2019/01/21

EEEB GU4055 Principles and applications of modern DNA sequencing

Term taught: Spring 2019
Class times: Mondays and Wednesdays, 11:40am-12:55pm.
Classroom location: Pupin Laboratories Room 420
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: Tuesday 12-1pm

Deren Eaton
de2356@columbia.edu
(212) 851 4064
Schermerhorn Extension 1007
1200 Amsterdam Ave.
Office hours: Tuesday 12-1pm

TA:

Natalie Niepoth
natalie.niepoth@columbia.edu
Jerome L. Greene Science Center
3227 Broadway, L3-051
Office hours: Friday 4-5pm in E3B conference room (Sch. Ext. 1015)

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 online 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 assignments (50%), a midterm (15%), class participation (15%), written project proposals (5%) and project presentations (5%) and field trip report (10%). Class participation consists on answering online polling questions and asking questions about course material. All assignments, midterms, and written proposals need to be completed in order to pass the course. No points for turning them in late.

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:
If you are a student with a disability and have a Disability Services-certified 'Accommodation Letter' please contact the instructors before the course starts to confirm your accommodation needs. If you believe that you might have a disability that requires accommodation, you should contact Disability Services at 212-854-2388 and disability@columbia.edu.

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. Students are allowed to discuss homework assignments but should respond to questions and tasks on their own, not using a group answer. 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: Intro to Course, Codio, Unix, Jupyter notebooks, and genome biology
Date: 1/23/2019 (Wed.)
Readings:

- Shendure et al. 2017. DNA sequencing at 40: past, present and future: https://www.nature.com/articles/nature24286

Assignment: Computational notebooks

Session: 2: Intro to Python objects, intro to homology/BLAST.
Date: 1/28/2019 (Mon.)
Readings:

- The Official Python Tutorial. Chapters: 1,3,4,5,6,7: https://docs.python.org/3/tutorial/ (Don't worry, it doesn't take long to read these chapters!)

Assignment: Computational notebook

Session 3: Python advanced, Python data science tools, genome structure
Date: 1/30/2019 (Wed.)
Readings:

- 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.

- Vanderplas, J. 2016. Python Data Science Handbook. Chapters 1-4: https://jakevdp.github.io/PythonDataScienceHandbook/

Assignment: Computational notebook

Session 4: Blast, homology, substitution models, alignment
Date: 2/04/19 (Mon.)
Readings:

- Waterhouse, Robert M., Fredrik Tegenfeldt, Jia Li, Evgeny M. Zdobnov, and Evgenia V. Kriventseva. 2013. "OrthoDB: A Hierarchical Catalog of Animal, Fungal and Bacterial Orthologs." *Nucleic Acids Research* 41 (Database issue): D358–65. https://doi.org/10.1093/nar/gks1116.

Assignment: Computational notebooks

Session 5: Phylogenetic inference, Sanger, subprocess
Date: 2/06/19 (Wed.)
Readings:

- 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 notebooks

Session: 6
Topic: Recombination and meiosis
Date: 2/11/2019 (Mon.)
Readings:

- 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

Session: 7
Topic: Genetic disease inheritance
Date: 2/13/2019 (Wed.)
Readings:

- 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

Session: 8
Topic: Introduction to next-gen sequencing and Illumina
Date: 2/18/2019 (Mon.)
Reading:

- 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

Session: 9
Topic: Long-read sequencing technologies
Date: 2/20/2019 (Wed.)
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

Session: 10
Topic: Shotgun genome assembly
Date: 2/25/19 (Mon.)
Reading:

- Compeau, Phillip E. C., Pavel A. Pevzner, and Glenn Tesler. 2011. "How to Apply de Bruijn Graphs to Genome Assembly." Nature Biotechnology 29 (11): 987–91. https://doi.org/10.1038/nbt.2023.

Assignment: Computational notebooks

Session: 11
Topic: Short and long read assembly synthesis: combining and comparing
Date: 2/27/19 (Wed.)
Reading:

- Lightfoot et al. (2017) Single-molecule sequencing and Hi-C-based proximity-guided assembly of amaranth (Amaranthus hypochondriacus) chromosomes provide insights

into genome evolution. BMC Biology 15:74 https://bmcbiol.biomedcentral.com/articles/10.1186/s12915-017-0412-4
- 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 notebooks

Session: 12
Topic: Genetic comparisons; paralogs, duplications, loss
Date: 3/4/19 (Mon.)
Reading:
- Panchy, Nicholas, Melissa Lehti-Shiu, and Shin-Han Shiu. 2016. “Evolution of Gene Duplication in Plants.” Plant Physiology 171 (4): 2294–2316. https://doi.org/10.1104/pp.16.00523.

