.

ALBACORE – base calling

Albacore is a command-line (some programming skills are required) base-calling software, developed for MinIon and accounts for specific sequencing errors

Numerous applications explored by MinION Access Program (MAP)

- Genomic DNA sequencing
- Metagenomic analysis
- Medical diagnostics (in development)
- Species identification in the field
- Splice variants identification
- Virus detection in the field
- Sequencing in space, etc … ☺

Comparison table

	454	Illumina	Ion Torrent	PacBio	MinIon
Method all sequence by synthesis	Pyrosequencing: pyrophosphates detection by chemoluminicent reaction (luciferase enzyme). Detector: CCD camera	Bridge amplification; detection of fluorescently labeled nucleotides. Detector: CCD camera	Ion semiconductor: label free detection of released protons. Detector: ion sensor	Single-molecule in real-time: detection of fluorescently labeled cleaved pyrophosphates. Detector: ZMW camera (sensitive!)	Nanopores: modified pore proteins detect current change when different nucleotides pass the pore. Detector: ASIC -measures ionic current flow

454: https://www.youtube.com/watch?v=nFfgWGFe0aA

Illumina: https://www.youtube.com/watch?v=fCd6B5HRaZ8

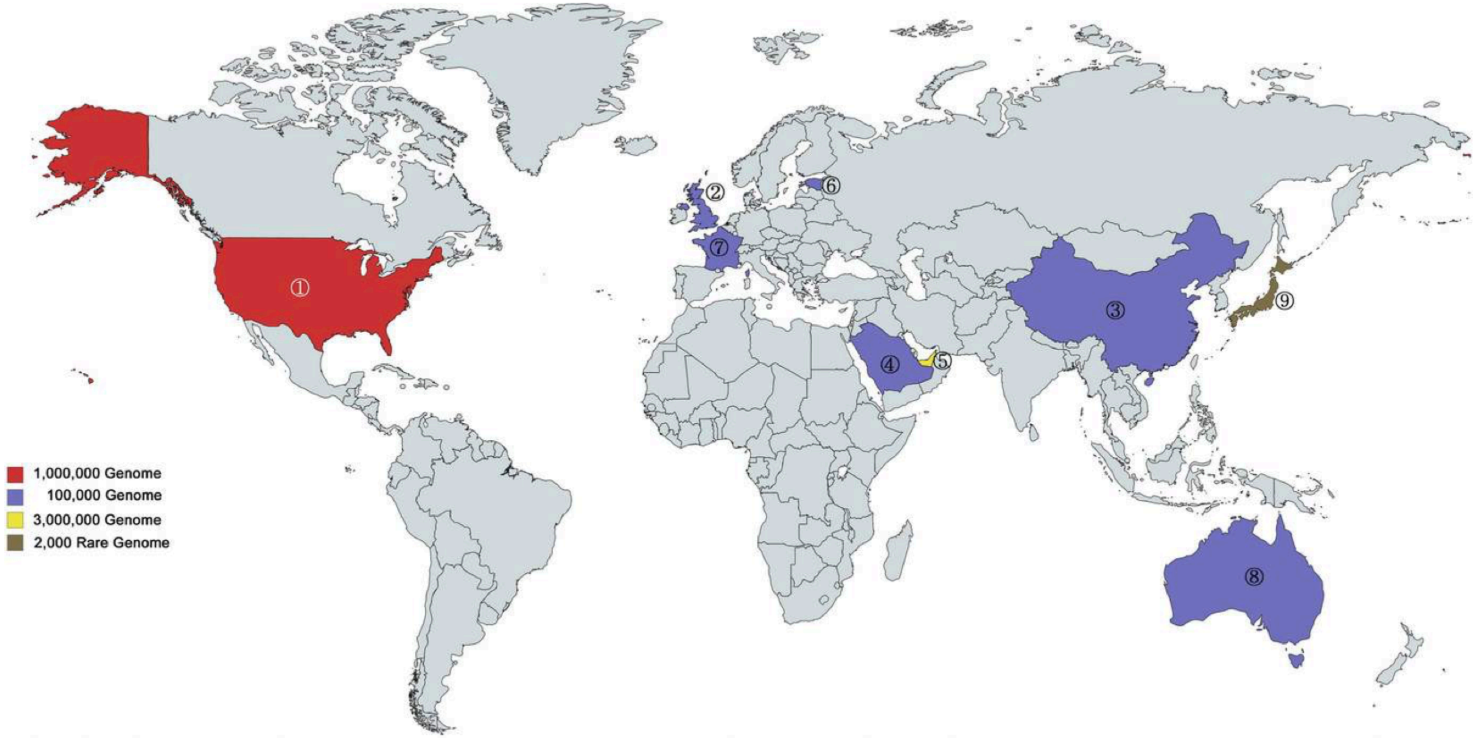
Ion Torrent: https://www.youtube.com/watch?v=WYBzbxIfuKs

PacBio: https://www.youtube.com/watch?v=_B_cUZ8hSYU

MioIon: https://nanoporetech.com/how-it-works

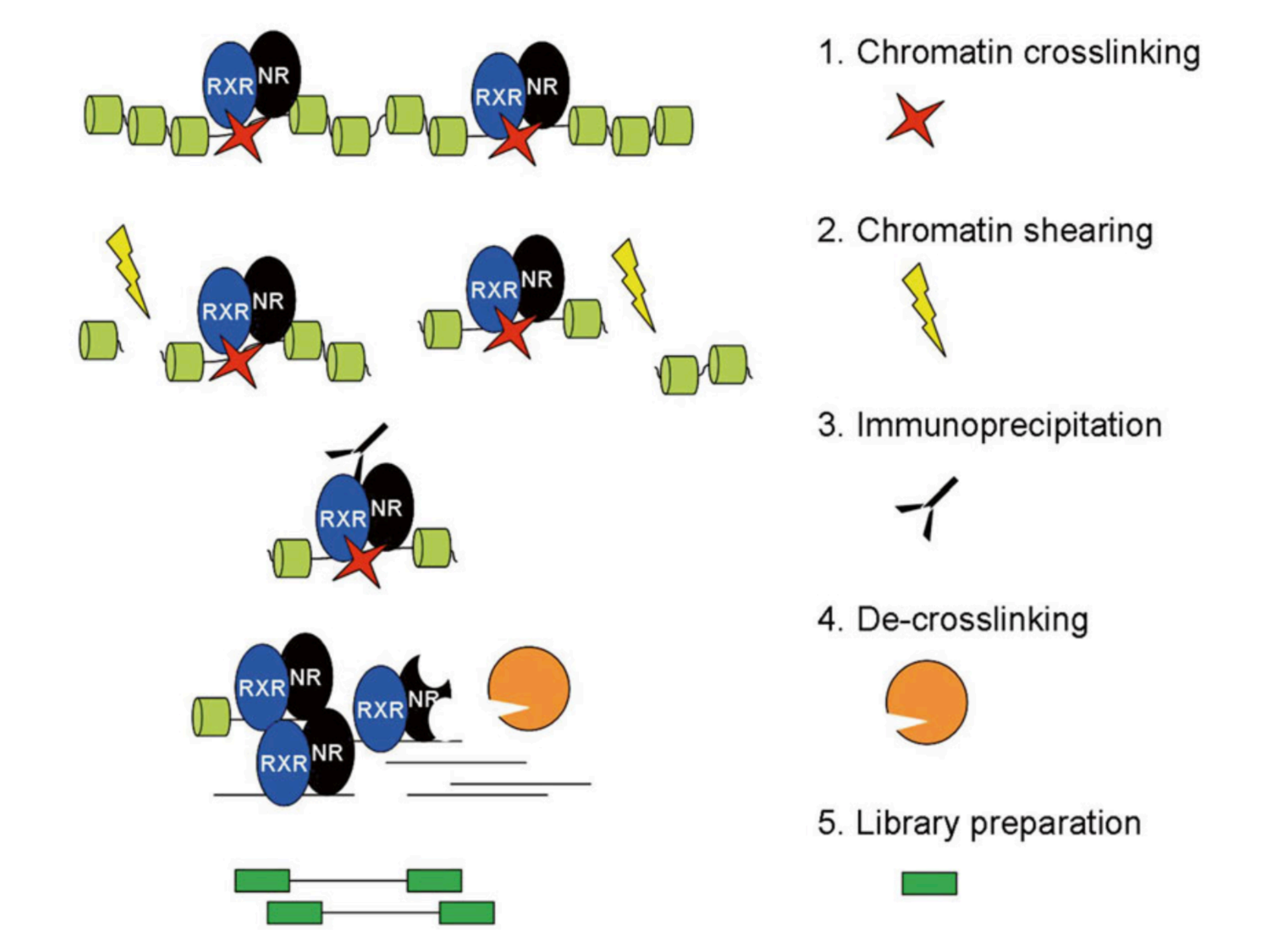
Comparison table

	454	Illumina	Ion Torrent	PacBio	MinIon
Read length	700 bp	50-250 bp	200 bp	3000-15000 bp	500-100000
Reads per run	1 million	up to 3 billion	up to 5 million	35000-75000	30-400 million
Time per run	24 hours	1-10 days	2 hours	30 min – 2 hours	6-48 hours
Cost per million bases	10$	0.05-0.15$	1$	2$	2$
Machine cost		120.000-650.000$	80.000$	695.000$	1500$
Error rate	0.1-1%	0.5-1%	1-2%	12%	5-10%



1.USA : 1,000,000 Genome @ Veterans Project & All of Us Reserach Program | 2.United Kingdom: 100,000 Genomes project | 3.China: 100,000 genomes project | 4.Saudi Arabia: 100,000 Genome Project (Saudi Genome) | 5.United Arab Emirates: 3,000,000 Genome Project | 6.Estonia: 100,000 Genome Project (Personalized Medicine Program) 7.France: 100,000 Genome Project (French Plan for Genomic Medicine 2025) | 8.Australia: 100,000 Genome Project (The Australian Genomics Health Futures Mission) 9.Japan: 2,000 Genome Project (Initiatives on Rare and Undiagnosed Diseases)

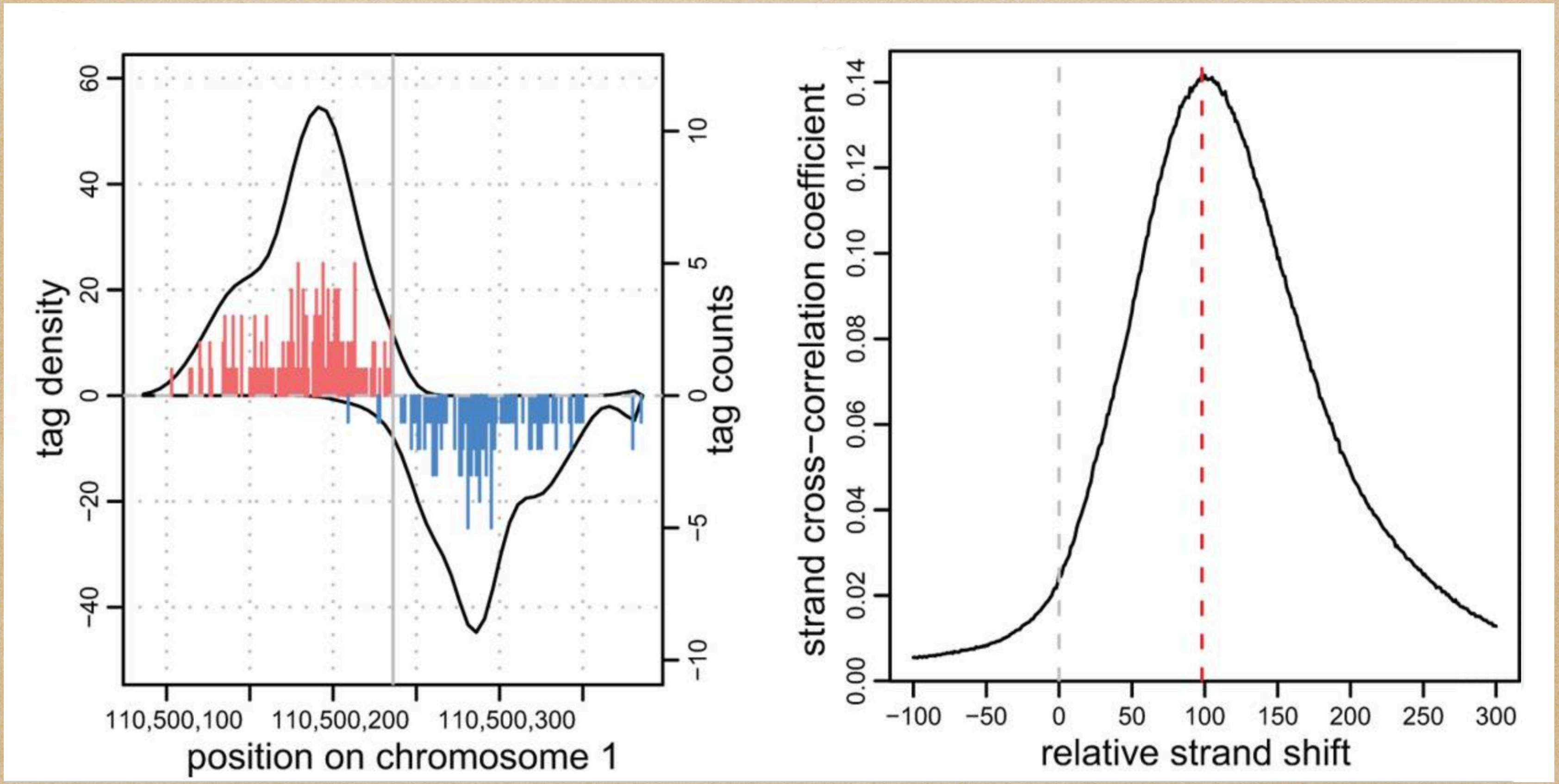
chip-seq experiments



1. Chromatin crosslinking
2. Chromatin shearing
3. Immunoprecipitation
4. De-crosslinking
5. Library preparation

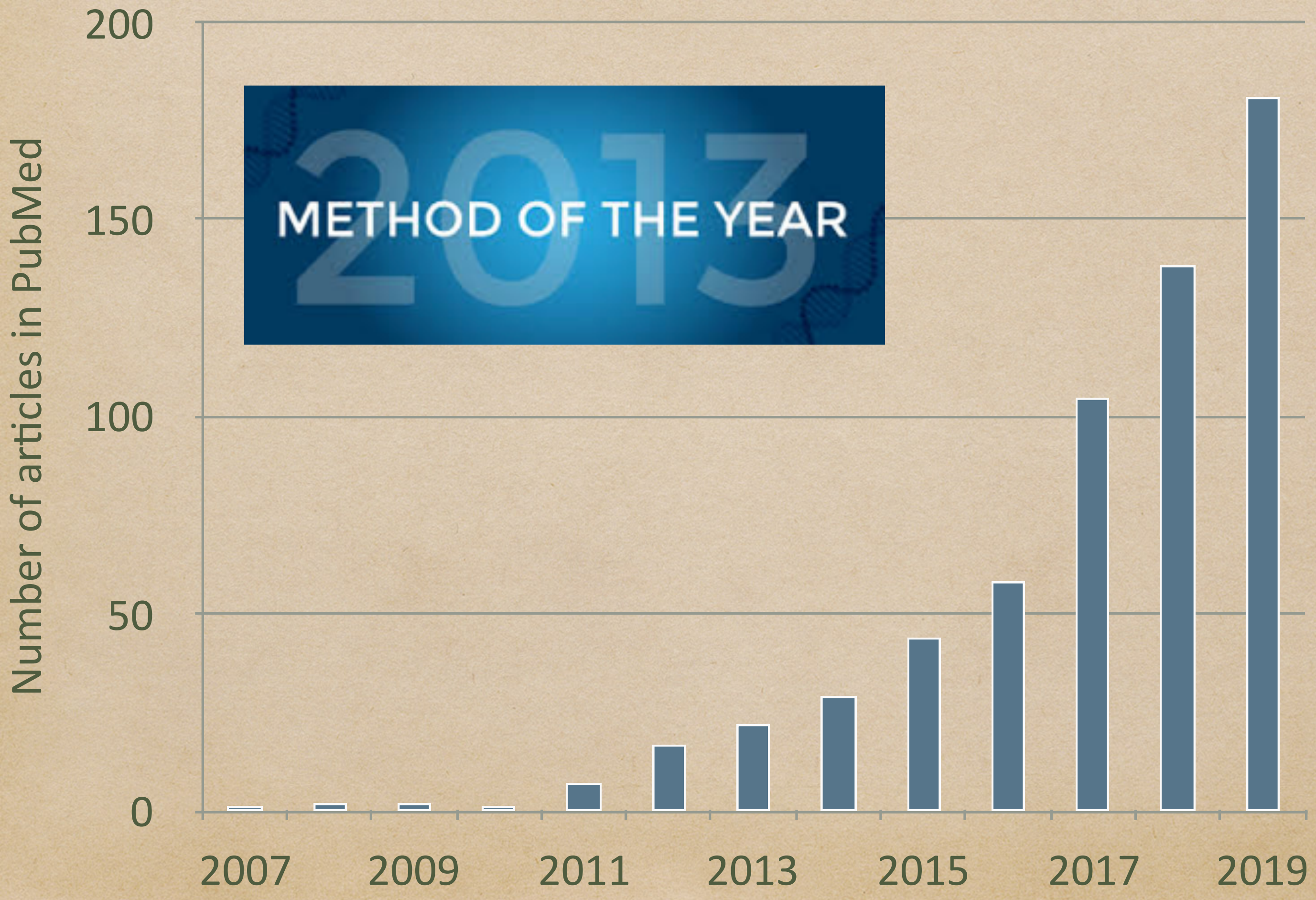
Methods in Molecular Biology 1204:15-24 (2014)

chip-seq experiments



Nat Biotechnol. 2008 Dec; 26(12): 1351–1359.

