USC University of Southern California

TRGN 515:

Advanced Human Genomic Analysis Methods

Lecture 1 – Week 1:
Administrivia & Expectations & Basics

Bilgenur Baloglu, Ph.D.

Clinical Instructor of Translation Genomics,

USC Keck School of Medicine

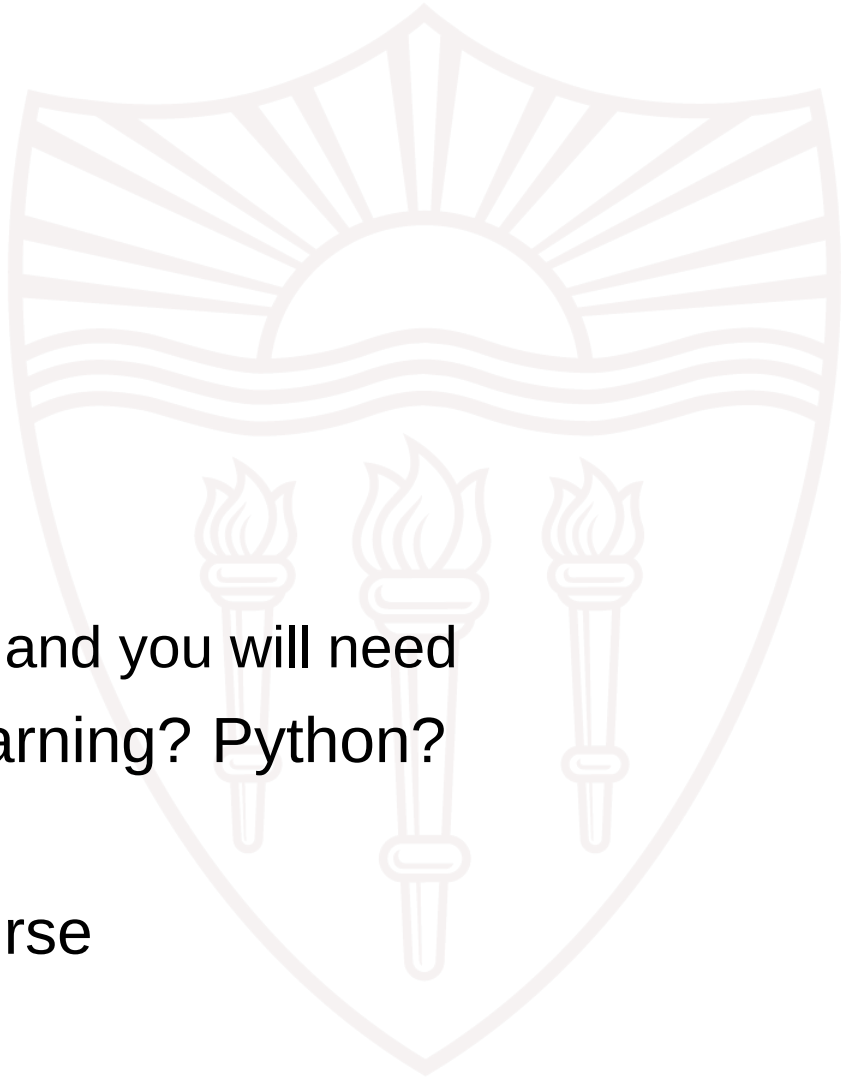
Bioinformatics scientist III, Thermo Fisher Scientific

Email: baloglu@usc.edu

USC

Outline for lecture 1

- Get to know your instructor
- Course things
  - Homework
  - Course etiquette
  - Course outline
  - Things you should know and you will need
- Bioinformatics? Machine learning? Python?
- Q&A
- Jump into Python crash course



USC

My academic path

- BSc in Molecular Biology and Genetics at Istanbul Technical University, Turkey
- Exchange year at Munich Technical University, Germany
- PhD in biological sciences at the National University of Singapore, Singapore
- Postdoc at the Centre for Biodiversity Genomics, University of Guelph, Canada

TMSI Tropical Marine Science Institute
National University of Singapore

PUB SINGAPORE'S NATIONAL WATER AGENCY

SINGAPORE INTERNATIONAL GRADUATE AWARD (SINGA)

