Bioanalytical chemistry

6. DNA sequencing

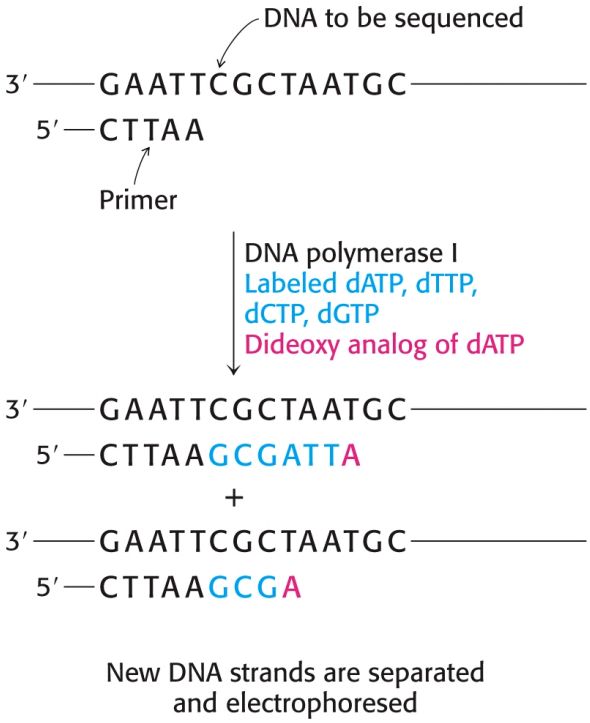
Some objectives for this section

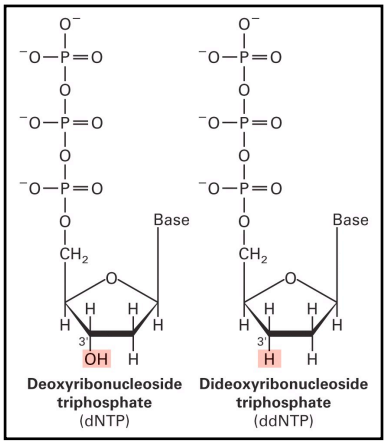
⇒ You will know how DNA was traditionally sequenced by 1 color, 4 lane methods
⇒ You will know how DNA was traditionally sequenced by 4 color, 1 lane methods
⇒ You will know how DNA was traditionally sequenced by capillary electrophoresis methods
⇒ You will appreciate the role that chemists had in developing the technology that made the Human Genome project possible.
⇒ You will have a good understanding of the 'Next Generation' methods for genome sequencing
⇒ You will understand how sequencing by synthesis and sequencing by ligation work

Primary Source Material

- Chapter 6 of Biochemistry: Berg, Jeremy M.; Tymoczko, John L.; and Stryer, Lubert (NCBI bookshelf).
- Chapter 7 of Molecular Cell Biology 4th ed. (Ch. 9, 5th ed.): Lodish, Harvey; Berk, Arnold; Zipursky, S. Lawrence; Matsudaira, Paul; Baltimore, David; Darnell, James E. (NCBI bookshelf).
- Chapter 12 of Introduction to Genetic Analysis Anthony: J.F. Griffiths, Jeffrey H. Miller, David T. Suzuki, Richard C. Lewontin, William M. Gelbart (NCBI bookshelf).
- Jonathan M Rothberg & John H Leamon, "The development and impact of 454 sequencing", *Nature Biotechnology* 26, 1117 - 1124 (2008).
- Elaine R. Mardis, "The impact of next-generation sequencing technology on genetics" *Trends in Genetics*, 24, 133-141 (2008)

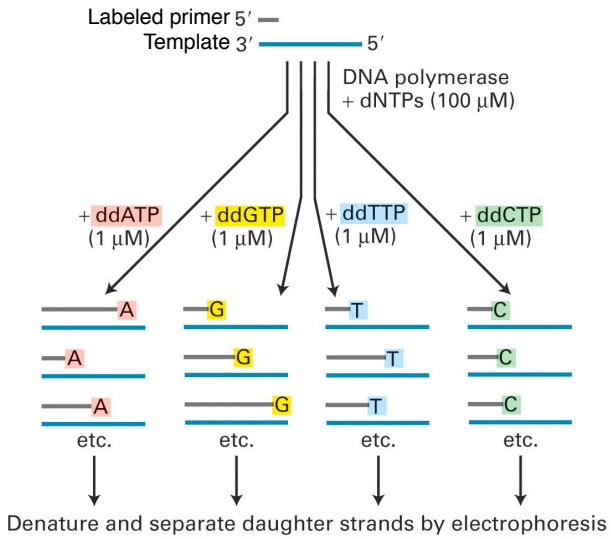
Early DNA sequencing

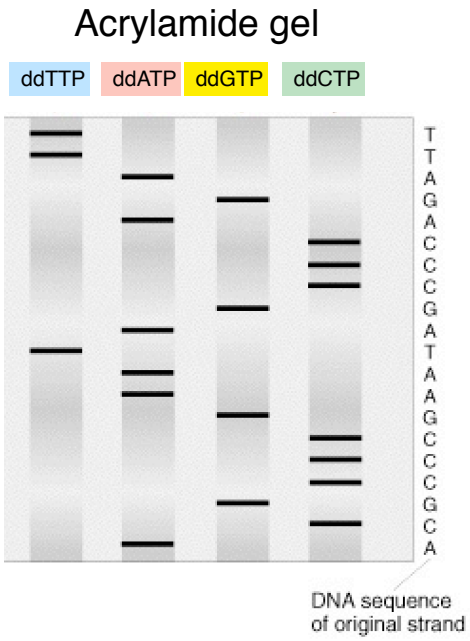


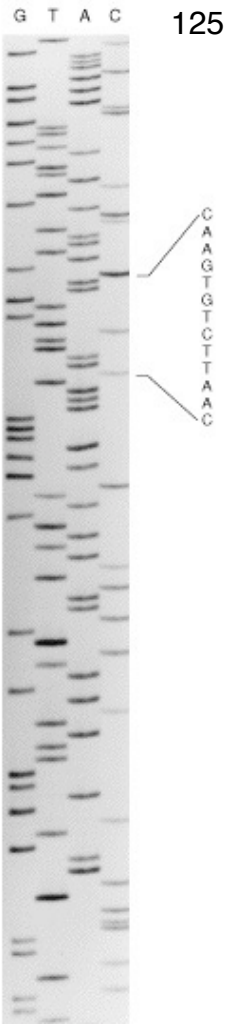


- The analysis of DNA structure and its role in gene expression also have been markedly facilitated by the development of powerful techniques for the *sequencing* of DNA molecules.
- The key to DNA sequencing is the generation of DNA fragments whose length depends on the last base in the sequence. Collections of such fragments can be generated through the *controlled interruption of enzymatic replication*, a method developed by Frederick Sanger and coworkers. This technique has superseded alternative methods because of its simplicity.
- The same procedure is performed on four reaction mixtures at the same time. In all these mixtures, a DNA polymerase is used to make the complement of a particular sequence within a single-stranded DNA molecule.
- The synthesis is primed by a fragment, usually obtained by chemical synthetic methods described on a previous slide, that is complementary to a part of the sequence known from other studies.
- In addition to the four deoxyribonucleoside triphosphates (radioactively labeled), each reaction mixture contains a small amount of the *2',3'-dideoxy analog* of *one* of the nucleotides, a different nucleotide for each reaction mixture. The incorporation of this analog blocks further growth of the new chain because it lacks the 3'hydroxyl terminus needed to form the next phosphodiester bond.
- http://www.dnalc.org/resources/animations/sangerseq.html

