# DNA Sequencing

***Mark Wamalwa***
*BecA-ILRI Hub, Nairobi, Kenya*
http://hub.africabiosciences.org/
*m.wamalwa@cgiar.org*

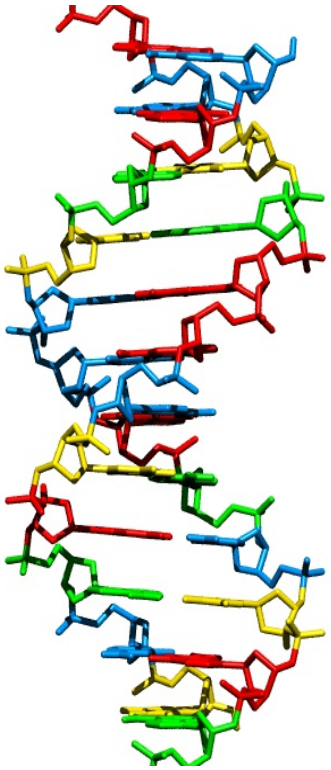
***Advanced Genomics - Bioinformatics Workshop***

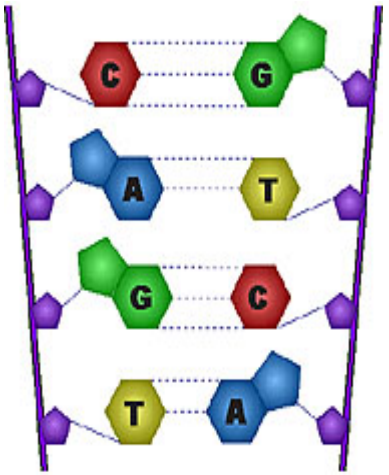
7th – 18th September 2015

Some slides borrowed from Serafim Batzoglou

# DNA sequencing

How we obtain the sequence of nucleotides of a species






...ACGTGACTGAGGACCGTG
CGACTGAGACTGACTGGGT
CTAGCTAGACTACGTTTTA
TATATATATACGTCGTCGT
ACTGATGACTAGATTACAG
ACTGATTTAGATACCTGAC
TGATTTTAAAAAAATATT...

# DNA Sequencing

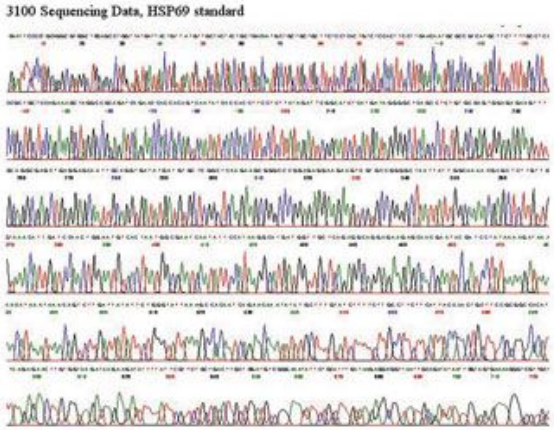
**Goal:**

Find the complete sequence of A, C, G, T's in DNA

**Challenge:**

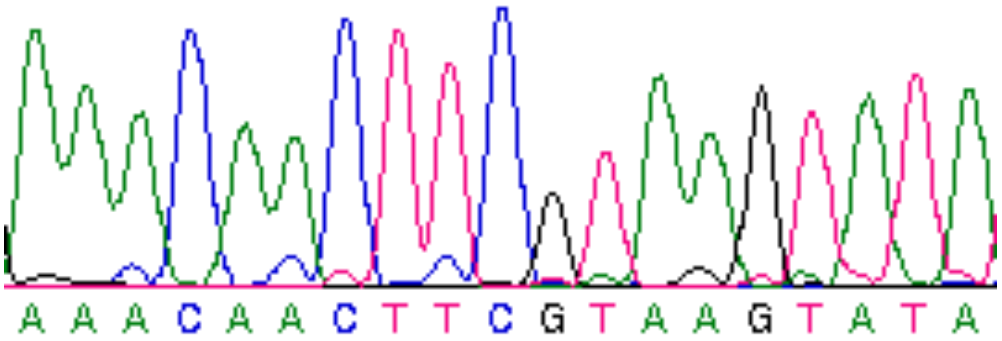
There is no machine that takes long DNA as an input, and gives the complete sequence as output

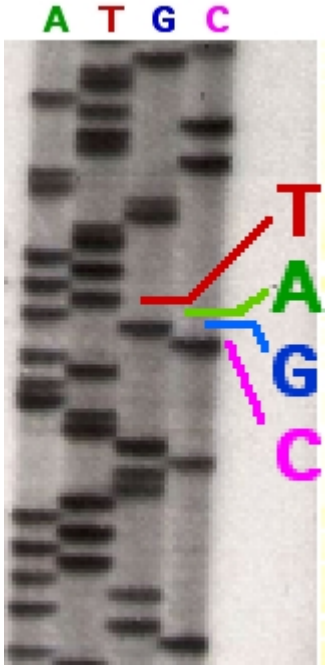
Can only sequence ~500 letters at a time




# Determining DNA Sequence

- Originally 2 methods were invented around 1976, but only one is widely used: the chain-termination method invented by Fred Sanger.
  - The other method is Maxam-Gilbert chemical degradation method, which is still used for specialized purposes, such as analyzing DNA-protein interactions.
- More recently, several cheaper and faster alternatives have been invented. It is hard to know which of these methods, or possibly another method, will ultimately become standard. We will discus two of them: 454 pyrosequencing and Illumina/Solexa sequencing



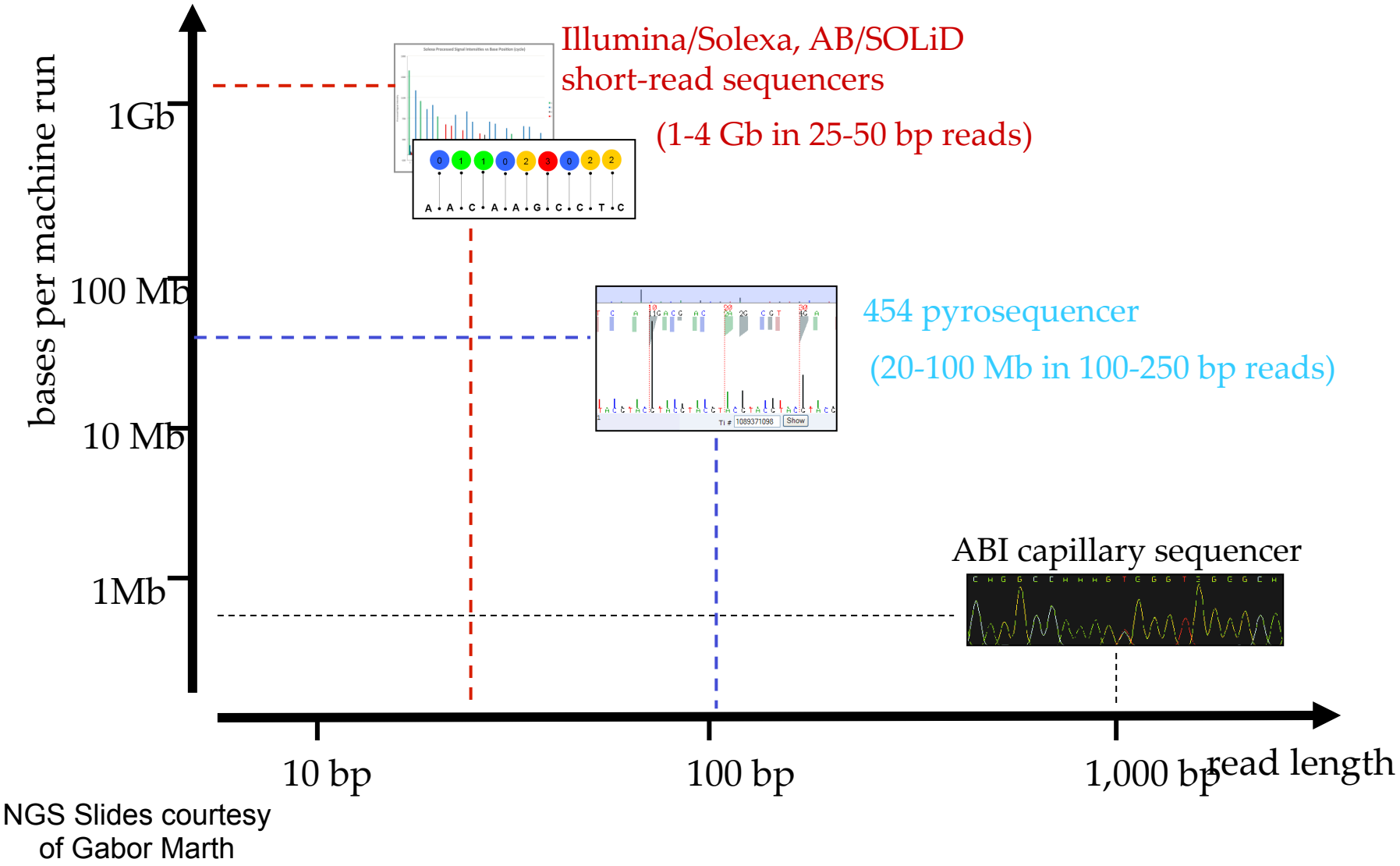



# History of DNA Sequencing

Adapted from Eric Green, NIH; Adapted from Messing & Llaca, *PNAS* (1998)

| Efficiency (bp/person/year) | | |
|---|---|---|
| 1 | 1870 | Miescher: Discovers DNA |
| | 1940 | Avery: Proposes DNA as 'Genetic Material' |
| | | Watson & Crick: Double Helix Structure of DNA |
| 15 | 1953 | Holley: Sequences Yeast $tRNA^{Ala}$ |
| 150 | 1965 | Wu: Sequences λ Cohesive End DNA |
| 1,500 | 1970 | Sanger: Dideoxy Chain Termination<br>Gilbert: Chemical Degradation |
| 15,000 | 1977 | Messing: M13 Cloning |
| 25,000 | 1980 | Hood et al.: Partial Automation |
| 50,000 | 1986 | |
| 200,000 | 1990 | • Cycle Sequencing<br>• Improved Sequencing Enzymes<br>• Improved Fluorescent Detection Schemes |
| 50,000,000 | 2002 | |
| 100,000,000,000 | 2009 | • Next Generation Sequencing<br>•Improved enzymes and chemistry<br>•New image processing |

