Genomes and their structures

CB 2010/6010

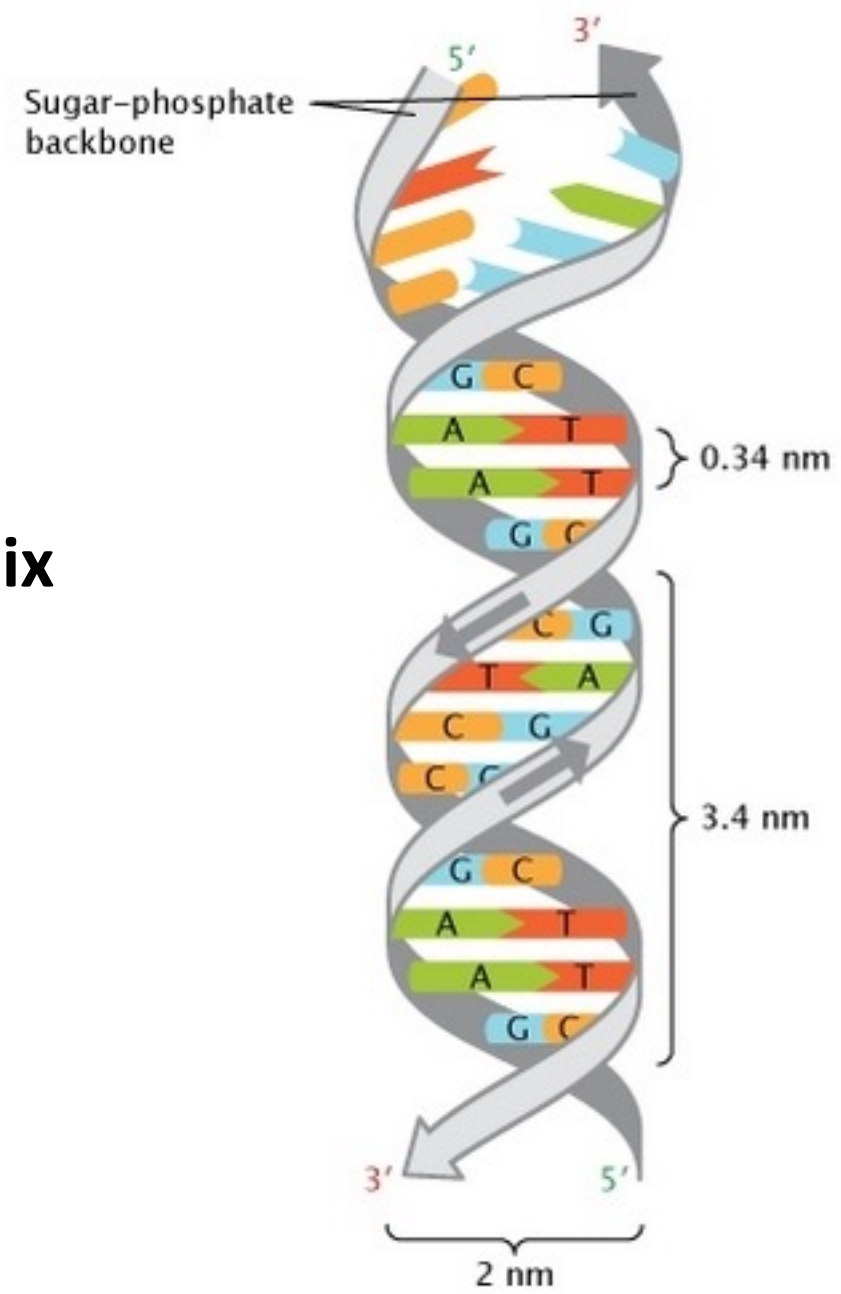
William KM Lai

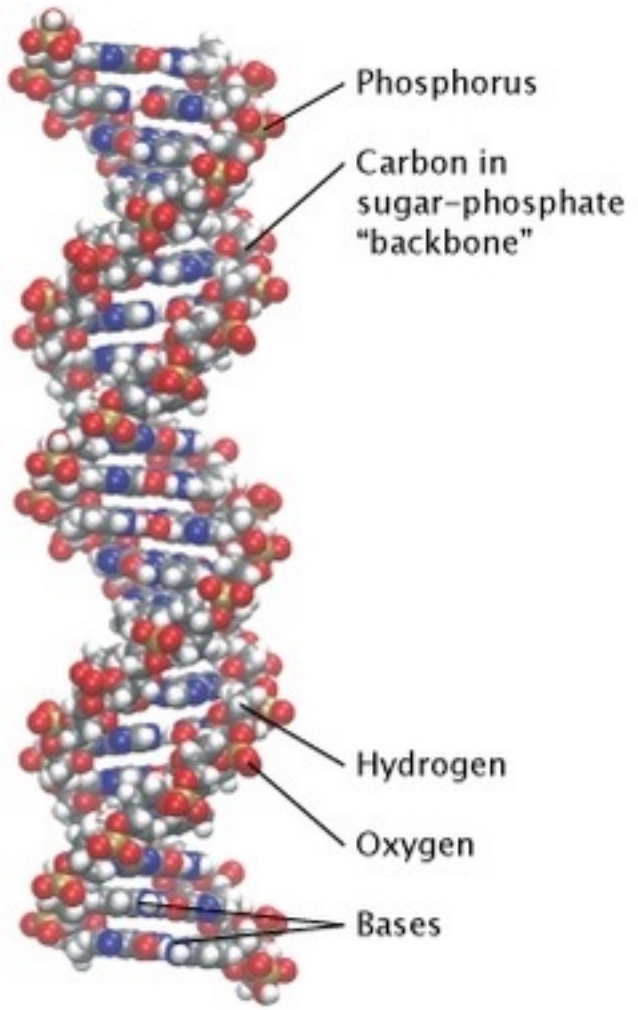
Learning objectives:

- Cover the basic terminology of molecular biology
- Overview of DNA
  - Basis of inheritance
  - Double helix - implications
- Basics of DNA sequencing
  - PCR
  - Sequencing by synthesis (SBS)
  - Nanopore sequencing

