Biomedical Big Data for ISU (BD4ISU)
2018 Summer Workshop Schedule

On days when workshops are being held for BD4ISU, the day is split between (a) time for the presenter to give a lecture and demonstration, and (b) lab time for participants to work through examples to solidify their learning. Lecture/demo are normally from 9-10:30am and 1-2:30pm; lab time is normally from 10:30am-noon and 2:30-4pm. Recorded lecture/demo videos will be posted to the BD4ISU blackboard site. A list of questions and answers from the Q&A time will be posted to the BD4ISU website or blackboard site.

Note – video lectures from the workshops are posted in a Blackboard site for the project. If you are an ISU student, faculty, or staff please write to us to ask to be added to the course.

Other Events

We keep a log of other events that the BD4ISU-supported students have taken part in.

- SURE Friday presentations – Fridays at noon from May 22 – August 2
- Meeting research team (Drs. Amaad, Schwab, Steding, Science Building) – May 29, 1-4pm
- Lab training (Dr. Steding, Science Building 213) - May 30 and June 6, 1-4pm
- The Ohio State University seminar (viewed remotely) - Thursdays at noon
- Advising checkin (Dr. Latimer) - Thursdays at 2pm
- SURE Symposium – Thursday August 2

Research Integrity Symposium

Date/time, location – May 21, 8am – 5pm in Science room 214

Organizers – Andrew-Sheppard Smith (Office of Sponsored Programs, Director),
Rusty Gonser (Professor, Dept Biology, Director of TCGA),

Lunch is provided for those who pre-registered. Organized by Andrew Sheppard-Smith and Rusty Gonser. Note that the day follows a different schedule than the normal BD4ISU workshop schedule. The day is broken into the following sessions; the presenter is listed for each.

- 8:00 - *Mentor/Mentee Responsibilities*
  Chris MacDonald, Professor, Dept of CDCSEP
- 9:00 - *Human Subject Research*
  Ryan Donlan, Interim Chairperson, Associate Prof, Dept of Ed Leadership, IRB Chairperson
- 9:30 - *Research Misconduct*
  Andrew Shepard-Smith, Director, Office of Sponsored Programs
- 10:30 - *Conflict of Interest*
  Andrew Shepard-Smith, Director, Office of Sponsored Programs
- 11:00 - *Export Control*
  Andrew Shepard-Smith, Director, Office of Sponsored Programs

- 12:00 - *Q and A session*
  Rusty Gonser, Professor, Dept of Biology, Director of TCGA
- 1:00 - *Data Use and Misuse*
  Andrew Shepard-Smith, Director, Office of Sponsored Programs
- 1:30 - *Data Acquisition, Management, Sharing, and Ownership*
  Matt Jenkins, Info Sec Security Officer, OIT
- 2:30 - *Biosafety and/or Chemical Hygiene*
  Catherine Steding, Assistant Professor, Dept of Biology
- 3:00 - *Grant Fraud*
  Andrew Shepard-Smith, Director, Office of Sponsored Programs
- 4:00 - *Animal Subject Research*
  Rusty Gonser, Professor, Dept of Biology, Director of TCGA
- 4:30 - *Peer Review*
  Shaad Ahmad, Assistant Professor, Dept of Biology

An Introduction to Web-based Bioinformatics

Date/time, location – May 23-24 9am-4pm, May 25 9am-noon, in University Hall 008R

Presenter – Jennifer Smith, microbiology

This workshop focuses on the usage of freely available, web-based bioinformatics programs and the interpretation of the data. We will explore a variety on concepts, such as how to obtain genome sequences online, how to run alignments, and how to visualize data.

Prerequisites – bring a laptop computer to work through examples. If possible, install the following software on your computer before the workshop starts -

- Putty - https://www.chiark.greenend.org.uk/~sgtatham/putty/latest.html
  - Note: if you use Linux or MacOS, you don't need Putty. On MacOS you will open the Terminal program, which you get to by opening Finder, then Applications, then Utilities, then Terminal.
- Notepad++ - https://notepad-plus-plus.org/download/
  - Note: for MacOS or Linux, use a different text editor (Notepad++ is only for Windows). A good cross-platform free text editor is Atom - https://atom.io
- FileZilla - https://filezilla-project.org/download.php?type=client
- Tablet from the James Hutton Institute - https://ics.hutton.ac.uk/tablet/download-tablet/

Resources that will be used

- NCBI (National Center for Biotechnology Information) - https://www.ncbi.nlm.nih.gov
- Flybase - http://flybase.org
- TimeTree - http://timetree.org
- Clustal/ClustalW - http://clustal.org
- Galaxy - https://usegalaxy.org
- Ensembl Genome Browser - https://useast.ensembl.org/index.html
- Ensembl Genomes - http://ensemblgenomes.org

- Circos Circular Genome Visualization - http://circos.ca
- Weblogo sequence logos - https://weblogo.berkeley.edu/logo.cgi
- Protein Data Bank - https://www.rcsb.org
- UniProt - http://www.uniprot.org
- BOLD - http://www.boldsystems.org
- Gene Expression Atlas - https://www.ebi.ac.uk/gxa/home
- Gene Ontology - http://pantherdb.org
- DAVID Gene Ontology - https://david.ncifcrf.gov

The rough schedule is the following.

