

# Welcome to Bioinformatics Applications

## 2021 Spring

### Overview

Bioinformatics Applications (PLPTH813)

Sanzhen Liu

1/26/2021

# Goal

PLPTH813, Bioinformatics Applications, will cover the basic principles of regular bioinformatics applications and emphasize the *practice* of bioinformatics.

The goal of this course is to help you to be prepared for next-generation biological research that often generates *large data* and requires researchers to have the capability in data management and data mining.

# Course materials are online

## Course site at Github

<https://github.com/liu3zhenlab/teaching/tree/master/PLPTH813Bioinformatics/2021>

- Course information
- Lecture slide files
- Labs files

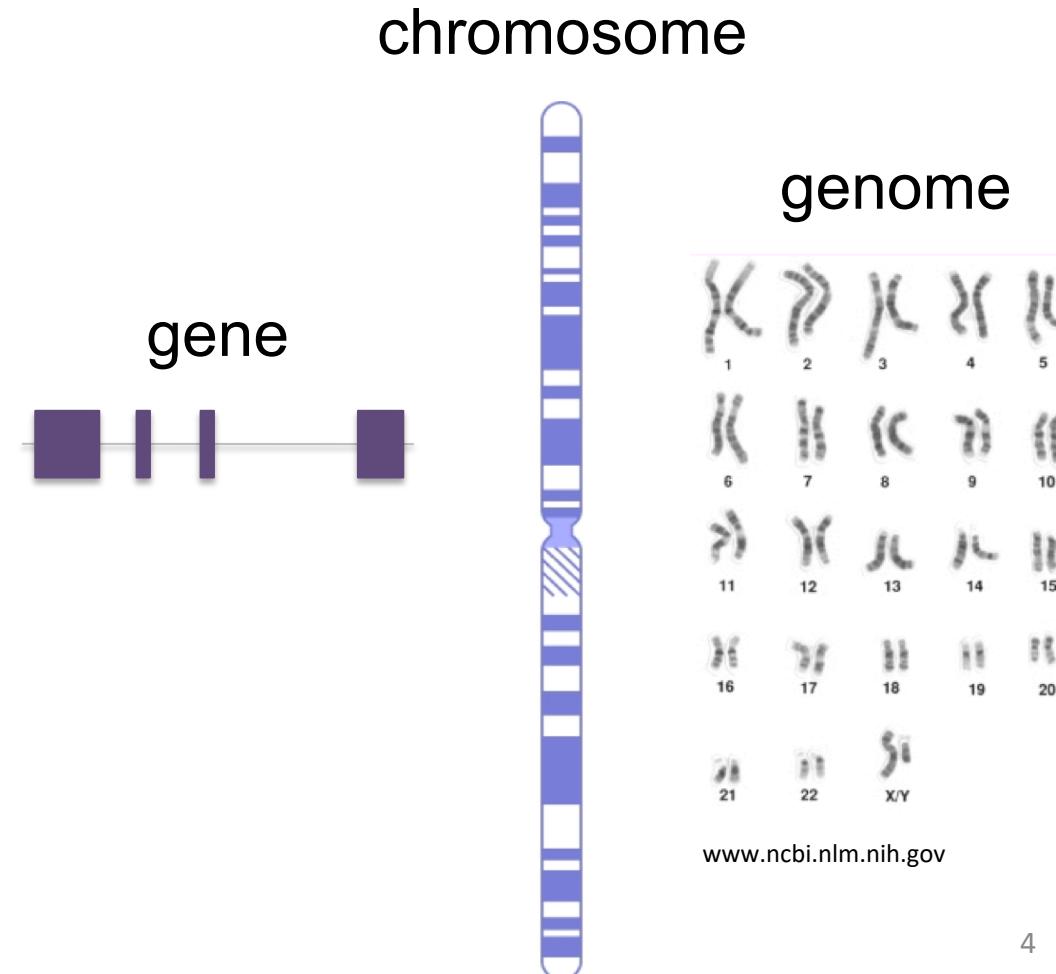
## K-State Canvas

# Bioinformatics

Bioinformatics is an interdisciplinary field that develops methods and software tools for understanding biological data.

## DNA sequencing data

```
1 ccttgaggct tttcgagcgc agcttccaa atcgatcca gatttccccg tccggggaa  
61 ggaggaccct gggaaaactg cgacgactat cttecccctgg ggcacatggac tccggggccca  
121 gctgggttc cagccggcccg tcgtcgccatg agcccgatga cctttttctg cggggccggaa  
181 gtaaaggccag cagccggcccg gccttcactcg gggggacccgt gtcttcgtcc acccccgatgt  
241 actggccggcc gggactgtggc gcccggatgc gcccggatcg ggggtctcgcc ggcggccatc  
301 ctggggggacat gcttggggacccatggggactggc agtggcttcgtcc gtcacggacccatggggatc  
361 cgtcgtggc ggttggctcgccatggggatcggccatggggatcggccatggggatcggccatggggatc  
421 tgacggatgt ggttggctcgccatggggatcggccatggggatcggccatggggatcggccatggggatc  
481 tggccatggatgt ggttggctcgccatggggatcggccatggggatcggccatggggatcggccatggggatc  
541 ttggccatggatgt ggttggctcgccatggggatcggccatggggatcggccatggggatcggccatggggatc  
601 tggggggatgt ggttggctcgccatggggatcggccatggggatcggccatggggatcggccatggggatc  
661 ctggggatgt ggttggctcgccatggggatcggccatggggatcggccatggggatcggccatggggatc  
721 cggccatggatgt ggttggctcgccatggggatcggccatggggatcggccatggggatcggccatggggatc  
781 ccggccatggatgt ggttggctcgccatggggatcggccatggggatcggccatggggatcggccatggggatc  
841 ccggccatggatgt ggttggctcgccatggggatcggccatggggatcggccatggggatcggccatggggatc  
901 ctggccatggatgt ggttggctcgccatggggatcggccatggggatcggccatggggatcggccatggggatc  
961 agggggatgt ggttggctcgccatggggatcggccatggggatcggccatggggatcggccatggggatc  
1021 acgtgtggatgt ggttggctcgccatggggatcggccatggggatcggccatggggatcggccatggggatc  
1081 actggggcccg gggccgttctg gggccatggggatcggccatggggatcggccatggggatcggccatggggatc  
1141 cttggccatggatgt ggttggctcgccatggggatcggccatggggatcggccatggggatcggccatggggatc  
1201 gggccatggatgt ggttggctcgccatggggatcggccatggggatcggccatggggatcggccatggggatc  
1261 gggggggatgt ggttggctcgccatggggatcggccatggggatcggccatggggatcggccatggggatc  
1321 ctggccatggatgt ggttggctcgccatggggatcggccatggggatcggccatggggatcggccatggggatc
```



