BINF6309 Final

## Instructions

Insert your answers to the questions below in this Word document and upload to the Canvas link under Modules. For open-ended questions, copy-and-paste answers or answers typed word-for-word from web resources or papers will not receive credit. You must rephrase in your own words to show understanding of the topic. Open-ended questions must be answered in complete, grammatically correct sentences with APA citations for external resources used; please include the citations at the end of the question, not the document. Answers to open-ended questions will be graded on completeness, clarity, grammar, spelling, formatting (e.g. body text not all in bold, only headers in bold, etc.), and proper citations.

There can be multiple correct answers to some of the questions, depending upon assumptions made. Explain your assumptions as part of your answer. If your answer is logical and complete given the assumptions you describe, you will receive credit for the answer.

If you feel that stating assumptions is not sufficient to clarify a question, you may reach out for clarification. The instructor’s response will be shared with the entire class so that all are on the same footing. Similarly, to maintain equity, no questions will be answered in the final day of the exam. That is, if the exam is due on 2/27, then no questions received after 11:59PM EST on 2/26 will be answered.

All questions are worth 5 points graded out of a 150 point scale.

For questions asking for code, you should insert the code in the Word document – taking some care to preserve the formatting (similar to the code in question 2).

1. What is linkage disequilibrium and how does it affect GWAS studies?

Linkage disequilibrium (LD) is the non-random association of alleles at different loci. It can be thought of as the tendency of a certain genetic variant to be inherited together.

In the context of genome-wide association studies (GWAS), LD mapping involves the use of dense maps of SNPs across the human genome to identify regions of the genome that contain functional DNA-sequence variants that influence a trait or disease.

If a significant frequency difference in allele frequency is found between cases (patients with a specific disease or individuals with a certain trait) and controls, it may indicate that the corresponding region of the genome contains functional DNA-sequence variants that influence that disease or trait.

LD is useful for mapping complex traits and diseases that do not follow simple Mendelian inheritance patterns, as it recognizes that a mutation that is shared by affected individuals through common descent will be surrounded by shared alleles at nearby loci, representing the haplotype of the ancestral chromosome on which the mutation first occurred.

1. The bash script below has multiple errors. Correct them so that this will run on the Discovery server. To make sure your answer is correct feel free to run the corrected script on the Discovery server. Assume that you are running the script in an interactive mode (i.e. that there is no sbatch wrapper script.)

#!/bin/bash  
# trimAll.sh  
#Initialize variable to contain the directory of un-trimmed fastq files  
fastqPath="/scratch/AiptasiaMiSeq/fastq/"  
#Initialize variable to contain the suffix for the left reads  
leftSuffix=".R1.fastq"  
rightSuffix=".R2.fastq"  
pairedOutPath="Paired/"  
unpairedOutPath="Unpaired/"  
#Loop through all the left-read fastq files in $fastqPath  
for leftInFile in $fastqPath\*$leftSuffix  
do  
 #Remove the path from the filename and assign to pathRemoved  
 pathRemoved="${leftInFile/$fastqPath/}"  
 #Remove the left-read suffix from $pathRemoved and assign to suffixRemoved  
 sampleName="${pathRemoved/$leftSuffix/}"  
 nice -n19 java -jar /usr/local/programs/Trimmomatic-0.36/trimmomatic-0.36.jar PE -threads 1 -phred33 \  
 $fastqPath$sampleName$leftSuffix \  
 $fastqPath$sampleName$rightSuffix \  
 $pairedOutPath$sampleName$leftSuffix \  
 $unpairedOutPath$sampleName$leftSuffix \  
 $pairedOutPath$sampleName$rightSuffix \  
 $unpairedOutPath$sampleName$rightSuffix \  
 HEADCROP:0 \  
 ILLUMINACLIP:/usr/local/programs/Trimmomatic-0.36/adapters/TruSeq3-PE.fa:2:30:10 \  
 LEADING:20 TRAILING:20 SLIDINGWINDOW:4:30 MINLEN:36  
done

1. Write RMarkdown code below to produce an html document with citations from bibliography.ris. When knitted the code should just print the citation for the .ris entry below. Assume bibliography.ris includes this:

TY - JOUR  
ID - Love  
DB - PubMed  
AU - Love, Michael I  
AU - Huber, Wolfgang  
AU - Anders, Simon  
...

1. When counting items in a dictionary, why should you make sure the loop to iterate over the dictionary and print the count results is not within the loop that does the counting? What happens if you incorrectly put the iterate/print loop within the counting loop?

