**Final Project for 500-level students**

**Important Dates:**

**November 19, 2018. Pre-proposal due**

**December 16, 2018. Final project due**

**Here we describe two possible project topics. Choose one of these two data/problem sets. See below for detailed instructions for report and expectations.**

**Project I: C. elegans single cell dataset**

**Background:**

*Caenorhabditis elegans* is a small nematode worm that is used as a model organism. One of its key characteristics is that it has ~1,000 cells total and the cells are produced by a cell lineage tree that is exactly the same for every individual (as opposed to our bodies where there is a substantial difference in the cell lineage trees between different individuals). Because of this characteristic, it is used to study developmental decisions and how small circuits of neurons work, among other studies. Wikipedia has a nice overview and the WormBook (<http://www.wormbook.org/>) is a comprehensive source of information about the organism.

Recently, in collaboration with Dr. John Murray in Genetics, we generated transcriptome data from single cells using a technology called 10X Genomics (<https://www.10xgenomics.com/>). The data (LINK) is a subset of our data collected from various parts of the worm at different stages of development. The raw data consists of cells (in rows) and counts of molecules for each gene (in columns).

**Project Specifications:**

With your group, **choose one of the design problems below**.

There are two files, one is the counts data and the second is an annotation file describing each cell type and sub-types. These annotations were created by expert curation and do not necessarily represent exact truth. That is, you may find some annotations incorrect (which might result in some isolated clusters in the unsupervised context of Question 1).

**Question 1:** Use the single cell data to create cell type classifications in an unsupervised learning setting. You may use any clustering method, combination of methods, or your own particular method. Report an assessment of accuracy by comparing appropriately against the annotation file. Note any unusual results or artifacts.

**Question 2:** Use the single cell data and the annotations files in a supervised learning setting to learn cell types and predict the types. Leave out a random selection of 20% of the cells to test your predictions. Assess accuracy. Try to minimize the number of genes needed to accurately predict the cell types.

**Notes and Resources:**

The included dataset is composed of raw counts of molecules. Single cell datasets often contain many zeros. Consider how one might treat these zeros. The raw counts have not been normalized. Consider strategies for normalizing the data to make the values comparable to each other between rows. Raw counts might not be the best scale to treat the data. Consider transformation of the data such as log transforms. The cell collected here are from various stages of development. It could be “body wall” cells may separate into early and late stages, as an example. Make efforts to biologically interpret the results.

DataSet:

C. Elegans Raw Counts: <https://canvas.upenn.edu/files/72913892/download?download_frd=1>

Labels for Data: <https://canvas.upenn.edu/files/72913893/download?download_frd=1>

**Some single cell transcriptome analysis packages:**

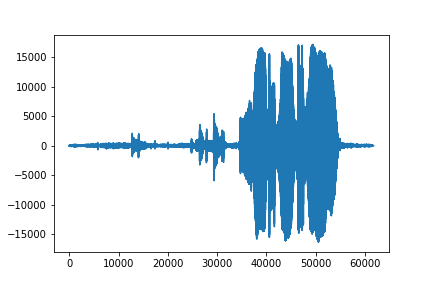
**Seurat---**<https://satijalab.org/seurat/>

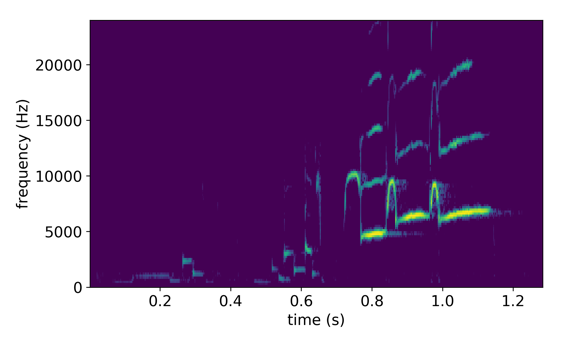
**PIVOT---**https://github.com/qinzhu/PIVOT

**Monocle---**http://cole-trapnell-lab.github.io/monocle-release/

**Project 2: Bird Song Patterns**

**Background:**

So far, we have discussed biological modeling primarily in the context of genomic data, however the principles we have discussed are broadly applicable. This project will apply these principles to behavioral data, namely, birdsong, either in Python or any other programming languages.

