Genome assembly & 2nd generation techniques

BIOL 425 Spring 2023

3/13/2023

Plan

- This week in antibiotics resistance
- Your projects
- Sequencing genomes
- 2nd generation sequencing

The New York Times

This Stomach Bug Isn't Responding to Antibiotics. Scientists Are Worried.

Five percent of Shigella infections are now resistant to standard antibiotics. This could be a predictor of what's to come with other bacteria.

Give this article 306

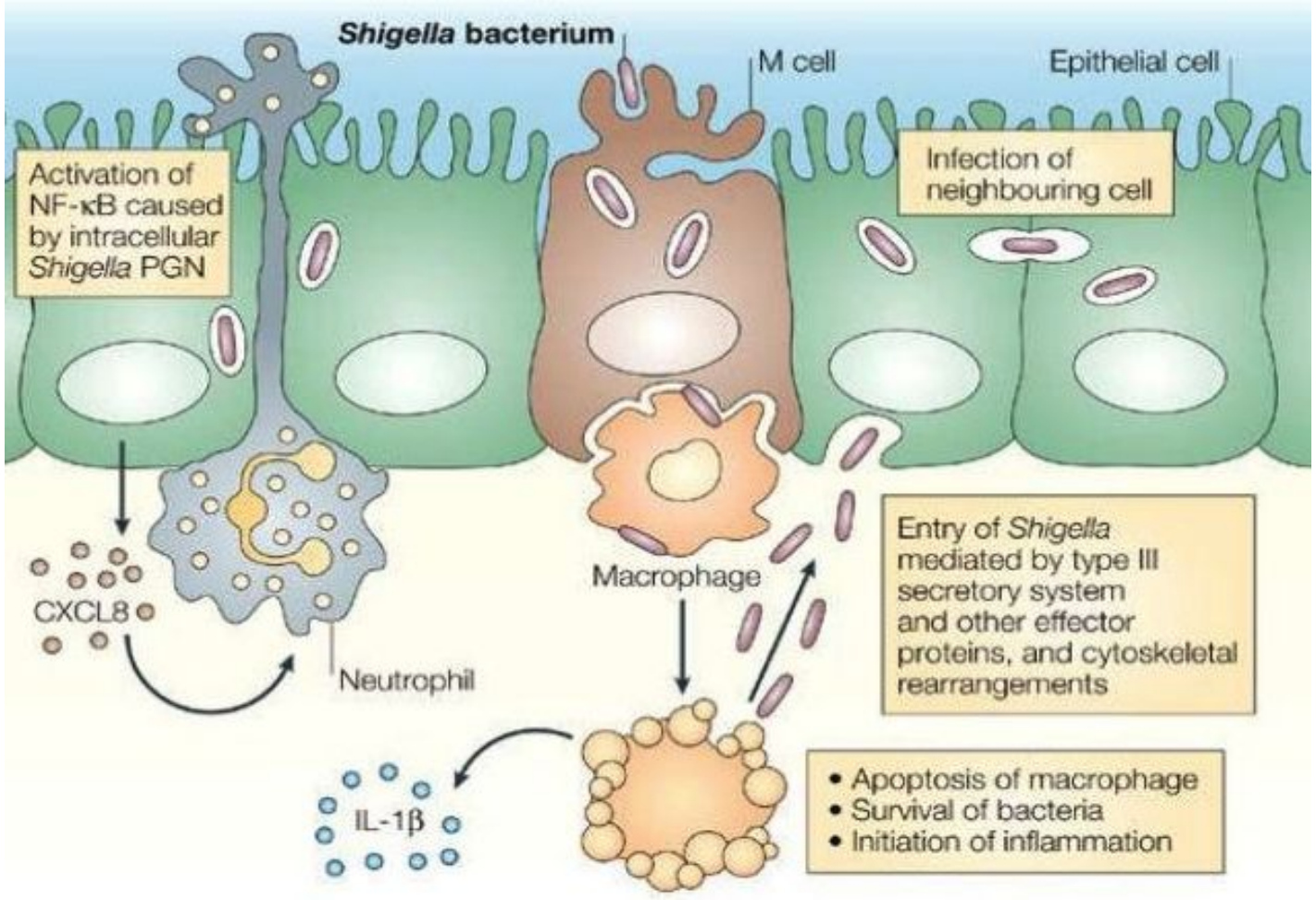
AP Photo/CDC

- 450,000 cases / yr in USA
- 6400 hospitalized

- 1st resistant case = 2016
- Now = 5% of all cases

- Causes bloody diarrhea
- Resolves in a few days

- Typically treated with azithromycin, cipro, ceftriaxone, trimethoprim-sulfamthoxazole and ampicillin
- Resistant to all 5 antibiotics



- 450,000 cases / yr in USA
- 6400 hospitalized

- 1st resistant case = 2016
- Now = 5% of all cases

- Causes bloody diarrhea
- Resolves in a few days

- Typically treated with azithromycin, cipro, ceftriaxone, trimethoprim-sulfamthoxazole and ampicillin
- Resistant to all 5 antibiotics

Increase in Extensively Drug-Resistant Shigellosis in the United States

Print

HAN HEALTH ALERT NETWORK

This is an official CDC HEALTH ADVISORY

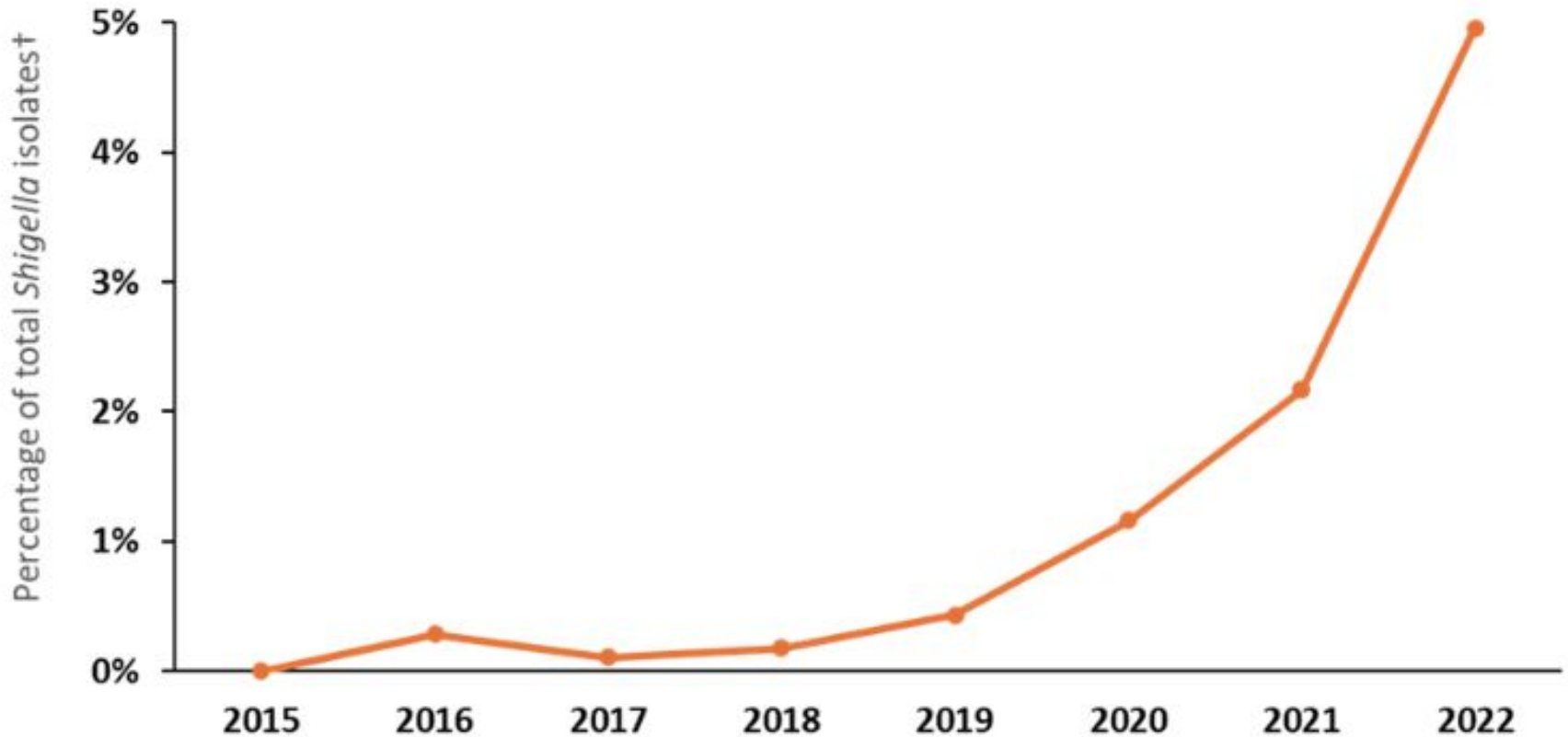
Distributed via the CDC Health Alert Network
February 24, 2023, 11:30 AM ET
CDCHAN-00486

Summary

The Centers for Disease Control and Prevention (CDC) has been monitoring an increase in extensively drug-resistant (XDR) *Shigella* infections (shigellosis) reported through national surveillance systems [1]. In 2022, about 5% of *Shigella* infections reported to CDC were caused by XDR strains, compared with 0% in 2015. Clinicians treating patients infected with XDR strains have limited antimicrobial treatment options. *Shigella* bacteria are easily transmissible. XDR *Shigella* strains can spread antimicrobial resistance genes to other enteric bacteria. Given these potentially serious public health concerns, CDC asks healthcare professionals to be vigilant about suspecting and reporting cases of XDR *Shigella* infection to their local or state health department and educating patients and communities at increased risk about prevention and transmission.

Shigellosis is an acute enteric infection that is an important cause of domestically acquired and travel-associated bacterial diarrhea in the United States. Shigellosis usually causes inflammatory diarrhea that can be bloody and may also lead to fever, abdominal cramping, and tenesmus. Infections are generally self-limiting; however, antimicrobial treatment may be indicated to prevent complications or shorten the duration of illness [2]. CDC defines XDR *Shigella* bacteria as strains that are resistant to all commonly recommended empiric and alternative antibiotics — azithromycin, ciprofloxacin, ceftriaxone, trimethoprim-sulfamethoxazole (TMP-SMX), and ampicillin. Currently, there are no data from clinical studies of treatment of XDR *Shigella* to inform recommendations for the optimal antimicrobial treatment of these infections. As such, CDC does not have recommendations for optimal antimicrobial treatment of XDR *Shigella* infections.

Figure: Percentage of *Shigella* isolates that showed an extensively drug resistant (XDR)* phenotype or genotype in the United States, by year, 2015–2022†



*XDR *Shigella* bacteria (n=239) are defined as resistant to azithromycin, ciprofloxacin, ceftriaxone, trimethoprim-sulfamethoxazole, and ampicillin.

Background

Shigella bacteria are transmitted by the fecal-oral route, directly through person-to-person contact including sexual contact, and indirectly through contaminated food, water, and other routes. *Shigella* bacteria are easily transmitted because of the low infectious dose (as few as 10–100 organisms), and outbreaks tend to occur among people in close-contact settings [2–6].

Historically, shigellosis has predominantly affected young children (age 1–4 years) in the United States. More recently, CDC has observed an increase in antimicrobial-resistant *Shigella* infections among adult populations [4,5] especially

- Gay, bisexual, and other men who have sex with men (MSM)
- People experiencing homelessness
- International travelers
- People living with HIV

Plan

- This week in antibiotics resistance
- Your projects
- Sequencing genomes
  - Understand how are genomes assembled from sequences.
- 2nd generation sequencing
  - Understand how 2nd generation sequencing technology works.
  - Contrast with Sanger sequencing.

Week 8 **March 6**	**<u>No Lecture:</u>** Independent project proposals	Independent project proposals during lab time	**Independent project** **Week 1: proposals**
Week 9 **March 13**	Sequencing and bioinformatics II	IP: Sample preparation & enumeration of bacteria	**Independent project** **Week 2: sample prep**
Week 10	Metagenomics	Metagenomics lab	**Metagenomics lab**

March 20	introduction lecture (Paul Scheid, milrd.org)		
Week 11 **March 27**	Lecture TBD, depending on independent projects	IP: AMR gene & 16S PCR,	**Independent project** **Week 3: AMR gene PCR**
Week 12 **April 3**	**No Lecture**	**No Lab Sections**	**Spring Break**
Week 13 **April 10**	No Lecture on Monday Lecture on Wednesday 4/12	Open Lab time	Open Lab time
Week 14 **April 17**	Presentation info Lecture TBD, depending on independent projects	IP: Gel of PCR from IP Week 3 and qPCR setup	**Independent project** **Week 4: qPCR & Gels**
Week 15 **April 24**	Lecture TBD, depending on independent projects	Final data analysis and presentation prep	**Independent project** **Week 5: data analysis and wrap up**
Week 16 **April 1**	Final Exam review, semester wrap-up	Presentations	

Plan

- This week in antibiotics resistance
- Your projects
- Sequencing genomes
  - Understand how are genomes assembled from sequences.
- 2^{nd} generation sequencing
  - Understand how 2nd generation sequencing technology works.
  - Contrast with Sanger sequencing.

Sanger Sequencing – developed in 1977

Fred Sanger (1918-2013)

2x Nobel Prize winner in chemistry.