Early DNA sequencing

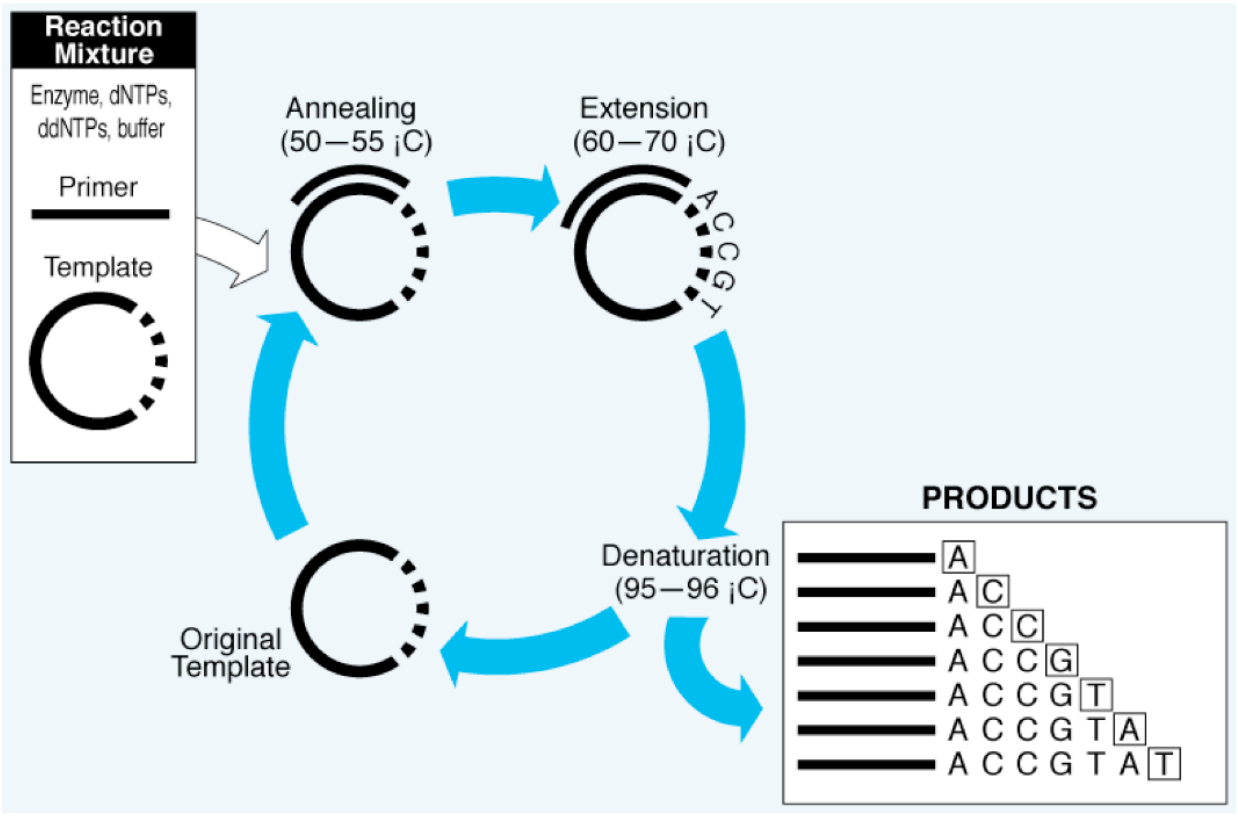






- The concentration of the dideoxy analog is low enough that chain termination will take place only occasionally. The polymerase will sometimes insert the correct nucleotide and other times the dideoxy analog, stopping the reaction.
- For instance, if the dideoxy analog of dATP is present, fragments of various lengths are produced, but all will be terminated by the dideoxy adenosine analog. Importantly, this dideoxy analog of dATP will be inserted only where a T was located in the DNA being sequenced. Thus, the fragments of different length will correspond to the positions of T.
- Four such sets of *chain-terminated fragments* (one for each dideoxy analog) then undergo electrophoresis, and the base sequence of the new DNA is read from the autoradiogram of the four lanes.
- *Question: In the Sanger method for sequencing, is the sequence that we read on the electrophoresis gel the complement of the ssDNA original sequence? Since in the picture above, it says that the order of the bases we read on the gel is the same as the sequence of the DNA, while I think it should be the complementary.*
- *Reply: The easiest way to think about this is to realize that the sequence being read off is for the strand that is the same as the primer.*

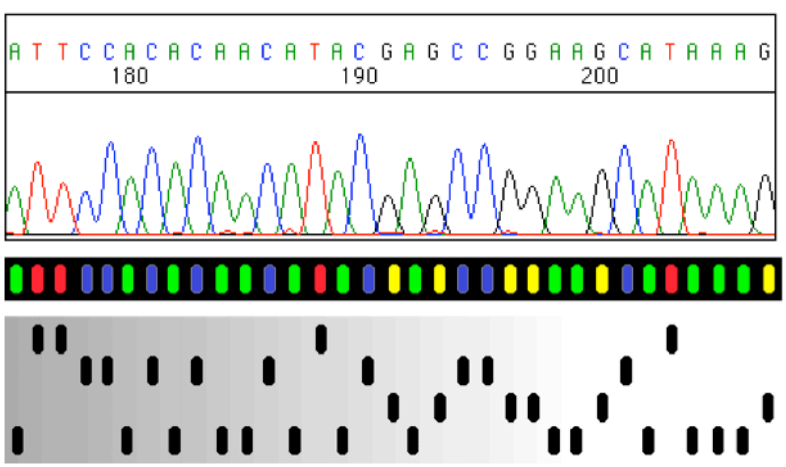
Cycle sequencing of DNA



Applied biosystems Chemistry guide to DNA sequencing (2.9 MB .pdf)

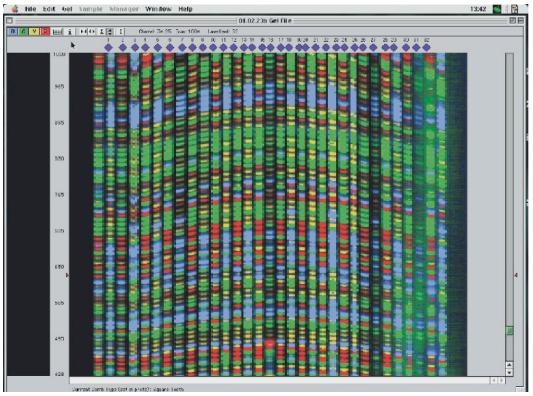
- Cycle sequencing is a simple method in which successive rounds of denaturation, annealing, and extension in a thermal cycler result in linear amplification of extension products. The products are then loaded onto a gel or injected into a capillary.
- For the sake of the above figure, assume that there is a single stranded circular DNA template. The primer will anneal to a specific site on the template and that is where the sequencing reaction will start. After the extension reaction has been allowed to progress for sometime the reaction is heated up to 96 ºC to denature (melt) the DNA into two separate strands.
- The reaction is then cooled back to 50 ºC to allow the primers (which are present in a large excess) to reanneal to the DNA template. The thermophilic DNA polymerase (such as Taq) is most active at 60-70 ºC.
- Advantages of cycle sequencing
  - Protocols are robust and easy to perform
  - Requires much less template DNA than single-temperature extension methods.
  - High temperatures reduce secondary structure, allowing for more complete extension.
  - High temperatures reduce secondary primer-to-template annealing
  - The same protocol is used for double- and single-stranded DNA

Multiple fluorescent labels enables 'one lane' fluorescence detection



4 color, 1 lane fluorescence sequencing

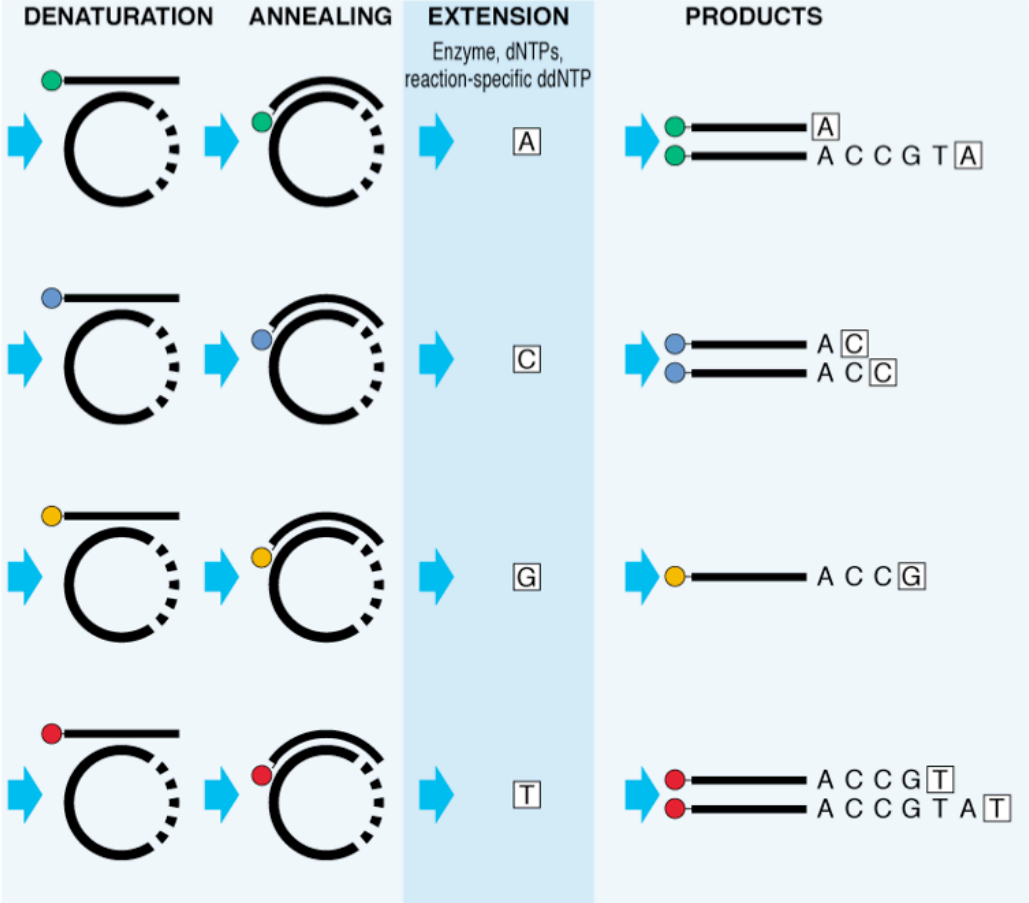
1 color, 4 lane radioactive sequencing



Screen shot of computer software for processing 4 color, 1 lane sequencing data. Note that each lane is a separate sample.

