Molecular & Cell Biology Lab Manual

Open Education Resource

Jeremy Seto

New York City College of Technology
Department of Biological Sciences

BIO3620L

https://openlab.citytech.cuny.edu/bio-oer

Types of Micropipettors

Pipettors are made by many different manufacturers and thus all do not look the same. Learning to correctly use one type of pipettor will provide you the knowledge to use others as they share the same method of distributing a small volumes. This lab will illustrate the Rainin Pipetman® micropipettors.

The top of the plunger shows the pipettor size for the Pipetman models. Each pipettor has its own volume range and it is **CRITICAL** to use a pipettor only in its proper volume range. The "P-number" represents the maximum volume in µL that the pipettor can measure. Pipettors are more accurate in the *upper* part of their range. 20 µL should be measured with a P20 rather than with a P200. The four pipettor sizes (P10, P20, P200, P1000) used in our lab will measure from 1 µL - 1000 µL as shown below.

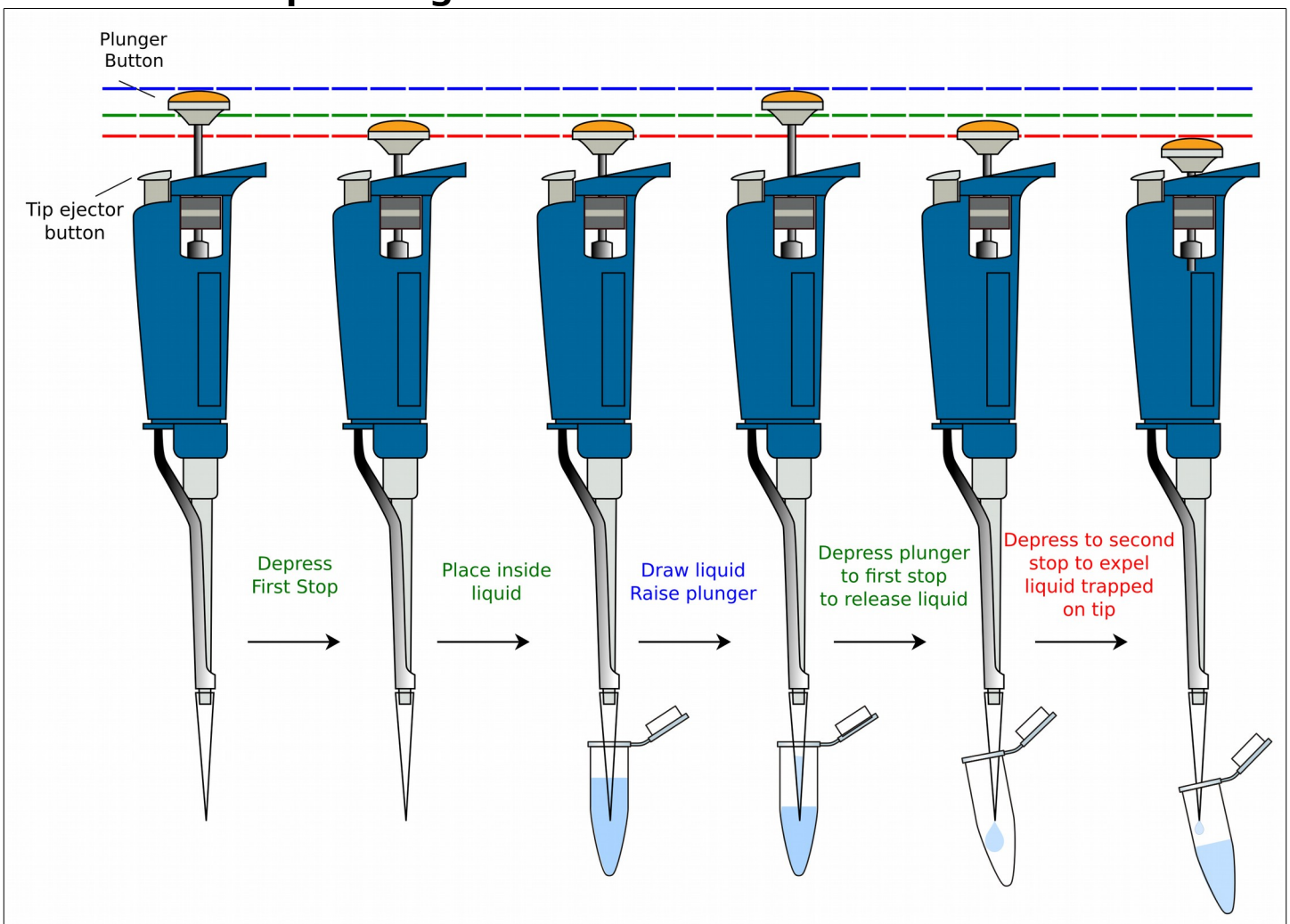
Correctly Adjusting the Pipettors

	P10	P20	P200	P1000
Volume Range (µL)	1-10	2-20	20-200 50-200 *	200-1000
	tens 0 ones 4 tenths 7 4.7µl	tens 1 ones 7 tenths 8 17.8µl	hundreds 1 tens 5 ones 4 154µl	thousands 0 hundreds 8 tens 6 860µl

Gilson Pipetman pipetting ranges chart. Note that the P200 officially has a range from 50-200µl

1 Jeremy Seto jseto@citytech.cuny.edu

Tutorial on Proper Usage



Pipetting sequence. Ensure the plunger is depressed outside of the liquid to displace air and avoid blowing bubbles into solution. Carefully draw the plunger up slowly and follow the liquid to avoid drawing air. Depress the plunger in destination tube to the first stop. Depress to the second stop if fluid remains in the tip to expel full volume.

https://www.youtube.com/watch?v=uEy_NGDfo_8

Jeremy Seto jseto@citytech.cuny.edu

Rules for use of the micropipettors:

These are precision instruments which can easily be damaged. Treat them with respect and care. They are essential for your success in this course and shared amongst numerous students.

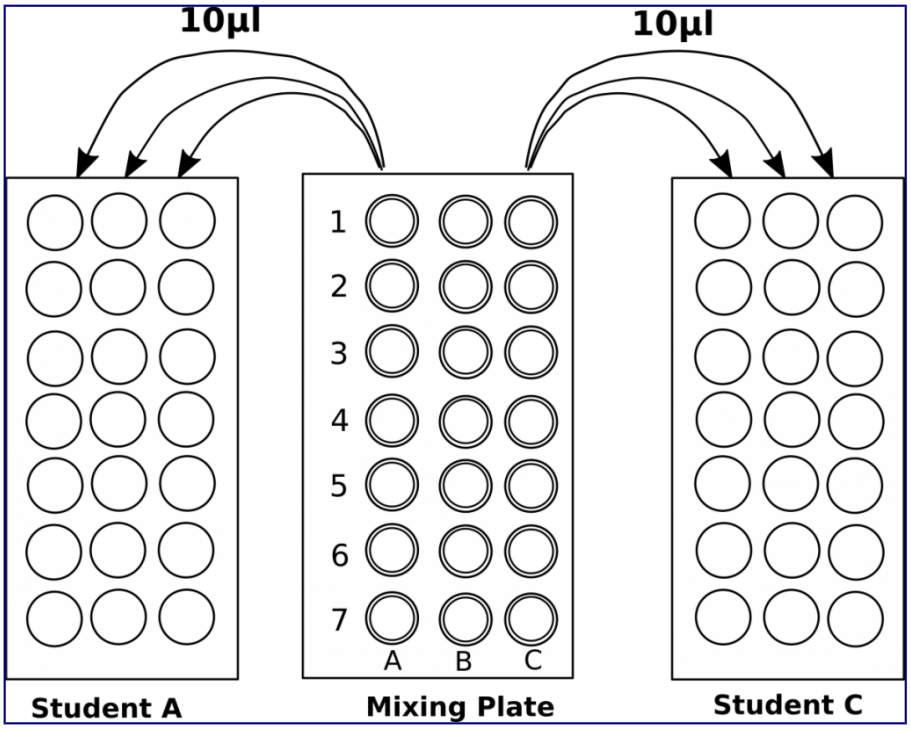
1. Never measure higher or lower than the range of the pipettor allows.
    - an exception to this rule is the P200 labeled 50-200 µl.
    - while we have P100 pipettors for this range, they appear too similar to P20 that they are often confused
    - originally, P200 were labeled 20-200 µl and we know that the lower range is less precise on these
2. Never turn the volume adjuster above or below this range indicated on the pipettor or you risk breaking the instrument.
3. Never allow liquid to get into the micropipettor.
    - this causes contamination
    - this weakens the seal on the o-rings and can damage them
4. Never use the micropipettor without a tip.
5. Never invert or lay down the micropipettor with liquid in the tip.
    - liquids will roll into the piston this way
6. Never let the plunger snap back when filling or ejecting liquid.
7. Never immerse the barrel in fluid.
    - this causes contamination
8. Never set the micropipettor on the edge of the bench; this may result in the micropipettor falling or being knocked onto the floor.

Jeremy Seto jseto@citytech.cuny.edu

Exercise: Pipetting Practice

1. Prepare seven dye mixtures as illustrated in the table below.
    1. One student mix samples in column A and a second mixes in column C
    2. Column B is left empty and used if one student makes a mistake
2. Each dye mixture prepared in the first well to reach a total volume of 45 µl.
3. Pipet 10 µl in triplicate from each well of the mixing plate into the center of the appropriate circles on the target card

Well #	Buffer (µl)	Red (µl)	Blue (µl)	Yellow (µl)	Glycerol (µl)	Alcohol (µl)	Total (µl)
1	40	5	-	-	-	-	**45**
2	15	-	-	10	10	10	**45**
3	20	5	10	10	-	-	**45**
4	30	-	15	-	-	-	**45**
5	22	13	-	-	-	10	**45**
6	25	-	10	-	10	-	**45**
7	20	6	-	6	13	-	**45**



4
Jeremy Seto jseto@citytech.cuny.edu

The Metric System

The **metric system** is an internationally agreed upon measurement system based on decimals or powers of 10. Scientists use a refined version called the **International System of Units** (abbreviated **SI**). In biology, you will often find a need to describe measurements of length, volume, mass, time, temperature or amount of substance.

International System of Units

BASIC SI UNITS		
MEASURE	**SI UNIT**	**SYMBOL**
length	meter	**m**
mass	kilogram	**kg**
time	second	**s**
temperature	Kelvin *(Celsius is used in Biology)*	**K** *(°C)*
quantity	mole	**mol**
current	Ampere	**A**
luminosity	candela	**cd**

Jeremy Seto jseto@citytech.cuny.edu

Metric Units:

- length: meter (**m**)
- volume: liter (**L**)
- mass: gram (**g**)
- time: second (**s**)
- temperature: Celsius (**°C**)
    - Kelvin (**K**) is a unit of thermodynamic temperature and is the SI unit. The Kelvin scale in the same as the Celsius or centigrade scale but offset by 273.16
    - Biology uses Celsius predominantly because of the range in which organisms live.
- amount of substance: mole (**mol**)
    - A mole is a number representing 6.022×10^{23} of something
    - Just as a pair of shoes equals 2 shoes, a mole of shoes is 6.022×10^{23} shoes
    - Just as a dozen eggs equals 12 eggs, a mole of eggs is 6.022×10^{23} eggs

METRIC PREFIXES IN EVERYDAY USE			
PREFIX	**SYMBOL**	**SCIENTIFIC NOTATION**	**FACTOR**
tera	T	10^{12}	1 000 000 000 000
giga	G	10^{9}	1 000 000 000
mega	M	10^{6}	1 000 000
kilo	k	10^{3}	1 000
hecto	h	10^{2}	100
deca	da	10^{1}	10
BASE UNIT	(none)	10^{0}	1
deci	d	10^{-1}	0.1
centi	c	10^{-2}	0.01
milli	m	10^{-3}	0.001
micro	μ	10^{-6}	0.000 001
nano	n	10^{-9}	0.000 000 001
pico	p	10^{-12}	0.000 000 000 001

Jeremy Seto jseto@citytech.cuny.edu

Strategy for conversions

1. What unit is being asked for?
    - 500ml = ____L → liters
2. What unit are you starting from?
    - 500ml = ____L → milliliters
3. Which unit is larger? By how much is that unit larger?
    - Liters are the larger unit. Liters are 1,000X (10^3) greater than milliliters.
4. Which direction are we moving?
    - Since we are moving to a larger unit, our value will be smaller. In this case, the value is smaller by 1,000X
    - In other words, the value is 1/1000 or 0.001 the value.
    - So what is the answer?

Factoring Out

Using the idea of factors of ten, you can assess the difference of the two units and cancel out the original unit algebraically to reach the desired final unit.

- 500ml=_____L
    - $1ml = \frac{1}{1000}L$ OR, $\frac{1L}{1000ml}$
        - which states 1000 milliliter in every 1 liter
    - $500ml \times \frac{1L}{1000ml} = \frac{500L}{1000} = 0.5L$
        - pay attention to the units and how we've canceled out the ml in the numerator of 500ml and in the denominator in the conversion of 1L in 1000ml

Additional Resources

https://youtu.be/w0nqd_HXHPQ

Jeremy Seto jseto@citytech.cuny.edu

ACCURACY and PRECISION

Accuracy refers to how closely a measured value agrees with the correct or target value.
Precision refers to how closely individual measurements agree with each other and reflects a repeatability in those measurements.

This illustrates accuracy. Measurements are on target.

This illustrates precision. Measurements are very close to each other and repeatable.

This illustrates Accuracy AND Precision. Each measurement is on target and also highly repeatable.

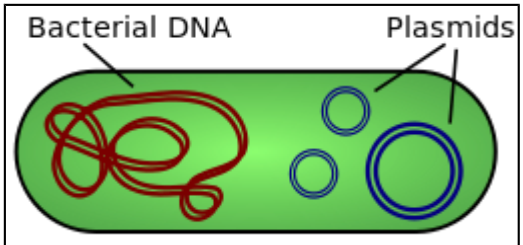
Instruments have a finite amount of accuracy and it is important to report measurements within that level of accuracy. **Significant figures**, report the number of digits that are known to some degree of confidence with the measuring device. With increased sensitivity of the equipment , the number of significant figures increases.

Jeremy Seto jseto@citytech.cuny.edu

Bacterial Genetics and Complementation

George Beadle and Edward Tatum first described the concept that each gene corresponded to an enzyme in a metabolic pathway by exposing the yeast *Neurospoara crassa* to mutagenic conditions (Beadle & Tatum, 1941). Following these procedures Josua lederberg continued these studies with Tatum where they generated two mutants strains in *Escherichia coli* . These bacteria were **auxotrophs**, unable to generate some basic nutrients necessary to sustain their growth. The two strains were described as met^{-} bio^{-} Thr^{+} Leu^{+} Thi^{+} (Strain A) and Met^{+} Bio^{+} thr^{-} leu^{-} thi^{-} (Strain B). Strain A can sufficiently synthesize the amino acids threonine, leucine and the cofactor thiamine while deficient in producing the cofactor biotin and the amino acid methionine while the converse was true of Strain B. When either of these two strains were plated onto minimal media, no growth occurred. Supplementing minimal media with methionine and biotin permitted Strain A to grow as normal. When the two strains were mixed together and plated on minimal media, there was growth of bacteria. The two strains were capable of complementing each other in some way as if a sexual exchange of genetic material had occurred (Lederberg & Tatum, 1946).

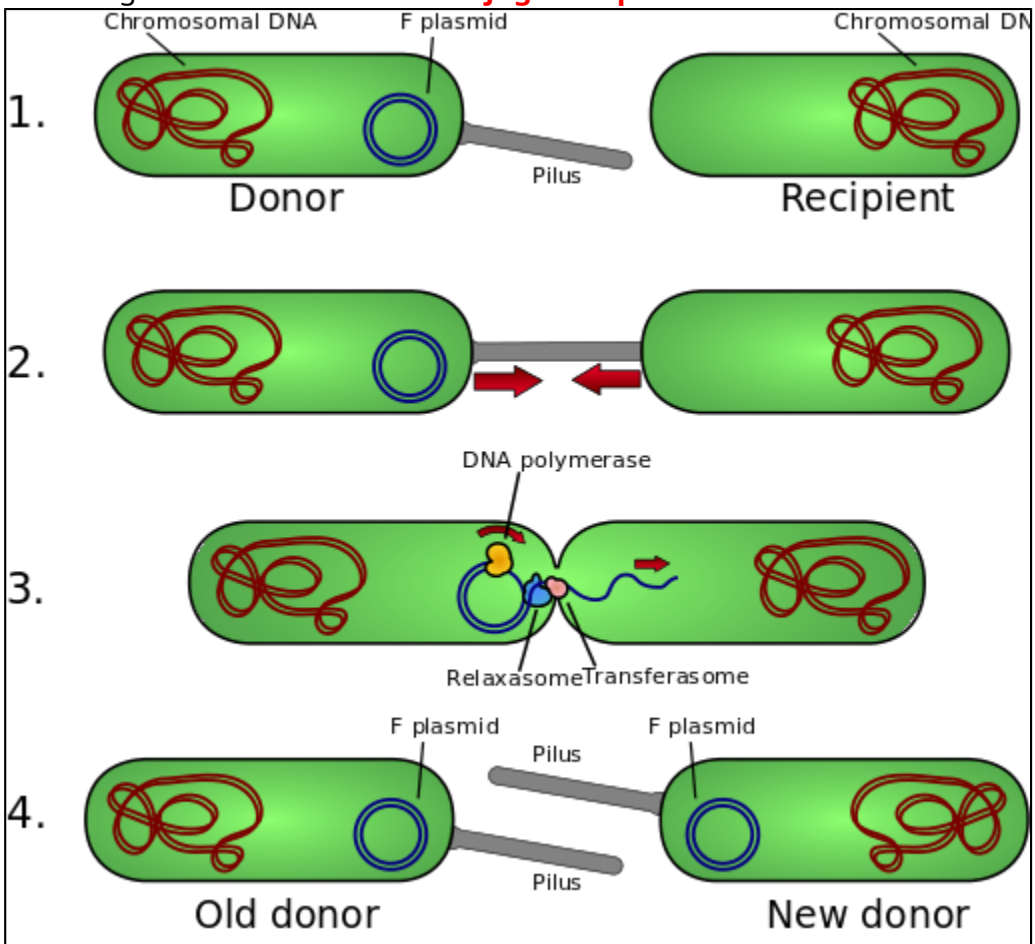
Bacteria are equipped with all the necessary capacities to replicate DNA. Common bacterial species have bee adapted for use in the lab to carry DNA and propagate it for uses in biotechnology. In addition to chromosomal DNA of the bacterial genome, bacteria also have extrachromosomal DNA called **plasmids**. These plasmids replicate independently of the bacterial chromosome and can occur in high copy. These circular pieces of DNA are modified in labs to carry specific pieces of DNA so they can be studied or used for expression into proteins. Plasmids can naturally carry important traits, including antibiotic resistance. Plasmids are relatively small, ranging in size from 1000 bases to 1,000,000 bases long (1kb-1000kb).



Bacterial DNA usually exists as a large circular chromosome (red). Plasmids are extrachromosomal and autonomously replicating pieces of DNA (blue).
Credit: Spaully (CC BY-SA 2.5)

Jeremy Seto jseto@citytech.cuny.edu

Through a process called **conjugation**, bacteria can “sexually” transfer genetic material to another by passing plasmids through a structure called a **conjugation pilus**.



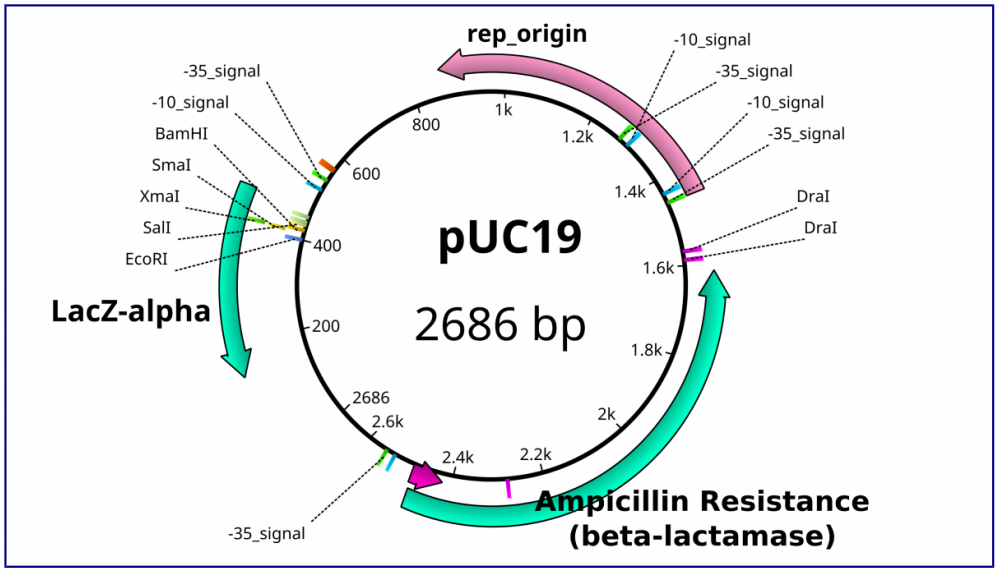
Conjugation process between a plasmid bearing donor and a plasmid-less recipient. The donor creates a conjugation pilus to create a cytosolic bridge with the donor where the plasmid is replicated into the recipient through the rolling circle method of replication. The recipient then becomes competent to act as a donor.

Credit: Adenosine (CC-BY-SA 3.0)

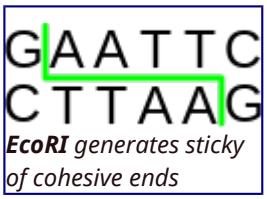
Jeremy Seto jseto@citytech.cuny.edu

Features of Plasmids

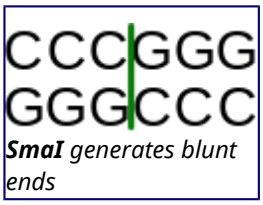
Plasmids that are designed by Biologists to shuttle pieces of DNA for study are referred to as **vectors**, because they *move* a piece of DNA.



These plasmid vectors have the same hallmarks as traditional plasmids with the capacity to replicate independently of the bacterial genome. The feature that allows these DNA's to replicate is called an **origin of replication** (ori) that is usually rich in A's and T's. However, these plasmid vectors have the additional properties that make them easy to work with and distinguishable from bacterial plasmids; a selection marker and a multiple cloning site. A **selection marker** usually comes in the form of a gene that encodes resistance to a specific antibiotic. In the pictured plasmid, Ampicillin resistance granted by the β-lactamase gene. The **multiple cloning site** (**MCS**), also known as the polylinker, is the location in which the DNA of interest is incorporated into the vector. MCSs are defined by a set of unique sites where the DNA can be cut by **restriction endonucleases** (**RE**). As the name implies, restriction enzymes are “restricted” in their ability to cut or digest DNA. The restriction that is useful to biologists is usually **palindromic** DNA sequences. Palindromic sequences are the same sequence forwards and backwards. Some examples of palindromes: RACE CAR, CIVIC, A MAN A PLAN A CANAL PANAMA. With respect to DNA, there are 2 strands that run antiparallelel to each other. Therefore, the reverse complement of one strand is identical to the other.



EcoRI *generates sticky of cohesive ends*

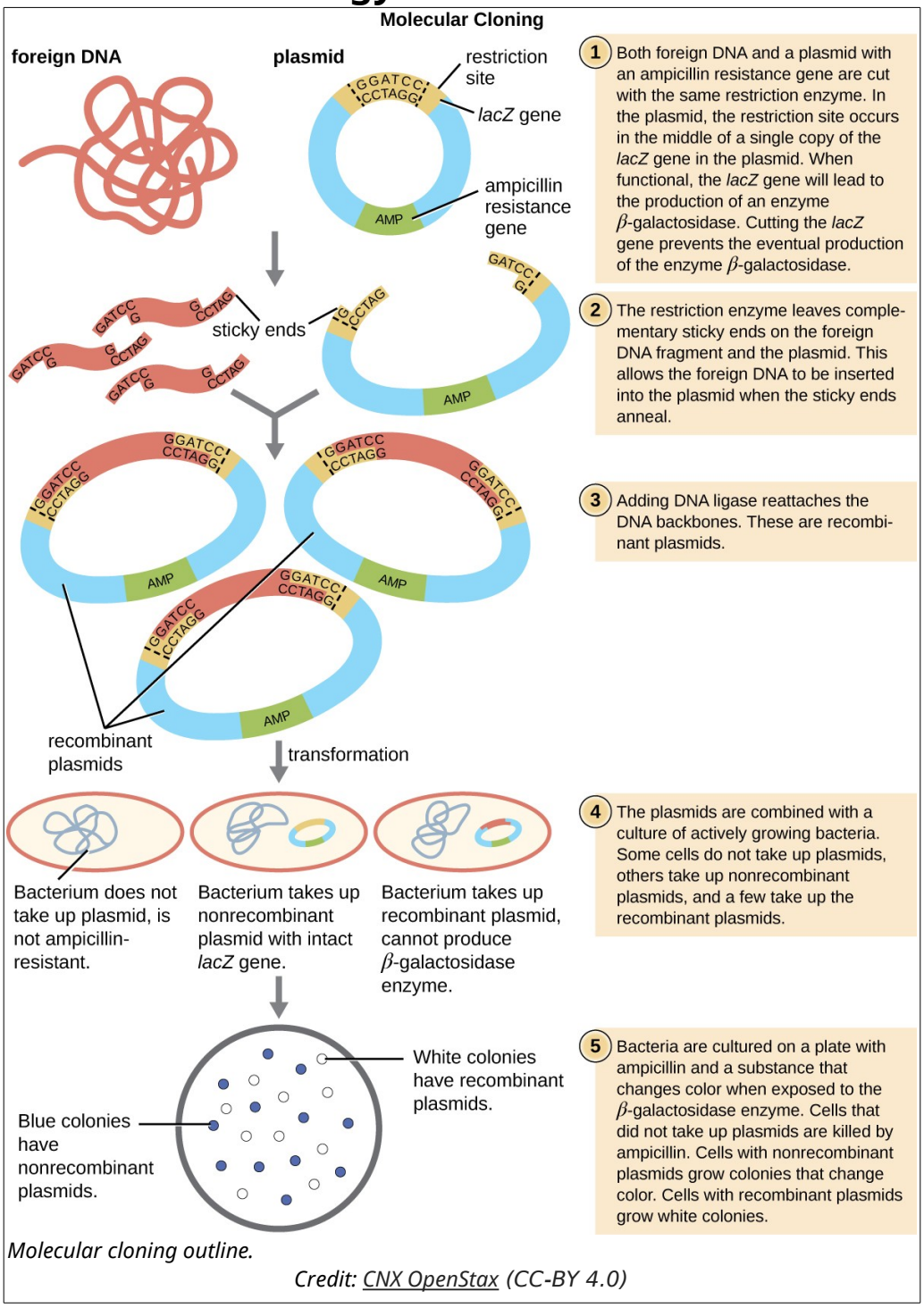


SmaI *generates blunt ends*

Restriction enzymes hydrolyze covalent phosphodiester bonds of the DNA to leave either “sticky/cohesive” ends or “blunt” ends. This distinction in cutting is important because an *EcoRI* sticky end can be used to match up a piece of DNA cut with the same enzyme in order to glue or ligate them back together. While endonucleases cut DNA, **ligases** join them back together. DNA digested with *EcoRI* can be ligated back together with another piece of DNA digested with *EcoRI*, but not to a piece digested with *SmaI*. Another blunt cutter is *EcoRV* with a recognition sequence of GAT | ATC.
By “cutting and pasting” DNA into vectors, we can introduce foreign or exogenous DNA into bacteria. This type of DNA is now called **Recombinant DNA** and is the heart of biotechnology.

Jeremy Seto jseto@citytech.cuny.edu

Recombinant DNA Technology



Molecular cloning outline.

Credit: CNX OpenStax (CC-BY 4.0)

Jeremy Seto jseto@citytech.cuny.edu

Additional Resources:

1. http://www.dnalc.org/resources/3d/20-mechanism-of-recombination.html
2. http://www.dnalc.org/view/16705-Animation-34-Genes-can-be-moved-between-species-.html

Questions for thought

1. Why do you think that origins of replication are made of A's and T's?
2. What is different about the types of bonds holding the double strands together versus phosphodiester bonds of the DNA backbone?
3. Can DNA digested with *SmaI* be ligated to DNA digested with *EcoRV*?
4. If so, which enzyme will be able to digest this new DNA?

References:

- *Beadle, G. W.; Tatum, E. L. (1941)*. "Genetic Control of Biochemical Reactions in Neurospora". *Proceedings of the National Academy of Sciences*. **27** (11): 499–506. doi:10.1073/pnas.27.11.499. PMC 1078370. PMID 16588492
- *Lederberg J, Tatum EL (1946). "Gene recombination in E. coli". Nature. **158** (4016): 558. doi:10.1038/158558a0*

Jeremy Seto jseto@citytech.cuny.edu

Alkaline Lysis

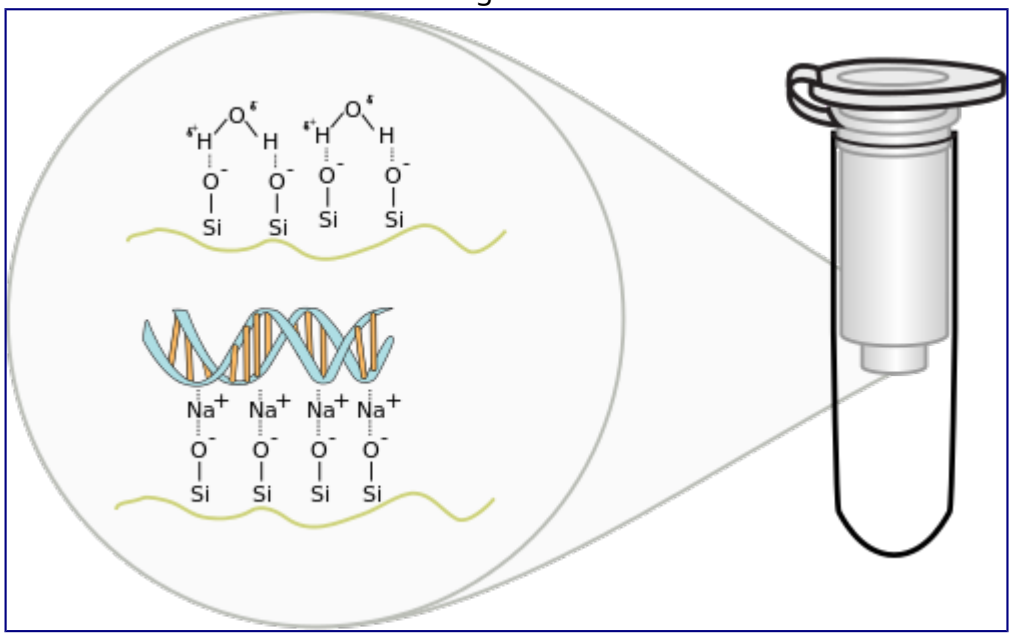
Once DNA is introduced and carried in bacteria, we would like to isolate the DNA again for further manipulation. In order to do so, bacteria containing the plasmid of interest is grown in a liquid culture of nutrient rich broth made of yeast extract called Luria-Bertani Broth (**LB**). These cultured bacteria are grown until they are of a high concentration over night. They are harvested through centrifugation and the broth is removed. The resulting pellet of bacteria is resuspended in a physiological buffer containing the chelator EDTA. A **chelator** is a chemical that removes divalent cations like Ca^{2+} or Mg^{2+} from solution. This is significant because divalent cations are necessary for DNA digesting enzymes to be active. By chelating the ions, the DNA we ultimately wish to purify will be safe from degradation.

After resuspension of the bacteria, an alkaline solution of 0.1N NaOH is mixed into the bacterial mix. This solution also contains an ionic detergent called sodium dodecyl sulfate (**SDS**) that aids in denaturing proteins and disrupting their interactions with the DNA. The mixture becomes viscous as the bacteria burst open and their contents leak into the solution. This basic solution is then neutralize with a potassium acetate buffer at pH5.5. As the solutions mix together, the pH approaches 7 and the potassium interacts with the SDS to cause a precipitation of the genomic chromosomal DNA and proteins. In order to separate the precipitate from the solution, the mixture is centrifuged at high speed to pellet the genomic DNA and protein. The **supernantant**, or solution, is transferred to a column containing a **silica membrane**. Under high salt conditions, DNA adheres to glass or silica. By passing the solution through this column, the plasmid DNA in the supernatant is trapped onto the silica membrane and removed from solution. Additional washes are used to removed stray contaminants and remove the excessive salt. Plasmid DNA is finally removed from the column through **elution** by a low salt buffer. This low salt buffer is Tris pH 8 with EDTA (**TE**). Plasmid DNA can be stored stably in TE buffer in the freezer for extended periods.

Exercise 1: Plasmid DNA Mini-Prep by Alkaline Lysis

1. **Inoculate** 2 ml of rich medium (LB, YT, or Terrific Broth) containing the appropriate antibiotic with a single colony of transformed bacteria. Incubate the culture overnight at 37°C with vigorous shaking. (This is what you were provided)
    1. Each group should take 2 cultures
2. **Centrifuge** culture tubes directly at maximum for 5 minutes.
    1. If incapable of spinning in these tubes, transfer 1.5 ml of the culture into a microfuge tube (Eppendorf Tube).
    2. Centrifuge at maximum speed for 30 sec.
3. When centrifugation is complete, pour the broth solution into a container of bleach
4. **Resuspend** the bacterial pellet in 250 µl of ice-cold P1 solution by vigorous shaking and transfer back into a microcentrifuge tube.
    - P1 is a physiological solution of 50mM Tris at pH 8
    - P1 contains a Chelator called EDTA
        - chelators bind up excess divalent cations that are required for DNAse activity
5. **Lyse**: Add 250µl of P2 solution to each bacterial suspension. Close the tube tightly, and mix the contents by inverting the tube gently five times. Do not vortex! Store the tube on ice.
    - This is the lysis buffer containing the detergent Sodium Dodecyl Sulfate and NaOH

Jeremy Seto jseto@citytech.cuny.edu

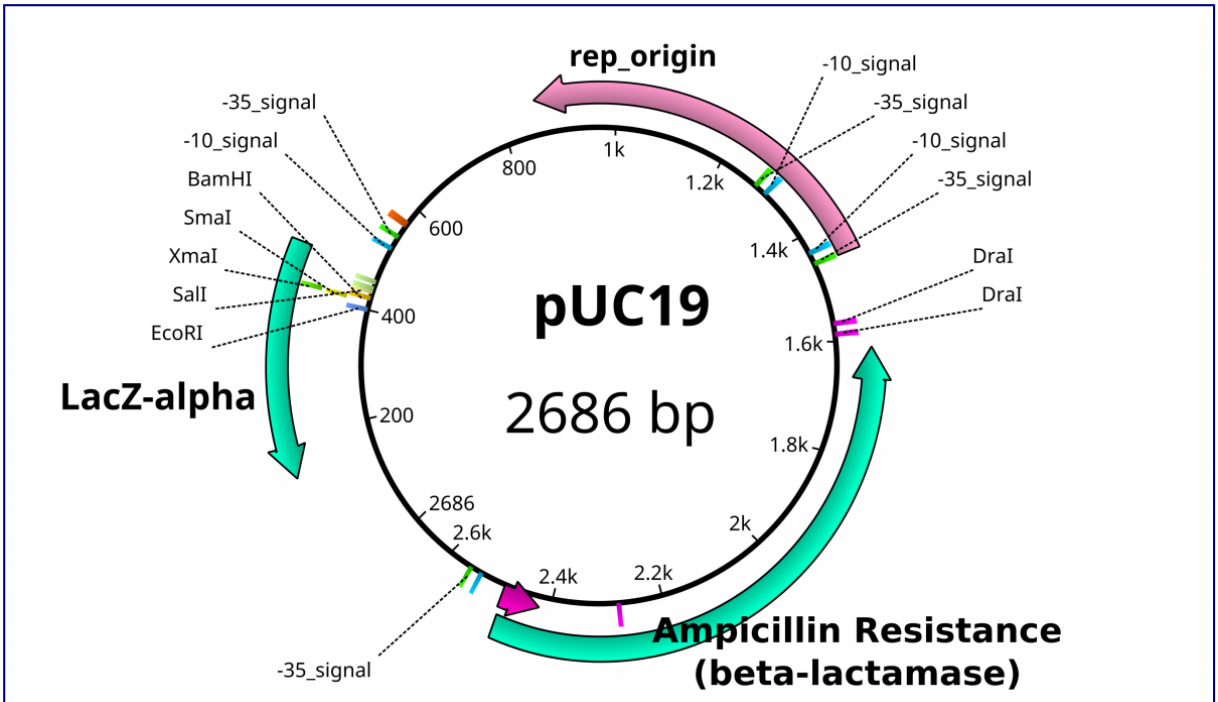
6. **Neutralize:** Add 350 µl of ice-cold P3 solution. Close the tube and disperse lysis solution by inverting the tube several times. Store the tube on ice for 3-5 minutes.
    - This is the neutralization buffer containing Potassium Acetate
    - Neutralization restores pH to near 7 and also causes the precipitation of genomic DNA and proteins into a gloopy mess (snot-like)
7. Centrifuge the bacterial lysate at maximum speed for 5 minutes in a microfuge.
    - Snot-like substances should be tightly packed into a pellet at the bottom of the tube after this step
    - the solution or supernatant contains the plasmid DNA
8. **Column Purification of DNA**: Transfer the supernatant to a fresh tube with silica-membrane column
    - DNA likes to bind to glass under high salt conditions
    - the white membrane is made of a glass fiber
    - 

9. Centrifuge the supernatant through column for 1 minutes at maximum speed in a microfuge.
    - The DNA will be bound to the membrane on the column (silica)
10. **Wash**: Discard flow-through and wash column with 500 µl PE. Centrifuge the supernatant through column for 1 minutes at maximum in a microfuge.
    - PE is a solution that helps to wash away the non-specifically bound substances
11. Discard flow-through. Wash Column with 700 µl PE. Centrifuge the supernatant through column for 1 minutes at maximum in a microfuge. Discard flow-through and repeat spin to dry column.
12. **Elute** the nucleic acids in 50 µl of TE (pH 8.0) by binding for 1 minute and spinning at maximum speed for 1 minute.

Jeremy Seto jseto@citytech.cuny.edu

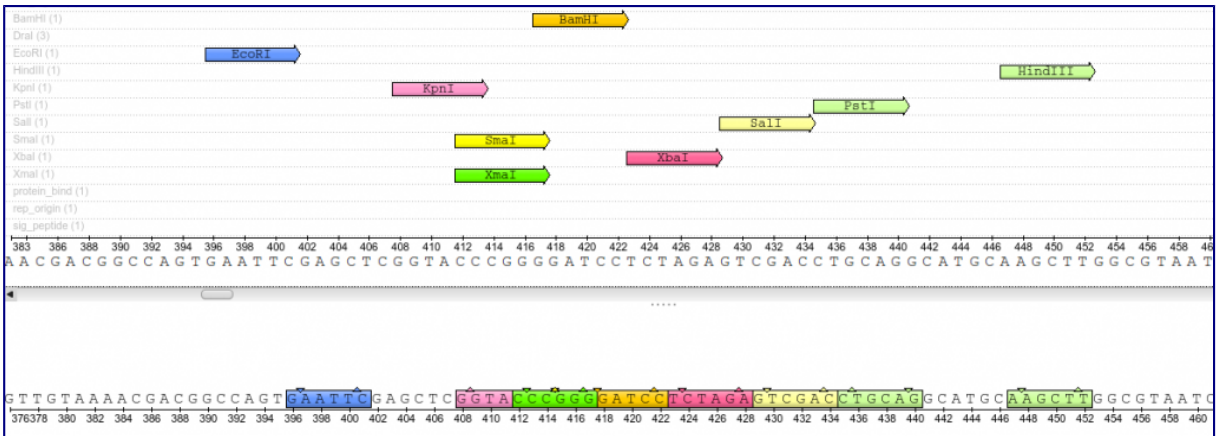
Identification of Plasmid DNA

Once plasmids are isolated, they require identification. Plasmid vectors have known sequences and are mapped of their major features. Knowing the sequence of these pieces of DNA means knowing the locations of RE digestion sites. By using Res, digesting plasmids into known sizes aids in verification of plasmid identity without the need to have the entire plasmid re-sequenced. A common plasmid is called pUC18 or pUC19. The “p” stands for plasmid, the “UC” stand for University of California (where it was designed) and 18 or 19 refer to the difference in the MCS. This plasmid is 2,686 base pairs or ~2.7kb (kilobase) long with a single *EcoRI* site in the MCS. Another plasmid of interest in learning Molecular Biology is called pGlo. This plasmid has a jellyfish gene in the MCS that codes for a protein that will fluoresce green when expressed under UV light. pGlo is 5.4kb long and contains a single *EcoRI* site. Once digested, by an enzyme, these plasmids can be identified based on size separation on an agarose gel. Usually, it is best to identify by using 2 different REs. Digestion is important before size comparisons since circular DNA migrates through agarose differently than linear DNA. Additionally, circular DNA can sometimes be “super-coiled” and lead to very rapid migration despite the size.

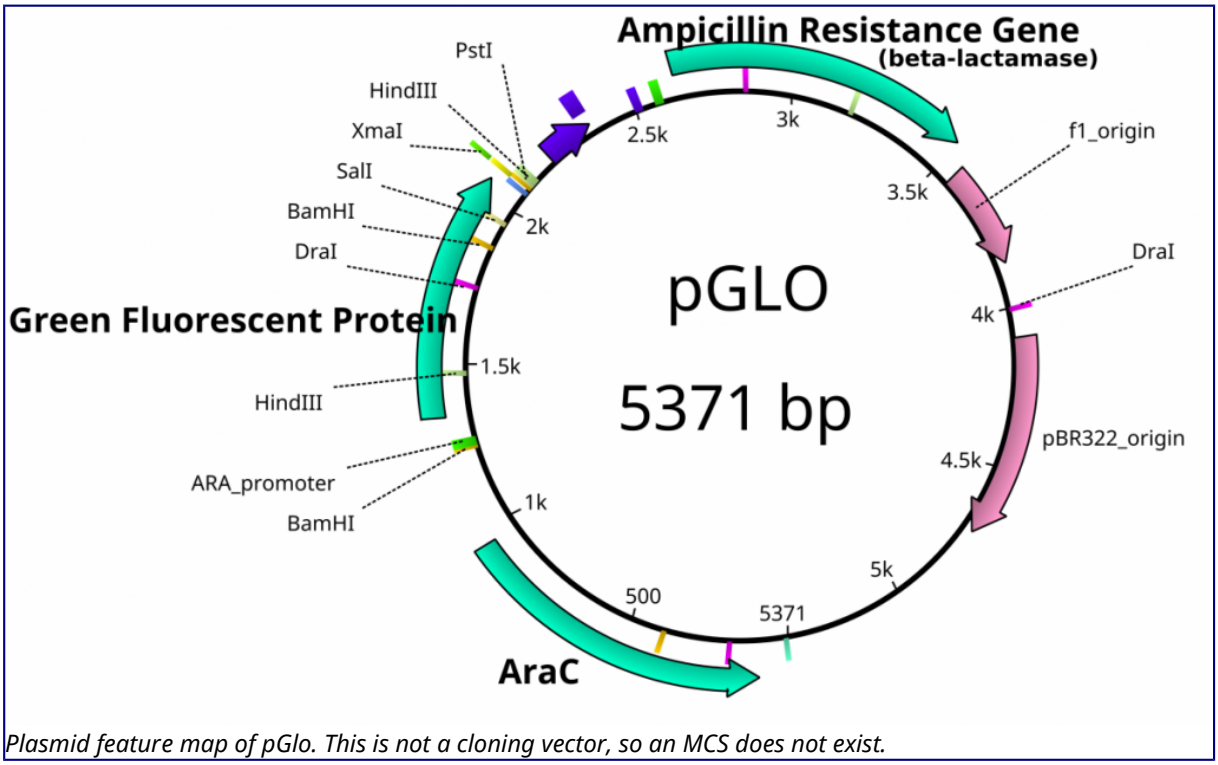


Feature map of pUC19 including the locations of some restriction enzymes.

Jeremy Seto jseto@citytech.cuny.edu



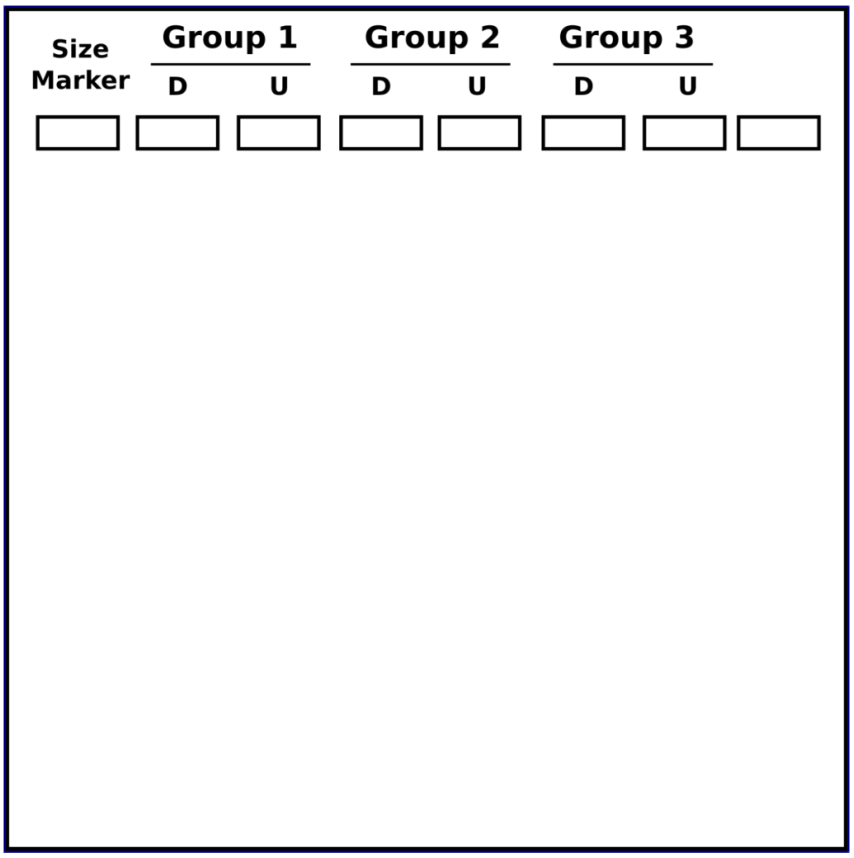
Close up view of the pUC19 Multiple Cloning Site (MCS). These restriction sites appear only once throughout the entirety of the plasmid sequence.



Plasmid feature map of pGlo. This is not a cloning vector, so an MCS does not exist.

Jeremy Seto jseto@citytech.cuny.edu

Exercise 2: Restriction Digestion Identification of Plasmids



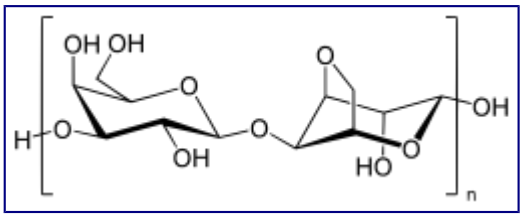
1. The class should prepare 2X 0.8% agarose gels by preparing 0.4g agarose in 50ml TBE buffer
    1. Melt agarose solution by microwaving for 1 minute
    2. Add 5µl Sybr Safe solution into 100ml gel solution
    3. Pour this solution into a casting tray inside the refrigerator
    4. insert a comb
2. To a new tube add 2µl of plasmid DNA to 8µl of *Eco*RI fast digest mixture
    1. 1µl Fast Digest Buffer
    2. 1µl Fast Digest *Eco*RI enzyme
    3. 6µl H_2O
3. Incubate at 37°C for 10 minutes
4. add 2µl of loading buffer to digestion mixture
5. in a separate tube combine 3µl plasmid DNA with 2µl loading buffer and 7µl of H_2O
6. Load gel with an appropriate size ladder in the first lane, in the next lanes load the Digested plasmid, then the undigested plasmid.
    - 3 groups can load onto one gel
    - U=undigested plasmid
    - D=digested plasmid
7. Run gel at 110V for 30 minutes and visualize on a UV transilluminator
8. Document with your camera

Additional Resources

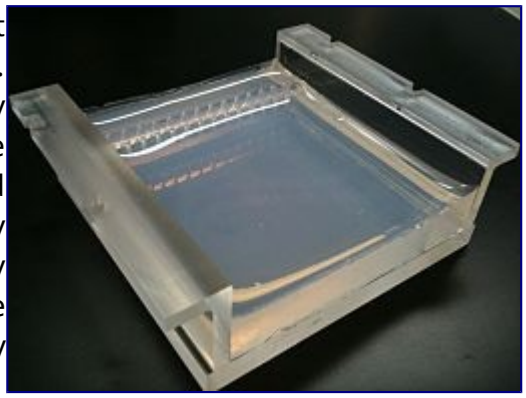
- For help on this problem, please try out the *In silico* digestion activity.

Jeremy Seto jseto@citytech.cuny.edu

Agarose Gel Electrophoresis

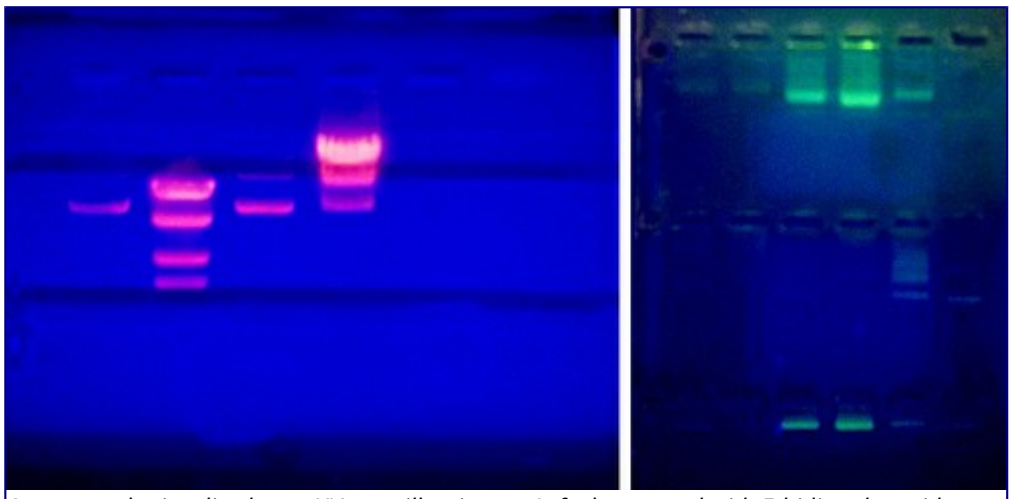


Agarose is a linear carbohydrate polymer purified from the cell walls of certain species of algae. Agar is a combination of the crude extract that contains agarose and the smaller polysaccharide agaropectin. When dissolved and melted in liquid, agarose strands become tangled together to form a netting that holds the fluid in a gel. Reduction of the fluid creates a higher percentage gel that is firmer and contains smaller pores within the netting.



Placing a **comb** within the melted agarose creates spaces that allow for the insertion of samples when the gel is solidified. Molecules can traverse through the pores as they are drawn by electrical currents. Charged compounds will migrate towards the electrode of opposite charge but migration rate will be influenced by the size of the molecules. Smaller compounds can easily traverse through the webbing while larger items are retarded by the pore size. Follow this simulation to get a better idea of how we use **Agarose Gel Electrophoresis** in molecular biology to study DNA fragments.

DNA molecules are not readily visible when **resolved** (separated) on an agarose gel. In order to visualize the molecules, a DNA dye must be administered to the gel. In research labs, a **DNA intercalating agent** called Ethidium Bromide is added to the molten gel and will bind to the DNA of the samples when run. Ethidium Bromide can then be visualized on a UV box that will fluoresce the compound and reveal bands where DNA is accumulated. Since Ethidium Bromide is known as a carcinogen, teaching labs will use a safer DNA intercalating agent known as Sybr Green. This can be visualized in a similar fashion, but will fluoresce a green color instead.



Agarose gels visualized on a UV transilluminator. Left shows a gel with Ethidium bromide. Right shows a gel with Sybr Green.

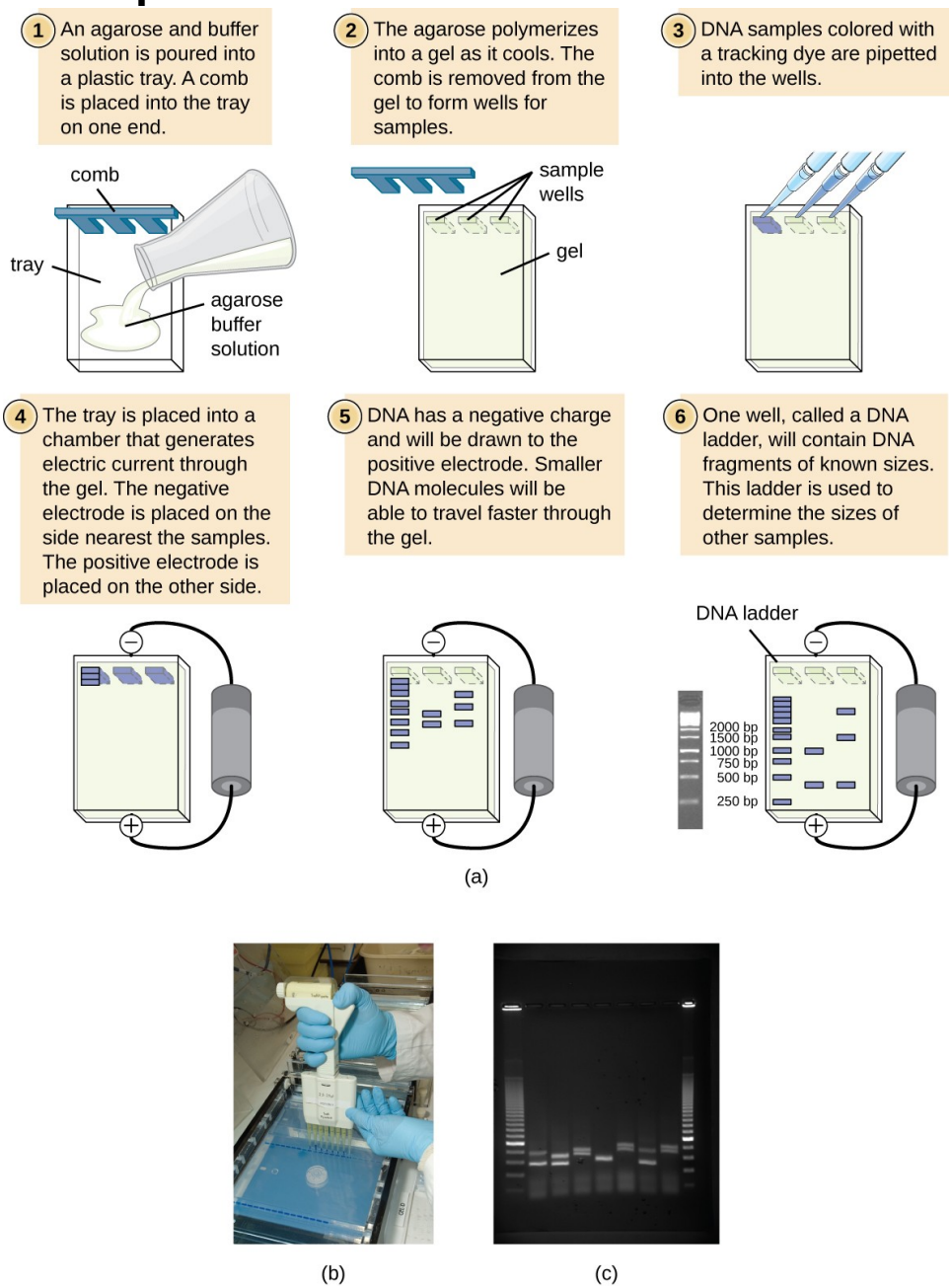
Agarose gels are made of and bathed in a buffered solution, usually of Tris-Borate-EDTA (**TBE**) or Tris-Acetate-EDTA (**TAE**). Regardless of buffer solution, the buffer provides necessary electrolytes for the current to pass through and maintain the pH of the solution.

