Lab Book

MDSC 301 CURE - Introduction to Bioinformatics

Winter 2024

Instructor: Dr. Tatiana Maroilley

Research Coach: Suzanne Ferris

Teaching Assistant: Shreya Tomar

Graduate Assistant: Rumika Mascarenhas

Introduction

Welcome to MDSC 301!

Here is your lab book that will guide you all along the semester. It contains mandatory readings for each week, information regarding assignments, tutorials, templates (abstract, proposal, manuscript) and protocols.

Important information

The laboratory class will meet in O1501 every Mondays and Wednesdays from 10:30 to 11:45am. The teaching team is composed of Dr. Tatiana Maroilley (tatiana.maroilley@ucalgary.ca), Rumika Mascarenhas (rumika.mascarenhas@ucalgary.ca), Suzanne Ferris (suzanne.ferris@ucalgary.ca), Shreya Tomar (shreya.tomar@ucalgary.ca), Dr. Maja Tarailo-Graovac (maja.tarailograovac@ucalgary.ca) and Dr. David Anderson (david.anderson1@ucalgary.ca).

Please note that all course communications must occur through your @ucalgary email.

Office hours:

Teaching Assistant Shreya Tomar: Wednesday 3-5pm

https://ucalgary.zoom.us/j/94607321572

Research Coach Suzanne Ferris: Tuesday 3-5pm

https://ucalgary.zoom.us/j/94486237619

Overview of Course

This course provides students with hands-on experience in experimental principles of Bioinformatics and multi-omics analyses. This course integrates Bioinformatics, genetics, genomics, transcriptomics, and biology and provides a foundation for further study in applied Bioinformatics and genomics.

Course Description

Develop independent investigative ability, technical skill, proficiency in scientific writing and presentation, and an in-depth understanding of research and the scientific approach through semester-long projects in genetics and molecular biology.

During the first part of the course, you will learn basic techniques of molecular genetics and bioinformatics and gain hands-on experience of genetics using *Caenorhabditis elegans* as a model organism. During the second part of the course, you will engage in a project that allows you to design, execute and evaluate experiments.

The course is divided into two parts. The first part (before Reading week) is composed of active learning activities, lectures and mandatory readings to introduce the principles needed in Biology and Bioinformatics. In addition, one designed genomics experiment will be conducted to introduce you to Bioinformatics principles, hypothesis testing, data analysis and reinforces your written and oral communications skills. You will present your work in a research paper following the conventions of scientific papers in biology and bioinformatics.

During the second part of the course (after Reading week), you will use the knowledge acquired in the first part of the course to develop your own novel independent research project using original transcriptomic datasets of *C. elegans* balancers.

Finally, you will submit an abstract and present their work in a conference with poster presentation.

An encouraging note

You might be thinking: "*80 confusing pages about a complex project in bioinformatics. Am I ever going to read the whole lab manual and understand it?"* Don't worry; we will explain the project to you as many times as needed. However, you should read the project descriptions several times during the semester. You will understand more each time and it will be (hopefully) rewarding to you to feel that you understand and participate in a complex research project such as this one.

I strongly recommend you to talk to your instructor and TAs if you have any problems, questions or curiosity (part of your grade, 10% Engagement item). We are here to help you and

to motivate you about research. We enjoy doing it! Think about us as a collaborative team working for your education. As long as you are engaged with the class, we will be there to help.

General Course Objectives

This course is a good opportunity for you to appreciate the challenges and rewards of research. At the completion of this course, we hope that you will be able to:

- ➔ Understand the Bioinformatics approach to analyze omics data and the use of sequencing technologies to explore the complexity of Biology
- ➔ Understand key features of analysis design
- ➔ Understand the importance of apprehending the biological questions, the origin of the data and the biases related to each technology and analysis
- ➔ Understand the organization of the scientific literature and scientific databases
- ➔ Communicate science and give presentations on scientific subjects

Student Learning Goals

At the end of this course, you should be able to:

- ➔ Interpret primary research literature
- ➔ Design and carry out novel Bioinformatics analysis investigating genomics or transcriptomics data
- ➔ Describe and explain concepts used in sequencing approaches
- ➔ Search for, collect, and evaluate scientific research articles relating to topics of interest.
- ➔ Statistically analyze and evaluate novel research findings
- ➔ Describe results in written papers and in oral presentations following established conventions for the field of biology.

General considerations

To make our time together as effective as possible, it is important that the lecture learning environment is one of mutual respect. I will do whatever I can to create and maintain that environment; my expectations of student conduct are outlined below:

- ➔ Everyone has the right to learn as well as the responsibility to not deprive others of their right to learn.
- ➔ Actions such as talking during instruction/lecturing, or using laptops and other electronic devices for non-class activities can be very distracting and affect others' learning. Please monitor your own behavior during classes and restrict your use of laptops and

other electronic devices to only those activities directly related to class to ensure that you do not distract others.

- ➔ Please arrive at class on time. Late arrivals and early departures can be disruptive and can result in you missing important information. I understand that there are special circumstances when you may have to arrive late or leave early; please make your arrival/departure as unobtrusive as possible and be sure to let your teammates know about your situation in advance of class.
- ➔ Please let me know right away if you are dealing with a problem or situation that is preventing you from performing at the level you want to be at in this class.
- ➔ Please treat your classmates, peer mentors, and me with respect. There may be times when you are frustrated with something that is going on in the course and find it difficult to be patient. However, to maintain a respectful and constructive environment in this class, I ask that you are **respectful to others in your words and actions.**

What you can expect from the teaching team:

- ➔ We will treat all students with respect and I do my best to make clear my expectations about how to succeed in this class.
- ➔ We will do our best to help your learning by designing clear assignments and assessments that provide you with timely feedback.
- ➔ I will start and end classes on time.
- ➔ We will be available outside of class time through office hours, appointments or email should you want to review concepts that are not clear, discuss study strategies, learn more about any topic or discuss concerns about any aspect of the course. Please note that we will aim to reply to emails within 24h, except on weekends.

Experimental Material and Projects

In both projects that you will work on, you will use the worm *C. elegans*. In the first part of the semester you will analyze genomes of *C. elegans* to detect structural variants and/or complex genomic rearrangements. During the second part of the course, you will develop your own research project based on transcriptomics data of *C. elegans*. You will design, execute and evaluate your own analyses.

General Organization of the Dry Lab

You will mostly work in teams. During the first part of the semester, the instructor, TAs and RC will guide you through the techniques that you will need to learn as you carry out the genomics project. You will then work more independently to complete the transcriptomics project (an analysis of your own design) during the remainder of the scheduled dry lab periods. Most Monday periods will begin with a quiz to assess knowledge acquired during the previous sessions. The procedures you are to follow in the experiments are described in detail in this lab book.

Grades

A student's final grade for the course is the sum of the separate assignments. It is not necessary to pass each assignment separately to pass the course.

The final grade for the course will be determined as follows:

Team work - Submitted to Peer evaluation

Proposal	10%	Proposal written with a template for genomics project
Manuscript	20%	Manuscript written with a template regarding genomics project
Poster	20%	Design of a scientific poster regarding the transcriptomics project

Individual work

Reflection Assignment 1	10%	Series of questions to help the students to reflect on their perception of research and their visit to the research lab
Reflection Assignment 2	10%	Series of questions to help the students to reflect on their experience of research in Bioinformatics in the class
Quizzes	10%	Quizzes through D2L assessing the necessary background knowledge for the student to be successful in this course
Engagement	10%	Assessed by the teaching team based on attendance and engagement in Active Learning Activities and Analyses
Abstract	10%	Abstract (conference-submission style) written with a template regarding the transcriptomics projects

Team work

In this class, we will be using a Team Based Learning (TBL) approach - the best way to learn Bioinformatics. In this process, you will spend many classes working in teams applying what you've learned. Teams in TBL are different from the kind of group work you may have done in other classes: the instructor forms the teams (as described below) which work together throughout the term to complete most course assignments; team members also evaluate each other's contributions to the group throughout the term.

We will be forming teams in the third week of class. Research shows that diverse teams function the best and produce the best outcomes. To help with this, we will be using a survey (https://survey.ucalgary.ca/jfe/form/SV_1TcBecwJi52IprE) to divide you into teams of 4-5 students based on previous courses you've taken, your program, work experience, and other factors that will help us form successful teams. These may feel like big teams at first, but research shows that teams of 4-7 individuals work best. As the term progresses, I am sure you will appreciate having the diversity of ideas and perspectives that come with a team of this size. Additionally, I will be putting measures in place (peer evaluation and instructions regarding the distribution of the workload) to ensure there is individual accountability to the team.

If you are having issues in your team, please don't hesitate to come and talk to the Teaching Team. Little problems can turn into big problems if not addressed. We are happy to facilitate a discussion with your team to help resolve issues.

Stay engaged – do not let others do all the work. We only learn Bioinformatics by doing it. Share the work. Even if someone has done the analysis before you did, still do it to compare results. Mistakes can be made at every step – it would be hazardous to rely on only one try. And there are always different ways to do the same thing. It is always interesting to compare different methods.

Peer Evaluation

Some notes regarding peer evaluation on Team assignments:

- ➔ Each student will be provided a Peer evaluation Rubric to provide an evaluation for each group member relating to Group assignments.
- ➔ Peer evaluation is part of the course grade.
- ➔ Students who do not submit a peer evaluation will receive a grade of zero on this component, regardless of the viewpoints of their fellow group members.
- ➔ Students will receive only an average group rating (not including self-rating) and will not be aware how each team member rated them.
- ➔ If a student's peer evaluation mark is > 10 (out of 20), they will receive 100% of the group assignment mark
- ➔ If a student's peer evaluation mark is >5 and ≤10, they will receive 50% of the group assignment mark
- ➔ If a student's peer evaluation mark is >2 and ≤5, they will receive 25% of the group assignment mark
- ➔ If a student's peer evaluation mark is between 0 and ≤2, they will receive 0% of the group assignment mark
- ➔ **A peer evaluation score of ≤5 will NOT be accepted unless concerns are expressed by the group to Dr. Maroilley no later than February 18th.**

Keys to success in Dry lab

- ➔ Follow directions
- ➔ Read the recommended readings
- ➔ Install the mandatory tools and created the mandatory accounts before coming to the lab
- ➔ Be organized: prepare for the lab by reading over the lab manual, thinking about what you did previously, so that you know what you're doing before you come to the lab.
- ➔ Keep good notes of what you do in the lab: every time you come to the lab, write down the key details of what you have done, save your scripts/protocols, how you manage to install the tools, save the useful bioinformatics commands
- ➔ Be meticulous: perform an analysis carefully and without rushing so it will be done correctly!
- ➔ Plan to complete what you need to for the week
- ➔ Consult with the instructor or TA with any questions

Preparation for Bioinformatic Lab sessions

Do the mandatory preparation steps (or at least try), so you can focus on analyses in class, and not in creating accounts on installing tools. If you cannot do it on your own before the class, reach out to the Instructor to get help BEFORE the class.

Attendance

Attendance in MDSC301 is primordial. Any absence should be justified and announced when possible. Groups that won't be attending the course will be disadvantaged, as the instructor, TA and RC will prioritize groups coming in person. The team nature of this class requires you to be in class and to do your part as a member of your Team. Quizzes missed without a valid excuse (medical or family emergency) will be awarded a mark of zero. Missed quizzes may not be written at a later time. Attendance at all labs is required for this course.

Of note, if you have a class right before or right after MDSC301, and that class is held online, please let me know, and I could get the Bioinformatics classroom available so you could attend the online class in O1501, and still come in person for MDSC301.

Engagement

Engagement in this class is the key to success. It is also part of your grade (10%). Engagement will be evaluated independently by each member of the teaching team, and your grade will be an average of each evaluation. Evaluation will take into account: attendance (5 pts), participation in class discussions (5 pts), involvement in the project (based on Reports) (5 pts).

Readings

Often in this class, you will be assigned readings that you are expected to complete before the start of the module. You will find in the lab book a Reading Summary of the topic you need to cover, with additional references. The summary is the bare minimum for you to go through. It is strongly recommended to have a look at the additional references suggested (articles, reviews, videos…). Those readings are key to success in this course.

Quizzes

At the beginning of several sessions, you will individually take a short (~10-15 questions) multiple choice test to see how well you've understood the concepts in the assigned reading and/or the learning sessions. Quizzes will make up 10% of your grade. Quizzes missed without a valid reason will be awarded a mark of zero. Missed quizzes may not be written at a later time. **The quiz with the lowest grade won't be taken into account.**

#	Date	Subject
Quiz 1	Jan. 15, 2024	*C. elegans* as model organism
Quiz 2	Jan. 22, 2024	Genomics
Quiz 3	Jan. 29, 2024	Bioinformatics genome analysis
Quiz 4	Feb. 05, 2024	Galaxy and HPC
Quiz 5	Feb 12, 2024	Validation
Quiz 6	Feb 26, 2024	RNA-Seq
Quiz 7	Mar. 04, 2024	R

Schedule Overview

- P = Peer evaluation
- Q = Quiz (about 10 min)
- R = Report (about 5 one sentence-length-answer questions)
- Grey: Individual assignment
- White: Group assignment

Date	Module / Topics	Instructor	Readings	Assignments & Due Dates
Jan. 08, 2024	Introduction to *C. elegans*	T. Maroilley		Team Building Survey
Jan. 10, 2024		T. Maroilley	Lab Book Pg 16 -18	
Jan. 15, 2024	Genome Sequencing, Structural Variants	T. Maroilley	Refer to Lab Book	Q1 (C. elegans)
Jan. 17, 2024		T. Maroilley		Reflection Assignment due **January 19, 2024 @ 11:59 PM**
Jan. 22, 2024	Designing genome analysis for SV detection	T. Maroilley	PMIDs: 34521941, 36617680	Q2 (Genomics)
Jan. 24, 2024		T. Maroilley		Proposal + P1 due **January 26, 2024 @ 11:59 PM**
Jan. 29, 2024	Genome analysis on Galaxy	T. Maroilley	Refer to Lab Book	Q3 (Genome Analysis)
Jan. 31, 2024		T. Maroilley	Refer to Lab Book	R1 by **Feb. 02, 2024 @ 11.59 PM**
Feb. 05, 2024	Introduction to IGV and PCR	R. Mascarenhas	Refer to Lab Book	Q4 (Galaxy)
Feb. 07, 2024		S. Ferris		R2 by **Feb. 09, 2024 @ 11.59 PM**
Feb. 12, 2024	Introduction to RNA-Seq	T. Maroilley	Refer to Lab Book	Q5 (Validation)
Feb. 14, 2024		T. Maroilley		Manuscripts for feedback on **Feb. 14, 2024 @ 11:59 PM**
Feb. 18,	Winter term			

Date	Module / Topics	Instructor	Readings	Assignments & Due Dates
2024	break			
Feb. 24, 2024				
Feb. 26, 2024	Introduction to R	S. Tomar	Refer to Lab Book	Q6 (RNA-Seq)
Feb. 28, 2024		T. Maroilley		Manuscript + P2 due on **Feb. 28, 2024 @ 11:59 PM**
Mar. 04, 2024	Project Design	T. Maroilley		Q7 (R)
Mar. 06, 2024		T. Maroilley		
Mar. 11, 2024	Project	T. Maroilley		Proposal due **Mar. 11, 2024 @ 11:59 PM**
Mar. 13, 2024		T. Maroilley		R3 by **Mar. 15, 2024 @ 11.59 PM**
Mar. 18, 2024	Project	T. Maroilley		
Mar. 20, 2024		T. Maroilley		Abstracts due on **March 22, 2024 @ 11:59 PM**
Mar. 25, 2024	Poster design	T. Maroilley		
Mar. 27, 2024	Poster design	T. Maroilley		Poster + P3 due on **Mar. 27, 2024 @ noon**
Apr. 01, 2024	Easter Monday – no classes			
Apr. 03, 2024	Conference preparation	T. Maroilley		
Apr. 08, 2024	Conference with poster presentation	T. Maroilley		Reflection assignment due on **April 08, 2024 @ 11:59 PM**

Description of the Multi-Omics Project

In the Multi-Omics Project, you will be using *Caenorhabditis elegans* as a genetic model to study structural variants. The first analysis is based on published work by the Tarailo-Graovac Lab at the University of Calgary (https://pubmed.ncbi.nlm.nih.gov/34521941 and https://pubmed.ncbi.nlm.nih.gov/36617680/). You will be analyzing whole genome sequencing data of one balancer strain of your choice among: SS746, KR2839, MT690, CZ1072, and RW6002. To perform your analysis, you will have tutorials (see Lab Book) and you will be guided by the instructors while doing them. The Genomic Analysis will guide you in the elaboration of genome analysis in Galaxy, a platform that allows user-friendly bioinformatics analysis, freely available online and on a HPC system. In the Transcriptomic Analysis, you will have the freedom to do a research project of your own with brand new RNA-Seq data – made available by Dr. Tarailo-Graovac, not yet analyzed or published. Everything you would have learned prior to that analysis will be key to accomplish that project. Of course, the instructor, TAs and RC will be there to help you and advise you during your independent work. The question that Analysis 2 has to answer is "**What effect structural variants have on gene expression?**". If the question is defined in this project, you will have to come up with the workflow of analyses to answer that question. It might be difficult for you right now to come up with an analysis design, but I assure you that by the middle of the semester you will have a better idea about how to design a bioinformatics analysis of sequencing data. You will work on the same strain you picked for the genome analysis and explore the impact of at least on the event you observed in the genome on the gene expression.

These articles (https://pubmed.ncbi.nlm.nih.gov/34521941 and https://pubmed.ncbi.nlm.nih.gov/36617680/) are the result of the project in which the MTG Lab is promoting short-read whole genome sequencing as an efficient method to detect structural variants (SVs) including copy number variants (CNVs), if combined with appropriate and tailored bioinformatics analyses.

This project is the founding research that leads to your project. After trying to retrieve SVs and CNVs from the DNA sequencing data (genomes) already published in that paper and publicly available, you will try to explore the effect of such variants using transcriptomics data of the same strain, data that no one has yet analyzed. It is then important for you to understand the content of that paper. If you have a question, please discuss it with your team, or other students, or with the teaching team (we are here to help and we will explain as long as you need it). I encourage you to read these papers several times over the next coming weeks, as Genomics and Transcriptomics Analyses are based on those data.

Schedule

The next pages of the lab book are organized according to the schedule, week after week. All required/recommended readings are outlined in the lab book.

A tiny organism with great power: Meet the nematode *Caenorhabditis elegans*

By Victoria R. A. Barbosa

These microscopic creatures with an almost unpronounceable offer more than what the eyes can see. Back in the sixties, when Sydney Brenner first explored this round worm as a model organism for genetic studies, I bet that not even him could foresee the impact they would later cause. They were the first multicellular organism to have its genome fully sequenced (C. elegans Sequencing Consortium 1998), have been used to understand many fundamental questions of developmental biology and neurobiology and led to three Nobel prizes in the past two decades. But what exactly makes them so special?