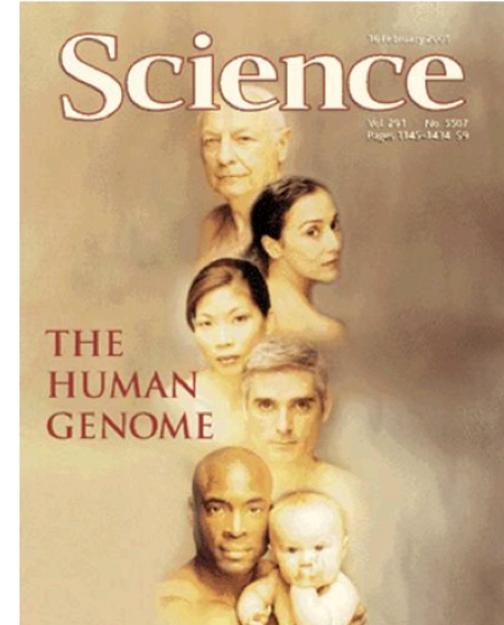
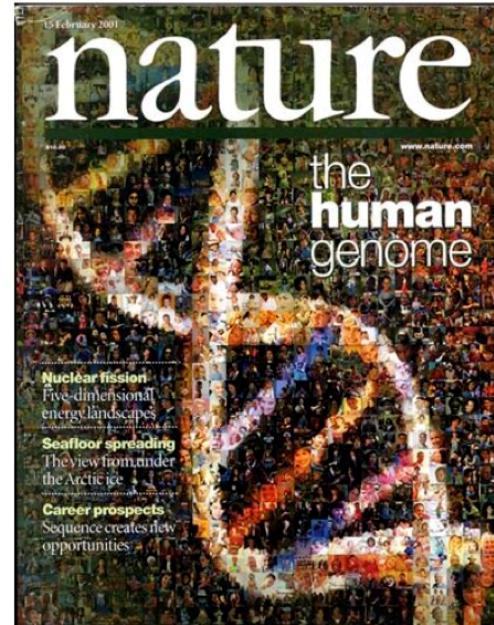
# Human Genome Project (HGP)

1st-gen sequence  
(Sanger)

- International Human Genome Sequencing Consortium  
Proposed 1985, endorsed in 1988; BAC-by-BAC (Francis Collins et al)
- Craig Venter & Celera Genomics:  
Founded 1998, finished in 3 years; whole genome shotgun

>sample  
TGCGGCCTACTAACCGGTTCTGAGAGTCTGAGATG  
AGAGAAATGCCCTACTAACCGGTTCTGAGAATGCC  
CTAACCGATGCCCTACTAACCGGTTCTGAGAGTTCTG  
AGCTGAGATGCCCTACTAACCGGTTCTGAGAATGCC  
TAACACCGATGCCCTACTAACCGGTTCTGAGAGTTCTG  
TGAGATGAGAGAAATGCCCTACTAACCGGTTCTGAGA  
ATGCCCTACTAACCGATGCCCTACTAACCGGTTCTGAG  
AGTCTGAGCTATGCCCTACTAACCGGTTCTGAGAATGCC  
CTAACCGATGCCCTACTAACCGGTTCTGAGA  
GTTCTGAGATGAGAGAAATGCCCTACTAACCGGTTCTG  
GAGAATGCCCTACTAACCGGTTCTGAGAATGCCCTACTAACCGG  
GAATGCCCTACTAACCGATGCCCTACTAACCGGTTCTG  
AGAGTCTGAGATGAGAGAAATGCCCTACTAACCGGTTCTG  
TCTGAGAATGCCCTACTAACCGATGCCCTACTAACCGG  
TTCTGAGAGTTCTGAGCTCCGATGCCCTACTAACCGG  
TTCTGAGAGTTCTGAGCTAATCTAACCGGTTCTG  
CTACTAACCGGTTCTGAGAAATGCCCTACTAACCGGTTCTG  
CTGAGAATGCCCTACTAACCGATGCCCTACTAACCGGTTCTG  
TCTGAGAGTTCTGAGATGAGAGAAATGCCCTACTAAC  
CGGTTCTGAGAAATGCCCTACTAACCGATGCCCTACTAAC  
CGGTTCTGAGAGTTCTGAGCTGAGAA

February 2001 - Publication of the first draft of the human genome



First draft

Polling time

Human genome questions

# DNA sequencing technology

1st-gen sequence  
(Sanger)

