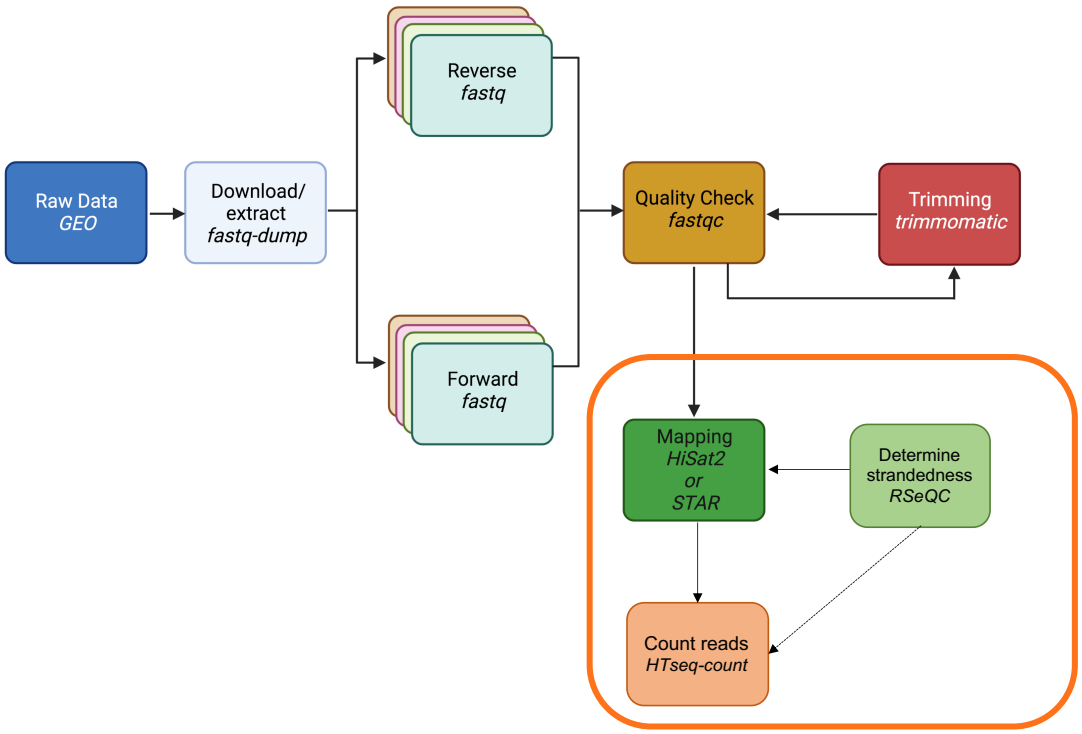
MMG3320/5320
Spring 2025
Homework #7
200 points

Guidelines for this homework:
All students have now: 1) downloaded raw data from GEO, 2) performed QC with fastqc and trimmed (*optional*), and 3) are ready to perform read alignment and counting. This will involve the most computationally expensive step of the entire pipeline, so please allocate the appropriate amount of time (*days not hours*) to complete prior to due date.



Due date: The due date for this homework is Friday, April 4th by 5:00pm.

PART A (50 points): BASH script (.sh) used to perform the alignment and create BAM file
PART B (50 points): BASH script (.sh) used to count reads or other downstream processing files.
PART C (50 points): Final MULTIQC (.html) output displaying plots from the alignment, counting, and RSeQC stage.
PART D (50 points): Fill out the table below for your samples. Submit the filled-out table in a separate Microsoft Word document (.docx).

Most of these file types will not be accepted by Brightspace, therefore, please send me ONE email with PART A-D included. Late homework will be docked 10% of the overall grade for every day that the assignment is late. An assignment is considered late if it is not submitted by the date and time specified. Three days past the due date (weekend included), the assignment will no longer be accepted, and the student(s) will receive a ZERO.

Please email princess.rodriguez@med.uvm.edu if you have any questions.

Instructions:

PART A (50 points): Customize the HISAT2 script provided to align FASTQ files. Please attach in the email your final bash script (.sh file).

PART B (50 points): Customize the HT-SEQ script provided to count reads. Please attach in the email your final bash script (.sh file). If you are performing another NGS analysis besides, RNA-Seq, please consult with Dr. Rodriguez.

PART C (50 points): Create a MULTIQC report using output files from both Part A, Part B, and RSeQC. The final multiqc should be submitted as a .html file. *It is okay to submit two separate multiqc outputs since RSeQC is giving issues.*

PART D (50 points): Fill out the table below for your samples. Submit the filled-out table in a separate Microsoft Word document (.docx).

Important Links:

Helpful Tips: https://prodriguez19.github.io/MMG3320-5320/assignments/Helpful_Tips_HW7/

Information about Index Genomes & GTF: https://prodriguez19.github.io/MMG3320-5320/assignments/genome_index-2025/

Category	Metric	Ideal Threshold	Comment	Your Samples (provide range)
General Statistics	Overall Alignment Rate	> 80%	lower suggests contamination or poor reference compatibility.	
	M Reads	> 20 M	To perform downstream statistically analysis > 20 M is required	
HTSeq Count	Percentage of Assigned Reads	> 70%	High percentage assigned to exonic regions.	
Read Distribution	Exonic Reads	> 70%	Most reads should map to exons.	
	Intergenic Reads	≤ 10%	Low intergenic mapping suggests minimal contamination.	
Infer experiment	Sense	will vary		
	Antisense	will vary		
	Undetermined	will vary		
HiSAT2	SE/PE mapped uniquely	> 70%	Reads should map uniquely to the genome	
	SE/PE multimapped	≤ 10%		
	SE/PE not aligned	≤ 15%	Higher rates suggest poor alignment	