DNA is a double helix

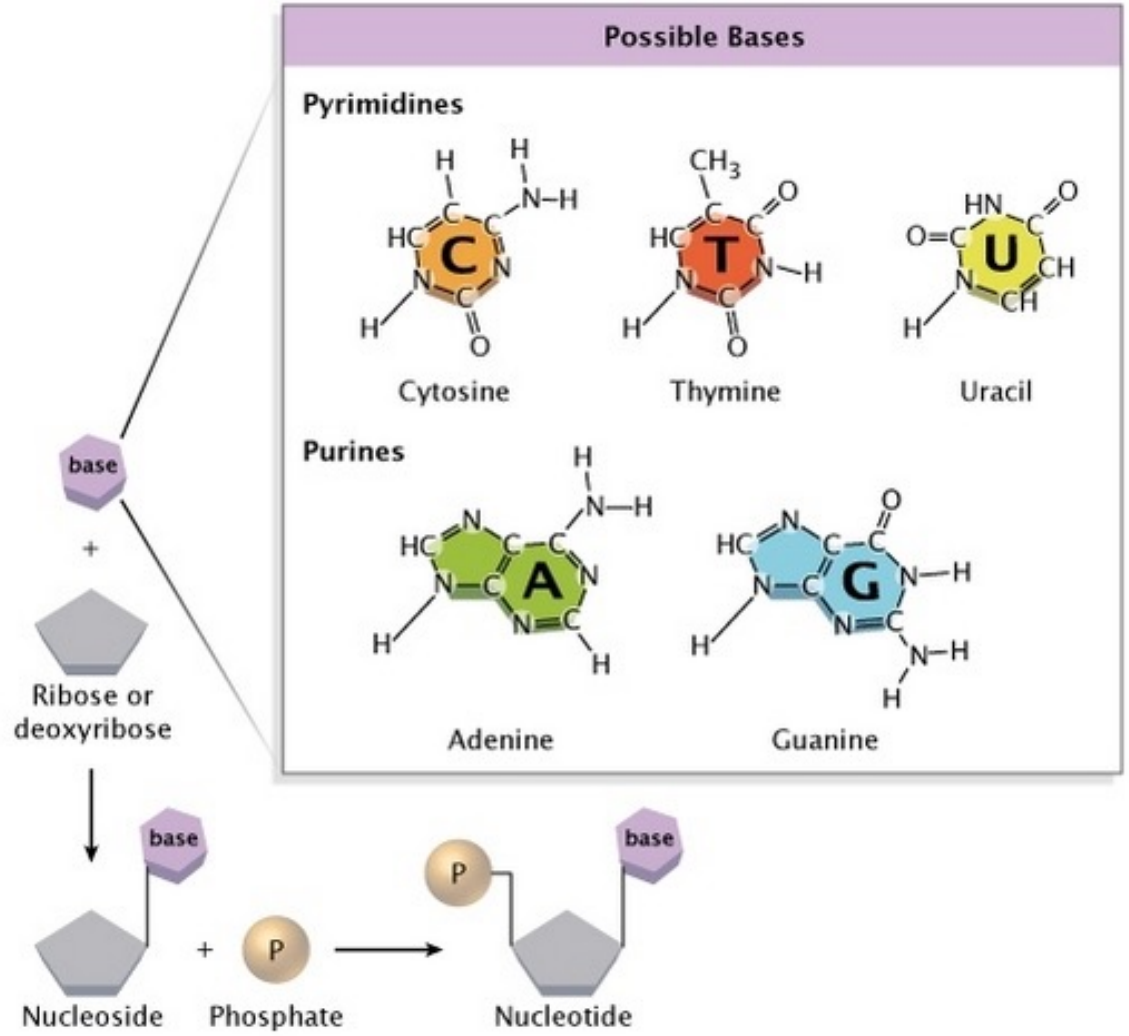
- A to T/U
- G to C





https://www.nature.com/scitable/content/ne0000/ne0000/ne0000/ne0000/104944953/73_1_2.jpg

The pattern of these in a particular order encodes the information for all life on this planet!



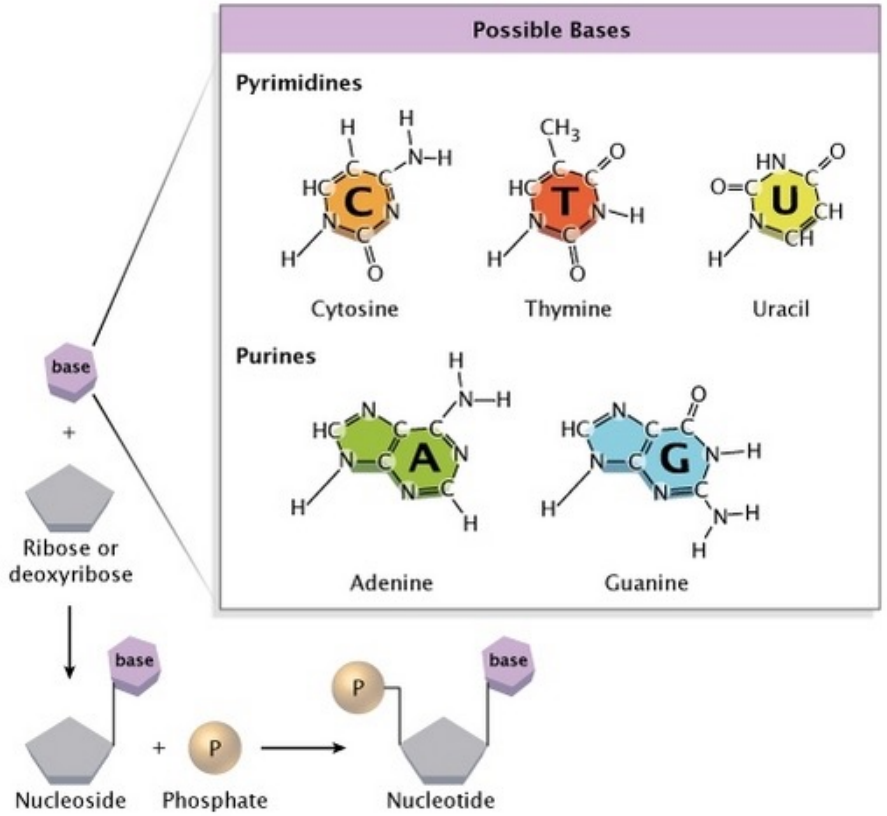
ATGCTAGCACGCATCGTCAGCGACTACTACGACGA….

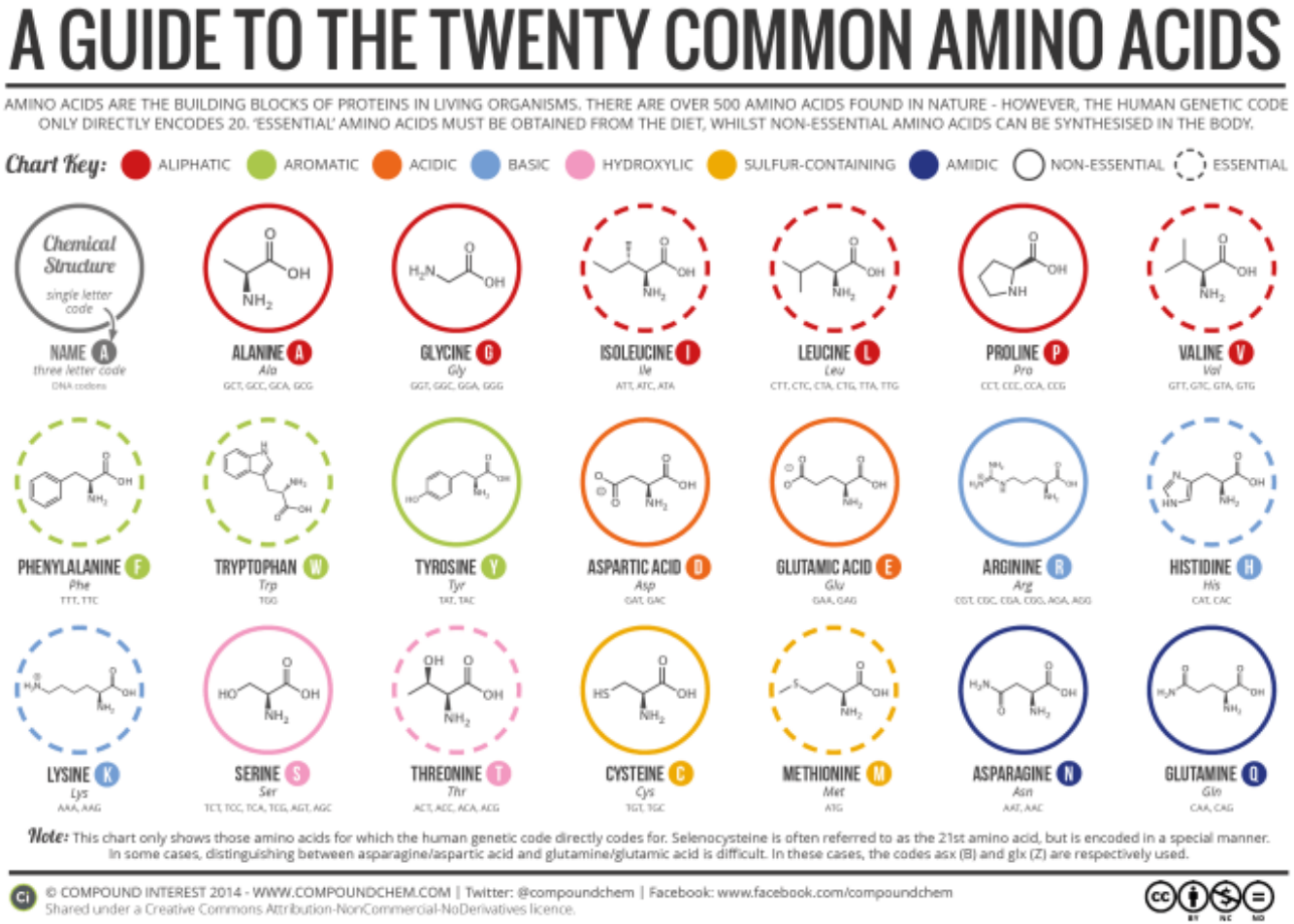
https://www.nature.com/scitable/content/ne0000/ne0000/ne0000/ne0000/104573606/70_1_2.jpg

