**SEAS COURSE DESCRIPTION TEMPLATE**

|  |  |
| --- | --- |
| **Course Number & Title** | **BE504: Biological Data Science II -- Data Mining Principles for Epigenomics** |
| **Credit Units** | 1 CU (3 semester hours) |
| **Class/Laboratory Schedule** | Lecture: 3 hrs/week  Towne 311 – 3-4:30pm – Tues/Thurs |
| **Instructor** | Jennifer E. Phillips-Cremins, PhD, Assistant Professor of Bioengineering  Office Hours: On request – Thursday 4:30-5:30pm Hayden 304 |
| **Prerequisites** | Undergraduates Require Senior Standing or Permission of the Instructor |
| **Course Satisfies**  **(check only one)** | [ ] Math  [ x] Science  [ ] Engineering  [ x] Technical Elective  [ ] TBS |
| **Text(s)/Required Materials** | Assigned reading will come from the primary literature. |
| **Catalog Description** | This course will teach upper level undergraduates and graduate students how to answer biological questions by harnessing the wealth of genomic and epigenomic data sets generated by high-throughput sequencing technologies. |
| **Topics Covered** | The purpose of this course is to provide students with skills to analyze and interpret biological data generated by high-throughput sequencing. Fundamentals in biostatistics, computational biology, chromatin biology and epigenetics will be taught through a series of case studies focused on cutting edge biological questions in biomedical research. Example case studies include: measuring gene expression changes during disease progression; analyzing epigenetic marks (histone modifications, DNA methylation) in stem cells and differentiated cells; or investigating changes in transcription factor binding across the genome in response to drug treatment. Each case study will be covered by 3-4 lectures that (1) provide background on the biology behind the epigenetic modification, (2) describe the molecular biology technique employed to query that particular epigenetic modification, (3) introduce statistical concepts important for analyzing specific types of high-dimensional data. This course will not cover comparative genomics, sequence alignment algorithms, genome assembly or population genetics. |
| **Course Objectives and Relationship to Program Education Objectives** | The knowledge in biomedical science and training in statistical principles imparted through this course will enable students to understanding emerging paradigms in systems biology, epigenetics and genomics.  The objectives include:   1. To appreciate statistical methods for analyzing high dimensional data 2. To encourage students to think about biology probabilistically 3. To format, summarize, explore, stratify, visualize and model big data 4. To select statistical methods that are appropriate to the scientific question of interest 5. To estimate parameters and quantify uncertainty for different types of genomic data 6. To assess statistical significance when multiple hypothesis tests are performed   Students will conduct a variety of analytical tasks using python or R programming languages and real world data sets. Coding experience is preferred, but not essential. Algorithms required for each homework will be provided, but students are expected to invest the time to work through each line of code independently. Students will be asked to adapt the code to solve the homework problems and also explore their own selected data. |
| **Contribution towards Program Outcomes** | Multidisciplinary Ability – High  Problem Solving Approach – High  Problem Solving Methods – Med  Experimentation – Med  Design – Med  Professional Orientation – High |
| **Contribution towards Professional Component** | 40% Biomedical Science; 30% Engineering science; 30% Engineering mathematics |
