Introduction to DNA Sequencing

(Relevant) Trivia

How many base pairs (bp) are there in a human genome?

How much did it cost to sequence the first human genome?

How long did it take to sequence the first human genome?

When was the first human genome sequence complete?

Whose genome was it?

(Relevant) Trivia

How many base pairs (bp) are there in a human genome?

~3 billion (haploid)

How much did it cost to sequence the first human genome?

~$2.7 billion

How long did it take to sequence the first human genome?

~13 years

When was the first human genome sequence complete?

2000-2003

Whose genome was it?

Several people's, but actually mostly a dude from Buffalo

Overview

- Prologue: **Assembly**
- The Past: **Sanger**
- The Present: **Next-Gen (454, Illumina, …)**
- The Future: **? (Nanopore, MinION, Single-molecule)**

Overview

- Prologue: **Assembly**
- The Past: Sanger
- The Present: Next-Gen (454, Illumina, …)
- The Future: ? (Nanopore, MinION, Single-molecule)

Method	Read Length
Sanger	
454	
Illumina	
Ion Torrent	

Method	Read Length
Sanger	600-1000 bp
454	
Illumina	
Ion Torrent	

Method	Read Length
Sanger	600-1000 bp
454	300-500 bp
Illumina	
Ion Torrent	

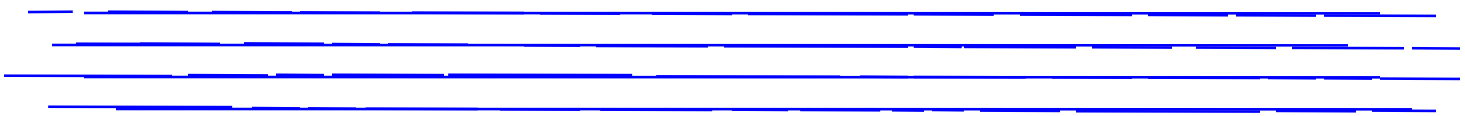
Method	Read Length
Sanger	600-1000 bp
454	300-500 bp
Illumina	~100 bp
Ion Torrent	

Method	Read Length
Sanger	600-1000 bp
454	300-500 bp
Illumina	~100 bp
Ion Torrent	~200 bp

But…

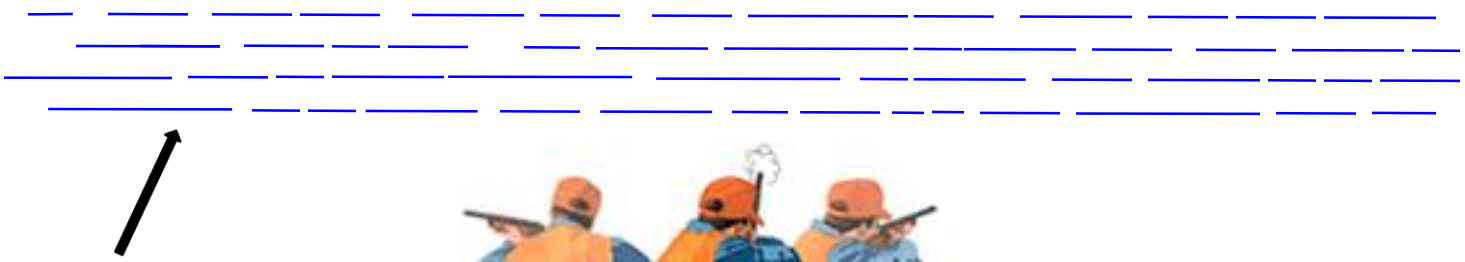
- Phage Genome: 30,000 to 500,000 bp
- Bacteria: Several million bp
- Human: 3 billion bp

Shotgun Genome Sequencing



Fragmented genome chunks

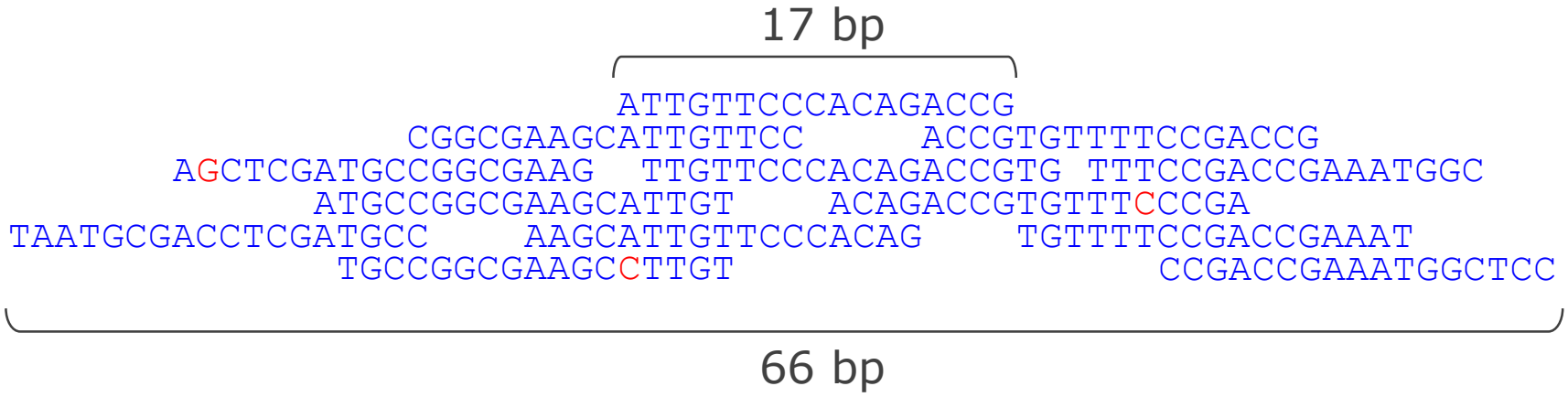
Shotgun Genome Sequencing



Fragmented genome chunks

NOT REALLY DONE BY DUCK HUNTERS
Hydroshearing, sonication, enzymatic shearing

Some Assembly Required



Some Assembly Required

Consensus:

TAATGCGACCTCGATGCCGGCGAAGCATTGTTCCCACAGACCGTGTTTTCCGACCGAAATGGCTCC

```
                          ATTGTTCCCACAGACCG
                 CGGCGAAGCATTGTTCC     ACCGTGTTTTCCGACCG
       AGCTCGATGCCGGCGAAG  TTGTTCCCACAGACCGTG TTTCCGACCGAAATGGC
             ATGCCGGCGAAGCATTGT    ACAGACCGTGTTTCCCGA
TAATGCGACCTCGATGCC    AAGCATTGTTCCCACAG    TGTTTTCCGACCGAAAT
              TGCCGGCGAAGCCTTGT                  CCGACCGAAATGGCTCC
```

Some Assembly Required

Consensus:

TAATGCGACCTCGATGCCGGCGAAGCATTGTTCCCACAGACCGTGTTTTCCGACCGAAATGGCTCC

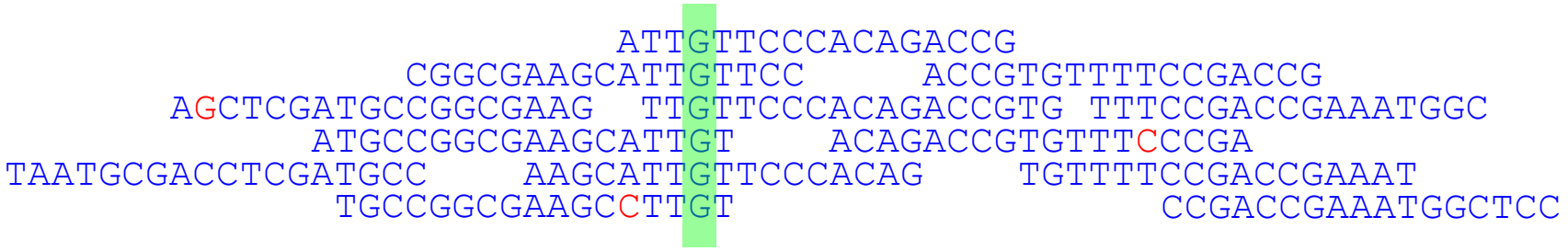
```
                          ATTGTTCCCACAGACCG
                 CGGCGAAGCATTGTTCC     ACCGTGTTTTCCGACCG
       AGCTCGATGCCGGCGAAG  TTGTTCCCACAGACCGTG TTTCCGACCGAAATGGC
             ATGCCGGCGAAGCATTGT    ACAGACCGTGTTTCCCGA
TAATGCGACCTCGATGCC    AAGCATTGTTCCCACAG    TGTTTTCCGACCGAAAT
              TGCCGGCGAAGCCTTGT                  CCGACCGAAATGGCTCC
```

Coverage: # of reads underlying the consensus

Some Assembly Required

Consensus:

TAATGCGACCTCGATGCCGGCGAAGCATTGTTCCCACAGACCGTGTTTTCCGACCGAAATGGCTCC



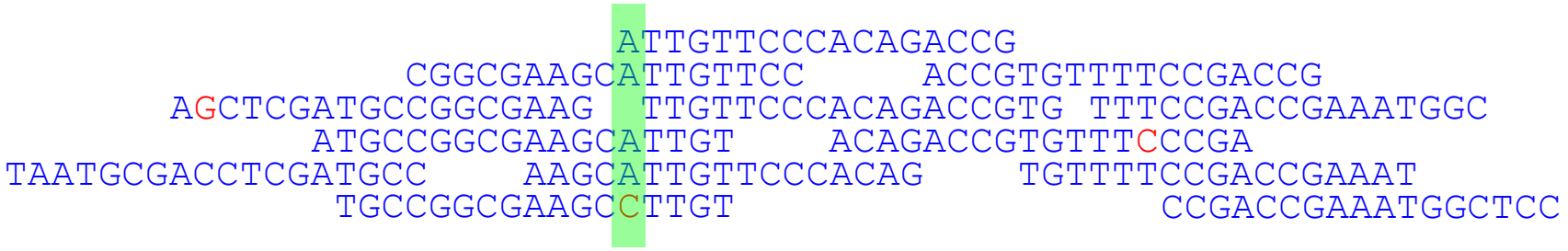
6x coverage
100% identity

Coverage: # of reads underlying the consensus

Assembly

Consensus:

TAATGCGACCTCGATGCCGGCGAAGCATTGTTCCCACAGACCGTGTTTTCCGACCGAAATGGCTCC



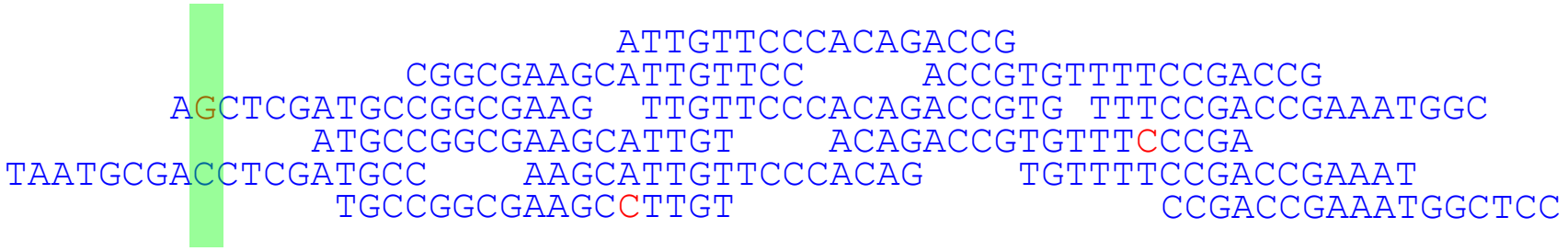
5x coverage
80% identity

Coverage: # of reads underlying the consensus

Assembly

Consensus:

TAATGCGACCTCGATGCCGGCGAAGCATTGTTCCCACAGACCGTGTTTTCCGACCGAAATGGCTCC



2x coverage
50% identity

Coverage: # of reads underlying the consensus

Assembly

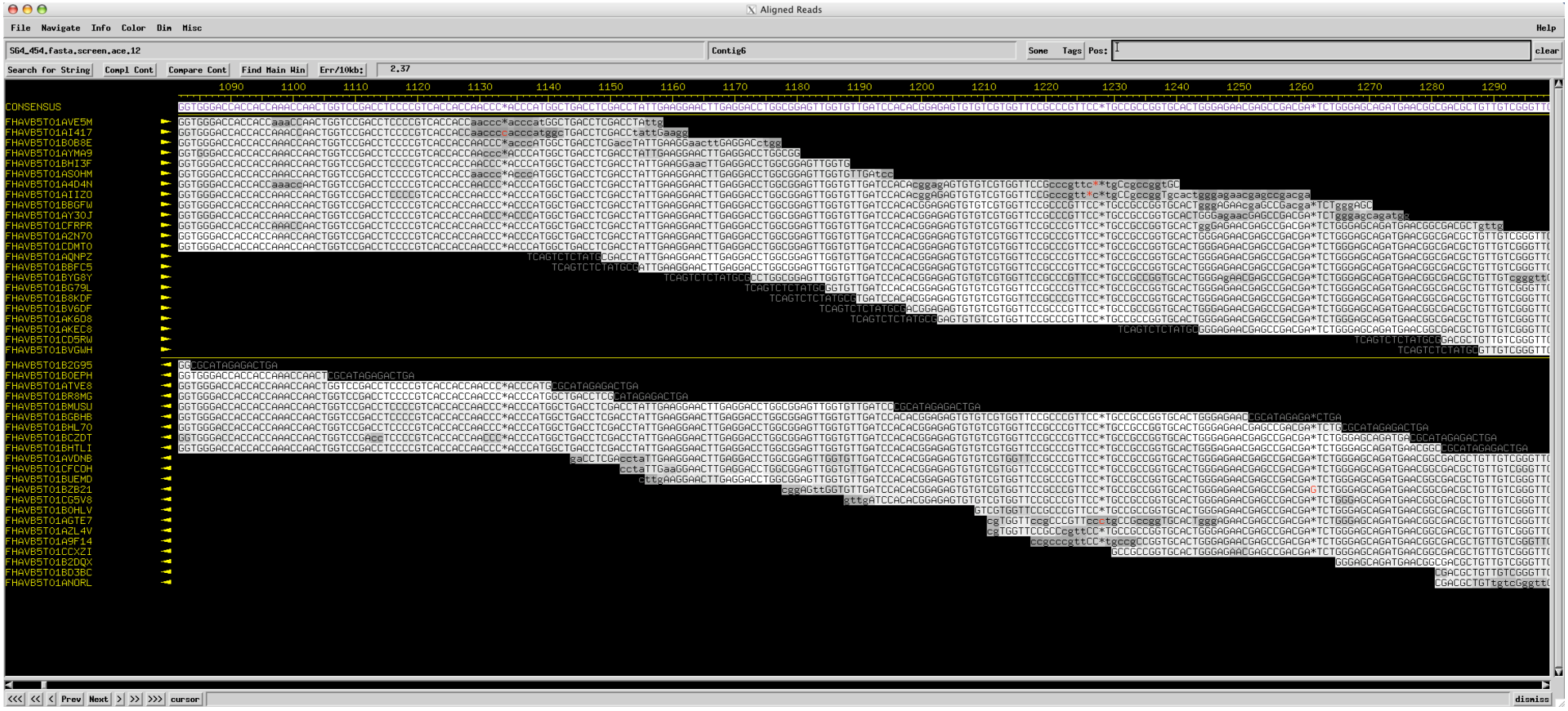
Consensus:

TAATGCGACCTCGATGCCGGCGAAGCATTGTTCCCACAGACCGTGTTTTCCGACCGAAATGGCTCC

1x coverage

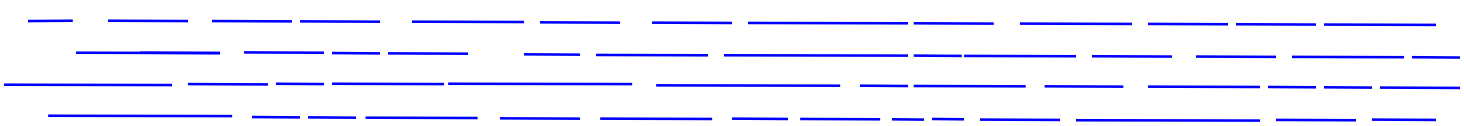
Coverage: # of reads underlying the consensus

Assembly

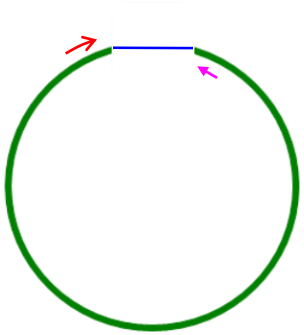

Aligned Reads
File Navigate Info Color Dim Misc
Help
SG4_454.fasta.screen.ace.12
Contig6
Some Tags Pos:
clear
Search for String
Compl Cont
Compare Cont
Find Main Win
Err/10kb:
2.37
1090 1100 1110 1120 1130 1140 1150 1160 1170 1180 1190 1200 1210 1220 1230 1240 1250 1260 1270 1280 1290
CONSENSUS
FHAVB5T01AVE5M
FHAVB5T01AI417
FHAVB5T01BOB8E
FHAVB5T01AYMA9
FHAVB5T01BHI3F
FHAVB5T01ASOHM
FHAVB5T01A4D4N
FHAVB5T01AIIZO
FHAVB5T01BGGFW
FHAVB5T01AY3OJ
FHAVB5T01CFPPR
FHAVB5T01AZN7O
FHAVB5T01CDMTO
FHAVB5T01AQNPZ
FHAVB5T01BBFC5
FHAVB5T01BYG8Y
FHAVB5T01BG79L
FHAVB5T01B8KDF
FHAVB5T01BV6DF
FHAVB5T01AK6O8
FHAVB5T01AKEC8
FHAVB5T01CD5RU
FHAVB5T01BVGWH
FHAVB5T01B2G95
FHAVB5T01BOEPH
FHAVB5T01ATVE8
FHAVB5T01BR8MG
FHAVB5T01BMUSU
FHAVB5T01BGBHB
FHAVB5T01BHL7O
FHAVB5T01BCZDT
FHAVB5T01BHTLI
FHAVB5T01AVDNB
FHAVB5T01CFCOH
FHAVB5T01BUEMD
FHAVB5T01BZBZ1
FHAVB5T01CB5V8
FHAVB5T01BOHLV
FHAVB5T01AGTE7
FHAVB5T01AZL4V
FHAVB5T01A9F14
FHAVB5T01CCXZI
FHAVB5T01B2DQX
FHAVB5T01BD3BC
FHAVB5T01ANORL
<<< << < Prev Next > >> >>> cursor
dismiss

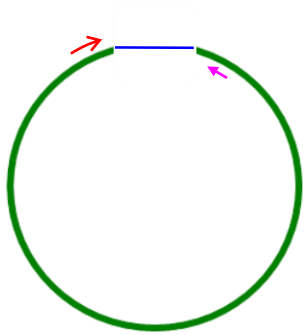
Overview

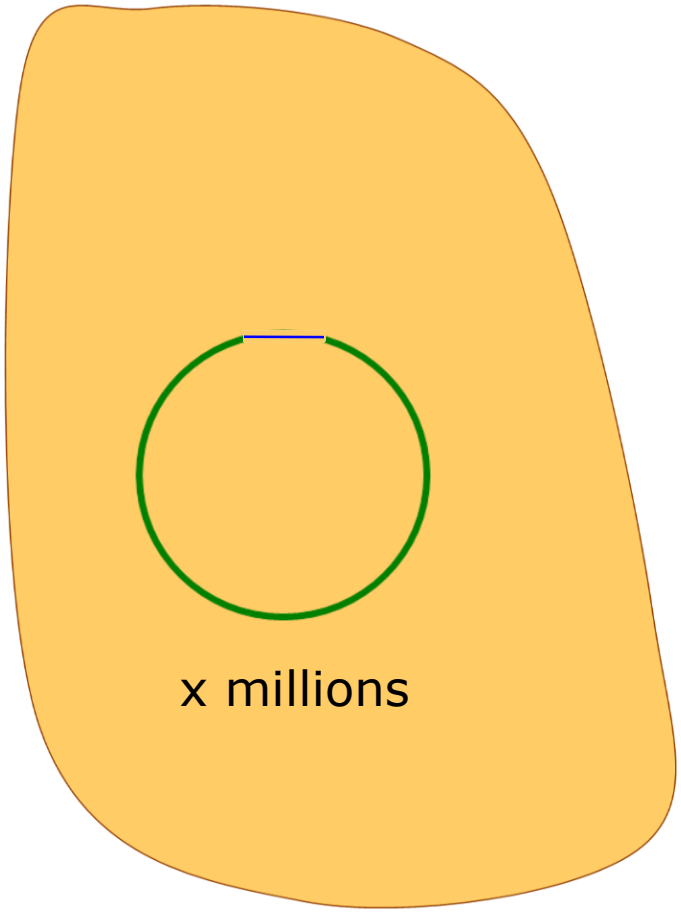
- Prologue: Assembly
- The Past: **Sanger**
- The Present: Next-Gen (454, Illumina, …)
- The Future: ? (Nanopore, MinION, Single-molecule)

