EEEB GU4055 Principles and applications of modern DNA sequencing

Term taught: Spring 2019
Class times: Mondays and Wednesdays, 11:40am-12:55pm.
Classroom location: Schermerhorn Extension 10th floor classroom
Course format: Lectures, discussions, computer exercises using Codio, laboratory sessions and a field trip.
Points for the course: 4
Level: Undergraduate and graduate
Prerequisites: Introductory biology or permission of the instructor
Maximum enrollment: 25
Instructor's permission required prior to registration: No

Instructors:

Andrés Bendesky
a.bendesky@columbia.edu
(212) 853 1173
Jerome L. Greene Science Center
3227 Broadway, L3-051
Office hours: TBD

Deren Eaton
de2356@columbia.edu
(212) 851 4064
Schermerhorn Extension 1007
1200 Amsterdam Ave.
Office hours: TBD

Course description and bulletin

Genome sequencing, the technology used to translate DNA into data, is now a fundamental tool in biological and biomedical research, and is expected to revolutionize many related fields and industries in coming years as the technology becomes faster, smaller, and less expensive. Learning to use and interpret genomic information, however, remains challenging for many students, as it requires synthesizing knowledge from a range of disciplines, including genetics, molecular biology, and bioinformatics. Although genomics is of broad interest to many fields—such as ecology, evolutionary biology, genetics, medicine, and computer science—students in these areas often lack sufficient background training to take a genomics course. This course bridges this gap, by teaching skills in modern genomic technologies that will allow students to innovate and effectively apply these tools in novel applications across disciplines. To achieve this, we implement an active learning approach to emphasize genomics as a *data science,* and use this organizing principle to structure the course around computational exercises, lab-based activities using state-of-the-art sequencing instruments, case studies, and field work. Together, this approach will introduce students to the principles of genomics by allowing them to generate, analyze, and interpret data *hands-on* while using the most cutting-edge genomic technologies of today in a stimulating and engaging learning experience.

Organization and learning outcomes

Learning objectives: The primary learning objectives of this course are to train students in the skills required to design, conduct, and analyze a genomic experiment—from the first step of generating raw data to the final steps of statistical analyses. The course builds upon subjects with increasing complexity. By the end of class students should be able to: (1) describe the structure of genomes and how information is represented in them; (2) choose the most appropriate sequencing technique for a particular question; and (3) analyze genomic data using computational methods. Each of these objectives can be measured -- using assignments, in-class discussions, the midterm exam, and projects -- and will form the basis of our assessments used to ensure students are learning as expected.

Format: The course will meet on Mondays and Wednesdays for 75 minutes. Each meeting will be a mix of lecture, in-class active learning exercises, and discussion. A few meetings will take place in a laboratory where students will learn simple molecular techniques. Most weeks, readings will be assigned between class periods with accompanying computational exercises. All readings are from the primary literature and can be accessed through the Columbia library portal (free). Computational exercises will be graded to assess students comprehension of materials and to reinforce lessons from the readings. The following meeting will begin with a review of topics from the readings and a group discussion among students to compare solutions to computational exercises. Every other class period will focus on solving an applied problem using the genetic techniques we have learned.

Assignments: There will be 20 assignments in which students complete computational exercises in on-line notebooks. These assignment can be worked on in groups. Code reviews will be done in class as group discussions. Online polling will be used in class to provide points for participation and to assess reading comprehension. Students will prepare an essay and give a short presentation for a project near the end of class in which they envision a new sequencing technology or application. This is to be completed individually and requires synthesizing knowledge from topics throughout the semester. Finally, students will write a final report following the field trip at the end of the course to summarize and describe their results.

Basis for grading: Grades will be composed of 20 assignments (50%), a midterm (15%), class participation (15%), written project proposals (5%) and project presentations (5%) and field trip report (10%).

Attendance policy: The course relies upon student participation in class and thus, attendance is expected. Absences will incur a grade penalty. Students who are unable to attend class for health or other personal reasons should reach out to the instructors.

