

Genomics

What we are going to cover

- DNA sequencing technology (1.5)
- Sequence assembly (1.5)
- Gene prediction (2)
- Functional annotation (1)

DNA Sequencing

(also include sequencing of RNA, which could be reverse-transcribed into cDNA)

<http://www.cs.colostate.edu/~cs680/Slides/lecture4.pdf>

http://genetics.stanford.edu/gene211/lectures/Lecture1_Genome_Sequencing-2013.pdf

http://fenchurch.mc.vanderbilt.edu/lab/bmif310/2012/BMIF310_Genome_Sequencing.pdf

<http://www.utwente.nl/tnw/mcbp/education/Next%20gen%20sequencing.pdf>

<http://www.ncbi.nlm.nih.gov/books/NBK21129/#A5997>

DNA recombination tech

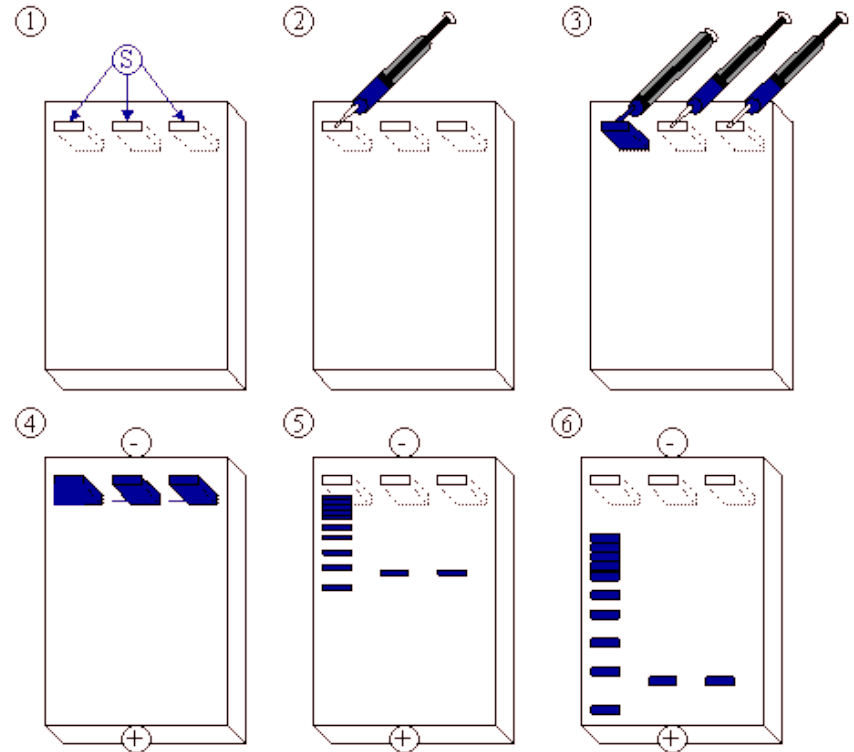
Isolating DNA

- The basic steps:
 1. Break the cells open
 2. Disrupt cell membranes with a detergent
 3. Remove proteins and other macromolecules
 4. Concentrate the DNA by precipitating it and re-suspending it in fresh buffer.



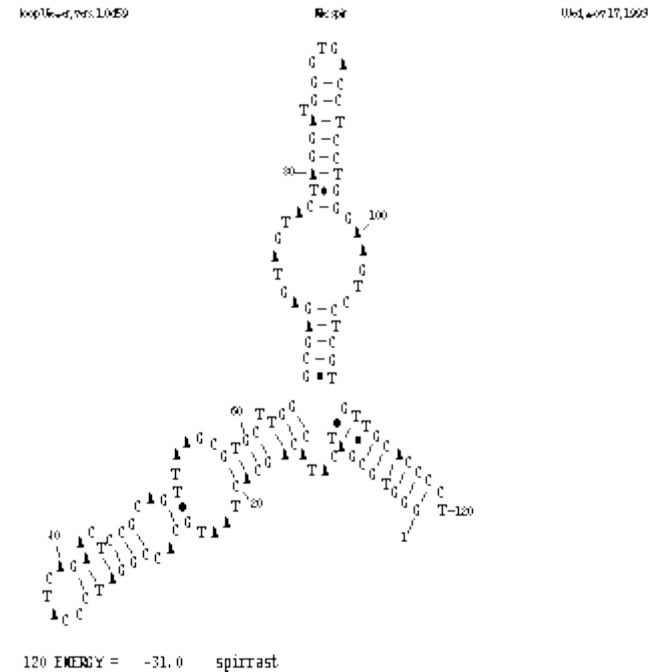
Electrophoresis

- Separation of charged molecules in an electric field.
- Nucleic acids have 1 charged phosphate (- charge) per nucleotide. This implies a constant charge to mass ratio. Thus, separation is based almost entirely on length: longer molecules move slower.



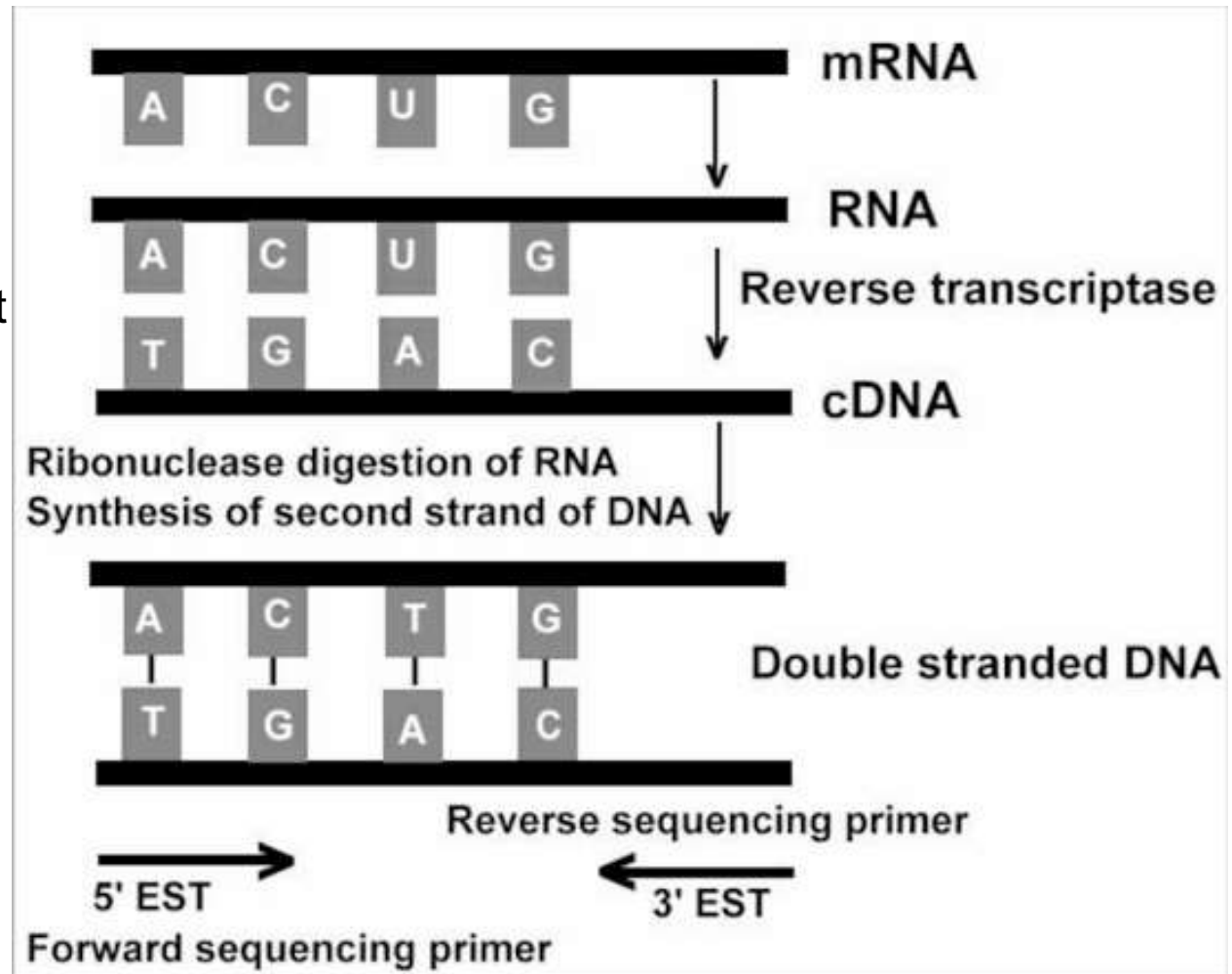
RNA

- Gene expression is studied using RNA. However, RNA has two annoying properties:
 - it is very easily degraded. A desirable property in the cell: allows rapid response to environmental changes
 - It usually has a lot of secondary structure. This means that migration speed in electrophoresis is not proportional to length. The same problem occurs with proteins.
- For sequencing, RNA is often converted to DNA (called cDNA).



cDNA Synthesis

- use oligo-dT primer, which binds to poly-A tail.
- You can also use random primers (short oligonucleotides that bind to random locations on the mRNA).
- make the first DNA strand from the RNA using reverse transcriptase



Expressed Sequence Tags

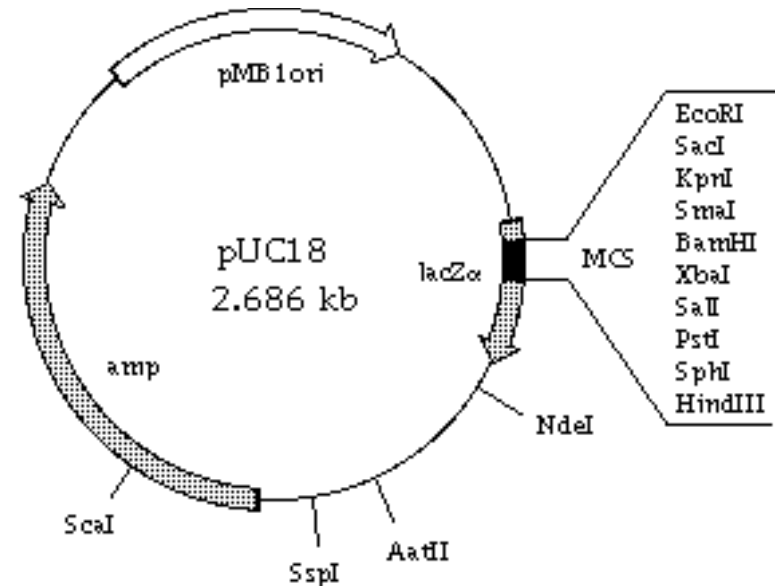
- ESTs are cDNA clones that have has a single round of sequencing done from one end.
- First extract mRNA from a given tissue and/or environmental condition. Then convert it to cDNA and clone.
- Sequence thousands of EST clones and save the results in a database.
- A search can then show whether your sequence was expressed in that tissue.
 - quantitation issues: some mRNAs are present in much higher concentration than others. Many EST libraries are “normalized” by removing duplicate sequences.
- Also can get data on transcription start sites and exon/intron boundaries by comparing to genomic DNA
 - but sometimes need to obtain the clone and sequence the rest of it yourself.

Cell-Based Molecular Cloning

- The original recombinant DNA technique: 1974 by Cohen and Boyer.
- Several key players:
 - 1. restriction enzymes. Cut DNA at specific sequences. e.g. EcoR1 cuts at GAATTC and BamH1 cuts at GGATCC.
 - Used by bacteria to destroy invading DNA: their own DNA has been modified (methylated) at the corresponding sequences by a methylase.
 - 2. Plasmids: independently replicating DNA circles (only circles replicate in bacteria). Foreign DNA can be inserted into a plasmid and replicated.
 - Plasmids for cloning carry drug resistance genes that are used for selection.
 - Spread antibiotic resistance genes between bacterial species
 - 3. DNA ligase. Attaches 2 pieces of DNA together.
 - 4. transformation: DNA manipulated in vitro can be put back into the living cells by a simple process .
 - The transformed DNA replicates and expresses its genes.

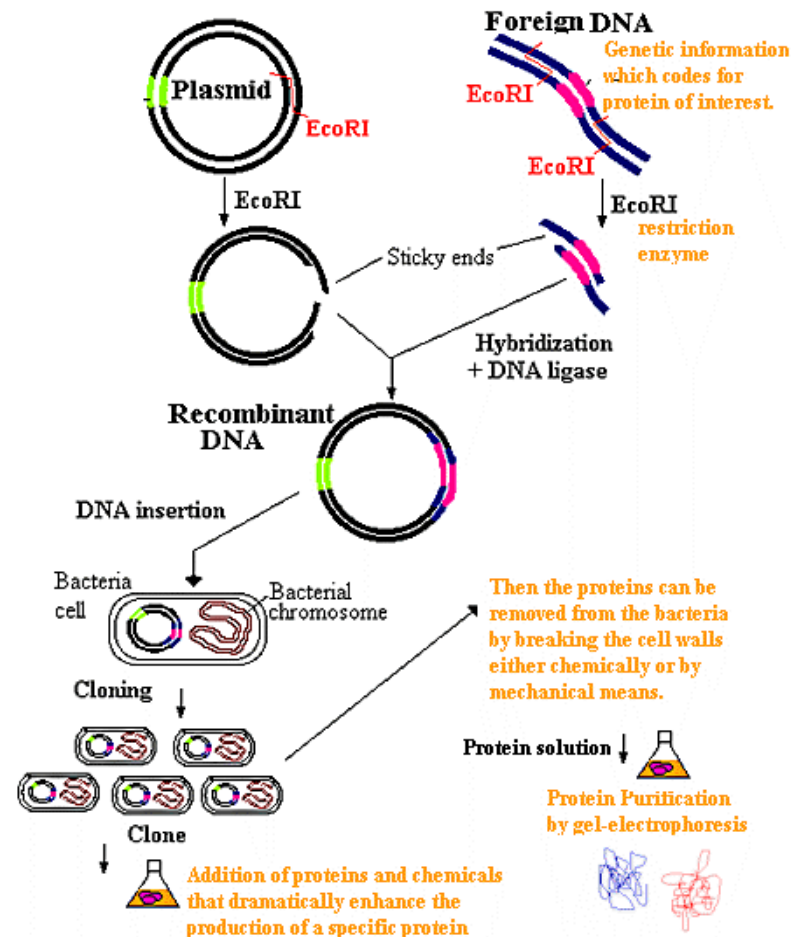
Plasmid Vectors

- To replicate, a plasmid must be circular, and it must contain a replicon, a DNA sequence that DNA polymerase will bind to and initiate replication. Also called “ori” (origin of replication).
 - Replicons are usually species-specific.
 - Some replicons allow many copies of the plasmid in a cell, while others limit the copy number or one or two.
- Plasmid cloning vectors must also carry a selectable marker: drug resistance. Transformation is inefficient, so bacteria that aren't transformed must be killed.
- Most cloning vectors have a multiple cloning site, a short region of DNA containing many restriction sites close together (also called a polylinker). This allows many different restriction enzymes to be used.
- Most cloning vectors use a system for detecting the presence of a recombinant insert, usually the blue/white beta-galactosidase system.



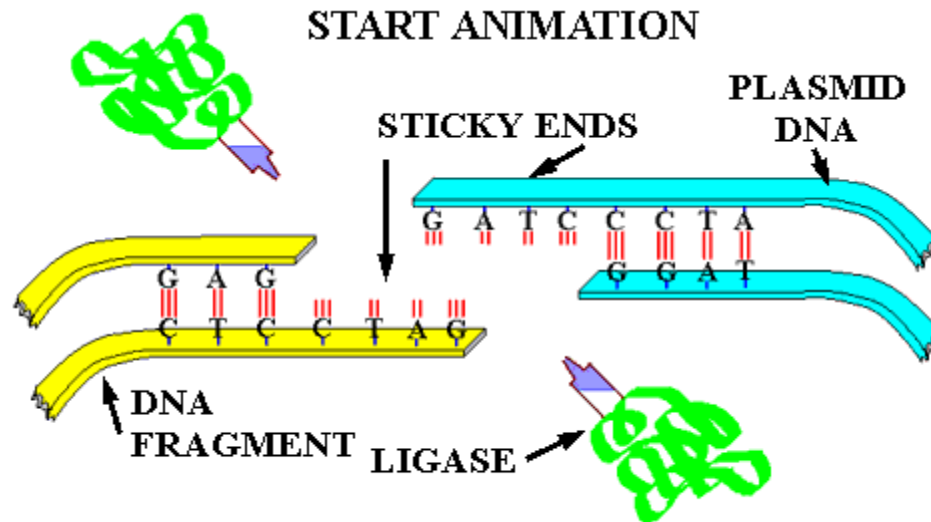
Basic Cloning Process

- Plasmid is cut open with a restriction enzyme that leaves an overhang: a sticky end
- Foreign DNA is cut with the same enzyme.
- The two DNAs are mixed. The sticky ends anneal together, and DNA ligase joins them into one recombinant molecule.
- The recombinant plasmids are transformed into E. coli using heat plus calcium chloride.
- Cells carrying the plasmid are selected by adding an antibiotic: the plasmid carries a gene for antibiotic resistance.



DNA Ligase in Action!

I hope



Cloning Vector Types

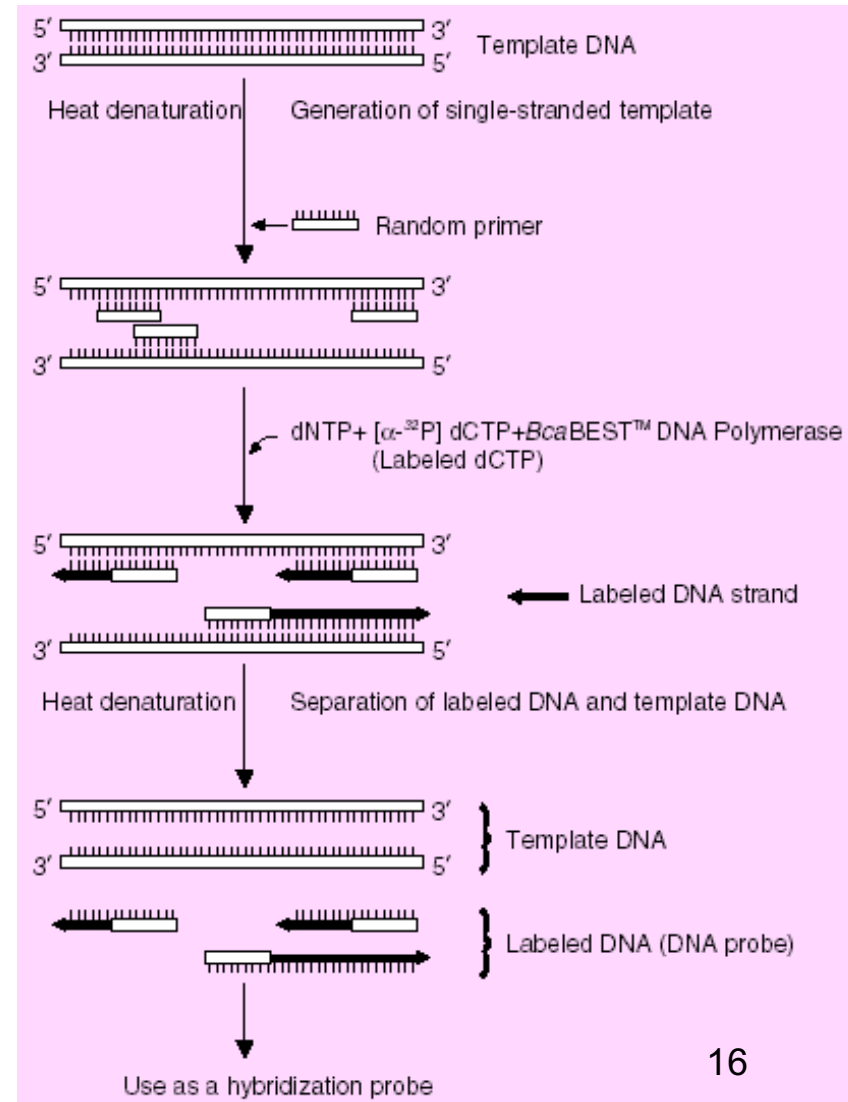
- For different sizes of DNA:
 - plasmids: up to 5 kb
 - phage lambda (λ) vectors (also cosmids): up to 50 kb
 - BAC (bacterial artificial chromosome): 300 kb
 - YAC (yeast artificial chromosome): 2000 kb
- Expression vectors: make RNA and protein from the inserted DNA
 - shuttle vectors: can grow in two different species, usually *E. coli* and something else

Hybridization

- The idea is that if DNA is made single stranded (melted), it will pair up with another DNA (or RNA) with the complementary sequence. If one of the DNA molecules is labeled, you can detect the hybridization.
- Basic applications:
 - Southern blot: DNA digested by a restriction enzyme then separated on an electrophoresis gel
 - Northern blot: uses RNA on the gel instead of DNA
 - *in situ* hybridization: probing a tissue
 - colony hybridization: detection of clones
 - microarrays

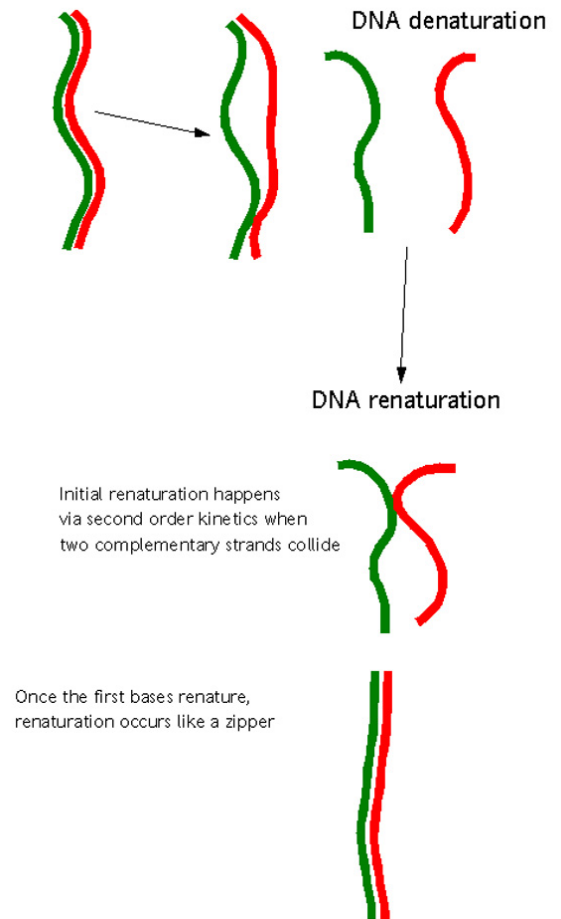
Labeling

- Several methods. One is random primers labeling:
 - use ^{32}P -labeled dNTPs
 - short random oligonucleotides as primers (made synthetically)
 - single stranded DNA template (made by melting double stranded DNA by boiling it)
 - DNA polymerase copies the DNA template, making a new strand that incorporates the label.
- Can also label RNA (sometimes called riboprobes), use non-radioactive labels (often a small molecule that labeled antibodies bind to, or a fluorescent tag), use other labeling methods.