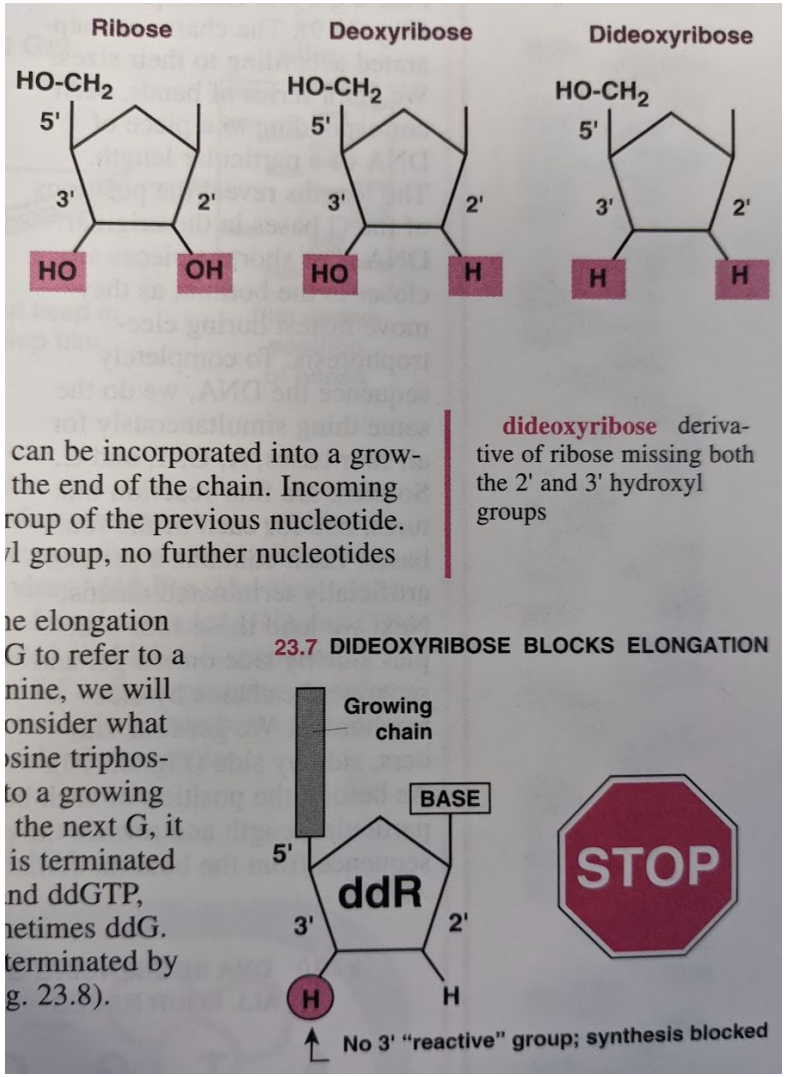
- 1958 - Amino acid sequence of insulin
  - Proved that proteins have defined amino acid sequences (1951)
  - Pre-DNA structure (1953)
- 1980 - DNA sequencing
  - Sequenced the first genome phage ΦX174 (5386 bp)

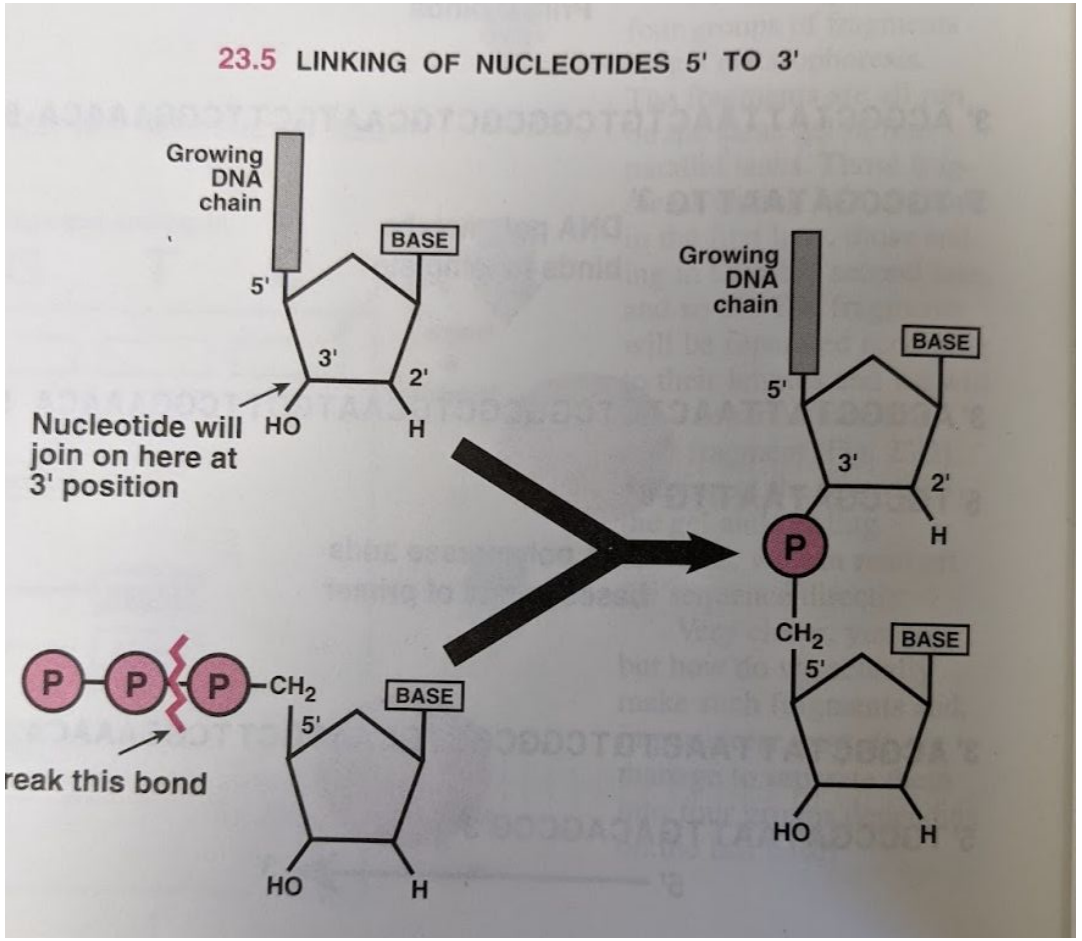
Review

- What reagent does modern Sanger sequencing rely on to label and generate sequences?

- A) Gel electrophoresis
- B) Fluorescently labeled primers
- C) Fluorescently labeled dNTPs
- D) Fluorescently labeled ddNTPs

ddNTPs





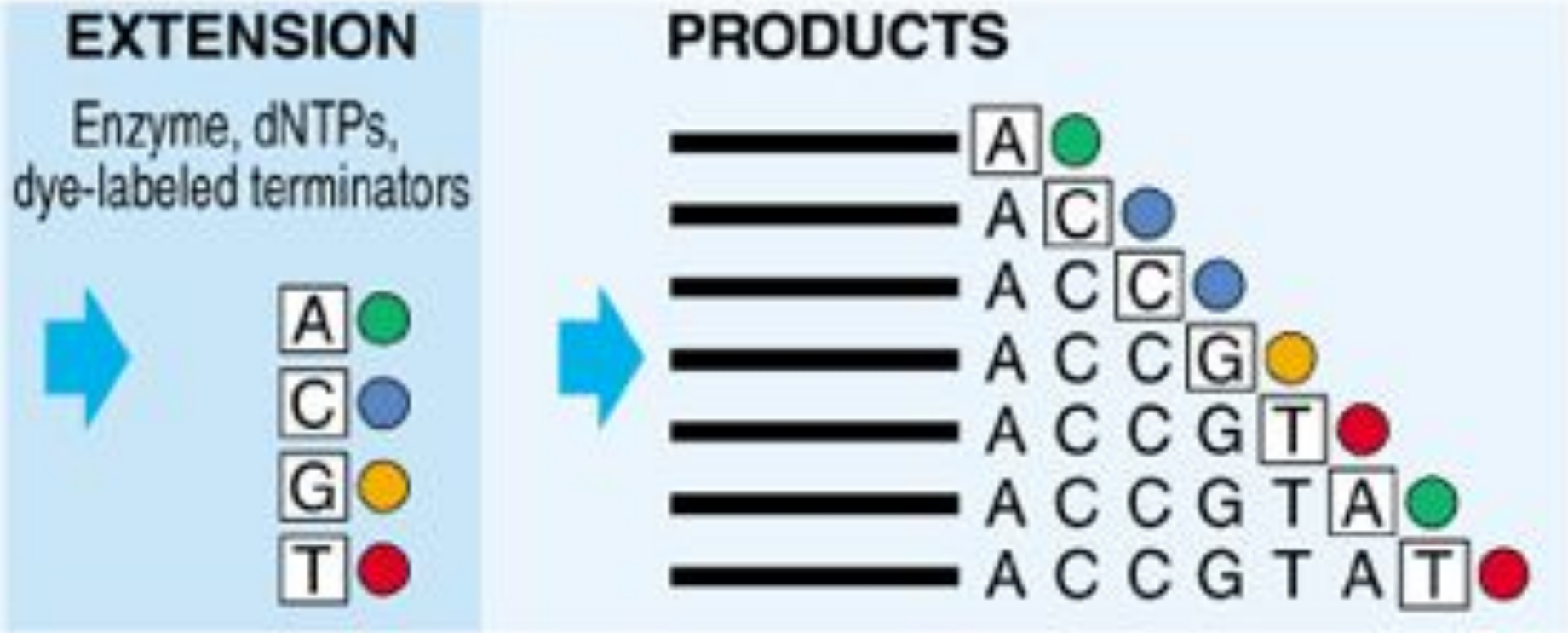
ddNTP = Chain Termination

- Prevents incorporation of the next base.

What if you set up a PCR with ddGTP <u>instead</u> of dGTP?

Dye Terminator Sequencing

- four fluorescently labeled ddNTP chain terminators
- The newly synthesized DNA fragments are different from each other in one nucleotide increments and labeled according to the incorporated ddNTP



Review

- About how long are sequencing reads from Sanger sequencing?

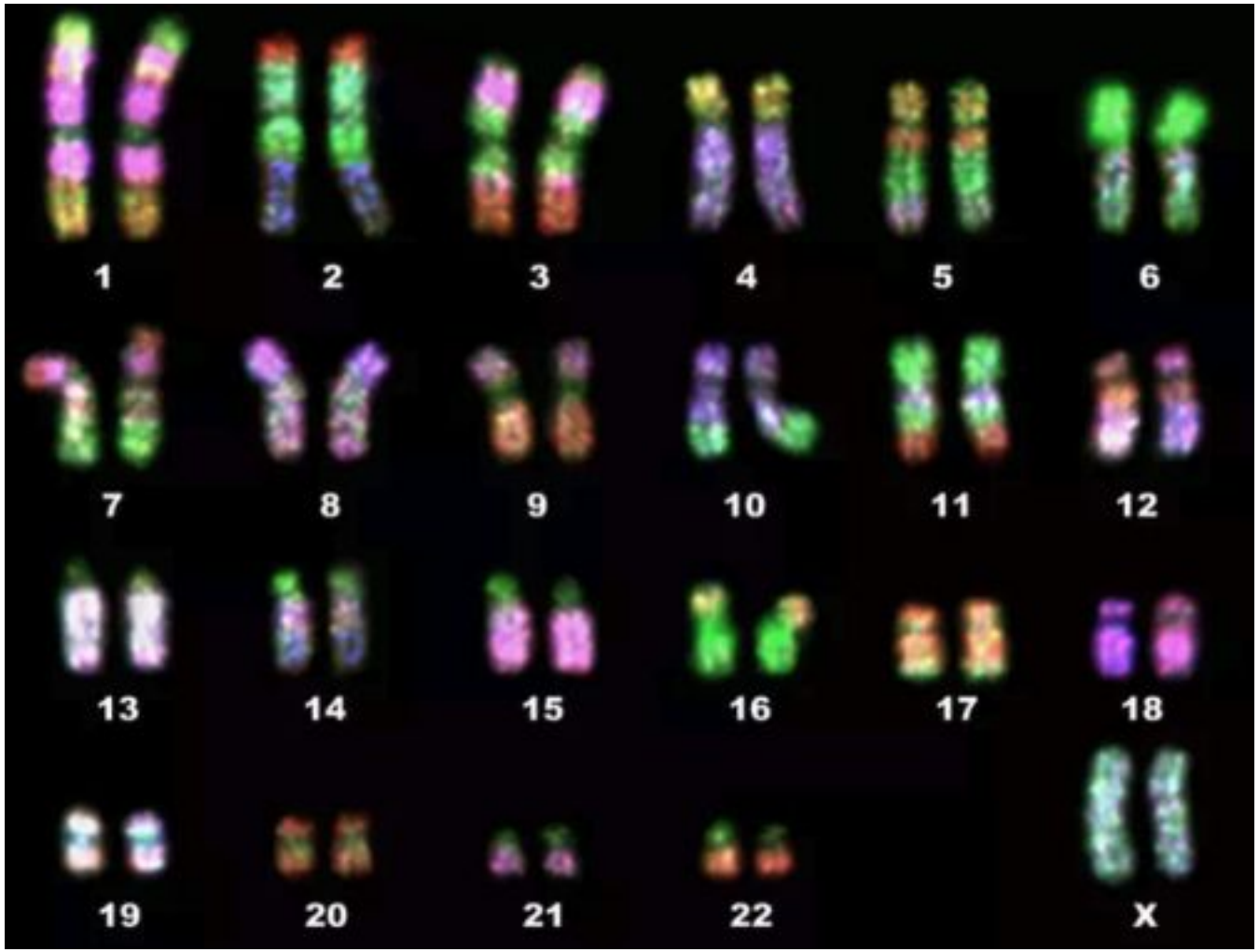
- A) 100 bp
- B) 1000 bp
- C) 10,000 bp
- D) 100,000 bp

Review

- About how long is the human genome?

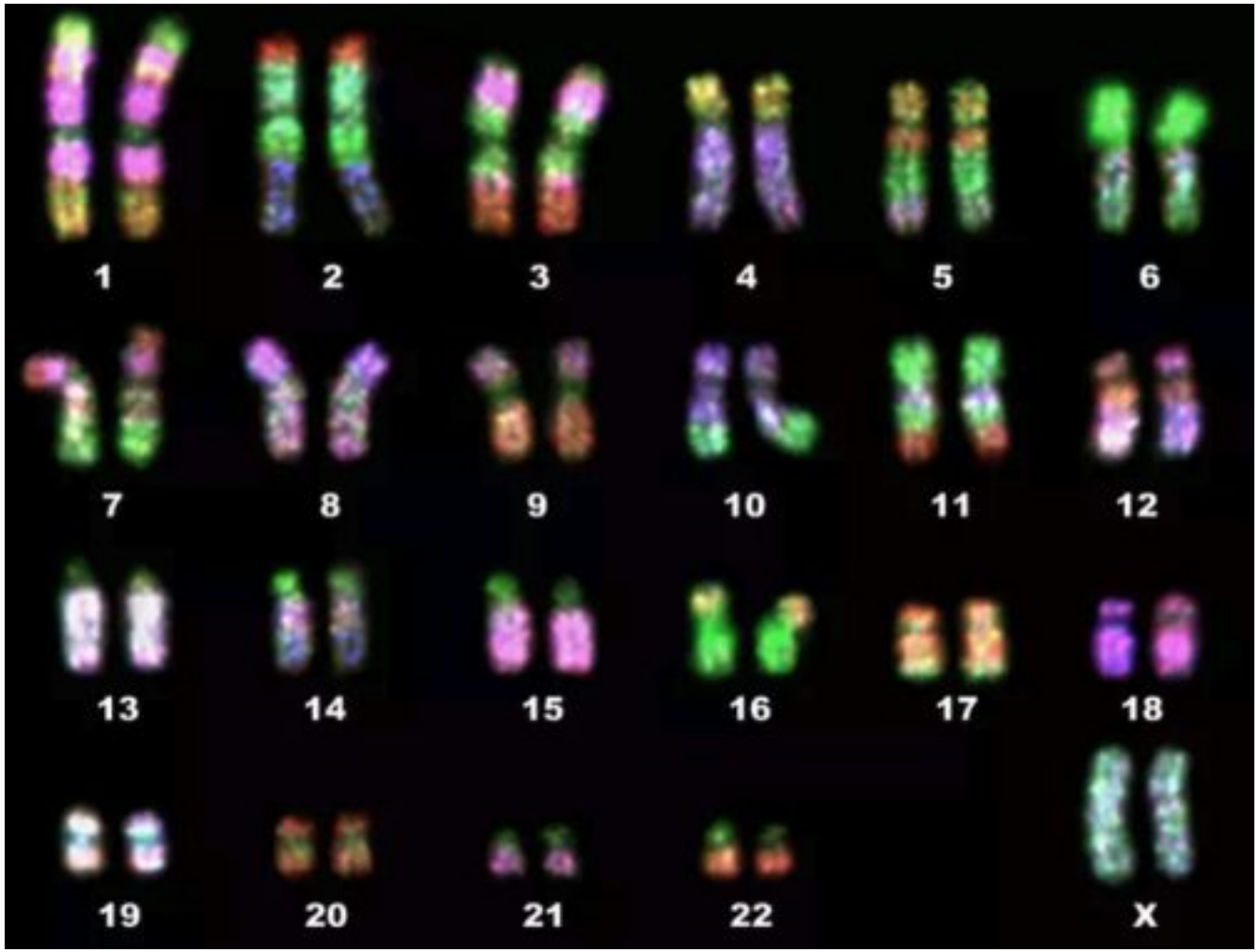
- A) 3,000,000
- B) 30,000,000 bp
- C) 300, 000, 000 bp
- D) 3, 000, 000, 000 bp

Sanger sequencing requires a primer, so how do we sequence an unknown genome?



