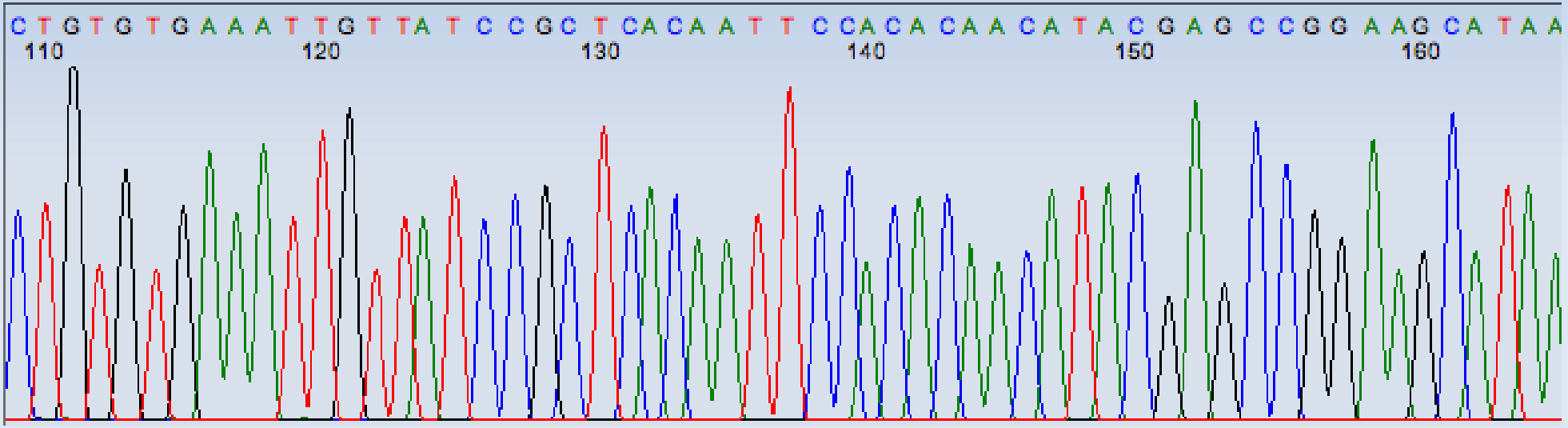
Topic 9.1: Genomes, Sequencing



Sequencing: find the order of nucleotides in a DNA strand

Genomes

Remember, our genome is information about each one of us as individuals – how our cells and bodies are built.

Some questions about eukaryotic genomes:

- Why are they fragmented?
  - Why are genes ”everywhere”? Why are they not ”organized”?

    DNA Transposons - ”jumping” genes found by Barbara McClintock in corn
  - Why is a gene ”in pieces”? And not all in one place?

    Why do we have exons and introns (unlike bacteria)?

    Why do the different types of eukayotes have different numbers and sizes of introns?

Yeast have few introns, plants have many, introns are 25% of our genomes!
1977: introns discovered; 2009: mechanism for creating introns found: ”introner” genomic parasitic DNA sequences that have been creating introns (by copying and pasting themselves into genomes) for billions of years!

Genome Sequencing

Number of base pairs in human genome:

~**3 billion** (haploid – single set of chromosomes)

Cost of sequencing 1st human genome: **$2.7 bn**

Started ~1990. Completed in stages:

Working draft: Feb 2001

92% complete: April 2003

Telomere-to-telomere completion: **April 2022**

Whose genome was it?

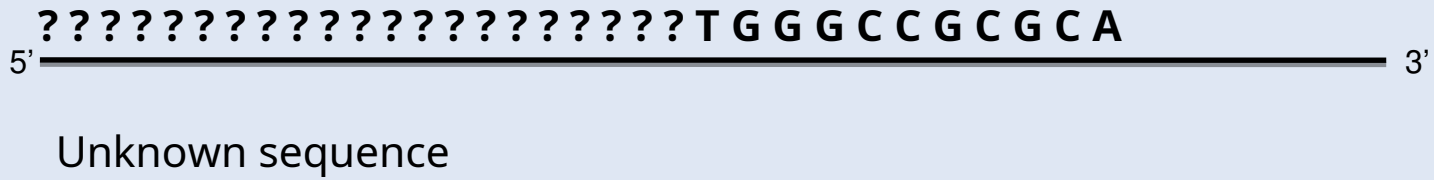
Many people, mainly one anonymous blood donor from Buffalo, NY

DNA Sequencing

DNA Sequencing is the process of taking a strand of DNA and figuring out the sequence of nucleotides in the strand

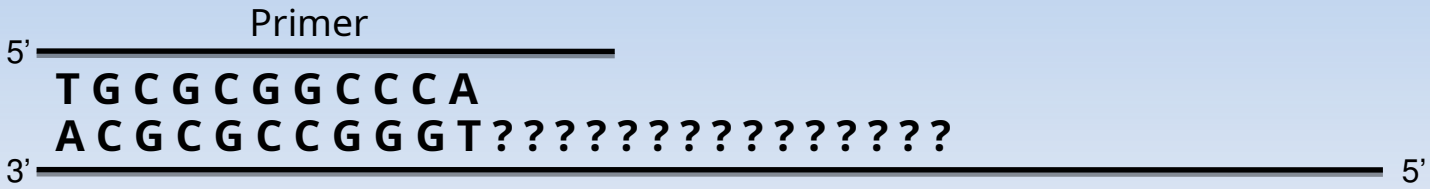
Chain termination method (Sanger):

We start with a mostly unknown sequence – suppose we know the 3' end or we add a known sequence at the 3' end:

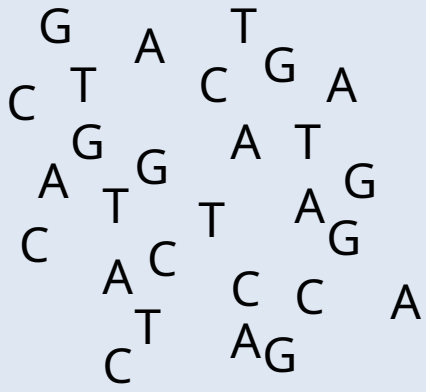


Sequencing: Sanger method

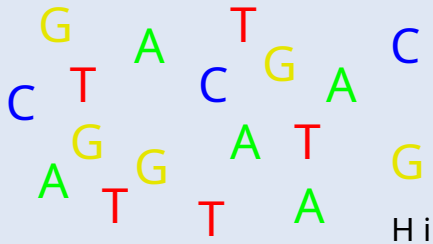
Prime unknown sequence (primer binds to vector)



Regular Nucleotides



Dideoxy Nucleotides

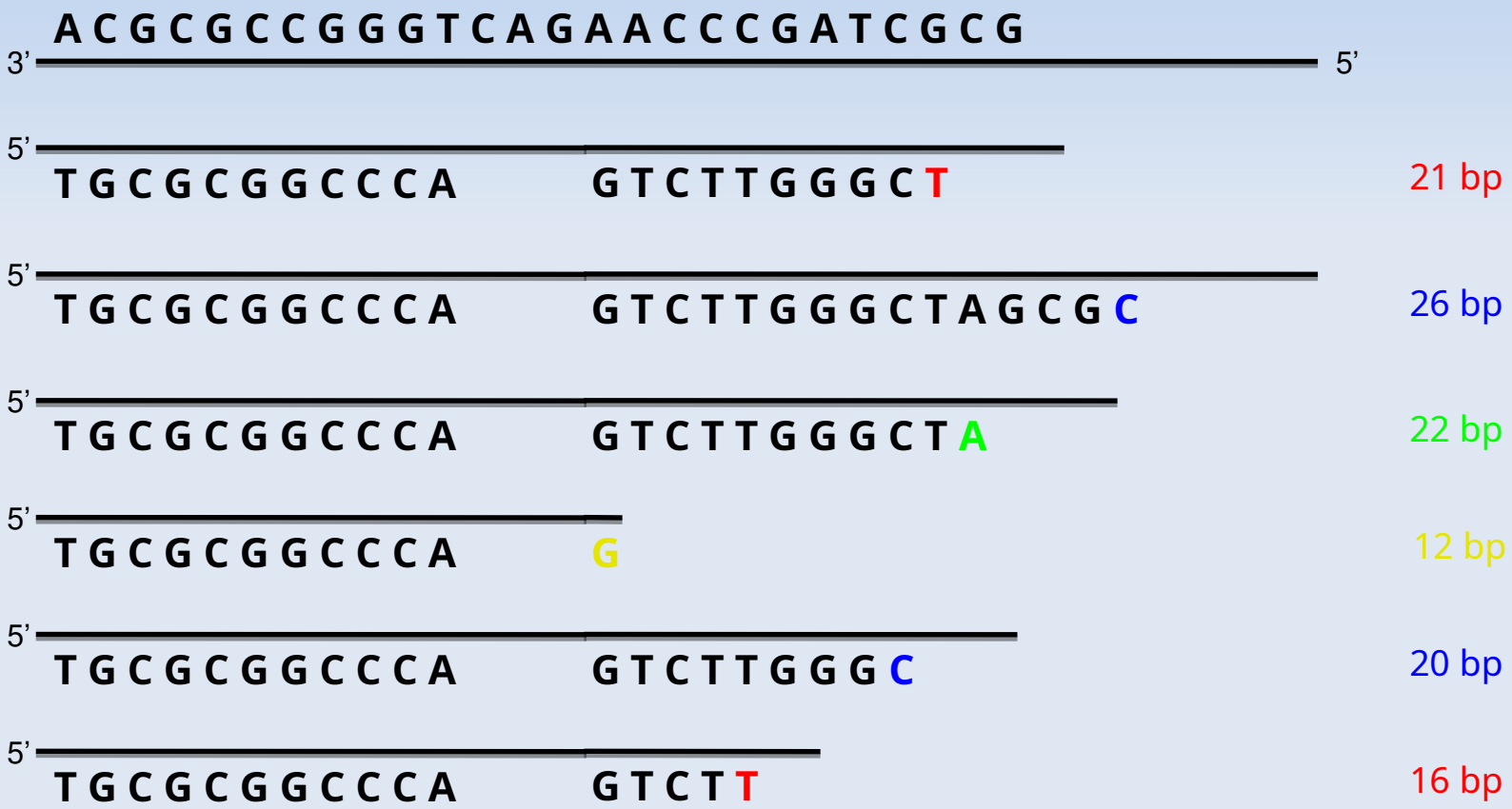


H instead of OH: cannot form phosphodiester bond with next nucleotide

Add polymerase enzyme, ”regular” nucleotides and labelled terminating dideoxy nucleotides

