Bioinformatics 6110: Genomic Methods for Bioinformatics

Course Outline

Please, prior to the first class period, apply for a SHARCNET account. Please send your SHARCNET username to Megan via email. When you apply, please use Prof. Lewis Luken's ID (sug-385-01) as the reference.

Class Dates: Classes are Tuesdays, 1:30 – 4:30 pm in SSC 3310.

Course coordinator: Steffen Graether (graether@uoguelph.ca). Email me if you ever have any questions or concerns about the course.

There will be three modules:

1. ChiP-seq: analysis of protein binding to promoter sequences

2. RNA-seq: transcriptome analysis

3. Computational methods for protein structure

Module 1: ChIP-seq analysis with Zhenhua Xu: zxu@uoguelph.ca Please prior to the first class period register for an iplant account (http://www.iplantcollaborative.org/) and send your user name for this account to Zhenhua via e-mail

Data for this module is based on:
Hudson D, Guevara D, Yaish MW, Hannam C, Long N, Clarke JD, Bi YM, Rothstein SJ. (2011). *GNC* and*CGA1* Modulate Chlorophyll Biosynthesis and Glutamate Synthase (*GLU1/Fd-GOGAT*) Expression in *Arabidopsis*. PLoS ONE 6(11): e26765.

In this paper, the authors used Arabidopsis (*Arabidopsis thaliana*) to study the functions of two GATA transcription factor, *AtGNC* and *AtCGA1*, in regulating chlorophyll biosynthesis. The myc-tagged GNC and CGA1 proteins were expressed in Arabidopsis. The myc-tagged overexpression lines were used for ChIP-DNA preparation and this DNA was sent for ChIP library construction and high-throughput sequencing. We will be only using the data generated from AtGNC-ChIP-seq to see what the procedure of ChIP-seq data analysis is and how to use ChIP-seq data to find downstream target genes of a transcription factor.

January 13: Discussion of the basics of gene regulation, promoter-DNA binding and searching for biological relevance; data files used for analysis and the iplant interface; sequence mapping to the reference genome; removing multiple mapped reads; peak calling.

January 20: Peak annotation; peak visualization; motif search; GO and Pathway analysis

January 27: **TENTATIVE:** Tour of genomic facilities on campus: Biodiversity Institute: at 1:30PM; Advance Analysis Centre Genomics Facility in Science Complex: at 3:00PM

Project 1: Analysis of ChIP-seq data: Using the data analysis you have done in class and your reading of the relevant literature address the following: what are the issues in the data generated from ChIP-seq data and how can these be addressed; based on the literature pick 5 target genes and try to build a regulatory model for this transcription factor; discuss the methods you would use to verify the data analysis. Your paper should be no more than five pages, double-spaced not including figures and references, and should follow standard journal format (Introduction, Materials & Methods, Results, Discussion, References). **(Due February 9 at 4 PM)**

Reference of the tools:

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.
(http://bowtie-bio.sourceforge.net/bowtie2/index.shtml)

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. *Nat. Biotechnol. 29*, 24–26.
Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinform. 14, 178–192. (http://www.broadinstitute.org/igv/)

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W and Liu XS. (2008). Genome Biology 9:R137
(http://liulab.dfci.harvard.edu/MACS/)

Chen TW, Li HP, Lee CC, Gan RC, Huang PJ, Wu TH, Lee CY, Chang YF and Tang P. (2014). *BMC Genomics*. 15:539 (http://chipseek.cgu.edu.tw/)

Du Z, Zhou X, Ling Y, Zhang ZH and Su Z. (2010). Nucl. Acids Res. 38 (suppl 2): W64-W70.
(http://bioinfo.cau.edu.cn/agriGO/)

Thimm O, Bläsing O, Gibon Y, Nagel A, Meyer S, Krüger P, Selbig J, Müller LA, Rhee SY and Stitt M. (2004). The Plant Journal. 37(6), 914–939.
(http://mapman.gabipd.org/web/guest;jsessionid=48F423C6E155DA04FDD11A090CB64738.ajp13_mapman_gabipd_org)

Module 2: RNA-Seq with Megan House: housem@uoguelph.ca. Please, prior to the first class period, apply for a SHARCNET account. Please send your SHARCNET username to Megan via email. If you are using a computer with a Windows operating system, you will need to install PuTTY (http://www.chiark.greenend.org.uk/~sgtatham/putty/) and WinSCP (https://winscp.net/eng/index.php) prior to the first class period (you will need these to

access Sharcnet and transfer files between Sharcnet and your computer). Please bring a laptop to all sessions. Please also install R (https://www.r-project.org/) on your computer prior to March 9th, and Mapman (http://mapman.gabipd.org/web/guest/home) prior to March 16th.

