Fundamentals of population genomics

Reid Brennan & Jenny Nascimento-Schulze

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
  - Mostly short lectures
- Hands on tutorial
  - → analyzing population genomic dataset together
- Independent project
  - In groups, analyze population genomic data and present results via poster

Independent project

- In groups of 2-3
- From VCF, analyze data to make population genetic inferences
- Poster on July 4
- Grading:
  - Use of tools covered in the course
  - Appropriate conclusions from your results
  - Overall quality of analysis, interpretation, and presentation
- You must hand in:
  - 1. Poster
  - 2. mark down/html document of your analysis.

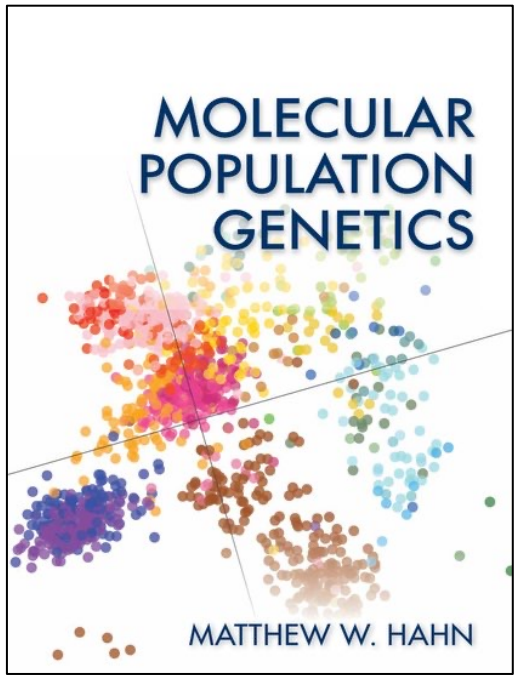
Schedule

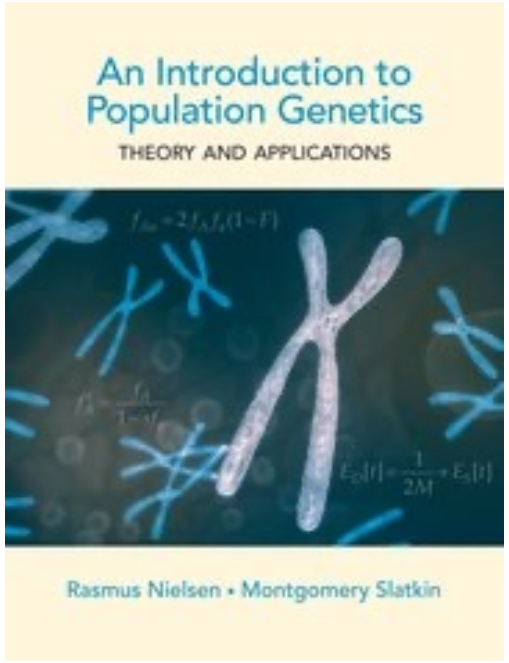
- Day 1-5: Tutorial
- Day 6-10: Project

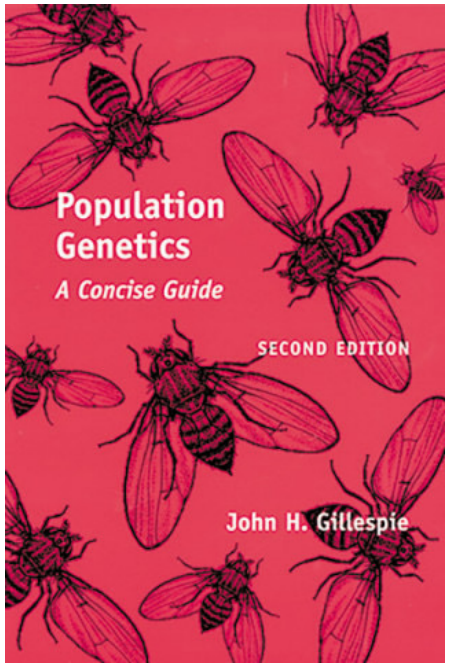
- For the most up to date information, slides, tutorials, see the course website:
  - https://rsbrennan.github.io/EvolutionaryGenomics_2024/

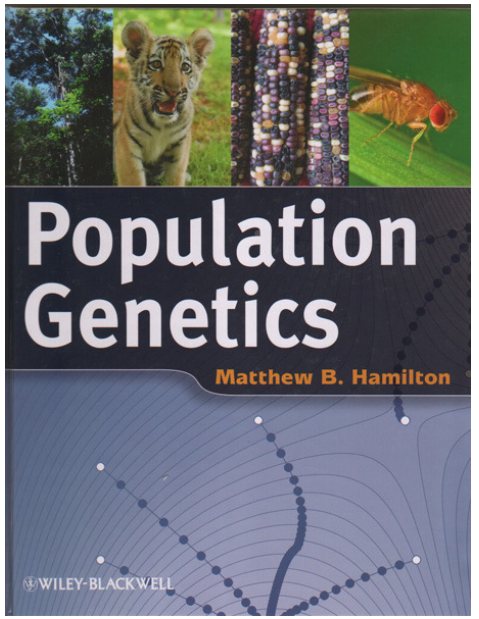
Resources

- Graham Coop's Population and Quantative Genetics notes
  - https://github.com/cooplab/popgen-notes/releases