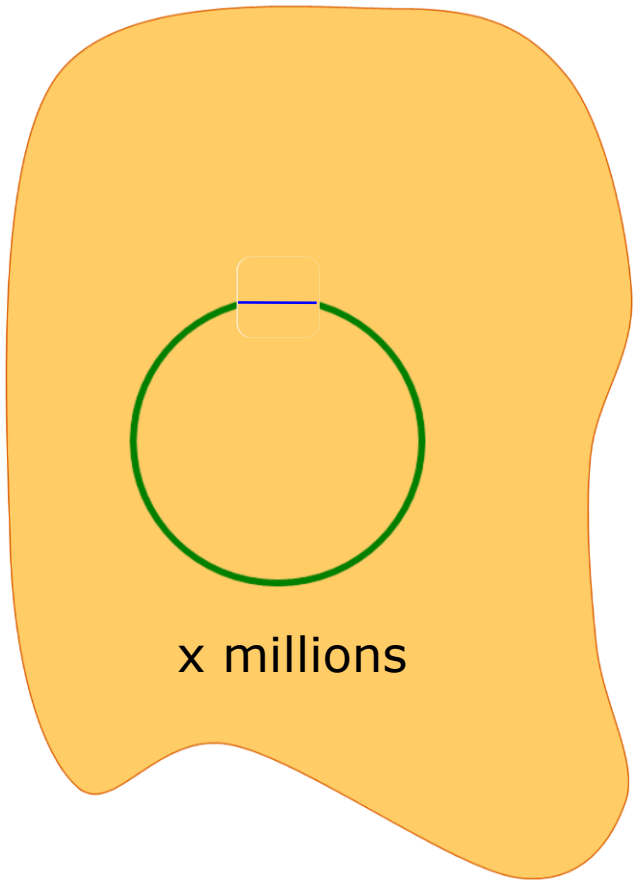


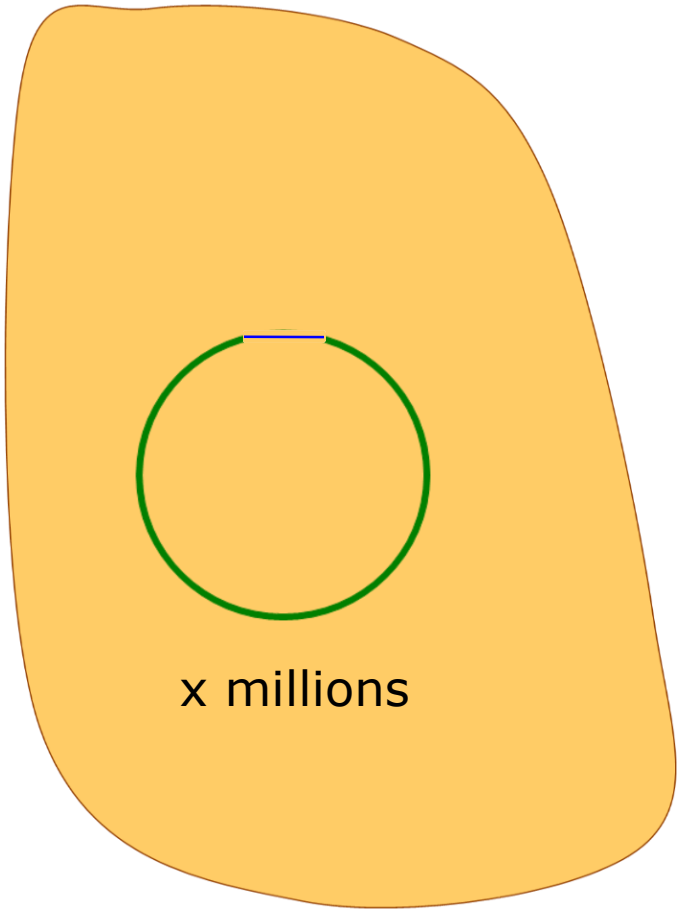
Fragments were cloned:

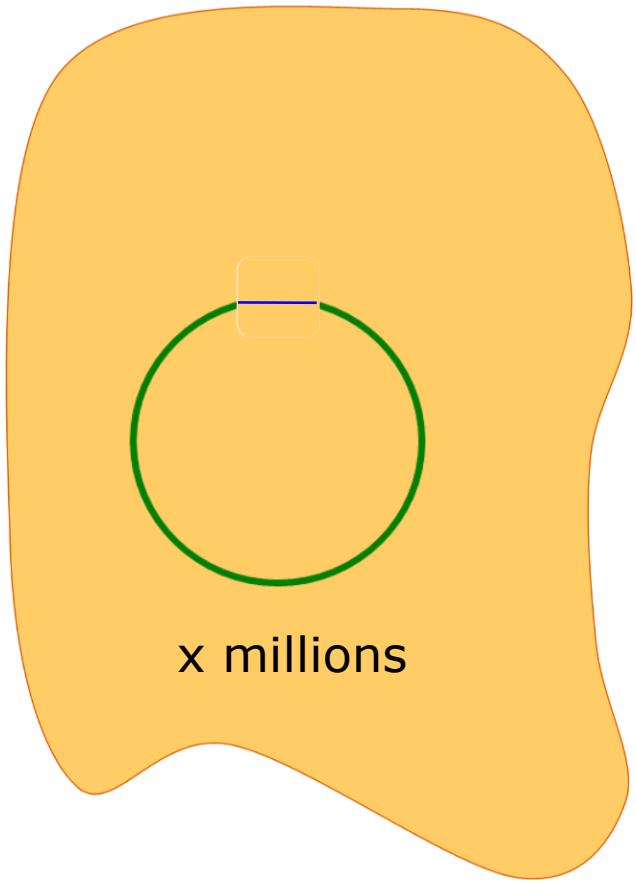





x millions


x millions


x millions


x millions

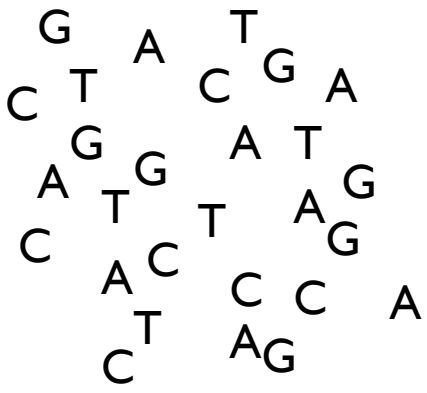
Sanger Sequencing Reactions

For given template DNA, it's like PCR except:

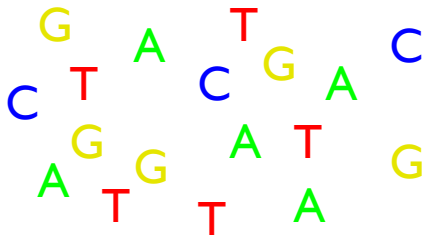
Uses only a single primer and polymerase to make new ssDNA pieces.

Includes regular nucleotides (A, C, G, T) for extension, but also includes dideoxy nucleotides.