Single-cell sequencing



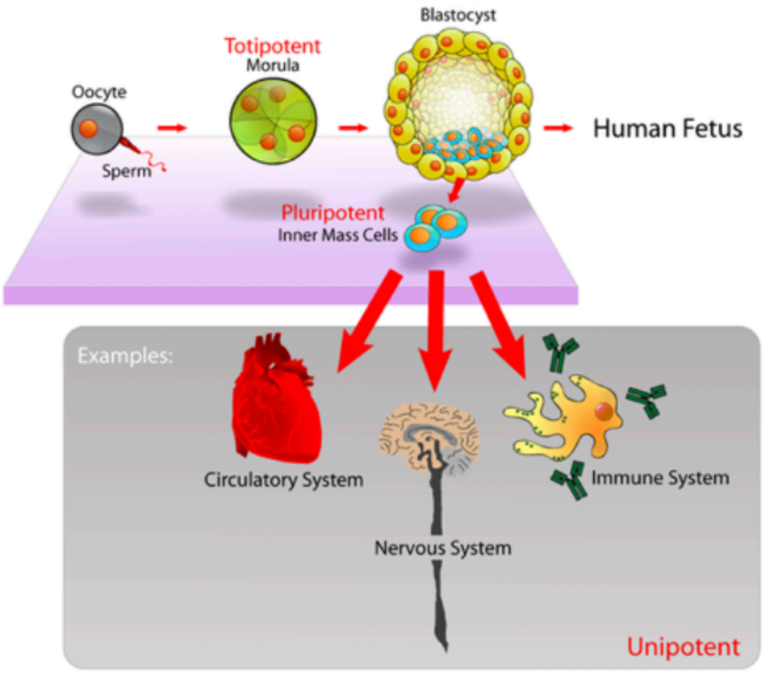
Single-cell sequencing

applications

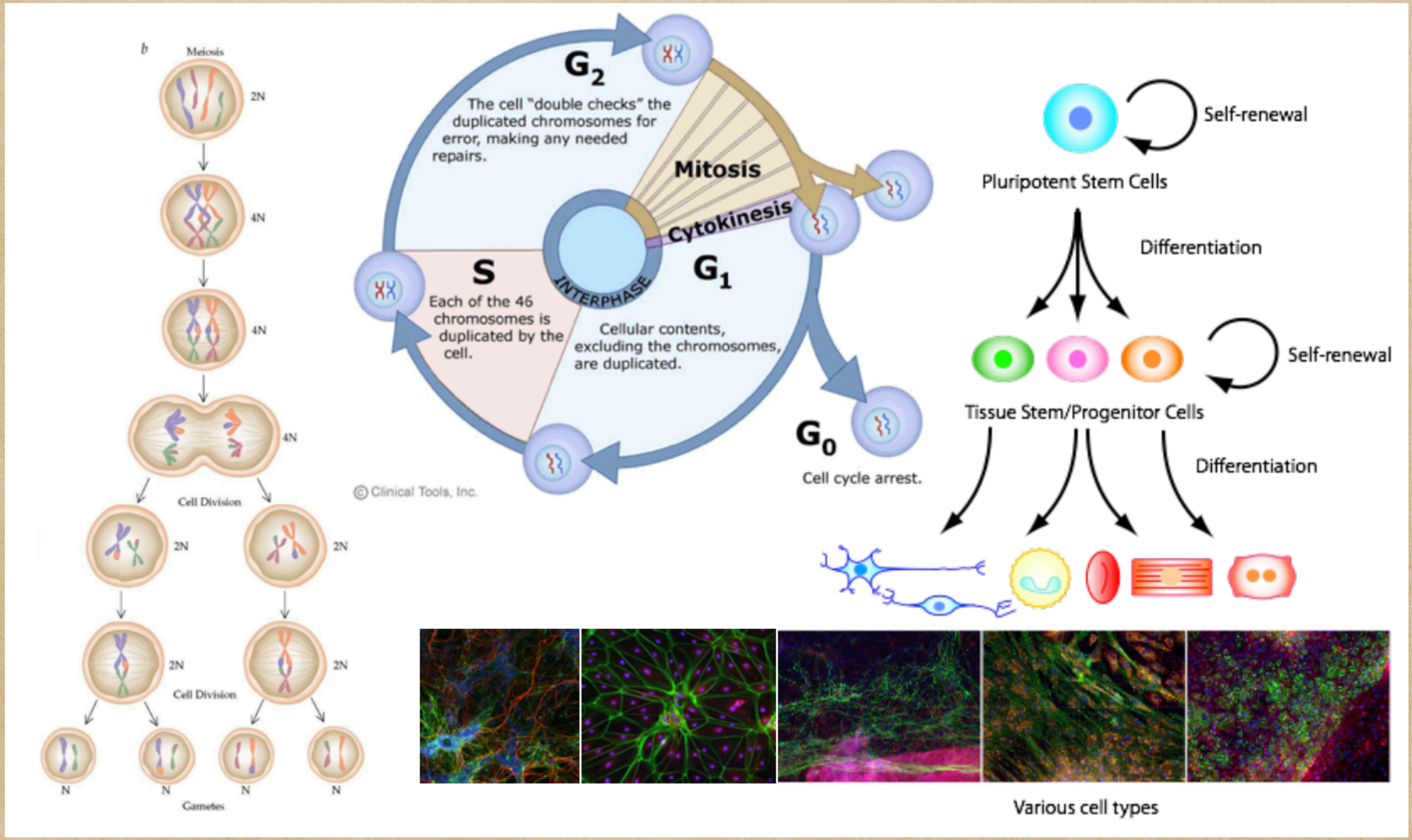
- Developmental Biology
- Cancer Biology
- Microbiology
- Neurology

Developmental Biology

How do animals grow and develop from a single cell?



Developmental Biology



b

Meiosis

2N

4N

4N

4N

Cell Division

2N

2N

2N

Cell Division

2N

N

N

N

N

Gametes

G_2

The cell "double checks" the duplicated chromosomes for error, making any needed repairs.

Mitosis

Cytokinesis

S

Each of the 46 chromosomes is duplicated by the cell.

INTERPHASE

G_1

Cellular contents, excluding the chromosomes, are duplicated.

G_0

Cell cycle arrest.

© Clinical Tools, Inc.

Self-renewal

Pluripotent Stem Cells

Differentiation

Self-renewal

Tissue Stem/Progenitor Cells

Differentiation

Various cell types

Developmental Biology

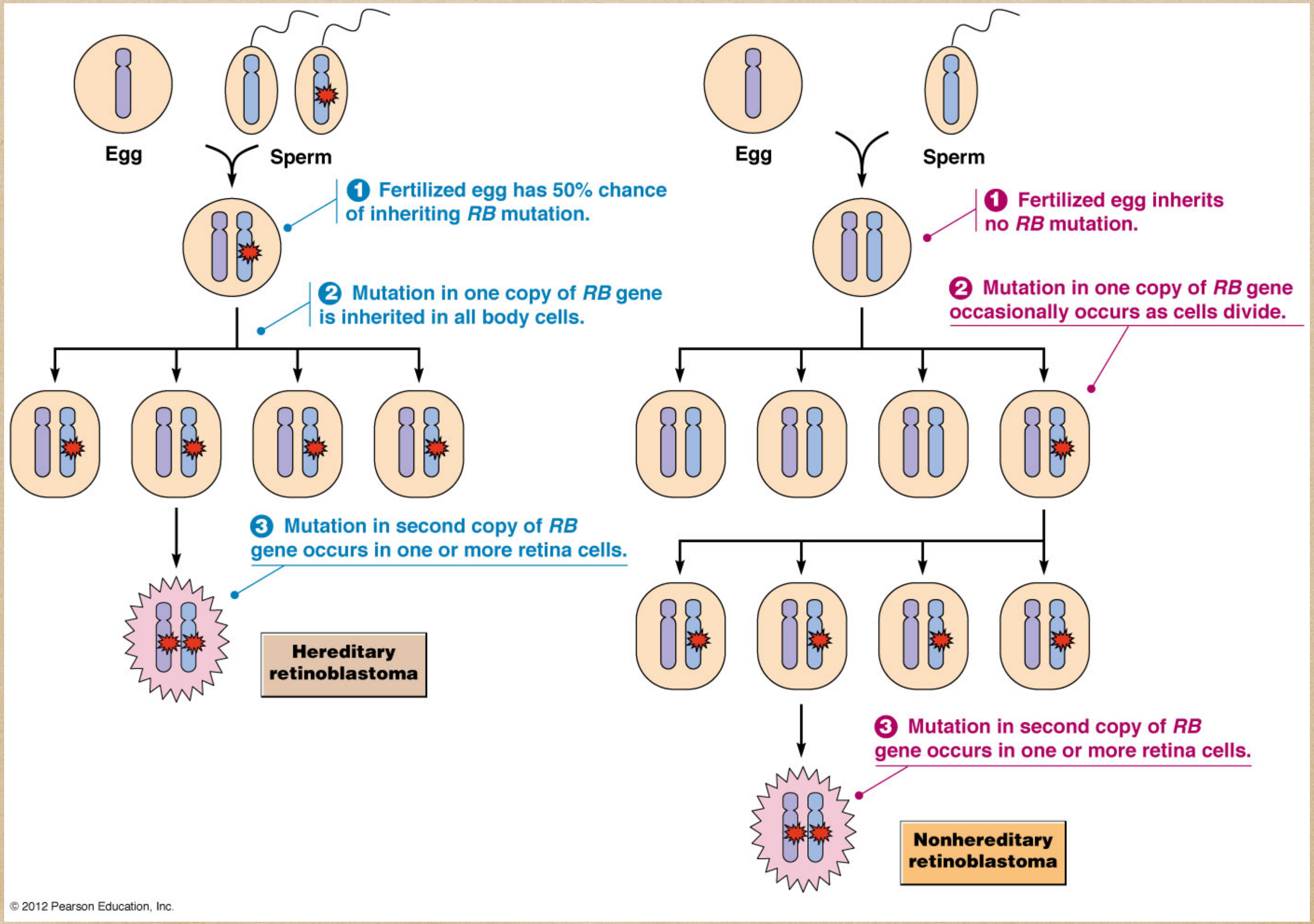
- We need single-cell resolution to:
  - Discover more complicated mechanisms in cellular development
  - Confirm the distinct gene expression signatures across different cell types
  - Identify functional differences among the same cell cell type

Single-cell sequencing

applications

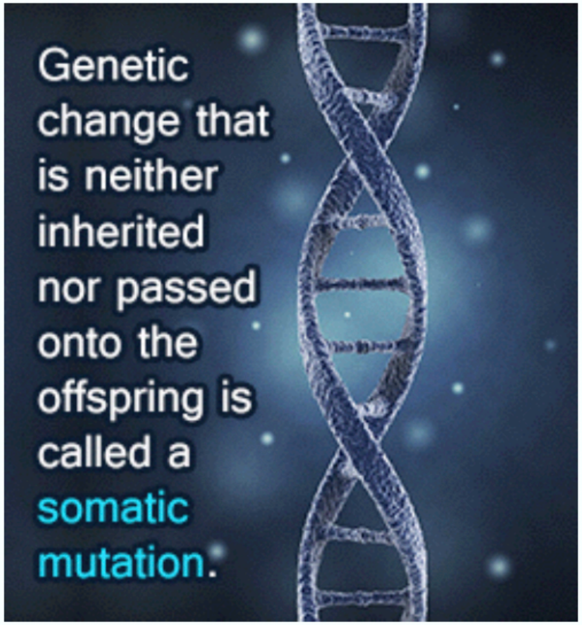
- Developmental Biology
- Cancer Biology
- Microbiology
- Neurology

Cancer Biology

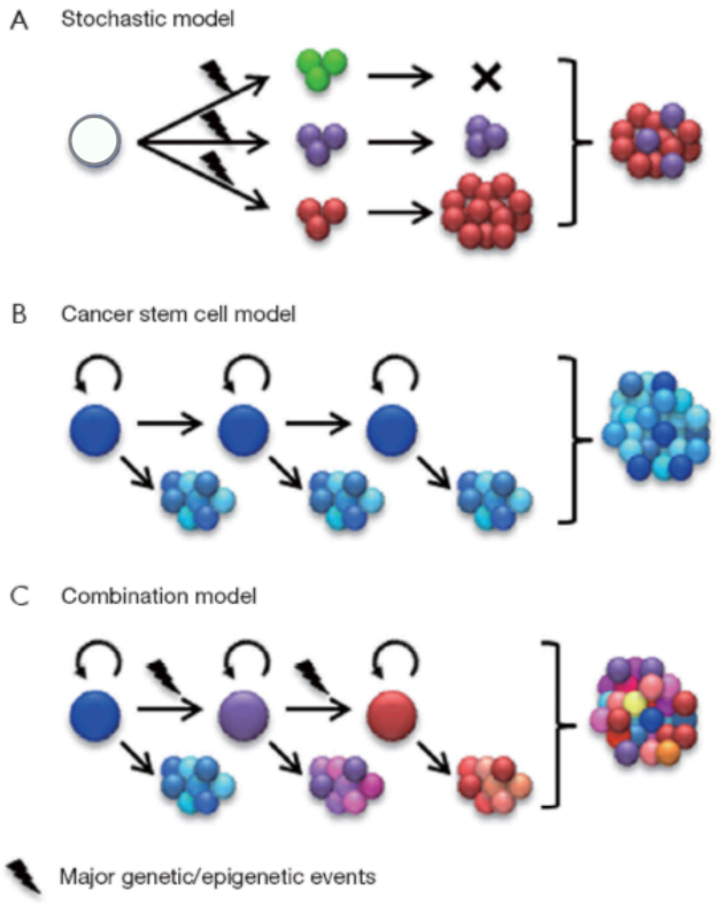

Egg
Sperm
❶ Fertilized egg has 50% chance of inheriting *RB* mutation.
❷ Mutation in one copy of *RB* gene is inherited in all body cells.
❸ Mutation in second copy of *RB* gene occurs in one or more retina cells.
Hereditary retinoblastoma
Egg
Sperm
❶ Fertilized egg inherits no *RB* mutation.
❷ Mutation in one copy of *RB* gene occasionally occurs as cells divide.
❸ Mutation in second copy of *RB* gene occurs in one or more retina cells.
Nonhereditary retinoblastoma