How did we figure out DNA encoded life?

(Basis of inheritance)

Historical context: there is something else in the cell with potentially greater information density



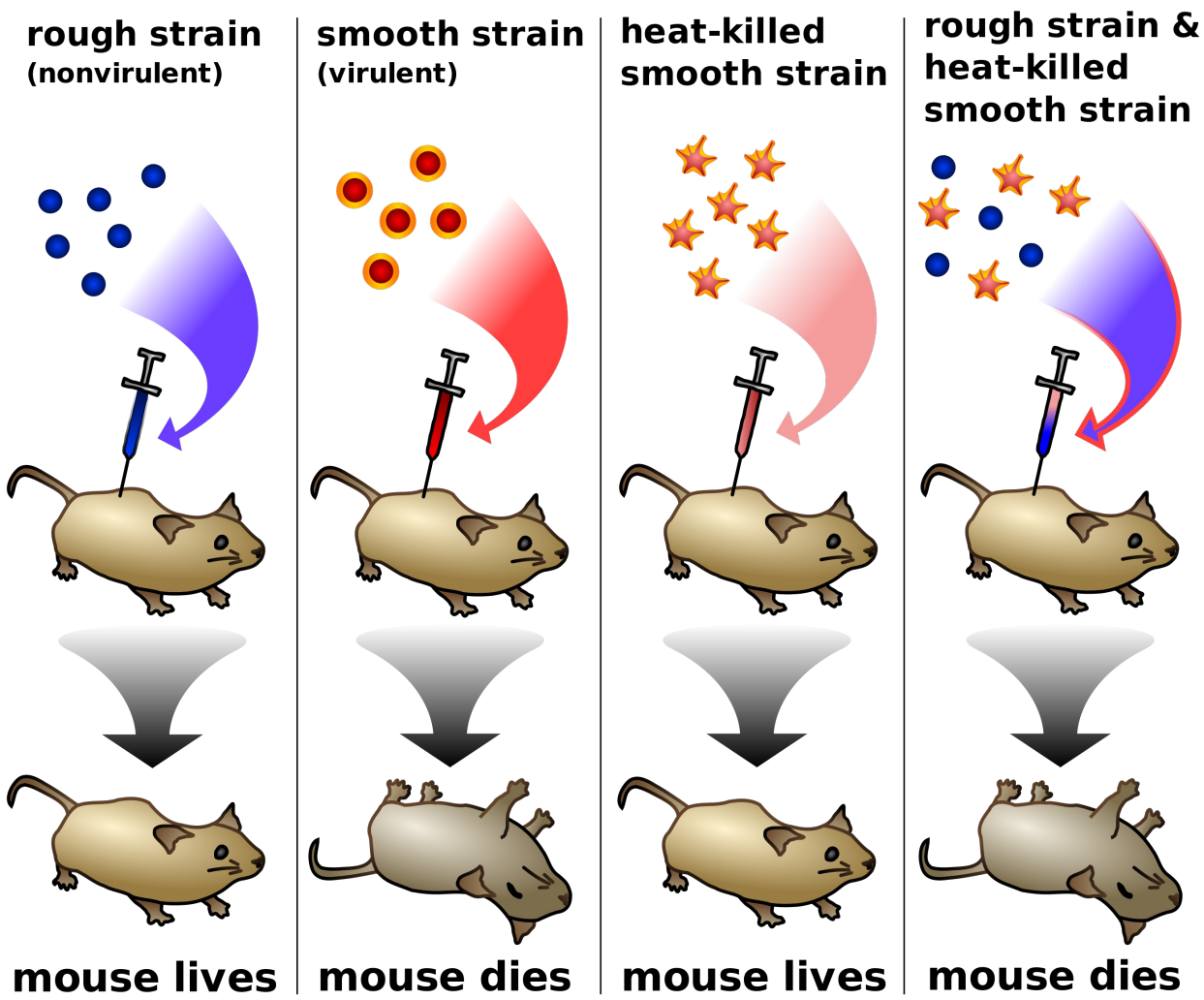


https://www.nature.com/scitable/content/ne0000/ne0000/ne0000/ne0000/104573606/70_1_2.jpg

https://i0.wp.com/www.compoundchem.com/wp-content/uploads/2014/09/20-Common-Amino-Acids-v3.png?ssl=1

So how did we figure out DNA was the basis of inheritance?

Griffith's Experiment (1928)



Like many major discoveries, incidental to the original purpose!!!

https://en.wikipedia.org/wiki/Griffith%27s_experiment#/media/File:Griffith_experiment.svg

Avery–MacLeod–McCarty experiment (1944)

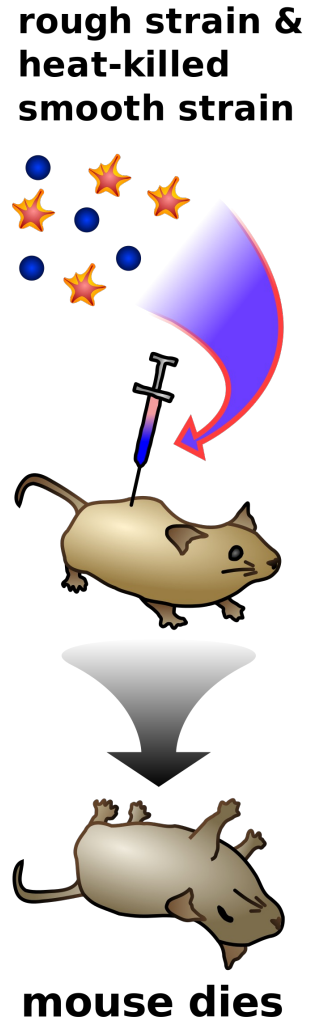
Basic experiment:

1. Grow up a lot of heat-killed smooth strain (~75 liters)

2. Biochemically purify:
   - Proteins
   - DNA
   - RNA

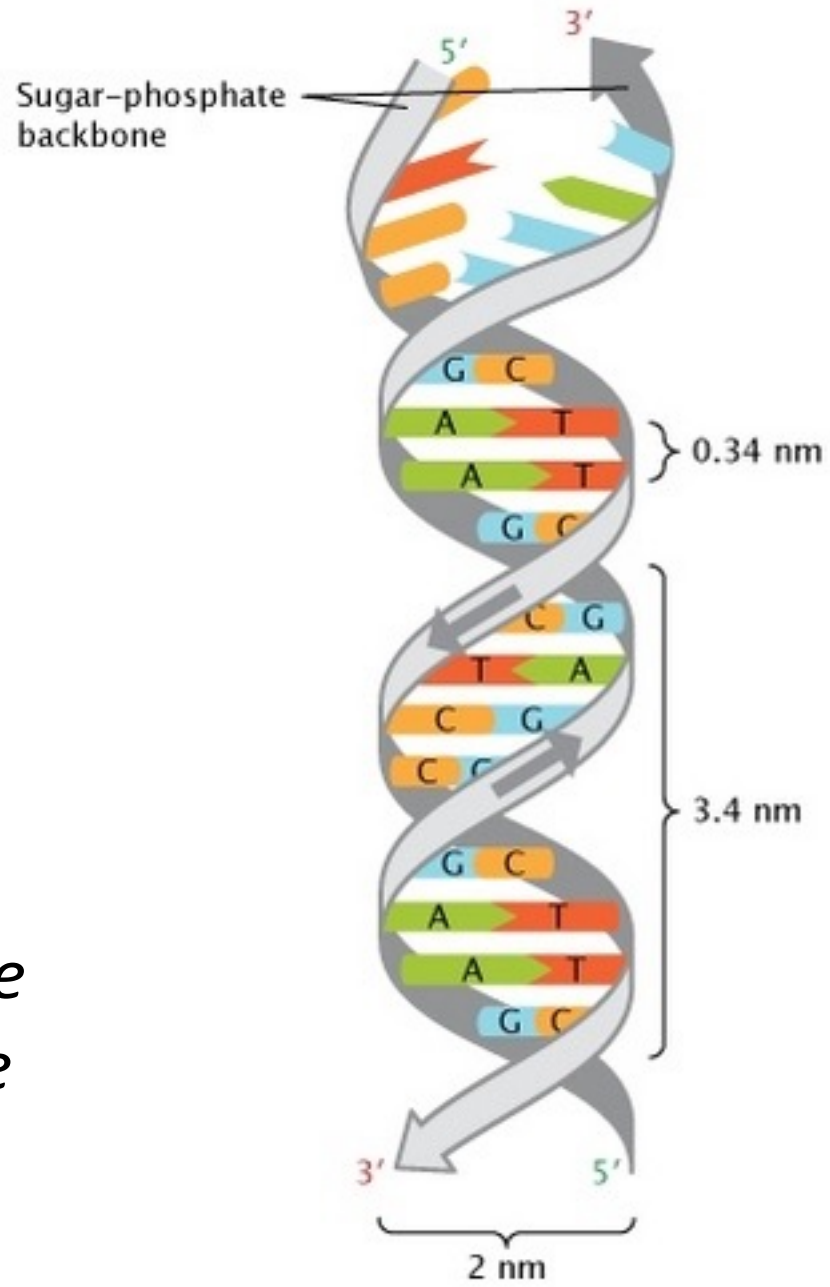
3. Repeat Griffith's experiment using ONLY protein / ONLY DNA / ONLY RNA + smooth strain

4. Mouse died only when DNA was added!



DNA is a double helix

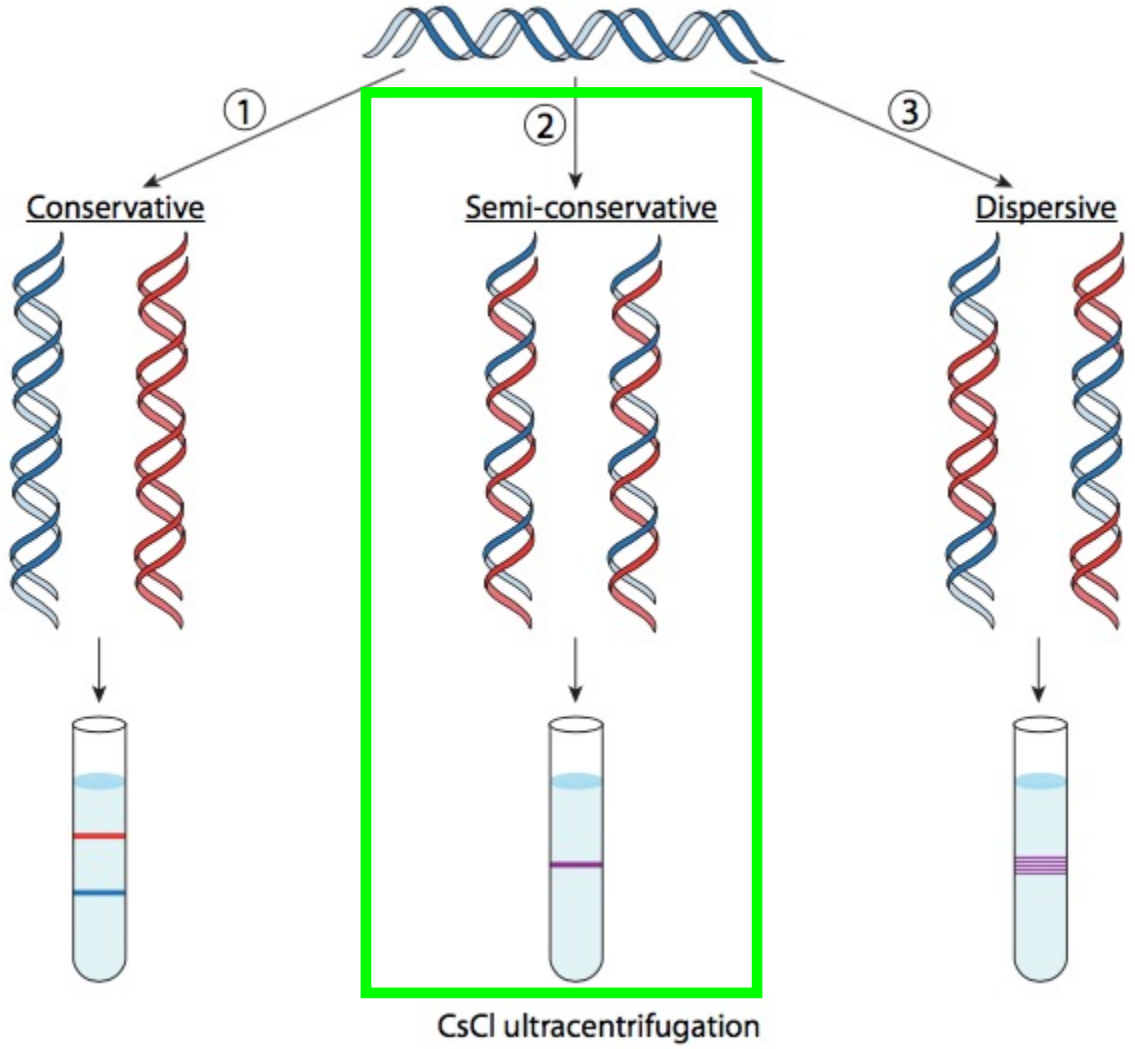
- So what?
- What does that mean?
- Why do we care?



