GENETIC SEQUENCING

By C. Kohn

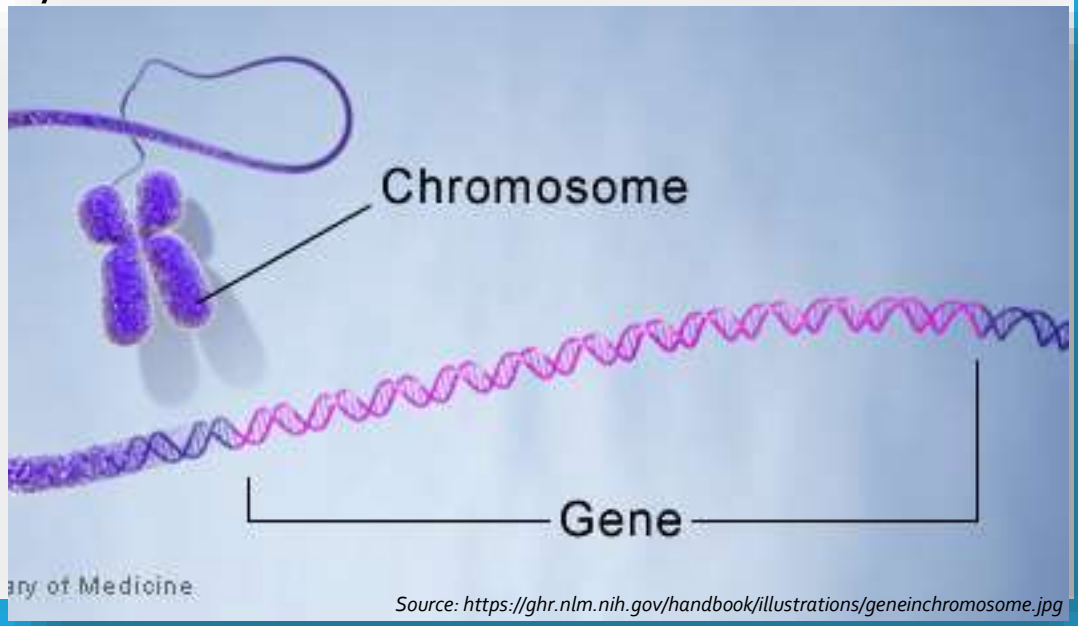
Agricultural Sciences

Waterford, WI

Source: www.pointersoft.in

DNA Sequencing

- **Gene sequencing is a process in which the order of nucleotide bases (A, G, C, and T's) is determined in a sample of DNA.**
  - Gene sequencing is when we "read" the letters of DNA to determine the genetic code that makes the genes of an organism.
- **A gene is a stretch of DNA that codes for how to assemble a specific protein for a specific trait.**
  - For example, your gene for eye color is the stretch of your DNA that indicates how the amino acids need to be assembled to create the protein that gives you your eye color.
- **A genome is all of the DNA that is found in an individual organism.**
  - A genome contains the complete set of genes found in the cells of an organism.



Source: https://ghr.nlm.nih.gov/handbook/illustrations/geneinchromosome.jpg

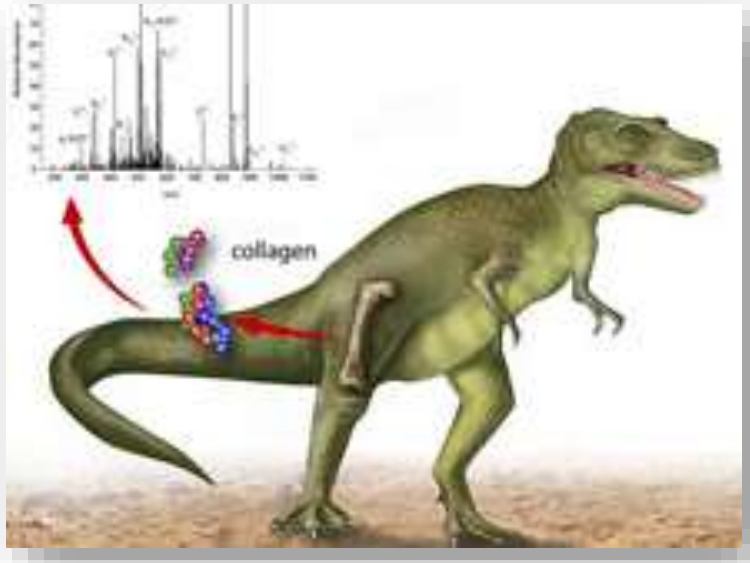
Applications of DNA Sequencing

- **The ability to read the DNA of different organisms allows researchers to identify the exact order of nucleotide bases that are used to assemble the proteins that are responsible for specific traits.**
  - For example, gene sequencing can tell a researcher if an individual possesses mutated or defective genes that could lead to a genetic disorder such as cancer, cystic fibrosis, or sickle cell anemia.
- **Gene sequencing also might help researchers improve agriculture.**
  - For example, genetic sequencing can help researchers understand which genes in cattle are responsible for excellent milk or meat production.
  - Genetic sequencing of plants with valuable traits (such as drought tolerance or a rapid growth) can enable scientists to determine which genes are responsible for those traits, enabling improvements to valuable crops like corn, wheat, or soybeans.

Source: www.illumina.com

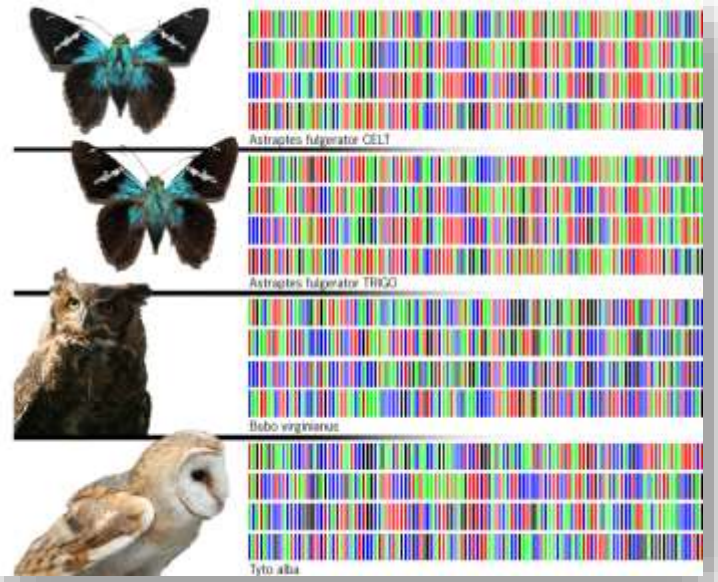
Applications of DNA Sequencing

- **Researchers could also use gene sequencing to determine the evolutionary origins of a given species.**
  - For example, genetic sequencing of tissue from the Tyrannosaurus rex has shown that modern-day chickens are genetically similar these dinosaurs.



Source: www.nsf.gov

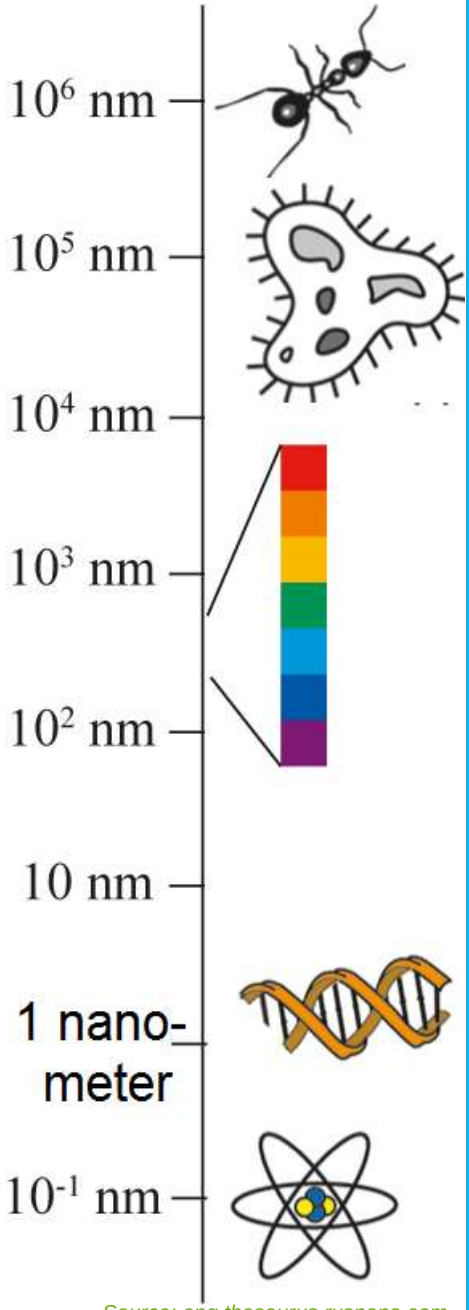
- **Genetic sequencing also allows researchers to better protect the environment.**
  - For example, ecologists can use genetic sequencing to determine the health of an ecosystem just by sampling for DNA to determine the level of biodiversity (instead of having to individually find and physically identify each different kind of species in that habitat).



Source: www.councilforresponsiblegenetics.org

Difficulty of DNA Sequencing

- **Until recently, DNA sequencing was very costly, difficult, and time consuming.**
  - What had made genetic sequencing exceptionally difficult is that DNA is EXTREMELY small.
- **DNA is only 2 nanometers wide; a <u>nanometer</u> is one billionth of a meter.**
  - In comparison, a human hair is 75,000 nanometers wide.
  - A wavelength of visible light is 400 nanometers wide, meaning that DNA hundreds of times smaller than a wavelength of visible light!
- **DNA is so small that we could never see its molecular structure with even the most powerful light microscope in the world.**
  - To read something that cannot be seen, scientists have to use specific tools that allow them to indirectly read each letter in a gene or a genome without actually seeing the DNA itself.

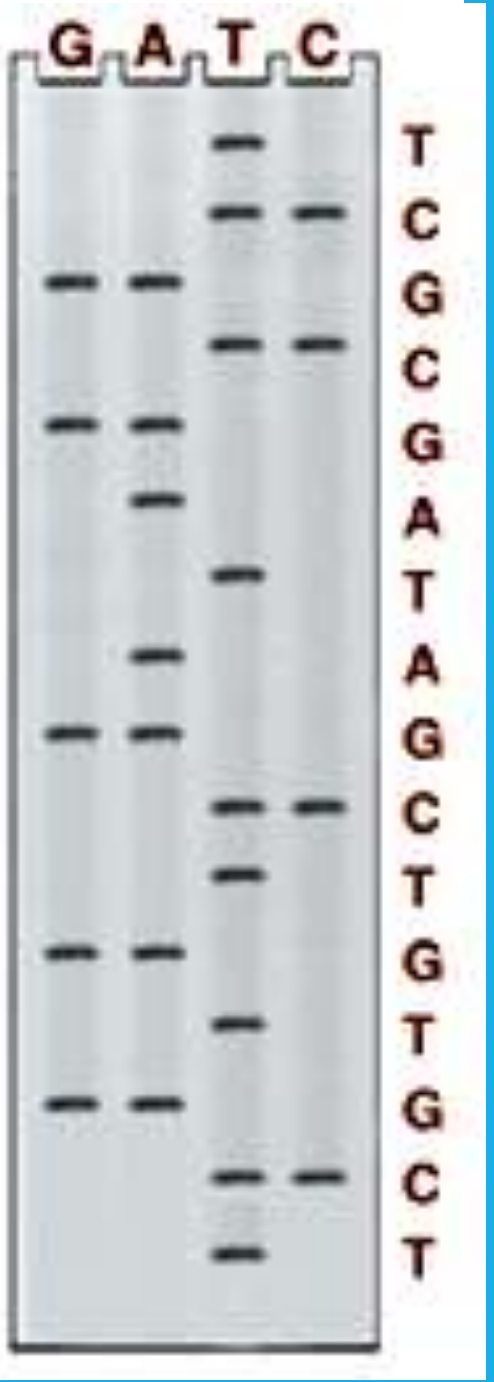


Source: eng.thesaurus.rusnano.com

Genetic Sequencing Techniques

- **Some of the most widely-used techniques to read the nucleotide bases in a sample of DNA include:**
  - The Sanger Sequencing Method
  - 454-Roche Pyrosequencing
  - Illumina Bridge Sequencing
  - Ion-Torrent Sequencing
  - Nanopore Sequencing.

