Advanced Topics in Computational Genomics

Spring 2018

Instructors:	Dan DeBlasio	**Time:**	TR 15:00 – 16:20
	Mingfu Shao	**Place:**	Gates-Hillman Complex 4301
	Heewook Lee	**Credit Hours:**	12 (9 for undergraduates)
Teaching Assistant:	Natalie Sauerwald		
Emails:	{deblasio,mingfus,heewookl,Natalie.Sauerwald}@cs.cmu.edu		

Course Website: The CMU Piazza system will be used for web-based course communications. The instructors will enroll the students before the first day of the course. If you enrolled after the first day of class please email one of the instructors to be added.

Office Hours: All three instructors will be available in GHC 7401 on Tuesdays from 10:30 to 11:30 or by appointment.

Main References: This course will take material from current publications that will be listed on the course webpage in a timely manner. If the selected papers are not available though the campus website copies will be made available before the discussion.

Objectives: This course is primarily designed for graduate students to gain exposure to emerging topics in genomics that are not covered in existing course offerings. The topics presented here may overlap with the students ongoing research, but at least one topic should be novel. At the end of the course it is expected that students able to demonstrate some knowledge of the topics presented, the goal is for each student to feel comfortable working with and discussing the topics being covered with anyone on the leading edge of the field. Additionally, the course may expose students to topics that could be of research interest to them later in their career and spur ideas for ongoing research opportunities. We encourage the participants to integrate the projects with their own research if the opportunity arises.

Prerequisites: 02-710 or equivalent. This course is designed for advanced graduate students in CBD, primarily in their second year or beyond. Highly advanced undergraduates who have passed 02-510 and graduate students from other departments are welcome with instructor approval.

Grading Policy: Course participation (50%), Projects (25% each, choose 2 of 3).

Important Dates:

First Class Meeting	January 16, 2018
Module 1 paper selection due	January 26, 2018[a]
Module 1 presentation schedule announced	by January 30, 2018
Module 2 paper selection due	February 23, 2018[a]
Module 2 presentation schedule announced	by February 27, 2018
Project 1 due	Mar 2, 2018[a]
Spring Break (no classes)	March 13 & 15, 2018
Module 3 paper selection due	March 23, 2018[a]
Module 3 presentation schedule announced	by March 27, 2018
Project 2 due	Apr 6, 2018[a]
Last Class Meeting	May 3, 2018
Project 3 due	May 4, 2018[a]

[a] by 5:00pm EST

Class Policy: Regular attendance is essential and expected. Due to the high emphasis of group discussion and dialogue all students are discouraged from missing classes. Missed course meetings will be noted and chronic absences may impact the students grade if not discussed with the course instructors.

The course projects are meant to expose the student to the topics being discussed. Because only two out of three projects is required, students may be asked to make additional paper presentations on topics in which they are not participating in the project. This will be determined by the final class size and at the module instructors discretion.

Late projects will not be accepted.

Tentative Course Outline: The course will be broken into three major topics, each covered by one instructor. The topics and order listed are tentative as are the start and end dates. Each module will be accompanied by a project which will assist in the understanding of the topic.

Week	Tuesday meeting	Thursday meeting
Jan 16–18	Detailed introduction to Module 1: "Alignment-free genomics" (DeBlasio)	
Jan 23–25	Detailed introduction to Module 2: "Single-cell RNA-seq analsysis" (Shao)	
Jan 30–Feb 1	Detailed introduction to Module 3: "Immunogenomics" (Lee)	
Feb 6–8	"Alignment-free genomics" module intro (DeBlasio)	Paper discussions
Feb 13–15	Paper discussions	
Feb 20–22	Paper discussions	
Feb 27–Mar 1	Project presentations	
Mar 6–8	"Single-cell RNA-seq analsysis" module intro (Shao)	Paper discussions
Mar 13–15	Spring Break	
Mar 20–22	Paper discussions	
Mar 27–29	Paper discussions	
Apr 3–5	Project presentations	
Apr 10–12	"Immunogenomics" module intro (Lee)	Paper discussions
Apr 17–19	Paper discussions	
Apr 24–26	Paper discussions	
May 1–3	Project presentations	

Academic Honesty: All students will be held to the up most standard of university academic integrity. The full description of CMU's policies can be found on the university webpage (https://www.cmu.edu/policies/student-and-student-life/academic-integrity.html) it is you're responsibility of understanding and following all guidelines.

Course project: Each student is expected to complete a projects related to two (2) of the three (3) topics being presented. The selection of proposed projects should be discussed with the module instructor, via email or in person, during or before the first week of each module (February 6–8, March 6–8, April 10–12). Each student may be asked to present their project to the class during the last week of each module (at the discretion of the module instructor). The final project deliverable (as defined by each instructor) will be due the Friday of the last week of the module.