Statement on policy for students with disabilities: http://www.college.columbia.edu/rightsandresponsibilities

Statement of academic integrity: Academic dishonesty is a serious offense and will not be tolerated in the class. Students are expected to reference sources appropriately in any work, including reference to third party software tools used in assignments or projects. Violation of the rules of academic integrity (e.g., plagiarizing materials) from Columbia College or the Graduate School of Arts and Sciences, will result in automatic failure of the course. Rules and consequences are outlined in Columbia College's Faculty Statement on Academic Integrity**:** http://www.college.columbia.edu/faculty/resourcesforinstructors/academicintegrity/statement

Schedule

Session: 1
Topic: Intro to course + Codio + Genome biology – organization and structure
Date: 1/23/2019 (Wed.)
Reading:

- International Human Genome Sequencing Consortium. (2001) Initial sequencing and analysis of the human genome. Nature 409:860-921 https://www.nature.com/articles/35057062
- Daniel P. Clark. (2013) Chapter 4: Genomes and DNA, from *Molecular Biology*. pp. 94-124 https://www.clinicalkey.com/#!/content/book/3-s2.0-B9780123785947000044

Assignment: Computational notebook 1 (unix, notebooks, grep, ge)

Session: 2
Topic: Jupyter review, papers discussion, blast and orthomcl introduction (sequence comparison algorithms)
Date: 1/28/2019
Reading:

- Madden, Thomas. 2013. *The BLAST Sequence Analysis Tool*. National Center for Biotechnology Information (US). https://www.ncbi.nlm.nih.gov/books/NBK153387/.
- Yandell, Mark, and Daniel Ence. 2012. "A Beginner's Guide to Eukaryotic Genome Annotation." *Nature Reviews Genetics* 13 (5): 329–42. https://doi.org/10.1038/nrg3174.
- Simão, Felipe A., Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. 2015. "BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs." *Bioinformatics* 31 (19): 3210–12. https://doi.org/10.1093/bioinformatics/btv351.

Assignment: Computational notebook 2 (blast, orthomcl, kmers, python, APIs)

Session: 3
Topic: Genetic variation and meiosis
Date: 1/30/2019
Reading

- Griffiths AJF, Miller JH, Suzuki DT, et al. (2000) Recombination, from *An Introduction to Genetic Analysis*. 7th ed. https://www.ncbi.nlm.nih.gov/books/NBK21889/,
- Griffiths AJF, Miller JH, Suzuki DT, et al. (2000) Linkage mapping by recombination in humans, from *An Introduction to Genetic Analysis*. 7th ed. https://www.ncbi.nlm.nih.gov/books/NBK21799/
- Lichten, M. (2015) Putting the breaks on meiosis. *Science (80-.).***350,**913. http://science.sciencemag.org/content/350/6263/913

Assignment: Computational notebook 3 (simulation of meiosis)

Session: 4

Topic: Genetic disease inheritance and meiosis

Date: 2/4/2019

Reading: Case study of genetic disease inheritance

- Scholl, U. I. et al. (2018) CLCN2 chloride channel mutations in familial hyperaldosteronism type II. *Nat. Genet*. 50, 349–354. https://www.nature.com/articles/s41588-018-0048-5).
- Mukherjee, S. Runs in the Family. *The New Yorker*. March 28, 2016 issue. https://www.newyorker.com/magazine/2016/03/28/the-genetics-of-schizophrenia

Assignment: Computational notebook 4 (reproduce an analysis from disease study).

Session: 5

Topic: Genetic linkage and coalescent histories

Date: 2/6/2019

Reading:

- Rosenberg, Noah A., and Magnus Nordborg. 2002. “Genealogical Trees, Coalescent Theory and the Analysis of Genetic Polymorphisms.” *Nature Reviews Genetics* 3 (5): 380–90. https://doi.org/10.1038/nrg795.
- Felsenstein, Joseph, Mary K. Kuhner, Jon Yamato, and Peter Beerli. 1999. “Likelihoods on Coalescents: A Monte Carlo Sampling Approach to Inferring Parameters from Population Samples of Molecular Data.” *Lecture Notes-Monograph Series* 33: 163–85.

Assignment: Computational notebook 5: (Coalescent simulations in msprime)

Session: 6

Date: 2/11/2019

Topic: Gene trees, species trees, phylogenetics.

