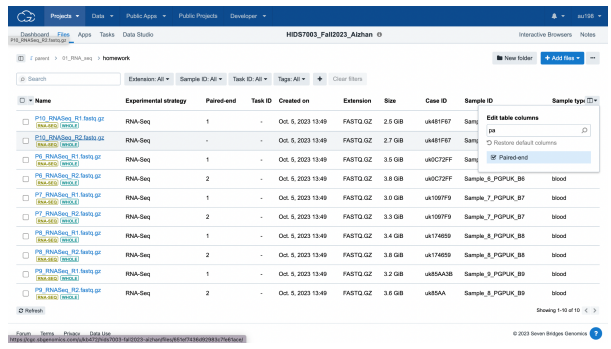
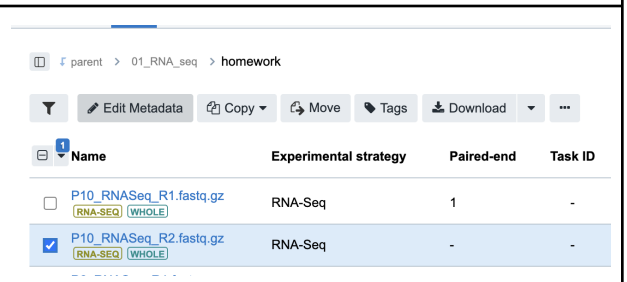
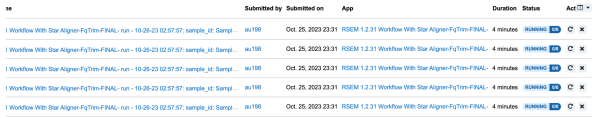
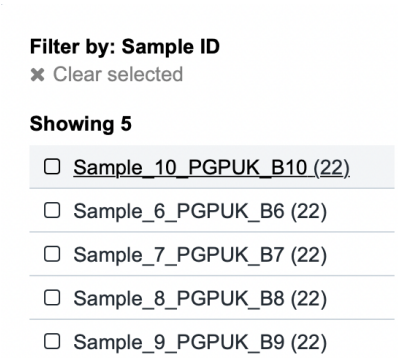
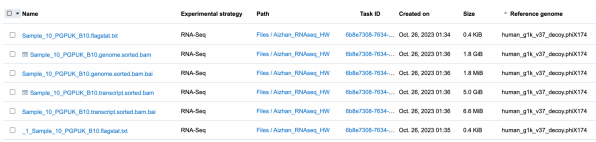
NetID: au198
Aizhan Uteubayeva

Homework 9: Data Processing of RNA-seq data
HIDS-7003-01

I	Edited the metadata Added the columns for: - Pair-end - Case ID - Sample ID	
II	First error appeared to be in the pair-end column The P10_RNASeq_R2 does not have the label that its a reversed end	
III	Selecting the file to edit the metadata	
IV	Changed the paired-end to 2, as it's a reversed read	
V	The second error appears to be in the Case ID and Sample ID With the reversed sample's case ID - uk85AA and sample ID Sample_9	3.2 GiB uk85AA3B Sample_9_PGPUK_B9 3.6 GiB uk85AA Sample_8_PGPUK_B9

VI	Similarly to the above steps, I have changed the Case ID to math with the forward sample	**Case** Case ID: uk85AA3B Case UUID:
VII	Similarly, I have changed the Sample ID to Sample_9	**Sample** Sample ID: Sample_9_PGPUK_B9 Sample UUID: Sample type: blood
VIII	1. Selected all the input files 2. Set the batch ON	Input Read Files Change selection Batch by: File metadata This task will be batched by file metadata (Sample ID) and this will create 5 groups. Sample_10_pgpuk_b10 (2 items) Sample_6_pgpuk_b6 (2 items) Sample_7_pgpuk_b7 (2 items) Sample_8_pgpuk_b8 (2 items) Sample_9_pgpuk_b9 (2 items)
IX	3. Set the annotation 4. Referenced to the genome	Annotation GTF Change selection Batch by: None Homo_sapiens.GRCh37.75.gtf Reference FASTA File or Index TAR Bundle * Change selection Batch by: None human_g1k_v37_decoy.phiX174_Homo_sapiens.GRCh37....

X	The output settings are set for us to obtain 12 categories of output/ or tasks	**Output Settings** *Genes results* — *No value* *Isoforms results* — *No value* *RSEM Plot Model PDF File* — *No value* *Transcript BAM* — *No value* *flagstat_metrics* — *No value* *flagstat_metrics_1* — *No value* *report* — *No value* *report_zip* — *No value* *report_zip_1* — *No value* *sample_name_genome_bam* — *No value* *star_log_files* — *No value* *trimmed_reads* — *No value*
XI	Screenshot of running the batch	
XII	Created a new folder for the output	
XIII	There are 110 files within the whole output showing in the output folder, with each patient having 22 (including the genome-referenced files)	
XIV	For speeding up and execution purposes, the following parameters could have been turned off: ● "Output genome BAM" ● "Sort BAM by coordinate" ● "Calculate credibility intervals"	