- May 23 AM - *Intro, assemblies, NCBI, alignments*
- May 23 PM - *CLUSTALW, phylogeny, Time Tree*
- May 24 AM - *Data acquisition - Ensembl, NCBI, Galaxy, sequencing, SRA and NCBI*
- May 24 PM - *Genomic variants, Visualization - Tablet, Weblogo*
- May 25 AM - *Proteins, BOLD, Gene Expression*

R Programming and Unix/Linux

Date/time, location – May 29 - May 31, 9am-4pm

Presenter – Jeff Kinne, Associate Prof, Dept of Math & Computer Science, Associate Director of TCGA

Prerequisites – bring a laptop computer to work through examples. If possible, install the following software on your computer before the workshop starts -

- R programming package (version 3.5) - https://cloud.r-project.org
- R studio (must install R programming package first) - https://www.rstudio.com/products/rstudio/download/#download
- Putty - https://www.chiark.greenend.org.uk/~sgtatham/putty/latest.html
    - Note: if you use Linux or MacOS, you don't need Putty. On MacOS you will open the Terminal program, which you get to by opening Finder, then Applications, then Utilities, then Terminal.
- FileZilla - https://filezilla-project.org/download.php?type=client

Resources that will be used -

- R for Data Science online book - http://r4ds.had.co.nz
- Harvard online PH525x series - Biomedical Data Science - http://genomicsclass.github.io/book/
- R on tutorialspoint - https://www.tutorialspoint.com/r/index.htm
- R introduction from R project - https://cran.r-project.org/doc/manuals/R-intro.html
- R reference card - https://cran.r-project.org/doc/contrib/Short-refcard.pdf
- R cheat sheet - https://www.rstudio.com/wp-content/uploads/2016/10/r-cheat-sheet-3.pdf
- Other cheat sheets - https://www.rstudio.com/resources/cheatsheets/

The rough schedule is *(still working out the precise list of topics)* the following -

- Supplement – videos provided for reference to use as needed

    - Bits, bytes, data files -
    - Windows command prompt -
    - Unix/Linux basic shell commands -
    - R data types -
    - R importing data -
    - R scripts – running R from terminal
    - R markdown -
    - R statistics -
    - R Programming – syntax errors, logical errors -
    - R Programming – debugging - "play computer", KISS -
    - Case study – mutations and binding sites in genes, next-gen RNA-seq data
    - R bioconductor -
    - http://rafalab.github.io/pages/harvardx.html#series_2
- May 29 AM -
    - Getting setup and started
        - Rstudio versus R builtin editor, R – installing packages (tidyverse)
        - R – use as calculator
        - Any programming language – start a list of what punctuation does what, and what functions are pre-defined, …. Here is an example of what I mean. There is no replacement to you doing this yourself, so If you want to be an R programmer you need to start keeping your own notes...
            - Punctuation - * + - / arithmetic, = variable assignment, %/% integer division, ? For help
            - Functions – mean, c to create a vector,
            - Rstudio tips – arrows for history, alt- for <-, ? For help, tab for auto-complete
            - R data types – numbers, vectors (like a list of things), data frame (excel table – rows of data, each column is the same data type)
            - Programming lingo – function, variable, parameters,
            - Base R cheatsheet - http://github.com/rstudio/cheatsheets/raw/master/base-r.pdf
    - Data visualization (aka plots, figures, graphs)
        - ggplot2 – follow along http://r4ds.had.co.nz/data-visualisation.html
        - Continue following along in chapter 3 there, do exercises, we answer some of the exercises tomorrow.
    - R programming basics –
        - First - http://r4ds.had.co.nz/workflow-basics.html
        - Basic overview - https://www.tutorialspoint.com/r/index.htm
        - More detailed overview - https://cran.r-project.org/doc/manuals/R-intro.html
    - Advice learning programming -
        - Patience, and attention to detail, and don't panic, and don't stare at the computer for too long
        - Warning – this takes a lot of time, you are the experiment. You need to spend at least a few hours per day on programming for it to start sinking in...