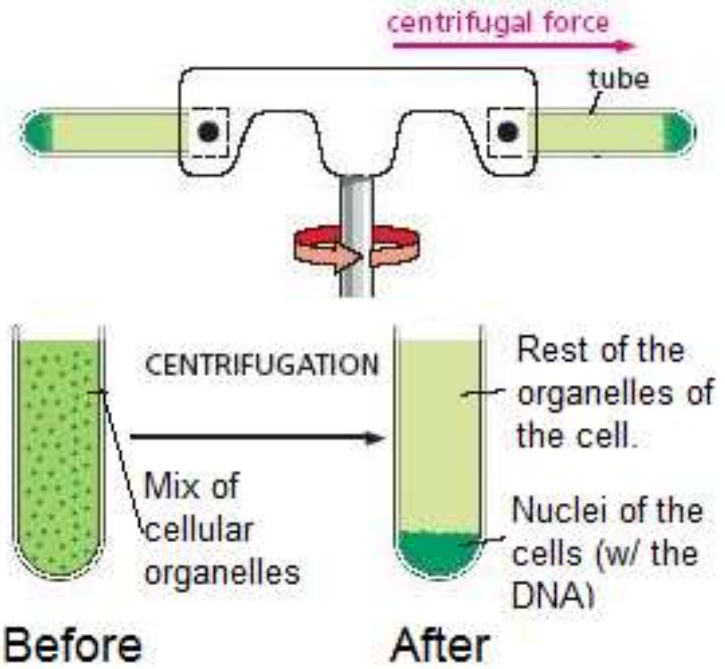
- **Each of these techniques works in a different way to read the nucleotide bases in a sample of DNA.**
  - However, all provide the same end result: they all determine the order of G's, C's, A's, and T's that make up any sample of DNA.



Source: www.nationaldiagnostics.com

Acquiring the DNA

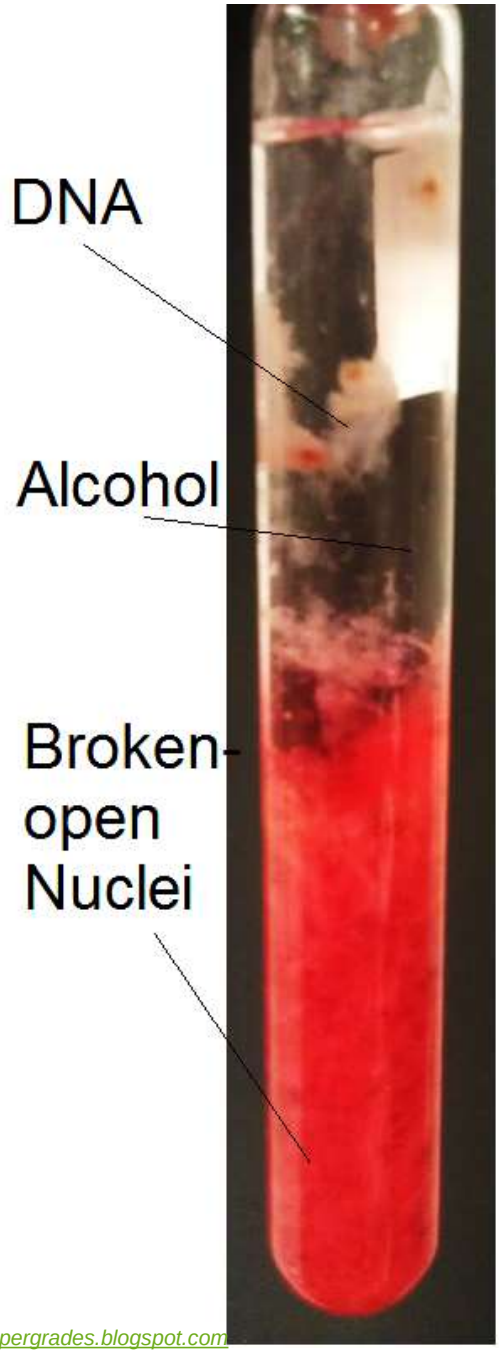
- **Regardless of the method that is used to read the order of nucleotide bases in a sample of DNA, the first step is always to get a pure sample of the DNA that we are interested in sequencing.**
  - To do this, scientists first collect cells from the species they intend to study.
  - The membranes of these cells are then broken apart using a detergent (similar to what you would find in dish soap).
  - Because the membranes of cells are made of fatty lipids, the same detergent that you use to get grease off of a frying pan could work to break apart the lipid cellular membranes that protect the DNA.
- **Once the cellular membranes are broken apart, the cellular mixture is <u>centrifuged</u> (or spun rapidly).**
  - This forces the heaviest parts of the cells to move to the bottom and the lightest parts of the cells move to the top of the solution.



Source: cellbiologyolm.stevegallik.org

Acquiring the DNA

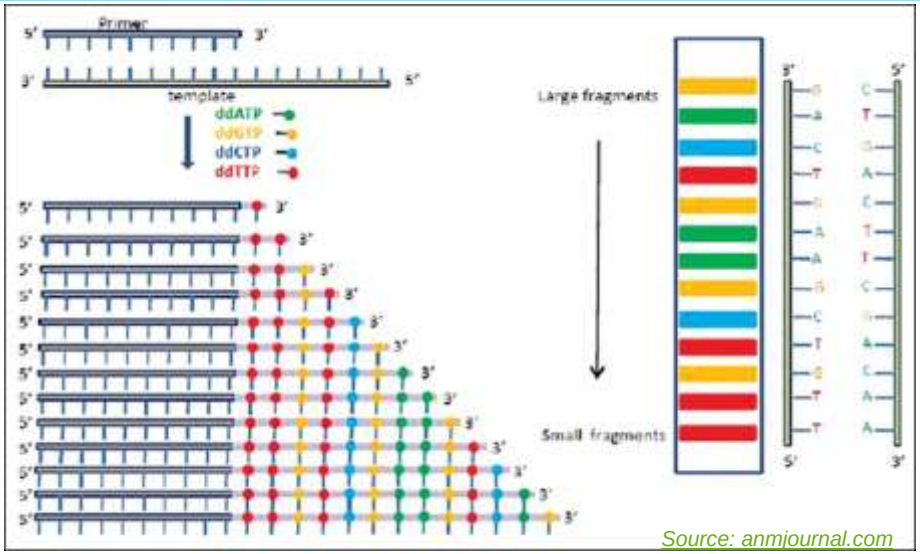
- **Centrifuging the sample of cells separates each of the different kinds of cellular organelles from each other so that the nuclei containing the DNA can be removed.**
  - Scientists can then extract the nuclei and break them open using the same process as before was used to break open the cell.
  - The nuclei of cells have the same kind of membrane as the cell itself.
- **Once the nuclei have been extracted and broken apart, the DNA can then be separated from the rest of the contents of the nuclei using cold alcohol.**
  - The DNA does not dissolve in the alcohol like the rest of the cell's contents.
  - Because of this, the addition of alcohol allows the researcher to extract the DNA without getting any other cellular materials.



Source creativeuppergrades.blogspot.com

SANGER SEQUENCING

The original method of sequencing DNA.

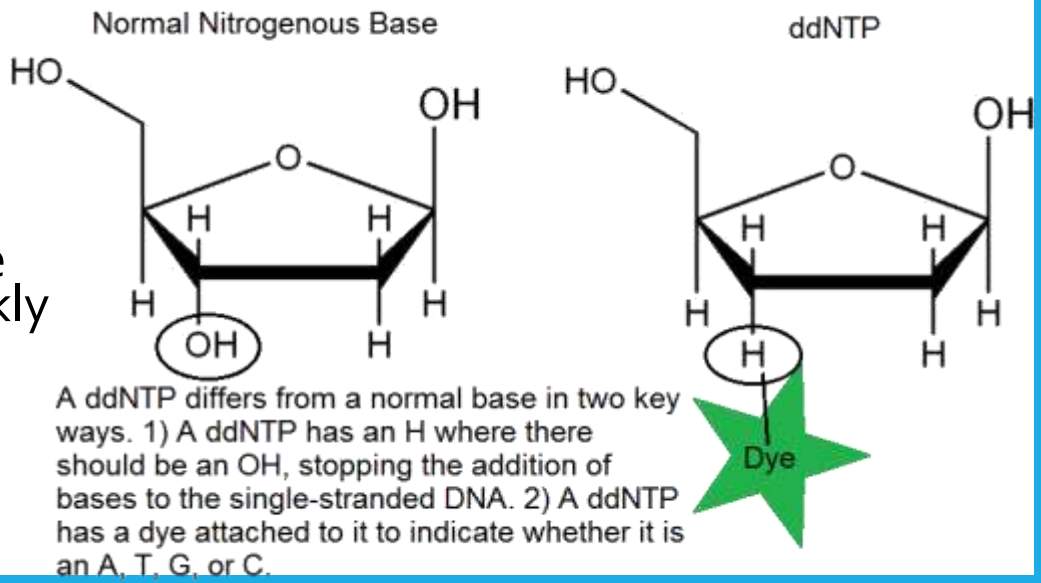


Source: anmjournal.com

Sanger Sequencing

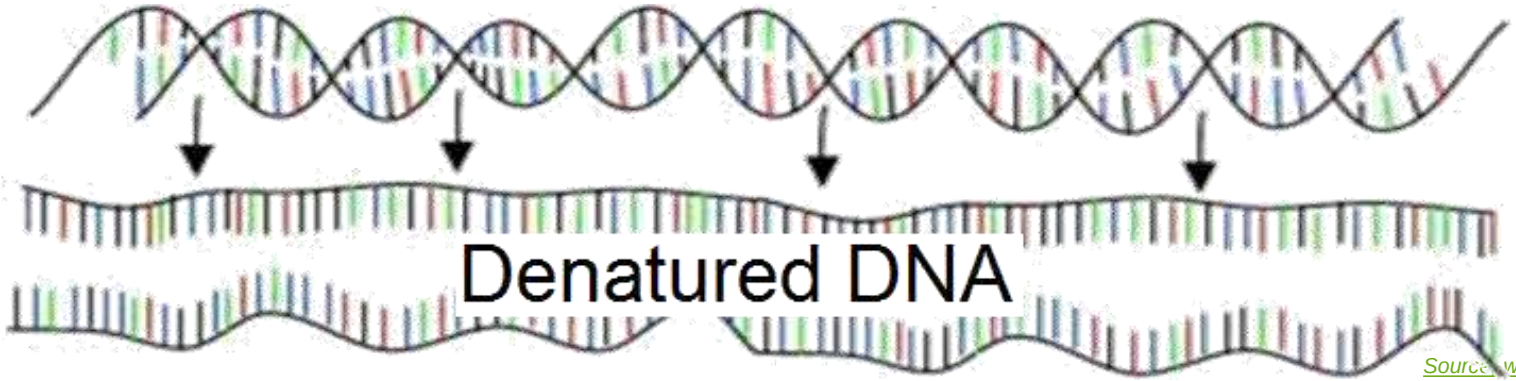
- **The Sanger Method of Sequencing was one of the original ways of reading each nucleotide base in a strand of DNA letter by letter.**
  - The Sanger Method is able to do this by using specially-made nucleotide bases.
  - These artificial bases, called ddNTP's, are dyed a specific color for each of the specific bases (A, G, C, and T) and stop the addition of additional bases.

