

Fundamentals of population genomics

Reid Brennan

rbrennan@geomar.de

Goals of course

1. Basics of population and quantitative genetic theory
2. Gain understanding of the technologies and data underlying genomics
3. Experience using the command line and R to analyze data
4. Ability to analyze and interpret population genetics data

Course format

- Lectures introducing fundamentals
 - Mix of traditional and short lectures
- Hands on tutorial
 - → analyzing population genomic dataset together
- Independent project
 - In groups, analyze population genomic data and present results

Independent project

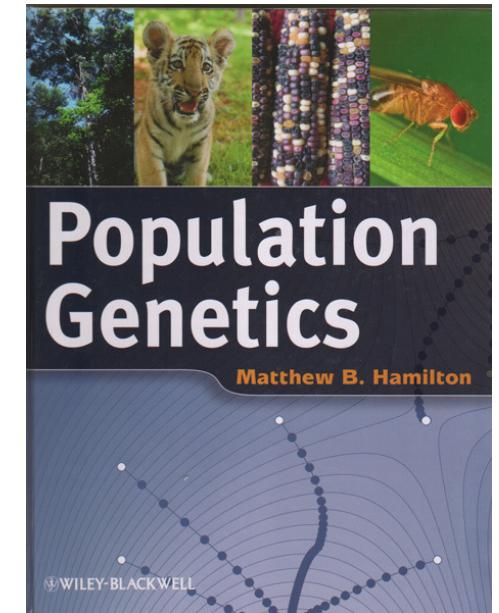
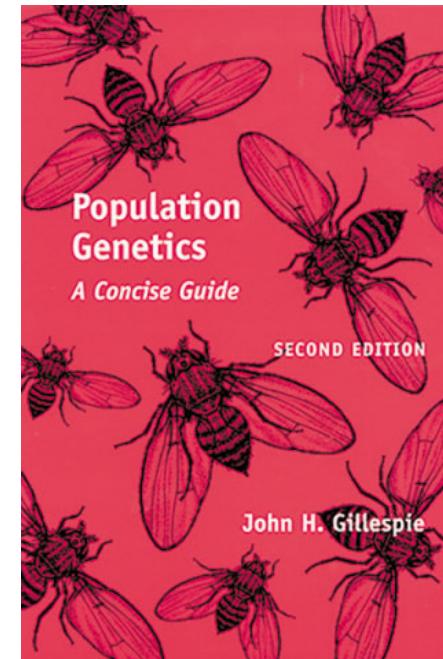
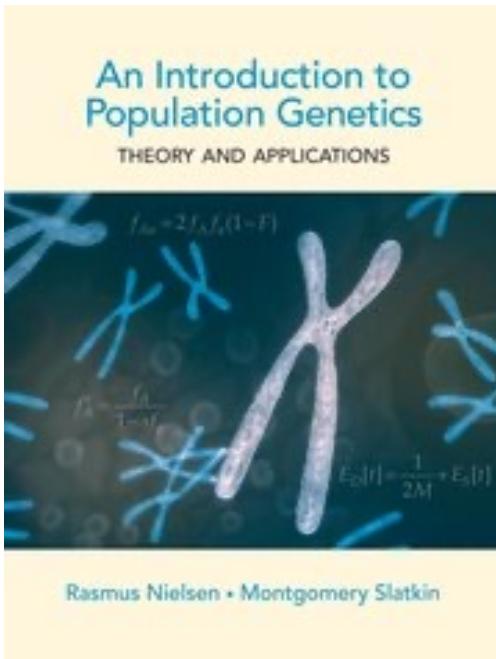
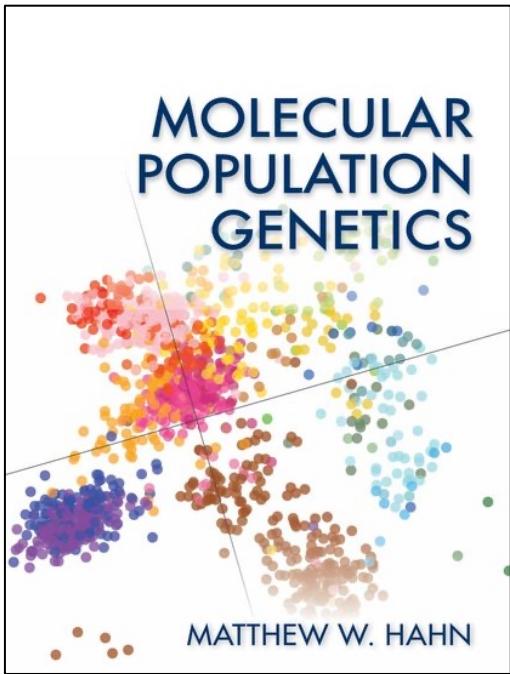
- In groups of...2? 3?
- From VCF, analyze data to make population genetic inferences
- Presentation of ~20 minutes
- Grading:
 - Use of tools covered in the course
 - Appropriate conclusions from your results
 - Overall quality of analysis, interpretation, and presentation

Schedule

- Day 1-5: Tutorial
- Day 6-9: Project
- Day 10: Presentations
- For the most up to date information, slides, tutorials, see the course website:
 - https://rsbrennan.github.io/EvolutionaryGenomics_2023/

