Course introduction

Population genomic inference from low-coverage whole genome sequencing data

Course goals

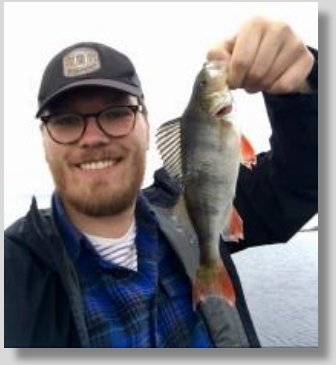
- Understand the power and challenges associated with using low-coverage whole genome sequencing data for population genomic analysis
- Become familiar with all steps involved from sample to inference
- Develop an intuition for the statistical framework implemented in ANGSD and associated programs
- Gain experience with building a bioinformatic pipeline to process low-coverage sequencing data to perform different types of population genomic analyses

Who we are

Nina Overgaard Therkildsen
Cornell University

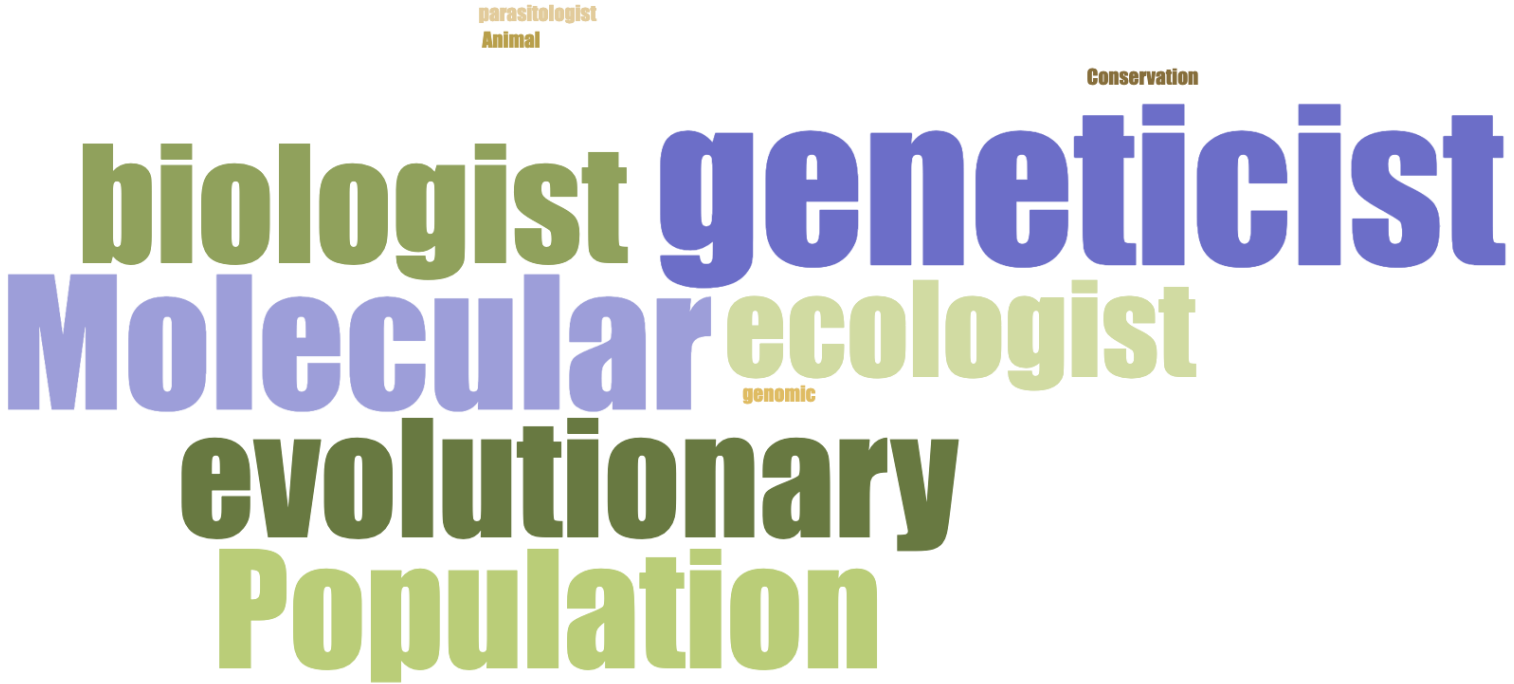
Tyler Linderoth
Michigan State University

Nicolas Lou
UC Berkeley



Arne Jacobs
University of Glasgow

Who you are



https://www.jasondavies.com/wordcloud/

Who you are

https://www.jasondavies.com/wordcloud/

Where are you?