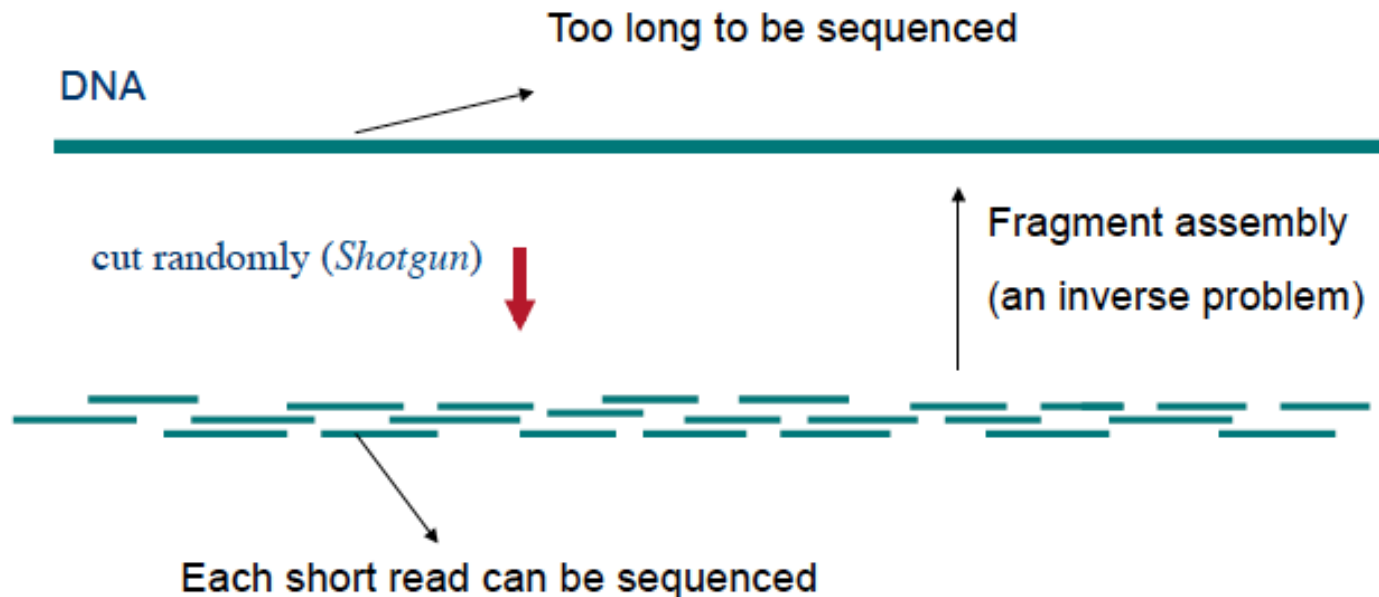
Hybridization Process

- All the DNA must be single stranded (melt at high temp or with NaOH). Occurs in a high salt solution at say 60°C. Complementary DNAs find each other and stick. Need to wash off non-specific binding.
- Stringency: how perfectly do the DNA strands have to match in order to stick together? Less than perfect matches will occur at lower stringency (e.g. between species). Increase stringency by increasing temp and decreasing salt concentration.
- Rate of hybridization depends on DNA concentration and time (Cot), as well as GC content and DNA strand length.
- Autoradiography. Put the labeled DNA next to X-ray film; the radiation fogs the film.



Overview of sequencing

- How genomes are sequenced?
 - Sample preparation
 - Sequencing
 - Assembly
- Sequence technologies
 - Sanger sequencing
 - Next generation sequencing (NGS)



Clone library preparation

Genomes need to be broken into pieces

Limitation on read length (30 – 1000 bp)

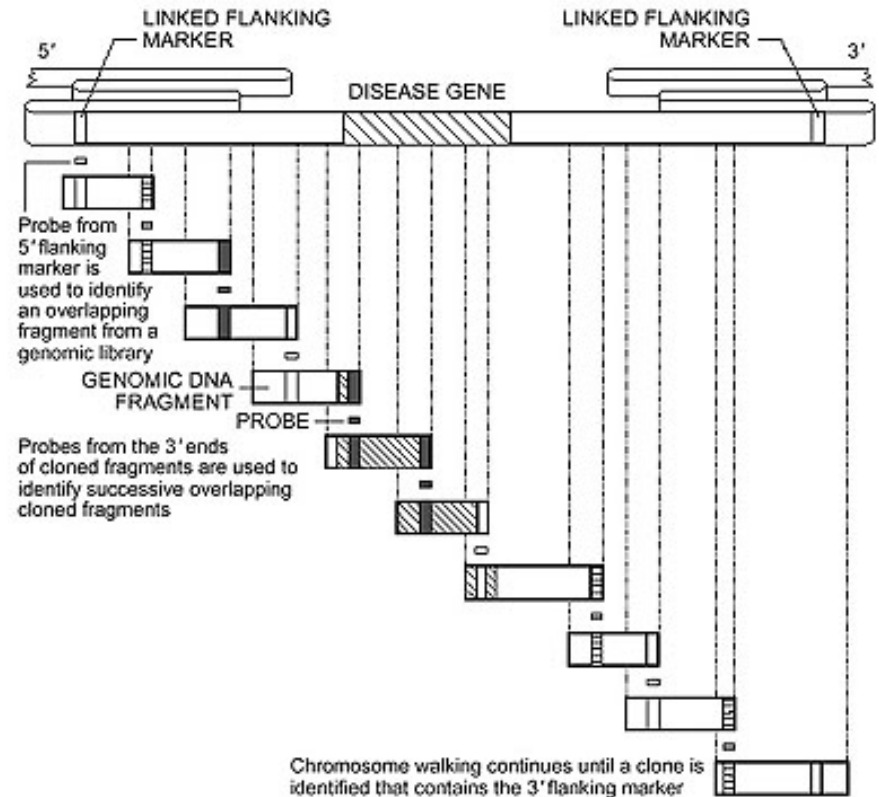
Sequencing of long DNA sequences (a chromosome or a whole genome) relies on sequencing of short segments

Short pieces will have a lot of overlaps and they will be put into cloning vectors for sequencing

Human genome is 3×10^9 bases long, so you will have at least 3×10^6 such pieces. If you want to have 10 coverages (overlaps for assembly), that is 3×10^7 pieces to be sequenced

Chromosome walking

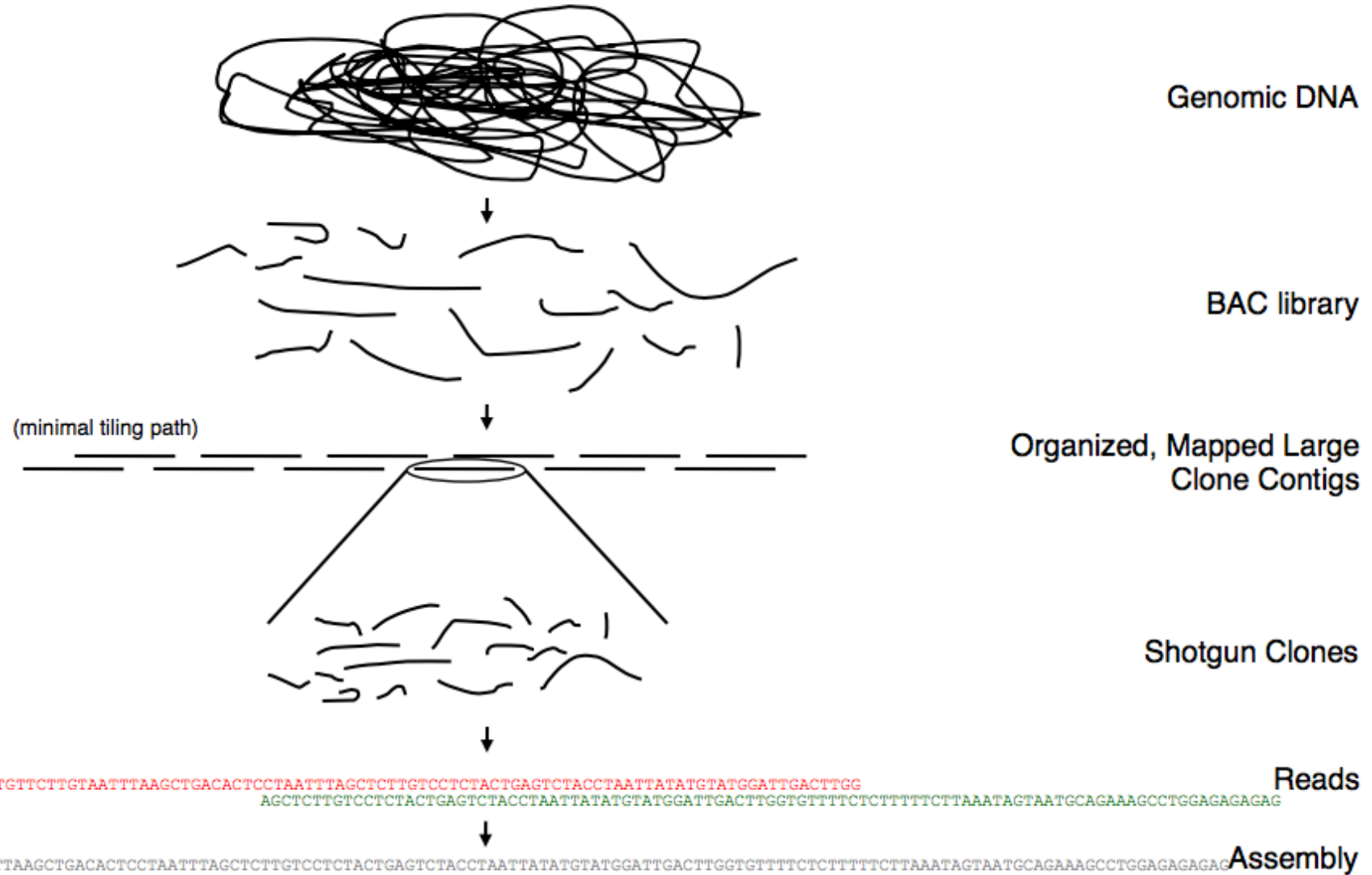
- DNA is sequenced in very small fragments: at most, 1000 bp. Compare this to the size of the human genome: 3,000,000,000 bp. How to get the complete sequence?
- In the early days (1980's), genome sequencing was done by chromosome walking (aka primer walking): sequence a region, then make primers from the ends to extend the sequence. Repeat until the target gene was reached.
 - The cystic fibrosis gene was identified by walking about 500 kbp from a closely linked genetic marker, a process that took a long time and was very expensive.
 - Still useful for fairly short DNA molecules, say 1-10 kbp.




Whole Genome Sequencing Approaches

Hierarchical Shotgun Approach

Clone library based



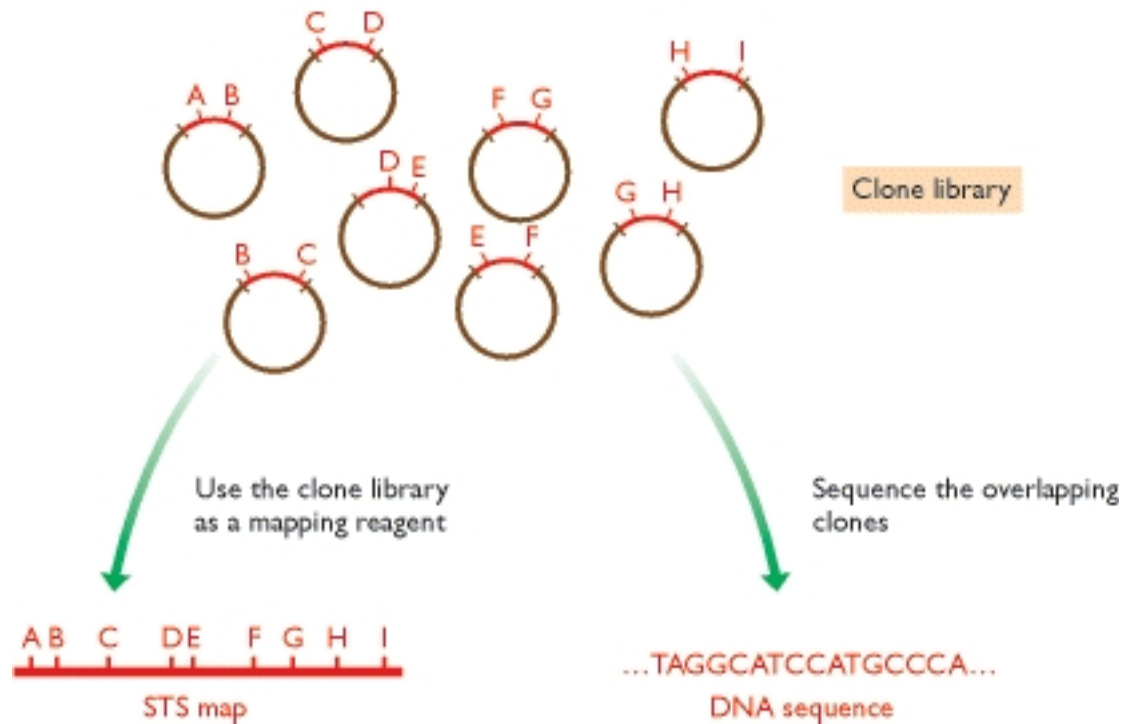
 The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

A **genomic library** is a collection of the total genomic [DNA](#) from a single [organism](#). The DNA is stored in a population of identical [vectors](#), each containing a different [insert](#) of DNA. In order to construct a genomic library, the organism's DNA is [extracted](#) from [cells](#) and then digested with a [restriction enzyme](#) to cut the DNA into fragments of a specific size. The fragments are then inserted into the vector using the enzyme, [DNA ligase](#).^[1] Next, the vector DNA can be taken up by a host organism- commonly a population of [Escherichia coli](#) or [yeast](#)- with each cell containing only one vector molecule. Using a host cell to carry the vector allows for easy [amplification](#) and retrieval of specific [clones](#) from the [library](#) for analysis.

Vector type	Insert size (thousands of bases)
Plasmids	up to 15
Phage lambda (λ)	up to 25
Cosmids	up to 45
Bacteriophage P1	70 to 100
P1 artificial chromosomes (PACs)	130 to 150
Bacterial artificial chromosomes (BACs)	120 to 300
Yeast artificial chromosomes (YACs)	250 to 2000

http://en.wikipedia.org/wiki/Genomic_library

The value of clone libraries in genome projects. The small clone library shown in this example contains sufficient information for an STS map to be constructed, and can also be used as the source of the DNA that will be sequenced

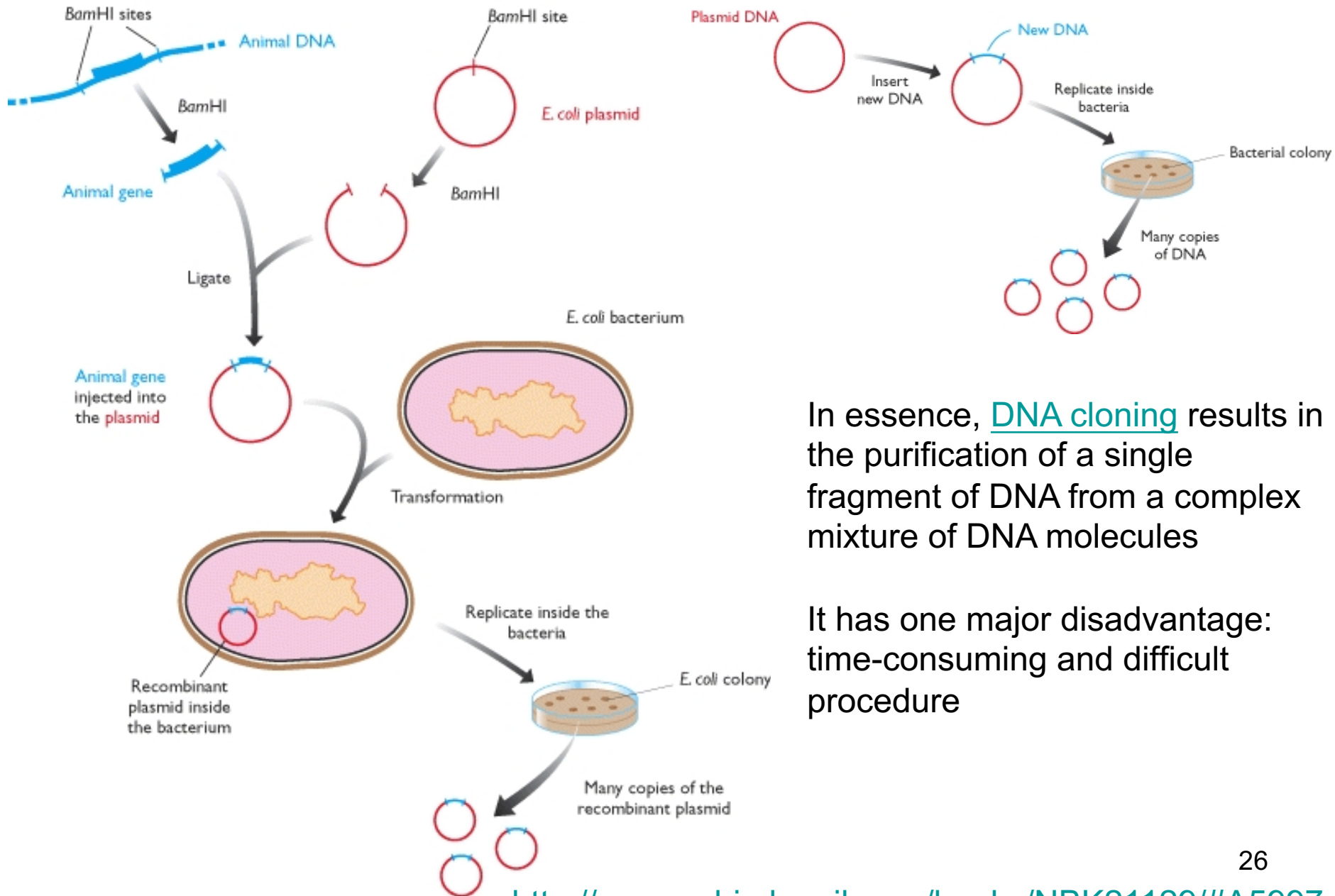


The STS (sequence tagged sites) data can be used to anchor this sequence precisely onto the physical map

Physical mapping technique has been responsible for generation of the most detailed maps of large genomes, e.g. [STS mapping](#). [A sequence tagged site](#) or **STS** is simply a short [DNA](#) sequence, generally between 100 and 500 [bp](#) in length, that is easily recognizable and **occurs only once in the chromosome** or genome being studied.

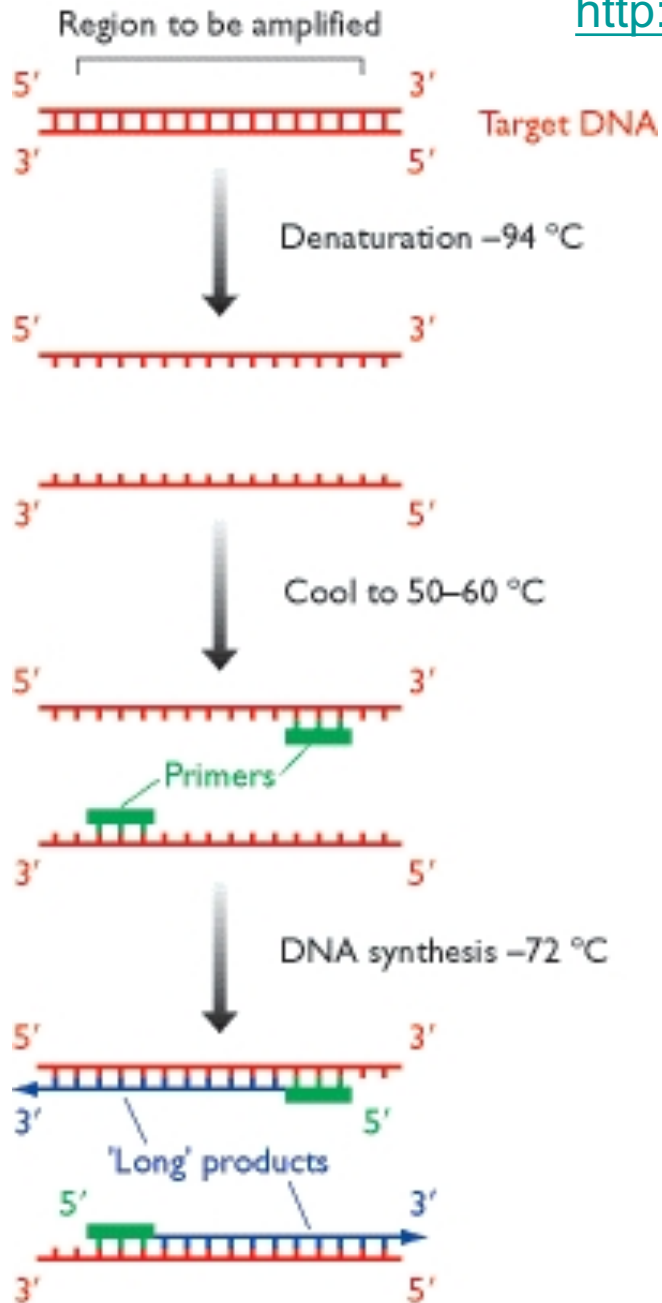
To qualify as an [STS](#), a [DNA](#) sequence must satisfy two criteria. The first is that its **sequence must be known**, so that a [PCR](#) assay can be set up to test for the presence or absence of the STS on different DNA fragments. The second requirement is that the STS must **have a unique location in the chromosome** being studied, or in the genome as a whole if the DNA fragment set covers the entire genome.

Molecular cloning to insert a genome fragment into ecoli plasmids and propagate



In essence, DNA cloning results in the purification of a single fragment of DNA from a complex mixture of DNA molecules

It has one major disadvantage: time-consuming and difficult procedure



PCR complements DNA cloning in that it enables the same result to be achieved - purification of a specified DNA fragment - but in a much shorter time, perhaps just a few hours (Saiki *et al.*, 1988).