Cancer Biology

Tumors are composed of genetically and phenotypically **heterogeneous** clones

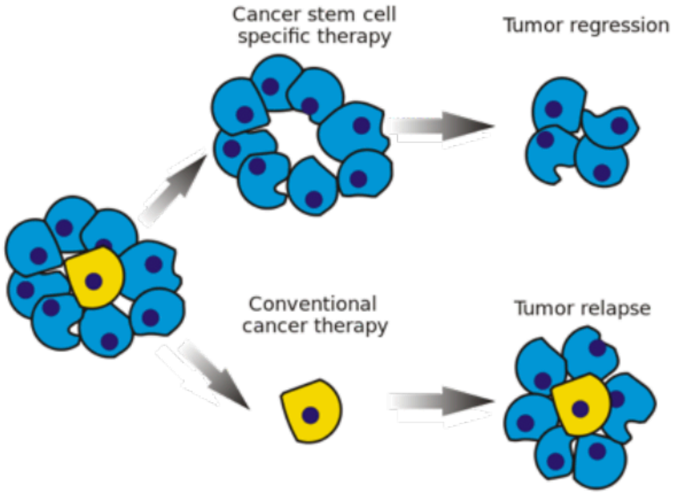


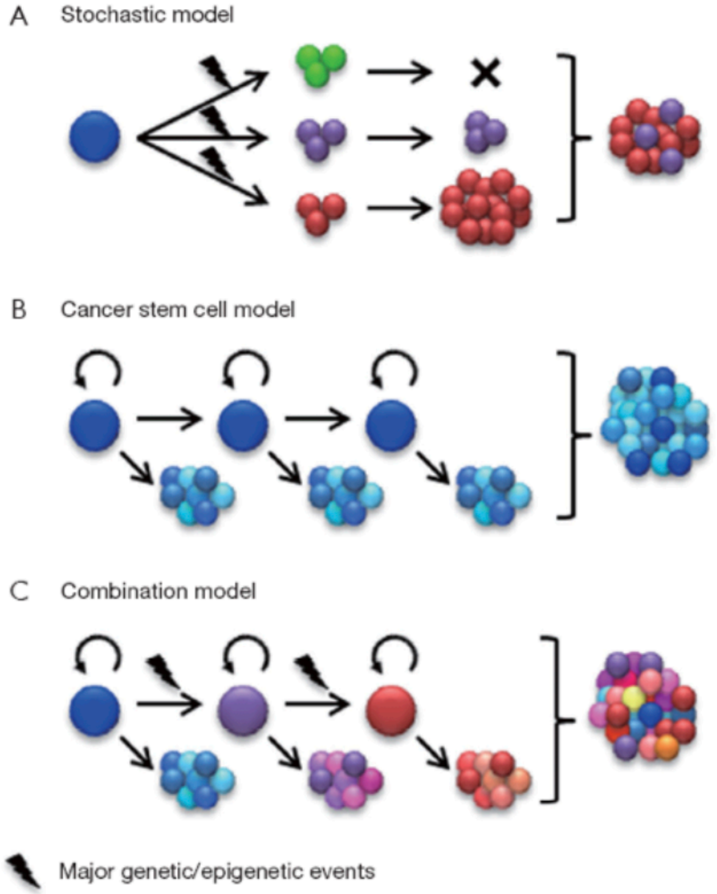
Genetic change that is neither inherited nor passed onto the offspring is called a somatic mutation.



http://www.thetcr.org/article/viewFile/1415/html/10439

Cancer Biology


Cancer stem cell specific therapy
Tumor regression
Conventional cancer therapy
Tumor relapse



Deep (bulk) sequencing can only capture 1% of the cell population (excluding some types such as circulating tumor cells).

http://www.thetcr.org/article/viewFile/1415/html/10439

Caner Biology

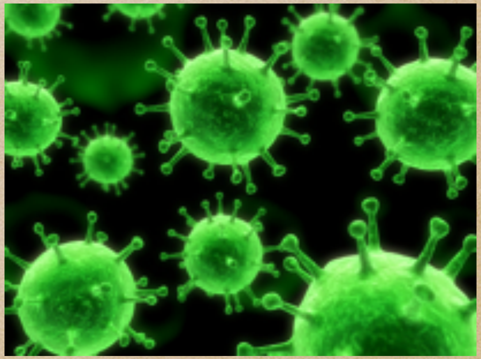
- We need single-cell resolution to:
    - Find evidence for models of cancer
    - Infer timing of mutations and the drivers
    - Evaluate effectiveness of targeted therapy

Single-cell sequencing

applications

- Developmental Biology
- Cancer Biology
- <u>Microbiology</u>
- Neurology

Microbiology



Microbiology

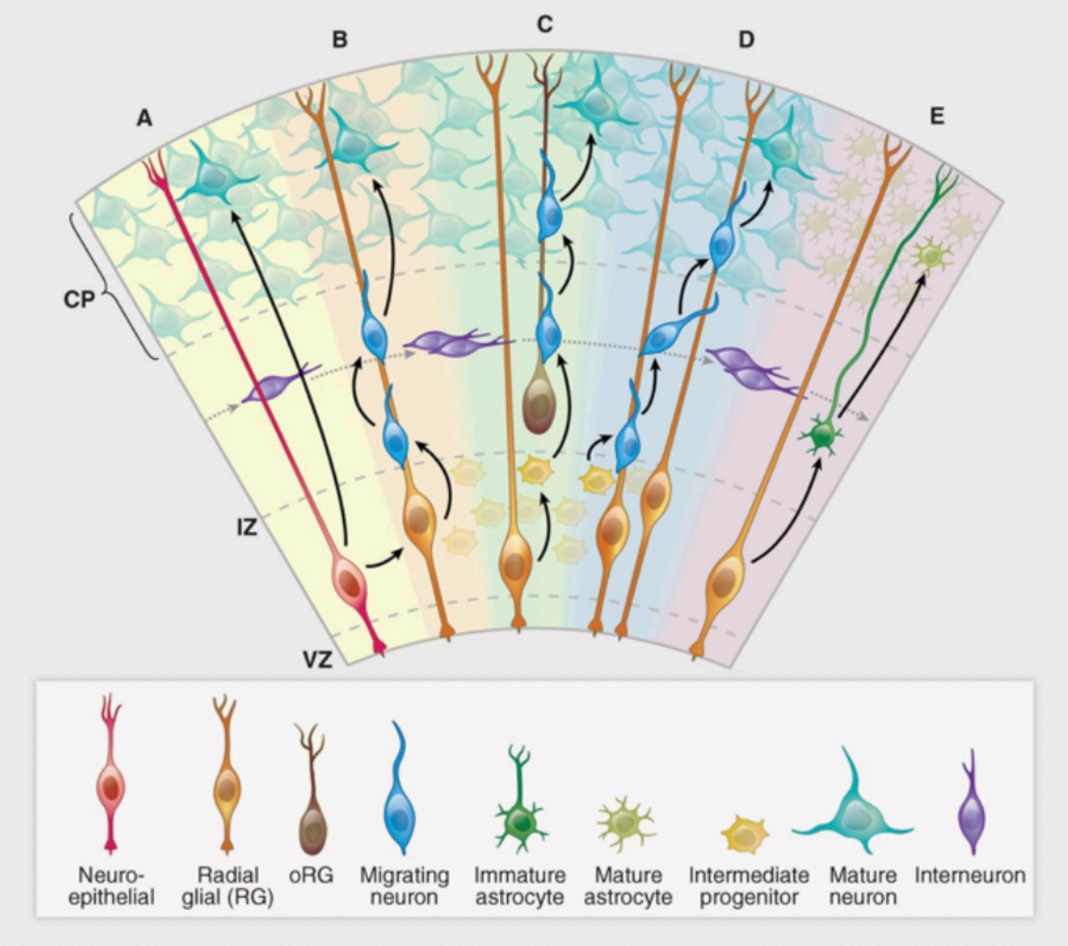
- We need single-cell resolution to:
    - Discover low-abundance species that are difficult to culture in vitro
    - Monitor transcriptional gene activation mechanisms for functional annotation

Single-cell sequencing

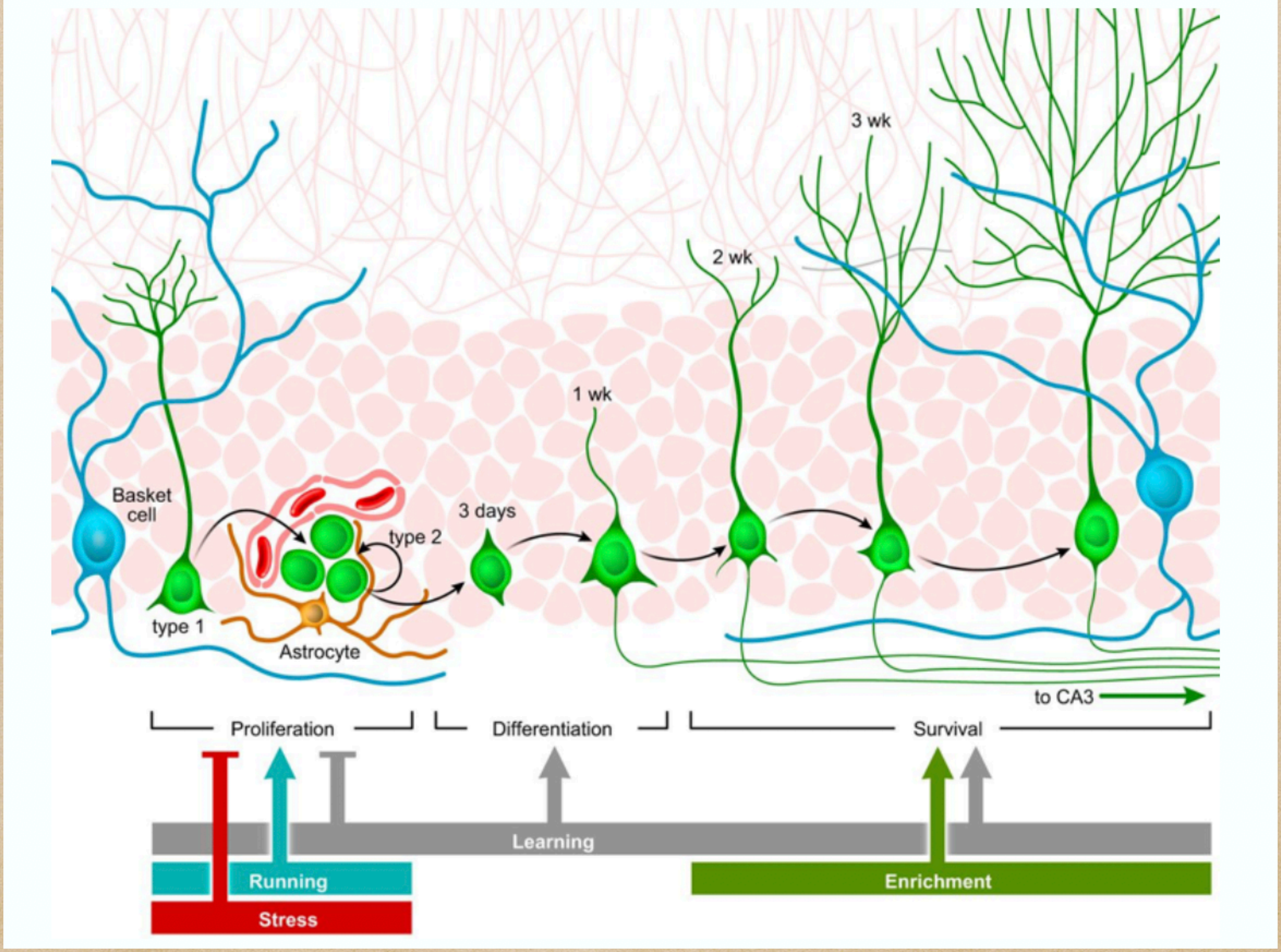
applications

- Developmental Biology
- Cancer Biology
- Microbiology
- Neurology

Microbiology



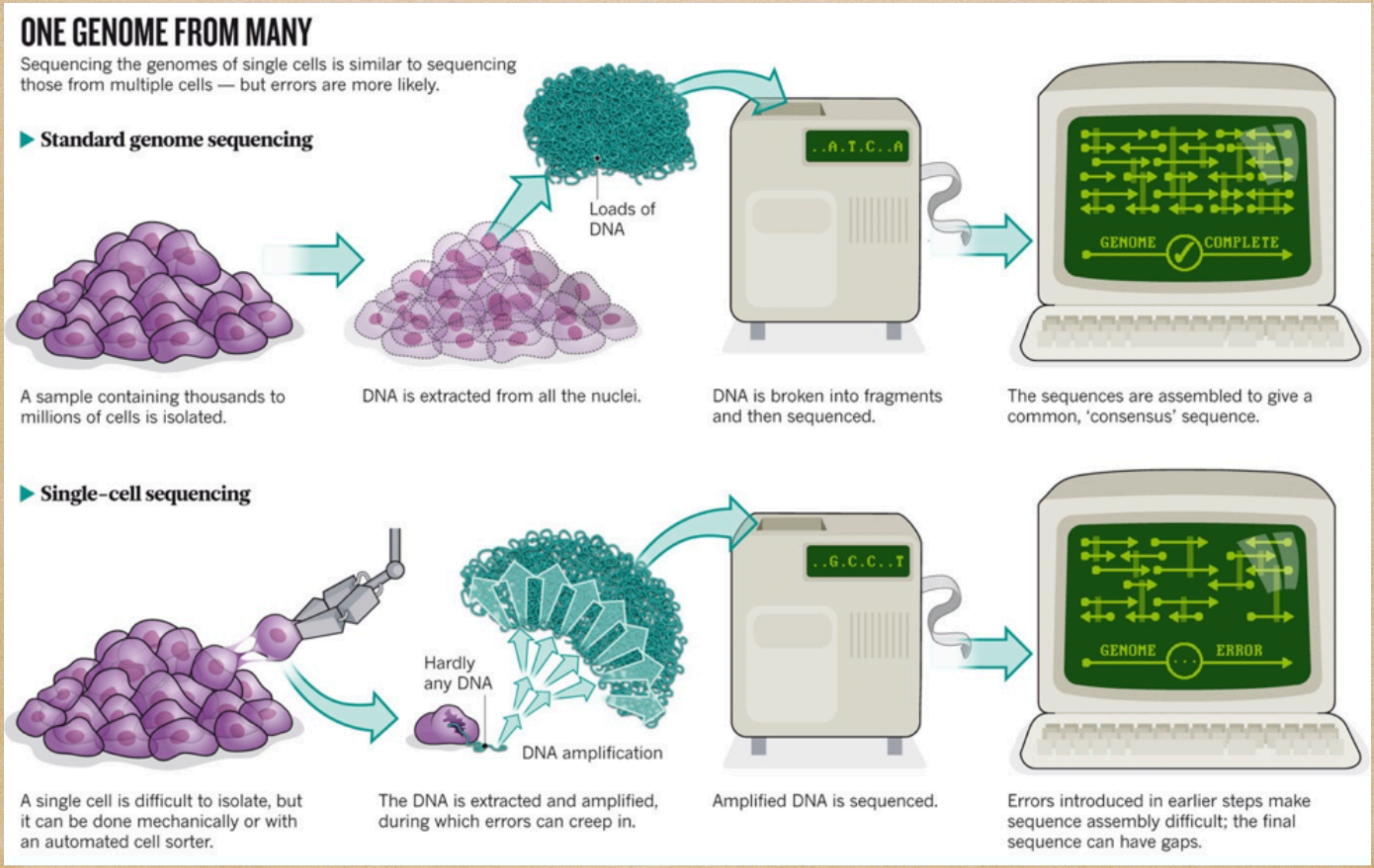
Microbiology



Neurology

- We need single-cell resolution to:
  - Study the mosaic genomes of individual neurons and compositions in the brain
  - Follow genetic variations during fetal development
  - Develop targeted therapy for neurological diseases for specific cell types

Traditional vs. Single-cell sequencing



ONE GENOME FROM MANY

Sequencing the genomes of single cells is similar to sequencing those from multiple cells — but errors are more likely.

Standard genome sequencing

A sample containing thousands to millions of cells is isolated.

DNA is extracted from all the nuclei.

DNA is broken into fragments and then sequenced.

The sequences are assembled to give a common, 'consensus' sequence.

Single-cell sequencing

A single cell is difficult to isolate, but it can be done mechanically or with an automated cell sorter.

The DNA is extracted and amplified, during which errors can creep in.

Amplified DNA is sequenced.

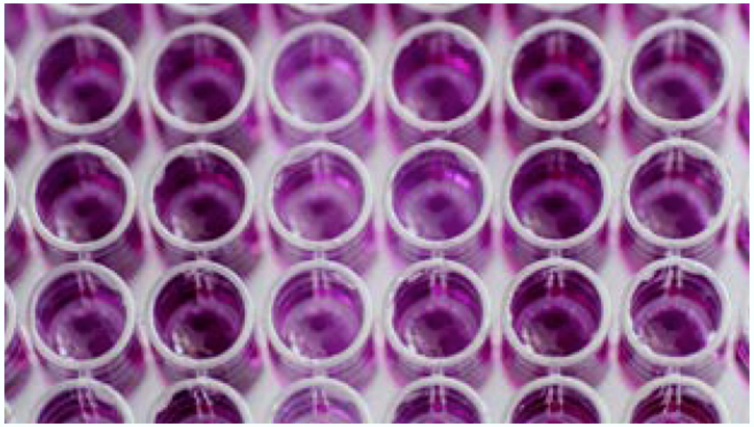
Errors introduced in earlier steps make sequence assembly difficult; the final sequence can have gaps.

