Welcome to Bioinformatics Applications Spring 2021

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/PLPTH813Bioinformatis/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.

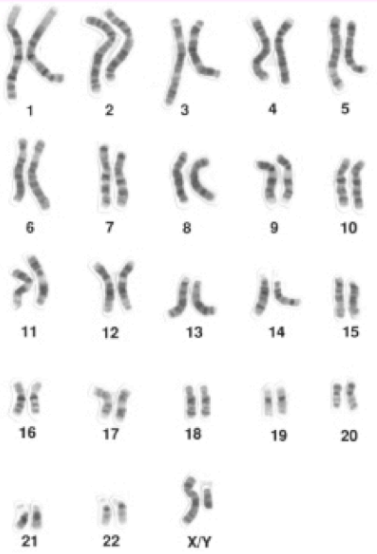
DNA sequencing data

```
   1 ccctgaggct tttcggagcg agctcctcaa atcgcatcca gattttcggg tccgagggaa
  61 ggaggaccct gcgaaagctg cgacgactat cttcccctgg ggccatggac tcggacgcca
 121 gcctggtgtc cagccgcccg tcgtcgccag agcccgatga ccttttctg ccggcccgga
 181 gtaagggcag cagcggcagc gccttcactg ggggcaccgt gtcctcgtcc accccgagtg
 241 actgcccgcc ggagctgagc gccgagctgc gcggcgctat gggctctgcg ggcgcgcatc
 301 ctggggacaa gctaggaggc agtggcttca agtcgtcctc gtccagcacc tcgtcgtcta
 361 cgtcgtcggc ggctgcgtcg tccaccaaga aggacaagaa gcaaatgaca gagccggagc
 421 tgcagcagct gcgtctcaag atcaacagcc gcgagcgcaa gcgcatgcac gacctcaaca
 481 tcgccatgga tggcctccgc gaggtcatgc cgtacgcaca cggcccttcg gtgcgcaagc
 541 tttccaagat cgccacgctg ctgctggcgc gcaactacat cctcatgctc accaactcgc
 601 tggaggagat gaagcgactg gtgagcgaga tctacggggg ccaccacgct ggcttccacc
 661 cgtcggcctg cggcggcctg gcgcactccg cgcccctgcc cgccgccacc gcgcacccgg
 721 cagcagcagc gcacgccgca catcaccccg cggtgcacca ccccatcctg ccgcccgccg
 781 ccgcagcggc tgctgccgcc gctgcagccg cggctgtgtc cagcgcctct ctgcccggat
 841 ccgggctgcc gtcggtcggc tccatccgtc caccgcacgg cctactcaag tctccgtctg
 901 ctgccgcggc cgccccgctg gggggcgggg gcggcggcag tggggcgagc gggggcttcc
 961 agcactgggg cggcatgccc tgcccctgca gcatgtgcca ggtgccgccg ccgcaccacc
1021 acgtgtcggc tatgggcgcc ggcagcctgc cgcgcctcac ctccgacgcc aagtgagccg
1081 actggcgccg gcgcgttctg gcgacagggg agccaggggc cgcggggaag cgaggactgg
1141 cctgcgctgg gctcgggagc tctgtcgcga ggaggggcgc aggaccatgg actgggggtg
1201 gggcatggtg gggattccag catctgcgaa cccaagcaat gggggcgccc acagagcagt
1261 ggggagtgag gggatgttct ctccgggacc tgatcgagcg ctgtctggct ttaacctgag
1321 ctggtccagt agacatcgtt ttatgaaaag gtaccgctgt gtgcattcct cactagaact
```

gene

chromosome

genome

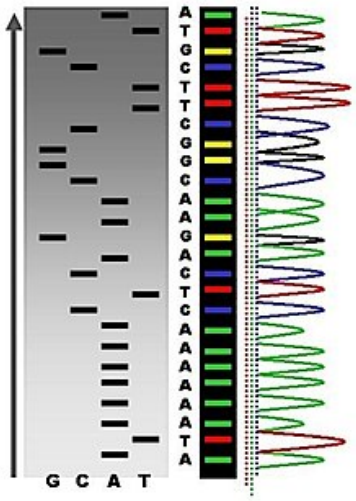


www.ncbi.nlm.nih.gov

Human Genome Project (HGP)

1st-gen sequence (Sanger)