- **To begin the Sanger method of DNA Sequencing, the strand of DNA we want to read must be copied millions of times.**
  - The DNA is copied over and over using bacteria.
  - The DNA to be sequenced is added to the bacterial cell, and the bacteria reproduce the DNA every time the bacterial cell divides.
  - Because bacteria can reproduce very quickly, they can very quickly (and efficiently) increase the number of copies of the strand DNA that we want to read.



Denaturing DNA

- **Once enough copies of the DNA fragment have been made by bacterial cell division, the DNA is removed from the bacterial cell.**
  - The DNA has to be purified so that we only have the kind of DNA that we are interested in sequencing.
- **Next, the DNA has to be <u>denatured</u>.**
  - This means that the DNA is changed from double-stranded to single-stranded.
  - This is necessary so that the dyed artificial bases can be added to fill in the other side of the DNA.
  - Polymerase will be used to 'fill in' the other side of the single stranded DNA (just as in ordinary DNA replication).

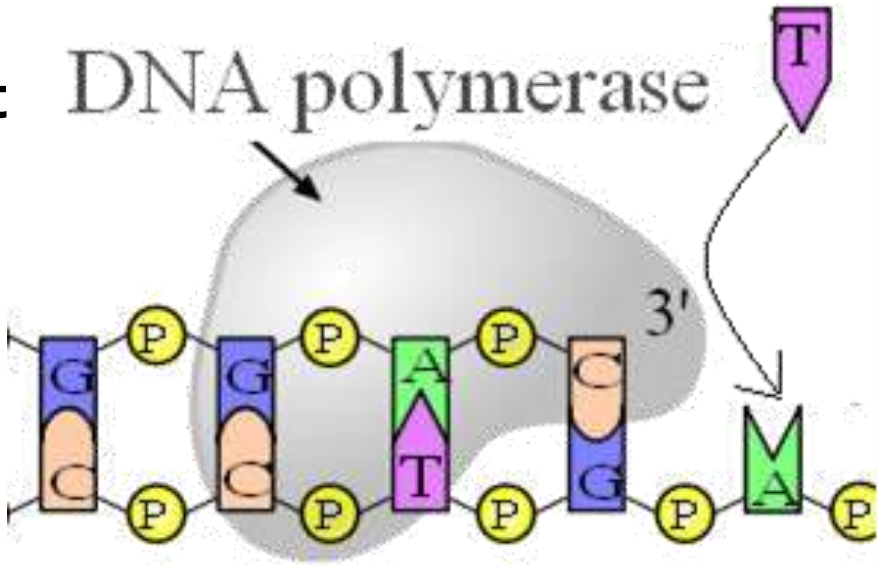


Source: www.btci.org

Components of the Sanger Method

- **Once we have a pure sample of single-stranded (or denatured) DNA that has been copied millions of times, we have to add four things to make the Sanger Method work:**
  - 1) Polymerase – the enzyme that makes copies of DNA.
  - 2) Primers – these tell polymerase where to start copying the DNA.
  - 3) Nucleotide Bases – A, T, G, and C.
  - 4) ddNTP's – artificial nucleotide bases that give off a color specific to which letter it represents.

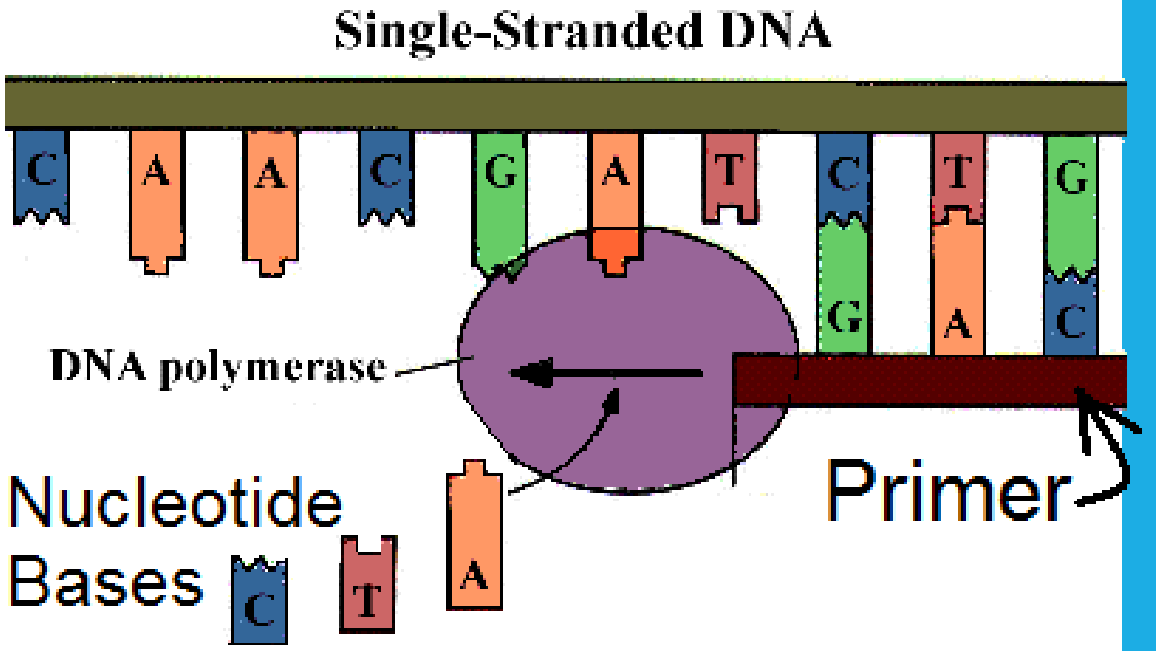
- **Polymerase is the enzyme that makes copies of DNA (the same polymerase that is used to make additional copies of DNA or RNA).**
  - Polymerase is what "fills in" the other side of the single stranded DNA to make it double stranded.



Source: researchguides.library.vanderbilt.edu

Components of the Sanger Method

- **The primer tells the polymerase where to start adding bases to the single stranded DNA.**
  - A primer guides polymerase to where it needs to be just like runway lights guide an airplane to its runway.

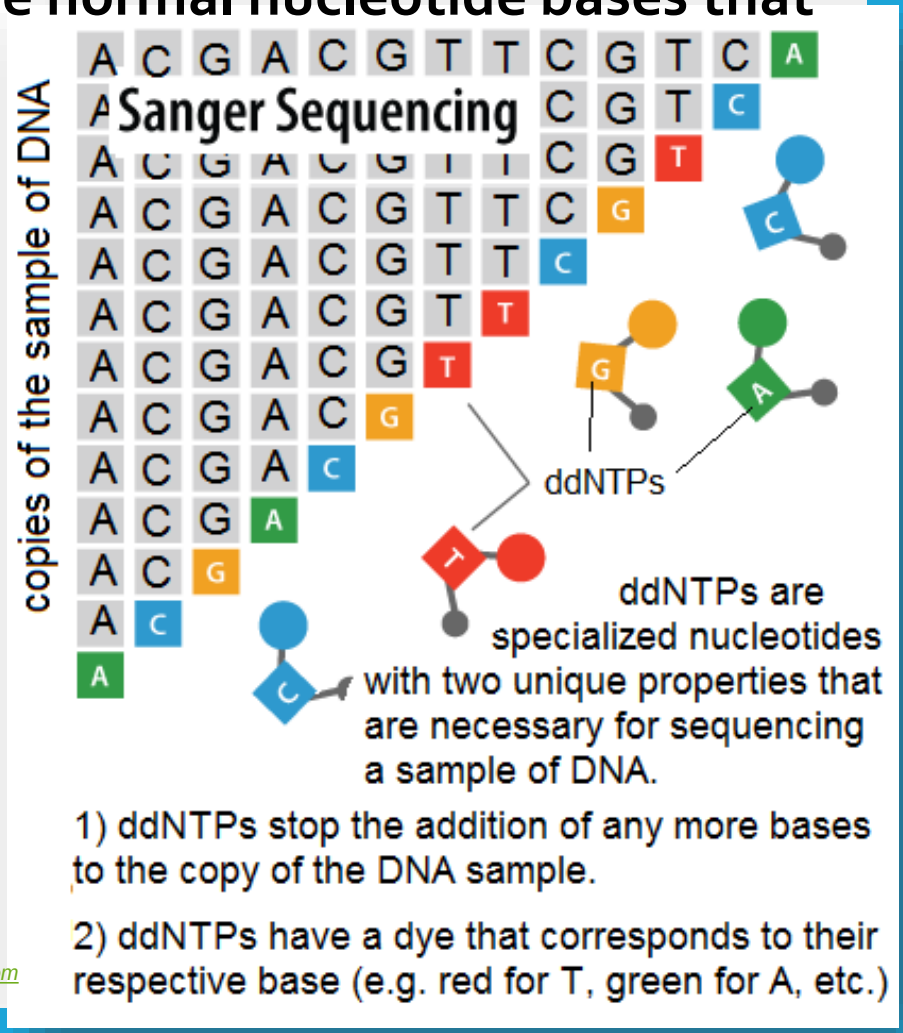


Source; www.gatewaycoalition.org

- **Nucleotide bases are what the polymerase enzyme adds to "rebuild" the other side of the single-stranded DNA.**
  - These are the A, G, T, and C "letters" that make up DNA.
  - For example, if the single stranded DNA had the sequence **G-G-T-T-C-C-A-A**, the polymerase enzyme would add the following nucleotide bases to fill in the other side: **C-C-A-A-G-G-T-T**