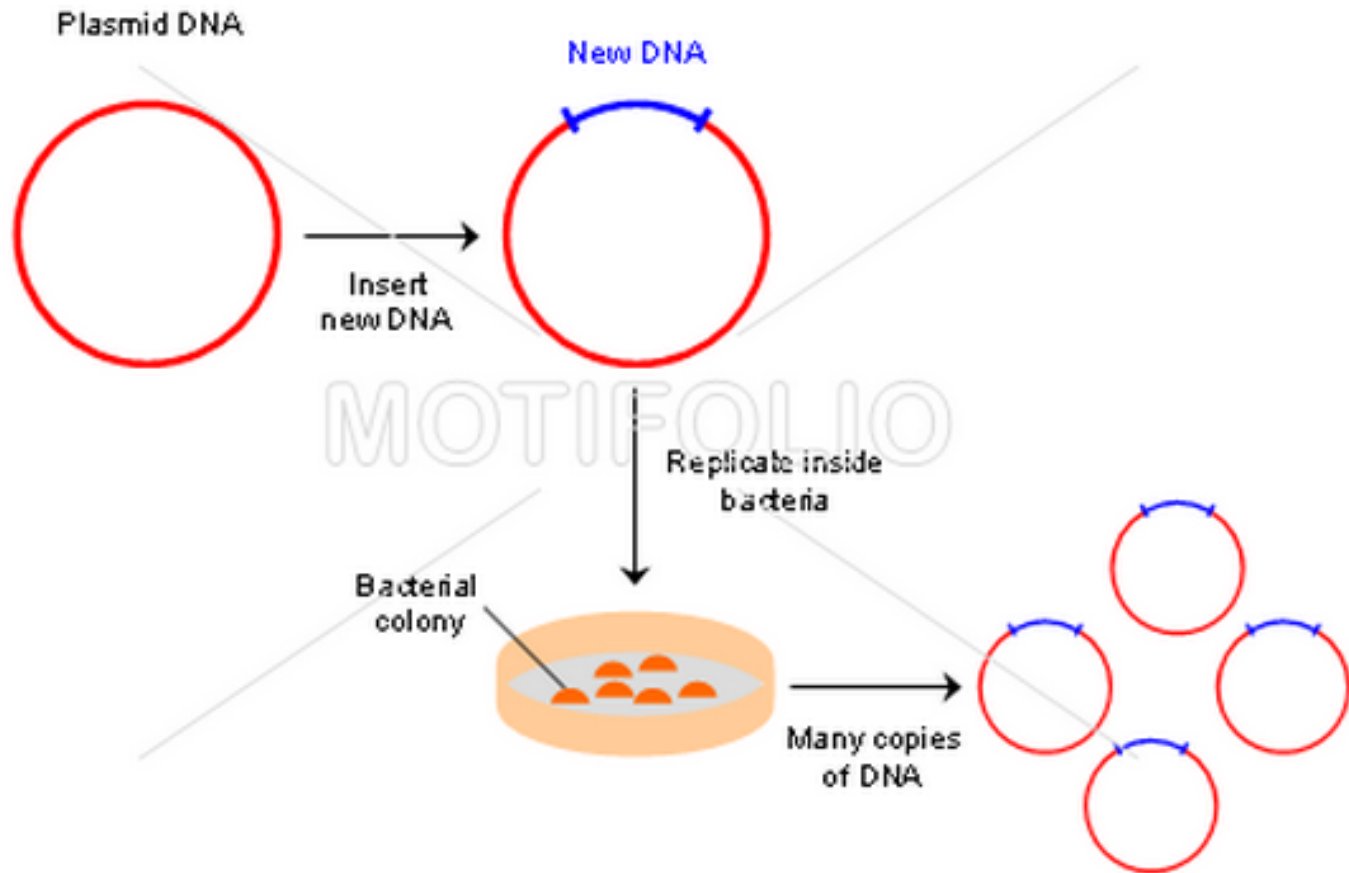
PCR is complementary to, not a replacement for, cloning because it has its own limitations, the most important of which is the need to know the sequence of at least part of the fragment that is to be purified

Cloning vectors

Cloning Vectors: DNA vehicles in which a foreign DNA can be inserted and stay stable

Various types:

- Cosmid (plasmid, containing 37-52 kbp of DNA)
- BAC (Bacterial Artificial Chromosome; takes in 100-300 kbp of foreign DNA)
- YAC (Yeast Artificial Chromosome)



Clone foreign DNAs into BAC

A bacterial artificial chromosome (BAC) is **an engineered DNA molecule used to clone DNA sequences in bacterial cells** (for example, *E. coli*). BACs are often used in connection with DNA sequencing. Segments of an organism's DNA, ranging from 100,000 to about 300,000 base pairs, can be inserted into BACs. The BACs, with their inserted DNA, are then taken up by bacterial cells. As the **bacterial cells grow and divide, they amplify the BAC DNA**, which can then be isolated and used in sequencing DNA

Whole Genome Sequencing Approaches

Shotgun Approach



Genomic DNA



Shotgun Clones



<http://www.bio.davidson.edu/genomics/method/shotgun.html>

GCAATGAAATATGTTCTTGAATTAAGCTGACACTCCTAATTTAGCTCTTGTCCCTCTACTGAGTCTACCTAATTATATGTATGGATTGACTTGG
AGCTCTTGTCCCTCTACTGAGTCTACCTAATTATATGTATGGATTGACTTGGTGTCTCTTTTCTTAAATAGTAATGCAGAAAGCCTGGAGAGAGAG

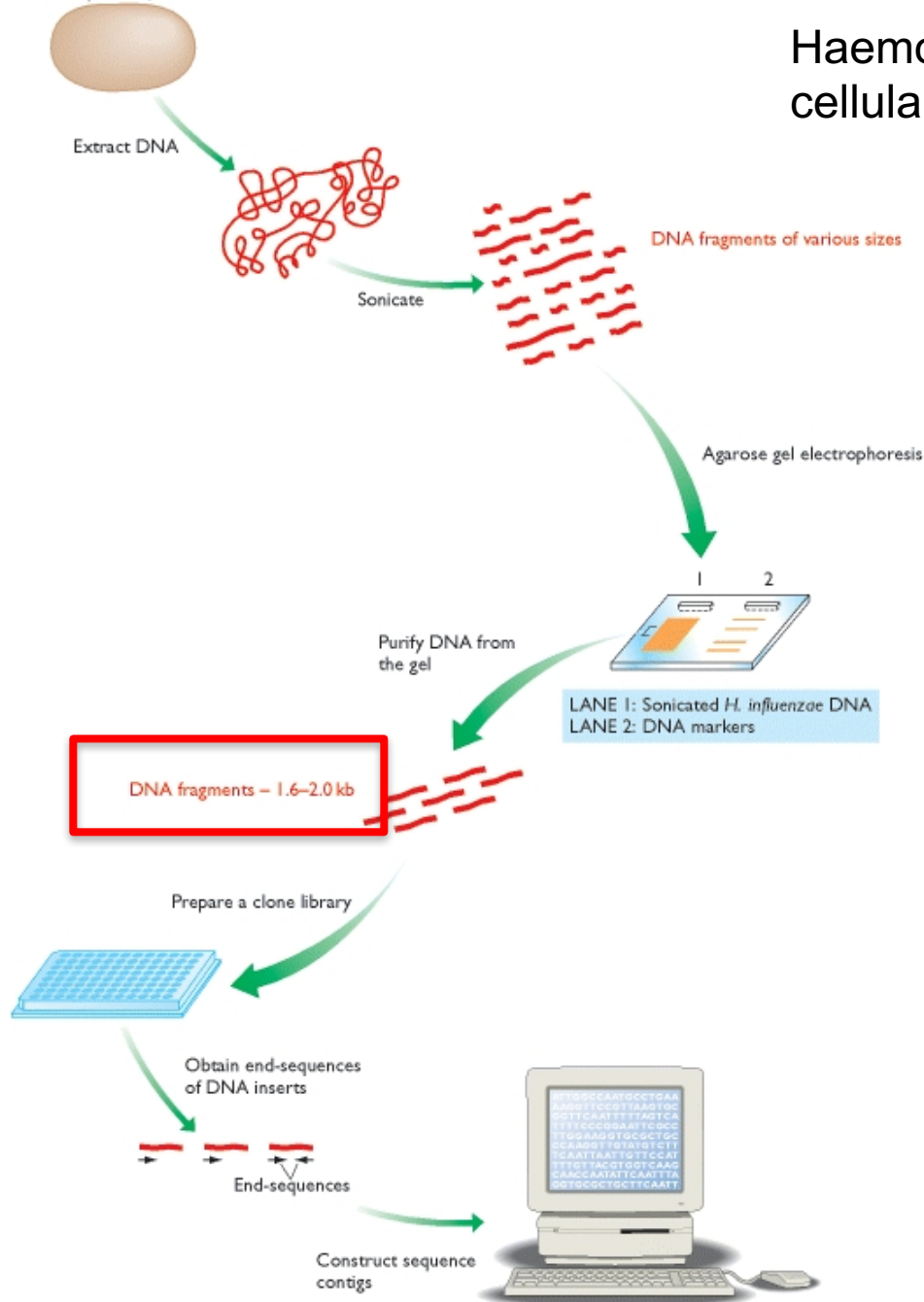
Reads



ATGTTCTTGAATTAAGCTGACACTCCTAATTTAGCTCTTGTCCCTCTACTGAGTCTACCTAATTATATGTATGGATTGACTTGGTGTCTCTTTTCTTAAATAGTAATGCAGAAAGCCTGGAGAGAGAG

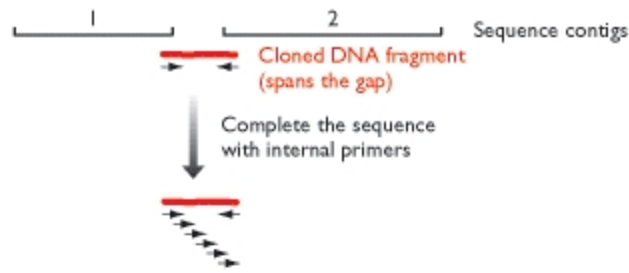
Assembly

Haemophilus influenzae: the first sequenced cellular organism ever, in 1995, using WGS

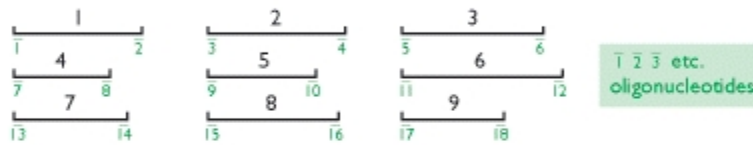


H. influenzae DNA was sonicated and fragments with sizes between 1.6 and 2.0 kb purified from an agarose gel and ligated into a plasmid vector to produce a clone library. End sequences were obtained from clones taken from this library, and a computer used to identify overlaps between sequences. This resulted in 140 sequence contigs, which were assembled into the complete genome sequence

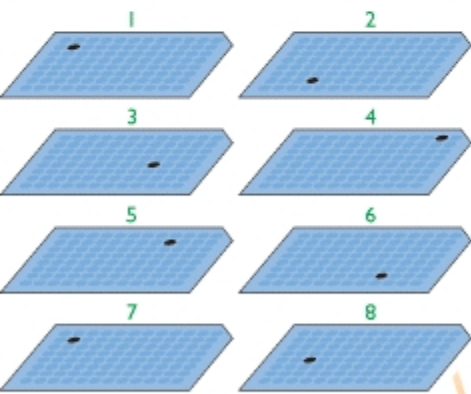
(A) Closing a 'sequence gap'



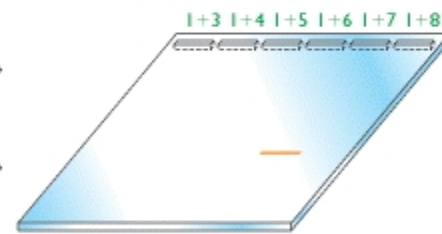
(B) Closing a 'physical gap'



Probe a second clone library with oligonucleotides



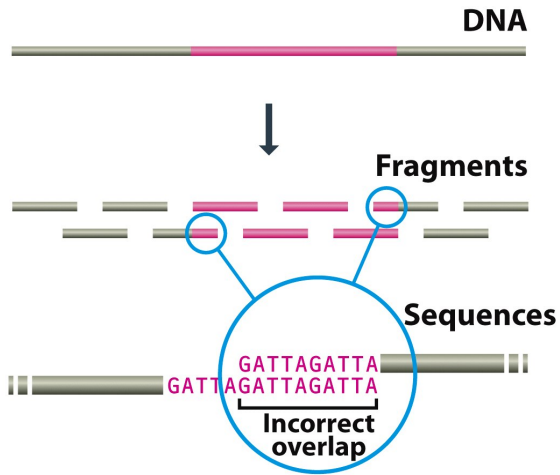
PCR with pairs of oligonucleotides



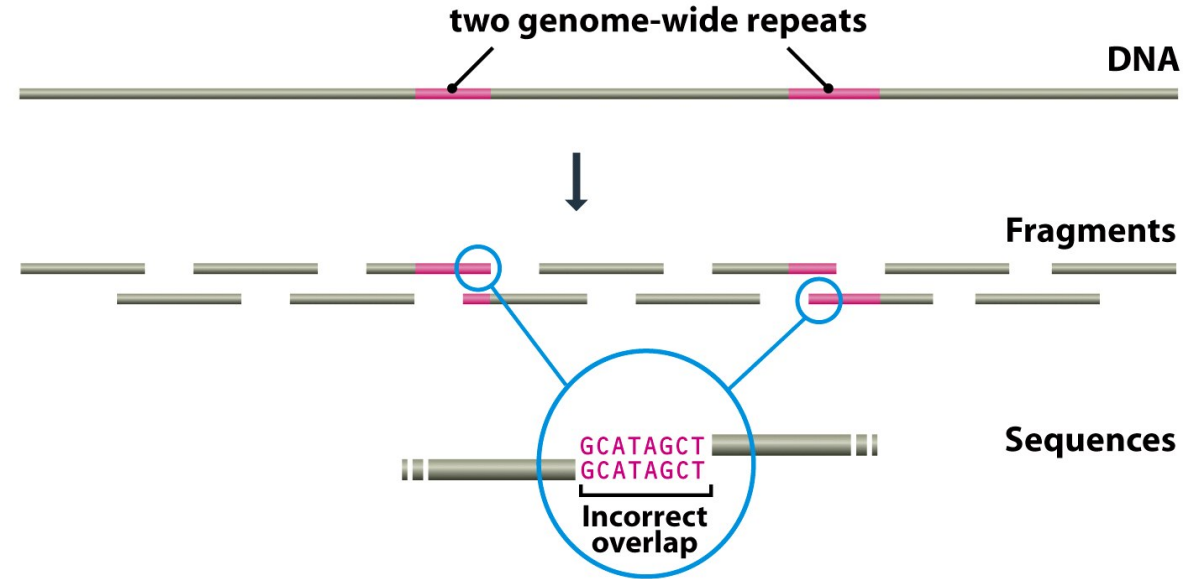
CONCLUSION:
Contigs 1 and 4 are adjacent

- (A) 'Sequence gaps' are ones which can be closed by further sequencing of clones already present in the library. In this example, the end-sequences of contigs 1 and 2 lie within the same plasmid clone, so further sequencing of this [DNA](#) insert with internal primers (see [Figure 6.5B](#)) will provide the sequence to close the gap.
- (B) 'Physical gaps' are stretches of sequence that are not present in the clone library, probably because these regions are unstable in the cloning vector that was used. Two strategies for closing these gaps are shown. On the left, a second clone library, prepared with a bacteriophage λ vector rather than a plasmid vector, is probed with oligonucleotides corresponding to the ends of the contigs. Oligonucleotides 1 and 7 both hybridize to the same clone, whose insert must therefore contain DNA spanning the gap between contigs 1 and 4. On the right, PCRs are carried out with pairs of oligonucleotides. Only numbers 1 and 7 give a [PCR](#) product, confirming that the contig ends represented by these two oligonucleotides are close together in the genome.

Problems with tandemly repeated DNA



Problems with genome-wide repeats



Problem with the shotgun method is that it can lead to errors when repetitive regions of a genome are analyzed. When a repetitive sequence is broken into fragments, many of the resulting pieces contain the same, or very similar, sequence motifs. It would be very easy to reassemble these sequences so that a portion of a repetitive region is left out, or even to connect together two quite separate pieces of the same or different chromosomes

The difficulties in applying the shotgun method to a large molecule that has a significant repetitive [DNA](#) content means that this approach cannot be used on its own to sequence a eukaryotic genome. **Instead, a genome [map](#) must first be generated. A genome map provides a guide for the sequencing experiments by showing the positions of genes and other distinctive features.** Once a genome map is available, the sequencing phase of the project can proceed in either of two ways ([Figure 5.3](#)):

1. By the **whole-genome shotgun** method ([Section 6.2.3](#)), but uses the distinctive features on the genome map as landmarks to aid assembly of the master sequence. Reference to the map also ensures that regions containing repetitive [DNA](#) are assembled correctly.
2. By the **[clone contig approach](#)** ([Section 6.2.2](#)). In this method the genome is broken into manageable segments, each a few hundred [kb](#) or a few [Mb](#) in length, which are short enough to be sequenced accurately by the shotgun method. Once the sequence of a segment has been completed, it is positioned at its correct location on the map. This step-by-step approach takes longer than whole-genome shotgun sequencing, but is thought to produce a more accurate and error-free sequence.

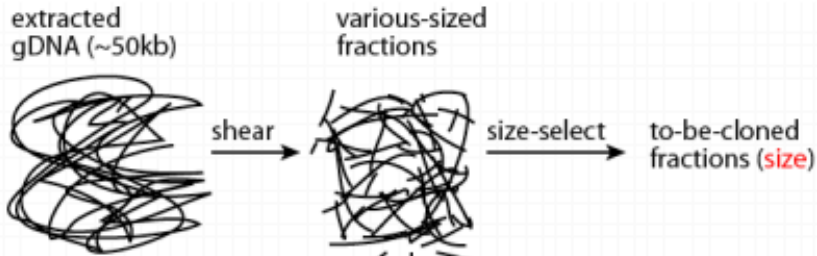
Rationale for Hierarchical Strategy

- Better for a repeat-rich genome
 - *less misassembly of finished genome*
 - *long-range misassembly largely eliminated and short-range reduced*
- Better for an outbred organism
 - *each clone from an individual and no polymorphisms in the final sequence.*
 - *(Added bonus: get SNPs from regions of overlapping clones)*
 - *Can also get some haplotype information, if individual BACs shotgun sequenced.*
- Better if there are cloning biases
 - *use minimum tiling path, so the same coverage for each region*
- Easier to identify and fill gaps (from unclonable regions) sooner

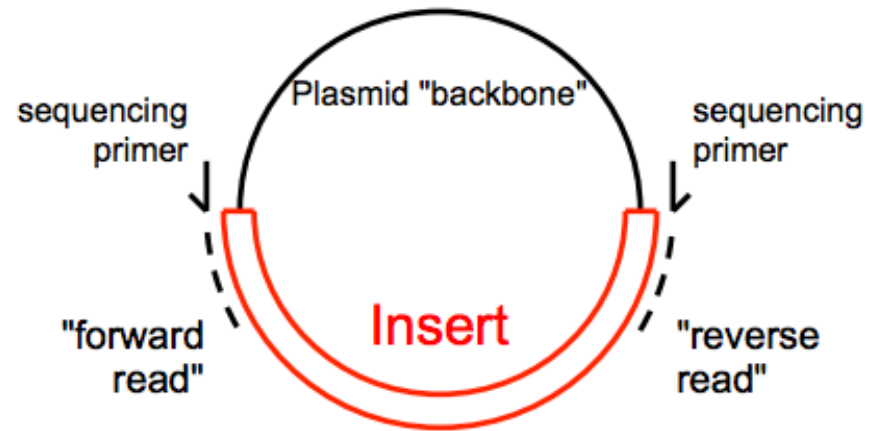
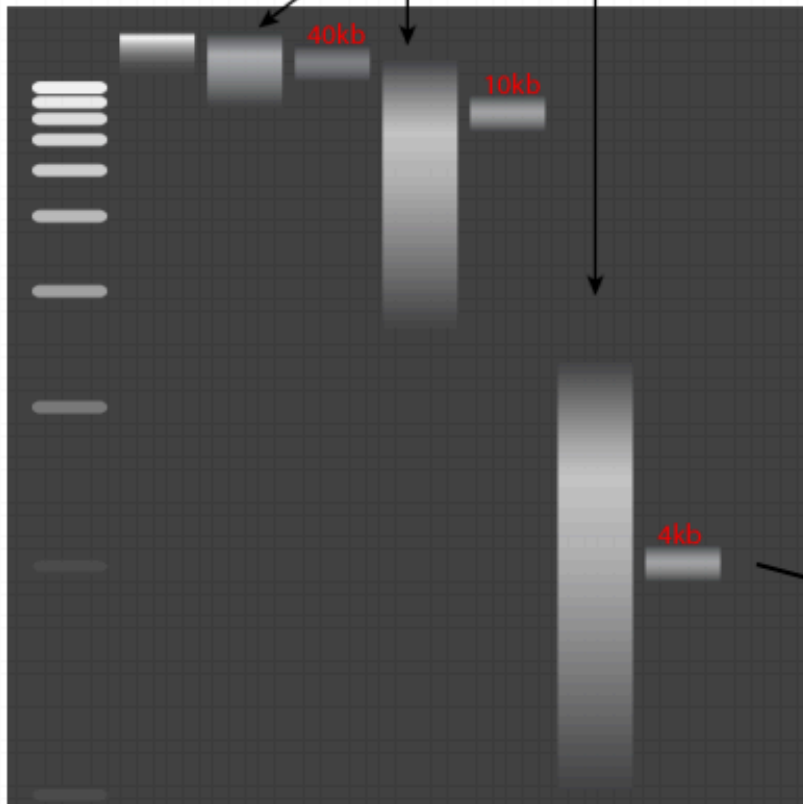
BUT

- **Time consuming and expensive to make minimum tiling path**

De Novo Whole Genome Sequencing



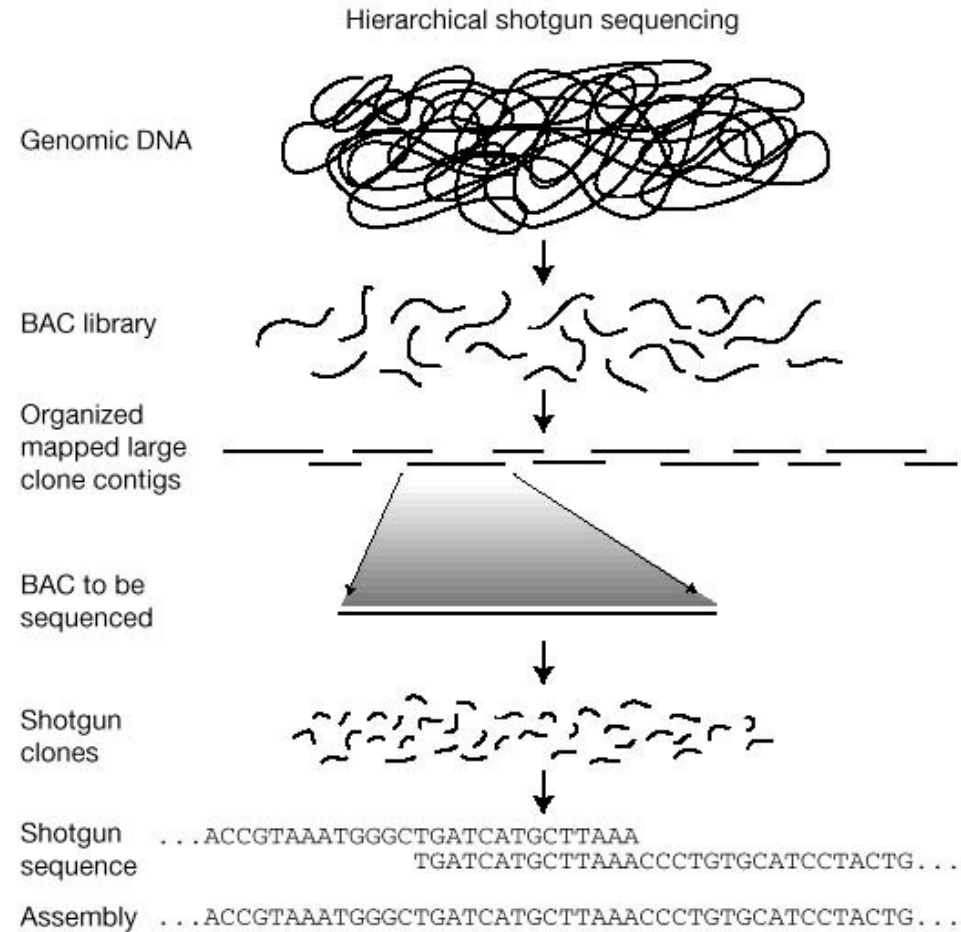
Make millions of random clones: "Shotgunning"



Sequencing tech

Shotgun Sequencing

- Shotgun sequencing is what is typically done today: DNA is fragmented randomly and enough fragments are sequenced so each base is read 10 times or more on average. The overlapping fragments (“reads”) are then assembled into a complete sequence.
- For large genomes, hierarchical shotgun sequencing is a useful technique: first break up the genome into an ordered set of cloned fragments (scaffolds), usually BAC clones. Each BAC is shotgun sequenced separately.