You should make sure to print the count outside of the loop so that when it iterates over the dictionary it does not print out the result with each iteration. It instead only prints the final count. It is not technically incorrect, but it would lead to more computing power to print each iteration

Ex: print count inside the loop:

my\_dict = {'a': 1, 'b': 2, 'c': 3}

count = 0

# prints inside the loop (incorrect)

for key, value in my\_dict.items():

count += 1

print(f'Count: {count}')

Output:

Count: 1

Count: 2

Count: 3

Ex: print count outside the loop

my\_dict = {'a': 1, 'b': 2, 'c': 3}

count = 0

for key, value in my\_dict.items():

count += 1

# Print the count after the counting loop

print(f'Total count: {count}')

1. Write a bash script below that will find files in the FASTQ directory (relative path) ending in **.fq** and print them to the console.

Run the file find\_fastq\_files.sh using the command ./find\_fastq\_files.sh. Bwlos is the bash script in the file. Use chmod +x find\_fastq\_files.sh if do not have proper permissions.

#!/bin/bash

# run using: ./find\_fastq\_files.sh

# use chmod +x find\_fastq\_files.sh

# Set the directory path

directory="FASTQ"

# Check if the directory exists

if [ -d "$directory" ]; then

# Find and print files ending in .fq

find "$directory" -type f -name "\*.fq" -exec echo {} \;

else

echo "Directory not found: $directory"

fi

1. Summarize the basic steps in Differential Expression analysis starting with a directory of un-trimmed RNA-Seq reads and a genome. I’m looking for general descriptions of the steps, not code.

The basic steps in Differential Expression analysis starting with a directory of un-trimmed RNA-seq reads and a genome is

1. Quality control/preprocessing
   1. Trim low quality reads – usually using trimmomatic
2. Read alignment to reference genome
   1. Process the reads to the reference genome using STAR, HISAT2, or Bowtie to generate a BAM/SAM file (alignment/map file) with mapped reads
3. Read quantification
   1. Quantify the gene or transcript expression by counting the reads mapped to gene/transcript – using featureCounts of HTSeq for
4. Normalization
   1. Normalize raw counts to account for variations in sequencing depths/technical biases to ensure that expression values are comparable between samples.
   2. Can use DESeq2 which employs median-of-ratios method to adjust for differences in library sizes and correct for systematic biases, trimmed means of M values (TMM) or reads per kilobase per million mapped reads (RPKM)
5. Statistical analysis
   1. Use statistical analysis to identify differentially expressed genes – DESeq2
6. Multiple testing correction
   1. Adjust p-values for multiple testing to control the false discovery rate (FDR)
7. Visualization
   1. Plot using volcano plots, heatmaps, MA plots, etc.
8. Functional enrichment analysis
   1. Use functional enrichment to identify biological pathways or gene ontology terms associated with differentially expressed genes
9. Annotation and interpretation
   1. Annotate differentially expressed genes with biological information – gene symbols, pathways, and biological functions.
10. What is a guide tree, how is it used in multiple sequence alignments, and how does its accuracy affect the alignment?

A guide tree is a hierarchical structure that is constructed to guide the alignment of multiple biological sequences in a multiple sequence alignment. The primary purpose of the guide tree is to represent the evolutionary relationship between the sequences being aligned and is used to see which sequences align to each other with the idea that more closely related sequences will be aligned first.

If the guide tree is accurate, it helps in aligning homologous regions correctly. However, if it is incorrect, it might lead to misalignments since sequences that are not closely related evolutionarily might be linked.

1. Describe the difference between Multinomial, Markov Sequence and Hidden Markov Models in a few sentences in your own words.

Multinomial sequence is a probability distribution model used for discrete data with multiple categories. It assumes each observation belongs to one of the predefined categories and the probabilities of each category are independent of each other.

Markov sequence assumes that the probability of transitioning from 1 state to another depends only on the current state and not on the sequence of events that preceded it and is good for modeling sequential dependencies.

Hidden Markov Model (HMM) is an extension of the Markov model with hidden states. In an HMM, observable outcomes are influenced by underlying states that are not directly observed. The system is able to evolve through a series of these hidden states by associating them with probability distribution for each observable outcome.

1. Explain in your own words why objects and classes are essential to understand OOP?

Objects and classes are the foundation of OOP (object oriented programming). They provide a powerful way to organize and structure code that is also reusable.

Classes are the blueprints/templates that can encapsulate properties (attributes) and behaviors (methods). The classes can be called on as objects in code and reused time and again with different attributes allowing for complex systems to be created by the class.

1. What type of object would sequences need to be in the piece of code shown below?

SeqIO.write(sequences, "kmers.fasta", "fasta")

The sequences would need to be SeqRecord objects which can be found in the BioPython package.

1. Summarize in your own words the differences between paired-end and mate-pair reads. Why are mate-pairs sometimes referred to as “outward pairs”?

Paired-end and mate-pair reads are 2 types of DNA sequencing methods looking at fragments of the DNA.

