



(<https://academicworks.cuny.edu>)

Open Educational Resources

YOU HAVE SUCCESSFULLY REVISED: SCB252 FUNDAMENTALS OF BIOTECHNIQUES LABORATORY MANUAL

[Revise Submission](#) ([ir_submit.cgi?context=lg_oers&article=1149&window=upload_article&edypass=&editpanel=](#))

[Make Another Submission](#) ([ir_submit.cgi?context=lg_oers&edypass=&editpanel=](#))

Submission Metadata

Metadata Field	Value
Title	SCB252 Fundamentals of Biotechniques Laboratory Manual
Authors	Na Xu, CUNY La Guardia Community College and Hanqi Ye, CUNY La Guardia Community College
Affiliations at CUNY	LaGuardia Community College, Natural Sciences
Document Type	Activity or Lab
Publication Date	Spring 5-15-2026
Embargo Period	5-15-2026
Keywords	Biotechniques, Biotechnology
Disciplines	Biotechnology
Language	English

Abstract This laboratory manual for SCB252: Fundamentals of Biotechniques provides an integrated, hands-on introduction to core concepts and methods in modern biological research using *Drosophila melanogaster* as a model organism. The manual emphasizes laboratory safety, research integrity, and responsible conduct of research, alongside foundational experimental skills. Students are guided through key biological techniques including fly husbandry, genetic crosses, mutant analysis, microscopy, DNA extraction, polymerase chain reaction (PCR), gel electrophoresis, sequencing technologies, and bioinformatics tools such as BLAST and genome databases. Advanced modules introduce gene structure annotation, RNA-seq data analysis, and primer design, while fluorescence microscopy and protein analysis techniques further expand technical proficiency.

The curriculum also integrates broader scientific themes, including virology, genetic engineering, cancer biology, and ethical considerations in biotechnology, fostering connections between laboratory practice and societal impact. Inquiry-based research projects and data analysis exercises support the development of experimental design, statistical reasoning, and scientific communication skills. Overall, this manual serves as a comprehensive resource for training students in fundamental biotechniques while preparing them for independent research and critical engagement with biological sciences.

Comments - empty -

Creative Commons License



(<https://creativecommons.org/licenses/by-nc-sa/4.0/>)

This work is licensed under a [Creative Commons Attribution-Noncommercial-Share Alike 4.0 License](https://creativecommons.org/licenses/by-nc-sa/4.0/) (<https://creativecommons.org/licenses/by-nc-sa/4.0/>).

CUNY OER Funding CUNY OER Initiative

Accessibility Check True

Upload File [Download](#) (https://academicworks.cuny.edu/cgi/viewcontent.cgi?article=1149&context=lg_oers)

Update Submission Revision

SCB252 Fundamentals of Biotechniques Laboratory Manual

Open Educational Resources (OER)

Na Xu and Hanqi Ye

Natural Sciences Department
LaGuardia Community College

Permissions and Licensing

Except where otherwise noted, this lab manual is protected under a **Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0)**.



You are permitted to:

- **Share:** Copy and redistribute the material in any medium or format.
- **Adapt:** Remix, transform, and build upon the material.

Under the following conditions:

- **Attribution (BY):** You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
- **NonCommercial (NC):** You may not use the material for commercial purposes.
- **ShareAlike (SA):** If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.

For the full license details, please visit:

<https://creativecommons.org/licenses/by-nc-sa/4.0/>

Important Note on Images

Please be aware that some images within this lab manual are credited to their original owners and **do not fall under the Creative Commons license of this manual**. Therefore, you should use these images with care and respect their individual copyrights. We are actively working to replace these third-party images with our own original content in future versions of this manual.

Lab 1 Laboratory Safety, Research Integrity, Responsible Conduct of Research, Intro to fruit fly

Na Xu , Hanqi Ye

Objectives

- To be familiar with lab safety, research integrity, and basic principles of responsible conduct of research.
- To complete Responsible Conduct of Research training through the Collaborative Institutional Training Initiative (CITI).
- Intro to the fruit fly and prepare new food, prepare for cross.

Introduction

Research integrity includes the use of honest and verifiable methods in proposing, performing, evaluating, and reporting research results, with particular attention to rules, regulations, guidelines, and commonly accepted professional codes or norms. During this class, students will learn the basic principles of **the Responsible Conduct of Research** (Steneck and Zinn, 2006): <https://ori.hhs.gov/education/products/RCRintro/index.html>

The City University of New York (CUNY) subscribes to the Collaborative Institutional Training Initiative's (CITI) research compliance courses (RCR) to fulfill the institution's educational requirements. All CUNY faculty members, staff, postdoctoral scholars, and graduate and undergraduate students involved in research are required to complete the **CITI RCR training** within six weeks of initiating their research. The RCR training certificate will be valid for four years. CUNY researchers are required to take a refresher CITI RCR training course every four years. Introduction to the RCR online course can be found here:

<https://about.citiprogram.org/series/responsible-conduct-of-research-rcr/>

Instructions to complete the RCR training through CITI:

1. Login (or register if you are a first-time user) at <https://www.citiprogram.org/Default.asp>
2. Click on "Add a course or update your learner groups"
3. Select the curriculum "Responsible Conduct of Research"

4. Select the learner group “CUNY Researchers” as your role in research.
5. Complete the required modules
6. Save a copy of your completion certificate.

Introduction to Fruit flies

Rearing flies in the laboratory: Ideally, flies are reared and handled in rooms that maintain a temperature between 20 and 25°C since higher or lower temperatures have adverse effects on fly viability and fecundity. Continued exposure to 30°C or above may cause sterility or death, whereas temperatures below 18°C delay the life cycle and may impair viability. An appropriate medium in suitable containers is required to maintain *Drosophila* in the laboratory. Semi-transparent plastic bottles and vials are plugged with non-absorbent cotton or synthetic foam plugs.

Flies require care to maintain optimal health. Overcrowding is a common concern since flies continue laying eggs and expand their numbers within the limited space and resources available in the bottle/vial. Furthermore, the build-up of dead flies adversely affects fly fitness. Decaying medium and fly debris attract infecting agents including bacteria, fungi, and mites. Therefore, the flies must be transferred regularly to fresh bottles/vials with a medium to maintain health. Starting a new culture with approximately 10–20 flies in vials and < 50 flies in bottles avoids overcrowding ensuring a healthy culture. As a good practice, flies reared at 25°C must be transferred to a fresh food source within 25–30 days if cultured in bottles and 15–20 days if cultured in vials.

Cultures should be handled carefully to avoid flies escaping from the containers during the maintenance process. Escaped flies may contaminate other cultures in the laboratory or fresh medium vials/bottles. To eliminate unwanted flies in the laboratory, fly traps should be placed in the laboratory area. A simple fly trap can be prepared with a fresh medium bottle and a funnel placed at its open end. A dilute solution of activated yeast with a few drops of propionic acid as a thin fluid layer on the food surface attracts the flies, which enter the trap through the funnel. Replace the fly traps at least once a week.

Killing Flies and Setting up a Morgue: This is an unfortunate necessity when using flies. A bottle or beaker with soapy water or mineral oil is generally used for this purpose. Place anesthetized flies directly into the soapy water or mineral oil where they drown. A bottle (beaker, or screw-capped jar) filled with ethanol or isopropanol can also be used as a morgue.

Lab Exercise I -Transfer files to new food bottles

Materials

30 vials of flies (6xEb,6xVg, 6xCn, 6xDp, 6xWT)
30 new vials
1 bag of cotton balls
4-24 instant food
Dry yeast
12 markers
6 weighing boats
6 digital scales
6 spoons
6 10 ml graduated cylinders

Methods

For each group of students, make 5 new vials, 1 for each: Eb, Vg, Cn, Dp, WT.

For each vial, do as follows:

1. Obtain 1 vial of flies, one funnel, one marker and 1 new bottle with plug.
2. Use marker to label new vial with the Mutant symbol, today's date and your name.
3. In the empty vial, add 1.5g 4-24 instant fly food, and then measure 5g water with a graduated cylinder.
4. Add water to the vial, quickly swirl to mix and then place it on the table, wait for 2 mins.
5. After the food solidifies, add 5-10 grains of yeast on top of the food.
6. Take the fly vial, tap the vial on the table several times to settle down flies.
7. Remove the cotton quickly and transfer the flies to the new vial by flipping the vial to the new vial.
8. Lift 2 vials together and hit the bench several times, cap both vials quickly.
9. Return vials to the instructor, and clean the bench.

Lab Exercise II - Make new food vials 100 ml per group (for 6 groups)

Materials

Nutri-Fly medium
Tegosept
Ethanol
Phosphoric acid
1 1000 ul micropipette and 1 P1000 tips
1 20-200 ul micropipette and 1 P200 tips
1 tray of empty vials (x100)
1 bag of cotton balls
6 1.5 ml microcentrifuge tubes for making 10% Tegosept
6 Hotplate

6 Distilled water bottle
6 Digital scale
6 small weighing boat
6 large weighing boat
6 250 mL flasks
6 Thermometer
6 pairs of Heat resistant gloves
Paper towels

REMINDER: New food vials or bottles should be placed under the hood overnight, covered by a sheet of paper towel and capped with cotton on the next day before putting away to the fridge.

Methods

1. Preheat the hotplate at setting 3, weigh 17.8 g of the Nutri-Fly medium
2. While the plate is heating, place a 250 mL flask on the plate, add 100 mL of distilled water, a magnetic stirrer and slowly pour in the 17.8 g of Nutri-Fly food. Turn on the stirring function, mixing constantly to break up any bubbles or lumps.
3. HOT Caution, wear a hot resistant glove. Place the flask on a hotplate and set to maximum power, turn on stir with setting 1. When it starts bubbling (use a thermometer if it reaches 90°C), set the hotplate back to 3. (Observe the meniscus carefully. If it starts to rise very fast, remove the flask off the hotplate.)
4. The medium should be at the boiling stage, steadily producing bubbles, but not boiling over the container. During this time, insert a thermometer such that the bulb is just below the surface but clear of the magnetic stirrer. Set the timer for 10 min.
5. During this 10 min, prepare 10% Tegosept (dissolve 0.1 g Tegosept in 1 mL ethanol in 1.5 ml microcentrifuge tube). Tegosept is an antifungal drug to prevent fungal infections in fly cultures.
6. After 10 min, place the food on the bench. Periodically swirl the flask until temperature drops to 75°C.
7. Remove the thermometer, add 1 mL of 10% Tegosept and 34 µL of phosphoric acid and mix thoroughly by . When the medium is ready, pour it into containers quickly before it solidifies. Phosphoric acid is used to acidify the medium and has fungicidal effects. Phosphoric acid is a strong acid and corrosive. Therefore, utilize gloves when handling phosphoric acid.
8. Pour 100 mL of food into containers. Pour approximately 5 mL per vial. Wipe excess food off the sides of the vial/bottle with paper towels.

9. Stack the trays onto each other or cover the top tray with paper towels to keep flies from the medium before you can plug.

10. Leave at room temperature overnight to dry.

11. On the next day, plug vials with cotton and place them in plastic bags. Store at room temperature for short-term use and in the refrigerator for long-term use (the laboratory technician or instructor will complete this).

Assignment 1

Students will spend a part of the class time completing the Responsible Conduct of Research (RCR) training through the Collaborative Institutional Training Initiative (CITI). A certificate will be generated upon completion of the training. Students will submit the certificate to mark the completion of this assignment.

Lab 2 Introduction to Use *Drosophila* as a Model System for Scientific Research

Na Xu, Hanqi Ye

Objectives: This lab provides basic information about the *Drosophila* life cycle, maintenance, and handling in the laboratory, and fly genetics.

Introduction I

The fly *Drosophila melanogaster* is among the most well-characterized organisms in biological sciences and serves as a model system for investigating many developmental and cellular processes common to higher eukaryotes, including humans. Of the ~1500 known *Drosophila* species, ***D. melanogaster***, the Pomace fly or fruit fly, is the most extensively used and versatile experimental model. C. W. Woodworth first cultivated the fruit fly in 1900 at the Harvard Zoological Laboratory. Subsequently, *D. melanogaster* became a popular model system for research, due to the following unique features: 1. Small body size requiring little space and facilities handling, which reduces the cost of large-scale rearing in the laboratory compared to other model organisms; 2. short generation time (less than 2 weeks) with large numbers of progeny; 3. sharply defined body phenotypes that facilitate the identification of numerous mutants; 4. small genomic size.

Drosophila Life Cycle: *Drosophila* has distinct stages of development including egg, larva, pupa, and adult (Figure 1). At 25°C, an adult emerges in approximately 10 days. The egg of *D. melanogaster* is white, approximately 0.5 mm in length, and is laid in the food source. Minuscule larvae hatch from the eggs. The larval period consists of three motile stages or instars, the first, second, and third instars. The different larval instars can be identified by their age and body size, spiracles, and mouth hooks (Figure 1). The first, second, and early third instar larvae largely remain within the food, actively feed, and increase their body mass. Late third instar larvae come out of food and crawl in search of a dry environment for pupation, which is usually the wall of a vial/bottle where they are grown in the laboratory. Larvae adhere to a dry surface using the glue protein secreted by their salivary glands (Figure 1). The pupa is formed within the last larval cuticle. Gradually the cuticle hardens and becomes darker in color (Figure 2).

During the next 4–5 days, the pupa undergoes complete metamorphosis within the pupal case. Most larval tissues undergo histolysis during the metamorphosis while the adult structures differentiate (Figure 2). At the end of pupal metamorphosis, the adult fly forces its way through the anterior end of the pupal case. A newly enclosed fly is elongated, lighter in color than later stage adults, and has unexpanded wings. A small dark-greenish mark of ‘meconium’, representing the decaying larval gut, is visible on the ventral side through the young fly’s abdominal wall (Figure 1). These characteristic features of newly emerged flies are used to isolate virgin females when producing a desired genetic cross. Within 2–3 hours, the wings expand, and the body darkens. Female flies become ready to mate in 4–6 hours after eclosion.

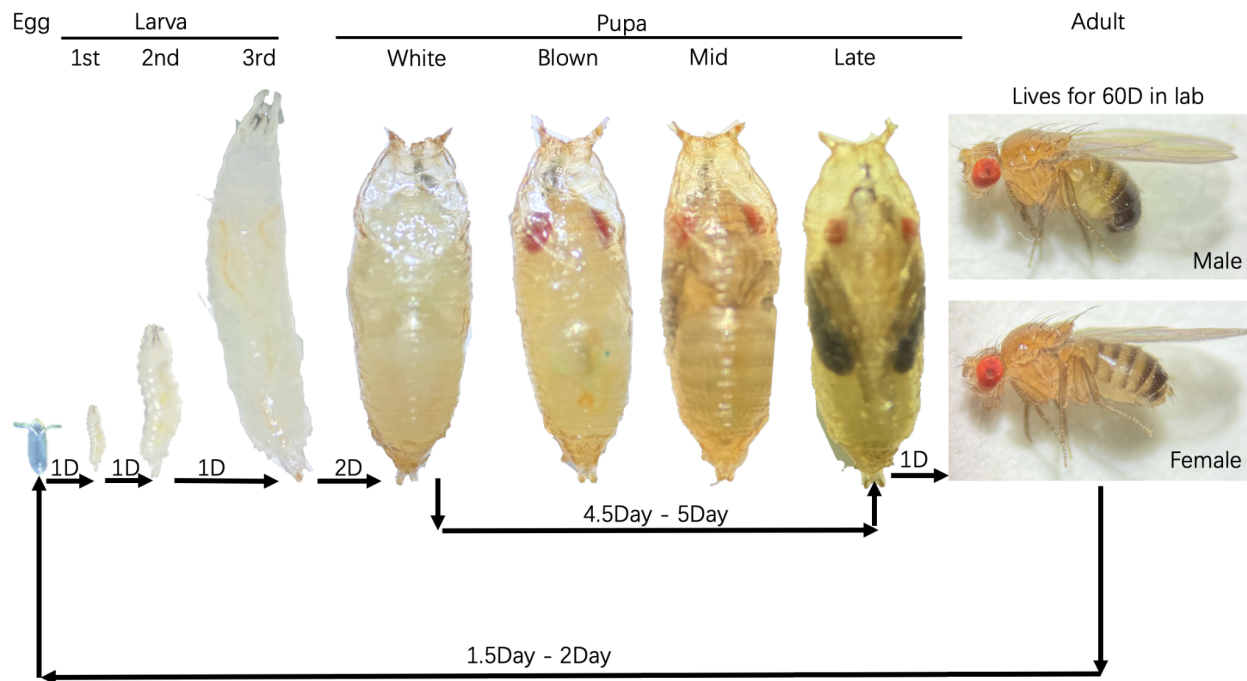


Figure 1: *Drosophila* Life Cycle. This figure shows the different stages of the *Drosophila* life cycle.

Identification of Sex: Male and female *D. melanogaster* flies (Figure 2) can be distinguished by the following features: 1. The abdomen of females has eight segments, with seven being visible. In males, the last three abdominal segments are fused, so that only five segments are visible. 2. The abdominal tip in males is considerably darker than in females. 3. Males and females have different genitalia, visible on the ventral surface of the abdominal tip, with the male genital plate appearing rougher and darker. 6. The males have sex combs (Figure 3 E'), which are a row of approximately ten stouts, and black bristles on the first tarsal segment of the first pair of legs.

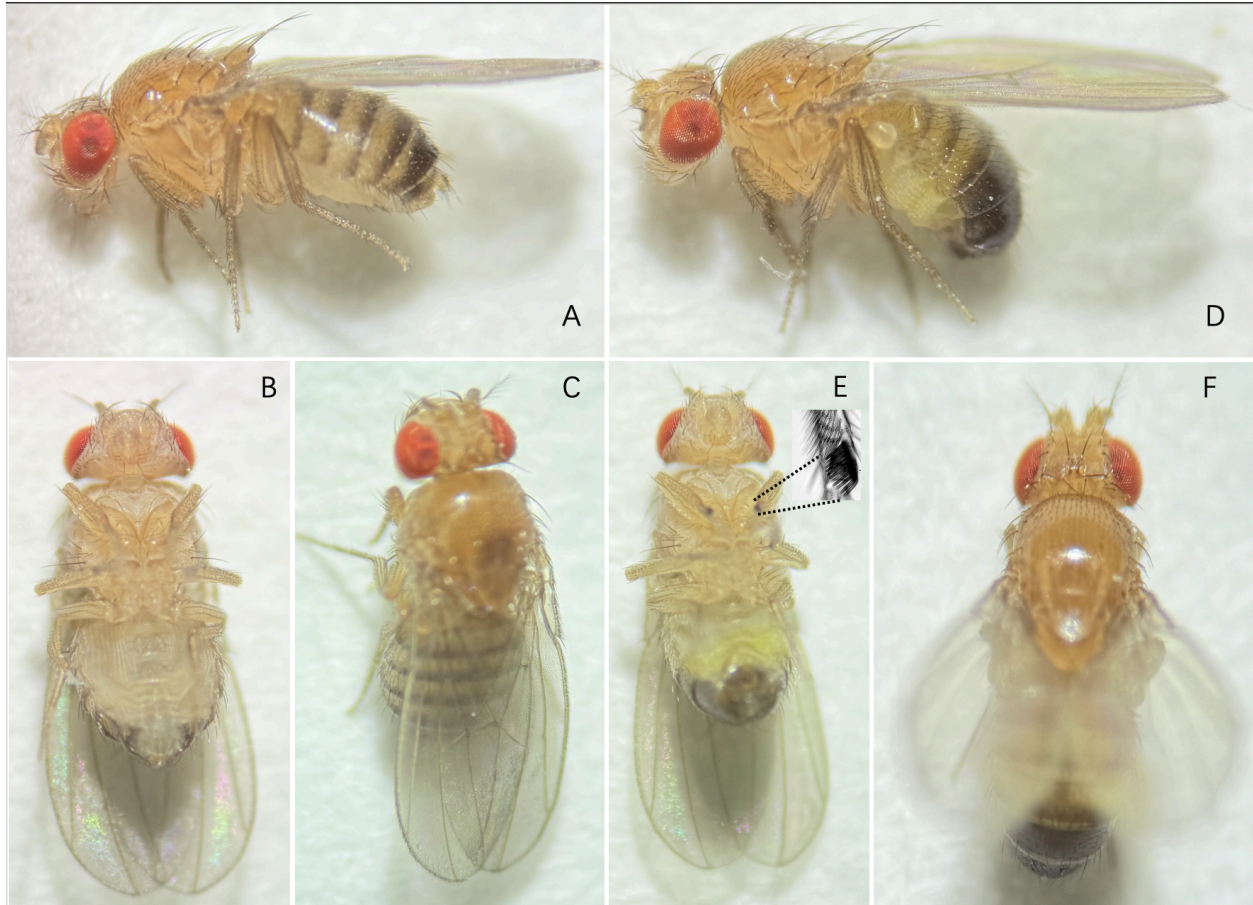


Figure 2: Female and Male Adult *D. melanogaster*. Wild type *D. melanogaster* female (A-C) and male (D-F) flies as seen from lateral, ventral and dorsal views. The inset in E shows a magnified view of the sex comb on the first tarsal segment of the prothoracic pair of legs.

Anatomy of Female and Male Adult Flies: Females (Figure 2B) have a pointed abdomen whereas males (Figure 2E) have a rounded abdomen. Males have greater and more uniform pigmentation on the abdominal tip. Females are typically larger than males, with light-colored, pointed genitalia, which are more rounded and rougher than males. Males characteristically exhibit a small patch of dense bristles on their forelegs, called the sex comb, which aids in grasping the female during mating.

Stereo Microscope: The **stereo** or **dissecting microscope** is designed for low magnification observation of a sample, typically using light reflected from the surface of an object rather than transmitted through it. The instrument includes two separate optical paths with two objectives and eyepieces to provide slightly different viewing angles for the left and right eyes. This arrangement allows three-dimensional visualization of the sample under examination. The stereo microscope is used to study the surfaces of solid specimens or to carry out fine work such as dissection, microsurgery, or inspection.

Lab Exercise I - Separating Male and Female *Drosophila*

Materials

- 1 Fly-Nap
- 6 Wild-type *D. melanogaster* male and female cultures in a vial
- 6 empty vials capped with cotton
- 6 Brushes
- 6 White paper
- 6 Stereo microscope

Methods

1. Work individually on your stereo microscope and become familiar with the operation of these microscopes.
2. You will be provided live “wild-type” cultures containing male and female flies and a piece of white paper to hold flies.
3. Observe the fly culture and identify the different stages of fly development, such as egg, larva, pupa, and adult.
4. Use “Fly-Nap” to anesthetize flies, when 95% flies are anesthetized, transfer them on paper.
5. Observe anesthetized flies under the stereo microscope, compare and separate female from male flies
6. Photograph and record this as part of the lab report

Record on Lab notebook:

1. Take pictures of male and female flies.
2. Describe the difference between female and male flies in writing.
3. In the fly culture provided to you, how many were females and how many were males?

Introduction II

In *Drosophila*, various organs and anatomical structures arise from ten pairs of imaginal discs and the genital disc. The brain of the *Drosophila* larva is composed of two hemispheres and the subesophageal ganglion

The salivary glands of *Drosophila* are the largest secretory organs in the organism. They are tubular organs that are made up of two types of cells: duct cells and secretory cells.

The salivary glands of *Drosophila* produce glue proteins that help attach the pupa to a solid surface before metamorphosis.

Polytene chromosomes are a strongly amplified form of interphase chromosomes, found, for example, in salivary gland cells of *Drosophila*. They are formed when multiple rounds of DNA replication produce many sister chromatids that remain fused together.

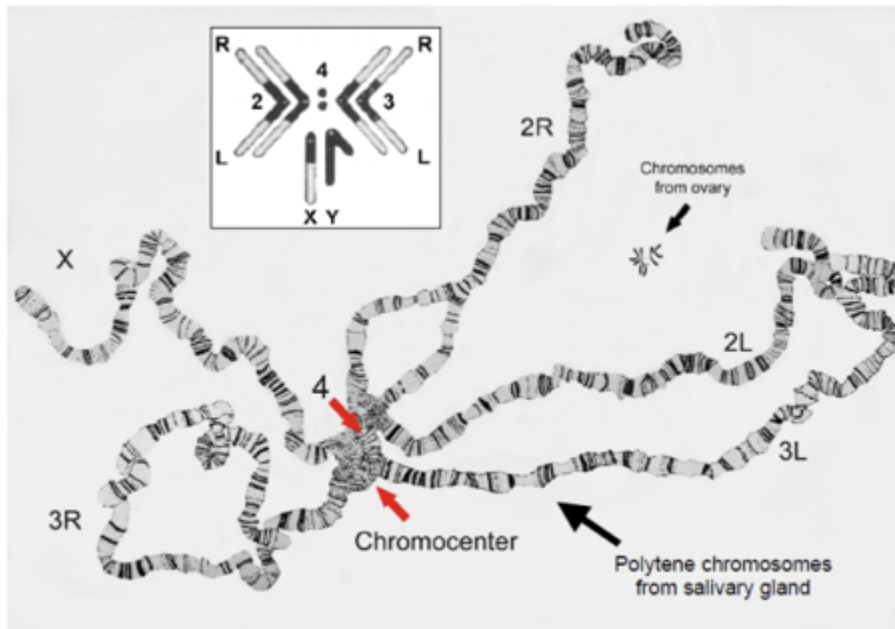


Figure 3 *Drosophila melanogaster* polytene chromosomes (from painter 1934)
Public domain in the U.S. Source: *Journal of Heredity*, 25(11), 465–476.

Polytene chromosomes were first reported by E.G. Balbiani in 1881. Polytene chromosomes are found in dipteran flies: the best understood are those of *Drosophila*, *Chironomus* and *Rhynchosciara*. They are present in another group of arthropods of the class Collembola, a protozoan group Ciliophora, mammalian trophoblasts and antipodal, and suspensor cells in plants. In insects, they are commonly found in the salivary glands when the cells are not dividing.

They are produced when repeated rounds of DNA replication without cell division forms a giant chromosome. Thus polytene chromosomes form when multiple rounds of replication produce many sister chromatids which stay fused together.

Polytene chromosomes, at interphase, are seen to have distinct thick and thin banding patterns. These patterns were originally used to help map chromosomes, identify small chromosome mutations, and in taxonomic identification. They are now used to study the function of genes in transcription.

In addition to increasing the volume of the cells' nuclei and causing cell expansion, polytene cells may also have a metabolic advantage as multiple copies of genes permits a high level of gene expression. In *Drosophila melanogaster*, for example, the chromosomes of the larval salivary glands undergo many rounds of endoreduplication to produce large quantities of adhesive mucoprotein ("glue") before pupation. Another example within the fly itself is the

tandem duplication of various polytene bands located near the centromere of the X chromosome which results in the Bar phenotype of kidney-shaped eyes.

The interbands are involved in the interaction with the active chromatin proteins, nucleosome remodeling, and origin recognition complexes. Their primary functions are: to act as binding sites for RNA pol II, to initiate replication and, to start nucleosome remodeling of short fragments of DNA.

Lab Exercise II Drosophila Larva Brain and Salivary Gland Dissection, Polytene chromosome staining

Materials

- 1 Aceto-alcohol(1:1) in a dropper bottle
- 1 Aceto-orcein stain in a dropper bottle
- 1 45% Acetic acid in a dropper bottle
- 1 box Cover slips
- 1 box of Slides
- 6 Immersion oil
- 6 Bibulous paper
- 6 Black paper
- 6 Stereo microscope
- 6 Compound microscope
- 6 Wild-type D. melanogaster larva in a vial
- 6 Distil water bottles
- 12 Fine-tip forceps

Methods

1. Obtain a black paper, place it under Stereo Microscope, then put a glass slide on the black paper..
2. Use forceps to take a 3rd instar larvae and place it on the center of the slide, then add 1-2 drops of water. During dissection, add water to prevent tissue dry out.
3. Dissect the salivary glands from the Drosophila larvae and take a picture of it.
4. Use forceps to separate the glands and remove any excess tissue.
5. Add 2 drops of aceto-alcohol(1:1) fixative to salivary glands for 2 min.
6. Remove the excess fixative with bibulous paper .
7. Add 1 drop of Aceto-orcein stain over the salivary glands
8. Cover the slide with a petri dish and wait for 10 minutes, then absorb the excess stain with bibulous paper.
9. Add 2 drops of 45% acetic acid over the salivary glands and wait 1 min, then absorb excess acetic acid with bibulous paper.

10. Add 1 drop of Aceto-orcein stain over the salivary glands and cover it with a coverslip.
11. Place the slide between the fold of bibulous paper and apply pressure with the thumb over the coverslip and roll it.
12. Observe under the compound light microscope for polytene chromosomes.
13. Try to find a sharp image under x40 or x100 with immersion oil.
14. Take a picture of the best polytene chromosome you find.

Record on lab notebook and answer the question:

1. Draw a sketch of your best polytene chromosome spread.
2. Are you able to see all five major chromosome arms (X, 2L, 2R, 3L, 3R)?
3. Are you able to see the small 4th chromosome?
4. Can you locate the chromocenter in your spread?
5. Are you able to track along a single chromosome arm from its centromere to telomere?

Assignment 2

Students will read the beginning chapters of [*"Genetics on the Fly: A Primer on the Drosophila Model System"*](#) (Hales et al., 2015) including the chapters titled "Natural History", "Life Cycle", "Food and Husbandry" and "Sex Determination". Download the assignment 2, answer questions in your own words and upload to Blackboard.

1. How did **Drosophila melanogaster** disperse globally, and why is it used to study adaptation to new environments?
2. What makes **Drosophila melanogaster** a powerful model organism for scientific research?
3. What are the key components of **Drosophila media** used in laboratory settings, and how is the firmness of the food adjusted?
4. How does sex determination in **Drosophila melanogaster** differ from that in mammals, especially regarding the role of the Y chromosome?

Lab 3 Classic *Drosophila* Mutations and Genetic tools

Na Xu, Hanqi Ye

Objective: This lab provides basic information about the adult *Drosophila* anatomy, genes, mutations, phenotypes, and fly genetics. Students will familiarize themselves with common visible mutations affecting different parts of *D. melanogaster* and with the genetic basis of these mutant phenotypes.

Introduction

Wild Type *Drosophila*: The wild type is ascribed to the most common phenotype seen in natural populations. Various natural fly populations have been collected from different geographical areas by researchers at different times and maintained in the laboratory as distinct wild type populations since the original time of collection.

Drosophila chromosomes: *D. melanogaster* has four pairs of chromosomes, which include one pair of sex chromosomes (a pair of X chromosomes in females and the X and Y pair in males) and three pairs of autosomes (chromosomes 2, 3, and 4) (Figure 1). The 4th chromosome is very small and often named the 'dot' chromosome (Figure 1). The two arms of the long X, second, and third, chromosomes are designated left (XL, 2L, 3L) and right arms (XR, 2R, 3R) while those of the Y chromosome are designated long (YL) and short (YS) arms (Figure 1).

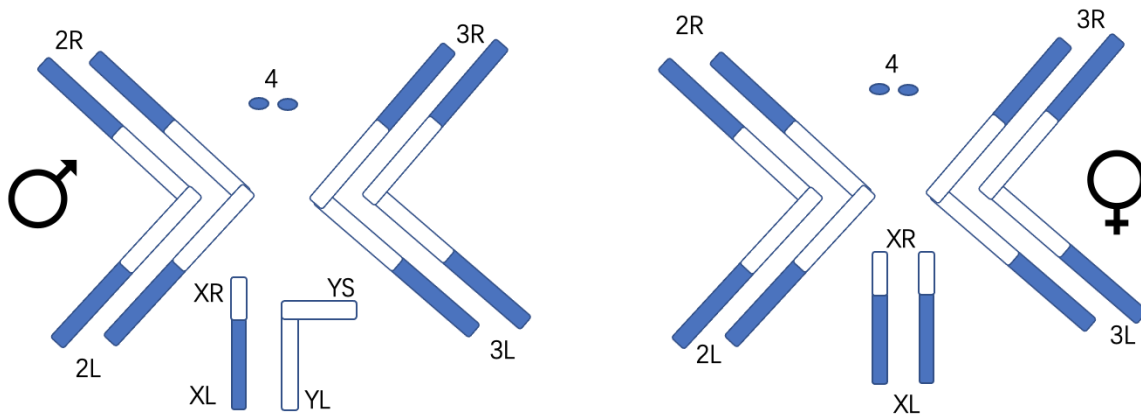


Figure 1: Schematic of Mitotic Chromosomes of *Drosophila melanogaster*(male and female). L and R following the chromosome name (X, 2, and 3) indicate left and right arms, respectively; the very short right arm of the acrocentric X is indistinguishable. YS and YL represent the short and long arms of the Y-chromosome, respectively. The 'dot' chromosome 4 does not show left or right arms. Light and dark-shaded regions represent heterochromatic and euchromatic regions, respectively.

Classic mutants and their phenotypes:

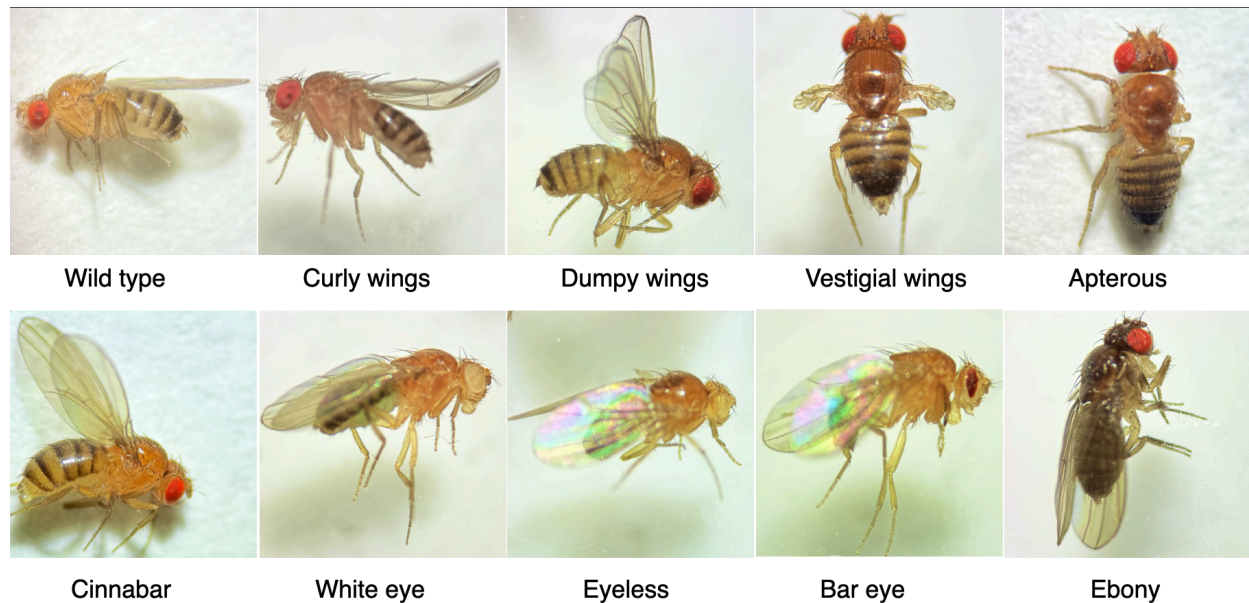


Figure 2: Common Mutations of fruit flies

Vestigial wings mutation: Gene name: *vestigial*; Symbol: *vg*; Affected chromosome: 2; Recessive mutation. In the classical *vg* mutants, the wings of homozygotes are reduced to vestiges and are usually held at right angles to the body; wing veins are visible (Figure 2). Temperatures of 29°C or greater appreciably increase wing size. Flies with vestigial wings cannot fly.

Ebony body mutation: Gene name: *ebony*; Symbol: *e*; Affected chromosome: 3; Recessive mutation. Body color varies from shining black to slightly darker than wild type flies (Figure 2). The ebony gene is responsible for building up the tan-colored pigments in the wild type fruit fly. If the ebony gene is defective, the black pigments accumulate on the entire body.

Dumpy wings mutation: Gene name: *dumpy*; Symbol: *dp*; Affected chromosome: 2; Recessive mutation. *Drosophila* Dumpy is a gigantic extracellular protein required to maintain tension at epidermal–cuticle attachment sites. Flies carrying the recessive viable *dumpy* (*dp*) mutation have shortened wings and altered paths of the wing veins (Figure 2) resulting from contraction of the pupal wing epithelia. This contraction occurs after patterning of the wing has been established and leads to both reduced cell size and altered cell positions. Similarly, vortex-like depressions on the notum (the second Thoracic segment) are observed with certain *dp* alleles (Figure 3).

Flies have a number of serially homologous segments in a tripartite body plan that comprises a Head or Cephalic region, a Thoracic region, and an Abdominal region. Each of the three pairs of legs is located on the ventral side of a different Thoracic segment. The wings are on the dorsal side of **the second Thoracic segment (notum)**. Abdominal segments 5-8 are compressed and modified for reproductive functions. The cephalic region includes segments that differentiate as paired antennae, eyes, and labial palps. All of the cuticular structures develop from imaginal discs present in the fly larva.

Cinnabar Mutants: Gene name: *Cinnabar*; Symbol: *cn*; Affected chromosome: 2; Recessive mutation. *cinnabar* (*cn*) encodes the enzyme that synthesizes 3-hydroxykynurenine (3-HK) as part of the kynurenine pathway of tryptophan degradation. 3-HK is an intermediate product in the formation of ommochromes, the brown pigments in the fly eye, and thus *cn* loss of function mutant flies exhibit altered eye color. Flies homozygous for *cinnabar* mutations have bright red eyes (Figure 2)

Collection of virgins: Many experiments require crosses between males and females of specifically desired genotypes. *Drosophila* females store sperm transferred during the mating in their spermatheca and use these sperm during their entire life. Therefore, selecting unmated virgin females is imperative to set crosses between flies of the desired genotypes. Female flies mate only 4–6 h after eclosion at 25°C. Therefore, segregating males and females in separate vials/bottles within 4 h of emergence ensures that the females remain virgins.

LPCR

ab Exercise I - Observe Mutations and Prepare for Genetic Crosses to Generate Flies Carrying Double Mutations

Materials

- 1 Fly-nap
- 6 Fly cultures with wild-type flies
- 6 vials each with a mutation in *vestigial*, *eyeless*, *bar* or *ebony*
- 6 Stereo microscopes
- 6 new Fly medium vials (made on lab 1)
- 6 Markers
- 6 White paper
- 6 100 ml beakers
- 12 Brushes (6 for dry and 6 for wet)

12 empty bottles capped with sponge
16 fly media bottle (made on lab 1)

Methods

1. Examine and identify mutant characters externally visible and match them with their phenotypes using the wild type (Figure 7) and different mutant flies provided.
2. Record your observations in writing with pictures on a lab notebook.
3. Design genetic crosses to generate a fly carrying double mutations
4. Collect five virgins from one mutant and mate them with five males from another mutant.
5. If not enough adult virgins, wet the brush in the beaker filled up with water, and obtain the female pupae which has no sex comb found on the forelegs.
6. Transfer the cross into a new vial and label the cross and date.
7. Leave the cross at room temperature for two weeks to observe F1 progeny and select flies carrying double mutations.

Introduction II Genetic toolkit for Drosophila

Balancer Chromosomes:

Balancer chromosomes are a special type of genetically modified chromosome used widely in *Drosophila melanogaster* studies. They help maintain recessive lethal (or sterile) mutations in living organisms without being affected by natural selection. These mutations can only survive in heterozygotes, and cannot be stably passed down through generations as they continually produce wild-type organisms. This can be avoided by substituting the wild-type chromosome with a balancer.

Typically, balancer chromosomes are engineered to carry their own recessive lethal mutations, eliminating homozygotes that don't carry the desired mutation. They also suppress meiotic recombination with their homologs, preventing the creation of new wild-type chromosomes. Additionally, they carry dominant genetic markers that can help identify rare recombinants and are useful for screening.

In diploid organisms, mutations without recessive lethal (or sterile) phenotypes can be bred to homozygosity and maintained indefinitely by crossing homozygotes. However, homozygotes for recessive lethal mutations are non-viable by definition, as the presence of the recessive lethal allele on both chromosomal homologs causes the organism to die early in development. An organism that is homozygous for a recessive mutation causing sterility yields a similar result (i.e., its genetic material cannot be passed onto offspring, even if the sterile individual survives to maturity). This issue compels geneticists studying recessive lethal/sterile mutations to maintain the mutation in heterozygous organisms instead.

Crosses between heterozygotes produce wild-type organisms, heterozygotes, and non-viable homozygotes. To maintain a purely heterozygous line, wild-type offspring must be identified and prevented from mating, which can be resource-intensive, especially for long-term maintenance of the recessive mutation.

Replacing a wild-type homolog of the chromosome carrying the recessive mutation with a balancer chromosome prevents the emergence of wild-type organisms. Firstly, a balancer carries its own independent recessive lethal mutation, making the organism non-viable if two balancer copies are inherited (i.e., no copy of the desired mutation). However, recombination between the balancer and the homolog containing the mutated allele may also result in the creation of a wild-type chromosome. To suppress recombination, balancers usually contain multiple, nested chromosomal inversions that disrupt synapsis between the homologous chromosomes.

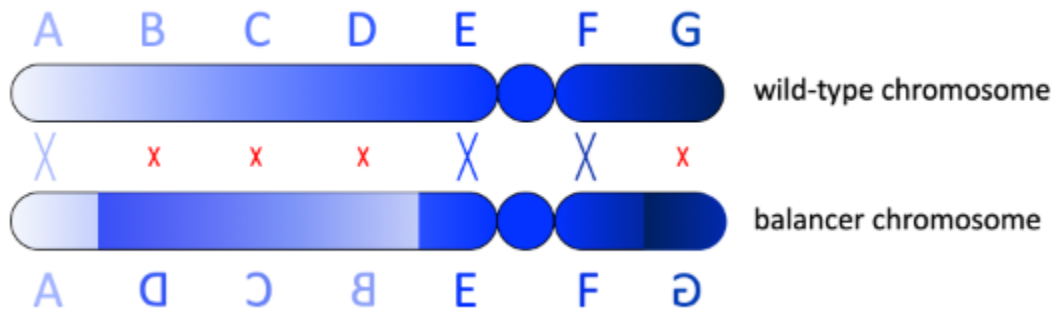


Figure 4: An illustration of a wild-type chromosome and a balancer chromosome. Image by **John Tuthill**, *Wikimedia Commons*, **CC BY-SA 4.0**, created 11 November 2019.

A balancer chromosome contains large inversions ([B,C,D] and [G]). Normal genetic recombination (blue X) is suppressed (red X) at these sites. If crossover happens, it's typically imbalanced, resulting in chromatids that are missing some genes and have duplicates of others. This can also create dicentric or acentric chromosomes (chromosomes with two or no centromeres), which are inherently unstable and often break up, mutate, or get lost during the next mitosis. These outcomes are usually lethal.

Flies with balancer chromosomes can be easily spotted by their genetic marker mutations, such as curly wings or stubble hair. These traits help researchers quickly identify flies carrying the balancer. If recombination occurs and is viable, the marker might disappear, serving as a signal to researchers.

P-element:

P elements, discovered in *Drosophila*, are transposable elements that cause genetic traits known as hybrid dysgenesis. Hybrid dysgenesis refers to the high mutation rate in germ line cells of *Drosophila* strains resulting from a cross of males with autonomous P elements (P Strain/P cytotype) and females that lack P elements (M Strain/M cytotype). The hybrid dysgenesis syndrome is marked by temperature-dependent sterility, elevated mutation rates. These elements, found only in wild flies and other eukaryotes, are responsible for the P trait. The term was first proposed by evolutionary biologist Margaret Kidwell and her colleagues, who studied hybrid dysgenesis in *Drosophila*.

P elements encode an enzyme called P transposase. Wild-type females, unlike lab-bred ones, are believed to also express an inhibitor of P transposase, produced by the same element. This inhibitor minimizes genome disruption caused by P elements, enabling offspring to be fertile. This is evidenced by crosses of lab females (lacking the P transposase inhibitor) with wild-type males (having P elements). Without the inhibitor, P elements can spread across the genome, disrupting many genes and often leading to lethal or sterile offspring.

P elements are frequently used in *Drosophila* genetic experiments as mutagenic agents. One benefit of this method is the ease of locating mutations. In hybrid dysgenesis, one *Drosophila* strain mates with another, producing hybrid offspring and causing chromosomal damage known as dysgenesis. Both parents must contribute to hybrid dysgenesis. For instance, in the P-M system, dysgenesis can occur when the P strain contributes paternally and the M strain maternally. The reverse cross, with an M cytotype father and a P mother, yields normal offspring, as it crosses in a P x P or M x M manner. P male chromosomes can induce dysgenesis when crossed with an M female.

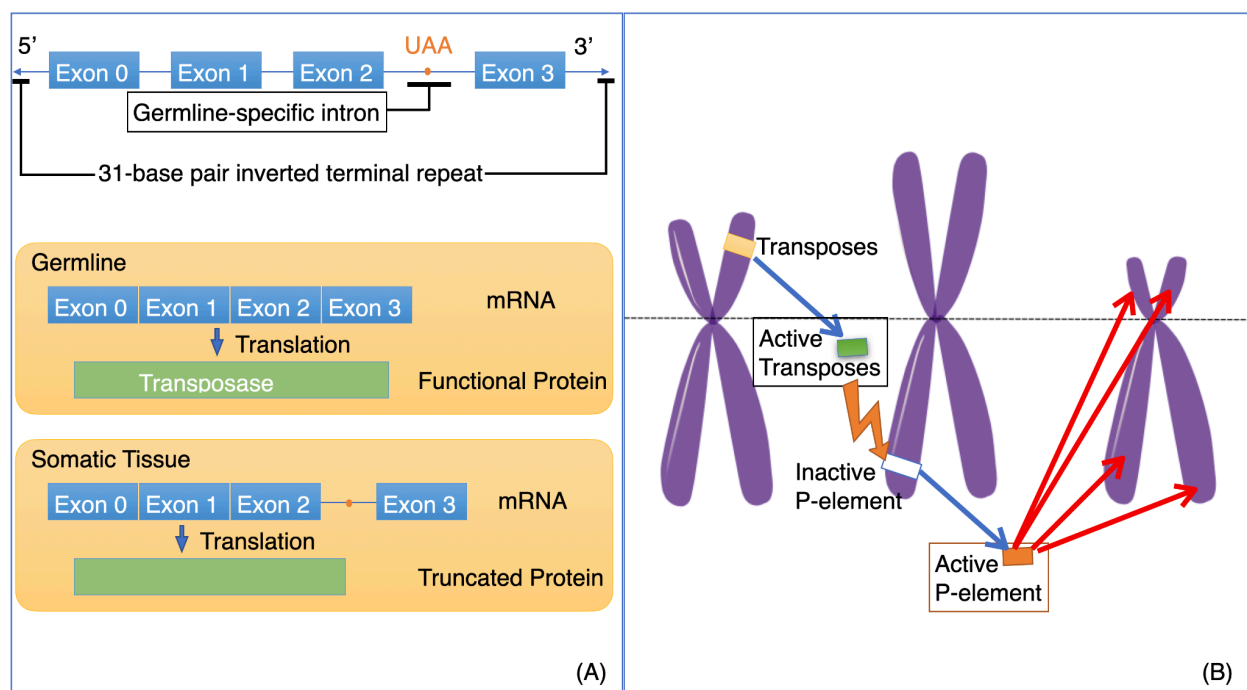


Figure 5 The structure of P-element and P-element-mediated mutagenesis. (A) P-element structure and its tissue-specific alternative splicing. (B) A cartoon showing P-element-mediated mutagenesis.