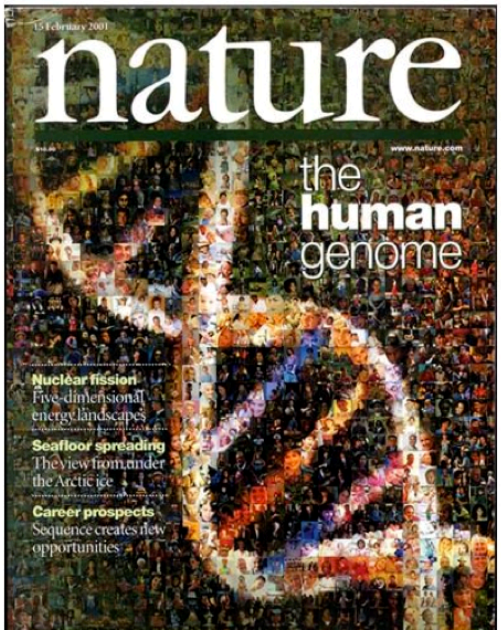
>sample
TGCCGCCTACTAACCGGTTCTGAGAGTTCTGAGATG
AGAGAATGCCTACTAACCGGTTCTGAGAATGCCTA
CTAACCGATGCCTACTAACCGGTTCTGAGAGTTCTG
AGCTGAGATGCCTACTAACCGGTTCTGAGAATGCC
TACTAACCGATGCCTACTAACCGGTTCTGAGAGTTC
TGAGATGAGAGAATGCCTACTAACCGGTTCTGAGA
ATGCCTACTAACCGATGCCTACTAACCGGTTCTGAG
AGTTCTGAGCTATGCCTACTAACCGGTTCTGAGAAT
GCCTACTAACCGATGCCTACTAACCGGTTCTGAGA
GTTCTGAGATGAGAGAATGCCTACTAACCGGTTCT
GAGAATGCCTACTAAATGCCTACTAACCGGTTCTGA
GAATGCCTACTAACCGATGCCTACTAACCGGTTCTG
AGAGTTCTGAGATGAGAGAATGCCTACTAACCGGT
TCTGAGAATGCCTACTAACCGATGCCTACTAACCGG
TTCTGAGAGTTCTGAGCTCCGATGCCTACTAACCGG
TTCTGAGAGTTCTGAGCTAATACTAACCGGTTATGC
CTACTAACCGGTTCTGAGAATGCCTACTAACCGGTT
CTGAGAATGCCTACTAACCGATGCCTACTAACCGGT
TCTGAGAGTTCTGAGATGAGAGAATGCCTACTAAC
CGGTTCTGAGAATGCCTACTAACCGATGCCTACTAA
CCGGTTCTGAGAGTTCTGAGCTGAGAA

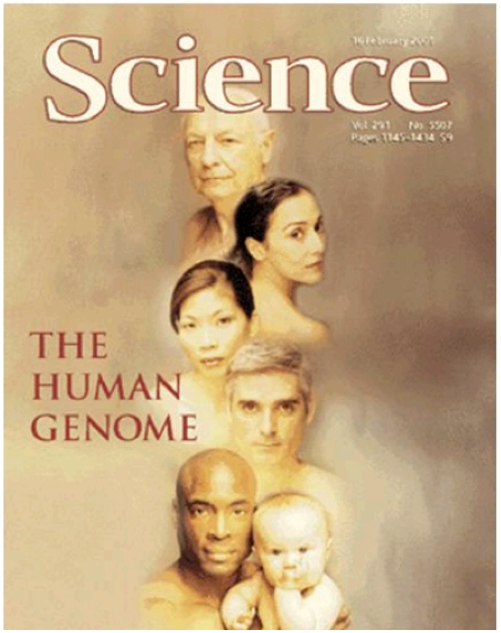


- International Human Genome Sequencing Consortium
Proposed 1985, endorsed in 1988; BAC-by-BAC (Francis Collins et al)

- Craig Venter & Celera Genomics:
Founded 1998, finished in 3 years; whole genome shotgun

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





First draft

Polling 1

Human genome questions

DNA sequencing technology

1st-gen sequence (Sanger)

1980 Nobel Prize

```
>sample
TGCCGCCTACTAACCGGTTCTGAGAGTTCTGAGATG
AGAGAATGCCTACTAACCGGTTCTGAGAATGCCTA
CTAACCGATGCCTACTAACCGGTTCTGAGAGTTCTG
AGCTGAGATGCCTACTAACCGGTTCTGAGAATGCC
TACTAACCGATGCCTACTAACCGGTTCTGAGAGTTC
TGAGATGAGAGAATGCCTACTAACCGGTTCTGAGA
ATGCCTACTAACCGATGCCTACTAACCGGTTCTGAG
AGTTCTGAGCTATGCCTACTAACCGGTTCTGAGAAT
GCCTACTAACCGATGCCTACTAACCGGTTCTGAGA
GTTCTGAGATGAGAGAATGCCTACTAACCGGTTCT
GAGAATGCCTACTAAATGCCTACTAACCGGTTCTGA
GAATGCCTACTAACCGATGCCTACTAACCGGTTCTG
AGAGTTCTGAGATGAGAGAATGCCTACTAACCGGT
TCTGAGAATGCCTACTAACCGATGCCTACTAACCGG
TTCTGAGAGTTCTGAGCTCCGATGCCTACTAACCGG
TTCTGAGAGTTCTGAGCTAATACTAACCGGTTATGC
CTACTAACCGGTTCTGAGAATGCCTACTAACCGGTT
CTGAGAATGCCTACTAACCGATGCCTACTAACCGGT
TCTGAGAGTTCTGAGATGAGAGAATGCCTACTAAC
CGGTTCTGAGAATGCCTACTAACCGATGCCTACTAA
CCGGTTCTGAGAGTTCTGAGCTGAGAA
```

800 letters

next-gen sequence (NGS)

billions of letters

1970's Sanger sequencing

Sequencing cost

cost per megabase

$5,000

Roche 454

$1000

$100

Illumina

$10

$1

Short reads (<600bp)

Illumina

$0.1

$0.05

Illumina

$0.02

cost

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”