Paired-end sequencing sequences from both ends of the DNA fragment. This is typically for shorter reads that are next to each other on the reference genome. They are good for mapping reads to the reference genome and for identifying structural variants (insertions, deletions, duplications)

Mate-pair sequencing sequences DNA fragments from a fixed distance apart. This is good for looking at larger structural variants and genome assembly. They are sometimes referred to as “outward pairs” because the reads are sequenced outward from a larger DNA fragment, capturing information about the genomic intervals between the paired reads. They are pointed away from each other on the original DNA fragment which gives insights into genomic distances and orientation.

1. Briefly explain MCMC methods and how relates to posterior distribution in your own words.

MCMC stands for Markov Chain Monte Carlo methods and it is a class of algorithms used to estimate probability distributions. It is used to estimate the posterior distribution of model parameters which represents beliefs about the parameters after taking into account prior knowledge/observed data.

1. What is the name for the version of a variant found in the reference genome?

Variants found in the reference genome are called either wild type allele or reference allele. The reference genome is the most commonly observed in a population and is used for a standard for comparison.

1. Summarize the key differences between 16s sequencing and shotgun metagenomic sequencing (not metatranscriptomic sequencing). What are the pros and cons of each? What are the key differences, pros, and cons between untargeted metagenomic NGS and 16S sequencing?

16s sequencing sequences the targets specific to the 16s rRNA gene which allows for identification of bacteria and archaea. This method is more coset-effective as it only looks at the 16s rRNA for specific bacteria and arachaea and is commonly used for community profiling. The cons is that it only looks at bacteria and archaea and does not have a wide taxonomic resolution.

Shotgun metagenomic sequencing analyzes the entire genetic material in a sample to find all of the organisms present in the sample (bacteria, fungi, archaea). The pros to this method is that it gives a comprehensive look at all genetic material. However, the cons are that it costs more and requires more computational power. Also, it has lower sensitivity for less abundant species inside the sample.

Untargeted metagenomic NGS analyzes all genetic material in the sample and is also able to find all information on bacteria, viruses, fungi, and archaea. This is also generally more expensive and requires more computational power/analysis.

1. Briefly describe one way to make an unrooted tree a rooted tree.

One way to make an unrooted tree a rooted tree is to include the outgroup species in the analysis. This group is known to be more distantly related that the species of interest. The root is believed to be located along the brand that leads to the outgroup so that the tree for the ingroup species is rooted.

Citation: ZIHENG YANG, & RANNALA, B. (2012). Molecular phylogenetics: principles and practice. *Nature Reviews. Genetics*, *13*(5), 303–314. https://doi.org/10.1038/nrg3186

1. Why is it better to translate a nucleotide sequence to an amino acid sequence for multiple sequence alignment?

It is better to translate a nucleotide sequence to an AA sequence for multiple sequence alignment because this is going to be a shorter sequence that has more biological meaning due to it emphasizing the functional and structural components of the sequence. It also has less gaps and gets rid of missense mutations (mutations that do not change AA translation).

1. Why might a SNP be correlated with a disease or trait, despite not having any causative effect on the disease or trait?

A SNP could be correlated with a disease/trait despite not having causative effect on the disease or trait by linkage disequilibrium where a SNP could be physically close to another disease-causing gene that could lead it to be correlated to a disease/trait since it is inherited together with the diseased mutation.

1. Do GWAS studies based on commercial SNP arrays generally identify causal variants? Why or why not?

Genome wide association studies based on commercial SNP arrays generally do not identify causal variants. GWAS is used to detect associations between alleles/genotype frequency trait status. In this method, causal variants are not directly genotyped but are in linage disequilibrium with genotyped SNPs. Functional characterization is necessary to move the statistical associations found in GWAS to causal variants.

Citation: Tam, V., Patel, N., Turcotte, M. *et al.* Benefits and limitations of genome-wide association studies. *Nat Rev Genet* **20**, 467–484 (2019). https://doi-org.ezproxy.neu.edu/10.1038/s41576-019-0127-1

1. How does scoring of alignment matches differ for nucleotide sequences versus amino acid sequences?

Scoring alignment matches differ for nucleotide sequences vs AA sequences due to the fundamental differences between DNA/RNA and proteins. Nucleotide sequence alignments look closely at mutations and indels while proteins look at structural variants. Nucleotide scoring can be based on the level of transitions(purine to purine, pyrimidine to pyrimidine)/transversions(purine to pyrimidine and other way around), gap penalties, and substitutions. AA sequences are scored also on substitutions and gap penalties but then look at sequence similarity (polar, non-polar, aromatic, etc.).

1. In terms of metagenomic shotgun sequencing, what is enrichment, and how can it affect the downstream analysis of the data?
2. Briefly define haplotype in your own words.

Haplotype is a set of genetic variations or alleles found on a single chromosome that are inherited together as a unit. It will be a combination of specific genetic markers/sequences located closely on the chromosomes.