Regular Nucleotides

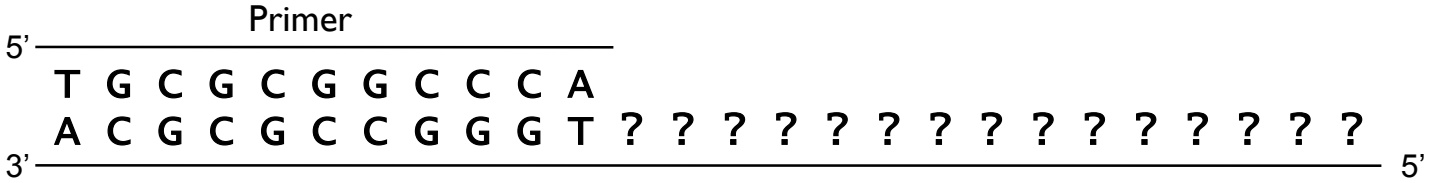


Dideoxy Nucleotides

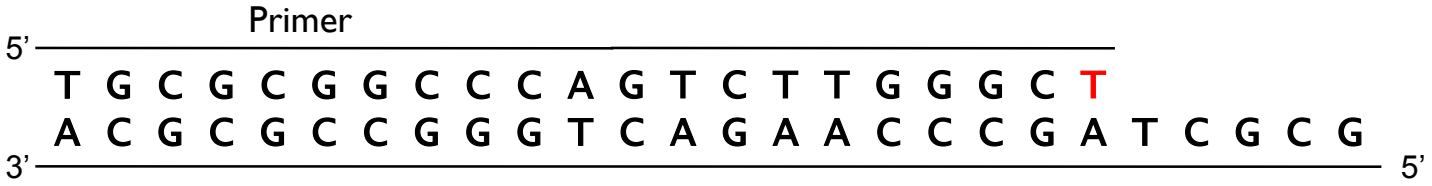


1. Labeled
2. Terminators

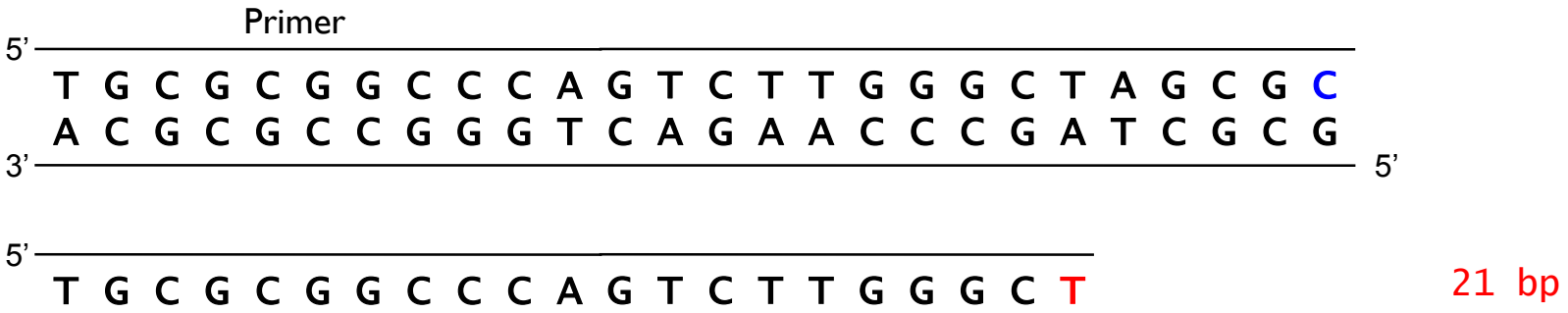
Sanger Sequencing



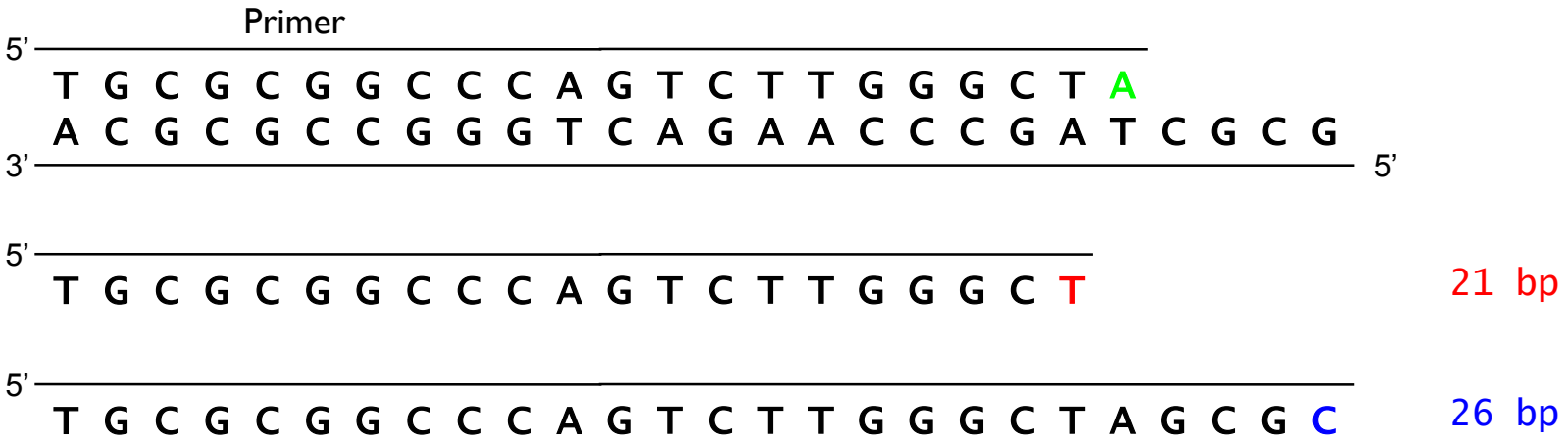
Sanger Sequencing



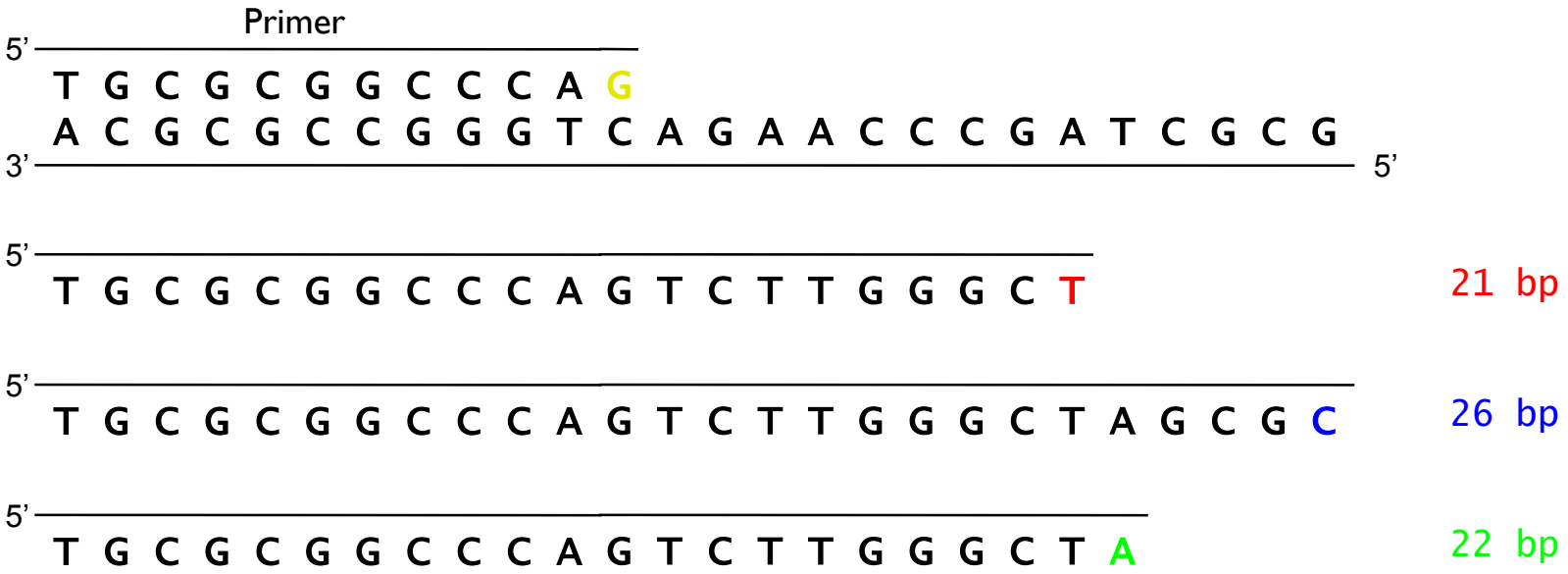
Sanger Sequencing



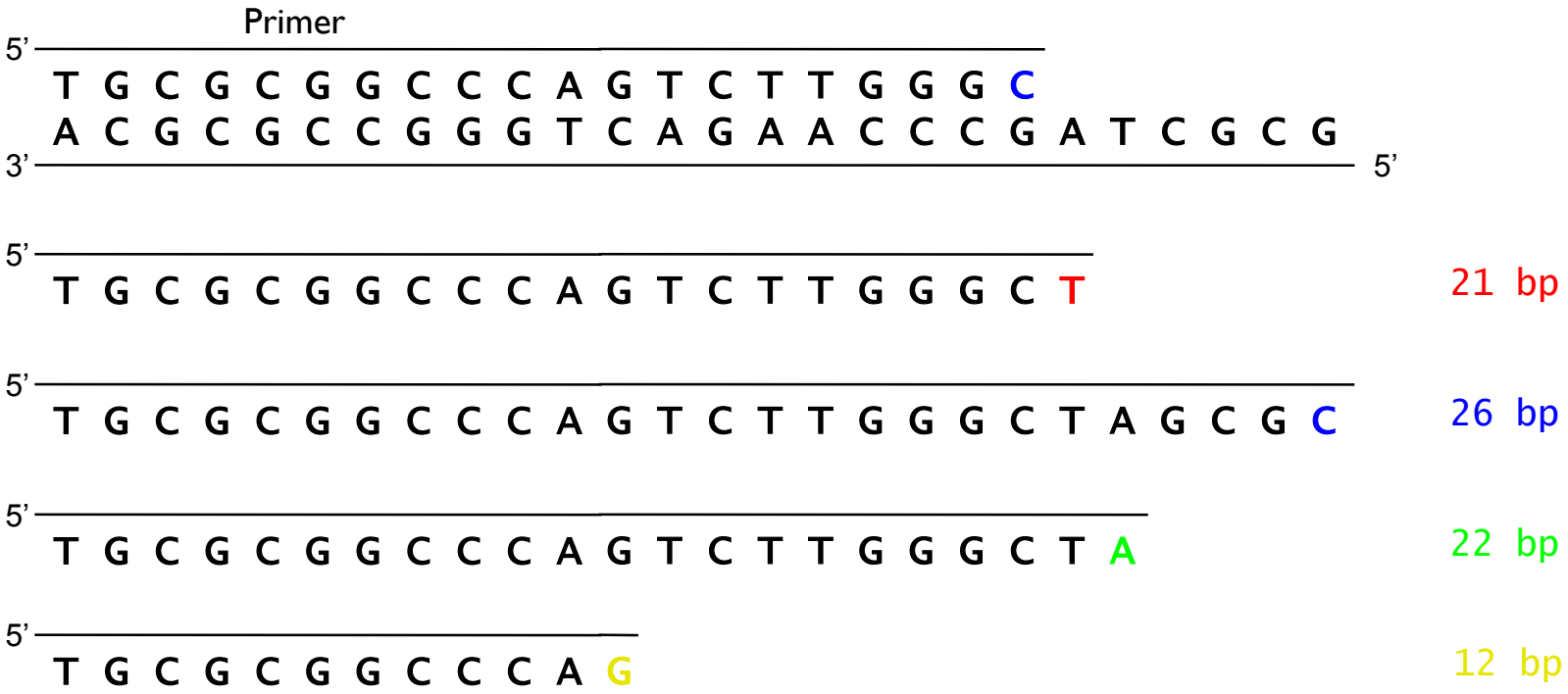
Sanger Sequencing



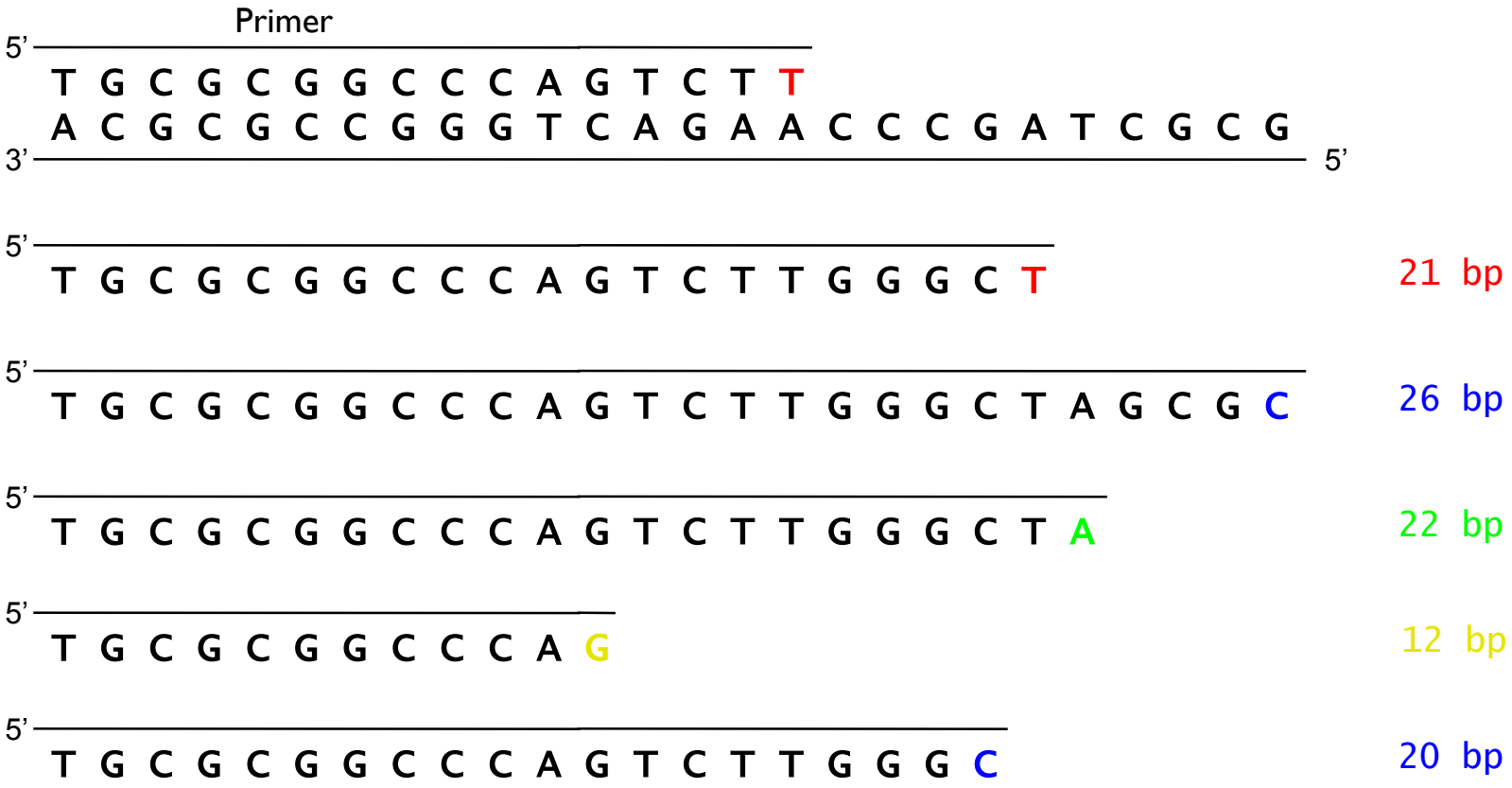
Sanger Sequencing



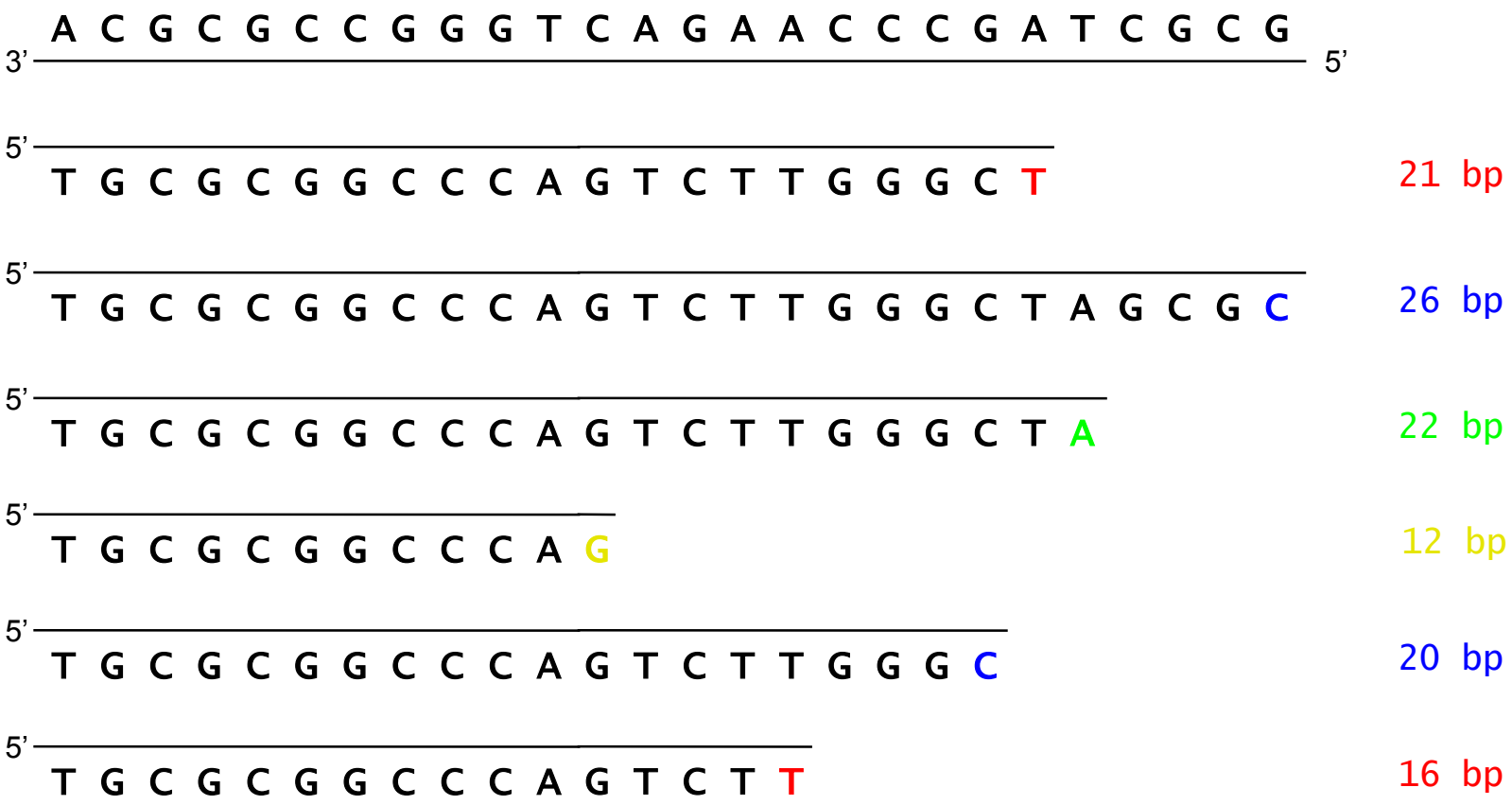
Sanger Sequencing



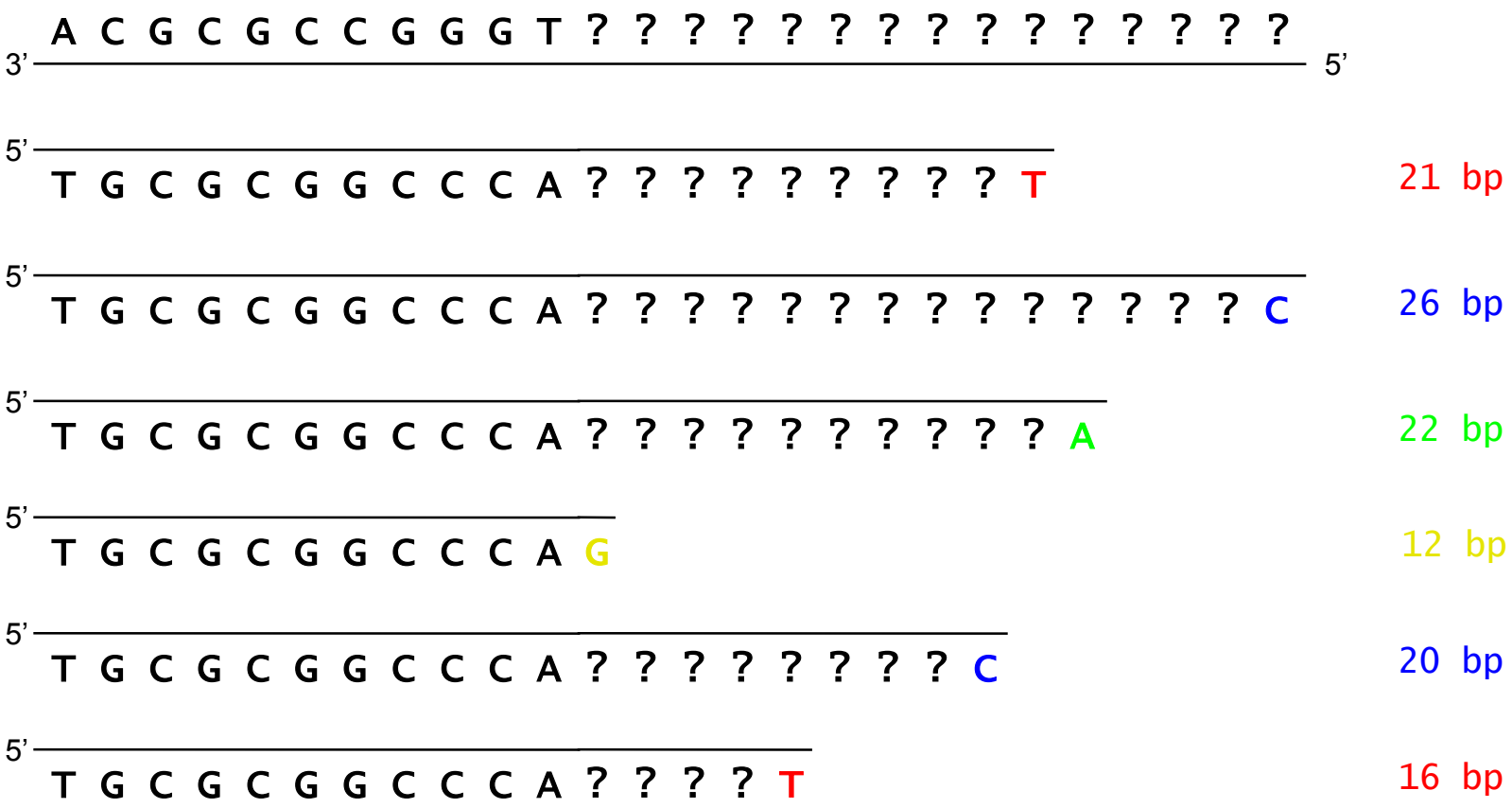
Sanger Sequencing



Sanger Sequencing



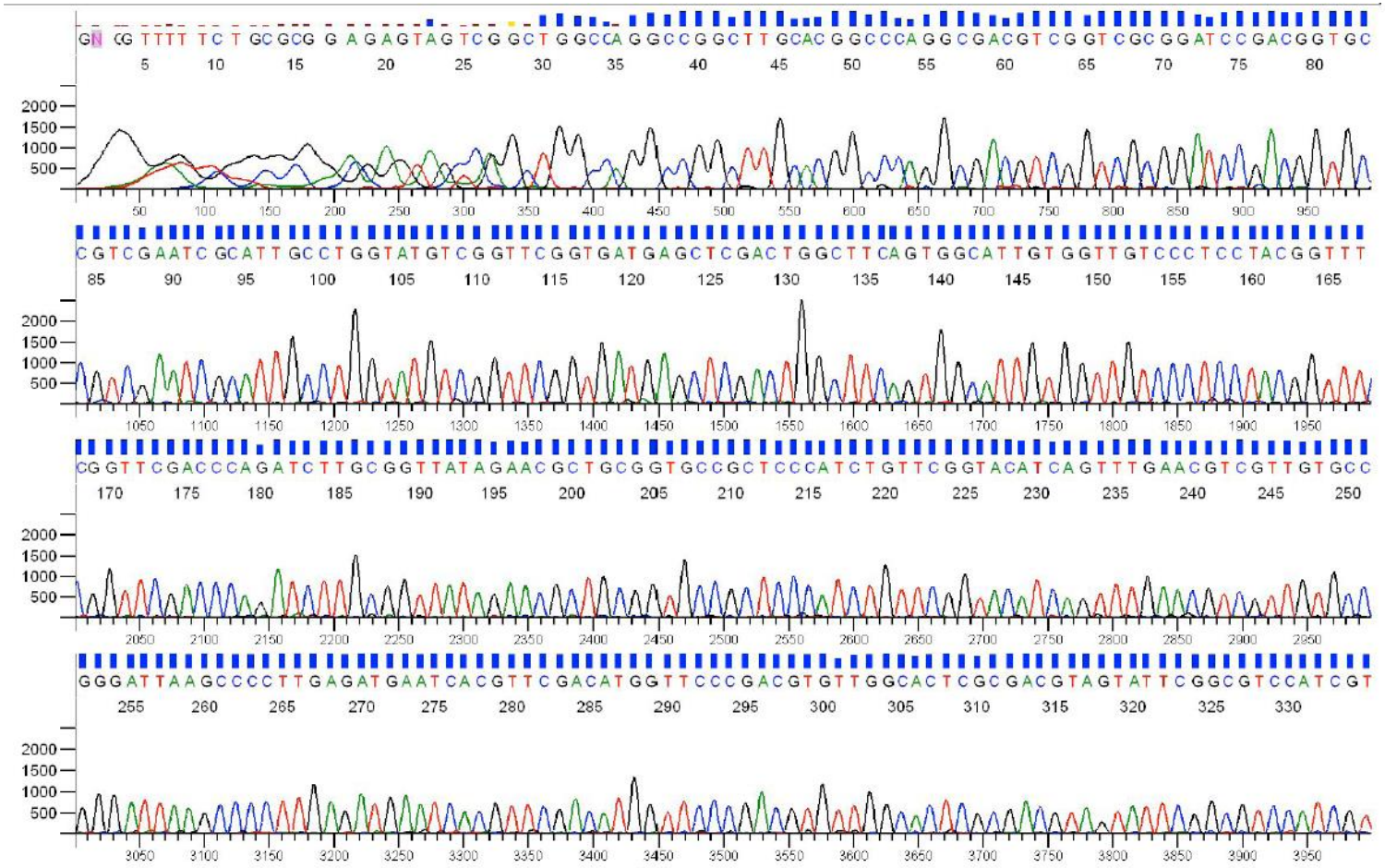
Sanger Sequencing



Sanger Sequencing

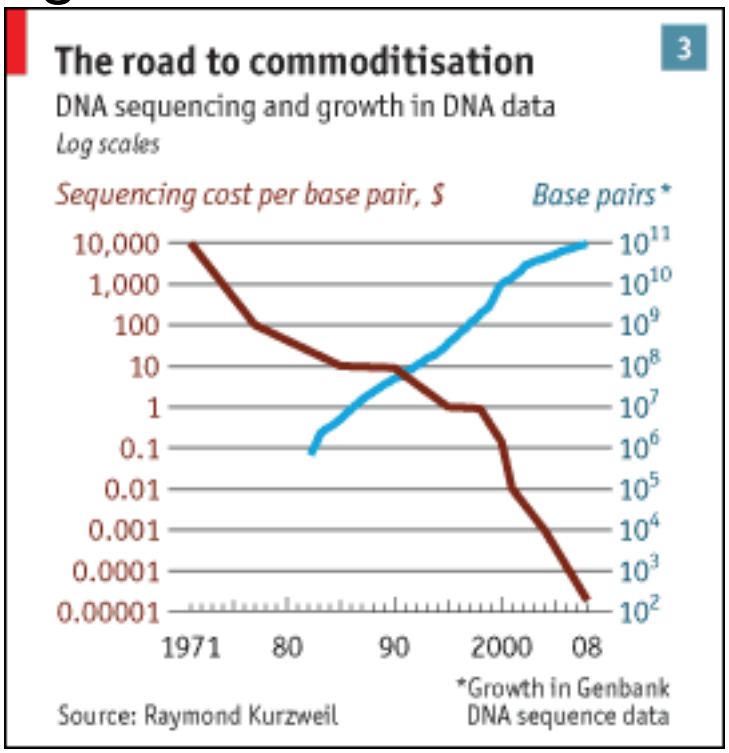
Sanger Sequencing Output

Each sequencing reaction gives us a **chromatogram**, usually ~600-1000 bp:



Sanger Throughput Limitations

- Must have 1 colony picked for every 2 reactions
- Must do 1 DNA prep for every 2 reactions
- Must have 1 PCR tube for each reaction
- Must have 1 gel lane for each reaction

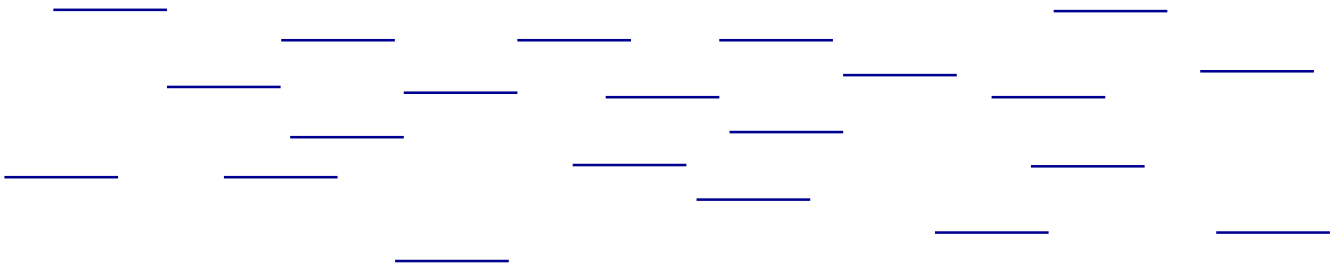


from *The Economist*

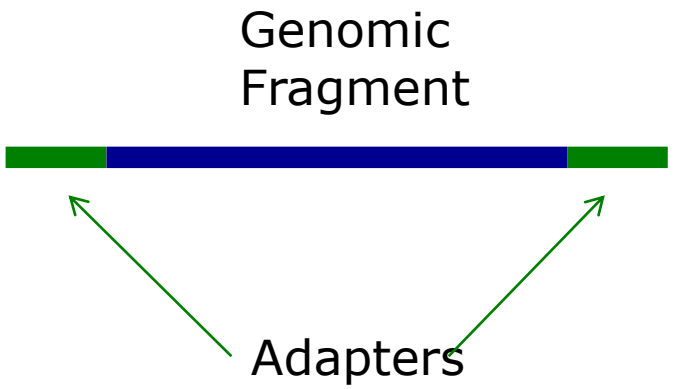
Overview

- Prologue: Assembly
- The Past: Sanger
- The Present: **Next-Gen (454, Illumina, …)**
- The Future: ? (Nanopore, MinION, Single-molecule)

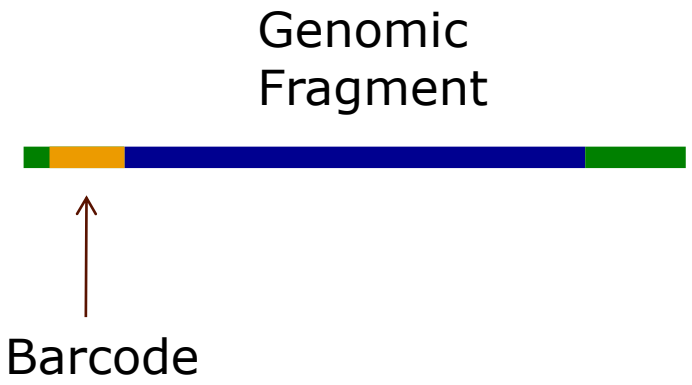
Shotgun sequencing by Ion Torrent Personal Genome Machine and 454