Article | Open Access | Published: 25 April 2018

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

RESEARCH ARTICLE

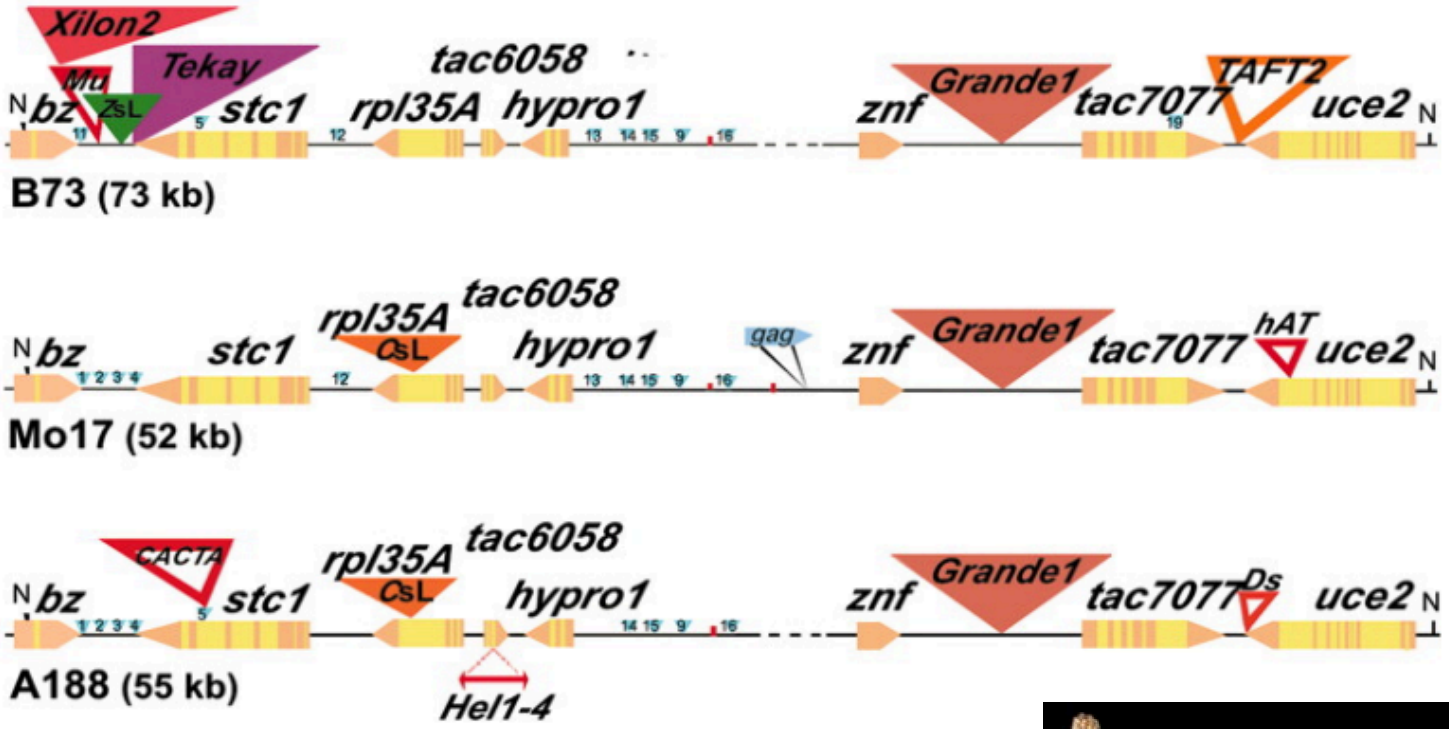
Deep sequencing of 10,000 human genomes

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

52,969

Nucleotide records

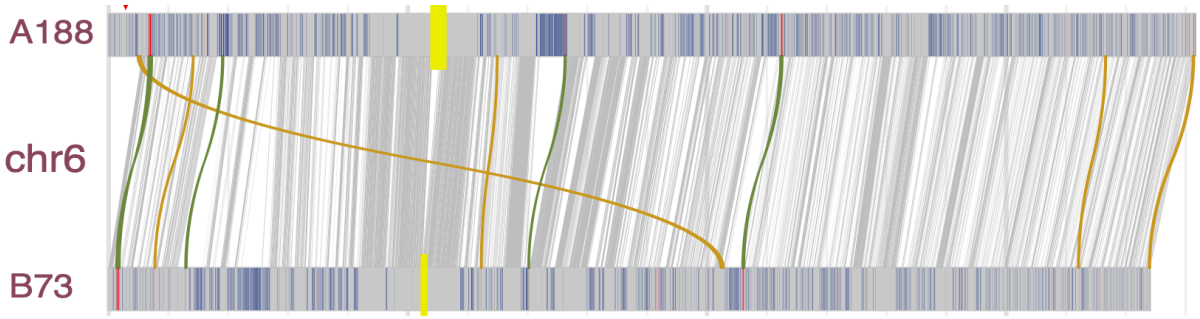
Comparative genomics (I)