Reading:

- Degnan, James H., and Noah A. Rosenberg. 2009. “Gene Tree Discordance, Phylogenetic Inference and the Multispecies Coalescent.” *Trends in Ecology & Evolution* 24 (6): 332–40. https://doi.org/10.1016/j.tree.2009.01.009.
- https://msprime.readthedocs.io/en/stable/tutorial.html

Assignment: Computational notebook 6: (reproduce results from gene tree study, measure genetic distances between markers)

Session: 7
Topic: Introduction to next-gen and Illumina
Date: 2/13/2019
Reading: BWA and read mapping

- Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25 (16): 2078–79. https://doi.org/10.1093/bioinformatics/btp352.
- Trapnell, C. & Salzberg, S. L. (2009) How to map billions of short reads onto genomes. Nat. Biotechnol. 27, 455–457. https://www.nature.com/articles/nbt0509-455.pdf

Assignment: Computational notebook 7: (read mapping, sam/bam formats, kmers, assembly)

Session: 8
Topic: Genome assembly -- de novo shotgun
Date: 2/18/2019
Reading: debruijn graphs, heterozygosity, doubled-haploids, metagenomics

- Sharpton, T. J. (2014) An introduction to the analysis of shotgun metagenomic data. Front. Plant Sci. 5, 1–14. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4059276/pdf/fpls-05-00209.pdf

Assignment: Computational notebook 8: (debruijn graphs, assemble yeast genome).

Session: 9
Topic: Genome assembly -- single Molecule
Date: 2/20/2019
Reading:

- Gordon, D. et al. (2016) Long-read sequence assembly of the gorilla genome. Science (80-.). 352. http://science.sciencemag.org/content/352/6281/aae0344
- Loomis, E. W. *et al.* (2013) Sequencing the unsequenceable : Expanded CGG-repeat alleles of the fragile X gene. *Genome Res.*121–128. https://genome.cshlp.org/content/23/1/121.full.pdf+html

Assignment: Computational notebook 9: (PoreCamp demo code).

Session: 10
Topic: Genome assembly -- single Molecule
Date: 2/25/2019
Reading: Hybrid assembly methods

- Sović, Ivan, Krešimir Križanović, Karolj Skala, and Mile Šikić. 2016. "Evaluation of Hybrid and Non-Hybrid Methods for de Novo Assembly of Nanopore Reads." *Bioinformatics* 32 (17): 2582–89. https://doi.org/10.1093/bioinformatics/btw237.
- Koren, Sergey, Brian P. Walenz, Konstantin Berlin, Jason R. Miller, Nicholas H. Bergman, and Adam M. Phillippy. 2017. "Canu: Scalable and Accurate Long-Read Assembly via Adaptive k-Mer Weighting and Repeat Separation." *Genome Research*, March, gr.215087.116. https://doi.org/10.1101/gr.215087.116.

Assignment: Computational notebook 10: (Canu and Falcon-unzip examples).

Session: 11
Topic: Genome Assembly synthesis and comparison
Date: 2/27/2019
Reading:

- Li, Fay-Wei, and Alex Harkess. n.d. "A Guide to Sequence Your Favorite Plant Genomes." *Applications in Plant Sciences* 6 (3): e1030. https://doi.org/10.1002/aps3.1030.

Assignment: Computational notebook 11: (Genome alignment)

Session: 12
Topic: Genetic comparisons: VCF, genome scans
Date: 3/4/2019
Reading: Fst, Dxy, ABBA-BABA

- Fumagalli, M. et al. (2015) Greenlandic Inuit show genetic signatures of diet and climate adaptation. Science (80-.). 349, 1343. http://science.sciencemag.org/content/sci/349/6254/1343.full.pdf
- Yi, X. et al. (2010) Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude. Science (80-.).329,75. http://science.sciencemag.org/content/329/5987/75

Assignment: Computational notebook 12: (Measure Fst in sliding windows on a VCF)

Session: 13
Topic: Genetic comparisons; phylogenetics
Date: 3/6/2019
Reading:

- McCormack, John E., Michael G. Harvey, Brant C. Faircloth, Nicholas G. Crawford, Travis C. Glenn, and Robb T. Brumfield. 2013. "A Phylogeny of Birds Based on Over 1,500 Loci Collected by Target Enrichment and High-Throughput Sequencing." *PLOS ONE* 8 (1): e54848. https://doi.org/10.1371/journal.pone.0054848.
- Federman, Sarah, Michael J. Donoghue, Douglas C. Daly, and Deren A. R. Eaton. 2018. "Reconciling Species Diversity in a Tropical Plant Clade (Canarium, Burseraceae)." *PLOS ONE* 13 (6): e0198882. https://doi.org/10.1371/journal.pone.0198882.