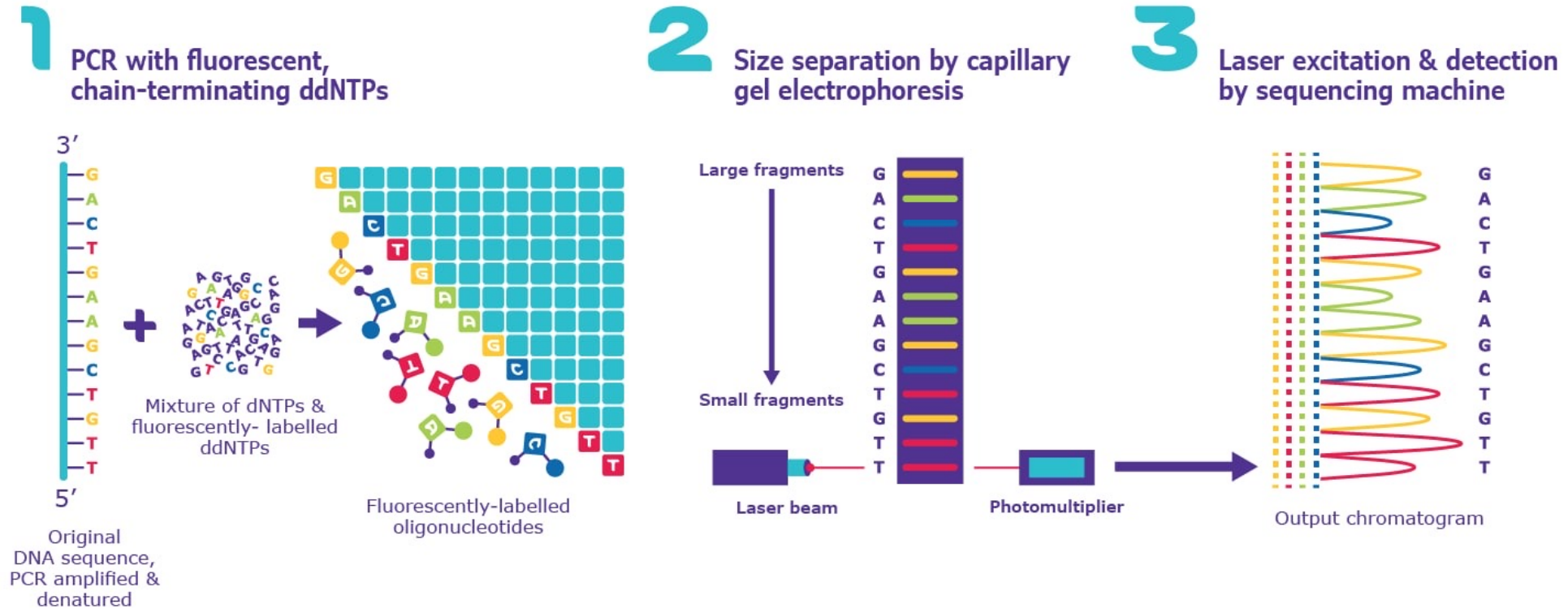




And many others….

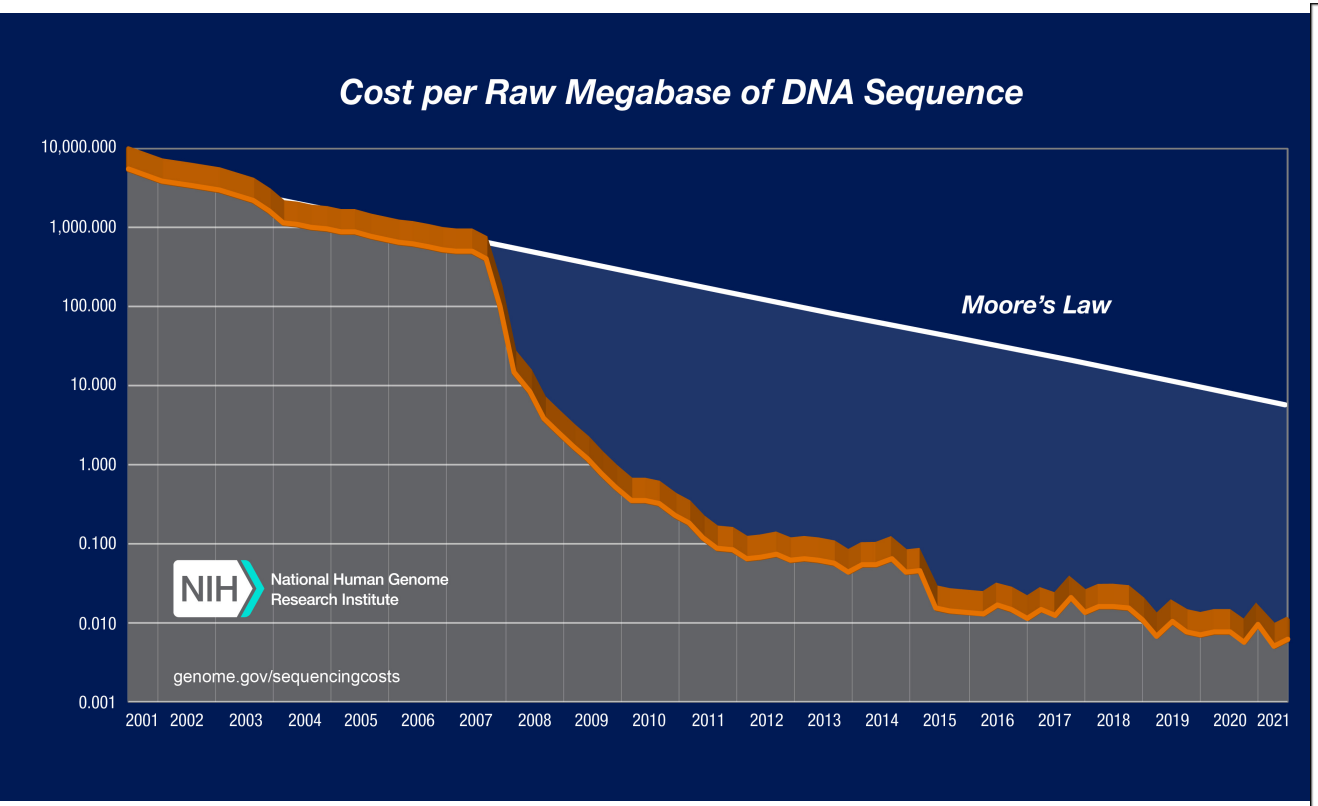
What is genomics anyway?