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY

churchillnorthernstudiescentre

CANADA FIRST RESEARCH EXCELLENCE FUND

APOGÉE CANADA FONDS D'EXCELLENCE EN RECHERCHE

NUS National University of Singapore

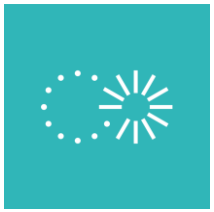
UNIVERSITY of GUELPH

CBG

USC

My industry career

- Won the All-Genetics award for industrial applications using DNA barcoding and DNA metabarcoding for environmental DNA study category
- Managed to get to the top 5% in the application process of 'Sci-Founder Fellowship' among 371 applicants for start up idea (then got rejected)
- Got accepted into Braid Theory's Celsius: Ocean Synthetic Biology Pre-Accelerator (currently debating)
- Bioinformatics lead, Sequential Skin (Feb – August 2021)
  - Developed bioinformatics pipelines to process MiSeq and ONT 16S data. Products: Python based primer designing algorithm, bash & R based MiSeq ecological analysis tool
- Bioinformatics scientist III, Thermo Fisher Scientific (August 2021 –)



ThermoFisher SCIENTIFIC

BRAID THEORY

USC

My path from academia to industry

Baloğlu *et al. Frontiers in Zoology* (2018) 15:31
https://doi.org/10.1186/s12983-018-0276-7

Frontiers in Zoology

RESEARCH Open Access

NGS barcoding reveals high resistance of a hyperdiverse chironomid (Diptera) swamp fauna against invasion from adjacent freshwater reservoirs

Bilgenur Baloğlu[1], Esther Clews[2] and Rudolf Meier[1,3*]

ROYAL SOCIETY OPEN SCIENCE

rsos.royalsocietypublishing.org

Research

Cite this article: Lim NKM, Tay YC, Srivathsan A, Tan JWT, Kwik JTB, Baloğlu B, Meier R, Yeo DCJ. 2016 Next-generation freshwater bioassessment: eDNA metabarcoding with a conserved metazoan primer reveals species-rich and reservoir-specific communities. *R. Soc. open sci.* **3**: 160635. http://dx.doi.org/10.1098/rsos.160635

Received: 25 August 2016

Next-generation freshwater bioassessment: eDNA metabarcoding with a conserved metazoan primer reveals species-rich and reservoir-specific communities

Nicholas K. M. Lim[1], Ywee Chieh Tay[1], Amrita Srivathsan[1], Jonathan W. T. Tan[1], Jeffrey T. B. Kwik[1], Bilgenur Baloğlu[1], Rudolf Meier[1,2] and Darren C. J. Yeo[1]

RESOURCE ARTICLE

WILEY MOLECULAR ECOLOGY RESOURCES

A MinION™-based pipeline for fast and cost-effective DNA barcoding

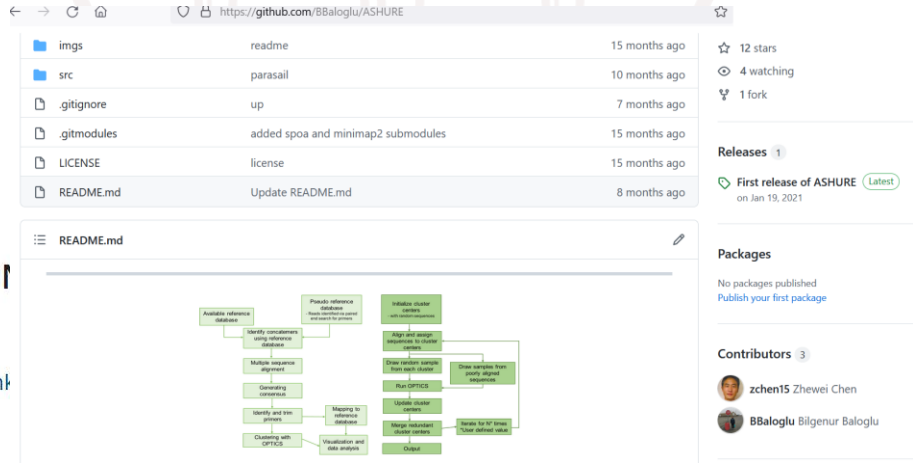
Amrita Srivathsan[1,*] | Bilgenur Baloğlu[1,*] | Wendy Wang[2] | Wei X. Tan[1] | Denis Bertrand[3] | Amanda H. Q. Ng[3] | Esther J. H. Boey[3] | Jayce J. Y. Koh[3] |

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

RESEARCH ARTICLE | Open Access

A workflow for accurate metabarcoding using nanopore MinION sequencing

Bilgenur Baloğlu, Zhewei Chen, Vasco Elbrecht, Thomas Braukmann, Shanna MacDonald, Dirk Steink