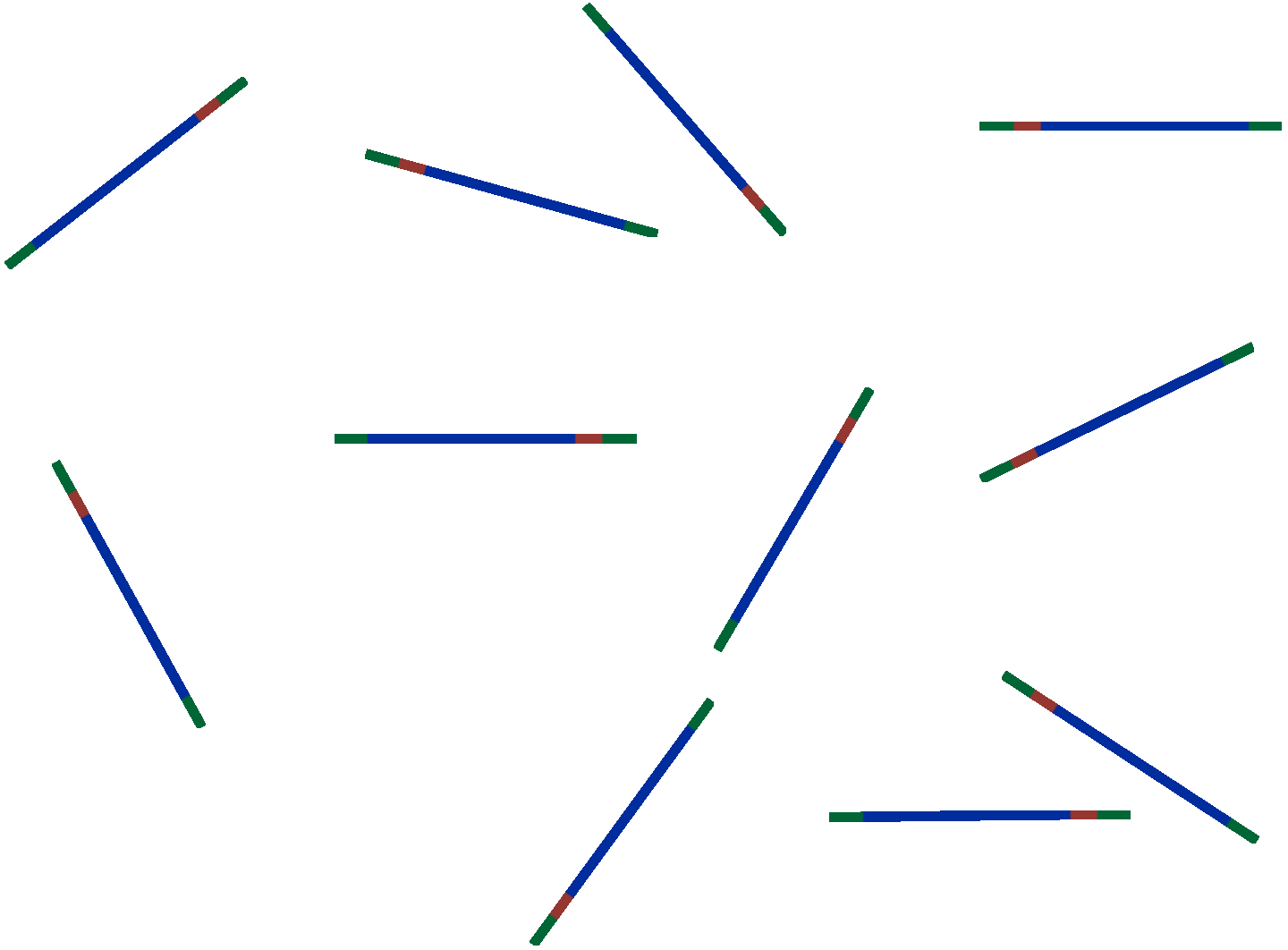
Shotgun sequencing by PGM/454



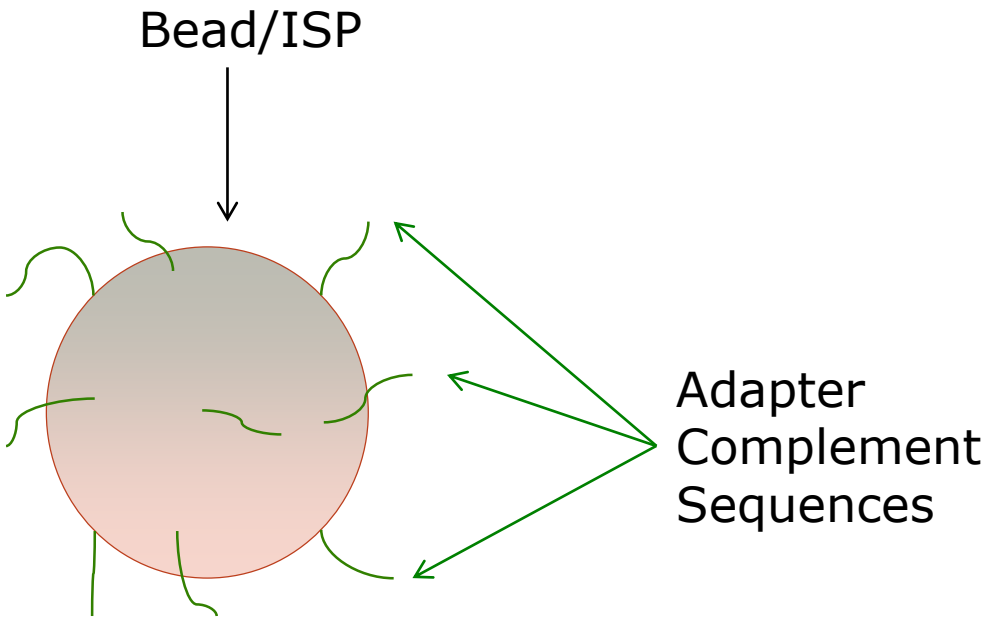
Shotgun sequencing by PGM/454



Shotgun sequencing by PGM/454

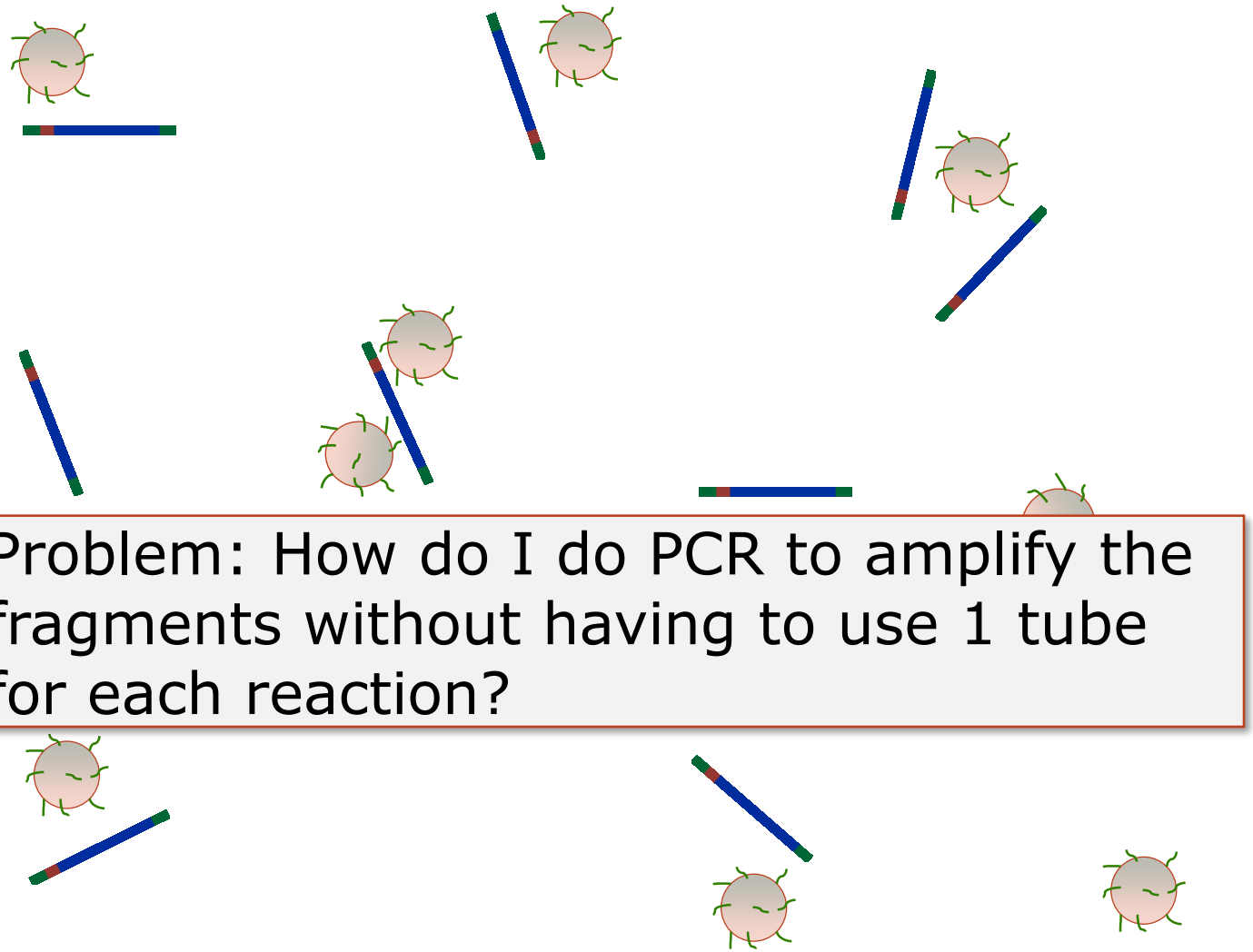


Shotgun sequencing by PGM/454



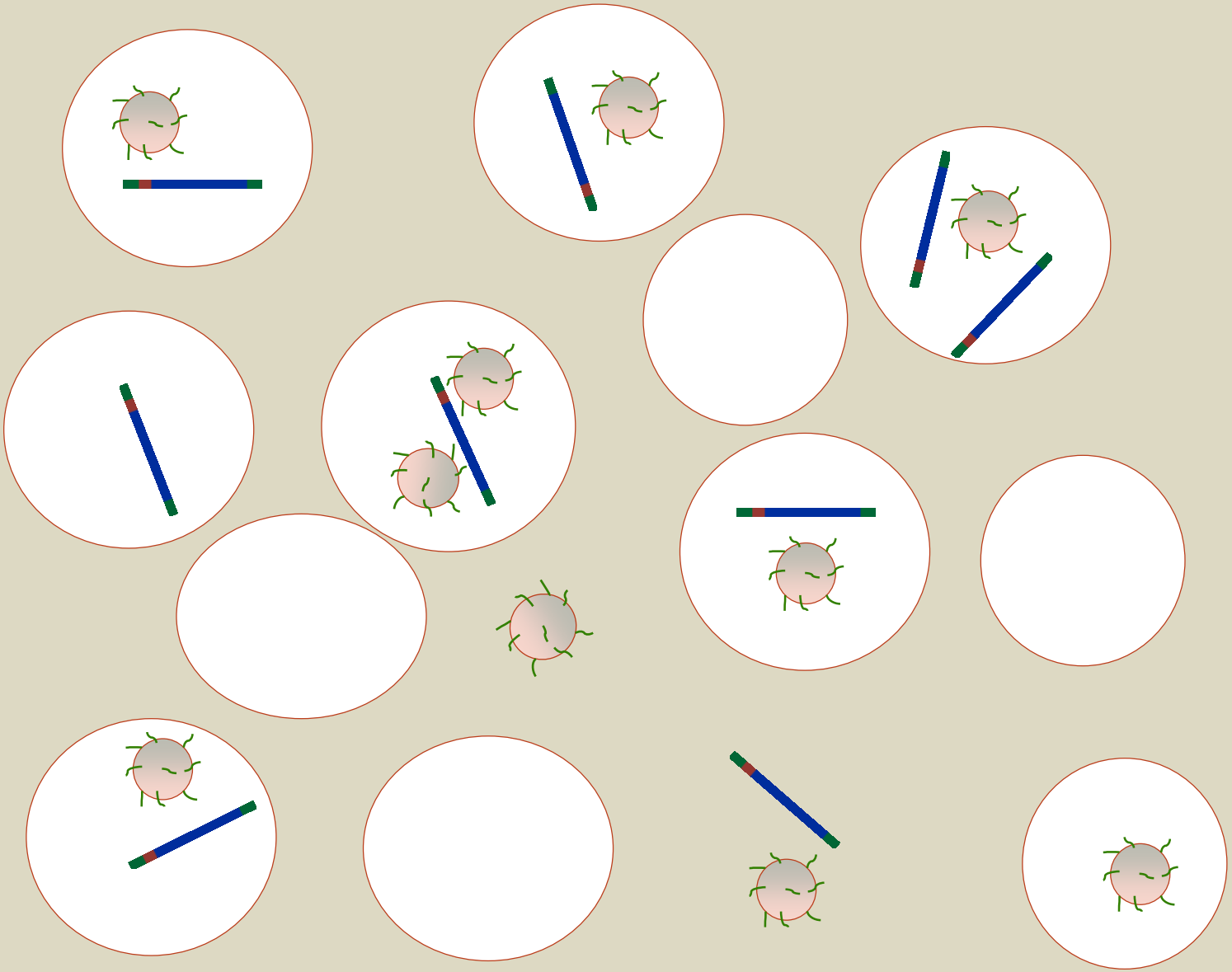
The idea is that each bead should be amplified all over with a SINGLE library fragment.

Shotgun sequencing by PGM/454

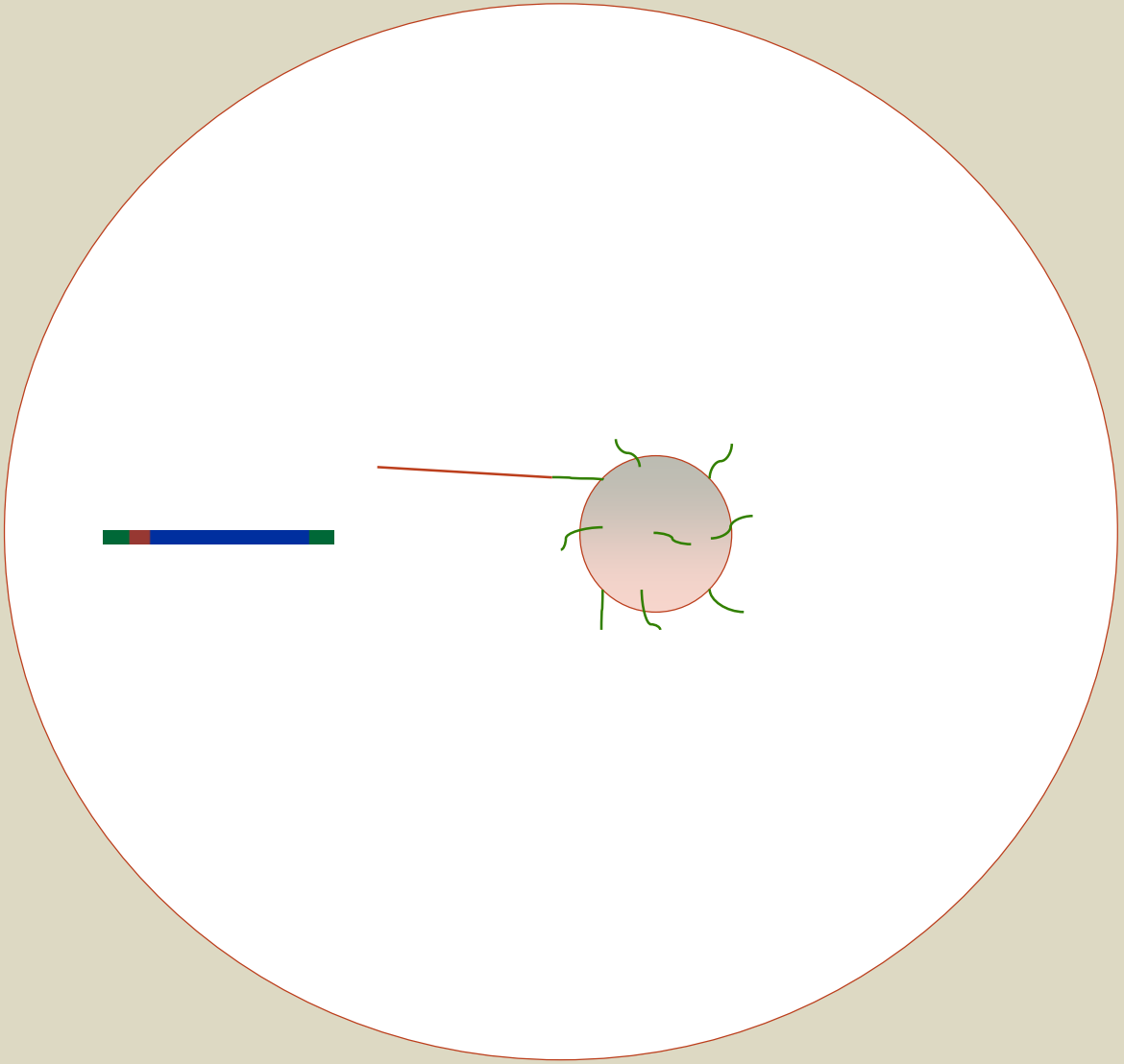


Problem: How do I do PCR to amplify the fragments without having to use 1 tube for each reaction?