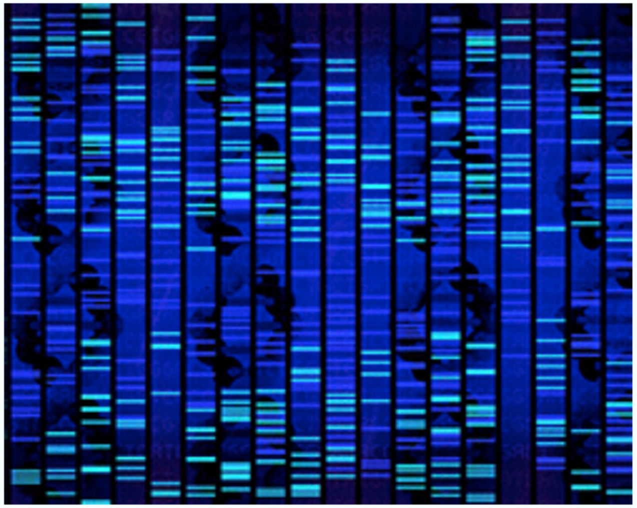
Single-Cell Technologies

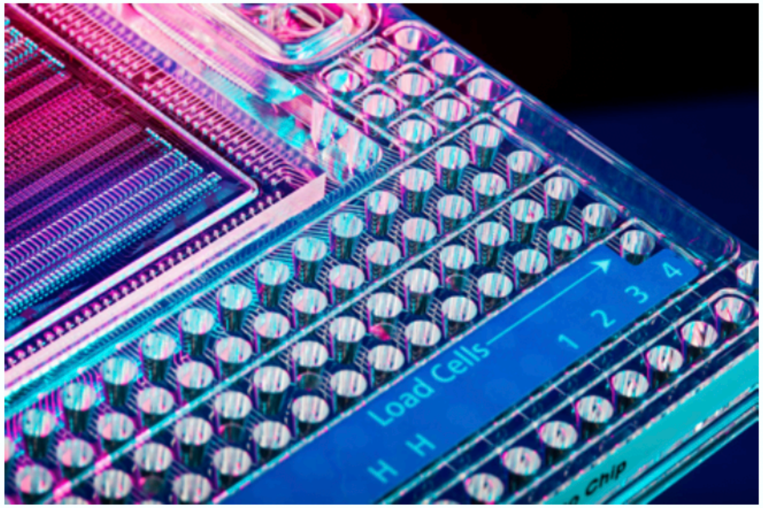
(i) isolate single cells

(ii) amplify genome efficiently

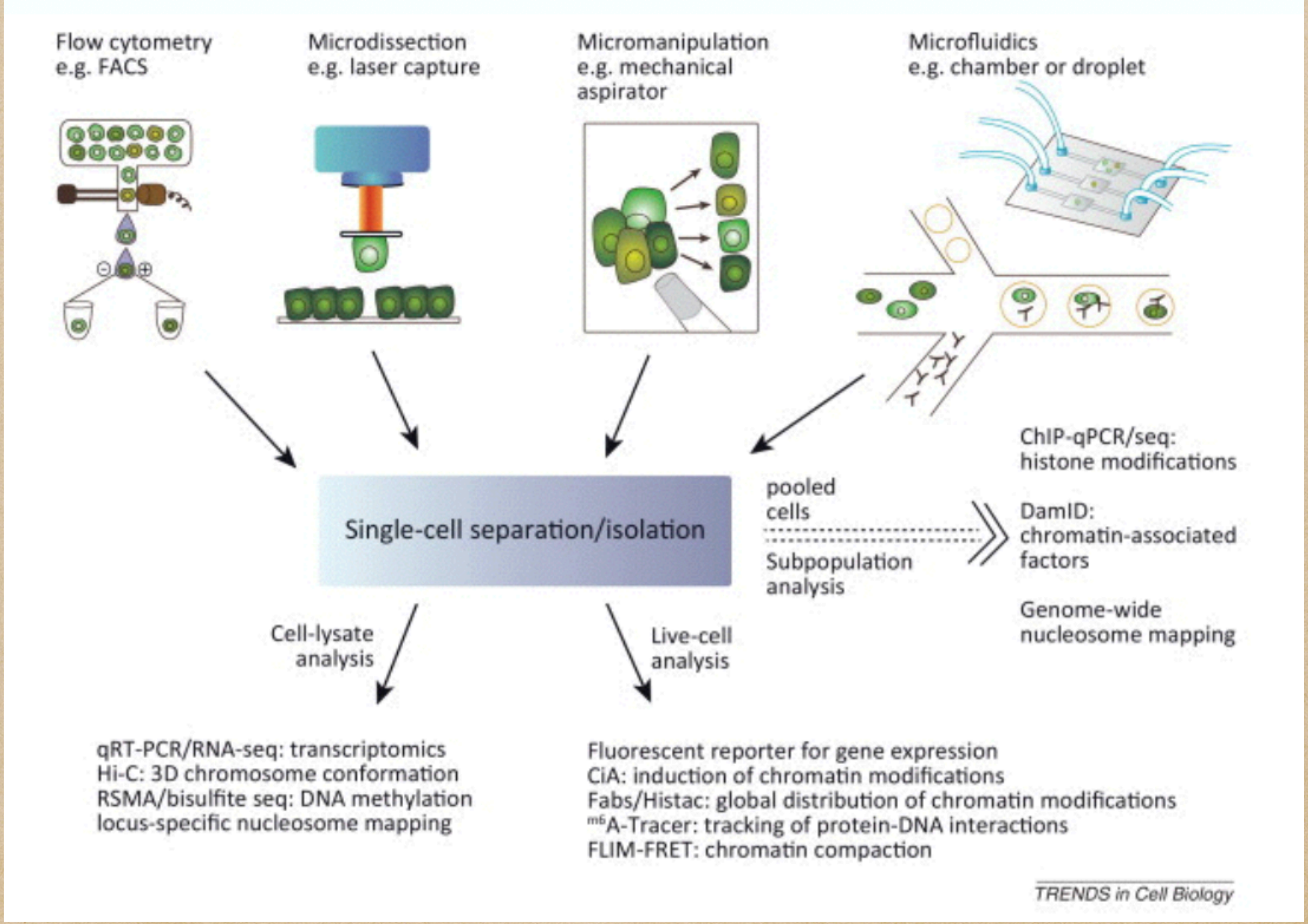
(iii) sequence DNA







Single-Cell Technologies



ChIP-qPCR/seq: histone modifications

DamID: chromatin-associated factors

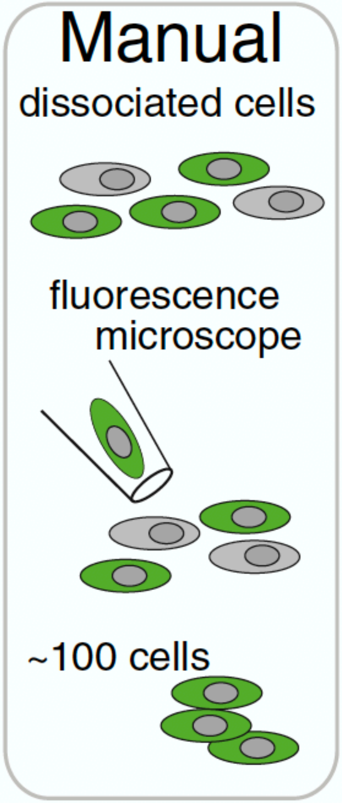
Genome-wide nucleosome mapping

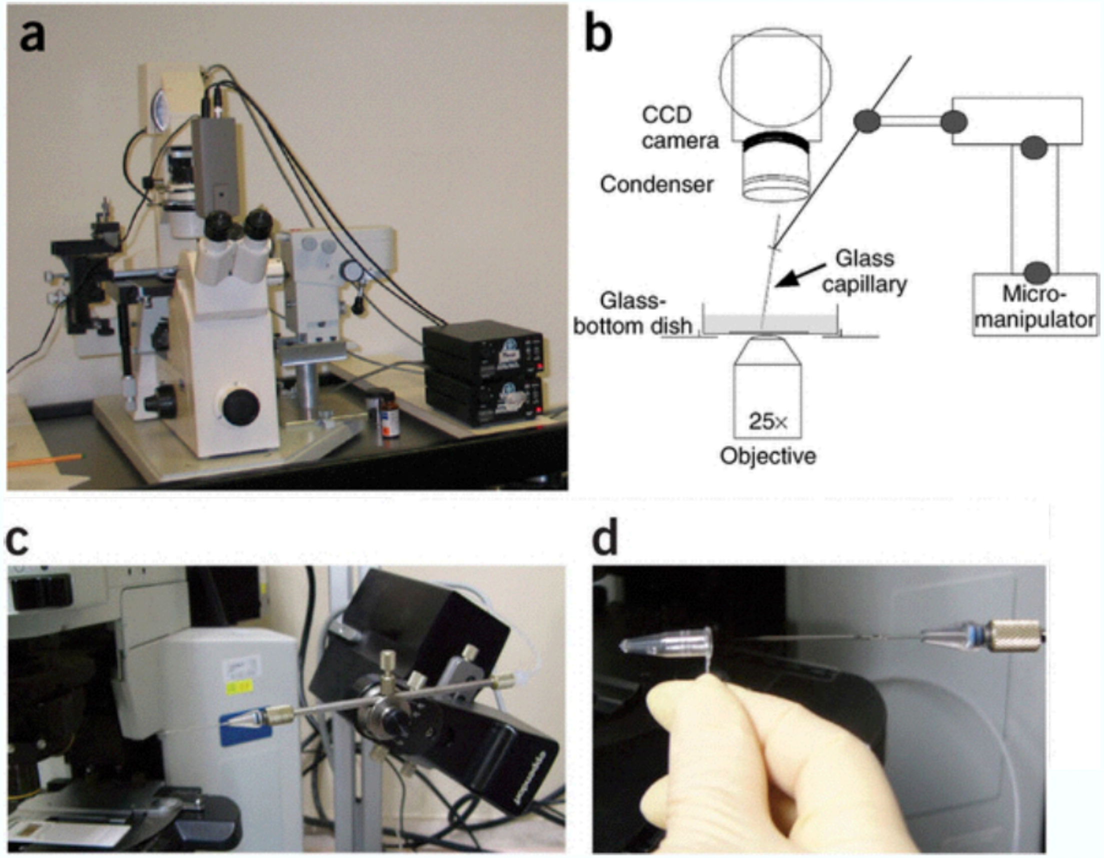
qRT-PCR/RNA-seq: transcriptomics
Hi-C: 3D chromosome conformation
RSMA/bisulfite seq: DNA methylation
locus-specific nucleosome mapping

Fluorescent reporter for gene expression
CiA: induction of chromatin modifications
Fabs/Histac: global distribution of chromatin modifications
^{m6}A-Tracer: tracking of protein-DNA interactions
FLIM-FRET: chromatin compaction