DNA samples are prepared in a buffer similar to the solution that it will be run in to ensure that the phosphate backbone of the DNA remains deprotonated and moves to the positive electrode. Additionally, **glycerol** or another compound is added to this buffer in

Jeremy Seto jseto@citytech.cuny.edu

order for the solution to sink into the wells without spreading out. A dye is often included in this loading buffer in order to visualize the loading in the wells and to track the relative progression of gel.

Agarose Gel Set-up



External Resources

- Gel Electrophoresis Simulation
- The Structure of DNA http://www.nature.com/scitable/topicpage/discovery-of-dna-structure-and-function-watson-397
- http://learn.genetics.utah.edu/content/science/forensics/

Jeremy Seto jseto@citytech.cuny.edu

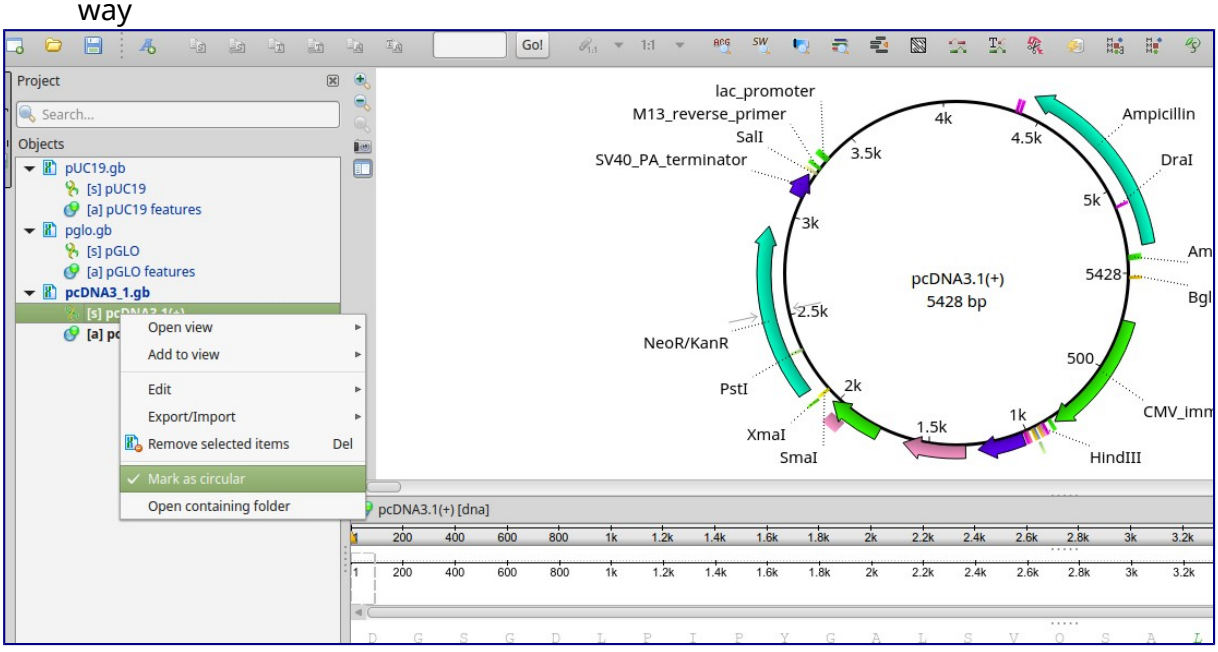
Concept

Restriction enzymes act as molecular scissors. The ones we use in Molecular Biology are those that cut within known sequences that occur often enough, yet rare enough to cut our DNA into analyzable fragments. Molecular Biologists often use 6-cutters. This means that the site of digestion is "restricted" to a recognition sequence of 6 nucleotides. These nucleotides are usually palindromic as discussed before.

Imagine a linear piece of DNA as a piece of string. When cutting the string once, you result in 2 pieces. Now consider a plasmid. This was already discussed to be a circular piece of DNA. With a circle, there are no ends. cutting the plasmid once results in 1 piece of DNA as opposed to 2. Keep this in mind when digesting circular plasmids. In this case, nucleotide 1 is adjacent to and contiguous with the last nucleotide of the sequence.

In silico digestion

1. This activity is meant to supplement Identification of DNA (activity)
2. Launch UGENE and open the following files:
    - **pGlo.gb**
    - **pUC19.gb**
    - **pcDNA3.1.gb**
    - **pUC with Insert**
3. In the **Objects** menu, right-click on the sequences and select "**Mark as circular**"
    - the sequences will now be treated as circular DNA
    - the first nucleotide and the last nucleotide become adjacent as a continuous sequence this way



4. From top menu, select **Actions** → **Analyze** → **Find restriction sites**

Jeremy Seto jseto@citytech.cuny.edu

- this will load a set of default restriction enzymes
- if none are loaded, select them individually (found in alphabetical order)
    - choose: BamHI, BglII, ClaI, DraI, EcoRI, EcoRV, HindIII, KpnI, PstI, SalI, SmaI, XbaI, XmaI, NotI

5. **Actions → Cloning → Digest into fragments**
    - Choose your enzyme
    - Try *Hind*III for each plasmid
    - Fragments will be added to the circular view and annotations for these fragments will be added
    - You can use this information to calculate how many fragments come from enzymes based on how many times they cut and by the nucleotide coordinates found in the annotation of the fragment

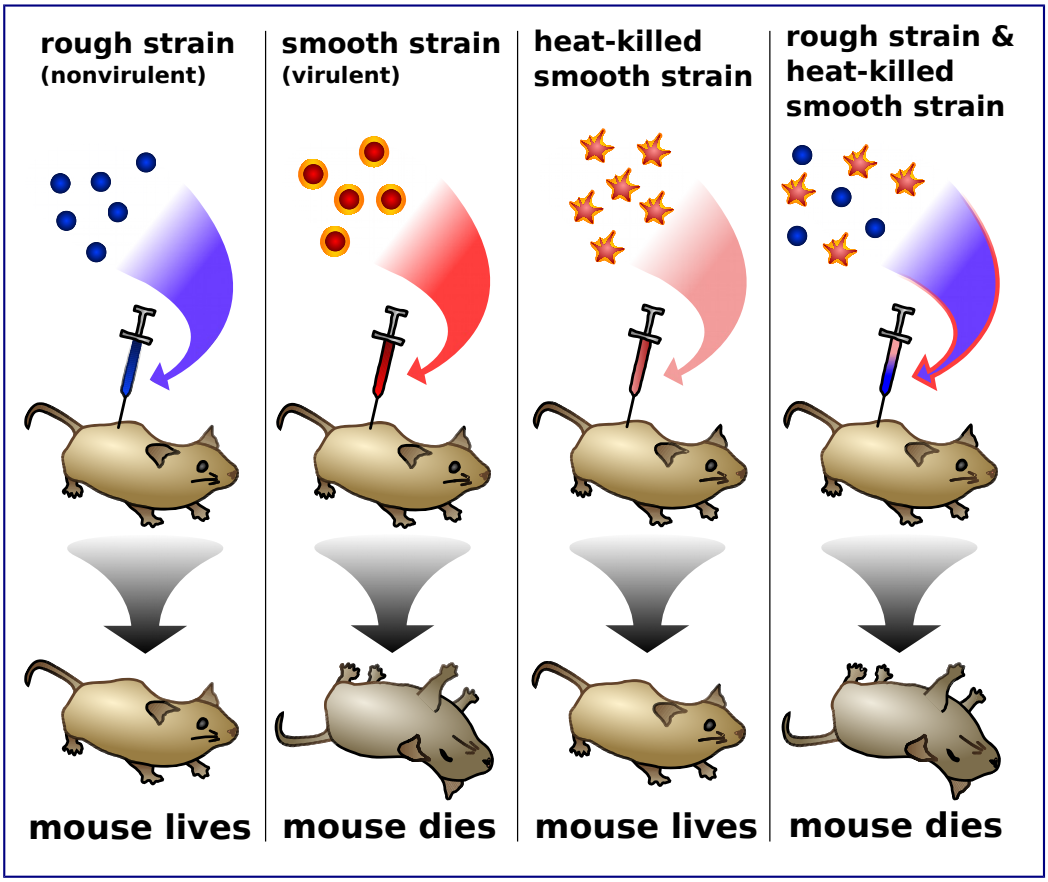
Jeremy Seto jseto@citytech.cuny.edu

History of Genetic Transformation

Any uptake of genetic information from the external environment into cells that results in the expression of new traits is called **genetic transformation**. This process can occur naturally. Some bacteria are referred to as being "competent" to indicate that they are capable of taking DNA into the cell from the environment. This is referred to as **natural competence**. Bacteria are also capable of receiving DNA through the process of conjugation where plasmids from one bacteria are sent to another through the **conjugation pilus**. Other methods of introduction of foreign DNA include direct injection into the cytosol or through the use of viruses in a process called **transduction**. In eukaryotic cells, we refer to the introduction of DNA as **transfection**.

Frederick Griffith and the Transforming Agent

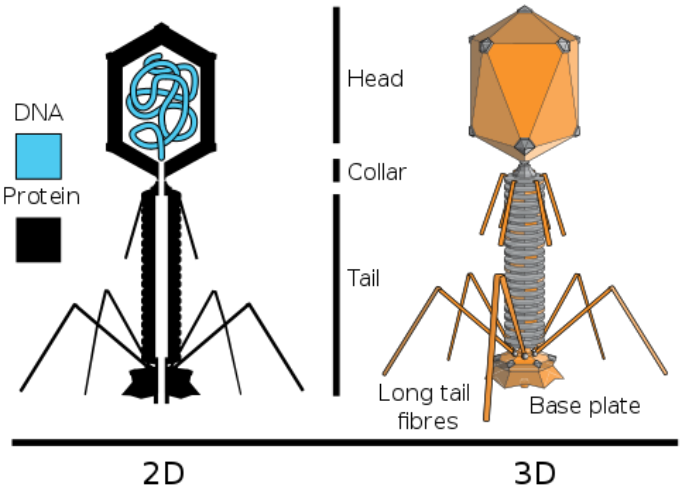
At the beginning of modern biology, the source of genetic material was not known to be DNA. In fact, many scientists thought DNA was too simple to perform this job. Scientists believed that proteins, with their 20 varied amino acids, were the carriers of genetic information. In an attempt to develop a vaccine for a bacterial induced pneumonia, Frederick Griffith was the first to describe the process of genetic transformation by accident in 1928. Griffith took a virulent strain of bacteria (smooth in appearance) that caused pneumonia and injected them into mice. This would result in death of the mice. He also observed that injection of a rough bacteria did not cause any disease. After heat-killing the smooth bacteria, he discovered that living bacteria of the virulent strain was required for the disease to progress. Finally, he observed that injecting the heat-killed virulent bacteria with living bacteria of the non-virulent strain resulted in pneumonia and death in the mice. From this experiment, a **transforming agent** with the capacity to pass on a trait was found to be within the contents of those dead cells. But no one knew this agent to be DNA at that point.



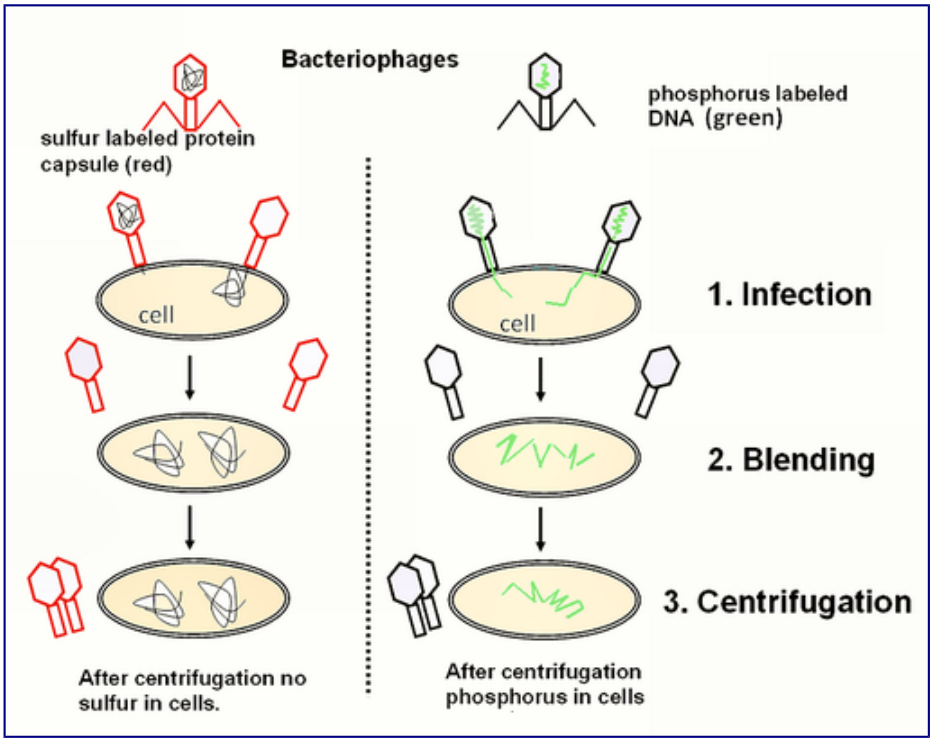
Jeremy Seto jseto@citytech.cuny.edu

Hershey and Chase

Hershey and Chase studied bacteriophage (phage=eater). Phage are bacterial viruses that infect bacteria and cause lysis of the cells. They have a very simple structure of a proteinaceous head/collar/tail and a DNA core. It was known that bacteria infected with phage were resistant to additional infection. In 1952 Hershey and Chase grew bacteriophage in conditions that would specifically label either the DNA or the protein with radioactivity. They subsequently took phage with radiolabeled DNA and infected bacteria. In parallel, they took phage with radiolabeled protein and infected another set of bacteria. After just enough time for infection, the bacterial cultures were placed into a blender to separate the bacteriophage from the bacteria.



T4 bacteriophage anatomy.
Credit: Adenosine, Pbroks13 (CC-BY-SA 2.5)



Solutions were centrifuged to isolate bacteria from the phage. Bacteria were radioactive only when the phage grown in conditions to radiolabel DNA infected the bacteria to indicate that DNA might be the transforming agent.

Questions to think about:

1. What isotope would be used to label protein?
2. What isotope would be used to label just DNA?

Jeremy Seto jseto@citytech.cuny.edu

Avery, McCarty and MacCleod

In 1944, Oswald Avery, Colin MacCleod and Maclyn McCarty repeated Griffith's experiment. Instead of using heat-killed bacteria, these scientists isolated protein, carbohydrates, lipids and nucleic acids from the virulent strain and co-injected with the non-virulent bacteria. Carbohydrate extracts were ineffective at transforming bacteria. Protein extracts were incapable of causing transformation. Lipid injections were unable to result in virulence. Only nucleic acid samples treated with RNase were capable of transforming bacteria. When co-injecting with DNase, bacteria were not transformed. Along with Hershey-Chase, this definitively illustrated that DNA was the transforming agent capable of transferring genetic information.

More Resources

- http://www.dnalc.org/view/16375-Animation-17-A-gene-is-made-of-DNA-.html

Jeremy Seto jseto@citytech.cuny.edu

Bacterial Transformation

Escherichia coli are commensal gram negative bacteria found in the guts of humans. They have the capacity to double every twenty minutes and make a favorable carrier of recombinant DNA. Plasmid DNA can be introduced into *E. coli* easily after making them competent. One method to achieve this is through chemical competence with heat shock. In this process, the bacteria are incubated in $CaCl_2$ solution on ice. The cold serves to slow down molecular motion of the plasma membrane while the Ca^{2+} ions remove the charge-charge repulsion between the phospholipids and the negatively charged DNA seeking to gain entry into the cell. Cells are placed for a short period of time at 42°C to induce **heat shock**. This heat shock results in the cell taking up the DNA. This method is very low efficiency so many bacteria do not take in any DNA. Cells are allowed to recover from heat shock at 37°C in rich nutrient broth to allow for the production of the antibiotic resistance proteins encoded on the vector as a selection marker. Transformed cells are then spread across an agar plate containing the antibiotic which will then kill all non-transformed cells. Only the bacteria containing the vector with the antibiotic resistance gene will survive and replicate to form small colonies on the surface of the agar.

- http://www.dnalc.org/view/15918-Transformation.html
- http://www.dnalc.org/view/15916-DNA-transformation.html

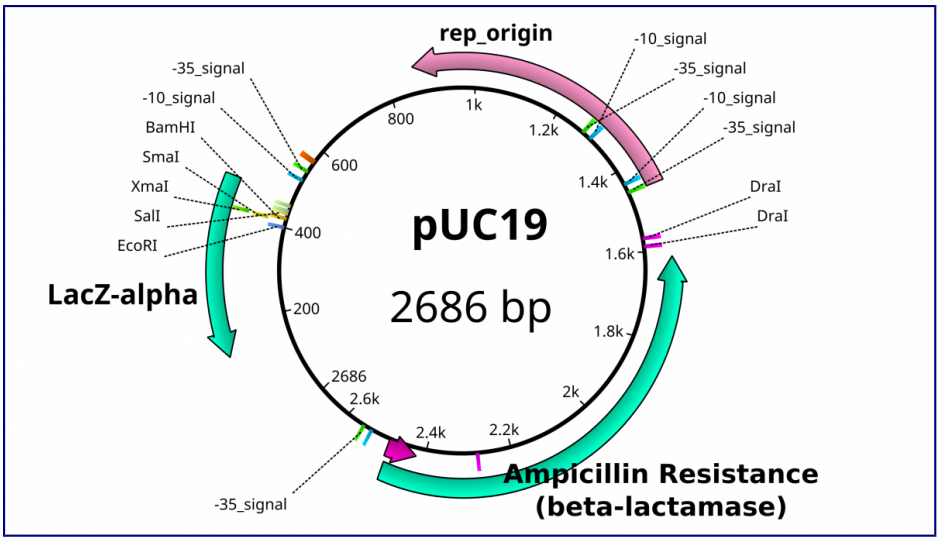
Exercise: Transformation of Bacteria with RE Identified Plasmids

1. Each group retrieves the 2 miniprepped plasmids from the previous week in the freezer and allow to thaw on ice.
2. Bring 2 agar plates to room temperature
    - 2 plates will contain antibiotic, X-Gal and arabinose
3. For each plasmid, obtain 250µl of transformation buffer (50mM $CaCl_2$) in microfuge tubes and place on ice for 10 minutes
4. Take an inoculating loop and remove a single colony of bacteria from a freshly streaked plate grown overnight
5. Swirl bacteria in each tube containing transforming solution to distribute bacteria throughout solution
6. Pipette 5 µl of plasmid into the tube and incubate on ice for 10 minutes
7. During this incubation, flip the warmed plates and label them with your group names.
8. Place transformation tubes into 42°C heatblock for 1 minute to heat shock the cells
9. Add 500µl fresh SOC media (or LB) and incubate at 37°C for 15 minutes.
10. Pipette 150µl of transformation solution onto each plate and spread across the plate.
11. Turn plates agar side up and place them into 37°C incubator overnight. (your instructor will retrieve them and place them into refrigerator)

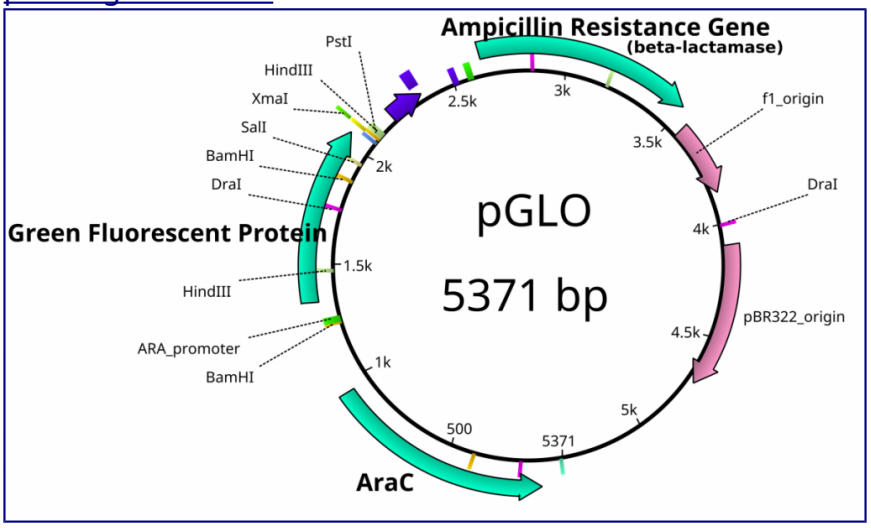
Hypothesize: What will I expect of my transformed cells?

From the previous lab, we can identify our plasmids. The plasmids are either pGlo, pUC18/19 or pUC18/19 with a 6kb insert disrupting the LacZ gene. pGlo contains a gene that encodes the protein GFP that will fluoresce green under UV light and is 5.4kb. pUC is typically 2.7kb in size. LacZ is a gene encoding the protein β-Galactosidase, the enzyme that hydrolyzes lactose into the monosaccharides galactose and glucose. X-Gal is a chemical resembling lactose, however upon hydrolysis, the molecule deposits a blue coloring into the cell.

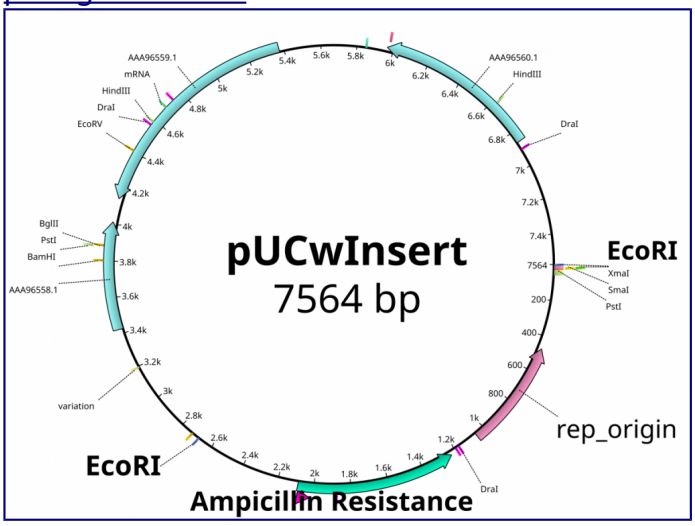
Jeremy Seto jseto@citytech.cuny.edu



- pUC19 genbank file



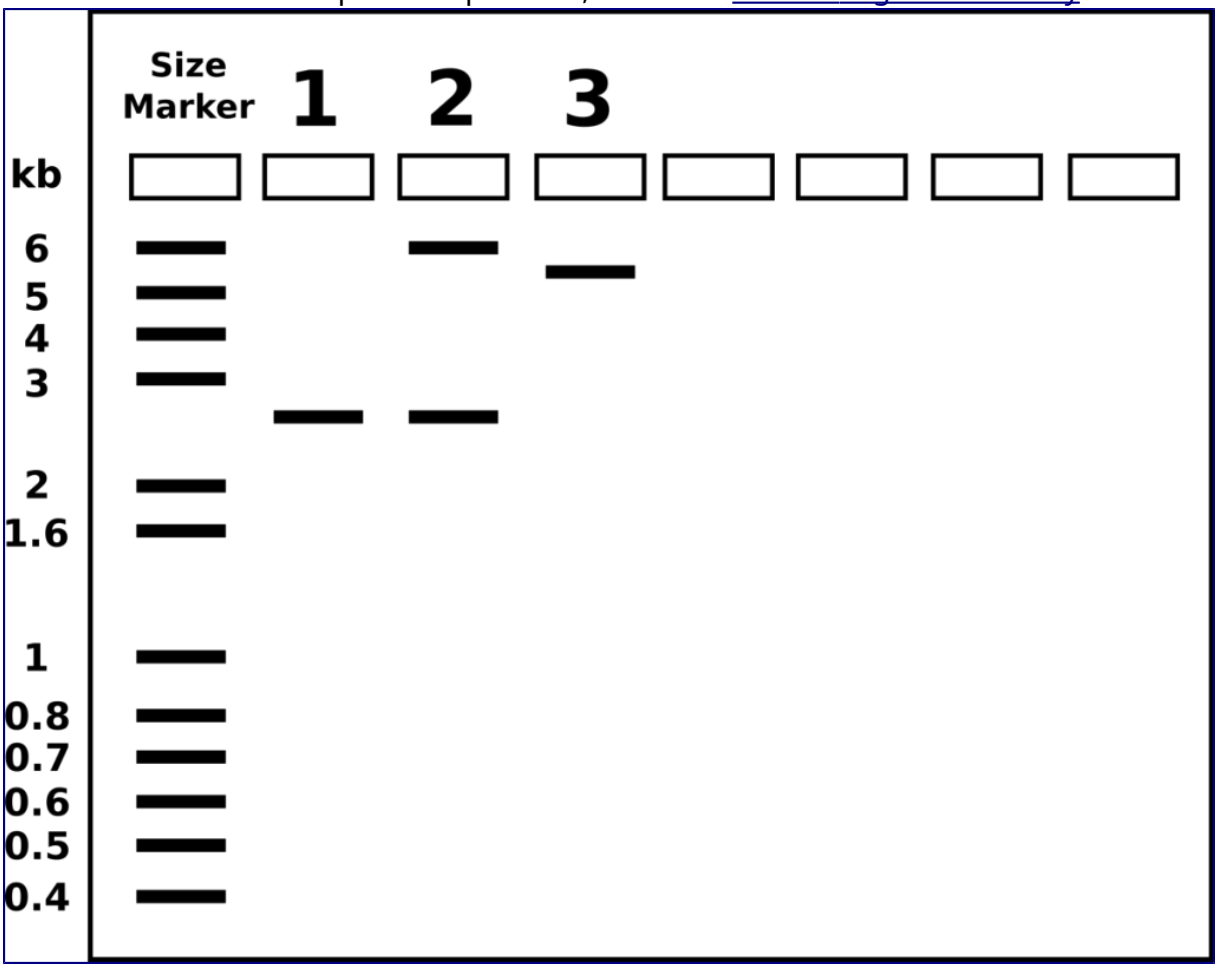
- pGlo genbank file



- pUC with insert file

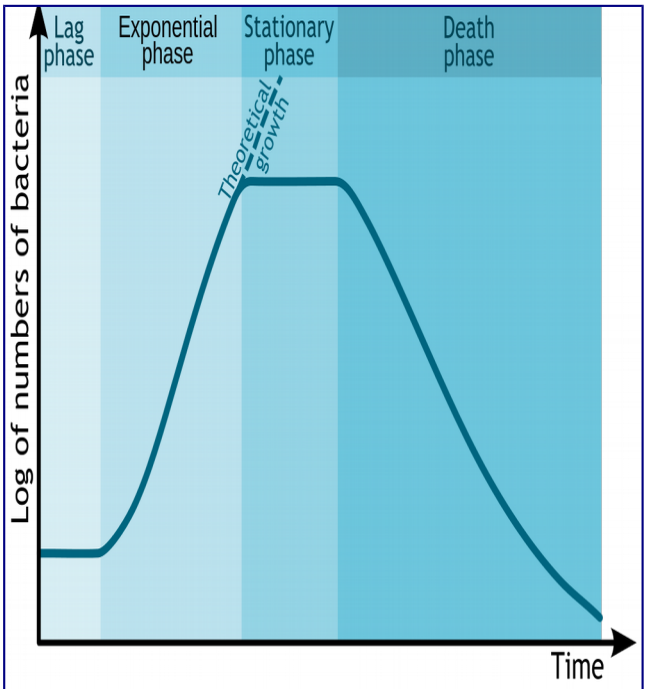
Jeremy Seto jseto@citytech.cuny.edu

1. If the previously mentioned plasmids were digested by *EcoRI*, label the lanes below with the appropriate plasmid (pGlo, pUC, pUC-inserted)
2. Predict if your transformants will be green under UV, white in all conditions or blue.
3. For additional help on this problem, utilize the *In silico* digestion activity



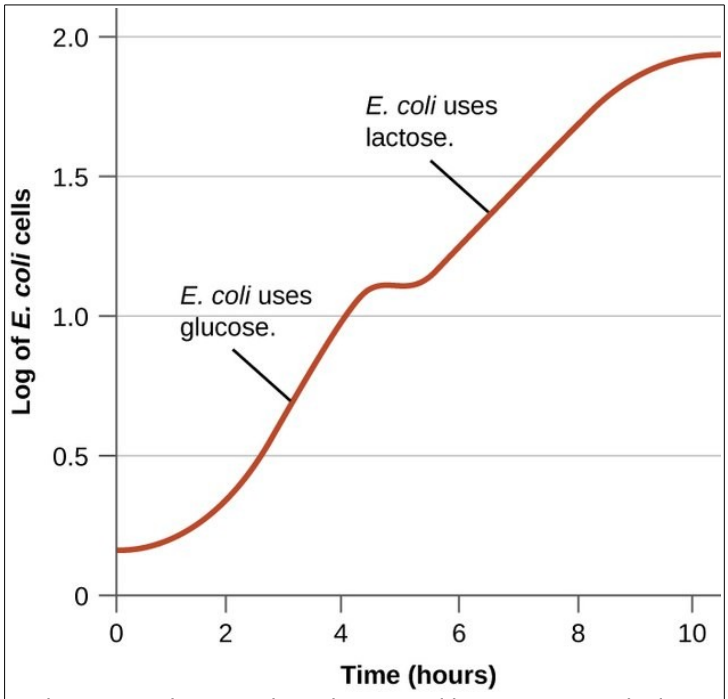
Jeremy Seto jseto@citytech.cuny.edu

The Lactose Intolerance of Bacteria



The standard growth kinetics of E. coli are described by the curve.

Credit: Michał Komorniczak (CC-BY-SA 3.0)

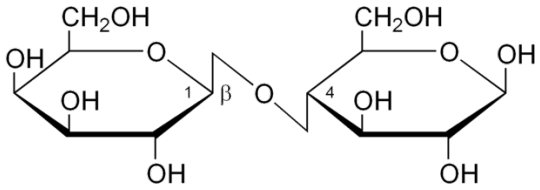


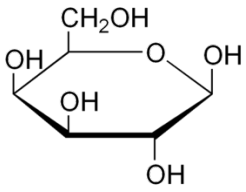
Biphasic growth curve when glucose and lactose are supplied to media

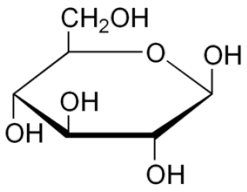
Credit: CNX OpenStax (CC-BY 4.0)

Glucose is the preferred energy source of cells. François Jacob and Jacques Monod sought to understand how bacteria made decisions to switch between different sugars as sources of energy. Jacob and Monod found that if glucose and lactose were presented as food for bacteria, there would be a **biphasic growth** pattern.

Jacob and Monod came to understand that the glucose would first be utilized (preferred source) and the lactose would be digested after the depletion of glucose. This occurred because, under normal situations the bacteria would not have access to lactose and would waste energy by creating enzymes to digest it. The enzyme **β-galactosidase**, which is responsible for digesting lactose to the monomers galactose and glucose would only be induced under the conditions of low glucose and high lactose. Monod found that when lactose was the sole sugar, the expression of the enzyme β-galactosidase was induced and displayed a monophasic growth with a delay.



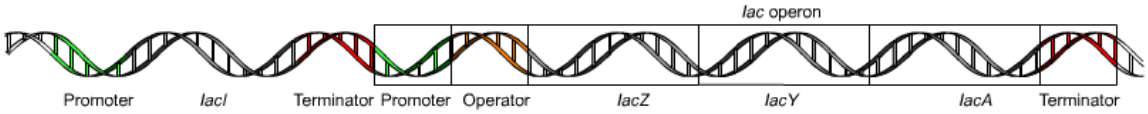




Jeremy Seto jseto@citytech.cuny.edu

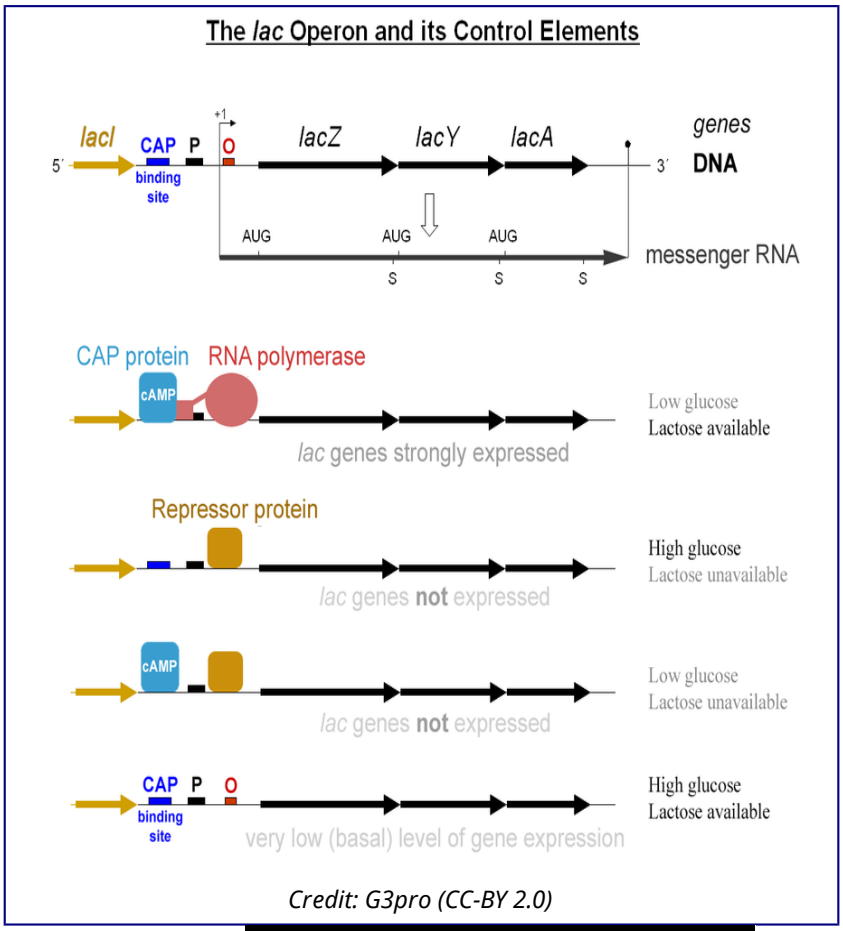
The Lac Operon

Jacob and Monod later found that the genes involved in utilizing lactose were clustered together in proximity under a coordinated control mechanism. This became known as the **Lac operon**.

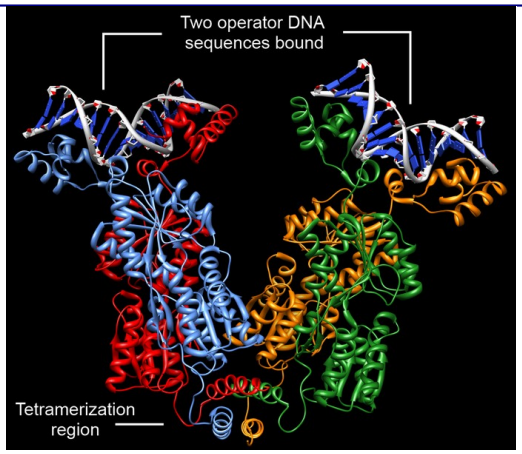


A schematic of the Lac Operon. LacZ, LacY and LacA are transcribed as a single mRNA.

The usage of lactose as a source of energy is preferred by bacteria when glucose is not present. In the presence of abundant glucose, it would be a waste of energy and cellular resources to commit to synthesizing the mRNA and the protein for β-galactosidase. Unless lactose is present, a protein binds to a portion of the Lac promoter referred to as the **operator**. This **repressor** protein is encoded by another gene (***LacI***) outside of the gene cluster. Occasionally, the repressor unbinds to the operator and RNA Polymerase is permitted to transcribe the ***LacZ*** gene (β-galactosidase), ***LacY*** gene (permease), and ***LacA*** gene (acetylase).This "leakiness" of expression is important since the permease protein is needed on the surface of the cell to allow lactose into the cell if it is present in the environment. The presence of lactose causes the repressor to fall off the operator to grant RNA pol access to the DNA. When glucose is low, a protein called **CAP** (Catabolite Activated Protein), binds to the Lac promoter and works as an recruiter of RNA pol. The coordinated effects of CAP activation and Lac Repressor inactivation yields high transcription of the operon.

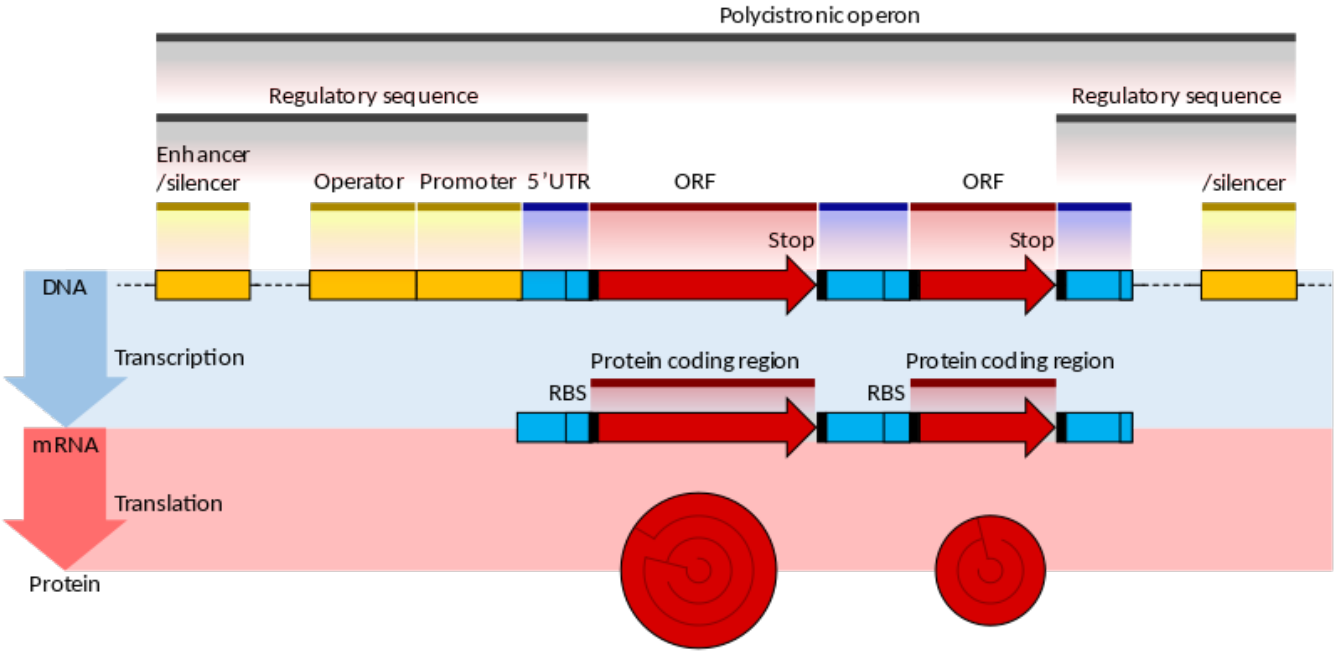


Credit: G3pro (CC-BY 2.0)



LacI bound to 2 DNA operator sequences.
Credit: SocratesJedi (CC-BY-SA 3.0)

Jeremy Seto jseto@citytech.cuny.edu

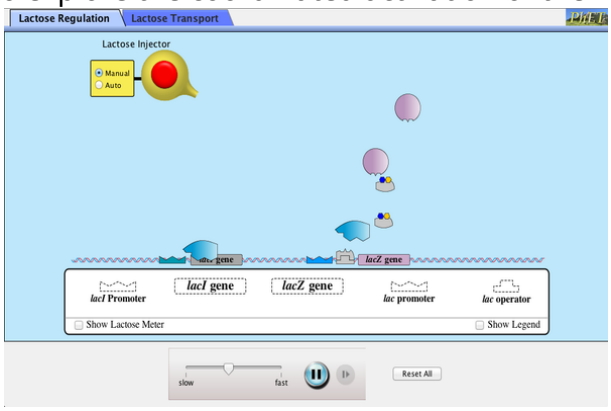


Central Dogma of Molecular Biology in Prokaryotes.

Credit: Thomas Shafee (CC-BY 4.0)

Lac Operon Simulation

Launch the simulation below to explore the coordinated activation of the Lac Operon.

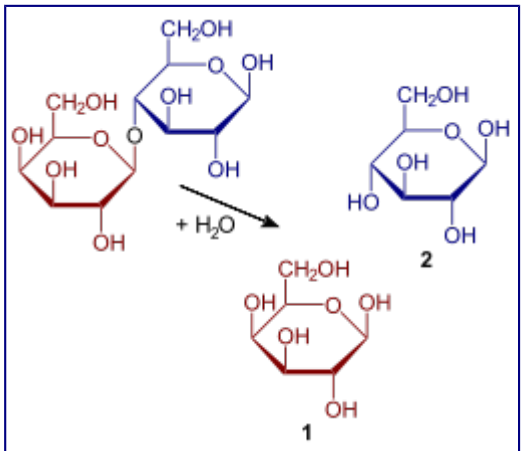


Jeremy Seto jseto@citytech.cuny.edu

LacZ as a reporter gene

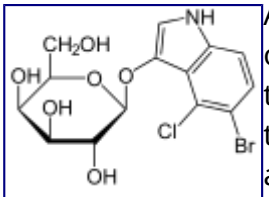
pUC19 contains LacZ DNA as a reporter gene to illustrate the presence of the functioning gene. Transcription of this gene is driven by the binding site for the RNA Polymerase subunit called **σ factor**. The σ factor binding site determines the directionality of the RNA polymerase, since there is an option of transcribing in 2 directions. The standard σ factor binding site is often denoted as **-35 TTGACA...TATAAT -10** from the transcription initiation.

The multiple cloning site within the plasmid provides a convenient location to shuttle a foreign piece of DNA. When no foreign DNA is inserted to this space, the LacZ gene product β-galactosidase is functional. Disruption of the reading frame for this gene likewise disables the functional product from being translated. By using chemical reporters, the integrity of this gene can be confirmed through enzymatic activity.



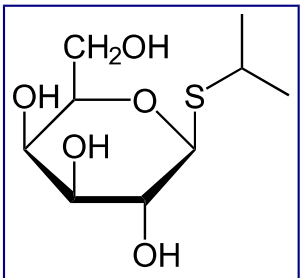
Hydrolysis of lactose to galactose and glucose

Two chemical reporters used to reveal the presence of a functioning LacZ are **X-Gal** (5-bromo-4-chloro-3-indolyl-β-D-galactoside5-bromo-4-chloro-3-indolyl-β-D-galactoside) and **ONPG** (orthonitrophenol).

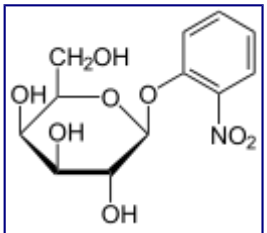


X-Gal (5-bromo-4-chloro-3-indolyl-β-D-galactoside) yields a blue color when cleaved by β-galactosidase

As in the case of the Lac operon, the *LacI* (repressor protein) will occupy the operator. This operator happens to be overlapping the -35 & -10 sequences. In order to fully activate these genes, the Lac repressor must be removed by binding to a lactose analog. In this case, the chemical **IPTG** (Isopropyl β-D-1-thiogalactopyranoside) is used since it is a non-cleavable analog that will perpetually bind to the Lac repressor.



IPTG (Isopropyl β-D-1-thiogalactopyranoside)

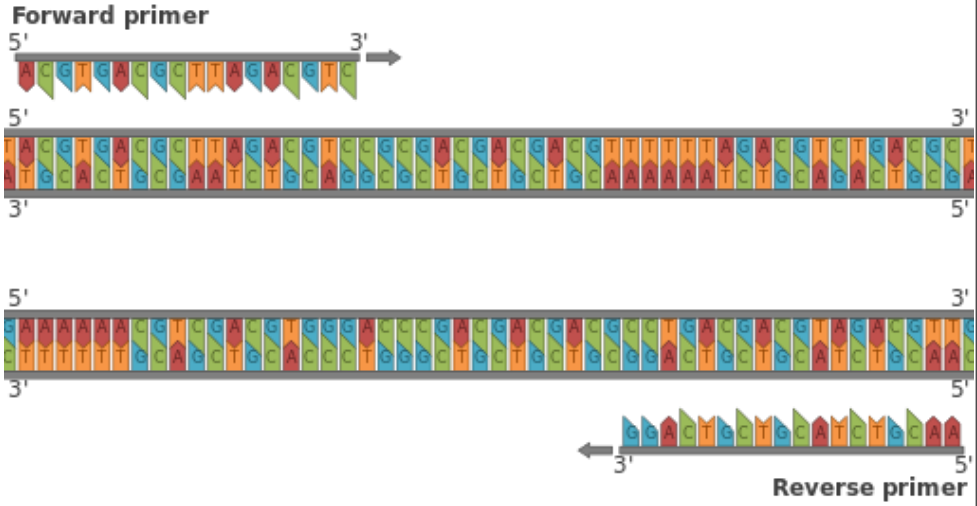


ONPG (orthonitrophenol) yields a yellow color upon cleavage by β-galactosidase

Jeremy Seto jseto@citytech.cuny.edu

DNA Replication
Polymerase Chain Reaction (PCR)

The **Polymerase Chain Reaction** (**PCR**) is a method of rapidly amplifying or copying a region of DNA in a tube. As the name implies, the technique uses a thermostable **DNA Polymerase** enzyme to mimic in a tube what happens within a cell during DNA replication. The chain reaction permits us to rapidly copy DNA from very minute source material in an exponential way. This technique is used in forensic science, genetic testing and cloning of rare genes. Because of the exponential copying process, a stray cell left behind can provide enough genetic material to make billions of copies of this DNA. The process of PCR can be observed in an animation found at Cold Spring Harbor Laboratory's DNA Learning Center website (http://www.dnalc.org/resources/3d/19-polymerase-chain-reaction.html).



Primers are reverse complimentary to one of the 2 strands of DNA. They flank the area of interest and become incorporated into the replicated product.

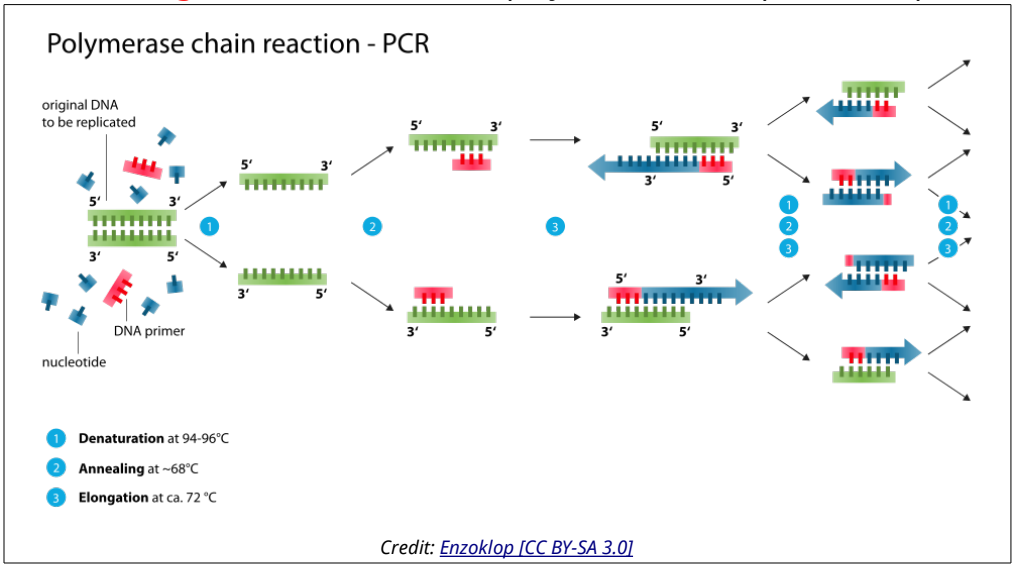
Credit: Zephyris [CC BY-SA 3.0 or GFDL]

As with any DNA replication process, one needs to start off with a **template**.The template is the source material that is meant for duplication. In this process, scientists are not interested in copying the entirety of the genome, just a small segment of interest. DNA polymerases require primers to begin the polymerization process. **Primers** are designed as small oligonucleotide segments that flank the area of interest. These are short strands of DNA that reverse complement to the DNA area of interest so that the DNA polymerase has a starting point and is guided only to the DNA segment of interest. These primers tend to be about 18-24 bases long.

However, a double stranded DNA molecule is already base paired together into a double helix so our primers can not interact. The first step of PCR is to separate the double-stranded DNA molecule by **denaturing** the H-bonds using high heat (95°C). The primer concentrations are much higher than the original template. The next step of PCR is called **annealing**. During this step, the temperature is reduced to a temperature of about 55°C. This temperature is still hot by our standards, but is necessary to enhance the stringency of the correct base pairing of the primers to their targets on the template. The DNA Polymerase used in this process is derived from a bacteria that lives in very high temperatures

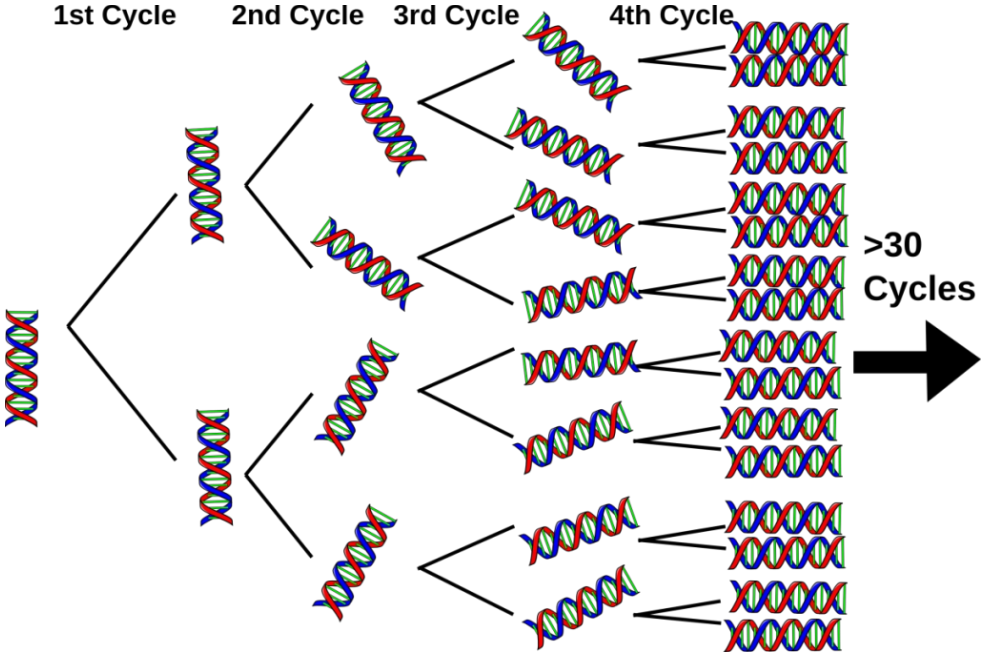
and does not denature as other proteins would under such conditions (thermostable). The original enzyme was isolated from an organism called *Thermus aquaticus*, so we call the enzyme Taq polymerase or just **Taq** for short. This bacteria lives in hot springs where the temperatures are about 50°C but it thrives at a range between 50-80°C. The temperature is raised again to a higher temperature of 72°C for the polymerase to **extend** (also called **elongation**) or continue the polymerization step from the primer.

Within this tube are all the components for the polymerase to act appropriately including buffer to maintain the pH, divalent cations like Mg^{2+}, primers and the supply of nucleotide monomers – **dNTPs** or deoxynucleoside triphosphates (dATP, dCTP, dTTP, dGTP).



Credit: Enzoklop [CC BY-SA 3.0]

PCR is accomplished by cycling rapidly between these three steps: denature, anneal, extension. The rate limiting step is the extension which limits the length of DNA to be copied. If the original template is only a single copy, then after the completion of a cycle, we would have 2 copies. The subsequent cycle would have 4 copies, then 8, then 16, 32, and so on. The doubling process is exponential so from 1 copy undergoing 30 cycles; we would have 2^{30} or 1,073,741,824 copies. This is over a billion copies in a few hours of time.



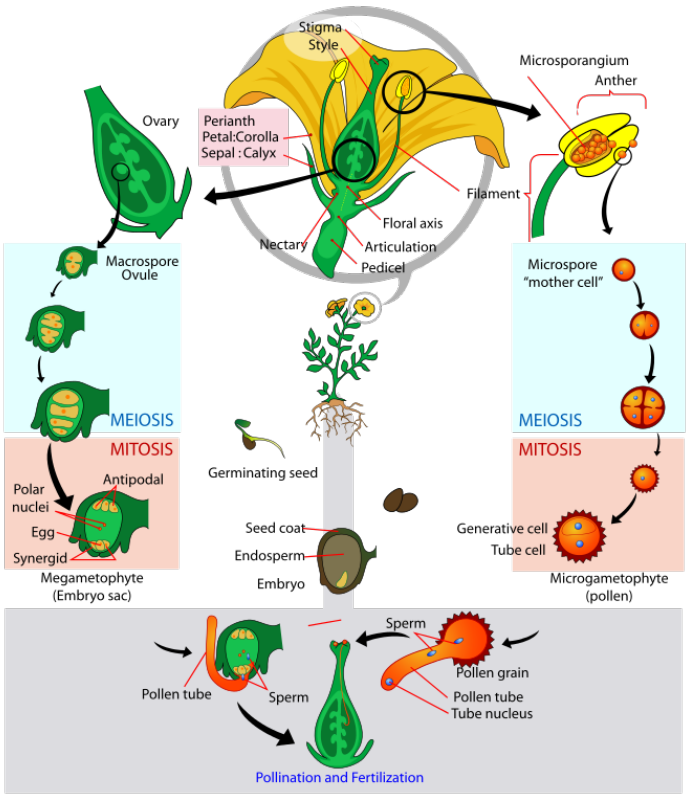
Writing the Rules of Heredity

In the mid 1800's, an Augustinian friar named Gregor Mendel formalized quantitative observations on heredity in the the pea plant. He undertook hybridization experiments that utilized purebred or ***true breeding*** plants with specific qualities over many generations to observe the passage of these traits. Some of these physical traits included: seed shape, flower color, plant height and pod shape.