Resources

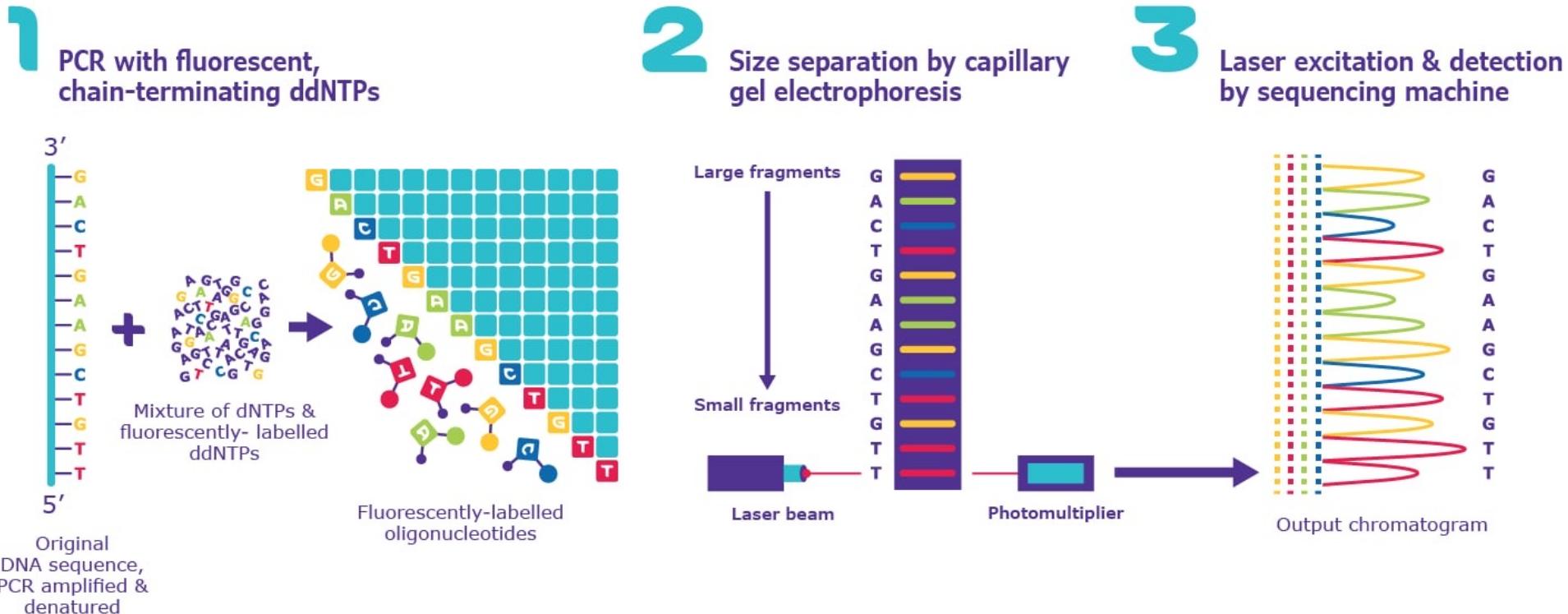
- Graham Coop's Population and Quantitative Genetics notes
 - <https://github.com/cooplabs/popgen-notes/releases>



And many others....

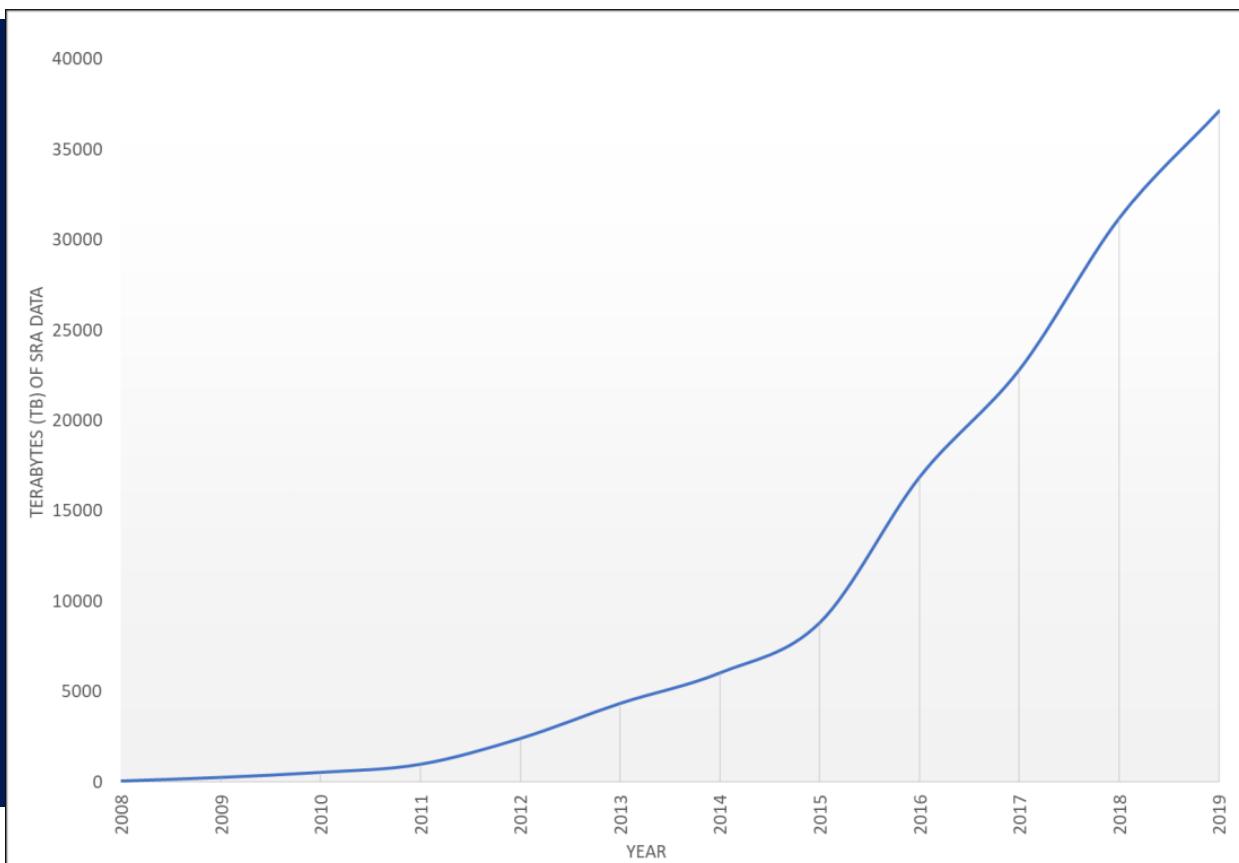
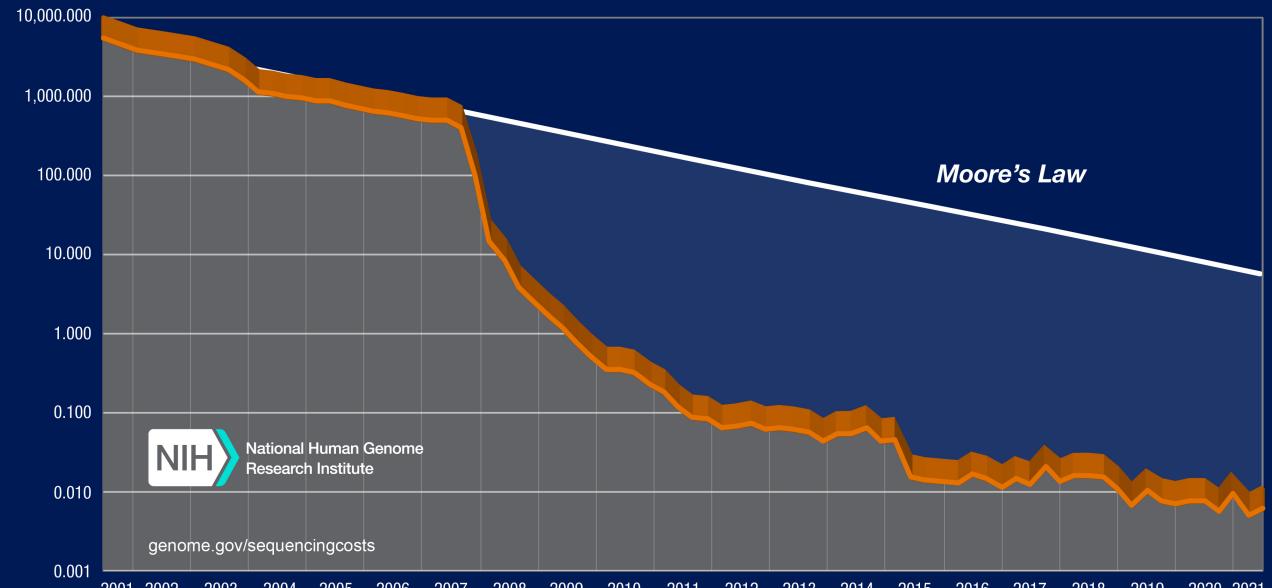
What is genomics anyway?