1980 Nobel Prize

```
>sample
TGCGGCCCTACTAACCGGTTCTGAGAGTTCTGAGATG
AGAGAATGCCTACTAACCGGTTCTGAGAAATGCCTA
CTAACCGATGCCTACTAACCGGTTCTGAGAGTTCTGAG
AGCTGAGATGCCTACTAACCGGTTCTGAGAAATGCC
TACTAACCGATGCCTACTAACCGGTTCTGAGAGTTCTGAG
TGAGATGAGAGAAATGCCTACTAACCGGTTCTGAGA
ATGCCTACTAACCGATGCCTACTAACCGGTTCTGAG
AGTTCTGAGCTATGCCTACTAACCGGTTCTGAGAAT
GCCCTACTAACCGATGCCTACTAACCGGTTCTGAGA
GTTCTGAGATGAGAGAAATGCCTACTAACCGGTTCTGAG
GAGAAATGCCTACTAACCGATGCCTACTAACCGGTTCTGAG
GAATGCCTACTAACCGATGCCTACTAACCGGTTCTGAG
AGAGTTCTGAGATGAGAGAAATGCCTACTAACCGGTTCTGAG
TCTGAGAAATGCCTACTAACCGATGCCTACTAACCGGTTCTGAG
TTCTGAGAGTTCTGAGCTCGATGCCTACTAACCGGTTCTGAG
TCTGAGAGTTCTGAGCTAATACTAACCGGTTCTGAG
CTGAGAAATGCCTACTAACCGATGCCTACTAACCGGTTCTGAG
TCTGAGAGTTCTGAGATGAGAGAAATGCCTACTAACCGGTTCTGAG
CGGTTCGAGAAATGCCTACTAACCGATGCCTACTAACCGGTTCTGAG
CCGGTTCTGAGAGTTCTGAGCTGAGAA
```

next-gen sequence (NGS)



800 letters

billions of letters

# Sequencing cost

cost per megabase

1970's Sanger sequencing

\$5,000

Roche 454

\$1000

\$100

cost

\$10

\$1

\$0.1

\$0.05

\$0.02

Illumina

Short reads (<600bp)  
Illumina

Illumina

Sep-01 Mar-02 Sep-02 Mar-03 Oct-03 Jan-04 Apr-04 Jul-04 Oct-04 Jan-05 Apr-05 Jul-05 Oct-05 Jan-06 Apr-06 Jul-06 Oct-06 Jan-07 Apr-07 Jul-07 Oct-07 Jan-08 Apr-08 Jul-08 Oct-08 Jan-09 Apr-09 Jul-09 Oct-09 Jan-10 Apr-10 Jul-10 Oct-10 Jan-11 Apr-11 Jul-11 Oct-11 Jan-12 Apr-12 Jul-12 Oct-12 Jan-13 Apr-13 Jul-13 Oct-13 Jan-14 Apr-14 Jul-14 Oct-14 Jan-15 Apr-15 Jul-15 Oct-15

time

Data source: genome.gov/sequencingcosts



# Long-read sequencing



Nanopore:MinION



Nanopore:PromethION



PacBio RSII



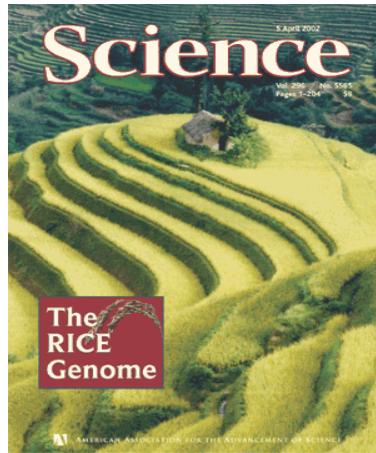
PacBio Sequel

- Produce long reads (>10 kb, single-molecule reads, high errors)
- PacBio Sequel (Repeated sequencing of a fragment to achieve high accuracy (e.g., 99%))