Let's start with the basics:

- ➔ *C. elegans* is a free-living nematode found worldwide that is approximately 1 millimeter long (adults). Some of the most studied strains were isolated in England and Hawaii, but wild isolates can be found in practically every continent. Who's up to travel around and try to find some new strains? I volunteer! 🙋🏻‍♀️
- ➔ Speaking of England, this is where the most used in labs *C. elegans* strain comes from: N2 was isolated in Bristol by Sydney Brenner himself and have been propagated in labs for decades after that, being considered as the wild-type control for any experiment in this species till this very day.
- ➔ Its life cycle is impressively fast, taking only 3 days for an egg to become an egg-laying adult when kept at 25°C. They have four larval stages (L1 – L4) before reaching adulthood. Curiously, if exposed to critical circumstances such as overpopulation and food shortage, the larvae enter an alternative life cycle called dauer, in which their food consumption and metabolism is reduced, allowing the population to survive for muuuch longer!
- ➔ Their genome carries 5 pairs (diploid) of autosomal chromosomes (often referred to as LG I, II, III, IV, V) + 1 pair of sex chromosomes (XX = hermaphrodites). Males can also be present in the population, but only in rare cases of chromosomal non-disjunction (<0.2%), in which they would carry only one copy of the X chromosome (XO). Males and hermaphrodites have different body features that are very easy to distinguish, such as their size, tail, and specific internal structures.
- ➔ They are self-fertilizing, meaning that hermaphrodites produce their own oocytes and sperm. This feature is extremely useful for genetic studies, since you would always have a homogeneous population (every mutation or variant will be homozygous) coming from the same "mother", with no interference of gametes from another individual. Hermaphrodites have a limited capacity to produce sperm and usually

generate around 300 eggs but are also able to mate with males and produce viable progeny, laying up to 500 eggs instead.

Okay, but how on earth do we find these worms, and more importantly, keep them in the lab? They can often be found on rotting vegetable matter (rich in bacteria, their favorite food), but luckily, we don't have to deal with any of this. Instead, scientists count with this huge facility in the US that provides all types and shapes of *C. elegans*: the CGC or Caenorhabditis Genetic Center. You simply order it online and in a matter of days you will have them mailed to you and ready to go! To maintain them under lab conditions, they are kept in Petri dishes that contain a thick layer of agar (a nutrient-rich "jelly"). Each of these plates is also seeded with bacteria, usually *Escherichia coli*, which is what the worms are fed with. They can grow in a range of temperatures from 12°C to 25°C, and that is directly related to how fast that happens. The higher the temperature, the faster they will go through the life cycle. And here comes the ~coolest~ part: they can be frozen for years and revived whenever needed!

Last, but not least, why use *C. elegans*? It is undeniable that they have many advantages in comparison to other model organisms, such as its small size, large brood size, ease of cultivation, low-cost maintenance, long-term cryopreservation, quick generation time, transparency, and invariant cell number and development. They also have defined tissues, such as epidermis, muscle cells, nervous, digestive, and reproductive systems. And above all, they carry many genes that are orthologous to humans (equivalent function), which allow for the study of a diversity of human conditions in a much simpler organism. It is also quite interesting how they carry differences in their genome according to the place in which they were isolated (like humans from different parts of the world, that also carry different traits). That is extremely useful to study the effects of each genetic background in specific mutations, as we have been exploring in our lab. In summary, they are extremely versatile and still have a huge impact in molecular genetics and developmental biology discoveries. How cool is that?

Some additional information:

- ➔ Gene names: usually written in lower-case, italicized letters, carrying three letters, a dash, and a number. For example: *nhl-2*, *cgh-1*, *dpy-10*. Proteins would be written with the gene name, but this time in capital letters, non italicized. Example: NHL-2.
- ➔ Their body is transparent, which allowed for some interesting discoveries: they have a determined and immutable number of autosomal cells (959, to be exact) and each of them was carefully mapped, meaning that this is one of the only species in which you know exactly each individual cell fate and can follow the life of individual cells. The number and connectivity of all the neurons in an adult worm have also been identified (302 neurons).
- ➔ Balancer chromosomes: as I mentioned, worms are a great tool for genetic studies, but what happens if you are investigating a lethal mutation (mutation that, if homozygous, will be deleterious enough for either kill the worm, prevent the larvae from hatching, or stop the development of the embryo)? How are you supposed to keep the population of

worms alive to assess that phenotype? One answer to that is by using balancer strains. Those strains are known to carry large structural variations (inversion, deletion, duplication, or complex rearrangements) either induced by X-rays, UVs or gamma radiations in a lab, or that occurred naturally (rare). Such large structural variants will "balance" the primary mutation: avoiding crossover during meiosis as the two copies of the chromosomes are too different, keeping all mutations in that region as heterozygous, instead of homozygous (see video on "Other resources" section). In other words, this will guarantee that you maintain a stable heterozygote population for the mutation of interest (mutation "A", for example). They can also carry a mutation that will alter a physical trait in the worm to make it possible to visualize which ones are heterozygotes for A and balancer variant so you can differentiate them from homozygous balancer genotype and still maintain worms carrying the mutation of interest in the population.

Still not convinced of *C. elegans* importance? Check some other great discoveries made in this species:

- ➔ In 2002, Sydney Brenner won the Nobel Prize in Physiology or Medicine for his discoveries on the mechanism of apoptosis, which was made possible by studying the individual cell fates in *C. elegans*.
- ➔ Later, in 2006, Andrew Fire and Craig Mello also received a Nobel Prize for the development of the RNA interference gene silencing method (RNAi) in *C. elegans*, a very useful tool for gene knock-down and genotype-phenotype association studies used later in many species.
- ➔ Additionally, in 2008, Martin Chalfie received the Nobel Prize in Chemistry for the discovery of the green fluorescent protein, GFP, and its use for tagging relevant genes first in *C. elegans*. GFP is found originally in a bioluminescent jellyfish (*Aaequorea victoria*), but with the use of *C. elegans*, Chalfie probed its use in other species for monitoring gene expression and protein localization.

References:

Brenner, S., 1974 The genetics of *Caenorhabditis elegans*. Genetics 77: 71-94.

Brenner, S., Horvitz, H. R., & Sulston, J. E. Genetic Regulation of Organ Development and Programmed Cell Death.

Chalfie, M. (1995). Green fluorescent protein. Photochemistry and photobiology, 62(4), 651-656.

Corsi, A. K., Wightman, B., & Chalfie, M. (2015). A transparent window into biology: a primer on *Caenorhabditis elegans*. Genetics, 200(2), 387-407.

Fire, A., S. Xu, M. K. Montgomery, S. A. Kostas, S. E. Driver, and C. C. Mello, 1998 Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans.* Nature 391: 806-811.

Sulston, J. E. and H.R. Horvitz, 1977 Post-embryonic cell lineages of the nematode, *Caenorhabditis elegans.* Dev. Biol. 56: 110-156.

Week 1 - Meet *C. elegans*

Deadlines Week 1: Team Survey Building due Jan. 8.

Monday, January 08, 2024

- Introduction to MDSC 301 CURE

We will go through the organization of the class, the schedule and the lab book.

- Introduction to the PROJECT
- Survey for Team Building

https://survey.ucalgary.ca/jfe/form/SV_1TcBecwJi52IprE

Reading for Wednesday, January 10, 2024

- ☐ **A tiny organism with great power: Meet the nematode *Caenorhabditis elegans* (Pages 16-18)**

Other resources (optional):

- ☐ Unusual Labmates: C. elegans
- ☐ Sydney Brenner - Caenorhabditis elegans: the perfect hermaphrodite (150/236)
- ☐ Wormbook (*C. elegans* “Wikipedia”) Introduction - http://www.wormbook.org/chapters/www_celegansintro/celegansintro.html
- ☐ Genetic Balancers in Wormbook - http://surl.li/dupoy
- ☐ Balancer Chromosomes Explained

Wednesday, January 10, 2024

- Meet the worms and the research team (dry lab and wet lab): During this session, you will get a chance to look at some worms through the microscope and meet trainees in Bioinformatics from Dr. Tarailo-Graovac lab, with very different backgrounds, but all applying bioinformatics analyses in their project.
    a. The session will be organized as a fair. Walk around and visit the different “booths" to see some worms under the microscope and talk to people working with Bioinformatics every day.

b. Take advantage of that time to ask them any question about lab work, *C. elegans*, balancers, bioinformatics…
c. See below the questions you will have to answer for the assignment and the rubric.
d. The assignment is individual.
e. The assignment is to be done on D2L (available under Assessment > Quizzes)
f. The assignment is due on **Friday, Jan. 19 - 11:59pm.**
g. In case of issue with submission through D2L, submission by email will be accepted before the deadline.

Assignment questions

1. What aspects of *C. elegans* were surprising to you?
2. What aspects of working in the dry lab were surprising to you?
3. Do you think *C. elegans* are a practical choice as a model organism for genomics? Explain why or why not?
4. Reflect on a collaborative project where both wet lab and dry lab techniques were utilized. What stood out to you about this collaboration?
5. Following the session, have your perspectives on research changed? If so, in what specific ways?
6. What insights did you gain about Bioinformatics in a real research environment?

Grading Rubric for Assignment questions (Each question equals 5 points)

6 questions - Total points: __/30

Criteria	0 Point	0.25 Point	0.5 Point	0.75 Point	1 Point
Relevance (1pt): Does your answer align to the question?	Response lacks alignment with the question; content is unrelated or minimally related to the prompt.	Limited relevance; some connection to the question but lacks a clear focus.	Generally relevant; addresses the question with a clear focus on key points.	Highly relevant; effectively addresses the question, demonstrating a clear and focused response.	Exceptionally relevant; addresses the question comprehensively, providing depth and insight.
Clarity (1pt)	Response is unclear, making it difficult for the reader to follow the line of thought.	Somewhat clear; the response has some coherence but may lack smooth transitions.	Clear and organized; the reader can easily follow the line of thought.	Very clear; well-structured with smooth transitions, facilitating easy understanding.	Exceptionally clear; exceptionally well-organized, with a logical flow and seamless transitions.
Significance (1pt)	Lacks personal reflection; the response is	Limited personal reflection; includes some	Demonstrates a deep personal reflection on the	Displays a significant and meaningful	Exhibits an outstanding personal

	superficial and lacks depth.	depth but lacks a profound connection to the proposed activity.	proposed activity, showing thoughtful consideration.	personal reflection, providing insights into the proposed activity.	reflection, demonstrating a profound connection to the proposed activity.
Mechanics (1pt)	Numerous typographical, spelling, and grammatical errors throughout the response.	Some typographical, spelling, or grammatical errors present, impacting overall clarity.	Consistently avoids major typographical, spelling, and grammatical errors, enhancing clarity.	Virtually error-free; minimal typographical, spelling, or grammatical errors.	Impeccable mechanics; response is free of typographical, spelling, and grammatical errors.
Length (1pt): 150-200 words per question.	Below 150 words or exceeds 200 words, deviating significantly from the specified range.		Falls within the 150-200 word range but may be slightly below or above the specified limits..		Optimal length; response is precisely within the 150-200 word range, demonstrating effective conciseness.

Week 2 - Genomics and Structural Variants

Deadlines Week 2: Quiz 1 on Jan. 15 + Reflection Assignment 1 (D2L - Quizzes) Jan. 19

Monday, January 15, 2024

- 10:30-10:45 - Quiz: Questions on *C. elegans*
  - The quiz starts at 10:30am and ends at 10:45am. Do not be late! You won't have extra time.
  - The quiz will be on D2L and password secured. The password will be delivered to you in class.
- Lecture: Introduction to Genomics and Structural Variants

Additional resources:

- Illumina Sequencing by Synthesis: https://youtu.be/fCd6B5HRaZ8
- Long-read sequencing (PacBio): https://youtu.be/_ID8JyAbwEo
- Ultra-long read sequencing (Nanopore): https://youtu.be/1_mER5qmaVk
- Optical Genome Mapping (Bionano Genomics): https://youtu.be/S2ng6glu04I
- Structural Variants: https://youtu.be/1p-txfOt_io

Reading for Wednesday, January 17, 2024

These readings are to be completed in groups. Each group member should read at least one article. You will then share what you have learned in class.

Sequencing

- ☐ **Larson NB, Oberg AL, Adjei AA, Wang L, A clinician's guide to bioinformatics for next-generation sequencing, Journal of Thoracic Oncology (2022), doi: https://doi.org/10.1016/j.jtho.2022.11.006.**
- ☐ **Lebo MS, Hao L, Lin CF, Singh A. Bioinformatics in Clinical Genomic Sequencing. Clin Lab Med. 2020 Jun;40(2):163-187. doi: 10.1016/j.cll.2020.02.003. PMID: 32439067.**
- ☐ **Slatko BE, Gardner AF, Ausubel FM. Overview of Next-Generation Sequencing Technologies. Curr Protoc Mol Biol. 2018 Apr;122(1):e59. doi: 10.1002/cpmb.59. PMID: 29851291; PMCID: PMC6020069.**
- ☐ **Hu T, Chitnis N, Monos D, Dinh A. Next-generation sequencing technologies: An overview. Hum Immunol. 2021 Nov;82(11):801-811. doi: 10.1016/j.humimm.2021.02.012. Epub 2021 Mar 19. PMID: 33745759.**

Large Variants

- ☐ **Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. Structural variant calling: the long and the short of it. Genome Biol. 2019 Nov 20;20(1):246. doi: 10.1186/s13059-019-1828-7. PMID: 31747936; PMCID: PMC6868818.**
- ☐ **Ho SS, Urban AE, Mills RE. Structural variation in the sequencing era. Nat Rev Genet. 2020 Mar;21(3):171-189. doi: 10.1038/s41576-019-0180-9. Epub 2019 Nov 15. PMID: 31729472; PMCID: PMC7402362.**

Wednesday, January 17, 2024

- Active Learning Activity

In teams, you will work on paper panels to create a poster regarding specific thematic questions listed below. Every 10 minutes, you will walk to another poster/theme and complete the poster started by other groups before you. The activity is open books - use any resource that you'd like.

    - Genomics - Analysis of DNA sequences with high-throughput technologies
        - Different technologies available for high-throughput analyses of genomic sequences (DNA)
        - Types of studies that can be conduct using high-throughput DNA technologies
        - Biases and error sources: in the design of a study, in the sample collection and library preparation…
        - Biases and error sources: in the bioinformatics analysis
        - Main steps/files of bioinformatics analyses
        - Library preparation: Main steps in the laboratory to prepare a sample for high-throughput sequencing
    - Structural Variants
        - Types of SVs
        - How can we detect SVs?: technologies and analyses
        - Potential biological Impact of SVs
        - How do SVs arise in a genome?

Introduction to Genomics and Structural Variants

By Rumika

DNA and Sequencing

Before we embark on our journey into the intricate world of genomics, let's start at the very beginning—with **DNA**. Deoxyribonucleic acid is the hereditary material in humans and almost all other organisms, consisting of a sequence of nucleotides—adenine (A), thymine (T), cytosine (C), and guanine (G). These nucleotides form the basis of the genetic code, dictating the instructions for building and maintaining living organisms. **The genome** is the complete set of genetic material (DNA in most organisms) present in an individual or organism. It includes all the genes, along with non-coding sequences, regulatory elements, and other structural components. For diploid organisms, such as *C. elegans* and humans, each sequence is present in two copies. **DNA sequencing** involves the process of determining the order of nucleotide bases (A, T, C, and G) within a DNA segment. Sequencing an entire genome, encompassing all the DNA of an organism, remains a complex undertaking. This process entails breaking down the genomic DNA into numerous smaller fragments, sequencing each fragment, and then piecing together the sequences to create a unified and comprehensive "consensus."

Sequencing Technologies

Numerous technologies have emerged to decode the sequence of DNA. In this module, we'll focus on the pioneering **Illumina sequencing technology**, a cornerstone in modern genomics. Other technologies, such as Pacific Biosciences and Oxford Nanopore, offer unique advantages, but Illumina's widespread use and efficiency make it still a central player in the field.

Before sequencing, DNA must undergo a meticulous preparation process known as **library preparation**. For short-read sequencing technologies like Illumina, this involves fragmenting the genomic DNA into short segments, attaching adapters, and amplifying these fragments to create a library of DNA fragments ready for sequencing. This step is critical for generating uniform, high-quality sequencing data. Illumina sequencing relies on a process called **sequencing by synthesis**. In a nutshell, it involves amplifying DNA fragments on a solid surface, attaching fluorescently labeled nucleotides, and capturing images as each nucleotide is incorporated into the growing DNA strand. This iterative process generates millions of short DNA sequences, or reads, in a massively parallel fashion. The key step involved are:

1. Library Preparation: DNA is fragmented, adapters are added, and the fragments are amplified to create a library of DNA molecules.
2. Cluster Generation: DNA fragments are immobilized in clusters on a solid surface.

3. Sequencing: Fluorescently labeled nucleotides are added, and images are captured as each nucleotide is incorporated.
4. Base Calling: The sequence is determined by decoding the fluorescent signals.
5. Data Analysis: Raw sequencing data undergoes bioinformatics analysis to align reads, identify variants, and extract meaningful biological information.

Bioinformatics Analysis of Sequencing Data

One of the main bioinformatics steps is **read alignment**, where the short DNA sequences (reads) are mapped to a reference genome. This step ensures accurate positioning of reads and forms the basis for downstream analyses.

Variant calling is one of the downstream analyses possible with sequencing datasets. It involves identifying differences (variants) between the sequenced DNA and the reference genome. Single nucleotide polymorphisms (SNPs) and small insertions/deletions (indels) are among the variants detected.

Additional steps are usually performed to ensure the quality of the analysis and increase the accuracy of the calling: trimming the reads, removing the adapters, marking the duplicates, filtering calls…

Structural Variants: Unraveling Genomic Architecture

Structural variants (SVs) are alterations in the structure of the genome that involve changes in the arrangement of DNA segments. These variations come in various forms. Some involve a change in the number of copies, some do not.

- Insertions: Additional DNA segments are incorporated (no change in copy number)
- Deletions: Segments of DNA are missing (one copy is missing)
- Duplications: Sections of DNA are repeated and inserted right before or after the original sequence, or somewhere else in the genome (one additional copy)
- Inversions: DNA segments are reversed (no change in copy number)
- Translocations: DNA segments move from one location to another (no change in copy number).

Being able to detect SVs is crucial as they can impact gene function, contribute to phenotypes and disease, and influence evolutionary processes. Bioinformatics tools are employed to detect and characterize these variations, providing insights into the genomic landscape. However, no tool is currently recognized in the field as able to detect with enough accuracy and efficiency all types of SVs. As you delve into the complexities of genomics and structural variants, you'll gain a profound appreciation for the technologies and analytical tools that unlock the mysteries of the genetic code.

Week 3 - Write a Proposal - Genomics analysis

Deadlines Week 3: Quiz Genomics (D2L) Jan 22 + Proposal (Group - D2L Dropbox) and Peer Evaluation 1 (Individual - D2L Dropbox) Jan. 26

Reading for Week 3

- ☐ **Maroilley T, Li X, Oldach M, Jean F, Stasiuk SJ, Tarailo-Graovac M. Deciphering complex genome rearrangements in *C. elegans* using short-read whole genome sequencing. Sci Rep. 2021 Sep 14;11(1):18258. doi: 10.1038/s41598-021-97764-9. PMID: 34521941; PMCID: PMC8440550.**
- ☐ **Maroilley T, Flibotte S, Jean F, et al. Genome sequencing of *C. elegans* balancer strains reveals previously unappreciated complex genomic rearrangements. *Genome Res*. 2023;33(1):154-167. doi:10.1101/gr.276988.122**

These articles are the result of the project in which the MTG Lab is promoting short-read whole genome sequencing as an efficient method to detect structural variants (SVs) including copy number variants (CNVs), if combined with appropriate and tailored bioinformatics analyses.