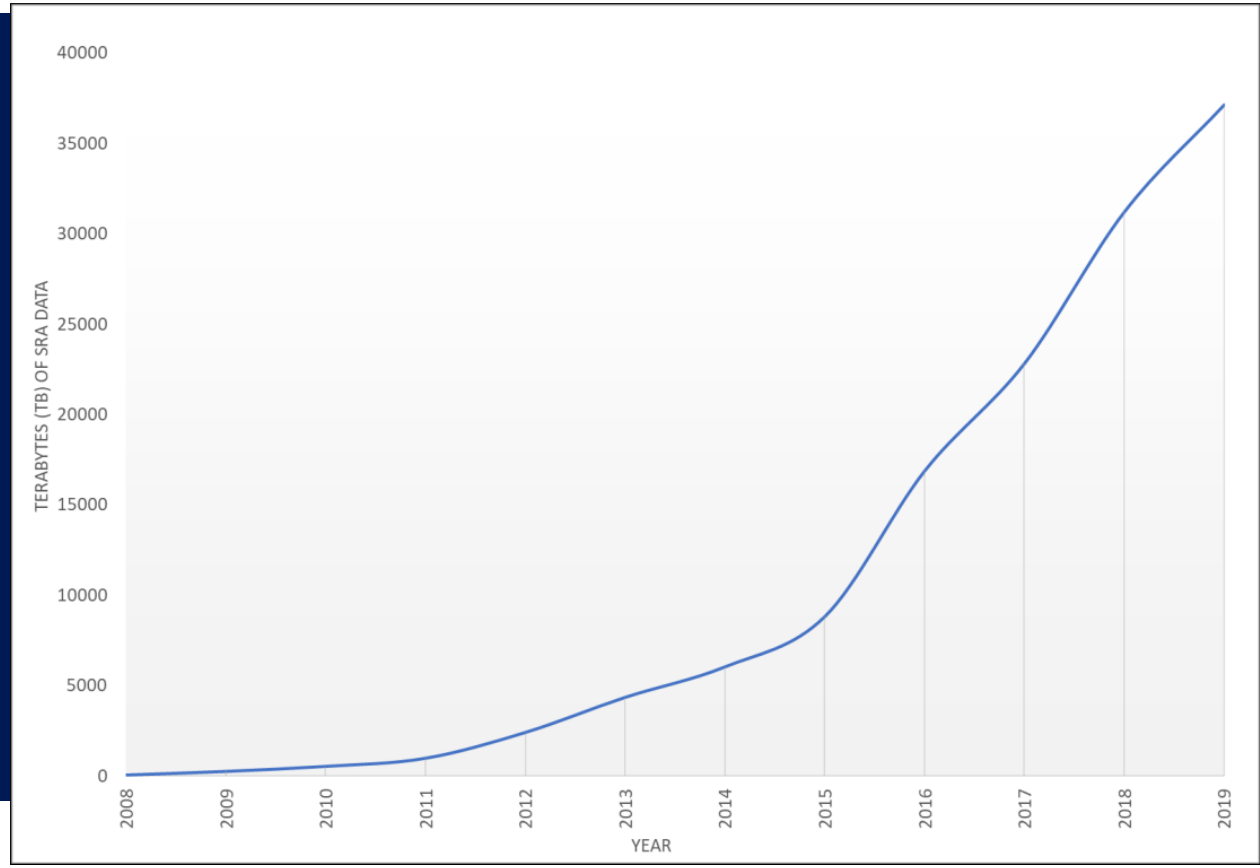
Sanger sequencing



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

https://www.sigmaaldrich.com/DE/de/technical-documents/protocol/genomics/sequencing/sanger-sequencing


Cost per Raw Megabase of DNA Sequence
10,000.000
1,000.000
100.000
10.000
1.000
0.100
0.010
0.001
2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts

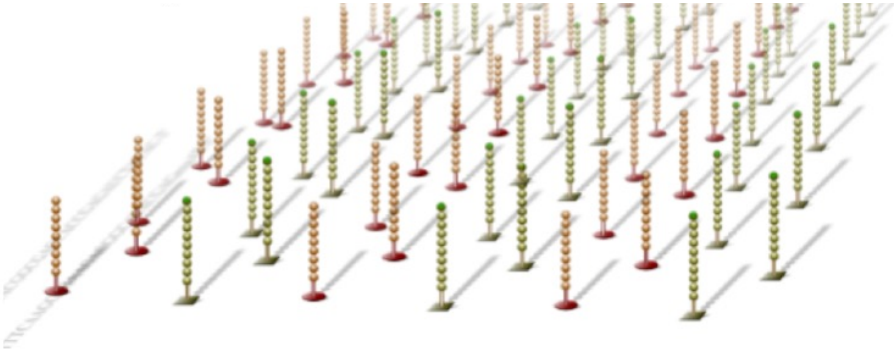

40000
35000
30000
25000
20000
15000
10000
5000
0
TERABYTES (TB) OF SRA DATA
2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019
YEAR