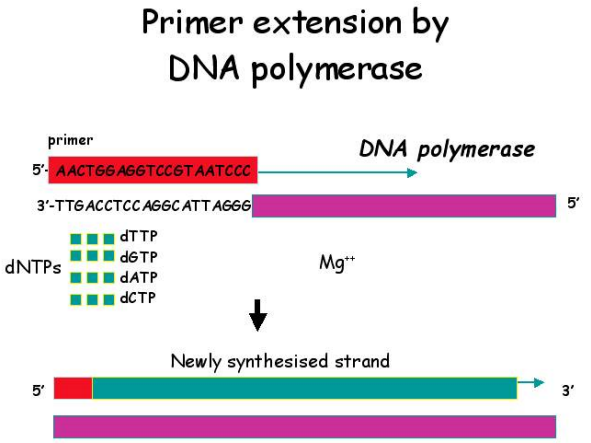
# Read length and throughput

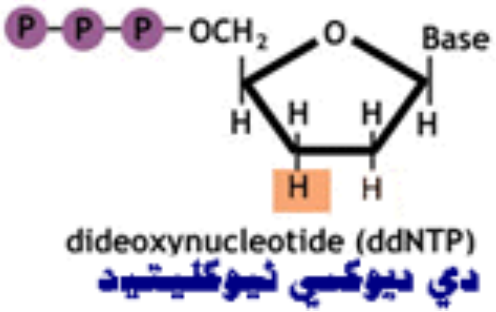


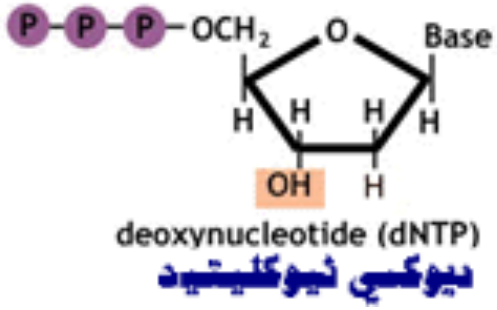

NGS Slides courtesy
of Gabor Marth

# Sanger Sequencing

- Uses DNA polymerase to synthesize a second DNA strand that is labeled. DNA polymerase always adds new bases to the 3’ end of a primer that is base-paired to the template DNA.
  - DNA polymerase is modified to eliminate its editing function
- Also uses chain terminator nucleotides: dideoxy nucleotides (ddNTPs), which lack the -OH group on the 3' carbon of the deoxyribose. When DNA polymerase inserts one of these ddNTPs into the growing DNA chain, the chain terminates, as nothing can be added to its 3' end.



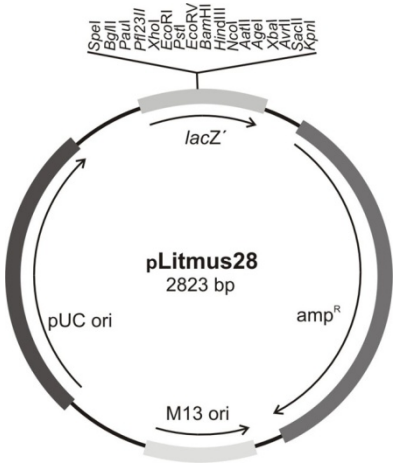



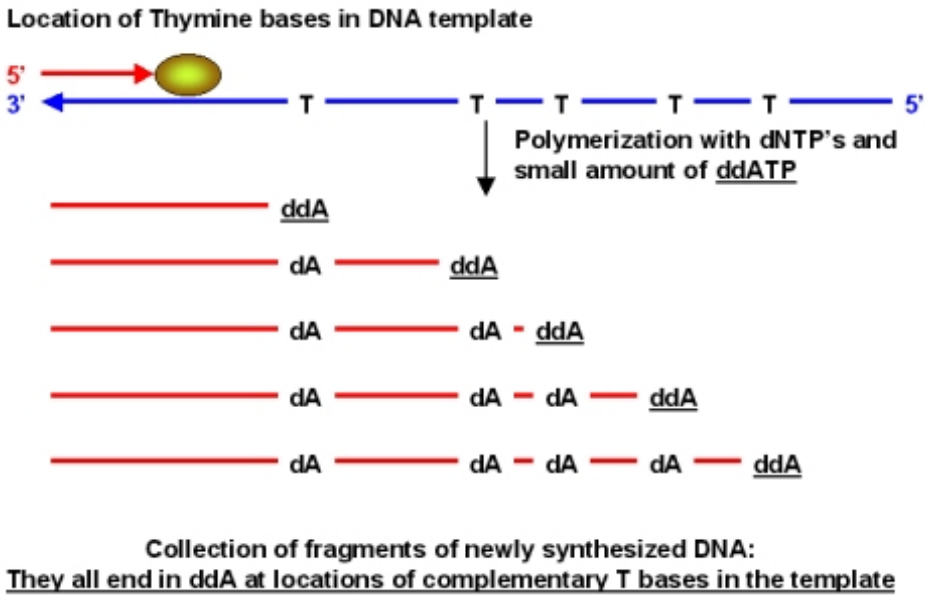



# Sequencing Reaction

- 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.

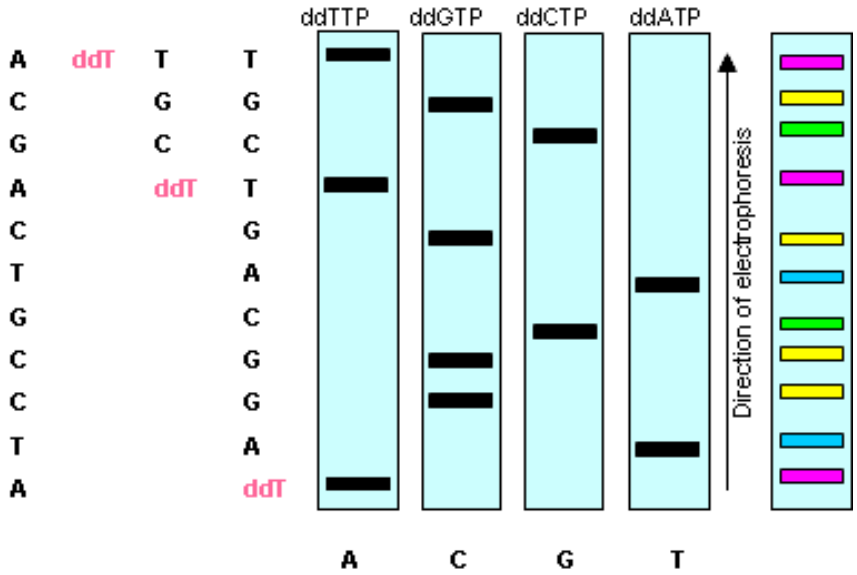
- Sequencing is done by having 4 separate reactions, one for each DNA base.
- All 4 reactions contain the 4 normal dNTPs, but each reaction also contains one of the ddNTPs.
- In each reaction, DNA polymerase starts creating the second strand beginning at the primer.
- When DNA polymerase reaches a base 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.