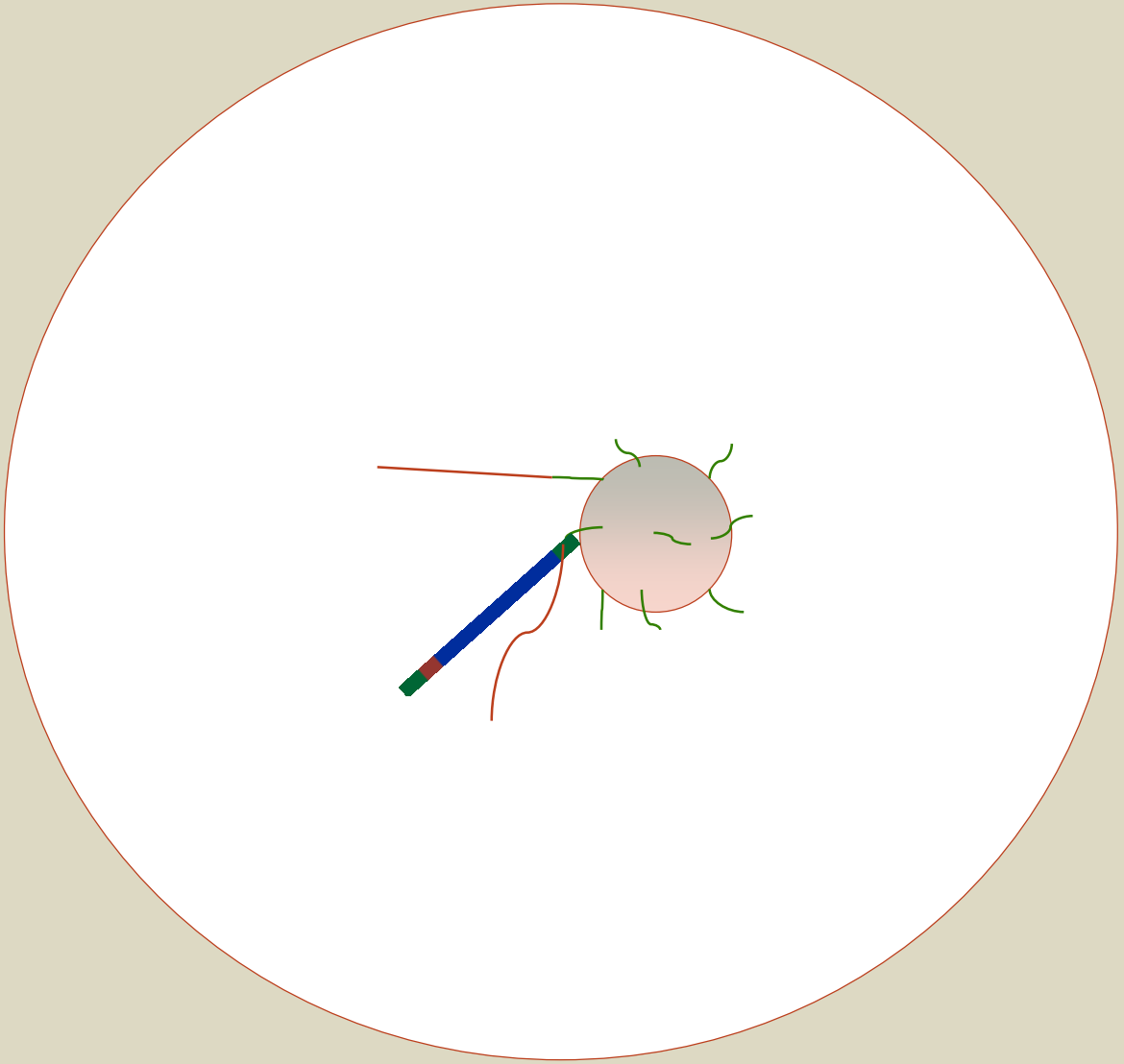
Shotgun sequencing by PGM/454



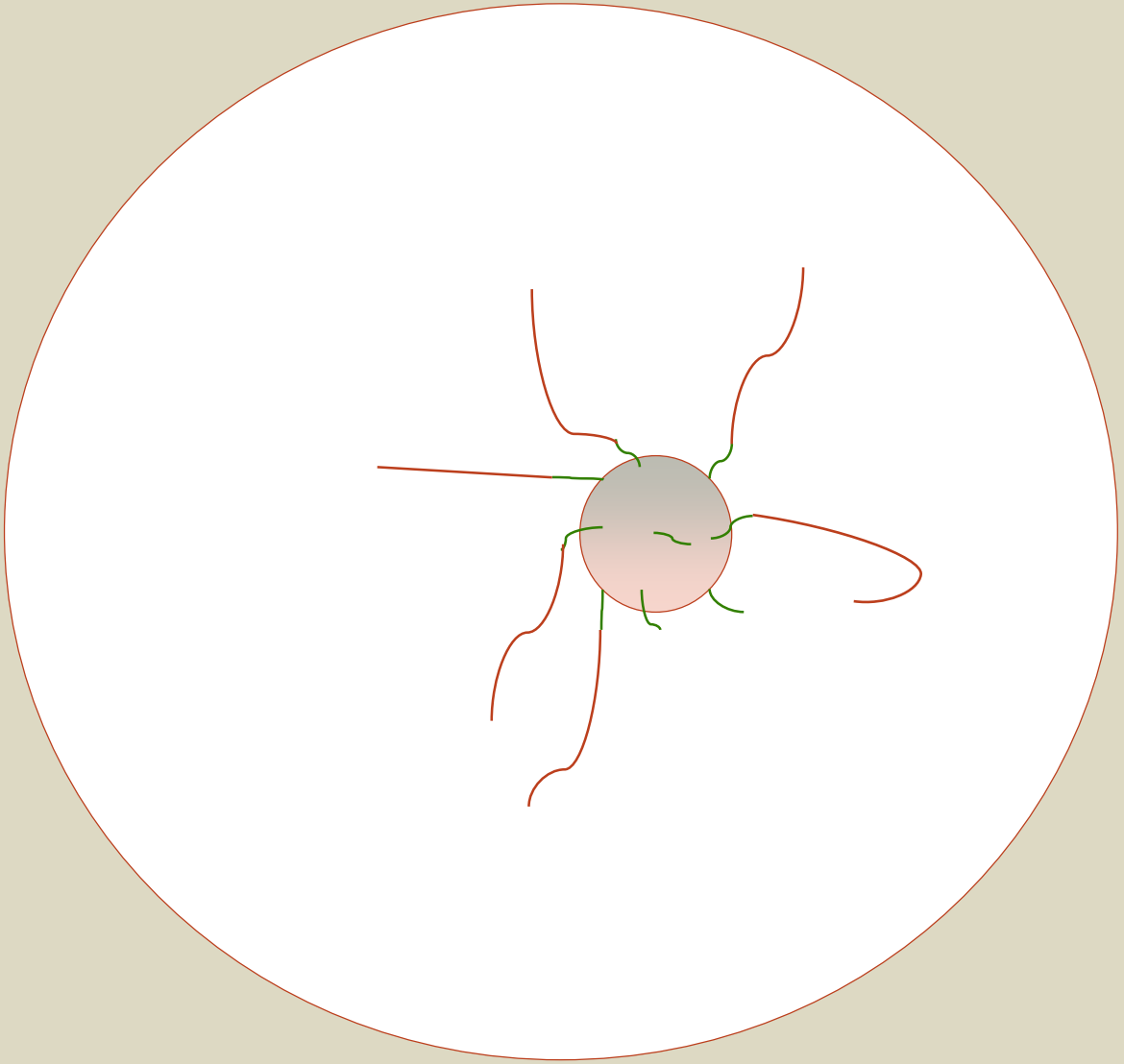
Shotgun sequencing by PGM/454



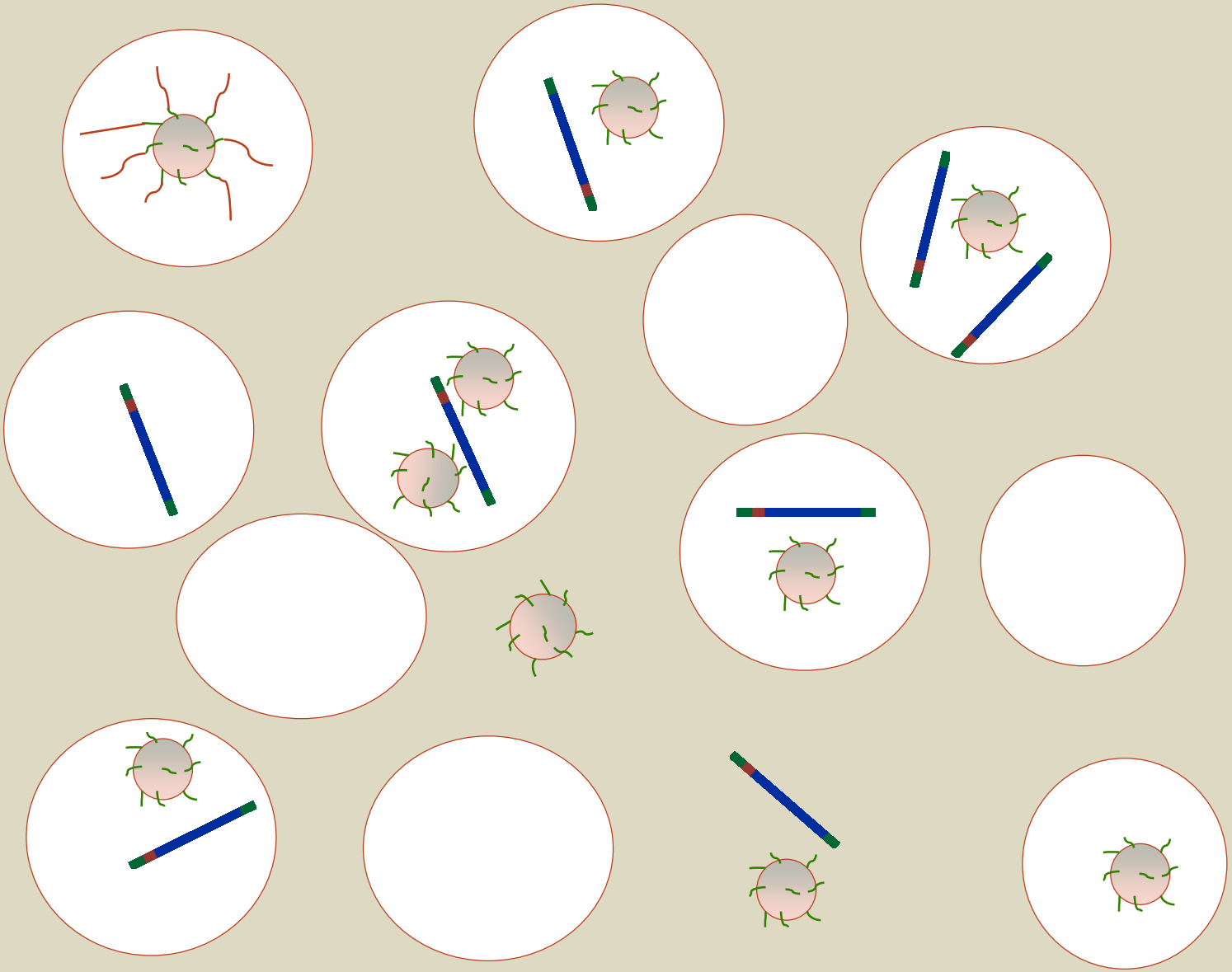
Shotgun sequencing by PGM/454



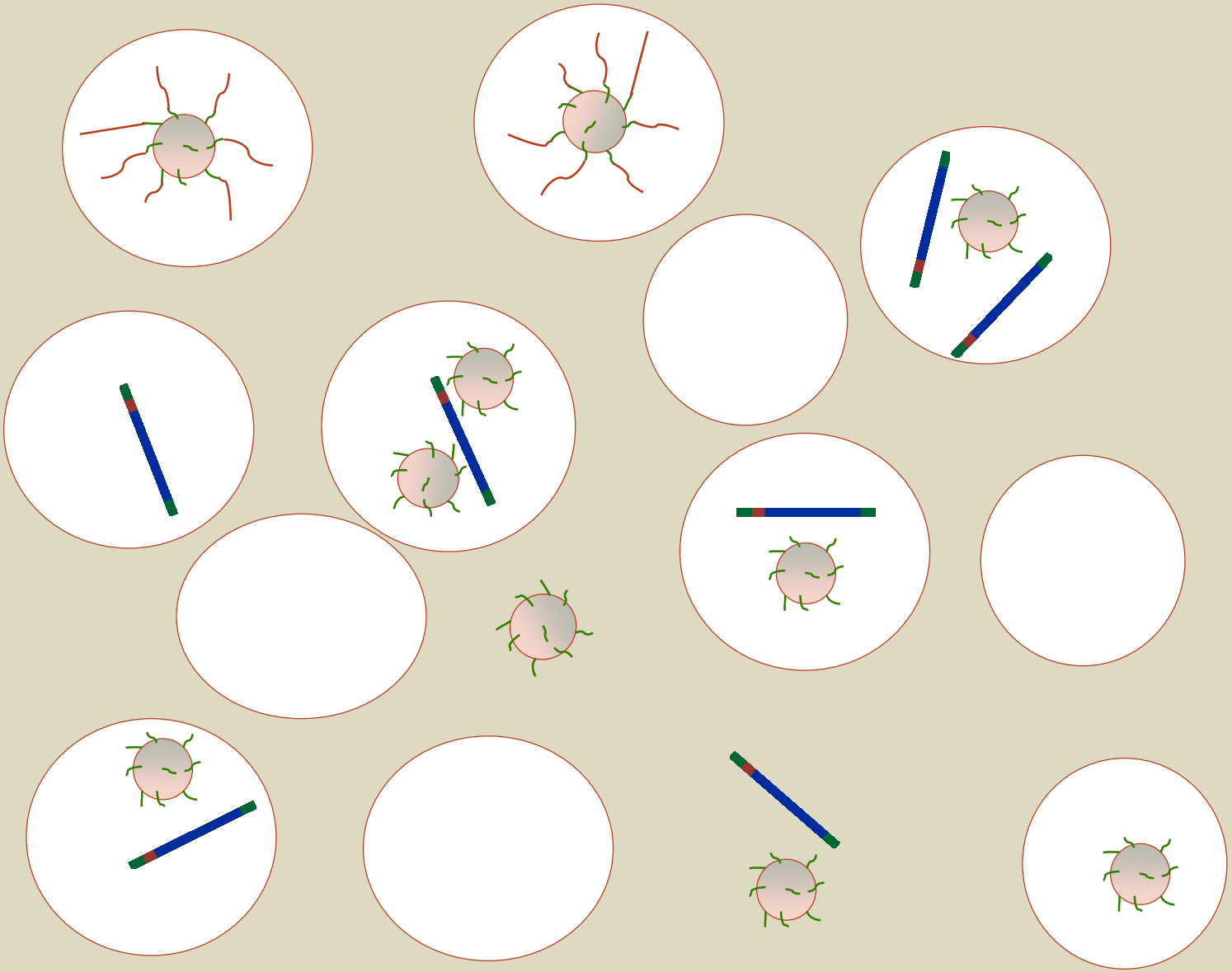
Shotgun sequencing by PGM/454



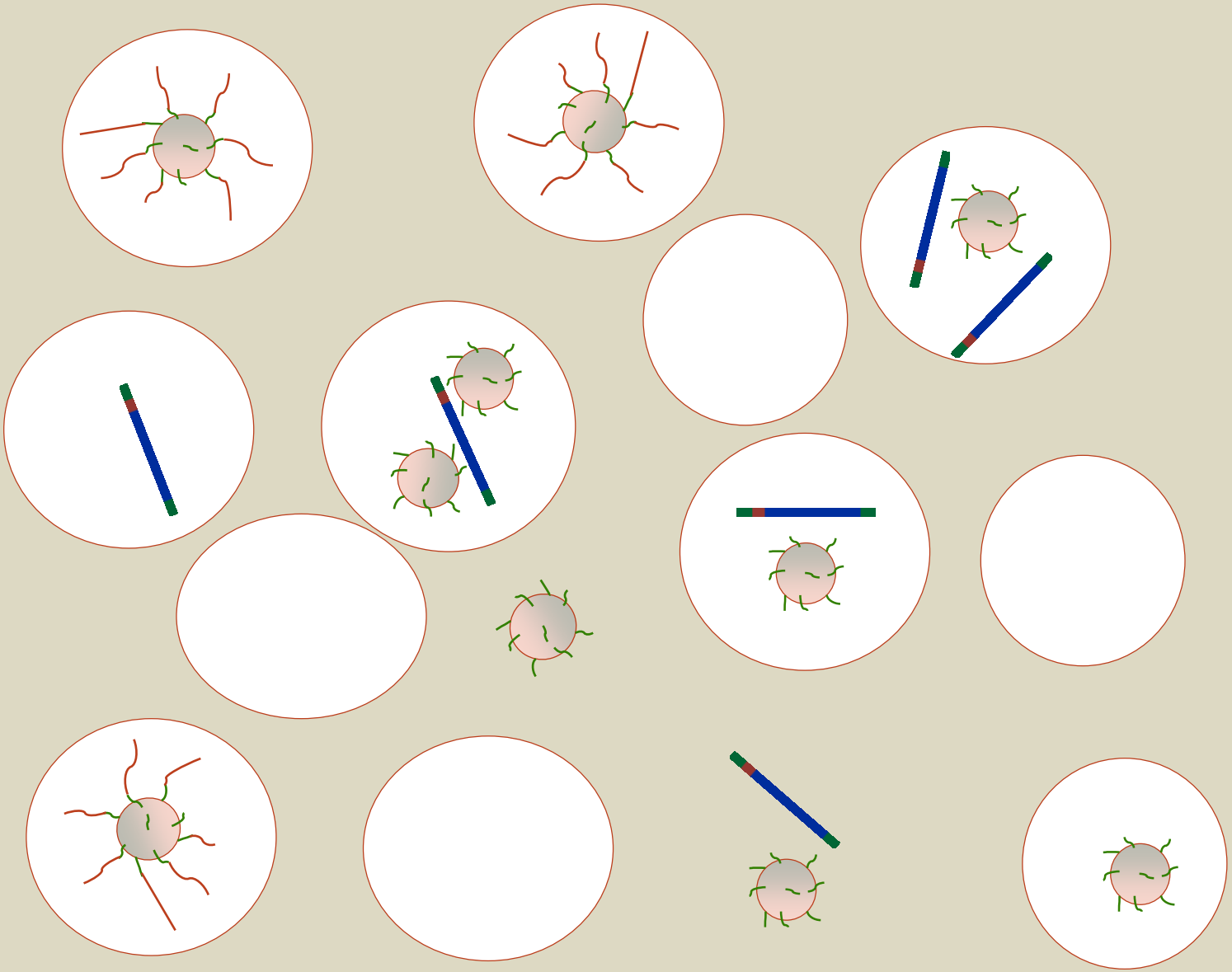
Shotgun sequencing by PGM/454



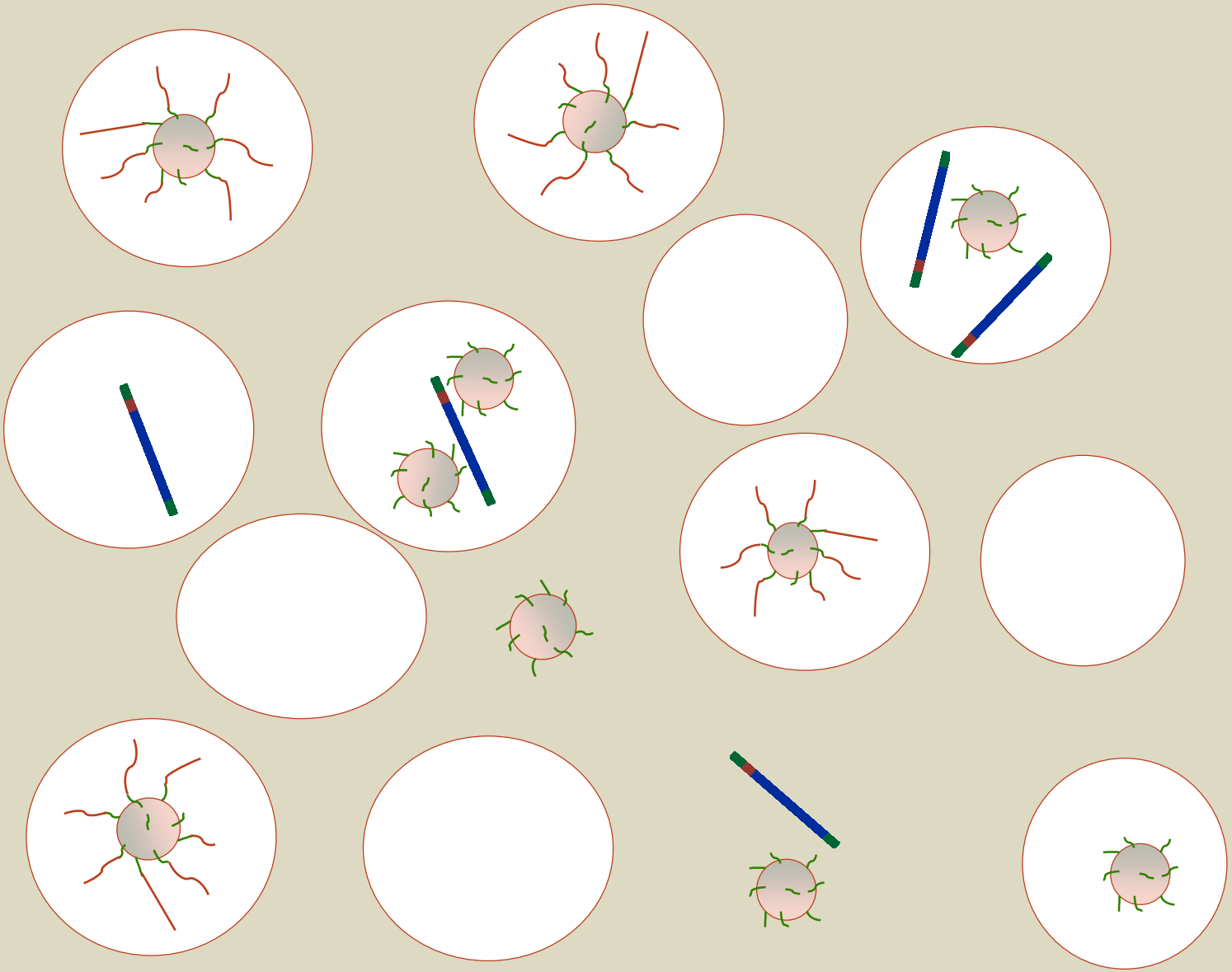
Shotgun sequencing by PGM/454



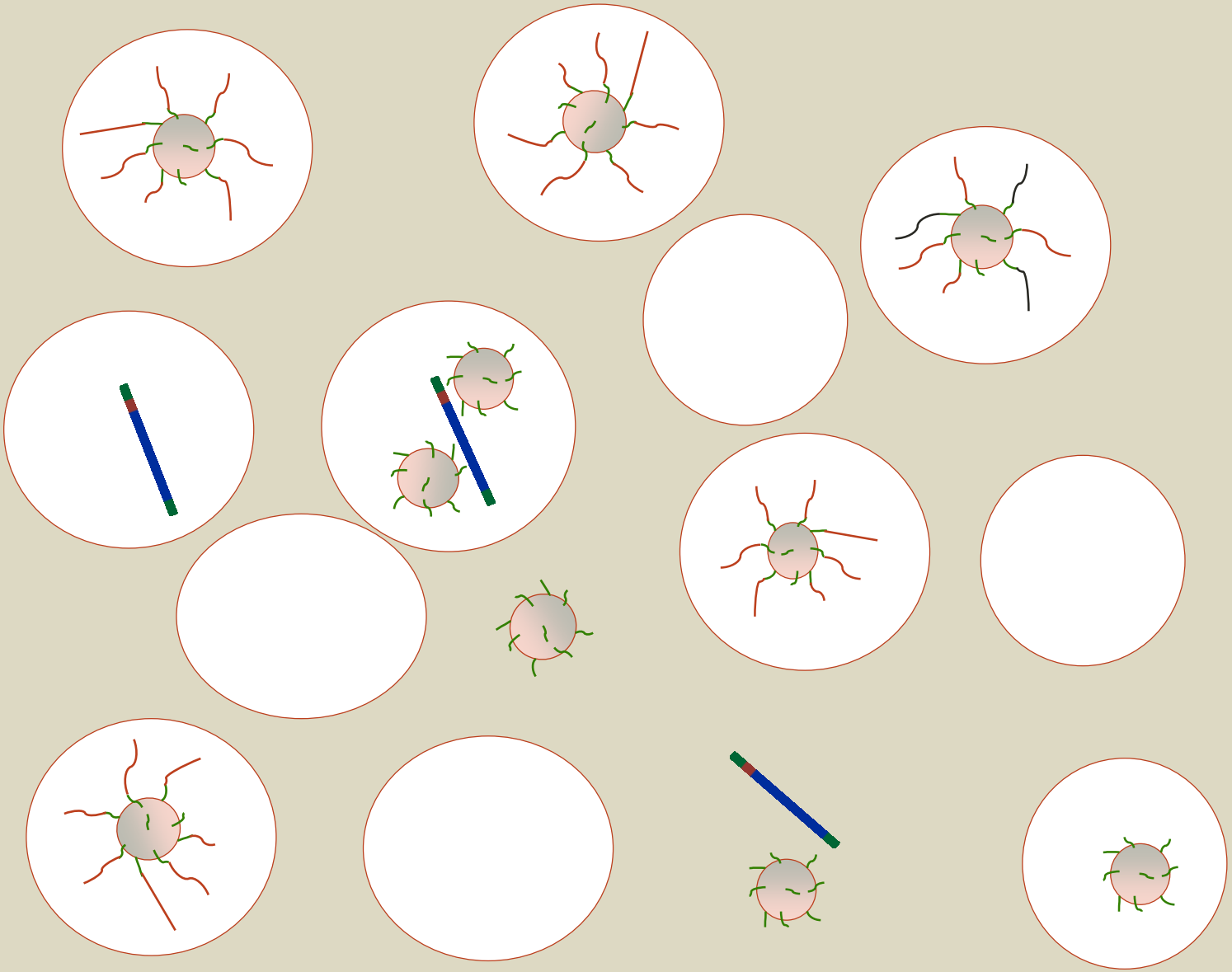
Shotgun sequencing by PGM/454



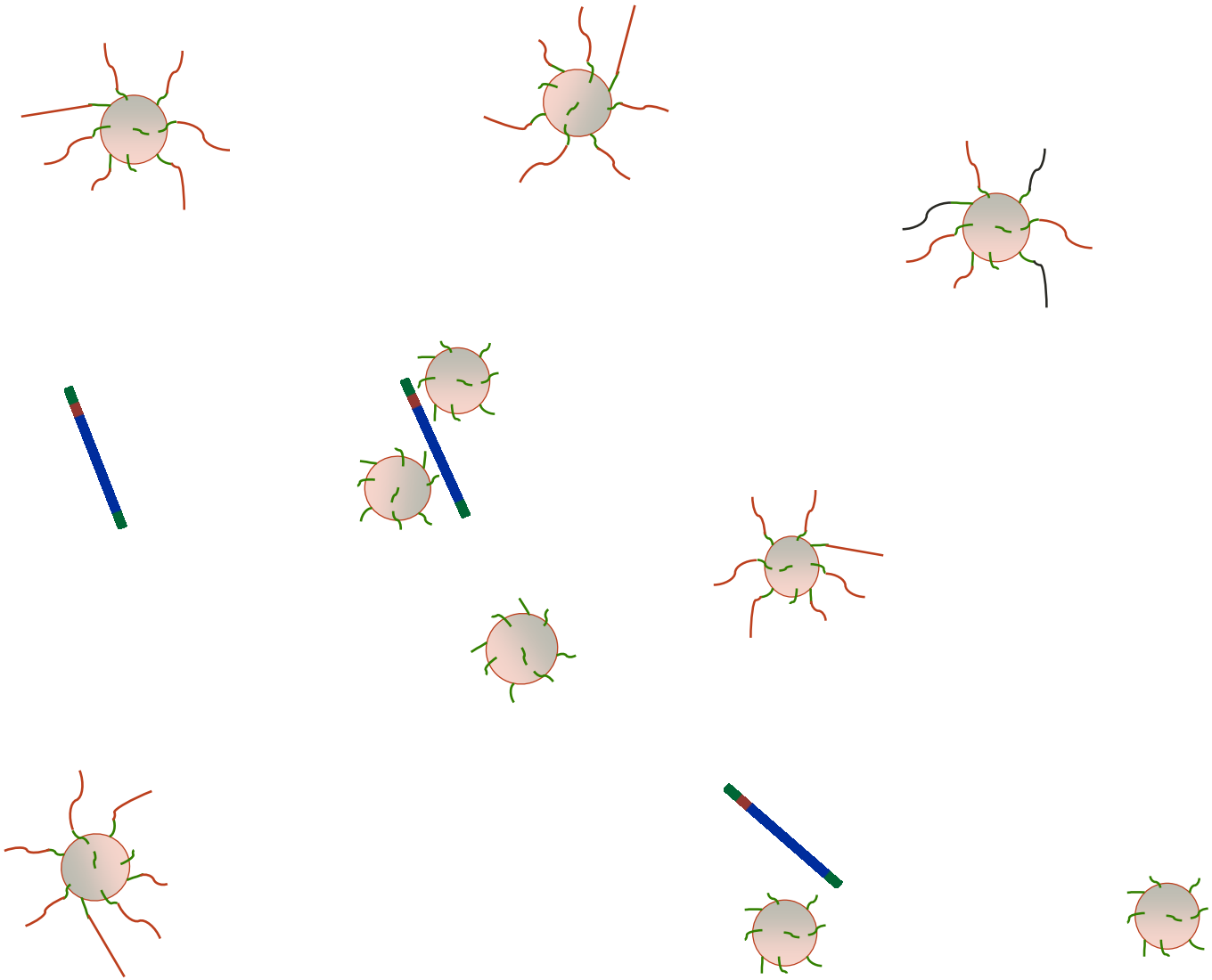
Shotgun sequencing by PGM/454



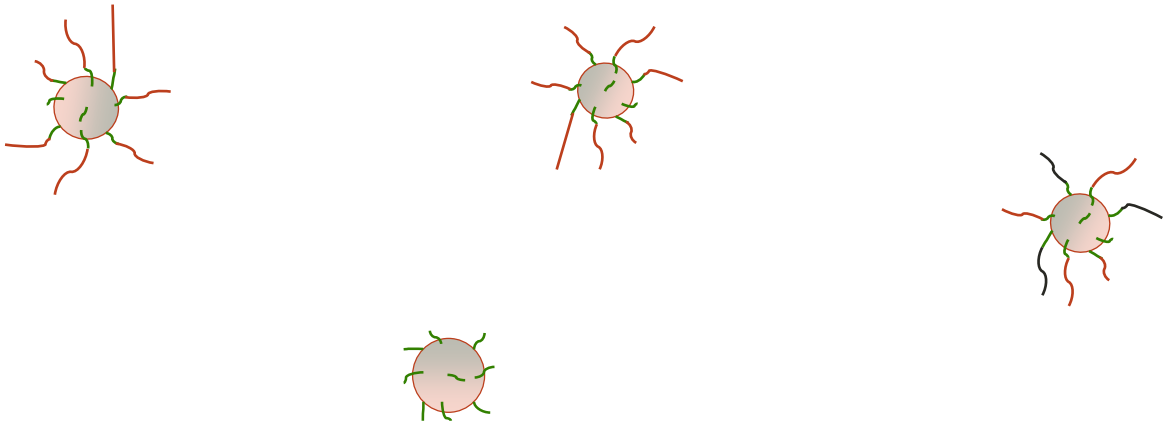
Shotgun sequencing by PGM/454

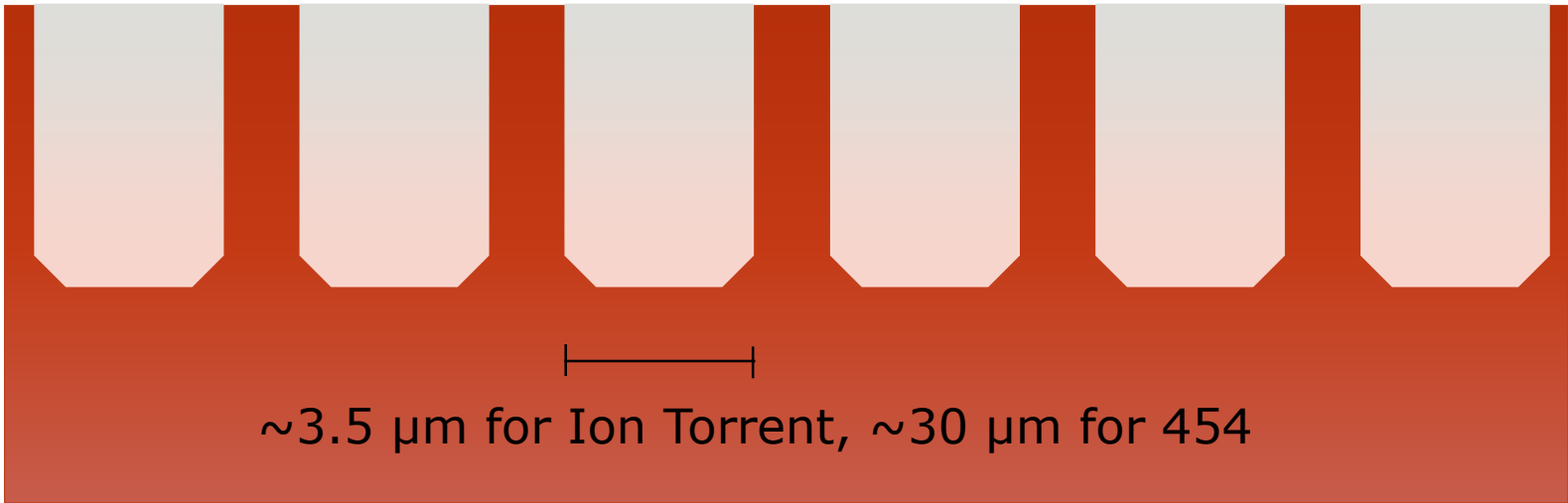


Shotgun sequencing by PGM/454



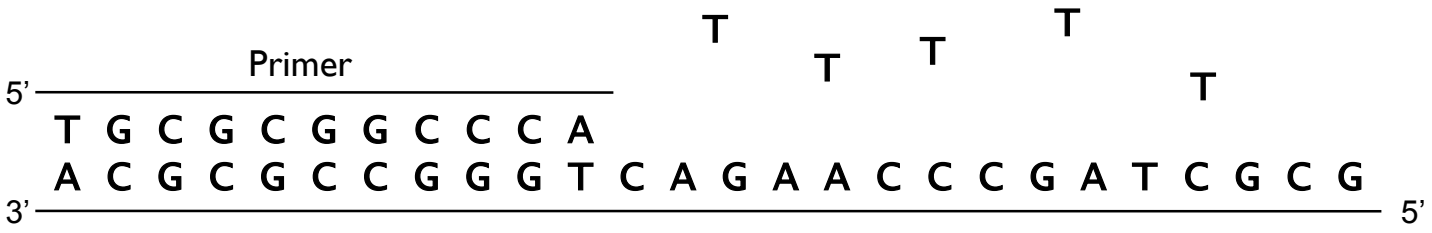
Shotgun sequencing by PGM/454



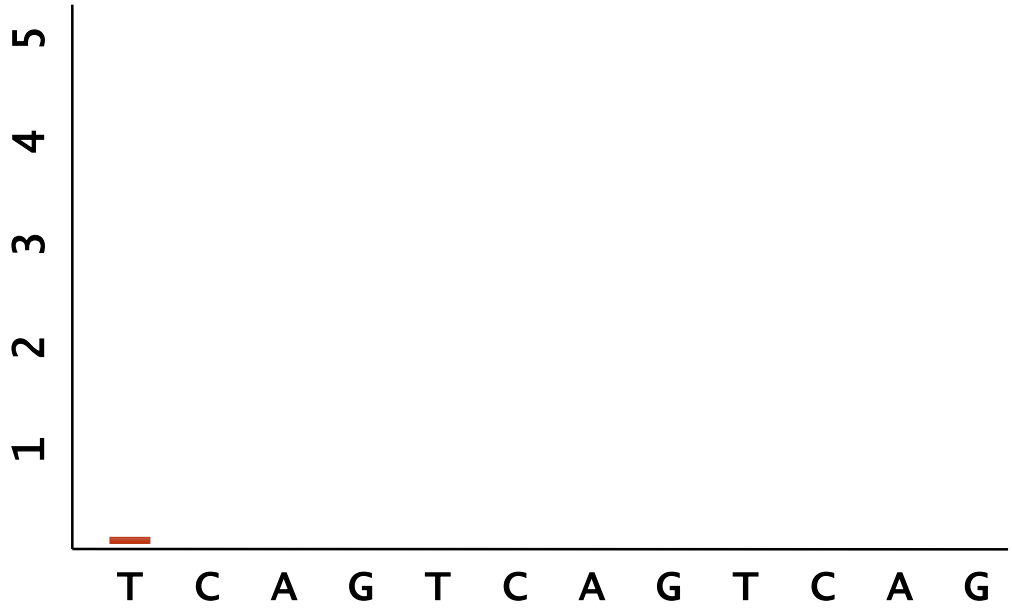


Shotgun sequencing by PGM/454

Only give polymerase one nucleotide at a time:

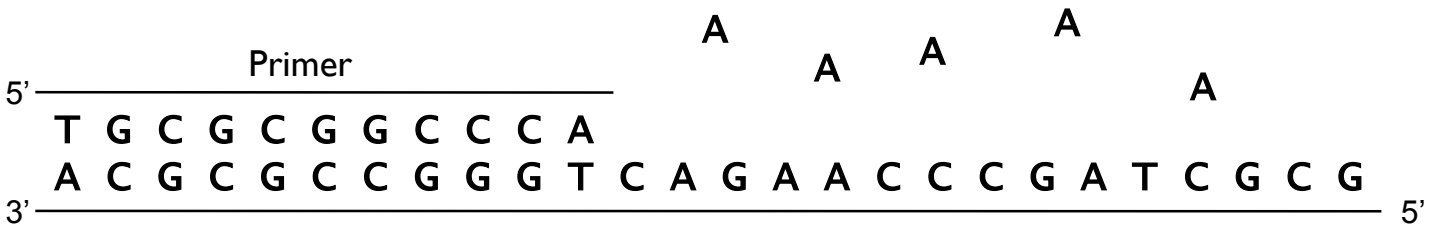


If that nucleotide is incorporated, enzymes turn by-products into light:

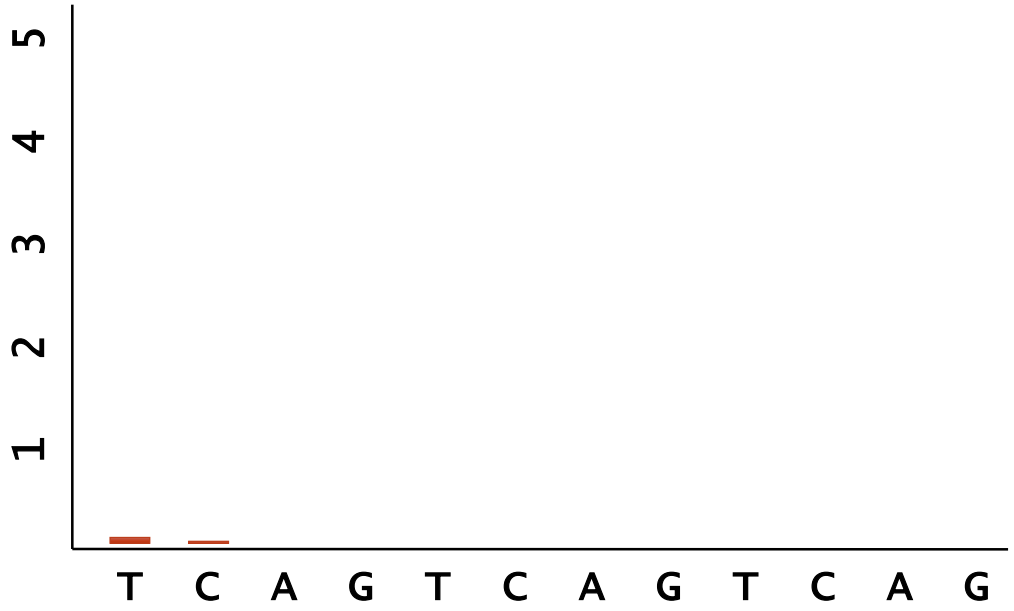


Shotgun sequencing by PGM/454