Data for this module is based on:

Horvath, D.P., Hansen, S.A., Moriles-Miller, J.P., Pierik, R., Yan, C., Clay, D.E., Scheffler, B. and Clay, S.A. (2015). RNAseq reveals weed-induced PIF3-like as a candidate target to manipulate weed stress response in soybean. *New Phytologist*, *207*(1): 196-210.

In this paper, the authors used soybean (*Glycine max*) to study the effect of weed competition on transcript abundance. Weed competition has been studied with respect to plant morphology and physiology in detail, however, the molecular basis for observed morphological and physiological differences are not well understood. We will be performing analyses to identify transcriptional changes that occur, particularly focusing on such biological processes as photosynthesis, biotic defense, and ROS (reactive oxygen species) scavenging.

February 3: RNA, Biological Questions, and Unix. A review on RNA and transcription. A review of research hypotheses and objectives, and how RNA-seq can be used to address these. Using UNIX in a High Performance Computing (HPC) environment; Navigating SHARCNET. Hands-on practice using basic command line.

February 10: Next generation sequencing, the RNA-seq pipeline, and Quality Control. Sequencing methods used to obtain short RNA reads. An overview of the main pipeline used for RNA-seq analysis. Next generation sequencing technology, terminology, and data structures; Quality control of RNA-Seq data; Hands-on quality assessment of RNA-Seq data using FastQC.

February 17: Reading week, no class.

February 24: Aligners, Bowtie and TopHat. How to map short reads to a genomic scaffold. Finding and using reference genomes; Hands-on assessment of Tophat output files using SHARCNET.

March 2: Gene Expression. Review of genes, isoforms, and splice variants. Quantification of mRNA using RNA-Seq data; Normalization of data, how read depth and gene length affect read counts; FDR correction for multiple testing. Hands-on assessment of Cuffdiff output files on SHARCNET to differences in gene expression between two samples.

March 9: HT-Seq and EdgeR. An overview of more complicated experimental designs that involve more than two samples. Using HT-Seq to generate raw read counts, and use of EdgeR to analyze raw read counts. Hands-on practice using EdgeR to identify differentially expressed reads.

March 16: Understanding the Biology using gene ontology and Mapman. Identification of genes with a shared biological role using GO terms. Determination of the biological function of genes using mapman; understanding the mapman mapping file. Hands-on analysis of GO and Mapman terms to assess the function of differentially expressed genes.

Project 2: Analysis of RNA-Seq data for differential expression. Each student will be provided with the same raw read counts, and each student will be assigned a unique research hypothesis. After determining an appropriate model, each student must analyze the data using edgeR to identify differentially expressed genes. You will then use a variety of tools (such as AgriGO and Mapman) to identify the genes (or groups of genes) that pertain to your research hypothesis. For instance, if your hypothesis is related to photosynthesis, you will GO and mapman terms to identify genes related to photosynthesis. Be sure to place your findings in the context of established biology, and describe how your findings relate to other relevant work in the field. Your paper should be no more than 15 pages, double-spaced not including figures and references, and should follow standard journal format (Abstract, Introduction, Materials & Methods, Results, Discussion, References) with figures attached at the back of the document. **(Due March 29 by 4 PM - please submit via the course Dropbox)**

Software used:

FastQC:
Andrews S: FASTQC. A quality control tool for high throughput sequence data. [http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/].

Bowtie and Bowtie2:
Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Tophat, Tophat2, and Cuffdiff:
Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. *7*, 562–578.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. *14*, R36.

HTSeq:
Simon Anders. HTSeq: Analysing high-throughput sequencing data with Python. http://www-huber.embl.de/users/anders/HTSeq/, 2011

EdgeR:
Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.

Mapman:
Thimm, O., Bläsing, O., Gibon, Y., Nagel, A., Meyer, S., Krüger, P., Selbig, J., Müller, L.A., Rhee, S.Y., and Stitt, M. (2004). mapman: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J. *37*, 914–939.

Usadel, B., Nagel, A., Thimm, O., Redestig, H., Blaesing, O.E., Palacios-Rojas, N., Selbig, J., Hannemann, J., Piques, M.C., Steinhauser, D., et al. (2005). Extension of the Visualization Tool MapMan to Allow Statistical Analysis of Arrays, Display of Coresponding Genes, and Comparison with Known Responses. Plant Physiol. *138*, 1195–1204.

Module 3: Structural bioinformatics with Steffen Graether: graether@uoguelph.ca

One week prior to the first lecture, download and install the following software:

Pymol: http://pymol.org/ep

Modeller: http://salilab.org/modeller/registration.html You will need to register to use Modeller. You will also need to download the sample data set at http://salilab.org/modeller/downloads/pdball.pir.gz

Should you have any difficulty with software or downloads, please email me.

March 23: A brief introduction to protein structures and structural methods (secondary and tertiary structure, X-ray crystallography and protein NMR)

March 30: Structural bioinformatics (protein structure visualization, structure analysis)

April 6: Modelling of protein structures (homology and *ab initio* modelling)

Background reading and software used: The readings are meant to help those students with a weaker biochemistry background.

