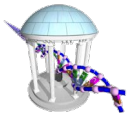
BCB 716 - Sequence Analysis

- I may have a guest lecturer on thursday
- Keep an eye on the course website for updates
- Also monitor it for login instructions

Sequencing Technologies

DNA Sequencing History



- **DNA sequencing was one of the most significant breakthroughs of the 20th century**
- **This was so inherently obvious it was awarded a Nobel prize only 3 years after its development**

Sanger method (1977):

Uses labeled dideoxynucleotide-triphosphates (ddNTPs) terminate DNA copying at random points.

Fredrick Sanger

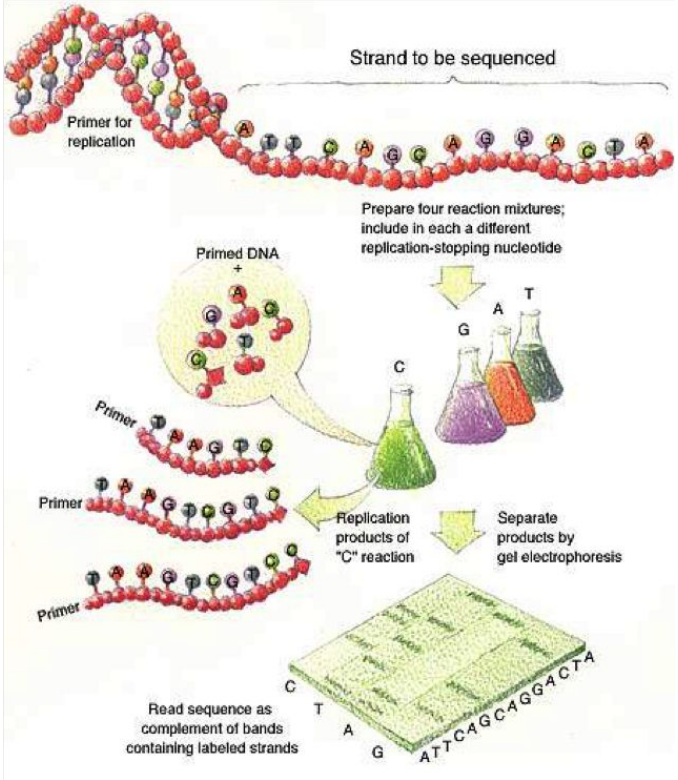
Gilbert method (1977):

Used various chemicals (Dimethyl Sulfate, Hydrazine) to modify and then cleave DNA at specific points (G, G+A, T+C, C).

Walter Gilbert

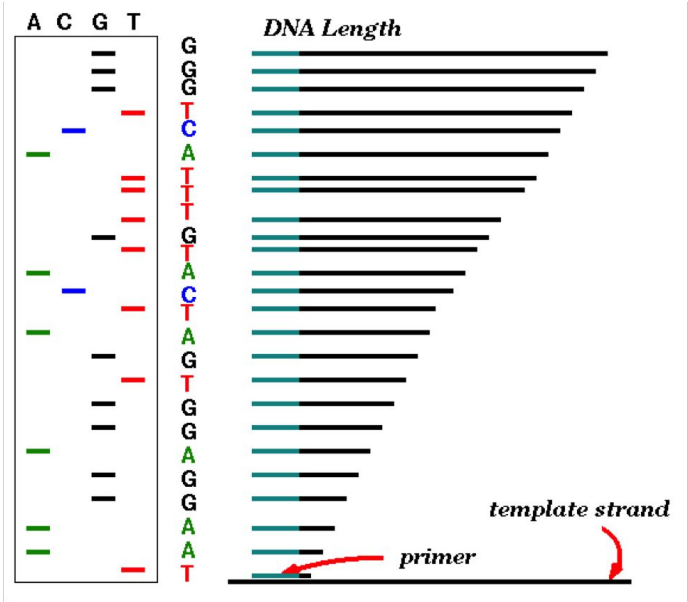
Sanger Method

1. Use the polymerase chain reaction (PCR) to make billions of copies of a DNA sequence
2. Starting at *custom* primer, sort of like our the *origin of replication*, we initiate one last replication
3. Include *chemically altered* and *fluorescently labelled nucleotides*, called dideoxynucleotide-triphosphates (ddNTPs)
4. If a ddNTP gets incorporated into a sequence it stops further replication
5. Separate replication products by length, using gel electrophoresis
6. Good for 500-1000 bases, then the error rates grow and extension rate slows
7. About 10 bases-per-second or 9.5 years to read an entire genome if we could do it from beginning to end

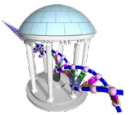


Sanger Method

1. Use the polymerase chain reaction (PCR) to make billions of copies of a DNA sequence
2. Starting at a *custom* primer, nearby sequence motif we initiate one last replication run
3. Include *chemically altered* and *fluorescently labelled nucleotides*, called dideoxynucleotide-triphosphates (ddNTPs)
4. If a ddNTP gets incorporated into a sequence it stops further replication
5. Separate replication products by length, using gel electrophoresis
6. Good for 500-1000 bases, then the error rates grow and extension rate slows
7. About 10 bases-per-second or 9.5 years to read an entire genome if we could do it in a single read from beginning to end



Assembling the Human Genome



In 1990, a moon-shot-like project was begun to sequence the entire Human Genome.

- It would require 30x coverage to provide enough sequences
- Recall there are sequence differences-- Approximately 1:1000 bases
- Redundancy was needed to find the majority base from 16 different individuals (32 genomes)
- Also needed the extra coverage to assure that there is enough overlap to assemble the 500 base-pair reads

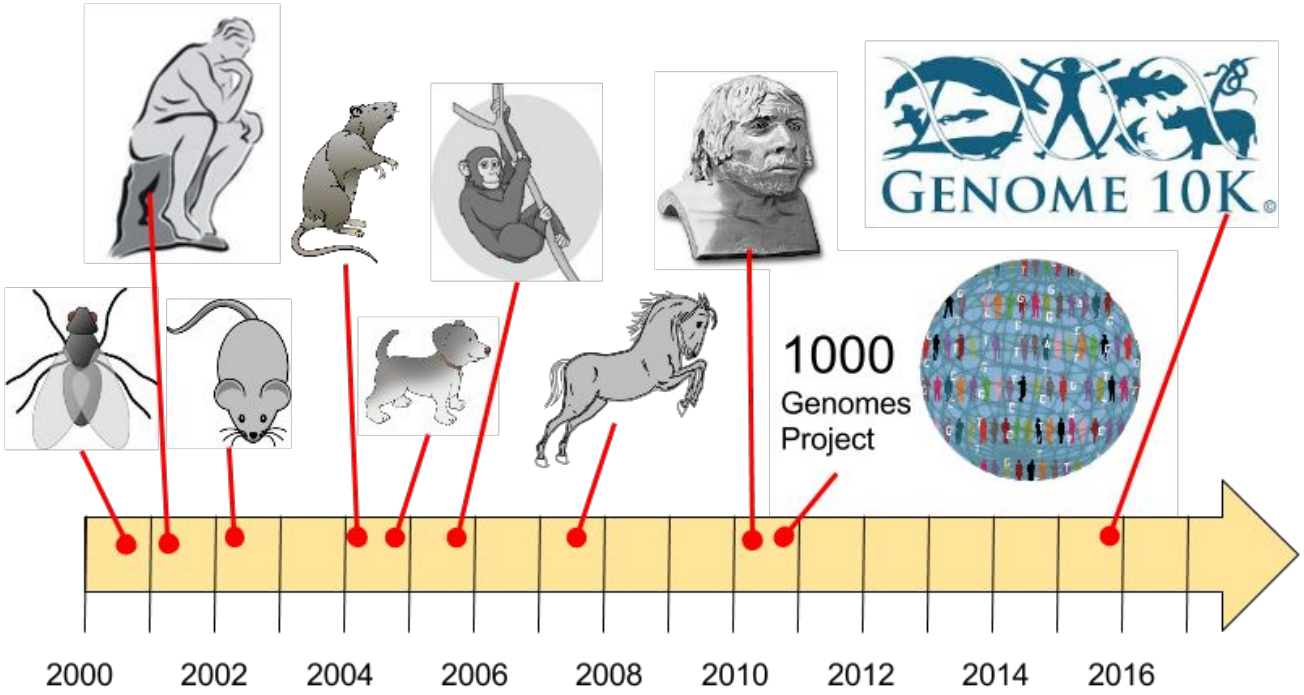
A $3 billion dollar NIH funded public effort led by Francis Collins with a 15-year plan. It would distribute the work across several labs in a community effort by assigning primers to groups on a first-come basis. New sequencing results yielded new primers, so the project required a central coordination.

