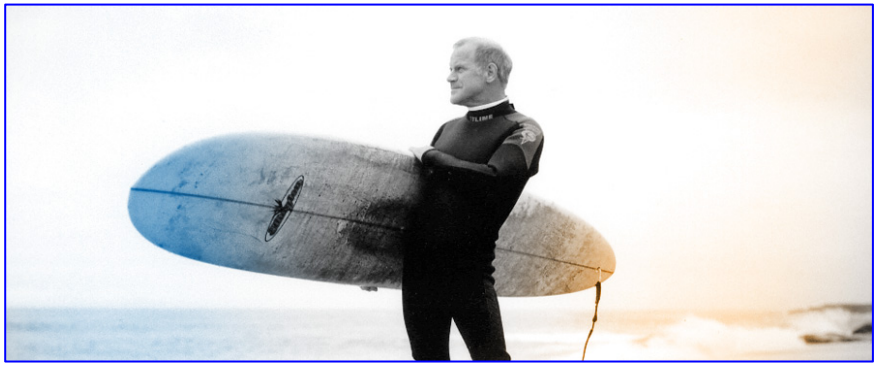


BIOL/CS 123B
Deck 1: Sequencing and Assembly

Spring 2021

Philip Heller

Topics

- Amplification
- 3 generations of sequencing
- Shotgun sequencing and assembly

Topics

- Amplification
- 3 generations of sequencing
- Shotgun sequencing and assembly

Amplification

- WHAT: Making many identical copies of a DNA strand.
- HOW: Subvert DNA's natural ability to self-replicate (cloning or PCR)
- WHY: It's the only way to determine the sequence…technology requires lots of copies.
- UNTIL: Single-molecule sequencing becomes an affordable reality

Cloning – Before there was PCR

- Long ago in a galaxy far far away…

Luke: You fought in the Clone Wars?

Obi-Wan: Yes. I was once a Jedi knight, the same as your father.

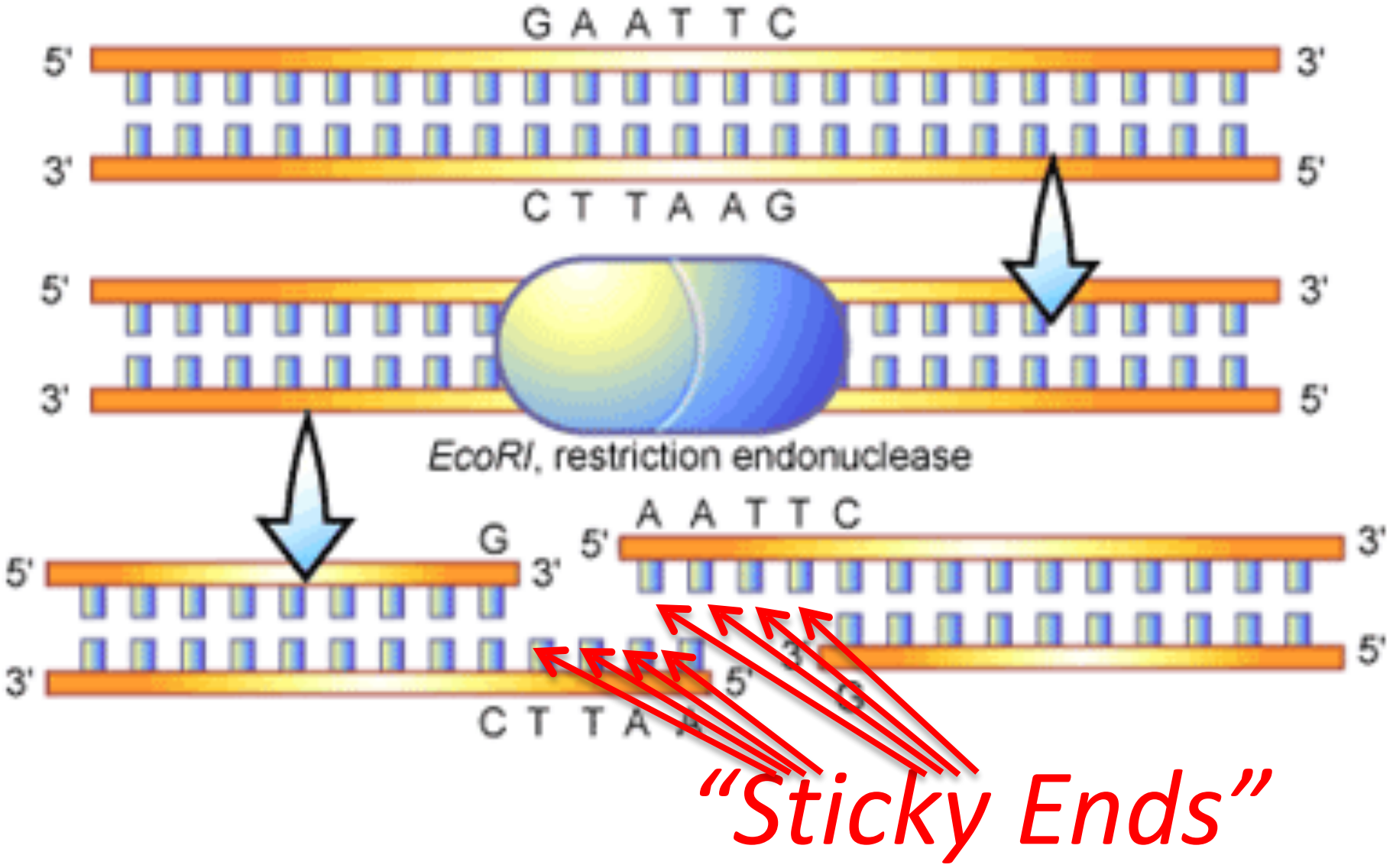
-- *Original "Star Wars", 1977*

Cloning: 1972

- Technology
  - Restriction enzymes: cut DNA molecule at a specific sequence
  - DNA ligases: glue together 2 DNA fragments

- Bacteria contain *plasmids*
  - Small (~5000 bp) circular DNA fragments
  - Reproduced during cell reproduction
  - Relatively easy to sequence, even with 1960s/1970s technology
  - Some sequences 100% known

Restriction Enzyme

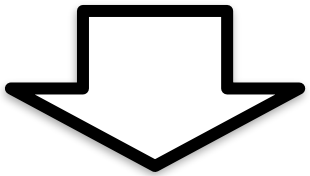


DNA Ligase

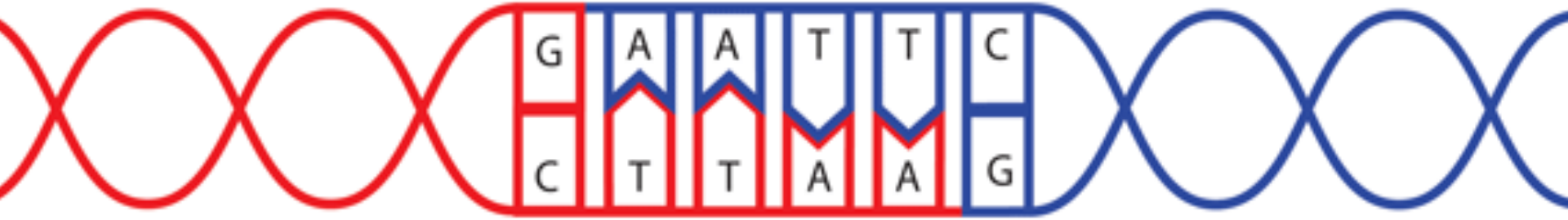
Sticky End

Sticky End

Ligation



Engineered (“Recombinant”) DNA

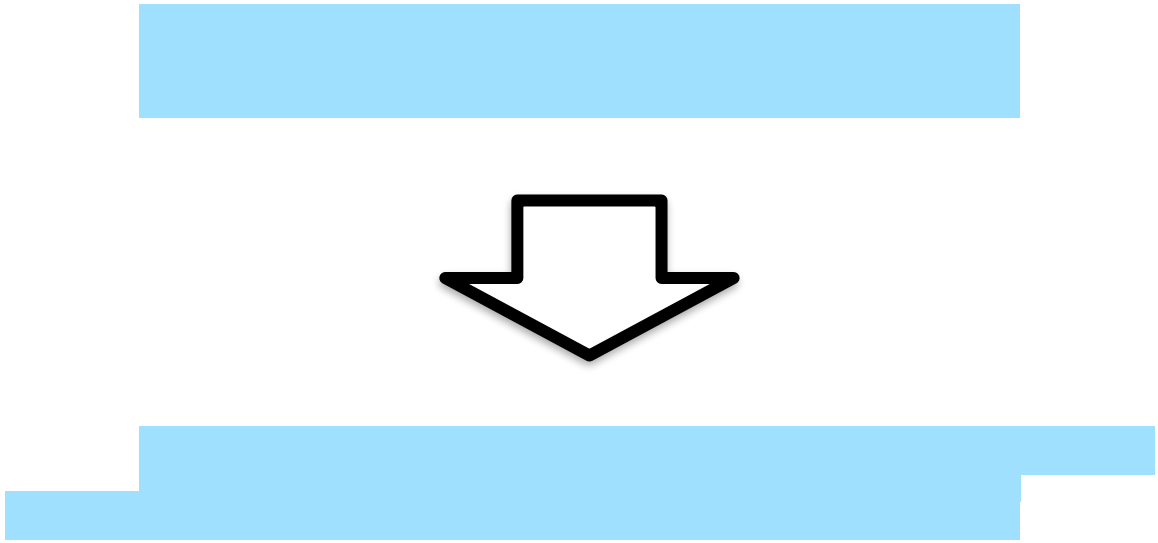


So …

- They could cut-and-paste known sequences
- The knew some plasmid sequences
- Plasmids amplify naturally
- → DNA cloning
- "Clone" = to make a genetically identical copy
  - Zygote => identical twins
  - Cutting from your neighbor's tomato plant => your own tomato plant
  - DNA => many copies

Cloning: Step 1

Add sticky ends to original gene



Cloning: Step 2

Cut a plasmid with restriction enzymes

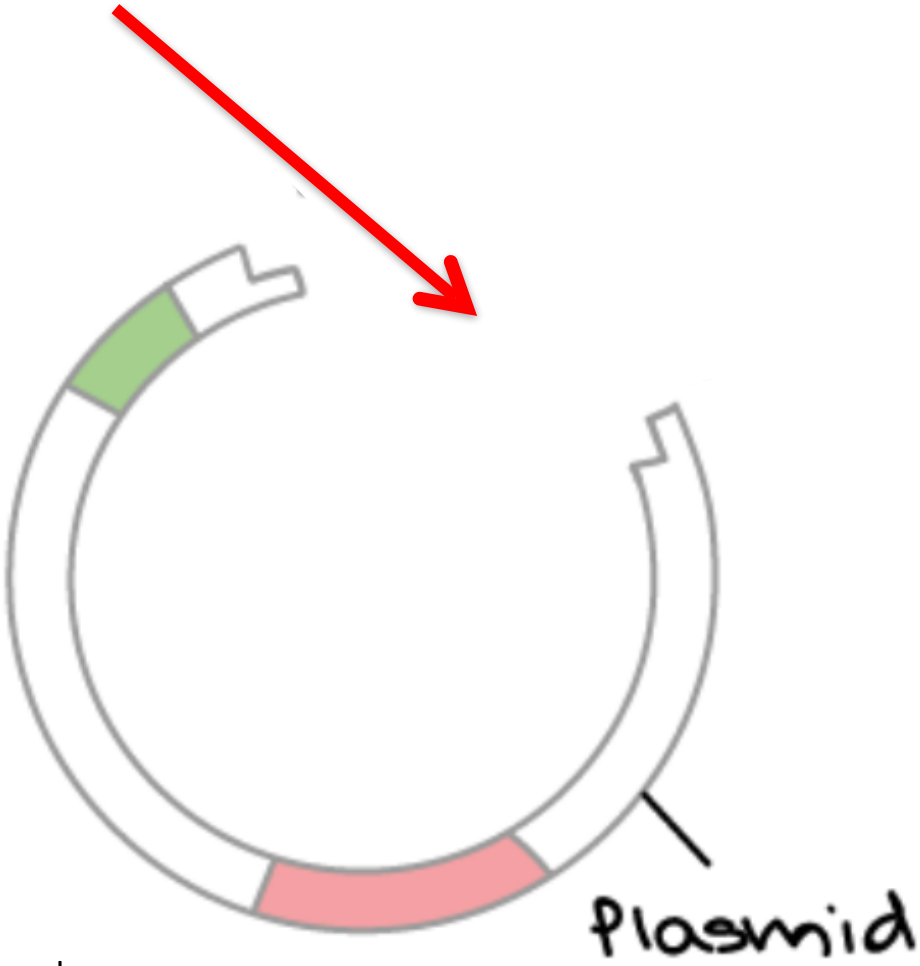
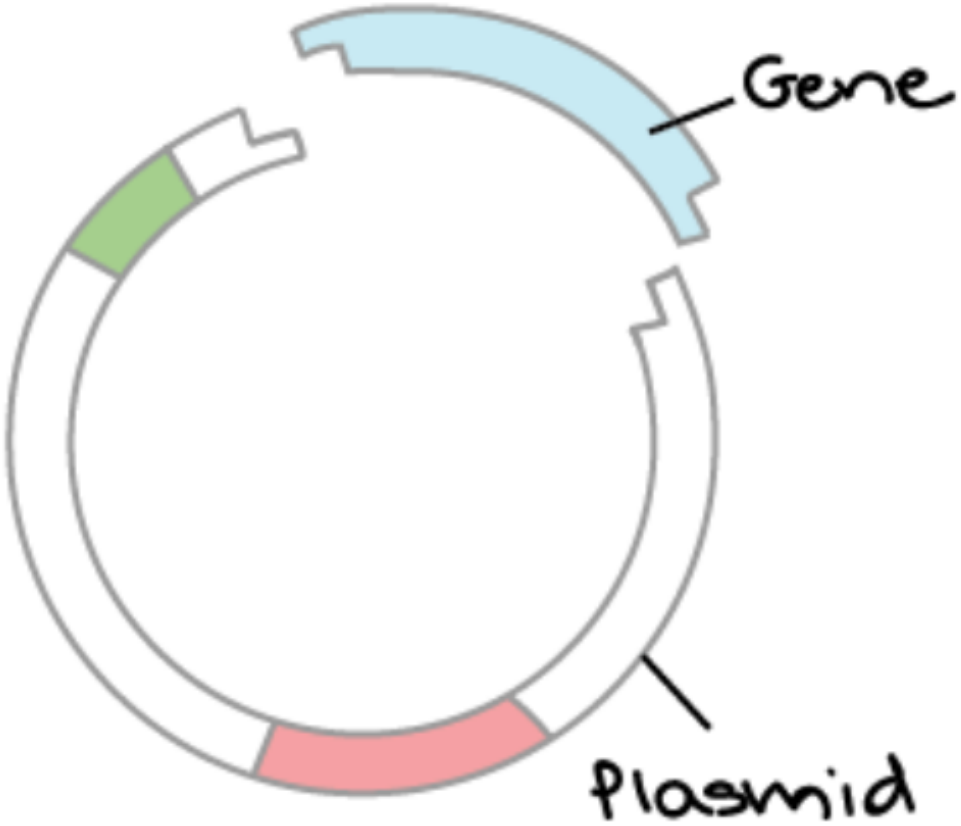


Image source: Khanacademy.org

Cloning: Step 3

Add gene with appropriate sticky ends
Add ligase

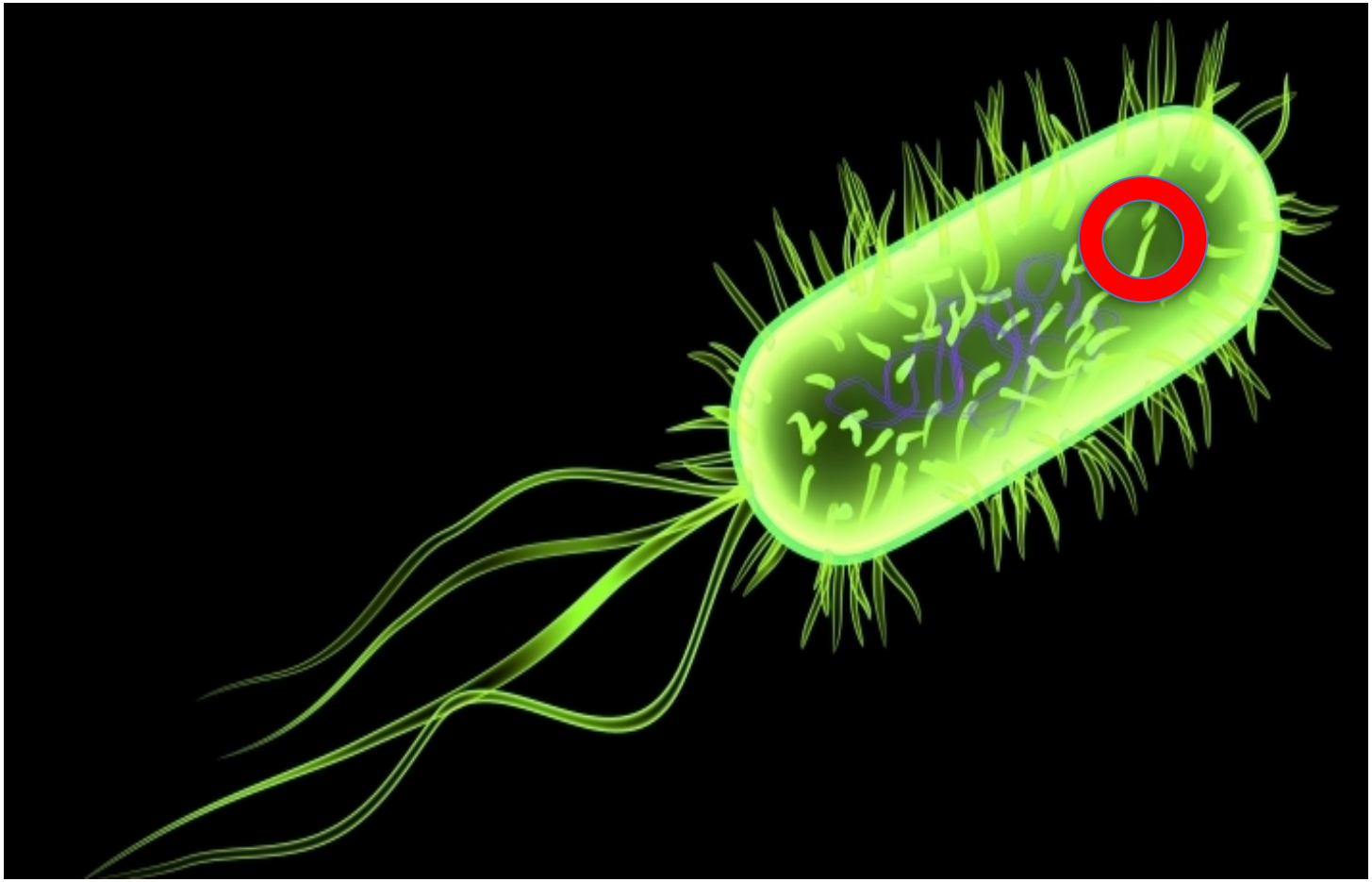


Cloning: Step 4

Plasmid and gene combine → recombinant DNA

Cloning: Step 5

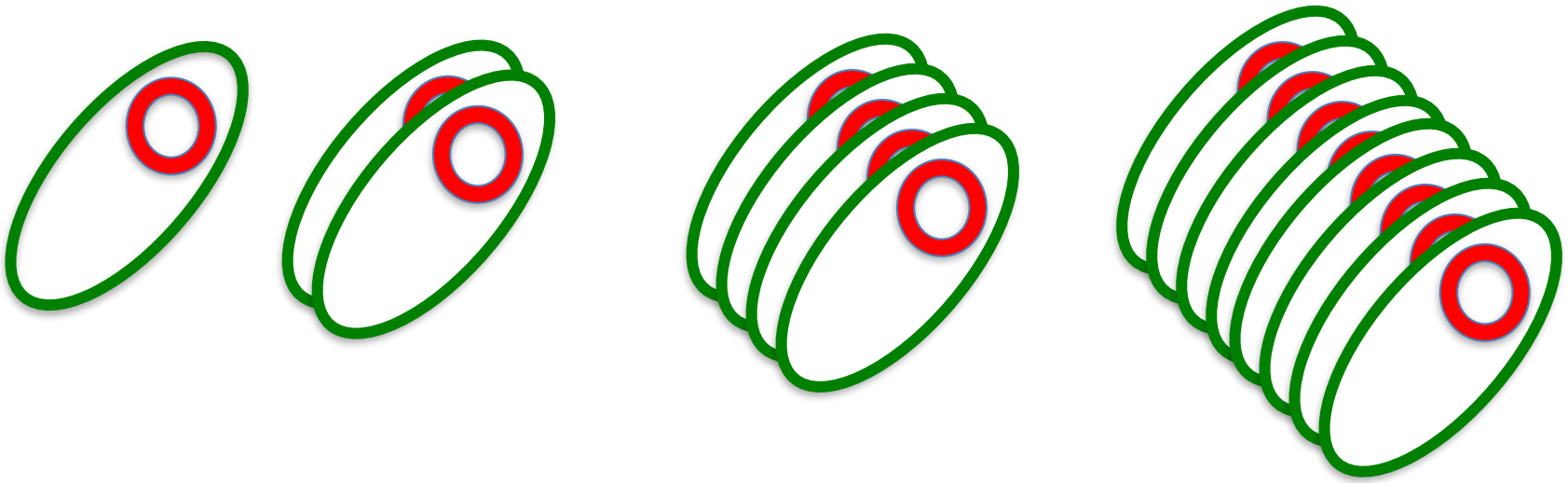
Insert engineered plasmid into a bacterium



Cloning: Step 6

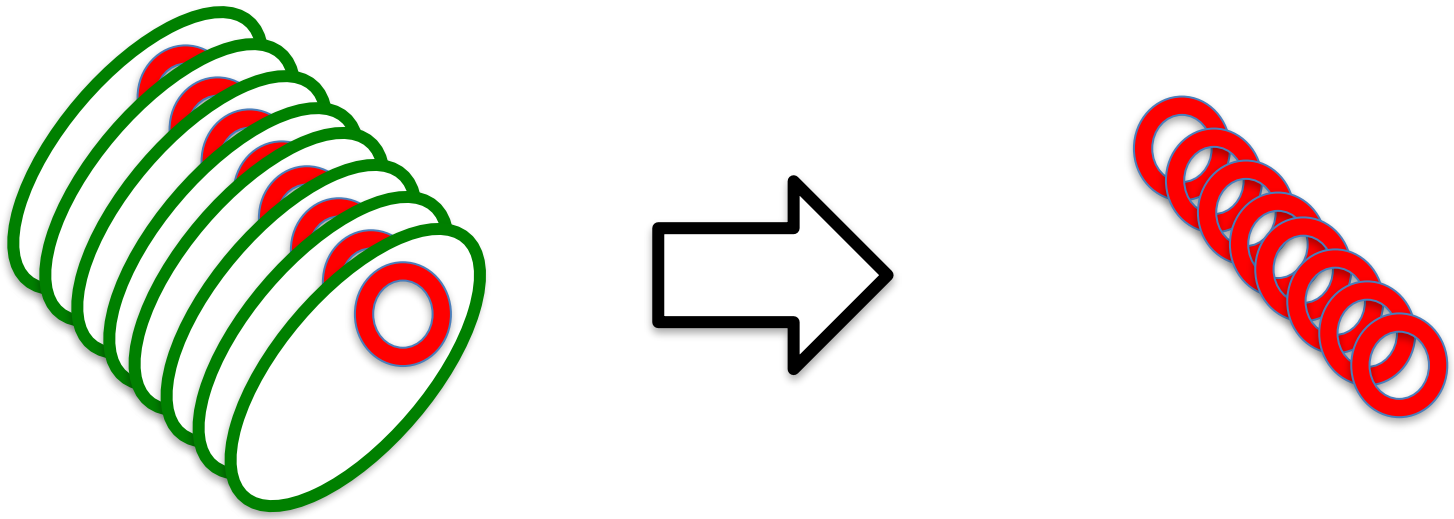
Bacterium reproduces, doubling population each generation