| **Weekly/Session Schedule** | **Overview Lectures**  **1/11 Course Outline – Epigenetics Overview**  **1/16 Epigenomics & Next Generation Sequencing Overview**  **Module 1: Transcription**  **1/18 Mechanisms of Gene Expression Regulation**  **1/23 RNAseq & Sequencing Read Mapping**  **1/25 Exploratory data analysis and correlations**  **1/30 Transformation, Normalization (seq depth, median-of-ratios, quantile, rpkm), Spikein Controls**  **2/1 Modeling continuous (Normal; Log-normal) vs. discrete distributions (Poisson; Negative Binomial)**  **2/6 Parameter estimation with maximum likelihood; Global mean-variance relationship**  **2/8 CLASS CANCELLED DUE TO EAGLE’S PARADE**  **2/13 Differential Gene expression Part 1. Hypothesis testing; T-test; Fisher’s Exact Test**  **2/15 QUIZ 3 and Titus coding Q&A session – Cremins accepting Kavli award in Irvine, CA**  **2/20 Differential Gene expression Part 2. ; Likelihood Ratio Test; Wald Test; Nonparametric tests (K-S; Mann**  **Whitney U)**  **2/22 Differential Gene expression Part 3. Randomization tests; Permutation tests; Bootstrapping; In class**  **exercise**  **2/27 Principal Component Analysis**  **3/1 Multiple Testing Correction and ROC curves**  **\*\*Note, By Friday 3/16 at 11:59pm you should have a 1 page talk sheet describing your plans for your final project with hypothesis, objective, data selected and methods with 10 general bullet points uploaded to canvas**  **Spring Break**  **3/6 SPRING BREAK – OFF**  **3/8 SPRING BREAK – OFF**  **Module 2: Histone Modifications**  **3/13 Histone modifications, chromatin signatures on enhancers/promoters**  **3/15 ChIP-seq**  **3/20 Peak calling punctate and diffuse chromatin marks**  **3/22 Finding biology in ChIPseq peaks – gene ontology, set theory, and pileup plots**  **Module 3: Chromatin Accessibility**  **3/27 DNA binding proteins and chromatin accessibility**  **3/29 ATAC-seq, DNA-seq, MNAse-seq**  **4/3 Computing a consensus sequence**  **4/5 Clustering**  **Module 4: Higher-order chromatin architecture**  **4/10 3-D genome folding**  **4/12 Hi-C, 5C, 4C, 3C, ChIA-PET, CaptureC, RICCseq**  **4/17 Heatmaps and calling TADs and subTADs**  **4/19 Distance-dependence background, donut expected, Calling “loops”**  **Module 5: DNA methylation**  **4/24 DNA methylation, hydroxymethylation, Bisulfite-seq, RRBS-seq**  **4/26 READING DAY – OFF**  **\*Note: Final Project Due: Wednesday, April 25, 2018 @ 11:59pm**  **Integrating multiple (2-3) epigenomic (or other genetics) data sets to test hypotheses about transcriptional (or genome) regulation. Students pick their own data sets – write a 3 page report with abstract, introduction, hypotheses, methods, results, conclusions and references. Include at least 5 figures with captions.** |
| **Grading Details** | 60% Homework and In-class quizzes - 6 quizes – 5 homeworks – lowest grade dropped  30% Final Project  10% Attendance and In-Class Participation |
| **Homework and Project**  **Due Dates** | Below are assignment dates and their topics. I will clarify due dates as each homework is rolled out.  **HOMEWORK 1 assigned on 1/29/2018 - Due 2/7/2018**  1/31 Exploratory data analysis and correlations  **HOMEWORK 2 - assigned on 2/1/2018 - Due 2/14/2018**  1/30 Transformation, Normalization (seq depth, median-of-ratios, quantile, rpkm), Spikein Controls  2/1 Modeling continuous (Normal; Log-normal) vs. discrete distributions (Poisson; Negative Binomial)  2/6 Parameter estimation with maximum likelihood; Global mean-variance relationship  **HOMEWORK 3 - assigned on 2/8/2018 - Due 2/22/2018**  2/13, 2/20, 2/22 Differential Gene expression Part 1. T-test; Fisher’s Exact Test; Likelihood Ratio Test; Wald Test  Differential Gene expression Part 2. Nonparametric tests (K-S; Mann Whitney U); Part 3. Randomization tests;  Permutation tests; Bootstrapping  **HOMEWORK 4 - assigned on 2/24/2018 - Due 3/3/2018**  2/27 Multiple Testing Correction  **Homework 5 - assigned on 2/24/2018 - Due 3/20/2018**  3/1 Principal Component Analysis  **FINAL PROJECT STRATEGY PLAN  (Formulate final project plan and preliminary data 3/1/2018-3/16/2018)** \*\*Note, By Friday 3/16 at 11:59pm turn in a 1 page talk sheet with hypothesis, objective, data selected and methods with 10 general bullet points to canvas  **FINAL PROJECT DUE -- WEDNESDAY, APRIL 25 @ 11:59pm**  Final Project: Integrating multiple (2-3) epigenomic (or other genetics) data sets to test hypotheses about transcriptional (or genome) regulation. Students pick their own data sets – write a 3 page report with abstract, introduction, hypotheses, methods, results, conclusions and references. Include at least 5 figures with captions. |
| **Prepared By/Date** | Jennifer Phillips-Cremins / January, 2018 |