Genomic Data Analysis

(BioSci 469) – Spring 2023

Class location: Lapham Hall S386

Dr. Peter Dunn
Office: LAP S497 E-mail: pdunn@uwm.edu

Live on Teams: Mondays at 11:30 - 1:30.
(recorded for those who cannot make it)

Course Description: This course is designed for students interested in learning current techniques for the analysis of **large-scale** genomic data sets. High-throughput sequencing has become widespread in biology and medicine over the past decade due to both rapid technological advances and decreases in overall cost. The class will discuss study design, choice of methods, including practical issues of sequencing facilities, cost and computing resources, and then proceed to **hands-on data analyses** used in whole genome (re)sequencing, transcriptome analysis, and reduced-representation sequencing (e.g., RAD-seq, GBS). The schedule below gives some introductory topics. Additional topics will be covered depending on the interests of students and time available.

This course is designed to build competence in the computing and statistical methods for analyzing high-throughput genomic data. **The only background assumed is a basic knowledge of statistics and genetics, familiarity with your computer and interest in learning current genomic methods.** Knowledge of Linux operating systems is desirable, but not necessary. All that is required is a willingness to work hard (ie, not quit after the first [or second] error message 😉).

Course Objectives: The primary objectives of this course are:
• to learn how genomic data are being used in biology, particularly evolutionary biology.
• to become familiar with the software and databases available for bioinformatics
• to develop the ability to formulate and investigate genomic research questions, and to effectively communicate your questions, methods, and results.

Prerequisites: Genetics (BioSci 325 or equivalent) and Biostatistics (465 or equivalent) or consent of instructor.

Computers: In the past we used a computer lab on campus, but it is no longer available, so now you will need to have access to a computer that can connect to the University servers.

Option 1 (the easy way).

To use the software under this option, you will need two programs (Putty and Filezilla) to connect to the server (remote computer) on campus. The server is called "peregrine". I previously used a server called "unixdev1", so if you see that name anywhere, just think "peregrine". These two programs allow you to move files and run programs on the server using your own computer (or one in the classroom). Note that **if you connect from off campus, you will also need to install a software program that provides a secure link to the server (ie, a virtual private network; VPN)**. More information about the software and how to connect will be presented in class.

Option 2 (more work to install, but no connection required).

It is also possible to do the exercises in class without connecting to the peregrine server (Option 1 above). Much of the software we use in class is also available in a (relatively) easy to install version through the **BioStars Handbook**, which is also the source of some exercises. So if you want to do the exercises in class with your own computer (a typical laptop is fine), I would recommend purchasing the license for the BioStars Handbook (see below) and following the directions under ""**2. Getting Started. How to set up your computer**."
Note that it is much easier to set up a Mac than a Windows (10+) computer because Macs already use a version of Linux, but you have to use additional software to get Linux on Windows (ie, use the "Linux subsystem for Windows" or install an emulation program like VMware Workstation Player). Talk to me if you want to use Option 2.

Textbook**: There is no required textbook for the course, but I highly recommend** "**The Biostar Handbook: A Beginner's Guide to Bioinformatics" (2017)** by Istvan Albert (available online at: https://read.biostarhandbook.com/). There is now a second edition. You get the updates with your license, which you can purchase at: https://biostar.myshopify.com/. A six month student license is $25. A 2 year license is $35. Additional papers and references are listed below.

Credits and Evaluation: This is a 2-credit course. Grades are based on assignments completed each week during the online lectures (this may be modified depending on the availability of computer resources). **There is no Final Exam.**
Graduate students will also receive 10% of their grade based on a short research project/paper/talk (2-3 pages for paper/12 min talk) to be agreed upon with the instructor. Final assessment is based on the cumulative grades, as follows:

Undergraduates: 10 in-class assignments 100% (10 pts each).

Graduate students: 10 in-class assignments 90% (~9 pts each), **research project 10%**. Graduate students will write a 2-3 page report on a software program that they tested or a topic of interest (chosen after consulting the instructor).

Attendance and Assignment requirements: *Assignments will be completed during the week of a lecture.*

Time investment for this course: Students should plan to spend an average of 4 hours outside of class per week reading in preparation for in-class assignments. This amount of

time is based on the campus credit hour policy (Faculty Document # 2838); ie, two hours out-of-class work for each credit hour per week of class.

Letter grades will be assigned based on the final total points listed below.

A	92 – 100%		C	71 – 75%
A-	89 – 91%		C-	68 – 70%
B+	86 – 88%		D+	65 – 67%
B	82 – 85%		D	61 – 64%
B-	79 – 81%		D-	56 – 60%
C+	76 – 78%		F	0 - 55%

Need for Special Accommodation-- Students who require note-taking or test-taking accommodations in order to meet any of the requirements of this course, please contact me as soon as possible to make suitable arrangements.