Components of the Sanger Method

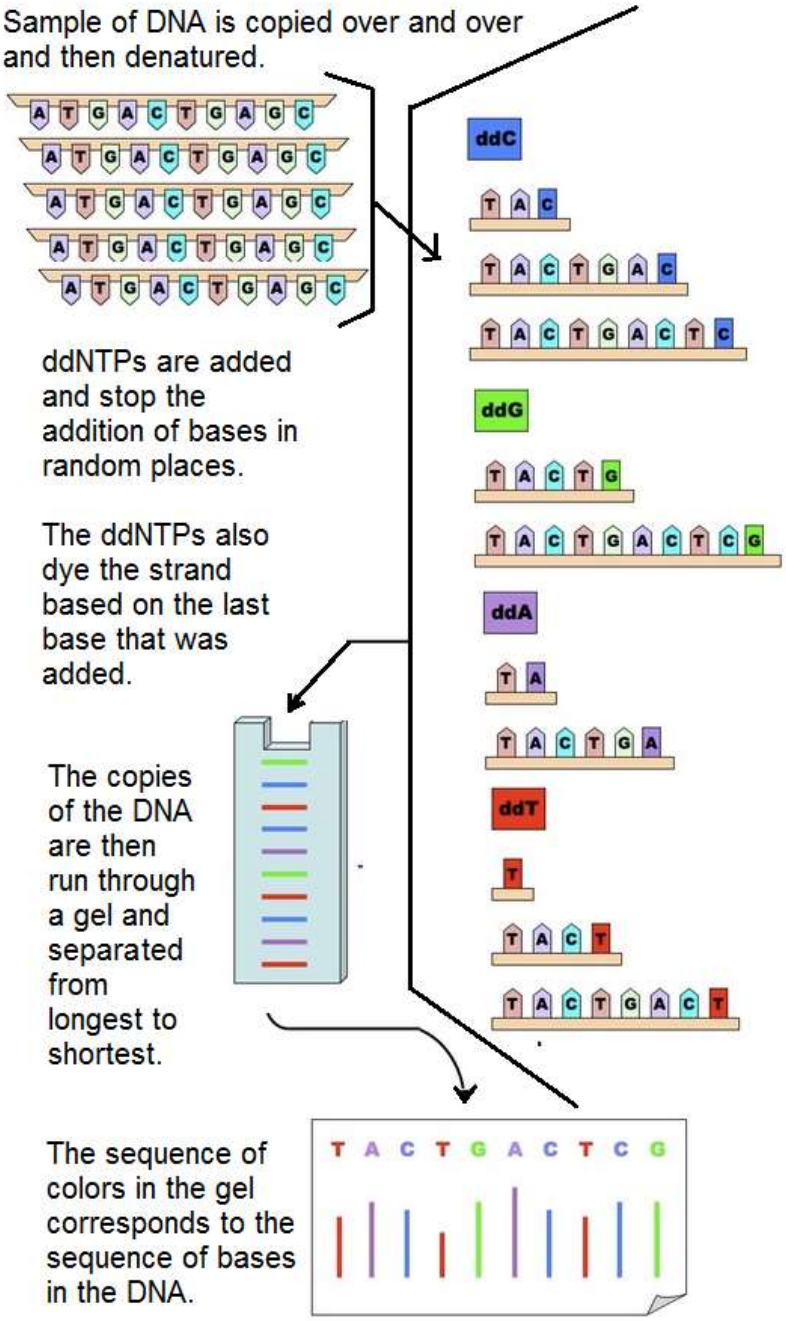
- **ddNTP's are modified artificial nucleotide bases.**
  - Like regular nucleotide bases, there are four kinds of ddNTP's: A, G, C, and T.
- **ddNTP's are almost identical to the normal nucleotide bases that are found in DNA except for two main differences:**
  - 1) ddNTP's have colored molecular "dye" attached to them; this dye gives off a specific color when excited by a laser.
  - 2) ddNTP's prevent the addition of any other nucleotide bases once they have been added to the single stranded DNA.
- **The ddNTP's are the most important part of the Sanger Method.**
  - It is precisely the way that ddNTP's are different from regular nucleotide bases that enables the Sanger Method to work.



Source: www.abmgood.com

ddNTP Differences

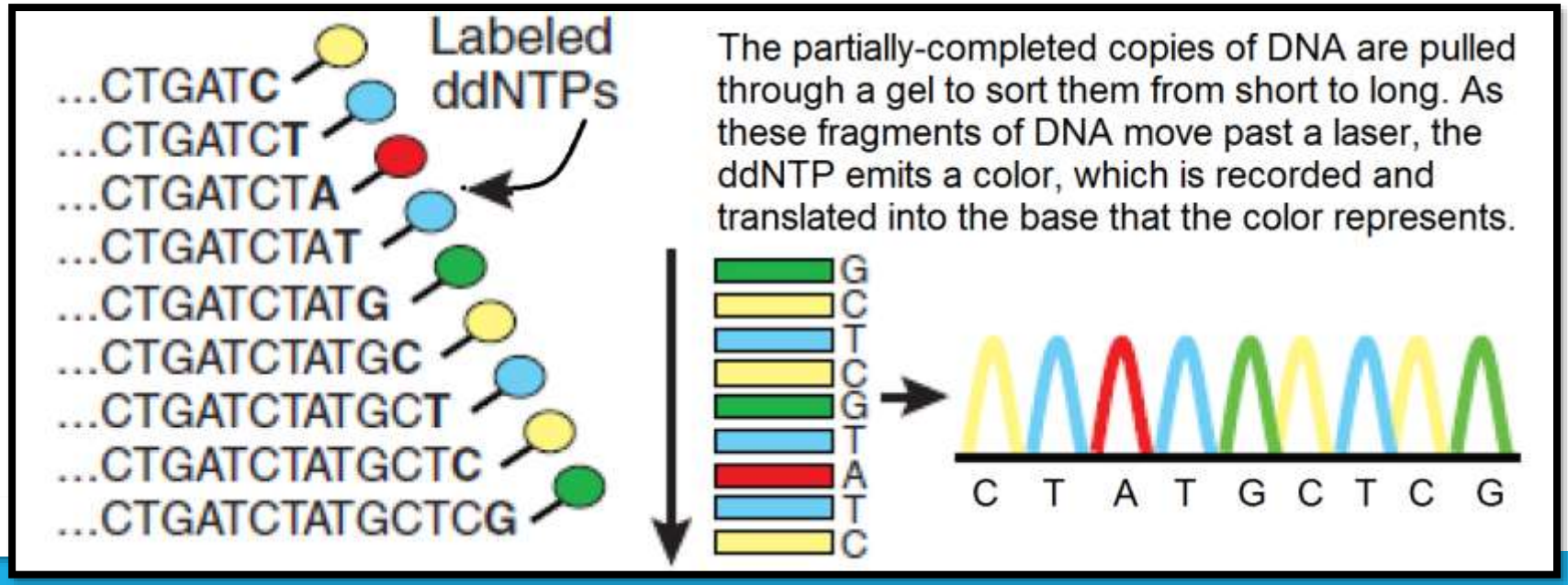
- **First, the dye attached to each ddNTP will indicate which 'letter' it represents.**
  - For example, yellow for G, green for A, red for T, and blue for C.
- **However, this alone is not enough to read an entire strand of DNA.**
  - If we only had one copy of that strand of DNA, we would only know one of the nucleotide letters that it contains!
- **Instead, millions of copies of that particular strand of DNA were made.**
  - In each copy of the DNA strand, the ddNTP is randomly added in a different spot.



Source: www.vce.bioninja.com.au

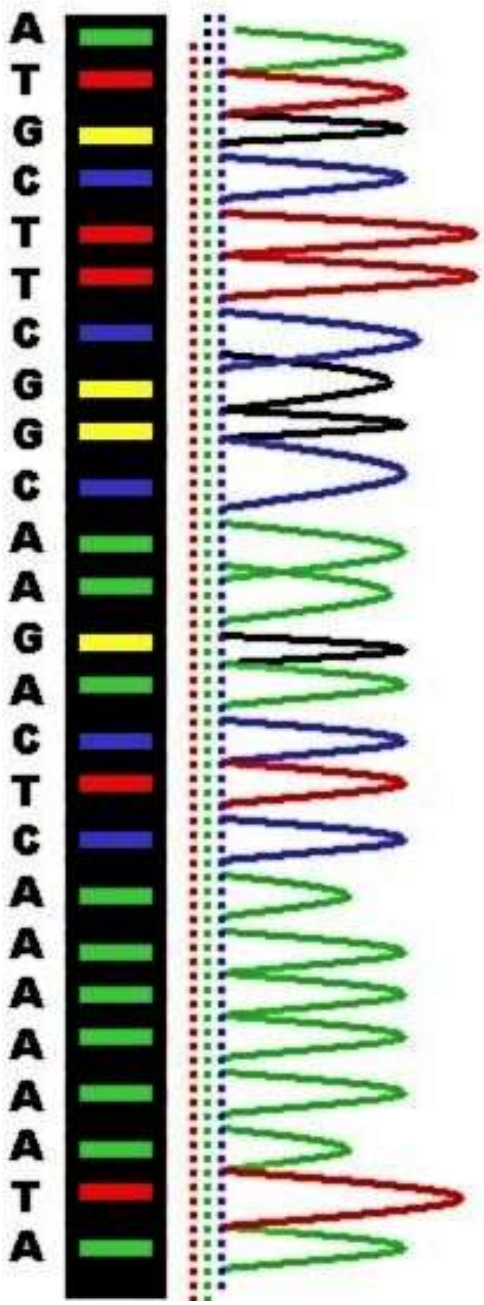
ddNTP's

- **Because the ddNTP prevents the addition of any other nucleotide bases once it has been added, we know that the color given off by that particular strand of DNA represents the last letter added in that sequence.**
  - Each of the millions of copies of that same sample of DNA has a ddNTP that was randomly added at a different spot.
  - Each of these millions of copies of DNA has a different length because nothing can come after the ddNTP.
  - If we line up the copies of that strand of DNA from shortest to longest, we can determine the order of nucleotide bases by looking at the order of the colors emitted by the ddNTPs.