Sanger sequencing



- A few hundred to ~1000 bp
- Low throughput
- Accurate

Cost per Raw Megabase of DNA Sequence



Data in NCBI

Massively parallel sequencing

- Aka
 - Next generation sequencing
 - Second generation sequencing
- Basically just Illumina currently
- Short reads
 - 50-150 bp
- Very high throughput

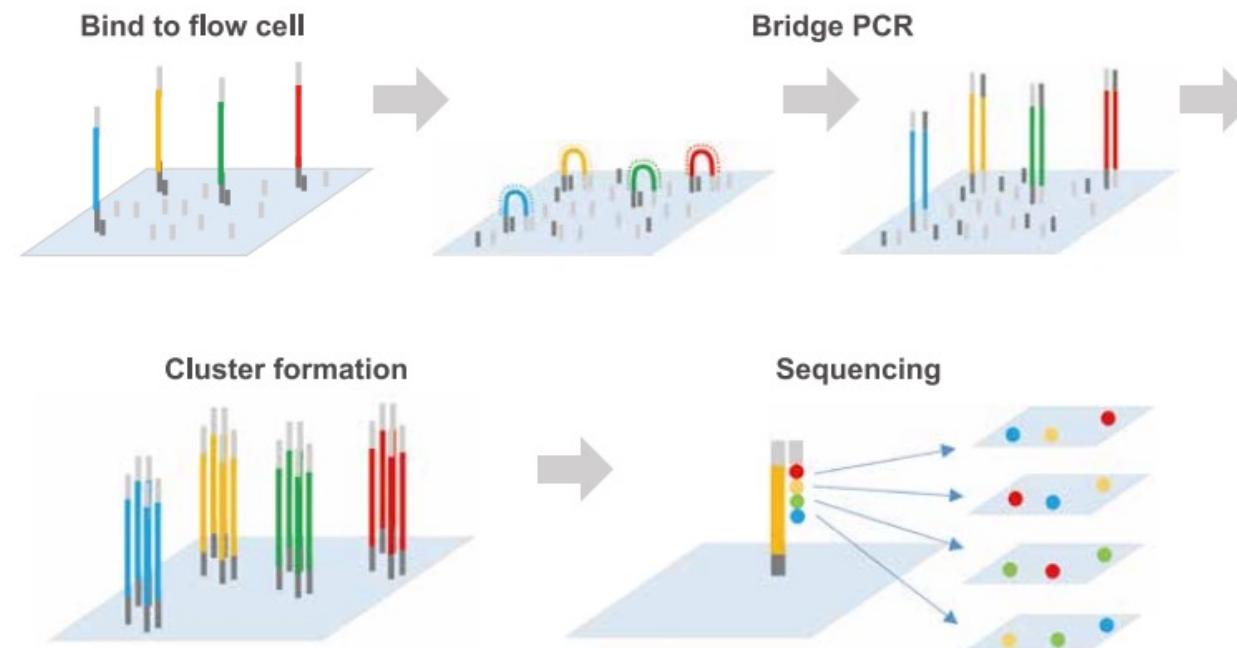
Fragment your DNA



Add adapters and amplify

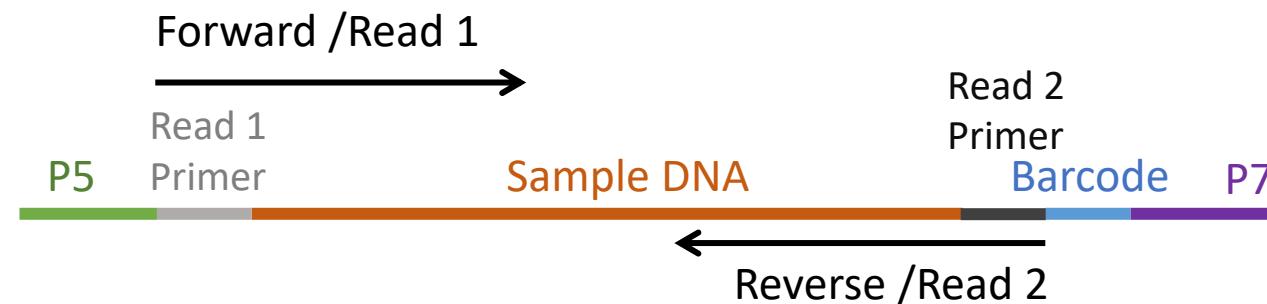


Sequence



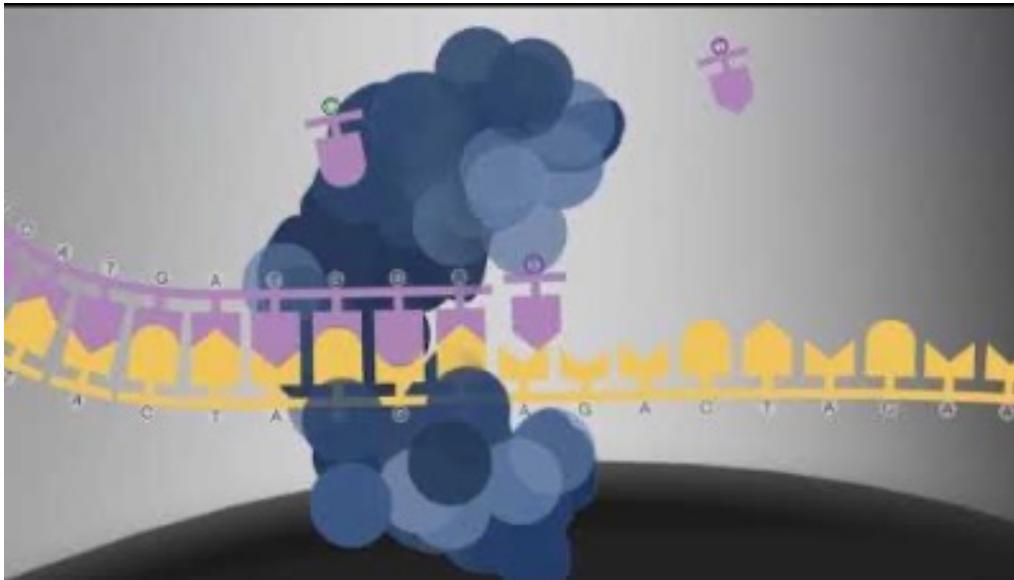
Some important concepts to understand

- Library
 - The prepared sequences ready for sequencing that represent your sample
 - We modify the DNA to make a sequencing library
- Read
 - One sequence from a library



Long read technologies

- PacBio