****Birdsong (along with other acoustic behaviors, like frog calls) have a rich history of analysis, given that any sound signal can be represented precisely in a one-dimensional vector, denoting pressure at each time point. This has allowed us to leverage the tools provided us by acoustics to assess motor coding, song similarity, vocal learning, information content, etc. Within the field, “song” usually refers specifically to the complex vocalizations produced by (mostly male) birds of the large clade Passerida. All birds also produce simpler vocalizations, generally referred to as “calls”. Within vocal learners (parrots, hummingbirds, and passerines), the song is almost always learned from conspecifics (often the fathers). Vocal learning generally refers to communication that relies on copying signals of some tutor (as seen humans). It is quite rare within the animal kingdom (seemingly limited to humans, bats, cetaceans, and 3 groups of birds). Most other animals produce all species-typical vocalizations having never heard them from a tutor.

A

**Acoustics:**

Sound is recorded as a measurement of pressure at each time (fig 1, top). The raw data is stored in a .wav file, which may be stereo (having two channels) or mono (one channel.). Stereo recordings allow us to produce the illusion of real sound (since we have two ears), however mono files are generally easier to analyze. Sounds are often visualized and analyzed using the Fourier transform, which separates a signal into its component frequencies. This gives us a matrix called a spectrogram (fig 1, bottom) with the amplitude of each frequency at every time bin. This is also largely representative of how the brain encodes sound via the cochlea.

**Analyzing Sound:**

Although sound transmission is essentially one dimensional, like DNA it is generated by higher dimensional processes, and thus we must decide how best to model the data. Song is generally modeled as a series of sub components (syllables, denoted by colored boxes in the spectrogram above.) which are parameterized into a set of acoustic features (pitch, spectral continuity, entropy, etc.). We can then compare songs by asking questions of acoustic similarity, or label syllables and compare songs on the basis of sequencing (for example, if a bird has 3 syllables in its song–we’ll call them ABC–and usually sings “ABC” we can observe variation within each syllable, A vs A’, and also ask the probability of seeing “AAC”,“BAC”, etc.). Syllable labeling can be somewhat subjective, but in general syllables are defined by continuous harmonics, and often begin and end with a short period of silence (when the bird takes a short breath).

**Project Specifications:**

With your group, **choose one of the design problems below**.

**Question 1:** Use the song data to create song classifications in an unsupervised learning setting. You may use any clustering method, combination of methods, or your own particular method. Report an assessment of accuracy by comparing appropriately against the species annotations. Note any unusual results or artifacts.

**Question 2:** Sound identification remains a challenge for both field study and laboratory work. Each study produces many hours of sound data, and hand labeling songs or syllables is a significant bottleneck and source of bias. Use the song data and annotations in a supervised learning setting to predict song class (e.g. species, calls vs song, location, etc). Leave out a random selection of 20% of the calls to test your predictions. Assess accuracy. Try to minimize the number of files needed to accurately predict your chosen classes.

**Question 3:** As with genetic data, many questions in behavior are simply questions of distance. Choose a model of distance and apply it to vocalizations in order to test a hypothesis you find interesting. You could, for example, ask whether song similarity is a function of geographic distance. Or whether song variability is higher in vocal-learners (i.e. passerida) than non vocal-learners (e.g. water-fowl). Or you might ask whether song similarity is a function of genetic distance. Annotations (sex, location, time of day, etc.) are available on Xeno-Canto.org. Phylogenetic information is available on BirdTree.org