Lining Up Shortest to Longest

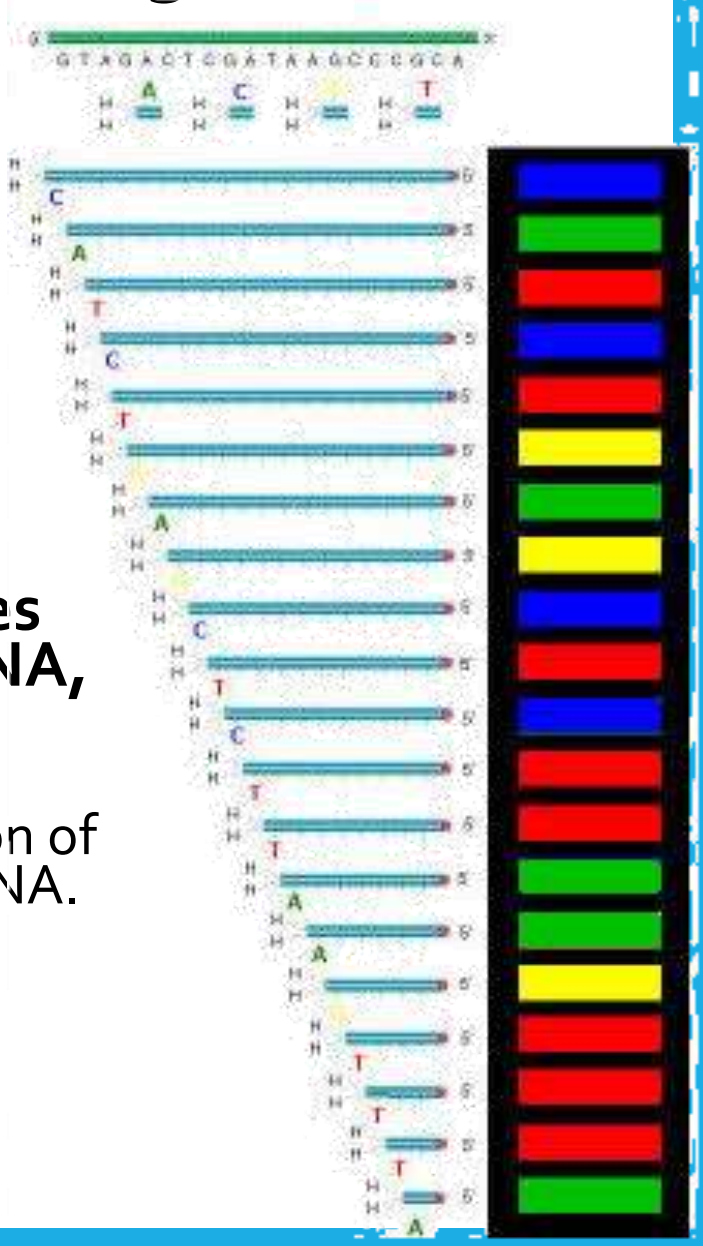
- **To line up the copies of DNA from shortest to longest, we put all of the partially-copied fragments of DNA into a specialized gel.**
  - We then pull the DNA through the gel using electricity (because DNA is negatively charged, electrical current will cause it to move towards a positive pole).
  - The smallest copies of the DNA strand will move more quickly through the gel than the longer copies (*just like a swimmer in a small speedo can move more quickly through the water than a swimmer in large board shorts*).
- **As the DNA fragments are pulled through the gel, they will move past a laser one by one.**
  - As the partial-copy of DNA moves past the laser, the ddNTP that was added to that particular sequence will flash a color indicating the last letter added.
- **The sequence of colors that occur as the DNA moves past the laser will correspond to the sequence of nucleotide bases (A, T, G, C) in the DNA.**



Source: en.wikipedia.org

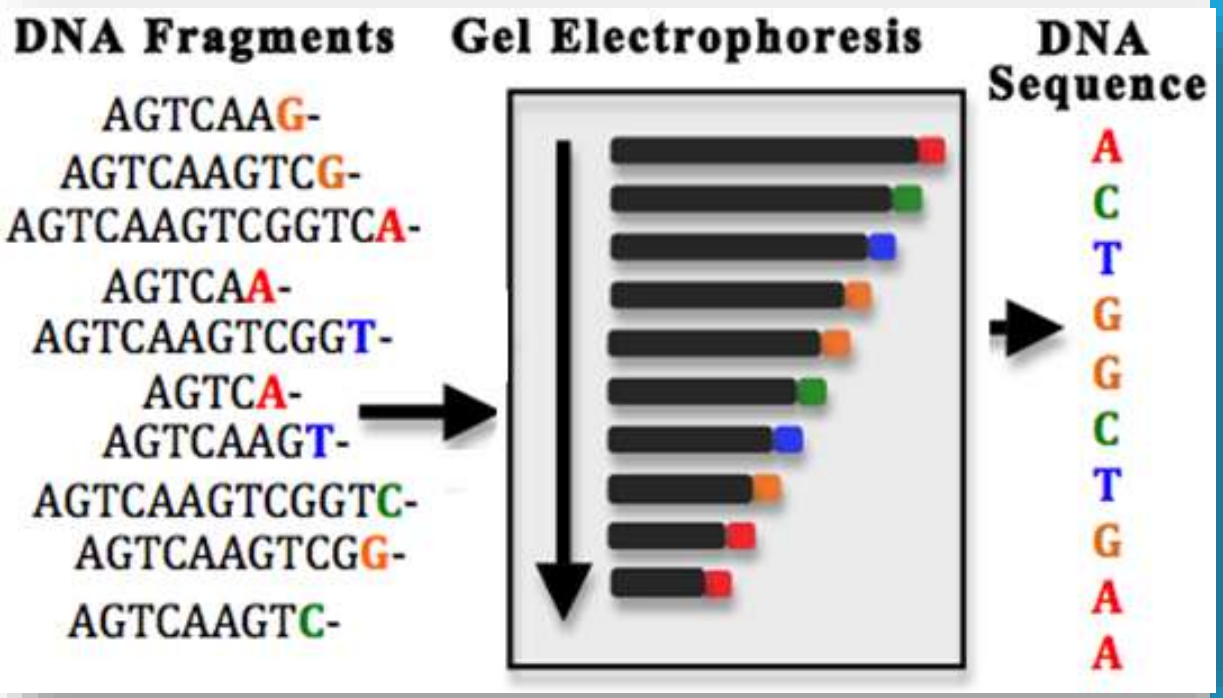
Summary of Sanger Sequencing

- **1. A sample of DNA is copied millions of times using bacteria.**
- **2. The DNA is changed from double stranded to single stranded (denaturing).**
- **3. Polymerase enzyme, primers, ddNTPs, and bases are added to the denatured sample of DNA.**
- **4. As polymerase adds nucleotide bases to the copies of the single stranded DNA, it will randomly add a ddNTP to each copy.**
  - The addition of the ddNTP stops the addition of any other bases to that particular copy of DNA.
  - This creates millions of partially-made fragments of the same strand of DNA, each being a different length.



Summary of Sanger Sequencing

- **5. The partially-made fragments of DNA are then pulled through a gel using electricity.**
  - This causes the DNA fragments to line up from shortest to longest.
- **6. As the fragments of DNA move through the gel, they will move past a laser.**
  - The laser causes the fragment to give off a color that represents the last letter added in that particular copy of the DNA.
- **7. The sequence of colors that occur corresponds to the sequence of nucleotide bases in that sample of DNA.**



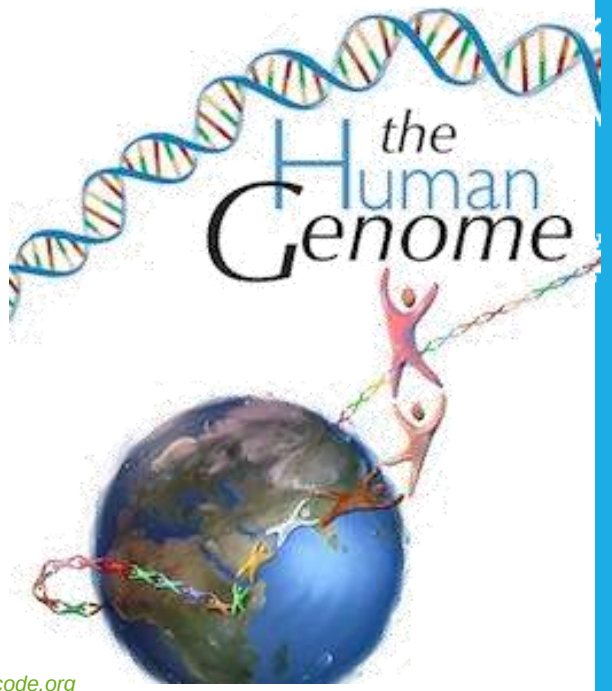
Source: hplusmagazine.com

THE HUMAN GENOME PROJECT

One of the greatest scientific accomplishments in history.

Human Genome Project

- **The development of the Sanger Sequencing Method in the 1970s made the prospect of reading the entire human genome feasible for the first time.**
  - In 1990, the National Institute of Health and the US Department of Energy launched an ambitious project to sequence the entire human genome within 15 years.
- **This goal for biological science was equivalent to what the moon landing was to aviation.**
  - It was incredibly ambitious and unprecedented in regard to its impact and importance.
- **In 1998, a corporation called Celera Genomics was founded to compete with the US government to become the first to sequence the entire genome.**
  - The race between the federal government and Celera Genomics was similar to the space race between the US and Russian governments during the 1960's.



Source: unlockinglifescode.org

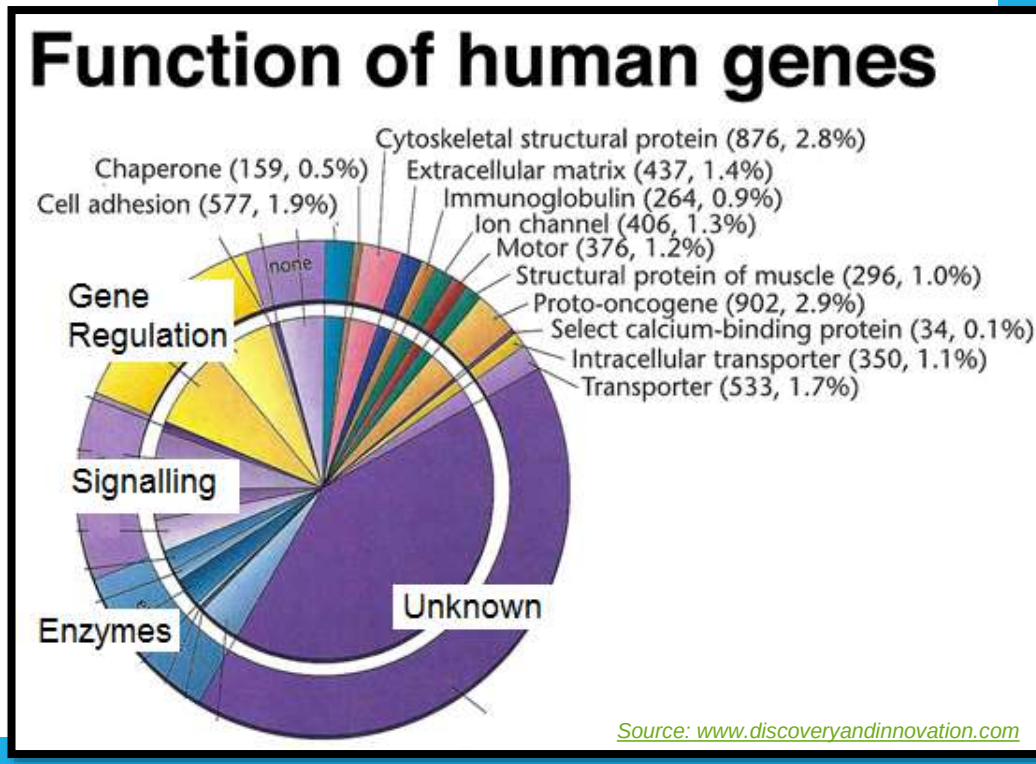
Human Genome Project

- **By 1999, the first human chromosome was sequenced.**
  - A chromosome is a "package" of tightly coiled DNA. Humans contain 23 pairs of chromosomes.
  - This chromosome was sequenced by a collaboration of scientists in the U.S., England, Japan, France, Germany, and China.
- **By 2001, both the collaboration of international governments as well as Celera Genomics had published their own drafts of the human genome.**
  - Both reports found that there are about 20,000 genes in the human body and that human DNA is 99.9% identical from person to person.
- **By 2003, all of the goals of the Human Genome Project had been completed.**
  - This project was completed over two years ahead of schedule and under budget.
  - Additional species' genomes have been sequenced since then, including the mouse, dog, cat, cow, pig, horse, and many more.