 - 10-50kb read length
 - Medium output
 - Read length limited by the longevity of the polymerase.
- Oxford Nanopore
 - 100kb or more
 - Low output
 - Read length limited by the length of the input DNA and decreasing yield with increasing DNA length



https://www.youtube.com/watch?v=_ID8JyAbwEo



<https://www.youtube.com/watch?v=E9-Rm5AoZGw>

Illumina

150-300bp read length
High output

Population genetics,
SNPs, RNAseq

PacBio

10-25kb read length
Medium output

Structural variation,
genome assembly

Nanopore

100kb or more
Low output

Structural variation,
genome assembly,
in field analysis

What would we use each technology for?

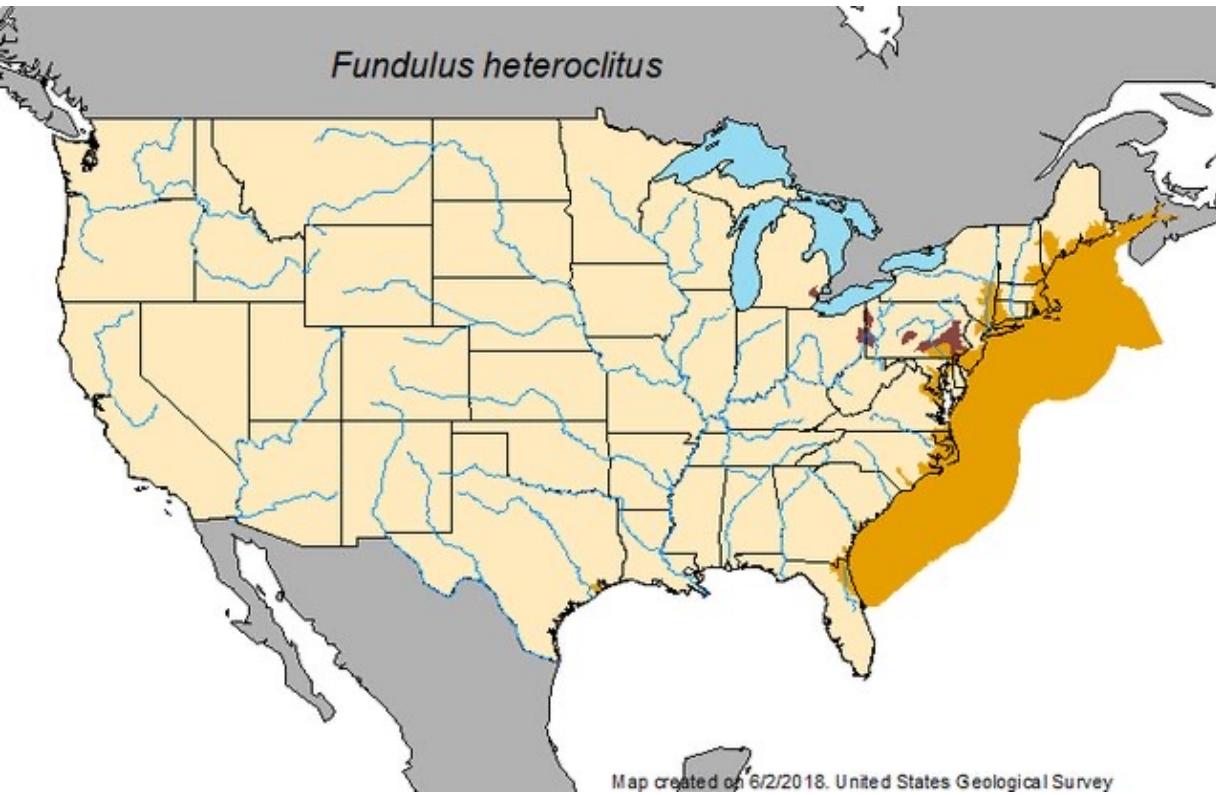
To study population genetics, we need to know allele frequencies → genomics!