**Bird Resources:**

**Data Set:** <https://www.xeno-canto.org/> - Hundreds of thousands of recordings of varying quality. This is an extremely powerful resource but will require some careful decisions on how to handle your data, which songs to include, and the scope of your project. This decision process should be driven by the question you have chosen and is an important part of your project. Depending on the amount of data you want to use, you will likely want to download it in some automatic way. Python handles data scraping quite well (try BeautifulSoup), and there’s actually an R package (called WarbleR) which includes a script to download files from Xeno-Canto. Other data sets exist, if you find something better suited to your question, you may use that (please check with us to be sure). Be sure to use data responsibly (e.g., provide wait times between requests to avoid overloading the server)

**Additional Resources:**

Phylogenetic tree of birds: <https://birdtree.org/>

More info on bird song: <http://www.birds.cornell.edu/Page.aspx%3Fpid%3D1059>

Analytical Packages you might consider:

* David Mets. “"SongD(KL), an automated approach to the quantitation of vocal    
  learning in the songbird". [https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006437&rev=2](https://urldefense.proofpoint.com/v2/url?u=https-3A__journals.plos.org_ploscompbiol_article-3Fid-3D10.1371_journal.pcbi.1006437-26rev-3D2&d=DwMFaQ&c=clK7kQUTWtAVEOVIgvi0NU5BOUHhpN0H8p7CSfnc_gI&r=XpGqnFKnJHmx0CQn9lpv8BoZo-WGWpT8WvlwieK2nBI&m=_x3T0XI8jJIfucd4dHm9udP-sXSwCebsDxixQVz3kEM&s=rVepqm_5yLa1lefXHj0UyAKfdAx43fpDUEq7e7b0bJo&e=)
  + Script1 (syllable counter) - <https://canvas.upenn.edu/files/72914333/download?download_frd=1>
  + Script2 (Song D(KL)) - <https://canvas.upenn.edu/files/72914334/download?download_frd=1>
  + Note, I edited these slightly to make them as is on python 3, all the code belongs to David Mets, and the original files can be found alongside the article.
* Frederic Theunnisen. “BioSound: a python library and notebooks to extract sound features for bio-acoustical analyses” [https://github.com/theunissenlab/soundsig](https://urldefense.proofpoint.com/v2/url?u=https-3A__github.com_theunissenlab_soundsig&d=DwMFaQ&c=clK7kQUTWtAVEOVIgvi0NU5BOUHhpN0H8p7CSfnc_gI&r=XpGqnFKnJHmx0CQn9lpv8BoZo-WGWpT8WvlwieK2nBI&m=_x3T0XI8jJIfucd4dHm9udP-sXSwCebsDxixQVz3kEM&s=CsmkxkEHrEdNoc9KAGQFOviWRuKwNzAg8VHSI0ycu2w&e=) [https://github.com/theunissenlab/BioSoundTutorial](https://urldefense.proofpoint.com/v2/url?u=https-3A__github.com_theunissenlab_BioSoundTutorial&d=DwMFaQ&c=clK7kQUTWtAVEOVIgvi0NU5BOUHhpN0H8p7CSfnc_gI&r=XpGqnFKnJHmx0CQn9lpv8BoZo-WGWpT8WvlwieK2nBI&m=_x3T0XI8jJIfucd4dHm9udP-sXSwCebsDxixQVz3kEM&s=gFk0CiRHxmhrndtfq5E7uur4baV3j7MKcMakB1DU5kM&e=) [https://link.springer.com/article/10.1007%2Fs10071-015-0933-6](https://urldefense.proofpoint.com/v2/url?u=https-3A__link.springer.com_article_10.1007-252Fs10071-2D015-2D0933-2D6&d=DwMFaQ&c=clK7kQUTWtAVEOVIgvi0NU5BOUHhpN0H8p7CSfnc_gI&r=XpGqnFKnJHmx0CQn9lpv8BoZo-WGWpT8WvlwieK2nBI&m=_x3T0XI8jJIfucd4dHm9udP-sXSwCebsDxixQVz3kEM&s=YouCgtpQsRdkEqueJE7G35q_Dqk-iPMPdDnzhjutOHk&e=)
* Tim Sainburg. “Generative modelling for birdsong: learning acoustic spaces for generating stimuli, labelling syllables, and visualizing sequences”. https://github.com/timsainb/AVGN
* Yarden Cohen. “Segmentation and annotation of birdsong with a hybrid recurrent-convolutional neural network”. [https://github.com/yardencsGitHub/tweetynet](https://urldefense.proofpoint.com/v2/url?u=https-3A__github.com_yardencsGitHub_tweetynet&d=DwMFaQ&c=clK7kQUTWtAVEOVIgvi0NU5BOUHhpN0H8p7CSfnc_gI&r=XpGqnFKnJHmx0CQn9lpv8BoZo-WGWpT8WvlwieK2nBI&m=_x3T0XI8jJIfucd4dHm9udP-sXSwCebsDxixQVz3kEM&s=7t3PYGWBT6lBsW7hlX2QB6YLoG4HTzwCyTIhTLR2Dz0&e=)  [https://www.abstractsonline.com/pp8/#!/4649/presentation/32476](https://urldefense.proofpoint.com/v2/url?u=https-3A__www.abstractsonline.com_pp8_-23-21_4649_presentation_32476&d=DwMFaQ&c=clK7kQUTWtAVEOVIgvi0NU5BOUHhpN0H8p7CSfnc_gI&r=XpGqnFKnJHmx0CQn9lpv8BoZo-WGWpT8WvlwieK2nBI&m=_x3T0XI8jJIfucd4dHm9udP-sXSwCebsDxixQVz3kEM&s=r0TPVWFK4xRTrEQ87GLOVD8CeUtsCemOBVaKYlyOHHc&e=)