Monday, January 22, 2024

- 10:30-10:45 - Quiz: Questions on “Genomics and Structural Variants”
  - The quiz starts at 10:30am and ends at 10:45am. Do not be late! You won’t have extra time.
- Lecture - Introduction to PROJECT DESIGN, FILES and PROCESS (~30 minutes)
- Literature review (use PubMed: https://pubmed.ncbi.nlm.nih.gov/)
  - Work with your team to explore the literature and put together a plan to analyze sequencing data.
  - The problem you are trying to resolve in this project is (**YOUR AIM**): **“Detection of structural variants in short-read whole genome sequencing data”**
  - Make sure that your pipeline will cover all the necessary steps to limit errors and biases
  - Your input files are FASTQs from one of these strains: SS746, KR2839, MT690, CZ1072, and RW6002. N2 strain is your control.
  - The output is expected to be a VCF file.
  - It is expected that your proposal will include a visualization of the variants identified.

Visualization tools:

- ☐ https://cmdcolin.github.io/awesome-genome-visualization/?latest=true
- ☐ https://github.com/cmdcolin/awesome-genome-visualization

Articles to help you design your analysis

- ☐ Tattini L, D'Aurizio R, Magi A. Detection of Genomic Structural Variants from Next-Generation Sequencing Data. Front Bioeng Biotechnol. 2015 Jun 25;3:92. doi: 10.3389/fbioe.2015.00092. PMID: 26161383; PMCID: PMC4479793.
- ☐ Yang L. A Practical Guide for Structural Variation Detection in the Human Genome. Curr Protoc Hum Genet. 2020 Sep;107(1):e103. doi: 10.1002/cphg.103. PMID: 32813322; PMCID: PMC7738216.

Tips to design a bioinformatics pipeline:

- ☐ Make sure that the tools you pick will work on your type of data: short-read whole genome sequencing
- ☐ Make sure that the tools you pick will be able to work with each other: Tool 1 will create File A that should be compatible with Tool 2. If not, plan to have intermediate steps (it is ok if you do not how exactly what those steps will be)
- ☐ Make sure that your pipeline will allow you to answer your original question: detecting SVs in short-read whole genome sequencing
- ☐ Include quality steps
- ☐ Instead of going blindly into PubMed, you can start by finding a research study that worked on a similar question and look at what tools they used
- ☐ Rely of studies that has compared multiple tools to pick the one that seems to most promising
- ☐ It is also ok to want to assess tools just recently published.

Wednesday, January 24, 2024

It is time now to write your proposal.

- ➔ Use the template provided below.
- ➔ The proposal + Peer evaluation are due on **Friday Jan. 26, 11:59pm.**
- ➔ **Please upload one proposal per group in D2L.**
- ➔ Accepted format: pdf.
- ➔ **Authors should be listed under the title.**
- ➔ **Make sure that all parts are ALIGNED (the methods should be applicable to the type of data you have, and should allow you to answer your question)**. A proposal has to make sense as a whole!

Research Proposal Template

Assignment due on Jan. 26, 2024, Prepared by Victoria R A. Barbosa

A research proposal is meant to answer three main questions:

- ➔ What are you doing? Your exact research topic and hypothesis.
- ➔ Why are you doing it? Background information and/or preliminary data that justifies your research question.
- ➔ How do you plan to do it? Detailed methodology, including the reasons for your choice of a method over another (if applicable).

A proposal is like a plan. It is ok to change the plan later, as long as you have a good reason to and that you are still meeting your objectives. In approximately **one page, single-spaced, Calibri 12pt (references and figure not included)**, briefly describe your research project and potential achievements **(2 points for respecting the length). Clarity and quality of writing** will count for **3 points**. The text should be divided into the following sections:

- ☐ **Title (1 point)**: should be objective and inform what your research is about. For example: "The use of scRNA-seq for investigation of pancreatic tumor in smokers". This title mentions the topic of study + methods applied + specific target/public. Keep it simple, but informative (and not too long).
- ☐ **List of authors**
- ☐ **Introduction (3 points)**: ~1 paragraph with brief background information on the research topic. It is in this part that you will start to describe the existent "problem" that your research will collaborate to solve. For example: *"Pancreatic cancer is one of the most aggressive forms of cancer and affects XXX people yearly. The average survival of a patient with a pancreatic tumor is 3-5 years after the tumor is installed. This aggressive cancer is often related to tobacco use, being the thought cause of disease in 25% of cases. The use of modern techniques, like scRNA-seq, might provide a better understanding of the tumor formation and comparison to surrounding healthy cells…"*
- ☐ **Hypothesis and aims (3 points):** in a sentence or two, explain what is the expected result for your research, and what you hope to achieve with it. For example: *"Using scRNA-seq, I expect to observe different RNA expression levels when comparing these divergent cell populations… Therefore, our main aim is to identify new genetic markers for pancreatic cancer that are specifically present in the tumor cells…"*
- ☐ **Preliminary data (if applicable):** if any initial experiment or analysis has been developed, it is worth mentioning what was done and how that impacts your hypothesis. For example: *"based on preliminary data observed through X experiment, it is known that cancer cells may vary even within the same tumor, exemplifying how a deeper*

analysis of these different populations might be essential for a better understanding of…"

- ☐ **Experimental plan (3 points):** in a paragraph or two, explain exactly how you plan to lead your research: methods that will be applied, order of experiments/analysis, what each technique will show you, etc. For example: *"tumor samples were provided from biopsies collected between January and September 2021, at Seattle Grace Mercy West Hospital, with patient consent. scRNA-seq will be done using Illumina Next-Seq 1000 technology…"*
- ☐ **Significance (3 points):** here, you will summarize in a couple sentences the importance of your work to the population or scientific community. For example: *"only up to 10% of patients with early diagnosis for pancreatic cancer are disease-free after treatment. A better understanding of differential gene expression for these cells might be relevant for new therapeutic approaches and represent a better prognosis for these patients…"*
- ☐ **References (2 points):** there is no secret here! Just list your references using APA or VANCOUVER style, references should be listed in alphabetical order. Please use a Reference Manager tool such as Zotero or EndNote.
- ☐ **BONUS -** Figure showing the bioinformatic process – 1 point (not included in the 1-page limit) – see example in this paper: https://doi.org/10.3389/fmicb.2019.00362

Grading Rubric for Research Proposal

Total point: __/20

Criteria	0 Point	1 Point	2 Points	3 Points
Length (2 points)	Exceeds one-page requirement.	Slightly over/under the one-page requirement.	Precisely meets the one-page requirement.	
Clarity and Quality of Writing (3 points)	Highly unclear with major communication issues and grammatical errors.	Somewhat clear, but has some grammatical errors and other issues in communication.	Somewhat clear, minor issues in communication.	Clear and effective communication.
Title (1 point)	Not present or entirely lacks clarity and informativeness.	Objective, informative, and clear.		
Introduction (3 points)	Lacks background information and fails to identify the research problem.	Some background provided but contains inaccurate information, has major issues with clarity.	Background provided but lacks clarity or effective identification of the research problem.	Concise background, clear, well-justified identification of the research problem.
Hypothesis and Aims (3 points)	No hypothesis or aims stated	Hypothesis and aims stated but entirely lacks specificity.	Hypothesis and aims stated but is somewhat unclear or lacks specificity.	Clearly stated expected results and outlines goals.

Experimental Plan (3 points)	No experimental plan provided.	Extreme lack of details regarding methods and expected outcomes or misalignment with the aim.	Somewhat unclear or lacks detail regarding methods and expected outcomes or crucial steps missing	Clearly outlines methods, experiments, and expected outcomes.
Significance (3 points)	No articulation of importance or significance of the project.	Significance is somewhat stated but is extremely lacking depth and clarity.	Significance is stated but lacks depth or clarity.	Clearly articulates importance with well-supported explanation of significance.
References (2 points)	No references provided or significantly deviates from APA or vancouver style.	References present but may not be in alphabetical order or in APA or vancouver style.	Listed in APA or vancouver style, in alphabetical order. Utilizes a Reference Manager tool.	
Bonus (1 point)	No figure showing the bioinformatic process included	Figure showing the bioinformatic process included		

Some additional writing tips:

- ➔ Avoid unnecessary abbreviations and terms that are only understood by specialists in that area (like someone with little Bioinformatics experience would read your text).
- ➔ Keep it short: often, fewer words are better. Write short and clear sentences with no more than 20 words each. One idea, one sentence.
- ➔ Avoid writing sentences in the third person. Instead of “fluorescence microscopy will be applied for data analysis that will be conducted by the lab technician…” say “the lab technician will analyze the data using fluorescence microscopy.”
- ➔ Write a single idea in each paragraph. Avoiding too much information in the same paragraph will make the text – and the idea – flow much better.
- ➔ In the Methods section, name the tools that you will be using (and reference them!), and explain briefly how they work, why you choose them.
- ➔ The Introduction section should never be the longest one. Your ideas are what matters, not the context! A proposal is NOT a literature review!
- ➔ Review your text for spell-checking (do not hesitate to use Grammarly!) and do not hesitate to ask for someone else’s help. Sometimes, we read our own words so often that we end up missing a detail or two. The point of view of another person might help with clarity, flow and even with catching those minor typos that our own eyes miss.
- ➔ Work as a team, like in a real lab: everyone should read, comment, and approve the final version of the proposal.

Check other tips at:
https://www.schulich.uwo.ca/biochem/research/docs/nserc_writing_guide_pres.pdf

Week 4 - Genomic Analysis on Galaxy and HPC

Deadlines Week 4: Quiz Genome Analysis (D2L) Jan 29 + Report 1 (ungraded - individual) Feb. 02

Requirements for Week 4

- ☐ **Create an account on Galaxy (see tutorial below) - with your @ucalgary email**
- ☐ **Upload the data from the shared history (See tutorial below)**

Optional readings

- Galaxy Tutorials | How to use Galaxy for Bioinformatics (Beginners)
- https://www.youtube.com/watch?v=bz93ReOv87Y
- https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/
- More about how to use a cluster (that webpage is about ARC, another cluster at the UofC. Everything works for TALC too): https://rcs.ucalgary.ca/ARC_Cluster_Guide#/scratch:_Scratch_file_system_for_large_job-oriented_storage
- More about SLURM, the scheduler: https://blog.ronin.cloud/slurm-intro/

Timeline for Week 4

By the end of Monday, each student should have the first steps running on GALAXY.

For Wednesday, each student should have obtained the BAM files for one strain + control strain (N2).

By the end of Wednesday, each student should have obtained the two VCFs (one for the strain + one for control strain).

For Week 5, students should have the VCFs ready and each team should have a strategy to filter the SVs (based on quality and uniqueness).

PROTOCOL: Analysis of genome sequencing with Galaxy and HPC

FOR THE ANALYSIS: Choose one strain (SS746, KR2839, MT690, CZ1072, and RW6002) and one SV from the list below.

YOU ARE NOT IMPLEMENTING YOUR PROPOSAL!

The next pages will walk you through the detection of SVs using Galaxy and HPC in short read data according to a certain protocol that might be different from your proposal.

Purpose of this analysis: test different SV callers and evaluate each tool in term of accuracy at detecting known events (the one you picked) and new/unique events (not in controls)

With your team, choose on which strain published in Maroilley et al. (2023) you will be working (you will work on that same strain for the rest of the semester). In addition, with your team, choose one SV/CNV reported and confirmed in Maroilley *et al*. (2022) in your strain that **overlaps/falls into a gene**.

No matter your choice, you will have to run one control strain (N2). *Tip: On Galaxy, run each strain in a different History to avoid mixing up samples.*

Strain	*Type*	*chr*	*start*	*end*	*genes*
SS746	*Reciprocal translocation*	*II/III*	*II:6296872*	*III:3635354*	*linc-95*
KR2839	*Inversion*	*I*	*11249952*	*15003739*	*F46A8.11, mtd-1*
KR2839	*Duplication*	*I*	*14573853*	*14580325*	*Y105E8A.25, Y105E8A.53, Y105E8A.55*
MT690	*Deletion*	*III*	*3607207*	*3695411*	*21 genes*
MT690	*Duplication*	*X*	*11845164*	*11852389*	*F54F7.2, F54F7.11, F54F7.9, F54F7.3, F54F7.14*
CZ1072 and KR2839	*Duplication*	*I*	*10565986*	*10577022*	*pas-5, rla-0, F25H2.16, tct-1, F25H2.14, F25H2.12, rtel-1*
CZ1072	*Duplication*	*III*	*3891219*	*3898488*	*tir-1*
RW6002	*Deletion*	*IV*	*5370371*	*5371720*	*F41H10.3*
RW6002	*Duplication*	*III*	*11830810*	*11834895*	*rsp-8, ntl-3*

The Genomic Analysis shall be conducted as a work group. Each student will be attributed a part of the analysis. Note that the assignment for this analysis is a manuscript (see template, requirements and rubrics pages). Start writing the manuscript as soon as possible by analyzing the results of each step as per requirements (refers to Manuscript template).

Follow the pace of the class - this analysis and assignment cannot be done last minute. Each step will require troubleshooting and computational time differently depending on your Internet connection and the usage of the Galaxy server. Do not under-estimate the necessary time!

This assignment lies into each student accomplishing his tasks on time and delivering it to the team. Communication is key! Plan regular check in on each other, use the time in class to help each other, plan and discuss the project.

Here is how to distribute the work in your team. If your team has only 4 members, ignore Student 5.

	Student 1	**Student 2**	**Student 3**	**Student 4**	**Student 5**
Samples	Your strain + N2	Your strain + N2	Your strain + N2	Your strain + N2	Your strain + N2
Control quality	GALAXY: Fast-QC	GALAXY: Fast-QC	GALAXY: Fast-QC	GALAXY: Fast-QC	GALAXY: Fast-QC
Trimming and remove adapters	GALAXY: Trimmomatic	GALAXY: Trimmomatic	GALAXY: Trimmomatic	GALAXY: Trimmomatic	GALAXY: Trimmomatic
Alignment	GALAXY: BWA-MEM	GALAXY: BWA-MEM	GALAXY: BWA-MEM	GALAXY: BWA-MEM	GALAXY: BWA-MEM
Mark Duplicates	GALAXY: MarkDuplicates (Picard)	GALAXY: MarkDuplicates (Picard)	GALAXY: MarkDuplicates (Picard)	GALAXY: MarkDuplicates (Picard)	GALAXY: MarkDuplicates (Picard)
Calling	GALAXY: Lumpy	GALAXY: Basil	GALAXY: DELLY	HPC: Manta	HPC: SeekSV

In this Analysis, each student will proceed as follow (Please see the detailed tutorial starting from Page 35!):

1. The data has been uploaded on Galaxy in a shared History. In Galaxy, use this link to access the History and import it in your account.
2. Check the quality of the FASTQ files
    a. Tool: FAST-QC
    b. Input: .fastq.gz (R1 and R2)
    c. Output: html report
3. Trim the reads to remove adapters and bad quality bases
    a. Tool: Trimmomatic
    b. Input: .fastq.gz (R1 and R2)
    c. Output: trimmed.fastq.gz

4. Check the quality of the FASTQ files has been improved after trimming and removing the adapters
    a. Tool: FAST-QC
    b. Input: .fastq.gz (R1 and R2)
    c. Output: html report
5. Align the reads on the reference genome
    a. Tool: BWA
    b. Input: trimmed.fast.gz (R1 and R2)
    c. Output: .bam
6. Mark the duplicated reads
    a. Tool: MarkDuplicates
    b. Input: sorted.bam
    c. Output: sorted.dedupped.bam
7. Check the quality of your alignment
    a. Tool FAST-QC
    b. Input: sorted.dedupped.bam
    c. Output: html report
8. **Call the SVs ON GALAXY - ONLY Students 1, 2, 3**
    a. Tools: Lumpy, Delly, Basil
    b. Input: sorted.dedupped.bam
    c. Outputs: VCF
9. **Call the SVs ON HPC - ONLY Students 4, 5**
    a. Tools: Manta, SeekSV
    b. Input: bam files in /work/TALC/mdsc301_2024w/data/bams/
    c. Outputs: VCF

Note: Write down any observation, difficulty, challenge, comment you have. This is part of the manuscript that you will have to submit as a team after reading week. The following pages will guide you through running the different tools necessary for detecting SVs in short-read whole genome sequencing data.

Monday, January 29, 2024

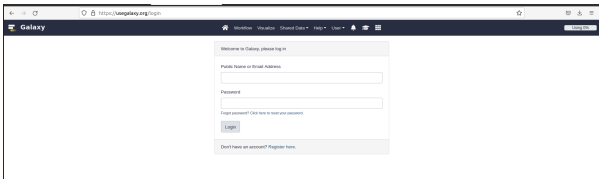
- 10:30-10:45 - Quiz: 10 questions on “Bioinformatics genome analysis”
    - The quiz starts at 10:30am and ends at 10:45am. Do not be late! You won’t have extra time. The quiz will be posted on D2L and password protected.
- Lecture (~20 minutes): Introduction to Galaxy and HPC (TALC)
- Running Genomic Analysis - Objective: up to BAM files

Wednesday, January 31, 2024

- ☐ Have BAM files ready
- ☐ We will work on getting the VCFs either by GALAXY or HPC

Galaxy

Dr. Tatiana Maroilley

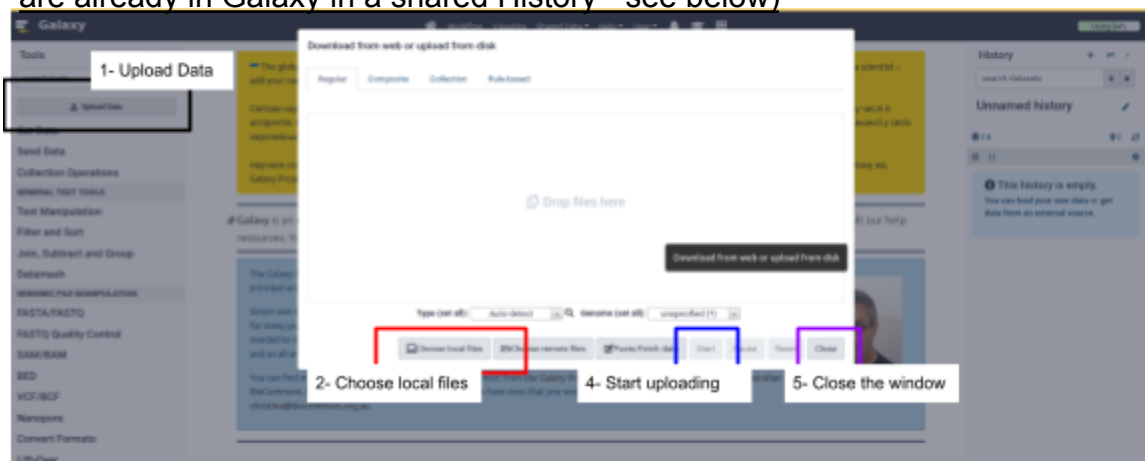


With your favorite browser, go to https://usegalaxy.eu/login and Register or create an account (WITH YOUR @ucalgary.ca email!!!)

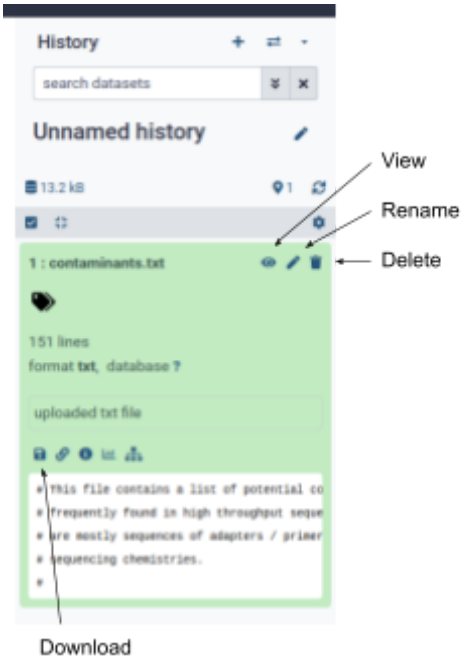
We will be using **Galaxy Europe** (because no SV calling tool available on the main usegalaxy.org).