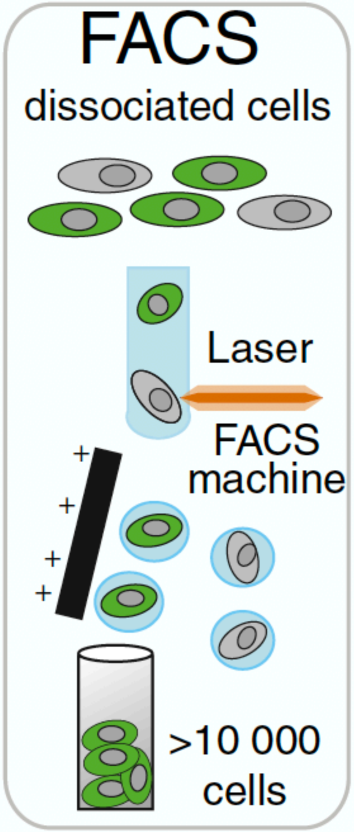
TRENDS in Cell Biology

Cell Sorting

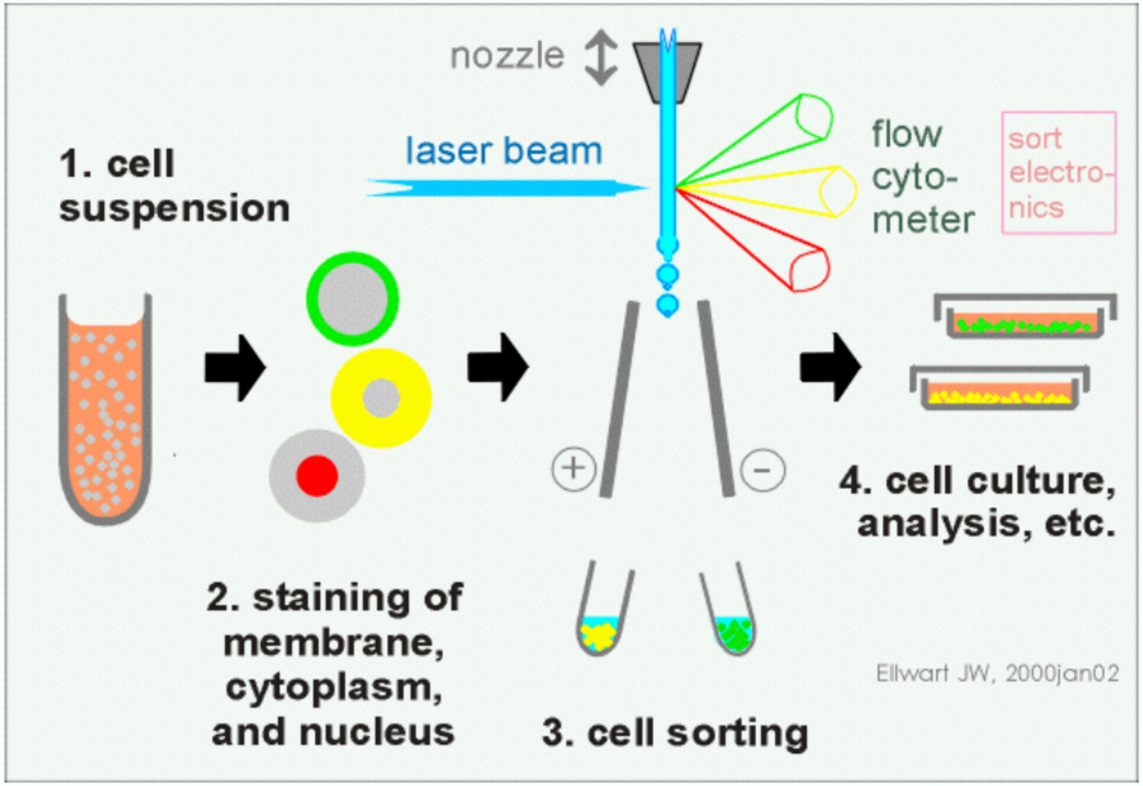




Cell Sorting

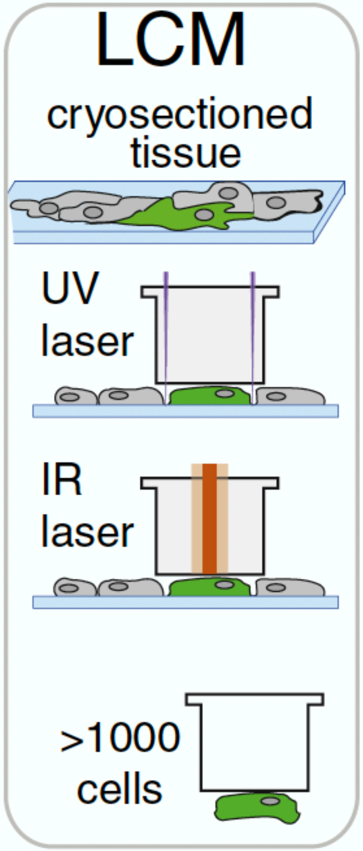


FACS: fluorescence activated cell sorting



Ellwart JW, 2000jan02

Cell Sorting



LCM

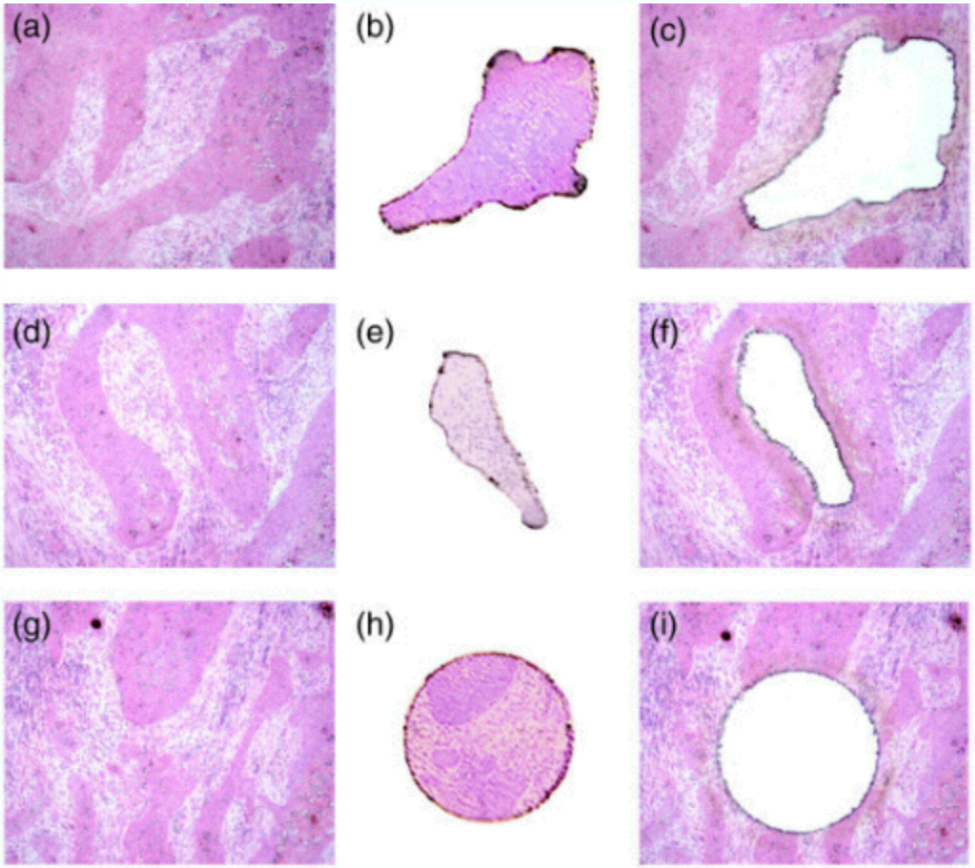
cryosectioned tissue

UV laser

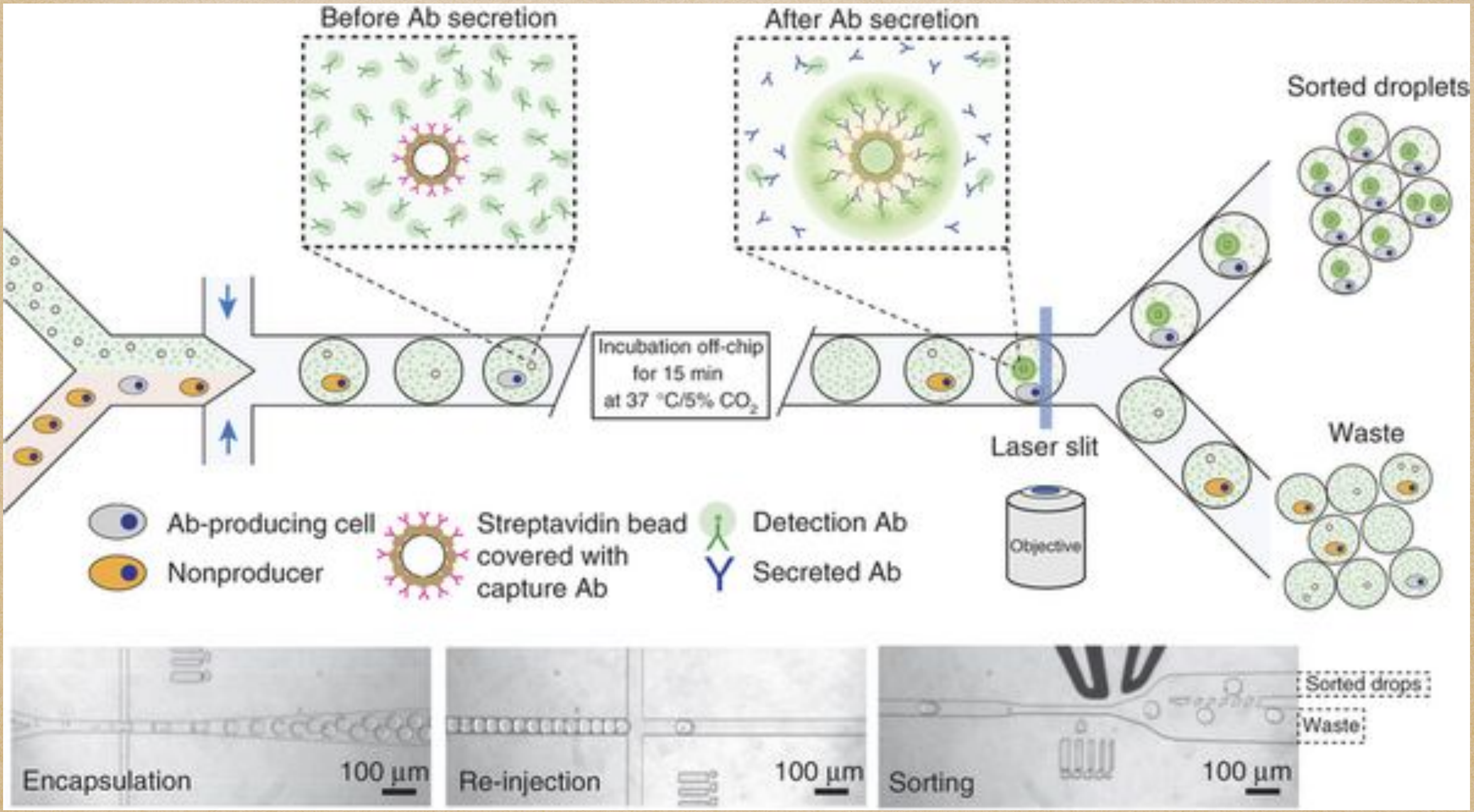
IR laser

>1000 cells

LCM: laser capture microdissection



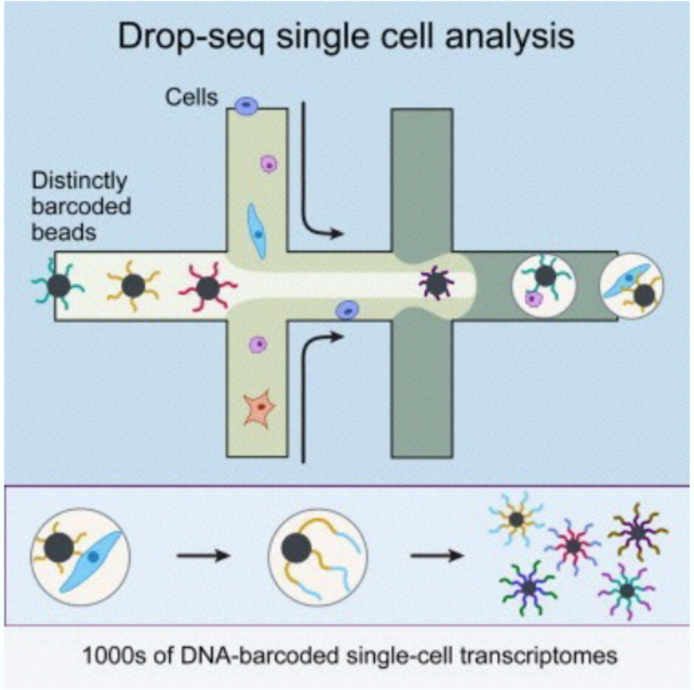
Cell Sorting



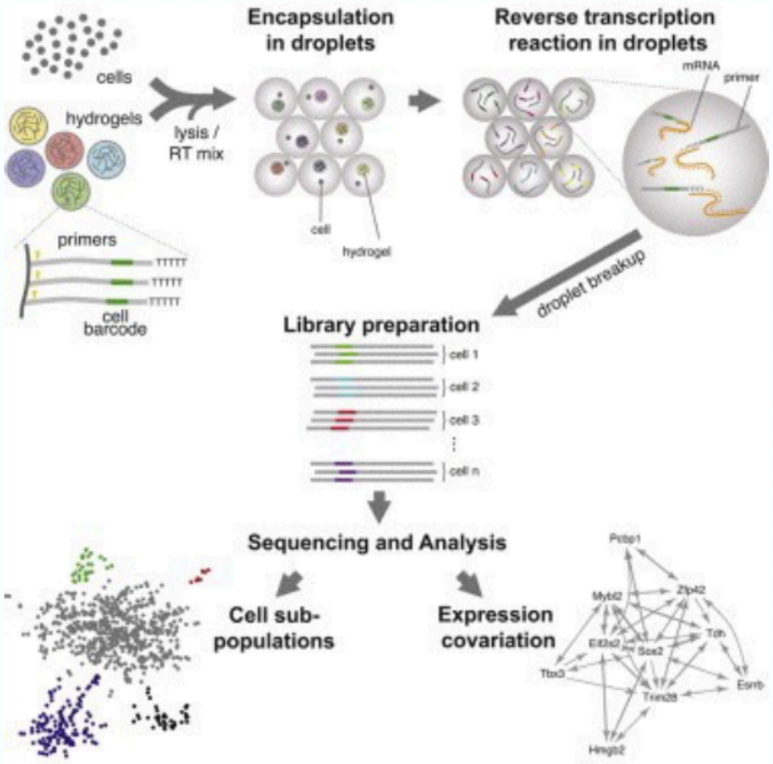
https://media.nature.com/full/nature-assets/nprot/journal/v8/n5/images_article/nprot.2013.046-F4.jpg

Cell Sorting

High-throughput (~100,000 cells)


Drop-seq single cell analysis
Cells
Distinctly barcoded beads
1000s of DNA-barcoded single-cell transcriptomes

Drop-seq



inDrop

http://www.cell.com/abstract/S0092-8674(15)00549-8

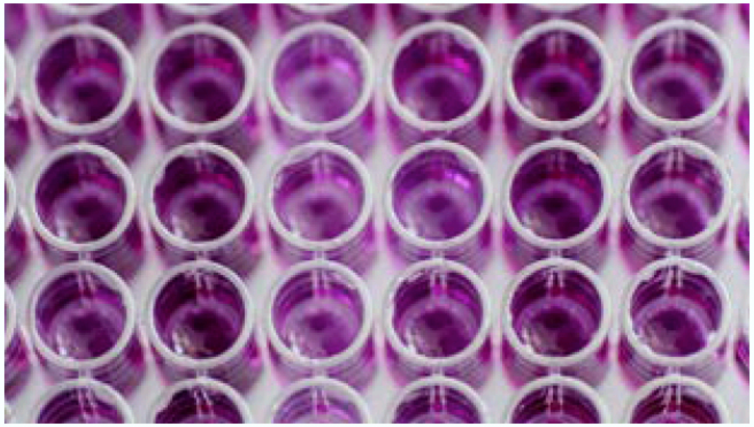
http://www.cell.com/cell/abstract/S0092-8674(15)00500-0

