M21-550 Course Syllabus

Summer 2017 (July 27-August 9)

Course Title:	**Introduction to Bioinformatics**
Course Master:	C. Charles Gu, Ph.D., Rosy Luo, Ph.D.
Guest Lecturer:	Gary Stormo, Ph.D.
Teaching Assistant:	Will Yang, Ph.D.
Schedule:	*R-Primer July 5~12 required of ALL students.* Bioinformatics Class - **Every weekday, 9:00am-12:00pm** and **1:30-4:30pm**.
Place:	MSIBS classroom & Computer Lab (Room 502 & 501, 5th floor, Becker's Library)
Format:	Small group lecture and extensive computer labs using real-world data
Grade Criteria:	Numerical score based on:

Quizzes	35%
Lab projects	45%
Final exam	20%

Reference Text:

1. Andreas Baxevanis and Francis Ouellett (eds.)**,** Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins**,** 3rd Edition, Wiley & Sons, 2005.
2. Robert Gentleman, Vincent Carey, Wolfgang Huber, Rafael Irizarry, Sandrine Dudoit (eds.), Bioinformatics and Computational Biology Solutions Using R and Bioconductor, Springer, 2005.
3. Malcolm Campbell and Laurie Heyer, Discovering Genomics, Proteomics and Bioinformatics (2nd Edition), Benjamin Cummings, 2007.
4. Richard Durbin, Sean Eddy, Anders Krogh, and Graeme Mitchison, Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, Cambridge University Press, 1998.

Other Readings:

Additional reading materials and papers will be assigned by lecturers.

Description: The course provides a broad exposure to basic concepts, methodologies and applications of bioinformatics. Students will learn online databases & mining tools, and acquire understanding of mathematical algorithms in sequence analysis (sequence alignment, gene finding, and hidden Markov models), gene expression microarray analysis (data QC & normalization, univariate & multivariate differential expression analysis), next generation sequence analysis (short-read data format and processing, variant calling algorithms), and topics on other high-throughput biomedical experiments. Students will become familiar with popular bioinformatics software, online tools, and R/BioConductor packages. We will discuss methods for high-dimensional data analysis including classification and clustering analysis, principal component analysis (PCA), statistical/machine learning, and Bayesian inference. There also will be seasonal additional lectures on topics such as proteomics and applications of bioinformatics to real studies of complex diseases.

As an important component of this course, students will conduct hands-on computer labs to learn basics of online bioinformatics databases and tools, and to practice computer programming. The labs require using the statistical computing environment R (i.e., the R primer is required) though introduction to BioConductor basics will be provided. Students will use specialized software and R packages to accomplish tasks including designing experiments, low-level analysis of expression levels, univariate differential expression analysis, and various multivariate analysis techniques taught in class. A variety of software will be used for NGS data analysis covering alignment, variants calls, differential analysis, and visualization of results. Through the lab exercises, students will learn how state of art computational tools are applied to solving bioinformatics problems in real studies of human diseases.

Schedule of Lectures:

Day	Morning (9:00am -12:00)		Afternoon (1:30pm - 4:30pm)	
	Topics	**Instructor**	**Computer lab**	**Instructor**
#1 Th 7/27	**Course overview:** summary of topics, grading policy, expectations	Gu		
*Qz-1**	**Sequence analysis I:** homology detection; pair-wise sequence alignment, alignment score statistics, BLAST search	Stormo, Gu	**Lab #1:** online databases & tools, BLAST	Gu
#2 F 7/28 *Qz-2*	**Sequence Analysis II:** multiple alignment, Hidden Markov Models (HMM) & applications to motif and gene finding	Gu	**Lab #2:** seq alignment, database/R tools; HMM motif/gene prediction	Gu
#3 M 7/31 *Qz-3*	**DNA variations analysis:** diseases & DNA variations; copy number variations (CNVs), array-/sequence-based CNVs detection & analysis	Gu	**Lab #3:** R refresher; R tools for CNV (*DNAcopy, ReadDepth,* etc.)	Gu
#4 T 8/1 *Qz-4*	**Microarray analysis I:** microarray & transcriptomics; high-dimensional analysis, bias & normalization; robust statistics; differential expression, false discovery rate, resampling statistics	Gu	**Lab #4:** microarray data preprocessing (*ReadAffy*); differential expression analysis (*siggenes*);	Gu
#5 W 8/2 *Qz-5*	**Microarray analysis II:** multivariate analysis & dimension reduction, PCA (principal component analysis); dissimilarity measures, clustering analysis (hierarchical, k-means)	Gu	**Lab #5:** R tools for PCA (*prcomp*), clustering analysis (*heatmap, hclust*)	Gu
#6 Th 8/3 *Qz-6*	**Microarray analysis III:** statistical learning (supervised/unsupervised), model/feature selection; classification analysis, LDA (linear discriminant analysis), Naïve Bayes classifier	Gu	**Lab #6:** R tools for LDA (*lda*), naive Bayes classification (*naiveBayes*)	Gu
#7 F 8/4 *Qz-7*	**NGS data analysis I:** intro to next generation sequencing (NGS), sequencing technology & platforms, short reads alignment algorithms & software	Luo	**Lab #7:** BWA, Bowtie, Galaxy; R tools for exploring sequencing data	Luo, Gu
#8 M 8/5 *Qz-8*	**NGS data analysis II:** NGS file formats, SAM/BAM; SAMtools; variants calling, VCF format; QC of NGS reads; NGS study design and statistical issues	Luo	**Lab #8:** SAMtools, and R tools for exploring sequencing data	Luo, Gu
#9 T 8/8 *Qz-9*	**NGS data analysis III:** RNA-seq based expression analysis, count-based normalization, transcript-level differential expression	Luo/Gu	**Lab #9:** R tools for RNA-seq analysis (*edgeR, DESeq*); TopHat, Cufflinks; Galaxy	Luo/Gu
#10 W 8/9	**Bioinformatics in medicine:** genetic circuits, environments & interactions; EHR & bigdata, emerging data science; precision medicine & public health initiatives	Gu	**FINAL EXAM (2:00 - 4:00pm)**	Gu

* *Qz-1* is an assessment of your proficiency in R; it will not be counted toward your final grade.