Midterm assignment
Practical Introduction to Metagenomics
AS 410.734
Due date (strict): October 23rd at midnight

The midterm is a practical project in which you'll use the data sources and analysis approaches covered so far to create a re-usable utility. Specifically, you're going to create a fragment recruitment plot generation tool. I'm going to try to be verbose here for the sake of clarity of the assignment and head off as many questions as I can.

Fragment recruitment plots were introduced in Week 02 (Global Ocean Survey paper) and then covered more fully in Week 06. There is no specific protocol you must follow. Rather, the general idea is that you have a set of assembled reference genomes upon which you'll align metagenomic reads sampled from your habitat of interest. You'll then write a re-usable script to process these alignment results and generate summary plots.

As you've seen by now, I try to approach the assignments in this course in a way that resembles how you'll face analysis tasks in your careers - it is up to you to choose the methods, languages and tools to complete your analysis. You can use any resources at your disposal as long as you honestly do the work yourself, and I will help guide you as needed along the way, providing options, suggestions and, where necessary, explanations.

There will be 3 major aspects of this assignment:

1. Data gathering
2. Alignment of reads to reference genomes
3. Generation of fragment recruitment plots using alignment data
4. Formal write-up

Data gathering:

Metagenomic sequences

In the homework for Week 05 you chose a metagenomic project of interest and described your chosen habitat. One of the criteria was that data be publicly available and accessible. Place the FASTQ or FASTA files in your project area here:

~/project/samples/

You can create whatever directory structure under there which best fits your data. How much should you download? You need a good representation of your data but you should probably stop at around 10GB if your sample site has that much. Some habitats will have far less (like air samples), so just get everything you can for those. If you're feeling industrious, and want to process *more* sequences to get a more complete picture that's OK, DIAG has the resources, but attacking too much data won't be a valid excuse for late submissions. It's wise to test your method with a smaller dataset and then scale up.

Reference genomes

These should have been chosen, explained, and downloaded in Week 05's homework. I'm posting this mid-term the day after that assignment was due, but I'll try to get to the grading in the next few days. In the meantime, if you have any questions or concerns that you got the right data to get started with the midterm please e-mail me.

Alignment of reads to reference genomes

We've covered a few ways to do this already, including FT-HIT, Genometa, and BLAST (GOS paper). You can choose any algorithm you want to do this as long as you can explain it in your write up and compare to at least one other alignment method. Most of these tools are installed under the /diag/software directory, but let me know if there's another you wish to try and would like centrally installed.

Algorithms:

FT-HIT
BLAT
Bowtie
Genometa

Each of these have their own native output formats, and you *can* write your plotting tool to expect that format, but it then ties your plotter to that tool. If, instead, your script expects some independent format, such as SAM (http://bioinformatics.oxfordjournals.org/content/early/2009/06/08/bioinformatics.btp352.full.pdf) then your tool will be more generally useful. Most of the alignment tools above either have options to export SAM format or come with conversion scripts.

Choose at least two read-to-reference genome alignment methods to run and compare how many reads were successfully aligned. The publications we've had in the course so far have specified runtime options to use for several of them, so your write-up should include documentation of the command-line parameters used and why you chose them.

Generation of fragment recruitment plots (FRPs) using alignment data

Your FRP creation script should parse the read alignment data from the step above to generate the plot. Again, you can use any languages or visualization method you wish here.

This has been detailed in a few of the papers we've covered, but the traditional FRP has percent identity on the X-axis and position in genome on the Y-axis. You parse the alignment scores for each read and mark the plot in the appropriate position. You should feel free to expand on this or modify it as you see fit for your habitat.

Your write-up should describe the methods used to generate the plots as well as thoroughly document their visual format layout for external users.

How to make these? That will probably be where you spend the most time in this assignment. It depends completely on the language and approach you want to take.

Perl – There are many graphics modules which mamy be appropriate here:
Chart::Graph::Gnuplot, GD::Graph, etc.

Python – Like perl, there are many libraries that can make plotting easy. You are free to use any you like, but one favorite is matplotlib.

Gnuplot – http://www.gnuplot.info/ - Gnuplot is a powerful command-line utility for drawing and displaying graphics.

R – This is magic to me and I'm guessing you could use it here. It's outside my realm of expertise so if you think you can use it, go for it.

Javascript – You truly can use anything you want for visualization, so if you're an expert at web development and want to use any one of many Javascript frameworks to display these data you can.

As the week carries forward I will post a few skeleton example scripts in some of these languages in case you need a primer, though a quick Google search will provide tutorials with demos for most of them.

Formal write-up and deliverables

Deliverables:

- Path and description of metagenomic sequence data files. If you chose a subset of a large sample, explain why and how the subset was created.
- Description of algorithms you chose to align the reads to your reference genomes along with the options/parameters used. You should also describe the output format for each that you used.
- A well-written, documented script that reads a read-to-reference alignment result file and produces a fragment recruitment plot.
- A write-up which includes these things as well as the analysis of your sample habitat after using the script on it. It should include screenshots or embedded images of your FRPs. Which of your reference genomes is most represented in your sample? What percentage of the genome is covered by your reads? If you used multiple samples from your chosen microbiome did their community profiles differ? What did you learn from this process?