Galaxy Main Page

Uploading data to Galaxy (from an outside source - just for your information - the data are already in Galaxy in a shared History - see below)



1. Click on “Upload Data” (Upper left corner)
2. Click on “Choose local files” in the pop-up window
3. Select the files you want to upload on Galaxy from your computer
4. Start uploading by clicking on “Start”
5. Close the window by clicking on “Close”



The file will appear in your history:

- Grey: in the queue, waiting to be uploaded
- Orange: Uploading
- Green: Uploaded successfully
- Red: An error has occurred

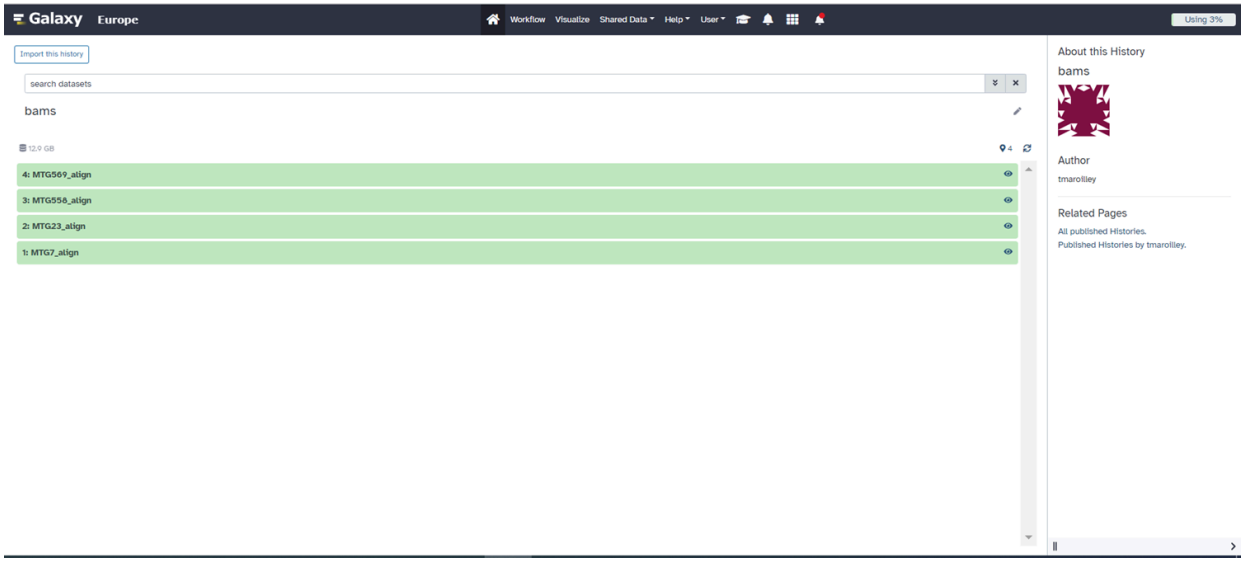
Click on a file in History to see options.

Get data from a shared History

The Instructor has created a shared History with all the necessary files already uploaded (FASTQs, reference genome, list of adapters…).

Link: https://usegalaxy.eu/u/tmaroilley/h/genomicsdata

1. Click on the link provided above - this will take you to the shared History (in case of permission error, please email the Instructor immediately)
2. Click on “Import this history” (see below). The History is now duplicated, with one copy in your own account.


Galaxy Europe
Workflow Visualize Shared Data Help User
Using 3%
Import this history
search datasets
bams
12.9 GB
4: MTG569_align
3: MTG558_align
2: MTG23_align
1: MTG7_align
About this History
bams
Author
tmarolliey
Related Pages
All published Histories.
Published Histories by tmarolliey.


Galaxy Europe
Workflow Visualize Shared Data Help User
Using 4%
Tools
search tools
Upload Data
Get Data
Send Data
Collection Operations
GENERAL TEXT TOOLS
Text Manipulation
Convert Formats
Filter and Sort
Join, Subtract and Group
GENOMIC FILE MANIPULATION
Convert Formats
FASTA/FASTQ
Quality Control
SAM/BAM
BED
VCF/BCF
Nanopore
COMMON GENOMICS TOOLS
Operate on Genomic Intervals
Fetch Sequences / Alignments
GENOMICS ANALYSIS
Paper Alert!
Scalable, accessible, and reproducible reference genome assembly and evaluation in Galaxy
Larivière, D. et al. 2023
READ NOW!
"Anyone, anywhere in the world should have free, unhindered access to not just my research, but to the research of every great and enquiring mind across the spectrum of human understanding." – Prof. Stephen Hawking
News
[GTN news] Feedback: Easy slide recordings
[GTN news] GTN has left Twitter/X as of October
Events
Dec 11 - Dec 15 2nd BioHackathon Germany
The 2nd BioHackaton Germany will take paice in Bielefeld and has 11 exciting projects.
Mar 4 - Mar 8 Workshop on High-Throughput Data Analysis with Galaxy
History
search datasets
Copy of 'bams'
12.9 GB
4: MTG569_align
3: MTG558_align
2: MTG23_align
1: MTG7_align
https://www.biorxiv.org/content/10.1101/2023.06.28.546576v1.full

GALAXY ANALYSIS TUTORIAL:

- FastQC

FastQC offers several options. Here are the ones you need to parameter:

- ☐ “Raw read data from your current history”: one FASTQ file you have been downloaded in your History
- ☐ “Contaminant list”: made available by FastQC - can be found in the shared history
- ☐ “Adapter list”: made available by FastQC - can be found in the shared history

Then, click on Run Tool.

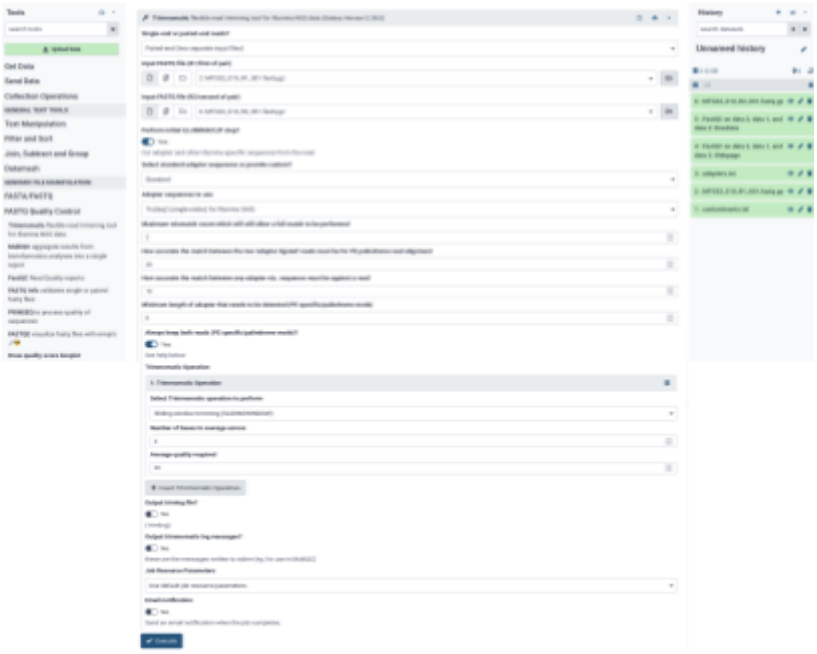
Trimmomatic

Trimmomatic offers several options. Here are the ones you need to parameter:

- ☐ “Single-end or paired-end reads?”: Choose paired-end.
- ☐ “Input FASTQ file(R1/first of pair)”: Select R1 Fastq file from your history.

- ☐ “Input FASTQ file(R2/first of pair)”: Select R2 Fastq file from your history.
- ☐ “Perform initial ILLUMINACLIP step?”: Yes
- ☐ “Adapter sequences to use” :”TrueSeq3 (paired-ended, for MiSeq and HiSeq)

Then, click on Run Tool.



Map with BWA-MEM

Map with BWA-mem offers several options. Here are the ones you need to parameter:

- ☐ “Will you select a reference genome from your history or use a built-in index?”: Select “Use a genome from history and build index”.
- ☐ “Use the following dataset as the reference sequence”: Select the reference genome uploaded in the shared history.
- ☐ ”Single or Paired-end reads”: Select “Paired”.
- ☐ “Select first set of reads”: Use the trimmed R1 FASTQ (R1 Paired). Should be in your history.
- ☐ “Select second set of reads”: Use the trimmed R2 FASTQ (R2 Paired). Should be in your history.
- ☐ “Set read groups information” : Select “Set read groups (Picard style)
- ☐ “Read group identifier (ID): Enter the name of your strain (i.e. BC986)
- ☐ “Read group sample name (SM): Enter the name of your strain (i.e. BC986)
- ☐ “Auto-assign”: Yes
- ☐ “Select analysis mode”: Select “1.Simple Illumina mode”.
- ☐ “Bam sorting mode”: Select “Sort by chromosomal coordinates”.

Then, click on Run Tool.

MarkDuplicates

MarkDuplicates offers several options. Here are the ones you need to parameter:

- ☐ “Select SAM/BAM dataset or dataset collection”: Use the Bam file created by BWA-mem. Should be in your history.
- ☐ “If true do not write duplicates to the output file instead of writing them with appropriate flags set”: Select “No”.
- ☐ “Assume the input file is already sorted”: Select “Yes”.

Then, click on Run Tool.

Delly call

Delly call offers several options. Here are the parameters you need to set:

- ☐ “Select input files(s)”: Use the BAM file created by MarkDuplicates. Should be in your history.
- ☐ “Select type(s) of structural variants to detect”: Select “All types”.
- ☐ “Select genome file”: Use the same reference genome as before from your history.
- ☐ “Select output files(s)”: Select VCF.

Then, click on Run Tool.

Basil

Basil offers several options. Here are the ones you need to parameters you need to set:

- ☐ “Reference Sequence File”: Use the same reference genome as before. Should be in your history.
- ☐ “Alignment File”: Use the BAM file created by MarkDuplicate. Should be in your history.
- ☐ “Minimum supporting reads, each side”: Increase to 5.

Then, click on Run Tool.

Lumpy

Preprocessing

LUMPY preprocessing offers several options. Here are the ones you need to parameters you need to set:

- ☐ “BAM input dataset”: Use the BAM file created by MarkDuplicates. Should be in your history.

- ☐ “Duplicate detection and removal”: Select “Remove duplicates marked in input data”.

Then, click on Run Tool.

Calling

LUMPY offers several options. Here are the ones you need to set parameters you need to set:

- ☐ “BAM input dataset”: Use the BAM file created by MarkDuplicates. Should be in your history.
- ☐ “Discordant reads input”: Use the discordant file created by LUMPY preprocessing. Should be in your history.
- ☐ “Split alignment input”: Use the split file created by LUMPY preprocessing. Should be in your history.

Then, click on Run Tool.

More details on Galaxy History and how to use it

Figure 3

Tip: Share History to share files on Galaxy

History Options > Share or Publish > Select ‘Make History accessible’ > click on ‘Share History with individual users’ > Add email addresses > Click on ‘Save’ > Send the link

High Performance Computing system - HPC called TALC

Dr. Tatiana Maroilley

A HPC is several computers put together to increase computational performance (memory and computational resources). You access an HPC system by connecting to the main server through your own machine. At UCalgary, you have access to TALC, a HPC for students. Once connected, everything you do is done on TALC, not on your computer.

To run an analysis, you will have to create scripts (text files with options and command lines) and submit them to the scheduler. The scheduler is the program that attributes the resources of the cluster to the different jobs that have been submitted by different users - prioritize some jobs over others based on available resources.

To access TALC, you need to be either physically on campus, or use a VPN called Forticlient. See here the Wiki page put together by UCalgary to install the VPN and connect to TALC: https://rcs.ucalgary.ca/Connecting_to_RCS_HPC_Systems

To install the VPN, floor the instructions at the section “Connecting using FortiClient VPN”. If that does not work, or you do not want to install the VPN, you can also access TALC through your browser: https://generalconnect.ucalgary.ca:10443/remote/login > Clik on Single Sign-On and sign in with your UCalgary credentials. Then “Quick connection” > “SSH”. In “Host” type: talc.ucalgary.ca > Launch. Then enter tour UCalgary credential (again, and usually it doe not work the first type - just do it again) and click (with the mouse - press Enter does not do the trick) on “Login”. You will be then connected to TALC.

Our shared directory is /work/TALC/mdsc301_2024w. You can do your analysis in your home directory (where you got connected originally). For convenience, BAM files have been produced and copied on TALC: /work/TALC/mdsc301_2024w/data/bams/.

Because you have already used Galaxy to get the bam files, we will start from there and run only SV callers. We will be running different ones than the ones on Galaxy, to extend the possibilities of comparison. **Manta** (https://github.com/Illumina/manta) and **seekSV** (https://github.com/qiukunlong/seeksv) have been installed and tested. The templates of scripts to run them are available in /work/TALC/mdsc301_2024w/scripts/.

If you have never use command lines in a UNIX/Linux system or if you need to refresh your memory, have a look at the resources on D2l > Linux

If you have a Windows machine, install MobaXterm: https://mobaxterm.mobatek.net/download-home-edition.html. Choose the portable edition.

Once open, click on “Start a local terminal”. To connect to TALC, type in the terminal user.name@talc.ucalgary.ca (user.name being your UCalgary username). Then you will be asked to enter your UCalgary password - type it (nothing will appear, that is ok) and press Enter. You are in! That should work if you are on Campus. If not, you will need to start the VPN.

Script Template

```
#!/bin/bash
#SBATCH --job-name=jobname
#SBATCH --cpus-per-task=12
#SBATCH --mem=50000
#SBATCH --time=24:00:00
#SBATCH --mail-user=your.name@ucalgary.ca
#SBATCH --mail-type=END
#SBATCH --output=jobname.out
#SBATCH --error=jobname.err
```

TALC Tutorial

T. Maroilley

1. Connect to TALC (from campus - if outside campus, please refer to how to use the VPN)
    a. Open a terminal (MobaXterm in Window or console in MAC)
    b. Type:

```
ssh -X YOUR_UofC_USERNAME@talc.ucalgary.ca
```

    c. Press Enter
    d. It will ask for your password
    e. Type your password (nothing will appear) and press Enter
2. You are in!
3. The first time you connect, you have to read the guidelines and agree to it.
4. You are in TALC in your /home. (/home/USERNAME/)
5. Look at what is there (should be empty) - type and press Enter:

```
ls
```

6. Here, create a folder from the first tools you want to run (manta)

```
mkdir manta
```

7. Check that the folder has been created by listing what is there once again

```
ls
```

8. Go into the folder manta/ that you have created and check what is there (should be empty)

```
cd manta
```

```
ls
```

9. Copy the script from the /work/TALC/mdsc301_2024w/scripts folder (the command for copy is cp, and it takes two arguments or parameters in that order: what do you ant to copy and where to you want to copy it)

```
cp /work/TALC/mdsc301_2024w/scripts/manta.sh /home/YOUR_USERNAME/
```

10. Check that the script was copied by listing what is there

```
ls
```

11. Modify the script with your own information **(everything between <> should be updated - <> have to be removed!!**)
    a. Open the script with the editor called nano through command line if you are on MAC, or with the editor in MobaXterm if you are on Windows

```
nano manta.sh
```

```
#!/bin/bash
#SBATCH --time=00-06:00         #time (DD-HH:MM)
#SBATCH --job-name=manta
#SBATCH --output=job.manta.out
#SBATCH --error=job.manta.err
#SBATCH --mail-user=<username>@ucalgary.ca
#SBATCH --mail-type=FAIL
#SBATCH --mail-type=BEGIN
#SBATCH --mail-type=END
#SBATCH --ntasks=4              #number of mpi processes
#SBATCH --mem-per-cpu=4G        #memory; default unit is megabytes
#SBATCH --cpus-per-task=1
```

This first part is the options you choose to give SLURM, the scheduler: how much memory, how much time you estimate that toll will need.

You are also defining the name of two files: manta.err - where all error messages from manta will end up, and manta.out - where every other message that manta is creating will go. Those files are useful when your job fails and you need to investigate why.

In --mail-user=EMAIL, change EMAIL by your ucalgary email address. The option --mail-type=END,FAIL will make TALC send you an email if your job has been successfully run or if it failed.

Each tool has requirements and you can find the list in the description of the tool: https://github.com/Illumina/manta/blob/master/docs/userGuide/README.md.

```
cd
mkdir manta
cd manta
```

Those three lines are to make sure that you will be launching manta from the right place: /home/USERNAME/manta. cd takes you back to your home/, mkdir create the manta folder (that you have already so it won't do it again), and then jump into manta/

```
# step1
/work/TALC/mdsc301_2024w/tools/manta/bin/configManta.py \
--bam /work/TALC/mdsc301_2024w/data/bams/<YOUR_BAM_FILE>.bam \
--referenceFasta /work/TALCmdsc301_2024w/references/c_elegans.PRJNA13758.WS265.genomic.fa \
--runDir ./

# step2
python runWorkflow.py
```

This line is how we get manta to run. I will let you have a look at manta documentation to understand each option (https://github.com/Illumina/manta/blob/master/docs/userGuide/README.md). You will have to change YOUR_BAM_FILE.bam by the complete path to the bam file which is: /work/TALC/mdsc301_2024w/data/bams/YOURSTRAIN.bam

12. Ctrl-X to close the file with nano + Yes (to the question do you want to save the changes) + Enter (to save under the same name)
13. Once the script is ready, make sure that your are in the folder where the script is and the content of your script:

```
ls
pwd
more your_script.sh
```

14. Then launch the script

```
sbatch manta.sh
```

15. Check if it is in the queue

```
squeue
```

16. If yes, good! But check its status (5th column, called "ST" - R is for running, PD for Pending - anything else, check the SLURM documentation) If not, you made a mistake - go check the manta.err and manta.out to see where the problem is. Fix it (probably a typo in the script) and relaunch the script with `sbatch`.

Manta produces a lot of files. Look at the documentation to figure out which VCF to use (ask for help if necessary).

Follow the same tutorial for seekSV (https://github.com/qiukunlong/seeksv).

How to get the VCF out of TALC to your computer

Optional: In a terminal connected in TALC, unzip the VCF (vcf.gz): gunzip XXX.vcf.gz. This will create a XXX.vcf file - uncompressed and readable.