Generations of sequencing technologies

- 1st gen: 1977-2007 Sanger sequencen
- 2nd gen 2007-2013 (Next gen) massive parallel sequencing
- 3rd gen 2012-? Single molecule sequencing

- NGS (2nd gen) platforms
 - Illumina (former: Solexa)
 - SoLID: Life technologies (former: Applied Biosystems) (SOLID)
 - Roche (Pyrosequencing)
 - Ion Torrent (semiconductor hydrogen ion detection)
- 3rd gen platforms
 - Helicos
 - Pac bio
 - Oxford Nanopore

Four Fundamentally Different Types of Chemistries for DNA Sequencing

“Chemistry”: Jargon; meaning the specific molecular biology and chemistry utilized in the sequencing reactions

- **Maxam-Gilbert**
 - chemical degradation of DNA
 - obsolete
- **Sequencing by synthesis (“SBS”)** Aka chain-termination method
 - uses DNA polymerase in a primer extension reaction
 - most common approach
 - Sanger developed it (“Sanger sequencing”)
 - 454, Illumina, Pacific Biosciences, Ion Torrent
- **Ligation-based**
 - sequencing using short probes that hybridize to the template
 - novel approach
 - SOLiD, Complete Genomics
- **Other**
 - Nanopore

DNA sequencing with chain-terminating inhibitors

(DNA polymerase/nucleotide sequences/bacteriophage ϕ X174)

F. SANGER, S. NICKLEN, AND A. R. COULSON

Medical Research Council Laboratory of Molecular Biology, Cambridge CB2 2QH, England

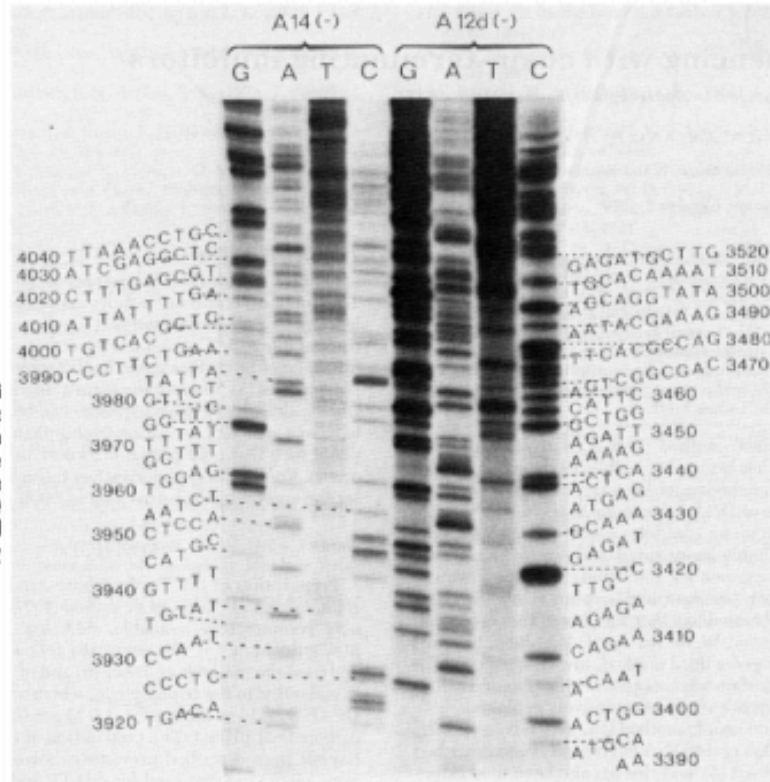
Contributed by F. Sanger, October 3, 1977

ABSTRACT A new method for determining nucleotide sequences in DNA is described. It is similar to the "plus and minus" method [Sanger, F. & Coulson, A. R. (1975) *J. Mol. Biol.* 94, 441-448] but makes use of the 2',3'-dideoxy and arabinonucleoside analogues of the normal deoxynucleoside triphosphates, which act as specific chain-terminating inhibitors of DNA polymerase. The technique has been applied to the DNA of bacteriophage ϕ X174 and is more rapid and more accurate than either the plus or the minus method.

a stereoisomer of ribose in which the 3'-hydroxyl group is oriented in *trans* position with respect to the 2'-hydroxyl group. The arabinosyl (ara) nucleotides act as chain terminating inhibitors of *Escherichia coli* DNA polymerase I in a manner comparable to ddT (4), although synthesized chains ending in 3' araC can be further extended by some mammalian DNA polymerases (5). In order to obtain a suitable pattern of bands from which an extensive sequence can be read it is necessary

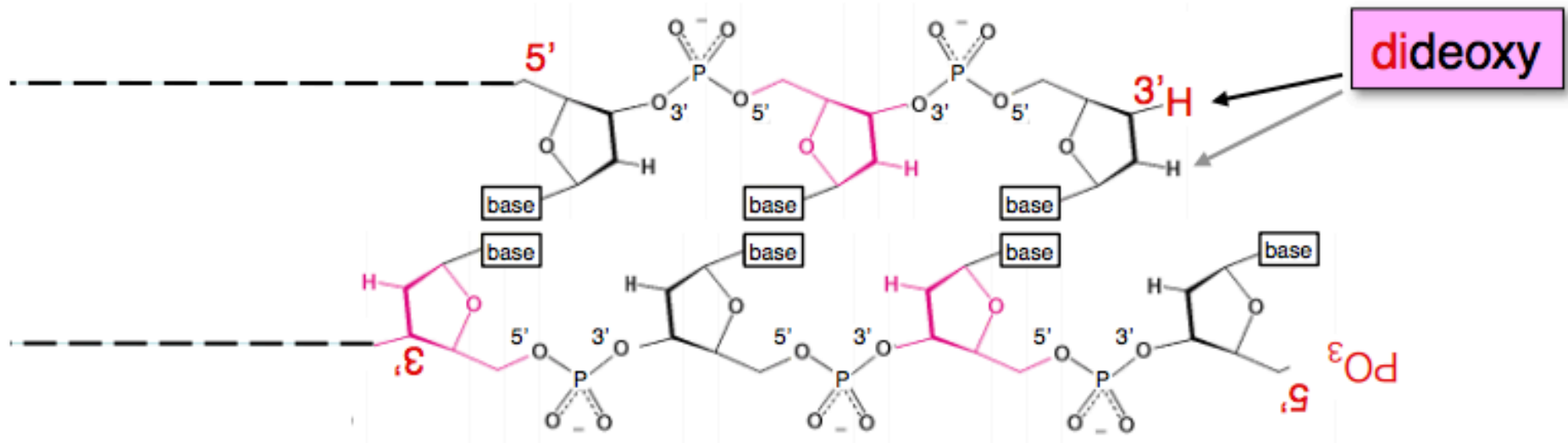


Frederick Sanger

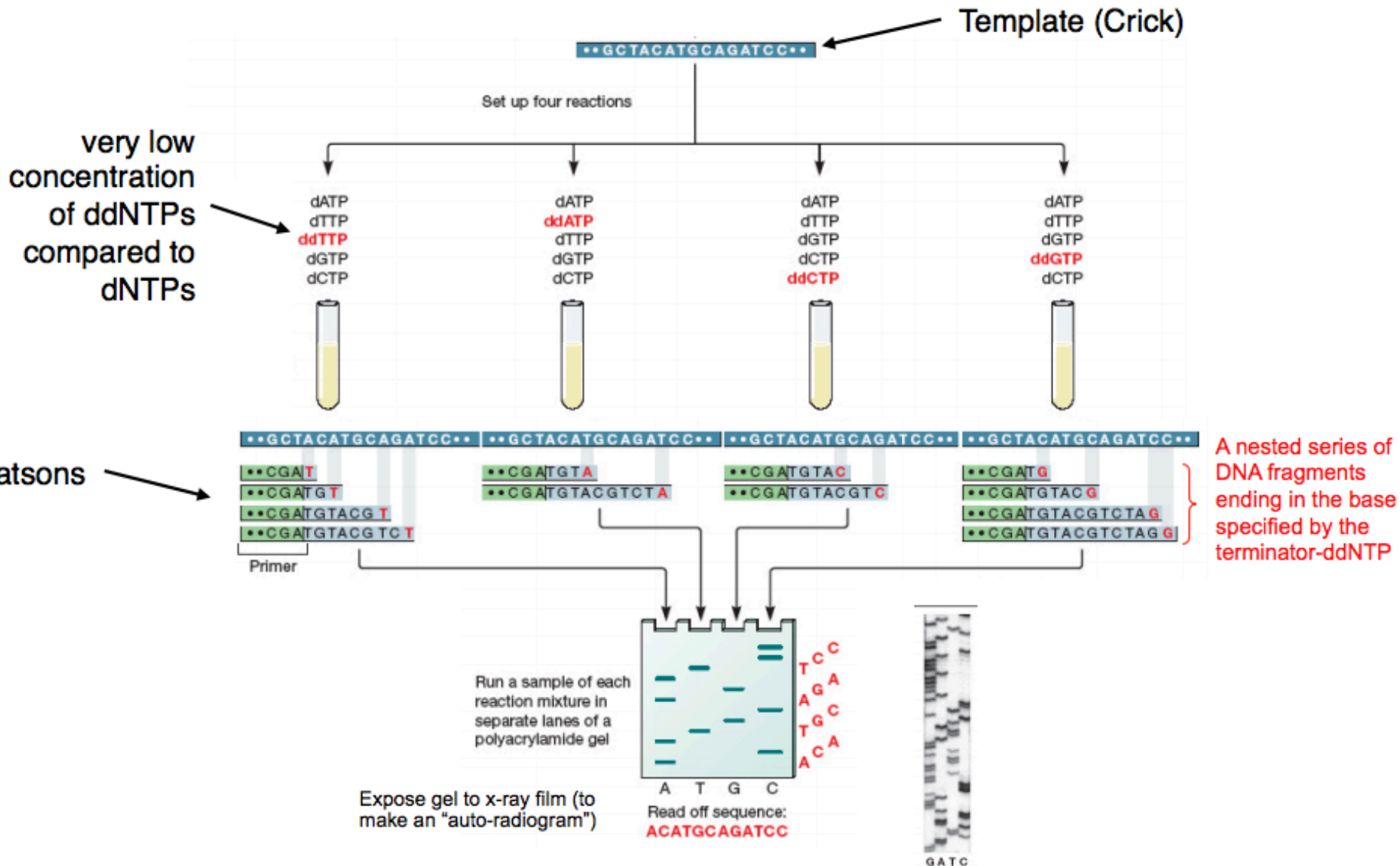


Chain Terminator: ddNTP

- Dideoxy nucleotides cannot be further extended, and so terminate the sequence chain



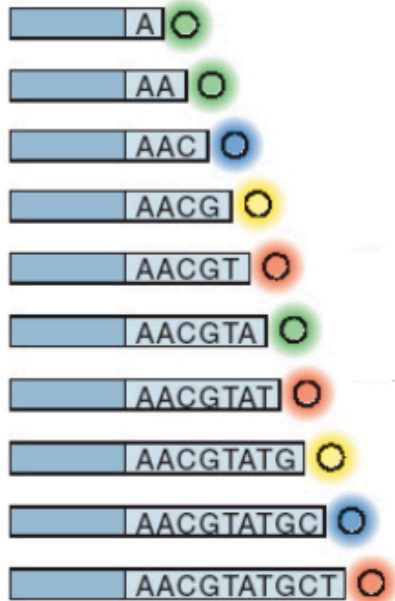
ddNTP is randomly incorporated and stops the synthesis



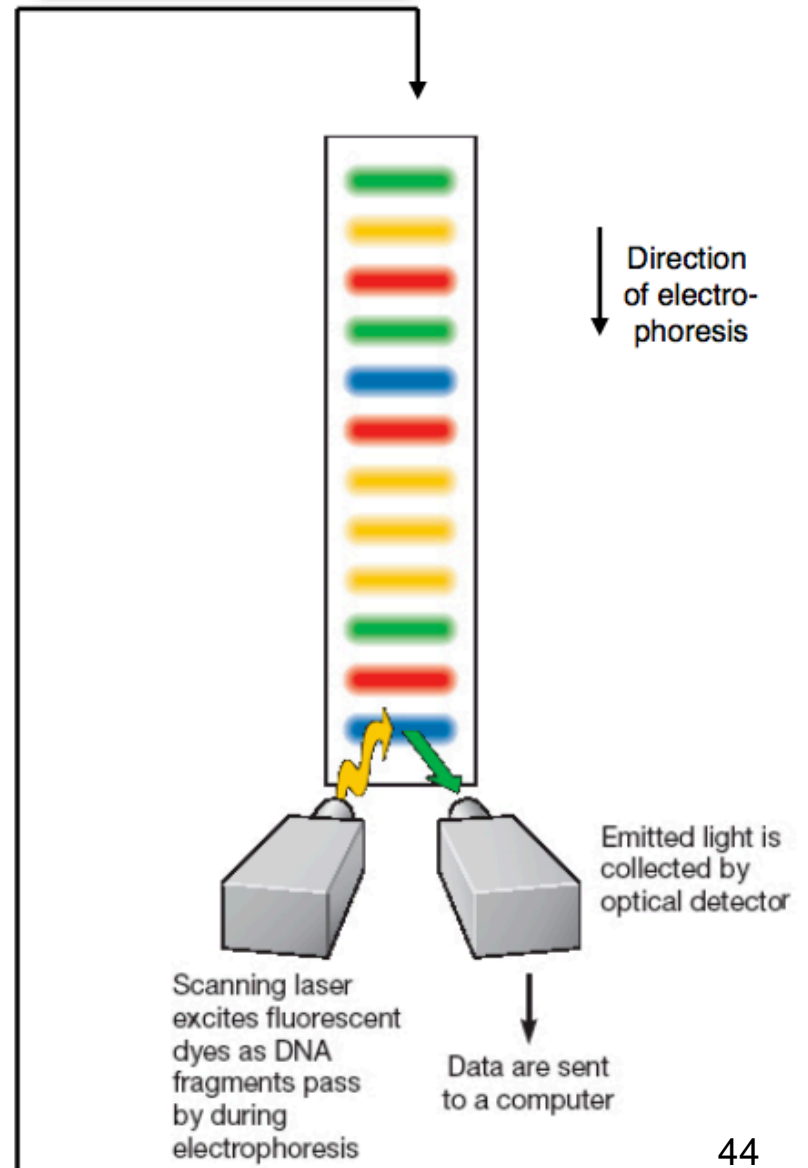
Each of the 4 ddNTPs is labeled with a different fluorescent dye

dGTP	+	ddGTP	○
dATP	+	ddATP	○
dTTP	+	ddTTP	○
dCTP	+	ddCTP	○

One-tube sequencing reaction
(note: cycle sequencing with modified Taq Polymerase)



Load on gel
(modern machines use capillaries, not slab gels)



Fluorescent Sanger Sequencing Trace

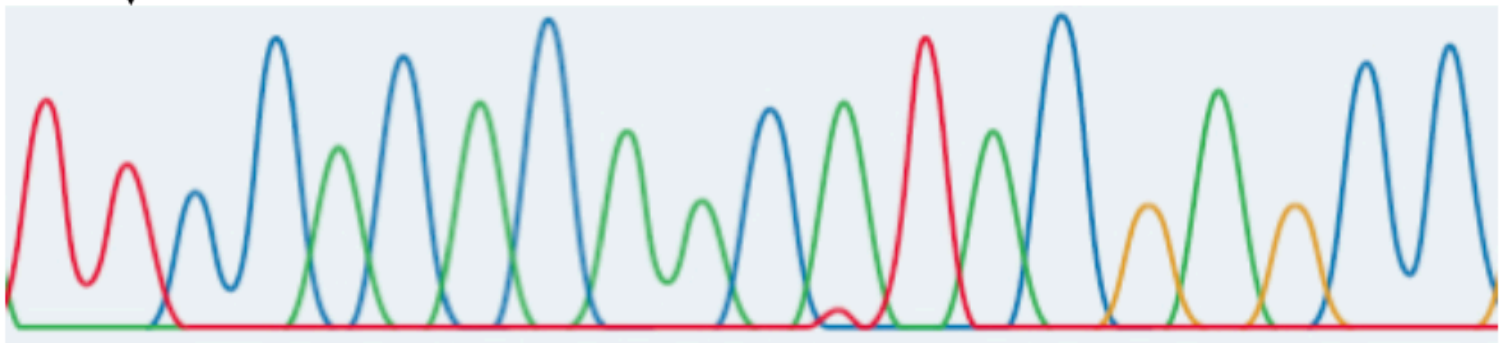
Lane signal



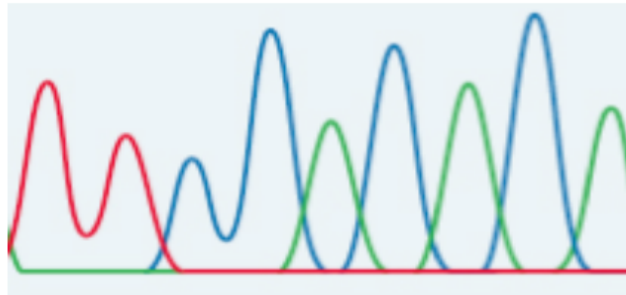
(Real fluorescent signals from a lane/capillary are much uglier than this).

Various algorithms to boost signal/noise, correct for dye-effects, mobility differences, etc., generates the 'final' trace (for each capillary of the run)

Trace



Base calling: image process to convert optical signal/noise to a base letter and a score



Sequence trace files contain the raw data output from automated sequencing instruments

After base calling, sequences are read out and that's why they are called "sequence reads"

Base Caller (Phred)

...	44	45	46	47	48	49	50	51	52	53	54	55	...	718	719	720	...
...	N	A	G	C	G	T	T	C	C	G	C	G	...	A	N	N	...
...	0	3	20	25	40	88	95	99	99	99	99	99	...	10	0	0	...

Quality score = $-10 * \log(\text{probability of error})$

For Q20, probability of error = $1/100$

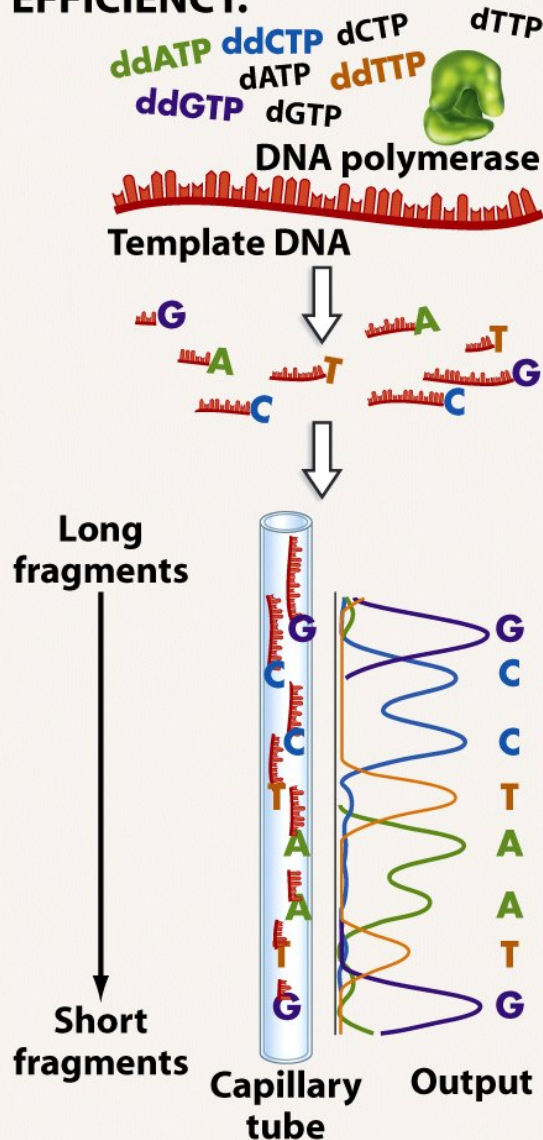
For Q99, probability of error $\sim 10^{-10}$

Phred: The base-calling program

Algorithm based on ideas about what might go wrong in a sequencing reaction and in electrophoresis

Tested the algorithm on a huge dataset of “gold standard” sequences (finished human and *C. elegans* sequences generated by highly-redundant sequencing)

FLUORESCENT MARKERS IMPROVE SEQUENCING EFFICIENCY.



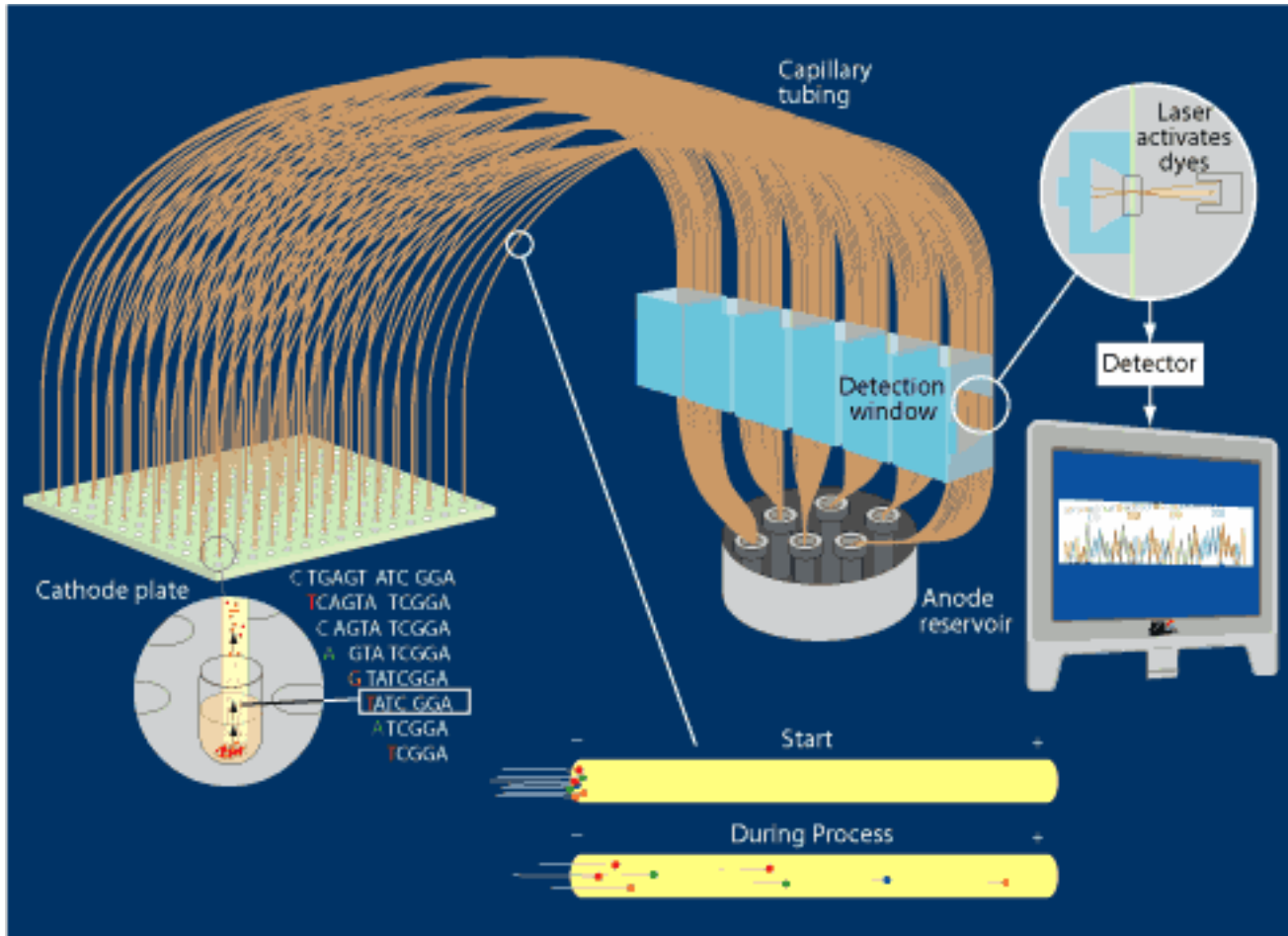
1. Do one sequencing reaction instead of four. Reaction mix contains ddATP, ddTTP, ddGTP, ddCTP with distinct fluorescent markers. (With radioactive labels, four reactions are needed—one labeled ddNTP at a time.)

2. Fragments that result have distinctive labels.

3. Separate fragments via electrophoresis in mass-produced, gel-filled capillary tubes. Automated sequencing machine reads output.