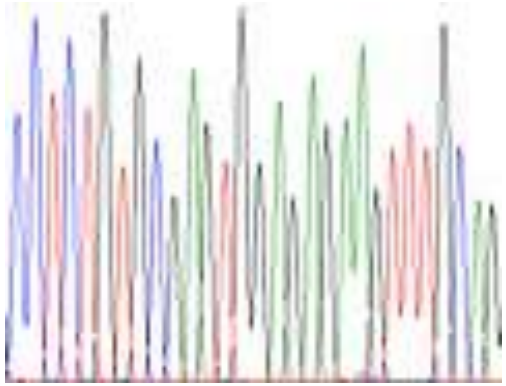


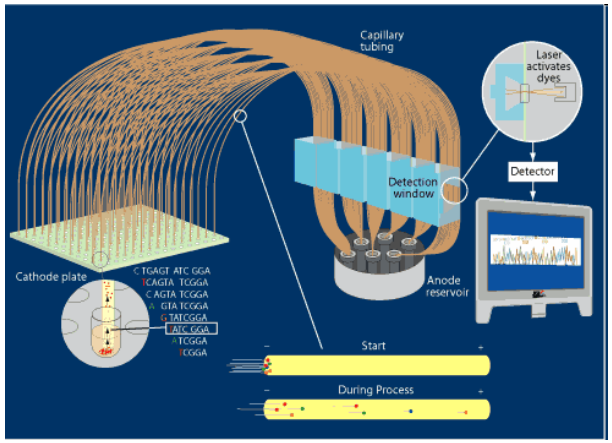



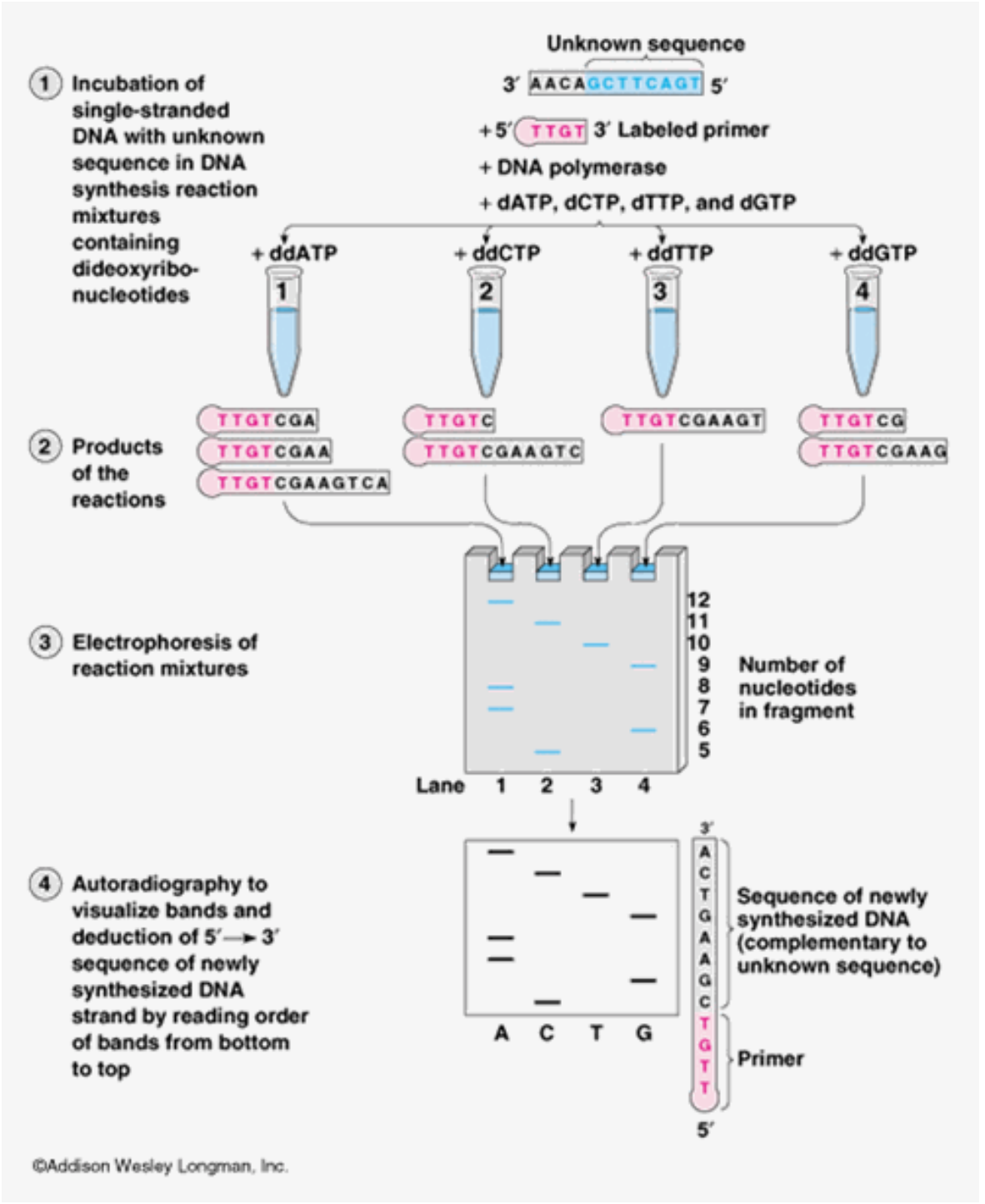
# Electrophoresis

- 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.

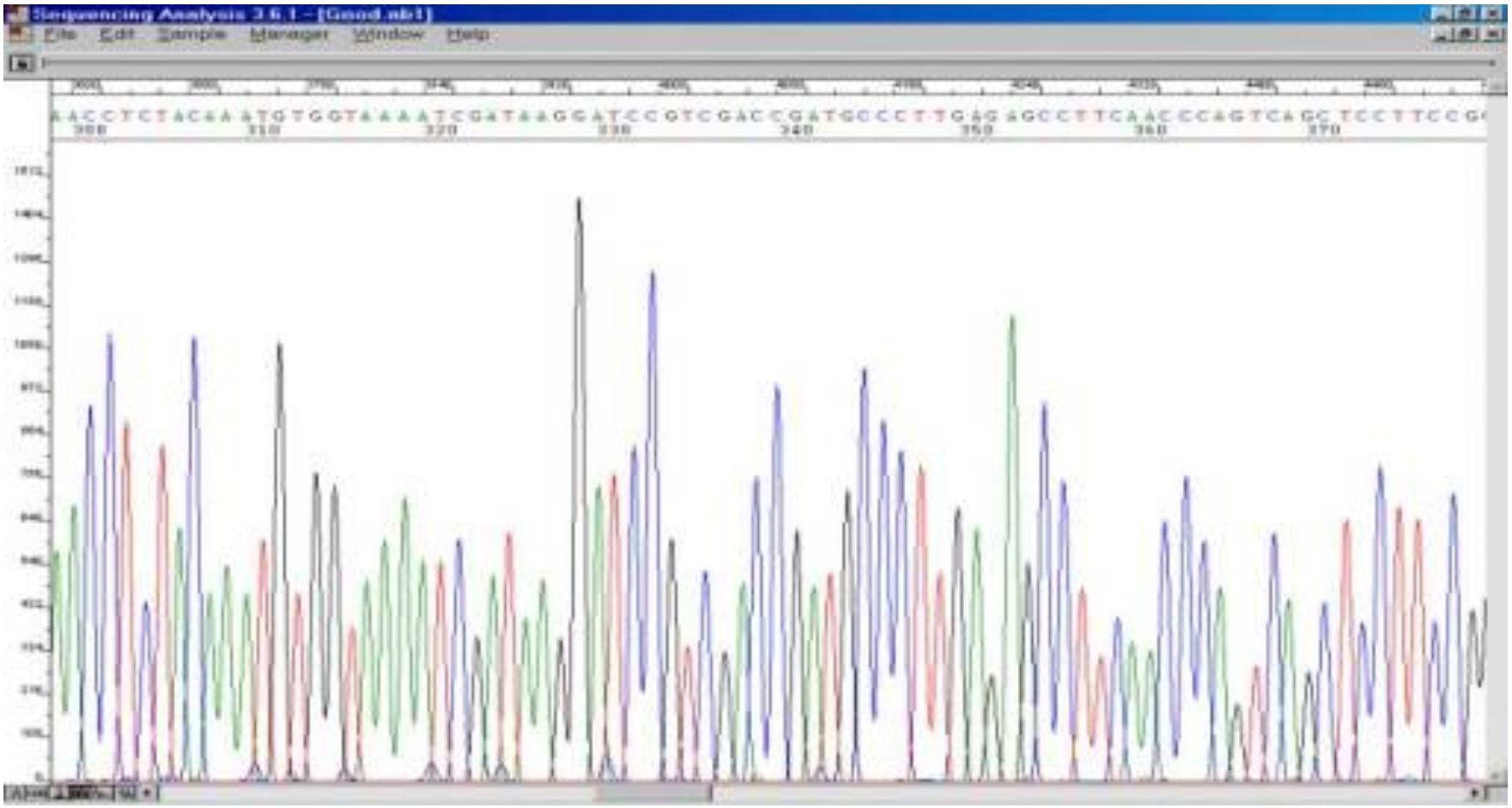


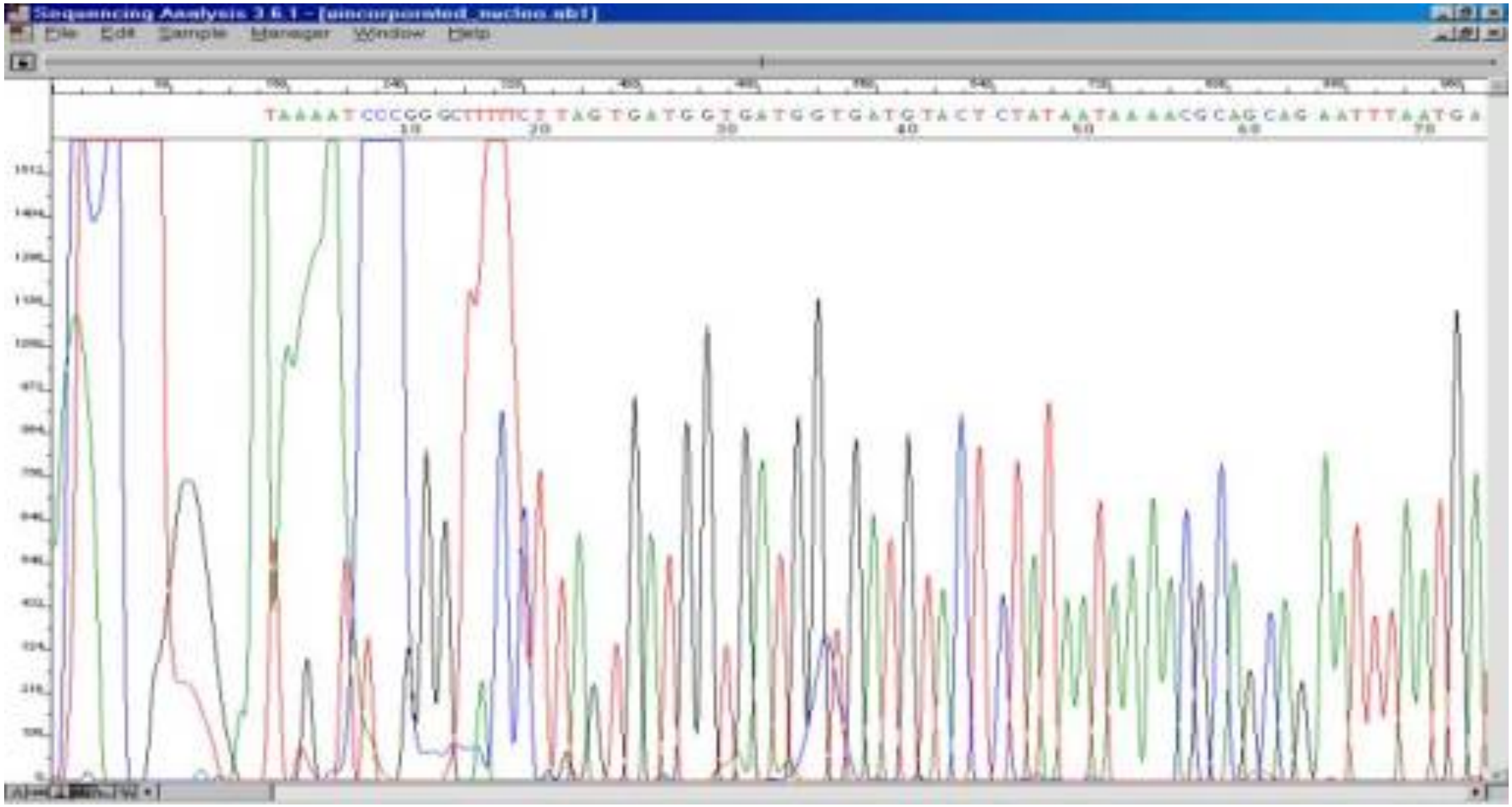



Unknown sequence
3′ AACAGCTTCAGT 5′
1 Incubation of single-stranded DNA with unknown sequence in DNA synthesis reaction mixtures containing dideoxyribo-nucleotides
+5′ TTGT 3′ Labeled primer
+ DNA polymerase
+ dATP, dCTP, dTTP, and dGTP
+ ddATP
+ ddCTP
+ ddTTP
+ ddGTP
1
2
3
4
2 Products of the reactions
TTGTCGA
TTGTCGAA
TTGTCGAAGTCA
TTGTC
TTGTCGAAGTC
TTGTCGAAGT
TTGTCG
TTGTCGAAG
3 Electrophoresis of reaction mixtures
12
11
10
9
8
7
6
5
Number of nucleotides in fragment
Lane 1 2 3 4
4 Autoradiography to visualize bands and deduction of 5′→3′ sequence of newly synthesized DNA strand by reading order of bands from bottom to top
A C T G
3′
ACTGAAGCTGTT
5′
Sequence of newly synthesized DNA (complementary to unknown sequence)
Primer
©Addison Wesley Longman, Inc.