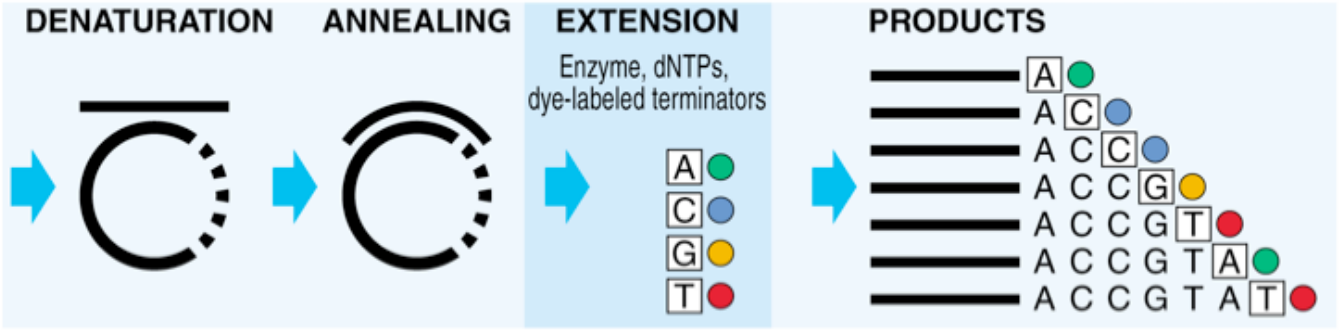
- Fluorescence detection is a highly effective alternative to autoradiography.
- **A fluorescent tag is attached to an oligonucleotide priming fragment**—a differently colored one in each of the four chain-terminating reaction mixtures (e.g., a blue emitter for termination at A and a red one for termination at C).
- The reaction mixtures are combined and subjected to electrophoresis together. The separated bands of DNA are then detected by their fluorescence as they emerge from the gel; the sequence of their colors directly gives the base sequence.
- Sequences of as many as 500 bases can be determined in this way.
- **Alternatively, the dideoxy analogs can be labeled**, each with a specific fluorescent label. When this method is used, all four terminators can be placed in a single tube, and only one reaction is necessary.
- Fluorescence detection is attractive because it eliminates the use of radioactive reagents and can be readily automated.

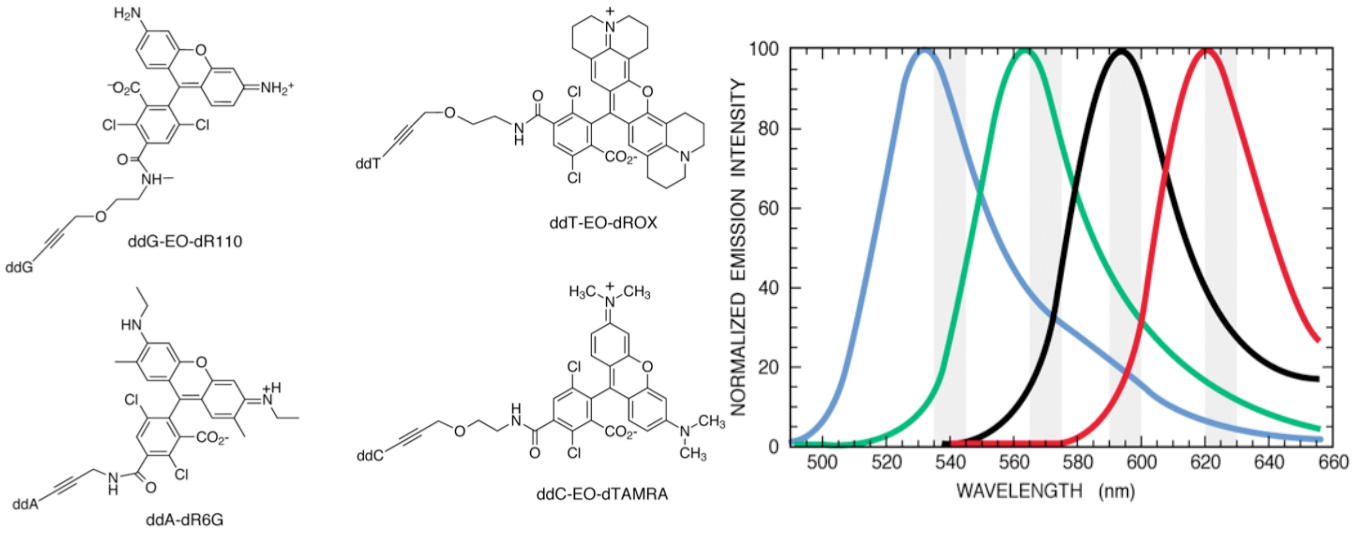
Sequencing DNA with labeled primers



- With dye primer labeling, primers are tagged with four different fluorescent dyes. Labeled products are generated in four separate base-specific reactions. The products from these four reactions are then combined and loaded into a single gel lane or capillary injection
- Dye primer chemistries generally produce greater signal intensities than dye terminator chemistries
- Labeled primers are available for common priming sites. Custom primers can also be labeled
- Four-color dye-labeled reactions are loaded onto a single lane of a gel (I.e. four color, one lane)

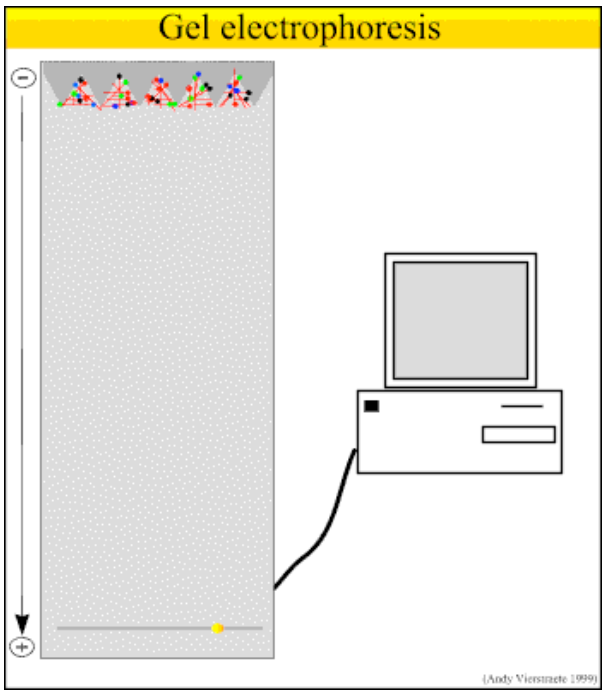
Sequencing DNA with labeled terminators

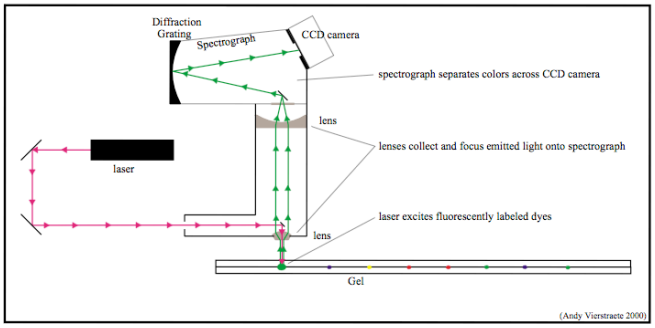


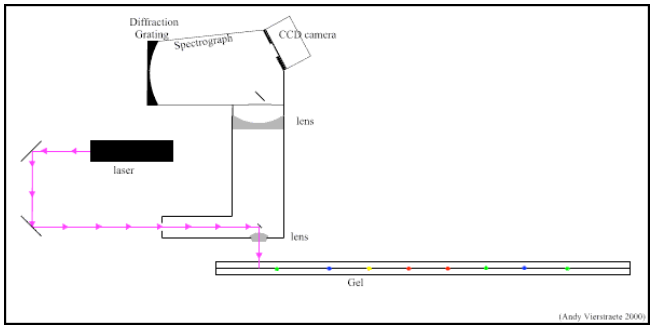


- With dye terminator labeling, each of the four dideoxy terminators (ddNTPs) is tagged with a different fluorescent dye. The growing chain is simultaneously terminated and labeled with the dye that corresponds to that base.
- Advantages include:
  - An unlabeled primer can be used.
  - All 4 reactions can be done in one tube which is then loaded in a single gel lane.
  - False stops are minimized, i.e. fragments that are not terminated by a dideoxynucleotide go undetected because no dye is attached.
- The spectra of the 4 dyes in the dRhodamine terminator set from Applied Biosystems is shown.
- Notice that spectral separation is not perfect for these dyes. Although each of these dyes emits its maximum fluorescence at a different wavelength, there is some overlap in the emission spectra between the four dyes. Multicomponent analysis is used to isolate the signal from each dye so that there is as little noise in the data as possible.
- **The sequencing reaction :**There are three major steps in a sequencing reaction (like in PCR), which are repeated for 30 or 40 cycles.
  1. **Denaturation** at 94°C :During the denaturation, the double strand melts open to single stranded DNA, all enzymatic reactions stop (for example : the extension from a previous cycle).
  2. **Annealing** at 50°C :In sequencing reactions, only one primer is used, so there is only one strand copied (in PCR : two primers are used, so two strands are copied).
  3. **Extension** at 60°C :This is the ideal working temperature for the polymerase. Normally it is 72 °C, but because it has to incorporate ddNTP's which are chemically modified with a fluorescent label, the temperature is lowered so it has time to incorporate the 'strange' molecules.
- Because only one primer is used, only one strand is copied during sequencing, there is a **linear** increase of the number of copies of one strand of the gene. Therefore, there has to be a large amount of copies of the gene in the starting mixture for sequencing. Suppose there are 1000 copies of the wanted gene before the cycling starts, after one cycle, there will be 2000 copies : the 1000 original templates and 1000 complementary strands with each one fluorescent label on the last base. After two cycles, there will be 2000 complementary strands, three cycles will result in 3000 complementary strands and so on.