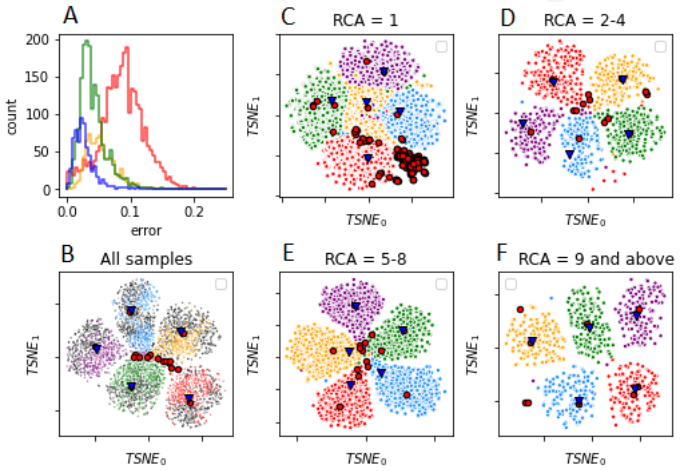
USC

My path from academia to industry

Algorithms

[1] **Baloglu, B.**, Chen, Z. Python-based bioinformatics algorithm for analyzing metabarcoding data obtained with Nanopore sequencing. Github site

[2] Srivathsan, A.*, **Baloglu, B.***, Wang, W., Tan, W.X., Bertrand, D., Ng, A.H.Q., Boey, E.J.H., Koh, J.J.Y., Nagarajan, N. and Meier, R. Python-based bioinformatics algorithm for analyzing DNA barcoding data obtained with Nanopore sequencing. Github site



USC

Course info

- Lecture (Tue/Thu)
  - 5-7 pm
- 6 Homeworks, ~20% of final grade
  - Due on Sunday night (11:59 pm) via Blackboard
  - Homework 1 is released as of this lecture!
    - Check /scratch/trgn515/Homework
    - Due this Sunday
  - 1 Midterm exam, at week 9, 20% of final grade
  - 1 Final project, due week 14, 30% of final grade
  - Final exam, week 15, 10% of final grade

Plan accordingly w/ trgn515!

Regarding homeworks

- If you have prior experience with Python
  - Should be pretty straightforward (2-3 hours)
- If you do not…
  - Might take a while (3-4 hours?)
  - But don’t worry, you will catch up as long as you follow the notebooks and run them on your own time

Late submission policy

- Up to 24 free late hours
- Specify # of late hours used when submitting

USC

Course etiquette

- Feel free to ask questions during lecture
- Adhere to the academic integrity
  - Do not copy each other's solutions

USC

Course website

- Link here (but you should create your own repositories and add me as a collaborator):
  - https://github.com/BBaloglu/TRGN515_USC
- Link to my personal website
  - https://bbaloglu.github.io/
- Up-to-date office hours, zoom links, additional reading, etc.:
  - Check Blackboard

Course outline

- Python crash course and Python libraries for data visualization
  - 3 weeks
- Introduction to Bioinformatics with Biopython library
  - 2 weeks
- Introduction to machine learning
  - 3 weeks
- Building machine learning models
  - 4-5 weeks
- Final project: Throughout the course with presentation in week 14

What should you know

- Pandas, numpy, Scikit
- Python
- Basic statistics
- Basic knowledge of biology and DNA sequencing
- Basic competency with UNIX/Linux (can use a terminal)
- Familiarity with some ML terminology
  - Don’t worry. You will get there!

What do you need

- Python 3.7
- Jupyter notebooks
- If you want to access Jupyter on your own PC, then you will also need Anaconda3 installation (version 4.5.12)
- GitHub account: This is where you will upload your homeworks and projects, where you will add me as a collaborator in order to share them with me
- CRITICAL THINKING!

How you will operate

- Make your own directories and subdirectories on the TRGN server in /scratch directory
- Copy the lecture and homework notebooks in your own directory
- Work on homework in your own directory (not in the shared directory)
- Copy or transfer the final homework in your Github repository (ideally keep the same directory system), which should notify me if I am added as a collaborator

What is bioinformatics?

- Dealing with biological challenges using computer science tools?
- Biggest part of bioinformatics: Dealing with DNA sequencing data
  - DNA sequencing is the process of reading biological material and translating it into a computer readable data representation.
  - The sequencing process is complex and introduces many challenges such as gaps between reads, lack of coverage and various other sequencing errors

What is machine learning?

"Machine learning is fitting a function to examples and using that function to generalize and make predictions about new examples."