Listed below are the projects for each topic being discussed. Each student must select a project outline and propose a detailed project to the instructors (in person or via email). It is expected that no two students will do the same exact project, and the instructors may ask that students slightly alter their proposed projects to not overlap with each other. These modifications will be made when proposals are submitted.

Alignment-free genomics: Sample contamination is major problem for sequencing experiments, the sources of these contaminants is sometimes unknown and can thus be hard to environmentally correct. The goal of this project will be to determine the "contaminome" of the Mellon Institute building and identify some of the causes of the noise introduces in several sequencing experiments conducted in labs in the building. Students will be provided with sequencing reads that do no match to the target genome that was sampled and will be required to attempt to cluster and identify the source of the remaining cells that were sequenced. At the end of this module each student will give a presentation to the class on their methods and findings. The method developed can either be novel or an adaptation of existing software. If existing software is used some significant change should be made to adapt the tool to the given experiment. At the end of the day the goal will be to present the most reasonable explanation for the given data and convince the instructors as well as fellow classmates that you're results are most correct.

Ideally all of the evidence collected will be given to the experimentors who generated the data to either correct the problem if possible or have a reference of known contaminants that can be removed from future experiments.

Single-cell RNA-seq analysis: The goal of this project to get familiar with the pipeline of scRNA-seq analysis. The description of the project is as follows.

1. Select two human scRNA-seq datasets related to cell differentiation.
2. Choose one transcript assembly program (either specifically for scRNA-seq, or for bulk RNA-seq) to assembly the transcripts in the cells and try to identify novel transcripts (i.e., those not in human refernece transcriptome). **NOTE:** each student should pick up different program.
3. Choose one program to identify cell types and subtypes in all cells. **NOTE:** different student should pick up different program.
4. Choose one trajectory inference method to construct and visualize the pesudotime of cell progress. **NOTE:** each student should pick up different program.

Immunogenomics: There are 2 projects to choose from:
Graph alignment and modification Many problems in immunogenomics require accurate identification/assembly of immune-related sequences (MHC, KIR, etc). The goal here is to implement a graph-based alignment module for HLA assembly/typing. The input to the module consists of a multiple sequence alignment of HLA alleles and sequencing reads from HLA regions. It should first construct a partial order graph from the given MSA and index the graph. The indxing can be constructed using BWT for graph by Siren

et al. (2014). The module should be able to align reads to the graph and perform graph modification when there are read-graph differences.

Evaluation of MHC binding affinity predictors

Prediction of the binding affinity between MHC molecules and peptides is an important problem in immunogenomics. Especially with promising results of neoantigen-based cancer vaccines in recent years, the problem is essnetial in automatically screening for neoantigens that can actually bind to patients MHC molecules. Lately, there has been many MHC biniding affinity predictors and the goal of this project is to evaulate some of the popular biniding affinity predictors and compare their performances.

Paper presentations: Each student will be required to present papers related to topics covered in the course. Depending on the enrollment each student will be asked to present one or more papers related to each module. It is expected that each student will begin by giving a review of the paper, pointing out what makes the method or result unique and what should be taken away from the publication. The student will then lead a discussion about the paper.

Below are a list of possible papers for students to present for each topic. A selection from each of the topics related to the students project should be chosen at least two (2) weeks before the start of a module and emailed to the instructors. Papers will be assigned based on both relevance to each students projects then on a first come first served basis. The talk order will be released as well as any changes that may need to be made to ensure that each paper is only presented once will be announce the week before each module starts. While these lists are suggested reading material, the selection of papers that are not listed but are relevant in these areas is not only allowed but encouraged.

Alignment-free genomics:

1. Patro, R, Duggal, G, Love, MI, Irizarry, RA, Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods* (2017)
2. Jain, C, Dilthey, A, Koren, S, Aluru, S, Phillippy, AM. A fast approximate algorithm for mapping long reads to large reference databases. *Lecture Notes in Computer Science* **10229** (2017)
3. Wood, DE, Salzberg, SL. `Kraken`: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology* **15** R46.(2014)
4. Grabowski, S, Raniszewski, M, Sampled suffix array with minimizers. *Software: Practice and Experience*, **47** 1755–1771 (2017)
5. Thankachan, SV, Chockalingam, SP, Liu, Y, Krishnan, A, Aluru, S. A greedy alignment-free distance estimator for phylogenetic inference. *BMC Bioinformatics* **18**(Suppl 8):238 (2017)
6. Luo, Y, Yu, YW, Zeng, J, Berger, B, Peng, J. Metagenomic binning through low density hashing. *bioRxiv* **133116** (2017)
7. Holley, G, Wittler, R, Stoye, J. Bloom Filter Trie - A Data Structure for Pan-Genome Storage. *Workshop on Algorithms in Bioinformatics (WABI)* (2015)
8. Lu, YY, Chen, T, Fuhrman, JA, Sun, F. COCACOLA: binning metagenomic contigs using sequence COmposition, read CoverAge, CO-alignment and paired-end read LinkAge. *Bioinformatics* **33**:6 791–798 (2017)
9. Liao, W, Ren, J, Wang, K, Wang, S, Zeng, F, Wang, Y, Sun, F. Alignment-free Transcriptomic and Metatranscriptomic Comparison Using Sequencing Signatures with Variable Length Markov Chains. *Scientific Reports* **6**:1 (2016)
10. Popic, V, Kuleshov, V, Snyder, M, Batzoglou, S. GATTACA: Lightweight Metagenomic Binning with Compact Indexing of Kmer Counts and MinHash-based Panel Selection. *Research on Computational Biology (RECOMB)* (2017)