Figure 20-1 Biological Science, 2/e

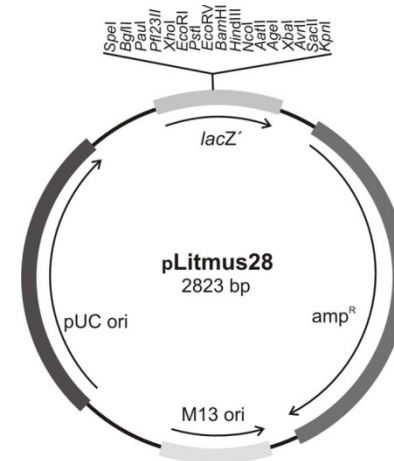
Use capillary sequencer to improve throughput



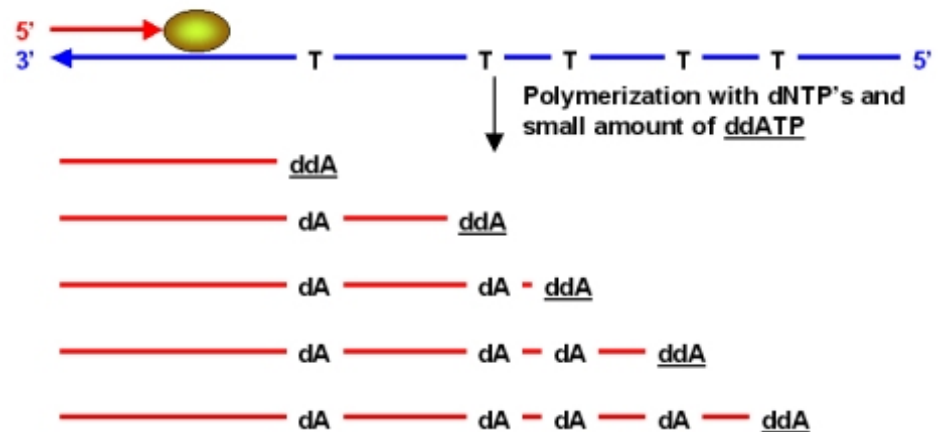
Capillary sequencers drastically sped up sequencing by both automation **and** multiplexing.

Sequencing Reaction

- The template DNA is usually single stranded DNA, which can be produced from plasmid cloning vectors that contain the origin of replication from a single stranded bacteriophage such as M13 or fd. The primer is complementary to the region in the vector adjacent to the multiple cloning site.
- Sequencing is done by having 4 separate reactions, one for each DNA base.
- All 4 reactions contain the 4 normal dNTPs, but each reaction also contains one of the ddNTPs.
- In each reaction, DNA polymerase starts creating the second strand beginning at the primer.
- When DNA polymerase reaches a base for which some ddNTP is present, the chain will either:
 - terminate if a ddNTP is added, or:
 - continue if the corresponding dNTP is added.
 - which one happens is random, based on ratio of dNTP to ddNTP in the tube.
- However, all the second strands in, say, the A tube will end at some A base: you get a collection of DNAs that end at each of the A's in the region being sequenced.



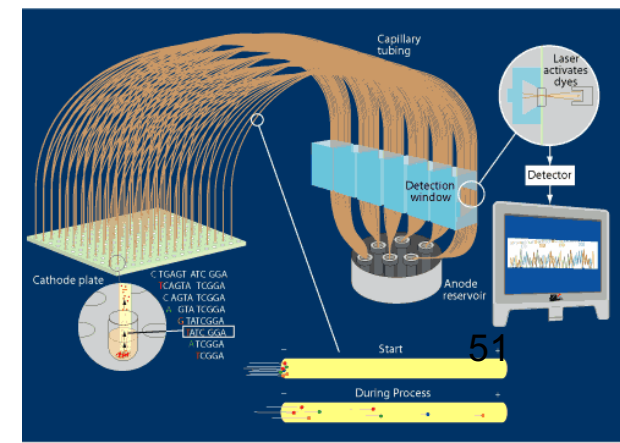
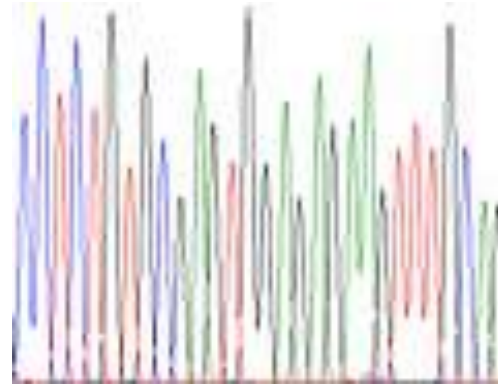
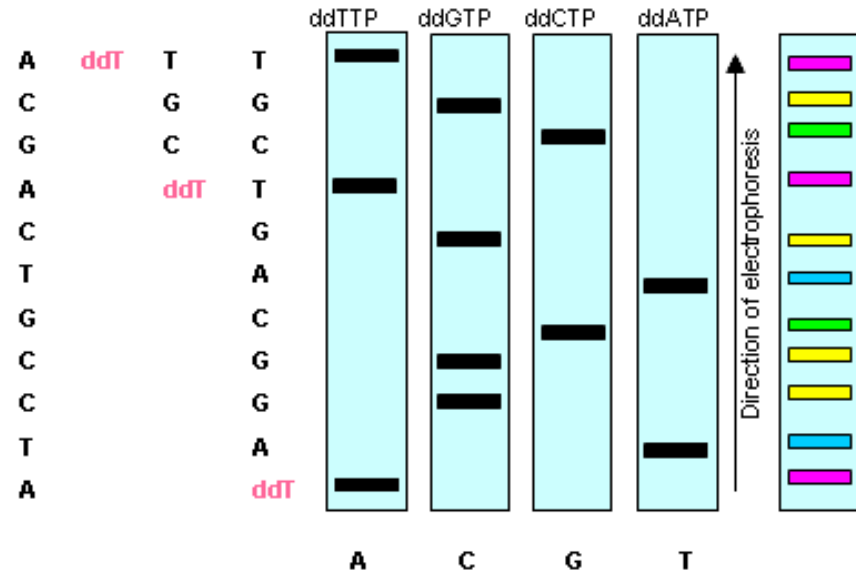
Location of Thymine bases in DNA template



Collection of fragments of newly synthesized DNA:
They all end in ddA at locations of complementary T bases in the template

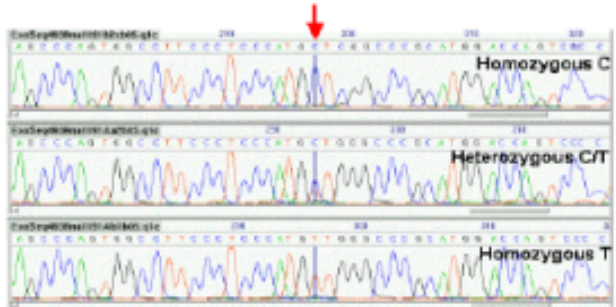
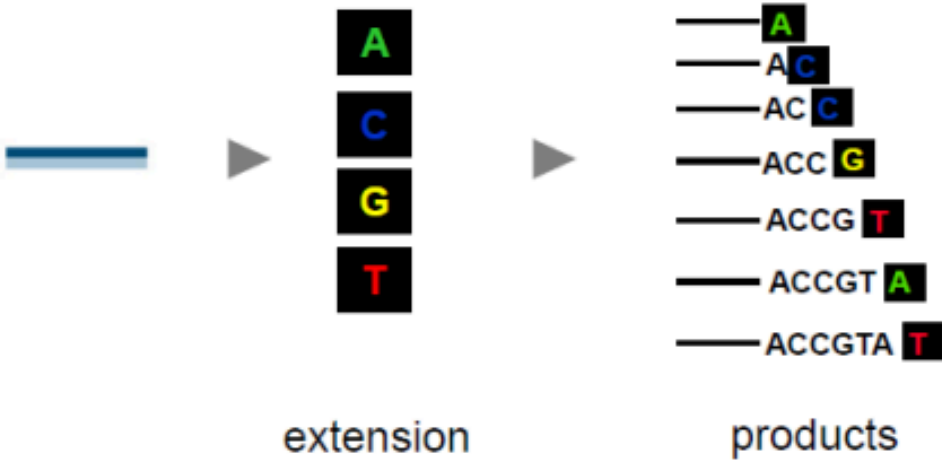
Electrophoresis

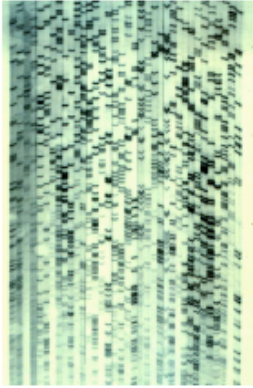
- The newly synthesized DNA from the 4 reactions is then run (in separate lanes) on an electrophoresis gel.
- The DNA bands fall into a ladder-like sequence, spaced one base apart. The actual sequence can be read from the bottom of the gel up.
- Automated sequencers use 4 different fluorescent dyes as tags attached to the dideoxy nucleotides and run all 4 reactions in the same lane of the gel.
 - Today's sequencers use capillary electrophoresis instead of slab gels.
 - Radioactive nucleotides (^{32}P) are used for non-automated sequencing.
- Sequencing reactions usually produce about 500-1000 bp of good sequence.



1987 first automated sequencer ABI370

DNA polymerase, dNTP's
and dye-labeled dideoxynucleotides (terminators)





**Radioactive
polyacrylamide
slab gel**

Low throughput,
labor intensive



**AB slab gel sequencers
(370, 373, 377)**

Fluorescent sequencing
1990-1999
6 runs/day
96 reads/run
500 bp/read
288,000 bp/day



**AB capillary sequencers
(3700, 3730)**

1998-now
24 runs/day
96 reads/run
550 – 1,000 bp/read
1-2 million bp/day

~1,000-fold increase in throughput since 1985 accomplished by incremental improvements of the same underlying technology

2nd Generation Sequencing Technologies have ~500 - 30,000x more throughput than 3730:
454 Pyrosequencing, Illumina, SOLiD (+ PacBio, Ion Torrent, Complete Genomics⁵³)

Celera, the company who competed with the government-funded human genome sequencing consortium

- 300 ABI DNA sequencing platforms
- 50 production staff
- 20,000 square feet of wet lab space
- 1 million dollars / year for electrical service
- 10 million dollars in reagents

Total cost of human genome: 2.7 Billion dollars

High-Throughput DNA Sequencing Technologies

454 / Roche

450 bp 1.5 Gbp / day



Illumina

150 bp 35 Gbp / day



Helicos

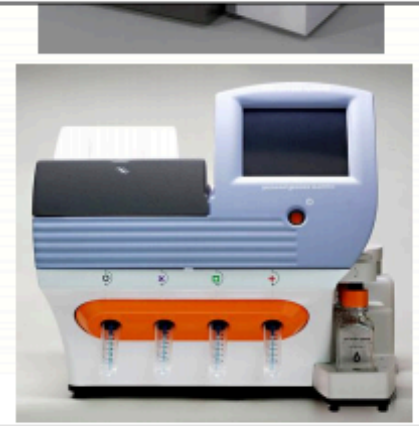
55 bp 4.5 Gb / day



ABI: "Traditional" Sanger Sequencing

650 bp

2 Mbp / day



SOLID ABI

75 bp 22 Gbp / day

PacBio

1000 bp 70 Gbp / day

Ion PGM

100 bp 120 Gbp / day

Error rate

Technology	Read length (bp)	Error rate
ABI/Solid	75	Low (~2%)
Illumina/Solexa	100–150	Low (<2%)
IonTorrent	~200	Medium (~4%)*
Roche/454	400–600	Medium (~4%)*
Sanger	Up to ~2,000 bp	Low (~2%)
Pacific Biosciences	Up to ~15,000 [‡]	High (~18%)

Different NGS platforms

- 454 Sequencing / Roche
 - GS Junior System
 - GS FLX+ System
- Illumina (Solexa)
 - HiSeq System
 - Genome analyzer Iix
 - MySeq
- Applied Biosystems - Life Technologies
 - SOLiD 5500 System
 - SOLiD 5500xl System
- Ion Torrent - Life Technologies
 - Personal Genome Machine (PGM)
 - Proton
- Helicos
 - Helicos Genetic Analysis System
- Pacific Biosciences
 - PacBio RS
- Oxford Nanopore Technologies
 - GridION System
 - MinION

Next Generation Sequencing
Amplified Single Molecule Sequencing

Third Generation Sequencing,
Next Next Generation Sequencing,
Single Molecule Sequencing

2013 NGS field guide

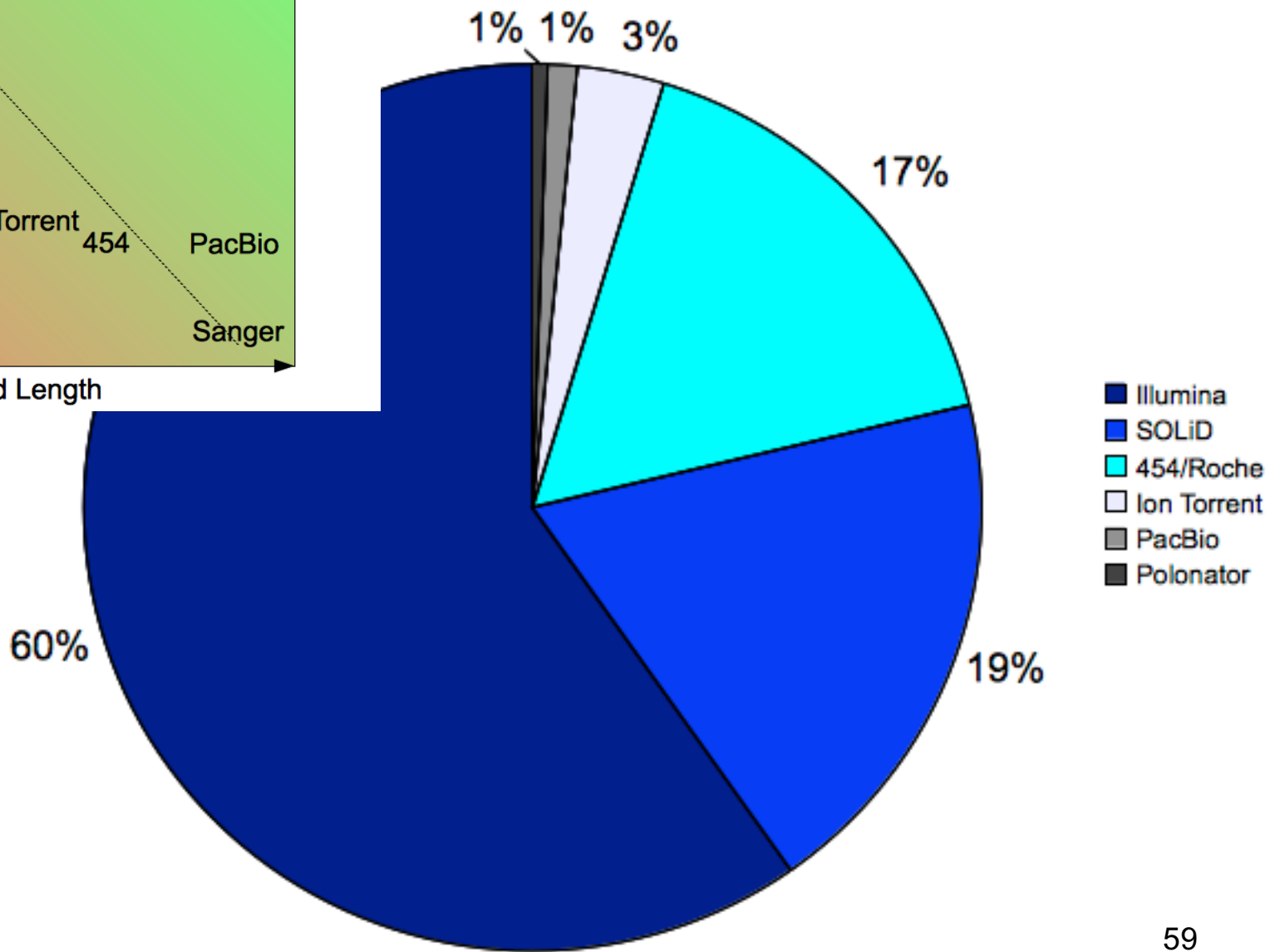
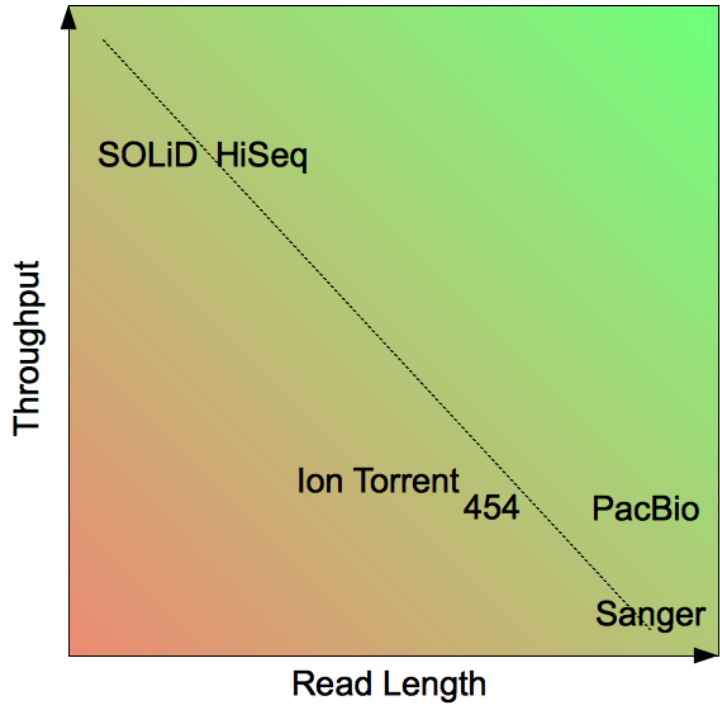
- <http://www.molecularecologist.com/next-gen-table-2b-2013/>

Instrument	Reagent Cost/run ^a	Reagent Cost/MB	Minimum Unit Cost (% run) ^b
3730xl (capillary)	\$144	\$2,308	\$6 (1%)
454 FLX Titanium	\$6,200	\$12	\$2,000 (12%)
PacBio RS	≥ \$300 ^c	\$2-17	\$500 (100%)
Ion Torrent – ‘316’ chip	\$739	\$1.20	~\$1,000 (100%)
Illumina MiSeq	\$1,040	\$0.70	~\$1,400 (100%)
Ion Torrent – ‘318’ chip	\$939	\$0.60	~\$1,200 (100%)
Ion Torrent – Proton I	\$1,050	\$0.09 ^d	? (100%)
SOLiD – 5500xl	\$10,503 ^e	< \$0.07	\$2,000 (12%)
Illumina HiSeq 2500 – rapid *	\$6,145 ^f	\$0.05	? (50%)

^aIncludes all stages of sample prep. for a single sample (i.e., library prep through sequencing; capillary = sequencing only)

^bTypical full cost (i.e., including labor, service contract, etc.) of the smallest generally available unit of purchase at an academic core lab provider for the longest available read (and percentage of reads relative to a full run, rounded to the nearest whole percentage)

<http://www.allseq.com/knowledgebank/sequencing-platforms/>



The key ideas of NGS to increase the throughput:

You don't run gel

You sequence a strand of DNA while keeping it held in place

You stick the DNA in a tiny little area on a chip/bead, so that the chip/bead could hold millions or billions of such dots (high density and throughput)

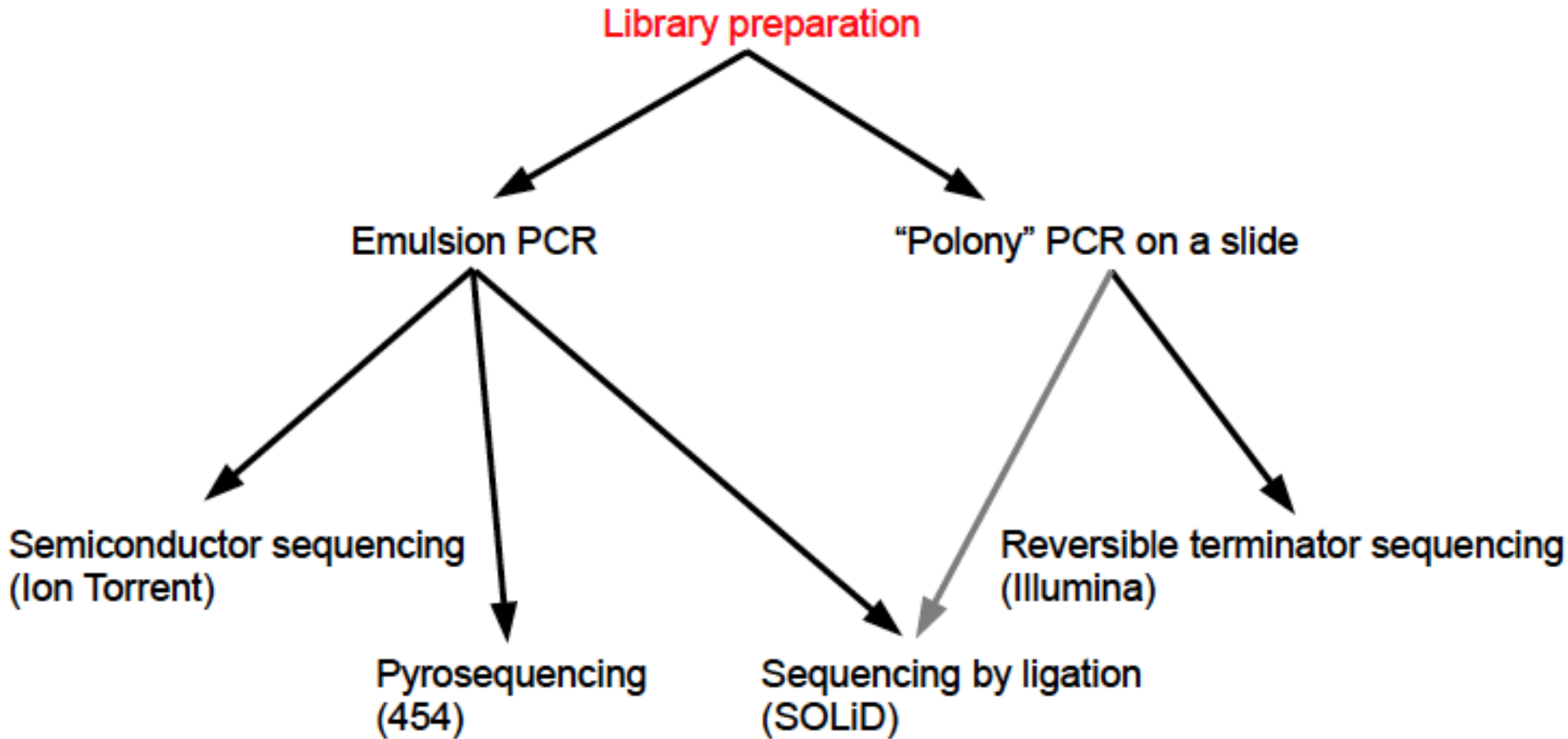
Each dot will hold many (thousands) molecules of the same DNA strand to be sequenced as a cluster (high signal/noise ratio)

You even skip the clone library preparation and use PCR to amplify DNA

The rest is similar to Sanger: adding a nt will emit fluorescent light and an optical device is used to capture the light signal.

