CSB1020HS, LEC 0141
Data Visualization and Advanced Graphics in R

Offered by the *Centre for the Analysis of Genome Evolution & Function (CAGEF)*

Instructors:
Dr. David S. Guttman, CSB, CAGEF david.guttman@utoronto.ca
Dr. Calvin Mok, CAGEF Bioinformatics calvin.mok@utoronto.ca

Time:
March 3 – April 14 (6 weeks, no class on March 17th)
Thurs, 2:00-5:00pm

Enrollment:
20 graduate students
Up to 40 auditors

Course Objectives
This is an intermediate to advanced level introduction to R and the packages associated with visualizing large or complex data sets. Participants are strongly encouraged to have prior experience in R (i.e., Introduction to R, CSB1020). Individuals who complete the course will be able to manipulate and prepare large datasets to produce publication-quality graphics. The goal of this course is to introduce the proper use and interpretation of simple, popular and complex data visualizations. Topics will include

- A deep dive into building relatable figures with the ggplot package.
- Analysis and visualization of large datasets from differential expression experiments.
- Popular visualization methods and packages for genes and genome analysis.

Each class will consist of a short introductory section followed by 'code-along' hands-on learning that will gradually build up the lecture's topic(s). Students are expected to have access to a computer during class and are encouraged to ask questions while coding along with the instructor. A homework assessment will be assigned after each class to reinforce the skills learned. The course will be provided through Quercus using Zoom for class lectures and the University of Toronto JupyterHub for lecture and assessment materials.

Course Availability
This course will be presented online and will be available to all graduate students, postdocs, staff, and faculty, although only registered students will be evaluated. The course will count as a single module (0.25 credits) for CSB graduate students. All graduate students interested in taking the course for credit should enroll through ACORN.
Anyone wishing to audit the course should fill out the request form at https://bit.ly/2VBFble

Evaluation

Item	Note	% Mark
Homework Assignments	6 weekly assignments ranging from 15-20% each	100%

Pre-requisites: CSB1020 *Introduction to R* (or equivalent). Access to a computer.

Reference Material: *R for Data Science* (http://r4ds.had.co.nz/)

Syllabus

Class	Topic
1	**Re-Introduction to R, RStudio, and Jupyter Notebooks**: R and RStudio basics, setting up R for Jupyter Notebooks, installing R packages, best practices for producing graphs, best coding practices, functions and syntax, data types and structures, importing and exporting data, tidy data formatting, saving data and plots.
2	**The grammar of graphics with ggplot:** box-, violin-, beeswarm-, and jitter plots, combining layers in ggplot, kernel density plots, and parallel coordinate plots.
3	**Finishing touches for ggplot**: themes, aesthetics, color palettes, mathematic annotation with expression() and bquote(), scaling data, error bars, handling outliers, and multi-panel plots.
4	**Visualizing differential expression data**: heatmaps, volcano plots, side-by-side boxplots, dotplots, and Upset plots.
5	**Common visualization methods for data classification/partitioning:** clustering, principal component analysis, multidimensional scaling, and linear projection with t-SNE plots and UMAP.
6	**Simplifying Genes and genomes:** sequence logos, phylogenetic trees, network graphs, Manhattan plots, Gviz, GenomeGraphs, gene model plots and other helpful packages.

Subject to change