1. Do multiple sequence aligners score every possible path to produce the optimal alignment? Explain why or why not.

Multiple sequence aligners do not score every possible path to produce the optimal alignment. This is too computationally complex as the number of possible alignments grows exponentially with the number of sequences being aligned which makes it impractical to find every possible path. So, instead aligners use heuristic methods and optimization algorithms such as ClustalW to approximate the optimal alignment.

1. Describe the problem of being “trapped in a local minima” for progressive alignment methods.

Progressive alignment methods is the alignment process that builds a guide tree that represents the evolutionary relationships between the sequences. The sequences are then aligned progressively, thus the name. When you are “trapped in a local minima” the algorithm converges to a suboptimal solution that is locally optimal, but not necessarily globally optimal. This is due to the guide tree construction being based on heuristics/approximation algorithms making it possible for the algorithm to get stuck in a locally optimal position. This problem generally occurs when the algorithm makes a suboptimal decision at an early stage in development and can propagate through the subsequent steps.

1. What are the three types of HMM variants? Briefly explain each.

The 3 types of HMM (hidden Markov models) are

1)profile-HMMs – HMMs with a specific architecture. These are strictly left to right structure that does not contain any cycles. It repetitevly uses 3 types of hidden states: match states (Mk), insert states (Ik), and delete states (Dk) to describe the model. They extend the traditional HMMs by incorporating information about position-specific residues. In sequence alignment

2)pair-HMMs – this HMM is a variant of the basic HMM and generates an aligned pair of sequences that goes between multiple states to simultaneously generate 2 aligned DNA sequences. These are employed in context specific alignment, particularly comparing 2 sequences since they generate the aligned pair.

3)context-sensitive HMMs – these HMMs can use past findings to adjust the probabilities at future states. So, it can use the previous context to make predictions. It is particularly good at long range correlations between symbols. It uses 3 states: single-emission states Sn, pairwise-emission states Pn, and context-sensitive states Cn along with pairwise-emission states and the context-sensitive states. This HMM allows us to describe any kind of pairwaise symbol correlations by properly arranging pairwise emission states with context-sensitive states.

Citation: Yoon BJ. Hidden Markov Models and their Applications in Biological Sequence Analysis. Curr Genomics. 2009 Sep;10(6):402-15. doi: 10.2174/138920209789177575. PMID: 20190955; PMCID: PMC2766791.

1. The number of possible phylogenetic trees for a given alignment is quite large. Do most phylogenetics packages evaluate every possible tree? If not, why not and how is this handled?

No, most phylogenetic packages do not evaluate all possible trees. This can be extremely large and thus it is unrealistic to generate each. Instead phylogenetic packages use heuristic methods to search through tree space efficiently and find trees that are likely to be good approximates of the true evolutionary history. These heurisitics include things such as maximum parsimony (considering possible arrangements of taxa with least total number of changes is most parsimonious) and Bayesian inference (estimates posterior distribution of trees)

1. What alignment methods are used to produce the guide tree for progressive alignments?

The alignment methods to produce the guide tree for progressive alignments is UPGMA and neighbor-joining.

1. How are pairwise distances between sequences used for iterative alignment methods?

Pairwise distances between sequences are used for iterative alignment methods in multiple sequence alignments (MSA) for constructing guide trees in methods like progressive alignment.

In iterative methods, they can be used to modify the alignment produced by progressive methods and to evaluate the quality of a given alignment. The guide tree in iterative methods can be modified and the distances recalculated for more accurate alignment. Pairwise alignments and alignments with outgroup sequences can also be used to maximize consistency and improve early alignment stages through a scoring scheme.

Citation: Biswanth C., Gautam G., A review on multiple sequence alignment from the perspective of genetic algorithm. Genomics. 2017, Volume 109, Issues 5–6, October 2017, Pages 419-431

1. Under maximum parsimony, what must be true of a site in an alignment for that site to be informative for tree comparison?

Maximum parsimony minimizes the number of changes on a phylogenetic tree by assigning character states to ingerior nodes on the tree. Under maximum parsimony, a true site in an alignment must exhibit at least two distinct characters being observed twice because only sites that are variable and exhibit 2 or more character states are informative.

Citation: ZIHENG YANG, & RANNALA, B. (2012). Molecular phylogenetics: principles and practice. *Nature Reviews. Genetics*, *13*(5), 303–314. https://doi.org/10.1038/nrg3186

1. What type of substitution involves the interchange of a purine and a pyrimidine?

Transversion is the interchange of a purine (2 rings A/G) to a pyrimidine (1 ring T/C).

1. Are somatic variations heritable? Explain why or why not.

No somatic variants are not heritable. Somatic variants are variants that are acquired throughout one’s life and are usually found in lower frequency compared to germline variations due to germline variations being present since birth.