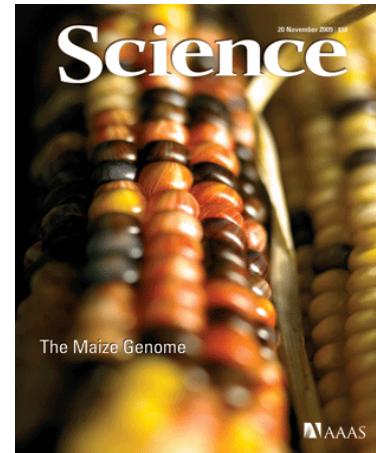
# Sequence genomes of model species



2000



2002



2009



2018



Sequence EVERY species

# Sequence “populations”

**nature**

Explore content ▾ Journal information ▾ Publish with us ▾

nature > articles > article

Article | Open Access | Published: 25 April 2018

## Genomic variation in 3,010 diverse accessions of Asian cultivated rice

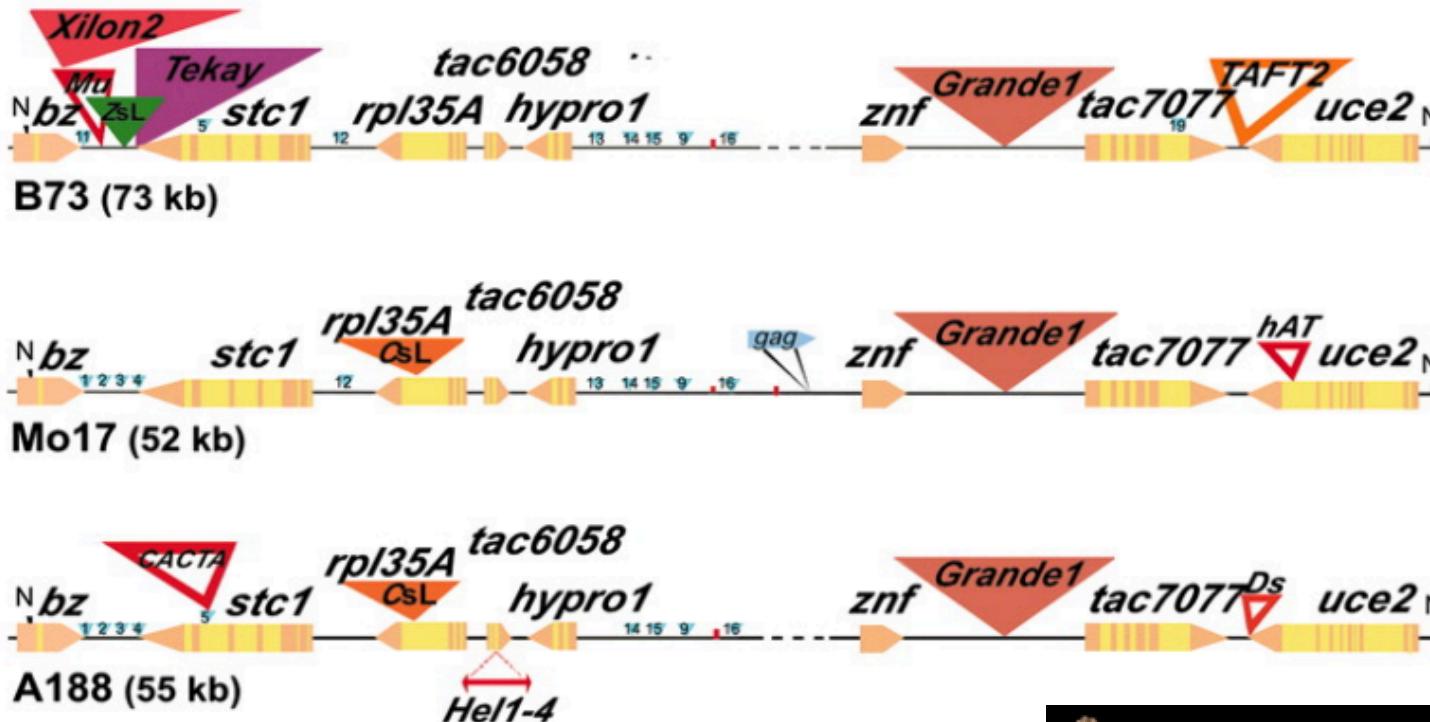
RESEARCH ARTICLE

## Deep sequencing of 10,000 human genomes

SARS-CoV-2  
(as of 1/24/2021)