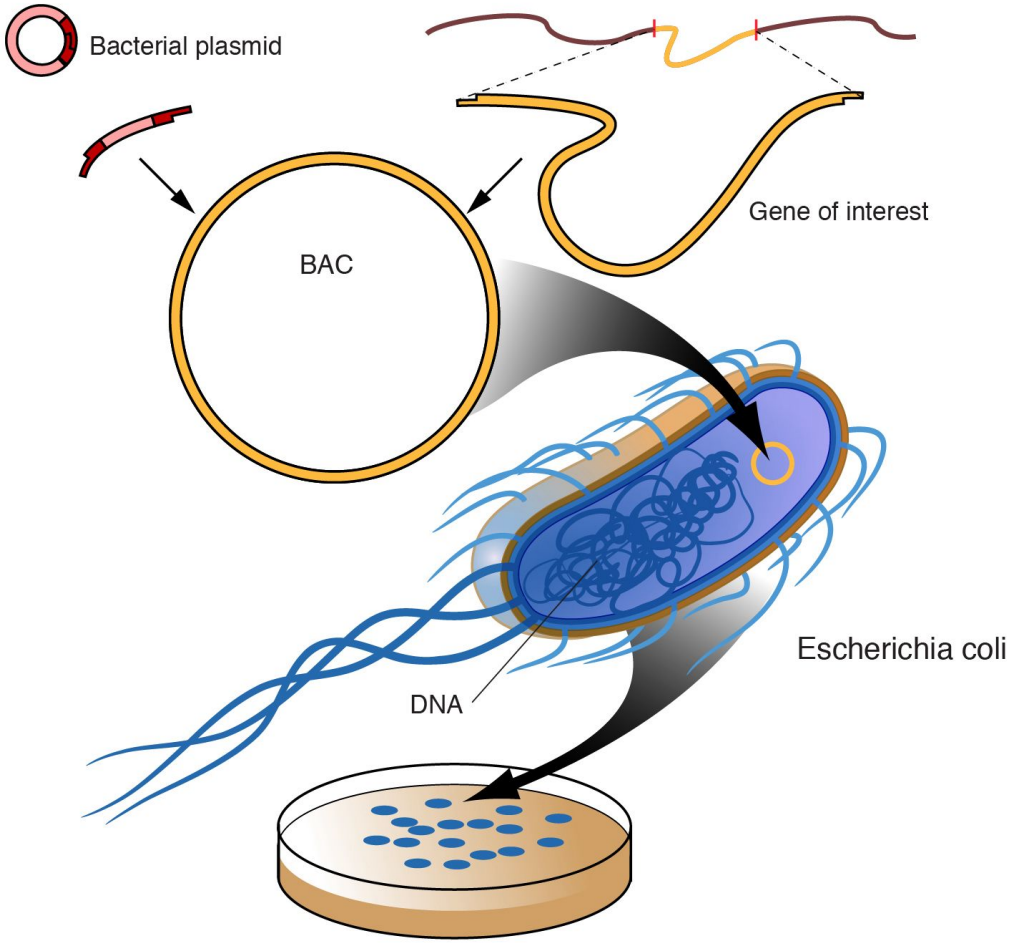
- 3,000,000,000 / 500 =
  - 6, 000, 000 Sanger sequence reads.

Sanger sequencing requires a primer, so how do we sequence an unknown genome?



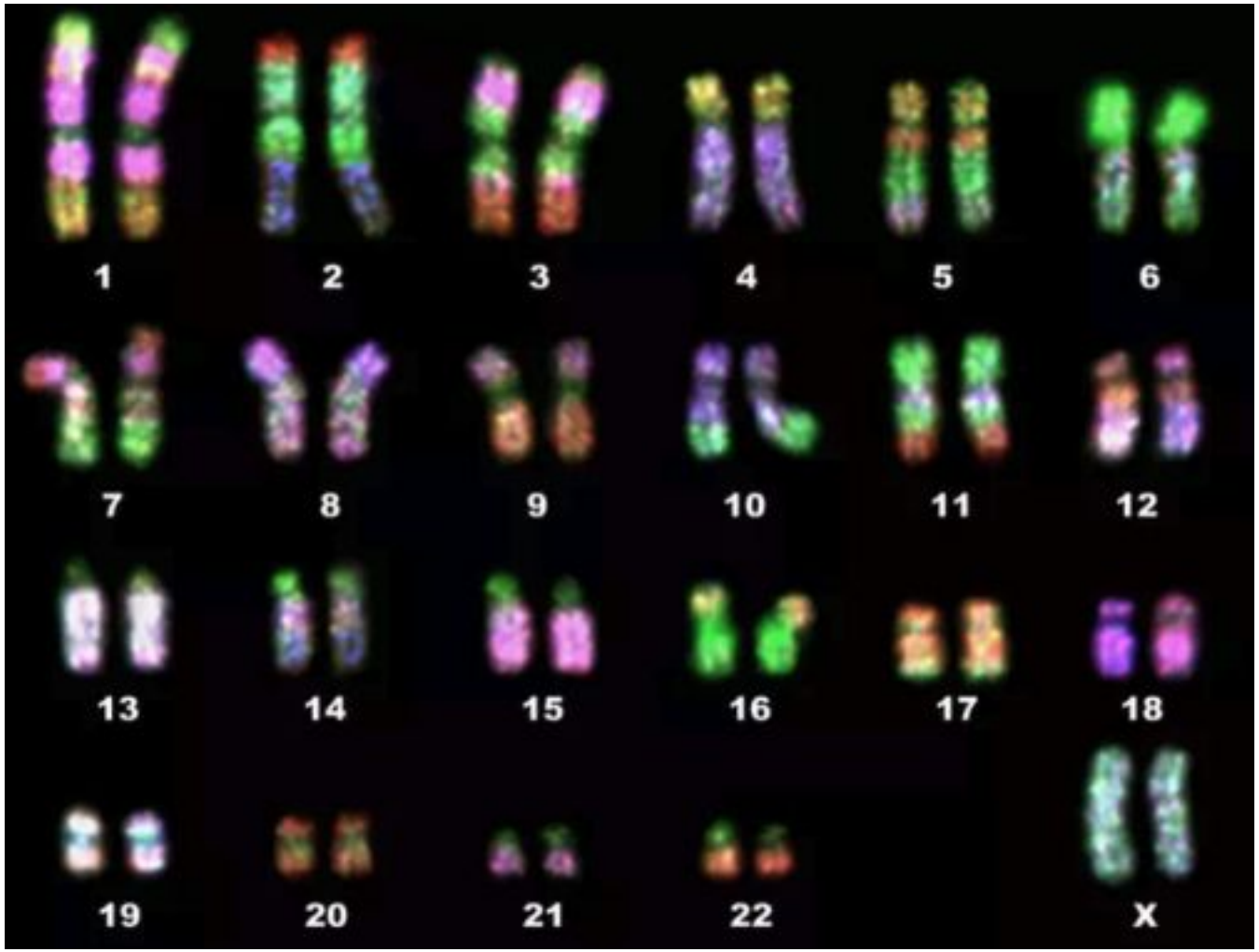
- 1) Break it all up into small pieces.
  - At the time Sanger was the only option so this needed to be done ~ 500bp at a time.
  - BUT, we need a primer.

Cloning DNA



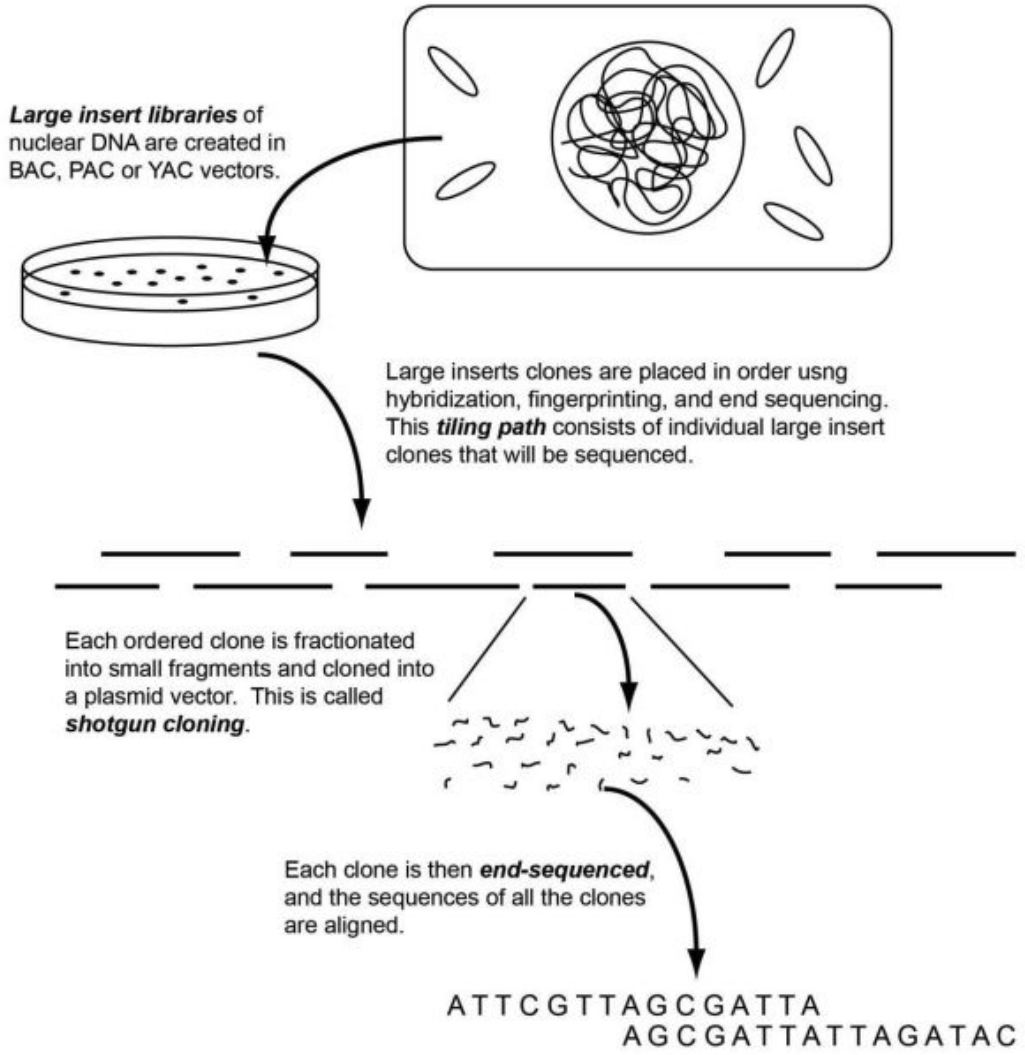
- Developed in the 1970s
- Uses plasmids that self-replicate in bacterial hosts
  - Usually *E. coli*
- Restriction enzymes:
  - Bacterial defense mechanisms that chop up DNA from viruses
  - Cut only at specific short sequences (4-8bp).
  - Allowed a cut-and-paste mechanism to insert DNA into plasmids.
- BAC = Bacterial artificial chromosomes
- The BAC plasmids <u>all had the same sequence</u>
  - Allows the use of the same primer for all unknown sequences.

Sanger sequencing requires a primer, so how do we sequence an unknown genome?



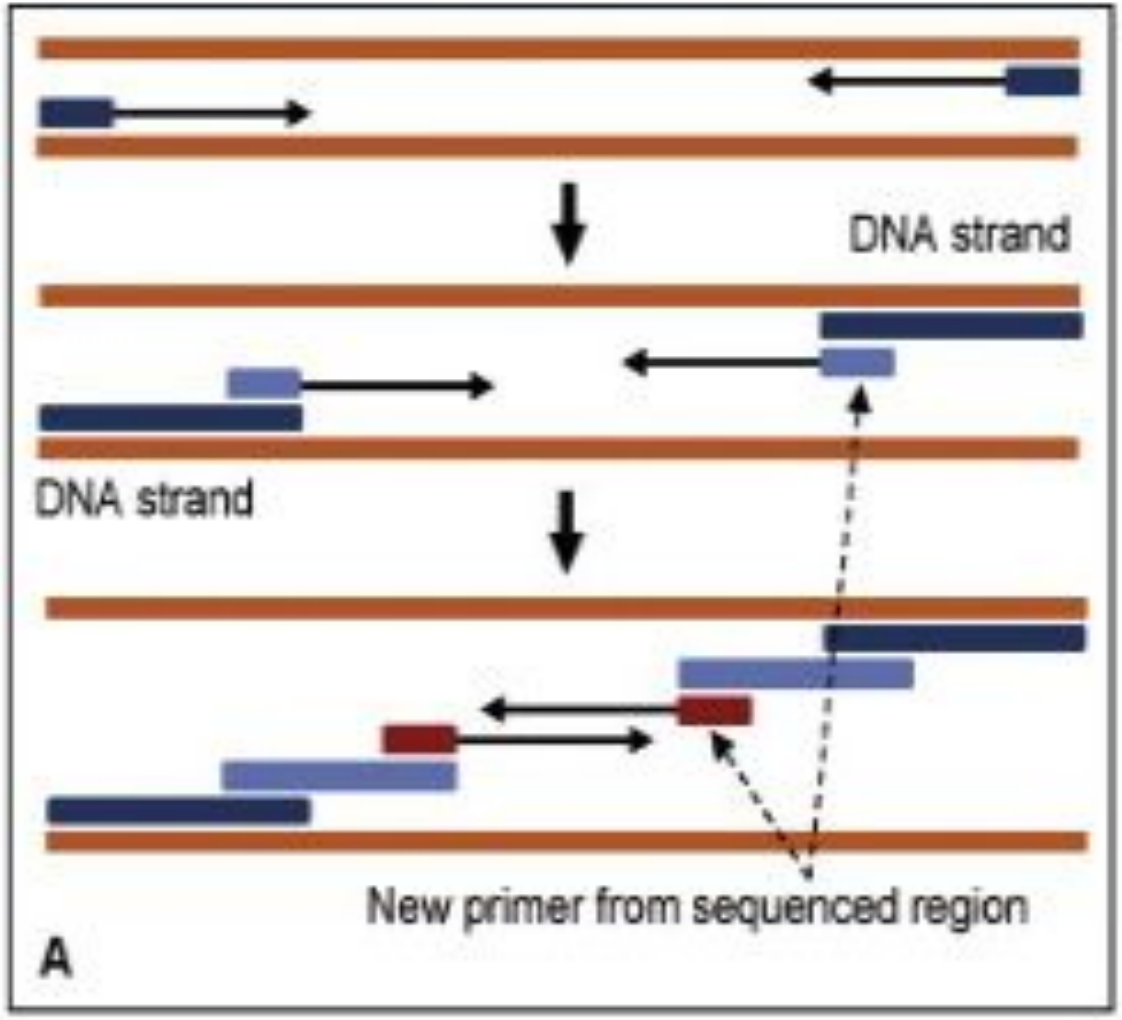
- 1) Break it all up into small pieces.
- 2) Clone it all into BACs