Study system for the
practical

Fundulus heteroclitus study system



Fundulus heteroclitus study system



Alcaraz-Hernandez and Garcia-Berthou

- Large population sizes
- Low movement
- Huge range of environments
 - Temperature
 - Salinity
 - Hypoxia
- Given the high population sizes and low migration:
 - what might we predict about populations of killifish from different regions of their distribution?

Received: 3 May 2018 | Accepted: 6 June 2018

DOI: 10.1111/gcb.14386

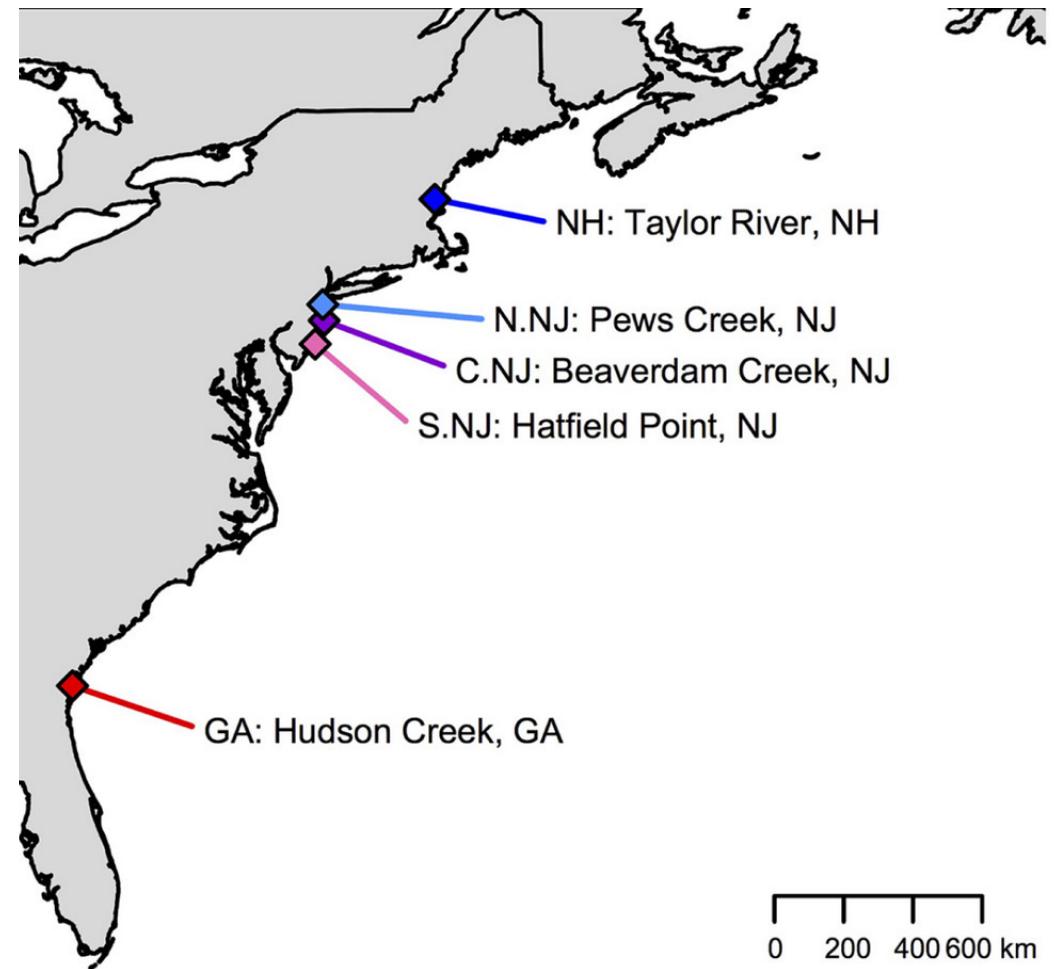
PRIMARY RESEARCH ARTICLE

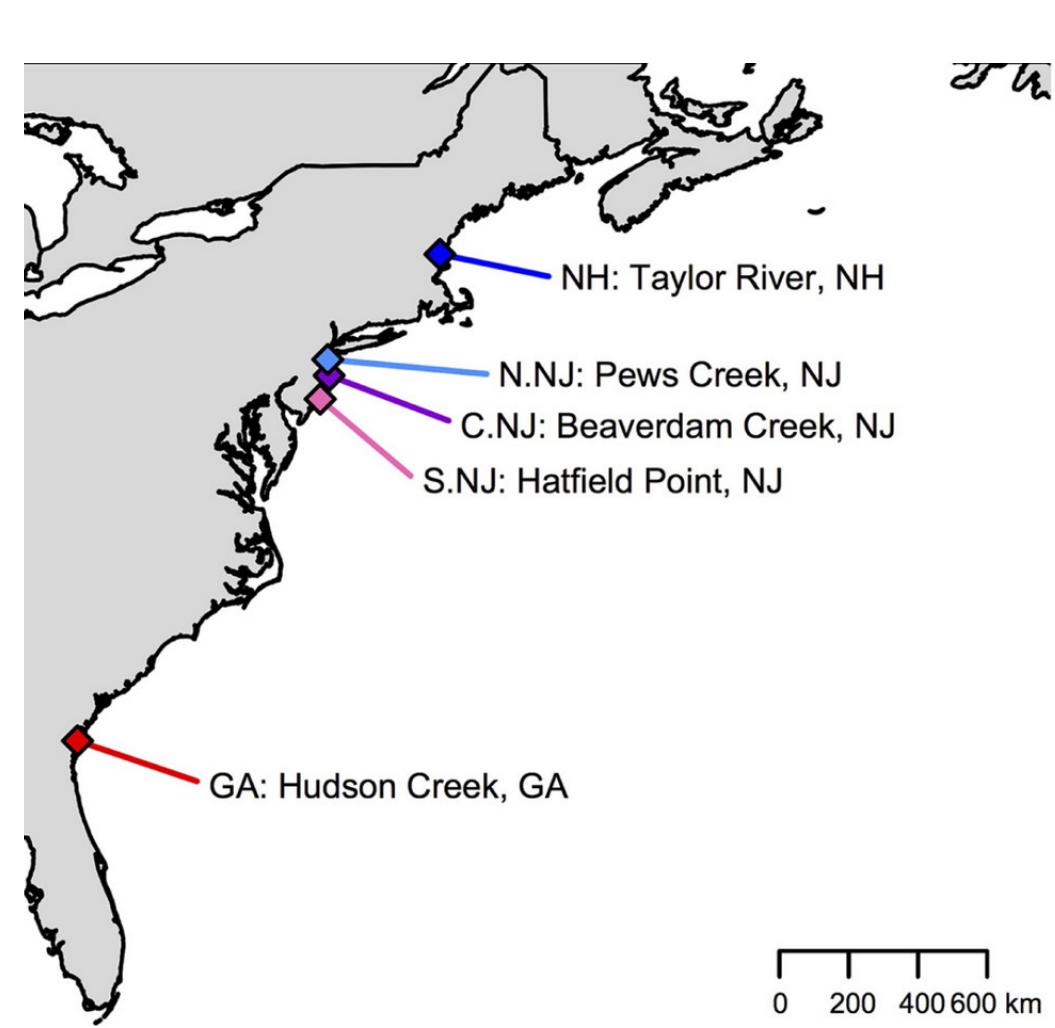
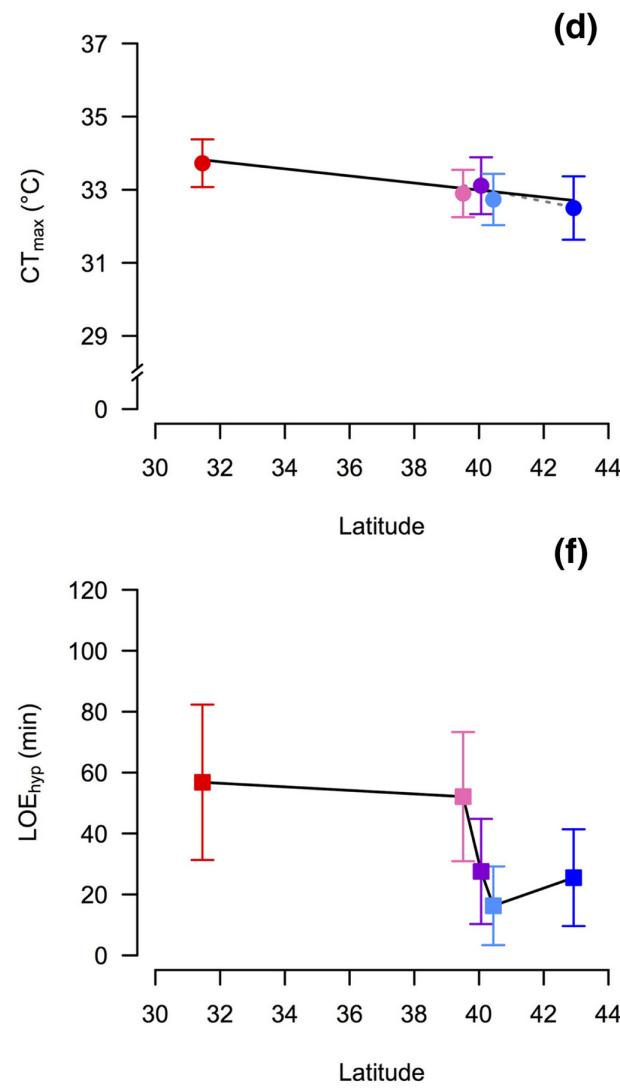
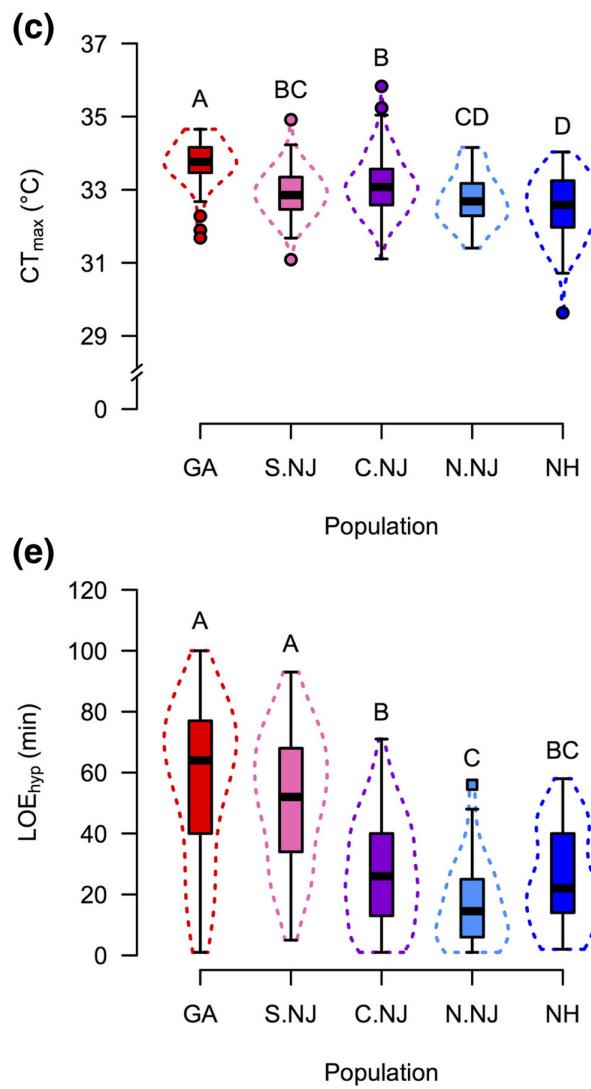
WILEY  Global Change Biology