The pea plant (*Pisum sativum*) offered a great advantage of being able to control the fertilization process and having large quantities of offspring in a short period of time. In a simple experiment of tracking the passage of a single trait (**monohybrid cross**) like flower color through multiple generations he was able to formulate rules of heredity. In this case, pea plants either produced white flowers or purple flowers for many generations (true breeding purple flower or true breeding white flower). These true breeding plants are referred to as the **Parental Generation** (**P**). By removing the male parts of the pea flower (anthers containing pollen), Mendel was able to control for self-pollination. The hybridization came from applying the pollen from one true breeding plant to the female part (the pistil) of the opposite true breeding plant. The subsequent offspring are referred to as the **First Filial Generation** (**F_1**). In the first generation, all flowers are purple. Permitting self-pollination generates a **Second Filial Generation** (**F_2**). This generation sees the re-emergence of the white flowered plants in an approximate ratio of 3 purple flowered to 1 white flowered plants.

Pea flowers (Pisum sativum)

Credit: Bmdavll [GFDL or CC BY-SA 4.0-3.0-2.5-2.0-1.0]

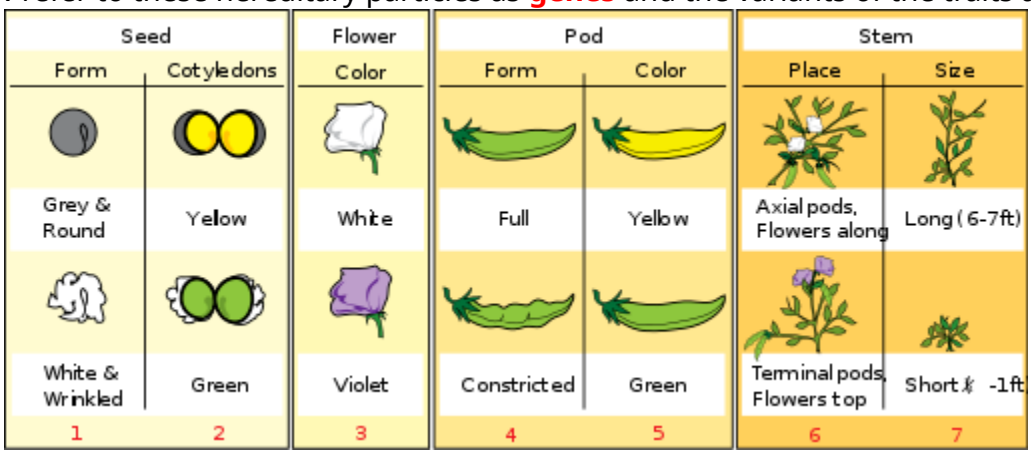


Male and female parts of flowers. Mendel removed the anthers containing pollen to prohibit self-pollination and selectively applied the pollen to stigmas in order to control the "hybridization".

Credit: LadyofHats Mariana Ruiz [Public domain]

Jeremy Seto jseto@citytech.cuny.edu

The loss of one variant on the trait in the F_1 plants with the re-emergence in the F_2 prompted Mendel to propose that each individual contained 2 hereditary particles where each offspring would inherit 1 of these particles from each parent. Furthermore, the loss of one of the variants in the F_1 was explained by one variant masking the other, as he explained as being **dominant**. The re-emergence of the masked variation , or **recessive** trait in the next generation was due to the both particles being of the masked variety. We now refer to these hereditary particles as **genes** and the variants of the traits as **alleles**.



Credit: Ladyof Hats [Public Domain]

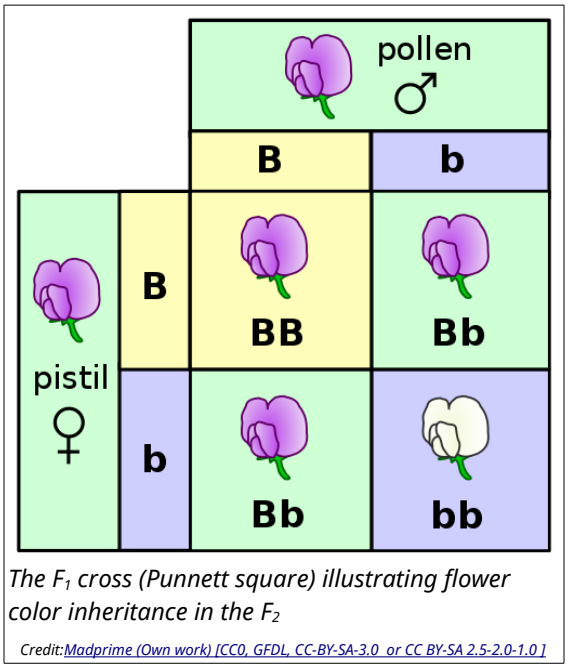
Mendel's Rules of Segregation and Dominance

The observations and conclusions that Mendel made from the monohybrid cross identified that inheritance of a single trait could be described as passage of genes (particles) from parents to offspring. Each individual normally contained two particles and these particles would separate during production of gametes. During sexual reproduction, each parent would contribute one of these particles to reconstitute offspring with 2 particles. In the modern language, we refer to the genetic make-up of the two "particles" (in this case, alleles) as the **genotype** and the physical manifestation of the traits as the **phenotype**. Therefore, Mendel's first rues of inheritance are as follows:

1. **Law of Segregation**
    - During gamete formation, the alleles for each gene segregate from each other so that each gamete carries only one allele for each gene
2. **Law of Dominance**
    - An organism with at least one dominant allele will have the phenotype of the dominant allele.
    - The recessive phenotype will only appear when the genotype contains 2 recessive alleles. This is referred to as **homozygous recessive**
    - The dominant phenotype will occur when the genotype contains either 2 dominant alleles (**homozygous dominant**) or on dominant and one recessive (**heterozygous**)

Jeremy Seto jseto@citytech.cuny.edu

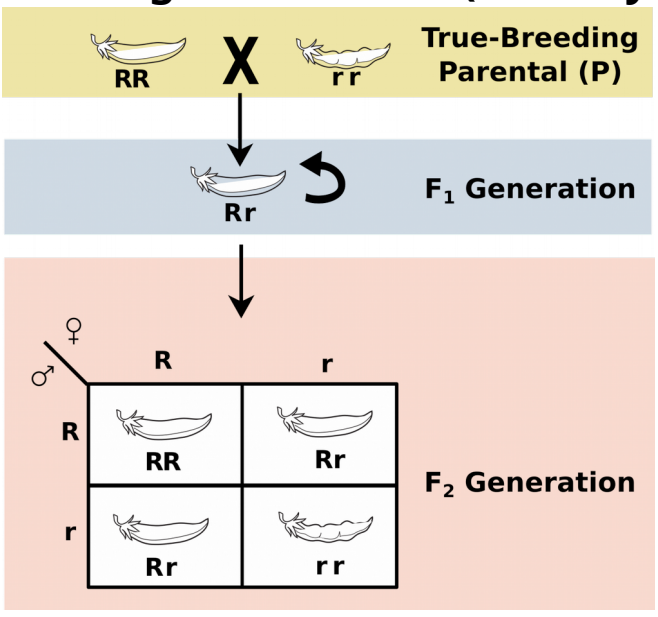
The Punnett Square is a tool devised to make predictions about the probability of traits observed in the offspring in the F_2 generation and illustrate the segregation during gamete formation.



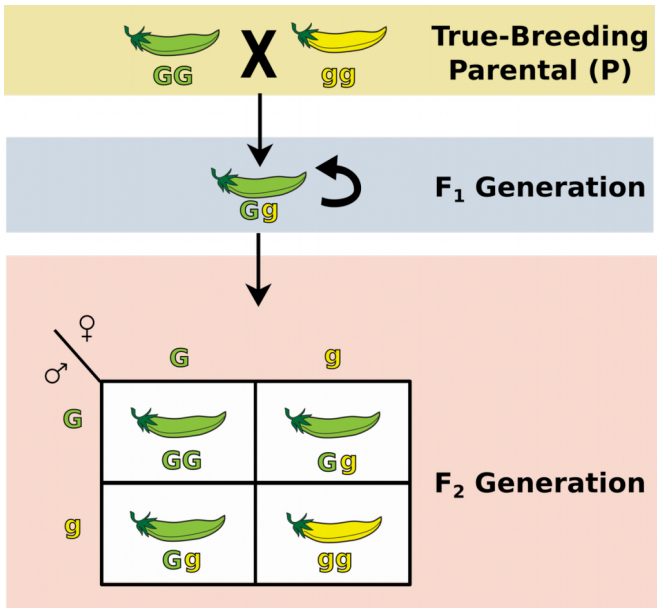
The F_1 cross (Punnett square) illustrating flower color inheritance in the F_2

Credit:Madprime (Own work) [CC0, GFDL, CC-BY-SA-3.0 or CC BY-SA 2.5-2.0-1.0]

The Single Trait Cross (Monohybrid Cross)



Monohybrid cross (one trait cross) observing the pod shape of peas

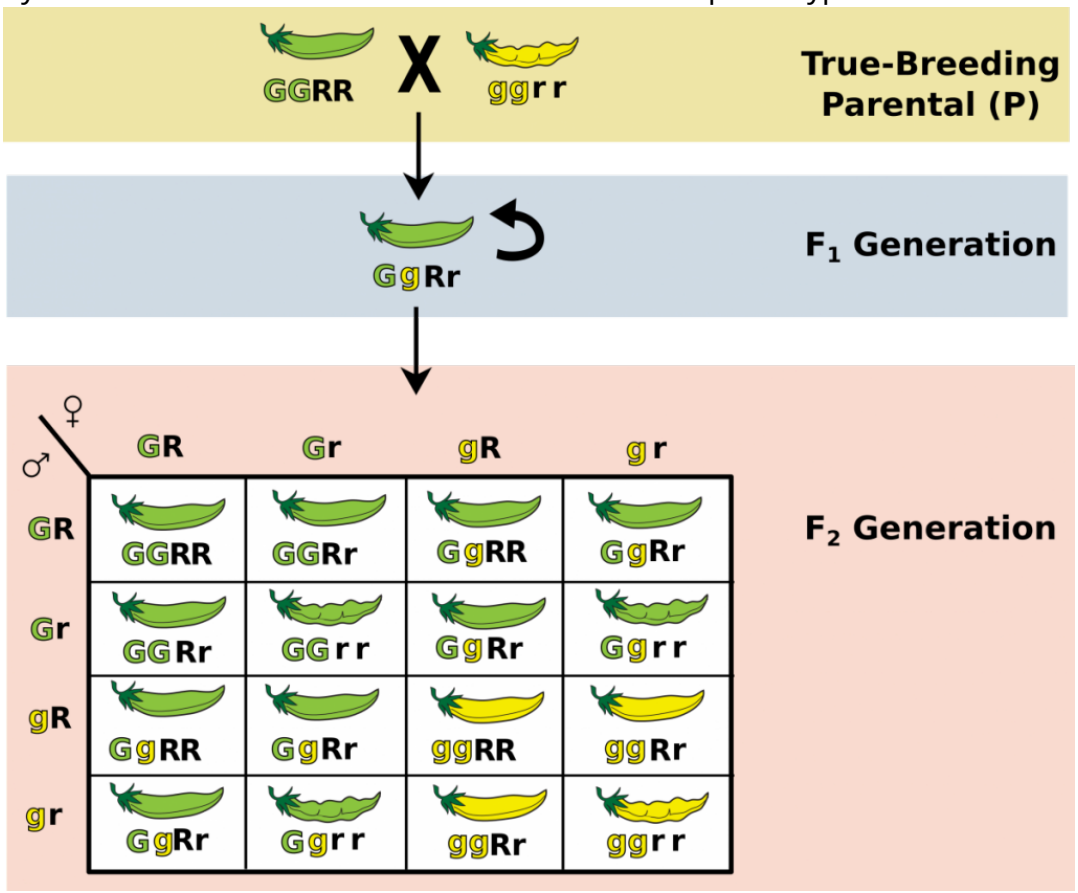


Monohybrid cross (on trait cross) observing the pod color of peas.

Jeremy Seto jseto@citytech.cuny.edu

The Two Trait Cross (Dihybrid Cross)

Mendel continued his experimentation where he looked at two traits. These two trait crosses are called **dihybrid crosses**. While the monohybrid cross would yield 3:1 ratios of the phenotypes, the dihybrid crosses would yield 9:3:3:1 ratios of all the combinations of each phenotype.



Mendel's Rule of Independent Assortment

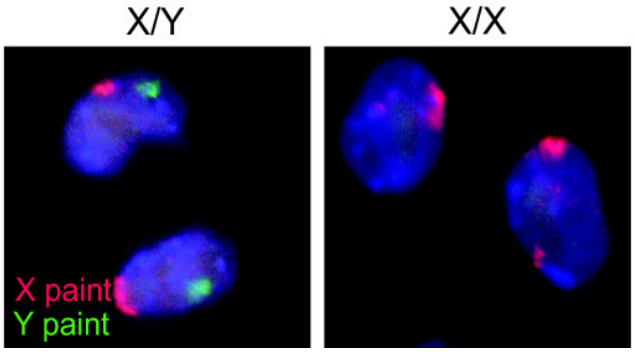
The dihybrid cross revealed another law of inheritance to Mendel. By observing the 9:3:3:1 ratio, Mendel concluded that traits were not tied to each other. That is to say, if a pea pod was yellow, it could still be either smooth or wrinkled in texture. This lack of linkage between genes yielding different characteristics was dubbed the **Law of Independent Assortment**. Genes for different traits can segregate independently during the formation of gametes.

Jeremy Seto jseto@citytech.cuny.edu

Sex-Linked Genes

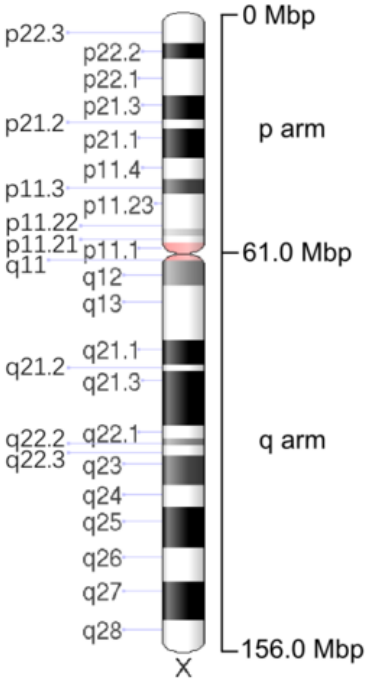
For the most part, mammals have gender determined by the presence of the Y chromosome. This chromosome is gene poor and a specific area called sex determining region on Y (**SRY**) is responsible for the initiation of the male sex determination. The X-chromosome is rich in genes while the Y-chromosome is a gene desert. The presence of an X-chromosome is absolutely necessary to produce a viable life form and the default gender of mammals is traditionally female.

Chromosomal painting techniques can reveal the gender origin of mammalian cells. By using fluorescent marker sequences that can hybridize specifically to X or Y chromosomes through Fluorescence In Situ Hybridization (FISH), gender can be identified in cells.



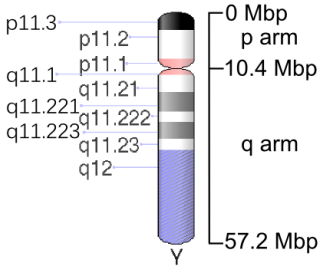
The male cells have an X and a Y while the female cells have X and X combination.

Credit: Janice Y Ahn, Jeannie T Lee [CC BY 2.0]



The human X-chromosome

Credit: NCBI [Public Domain]

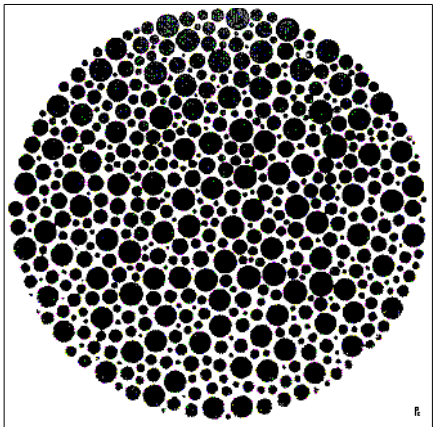


The minuscule human Y-chromosome

Credit: NCBI [Public Domain]

Jeremy Seto jseto@citytech.cuny.edu

Ishahara tests (Activity)



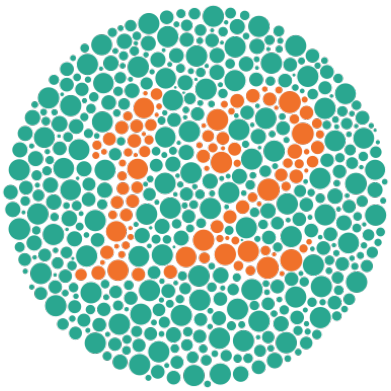
Monochromatic representation of Ishahara test to a color blind individual as it emerges to something visible to a color-sighted individual

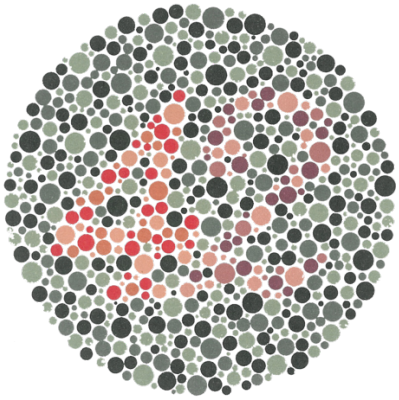
Credit: [Public domain]

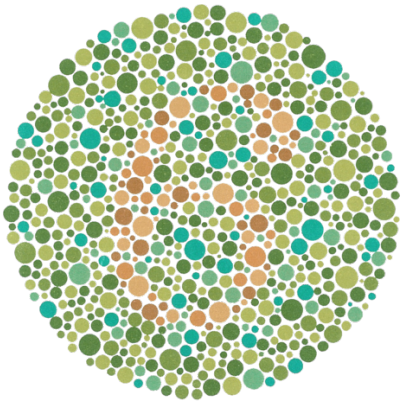
The genes encoding photoreceptor proteins for the long wave-length (reds) and middle wave-lengths (greens) reside on the X chromosome at Xq25. Since the Y-chromosome is not homologous, any mutation to either of these genes that render them non-functional results in an inability to perceive either of those colors. Men are more susceptible to the condition of red-green colorblindness since they are **hemizygous**. This means that there is no corresponding gene that could complement a deficient red or green photoreceptor gene.

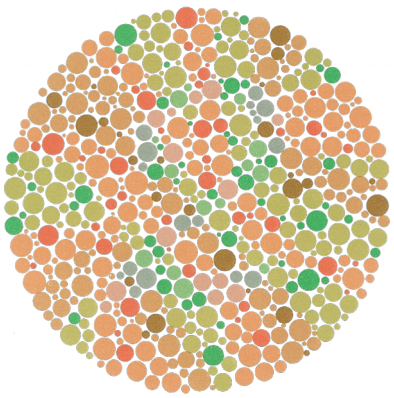
Dr. Shinobu Ishihara published his test for color perception in 1917 and this test is widely used to detect deficits in color perception. Below are examples of Ishahara plates. Record the number that you perceive in each plate and discuss with the rest of the class.

1. As you go through the plates above, note the number that you see (if any).
2. The genes for the Red and Green receptors are on the X-chromosome, who are most affected by mutation? Create a Punnet Square to illustrate how this works.
3. Can women be color-blind for red/green?
4. Humans have 3 color light receptors and have trichromatic vision. Some women are described as possibly having tetrachromatic vision (seeing 4 colors) and being able to discriminate colors invisible to the rest of us. Describe a mechanism for why this could happen. Why is there a possible gender bias?





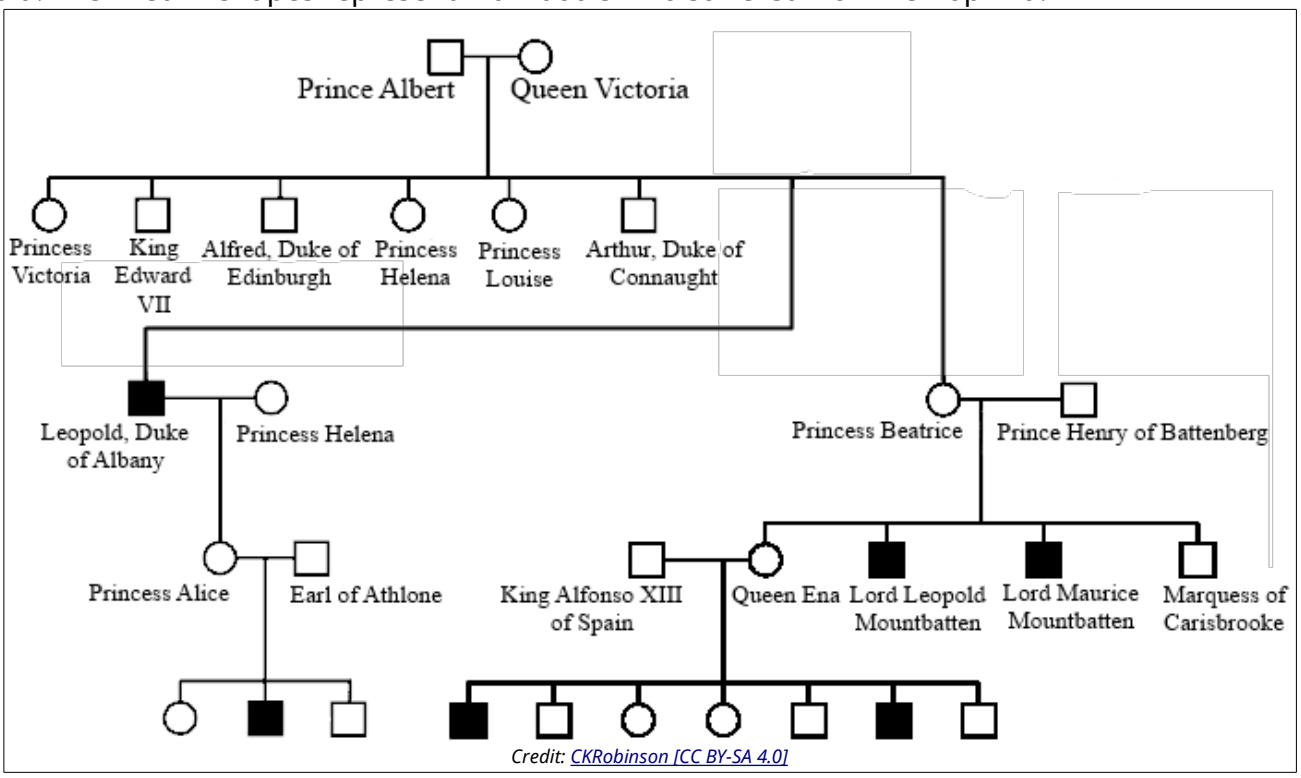




Jeremy Seto jseto@citytech.cuny.edu

The case of Queen Victoria

Hemophilia literally translates to *blood loving*. This is a description of a series of disorders where an individual has an inability to clot blood after a cut. In modern times, clotting factors may be administered to an afflicted individual, but a prior treatment involved blood transfusions. A very famous family had a genetic predisposition to hemophilia and due to the proliferative nature of this family, we have some statistical power to verify predictions on the probabilities of passing the disease state. Below is a partial pedigree for Victoria, Queen of the United Kingdom of Great Britain and Ireland and Empress of India. The filled in shapes represent individuals who suffered from hemophilia.



Credit: CKRobinson [CC BY-SA 4.0]

1. From the pedigree above, what can you say about this form of hemophilia with respect to dominance?
2. From this pedigree, can you comment on the probable chromosome where the deficiency occurs?
3. Assign genotypes for Prince Albert and Queen Victoria and perform a Punnet Square to illustrate if their offspring reflect your statements on dominance and chromosome location.
4. Albert and Victoria were 1st cousins. Do you believe this had anything to do with the propagation of this disease? What does your Punnet Square tell you?
5. Highlight the definitive carriers of the disease gene in the pedigree above.

Additional Resources

- Full case study can be acquired at the National Center for Case Study Teaching in Science.
- Human Factor IX mRNA sequence

Jeremy Seto jseto@citytech.cuny.edu

X-inactivation

The mammalian X-chromosome contains significantly more genetic information than the Y-chromosome. This gene dosage is controlled for in females through a process called **X-inactivation** where one of the X-chromosomes is shut down and highly condensed into a **Barr body**. Inactivation of the X-chromosome occurs in a stochastic manner that results in females being cellular mosaics where a group of cells have inactivated the paternal X-chromosome and other patches of cells have inactivated the maternal X-chromosome. The most striking example of **mosaicism** is the calico cat. A calico cat (tortoise shell cat) is always a female. One of the genes that encodes coat color in cats resides on the X-chromosome and exist as either orange or black alleles. Due to the stochastic inactivation, the patterning of orange and black fur is a distinctive quality of calicos.

Credit: Howcheng [GFDL CC-BY-SA-3.0 or CC BY-SA 2.5]

While the genetic information for the the orange or black coat color exists in all cells, they are not equally expressed. This type of heritable trait in spite of the presence of the genetic material (DNA) is called **epigenetic** to imply that it is "above" (epi) genetics .

Jeremy Seto jseto@citytech.cuny.edu

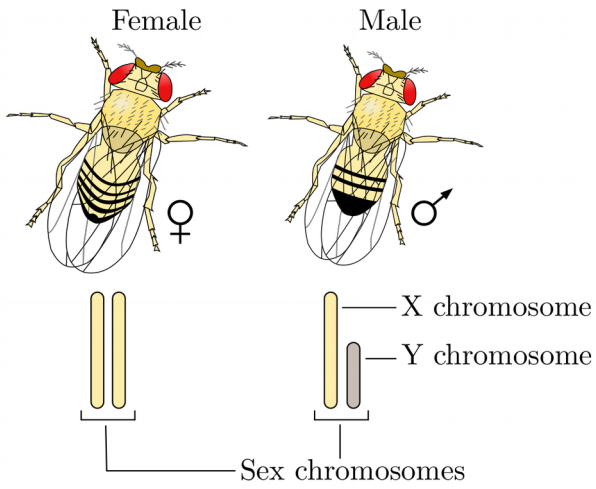
Drosophila: Thomas Hunt Morgan

Eye colors (clockwise): brown, cinnabar, sepia, vermilion, white, wild. Also, the white-eyed fly has a yellow body, the sepia-eyed fly has a black body, and the brown-eyed fly has an ebony body. White-eyed flies have a gender imbalance and occur mostly in males.

Credit: [Public Domain]

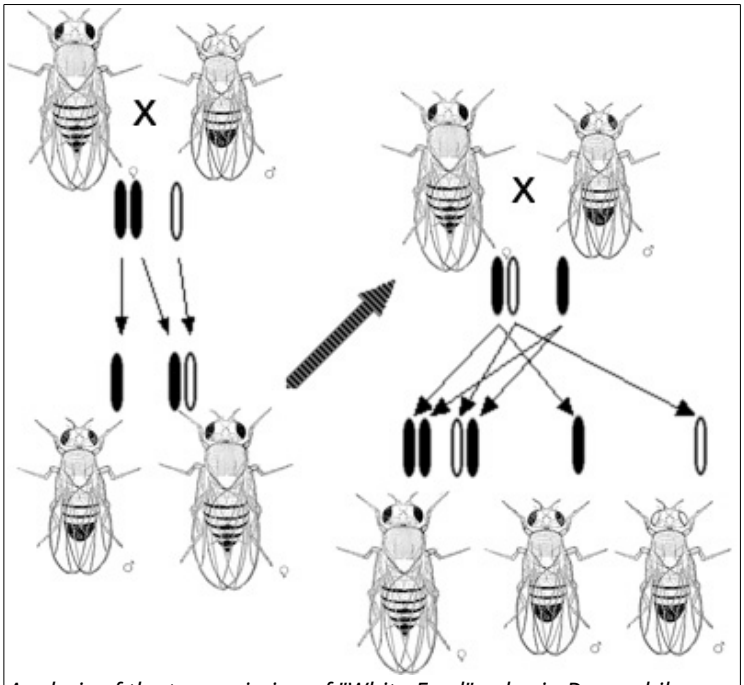
Around 1908, Thomas Hunt Morgan began to explore the genetics of what was to become a model organism, *Drosophila melanogaster* (Fruit fly). This small organism had a relatively short life cycle, great fecundity and was easily managed. From these flies that normally have red eye coloring, he and his students found white-eyed mutants. The lab noted that white-eyed flies were almost exclusively male. This gender imbalance lead Morgan to believe that the trait was sex-linked. In 1911, Morgan published a paper that described the inheritance patterns of 5 eye-colors in *Drosophila* (Morgan, 1911).

While DNA was not yet known as the source of genetic information, Morgan's studies revealed that the location of genes most likely resided on the chromosomes. By cataloging many mutations in the lab, he was able to construct a map of gene locations. His 1922 paper specifically stated that some traits were sex-linked and therefore residing on the sex chromosome. When performing crosses of white-eyed males to wild-type females, he continued to find white-eyed trait only in males. However, in the subsequent cross of females from that generation with white-eyed males, the presence of white-eyed males and females were revealed. This indicated that the white-eyed trait was recessive and resided on the X chromosome.



Drosophila follow a sex determination based on the the ratio of X:A chromosomes and not by the presence of a Y as in mammals. A 1:1 ratio results in a female and a 1:2 ratio results in a male where the Y is ignored. The standard 2N = 8.

Credit: GYassineMrabet [GFDL or CC BY-SA 4.0-3.0-2.5-2.0-1.0]



Analysis of the transmission of "White-Eyed" color in Drosophila

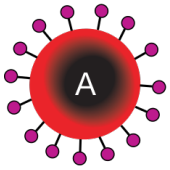
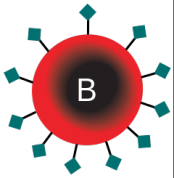
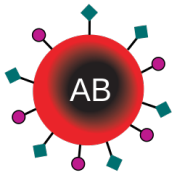
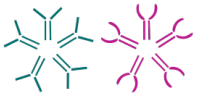
Credit: [Public Domain]

Jeremy Seto jseto@citytech.cuny.edu

Non-Mendelian Genetics

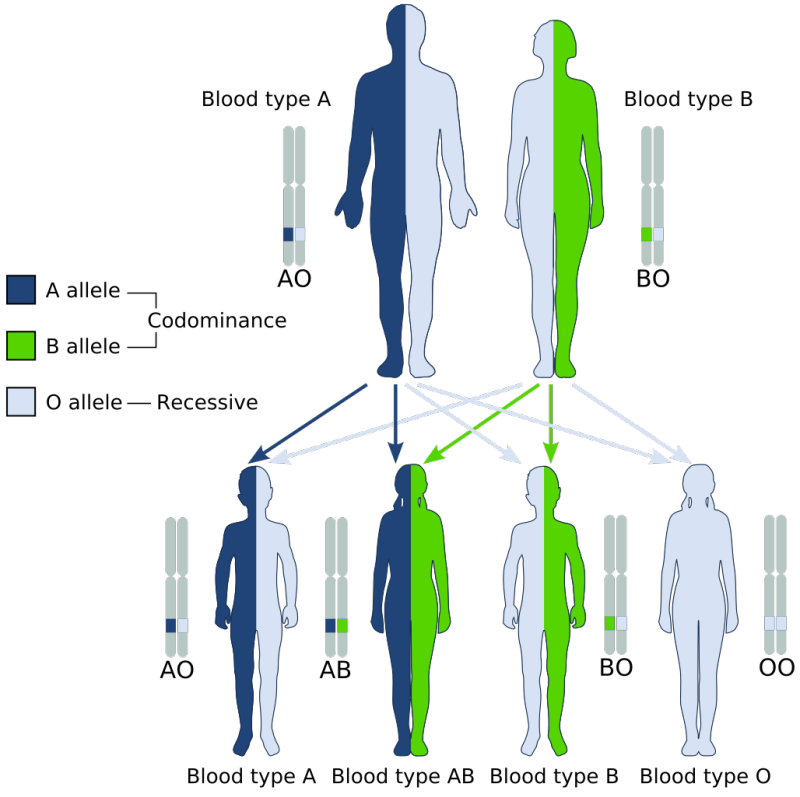
Co-Dominance and multiple alleles

Co-dominance is said to occur when there is an expression of two dominant alleles. The prototypical case for this is the human ABO blood grouping.

	Group A	Group B	Group AB	Group O
Red blood cell type				
Antibodies in Plasma	 Anti-B	 Anti-A	None	 Anti-A and Anti-B
Antigens in Red Blood Cell	 A antigen	 B antigen	 A and B antigens	None

Credit: InvictaHOG [Public Domain]

Three alleles exist in the ABO system: A, B and O. This results in four blood types: A, B, O and the blended AB.

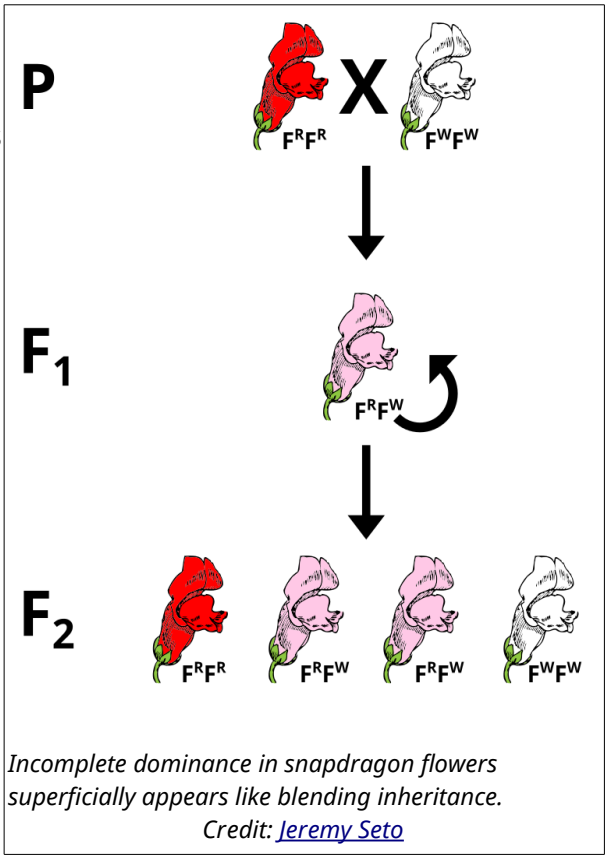


Credit: YassineMrabet [Public Domain]

Jeremy Seto jseto@citytech.cuny.edu

Incomplete Dominance

During Mendel's time, people believed in a concept of blending inheritance whereby offspring demonstrated intermediate phenotypes between those of the parental generation. This was refuted by Mendel's pea experiments that illustrated a Law of Dominance. Despite this, non-Mendelian inheritance can be observed in sex-linkage and co-dominance where the expected ratios of phenotypes are not observed clearly. **Incomplete dominance** superficially resembles the idea of blending inheritance, but can still be explained using Mendel's laws with modification. In this case, alleles do not exert full dominance and the offspring resemble a mixture of the two phenotypes.



Incomplete dominance in snapdragon flowers superficially appears like blending inheritance.
Credit: Jeremy Seto

The most obvious case of a two allele system that exhibits incomplete dominance is in the snapdragon flower. The alleles that give rise to flower coloration (Red or White) both express and the heterozygous genotype yields pink flowers. There are different ways to denote this. In this case, the superscripts of R or W refer to the red or white alleles, respectively. Since no clear dominance is in effect, using a shared letter to denote the common trait with the superscripts (or subscripts) permit for a clearer denotation of the ultimate genotype to phenotype translations.

Problem: Incomplete Dominance

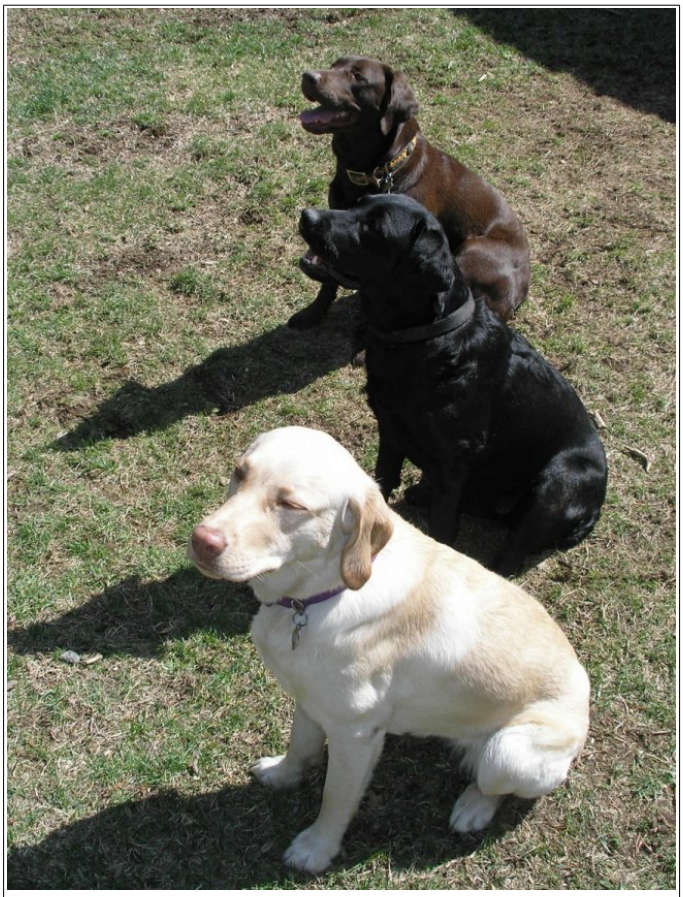
If pink flowers arose from blending inheritance, then subsequent crosses of pink flowers with either parental strain would continue to dilute the phenotype. Using a Punnet Square, perform a test cross between a heterozygous plant and a parental to predict the phenotypes of the offspring.

Jeremy Seto jseto@citytech.cuny.edu

Epistasis and Modifier Genes

Interplay of multiple enzymes in a biochemical pathway will alter the phenotype. Some genes will modify the actions of another gene.

Genes do not exist in isolation and the gene products often interact in some way. **Epistasis** refers to the event where a gene at one locus is dependent on the expression of a gene at another genomic locus. Stated another way, one genetic locus acts as a modifier to another. This can be visualized easily in the case of labrador retriever coloration where three primary coat coloration schemes exist: black lab, chocolate lab and yellow lab.

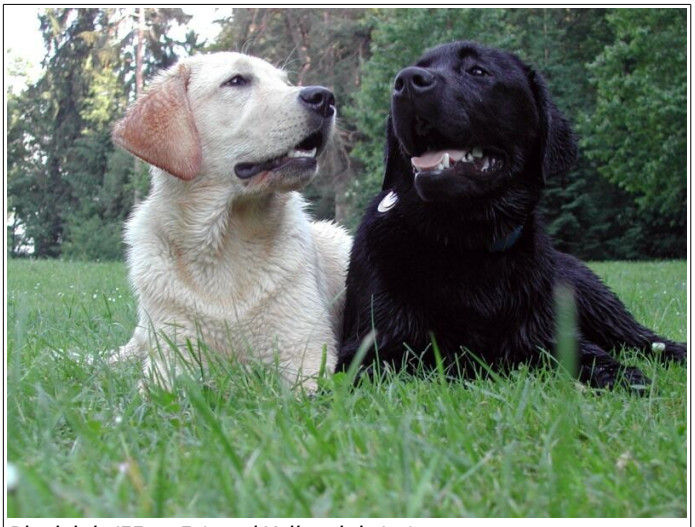


Chocolate lab (top), Black lab (middle), Yellow lab (bottom) coat colorations arise from the interaction of 2 gene loci, each with 2 alleles.

Credit: Erikeltic [CC BY-SA 3.0 or GFDL]

Black lab (BB or Bb) and Chocolate lab (bb)

Credit: dmealiffe[CC BY-SA 2.0]



Black lab (EE or Ee) and Yellow lab (ee)

Credit: [Public Domain]

Jeremy Seto jseto@citytech.cuny.edu

Two genes are involved in the coloration of labradors. The first is a gene for a protein called TYRP1, which is localized to the melanosomes (pigment storing organelles). Three mutant alleles of this gene have been identified that reduce the function of the protein and yield lighter coloration. These three alleles can be noted as "***b***" while the functioning allele is called "***B***". A heterozygous (Bb) or a homozygous dominant individual will be black coated while a homozygous recessive (bb) individual will be brown.

The second gene is tied to the gene for Melanocortin 1 Receptor (MC1R) and influences if the eumelanin pigment is expressed in the fur. This gene has the alleles denoted "***E***" or "***e***". A yellow labrador will have a genotype of either *Bbee* or *bbee*.

The interplay between these genes can be described by the following diagram:

Black lab (B_E_, Chocolate lab (bbE_), Yellow lab with dark skin where exposed (B _ee) and Yellow lab with light skin where exposed.
Credit: Jeremy Seto

Jeremy Seto jseto@citytech.cuny.edu

Using Cytobrush

1. Use sterile cytobrush and insert into mouth
2. brush cytobrush on inside of cheek 25 times
3. Swirl cytobrush in 100 µl of Chelex suspension (10% w/v)
4. Place centrifuge tube with Chelex and cell suspension on 100 °C heat block for 10 minutes
5. Centrifuge tubes at maximum speed for 5 minutes
6. DNA is in the supernatant. (avoid beads at bottom)
7. Store DNA in -20 °C

PCR with PCR Beads

1. Add 22 µl of primer mix (forward and reverse) to beads
2. Ensure that the the bead is dissolved
3. Add 3 µl of DNA

Jeremy Seto jseto@citytech.cuny.edu

Polymorphisms

The difference in nucleotide sequences between humans lies between 0.1-0.4%. That means that people are greater than 99% similar. But when you look around the room at your classmates, you can see that that small difference amounts to quite a bit of variation within our species. The bulk of these differences aren't even within the coding sequences of genes, but lie outside in regulatory regions that change the expression of those genes. Imagine if there were mutations to the coding sequences, this could be very deleterious to the well-being of the organism. We say that the coding sequences of genes that ultimately lead to proteins has a **selective pressure** to remain the same. The areas outside of the coding sequences have a reduced and sometimes non-existent selection pressure. These areas are allowed to mutate in sequence and even expand or contract. Areas of changes or differences are called **polymorphic** (many forms). If you were to read a repetitive set of sequences and count the repetition, you'd make mistakes and lose count. Likewise, DNA polymerase will make errors or stutter in areas of repetitiveness and produce polymorphic regions.

Tandem Repeats

A type of polymorphism occurs due to these repeats expanding and contracting in non-coding regions. These regions are called variable number tandem repeats (**VNTRs**) or sometimes short tandem repeats (**STRs**). Any region or location on a chromosome is referred to as **locus** (loci for plural). Scientists use polymorphic loci that are known to contain VNTRs/STRs in order to differentiate people based on their DNA. This is often used in forensic science or in maternity/paternity cases. Any variation of a locus is referred to as an **allele**. In standard genetics, we often think of an allele as a variation of gene that would result in a difference in a physical manifestation of that gene. In the case of STRs, these alleles are simply a difference in number of repeats. That means the length of DNA within this locus is either longer or shorter and gives rise to many different alleles. VNTRs are referred to as **minisatellites** while STRs are called **microsatellites**.

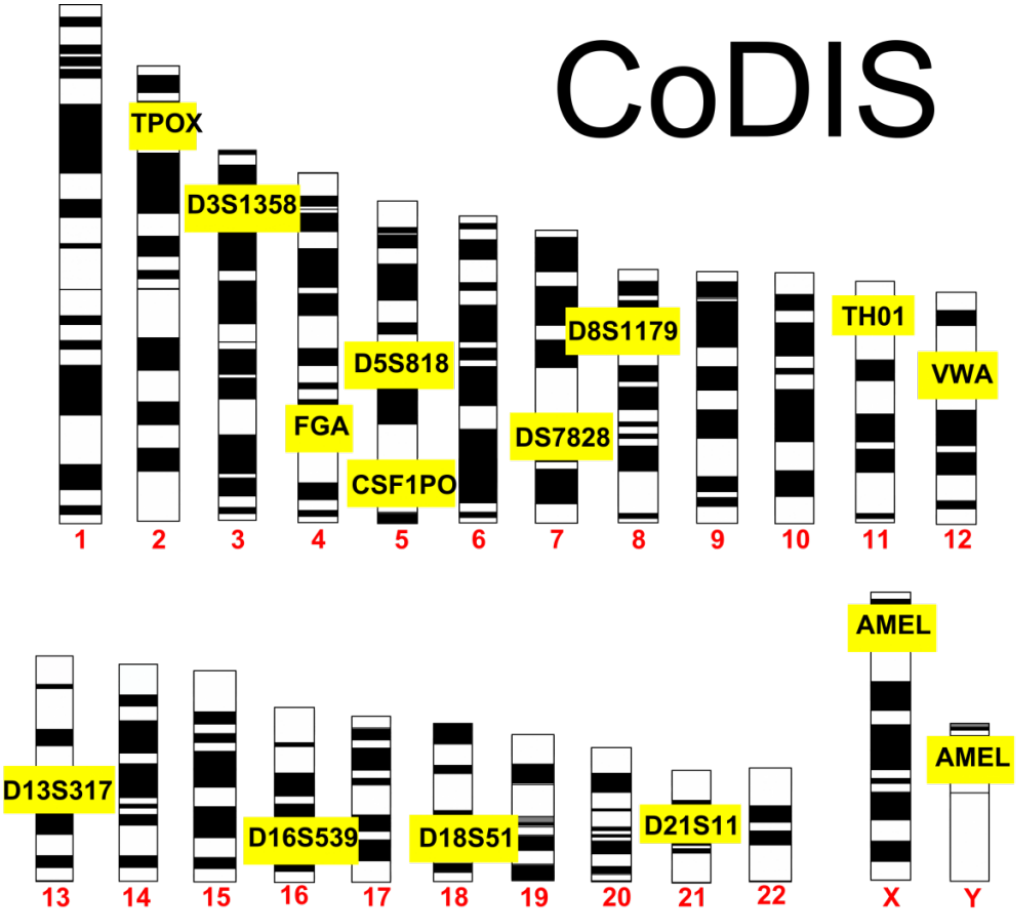
CoDIS

The FBI and local law enforcement agencies have developed a database called the Combined DNA Index System (**CoDIS**) that gathers data on a number of STRs. By establishing the number of repeats of a given locus, law enforcement officials can differentiate individuals based on the repeat length of these alleles. CoDIS uses a set of 13 loci that are tested together. As you

Credit: Shruthi.n.christ [CC BY-SA 4.0]

Jeremy Seto jseto@citytech.cuny.edu

would imagine, people are bound to have the same alleles of certain loci, especially if they were related. The use of 13 different loci makes it statistically improbable that 2 different people could be confused for each other. Think about this in terms of physical traits. As you increase the number of physical traits used to describe someone, you are less likely to confuse that person with someone else based on those combinations of traits. Using the CoDIS loci increases the stringency since there are many alleles for each locus. The thirteenth locus in CoDIS (called AMEL) discriminates between male and female.

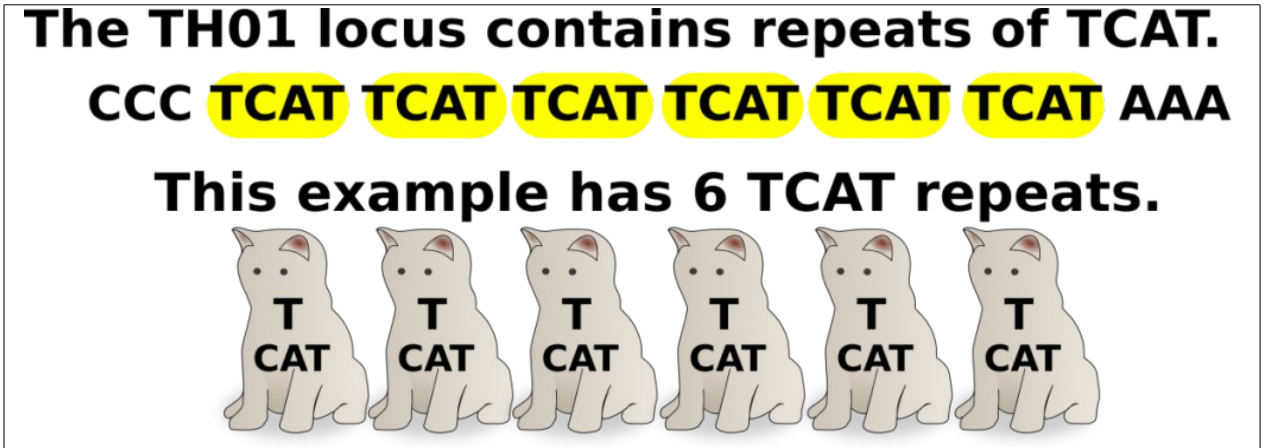


CoDIS STRs: *The FBI utilizes 13 different loci to discriminate between people. AMEL discriminates by gender and is located on the X & Y.*

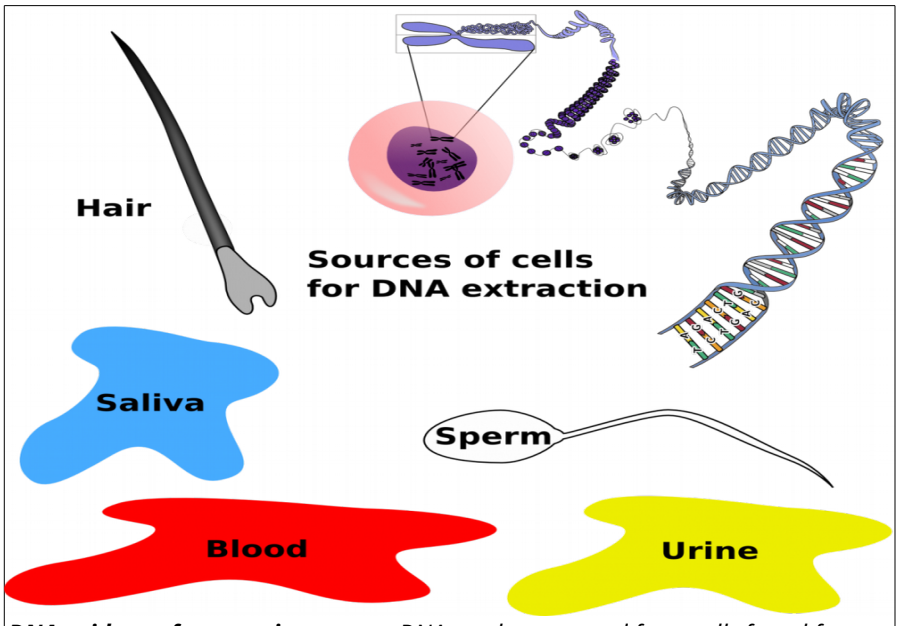
Jeremy Seto jseto@citytech.cuny.edu

Crime Scene Investigation

This lab uses a CoDIS locus called TH01. TH01 is a locus on chromosome 11 that has a repeating sequence of TCAT. There are reported to be between 3-14 repeats in this locus. With the exception of X and Y in a male, all chromosomes have a homologous partner. Therefore, each individual will have 2 alleles for each CoDIS locus.



TH01 STR: *Outside of the STR, there is flanking areas of known sequence. The primers that amplify TH01 in PCR recognize these flanking sequences to amplify the TCAT repeats.*

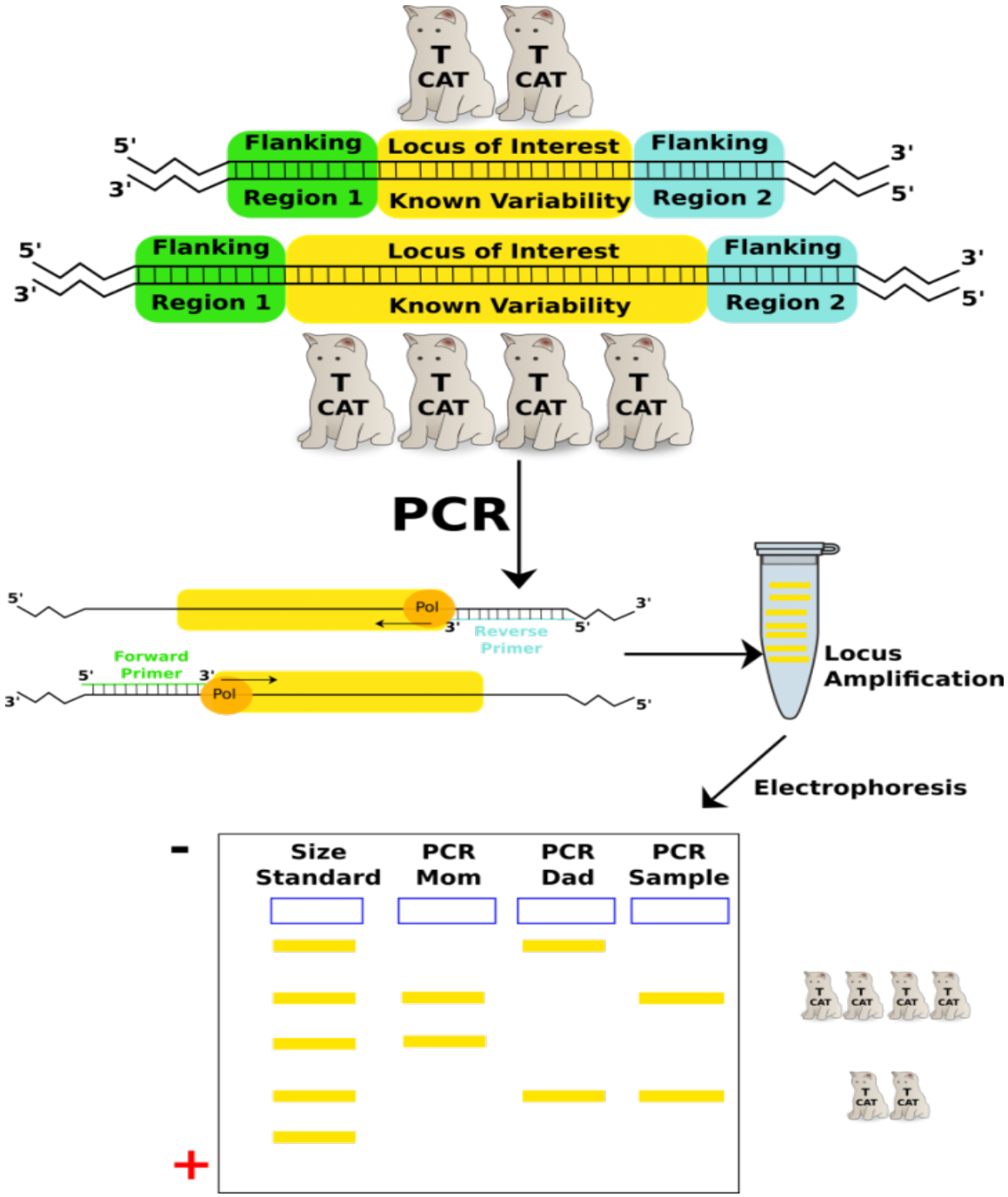


DNA evidence from a crime scene: *DNA can be extracted from cells found from various sources at a crime scene. PCR can amplify this small amount of DNA.*

At a crime scene, criminals don't often leave massive amounts of tissue behind. Scant evidence in the form of a few cells found within bodily fluids or stray hairs can be enough to use as DNA evidence. DNA is extracted from these few cells and amplified by PCR using the specific primers that flank the STRs used in CoDIS.