The P element, a class II transposon, moves via a DNA-based “cut and paste” mechanism. The recognition sequence consists of four exons separated by three introns (Figure 5). Full splicing of the introns produces the transposase enzyme, while alternative partial splicing of introns 1 and 2, leaving only intron 3 in the mRNA transcript, encodes the P element repressor. The complete, autonomous P element encodes a transposase enzyme, which recognizes the 31-bp terminal inverted repeats at either end of the P element and catalyzes P element excision and re-insertion. Non-autonomous P elements contain an internal deletion of varying length which abolishes transposase production, but such elements can still be mobilized if a functional transposase is encoded elsewhere in the genome. P element insertion and subsequent excision necessarily leaves behind 8-bp direct repeats at the excision site; thus the presence of such repeats is indicative of previous P element activity.

All P elements have a canonical structure containing 31-bp terminal inverted repeats and 11-bp internal inverted repeats located at the THAP domain of the transposase. The shortest and longest P elements are non-autonomous elements. The longest P elements encode transposase needed for transposition. The same sequence that encodes the transposase also encodes a suppressor of transposition, which accumulates in the cytoplasm during the development of cells. Thus, in a cross of a P or M male with a P female, the female cytoplasm contains the suppressor, which binds to any P elements and prevents their transposition.

Gal4/UAS system:

The GAL4-UAS system is a biochemical technique used to investigate gene expression and function in organisms like the fruit fly. It was discovered by Hitoshi Kakidani and Mark Ptashne, and Nicholas Webster and Pierre Chambon in 1988 that the binding of Gal4 to UAS sequences triggers gene expression. This method, introduced into flies by Andrea Brand and Norbert Perrimon in 1993, is a powerful tool for studying gene expression. The system consists of two parts: the Gal4 gene, which encodes the yeast transcription activator protein Gal4, and the UAS (Upstream Activation Sequence), an enhancer that GAL4 specifically binds to activate gene transcription.

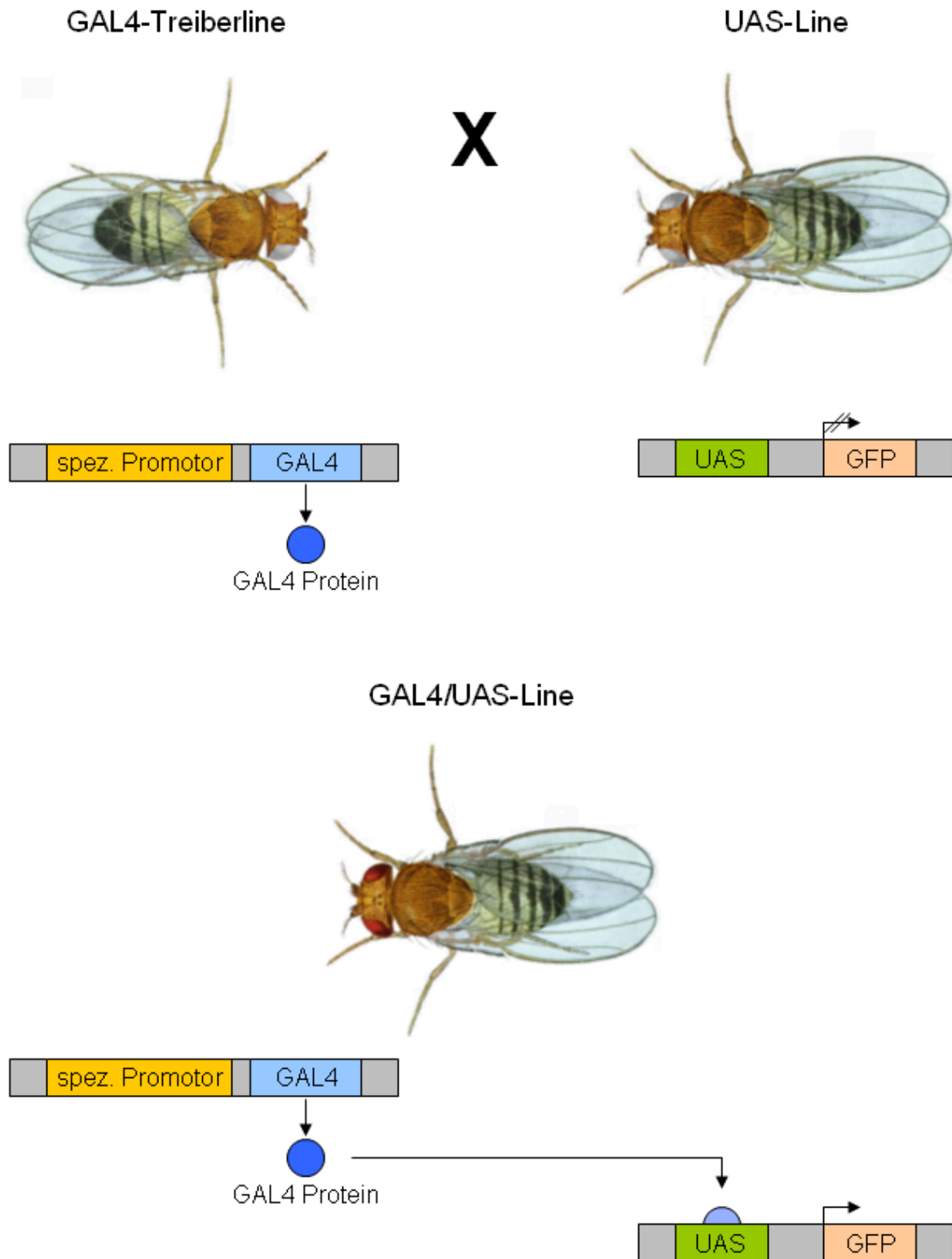


Figure 6 An example GAL4-UAS system, with GAL4 lines and UAS reporter lines. (Image by **Johannes82**, Wikimedia Commons, **CC BY-SA 3.0**.)

The Gal4 system separates the challenges of identifying which cells express a gene or protein and what the researcher wants to do with this information. Geneticists have created genetic variants of model organisms (usually fruit flies), known as GAL4 lines, each expressing GAL4 in a subset of the animal's tissues. For instance, some lines might express GAL4 only in muscle cells, nerves, or antennae. There are tens of thousands of such lines in fruit flies, with the most useful ones expressing GAL4 in a very specific subset of the animal. The presence of GAL4 in these cells has little or no effect, as GAL4's main function is to bind to a UAS region, and most cells have no (or harmless) UAS regions (Figure 6).

Since Gal4 is not visible and has little effect on cells, "reporter lines" are also necessary. These are fly strains with the special UAS region next to a desired gene. These genetic instructions are present in every cell of the animal, but in most cells nothing happens since that cell is not producing GAL4. However, in the cells that are producing GAL4, the UAS is activated, the adjacent gene is turned on, and it starts producing its resulting protein. This may inform the investigator which cells are expressing GAL4, hence the term "reporter line", but genes intended to manipulate cell behavior are often used as well.

Common reporter genes include:

- Green (GFP) or red fluorescent proteins (RFP), which allow scientists to identify which cells express Gal4
- Channelrhodopsin, which enables light-sensitive activation of nerve cells
- Halorhodopsin, which allows light to inhibit the firing of neurons
- Shibire, which deactivates neurons, but only at higher temperatures (30 °C and above). Flies with this gene can be raised and tested at lower temperatures where their neurons will behave normally. Then the body temperature of the flies can be raised (since they are cold-blooded), and these neurons turn off. If the fly's behavior changes, this provides a strong clue to what those neurons do.
- GECI (Genetically Encoded Calcium Indicator), often a member of the GCaMP family of proteins. These proteins fluoresce when exposed to calcium, which, in most cells, happens when the neuron fires. This allows scientists to take pictures, or movies, that show the nervous system in operation.

For instance, scientists can first visualize a class of neurons by selecting a fly from a GAL4 line that expresses GAL4 in the desired set of neurons, and crossing it with a reporter line that expresses GFP. In the offspring, the desired subset of cells will produce GAL4, and in these cells the GAL4 will bind to the UAS, and enable the production of GFP. So the desired subset of cells will now fluoresce green and can be tracked with a fluorescence microscope. Next, to determine what these cells might do, the experimenter might express channelrhodopsin in each of these cells, by crossing the same GAL4 line with a channelrhodopsin reporter line. In the offspring the selected cells, and only those cells, will contain a channel.

Record on Lab notebook:

1. Students will describe which mutant flies were observed and their features compared to the wild-type flies.
2. Students should include pictures of the wild-type and mutant flies.

Biotechnology and Society Module I: Influenza

Viruses are basically parasites of our cells. The purpose of a virus is to find a host cell, and then hijack the metabolic machinery of that cell to produce more copies of itself. Viruses need a host cell in order to replicate. Viruses vary in their shape and in the types of cells they infect.

Despite their differences, all viruses have the following two components: **Genetic material** - The genetic material, which may be in the form of DNA or RNA, contains a relatively small number of genes in comparison to a cell. **Capsid** - A protein coat that surrounds the genetic material. The capsid may have proteins attached to it, such as tail fibers or spikes, that assist the virus in attaching to the target host cell.

It is important to recognize that there are several different forms of influenza: influenza A, B, C, and D. Of these, *influenza A* and *influenza B* are the two most prevalent forms of the virus that infect humans.

Influenza viruses are typically named by the *World Health Organization (WHO)* according to their host (in case it is not human), type (A or B), strain, and year of isolation. If the virus is *influenza A*, this is followed by a nomenclature on the types of **spikes** found on the surface of the virus. *Influenza B* viruses have a lineage designation (either B/Yamagata or B/Victoria) that helps identify the history of the virus. On the surface of an influenza A virus are two different types of **spikes**. These spikes are a type of glycoprotein (a protein with a sugar attached) that assists the virus in identifying and entering its host cell.

For influenza A viruses, there are two important forms of these spikes: **H spikes and N spikes**. H spikes are associated with hemagglutinin, a type of glycoprotein that assists the virus in identifying the receptors on the host cell. The N spikes are an enzyme called neuraminidase. N spikes help break down the mucous material surrounding host cells in the respiratory tract and initiate penetration of the virus into the host cell. These spikes are not the same in all influenza A viruses. There are currently 17 different known forms of H spikes, and 10 different variations in N spikes. These variations are what give influenza A viruses their distinct name. In humans, H1N1 and H3N2 are the two forms of influenza A that commonly contribute to the normal seasonal flu outbreak. So, for example, the H1N1 virus has hemagglutinin form 1 and neuraminidase form 1. As we will see, these spikes also play a role in identifying how the influenza virus is evolving over time.

Most viruses are very specific in the types of cells they infect. This specificity is the function of the proteins located on the capsid of the virus. Proteins often have complex, three-dimensional

shapes, and in the case of the capsid proteins, these complex shapes allow the virus to target specific molecules on the surface of the target cell. These molecules may be other proteins, receptors on the surface of the cell, or other molecules integrated into the membrane of the host cell. Once the virus has entered the target cell, there are two pathways, or cycles, it can follow: It can either immediately begin using the resources of the host cell to manufacture new viruses (**the lytic cycle**), or it can integrate directly into the host cell genome and delay its own reproduction until a later time (**the lysogenic cycle**).

The lytic cycle consists of five stages:

1. **Attachment**. As we previously discussed, the structure of the proteins in the capsid allows it to recognize and target specific cells for infection.

2. **Penetration**. Once the virus has identified the host, its genetic material enters the cell. This can occur by injecting the genetic material directly into the host cell (leaving the capsid outside), or by convincing the host cell to engulf the entire virus (a process called endocytosis). In the case of influenza, the latter process occurs and once in the cell the virus releases its genetic material into the host cell.

3. **Biosynthesis**. At this point, the virus hijacks the host cell's machinery to manufacture new capsid proteins and copies of its genetic material. Energy and materials for these processes are all provided by the host cell.

4. **Maturation**. Also known as assembly, this is when the pieces of the virus are assembled.

5. **Release**. Finally, the new viruses are ready to leave the cell. This is possible by either causing the host cell to burst (called lysis) or by exiting through the plasma membrane (exocytosis). Viruses that possess an outer envelope leave by endocytosis. As they exit, they often create an envelope about themselves by acquiring pieces of the host cell's plasma membrane.

In the **lysogenic cycle**, following penetration into the cell, the viral DNA integrates into the host cell's DNA. Unlike the lytic cycle, the virus remains *inactive* (or latent).

As the host cell divides, the viral DNA is replicated along with the host cell DNA, thus creating new cells that contain the virus. The virus remains latent until some environmental factor signals it to re-enter the lytic cycle and begin the process of biosynthesis and maturation.

HIV is an example of a virus in humans that uses the lysogenic cycle. Scientists have estimated that almost 8% of the human genome is actually viral in nature, and thus evidence of how active the lysogenic cycle has been in our evolutionary history. HIV, influenza A, and SARS-CoV-2 are some examples of RNA viruses. This presents a problem for the virus, because the host cell's protein synthesis processes require DNA as the starting template.

RNA viruses can contain enzymes that convert the viral RNA into DNA, as in HIV, or convert the viral RNA into mRNA, as in the influenza virus. The primary targets of the influenza virus are the epithelial cells lining the respiratory system. **Epithelial cells** form the boundary between the tissues of our bodies and internal environments, such as the respiratory tract and digestive tract. Influenza viruses enter the respiratory system through either the mouth or nose. Often, viruses are picked up on the hands and introduced to the body when we touch our faces or

mouth or handle food. From here, they quickly enter into the respiratory system and travel down the pharynx into the trachea and lungs.

Most other flu symptoms are not the direct result of the influenza virus, but are, as we will see, due to the body's response to the virus—that is, activation of the immune system.

This is due primarily to chemical signals released by the immune system in response to an infection. The white blood cells of the immune system release proteins called cytokines. Cytokines act as an alarm to the organ systems of the body, alerting them that a pathogen is present.

Vaccinations contain either inactivated or weakened viruses, plus additives to preserve the vaccine's contents before use. The most common components are listed in the table. The addition of aluminum salts, formaldehyde, and thimerosal often causes some concern among people. However, these ingredients are all present in very low levels and are usually quickly dispersed by the body. For example, while formaldehyde is a chemical that is toxic at high levels; it is also naturally present in the tissues of our bodies as a by-product of our normal metabolism. The levels of formaldehyde in a vaccine are lower than those naturally found in the tissues of our bodies.

Two different forms of vaccine are used each year. The most common vaccine contains three inactivated viruses. This is called a trivalent vaccine. The most common method of producing an influenza vaccine has been to inject the viruses into fertilized chicken eggs and allow them to incubate for a few days, during which time the viruses begin to replicate. The liquid portion of the infected egg is then extracted and the viruses in it are killed. The virus particles are then isolated, purified, and used to make the vaccines. While this process is effective in producing vaccines (and has been used for decades), it requires the longest development time.

The newest process is called a **recombinant process** because it only uses a small portion of the influenza virus. In the case of influenza, the portion of the virus used is a protein associated with the H spikes, because this is what is identified by our immune systems. This protein is added to a virus that infects insect cells and is then cultured to produce large amounts of the H spike protein. The complete virus is not produced, and the extracted H spike protein is used to make the viruses. This is the most rapid method of producing a vaccine and also allows for the production of egg-free vaccines.

For further information regarding alternative vaccine production direct students' attention to the company Medicago (<http://medicago.com/>) which produces the flu vaccine using tobacco plants.

The idea Medicago's scientists have will make more of the vaccine at a faster rate. Additionally, the plants and vaccine grow within 5-6 weeks as opposed to the 5-6 months required to prepare the vaccine in chicken eggs.

Videos:

- Attenuation: How Scientists Make Live Vaccines:

<https://www.youtube.com/watch?v=szfVMtyueks>

- Using Genetic Engineering to Make Vaccines

<https://www.youtube.com/watch?v=DgDDw2gQJx0>

- RNA Vaccines (mRNA Vaccine) - Basis of Pfizer and Moderna COVID-19 vaccines, Animation

<https://www.youtube.com/watch?v=oMXGGmBfkf8>

By studying and classifying life on this planet, scientists have been able to recognize that the organization of life provides evidence of change over time. The question has not been is **evolution** actually happening, but rather, what is the mechanism of evolutionary change?

Viruses, specifically the influenza virus, are an excellent example of how evolution occurs over time. Human influenza viruses originate in animal hosts, such as the pig and birds. As the viruses exchange genetic material, they evolve into new forms that can then be transmitted to humans, causing flu outbreaks.

Natural selection was provided independently by both Charles Darwin and Alfred Russel Wallace. Their explanation of evolutionary processes forever changed the science of biology.

Individuals vary in their traits. Variation exists in a population for heritable traits. This does not just include physical characteristics, such as beak shape, but also physiology and behavior.

Individuals struggle to exist. Populations produce more offspring than can survive, which creates competition for available resources.

Individuals differ in their fitness. Individuals who have favorable traits will survive at a greater rate than those without them. When these individuals reproduce, they pass these traits on to the next generation. The term fitness is often used to describe this reproductive success.

Populations become adapted to the environment. Over time, natural selection increases the frequency of the advantageous trait in the population and decreases the frequency of detrimental traits. Over multiple generations, the population becomes adapted to the new environment.

Darwin proposed that **natural selection** allows populations to adapt to a new environment, and is the basis of the process of evolution.

Each year, 3 to 5 million people are infected with influenza worldwide resulting in an average of 500,000 deaths. In 1918, a variant of the H1N1 influenza virus (the "Spanish flu") infected over 500 million people across the globe, or one-third of the world's population at that time. The death toll worldwide was estimated to be between 30 and 50 million.

There are two basic methods by which the influenza virus changes its genetic information: antigenic *shift* and *antigenic drift*: **Antigenic drift** occurs when there are small changes in the surface spikes of the influenza virus. The second mechanism the influenza virus has of evading

the immune system—**antigenic shift**—has greater consequences for humans. In an antigenic shift, two different forms of the virus infect the same cell.

Our immune system has several levels of defenses, from nonspecific physical barriers to cellular mechanisms of protecting us from specific pathogens, such as the influenza virus.

The defenses that act as general defenses, meaning they are not targeting a specific pathogen, are collectively called the **innate defenses**. **Adaptive immunity** targets specific antigens associated with a pathogen and also has the ability to remember past responses. Adaptive immunity has the ability to target cells of the body that are infected with the pathogen, while also responding to any pathogens that are free in the blood or interstitial fluid of the body.

Cell-mediated immunity targets cells of our bodies that have become infected with pathogens and are presenting the antigen. Cell-mediated immunity makes use of T lymphocytes, which are white blood cells produced in the bone marrow. After being formed, T lymphocytes travel to the thymus where they mature. As they mature, each T lymphocyte develops receptors that will allow it to recognize a single form of antigen.

When the influenza virus enters into an epithelial cell on the surface of our respiratory tract, the infected cell responds by posting an antigen on its surface, attached to an MHC-I marker. This acts as a flag for the cytotoxic T cells, which were just activated by the helper T cells. At the same time that helper T cells and cytotoxic T cells are working to target infected cells, the B lymphocytes begin the production of antibodies. This is the second form of adaptive immunity and is often referred to as **antibody-mediated immunity**.

The adaptive immune response also has the ability to remember past exposures to pathogens. This characteristic is called **memory**. Some of the activated B cells are held in reserve and not used in the first response to the antigen. Instead, they wait for future exposure.

This is useful because if you are exposed to the same pathogen again—for example, the same influenza virus—the adaptive response can be much faster and more efficient. Memory T cells are also produced by the cell-mediated immunity pathway. Scientists are researching ways of getting around the ability of the virus to evolve, including the use of antiviral agents and the development of a universal flu vaccine.

Assignment 3

Students will answer 2 questions based on the mutant flies they chose:

1. What are the guidelines for selecting parent flies when setting up a genetic cross to obtain a double mutant?
2. Draw 2 Punnett squares to demonstrate genotypes and phenotype ratio for F1 and F2 respectively.

Lab 4 Classic Drosophila Mutations and Genetic tools

Na Xu, Hanqi Ye

Objective:

1. Understand the principles of Drosophila genetics and its significance in studying pattern formation.
2. Explore the role of fluorescence microscopy in visualizing gene expression patterns and protein localization in Drosophila embryos.
3. Develop practical skills in fluorescence microscopy techniques, including sample preparation, imaging.

Introduction

It is the fly *Drosophila melanogaster*, more than any other organism, that has transformed our understanding of how genes govern the patterning of the body. The discovery of pattern formation genes and the subsequent analysis of their functions was a famous tour-de-force of developmental genetics. It has had a revolutionary impact on all of developmental biology by showing the way toward a systematic, comprehensive account of the genetic control of embryonic development. In this session, students will review the following book chapter and understand the discovery and the fundamental roles of pattern formation genes in flies:

- The Insect Body Is Constructed as a series of Segmental Units (Alberts et al., 2002)

The period of **embryonic development** begins at **fertilization** and takes about a day, at the end of which the embryo hatches out of the **eggshell** to become a **larva**. The larva then passes through three stages, or **instars**, separated by molts in which it sheds its old coat of cuticle and lays down a larger one. At the end of the third instar, it pupates. Inside the **pupa**, a radical remodeling of the body takes place—a process called **metamorphosis**. Eventually, about nine days after fertilization, an **adult fly, or imago**, emerges.

The fly consists of a **head, with mouth, eyes, and antennae**, followed by **three thoracic segments** (numbered T1 to T3), and **eight or nine abdominal segments** (numbered A1 to A9). Each segment, although different from the others, is built according to a similar plan. Segment T1, for example, carries a pair of legs, T2 carries a pair of legs plus a pair of wings, and T3 carries a pair of legs plus a pair of halteres—small knob-shaped balancers important in flight, evolved from the second pair of wings that more primitive insects possess. The quasi-repetitive segmentation develops in the early embryo during the first few hours after fertilization, but it is more obvious in the larva, where the segments look more similar than in the adult. In the embryo, it can be seen that the rudiments of the head, or at least the future adult mouthparts, are likewise segmental. At the two ends of the animal, however, there are highly specialized terminal structures that are not segmentally derived.

- *Drosophila* Begins Its Development as a Syncytium (Alberts et al., 2002)

The **egg** of *Drosophila* is about 0.5 mm long and 0.15 mm in diameter, with a clearly defined polarity. Like the eggs of other insects, but unlike vertebrates, it begins its development in an unusual way: a series of nuclear divisions without cell division creates a **syncytium**. The early nuclear divisions are synchronous and extremely rapid, occurring about every 8 minutes. The first nine divisions generate a cloud of nuclei, most of which migrate from the middle of the egg toward the surface, where they form a monolayer called the **syncytial blastoderm**. After another four rounds of nuclear division, plasma membranes grow inward from the egg surface to enclose each nucleus, thereby converting the **syncytial blastoderm** into a **cellular blastoderm** consisting of about 6000 separate cells. About 15 of the nuclei populating the extreme posterior end of the egg are segregated into cells a few cycles earlier; these **pole cells** are the **germ-line precursors (primordial germ cells)** that will give rise to eggs or sperm.

Up to the cellular blastoderm stage, **development** depends largely—although not exclusively—on stocks of **maternal mRNA and protein** that accumulated in the egg before fertilization. The frantic rate of DNA replication and nuclear division evidently gives little opportunity for

transcription. After cellularization, **cell division** continues in a more conventional way, asynchronously and at a slower rate, and the rate of transcription increases dramatically. **Gastrulation** begins a little while before cellularization is complete when parts of the sheet of cells forming the exterior of the embryo start to tuck into the interior to form the gut, the musculature, and associated internal tissues. A little later and in another region of the embryo, a separate set of cells move from the surface epithelium into the interior to form the central nervous system. By marking and following the cells through these various movements, one can draw **a fate map for the monolayer of cells on the surface of the blastoderm**.

- Genetic Screens Define Groups of Genes Required for Specific Aspects of Early Patterning (Alberts et al., 2002)

Sometimes the developmental functions revealed by mutant phenotypes are those that one would expect; sometimes they are a surprise. A large-scale **genetic screen** focusing on early ***Drosophila* development** revealed that the key genes fall into a relatively small set of functional classes defined by their mutant phenotypes. Some—**the egg-polarity genes**—are required to **define the anteroposterior and dorsoventral axes** of the embryo and mark out its two ends for special fates, by mechanisms involving interactions between the oocyte and surrounding cells in the ovary. Others, the **gap genes**, are required in specific broad regions along the anteroposterior axis of the early embryo to allow their proper development. A third category, **the pair-rule genes**, is required, more surprisingly, for the development of alternate body segments. A fourth category, the **segment polarity genes**, is responsible for organizing the anteroposterior pattern of each individual segment.

In the stages before fertilization, **the anteroposterior axis** of the future embryo becomes defined by three systems of molecules that create landmarks in the oocyte. Following fertilization, each landmark serves as a beacon, providing a signal, in the form of **a morphogen gradient**, that organizes the developmental process in its neighborhood. Two of these signals are generated from localized deposits of specific mRNA molecules. The future **anterior end** of the embryo contains a high concentration of **mRNA for a gene regulatory protein** called **Bicoid**; this mRNA is translated to produce Bicoid protein, which diffuses away from its source to form a concentration gradient with its maximum at the anterior end of the egg.

The future **posterior end** of the embryo contains a high concentration of mRNA for a regulator of translation called **Nanos**, which sets up a posterior gradient in the same way.

The third signal is generated symmetrically at both ends of the egg, by local activation of a transmembrane tyrosine kinase receptor called **Torso**. The activated receptor exerts its effects over a shorter range, marking the sites of **specialized terminal structures** that will form at the head and tail ends of the future larva and defining the rudiments of the future gut.

The three sets of genes responsible for these localized determinants are called the **anterior, posterior, and terminal** sets of **egg-polarity** genes.

A fourth landmark defines **the dorsoventral axis**: a protein that is produced by follicle cells underneath the future ventral region of the embryo that leads to localized activation of another transmembrane receptor, called **Toll**, in the oocyte membrane. The genes required for this function are called dorsoventral egg-polarity genes.

All the egg-polarity genes in these four classes are **maternal-effect genes**: it is the mother's genome, not the zygotic genome, that is critical. Thus, a fly whose chromosomes are mutant in both copies of the *bicoid* gene but who is born from a mother carrying one normal copy of *bicoid* develops perfectly normally, without any defects in the head pattern. However, if that daughter flies is a female no functional *bicoid* mRNA can be deposited into the anterior part of her own eggs, and all of these will develop into headless embryos regardless of the father's genotype.

- Three Classes of Segmentation Genes Refine the Anterior- Posterior Maternal Pattern and Subdivide the Embryo

After the initial gradients of **Bicoid** and **Nanos** are created to define the **anteroposterior axis**, the **segmentation genes** refine the pattern. Mutations in any one of the segmentation genes alter the number of segments or their basic internal organization without affecting the global polarity of the embryo. Segmentation genes are expressed by subsets of cells in the embryo, so their products are the first components that the **embryo's own genome**, rather than the maternal genome, contributes to embryonic development. They are therefore called **zygotic-effect genes** to distinguish them from the earlier **maternal-effect genes**.

The segmentation genes fall into three groups according to their mutant phenotypes and the stages at which they act. First comes a set of at least six **gap genes**, whose products mark out coarse subdivisions of the embryo. Mutations in a gap gene eliminate one or more groups of adjacent segments, and mutations in different gap genes cause different but partially overlapping defects. In the mutant **Krüppel**, for example, the larva lacks eight segments, from T1 to A5 inclusive.

The next segmentation genes to act are a set of eight **pair-rule genes**. Mutations in these cause a series of deletions affecting alternate segments, leaving the embryo with only half as many segments as usual. The pair-rule mutant **even-skipped (eve)**, for example, lacks the whole of each odd-numbered parasegment, while the pair-rule mutant **fushi tarazu (ftz)** lacks the whole of each even-numbered parasegment, and the pair-rule mutant **hairy** lacks a series of regions that are of similar width but out of register with the parasegmental units.

Finally, there are at least 10 **segment-polarity genes**. Mutations in these genes produce larvae with a normal number of segments but with a part of each segment deleted and replaced by a mirror-image duplicate of all or part of the rest of the segment. In **gooseberry mutants**, for

example, the posterior half of each segment is replaced by an approximate mirror image of the adjacent anterior half-segment. The phenotypes of the various segmentation mutants suggest that the segmentation genes form a coordinated system that subdivides the embryo progressively into smaller and smaller domains along the anteroposterior axis, distinguished by different patterns of gene expression. Molecular genetics has helped to reveal how this system works.

Videos:

- <https://www.ibiology.org/speakers/eric-wieschaus/>
- <https://www.ibiology.org/speakers/christiane-nusslein-volhard/>
- Collaborating to Find Developmental Genes:
<https://www.ibiology.org/development-and-stem-cells/developmental-genes/>
- Finding Genes That Control Development:
<https://www.ibiology.org/development-and-stem-cells/embryonic-development/>

Introduction of fluorescence microscopy

A fluorescence microscope is a type of optical microscope that uses fluorescence, in addition to or instead of scattering, reflection, and attenuation or absorption, to examine the properties of organic or inorganic substances. The term “fluorescence microscope” applies to any microscope that uses fluorescence to create an image, whether it’s a simple setup like an epifluorescence microscope or a more complex design like a confocal microscope, which uses optical sectioning to improve the resolution of the fluorescence image.

Principle: the specimen is lit with light of a specific wavelength (or wavelengths) which is absorbed by the fluorophores, causing them to emit light of longer wavelengths (i.e., a different color than the absorbed light). The illumination light is separated from the much weaker emitted fluorescence using a spectral emission filter. Common components of a fluorescence microscope include a light source (xenon arc lamp or mercury-vapor lamp are common; more advanced versions use high-power LEDs and lasers), the excitation filter, the dichroic mirror (or dichroic beamsplitter), and the emission filter. The filters and the dichroic beamsplitter are selected to align with the spectral excitation and emission characteristics of the fluorophore used to label the specimen. In this way, the distribution of a single fluorophore (color) is imaged at a time. Multi-color images of several types of fluorophores are created by combining several single-color images.

Most fluorescence microscopes in use today are epifluorescence microscopes, where the excitation of the fluorophore and the detection of the fluorescence occur through the same light path (i.e., through the objective). These microscopes are extensively used in

biology and form the foundation for more advanced microscope designs, such as the confocal microscope and the total internal reflection fluorescence microscope (TIRF).

Fluorescence Microscopy

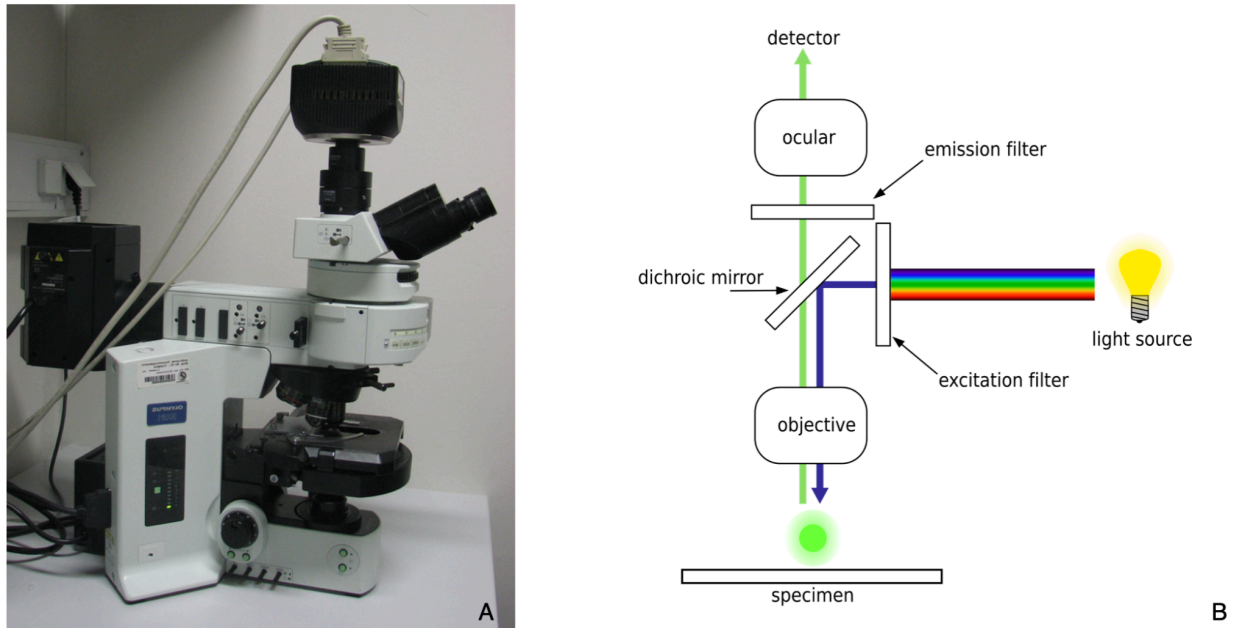


Figure 1 (A) Typical components of a fluorescence microscope (Olympus-BX61 fluorescence microscope, Author: Masur, CC BY-SA 3.0, via Wikimedia Commons) and (B) diagram (FluorescenceFilters 2008-09-28.svg, Author: Henry Mühlpfordt / Krzysztof Blachnicki, CC BY-SA 3.0, via Wikimedia Commons.)

1. Fluorescent Dyes (Fluorophore): A fluorophore is a chemical compound that exhibits fluorescence, meaning it can absorb light and then re-emit it. These compounds often contain multiple aromatic groups or planar or cyclic molecules with several π bonds. A variety of fluorescent stains have been developed for different biological molecules. Some of these stains are small molecules that naturally exhibit fluorescence and can bind to a specific biological molecule. Key examples include nucleic acid stains like DAPI and Hoechst, and phalloidin, which is used to stain actin fibers in mammalian cells.

2. Light Source: There are four main types of light sources used, including xenon arc lamps, mercury-vapor lamps with an excitation filter, lasers, and high-power LEDs. Lasers are primarily used for advanced fluorescence microscopy techniques, while xenon lamps, mercury lamps, and LEDs with a dichroic excitation filter are typically used for wide-field epifluorescence microscopes.

3.Excitation Filter: The exciter is usually a bandpass filter that only allows wavelengths absorbed by the fluorophore to pass through. This minimizes the excitation of other fluorescence sources and blocks excitation light in the fluorescence emission band.

4.Dichroic Mirror: A dichroic filter, or thin-film filter, is a highly precise color filter that selectively allows light of a narrow color range to pass through while reflecting other colors.

5.Emission Filter: The emitter is typically a bandpass filter that only allows wavelengths emitted by the fluorophore to pass through, blocking all undesired light outside this band, especially the excitation light. By blocking unwanted excitation energy (including UV and IR) or autofluorescence from the sample and system, optical filters ensure the darkest background.

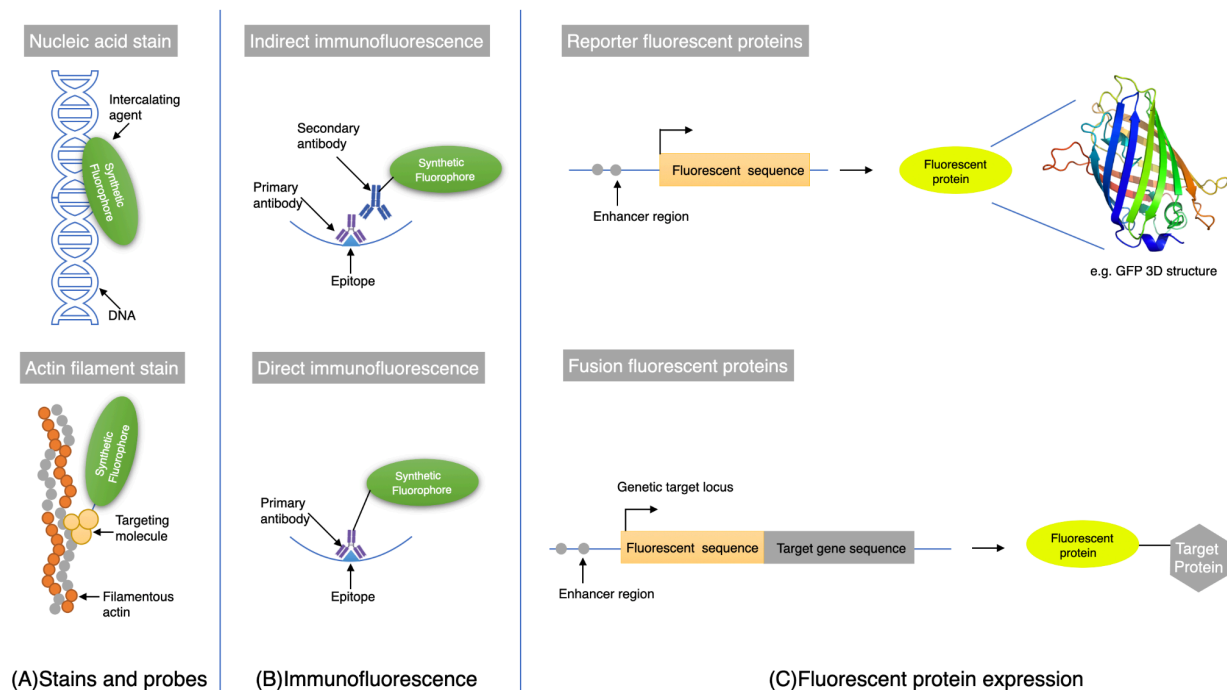


Figure 2 Fluorescent labeling techniques. The great diversity of synthetic fluorophores facilitates the visualization of cellular components by different fluorophore-based labeling techniques. (A) Fluorescent stains and probes directly interact with cellular components (e.g., by intercalation into the DNA double helix) or may be fused to a targeting molecule (e.g., phalloxins binding to actin filaments). (B) In immunofluorescence, synthetic fluorophores are coupled to immunoglobulins (e.g., IgG and IgM) that selectively bind to specific antigens of target proteins. Synthetic fluorophores may be coupled to a primary antibody that directly interacts with the target protein (direct immunofluorescence). An unlabeled primary antibody

can also serve as a platform for binding of multiple secondary antibodies that are coupled to synthetic fluorophores, thereby amplifying the signal intensity (indirect immunofluorescence). (C) The endogenous expression of fluorescent proteins (FPs) provides a genetically encoded avenue for the visualization of cellular components. FPs may be expressed under the control of regulatory elements (i.e., reporter FPs) or fused in-frame with the genetic sequence of a protein-coding gene to create a tagged version of the target protein (i.e., fusion FPs). ("GFP structure." Image by Richard Wheeler (User: Zephyris), licensed under Creative Commons Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0). Via Wikimedia Commons: https://commons.wikimedia.org/wiki/File:GFP_structure.png)

Sample preparation:

For a sample to be compatible with fluorescence microscopy, it needs to exhibit fluorescence. There are multiple ways to achieve this; the primary methods include labeling with fluorescent stains or, for biological samples, expressing a fluorescent protein. Another option is to utilize the sample's inherent fluorescence, also known as autofluorescence. In life sciences, fluorescence microscopy serves as a potent tool that enables specific and sensitive staining of a specimen, facilitating the detection of protein distribution or other molecules of interest. Consequently, there is a wide array of techniques available for the fluorescent staining of biological samples.

Biological Fluorescent Stains: Numerous fluorescent stains have been developed for a variety of biological molecules. Some of these are small molecules that naturally exhibit fluorescence and can bind to a specific biological molecule. Key examples include nucleic acid stains such as DAPI and Hoechst, which are excited by UV light, and DRAQ5 and DRAQ7, which are optimally excited by red light. All these stains bind to the minor groove of DNA, thereby labeling the cell nuclei. Other stains are drugs, toxins, or peptides that bind to specific cellular structures and have been modified with a fluorescent reporter. A prime example of this type of fluorescent stain is phalloidin, used to stain actin fibers in mammalian cells. A new peptide, known as the Collagen Hybridizing Peptide, can be conjugated with fluorophores and used to stain denatured collagen fibers. Plant cell walls are stained using stains or dyes that bind to cellulose or pectin. The search for fluorescent probes with high specificity that also allow live imaging of plant cells continues.

There are many fluorescent molecules, known as fluorophores or fluorochromes, such as fluorescein, Alexa Fluors, or DyLight 488, which can be chemically linked to another molecule that binds to the target of interest within the sample.

Immunofluorescence: Immunofluorescence is a technique that uses the highly specific binding of an antibody to its antigen to label specific proteins or other molecules within the cell. A sample is treated with a primary antibody that is specific for the molecule of interest. A fluorophore can be directly conjugated to the primary antibody. Alternatively, a secondary antibody, conjugated to a fluorophore, which binds specifically to the first antibody can be used.

For instance, a primary antibody raised in a mouse that recognizes tubulin, combined with a secondary anti-mouse antibody modified with a fluorophore, could be used to label microtubules in a cell.

Fluorescent Proteins: With the modern understanding of genetics and the techniques available for modifying DNA, scientists can genetically modify proteins to carry a fluorescent protein reporter. In biological samples, this allows a scientist to make a protein of interest directly fluorescent. The location of the protein can then be directly tracked, including in live cells.

Fluorescence microscopy limitations:

1. Photobleaching: Fluorophores can degrade or bleach when exposed to intense light, limiting the duration of observation and requiring careful control of light exposure.
2. Phototoxicity: Continuous exposure to the excitation light can induce cellular damage or alter physiological processes, limiting live-cell imaging experiments.
3. Background autofluorescence: Biological samples may exhibit autofluorescence due to endogenous fluorophores or impurities, which can obscure the signal from the fluorophores of interest.
4. Resolution limitations: While super-resolution techniques have improved resolution beyond the diffraction limit, traditional fluorescence microscopy is limited by the diffraction of light, restricting the ability to resolve structures smaller than approximately half the wavelength of the excitation light.
5. Cross-talk between channels: When using multiple fluorophores with overlapping emission spectra, spectral bleed-through or cross-talk between channels can occur, complicating the interpretation of the results.

Lab Exercise I - Discard the parental flies from lab 3.

Material

Genetic crosses performed in lab 3

Methods

1. Observe the Parental progeny of the crosses produced in lab 3.
2. Prepare a fly morgue and discard the parental flies into it.
3. Discard the old label on the old vial, relabel it with "F1", group name and new date.
4. Incubate for another week.
5. Record all results in a lab notebook.

Lab Exercise II - Stain the polytene chromosome with DAPI

Material

- 1 box of Slide
- 1 box of Cover slip
- 1 roll of aluminum foil
- 1 300nM DAPI stain solution in dropper bottle
- 6 PBS (Phosphate Buffered Saline) in dropper bottles
- 6 Wild-type *D. melanogaster* larvae vials
- 6 Bibulous paper
- 6 Distil water bottles
- 12 Fine-tip forceps

300nM DAPI solution preparation (complete by lab technician)

1. Add 2 mL of deionized water (diH₂O) or dimethylformamide (DMF) to the entire contents of the DAPI vial to make a 14.3 mM (5 mg/mL) DAPI stock solution. Note: DAPI has poor solubility in water, so sonicate is necessary to dissolve. The 5 mg/mL DAPI stock solution may be stored at 2–6°C for up to 6 months or at $\leq -20^{\circ}\text{C}$ for longer periods.
2. Add 2.1 μL of the 14.3 mM DAPI stock solution to 100 μL PBS to make a 300 μM DAPI intermediate dilution
3. Dilute the 300 μM DAPI intermediate dilution 1:1,000 in PBS as needed to make a 300 nM DAPI stain solution.

Methods

1. Obtain a third instar larva and dissect out the salivary gland.
2. Transfer the salivary gland onto a glass slide and wash it 1–3 times with phosphate-buffered saline (PBS).
3. Absorb any excess PBS using bibulous paper.
4. Apply a sufficient amount of 300 nM DAPI stain solution to cover the cells.
5. Incubate the slide for 20 minutes, ensuring it is covered with aluminum foil to protect it from light.
6. Absorb the DAPI stain solution with bibulous paper.
7. Wash the cells 2–3 times with PBS.
8. Place the slide under a fluorescence microscope with Excitation/Emission 358/461 nm to image the cells.

Record on lab notebook and answer the questions:

1. Record the fluorescence image
2. Compare this image with those obtained using traditional staining methods for salivary gland cells. Any differences in staining patterns, clarity, and contrast between these two methods?
3. Are you able to see all five major chromosome arms (X, 2L, 2R, 3L, 3R)? Are you able to see the small 4th chromosome? Can you locate the chromocenter in your spread? Are you able to track along a single chromosome arm from its centromere to telomere?

Assignment 4

1. After reading the chapter: "Using the Power of Drosophila Genetics to Study Developmental and Cell Biology" in the paper "Genetics on the Fly: A Primer on the Drosophila Model System" (Hales et al., 2015)

(<https://academic.oup.com/genetics/article/201/3/815/5930114>).

Summarize the process of key embryonic developmental steps and cell biological processes in your own words.(3 points)

2. Describe the method used to find the genes controlling embryonic development.

After watching this video: "Finding Genes that Control Development". (1 point)

(<https://www.ibiology.org/development-and-stem-cells/embryonic-development/>)

Lab 5 Introduction to Bioinformatics and sequencing techniques

Hanqi Ye and Na Xu

Objective:

- To become familiar with the NCBI and Flybase database.
- To practice and use the fundamental bioinformatics tool: basic local alignment search tool (BLAST)
- To understand some sequencing techniques

Introduction to sequencing techniques

First-generation sequencing technology or early sequencing methods didn't emerge until later, but in 1965, significant progress towards establishing a DNA sequencing method was made following the publications of "Structure of a Ribonucleic Acid" by Robert Holley and "A Two-dimensional Fractionation Procedure for Radioactive Nucleotides" by Frederick Sanger (Figure1). They focused their efforts on sequencing RNA fragments, considering RNA less intricate and more manageable than DNA. RNA bacteriophages, lacking a complementary strand, could be more easily produced in culture. Employing enzymes to cleave RNA chains at specific sites allowed comparisons between fully and partially degraded RNA fragments to discern the sequence of bases.