Tolerance traits related to climate change resilience are independent and polygenic

Timothy M. Healy¹  | Reid S. Brennan²  | Andrew Whitehead²  |

Patricia M. Schulte¹ 

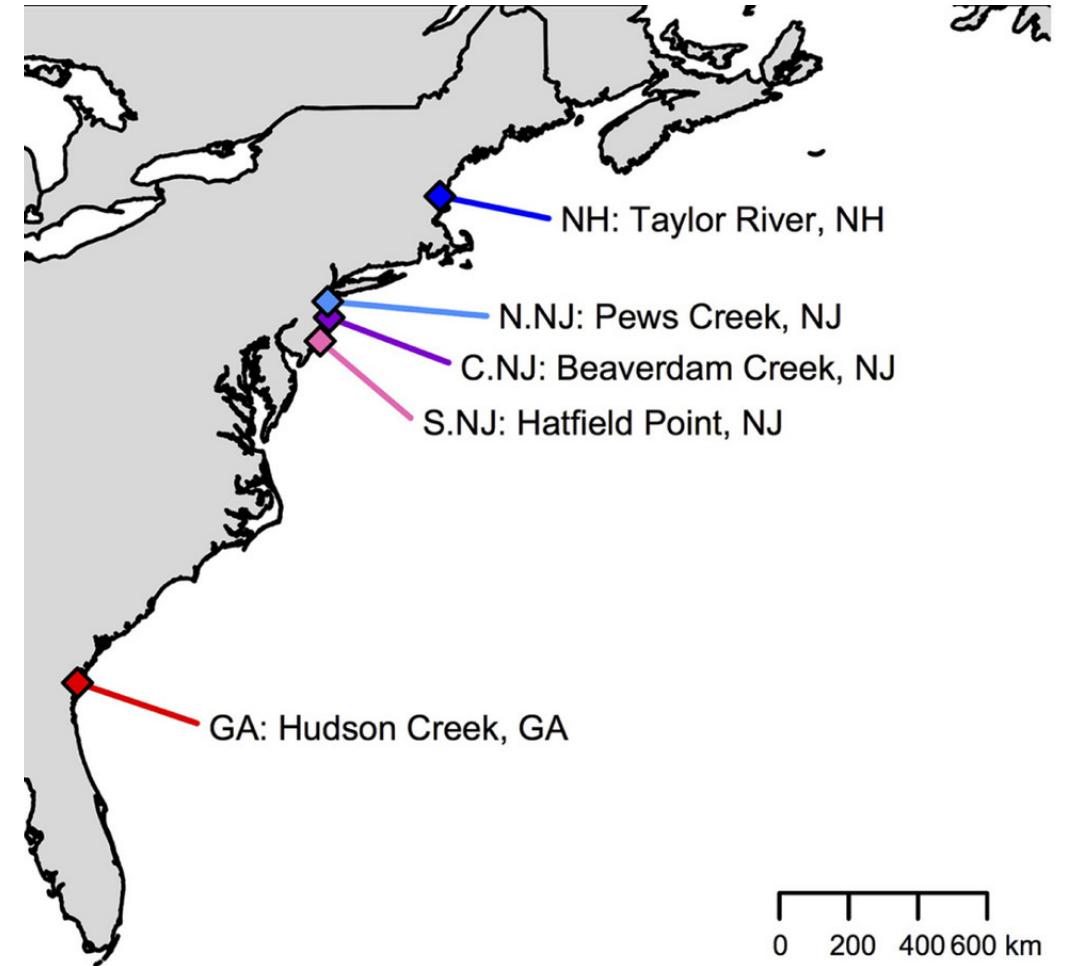




Dataset

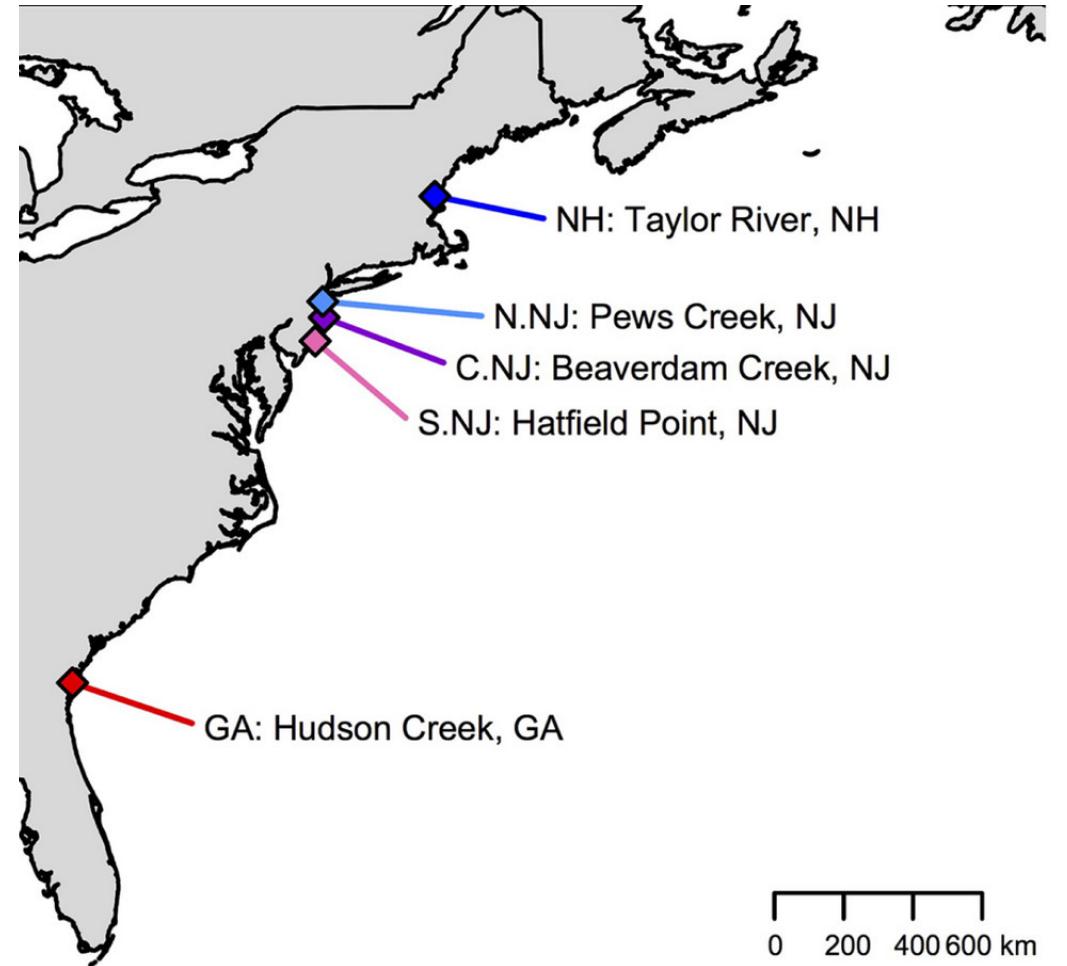
Population Location	Population ID	N
GA	GA	25
S.NJ	HP	46
C.NJ	BC	50
N.NJ	PC	44
NH	TR	45

RAD-seq



Dataset

- Population structure
- Genetic diversity
- Selection



Reduced representation

- In an ideal world...
 - Sequence full genomes of all individuals from many populations
 - Expensive, sometimes unnecessary, limited genomic resources
- RADseq, GBS, sequence capture, ddRAD, 2b-RAD...
- Useful for population structure and similar
- But... misses lots of the genome
- Cut anywhere you see the restriction site:
 - CCTGCAGG

Methods

Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers

Michael R. Miller,¹ Joseph P. Dunham,² Angel Amores,³ William A. Cresko,² and Eric A. Johnson^{1,4}

¹Institute for Molecular Biology, University of Oregon, Eugene, Oregon 97403, USA; ²Center for Ecology & Evolutionary Biology, University of Oregon, Eugene, Oregon 97403, USA; ³Institute of Neuroscience, University of Oregon, Eugene, Oregon 97403, USA

2007

Reduced representation

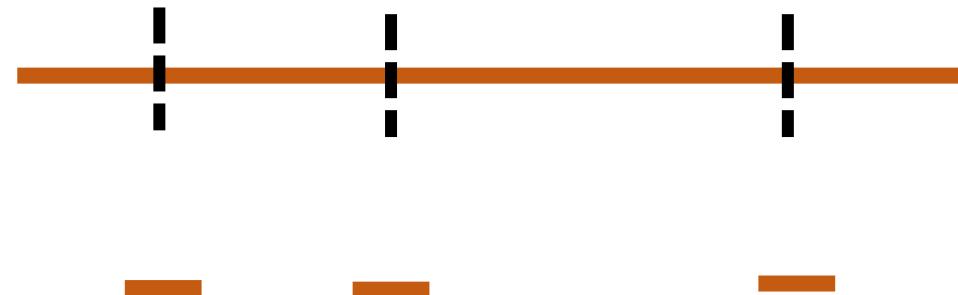
- In an ideal world...
