Hello to all R workshop participants!

There’ll be 2 exercises that we’ll cover during my R workshop session. Here’s some instructions on how to get your system ready prior to the session.

Exercise 1: Visualization of multiple RNAseq datasets in KEGG

Purpose: We'll take RNA-seq data quantifying transcript abundances in treatment vs control conditions for 1000s of genes and visualize them in various molecular pathways using KEGG (Kyoto Encyclopedia of Genes and Genomes). This is superior to the Search&Color visualization tool provided online (<http://www.genome.jp/kegg/tool/map_pathway2.html>) since it 1) enables much finer visualization control, 2) allows us to visualize >1 dataset per pathway graph, and 3) allows for easy customization.

We’ll need to install 2 packages called ‘pathview’ and ‘KEGGREST’. ‘KEGGREST’ has a nifty function for retrieving KEGG pathway IDs that we’ll want to use. ‘pathview’ is required for the actual visualization of the KEGG pathways. You can install these packages using the Bioconductor installer (copy the following 3 lines into RStudio and run them):

source("https://bioconductor.org/biocLite.R") # gets the newest Bioconductor installer

biocLite("KEGGREST") # installs the package 'KEGGREST' (case-sensitive)

biocLite("pathview") # installs the package 'KEGGREST' (case-sensitive)

There’s also 2 additional, optional packages (‘stringr’ and ‘gage’), both of which provide alternative to ‘KEGGREST’ for retrieving the KEGG pathway IDs. You can install both of these in a similar fashion:

source("https://bioconductor.org/biocLite.R") # gets the newest Bioconductor installer

biocLite("stringr")

biocLite("gage")

Exercise 2: CDS extraction from contig sequences

Purpose: We've identified some contigs of interest in our RNA-seq transcriptome. Contigs are contiguous sequences that correspond to partial or full transcripts, and thus include untranslated regions and coding sequences (CDSs). For many down-stream applications we are interested in using the CDS only. Identifying the CDS is a laborious process that involves multiple steps including alignments against orthologs using blast, identifying the exact start and stop nucleotide positions, and trimming both the 5’ and 3’UTRs (if they are present). This becomes tedious when you have to process a decent number of contigs, for example for designing primers to use in real-time PCR validation of your RNA-seq data. Other uses for the CDSs include protein sequence prediction, ortholog alignments and phylogenetic tree construction, and construction of protein expression plasmids. This tool can be used in conjunction with a local blast database to greatly facilitate the identification of CDSs.

We’ll be working with nucleic acid sequences, which R does not understand how to manipulate natively. In order to be able to perform manipulations such as reverse-complementing we’ll be using the ‘Biostrings’ package to handle those sequences. You can install the package using the following lines:

source("https://bioconductor.org/biocLite.R")

biocLite("Biostrings")

For the purposes of this exercise we have already retrieved the sequences for several contigs. You will also be provided already with a file containing results from a local blastx query against the non-redundant protein sequences (nr) database for all of these sequences. If you don’t have a local blast database set up in your lab but are interested in doing let me know and I can provide you with a how-to manual that explains the process. You can also use the online blastx tool (<http://blast.ncbi.nlm.nih.gov/Blast.cgi>) though it will be more laborious since you have to manually determine the start and end nucleotide positions for the CDS.