Data in NCBI

Massively parallel sequencing

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

Illumina flow cell

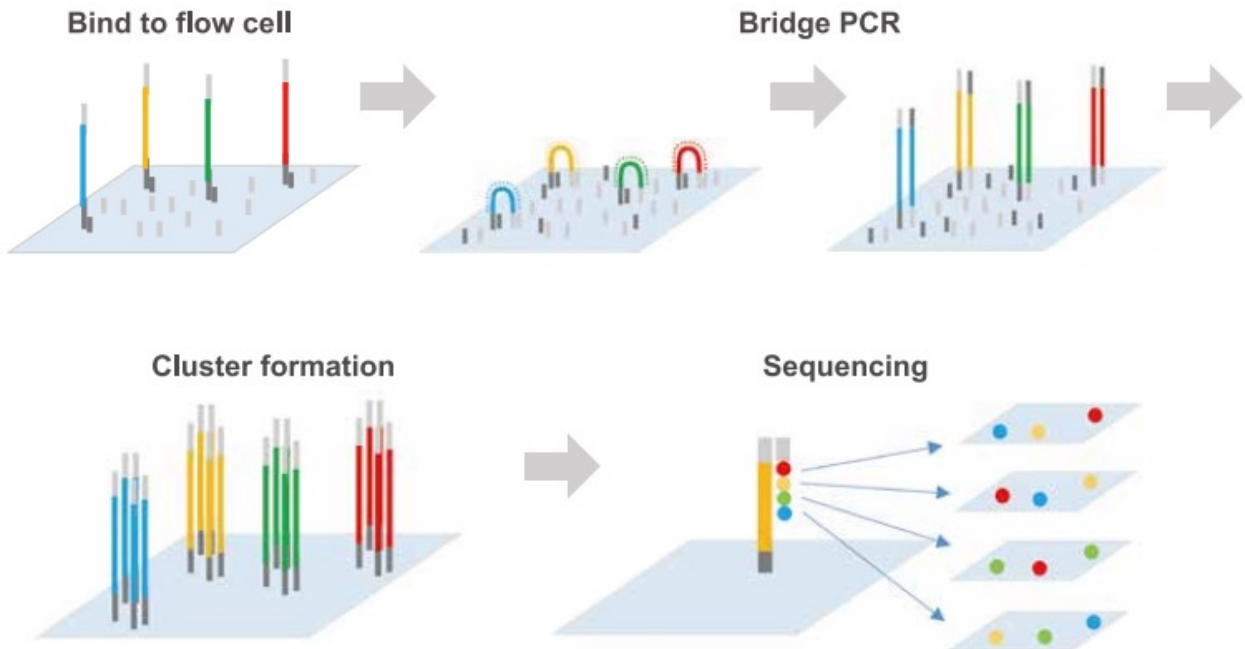


http://en.biomarker.com.cn/platforms/illumina

Fragment your DNA

Add adapters and amplify

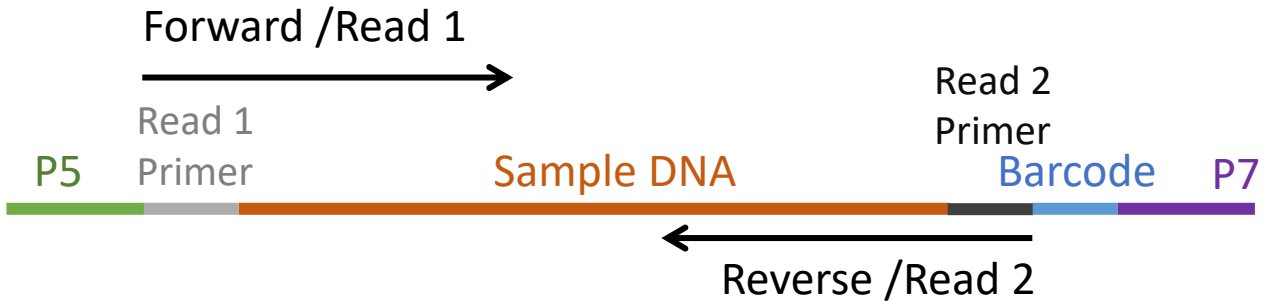
Sequencing by synthesis



https://www.genscript.com/advancing-genomics-medicine-and-health-together-by-semiconductor-dna-synthesis-technology-summary.html

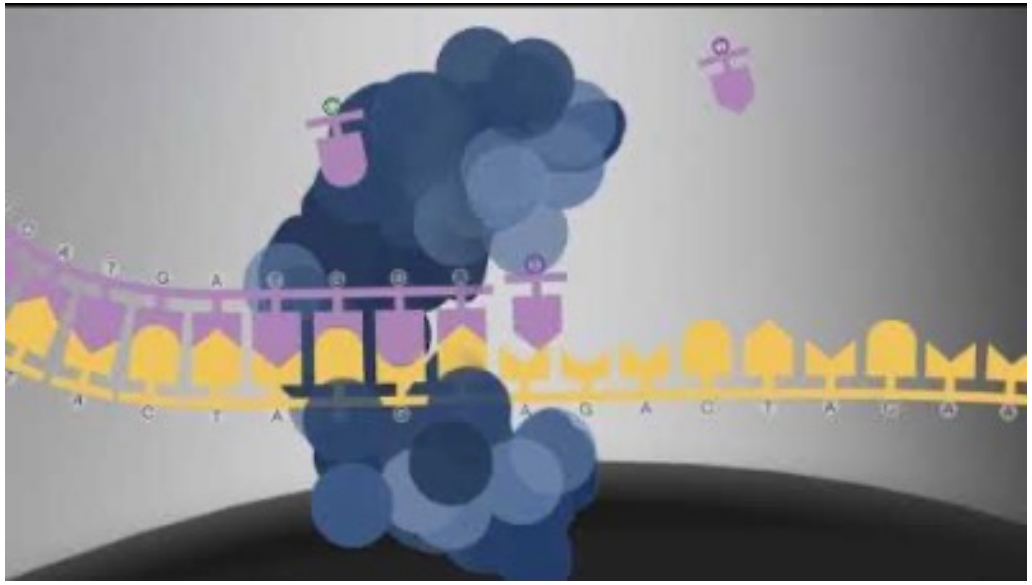
Some important concepts to understand

- 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=_lD8JyAbwEo

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

Illumina	PacBio	Nanopore
150-300bp read length	10-25kb read length	100kb or more
High output	Medium output	Low output
Population genetics, SNPS, RNAseq	Structural variation, genome assembly	Structural variation, genome assembly, in field analysis

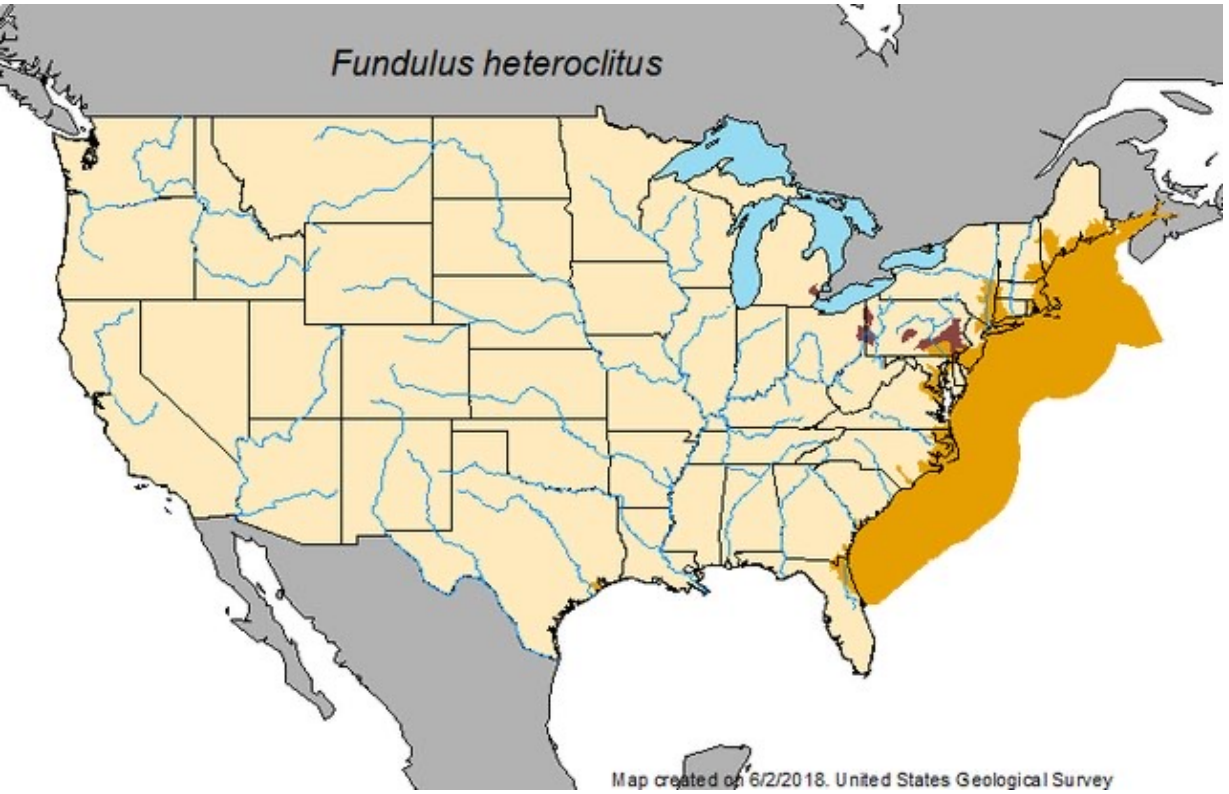
What would we use each technology for?