 - Sequence full genomes of all individuals from many populations
 - Expensive, sometimes unnecessary, limited genomic resources
- RADseq, GBS, sequence capture, ddRAD, 2b-RAD...
- Useful for population structure and similar
- But... misses lots of the genome
- Cut anywhere you see the restriction site:
 - CCTGCAGG

5' CCTGCAGG 3'
3' GGACGTCC 3'

Ligate
adapters
for
sequencing

5' CCTGCA|GG 3'
3' GG|ACGTCC 3'

GG.....
ACGTCC.....



How much of the genome does RADseq cover?

- How big is the genome?
 - 1.5 billion bases
- How often do we find a cut site?
 - Assume bases are randomly distributed
 - 0.25^8
- $1.5e+09 * (0.25^8) = 22,888$

Unix, cloudb, and R

We have lots of data

- Normal to get back 100's of GB of data.
 - You can't have this on your computer!
- Computer clusters are needed
 - These require command line, unix, etc.

Why unix/command line?

- Most software is written for unix
- Computing clusters use unix
- Good at working with large data

Why R?

- Very popular in biology (good community)
- Good for statistics
- Good for plotting

Why Both?

- Reproducible!!

How to learn?

- By trying
- Workshops/courses
- Google
 - Stackoverflow; seqanswers; others
- ChatGPT... but be careful

“the cloud”

- Aka, a computer somewhere else



- We will use Cloudlab → hosted by CAU

Cloudlab

Today's tutorial

- Unix for biologists