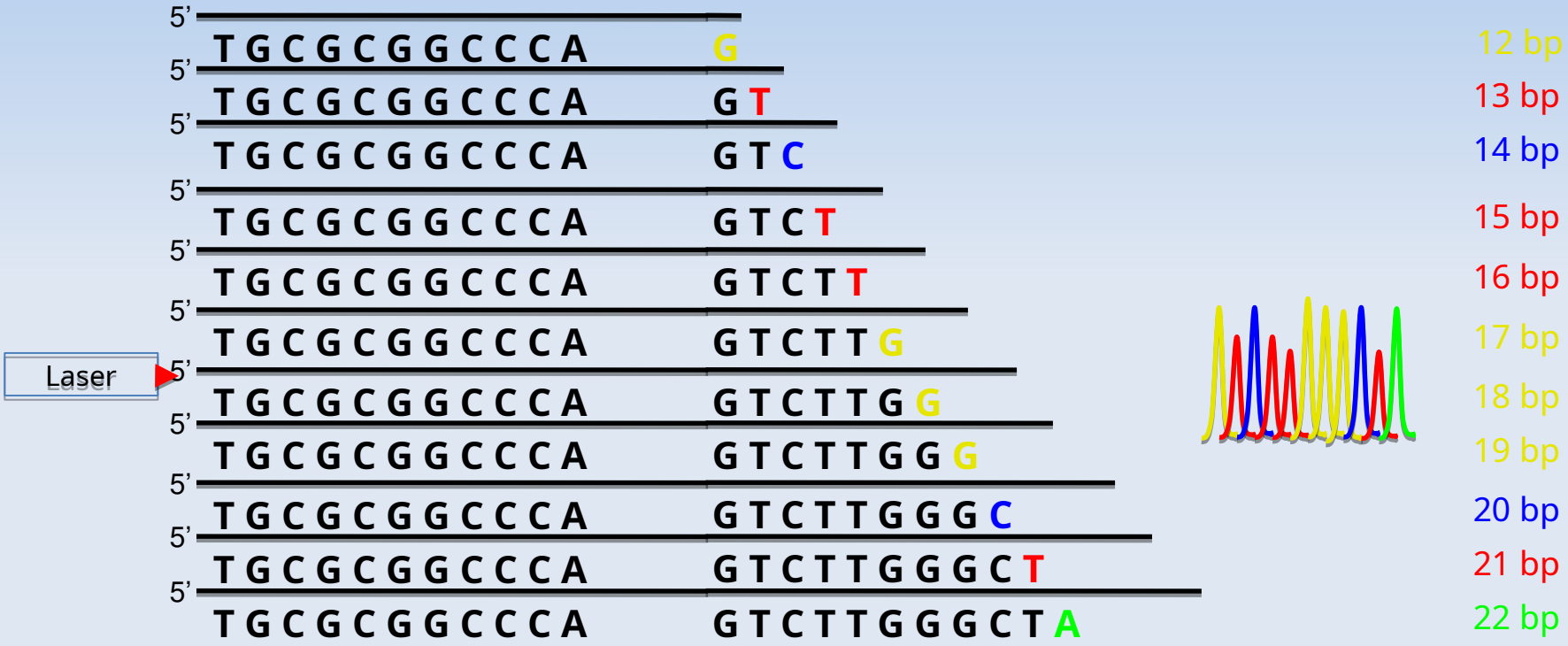
Sequencing: Sanger method

Get many terminated sequences of different lengths

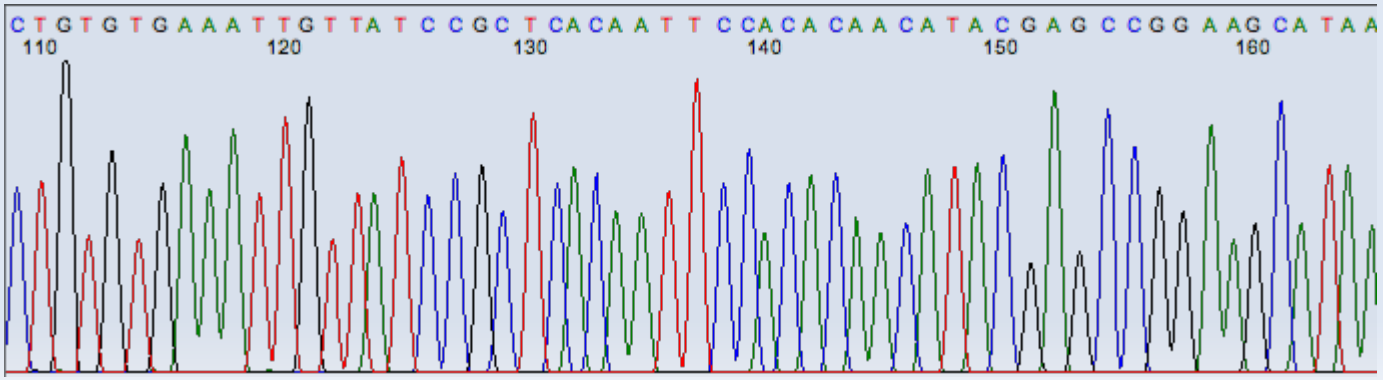


Sequencing: Sanger method

Gel electrophoresis separates products by size; Laser excitation of fluorescent tag → detector ”reads” the terminating nucleotides



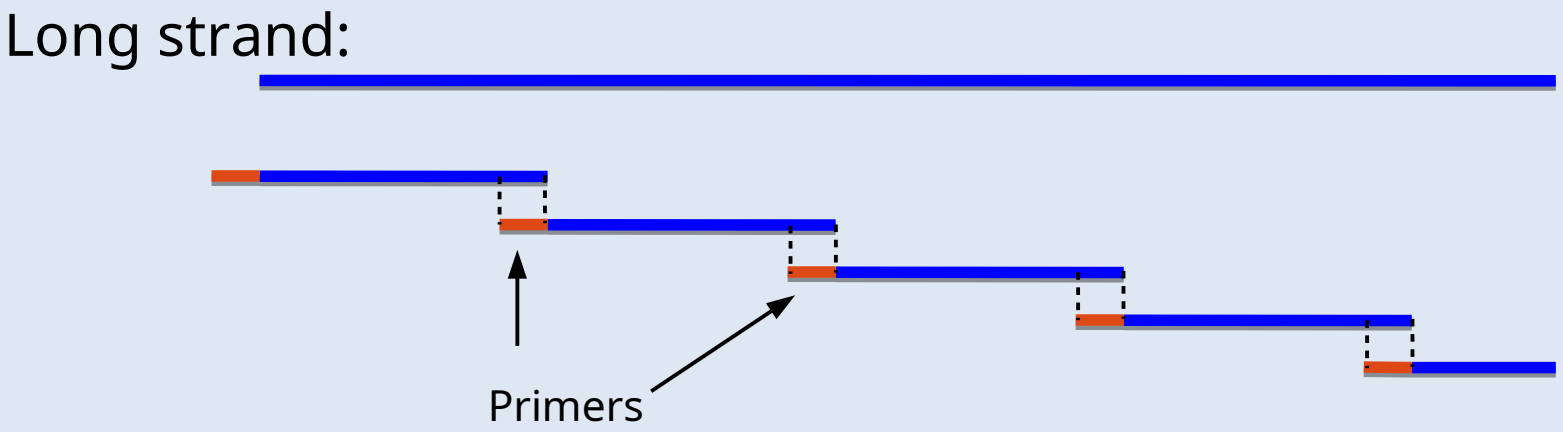
Gives us a chromatogram 600-1000 bp



Primer walking

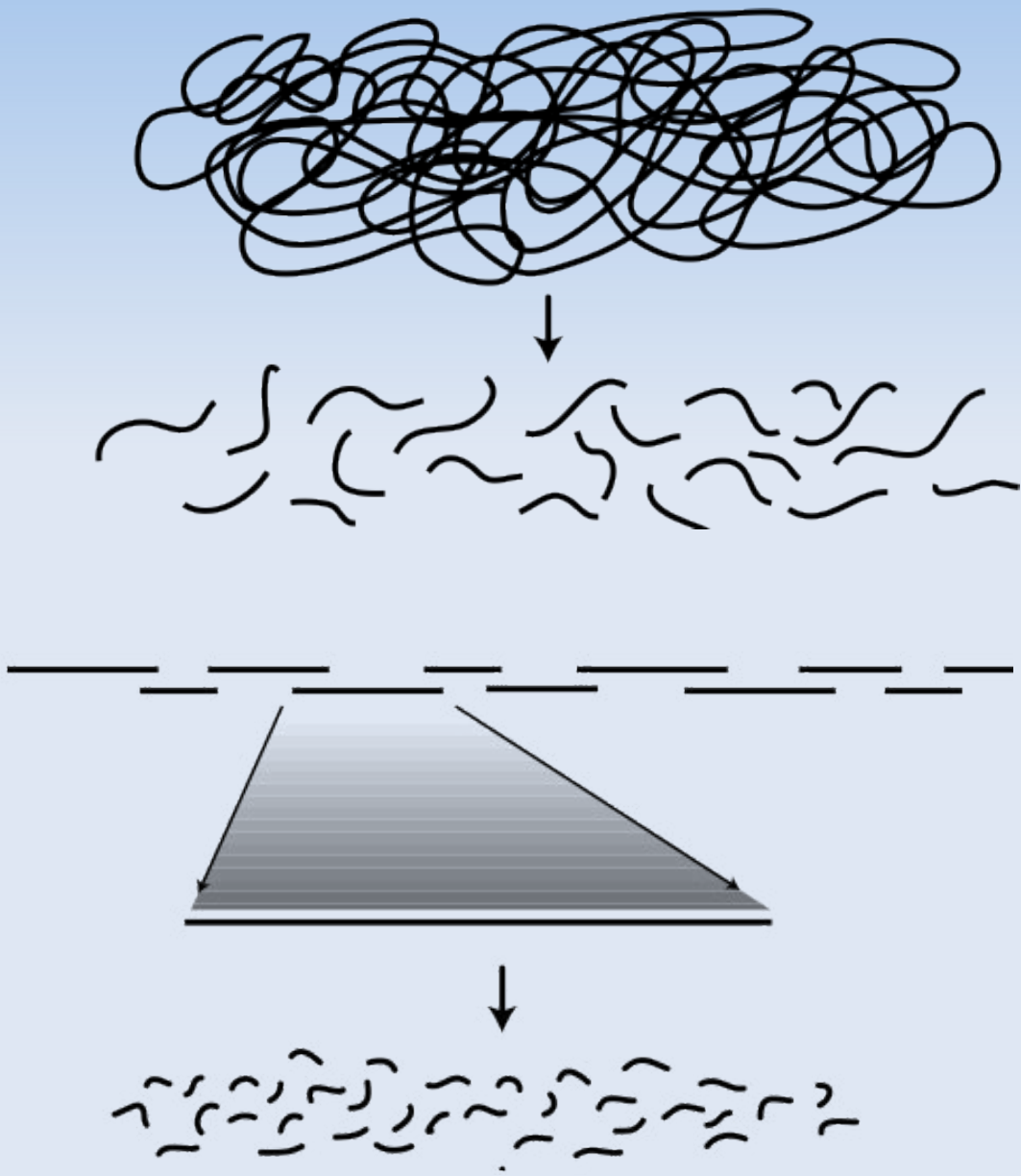
Sanger method works for sequences up to 1 Kbp