Amplified DNA will be separated by gel electrophoresis and analyzed. Size reference standards and samples from the crime scene and the putative suspects would be analyzed together. In a paternity test, samples from the mother, the child and the suspected father would be analyzed in the same manner. A simple cheek swab will supply enough cells for this test.

Jeremy Seto jseto@citytech.cuny.edu



TH01 locus used in a Paternity/Maternity test: *Individual PCR reactions are run for each sample (mom, dad, child). The TH01 primer pair specifically amplifies the locus. Each amplified sample is run on the same gel to resolve the different alleles of TH01 from each individual. From this test the sample could be the offspring from these 2 parents but use of more STRs would make it more definitive. Count the TCATs.*

External Resources

- Flash animation walking through what a STR is http://www.dnalc.org/view/15981-DNA-variations.html

Jeremy Seto jseto@citytech.cuny.edu

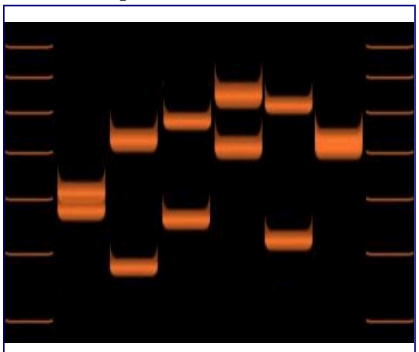
Minisatellite Activity

The minisatellite marker D1S80 is located at 1p35-p36. This VNTR is 16 bases long. With a variation of alleles between 3-24 repeats, the locus displays enough diversity to aid in distinguishing between people. Although this is not a CoDIS marker, use of multiple loci are required to definitively identify samples. The large repeat (16bp) permits the use of standard agarose gel electrophoresis to explore the diversity of this locus in our lab. PCR products range between 430 to 814bp long.

- D1S80-for: 5'-GAAACTGGCCTCCAAACACTGCCCGCCG-3'
- D1S80-rev: 5'-GTCTTGTTGGAGATGCACGTGCCCCTTGC-3'

1. PCR the DNA samples extracted from cheek cells using the PCR Beads
2. Pour 2% agarose into casting apparatus in refrigerator
    - 2 gels per class need to be made → 100ml of TBE with 2g agarose
    - add 5µl SYBR safe solution into the molten agarose before casting
    - place 2 sets of combs into the gel → at one end and in the middle
3. load gel with DNA ladder
    - Sample is from the PCR
4. Run gel at 120V for 20 minutes
5. Visualize on UV transilluminator

Example Results



By: PaleWhaleGail (CC BY-SA 3.0)

1. How many alleles are visible in each lane?
2. Are the genotypes distinguishable between individuals?
3. Are any of the alleles common between individual samples?

Jeremy Seto jseto@citytech.cuny.edu

Genetics leaves a bad taste in my mouth... or not

Some of our personal preferences arise from the way we were brought up. Culture plays a role in our likes and dislikes. Likewise, our experiences play a role in how we respond to certain stimuli. Another major factor that plays a role into our preferences comes wired in our genome. The DNA in our cells is the instruction manual for who we are. We are programmed to seek out things of a nutritive values in order to acquire raw materials like carbohydrates, proteins and lipids. In our search for nutritive compounds we have learned to avoid things that don't taste good. Bitter things have a tendency to be associated with toxic compounds in nature. When eating a food item for the first time, molecules hit our tongue and stimulate multiple sensations: sweet, sour, salty, savory and bitter. Attributed to these multiple taste types are a diverse family of receptors that bind to the molecules that result in our perception of these sensations. Something bitter might make us learn to avoid this food item in the future.

One type of bitter receptor senses the presence of a chemical called phenylthiocarbamide (**PTC**). This chemical chemically resembles toxic compounds found in plants but is non-toxic. The ability to taste PTC is comes from the gene called *TAS2R38*. This gene encodes a protein that on our tongues that communicates the bitterness of this chemical. There are two common alleles of this gene with at least five more uncommon variants. Within the two common forms, a **single nucleotide polymorphism** (**SNP**) is responsible for changing one amino acid in the receptor. It's this difference of one amino acid that results in the ability of the receptor to either respond or not respond to PTC. We inherit one copy of the gene from our father and one copy from our mother. Based on how our parents gametes were formed and what alleles we received during the fertilization event determines how we respond to this chemical. Because we each have 2 copies of this gene, we can utilize simple Mendelian genetics to understand which allele is dominant or recessive.

1. Place a piece of “Control” paper on the tongue and indicate if there is a taste
2. Place a piece of “PTC” paper on the tongue and indicate if there is a taste and the taste severity
3. Fill out the table for the class to identify how many non-tasters, tasters or super-tasters there are.
4. Indicate if you believe the trait is dominant or recessive (ability to taste or not taste)
5. Assign a descriptor allele for the dominant (a capital letter) or the recessive (a lowercase letter) and draw a Punnet square for the F_2 generation of 2 Heterozygous parents.
6. Compare the class tally of tasters and non-tasters in the class and discuss with your instructor if there is a clear dominance of this trait.

Table: PTC Tasting Tally

Phenotypes	Number	% Total
PTC Tasters (Dominant or Recessive)		
PTC Non-Tasters (Dominant or Recessive)		
Total		

Jeremy Seto jseto@citytech.cuny.edu

Questions:

1. How do you explain the presence of those who can't taste PTC, those who can taste it and those who really can't stand the taste of it?
2. This chemical is non-toxic and doesn't exist in nature. Do you think there is a **selective pressure** that confers an advantage to those who do taste it?

Exercise: Coding Bitterness

Prior to this exercise, review the **Central Dogma**.

The full coding sequence of ***TAS2R38*** is 1,002 bases (334 amino acids) long. A segment of the gene is shown below where the SNP (in red) occurs. Variant 1 is the version of the gene that encodes for the ability to taste PTC. Variant 2 is the version of the gene that is unable to bind to PTC. This SNP mutation is called a **missense** mutation because it changes the amino acid. Some mutations cause the insertion of a premature stop codon. This **nonsense mutation** results in a truncated protein and can be disastrous to the function. We already know that the simple substitution of one nucleotide translates to a change in one amino acid and determines the ability to taste PTC. Imagine if a large group of amino acids from the protein was missing.

With template strand (“Complement”) information:

1. Write the sequence of the coding strand.
2. Write the sequence of the mRNA
3. Use the Genetic Code Chart to translate the amino acid sequence

Variant 1

Coding Strand: 5'-

Complement : 3'-TTC TCC GTC CGT GAC TCG-5'

mRNA : 5'-

Amino Acid :

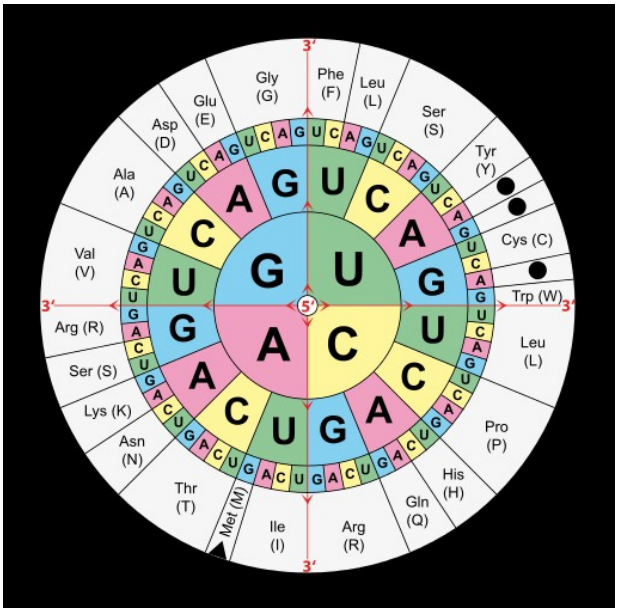
Variant 2

Coding Strand: 5'-

Complement : 3'-TTC TCC GTC GGT GAC TCG-5'

mRNA : 5'-

Amino Acid :



Jeremy Seto jseto@citytech.cuny.edu

PCR Genotyping the TAS2R38 PTC receptor

- 5’-CCTTCGTTTTCTTGGTGAATTTTTGGGATGTAGTGAAGAGGCGG-3’ (Forward Primer)
- 5'-AGGTTGGCTTGGTTTGCAATCATC-3' (Reverse Primer)

1. PCR the DNA samples extracted from cheek cells using the PCR Beads
2. Pour 2% agarose into casting apparatus in refrigerator
    - 2 gels per class need to be made → 100ml of TBE with 2g agarose
    - add 5µl SYBR safe solution into the molten agarose before casting
    - place 2 sets of combs into the gel → at one end and in the middle
3. Digest PCR product with *Hae*III
    - remove 10µl of PCR product into a fresh tube
    - add 1µl of *Hae*III enzyme into tube
    - incubate for 10 minutes at 37°C
4. load gel with DNA ladder, Digested and Undigested
    - Undigested sample is from the original PCR
5. Run gel at 120V for 20 minutes
6. Visualize on UV transilluminator

SNP detection

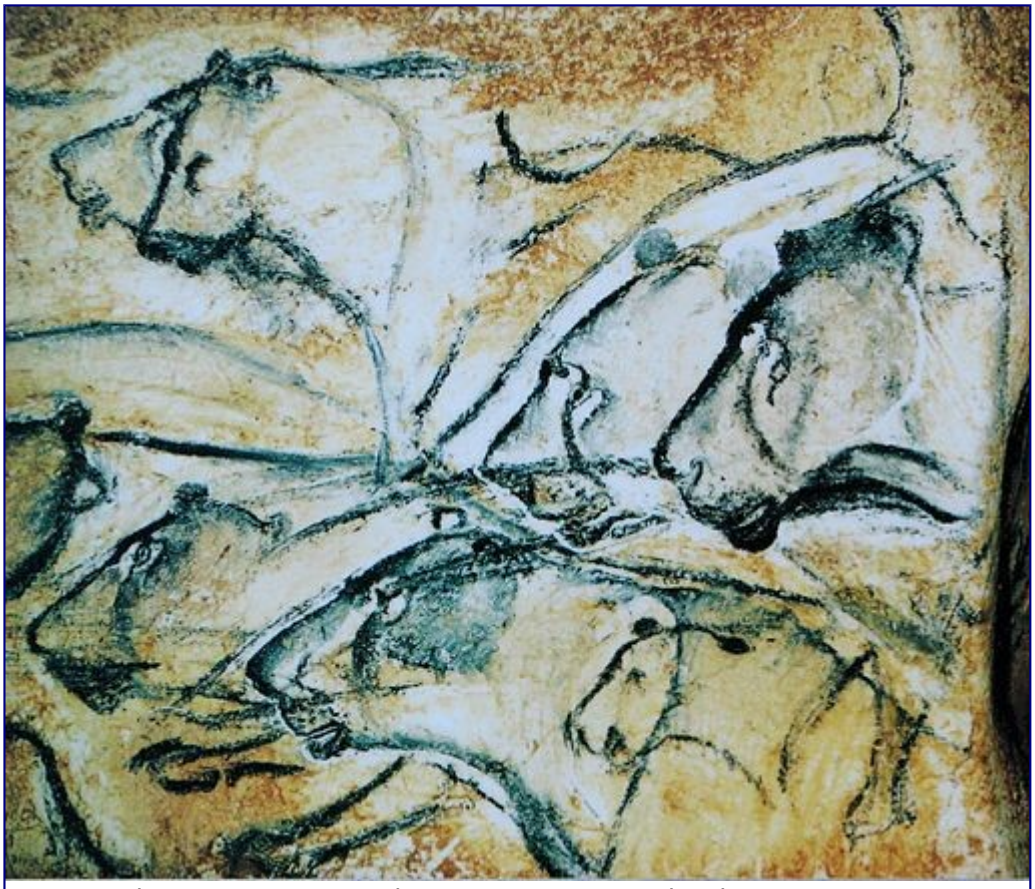
The longer primer ends with the sequence "GG". Both alleles at this locus will amplify equally well with this primer set, however, one allele will have the sequence "GGGC" and another "GGCC". "GGCC" is the restriction site for the enzyme *Hae*III. The digestion of this amplified DNA will be digestible for one allele and yield a DNA fragment the size of the large primer (44 bp) as well as the remainder of the amplicon. Because of this difference in digestion profile of the amplicon, we can identify the 2 alleles at this locus.

Analysis questions

1. What is the size of the PCR product?
    - Perform an *in silico* PCR on the *Tas2R38* gene and identify the size of the amplicon.
2. The long primer is 44bp. If the amplicon of the allele digests, what are the sizes of fragments expected following *Hae*III?
3. Which allele is the one that can be identified through *Hae*III digestion?
    - use the results of the PTC paper test
4. Some lanes contain 3 bands instead of 1 or 2. Can you explain this?

Jeremy Seto jseto@citytech.cuny.edu

Being Human



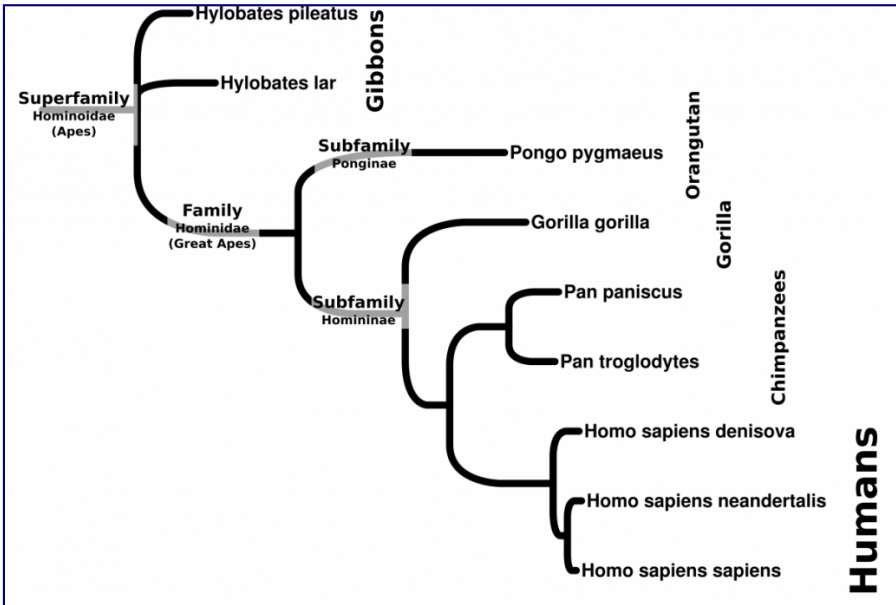
Drawings dating to approximately 30,000 years ago in the Chauvet Cave

What constitutes being human? Many will point at a cultural identity and leaving long-standing remnants of that culture. Such prehistorical artifacts like cave drawings and tools provide an anthropological framework for identifying what it is to be human, but the biological identity remains locked in the history of our DNA.

Spear points of the Clovis Culture in the Americas dating to approximately 13,000 years ago.

Jeremy Seto jseto@citytech.cuny.edu

The Great Apes



Phylogenetic tree generated with Cytochrome Oxidase I (COI) genes.

Homo sapiens represent a branch of primates in the line of Great Apes. The family of Great Apes consists of four extant genera: *Homo, Pan, Gorilla, Pongo*. Karyotype analysis (Yunis *et al.*, 1982) reveals a shared genomic structure between the Great Apes. While humans have 46 chromosomes, the other Great Apes have 48. Molecular evidence at the DNA level indicates that Human Chromosome 2 is a fusion of 2 individual chromosomes. In the other Great Apes, these 2 Chromosomes are referred to as 2p and 2q to illustrate their synteny to the human counterpart.

Chimpanzees (*Pan*) are the closest living relatives to modern humans. It is commonly cited that less than 2% differences in their nucleotide sequences exist with humans (Chimpanzee Sequencing and Analysis Consortium, 2005). More recent findings in comparing the complement of genes (including duplication and gene loss events) now describes the difference in genomes at about 6% (Demuth JP, et al., 2006).

The Pan-Homo divergence. A display at the Cradle of Humankind illuminates the skulls of two extant Hominini with a series of model fossils from the Hominina subtribe of Australopithecina and Homo.

Credit: Jeremy Seto (CC-BY-NC-SA) https://flic.kr/p/SmhHTd

Jeremy Seto jseto@citytech.cuny.edu

The Genus *Homo*

An underground lake at inside the Sterkfontein Cave system at the Cradle of Humankind (South Africa)

Credit: Jeremy Seto (CC-BY-NC-SA) https://flic.kr/p/RczrEg

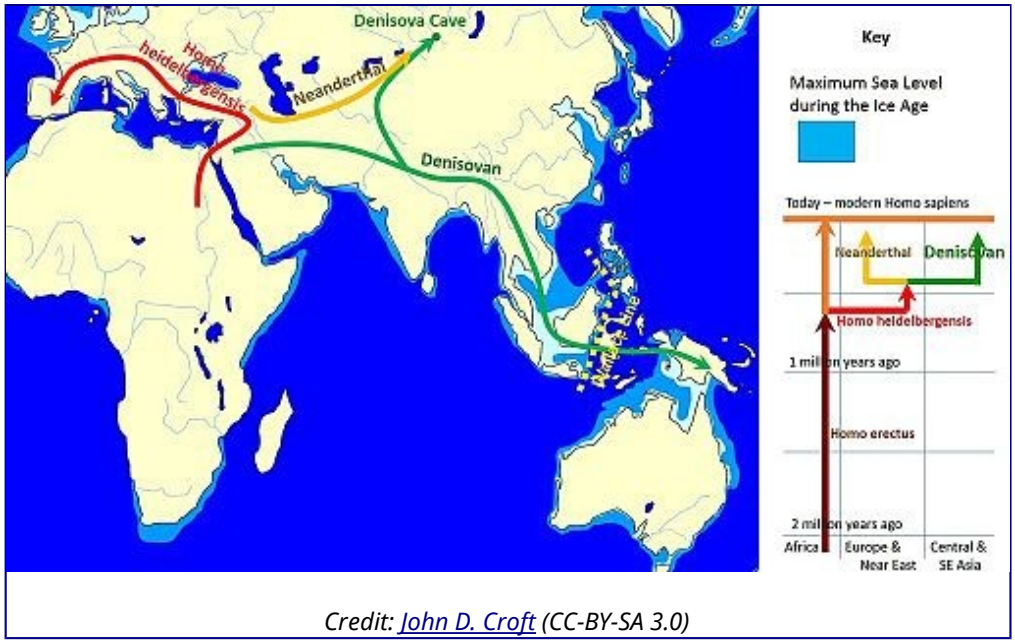
The rise of the human lineage is thought to arise in Africa. Fossils of *Austroloptihs* (southern apes) found in death traps, like those at the Cradle of Humankind, reveal a historical record of organisms inhabiting the landscape. The breaks in the ceiling of the caves provide opportunities for animals to fall inside these caves to their death. The limestone deposits of the caves serve as an environment for fossilization and mineralization of their remains. An abundance of fossilized hominids in these caves including *Australopithecus africanus, Australopithecus prometheus, Paranthropus boisei,* and the newly discovered *Homo naledi* continue to reveal the natural history of the genus *Homo* from 2.6 million to 200,000 years ago.

The entrance to the archaeological site at Sterkfontein, Cradle of Humankind (South Africa).

Credit: Jeremy Seto (CC-BY-NC-SA) https://flic.kr/p/ULs2Sv

Jeremy Seto jseto@citytech.cuny.edu

Ancient DNA of Humans



Credit: John D. Croft (CC-BY-SA 3.0)

In 2008, a piece of a finger bone and a molar from a Siberian Cave were found that differed slightly from that of modern humans. The cave, called Denisova Cave, maintains an average temperature of 0ºC year round and was suspected to contain viable soft tissue. Bones in this cave were discovered that had similarities to modern humans and Neandertals. An initial mitochondrial DNA analysis revealed that these beings represented a distinct line of humans that overlapped with them in time (Krause *et al.*, 2010). Analysis of the full nuclear genome followed and indicated that interbreeding existed between these Denisovans, Neandertals and modern humans (Reich *et al.*, 2010). Furthermore, analysis of DNA from a 400,000 year old femur in Spain revealed that these three lines diverged from the species *Homo heidelbergensis* and that Denisovans were closest in sequence (Meyer *et al.*, 2016).

Between modern humans, markers found in the mtDNA can be used to trace the migrations and origins along the maternal line. Similarly, VNTRs found on the Y chromosome have revealed migration patterns along paternal lines within men. Other markers, like the insertion points of transposable elements can be used to further describe the genetics and inheritance of modern humans while providing a snapshot into evolutionary history.

Jeremy Seto jseto@citytech.cuny.edu

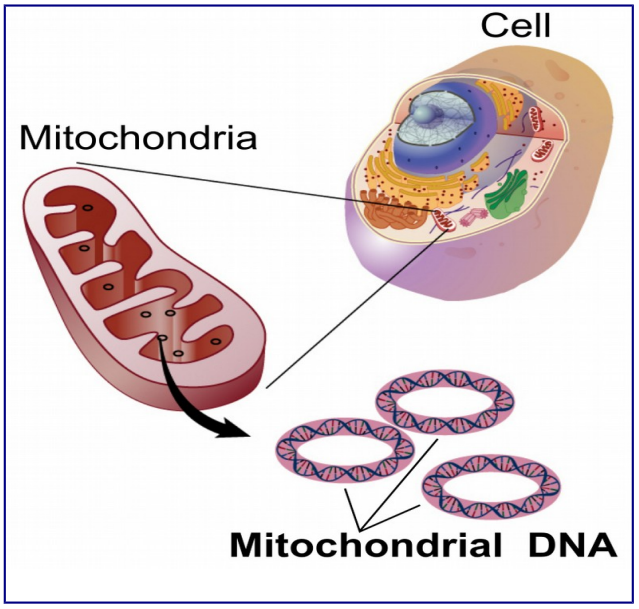
Other Resources

- **Great Ape Mitochondrial Sequences**
- http://media.hhmi.org/biointeractive/click/Origins/01.html
- Yunis JJ, Prakash O. The origin of man: a chromosomal pictorial legacy. Science. 1982 Mar 19;215(4539):1525-30.
- Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature. 2005 Sep 1;437(7055):69-87.
- Demuth JP, De Bie T, Stajich JE, Cristianini N, Hahn MW. **The evolution of mammalian gene families**. PLoS One. 2006 Dec 20;1:e85.
- Krause, Johannes; Fu, Qiaomei; Good, Jeffrey M.; Viola, Bence; Shunkov, Michael V.; Derevianko, Anatoli P. & Pääbo, Svante (2010), "The complete mitochondrial DNA genome of an unknown hominin from southern Siberia", Nature, **464** (7290): 894–897, doi:10.1038/nature08976, PMID 20336068
- Reich, David; Green, Richard E.; Kircher, Martin; Krause, Johannes; Patterson, Nick; Durand, Eric Y.; Viola, Bence; Briggs, Adrian W. & Stenzel, Udo (2010), "Genetic history of an archaic hominin group from Denisova Cave in Siberia", *Nature*, **468** (7327): 1053–1060, doi:10.1038/nature09710, PMID 21179161
- Meyer M, Arsuaga JL, de Filippo C, Nagel S, Aximu-Petri A, Nickel B, Martínez I, Gracia A, Bermúdez de Castro JM, Carbonell E, Viola B, Kelso J, Prüfer K, Pääbo S. **Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins.** Nature. 2016 Mar 24;531(7595):504-7. doi: 10.1038/nature17405. PMID:26976447

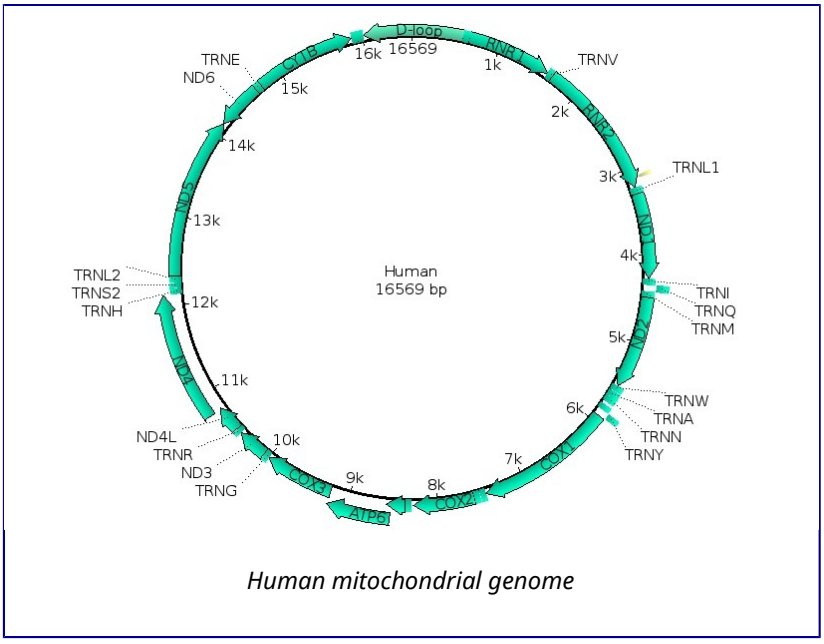
Jeremy Seto jseto@citytech.cuny.edu

Mitochondrial and Maternal Inheritance

In addition to the 23 chromosomes inherited from mother and 23 chromosomes inherited from father, humans have an additional genome that is only inherited from the mother. This genome comes from the **endosymbiotic** organelle, the mitochondrion.



Mitochondria are thought to have arisen in the eukaryotic line when bacteria capable of detoxifying the deadly effects of atmospheric oxygen were engulfed by a eukaryote that did not proceed to consume it. Over the course of time, these formerly free-living bacteria became dependent on the eukaryotic cell environment while providing the benefit to the host cell of aerobic respiration. Hallmarks of this endosymbiotic event include: the inner prokaryotic membrane surrounded by the outer eukaryotic membrane, the presence of prokaryotic ribosomes and most significantly, the circular prokaryotic chromosome. Mitochondria still replicate independently of the host cell but can not survive outside of this cellular environment. Animal mitochondria have the simplest genomes of all mitochondrial genomes, ranging from 11-28kb. The human mitochondrial genome consists of 37 genes which are almost all devoted to processing ATP through oxidative phosphorylation.

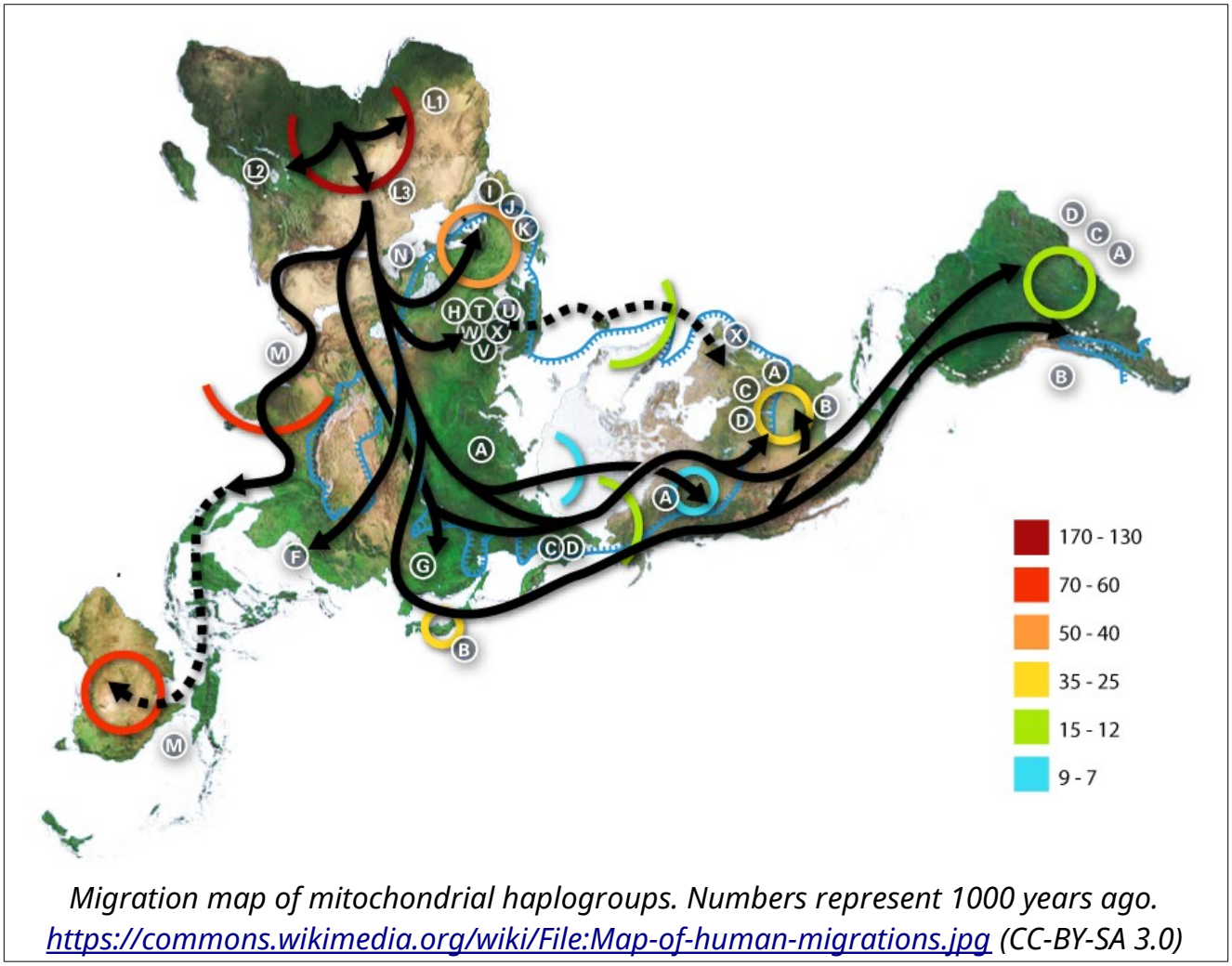


Human mitochondrial genome

The human mitochondrial genome (genbank file) consists of 16,569 nucleotides (16.6kb). While most of this 16.6kb genome consists of protein encoding genes, approximately 1.2kb non-coding DNA takes part in signals that control the expression of these genes and replication processes. It is the area of DNA where the double-strandedness is displaced and having the name **D-loop** (displacement loop). Mutations in this area generally have very little effect on the functioning of the mitochondria. Because of this reduced selection pressure on this area, this **control region** is also referred to as the **hypervariable region**. This hypervariable region actually has 10 times more SNPs than the nuclear genome. Due to this abundance of mutations, it is possible to track down the maternal line of an individual. Why just maternal? The human oocyte contains many mitochondria while sperm cells only contain mitochondria that power the flagellar motion. Upon fertilization, the flagellum and the associated mitochondria are lost, leaving the zygote with only maternal mitochondria.

Jeremy Seto jseto@citytech.cuny.edu

The cluster of SNPs found in the mitochondrial control region are linked and are always inherited together. Because of the lack of paternal contribution, this linkage is referred to as a **haplotype**, or "half-type". Tracking these polymorphic haplotypes, a family tree of humans was developed in the 1980s which concluded that humans arose from a metaphorical "Mitochondrial Eve" 200,000 years ago. As a metaphor to the Biblical Eve, this alludes to an origin but unlike the Biblical event, this does not mean that it was a single woman that gave rise to all of modern humanity. On the contrary, the metaphor merely indicates that a series of females; sisters and cousins, of this line gave rise to modern humans.



Migration map of mitochondrial haplogroups. Numbers represent 1000 years ago. *https://commons.wikimedia.org/wiki/File:Map-of-human-migrations.jpg (CC-BY-SA 3.0)*

The use of mitochondria for this analysis provides great flexibility, especially from ancient sources. Unlike the nuclear genome which only has 2 copies of DNA per cell, the mitochondria are abundant in number and provide many copies of genome per cell. Ancient sources of DNA in fossils will most often have degradation of the DNA. The mitochondrial genome is just as likely to undergo degradation over time, however the high copy number allows for gaps to be filled in easily. SNPs do not alter the overall size of the hypervariable region, therefore amplification by PCR can not resolve these differences based on agarose gel migration. However, **amplicons** (amplified copies) can be sent for sequencing whereby each nucleotide can be called out in succession and reveal the specific SNPs.

Jeremy Seto jseto@citytech.cuny.edu

The PCR amplification of the mitochondrial control region

There are 2 hypervariable regions within the control region of the mitochondria. This exercise amplifies just one of these. For more definitive results, both should be amplified and sequenced. This exercise will permit us to have a rough idea of the origins of our maternal line and we will be able to attribute ourselves to various tribes throughout the world. The human mitochondrial genome (genbank file).

Forward Primer 5'-TTAACTCCACCATTAGCACC-3'
Reverse Primer 5'-GAGGATGGTGGTCAAGGGAC-3'

1. PCR the previously extract DNA samples
    - Pour 2% agarose into casting apparatus in refrigerator
    - 2 gels per class need to be made → 100ml of TBE with 2g agarose
    - add 5µl SYBR safe solution into the molten agarose before casting
    - place 2 sets of combs into the gel → at one end and in the middle
2. load gel with DNA ladder and PCR
3. Run gel at 120V for 20 minutes
4. Visualize on UV transilluminator
5. Document with camera to verify amplification
6. The instructor will submit the viable reactions for sequencing
7. **Analyze data during Bioinformatics Lab session**
    1. Using NYCCT email address, register for account at http://dnasubway.iplantcollaborative.org/
    2. retrieve reference mitochondrial sequences
    3. perform multiple sequence alignment using MUSCLE
    4. draw phylogenetic trees using PHYLIP and visualize using FigTree

Jeremy Seto jseto@citytech.cuny.edu

Transposable Elements

Mobile genetic elements called **transposable elements** or **transposons** are located throughout the genome. These elements were first described in maize by Barbara McClintock at the Cold Spring Harbor Laboratory where she observed a disruption of coloring in corn kernels that did not follow simple Mendelian inheritance.

McClintock's Corn Kernels

Each kernel represents a distinct new individual organism. Kernel color is described through simple Mendelian inheritance where purple is dominant over yellow. Dr. McClintock noticed that some kernels contained spots. She noticed that the coloration disruption could later reverse in subsequent generations. She described the phenomenon of this break in the Mendelian characteristics as a "genetic instability". Over time, she would come to realize that the spots in these kernels arose from the insertion of DNA into the area of genes that were involved in controlling kernel coloration.

Dr. McClintock's description of this phenomenon and the underlying mechanisms was extremely unpopular as it violated what was already known about the *fixity* of genetics. Though initially skeptical, biology has found that these “jumping genes” are found in every taxa including prokaryotes (where they are often associated with genes conferring antibiotic resistance) and she was later awarded the Nobel Prize. Approximately half of the human genome consists of transposons, making up the bulk of what was previously referred to as “junk DNA”.
(See Animation: http://www.dnaftb.org/32/animation.html)

Advanced Video of gene disruption by jumping genes

https://youtu.be/_cJfsWYR42M

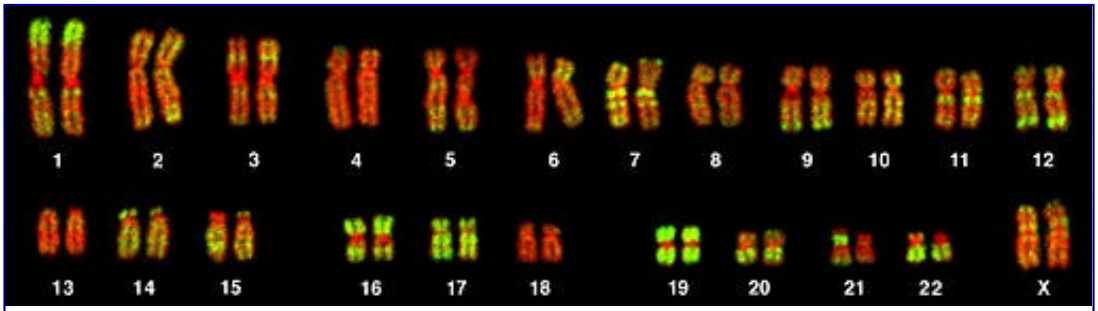
Classes of Transposons

Transposons can be autonomous or non-autonomous. **Autonomous transposons** encode their own **transposase** enzyme that facilitates the jumping of the gene while **non-autonomous transposons**

Jeremy Seto jseto@citytech.cuny.edu

require the transposase activity of another transposable element. Functional DNA transposons are autonomous and work through a "*cut and paste*" mechanism. DNA transposons are delineated by flanking terminal repeats that mark the location that the transposase excises the DNA. These DNA elements then re-integrate at a different location within the genome. The excision from DNA leaves marks of these flanking repeats that can be used to study the rate and level of DNA transposition events within a genome. The insertion of these transposons can effect the expression of nearby genes and can completely disrupt genes they land into as evidenced in the speckled corn kernels that McClintock described.

RNA transposons are called **retrotransposons** because they are transcribed into an mRNA and require a reverse transcription to integrate into the genome. The most common mobile element in the human genome are the Long Interspersed Nuclear Elements (**LINEs**) and the Short Interspersed Nuclear elements (**SINEs**). These retrotransposons are most abundantly represented by the autonomous LINE1 (L1) and non-autonomous *Alu* elements, respectively. *Alu* elements rely on the expression of the L1 in order to be reverse-transcribed and integrated into the genome. These retrotransposons work in a "*copy and paste*" mechanism and are responsible for genomic expansion. As their classifications signify, LINEs are longer than SINEs. This is due to the presence of a second reading frame that encodes the transposase.



The Human Genome is littered with "junk" DNA. Green coloration corresponds to Alu sequences in this karyogram of the human genome.
Credit: Andreas Bolzer, Gregor Kreth, Irina Solovei, Daniela Koehler, Kaan Saracoglu, Christine Fauth, Stefan Müller, Roland Eils, Christoph Cremer, Michael R. Speicher, Thomas Cremer (CC-BY 2.5)

- Human L1 sequence (fasta)
- Human Ya5 Alu sequence (fasta)

Advanced Video of Classification of Transposons and the Evolution of Genomes

https://youtu.be/IOXvZXtc93U

Further Reading

- http://www.nature.com/scitable/topicpage/barbara-mcclintock-and-the-discovery-of-jumping-34083
- http://www.nature.com/scitable/topicpage/functions-and-utility-of-alu-jumping-genes-561

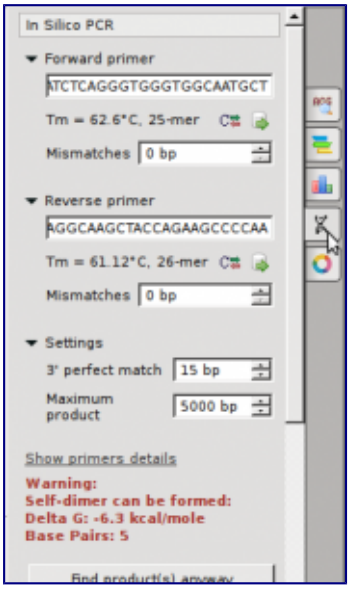
Jeremy Seto jseto@citytech.cuny.edu

Alu's are unique SINEs that appear in the primate lineage and reveal the lineage and diversification of primates. While retrotransposons can disrupt gene (as in some cases of hemophilia), they often land outside of genes or within introns without effect. One example of a non-disruptive Alu element in humans is found in the location called PV92 on chromosome 16. This element is of the youngest subfamily of Alu, called Ya5.

Since PV92 does not cause any deleterious effects, it can be used as a non-selected marker to illustrate lineage. Some people have an *Alu* element int his location while others do not. The presence or absence of this marker is viewed as an allele. This lab uses primer that flank the location of the *Alu* insertion that span 416 bp. If an *Alu* is present, the amplified DNA will be 300bp larger (the size of an *Alu*) at 731bp.

Exercise: In silico PCR of PV92

Forward primer: 5' GGATCTCAGGGTGGGTGGCAATGCT 3'
Reverse primer: 5' GAAAGGCAAGCTACCAGAAGCCCCAA 3'

1. Perform Virtual PCR **Informatics Exercise/Discussion**
2. Visit BLAST: https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch
3. Paste both primers: GGATCTCAGGGTGGGTGGCAATGCT GAAAGGCAAGCTACCAGAAGCCCCAA
4. Choose "Somewhat Similar"
    - Locate the locus of the product and the size
5. Find the PCR fragments in Ugene
    - Download the sample FASTA file: PV92 sample
    - Open the file in Ugene and select option "As Separate Sequences in Viewer"
    - Select the "In Silico PCR" button on the far right (double helix button) and insert the primers
    - 

    - A PCR product should be noted for one of the sequences after pressing "Find Products anyway"
    - Click on the second sequence in the viewer and Press "Find Products anyway"

Jeremy Seto jseto@citytech.cuny.edu

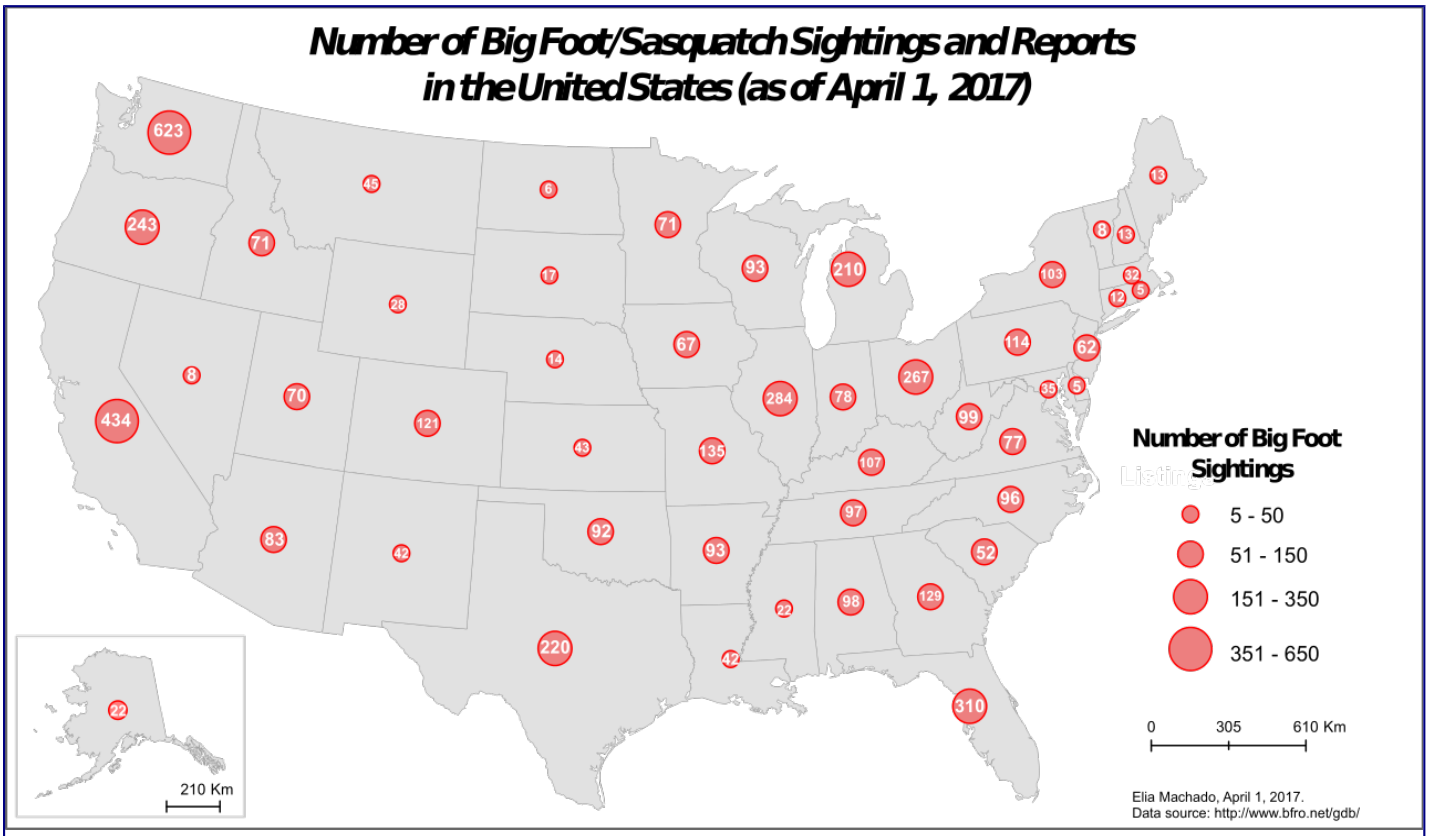
Exercise: PCR genotype PV92 locus

1. PCR the individual samples
2. Pour 2% agarose into casting apparatus in refrigerator
    - 2 gels per class need to be made → 100ml of TBE with 2g agarose
    - add 5µl SYBR safe solution into the molten agarose before casting
    - place 2 sets of combs into the gel → at one end and in the middle
3. Load DNA ladder and PCR samples
4. Run gel at 120V for 30 minutes
5. Visualize on UV transilluminator
6. Score gels for the presence/absence of the alleles to determine genotype frequency in the class

Jeremy Seto jseto@citytech.cuny.edu

Cryptozoology

Cryptozoology is a pseudoscience centered around the description of animals that have little or no evidence of existing. These mythical beasts include: Bigfoot, Yeti, Sasquatch, jackelope, Loch Ness Monster and chupacabra. Little evidence exist to illustrate their existence other than folklore.



Sightings of Bigfoot in North America.

Credit: Elia Machado (CC-BY-NC-SA)

Sometimes, physical evidence is left behind like hair or feces. With DNA evidence, we can help to confirm the existence of these unknown creatures. Below features a table from Sykes et al. displaying results on supposed cryptic Apes (Bigfoot/Yeti) and what DNA evidence has revealed them to be.

Jeremy Seto jseto@citytech.cuny.edu

ref. no.	location	attribution	GenBank sequence match	common name
25025	Ladakh, India	*yeti*	*U. maritimus*	polar bear
25191	Bhutan	*yeti/migyhur*	*U. maritimus*	polar bear
25092	Nepal	*yeti*	*Capricornis sumatraensis*	serow
25027	Russia	*almasty*	*U. arctos*	brown bear
25039	Russia	*almasty*	*Equus caballus*	horse
25040	Russia	*almasty*	*Bos taurus*	cow
25041	Russia	*almasty*	*Equus caballus*	horse
25073	Russia	*almasty*	*Equus caballus*	horse
25074	Russia	*almasty*	*U. americanus*	American black bear
25075	Russia	*almasty*	*P. lotor*	raccoon
25194	Russia	*almasty*	*U. arctos*	brown bear
25044	Sumatra	*orang pendek*	*Tapirus indicus*	Malaysian tapir
25035	AZ, USA	*bigfoot*	*P. lotor*	raccoon
25167	AZ, USA	*bigfoot*	*Ovis aries*	sheep
25104	CA, USA	*bigfoot*	*U. americanus*	American black bear
25106	CA, USA	*bigfoot*	*U. americanus*	American black bear
25081	MN, USA	*bigfoot*	*Erethizon dorsatum*	N. American porcupine
25082	MN, USA	*bigfoot*	*U. americanus*	American black bear
25202	OR, USA	*bigfoot*	*U. americanus*	American black bear
25212	OR, USA	*bigfoot*	*C. lupus/latrans/domesticus*	wolf/coyote/dog
25023	TX, USA	*bigfoot*	*Equus caballus*	horse
25072	TX, USA	*bigfoot*	*Homo sapiens*	human
25028	WA, USA	*bigfoot*	*U. americanus*	American black bear
25029	WA, USA	*bigfoot*	*C. lupus/latrans/domesticus*	wolf/coyote/dog
25030	WA, USA	*bigfoot*	*Bos taurus*	cow
25069	WA, USA	*bigfoot*	*Odocoileus virginianus/hemionus*	white-tailed/mule deer
25086	WA, USA	*bigfoot*	*Bos taurus*	cow
25093	WA, USA	*bigfoot*	*C. lupus/latrans/domesticus*	wolf/coyote/dog
25112	WA, USA	*bigfoot*	*Bos taurus*	cow
25113	WA, USA	*bigfoot*	*C. lupus/latrans/domesticus*	wolf/coyote/dog

Cryptozoological samples of hair believed to arise from legendary animals like Bigfoot, Sasquatch, Yeti, etc.
Table taken from: http://rspb.royalsocietypublishing.org/content/281/1789/20140161 CC-BY

The Need for Barcoding

Taxonomy of living things was created by Carl von Linné, who formalized it by using a binomial classification system to differentiate organisms. Binomial nomenclature was used to describe a genus and a species name to each organism to provide an identity. These days, classification of organisms is becoming increasingly important as a measurement of diversity in the face of habitat destruction and global climate change. There is no consensus on how many life forms exist on this planet, but the

Jeremy Seto jseto@citytech.cuny.edu

estimation of extinction rates is about 1 species per 100-1000 million species. Classification in Linné's day was mostly performed by morphological differences. This was carried on in fossils. However, morphology has many drawbacks, especially in sexually dimorphic species or species with multiple developmental morphologies.

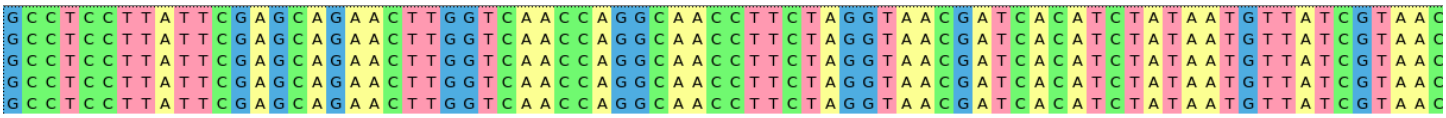
Larva (top) of the Green Lacewing and the adult (bottom).

Molecular biology and DNA technologies have revolutionized the classification system of living things especially in providing the ability to match relatedness of these species. **DNA barcoding**, like the name implies, seeks to utilize DNA markers to differentially identify organisms. But what DNA markers should be used? What criteria do we use to develop barcodes? Discrimination, Universality and Robustness are the criteria used to define the usefulness of barcodes.

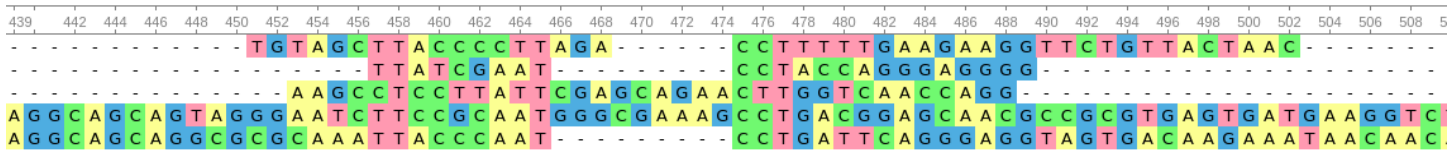
Since the goal of barcoding is to define specific organisms, discrimination is the *primary objective*. **Discrimination** refers to the difference of sequences that occur between species. However, science is easier when there is some universality in the locus used for discrimination. As it sounds, **universality** is an attempt to use the same locus in disparate genomes. While discrimination is about uniqueness of sequences, universality seeks to use a single set of PCR primers that will be able to amplify that same distinct region with variable sequence similarity. If some region of DNA has absolutely no sequence deviation between species, this has great universality but poor discrimination. But if a sequence has very low sequence similarity, this is

Jeremy Seto jseto@citytech.cuny.edu

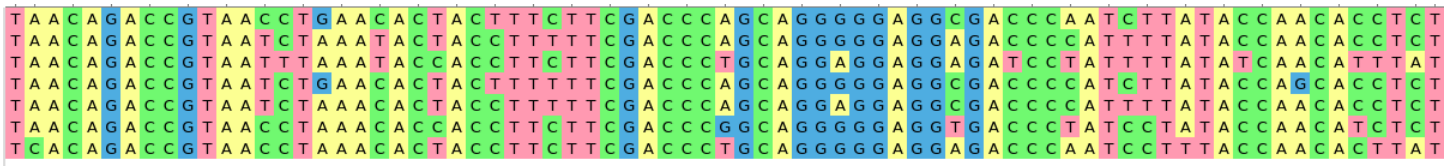
great for discrimination but has absolutely no universality and can not be amplified with the same set of primers. **Robustness** refers to the reliability of PCR amplification of a region. Some regions of DNA just don't amplify well or it is too difficult to design appropriate and unique primers for that locus.



A case where there is universality for designing primers, but not an area where discrimination can occur.



While discrimination of different organisms can occur in this situation, the lack of similarity in sequence would make it difficult to design primers. That is, the lack of universality in sequence would also make this PCR not robust.



Enough variability in these sequences gives us the ability to discriminate between species. The high similarity provides us the universality required to design primers that may be robust enough to amplify by PCR.

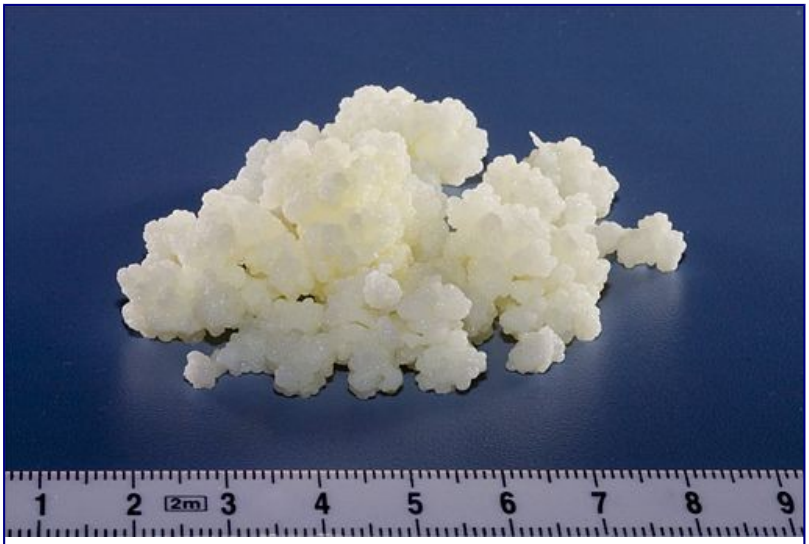
Sometimes, species are so similar for one sequence that a second marker is required. Just as the standard UPC barcode has a series of vertical line of different spacing and width, a 2-dimensional barcode adds that second dimension of information into a square of dots like in a QR code (Quick Response code). We can also utilize a second or a third or a fourth set of loci that will aid in increased discrimination just as CoDIS utilizes multiple STR sites to define individual people. In animals, the most commonly used barcode is the mitochondrial gene, Cytochrome Oxidase I (**COI**). Since all animals have mitochondria and have this mitochondrial gene, it offers high universality. It is a robust locus that is easy to amplify and has high copy number with enough sequence deviation between species to discriminate between them. Animal mitochondrial genomes vary from 16kb-22kb. However, plants, fungi and protists have wildly different and larger mitochondrial genomes. For plants, we use a chloroplast gene, ribulose-bisphosphate carboxylase large subunit (**rbcL**) or maturase K (**matK**) (Hollingsworth et al. 2011). Prokaryotes are often discriminated by their **16s** rRNA gene while eukaryotes can be identified by **18s** rRNA. Because of hybrid animals (mules, ligers, coydogs, etc.), COI (a maternally transmitted gene) will not create a clear picture of species identity. sometimes, closely related species are also indistinguishable by a single barcode, so the inclusion of 18s

Jeremy Seto jseto@citytech.cuny.edu

with COI may be necessary to define the identity of the species. Since it is so difficult to meet the three criteria (robustness, universality and discrimination) for all species, having these multiple barcodes is important. Fungi prove to be difficult in identification by COI, so another marker called the internal transcribed spacer (**ITS**) is used to aid in their identification. We must also remember that not everything with chloroplasts are plants and therefore additional markers are used to identify protists.