Hierarchical Shotgun Sequencing



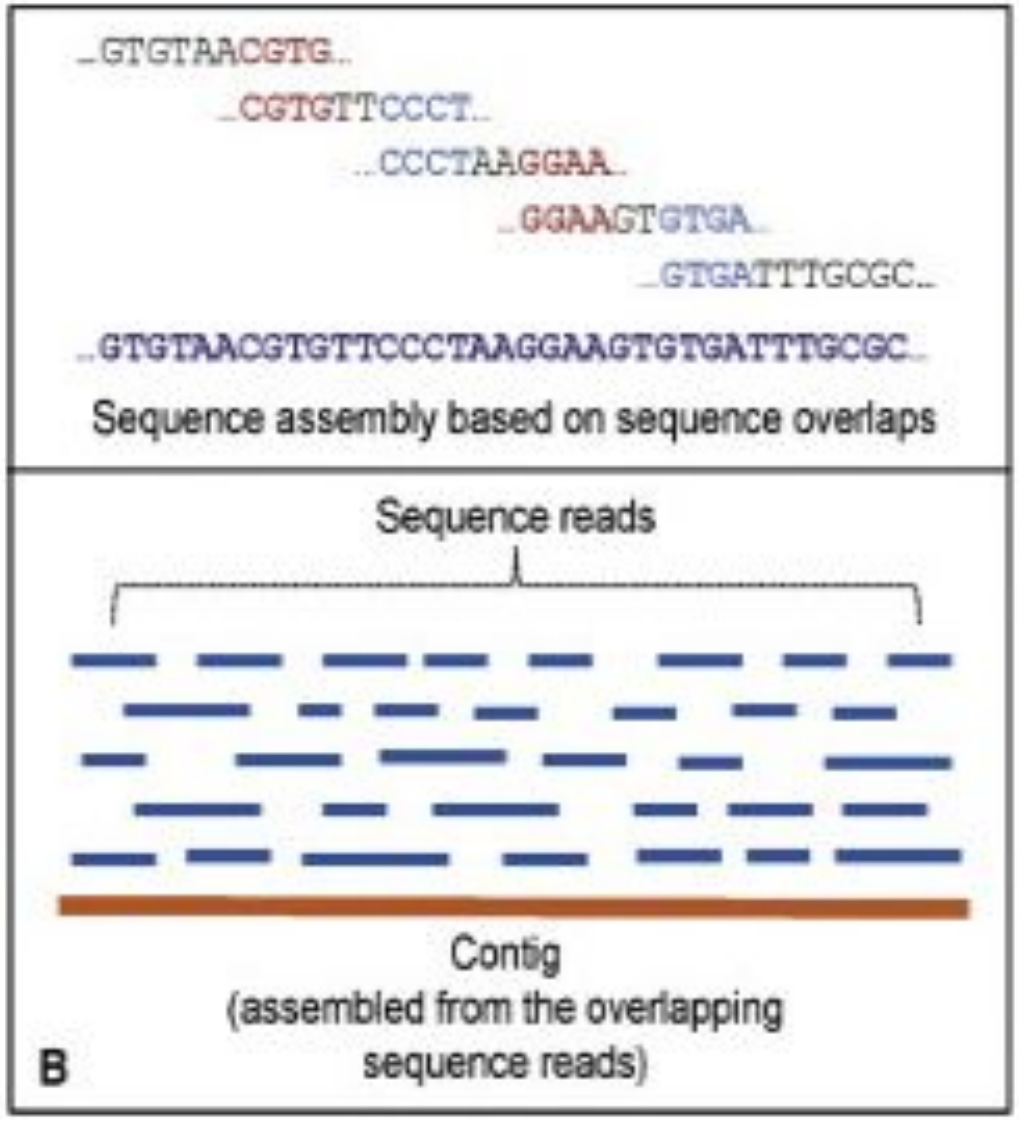
- 1) Break it all up into small pieces.
- 2) Clone it all into BACs
- 3) Sequence the ends: Tiling
  - Use a primer directed to the BAC
  - Then align the sequenced ends computationally
  - Produces an ordered sequence of BACs

Hierarchical Shotgun Sequencing



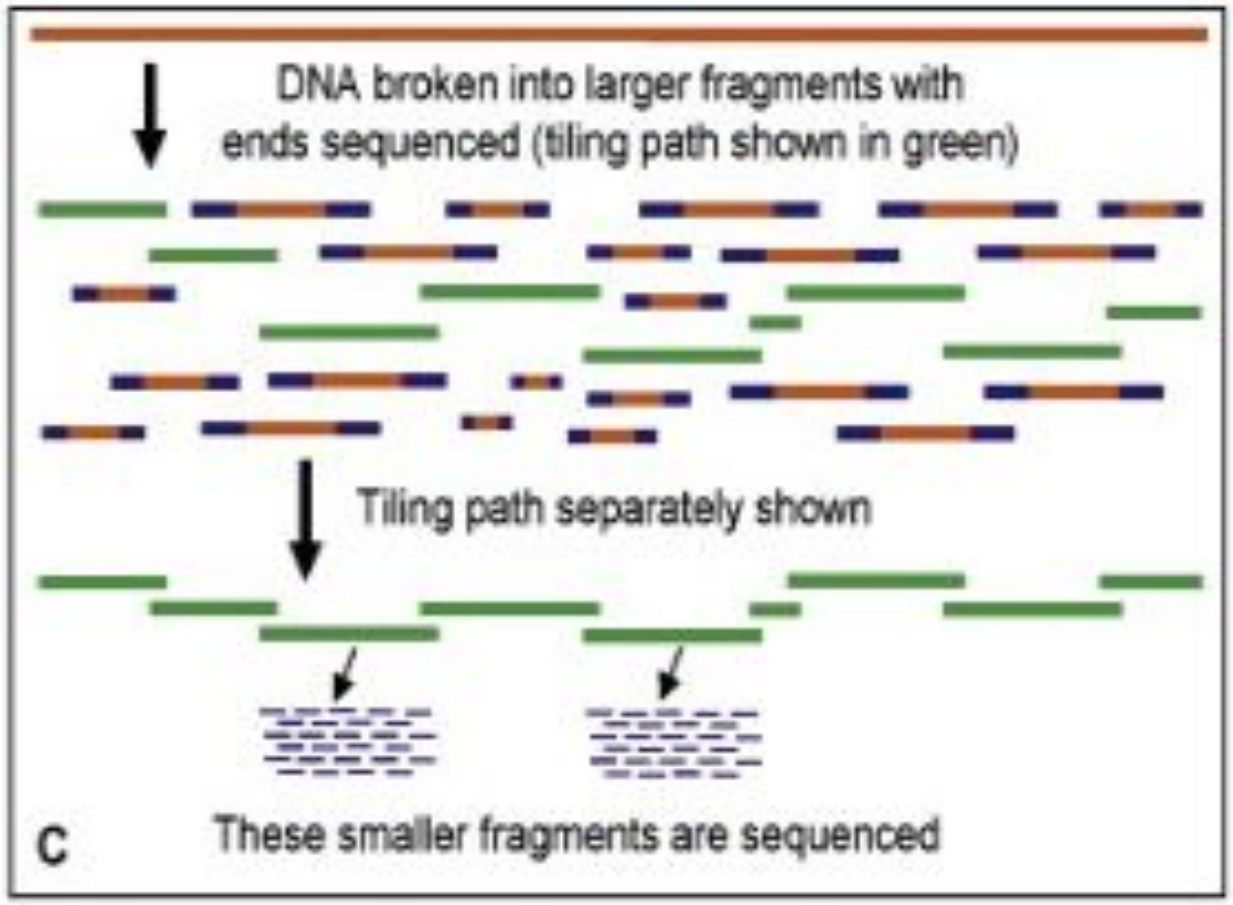
- 1) Break it all up into small pieces.
- 2) Clone it all into BACs
- 3) Sequence the ends
  - Use a primer directed to the BAC
  - Then align the sequenced ends computationally
  - Produces an ordered sequence of BACs
  - BUT, BAC cloned DNA is large: > 10,000 base pairs
  - AND, Sanger can only do ~ 500bp at a time
- 4) Primer walking
  - Design new primers after each Sanger sequencing
  - Fills in the BAC sequence

Hierarchical Shotgun Sequencing



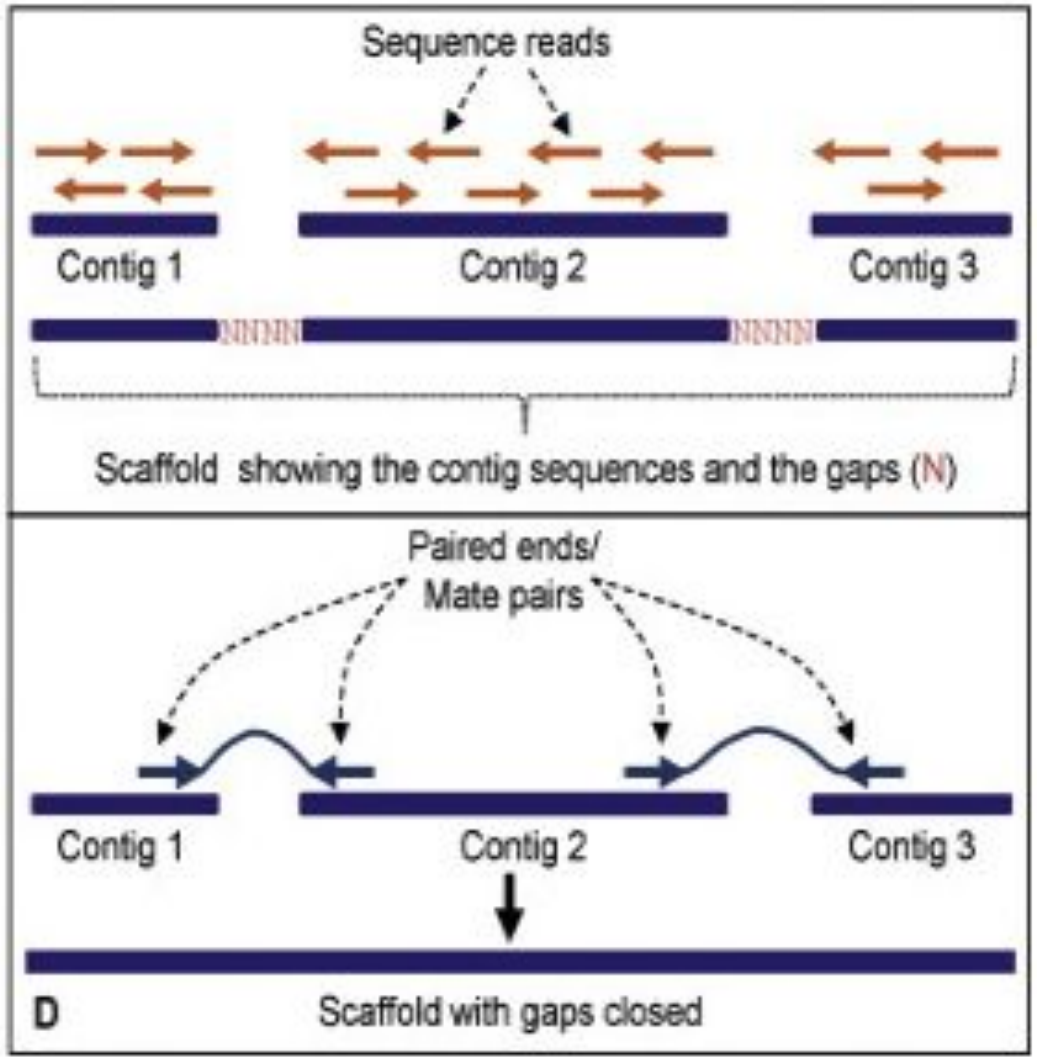
- 1) Break it all up into small pieces.
- 2) Clone it all into BACs
- 3) Sequence the ends
- 4) Primer walking
  - Contigs
  - ‘Contiguous overlapping sequences’

Hierarchical Shotgun Sequencing



- 1) Break it all up into small pieces.
- 2) Clone it all into BACs
- 3) Sequence the ends: Tiling
- 4) Primer walking
  - Contigs
  - ‘Contiguous overlapping sequences’