For longer sequences we have to use primer walking: the tail end of one sequencing run as the primer for the next:



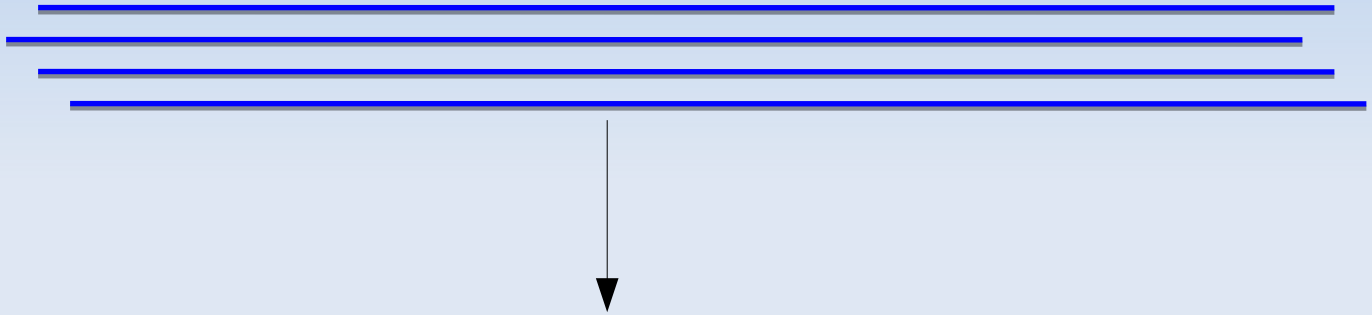
First Human Genome

- The public human genome project took a slow-but-sure approach of dividing the genome into large fragments (100-200 Kbp).
- A second cutting step of each BAC contig into 2Kbp fragments allowed sequencing

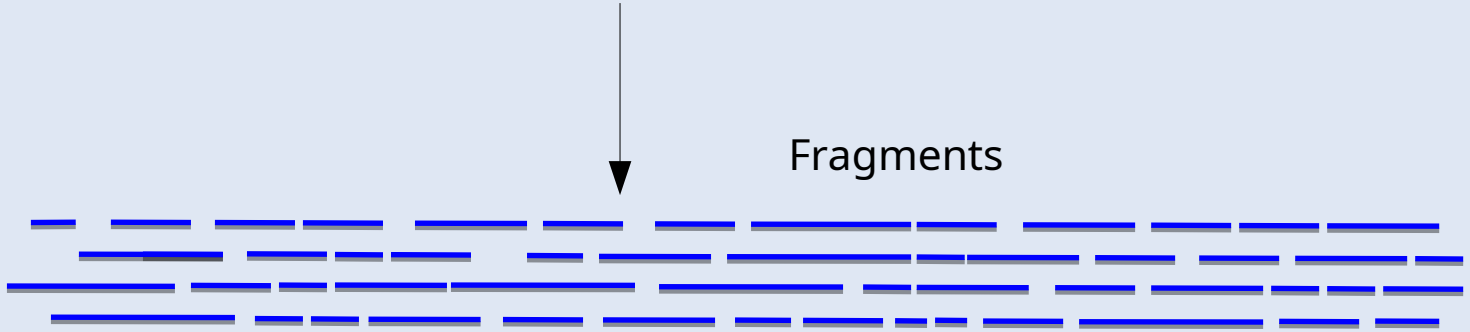


Shotgun Sequencing

Shotgun sequencing starts with many copies of large sequences:



Make enzymatic cuts, sonicate, or hydroshear

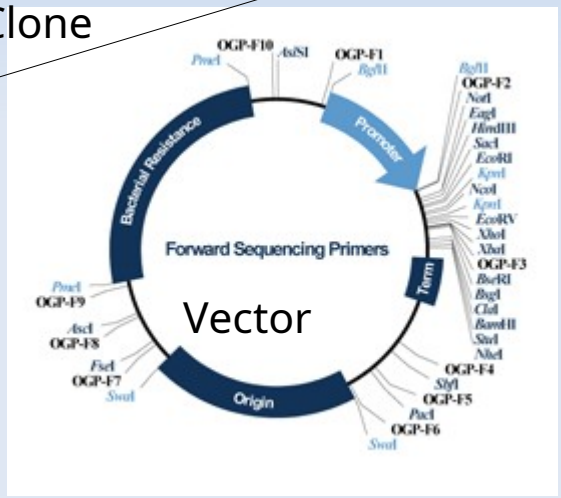


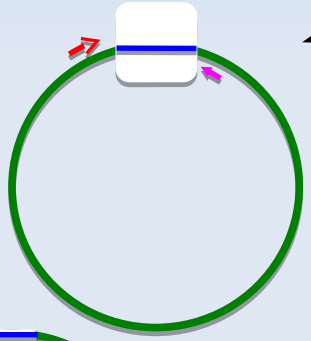
Shotgun Sequencing

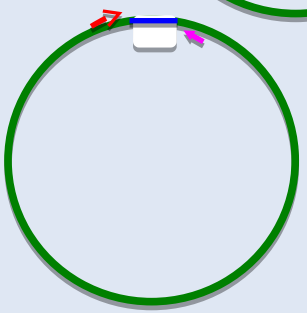
Clone fragments into a vector, amplify, isolate

Fragments:

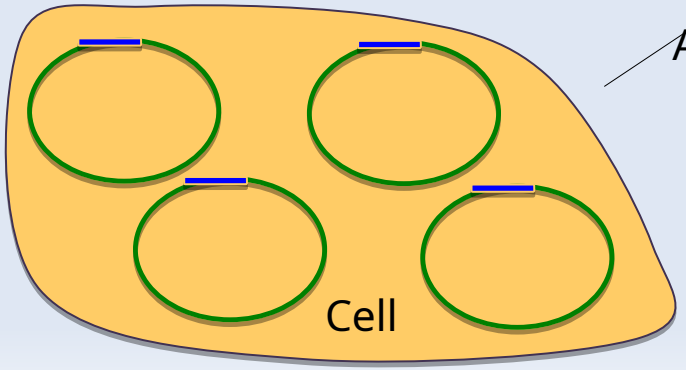
Clone



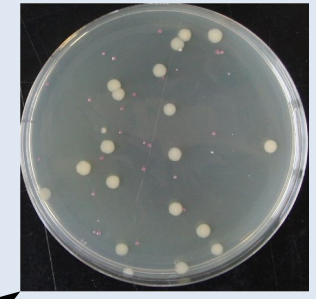




plate



Amplify



Isolate and sequence inserts

Celera Approach

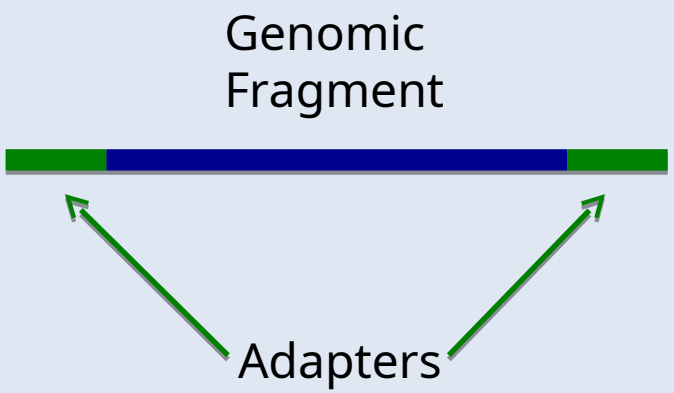
The private (Celera) genome project took the riskier ”**whole genome shotgun**” approach

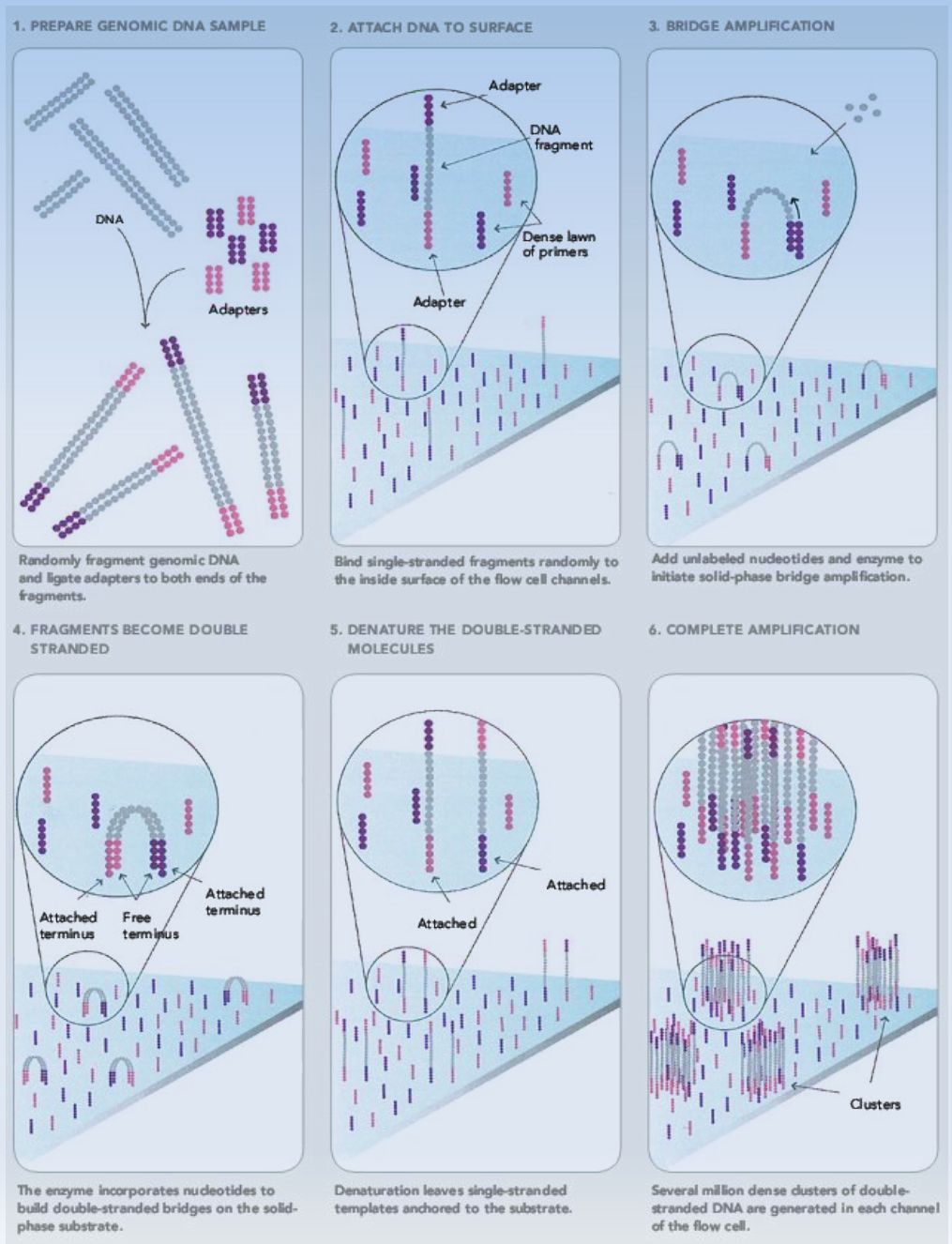
This was faster but required a lot more bioinformatics computations to ”assemble” the genome

At the time, the computational work was faster than the chemical sequencing part.