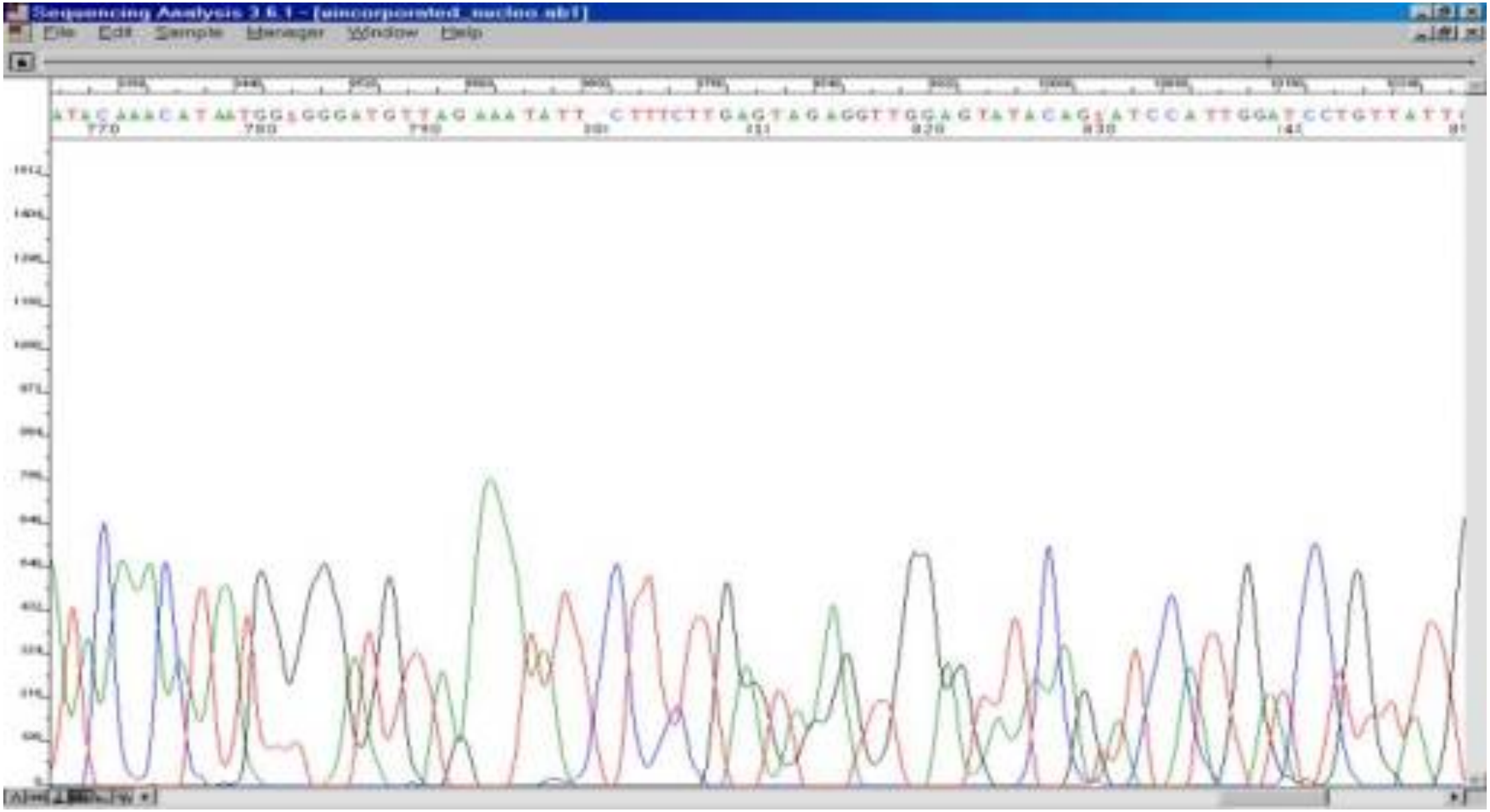
# Electrophoresis diagrams



# Challenging to read answer



# Challenging to read answer



# Reading an electropherogram

1. Filtering
2. Smoothening
3. Correction for length compressions
4. A method for calling the letters – **PHRED**

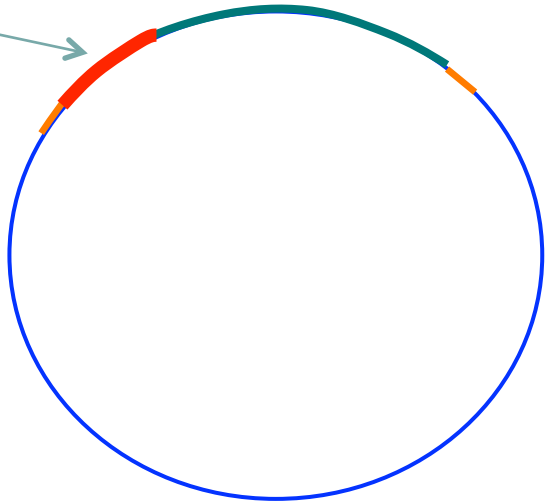
**PHRED** – **PH**il's **R**ead **ED**itor (by Phil Green)

Based on dynamic programming

Several better methods exist, but labs are reluctant to change

# Output of PHRED: a read

A **read**: 500-700 nucleotides



```
A  C  G  A  A  T  C  A  G  …A
16 18 21 23 25 15 28 30 32 …21
```

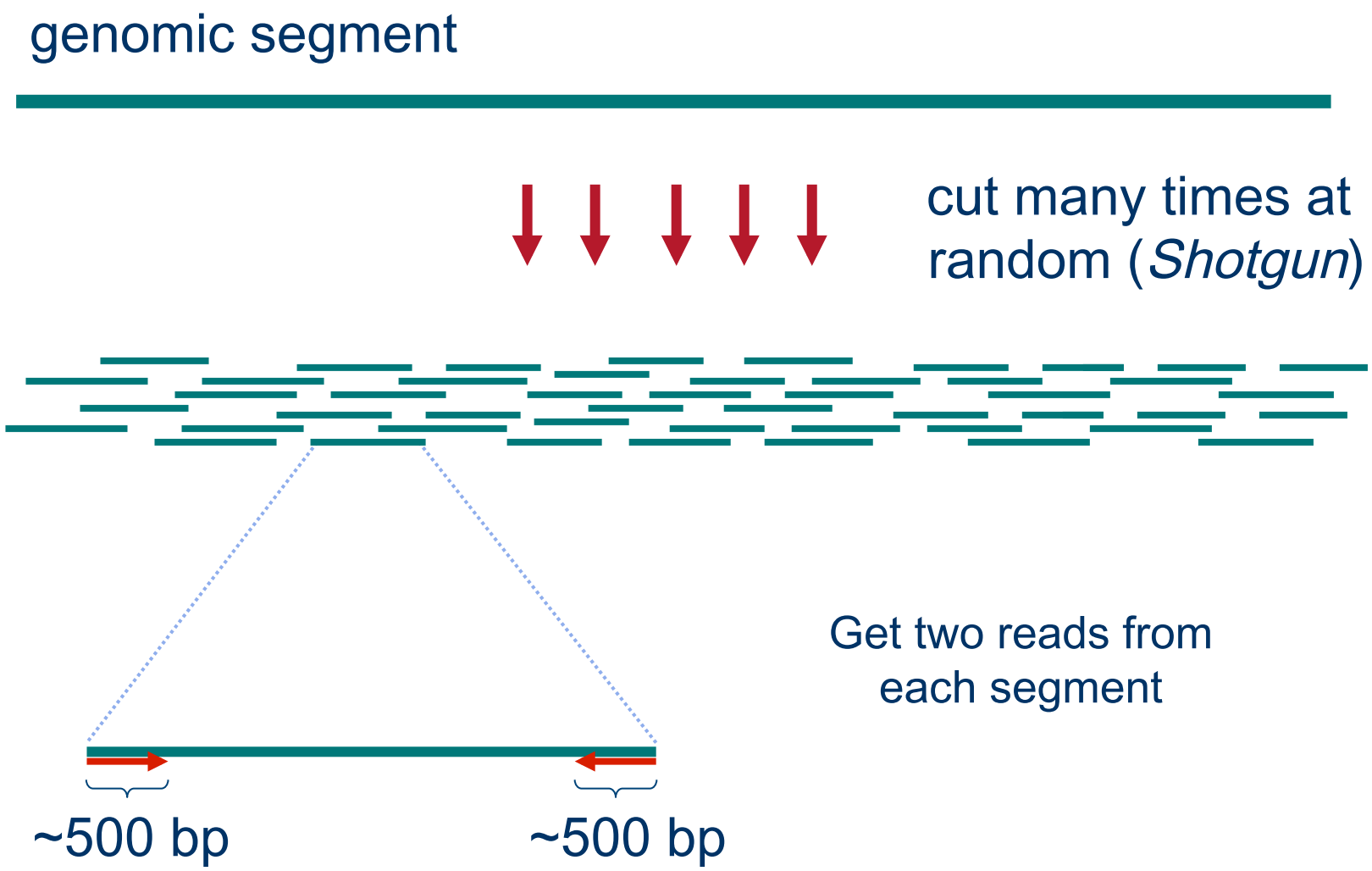
Quality scores: $-10*\log_{10}\text{Prob(Error)}$