Structure == function

“It has not escaped our notice that the pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.” – Watson and Crick (1953)

Meselson – Stahl experiment (1953)

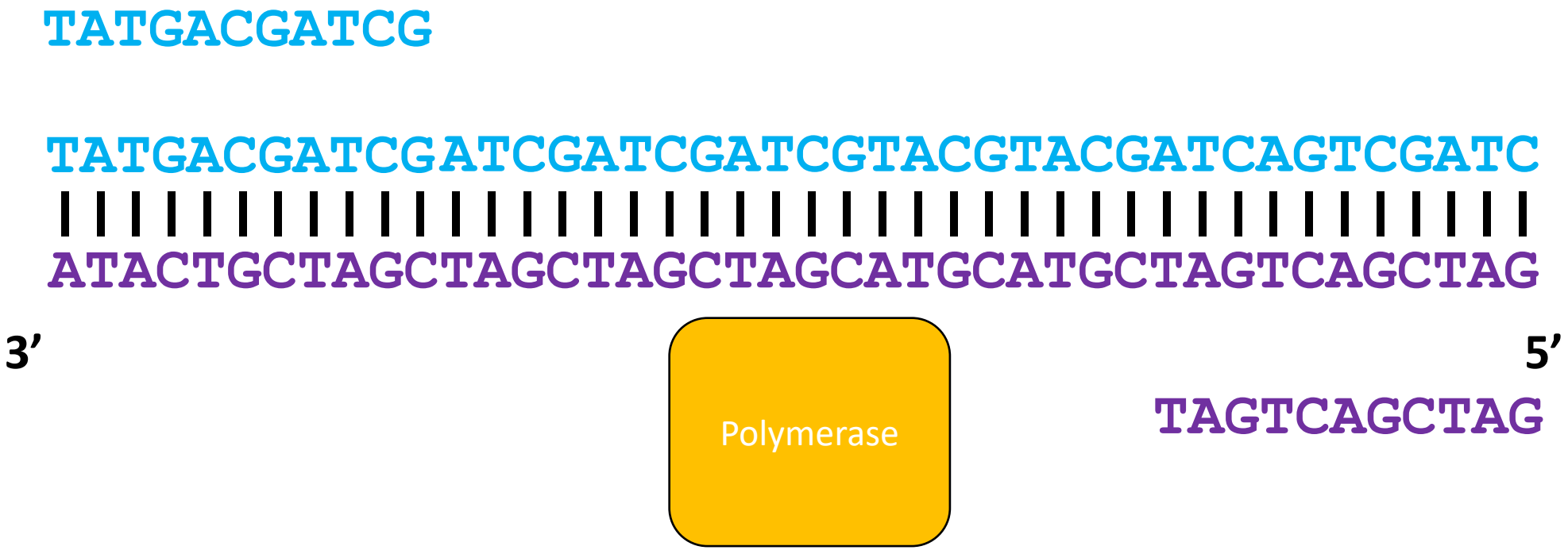


https://bio.libretexts.org/@api/deki/files/13551/Screen_Shot_2018-12-23_at_8.19.48_PM.png?revision=1

Polymerase Chain Reaction – PCR (1983)

Polymerase Chain Reaction – PCR (1983)

Polymerase Chain Reaction – PCR (1983)



- Unlimited DNA!!!
- Change the ‘primers’ and we can add/remove mutations
- Reintroduce this DNA back into the genome and we can test out what changes in DNA do to the organism at precision locations and with precise mutations

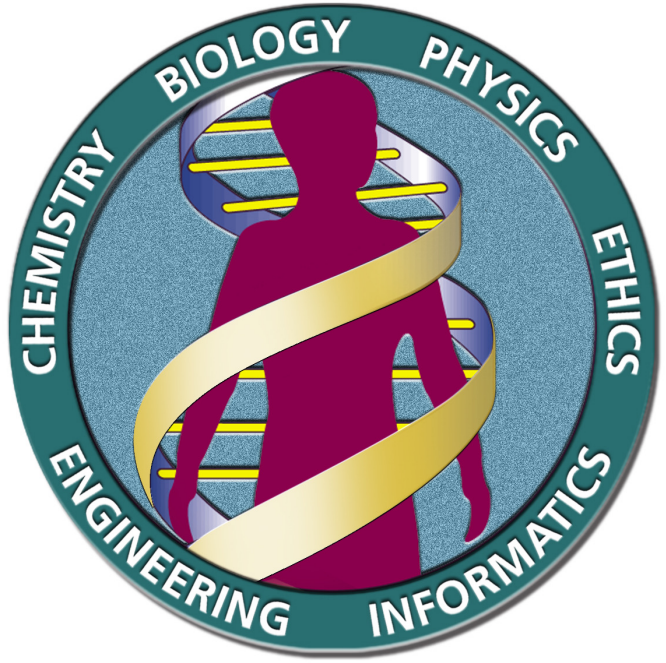
So now we have everything we need to examine everything in the genome!

Right??

We need a map!

Human Genome Project (1990-2000ish)

- Cost $3 Billion (1990) dollars and 10 years



MEDTECH

Illumina pitches $200 genomes with new line of DNA sequencers

By **Conor Hale** · Sep 30, 2022 12:06pm

Element Biosciences

Products Applications Technology Reso

Cheating!!!

Element Delivers $200 Genome on AVITI™ Benchtop Sequencing System

The complete sequence of a human genome

SERGEY NURK, SERGEY KOREN, ARANG RHIE, MIKKO RAUTIAINEN, ANDREY V. BZIKADZE, NICOLAS ALTEMOSE, LEV URALSKY, [...], AND ADAM M. PHILLIPPY +90 authors Author

SCIENCE · 31 Mar 2022 · Vol 376, Issue 6588 · pp. 44-53 · DOI: 10.1126/science.abj6987

Wait, what?

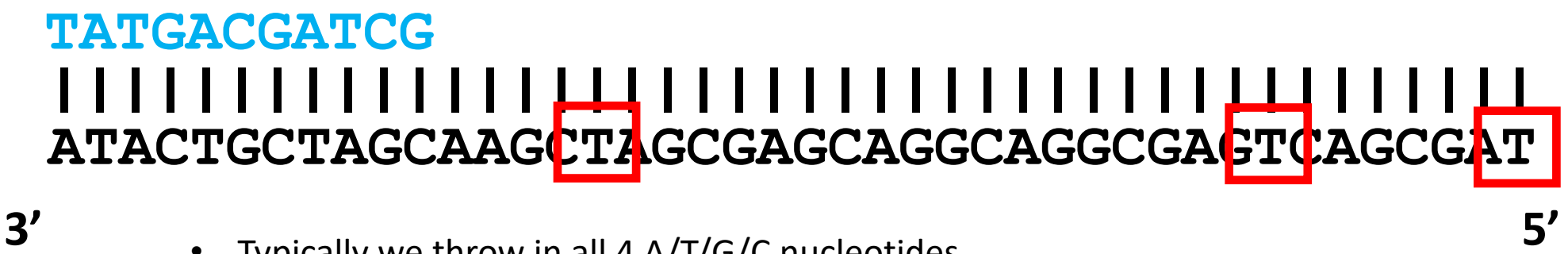
Home / 2023 / August / Scientists release the first complete sequence of a human Y chromosome

Scientists release the first complete sequence of a human Y chromosome

August 23, 2023

By Emily Cerf

How did we do it?



