Homework Week 7

Friederike Dündar and Luce Skrabanek

ANGSD Course 2020

RNA properties (**5pts**)

1. Why are researchers interested in studying mRNA using RNA-seq, i.e. what can we learn from an RNA-seq experiment in contrast to a genomic DNA sequencing experiment? (**1pt**)
2. Explain the differences between coding and non-coding RNAs (**2pts**):
    - which functions do they have?
    - which structural features set them apart from each other? (Think, for example, about properties that might be exploited to enrich one population of RNA over another)
3. Why can genetic mutations in introns be harmful? Name two possible scenarios that involve different underlying mechanisms. (**1pt**)
4. Describe the difference between the transcription start site and the promoter. (**1pt**)

Alignment QC (**5pts**)

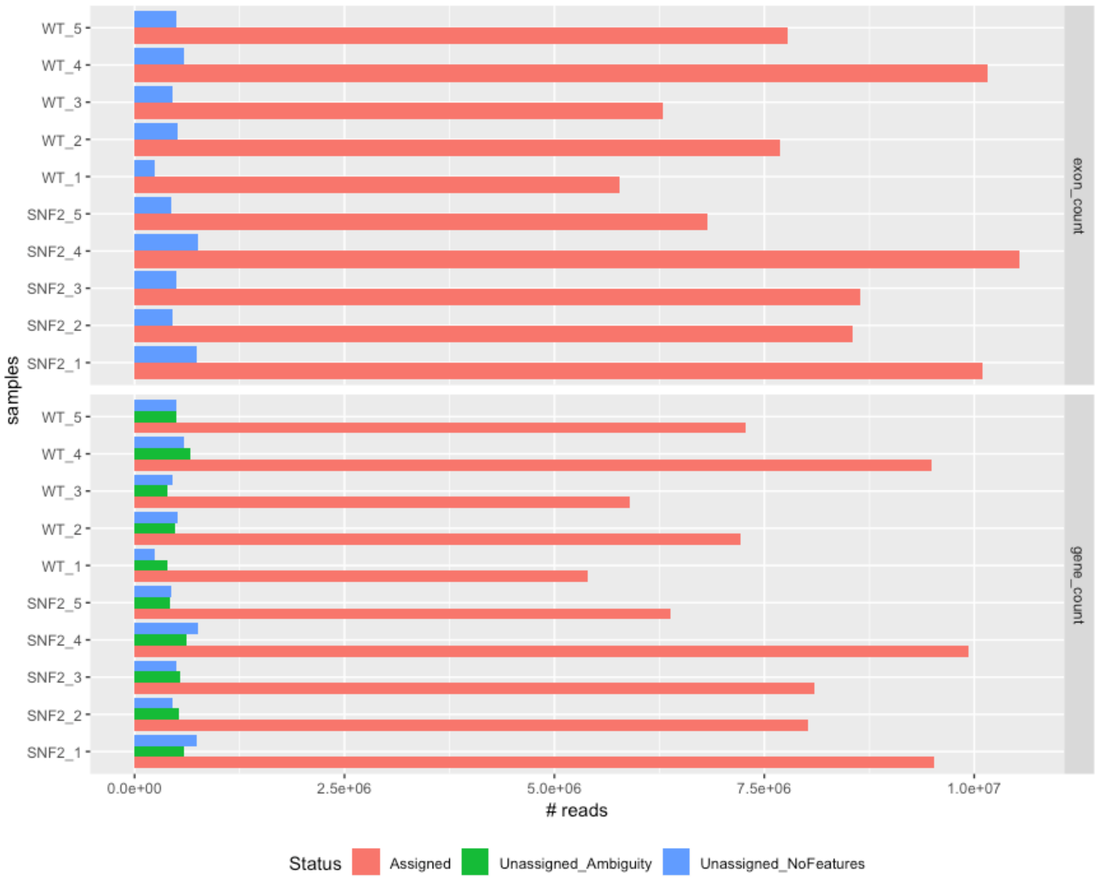
1. Which problem of mRNA-sequencing can be detected with `FastQC`? Explain how. (**2pts**)
    - rRNA contamination
    - mRNA degradation
    - genomic DNA contamination
2. Generate plots for (a) assessing the read distribution across exons/introns/etc. and (b) to determine the average gene body coverage. You can use either tool that we mentioned in class. You can use any of the Gierlinski BAM files that you have generated, or use one from `/home/frd2007/ANGSD_2019/alignment`. (**2pts**)
3. Why is a 3’ bias in the gene coverage plot evidence of RNA degradation? What type of enrichment step is particularly prone to overrepresenting the 3’ ends? (**1pt**)

Counting reads (**7pts**)

1. Use `featureCounts` to count the number of reads that overlap with every **exon**. As usual, keep track of all the commands. You can use the BAM files from `/home/frd2007/ANGSD_2019/alignment` if you don't have them in your home directory. Briefly explain at least 2 parameters (and their consequences!) that you're using (can include parameters set to default mode, but not parameters that specify input and output file names). (**2 points**)
2. Read the *summary* files generated by the `featureCounts` run shown during class and the one you just did into R. Generate a bar plot (using `ggplot2`) that displays the numbers of assigned and unassigned reads for either `featureCounts` run. The plot below is an example, you do not need to generate an exact replicate (you may also have run your `featureCounts` with slightly different parameters, so you may not even be able to replicate it exactly). (**2 points**)
3. Describe at least two observations from the plot including an explanation of what they mean. (**2 points**)
4. Download an annotation file (`GTF`) for one mammalian model organism of your liking. Determine the different types of loci that are annotated within that file and how many times each type is present in that file (you may want to look into the `uniq` UNIX command). (**1pt**)

Example plot based on featureCounts summary files

featureCounts was run in 2 modes: counting reads per genes or counting reads per exons



Compile the `.Rmd` file and send both the `.Rmd` and the `HTML` files to **angsd_wmc@zohomail.com** by Saturday night. If you need support, get in touch with Merv on Thursday, 3-4pm.
(*17pts*)