Human Genome Project

- **To this date, the Human Genome Project is recognized as one of the most outstanding and successful achievements in the history of science.**
  - However, the scientific research related to the human genome (and the genomes of other species) has only just begun.

- **Now that scientists know the sequence of A's, G's, C's, and T's in the genomes of many species, they can begin to determine how these genes function, what proteins they code for, and finally begin to understand how many of biology's most complicated processes function.**
  - These unknown processes include how a single cell fertilized egg can become an adult with trillions of cells, how genes regulate the functions of organs, how diseases occur in otherwise healthy people, and how the human brain works.



NEXT GENERATION SEQUENCING METHODS

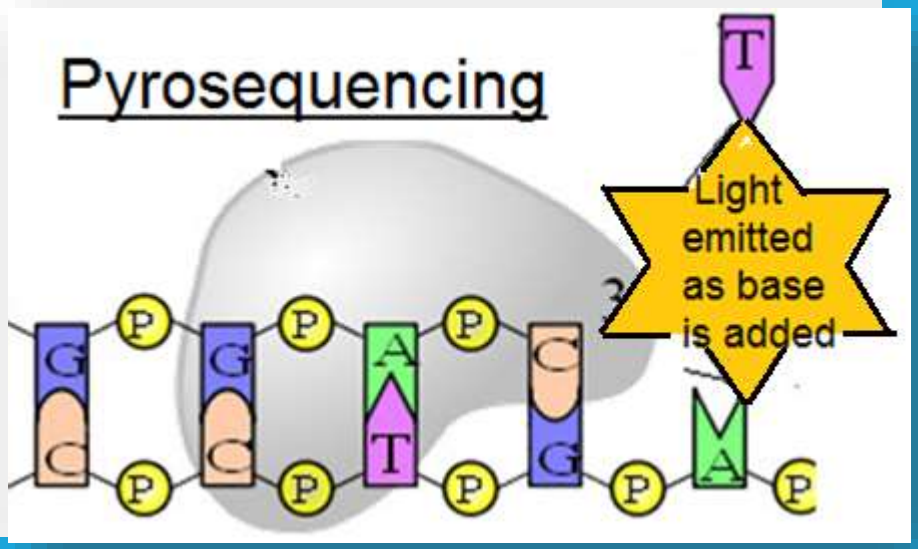
Making DNA sequencing faster and cheaper.

Next Generation Sequencing

- **The Sanger Sequencing Method made the Human Genome Project and other major scientific advances in genomics possible.**
  - However, the Sanger Method is slow and expensive, leading researchers to seek other ways to more effectively read the nucleotide bases that make up genes and genomes.
  - These new methods are called Next-Generation Sequencing, or NGS.
- **Next-Generation Sequencing is not a specific method of genetic sequencing.**
  - NGS is a collective term for all of the newest methods of reading genes and genomes that are faster and cheaper than the Sanger Method.
  - Two of the most widely-used NGS methods include 454-Roche Pyrosequencing and Illumina Bridge Sequencing.
  - Others include Ion-Torrent Sequencing and Nanopore Sequencing.

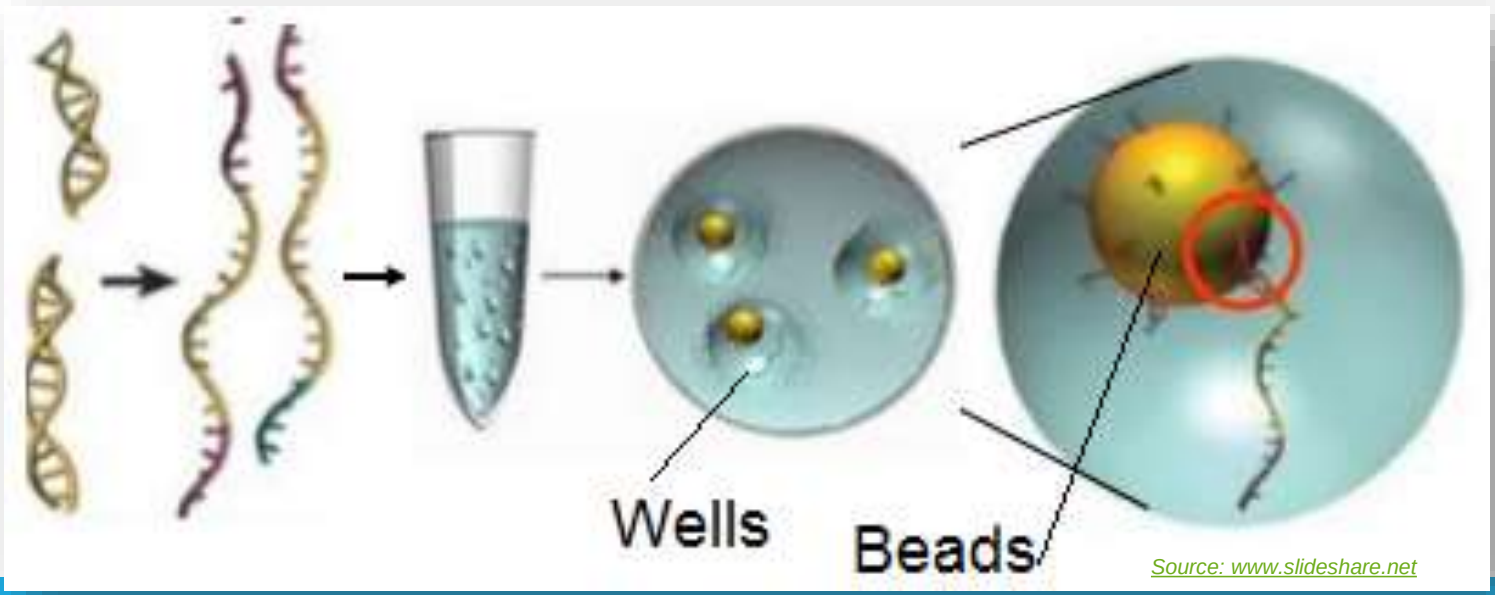
454 Roche Pyrosequencing

- **<u>454-Roche Pyrosequencing</u> is a method of genetic sequencing in which a flash of light is given off every time a nucleotide base is added to a single-stranded section of DNA.**
  - Roche is pronounced "Roash" (like "roast" but with an 'sh').
  - "Pyro" means fire or light, so pyrosequencing is a version of sequencing that uses flashes of light to determine the order of bases.
- **By recording whether or not light has been emitted when each base is added, a scientist can read the genome.**
  - For example, if "A" nucleotide bases are added and a flash of light was given off, then we would know that the next nucleotide base in a sequence is "A".
  - If no light was given off when the A's are added, then "A" is not the next nucleotide base in a sequence.
  - All four bases would be added until a flash of light is given off (indicating which of the four bases is next in the sequence of DNA).



454 Roche Pyrosequencing

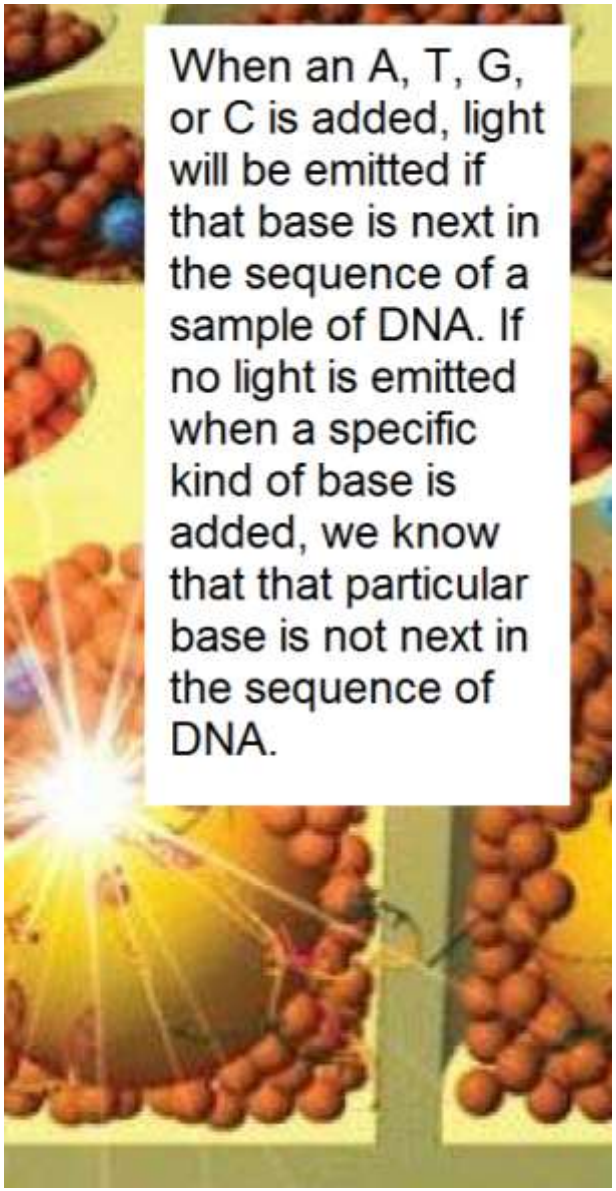
- **The 454-Roche method begins by taking the sample of DNA and breaking it into much smaller pieces.**
  - The "chunks" of DNA are then attached to specialized beads.
  - Primers on each bead ensure that only one kind of chunk of DNA attaches to each bead.
- **The DNA that has attached to the bead is then denatured so that it is single stranded.**
  - The beads (each with a different chunk of the DNA sample) are then placed in the depressions on a specialized plate.
  - These depressions are called "wells".



Source: www.slideshare.net

454 Roche Pyrosequencing

- **Polymerase enzymes are added to each well so that they can add nucleotide bases to the single-stranded DNA.**
  - Nucleotide bases are added one at a time to the wells (for example: A, then G, then C, then T).
- **The nucleotide bases used for pyrosequencing are different from ordinary bases.**
  - These bases will flash a color of light when they are added to the single-stranded (denatured) DNA.
- **If a flash of light is given off when the base is added to a well with the DNA, this indicates that the particular base that was added is next in the sequence of DNA.**
  - If no light is given off when a base is added to a well with the sample of DNA, it indicates that this particular base is *not* next in the sample of DNA.