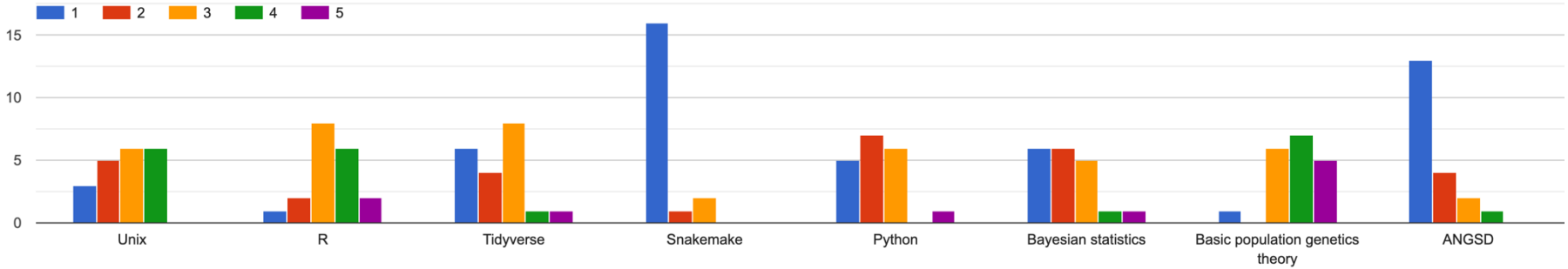
- Please indicate your location on this map https://www.google.com/maps/d/edit?mid=1uwCs3a6EwGY4LP_RnqFkLvlJApN3dhI&usp=sharing

Round of introductions

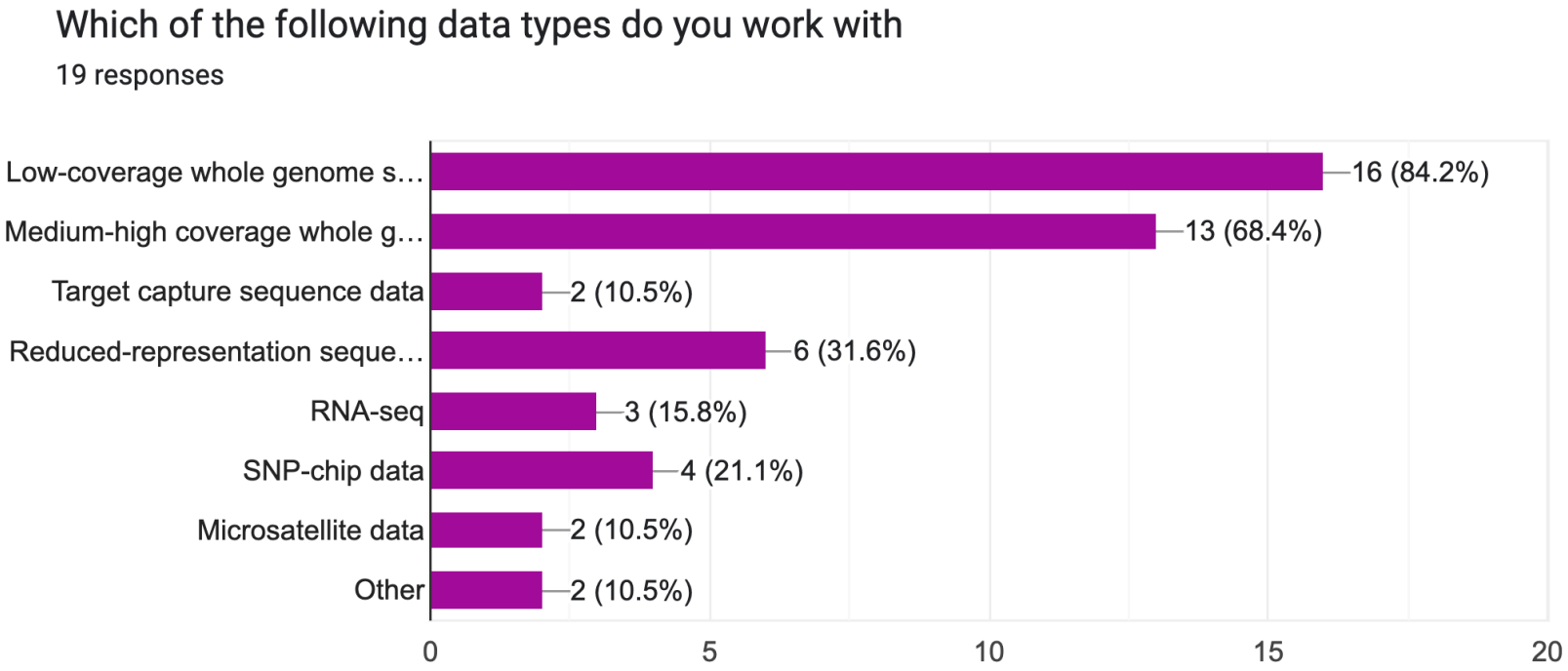
- Please share briefly
  - Your name and current affiliations
  - What kind of organism(s) you work on
  - Why you're interested in this course