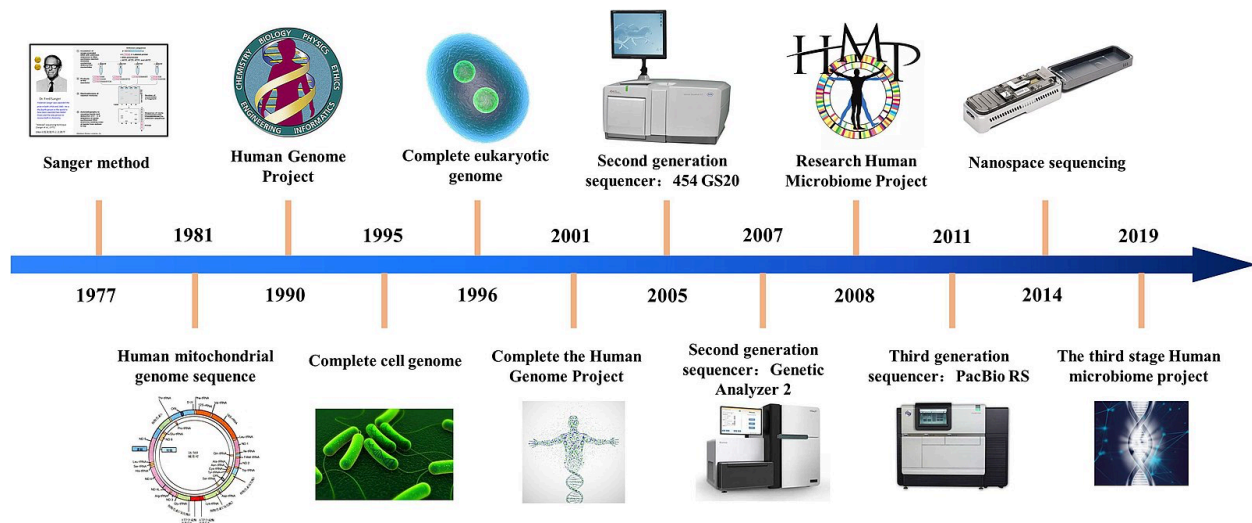


Figure 1 History of sequencing technology by Aimin Yang, Wei Zhang, Jiahao Wang, Ke Yang, Yang Han, and Limin Zhang, CC BY-SA 4.0, via [Wikimedia Commons](#).

A decade later, Sanger and Harvard scientists Allan Maxam and Walter Gilbert revolutionized genomics as they independently introduced their own methods of sequencing DNA in 1977. Frederick Sanger was the first to sequence the complete DNA genome of a bacteriophage, called phi X174 with the method of 'DNA sequencing with chain-terminating inhibitors' which we call sanger sequencing now. In the same year, Allan Maxam and Walter Gilbert produced 'DNA sequencing by chemical degradation', also known as maxam gilbert sequencing. It is a method that involves chemically cleaving DNA molecules at specific nucleotide positions using different chemical reagents. By selectively breaking the DNA backbone at predetermined sites, researchers could deduce the sequence of bases.

Ultimately, Sanger would become the widely adopted approach to DNA sequencing, so much so that Applied Biosystems commercialized and automated the process by the following decade in 1987. These two methods are classified as first-generation sequencing due to their reliance on extensive manual labor and repetitive tasks. Sanger sequencing is based on the use of fluorophore-labeled dideoxynucleotides (ddNTPs) in combination with regular deoxynucleotides (dNTPs). Because ddNTPs lack a hydroxyl group needed for nucleotide binding, the addition of ddNTPs by the DNA polymerase during chain extension terminates the DNA strand elongation. Reiteration of primer annealing and DNA extension cycles results in fragments with a fluorophore-labeled nucleotide at each position, thereby identifying every nucleotide in the DNA template (Figure 2 left).

Next-generation sequencing (NGS) or second generation sequencing: The emergence of the second generation of DNA sequencing started following the introduction of pyrosequencing in 1996. The process was based on sequencing-by-synthesis technology, that measured the luminescence generated from enzymatic conversion of released pyrophosphate into light (Nat Protoc 2). The sequencing-by-synthesis (SBS) method refers to sequential addition and detection of individual nucleotides. Some of the major post-Sanger sequencing techniques include Roche/454, SOLiD, and SOLEXA sequencing technologies. In all these processes, the DNA is sheared prior to attachment of adaptors to the ends. During 454 sequencing, each fragment is attached to a bead, and the DNA on each bead is PCR amplified.

For clonal amplification, primers anneal to single-stranded DNA templates and a DNA polymerase initiates a reaction generating multiple copies of the same sequence. During the sequencing reaction, each nucleotide is added sequentially, and the signal emitted by an individual nucleotide is quantified and classified. SBS was subsequently adopted by several sequencing platforms such as Illumina, IonTorrent, and Pacific Biosciences. The signal measured varies by instrument and includes release of pyrophosphate, fluorescence emission, or a change in electric current..

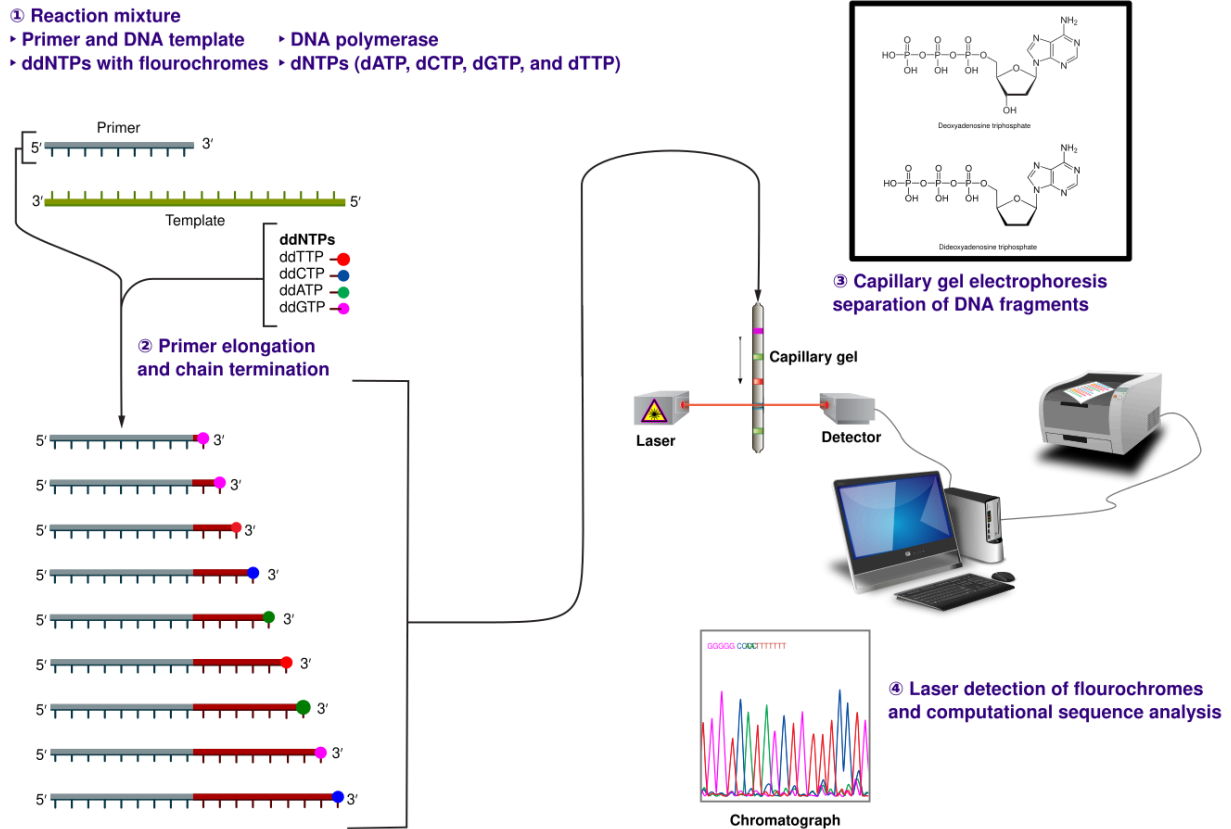


Figure 2 Sanger (chain-termination) method for DNA sequencing” by Estevezj, CC BY-SA 3.0, via Wikimedia Commons. (Source: <File:Sanger-sequencing.svg>)

NGS is similar to Sanger sequencing as they both work with DNA fragments. But in NGS, it's massively parallel. This allows millions of fragments to be sequenced in a single run versus sanger sequencing which only produces one forward and reverse read. Researchers today have many choices when deciding which sequencing technology to use for their clinical research. Sanger sequencing with 99.99% accuracy is the “gold standard” for clinical research sequencing. However, newer NGS technologies are also becoming common in clinical research labs due to their higher throughput capabilities and lower costs per sample.

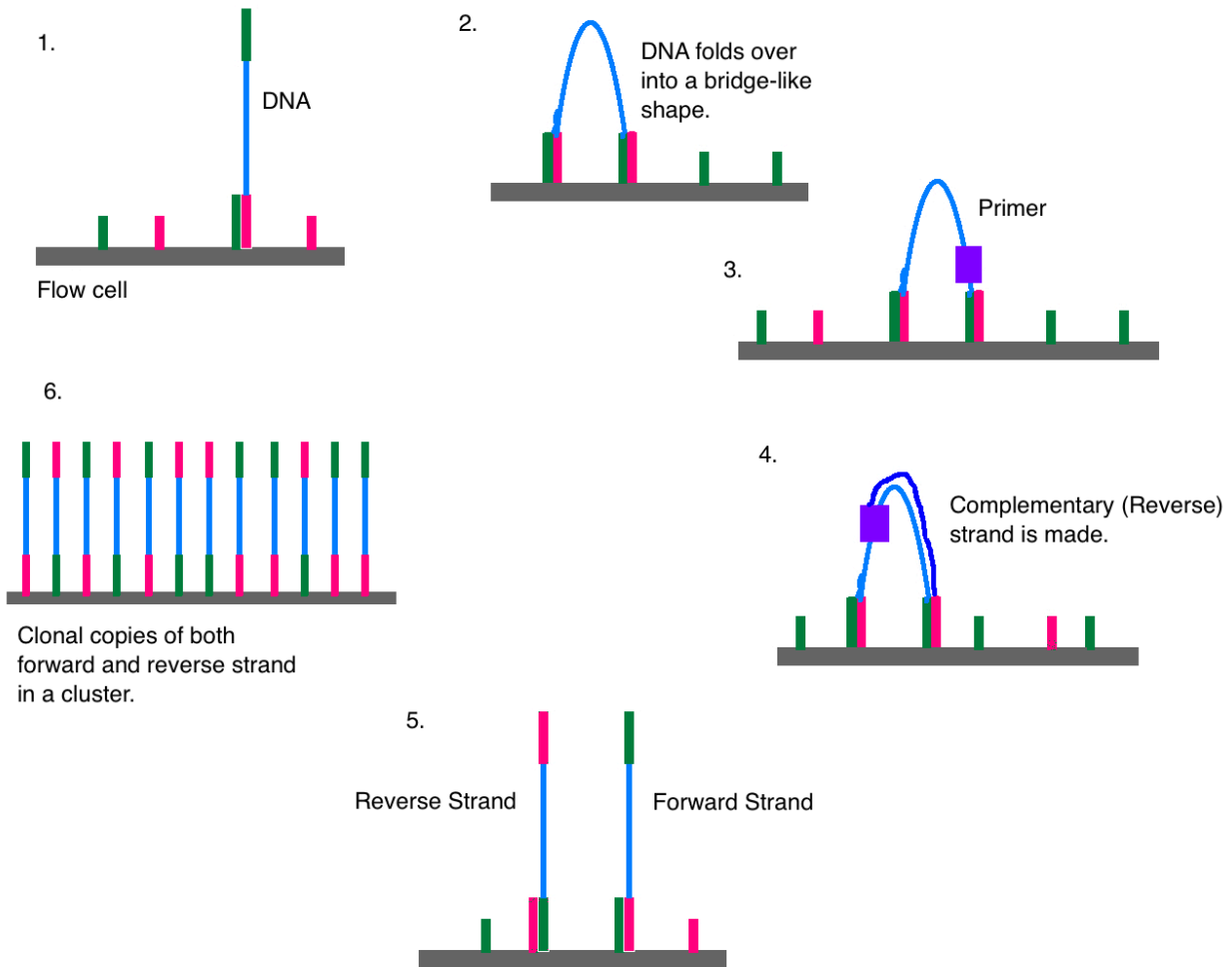


Figure 3 *Cluster Generation.png* by DMLapato, published 28 September 2015. Licensed under CC BY-SA 4.0. Via Wikimedia Commons:

https://commons.wikimedia.org/wiki/File:Cluster_Generation.png

In figure 3, after step 1 DNA extraction, the 2nd step is library preparation which begins with DNA fragmentation. DNA fragmentation refers to the deliberate breaking down of targeted DNA into numerous shorter segments, typically ranging from 100 to 300 base pairs (bp) in length. Various techniques can accomplish this task, including mechanical methods, enzymatic digestion, and other approaches. For instance, sonication is employed to fragment DNA into shorter segments. Subsequently, relevant short segments pertaining to the targeted DNA sequences are isolated using specific complementary probes, a method commonly known as the hybridization capture assay. Alternatively, the polymerase chain reaction (PCR) amplification method utilizes multiple pairs of primers to amplify the targeted DNA segments, generating PCR products that serve as short segments of the targeted DNA, a method referred to as the amplicon assay. These DNA segments are then utilized for library preparation.

Library preparation (step 2) involves modifying DNA segments to incorporate sample-specific indexes, facilitating the identification of the patient from whom DNA sequencing was conducted. Additionally, this process facilitates the attachment of sequencing adaptors to the DNA segments, enabling sequencing primers to bind to all DNA segments for subsequent massive parallel sequencing.

Massive parallel sequencing (step 3) is executed using next-generation sequencing (NGS) platforms. The library is loaded onto a sequencing matrix specific to the sequencer being used, such as flow cells for Illumina NGS sequencers and sequencing chips for Ion Torrent NGS sequencers. Despite the variations in sequencing matrices among different platforms, the objective remains consistent: to enable simultaneous massive parallel sequencing of all DNA segments. The resulting sequence data is analyzed using bioinformatics software.

Bioinformatics analysis (step 4) encompasses several steps, including base calling, read alignment, variant identification, and variant annotation. During this phase, the sequence information is compared to a human genome reference sequence to detect any variants or mutations within the targeted sequences. The information from each sequenced segment is integrated to produce final sequencing results for the entire length of the targeted DNA. These final results are then provided to the user for interpretation. The annotation and interpretation processes aim to identify each variant and assess its potential biological or clinical significance.

Third-generation sequencing refers to the latest advancements in DNA sequencing technology that offer improved accuracy, throughput, and efficiency compared to previous generations. These methods typically involve single-molecule sequencing, where DNA molecules are directly sequenced without the need for amplification or fragmentation.

One example of third-generation sequencing technology is Nanopore sequencing, developed by Oxford Nanopore Technologies, involves passing DNA strands through protein nanopores embedded in a membrane. As the DNA moves through the nanopore, changes in electrical current are measured, allowing for real-time sequencing of the DNA molecule. This technology offers long read lengths over 300 kb in length and portability, making it suitable for a wide range of applications, including fieldwork and clinical diagnostics.

First-generation Sequencing: This traditional method is commonly used for sequencing short to medium-length DNA fragments. It's useful for verifying sequences, identifying mutations, and sequencing individual genes. However, it's relatively slow and expensive compared to newer techniques.

Next-Generation Sequencing (NGS): NGS methods, such as Illumina sequencing, are used when high-throughput sequencing is required. They're ideal for whole-genome sequencing, RNA sequencing (RNA-seq), chromatin immunoprecipitation sequencing (ChIP-seq), and other applications where large amounts of data need to be generated quickly and cost-effectively.

Third-Generation Sequencing (Long-Read Sequencing): Technologies like PacBio and Oxford Nanopore produce longer reads compared to NGS platforms. Long-read sequencing is particularly useful for de novo genome assembly, structural variant detection, and studying complex genomic regions, such as repetitive sequences or structural variations.

The history of genome sequencing reflects one of the most remarkable technological advances in modern biology. Early sequencing methods, developed in the 1970s, laid the foundation for decoding small fragments of DNA, allowing scientists to begin uncovering the genetic instructions that underlie life. With the introduction of Sanger-based shotgun sequencing, researchers gained the ability to sequence and assemble much larger genomes. This approach played a central role in landmark achievements such as the Human Genome Project, which demonstrated the power of large-scale DNA sequencing. In the early 2000s, the emergence of next-generation sequencing (NGS) revolutionized the field by making sequencing faster, more accurate, and affordable. These high-throughput technologies opened new possibilities for studying genetic variation, disease, and evolution across countless species. More recently, third-generation sequencing has extended these capabilities even further by producing much longer reads of DNA, improving the completeness and accuracy of genome assemblies. Together, these advancements have transformed genomics from a slow, labor-intensive process into a rapid and comprehensive approach for exploring the complexity of life at the molecular level. *it becomes evident that as sequencing techniques advance, the capacity to sequence larger genomes increases while the cost per human genome declines at a rate surpassing the trajectory predicted by Moore's Law, particularly since 2008, as illustrated in Figure 5.*

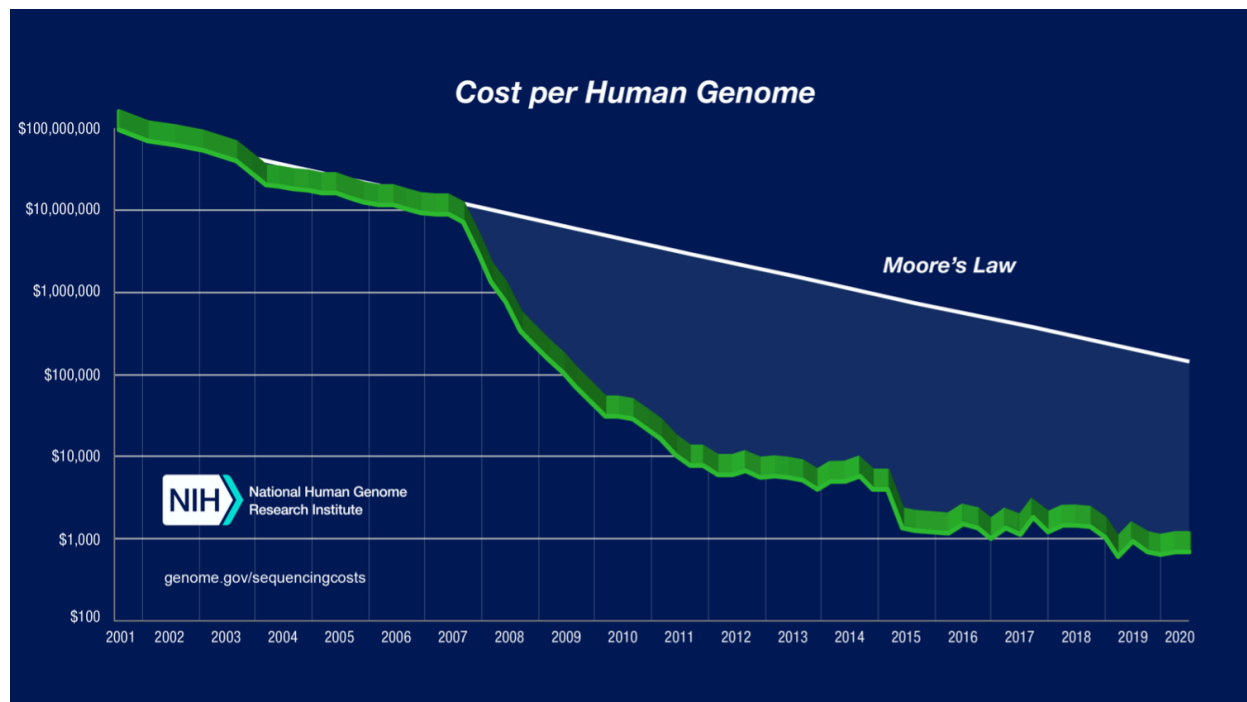


Figure 4 Approximate cost of sequencing a human-sized genome since September 2001" (NHGRI, via Wikimedia Commons). License: [insert licence, e.g., CC BY-SA 4.0]

This remarkable trend underscores the pivotal role of bioinformatics in genomics. As sequencing technologies evolve, more genome would be sequenced so bioinformatics tools and algorithms will play a crucial role in managing, analyzing, and interpreting the vast amounts of genomic data generated.

Bioinformatics is an interdisciplinary field that develops methods and software tools for understanding biological data. The main driving force for the emergence and rapid development of bioinformatics comes from the application of high-throughput technologies such as next-generation sequencing in the field of life science. At present, the genomes of humans, animals, insects, and plants have been sequenced. The number of known DNA sequences exceeds 20 trillion bases, doubling every five months. The unprecedented amount of "big data" contains a significant number of unknown new discoveries and new knowledge, providing new historical opportunities for life science research.

At the largest scale, bioinformatics is divided into databases and software tools. The databases are classified as large centralized resources such as the **National Center for Biotechnology Information (NCBI)** (<https://www.ncbi.nlm.nih.gov/>) of the United States, the **European Bioinformatics Institute (EBI)** (<https://www.ebi.ac.uk/>) of Europe, the **Genome Browser of the University of California Santa Cruz (UCSC)** (<https://genome.ucsc.edu/>), and specific independent resources built by research groups like, such as **Flybase** (<http://flybase.org>) which is a joint project with the Berkeley Drosophila Genome Project.

FlyBase is the leading database and web portal for genetic and genomic information on the fruit fly *D. melanogaster* and related fly species. All databases offer powerful software to analyze and interpret big biological data. One of the most useful tools is the **BLAST**. BLAST finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families. This tool enables thorough analyses of large and complex data sets.

Lab Exercise I - Observe F1 Generation Phenotypes from Lab 4

Material

Genetic crosses performed in Lab 4

12 new food vials

6 markers

Fly nap

Methods

1. Observe the F1 progeny of the crosses produced in lab 4 and their phenotypes.
2. Get rid of all parent-alike flies and keep only the wild type like flies.
3. Count the total number of F1 generation, separate flies based on their gender.
4. Transfer 5 male and 5 female to a new food vial.
5. Transfer the rest male and female to another food vial.
6. Label new food with F2, date, name of group and name of the cross.

Lab Exercise II - Get to know NCBI: Use BLAST to Analyze Unknown Genes and Protein Sequences

Material

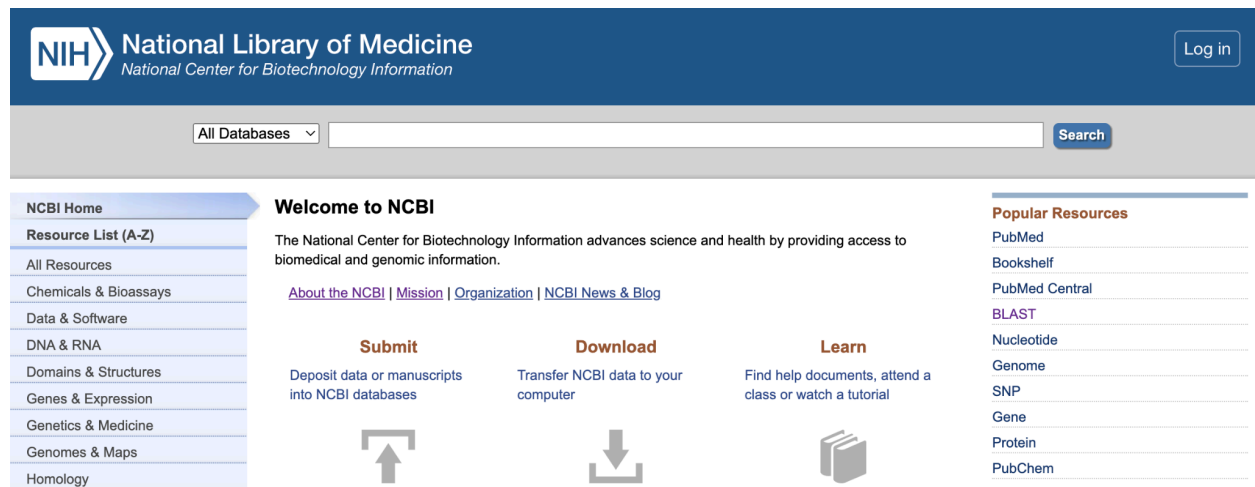
Laptops

Two unknown sequences

NCBI home page (<https://www.ncbi.nlm.nih.gov/>)

Methods

1. Download two unknown sequences [here](#).
2. Open the NCBI home page (<https://www.ncbi.nlm.nih.gov/>), click on BLAST at the right side of the page under popular resources.



3. Use BLAST to search nucleotide sequences and BLASTp to search protein sequences. BLASTx searches protein databases using a translated nucleotide query. Verify the unknown gene sequence using BLASTx results since it is a protein-coding gene.

Web BLAST



- In the blast page, upload the unknown sequences, use all default parameters, check the box "show result in a new window", and click on the BLAST button. The result page will refresh itself until you see the BLAST summary.

Enter Query Sequence

BLASTN programs search nucleotide databases using a nucleotide query.

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#) [Clear](#)

Query subrange [?](#)

From

To

Or, upload file demo_scaffolds.fasta [?](#)

Job Title

Enter a descriptive title for your BLAST search [?](#)

☐ Align two or more sequences [?](#)

- Find the optimal hit of the query sequences. As shown below, the BLAST query alignments are displayed in rank order according to their "Max Score". For this example, the top hit from the Demo scaffold sequence was *Escherichia coli* strain 2011C-3911 with Accession #CP015240.1.

Descriptions

Graphic Summary

Alignments

Taxonomy

Sequences producing significant alignments

Download

Manage Columns

Show100

☒

select all

100 sequences selected

GenBank

Graphics

Distance tree of results

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☒	Escherichia coli strain 2011C-3911, complete genome	55042	56150	100%	0.0	99.85%	CP015240.1
☒	Escherichia coli strain ESBL 15 chromosome, complete genome	54765	55410	100%	0.0	99.68%	CP041678.1
☒	Escherichia coli strain AMSHJX01 chromosome, complete genome	54748	55393	100%	0.0	99.67%	CP030939.1
☒	Escherichia coli strain S21, complete genome	54741	55856	100%	0.0	99.67%	CP010230.1
☒	Escherichia coli strain H14, complete genome	54741	55856	100%	0.0	99.67%	CP010177.1
☒	Escherichia coli strain 2015C-3905 chromosome, complete genome	54732	55376	100%	0.0	99.66%	CP027371.1
☒	Escherichia coli strain 2014C-3097 chromosome, complete genome	54726	55371	100%	0.0	99.66%	CP027449.1
☒	Escherichia coli strain ATCC 43896 chromosome, complete genome	54726	55847	100%	0.0	99.66%	CP024278.1

- Clicking the Description name for one of the BLAST hits will display a sequence alignment of the query sequence versus the NCBI nucleotide database sequences, as shown in the figure below. This result includes the Karlin-Altschul Expect value, the percent identity of the alignment, the number of gaps in the alignment, and other pertinent information.

Escherichia coli strain 2011C-3911, complete genome

Sequence ID: [CP015240.1](#) Length: 4863599 Number of Matches: 4

Range 1: 4671753 to 4701689 [GenBank](#) [Graphics](#)

▼ [Next Match](#) ▲ [Previous Match](#)

Score	Expect	Identities	Gaps	Strand
55042 bits(29806)	0.0	29896/29940(99%)	3/29940(0%)	Plus/Plus
Query 1	CCTCCCAA	ACTGGCGATATCCAGTCTGGTAAACAGCCTCAAAGGCGTATCGGGTAGGTTA	60	
Sbjct 4671753	CCTCCCAA	ACTGGCGATATCCAGTCTGGTAAACAGCCTCAAAGGCGTATCGGGTAGGTTA	4671812	
Query 61	CTGCGACGAGATCGACCAGATATTGCAGTCAGGTATTACTACAAAGGCGTTTTGTGGAGT	120		
Sbjct 4671813	CTGCGACGAGATCGACCAGATATTGCAGTCAGGTATTACTACAAAGGCGTTTTGTGGAGT	4671872		
Query 121	CCTGGCTATTTTGCCAGTAGCTGCGGAGGTGCGCCAATATCCGTCATCCGCCAATACATT	180		
Sbjct 4671873	CCTGGCTATTTTGCCAGTAGCTGCGGAGGTGCGCCAATATCCGTCATCCGCCAATACATT	4671932		

Lab Exercise III - Search and Download Required *Drosophila melanogaster* Gene Sequences

Material

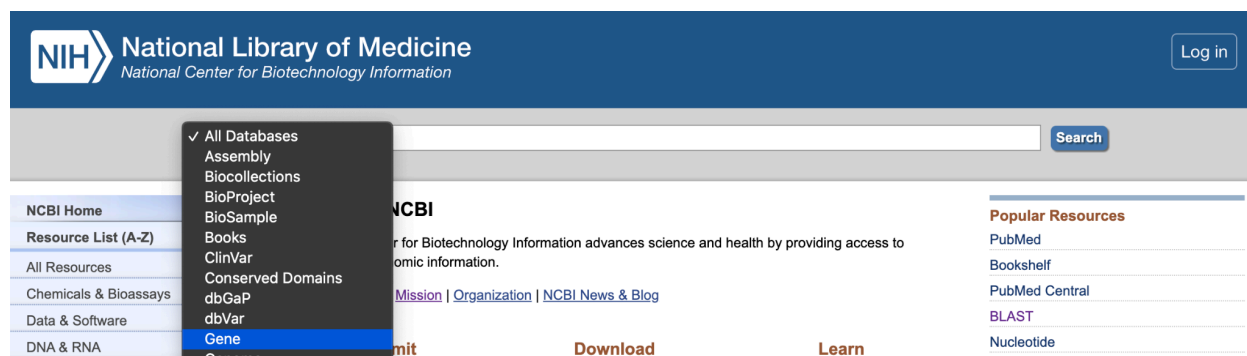
Laptops

NCBI home page (<https://www.ncbi.nlm.nih.gov/>)

Flybase (<http://flybase.org>)

Methods

1. Open **NCBI**, click on the top dropdown box and choose “Gene”.



2. At the search box, search four *D. melanogaster* genes: **bar**, **eyeless**, **vestigial**, and **ebony**.

Gene [Create RSS](#) [Save search](#) [Advanced](#)

Tabular ▾ 20 per page ▾ Sort by Relevance ▾ Send to: ▾

See [ADRB2 \(BAR\) adrenoceptor beta 2](#) in the Gene database
bar in [Homo sapiens \(3\)](#) [All 16 Gene records](#)

Search results
Items: 1 to 20 of 73489 << First < Prev Page 1 of 3675 Next > Last >>

[See also 1262 discontinued or replaced items.](#)

Name/Gene ID	Description	Location	Aliases	MIM
☐ hh ID: 42737	hedgehog [<i>Drosophila melanogaster</i> (fruit fly)]	Chromosome 3R, NT_033777.3 (23128169..23141906, complement)	Dmel_CG4637, CG4637, DmelCG4637, Dmhh, HH, Hg, Hh, Mir, Mrt, anon-WO0134654.19, anon- WO0182946.19, bar, bar-3, bar3, hg, l(3)hh, l(3)neo56, l(3)neo57	
☐ B ID: 44798	Bar [<i>Drosophila melanogaster</i> (fruit fly)]		Bar, BarH1, bar	
☐ B-H1 ID: 32724	BarH1 [<i>Drosophila melanogaster</i> (fruit fly)]	Chromosome X, NC_004354.4	Dmel_CG5529, BH1, Bar, Bar H1, Bar- H1, BarH, BarHl, Barh1, CG5529,	

clear

Filters: [Manage Filter](#)

Results by taxon

Top Organisms [Tr](#)
 Apaloderma vittatu
 Caenorhabditis ele
 Salmo trutta (206)
 Oncorhynchus myl
 Cyprinus carpio (2
 All other taxa (580
 More...

Find related data
 Database:

Search details

- Click on the optimal hit and find the download datasets button. Download gene, transcript, and protein sequences. Rename the sequence files in the format of '**Gene name_gene/rna/protein**' and keep the original file extension.

B-H1 BarH1 [*Drosophila melanogaster* (fruit fly)]

[Download Datasets](#)

Gene ID: 32724, updated on 23-Jun-2022

Summary

Official Symbol B-H1 provided by [FlyBase](#)

Official Full Name BarH1 provided by [FlyBase](#)

Primary source [FLYBASE:FBgn0011758](#)

Locus tag Dmel_CG5529

See related [AllianceGenome:FB:FBgn0011758](#)

Gene type protein coding

RefSeq status REVIEWED

Organism [Drosophila melanogaster](#)

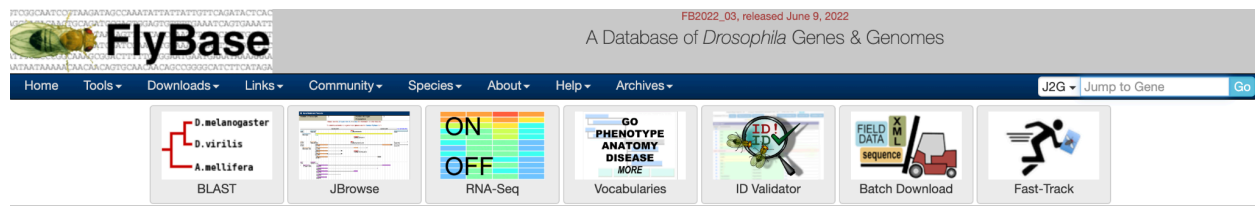
Lineage Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta; Pterygota; Neoptera; Holometabola; Diptera; Brachycera; Muscomorpha; Ephydroidea; Drosophilidae; Drosophila; Sophophora

Also known as Bar; Bar H1; bar-h1; Bar-H1; BarH; barh1; barH1; Barh1; BarHl; BH1; CG5529; DmelCG5529

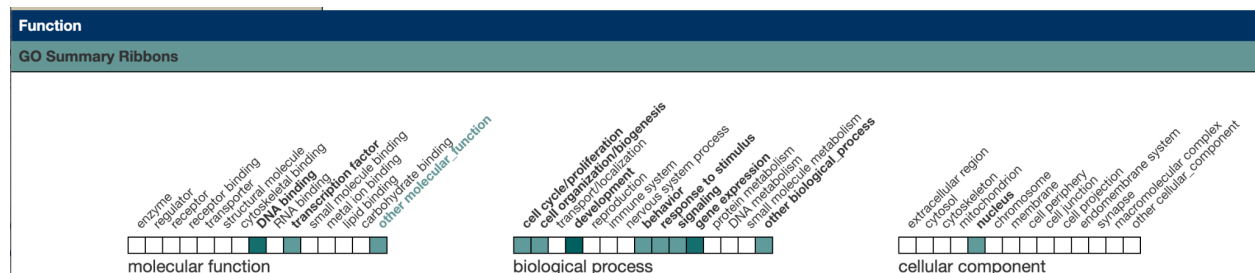
Summary Enables DNA-binding transcription factor activity, RNA polymerase II-specific. Involved in several processes, including animal organ morphogenesis; eye pigment granule organization; and leg disc proximal/distal pattern formation. Located in nucleus. Is expressed in several structures, including extended germ band embryo; larval segment; nervous system; presumptive embryonic/larval system; and sensory nervous system primordium. Human ortholog(s) of this gene implicated in oral squamous cell carcinoma. Orthologous to several human genes including BARHL2 (BarH like homeobox 2). [provided by Alliance of Genome Resources, Apr 2022]

NEW Try the new [Gene table](#)
Try the new [Transcript table](#)

- Read the hit report and fill up the table in the **results section** below.
- When looking for Gene Ontology (GO), use Flybase as it provides more details.



6. In **Flybase**, click on the drop box at the top right corner of the page and then choose "Jump to Gene(J2G)". Subsequently, type the desired gene name or symbol in the search box and click on the search icon.
7. Find the section of GO Summary Ribbons and pick the main (with the deepest color) molecular function, biological process, or cellular component.



Biotechnology and Society Module II: Genetically Modified Organisms (GMO)

For thousands of years, humans have been using traditional modification methods such as **selective breeding and crossbreeding** to breed plants and animals with desirable traits. For example, early farmers developed crossbreeding methods to grow corn with a range of colors, sizes, and uses. Currently, strawberries are a cross between a strawberry species native to North America and a strawberry species native to South America.

Changing plants and animals through traditional breeding can take a long time, and it is difficult to produce very specific changes. The development of genetic engineering in the 1970's enabled similar changes in a more specific manner and in a shorter amount of time. Further reading on the history of major events can be found in "Genetic Modification in Modern Agriculture" through this link: <https://www.fda.gov/media/135276/download>.

What Is a GMO?

According to the World Health Organization (WHO), a **GMO** is an "organism(s) in which the genetic material (DNA) has been altered in a way that does not occur naturally by mating and/or natural recombination." In other words, the variation in a GMO is the result of genetic manipulation in a laboratory environment and not natural processes such as reproduction.

The prefix “trans-” means “across”; therefore, a transgenic organism has DNA that is across species. In a **transgenic** organism, the DNA of a target species is modified by the insertion of a gene, or a portion of a gene, from a different species (such as a bacterium).

The prefix “cis-” means “the same.” In **cisgenic** GMOs, foreign DNA is not integrated into the target species, but rather the DNA of the target organism is modified to alter the physical characteristics of the target organism.

“Intra-” means “within.” In an **intragenic** GMO, pieces of the organism’s own genetic information have been rearranged to improve a characteristic.

Genetically modified (GM) foods are foods derived from organisms whose genetic materials are modified in a way that does not occur naturally. Transferring genes into crop plants confers resistance to herbicides, insect pests, and plant diseases. Gene transfer can also be used to increase the nutritional value of foods. The technology used is called “**recombinant DNA technology**” or “**genetic engineering**”. Currently, available GM foods stem mostly from plants. However, in the future, foods derived from GM microorganisms or GM animals are likely to be introduced in the market. GM foods can also allow reductions in food prices through improved yields and reliability. All GM foods should be assessed before being approved for sale. The Food and Agriculture Organization of the United Nations (FAO)/World Health Organization (WHO) Codex guidelines exist for risk analysis of GM food. All genetically modified foods currently available on the international market have passed safety assessments. When discussing and developing GM foods, three main safety issues are considered. The first is allergenicity, meaning the ability of the edited genes or food product to cause an allergic reaction. No allergic effects have been found relative to GM foods currently on the market. The second is the transfer of genes from GM foods to the human digestive tract. The third is outcrossing, meaning the transfer of genes from GM organisms to other species. Several countries have developed preventative strategies, including a clear separation of GM and non-GM food crops.

Currently, many GM foods are available in grocery stores. For instance, soybean, cotton, canola, and sugar beets in the market have been genetically modified to have a natural resistance to an herbicide. This increased resistance to the herbicide allows farmers to use more herbicides to kill weeds. Most cotton has been genetically modified to increase yield and resistance to disease. In 2018, GM soybeans comprised 94% of all planted soybeans, GM cotton comprised 94% of all planted cotton, and 92% of corn were genetically modified organisms (GMO). In 2013, GM canola comprised 95% of planted canola while GM sugar beets comprised 99.9% of all harvested sugar beets. Most GMOs are used to produce ingredients that are subsequently used in other food products, such as cornstarch, derived from GM corn, or sugar derived from GM sugar beets. GM soy is also used as an ingredient (lecithin, emulsifiers, and proteins) in processed foods.

Corn is the most-grown crop in the United States, and most of it is GMO. Most GM corn is created to resist insect pests or tolerate herbicides. *Bacillus thuringiensis* (Bt) corn is GM corn

that produces proteins that are toxic to certain insect pests but not to humans, pets, livestock, or other animals. These are the same types of proteins that organic farmers use to control insect pests, and they do not harm other, beneficial insects such as ladybugs. GM Bt corn reduces the need for spraying insecticides while still preventing insect damage. While a significant amount of GM corn is used in processed foods and drinks, most of it is used to feed livestock, such as cattle, and poultry including chickens.

By the 1990s, ringspot virus disease had nearly eliminated Hawaii's papaya crop. GM papaya, named Rainbow papaya, was created to resist the ringspot virus. This GMO saved papaya farming on the Hawaiian Islands.

A few varieties of GM apples were developed to resist browning after harvesting. This helps reduce food waste, as many consumers believe brown apples are spoiled.

More than 95% of animals used for meat and dairy in the United States are fed GMO crops. The DNA in GMO food does not transfer to the animal that consumes it. This means that animals that consume GMO food do not turn into GMO. Similarly, the DNA from GMO animal food is not present in the meat, eggs, or milk of the animal. Research shows that products such as eggs, dairy products, and meat derived from animals fed with GMO have the same nutritional value, safety, and quality as those derived from animals that consume only non-GMO food. **GMO animals** will soon be available in the food supply chain. The FDA has approved an application that allows the marketing of AquAdvantage Salmon, an Atlantic salmon genetically modified for accelerated growth. The FDA determined that AquAdvantage Salmon is safe and has the same nutritional value as non-GMO Atlantic salmon.

Bacteria were the first organisms to be genetically modified. This is primarily due to the small size of a bacterial genome (often less than a few hundred genes), the presence of plasmids (small pieces of circular DNA), and their rapid generation time (as often as every 15 minutes), which make them ideal for genetic manipulation.

Production of biotechnology products. GM bacteria have been used in the production of medicine and pharmaceutical compounds since the beginning of medically related GMOs. One of the first human compounds produced by bacteria was human insulin. GM bacteria also produce vaccines for hepatitis B, clotting factors for hemophilia, human growth hormone, and interferons used in cancer treatment, just to name a few.

Protection of plants. Bacteria have been engineered to produce proteins that protect plants from the effects of freezing. Also, chemicals produced by the *Bacillus thuringiensis* (Bt) bacteria have been inserted into plants to protect them from insect predation.

Bioremediation. By enhancing enzymes and other metabolic pathways, GM bacteria can be used to break down toxic chemicals in the environment. This approach has been used successfully for large oil spills, and against toxic levels of mercury.

Synthesis of organic chemicals. Along with algae (a eukaryotic ancestor of plants), GM bacteria are able to produce biofuels and other organic chemicals for use in manufacturing.

How Can Genetic Modification Improve Fruit Ripening?

Fruits can over-ripen and spoil before being purchased by customers. In apples, the browning reaction is caused by exposure to chemicals in the fruit called **polyphenols**. An enzyme in the fruit, **polyphenol oxidase (PPO)**, facilitates a reaction between the polyphenols and oxygen, causing the fruit to brown. In some GM fruits, the gene responsible for the PPO enzyme is silenced, and ripening is slowed down.

How Can Genetic Modification Increase Crop Yield?

With almost 7.9 billion people on the planet and only a limited amount of farmable land, humans have a real need to increase the amount of food produced per acre, also referred to as **crop yield**. We have already learned that transgenic corn that contains *Bacillus thuringiensis* (BT) endotoxin genes is capable of controlling many insect pest species of plants. The same approach applies to making plants resistant to the herbicides (namely **glyphosate**, also called RoundUp) that control weed populations.

How Can Genetic Modification Produce More Nutritious Food?

In the early 1990s, a group of researchers realized it may be possible to produce a variety of rice that included beta-carotene, a nutrient that is the precursor of vitamin A, an important vitamin in human development. It is estimated that over 600,000 children die annually due to vitamin A deficiencies, and because rice is a staple food source for a large part of the human population, it was ideal as a method for delivering the nutrient. Because rice does not normally produce beta-carotene, it was necessary to genetically modify its existing metabolic pathways to allow it to manufacture the compound. This process involves the addition of two genes to the rice genome: one from a daffodil; the other from a soil bacterium.

Concerns about GMOs may include:

- Are foods containing new proteins safe to eat?
- Is it Safe to eat food carrying part of a viral gene that switches on transgenes?
- Will pesticide-resistant insects develop?
- Will disease-causing bacteria acquire antibiotic-resistance genes used as markers?

**Students will use the material in the "Biotechnology and Society" modules to complete the "Biotechniques Reflection Assignment" for this course.*

Assignment 5

1. Lab Exercise II: Complete the table according to the BLAST results.

Query sequence	Blast method	Species	Possible gene or protein name
Unknown gene sequence			
Unknown protein sequence			

2. Lab Exercise III: Download these required genes and fill out the table below.

Gene name	Gene type	Gene Length	Exon count	Chromosome	GO-molecular function	GO-Biological process	GO-cellular component
Bar (B/BH1)							
Eyeless (ey)							
Vestigial (vg)							
Ebony (e)							

3. Create a Word file, screenshot the sequence files you have downloaded with the correct name format, and copy the above two tables to the Word file.

Assignment 5 Instruction Video:

https://cuny907-my.sharepoint.com/personal/hanqi_ye06_login_cuny_edu/_layouts/15/stream.aspx?id=%2Fpersonal%2Fhanqi%5Fye06%5Flogin%5Fcuny%5Fedu%2FDocuments%2FSCB252%20videos%2FLab5%2Dvideo%20demo%2Emp4&ga=1

Lab 6 Gene structural annotation and sequence analysis

Hanqi Ye

Objective

- To understand gene structure and be able to identify different gene elements.
- To be familiar with alternative splicing and how to predict splice donor and acceptor sites in a gene.
- To practice web-based gene sequence analysis tools: genome browsers and CD-search.

Introduction

Gene structure is the organization of specialized sequence elements within a gene. Genes contain most of the information necessary for living cells to survive and reproduce. A gene is transcribed (copied) from DNA into RNA, which can either be non-coding (ncRNA) with a direct function, or an intermediate messenger (mRNA) that is then translated into a protein. Each of these steps is controlled by specific sequence elements, or regions, within the gene. Every gene, therefore, requires multiple sequence elements to be functional (Figure 1).

Much of the gene structure is broadly similar between eukaryotes and prokaryotes. These common elements largely result from the shared ancestry of cellular life in organisms over 2 billion years ago. Understanding gene structure is the foundation to understanding gene annotation, expression, and function.

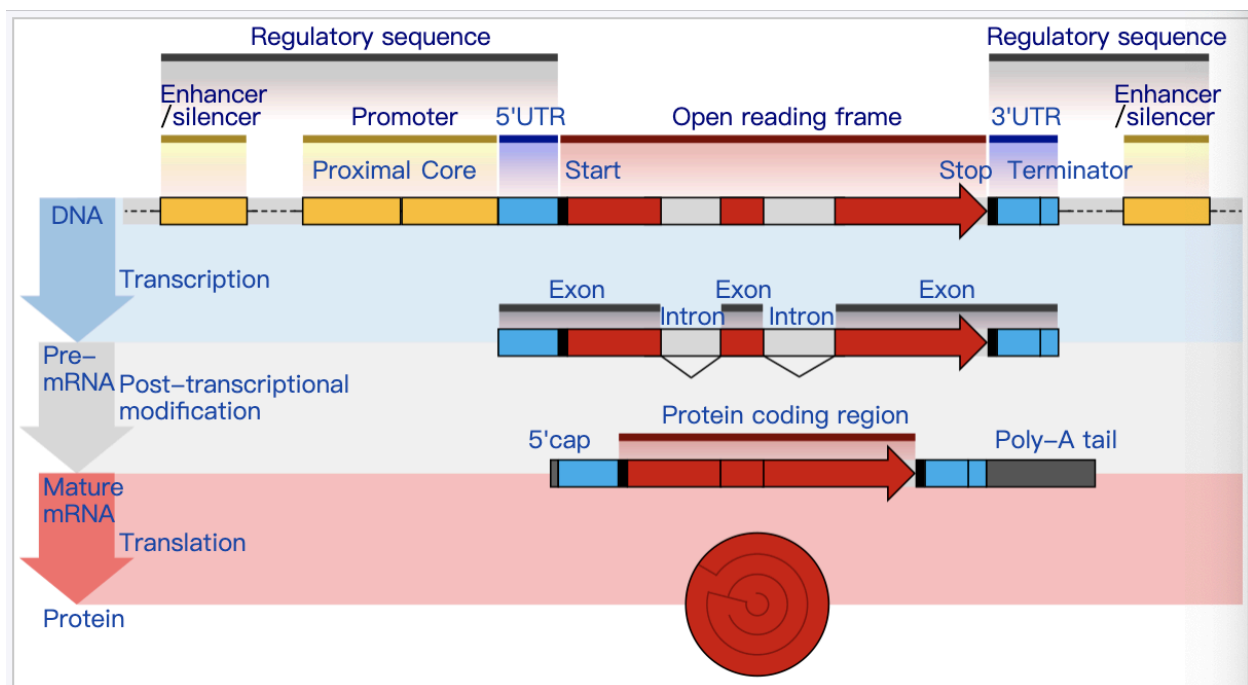


Figure 1. "Gene structure eukaryote 2 annotated.svg" by Thomas Shafee (User: Evolution and evolvability, licensed under CC BY 4.0, via Wikimedia Commons.)