Deep or 2nd Generation Sequencing: Solexa sequencing by synthesis

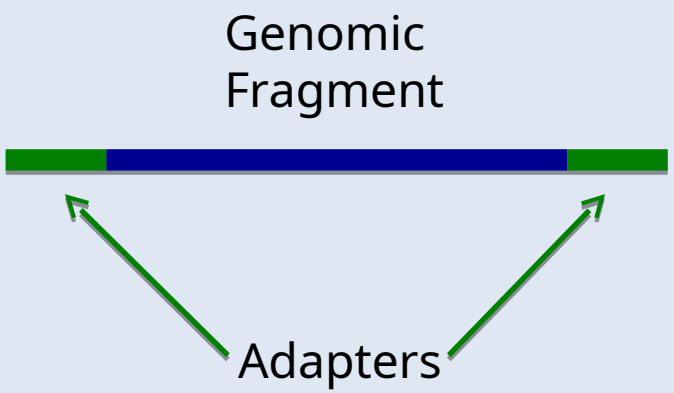
High throughput – many (smaller) fragments, use computational techniques to stitch fragments together.

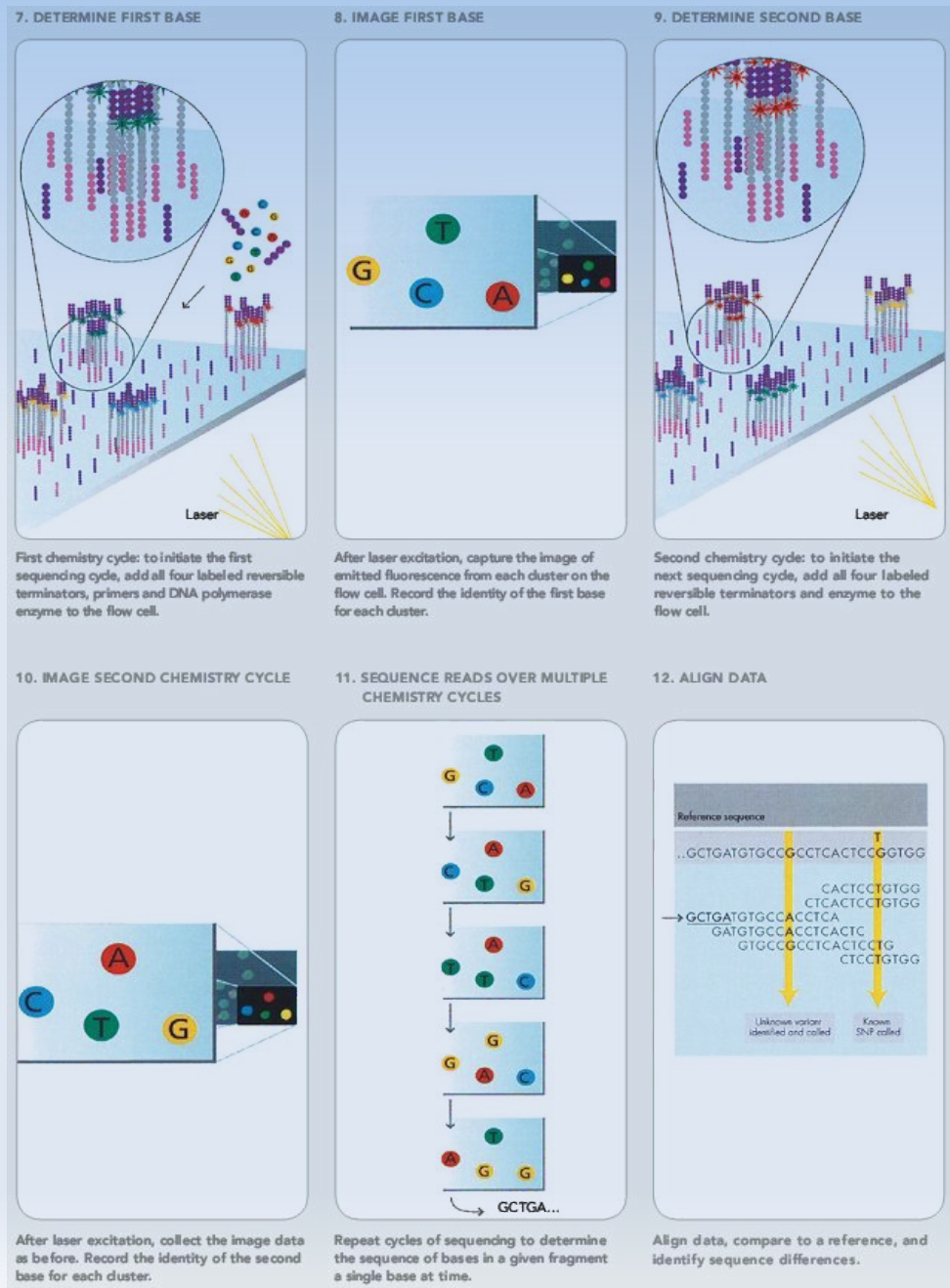




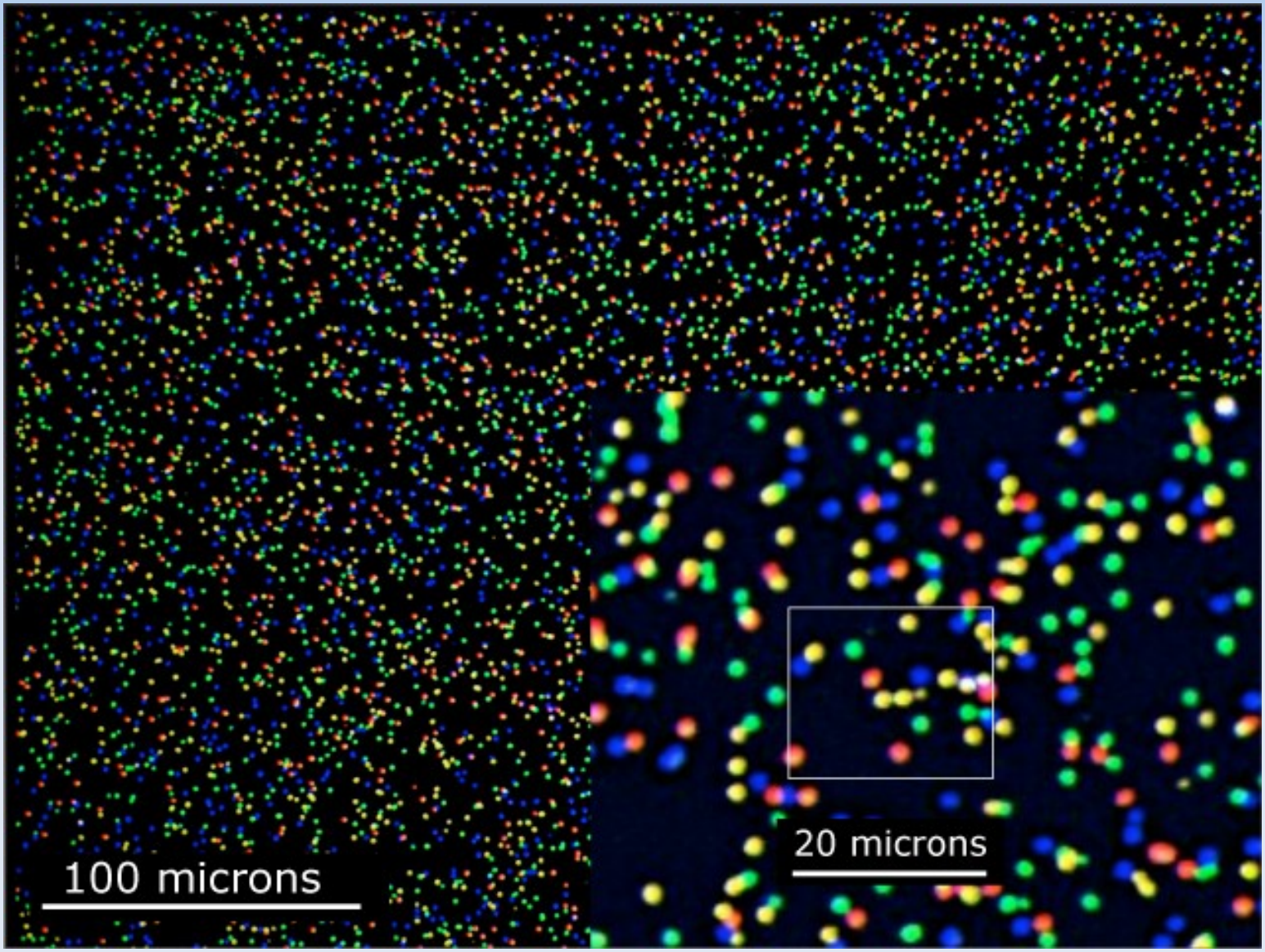
2nd Generation Sequencing

High throughput – many (smaller) fragments, use computational techniques to stitch fragments together.





2nd Generation Sequencing



Shotgun Sequencing by Hybridization

(a) Repeats of identical ~200-base DNA fragments are immobilized on a glass surface.