Gel electrophoresis and detection on an automated sequencer ABI 377







Homepage of Andy Vierstraete
http://users.ugent.be/~avierstr/index.html

- **Separation of the molecules:** After the sequencing reactions, the mixture of strands, all of different length and all ending on a fluorescently labelled ddNTP have to be separated.
- This is done with acrylamide gel electrophoresis which is capable of separating DNA molecules that differ in length by as only one base.
- **Detection on an automated sequencer:** The fluorescently labeled fragments that migrate through the gel, are passing a laser beam at the bottom of the gel. The laser excites the fluorescent molecule, which sends out light of a distinct color. That light is collected and focused by lenses into a spectrograph.
- Each base has its own color, so the sequencer can detect the order of the bases in the sequenced gene.

Norm Dovichi is a pioneer of high-speed DNA sequencing

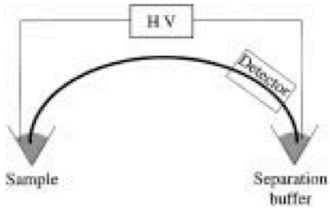


Figure 4. Single-capillary electrophoresis instrument. A fused-silica capillary is used for the separation. One end of the capillary is dipped into the sample or buffer reservoir, while the other end passes through a detector before being placed in a buffer-filled reservoir. High voltage (HV) is applied through platinum electrodes. The high voltage end of the capillary is held in a safety interlock equipped chamber.

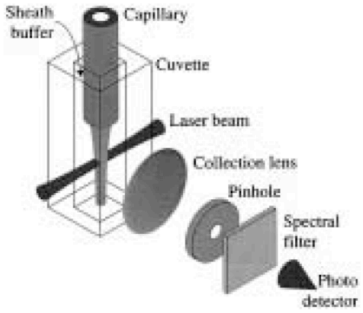


Figure 5. Single-capillary sheath-flow cuvette. A fused-silica capillary is placed inside of a square quartz flow chamber with high optical quality windows. Sheath fluid is pumped in the space between the capillary and the window, and this fluid draws the analyte into a thin stream in the center of the flow chamber. A laser beam is focused beneath the capillary tip on the sample stream. Fluorescence is collected by a lens, spectrally filtered, imaged onto an aperture, and detected with a photomultiplier tube.

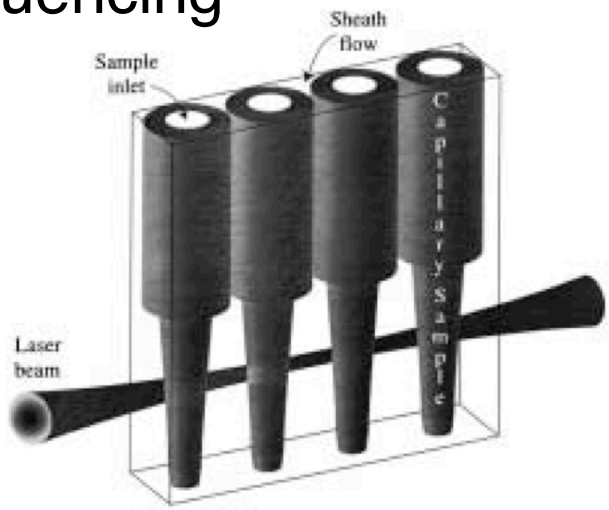


Figure 6. Capillary linear-array sheath-flow cuvette. A linear array of fused-silica capillaries is placed inside of a rectangular glass flow chamber. Sheath fluid draws the analyte into a thin streams in the center of the flow chamber, with one stream produced down-stream from each capillary. A laser beam is focused beneath the capillary tips on the sample streams. Fluorescence is collected by a lens, spectrally filtered, and detected with either an array of photodiodes or a CCD camera.

Suggested reading: "How capillary electrophoresis sequenced the human genome" N.J. Dovichi and J.Z. Zhang, *Angewandte Chemie International Edition* 39, 4463-4468 (2000).

(http://faculty.washington.edu/dovichi/)

- Slab gel electrophoresis was eventually completely replaced by capillary gel electrophoresis for size-based separations of DNA fragments.
- Instrumentation for capillary electrophoresis is quite simple. The capillary holds a sieving medium, which allows separation of the DNA fragments based on their size. The sample is injected into the capillary by dipping the capillary and an electrode into a sample solution and briefly applying electric current through the capillary so that DNA fragments migrate onto the tip of the capillary. Once injection is complete, the sample is replaced with running buffer and electric field is reapplied to drive the DNA fragments through the capillary. A laser-based fluorescence detector near the end of the capillary records the fluorescence signal in four different spectral channels to resolve the fluorescence signature from the four dyes.
- While the performance of capillary electrophoresis is superior to that of conventional slab-gel electrophoresis, a single-capillary instrument does not offer significant advantages over a multi-lane slab-gel system. By separating many samples simultaneously, the slab-gel system produces much more data than is possible from a single-capillary instrument.
- Professor Norm Dovichi of the University of Alberta Chemistry Department made key contributions and overcame this limitation of capillary electrophoresis for DNA sequencing. He developed linear arrays of capillaries that could be excited with a single laser line. This approach led to the development of instruments that could analyze 96 reactions simultaneously.

The Human Genome was first sequenced by multi-capillary electrophoresis

This product is no longer manufactured. more info...

Replacement Product:

Applied Biosystems 3730xl DNA Analyzer

ABI PRISM® 3700 DNA Analyzer

- Production-scale throughput at up to 12 runs per day.
- Walkaway automation frees personnel to focus on other critical activities.
- Fragment analysis and sequencing applications provide system workload flexibility.
- Single platform lowers operating costs

The ABI PRISM® 3700 DNA Analyzer is a fully automated, multi-capillary electrophoresis instrument designed for use in production-scale DNA analysis. It can automatically analyze multiple runs of 96 samples, which enables 24-hour unattended operation.

Science 2001 February 16; 291: 1304-1351

http://www.appliedbiosystems.com

- Sanger and coworkers determined the complete sequence of the 5386 bases in the DNA of a DNA virus in 1977, just a quarter century after Sanger's pioneering elucidation of the amino acid sequence of a protein. This accomplishment is a landmark in molecular biology because it revealed the total information content of a DNA genome.
- This tour de force was followed several years later by the determination of the sequence of human mitochondrial DNA, a double-stranded circular DNA molecule containing 16,569 base pairs. It encodes 2 ribosomal RNAs, 22 transfer RNAs, and 13 proteins.
- In recent years, the complete genomes of free-living organisms have been sequenced. The first such sequence to be completed was that of the bacterium *Haemophilus influenzae*. Its genome comprises 1,830,137 base pairs and encodes approximately 1740 proteins.
- The first eukaryotic genome to be completely sequenced was that of baker's yeast, *Saccharomyces cerevisiae*, which comprises approximately 12 million base pairs, distributed on 16 chromosomes, and encodes more than 6000 proteins. This achievement was followed by the first complete sequencing of the genome of a multicellular organism, the nematode *Caenorhabditis elegans*, which contains nearly 100 million base pairs.
- The human genome is considerably larger at more than 3 billion base pairs, but it has now been completely sequenced. *The ability to determine complete genome sequences has revolutionized biochemistry and biology*.

The race to the < $1000 genome is probably nearing the finish line!

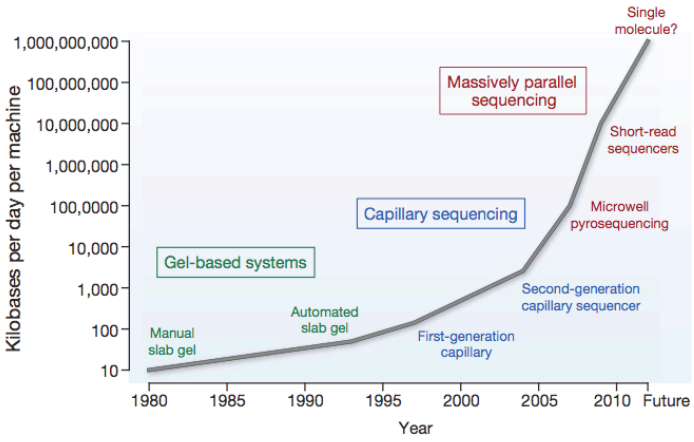
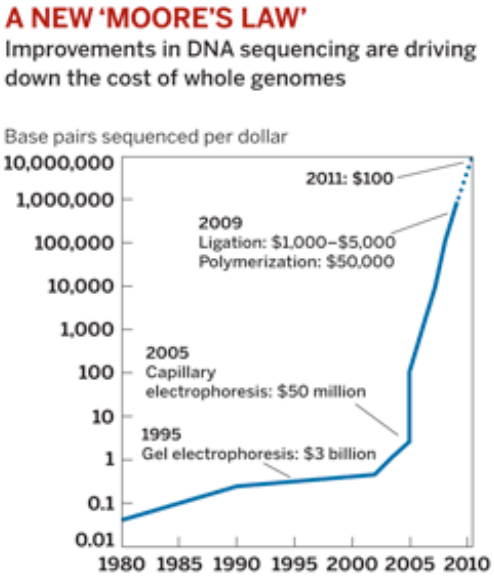