Hierarchical Shotgun Sequencing

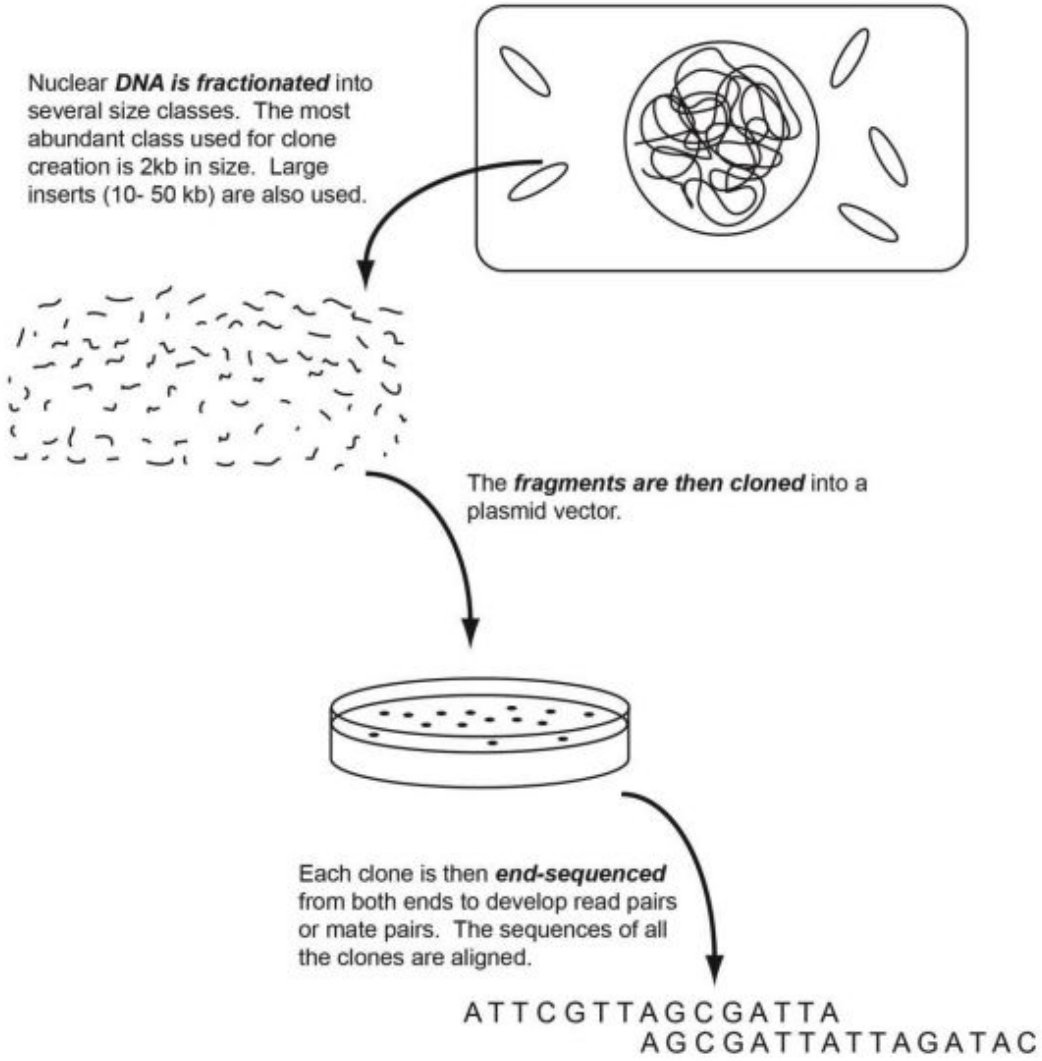


- 1) Break it all up into small pieces.
- 2) Clone it all into BACs
- 3) Sequence the ends: Tiling
- 4) Primer walking
- 5) Gap closure
  - Most time-consuming
  - Still going on today

Hierarchical Shotgun Sequencing

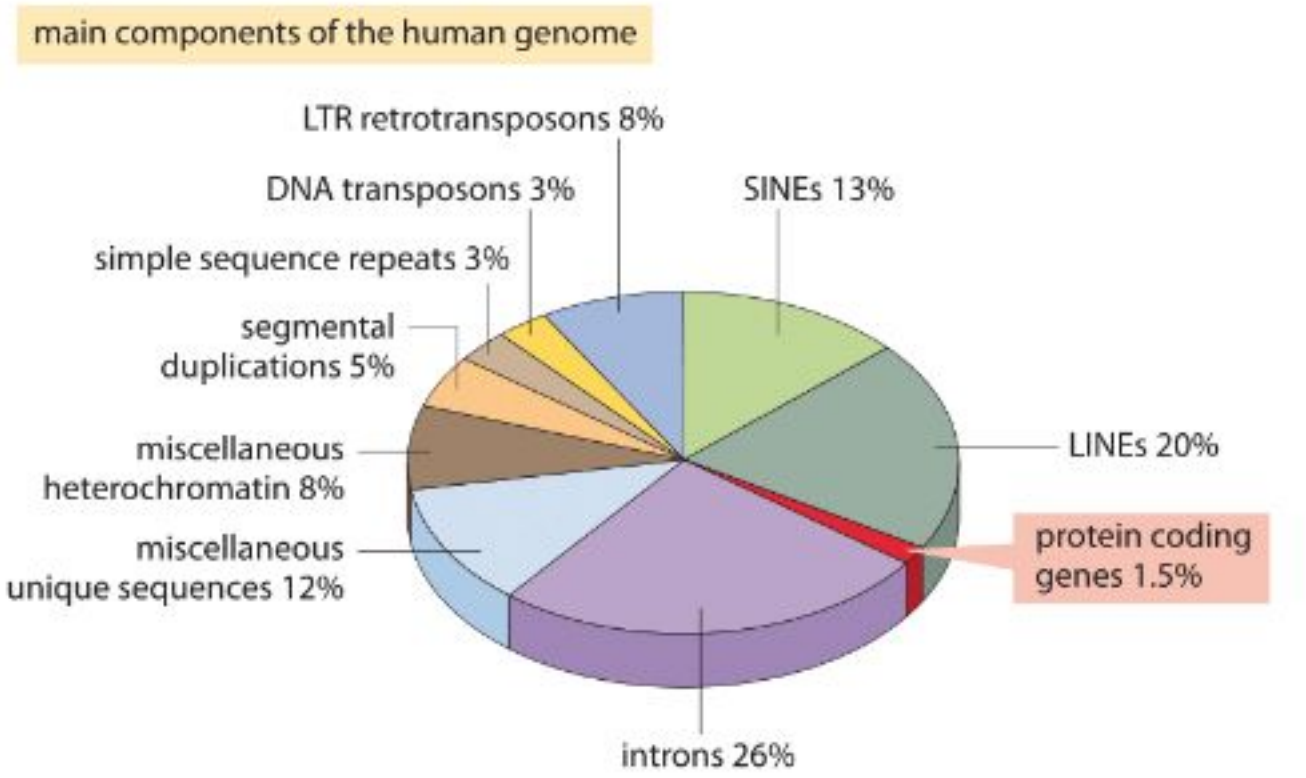
- 1) Break it all up into small pieces.
- 2) Clone it all into BACs
- 3) Sequence the ends: Tiling
- 4) Primer walking
- 5) Gap closure
  - Most time-consuming
  - Still going on today

A second approach: Shotgun sequencing



- 1) Break it all up into a variety of small pieces.
  - 100s bp to ~ 2000 bp
  - Also some large clones
- 2) Clone into a vector (no BACs)
- 3) Sequence
- 4) Let the computer put it together

A second approach: Shotgun sequencing



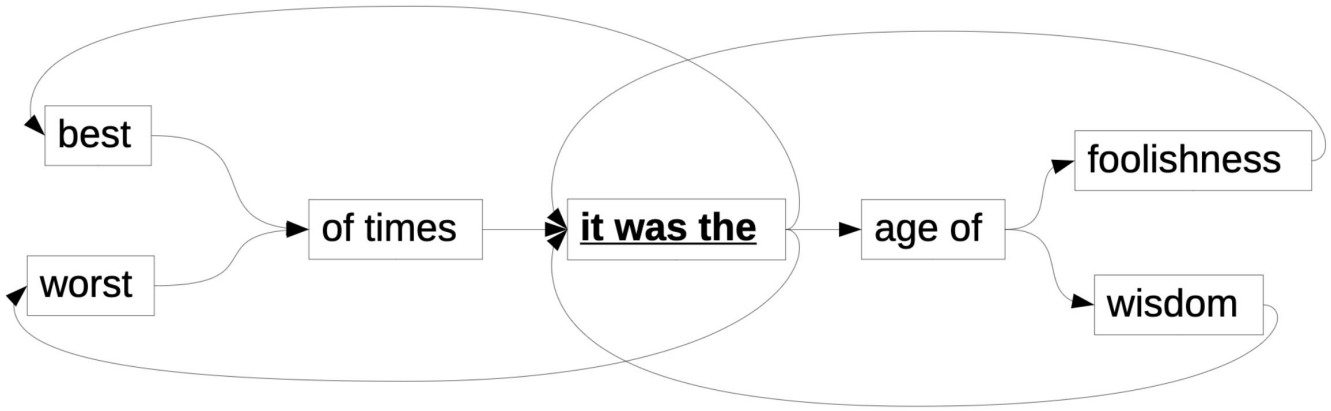
- 1) Break it all up into a variety of small pieces.
  - 100s bp to ~ 2000 bp
  - Also some large clones
- 2) Clone into a vector (no BACs)
- 3) Sequence
- 4) Let the computer put it together
- Problem: repetitive DNA > 50%
  - Telomeres
  - LINES, SINES
  - Centromeres

A second approach: Shotgun sequencing

Read = 3 “words” (<= length of repeat)

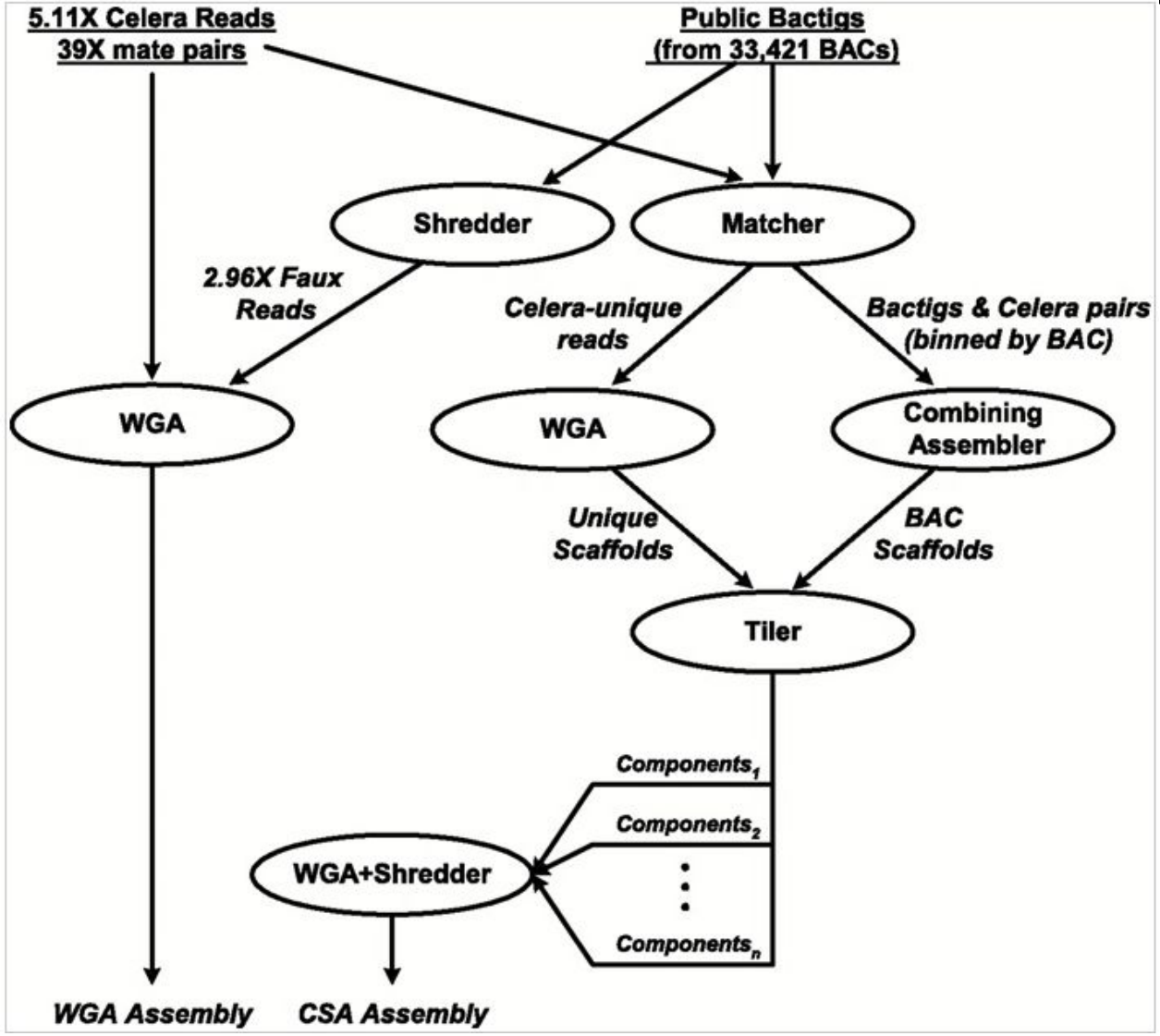
it was the best *of times* it was the worst *of times*
it was the *age of* wisdom it was the *age of* foolishness

it was the, was the best, was the worst, was the age
the age of,...



- 1) Break it all up into a variety of small pieces.
  - 100s bp to ~ 2000 bp
  - Also some large clones
- 2) Clone into a vector (no BACs)
- 3) Sequence
- 4) Let the computer put it together
- Problem: repetitive DNA > 50%
  - Telomeres
  - LINES, SINES
  - Centromeres
- Difficult to align computationally

A second approach: Shotgun sequencing



- 1) Break it all up into a variety of small pieces.
  - 100s bp to ~ 2000 bp
  - Also some large clones
- 2) Clone into a vector (no BACs)
- 3) Sequence
- 4) Let the computer put it together
- Celera: private company founded after personality conflicts in the public consortium effort
- Used public BAC assembly data to finish its computational assembly.

Announced at The White House in June 2000

Abstract

Since its initial release in 2000, the human reference genome has covered only the euchromatic fraction of the genome, leaving important heterochromatic regions unfinished. Addressing the remaining 8% of the genome, the Telomere-to-Telomere (T2T) Consortium presents a complete 3.055 billion–base pair sequence of a human genome, T2T-CHM13, that includes gapless assemblies for all chromosomes except Y, corrects errors in the prior references, and introduces nearly 200 million base pairs of sequence containing 1956 gene predictions, 99 of which are predicted to be protein coding. The completed regions include all centromeric satellite arrays, recent segmental duplications, and the short arms of all five acrocentric chromosomes, unlocking these complex regions of the genome to variational and functional studies.