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

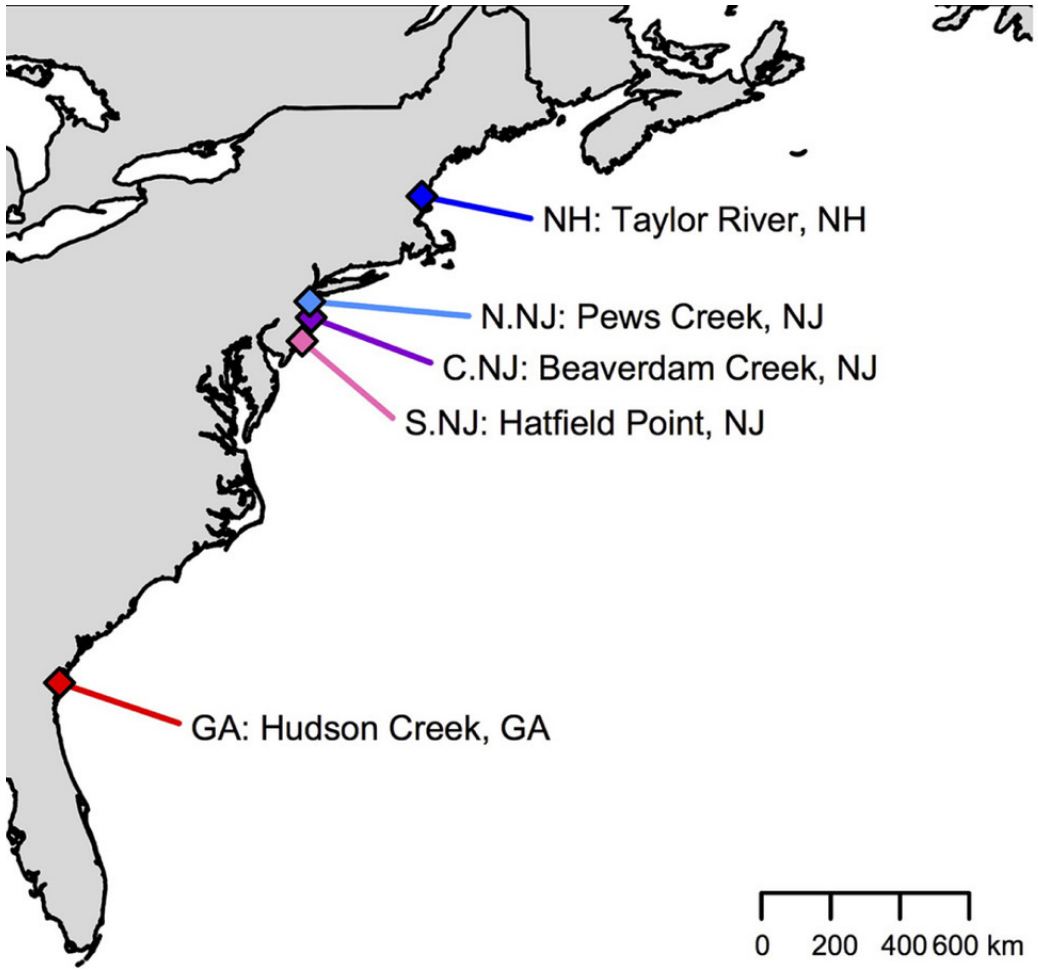
DOI: 10.1111/gcb.14386

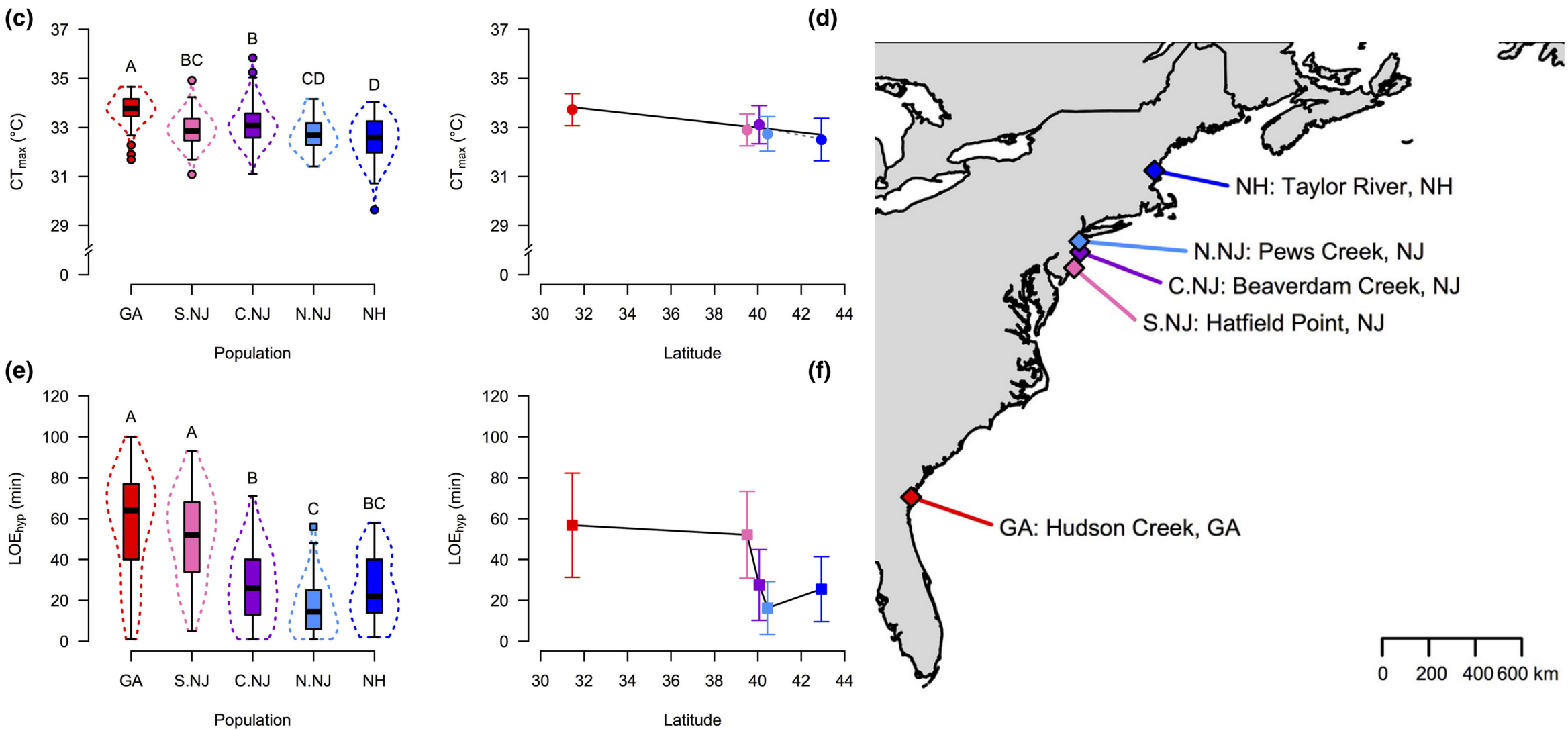
PRIMARY RESEARCH ARTICLE

WILEY Global Change Biology

Tolerance traits related to climate change resilience are independent and polygenic