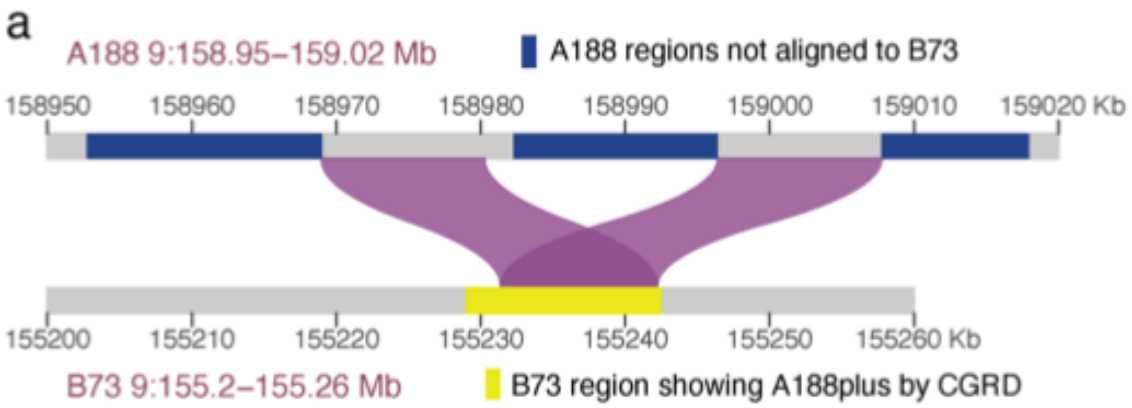


Remarkable variation in maize genome structure at the *bz* locus

PNAS, 2006, 103:17644-49

Comparative genomics (III)





Lin et al., BioRxiv 2020

NGS is changing the way to discover genetic variants

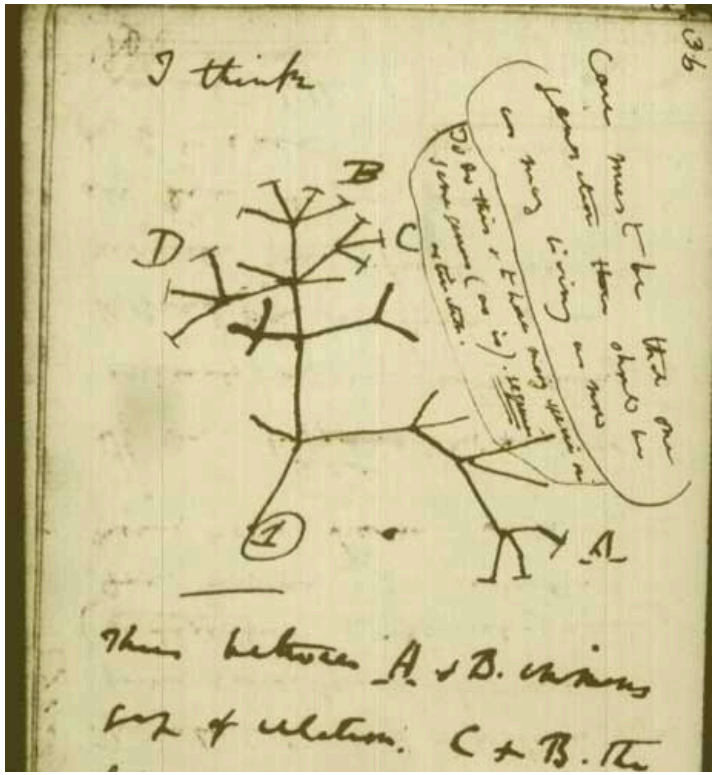
Reference ATCGCTGCCGATCTGCGTCATACGGAATCGTCGGCTTCAG