From a terminal (not connected to TALC), type:

```
scp
YOUR_UofC_USERNAME@talc.ucalgary.ca:/home/USERNAME/PATH/TO/THE/FILE/YOU/WANT.vcf /PATH/ON/YOUR/COMPUTER/WHERE/YOU/WANT/THE/FILE
```

Week 5 - Visualization (IGV) and Validation (PCR)

Deadlines Week 5: Quiz Galaxy (D2L) Feb 05 + Report 2 (ungraded - individual) Feb. 09

Requirements for Week 5

- ☐ **Install IGV (https://software.broadinstitute.org/software/igv/download)**

Optional readings

- ☐ IGV tutorial: https://training.galaxyproject.org/training-material/topics/introduction/tutorials/igv-introduction/tutorial.html

Timeline for Week 5

This week is dedicated to filtration of the VCFs, interpretation of the SVs and validation by PCR. One student per team will run a PCR of the SV you have picked (already published and validated). The primers have been designed and ordered. The other students will work on the VCFs to evaluate if:

- The tools were able to detect, and with which accuracy (right position and right characterisation - DEL, DUP…), the SV you picked
- The tools were able to report anything unique in your strain, when compared to the control (N2)
- Which tool seem to perform better/worse

Monday, February 05, 2024

- 10:30-10:45 - Quiz: 10 questions on “Galaxy”
  - The quiz starts at 10:30am and ends at 10:45am. Do not be late! You won’t have extra time.
- Lecture (~20 min) “How to interpret and filter a VCF?” (Rumika)
- Lecture (~20 min) “How to validate a SV with a PCR?” (Sue)
- Filtration of SVs reported in VCFs

Wednesday, February 07, 2024

- Filtration of SVs reported in VCFs
- PCR

Filtering and Interpreting VCFs by data visualization on IGV

Dr. Tatiana Maroilley

Filtering

1. Quality: For each tool, discard all calls with low quality flags
2. Uniqueness: For each tool, remove from the VCF all coordinates also reported in control (N2)
3. Combine the remaining calls from all tools - but keep this information regarding which tool reported which SV.

Note: calls reported by multiple tools have potentially more probability to be a true positive but a tool could outperform the others and report something missed by all other tools.

Is any tool reporting the SV you picked? Accurately?

Manual curation

Check each position on IGV on both your strain and control (N2) and rule:

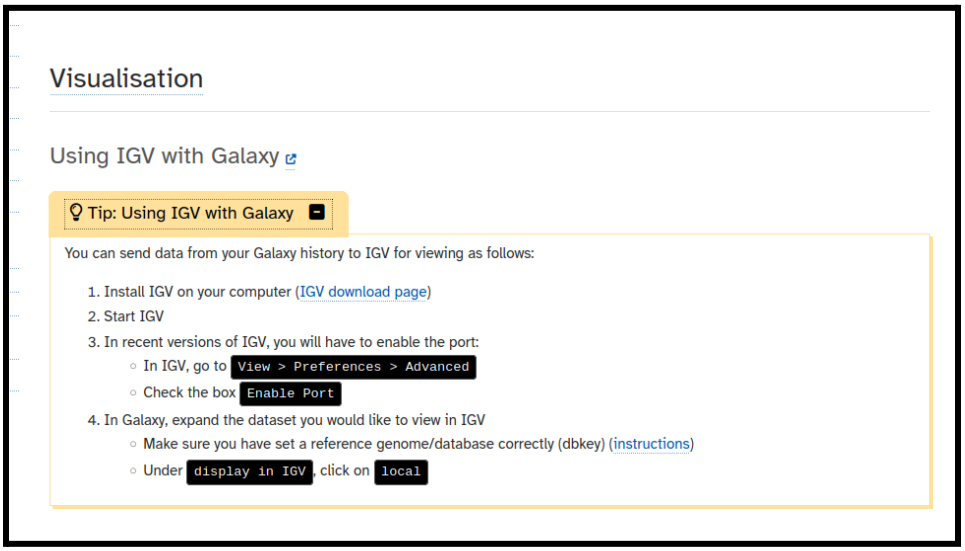
1. Do the reads seem to show the signature of a SV at this position? If not, discard it.
2. Does the alignment of the reads look similar in control? If yes, you can discard it.
3. Keep a screenshot of SVs unique and of good quality
4. Does it seem that the position reported by the tool is the exact position of the breakpoint?
5. What kind of SV can it be? Did any tool characterize it well?

Tip: Each team mate could do at least one chromosome.

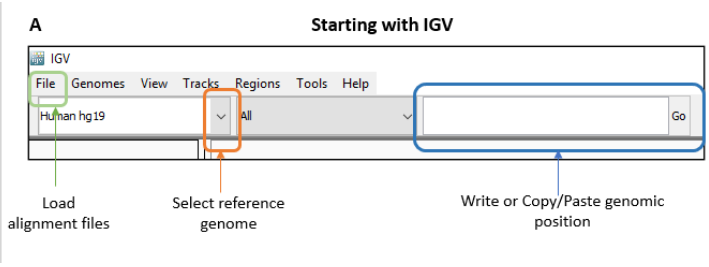
Tips to read a VCF

- ☐ You can copy/paste the content of the VCF in an Excel sheet. This will facilitate the manual curation of the results.
- ☐ The first columns give you the position of the beginning of the SV. The END is usually in the eighth column (INFO).
- ☐ Most calling tools have their own filters and thresholds. It is recommended to filter out and ignore SVs that do not pass those internal filters - look at column FILTER of the VCF and only keep the calls that “PASS”. Everything else should be ignored (e.g., LOW QUAL).
- ☐ Visualize SVs unique to a balancer strain. Anything also present in VCFs from control can be ignored.

The BAM files should be available in your Galaxy History. Follow the steps in the screenshot below to open them with the IGV installed on your machine.



Looking at SV on IGV



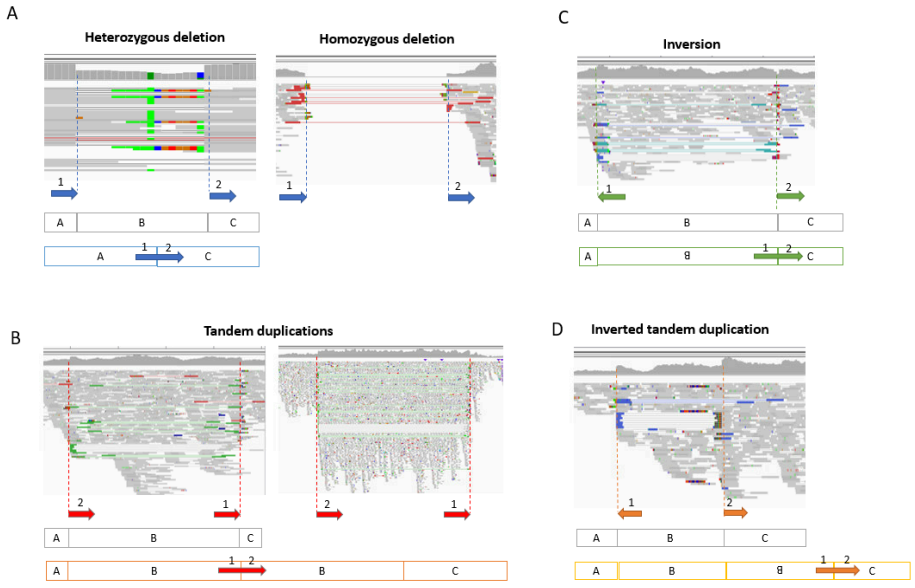
On the main window of IGV, you might need to change the Reference Genome. Click on the drop-down menu in the top left corner to change the reference genome to the species you are analyzing. Then upload the alignment files or bam files (previously downloaded from your Galaxy history). Note that IGV needs to index files that can be downloaded along with the bam files from your Galaxy history. The indexes must be stored in the same folder as the alignment files. To load your alignment files, click on “File” in the bar Menu on the top left corner of the main IGV window. This will open a drop-down menu. Choose “Load from file” and select your bam files. Note that IGV allows you to open several alignment files at the same files, facilitating comparison between samples. Load your sample of interest along with at least N2 and Hawaiian strains.

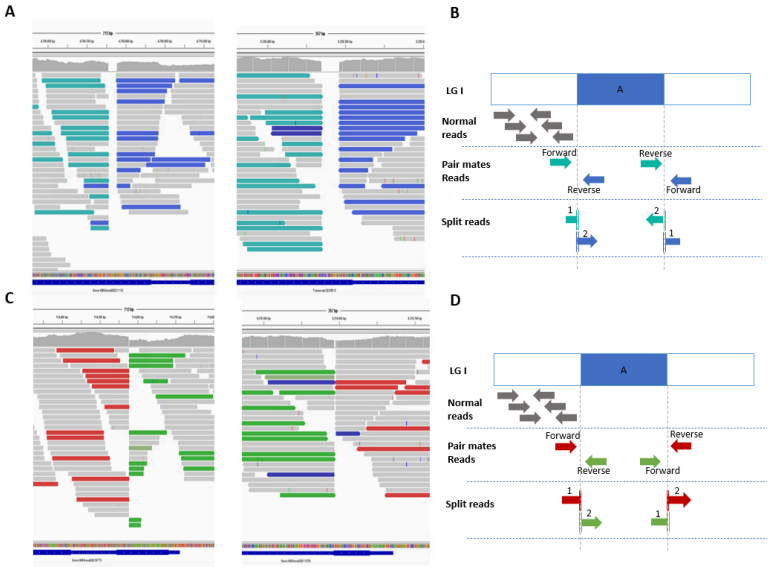
At the bottom of the main window, the reference genome is displayed (genes structures and sequences). In brief, samples will be displayed in two tracks. The upper track represents a histogram of the coverage (the number of reads aligned at this position). The lower track is the visualization of the reads themselves. They are represented by arrows (in different directions for paired-end libraries, depending on if they are forward or reverse reads). By default, IGV colors them in grey. If they display specific characteristics, they will be colored differently. When a read carries a base different from the reference sequence, it is written on the read. In addition, when the alternate allele is A, bases are colored in green, T in red, G in orange and C

in blue. When several reads are carrying an alternative allele, the histogram on the top track will display a color (same base color code), correlated with the percentage of those reads compared to the coverage.

To visually assess variants, write copy/paste the genomic location of a variant in the text box in the middle on the upper part of the main window and click on “Go” (or press Enter on your keyboard). See below the signature of split reads at breakpoints you could observe and how to interpret them.

See below how to interpret IGV color code:





A and B show reads colored by IGV in blue because pair mates reads have been aligned in the same direction, when they should be in opposite directions (possible inversion). C and D show reads colored by IGV in red because pair mates reads are closer to each other than usual (possible deletion) and in green because the reverse is aligned before the forward (possible tandem duplication).

PCR protocol

By Suzanne Ferris

PCR Preparation

- In a 1.5 mL microcentrifuge tube add the PCR reagents in the following order: Nuclease-free water, 5X EZ PCR Master Mix, primer stock (contains forward and reverse primers) (10µM), and worm lysate (DNA template) (0.6µL).

- PCR reagent quantities for a 1X 20µL reaction:

PCR Reagents	1X (µL)
Nuclease Free Water	15
5X EZ PCR Master Mix	4
Primer (10uM)	0.4
Worm lysate (DNA template)	0.6
Worm lysate	0.6
Total:	20 µL/tube

- Aliquot 20µL of the sample into 0.2mL PCR tubes
    - Make sure to have a tube for mutant sample, wildtype sample, and a no template control
- Centrifuge tubes to ensure proper mixing of sample and removal of bubbles
- Place tubes into PCR thermocycler and run the overall program to amplify sample:

Steps		Temperature	Time
Initial Denaturation		95˚C	3 minutes
Denaturation	**Repeat steps for 30-35 cycles**	95˚C	30 seconds
Primer Annealing		48-68˚C*	30 seconds
Extension		72˚C	1 minute per kb
Final Extension		72˚C	5 minutes
Hold at Low Temperature		4-10˚C	

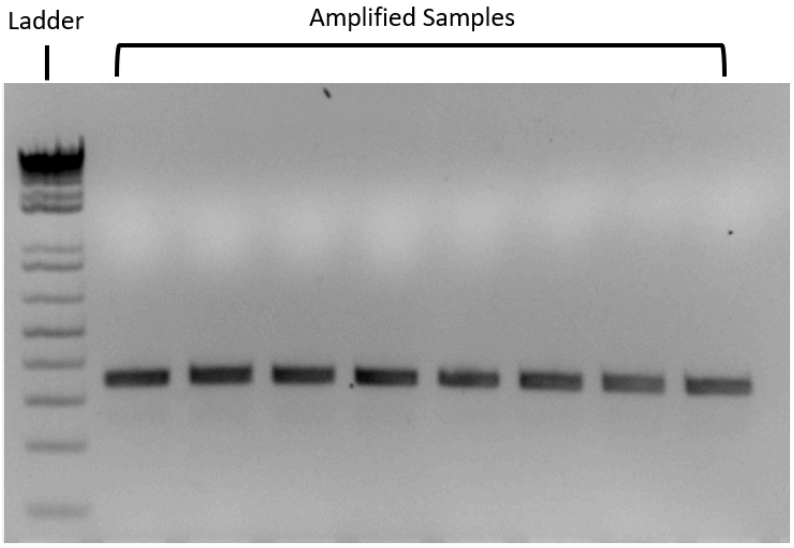
*temperature depends on the annealing temperature of your primers

Gel Electrophoresis Visualization

- Make 2% agarose gel

  - In a 250 mL Erlenmeyer flask, add 1 SeeGreen all-in-one agarose tablet with 20 mL of distilled water and stir until fully dissolved.
  - Heat solution in microwave until solution is clear (roughly 30 seconds)
  - Let solution cool until flask is warm enough to touch
  - Pour solution into mold and let cool for ~20 minutes to solidify
- Place gel in electrophoresis apparatus and fill with buffer until wells are full
  - The end of the gel with the wells must be placed towards the negative electrode.
- Load 5μL of sample and 3μL of 1kb ladder into each well
- Run gel in apparatus for 15-20 minutes
- Visualize the gel under blue light and take an image

Example of gel picture:



Manuscript Template

By Victoria R. A. Barbosa

Throughout this template, you will find the *requirements* for this Assignment. The requirements are meant to be the core of your manuscript. You can expand your manuscript to other results or discussion points.

The manuscript should have between **10-20 pages, double spaced (except references) and figures included,** and contain a detailed explanation of your whole project. Generally speaking, it will follow the same structure and sections of the research proposal, but this time you can be more detailed and report everything you have done. Also, the idea behind the research proposal is to briefly explain your project plan and convince the reader that it is feasible and significant. However, the idea behind the manuscript is to communicate your core findings and the innovative methods you applied to get to that result.

Table 1 Recommended fonts and sizes.

Style name	Brief description
Article Title	**16 pt, bold**
Author Names	**12 pt, bold**
Author Affiliations	10 pt
Abstract	10 pt
Keywords	10 pt
Heading 1	**12 Pt, bold**
Heading 2	*12 pt, italic*
Heading 3	*11 pt, italic*
Body Text	12 pt
Figure caption	10 pt
Table caption	10 pt

Formatting: use Calibri, size 12, double spacing, justified.

The aims of this Assignment are for you to summarize your findings on the balancer strain you have selected and reflect on the pros/cons of using Galaxy vs. HPC (TALC) system. Each aim will be crucial for the design and the implementation of your RNA-Seq project.

- ➔ Use the template provided below.
- ➔ The manuscript is due on **Friday, March 1st, at 11:59pm.**
- ➔ **Please upload one manuscript per group in D2L.**
- ➔ Accepted format: pdf.
- ➔ **Authors should be listed under the title.**
- ➔ Work in a team. Every team member should provide input, comments, suggestions and approve the final version, as it is mandatory in collaboration in research.
- ➔ If you divide the workload (one does introduction, one does methods…), make sure that one of you is in charge of checking that all parts align.

Grading Rubric for Manuscript:

Total Points: ____/20

Criteria	0 Point	1 Point	2 Points	3 Points
Length (2 points)	Significantly over or under the 10-20-page requirement.	Slightly over or under the 10-20-page requirement.	Precisely meets the 10-20-page requirement.	
Meaningful Title (1 point)	Title is unclear, lacks meaning, or does not effectively convey the focus.	Title is clear, meaningful, and effectively conveys focus.		
Introduction (3 points)	Introduction lacks coverage of key project points or is missing key points.	Covers main points but may lack depth or context.	Comprehensive and mostly clear, provides essential context.	Comprehensive and exceptionally clear, providing essential context. Thoroughly covers all main points with exceptional depth and clarity.
Results (3 points)	Unclear, disorganized, and lacking key findings.	Somewhat clear but not concise; key findings mentioned but not effectively highlighted.	Clear and concise presentation of results.	Exceptionally clear, concise, and to the point.
Discussion (3 points)	Lacks detail, organization, and reference support.	Somewhat detailed but lacks reference support.	Detailed discussion with some reference support.	Detailed, well-supported discussion with relevant references.
Methods (3 points)	Incomplete, lacking necessary details, and unclear.	Somewhat clear but lacks necessary details.	Methods cover essential aspects but may lack detail.	Comprehensive coverage of all methods with clear explanation.
References (2 points)	References are not correctly formatted, significant issues.	References are correctly formatted but may have issues.	References are correctly formatted, organized, and follow the required citation style.	
Clarity and Quality of Writing (3 points)	Lacks clarity and has significant issues in communication.	Somewhat clear with minor issues in communication.	Clear and effective communication.	Writing is exceptionally clear and of high quality, enhancing content understanding.

Sample manuscript for MDSC301 (title)

First Author,[a] Second Author,[a] Third Author,[b] Fourth Author[a,b,*]

[a] University Name, Faculty Group, Department, Street Address, City, Country, Postal Code

[b] Company Name, Street Address, City, Country, Postal Code

Introduction -

"These are the information and context that led to my project and are necessary to understand it"

Here, give all the background information relevant to your project and even previous published works that might have led to the research question you are investigating. You should spend up to 1-2 pages on it, but remember: this should **not** be your longest section. At the end, dedicate a paragraph to outline aims and highlights of your project before moving to the next section.

Requirements:

- ☐ *Importance of SVs?*
- ☐ *Literature review on detection of SV - challenges and opportunities?*
- ☐ *Why use C. elegans?*
- ☐ *Last paragraph must summarize your study and highlights*

Results

"This is what I found"

Here, you will describe all your findings in 5-10 pages. Organize your findings, and explain them in detail, one at a time, into subsections. Use subsection headings to help the reader (subsection headers should be italicized, as exemplified in the Methods section of this template). It is also the section where to put figures and tables (see examples below)… Just don't forget to write the captions! These items could be close to the part of the text in which they are mentioned and should follow a numeric order as they are called out (Figure 1, Figure 2…).

Figure 1 (bold): *C. elegans.* P.S.: caption for figures should be placed under the image, font size: 10.

If you are working with a table instead, the captions should be above it, also with font size 10 (Table 1).

Table 1: Common *C. elegans* strains and isolation site.

STRAIN	ISOLATION SITE
CB4856	Hawaii
KR314	Canada
AB1	Australia

As seen in Fig. 2, you can even have a figure with multiple items that should be labeled with letters (following alphabetic order) and explain all of them in the caption and text.

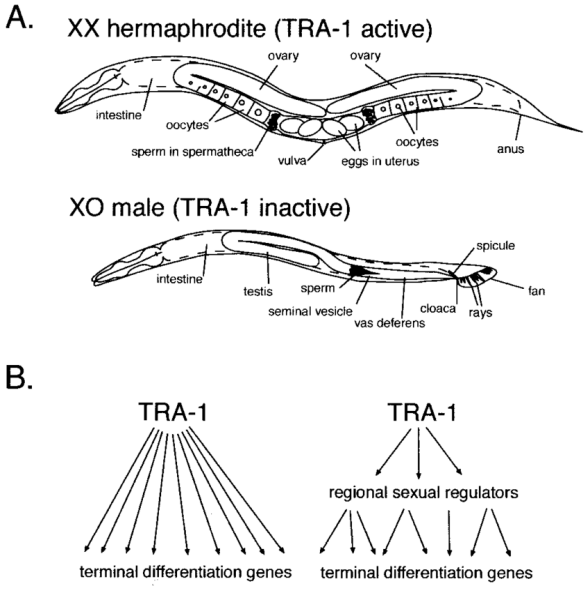


Figure 2: Sexual dimorphism in *C. elegans.* (A) depicts morphological differences between males and hermaphrodites. (B) shows potential TRA-1 functions on sexual determination. Adapted from: Harley, 2002.