1 → 2 → 4 → 8 → 16 → 32 …



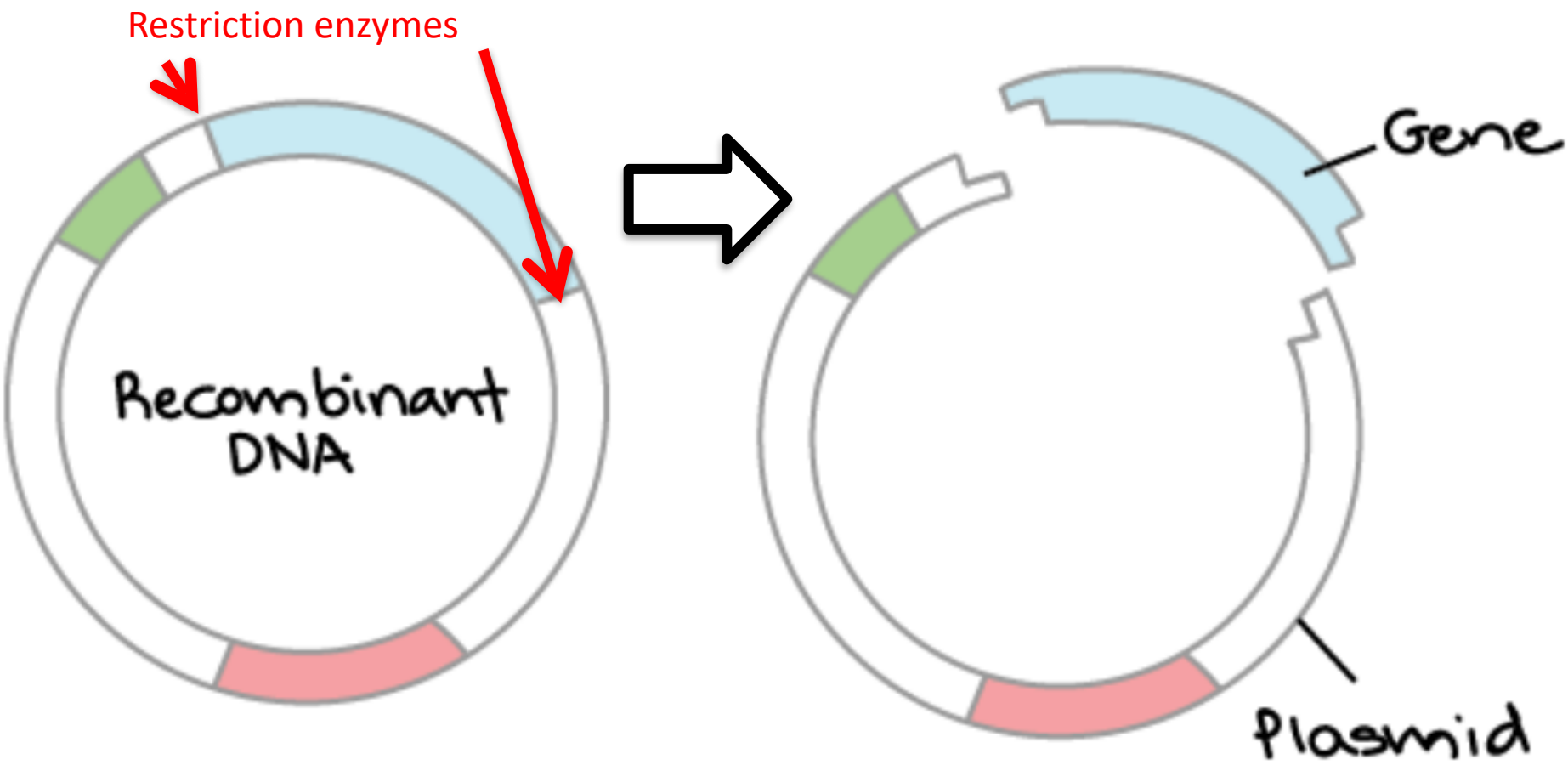
Cloning: Step 7

Extract all the replicated plasmids, discard everything else



Cloning: Step 8

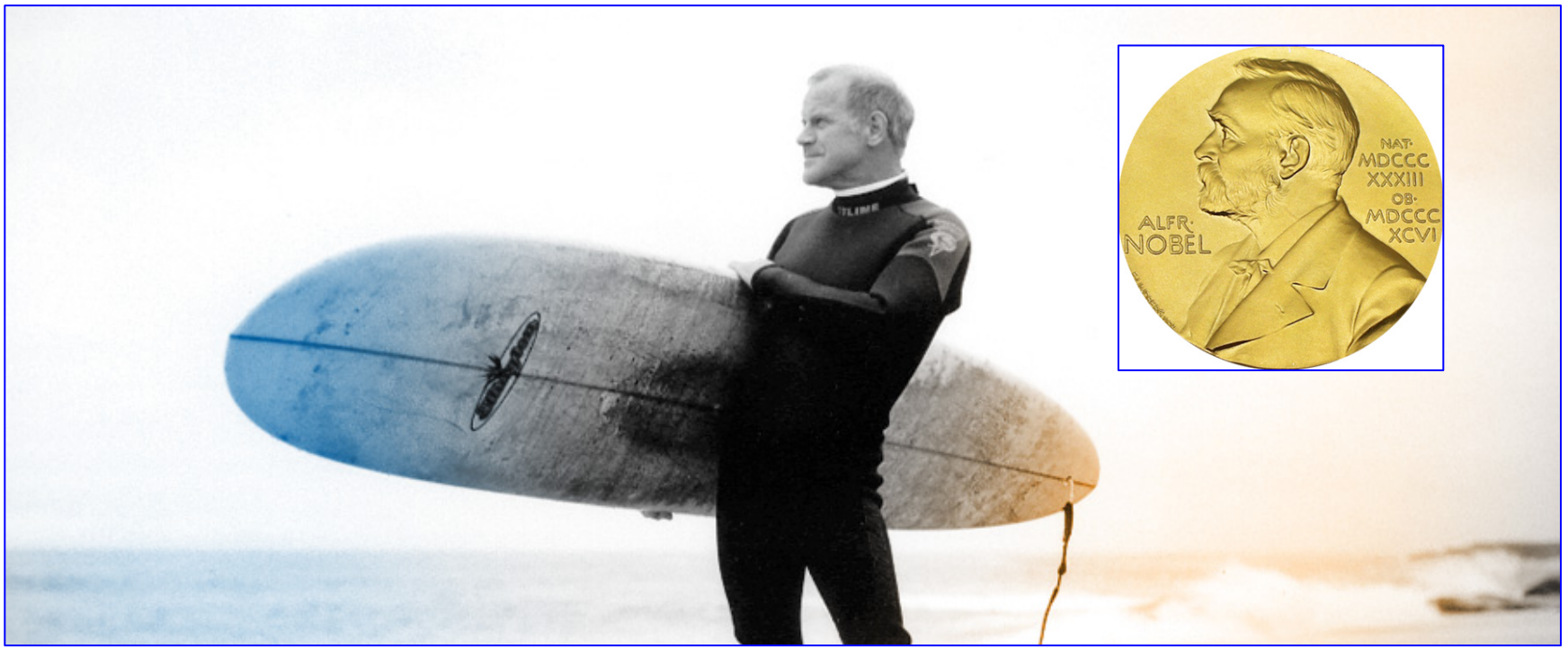
Extract target gene from each replicated plasmid, discard everything else



Drawbacks of cloning as an amplification technology

- To replicate a gene, you need plasmids and bacteria.
- The doubling time is limited by the generation time of the bacteria (best case = ~1/2 hour).
- → If only the target gene could be made to replicate directly!
  - No bacteria.
  - No plasmids.

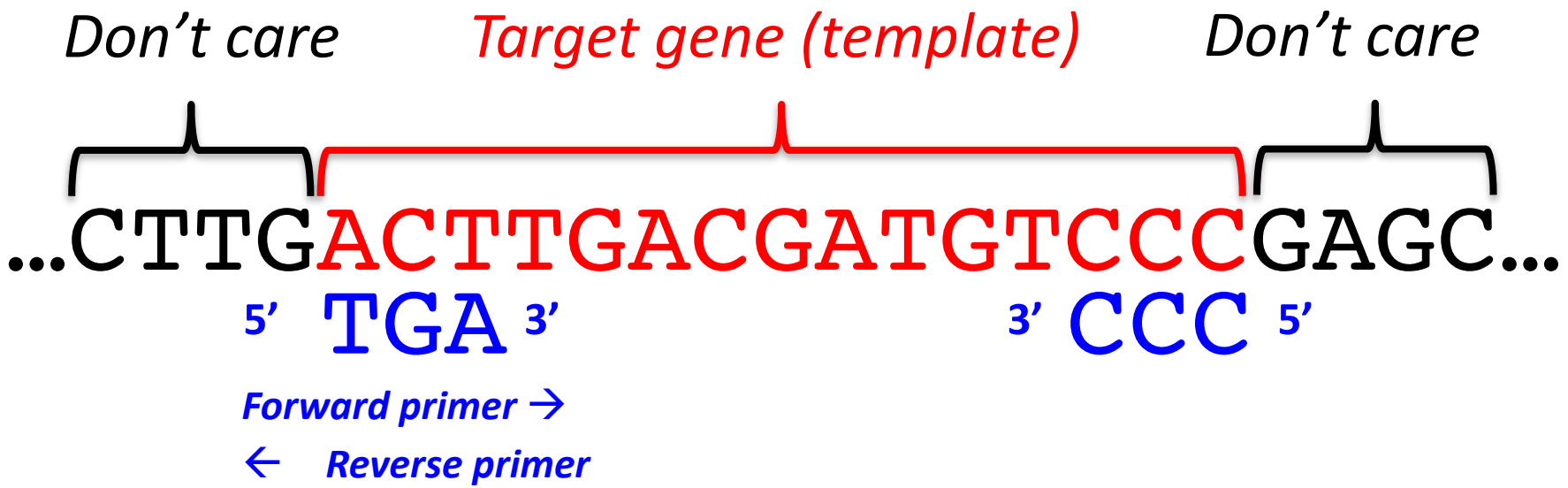
1983: Kary Mullis develops Polymerase Chain Reaction (PCR)



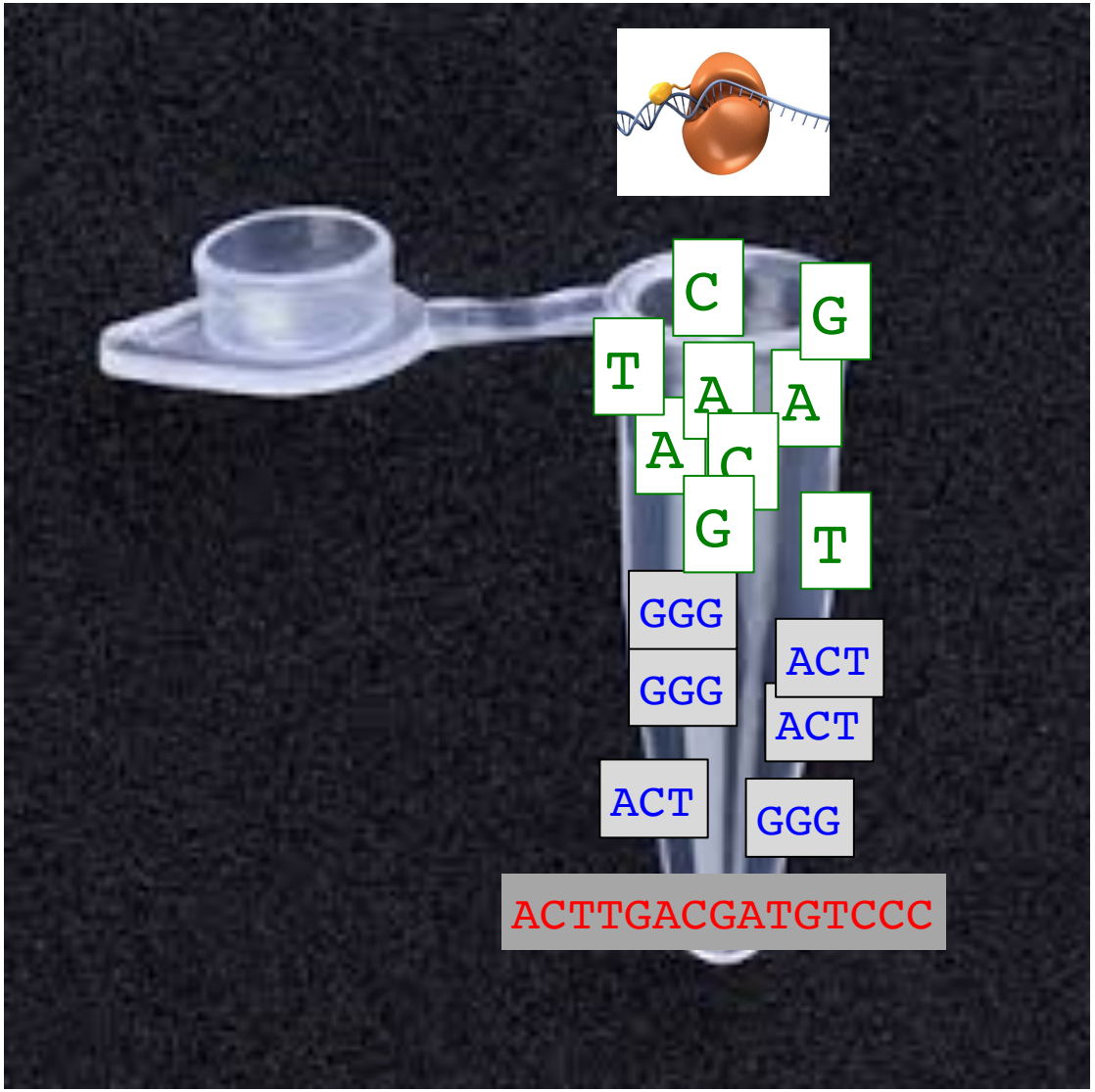
No more plasmids, no more bacteria

PCR Primers (short single strands)

- If you know the sequence of the gene you want to amplify…
- Or actually just the beginning and end of the sequence …
- Make 2 primers
  - Reverse complement of first n bases of the gene
  - Last n bases of the gene
  - n ~= 10 – 30 (but n = 3 in these simple examples)



The PCR Tube: where it all happens



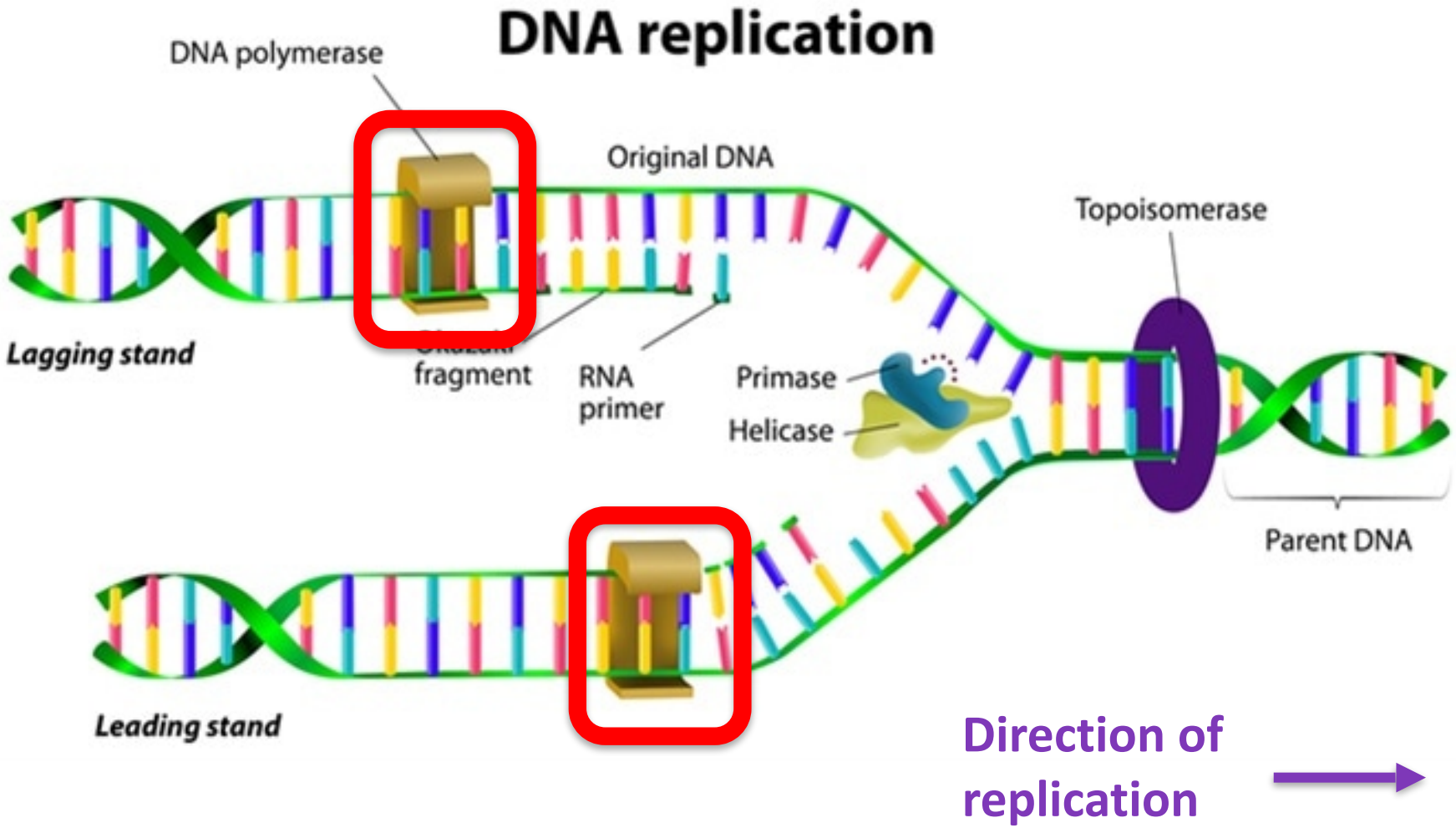
Add DNA template

Add many copies of primers

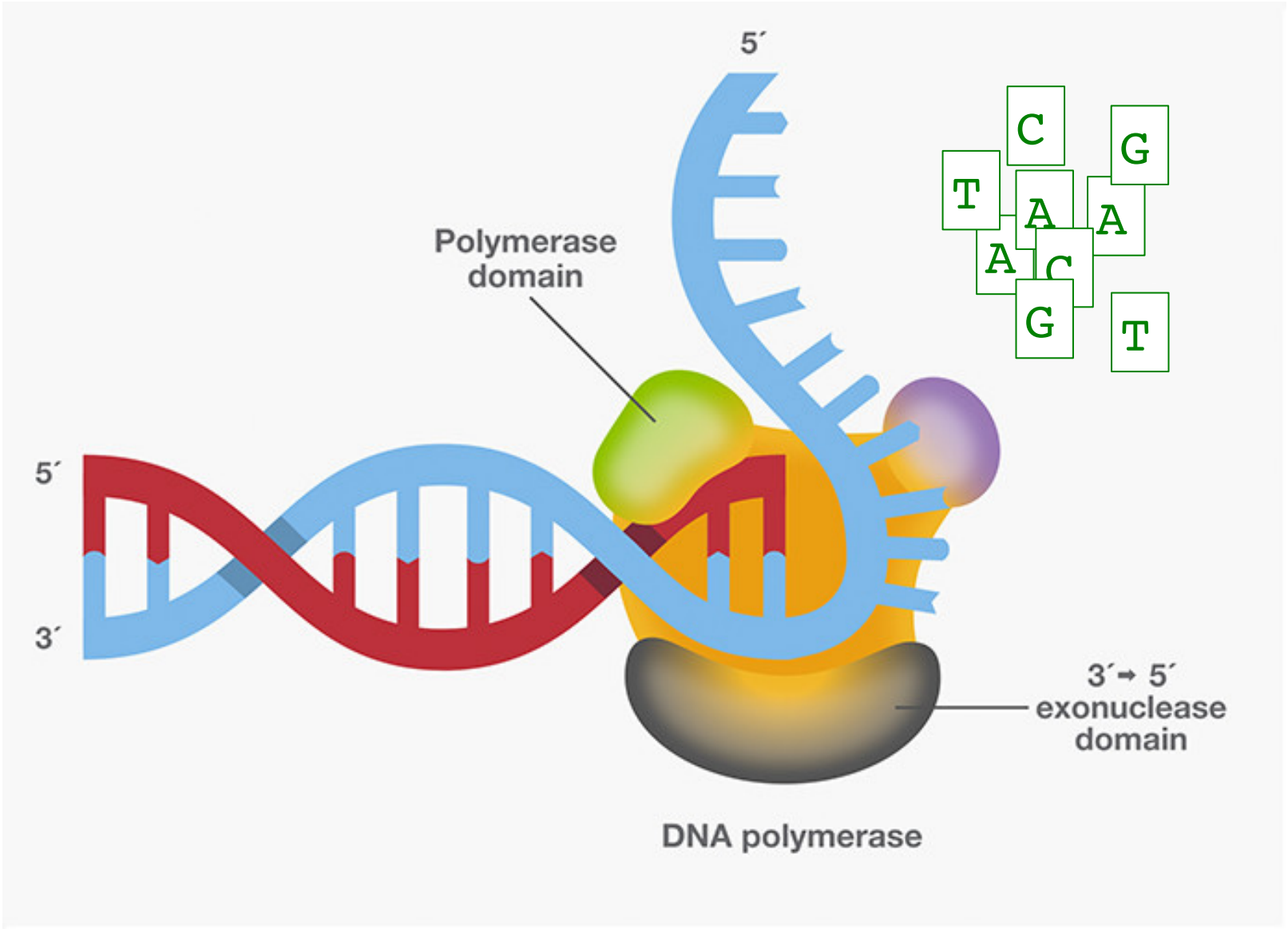
Add lots of As, Cs, Gs & Ts

Add DNA polymerase and other enzymes

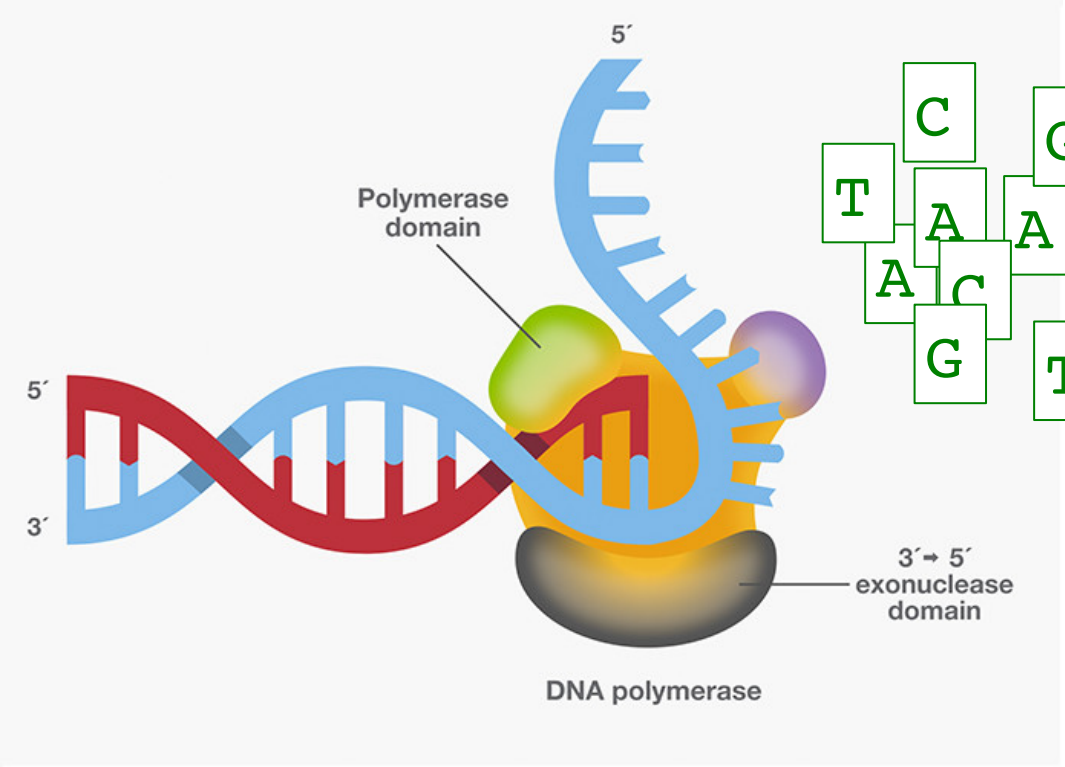
DNA Polymerase: quick review



DNA Polymerase: quick review



DNA Polymerase: quick review



- Polymerase moves 3’ → 5’ along a single **strand**
- Attaching matching nucleotides to the new **strand**
- Which grows 5’ → 3’