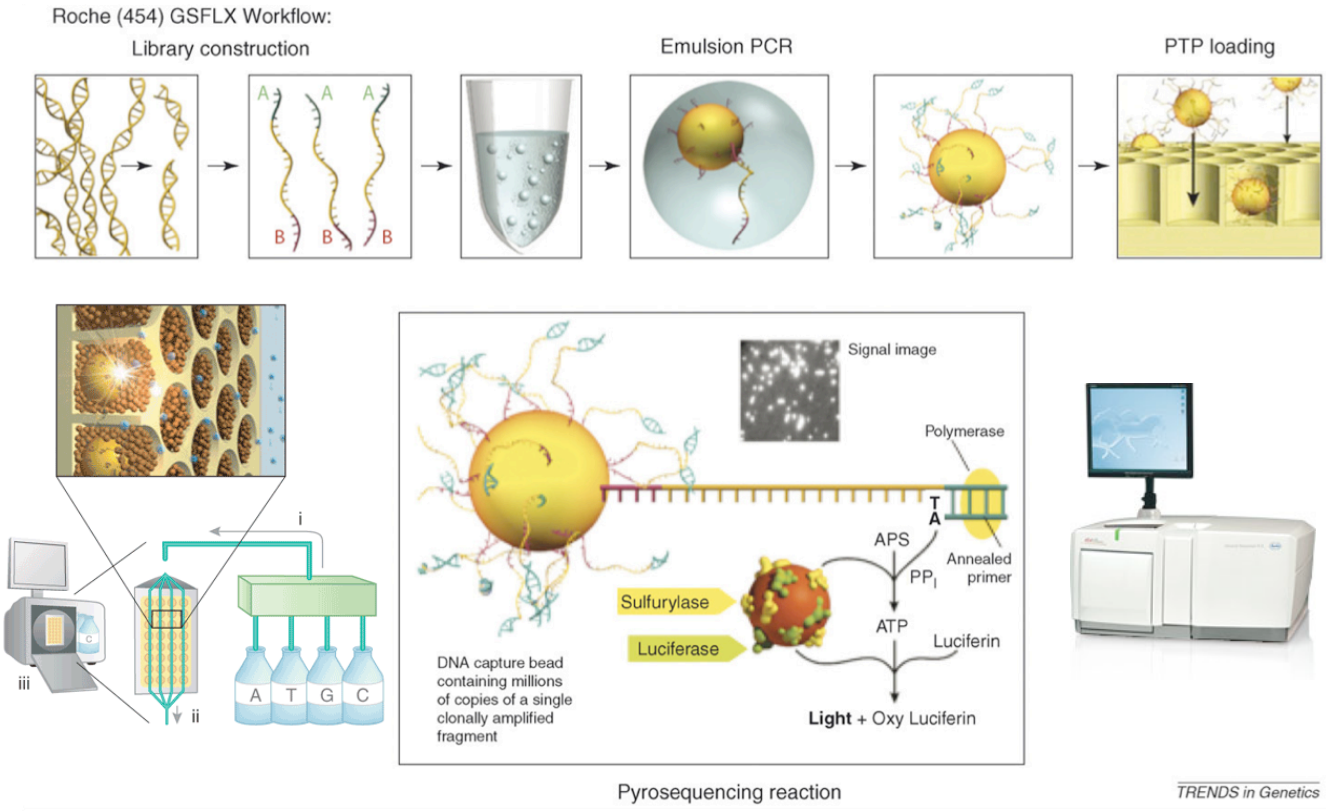
Figure 3 | Improvements in the rate of DNA sequencing over the past 30 years and into the future. From slab gels to capillary sequencing and second-generation sequencing technologies, there has been a more than a million-fold improvement in the rate of sequence generation over this time scale.



Nature **458**, 719-724 (9 April 2009) | doi:10.1038/nature07943

- With the completion of the human genome, a new race began. The finish line of this race is a technology that can sequence a human genome for less than $1000. It is quite clear that traditional dideoxy based sequencing is not a competitor in this race.
- There are now a growing number of new approaches that dramatically decrease the cost of sequencing a human genome. Right now the leaders are somewhere in the range of $30-50k per genome.
- http://pubs.acs.org/cen/coverstory/87/8750cover2.html
- One of the limiting factors is rapidly becoming the overwhelming amount of data that is being generated (http://www.nytimes.com/2011/12/01/business/dna-sequencing-caught-in-deluge-of-data.html?_r=1)

454 (Roche): Sequencing by synthesis with **bioluminescence** detection



- 454 Life Sciences was the first 'Next generation' company to develop and market a DNA sequencing instrument that did not use the traditional dideoxy methodology. 454 Life Sciences relies on sequencing by synthesis. Essentially, the goal is to 'watch' DNA polymerase add nucleotides to a primer that is bound to a template strand. By observing which nucleotides are incorporated into the growing primer, the sequence of the DNA can be read.
- The overall steps are as follows:

  1. Genomic DNA is isolated, fragmented, ligated to adapters (different adapters, A and B, at each end) and separated into single strands.

  2. The DNA fragments are incubated with beads that have previously been covalently coupled with a primer that is complementary to the 'A' adaptor. This is done under relatively dilute conditions so only one DNA fragment is bound to each bead. The beads are isolated and compartmentalized in the droplets of a PCR-reaction-mixture-in-oil emulsion and PCR amplification occurs within each droplet, resulting in beads each carrying ten million copies of a unique DNA template. A primer attached to the bead is complimentary to the A adaptor. A second primer in solution is complementary to the B adaptor.

  3. The emulsion is broken, the DNA strands are denatured, and beads carrying single-stranded DNA templates are deposited into the wells of a picotiter plate that has > 1.5 million wells on it.

  4. Smaller beads carrying immobilized enzymes (sulfurylase and luciferase) required for a solid phase pyrophosphate sequencing reaction are deposited into each well.

  5. The plate is placed into the instrument. The 454 sequencing instrument consists of the following major subsystems: a fluidic assembly (i), a flow cell that includes the well-containing fiber-optic slide (ii), and a CCD camera-based imaging assembly for imaging of the whole picotiter plate.

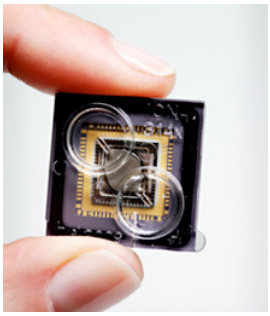
  6. Each of the 4 nucleotides are flowed over the picotiter plates consecutively and for many cycles. If the nucleotide is complementary to the next nucleotide of the template strand, the primer is extended by one nucleotide. This reaction releases pyrophosphate. The enzyme sulfurylase converts adenosine 5´-phosphosulfate (APS) and pyrophosphate into ATP. The enzyme luciferase converts ATP and luciferin into AMP, pyrophosphate, oxy-luciferin and a photon of light.

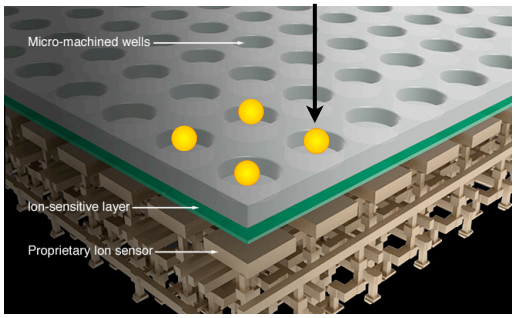
  7. The CCD camera records the location of those wells that produce a flash of bioluminescence when a particular nucleotide is washed over the plate. By repeating this over many cycles, a sequence of about 500 bases can be read out for each bead in a well on the picotiter plate.

  8. These 500 base long fragments of sequence must be computationally reassembled into a complete genome. This process requires that there be many overlapping fragments. Alternatively, the complete genome can be assembled by relying on a reference genome for helping to stitch the fragments together.
- Jonathan M Rothberg & John H Leamon, "The development and impact of 454 sequencing", Nature Biotechnology 26, 1117 - 1124 (2008).
- Elaine R. Mardis, "The impact of next-generation sequencing technology on genetics" Trends in Genetics, 24, 133-141 (2008)

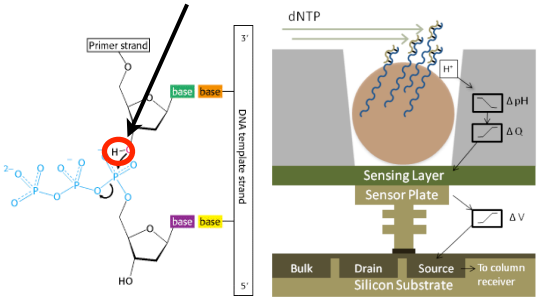
Ion torrent: sequencing by synthesis with **proton** detection

Bead coated with PCR-amplified DNA

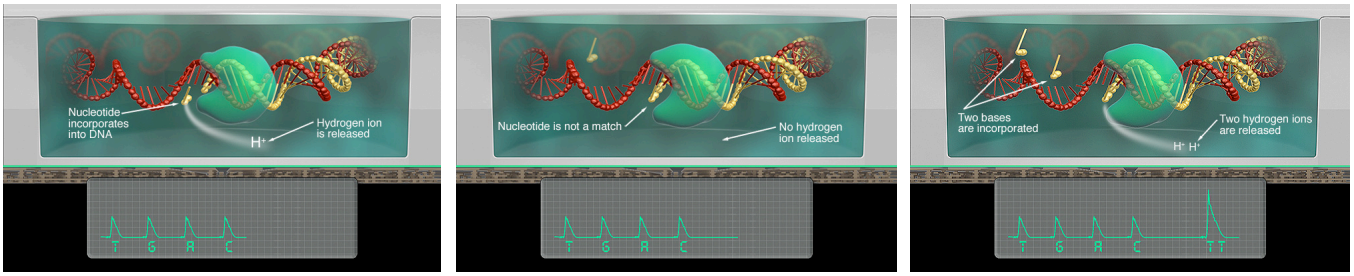




H^+ is released with each reaction



The chip is flooded with one nucleotide at a time.
For example: T then G then A then C then T then G then A etc.