In 1997 a private company, Celera, lead by Craig Venter, suggested they could beat the public effort by dispensing with primers. They'd just randomly fragment DNA and sequence each with no idea of the how sequenced fragments would fit together. In other words, they were going to rely on computer science to assemble their reads algorithmically.

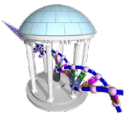
The result was that, despite tensions, the groups ended up sharing data and technologies.
And the competition led to a completed draft 5 years ahead of schedule.

The Sequencing Race

Since the Human Genome project there have been an explosion of genomes sequenced. Initially, the focus was on model organisms, then favorites, then all of human diversity, and finally a catalog of life's diversity.

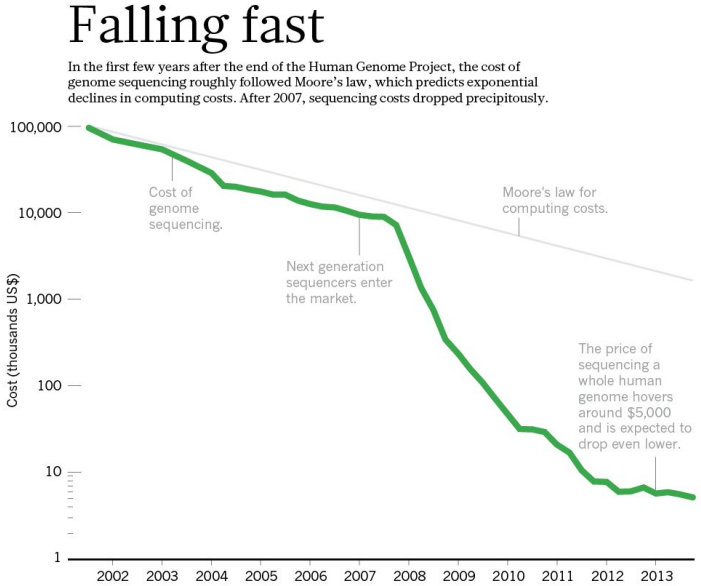


The secret behind this explosion of genomes

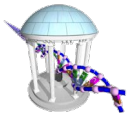


Next generation sequencing machines have revolutionized the DNA sequencing process. They work in various ways including massively-parallel single-base extension methods, to captured Dnases whose motions suggest a the base being replicated, to microholes that only a single DNA molecule can pass through, and the bases are determined by detectable charge differences.

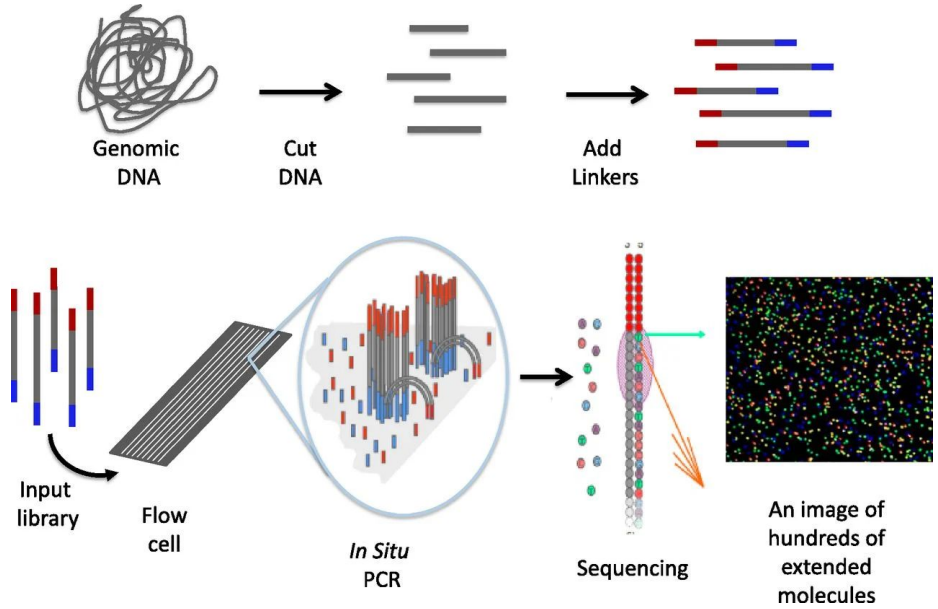
In a way, the *genome moonshot* was far more successful than the real moonshot. The rate at which genomes can be sequenced, and the cost per base has seen unprecedented improvements. Faster than even Moore's Law.



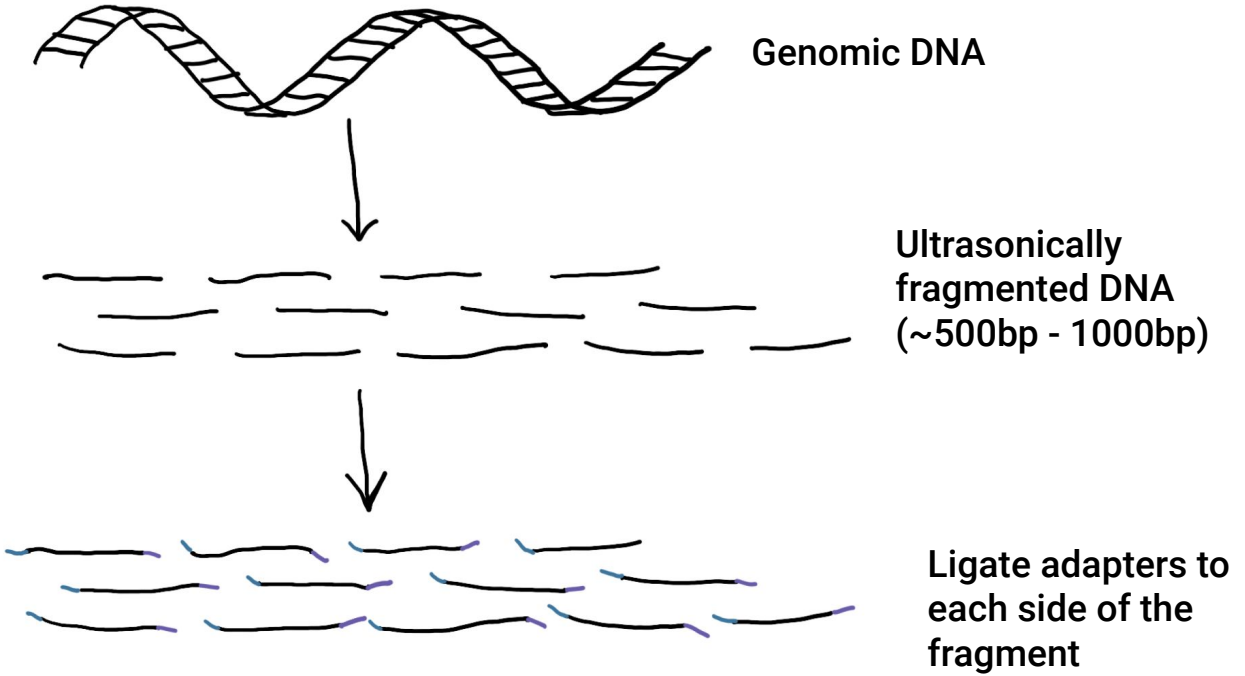
Sequencing by Synthesis (Illumina-like)



1. DNA is fractured into fragments
2. Adaptors are ligated on each end
3. Fragments are captured within flow cell
4. Amplified
5. Sequenced