Mixtures of organisms

Lichens are composite organisms composed of cyanobacteria or other algae with fungi. In this case, a single barcode would incorrectly identify the species.



Kefir granules represent colonies of mixed microbes that are used to generate kefir.

Credit: A. Kniesel (CC-BY-SA 3.0)

A symbiotic colony of bacteria and yeast is used to ferment kombucha. As the name implies, this is a complex composite colony of multiple species that contribute to the qualities of the kombucha

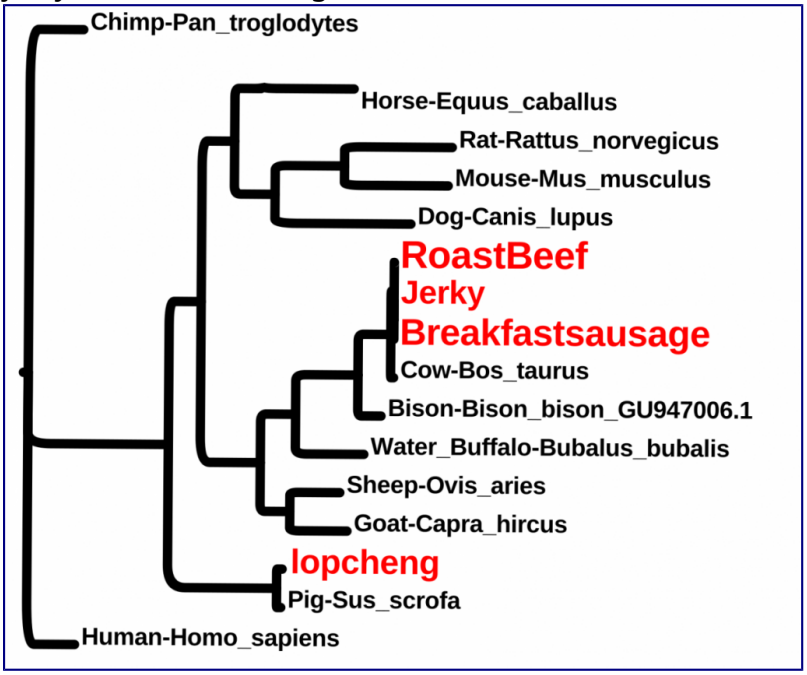
Credit: Lukas Chin (CC-BY-SA 4.0)

Metabarcoding and Microbiomes

Jeremy Seto jseto@citytech.cuny.edu

Class Results

Students wanted to check some food items. These included, breakfast sausage from a Halal cart, "beef jerky" from the vending machine, roast beef from the cafeteria and a Chinese sausage (lopcheng).



For more class results, please visit https://openlab.citytech.cuny.edu/dna-barcodes/

Further Resources

- Reference Sequences
  - Gorilla COI: FASTA
  - Apple rbcL: FASTA
  - *Staphylococcus* 16S: FASTA
  - *Toxoplasma* 18S: FASTA
  - Porcinni mushroom ITS: FASTA
- Sykes BC, Mullis RA, Hagenmuller C, Melton TW, Sartori M. Genetic analysis of hair samples attributed to yeti, bigfoot and other anomalous primates. Proc Biol Sci. 2014 Aug 22;281(1789):20140161. doi: 10.1098/rspb.2014.0161.
- Hollingsworth PM, Graham SW, Little DP (2011) Choosing and Using a Plant DNA Barcode. PLOS ONE 6(5): e19254. https://doi.org/10.1371/journal.pone.0019254
- http://www.dnabarcoding101.org/

Jeremy Seto jseto@citytech.cuny.edu

DNA barcoding of samples

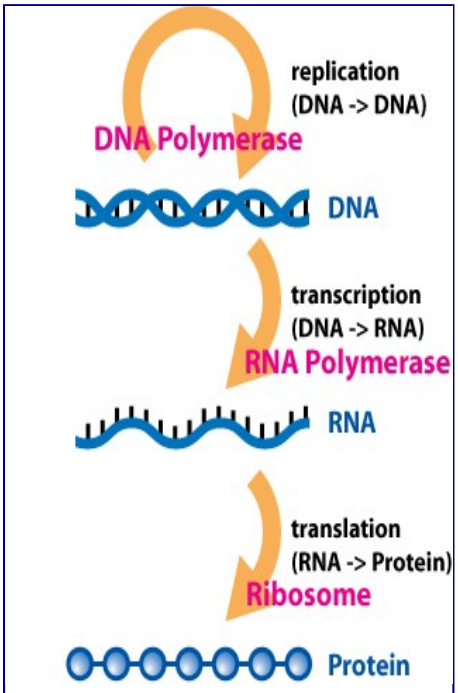
1. Place sample in a clean 1.5 mL tube
2. Add 100 µl of nuclear lysis solution to tube.
    - Twist a clean plastic pestle against the inner surface
3. Add 500 µl more nuclear lysis solution to tube.
4. Incubate the tube in a water bath or heat block at 65°C for 5-15 minutes.
5. [Optional] Add 200 µl of protein precipitation solution to each tube incubate on ice for 5 minutes
6. Centrifuge for 4 minutes at maximum speed to pellet protein and cell debris
7. Transfer 600 µl of supernatant to a clean labeled tube.
8. Add 600 µl of isopropanol
9. Centrifuge for 2 minute at maximum speed to pellet the DNA.
10. Pour off the supernatant and add 600 µl of 70% ethanol to wash the pellet
11. Centrifuge the tube for 2 minute at maximum speed and carefully remove the solution
12. Air dry the pellet for 10 minutes and add 100 µl of the DNA rehydration solution (TE)
13. Incubate the DNA at 65°C for 5-10 minutes to dissolve
14. Obtain PCR tube containing Ready-To-Go PCR Bead. Label the tube with your identification number.
15. Use a micropipette with a fresh tip to add 23 µL of one of the following primer/loading dye mixes to each tube. Allow the beads to dissolve for 1 minute.
    - Plants: rbcL primers (rbcLaF / rbcLa rev)
    - Fish: COI primers (VF2_t1/ FishF2_t1/ FishR2_t1/ FR1d_t1)
    - Insects: (LepF1_t1/ LepR1_t1)
16. Add 2 µl of your DNA directly into the appropriate primer/loading dye mix.
17. Place tubes in a Thermal cycler
18. Pour 2% agarose into casting apparatus in refrigerator
    1. 2 gel per class need to be made → 100ml of TBE with 2g agarose
    2. add 5 µl SYBR safe solution into the molten agarose before casting
    3. place 2 sets of combs into the gel → at one end and in the middle
19. Load DNA ladder and PCR samples
20. Run gel at 120V for 30 minutes
21. Visualize on UV transilluminator
22. Document with camera
23. Send amplicons of verified samples for sequencing

- Plant rbcL gene
    - rbcLaf 5'- ATGTCACCACAAACAGAGACTAAAGC-3' (forward primer)
    - rbcLar 5'- GTAAAATCAAGTCCACCRCG-3' (reverse primer)
- Animal coi gene
    - lepF1 5'- ATTCAACCAATCATAAAGATATTGG -3' (forward primer)
    - lepR1 5'- TAAACTTCTGGATGTCCAAAAAATCA-3'(reverse primer)
    - vf1f 5'- TCTCAACCAACCACAAAGACATTGG-3' (forward primer)
    - vf1r 5'- TAGACTTCTGGGTGGCCAAAGAATCA-3' (reverse primer)

Jeremy Seto jseto@citytech.cuny.edu

The Central Dogma

DNA was described as a molecule consisting of 2 anti-parallel strands in a double helix by Francis Crick and James Watson. The elegant model illustrated the intrinsic redundancy that made DNA a suitable data storage vessel for genetic information. Francis Crick later posited a notion of how this information went from storage to an actual program that runs cells. Crick first posited it as a "sequence hypothesis". This idea of information flow is called the **Central Dogma of Molecular Biology**. DNA stores the information that is expressed as an intermediate message of RNA. This RNA is then translated in amino acids to yield proteins.

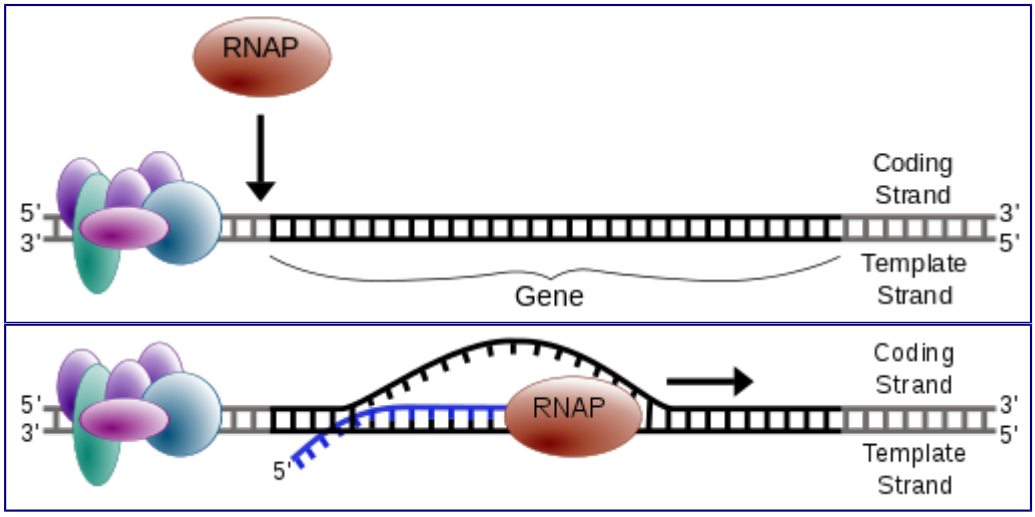


Flow of information in cells. DNA serves as a template for copying itself (replication). DNA can also serve as a template for RNA (transcription). RNA is decoded into amino acids to generate proteins (translation). Credit Daniel Horspool (CC-BY-SA 3.0)

Transcription

https://youtu.be/SMtWvDbfHLo

DNA is simply a storage vessel of genetic information. It sits in the nucleus and must be called upon through a process of **transcription** where an enzyme called **RNA Polymerase**"reads aloud" the stored information into a molecule called messenger RNA (**mRNA**). Since DNA is double-stranded in an **anti-parallel** fashion, we automatically know the sequence of the second strand by knowing the first. The mRNA is made through complimentary base-pairing to the **template strand**, which is the reverse complement of the coding strand. The **coding strand** is the strand that reads identical in sequence to the mRNA with the exceptions of T's being replaced by U's.

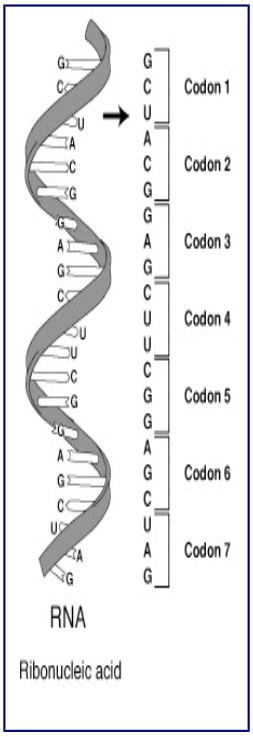


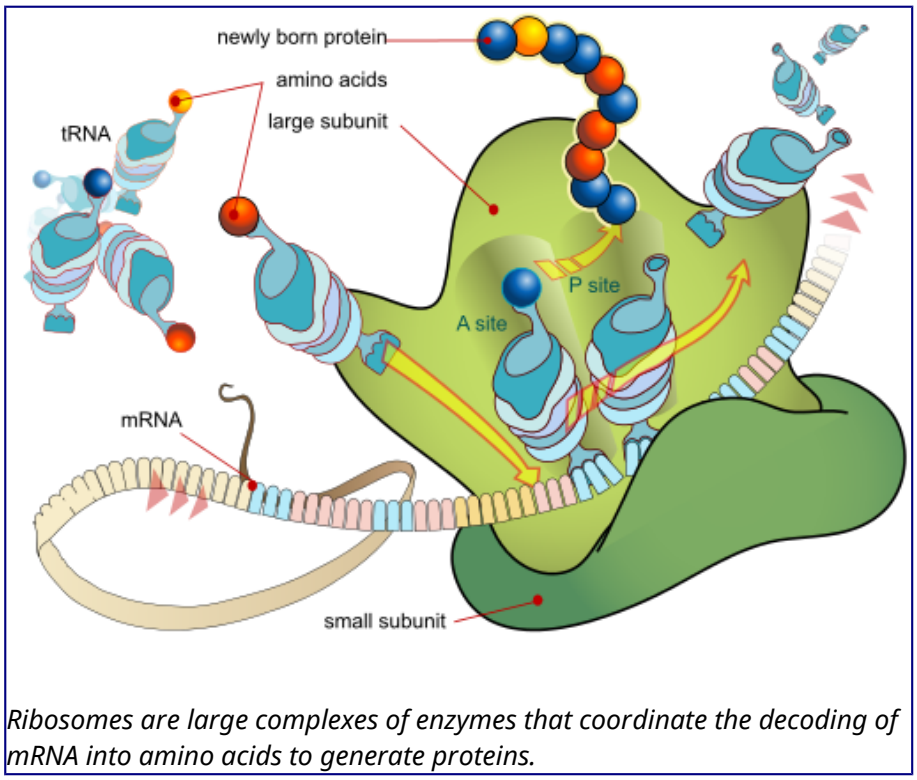
Jeremy Seto jseto@citytech.cuny.edu

Translation

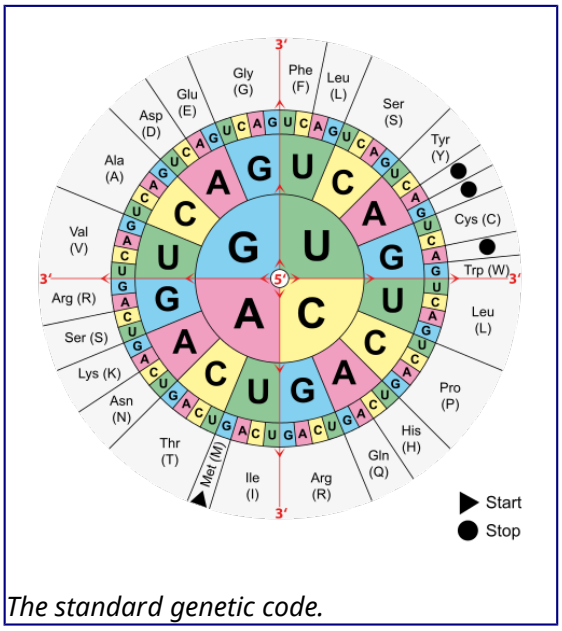
https://youtu.be/TfYf_rPWUdY

This coding strand is later decoded by the **ribosomes** with the help of transfer RNA's **tRNA**'s) that act as a decoder of the information and protein assembler in a process called **translation**. The ribosome scans along the mRNA and recognizes nucleotides in batches of 3 . These batches of 3 can be translated into an amino acid and is known as a **codon**. Since there are 4 types of bases and they are read as groups of 3, there are 4^3 (or 64) combinations of these codons. However, there are only 20 amino acids used to build proteins. This indicates that there is room for redundancy. Three of these codons tell the ribosome to stop, like a period in a sentence. These are called **stop codons**. There is one special codon that performs double duty: ATG. The codon (ATG) that encodes the amino acid Methionine also acts as a **start codon** that tells the ribosome where to start reading from. Like nucleic acids, proteins have a polarity and are synthesized in an amino to carboxyl direction. We abbreviate this by terming the beginning of the protein sequence, N-terminal, and the ending of the sequence as the C-terminal.

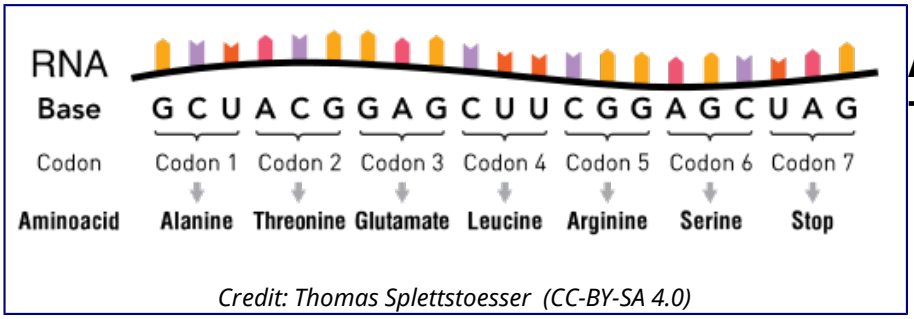




Ribosomes are large complexes of enzymes that coordinate the decoding of mRNA into amino acids to generate proteins.



The standard genetic code.



Credit: Thomas Splettstoesser (CC-BY-SA 4.0)

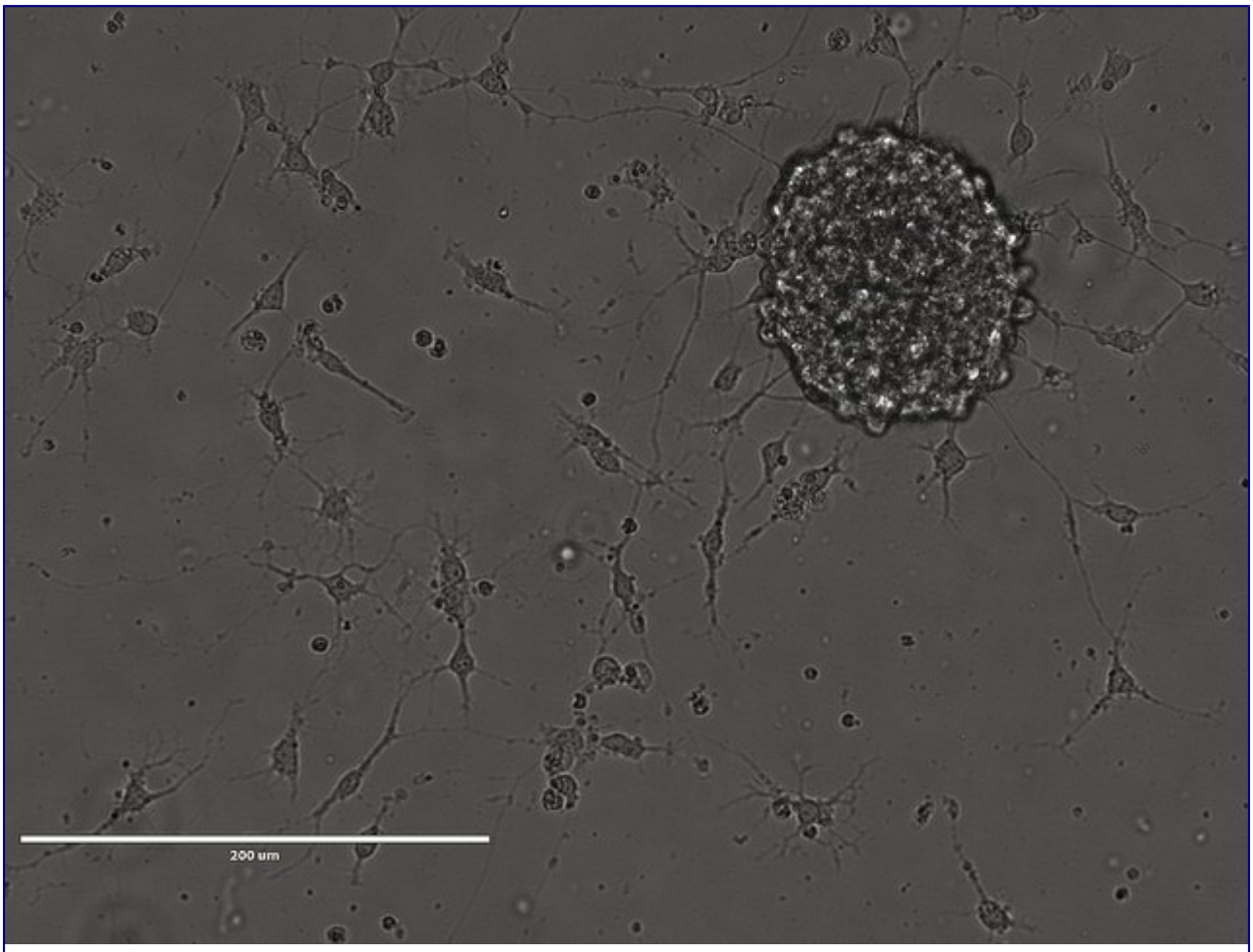
Advanced video of Translation

https://youtu.be/7EZ87bIvCOM

Jeremy Seto jseto@citytech.cuny.edu

Decisions... decisions...

What kinds of decisions are made for stem cells to differentiate into different cell types? What types of regulation occur during this process?



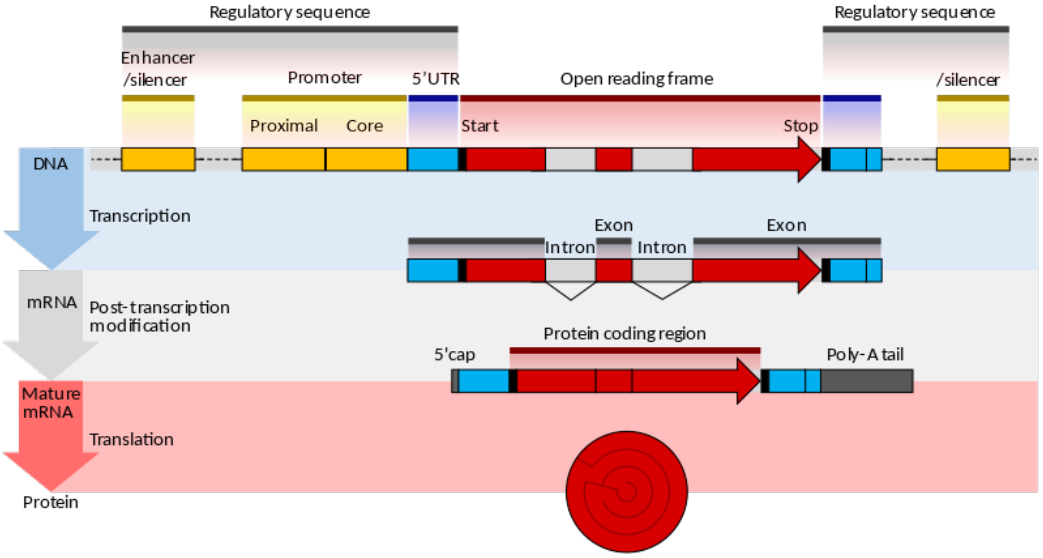
A cluster of neuronal progenitor cells (neurosphere) dissociates and differentiates into neurons.
Credit: Jeremy Seto (CC-BY-NC-SA) https://flic.kr/p/LJR6pY

Video Review

- Transcription Simulation (MIT License)
- **Transcription (CC-BY-NC-ND Cold Spring Harbor Lab - DNA Learning Center)**
  https://www.dnalc.org/content/c15/15510/transcription_basic.mp4
- **Translation (CC-BY-NC-ND Cold Spring Harbor Lab - DNA Learning Center)**
  https://www.dnalc.org/content/c15/15501/translation_basic.mp4

Jeremy Seto jseto@citytech.cuny.edu

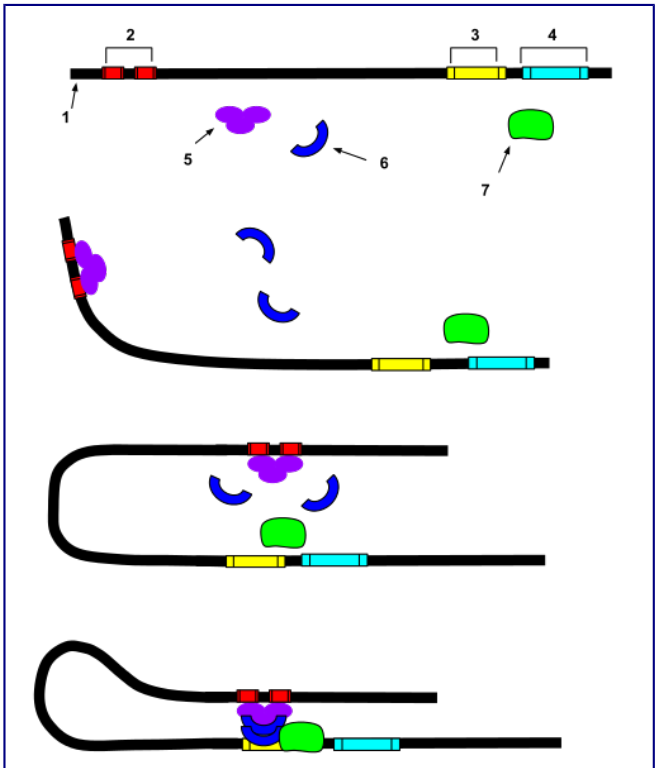
Eukaryotic gene expression



The Central Dogma in Eukaryotes. Genomic DNA of genes often contain introns that are spliced out when an RNA matures to a mRNA. This excision of introns can result in splice variants of the same gene with variants of the same protein.
Credit: Thomas Shafee (CC-BY 4.0)

Unlike prokaryotic genes, expression of genes in eukaryotic cells have complex systems of transcription factors that act on promoters to recruit RNA polymerases. Additionally, **enhancer elements** may reside many kilobase upstream of the promoter. These enhancers strengthen the transcription of the gene. In this case, **transcription activator proteins** or trans-activators augment the promoter activity.

Mediator proteins (**coactivators**) form a multiprotein complex with the activators to recruit RNA polymerase to the promoter.

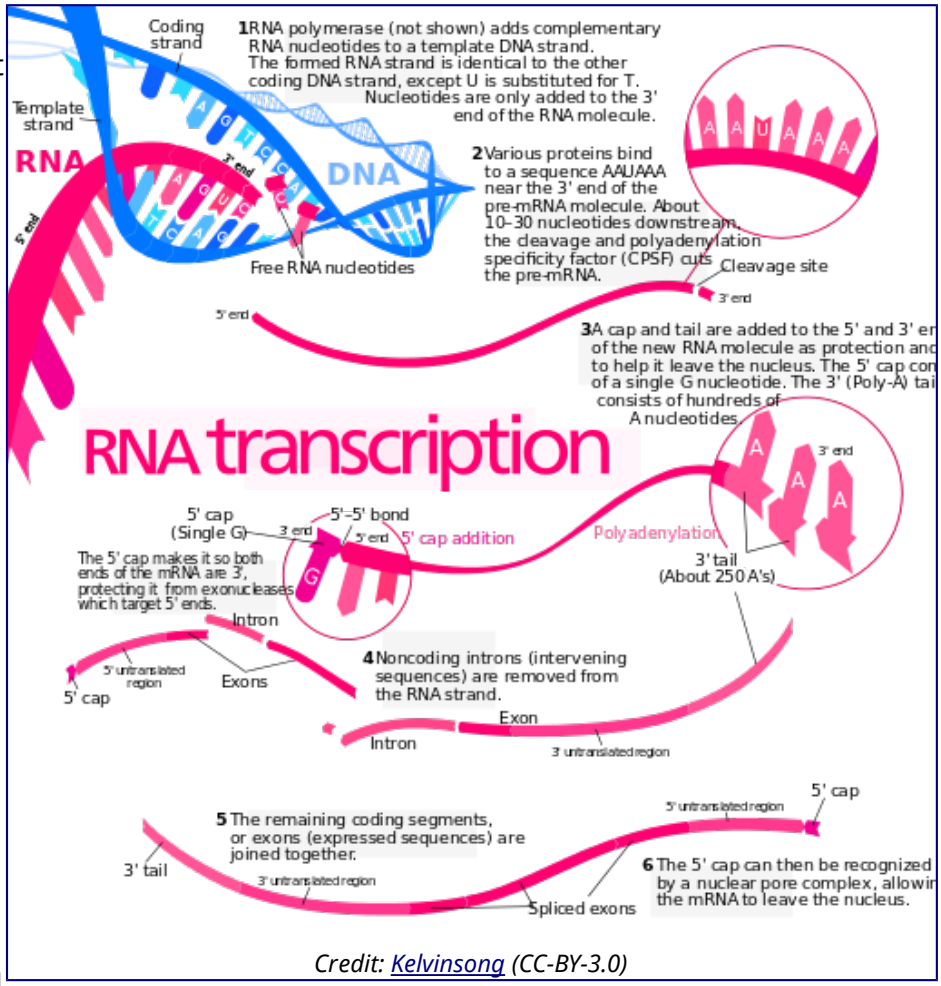


1. DNA 2. Enhancer 3. Promoter 4. Gene 5. Transcription Activator Protein 6. Mediator Protein 7. RNA Polymerase *Credit:* Jon Cheff (CC-BY-SA 4.0)

Jeremy Seto jseto@citytech.cuny.edu

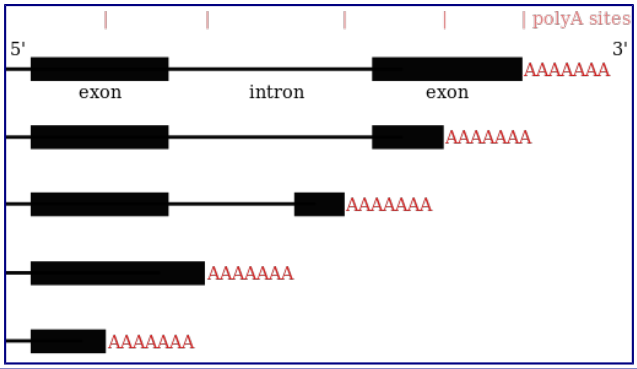
Eukaryotic mRNA

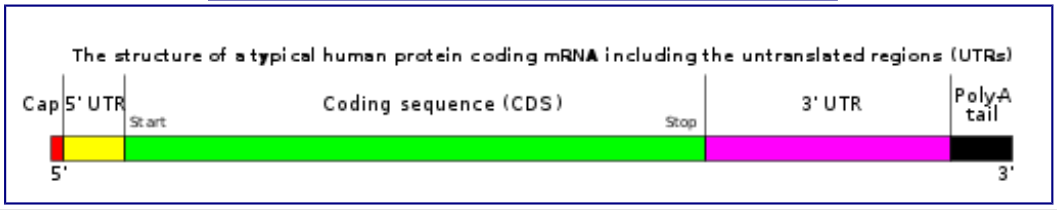
Eukaryotic genes may often contain **introns** (non-coding sequences) that are spliced out from the **exons** (coding sequences). This complexity permits for increased variety of gene products. Mature eukaryotic mRNAs conatins a 5'-methyl-Guanine followed by an untranslated leader sequence (**5'-UTR**), the coding sequences (**cds**), a 3'-untranslated region (**3'-UTR**) and a long stretch of Adenines (polyA tail).



Credit: Kelvinsong (CC-BY-3.0)

Expression is most easily measured with RNA since nucleic acid manipulation is fairly simple with 4 different nucleotides. In eukaryotes, the messenger RNA (mRNA) intermediate that is transcribed from DNA contains a polyA tail that is used to separate these messages from other types of RNA that are abundant within cells (like ribosomal RNA). Through the use of an enzyme called **reverse transcriptase** (RT) and primers composed of deoxy-Thymidines (**oligo-dT** or dT_{18}), mRNA can be converted into a single strand of DNA that is complimentary to the mRNA. This complimentary DNA is called **cDNA**. cDNA is very stable compared to the highly labile mRNA and is used for subsequent processing.





Jeremy Seto jseto@citytech.cuny.edu

Advanced Video of Eukaryotic Transcription Regulation

The first video describes the discovery of transcription factors that regulate the expression of eukaryotic genes.

https://youtu.be/ugMJrhQSfm8

The second video describes the complexity of gene expression that involve chromatin remodeling and enhancers. This video explores the the roles and outcomes of differential gene expression.

https://youtu.be/vDYO7V4xS_A

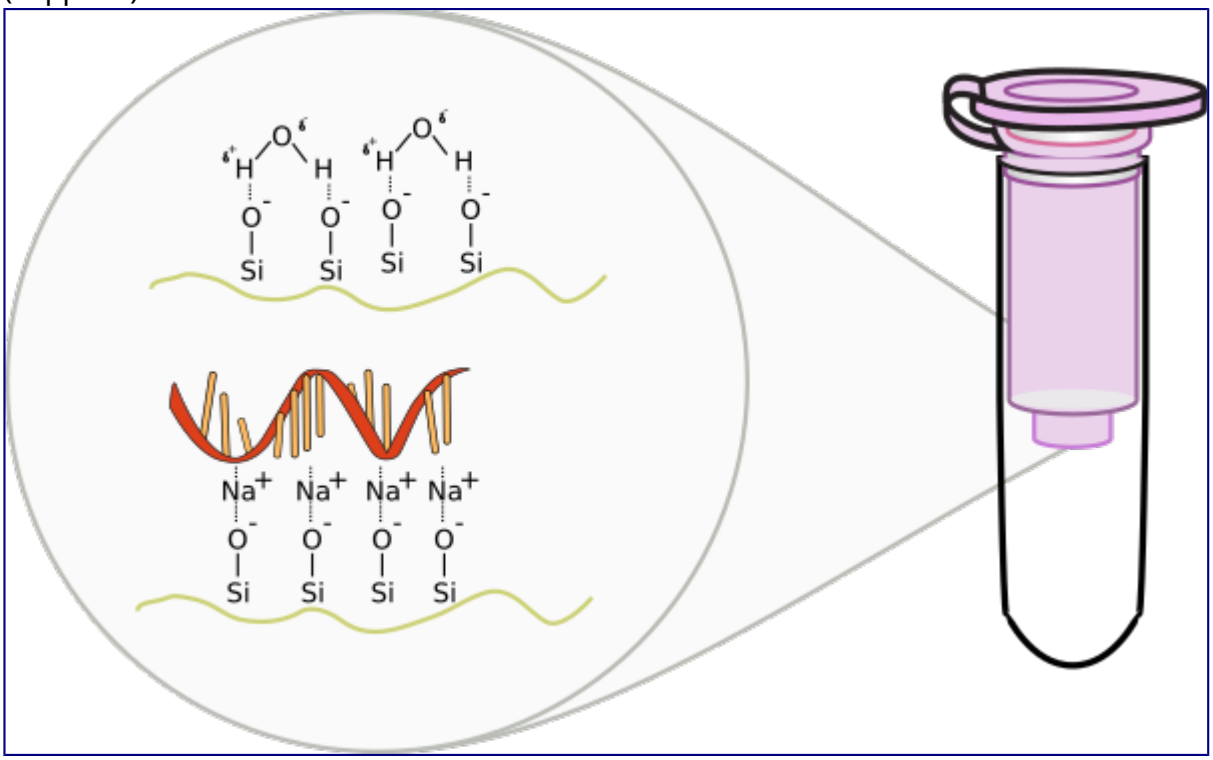
Jeremy Seto jseto@citytech.cuny.edu

Exercise: RNA miniprep

RNA purification occurs similar to DNA preparations. A silica based column is used where DNA is excluded from binding based on size and through an additional DNA digestion step using the enzyme DNase I. RNA is extremely fragile and prone to degradation. Because of this, separate pipettes and plastics are usually used in labs to reduce the amount of exposure to environmental or experimental RNase. When handling RNA, be extremely careful of contaminating the buffers or samples. Always wear gloves as skin carries RNase enzymes. Refrain from talking as to not contaminate the area with RNase found in saliva.

RLT = RNA lysis buffer: contains guanidine, a harsh denaturant
RW1 = RNA wash buffer
RPE = Second RNA wash buffer with Ethanol
RDD = DNase digestion buffer

1. **Harvest** a maximum of 1 x 10^7 cells, as a cell pellet or by direct lysis in the vessel. Add the appropriate volume of Buffer RLT and vortex vigorously.
    - If < 5 x 10^6 cells → 350 µl RLT (< 6cm plate)
    - if ≤ 1 x 10^7 cells → 600 µl RLT (6-10cm plate)
2. Add 1 volume of 70% ethanol to the lysate, and mix well by pipetting. Do not centrifuge. Proceed immediately to next step.
3. Transfer up to 700 µl of the sample, including any precipitate, to an spin column placed in a 2 ml collection tube
(supplied).



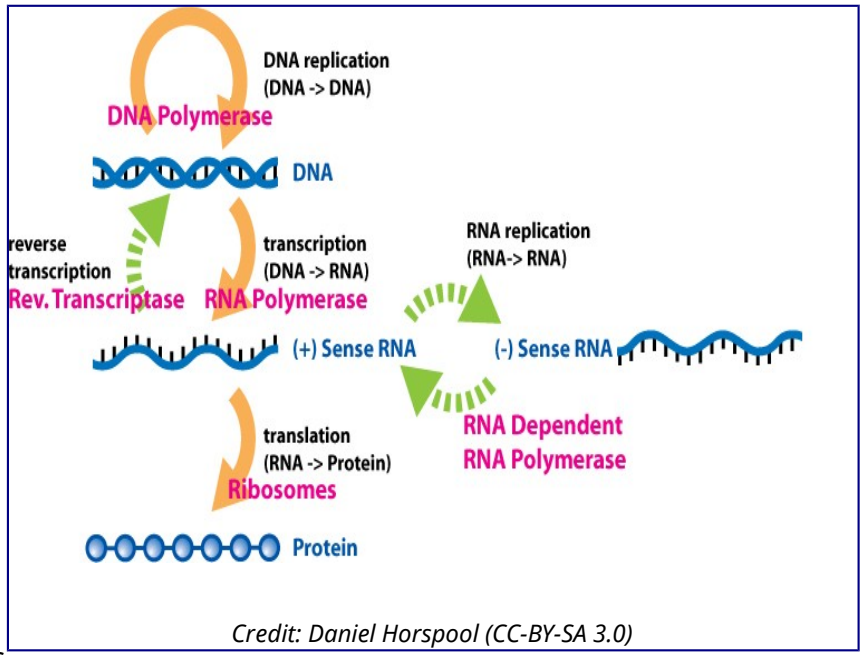
1. Close the lid, and centrifuge for 15 s at ≥8000 x g.
2. Discard the flow-through.

Jeremy Seto jseto@citytech.cuny.edu

4. **Wash**: Add 350 µl Buffer RW1 to spin column, close lid, centrifuge for 15 s at ≥8000 x g (≥10,000 rpm). Discard flow-through.
5. Add 10 µl DNase I stock solution (see above) to 70 µl Buffer RDD. Mix by gently inverting the tube
6. **Remove DNA** (optional): Add DNase I incubation mix (70 µl) directly to spin column membrane, and place on benchtop (20–30°C) for 15 min.
7. **Wash**: Add 350 µl Buffer RW1 to spin column, close lid, centrifuge for 15 s at ≥8000 x g. Discard flow-through.
8. Add 700 µl Buffer RW1 to the spin column. Close the lid, and centrifuge for 15 s at ≥8000 x g. Discard the flow-through.
9. Add 500 µl Buffer RPE to the spin column. Close the lid, and centrifuge for 15 s at ≥8000 x g. Discard the flow-through.
10. Add 500 µl Buffer RPE to the spin column. Close the lid, and centrifuge for 2 min at ≥8000 x g.
11. Discard all flow-through and centrifuge at full speed for 1 min to dry the membrane.
12. Place the spin column in a new 1.5 ml collection tube. Add 30 µl RNase-free water directly to the spin column membrane.
13. Close the lid, and centrifuge for 1 min at ≥8000 x g to elute the RNA.
14. Add 30 µl RNase-free water directly to the spin column membrane. Close the lid, and centrifuge for 1 min at ≥8000 x g to elute the RNA.

Reverse Transcription

The **Central Dogma** of Molecular Biology was proposed by Francis Crick, the co-describer of the double stranded helical structure of DNA. This “dogma” was a statement to describe the flow of genetic information to show that DNA houses or stores data that is transcribed into RNA that is subsequently translated from nucleotides into amino acids through the machinery of the ribosomes. Since DNA is relatively static in it's ability to store genetic information, the expression of this stored data into the intermediate RNA or to the final protein product is of great significance. Imagine that the DNA in the nucleus of your cheek cells is identical to the DNA of the nucleus of cells in your liver. While the instructions are identical, these are clearly different cells that have a difference in expression of proteins. Imagine a hard drive on a computer that stores information as 1's and 0's. These 1's and 0's do not have meaning until specific programs are called upon to act on this information. Likewise, different programs are called to use the instructions of your DNA to make a cheek cell different than a liver cell.



Credit: Daniel Horspool (CC-BY-SA 3.0)

Jeremy Seto jseto@citytech.cuny.edu

In 1970, Howard Temin and David Baltimore independently isolated an enzyme from the Rous Sarcoma Virus and Murine Leukemia Virus, respectively. This enzyme was capable of violating the Central Dogma. The genomes of these viruses consist of RNA, not DNA. During the infection process, this enzyme is responsible converting the RNA into DNA in a process called **reverse transcription**. This enzyme is logically called **reverse transcriptase** (RT). This discovery was rewarded with Nobel Prize in 1975. Later on, more viruses were discovered that were composed of RNA genomes that utilized this process, including HIV. Other enzymes within cells were also recognized to have reverse transcriptase activity, such as telomerase and retrotransposases. In molecular biology, these enzymes are used to convert mRNA into complementary copies of DNA called cDNA. The sum total of everything that is transcribed into RNA is referred to as the **transcriptome**. Synthesis of cDNA from any transcribed RNA can then be used for transcriptome analysis.

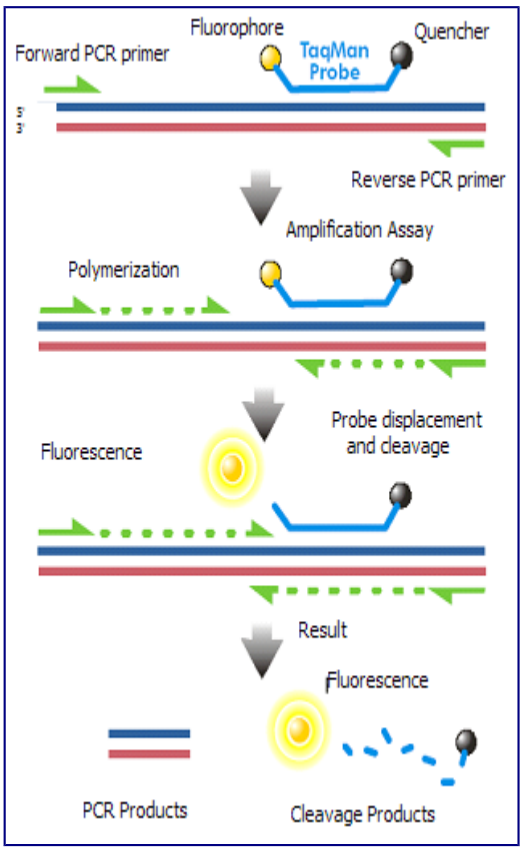
Exercise: Reverse Transcription of Eukaryotic mRNA

mRNA from eukaryotes are modified with 3' polyadenylated tails. Oligo-dT primers can be used to prime the reverse transcription process of all mRNAs. All solutions should be kept on ice.

1. Determine the concentration of total RNA
2. Adjust concentration of RNA to 0.1mg/ml using Rnase free water
3. Combine 10 ml RNA, 1 ml Oligo-dT (50µM), and 1 ml dNTP Mix (10 mM each)
4. Denature mixture at 65°C for 5 minutes and then place on ice
5. Combine the following in a separate tube
    1. 4µl Buffer 5X → contains all salts and pH buffer
    2. 2µl 0.1 M DTT → a reducing agent to mimic the cellular environment
    3. 1µl RNaseOUT (40U/µl) → an RnaseA inhibitor
    4. 1µl SuperScript III RT → the reverse transcriptase enzyme
6. After the denatured mixture has been sufficiently cooled, ad 8µl enzyme mixture
7. incubate 45°C for 1 hour
8. deactivate enzyme by incubating 75°C for 10 minutes
9. Store your cDNA in the freezer

Jeremy Seto jseto@citytech.cuny.edu

Quantitative PCR (qPCR)



Measurements can be made of individual genes of interest through PCR of those specific genes. A process known as Real-Time PCR or quantitative PCR (**qPCR**) is used to measure individual genes using fluorescence measurements. An intercalating agent that binds only to double stranded DNA called **Sybr Green** is used in a qPCR machine that is measuring fluorescence after each cycle of PCR indirectly indicates the amount of amplified product. However, non-specific products of amplification may also be measured and not discriminated from the authentic amplicon.

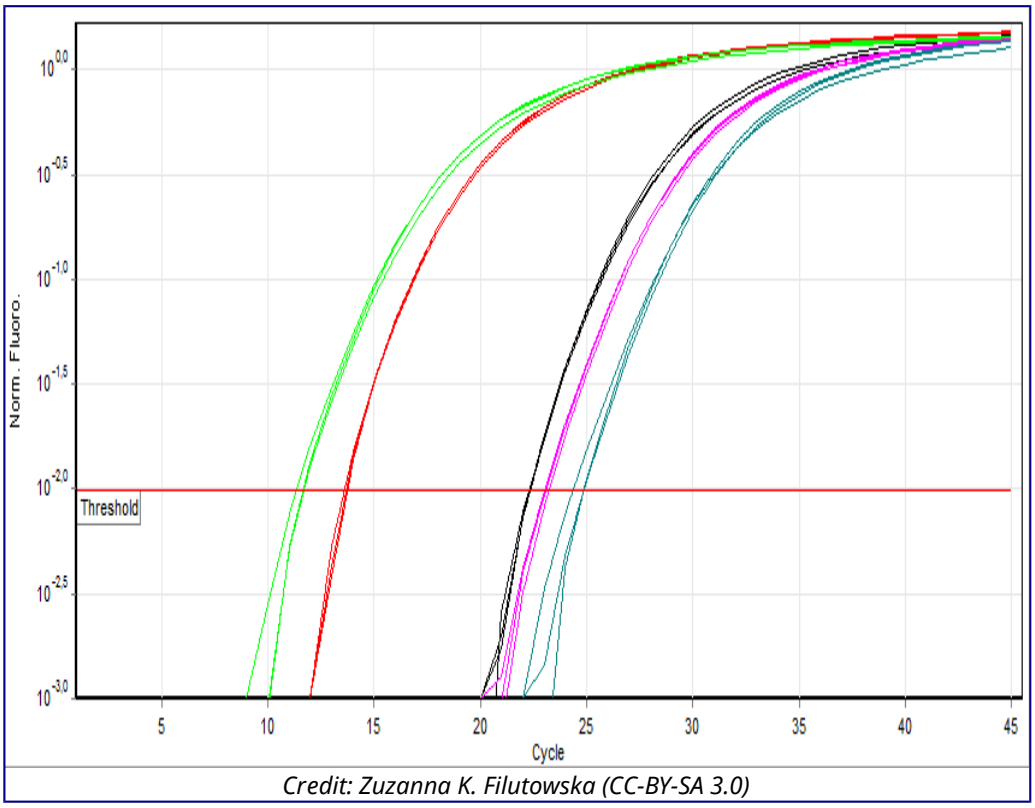
An alternative to Sybr Green is exemplified by the TaqMan technology. With TaqMan, a third primer (**TaqMan probe**) is designed in the middle of the area to be amplified. This middle primer is designed with a hairpin self-complimentarity so that the 5' and 3' ends are in close proximity. At one end, a **fluorescent reporter** is attached while the other terminus has a **quencher** that absorbs any fluorescence signal. Under normal circumstances, measurements of fluorescence will be very low. When PCR extension occurs, the Polymerase hydrolyzes this middle primer, thereby separating the quencher and reporter. The name TaqMan is a play on words since it is imagined that the polymerase is chewing up the probe like Pacman. With increased distance between quencher/reporter, fluorescence signal from this probe can now be measured. This method is much more specific than Sybr Green, however the use of specific probes increases the cost considerably.

Threshold Cycles (C_t)

Fluorescence measurement early during the PCR process will be very low due to small number of dsDNA molecules (Sybr Green) or most TaqMan primers being quenched. During this exponential DNA production, a threshold will be reached in which the fluorescence will linearly increase. A specific point where the fluorescence is clearly measurable called the **Threshold Cycle** (C_t) is used as a reference point to compare expression values.

Jeremy Seto jseto@citytech.cuny.edu

Looking at the example of Sybr Green qPCR above, it can be observed that samples exponentially increasing at a lower Cycle number (C_t) has a higher level of mRNA expression (towards the left) of that gene than samples with higher cycle number (towards the right). Notice that the fluorescence eventually plateaus and stops increasing. This is due to the depletion of raw materials for DNA production like dNTPs.

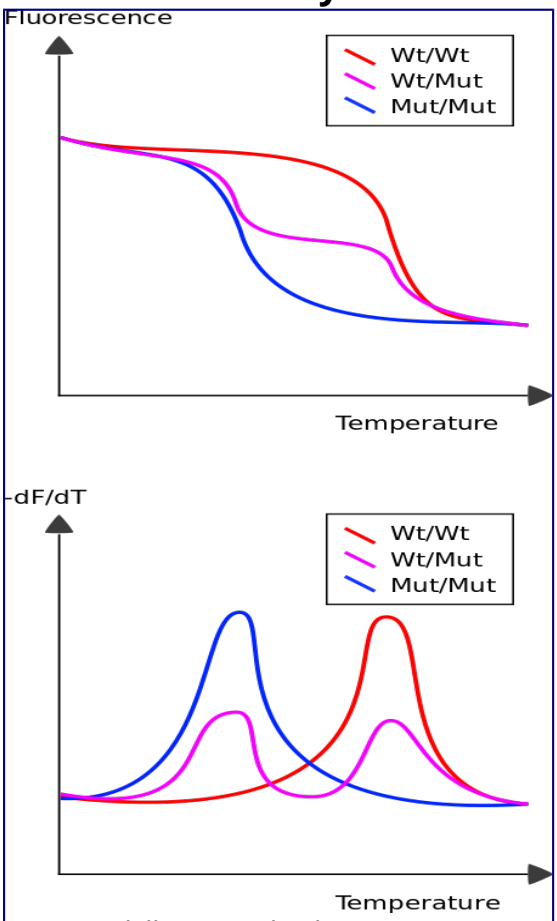


Credit: Zuzanna K. Filutowska (CC-BY-SA 3.0)

Since the PCR reactions theoretically represent a doubling of DNA after each cycle, the C_t values can be interpreted on a base 2 system. If there is a difference in C_t between two samples (**ΔC_t**) of 5 cycles, this corresponds to 2^5 or 32 fold difference. We can control for variations in the RNA preparation through comparing the fluorescence values of our gene of interest to a housekeeping gene like actin. The use of a house-keeping gene to normalize the initial input to the reactions and comparison between samples is referred to as **Relative Quantification**.

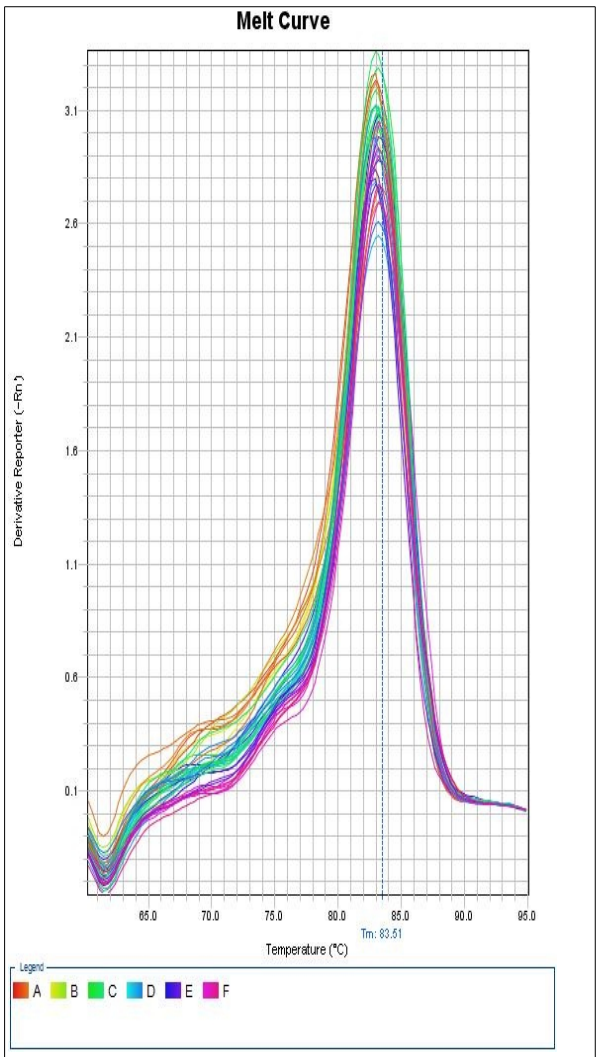
Jeremy Seto jseto@citytech.cuny.edu

Melt Curves for Sybr Green



Top panel illustrates the decrease in fluorescence as the temperature increases due to the dissociation of double stranded DNA. Bottom panel illustrates the first derivative plot. Each peak in this example illustrates a different allele. The double peaks represent the presence of the 2 distinct alleles in the amplification products.

Credit: Seans Potato Business (CC-BY-SA 3.0)



This melt curve illustrates each sample contains the same specific product with a melting temperature of 83.51°C.

When using Sybr Green, we need to ensure that the PCR is specific so that the fluorescence measurement truly reflect amplification of our gene of interest. At the end of each qPCR run (~40 cycles), a melt curve is performed. A **melting curve** (or dissociation curve) comes from constant measurements as the temperature is increased. As temperature increases, the DNA strands start to denature and fluorescence will begin to decrease. After complete separation of DNA strands, the fluorescence will again remain constant. The way this curve is viewed is through a derivative plot where the inflection in fluorescence reading is reported as the **melting temperature (T_m)**.

Any peaks in this plot refer to a specific PCR product. If multiple peaks appear, the results will not be valid as they do not directly measure a single product.

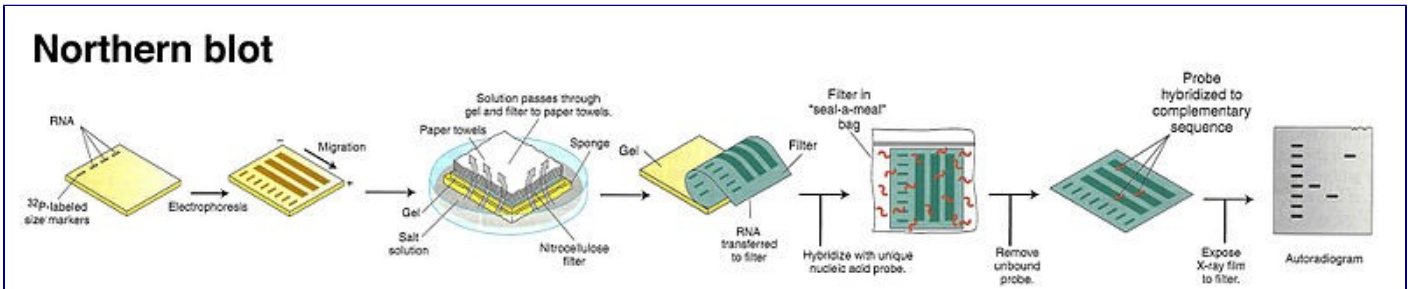
Jeremy Seto jseto@citytech.cuny.edu

Expression measurements

Differential gene expression refers to transcriptional programs activated by the cell under various conditions. "Differential" refers to a comparison of two or more states or timepoints. Using mRNA as an indirect measurement of protein, one can ascertain which proteins are linked to these different states. In eukaryotes, this can be assessed by enriching total RNA for polyA-containing mature mRNA. Through the use of oligo-(dT) containing resin, mRNA can be separated from non-protein encoding RNA. Likewise, performing a reverse-transcription using an oligo-(dT) primer will create a stable complimentary DNA (cDNA) molecule that can be used with PCR. Using qPCR in this way is called **RT-PCR** or reverse-transcription polymerase chain reaction where specific primer pairs are used to amplify a small portion of a known gene.