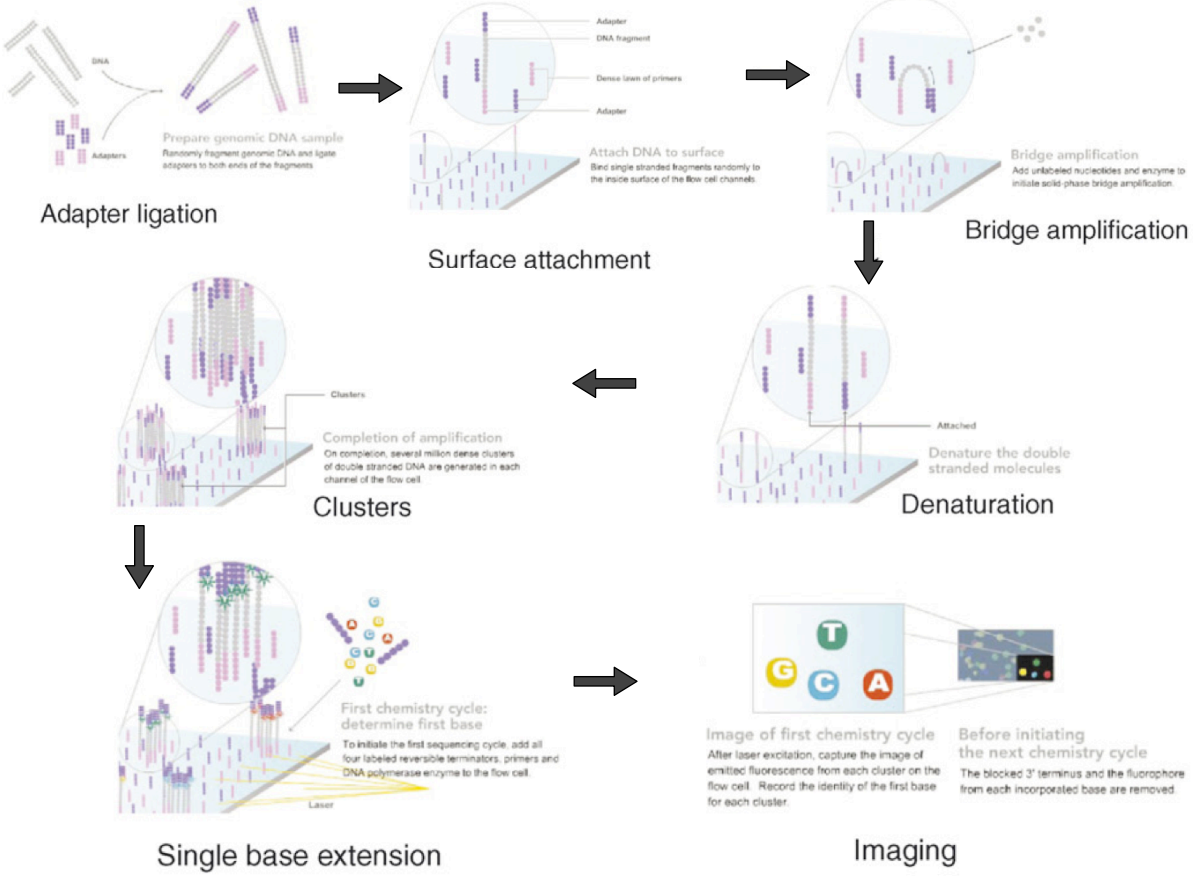
- If a reaction occurs in a particular well, a temporary increase in [H^+] is measured
- If the nucleotide is not a match, no H_+ is released and no increase is measured
- The overall process is very similar to the 454 technology, except [H+] is recorded rather than bioluminesence.

- similar to the 454 system, Ion torrent uses an emulsion PCR based amplification of single molecules of DNA to create beads that are coated with many copies of the same sequence.
- these beads are deposited in microwells (~1.5 million per chip) on a specially fabricated microchip. Under each well is an individual pH sensor that is extremely sensitive.
- this technology is used in the Ion torrent (now Invitrogen Life Sciences) Personal genome machine (PGM) which is aiming to be the lowest cost 2nd generation sequencer.
- In July 2011, this company published a paper on this technology in Nature (An integrated semiconductor device enabling non-optical genome sequencing, Nature 475, 348–352 (2011)). In this paper they describe the full sequencing of the Intel co-founder Gordon Moore (i.e., Moore's Law). An article about the Nature paper in the New York Times includes this statement:

"*George Church, a genome technologist at the Harvard Medical School, said he estimated the cost to sequence Dr. Moore's genome at $2 million. This is an improvement on the $5.7 million it cost in 2008 to sequence Dr. Watson's genome on the 454 machine, but not nearly as good as the $3,700 spent by Complete Genomics to sequence Dr. Church's genome and others in 2009. Dr. Rothberg [the inventor of this technology] said he had already reduced the price of his chips to $99 from $250, and today could sequence Dr. Moore's genome for around $200,000. Because of Moore's Law — that the number of transistors placeable on a chip doubles about every two years — further reductions in the cost of the DNA sequencing chip are inevitable, Dr. Rothberg said." (http://www.nytimes.com/2011/07/21/science/21genome.html)*

- http://www.wired.com/wiredscience/2011/07/how-accurate-is-the-new-ion-torrent-genome-really/

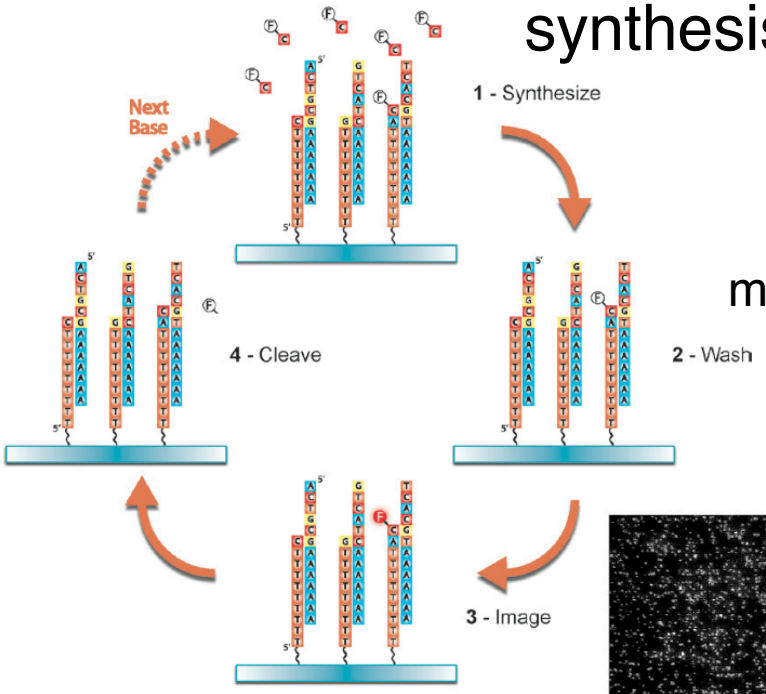
Illumina: Sequencing by synthesis with **fluorescence** detection



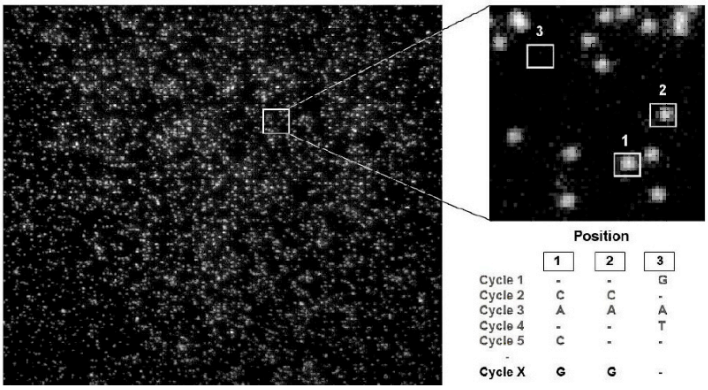
- Illumina is a San Diego Based company that uses a related sequencing by synthesis technology. They sell instruments for sequencing and also offer a personal genome sequencing service for ~$50,000.
- The two major differences between Illumina and the 454 technology are listed here:
- 1. Illumina does not immobilize DNA on beads. Instead they use a surface on which both primers are immobilized (primers complementary to the adapters at both ends of the fragments). These immobolized primers are used for 'bridge amplification' which produces a small cluster of PCR products attached to the surface in the vicinity of where a single original template was bound.
- 2. Instead of using bioluminescence, Illumina uses fluorescence. The nucleotides that are washed over the plate have a fluorophore attached via the 3' hydroxyl. Each type of nucleotide (ATCG) has a specific color of fluorophore attached to it. Only one type of nucleotide can be added at each cluster - the nucleotide that happens to be complementary to the next nucleotide of the template. This means that each cluster will incorporate a certain color of fluorophore that can be imaged. The color reveals the identity. Following imaging, the 3' hydroxyl is deprotected and the fluorophore is removed. This sets up the primer for the next cycle of washing over the nucleotides. Read lengths are limited to about 35-50 bases.