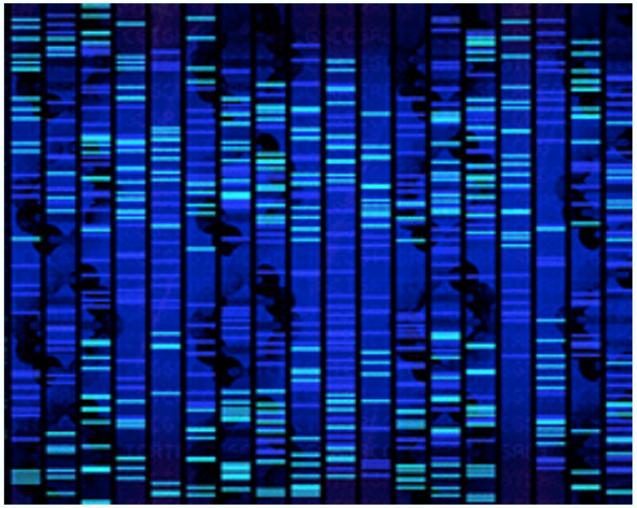
Single-Cell Technologies

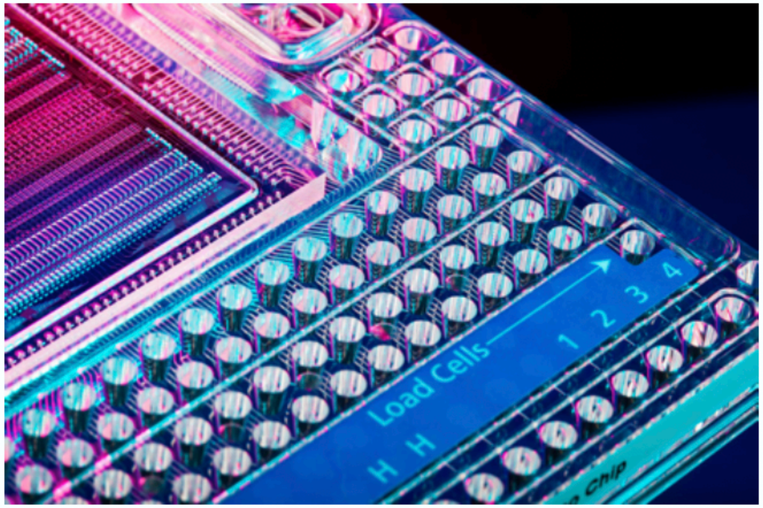
(i) isolate single cells

(ii) amplify genome efficiently

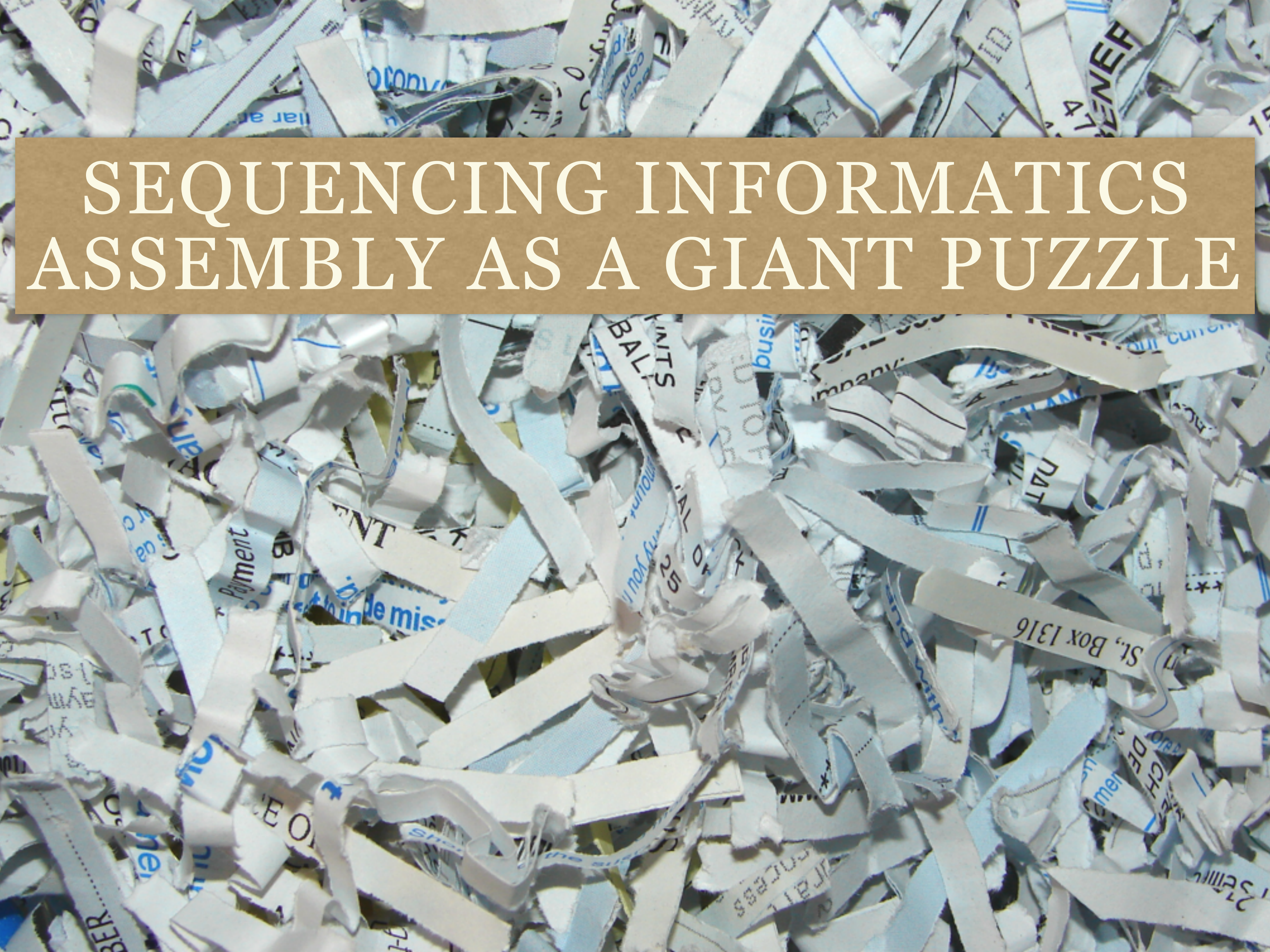
(iii) sequence DNA







SEQUENCING INFORMATICS
ASSEMBLY AS A GIANT PUZZLE



Sequencing informatics

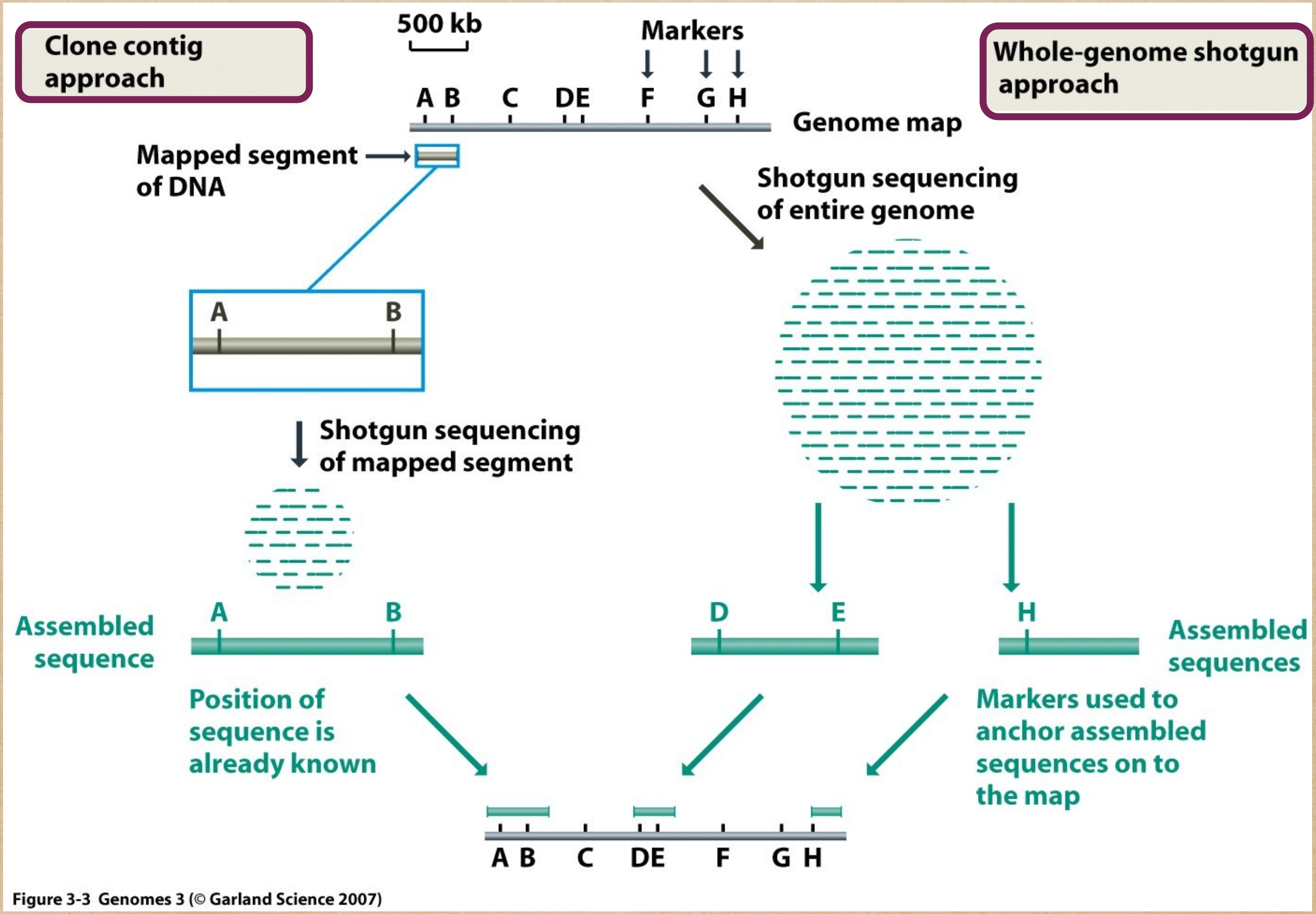

Clone contig approach
500 kb
Markers
Whole-genome shotgun approach
A B C DE F G H
Genome map
Mapped segment of DNA
Shotgun sequencing of entire genome
A
B
Shotgun sequencing of mapped segment
Assembled sequence
A
B
D
E
H
Assembled sequences
Position of sequence is already known
Markers used to anchor assembled sequences on to the map
A B C DE F G H

Figure 3-3 Genomes 3 (© Garland Science 2007)

Sequencing informatics

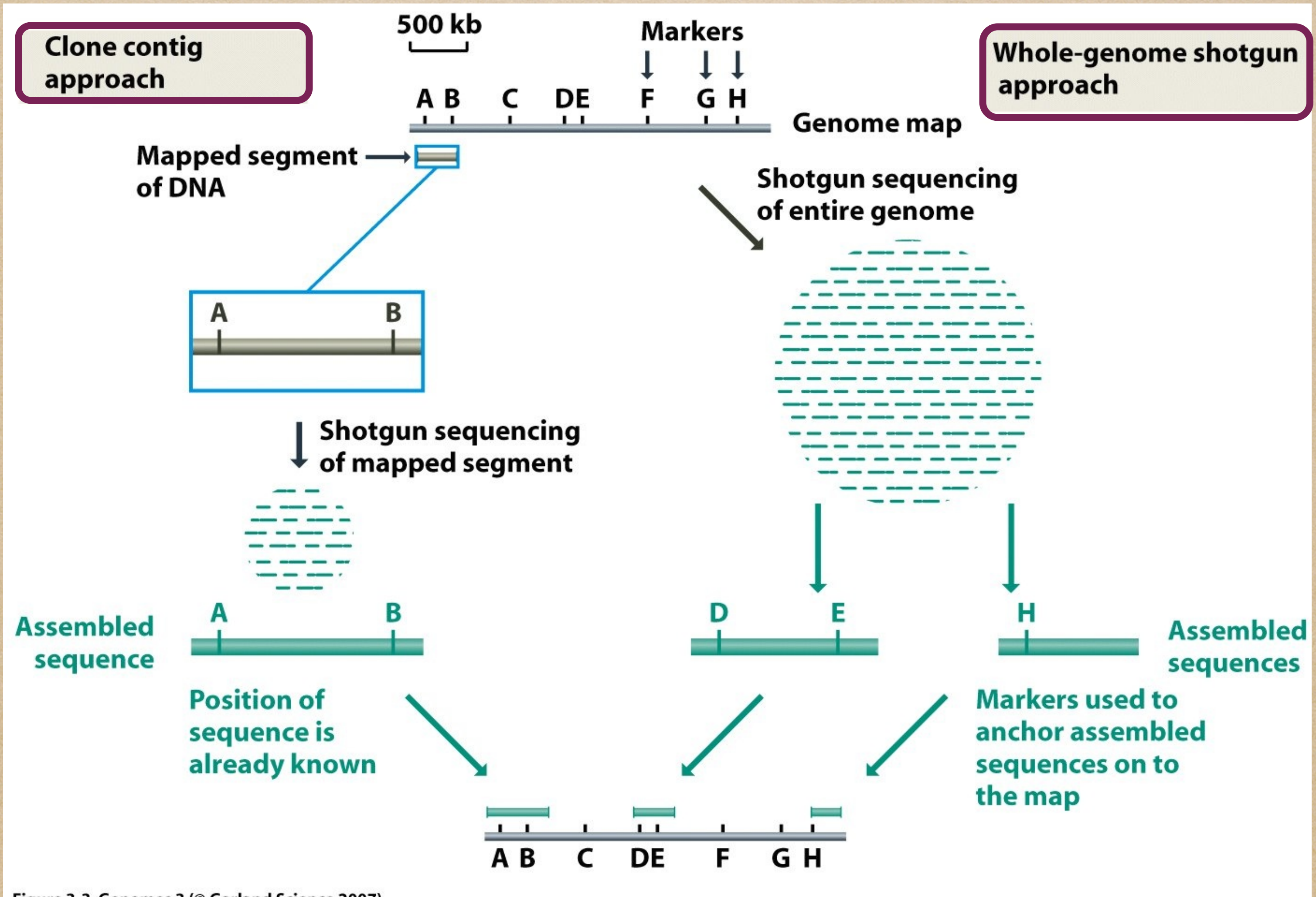
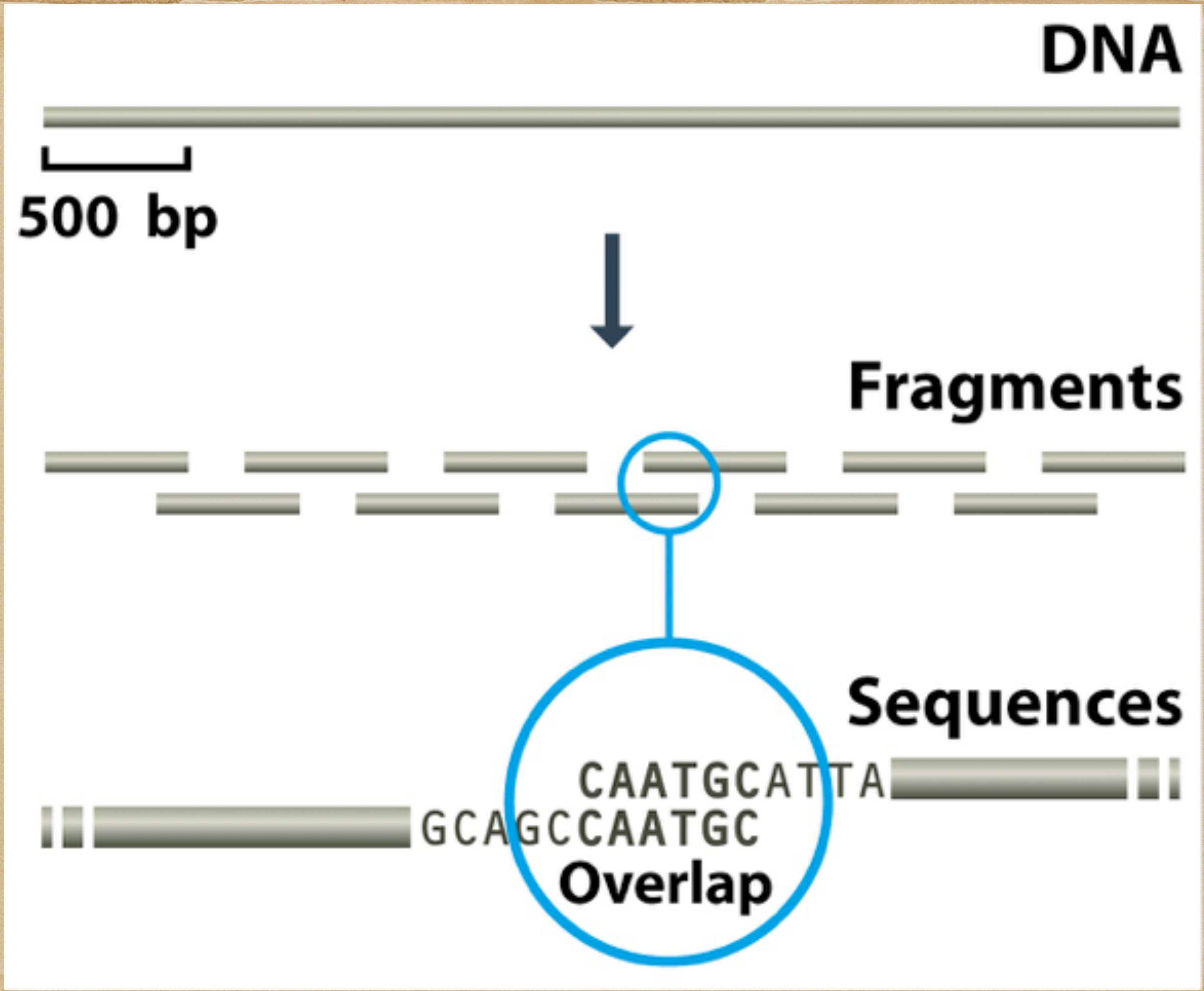

Clone contig approach
500 kb
Markers
A B C DE F G H
Genome map
Whole-genome shotgun approach
Mapped segment of DNA
Shotgun sequencing of entire genome
A
B
Shotgun sequencing of mapped segment
Assembled sequence
A
B
D
E
H
Assembled sequences
Position of sequence is already known
Markers used to anchor assembled sequences on to the map
A B C DE F G H

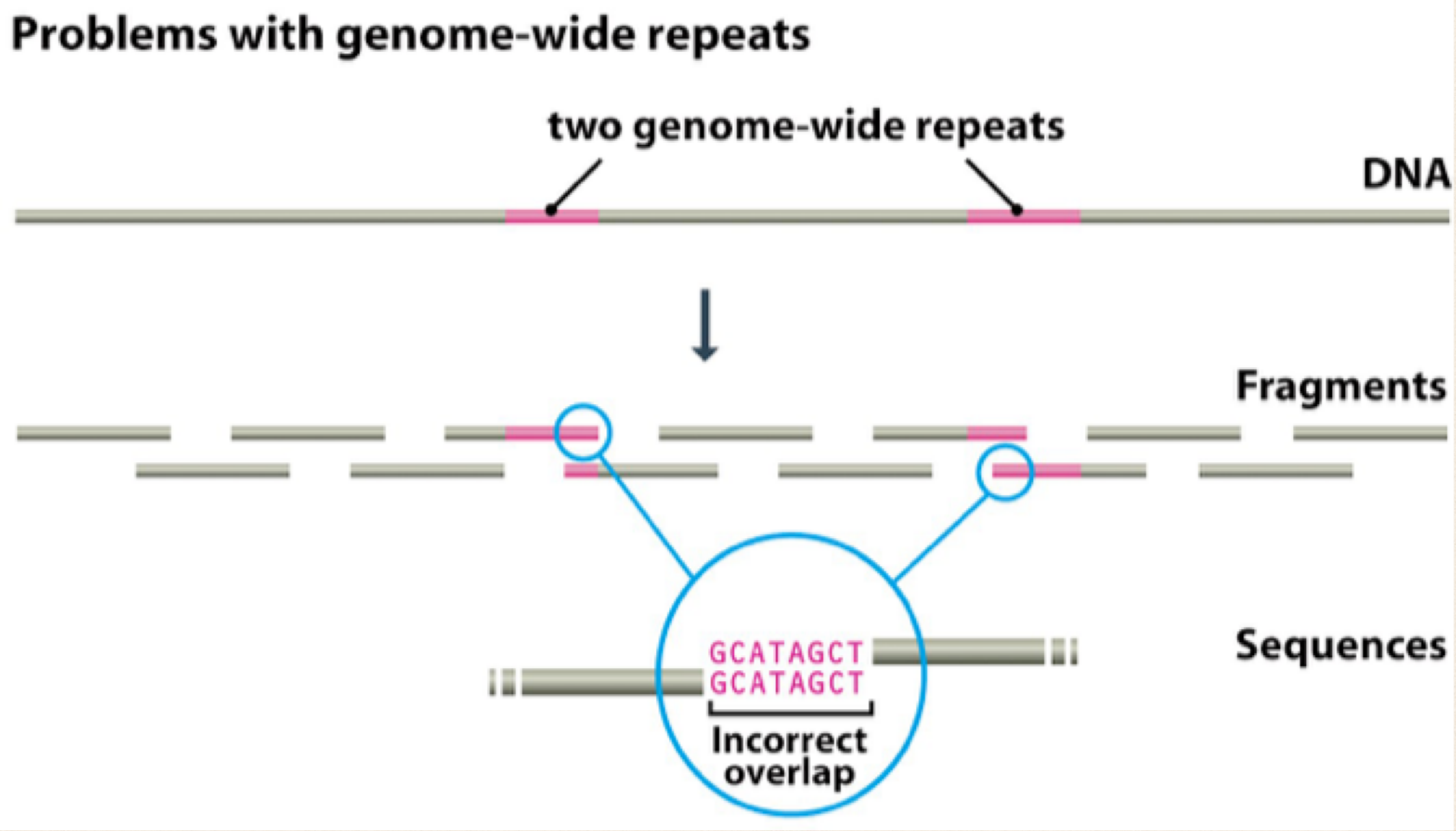
Figure 3-3 Genomes 3 (© Garland Science 2007)