Still not quite done 22 years later, but getting closer…

Hierarchical Shotgun Sequencing

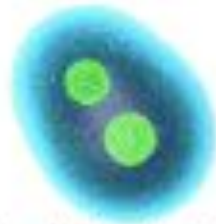
- 1) Break it all up into small pieces.
- 2) Clone it all into BACs
- 3) Sequence the ends: Tiling
- 4) Primer walking
- 5) Gap closure
  - Most time-consuming
  - Still going on today

Sanger method

1977

Human Genome Project

1990



Complete eukaryotic genome

1996

Second generation sequencer：454 GS20

2005

Research Human Microbiome Project

2008

Nanospace sequencing

2014

1981

Human mitochondrial genome sequence

1995

Complete cell genome

2001

Complete the Human Genome Project

2007

Second generation sequencer：Genetic Analyzer 2

2011

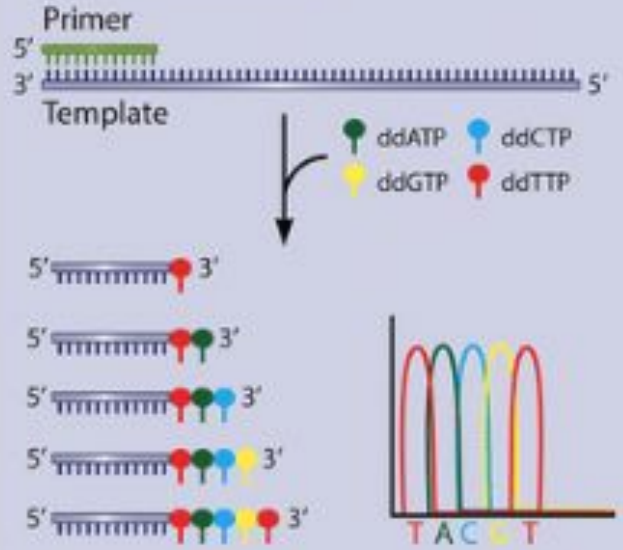
Third generation sequencer：PacBio RS

2019

The third stage Human microbiome project

Generational changes

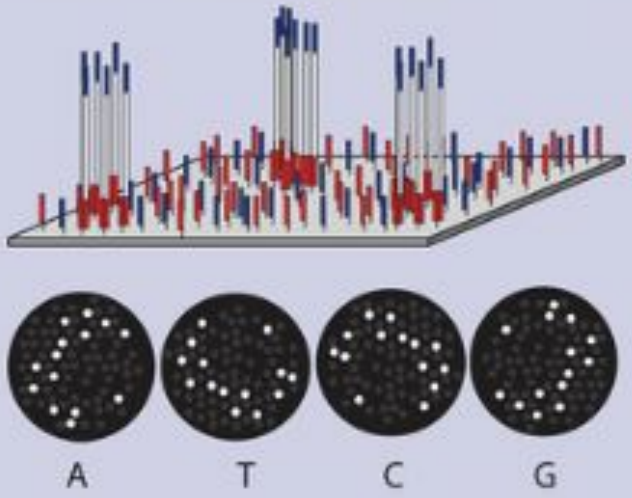
First Generation
Shotgun Sequencing



- Sequencing by synthesis
- High accuracy
- Long read lengths
- Relatively small amount of data generated

e.g., ABI capillary sequencer (ABI)

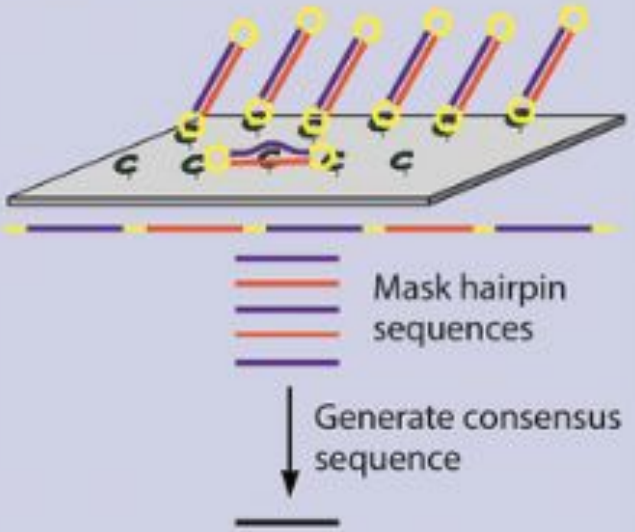
Second Generation
Massively Parallel Sequencing



- Sequencing by synthesis
- Amplified templates are generated during sequencing, reducing the requirements for starting material
- High accuracy
- Short read lengths

e.g., MiSeq (Illumina), Ion Torrent (Thermo Fisher Scientific)

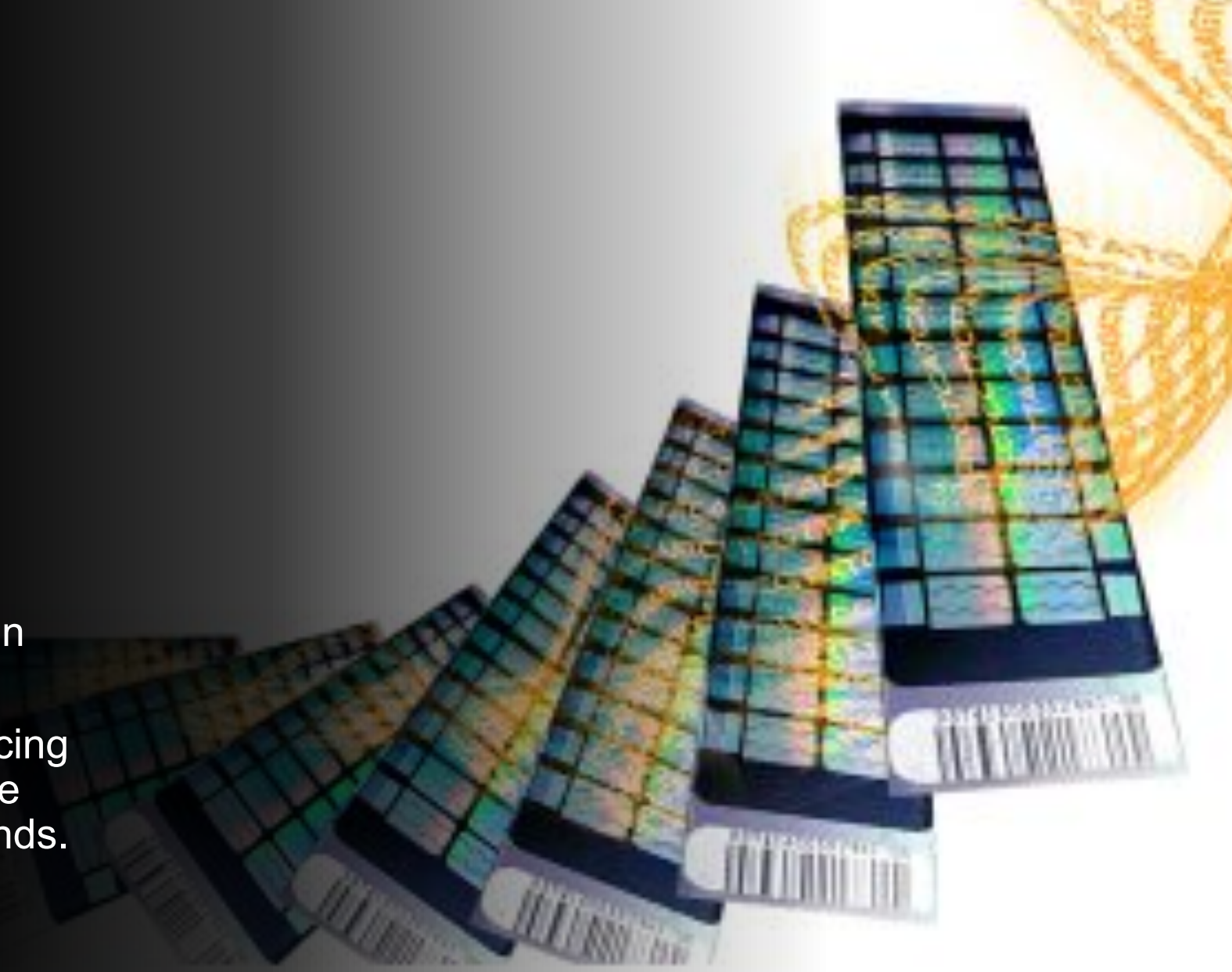
Third Generation
Single-molecule Sequencing



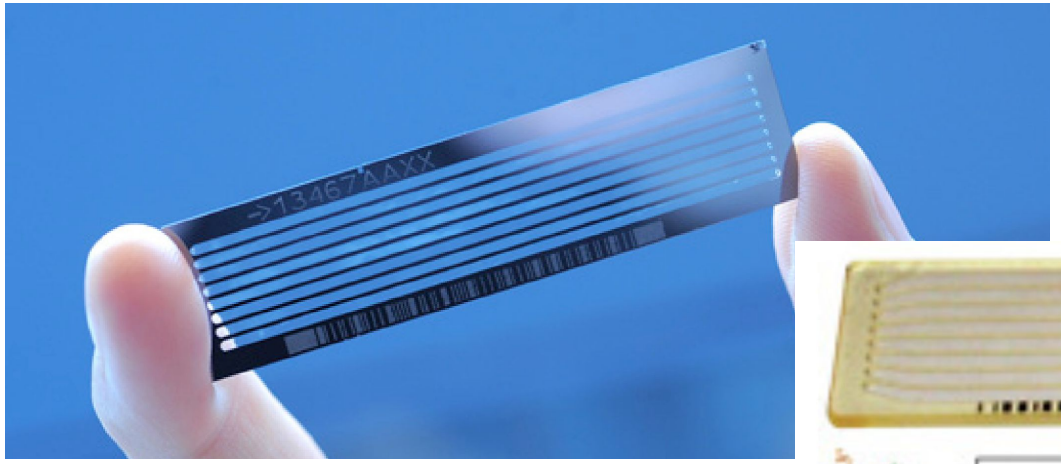
- Single-molecule templates
- Low accuracy
- Long read lengths

e.g., Single-Molecule Real-Time (SMRT) — Sequencing (Pacific Biosciences), MinION (Oxford Nanopore Technologies)

- Illumina sequencing technology is based on sequencing by synthesis (SBS)
- responsible for generating more than 90% of the world's sequencing data
- supports massively parallel sequencing that detects single bases as they are incorporated into growing DNA strands.



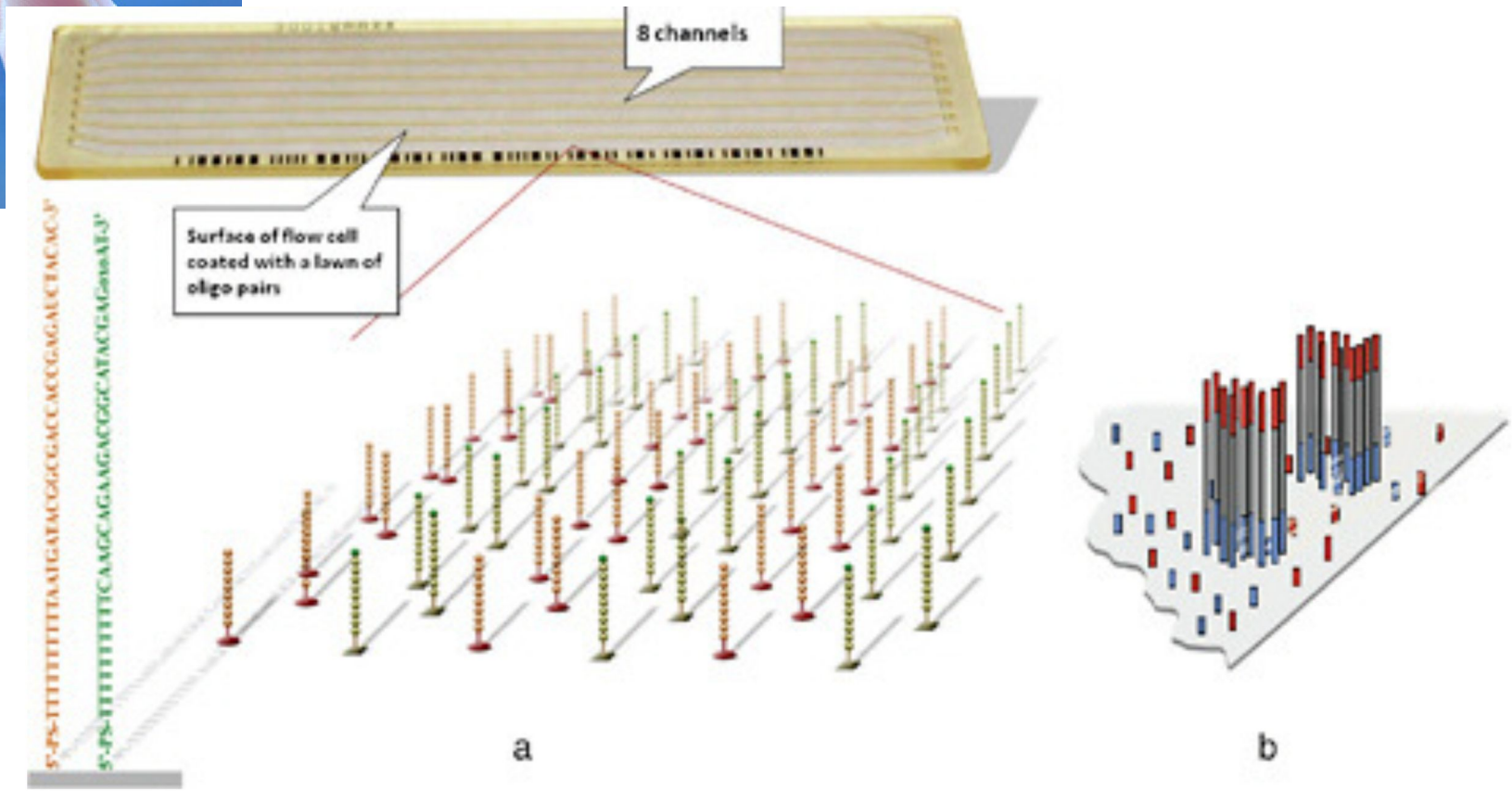
Illumina Sequencing

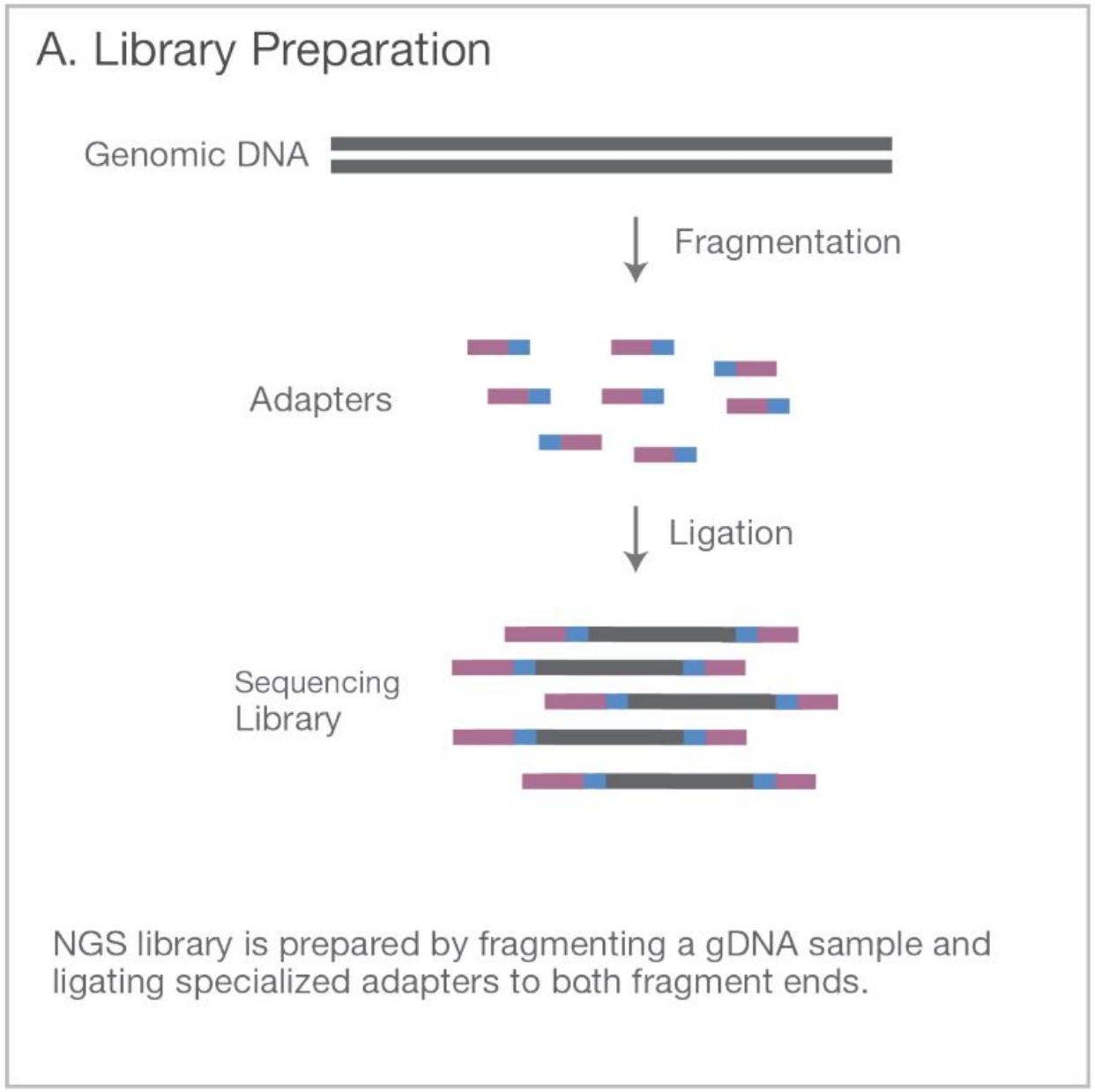


- Target DNA is immobilized in a ‘flow cell’