Remember, all Figures and Tables should serve a specific purpose and should be referred to in the main text.

Requirements:

- ☐ *Section regarding the quality of the data*
  - ☐ *Report, comment and explain*
    - ☐ *The number of reads (or sequences) before/after trimming and after alignment*
    - ☐ *The average length of the reads before/after trimming*
    - ☐ *The number of k-mers before/after trimming*
    - ☐ *The quality of the reads before/after trimming (per base)*
  - ☐ *Figure 1: Control quality of the data (FastQC)*
    - ☐ *Panel of screenshots of graphs from FastQC before/after trimming for your strain (no need to report QC of N2)*
  - ☐ *Table 1: Size of the FASTQ files, Number of reads in FASTQs before/after trimming Size of the BAM files, Number of reads aligned on the genome, Size of the reads before/after trimming*
- ☐ *Section regarding your SV of interest*
  - ☐ *Which tool has reported it? With which accuracy? With proper characterization? Which tool missed it?*
  - ☐ *Figure 2: Screenshots of IGV of your SV of interest and PCR gel to confirm it*
- ☐ *Section regarding additional SVs you might identify*
  - ☐ *Figure 3: Screenshots of IGV of any interesting additional SV (pick a few, no need to show them all)*
  - ☐ *Table 2: Number of calls of each tool used either on Galaxy or HPC, number of calls of "good quality" calls after visual inspection, Number of good SVs previously reported in Maroilley et al. 2023, Number of good SVs not previously reported*
- ☐ *Section comparing each tool*
  - ☐ *Comment on the outcome of each tool: Is there a tool that seems more accurate? More sensitive? More precise in defining the structure? More precise in finding the right genomic position? Is there a tool that has a better/worse balance of true positives/false positives? Is there a tool that is excellent/terrible at giving the right structure? Is there a tool for which VCF structure is easier/more difficult to manually analyze?*

- ☐ *Figure 4: Venn diagrams comparing calls from different tools (number of calls before any filtering, after filtering for quality and presence in control, after visualization on IGV) - you can use this tool: https://jvenn.toulouse.inrae.fr/app/example.html (remember to cite it)*

Discussion

"This is what it means"

In this section, you will reflect on the results you found and explain how that relates to your initial aim (that you mentioned in the intro) and if the results you got were aligned to what you were expecting (hypothesis) or if you actually ended up with some different findings. Here, you do not use figures or tables and will just argue on the ideas you presented. Spend up to 1-2 pages on this section. Organize your discussion: one paragraph for each idea. Try to be concise but clear. You can still add references to the literature here. Discuss limitations of your methods that could impact the interpretation of your results.

Requirements:

- ☐ *Discuss the performances of the different tools you have tested - which one performed better/worst and hypothesis on why*
- ☐ *Discuss pros/cons of using Galaxy vs. a HPC system*
- ☐ *Discuss possible impact of your SV of interest on transcriptome and protein and how you would test your hypothesis*
- ☐ *Discuss performance of short reads in detecting SV and what could be done to improve it*

Conclusion

"This is why it is important"

Finally, you will briefly (1 paragraph) wrap up your work by mentioning its relevance and how it solved your question raised in the intro. You can also point out future directions if those are feasible.

Methods

"This is how I did it"

In this section, list all the steps you took (literally all of them – do not let anything out – this should allow someone to reproduce your work). Try to follow a logical order and even organize it in subsections (let's say, one for each tool you used, or for each stage of the project). For example:

Pre-processing - FastQ files were submitted to quality check using… + trimming

Alignment - Reads were aligned with… on the reference genome…

Calling - SVs were called with… (*Briefly describe how each caller works)* The VCFs were filtered as follows:...

Validation - PCR protocol

References

"This is where I got information from"

Last, but not least, you will add all the material you have cited in the document, organized by alphabetical order (Vancouver style). Use a Reference Manager again (Mendeley, Zotero, etc.). These are life savers! See below a tutorial for Mendeley. Each tool must be cited at least once (including Galaxy).

Other important tips:

- At the first occurrence of an acronym, spell it out followed by the acronym in parentheses, e.g., whole genome sequencing (WGS). After that, you can refer to that term only by the acronym.
- Remember the writing tips from the Research Proposal:
- Avoid unnecessary abbreviations and terms that are only understood by specialists in that area. Consider as if someone with little Bioinformatics experience would read your text, for example.
- Keep it short: often, fewer words are better. Write short and clear sentences with no more than 20 words each. One idea, one sentence.
- Avoid writing sentences in the third person. Instead of "fluorescence microscopy will be applied for data analysis that will be conducted by the lab technician…" say "the lab technician will analyze the data using fluorescence microscopy."
- Write a single idea in each paragraph. Avoiding too much information in the same paragraph will make the text – and the idea – flow much better.
- Review your text for spell-checking (do not hesitate to use Grammarly!) and do not hesitate to ask for someone else's help. Sometimes, we read our own words so often that we end up missing a detail or two. The point of view of another person might help with clarity, flow and even with catching those minor typos that our own eyes miss.
- Work as a team, like in a real lab: everyone should read, comment, and approve the final version of the proposal.
- Check other tips at:

https://www.schulich.uwo.ca/biochem/research/docs/nserc_writing_guide_pres.pdf

Mendeley 101 - Tutorial

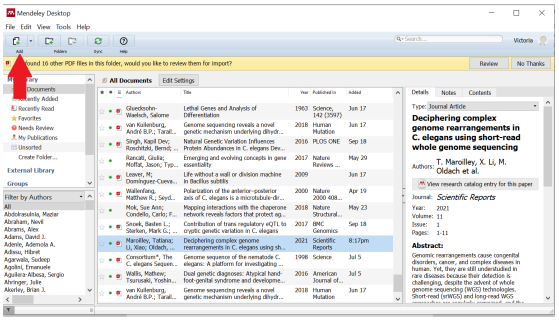
By Victoria R. A. Barbosa

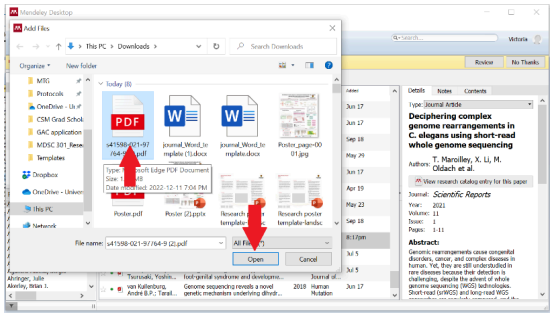
- Go to https://www.mendeley.com/download-reference-manager/windows
- Download the appropriate version (Windows, Mac or Linux)
- Follow the instructions for signing up

You will first have to add the paper in the Mendeley app, so you can then use it in any file on word. To do so, you have 2 options:

Directly through the app:

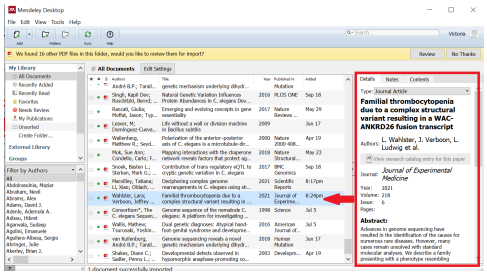
1. Download the paper pdf version
2. Open the desktop app and go to the “Add” option (top left)
3. Select the desired pdf
4. All set! It should appear on your list and a preview can be seen on the right





Using the Web Importer

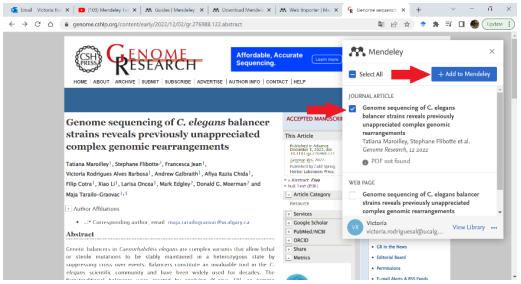
Or, you can count with a Google Chrome extension (very practical) that allows you to save the paper into your Mendeley account without having to download it (or even having the pdf open, just the web version will do). To do so, follow this link

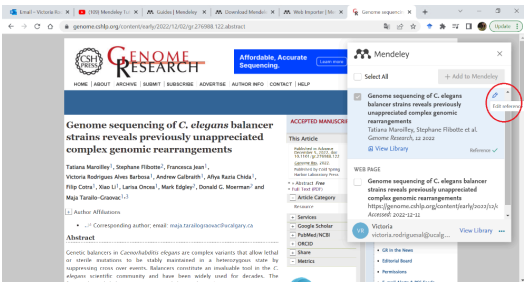


https://www.mendeley.com/reference-management/web-importer to download the Web Importer.

1. Once installed, the icon should appear on your navigator extensions (top right)

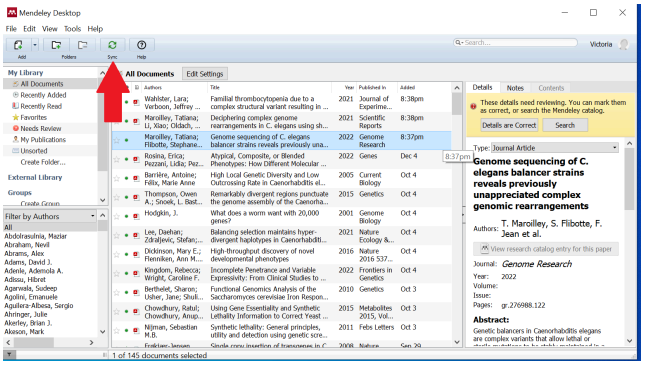
2. Found an interesting paper you need to cite? Simply click on the icon and open Mendeley
3. Log in first if you hadn't done that yet)
4. Select the paper you want (the tool will often recognize other papers and references listed on the same webpage) and click on "Add to Mendeley"





P.S.: Before adding it to Mendeley, I usually open the "edit" option, don't change anything, but just double check if the tool is getting all the right information (title, year of publication, authors, etc)

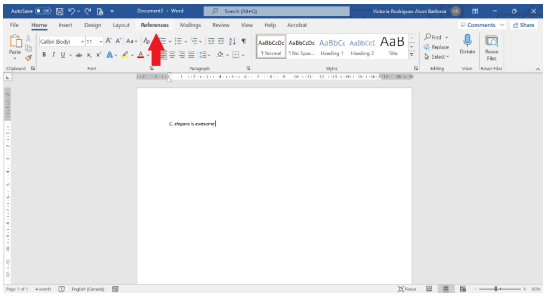
5. Done! Since the Chrome extension also uses your account, which is linked to the app, the paper should automatically appear on the Desktop App as well. This might take a few minutes. If it doesn't show up, go to the "sync" option on your desktop app

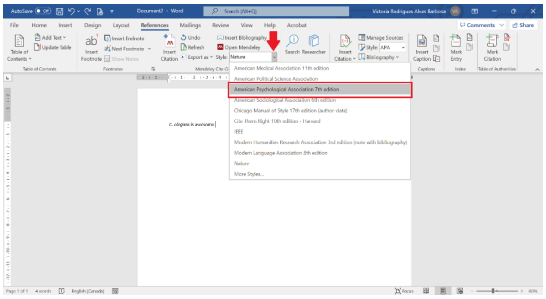


Now comes the fun part: you can use Mendeley while writing your documents on Word and not stress out about citations ever again (JK, it might drive you nuts sometimes, but it still saves a ton of time). HOWEVER, it will only work if you have the Desktop App installed (meaning, Chrome extension is optional, but desktop app is essential for it to work).

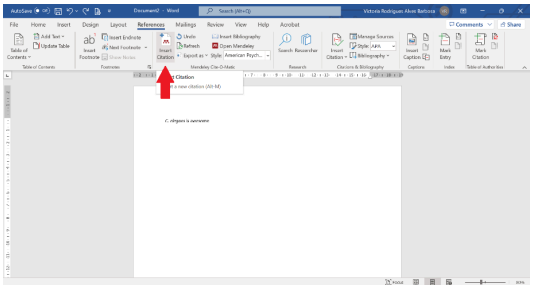
All you have to do is…

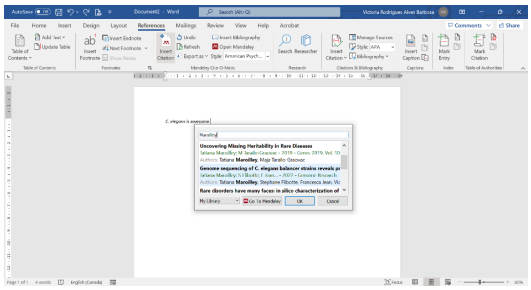
1. On the part of the text you want to add the citation, go to References



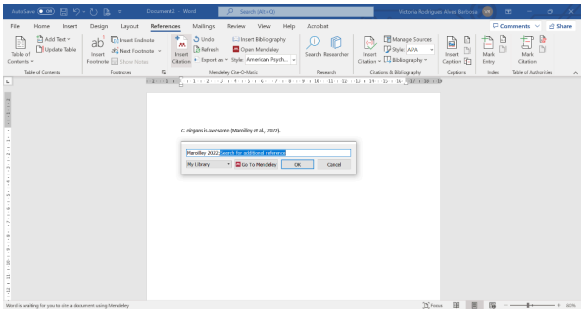


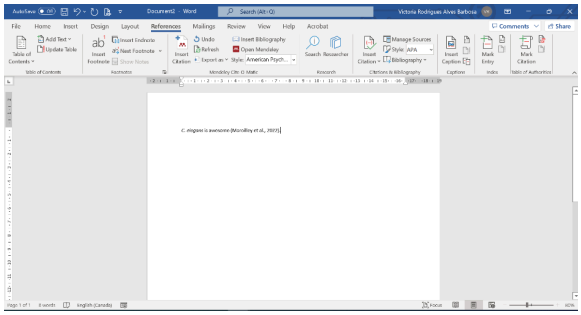
2. For the first time you use it, you might have to select a reference style first. Go for the APA 7th edition:



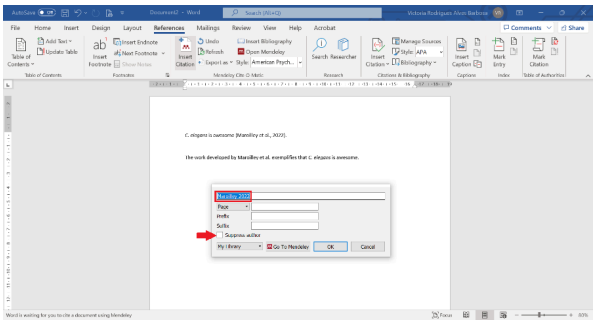


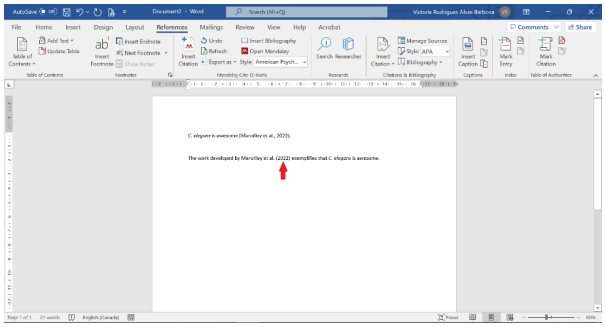
3. Then, click on Insert Citation
4. Find the paper you desire by either looking for the title or the author (usually easier by the author). Note: if you added it to Mendeley but can't find it through Word, try to open Mendeley and re-sync the folder, then press Refresh on Word (right besides the "Insert Citation" icon.
5. Click in Ok. Note: if you are citing more than one paper for the same sentence, just keep searching for the authors, then finally press Ok.
6. After you click on the paper and press Ok, this is what it should look like on text.



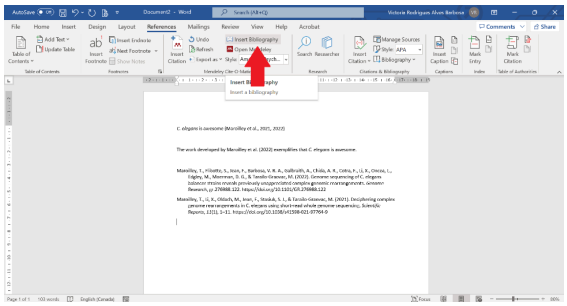


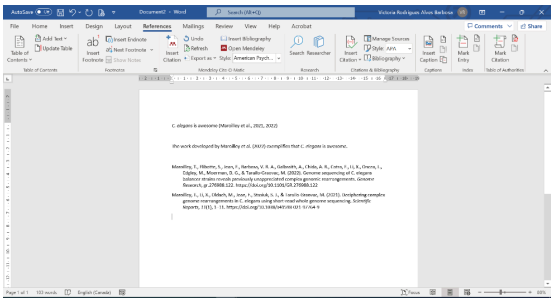
7. If you are doing a direct citation (as exemplified below), you can also use Mendeley to add just the year of the paper (this is necessary because if you just type the citation yourself, Mendeley will not recognize it and it won't add it to the references list at the end). To do so, after you search for the desired paper, click on the author's name, then select Suppress Author and finally click Ok.
8. This is what it will look like:





9. Whenever your document is finished, you can use Mendeley to list all the References in alphabetical order :o Just go to wherever you want to place your references, then click Insert Bibliography. And THAT'S IT! Simple as that. One click. A real lifesaver! But remember, it will only list the citations you added in the text using Mendeley. Any others that were manually added won't be there ;)





Additional resources:

- Introduction to Mendeley: The Complete Guide - https://www.youtube.com/watch?v=YEZ-mkJ770E
- https://www.mendeley.com/guides

Week 6 - Introduction to RNA-Seq

Deadlines Week 6: Quiz Validation (D2L) Feb 12 + Manuscript (Group - for Feedback only) Feb. 14

Reading for Monday, February 12, 2024

Each group member should read one for the list below. Make sure that no one in your group will read the same paper as you will.