- Typically we throw in all 4 A/T/G/C nucleotides
- What happens if we throw in a ‘broken’ A that wrecks the rest of the molecule?

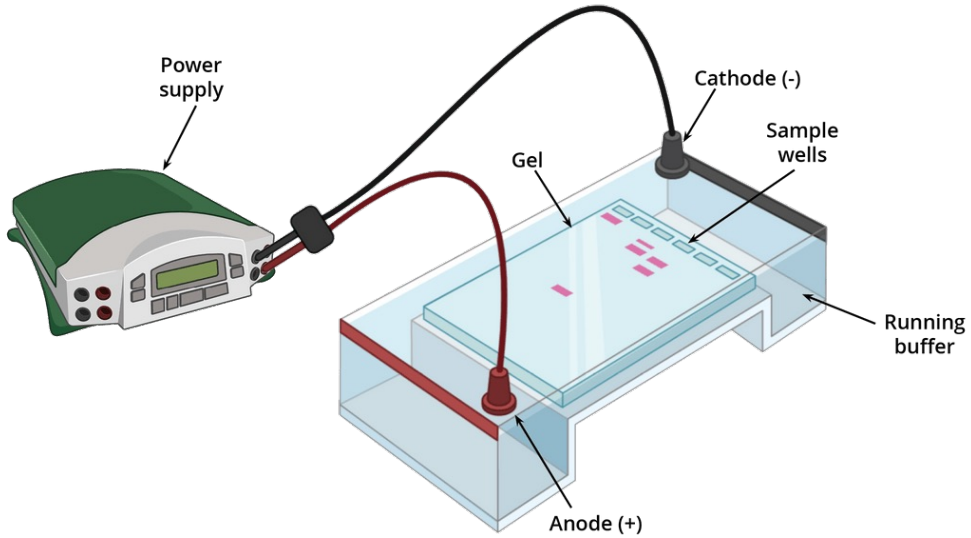
Dideoxynucloeotide - https://en.wikipedia.org/wiki/Dideoxynucleotide

TATGACGATCGTTCGA

TATGACGATCGTTCGATCGCTCGTCCGTCCGCTCA

TATGACGATCGTTCGATCGCTCGTCCGTCCGCTCAGTCGCTA

Gel Electrophoresis



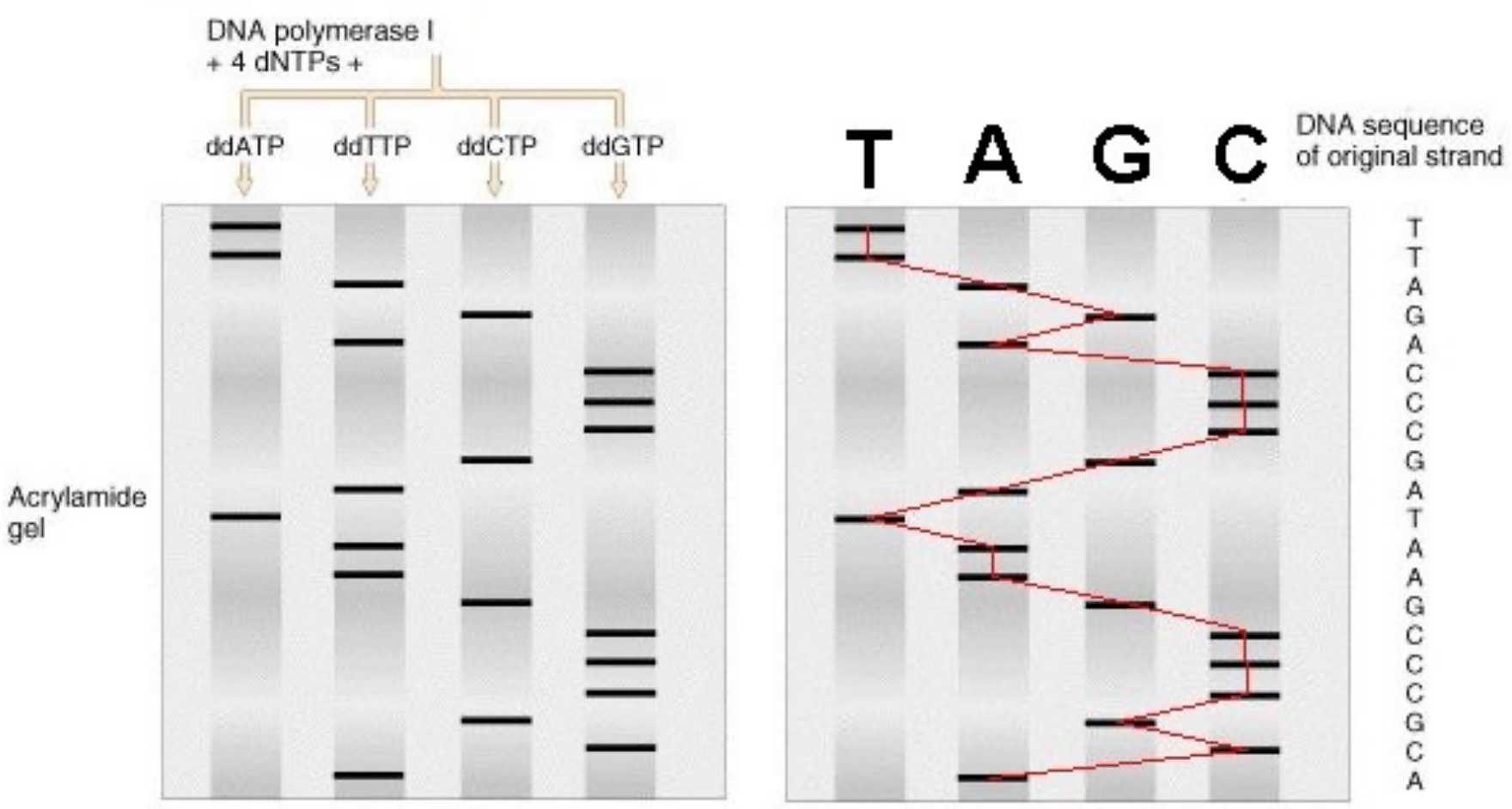
- DNA is negatively charged
- Putting it inside a gel and adding an electric current will make it move
- Small things move fast
- Heavy things moves slow

