R for functional genomics (MCB4934/CB6940)
(3 credits)

Instructor	Ana Conesa	Phone	352 273 8230
Office	Genetics Institute 402 / 351b	E-mail	aconesa@ufl.edu
Location	ON-LINE Fall Semester 2015	Office hours	Tue and Fr 3-4 pm, GI room 351b

Motivation
Today's genome research is fundamentally quantitative. New advances in our understanding of genome function and its applications to areas such as personalized medicine or biotechnology have come from the generation and processing of large volumes of data. The progress of this field largely depends on new genome professionals that are comfortable with the notion of Big Data and skilled in the analysis of large datasets. One prerequisite for this is knowledge of a programming language that is also suitable for statistical analysis. R has imposed itself as THE open source statistical programming language for the genome research community. The goal of this course is to train students in R programming skills and in the analysis of functional genomics data.

Description and Student Learning Objectives
This is a 3 modules, 3 credits course that introduces students to the basics of the R language and that teaches the most important statistical concepts and algorithms used in functional genomics data analysis. Most of the course will deal with R scripts and packages although other software resources may be used for specific classes.
The three modules are:
Module I: Basics of the R language
Module II: Statistics for functional genomics
Module III: Functional genomics data analysis

The course is provided as a 4000 and 6000 level class. For the 6000 level class, students will need two complete two additional lecture modules and essays.

By the end of the course, the student should be able to
For 4000 level class (MCB4934):
- write basic R scripts
- utilize packages at R-CRAN and Bioconductor repository
- perform a complete RNA-seq data analysis pipeline using free software.

For 6000 level class (CB6940), additionally:
- Write a basic R package
- Perform a complete analysis pipeline for additional *.seq technologies

Requirements
Students should have background knowledge in Biology and Genetics. Students are expected to be familiar with the principles, of gene expression, the first dogma of molecular biology and the basic structures of the cell. Recommended courses: BSC2010, BSC2011, MCB3020, MCB3023, BCM4024, or CHM3218.

Moreover, students should also have basic knowledge in statistics Students are expected to be acquainted with basic summary statistics concepts such as mean, standard deviation, distribution, correlation, histograms, scatter plots, etc. Recommended courses are: STA6166 & STA6167. Students without the specific prerequisites can enroll with permission from the instructor.

Communication with instructor

Students can either formulate questions via email (indicating course code in the subject) or personally at office hours. Skype discussion sessions will be organized the first and third Tuesday of the month from 2:30 to 3:30. Students will need to provide their SkypeID and intention to participate in the discussion by the previous Monday at 12 pm. The SkypeID for the course is **MCB4934 CB6940.**

Software

Students will need to install the free-software R (http://www.r-project.org) and Rstudio (http://www.rstudio.com/) and a Linux virtual machine (provided by the Instructor at the beginning of Module III) in their personal computers.

Recommended books

R Programming for Bioinformatics. Robert Gentelman. July 14, 2008 by Chapman and Hall/CRC. ISBN 9781420063677

Bioconductor Case Studies. Hahne, F., Huber, W., Gentleman, R., Falcon, S. ISBN 978-0-387-77240-0

RNA-seq Data Analysis: A Practical Approach. Eija Korpelainen, Jarno Tuimala, Panu Somervuo, Mikael Huss, Garry Wong. September 19, 2014 by Chapman and Hall/CRC. ISBN 9781466595002

Evaluation and Grading

Lectures and exercises will be posted weekly. Students must complete exercises assigned each week by the Wednesday of the following week. Each module will be evaluated by an essay consisting of writing a script with functions and concepts discussed in the module.

Weekly exercises: 400 points
Module essays: 600 points

Each module will be evaluated independently and final grade will be the sum of points obtained in each module. Essays for module I, II, and III are worth 150, 200, and 200-250 points respectively. Extra points can be won by including essays for additional specific subjects of the course.

Grading Scale:

A	900 or above
A-	860-899
B+	830-859
B	790-829
B-	750-789
C+	720-749

C	690-719
C-	660-689
D+	630-659
D	600-629
D-	570-599
E	560 or below

The grading scale may be adjusted slightly, based on class performance.

Academic Honesty
As a result of completing the registration form at the University of Florida, every student has signed the following statement: "I understand that the University of Florida expects its students to be honest in all their academic work. I agree to adhere to this commitment to academic honesty and understand that my failure to comply with this commitment may result in disciplinary action up to and including expulsion from the University.