**52,969**  
Nucleotide records

# Comparative genomics (I)

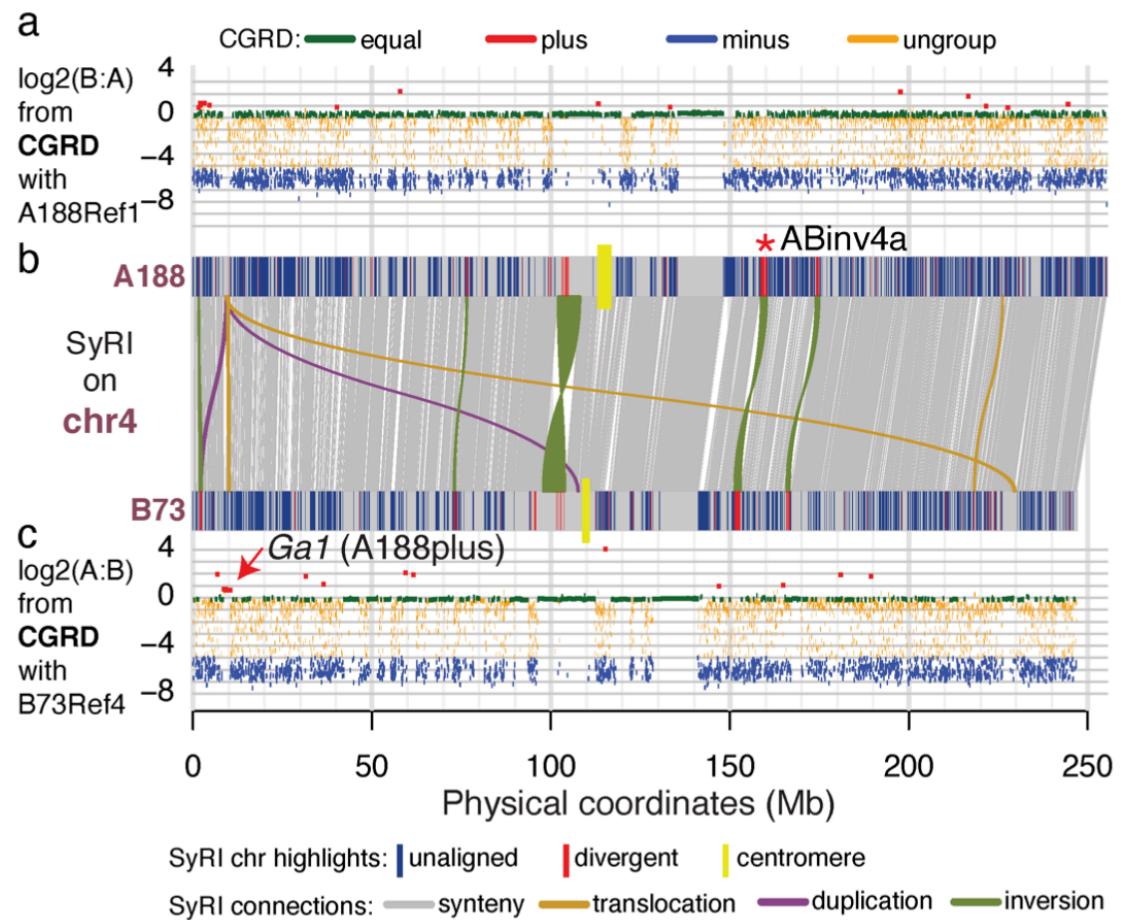


Remarkable variation in maize genome structure at the *bz* locus

PNAS, 2006, 103:17644-49



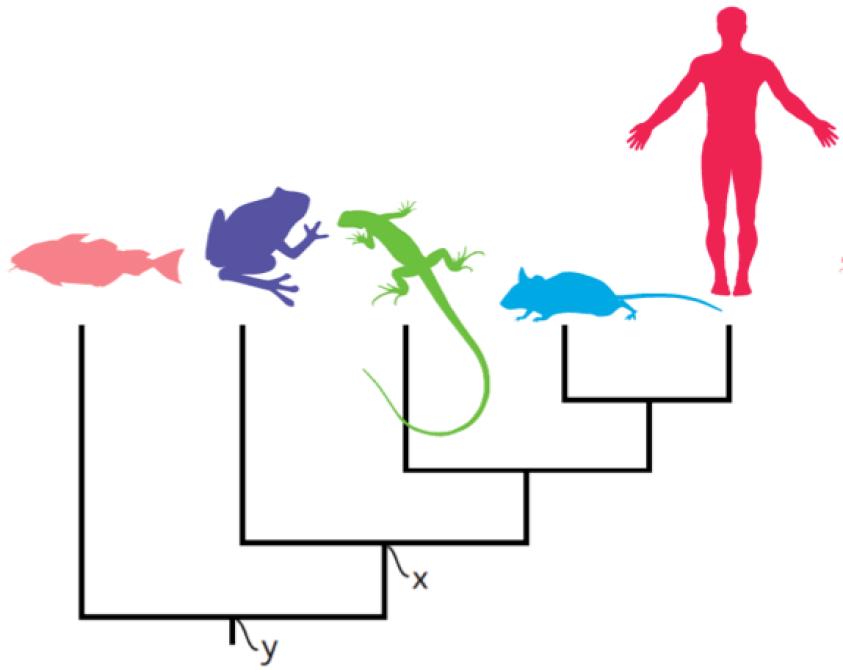
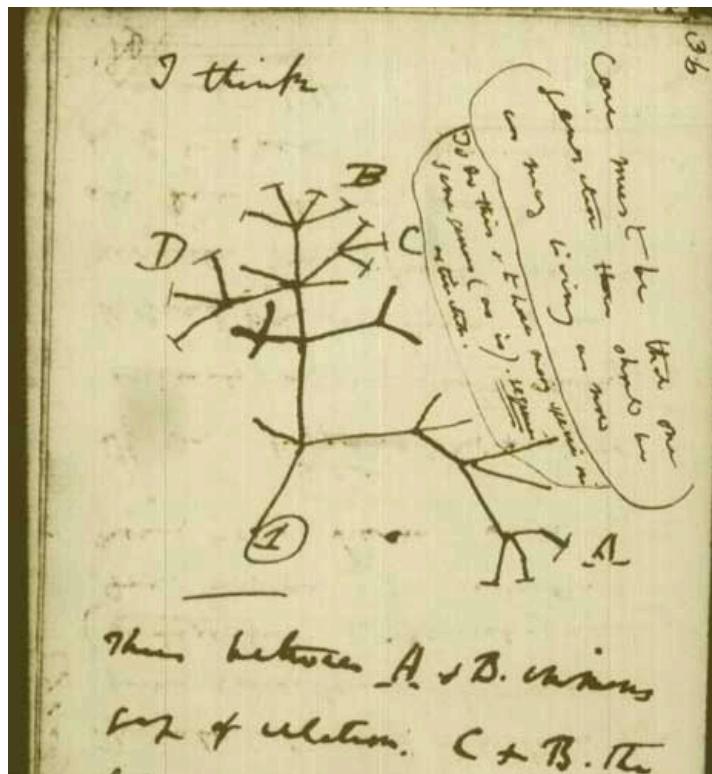
# Comparative genomics (II)



# NGS is changing the way to discover genetic variants

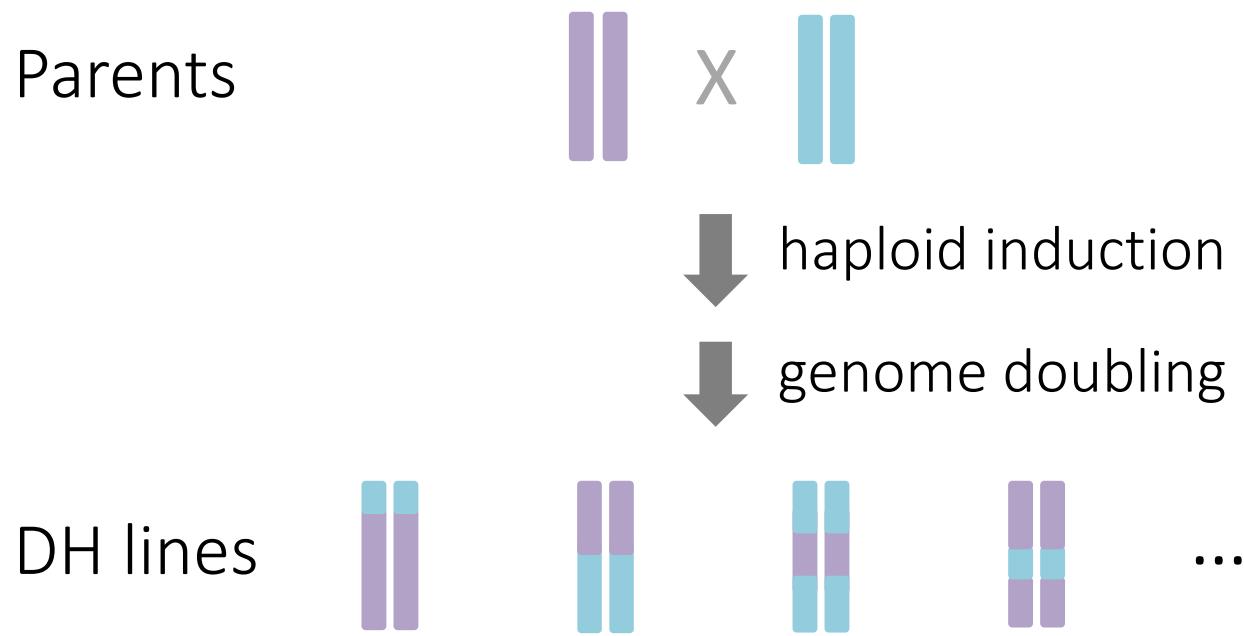
Reference	ATCGCTGCCGATCTGCGTCATA CGGAATCGTCGGCTTCAG
Sequences	ATCGCTGCCGATCTGCGTCATA CGGAATCGTCGGCTTCAG
	ATCGCTGCCGATCTGCGT <b>G</b> ATAC CGGAATCGTCGGCTTCAG
	ATCGCTGCCGATCTGCGTCATA CGGAATCGTCGGCTTCAG
	ATCGCTGCCGATCTGCGT <b>G</b> ATAC CGGAATCGTCGGCTTCAG
	ATCGCTGCCGATCTGCGT <b>G</b> ATAC CGGAATCGTCGGCTTCAG
Sequences	ATCGCTGCCGATCTGCGT <b>G</b> ATAC CGGAATCGTCGGCTTCAG
	ATCGCTGCCGATCTGCGTCATA CGGAATCGTCGGCTTCAG
	ATCGCTGCCGATCTGCGTCATA CGGAATCGTCGGCTTCAG
	ATCGCTGCCGATCTGCGT <b>G</b> ATAC CGGAATCGTCGGCTTCAG
	ATCGCTGCCGATCTGCGTCATA CGGAATCGTCGGCTTCAG
Genotype	----- <b>C/G</b> -----