Assignment: Computational notebooks

Session: 13
Topic: Reduced representation sequencing: Anchored enrichment
Date: 3/6/19 (Wed.)
Reading:
- Faircloth, Brant C., John E. McCormack, Nicholas G. Crawford, Michael G. Harvey, Robb T. Brumfield, and Travis C. Glenn. 2012. “Ultraconserved Elements Anchor Thousands of Genetic Markers Spanning Multiple Evolutionary Timescales.” *Systematic Biology* 61 (5): 717–26. https://doi.org/10.1093/sysbio/sys004.
- Gasc, Cyrielle, Eric Peyretaillade, and Pierre Peyret. 2016. “Sequence Capture by Hybridization to Explore Modern and Ancient Genomic Diversity in Model and Nonmodel Organisms.” *Nucleic Acids Research* 44 (10): 4504–18. https://doi.org/10.1093/nar/gkw309.

Assignment: Computational notebooks

Session: 14
Topic: Reduced representation sequencing: RAD-seq
Date: 3/11/19 (Mon.)
Reading:
- Hohenlohe, Paul A., Susan Bassham, Paul D. Etter, Nicholas Stiffler, Eric A. Johnson, and William A. Cresko. 2010. “Population Genomics of Parallel Adaptation in Threespine Stickleback Using Sequenced RAD Tags.” *PLOS Genetics* 6 (2): e1000862. https://doi.org/10.1371/journal.pgen.1000862.
- Andrews, Kimberly R., Jeffrey M. Good, Michael R. Miller, Gordon Luikart, and Paul A. Hohenlohe. 2016. “Harnessing the Power of RADseq for Ecological and Evolutionary Genomics.” Nature Reviews Genetics 17 (2): 81. https://doi.org/10.1038/nrg.2015.28.

Assignment: Computational notebooks

Session: 15
Midterm exam
Date: 3/13/2019 (Wed.)
Reading: None
Assignment: None

Session: 16
Topic: Genetic differentiation (variant calling and F_{ST})
Date: 3/25/2019 (Mon.)
Reading:

- 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

Session: 17
Topic: *De novo* mutations
Date: 3/27/2019 (Wed.)
Readings:

- 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

Session: 18
Topic: Low-frequency (e.g. somatic/cancer) mutations
Date: 4/1/2019 (Mon.)
Readings:

- Stephens, P. J. et al. (2009) Complex landscapes of somatic rearrangement in human breast cancer genomes. Nature 462, 1005–1010. https://www.nature.com/articles/nature08645.pdf
- Wang, J. et al. (2016) Clonal evolution of glioblastoma under therapy. Nature Genetics **48,**768–776. https://www.nature.com/articles/ng.3590.pdf

Assignment: Computational notebook

Session: 19
Topic: Applications of low and high coverage sequencing

Date: 4/3/2019 (Wed.)
Readings:

- Joe Pickrell: It is time to replace genotyping arrays with sequencing. Gencove Blog https://medium.com/the-seeq-blog/it-is-time-to-replace-genotyping-arrays-with-sequencing-73535efa66ed
- 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

Session: 20
Topic: Counting methods
Date: 4/8/2019 (Mon.)
Readings:

- 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
- Mele, M. et al. (2015) The human transcriptome across tissues and individuals. Science 348(6235): 660-665 http://science.sciencemag.org/content/sci/348/6235/660.full.pdf

Assignment: Computational notebook

Session: 21
All presentations due today on Courseworks by 11am
Project presentations
Date: 4/10/2019 (Wed.)

Session: 22
Project presentations
Date: 4/15/2019 (Mon.)
Reading:

- 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.

Session: 23
Date: 4/17/2019 (Wed.)

Prepare for field trip: Laboratory part I: DNA extractions, pipetting, library preparations.
Assignment: Code to analyze ONT data.

Session: 24
Date: 4/22/2019 (Mon.)
Prepare for field trip: Laboratory part II: ONT sequencers.
Discussion: improving code, hypotheses to test in field
Assignment: Test refined code with your sequencing data

Session: 25
Date: 4/24/2019 (Wed.)
Prepare for field trip: Laboratory review
Assignment: Adapt code to work in the field

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 (Wed.)
Preliminary trip reports due today before class (submit on Courseworks)
Review results of field trip data
Assignment: Write report

Session: 28
Date: 5/6/2019 (Mon.)
Final trip reports due today before class (submit on Courseworks)
Review results of field trip reports. Discussion of genomics best practices and future expectations.