Schedule of Topics:

Topics (usually 1 week each; subject to change)

1. Introduction. What is your question? DNA or RNA? What technique should I use? Methods to analyze genomes, transcriptomes, meta-genomes & SNPs. What sequencers, data and software are available? Lec/Lab 23 Jan 2033. No Exercise due.

2. How do I get sequences and genomes? NCBI, Ensembl etc. FASTA and other files. **Exercise #1**.- using NCBI and genome browsers. Lec/Lab 30 Jan. Exercise 1 due 3 Feb by 5 PM. Upload to Canvas

3. Computing resources and capabilities (stand-alone, UWM cluster, Amazon and other cloud resources). How to install software without going crazy.
   How to use Linux. **Exercise #2** – using unixdev1, Filezilla, and Putty, and cleaning sequences with Trimmomatic. Lec/Lab 6 Feb. Exercise 2 due 10 Feb by 5 PM. Upload to Canvas

4. Short-read sequence alignment to reference genomes. SNP calling. **Exercise #3**. - using BWA & SAMtools to align Ebola virus sequences and bcfTools to call SNPs. Lec/Lab 13 Feb. Exercise due 17 Feb by 5 PM. Upload to Canvas

5. SNP calling / genotyping and viewing. **Exercise #4**. - Using IGV and the UCSC Genome Browser to view alignments and SNPs. Lec/Lab 20 Feb. Exercise due 24 Feb by 5 PM. Upload to Canvas

6. Searching databases using BLAST on Ebola sequences. **Exercise #5** Lec/Lab 27 Feb. Exercise due 3 March by 5 PM. Upload to Canvas

7. Genome wide association studies (GWAS). **Exercise #6**.- using PLINK. Lec/Lab 6 Mar. Exercise due 10 March by 5 PM. Upload to Canvas

8. Transcriptome analysis overview. de novo or reference-guided. Sampling strategies. **Exercise # 7** Lec/Lab 13 Mar. Exercise due 17 March by 5 PM. Upload to Canvas

19-26 March SPRING BREAK

9. Differential gene expression. **Exercise # 8**.- using the new 'Tuxedo' Suite on Zika data. Lec/Lab 3 27 Mar. Exercise due 31 Mar by 5 PM. Upload to Canvas

10. Gene ontology. **Exercise #9**.- using EdgeR (Chen et al. tutorial). Lec/Lab 3 Apr. Exercise due Exercise due 7 April by 5 PM. Upload to Canvas

11. Network analysis with KEGG, DAVID, Cytoscape etc. **Exercise #10.** Lec/Lab 10 Apr. Exercise due 14 April by 5 PM. Upload to Canvas

12. Metagenomics: Microbiome analysis of mice with Mothur and Microbiome Analyst. Exercise #11. Lec/Lab 17 Apr. Exercise due 21 April by 5 PM. Upload to Canvas

References for corresponding Topics

Some of the class material is based on: "**The Biostar Handbook: A Beginner's Guide to Bioinformatics" (2017)** by Istvan Albert (available online at: https://read.biostarhandbook.com/)

1. Bild, A. H., J. T. Chang, W. E. Johnson, and S. R. Piccolo. 2014. A field guide to genomics research. PLoS Biol 12:e1001744.
   Kell, D. B. and S. G. Oliver. 2004. Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. Bioessays 26:99-105.
   Ekblom, R. and J. Galindo. 2011. Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity 107:1-15.
   Todd, E. V., M. A. Black, and N. J. Gemmell. 2016. The power and promise of RNA-seq in ecology and evolution. Mol. Ecol. DOI:10.1111/mec.13526.
2. Unix primer for Biologists: http://korflab.ucdavis.edu/unix_and_Perl/
3. Chapter 7 in The Biostar Handbook.
4. Gire, S. K., et al. 2014. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science: DOI: 10.1126/science.1259657.
5. Andrews, K. R., J. M. Good, M. R. Miller, G. Luikart, and P. A. Hohenlohe. 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. Nat Rev Genet 17:81-92.
6. McKinney, G. J., W. A. Larson, L. W. Seeb, and J. E. Seeb. 2016. RADseq provides unprecedented insights into molecular ecology and evolutionary genetics: comment on Breaking RAD by Lowry et al. (2016). Mol. Ecol. Resour. DOI:10.1111/1755-0998.12649.
   Lowry, D. B., S. Hoban, J. L. Kelley, K. E. Lotterhos, L. K. Reed, M. F. Antolin, and A. Storfer. 2016. Breaking RAD: An evaluation of the utility of restriction site associated DNA sequencing for genome scans of adaptation. Mol. Ecol. Resour. DOI:10.1111/1755-0998.12635.