Library Prep



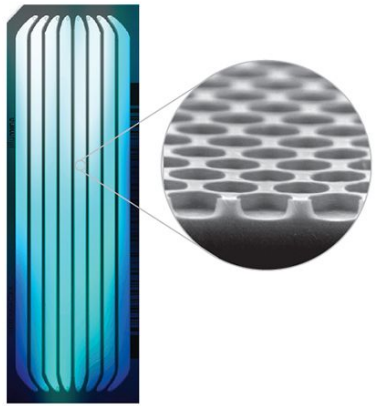
Linker/Adapter Schematic

- The primer sequences are used to start the polymerase sequencing
- Index regions are used for various purposes such as combining DNA from different samples in the same lane
- P5 and P7 adapters are used to capture fragments within a flow cell

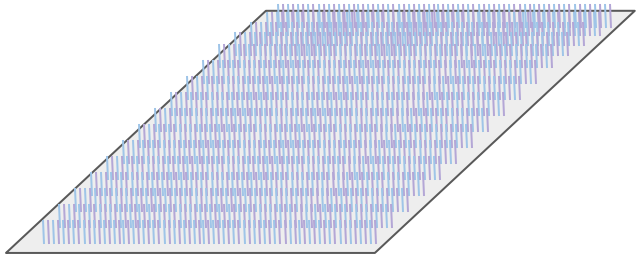
```
P5: 5' AATGATACGGCGACCACCGA 3'
P7: 5' CAAGCAGAAGACGGCATACGAGAT 3'
```

Cluster Generation

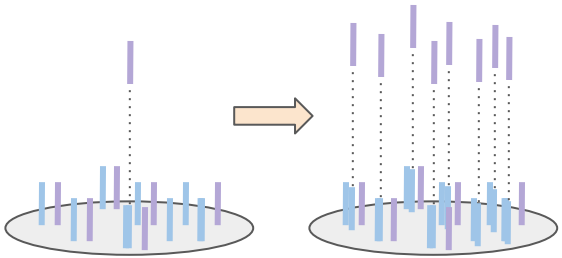
Each fragment is loaded into a flow cell where many (1000-4000) copies of it are made in a small localized region. These are called clusters.



Typical Lane

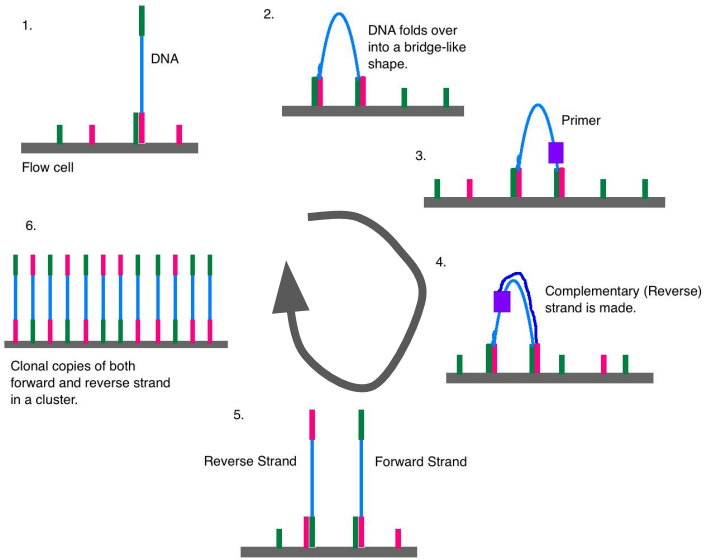


Each lane is filled with a "lawn" of P5/P7 capture oligos



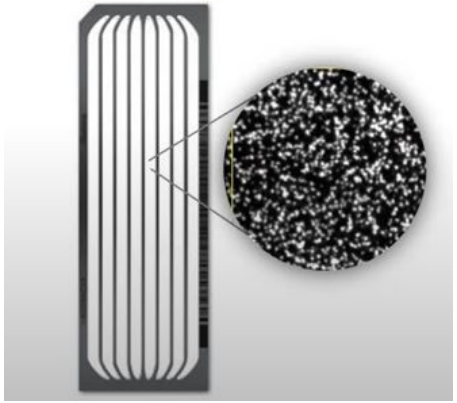
Lane sparsely captures DNA fragments, which are then amplified until all/most probes are filled

Amplification and Resulting Cluster Pattern



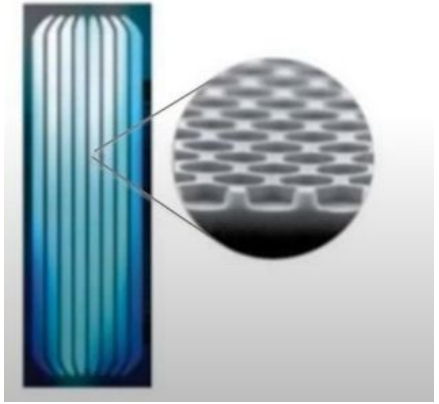
Random Flow Cell

- HiSeq™ 2500, MiSeq™, NextSeq™, MiniSeq™
- Randomly spaced clusters
- Variable Insert Sizes
- Lower Duplication Rates

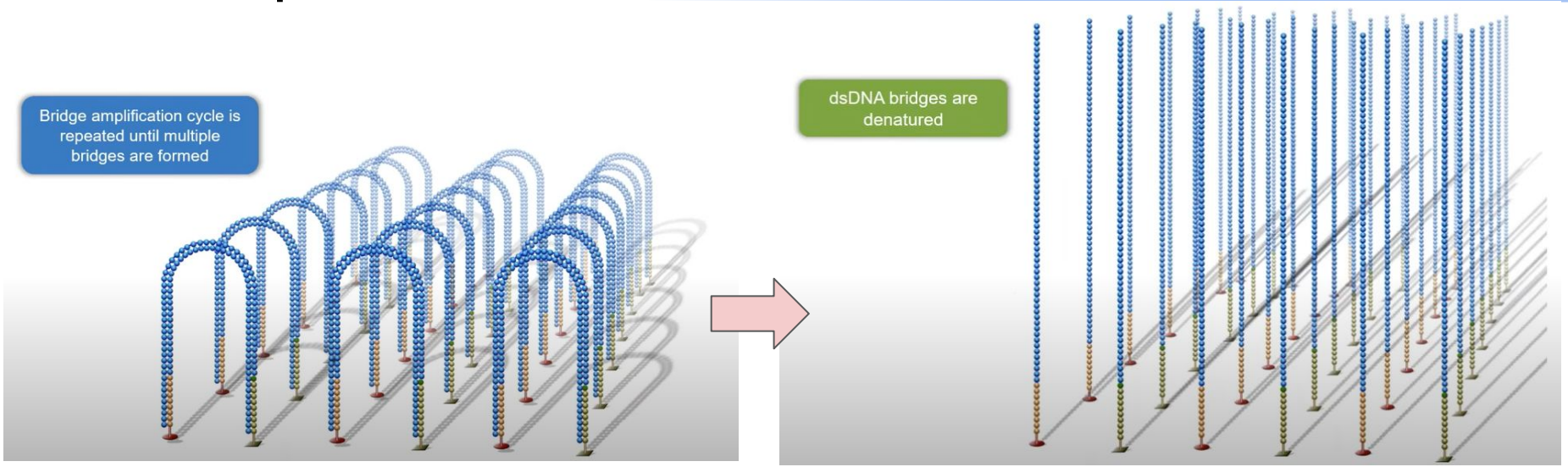


Patterned Flow Cell

- HiSeq 3K/4K/X, NovaSeq™ 6000, iSeq™ 100
- Defined size and spacing
- Increased Cluster density
- Simplified imaging