Reads can be obtained from leftmost, rightmost ends of the insert

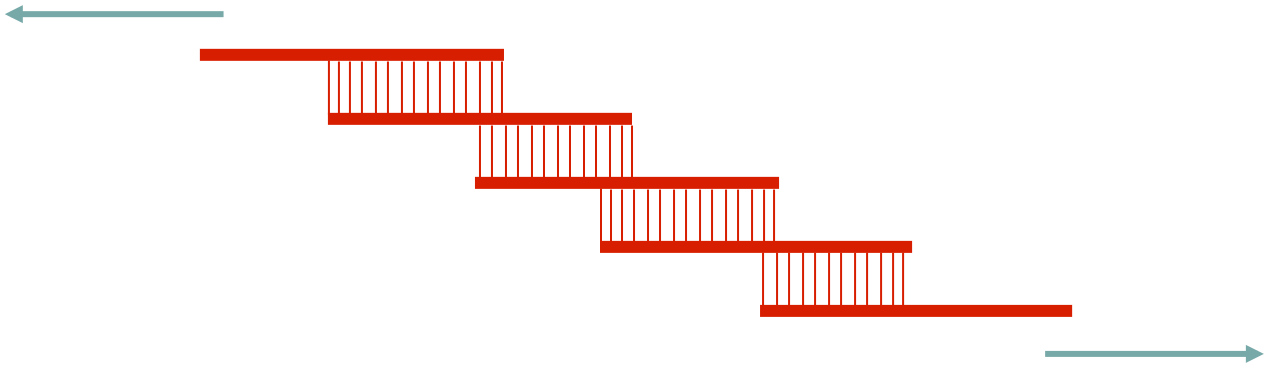
**Double-barreled sequencing:**

Both leftmost & rightmost ends are sequenced

# Method to sequence longer regions



# Reconstructing The Sequence

Cover region with ~7-fold redundancy (7X)

Overlap reads and extend to reconstruct the original genomic region

# Definition of Coverage


Length of genomic segment: **L**
Number of reads: **n**
Length of each read: **l**

**Definition:** Coverage **C = n l / L**

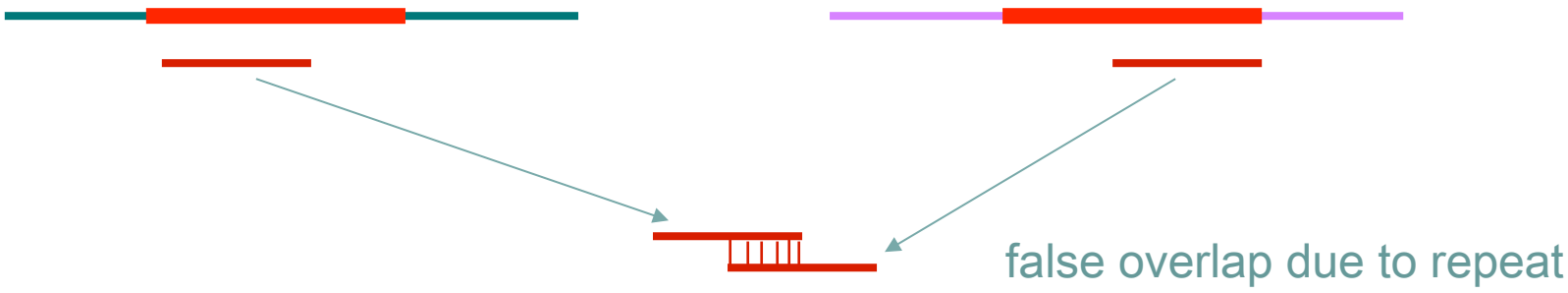
How much coverage is enough?

**Lander-Waterman model:**
Assuming uniform distribution of reads, C=10 results in 1 gapped region /1,000,000 nucleotides

# Challenges with Fragment Assembly

- Sequencing errors
  ~1-2% of bases are wrong

- Repeats




- Computation: ~ O( $N^2$ ) where N = # reads

# Repeats

Bacterial genomes: 5%
Mammals: 50%

**Repeat types:**

- **Low-Complexity DNA** (e.g. ATATATATACATA…)
- **Microsatellite repeats** $(a_1 \ldots a_k)^N$ where k ~ 3-6
  (e.g. CAGCAGTAGCAGCACCAG)
- **Transposons**
  - **SINE** (Short Interspersed Nuclear Elements)
    e.g., ALU: ~300-long, $10^6$ copies
  - **LINE** (Long Interspersed Nuclear Elements)
    ~4000-long, 200,000 copies
  - **LTR retroposons** (Long Terminal Repeats (~700 bp) at each end)
    cousins of HIV
- **Gene Families** genes duplicate & then diverge (paralogs)
- **Recent duplications** ~100,000-long, very similar copies

# Next Generation Sequencing

# High Throughput Sequencing Platforms

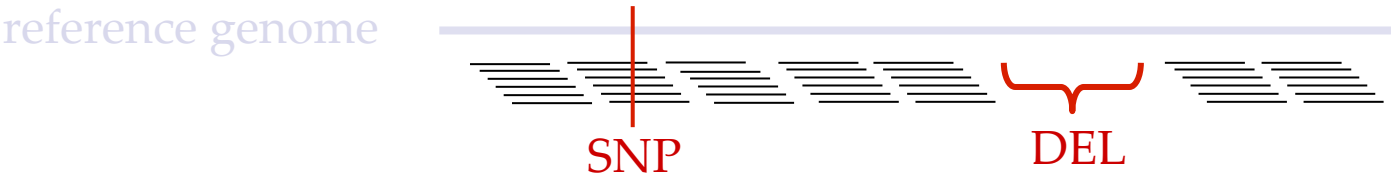
- Illumina HiSeq 1000 and HiSeq 2000
- Life Sciences/Roche 454 pyrosequencing
- ABI Solid Sequencing System *
- Pacific Biosciences *
- Ion Torrent
- Cambridge Nannopore (late 2012?)

# Features of NGS data

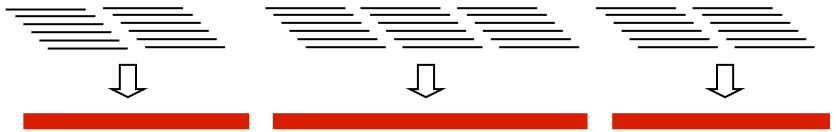
- Short sequence reads
  - –100-200bp: 454 (Roche)
  - –35-120bp Solexa(Illumina), SOLiD(AB)
- Huge amount of sequence per run
  - –Gigabases per run
- Huge number of reads per run
  - –Up to billions
- Higher error (compared with Sanger)
  - –Different error profile

# Current and future application areas

Genome re-sequencing: somatic mutation detection, organismal SNP discovery, mutational profiling, structural variation discovery

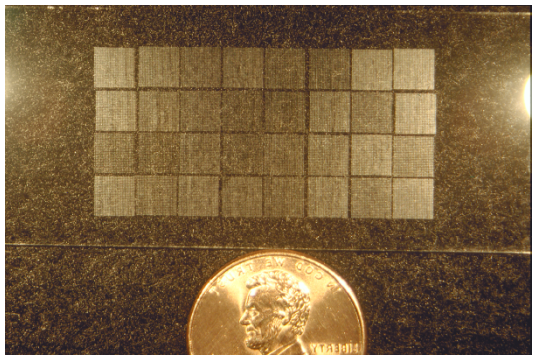



De novo genome sequencing



Short-read sequencing will be (at least) an alternative to microarrays for:

- DNA-protein interaction analysis (CHiP-Seq)
- novel transcript discovery
- quantification of gene expression
- epigenetic analysis (methylation profiling)

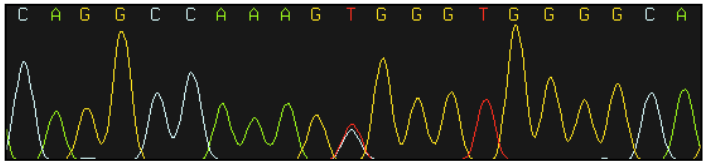


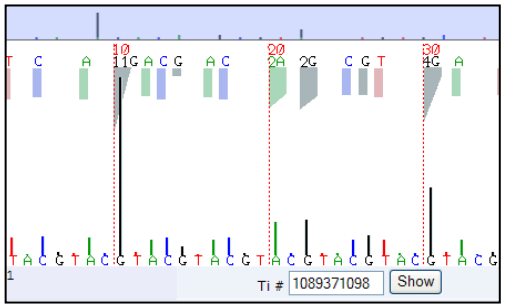
# What can we use them for?

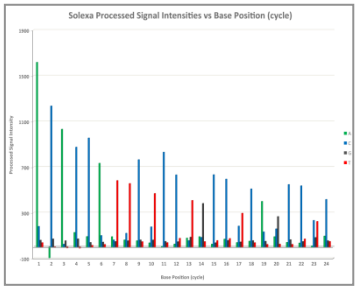
| | SANGER | 454 | Solexa | AB SOLiD |
|---|---|---|---|---|
| De novo assembly | Mammal ($3*10^9$) | Bacteria, Yeast | Bacteria | Bacteria? |
| SNP Discovery | Yes | Yes | 90% of human | 90% of human |
| Larger events | Yes | Yes | Yes | Yes |
| Transcript profiling (rare) | No | Maybe | Yes | Yes |

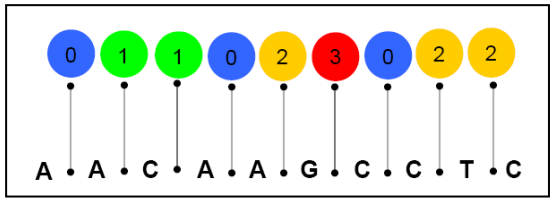
# Next Gen: Raw Data

- Machine Readouts are different



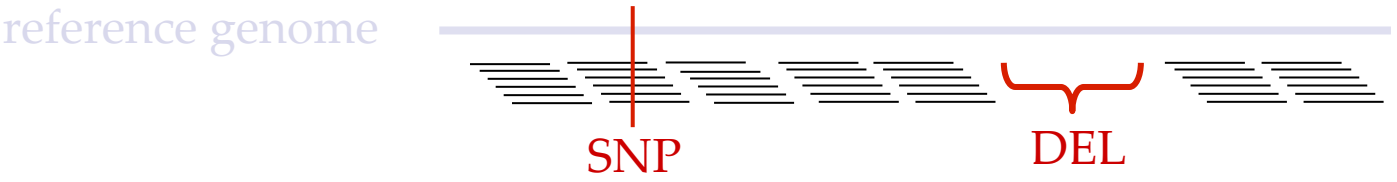




- Read length, accuracy, and error profiles are variable.