Source: http://www.mdpi.com/microarrays/microarrays-03-00245/article_deploy/html/images/microarrays-03-00245-g005-1024.png

454 Roche Pyrosequencing

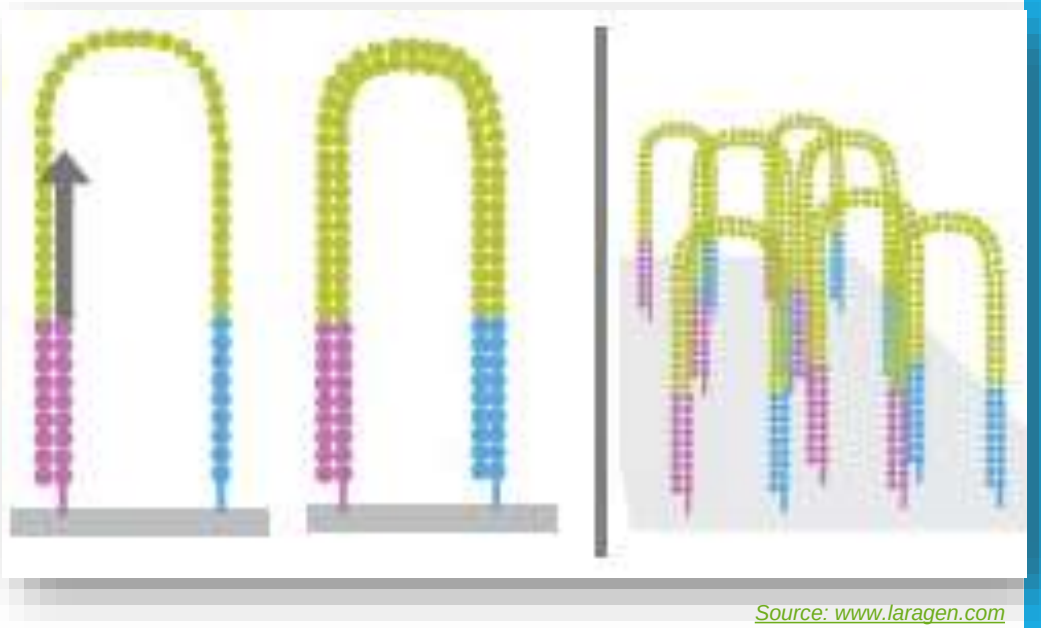
- **By knowing when and which nucleotide bases were added, and by recording when and where light was given off, a computer can then use this recorded data to read the sequence of A, G, C, and T in that sample of DNA.**

- **The 454-Roche Pyrosequencing method can read as much DNA in one day as the Sanger method can read in an entire year!**
  - The 454-Roche Pyrosequencing method is also much cheaper.
  - The Sanger method costs 5 times as much as the 454-Roche method to read the same amount of DNA.

Source: http://www.symposcium.com/wp-content/uploads/2013/05/454-pyrosequencing-picotiter-beads.png

Illumina Sequencing

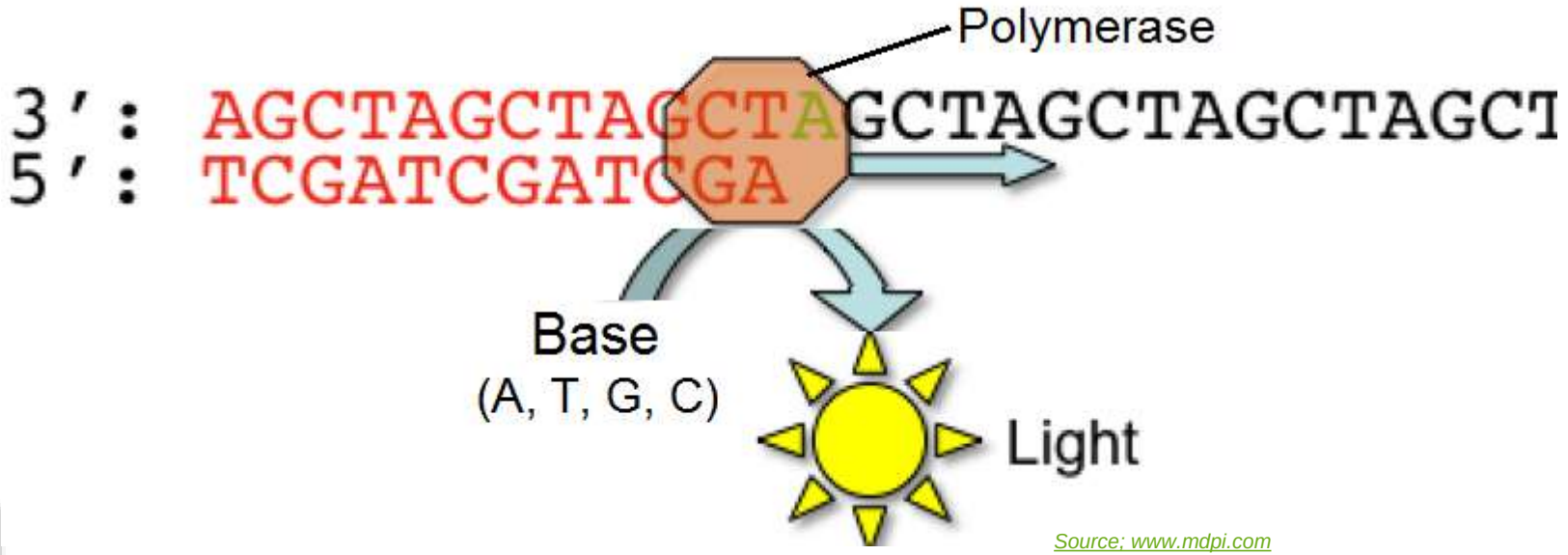
- **<u>Illumina sequencing</u> is another type of Next Generation Sequencing that can be used to rapidly and cheaply read a gene or genome.**
  - Illumina sequencing is similar to the 454-Roche method in that light is given off when a nucleotide base is added to the denatured sample of DNA.
- **However, instead of attaching the DNA to specialized beads, the Illumina method works by attaching each end of the DNA to a specialized plate.**
  - This forms what looks like an arch of DNA.
  - Polymerase and specialized nucleotide bases are then added to the chunks of DNA anchored to the plate.



Source: www.laragen.com

Illumina Sequencing

- **Illumina sequencing works very similarly to 454-Roche sequencing by using flashes of light when a base is added to determine the order of bases in a sample of DNA.**
  - The main difference is that 454-Roche uses beads while Illumina uses specialized plates to form 'arches' or 'bridges' of DNA.
- **Like 454-Roche, Illumina sequencing has drastically lowered the cost and difficulty of reading DNA.**
  - Illumina sequencing was the first to be able to sequence a human genome for less than $1000, currently making it the most inexpensive option in Next Generation Sequencing.



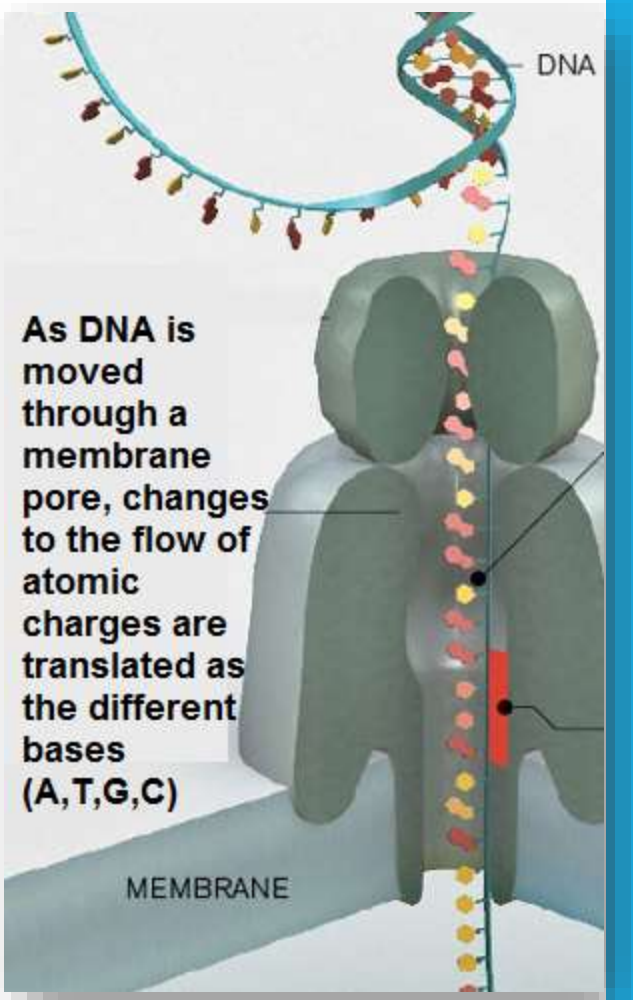
Source; www.mdpi.com

Ion-Torrent Sequencing

- **Some of the newest methods of Next Generation Sequencing use entirely new methods of reading the sequence of nucleotide bases in a sample of DNA.**
  - These include Ion-Torrent Sequencing and Nanopore Sequencing.
- **Ion-Torrent Sequencing is a method of sequencing DNA that relies on the fact that a hydrogen ion is released every time a nucleotide base is added to single-stranded DNA.**
  - The release of a hydrogen ion can be detected by a specialized sensor.
- **The sample DNA that needs to be sequenced is denatured and added to a plate with wells (or depressions).**
  - The wells with DNA are flooded with a single kind of nucleotide (an A, G, C, or T).
- **If the sensor detects a hydrogen ion has been released, it knows that that nucleotide base has been added and is the next in the sequence of the sample of DNA.**
  - If no hydrogen ion is sensed, the base that was added to the wells is *not* next in the sequence of DNA.

Nanopore Sequencing

- **Nanopore Sequencing is a method in which the DNA is fed through a molecular-sized pore that is only just big enough for the denatured DNA to fit through.**
  - As the DNA passes through this pore, the ion current (or flow of atomic charges) changes depending on which type of nucleotide base is passing through the pore at any given time.
  - This is because each of the nucleotide bases has a slightly different shape and size.
  - Because of this, each nucleotide base will interrupt the ionic current in a slightly different way.
  - Measuring how the ionic current changes as the sample of DNA moves through the pore allows the sequence of bases in the sample of DNA to be read.



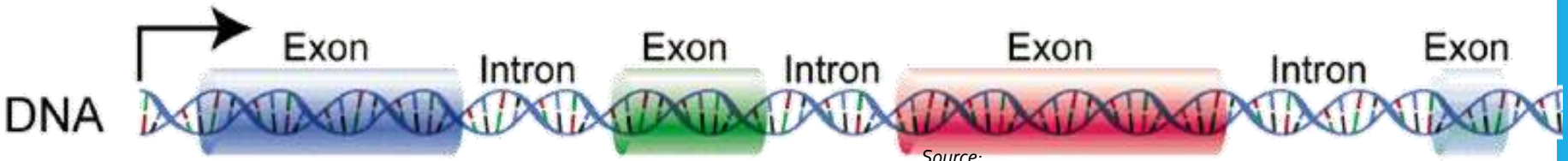
Source: www2.technologyreview.com

USING GENETIC SEQUENCES

Making sense of a billion bases.