Hybridization based methods and Microarrays

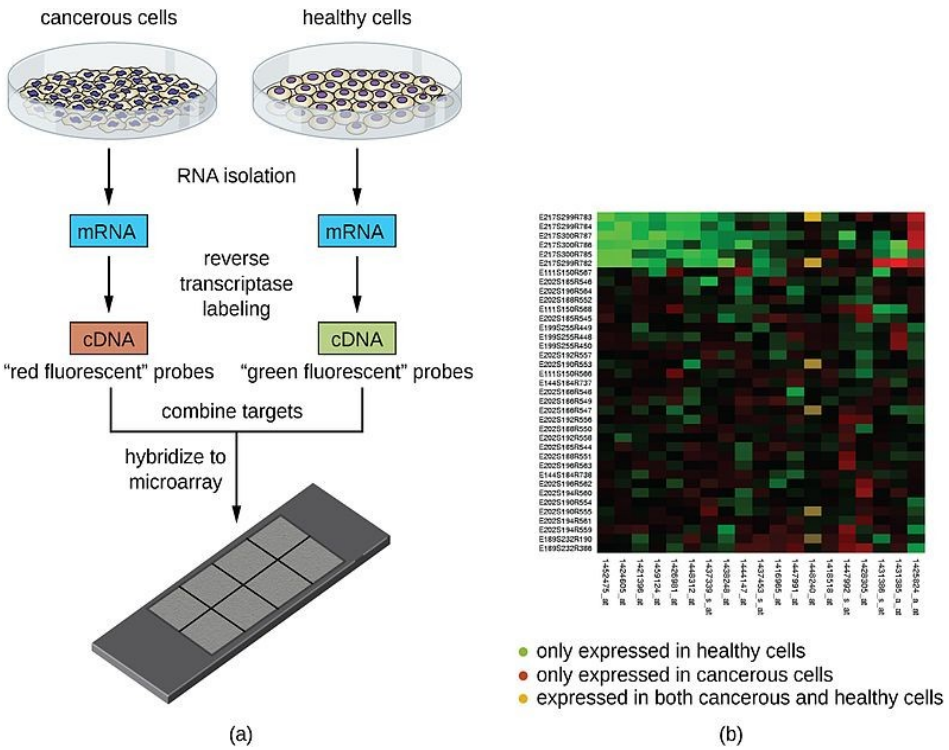
Prior to RT-PCR, expression of individual genes was assessed through a hybridization-based approach. This method called for running RNA on an agarose gel and transferring the size-fractionated RNAs onto a membrane through a method called "blotting". This transferred RNA was then hybridized to a radioactively labelled probe for a specific gene (corresponding to the reverse complimentary sequence) and visualized by exposure to X-ray film in a process called **Northern Blotting**. The intensity of the band would be proportional to the amount of mRNA corresponding to the gene of interest. Re-probing with a housekeeping gene like actin would be used as a loading control to illustrate that a similar amount of total RNA was loaded into each well. Differences in sizes of the mRNA on the Northern Blot also revealed differences in splice variants of mature mRNA in the different states.



Credit: FrozenMan (CC-BY-SA 4.0)

This technique was later adapted using non-radioactive methods. Using these non-radioactive methods, the reverse protocol was developed to measure multiple gene targets. By systematically immobilizing gene specific probes onto a membrane or a microscope slide, an array of targets can be produced. In the simplest paradigm of having 2 states (control or experimental), cDNA from each sample can be used to generate fluorescent RNA that can hybridize to immobilized probes. Using 2 different fluorescent markers allows for the competitive hybridization onto the array whereby the fluorescent signal in each channel can reveal the differential gene expression of the two states in a 2-color **microarray**.

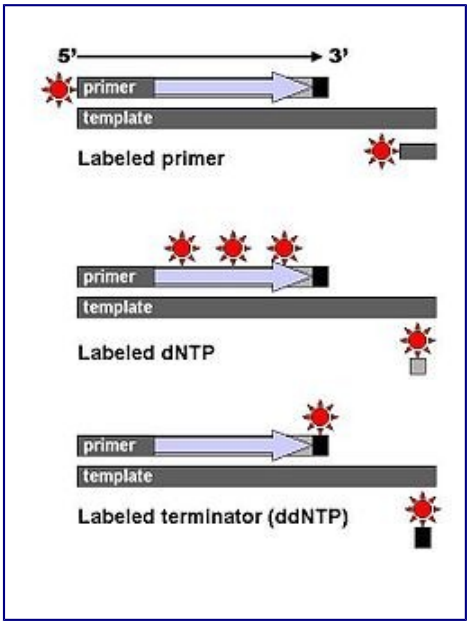
Jeremy Seto jseto@citytech.cuny.edu



Credit: CNX OpenStax (CC-BY 4.0)

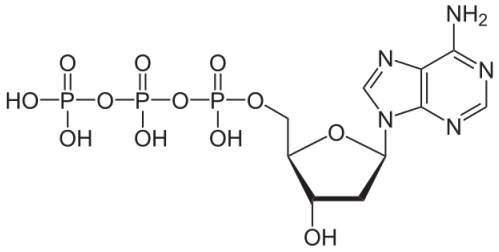
Jeremy Seto jseto@citytech.cuny.edu

Radioactive Chain Termination

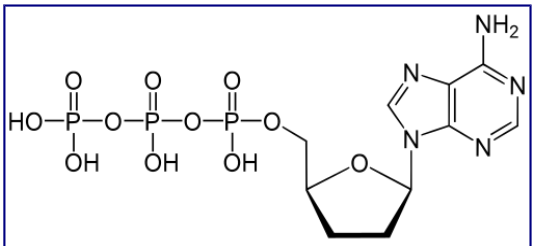


The polymerization of nucleic acids occurs in a 5' → 3' direction. The 5' position has a phosphate group while the 3' position of the hexose has a hydroxyl group. Polymerization depends on these 2 functional groups in order for a dehydration synthesis reaction to occur and extend the sugar-phosphate backbone of the nucleic acid. In the 1970's, Fred Sanger's group discovered a fundamentally new method of 'reading' the linear DNA sequence using special bases called **chain terminators** or **dideoxynucleotides**. The absence of a hydroxyl group at the 3' position blocks the polymerization resulting in a termination. This method is still in use today it is called: Sanger dideoxynucleotide chain-termination method. This method originally used a radioactively labeled primer to initiate the sequencing reaction. Four reactions take place where each reaction is intentionally "poisoned" with a dideoxy chain terminator. For example, 1 reaction will have all 4 dNTPs (deoxynucleotide triphosphates)with the addition to a small amount of ddATP (dideoxyadenosine triphosphate). This reaction will result in a series of premature terminations of the polymerization specifically at different locations where an Adenine would be incorporated.

dATP is a natural monomer used in the polymerization of DNA. The 3'-OH is the attachment point of the next subsequent nucleotide.



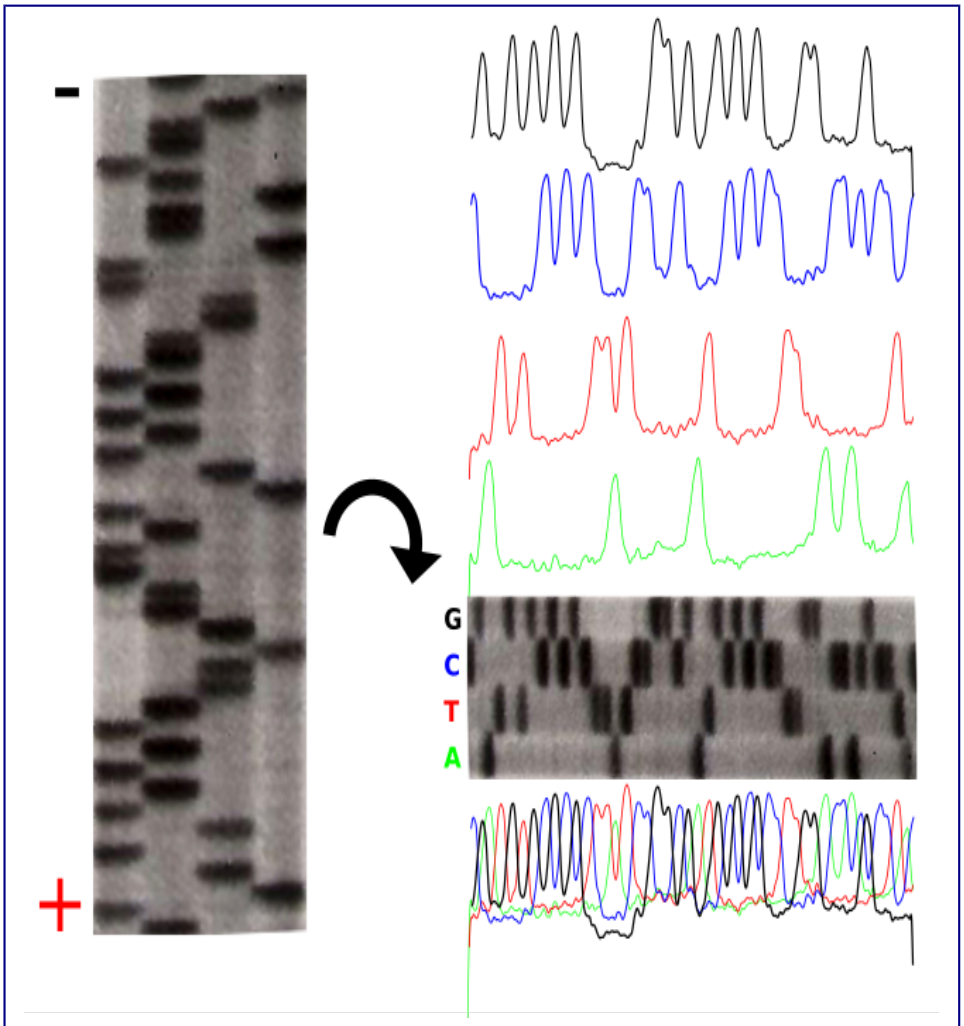
dATP is a natural monomer used in the polymerization of DNA. The 3'-OH is the attachment point of the next subsequent nucleotide.



The lack of a 3'-OH in this molecule of ddATP makes it a chain terminator that will prohibit the addition of another nucleotide to the DNA polymer

The product of these 4 separate sequencing reactions is run on a large polyacrylamide sequencing gels. The smallest fragments run through the gel the fastest and create a ladder-like pattern. This can be visualized through use of an x-ray film that is sensitive to the radioactivity. Each lane of the gel corresponds to one of the four chain terminating reactions. The bases are read sequentially from the bottom up and reveals the sequence of the DNA.

Jeremy Seto jseto@citytech.cuny.edu

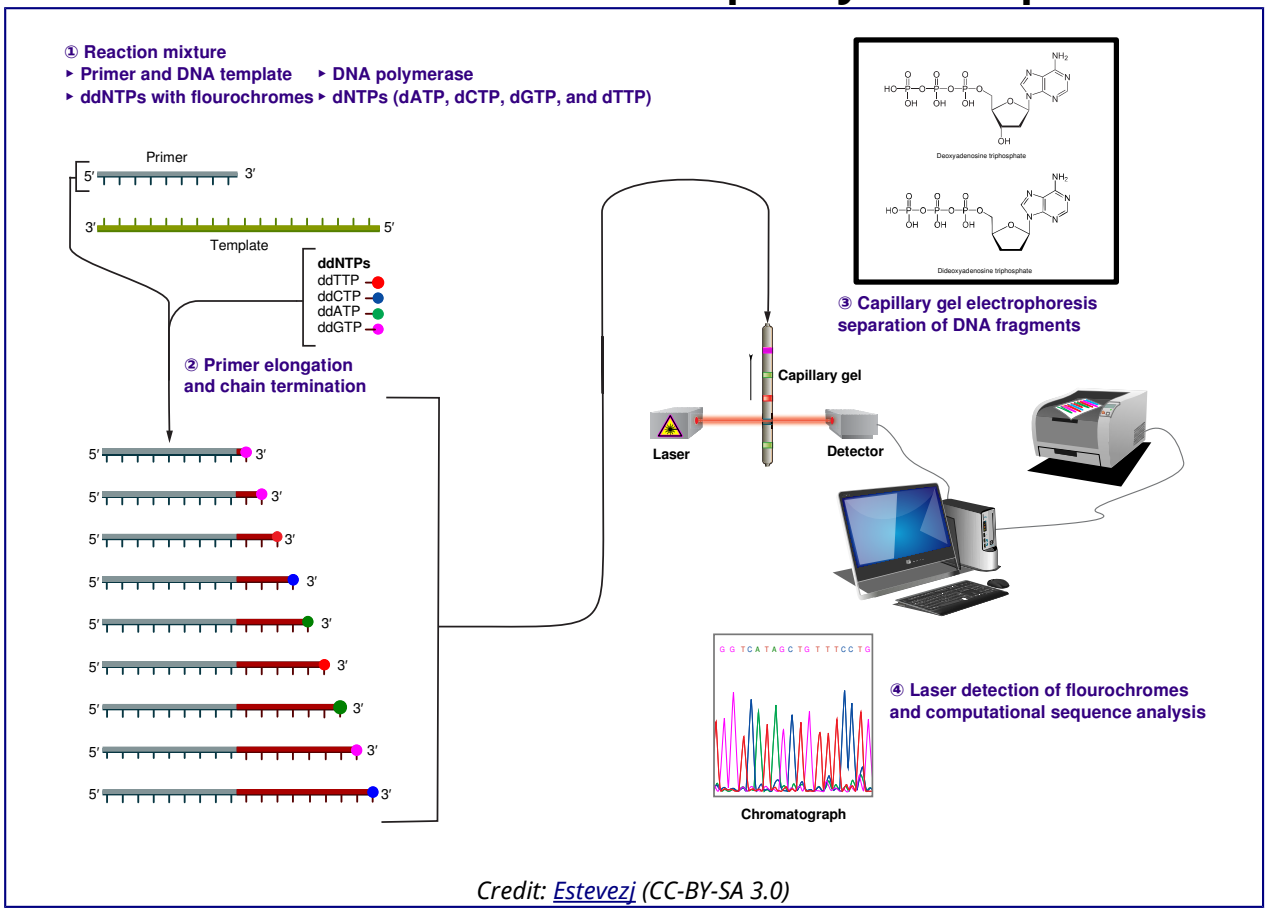


The sequencing gel can be manually scored. The profiles of each lane have been created using ImageJ to illustrate the banding pattern and subsequent sequence.

Credit: John Schmidt & Jeremy Seto (CC-BY-SA 3.0)

Jeremy Seto jseto@citytech.cuny.edu

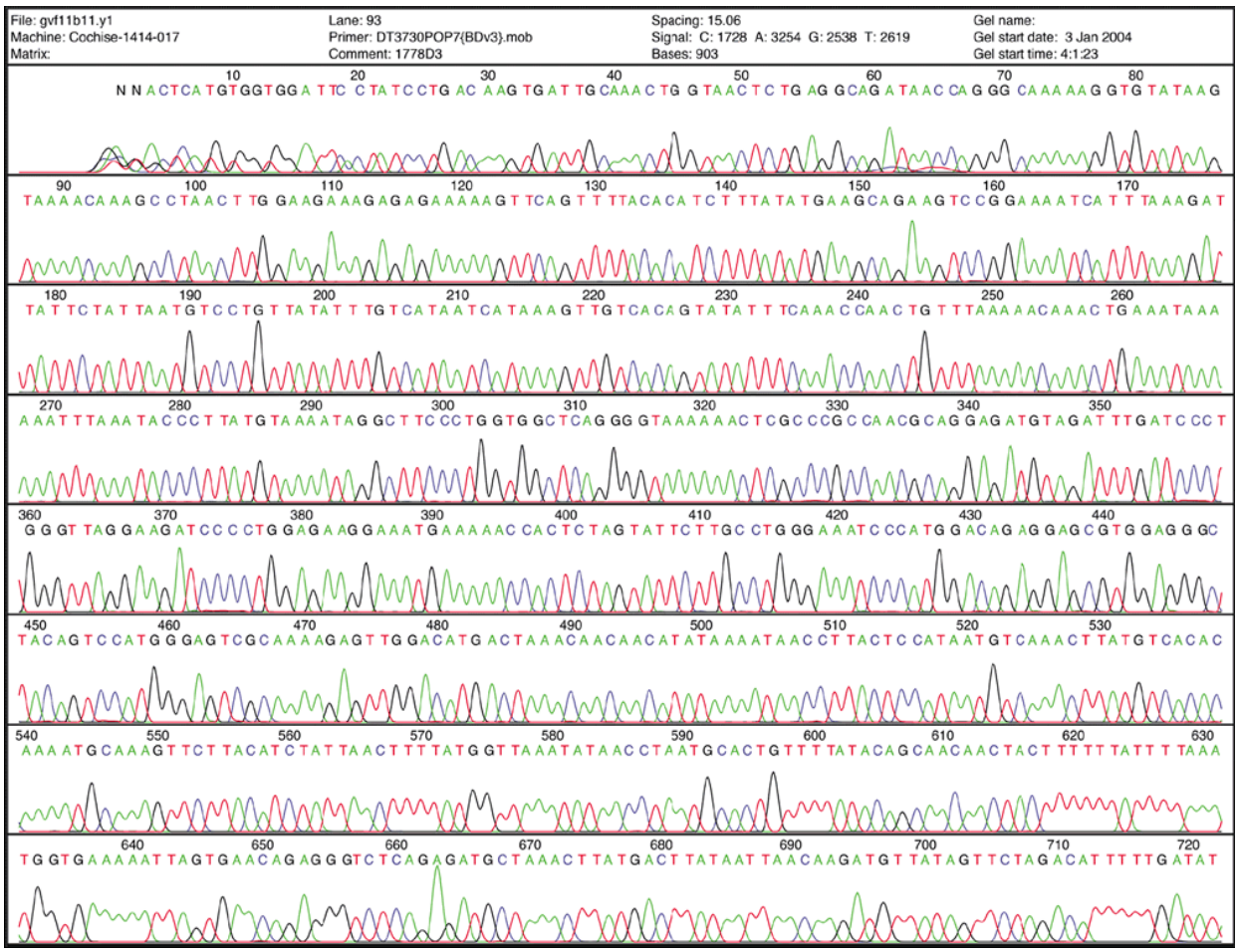
Fluorescent Chain Termination and Capillary Electrophoresis



Credit: Estevezj (CC-BY-SA 3.0)

Radioactivity is dangerous and undesirable to work with so chain terminators with fluorescent tags were developed. This method synthesizes a series of DNA strands that are specifically fluorescent at the termination that is passed through a capillary electrophoresis system. As the fragments of DNA pass a laser and detector, the different fluorescent signal attributed to each ddNTP is identified and generates a chromatogram to represent the sequence. Fluorescent Chain Terminators are now used in reactions and run through a small capillary. The smallest fragments run through first and are detected to reveal a chromatogram .

Jeremy Seto jseto@citytech.cuny.edu



Fluorescent Chromatograms are sued to score the the nucleotide chain termination. The amplitude of each peak corresponds to the strength or certainty of the nucleotide call. Chromatogram files are usually provided alongside the sequence file with the extension ***.ab1** while the sequence files are provided as a text file in the **fasta** format. More about these files can be found **here**. The ab1 files are extremely important to analyze when there is ambiguity or sequencing errors. These ab1 files can also be used to ascribe a quality score on the base call.

When there is too much ambiguity in the signal because of multiple peaks, you will often find a **N** in place of one of the 4 nucleotides (A,T,C,G).

This video (source: www.yourgenome.org CC-BY) illustrates the mechanism of fluorescent chain termination and capillary electrophoresis.

https://youtu.be/ONGdehkB8jU

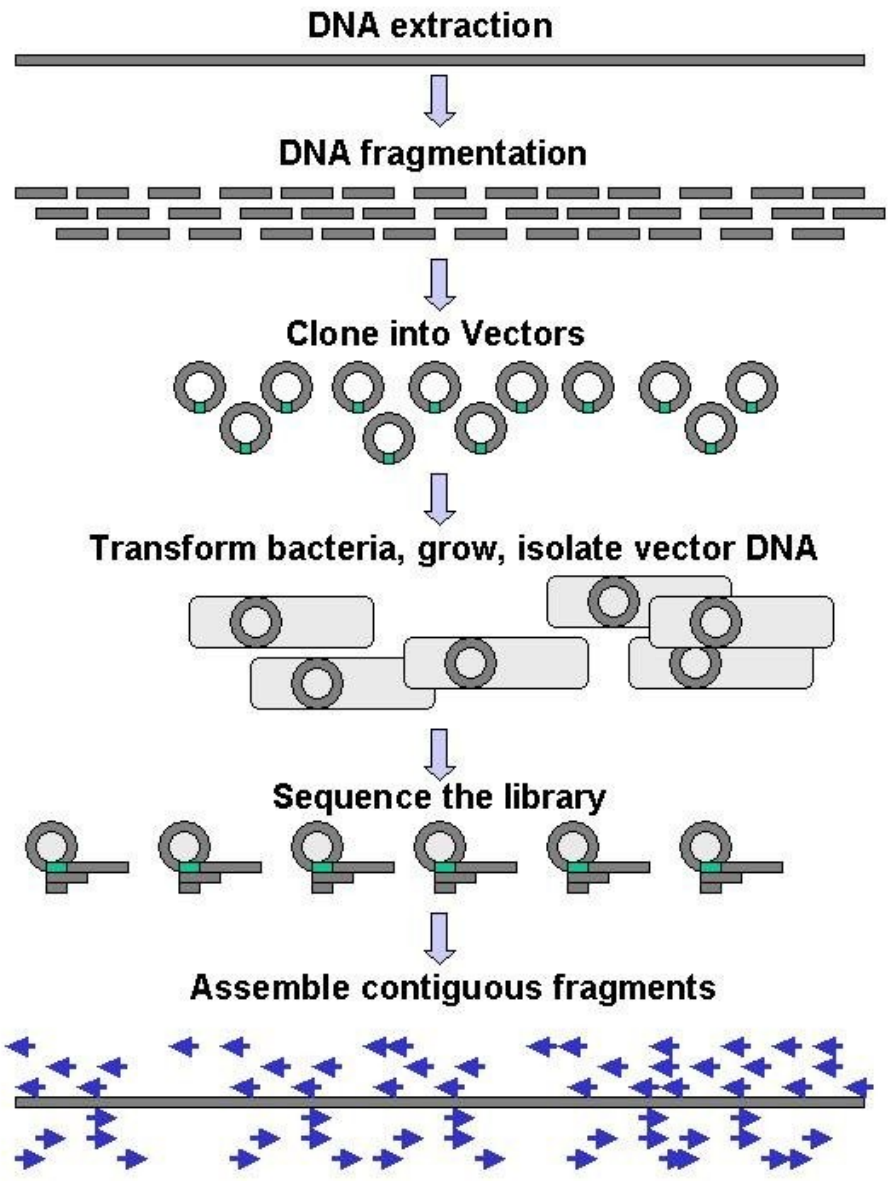
Jeremy Seto jseto@citytech.cuny.edu

Sequencing Genomes

Traditional sequencing of genomes was a long and tedious process that cloned fragments of genomic DNA into plasmids to generate a genomic DNA library (**gDNA**). These plasmids were individually sequenced using Sanger sequencing methodology and computational was performed to identify overlapping pieces, like a jigsaw puzzle. This assembly would result in a draft scaffold.

The video below is taken from yourgenome.org (CC-BY) and illustrates the sequencing of the human genome through the shotgun sequencing approach.

https://youtu.be/WX8V1SWQbFw

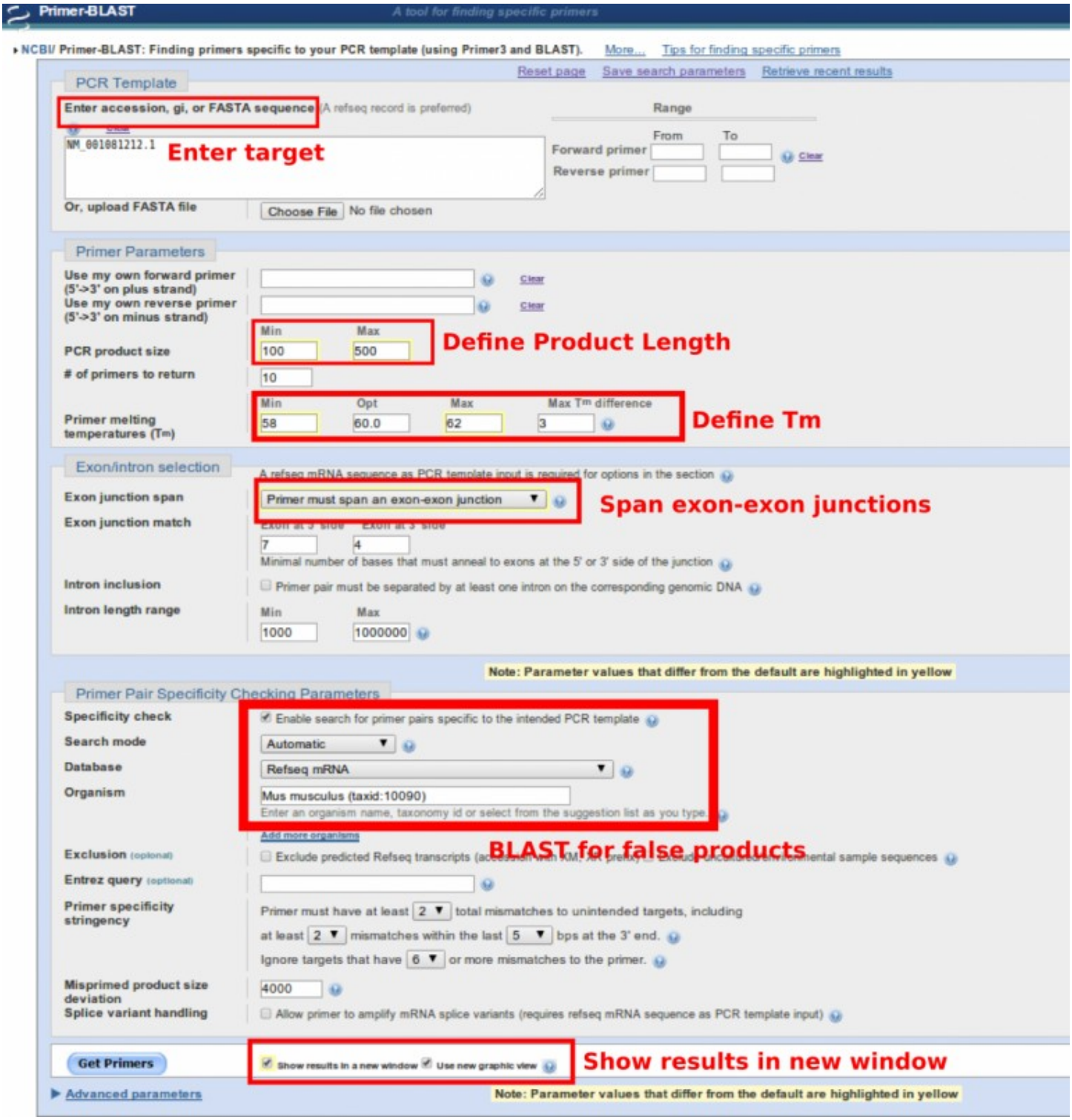


Jeremy Seto jseto@citytech.cuny.edu

Primer-BLAST is a combination of a program called **Primer3** that aids in the design of primers with specific properties and BLAST. Primer-BLAST allows for the construction of primers for qPCR where the user can specify the melting temperature, reduce the amount of self-priming and to span exon-exon junctions in order to avoid amplification of contaminating genomic DNA. After the design of primers, each primer pair is sent into BLAST to identify if similar products within the genome of the model organism will also be primed and amplified. This process ensures that the primers designed fall within your design parameters and most likely only amplify your gene of interest.

1. Enter the sequence OR the NCBI accession number for the gene of interest
2. Define the PCR product length
    1. limiting the product between 100-500 permits for good efficiency in qPCR
    2. longer products may not be efficiently replicated depending on your cycling protocol
3. Define the desired melting temperature (T_m) of the primers (minimum, optimal, maximum, difference between the set)
    1. 60ºC is fairly high and will aid in the enhanced specificity of the primer with the target during amplification to avoid false priming
    2. Try to have the T_m as close as possible so that they are annealing about equally
4. Choose the option "Primer must span an exon-exon junction"
    1. This aids in amplifying cDNA and not genomic DNA that may be contaminating
    2. do not select this if it a single exon gene as this will fail
5. Select Refseq mRNA as the database to search against.
    1. **Refseq** provides sequences to naturally occurring sequences.
    2. Things like plasmid sequences or vector constructs do not show up in Refseq
6. Select the organism you are BLASTing against
    1. there are options for model organisms as well as cell lines
    2. If you are using something like PC12 cells, you may use *Rattus norvegicus* or PC12 genome since that is also an option in the database
7. Evaluate the location of the primers and the other parameters. We generally choose primers at the 3' end of the RNA since RT reactions often have a 3' bias in eukaryotes by using oligo-dT priming in the reverse transcription

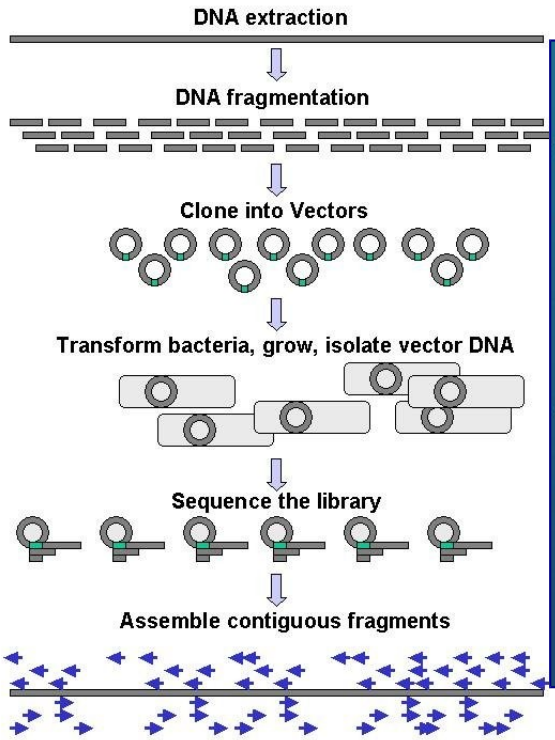
Jeremy Seto jseto@citytech.cuny.edu

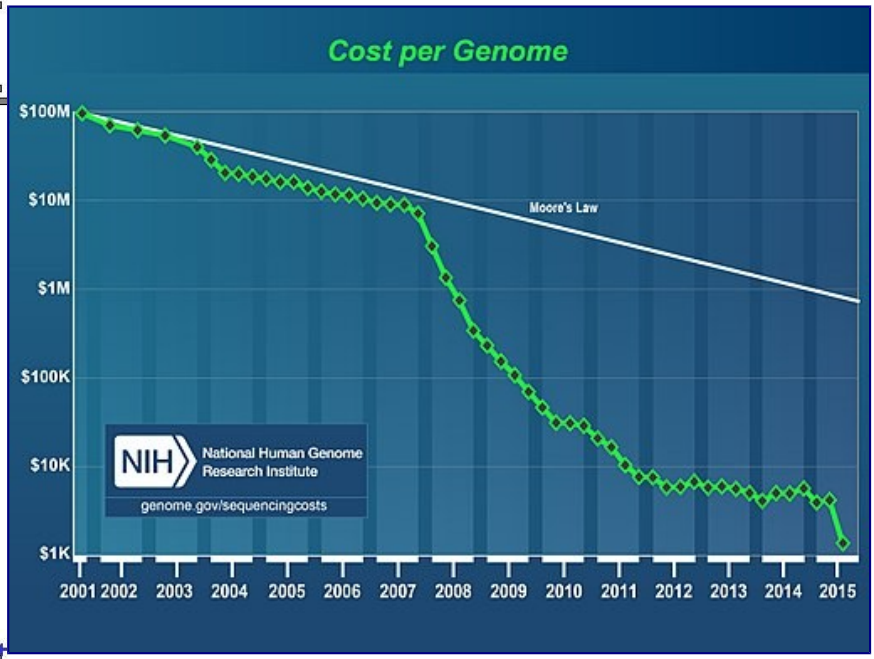


Jeremy Seto jseto@citytech.cuny.edu

Expansion of sequencing technology

Traditional sequencing of genomes was a long and tedious process that cloned fragments of genomic DNA into plasmids to generate a genomic DNA library (**gDNA**). These plasmids were individually sequenced using Sanger sequencing methodology and computational was performed to identify overlapping pieces, like a jigsaw puzzle. This assembly would result in a draft scaffold.

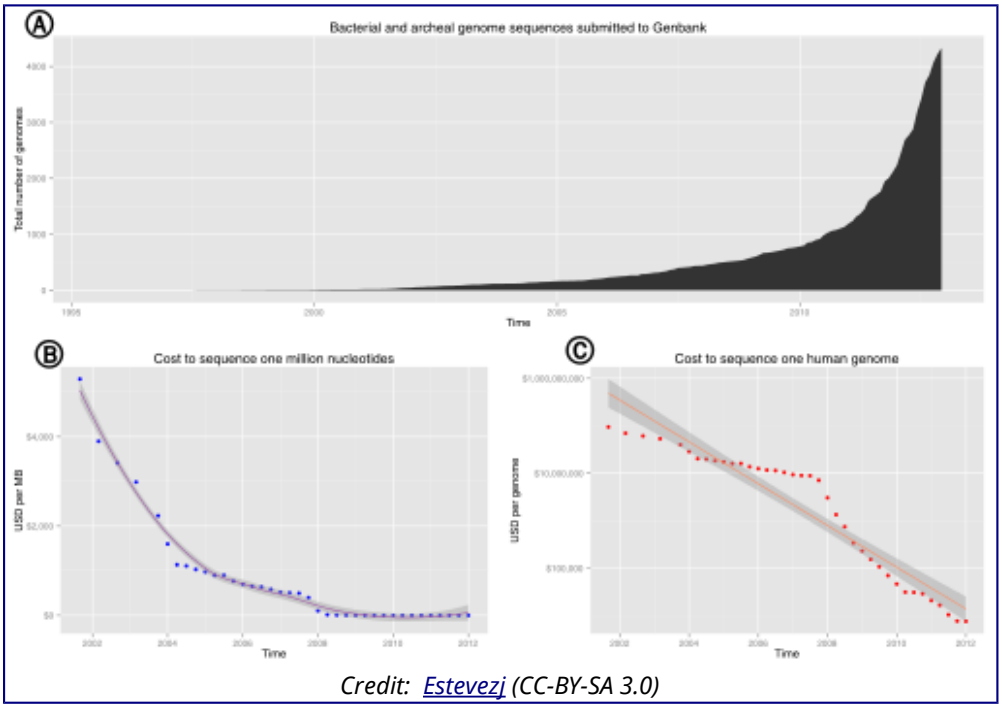




As technology improved, the cost of sequencing genomes became less expensive. This technology outpaced the Moore's Law, a semiconductor projection about the the speed of computers as time progressed. A dramatic price decrease in cost of genome sequencing occurred around 2008 due to technical advances.

Jeremy Seto jseto@citytech.cuny.edu

As the cost of genome sequencing decreased, a dramatic increase in genome deposition into Genbank was observed. These deposits reflected small genomes of bacteria and archaea.

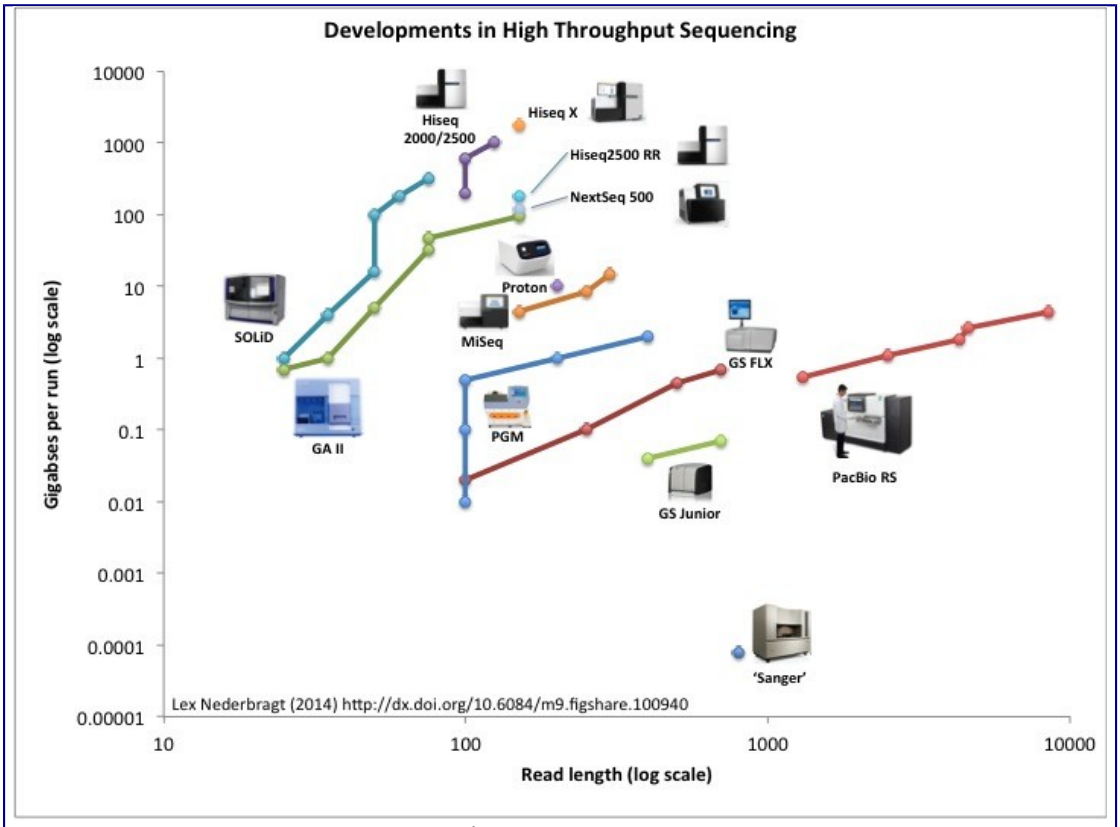


Credit: Estevezj (CC-BY-SA 3.0)

The decrease in per nucleotide sequencing cost came from the parallelization of sequencing. Whereas Sanger Sequencing is capable of sequencing one stretch at a time, a parallel assembly of sequencing reactions has lead to high throughput sequencing often dubbed Next Generation Sequencing (**NGS**).

Jeremy Seto jseto@citytech.cuny.edu

The Next Generation of Sequencing: High-Throughput Technologies



Credit: Jorjor7 (CC-BY 3.0)

High Throughput Sequencing Applied to Genome Sequencing (TEDed CC BY-NC-ND 4.0)

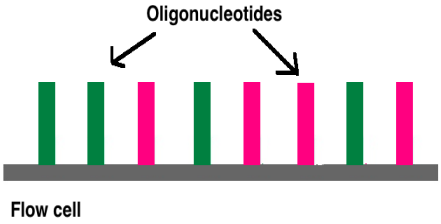
https://youtu.be/MvuYATh7Y74

Jeremy Seto jseto@citytech.cuny.edu

Short Read Sequencing by Synthesis

Illumina

Illumina short read sequencing uses flow cell technology where oligonucleotides complimentary to adapter primers are physically seeded.

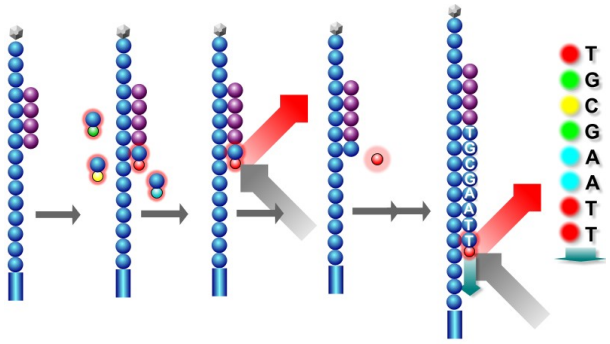


Flow cell surface with the adapter oligonucleotides.
Credit: DMLapato (CC-BY-SA 4.0)

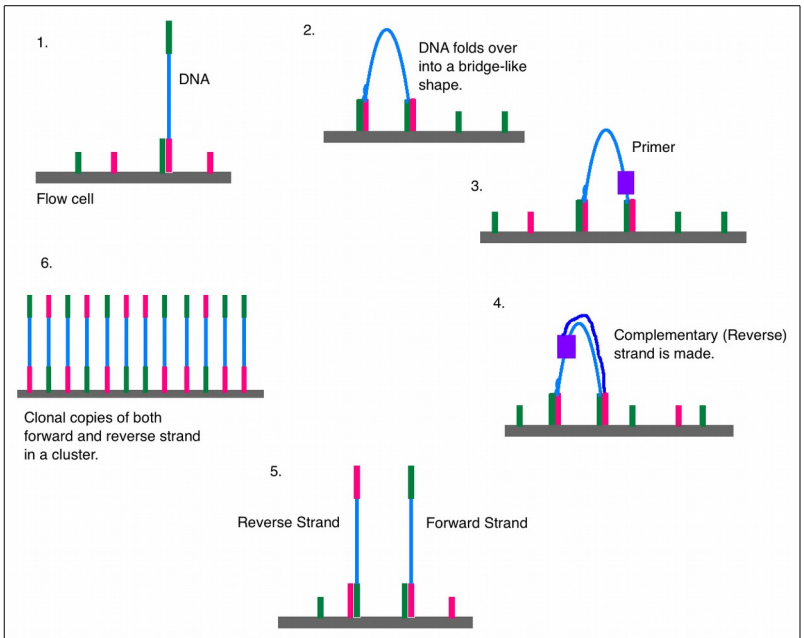
Fragmented DNA sequences are adapted with primers through ligation and hybridized to the flow cell. To increase the signal from sequencing, the short DNA sequences are amplified through a process called bridge amplification or cluster generation.

The flow cell undergoes successive rounds of flooding with a fluorescent nucleotide, permitted to incorporate with a DNA polymerase and washed away. After each flood/wash cycle, fluorescent signals are measured to indicate the incorporation. Specific locations of fluorescence are tracked and consolidated to indicate the sequence at each registered point.

https://youtu.be/womKfikWlxM



Each flow cycle introduces a fluorescent nucleotide for incorporation.
Credit: Abizar Lakdawalla (CC-BY 3.0)



Cluster generation through bridge amplification. A low number of PCR cycles is used. Cluster generation aids in subsequent signal/noise determination.
Credit: DMLapato (CC-BY-SA 4.0)

Jeremy Seto jseto@citytech.cuny.edu

Ion Torrent

Fragmented DNA is ligated to adapter sequences and adhered onto microbeads. The beads are embedded into microwells on a semiconductor. Ion Torrent performs the sequencing reactions in an unbuffered solution since the semiconductor acts as a pH meter to identify nucleotide incorporation. Standard nucleotides are flooded onto the chip and incorporated. Because nucleotide incorporation creates a proton (H^+), a microenvironment of low pH is detected in the unbuffered solution.

https://youtu.be/WYBzbxIfuKs

Single Molecule Real Time Sequencing

Pac Bio

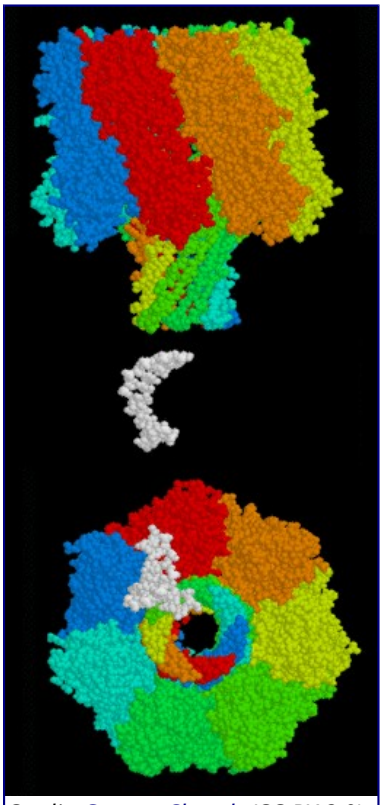
Pac Bio uses nanowells with covalent bonded DNA polymerase to sequence individual molecules of DNA. Fluorescent nucleotides are incorporated during synthesis reactions and a real-time incorporation can be measured. Pac Bio sequencing has the advantage of sequencing fragments of 10-20kb, in stark contrast to the short read methods.

https://youtu.be/NHCJ8PtYCFc

Oxford Nanopore

Oxford Nanopore utilizes the protein alpha-hemolysin integrated onto a semiconductor chip. The pore size of the protein is the correct ssize for a single DNA molecule to fit through. A DNA Polymerase molecule is linked to the opening of the pore where the replicated DNA is fed through. As the DNA traverses the pore, the voltage changes are measured and mapped to the qualities of specific bases.

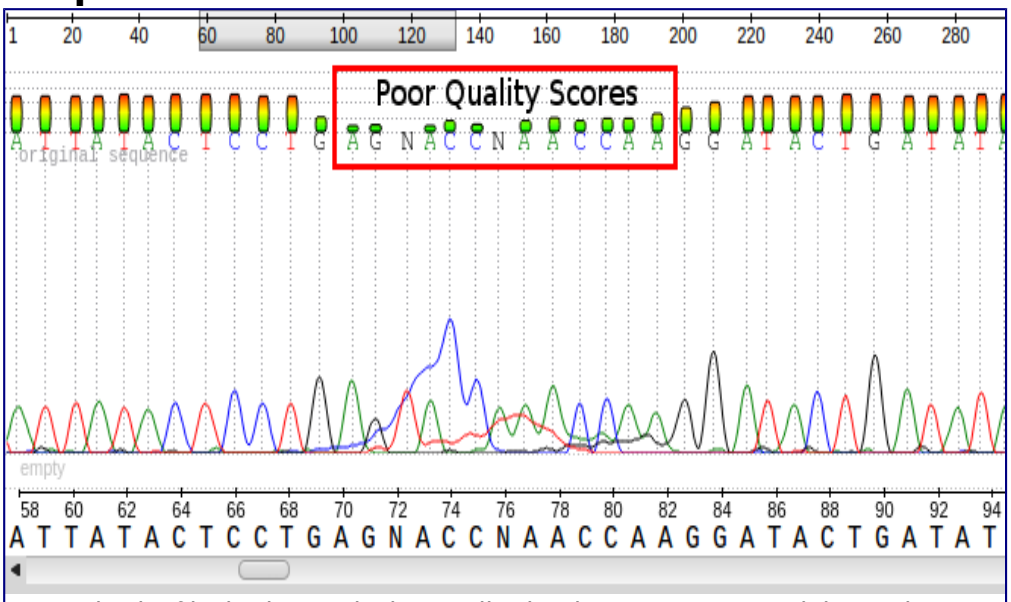
https://youtu.be/BNz880V52rQ



Credit: George Church (CC-BY 3.0)

Jeremy Seto jseto@citytech.cuny.edu

Sequence output



A sample ab1 file displaying the base calls, the chromatograms and the quality scores for each base. Notice the poor quality in the red box and the corresponding peaks/bases (CC-BY-NC-SA Jeremy Seto)

The output file of next generation sequencing methods utilize the **fastq** format. Like a **fasta** file, there is a header that describes the sequence. The first line is the header or title line which begins with '**@**' (remember that fasta begins with '**>**'). The second line is the actual raw sequence (once again similar to fasta). The third line has no meaning while the fourth line is filled with symbols as long as the sequence line. This last line is the quality score of the base call. As with the Sanger sequencing, there may be ambiguity with the base call of the sequence and the certainty is maintained in the quality score.

```
@GWZHISEQ02:207:H3VYHBCXX:1:1101:1465:1989 1:N:0:GTGGCC
AGAAGCCAAACGCATCATGGAAAAGCAGTTGCTCGAAGAACTCGAGCGACAGAGACAAGCTGAGCTTGCAGCACAAAAAGCCAGAGAGGAGGAAGAGAGAG
+
DBDDDIHIHHIIIHIHHHHIIIIGHIIIIIIIIIIHIHHHEHHIIHIIIIIIHIHIIIIIHHIIIHHHIIIHIHIHIHIIHHH?@HHHIIHIHHHIGG
@GWZHISEQ02:207:H3VYHBCXX:1:1101:1715:1978 1:N:0:GTGGCC
NTACAGCCACGTGGAACCTGAGCACCATGAAGTCCCCCAAACCAGTGGGTTTTAAAGCAATCACCACAAAACTCCCTTTTTAGCTTTTGCTTGGGGTGGGG
+
#<DDDIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIHIIIHIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@GWZHISEQ02:207:H3VYHBCXX:1:1101:1695:1987 1:N:0:GTGGCC
GCCAAGCAAGCCTGCCCCTCAGCCCCCAGAGTGCACCCAATGACCCCCCTCCCTTGGGGTGACTCCTGATGAAGCACAACTCCCCACAGGGCCCCTAACCC
+
@BDADHIIIHIHIIIIHHIHIIHIIICHDHIIHIGEHIIHGFEHHHIIIIIIIIHHHHCHDIHHHIIHHIHIICCHFEHI?FGHIGIGGHGHHHHHIHHHH
@GWZHISEQ02:207:H3VYHBCXX:1:1101:1915:1948 1:N:0:GTGGCC
NAGTGTCTTTATTATTTATATTTTTTAAAAGCCAAACAATTCCAGCATACTCGTGATAACTTTGAGGATTAGATACATTTTCTTAACATTTTTTTCCCTTT
+
#<<<DGHIIIGIIIIIII1GHIIIIIIHIIIHEHHIIHHIIIHGEHIHIIIIIGIIIIIIFHHEHIIIIGHIEEHIIIIIIIIIIHIHHHHIIIIIGHIHII
@GWZHISEQ02:207:H3VYHBCXX:1:1101:2116:1940 1:N:0:GTGGCC
NGAGGGATAATCAGGCCGGACCTGAAGGGAGCGGAGCTGCAGCTACCGGAGTTGGATCGGTCATGGGACTTTGCAAGTGCCCCAAGAGGAAGGTGACGAAC
+
#<DBDIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIHIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
```

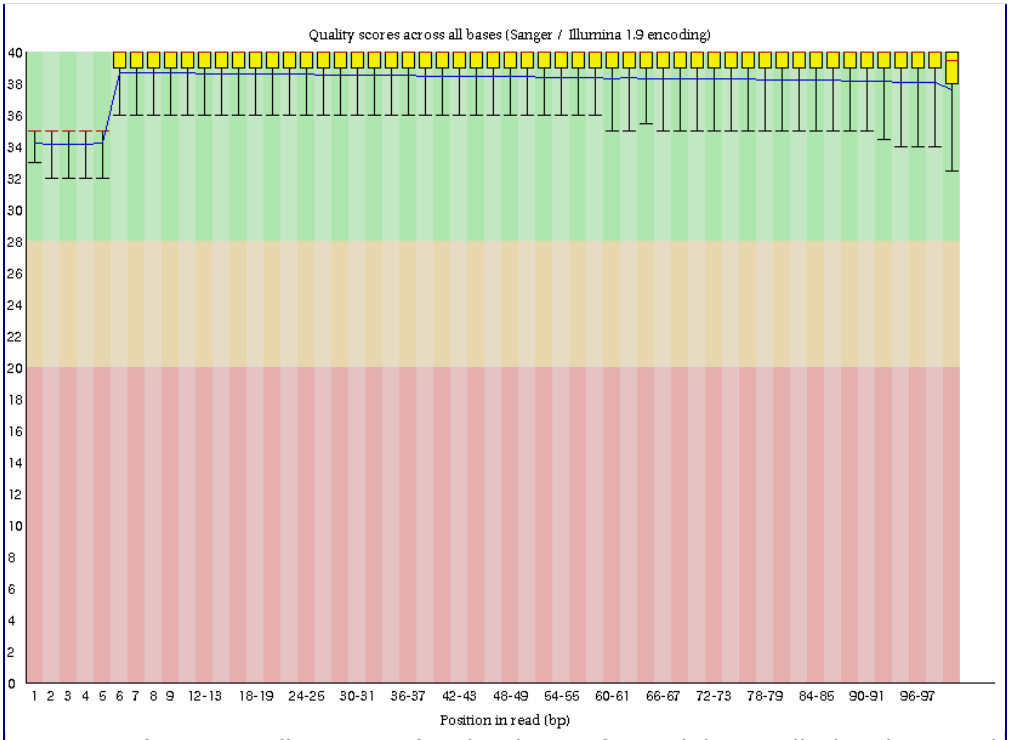
Sample fastq file displaying 5 short read sequences (CC-BY-NC-SA Jeremy Seto)

Phred scores were developed to assess the quality of the base calls arising from fluorescent Sanger sequencing during the Human Genome Project. The phred program scans the peaks of the chromatogram and scores based on certainty or accuracy of the call. The scores are logarithmically based and scores greater than 20 represent greater than 99% accuracy of the base call.

Jeremy Seto jseto@citytech.cuny.edu

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

Using the phred scores embedded in the last line of fastq files, poor quality reads can be removed. Using a program like FastQC permits the assessment of the reads and produces graphical representation of quality.