# Phylogeny



in-class project: how to build a phylogenetic tree from sequencing data

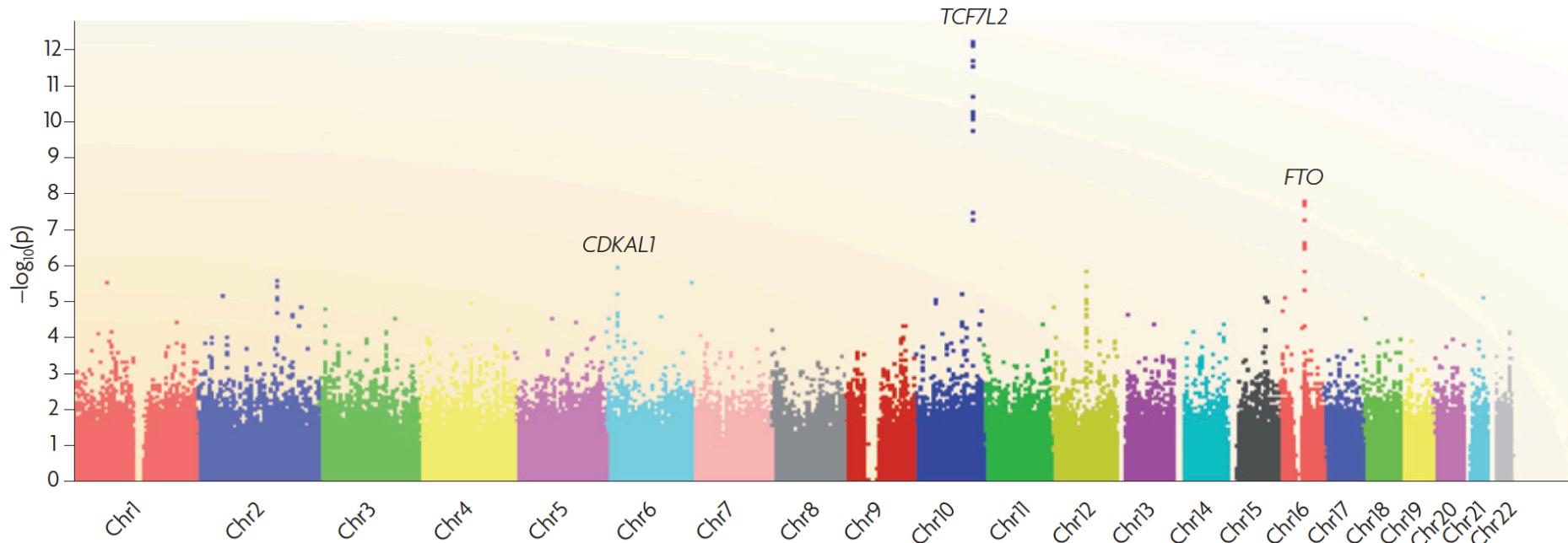
# Connect genotype with phenotype (I)



Map QTL using phenotype and genotype data

# Connect genotype with phenotype (II)

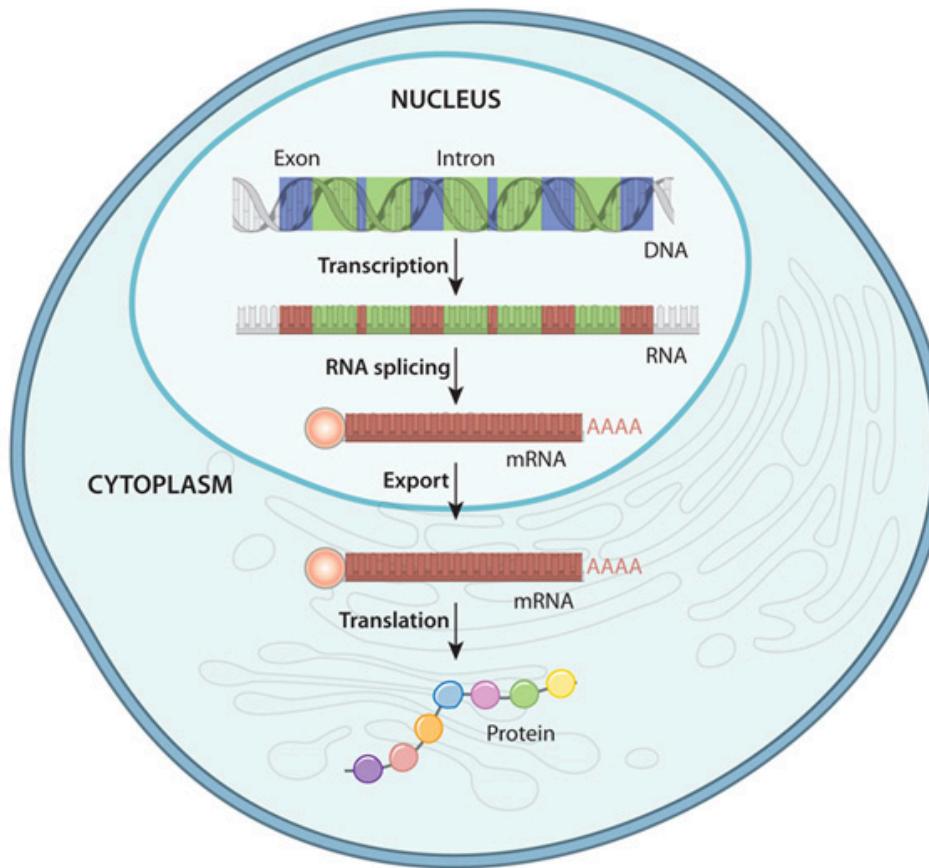
A Manhattan plot from Genome-wide association mapping (GWAS)