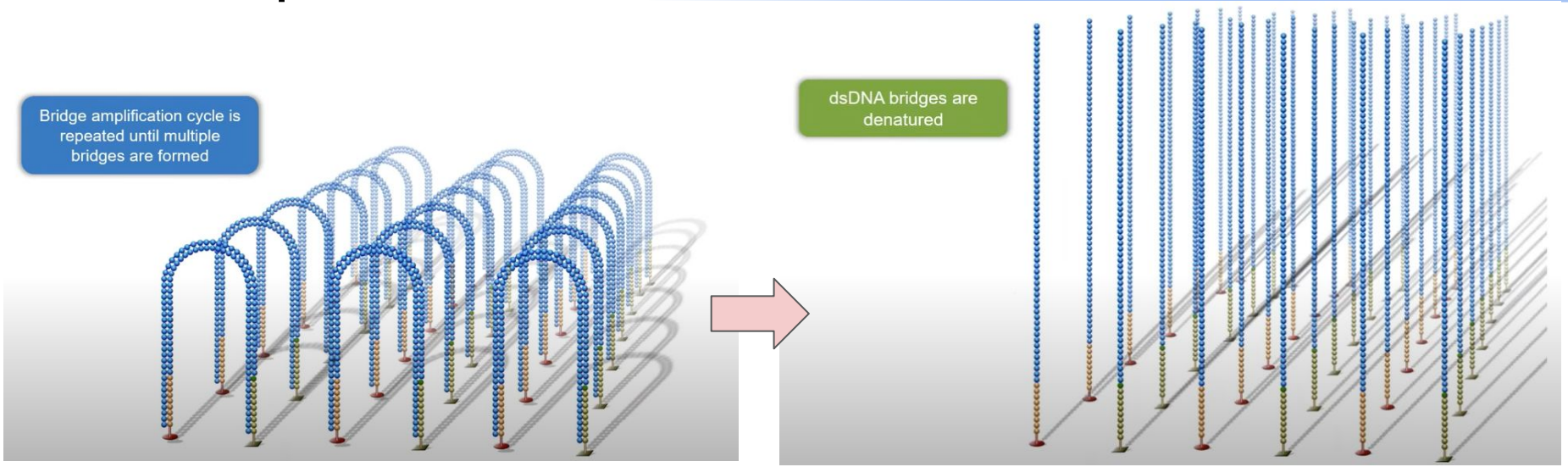


Post Amplification Linearization



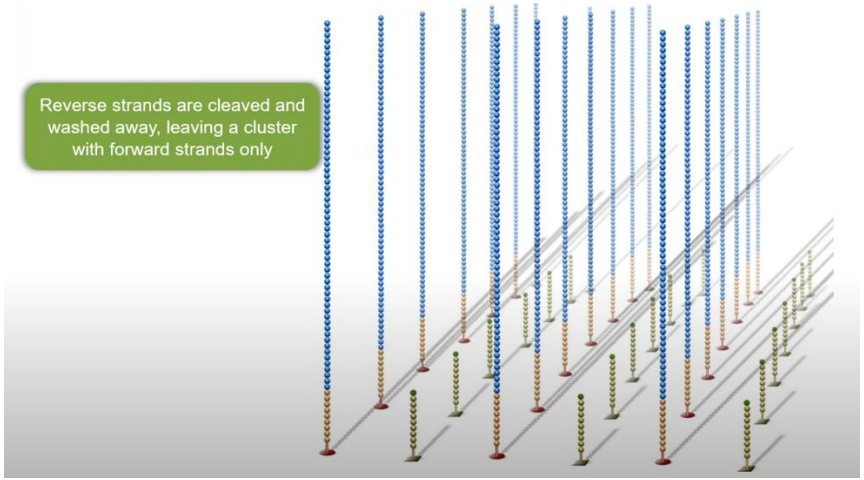
Many copies of the DNA are made by repeated bridge implications once all of the oligo's are occupied ampliation stalls. Finally, the double-stranded DNA bridges are denatured to produce single-strand DNA which are used a a template for sequencing

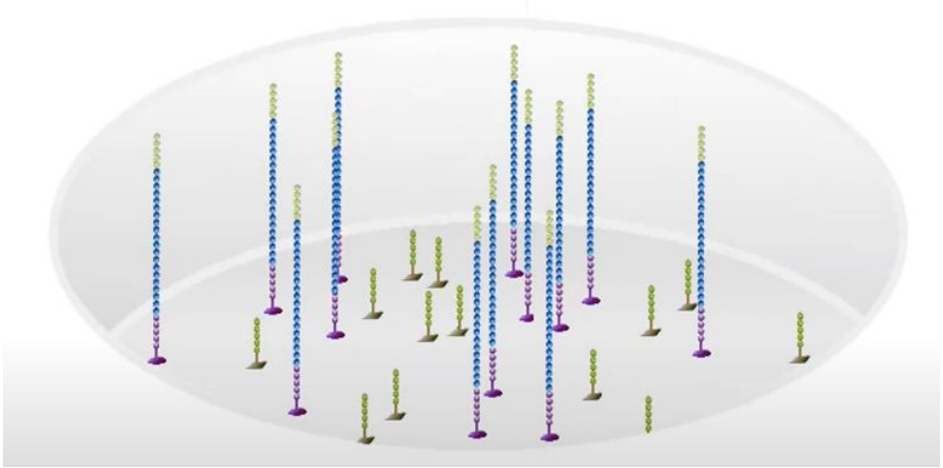
Post-Amplification Linearization



Many copies of the DNA are made by repeated bridge implications once all of the oligo's are occupied ampliation stalls. Finally, the double-stranded DNA bridges are denatured to produce single-strand DNA which are used a a template for sequencing

Post-Amplification Linearization





One of the two strands are chemically released and washed away. These will be the DNA clusters that are sequenced using fluorescently labeled nucleotides. Shown side-by-side are the random clustering lanes and the more recent patterned clustering lanes.

Sequencing

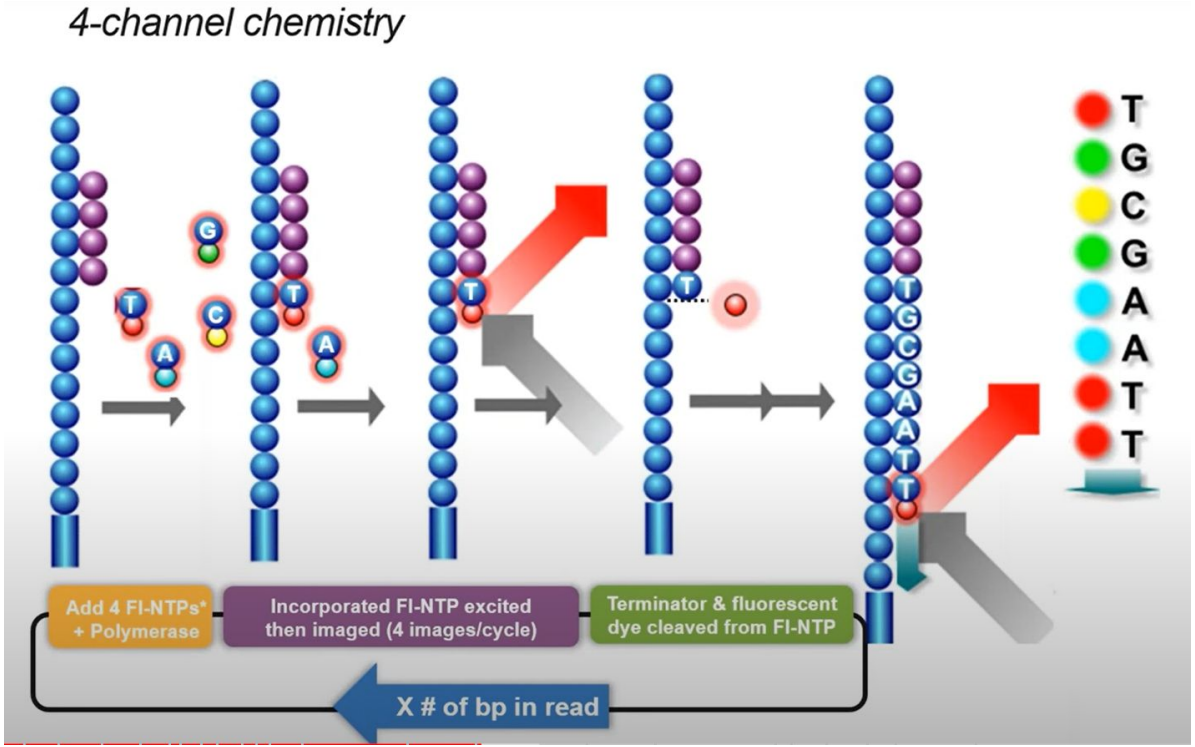
Each FI-NTP is labeled with a unique fluorophore.

Two emit under green laser light (G & A) and the other two under red light (C & T).

