December 2018

Assignment 3, Genome Informatics Module, Computational Biology MPhil

This assignment is to be carried out individually.

Due by **11.45pm Friday 14th December**. This is a fixed deadline and there will be no extensions as per the MPhil handbook.

Style: This assignment should be written in the style of a scientific paper, with sections for abstract (max. 250 words), introduction, results, discussion and methods.

Page limit: There is a page limit of 5, this is the absolute maximum and any additional pages will not be read. Please keep answers short, and use figures and tables where appropriate. It is in your interests to be lucid and concise, and to think about the most informative way to present your findings - i.e. summarise results rather than show all the details.

Submission: write-ups and code must be submitted via Moodle. Submit your write-up as a pdf file named gi3_eadc2.pdf, where eadc2 is your crsid and your code as a separate text file. The code must also be documented in the write-up, where it will not count against your page limit. **EVERYTHING ELSE WILL**.

If you have any technical difficulties with the assignment please email Dr Alastair Crisp (acrisp@mrc-lmb.cam.ac.uk).

Assignment Description

For this assignment you will be given a gene and have to investigate it using the methods from this course. Use at least two methods that fit into at least two of the following three categories (some examples listed, methods may cross categories).

Functional annotation

e.g. Identify functional domains in your gene and its orthologs in different species or paralogs in this species. How do the domains vary between homologs? Has the function of the gene varied across species / with time? How does it vary between paralogs?

Expression

e.g. Map reads to the genome to determine how the expression of your gene varies over time or across different tissues or how the expression of orthologs varies in different species or expression of paralogs in this species.
e.g. Determine which Transcription factors have binding sites in the 500 bp upstream of your gene and link this to gene function or spatial/temporal patterns of gene expression. Do the same for homologs of your gene.

Variation / Phenotype

e.g. Locate SNPs in/near your gene (all types of SNP) or that influence the expression of your gene. Link these SNPs to phenotypes where possible. Do the same for homologs of your gene.

Combine data across techniques to ”tell a story” about the genes.

You should use papers to guide your work, but you must perform some computational analysis yourself. You will not necessarily get the same results as published work, in this case comment on the differences and why they occur.

You do not have to provide data from all methods you use (if more than two), though briefly mentioning them and adding ”(data not shown)” may be appropriate, some results will be more interesting than others and

you should use your limited space accordingly.

To clarify, I would describe the two examples for the Expression category to use/be different methods, though obviously they belong to the same category.

Help and Guidance

Connecting to the server

You can connect to the server using secure shell (SSH)
From the linux command line you connect with the following command.
ssh <username>@subliminal.maths.cam.ac.uk
Windows users can use PUTTY `http://www.chiark.greenend.org.uk/~sgtatham/putty/`
There are plenty of HOW-TO guides online.

Using (or being) nice

To promote the sharing of resources when multiple users are trying to use the same server you can add:
nice -n x
where x is a number from 1 to 19 (1 highest priority) and the server will assign CPU time based on the priorities. This means you can run your programs using all 64 cores of subliminal (where possible) without worrying about preventing other people using the server and your programs will automatically fill the avaliable CPUs if other programs finish. This only works well if EVERYONE uses it. I would suggest you all use the same priority, say 5.

SFTP - SSH file transfer protocol

SFTP can be used to move files to and from the server.

GNU Screen

Screen is a terminal multiplexer which allow processes to continue running even after the client disconnects from the server. You can then reconnect to the screen session when you next login. After SSHing into the server, type 'screen' to start a new screen session. There are numerous screen tutorials online to help you use its full functionality.

Obtaining Data

For details on human genes see http://www.genenames.org/

Sequence data and annotation (where avaliable) for each species may be obtained from (among other sites) the list of websites found at the end of Lecture 5 and in the topic specific lectures.

For functional annotation all InterProScan databases may be searched at:
http://www.ebi.ac.uk/interpro/search/sequence-search

For tissue- and time-point-specific reads ENCODE and MODENCODE are your best sources.

For transcription factor motifs there are many databases, e.g.: http://jaspar.genereg.net/
http://motifmap.ics.uci.edu/
http://floresta.eead.csic.es/footprintdb/?databases (contains JASPAR)

Bioinformatics Software

You may use any programs you like. The following programs are avaliable in
/local/data/public/genome_informatics_2018/programs/ :
BLAST
HMMer
Tuxedo suite - n.b. To use TopHat Bowtie2 must be in $PATH - add
export PATH=/local/data/public/genome informatics_2018/programs/bowtie2-2.2.9:$PATH to your bash_profile