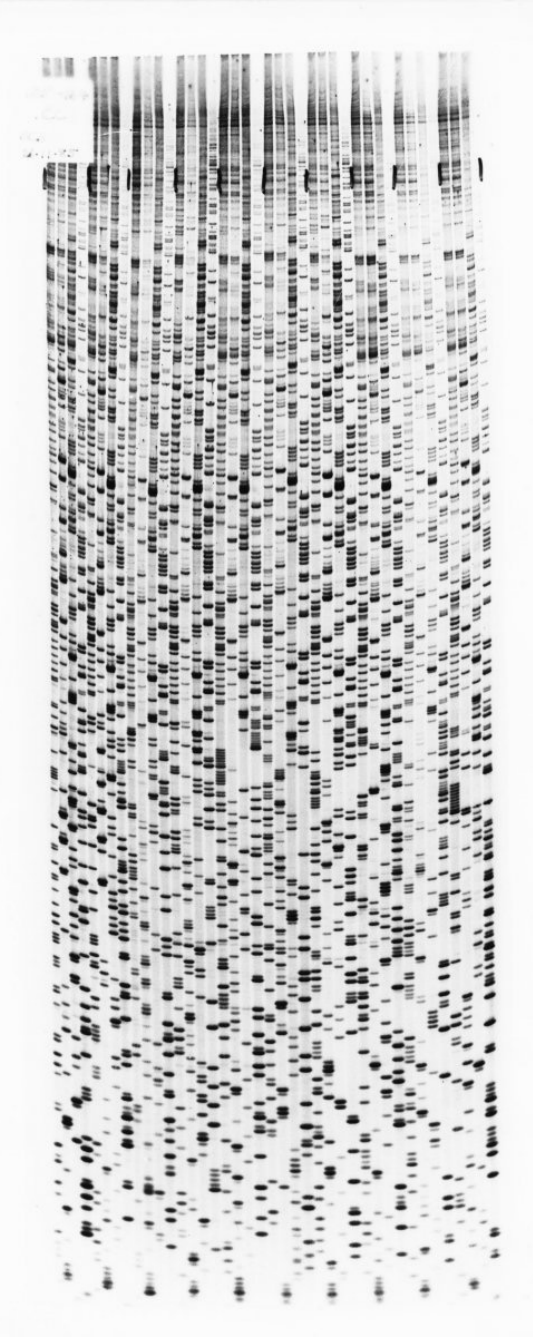
https://images.aatbio.com/universal/catalog/GelElectrophoresis/GelBoxDiagram.png

TATGACGATCGTTCGA

TATGACGATCGTTCGATCGCTCGTCCGTCCGCTCA

TATGACGATCGTTCGATCGCTCGTCCGTCCGCTCAGTCGCTA





https://www.mun.ca/biology/scarr/F12-22smc2.jpg

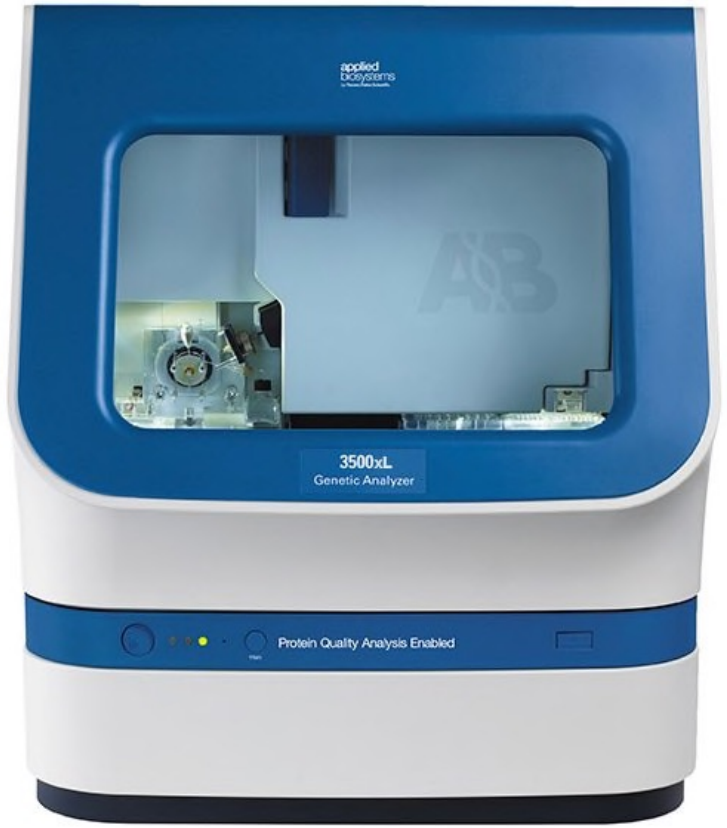
https://www.yourgenome.org/facts/where-did-dna-sequencing-begin/

Sanger sequencing

We've automated this!

Let's look at the numbers:

- A sanger sequencer can sequence up to 384 samples at a time
- Each sample can be ~800 bp
- Each run of the instrument generates ~300,000 bp



Human genome is 3,000,000,000 unique base pairs

We'd need to run the instrument AT LEAST 10,000 times to sequence the human genome

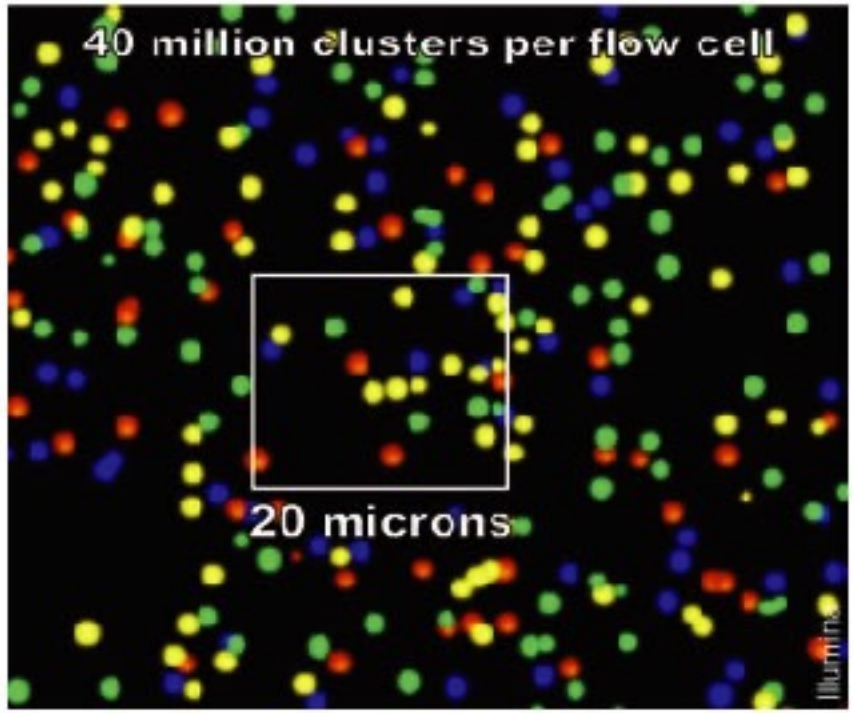
Sequencing by synthesis (SBS)

Order of events:

- Add in a dideoxynucleotide with a unique color
- Take a picture!
- Cut off the fluorophore and modify the nucleotide so it's normal again
- Rinse and repeat (literally)

A T
T A
G C
A T
C G
G C
T A
A T
A T
A T
T A
G C

What it looks like!



http://tucf-genomics.tufts.edu/home/faq

Sequencing by synthesis (SBS)

SBS can sequence billions of unique DNA strands per run

SBS chemistry is *usually* limited to ~500bp per DNA molecule

Let's look at the numbers*:

- 1 billion DNA molecules are sequenced per run
- Each sample can be ~500 bp
- Each run of the instrument generates ~500 BILLION bp

Human genome is 3 billion unique base pairs

We should be done right??

*actual numbers can be VERY different for many reasons

Why is this cheating?

MEDTECH

Illumina pitches $200 genomes with new line of DNA sequencers

By **Conor Hale** · Sep 30, 2022 12:06pm

Element Biosciences

Products Applications Technology Resou

Element Delivers $200 Genome on AVITI™ Benchtop Sequencing System

And didn't we 'complete' the genome 20+ years ago?

