Advanced Computing and Bioinformatics for Conservation Genomics. CSU FWCB/BZ 5XX

Instructors: *Eric C. Anderson and Kristen C. Ruegg*

Site Last Updated: *2019-12-05*

Contents

Course Overview

Welcome to the course web page for Advanced Computing and Bioinformatics for Conservation Genomics, an experimental course being offered for the first time in Winter Semester 2020. We will be covering computing, analysis and data-organization strategies for bioinformatics and analysis of high-throughput sequencing data for ecology, evolution, and conservation.

Modern high-throughput sequencing can provide extraordinary amounts of data, enabling researchers to tackle a wide range of questions and problems in ecology, evolution, conservation, and fisheries and wildlife management. Preparing and processing these data for use, however, requires multiple bioinformatic steps, and subsequent analysis of these large, complex data sets must rely on specialized computer programs. Mastering these skills presents a high bar for students originating from outside of computer science and related fields. At present, in many institutions, such skills are typically learned from peers within experienced laboratories, or through a variety of workshops. This course aims to comprehensively teach the computing and analytical skills necessary to use genomic data from high-throughput sequencing in the context of ecological research. During the first 2/3 of the course, the focus is on aligning DNA sequence data and identifying variants across multiple individuals. In the last 1/3 of the course we consider a series of case studies in how such data are used to make inference for applications in fisheries, wildlife, and conservation. Outside of the bioinformatic utilities that run within a Unix framework, emphasis is placed on using the R programming language and RStudio for project management and documentation.

The proposed course topics appear, by week, in the table below. Each week of the course is structured as a different chapter in the navigation panel on the left. In order to figure out what we are doing in the course each week, that will be the first place to check. The week's objectives, readings, and exercises will be listed there.

Course Learning Objectives

Upon successful completion of the course, students will be able to:

1. Organize and execute a complex bioinformatic data-analysis project in a manner that makes it easily understood and reproduced by others.
2. Describe the main data formats used in genomic analysis, and know how to generate and manipulate them.
3. Work with a wide range of the bioinformatic tools available in the Unix environment and understand how to script these tools into pipelines for DNA sequence alignment, variant calling, and analysis.
4. Understand how to break down complex genomic analysis projects into small, independent chunks and execute those using job arrays on a high performance computing cluster.
5. Perform a variety of computational analyses central to conservation genetics.

Assessment

Assessment will be based mostly on weekly problem sets. These will sometimes require considerable time and thought, but they will be critical for solidifying the concepts and procedures in the course. Students will also be undertaking individual projects in which they apply the skills they have learned in the course to a data set relevant in some way to their own research or to an interesting question relevant to some existing data, after discussion with the instructors (see below). Finally, students are expected to contribute to discussion and participation in the course, including (and most importantly) being helpful to one another in order to learn challenging material, together, in a supportive environment.

> "It is literally true that you can succeed best and quickest by helping others to succeed."
>
> —- Napolean Hill

Assessment Component	Percentage of Grade
Problem sets	60%
Individual analysis projects	30%
Class participation	10%

Weekly Schedule

The schedule is subject to change as the semester proceeds, but this is what we are shooting for.

Week	Lecture Component	Lab Component
1	Rstudio Projects, GitHub	Rmarkdown, git and GitHub
2	Unix, directory structure, utilities	Unix, data compression
3	Next generation sequencing, alignment conventions, FASTA, SAM	Samtools (faidx)
4	Shell scripting and awk	Shell scripting and awk
5	Remote computers, ssh, rclone, HPC, SLURM, SGE	Connecting to Summit Cluster
6	Sequence alignment, job arrays, parallelization over individuals	bwa, samtools (sort, merge, cat, index)
7	Variant calling – I, fundamental concepts	GATK, Base quality score recalibration
8	Variant calling – II, parallelization over regions	GATK, gVCFs, VCFs
9	Variant Filtering, manipulation, exploration, LD and pop-gen statistics	bcftools, bedtools, R package 'whoa', plink
10	Restriction-associated digest (RAD) sequencing	STACKS2, R package 'radiator'
11	Visualization of genomic data in space	R packages: 'ggplot', 'sf'
12	Visualization of genomic data: trees	R packages: 'ape', 'ggtree'
13	Population structure	R package 'srsStuff'
14	Inbreeding, runs of homozygosity	bcftools roh, plink
15	Genome wide association studies	ANGSD, snpEff
16	Genotype-environment association	Gradient Random Forest, RDA