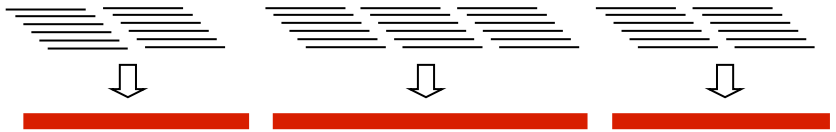
- All parameters change rapidly as machine hardware, chemistry, optics, and noise filtering improves

# Current and future application areas

Genome re-sequencing: somatic mutation detection, organismal SNP discovery, mutational profiling, structural variation discovery

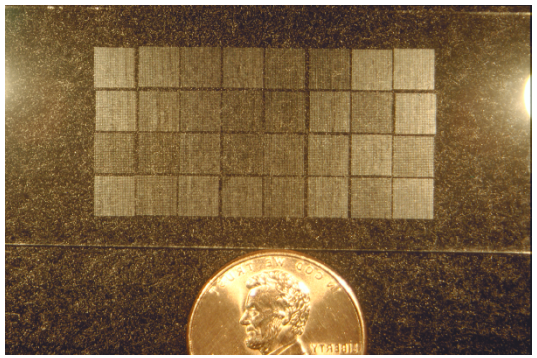



De novo genome sequencing



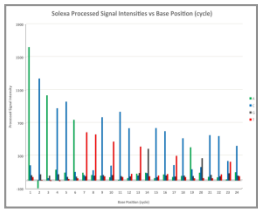
Short-read sequencing will be (at least) an alternative to microarrays for:

- DNA-protein interaction analysis (CHiP-Seq)
- novel transcript discovery
- quantification of gene expression
- epigenetic analysis (methylation profiling)



# Fundamental informatics challenges

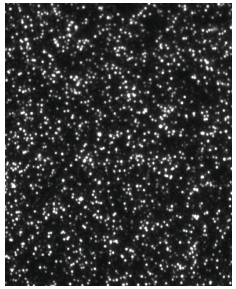
1. Interpreting machine readouts – base calling, base error estimation



```
>B_TITR_1_1_668_35 TIME: Tue Feb 20 02:26:06 2007
ATATCGGATGACACAATATGGGAGGTTGAC
>B_TITR_1_2_843_403 TIME: Tue Feb 20 02:26:06 2007
TGTAGCTTTTCATGACAATTTTATAGGTGT
>B_TITR_1_1_668_35
27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 6
>B_TITR_1_2_843_403
26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 24
>B_TITR_1_3_618_922
```

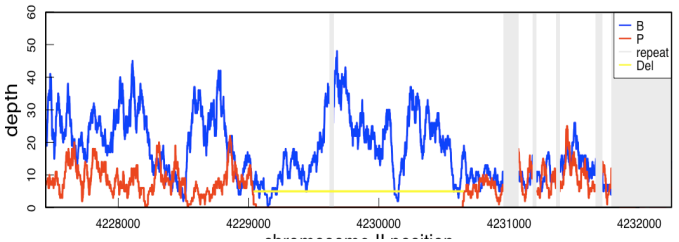
2. Data visualization

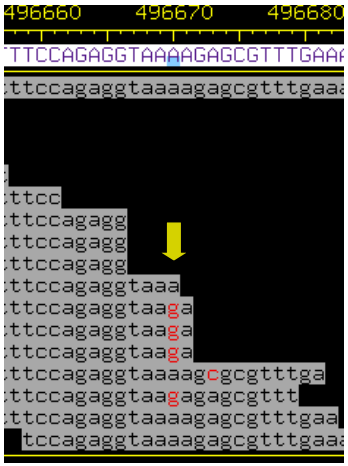
3. Data storage & management



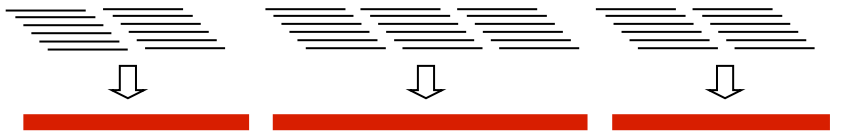
# Informatics challenges (cont'd)

4. SNP and short INDEL, and structural variation discovery

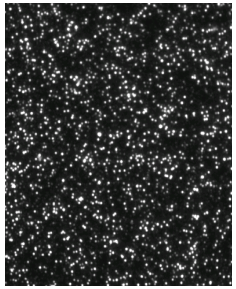



5. De novo Assembly

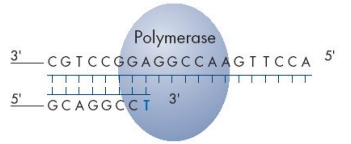


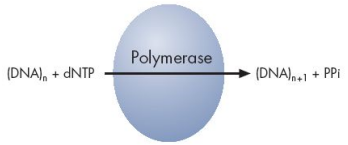
6. Data storage & management

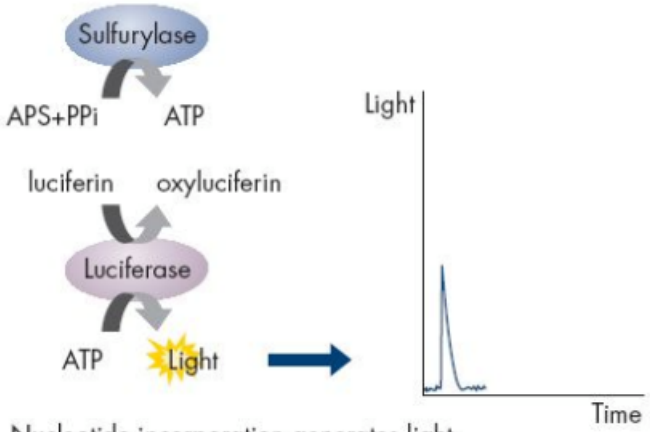


# Pyrosequencing Biochemistry

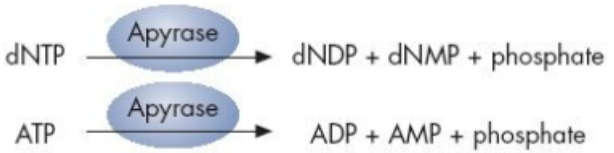
- In DNA synthesis, a dNTP is attached to the 3’ end of the growing DNA strand. The two phosphates on the end are released as pyrophosphate (PPi).
- ATP sulfurylase uses PPi and adenosine 5’-phosphosulfate to make ATP.
  - ATP sulfurylase is normally used in sulfur assimilation: it converts ATP and inorganic sulfate to adenosine 5’-phosphosulfate and PPi. However, the reaction is reversed in pyrosequencing.
- Luciferase is the enzyme that causes fireflies to glow. It uses luciferin and ATP as substrates, converting luciferin to oxyluciferin and releasing visible light.
  - The amount of light released is proportional to the number of nucleotides added to the new DNA strand.
- After the reaction has completed, apyrase is added to destroy any leftover dNTPs.




Nucleotide incorporation generates light seen as a peak in the Pyrogram trace



# 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=kYAGFrbGl6E

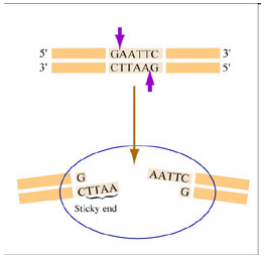
# 454 Technology

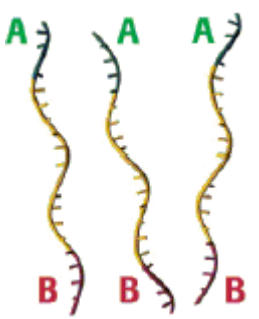
- To start, the DNA is sheared into 300-800 bp fragments, and the ends are “polished” by removing any unpaired bases at the ends.

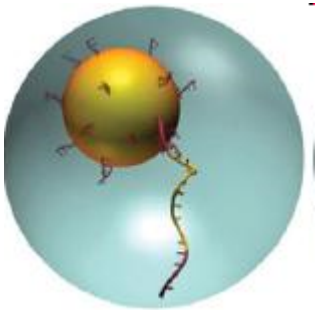
- Adapters are added to each end. The DNA is made single stranded at this point.

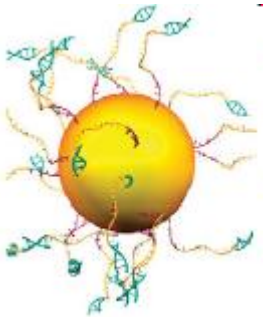
- One adapter contains biotin, which binds to a streptavidin-coated bead. The ratio of beads to DNA molecules is controlled so that most beads get only a single DNA attached to them.

- 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.