McCarthy et al., Nature Review Genetics, 2008: 9:356-369

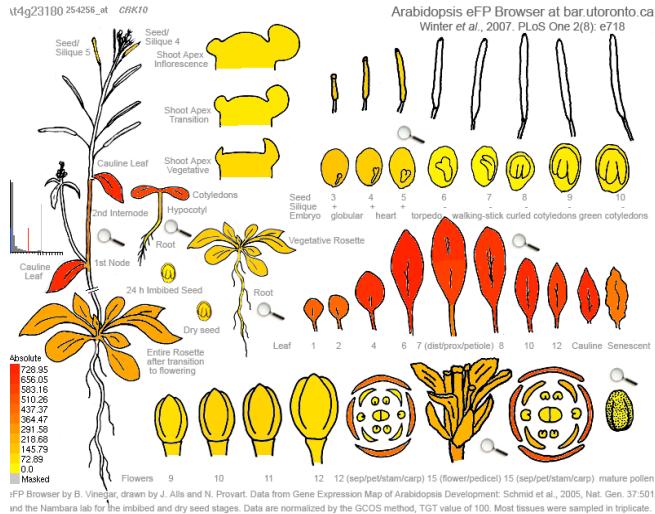
- To determine the genetic basis of a trait based on:
  - Genotyping data (from sequencing or other sources)
  - Phenotyping data (all kinds of trait data)

# Complexity of transcriptome



In many eukaryotic organisms, the majority of genes are alternatively spliced to produce multiple transcripts, or isoforms.

# Transcriptome analysis



Expression profiles in different tissues



Response to biotic stress

1. What are sequences of transcripts?
2. What is the expression level of each transcript?

RNA-Seq addresses both questions pretty well

# Environmental microbiomes



Water



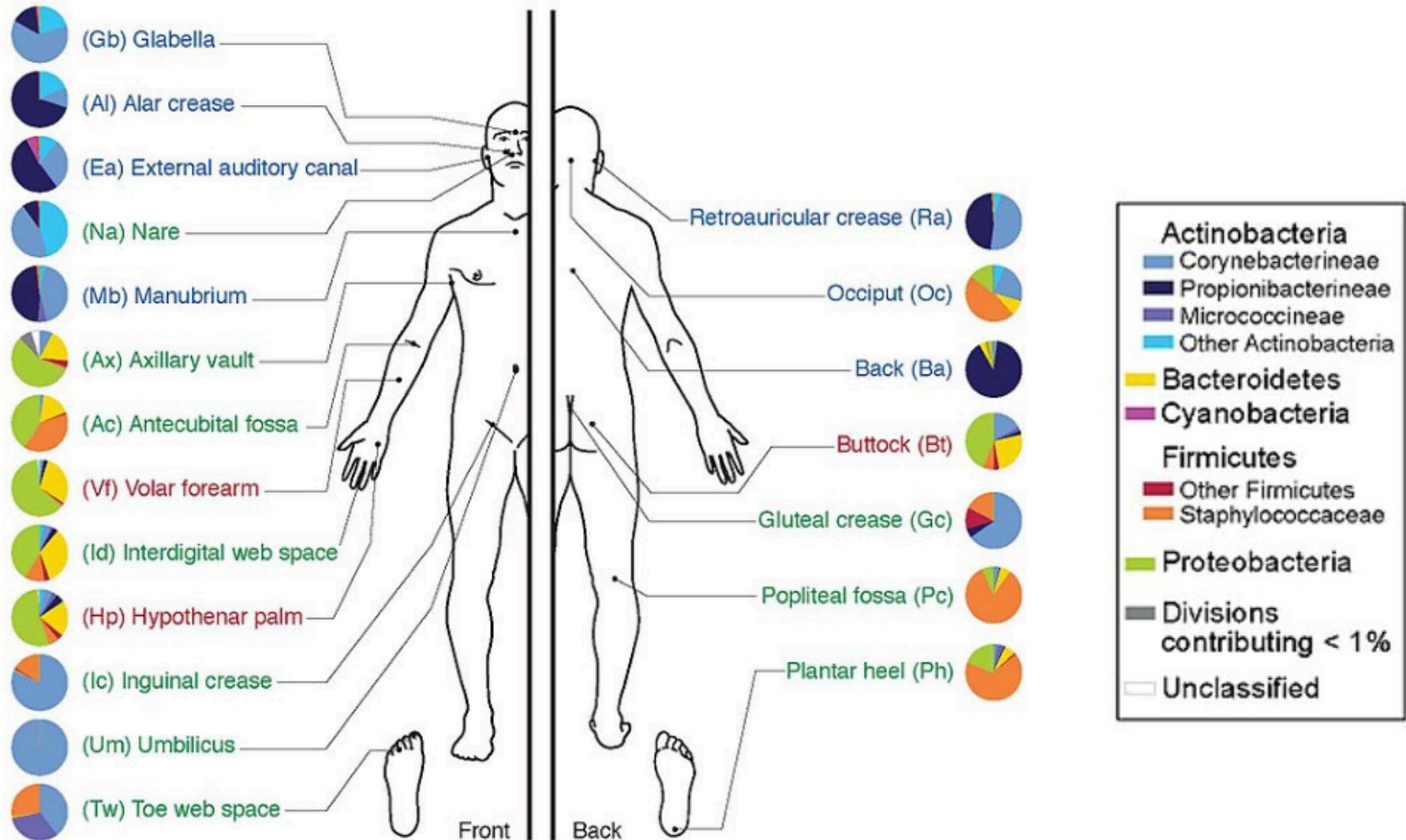
Plants



Soils



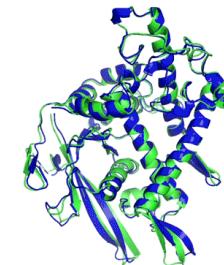
# Human microbiomes



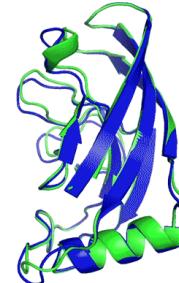
# Deep learning



autopilot



T1037 / 6vr4  
90.7 GDT  
(RNA polymerase domain)