- A solid support that allows reagents to flow over the DNA to be sequenced

- Immobilization requires an ‘adapter’ homologous to anchor DNA in the flow cell





Illumina sequencing

DNA is isolated and fragmented into small pieces

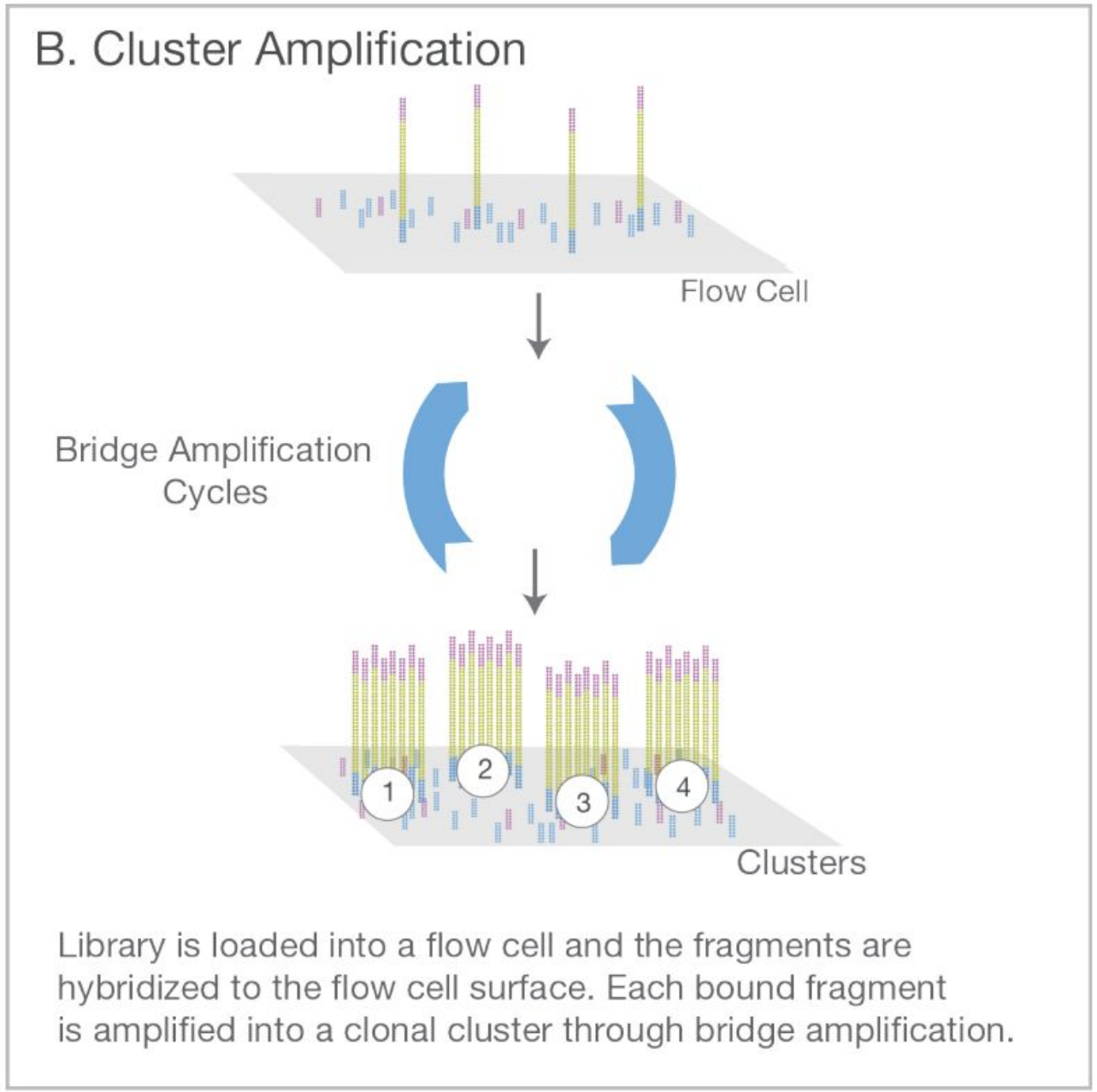
Adapters are ligated to the pieces of DNA

Allow binding to the flow cell slide

B) Bound DNA is amplified to create clonal clusters that increase signal-noise during sequencing/imaging

C) Sequential rounds of individual addition of reversible dye-terminator bases

D) Translation of imaging records to sequence data. Export and analysis of sequence data.



Illumina sequencing

DNA is isolated and fragmented into small pieces

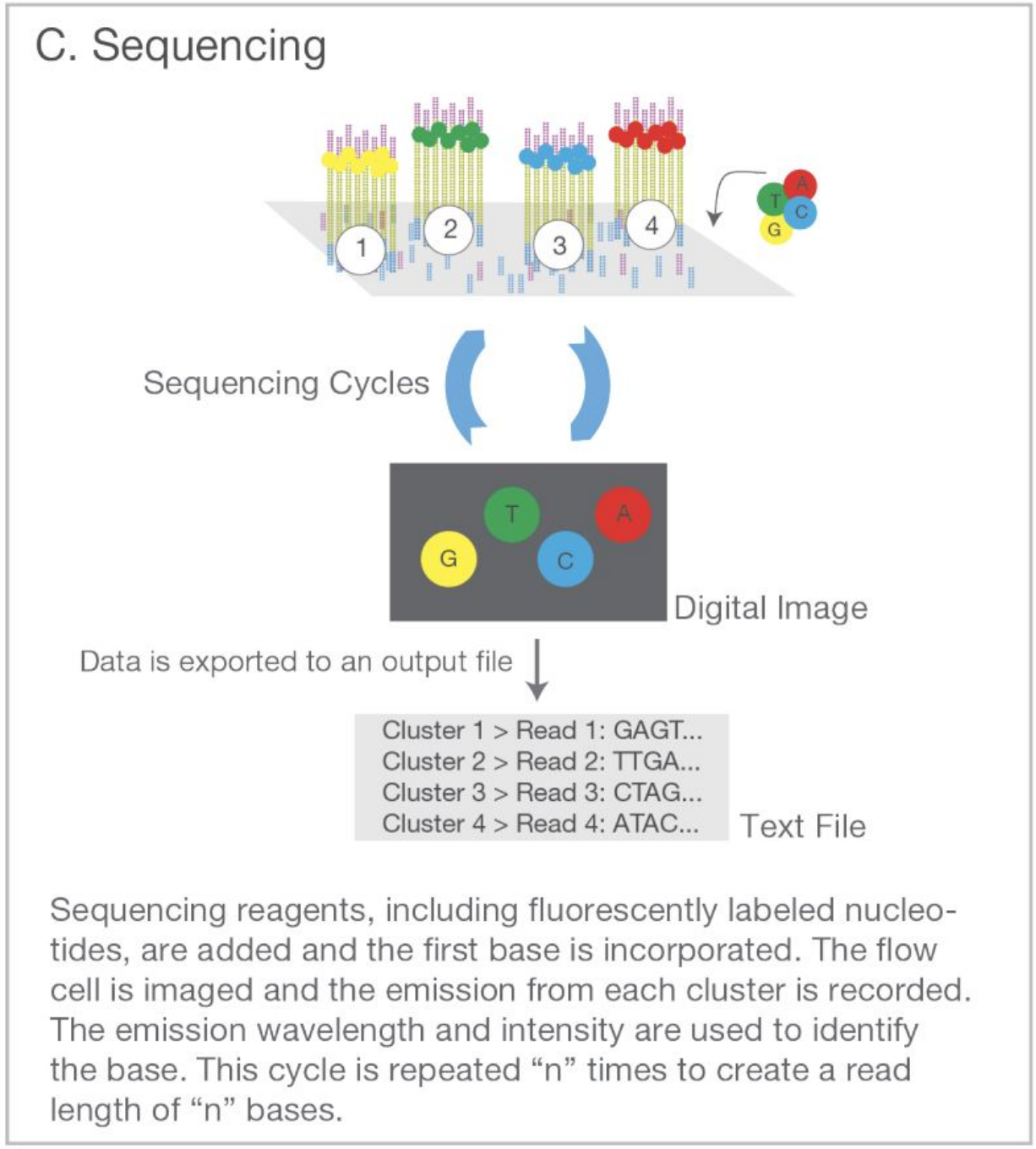
Adapters are ligated to the pieces of DNA

- Allow binding to the flow cell slide

B) Bound DNA is amplified to create clonal clusters that increase signal-noise during sequencing/imaging

C) Sequential rounds of individual addition of reversible dye-terminator bases

D) Translation of imaging records to sequence data. Export and analysis of sequence data.



Illumina sequencing

DNA is isolated and fragmented into small pieces

Adapters are ligated to the pieces of DNA

- Allow binding to the flow cell slide

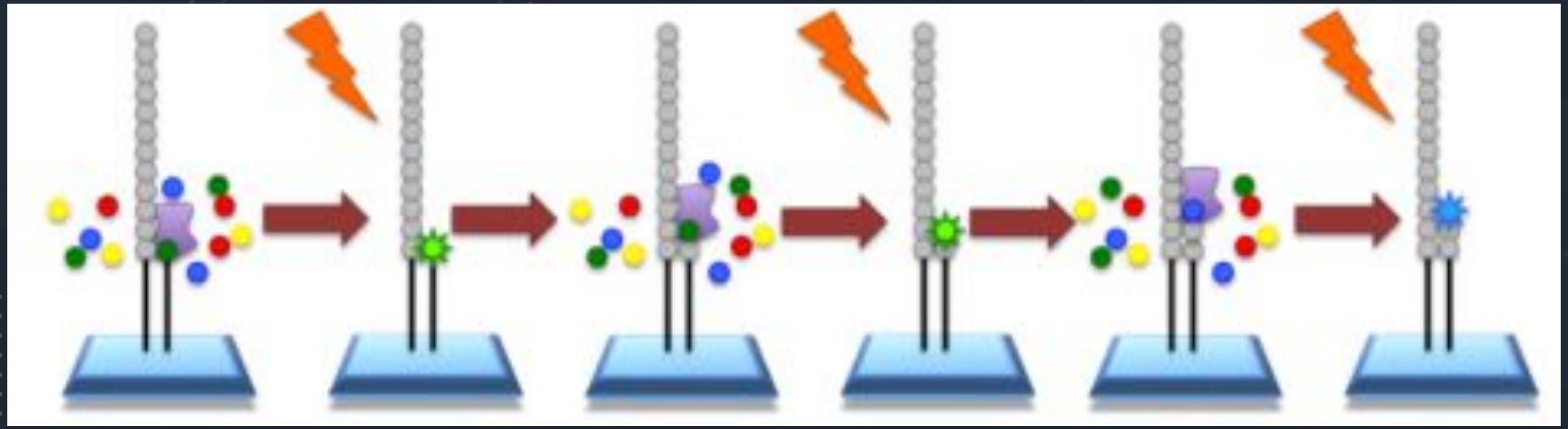
B) Bound DNA is amplified to create clonal clusters that increase signal-noise during sequencing/imaging

C) Sequential rounds of individual addition of reversible dye-terminator bases

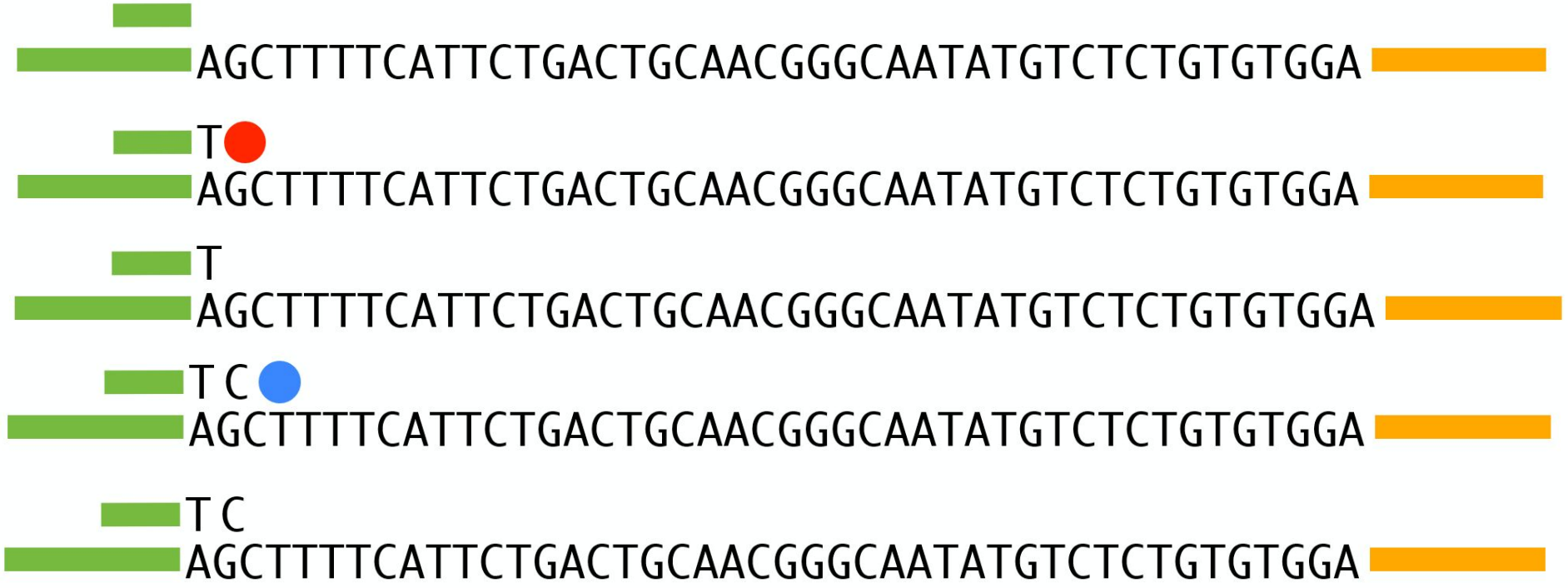
D) Translation of imaging records to sequence data. Export and analysis of sequence data.