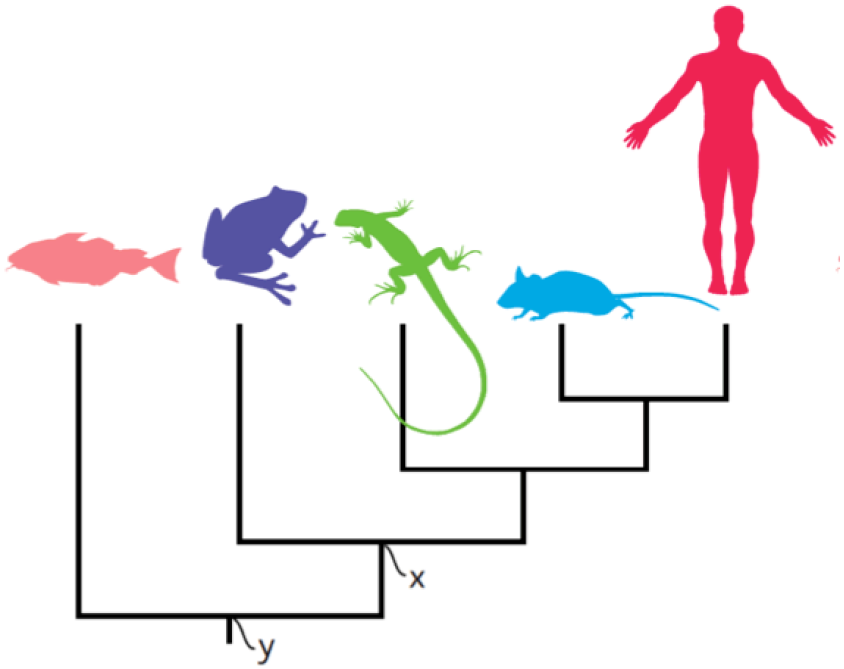
Sequences

ATCGCTGCCGATCTGCGTCATACGGAATCGTCGGCTTCAG
ATCGCTGCCGATCTGCGTGATACGGAATCGTCGGCTTCAG
ATCGCTGCCGATCTGCGTCATACGGAATCGTCGGCTTCAG
ATCGCTGCCGATCTGCGTGATACGGAATCGTCGGCTTCAG
ATCGCTGCCGATCTGCGTGATACGGAATCGTCGGCTTCAG
ATCGCTGCCGATCTGCGTGATACGGAATCGTCGGCTTCAG
ATCGCTGCCGATCTGCGTCATACGGAATCGTCGGCTTCAG
ATCGCTGCCGATCTGCGTCATACGGAATCGTCGGCTTCAG
ATCGCTGCCGATCTGCGTGATACGGAATCGTCGGCTTCAG
ATCGCTGCCGATCTGCGTCATACGGAATCGTCGGCTTCAG

Genotype ----------C/G----------

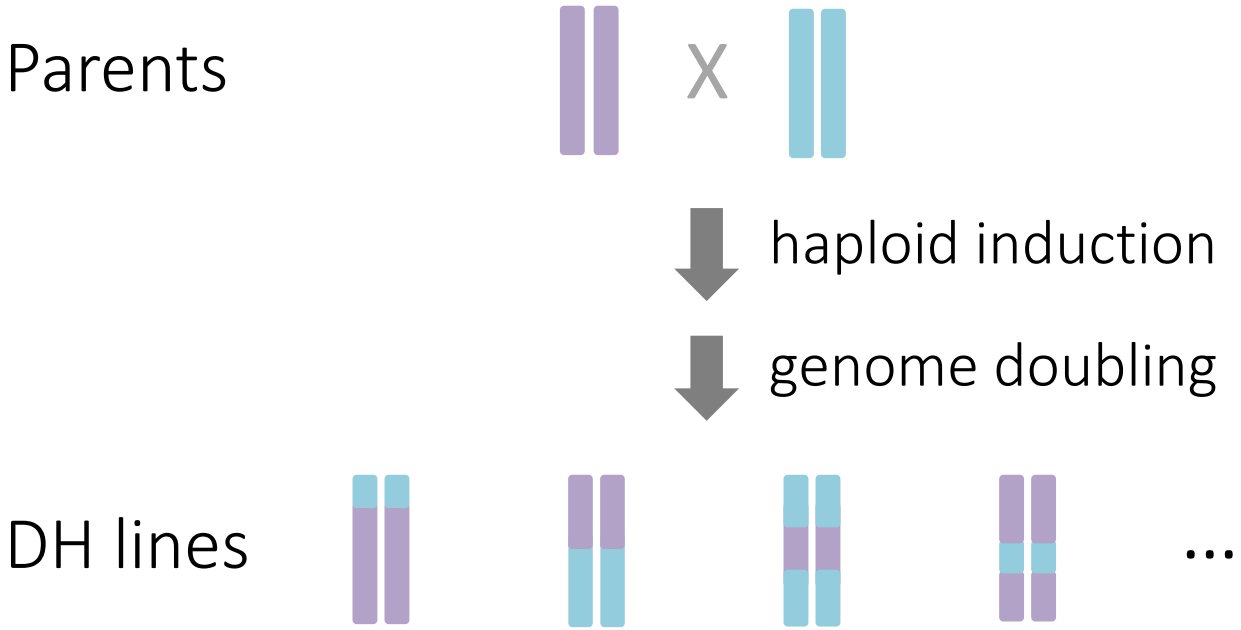
Phylogeny





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

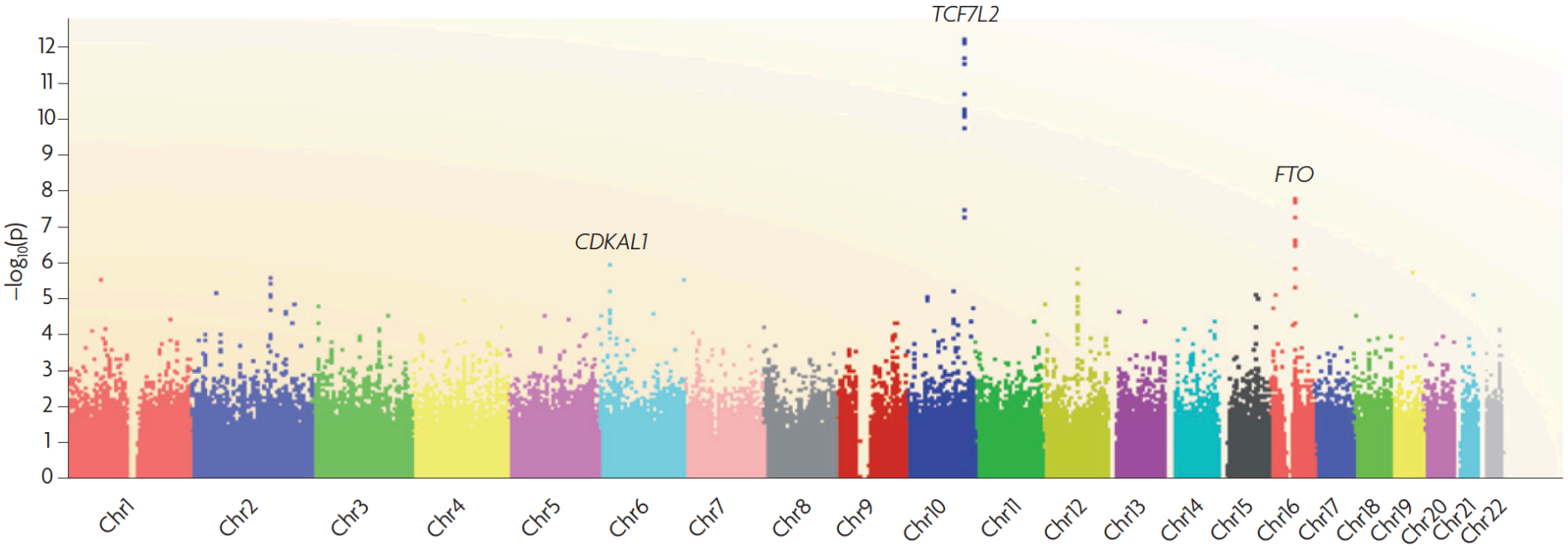
Connect genotype with phenotype (I)