Common genomic features are listed below:

Regulatory sequence: a segment of a nucleic acid molecule that can increase or decrease the expression of specific genes within an organism. Regulation of gene expression is an essential feature of all living organisms and viruses.

Open Reading Frame (ORF): spans the DNA sequence between the start and stop codons.

Promoter: a sequence of DNA to which proteins bind to initiate transcription of a single RNA transcript from the DNA downstream of the promoter.

Enhancer: a short (50–1500 bp) region of DNA that can be bound by proteins (activators) to increase the likelihood that transcription of a particular gene will occur.

Silencer: a sequence-specific element that induces a negative effect on the transcription of its gene. The most common position is found 20 bp to -2000 bp upstream of the target gene where it can help repress the transcription of the gene. Certain silencers can be found downstream of a promoter located within the intron or exon of the gene itself. Silencers have also been found within the three prime untranslated regions (3' UTR) of mRNA.

Terminator: a section of the nucleic acid sequence that marks the end of a gene or operon in genomic DNA during transcription.

Untranslated region (UTR): an UTR refers to either one of two sections, one on each side of a coding sequence on a strand of mRNA. If it is found on the 5' end, it is called the 5' UTR (or leader sequence). If it is found on the 3' end, it is called the 3' UTR (or trailer sequence).

Exon: any part of a gene that will form a part of the final mature RNA produced by that gene after introns have been removed by RNA splicing

Intron: any nucleotide sequence within a gene that is removed by RNA processing during the production of the final RNA product.

Start Codon: the first codon (nucleotide triplet within messenger RNA) of a messenger RNA (mRNA) transcript translated by a ribosome. The standard start codon is **AUG**.

Stop Codon: a codon (nucleotide triplet within messenger RNA) that signals the termination of the translation process of the current protein. There are 3 STOP codons in the genetic code - **UAG, UAA, and UGA**.

Protein coding region: The coding region of a gene, also known as the coding sequence (CDS), is the portion of a gene's DNA or RNA that codes for protein.

Key differences in gene structure between eukaryotes and prokaryotes reflect their divergent transcription and translation machinery. For example, a key feature of the structure of eukaryotic genes is that their transcripts are typically subdivided into exons and introns. Exons are retained in the final mature mRNA molecule, while introns are spliced out (excised) during post-transcriptional processing.

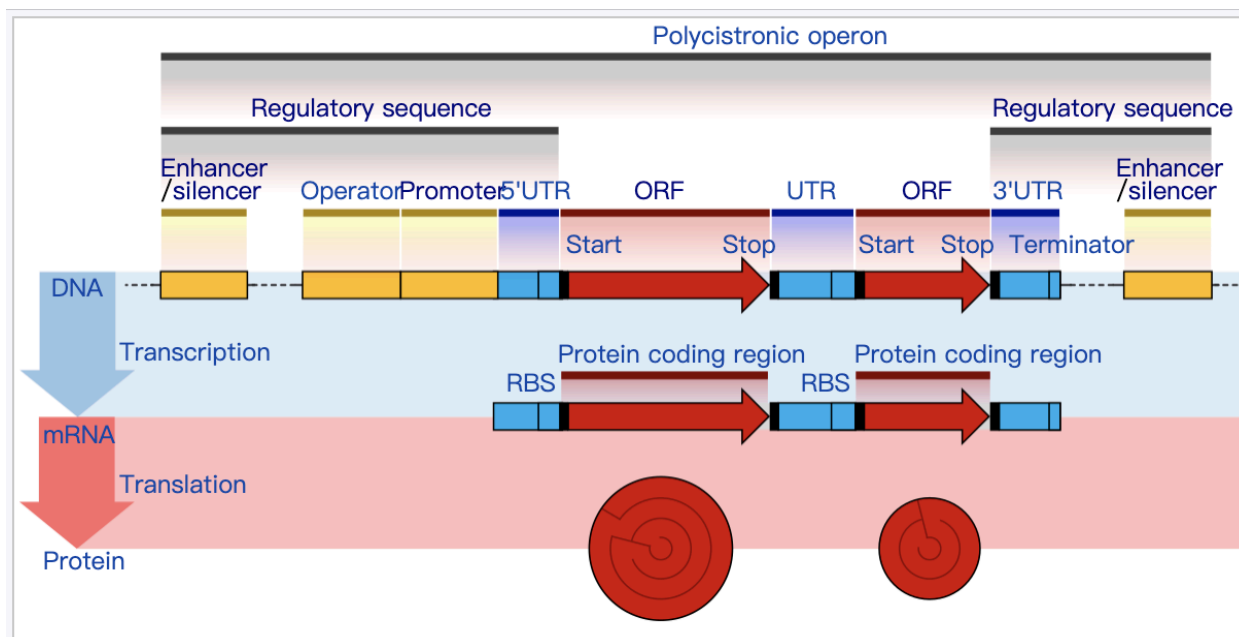


Figure 2. "Gene structure prokaryote 2 annotated." by Thomas Shafee (User: Evolution and evolvability, licensed under CC BY 4.0, via Wikimedia Commons.)

Within introns, a splice donor site (5' end of the intron), a branch site (near the 3' end of the intron), and a splice acceptor site (3' end of the intron) are required for splicing. The splice donor site includes an almost invariant sequence **GU** (or **GT** when represented in DNA) at the 5' end of the intron, within a larger, less highly conserved region. The splice acceptor site at the 3' end of the intron terminates the intron with an almost invariant **AG** sequence. Upstream (5'-ward) from the AG there is a region high in pyrimidines (C and U), or polypyrimidine tract. Further upstream from the polypyrimidine tract is the branchpoint, which includes an adenine nucleotide involved in lariat formation. The consensus sequence for an intron (in IUPAC nucleic acid notation, see the table on the right side) is G-G-[cut]-G-U-R-A-G-U (donor site) ... intron sequence ... Y-U-R-A-C (branch sequence 20–50 nucleotides upstream of the acceptor site) ... Y-rich-N-C-A-G-[cut]-G (acceptor site).

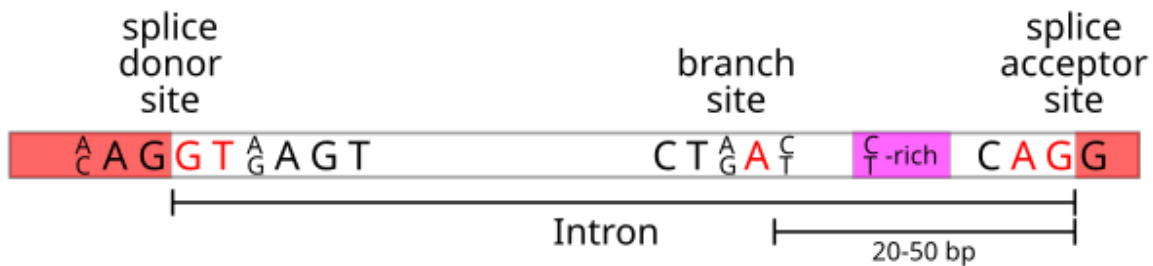


Figure 3. Schematic view of the consensus sequence in RNA splicing. ("RNA splicing.svg" by Michael Jeltsch (User: Mjeltsch), licensed under Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0), via Wikimedia Commons.)

Alternative splicing: (alternative RNA splicing or differential splicing) is an alternative splicing process during gene expression that allows a single gene to code for multiple proteins. In this process, particular exons of a gene may be included within or excluded from the final, processed mRNA produced from that gene. This means the exons are joined in different combinations, leading to different (alternative) mRNA strands. Consequently, the proteins translated from alternatively spliced mRNAs will contain differences in their amino acid sequence and, often, in their biological functions (Figure 3). This ensures that exons in the same gene are combined in different forms, so that a gene can produce different proteins at different times and in different environments, increasing the complexity or adaptability of the system under varying physiological conditions.

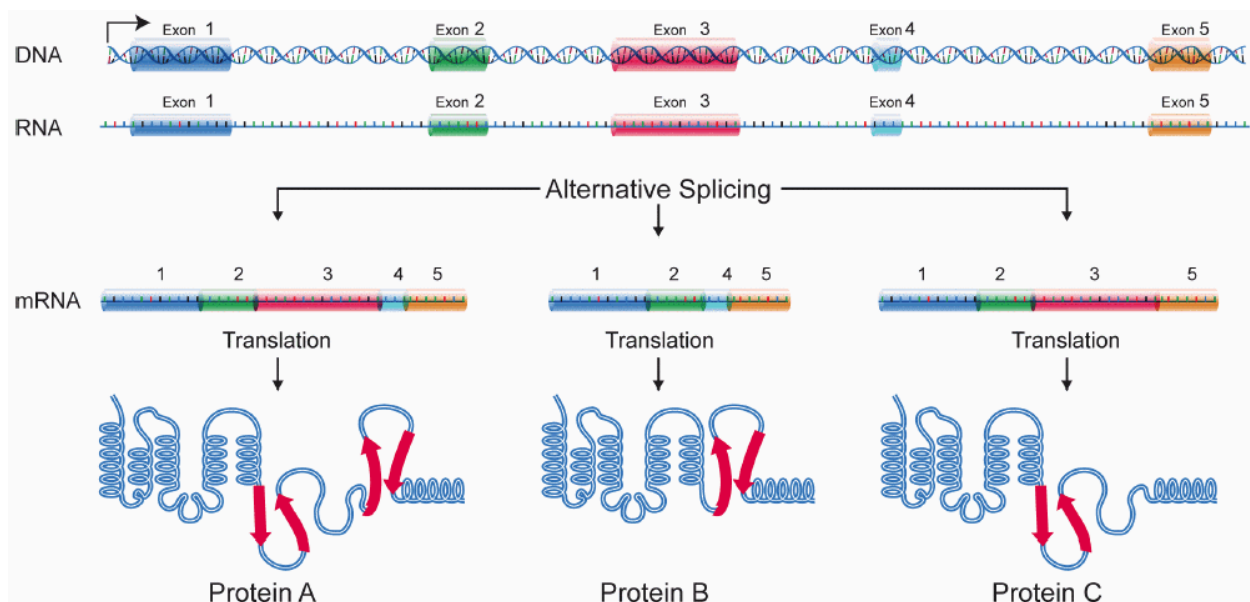


Figure 4. Alternative splicing produces three protein isoforms. Protein A includes all the exons, whereas Proteins B and C result from exon skipping. "DNA alternative splicing.gif" by National

Human Genome Research Institute (US Government), Public Domain (PD-USGov), via Wikimedia Commons.

There are five common alternative splice modes.

Exon skipping in this case, an exon may be spliced out of the primary transcript or retained. This is the most common mode in mammalian pre-mRNAs.

Intron retention: A sequence may be spliced out as an intron or simply retained. This is distinguished from exon skipping because the retained sequence is not flanked by introns. If the retained intron is in the coding region, the intron must encode amino acids in the frame with the neighboring exons, otherwise a stop codon or a shift in the reading frame will cause the protein to be non-functional. This is the rarest splice mode in mammals but the most common in plants.

Mutually exclusive exons: Only one of two exons is retained in mRNAs after splicing.

Alternative donor site: An alternative 5' splice junction (donor site) is used, changing the 3' boundary of the upstream exon.

Alternative acceptor site: An alternative 3' splice junction (acceptor site) is used, changing the 5' boundary of the downstream exon.

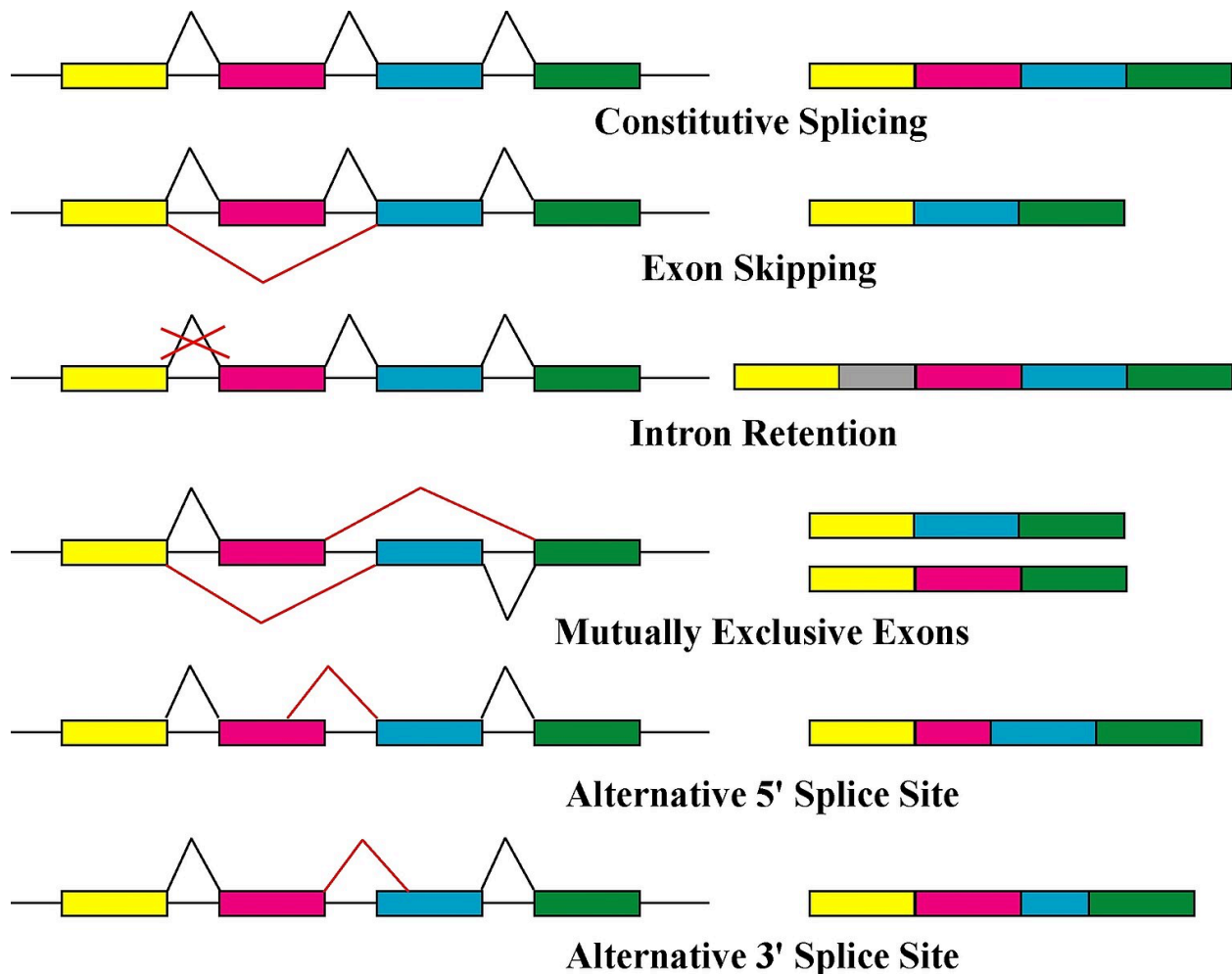


Figure 5. Constitutive splicing and different alternative splicing mode, "DNA alternative splicing.gif" by National Human Genome Research Institute (US Government), Public Domain (PD-USGov), via Wikimedia Commons.

Source: https://commons.wikimedia.org/wiki/File:DNA_alternative_splicing.gif

Lab Exercise I - Identify Gene Elements.

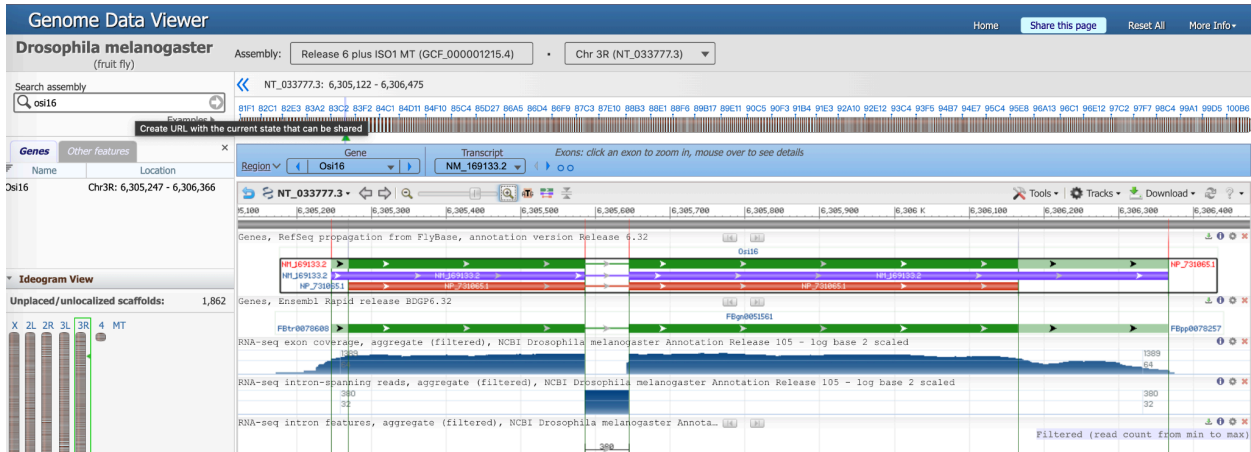
Material

Laptop

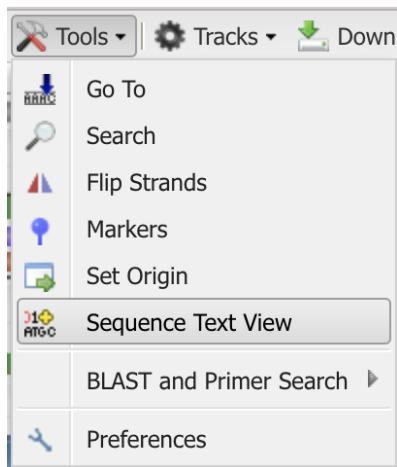
[NCBI genome data viewer](#)

Methods

1. Use the *Osi16* gene as an example. Open NCBI genome data viewer and search for the *Osi16* gene in the fruit fly genome.



2. Click on the Tool icon and choose the Sequence text view



3. Once the text view is ready, move your mouse to locate the *Osi16* gene. Purple highlighted sequences are UTRs. The Green highlighted sequence is the intron. Pink sequences are exons.

```
CGAGATGATGCAGACTCACAACAGTCTCCATTGCAGTAGATAGCGAAGATGTTGCTGACCAAGACCGTGAAATATCTTTTACTTGGCCTTATTCGCC
M L L T K T V K Y L F Y L A L F A
TTCATGTGCAAAATATGCCACGGCTGCCAGTGTCCAGCCCACCGAAGTGCTTGCAGTTGCTCCGGAAACCATTAGAATACCCCAAGGGCAGAGACCTTC
F M C K Y A T A A S V Q P T E V P A V A P E T I R I P Q R A E S L
TCAGTGGCTGTGAAGCATCTCTCCTTCAGTTGGATGTGCCTCAAGATCGAGTTTGTCAAGATTCATGAAAAGCTTGCCGAAACAGGAGGAACCTAACGTTT
L S G C E A S S F S W M C L K I E F V K I M E K L A E Q E E L N V L
GCCAGGAATCAGCGTGTCAAGGATGAGAACCACCCAGCACTCAAGACATCGGAGCTCATGGCAGGTATGTTCTTGAAAAACAAAAATAAGACAAAAA
P G I S V V K D E N A T E L K T S E L M A
ATTTTAACTTTTCGACAATATACAGAAGTTGCTCGCAGCTATCCCAGTGACCCGAGCACTCGCTTGAATGGCTATATAGTAGCCAACTGGAAAACCTTCT
E V A R S Y P S D P S T R L N G Y I V A K L E N L L
GAGAACCGGGTTCTACGATTTCGGCTGCTGGATGACAAGTCCCTCGTGAAGGTCGCAAGCACAAAGTTCGGCAAGAAAGGAGGCTTGAGGCCCTAGTG
R T R F L R F R L L D D K S L V E G R K H K F F G K K G G L E A L V
GCGGCTGGAGTCAATGATGAAGGAATGCTTATGGCCATGGGCCCTCGGGGCCATCGCTCTGATGGCCGGAAGGCTCTGATGACAGCTTAAATGGCCTTGA
A A G V M M K G M L M A M G L L G A I A L M A G K A L M T A L M A L
CCCTGTCTGGAGTCTCGGTCTAAAGAGCTTGGCCGGTGGCGGGGAAAGTCCACCACATACGAAATAGTGGCTAAACCAATTTACACCTCCAGCCACTC
T L S G V L G L K S L A G G G G K S T T Y E I V A K P I Y T S S H S
GCACTCGGTCACTCACGAAGATGGAGGTCACTCCACAGTCCCCACTTTGCCGACGAGGTGGGGGTGGTGGTACTGCCAGTGGCTTTGGCTACGGAGGA
H S V T D G H S H S S P H F A A G G G G G G G T A S G F G Y G G
TATGCGAGGTCCCTGAAAGTGGATCAATCAGCGAACAATACTAATAATTTAGAAAAGAACCACAGTATCATTACCTAATAACTAATTAATTTATTGT
Y A R S L K V D Q S A N K I *
ACTTTAAGAAACACAATCTACGAAACTCAGCAATGTTTGTGTTTATGTGTGAACAGACAGTTAAATAGACGTTGCCGATCTTCCAAATAATCATT
AGTCAGATTTTATTACTTTTAAATATGCCCAAGGTTTTTATAAACGACTATTAAAGCCACAAACGTAAAAAATGTTTAAATTTTATTGAAATAAT
```

4. Screenshot the whole gene text for the assignment 6 report and place sequence element (5' UTR, 3' UTR, Exons, Introns, start codon by a red square, stop by a blue square, splice donor site by red cycle, and splice acceptor by blue cycle) labels on it.

5' UTR

CGAGATGATGCAGACTCACAACAGTCTCCATTGCAGTAGATAGCGAAGATGCTGCTGACCAAGACCGTGAAATATCTTTTACTTGGCCTTATTCGCC

Exon1

Intron

Exon2

3' UTR

TTTCATGTGCAAAATATGCCACGGCTGCCAGTGTCCAGCCCACCGAAGTGCCTGCAGTTGCTCCGGAACCATCCCCAAAGGCGAGAGGCCTTC

F M C K Y A T A A S V Q P T E V P A V A P E T I K I P Q R A E S L

TCAGTGGCTGTGAAGCATCTTCCTTCAGTTGGATGTGCCTCAAGATCGAGTTTGTCAAGATCATGGAAAAGCTTGCCGAACAGGAGGAACCTAACGTTTT

L S G C E A S S F S W M C L K I E F V K I M F K L A E Q E E L N V L

GCCAGGAATCAGCGTGGTCAAGGATGAGAACGCCACCGAACTCAAGACATCGGAGCTCATGGCAGGTGTGTCTTGAAAACACAAA

P G I S V V K D E N A T E L K T S E L M A

ATTTTAACTTTCGACAATATACAGAGTTTGCTCGCAGCTATCCCAGTGACCCGAGCACTCGCTTGAATGGCTATATAGTAGCCAAACTTGAAAACCTTCT

E V A R S Y P S D P S T R L N G Y I V A K L E N L L

GAGAACGCGGTTCTACGATTTTCGCTGTGGATGACAAGTCCCTCGTAGAAGGTGCGAAGCACAAAGTTCGGCAAGAAAAGGAGGCTTGAGGCCCCTAGTG

R T R F L R F R L L D D K S L V E G R K H K F G K K G G L E A L V

GCGGCTGGATCATGATGAAGGAATGCTTATGGCCATGGGCCATCGCTCTGATGGCCGAAAGGCTCTGATGACAGCTTAAATGGCCTTGA

A A G V M M K G M L M A M G I A L M A G K A L M T A L M A L

CCCTGTCTGGAGTCTCGGTCTAAAGAGCTTGGCCGGTGGCGGGGAAAGTCCACCACATACGAAATAGTGGCTAAACCAATTACACCTCCAGCCACTC

T L S G V L G L K S L A G G G G K S T T Y E I V A K P I Y T S S H S

GCACCTCGGTCACTCACGAAGATGGAGTCACTCCACAGTCCCACTTTTGGCGCAGGAGGTGGGGGTGGTGGTACTGCCAGTGGCTTTGGCTACGGAGGA

H S V T H E D G G H S H S P H F A A G G G G G G T A S G F G Y G G

TATGGCAGGTCCTGAAAGTGGATCAATCAGCGAACAATACTAAATTTAGAAAAGAACCACAGTATCATTACCTAATACTAATTAATTTATTTG

Y A R S L K V D Q S A N K I *

ACTTTAAGAAACACAATACTCAGCAAACTCAGCATTGTTTTTTGTTTTATTGTGTGAACAGAACAGTTAAATAGACGTTGCCGATCTTCCAATAATCATTT

AGTCAGATTTTATTACTTTTAATATGCCAAGGTTTTTATAAACGACTATTAAGCCCAAAACGTAAAAAATGTTTTAATTTTTTTTGAAATAAT

Lab Exercise II - Identify Alternative Splicing Modes of a Gene and Explore RNA Expression in Flybase Jbrowser.

Material

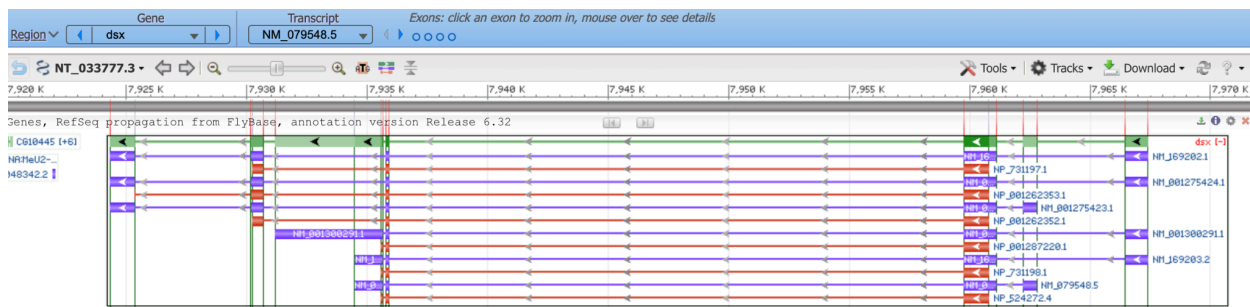
Laptop

[NCBI genome data viewer](#)

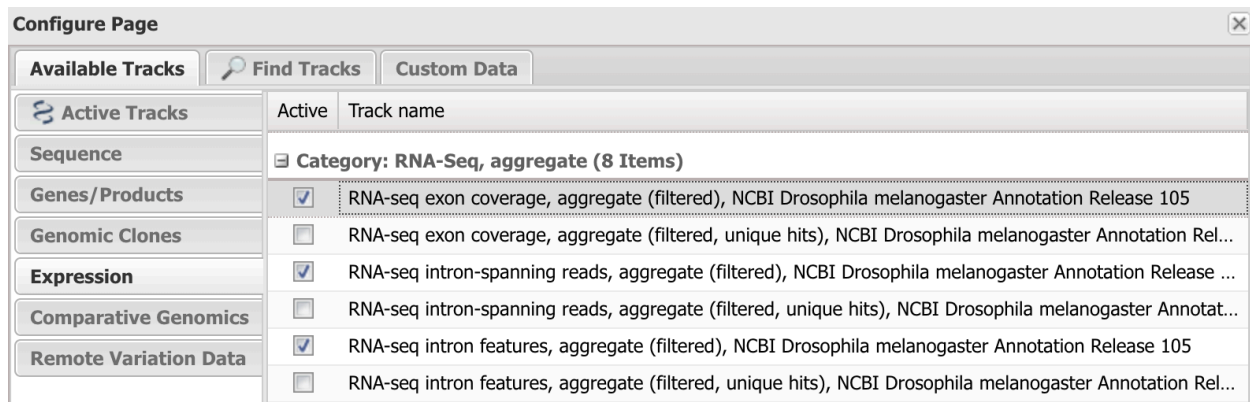
[Flybase Jbrowser](#)

Methods

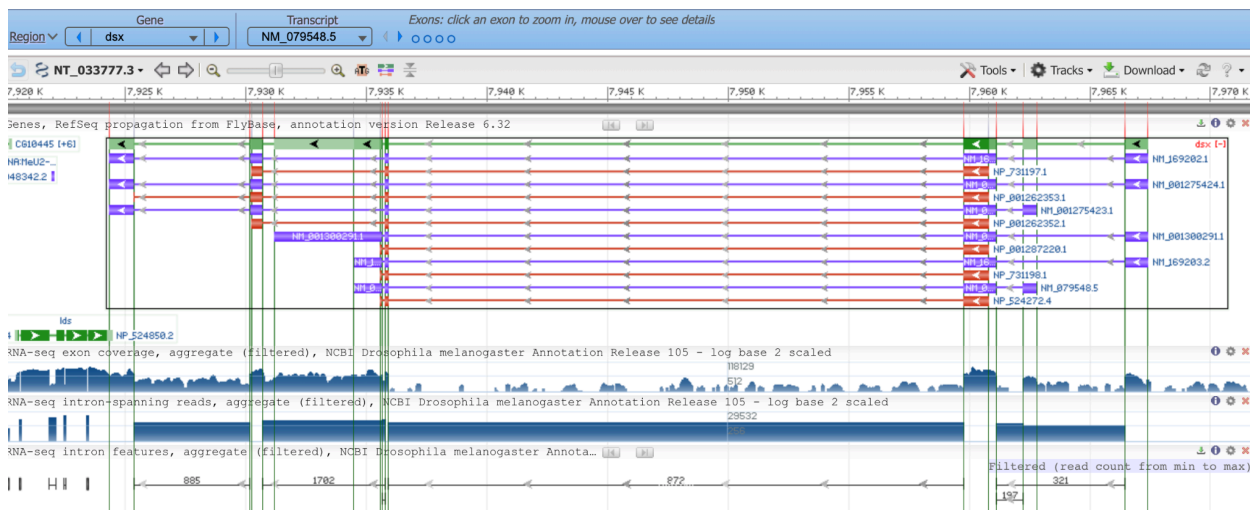
1. Take the *doublesex* gene as an example. Search dsx in NCBI genome data viewer. Expand the green bar of the gene.



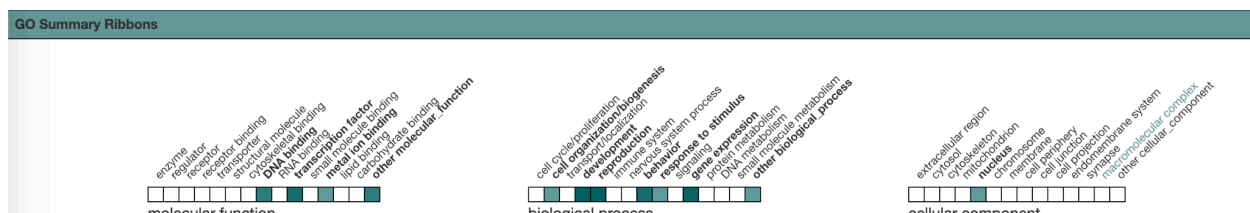
- Click on the Tracks icon and choose "Configure tracks". Make sure you choose "RNA-seq exon coverage", "RNA-seq intron-spanning reads", and "RNA-seq intron features".



- Save the screenshot below for the assignment 6 report. In the report, determine which alternative splicing mode occurs in each isoform of mRNA (ID starts with NM) compared with the green constitutive splicing. For example, the *D. melanogaster* gene *dsx* contains seven exons and reported six isoforms of mRNA. Isoform NM_169202.1 is the exon skipping mode of alternative splicing since it has no exon 3 and 6.



- Continue to use the *dsx* gene as an example. Go to Jbrowser and search for the *dsx* gene. Take a screenshot of the GO summary for the assignment 6 report.



- In the left navigation bar, you can add more tracks to the viewer. According to the GO summary, the *dsx* gene is involved in development and reproduction. It can be found

under the Expression section. Additionally, choose the Reproductive system and developmental stages RNA expression tracks. Screenshot the result for the assignment.



- In assignment 6, write a summary about this gene according to the RNA expression result. For example, the *dsx* gene shows the differential RNA expression pattern between males and females, which is consistent with its function of regulating sexual development.

Lab Exercise III - Comparative genomics in UCSC Genome Browser and Protein Conservation Analysis in CD-search NCBI Tool.

Material

Laptop

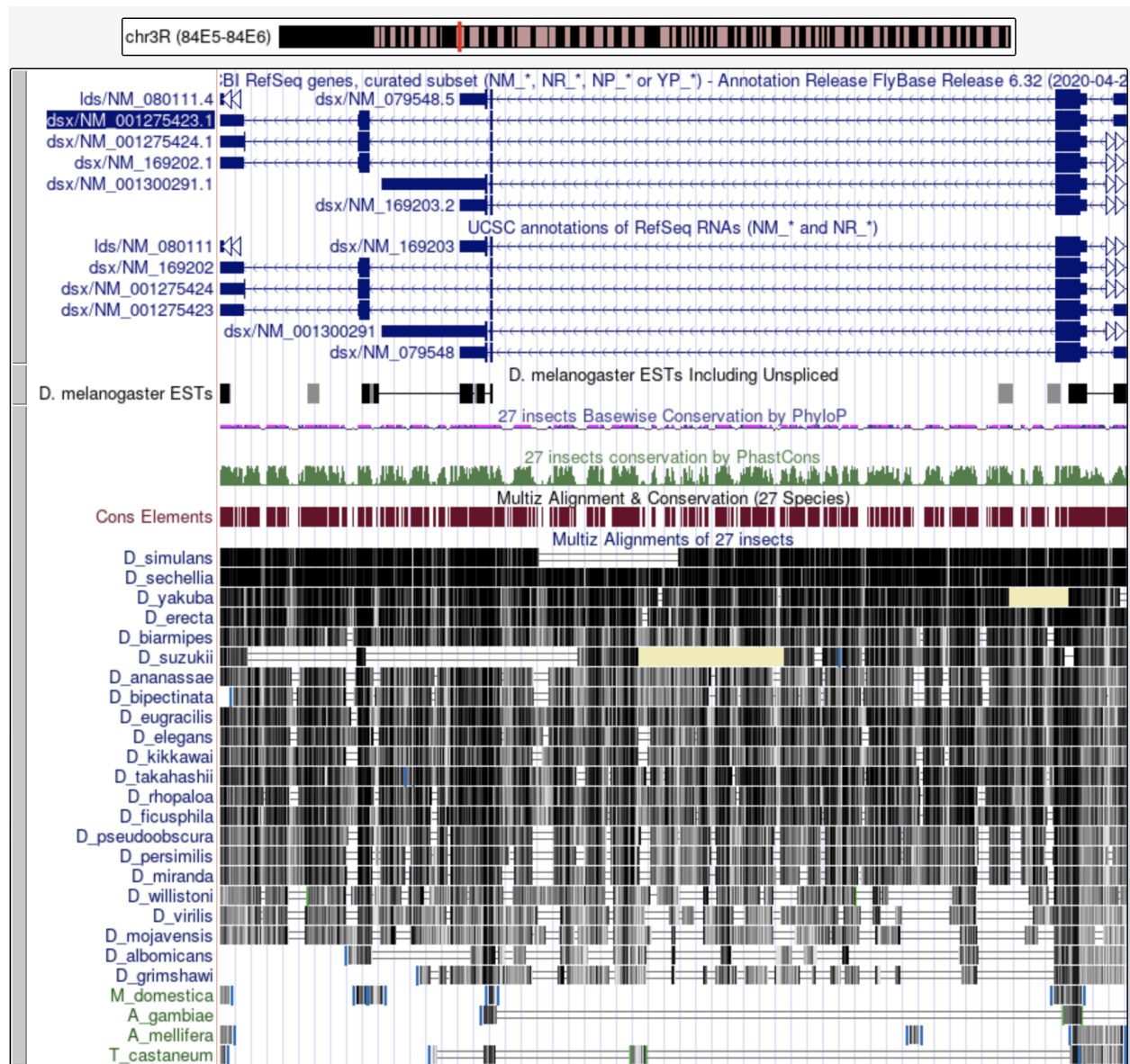
Target gene's protein sequence file was downloaded in lab 5.

[UCSC genome browser](#)

[NCBI Conserved Domains search \(CD-search\)](#)

Methods

1. For comparative genomics, please go to the UCSC genome browser and search the *dsx* gene. Scroll down the page and find track options. Choose only NCBI Refseq, *D. melanogaster* ESTs, and Conservation (pack) tracks. Save the screenshot for assignment 6.



2. Open the NCBI CD-search tool to find the conserved protein domain in a protein sequence. Find the *dsx* protein sequence file and copy and paste only one sequence into the entry box. Use the default parameters set in the options and click on "Submit".

NCBI

HOME SEARCH GUIDE Structure Home 3D Macromolecular Structures Conserved Domains

Search for **Conserved Domains** within a protein or coding nucleotide sequence

Enter **protein** or **nucleotide** query as accession, gi, or sequence in **FASTA format**. For multiple protein queries, use **Batch CD-Search**.

NP_001262352.1 dsx [organism=Drosophila melanogaster] [GeneID=40940] [isoform=D]

Submit Reset Help

OPTIONS

Search against database: CDD v3.19 - 58235 PSSMs

Expect Value threshold: 0.010000

Apply low-complexity filter

Composition based statistics adjustment

Force live search

Rescue borderline hits

Suppress weak overlapping hits

Maximum number of hits: 500

Result mode: Concise Standard Full

3. Here is the result, expand all the domain hits. Screenshot the result and save in the assignment 6 report.

Conserved domains on [lcl|seqsig_MVSEE_991785ce39c4f50b28271b0bf2189f88] View Concise Results

NP_001262352.1 dsx [organism=Drosophila melanogaster] [GeneID=40940] [isoform=D]

Protein Classification

DM and DSX_dimer domain-containing protein (domain architecture ID 10465243)

DM and DSX_dimer domain-containing protein

Graphical summary Zoom to residue level show extra options

Query seq. Specific hits Superfamilies

DM DSX_dimer

Search for similar domain architectures Refine search

List of domain hits

Name	Accession	Description	Interval	E-value
DM	pfam00751	DM DNA binding domain; The DM domain is named after dsx and mab-3. dsx contains a single ...	41-86	3.08e-26
DM DNA binding domain; The DM domain is named after dsx and mab-3. dsx contains a single amino-terminal DM domain, whereas mab-3 contains two amino-terminal domains. The DM domain has a pattern of conserved zinc chelating residues C2H2C4. The dsx DM domain has been shown to dimerize and bind palindromic DNA.				
Pssm-ID: 395608 Cd Length: 47 Bit Score: 100.82 E-value: 3.08e-26				
<pre> 10 20 30 40 ...*... ...*... ...*... ...*... ...*... seqsig_MVSEE_991785ce39c4f50b28271b0bf2189f88 41 PPNCARCRNHGLKITLKGHKRYCKFRYCTCEKRLTADRQRVMAAQ 86 Cdd:pfam00751 2 SPKCARCRNHGVIVPLKGHKRYCPYRDCTCPKCLLVAERQVRVMAAQ 47 </pre>				
DSX_dimer super family	d07436	Doublesex dimerization domain; Doublesex (DSX) is a transcription factor that regulates ...	354-403	9.34e-16
Doublesex dimerization domain; Doublesex (DSX) is a transcription factor that regulates somatic sexual differences in Drosophila. The structure of this domain has revealed a novel dimeric arrangement of ubiquitin-associated folds that has not previously been identified in a transcription factor.				
The actual alignment was detected with superfamily member pfam08828:				
Pssm-ID: 370143 Cd Length: 60 Bit Score: 71.64 E-value: 9.34e-16				
<pre> 10 20 30 40 50 ...*... ...*... ...*... ...*... ...*... seqsig_MVSEE_991785ce39c4f50b28271b0bf2189f88 354 DVFLDYCQKLLKFRYPWELMPLMYVILKADANIEASRIIEARVEIN 403 Cdd:pfam08828 2 DELVKRAQLLEKLYPWEMLPLMYVILKSDGVQKAQRIDEDQAEIH 51 </pre>				

4. Write a summary about the conservation of the gene. For example, the dsx gene is more conserved in *D. sechellia* than in other insects since it has less gaps and a darker color (denoting higher conservation) as shown in the UCSC genome browser. In addition, the dsx conserved domain DM and DSX_dimer locations match the conservation signals in the UCSC genome browser Conservation track.

Lab Exercise IV – Discard F1 Generation from the Lab 5 Crosses (if applicable)

Material

Genetic crosses performed in lab 5

Methods

1. Observe the F1 progeny of the crosses produced in lab 5.
2. Prepare a fly morgue and discard all F1 flies into it
3. Incubate for another week.
4. Record all results in a lab notebook.

Lab Exercise V - Isolate Genomic DNA from Drosophila

Students will be splitted into 4 groups.

Materials

Group A: 5 Wild type flies
Group B: 5 Eb ONLY flies
Group C: 5 Vg ONLY flies
Group D: 5 Eb&Vg double mutant flies
Quick-DNATM Miniprep Kit (Zymo Research)
4 micro pestle with 4 1.5 mL sterile microtubes
1 100-1000 ul micropipettes with 1 box of P1000
1 20-200 ul micropipettes with 1 box of P20-200
1 Centrifuge
1 NanoDrop Spectrophotometer
1 37°C incubator or water bath
1 floating microcentrifuge tube holder

Methods

1. Label the 4 tubes as WT, Eb, Vg, Eb&Vg. Place 5 adult flies accordingly in a 1.5 mL microtube fitted with a micro pestle and thoroughly grind in 500 µL of Genomic Lysis Buffer. Mix gently by inverting the microtube. Incubate for 1 h at 37°C.
2. Centrifuge the lysate at maximum speed (10,000 x g) for 5 min. Making sure not to disturb the pelleted debris, transfer the supernatant to a Zymo-Spin™ IICR Column in a Collection Tube and centrifuge at 10,000 x g for one minute. Discard the Collection Tube containing the flow through.
3. Transfer the Zymo-Spin™ IICR Column to a new Collection Tube. Add 200 µL of DNA Pre-Wash Buffer to the spin column. Centrifuge at 10,000 x g for one minute.
4. Add 500 µL of g-DNA Wash Buffer to the spin column. Centrifuge at 10,000 x g for one minute.

5. Transfer the spin column to a clean microcentrifuge tube. Add 50 μL of DNA Elution **Buffer** or water to the spin column. Incubate for 2–5 min at room temperature and then centrifuge at maximum speed for 30 s to elute the DNA. The eluted DNA can be used immediately for molecular-based applications or stored at $\leq -20^{\circ}\text{C}$ for future use.
6. Quantify the DNA using **NanoDrop Spectrophotometer** and record the quantification on the DNA storage tubes. Use a single drop on the NanoDrop machine and blank the machine using distilled water.
7. **Record the Experimental Procedure and Results in your Lab Notebook.**

Biotechnology and Society Module III: Cancer

Cancer is a disease that is caused by the inability of the body to control the division of cells. Cancer can be caused by genetic or environmental factors. Most cancers are named based on the type of cells or tissues where they originate.

The body's organ systems have specific physiological roles that maintain homeostasis. As the number of cancer cells grows, physiological changes occur in tissues or organs, impacting their role in homeostasis. During development, cells receive genetic information that allows them to become specialized, meaning they have a specific purpose in the body. **DNA** (deoxyribonucleic acid) is a nucleic acid that stores genetic information. It is the order of the nitrogen bases in DNA that stores the genetic information for the cell. The sum of all genetic information in a cell is called the **genome**. **Genes** are units of genetic information that provide instructions for making proteins

Gene expression occurs in 2 stages: **transcription and translation**.

During transcription, within the nucleus, RNA polymerase exposes the nucleotide bases of the DNA to be transcribed. A copy of mRNA is made from the DNA strand. Uracil (U) is used instead of thymine (T). mRNA is edited and leaves the nucleus.

During translation, within the cytoplasm, mRNA moves to a ribosome. tRNA reads codons (3 nucleotides) of the mRNA. For each codon, the corresponding amino acid is brought to the ribosome by tRNA and assembled. The polypeptide is released and the ribosome and mRNA may be used again.

There are **four levels of protein structure**: *Primary structure*: the linear sequence of amino acids. *Secondary structure*: Amino acids interact based on weak chemical interactions. *Tertiary structure*: 3-D molecule, functional. *Quaternary structure*: Multiple protein chains interact to form larger molecules with more complex functions.

The cell cycle describes the organized series of events required for cell division. The cell cycle consists of *Interphase* (orange) – The majority of the cell's time; normal functions occur. The cell cycle consists of G_1 (growth), S (synthesis), G_2 (growth) phases, and *M phase*– the portion where cell division occurs; consists of mitosis and cytokinesis. **Checkpoints** function to regulate the speed at which the cell moves through the cell cycle and ensure the cell is ready for cell

division. If a cell is not ready for cell division, it will be put into the G₀ phase to attempt to repair its DNA. If that is not possible, the cell may undergo **apoptosis** (programmed cell death).

DNA is compacted and organized into **chromosomes** that are found in the nucleus of cells.

Telomeres are repeating sections of nucleotides found at the end of chromosomes. They shorten with each cell cycle. When they are too short for replication, the cell undergoes apoptosis. Humans have 23 pairs of chromosomes. Of each pair, one comes from the mother and one comes from the father. Pairs of chromosomes are **homologous**, meaning they code for similar traits, even though they may have different alleles. During the S phase, DNA replication occurs. Each homologous chromosome is copied, forming a **sister chromatid**. Sister chromatids are held in place by a centromere until the M phase.

Mutations are any changes to the nucleotide sequence of DNA. Mutations can be caused by chemicals (called carcinogens), radiation exposure, or viruses. Mutations may or may not impact protein structure.

Tumor-suppressor genes prevent unregulated cell growth by slowing down the progression of the cell through the cell cycle. Most code for proteins that survey DNA for damage during the G₁ checkpoint. Examples associated with breast cancer include p53 and BRCA1.