Polymerase Chain Reaction

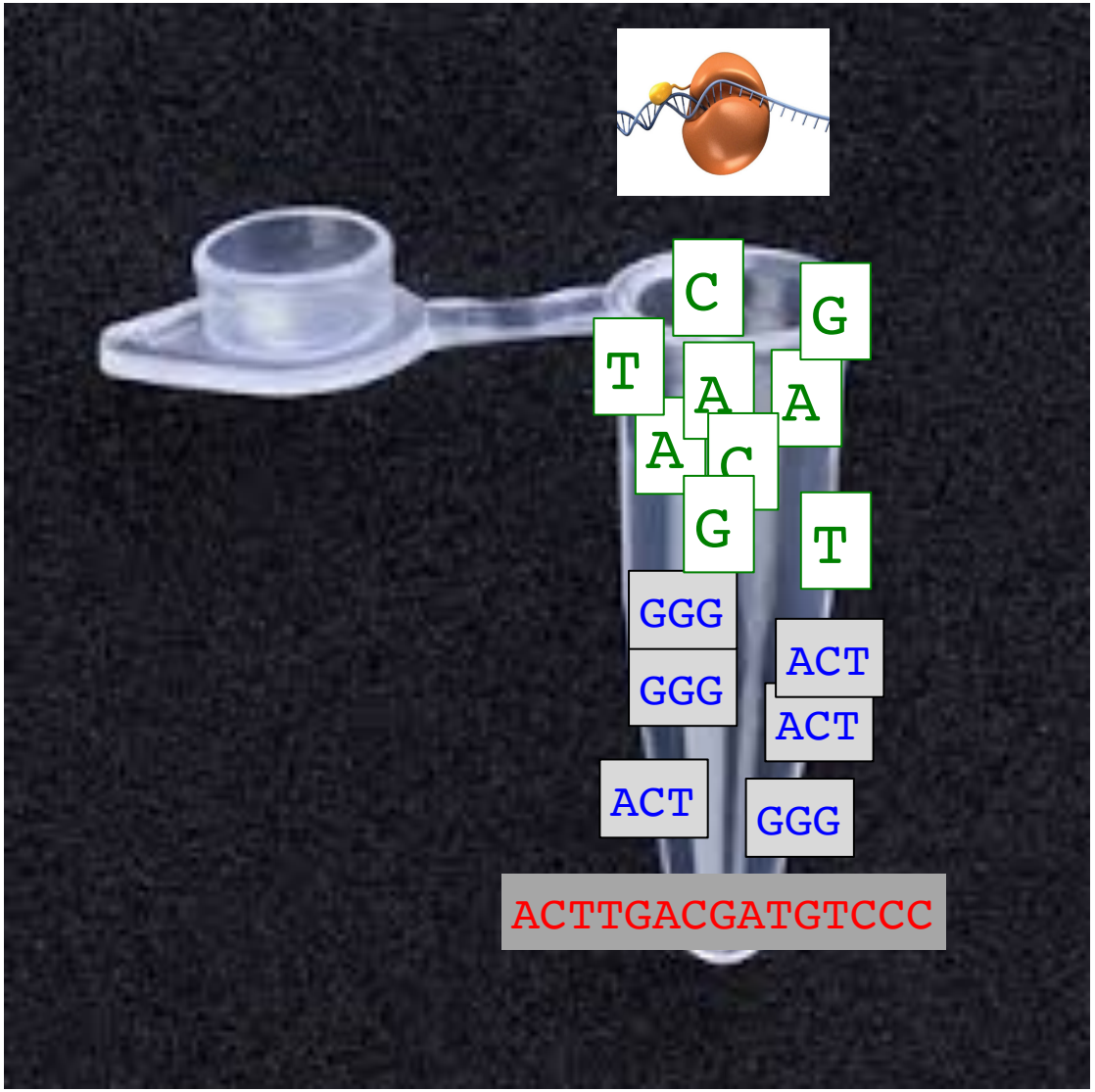
- Chain reaction: a reaction product catalyzes more reaction
- PCR: DNA fragment + primers + nucleotides + polymerase

→

2 DNA fragments + polymerase

Reaction builds more and more DNA fragments until you run out of primers or nucleotides.

The PCR Tube: where it all happens



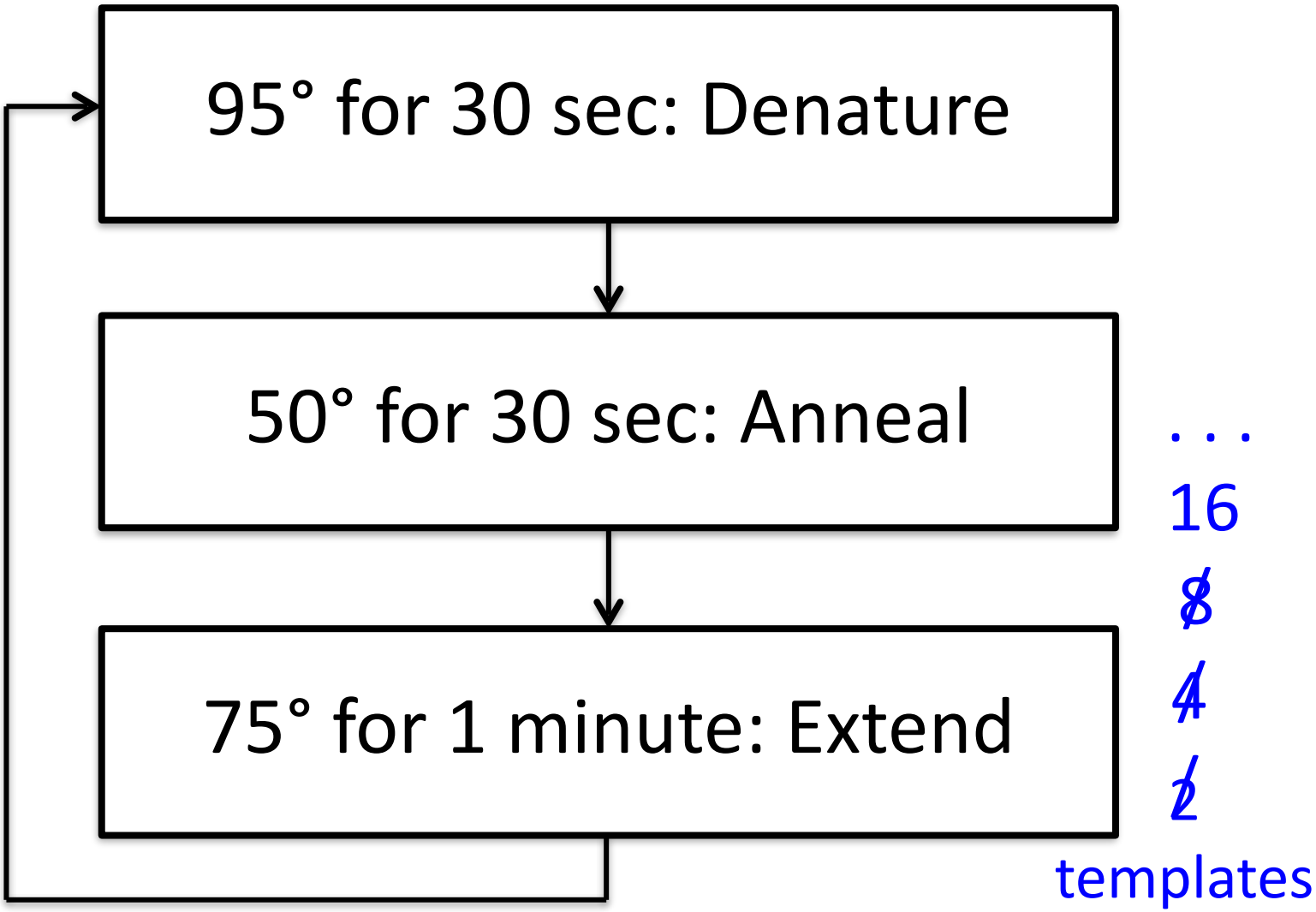
Add DNA template

Add many copies of primers

Add lots of As, Cs, Gs & Ts

Add DNA polymerase

Now just cycle the temperature
Temps & times are rough approximates

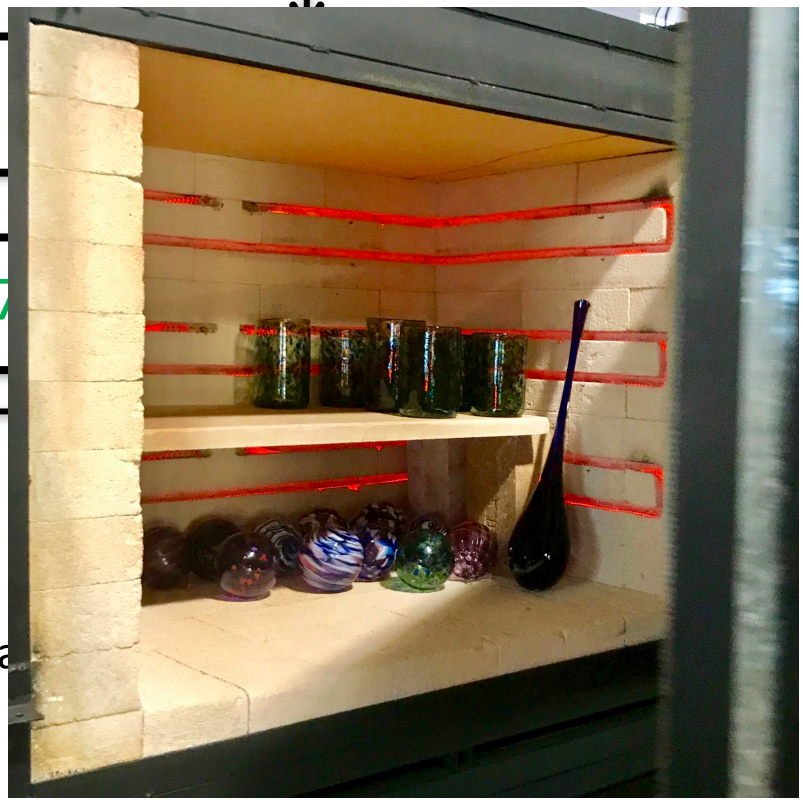


Nearly boiling

95° for 30 sec: Denature

Coldest

"Anneal" = to cool slowly

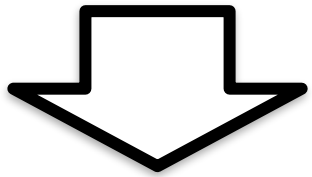


Cycle times are << shorter tha

PCR Cycle, Step 1: Denature (Separate the 2 DNA strands)

ACTTGACGATGTCCC

TGAACTGCTACAGGG

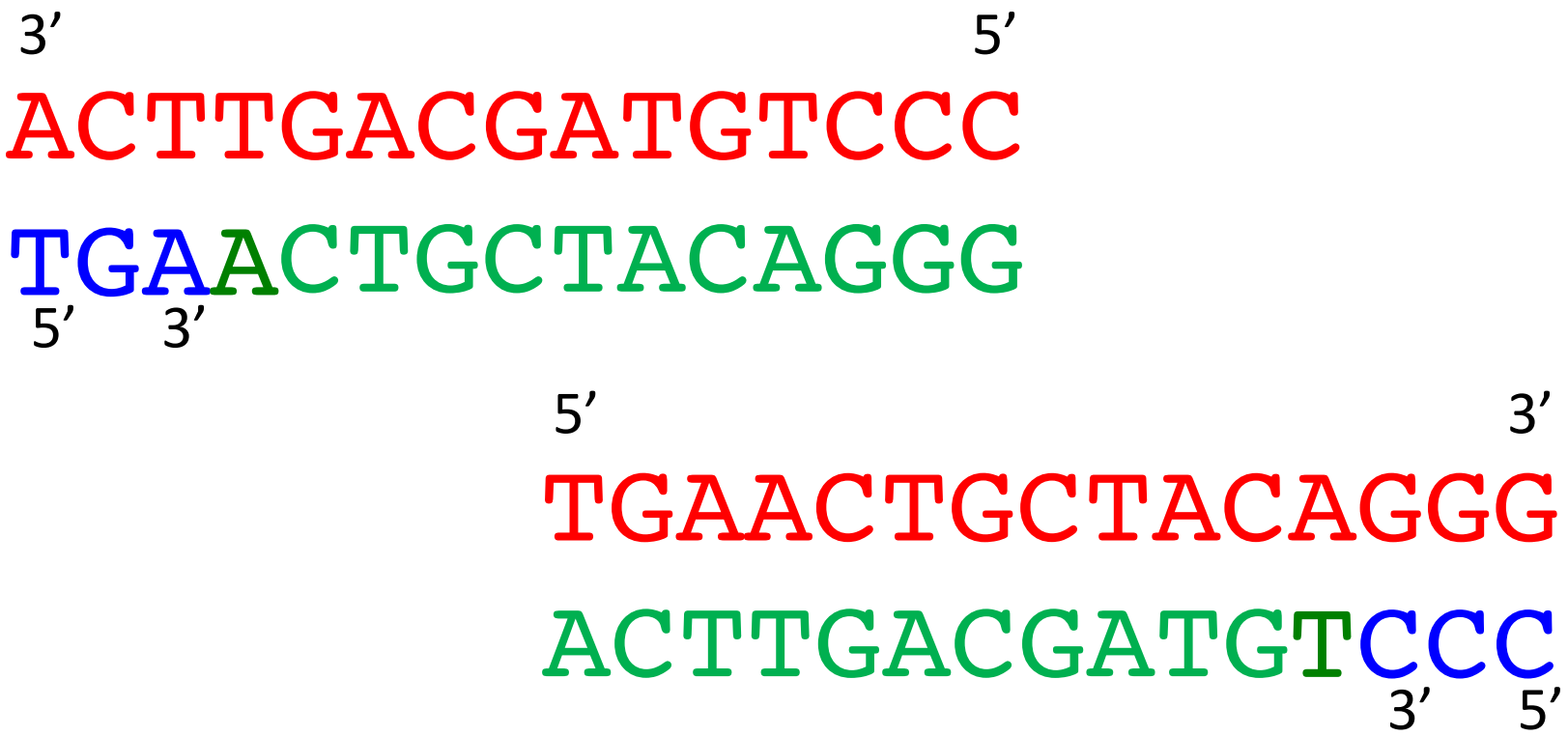


ACTTGACGATGTCCC

TGAACTGCTACAGGG

PCR Cycle, Step 2: Anneal (Attach primers to 5’ ends)

PCR Cycle, Step 3: Extend
(Free-floating bases extend primers at 3')



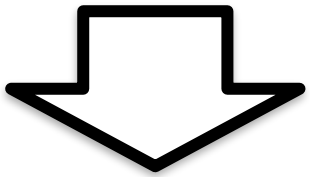
PCR Cycle, Back to Step 1: Denature

ACTTGACGATGTCCC

TGAACTGCTACAGGG

TGAACTGCTACAGGG

ACTTGACGATGTCCC



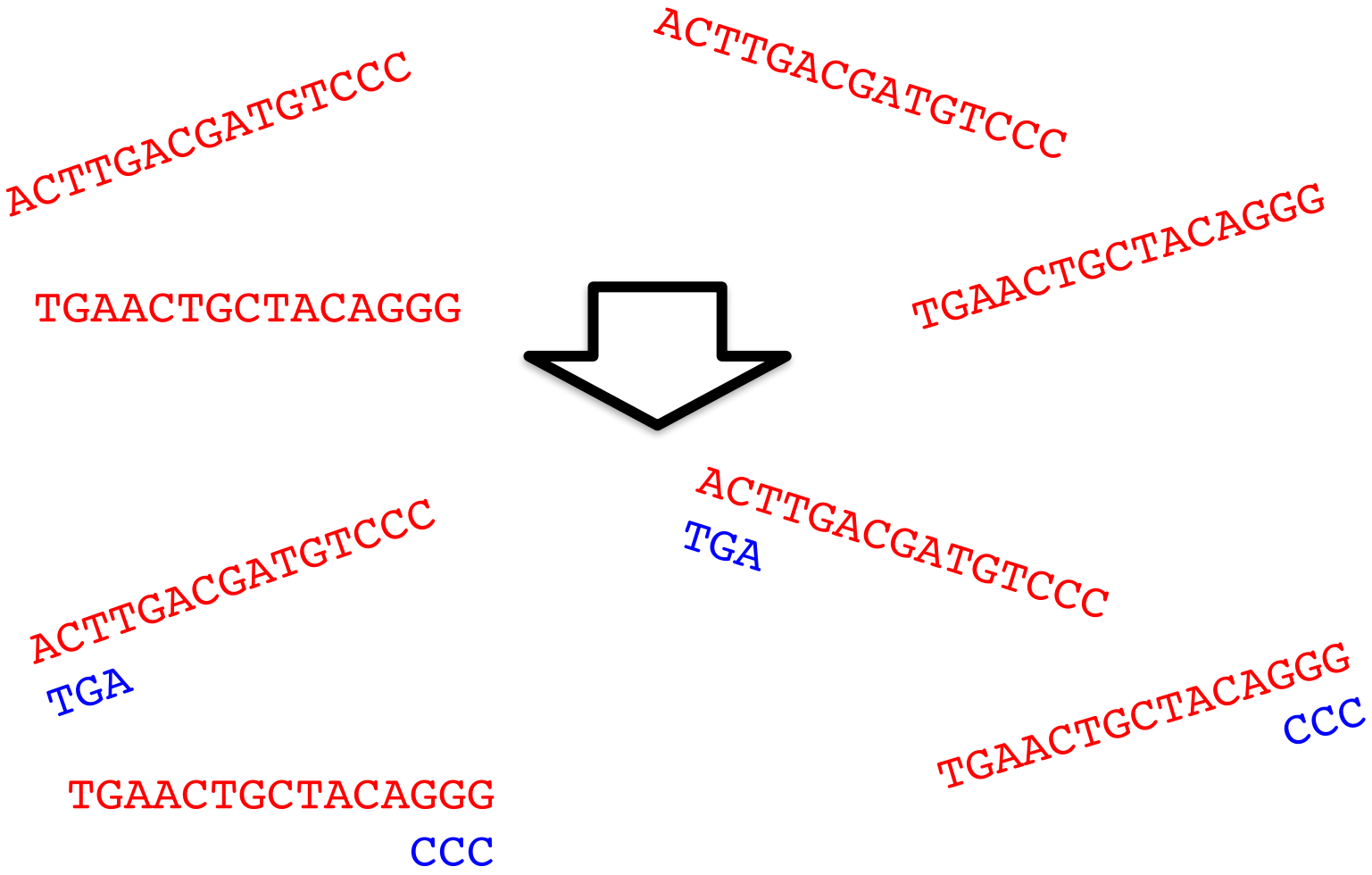
ACTTGACGATGTCCC

ACTTGACGATGTCCC

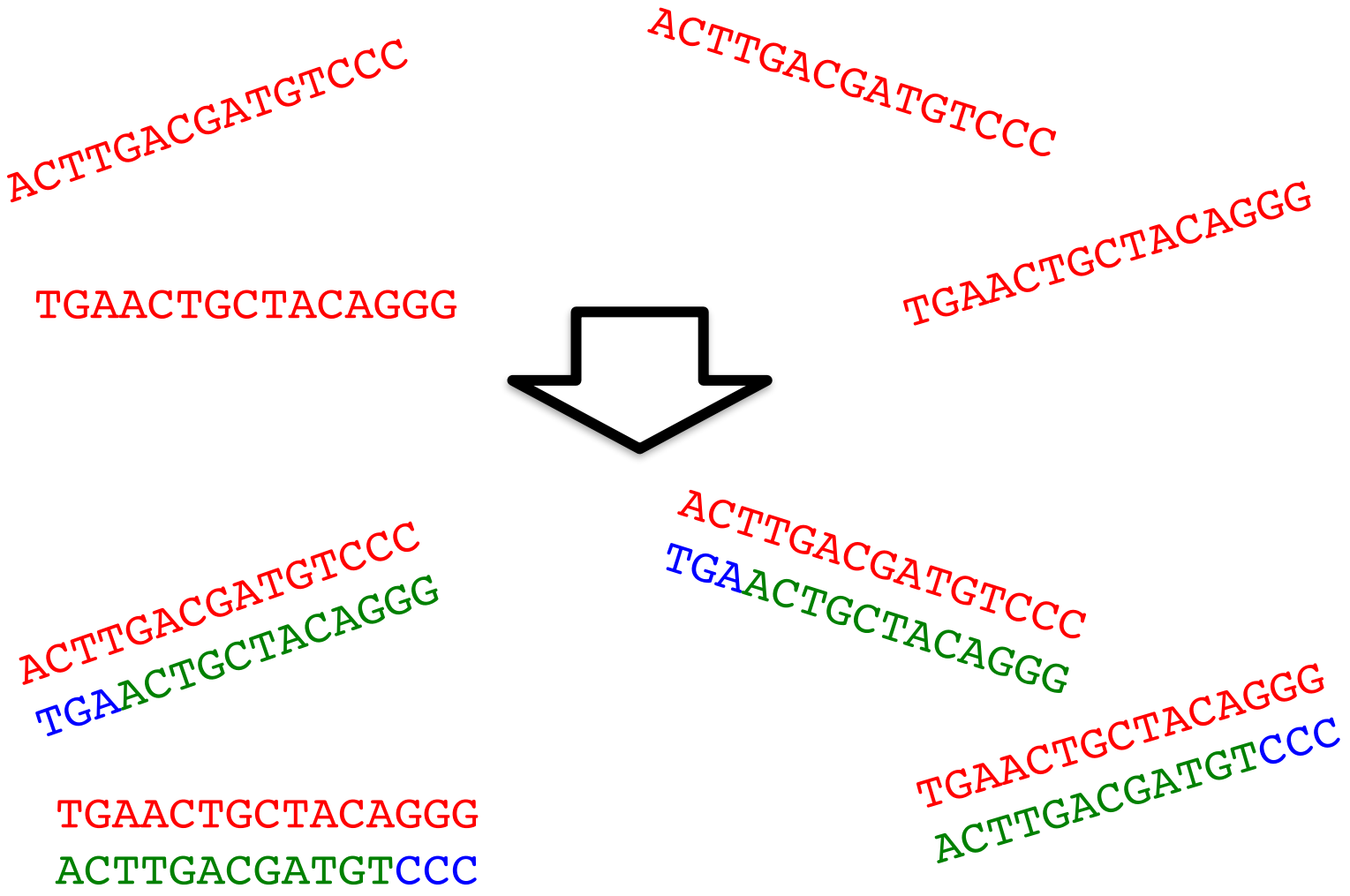
TGAACTGCTACAGGG

TGAACTGCTACAGGG

PCR Cycle, 2nd Step 2: Anneal



PCR Cycle, 2nd Step 3: Extend



Topics

- Amplification ✔
- 3 generations of sequencing
- Shotgun sequencing and assembly