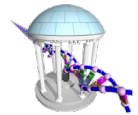
Four photos per base extensions

3-steps:

1. Incorporate
2. Image
3. Cleave & Wash away



Sequencing

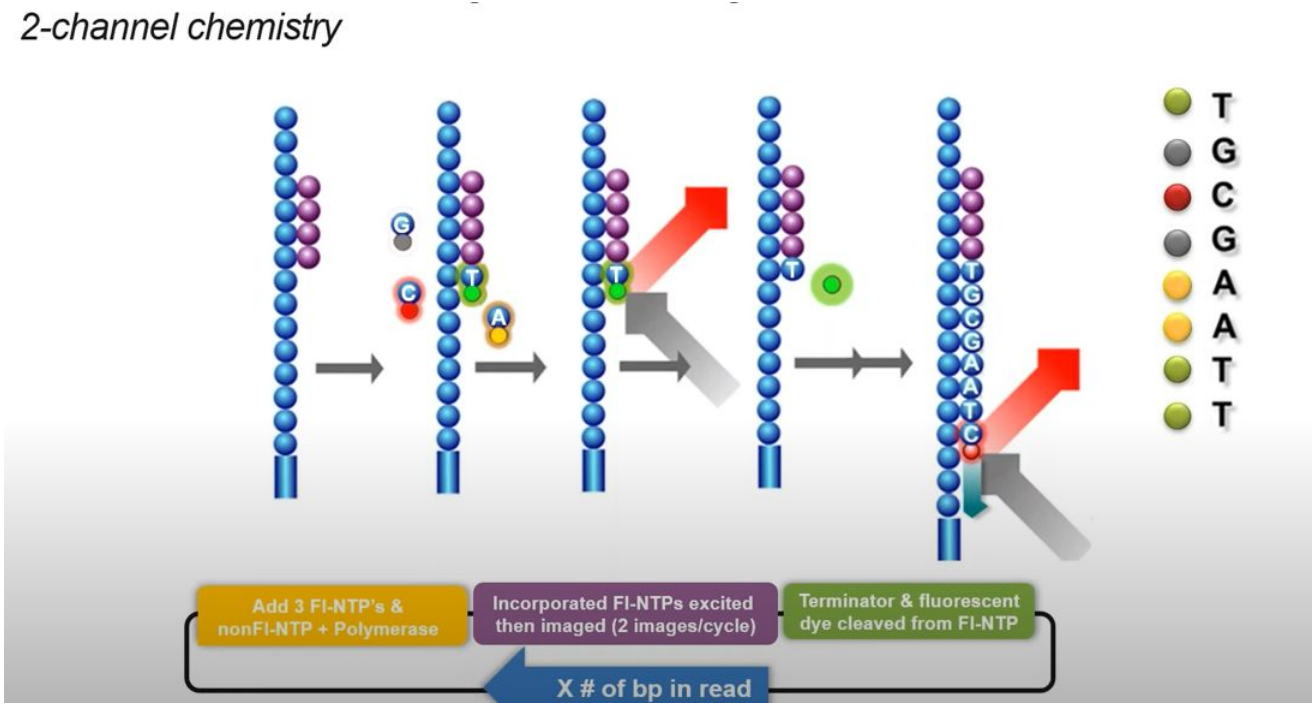


Each FI-NTP is labeled with a 1, 2, or no fluorophore.

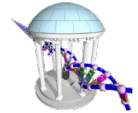
T with green, C with red, A with green and red, and G with none.

1 color photo per base extension

Faster, but has problems with long runs of Gs



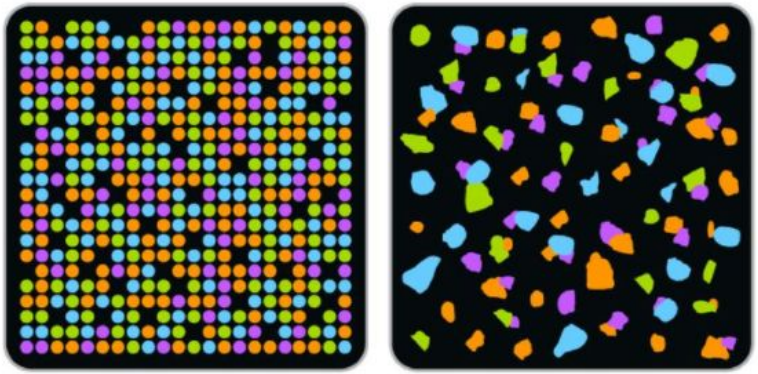
Flow Cell Cycle Images



Cartoons of imaging cycles for both random and patterned flow cells

Images are processed immediately, and not stored

Process is repeated 100-300 times

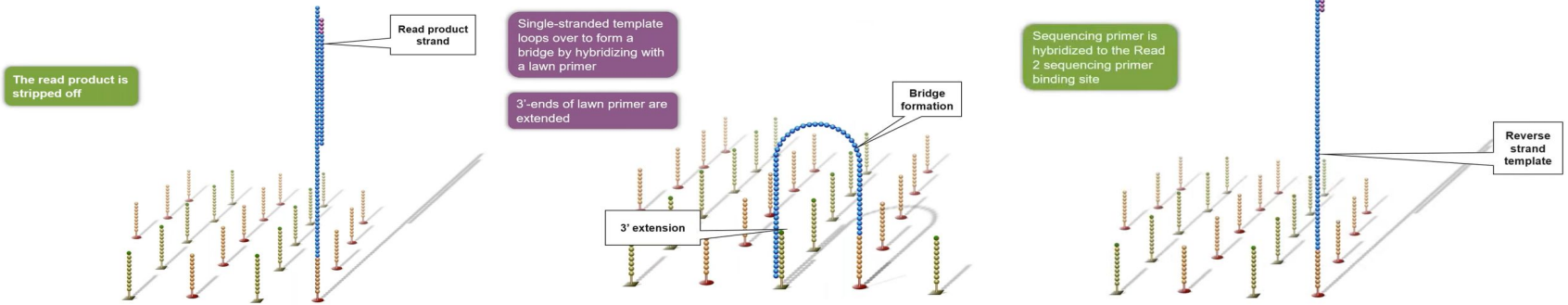


A. Patterned B. Non-Patterned

Cluster centers are found on the fly with non-patterned flow cells

Paired-end processing

After imaging cycles the chemistry reenables the fragment bridging

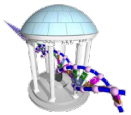


Only showing one strand, but all forward oligos have the read product.

Bridging and synthesis restarts

Forward strand is cleaved, and sequencing of the reverse complement fragment begins

Final result

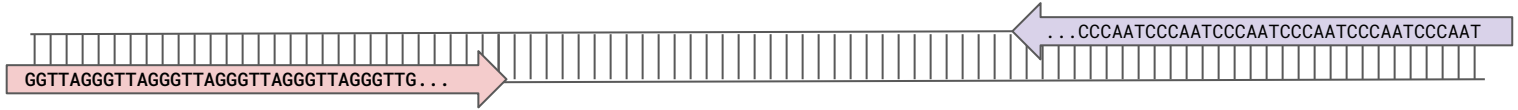


FASTQ file: *ERR1443441_1.fastq*

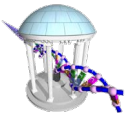
```
@ERR1443441.1 HX8_19333:3:2221:2909:63032/1
GGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTGGGGTAACGGACAGGGGTAGGGTAAGCGATAGGGTTAGGGGTAGGGTCAGGGTTAGGGTTAGGGTTATGGGTAGGGTTAGTGTTGGGGGTAGGGTTAGGGGTGGGGTTAGGGTTAGG
+
AAAFFKKKFKKKKKFFKKKKKKKKKKKKFAFKAF,F,,,,7,7,,,,,,,(7A<,,,,,,,,,,7,,A77F,,7,7,,,A<,AF,,,7<,,,FF,,,,7,A,,,,,7,,A7,7A,<,,,7AF((AF,7,,,,,,,,(,(<,,,,,,,,7<<
@ERR1443441.2 HX8_19333:3:1104:32644:17606/1
CTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTACCCCAAACCCAACCCAAACCCAAACCCAAACCGAACCCTAAACCCAACCCCAACCCCCCCCCCCCCCCCGCACCCCCCCCCCACCACCACCCCACCCACAA
+
AAFFFKKKKKKKKKKKKKKKKKKKKKKKKKKAKKKKKKKKKKKKKKKKKK7F,AK,,,,A<,7,7A7,7,AK,,F,AF,,,,AA,7FA,,,7,,,A,,7A<A7A,,A,,(((,,(((((((,((7((<,7(((,(,<,,<,,A,(,,(,,,,,
@ERR1443441.3 HX8_19333:3:2105:5010:65687/1
GGGTTAGGGTTAGGGTTGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGAGATCGGAAGAGCACACGTCTGAACTCCAGTCACCAATGGAAATCGCGTATGCCGTCGGCTGCTTGAAAAGAAGGGGCATGGGGTAGGGGTGTGGG
        .
        .
        .
```