Proto-oncogenes speed up the cell cycle. Associated with signaling pathways: growth hormones interact with target cell receptors, proto-oncogenes are activated, and cell division is promoted. Proto-oncogenes become oncogenes when mutations cause them to be “stuck”. Only a single mutation is required for this. Examples associated with breast cancer include *ERBB2*.

Cancer cells are different from typical cells as they look different (irregular shape, larger nuclei with more genetic material), are immortal (because of the telomerase enzyme that repairs telomeres), and lose the ability to specialize.

Masses of unspecialized cells are called **tumors**. They can be **benign or malignant**. As tumor cells grow, they require oxygen and nutrients. **Angiogenesis** occurs as growth factors are released and capillaries grow toward the tumor. Capillaries (red) and lymphatic vessels (green) may carry tumor cells to other parts of the body where angiogenesis may occur. This spread of cancer to other parts of the body is called **metastasis**.

Mutations in cancer-causing genes can be tested for. After blood is drawn, DNA is isolated from white blood cells and subjected to a **polymerase chain reaction (PCR)** in order to make copies of the gene of interest. Once copies of a gene have been made, the exact sequence of nucleotides is determined through **DNA sequencing**. Once sequenced, the nucleotide order can be compared with known sequences to determine if mutations are present.

DNA microarrays, or gene chips, can screen genetic information, allowing doctors to look for specific mutations or differences in gene expression between normal cells and cancer cells. Synthetic DNA that corresponds to known cancer-related mutations is placed on a chip. mRNA is isolated from cells and used to create fluorescently labeled complementary DNA (cDNA).

Once hybridized with the synthetic DNA on the chip, the cDNA will bind to its complementary match, indicating that those cancer-related genes are being expressed.

Blood tests called **complete blood counts (CBCs)** look at the presence of normal blood cells and basic blood chemistry. Cancers may elevate the number of white blood cells and tumor cells may even be present in the blood. Changes in blood calcium, enzymes, or ion concentrations may also indicate that cancer is interrupting the function of an organ. Specific proteins associated with cancer can be tested for in blood, including CA 15-3 and CA 125.

Chemotherapy is the use of drugs to treat disease. In the case of cancer, chemotherapy works by targeting the cell cycle to prevent cell division in rapidly dividing cells. Chemotherapy affects all cells of the body, so it can be effective in treating and preventing metastasis, but also has many negative side effects.

Radiation is the movement of energy in waves or particles. The shorter the wavelength, the more energy is present. Radiation therapy uses high-energy gamma rays and X-rays or particles, including alpha and beta particles. Radiation therapy damages the DNA within a cell, slowing down or stopping cell division. It can target specific areas but isn't as effective at metastasized cancer cells.

The immune system functions to identify "self" vs. "nonself" and then target any "nonself" cells for destruction. **Immunotherapy** involves convincing the body's immune system to identify and eliminate cancer cells as "nonself" instead of "self." It involves identifying antigen differences between "self" and cancer cells. Cancer vaccines use modified viruses to initiate an immune response against cancer cells. Monoclonal antibodies can be created to target specific cancer antigens.

Genome editing targets specific sequences in DNA for editing. **Clustered regularly interspaced short palindromic repeats (CRISPR)** are most commonly used. Enzyme Cas9 uses an RNA molecule to identify and bind to specific complementary sequences of DNA. PAM regulatory regions help make sure Cas9 has the correct sequence. CRISPR can remove nucleotides, inactivate the gene, or insert new nucleotides, changing the function of the gene. CRISPR has been used in cancer treatments.

Pre-symptomatic gene testing is performed on an individual who does not show symptoms of a disorder but is at risk of developing it. Pre-symptomatic DNA testing for autosomal dominant hereditary breast/ovarian cancer (HBOC) became an option after the identification of the *BRCA1* and *BRCA2* genes in 1994–1995. Healthy mutation carriers have a high lifetime risk for breast (56–87%) or ovarian cancer (10–60%) and may opt for intensive surveillance or prophylactic surgery (mastectomy/oophorectomy).

**Students will use the material in the "Biotechnology and Society" modules to complete the "Biotechniques Reflection Assignment" for this course.*

Assignment 6

1. Complete **Lab exercise I** with the **Bar** gene(search for B-H1),The B-H1 gene has two protein isoforms, choose the shorter one (B-H1/NM_001272713.1/NP_001259642.1) to label and save the results.

Screenshot the whole gene text for the assignment 6 report and place sequence elements labels:

- Annotation: Mark 5' UTR, 3' UTR, Exons, Introns
- Annotation: Mark the start codon with a **red square**, stop with a **blue square**, the splice donor site with a **red circle**, and the splice acceptor with a **blue circle**)

2. Complete **Lab Exercise II** using the vestigial and ebony genes, and **save the screenshots** from [NCBI genome data viewer](#) & [Flybase Jbrowser](#). **Write a summary** for each gene searched (vestigial and ebony), including **alternative splicing mode** for each isoform and the differential RNA expression patterns, and how this information correlates with their biological functions.

3. Complete **Lab Exercise III** for the ebony gene and **save the screenshots** from [UCSC genome browser](#) & [NCBI Conserved Domains search \(CD-search\)](#). **Write a summary** about the conservation of the gene. **Also, write a summary** of your findings using the NCBI CD-search.

Ebony Protein Sequence:

MGSLPQLSIVKGLQQDFVPRALHRIFEEQQLRHADKVALIYQPSTTGQGMAPSQSSYR
QMNERANRAARLLVAETHGRFLQPNSDGDFIVAVCMQPSEGLVTTLLAIWKAGGAYLPI
DPSFPANRIHHILLEAKPTLVIRDDIDAGRFGQTPTLSTTELYAKSLQLAGSNLLSEML
RGGNDHIAIVLYTSGSTGVPKGVRPLPHESILNRLQWQWATFPYTANEAVSVFKTALTFFV
DSIAELWGPLMCGLAIVVPKAVTKDPQRLVALLERYKIRRLVLVPTLLRSLLMYLKMEG
GGAAQKLLYNLQIWVCSGEPLSVSLASSFFDYFDEGVHRLYNFYGSTEVLDGVTYFAC
ESKKQLSLYDNVPIGIPLSNTVVYLLDADYRPVKNGEIGEIFASGLNLAAGYVNGRDPER
FLENPLAVEKKYARLYRTGDYGSLKNGSIMYEGRTDSQVKIRGHRVDLSEVEKNVAELP
LVDKAIVLCYHAGQVDQAILAFVKLRDDAPMVTEMQMEARLKDKLADYMTQPQVVILEHV
PLLVNGKVDRQALLKTYETANNNEGDSSIVLDFDYSQVPEDLKLTARDLFETVGGVIGR
STRATLAPHSNFYELGGNSLNSIFTVTLLREKGYNIGISEFIAAKNLGEIIEKMAANHDAV
QLEESLNACPHLKMEAVPLRLEHRQEVIDIIVASFYNKADLEQWLKPGVLRDYSIDILN
DIWNVLVERDLSFVVYDTNTDRIIGTALNFDARNEPEVDIKSKLLIVFEFLEFCEGPIRDN
YLPKGLNQLHSFMMGTAEKLNPRENIACMHFMEHEVLRVAREKQFAGIFTTNTSPLTQ
QLADVYHYKTLLNFQVNEYVHSDGSRPFGDAPDEQRAIVHWKEVGK

Assignment 6 Instruction video:

https://cuny907-my.sharepoint.com/personal/hanqi_ye06_login_cuny_edu/_layouts/15/stream.aspx?id=%2Fpersonal%2Fhanqi%5Fye06%5Flogin%5Fcuny%5Fedu%2FDocuments%2FSCB252%20videos%2FLab6%2Dvideo%20demo%2Emp4&ga=1

Lab 7 PCR Primer design and PCR amplification

Hanqi Ye

Objective

- To become familiar with the principle of designing a primer
- To practice primer design and validation using NCBI primer BLAST
- To become familiar with FlyPrimerBank.
- This lab provides basic information about genomic DNA and technology to isolate *Drosophila* genomic DNA

Introduction to PCR

PCR stands for Polymerase Chain Reaction. It's a laboratory method that makes many copies of a specific DNA sample. PCR is based on DNA polymerase's ability to create a new DNA strand that's complementary to the template strand. The amplification procedure involves: Denaturing, Annealing, Extension. It is also called a Thermal cycler which exposes reagents to repeated cycles of heating and cooling to permit different temperature-dependent reactions—specifically, DNA melting and enzyme-driven DNA replication.

PCR was invented in 1983 by American biochemist Kary Mullis at Cetus Corporation. Mullis and biochemist Michael Smith, who had developed other essential ways of manipulating DNA, were jointly awarded the Nobel Prize in Chemistry in 1993. The evolution of PCR has led to a rich tapestry of techniques, each tailored to specific research needs. Most popular ones like real-time quantitative PCR (qPCR) and reverse transcription PCR (RT-PCR), these variants continue to shape molecular biology laboratories worldwide.

PCR principle and procedure

PCR, or Polymerase Chain Reaction, is a technique used to magnify a specific region of a DNA strand, known as the DNA target. The majority of PCR methods can amplify DNA fragments ranging from 0.1 to 10 kilo base pairs (kbp) in length, although some methods can amplify fragments up to 40 kbp. The quantity of the amplified product is dictated by the substrates available in the reaction, which become limiting as the reaction progresses.

The basic setup for a PCR requires several components and reagents, including:

- A DNA template that includes the DNA target region to be amplified
- A DNA polymerase, an enzyme that creates new DNA strands. Taq polymerase, which is heat-resistant, is particularly common as it is more likely to stay intact during the high-temperature DNA denaturation process
- Two DNA primers that are complementary to the 3' (three prime) ends of each of the sense and antisense strands of the DNA target. DNA polymerase can only bind to and elongate from a double-stranded region of DNA, so without primers, there is no

double-stranded initiation site for the polymerase to bind. Specific primers that are complementary to the DNA target region are chosen in advance and are often custom-made in a lab or bought from commercial biochemical suppliers

- Deoxynucleoside triphosphates, or dNTPs (sometimes referred to as “deoxynucleotide triphosphates”; these are nucleotides that contain triphosphate groups), are the building blocks that the DNA polymerase uses to create a new DNA strand
- A buffer solution that provides a suitable chemical environment for the DNA polymerase to function optimally and remain stable
- Bivalent cations, typically magnesium (Mg) or manganese (Mn) ions. Mg^{2+} is the most common, but Mn^{2+} can be used for PCR-mediated DNA mutagenesis, as a higher Mn^{2+} concentration increases the error rate during DNA synthesis. Monovalent cations, typically potassium (K) ions, are also needed.

The reaction is usually carried out in a volume of 10–200 μ L in small reaction tubes (0.2 – 0.5 mL volumes) in a thermal cycler (Figure 7). The thermal cycler heats and cools the reaction tubes to achieve the temperatures required at each step of the reaction. Many modern thermal cyclers use a Peltier device, which allows both heating and cooling of the block holding the PCR tubes simply by reversing the electric current of the device. Thin-walled reaction tubes allow for favorable thermal conductivity, enabling rapid thermal equilibrium. Most thermal cyclers have heated lids to prevent condensation at the top of the reaction tube. Older thermal cyclers that lack a heated lid require a layer of oil on top of the reaction mixture or a ball of wax inside the tube.

PCR involves a series of 20-40 thermal cycles, each consisting of two or three temperature steps. The process begins with an initialization step, which is necessary for DNA polymerases that require heat activation. This is followed by the denaturation step, where the DNA template is heated to break the hydrogen bonds between complementary bases, resulting in two single-stranded DNA molecules.

Next is the annealing step, where the temperature is lowered to allow the primers to bind to each of the single-stranded DNA templates. The temperature for this step is critical as it affects the efficiency and specificity of the process.

The extension/elongation step follows, where the DNA polymerase synthesizes a new DNA strand complementary to the DNA template strand. The number of DNA target sequences doubles at each extension/elongation step, leading to exponential amplification of the specific DNA target region.

The process of denaturation, annealing, and elongation constitutes a single cycle. Multiple cycles are required to amplify the DNA target to millions of copies.

After the cycles, a final elongation step is performed to ensure that any remaining single-stranded DNA is fully elongated. The final hold step cools the reaction chamber for short-term storage of the PCR products. The number of DNA copies formed after a given

number of cycles is calculated using the formula 2^n , where n is the number of cycles. For example, a reaction set for 30 cycles results in 2^{30} , or 1,073,741,824, copies of the original double-stranded DNA target region.

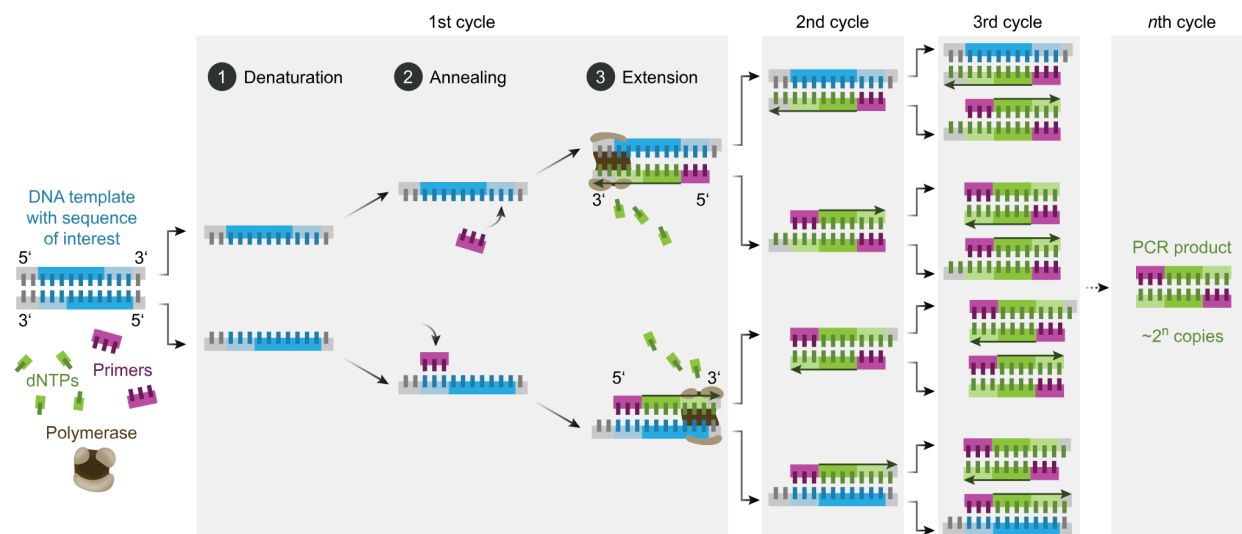


Figure 7. Schematic mechanism of the polymerase chain reaction (PCR) showing denaturation, annealing and extension phases. (*"Polymerase chain reaction-en.svg"* by Enzoklop, licensed under CC BY-SA 4.0, via Wikimedia Commons.)

It's crucial to address the frequent misunderstanding that RT-PCR, qPCR, and RT-qPCR are interchangeable. The similarities among these closely related techniques often lead to incorrect usage of the acronyms. To prevent this, the Minimum Information for Publication of Quantitative Real-Time PCR Experiments (MIQE) guidelines, first published in 2009, proposed a standardization of abbreviations. They suggested that 'RT-PCR' should only be used to denote reverse transcription PCR, not real-time PCR, as is often misunderstood.

Reverse transcription PCR (figure 8A) uses RNA as a template to create complementary DNA (cDNA). A single-stranded copy of cDNA is produced using the reverse transcriptase enzyme. This can then be amplified by a DNA polymerase, creating double-stranded cDNA, which feeds into a standard PCR-based amplification process. This technique can be used in molecular cloning of genes of interest (GOIs), but it most commonly serves as the first step in RT-qPCR.

According to MIQE, the term 'qPCR' describes quantitative real-time PCR (figure 8B), which is the PCR amplification of DNA in real time, measured by a fluorescent probe, most commonly an intercalating dye or a hydrolysis-based probe, enabling quantification of the PCR product. This technique is used to detect the presence of pathogens and to determine the copy number of DNA sequences of interest.

The final acronym 'RT-qPCR' is used for reverse transcription quantitative real-time PCR (figure 8C). This is a technique that combines RT-PCR with qPCR to enable the measurement of RNA levels through the use of cDNA in a qPCR reaction, thus allowing rapid detection of gene expression changes.

Despite these standardized abbreviations, it's important to note that this nomenclature guideline is not always followed, and qPCR is commonly used to describe RT-qPCR. Similarly, RT is used to denote real-time PCR rather than reverse transcription, thus causing confusion over which method is being described.

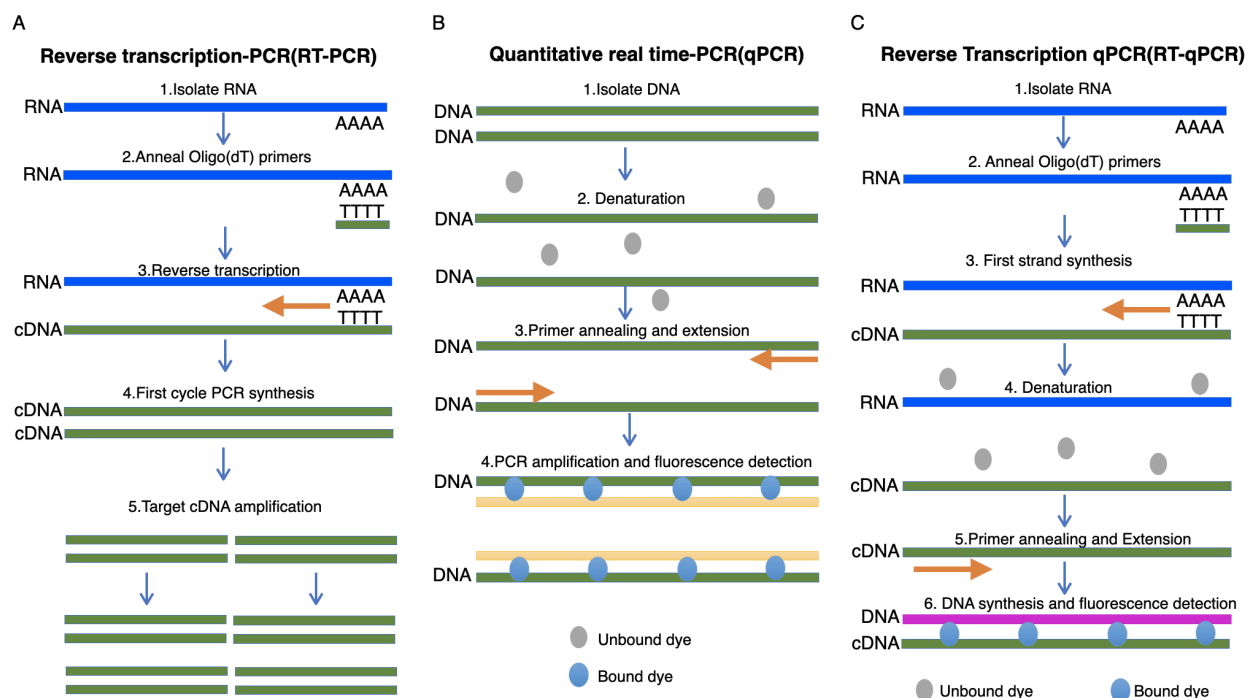


Figure 8. Schematic comparing RT-PCR, qPCR and RT-qPCR. (A) RT-PCR workflow. RNA is isolated and cDNA is generated via reverse transcription (RT); PCR is then carried out to amplify areas of interest. (B) qPCR schematic. DNA is isolated and amplified; amplification is quantitated using a probe which fluoresces upon intercalation with double-stranded DNA. (C) RTqPCR procedure. RNA is isolated and cDNA generated before commencing a qPCR procedure.

In RT-PCR, an RNA template is first converted into a complementary DNA (cDNA) using a reverse transcriptase (RT). This cDNA then serves as a template for exponential amplification via PCR. The application of RT-PCR for RNA transcript detection has significantly advanced gene expression studies by:

- Making it theoretically possible to detect the transcripts of virtually any gene
- Enabling sample amplification, thereby eliminating the need for abundant starting material required in northern blot analysis
- Providing tolerance for RNA degradation, as long as the RNA spanning the primer remains intact

Quantification of mRNA using RT-PCR can be performed as either a one-step or a two-step reaction. The two-step reaction necessitates that the reverse transcriptase reaction and PCR amplification be conducted in separate tubes, which increases the risk of contamination due to more frequent sample handling. Conversely, the one-step approach carries out the entire reaction, from cDNA synthesis to PCR amplification, in a single tube. This method is believed to minimize experimental variation by conducting all enzymatic reactions in a single environment and eliminates the labor-intensive and contamination-prone steps of pipetting cDNA product to PCR reaction.

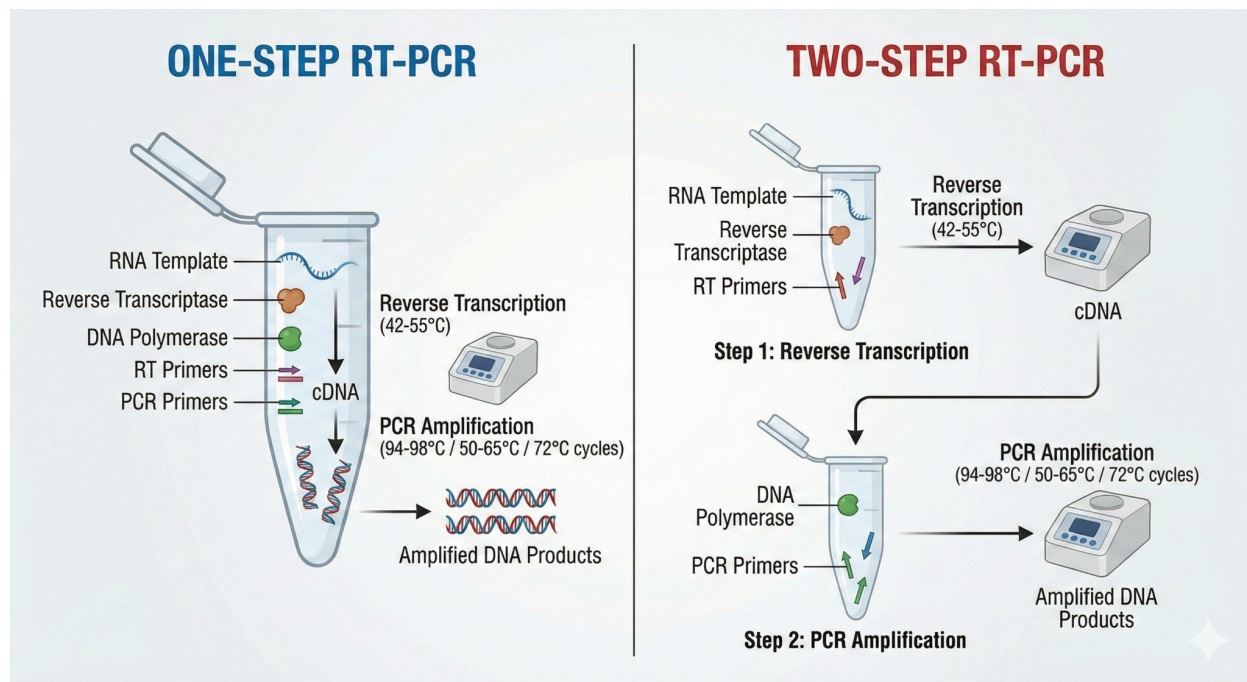


Figure 9. one-step RT-PCR and two-step RT-PCR

The one-step approach supports the reverse transcription of RNA from unpurified or crude samples, such as whole blood and serum, with the use of inhibitor-tolerant polymerases and polymerase enhancers under optimized one-step RT-PCR conditions. However, the starting RNA templates are susceptible to degradation in the one-step approach, and this method is not recommended when repeated assays from the same sample are required. Additionally, the one-step approach is reported to be less accurate compared to the two-step approach. It is also the preferred method of analysis when using DNA binding dyes like SYBR Green, as the elimination of primer-dimers can be achieved through a simple change in the melting temperature. Despite these considerations, the one-step approach offers a relatively convenient solution for the rapid detection of target RNA directly in biosensing.

Real-time PCR (qPCR) is conducted in a thermal cycler that has the ability to shine a light of at least one specific wavelength on each sample and detect the fluorescence emitted by the

excited fluorophore. The thermal cycler can also swiftly heat and cool samples, making use of the physicochemical properties of nucleic acids and DNA polymerase.

Due to the small size of the fragments, the last step is often omitted in this type of PCR as the enzyme can replicate the DNA amplicon during the transition between the alignment stage and the denaturing stage. In four-step PCR, the fluorescence is measured during short temperature phases lasting only a few seconds in each cycle, at a temperature of, for example, 80 °C, to reduce the signal caused by the presence of primer dimers when a non-specific dye is used.

Quantitative PCR, with or without a reverse transcription step, is a standard tool in molecular biology laboratories and has significantly transformed research practices due to its relatively straightforward process. Its benefits over traditional PCR include the ability to monitor reactions in real time and eliminate the need for an agarose gel. It also enables truly quantitative analysis. One of the most frequent applications of qPCR is to determine the copy number of a specific DNA sequence. By using absolute quantification, users can determine the target copy numbers relative to a standard curve of defined concentration, providing a much more accurate method than previously possible.

Quantitative reverse transcription polymerase chain reaction, also called RT-qPCR, is used to detect and quantify RNA. Total RNA or mRNA is first transcribed into complementary DNA (cDNA). The cDNA is then used as the template for the quantitative PCR or real-time PCR reaction (qPCR). In qPCR, the amount of amplification product is measured in each PCR cycle using fluorescence. RT-qPCR is used in a variety of applications including gene expression analysis, RNAi validation, microarray validation, pathogen detection, genetic testing, and disease research. RT-qPCR enables the study of changes in gene expression following the treatment of model systems with inhibitors, stimulants, small interfering RNAs (siRNAs), or knockout models, among others. This technique is also routinely used to detect changes in expression both before (as a quality control measure) and after (to confirm changes) RNA-Seq experiments.

There are two methods of detection in real-time PCR (qPCR): non-specific detection with double-stranded DNA-binding dyes as reporters, and specific detection using fluorescent reporter probe methods.

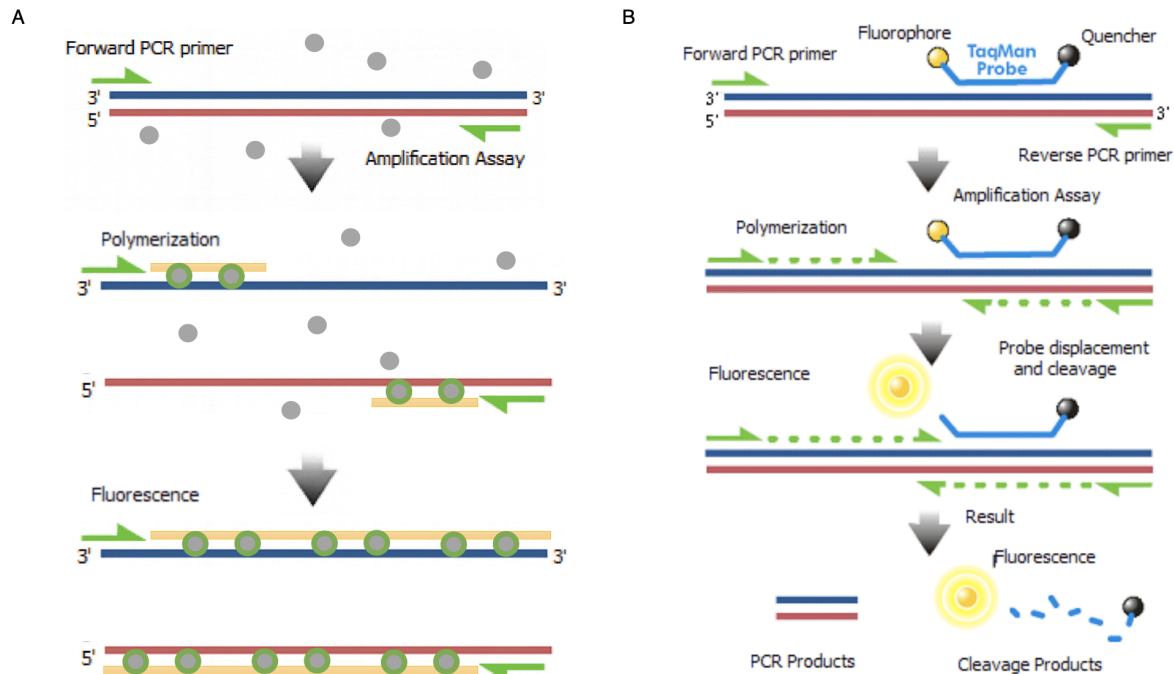


Figure 10 (A) Non-specific detection with double-stranded DNA-binding dyes such as SYBR Green and (B) specific detection using fluorescent reporter probe methods using TaqMan.

In non-specific detection, a DNA-binding dye binds to all double-stranded DNA in PCR, leading to an increase in fluorescence intensity with each cycle. However, these dyes, such as SYBR Green, will bind to all double-stranded DNA PCR products, including non-specific ones, which can interfere with accurate monitoring of the intended target sequence.

In specific detection, fluorescent reporter probes are used that only detect the DNA containing the sequence complementary to the probe (figure 10). This increases specificity and allows the technique to be performed even in the presence of other double-stranded DNA. The method involves a DNA-based probe with a fluorescent reporter at one end and a quencher of fluorescence at the opposite end. The close proximity of the reporter to the quencher prevents detection of its fluorescence. During PCR, the probe is degraded by the Taq polymerase and the fluorescent reporter is released, resulting in an increase in fluorescence. This increase is used to determine the quantification cycle in each reaction. This method, however, does not prevent the inhibitory effect of primer dimers, which may depress accumulation of the desired products in the reaction.

Introduction - PCR Primers

PCR is a standard technology in biological laboratories. Biotechniques such as gene quantification, cloning, assembly, mutation, and sequencing depend on PCR. At present,

PCR-related reagents, consumables, and equipment have been well developed, including the optimization of polymerases and buffers, which have greatly improved the amplification capability, fidelity, and compatibility of the process. Therefore, optimization of the PCR reaction, including evaluations of thermal cycle conditions, substrate concentrations, and others is no longer required. In other words, all the PCR experimental conditions can be standardized, without having to re optimize for each experiment. However, for the PCR experiment, the primer design must be performed entirely by the researcher without aid from the PCR reagents manufacturer.

PCR primers are short sequences of single-stranded DNA, usually approximating 20 nucleotides in length. Two primers are used in each PCR reaction, and they are designed to flank the target region (a region that will be amplified). That is, they are given sequences that will make them bind to opposite strands of the template DNA at the edges of the region to be copied. The primers bind to the template by complementary base pairing rules. When the primers are bound to the template, they can be extended by the polymerase, and the region that lies between them will get copied.

PCR primers are made by combining nucleotides in a chemical reaction. When adding a nucleotide one at a time, the reactive groups must be deliberately blocked and unblocked frequently. Primers' key characteristic is that they must match sequences on the template molecule (must be complementary to the template strand).

The melting temperature, or T_m , is a critical parameter to consider when designing and performing many molecular biology experiments, including PCR and qPCR. The T_m of an oligonucleotide duplex is the temperature at which half of the oligonucleotide molecules are single-stranded and half are double-stranded, i.e., the oligonucleotide is 50% annealed to its exact complement. Primers with melting temperatures in the range of 52-58°C generally produce better results than primers with lower melting temperatures. While the annealing temperature can go as high as 72°C, primers with melting temperatures above 65°C have a higher potential for secondary annealing.

Accurate prediction of T_m identifies duplexes that are likely to form at specific temperatures, allowing you to determine appropriate thermal cycling parameters. During the annealing phase of PCR, the reaction temperature needs to be sufficiently low to allow both forward and reverse primers to bind to the template, but not so low as to enable the formation of undesired, non-specific duplexes or intramolecular hairpins, both of which reduce reaction efficiency. Both primers in PCR should be chosen to have a similar T_m . IDT recommends selecting an annealing temperature 5°C below the lowest primer T_m .

Primer length and sequence are of critical importance in designing the parameters of a successful amplification. The melting temperature of a nucleic acid duplex increases both with its length and with increasing GC content. A simple formula for calculation of the (T_m) is:

$$T_m = 4(G + C) + 2(A + T) \text{ } ^\circ\text{C}$$

The actual T_m is influenced by the concentration of Mg^{2+} , K^+ , and cosolvents. There are numerous computer programs to assist in primer design. The formula given above for T_m is simplistic; there are many primer design programs which use more complex nearest-neighbor thermodynamics values.

Thus, different target genes require specific primer pairs. The first exercise of this lab will use one gene (*eyeless*) as an example and teach you to design primers for PCR experiments by using primer BLAST in NCBI. In the following exercises, we will briefly describe how to validate primers using the same tool and you will practice downloading primers from FlyprimerBank and validating them.

Lab Exercise I - PCR Primer design.

Material

Student needs to bring their own laptop

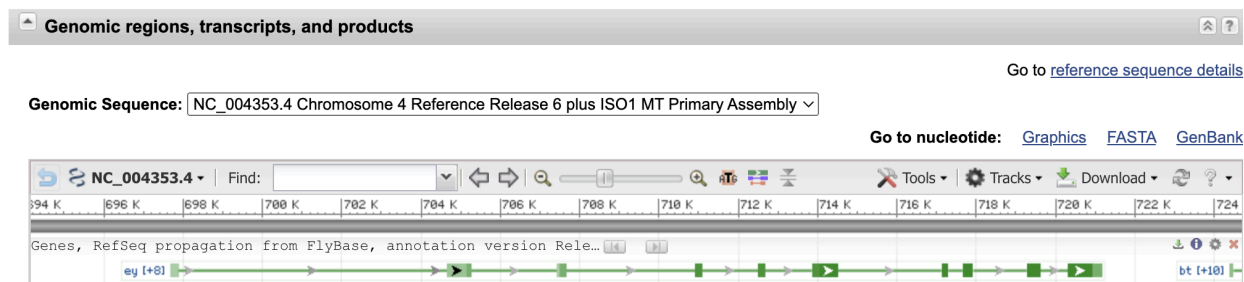
NCBI website (<https://www.ncbi.nlm.nih.gov/>)

NCBI Primer-BLAST (<https://www.ncbi.nlm.nih.gov/tools/primer-blast/index.cgi>)

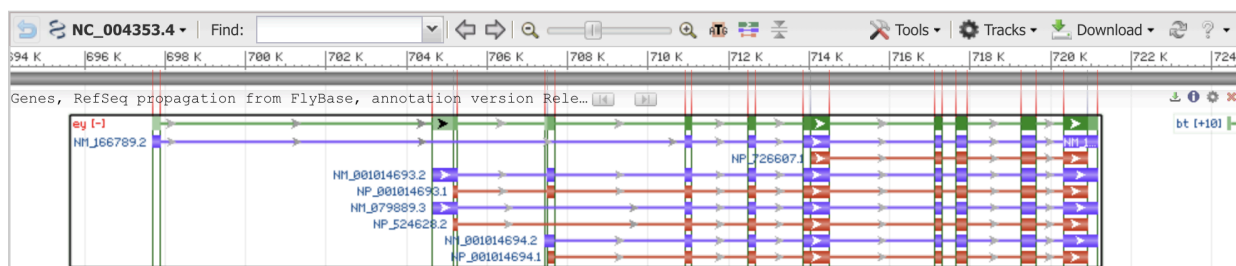
Methods

A) Design a primer for a specific transcript.

1. Open NCBI and search for the gene *eyeless* (*ey*). Choose the gene and scroll down the result page, find the genomic viewer, and click on the key [+8] green bar shown below.



2. More information will show up including four transcript variants for this gene and four protein variants. (NM is mRNA, NP is protein, NC is a nucleotide.)



3. Continue scrolling down and find **NCBI Reference Sequences (RefSeq)**, **mRNA**, and **protein** sections. Specific information about all isoforms will be displayed if there is more than one mRNA transcript. In this case, there are A to D *eyeless* transcripts.

mRNA and Protein(s)

1. [NM_166789.2](#) → [NP_726607.1](#) eyeless, isoform B [Drosophila melanogaster]

[See identical proteins and their annotated locations for NP_726607.1](#)

Status: REVIEWED

4. Using isoform B as an example, this primer is designed specifically for amplifying eyeless isoform B cDNA during the PCR. click on the hyperlink [NM_166789.2](#) to go to the following page.

Drosophila melanogaster eyeless (ey), transcript variant B, mRNA

NCBI Reference Sequence: NM_166789.2

[FASTA](#) [Graphics](#)

Go to: 

LOCUS NM_166789 2854 bp mRNA linear INV 22-APR-2020
 DEFINITION Drosophila melanogaster eyeless (ey), transcript variant B, mRNA.
 ACCESSION NM_166789
 VERSION NM_166789.2
 DBLINK BioProject: [PRJNA164](#)
 BioSample: [SAMN02803731](#)
 KEYWORDS RefSeq.

Customize view 

Analyze this sequence 

Run BLAST

Pick Primers

Highlight Sequence Features

Find in this Sequence

Show in Genome Data Viewer

5. Click the “Pick Primers” on the right side under Customize the view to enter the Primer BLAST page. Once you see the page, the PCR Template has been filled out for you. For now, leave everything else blank and change the maximum value for “**PCR product size**” to 350 bp. In qPCR reactions, the extension time generally does not exceed 30 s. According to the conventional Taq enzyme reaction rate of 1 Kb /min, 30 s can synthesize 500 bp. However, the enzyme activity will decrease with the prolongation of the reaction time. Therefore, it is more appropriate to set the maximum product length as 350 bp or even lower to be conservative. An overly short product cannot be separated from the primer dimer. Therefore, the lower limit is 70 bp. If you cannot find any suitable primers, then increase this number, preferably below 500 bp, or the qPCR reaction conditions (extension time) will have to be changed.

Primer-BLAST

A tool for finding specific primers

Finding primers specific to your PCR template (using Primer3 and BLAST).

Primers for target on one template

Primers common for a group of sequences

Retrieve recent results

Publication

Tips for finding specific primers

PCR Template

Enter accession, gi, or FASTA sequence (A refseq record is preferred) ?

Clear

NM_166789.2

Range ?

Clear

From

To

Forward primer

Reverse primer

Or, upload FASTA file

选择文件

未选择任何文件

Primer Parameters

Use my own forward primer (5'->3' on plus strand)

Use my own reverse primer (5'->3' on minus strand)

PCR product size

of primers to return

Primer melting temperatures (T_m)

Min

Max

Min

Opt

Max

Max T_m difference

Clear

Clear

70

1000

10

57.0

60.0

63.0

3

?

- The “**Exon junction span**” option is changed to “Primer must span an exon-exon junction”. This means that at least one of the primers crosses the junction between the exons. This way, even if there is genomic DNA contaminating the cDNA, it will not be amplified, ensuring accurate quantification. Nonetheless, if a gene does not have introns, changing this option will result in an error and another option must be chosen.

Exon/intron selection

A refseq mRNA sequence as PCR template input is required for options in the section ?

Exon junction span

Exon junction match

Intron inclusion

Intron length range

✓ No preference

Primer must span an exon-exon junction

Primer may not span an exon-exon junction

Minimal and maximal number of bases that must anneal to exons at the 5' or 3' side of the junction ?

☐ Primer pair must be separated by at least one intron on the corresponding genomic DNA ?

Min

Max

1000

10000

?

- The following options are set as default. “Specificity Check” is chosen as default, which helps you verify the specificity of a primer. The organism is automatically filled in for you as “*D.mel*” (7227 is for fruit fly *D. melanogaster*, add more species if needed). The last option, “Allow splice variants”, is not used to design a specific primer for this chosen isoform.

Primer Pair Specificity Checking Parameters

Specificity check
☒ Enable search for primer pairs specific to the intended PCR template ?

Search mode
 Automatic ?

Database
 Refseq mRNA ?

Exclusion
☐ Exclude predicted Refseq transcripts (accession with XM, XR prefix) ☐ Exclude uncultured/environmental sample sequences ?

Organism
 7227 [Add organism](#)
 Enter an organism name (or organism group name such as enterobacteriaceae, rodents), taxonomy id or select from the suggestion list as you type. ?

Entrez query (optional)
 ?

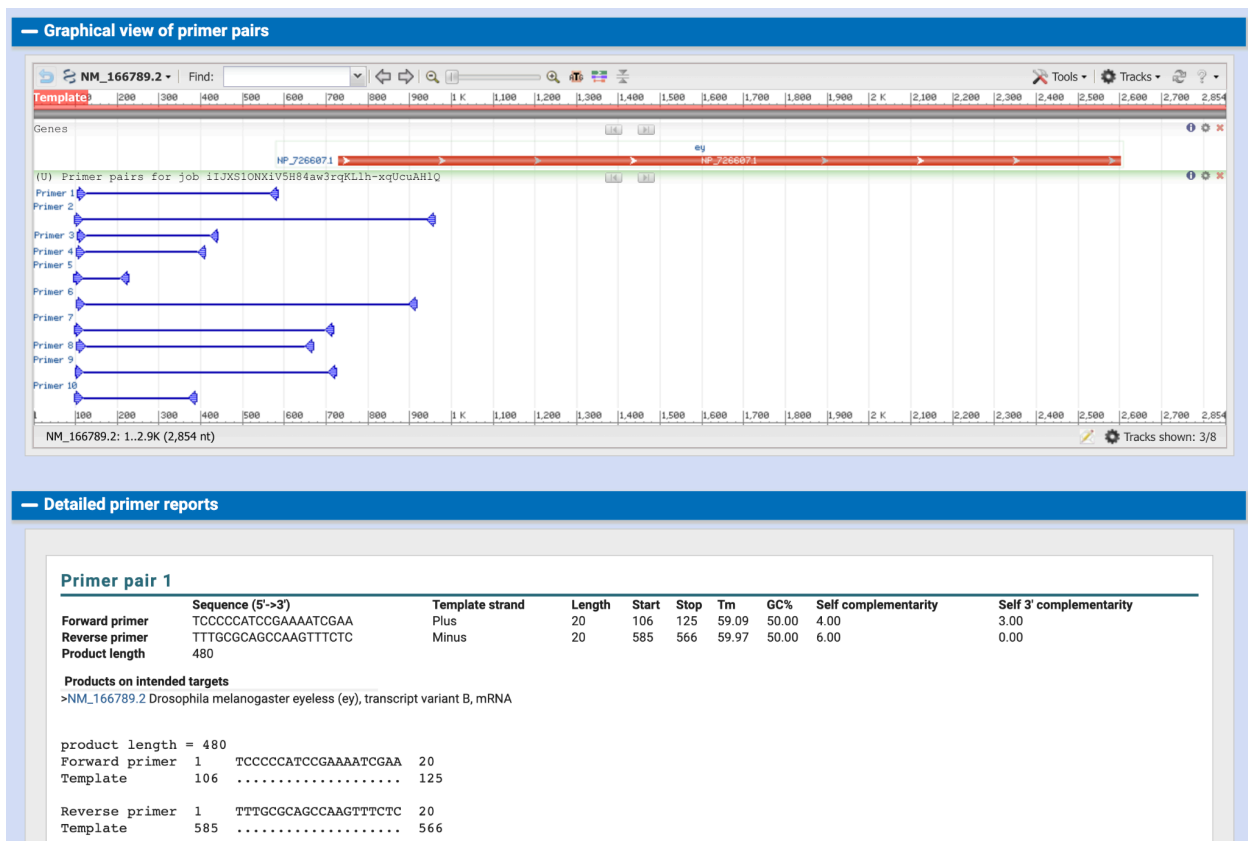
Primer specificity stringency
 Primer must have at least 2 total mismatches to unintended targets, including at least 2 mismatches within the last 5 bps at the 3' end. ?
 Ignore targets that have 6 or more mismatches to the primer. ?

Max target amplicon size
 4000 ?

Allow splice variants
☐ Allow primer to amplify mRNA splice variants (requires refseq mRNA sequence as PCR template input) ?

[Get Primers](#) ☐ Show results in a new window ☒ Use new graphic view ?

8. Click the “Get Primers” button and wait for a few seconds or minutes depending on the number of search requests. Subsequently, this page returns 10 pairs of primers.

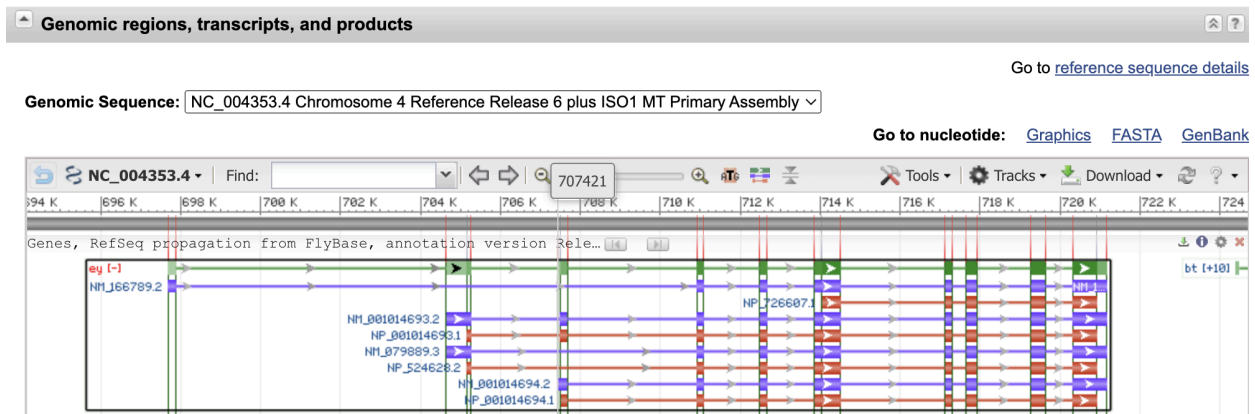


9. Save the screenshot of the graphical view of primer pairs and copy the primer pair 1 text to your word report. **Be sure to indicate for which gene this primer was designed.**

B) Design a primer for multiple isoforms

Change the range of the forward or reverse primer. The goal is to make sure that the primer falls in the common interval of all the transcripts since these isoforms have strong homology, almost only the 5' or 3' ends are slightly different.

1. Find the common interval of all transcripts in the genomic regions viewer at the top of the gene page. In this case, the 3' ends in all isoforms are the same, only the 5' end is different. Choose the longest transcript isoform (NM_166789.2) as the PCR template but use the range of the shortest transcript isoform (NM_001014694.2).



2. Compared with other isoforms, NM_001014694.2 is missing two exons. Move the mouse to the red lines indicating the range of exons to calculate the common range of all transcript variants. In this case, the first exon range is 697673–697874, the second exon range is 704621–705257.
3. Click on the hyperlink of the longest isoform (isoform B [NM_166789.2](#)) and choose “pick primers”. The forward primer range begins from 873.

Primers for target on one template

Primers common for a group of sequences

PCR Template

Retrieve recent results Publication Tips for finding specific primers

Enter accession, gi, or FASTA sequence (A refseq record is preferred) ?

Clear

Range ?