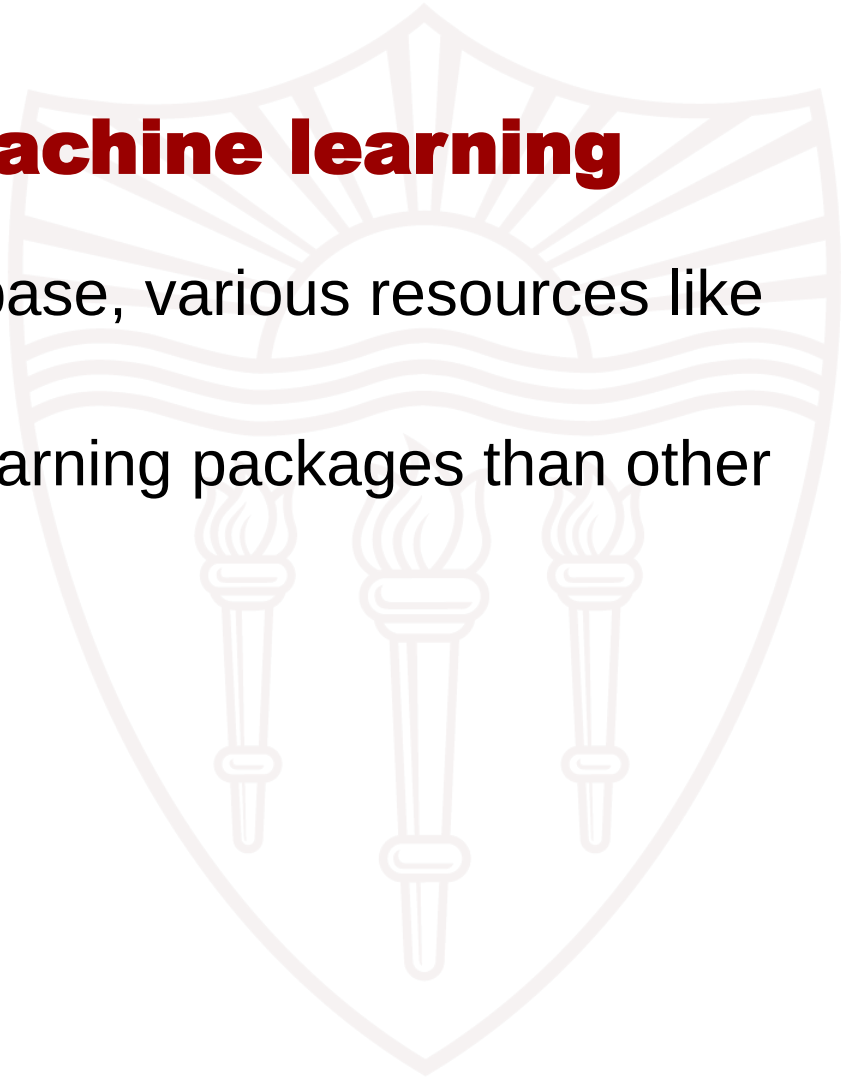
Derek Jedamski, GitHub

What is machine learning?

- Make a machine (computer) learn a model (hypothesis) with enough data of a given type, so it becomes able to identify one or more patterns within it.
- Identified (learned) patterns can then be used for making estimates (predictions) on unseen data of similar type as the data which was used to learn the pattern.
- The amount of required data may vary based on the difficulty of the pattern to learn.
- The learning process is often referred to as training, while the process of making decisions is called classification

Why use Python for Machine learning

- Popular and has large user base, various resources like stack overflow
- Python has more machine learning packages than other languages
- Easy to learn, easy to use



USC

Should we even use machine learning?

- Is this a type of problem that can be solved using machine learning?
  - Does this problem require a prediction or some type of bucketing into categories?
- Do you have all the components needed to build a model?
  - Do I have data with labels?
  - Do I have the ability to assess the quality of the model?
  - Do I know what an acceptable accuracy threshold looks like?

Common challenges with machine learning

- Problem scoping
  - Dealing with the wrong problem
  - Tolerance threshold (i.e., accuracy %) not determined
- Data
  - Lack of data
  - Too much data
  - Lack of labels in the data
  - Data is noisy, dirty etc.
- Infrastructure
  - Lack skills to automate
  - Not enough compute power
  - Inability to test quality of the model
- Latency
  - Model takes too long to train
  - Model takes too long at inference time

Exploratory data analysis

Why?

- Understand the shape of the data
- Learn which features of might be useful
- Inform the cleaning that will come next

What?

- Counts or distribution of all variables
- Data types of each feature
- Missing data
- Correlations
- Duplicates

Data cleaning

Why?

- No data out there are served clean
- Shape data so model can pick the best signal
- Remove irrelevant data
- Adjust features to be acceptable for a model

What?

- Encode categorical variables
- Fill missing data
- Scale data to account for outliers