<http://www.genetic-inference.co.uk/blog/2009/08/basics-sequencing-dna-part-2/>

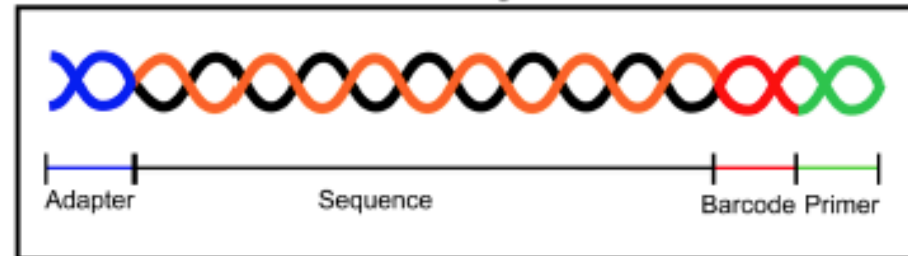
Next Generation Sequencing : Amplified Single Molecule Sequencing



Next Generation Sequencing : Amplified Single Molecule Sequencing

Library preparation

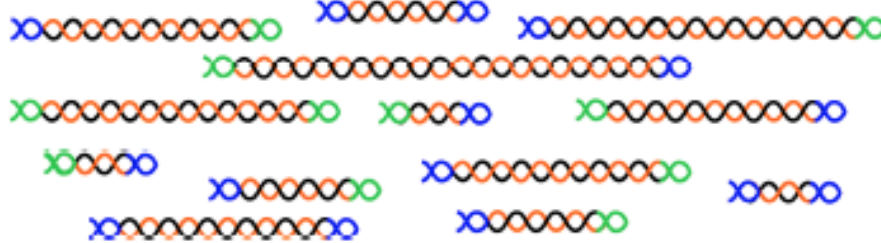
Good fragments :



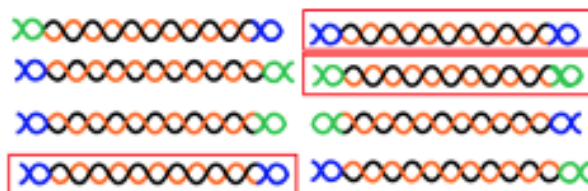
Fragmentation of DNA
(sonication or enzymatic)



Ligation of adapter and primer (or barcode)



Size-select the fragments



The principle

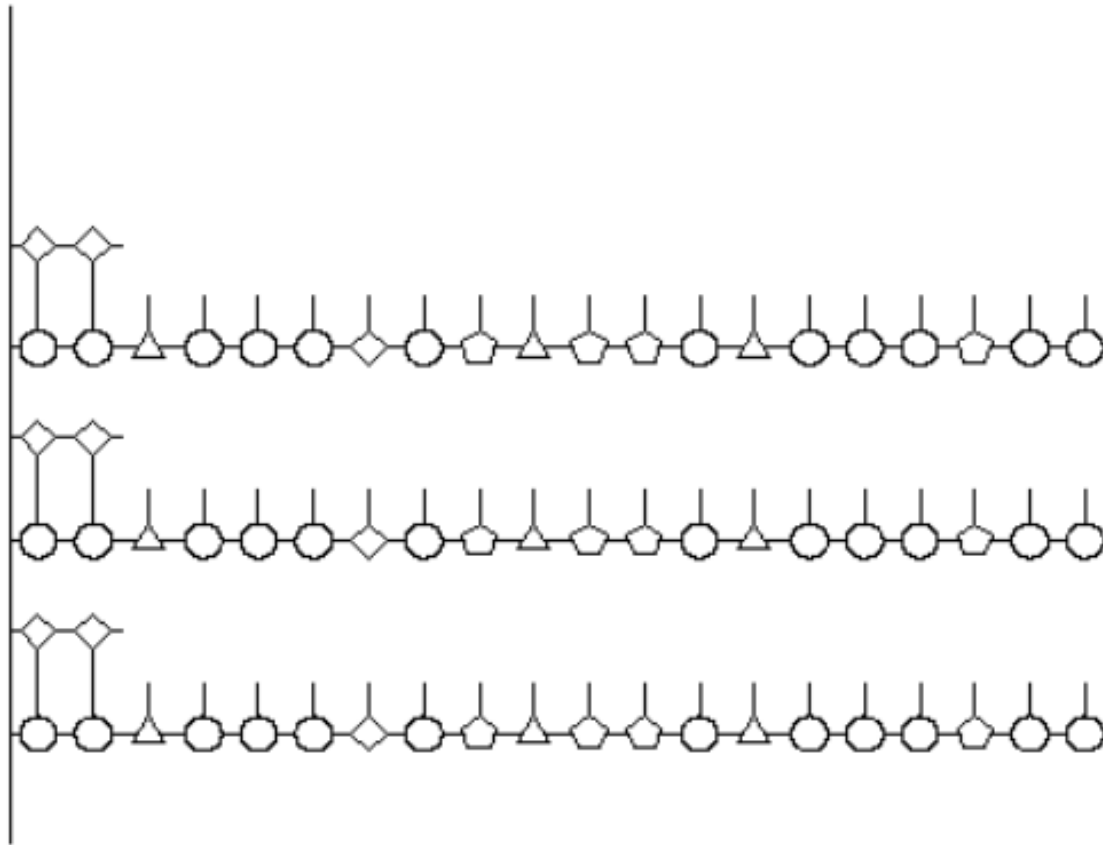
Detecting base incorporation during DNA strand synthesis at a massively parallel scale

Base detection and strand synthesis are key differences between competing technologies

How it works:

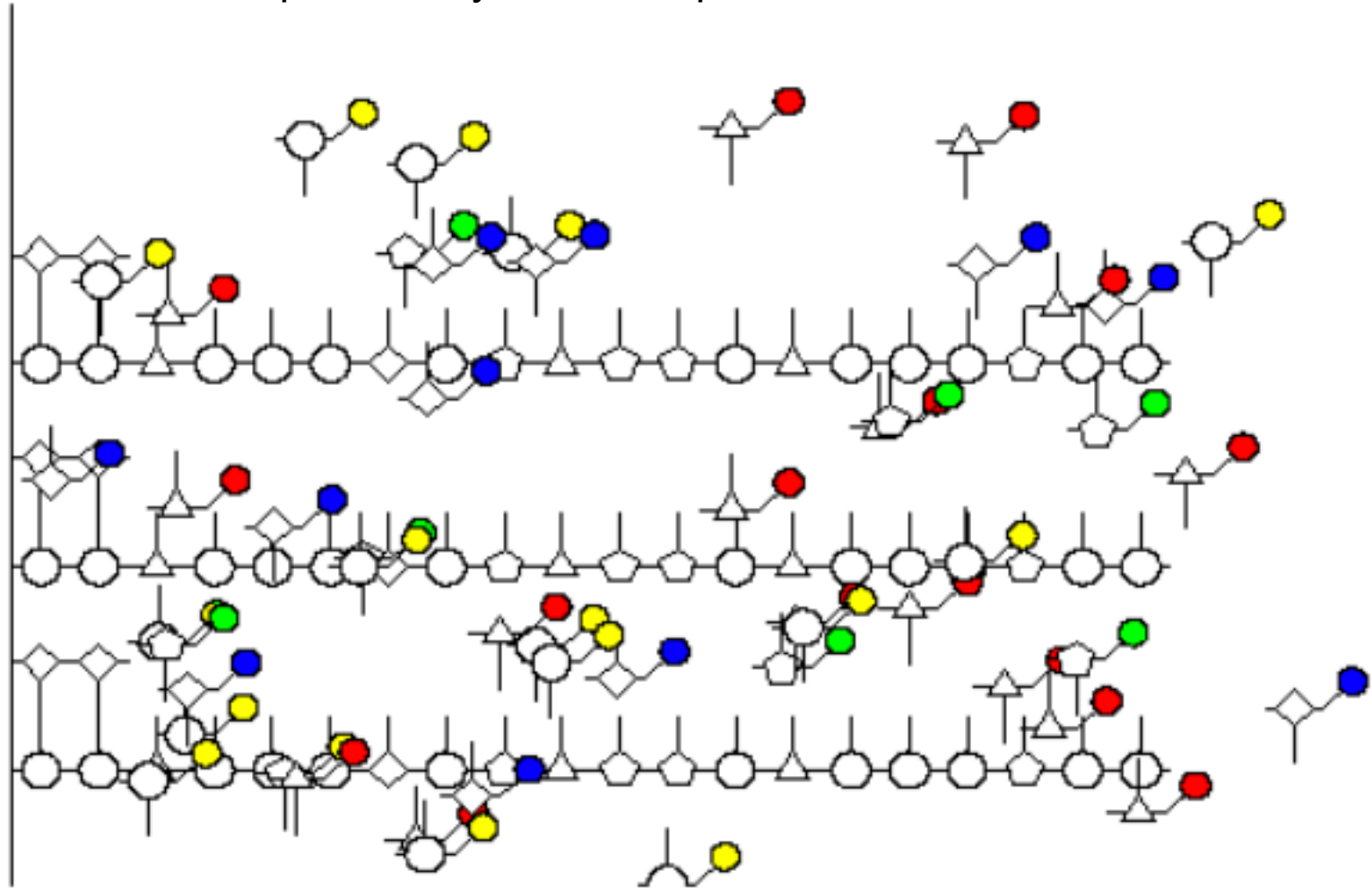
1. Physically separate thousands to millions single-stranded DNA fragments
2. Fix the fragments' location on a substrate & amplify
3. Detect the incorporation of each base in each location through a signal
4. Repeat step #3 for many cycles

Illumina: we multiply up the DNA molecules as a cluster of them. All molecules are single stranded and stick on a few bases of “adapter” sequences, which is attached on the surface of the chip, holding the DNA in place while being sequenced.

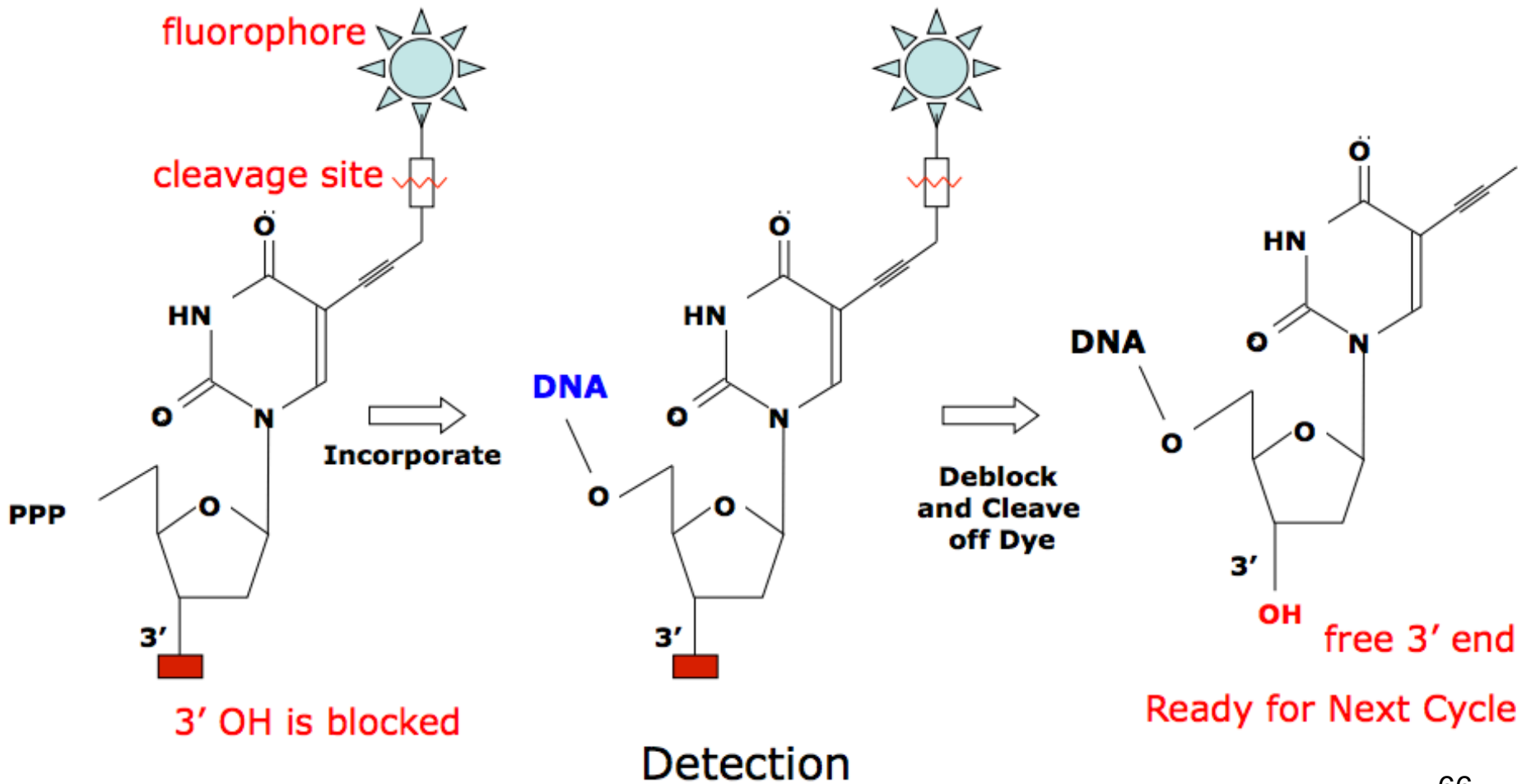


High density of integration on the chip or bead so that huge amounts of DNA could be sequenced simultaneously

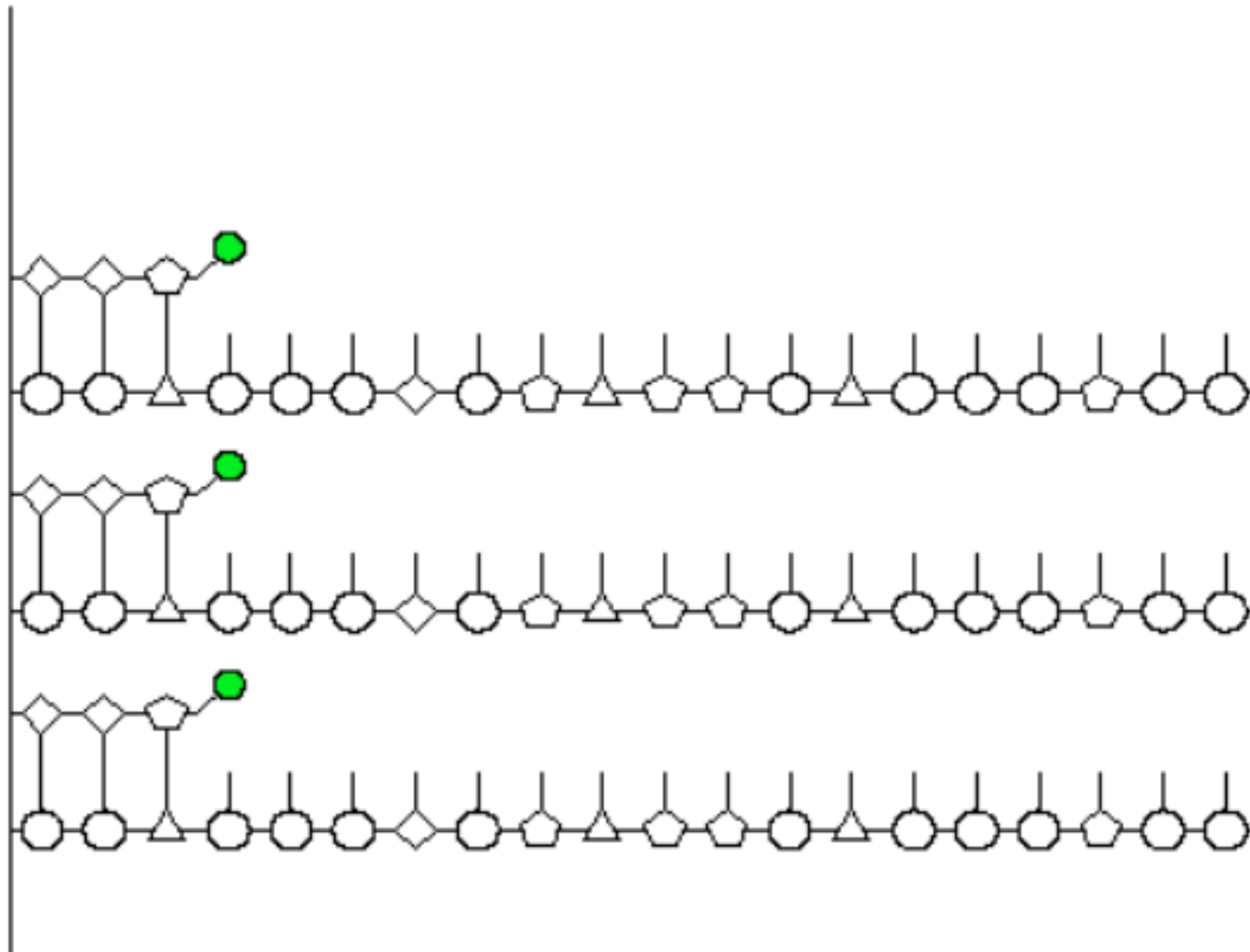
Flood the chip or bead with fluorescent nucleotides (**reversible terminator bases**), add a polymerase enzyme, which incorporates the RT-base into the new strand that is complementary to the template strand



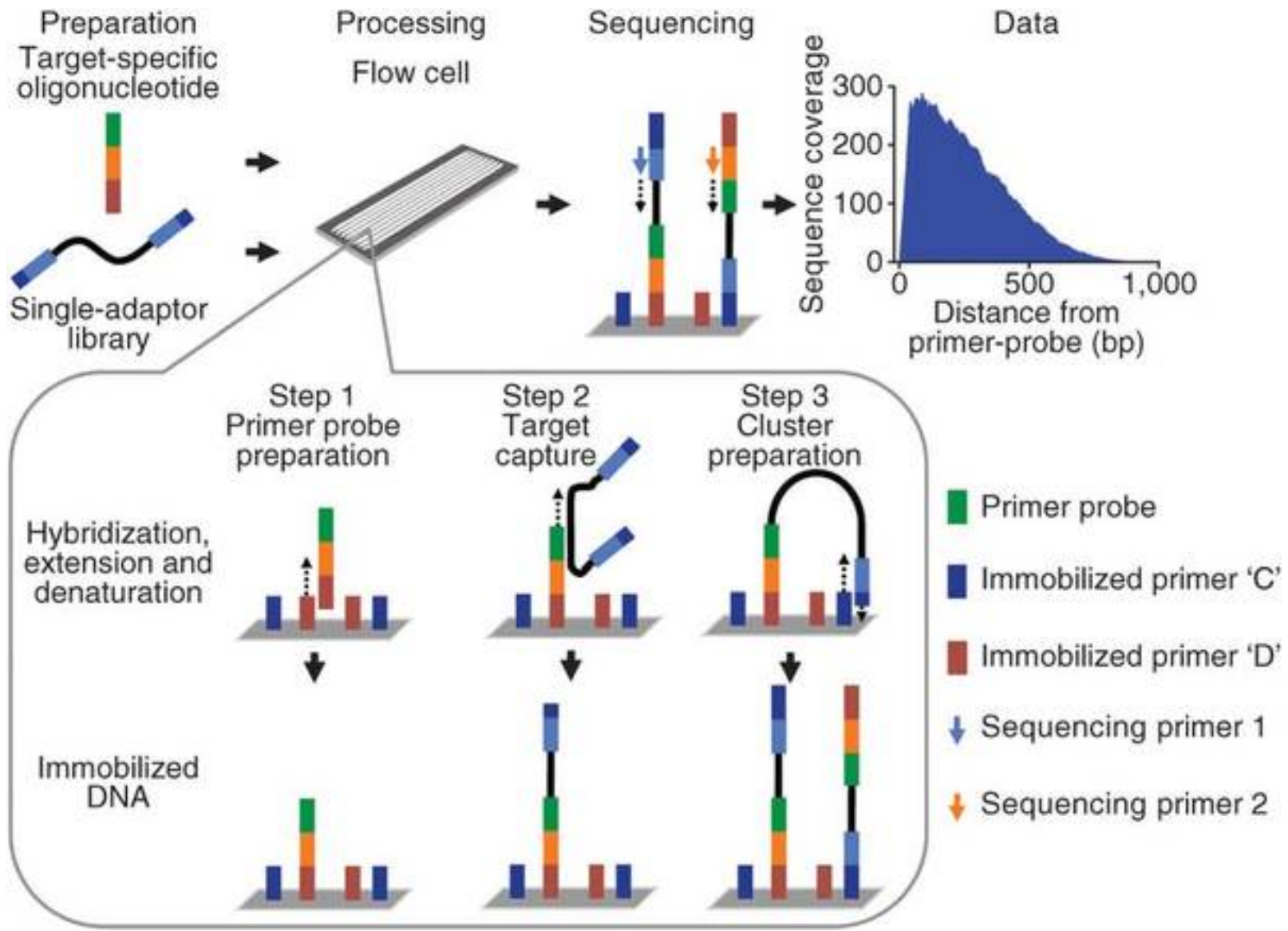
Illumina Sequencing: Reversible Terminators



We then wash away all the RT-bases, leaving just those that were incorporated into the new strand; we can read off what base this is by looking at the colour of the dye:

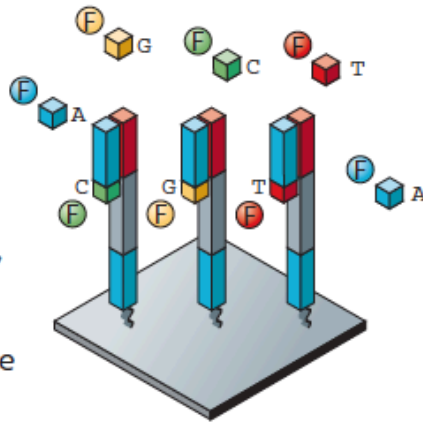


- Finally, we send in the cleavage enzyme, which cuts off the terminator region and the dye, leaving a normal base pair. We can then start again to sequence the next base pair.
- In a single Illumina machine we have hundreds of millions of these clusters; cameras look at all of these dots and record how they change color over time, allowing you to determine the sequence of bases of millions of bits of DNA at once.
- Sequencing method is actually pretty inefficient, however, the machine is capable of sequencing millions of fragments of DNA at once.