Sequencing through the ages

1977

1990

Now

∞

1st Generation
Sanger Sequencing

2nd Generation
"NextGen"
Pyrosequencing

3rd Generation
Single Cell
Nanopores

You are here

Sequencing through the ages

1977

1990

Now

∞

1st Generation Sanger Sequencing

Assembly required

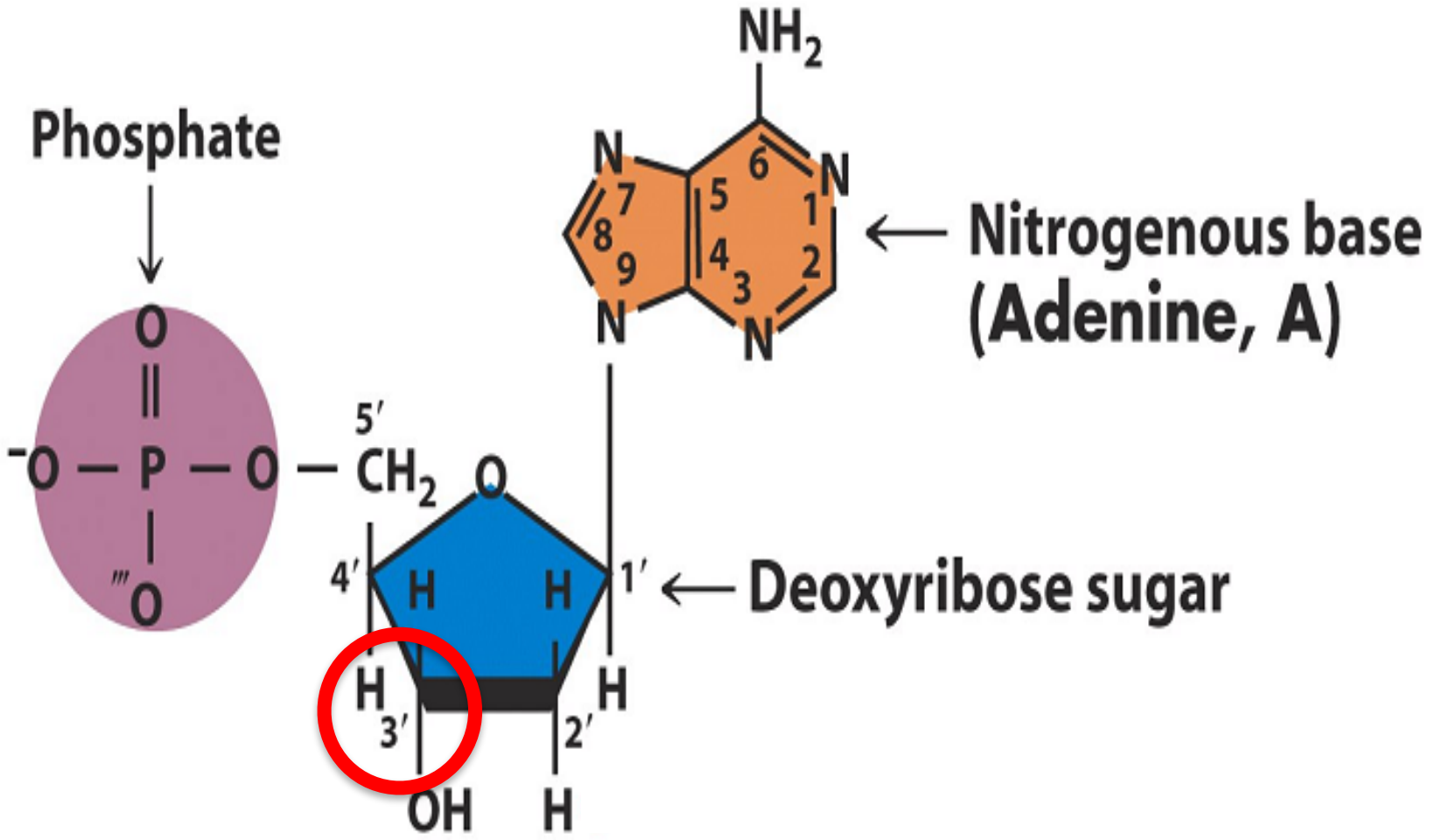
2nd Generation "NextGen" Pyrosequencing

Assembly required

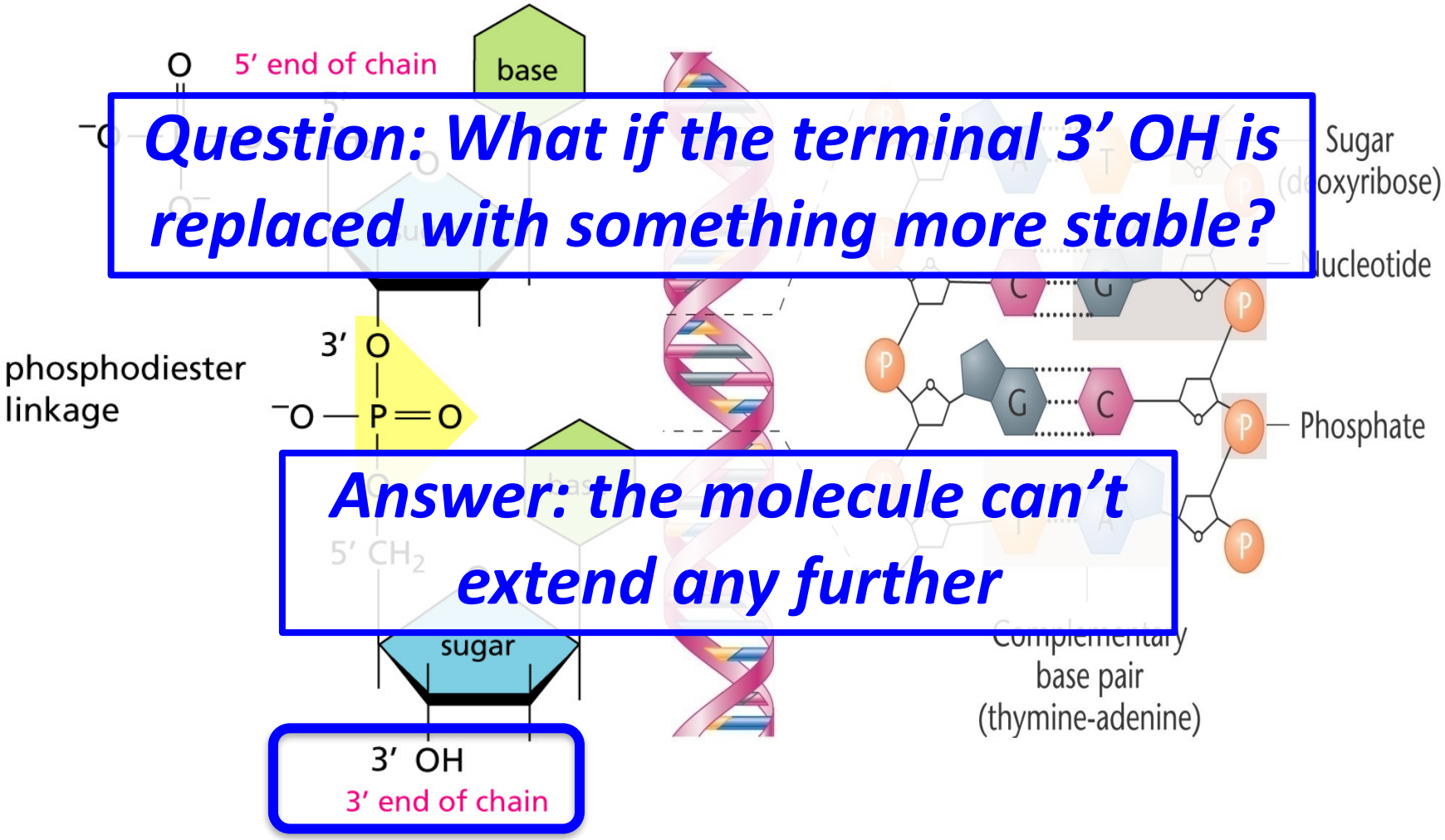
3rd Generation Single Cell Nanopores

Not much assembly required

The chemistry of sequencing: It's all about 3' engineering



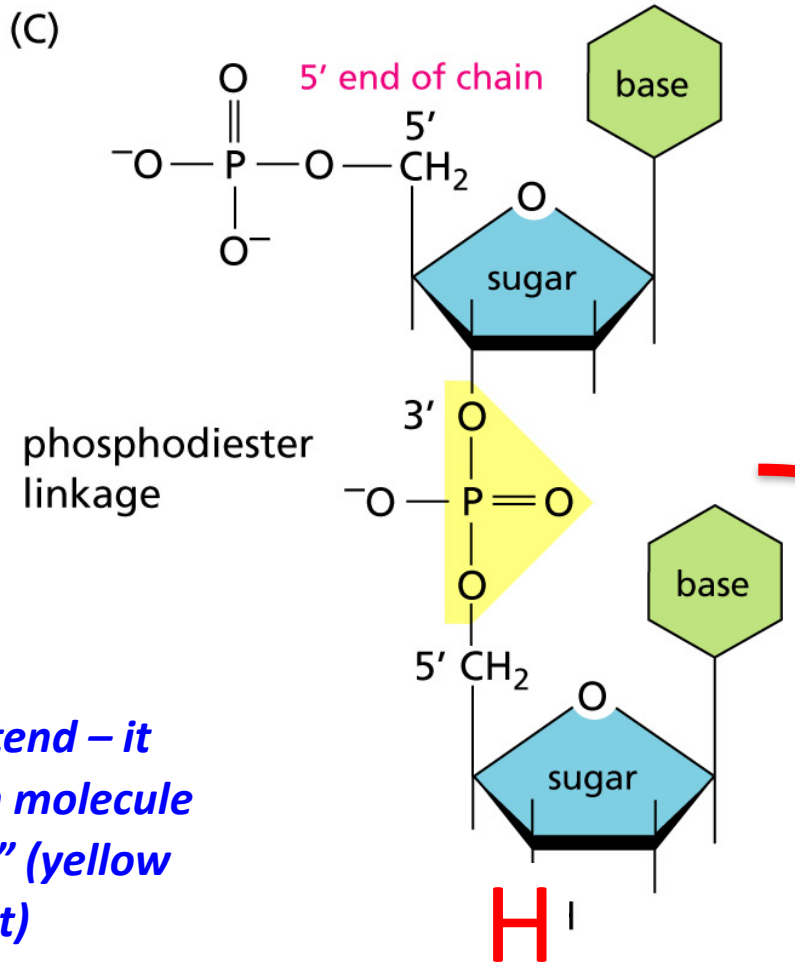
The chemistry of sequencing: It's all about 3' engineering



Sanger Chain-Termination Sequencing

- Developed ~1977 by Fred Sanger
  - 1918 – 2013
  - Won Nobel Prize twice
  - Only 3 others have ever done that: Curie, Pauling, Bardeen
- ddNTP

The chemistry of sequencing: It's all about 3' engineering

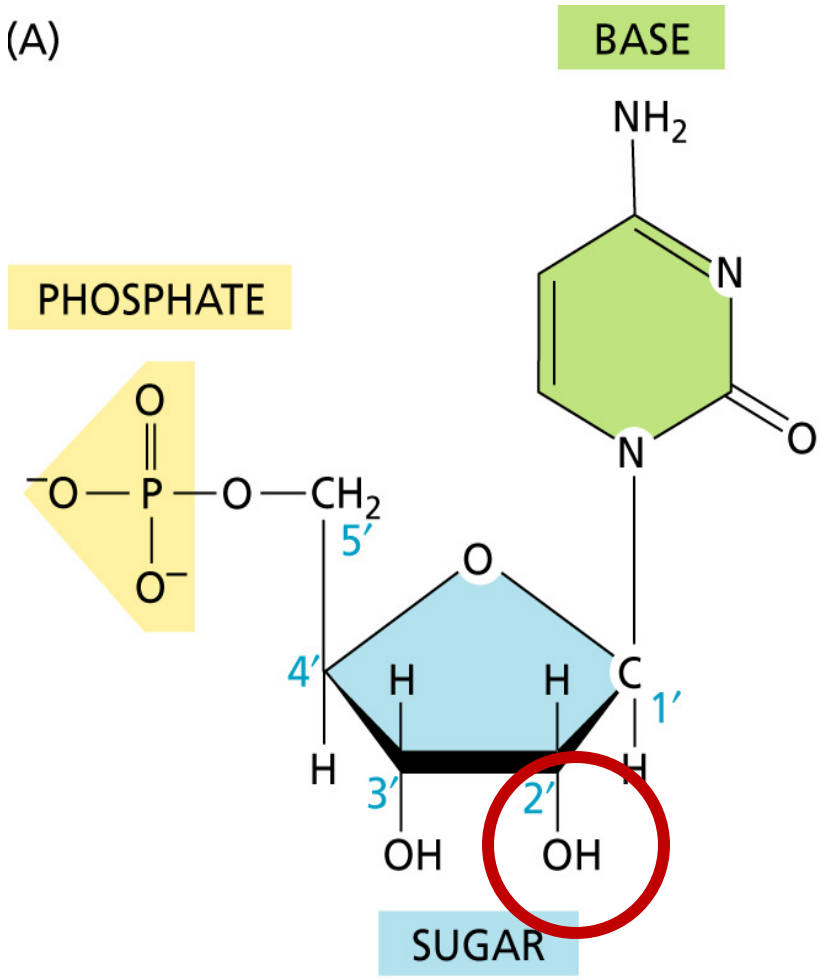


ddNTP

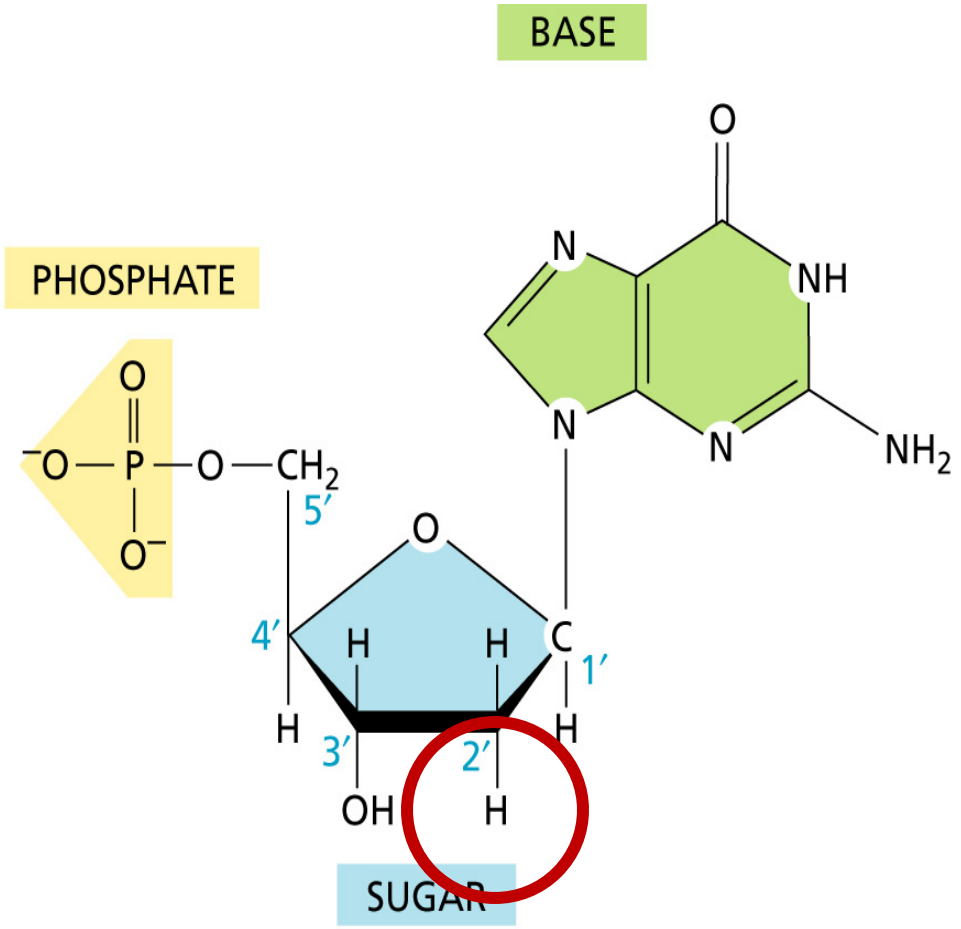
Last sugar can't extend – it needed the Oxygen molecule for attaching "glue" (yellow phosphodiester unit)

"dd" = "di-deoxy"
2 Oxygens are missing: from 2' and 3'

DNA sugar is “deoxy” with respect to RNA (at 2’ Carbon)

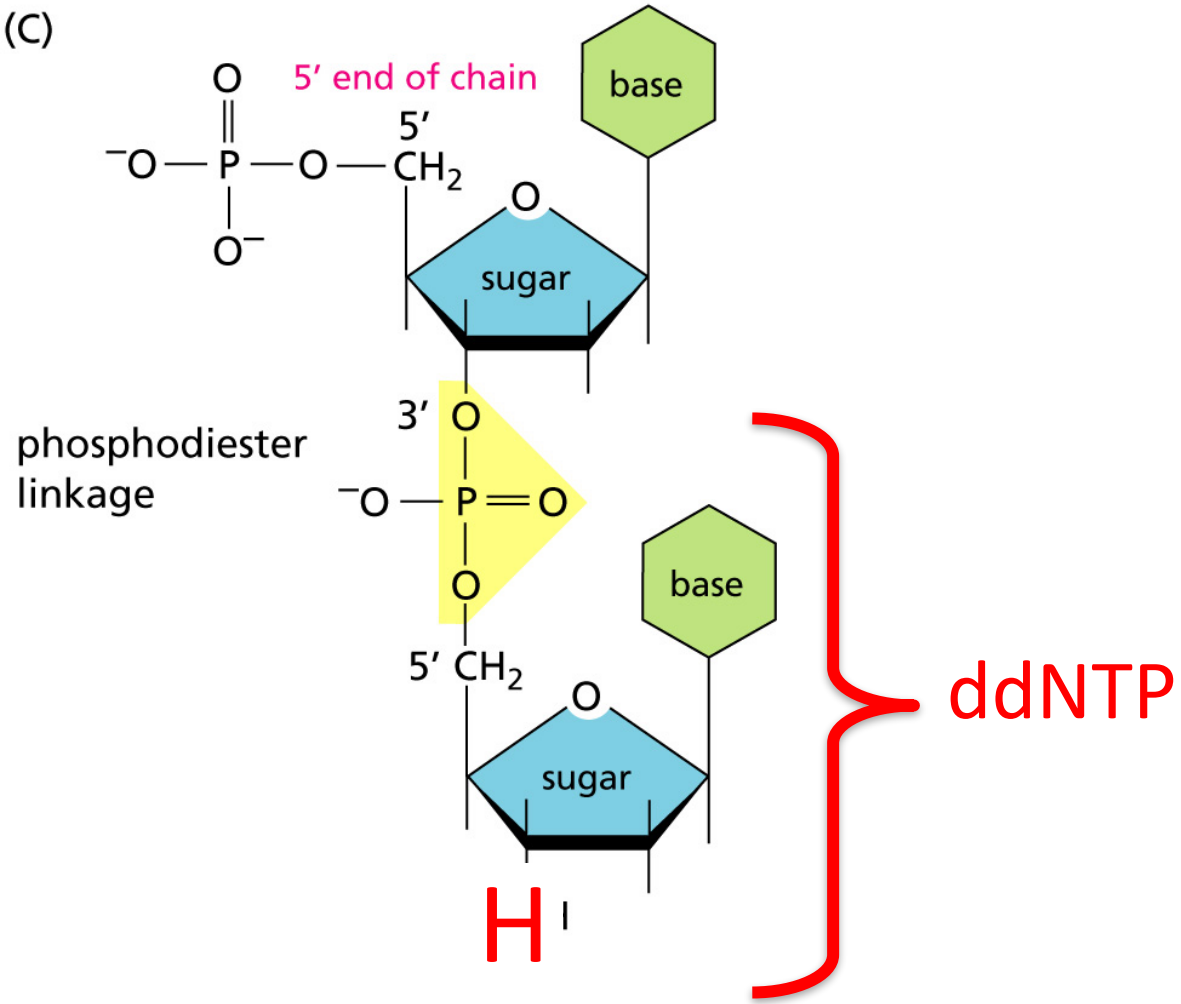


Pentose sugar present in **RNA**



Pentose sugar present in **DNA**

ddNTP is “di-deoxy” because 3’ carbon of sugar *also* lost an O



4 kinds of ddNTP

- Chain-terminating Adenine = ddATP
- Chain-terminating Cytosine = ddCTP
- Chain-terminating Guanine = ddGTP
- Chain-terminating Thymine = ddTTP

- ➔ “N” in “ddNTP” is a wildcard char
  - CS majors: think of it as “dd*TP”
- "TP" in "ddNTP" = TriPhosphate
  - We won't go into that chemistry

Sanger Sequencing: Preparation

- Combine
  - Amplified target gene copies
  - Forward primer for target gene
    - 1st n bases of coding strand
    - Will anneal to, and extend along, template strand
  - Individual nucleotides (As, Cs, Gs, Ts)
  - Adenine ddNTP (ddATP)
  - DNA polymerase