Only give polymerase one nucleotide at a time:

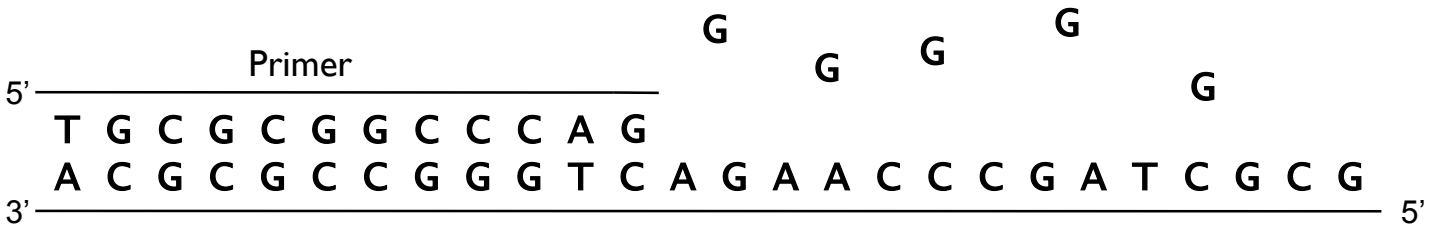


If that nucleotide is incorporated, enzymes turn by-products into light:

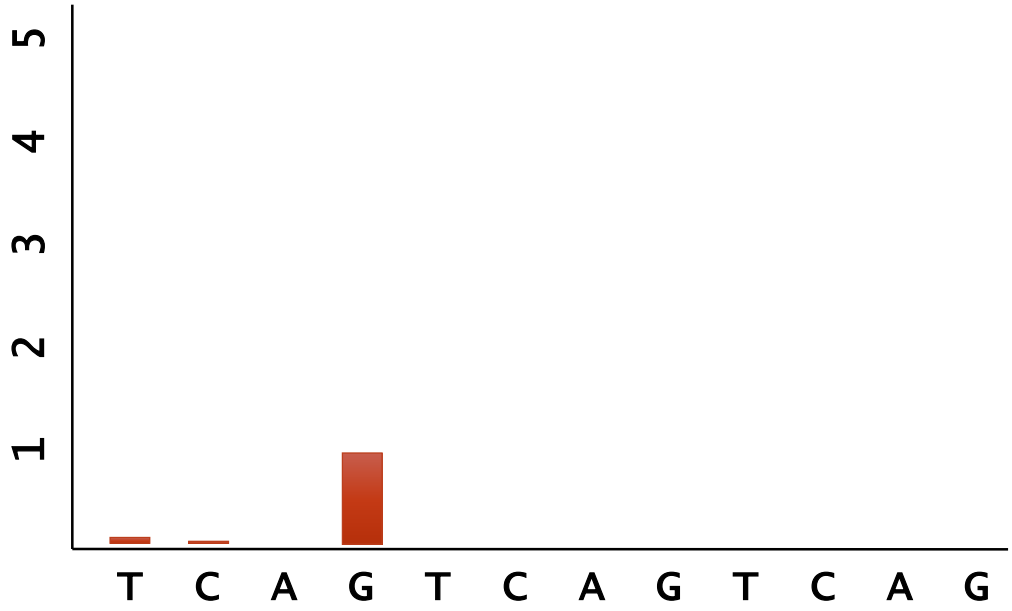


Shotgun sequencing by PGM/454

Only give polymerase one nucleotide at a time:

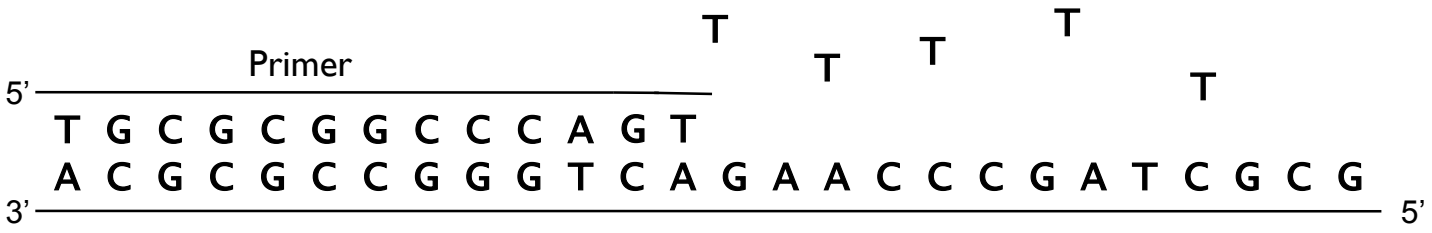


If that nucleotide is incorporated, enzymes turn by-products into light:

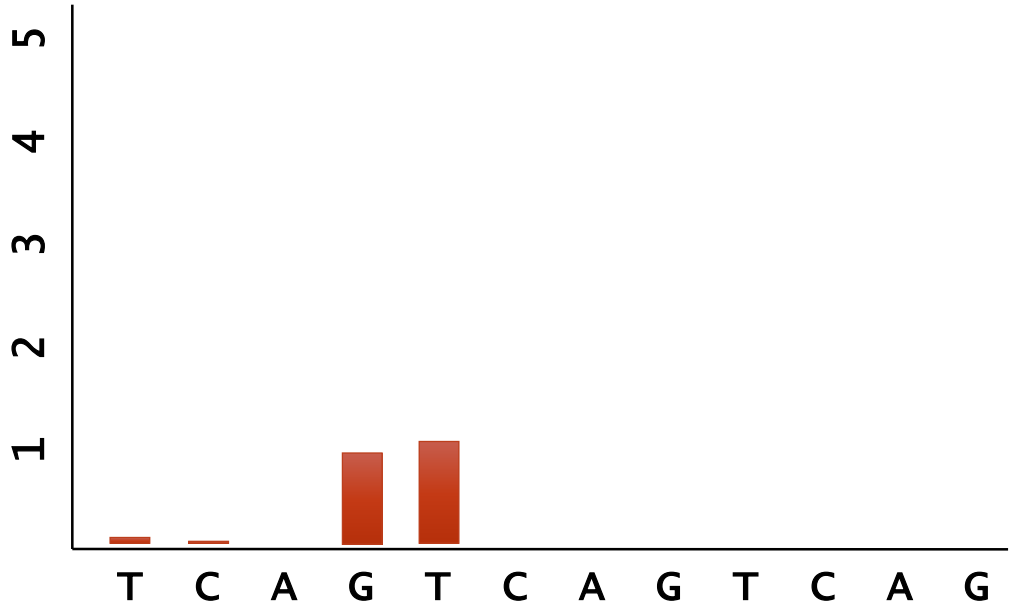


Shotgun sequencing by PGM/454

Only give polymerase one nucleotide at a time:

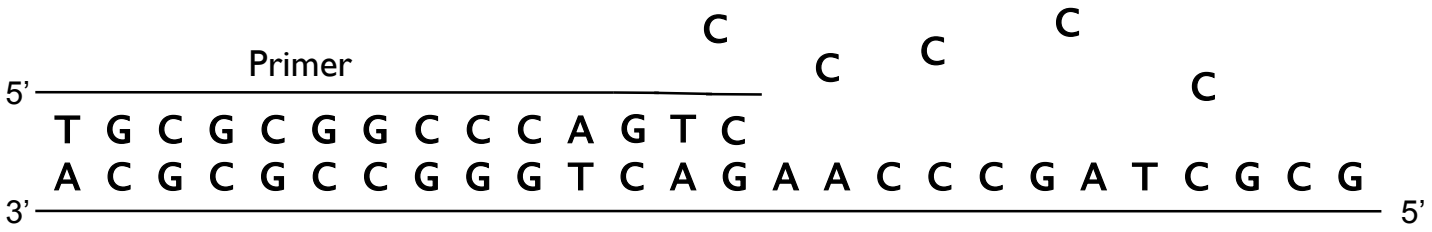


If that nucleotide is incorporated, enzymes turn by-products into light:

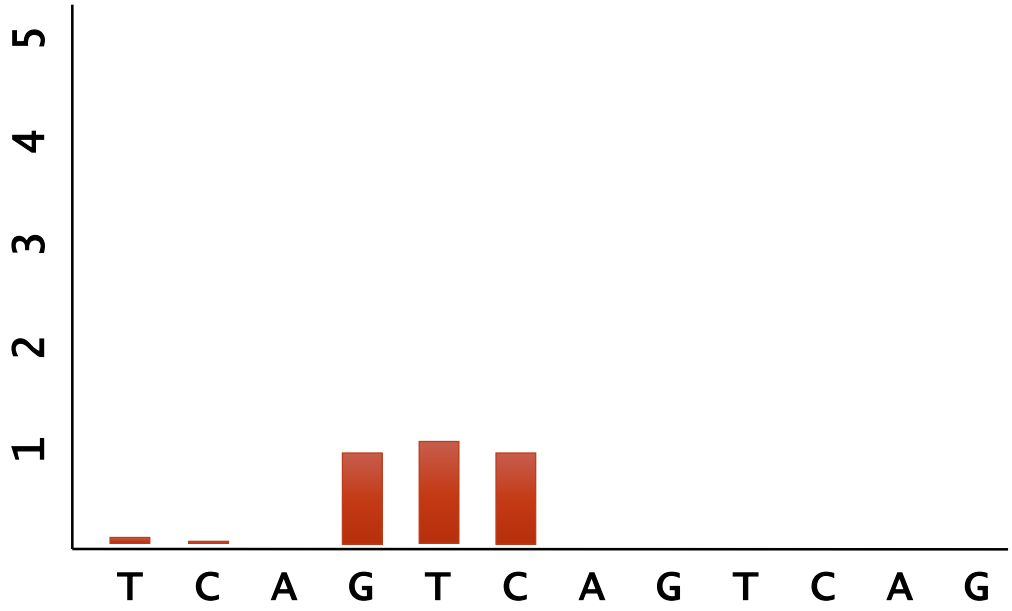


Shotgun sequencing by PGM/454

Only give polymerase one nucleotide at a time:

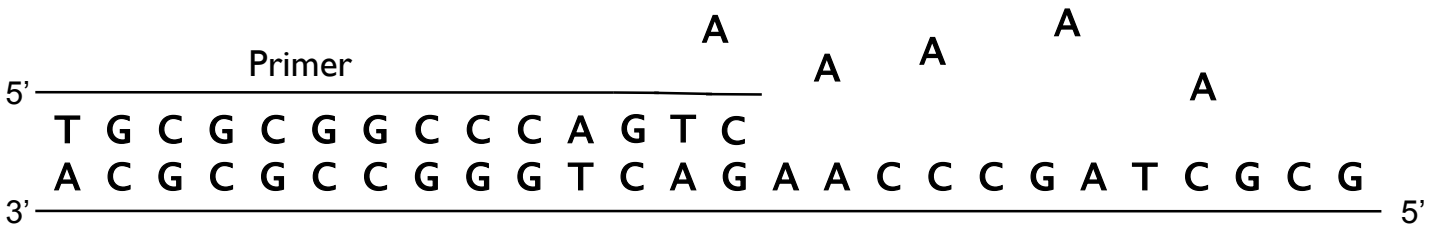


If that nucleotide is incorporated, enzymes turn by-products into light:

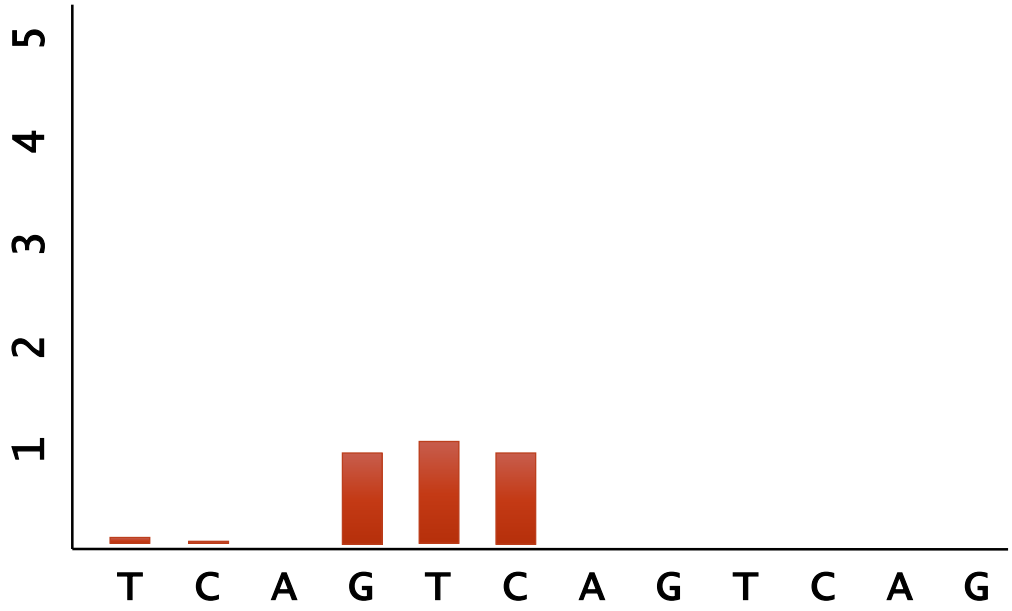


Shotgun sequencing by PGM/454

Only give polymerase one nucleotide at a time:

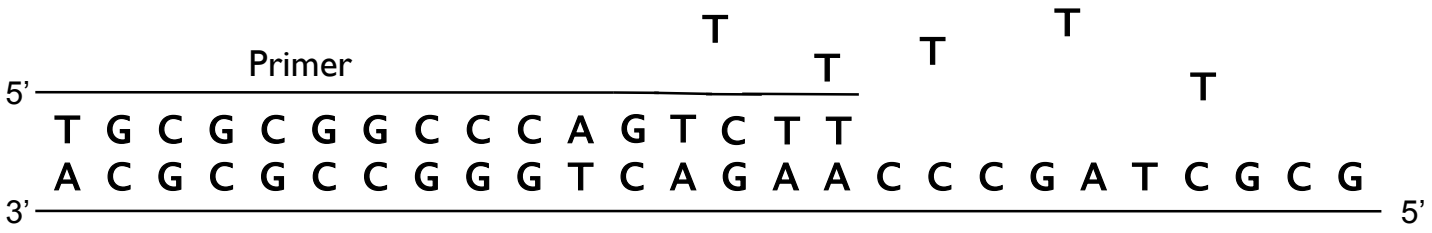


If that nucleotide is incorporated, enzymes turn by-products into light:

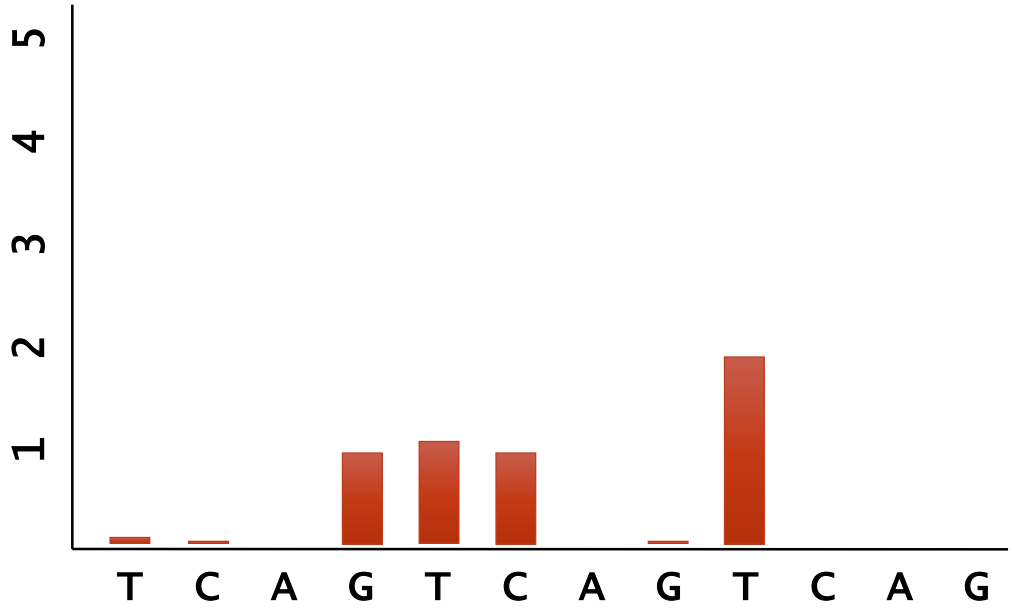


Shotgun sequencing by PGM/454

Only give polymerase one nucleotide at a time:

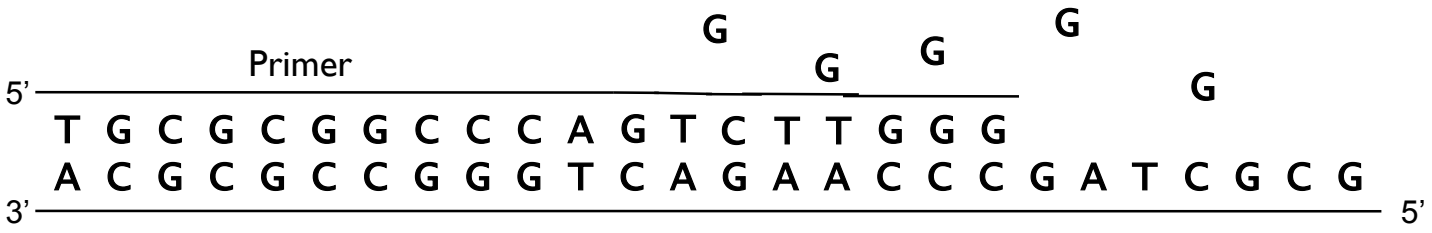


If that nucleotide is incorporated, enzymes turn by-products into light:

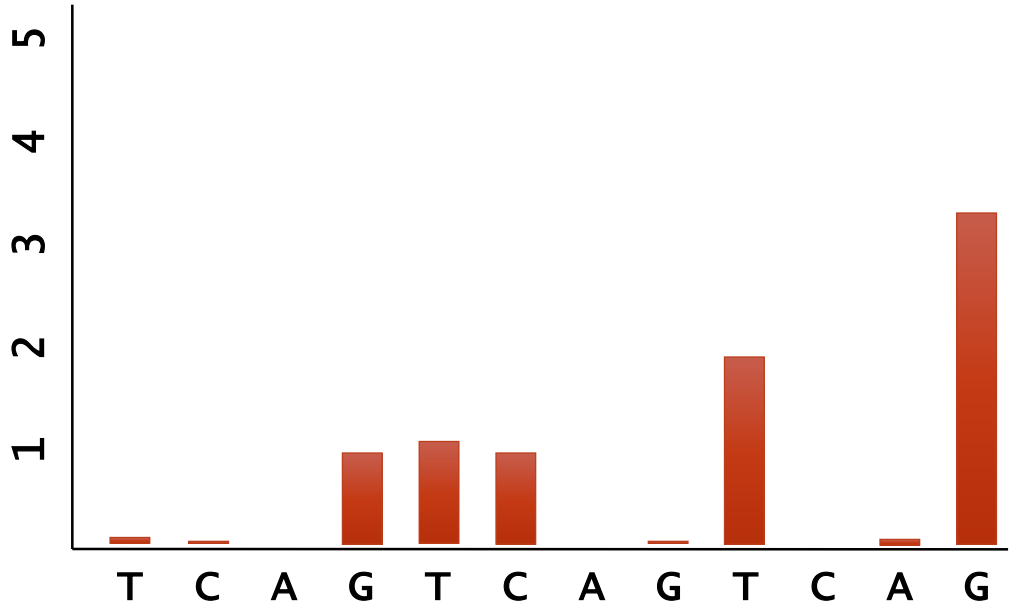


Shotgun sequencing by PGM/454

Only give polymerase one nucleotide at a time:

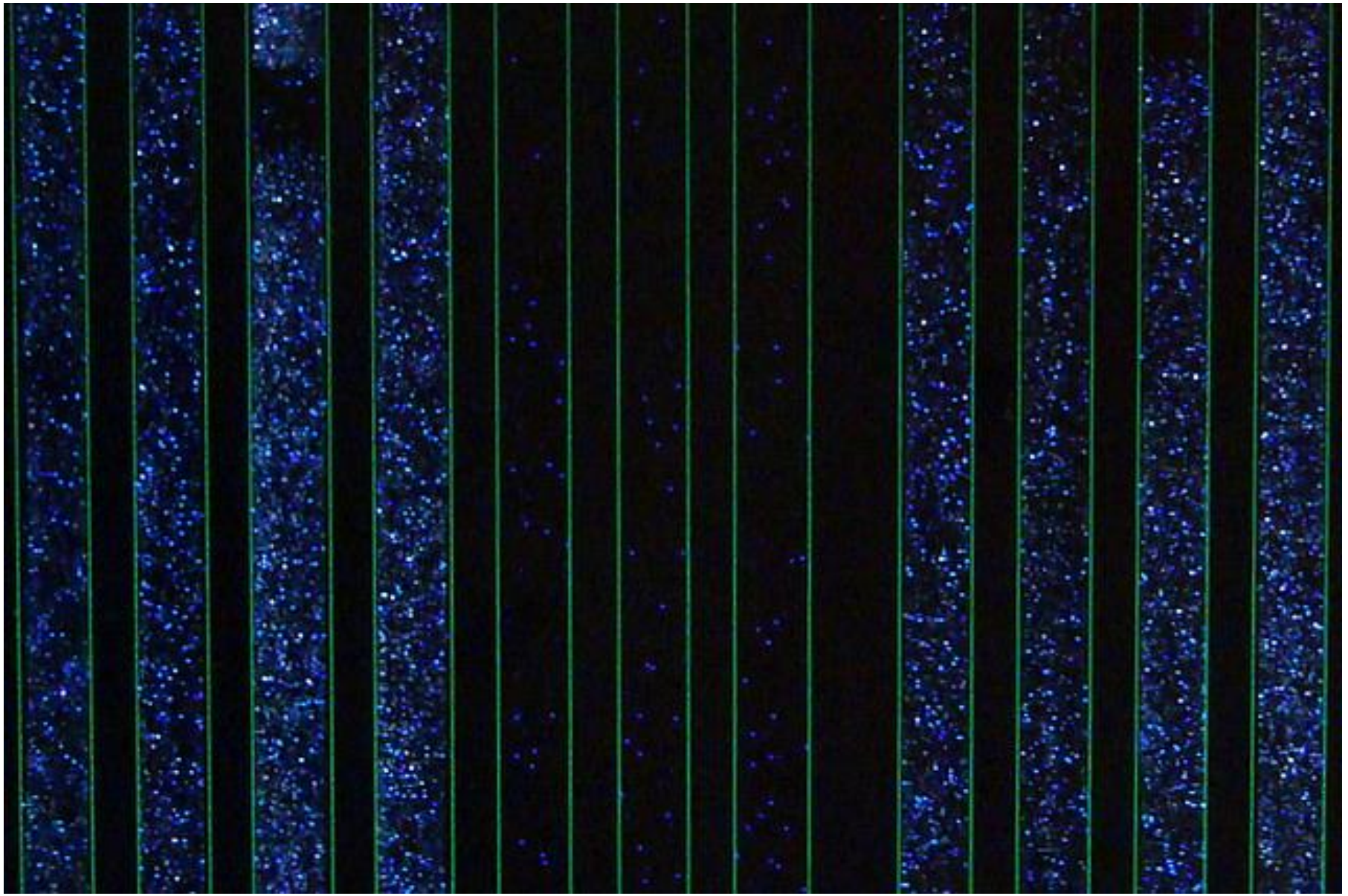


If that nucleotide is incorporated, enzymes turn by-products into light:

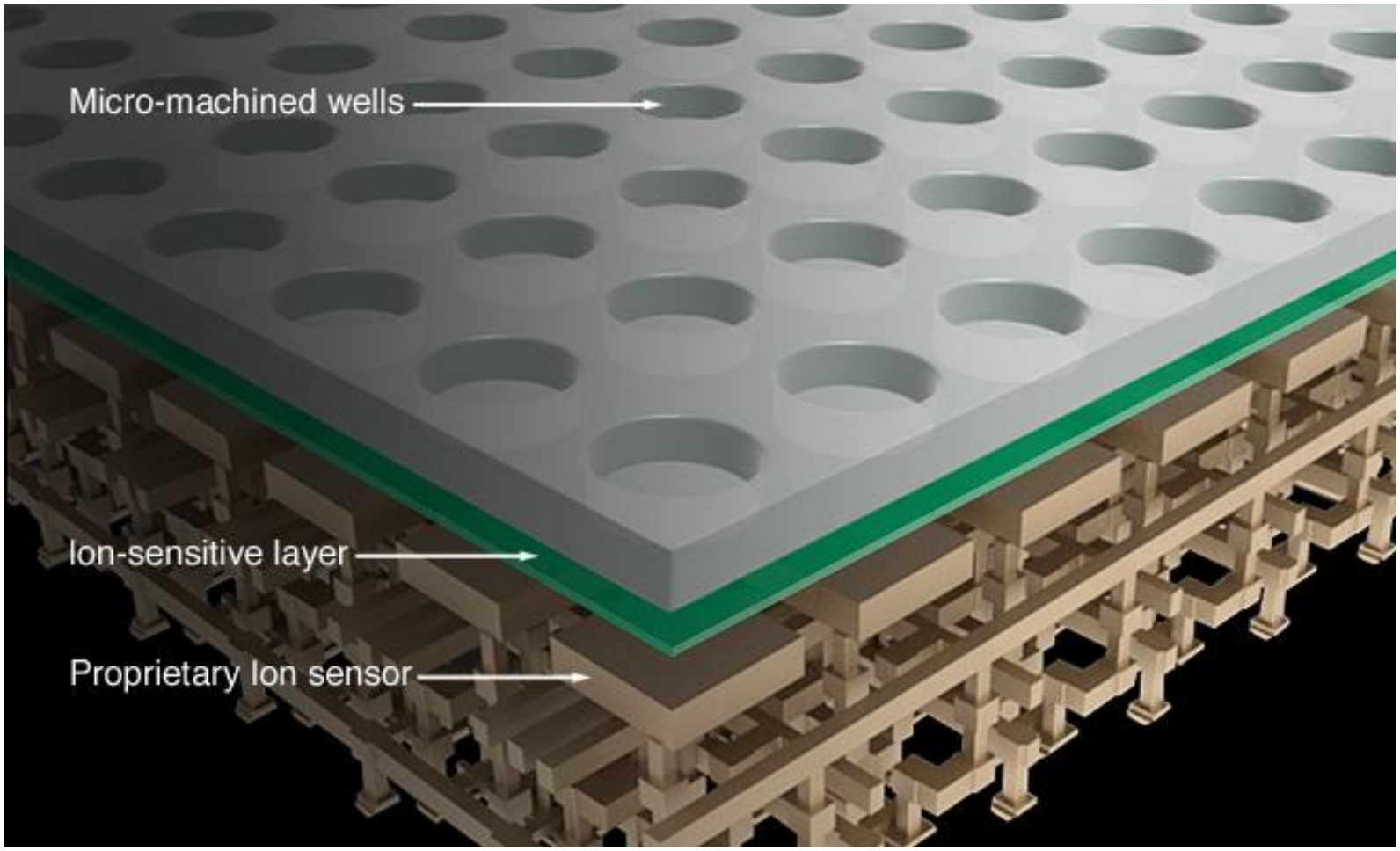


The real power of this method is that it can take place in millions of tiny wells in a single plate at once.

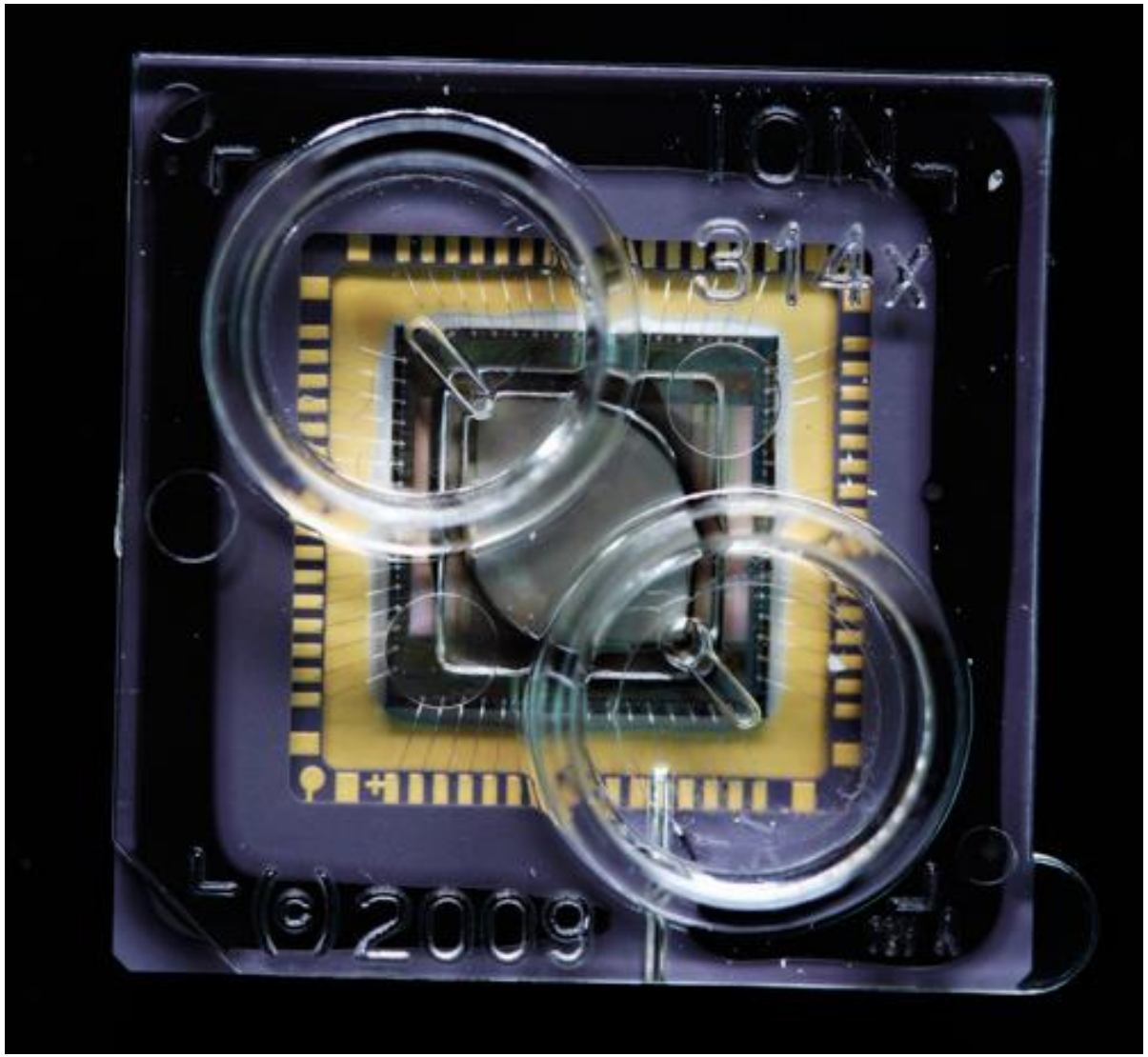
Raw 454 data



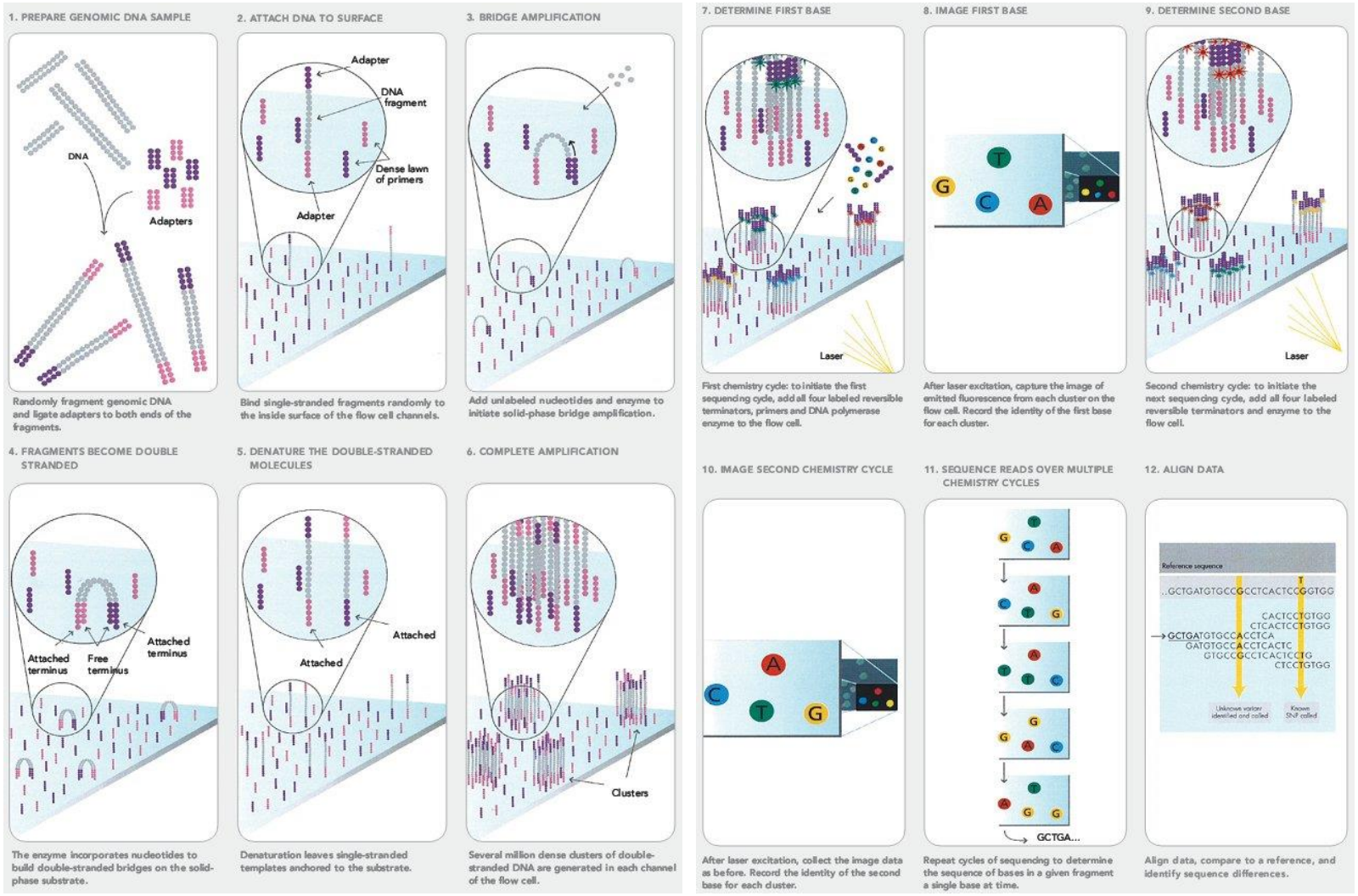
Ion Torrent Sequencing



Ion Torrent Sequencing



Illumina Sequencing



Next-Gen Sequencing

Take home message: Massively Parallel

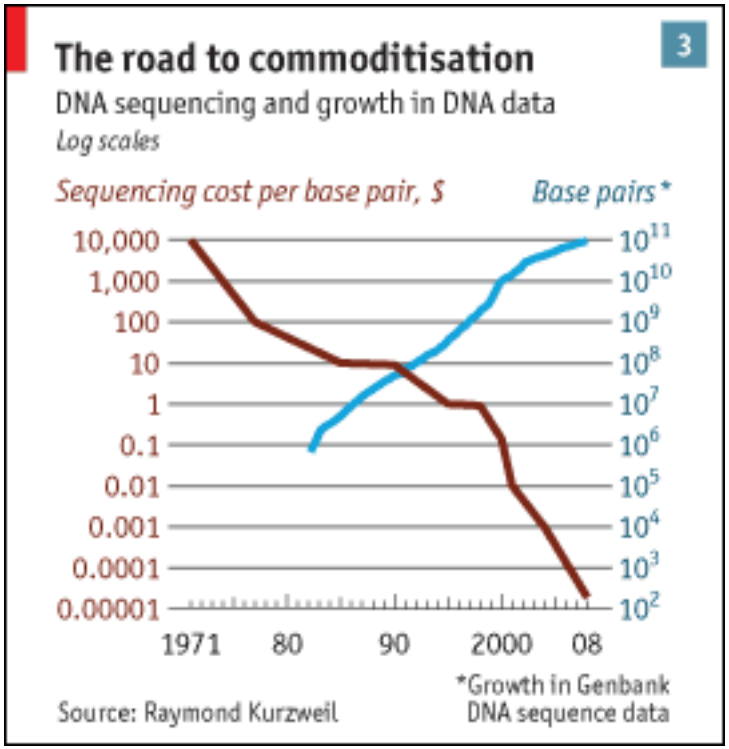
1,000 monkeys at 1,000 typewriters is nothing