- ☐ **Kukurba KR, Montgomery SB. RNA Sequencing and Analysis. Cold Spring Harb Protoc. 2015 Apr 13;2015(11):951-69. doi: 10.1101/pdb.top084970. PMID: 25870306; PMCID: PMC4863231.**
- ☐ **https://hbctraining.github.io/Intro-to-rnaseq-hpc-salmon/lessons/Intro-to-RNAseq.html**
- ☐ **Batut B, van den Beek M, Doyle MA, Soranzo N. RNA-Seq Data Analysis in Galaxy. Methods Mol Biol. 2021;2284:367-392. doi: 10.1007/978-1-0716-1307-8_20. PMID: 33835453.**
- ☐ **Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak MW, Gaffney DJ, Elo LL, Zhang X, Mortazavi A. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016 Jan 26;17:13. doi: 10.1186/s13059-016-0881-8. Erratum in: Genome Biol. 2016;17(1):181. PMID: 26813401; PMCID: PMC4728800.**

Monday, February 12, 2024

- 10:30-10:45 - Quiz: Questions on “Visualisation and Validation”
  - The quiz starts at 10:30am and ends at 10:45am. Do not be late! You won’t have extra time.
- Lecture (~30min): Introduction to RNA-Seq (Shreya)
- Time left: Working on Manuscript

Wednesday, February 14, 2024

- Walking Gallery - Topics: (**Be careful - RNA-Seq IS NOT Single-cell RNA-Seq**)
  - Differences in library preparation in RNA-Seq compared to DNA-Seq
  - Differences in design analysis in RNA-Seq compared to DNA-Seq
  - Main steps of the Bioinformatics pipeline for RNA-Seq data
  - Examples of analyses made possible with RNA-Seq data
  - Examples of tools for RNA-Seq analysis
  - Source of bias and errors in RNA-Seq
  - Different technologies and protocols to study gene expression

Week 8 - Introduction to R

Deadlines Week 8: Quiz RNA-Seq (D2L) Feb 26 + Manuscript (for feedback only) Feb. 26th + Manuscript (Final submission - Group) Mar. 01 + Peer Evaluation 2 (Individual) Mar. 01

Optional readings

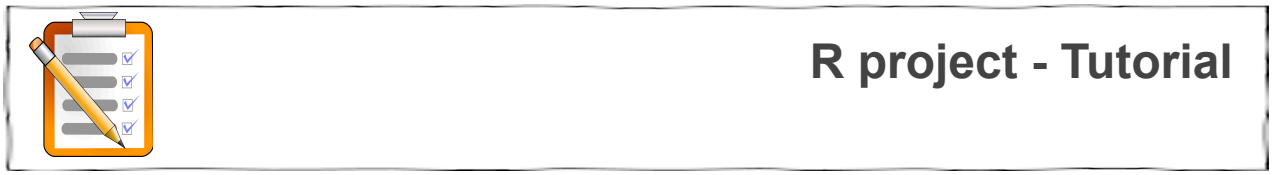
- ☐ https://training.galaxyproject.org/training-material/topics/galaxy-interface/tutorials/rstudio/tutorial.html
- ☐ https://intro2r.com/why-an-open-book.html

Monday, February 26, 2024

- 10:30-10:45 - Quiz: Questions on “RNA-Seq”
  - The quiz starts at 10:30am and ends at 10:45am. Do not be late! You won’t have extra time.
- Lecture (~30min): Introduction to R
- Time left: Working on Manuscript

Wednesday, February 28, 2024

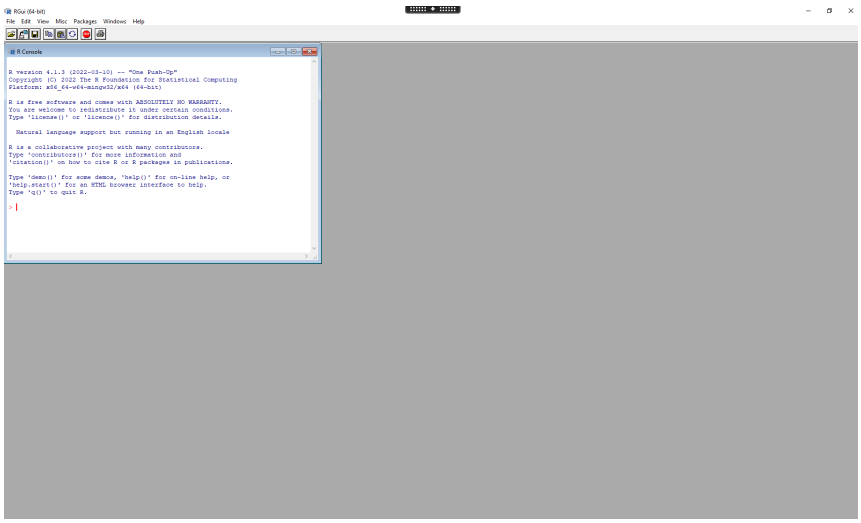
- R tutorial



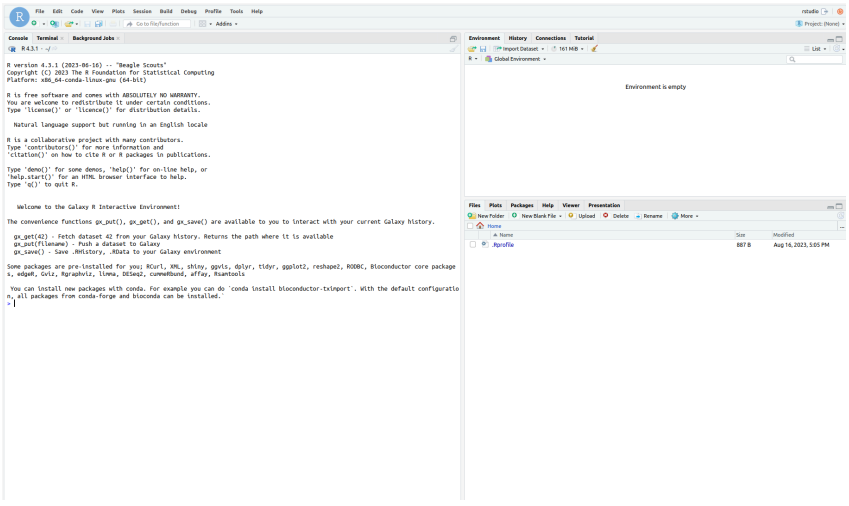
R project - Tutorial

R is a programming language for statistical computing and graphics. R software is open-source free software. R has a native command line interface. Moreover, multiple third-party graphical user interfaces are available, such as RStudio.

R interface



RStudio Interface (Galaxy)



Launch RStudio on Galaxy

To avoid any installation on your machine, you can use R via RStudio in Galaxy.
Tools > Search Tools > RStudio > Run Tool > Open

Launch R on TALC

Open a terminal and type:

```
module load R/3.6.2
R
```

Launch R on Windows

Click on icon

Quit R on terminal (laptop or TALC)

```
> quit()
Save workspace image? [y/n/c]: n
```

(if you say yes, next time you will open R, you will end up on the same session (variables still existing...etc). If you say no, next time you will start from scratch)

Basic commands

See current working directory

```
getwd()
```

Move to another directory

```
setwd("path/where/you/want/to/go")
```

Installation packages/libraries: R is open source. Anyone can create libraries (group of functions that will allow a complex analysis) and make it available in the CRAN repository or BiocManager. To know how to install a library, refer to the manual of each library.

```
install.packages ('<name_library>')
```

OR

```
if (!require("BiocManager", quietly = TRUE))
      install.packages("BiocManager")
BiocManager::install("<LIBRARY>")
```

Differential expression analysis tutorial

Background of the data: We are comparing gene expression (transcriptomic profiles) of two different tissues in pigs: peripheral whole blood and mesenteric lymph node. All tissues have been extracted from 4 different pigs. RNA-Seq has been done with Illumina short-read technology. Reads have been aligned with STAR, and read counts have been obtained with HTSeqCount. Data are available on Galaxy in a shared History: https://usegalaxy.eu/u/tmaroilley/h/rtutorial

DESeq2 tutorial:
http://www.bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html#htseq

############ In R ##
https://bioconductor.org/packages/release/bioc/html/DESeq2.html
```
if (!require("BiocManager", quietly = TRUE))
      install.packages("BiocManager")
BiocManager::install("DESeq2")


gx_get(1)
…
```

Load packages/libraries
```
library(DESeq2)
```

Creation of the table (called dataframe)
```
sampleName = c("Pig1MLN", "Pig1SG", "Pig2MLN", "Pig2SG","Pig3MLN",
"Pig6SG","Pig4MLN", "Pig4SG")
sampleFiles =
c("Pig1_MLN.htseqcount.txt","Pig1_SG.htseqcount.txt","Pig2_MLN.htseqco
unt.txt","Pig2_SG.htseqcount.txt","Pig3_MLN.htseqcount.txt","Pig3_SG.h
tseqcount.txt", "Pig4_MLN.htseqcount.txt", "Pig4_SG.htseqcount.txt")

sampleFiles = c("1", "2", "3", "4", "5", "6", "7", "8")

sampleCondition = c("MLN", "SG", "MLN", "SG", "MLN", "SG", "MLN",
"SG")
sampleTable = data.frame(sampleName = sampleName, fileName =
sampleFiles, condition = sampleCondition)
sampleTable$condition = factor(sampleTable$condition)
```

display top lines of the dataframe
```
head(sampleTable)
```

dimensions of the dataframe
```
dim(sampleTable)
```

Extract data from files to build an object (structure of data, specific to DESeq2, containing various types of information)
```
ddsHTSeq = DESeqDataSetFromHTSeqCount(sampleTable = sampleTable,
directory = "/import/", design= ~ condition)
```

```
# have a look at the object
ddsHTSeq
# class of the object
class(ddsHTSeq)
# to access the matrix of read counts
head(counts(ddsHTSeq))
# dimension of the read counts matrix
dim(counts(ddsHTSeq))
# header
colnames(counts(ddsHTSeq))
# to simplify, you can save the matrix under another variable
data = counts(ddsHTSeq)
#type and class of the variable
class(data)
typeof(data)
```

What is the size of the read counts matrix? (number of columns and rows) Deduce the number of genes.
What is the name of the columns?
What type of data are in "data"?
Of what class is "data"?

Filter genes - keep only the ones that are "expressed" (at least 10 reads in the whole cohort - arbitrary)

```
# function to calculate the sum of each row (meaning how many reads are fragments of transcripts expressed in at least one sample)
rowSums(counts(ddsHTSeq))
#see how many genes are present
length(rowSums(counts(ddsHTSeq)))
#keep only genes with more than 10 reads
rowSums(counts(ddsHTSeq)) >= 10
# keep list of genes with at least 10 reads (TRUE) in a variable and filter the DESeq2 object
keep = rowSums(counts(ddsHTSeq)) >= 10
ddsHTSeq <- ddsHTSeq[keep,]
```

What is the size of the read counts matrix after filtration? (number of columns and rows) Deduce the number of genes with >= 10 reads (expressed).
What is the name of the columns?

Run differential expression analysis - for each gene (row), is there a difference in the number of reads (in the level of expression) between 2 groups (blood and lymph node)?

```
ddsHTSeq <- DESeq(ddsHTSeq)

# extract the result and save it in a variable
res <- results(ddsHTSeq)
res
```

```
# What information can be found in "res"?

###### Filter genes with no significantly differentially expressed between blood and lymph node (p.value adjust < 0.05)
res05 <- results(ddsHTSeq, alpha=0.05)
# how many genes are differentially expressed?
sum(res05$padj < 0.05, na.rm=TRUE)
# have a look at "res05" - still ordered by Ensembl gene ID
head(res05)
# order "res05" according to p.value adjusted (level of significance) - the lower the p.value (or adjusted) is, the better
head(res05[order(res05$padj),])

# What are the 5 most significantly differentially expressed genes? Give Ensembl ID (the one available in the data, and the gene name - use ensembl.org)

###### Plots of normalized read counts by gene
# to create the plots and see it in a window or in R studio
# on TALC, your graph will automatically be saved in a Rplots.pdf file (in your working directory). To copy it to your laptop, type in your laptop's terminal or mobaxterm: scp <username>@talc.ucalgary.ca:/home/<username>/Rplots.pdf .
plotCounts(ddsHTSeq, gene="ENSSSCG00000012347", intgroup="condition")
plotCounts(ddsHTSeq, gene="ENSSSCG00000000001", intgroup="condition")
plotCounts(ddsHTSeq, gene="ENSSSCG00000000006", intgroup="condition")

# to save the plots in a file (in your working directory)
png("ENSSSCG00000012347.png")
plotCounts(ddsHTSeq, gene="ENSSSCG00000012347", intgroup="condition")
dev.off()

# Comment and compare those three plots

###### Heatmap of the 20 top genes
# normalization and selection of the 20 most differentially expressed genes
ntd <- normTransform(ddsHTSeq)
select <- order(rowMeans(counts(ddsHTSeq,normalized=TRUE)),
decreasing=TRUE)[1:20]
# draw the heatmap
heatmap(assay(ntd)[select,])

# What is a heatmap?
# to draw the heatmap with every gene
png("heatmap.png")
heatmap(assay(ntd))
dev.off()
```

Principal component analysis (PCA)
```
plotPCA(ntd, intgroup=c("condition"))
```

Interpret the PCA graph

Week 9 - Project Design

Deadlines Week 9: Quiz R (D2L) Mar. 04

The project you will be working on during the remainder of the semester aims to uncover SVs effect of gene expression. You will explore the transcriptome profile of your strain by analyzing RNA-Seq data. The data (fastq files and references) are available in Galaxy (shared History): https://usegalaxy.eu/u/tmaroilley/h/transcriptomicsdata.

Monday, March 04, 2024

- 10:30-10:45 - Quiz: Questions on “R”
    - The quiz starts at 10:30am and ends at 10:45am. Do not be late! You won’t have extra time.
- Lecture (~20min): Introduction to the RNA-Seq project and Biostatistics of transcriptome analyses
- Project design

Wednesday, March 06, 2024

- Project design
- Start analysis
- Presentation SoTL study

As you did for the analysis of a genome, you will work with your team to put together an outline of a proposal to analyze the transcriptome profile of your strain, made available to you by Dr. Tarailo-Graovac Lab, as RNA-Seq data. **The aim of your project is to explore the effect of SVs on the transcriptome.** You should first focus on the SV you have been working on while analyzing the genome of your strain, and then extend your project to additional SVs you have found, if any. Analyzing RNA-Seq data is based on many statistics concepts. Bioinformatics and Biostatistics often work together. I encourage you to sign in for a Biostatistics class as soon as you can. Those concepts are widely used and applied in biological and medical

research and avoid doing bad sciences by misinterpreting data. When it comes to analyzing transcriptomes, not only do you have to know a bit about coding, like in R, but you also need to understand the statistics of each library you are using to use it properly. Because mistakes can be made and statistics can be tricky, it is always recommended to validate (qPCR) some results before interpreting and publishing.

Note: You might be tempted to run DESeq2 to perform a differential expression analysis. The statistics beyond this library is based on comparing groups (at least 2 replicates per group). Here, each strain has been sequenced **only once and only one sample carries each SV, so you do not have groups. Other libraries give you more flexibility, such as limma. You can also check methods that allow you to create replicates to increase power (see bootstrap method for instance).**

If the question of the project is defined, as well as the data you will be working on, however, it is up to you and your team to develop the analysis based on literature. Remember that different Galaxy have different tools. Make sure to have a proposal including tools already implemented in Galaxy. Feel free to plan your project by using Galaxy, TALC and your machine for different parts of your analysis.

Note: Most popular R libraries are implemented in Galaxy in the ToolShed.

Resources to guide you in designing your RNA-Seq project

- ☐ **https://youtu.be/dx9-N8b9Yj4**
- ☐ **Batut B, van den Beek M, Doyle MA, Soranzo N. RNA-Seq Data Analysis in Galaxy. Methods Mol Biol. 2021;2284:367-392. doi: 10.1007/978-1-0716-1307-8_20. PMID: 33835453.**
- ☐ **Waters EV, Tucker LA, Ahmed JK, Wain J, Langridge GC. Impact of *Salmonella* genome rearrangement on gene expression. Evol Lett. 2022 Nov 19;6(6):426-437. doi: 10.1002/evl3.305. PMID: 36579163; PMCID: PMC9783417.**
- ☐ **Ma C, Shao M, Kingsford C. SQUID: transcriptomic structural variation detection from RNA-seq. Genome Biol. 2018 Apr 12;19(1):52. doi: 10.1186/s13059-018-1421-5. PMID: 29650026; PMCID: PMC5896115.**
- ☐ **Cmero M, Schmidt B, Majewski IJ, Ekert PG, Oshlack A, Davidson NM. MINTIE: identifying novel structural and splice variants in transcriptomes using RNA-seq data. Genome Biol. 2021 Oct 22;22(1):296. doi: 10.1186/s13059-021-02507-8. PMID: 34686194; PMCID: PMC8532352.**
- ☐ **Han L, Zhao X, Benton ML, Perumal T, Collins RL, Hoffman GE, Johnson JS, Sloofman L, Wang HZ, Stone MR; CommonMind Consortium; Brennand KJ, Brand H, Sieberts SK, Marenco S, Peters MA, Lipska BK, Roussos P, Capra JA, Talkowski M, Ruderfer DM. Functional annotation of rare structural variation in the human brain. Nat Commun. 2020 Jun 12;11(1):2990. doi: 10.1038/s41467-020-16736-1. PMID: 32533064; PMCID: PMC7293301.**

- ☐ **Brechtmann F, Mertes C, Matusevičiūtė A, Yépez VA, Avsec Ž, Herzog M, Bader DM, Prokisch H, Gagneur J. OUTRIDER: A Statistical Method for Detecting Aberrantly Expressed Genes in RNA Sequencing Data. Am J Hum Genet. 2018 Dec 6;103(6):907-917. doi: 10.1016/j.ajhg.2018.10.025. Epub 2018 Nov 29. PMID: 30503520; PMCID: PMC6288422.**
- ☐ **Mertes, C., Scheller, I.F., Yépez, V.A. *et al.* Detection of aberrant splicing events in RNA-seq data using FRASER. *Nat Commun* 12, 529 (2021). https://doi.org/10.1038/s41467-020-20573-7**
- ☐ **Yépez VA, Mertes C, Müller MF, Klaproth-Andrade D, Wachutka L, Frésard L, Gusic M, Scheller IF, Goldberg PF, Prokisch H, Gagneur J. Detection of aberrant gene expression events in RNA sequencing data. Nat Protoc. 2021 Feb;16(2):1276-1296. doi: 10.1038/s41596-020-00462-5. Epub 2021 Jan 18. PMID: 33462443.**

Week 10 - Project - Part 1

Deadlines Week 10: Proposal (Group - Ungraded) Mar. 11 + Report 3 (Individual - Ungraded - D2L Survey) Mar. 15

Monday, March 11, 2024

It is now time to work on the implementation of your project.

The Teaching team will be available to guide you at every step of this project. Use the in person session to ask for help and support. Take advantage of the office hours to reach out. **Your proposals are to be submitted to D2L by tonight, March 11 - 11:59pm.**

Wednesday, March 13, 2024

By the end of this week, you should have at least aligned your data and obtained a BAM file, available for visualization on IGV and further downstream analyses.

As a mandatory milestone, and in order to track every group progress, **please submit a report (R3) on D2L (>Survey) by Friday, March 15 - 11:59 pm.**

Week 11 - Project - Part 2

Monday, March 18, 2024

This is the last week for you to work on your project. By the end of this session, you should have at least explored the effect of one SV on gene expression. Use that opportunity to put in writing questions/issues that you might have. You will receive feedback by Wednesday to help you pursue your project.

Wednesday, March 20, 2024

Keep in mind that an **abstract is to be submitted this April 12th**.

Abstract Template

By Victoria R. A. Barbosa

The abstract should contain around 250 words that summarize all your work. Keep in mind that whoever read this might be hearing from your project for the very first time, so you want to give all the important information that will make them understand what are you doing, why are you doing it (background information that justifies this project/and what impact will it have) and how are you doing it. Also, make it short and simple and do not complicate it by using terms that are too specific of your area of study and that might be missed by people with a broader knowledge on that area. Check below an example of an abstract submitted to the Population, Evolutionary and Quantitative Genetics Conference (2022).