- The read lengths in sequencing by synthesis are relatively short. There are probably several contribution reasons. One is the loss of signal due to incomplete removal of the protecting group after each round of dNTP addition. This might also cause a portion of the growing primers to get out of frame. Another reason might be the accumulation of photo-induced damage of the DNA template. These limiting factors are all technical details. If the system worked perfectly, read lengths could continue indefinitely.

Helicos: **Single molecule** sequencing by synthesis

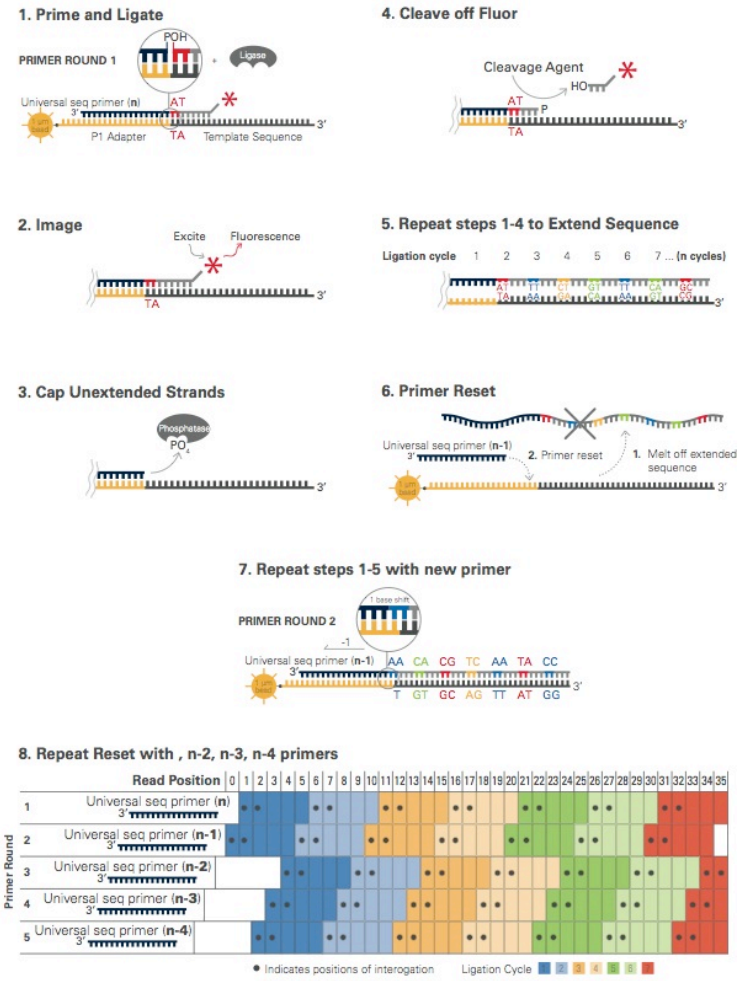


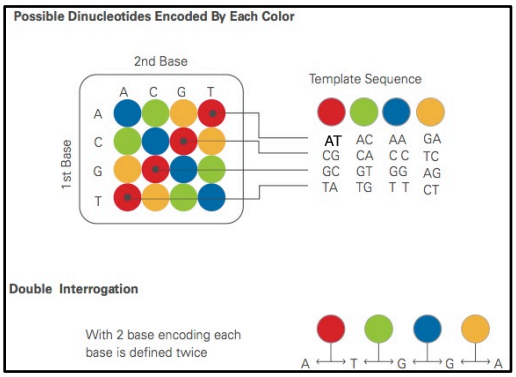
Advantages: no amplification step and many more ‘spots’ per run



- Helicos Biosciences is a Cambridge Massachusetts-based company that is pursuing true single molecule sequencing. Essentially their approach is identical to that Illumina, but they don’t bother with the bridge amplification step. Each ‘spot’ that they image is a single dye molecule that is linked to a single nucleotide and incorporated into a single primer strand on a single template strand.
- Original DNA samples are first fragmented, the DNA double-helix is melted into single strands and a polyA tail is added to these DNA molecules. Billions of these single DNA molecules are captured on a proprietary surface within a flow cell and serve as templates for the sequencing-by-synthesis process.
- Fluorescently-labeled nucleotides are added one at a time and incorporated into the growing complementary strand by a DNA polymerase enzyme, based on the sequence of the template. Unused nucleotides are washed away. Upon illumination with a laser, the incorporated nucleotides emit light that is detected by the HeliScope Single Molecule Sequencer. The HeliScope Sequencer captures thousands of images across the flow cell surface to record which strands incorporated which nucleotides. These images containing tens of thousands of single flurescent molecules are akin to star fields, as pictured at right. Once the imaging of the flow cell is complete, the fluorescent label is removed before the next nucleotide is added to continue the cycle. Tracking nucleotide incorporation on each strand determines the exact sequence of each individual DNA molecule. [From the Helicos tSMS Technology Primer at www.helicosbio.com]
- They currently claim that they can get > 30 gigabases (billions of bases) of sequence information per run. This number is substantially larger than a human genome, though there does have to be a high degree of redundancy in order to assemble the short fragments of sequence (~35 bases each) into a complete genome. They have also reported that they have sequenced a human genome (Stephen Quake’s) for under $50,000. The main drawback of this approach are the relatively high error rates.

The SOLiD System from Applied Biosystems: Sequencing by **ligation**





Another example: Compete Genomics is a whole genome sequencing company that uses a *proprietary sequencing by ligation method*. They don't sell instruments, and are aiming at providing full genome sequencing for $5000 (and ultimately $1000).

- Applied Biosystems (now part of Life Technologies after merging with Invitrogen), has taken a different strategy for sequencing by synthesis. They don't use a DNA polymerase at all, but rather use a DNA ligase in a technique known as SOLiD. I will freely admit that this technique is very hard to understand. However, it is one of the leading technologies in the race for the $1000 genome. One of the main advantages of this technique is the very low error rate. A disadvantage are the short reads (35 bases).
- Sample preparation is similar to the 454 approach, but SOLiD uses a mixture of labeled oligonucleotides and queries the input strand with ligase (not a polymerase). Understanding the labeled oligo mixture is key to understanding SOLiD technology.
- The key reagents in SOLiD technique are 16 different fluorescently labeled oligonucleotides. Each of the 16 different oligos (8 mers) is fully degenerate (all N's) at all positions other than the two nucleotides closest to the 3' end. At the two positions closest to the 3' end, there is one of the 16 possible dinucleotide sequences. Between nucleotides 5 and 6, there is a chemically cleavable linkage that mimics the phosphodiester bond. At the 5' end of the oligo is attached one of 4 fluorescent dyes. This is confusing for many people - why are there 16 different oligonucleotides but only 4 different colors of dye? The answer is below.
- The sequencing involves:

1. Anneal a primer, then hybridize and ligate the mixture of the 16 different fluorescent oligos (8-mers). Only one of the 16 oligos will be a perfect match and ligate. This will introduce one of the 4 fluorescent colors into the cluster of templates on the surface of the chip. Some of the probes will almost certainly bind to other locations along the template. However, they will not be properly positioned to ligate with the primer. Following ligation the probes that are non-specifically hybridized can be washed away, leaving only those that are covalently linked to the primer.
2. Image the fluorescence on the surface to reveal which fluorophore has become associated with each spot on the chip. Observation of a certain color of fluorescence tells you that the two nucleotides of the template must be one of 4 possibilities (see 'Possible dinucleotides encoded by each color')
3. Chemical cleavage of the three 5' bases removes the fluorophore and leaves behind a 5 base ligated probe, with a 5' phosphate
4. Repeat, this time querying the 6th & 7th bases of the template
5. After 5-7 cycles of this, perform a "reset", in which the initial primer and all ligated portions are melted from the template and discarded.
6. Next a new initial primer is used that is N-1 in length. Repeating the initial cycling (steps 1-5) now generates an overlapping data set (bases 1/2, 6/7, etc.).