From the pre-course survey

Rate your familiarity with the following on a scale of 1 (beginner) to 5 (wizard):



From the pre-course survey



Graphic: https://allarewelcomehere.us/

Code of Conduct

- We are dedicated to providing a **welcoming and supportive environment for everyone**, regardless of background, identity and prior experience. Everyone in this course will be coming from a different place with different experiences and expectations. We will not tolerate any form of language or behavior used to exclude, intimidate, or cause discomfort. This applies to all course participants (instructor, students, guests). In order to foster a positive and professional learning environment, we encourage the following kinds of behaviors

Behaviors we encourage

- Use welcoming and inclusive language
- Be respectful of different viewpoints and experiences
- Show courtesy and respect towards others
- Help each other - you may well learn something or reinforce your own skills in the process

Approximate daily schedule

Berlin time	US eastern	Activity
14 – 15.15	8 - 9.15	Session 1
		BREAK
15.30 – 16.45	9.30 -10.45	Session 2
		BREAK
17.15 – 18.30	11.15 - 12.30	Session 3
		BREAK
18.45 - 20	12.45 - 2	Session 4

Course schedule

- Day 1
  - Welcome!
  - Introduction to low-coverage whole genome sequencing
  - From sample to fastq
  - From fastq to bam

- Day 2
  - Recap on exercises from day1
  - Genotype likelihoods
  - SNP calling
  - Allele calling

Course schedule

- Day 3
  - Linkage disequilibrium
  - Population structure (PCA and admixture analysis)

- Day 4
  - The site frequency spectrum (1d and 2d)
  - Fst and diversity statistics
  - Automation with snakemake
  - Overview of other applications and future perspective

Daily practicals

- Will be available in a GitHub repo that you can keep accessing after the course
- Will work in breakout rooms
- Options to choose the type of breakout room you prefer
  - Quiet room (everyone works independently)
  - Semi-quiet room (people mostly work independently, but can ask each other questions)
  - Collaborative room (you work through the exercises together)
- Ask questions!

Asking questions

- Please ask! If you're unsure about something, others may be as well and you might help them by asking

- Zoom chat is disabled – Make sure you have joined the Slack workspace
  - Slack works better because
    - Easier to follow up with answers directly under the question and continue back-and-forth in a thread where needed
    - We'll continue to have the record after the Zoom call ends

- Please raise your hand on Zoom or use Slack for asking questions

Make sure you have access to the server

- If you haven't logged on already, please do that during one of the breaks today so we can help or troubleshoot if there are issues

What we will not have time to cover

- Genotype – phenotype association
- Imputation
- Methods specific to ancient DNA
- Structural variant detection
- Assembly of reference sequence

Course goals

- Understand the power and challenges associated with using low-coverage whole genome sequencing data for population genomic analysis
- Become familiar with all steps involved from sample to inference
- Develop an intuition for the statistical framework implemented in ANGSD and associated programs
- Gain experience with building a bioinformatic pipeline to process low-coverage sequencing data to perform different types of population genomic analyses