Timothy M. Healy[1] | Reid S. Brennan[2] | Andrew Whitehead[2] | Patricia M. Schulte[1]

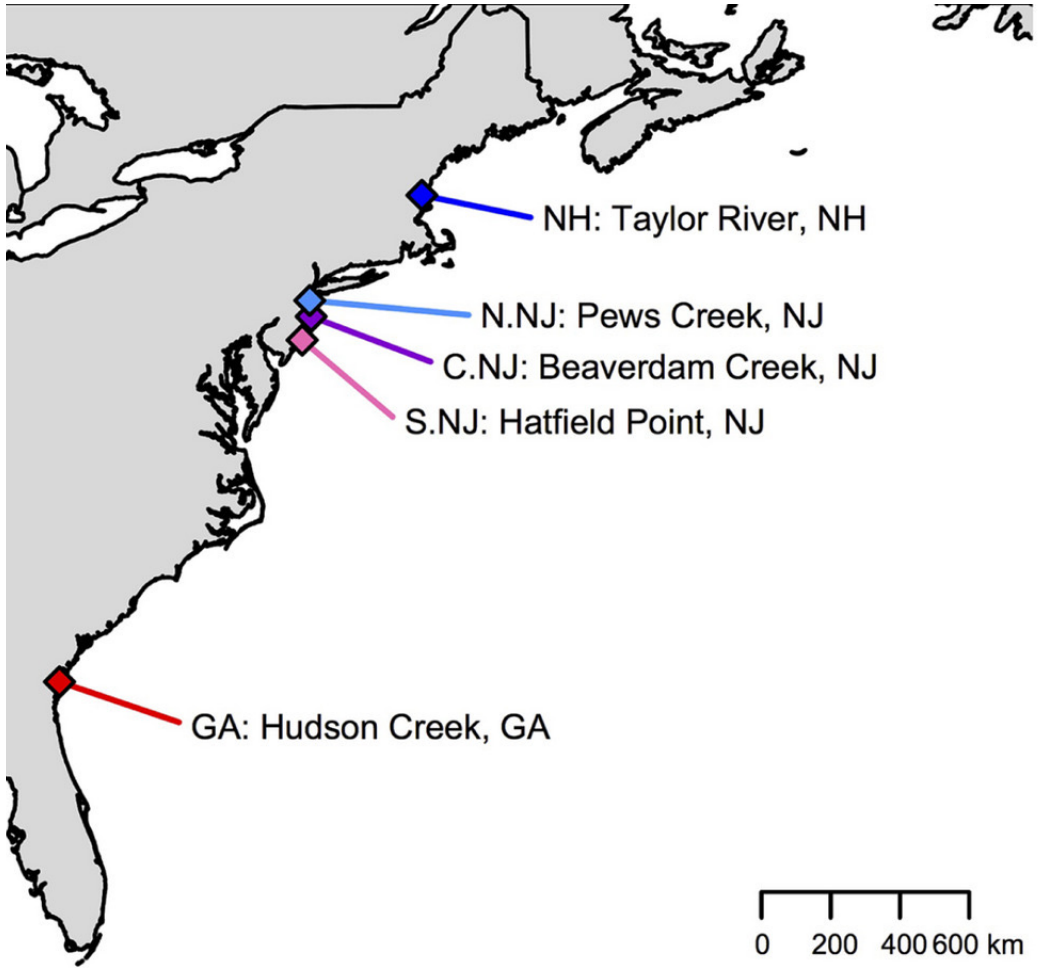



(c)
CTmax (°C)
37
35
33
31
29
0
A
BC
B
CD
D
GA
S.NJ
C.NJ
N.NJ
NH
Population
(d)
CTmax (°C)
30
32
34
36
38
40
42
44
Latitude
(e)
LOEhyp (min)
120
100
80
60
40
20
0
A
A
B
C
BC
GA
S.NJ
C.NJ
N.NJ
NH
Population
(f)
LOEhyp (min)
Latitude
NH: Taylor River, NH
N.NJ: Pews Creek, NJ
C.NJ: Beaverdam Creek, NJ
S.NJ: Hatfield Point, NJ
GA: Hudson Creek, GA
0
200
400
600 km