Protein structures: Any introductory biochemistry textbook that discusses amino acids, protein sequences, secondary and tertiary structure.

X-ray crystallography:
http://chemwiki.ucdavis.edu/Analytical_Chemistry/Instrumental_Analysis/Diffraction/X-ray_Crystallography

Protein NMR review: Barrett, P. J. et al. The Quiet Renaissance of Protein Nuclear Magnetic Resonance. Biochemistry 52, 1303–1320 (2013).

Pymol: http://www.pymolwiki.org/index.php/Main_Page

Modeller: N. Eswar, M. A. Marti-Renom, B. Webb, M. S. Madhusudhan, D. Eramian, M. Shen, U. Pieper, A. Sali. Comparative Protein Structure Modeling With MODELLER. Current Protocols in Bioinformatics, John Wiley & Sons, Inc., Supplement 15, 5.6.1-5.6.30, 2006.

Project 3: Visualization of protein structures and homology modelling. The student will select a protein sequence with 50-75% sequence identity (over the entire sequence) with a solved protein structure. Students will select their structure themselves. The write up should be a brief report and include background information on the protein you have selected, figures of the modelling result and analysis of the structure, and a basic discussion of the results. The total written report should be no longer than five pages, not including the references. **(Due April 20th at 4 pm)**

Assessment:

First project: 25%
Second Project: 40%
Third Project: 25%
Courselink Quizzes (online): Due no later than 11:59 pm on Feb 16 (quiz 1), Mar 1 (quiz 2), Mar 8 (quiz 3), Mar (quiz 4) and Mar 22 (quiz 5), 10% total

Assignments are due as noted in the course outline. Late assignments will be penalized at 10% per day late or less. Assignments will not be accepted after 5 days. If there is a valid reason why this cannot be achieved see below for the university guidelines.

University Policies

When You Cannot Meet a Course Requirement
When you find yourself unable to meet an in-course requirement because of illness or compassionate reasons, please advise the course instructor (or designated person, such as a teaching assistant) in writing, with your name, id#, and e-mail contact. See the graduate calendar for information on regulations and procedures for Academic Consideration: http://www.uoguelph.ca/registrar/calendars/undergraduate/current/c08/c08-ac.shtml

Accessibility
The University of Guelph is committed to creating a barrier-free environment. Providing

services for students is a shared responsibility among students, faculty and administrators. This relationship is based on respect of individual rights, the dignity of the individual and the University community's shared commitment to an open and supportive learning environment. Students requiring service or accommodation, whether due to an identified, ongoing disability or a short-term disability should contact the Centre for Students with Disabilities as soon as possible.

For more information, contact CSD at 519-824-4120 ext. 56208 or email mailto:csd@uoguelph.ca or see the website: http://www.csd.uoguelph.ca/csd/

Academic Misconduct

The University of Guelph is committed to upholding the highest standards of academic integrity and it is the responsibility of all members of the University community – faculty, staff, and students – to be aware of what constitutes academic misconduct and to do as much as possible to prevent academic offences from occurring. University of Guelph students have the responsibility of abiding by the University's policy on academic misconduct regardless of their location of study; faculty, staff and students have the responsibility of supporting an environment that discourages misconduct. Students need to remain aware that instructors have access to and the right to use electronic and other means of detection.

Please note: Whether or not a student intended to commit academic misconduct is not relevant for a finding of guilt. Hurried or careless submission of assignments does not excuse students from responsibility for verifying the academic integrity of their work before submitting it. Students who are in any doubt as to whether an action on their part could be construed as an academic offence should consult with a faculty member or faculty advisor.

The Academic Misconduct Policy is detailed in the Undergraduate Calendar: http://www.uoguelph.ca/registrar/calendars/undergraduate/current/c08/c08-amisconduct.shtml

E-mail Communication

As per university regulations, all students are required to check their <uoguelph.ca> e-mail account regularly. E-mail is the official route of communication between the University and its students.

Drop Date

The last date to drop one-semester courses, without academic penalty, is the 40th class day. To confirm the actual date, please see the schedule of dates in the Graduate Calendar.

Copies of out-of-class assignments

Keep paper and/or other reliable back-up copies of all out-of-class assignments: you may be asked to resubmit work at any time.

Recording of Materials

Presentations which are made in relation to course work—including lectures—cannot be recorded or copied without the permission of the presenter, whether the instructor, a classmate or guest lecturer. Material recorded with permission is restricted to use for that course unless further permission is granted.

Resources

The Academic Calendars are the source of information about the University of Guelph's procedures, policies and regulations which apply to undergraduate, graduate and diploma programs:
http://www.uoguelph.ca/registrar/calendars/index.cfm?index