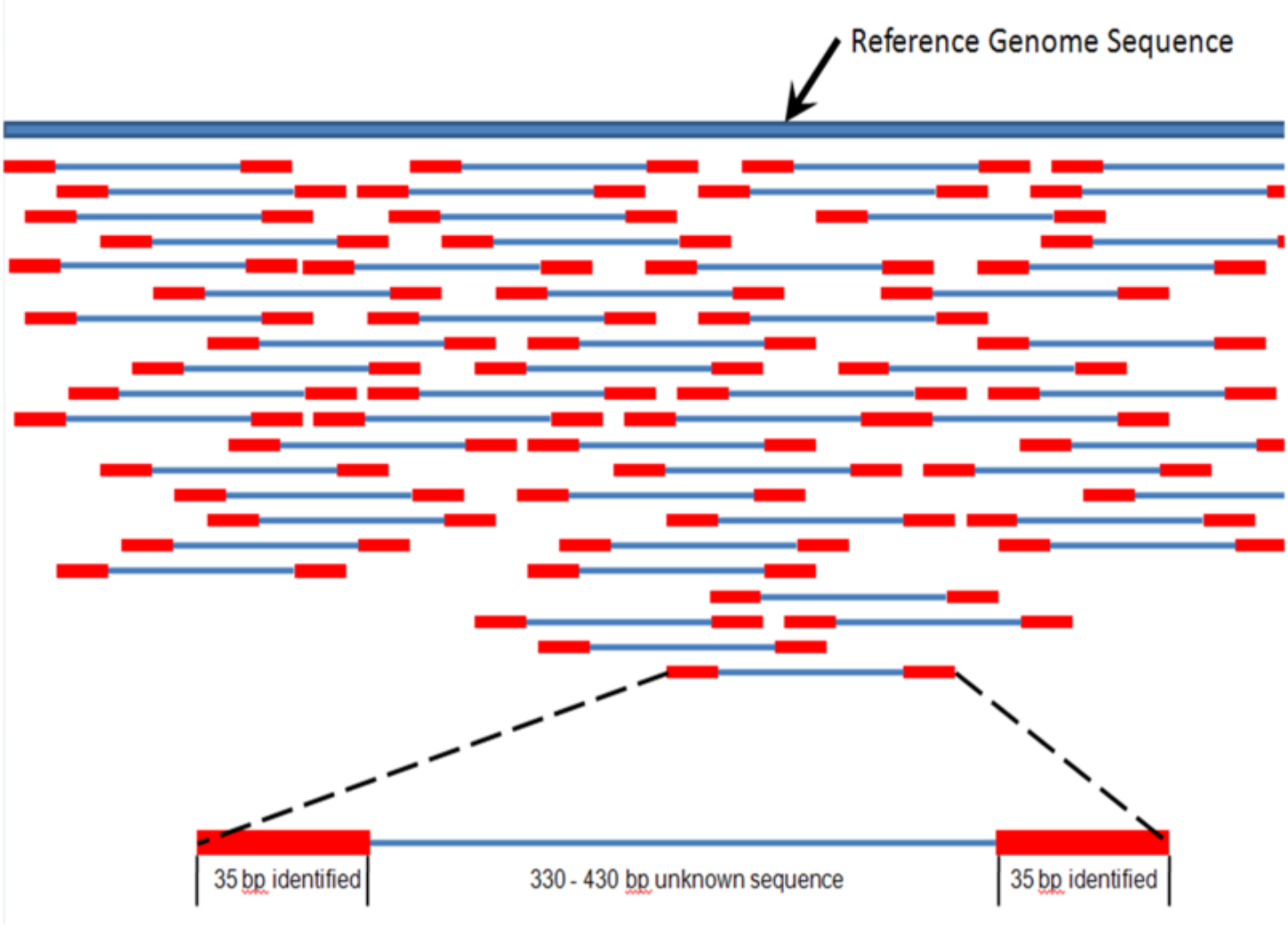


FastQC quality output illustrating the Phred score for each base call. This short read sequence of about 100 nucleotides has all bases made at greater than 30, or > 99.9% accuracy.

Jeremy Seto jseto@citytech.cuny.edu

Assembly and Alignment

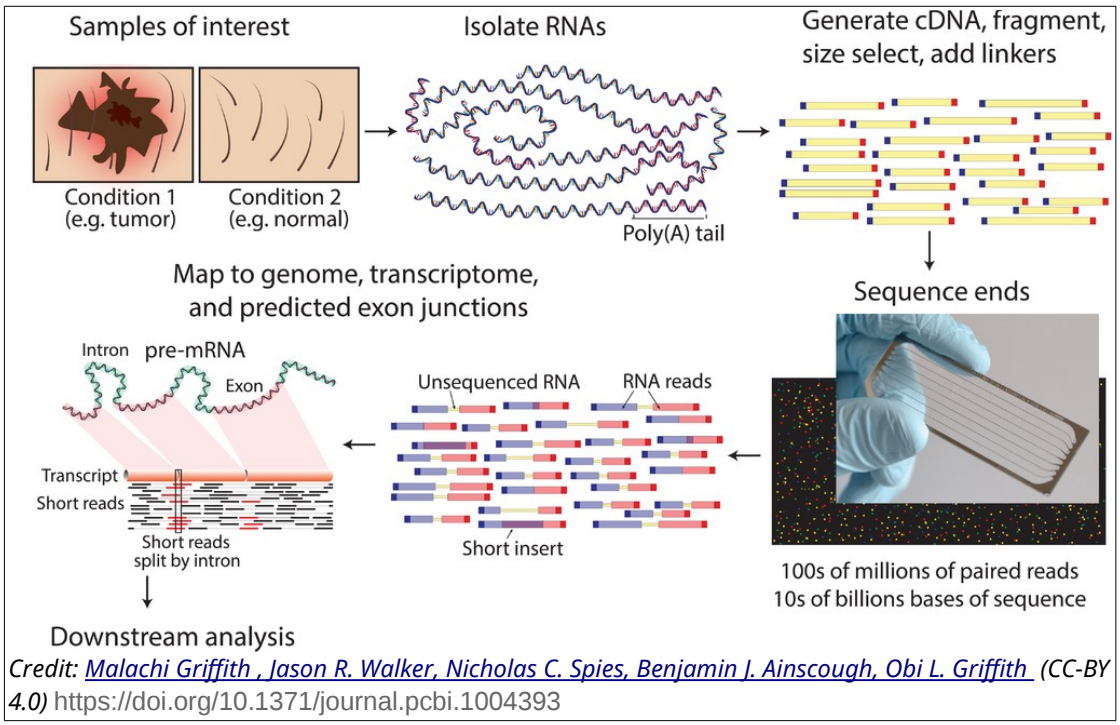
Sequences from short reads must be assembled into a usable sequence. To do so, a **reference genome** may aid in the assembly after adapter sequences are trimmed using automated methods. In the case that there is no reference genome, a related species may be used or a more computationally intensive process of ***de novo* assembly** must take place. With *de novo* assembly, it may be useful to have some long reads performed with PacBio to create scaffolds for generating the assembly into contiguous sequences, or **contigs**.



Jeremy Seto jseto@citytech.cuny.edu

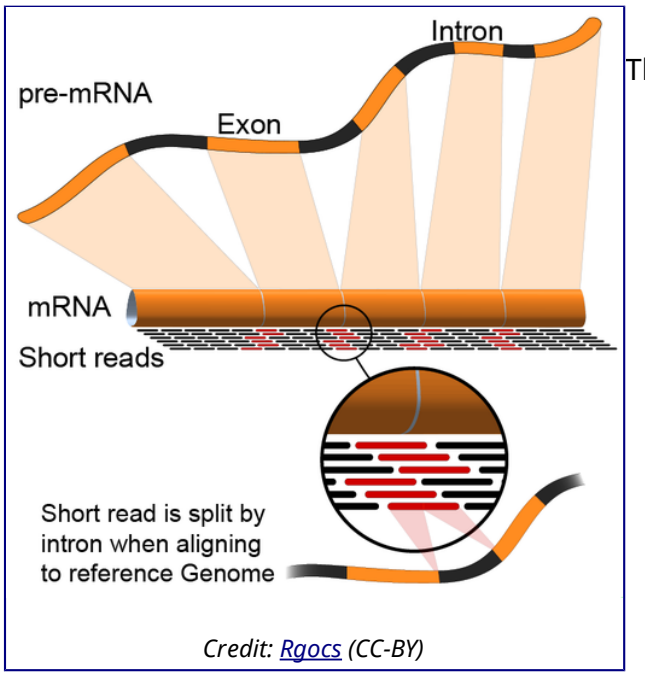
RNA-Seq

RT-PCR and **RT-qPCR** can be used to measure the abundance of specific transcripts in a fairly low throughput way. Leveraging the the concept of Reverse Transcription and coupling that to high-throughput sequencing technologies, transcripts can be sequenced and mapped to a genome to depict the quantity of transcripts as represented by number of reads.



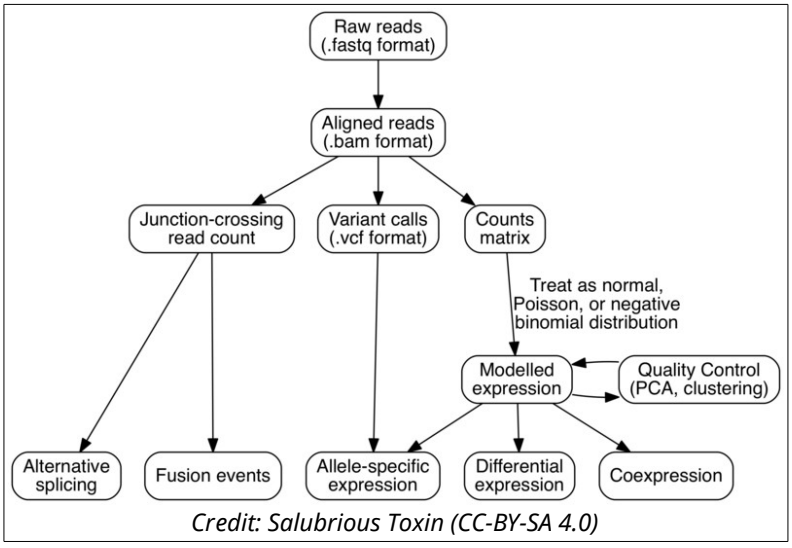
Credit: Malachi Griffith , Jason R. Walker, Nicholas C. Spies, Benjamin J. Ainscough, Obi L. Griffith (CC-BY 4.0) https://doi.org/10.1371/journal.pcbi.1004393

Given sufficient read coverage, novel splice isoforms can also be identified as different exon-exon junctions are identified.



Credit: Rgocs (CC-BY)

The general workflow of RNA-Seq analysis follows:



Credit: Salubrious Toxin (CC-BY-SA 4.0)

Jeremy Seto jseto@citytech.cuny.edu

Advanced Video of Variant Calling from NGS to Decipher a Genetic Susceptibility

https://youtu.be/T6hC-C4TasQ

Jeremy Seto jseto@citytech.cuny.edu

Genetic Manipulation (selection)

Genetic modification of organisms has been occurring through human manipulation since the beginning of agriculture. Humans selectively bred crops and livestock to propagate desirable traits in a process termed **artificial selection**. The original grass that gave rise to domesticated corn called teosinte hardly resembles what we think of when imagining modern maize.

Teosinte, the progenitor of maize. Corn came about due to selective breeding.
Credit: John Doebley (CC-BY)

Jeremy Seto jseto@citytech.cuny.edu

Variation: Crop domestication

Selective breeding can yield a variety of features even within the same species. Below is selection of vegetables of the species *Brassica oleracea* that have been developed into different varieties over the course of agricultural history.

Cabbage: Brassica oleracea var. capitata
Credit: Forest & Kim Starr ***(CC-BY 3.0)***

Broccoli: Brassica oleracea var. italica
Credit: Coyau *(CC-BY-SA 3.0)*

Kohlrabi: Brassica oleracea var. gongylodes
Credit: *Coyau* (CC-BY-SA 3.0)

Romanesco: Brassica oleracea var. botrytis
Credit: Richard Bartz (CC-BY-SA 2.5)

Jeremy Seto jseto@citytech.cuny.edu

Variation: Animal domestication

Companion animals like dogs underwent thousands of years of domestication and selection for traits that were desirable for different circumstances. A high degree of morphological diversity exists between dog breeds and their ancestral grey wolf progenitor.

Credit: Mary Bloom, American Kennel Club (CC-BY-SA 4.0)

Jeremy Seto jseto@citytech.cuny.edu

Genetic Manipulation (engineered)

Artificial selection takes multiple generations over a long period of time. With the advent of recombinant DNA and biotechnology, scientists can now genetically modify organisms through introduction of foreign genes to provide desirable characteristics within one generation. This process does not require traits to naturally arise in a species.

GloFish® are novelty pets that have the insertion of various cnidarian fluorescent protein genes into the genome. These fish were released in the United States in 2003 and have subsequently been developed in red, orange, and blue varieties. Black tetras and tiger barbs are also now available.

Black tetra (Gymnocorymbus ternetzi)

Wild-type Black Tetra

Credit: Fernandograu (CC-BY-SA 3.0)

GloFish are transgenic zebra fish (Danio rerio) expressing variants of GFP. Bottom features a wild-type fish.

Credit: Azul (CC-BY)

Jeremy Seto jseto@citytech.cuny.edu

Genetic Engineering in Plants

With the advent of agribusiness, agriculture has become a profit driven venture independent of food production. In this case, high production is paramount. Whereas traditional agriculture and artificial selection was slow and methodical, genetic modification in the context of agribusiness is instantaneous through **genetic engineering**. The ***objective*** of genetic engineering is to transfer the DNA encoding a useful or favorable gene from an organism that carries that gene to one that does not. Simply inserting DNA into an organism does not result in expression. An appropriate promoter for the transgenic organism must be upstream of the gene of interest in order to drive transcription. In mammals, a strong promoter that will result in expression in every cell is the CMV promoter that is derived from cytomegalovirus. Likewise in plants, a strong promoter that works in every cell is derived from viral promoters like the **CaMV promoter** from cauliflower mosaic virus from the **35S gene** (a ribosomal RNA). (CaMV 35S sequence on NCBI)

Examples of useful traits include:

- degrade herbicides
- kill agricultural pests
- synthesize critical nutrients
- to improve color and taste
- resist damage during transit or prolonged storage.
- Increase size
- reduce time to market (more rapid growth or maturation)

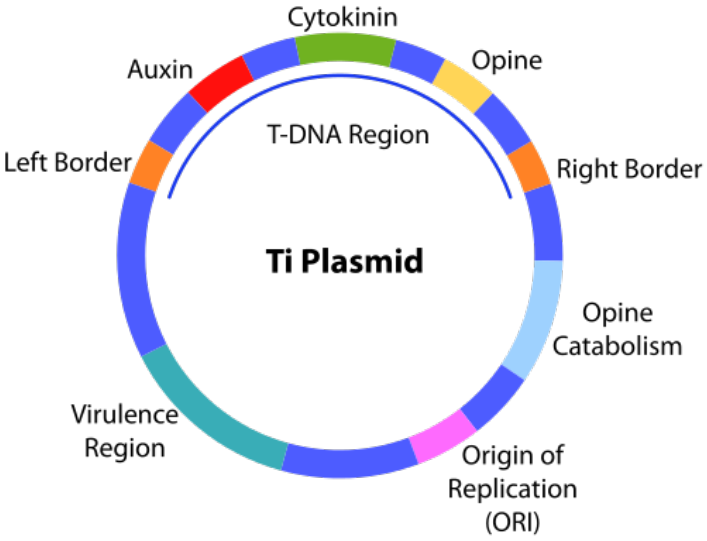
Genetically Modified foods have become a hot topic of contention in recent times. These crops are generated through the infection of plant cells by a bacterium called ***Agrobacterium tumefaciens***. *Agrobacterium* are gram-negative alphaproteobacteria of the family Rhizobiaceae, which include symbiotic nitrogen fixers found in legumes. Unlike those symbionts, *Agrobacterium* is a pathogenic soil bacterium known as a causative agent of crown galls (tumors).

Crown gall on a Kalanchoë infected with Agrobacterium tumefaciens
Credit: Bhai (CC-BY-SA 3.0)

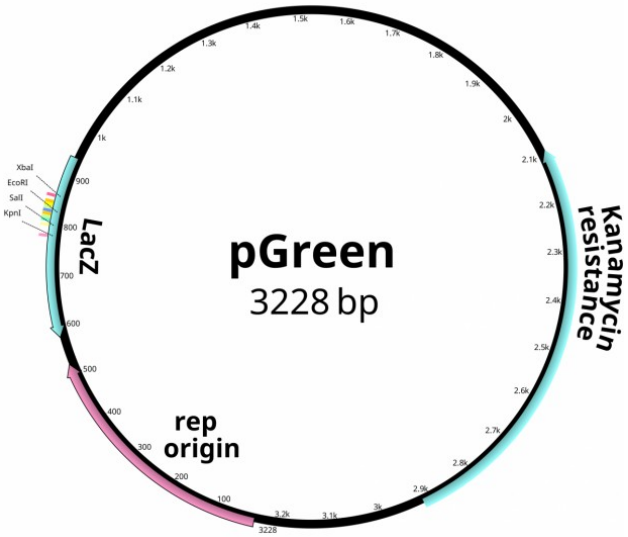
Jeremy Seto jseto@citytech.cuny.edu

The tumors are caused by the infection of plant cells by the bacterium and the subsequent insertion of the **T-DNA** ("Transfer DNA") that has a tumor inducing capability (**Ti**). Through the engineering of the T-DNA in a plasmid, selected genes can be delivered to plant cells through infection of transformed bacteria.

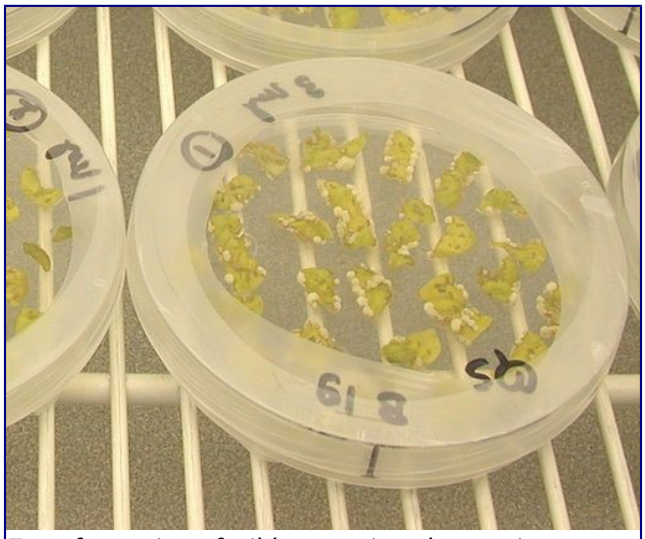
A modified Ti plasmid called **pGreen** was engineered to provide a MCS and selection marker for insertion of foreign genes of interest.In order for these genes to be expressed, the are driven by strong plant promoters like those from the CaMV 35S gene.



Ti plasmid has the T-DNA region replaced by the transgene. Ti is transformed into Agrobacterium which transduce the DNA into plant cells.



The plant to be engineered is cultured and infected with the transformed *Agrobacteria* that will then induce cysts that eventually root. The strong promoter of the CaMV 35S will constitutively express the gene in all cells of the plant.

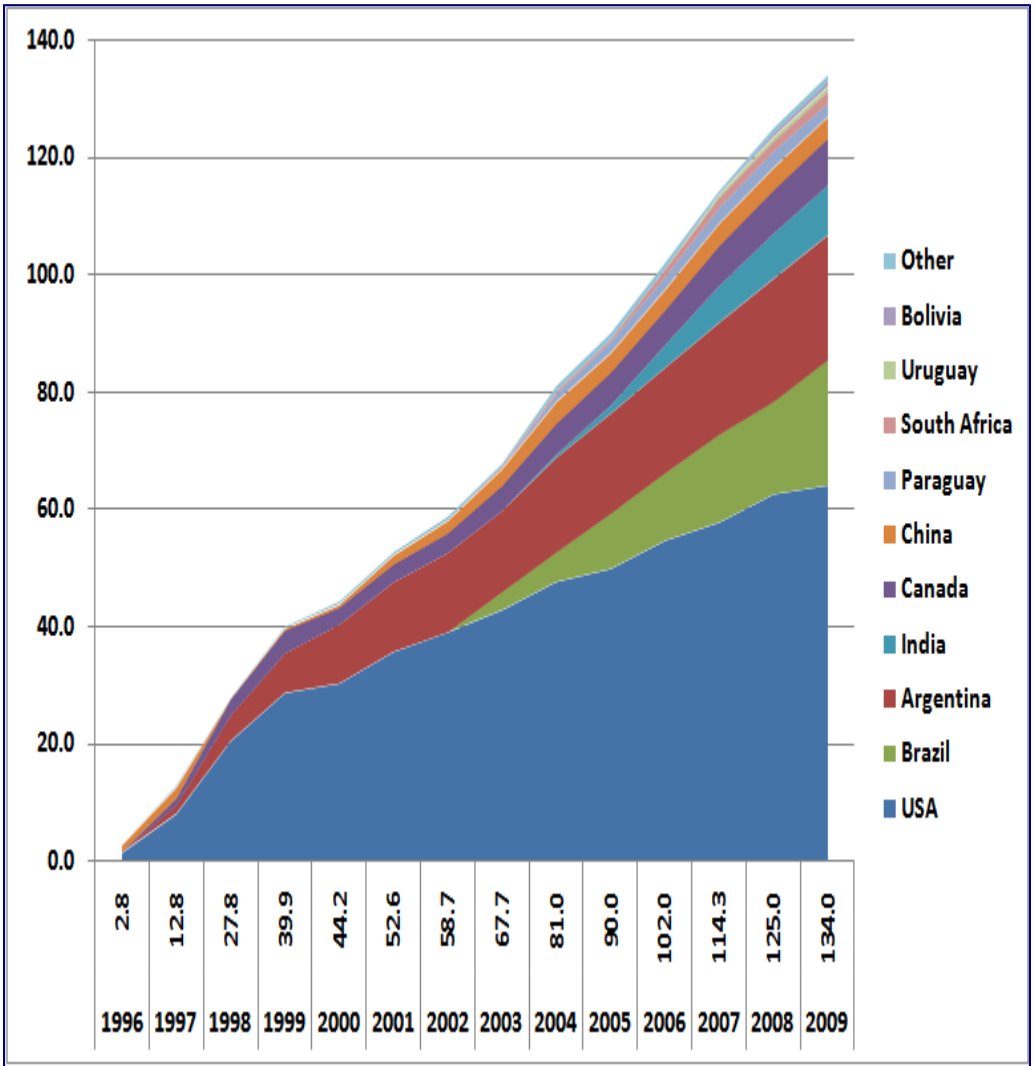


Transformation of wild potato in culture using agrobacterium

Credit: Seb951 (CC-BY-SA 3.0)

Jeremy Seto jseto@citytech.cuny.edu

Growth of GMO Crops



Land area use in millions of hectares

Credit: Fafner (CC-BY-SA 3.0)

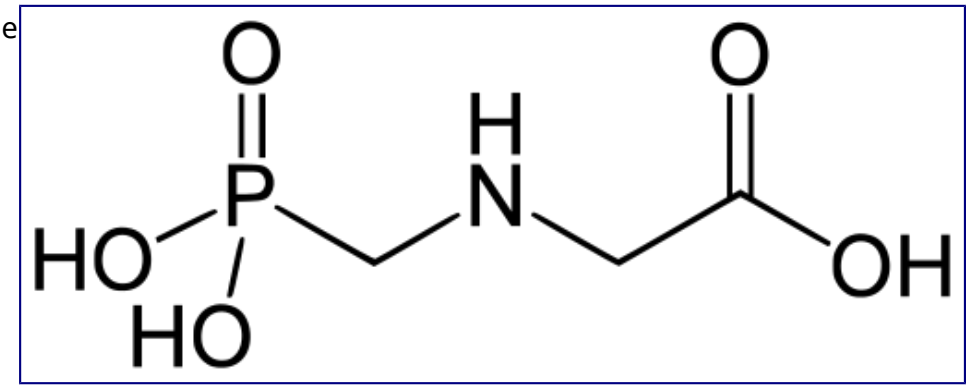
Damage Resistance

The first genetically modified crop FDA approved for sale was known as the Flavr Savr tomato. Tomatoes are prone to damage during shipping and are therefore picked before ripening. However, vine ripened tomatoes have richer flavor. Calgene, developed the Flavr Savr and sold it to market between 1994-1997 in the U.S. Flavr Savr was modified by inserting an antisense into the genome that knocked-down the expression of polygalacturonase. Polygalacturonase degrades pectin in the cell walls of the fruit that results in softening, proclivity to damage and eventual rotting.

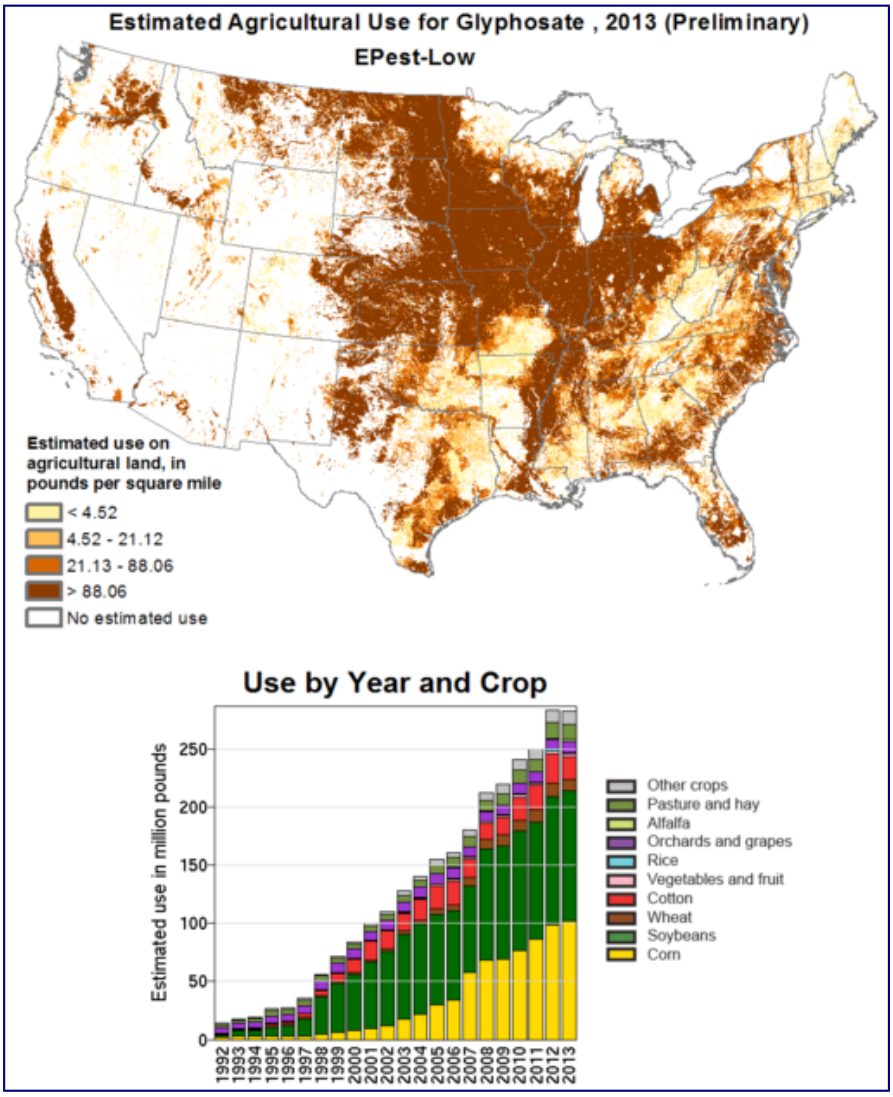
Credit: WikiPedant *(CC-BY-SA 3.0)*

Jeremy Seto jseto@citytech.cuny.edu

Herbicide Resistance



Roundup is the trade-name of the herbicide **glyphosate** used in agriculture to control weed populations developed and patented by Monsanto. Glyphosate is absorbed through the foliage of plants and interferes with the enzymes that aid in production of tyrosine, phenylalanine and tryptophan. Plants and lower organisms generate aromatic amino acids through an enzyme 5-enolpyruvylshikimate-3-phosphate (EPSP) synthase which is the target of this compound. A series of Roundup Ready crops were designed by the insertion of the *Agrobacterium* EPSP gene driven by the CaMV 35S promoter. This version of the gene is inherently resistant to glyphosate poisoning.



Jeremy Seto jseto@citytech.cuny.edu

Emergence of superweeds

Palmer amaranth (Amaranthus palmeri), commonly referred to as pig weed is a pest species in cotton and soy fields that has become glyphosate resistant.

Credit: Pompilid *(CC-BY-SA 3.0)*

Pest Resistance

Crystals from the *Bacillus thuringiensis* (**Bt**), called ***Cry*** protein, are toxic to various insects: moths & butterflies, beetles, ants, wasps, flies & mosquitoes, bees, nematodes. The insertion of this gene into plants like corn (Bt-Corn) were designed to be resistant to pests.

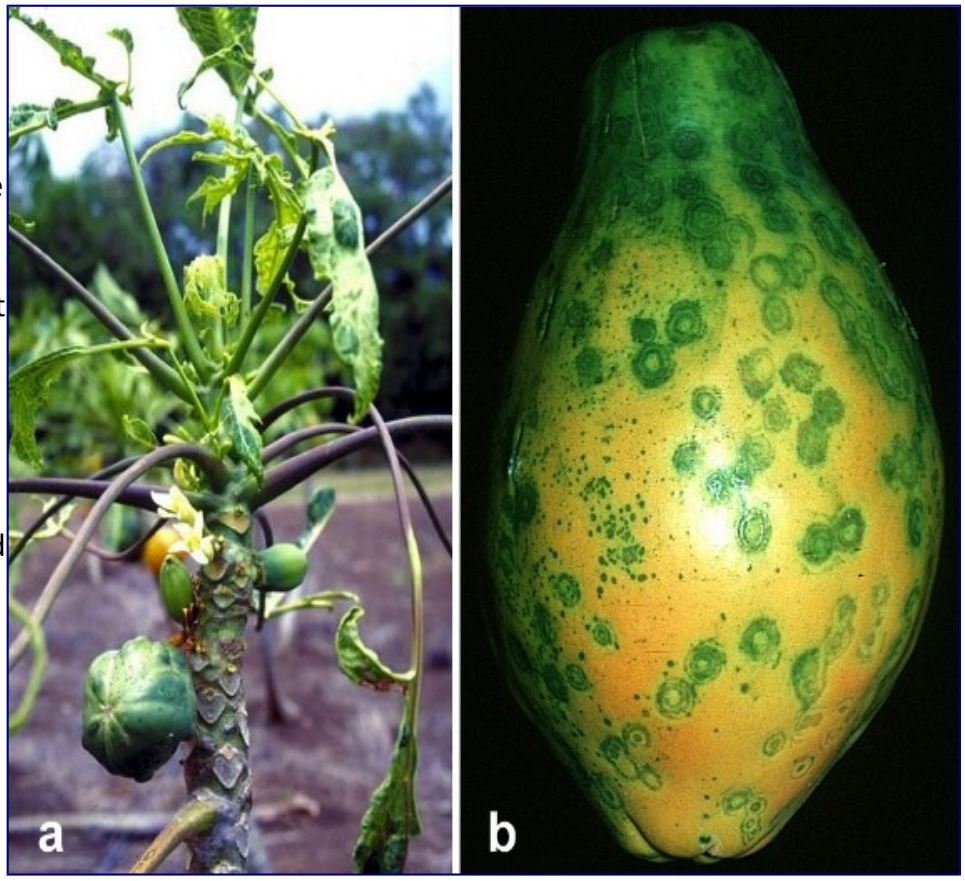
Credit: Cyndy Sims Parr *(CC-BY-SA)*

Jeremy Seto jseto@citytech.cuny.edu

Disease resistance

Papayas in the United States primarily come from Hawaii. A virus known as Papaya Ringspot Disease threatened the papaya crop in Hawaii. To combat this, papaya were genetically engineered to block viral entry into the papaya cells. Papayas purchased in Latin markets are most likely unmodified and usually come from Mexico where the ringspot disease is not yet a problem.

Plum Pox Virus is a threat to the genus *Prunus*. A genetically modified plum plant has been developed called C5. The cells in these plants silence the expression of plum pox coat protein if infected to block propagation of the virus.



C5 resistant plums[/caption]

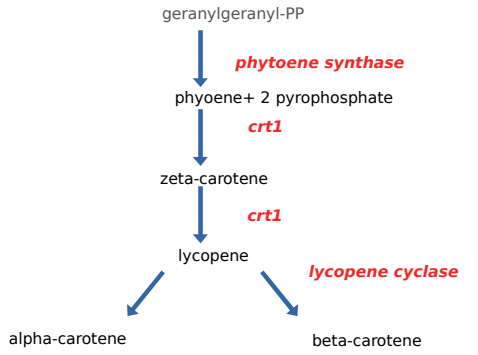
An apricot infected with plum pox.

C5 resistant plums

Jeremy Seto jseto@citytech.cuny.edu

Nutritional Engineering

Golden rice is a genetically engineered rice that is meant to address Vitamin A deficiency. It introduces enzyme genes from other species involved in the biosynthetic pathways for β-carotene production, a vitamin A precursor. It is estimated that millions of deaths and irreversible blindness occur in the third world each year due to Vitamin A deficiency and creating this rice was meant to address the problem.



Credit: en:user:Petaholmes, WilliamTheaker (cc-BY-SA 3.0)

Rice is a staple in many cultures, therefore it is a good delivery system. Many controversies exist surrounding Golden Rice due to anti-GMO sentiment (from patenting systems), cultural sensitivities (white rice revered in certain cultures), and Vitamin A content/conversion doubts.

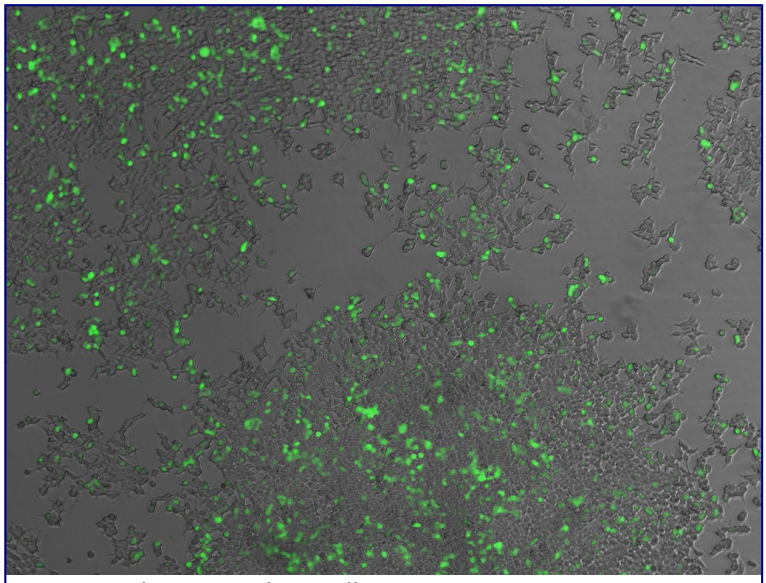
Credit: International Rice Research Institute (CC-BY-2.0)

Explore the debate

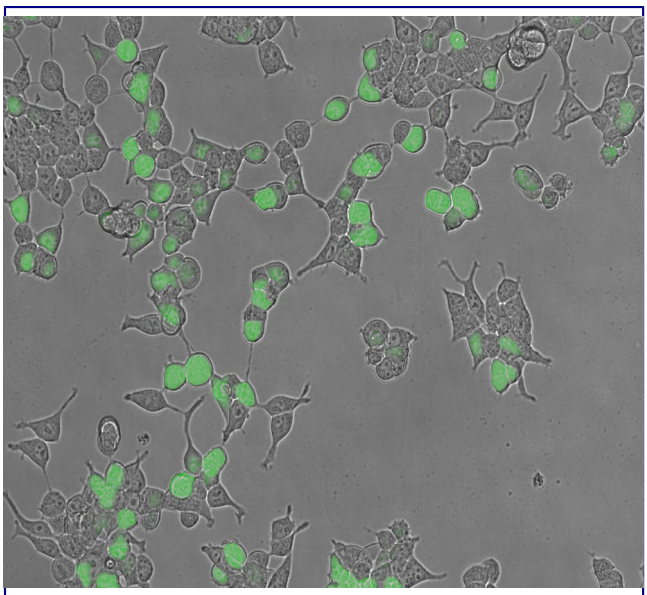
- Case studies reveal the complex truths behind GM crop myths.
- Case studies: A hard look at GM crops. Gilbert N. Nature. 2013 May 2;497(7447):24-6. doi: 10.1038/497024a

Jeremy Seto jseto@citytech.cuny.edu

Heterologous expression in Tissue Culture

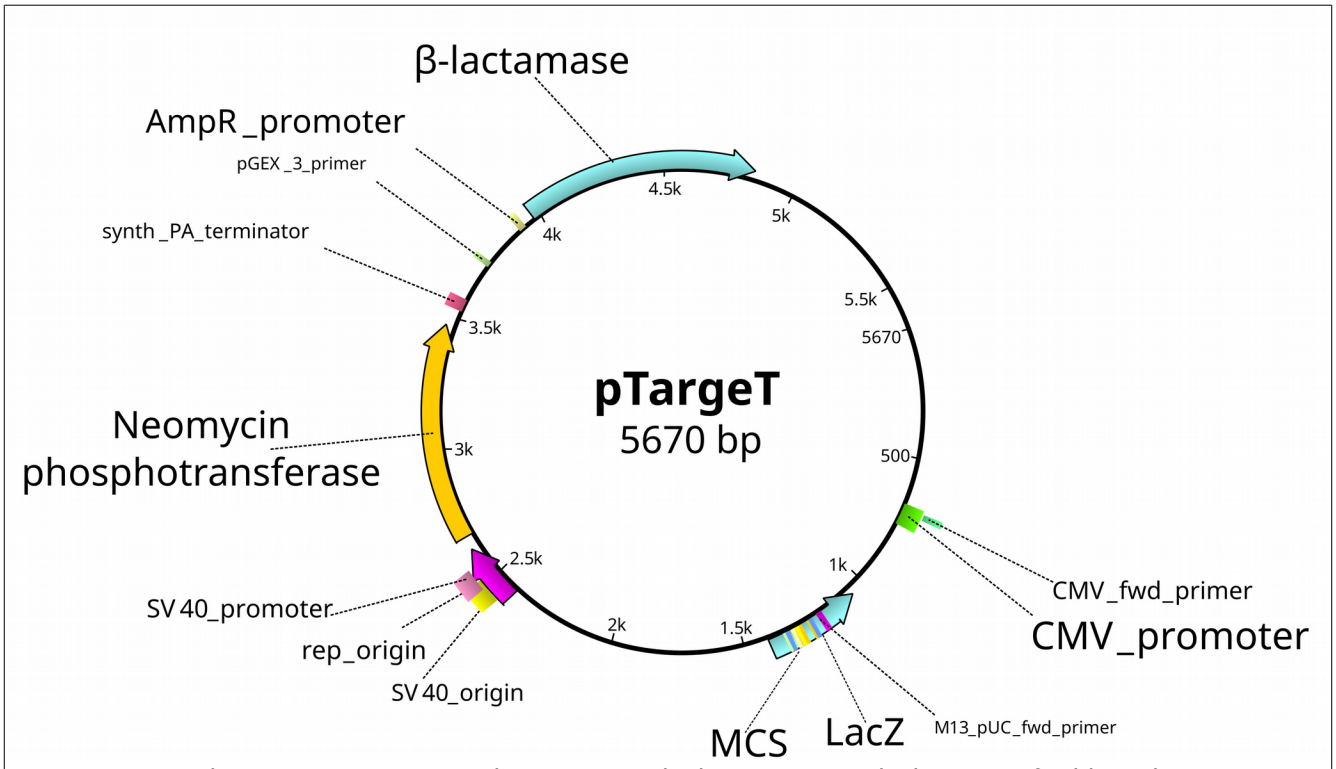


Human Embryonic Kidney cells (HEK293T) expressing green fluorescent protein.



Higher magnification of HEK293T expressing green fluorescent protein.

Plasmid Structure



pTarget mammalian expression vector. The MCS is inside the LacZ gene which permits for blue/white screening of the bacteria after cloning. The LacZ runs in the opposite orientation as the CMV promoter which will drive the gene transcription when inside a mammalian cell. SV40 origin servers as a plasmid replication origin if the cell line also expresses SV40 large T antigen, such as HEK293T cells.

Jeremy Seto jseto@citytech.cuny.edu

Mammalian expression vectors contain the same hallmark features as bacterial plasmids: bacterial replication of origin and bacterial antibiotic resistance gene (β-lactamase or **Amp_R**). General bacterial plasmid features allow for the carrying and propogation of the plasmid in a bacterial cell. Mammalian expression vectors additionally include a strong mammalian promoter (like **CMV** from the cytomegalovirus immediate early promoter) upstream of a multiple cloning site (**MCS**). Plasmids transfected into cells are transient in nature unless the DNA is selected for. The inclusion of a mammalian antibiotic resistance gene, like neomycin phosphotransferase (**Neo_R**), allows for the integration of the plasmid into the genome of the cell by using high concentrations of Neomycin or the analog G418.

Lipofection

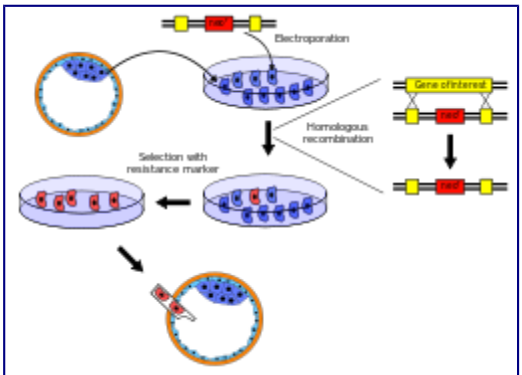
Cationic lipids can encapsulate plasmid DNA in liposomes. The cationic portions interact with the negatively charged plasma membrane to deliver the DNA into cells.

Calcium Phosphate Transfection

Calcium chloride solution can be used to incubate with plasmid DNA. When this solution is mixed with a HEPES-buffered saline solution (HeBS) containing phosphate ions, the solution precipitates onto the surface of mammalian cells where they are taken up with the DNA.

Knock-Out & Transgenesis

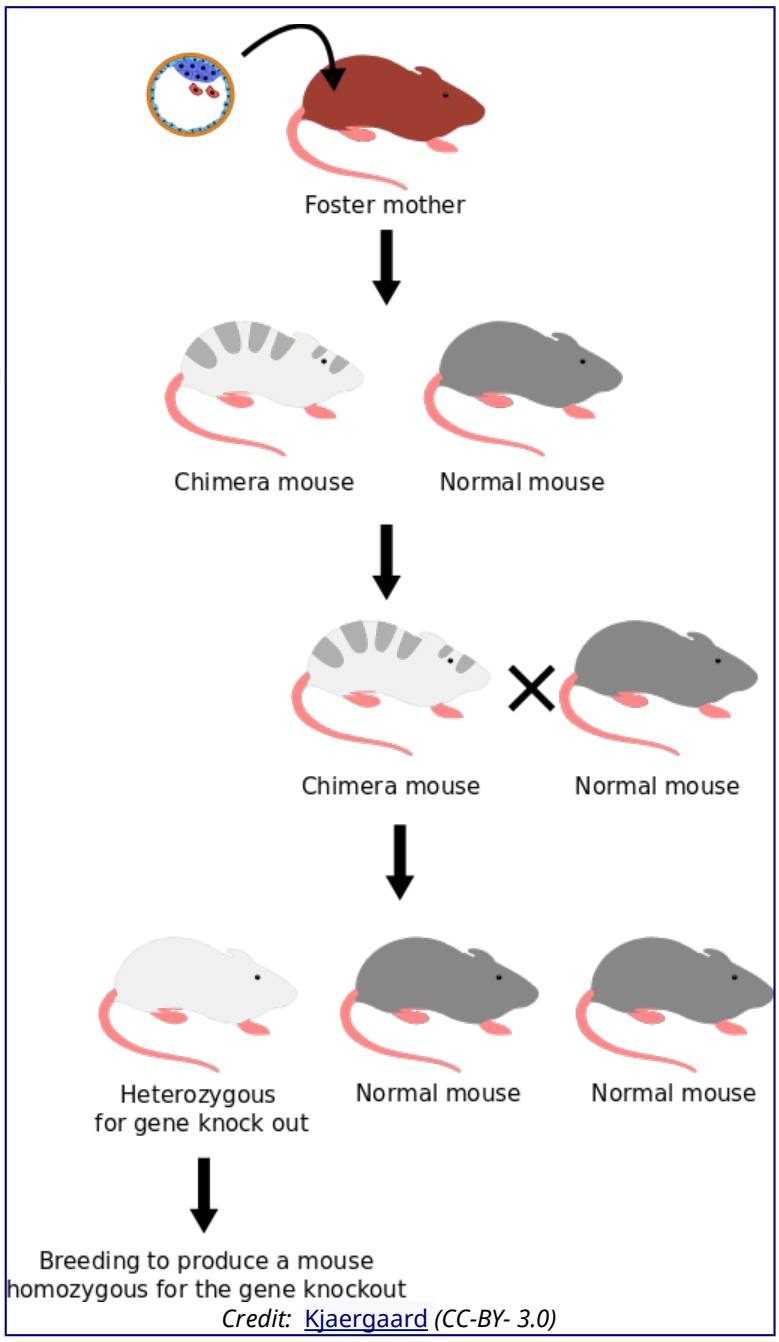
In the laboratory, model organisms are modified in order to understand the basic mechanisms of genes. The transformation of recombinant DNA into bacteria is an example of a genetic modification. Other model organisms, like mice, are used to study genes. Through recombinant DNA scientists can selectively ablate a gene, or create a **knock-out** (KO). Embryonic Stem (**ES**) cells are pluripotent cells with the capacity to differentiate into other cell types. Cultured ES cells can be transfected with plasmid DNA in order to genetically alter them. Linearized vectors containing a disrupted gene can homologously recombine with the native gene to replace it. Selection of cells with the disrupted gene by an antibiotic (like G418) enables the isolation and propagation of engineered ES cells.



Credit: Kjaergaard *(CC-BY- 3.0)*

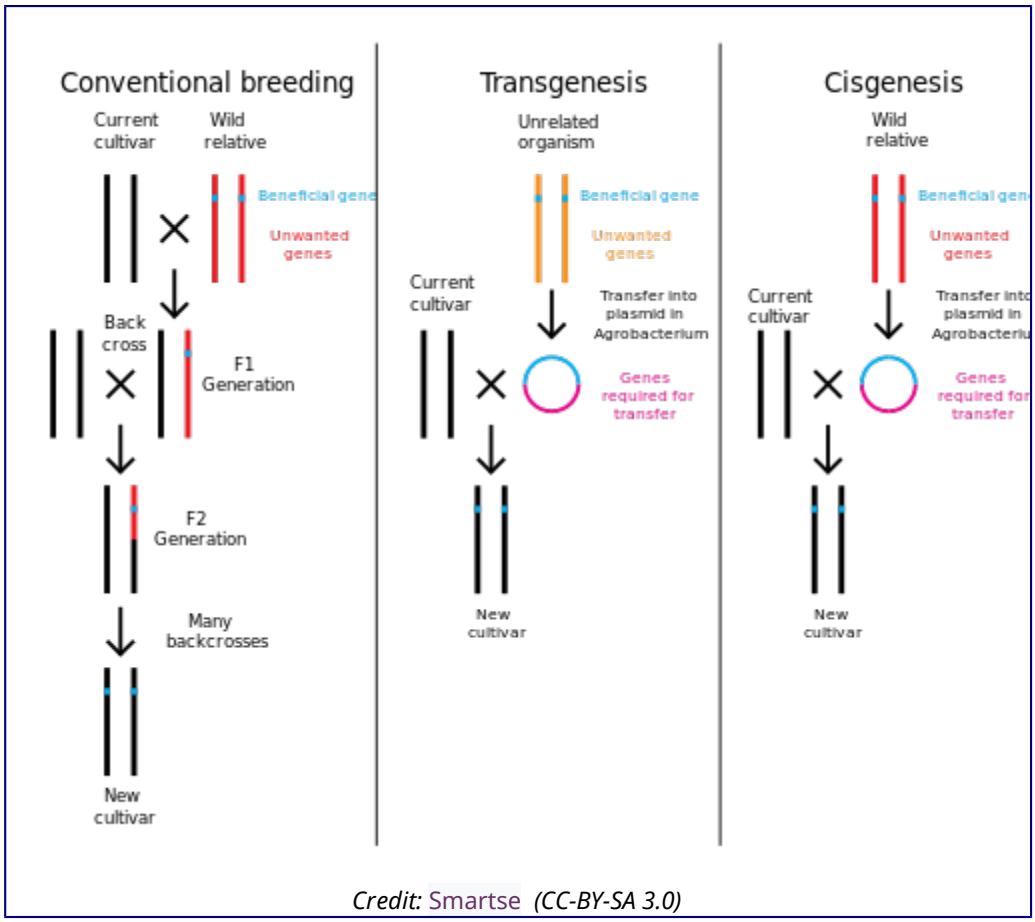
Jeremy Seto jseto@citytech.cuny.edu

ES cells can be injected into mouse blastocysts and partially contribute to the subsequent mouse upon implantation into a mouse. These first mice are referred to as **chimeras** because they arise from mixtures of cells from 2 genetic sources. Germ-line transmission of the modified cells is desired and breeding of the chimera reveal heterozygous offspring of the engineered background. Full knock-out mice can be generated in the subsequent generation of breeding.



Credit: Kjaergaard *(CC-BY- 3.0)*

Jeremy Seto jseto@citytech.cuny.edu

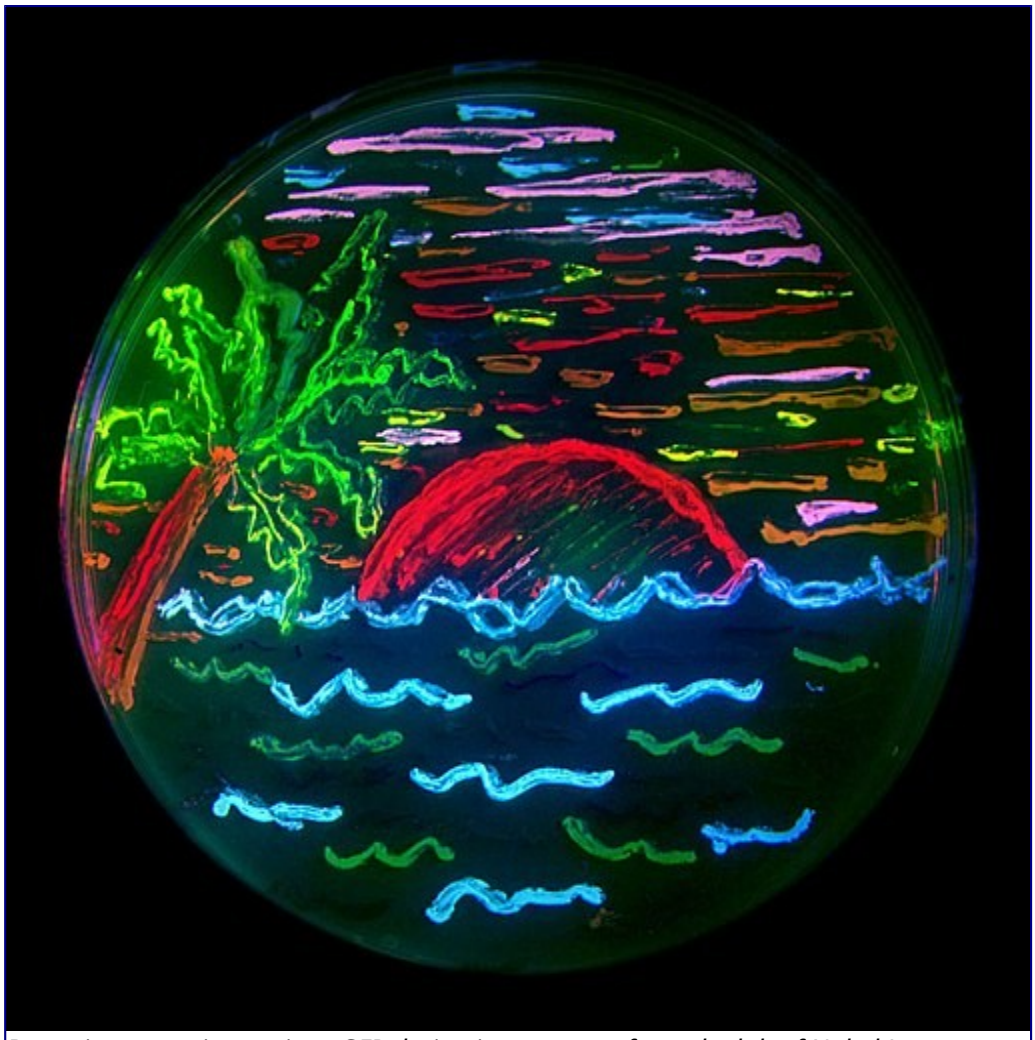


Credit: Smartse *(CC-BY-SA 3.0)*

Scientists can also overexpress or heterologously express foreign genes in what are termed **transgenic** organisms. As the name sounds, transgene refers to a gene from one place brought across into another.

Transgenic and KO models permit scientists to study the roles of genes inside the organism and understand basic functioning.
Through mutagenesis, derivatives of the green fluorescent protein (GFP) have been produced to provide a palette of colors. Additionally, the subsequent discovery of similar genes from other cnidarian species have aided biotechnology by providing tracer molecules within developing organisms or within cells.

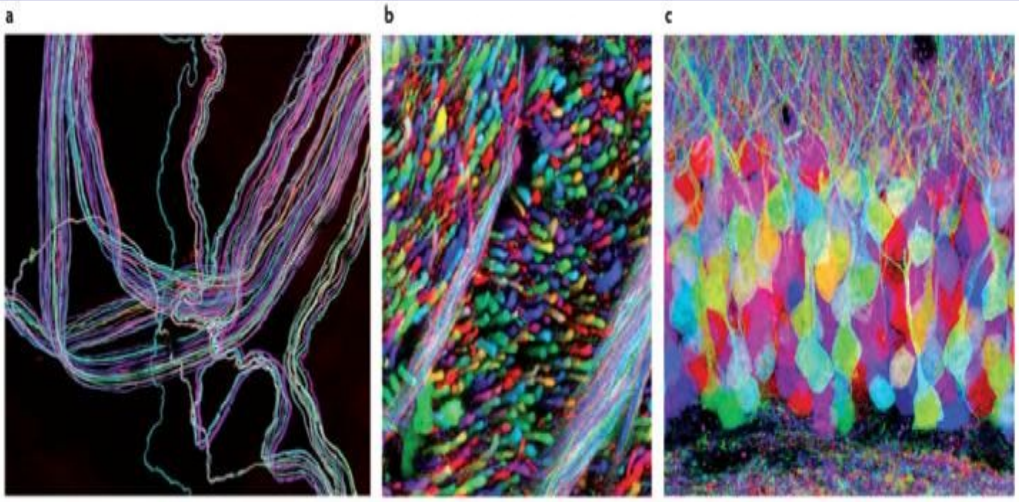
Jeremy Seto jseto@citytech.cuny.edu



Bacteria expressing various GFP derivatives on agar from the lab of Nobel Laureate Roger Tsien.