Dataset

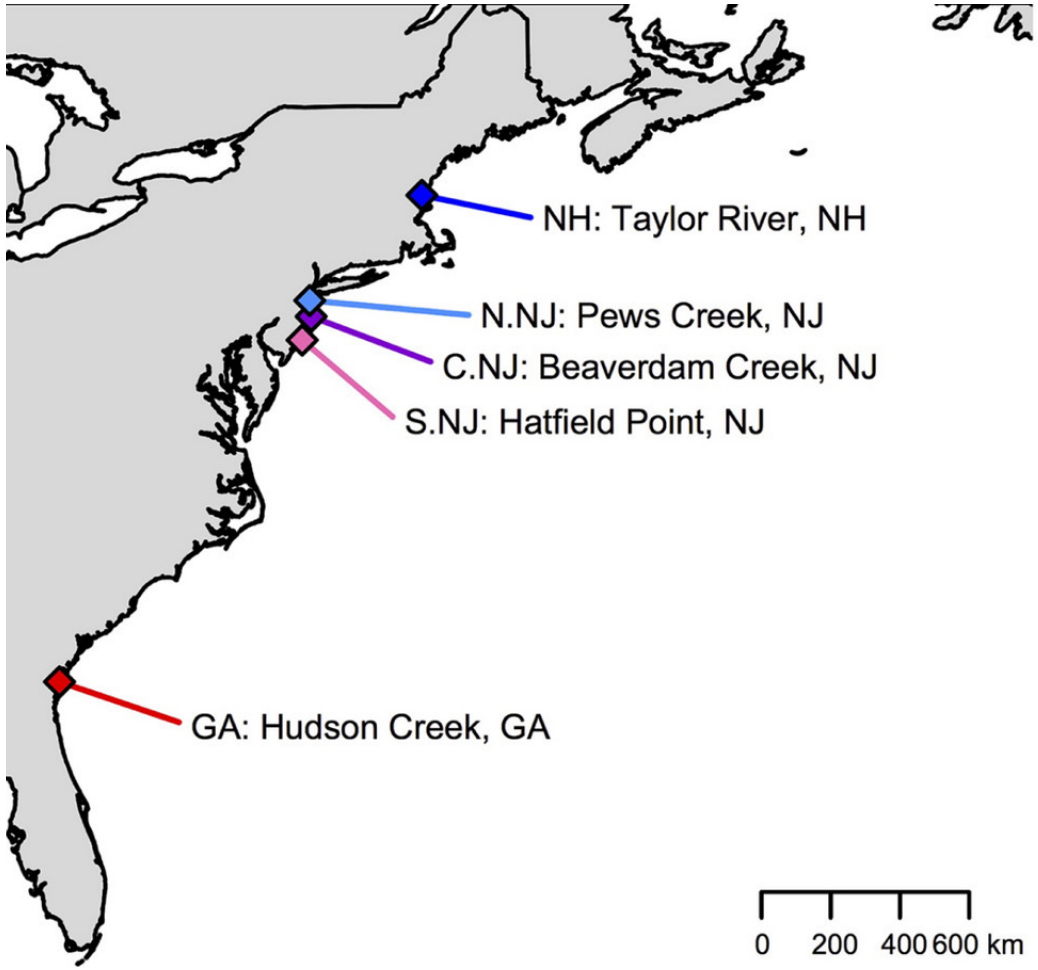
Population Location	Population ID	N
GA	GA	25
S.NJ	HP	46
C.NJ	BC	50
N.NJ	PC	44
NH	TR	45

RAD-seq



Dataset

- Population structure
- Genetic diversity
- Selection



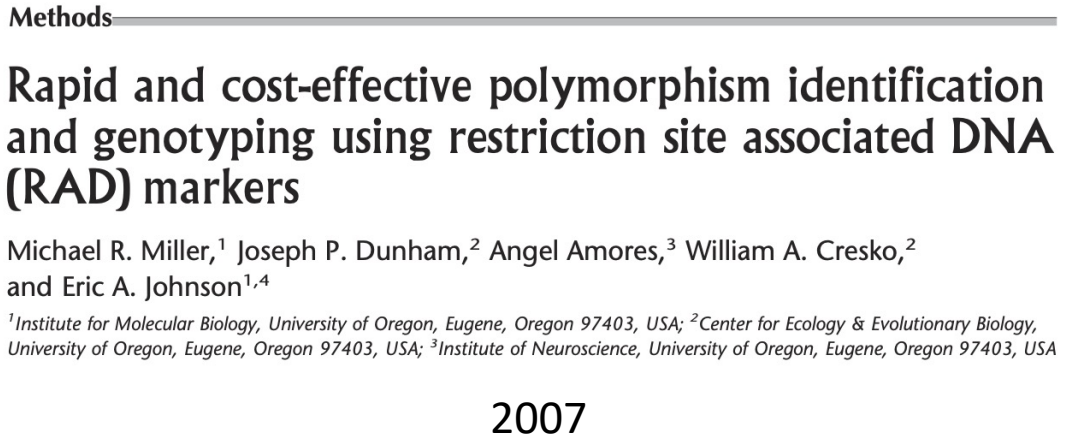
Reduced representation

- 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]

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

2007

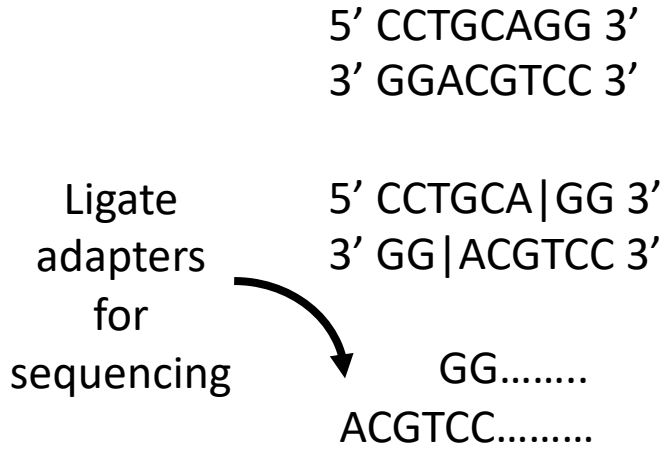
Reduced representation

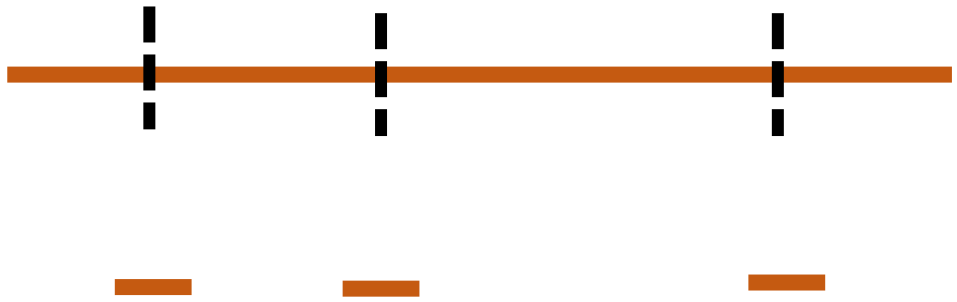
- In an ideal world…
  - Sequence full genomes of all individuals from many populations
  - Expensive, sometimes unnecessary, limited genomic resources

- 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





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, cloudlab, 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

“the cloud”