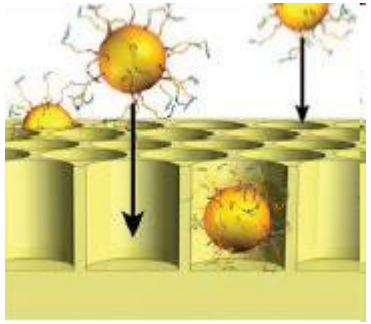





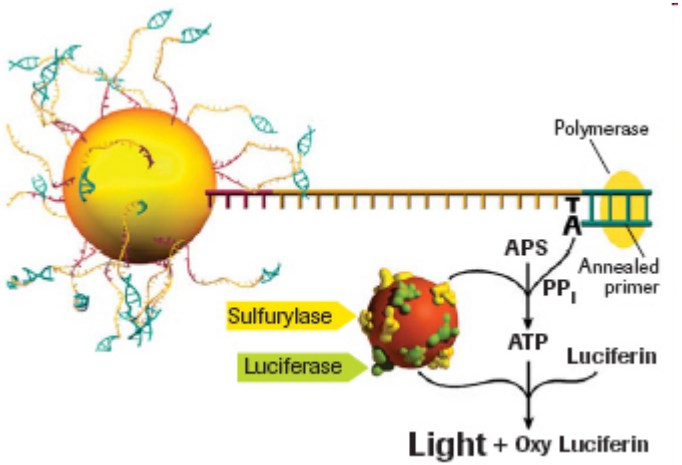
# More 454 Technology

- 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.



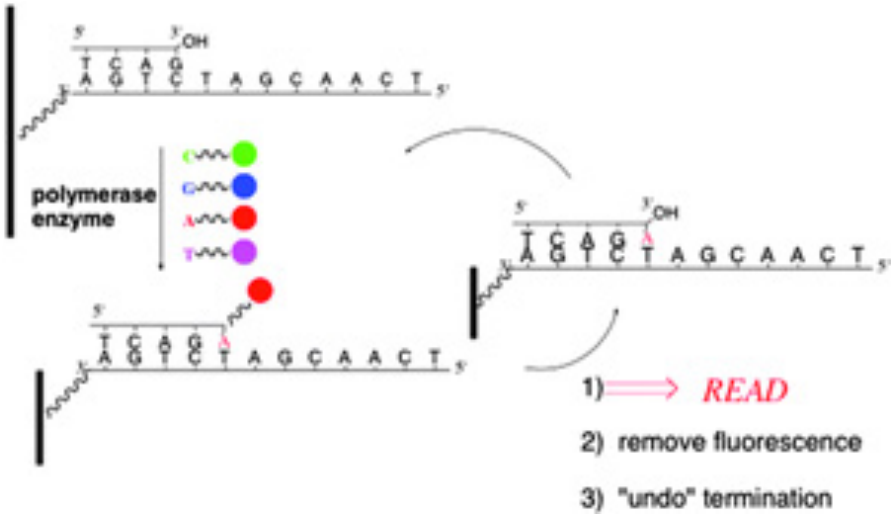

- The plate is coupled to a fiber optic chip. A CCD camera records the light flashes from each well.

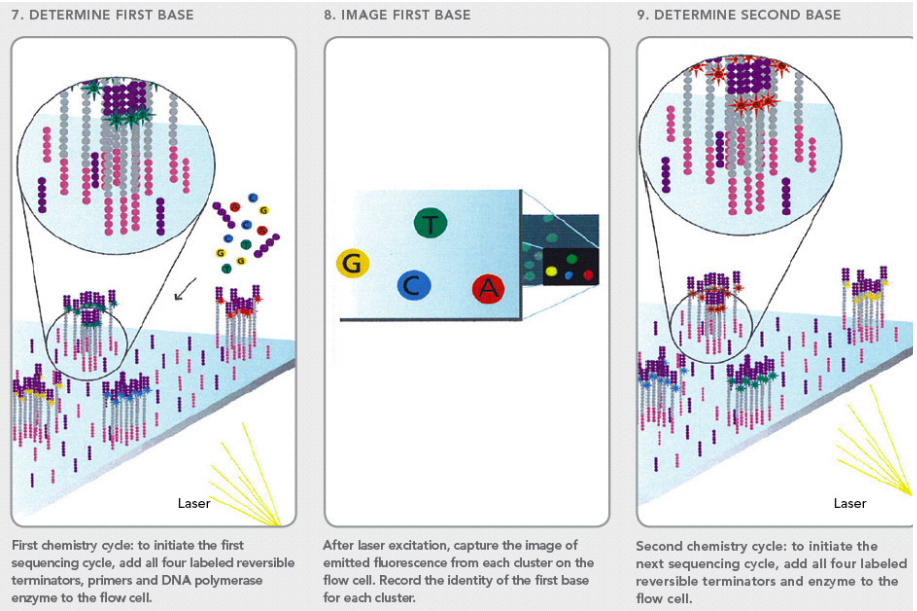
# Illumina/Solexa Sequencing

- Another high throughput Next Generation Sequencing method.
- http://www.youtube.com/watch?v=77r5p8IBwJk&NR=1
- http://www.youtube.com/watch?v=HtuUFUnYB9Y

# Illumina Sequencing Chemistry

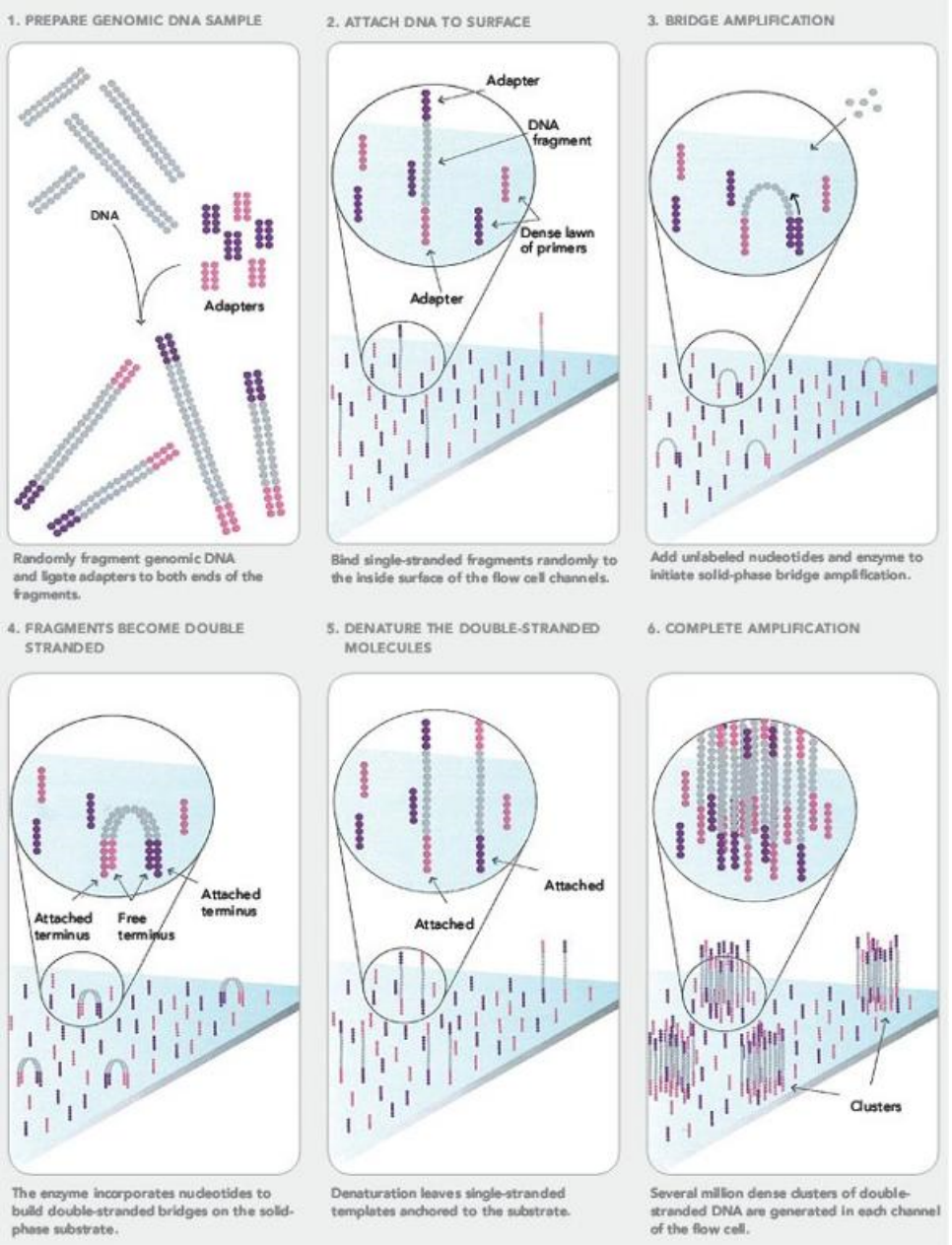
- uses the basic Sanger idea of "sequencing by synthesis" of the second strand of a DNA molecule. Starting with a primer, new bases are added one at a time, with fluorescent tags used to determine which base was added.
- The fluorescent tags block the 3'-OH of the new nucleotide, and so the next base can only be added when the tag is removed.
  - So, unlike pyrosequencing, you never have to worry about how many adjacent bases of the same type are present.
  - The cycle is repeated 50-100 times.



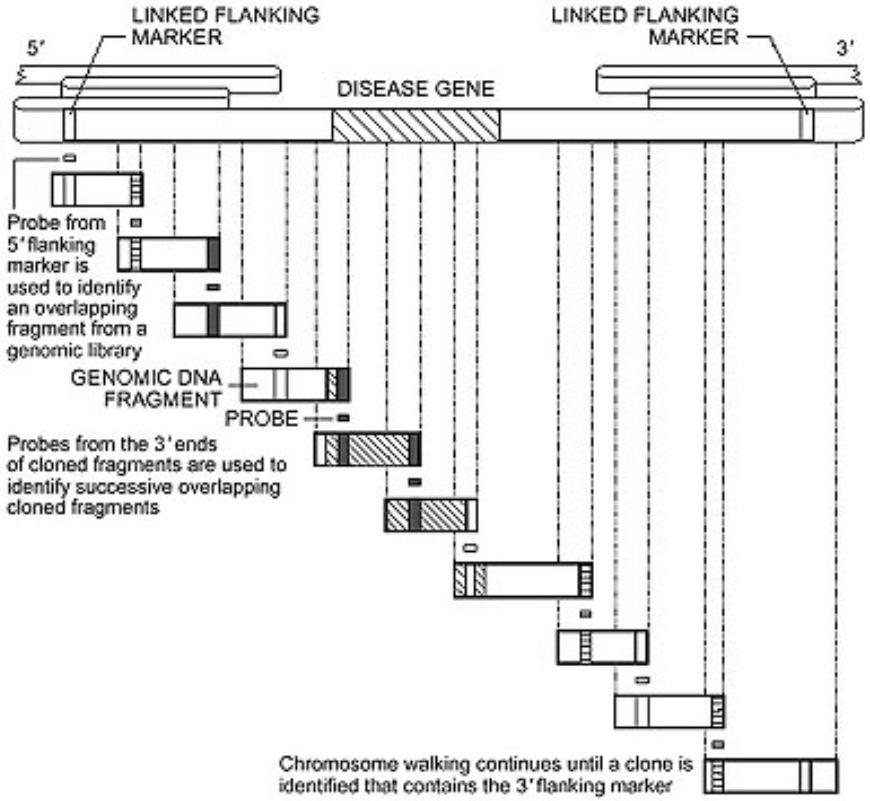
# Illumina Massively Parallel System

- 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.