FASTQ file: *ERR1443441_2.fastq*

```
@ERR1443441.1 HX8_19333:3:2221:2909:63032/2
TAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCCAACCCTAACCCCAACCCATACCCCAACCCCAACCCTAACCCTAAGCCTAACCCCAGGCCGAAGCCCACCCGTGACCCTACGACT
+
AAFFFKKKKKKKKKKKKKKKKKKKKKKKKKFFKKKFKAF<7AFAFFK,<FKK,,<,7,,77FAK,,A,,77,,,7<7,F7<,,,,,,,,,7,<<,FA,<A,,<,,,7,A,,7,,A,<,,,<,,((,,(,,,,,,,((((,,,,,,,,,,,,,
@ERR1443441.2 HX8_19333:3:1104:32644:17606/2
GTTAGGGTTAGGGTTAGGGTTGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGATGGGGAGGGCGAGGGGGAGGGATAGGGGGTAGGTGTCGGGGGTGGCGGTAGGGGGGAAGGGGGAGG
+
A,AFFKKKKKKKFKFKFKK<FAKK7FAKKKAFFKKK<FFFKKF7FFKK<AAFKKKFFKKKKKKKKAFKF,FF,7,A7,,7,,A,A,,77,AA(A(,,(,A<,,7AA,,7<,,,,(,,<,7<,,,(7,(,7,7A<F,,,,,((,,,,,7(,(
@ERR1443441.3 HX8_19333:3:2105:5010:65687/2
CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCAACCCTAACCCTAACCCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATTAAAAAAGAAACGCGAGTACTAGACTGCAGTCTCAATCA
```



FASTQ Quality strings



A string interpreted as 40 numbers

Formula: score + offset => look for American Standard Code for Information Interchange (ascii) symbol
Two variants: offset=64(Illumina 1.0-before 1.8); offset=33(Sanger, Illumina 1.8+).
A quality score is typically: [0, 40]

P = error probability

```
(33):  !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
(64):  @ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefgh
```

Sanger, Illumina v1.3 to 1.7 (ASCII_BASE=64)

Q	ASCII	P	Q	ASCII	P	Q	ASCII	P	Q	ASCII	P
1	A	0.79433	12	L	0.06310	23	W	0.00501	34	b	0.00040
2	B	0.63096	13	M	0.05012	24	X	0.00398	35	c	0.00032
3	C	0.50119	14	N	0.03981	25	Y	0.00316	36	d	0.00025
4	D	0.39811	15	O	0.03162	26	Z	0.00251	37	e	0.00020
5	E	0.31623	16	P	0.02512	27	[	0.00200	38	f	0.00016
6	F	0.25119	17	Q	0.01995	28	\	0.00158	39	g	0.00013
7	G	0.19953	18	R	0.01585	29	]	0.00126	40	h	0.00010
8	H	0.15849	19	S	0.01259	30	^	0.00100			
9	I	0.12589	20	T	0.01000	31	_	0.00079			
10	J	0.10000	21	U	0.00794	32	`	0.00063			
11	K	0.07943	22	V	0.00631	33	a	0.00050			

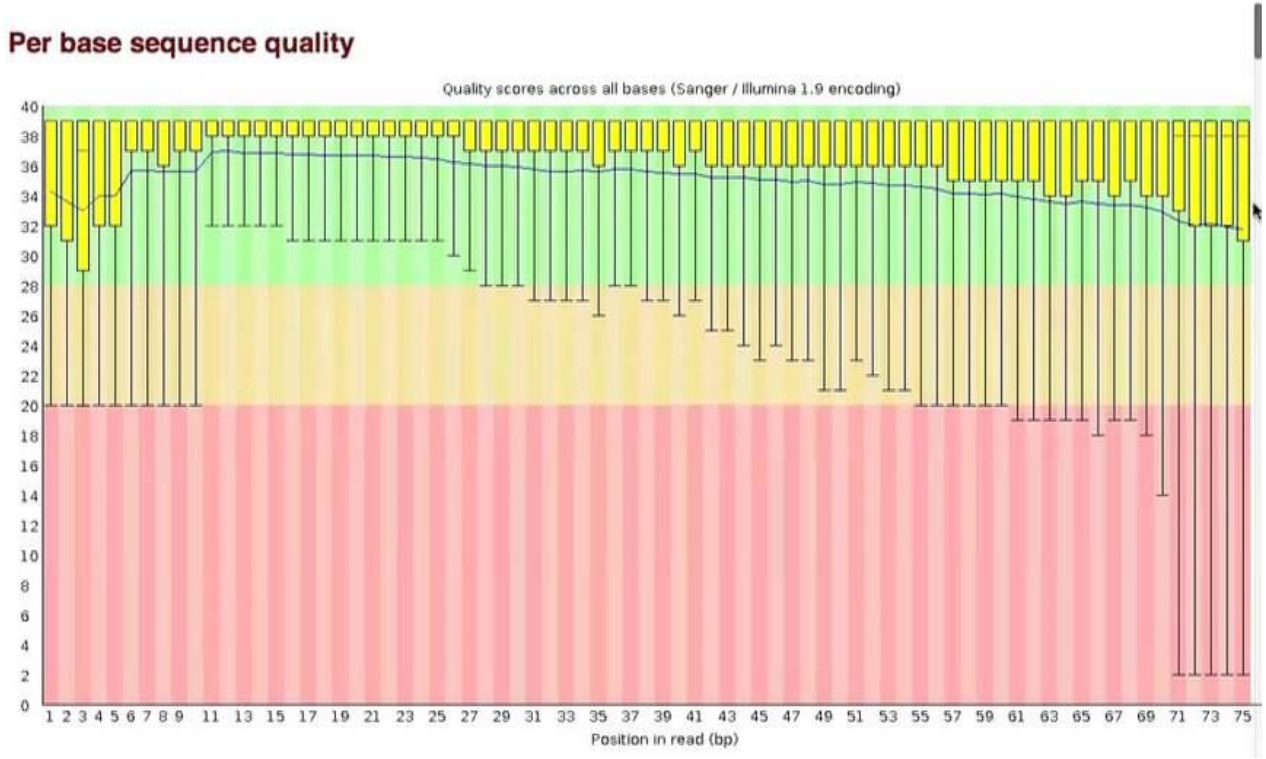
Illumina v1.8 and later (ASCII_BASE=33)

Q	ASCII	P	Q	ASCII	P	Q	ASCII	P	Q	ASCII	P
1	"	0.79433	12	-	0.06310	23	8	0.00501	34	C	0.00040
2	#	0.63096	13	.	0.05012	24	9	0.00398	35	D	0.00032
3	$	0.50119	14	/	0.03981	25	:	0.00316	36	E	0.00025
4	%	0.39811	15	0	0.03162	26	;	0.00251	37	F	0.00020
5	&	0.31623	16	1	0.02512	27	<	0.00200	38	G	0.00016
6	'	0.25119	17	2	0.01995	28	=	0.00158	39	H	0.00013
7	(	0.19953	18	3	0.01585	29	>	0.00126	40	I	0.00010
8	)	0.15849	19	4	0.01259	30	?	0.00100	41	J	0.00008
9	*	0.12589	20	5	0.01000	31	@	0.00079			
10	+	0.10000	21	6	0.00794	32	A	0.00063			
11	,	0.07943	22	7	0.00631	33	B	0.00050			