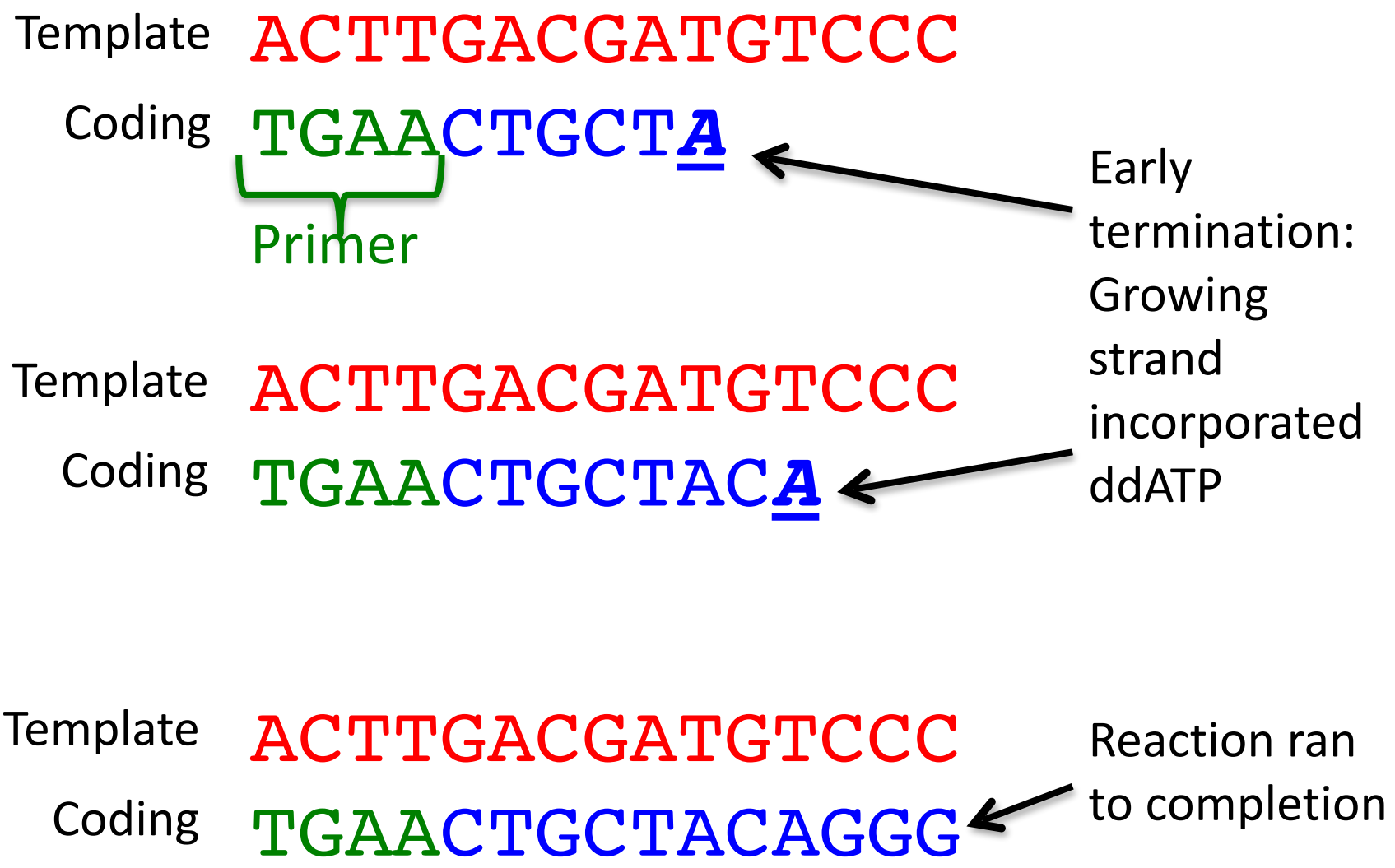
Sanger Sequencing

- Denature (separate) target DNA strands (like PCR)
- Primer will anneal to 3’ end of template strand (like PCR)
- Nucleotides in solution will extend from the primer (like PCR) until …

Sanger Sequencing

- Denature (separate) target DNA strands (like PCR)
- Primer will anneal to 3’ end of template strand (like PCR)
- Nucleotides in solution will extend from the primer (like PCR) until …
- Until the growing strand incorporates a ddATP, which terminates extension

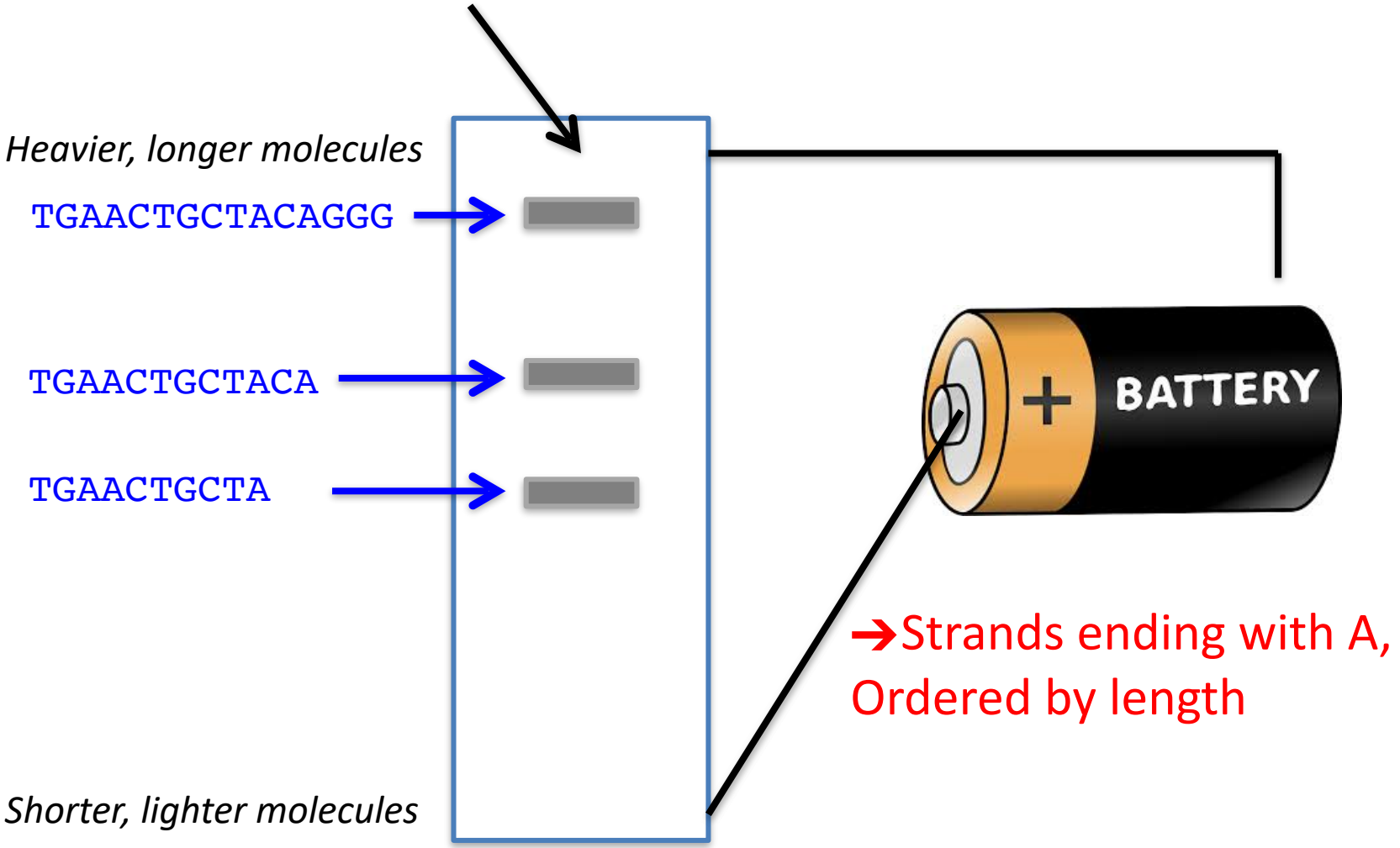
After the reaction:



After the reaction

- In roughly equal proportions we have:
  - Each sub-sequence of the target gene that begins with the primer and ends with A
  - The entire target gene
- Mass of each molecule is ~α number of bases
- Separate molecules by mass
  - Inject into gel (high-tech clear unflavored Jell-O) … sort of a solid, sort of a liquid
  - Apply negative charge to injection side of gel
  - Apply positive charge to far side
  - DNA is negatively charged, will migrate toward far side
    - Really slowly
  - Smaller lighter strands move faster
    - But still really slowly

TGAACTGCTA +
TGAACTGCTACA +
TGAACTGCTACAGGG



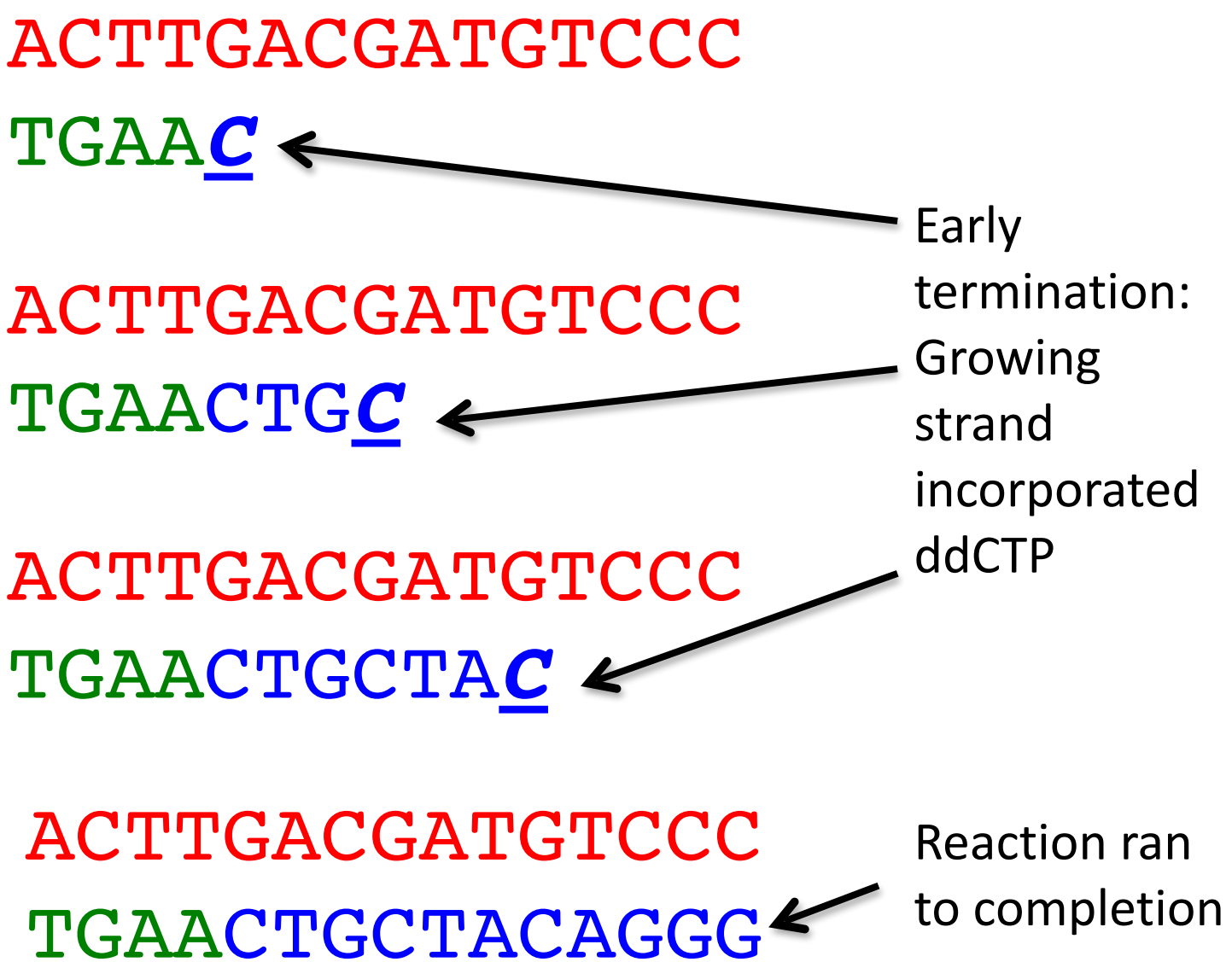
→Strands ending with A,
Ordered by length

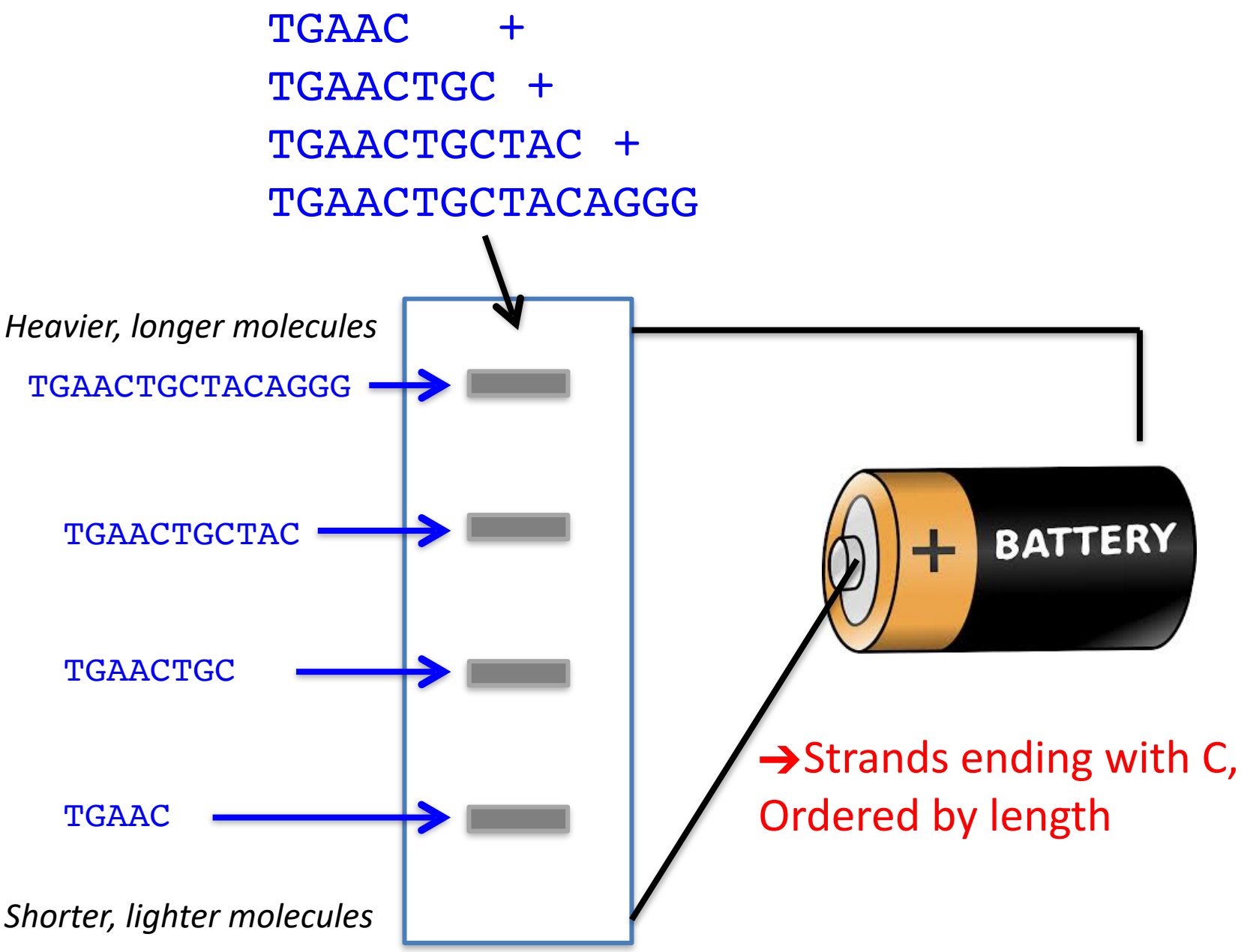
Sanger Sequencing: More Preparation

- Combine
  - Amplified target gene copies
  - Primer for 5’ end of coding strand of target gene
  - Individual nucleotides
  - Cytosine ddNTP (ddCTP)
  - DNA polymerase

The only difference (was ddATP, now C instead of A)

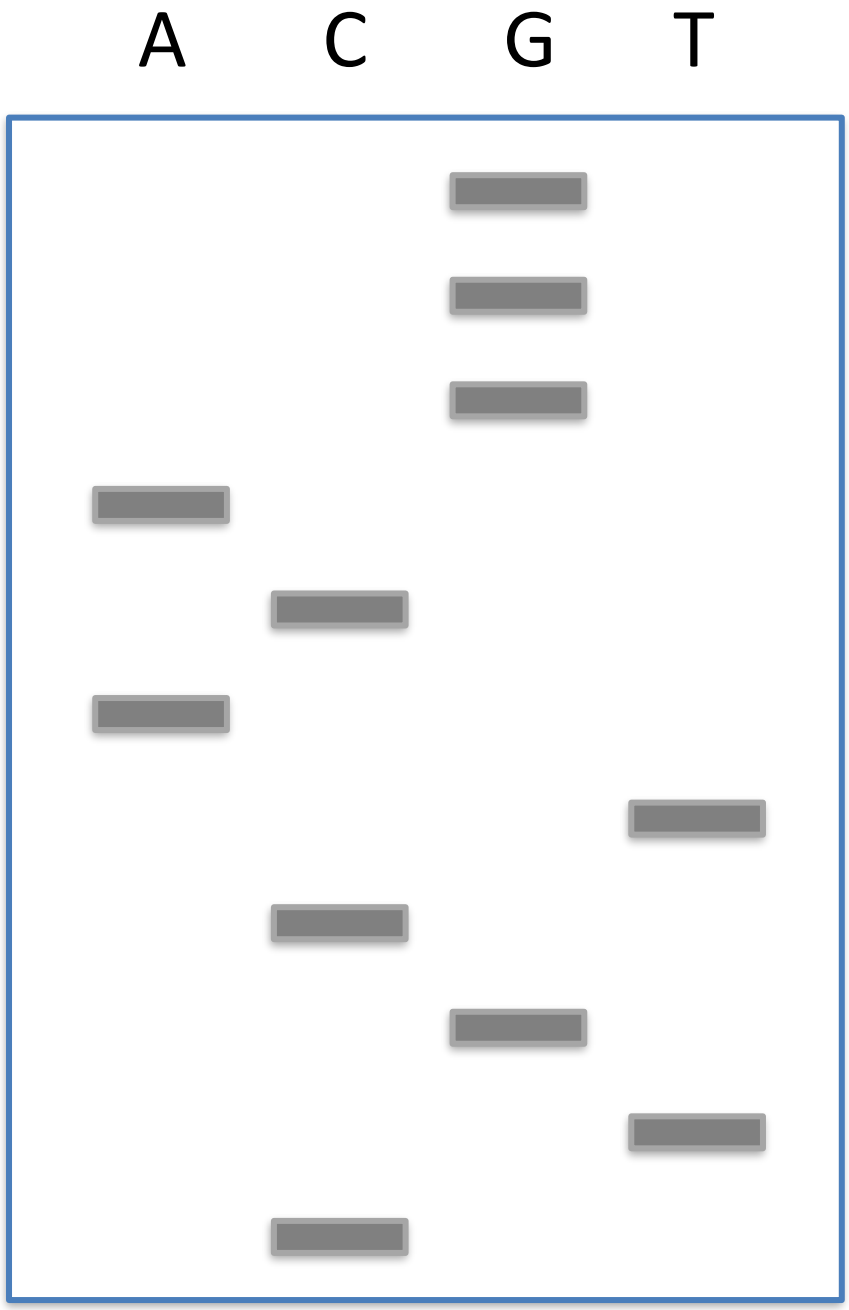
After the reaction:

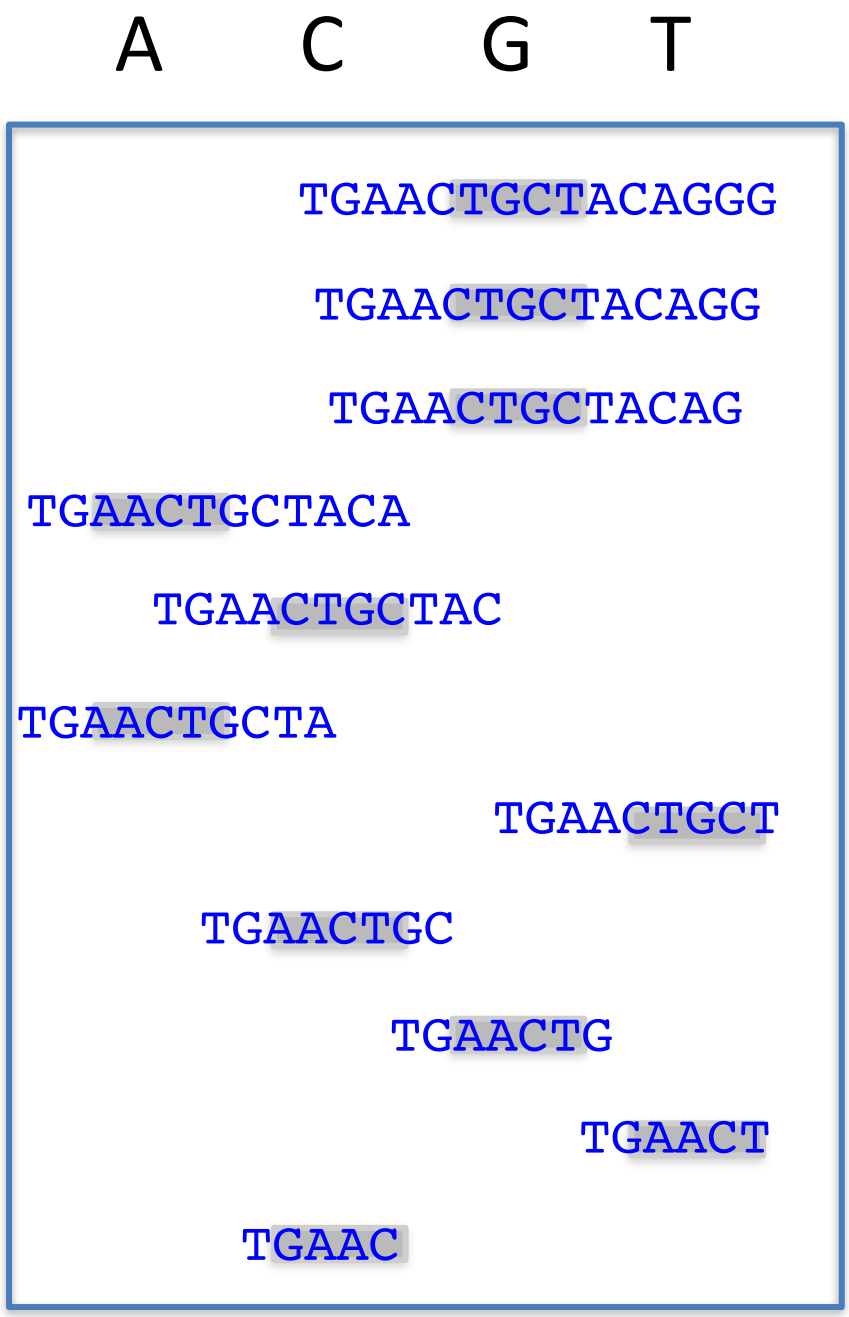



TGAAC +
TGAACTGC +
TGAACTGCTAC +
TGAACTGCTACAGGG
Heavier, longer molecules
TGAACTGCTACAGGG
TGAACTGCTAC
TGAACTGC
TGAAC
Shorter, lighter molecules
BATTERY
→Strands ending with C,
Ordered by length