Clear

NM_166789.2

Forward primer From 873 To

Reverse primer

- Other settings are the same as the single isoform primer design except for choosing **"Allow splice variants"**. Click on "Submit" and wait until you see a warning.

Allow splice variants

☒ Allow primer to amplify mRNA splice variants (requires refseq mRNA sequence as PCR template input) ?

- The following warning is asking which variants you want to include together for this primer. Choose the desired variants and click "Submit" again.

Primer-BLAST

A tool for finding specific primers

Finding primers specific to your PCR template (using Primer3 and BLAST)

Input PCR template NM_166789.2 Drosophila melanogaster eyeless (ey), transcript variant B, mRNA

Range 873 - 2854

Your PCR template is highly similar to the following sequence(s) from the search database. To increase the chance of finding specific primers, please review the list below and select all sequences (within the given sequence ranges) that are intended or allowed targets.

Select: All None Selected:3

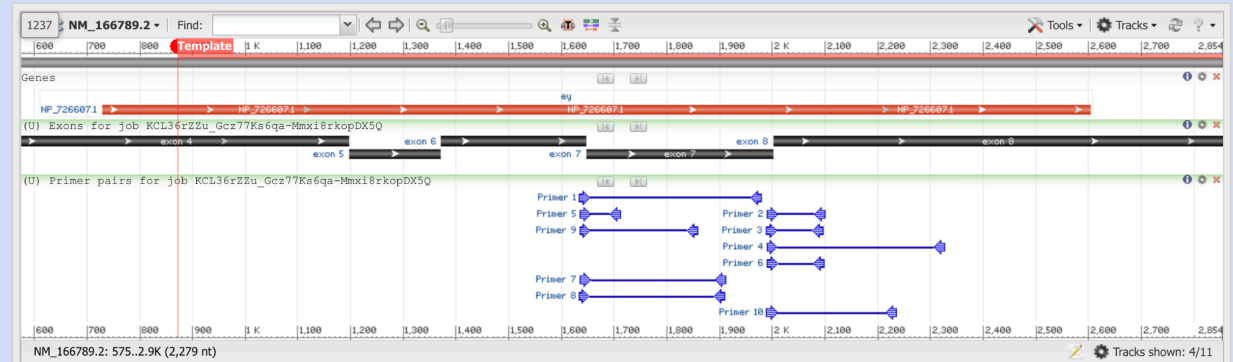
Accession	Title	Identity	Alignment length	Seq. start	Seq. stop
☒ NM_001014693.2	Drosophila melanogaster eyeless (ey), transcript variant D, mRNA	100%	1982	1481	3462
☒ NM_001014694.2	Drosophila melanogaster eyeless (ey), transcript variant C, mRNA	100%	1982	921	2902
☒ NM_079889.3	Drosophila melanogaster eyeless (ey), transcript variant A, mRNA	100%	1982	1301	3282

Submit

☒ Show results in a new window

- Wait until you see the result with 10 pairs of candidate primers.

Graphical view of primer pairs



Detailed primer reports

Primer pair 1

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	GGTCCCTCAGTCAGTACCA	Plus	20	1634	1653	60.25	60.00	5.00	1.00
Reverse primer	CGCCGATATGATGATGTTGG	Minus	20	1979	1960	60.04	55.00	6.00	0.00
Product length	346								
Exon junction	1647/1648 (forward primer) on template NM_166789.2								

- Save the screenshot of the graphical view of primer pairs and copy the primer pair 1 text to your word report. **Make sure to indicate for which isoform this primer was designed.**

Lab Exercise II - Validation of Previously Designed Primers

Material

Laptop

Previous Primers provided by instructor

NCBI Pick Primers tool (<https://www.ncbi.nlm.nih.gov/tools/primer-blast/index.cgi>)

Methods

An **eyeless** transcript primer will be used as an example.

- Pick your desired isoform, click on its NM hyperlink, and then click on “pick primers”.

mRNA and Protein(s)

- [NM_166789.2](#) → [NP_726607.1](#) **eyeless, isoform B [Drosophila melanogaster]**
[See identical proteins and their annotated locations for NP_726607.1](#)

- In the pick primers page, input the candidate forward and reverse primers for validation. Also, in “Exon junction span” choose “Primer must span an exon-exon junction”.

Primers for target on one template

Primers common for a group of sequences

Retrieve recent results

Publication

Tips for finding specific primers

PCR Template

Enter accession, gi, or FASTA sequence (A refseq record is preferred) ? Clear

NM_166789.2

Or, upload FASTA file 选择文件 未选择任何文件

Range ? Clear

Forward primer From To

Reverse primer From To

Primer Parameters

Use my own forward primer (5'→3' on plus strand) ? Clear

Use my own reverse primer (5'→3' on minus strand) ? Clear

PCR product size

Min 70 Max 1000

of primers to return 10

Primer melting temperatures (T_m)

Min 57.0 Opt 60.0 Max 63.0 Max T_m difference 3 ?

Exon/intron selection

A refseq mRNA sequence as PCR template input is required for options in the section ?

Exon junction span Primer must span an exon-exon junction ?

- If you see this warning, you can broaden the search by unchecking the “Exon junction span” option and redoing the search.

Primer-BLAST » JOB ID:J5LV1SWWT5-BMKbXGHM756_AGTaecckg

Primer-BLAST Results ?

Warnings:
 ⚠ No primers satisfying exon/intron selection parameters were found. Try to loosen the selection criteria and/or increase value for “Max primer pairs to screen” parameter.

Input PCR template NM_166789.2 Drosophila melanogaster eyeless (ey), transcript variant B, mRNA
 Range 1 - 2854
 Other reports [Search Summary](#)

No primers were found...see explanation below: Primer3 info:
 Left primer: considered 1, high end compl 1, ok 1.
 Right primer: considered 1, ok 1.
 Primer pairs: considered 2, no overlap of required point 1, ok 1

- You can read the result to see how this primer fits your template DNA. This result also includes the products on potentially unintended templates. The fewer the unintended templates the more specific the primer is to your desired gene.

Graphical view of primer pairs

Detailed primer reports

Primer pair 1

	Sequence (5'→3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	TGGTAGGTC AATCACCCAACC	Plus	21	886	906	59.65	52.38	5.00	
Reverse primer	GCTGCTGTAGTGCCTGATGG	Minus	20	969	950	61.09	60.00	3.00	1.00
Product length	84								

Products on intended targets
 >NM_166789.2 Drosophila melanogaster eyeless (ey), transcript variant B, mRNA

product length = 84
 Forward primer 1 TGGTAGGTC AATCACCCAACC 21
 Template 886 906

Reverse primer 1 GCTGCTGTAGTGCCTGATGG 20
 Template 969 950

Products on potentially unintended templates
 >NM_001014693.2 Drosophila melanogaster eyeless (ey), transcript variant D, mRNA

product length = 84
 Forward primer 1 TGGTAGGTC AATCACCCAACC 21
 Template 1494 1514

Reverse primer 1 GCTGCTGTAGTGCCTGATGG 20
 Template 1577 1558

>NM_001014694.2 Drosophila melanogaster eyeless (ey), transcript variant C, mRNA

- Save the screenshot of the graphical view of primer pairs and copy all the detailed primer reports (all the products on intended targets or potentially unintended templates) to the report. Remember to evaluate this primer and provide a suggestion. **Be sure to indicate which primer was evaluated and its origin.**

Lab Exercise III - FlyPrimerBank Primer Validation.

Material

Laptop

FlyPrimerBank (<https://www.flyrnai.org/flyprimerbank>)

NCBI Primer-BLAST (<https://www.ncbi.nlm.nih.gov/tools/primer-blast/index.cgi>)

Methods

1. Download a pair of primers from FlyPrimerBank. Type in the "Enter **gene identifiers/Primer IDs**" box, for example **ey** for eyeless.

Search FlyPrimerBank for PCR Primers in *D. melanogaster*

[Documentation / Demo Video](#)

[Details](#)

Enter gene identifiers/primer IDs ([help](#)):

ey



Or upload from a file of gene IDs:

2. Click on the search button. Suggested primers will be displayed in seconds. Choose the first pair. Save the primer text in your report.

Search Term	Primer Pair	Forward Primer	Fwd Pri Temp	Reverse Primer	Rev Pri Temp	PCR Product	FBgn	Symbol	Isoforms Targeted	Region	Exon Spanning?
ey	PP21231	TGGTAGGTCAATCACCCAACC	61.4°C	GCTGCTGTAGTGCTGATGG	62.9°C	84	FBgn0005558	ey	4/4	CDS	No
ey	PP33244	TCAGGCACTACAGCAGCATC	61.9°C	TTTCGCTAAGAGAGGTGGGGT	62.6°C	83	FBgn0005558	ey	4/4	CDS	No
ey	PP8355	CGGCTGTTAAATACTCAGCAGC	61.6°C	TGGGTGATTGACCTACCAAGG	61.1°C	82	FBgn0005558	ey	4/4	CDS	No

3. Validate the FlyPrimerBank primer by using NCBI Primer-BLAST ("pick primers"). Leave the PCR template blank, copy-paste the primers from the FlyPrimerBank result and

choose the organism (*Drosophila melanogaster* - taxid:7227). Submit and wait for the result. This is also a simulation of the PCR experiment using a certain primer.

Primers for target on one template

Primers common for a group of sequences

PCR Template

Enter accession, gi, or FASTA sequence (A refseq record is preferred) ?

Clear

Or, upload FASTA file

选择文件

未选择任何文件

Range ?

Clear

From

To

Forward primer

Reverse primer

Primer Parameters

Use my own forward primer (5'->3' on plus strand)

Clear

Use my own reverse primer (5'->3' on minus strand)

Clear

PCR product size

Min

Max

of primers to return

Primer melting temperatures (T_m)

Min

Opt

Max

Max T_m difference

Exon/intron selection

Exon junction span

?

Exon junction match

Min 5' match

Min 3' match

Max 3' match

Intron inclusion

☐

Primer pair must be separated by at least one intron on the corresponding genomic DNA ?

Intron length range

Min

Max

Note: Parameter values that differ from the default are highlighted in yellow

Primer Pair Specificity Checking Parameters

Specificity check

☒

Enable search for primer pairs specific to the intended PCR template ?

Search mode

?

Database

?

Exclusion

☐

Exclude predicted Refseq transcripts (accession with XM, XR prefix)

☐

Exclude uncultured/environmental sample sequences ?

Organism

Add organism

Enter an organism name (or organism group name such as enterobacteriaceae, rodents), taxonomy id or select from the suggestion list as you type. ?

4. Save the screenshot of the graphical view of primer pairs and copy all the detailed primer reports (all the products on intended targets or potentially unintended templates) in the report. Remember to evaluate this primer and provide a suggestion. **Be sure to indicate which primer was evaluated and its origin.**

Introduction DNA isolation and amplification

Genomic DNA: A cell's DNA, packaged as a double-stranded DNA molecule, is called its genome. Genomic DNA **constitutes the total genetic information of an organism**. Genomic DNA molecules are generally large, and, in most organisms, these are organized into DNA–protein complexes called chromosomes. In research, genomic DNA is a useful tool in applications such as polymerase chain reaction (PCR), library construction, Southern blotting, hybridizations, single nucleotide polymorphism (SNP) analysis, and molecular diagnostic assays.

Genome Sequence of *Drosophila*: The genome of *D. melanogaster* is approximately 180 megabase (Mb) in size, with roughly two-thirds euchromatic and one-third heterochromatic. Euchromatin contains approximately 98% of the protein-coding genes in the genome; heterochromatin is largely composed of simple sequence repeats.

DNA isolation, also known as DNA extraction, is a fundamental technique in molecular biology that involves separating DNA from cells or tissues for various downstream applications. The isolation process aims to obtain pure DNA free from contaminants such as proteins, lipids, and other cellular components.

The extracted DNA can be used in a wide range of applications including PCR (Polymerase Chain Reaction), DNA sequencing, genetic testing, forensics, and cloning. The success of these downstream applications relies heavily on the quality and purity of the isolated DNA.

Several methods exist for DNA isolation, each tailored to different sample types and experimental requirements. Commonly used techniques include phenol-chloroform extraction, silica column-based purification, and magnetic bead-based extraction.

DNA Quantification: A reliable measurement of DNA concentration is important for many applications in molecular biology. Several methods are commonly used to measure both genomic and plasmid DNA concentrations.

1. **Spectrophotometry** can be used to measure microgram quantities of pure DNA samples (i.e., DNA that is not contaminated by proteins, phenol, agarose, or RNA). DNA concentrations can be determined by measuring the absorbance at 260 nm (A_{260}) in a spectrophotometer using a quartz cuvette. For optimal accuracy, readings should be between 0.1 and 1.0. An absorbance of 1.0 at 260 nm corresponds to 50 $\mu\text{g/mL}$ of genomic DNA ($A_{260} = 1$ for 50 $\mu\text{g/mL}$; based on a standard 1 cm path length. This relation is valid only for measurements made at neutral pH, therefore, samples should be diluted in a low-salt buffer with neutral pH [e.g., Tris·Cl, pH 7.0]). Spectrophotometric measurements do not differentiate between DNA and RNA. Therefore, RNA contamination can lead to an overestimation of the DNA concentration. The **NanoDrop Spectrophotometer** is a spectrophotometer that enables highly accurate analyses of 1 μL samples of DNA, RNA, protein, pigments, the whole ultraviolet (UV)-visible (220-700 nm) spectrum, and cell density with remarkable reproducibility.
2. **Fluorometry** is more sensitive, allowing the measurement of nanograms of DNA. The Hoechst 33258 dye has little affinity for RNA, allowing accurate quantification of DNA samples that are contaminated with RNA. This dye shows increased emission at 458 nm when bound to DNA. DNA standards and samples are mixed with Hoechst 33258 and measured in glass or acrylic cuvettes using a scanning fluorescence spectrophotometer or a dedicated filter fluorometer set at an excitation wavelength of 365 nm and an

emission wavelength of 460 nm. The sample measurements are then compared to the standards to determine DNA concentration.

3. **Agarose gel analysis** enables quick and easy quantification of DNA, especially for small DNA fragments (such as PCR products). As little as 20 ng of DNA can be detected by agarose gel electrophoresis with ethidium bromide staining. The DNA sample is run on an agarose gel alongside known amounts of DNA of a similar size. The amount of sample DNA loaded can be estimated by comparison of the band intensity with the standards using a scanner or imaging system. Standards of roughly the same size as the fragment of interest should be used to ensure a reliable estimation of the DNA quantity, since large fragments intercalate more dye than small fragments producing greater band intensity.

Lab Exercise IV - Observe F2 Generation Phenotypes from the Lab 5 (if applicable)

Materials

Genetic crosses performed in lab 5

Methods

1. Observe the F2 progeny of the crosses produced in lab 5 and their phenotypes.
2. Count the total number of F2 generation.
3. Separate flies based on their phenotypes.
4. Count the number of flies in each phenotype.

In Lab notebook:

1. Calculate F2 phenotype and genotype ratios and describe how these observations may fit your original plan for genetic crosses.
2. Reflect on how this observation fits the law of inheritance, such as Mendel's law of segregation and independent assortment.

Lab Exercise V PCR amplification

Materials

Ice

Microcentrifuge

Vortex

Thermal cycler with special program named 'FlyPCR'

4 DNA samples from Lab exercise V (on Ice)

Eb Primers(Forward: CCAGGGAAGTCCCACGTTATC; Reverse: GGAGCCGATTGAGGATGCT)

Vg Primers(Forward: GGGGCAATACCCAAGAATCTG; Reverse: TTGGTGAACACAACGCAGGA)

2 PCR machines

2 PCR Master Mix (2X) (On ice) stores at -20°C

2 Nuclease-free water (come with the PCR master Mix) (On ice)
 2 2-20 ul micropipettes
 2 20-200 ul micropipettes
 2 box of P20-200
 2 PCR tubes rack
 8 0.2ml PCR tubes
 2 fine-tip markers
 2 beakers for pipettes disposal

'FlyPCR' Program conditions:

Step	Temperature, °C	Time	Number of cycles
Initial denaturation	95	2 min	1
Denaturation	95	30 s	33
Annealing	55	30 s	
Extension	72	30 S	
Final Extension	72	5 min	1
Infinite Hold	4	Infinite	1

Methods

1. Gently vortex and briefly centrifuge PCR Master Mix (2X) after thawing (Complete by instructor).

2. Each group obtain 2 PCR tubes and label them as

Group A: ① WT-E ② WT-V

Group B: ① Eb-E ② Eb-V

Group C: ① Vg-E ② Vg-V

Group D: ① 2-E ② 2-V

3. Place PCR tubes on ice and add following components:

PCR Master Mix (2X)	25 µL
Forward primer	0.1-1.0 µM
Reverse primer	0.1-1.0 µM
Template DNA	10 pg - 1 µg
Water, nuclease-free	to 50 µL
Total volume	50 µL

4. The concentration of each primer is 10uM, use the formula **C1V1 = C2V2** to find out the volume of primers for this PCR.
5. DNA samples are measured by nanodrop with a unit of ng/ul. Make sure the mass of the total DNA template falls in the range of 10pg - 1ug. (1 ug = 1000 ng = 1000,000 pg). Calculate the volume of DNA template for this PCR.
6. Add nuclease-free water to match the total volume of 50uL.
7. Gently vortex the samples and spin down with a microcentrifuge.
8. Put PCR tubes in Thermal cycler and set up the program named 'FlyPCR'.
9. After PCR is done, Lab technician will pick them up and stores at -20°C

Assignment 7

1. Download assignment 7 template and complete below questions.
2. Lab Exercise I: Design primers for **Vestigial and Ebony** genes. If the gene has multiple isoforms, design a primer to fit all of them. Follow the Lab Exercise I method and save each result to the report.
3. Lab Exercise II: Do primer validation for each pair of primers given below:
Eyeless: CGGCTGTAAATACTCAGCAC (Forward Primer);
TGGGTGATTGACCTACCAAGG (Reverse Primer)
BarH1: GACTCGATGAGCATACTCACCC (Forward Primer);
GACCGATTGGGCTTACCGAT (Reverse Primer)
4. Lab Exercise III: Do primer validation for one pair of primers chosen by yourself for **Vestigial and Ebony** genes.

Save the screenshot of the graphical view of primer pairs and copy all the detailed primer reports (all the products on intended targets or potentially unintended templates) in the report. Remember to evaluate this primer and provide a suggestion. Be sure to indicate which primer was evaluated and its origin.

Lab notebook

1. Calculate F2 phenotype and genotype ratios and describe how these observations may fit your original plan for genetic crosses.
2. Reflect on how this observation fits the law of inheritance, such as Mendel's law of segregation and independent assortment.
3. Record the DNA isolation procedure, DNA quantification result, PCR procedure and T_m calculation and DNA sample volume.

Lab 8 Gel electrophoresis, Microarray, RNAseq analysis

Hanqi Ye

Objective

- To practice casting a gel and interpret the result
- To understand the microarray and RNA-seq and their difference
- To practice analysing mRNA expression dataset with Galaxy
- To be able to interpret the visualization

Introduction

Gel electrophoresis is a technique employed to separate and analyze biomacromolecules (such as DNA, RNA, and proteins) and their fragments, utilizing differences in size and charge. This method finds applications in various fields: in clinical chemistry, it's used to segregate proteins according to charge or size (for instance, IEF agarose, which is essentially size-independent); and in biochemistry and molecular biology, it's utilized to separate mixed DNA and RNA fragments by length, estimate the size of DNA and RNA fragments, or segregate proteins based on charge.

Nucleic acid molecules are separated by applying an electric field that propels the negatively charged molecules through a matrix composed of agarose or other substances. Shorter molecules move faster and migrate further compared to longer ones due to their enhanced ability to navigate through the gel's pores. This phenomenon is known as sieving. Conversely, proteins are segregated based on charge within agarose, as the gel's pores are too large to effectively sieve proteins. Gel electrophoresis can also facilitate the separation of nanoparticles.

DNA gel electrophoresis is typically conducted for analytical purposes, often following DNA amplification by polymerase chain reaction (PCR). It may also be used as a preparative technique preceding other methods such as mass spectrometry, restriction fragment length polymorphism (RFLP), PCR, cloning, DNA sequencing, or Southern blotting to further characterize the DNA.

Electrophoresis experiments often specify a particular percentage of agarose to be used in the gel. Typically, scientists utilize agarose gels ranging from 0.5% to 2% in concentration. Agarose gels are prepared by dissolving agarose powder in a liquid buffer. The gel percentage is

expressed as a weight to volume (w/v) ratio. For instance, dissolving 1 gram of agarose powder in 100 ml of buffer yields a 1% agarose gel.

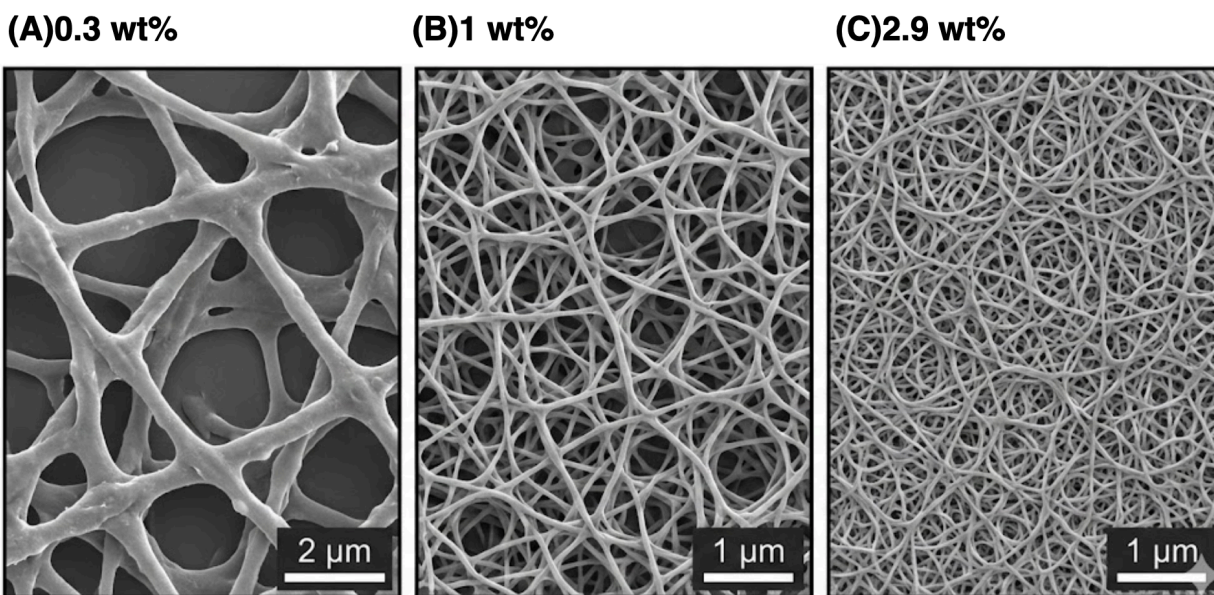


Figure 1 . Simulated scanning electron microscopy (SEM) images of polymer network structures formed at different concentrations. (a) **0.3 wt%**: A sparse and loosely connected fibrous network with large pores and thick, irregular strands. Scale bar = 2 μm . (b) **1 wt%**: A more interconnected fibrillar network with increased fiber density and reduced pore size relative to the low-concentration sample. Scale bar = 1 μm . (c) **2.9 wt%**: A highly dense and uniform nanofibrous network with tightly packed structures and minimal open space. Scale bar = 1 μm .

The progressive increase in concentration results in a corresponding increase in network density, fiber interconnectivity, and overall structural compactness.

In a gel, agarose forms a mesh that functions as a molecular sieve. Gels with higher percentages of agarose create a tighter mesh with smaller pores. In figure 1, a 1% gel (middle) exhibits significantly larger pores compared to a 2.9% gel (right).

Generally, higher percentage agarose gels are more effective for separating smaller DNA fragments, whereas lower percentage gels are preferable for separating larger DNA fragments. The table below summarizes common agarose gel percentages and the corresponding range of DNA fragment sizes that can be effectively separated:

Agarose Gel Concentration (%)	Effective Separation Range (bp)
0.3%	5,000 – 60,000 bp
0.5%	1,000 – 30,000 bp
0.7%	800 – 12,000 bp

1.0%	500 – 10,000 bp
1.2%	400 – 7,000 bp
1.5%	200 – 3,000 bp
2.0%	100 – 2,000 bp
3.0%	50 – 1,000 bp
4.0%	20 – 400 bp

Table 1. Agarose gel concentration: Range of Effective Separation (bp).

In addition to the agarose gel percentage, several other factors can significantly influence band resolution in gel electrophoresis experiments. These factors include:

1. **Voltage and Run Time:** The voltage applied during electrophoresis and the duration of the run impact band resolution. Higher voltages can lead to faster migration of DNA fragments but may result in increased heating and band smearing. Longer run times can improve resolution but may also cause diffusion and broadening of bands.
2. **Buffer Composition:** The type and concentration of the electrophoresis buffer used can affect band resolution. Buffers such as TAE (Tris-acetate-EDTA) or TBE (Tris-borate-EDTA) provide stable pH and ionic strength, which offers better resolution for smaller DNA fragments.
3. **Gel Thickness:** The thickness of the gel can influence resolution. Thinner gels allow for better heat dissipation and more efficient separation of DNA fragments compared to thicker gels, which may experience uneven heating and slower migration rates.
4. **DNA Loading Amount:** Overloading the gel with DNA samples can lead to overcrowded bands and poor resolution. Optimizing the amount and volume of DNA loaded onto the gel ensures clear and distinct bands.
5. **Staining and Visualization:** The choice of DNA stain and the method of gel visualization (UV transillumination, fluorescence imaging, etc.) can affect the clarity and visibility of DNA bands. Proper staining and imaging techniques are essential for accurate band analysis.
6. **Quality of DNA Samples:** The quality and purity of DNA samples used in electrophoresis influence band resolution. Contaminants or degraded DNA can result in smearing or indistinct bands.

By carefully considering and optimizing these factors, researchers can enhance band resolution and overall performance in gel electrophoresis experiments, leading to more reliable and interpretable results.

Lab Exercise I Gel electrophoresis

Material

1 bucket of ice(thaw below 3 items on ice)

8 PCR samples from previous lab 7

2 pBE322/BstNI Marker

2 5X loading dye

1 20xTBE buffer

2 250ml flask

2 SeeGreen agarose tabs

2 2-20ul micropipette with 1 box of P20-200

2 100 ml graduated cylinder

2 bottle of 500ml Distilled water

2 gel apparatus with 9 gel comb with 9 wells

2 BlueGel apparatus

1 set of 4 combs practice gel in a apparatus filled with water

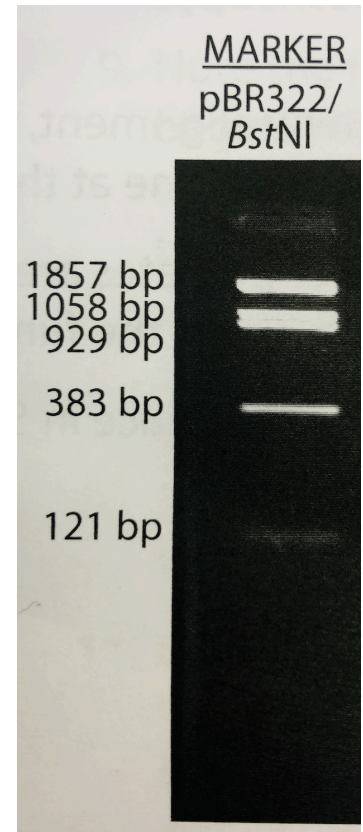


Figure 2 pBE322/BstNI Marker

Methods

1. Add 1 SeeGreen all-in-one agarose tab, 20ml Distilled water to 250ml flask, swirl to mix and microwave for 30s.

2. Place the flask on a bench to cool down for 1min, then pour the gel into the gel casting tray.

3. Insert the 9 well comb into the gel casting tray, each well holds up to 20 ul samples.

4. Let gel stand for ~10 minutes until completely set. For a faster set time, place the casting platform with the gel in a refrigerator. Do not disturb the gel during this time.

5. While waiting for the gel, prepare the DNA samples for loading.

6. To each PCR sample tube, add 12.5ul 5X loading dye and mix by pipetting in and out.

7. After the gel has solidified, remove comb/s gently by pulling straight upward.

8. Remove the gel tray from the casting platform. If a small amount of gel has formed underneath the gel tray, wipe it off and discard it.

9. Place the gel tray containing a gel in the buffer chamber and place the buffer chamber inside the blueGel™ base. The wells should be closest to the (-) end.

10. Add 40 ml of 1X TBE buffer in the buffer chamber. The buffer should just cover the agarose gel.
11. Remove air bubbles (if any) trapped between the gel and the gel tray, or between the gel tray and the buffer chamber.
12. Load 15ul pBR322/BstNI marker to the 1st line of the gel.
13. Load 15ul of each PCR sample to the rest of the wells.
14. Run the gel for approximately 40 mins.
15. Observe the result of the gel by turning on the UV light.
16. Take a picture of the result and interpret the result on a lab notebook.

Introduction to Microarray and RNAseq

Microarray technology is a common lab technique that involves attaching a large array of known nucleic acid fragments, ranging from thousands to millions, onto a solid surface known as a “chip”. This chip is then exposed to DNA or RNA extracted from a sample under study, such as cells or tissues. The interaction between the sample and the fragments immobilized on the chip, through complementary base pairing, generates light via fluorescence. This light can be detected using a specific machine. Microarray technology has a wide range of applications in both research and clinical studies, including the measurement of gene expression and the identification of specific DNA sequences, such as single-nucleotide polymorphisms (SNPs).

RNA-Seq (named as an abbreviation of RNA sequencing) is a technique that uses next-generation sequencing to reveal the presence and quantity of RNA molecules in a biological sample, providing a snapshot of gene expression in the sample, also known as transcriptome.

RNA-Seq is transforming our understanding of the transcriptome. As a highly precise and sensitive method for quantifying expression across the transcriptome, it offers researchers insights into previously unseen changes that occur in various states of disease, in reaction to therapeutic treatments, under diverse environmental conditions, and in a multitude of other experimental designs. RNA-Seq empowers scientists to identify both recognized and new features in a single test. This includes the detection of transcript isoforms, gene fusions, single nucleotide variants, and other elements, all without the constraint of pre-existing knowledge.

Microarrays allow genomic analysis without sequencing, which tremendously reduces the cost of doing large studies across a wide area of biology and biomedicine (Figure 3). They enable the examination of gene expression, or the quantity of RNA product from any identified gene in a bio sample and facilitate the easy detection of Single Nucleotide Polymorphisms SNPs, which are crucial for genome-wide association studies (GWASs).

RNA-Seq has several advantages over microarrays. It provides a very wide dynamic range, improving the sensitivity of detection for rare transcripts. It also allows for the examination of more than just differential gene expression. With RNA-Seq, you can look at coding and non-coding RNA, splicing, allele-specific expression, and even full-length cDNA sequences. Moreover, RNA-Seq data can be reanalyzed as more information about the transcriptome becomes available. This makes it a powerful tool for studying gene expression and its changes over time, or differences in gene expression in different groups or treatments (Figure 3).

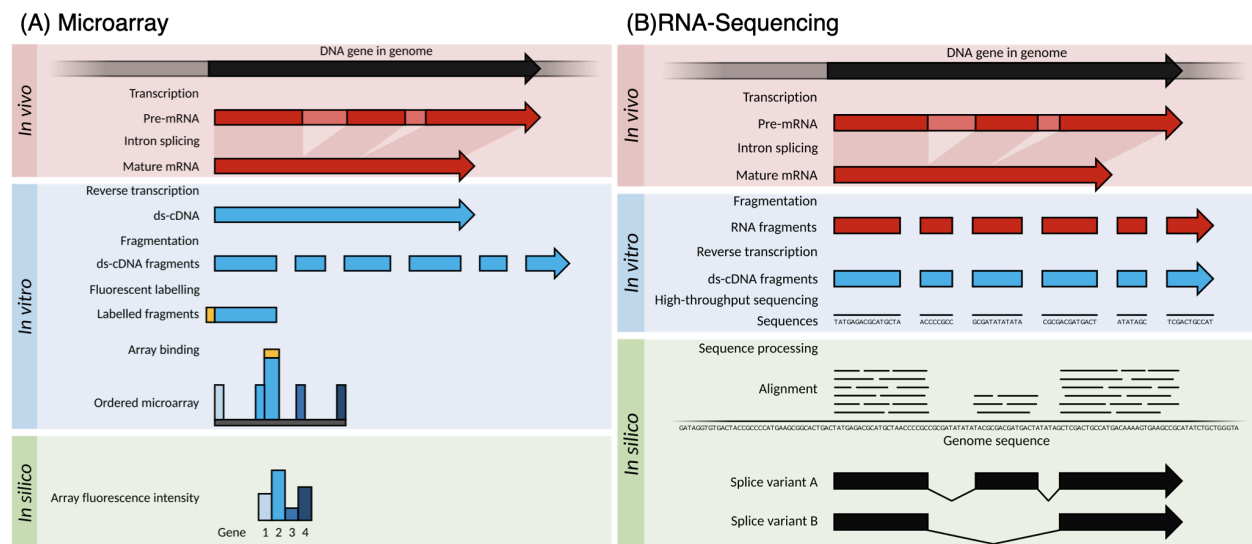


Figure 3 Microarray(left) vs RNA-seq(right).Microarray adapted from Summary of RNA Microarray.svg by Thomas Shafee – CC BY 4.0 (via Wikimedia Commons); RNA-seq diagram adapted from Summary of RNA-Seq.svg by Thomas Shafee – CC BY 4.0 (via Wikimedia Commons).

Introduction to RNAseq pipeline

Creating a pipeline for RNA-Seq analysis involves several steps, each with its own purpose. Here's an overview of a typical RNA-Seq pipeline:

1. Quality Control (QC)

- **Purpose:** Assess the quality of the raw sequencing reads to identify potential issues such as low-quality bases, adapter contamination, or duplicated sequences.
- **Tools:** FastQC, MultiQC
- **Why:** Ensuring data quality at the beginning prevents erroneous downstream analyses and saves computational resources.

2. Trimming & Filtering

- **Purpose:** Remove low-quality bases and adapter sequences from the reads.
- **Tools:** Trim Galore!, Trimmomatic, Cutadapt

- **Why:** Low-quality sequences and adapter contamination can introduce bias and inaccuracies in downstream analysis, such as alignment.

3. Read Alignment

- **Purpose:** Map the quality-controlled reads to a reference genome or transcriptome.
- **Tools:** *STAR*, *HISAT2*, *Bowtie2*
- **Why:** Mapping reads to a reference genome is necessary to identify which genes are being expressed and how abundantly.

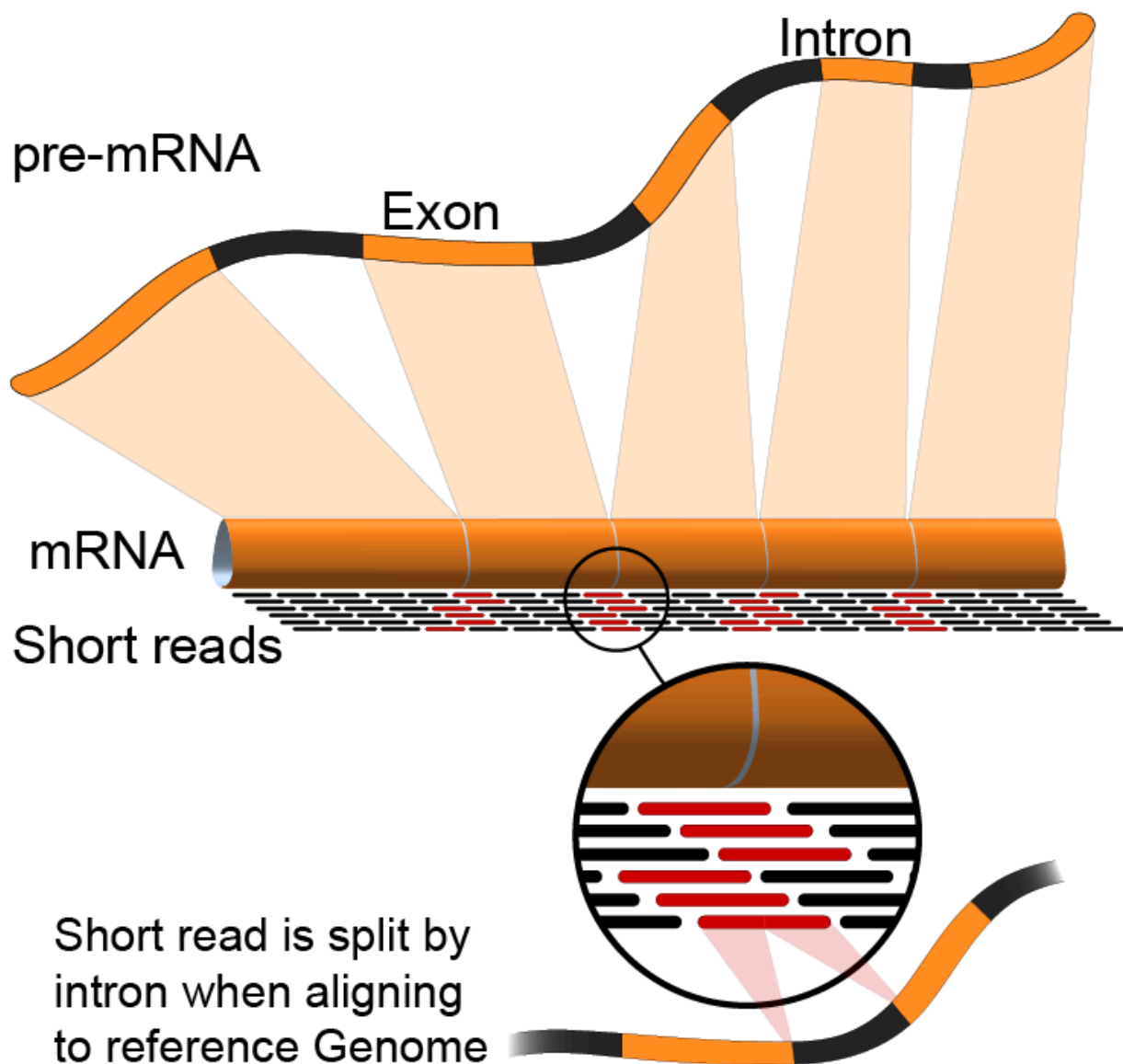


Figure4. Overview of RNA-seq read alignment to a reference genome.

Short sequencing reads generated from RNA-seq are aligned to the reference genome to determine their genomic origin and quantify transcript abundance. Exonic reads map directly to coding regions, while splice-junction reads span across exon–exon boundaries, indicating alternative splicing or transcript isoforms. Regions of higher read density correspond to higher expression levels. Accurate alignment enables downstream analyses such as differential gene expression, transcript structure reconstruction, and isoform discovery. *(Image adapted from RNA-Seq-alignment.png by Rgocs – CC BY (via Wikimedia Commons).)*

4. Post-alignment QC

- **Purpose:** Assess the quality of the aligned reads.
- **Tools:** [Samtools](#), [Qualimap](#)
- **Why:** To ensure that the majority of the reads align to the reference genome properly, which is crucial for accurate quantification.

5. Transcript Assembly & Quantification

- **Purpose:** Quantify gene or transcript expression levels from the aligned reads.
- **Tools:** [StringTie](#), [Salmon](#), [Kallisto](#), [FeatureCounts](#)
- **Why:** The aim is to calculate the abundance of transcripts, which will be used for downstream analysis like differential expression.

6. Differential Expression Analysis

- **Purpose:** Identify genes that are differentially expressed between conditions (e.g., control vs. treated).
- **Tools:** [DESeq2](#), [edgeR](#), [limma](#)
- **Why:** Differential expression analysis helps uncover how experimental conditions affect gene expression.

7. Functional Enrichment Analysis

- **Purpose:** Determine the biological significance of the differentially expressed genes (DEGs).
- **Tools:** [DAVID](#), [GSEA](#), [ClusterProfiler](#)
- **Why:** Enrichment analysis reveals pathways or processes that may be regulated under the experimental conditions.

8. Visualization

- **Purpose:** Create visual representations of the data, such as heatmaps, volcano plots, and PCA plots.
- **Tools:** [ggplot2](#), [heatmap](#), [ComplexHeatmap](#)

- **Why:** Visualization helps interpret and communicate the results clearly.

9. Validation

- **Purpose:** Confirm the differential expression results using a complementary method.
- **Tools:** qPCR, Western blot
- **Why:** Validation ensures that the RNA-Seq findings are accurate and biologically relevant.

This pipeline provides a comprehensive workflow for RNA-Seq analysis, addressing both the technical aspects (like quality control and alignment) and the biological interpretation (differential expression and functional enrichment). Each step serves to refine the data and increase the accuracy of biological insights derived from the experiment.

Lab Exercise II RNA-seq data analysis - Part one

Materials:

Galaxy: <https://usegalaxy.org/>

Dataset : <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE241397>

Paper: <https://onlinelibrary.wiley.com/doi/10.1002/ece3.10976>

RNA-seq data analysis part 1 worksheet word file (download from the blackboard or this [link](#))

Methods:

1. Click on the link [Dataset](#) and read the research description and click on the link [paper](#), to find the diet section in the Methodology, **answer the Q1 in data analysis assignment**.
2. Click on the [SRA Run Selector](#), it will jump to the dataset download page, read the information about dietary condition and **answer the Q2 in data analysis assignment**
3. In the SRA Run Selector page, click on the Accession list in the select section, it will download the list of all accession numbers.

Select	Runs	Bytes	Bases	Download	Cloud Data Delivery	Computing
Total	12	11.34 Gb	39.08 G	Metadata or Accession List		
Selected	0	0	0	Metadata or Accession List or JWT Cart	Deliver Data	Galaxy

4. Click on Galaxy: <https://usegalaxy.org/> and sign up an account and log in.
5. Click on the History + icon to create a new history and rename it as RNAseq.
6. Click on the **Upload** and **Choose local file**, upload the accession list just downloaded.
7. Once complete uploading, In the **Tool** section, click on the **Get data**, choose **Download and Extract Reads in FASTQ**

8. In the tool parameters section, under select input type, choose **List of SRA accession, one per line** and make sure the uploaded accession list is chosen. Select output format as gzip compressed fastq. Enable the Email notification and click on **Run tool**. It takes about **3 hours** to complete the download process.

Download and Extract Reads in FASTQ format from NCBI SRA (Galaxy Version 3.1.1+galaxy1)

Tool Parameters

select input type

List of SRA accession, one per line

sra accession list *

1: SRR_Acc_List.txt

accepted formats

Select output format *

☒ gzip compressed fastq

☐ Uncompressed fastq

☐ bzip2 compressed fastq

Compression will greatly reduce the amount of space occupied by downloaded data. Downstream applications such as a short-read mappers will accept compressed data as input. Consider this example: an uncompressed 400 Mb fastq datasets compresses to 100 Mb or 80 Mb by gzip or bzip2, respectively. (`--gzip --bzip2`)

Advanced Options

Additional Options

Email notification

☒ Yes

Send an email notification when the job completes.

9. In the right side under the RNAseq history, you will see the **Single-end data** and **pair-end data**. Delete the **Single-end data** collection as it is empty and then search for tool **Flatten collection**, choose **pair-end data** as input collection. Click on **Run Tool**, this takes 20 seconds.

Flatten collection (Galaxy Version 1.0.0)

Tool Parameters

Input Collection *

accepted formats 2: Paired-end data (fastq-dump)

Join collection identifiers using *

underscore (_)

Separator for merging dataset identifiers

10. In the **Tool** section, search for **FASTQC**, click on it and under the **Tool Parameters**, “Raw read data from your current history” click on **Dataset collection** and choose: Output of Flatten collection. Keep the rest setting as default and **Run Tool**. This takes about **10 mins**.

Tool Parameters

Raw read data from your current history *

📄 📄 📁 ⋮

106: data 27, data 26, and others (flattened)

⌵

accepted formats ▼

11. Inspect the webpage output of FastQC tool for the SRR25724539 sample (forward and reverse), answer the **Q3 in the assignment**.
12. As it is tedious to inspect all these reports individually we will combine them with tool **MultiQC**, parameters as follows:

Tool Parameters

Results

1: Results
⬆ ⬇ 🗑

Which tool was used generate logs?

FastQC

Software name

FastQC output

1: FastQC output
⬆ ⬇ 🗑

Type of FastQC output? *

Raw data

Raw data

FastQC output *

📄 📄 📁 ⋮

108: FastQC on collection 106: RawData

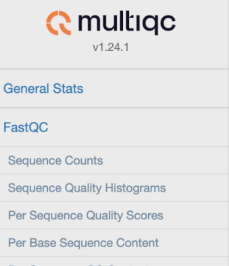
✕

⌵

accepted formats ▼

switch to column select ▼

13. Under the History, to inspect the MultiQC webpage result, click on the **eye** icon. Inspect the result and answer **the Q4 in the assignment**.



General Stats

FastQC

Sequence Counts

Sequence Quality Histograms

Per Sequence Quality Scores

Per Base Sequence Content

Per Sequence GC Content

General Statistics

📄 Copy table
⚙️ Configure columns
📊 Scatter plot
📊 Violin plot
📄 Export as CSV

Showing 24/24 rows and 3/6 columns.

Sample Name	% Dups	% GC	M Seqs
SRR25724544_reverse	54.8 %	50 %	10.7 M
SRR25724544_forward	54.0 %	50 %	10.7 M
SRR25724545_forward	54.3 %	50 %	11.5 M
SRR25724545_reverse	53.7 %	51 %	11.5 M
SRR25724546_forward	51.0 %	50 %	9.2 M

History

search datasets

RNAseq

17.4 GB

158: MultiQC on data 156, data 154, and others: Stats

157: MultiQC on data 156, data 154, and others: Webpage

14. Use **Trim Galore!** tool to remove the adaptors, the parameters are as below picture **EXCEPT Adapter sequence to be trimmed** choose **the one you answered in Q4**. Run the tool and it takes about 20 mins.

Trim Galore! Quality and adapter trimmer of reads (Galaxy Version 0.6.7+galaxy0) ☆ 🔗 ▼ 📄 ▶ Run Tool

Tool Parameters

Is this library paired- or single-end?

Paired Collection ▼

Select a paired collection *

accepted formats ▼ 2: Paired-end data (fastq-dump) ▼

See help section for an explanation of dataset collections

Adapter sequence to be trimmed

Automatic detection ▼

Trims 1 bp off every read from its 3' end.

☒ Yes

15. Use **Flatten collection** to flatten the collocation of Trimmed reads, then use the **FastQC** to check the quality for post-trimmed reads, lastly use **MultiQC** to inspect the result of the adapter content again, and answer **the Q5 in the assignment.**
16. Download the Ensembl gene annotation for Drosophila melanogaster by click on this link: https://zenodo.org/record/6457007/files/Drosophila_melanogaster.BDGP6.32.109_UCSC.gtf.gz
17. Upload this file to Galaxy and verify that the datatype is **gtf**, and that the database is **dm6**. Click on Save to upload.