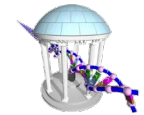
Typical Quality String trends

Base calling quality begins slightly low, Improves, and then falls off later in the sequence

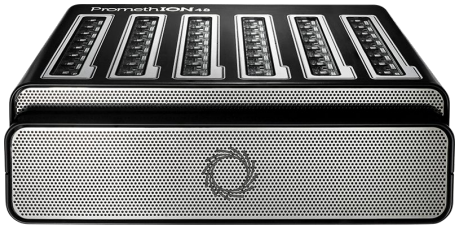
Output of a common sequence analysis tool FASTQC



Long Read Sequencing



- The Sequencing-by-Synthesis technologies are reliable and scalable
- Fidelity falls off as the sequencing process extends longer and longer
- Currently, limited to ~300bps per side
- New technologies developed for long reads

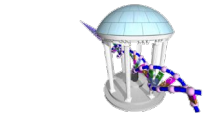


PromethION ($195K)

GridION ($50K)

MinION $1K

Nanopore Libraries

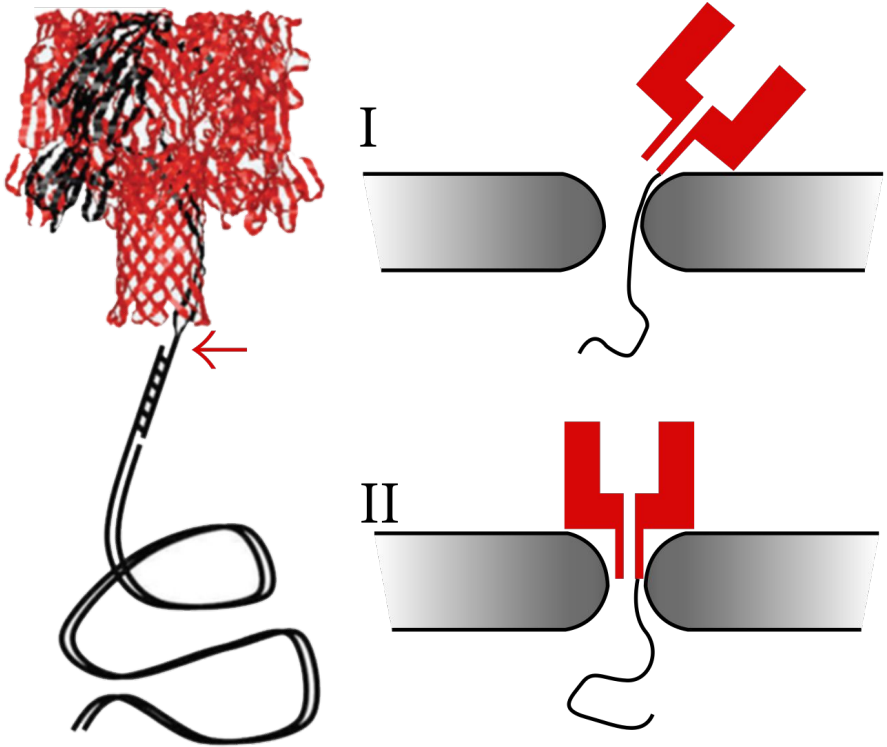


Long (20k-50k mean) bp fragments are fragmented from genomic DNA

An adapter sequence is ligated onto one end

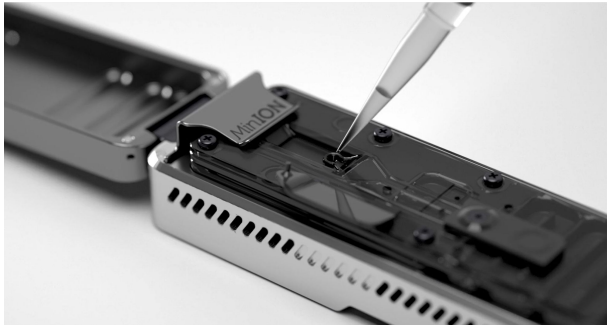
An engineered helicase motor protein is then mated to each fragment

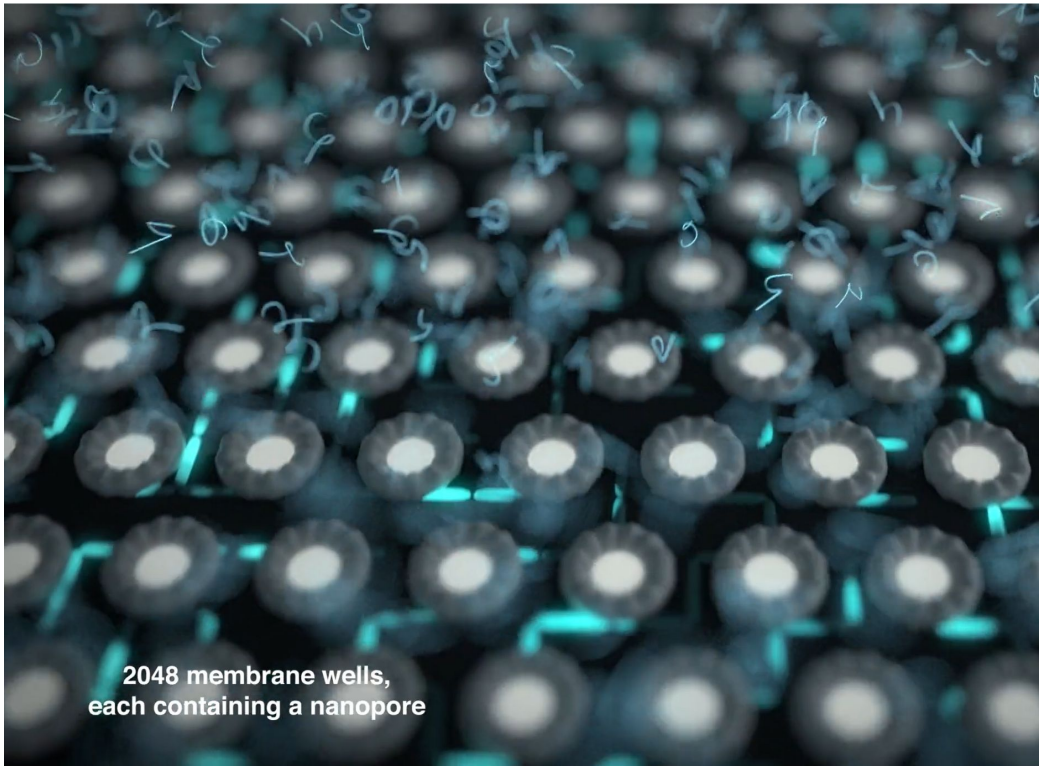
This DNA-protein complex is then docked into a membrane substrate of nano wells



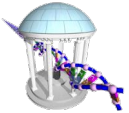
Sample added to Sequencer

Both the library kits and membrane of wells, also called a flow cell, are consumables





A single DNA strand is pushed through the pore

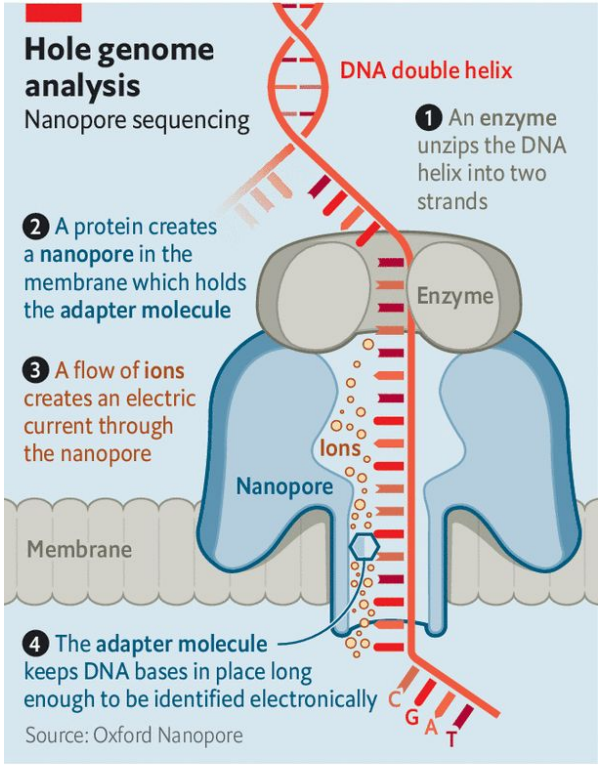


An aqueous solution of ions is used to straighten and measure minute changes in current as the DNA is pushed through the pore.