- Makes use of reversible dye-terminators
  - four types of fluorescently labeled ddNTPs
  - terminal 3' blocker is chemically removed
- DNA can only be extended one nucleotide at a time:
  - A, T, C, G, then A, T, C, G then A, T….

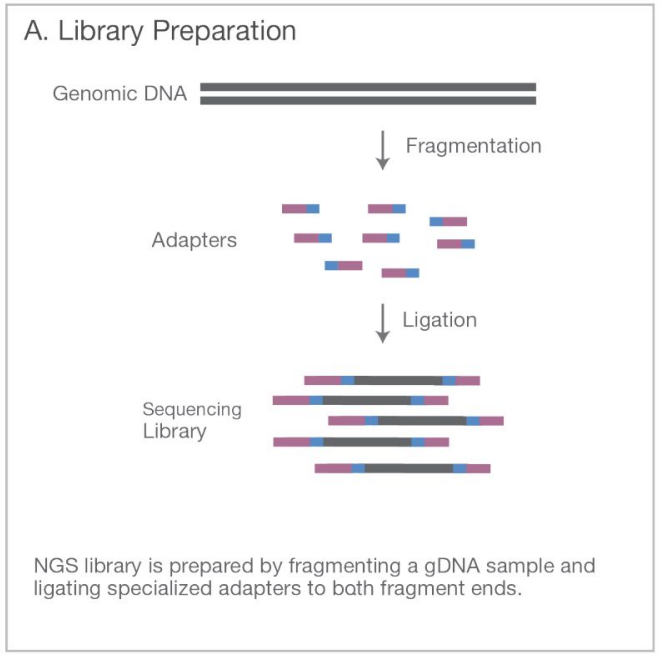
- A camera takes images of the fluorescently labeled nucleotide after each addition of NTPs to the DNA
  - An Imaging based technique
- then the dye along with the terminal 3' blocker is chemically removed from the DNA, allowing the next cycle
- Several 100 Mio sequences

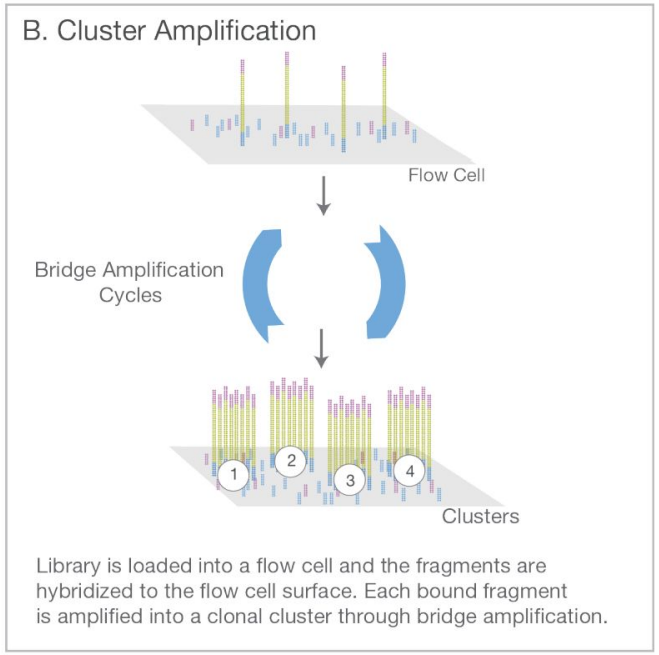


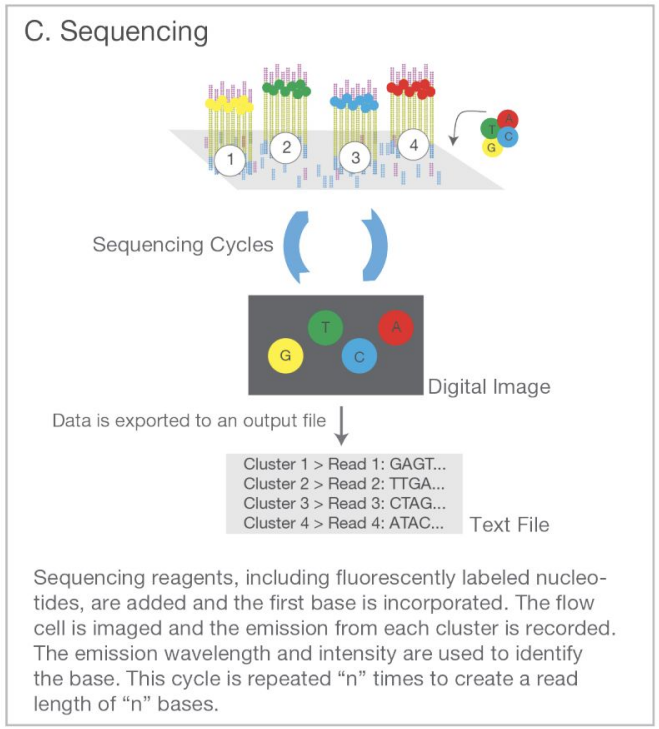
Illumina Sequencing



To be continued……







Illumina sequencing

DNA is isolated and fragmented into small pieces

Adapters are ligated to the pieces of DNA

- Allow binding to the flow cell slide

B) Bound DNA is amplified to create clonal clusters that increase signal-noise during sequencing/imaging

C) Sequential rounds of individual addition of reversible dye-terminator bases

D) Translation of imaging records to sequence data. Export and analysis of sequence data.

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

Sanger Sequencing vs Illumina sequencing

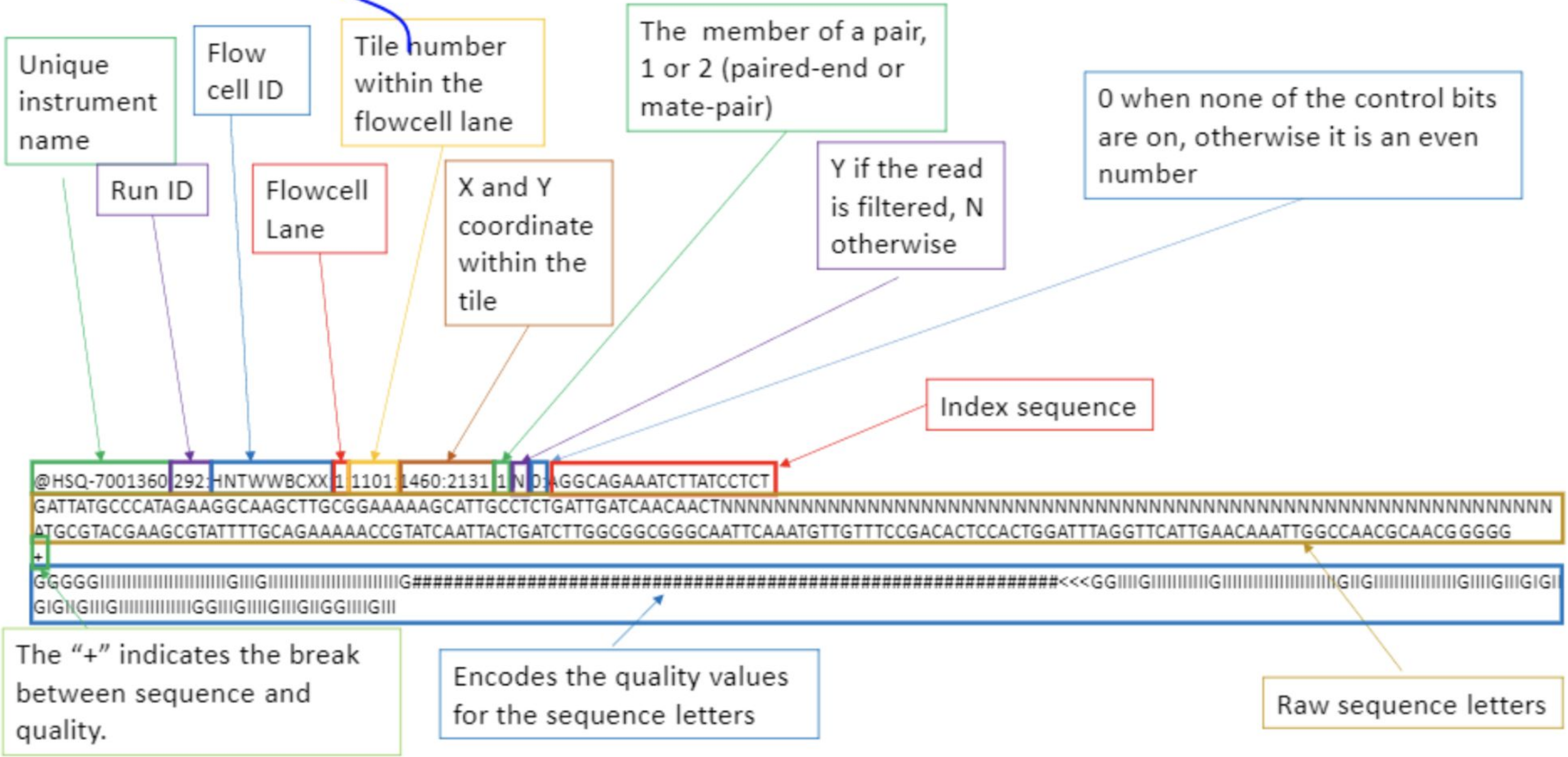
Sanger sequencing

- Accurate and reproducible
- Expensive, time-consuming, and labor-intensive
- Requires capillaries or sequencing gels
- But, method capable of producing long sequencing runs (~800 bp)

Illumina sequencing

- accurate
- fully automated
- Expensive but produces 10s-100s Million reads per run
- Read lengths are around ~100-300 bp
- Suitable for genome sequencing and transcriptomic studies

FASTQ File Format Analysis

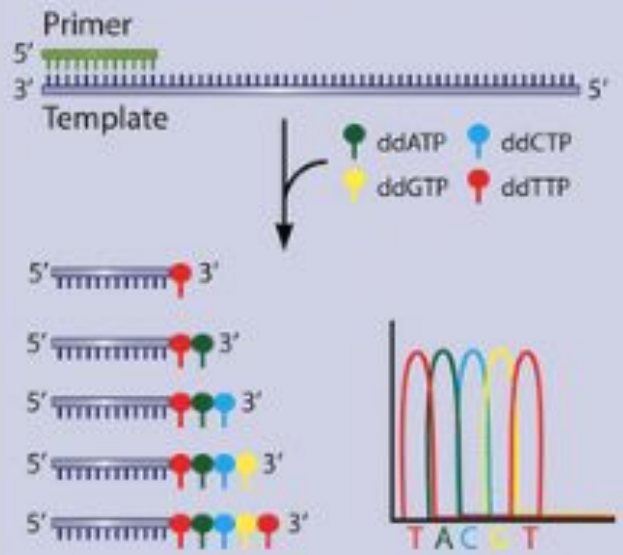


Week 8 **March 6**	**No Lecture:** Independent project proposals	Independent project proposals during lab time	**Independent project** **Week 1: proposals**
Week 9 **March 13**	Sequencing and bioinformatics II	IP: Sample preparation & enumeration of bacteria	**Independent project** **Week 2: sample prep**
Week 10	Metagenomics	Metagenomics lab	**Metagenomics lab**

March 20	introduction lecture (Paul Scheid, milrd.org)		
Week 11 **March 27**	Lecture TBD, depending on independent projects	IP: AMR gene & 16S PCR,	**Independent project** **Week 3: AMR gene PCR**
Week 12 **April 3**	**No Lecture**	**No Lab Sections**	**Spring Break**
Week 13 **April 10**	No Lecture on Monday Lecture on Wednesday 4/12	Open Lab time	Open Lab time
Week 14 **April 17**	Presentation info Lecture TBD, depending on independent projects	IP: Gel of PCR from IP Week 3 and qPCR setup	**Independent project** **Week 4: qPCR & Gels**
Week 15 **April 24**	Lecture TBD, depending on independent projects	Final data analysis and presentation prep	**Independent project** **Week 5: data analysis and wrap up**
Week 16 **April 1**	Final Exam review, semester wrap-up	Presentations	

Generational changes

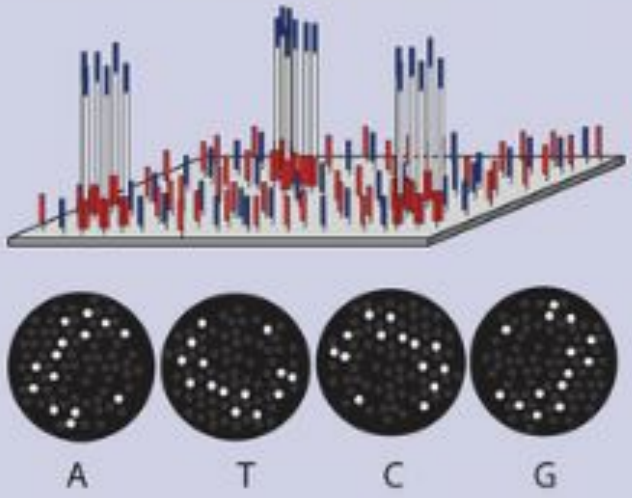
First Generation
Shotgun Sequencing



- Sequencing by synthesis
- High accuracy
- Long read lengths
- Relatively small amount of data generated

e.g., ABI capillary sequencer (ABI)

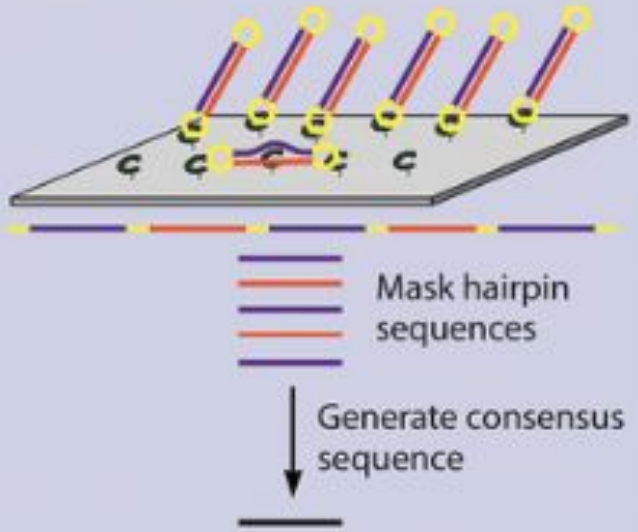
Second Generation
Massively Parallel Sequencing



- Sequencing by synthesis
- Amplified templates are generated during sequencing, reducing the requirements for starting material
- High accuracy
- Short read lengths

e.g., MiSeq (Illumina), Ion Torrent (Thermo Fisher Scientific)

Third Generation
Single-molecule Sequencing



- Single-molecule templates
- Low accuracy
- Long read lengths

e.g., Single-Molecule Real-Time (SMRT) — Sequencing (Pacific Biosciences), MinION (Oxford Nanopore Technologies)