- Thus, 5-7 ligation reactions followed by 5 primer reset cycles are repeated generating sequence data for ~35 contiguous bases, in which each base has been queried by two different oligonucleotides.
- What's going on with the 16 possible dinucleotides (4^2) only being encoded using 4 fluorescent dyes...? Detection of a certain color of fluorescence at a particular spot does not tell you what base is at a given position. This is where the brilliance (and potential confusion) comes about with regard to SOLiD. There are 4 oligos for every dye, meaning there are four dinucleotides that are encoded by each dye. For example, the dinucleotides CA, AC, TG, and GT are all encoded by the green dye. Because each base is queried twice it is possible, using the two colors, to determine which bases were at which positions (see 'Double Interrogation'). [SOLiD_Brochure.pdf (http://marketing.appliedbiosystems.com]
- It seems that you need even a series of 3 or more consecutive colours is not enough information to unambiguously assign the sequence. You need a reference nucleotide that you know the identity of and which let's you assign the remainder of the colour coded sequence. This reference nucleotide is the one in position '0' (refer to step 8 on slide 343). Notice that this is part of the adapter sequence that the Universal primer anneals to. In the second round (when using the n-1 primer), the first read is of positions 0 and 1. Since you already know the identity of the nucleotide at position 0 (as it is part of the adapter), the colour of the primer that is ligated in the first ligation cycle tells you the exact identity of the nucleotide in position 1. From there you should be able to figure out the remainder of the nucleotide sequence.

The race for the lowest cost sequencers: Ion torrent Ion PGM vs. Illumina MiSeq

Ion Torrent (now Life Technologies)
Ion Personal Genome Machine (PGM) ~
$50k
100 bp read length
$500 consumables/run
75,000,000 bp/run

Illumina
MiSeq
~$125k
35 - 150 bp read length
$500+ consumables/run
500,000,000 to 5,000,000,000 bp/run

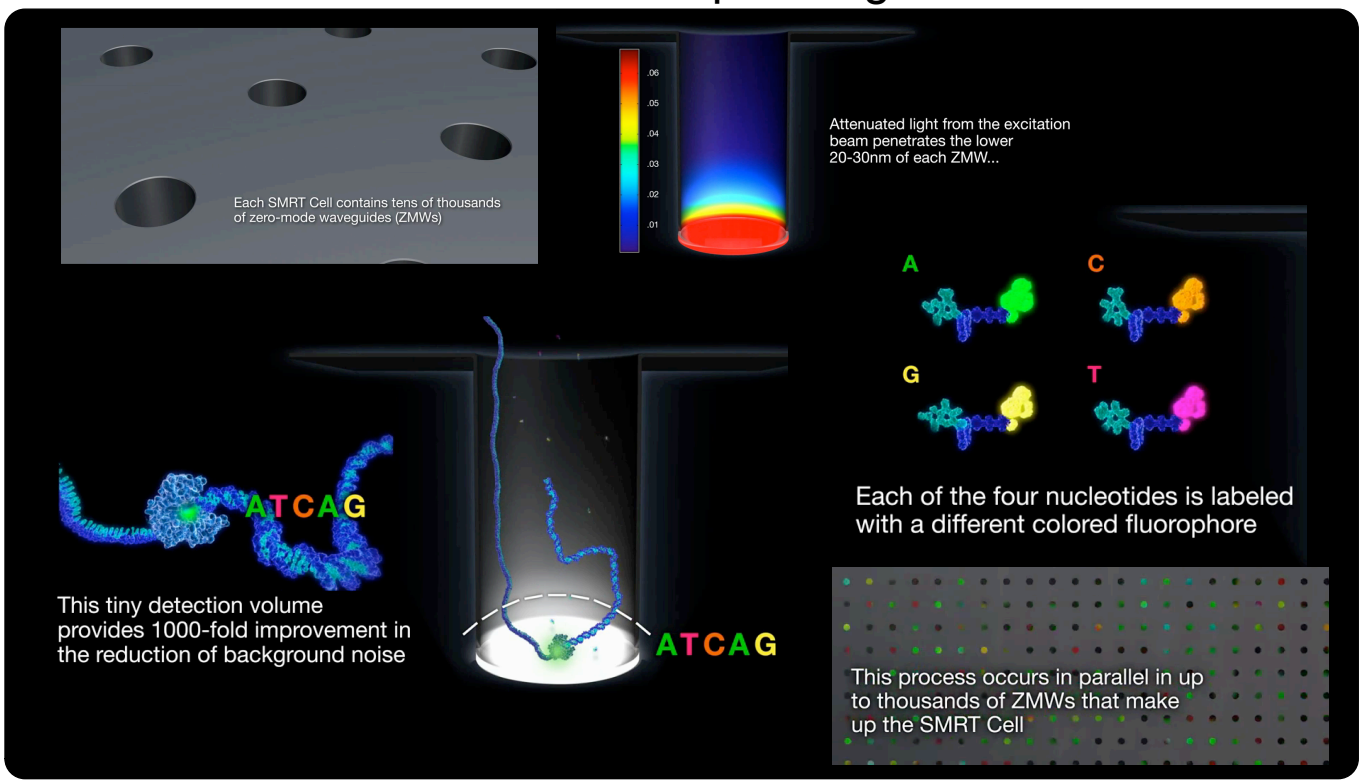
- http://www.illumina.com/systems/miseq.ilmn
- http://www.iontorrent.com/

- These two companies are currently in a battle to prove that they have the best technology for genome sequencing. It's hard to tell who is winning, but it seems as though Illumina has a bit of an advantage right now. The Ion torrent instrument seems to be limited by a higher error rate (~99% for IT vs. 99.9% for MiSeq) that decreases the quality of the sequence and requires greater sequencing redundancy in order to achieve a high quality genome.

- There are even attack ads! (http://www.youtube.com/watch?feature=player_embedded&v=GUr17pHezUo)

Third generation sequencing technologies are on (maybe just over) the horizon

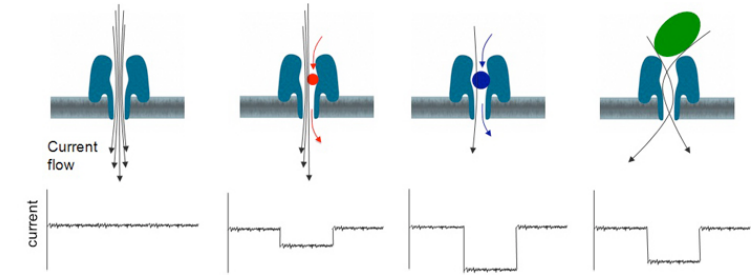
1. Pacific Biosciences: Sequencing in **nanowells**

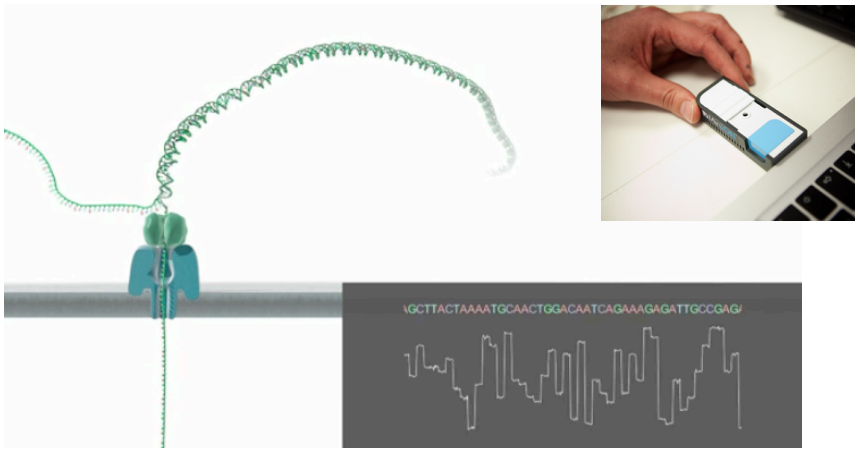


- http://www.pacificbiosciences.com/
- Pacific Biosciences is currently providing a limited number of select customers with instruments for single molecule DNA sequencing. The instrument is not yet widely available for general purchase. They call their technology SMRT (single molecule, real-time) sequencing. It is based on the detection of the fluorescence of single molecules of fluorescently labeled nucleotides as they bind to the polymerase and are incorporated into the growing primer strand.
- Images from: http://www.youtube.com/watch?v=NHCJ8PtYCFc&list=UU2y78sjVOumGc2da1tN629g&index=4&feature=plcp

Third generation sequencing technologies are on (maybe just over) the horizon

2. Oxford Nanopore: Sequencing with **nanopores**





- http://www.nanoporetech.com/home

- Oxford Nanopore issued a press release on Feb. 17, 2012, in which they claimed that they will ship the MinION USB stick-type DNA sequencer. They claim that it will be disposable and cost ~$900 per device. They will also ship a larger device known as the GridION
- The big advantage of nanopore sequencing is the essentially unlimited read lengths. The company has demonstrated read lengths of 10,000 bases and claims that it could go as high as 100,000,000.

DNA sequencing

- Short stretches of DNA, commonly known as oligos or primers, can be synthesized by solid phase methods.
- Long stretches of DNA (~500 to 1000 bases) can be sequenced through the use of DNA polymerases and dideoxy nucleotides that lack a 3' hydroxyl.
- Occasional incorporation of the dideoxy nucleotide results in termination of the DNA replication. Submitting the reaction mixture to analysis by electrophoresis results in a 'ladder' of truncated DNA sequences.
- Fluorescent labels have completely replaced radioactivity based DNA detection.
- Norm Dovichi contributed to the the technology of capillary gel electrophoresis for sequencing of DNA. His contribution was an important step towards the sequencing of the human genome.
- There is currently a race to develop the technology that will enable a human genome to be synthesized for less than $1000. Traditional dideoxy sequencing is not a serious competitor in this race!
- The leader in the race are techniques that are collectively known as 'sequencing by synthesis'. This can be now be done on the single molecule scale.