Assignment: Computational notebook 13: (UCE, transcriptomes, RADseq Assembly)

Session: 14
Topic: Midterm Exam Review
Date: 3/11/2019
Reading: None
Assignment: Midterm

Session: 15
Topic: Genetic comparisons; phylogenetics
Date: 3/13/2019
Reading:
Assignment: Computational notebook 14: (UCE, transcriptomes, RADseq notebooks)

Session: 16
Topic: de novo mutations
Date: 3/25/2019
Reading: Mutation rate.

- Turner, T. N. et al. (2017) Genomic Patterns of De Novo Mutation in Simplex Autism. Cell 171, 710–722.e12. https://www.cell.com/action/showPdf?pii=S0092-8674%2817%2931006-1
- Gao, Z., Moorjani, P., Amster, G. & Przeworski, M. (2018) Overlooked roles of DNA damage and maternal age in generating human germline mutations. bioRxiv. https://www.biorxiv.org/content/early/2018/05/22/327098

Assignment: Computational notebook 15: (samtools to map mutations in trios)

Session: 17
Topic: de novo mutations
Date: 3/27/2019
Reading: None
Assignment: Computational notebook 16: (samtools to map mutations in trios)

Session: 18
Topic: mapping traits
Date: 4/1/2019
Reading: GWAS classic reading

- Risch, N. & Merikangas, K. (1996) The Future of Genetic Studies of Complex Human Diseases. Science (80-.). 273, 1516–1517. http://science.sciencemag.org/content/sci/273/5281/1516.full.pdf
- Visscher, P. M. et al. (2017) 10 Years of GWAS Discovery: Biology, Function, and Translation. Am. J. Hum. Genet. 101, 5–22. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5501872/

- Hamer, D. H. & Sirota, L. (2000) Beware the chopsticks gene. Mol. Psychiatry 5, 11–13. https://www.nature.com/articles/4000662

Assignment: Computational notebook 17: (plink examples intro)

Session: 19
Topic: mapping traits
Date: 4/3/2019
Reading: GWAS modern applied example in humans

- Kathiresan S, Willer CJ, Peloso G, et al. (2009) Common variants at 30 loci contribute to polygenic dyslipidemia. Nature genetics. 41(1):56-65. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3178302/

Assignment: Computational notebook 18: (reproduce study)

Session: 20
Topic: counting methods
Date: 4/8/2019
Reading: RNAseq, Chipseq, HiC

- Chiu, R. W. K. et al. (2008) Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma. Proc. Natl. Acad. Sci. U. S. A. 105, 20458–20463. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2600580/pdf/zpq20458.pdf
- Buenrostro, Jason, Beijing Wu, Howard Chang, and William Greenleaf. 2015. “ATAC-Seq: A Method for Assaying Chromatin Accessibility Genome-Wide.” *Current Protocols in Molecular Biology / Edited by Frederick M. Ausubel ... [et Al.]* 109 (January): 21.29.1-21.29.9. https://doi.org/10.1002/0471142727.mb2129s109.

Assignment: Computational notebook 19: (reproduce study)

Session: 21
Topic: counting methods
Date: 4/10/2019
Reading:

- Stegle, Oliver, Sarah A. Teichmann, and John C. Marioni. 2015. “Computational and Analytical Challenges in Single-Cell Transcriptomics.” *Nature Reviews Genetics* 16 (3): 133–45. https://doi.org/10.1038/nrg3833.

Assignment: Computational notebook 20: (kallisto to measure gene expression)

Session 22:
Project presentations
Date: 4/15/2019

Session 23:

Project presentations
Date: 4/17/2019

Session 24:
Date: 4/22/2019
Prepare for field trip: Laboratory part I: DNA extractions, pipetting, library preparations.
Assignment: Write code to blast species based on ONT data.

Session 25:
Date: 4/24/2019
Prepare for field trip: Laboratory part II: ONT sequencers.
Discussion: improving code, hypotheses to test in field.
Assignment: Test refined code

Session 26:
Date: 4/26/2019 (Special Friday Meeting)
Class meets on Friday for field trip instead of normal time.
Field trip: Trap rodents, collect plants, identify species with keys and using MinION sequencing.

Session 27:
Date: 5/1/2019
Review results of field trip data
Assignment: Write report

Session 28:
Date: 5/6/2019
Review results of field trip reports. Discussion of genomics best practices and future expectations.