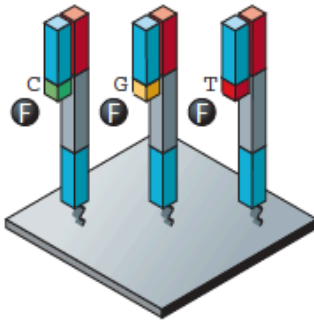


a Illumina/Solexa — Reversible terminators

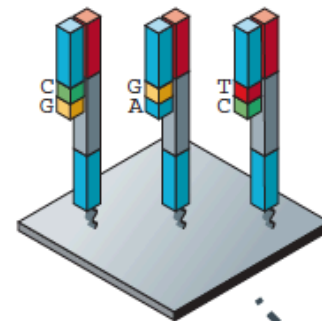
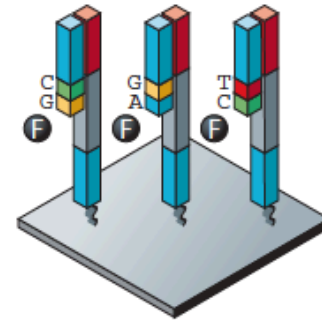
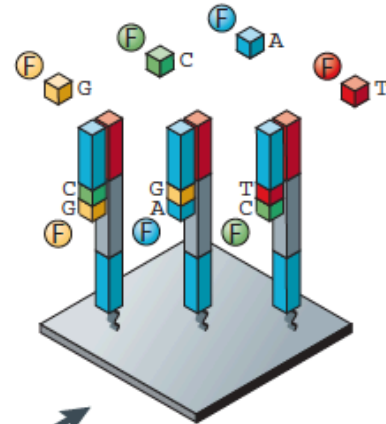
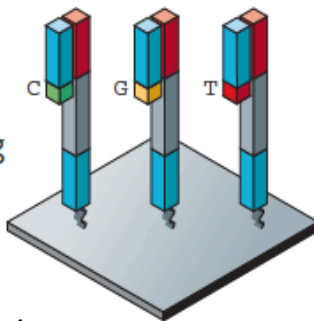
Incorporate all four nucleotides, each label with a different dye



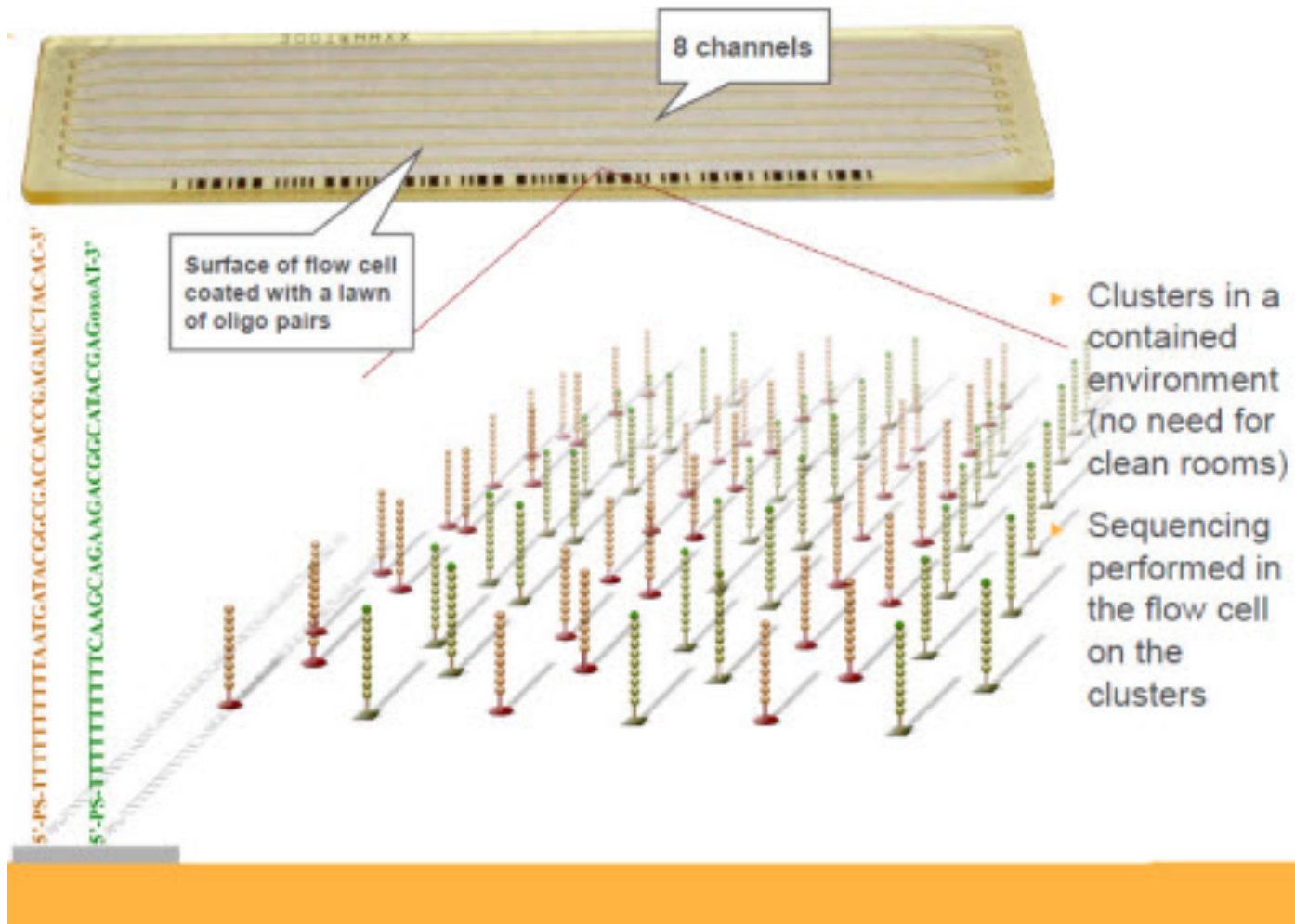
Wash, four-colour imaging

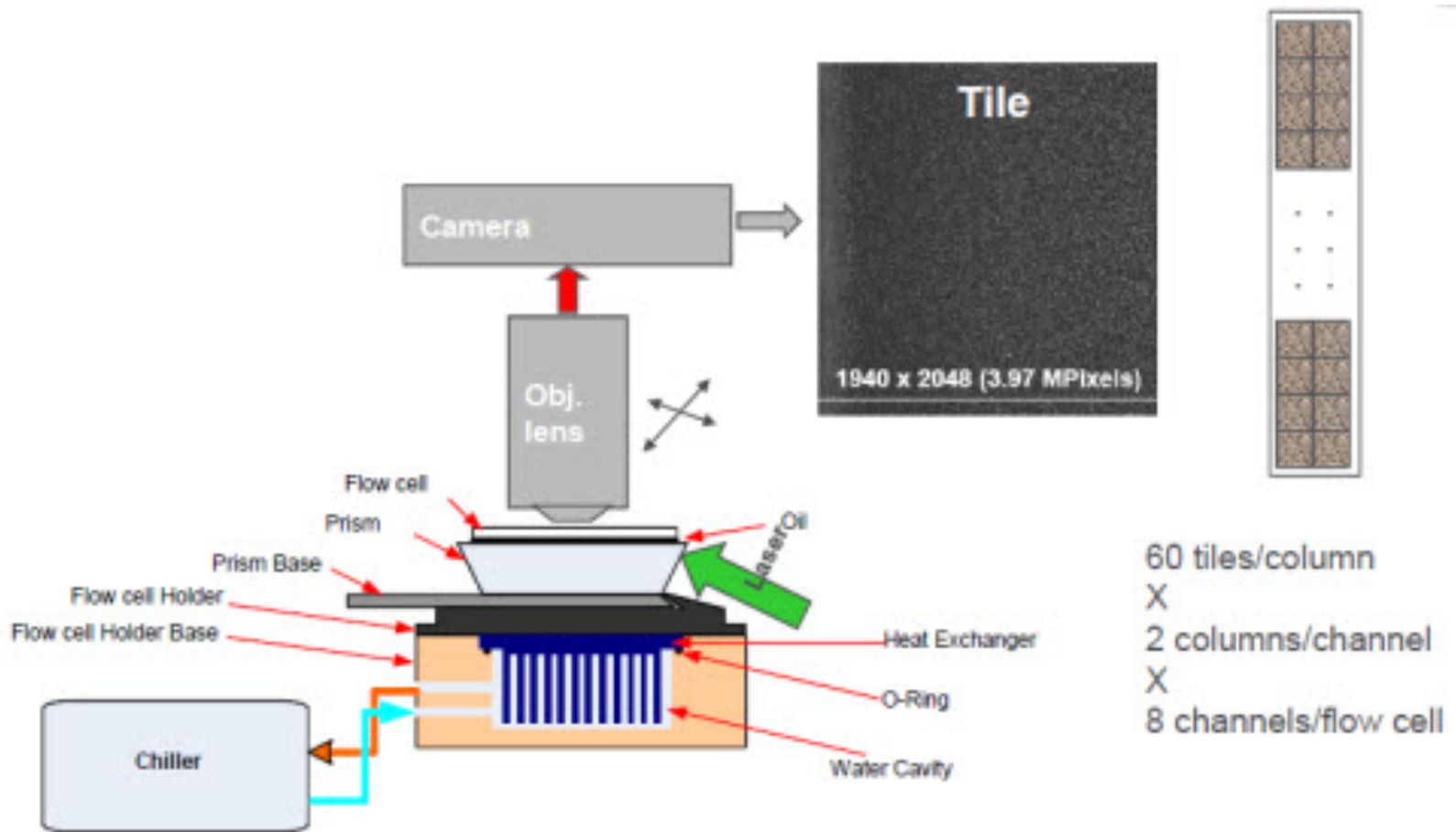


Cleave dye and terminating groups, wash



Repeat cycles

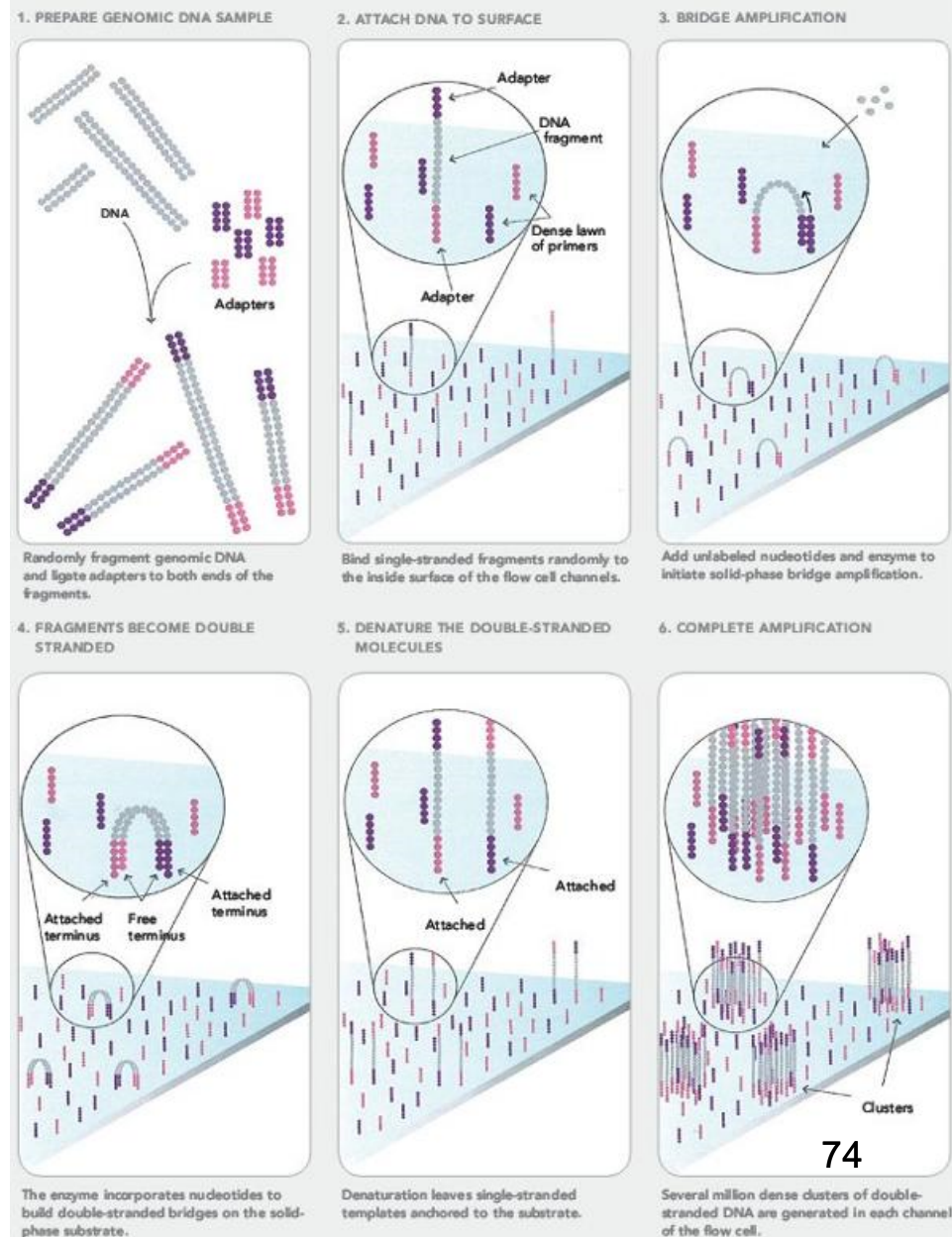




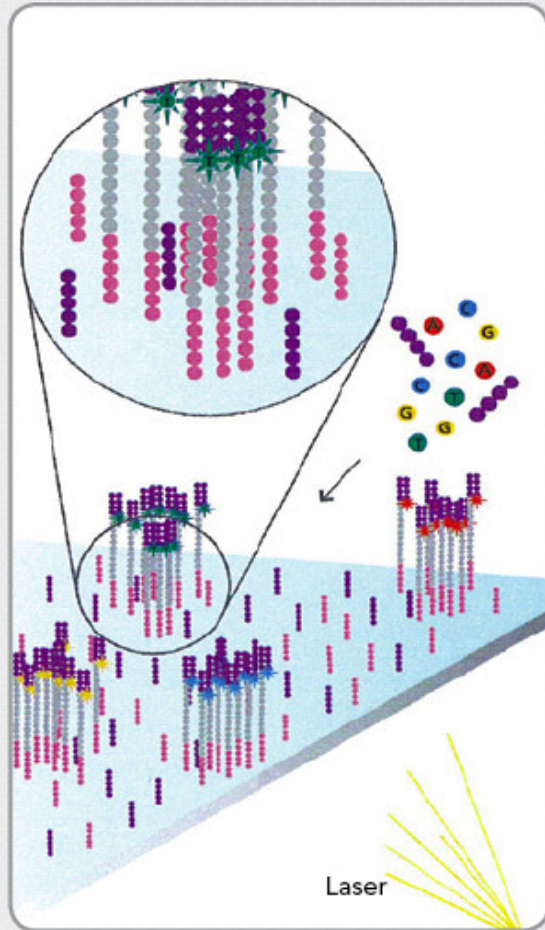
http://www.ohsu.edu/xd/research/research-cores/mpssr/project-design/mpssr_sequencing_technology.cfm

Illumina Massively Parallel System

- The idea is to put 2 different adapters on each end of the DNA, then bind it to a slide coated with the complementary sequences for each primer. This allows “bridge PCR”, producing a small spot of amplified DNA on the slide.
- The slide contains millions of individual DNA spots. The spots are visualized during the sequencing run, using the fluorescence of the nucleotide being added.

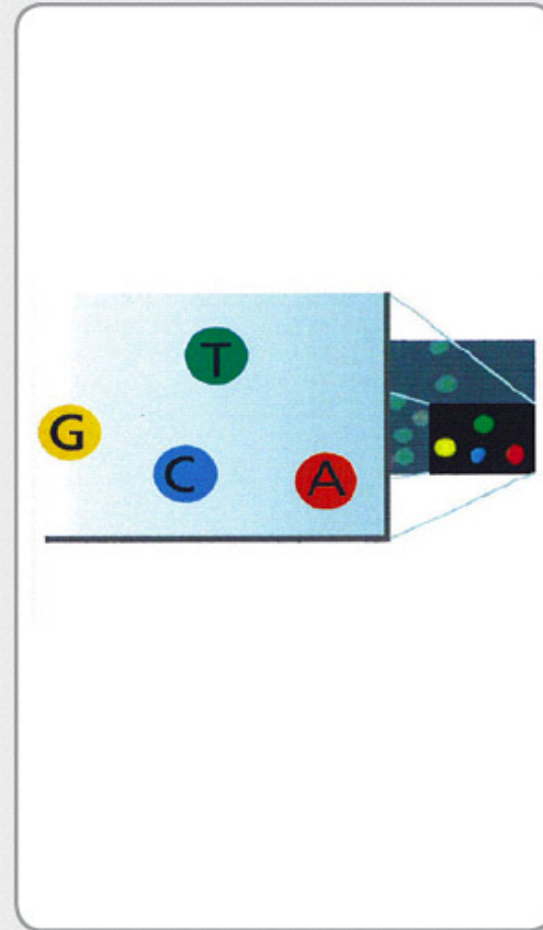


7. DETERMINE FIRST BASE



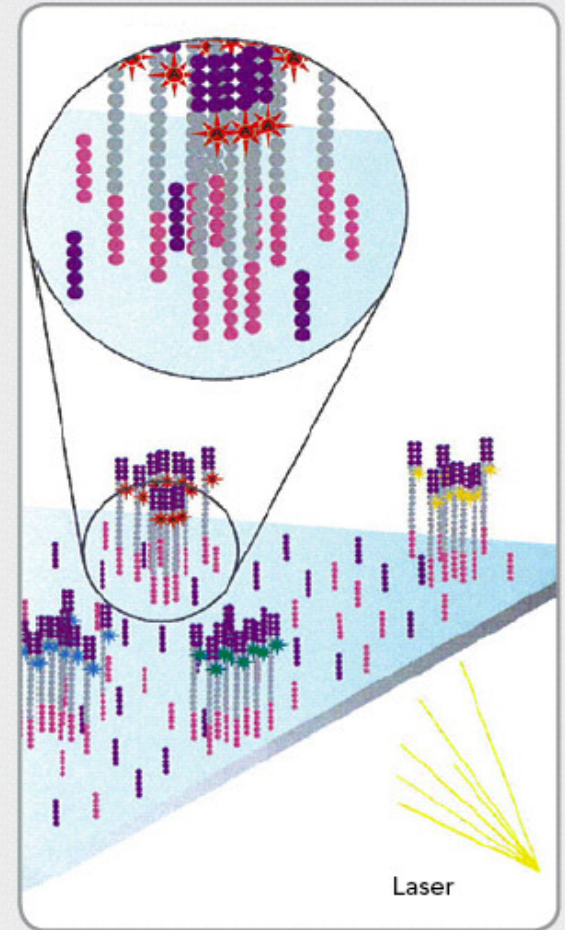
First chemistry cycle: to initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.

8. IMAGE FIRST BASE



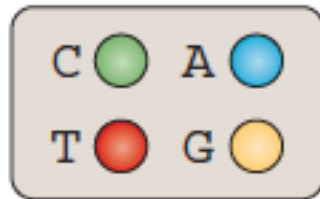
After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

9. DETERMINE SECOND BASE



Second chemistry cycle: to initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.

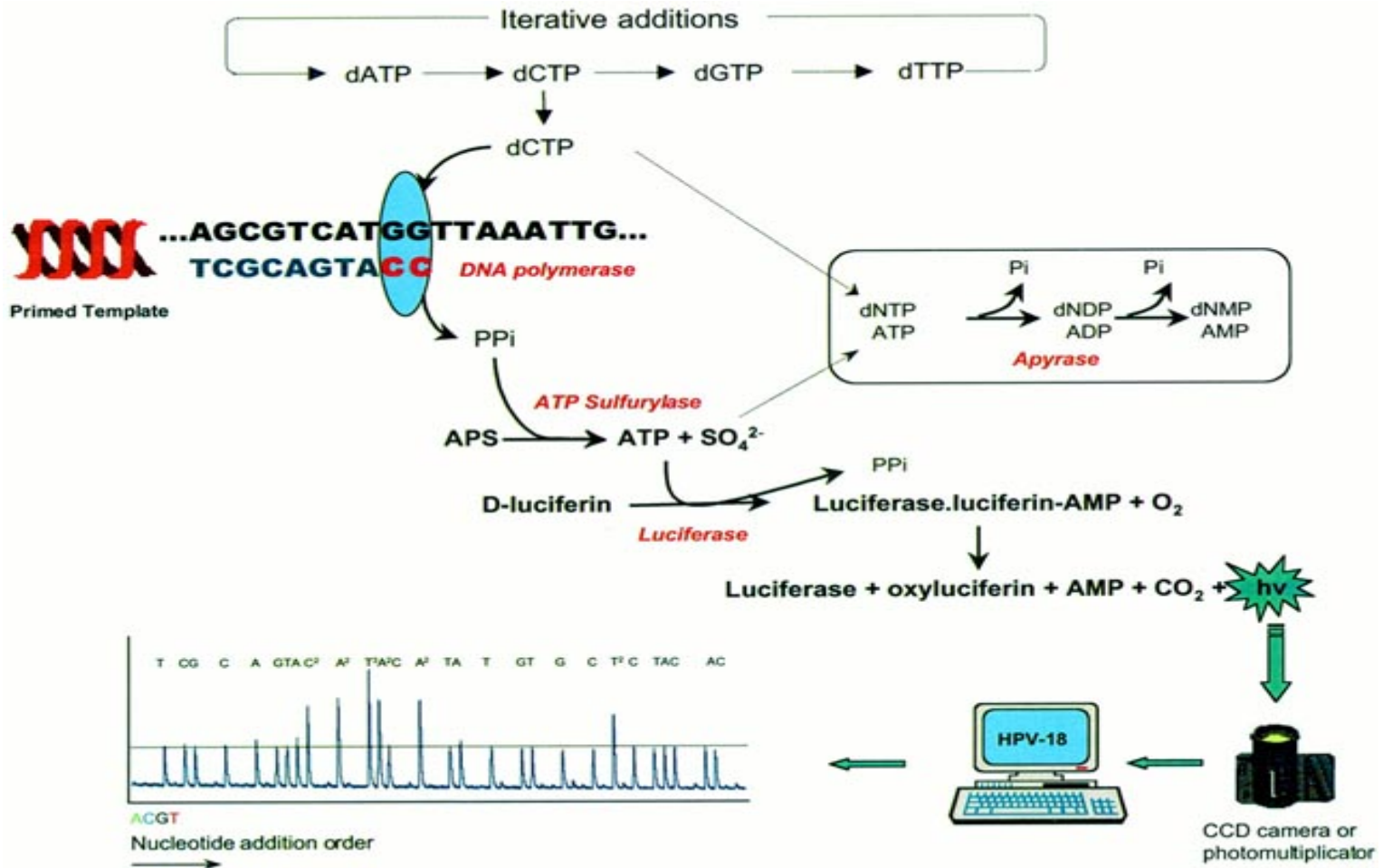
b



Top: CATCGT
Bottom: CCCCC

454 Pyrosequencing Biochemistry

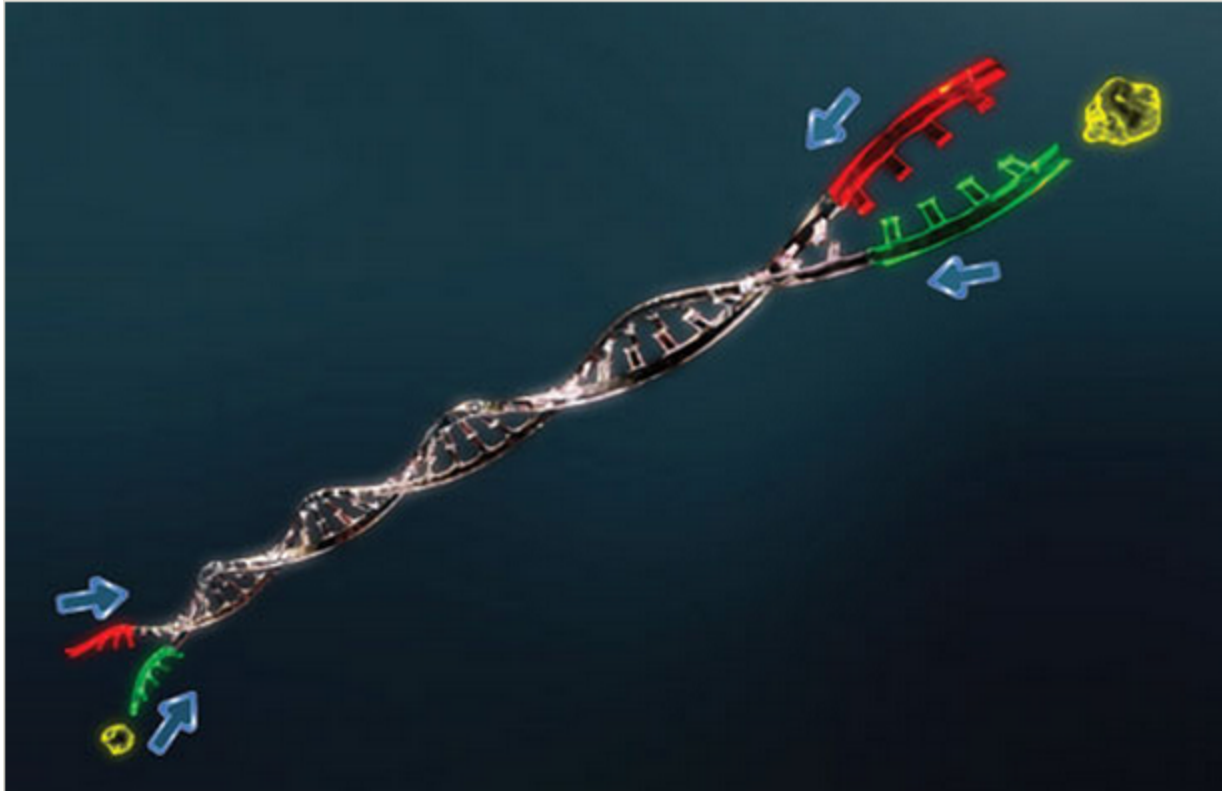
- In DNA synthesis, a dNTP is attached to the 3' end of the growing DNA strand. The two phosphates on the end are released as pyrophosphate (PPi).
- ATP sulfurylase uses PPi and adenosine 5'-phosphosulfate to make ATP.
- Luciferase uses luciferin and ATP as substrates, converting luciferin to oxyluciferin and releasing visible light.
- After the reaction has completed, apyrase is added to destroy any leftover dNTPs.