# Sequence Assembly

- 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.
  - Still useful for fairly short DNA molecules, say 1-10 kbp.

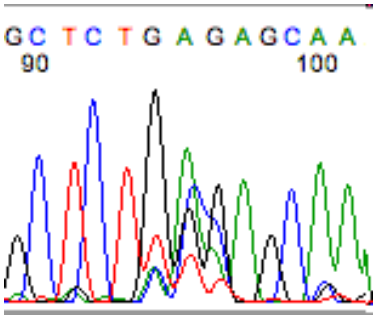


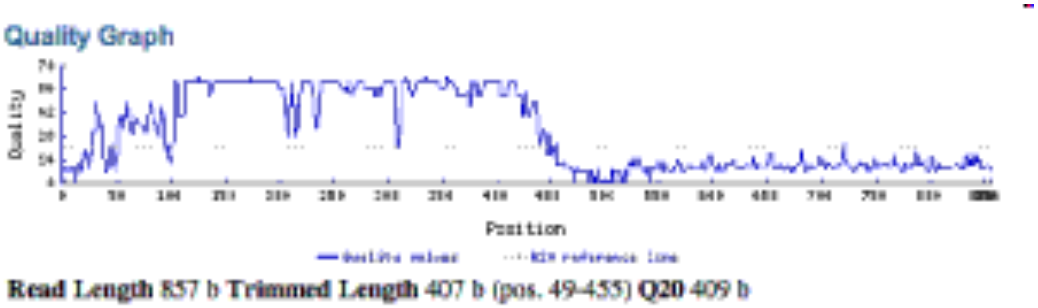
# Assembly Problems

- In principle, assembling a sequence is just a matter of finding overlaps and combining them.
- In practice:
  - most genomes contain multiple copies of many sequences,
  - there are random mutations (either naturally occurring cell-to-cell variation or generated by PCR or cloning),
  - there are sequencing errors and misreadings,
  - sometimes the cloning vector itself is sequenced
  - sometimes miscellaneous junk DNA gets sequenced
- Getting rid of vector sequences is easy once you recognize the problem: just check for them.
- Repeat sequence DNA is very common in eukaryotes, and sequencing highly repeated regions (such as centromeres) remains difficult even now. High quality sequencing helps a lot: small variants can be reliably identified.
- Sequencing errors, bad data, random mutations, etc. were originally dealt with by hand alignment and human judgment. However, this became impractical when dealing with the Human Genome Project.
- This led to the development of automated methods. The most useful was the phred/phrap programs developed by Phil Green and collaborators at Washington University in St. Louis.

# Phred Quality Scores

- Phred is a program that assigns a quality score to each base in a sequence. These scores can then be used to trim bad data from the ends, and to determine how good an overlap actually is.
  - there are much improved algorithms now, but the phred quality score is still widely used.
- Phred scores are logarithmically related to the probability of an error: a score of 10 means a 10% error probability; 20 means a 1% chance, 30 means a 0.1% chance, etc.
  - Q = -10 log P, where Q is the phred score and P is the probability that the base was called incorrectly.
  - A score of 20 is generally considered the minimum acceptable score.



# Phred Algorithm

- Phred uses Fourier analysis (decomposing the data into a series of sine waves) to examine chromatogram trace data.
  - First, find the expected position of each peak, assuming they are supposed to be evenly spaced, with no compressions or other factors altering peak positions.
  - Next, find actual center and area of each peak.
  - Finally, match observed peaks to expected. This involves splitting some peaks and ignoring others.
- Assigning a quality score involves comparing various parameters determined for each peak with data that was generated from known sequences run under a wide variety of conditions (i.e., based on ad hoc observations and not theory).
- Since all four traces (A, C, G, T) are examined separately, phred generates a best-guess sequence. Phred output is a file where each line contains one base and its quality score.

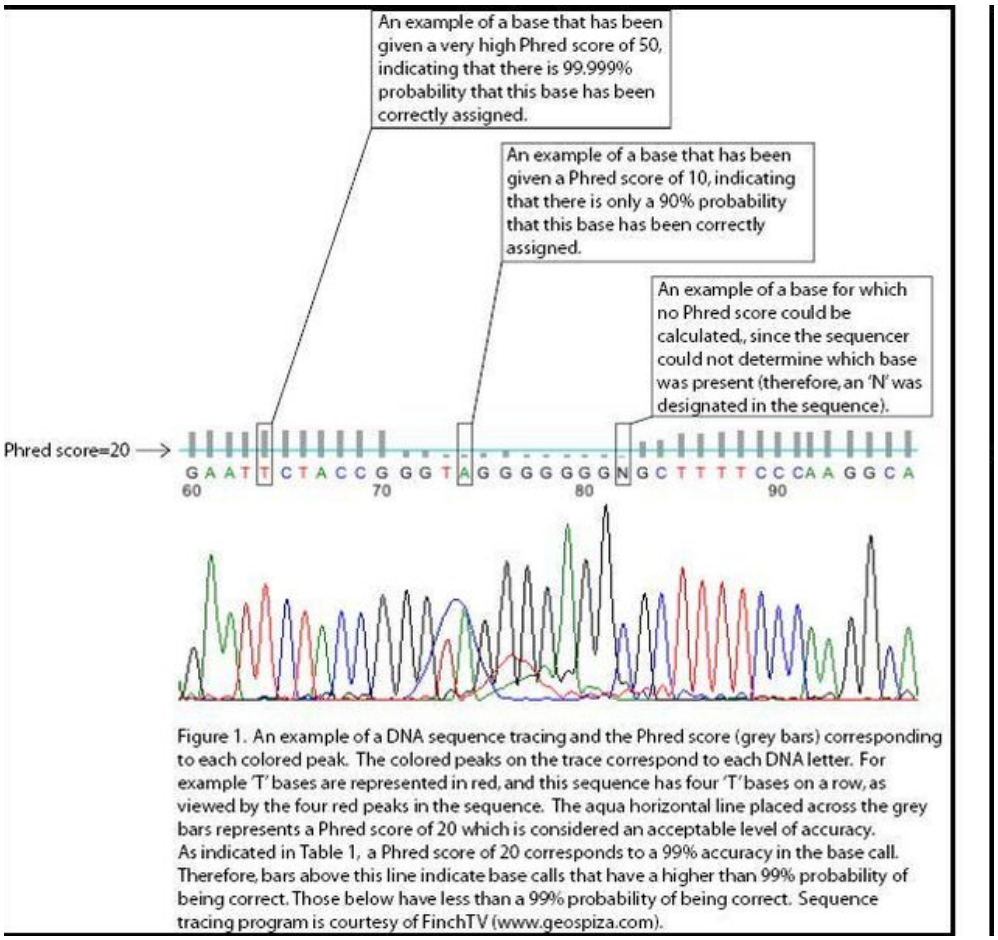



Figure 1. An example of a DNA sequence tracing and the Phred score (grey bars) corresponding to each colored peak. The colored peaks on the trace correspond to each DNA letter. For example 'T' bases are represented in red, and this sequence has four 'T' bases on a row, as viewed by the four red peaks in the sequence. The aqua horizontal line placed across the grey bars represents a Phred score of 20 which is considered an acceptable level of accuracy. As indicated in Table 1, a Phred score of 20 corresponds to a 99% accuracy in the base call. Therefore, bars above this line indicate base calls that have a higher than 99% probability of being correct. Those below have less than a 99% probability of being correct. Sequence tracing program is courtesy of FinchTV (www.geospiza.com).

# Combining Sequences with Phrap

- Phrap first examines all reads for matching "words": short sections of identical sequence. The matching words need to be in the same order and spacing.
  - Sequences in both orientations are examined, using the reverse-complement sequence if necessary.
- The entire sequences of pairs with matching words are then aligned using the Smith-Waterman algorithm (a standard technique we will look at later).
  - Phrap then looks for discrepancies in the combined sequences, using phred scores to decide between alternatives. Phrap generates quality scores from the combined phred data.
  - Sequencing errors are not necessarily random: homopolymeric regions (several of the same base in a row) are notoriously tricky to sequence accurately. Using the opposite strand often helps resolve these regions. Also using a different sequencing technology or chemistry.
- Sequences are combined with a greedy algorithm: all pairs of fragments are scored for the length and quality of their overlap region, and then the largest and best-matched pair is merged. This process is repeated until some minimum score is reached.
- The result is a set of contigs: reads assembled into a continuous DNA sequence.
  - The ideal result is the entire chromosome assembled into a single contig.

# 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.

# Comparison of the technologies

| | SANGER | 454 | Solexa | AB SOLiD |
|---|---|---|---|---|
| Output | Sequence | Flowgram | Sequence | Colors |
| Read Length | 500-700 | 250-500 | 35-70 | 35-50 |
| Error rate | 2% | 3% (indels) | 1% | 4% or 0.06% |
| Mb per run | 0.8 | 20 | 10000 | 20000 |
| Cost per Mb | $1000 | $50 | $0.15 | $0.05 |
| Paired? | Yes | Sort of | Yes (<1k) | Yes (<10k) |

# THANKS