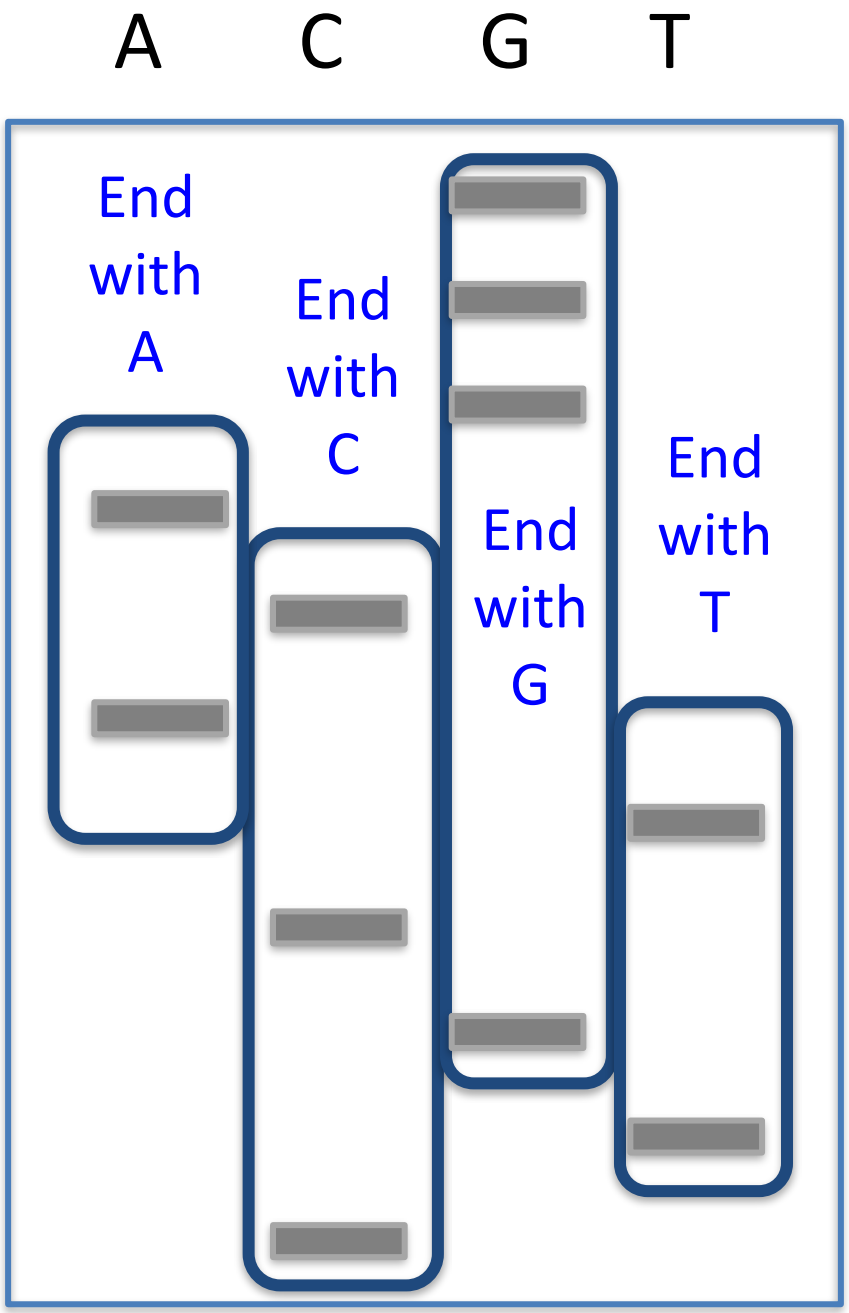
Well, actually …

- Do the Sanger rxn 4 times simultaneously, once for each ddNTP base
- Use a single gel, with a column (“lane”) for each base

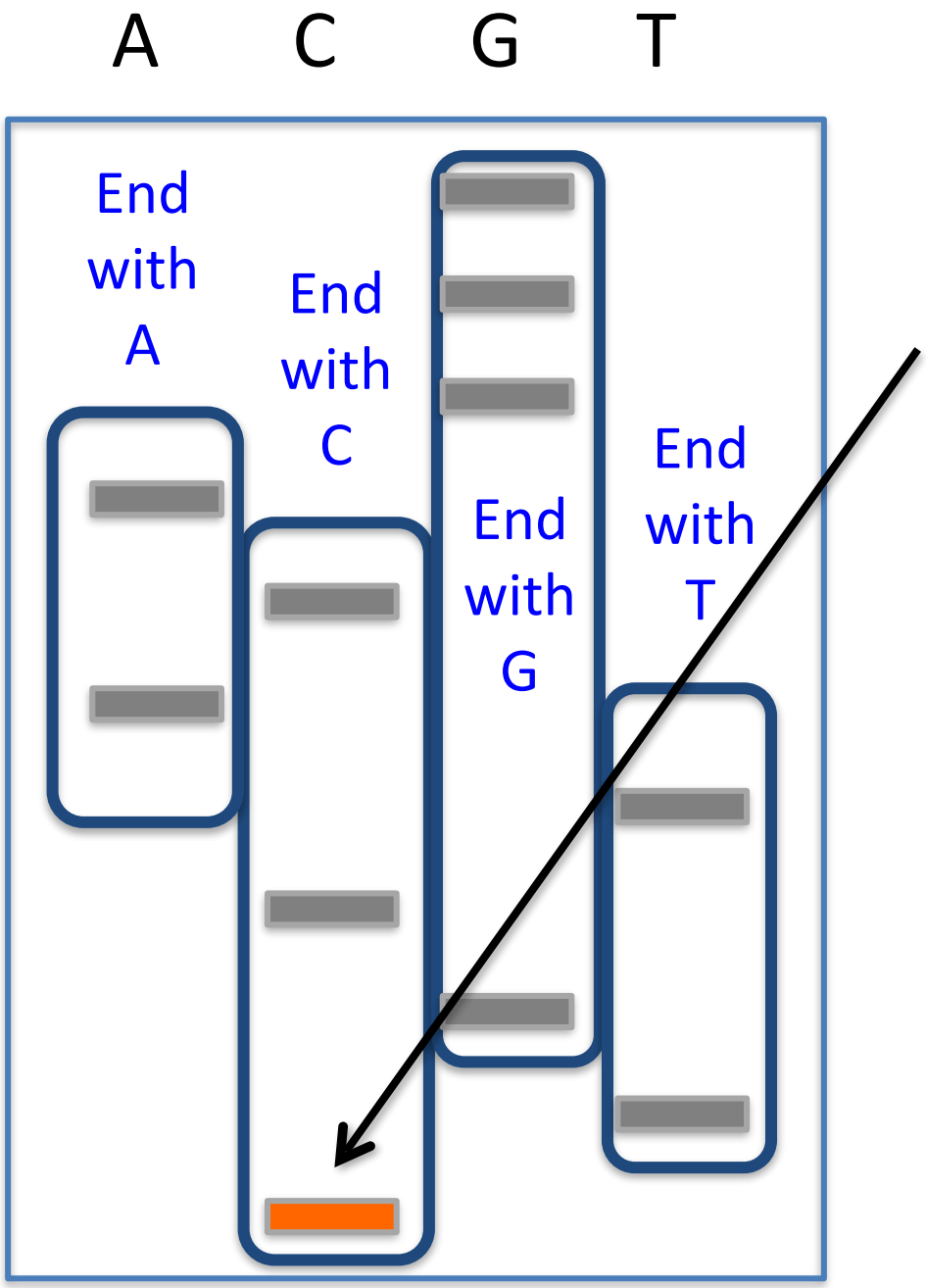

A
C
G
T



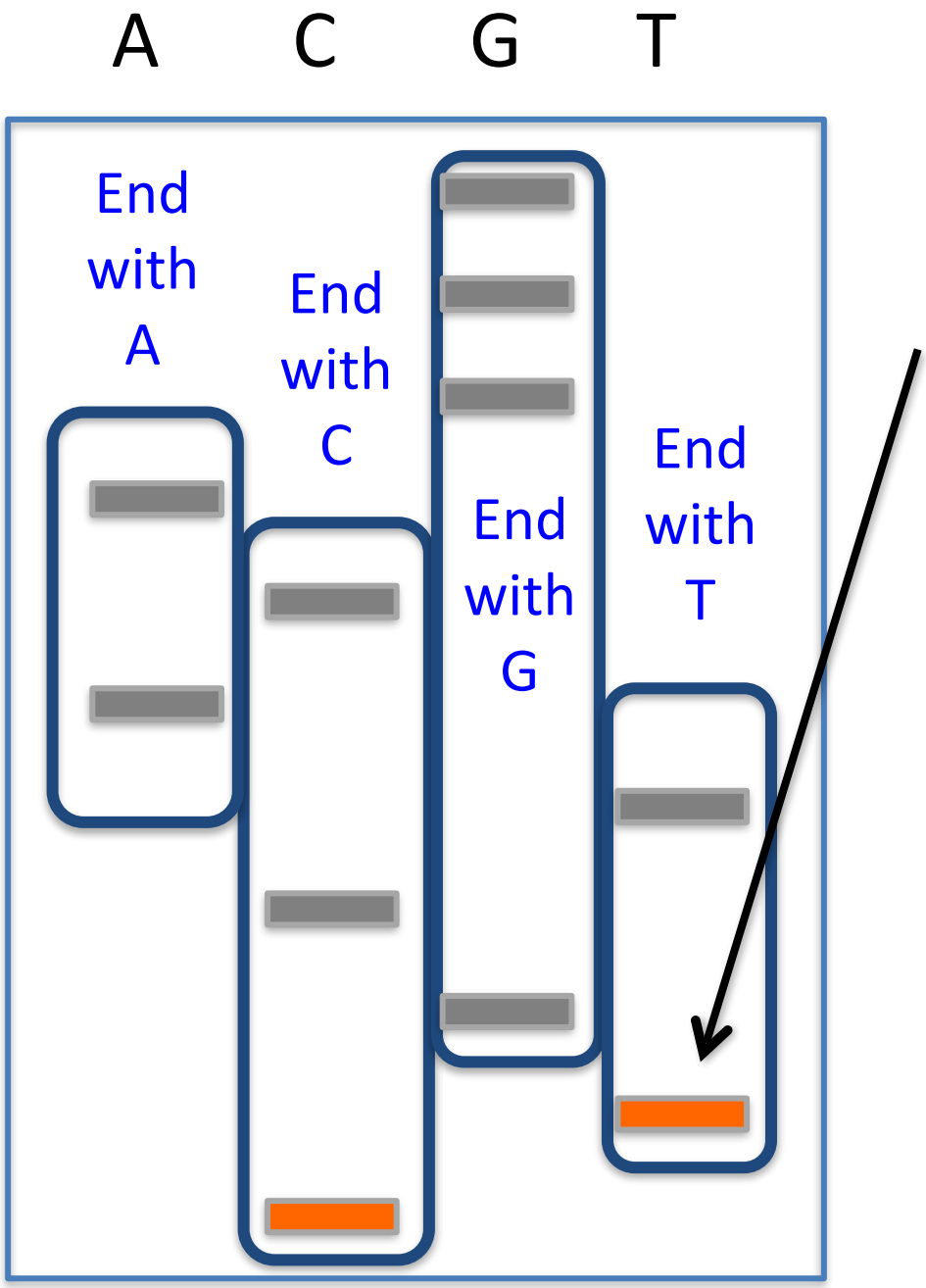
- Shorter toward the bottom
- Longer toward the top
- The very bottom: primer (TGAA) + 1 base
- The very top: entire target



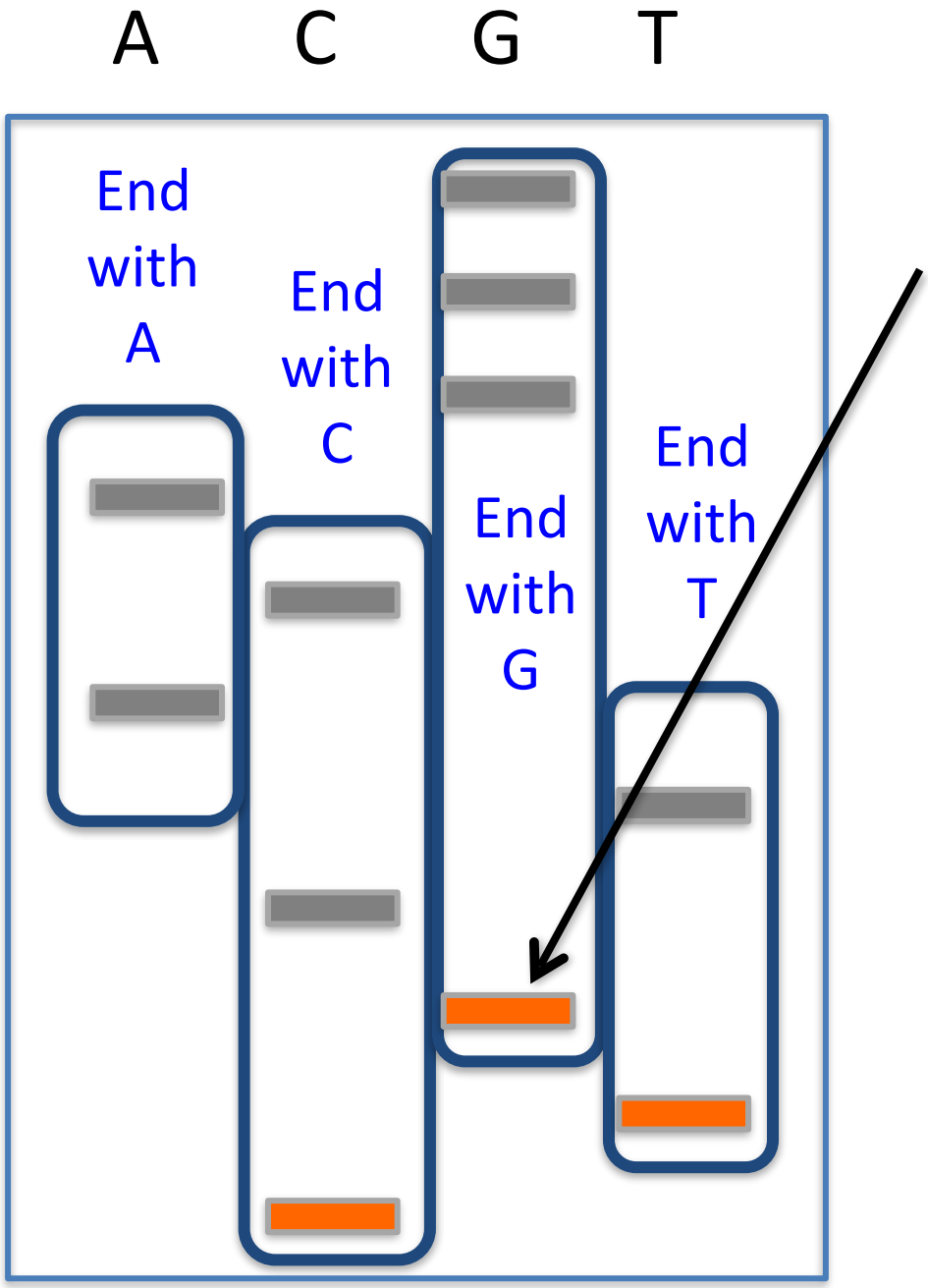
- All the investigator knows
- How to infer the target sequence?



- Shortest subseq (closest to bottom)
- Starts with primer
- Ends with C
- Has to be (primer)C

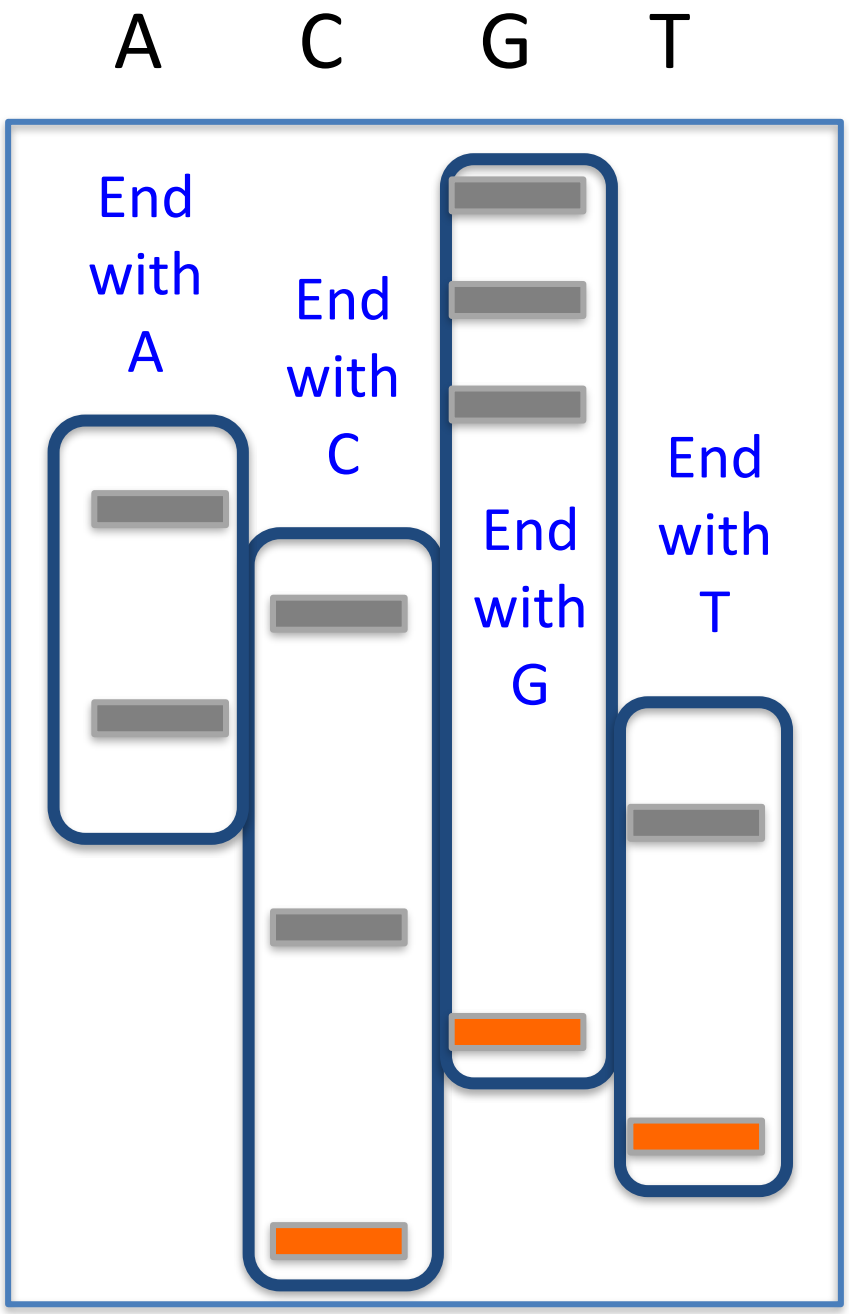


- 2nd-shortest subseq (2nd-closest to bottom)
- Starts with (primer)C
- Ends with T
- Has to be (primer)CT



- 3rd-shortest subseq (3rd-closest to bottom)
- Starts with (primer)CT
- Ends with G
- Has to be (primer)CTG

And so on…



Read up from bottom:

CTGCTACAGGG

Template strand ACTTGACGATGTCCC

Coding strand TGAACTGCTACAGGG

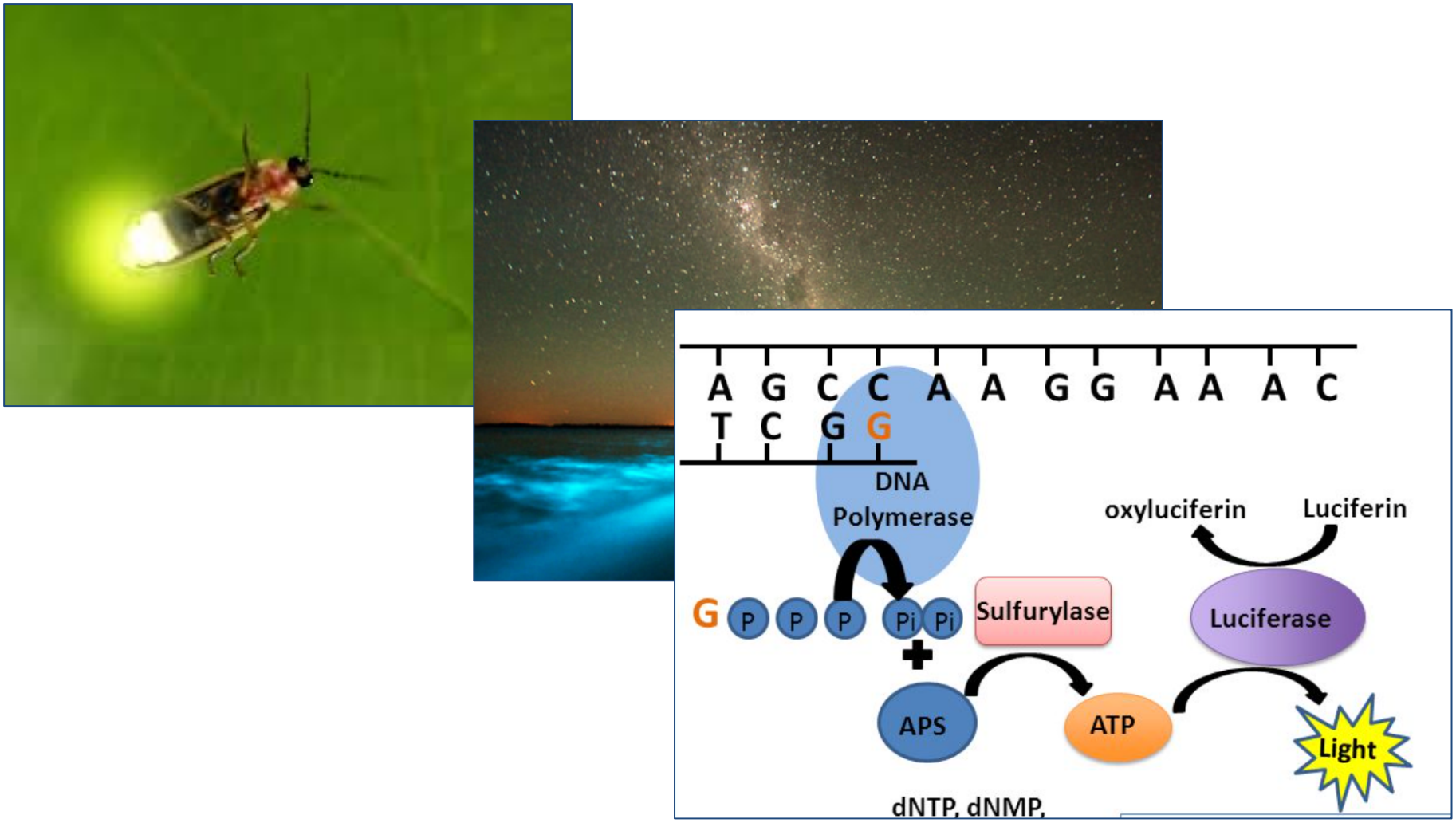
Primer = 1st n bases of coding strand, so gel readout is remainder of coding strand

Don’t complement, reverse, or reverse complement

That was great for 1977 but…

- 1-2 hours for bands to spread across gel
- Lots of human handling → risk of errors
- Macro scale → can't miniaturize/parallelize

1990: Pyrosequencing



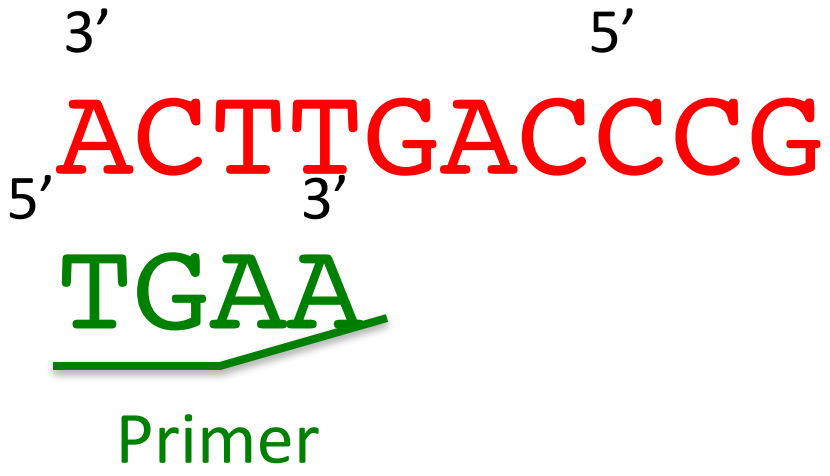
Pyrosequencing

- “Sequencing by synthesis”.
- Like Sanger, denature the DNA, add a primer, get information from how the primer extends.
- Like Sanger, use special free nucleotides with added triphosphorus:
  - dnATP, dnCTP, dnGTP, dnTTP
  - Not chain-terminating
  - No “normal” As/Cs/Gs/Ts
- Also use luciferase = the enzyme that causes bioluminescence.
- When a free nucleotide is incorporated into the extending strand, energetic phosphorus is released, causing luciferase to glow.