7. Toews, D. P., S. A. Taylor, R. Vallender, A. Brelsford, B. G. Butcher, P. W. Messer, and I. J. Lovette. 2016. Plumage genes and little else distinguish the genomes of hybridizing warblers. Curr Biol 26:2313-2318.
8. Kukurba, K. R. and S. B. Montgomery. 2015. RNA Sequencing and Analysis. Cold Spring Harbor protocols: DOI: 10.1101/pdb.top084970.
Martin, J. A. and Z. Wang. 2011. Next-generation transcriptome assembly. Nat Rev Genet 12:671-682.
Schurch, N. J., et al. 2016. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA 22:839-851.
Wang, Z., M. Gerstein, and M. Snyder. 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nature Genetics Reviews 10:57-63.
9. Trapnell, C., et al. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protocols 7:562-578.
Chapters 19-21 in the Biostars handbook.
Wolf, J. B. W. 2013. Principles of transcriptome analysis and gene expression quantification: an RNA-seq tutorial. Mol. Ecol. Resour. DOI:10.1111/1755-0998.12109.
10. Chen, Y., A. Lun, and G. Smyth. 2016. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. F1000Research 5:1438.
Zhang, Z. H., et al. 2014. A Comparative Study of Techniques for Differential Expression Analysis on RNA-Seq Data. PLoS ONE 9:e103207.
11. Huang, D. W., B. T. Sherman, and R. A. Lempicki. 2008. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4:44-57.
Raudvere et al. (2019) **g:Profiler: a web server for functional enrichment analysis and conversions of gene lists** Nucleic Acids Research 2019; doi:10.1093/nar/gkz369 https://biit.cs.ut.ee/gprofiler/gost
12. KEGG: Kyoto Encyclopedia of Genes and Genomes. https://www.genome.jp/kegg/
Langfelder, P. andHorvath, S. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics,* 9**,** 559.
13. Knight, R., et al. 2018. Best practices for analysing microbiomes. *Nature Reviews Microbiology*.
Thompson, L.R.,et al. and The Earth Microbiome Project 2017. A communal catalogue reveals Earth's multiscale microbial diversity. *Nature,* 551**,** 457-463.
Schloss, P.D. et al. 2009. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Applied and Environmental Microbiology,* 75**,** 7537-7541. https://mothur.org/

University Guidelines of Interest

See: http://uwm.edu/secu/wp-content/uploads/sites/122/2016/12/Syllabus-Links.pdf

1. *Students with disabilities.* Notice to these students should appear prominently in the syllabus so that special accommodations are provided in a timely manner.
http://www4.uwm.edu/arc

2. *Religious observances.* Accommodations for absences due to religious observance should be noted. http://www4.uwm.edu/secu/docs/other/S1.5.htm

3. *Students called to active military duty.* Accommodations for absences due to call-up of reserves to active military duty should be noted.
Students: http://www4.uwm.edu/academics/military.cfm

4. *Incompletes***.** A notation of "incomplete" may be given in lieu of a final grade to a student who has carried a subject successfully until the end of a semester but who, because of illness or other unusual and substantiated cause beyond the student's control, has been unable to take or complete the final examination or to complete some limited amount of term work. https://www4.uwm.edu/secu/docs/other/S_31_INCOMPLETE_GRADES.pdf

5. *Discriminatory conduct (such as sexual harassment*). Discriminatory conduct will not be tolerated by the University. It poisons the work and learning environment of the University and threatens the careers, educational experience, and well-being of students, faculty, and staff. https://www4.uwm.edu/secu/docs/other/S_47_Discrimina_duct_Policy.pdf

6. *Academic misconduct.* Cheating on exams or plagiarism are violations of the academic honor code and carry severe sanctions, including failing a course or even suspension or dismissal from the University. http://uwm.edu/academicaffairs/facultystaff/policies/academic-misconduct/

7. C*omplaint procedures.* Students may direct complaints to the head of the academic unit or department in which the complaint occurs. If the complaint allegedly violates a specific university policy, it may be directed to the head of the department or academic unit in which the complaint occurred or to the appropriate university office responsible for enforcing the policy. https://www4.uwm.edu/secu/docs/other/S_47_Discrimina_duct_Policy.pdf

8. *Grade appeal procedures*. A student may appeal a grade on the grounds that it is based on a capricious or arbitrary decision of the course instructor. Such an appeal shall follow the established procedures adopted by the department, college, or school in which the course resides or in the case of graduate students, the Graduate School. These procedures are available in writing from the respective department chairperson or the Academic Dean of the College/School. http://www4.uwm.edu/secu/docs/other/S28.htm

9. O*ther* The final exam requirement, the final exam date requirement, etc. http://www4.uwm.edu/secu/docs/other/S22.htm