Map QTL using phenotype and genotype data

Connect genotype with phenotype (II)

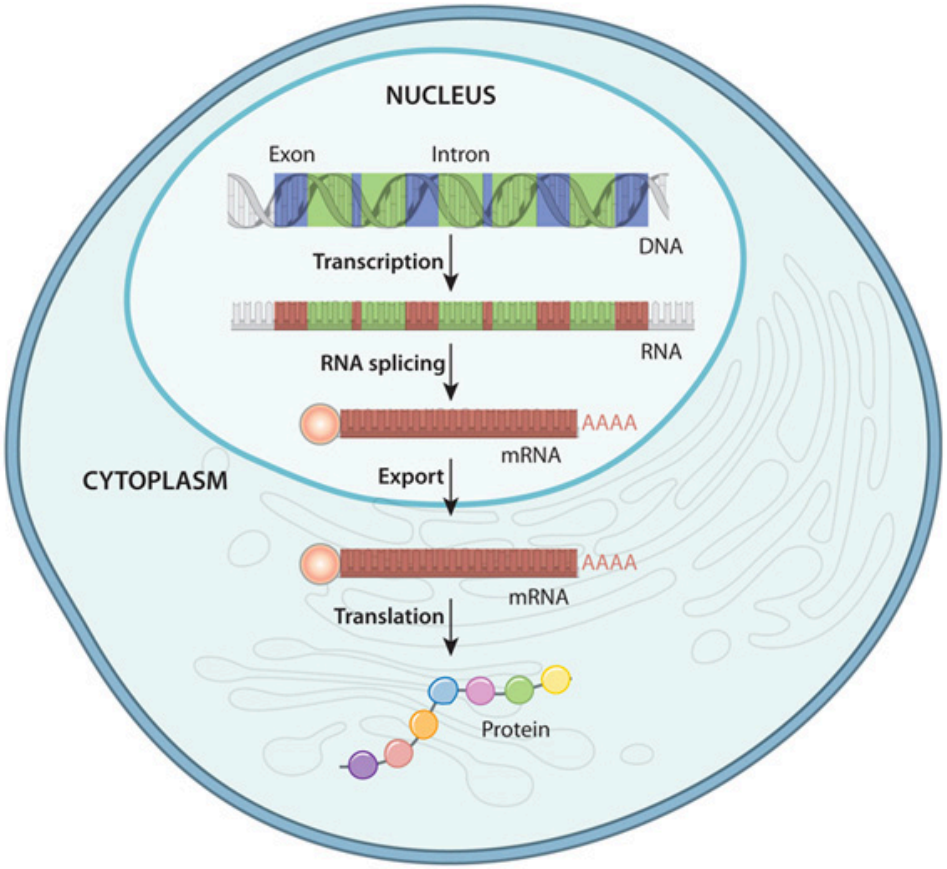
A Manhattan plot from Genome-wide association mapping (GWAS)



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)

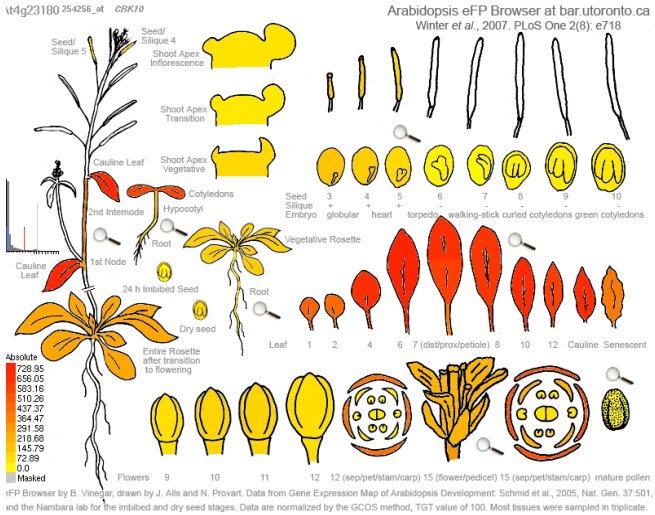
Complexity of transcriptome



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

nature.com/scitable/topicpage/gene-expression-14121669

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?