dnATP, dnCTP, dnGTP, dnTTP

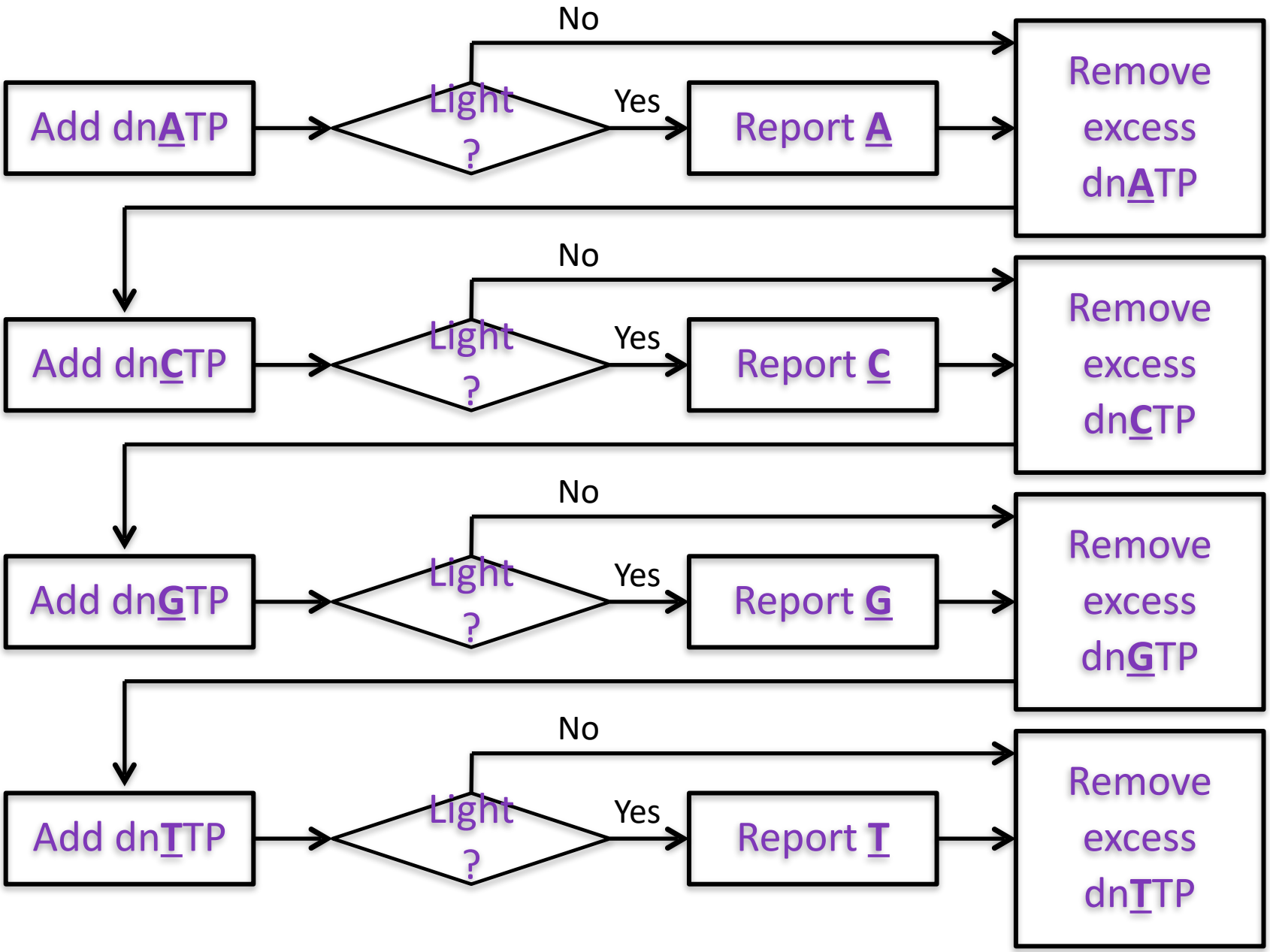
- Not chain-terminating.
- "TP" means tri<u>phosphate</u>.
- <u>Di</u>phosphate molecules are stable.
- Adding a 3rd phosphate to a diphosphate requires energy (like loading a nerf gun).
- Triphosphate molecules are less stable
  - Under the right conditions, release 1 phosphate and energy (like pulling the trigger of the nerf gun.

Pyrosequencing: Before 1st cycle



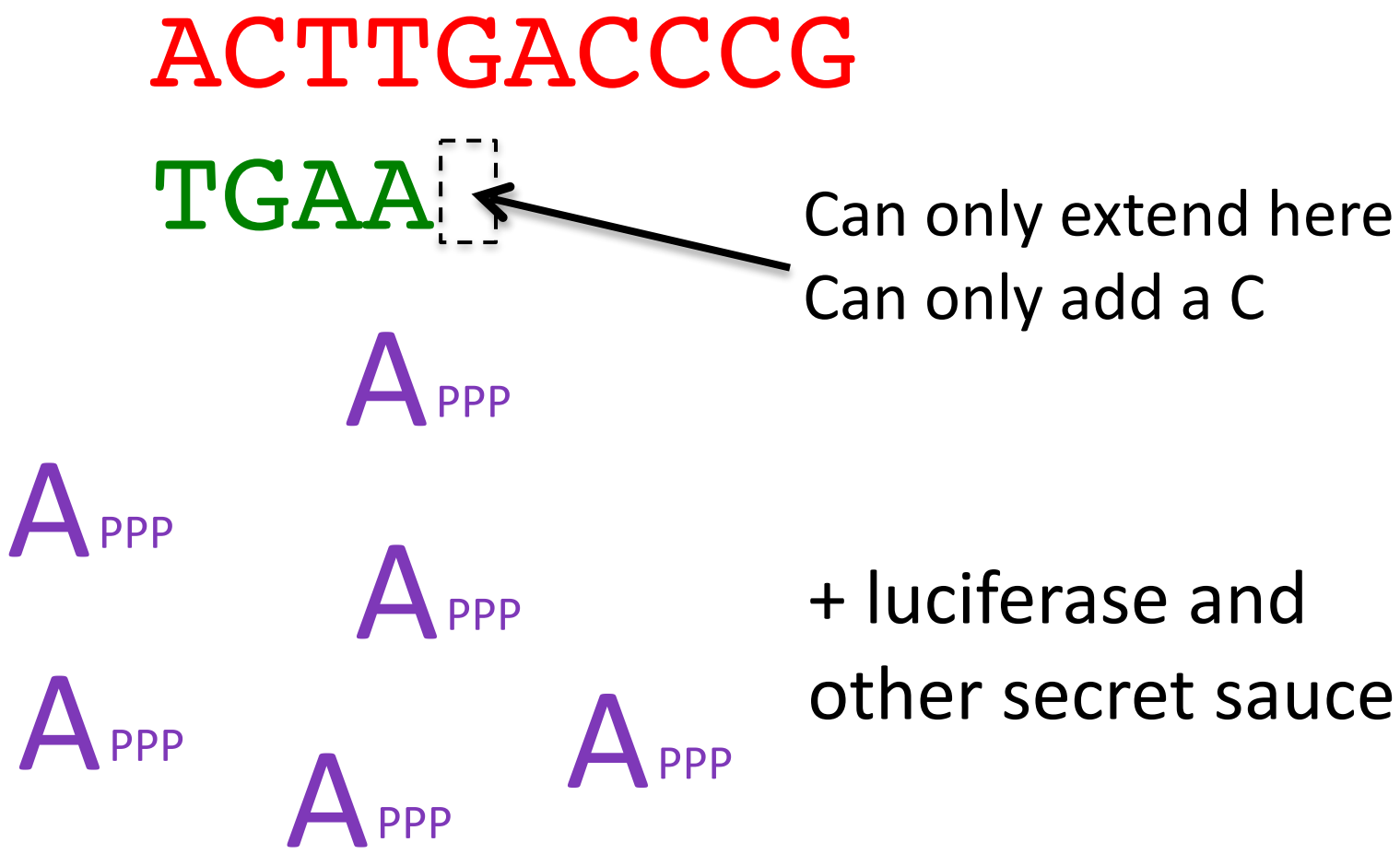
+ luciferin,
luciferase,
and
other secret sauce

Pyrosequencing Cycle



Pyrosequencing Cycle: add dnATP

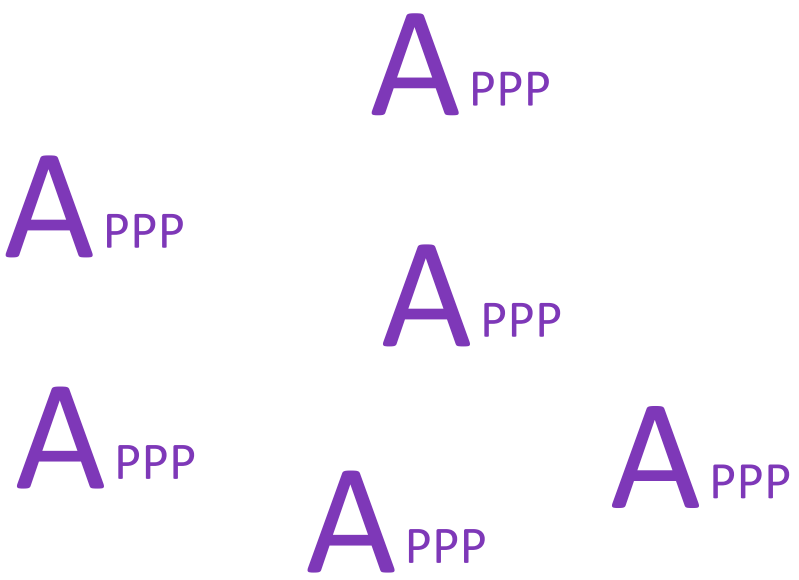
→ Nothing happens



Pyrosequencing Cycle: remove excess dnATP

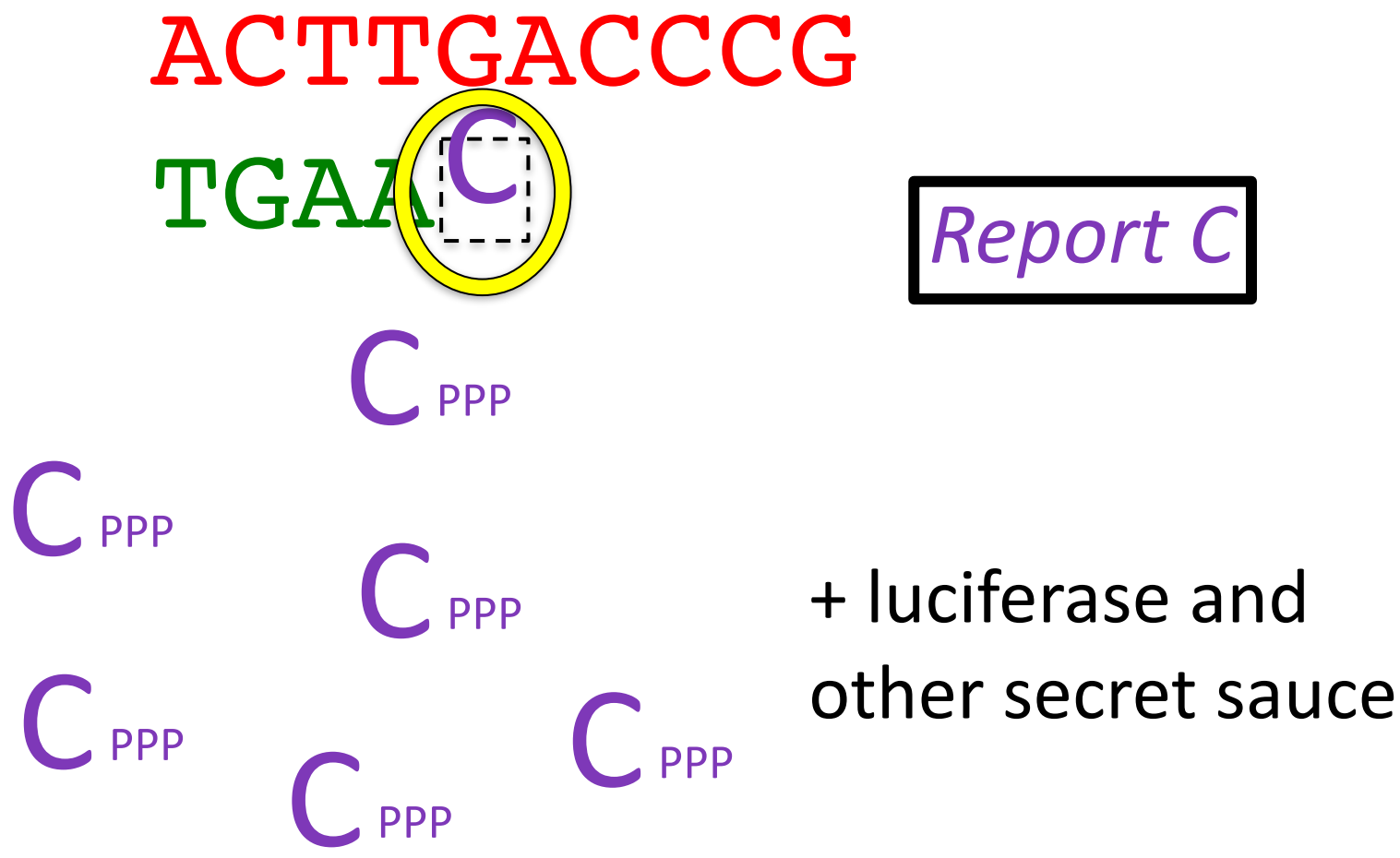
ACTTGACCCG

TGAA



+ luciferase and other secret sauce

Pyrosequencing Cycle: add dnCTP
➔Incorporates, causes a flash



Pyrosequencing Cycle: remove excess dnCTP

ACTTGACCCG

TGAAC

C PPP

C PPP

C PPP

C PPP

C PPP

C PPP

+ luciferase and other secret sauce

Pyrosequencing Cycle: add dnGTP

→ Nothing happens, remove it

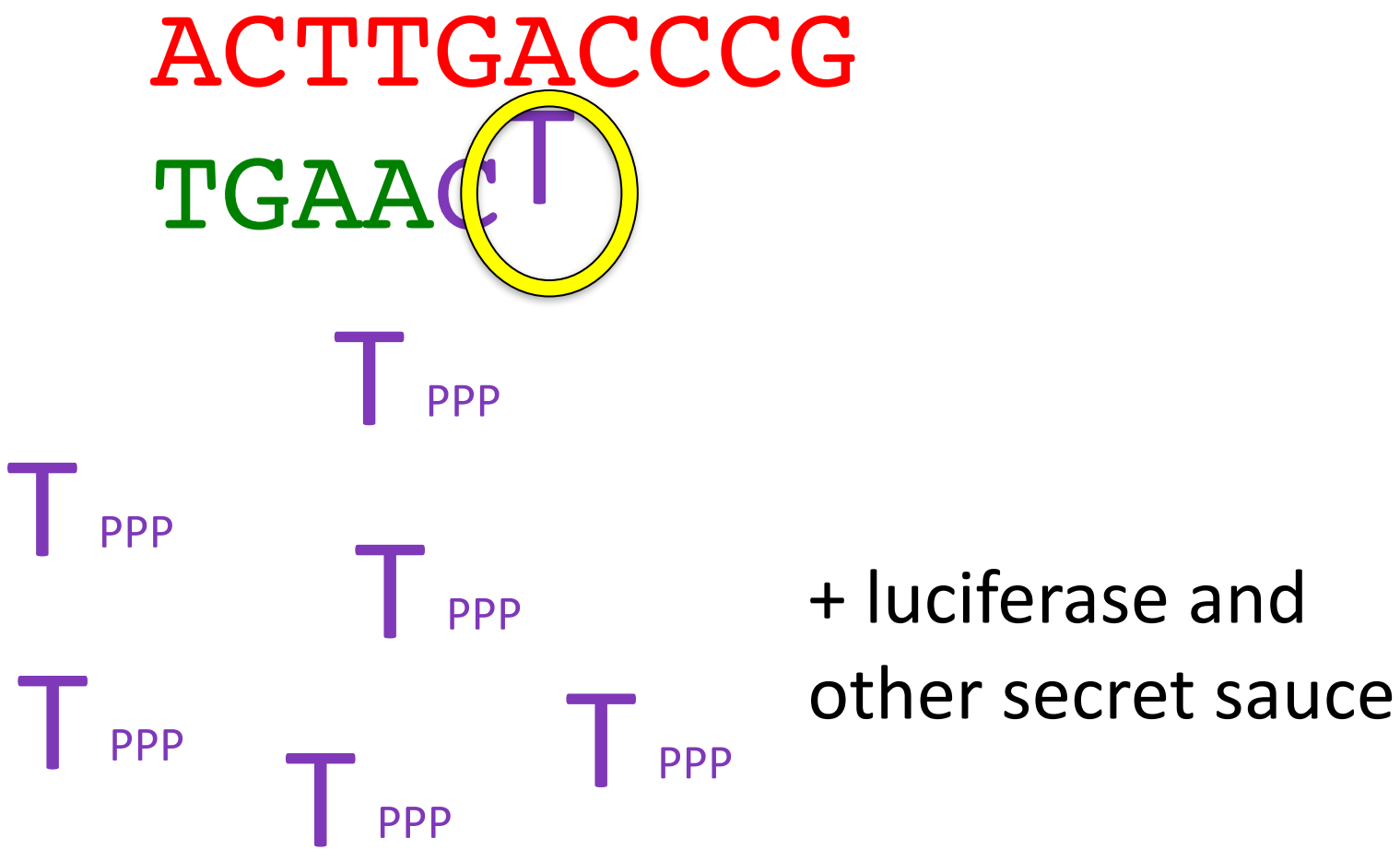
ACTTGACCCG

TGAAC

+ luciferase and other secret sauce

Pyrosequencing Cycle: add dnTTP

➔ Incorporates/flash/report T



Pyrosequencing Cycle: remove dnTTP, add & remove dnATP → nothing

ACTTGACCCG

TGAACT

Ok, you get the point,
But watch what
happens next …

+ luciferase and
other secret sauce

Pyrosequencing Cycle: add dnGTP

➔Incorporates 3 times, bright flash

G_{PPP} G_{PPP} G_{PPP} G_{PPP} G_{PPP}

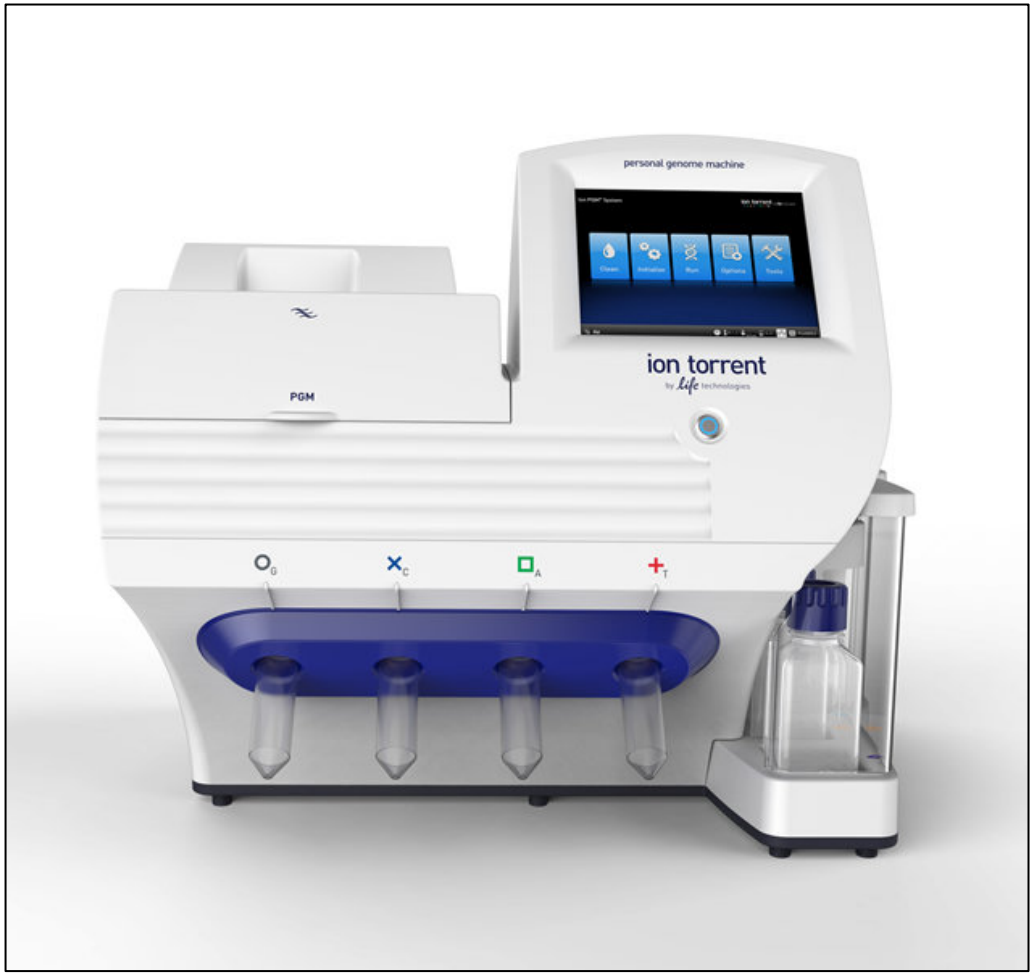
Report GGG

+ luciferase and other secret sauce

And so on to the end of the strand

- Fast: manufacturers claim 400 Mb in 10 hours (with parallelization)
- Small: light flashes detected by miniature photoelectric cells
- Software interprets flashes, converts to A/C/G/T
- Output is a fastq file

ION Torrent PGM: $80,000

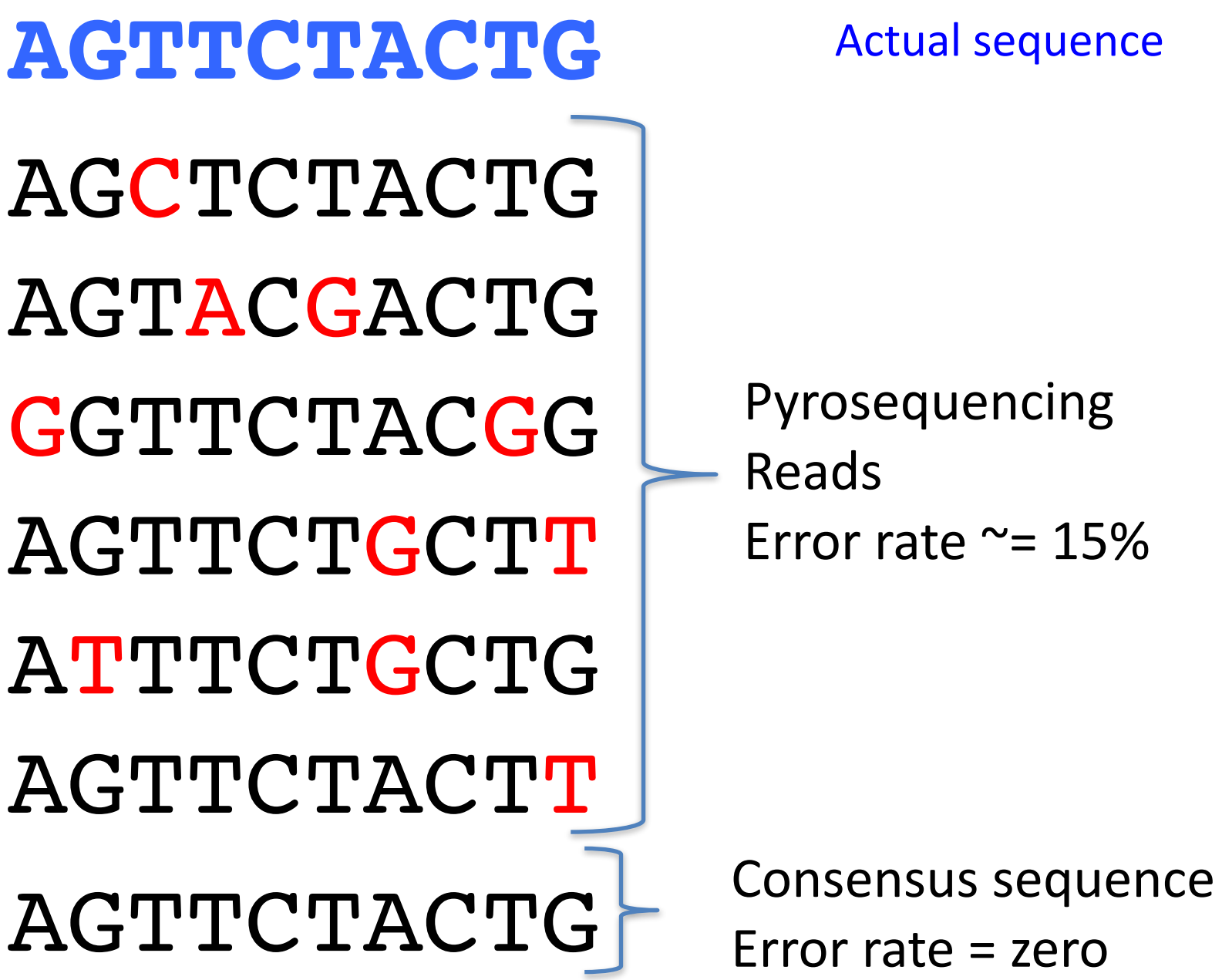


The 4th law of thermodynamics

(As formulated by former engineering consultants like me)