Sequence assembly

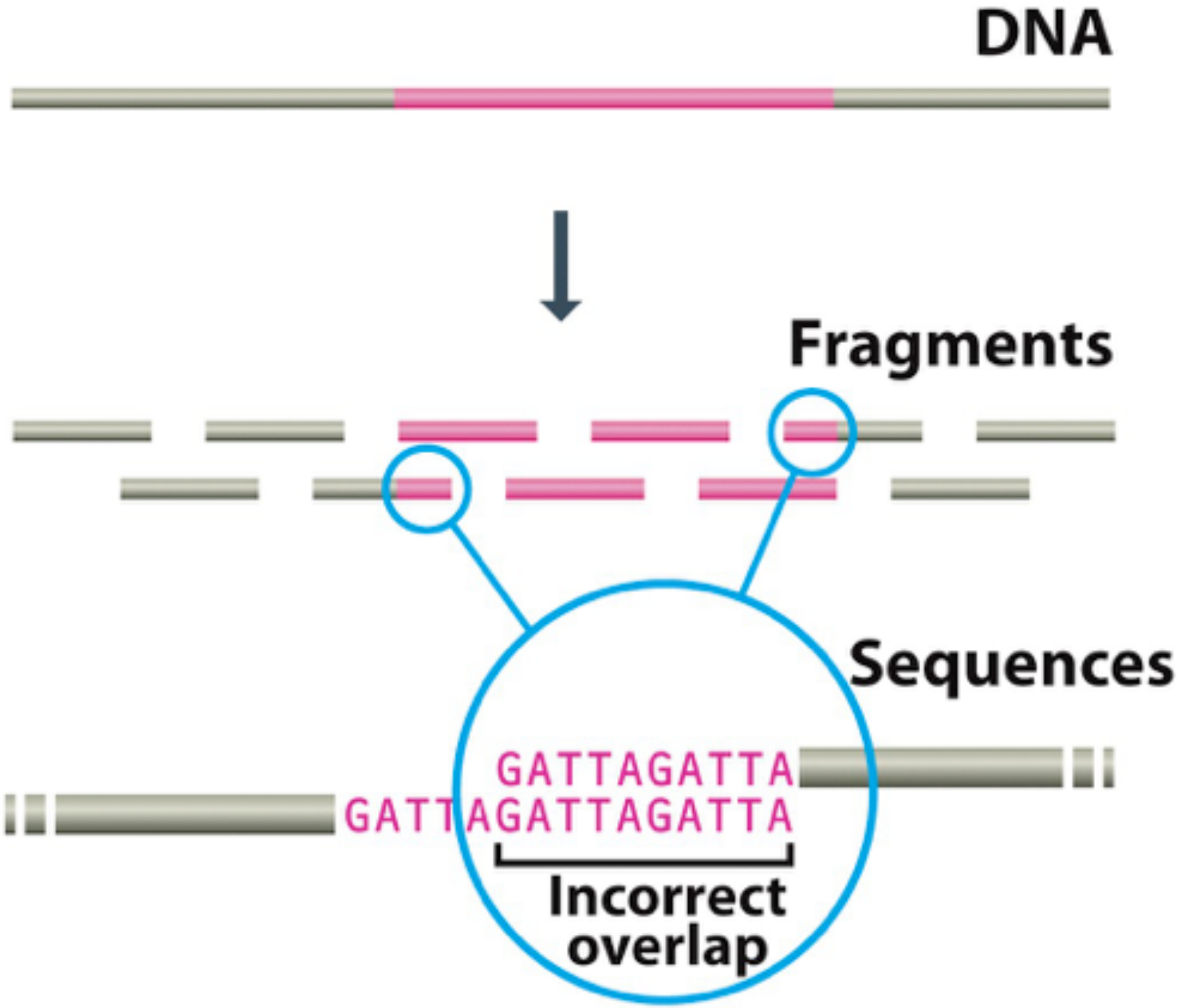
- A fundamental goal of DNA sequencing has been to generate large, continuous regions of DNA sequence – CONTIGS
- In principle, assembling a sequence is just a matter of finding overlaps and combining them.
- In practice:
    - most genomes contain multiple copies of many sequences,
    - there are random mutations (either naturally occurring cell-to-cell variation or generated by PCR or cloning),
    - there are sequencing errors


DNA
500 bp
Fragments
Sequences
CAATGCATTA
GCAGCCAATGC
Overlap

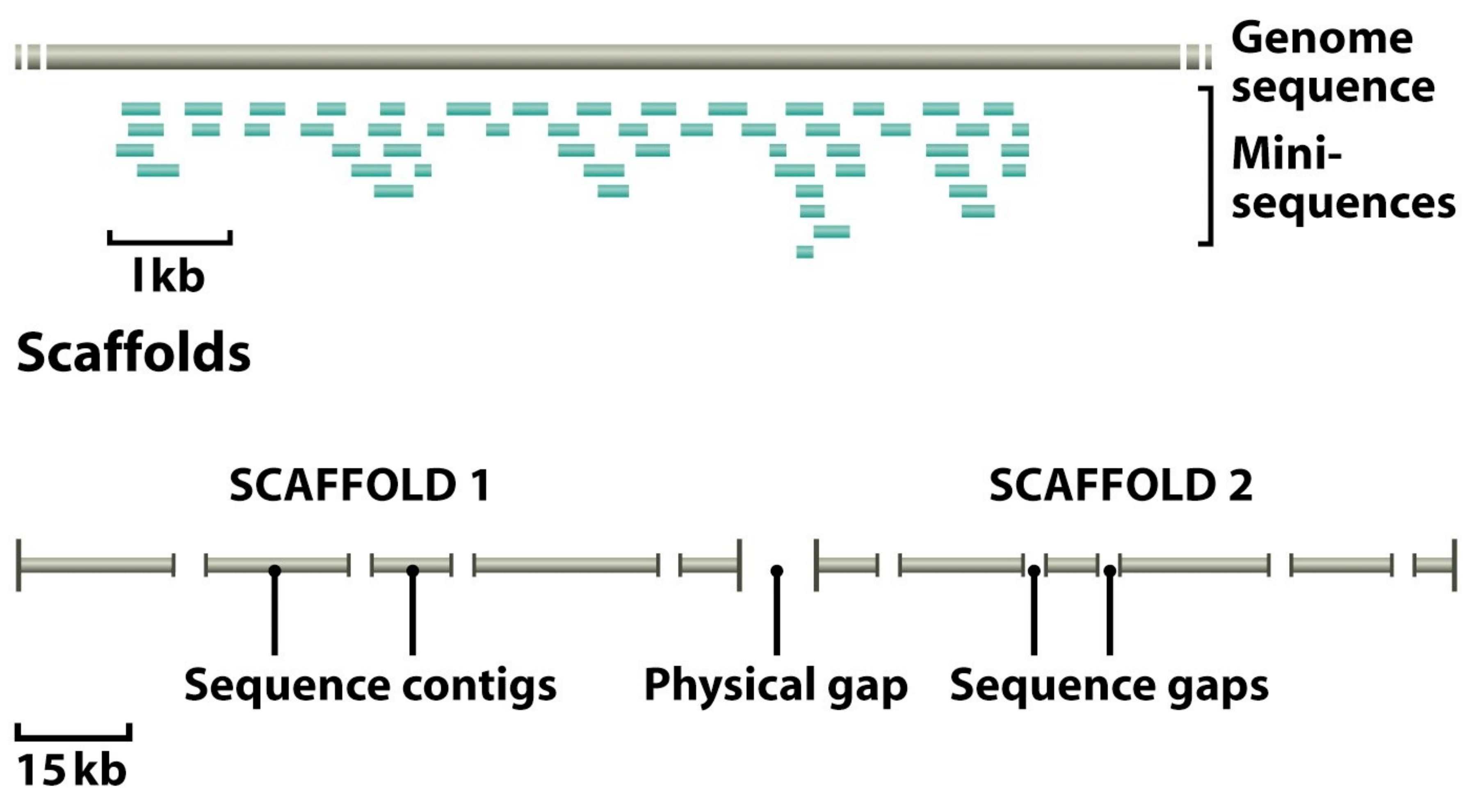
Assembly problems



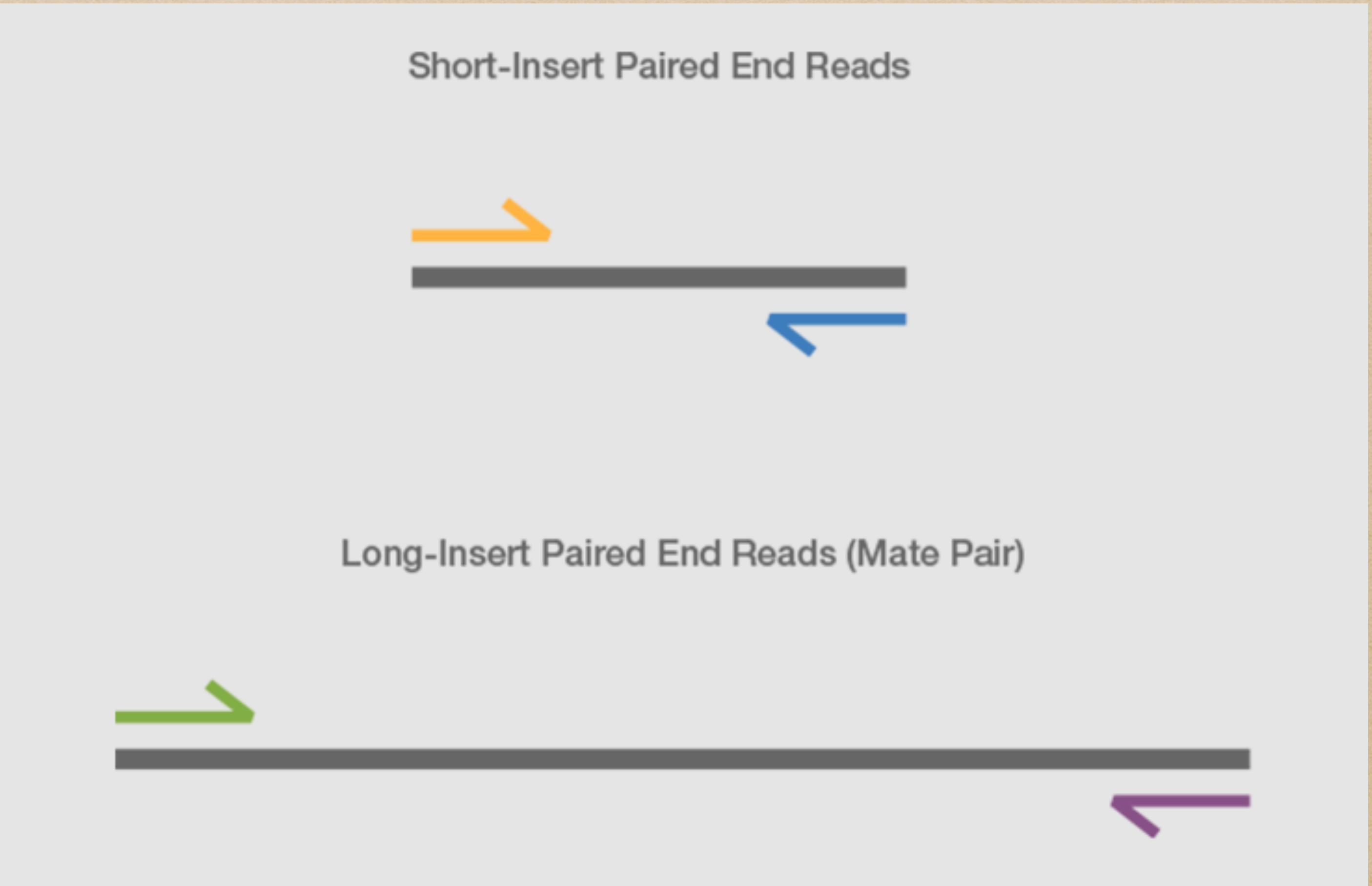
Problems with tandemly repeated DNA



Assembly problems: sequencing gaps



Sequencing gaps - pair end reads to the rescue



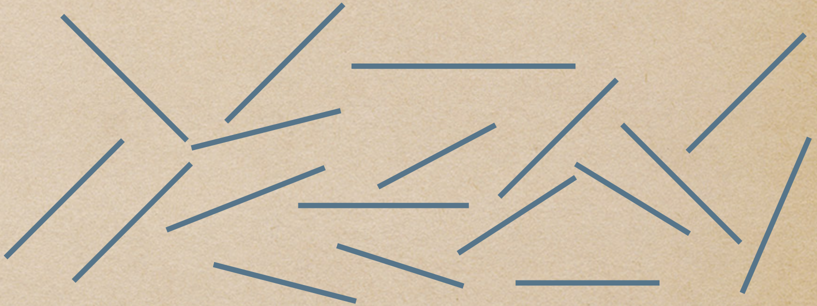
Short-Insert Paired End Reads

Long-Insert Paired End Reads (Mate Pair)

Overview of genome assembly (1)

Sample collection

DNA sequencing



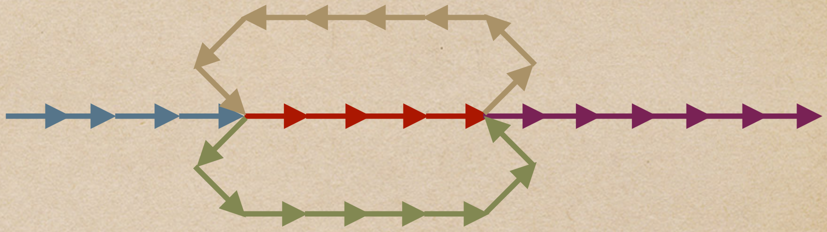
Pairwise read overlaps

```
…AGCTTTAGGCTAGCAATGC
            GCAATGCTATAGGCCT…
```

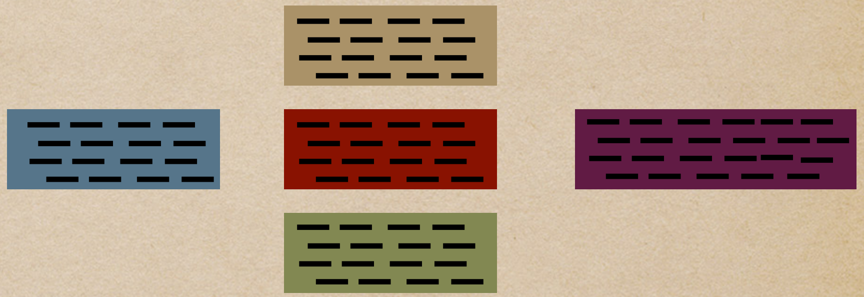
Based on Fig. 1 at Sedlazeck et al. 2018

Overview of genome assembly (2)