Introns & Exons

- **A problem we face when reading DNA is that most of an organism's DNA codes for nothing.**
  - Lengths of non-coding DNA are called introns.
  - Introns are genes for nothing – they aren't used to create proteins (they are incapable of creating proteins). They are sort of like junk DNA.
- **Most researchers are interested in the genes that actually code for a protein.**
  - The portions of DNA that code for an actual protein are called exons.
  - Exons are exceptional at creating proteins.
- **An exon can be found by looking for an Open Reading Frame, or ORF.**
  - An ORF simply means that there are no STOP commands between the beginning and end of a sequence of DNA.
  - If no stop commands are found until the end of a sequence of DNA, we would know that this particular stretch of DNA is a useful, protein-coding exon



Source: simple.wikipedia.org

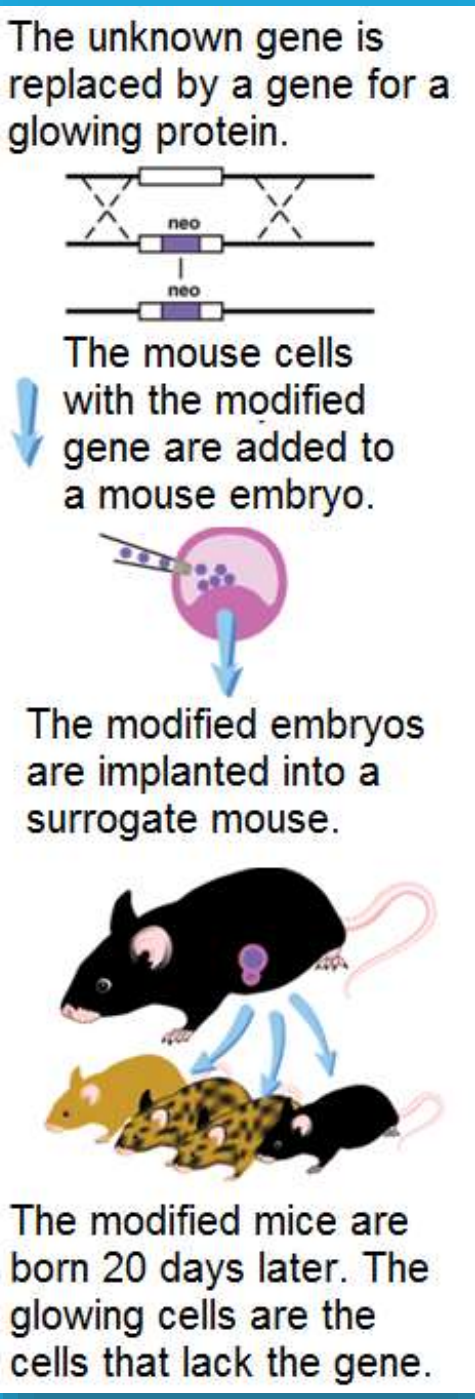
Using this Information

- **Once a gene or genome has been sequenced and its exons have been identified, the only information available is the order of A's, G's, C's, and T's that comprise those genes.**
  - This does not provide any information to the researcher about what a specific gene sequence actually does.
- **To determine the function of a specific gene, or of all of the genes in a genome, researchers have several options.**
  - One of the most widely-used methods to determine the function of a gene is called a knockout mouse.
  - A knockout mouse is a mouse that had the gene of interest deleted from its genome.

Source: jamesfindonblog.wordpress.com

Knockout Mice

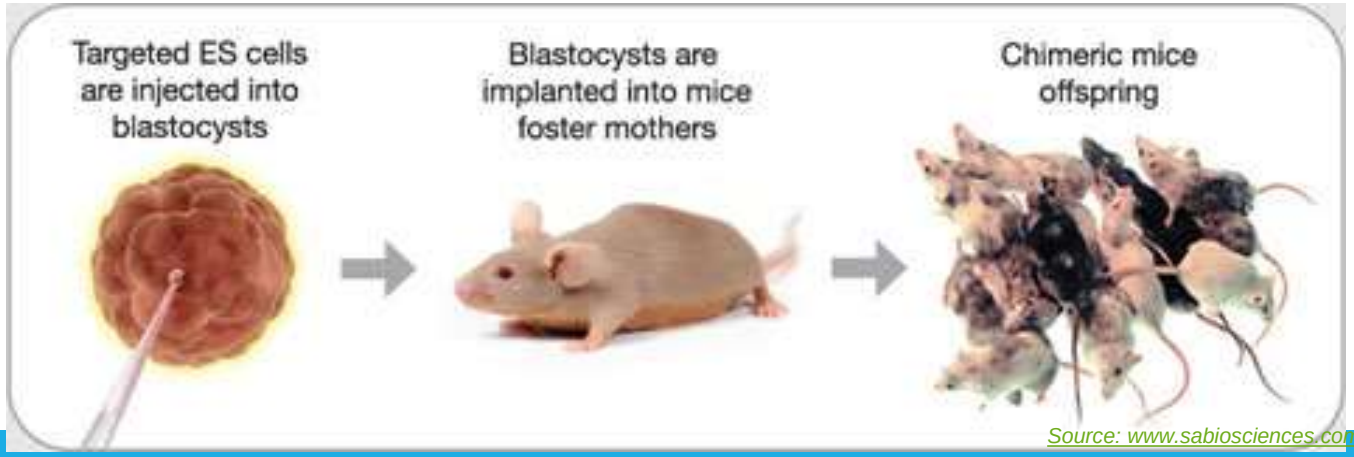
- **To create a knockout mouse, scientists create an artificial sequence of DNA in which the gene with the unknown function is deleted (or "knocked out").**
  - This removed gene is then replaced by a gene for a glowing protein.
  - The artificial DNA (with the glowing protein) is then inserted into fertilized mouse embryos.
- **Some of the embryonic cells will have the artificial gene with the sequence to make the glowing protein.**
  - Other embryonic cells will be unchanged and will still have a functional gene with the unknown function.
- **The embryo is implanted in the womb of a mouse and the baby mouse is born after 20 days.**



Source: http://healthsciences.ucsd.edu/research/moores/shared-resources/transgenic-core/services/PublishingImages/targetingschematic.gif

Knockout Mice

- **This baby mouse is now a chimera, or an organism with two different genomes.**
  - Some of the cells in this baby mouse will have the normal, unchanged genome.
  - Other cells will have the gene for glowing the protein instead of the gene we are researching.
- **Scientists know that the cells that glow do not have the gene that they are trying to understand.**
  - By looking at how the cells that are lacking the gene (and glow) perform differently than the cells that have the gene (and do not glow), they can understand the function of that particular gene by looking at what does not happen in the glowing cells compared to what does happen in the non-glowing cells.



Source: www.sabiosciences.com

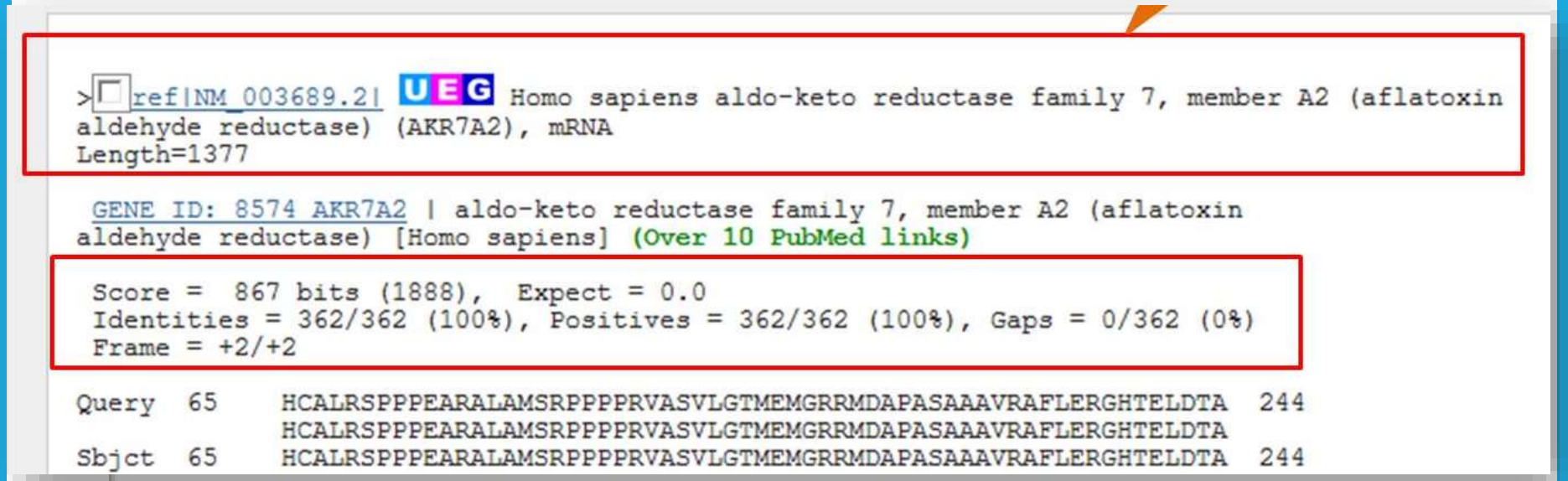
Knockout Mice

- **For example, if the glowing cells do not produce an important protein for the breakdown of sugars inside the cell, we would know that the gene we deleted is necessary for the breakdown of sugar.**
  - On the other hand, if the glowing cells do not produce an important protein for moving substances inside the cell, we would know that the deleted gene was responsible for a transport protein.
  - Whatever function that does not occur in the glowing cells is likely related to the gene that was deleted.
- **In addition to using knockout mice, scientists can also use a tool called BLAST to understand the function of a newly sequenced gene.**
  - BLAST is a search engine for genetic sequences.
  - BLAST for DNA is sort of what Google is like for the Internet.

Source: http://www.intechopen.com/source/html/19263/media/image3.jpeg

BLAST

- **BLAST allows a researcher to compare the sequence of DNA that they have acquired to all other sequences of DNA that have been found.**
  - BLAST will then identify all other known sequences in other species that are similar.
  - BLAST will also report the function of those known sequences.
  - The downside of using BLAST is that if a gene is unlike anything that has ever been discovered before, a BLAST search will not turn up any results and scientists will have to rely on a knockout mouse or similar method to determine the function of that particular gene.



```
>ref|NM_003689.2| UEG Homo sapiens aldo-keto reductase family 7, member A2 (aflatoxin
aldehyde reductase) (AKR7A2), mRNA
Length=1377

 GENE ID: 8574 AKR7A2 | aldo-keto reductase family 7, member A2 (aflatoxin
aldehyde reductase) [Homo sapiens] (Over 10 PubMed links)

 Score =  867 bits (1888),  Expect = 0.0
 Identities = 362/362 (100%), Positives = 362/362 (100%), Gaps = 0/362 (0%)
 Frame = +2/+2

Query  65    HCALRSPPPEARALAMSRPPPPRVASVLGTMEMGRRMDAPASAAAVRAFLERGHTELDTA  244
            HCALRSPPPEARALAMSRPPPPRVASVLGTMEMGRRMDAPASAAAVRAFLERGHTELDTA
Sbjct  65    HCALRSPPPEARALAMSRPPPPRVASVLGTMEMGRRMDAPASAAAVRAFLERGHTELDTA  244
```


Source: woratanti.wordpress.com