The complete sequence of a human genome

SERGEY NURK, SERGEY KOREN, ARANG RHIE, MIKKO RAUTIAINEN, ANDREY V. BZIKADZE, NICOLAS ALTEMOSE, LEV URALSKY, [...], AND ADAM M. PHILLIPPY +90 authors Author

SCIENCE · 31 Mar 2022 · Vol 376, Issue 6588 · pp. 44-53 · DOI: 10.1126/science.abj6987

Home / 2023 / August / Scientists release the first complete sequence of a human Y chromosome

Scientists release the first complete sequence of a human Y chromosome

August 23, 2023

By Emily Cerf

Let’s assemble a genome!

```
AGGGATCG
        TCGAGC
                GATCG
                     CGCGGCA
```

```
AGGGATCGAGCGATCGCGGCA
```

Let’s assemble a harder genome!

AAAAAAAAAAAAAAAAA

AAAAAAAAAAAAAAAAA

AACATCGTACGTCTAA

AAAAAAAAAAAAAAAAA

AACATCGTACGTCTAA

AAAAAAAAAAAAAAAAA

AAAAAAAAAAAAAAAAAACATCGTACGTCTAAAAAAAAAAAAAAAAA

Let’s assemble a harder genome!

AAAAAAAAAAAAAAAAA

AAAAAAAAAAAAAAAAA

AACATCGTACGTCTAA

AACATCGTACGTCTAA

AAAAAAAAAAAAAAAAA

AAAAAAAAAAAAAAAAA

AACATCGTACGTCTAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

Let’s assemble a harder genome!

AAAAAAAAAAAAAAAAA

AAAAAAAAAAAAAAAAA

AACATCGTACGTCTAA

AACATCGTACGTCTAA

AAAAAAAAAAAAAAAAA

AAAAAAAAAAAAAAAAA

AACATCGTACGTCTAAAAAAAAAAAAAAAAAAA

Having a pre-existing reference genome lets you easily figure out where your DNA is supposed to go!

AAAAAAAAAAAAAAAAACATCGTACGTCTAAAAAAAAAAAAAAAAA

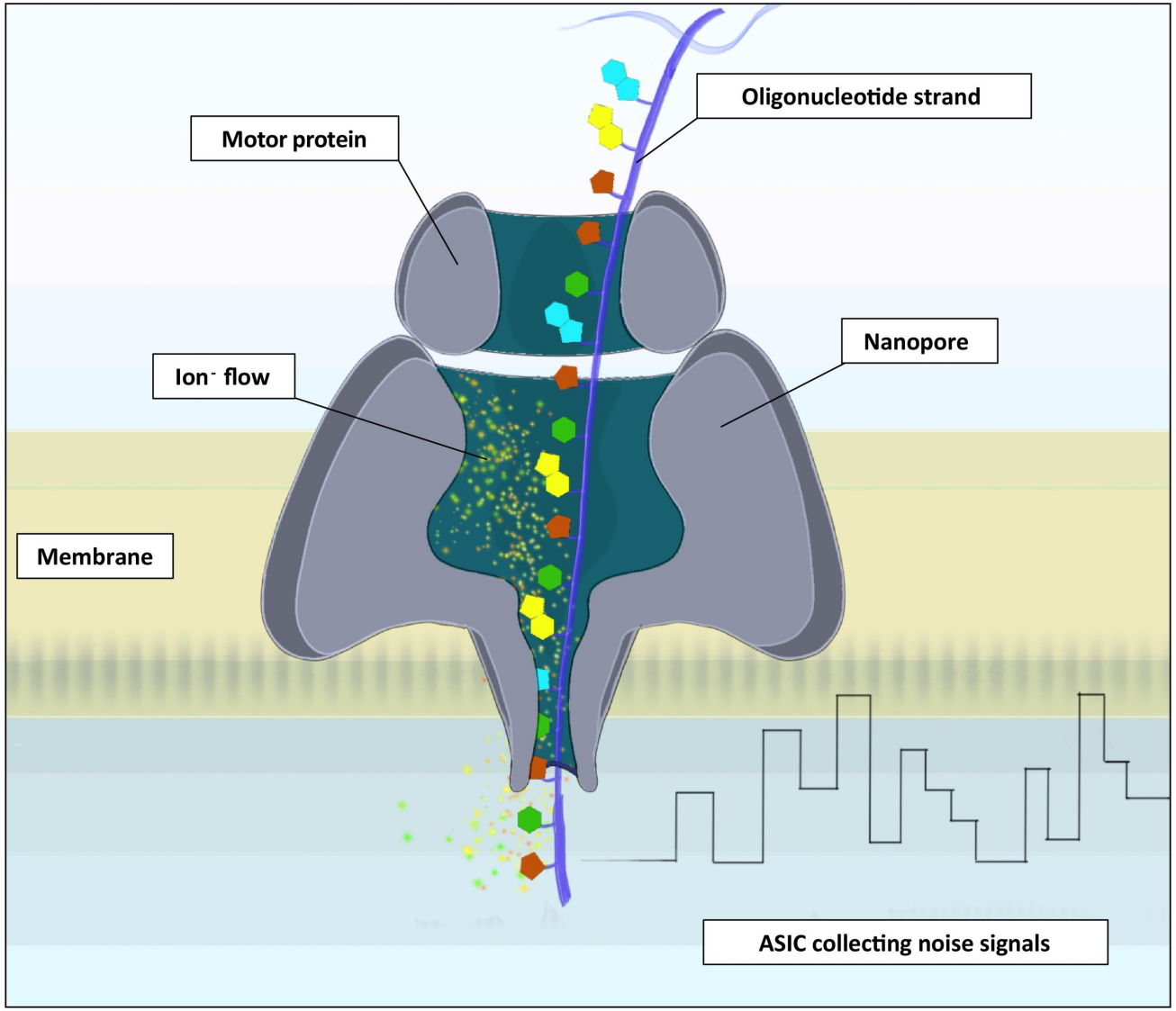
vs

AACATCGTACGTCTAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

The reason we keep 'completing' the human genome has a lot to do with how common low-complexity sequences occur in the genome…

Confounded by the fact that these regions change an awful lot person to person…

Nanopore sequencing



https://www.cell.com/trends/genetics/fulltext/S0168-9525%2821%2900306-1

Nanopore sequencing

Nanopore sequencing can sequence ~1 million DNA molecules

Exact length varies by company, but let's say 50,000 bp per molecule

Let's look at the numbers:

- 1 million DNA molecules are sequenced per run
- Each sample can be ~50,000 bp
- Each run of the instrument generates ~5 BILLION bp

Human genome is 3 billion unique base pairs

We should be done right??

SBS + Nanopore sequencing

SBS sequencing is high quality but terrible with low-complexity regions

Nanopore sequencing is lower quality by better with low-complexity regions

Most modern approaches use both!!!

Random Notes (things I didn't cover but are cool to know):

- Thousands of bacterial artificial chromosomes (BACs) containing huge chunks of the human genome gave us unlimited DNA samples that were used for sequencing
  - https://www.genome.gov/genetics-glossary/Bacterial-Artificial-Chromosome
- Sequencing technology is still actively evolving and new companies appear every year with new ideas on how to solve biochemical and computational problems
  - https://www.elementbiosciences.com/
  - https://singulargenomics.com/