Software Use
Students will need to install in their computers three free software packages (R, RStudio and VMware).
All faculty, staff and students of the University are required and expected to obey the laws and legal agreements governing software use. Failure to do so can lead to monetary damages and/or criminal penalties for the individual violator. Because such violations are also against University policies and rules, disciplinary action will be taken as appropriate.

UF Counseling Services are available on-campus for students having personal problems or lacking clear career and academic goals includes:
• University Counseling Center, 301 Peabody Hall, 392-1575, personal and career counseling
• Student Mental Health, Student Health Care Center, 392-1171, personal counseling.
• Sexual Assault Recovery Services (SARS), Student Health Care Center, 392-1161, sexual assault counseling
• Career Resource Center, Reitz Union, 392-1601, career development assistance and counseling

Each online distance learning program has a process for, and will make every attempt to resolve, student complaints within its academic and administrative departments at the program level. See http://distance.ufl.edu/student-complaints for more details

Week start	Topic	Description
MODULE I Basic R		
8/24/2015	Introduction to R	The R language, Rstudio, Bioconductor, Installation of packages, Help in R Language elements: vectors, matrices,

		lists, functions, data frames, factors. The "as." Function. Language elements: brackets, curly brackets, arrow, :, ;, comments
8/31/2015	Basics of R	Assign value to a variable, Browse data, Basic operators, Logic operators, Read and Write, *Data generation*
9/7/2015	Basic Functions I	Basic statistics (mean, max, etc), Compare sets, Order, string manipulations, Matrix Subsetting
9/7/2015	Basic Functions II	Apply family, Loops, *Creating functions*
9/21/2015	Graphs in R	ggplot function, graphical parameters, *high-quality graphs*
Evaluation: Students will create a personal dataset with at least 5 variables and 20 observations. Create an R script to analyze proprietary data. At least 20 different functions have to be used, including functions of weeks 4 and 5. Script will start with an explanation of the data, followed by the analysis to be performed and the conclusions of the analysis. Points: 150. 30 extra points by including 2 functions from: *data generation, creating functions, high-quality graphs.*		
MODULE II Statistics for genomics		
9/28/2015	Hypothesis testing	sample vs population, Reference distributions, pvalue, type I and II errors t-test
10/5/2015	Univariate statistics	Linear-models, Fisher-exact test, Correlation, *Willconxon Kolmogorov–Smirnov test*
10/12/2015	Multivariate statistics	Clustering, Heatmaps, PCA, Multiple testing, *Bootstrap*
EXTRA 6000	Create an R package	S4 classes, package skeleton, document packages
Evaluation 4000 class: Students will be given a gene expression dataset. They will create an R script to analyze these data. At least 2 univariate and 2 multivariate methods should be used. Script will start with an explanation of the data, followed by the analysis to be performed and the conclusions of the analysis. Points: 200. 50 extra points by including either: *Willconxon, Kolmogorov–Smirnov test or Bootstrap.* Evaluation 6000 class: Create a simple R package to analyze the same data. This essay replaces the 4000 class essay. The same requirements in terms of statistical methods and functions hold.		
MODULE III Functional genomics		
10/9/2015	NGS and Functional Genomics	Sequencing machines, *.seq assays Large NGS projects, Repositories NGS data
10/26/2015	Introduction to Linux	Virtual machine, Bash commands, Install programs
11/2/2015	RNA-seq analysis I	Experimental design, RNA-seq pipelines Sam/Bed tools, IGV
11/09/2015	RNA-seq analysis II	Mapping, Quantification

11/16/2015	Differential expression	2 class-comparisons, *Time series analysis*
11/23/2015	Functional profiling	Functional databases, Functional Enrichment (DAVID), *GSA, Network Analysis*
11/23/2015	RNA-seq analysis III	*Assembly*
12/7/2015	Functional annotation	*Blast2GO*
EXTRA 6000	Other *.seq data	ChiP-seq and Methyl-seq. Peak calling algorithms
Evaluation: Students will be given RNA-seq .bam files of either case-control, time series, or reference free transcriptomics experiment (student choice). Complete RNA-seq analysis pipeline from Quantification to Functional Profiling. A student script will start with an explanation of the data, followed by the analysis to be performed and the conclusions of the analysis. Points: 200 for case-control study, 250 for time series or reference free analysis. 50 extra points by including either: *GSA, Network Analysis, Blast2GO.* Evaluation 6000 class (extra essay): Analyze ChiP-seq with peak caller. 250 points.		