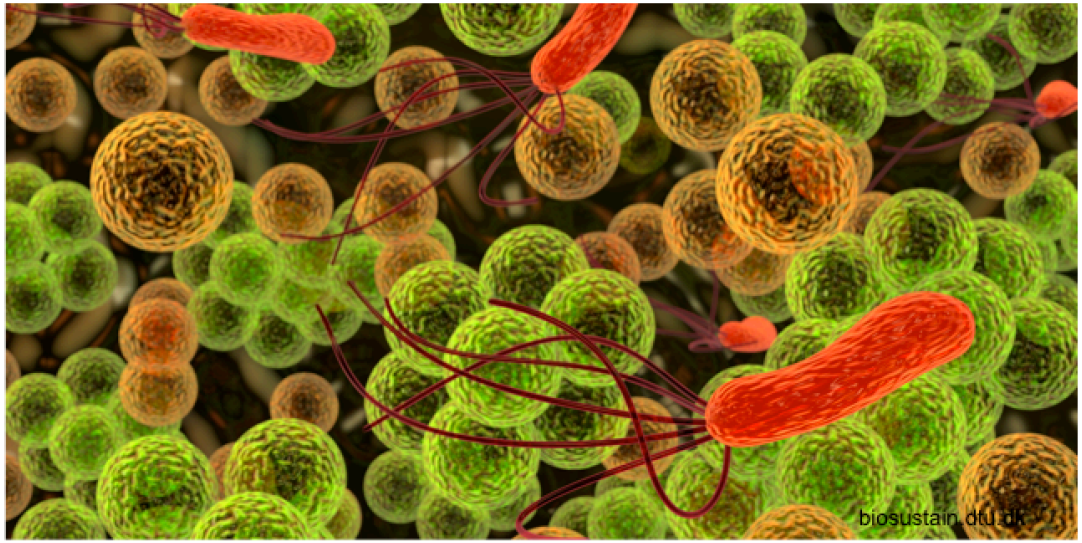
RNA-Seq addresses both questions pretty well