Attributes
Datatypes
Permissions

Name

Drosophila_melanogaster.BDGP6.32.109_UCSC.gtf.gz

Info

uploaded gtf file

Annotation - optional

Add an annotation or notes to a dataset; annotations are available when a history is viewed.

Database/Build - optional

D. melanogaster Aug. 2014 (BDGP Release 6 + ISO1 MT/dm6) (dm6)

! Please provide a value for this option.

Number of comment lines

Save
Auto-detect

18. Use **RNA STAR** with the following parameters to map your reads on the reference genome: (This process takes about 2 hours 40 mins)
 - "Single-end or paired-end reads": **Paired-end (as collection)**
 - param-collection "RNA-Seq FASTQ/FASTA paired reads": **the Trim Galore! on collection N: Reads** (output of Trim Galore! tool)
 - "Custom or built-in reference genome": **Use a built-in index**
 - "Reference genome with or without an annotation": **use genome reference without builtin gene-model but provide a gtf**
 - "Select reference genome": **Fly (Drosophila melanogaster): dm6 Full**
 - param-file "Gene model (gff3,gtf) file for splice junctions": **the imported Drosophila_melanogaster.BDGP6.32.109_UCSC.gtf.gz**
 - "Length of the genomic sequence around annotated junctions": **36**
 - "Per gene/transcript output": **Per gene read counts (GeneCounts)**
 - "Compute coverage": **Yes in bedgraph format**
19. Use **MultiQC** to aggregate the STAR logs with the following parameters, and answer **the Q6 in the assignment.**

Tool Parameters

Results

1: Results ▲ ▼ 🗑

Which tool was used generate logs?

STAR ▼

Software name

STAR output

1: STAR output ▲ ▼ 🗑

Type of STAR output?

Log ▼

STAR log output *

📄 📁 ...

705: RNA STAR on collection 602: log ×

accepted formats ▼
switch to column select ▼

20. Inspect the counts from any sample in the collection RNA STAR on collection N: **reads per gene**, and answer the **Q7 in the assignment**.

21. Use **Select last lines from a dataset(tail)** from a dataset to remove the first 4 lines with the following parameters:

Tool Parameters

Text file *

📄 📄 📁 ...

708: RNA STAR on collection 602: reads per gene ▼

accepted formats ▼

! This is a batch mode input field. Individual jobs will be triggered for each dataset.

Operation *

Keep everything from this line on ▼

Number of lines *

5

These will be kept (depending on 'operation'). (--lines)

22. **Cut columns from a table** with the following parameters:

 **Cut columns from a table** (Galaxy Version 1.0.2)    

Tool Parameters

Cut columns - optional

c1,c2

Delimited by *

Tab

From *



811: Select last on collection 708

accepted formats ▼

23. Rename the collection as **FeatureCount-like files**

24. Use the tool **Gene length and GC content**, to extract the gene lengths for each gene, parameters are used as follows:

 **Gene length and GC content** from GTF and FASTA file (Galaxy Version 0.1.2)    

Tool Parameters

Select a built-in GTF file or one from your history

Use a GTF from history

Choose history if you don't see the correct GTF

Select a GTF file *



704: Drosophila_melanogaster.BDGP6.32.109_UCSC.gtf.gz

accepted formats ▼

Make sure that the GTF corresponds to the same genome as the FASTA

Analysis to perform

gene lengths only

25. Use tool **Sort**, to sort the gene with descending order, inspect the result and **answer the Q8 in the assignment**. Parameters are used as follows:

Sort data in ascending or descending order (Galaxy Version 9.3+galaxy1)

Tool Parameters

Sort Query *

838: Sort on data 837

accepted formats ▼

Number of header lines *

0

These will be ignored during sort.

Column selections

1: Column selections

on column - optional

Column: 2

in *

☐ Ascending order
 ☒ Descending order

Flavor *

☒ Fast numeric sort (-n)
 ☐

Assignment 8 (2%)

- Complete the Microarray simulation:
<https://learn.genetics.utah.edu/content/labs/microarray>
- Answer the questions below:
 - What's the purpose of using the green and red labeling mix?
 - In the result, what does green, red, yellow and black mean?

DANOVA

aTukeyta analysis assignment part 1/2 (5%)

- Download the data analysis assignment part one word file.
- Complete the questions and submit them before the next class starts.

Lab 9 Scientific method and research project

Objective

- To become familiar with the scientific method, hypothesis, controls, constants, independent, and dependent variables.
- To practice experimental design.
- To become familiar with scientific literature, resources, and citations.
- To organize the experiment for the group research project

Introduction

Scientific methods are the technique used in the construction and testing of a scientific hypothesis ("an educated guess"). The scientific method is critical to the development of scientific theories, which explain empirical (experiential) laws in a scientifically rational manner. In a typical application of the scientific method, a researcher develops a hypothesis, tests it through various means, and then modifies the hypothesis based on the outcome of the tests and experiments. The modified hypothesis is then retested, modified, and tested again until it becomes consistent with observed phenomena and testing outcomes.

Control groups are the standard against which comparisons are made in an experiment. Many experiments are designed to include positive and/or negative control groups and one or more experimental groups. The **positive control** contains the variable for which you are testing. This group produces a positive reaction and demonstrates the test's ability to detect the expected target. A positive reaction to a positive control demonstrates that your test functioned correctly. A **negative control** does not contain the variable for which you are testing. It contains only the solvent (often distilled water with no solute) and does not react in the test. A negative control demonstrates the expected negative result.

Constant is a quantity that does not change. Although you can measure a constant, you cannot alter the constant during an experiment. Conversely, an experimental independent or dependent variable is the part of an experiment that is altered.

An Independent variable is a variable that is altered or manipulated in a scientific experiment.

A *Dependent variable* reflects the effects of the dependent variable. Its value depends on changes in the independent variable.

A *hypothesis* is a precise, testable statement of what the researcher(s) predict will be the study's outcome. It is stated at the start of the study. The hypothesis usually involves proposing a possible relationship between the independent variable (what the researcher manipulates) and the dependent variable (what the research measures). A fundamental requirement of a hypothesis is that it can be tested and supported or rejected. There are two types of hypotheses: 1) a **null hypothesis** states that there is no relationship between the two variables under investigation (one variable does not affect the other), thus, there will be no changes in the dependent variable due to the manipulation of the independent variable; 2) the **alternative hypothesis** states that there is a relationship between the two variables under investigation (the independent variable causes an effect on the dependent variable), thus, one variable has an effect on the other. The alternative hypothesis can further be divided into a non-directional or directional hypothesis. A **non-directional (two-tailed) hypothesis** predicts that the independent variable will affect the dependent variable, but the direction of the effect is not specified. A **directional (one-tailed) hypothesis** predicts the nature of the effect of the independent variable on the dependent variable. This hypothesis predicts in which direction the change will take place, i.e., greater, smaller, less, or more.

Core elements of optimal experimental design: 1. **Comparison/control/baseline:** In an experiment to measure the *dependent variable* and the effect of treatment with the *independent variable*, it is important to assess the effect produced without the *independent variable*. Therefore, it is a good practice to always include a *negative* and *positive control* in your experimental design to ensure your data have a baseline for comparison. The controls also prove that your experimental materials were used properly, and procedures were conducted correctly. 2. **Blocking/controlling for confounders:** This is appropriate in an experiment with possible external sources of variability other than your *independent variable*. In statistics, a **confounder** is a variable that influences both the *dependent* and *independent variables*, causing a possible association. Controlling for *confounders* is best achieved through blocked design. In this design, we divide subjects into groups corresponding to the levels of a confounding factor and repeat the experiment for each group. By accounting for the effect of confounding variables, we can avoid being misled by our data. For example, an educational intervention program may have different effects on students of different genders. Gender in this case is a *confounder* that could influence the results and students should be divided by gender when conducting the study. 3. **Replication:** An optimal experiment should be reproducible, yielding consistent results in each replicate. Increasing sample numbers and repeating experiments at least three times will demonstrate the reproducibility and accuracy of results.

Placebo effect: When you administer a treatment to a subject, that treatment may be ineffective but can still produce a significant effect due to the existence of the treatment. For example, a fake surgery can produce the same effect as actual knee surgery. Additionally, providing an individual with nonalcoholic beverages but telling them the beverages are alcoholic can produce

the symptoms of inebriation. Because of the placebo effect, the development of medical treatments demands a robust standard of outperforming a placebo rather than outperforming the absence of treatment. This is considering that a placebo alone could result in a startling improvement of a test subject's condition, often largely due to psychological factors. To examine the benefit of a placebo, an experiment could have a control group that receives no treatment, a placebo group that receives a sham treatment, and a treatment group that receives the actual treatment under investigation. This ensures that the researcher can measure both the effect of the placebo over the baseline and the effect of the treatment over the placebo.

Hawthorne effect: The Hawthorne effect is a type of reactivity in which individuals modify an aspect of their behavior in response to their awareness of being observed. In a famous experiment in the early 20th century, a factory called Hawthorne Works sought to measure the effect of lighting on productivity. An experimental group had their light bulbs changed, and the experimenters measured the effect of this on productivity. A control group saw workers change their bulbs, but the new bulbs were identical to the old ones. However, both groups had improved productivity after the "new" bulbs were provided. Therefore, the control group improved purely due to their belief that new bulbs were in place.

Scientific Literature and Publication

Sharing the results of scientific experiments is one of the most important aspects of science. Our current knowledge of science and the world around us is dependent upon scientists sharing their results. Scientists share their discoveries with the public through scientific publications in journals and books. Other scientists can then find this information and use it to help them design experiments, contextualize their own research, and increase our overall knowledge base. One of the most common ways for scientists to formally share their research is to publish a **journal article**. The journal article is a part of the formal written record of the scientific process and results. A credible journal is usually peer-reviewed. **Peer-reviewed publications** are written by scholars for scholars and are reliable sources of information on a topic or discipline. These journals can be found either in the library's online databases or in the library's local holdings. **Peer review** is a process where an article is verified by a group of scholars before it is published. When an author submits an article to a peer-reviewed journal, the editor distributes the article to a group of scholars in the related field (the author's peers). These peers review the article, ensuring that its sources are reliable and that the information it presents is consistent with the research. The reviewers determine whether the article will be accepted for publication, requires revisions, or should be rejected. The peer-review process ensures that only research meeting a quality standard is published. There are two types of scientific articles that can be published: **review articles** and **original research articles**. If the authors are discussing research and experiments that they carried out, and providing an outline of the experiment, it will be an **original research article**. In **review articles**, authors discuss experiments carried out by other scholars. Review articles usually summarize the findings of many original research articles regarding a specific topic. Conversely, original research articles describe the original research methods, materials, and data of a study.

To search for credible scientific publications, we usually recommend you use the “**OneSearch**” at LaGuardia’s online library (<https://library.laguardia.edu/>) or **PubMed**, a free search engine accessing primarily the MEDLINE database of references and abstracts on life sciences and biomedical topics, developed by the United States National Library of Medicine at the National Institutes of Health. Searching for information on a research topic using popular search engines may reveal unreliable resources. Shorter news articles may appear in certain scientific and popular publications. These news articles are not peer-reviewed and are normally written by science journalists, not researchers. The news articles may be easier to read, but since they are not original research, they may be an unreliable source for your paper or project. However, news articles can lead you to original research and help you stay informed on recent research developments. If you conduct your searches online, via Google, Yahoo, or another popular search engine, you may find journal articles, but you may also come across other scientific information, which can take many forms. Wikis, blogs, and personal websites can often contain a significant amount of scientific information, but these resources are generally very far removed from the original research. Each of these sources requires careful evaluation to determine their credibility. Additionally, these sources are not suitable for a research paper.

A **scientific literature review** is a critical account of published material on a topic by accredited researchers. We often need to write a scientific literature review for assignments, an introduction to an essay, a report, a thesis, and as part of research or grant proposals. Writing a literature review will improve your knowledge of a topic, provide new insight on the topic to readers, and help you develop literature searching, critical analysis, communication, and writing skills. A scientific literature review needs to follow the style of scientific writing. Writing requires precision, clarity, objectivity, and simple language.

Lab Exercise I - Experimental Design Practice

Activity One: Identifying and creating scientifically testable hypotheses; practicing optimal experimental design

1. Evaluate the following hypotheses and formulate experimental designs to test them. While designing experiments, identify and include proper controls, independent variables, dependent variables, constants, and confounders. Additionally, consider the potential placebo and Hawthorne effect.
 - 1). Classic music helps plants grow.
 - 2). Exercise helps people be healthier.

3). Men are stronger than women.

4). DNA was brought to the planet by aliens.

2. Formulate a hypothesis based on the following case study and design experiments to test your hypothesis. While designing experiments, identify and include proper controls, independent variables, dependent variables, constants, and confounders. Additionally, consider the potential placebo and Hawthorne effect.

Case Study: During the 2004 Tsunami disaster in Asia, there were reports that trained elephants had moved away from the coastline before the tsunami struck. Scientists wondered if elephants may have sensed the oncoming wave and earthquake that created it. Elephant trunks have two kinds of nerve endings that are sensitive to pressure differences, but little is known about nerve ending distribution in the elephant's skin or feet. An elephant has recently died at a nearby zoo and scientists have decided to dissect one of its feet to reveal any special nerve endings. Results indicate a high density of pressure-sensitive nerve endings in the front of the foot and along the edges, which supports the notion that elephants can sense seismic signals produced by earthquakes.

Activity Two: Evaluating Experimental Design

Case study: The Food and Drug Administration (FDA) asked two groups of scientists to design an experiment to test a new drug that potentially helps people lose weight. Read the experiments designed by these two groups of researchers and answer the following questions.

Group 1: 50 women and 50 men with body mass index (BMI) between 30 and 33% were chosen as the participants for this experiment. All participants had similar health, age, and ethnicity. All participants received instructions on their diets, exercise, and sleep time before starting the experiment. During the 12-week period of the experiment, half the participants (25 women and 25 men) received identical pills without medication (placebos). None of the participants in either group knew which pill they were receiving. The scientists who analyzed the data from this experiment did not know which participant received either the treatment or placebo pills.

Group 2: 50 women and 50 men with BMI between 30 and 33% were chosen as the participants for this experiment. The participants were in the same age range. All participants received instructions on their diets, exercise, and sleep time before starting the experiment. During the 12-week period of the experiment, half the participants (25 women and 25 men) received a pill every 12 h. The other half of the participants did not receive any medication.

1. Which experiment design is more reliable between the two groups? Provide three reasons
2. Which was the control group for each of these experimental designs? Which group designed the better control group? Why?
3. What were the independent and dependent variables?
4. What confounders were included in this experimental design?
5. Why is it important that neither the participants nor the scientists knew which participant received the weight loss medication?

Lab Exercise II - Formulate a Hypothesis and Design an Experiment Using *Drosophila*

Background information: Diet-mediated diseases represent an increasing challenge in Western society. Research has focused on carbohydrate consumption, which has increased significantly in the past three decades. Dietary sugars have been linked to a variety of metabolism-related problems including obesity, insulin resistance, metabolic syndrome, type 2 diabetes mellitus (T2DM), and cardiovascular disease. *Drosophila melanogaster* provides a useful complement to mammalian models. While flies show important differences from mammalian model systems, they also show marked similarities. For example, the fly genome encodes seven insulin-like peptide genes (*dilp* 1–7) that activate classical insulin pathway signaling. Ablation of insulin-producing cells phenocopied several aspects of type 1 diabetes. Furthermore, *Drosophila* is susceptible to diet-mediated metabolic and cardiac organ dysfunctions that are similar to those reported in mammals. Therefore, *Drosophila* serves as a proper model system to study the impact of dietary sugar on physiology and behavior.

The objective of the study: Work in groups to identify the impact of increased dietary sugar on the weight, life span, and reproduction of *Drosophila*.

Group activity: Students are divided into 5-6 groups, discuss, formulate a testable hypothesis, and draft a research plan, including proper controls, independent variables, dependent variables,

constants, and confounders. Subsequently, students will start conducting the research, collect data and report findings during the next few labs.

Lab Exercise III - Set up your experiment

Need 5-6 groups with the same set up. Based on the Hypothesis, set up your experiment.

Materials:

Formula 4-24® Instant Drosophila Medium

1 box of new vials

1 bag of cotton balls

1 Fly nap

1 Sucrose

1 Dry yeast

6 Distilled water bottles

6 digital scales

6 big weighing boats

6 small weighing boats

6 graduated cylinder(for measuring 5ml water)

6 spoons

Methods:

1. Obtain 5 vials and label them with : 0% , 1%, 2%, 5%, and 10% **sucrose**.

(For yeast: Obtain 5 vials and label them with : 0% , 0.2%, 1%, 2%, and 5% **Yeast**.)

2. Use a big weighing boat to weigh 1.2g of Formula 4-24 white powder and add to each vial.

3. Use small weighing boat to weigh certain Sucrose/Yeast according to the concentration of each vial. (**1% x 5ml = 0.05g Sucrose; 0.2% x 5ml = 0.01g yeast**)

4. Add the Sucrose/Yeast to its vial and swirl to mix.

5. Measure 5 ml water with graduated cylinder.

6. **Quickly** pour all the water to the vial, **quickly** swirl the vial to mix the media evenly.

7. Do step 6 for each vial. Wait 3 mins for food solidification.

8. Transfer 5 male and 5 female (same strain) adult flies to each vial.

9. Cap vials with cotton balls, incubate for 1 week.

Research project essay -10%

Following the Scientific Publication Format, minimum 1200 words and minimum 2 citations.

1. **Title of your research**

2. **Authors and Affiliations**

3. **Introduction/Background**

4. **Methods** (Refer to Lab 9 and 10, 11 lab content)
5. **Results:** including text explanation, non-textual elements (tables, pictures, charts, graphs). All tables and figures must have their titles and figure legends. We expect you to include descriptive statistics, histograms, bar graphs, and tables to illustrate your results. Make sure you explain clearly and label accurately.
6. **Conclusions** (one paragraph to summarize key findings)
7. **Discussions** (one or two paragraphs)
8. **References**(Proper citations are required in APA format, Minimum citations: 2)

Complete research essay by this week part 1/3:

1. Introduction – Students will draft the Introduction of their research project with literature review about your hypothesis. Including the background of your research and addressing the importance of this research and how the issue addressed by the research is a global issue.
2. Method – Start the Methods section (how to set up fruit fly vials.)

Lab 10 Methods, Results, Animal Behavioral Analysis, Data Collection and Research Project (Continued)

Na Xu

Objective

- To become familiar with the Data Analysis, and Method section of a research paper
- To conduct pupa analysis and animal behavioral analysis
- To continue the experiment for the group research project

Introduction Methodology of Research Projects

The methods section of a research paper describes how experiments are conducted and provides the information by which a study's validity is judged. The method section answers three main questions: 1) How was the data collected or generated? 2) What materials were used? 3) How were the data analyzed? The writing should be direct, precise, and written in the past tense. **An effectively written methodology section should:**

- **Introduce the overall methodological approach for investigating your research problem.** You need to include all the materials used and list all the steps of the experiments.
- **Indicate how the approach fits the overall research design.** Your methods should have a clear connection with your research problem. In other words, make sure that your methods directly address the problem. You should provide background and a rationale for methodologies that are unfamiliar to the readers. You should also provide a rationale for subject selection and sampling procedures. For instance, if you propose to conduct interviews, how do you intend to select the sample population? If you are analyzing texts, which texts have you chosen, and why? If you are using statistics, why is this set of statistics being used? If other data sources exist, explain why the data you chose is most appropriate. You need to clearly state the controls and variables in your experiment and explain the rationale for using these chosen controls and variables.
- **Describe the specific methods of data collection,** such as experiments, surveys, interviews, questionnaires, observation, and archival research. If you are analyzing existing data, such as a data set or archival documents, describe how it was originally created or gathered and by whom.
- **Explain how you intend to analyze your results.** You should include any statistical analysis, software, and tools used to analyze data, as well as the rationale for using those tools.
- **Address problems and limitations.** It is very likely that you will encounter problems when collecting or generating your data. Do not ignore these problems. Often, documenting how you overcame obstacles can form an interesting part of the methodology. This demonstrates to the reader that you can provide a coherent rationale for the decisions you made to minimize the impact of any problems that arose. Nearly all methodologies have limitations. Are there any practical limitations that could affect your data

collection? How will you attempt to control for potential confounding variables and errors? You should openly state the problems that may arise due to the methodology.

The methods section of a research paper describes how experiments are conducted and provides the information by which a study's validity is judged. The method section answers three main questions:

- 1)How was the data collected or generated?
- 2)What materials were used?
- 3)How were the data analyzed?

The writing should be direct, precise, and written in the past tense. An effectively written methodology section should:

Introduce the overall methodological approach for investigating your research problem. You need to include all the materials used and list all the steps of the experiments.

Indicate how the approach fits the overall research design. Your methods should have a clear connection with your research problem. In other words, make sure that your methods directly address the problem. You should provide background and a rationale for methodologies that are unfamiliar to the readers. You should also provide a rationale for subject selection and sampling procedures.

For instance, if you propose to conduct interviews, how do you intend to select the sample population? If you are analyzing texts, which texts have you chosen, and why? If you are using statistics, why is this set of statistics being used? If other data sources exist, explain why the data you chose is most appropriate. You need to clearly state the controls and variables in your experiment and explain the rationale for using these chosen controls and variables.

Describe the specific methods of data collection, such as experiments, surveys, interviews, questionnaires, observation, and archival research. If you are analyzing existing data, such as a data set or archival documents, describe how it was originally created or gathered and by whom.

Explain how you intend to analyze your results. You should include any statistical analysis, software, and tools used to analyze data, as well as the rationale for using those tools.

Address problems and limitations. It is very likely that you will encounter problems when collecting or generating your data. Do not ignore these problems. Often, documenting how you overcame obstacles can form an interesting part of the methodology. This demonstrates to the reader that you can provide a coherent rationale for the decisions you made to minimize the impact of any problems that arose. Nearly all methodologies have limitations. Are there any practical limitations that could affect your data collection? How will you attempt to control for potential confounding variables and errors? You should openly state the problems that may arise due to the methodology.

The Results section

Data analysis typically involves multiple activities such as collecting, evaluating, and organizing the data. Data analysis is usually shown in the “**Results**” section of the paper. The **Results** section of the research paper is where you report the findings of your study based on the information gathered from your experiments. The results section should simply state the findings, without bias or interpretation, and be arranged in a logical sequence. The results section should always be written in the past tense. A section describing results or findings is particularly necessary if your paper includes data generated from your own research. **In general, the content of your results section should include the following elements:**

- An introductory context for understanding the results by restating the research problem that underpins the purpose of your study.
- Inclusion of **non-textual elements**, such as figures, charts, photos, maps, and tables, among others to further illustrate the findings, if appropriate.
- Inclusion of **figure legends** explaining the non-textual elements presented.
- In the text, a systematic description of your results, highlighting for the reader the most relevant observations for the topic under investigation.
- Use the past tense when referring to your results.

Using Non-textual Elements: You can place figures, tables, and charts, among others within the text of the result at the end of the report as a separate section called “Figures”. Below are several rules for using non-textual elements in your report.

- In the text, refer to each non-textual element in numbered order (e.g., Table 1, Table 2; Chart 1, Chart 2; Map 1, Map 2; Figure 1, Figure 2).
- Each non-textual element must be titled, numbered consecutively, and complete with a figure legend explaining the results
- In proofreading your results section, be sure that each non-textual element is sufficiently complete so that it could stand alone, separate from the text.
- Incorporate non-textual elements well into the textual writings.

Problems to avoid: When writing the results section, avoid the following:

- **Ignoring negative results.** If some of your results fail to support your hypothesis, do not ignore them. Document them, then state in your discussion section why you believe a negative result emerged from your study. Note that negative results, and how you handle them, often provide you with the opportunity to write a more engaging discussion section, therefore, don't be afraid to highlight them.
- **Including raw data or intermediate calculations.** All the data in the Result section should be analyzed and organized. Ask your professor or supervisor if you need to include any raw data generated by your studies, such as transcripts from interviews or data files. If raw data is to be included, place it in an **Appendix** or **Supplementary Material** section that is referred to in the text.

- **Using vague, not confident, or unspecific wording.** Do not use phrases that are vague or unspecific, such as, "appeared to be greater or lesser than..." or "demonstrates promising trends that...."
- **Presenting the same data twice or repeating information.** If certain data needs to be highlighted, you will have a chance to do so in the discussion section.
- **Discuss your results.** Save all discussion for the **"Discussion"** section of your paper.

Pupa counts- Records your data in Excel

The pupa stage in the life cycle of *Drosophila*, commonly known as fruit flies, is an intermediate stage between the larva and adult stages. During this stage, the larva undergoes a process called metamorphosis to transform into an adult fly. After the larva has completed its growth and development, it molts for the last time and enters the pupa stage. The pupa is enclosed within a protective case called the puparium, which is formed by the hardening and darkening of the larval skin. Inside the puparium, the insect undergoes significant changes in its body structure and organs. During the pupa stage, the fly undergoes a complete transformation, including the formation of adult body parts, wings, legs, and internal reproductive organs. However, these structures are not visible from the outside of the puparium. Inside, the pupa undergoes a process of remodeling and reorganization of its tissues and organs.

The duration of the pupa stage varies depending on various factors such as temperature, nutrition, and genetic factors. In favorable conditions, the pupa stage of *Drosophila* typically lasts for about four to six days. At the end of this stage, the adult fly emerges from the puparium by splitting it open. The newly emerged adult fly is initially soft and pale but quickly expands and darkens as its cuticle hardens.

Animal Behavioral Studies

Ethology is the scientific study of animal behavior, particularly in their natural habitats. It seeks to understand and explain the behaviors and activities of animals, including their interactions with the environment, other individuals of the same species, and different species.

Ethologists examine a wide range of behaviors, including those related to feeding, reproduction, communication, social interactions, migration, territoriality, aggression, and learning. They aim to uncover the underlying mechanisms, evolutionary significance, and adaptive functions of these behaviors.

Ethologists employ various research methods and techniques to study animal behavior. These may include observational studies conducted in the field or in controlled laboratory settings, experiments to test specific hypotheses, and the use of sophisticated technologies such as tracking devices, genetic analyses, and neurophysiological recordings.

One of the fundamental concepts in ethology is the idea that behavior is influenced by both innate (genetically programmed) and learned factors. Innate behaviors are those that are instinctive and occur without prior experience or learning. These behaviors are typically associated with survival, reproduction, and basic physiological needs. Learned behaviors, on the

other hand, are acquired through experience, observation, and interactions with the environment and conspecifics (members of the same species).

Ethology also examines the ecological and evolutionary factors that shape animal behavior. It explores how behaviors are influenced by factors such as resource availability, predation pressure, competition for mates and territories, and social dynamics within groups or communities.

The study of ethology has contributed significantly to our understanding of animal behavior, providing insights into the diversity, complexity, and adaptive nature of behaviors across different species. It has practical applications in fields such as conservation biology, animal welfare, and the management of human-animal interactions, helping us develop a deeper appreciation and understanding of the animal world.

What is ethology?

Ethology, or animal behavior, is the study of how animals interact with each other and their environment(s).

What are taxis and kinesis?

Taxis and kinesis are both directional behaviors. Taxis is stimulus-induced movement towards a specific direction. Conversely, kinesis is stimulus-induced movement in random directions.

How do we name taxis behaviors?

Taxis behaviors are named after the type of triggering stimuli and the direction of movement. For example, both food and bug spray are chemical stimuli. Movement towards food is positive chemotaxis, whereas movement away from the bug spray is negative chemotaxis.

What is a choice chamber?

A choice chamber is a device with two sides that allows detection of response to a stimulus by presenting the stimulus on one side and counting the number of flies on both sides at certain time points.

Lab Exercise I: Data collection

Materials

Vials prepared in lab 9

1 Fly nap

1 box of new vials

1 bag of cotton balls

1 roll of Aluminum foils

6 choice chambers(made with 2 50ml falcon tube)

6 Stereo microscopes

6 small Weighing boats

6 digital scales

6 Forceps
6 brush
6 scissors

Methods:

1. Pupa Analysis:

To evaluate the impact of independent variables on fly development, you can count how many pupa were generated at day 5, 6, 7, etc, and record the data in the following table:

Table 1: Number of Pupa

<i>Number of Pupa</i>	Flies grown in normal medium	Flies grown in condition 1	Flies grown in condition 2
Day 5			
Day 6			
Day 7			
Day 8			

2. Test for Geotaxis

- Once the *Drosophila* is in the choice chamber, close the tape flap, rotate the chamber so that it stands vertically with **side A on the bottom, and start the timer**.
- Immediately**, count the number of *Drosophila* that are **moving up and the number that are moving down**. Record these data **for time point 0** in the table.
- After 30 seconds**, rotate the chamber so that **side B is on the bottom** and count the number of *Drosophila* that are moving up or down. Record these data for time **point 30** in the table.
- After 60 seconds**, rotate the chamber so that **side A is on the bottom** again and record the number of moving flies for **time point 60**.

Table 2: Geotaxis analysis

Time (sec)	Side A Orientation	# Flies moving down (control)	# Flies moving up (control)	# Flies moving down (condition 1)	# Flies moving up (condition 1)	# Flies moving down (condition 2)	# Flies moving up (condition 2)
0	Down						
30	Up						
60	Down						

3. Test for Phototaxis

- To perform a phototaxis analysis, first cover **side A of the choice chamber with tin foil**, taking care that the foil reaches to the exact center between the two sides.
- Set the timer to sound an alarm at 10 minutes.
- When the alarm sounds, stop the timer, and count the number of flies in the open side B of the chamber.
- Subtract that number from the total number of flies added to the choice chamber to calculate the number of flies in the covered side A.
- Record these data in the table for **time point 10**.

Table 3 Phototaxis analysis

Time (min)	# Flies side A (dark) control	# Flies side B (light) control	# Flies side A (dark) condition 1	# Flies side B (light) condition 1	# Flies side A (dark) condition 2	# Flies side B (light) condition 2
10						

4. Behavioral Analysis: Fly mobility

Tap down flies and take pictures every 2 seconds to record how many flies fly up at each time point: flies with less mobility do not fly actively.

Table 4 Fly mobility analysis

	Flies grown in normal medium	Flies grown in condition 1	Flies grown in condition 2	Flies grown in condition 3	Flies grown in condition 4
Number of flies fly up at 2 sec					
Number of flies fly up at 4 sec					
Number of flies fly up at 6 sec					

Number of flies fly up at 8 sec					
---------------------------------	--	--	--	--	--

5. Measuring the weight of adult flies(Parental or F1)

- 1) Use fly nap to anesthetize the adult flies and analyze the aspects listed below for the F1 progeny.
- 2) Measuring the **average weight**
total weight
number of flies
number of females vs males

$$\text{Average mass (g)} = \text{total mass (g)} / \text{number of flies}$$
- 3) You can also compare the **total number of F1 flies** in different conditions to evaluate reproduction efficiency. These results are supposed to be consistent with your pupa counts.
- 4) Record your data in Excel and keep pictures.

Table 5 Average Mass and Survivability

	Flies grown in normal medium	Flies grown in condition 1	Flies grown in condition 2
Total Mass (g)			
Number of flies			
Average mass (g)			
Standard Deviation			

Introduction of data analysis

Data analysis is the process of inspecting, transforming, and modeling data to discover useful information, draw conclusions, and support decision-making. It involves various techniques and methods to extract meaningful insights from raw data. Here are some key steps typically involved in data analysis:

1. **Data Collection:** Gathering relevant data from various sources, such as databases, spreadsheets, surveys, or online repositories.

2. **Data Cleaning:** Removing errors, inconsistencies, or irrelevant information from the dataset. This step often includes handling missing data, dealing with outliers, and standardizing formats.
3. **Data Exploration:** Examine the dataset to understand its structure, patterns, and relationships. This may involve summarizing data, creating visualizations, and calculating descriptive statistics.
4. **Data Preprocessing:** Transforming the data into a suitable format for analysis. This can include tasks like normalization, feature scaling, or encoding categorical variables.
5. **Statistical Analysis:** Applying statistical techniques to draw inferences, identify correlations, and test hypotheses. This may involve techniques like regression analysis, hypothesis testing, or analysis of variance (ANOVA).
6. **Data Modeling:** Building models or algorithms to make predictions or gain a deeper understanding of the data. Common techniques include machine learning algorithms, time series analysis, or clustering algorithms.
7. **Interpretation and Visualization:** Communicating the results of the analysis to the audience in a clear and meaningful way. This often involves creating visualizations, charts, or reports to present the findings.

Throughout the data analysis process, it is important to ensure data privacy, maintain data integrity, and apply ethical considerations when handling sensitive or personal information. Various tools and programming languages are commonly used for data analysis, including Python, R, SQL, Excel, and Tableau, among others. These tools provide functionalities for data manipulation, visualization, and statistical analysis, making the data analysis process more efficient and effective.

Excel is a widely used software tool that offers many built-in features for performing statistical analysis. Here are some common statistical tasks you can accomplish using Excel:

1. **Descriptive Statistics:** Excel provides functions like AVERAGE, MEDIAN, MODE, MIN, MAX, COUNT, and SUM for calculating basic descriptive statistics on a dataset. These functions can help you understand the central tendency, dispersion, and other summary measures of your data.
2. **Data Visualization:** Excel has various chart types, including histograms, scatter plots, bar charts, and line graphs, which you can use to create visual representations of your data. These charts can help you identify patterns, trends, and outliers in your dataset.
3. **Data Analysis ToolPak:** Excel includes the Data Analysis ToolPak add-in, which provides a wide range of advanced statistical analysis tools. It offers features for regression, correlation, t-tests, ANOVA, exponential smoothing, and more. You can activate the ToolPak by going to File > Options > Add-Ins > Manage Excel Add-ins.

These are just a few examples of what you can do with Excel for statistical analysis. Excel's flexibility and ease of use make it a popular choice for performing basic to intermediate statistical tasks. However, for more complex analyses or specialized statistical techniques, you

may need to consider using dedicated statistical software or programming languages like R or Python.

Lab Exercise II In-class data analysis using Excel

Hypothesis: the availability of light impacts drosophila body weight

Experiments: A number of male and female flies are grown in the same food resources for 14 days. One group of flies was exposed to light constantly. The other group of flies were kept in the dark and were never exposed to any light. At the end of the 14 days, all flies' body weights were measured.

Sample Data:

<https://www.dropbox.com/s/ya96v722bra1ojh/Assignment%209%20data%20set.xlsx?dl=0>

Sorting the data:

1. Open Data set.
2. Arrange the data for analysis:
 - 1) Select data from B4 to D22
 - 2) Click on Sort "A->Z". You should be able to sort the data based on whether the fly was exposed to light or not.
 - 3) Select data from C4 to D12, for only "no light" group.
 - 4) Click on sort "AàZ". You should be able to sort the data based on the sex of the flies.
 - 5) Select data from C13 to D22, for only the "exposed to light" group.
 - 6) Click on sort "AàZ". You should be able to sort the data based on the sex of the flies.

Add Excel Tools for Data Analysis

1. Click "File"
2. Options
3. Add-Ins
4. Manage "Excel Add-ins", Click on "Go"
5. Select "Analysis ToolPack" and "Solver Add-in". Click "OK".

Apply Descriptive Statistics

Descriptive statistics refer to summary measures that provide an overview and description of a dataset. They help to understand the central tendency, dispersion, shape, and other characteristics of the data. Excel offers various built-in functions to calculate descriptive statistics. Here are some commonly used descriptive statistics in Excel:

1. **Mean:** The mean (average) represents the sum of all values divided by the number of observations. You can calculate the mean using the AVERAGE function in Excel. For example, if your data is in cells A1 to A10, you can use the formula "`=AVERAGE(A1:A10)`" to calculate the mean.
2. **Median:** The median is the middle value in a sorted dataset. To calculate the median, you can use the MEDIAN function in Excel. For instance, if your data is in cells A1 to A10, the formula "`=MEDIAN(A1:A10)`" will give you the median.
3. **Mode:** The mode represents the value that appears most frequently in a dataset. Excel does not have a built-in function for mode, but you can use a combination of functions like MODE, MODE.MULT, or MODE.SNGL (depending on whether there is one mode or multiple modes) available in the Data Analysis ToolPak add-in.
4. **Standard Deviation:** The standard deviation measures the dispersion or variability of data around the mean. You can calculate the standard deviation using the STDEV.S or STDEV.P function in Excel. STDEV.S is used for a sample, while STDEV.P is used for a population. For example, if your data is in cells A1 to A10, you can use the formula "`=STDEV.S(A1:A10)`" to calculate the sample standard deviation.
5. **Variance:** Variance is the average of squared deviations from the mean. Excel provides the VAR.S and VAR.P functions to calculate the sample variance and population variance, respectively. The formulas "`=VAR.S(A1:A10)`" and "`=VAR.P(A1:A10)`" can be used to calculate the sample and population variances.
6. **Minimum and Maximum:** The MIN and MAX functions in Excel can be used to find the smallest and largest values in a dataset, respectively. For instance, if your data is in cells A1 to A10, "`=MIN(A1:A10)`" and "`=MAX(A1:A10)`" will give you the minimum and maximum values.
7. **Count:** The COUNT function in Excel can be used to count the number of non-empty cells in a range. For example, "`=COUNT(A1:A10)`" will give you the count of non-empty cells in the range A1 to A10.
8. **Range** equals Maximum – Minimum.

Steps to apply Descriptive Statistics:

1. Click on "Data" on top of the Excel
2. Click "Data analysis"
3. Click "Descriptive Statistics" and "OK"
4. Input Range: Select data from D4 to D8 for Female flies in dark
5. The Output range can be where you want analysis results to show up
6. Click "Summary Statistics".
7. Change Column 1 to "Female in Dark"
8. Repeat Steps 2–7 for the "Male in Dark", "Female in Light" and "Male in Light" groups. Put all your output next to each other so that your data can be shown as a table.

Histogram

A **histogram** is a graphical representation of the distribution of a dataset. It provides a visual depiction of the frequency or count of values falling within different intervals or bins. Excel provides built-in features to create histograms easily. Here's how you can create a histogram in Excel.

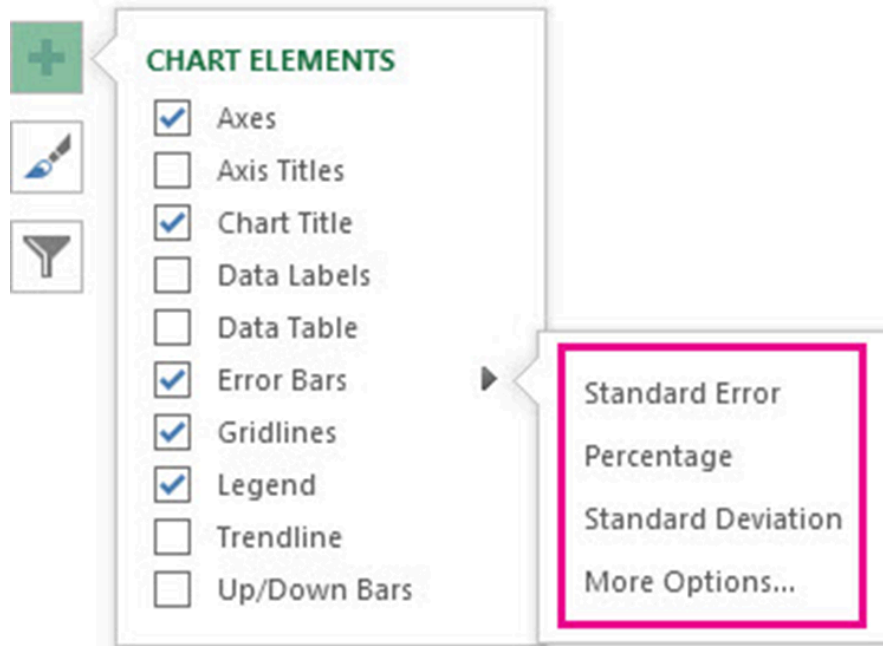
1. Write your Bin Range in Excel:
2. Click "Data Analysis"
3. Click "Histogram"--OK
4. Input Range selected from D3 to D8, for the "Female in Dark" group.
5. Select Bin range
6. Check "Labels" because, in the range, you included labels of the data "body weight " and "bins".
7. Select a space for the output range
8. Check "Pareto" and "Chart output", OK
9. Repeat Steps 2-8 for the " Male in Dark", "Female in Light" and " Male in Light" groups. You should make 4 histograms here and make sure you include a title for each histogram.

Comparing Means

1. Select the labels and data.
2. Click "Insert" and choose "Column or Bar Chart"
3. Change the title of the chart as the Average Body Weights (mg).

Add Standard Errors in the Chart

1. Select your chart by clicking on it once.
2. Click on the **Chart Elements** button next to the chart represented by a + sign, then check the **Error Bars** box. In the future, if you wish to remove error bars, simply clear the box.



Answer the questions with the graphs:

1. Copy and paste your sorted data
2. Show descriptive statistics for all four groups ("Female in Dark", "Male in Dark", "Female in Light", and "Male in Light") in a table format. Include a short explanation of this table in your own words.
3. Show histograms for four groups. You should make 4 histograms and make sure you include the correct titles for each histogram.
4. Use the Mean of all four groups to generate a chart comparing the average body weights. Include standard errors in your chart. Label the chart properly and clearly, including sample numbers. The chart should have a proper title.
5. Explain your data in detail according to histograms and charts. Draw a conclusion based on your data. Does exposure to light have an impact on flies' body weight? If Yes, does this impact depend on the sex of the fly?
6. Organize all contents and write it in a Word file. Do not submit raw analysis from Excel.

standard deviation

Lhistogram

ab Exercise III ANOVA and Tukey's HSD analysis

ANOVA is a test used to determine differences between research results from three or more independent samples or groups.

If the p-value from the ANOVA is less than the significance level, we can reject the null hypothesis and conclude that we have sufficient evidence to say that at least one of the means of the groups is different from the others.

However, this doesn't tell us which groups are different from each other. It simply tells us that not all of the group means are equal. In order to find out exactly which groups are different from each other, we must conduct a post hoc test.

The most commonly used post hoc test is the Tukey-Kramer test, which compares the mean between each pairwise combination of groups.

Material:

ANOVA&HSD: <https://www.socscistatistics.com/tests/anova/default2.aspx>

Method:

1. Input your data to each treatment box. **Please ensure that at least two consecutive text boxes are populated with data (a minimum of 5 values each), and that there are no more than 300 values in your sample.**

Treatment 1	Treatment 2	Treatment 3	Treatment 4	Treatment 5

- Pick the significant level, 0.05 for most of the case.

Significance Level:

☐ .01
☒ .05
☐ .10

- Click on the **Calculate** button.
- The output of this calculator is pretty straightforward. The values of f and p appear at the bottom of the page. If the text is blue, your result is significant; if it's red, it's not.

5. In the Post Hoc Tukey HSD results. The Tukey's HSD (honestly significant difference) procedure facilitates pairwise comparisons within your ANOVA data. A blue value for Q indicates a significant result.

Research essay part 2/3

1. Continue working on methods with the labs done today.
2. Collect data and use Excel or online tool for statistical analysis.

1st draft due on week 12

Lab 11 Discussion Sections of Research Papers, triglyceride-assay and Research Project (Continued)

Na Xu, Hanqi Ye

Objective

- To become familiar with the discussion sections of a research paper
- To collect data for the group research project
- To understand the spectrometry and metabolite analysis

Introduction

The discussion section is where you interpret and describe the significance of your findings considering what is already known about the research problem under investigation. Additionally, this section will explain any new understanding or insights revealed about the problem after your research. The discussion will always connect to the introduction by way of the research questions or hypotheses posed and the literature reviewed. However, it does not simply repeat or rearrange the introduction; the discussion should always explain how your study has improved the reader's understanding of the problem considering what is already known. This section is often considered the most critical part of a research paper because it most effectively demonstrates your ability as a researcher to think critically about an issue, develop creative solutions to problems based on the findings, and provide a deeper understanding of the research problem under investigation. The discussion section is where you **explore the underlying meaning of your research**, possible implications in other areas of study, and possible improvements to further mitigate the concerns of your research.

This is the section where you need to present **the importance of your study** and how it contributes to the field. If appropriate, the discussion section is also where you state how the

findings from your study **revealed new gaps** in the literature that had not been previously exposed or adequately described. A good discussion section includes an **analysis of any unexpected findings**. If more than one unexpected finding emerged during the study, describe each in the order they appeared as you gathered the data.

The content of the discussion section of your paper usually includes

1. **Explanation of results:** briefly reiterate the research problem or problems under investigation and the methods used to investigate them, then describe the major findings of the study. Write a direct, declarative, and succinct proclamation of the study results. You can comment on whether the results were expected and present explanations for the results. Provide more detailed explanations for findings that were unexpected or particularly significant. If appropriate, note any unusual or unanticipated patterns or trends that emerged from your results and explain their meaning. When writing the discussion section, you should carefully consider all possible explanations for the study results, rather than only those that fit prior assumptions or biases.
2. **References to previous research:** compare your results with findings from other studies or use such studies to support a claim. This can include re-visiting key sources already cited in the literature review section, or, saving them to cite later in the discussion section if they are more important for comparisons with your results than as part of the background information.
3. **Identification of potential limitations and weaknesses:** comment on the limitations and weaknesses of the research and their relative importance in relation to your overall interpretation of the results. Additionally, if necessary, note how the limitations and weaknesses may affect the validity of the findings. Avoid using an apologetic tone; however, be honest and self-critical. Describe the generalizability of your results to other situations, if applicable to the method chosen, then describe in detail problems you encountered in the method(s) used to gather data. Moreover, consider any unanswered questions or issues the study did not address.
4. **Deduction:** based on your research, you can propose a general claim or possible conclusion for the results. You can also explain how the results can be applied more generally. For example, describing lessons learned, proposing recommendations that can help improve a situation, or recommending best practices. You should explain the meaning of the findings and why they are important.
5. **Making suggestions for future studies:** although your study may offer important insights about the research problem, other questions likely remain unanswered or arise because of your study. Therefore, you should make suggestions for further research in the discussion section.

(TBC)Lab Exercise I: Triglyceride-assay

Material:

Fluorescence microplate readers(Not available)

Centrifuge

Microcentrifuge
Vials from lab 10
5% NP-40
Abcam fluorometric kit (AB65336)

Method:

Reference triglyceride levels were determined using the Abcam fluorometric kit (AB65336) following the manufacturer's instructions. Briefly, triglycerides were hydrolyzed to free fatty acids and glycerol. The glycerol reacts with the triglyceride enzyme mix to form an intermediate product, which in turn reacted with the PicoProbe and developer to generate fluorescence that can be detected at Ex/Em = 537/587 nm.

1. Experimental samples were prepared by grinding each adult fly in 100 µl of 5% NP-40/ddH₂O.
2. Samples were slowly heated to 85°C for 5 min using and then cooled down to room temperature.
3. The heating and cooling process was repeated twice, and samples were centrifuged for 2 min at 4,000 g to remove insoluble materials.
4. Triglyceride level was determined in a 96-well microplate, with each well containing 25 µl of samples and 25 µl of working buffer (the working buffer consists of 23.8 µl of triglyceride assay buffer, 0.2 µl of triglyceride probe, and 1 µl of triglyceride enzyme mix).
5. All measures are detected at Ex/Em = 537/587 nm and performed at room temperature(23-25°C).