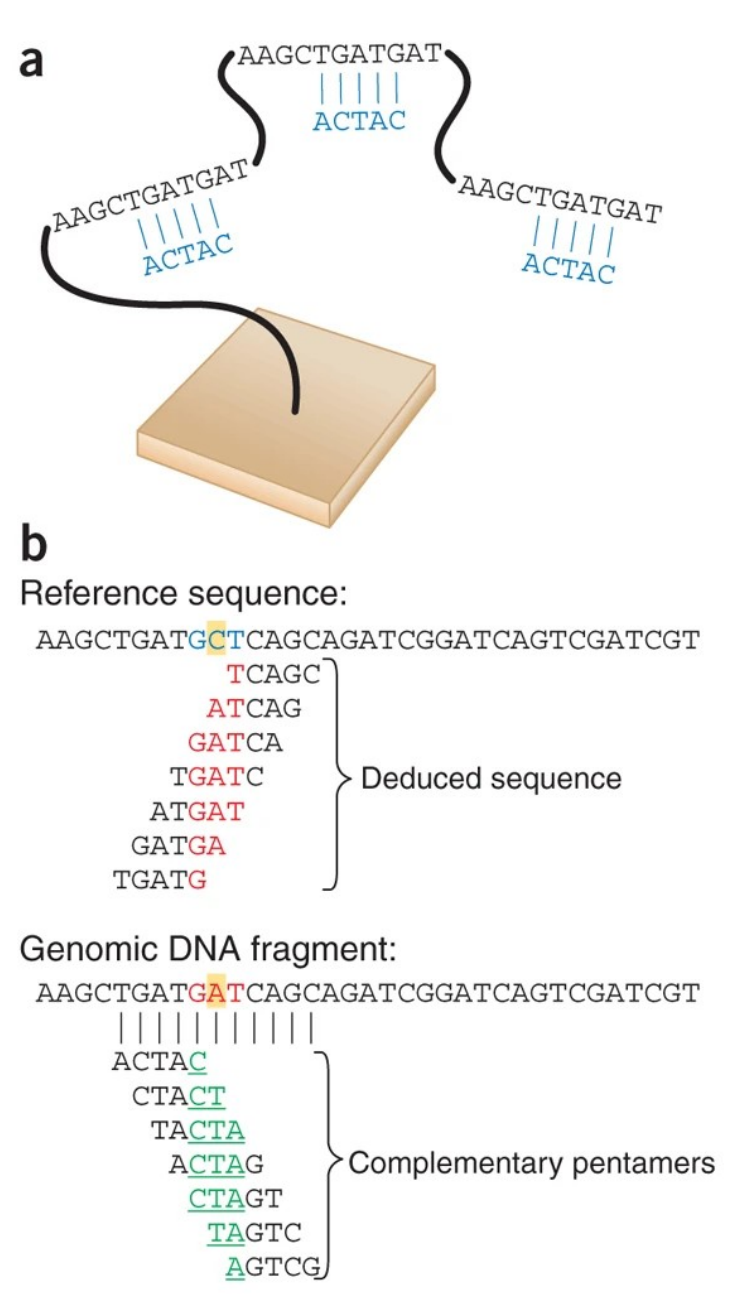
SBH relies on sequential hybridization of different pentamers; one hybridization round with the 3′-ACTAC-5′ pentamer is shown.

(b) Detection of a point mutation in the trinucleotide sequence GCT (blue in the reference sequence), which shows a change to GAT (red) in a genomic DNA fragment by shotgun-SBH.
The single base change is deduced from five independent complementarity matches (underlined, green) using antiparallel pentamer sequences

The deduced trinucleotide sequence shown in red.

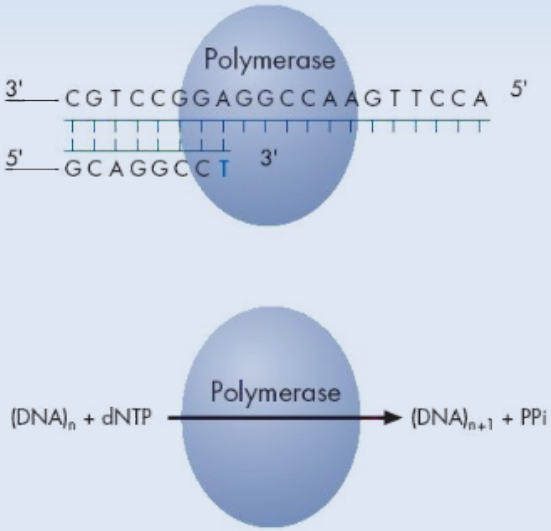
Correctly decoded ~96% of the 48.5-kbp Bacteriophage λ genome and ~80% of the 4.6-Mbp Escherichia coli genome in 2008

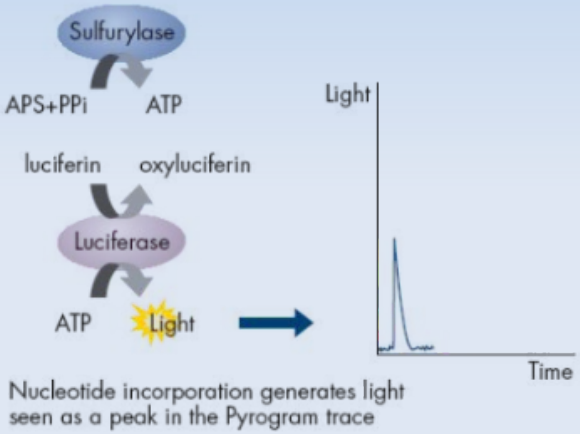
Worked well for resequencing for the discovery of **single nucleotide polymorphisms**, mutations in coding/noncoding exons

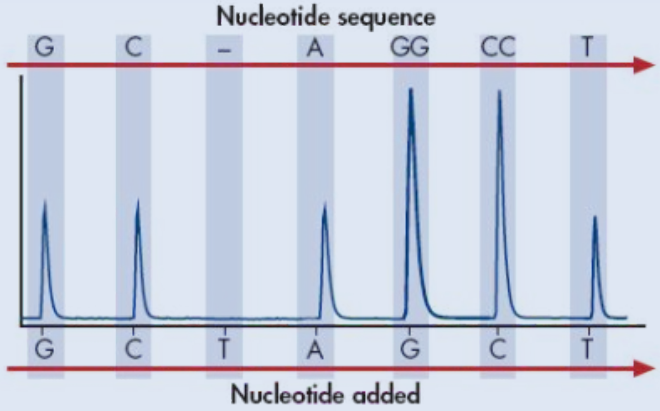


2nd GS: Pyrosequencing

Add nucleotide, Pyrophosphate (PPi) released when nuc. is incorporated into strand:







'pyrogram'

Detect PPi by converting it to ATP, Luciferase converts it to light signal

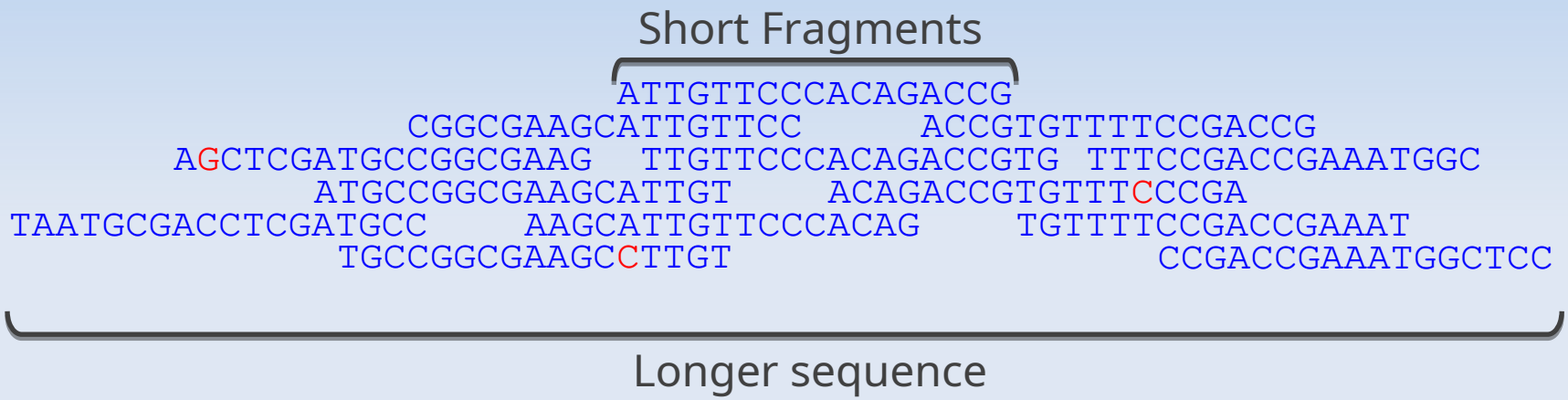
2nd Gen Sequencing

Current sequencers can reliably work with short fragments:

Method	Read Length
Sanger	600-1000 bp
454	400-800 bp
Illumina	75-300 bp
Ion Torrent	~200 bp

Assembly

Assemble fragments into a long sequence



Consensus:

TAATGCGACCTCGATGCCGGCGAAGCATTGTTCCCACAGACCGTGTTTTCCGACCGAAATGGCTCC

Coverage: # of reads underlying the consensus

Average coverage: Total bases / consensus length

13 reads * 17 bases per read = **221 bases**

221 bases / 66 base consensus = **3.35-fold coverage**

Assembly

We may luck out

6x coverage
100% identity

Coverage: # of reads underlying the consensus

Assembly

Or not

5x coverage
80% identity

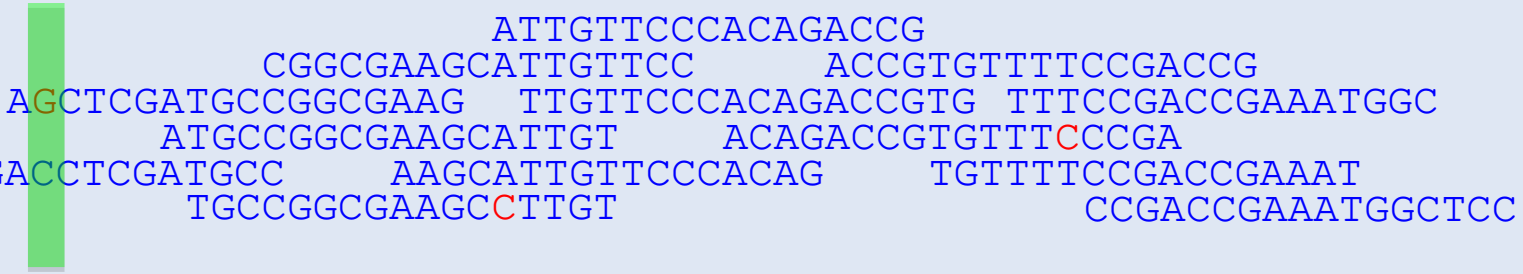
Coverage: # of reads underlying the consensus

Assembly

Or not

Consensus:

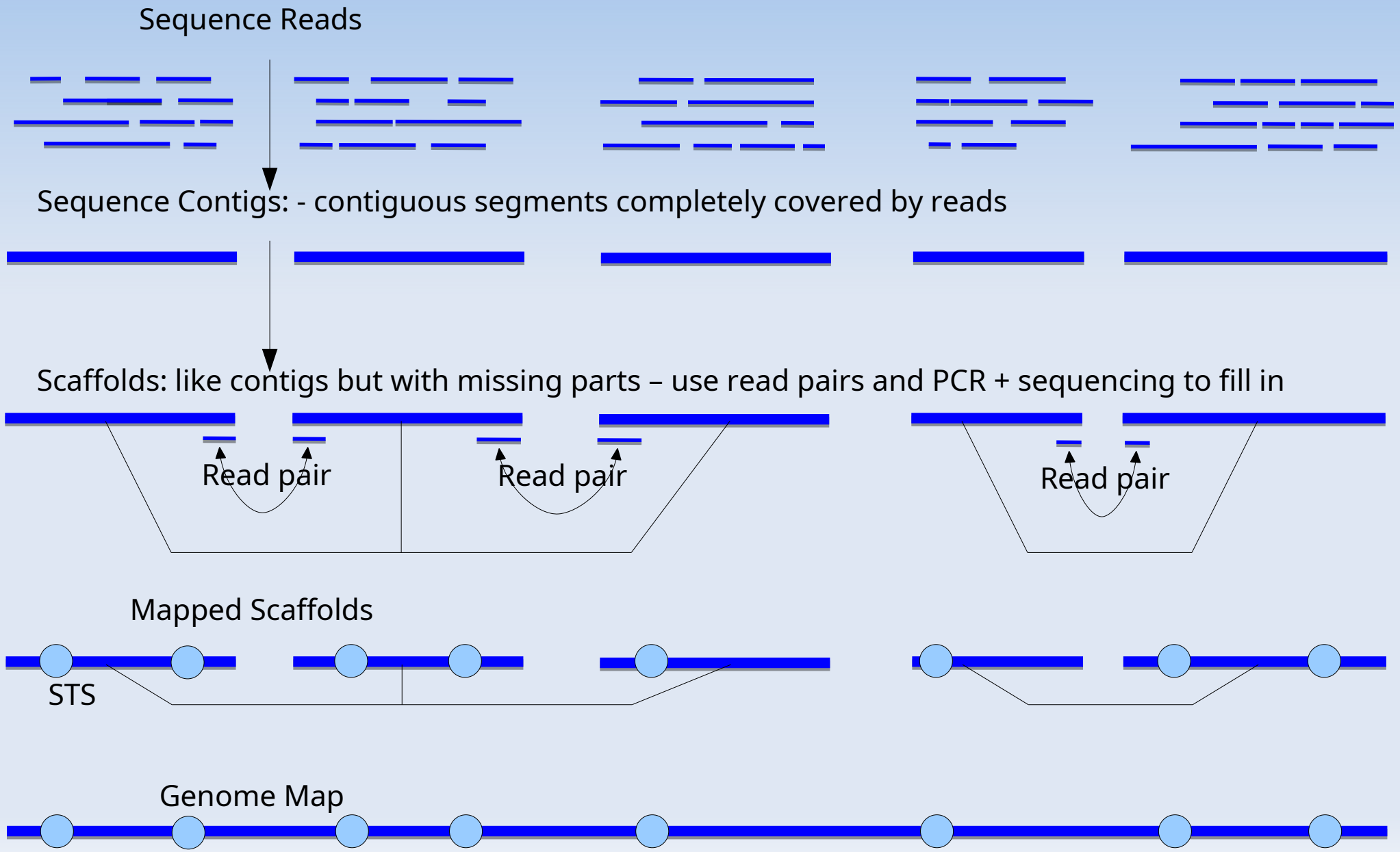
TAATGCGACCTCGATGCCGGCGAAGCATTGTTCCCACAGACCGTGTTTTCCGACCGAAATGGCTCC



2x coverage
50% identity

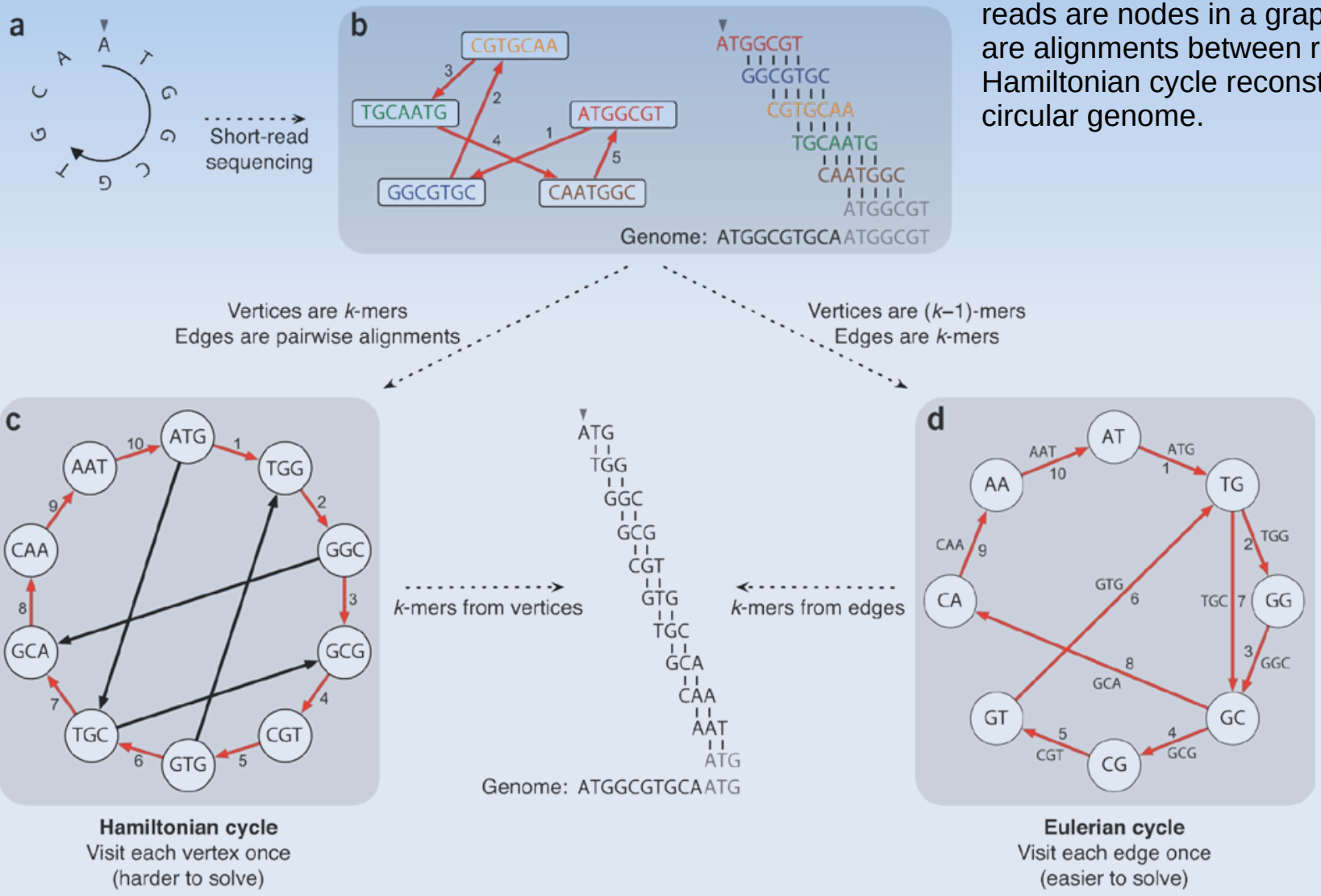
Coverage: # of reads underlying the consensus

Assembly


Sequence Reads
Sequence Contigs: - contiguous segments completely covered by reads
Scaffolds: like contigs but with missing parts – use read pairs and PCR + sequencing to fill in
Read pair
Read pair
Read pair
Mapped Scaffolds
STS
Genome Map

Graphs and Genome Assmbly

(a) An example small circular genome.



(b) Traditional sequencing algorithms: reads are nodes in a graph, edges are alignments between reads. A Hamiltonian cycle reconstructs the circular genome.

(c) An alternative technique splits reads into k-mers. A Hamiltonian cycle - red - reconstructs the genome by forming an alignment: each successive k-mer is shifted by one position.

(d) Modern short-read assembly algorithms construct a de Bruijn graph. An **Eulerian** cycle reconstructs the genome sequence.

3rd Gen Genome Sequencing

Single DNA molecule sequencing:

- Pacific Biosciences
- Oxford Nanopore
- Quantapore
- Stratos

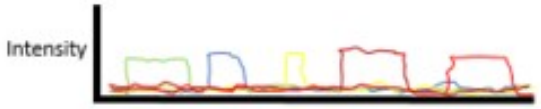
Main idea is to make longer reads – 10,000 to 100,000 bp reads → fewer reads needed for assembly → computationally simpler and faster

3rd Gen Genome Sequencing

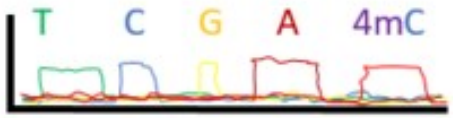
PacBio SMRT seq

DNA passes thru polymerase in an illuminated volume

Raw output is fluorescent signal of the nucleotide incorporation, specific to each nucleotide

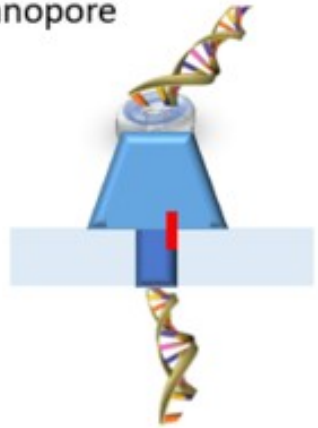


A,C,T,G have known pulse durations, which are used to infer methylated nucleotides

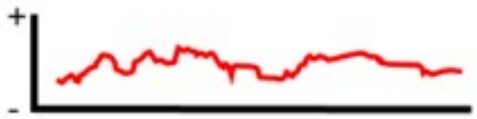


Oxford Nanopore

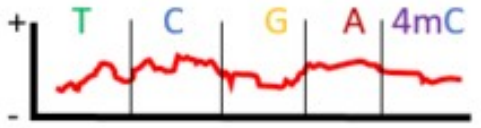
DNA passes thru nanopore



Raw output is electrical signal caused by nucleotide blocking ion flow in nanopore



Each nucleotide has a specific electric "signature"



Genome Sequencing

Number of base pairs in human genome:

~**3 billion** (haploid – single set of chromosomes)

Current cost of sequencing a human genome

$1000

Time needed to sequence a human genome in 2022:

1-3 days

Currently: 1000s of human genomes sequenced so far

Genome Projects

Currently there are ~500,000 organisms with genomes sequenced or being sequenced.

Mostly – ~400,000 – bacteria

~20,000 viruses

~50,000 Eukarya

~200 animals

https://gold.jgi.doe.gov/index

Comparative Genomics

By comparing genomes of related animals, we may be able to see what sequences are conserved

5-10% of our genome is conserved cf. other animals

Remember: <1% of our genome is ”coding DNA”

(exons that code for protein)

→ a LOT of non-coding functional sequence information for chromosome packaging, segregation, and replication; non-coding RNA, and gene regulation

Viewing Genomes

Sites that allow us to view various genomes:

- NCBI: Genome Data Viewer
- UC Santa Cruz Genome Browser
- JBrowse http://jbrowse.org/
- Ensemble's Genomes

The free Integrative Genomics Viewer (IGV) program allows us to access the same data but runs as a standalone program.

Other web sites:

The gnomAD web site allows us to look at variations in 120,000 genomes.