- Aka, a computer somewhere else

- We will use Cloudlab → hosted by CAU

R Markdown

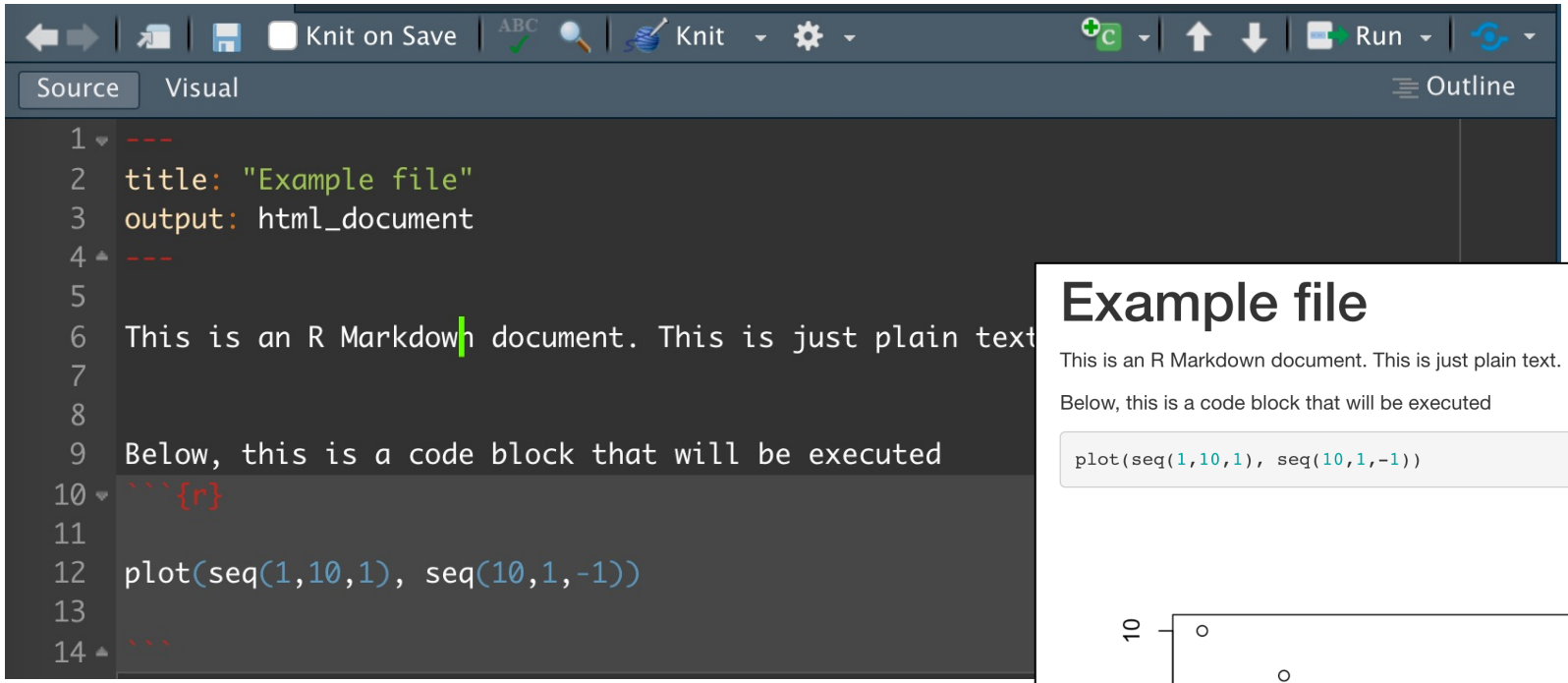
- You want to have a record of you analyses → markdown
- plain text formatting syntax designed to be converted to HTML and others
- R markdown is an extended markdown, that can embed and run R code, generate plots, etc.
- Why?
  - Reproducibility

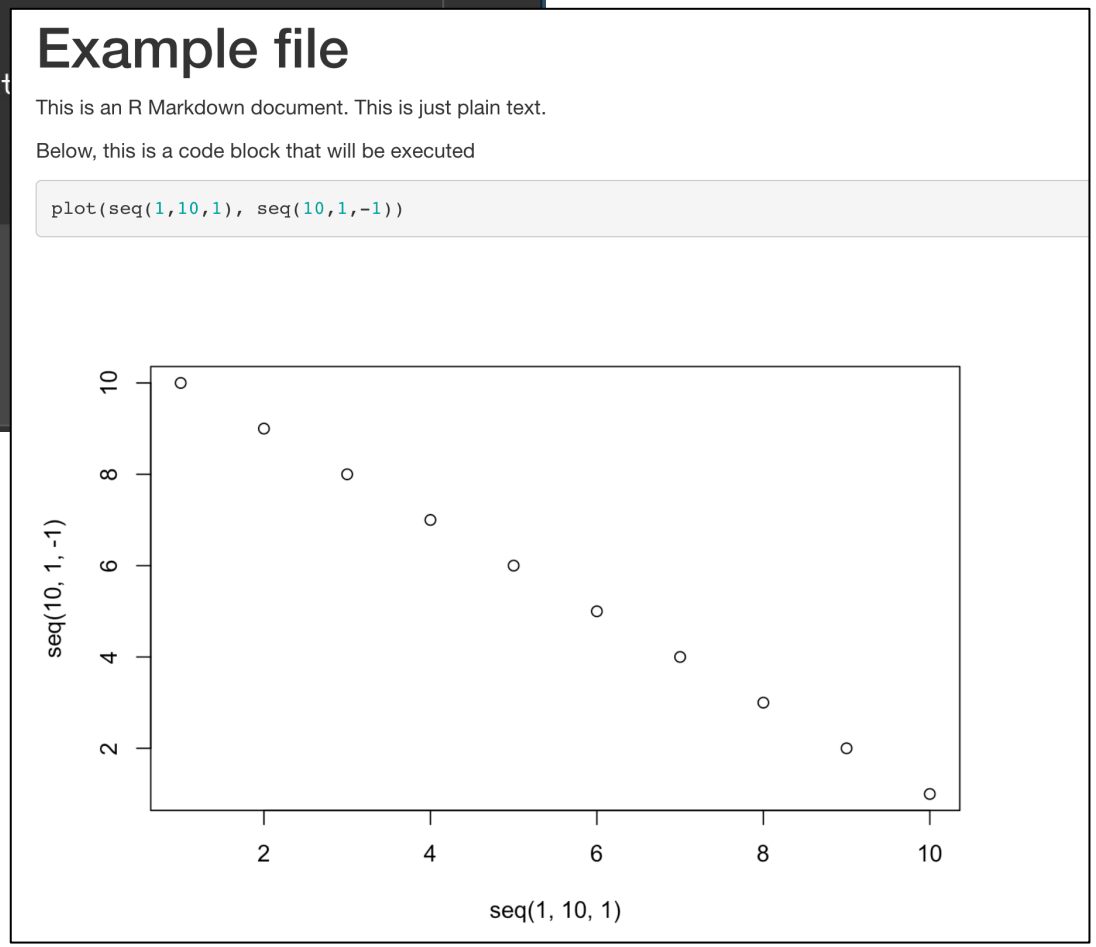
You will turn in a R markdown file with your final project

metadata

text

code chunks





Today’s tutorial

- Go to cloudlab together.
- Start exploring Rmarkdown
- Unix for biologists