Lab Exercise II RNA-seq data analysis - Part two

Material:

Galaxy: <https://usegalaxy.org/>

Header file:

https://drive.google.com/file/d/1hmnePtue7-_RNKTbERh4JsG23E_uV38y/view?usp=sharing

RNA-seq data analysis part 2 worksheet word file (download from the blackboard or this [link](#))

Method:

1. Create a new history by clicking on the + icon at the top of the history panel and rename as RNAseq-2

- From RNASeq history download **FeatureCount-like files** and **Gene length** files to your local computer, rename the feature-count files as follows according to the sample and diet composition table in Part 1 assignment

 SRR25724550.tabular	 PC25_1
 SRR25724549.tabular	 PC25_2
 SRR25724548.tabular	 PC25_3
 SRR25724547.tabular	 PC40_1
 SRR25724546.tabular	 PC40_2
 SRR25724545.tabular	 PC40_3
 SRR25724544.tabular	 PC55_1
 SRR25724543.tabular	 PC55_2
 SRR25724542.tabular	 PC55_3
 SRR25724541.tabular	 PC70_1
 SRR25724540.tabular	 PC70_2
 SRR25724539.tabular	 PC70_3

- Upload renamed files to RNAseq-2 history with Type (set all) as : **tabular**.

Upload from Disk or Web to **RNAseq-2**

Regular
Composite
Collection
Rule-based

	PC25_1	353 KB	Auto-detect		unspecified (?)		100%	
	PC25_2	351.5 KB	Auto-detect		unspecified (?)		100%	
	PC25_3	351 KB	Auto-detect		unspecified (?)		100%	
	PC40_1	351.9 KB	Auto-detect		unspecified (?)		100%	
	PC40_2	350.9 KB	Auto-detect		unspecified (?)		100%	
	PC40_3	351.3 KB	Auto-detect		unspecified (?)		100%	
	PC55_1	350.2 KB	Auto-detect		unspecified (?)		100%	

Type (set all): tabular
Reference (set all): unspecified (?)

 Choose local file
 Choose remote files
 Paste/Fetch data
Start
Pause
Reset
Close

Upload from Disk or

- Search tool **DESeq2**, and use the following parameters, this takes about 1 min:
 - ❖ "how": **Select datasets per level**
 - ❖ In "Factor":
 - "Specify a factor name, e.g. effects_drug_x or cancer_markers": **Diet**
 - In "1: Factor level":
 - "Specify a factor level": **PC70**
 - In "Count file(s)": Select all the P:C_0.70 count files (3 of them).
 - In "2: Factor level":
 - "Specify a factor level": **PC55**
 - In "Count file(s)": Select all the P:C_0.55 count files (3 of them).
 - ❖ param-repeat "Insert Factor level"
 - In "3: Factor level":
 - "Specify a factor level": **PC40**
 - In "Count file(s)": Select all the P:C_0.40 count files (3 of them).
 - ❖ param-repeat "Insert Factor level"
 - In "4: Factor level":
 - "Specify a factor level": **PC25**
 - In "Count file(s)": Select all the P:C_0.25 count files (3 of them).
 - ❖ "Files have header?": **Yes**
 - ❖ "Choice of Input data": **Count data (e.g. from HTSeq-count, featureCounts or StringTie)**
 - ❖ In "Output options":

- “Output selector”: Generate plots for visualizing the analysis results and Output normalised counts
5. Inspect the DESeq2 plots **PCA**, **answer Q1 in Data analysis assignment Part two**.
 6. Inspect the DESeq2 plots **Heatmap**, **answer Q2 in Data analysis assignment Part two**.
 7. Inspect the DESeq2 results, **answer Q3 in Data analysis assignment Part two**.
 8. **Answer Q4 in Data analysis assignment Part two**: Manually check for a gene by running tool **Filter data on any column using simple expressions** with below parameters, change the gene ID to your interest gene:

 **Filter data on any column using simple expressions** (Galaxy Version 1.1.1)   

Tool Parameters

Filter *

45: DESeq2 result file on data 44, data 43, and others 

accepted formats ▼

Dataset missing? See TIP below.

With following condition *

c1==The gene ID 

Double equal signs, ==, must be used as shown above. To filter for an arbitrary string, use the Select tool.

9. Open the **History Multiview**, and drag the below file from the RNAseq to RNAseq-2 **Drosophila_melanogaster.BDGP6.32.109_UCSC.gtf.gz**

Upload

Tools

Workflows

Workflow Invocations

Visualization

Histories

History Multiview

Datasets

History Multiview

Search datasets and collections in selected histories

RNAseq-2

SRA

last edited less than a minute ago

207 MB

37

15

24

64: **Drosophila_melanogaster.BDGP6.32.109_UC SC.gtf.gz**

63: **Filter on data 45**

62: **Normalized counts file on data 50, data 49, and others**

RNAseq

SRA

last edited 6 days ago

53.8 GB

27

571

756

a list with 12 **interval** datasets

705: **RNA STAR on collection 602: log**

a list with 12 **txt** datasets

704: **Drosophila_melanogaster.BDGP6.32.109_UC SC.gtf.gz**

10. Use tool **Annotate DEseq2/DEXSeq output tables**, with below parameters:

141

Annotate DESeq2/DEXSeq output tables



 Run Tool

Append annotation from GTF to differential expression tool outputs
(Galaxy Version 1.1.0)

Tool Parameters

Tabular output of DESeq2/edgeR/limma/DEXSeq *



60: DESeq2 result file on data 50, data 49, and others

accepted formats ▼

(-in)

Input file type *

DESeq2/edgeR/limma

(-m)

Reference annotation in GFF/GTF format *



64: Drosophila_melanogaster.BDGP6.32.109_UCSC.gtf.gz

accepted formats ▼

11. Add header to the output, first click on the header [link](#) to download the file to your computer, then upload to Galaxy by clicking on the **Upload** in the left panel and **Choose local file** and Change Type from “Auto-detect” to **tabular**. At last, press Start and Close the window.

12. Use tool **Concatenate datasets**, with parameters as follows:

Concatenate datasets tail-to-head (Galaxy Version 1.0.0)



 Run Tool

Tool Parameters

Concatenate Dataset *



66: Header

accepted formats ▼

Dataset

1: Dataset   

Select *



65: Annotate DESeq2/DEXSeq output tables on data 64 and data 60

accepted formats ▼

13. Use tool **Filter data on any column using simple expressions** to extract only gene with significant differential expression with parameters below:

 **Filter** data on any column using simple expressions (Galaxy Version 1.1.1)     **Run Tool**

Tool Parameters

Filter *

67: Concatenate datasets on data 65 and data 66 

accepted formats ▼

Dataset missing? See TIP below.

With following condition *

c7<0.05

Double equal signs, ==, must be used as shown above. To filter for an arbitrary string, use the Select tool.

Number of header lines to skip *

1

14. Use tool **sort** with below parameters and then rename file as **Genes with significant adj p-value** and **answer Q5 in Data analysis assignment Part two.**

 **Sort** data in ascending or descending order
(Galaxy Version 9.3+galaxy1)



 Run Tool

Tool Parameters

Sort Query *



69: Filter on data 65

accepted formats ▼

Number of header lines *

1

These will be ignored during sort.

Column selections

1: Column selections



on column - optional

Column: 3

in *

- ☐ Ascending order
☒ Descending order

15. Use tool **Filter** with below parameters and rename the file as **Genes with significant adj p-value & abs(log2(FC)) > 1** and **answer Q6 in Data analysis assignment Part two**

 **Filter** data on any column using simple expressions (Galaxy Version 1.1.1)



 Run Tool

Tool Parameters

Filter *



70: Genes with significant adj p-value

accepted formats ▼

Dataset missing? See TIP below.

With following condition *

abs(c3) > 1

Double equal signs, ==, must be used as shown above. To filter for an arbitrary string, use the Select tool.

Number of header lines to skip *

1

16. Before visualizing the normalized counts, we need to extract the normalized counts for the interesting genes. We join the normalized count table generated by DESeq2 with the table we just generated. We will then keep only the columns corresponding to the normalized counts. Use tool **Join two Datasets** with below parameters:

 **Join two Datasets** side by side on a specified field (Galaxy Version 2.1.3)    Run Tool

Tool Parameters

Join *

62: Normalized counts file on data 50, data 49, and others

accepted formats ▼

using column *

Column: 1

▼

with *

72: Genes with significant adj p-value & abs(log2(FC)) > 1

accepted formats ▼

and column *

Column: 1

▼

Keep lines of first input that do not join with second input *

No

▼

Keep lines of first input that are incomplete *

No

▼

Fill empty columns

No

▼

Keep the header lines *

Yes

▼

17. Use **Cut** to retain the gene ID with 12 samples count numbers. Then rename the file as **Normalized counts for the most differentially expressed genes**

Cut columns from a table (Galaxy Version 1.0.2)
 ☆ ▼ 📄
▶ Run Tool

Tool Parameters

Cut columns - optional

c1-c13

Delimited by *

Tab ▼

From *

73: Join two Datasets on data 72 and data 62 ▼

accepted formats ▼

18. Plot the heatmap of the normalized counts of these genes for the samples. Use `heatmap2` with below parameters. Inspect the plot and **Answer Q7 in Data analysis assignment Part two.**

- "Input should have column headers": `Normalized counts for the most differentially expressed genes`
- "Data transformation": `Log2(value+1) transform my data`
- "Enable data clustering": `Yes`
- "Labeling columns and rows": `Label columns and not rows`
- "Type of colormap to use": `Gradient with 2 colors`

19. Visualize the Z-score plot by using `heatmap2` with below parameters. Inspect the plot and **Answer Q8 in Data analysis assignment Part two.**

- "Input should have column headers": `Normalized counts for the most differentially expressed genes`
- "Data transformation": `plot the data as it is`
- "Compute z-scores prior to clustering": `Compute on rows`
- "Enable data clustering": `Yes`
- "Labeling columns and rows": `Label columns and not rows`
- "Type of colormap to use": `Gradient with 3 colors`

19. Prepare the first dataset for GO Functional enrichment analysis, use tools Compute, cut and change case with parameters as follows:

1. **Compute** (Galaxy version 2.0) on rows with the following parameters:
 - "Input file": the **DESeq2 result file** (output of **DESeq2**)
 - In "Expressions":
 - "Add expression": **bool(float(c7)<0.05)**
 - "Mode of the operation?": **Append**
 - Under "Error handling":
 - "Autodetect column types": **No**
 - "If an expression cannot be computed for a row": **Fill in a replacement value**
 - "Replacement value": **False**
2. **Cut** columns from a table with the following parameters:
 - "Cut columns": **c1,c8**
 - "Delimited by": **Tab**
 - "From": the output of the **Compute**
3. **Change Case** with
 - "From": the output of the previous **Cut**
 - "Change case of columns": **c1**
 - "Delimited by": **Tab**
 - "To": **Upper case**
4. Rename the output to **Gene IDs and differential expression**

20. Change the Gene ID of gene length file to upper case. Then rename it as **Gene IDs and length**.

Change Case of selected columns (Galaxy Version 1.0.0) ☆ 📁 ▶ Run Tool

Tool Parameters

From *

📄 📄 📁 ... 1: Gene length ▼

accepted formats ▼

Change case of columns - optional

c1

Delimited by *

Tab ▼

To *

Upper case ▼

21. Run the GO functional analysis with tool **goseq** with below parameters:

goseq (Galaxy version 1.50.0+galaxy0) with

- "Differentially expressed genes file": **Gene IDs and differential expression**
- "Gene lengths file": **Gene IDs and length**
- "Gene categories": **Get categories**
 - "Select a genome to use": **Fruit fly (dm6)**
 - "Select Gene ID format": **Ensembl Gene ID**
 - "Select one or more categories": **GO: Cellular Component**, **GO: Biological Process**, **GO: Molecular Function**
- In "Output Options"
 - "Output Top GO terms plot?": **Yes**
 - "Extract the DE genes for the categories (GO/KEGG terms)?": **Yes**

22. Inspect the top 10 over-represented GO terms plot, and **Answer Q9 in Data analysis assignment Part two.**

23. Inspect the Ranked categories list, and **Answer Q10 in Data analysis assignment Part two.**

Data analysis part 2/2 (5%)

1. Complete the data analysis part two and answer the questions.
2. Submit the Data analysis assignment part two to blackboard (due week 12)

Research essay part 3/3

1. Complete the result and discussion.
2. Incorporate the data analysis assignment part two result into your discussion.
3. 1st draft due on week 12

Lab 12 Protein Assay, Analysis Using ELISA, Western blot, Alpha-fold3

Na Xu, Hanqi Ye

Objective:

- To explore infectious diseases using a hands-on simulated enzyme-linked immunosorbent assay (ELISA) assay.
- To understand the basic principles of antibody-mediated immunity.
- To understand the steps used in an ELISA assay and how it is used as a diagnostic tool.
- To understand the disease-causing agent and transmission patterns of certain infectious diseases.
- To become familiar with the Methodology, Data Analysis, and Results sections of a research paper
- To continue the experiment for the group research project

Introduction

Pathogens: Pathogens are microscopic, disease-causing agents including viruses, bacteria, protozoa, molds, and other microorganisms. Pathogens invade the body and multiply, causing sickness or even death of the infected individual.

The Body's Defense System: The body has three lines of defense to fight infections. The first two modes are nonspecific: the body's physical barriers and the nonspecific immune system. The third defense mechanism is the specific immune system that is tailored to the type of invading pathogen. The external protective covering of the body includes skin, the mucous membranes that line the mouth, nostrils, and other potential gateways for pathogens. Secretions from the skin and mucous membranes such as sweat, tears, and saliva also prevent pathogens from entering the body. The human body has a series of nonspecific defenses that make up the **innate immune system**. These defenses are not directed against any one pathogen but instead, provide protection against all infections. If the exterior barriers are breached, nonspecific immune defenses are launched, including fever since many pathogens cannot function well at higher body temperatures. An **inflammatory response** begins when a pathogen stimulates an increase in blood flow to the infected area. Blood vessels in that area expand, and white blood cells exit the vessels to invade the infected tissue. These white blood cells, called *phagocytes*, *engulf* and destroy bacteria. The affected area often becomes red, swollen, and painful during an inflammatory response. Cells involved in the innate immune system encompass phagocytic cells, epithelial and endothelial cells, natural killer cells, innate lymphoid cells, and platelets. Phagocytic cells consist of granulocytes (i.e., neutrophils, eosinophils, basophils, and mast cells), monocytes/macrophages, and dendritic cells. These cells participate in phagocytosis and in the inflammatory process.

When pathogens bypass innate immune defenses, the **adaptive immune system** is activated. The innate immune system can distinguish between different classes of pathogens and recruit the most effective form of adaptive immune response to eliminate them. Unlike innate immune responses, adaptive responses are highly specific to the pathogen that induced them. They can also provide long-lasting protection. Specific immune responses are triggered by antigens. **Antigens** are usually found on the surface of pathogens and are unique to each pathogen.

Adaptive immune responses are carried out by white blood cells called lymphocytes. There are two broad classes of such responses—*antibody responses* and *cell-mediated immune responses*, and they are carried out by different classes of lymphocytes, called B cells and T cells, respectively (Figure 1). B cells are produced and mature in the bone marrow. In **antibody responses**, B cells are activated to secrete antibodies that circulate in the bloodstream, permeate other bodily fluids, and bind specifically to the foreign antigen (Figure 1). The binding of antibodies inactivates viruses and microbial toxins by blocking their ability to bind to receptors on host cells. Antibody binding also marks invading pathogens for destruction, mainly by facilitating phagocytosis by phagocytic cells of the innate immune system.

T cells are also produced in the bone marrow, but they mature in the thymus. In cell-mediated immune responses, activated T cells react directly against a foreign antigen that is presented to them on the surface of a host cell (Figure 1). The T cell, for example, might kill a virus-infected host cell that has viral antigens on its surface, thereby eliminating the infected cell before the virus can replicate. In other cases, the T cell produces signal molecules that activate macrophages to destroy the invading microbes that they have phagocytosed.

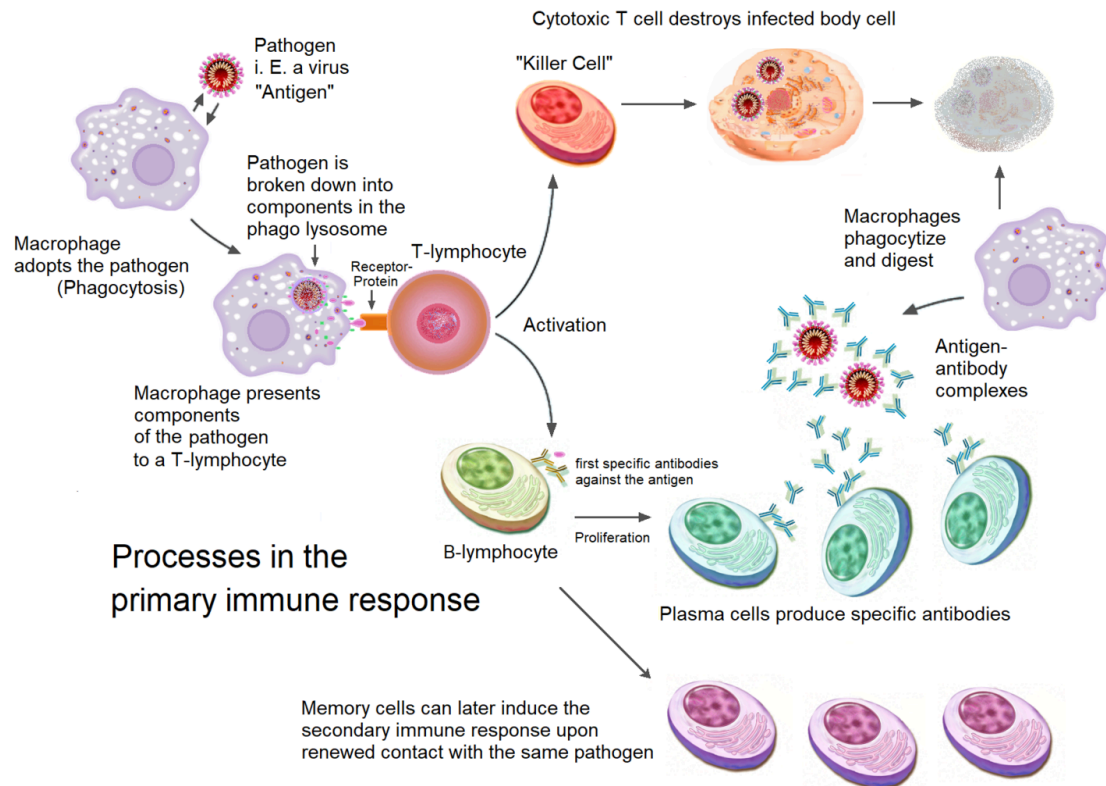


Figure 1: *Schematic overview of the primary adaptive immune response.* The diagram illustrates the sequence of immunological events following first exposure to a novel antigen: pathogen ingestion by antigen-presenting cells (e.g., macrophages), antigen processing and presentation on MHC-II, activation of naïve helper T-cells, subsequent activation of B-cells specific to the antigen, differentiation into antibody-producing plasma cells, and the eventual secretion of antibodies that target the pathogen. Naïve lymphocytes (B and T cells) must first be activated, which produces a lag before detectable antibodies appear in serum; antibody levels gradually rise, peak, and then decline as the pathogen is cleared. The figure emphasises that this "first encounter" response is slower and of lower magnitude compared to responses upon re-exposure.

Image credit: Sciencia58 et al., Primary immune response 1.png – CC BY-SA 4.0 (via Wikimedia Commons).

Western blot

The term “Western blot” is a pun on the Southern blot, a DNA detection technique named after its inventor, the English biologist Edwin Southern. Similarly, RNA detection is referred to as the Northern blot. The name “Western blot” was coined by W. Neal Burnette in 1981, although the technique itself was independently developed in 1979 by Jaime Renart, Jakob Reiser, and George Stark at Stanford University, and by Harry Towbin, Theophil Staehelin, and Julian Gordon at the Friedrich Miescher Institute in Basel, Switzerland. The Towbin group also used secondary antibodies for detection, making their method very similar to the one that is almost universally used today. Between 1979 and 2019, it has been mentioned in the titles, abstracts, and keywords of over 400,000 PubMed-listed publications and may still be the most widely used protein-analytical technique.

The Western blot, also known as the protein immunoblot, is a commonly employed method in molecular biology and immunogenetics. It is used to identify specific proteins in a tissue homogenate or extract sample. In addition to identifying proteins, this method is also used to visualize, differentiate, and quantify various proteins in a complex protein mixture.

The Western blot technique employs three key components to separate a specific protein from a complex: size separation, protein transfer to a solid medium, and target protein marking using primary and secondary antibodies for visualization. An antibody, either synthetic or derived from animals (referred to as the primary antibody), is produced to recognize and bind to a specific target protein. The electrophoresis membrane is then washed in a solution containing the primary antibody, and any excess antibody is washed away. A secondary antibody is then added, which recognizes and binds to the primary antibody. The secondary antibody can be visualized using various methods such as staining, immunofluorescence, and radioactivity, allowing for indirect detection of the specific target protein (Figure 2).

Related techniques include dot blot analysis, quantitative dot blot, immunohistochemistry, and immunocytochemistry, where antibodies are used to detect proteins in tissues and cells through immunostaining, and the enzyme-linked immunosorbent assay (ELISA).

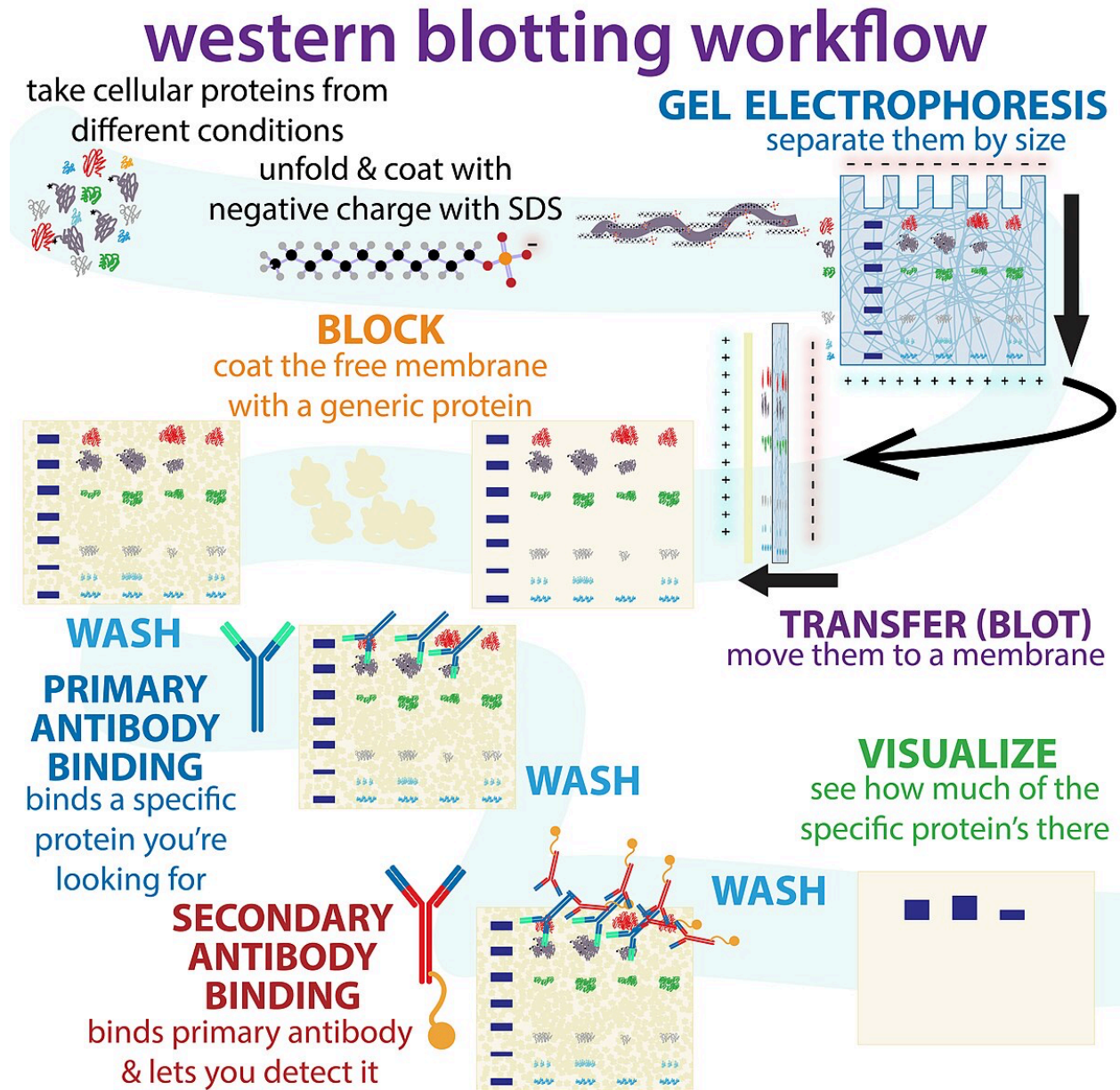


Figure 2. Workflow of a standard Western blot (immunoblot) procedure. The schematic illustrates the major steps typically involved in a Western blot experiment. Image credit: "Western blot workflow" by Biochemlife – CC BY-SA 4.0 (via Wikimedia Commons).

Sample preparation — proteins are extracted from cells or tissues and prepared with loading buffer (containing SDS and reducing agent), then denatured to ensure uniform charge-to-mass ratio.

SDS-PAGE (gel electrophoresis) — the denatured proteins are separated by molecular weight as they migrate through a polyacrylamide gel under an electric field.

Transfer (blotting) – the separated proteins are transferred from the gel onto a solid membrane (nitrocellulose or PVDF) by electroblotting, immobilizing them for detection.

Blocking & immunodetection – the membrane is first blocked to prevent nonspecific binding, then incubated with a primary antibody specific for the target protein, followed by a labeled secondary antibody (e.g., enzyme- or fluorophore-conjugated) to enable detection.

Signal development & visualization – after washes, the membrane is exposed to a detection reagent (e.g., chemiluminescent substrate), producing a signal where the target protein resides; this signal can then be imaged and analyzed to determine presence and relative abundance of the protein.

ELISA is an immunological assay commonly used to measure antibodies, antigens, proteins, and glycoproteins in biological samples. ELISA assays are based on the fact that antibodies are produced in response to infection and that these antibodies are designed to specifically target and bind tightly to specific antigens. ELISA is applied widely in diagnosis including diagnosis of HIV infection, pregnancy tests, and measurement of cytokines or soluble receptors in cell supernatants or serum. In a direct ELISA assay, the antigen is immobilized on the surface of the multi-well plate and detected with an antibody specific to the antigen. The antibody is conjugated with horseradish peroxidase (HRP) or other detection molecules. The direct ELISA procedure can be summarized into four steps: 1) plate coating with antigen, 2) blocking any remaining protein-binding sites, 3) antibody incubation, and 4) detection.

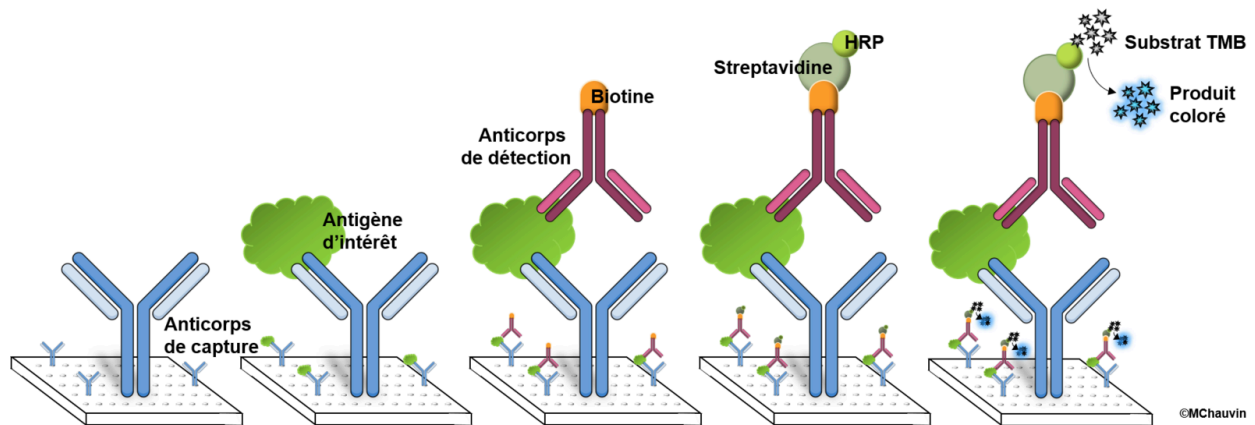


Figure 3: *Principle and workflow of a standard ELISA (Enzyme-Linked Immunosorbent Assay).* The figure illustrates the sequential steps used to detect and quantify target antigens or antibodies in a biological sample. First, the antigen or capture antibody is passively adsorbed onto the surface of a polystyrene microplate well. After a blocking step to minimize nonspecific interactions, the sample is introduced, allowing analyte molecules to bind specifically to the immobilized capture component. A detection antibody conjugated to an enzyme then binds to the target, forming a “sandwich” or direct complex depending on assay design. Upon addition of an appropriate chromogenic substrate, the enzyme catalyzes formation of a colored product,

whose optical density is measured and correlated to analyte concentration using a standard curve.(Adapted from "ELISA principle.tif," CC BY 4.0.)

Human immunodeficiency virus (HIV) is a virus that attacks the body's immune system, making a person more vulnerable to other infections and diseases. It is spread by contact with certain bodily fluids of an HIV-infected person, most commonly during unprotected sex, or through needle sharing. If HIV is not treated, it can lead to acquired immunodeficiency syndrome (AIDS). There is currently no effective cure. Once people acquire HIV, they remain infected for life. However, with proper medical care, HIV infections can be controlled.

Lyme disease is caused by the bacterium *Borrelia burgdorferi* and rarely, *Borrelia mayonii*. It is transmitted to humans through the bite of infected black-legged ticks. Typical symptoms include fever, headache, fatigue, and a characteristic skin rash called erythema migrans.

Avian influenza (bird flu) is caused by infection with avian (bird) influenza (flu) Type A viruses. These viruses naturally spread among wild aquatic birds worldwide and can infect domestic poultry and other bird and animal species. Bird flu viruses do not normally infect humans. However, sporadic human infections with bird flu viruses have occurred. Symptoms begin within two to eight days and can be similar to those of the common flu. Cough, fever, sore throat, muscle aches, headache, and shortness of breath may occur. The disease can have high mortality rates in humans. Certain antiviral drugs, if administered within two days of symptoms, may mitigate the infection.

West Nile Virus (WNV) is the leading cause of mosquito-borne illnesses in the continental United States. It is most spread to people by the bite of an infected mosquito. There are no vaccines to prevent or medications to treat WNV in people. Fortunately, most people infected with WNV do not develop symptoms. Approximately 1 in 5 people infected develop a fever and other symptoms such as headache, body aches, joint pains, vomiting, diarrhea, or rash. Most people with febrile illness due to WNV recover completely, although fatigue and weakness can last for weeks or months. Approximately 1 out of 150 infected people develop a serious, occasionally fatal, illness.

Lab Exercise I -ELISA Simulation

Materials

Positive control

Negative control

Patient A–F samples (mimic blood serum)

Simulated antigen, antibody, and chromogen (a colorless chemical compound that can be converted by chemical reaction into a colored compound)

16 microtiter plates

180 plastic pipettes

Methods

Refer to the provided students' guide.

Expected Results

Record data in the following table and handouts:

Sample	Color	Test Result
Positive control		
Negative control		
Patient A		
Patient B		
Patient C		
Patient D		
Patient E		
Patient F		

* Submitting the handouts and data records will count for 5 points for Assignment 10. The students' lab notes for results will be evaluated by the instructor.

Lab Exercise II - Western blot simulation

Material:

Protein gel Simulation: [SDS-PAGE gel](#)

Western blot Simulation: [Western Blot: Testing Protein Expression Levels - LabXchange](#)

Methods:

1. Click on the SDS-page gel simulation link, choose the level one.
2. Follow the instruction and complete the simulation,answer the questions.
3. Click on the Western Blot simulation link, choose the level one.
4. Follow the instruction and complete the simulation,answer the questions.

Introduction-3D protein structure modeling

Recently, Google Deepmind announced a significant update: their protein analysis AI tool, AlphaFold, has been upgraded to its third generation. AlphaFold 3 has significantly improved analytical accuracy compared to its predecessor, and it can predict the structure of protein complexes of “almost all molecular types” in addition to the protein itself. News reports suggest that this tool will bring about “dramatic” changes in the field of biomedical research.

Why is knowing protein structure important? If life science were a book, there would be a message from Feynman on the flyleaf: All the behavior of the living world can be understood as the quivering and twisting of atoms. This implies that to truly understand proteins, you have to delve from three dimensions down to atomic resolution, which is the scale at which proteins function. Structural biology was born out of such a concept.

Before AI got involved, How was protein structure studied? The primary methods to obtain protein structure fall into two categories: experimental determination and computational prediction. Before AI entered the scene, computers were already being used to analyze proteins. However, at that time, prediction accuracy was low, and structural details were often overlooked, but these predicted structures were sometimes used as templates for crystal structure analysis. The actual confirmation of protein structure relies on experimental determination, including protein crystal X-ray diffraction, cryo-electron microscopy, nuclear magnetic resonance, and so on. Obtaining high-resolution protein structures experimentally is slow and challenging: in the crystal age, figuring out a single crystal structure often consumed a PhD student’s entire career.

The structures are then uploaded to the Protein Database (<https://www.rcsb.org/>), to which about 200,000 structures have been submitted over the past 50 years. These past data are important learning materials when training AlphaFold, and it can be said that the birth of AlphaFold is a reward for the hard work of several generations of structural biologists.

AlphaFold was born, How much has it changed the study? The milestone event for protein structure prediction was CASP14 in 2020, when AlphaFold2 predicted proteins with near-atomic resolution. What’s more, these structures are “right” in detail - that is, in research situations such as drug screening, AI’s predictions become useful.

In terms of speed, AlphaFold2 is also far beyond experimental means: a structure that would take a PhD student months or even years to achieve can be approximated in half an hour by a GPU, which was previously unimaginable. Thanks to this efficiency, AlphaFold was able to predict 600 to 700 million protein structures in a very short time. Before that, through the efforts of generations of structural biologists, humans had accumulated only 200,000 protein structures in half a century.

What has AlphaFold improved from Generation 1 to Generation 3? Compared with past predictive models, the emergence of Generation 1 AlphaFold is like changing from “pipeline” to “all-in-one”. The previous structure prediction method is homologous modeling, which is more

like a pipeline, composed of many modules. After module A is processed, it is handed to B, and after module B is processed, it is handed to C. As a result, errors generated in each step will accumulate and magnify each other, resulting in a large deviation of the final model.

AlphaFold is a neural network method, which is “end-to-end”, the original pipeline is replaced by an integrated molding process, and the problem of error accumulation is perfectly solved. The original AlphaFold was a simple model, but it was already ahead of the curve at the time. Since then, AlphaFold 2 has been an epoch-making project - it has achieved unprecedented accuracy. AlphaFold uses the classic convolutional neural network architecture, while Alphafold2 brings in the more capable Transformer architecture. With the blessing of a deeper and stronger network, AlphaFold2 allows humans to predict protein structures at atomic resolution for the first time.

AlphaFold3's advance is that it can not only process proteins, but also feed nucleic acids, small molecules, metal ions and other substances at the same time, and predict how they will bind to the protein. Having this ability is extremely important for structural prediction, because proteins cannot function in isolation; they must communicate with other substances. This process of communicating with other substances is often what scientists are most concerned about and is the most important process in drug design.

AlphaFold made the following changes in our research:

1. Improve efficiency: For structural biology researchers, before AlphaFold, a lot of effort was devoted to model building. Especially in the crystal age, due to the limitations of the method itself, not all diffraction data can be obtained protein structure, often requiring a relatively correct template. It can be said that AlphaFold completely solves this problem, and since its birth, crystal structures that cannot be solved due to the lack of templates have become very rare. And it frees researchers from the heavy work of modeling, and even reduces the occurrence of modeling errors from the source.
2. Workflow simplification: For structural biologists and drug design engineers, the starting point for work is often a high-resolution structure. In the past, they spent a lot of energy and effort to get the structure, and then to design and verify the experiment. But now this most time-consuming process can be omitted or bypassed to design experiments, explain mechanisms and develop drugs from predicted protein structures. The advent of Alphfold has fundamentally changed and simplified the flow of work.

If we consider life as an assembly of Lego blocks, then large molecules such as proteins and nucleic acids are akin to individual Lego pieces. For the past 50 years, structural biologists have primarily focused on understanding the shape of each Lego piece. It's not that they aren't interested in how the pieces fit together; they simply lacked the capability - without knowing

what the parts look like, it's impossible to comprehend the shape of the whole assembly. For a long time, understanding the shape of each part was a dream for structural biologists. AlphaFold has turned this dream into reality. Today's AlphaFold doesn't just reveal the shape of each part, it also begins to show us how the parts interconnect. It offers researchers the possibility that we might be able to peer into the secrets of everything from individual parts to whole assemblies.

AI technologies, particularly in drug design, have made significant strides. They show promising results in early drug development stages. However, drug development is a long process, and the initial steps involving protein research and drug molecular screening with AI will require substantial time for clinical validation. As a result, it may take several years before we see AI-developed drugs.

AlphaFold3, an advanced tool, has demonstrated accurate prediction of small molecular-protein complexes, a task that has always been challenging yet crucial. It has performed well in the few completed cases, with results surprisingly close to experimental outcomes. The tool has also improved the prediction accuracy of antibody-protein complexes, which is vital for the development of biomacromolecular drugs. However, there are still significant differences in details that need further improvement. AlphaFold3 has also shown accurate predictions in protein-DNA complexes based on test results.

Lab Exercise III - 3D protein structure modeling using Alpha fold 3

Material:

Alpha fold server: <https://golgi.sandbox.google.com/>

NCBI: <https://www.ncbi.nlm.nih.gov/>

Ebony protein structure

Paper: https://www.pnas.org/doi/10.1073/pnas.1811194116?url_ver=Z39.88-2003&rfr_id=ori%3Arid%3Acrossref.org&rfr_dat=cr_pub++0pubmed

Method:

1. open NCBI and search for a protein sequence of your interest from Vg, Eb, Cn, Dp. Click on the FASTA link to copy the sequence.

NIH National Library of Medicine
National Center for Biotechnology Information

Protein
Advanced

GenPept Send to:

ebony protein isoform X1 [Bombyx mori]
NCBI Reference Sequence: XP_037876452.1
[Identical Proteins](#) [FASTA](#) [Graphics](#)

[Analyze this sequence](#)

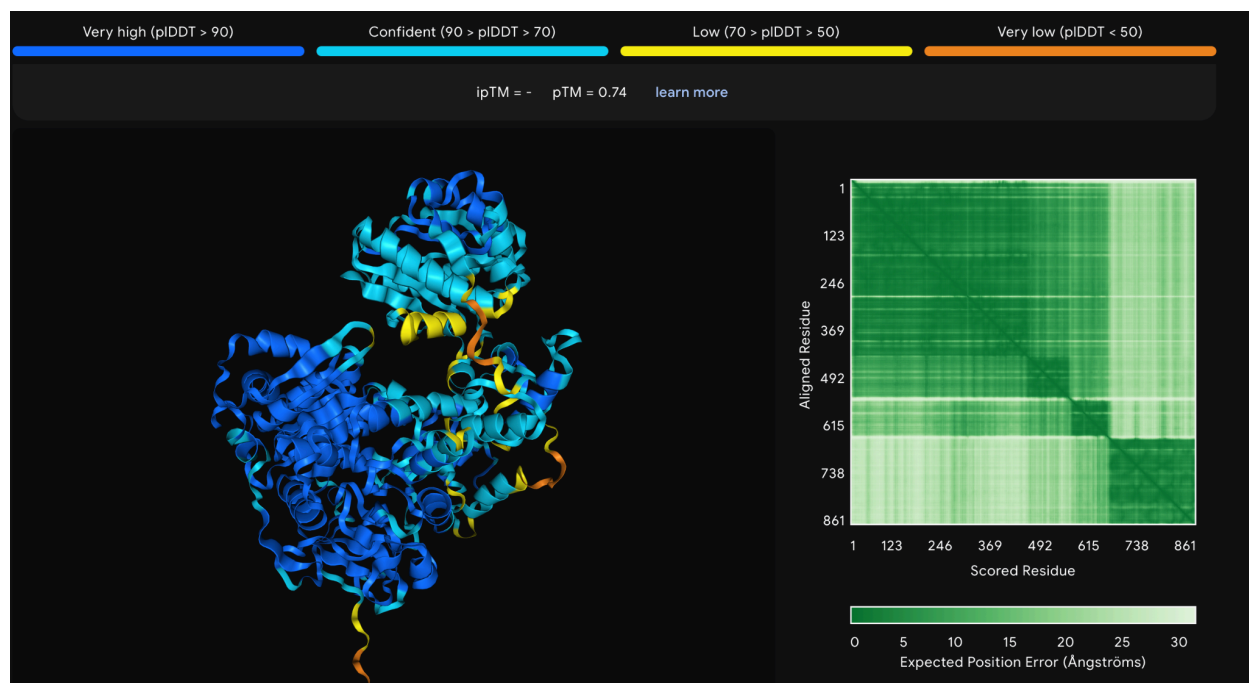
2. Copy the protein sequence and paste it to the alpha fold server in the sequence input.

AlphaFold Server BETA Server About FAQs

Remaining jobs: 9

AlphaFold Server allows you to model a structure consisting of many biological molecules [Learn more](#)

3. Generate the 3D model.



4. You can generate different protein isoforms 3D structure for comparison if a gene has multiple alternative splicing derived mRNA.

References

Alberts B, Johnson A, Lewis J, et al. Molecular Biology of the Cell. 4th edition. New York: Garland Science; 2002. Chapter 24, The Adaptive Immune System. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK21070/>

Alberts B, Johnson A, Lewis J, et al. Molecular Biology of the Cell. 4th edition. New York: Garland Science; 2002. Drosophila and the Molecular Genetics of Pattern Formation: Genesis of the Body Plan. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK26906/>

Alternative splicing - Wikipedia. (2022, June 23). Alternative Splicing - Wikipedia; [en.wikipedia.org](https://en.wikipedia.org/wiki/Alternative_splicing). https://en.wikipedia.org/wiki/Alternative_splicing

Celniker SE, Rubin GM. The Drosophila melanogaster genome. Annu Rev Genomics Hum Genet. 2003;4:89-117. doi: 10.1146/annurev.genom.4.070802.110323. PMID: 14527298.

Gene structure - Wikipedia. (2022, June 2). Gene Structure - Wikipedia; [en.wikipedia.org](https://en.wikipedia.org/wiki/Gene_structure). https://en.wikipedia.org/wiki/Gene_structure

Edward W. Green, Susanna Campesan, Carlo Breda, Korrapati V. Sathyasaikumar, Paul J. Muchowski, Robert Schwarcz, Charalambos P. Kyriacou & Flaviano Giorgini (2012) Drosophila eye color mutants as therapeutic tools for Huntington disease, Fly, 6:2, 117-120, DOI: 10.4161/fly.19999

Gupta RM, Musunuru K. Expanding the genetic editing tool kit: ZFNs, TALENs, and CRISPR-Cas9. J Clin Invest. 2014 Oct;124(10):4154-61. doi: 10.1172/JCI72992. Epub 2014 Oct 1. PMID: 25271723; PMCID: PMC4191047.

Hales K, KoreyC, Larracuenta A, Roberts D, Genetics on the Fly: A Primer on the *Drosophila* Model System, *Genetics*, Volume 201, Issue 3, 1 November 2015, Pages 815–842, <https://doi.org/10.1534/genetics.115.183392>

Hsu PD, Lander ES, Zhang F. Development and applications of CRISPR-Cas9 for genome engineering. Cell. 2014 Jun 5;157(6):1262-1278. doi: 10.1016/j.cell.2014.05.010. PMID: 24906146; PMCID: PMC4343198.

Steneck N and Zinn D. ORI introduction to the responsible conduct of research. 2006. <https://ori.hhs.gov/education/products/RCRintro/index.html>

Susan E. Celniker and Gerald M. Rubin. The Drosophila Melanogaster Genome. Annual Review of Genomics and Human Genetics 2003 4:1, 89-11

Tebas P, Stein D, Tang WW, Frank I, Wang SQ, Lee G, Spratt SK, Surosky RT, Giedlin MA, Nichol G, Holmes MC, Gregory PD, Ando DG, Kalos M, Collman RG, Binder-Scholl G, Plesa G, Hwang WT, Levine BL, June CH. Gene editing of CCR5 in autologous CD4 T cells of persons infected with HIV. N Engl J Med. 2014 Mar 6;370(10):901-10. DOI: 10.1056/NEJMoa1300662. PMID: 24597865; PMCID: PMC4084652.

Wilkin MB, Becker MN, Mulvey D, Phan I, Chao A, Cooper K, Chung HJ, Campbell ID, Baron M, MacIntyre R. *Drosophila dumpy* is a gigantic extracellular protein required to maintain tension at epidermal-cuticle attachment sites. *Curr Biol*. 2000 May 18;10(10):559-67. doi: 10.1016/s0960-9822(00)00482-6. PMID: 10837220.

Dunst S, Tomancak P. Imaging Flies by Fluorescence Microscopy: Principles, Technologies, and Applications. *Genetics*. 2019 Jan;211(1):15-34. doi: 10.1534/genetics.118.300227. PMID: 30626639; PMCID: PMC6325693.

Young, Andrew & Gillung, Jessica. (2019). Phylogenomics – principles, opportunities and pitfalls of big-data phylogenetics. *Systematic Entomology*. 45. 10.1111/syen.12406.

Batut, Bérénice, et al. "Community-driven data analysis training for biology." *Cell systems* 6.6 (2018): 752-758

Hiltemann, Saskia, Rasche, Helena et al. "Galaxy Training: A Powerful Framework for Teaching!." *PLOS Computational Biology* (2023) 19(1): e1010752

WC Aw, JWO Ballard. "Near-Infrared Spectroscopy for metabolite quantification and species identification" *Ecology and Evolution*(2018) doi: 10.1002/ece3.4847

T. Izoré, J. Tailhades, M.H. Hansen, J.A. Kaczmarek, C.J. Jackson, & M.J. Cryle, *Drosophila melanogaster* nonribosomal peptide synthetase *Ebony* encodes an atypical condensation domain, *Proc. Natl. Acad. Sci. U.S.A.* 116 (8) 2913-2918, <https://doi.org/10.1073/pnas.1811194116> (2019).