        - R – it's all about knowing the packages and functions
    - Today's biology data
        - Install bioconductor (after installing R, Rstudio) - http://bioconductor.org/install/
        - Run biocLite("GEOquery")
        - Today's data – gene expression in microarray data
            - See separate document for info about the data
- May 30 AM -
    - Continuing from yesterday
        - First introduction to useful linux commands
            - Commands -
            - Linux quick start, linux cheat sheet
            - Get to a terminal / shell / console / command prompt (Windows)
            - Today – Linux and Mac
            - Directories and files ("doing" Finder or File Explorer) - pwd, cd .., ls, cp, rm, mkdir, rmdir, rm –R –f, ls –l –h, ls -ltr
            - Note – there is no recycle bin
            - Other – clear
            - Note – tab auto-completion, and up/down arrows
            - Simple text editor – pico or nano
            - More full-featured text editor (not GUI, do most of the things that Notepad++, etc. Do) – vim, emacs, jove
            - Searching/viewing text files – head, tail, grep "AAAA" mouse.fa | wc
            - Kill a program - ^C (that's ctrl-c)
        - Running rstudio from CS server
            - From terminal/shell/console on mac or linux, run
              ssh username@cs.indstate.edu
              ssh –Y username@cs.indstate.edu if you are going to use graphical programs (only do this from on campus since this takes a lot of bandwidth)
            - Then do
              rstudio &
        - R files – saving in R script, executing the file. Use Rscript command if in putty/terminal
        - Exercises from chapter 3 of R for Data Science
        - Import micro array data as csv, use ggplot2 and dplyr filtering to look at
            - On the CS computers
              cd /u1/junk/bd4isu
        - Logging into CS from home -
- May 31 AM -
    - Homework, things to do -
        - Keep looking in R for Data Science, and apply that to the gene expression data
        - Use R to identify genes that have high variation – statistics of each row

        - Picking groups of genes that are related – heart-related genes, brain-related genes – looking at the gene/row information and pick out some by hand, or use R to pick them out automatically
        -
    - R package - bioconductor installation
    - Ggploting
    - Look at cheatsheets again

Statistics

Date/time, location – June 4-8, 9am-noon in Root Hall A-008

Presenter – Liz Brown, Professor and Chairperson, Dept Math & Computer Science

Note – look in the blackboard course for the project for content for this session. Note also that we chose to not record this session.

Collaborative Bioinformatics

Date/time, location – June 14 9-11:15am; 1-4pm, July-15 9-11:30am in University Hall 008R

Presenter – Kevin Coombes, Professor, Dept Biomedical Informatics, The Ohio State University

The presenter will lead the group through all the parts of setting up a collaborative project – accounts on github, sharing code, documenting the whole analysis process, downloading data and making everything reproducible.

Prerequisites – R, Rstudio, github

See the zip file in Blackboard for this week for what needs to be done to get ready for Dr. Coombes's workshops.

Bioinformatics in RNA-seq Data

Date/time, location – June 21 9-11:15am, 1-4pm; July 22 9-11:30am in University Hall 008R

Presenter – Guy Brock, Research Assoc Prof, Dept Biomedical Informatics, The Ohio State University

The presenter will guide the group through the use of R to analyze RNA-seq data, including using logistic regression and the analysis of gene expression.

Prerequisites – R, logistic regression, RNA-seq data formats

Work on Projects

Date/time, location – June 25-29, 9am-4pm

Presenter – none, students work with their research advisors

Work on Projects

Date/time, location – July 2-6, 9am-4pm

Presenter – non, students work with their research advisors

Bioinformatics and Machine Learning

Date/time, location – July 12 9-11:15am, 1-4pm; July 13 9-11:30am in University Hall 008R

Presenter – Fuhai Li, Assistant Professor, Dept Biomedical Informatics, The Ohio State University

List of topics – convolution, deep learning (e.g., CNN, RNN), network analysis (e.g., centrality and shortest distance), gene set enrichment analysis, drug and targets (e.g., drug bank), signaling pathways (e.g., KEGG)

Summer Honors, Wet Lab

Date/time, location – July 16-20

Presenter – Catherine Steding, Assistant Professor, Dept Biology

During this week TCGA hosts a Summer Honors course for high school students. BD4ISU-supported students will participate in some of the sessions to receive some wet lab training from Catherine Steding.

Shiny Apps and Metabolite-Gene Integration Tools

Date/time, location – July 26 9-11:15am, 1-4pm; July-27 9am-11:30am in University Hall 008R

Presenter – Ewy Mathe, Assistant Professor, Dept Biomedical Informatics, The Ohio State University

Prerequisites - R, Rstudio, reading in and dealing with high throughput data, linear modeling

R packages that will be used: RShiny, Plotly, Highcharter, RMySQL

Additional tools that will be used: https://github.com/Mathelab/RaMP-DB, https://github.com/Mathelab/IntLIM
Other background content: over-representation pathway analysis

Symposium

Date/time, location – July 30 – August 3

During the last week of the summer program, students finalize posters, present the posters at the SURE Symposium, and move out of their summer housing. The poster presentation is Thursday, August 2.