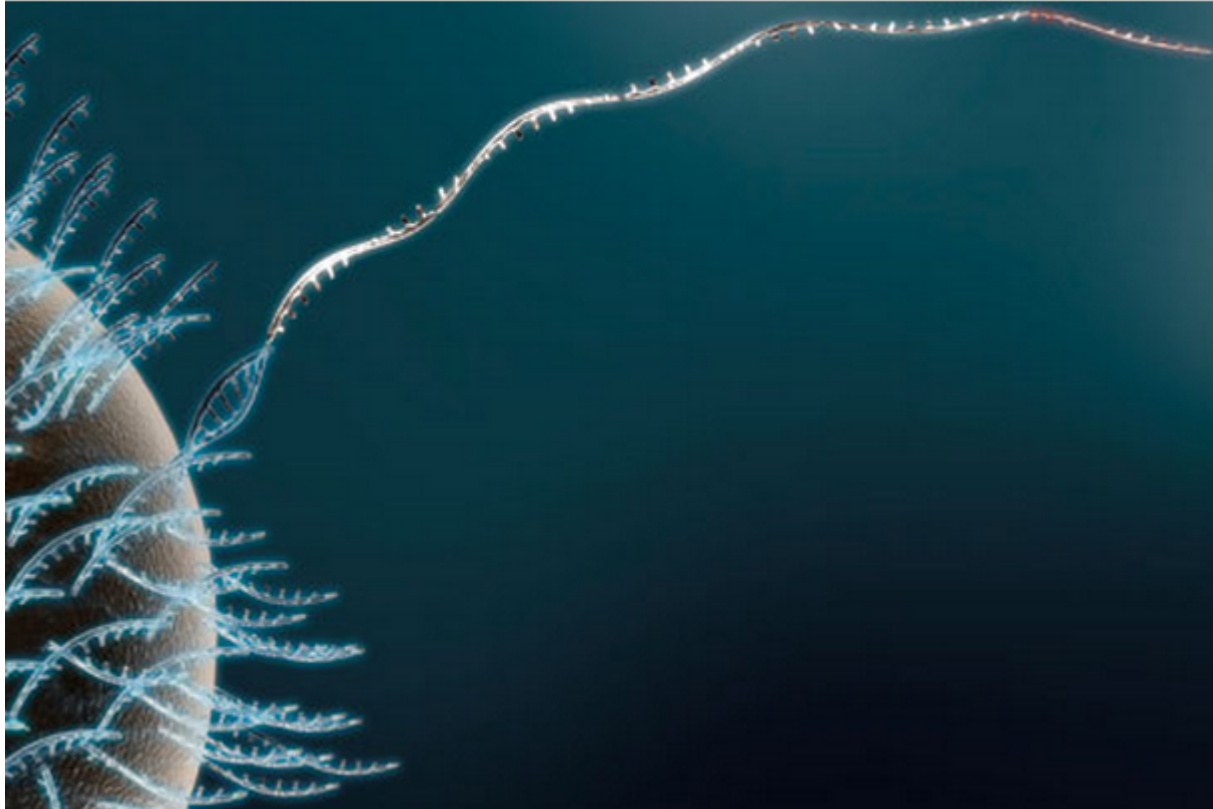
Sample Input & Fragmentation

The GS FLX and GS Junior Systems support the sequencing of samples from a wide variety of starting materials, including genomic DNA, PCR products, BACs and cDNA. For shotgun, paired end or cDNA libraries, start with as little as 500 ng of sample DNA. Nebulize longer DNA samples to create shorter library fragments



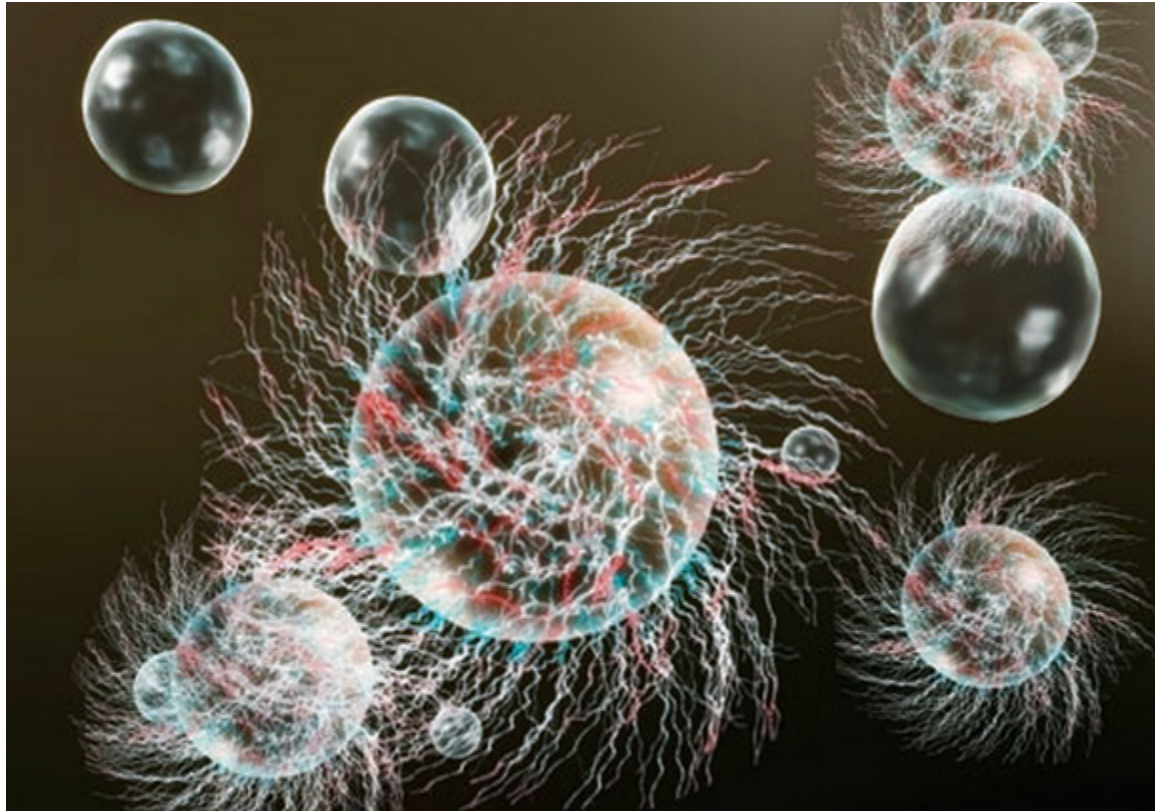
Library Preparation

Ligate Rapid Library Adaptors to the fragments for use in subsequent purification, quantitation, amplification and sequencing steps. For amplicon libraries, create PCR products by amplifying with specific fusion primer containing 454 Sequencing adaptor sequences.



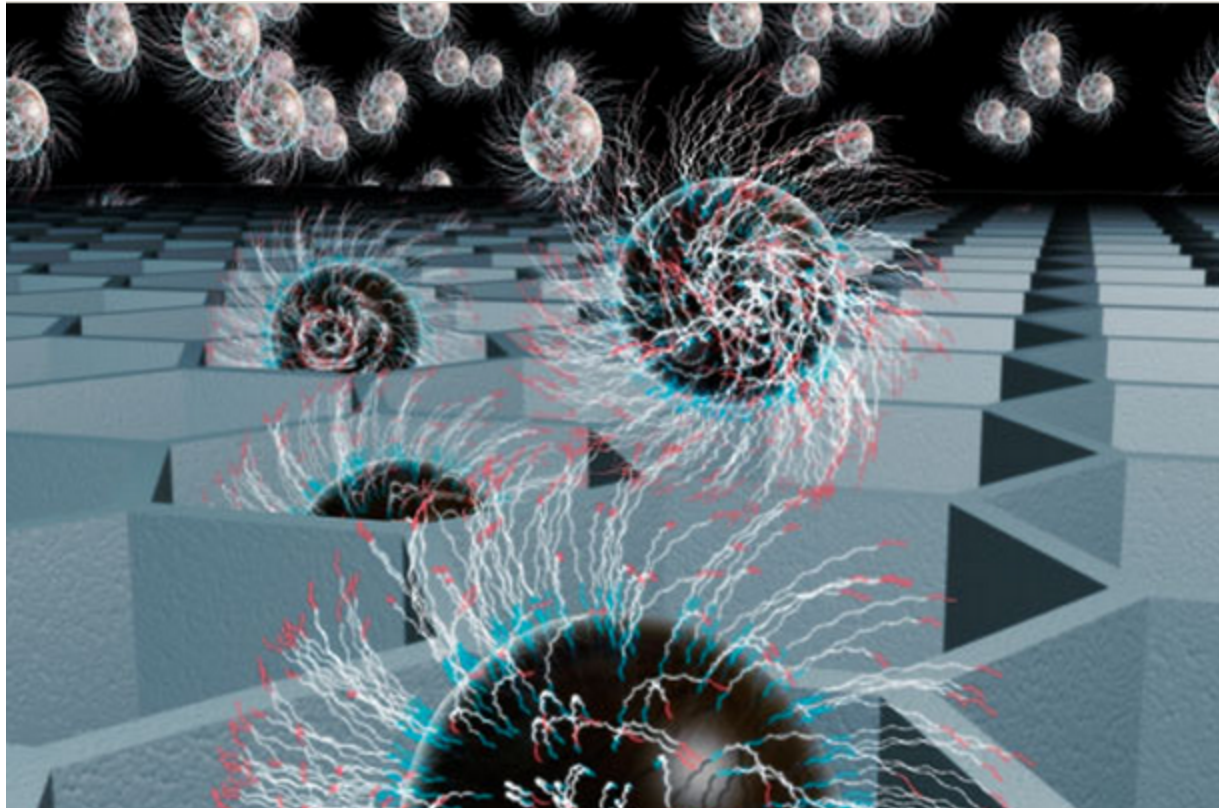
One Fragment = One Bead

Attach library to DNA Capture Beads. Each bead carries a unique single-stranded library fragment. Emulsify beads with amplification reagents in a water-in-oil mixture to trap individual beads in amplification microreactors.



emPCR: Emulsion PCR Amplification

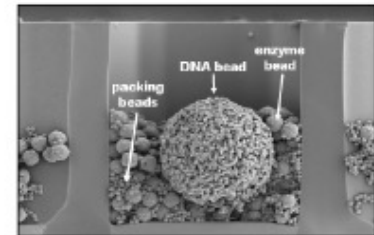
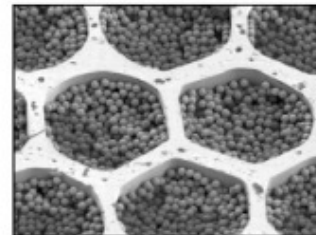
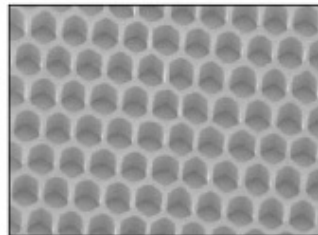
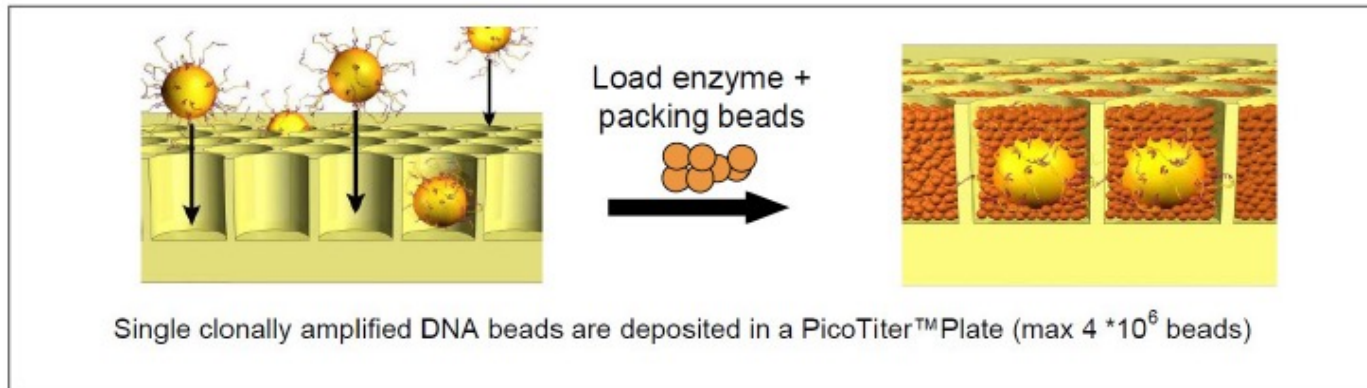
Amplify the entire emulsion in parallel to create millions of clonally copies of each library fragment on each bead. Break the emulsion while the amplified fragments remain bound to their specific beads.



Sequencing: One Bead = One Read

Load the beads onto the PicoTiterPlate device, where the surface design allows for only one bead per well. The PTP Device is then loaded in instrument for sequencing. Individual nucleotides are flowed in sequence across the wells. Each incorporation of a nucleotide complementary to the template strand results in a chemiluminescent light signal recorded by the camera.

454 Sequencing / Roche Pyrosequencing

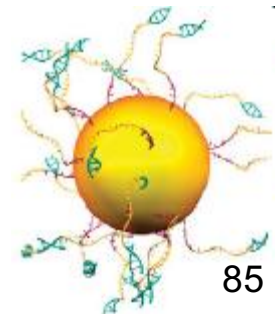
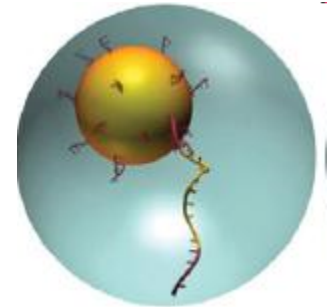
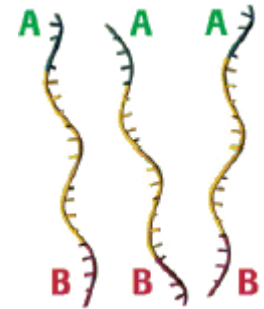
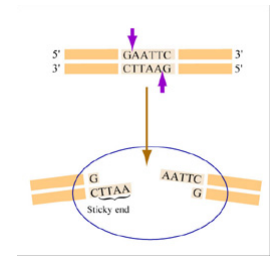


More Pyrosequencing

- The four dNTPs are added one at a time, with apyrase degradation and washing in between.
- The amount of light released is proportional to the number of bases added. Thus, if the sequence has 2 A's in a row, both get added and twice as much light is released as would have happened with only 1 A.
- The pyrosequencing machine cycles between the 4 dNTPs many times, building up the complete sequence. About 300 bp of sequence is possible (as compared to 800-1000 bp with Sanger sequencing).
- The light is detected with a charge-coupled device (CCD) camera, similar to those used in astronomy.
- YouTube animation (with music!):
<http://www.youtube.com/watch?v=kYAGFrbGI6E>

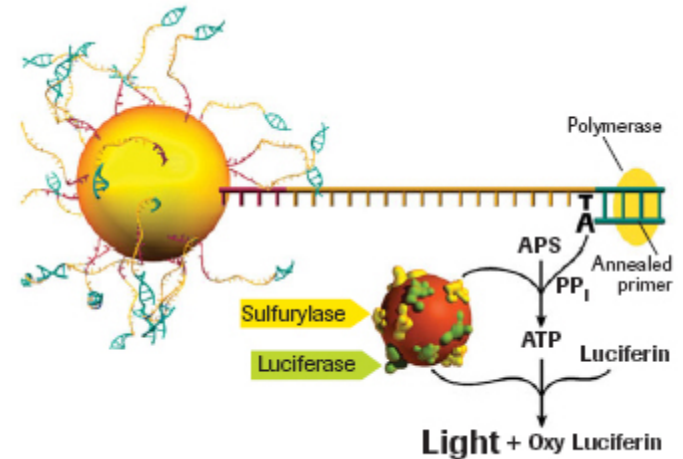
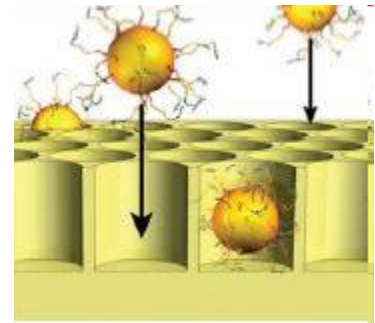
454 Technology

- To start, the DNA is sheared into 300-800 bp fragments, and the ends are “polished” by removing any unpaired bases at the ends.
- Adapters are added to each end. The DNA is made single stranded at this point.
- One adapter contains biotin, which binds to a streptavidin-coated bead. The ratio of beads to DNA molecules is controlled so that most beads get only a single DNA attached to them.
- Oil is added to the beads and an emulsion is created. PCR is then performed, with each aqueous droplet forming its own micro-reactor. Each bead ends up coated with about a million identical copies of the original DNA.



More 454 Technology

- After the emulsion PCR has been performed, the oil is removed, and the beads are put into a “picotiter” plate. Each well is just big enough to hold a single bead.
- The pyrosequencing enzymes are attached to much smaller beads, which are then added to each well.
- The plate is then repeatedly washed with the each of the four dNTPs, plus other necessary reagents, in a repeating cycle.
- The plate is coupled to a fiber optic chip. A CCD camera records the light flashes from each well.



454 Advantages and Limitations

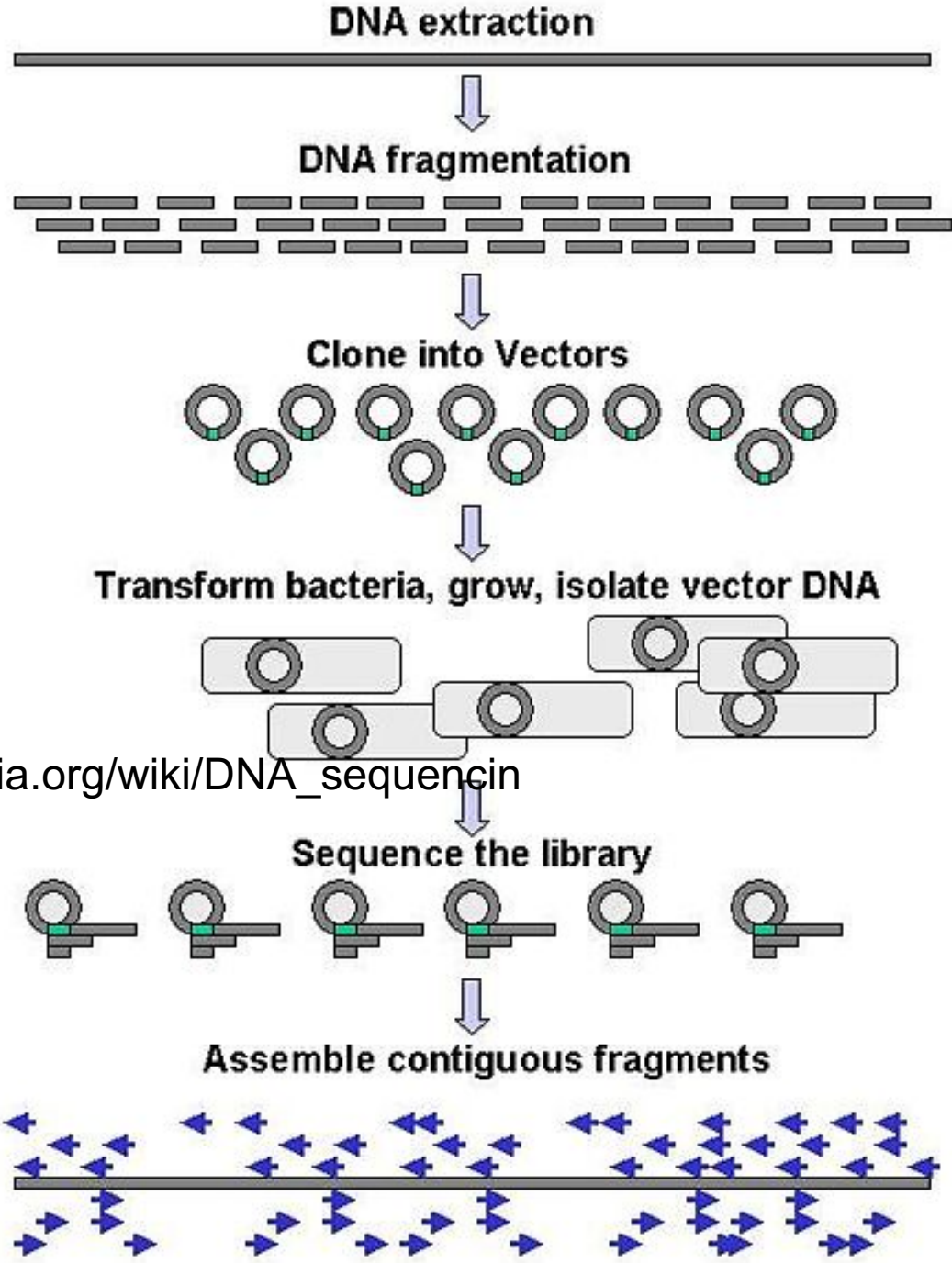
- Advantages
 - Fast, accurate
 - Great for small, simple genomes
 - Reads relatively long now (up to 1000bp)
- Disadvantages
 - Doesn't yet work well for de novo sequencing of large genomes
 - Homopolymer stretches (8+) are difficult to read

Pacific Biosciences

- Single Molecule Real Time DNA Sequencing
- Read lengths now averaging ~5kb, max 20kb
- Strobe sequencing
- Observation of DNA modifications
- Throughput per run is low, but run time is short
- Initial release in late 2010; up to 80,000 reads, of ~1.5kb each in ~15 minutes for \$100.
- Error rate is high, though hybrid approaches can significantly improve assemblies generated by short reads alone.

Finishing the Sequence

- Shotgun sequencing of random DNA fragments necessarily misses some regions altogether.
 - Also, for sequencing methods that involve cloning (Sanger), certain regions are impossible to clone: they kill the host bacteria.
- Thus it is necessary to close gaps between contigs, and to re-sequence areas with low quality scores. This process is called finishing. It can take up to 1/2 of all the effort involved in a genome sequencing project.
- Mostly hand work: identify the bad areas and sequence them by primer walking.
 - Sometimes using alternative sequencing chemistries (enzymes, dyes, terminators, dNTPs) can resolve a problematic region.
- Once a sequence is completed, it is usually analyzed by finding the genes and other features on it: annotation. We will discuss this later.
- Submission of the annotated sequence to Genbank allows everyone access to it: the final step in the scientific method.



http://en.wikipedia.org/wiki/DNA_sequencing