XV	Looking specifically at the patient task of the sample There are 16 files generated as an output However, as the final output we would only use **five output files:** *1) Genes results* *2) Isoforms results* *3) RSEM Plot Model PDF file* *4) Transcript BAM* *5) Genome BAM*	au198 **Output Settings** **Genes results** Sample_10_PGPUK_B10.genes.results **Isoforms results** Sample_10_PGPUK_B10.isoforms.results **RSEM Plot Model PDF File** Sample_10_PGPUK_B10_plot_model.pdf **Transcript BAM** Sample_10_PGPUK_B10.transcript.sorted.bam **flagstat_metrics** Sample_10_PGPUK_B10.flagstat.txt **flagstat_metrics_1** _1_Sample_10_PGPUK_B10.flagstat.txt *report* *No value* **report_zip** P10_RNASeq_R1_fastqc.zip P10_RNASeq_R2_fastqc.zip **report_zip_1** P10_RNASeq_R2.trimmed_fastqc.zip P10_RNASeq_R1.trimmed_fastqc.zip **sample_name_genome_bam** Sample_10_PGPUK_B10.genome.sorted.bam **star_log_files** Sample_10_PGPUK_B10Log.progress.out Sample_10_PGPUK_B10Log.out Sample_10_PGPUK_B10Log.final.out **trimmed_reads** P10_RNASeq_R1.trimmed.fastq P10_RNASeq_R2.trimmed.fastq

To summarize the output I will be looking at all files of the sample 10

1.	***Gene expression results*** This file contains data and information about the genes detected in the RNA sequencing experiment. It typically includes gene expression levels and related statistics.	**Genes results** Sample_10_PGPUK_B10.genes.results
2.	***Isoforms results*** This file stores results related to RNA isoforms. It contains data and information about the various alternative splicing variants of genes detected in the RNA sequencing analysis.	**Isoforms results** Sample_10_PGPUK_B10.isoforms.results

3.	***RSEM Plot (PDF)*** This is the output file that contains a PDF representation of the plot model generated by the RNA-Seq by Expectation-Maximization (RSEM) analysis. It visualizes gene expression data.	**RSEM Plot Model PDF File** Sample_10_PGPUK_B10_plot_model.pdf
4.	***Transcript BAM file after alignment (transcript.sorted.bam)*** This is a BAM file containing transcript-level data, specifically aligned RNA-Seq reads that correspond to the transcribed regions of the genome. It's a critical intermediate file in RNA-Seq analysis.	**Transcript BAM** Sample_10_PGPUK_B10.transcript.sorted.bam
5.	***Quality check of BAM files after alignment (flagstat.txt)*** This TXT file contains metrics for the genome BAM file, providing alignment statistics for the entire genome	**flagstat_metrics** Sample_10_PGPUK_B10.flagstat.txt
6.	***Quality check of BAM files after alignment (flagstat.txt)*** This TXT file contains metrics generated by the 'flagstat' command for the transcript BAM file. It provides statistics on the alignment quality and characteristics of the transcript-level data.	**flagstat_metrics_1** _1_Sample_10_PGPUK_B10.flagstat.txt
7.	***Report (TXT)*** The report summarizing the results and findings of the RNA sequencing analysis, which may include quality control and statistics was NOT GENERATED	*report* *No value*

8.	***Quality check of input FASTQ files (fastqc.zip) for forward (R1) and reversed (R2) ends*** FastQC is used for quality control of RNA-Seq data, helping to assess the quality and characteristics of the sequencing reads.	**report_zip** P10_RNASeq_R1_fastqc.zip P10_RNASeq_R2_fastqc.zip
9.	***Quality check after trimming (trimmed_fastqc.zip)*** Similar to the previous item, this ZIP file array contains FastQC reports specifically generated after trimming the sequencing reads, helping to assess the quality of the post-processed data	**report_zip_1** P10_RNASeq_R2.trimmed_fastqc.zip P10_RNASeq_R1.trimmed_fastqc.zip
10.	***Genome BAM file after alignment (genome.sorted.bam)*** This BAM file contains RNA-Seq reads mapped to the entire genome. It represents the alignment of reads to the genomic regions.	**sample_name_genome_bam** Sample_10_PGPUK_B10.genome.sorted.bam
11.	***Log files (x3)*** These are log files generated during the alignment process using the STAR aligner. They provide information about the alignment and mapping of sequencing reads to the reference genome for RNA-Seq analysis.	**star_log_files** Sample_10_PGPUK_B10Log.progress.out Sample_10_PGPUK_B10Log.out Sample_10_PGPUK_B10Log.final.out
12.	***Trimmed FASTQ files (trimmed.fastq) for forward (R1) and reversed (R2) ends*** This is an array of file types, such as FASTQ and related formats, containing RNA-Seq reads after quality trimming and adapter removal. It represents the cleaned and processed sequencing data.	**trimmed_reads** P10_RNASeq_R1.trimmed.fastq P10_RNASeq_R2.trimmed.fastq