Pyrosequencing is error prone

- Flashes are brief
- Photodetectors are imperfect
- Hard to distinguish XXXXX from XXXXXX
- It's cheap and it's fast, but its quality is lower than Sanger sequencing
- → Use PCR to quickly/cheaply make lots of copies, which are quickly/cheaply sequenced
- → Trust the majority

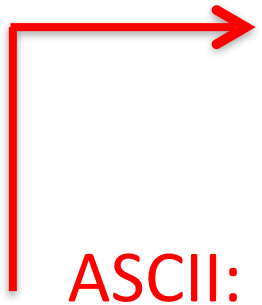

AGTTCTACTG
Actual sequence
AGCTCTACTG
AGTACGACTG
GGTTCTACGG
AGTTCTGCTT
ATTTCTGCTG
AGTTCTACTT
Pyrosequencing
Reads
Error rate ~= 15%
AGTTCTACTG
Consensus sequence
Error rate = zero

Sequencing devices estimate the quality of every base they report

- Users can ignore data whose quality is too low
- “Too low” varies with what you’re doing
- Quality is a single character
- Sequencers output files in “fastq” format
  - 4 fields per read
    - Unique identifier (“defline”), starts with @
    - Nucleotide sequence
    - +
    - Quality sequence, same length as nucleotide sequence

Fastq record: toy example

```
@This is read #1
ACGTACGTTGACTAGC
+
7887MN#+;;,87837
```



ASCII: Hard for humans to interpret

Fastq example from the real world

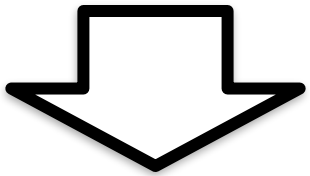
```
@HWI-M01367R:96:000000000-ACF47:1:1101:15126:1708#0/1
GTACACACCGCCCGTCGCTCCTACTGATTTCGAGTGGTCCGGTGAACCATTTGGACCGGCGCC
GCCTCGTGCTAAGCCGGGAAGTCTAGTAAACCACATCACAGAGAGGAAGGAGAAGTCGTAACA
AGGTTTCCGTAGGTGAACCTGCAGAAGGATCAAGATCGGAAGAGCACACGTCTGAACTCCAGT
CACGTGAAACGATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAAACAACATAAAGAAGC
+
BBBBBFFBBBBBGGGGGG?
GGGHHHHHHHHHHGHEHGHGHGGAEFHHHHHHHHGGHHGGGGGGGGGGGGHHGGHHHHHGGGG
GGFFHHHFGHFHHHGHHHHHHHHHGGGHHGGHHHHHFGGFFGGHHFHGFHHHEHFDD0DHHHH
HHHHHGHHHHHHHHHHHHE:E?
DGGGGGGGFGGGGGGGGGGGGGGFFFFEEFFFFFEAFFFDFFFFFACFFFFFFFFFFFBFF
FFF##################
```

Quality trimming … ***not the only way to do it!***

- 1st step of many analyses
- Determine quality threshold “q”
- Keep records where at least 95% of qualities are >= q
- Convert retained records to fasta format
  - Only 2 lines per read: Defline and nucleotide sequence
  - Defline starts with “>”, not “@”
  - No longer need quality info
- Done with fastq, analyze fasta, trust quality

Fasta record: toy example

```
@This is read #1
ACGTACGTTGACTAGC
+
7887MN#+;;,87837
```



```
>This is read #1
ACGTACGTTGACTAGC
```

ASCII ("ask-key"): How computers store text

- Computers can only store integer numbers in a limited range.
- They can't store anything else.
- Wait, what? What about text, sound, pix, video? What about non-integer numbers?
- Computer knows a certain number in memory represents e.g. a character. There's a standard universally accepted way to interpret a computer's integer as a character: American Standard Code for Information Interchange.

Invisible (non-printing)

Decimal - Binary - Octal - Hex – ASCII Conversion Chart

Used for quality

Decimal	Binary	Octal	Hex	ASCII	Decimal	Binary	Octal	Hex	ASCII	Decimal	Binary	Octal	Hex	ASCII	Decimal	Binary	Octal	Hex	ASCII
0	00000000	000	00	NUL	32	00100000	040	20	SP	64	01000000	100	40	@	96	01100000	140	60	`
1	00000001	001	01	SOH	33	00100001	041	21	!	65	01000001	101	41	A	97	01100001	141	61	a
2	00000010	002	02	STX	34	00100010	042	22	“	66	01000010	102	42	B	98	01100010	142	62	b
3	00000011	003	03	ETX	35	00100011	043	23	#	67	01000011	103	43	C	99	01100011	143	63	c
4	00000100	004	04	EOT	36	00100100	044	24	$	68	01000100	104	44	D	100	01100100	144	64	d
5	00000101	005	05	ENQ	37	00100101	045	25	%	69	01000101	105	45	E	101	01100101	145	65	e
6	00000110	006	06	ACK	38	00100110	046	26	&	70	01000110	106	46	F	102	01100110	146	66	f
7	00000111	007	07	BEL	39	00100111	047	27	‘	71	01000111	107	47	G	103	01100111	147	67	g
8	00001000	010	08	BS	40	00101000	050	28	(	72	01001000	110	48	H	104	01101000	150	68	h
9	00001001	011	09	HT	41	00101001	051	29	)	73	01001001	111	49	I	105	01101001	151	69	i
10	00001010	012	0A	LF	42	00101010	052	2A	*	74	01001010	112	4A	J	106	01101010	152	6A	j
11	00001011	013	0B	VT	43	00101011	053	2B	+	75	01001011	113	4B	K	107	01101011	153	6B	k
12	00001100	014	0C	FF	44	00101100	054	2C	,	76	01001100	114	4C	L	108	01101100	154	6C	l
13	00001101	015	0D	CR	45	00101101	055	2D	-	77	01001101	115	4D	M	109	01101101	155	6D	m
14	00001110	016	0E	SO	46	00101110	056	2E	.	78	01001110	116	4E	N	110	01101110	156	6E	n
15	00001111	017	0F	SI	47	00101111	057	2F	/	79	01001111	117	4F	O	111	01101111	157	6F	o
16	00010000	020	10	DLE	48	00110000	060	30	0	80	01010000	120	50	P	112	01110000	160	70	p
17	00010001	021	11	DC1	49	00110001	061	31	1	81	01010001	121	51	Q	113	01110001	161	71	q
18	00010010	022	12	DC2	50	00110010	062	32	2	82	01010010	122	52	R	114	01110010	162	72	r
19	00010011	023	13	DC3	51	00110011	063	33	3	83	01010011	123	53	S	115	01110011	163	73	s
20	00010100	024	14	DC4	52	00110100	064	34	4	84	01010100	124	54	T	116	01110100	164	74	t
21	00010101	025	15	NAK	53	00110101	065	35	5	85	01010101	125	55	U	117	01110101	165	75	u
22	00010110	026	16	SYN	54	00110110	066	36	6	86	01010110	126	56	V	118	01110110	166	76	v
23	00010111	027	17	ETB	55	00110111	067	37	7	87	01010111	127	57	W	119	01110111	167	77	w
24	00011000	030	18	CAN	56	00111000	070	38	8	88	01011000	130	58	X	120	01111000	170	78	x
25	00011001	031	19	EM	57	00111001	071	39	9	89	01011001	131	59	Y	121	01111001	171	79	y
26	00011010	032	1A	SUB	58	00111010	072	3A	:	90	01011010	132	5A	Z	122	01111010	172	7A	z
27	00011011	033	1B	ESC	59	00111011	073	3B	;	91	01011011	133	5B	[	123	01111011	173	7B	{
28	00011100	034	1C	FS	60	00111100	074	3C	<	92	01011100	134	5C	\	124	01111100	174	7C	|
29	00011101	035	1D	GS	61	00111101	075	3D	=	93	01011101	135	5D	]	125	01111101	175	7D	}
30	00011110	036	1E	RS	62	00111110	076	3E	>	94	01011110	136	5E	^	126	01111110	176	7E	~
31	00011111	037	1F	US	63	00111111	077	3F	?	95	01011111	137	5F	_	127	01111111	177	7F	DEL

Interpreting a quality score

- Lower ASCII value → lower confidence in the call → higher probability of error
- Let variable p be the probability of error at some position
- Smallest printing ASCII char = 31 = '!', largest = 126
- For any quality char
  - Convert to ASCII
  - Subtract 30 → Call it Q, in range 1-96

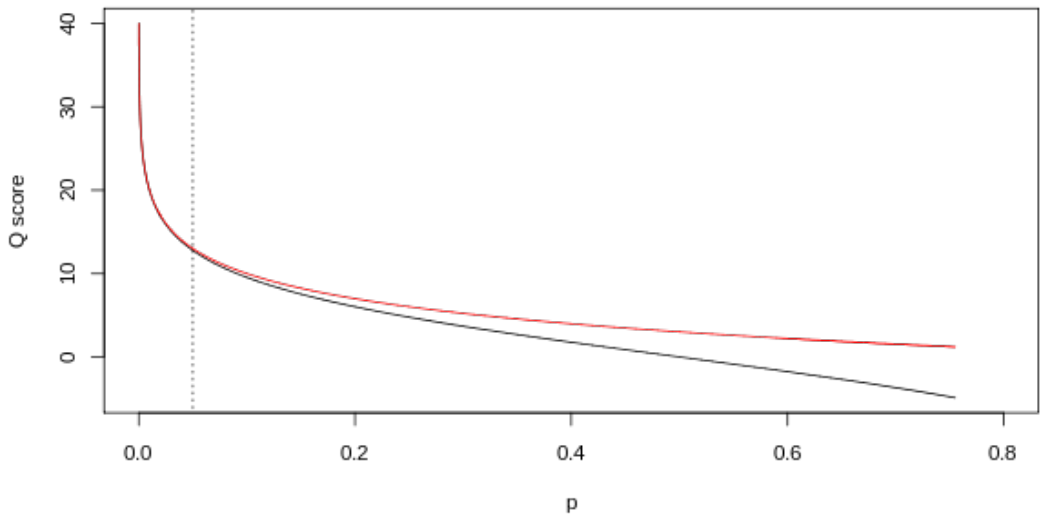
$$Q = -10 \log_{10}(p)$$

Computing p from Q

$$Q = -10 \log_{10}(p)$$

$$-.1 * Q = \log_{10}(p)$$

$$10^{-.1 * Q} = 10^{\log_{10}(p)} = p$$

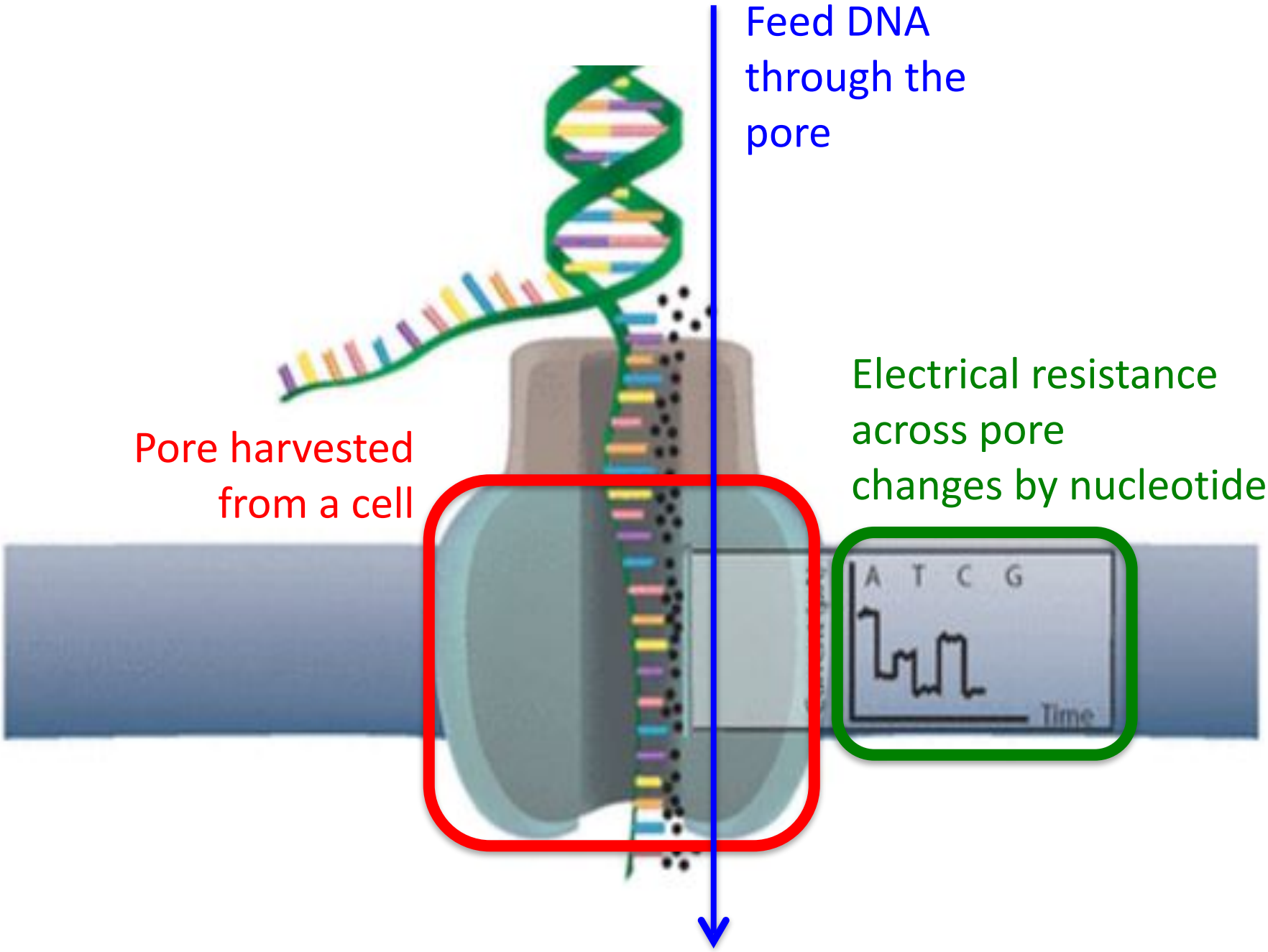


Source: Wikipedia
Ignore the black line

Example: char = 'M'

- 'M' in ASCII = 77
- Q = 77-30 = 47
- $P = 10^{-.1Q} = 10^{-.1*47} = 10^{-4.7} = 0.00001995$
- = Prob (the corresponding base call is wrong)

The future of sequencing, maybe: Nanopores


Feed DNA
through the
pore
Pore harvested
from a cell
Electrical resistance
across pore
changes by nucleotide
A T C G
Time

The Oxford Nanopore MinION

?? No assembly required ??

Topics

- Amplification ✔
- 3 generations of sequencing ✔
- Shotgun sequencing and assembly

Sanger & Pyro sequencing are length-limited

- Sanger: maximum read length ~= 1000 bases
- Pyro: maximum read length ~= 500 bases
- Prokaryote genomes are millions of bases
- Eukaryote genomes are billions of bases
- How can we sequence anything useful?????

Shotgun pellets fragment targets in random patterns and sizes

Shotgun sequencing

- Amplify
- Using restriction enzymes, fragment amplified DNA into random short strands
- As if the DNA had been shot by a shotgun
- Sequence
- Assemble

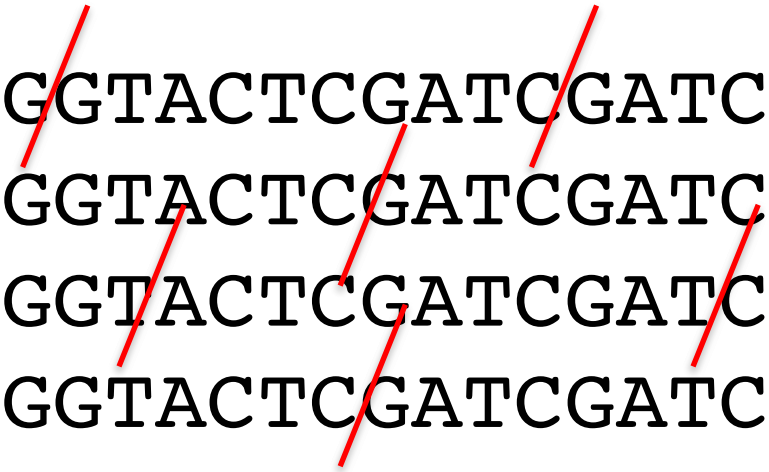
Shotgun Sequencing: Tiny Example

Original

ACGGTACTCGATCGATCGGC

Copies

ACGGTACTCGATCGATCGGC
ACGGTACTCGATCGATCGGC
ACGGTACTCGATCGATCGGC
ACGGTACTCGATCGATCGGC



After Shotgun Fragmentation

ACG GTACTCGATC GATCGGC
ACGGTACTC GATCGATCGGC
ACGGT ACTCGATCGAT CGGC
ACGGTACTCG ATCGATCGGC

Shotgun Sequencing: What we see

```
ACG
GTACTCGATC
GATCGGC
ACGGTACTC
GATCGATCGGC
ACGGT
ACTCGATCGAT
CGGC
ACGGTACTCG
ATCGATCGGC
```

How do we recover the original sequence?

A naïve approach: find longest overlap, then merge 2 reads

ACG

GTACTCGATC

GATCGGC

ACGGTACTC

GATCGATCGGC

ACGGT

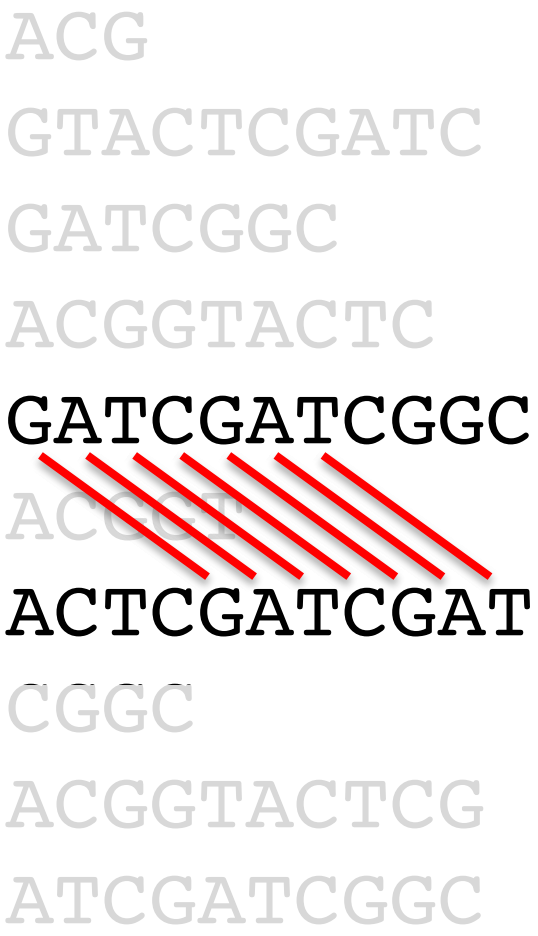
ACTCGATCGAT

CGGC

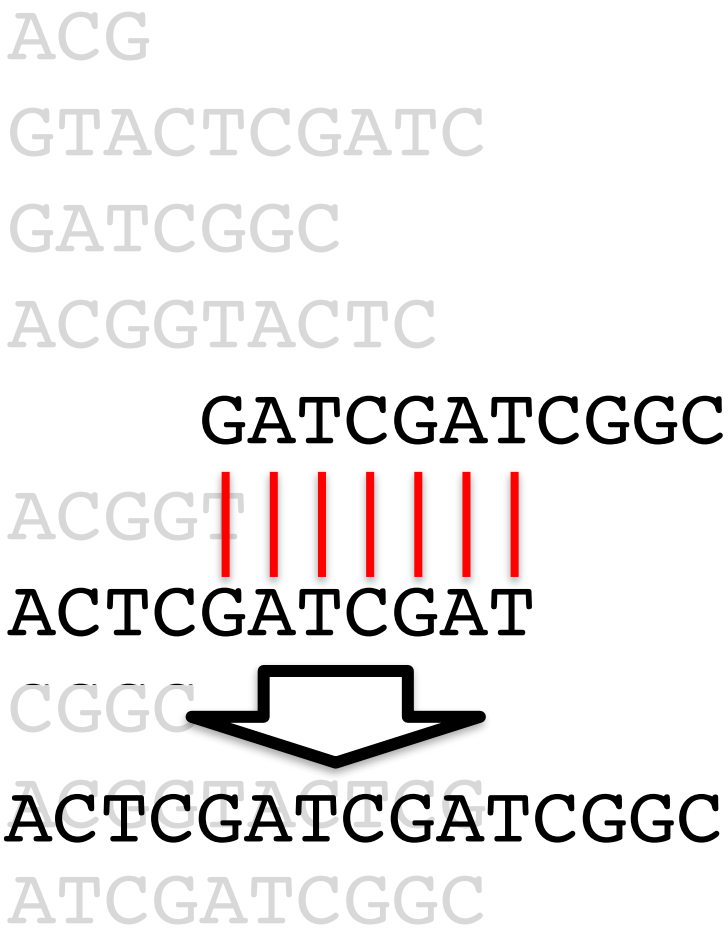
ACGGTACTCG

ATCGATCGGC

A naïve approach: find longest overlap, then merge 2 reads



A naïve approach: find longest overlap, then merge 2 reads



Replace 2 short reads with 1 longer read

ACG

GTACTCGATC

GATCGGC

ACGGTACTC

~~GATCGATCGGC~~

ACGGT

~~ACTCGATCGAT~~

CGGC

ACGGTACTCG

ATCGATCGGC

ACTCGATCGATCGGC

Repeat until everything has been merged

```
ACG
GTACTCGATC
GATCGGC
ACGGTACTC
ACTCGATCGATCGGC
ACGGT
CGGC
ACGGTACTCG
ATCGATCGGC
```

```
ACGGTACTCG
ATCGATCGGC
```

Only 3 problems

- If you're not careful, small sequencing errors can produce big assembly errors
- If you're not careful, execution time = O(n!) where n = number of reads
- ➔ Writing an assembler is a high-level skill
- ➔ Even *using* assembler apps requires skill and patience
- ➔ Too many issues for an undergrad course (whew!)

Topics

- Amplification ✔
- 3 generations of sequencing ✔
- Shotgun sequencing and assembly ✔

That's all for this topic.

The rest of this course is about identifying sequences that you've amplified, sequenced, and assembled.