S. Mediterranea Gene Regulation of Brain Regeneration Analysis of RNA-Sequencing Data
[Project Proposal]

Name: Erikka Linn

Biological Question
Planarians have amazing regeneration powers as it can create a clone of itself complete with a new brain from a small tail piece. In contrast, humans are unable to restore lost neural tissue which is the main reason individuals suffer immensely from brain injuries/diseases in comparison to other organ tissues. This project will investigate the regeneration capabilities in the brain tissue of the *Schmidtea mediterranea* species from the Planarian family. Though not much research has been published in this field, this project looks to confirm the results of previous studies that have identified potential genes associated with the regulation of brain regeneration. The ultimate goals of this study are to see what gene(s) may regulate *S. mediterranea* brain regeneration and whether the expression of such differs from species like humans that cannot regenerate neurons well.

Dataset
The dataset will include 3 RNA-seq Illumina paired-end samples from *S. mediterranea (clonal line BCN-10)* at different time periods from being cut. Each of the three libraries selected were analyzed in the same batch of Illumina sequencing with accession number [E-MTAB-607]. Library 1 (ER32066: regenerative brain tissue, 30 min - 1 h, 4 animals), 3 (ER32068: regenerative brain tissue, 4 - 8 h, 4 animals), and 6 (ER32071: non-regenerative tissue, 0 h, 8 animals) from the samples provided by the study: https://genomebiology.biomedcentral.com/articles/10.1186/gb-2011-12-8-r76#Sec13. Since each library is originally around 1.4 GB each in fastq file format, they will be down-sampled to around 400 - 500 MB each, totally at 1.5 GB for all libraries. Additionally, this same study provided the fasta files of genes it identified to regulate brain regeneration: smed-runt-like1 [GenBank:JF720854] and smed-egr-like1 [GenBank:JF914965] and will be used for later cross-checking in VarScan. Libraries 1, 3, and 6 will be aligned in TopHat with the fasta file, *S. mediterranea* reference genome [SRX021585] sized 430 MB from the study: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0015617. Depending on the time left and genes identified, data like Human EGR-1 RNA-seq data: [SRX2442361] around 150 MB in fasta file format from other species will be found and used for a multiple sequence alignment.

Bioinformatics pipeline
The pipeline for this project will be similar to the Week 4 lab: *'Why don't snakes have legs?'*. First, the RNA-seq data will be down-sampled via **seqtk** and quality will be checked using **fastqc**. A custom index for *S. mediterranea* will be made in **bowtie2** and the libraries will be aligned to the reference genome [SRX021585] by using **TopHat. Kallisto** and **sleuth** will then be used perform a differential gene analysis to identify potential genes that may regulate brain tissue regeneration. Finally, these genes will be compared to the reference genes (med-runt-like1 [GenBank:JF720854] and smed-egr-like1 [GenBank:JF914965]) via **VarScan** as well as visually inspected in an *S. mediterranea* specific genome browser, **smed454**. If there is enough time and depending on the genes identified, a promising gene enhancer region will be looked at more in depth in the genome browser using ChIP-seq data and/or a multiple sequence alignment of enhancer sequences from different organisms will be run with mafft to see if there are any major regional mutations that might be correlated with brain regeneration.