Hint: for word count, you can simply select the text and the word number should appear at the bottom left on your screen :)

Identifying the genetic factors in natural genome backgrounds that modulate essential phenotypic outcomes using *C. elegans*

→ Title

Afiya Razia Chida[1,2]*, Victoria Rodrigues Alves Barbosa[1,2]*, Xiao Li[1,2], Tatiana Maroilley[1,2], Francesca Jean[1,2], Tahsin Hassan Rahit[1,2], Andrew Galbraith[1,2], Filip Cotra[1,2], Larisa Oncea[1,2], Maja Tarailo-Graovac[1,2]

1. Departments of Biochemistry, Molecular Biology and Medical Genetics, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada T2N 4N1, Canada

2. Alberta Children's Hospital Research Institute, University of Calgary, Calgary, AB T2N 4N1, Canada

*Equal contribution

→ Authors and affiliations

Essential genes are vital for cellular development, metabolism, and reproduction; mutations with detrimental effects in these genes can cause severe defects. Essentiality varies based on different conditions, including genetic background. Our aim is to search for genetic factors that modulate gene essentiality in six diverse *Caenorhabditis elegans* natural isolate backgrounds - CB4856 (Hawaii, USA), JUI400 (Seville, Spain), AB1 (Adelaide, Australia), GXW1 (Wuhan, China) and KR314 (Vancouver, Canada), and N2 (Bristol, UK). The plasticity of essentiality might be influenced by genetic modifiers - variants that affect the causative gene by ameliorating or exacerbating a trait or disease. Our research focusses on identifying modifiers of two essential genes – *mat-1* (Metaphase-to-Anaphase Transition defect), which is involved in cell division, and *cgh-1* (Conserved Germline Helicase), which is important in oocyte development and gametogenesis. Previous studies with loss of CGH-1 and MAT-1 by RNA interference have suggested the presence of genetic modifiers in natural isolates (CB4856 and N2). To further investigate that, we individually knocked in conditional lethal temperature sensitive alleles of *cgh-1(tn691)* and *mat-1(ye121)* using CRISPR/Cas9 in all six isolates. At the permissive temperature (15°C), we observed about 100% hatch-rate in all isolates. However, at the restrictive temperatures (23°C & 25°C) both *cgh-1(tn691)* and *mat-1(ye121)* show evidence of phenotypic variability in different natural isolate backgrounds, ranging from 100% embryonic lethal to viable, indicating the presence of modifiers. Thus, to determine the genetic factors modulating the essentiality, the strains were subjected to whole genome sequencing followed by advanced bioinformatics analysis. We have detected and annotated a spectrum of variants, including single nucleotide variants and more complex structural variants, such as translocations and inversions. To identify potential genetic modifiers among them, we use genetic interaction network analysis, as well as in house developed machine learning approaches. To test the candidate modifiers, we are using CRIPSR/Cas9. Studying genetic modifiers using *C. elegans* natural backgrounds may improve the genomic approaches in the discovery of genetic modifiers in human diseases. This is crucial for proper diagnosis, prognosis, and more precise patient management.

Brief introduction on the theme

→ Specific topic of study and project aim

Previous literature that supports our hypothesis

→ Methods/experiments and results obtained (preliminary data)

→ Methods/what has been currently worked on

Significance

Additional tips from the Worm2023 meeting:

In all this last-minute editing and formatting, don't forget to pay attention to your abstract title. A seemingly small detail, the title can significantly impact the reader's impression of your abstract.

To help you come up with an abstract title that draws more attention to your work, we have some tips:

- ***Don't bury the lede.*** Start with the topic of the paper, not with the name of the gene or organism you studied.
- ***Entice the reader.*** Make what you learned seem exciting.
- ***Avoid jargon.*** Jargon is hard to avoid in technical publications, but you should do your best to purge it from the title.
- ***Be concise.*** Readers have a limited attention span. 13 words are more digestible than 30. In fact, most readers are unlikely to read past about a dozen words before their eyes wander away from your title.
- ***Don't give away the ending.*** Some authors treat the title as a one-sentence abstract, but that's a mistake. The purpose of the title is to entice readers with the question under investigation so they'll want to read more, not to tell the whole story. Don't give the conclusion of your story in the title.

Grading Rubric for Abstract:

Total points: _/20

Criteria and Description	0 Points	1 Point	2 Points	3 Points	4 Points
Title (1 point): Original and Short	Lacks originality and length issues.	Original and appropriately short.			
Introduction (4 points): Clarity and Efficiency - Introduce enough to provide context without overwhelming the reader.	No introduction written	Lacks clarity and overloads with information.	Somewhat unclear or overwhelming.	Clear but slightly overwhelming.	Clear, and well-balanced.
Methods (4 points): Provide a brief, clear description of the methods used, with visuals as a plus.	No methods provided	Extremely verbose or too little detail, making it unclear.	Excessive detail or too little detail, making it somewhat clear.	Appropriate length with mostly clear description.	Brief and clear description of methods with appropriate detail.
Results (4 points): Select relevant results	No results provided	Inappropriate results.	Presents some irrelevant results.	Mostly relevant results.	Well-selected relevant results.
Discussion/Conclusion (4 points): Clarity - Keep the discussion and conclusion short but clear.	No discussion provided	Unclear and lengthy.	Somewhat unclear or overly brief.	Clear and moderately concise.	Clear and effectively concise, conveying key points with precision.
Formatting (3 points): Follow formatting requirements (Calibri, single-spaced, 12pt, <= 250 words).	Does not meet formatting requirements.	Partially meets formatting requirements.	Mostly meets formatting requirements.	Meets formatting requirements.	

Week 12 - Poster Design

Deadlines Week 12: Poster (Group - D2L Dropbox) Mar. 27 + Peer Evaluation (Individual - Ungraded - D2L Dropbox) Mar. 27

Monday, March 25, 2024

It is time to finalize your project and draft your abstract. As if you were a student in a lab, you will submit an abstract to the conference organization team. Usually, the abstracts are reviewed by panels of researchers, to select the best abstracts for a presentation or a poster session.

Here, we organize a mock abstract submission. All abstracts will be accepted for a poster. But it is a good exercise to try to summarize your project and your results in an abstract.

Wednesday, March 27, 2024

Now is the time to start designing your poster. See below template and advice prepared for you by Victoria. Use our time together during this session to discuss with us your ideas and questions. Use the office hours of Shreya and Sue wisely to get help if you need it.

It is essential to submit your posters by March 27 on D2L - 11 am, so that they could be printed in time for our conference. Posters will be part of your grade. But also, they would be displayed during the conference and submitted to evaluation by judges. The best three posters will receive a prize.

Dimensions Poster: 24x36

Grading Rubric for Poster:

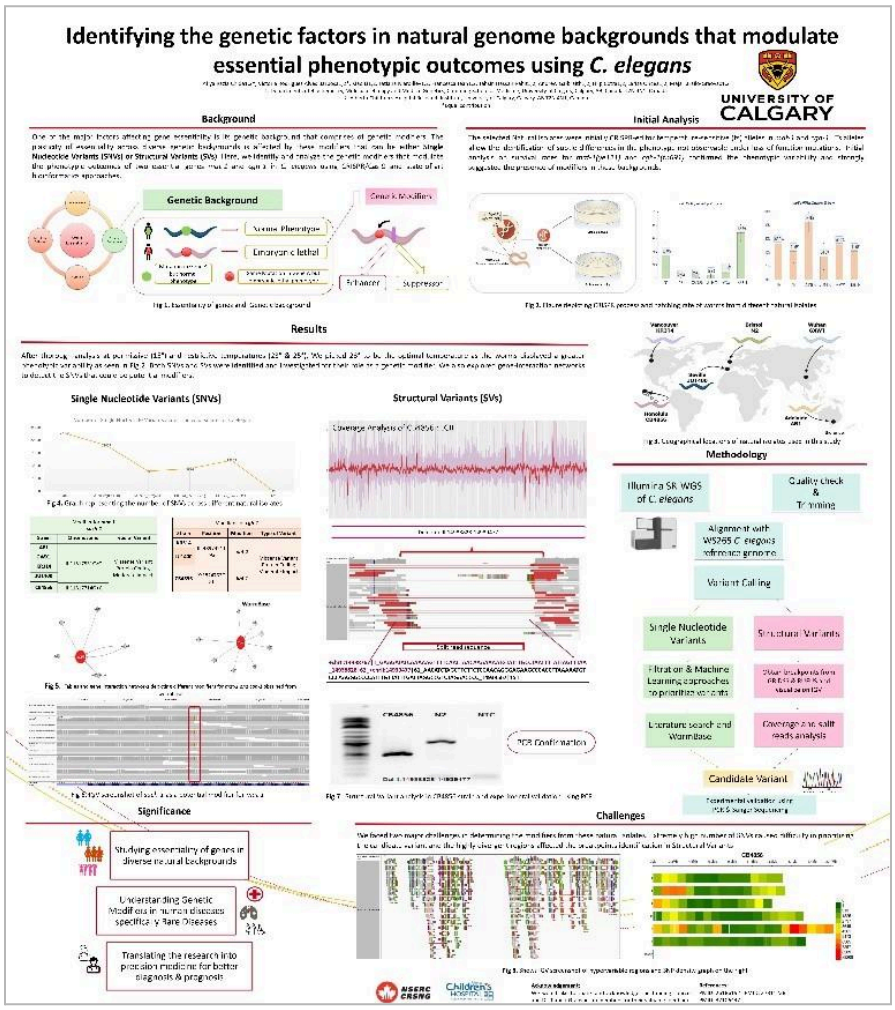
Total points: _/20

Criteria	0 Points	1 Point	2 Points	3 Points	4 Points
Title: Original and Short	Lacks originality and length issues	Somewhat original, but lengthy	Original and appropriately short		
Introduction: Clarity and Efficiency (introduce enough so that we can understand the context, but not too much to not overwhelm the reader and keep a good balance in the poster)	No introduction provided	Lacks clarity and overloads with information	Somewhat unclear or overwhelming	Clear and well-balanced	
Methods: Brief description of the methods used. Visuals are encouraged to describe the methods.	No Methods provided	Extremely verbose or unclear description. No visuals or irrelevant visuals	Excessive detail, somewhat clear description. Limited visuals with minor relevance	Appropriate length, and clear. Relevant visuals to support methods	
Results: Choice and Quality of Presentation	Missing results	Inappropriate results. Figures are of poor quality or missing figures	Presents some irrelevant results. Figures are of limited quality or relevance	Mostly relevant results. Figures are of good quality and relevance	Well-selected relevant results. High quality and relevant figures
Discussion/Conclusion: Clarity	Unclear and lengthy.	Somewhat unclear or overly brief	Clear and concise		
Formatting: Calibri, Title can be up to 66pt, Author list: 36pt, Main text: 36 pt, Subtitles: 40pt	Text is not readable, Title and subtitles do not stand out	Text is challenging to read on some part of the poster	Text is readable even though not 36 pt		
Design: Readability and organization of the poster	Not readable/Not organized	Extremely difficult to read and follow	Somewhat challenging to read and follow	Mostly easy to read and follow	Easy to read, well-organized and follow

Poster Template

By Victoria R. A. Barbosa

Now is your time to shine! Time to gather all your results and present them as a poster at our mini-conference. Bear in mind that your poster can probably not fit all you have done. It is totally okay (and actually recommended) that you focus on the most important information only. Any extra explanation can be given by you in case it's needed. However, imagine that you walk out for a minute, and someone comes to check your work, but you're not there: the person needs to be able to understand your research by himself just based on what you have on the poster. Therefore, try to keep the poster organized, give enough key-information, and even exemplify ideas with figures (or screenshots from your SVs on IGV, for example).



Remember that you are communicating your ideas through a visual form, and some text is necessary, but should not represent the whole poster. Be creative! Make use of figures, diagrams, logos for programs you've used, anything that can facilitate understanding and retain people's attention.

To make things clearer, separate your information into different sections: Introduction, Materials and Methods, Results, Discussion , Significance

- Font choice: Keep it simple, use Calibri.

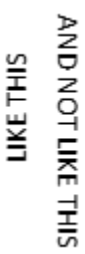
- Use a creative title: that is usually what people see first and can help a lot with drawing the public's attention!
- As always, try to keep the Introduction short (no more than a 1/6 of your poster).
- Use blank spaces wisely: not too much or your poster will look empty, but just enough so that the structure of your poster is clear, even for people looking at it from afar (or over someone else's shoulder).
- Tip: Print your poster on an A4 page to help with visualization.

UofC offers the following templates for conference posters (feel free to adapt them to your needs):
https://uofc.sharepoint.com/:p:/r/sites/spo-dept-advancement/UCalgary%20Brand/Templates%20(PowerPoint)/Research%20posters/Research%20poster%20template-portrait-36x48-October%202018.potx

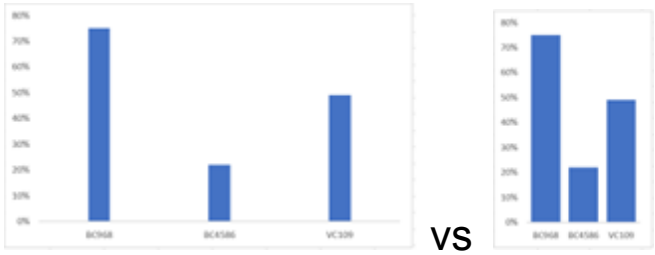
Figures make your work memorable! They draw attention and sometimes are the first reason why someone decides to further read a paper or stop by a poster. They are always welcome to present information through any form, as they facilitate the understanding of what is being communicated. “A picture is worth a thousand words”, they said ;)

You can, for example, create a graphical abstract whose goal is to be a visual summary of all your work. But keep in mind that you should only add figures where that is really feasible and necessary. They should act as a hook to your main idea and an extra help on explaining your main point. If you decide to create one, check the following tips:

- Pay attention to the font choice: focus on Calibri as these were shown to be clearer and facilitate reading; Also, make sure you are choosing a reasonable font size and keep it consistent throughout the whole figure (except for titles, which can be a little bigger).
- P.S.: take a look at your document with 100% zoom to see how it will look in real life. Especially when preparing a poster, the computer screen might be deceiving.
- Avoid abbreviations if feasible, as usually a poster will target a broader audience that might not be as familiar with those terms as you are
- Keep in mind that a good figure can usually stand out for itself and be understood even out of context. Use a descriptive title (that can even say exactly what the take-home-message of that figure or graph is). For example: instead of saying: “Comparative graph of mRNA expression levels of genes X and Y from pancreatic tumor”, you can simply say “Gene X is overexpressed in pancreatic tumors in comparison to gene Y” (that will better direct the reader towards the point you want to make).
- If creating a graph or a figure in which the text needs to be rotated, give preference to do it clockwise, which makes the reading easier:

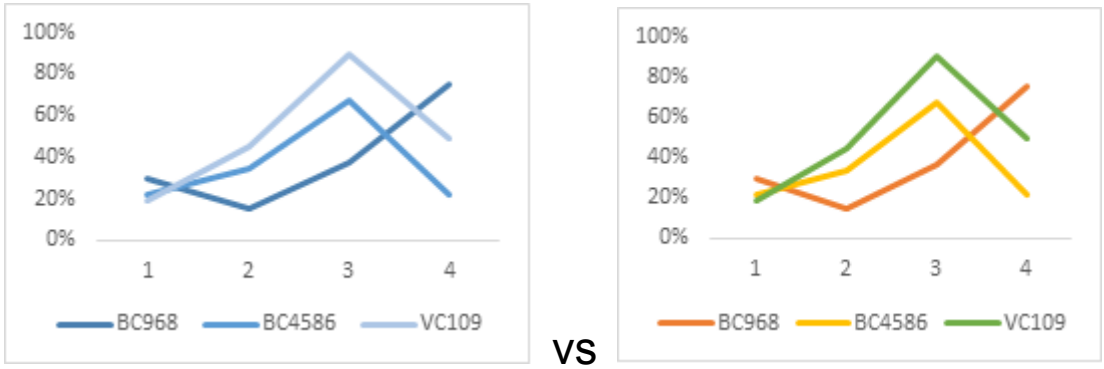


- Don’t hesitate to make use of shading or annotations to highlight a point of interest in your figure (like using an arrow to show the location of a variant of interest, for example). This will also help the reader to look for the right things while going through your data.
- Reduce excessive negative space - this might save you some space and make the graph/figure more appealing:



VS

- Last, but not least: as tempting as it is to use all colors of the rainbow in your figures, use them whenever they have a purpose, and be consistent with them. For example, in a line graph or scatter plot, it might be interesting to use different colors to better differentiate the values, as exemplified below:



VS

If you pick a color to a specific variable on the first figure (for example, blue for strain A, green for strain B), be consistent and use the same colors on the next figures to avoid any misunderstanding.

Follow these tips and voilà! You’ll have a beautiful and captivating figure to work with :)

Week 13 - Conference Preparation

Deadlines Week 13: None

It is usual to prepare a speech of a few minutes to present your poster to people coming to it during the conference. In addition, some student events, such as Summer Students Research Days usually ask the participants to prepare a 3-minute speech to the judges in order to help decide the winner of the Best Poster Award.

Monday, April 01, 2024

Holiday.

Wednesday, April 03, 2024

On Wednesday, you will practice your speech in front of the entire class. After each group, the class and the Teaching team will provide constructive feedback. The objective of this session is to get you ready to present your poster during our conference the following week.

Week 14 - Conference

Deadlines Week 14: Reflection Assignment 2 (individual - D2L Quizzes) + Abstract (individual - D2l Dropbox) Apr. 12

Monday, April 08, 2024

This is Conference Day! Let's meet in the Hall at 10:30 am to set up the posters and get you ready. By 10:45 am, students, postdocs and PIs will come to visit the poster at the conference. Be ready to give your speech and answer questions! Among them, the judges will pass and evaluate your poster, your presentation and your ability to answer questions.

We will then organize a Award Ceremony at 11:40 am, right before the end of our session and our semester together!

Reflection Assignment 2: Available in D2L (Quiz)

Min. 150 words per question

Questions:

1. Was the research project an enriching educational experience for you? Explain.
2. Did the research project change your perception of the Bioinformatics field? In what way? If not, why?
3. Was the research project conducive to your learning style?
4. One thing you learned that was unexpected?
5. One thing you learn that will probably be very useful in the future?
6. One thing you would have done differently as a team mate?
7. One thing you would have done differently in your research project?

Grading Rubric for Reflection Assignment Questions

7 questions (5 points each) - Total: __/35

Criteria	0 Point	0.25 Point	0.5 Point	0.75 Point	1 Point
Relevance (1pt)	Response lacks alignment with the question; content is unrelated or minimally related to the prompt.	Limited relevance; some connection to the question but lacks a clear focus.	Generally relevant; addresses the question with a clear focus on key points.	Highly relevant; effectively addresses the question, demonstrating a	Exceptionally relevant; addresses the question comprehensively,

				clear and focused response.	providing depth and insight.
Clarity (1pt)	Response is unclear, making it difficult for the reader to follow the line of thought.	Somewhat clear; the response has some coherence but may lack smooth transitions.	Clear and organized; the reader can easily follow the line of thought.	Very clear; well-structured with smooth transitions, facilitating easy understanding.	Exceptionally clear; exceptionally well-organized, with a logical flow and seamless transitions.
Significance (1pt)	Lacks personal reflection; the response is superficial and lacks depth.	Limited personal reflection; includes some depth but lacks a profound connection to the proposed activity.	Demonstrates a deep personal reflection on the proposed activity, showing thoughtful consideration.	Displays a significant and meaningful personal reflection, providing insights into the proposed activity.	Exhibits an outstanding personal reflection, demonstrating a profound connection to the proposed activity.
Mechanics (1pt)	Numerous typographical, spelling, and grammatical errors throughout the response.	Some typographical, spelling, or grammatical errors present, impacting overall clarity.	Consistently avoids major typographical, spelling, and grammatical errors, enhancing clarity.	Virtually error-free; minimal typographical, spelling, or grammatical errors.	Impeccable mechanics; response is free of typographical, spelling, and grammatical errors.
Length (1pt)	Below 150 words or exceeds 200 words, deviating significantly from the specified range.	Falls within the 150-200 word range but may be slightly below or above the specified limits.	Within the 150-200 word range, meeting the specified length requirements.	Well-balanced; maintains a concise and focused response within the specified word range.	Optimal length; response is precisely within the 150-200 word range, demonstrating effective conciseness.