The Galaxy web site has one of the largest collection of algorithms for analyzing genomes

Data in Genome Browsers

Besides the actual genomic sequences themselves, these genome browsers usually also have

- RefSeq genes (mRNA and protein)
- Ensembl genes
- Expressed sequence tags
- Gene predictions
- SNPs from a database like dbSNP
- Non-coding functional elements as found by projects like ENCoDE

ENCoDE Project

ENCyclopedia **o**f **D**NA **E**lements: Goal is to identify all functional elements in human, mouse genomes

Encode portal:

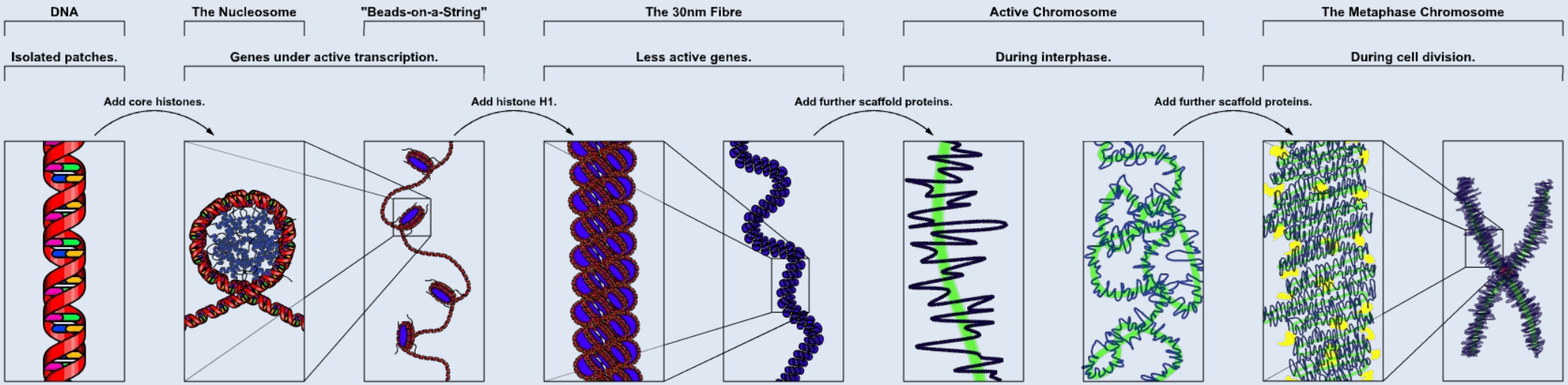
- integrative-level annotations, including a registry of candidate cis-regulatory elements
- ground-level annotations derived directly from experimental data

ENCoDE Project Phase II

ENCyclopedia **o**f **D**NA **E**lements: Identify all functional elements in human genome

Phase 4 underway now:

- 80% of human genome participates in biochemical RNA or chromatin (30 nm) process



- 95% of genome lies within 8 kb of a DNA-protein interaction.

Single Nucleotide Polymorphism
SNP

An SNP is a confirmed substitution, insertion, or deletion in single nucleotide in a gene (compared to a reference genome)

There are 3-5 million human SN variants

A single individual may have ~60 variants not seen in either parent – mostly biologically silent.

SNPs and Haplotypes

Haplotype – a group of genes that an individual inherits as a unit from a parent; can also refer to a set of SNPs that are inherited together.

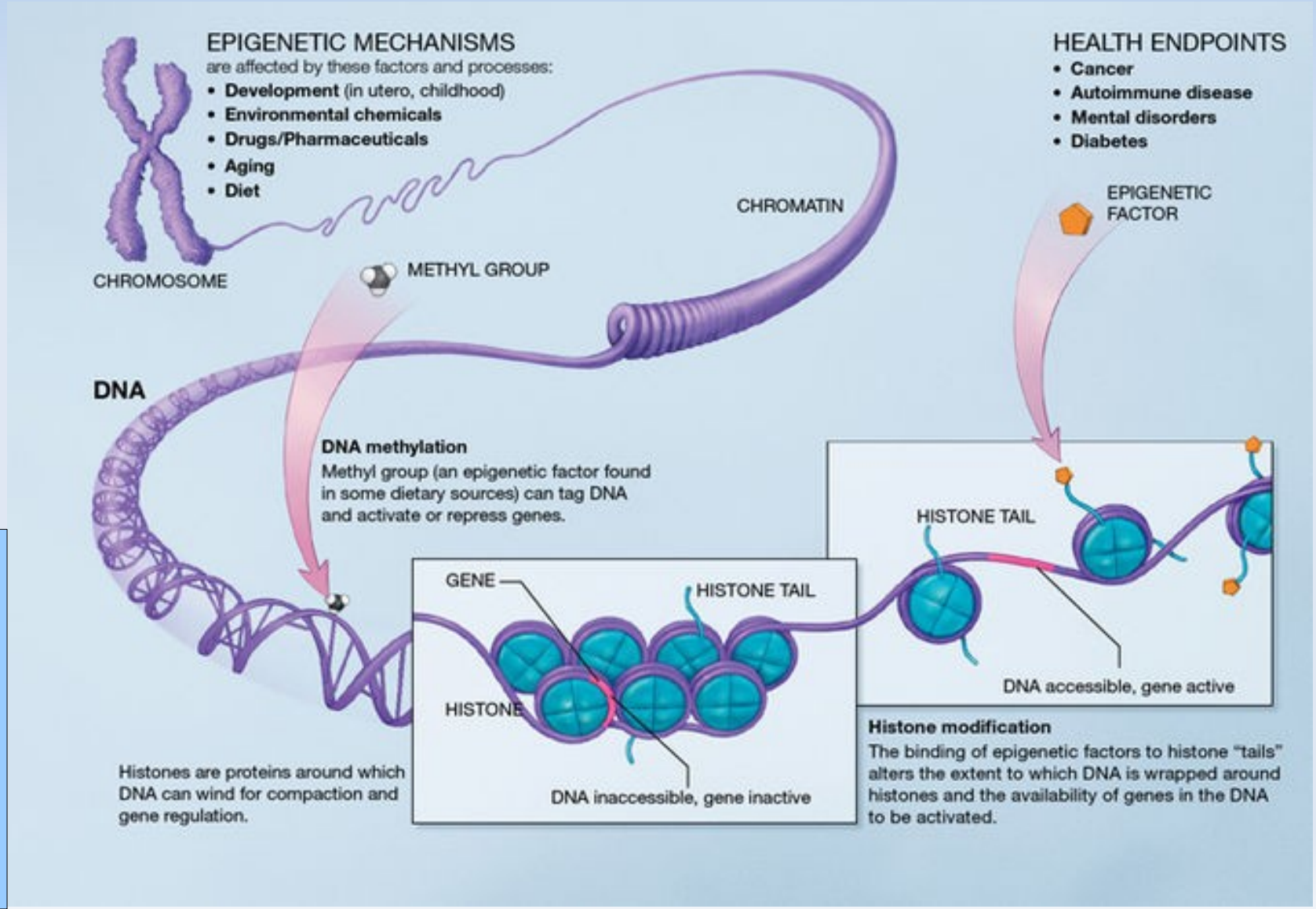
A **Haplogroup** is a set of SNPs that represent a clade or group of people with a common ancestor.

→ The HapMap project https://hapmap.ncbi.nlm.nih.gov/

→ 1000 Genomes project – has 2535 human genomes from 26 groups world wide http://www.1000genomes.org/

Epigenetics

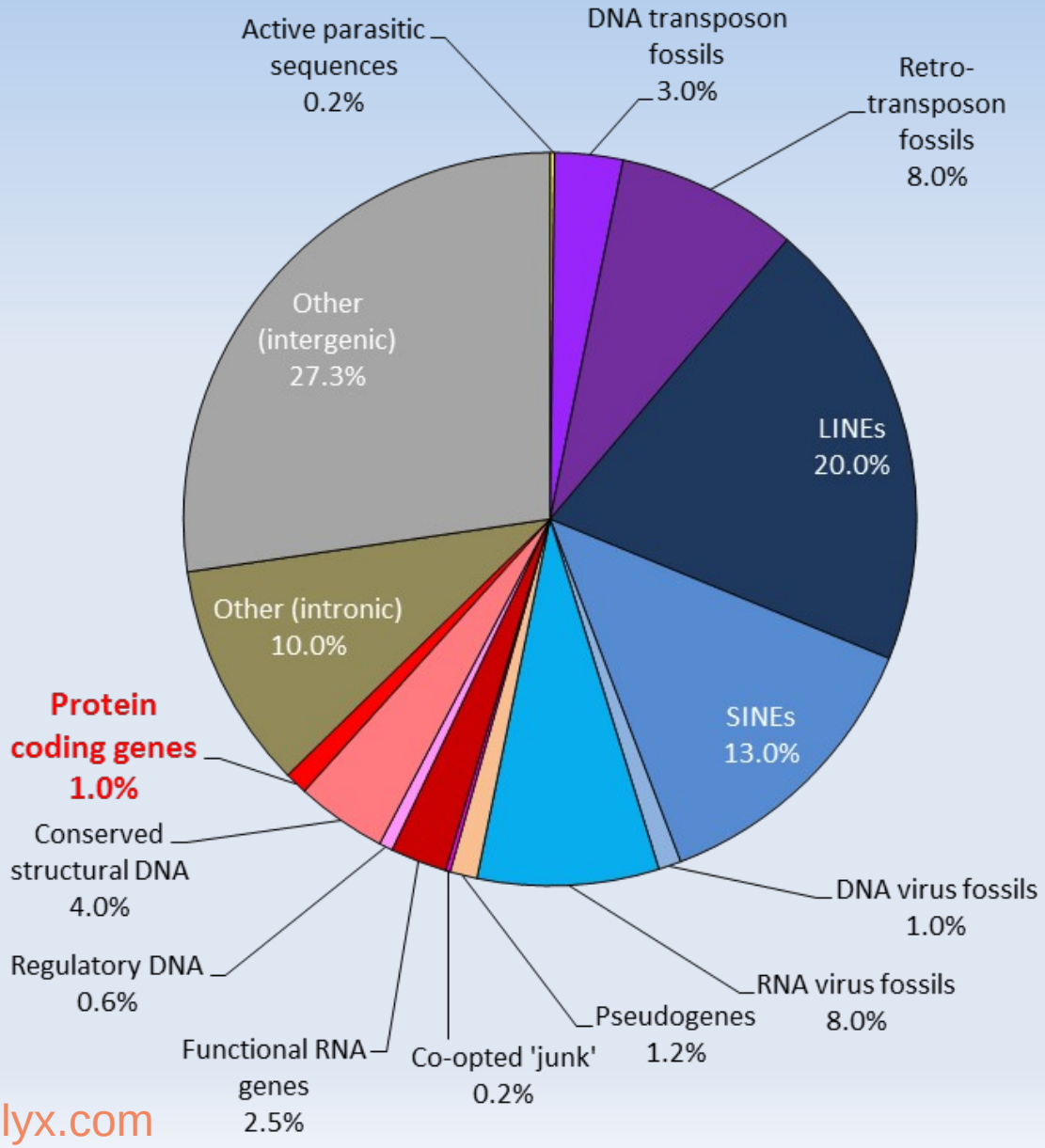
Epigenetics: study of variations in inherited traits caused by environmental factors → 2nd Genomic code



Can get genome wide epigenetic information. e.g. Bisulfite treatment of DNA: converts cytosine to Uracil; methylcytosine is not affected

What's in our genome?