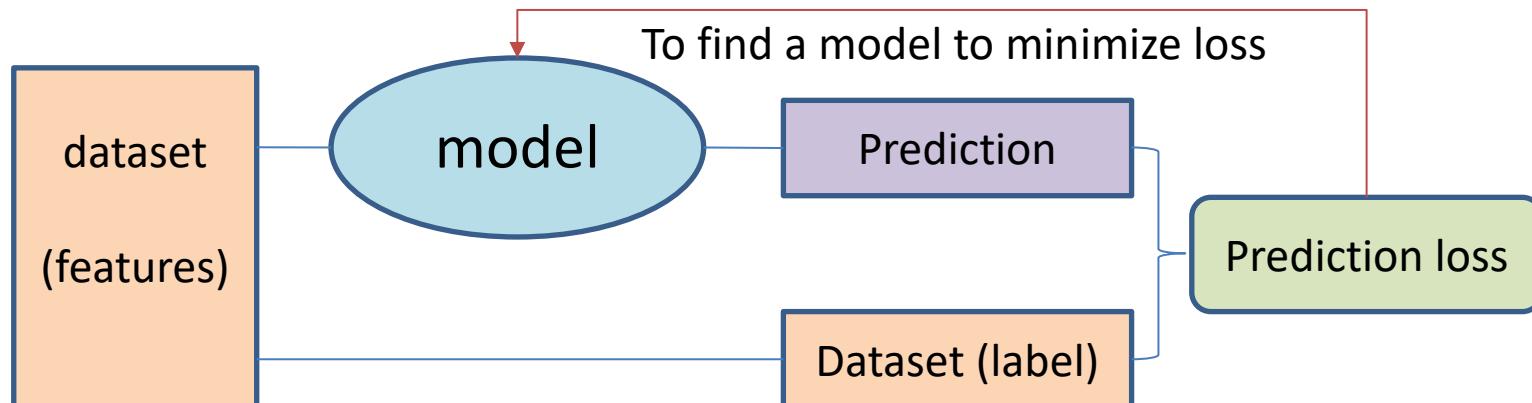


T1049 / 6y4f  
93.3 GDT  
(adhesin tip)

● Experimental result  
● Computational prediction

AlphaFold

## optimization procedure



# Reasons for command-lines analyses

- To perform *efficient* and *reproducible* data analyses
- To use advanced tools in research projects (most genomic software packages are run in the Unix system)
- To access to powerful computer servers (e.g., beocat)

Excel:

Order	Group1	Group2
1	12	1
2	10	5
3	35	
mean	19	=AVERAGE(C2:C4)

R program:

```
mean(group1)  
mean(group2)
```

# Lecture topics

1. Basic Unix
2. Basic R
3. Introduction of NGS and NGS bioinformatics tools
4. DNA sequence alignment
5. Genome variants
6. Phylogeny
7. QTL and GWAS
8. Genome assembly
9. Comparative genomics
10. Metagenomics
11. RNA-Seq
12. Deep learning

# Student Projects

March 25<sup>th</sup>: 5 min presentation to talk about project plan

Students are expected to design their projects after February.

The project can be related to students' own research projects or utilize public data for a meaningful analysis.

Ideas will be proposed in a 5-minute presentation during the class (March 25<sup>th</sup>). The presentation should include the goal, the rationale, the data source, and the expected result.

Before the final week, students will present the results from the projects. Each presentation will take ~15 minutes.

# Project examples

- Comparison of algorithms for genome assemblies
- GWAS of a trait
- Differential expression of wheat plants under cold conditions as compared to wheat plants under a normal condition
- Reproduce a work from a paper

# Grading

- Grading

Participation 10%, Homework 30%, Midterm Exam 20%,  
Student project 15%, final Exam 25%

- Homework: 8+ times

- Project presentation

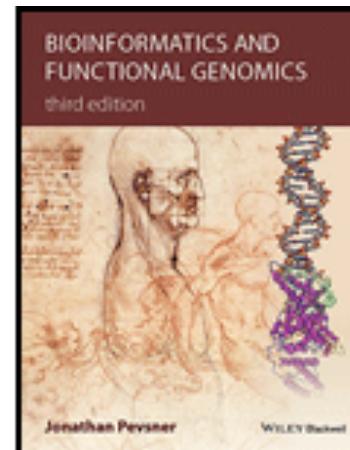
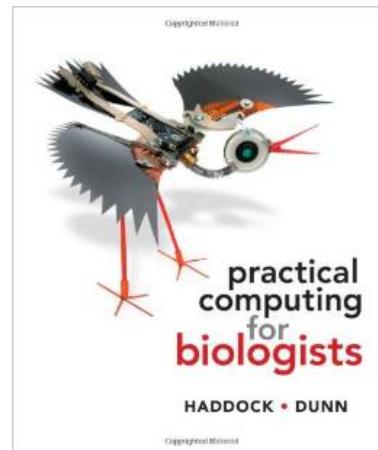
5 mins + 15 mins

- Two exams

Midterm (March 11<sup>th</sup>, in class) and final exam

# References

- Papers
- Online resources (e.g., Wikipedia)
- Practical computing for biologists, Haddock and Dunn, 2010
- Bioinformatics and Functional Genomics, Pevsner, 2015



# Schedule

ZOOM: <https://ksu.zoom.us/j/95782921551>

**Time:**

Tuesday, Thursday 10:30am-11:20pm (lectures);

Thursday 12:30-2:00pm (lab)

**Office hours:** Tuesday 12:30-1:30pm

<https://ksu.zoom.us/j/8468443307> (appointment is required)