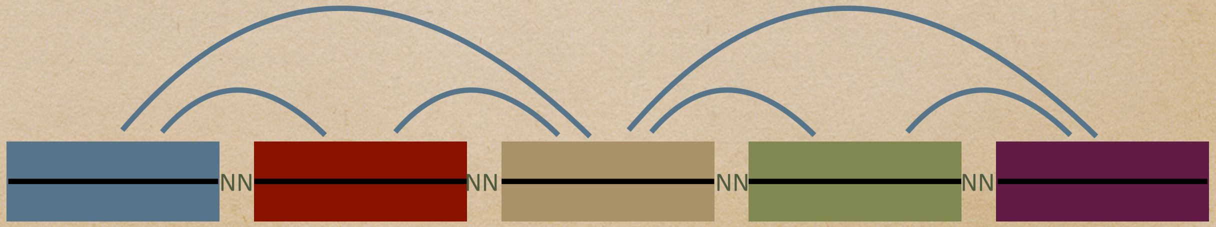
String graph construction



Contig construction

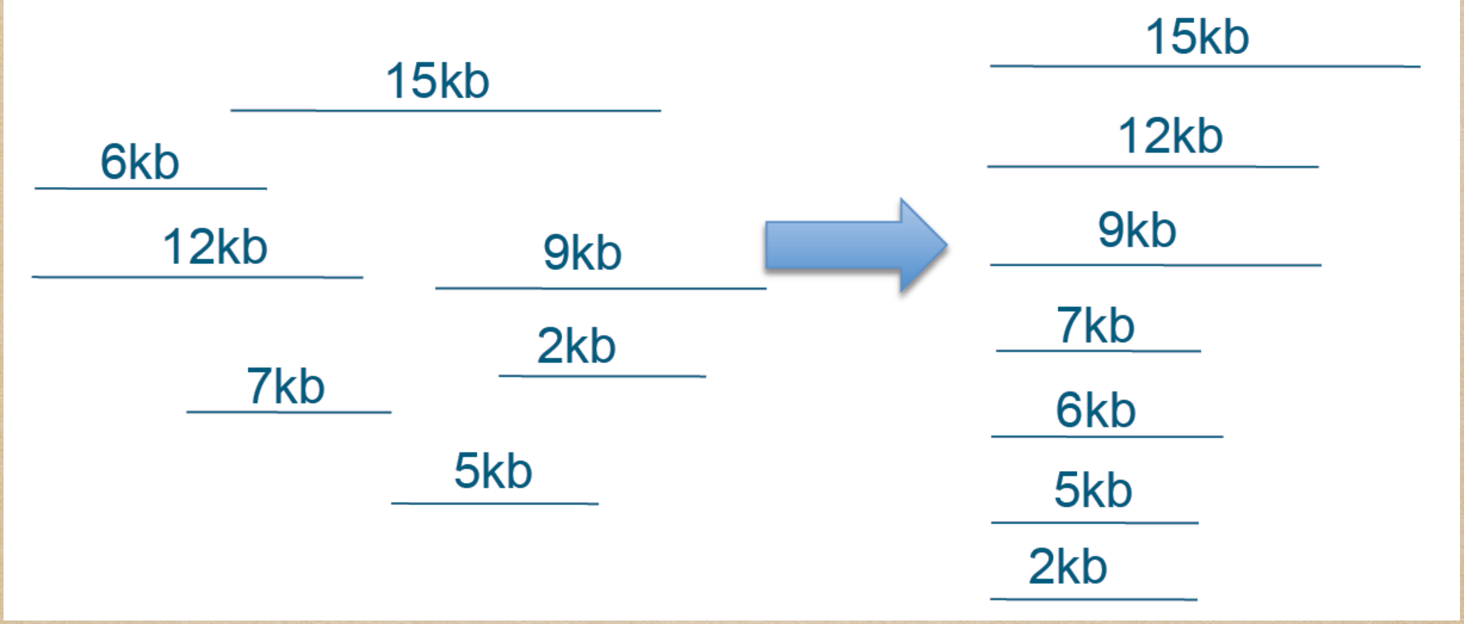


Scaffold construction



Based on Fig. 1 at Sedlazeck et al. 2018

Assembly evaluation - N50



If one orders the set of contigs produced by the assembler by size, then N50 is the size of the contig such that 50% of the total bases are in contigs of equal or greater size.

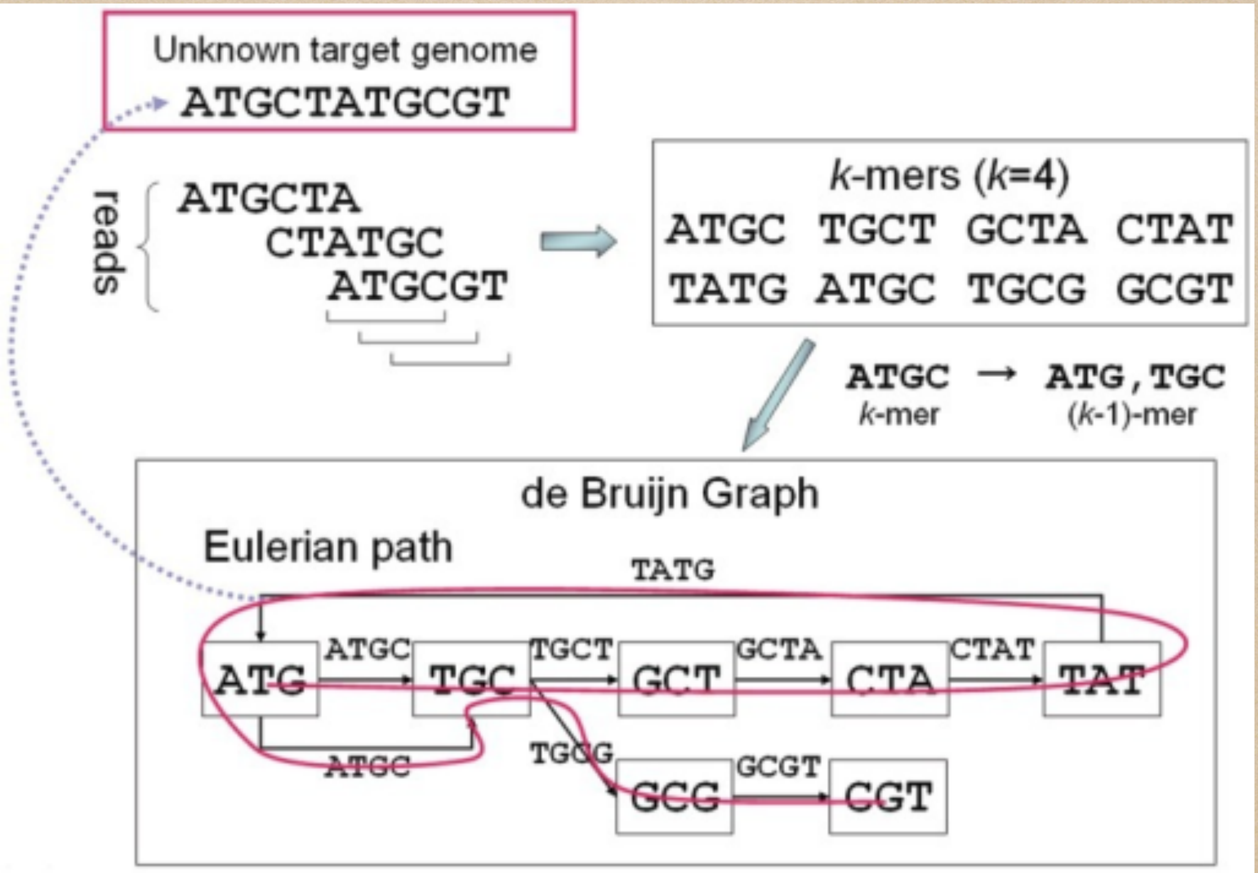
15+12+9+7+6+5+2 = 56.

56/2 = 28 -> N50 is 9kb (15+12 = 27 is less than 50%)

Sequence assembly
NGS case

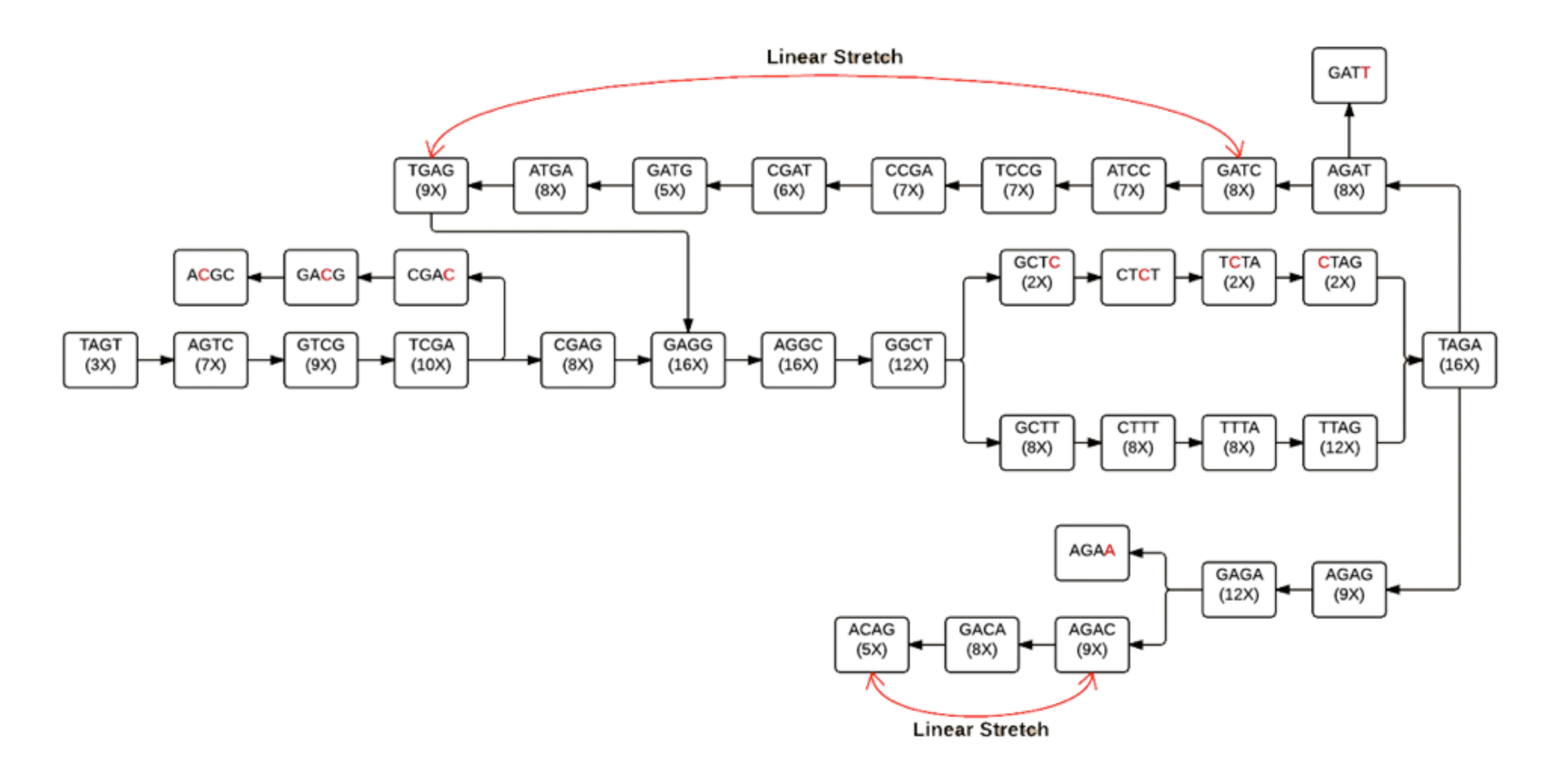
- Volume and read length of data from next-gen sequencing machines meant that the read-centric overlap approaches were not feasible
- Already in 1980's Pevzner et al. introduced an alternative assembly framework based on de Bruijn graph
- Based on a idea of a graph with fixed-length subsequences (k-mers)
- Key is that not storing read sequences – just k-mer abundance information in a graph structure

De bruijn graph construction



- continuous linear stretches within the graph
- assembler keeps information about reads coverage for each k-mer/node.

Flicek & Birney (2009) Nat Meth, 6: S6-S12.



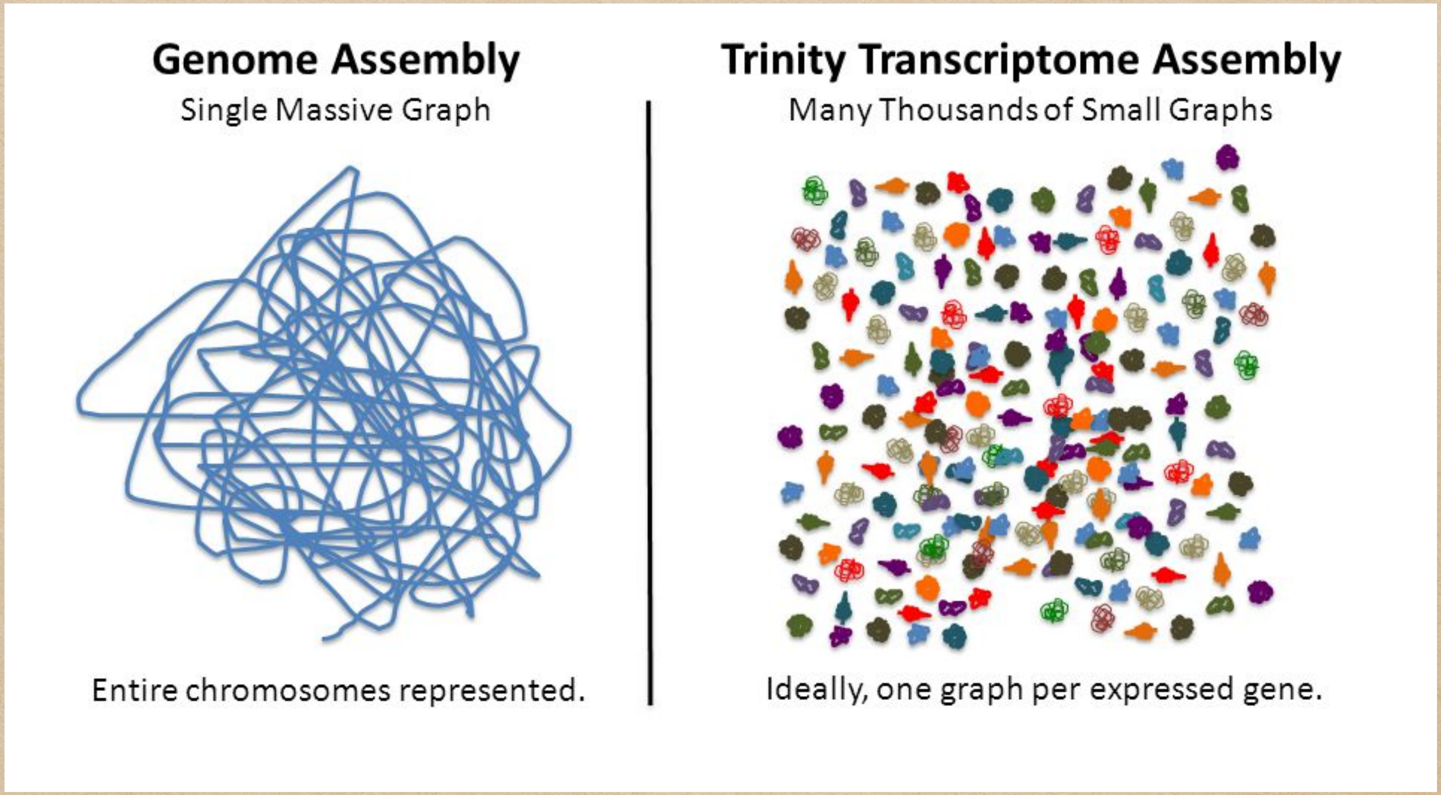
Graph is simplified to combine nodes that are associated with the continuous linear stretches into single, larger nodes of various k-mer sizes.

Error correction removes the tips and bubbles that result from sequencing errors.

Sequencing errors are low frequency tips in the graph.

Flicek & Birney (2009) Nat Meth, 6: S6-S12.

Sequence assembly: genome or transcriptome


Genome Assembly
Single Massive Graph
Entire chromosomes represented.
Trinity Transcriptome Assembly
Many Thousands of Small Graphs
Ideally, one graph per expressed gene.

Next-gen assemblers

- First de Bruijn based assembler was Newbler developed by 454 Life Sciences
  - Adapted to handle main source of error in 454 data – indels in homopolymer tracts
- Many de Bruijn assemblers subsequently developed
  - SHARCGS, VCAKE, VELVET, EULER-SR, EDENA, ABySS and ALLPATHS, SOAP
  - Most can use pair-mate information
- Slightly different approach to transcriptome assembly:
  - It has to allow many discontinuous graphs representing single transcript, including paralogs and alternatively spliced ones.
  - SOAP-Trans, Trinity

BIOINFORMATICS CREED

- Remember about biology
- Do not trust the data
- Use comparative approach
- Use statistics
- Know the limits
- Remember about biology!!!