Environmental microbiomes

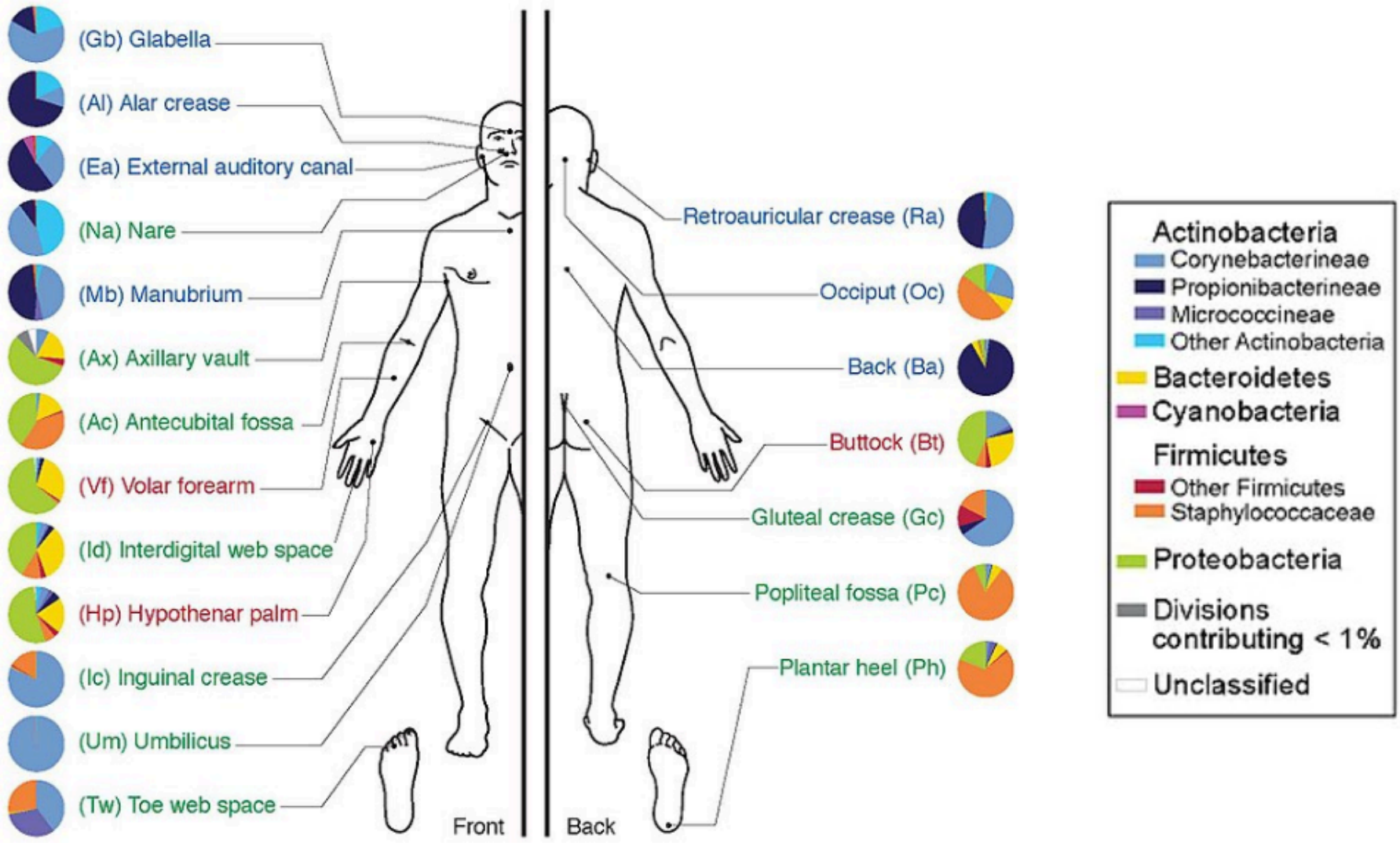
Water

Plants

Soils



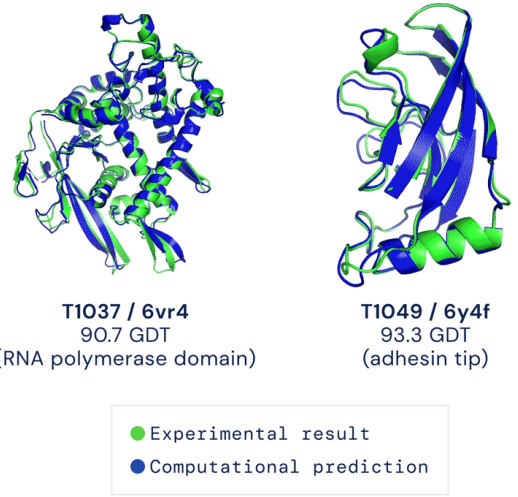
Human microbiomes



The human ***skin*** microbiota (wikipedia)

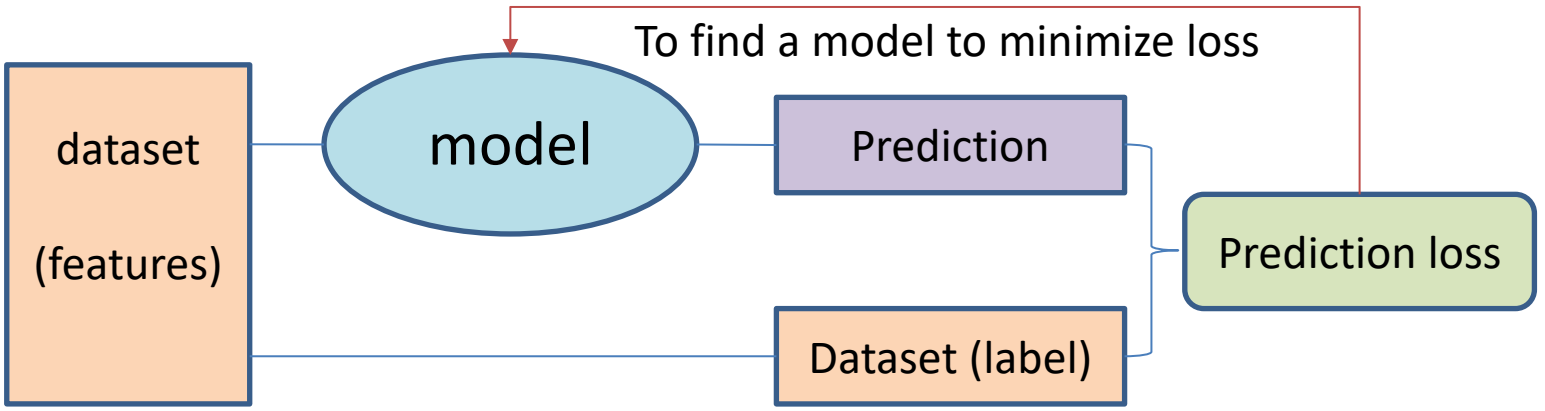
Deep learning

autopilot



AlphaFold

optimization procedure



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

Polling

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)
```

Student Projects

March 25th: 5 min presentation to talk about project plans

Students are expected to design their projects after February.

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

Ideas will be proposed in a 5-minute presentation during the class (March 25th). 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

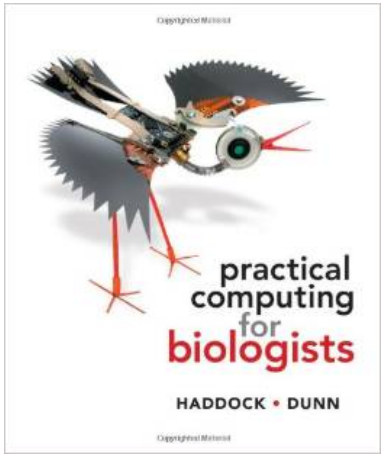
5 mins + 15 mins

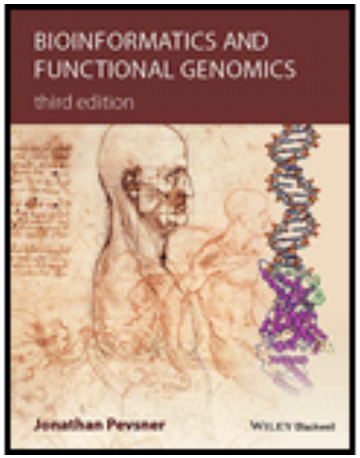
- Two exams

Midterm (March 11th, 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)