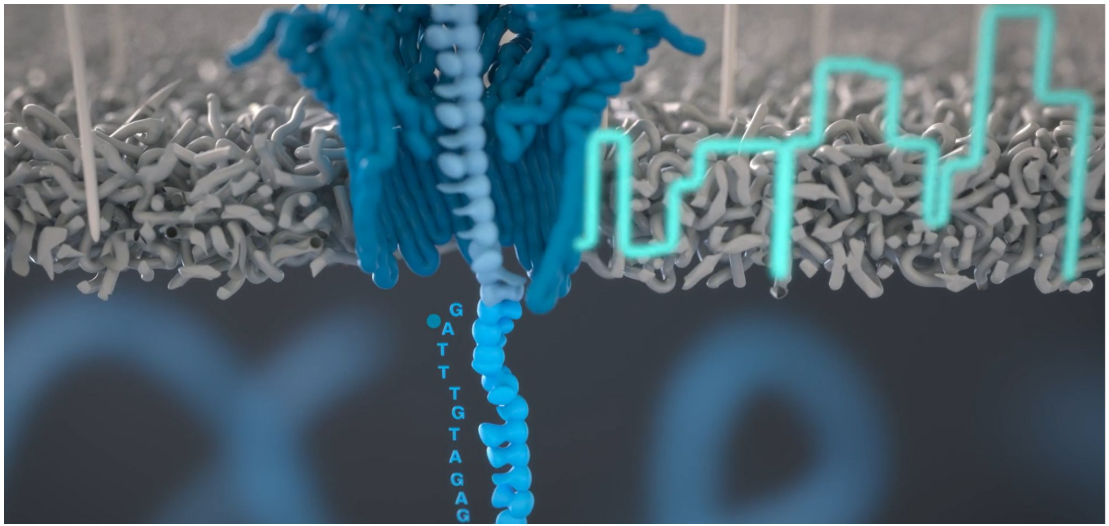
About 400 bases/sec

x2000 pores/flowcell yields 800,000 bp/sec

Higher end products have more flowcells



Deconvolution of signal



Each electrical reading encompases 4-5 bases

Strand is advanced one base for the next reading

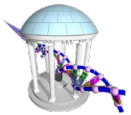
Each base is determined by deconvolving a sequence of 4-5 electrical pulses.

The electrical pulses are kept and converted to FASTQ via various base calling algorithms

Small current differences over longer windows can also detect methylation

GATTTGTAGAGCATTAGAAAA

Another Long-read Technology

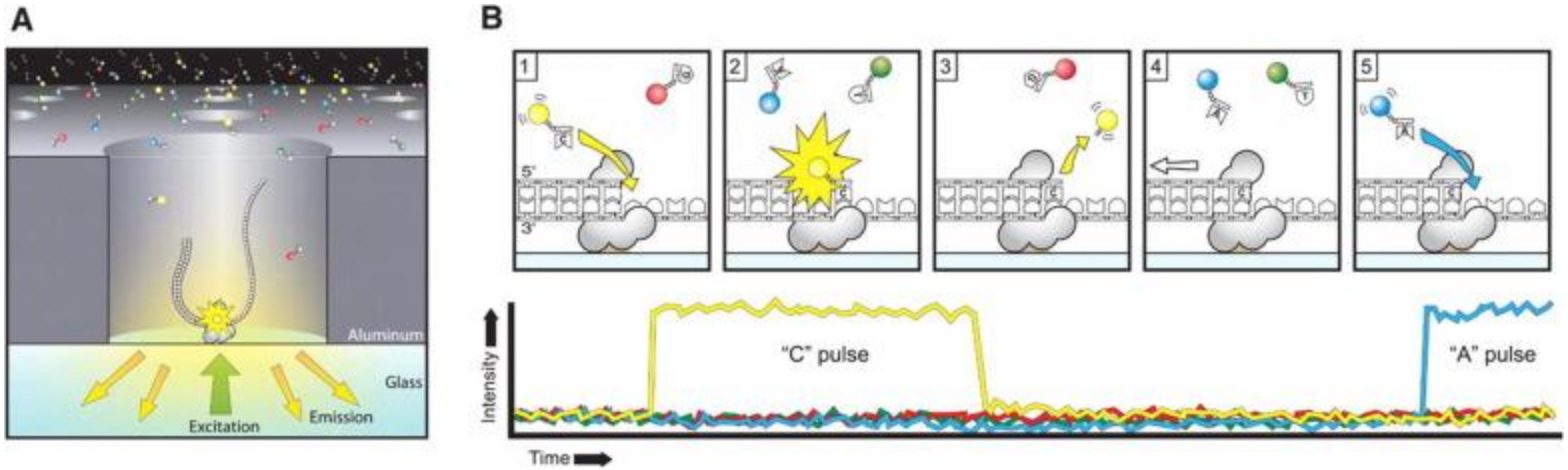


Another company, Pacbio Sequencing, has developed a sequencing technology that coops a DNA polymerase too do the work of sequencing. The technique is called Single-molecule real-time (SMRT) sequencing.

Recall that DNA polymerases are responsible for replicating DNA by all lifeforms. The also react differently when integrating the next base (A,C,T or G) of a sequence

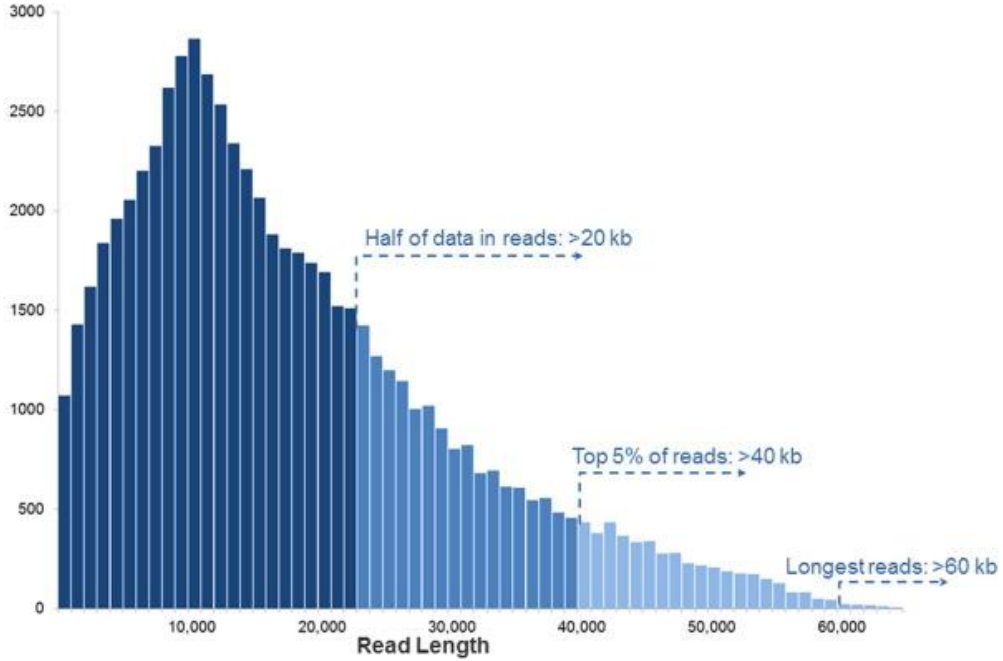
Captured modified DNA Polymerases

A laser is used to detect the small changes that appear when each base is added. This process can progress at a constant rate, so it doesn't stop and read like other technologies.

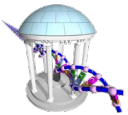


A large range of DNA

The accuracy can also be very high is the DNA strand is circularized and read multiple times.



Long Read FASTQ



A single, shorter than typical, nanopore read

Next Time

Sequence Alignment

Where sequencing meets genomes