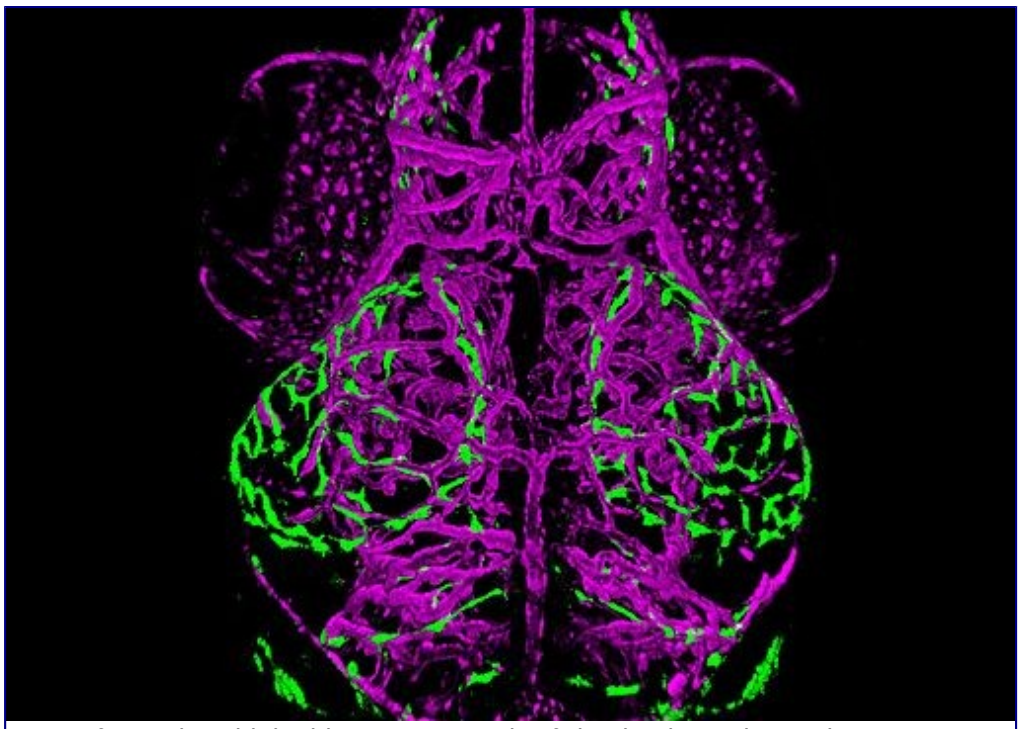
While commercial organisms like GloFish are a novelty, directed insertion of GFP and the variants into the genome under different promoter systems allow scientists to understand the cell-specific functioning or contribution to the organism. An example of this can be found in developmental neurobiology where individual axons can be traced.

Jeremy Seto jseto@citytech.cuny.edu



A "brainbow" is a system where a cassette of GFP variant genes are placed downstream of neuronal promoter to permit the tracing of individual neurons and their axons in mice.

Credit: Jeff W. Lichtman and Joshua R. Sanes (CC-BY 3.0)



Brain of a 10-day old double transgenic zebrafish. Blood vessels are shown in magenta (Kdrl:mcherry) and a novel population of perivascular endothelial cells are shown in green (MRC1a:eGFP).

Further Reading

- CRISP-Cas9 https://www.neb.com/tools-and-resources/feature-articles/crispr-cas9-and-targeted-genome-editing-a-new-era-in-molecular-biology

Jeremy Seto jseto@citytech.cuny.edu

PCR detection of GM food

Briefly, genomic DNA will be isolated from food items derived from vegetation. Genetic modification will then be identified by PCR of the plant promoter used in genetic engineering, **CaMV 35S**. As a **positive control** for the appropriate extraction of DNA, PCR for **plant specific tubulin** will be used.

1. Add 100 µL of lysis buffer to each tube containing the plant or food material.
2. Twist a clean plastic pestle against the inner surface of the 1.5-mL tube to forcefully grind the plant tissue or food product for 1 minute.
3. Add 900 µL of lysis buffer to each tube containing
4. Boil the samples for 5 minutes in a water bath
5. spin for 2 minutes to pellet cell and food debris.
6. Transfer 350 µL of each supernatant to a fresh tube
7. Add 400 µL of isopropanol to each tube
8. Mix and leave at room temperature for 3 minutes.
9. Spin for 5 minutes.
10. Carefully pour off and discard the supernatant from each tube. Air dry pellet.
11. Add 100 µL of TE buffer to each tube. 5 min at room temperature then keep on ice.

PCR for 35S Promoter:

1. Label one tube “35S FP” for food product.
2. Label one tube “35S WT” for wild-type soy: Negative control.
3. Label “35S RR” for Roundup Ready® soy plant: Positive control.
4. Different groups will only do one control.
5. Add **22.5 µL** of the 35S primer/loading dye mix to each tube containing PCR bead.
    - 5'-CCGACAGTGGTCCCAAAGATGGAC-3' (Forward Primer)
    - 5'-ATATAGAGGAAGGGTCTTGCGAAGG-3' (Reverse Primer)
6. Add **2.5 µL** of food product DNA to the reaction tube marked "35S FP.”
7. Add **2.5 µL** of wild-type or Roundup Ready® soybean DNA to the appropriate reaction tube marked "35S WT" or "35S RR.”

PCR for Tubulin: (positive control for DNA quality and PCR conditions)

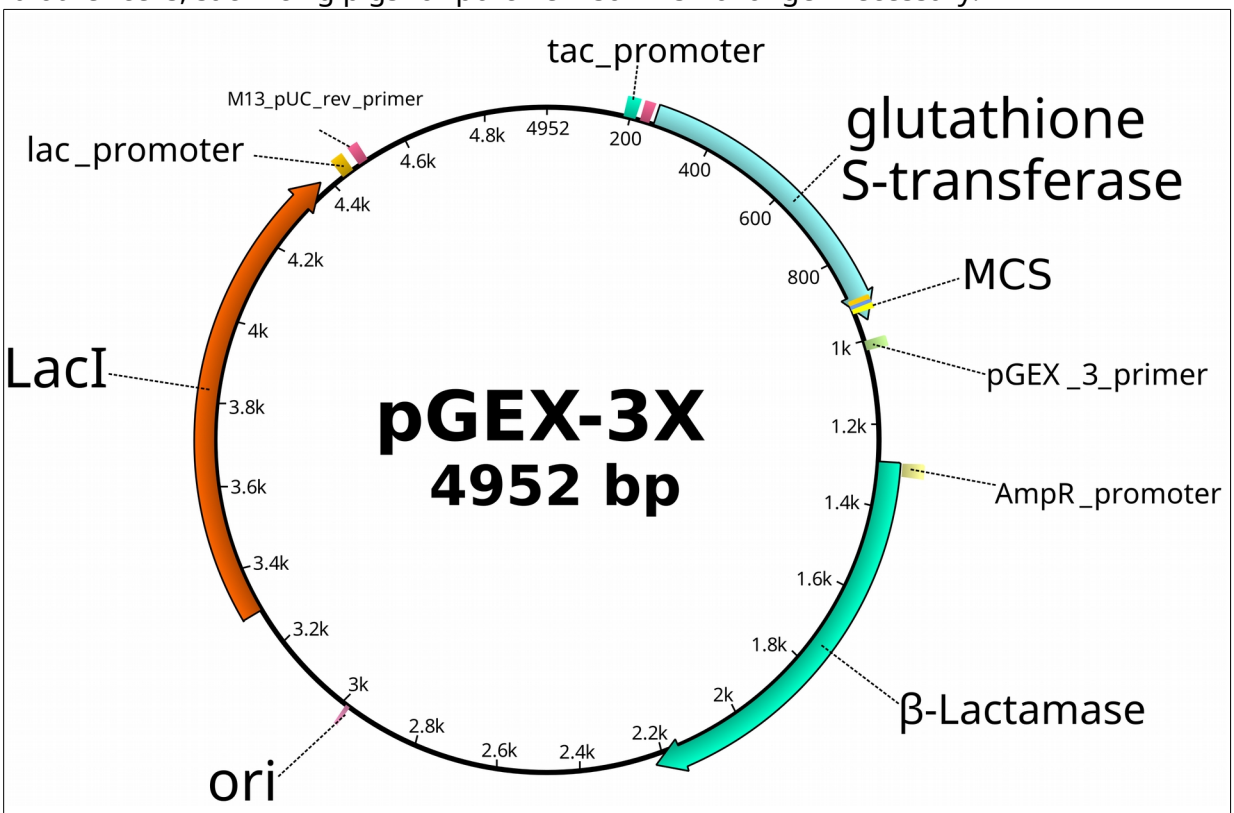
1. Label one tube “T FP” for food product.
2. Label one tube “T WT” for wild-type soy.
3. Label “T RR” for Roundup Ready® soy plant.
4. Different groups will only do one control.
5. Add **22.5 µL** of the Tubulin primer/loading dye mix to each tube containing PCR bead.
    - 5'-GGGATCCACTTCATGCTTTCGTCC-3' (Forward Primer)
    - 5'-GGGAACCACATCACCACGGTACAT-3' (Reverse Primer)
6. Add **2.5 µL** of food product DNA to the reaction tube marked “T FP.”
7. Add **2.5 µL** of wild-type or Roundup Ready® soybean DNA to the appropriate reaction tube marked “T WT" or “T RR.”

Jeremy Seto jseto@citytech.cuny.edu

Protein Expression

https://youtu.be/y4FEKGOHWLM

Recombinant DNA technology has many uses in basic scientific research to better understand the nature of of living things. As a tool, recombinant DNA technology can be used to express proteins towards medical applications. Prior to biotechnology, type I diabetes (insulin-dependent) was treated by injection of insulin isolated from the pancreas of pigs. With the ability to express human proteins inside bacteria, yeast and other cells, sacrificing pigs for porcine insulin is no longer necessary.



Bacterial expression vector pGEX-3X contains the AmpR gene, origin of replication, MCS downstream of the hybrid lac/trp promoter (tac) and the coding sequence for glutathione-S-transferase (GST). GST acts as a tag that is fused directly with the protein from the gene of interest and used to purify the protein with a glutathione resin.

Bacteria or other cells can be engineered to express proteins through the process of cloning and transformation. Bacteria are advantageous because of their rapid life cycle and ease of growth. A bacterial expression vector contains the basic plasmid features: origin of replication as well as antibiotic resistance gene. Often, an **affinity tag** will be used to aid in purification of the protein. An example in the vector above shows the **GST** (glutathione-s-transferase) tag that can be purified with glutathione resin. Expression is only the first problem since bacteria are also synthesizing proteins that are required for the bacteria to grow and divide. Injecting these proteins in addition to insulin would cause an immune reaction that could be deadly. Therefore, it is required that overexpressed proteins be purified and isolated from other undesirable proteins.

Credit: Stewart EJ, Madden R, Paul G, Taddei F (CC-SA 3.0)

Jeremy Seto jseto@citytech.cuny.edu

Criteria for Choosing an Expression System

Protein expression systems have inherent advantages and disadvantages. The table above summarizes the comparison of the various cellular systems of production (Fernandez & Hoeffler, 1999).

Desired characteristics	Bacteria	Yeast	Insect (baculovirus)	Mammalian cell culture (COS, CHO, HEK)
Cell growth	Rapid	Rapid	Slow	Slow
Complexity of growth medium	Minimum	Minimum	Complex	Complex
Cost of growth medium	Low	Low	High	High
Expression level	High	Low to high	Low to high	Low to moderate
Extracellular expression	Secretion to periplasm	Secretion to medium	Secretion to medium	Secretion to medium
Protein folding	Refolding usually required	Refolding may be required	Proper	Proper
Glycosylation	No	Yes	Some	Yes

Gene Expression Systems. Using nature for the art of expression (Fernandez, J.M. & Hoeffler, J.P., eds), Academic Press, San Diego, 1999.

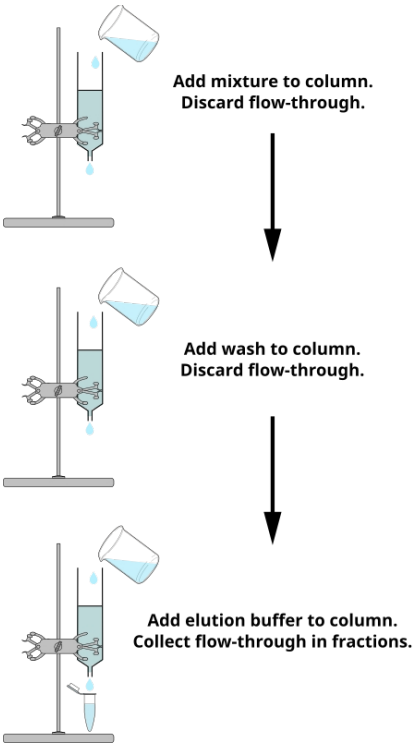
A bioreactor or fermenter that is used to grow large amounts of bacteria for the production of protein.

Credit: Miropiro CC-BY-SA 3.0)

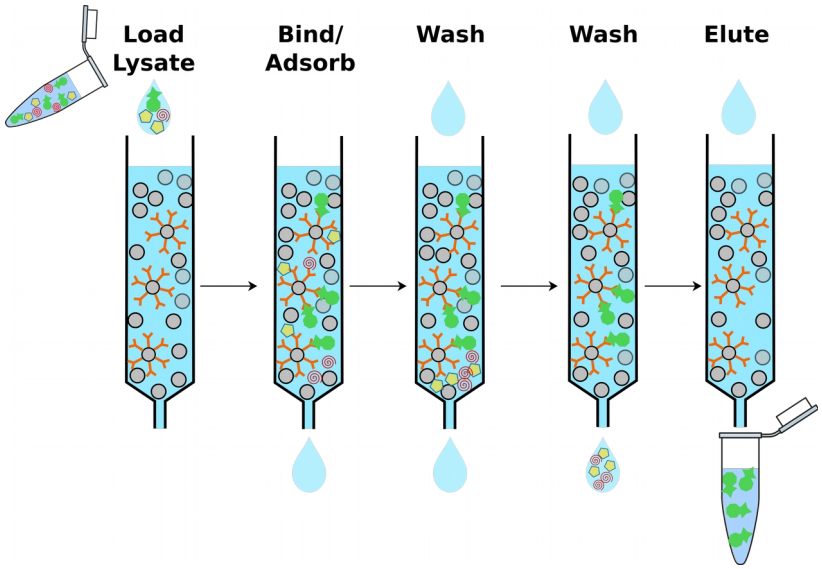
Jeremy Seto jseto@citytech.cuny.edu

Purification

Different methods of isolation can be applied depending on the properties of the protein. **Ion exchange chromatography** is useful if the protein of interest has a specific charge that will interact with a resin packed with the opposite charge.



Immunoprecipitation

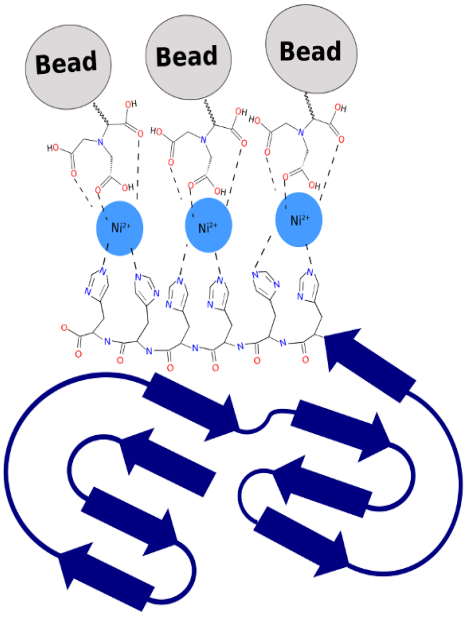


Immunoprecitation: *Column is packed with Protein-A agarose which binds to antibodies. Cell lysates are then loaded onto the columns where they flow through and are allowed to interact with the antibody. Washes are performed to remove the non-specifically bound proteins. An elution buffer is used to disrupt the interaction of the antibody to the protein target.*

Jeremy Seto jseto@citytech.cuny.edu

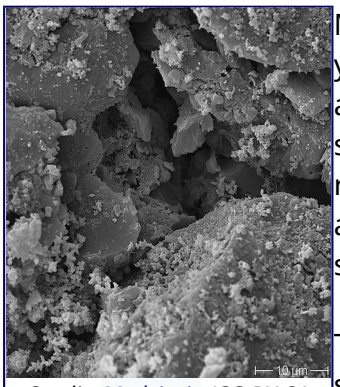
Affinity Purification

Affinity purification employs the use of specific antibodies that bind to the protein of interest very tightly to retain it on a column. With these techniques, the protein retained on the resin is washed numerous times to remove other proteins that are non-specifically sticking. A change in pH or ionic conditions then is used to disrupt the interaction with the resin and **elute** the proteins from the column. Proteins that are engineered to contain tags can be purified by antibodies specific to those tags. Also, the addition of 6 or more consecutive Histidine residues to the end of a protein make them susceptible to purification with Nickel-NTA resin or Cobalt purification. In these cases, the 6XHis tag associates with these metal ions on the resin and are selectively adhered.



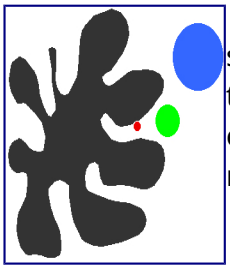
Nickel NTA resin coordinating the capture of a 6His tagged protein

Size Exclusion



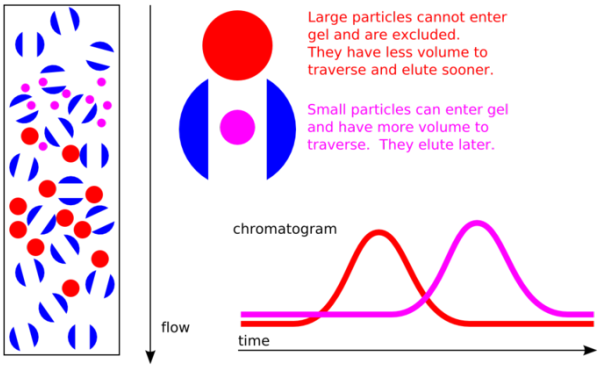
Credit: Mydriatic (CC-BY-SA 3.0)

Most of you are familiar with water purification filters. Before using these filters, you soak them in water and dark residue leaks out. This dark residue is activated charcoal. The activated charcoal has tiny microscopic pores that trap small items like ions and other particles. The primary goal of these filters is to remove metals and chlorine that are found in tap water. The porous nature of activated charcoal renders it useful for trapping molecules in water purification systems .



The process used to trap these small particles is called **size exclusion**. Unlike agarose gel electrophoresis where the smaller particles navigate through the matrix faster, size exclusion resins trap the smaller molecules.

The smaller the molecule, the longer they spend within the pores as they traverse through the matrix.



Credit: Takometer (CC-BY 2.5)

Jeremy Seto jseto@citytech.cuny.edu

Significance of Purification

All injectible drugs must be clean of endotoxins from bacteria. Purification of the protein of interest from bacterial lysates removes the dangerous pathogenic materials from that would otherwise activate host immune reactivity.

Credit: Hans Hillewaert (CC-BY)

The horseshoe crab (*Limulus polyphemus*) performs a special function in the ecosystem by providing eggs for migratory birds to feed on. This organism also houses a special cell type in its hemolymph. The limulus amoebocyte lysate (**LAL**) test is the most sensitive assay of detecting endotoxins from bacteria. Amoebocytes are collected from these organisms for use on testing batches of injectible drugs to ensure proper purification and safety.

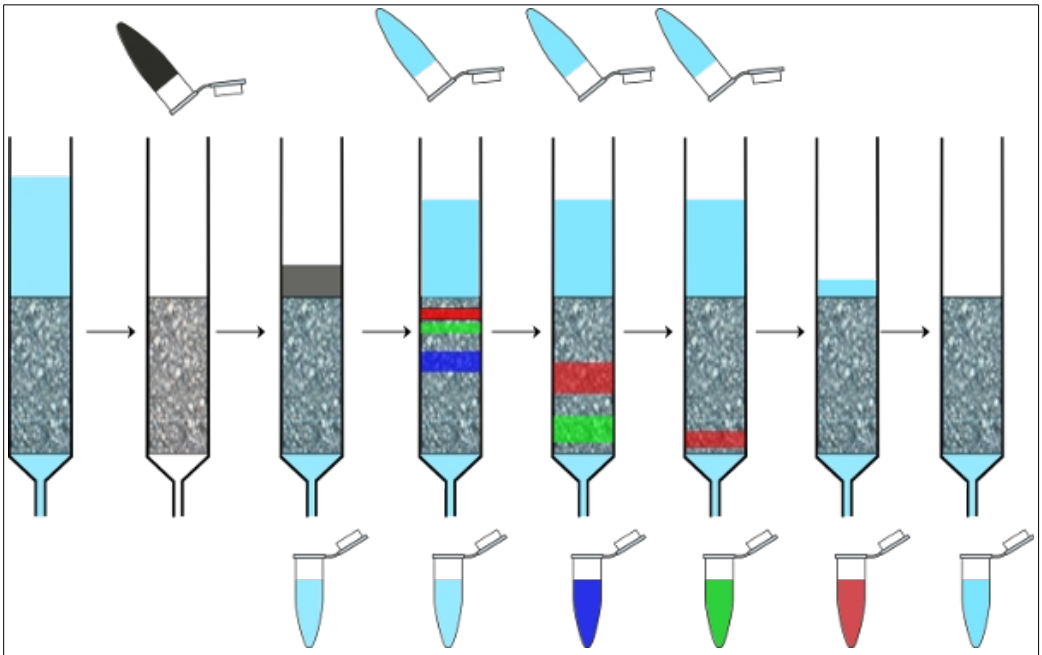
References

- Joseph M. Fernandez and James P. Hoeffler. Introduction: SO MANY POSSIBILITIES: HOW TO CHOOSE A SYSTEM TO ACHIEVE YOUR SPECIFIC GOAL, In Gene Expression Systems. Academic Press, San Diego. 1999, Pages 1-5, ISBN 9780122538407. http://dx.doi.org/10.1016/B978-012253840-7/50001-8.

Jeremy Seto jseto@citytech.cuny.edu

Size-exclusion of dye molecules

As a demonstration, the instructor may illustrate the concept of size exclusion on a set of mixed food coloring.



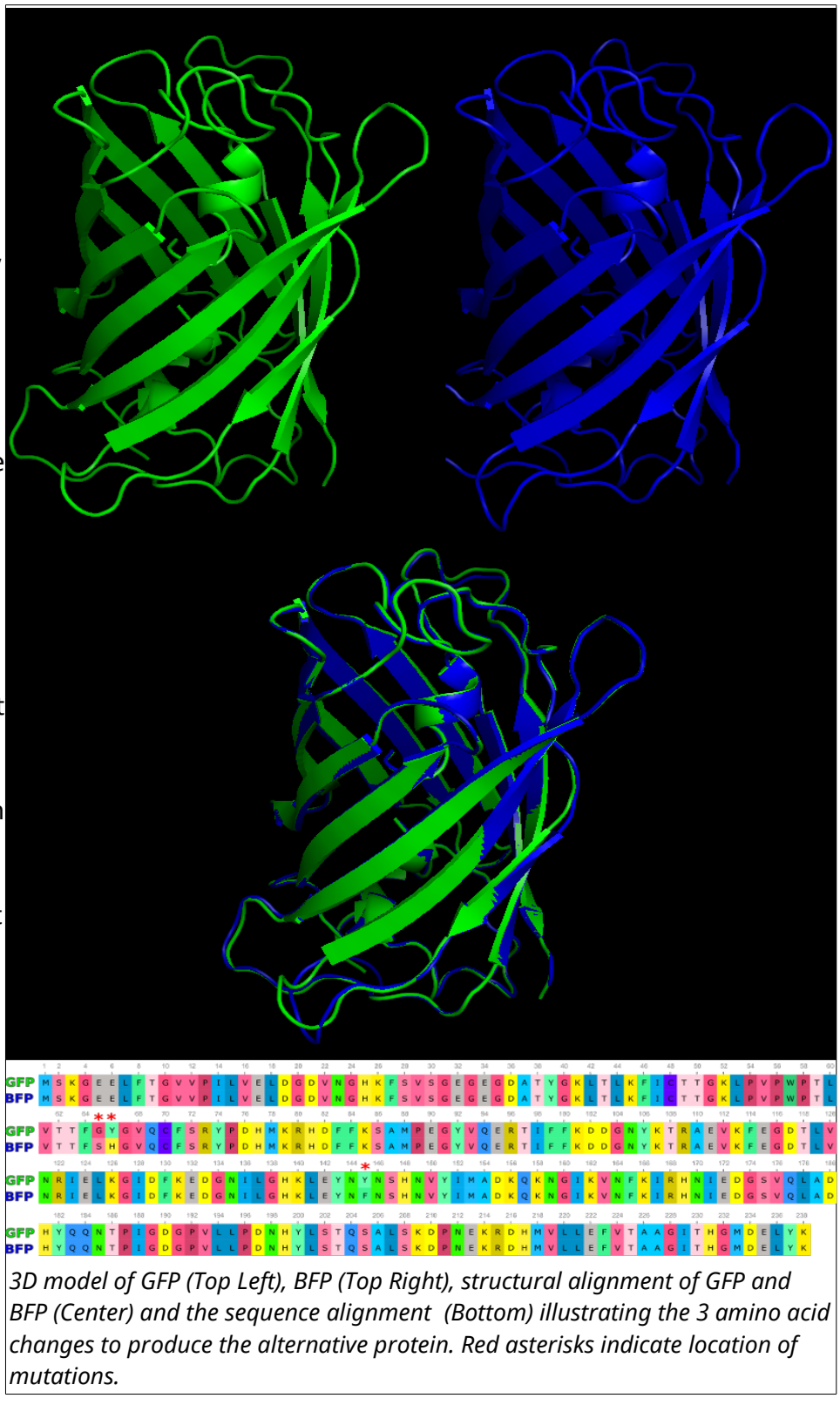
Size exclusion chromatography of food coloring

1. Pack column with 3 ml sepharose slurry
2. Let the column empty over a beaker
3. Carefully load 0.2 ml of food coloring mixture onto the column
4. Place 10 tubes on a rack under the column
5. Place 1 ml buffer on column and collect 0.5 ml fractions
6. Continue to add buffer 1 ml at a time until all fractions have been collected

Jeremy Seto jseto@citytech.cuny.edu

Size-exclusion of Proteins

This exercise seeks to purify Green Fluorescent Protein (GFP) or Blue Fluorescent Protein (BFP) from bacterial lysate. These proteins have a specific size of 238 amino acids and are 40,000 daltons (40kD). Based on their specific size, they will have a specific rate of migration through the size exclusion resin. Remember that the bacterial lysate is full of additional proteins that are not your protein of interest that we are attempting to isolate.



3D model of GFP (Top Left), BFP (Top Right), structural alignment of GFP and BFP (Center) and the sequence alignment (Bottom) illustrating the 3 amino acid changes to produce the alternative protein. Red asterisks indicate location of mutations.

Drops of fluid will be collected in fractions. The fractions containing the fluorescent proteins will be found only in specific fractions that will be visible under UV illumination.

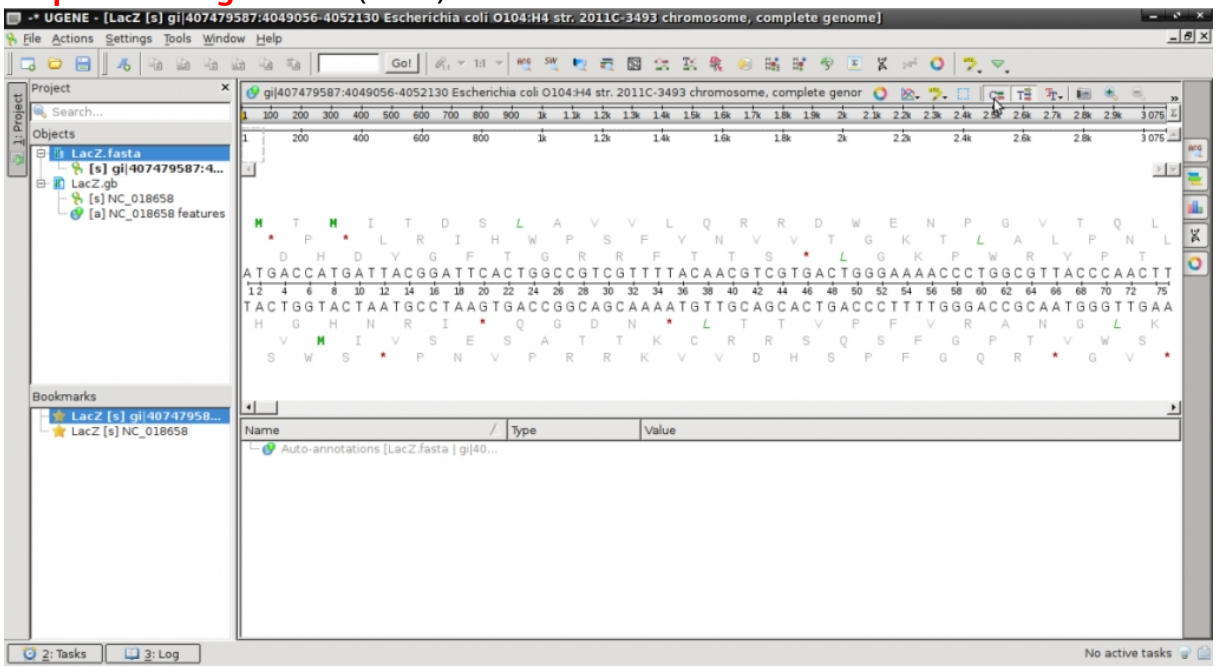
1. Vertically mount the column on a ring stand. Make sure it is straight.
2. Slide the cap onto the spout at the bottom of the column.
3. Mix the slurry (molecular sieve) thoroughly by swirling or gently stirring.
4. Carefully pipet 2 ml of the mixed slurry into the column by letting it stream down the inside walls of the column.
5. Place an empty beaker under the column to collect wash buffer.
6. Remove the cap from the bottom of the column and allow the matrix to pack into the column.

Jeremy Seto jseto@citytech.cuny.edu

7. Label eight microcentrifuge tubes #1-8.
8. Slowly load the column with 0.2ml of the GFP extract. Allow the extract to completely enter the column.
9. Add 1ml of elution buffer on top of resin without disturbing the resin
    - Add buffer slowly (several drops at a time) to avoid diluting the protein sample.
    - Using the graduated marks on the sides of the tubes, collect 0.5ml fractions in the labeled microcentrifuge tubes.
    - Continue to add 1ml buffer and collect fractions until all tubes are full
10. Check all fractions by using long wave U.V. light to identify tubes that contain the fluorescent GFP or BFP proteins.
11. Further purification may be performed with a different resin with the few fractions containing the protein of interest
12. Protein samples should be run on an acrylamide gel and stained against all proteins to check the purity of the sample or fluorescence measurements taken

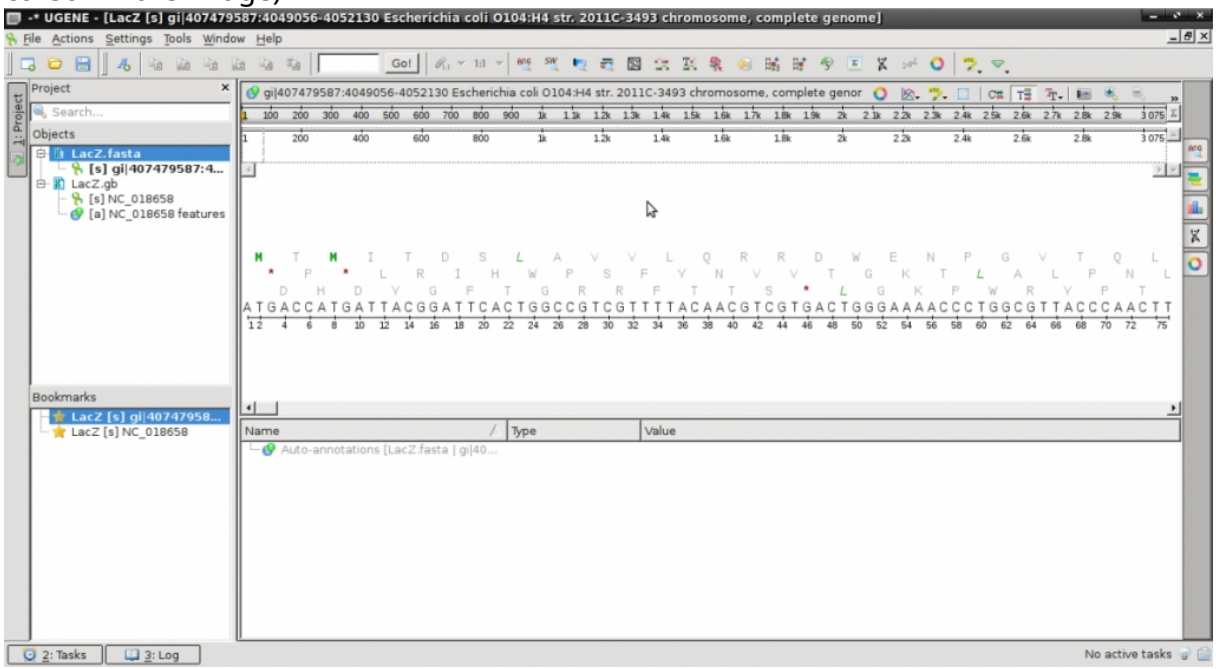
Jeremy Seto jseto@citytech.cuny.edu

1. Download the file LacZ.gb and open in a text editor.
    - This is a **Genbank format** file that contains the sequence following the word '**ORIGIN**' and terminating with '**//**'.
    - Prior to the sequence is a batch of descriptive information including references, organism and database cross-reference identifiers. While these don't mean much to you, the appropriate database within Genbank can be queried to reveal more information about the sequence.
2. Download the file **LacZ.fasta** and open in a text editor (NotePad or Texedit).
    - Notice the simple structure of the **fasta** file beginning with the '>' and description of the sequence.
    - This is a DNA sequence. ***But DNA is usually double stranded!*** We can assume the sequence of the second strand because it will be complimentary to this one.
        - By convention: we know that this sequence is **5' → 3'**
    - This text contains a portion of the *E. coli* genome that includes a gene called LacZ.
    - This file does not contain any annotation to indicate where the gene sequence actually begins or ends.
3. Launch UGENE and open both files. They will appear on the left side "Objects" pane.
    - The default display automatically shows the reverse compliment of the DNA strand and all 6 **Open Reading Frames** (ORFs).



Jeremy Seto jseto@citytech.cuny.edu

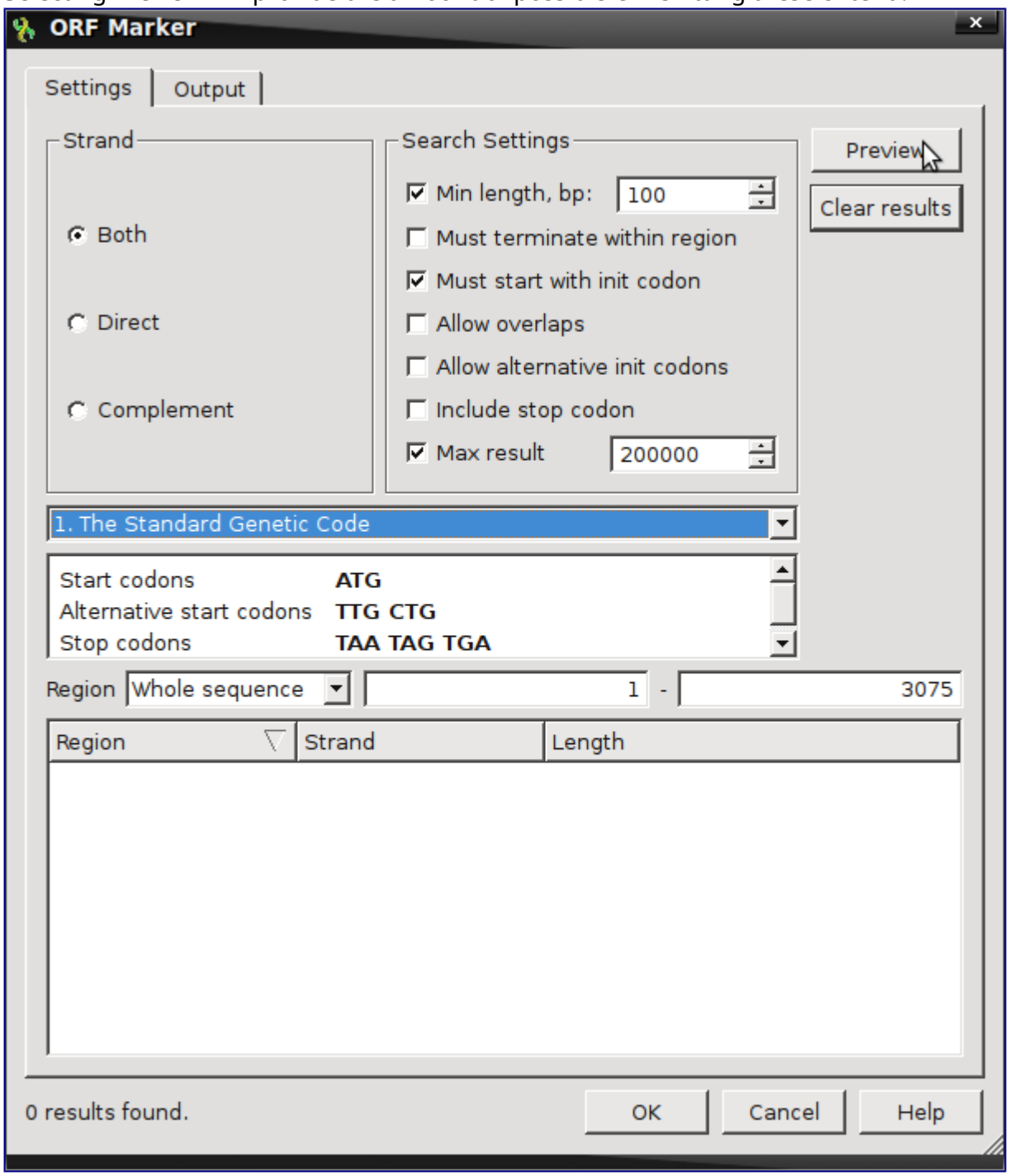
- To simplify the view, click on the '**C**' to remove the complimentary strand (look at the cursor in the image)



4. **Count the ORFs** :
   - Find ORFs by right-clicking on the sequence and select "**Analyze → Find ORFs**"
   - Default setting looks for ORFs on both strands with a minimum length of 100 nucleotides
   - The **Open Reading Frame** here is defined as something beginning with initiation or start codons from the Standard Genetic Code (**ATG**) and two additional alternative start codons (**TTG** & **CTG**) that is terminated by any one of the three standard stop codons (**TAA**, **TAG**, **TGA**)

Jeremy Seto jseto@citytech.cuny.edu

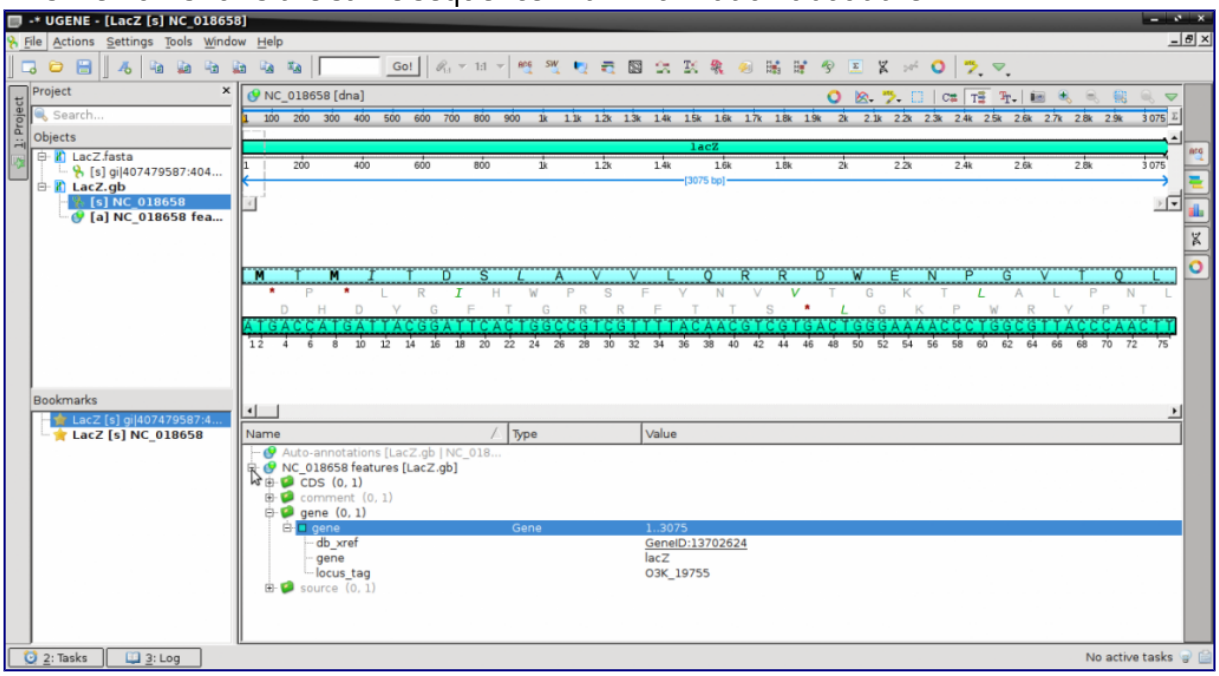
- Selecting Preview will provide the amount of possible ORFs fitting these criteria.



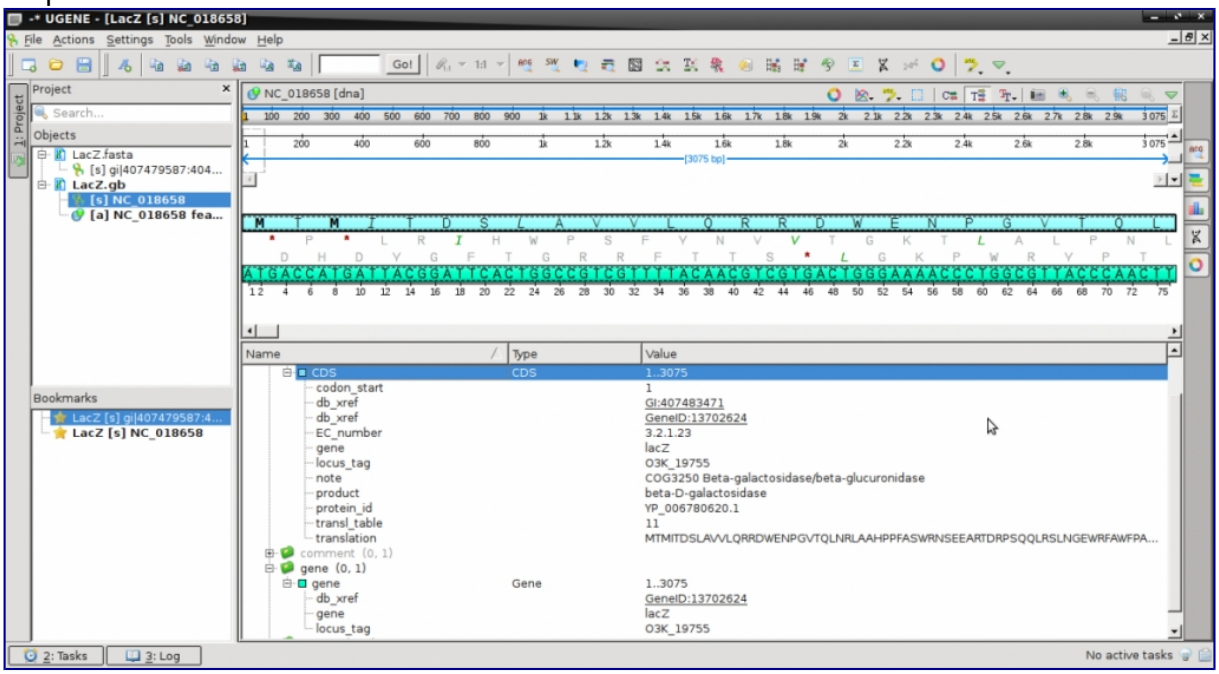
5. Double click on the **LacZ.gb** in the **Objects** panel to activate the view.

Jeremy Seto jseto@citytech.cuny.edu

- This file now shows the same sequence with information about the DNA



- Expand the various features in the Annotations pane at the bottom to explore the sequence features.

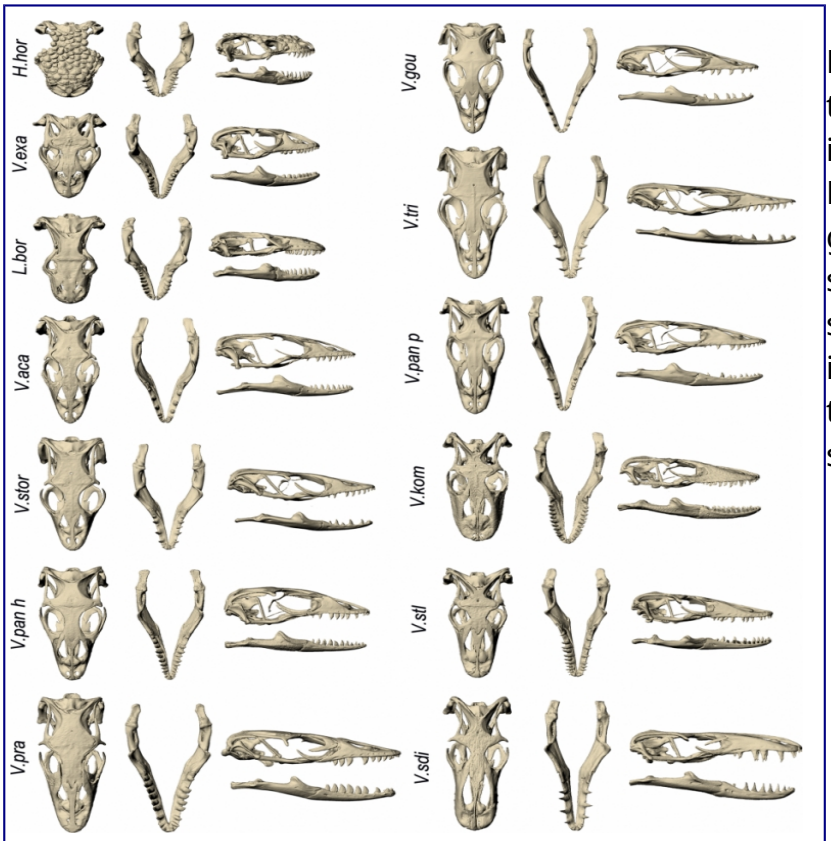


Jeremy Seto jseto@citytech.cuny.edu

Morphometrics and physical markers

Morphometrics (*morpho*- shape; *metrics*- measurements) is the use of physical measurements to determine the relatedness of organisms. With extinct organisms that have died out long ago, DNA extraction proves to be difficult. Likewise, prior to DNA technologies to analyze species, **Linnean taxonomy** was ascribed to organisms based on similarities in features.

Describing Species and Variation of Morphologies

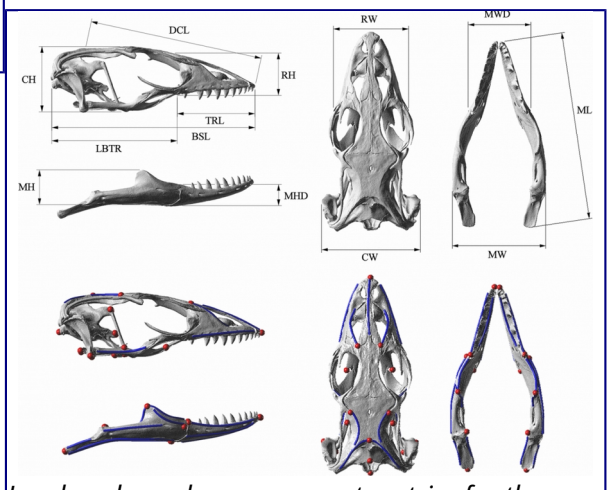


Skulls of the species involved in this analysis.
Credit: McCurry et al. (2015) (CC-BY)

Below are images of skull landmarks of the lizard family Varanidae. This family includes monitor lizards and Komodo Dragons.As can be seen below, the general morphology of the skulls are similar enough that they all retain the same landmarks. The figure below also illustrates the diversity in these lizards that illustrate a large variety between species.

Landmarks Standardize measurements

Having a set of shared landmarks provides the opportunity to make systematic measurements of morphometric features.



Landmarks and measurement metrics for the morphometric analysis of skulls.
Credit: McCurry et al. (2015) (CC-BY)

Jeremy Seto jseto@citytech.cuny.edu

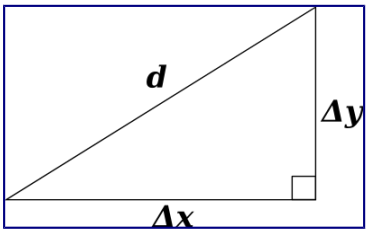
Euclidean distance to measure relatedness

Euclidean distance is a measurement derived from Pythagorean geometry that describes the shortest distance (d) between 2 points (A & B) as a straight line using triangulation. In a cartesian space, the points can be defined:

$$A = (x_A, y_A) \text{ and } B = (x_B, y_B)$$

Standard pythagorean theorem can be expressed as:

$$x^2 + y^2 = d^2$$



To find the distance between the 2 points, we utilize algebra to calculate for d.

$$d = \sqrt{x^2 + y^2}$$

In this case, we expand to comparing the coordinates of the two points:

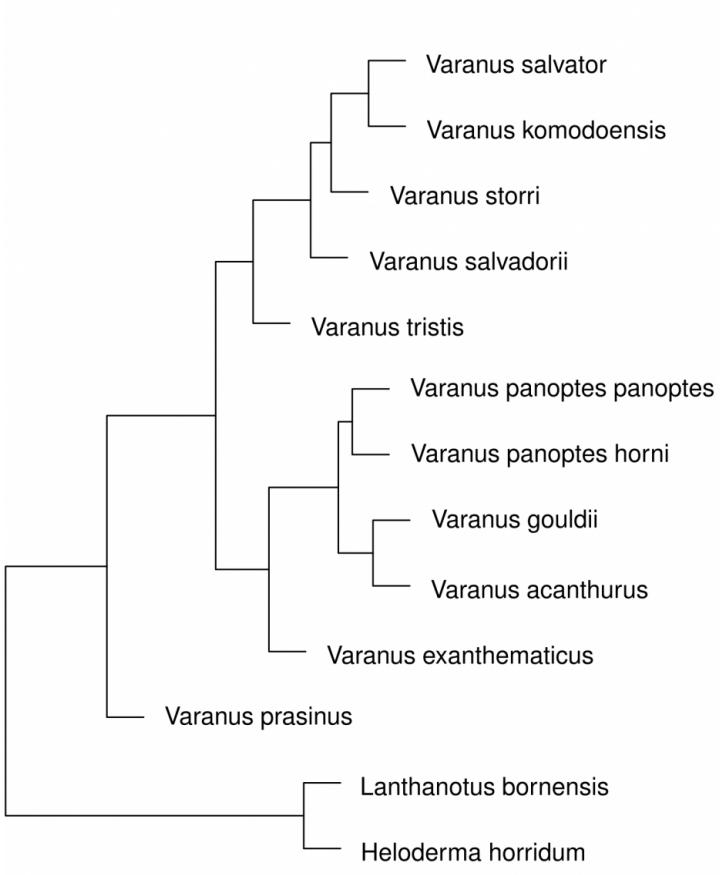
$$\Delta x = x_B - x_A \text{ and } \quad \Delta y = y_B - y_A$$

We can then expand this idea to include the differences of data points that describe the comparisons of multiple measurements.

$$d(\mathbf{X_i}, \mathbf{X_j}) = \sqrt{\sum_{k=1}^{p} (X_{ik} - X_{jk})^2}$$

Calculating distance with R

1. Download the **dataset** (McCurry *et al.* 2015) associated with this activity (a Comma Separated Value .csv file). This can be used in a spreadsheet or in a text editor. This data can be imported into R to determine the euclidean distances of landmarks.
2. The following code in R will download the data set into a variable called "varanoid", measure euclidean distance and save a plot into a PDF file in a directory called "/tmp".



Varanoid Dendrogram from R

Jeremy Seto jseto@citytech.cuny.edu

```
## install curl for fetching from internet if it isn't
install.packages('curl')
## Load the curl library
library(curl)
## read the data of measurements and assign it to a variable 'varanoid'
varanoid = read.csv(curl('https://raw.githubusercontent.com/jeremyseto/bio-
oer/master/data/varanoid.csv'))
## set the row names to the Species column
row.names(varanoid) = varanoid$Species
## remove the first column of the table to have purely numeric data
varanoid_truncated = (varanoid[,2:14])
## calculate distance using euclidean as the method
dist_measure = dist(varanoid_truncated, method='euclidean')
## display dist_measure to look at the comparisons
dist_measure
varanoid_cluster = hclust(dist_measure)
## open PDF as a graphics device  to save a file in the '/tmp' directory
pdf(file='/tmp/varanoid_tree.pdf')
plot(varanoid_cluster)
dev.off()
## close the device to save the plot as pdf
```

DNA Analysis

Before starting this activity, review **bioinformatics** and **sequence analysis**.

1. Search NCBI for mitochondrial sequences from the species involved in McCurry 2015. The data has been submitted by Ast (2001).
2. Find the sequences and identify/extract elements that are common to all
3. Assemble the shared sequences in a text editor as a single FASTA file where each species is separated by a header (">Species A")
    - Notepad on Windows (but it's better to download notepad++)
    - Textedit on Mac (but probably better to download TextWrangler)
    - Gedit on Linux
4. Save the file as "something.fasta"
5. Perform a multiple sequence analysis using **UGENE**
6. Generate a phylogenetic tree using UGENE. For this exercise, use Maximum Likelihood (PhyML) as the algorithm. File the tutorial below.
7. Compare the DNA with the morphometric analyses. What problems could we imagine arise if we rely solely on morphometry.

https://youtu.be/Z-dlww1V9Y8

Jeremy Seto jseto@citytech.cuny.edu

References

- McCurry MR, Mahony M, Clausen PD, Quayle MR, Walmsley CW, Jessop TS, Wroe S, Richards H, McHenry CR. (2015) **The Relationship between Cranial Structure, Biomechanical Performance and Ecological Diversity in Varanoid Lizards.** *PLoS ONE* 10(6): e0130625. doi: 10.1371/journal.pone.0130625
- Ast, Jennifer C. (2001) **Mitochondrial DNA Evidence and Evolution in Varanoidea (Squamata**). *Cladistics* 17(3): 211–26. http://www.sciencedirect.com/science/article/pii/S0748300701901690
- Fisher, R.A. (1936) **The use of multiple measurements in taxonomic problems.** *Annals of Eugenics,* 7: 179–188. doi:10.1111/j.1469-1809.1936.tb02137.x

UGENE

The following video illustrates the tree building process using MUSCLE and PhyML in **UGENE**.

https://youtu.be/Z-dlww1V9Y8

Command Line

The following requires:

- A UNIX-like environment like Linux or MacOS
- **MUSCLE** to perform a multiple sequence alignment
- **PhyML** to generate Maximum Likelihood
- **FigTree** to manipulate the tree

Download the example file **oranges**. In the download directory, perform the following:

```
[bash]
unzip orange.zip
cd orange
cat ./*txt >> oranges.fasta  ## merges all files into a single fasta file
muscle -in oranges.fasta -phyiout oranges.phy
## -phyiout tells muscle to use the interleaved phylip format for output
phyml -i oranges.phy -m HKY85
## -m is for method and HKY85 is the default nucleotide method we used in UGENE
mv oranges.phy_phyml_tree.txt oranges.nwk
## change the name of the output to reflect it is a nwk file
[/bash]
```


Jeremy Seto jseto@citytech.cuny.edu