Individual Projects

The purpose of the individual projects is to allow the students to use many of the skills learned, and to gain experience in preparing a reproducible research project. Some students likely already have their own data sets that they are working on, but we expect that many will not. We will be able to provide data and interesting questions to tackle from our own research. Additionally, we will

encourage students to take on related projects so that they can work together on different parts of a single question.

Chapter 1

Getting setup with Rstudio and GitHub

It might seem strange to start a bioinformatics course with a segment on RStudio, since most bioinformatics takes place on large computer clusters which can't always be configured by individual users to play well with RStudio running on a server. Not to mention the fact that R is not typically central to hard-core bioinformatic tasks in a Unix environment. However, bear with us. Using RStudio sets you up with a great mechanism for recording your work in RMarkdown Notebooks, it provides a natural way to organize projects, it makes it trivial to keep your work version-controlled, and finally, when you have done all your bioinformatics, it is great for analysis of and figure preparation from the generated data in the R language.

Required Readings

1. This
2. That
3. Etc

Recommended Readings

1. This
2. That
3. More

Exercises due by MM/DD/YYY

1. Links to notebooks that have the assignments. Will make it easy for students to download extra material for each one.

1.1 First topic for lecture

1.2 Second topic for lecture

Boing

Chapter 2

A one-week Unix crash course

Preamble, preamble, preamble.

Required Readings

1. ECA-bioinf-handbook-Chapter-4. Read the whole thing before coming to class on Tuesday, and commit the terms in the "Unix study guide" (the long table at the end) to memory.

Recommended Readings

1. This
2. That
3. More

Exercises due by MM/DD/YYY

1. Links to notebooks that have the assignments. Will make it easy for students to download extra material for each one.

2.1 First topic for lecture

2.2 Second topic for lecture

Boing

Chapter 3

Next generation sequencing and associated conventions

Preamble, preamble, preamble.

Required Readings

1. Write another chapter in the handbook for this

Recommended Readings

1. This
2. That
3. More

Exercises due by MM/DD/YYY

1. Links to notebooks that have the assignments. Will make it easy for students to download extra material for each one.

3.1 First topic for lecture

3.2 Second topic for lecture

Boing

Chapter 4

Unix scripting, awk, regular expressions

Preamble, preamble, preamble.

Required Readings

1. Handbook, chapters 5 and 6.

Recommended Readings

1. This
2. That
3. More

Exercises due by MM/DD/YYY

1. Links to notebooks that have the assignments. Will make it easy for students to download extra material for each one.

4.1 First topic for lecture

4.2 Second topic for lecture

Boing

Chapter 5

Remote computers and HPC

Preamble, preamble, preamble.

Required Readings

1. Handbook, parts of chapter 7.

Recommended Readings

1. This
2. That
3. More

Exercises due by MM/DD/YYY

1. Links to notebooks that have the assignments. Will make it easy for students to download extra material for each one.

5.1 First topic for lecture

5.2 Second topic for lecture

Boing

Chapter 6

Sequence alignment

Preamble, preamble, preamble.

Required Readings

1. Hmm…

Recommended Readings

1. This
2. That
3. More

Exercises due by MM/DD/YYY

1. Links to notebooks that have the assignments. Will make it easy for students to download extra material for each one.

6.1 First topic for lecture

6.2 Second topic for lecture

Boing