A lot of what we thought was "junk" but only now starting to understand!

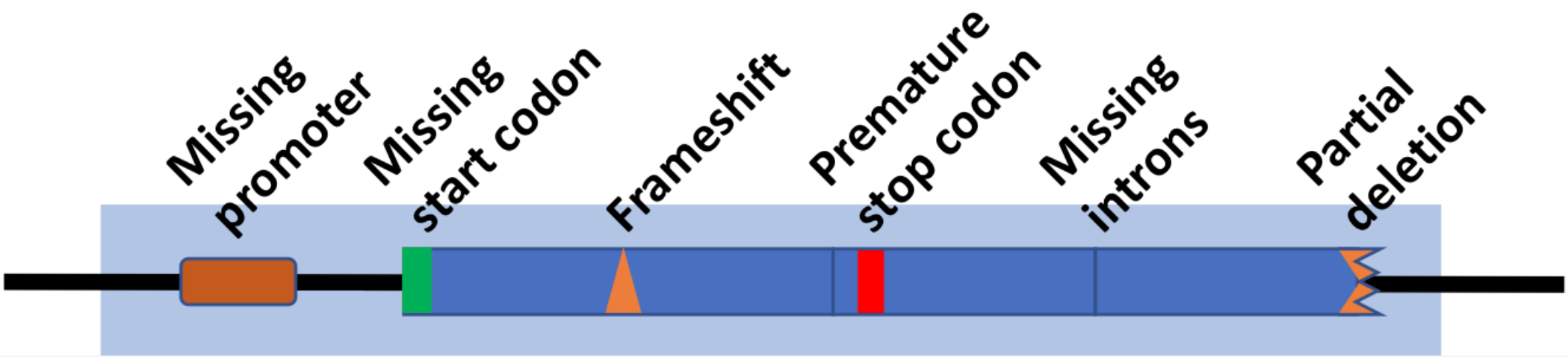


Pseudogenes

About the same fraction of space in our genome that protein coding genes take up is also taken up by pseudogenes

→sequences similar to either some of our own other genes (by gene duplication) or to genes found in other species but have become inactivated by mutations in our genome.

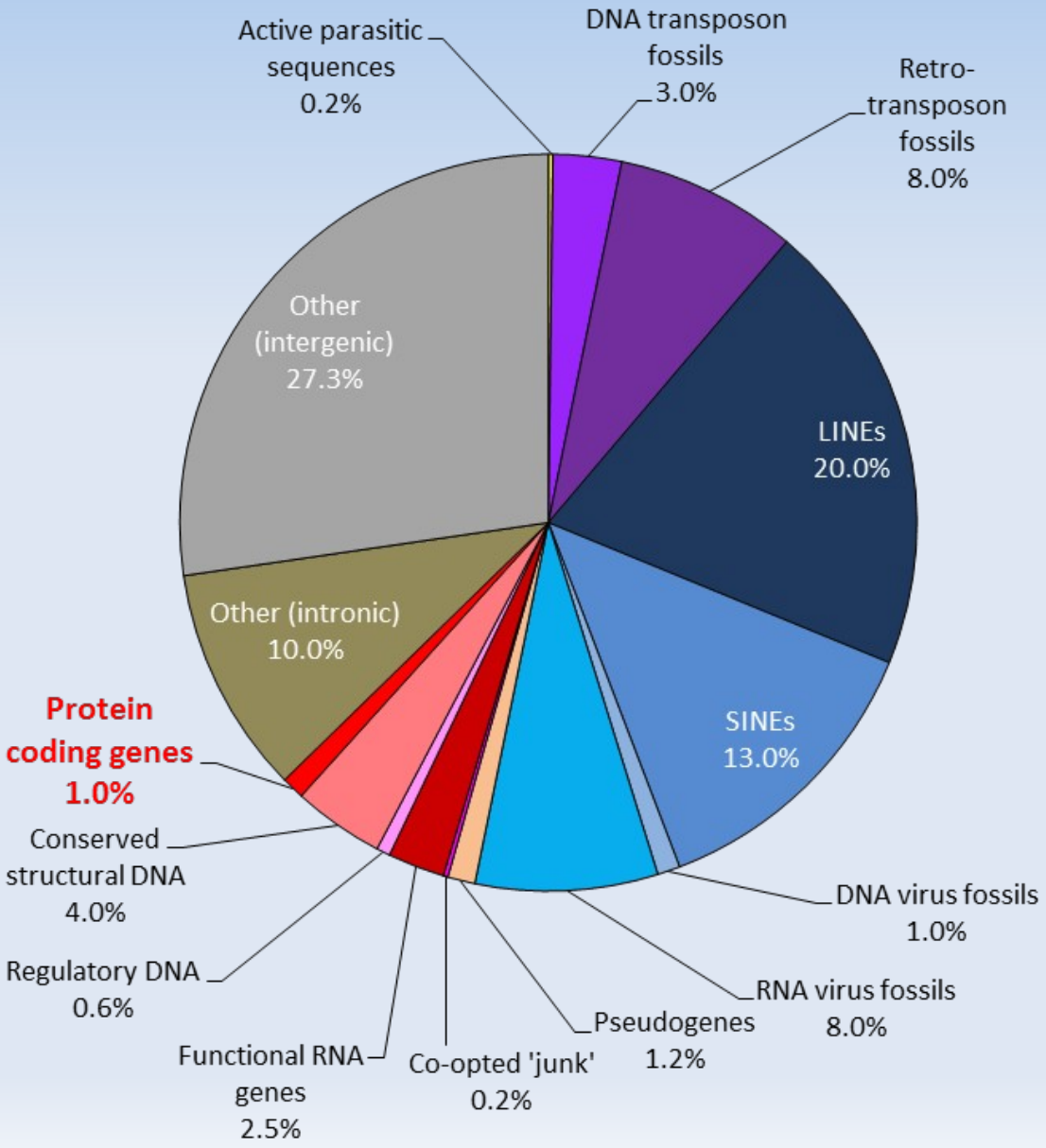
Common defects of pseudogenes:



Introns and intergenic sequences

Introns – the space between exons – may make up 20 times what exons take up

Intergenic sequences – the space between genes with unknown function – may make up another 20 times as much as exons



Transposable Elements

About half of the human genome is taken up by transposable elements or "jumping genes"

Transposable elements are genetic sequences that can change position within our genome. They include

Retrotransposons ~40%

DNA transposons ~3%

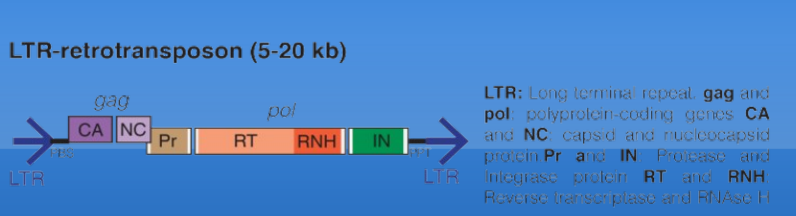
Retrotransposons

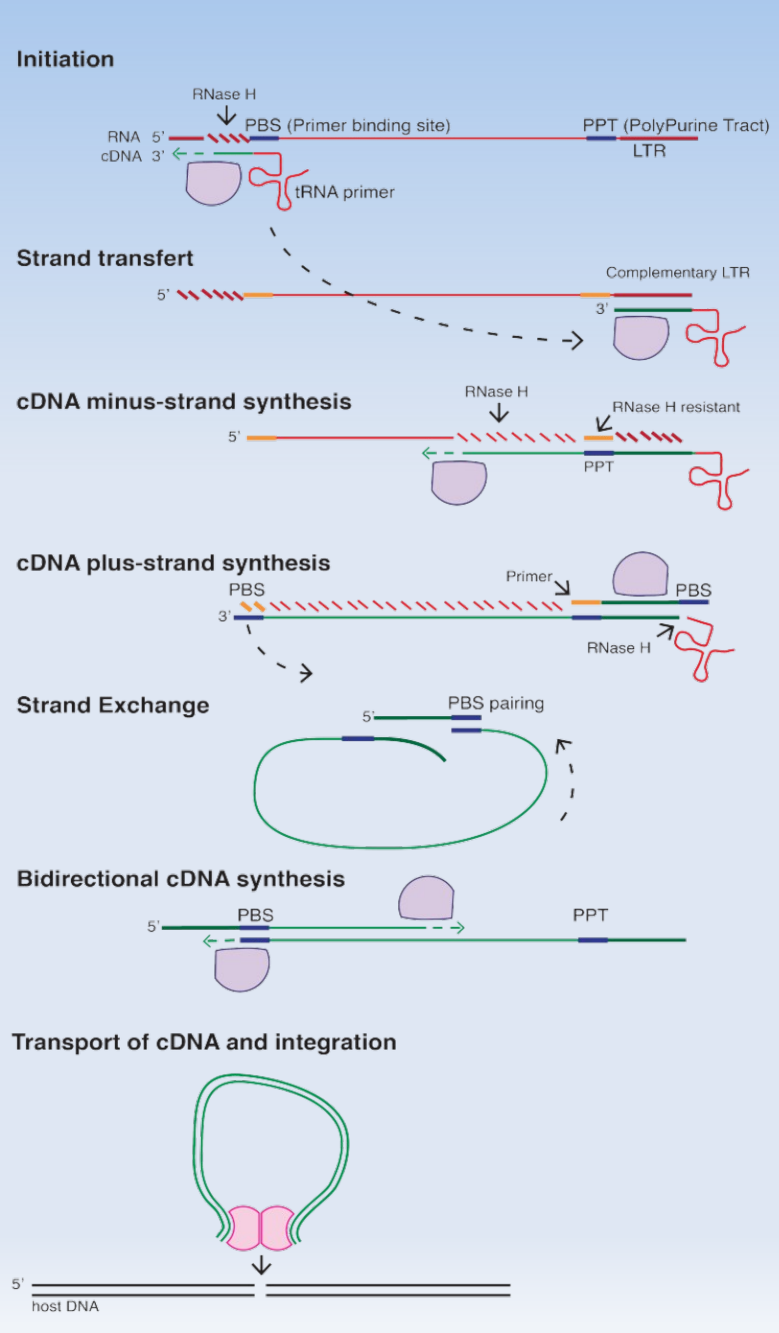
Retrotransposon: genomic DNA that is transcribed to RNA and then reverse-transcripted back to DNA to be copied and pasted into different genomic locations

Long Terminal Repeat (LTR) retrotransposons are ~8% of genome

Non-LTR ~33%
- LINEs
- SINEs

Mutations have made most retrotransposons ”dead” or dying, no longer capable of ”moving”





Long interspersed Nuclear Element (LINE) retrotransposons

LINE retrotransposons make up about 21% of the human genome

We each have ~800,000 LINEs mostly non-functional

We may have ~100 functional retrotransposons, each ~7000 bp segments

LINE gene-complexes contain genes for reverse transcriptase, RNA-binding protein & nucleases but activity is regulated by host cell epigenetics and siRNA