Single-cell RNA-seq Analysis

1. T. Smith, A. Heger, and I. Sudbery. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Research*, 27(3):491–499, 2017.

2. F. Buettner, K.N. Natarajan, F.P. Casale, V. Proserpio, A. Scialdone, et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology*, 33(2):155–160, 2015.

3. E.D. Amir, K.L. Davis, M.D. Tadmor, E.F. Simonds, J.H. Levine, et al. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nature Biotechnology*, 31(6):545–552, 2013.

4. C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N.J. Lennon, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, 32(4):381–386, 2014.

5. E. Marco, R.L. Karp, G. Guo, P. Robson, A.H. Hart, L. Trippa, and G.-C. Yuan. Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proceedings of the National Academy of Sciences*, 111(52):E5643–E5650, 2014.

6. J.D. Welch, Y. Hu, and J.F. Prins. Robust detection of alternative splicing in a population of single cells. *Nucleic Acids Research*, 44(8):e73–e73, 2016.

7. P.V. Kharchenko, L. Silberstein, and D.T. Scadden. Bayesian approach to single-cell differential expression analysis. *Nature Methods*, 11(7):740–742, 2014.

8. T.J. Nowakowski, A. Bhaduri, A.A. Pollen, B. Alvarado, M.A. Mostajo-Radji, et al. Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science*, 358(6368):1318–1323, 2017.

Immunogenomics

1. Euan A Ashley. Towards precision medicine. *Nature Reviews Genetics*, 17(9):507–522, 2016.

2. Patrick A Ott, Zhuting Hu, Derin B Keskin, Sachet A Shukla, Jing Sun, David J Bozym, Wandi Zhang, Adrienne Luoma, Anita Giobbie-Hurder, Lauren Peter, et al. An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature*, 547(7662):217, 2017.

3. Ugur Sahin, Evelyna Derhovanessian, Matthias Miller, Björn-Philipp Kloke, Petra Simon, Martin Löwer, Valesca Bukur, Arbel D Tadmor, Ulrich Luxemburger, Barbara Schrörs, et al. Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature*, 547(7662):222, 2017.

4. Heewook Lee and Carl Kingsford. `Kourami`: Graph-guided assembly for novel HLA allele discovery. *Genome Biology*, 2017.

5. Dilthey AT, Gourraud PA, Mentzer AJ, Cereb N, Iqbal Z, McVean G. High-accuracy HLA type inference from whole-genome sequencing data using population reference graphs. PLoS Comput Biol. 2016;12(10):e1005151.

6. András Szolek, Benjamin Schubert, Christopher Mohr, Marc Sturm, Magdalena Feldhahn, and Oliver Kohlbacher. OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics*, 30(23):3310–3316, 2014.

7. Ibrahim Numanagić, Salem Malikić, Victoria M Pratt, Todd C Skaar, David A Flockhart, and S Cenk Sahinalp. Cypiripi: exact genotyping of CYP2D6 using high-throughput sequencing data. *Bioinformatics*, 31(12):i27–i34, 2015.

8. Greyson P Twist, Andrea Gaedigk, Neil A Miller, Emily G Farrow, Laurel K Willig, Darrell L Dinwiddie, Josh E Petrikin, Sarah E Soden, Suzanne Herd, Margaret Gibson, et al. Constellation: a tool for rapid, automated phenotype assignment of a highly polymorphic pharmacogene, CYP2D6, from whole-genome sequences. *NPJ Genomic Medicine*, 2:16039, 2017.

9. Massimo Andreatta and Morten Nielsen. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics*, 32(4):511–517, 2015.

10. Vanessa Isabell Jurtz, Sinu Paul, Massimo Andreatta, Paolo Marcatili, Bjoern Peters, and Morten Nielsen. NetMHCpan 4.0: Improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *bioRxiv*, page 149518, 2017.

11. Timothy O'Donnell, Alex Rubinsteyn, Maria Bonsack, Angelika Riemer, and Jeffrey Hammerbacher. MHCflurry: open-source class I MHC binding affinity prediction. *bioRxiv*, page 174243, 2017.

Disclaimer: This course syllabus is subject to change at the instructors discretion. Any changes will be discussed in class and/or posted on the course website.