We’re talking 100,000 to 100 million concurrent reads

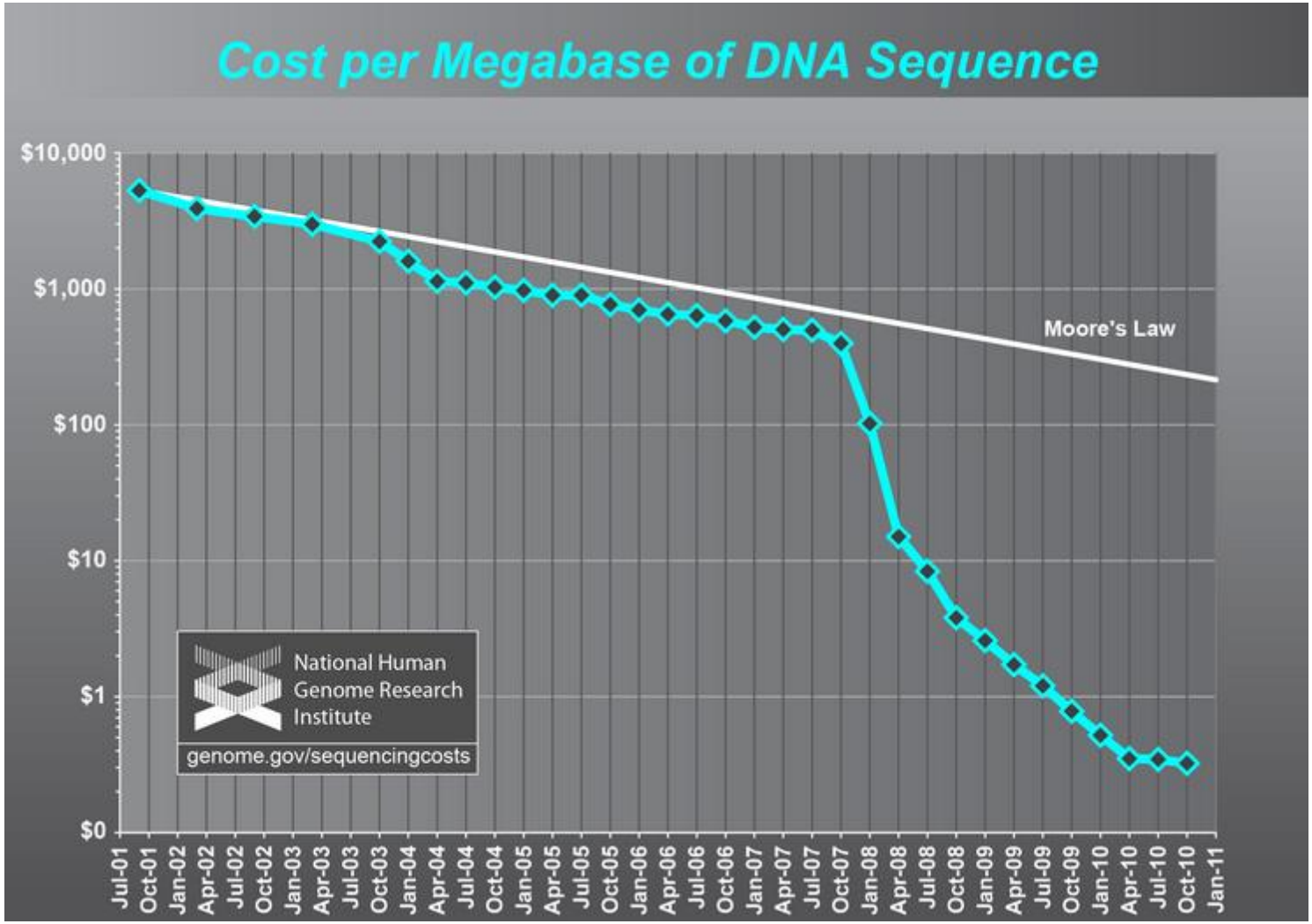
Overview

- Prologue: Assembly
- The Past: Sanger
- The Present: Next-Gen (454, Illumina, …)
- The Future: **? (Nanopore, MinION, Single-molecule)**

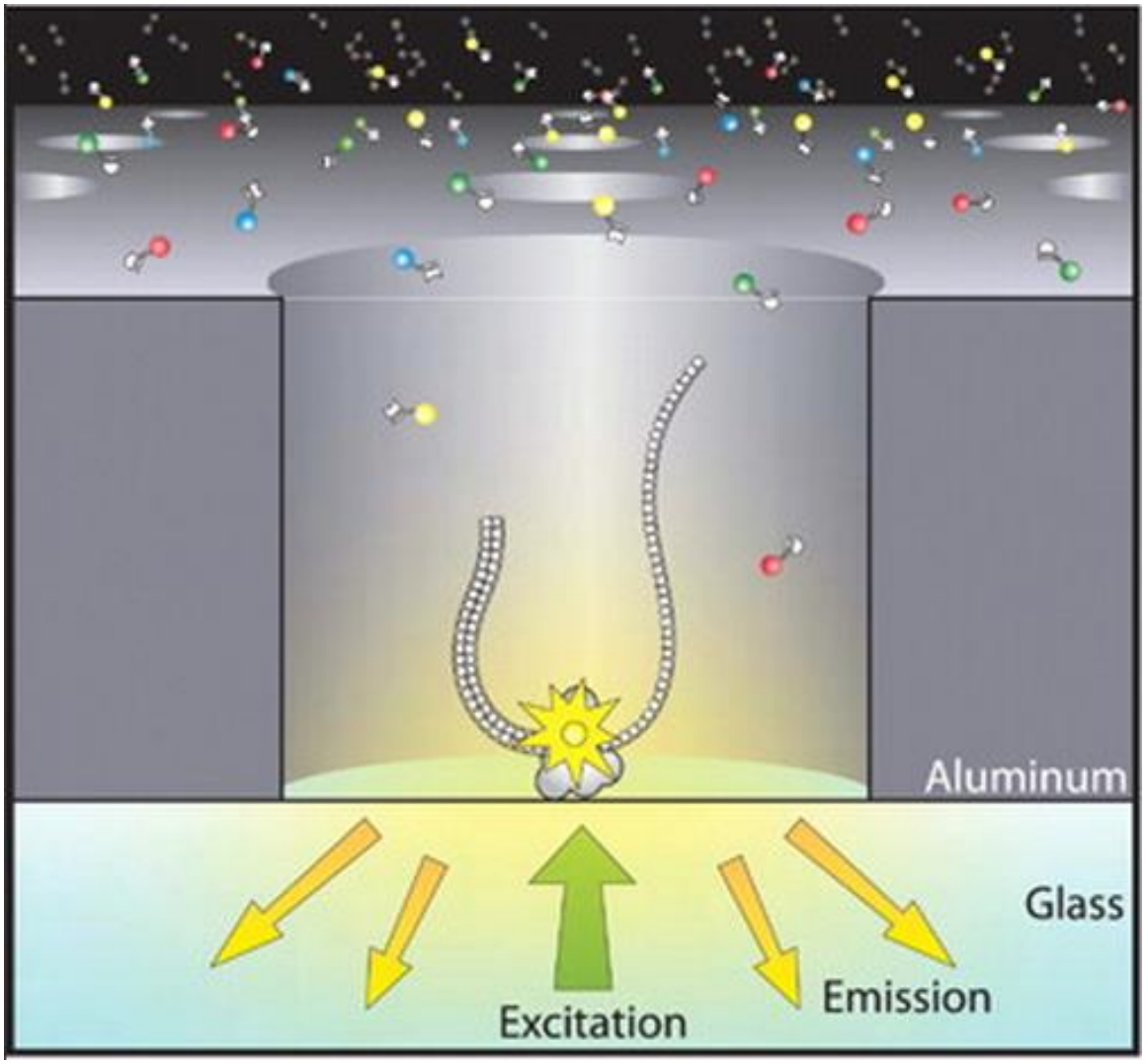
DNA Sequencing over Time



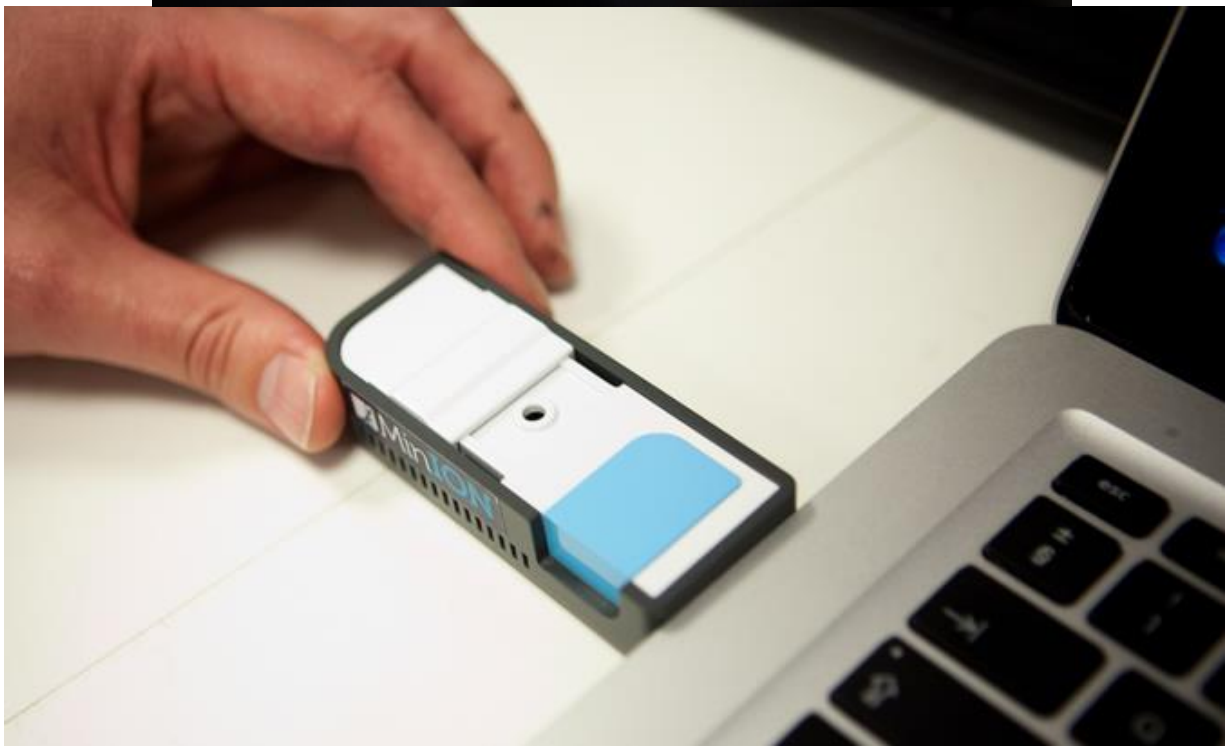
from *The Economist*


Cost per Megabase of DNA Sequence
$10,000
$1,000
$100
$10
$1
$0
Moore's Law
National Human Genome Research Institute
genome.gov/sequencingcosts
Jul-01
Oct-01
Jan-02
Apr-02
Jul-02
Oct-02
Jan-03
Apr-03
Jul-03
Oct-03
Jan-04
Apr-04
Jul-04
Oct-04
Jan-05
Apr-05
Jul-05
Oct-05
Jan-06
Apr-06
Jul-06
Oct-06
Jan-07
Apr-07
Jul-07
Oct-07
Jan-08
Apr-08
Jul-08
Oct-08
Jan-09
Apr-09
Jul-09
Oct-09
Jan-10
Apr-10
Jul-10
Oct-10
Jan-11

Single Molecule Sequencing




MinION


MinION

“The MinION has been used to successfully read the genome of a lambda bacteriophage, which has 48,500-ish base pairs, twice during one pass. That's impressive, because reading 100,000 base pairs during a single DNA capture has never been managed before using traditional sequencing techniques.

The operational life of the MinION is only about six hours, but during that time it can read more than 150 million base pairs. That's somewhat short of the larger human chromosomes (which contain up to 250 million base pairs), but Oxford Nanopore has also introduced GridION -- a platform where multiple cartridges can be clustered together. The company reckon that a 20-node GridION setup can sequence a complete human genome in just 15 minutes.”

—Wired

(Relevant) Trivia

How many base pairs (bp) are there in a human genome?

~3 billion (haploid)

How much did it cost to sequence the first human genome?

~$2.7 billion

How long did it take to sequence the first human genome?

~13 years

When was the first human genome sequence complete?

2000-2003

Whose genome was it?

Several people's, but actually mostly a dude from Buffalo

Final Thoughts

- DNA sequencing is becoming vastly faster and more affordable
- Generating data is no longer the bottleneck, understanding it is
- Bioinformatics types should be in high demand in the near future

Comparing Different Technologies

Sanger Sequencing

Advantages	Disadvantages
Lowest error rate	High cost per base
Long read length (~750 bp)	Long time to generate data
Can target a primer	Need for cloning
	Amount of data per run

Comparing Different Technologies

454 Sequencing

Advantages	Disadvantages
Low error rate Medium read length (~400-600 bp)	Relatively high cost per base Must run at large scale Medium/high startup costs

Comparing Different Technologies

Ion Torrent Sequencing

Advantages	Disadvantages
Low startup costs Scalable (10 – 1000 Mb of data per run) Medium/low cost per base Low error rate Fast runs (<3 hours)	New, developing technology Cost not as low as Illumina Read lengths only ~100-200 bp so far

Comparing Different Technologies

Illumina Sequencing

Advantages	Disadvantages
Low error rate Lowest cost per base Tons of data	Must run at very large scale Short read length (50-75 bp) Runs take multiple days High startup costs De Novo assembly difficult

Comparing Different Technologies

PacBio Sequencing

Advantages	Disadvantages
Can use single molecule as template Potential for very long reads (several kb+)	High error rate (~10-15%) Medium/high cost per base High startup costs

Sequencing

illumina®

Platform	Reads per run	Read length (mode or average)	Bases per run (gigabases)
ABI Sanger	96	800	0.0000768
454	1 millions	700	0.7
IonTorrent	75 millions	200	15
SOLiD	3 billions	75	320
Illumina	**600 millions to 6 milliards**	**100 to 300**	**7.5 to 2 000**

MiniSeq System

Up to 7.5 Gb

MiSeq Series ⊕

Up to 15 Gb

NextSeq Series ⊕

Up to120 Gb

NextSeq Series ⊕

Up to120 Gb

HiSeq Series ⊕

Up to 750 Gb

HiSeq X Series†

Up to 800 Gb

NovaSeq Series ⊕

Up to 2 Tb

Sequencing

illumina®

Platform	Reads per run	Read length (mode or average)	Bases per run (gigabases)
ABI Sanger	96	800	0.0000768
454	1 millions	700	0.7
IonTorrent	75 millions	200	15
SOLiD	3 billions	75	320
Illumina	**600 millions to 6 milliards**	**100 to 300**	**7.5 to 2 000**

MiniSeq System

50,000 $

MiSeq Series

99,000 $

NextSeq Series

250,000 $

NextSeq Series

250,000 $

HiSeq Series

690,000 $-
900,000 $

HiSeq X Series†

1,000,000 $
(X10= 10,000,000 $)

NovaSeq Series

850,000 $-
985,000 $

Single molecule Long Read Sequencing

PACIFIC BIOSCIENCES®

Platform	Year	Reads per run	Read length (mode or average)	Bases per run (gigabases)
ABI Sanger	2002	96	800	0.0000768
454	2011	1 millions	700	0.7
SOLiD	2013	3 milliards	75	320
IonTorrent	2015	75000000	200	15
Illumina	2016	600 millions to 6 milliards	100 to 300	7.5 to 2 000
PacBio	**2014**	**660000**	**13500**	**20**

RSII 700,000 $

Sequel 350,000 $

Roche

Single molecule Long Read Sequencing

Oxford NANOPORE Technologies®

Platform	Reads per run	Read length (mode or average)	Bases per run (gigabases)
ABI Sanger	96	800	0.0000768
454	1 millions	700	0.7
SOLiD	3 milliards	75	320
IonTorrent	75000000	200	15
Illumina	600 millions to 6 milliards	100 to 300	7.5 to 2 000
PacBio	660000	13500	12.000
Oxford Nanopore	**4400000**	**9545**	**42**

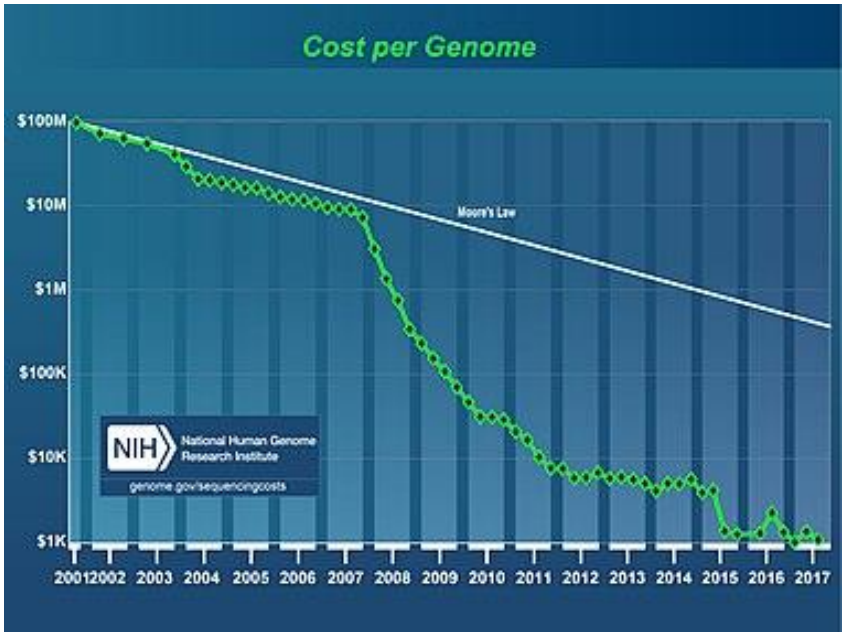
MinION Mkl: portable, real time biological analyses

MinION 1,000 $

Sequencing Technologies are getting Faster and Cheaper

The use of nucleic acid sequencing has increased exponentially as the ability to sequence has become accessible to research and clinical labs all over the world

Several Sequencing Technologies Applications "–Seq" : RNA-Seq, Chip-Seq, SingleCell-Seq, etc.



This demand has driven the development of High Troughput Sequencing (HTS), that are becoming exponentially cheaper.

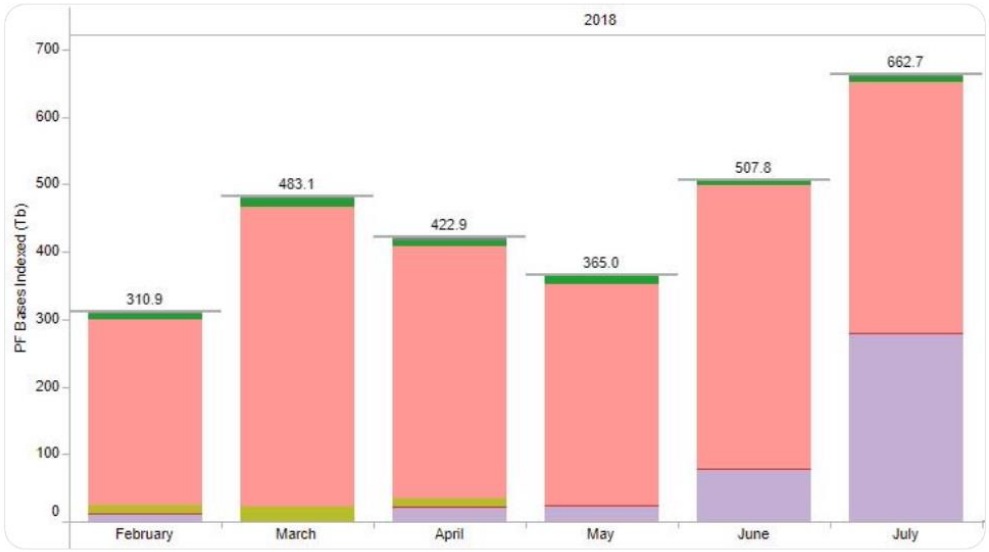
The exponential growth of genomic data unfortunately is not followed by an exponential growth of storage

A new Human Genome every 6 min

Broad Genomics
@BroadGenomics

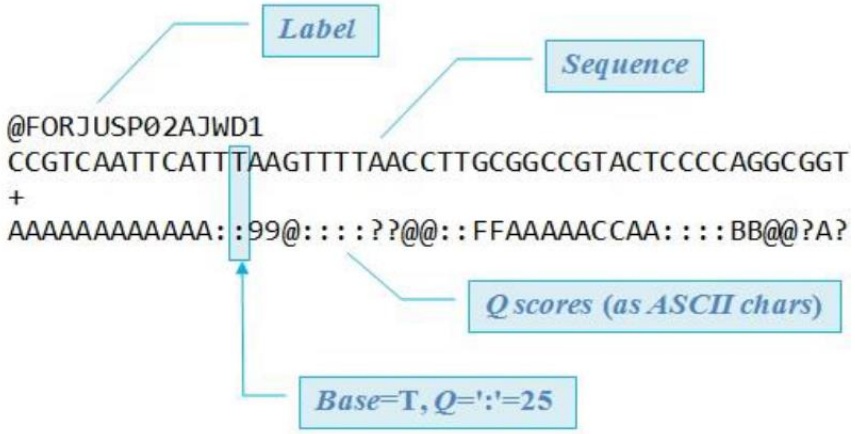
Segui

The sequencing lab just rang in its biggest month of sequence data generation EVER! 663 Terabases in the month of July. Equivalent to a human genome every 6 minutes.
#NewBenchmark #QualityandQuantity
#CantStopWontStop



What is the output of HTS

FastQ File



- In many applications compression is required to reduce space and improve throughput. e.g. The most popular reads mapper (BWA) can operate directly on the compressed FastQ.
- The DNA sequence exposes a high redundancy, especially on large reads collections with high coverage, and thus it is highly compressible
- Quality values have much higher entrophy than a genomic sequence because their alphabeth usually span a much larger range of values (e.g. [1 - 40]), and they are not repetitive.
- When FASTQ files are compressed quality values account for 70% of the total space.

Background on Quality Values

Quality Value	Probability of incorrect base call	Base call accuracy
10	**1 in 10**	**90%**
20	**1 in 100**	**99%**
30	**1 in 1000**	**99.9%**
40	**1 in 10,000**	**99.99%**

- Q(i): Phred-scaled probability that the i-th base of the read being wrong
- Q(i) = -10 $\log_{10}$ Prob{the base i of read is wrong}

Applications:

- SNP/Mutation Detection
- Removal of low-quality reads
- Reads Mapping
- Detection of Overlapping reads
- Error Correction
- Compression
- etc.