Online GPU via google: <https://cloud.google.com/gpu/>

Accoustics: Audacity ([www.audacityteam.org/download/](http://www.audacityteam.org/download/)) is a great, free package for working with sound of any kind.

Additional data sets:

<https://drive.google.com/drive/folders/1etvuPjaNHV9oFPgUAuLxP3bk1aWfj3Pl> - 3 Thousand annotated zebra finch calls from Theunnisen et al.

Jupyter Notebook to walk you through some initial processing steps:

<https://canvas.upenn.edu/files/72914332/download?download_frd=1>

Ammon has some experience in this field, reach out to him with questions if you are stuck.

**Reporting Format and Instructions**

**Part 1:**

Please submit a 1-page proposal for the question and specific problems within the question that your team proposes to address by midnight **November 19, 2018**. In this document, write your question, the set of hypothesis or analysis problems you propose to answer. We will evaluate and approve your project. (we will almost certainly say yes, but we may give you some suggestions to either expand or limit your scope to make sure it’s a good project.)

**Part 2:**

Carry out the project you propose in Part 1 by accessing the data source (below), generating programs, and using public analysis packages (see below of suggested resources).

**Submit (by December 16, 2018):**

1. Report consisting of the following sections (up to 3 single space pages, figures and tables not included in page limit):
2. Aims of your study—concise 1 paragraph description of your analysis goals
3. Approaches and the rationale for the approaches
4. Results
5. Discussions
6. Citations (not included in 3 page limit)
7. All code and raw data in a single zip file.

**Final thoughts:**

Many of these problems are important biological questions and a good answer to these would be publishable in its own right. You will find that many people have approached similar questions in the past. With any of these questions, you may draw inspiration from past work, both in terms of hypotheses and methods, however be sure to appropriately cite prior work, and do your best to ensure that your project is original (i.e., don’t just validate some past method you find: use an existing method to answer a new question, or use a new the method and compare its success). Your analytical approach does not need to be perfect. You will be graded for demonstrating that you understand the power, assumptions, and limitations of the method you chose, effectively executing your approach, and understanding the implications of your results, rather than your ability to successfully prove or disprove your hypothesis. Have fun!