

Welcome to Bioinformatics Applications 2019 Spring

Overview

Bioinformatics Applications (PLPTH813)

Sanzhen Liu

1/22/2019

Goal

PLPTH813 will cover **the basic principle** of regular bioinformatics applications and emphasize **the practice of bioinformatics**.

The ultimate 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/2019>

- Course information
- Lecture slide files
- Labs slide files

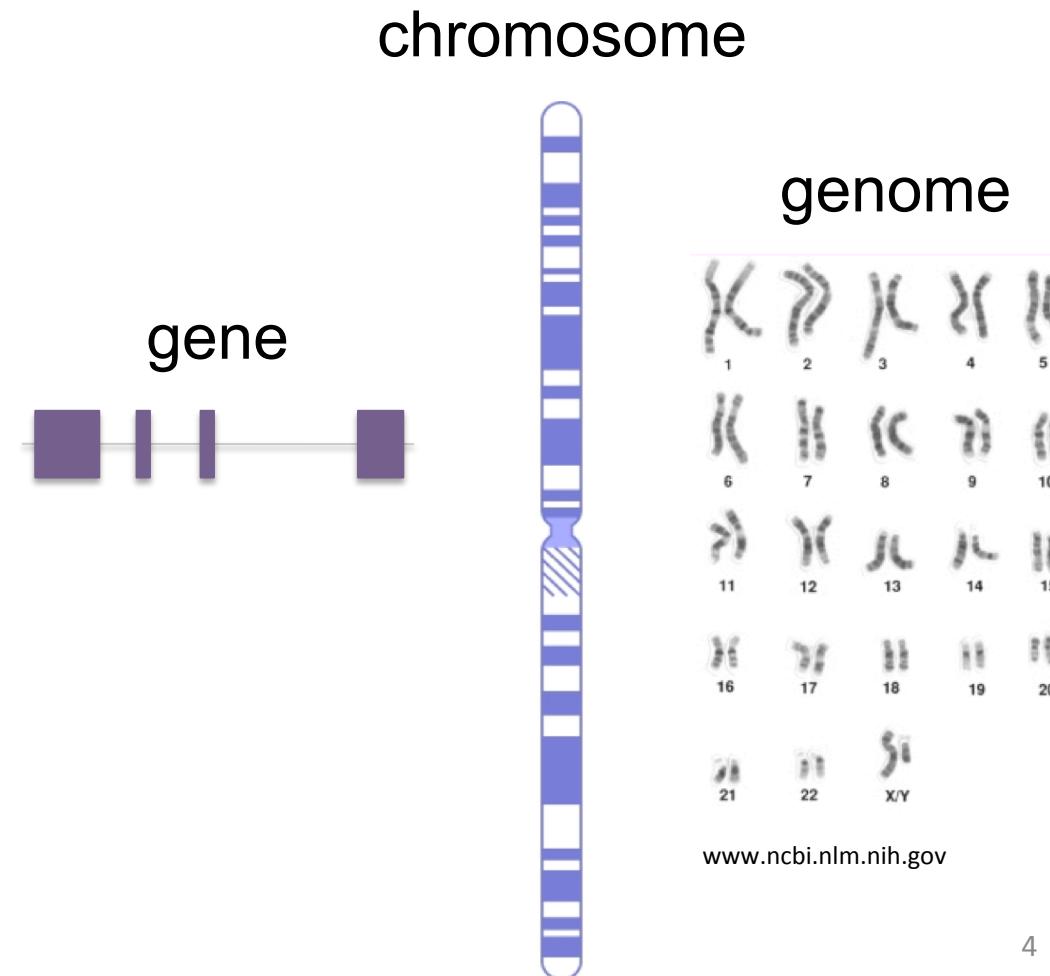
Bioinformatics

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

Sequencing
technology

1980 Nobel Prize

```
1 ccctgaggct tttcgagcgc agctcccaa atccatcca gatttcccgg tcggaggaa
 61 ggaggaccct gggaaaactg cgacgactat cttccccctgg ggcacatggac tcggacgcca
121 gcctgggtgc cagccgcggc tcgtcgccag agcccgatga ccctttctg ccggccggaa
181 gtaaggccgac cagccggcago cccttcactg ggggacccgt gtctctgtcc acccccgatgt
241 actgcggccg ggagctgagc gccgagatgc gcccgcgtat gggctctgcg ggccgcgcata
301 ctggggacaa gtagggggcggc agttggcttc agtcgtctc gtcacagcacc tcgtcgctca
361 cgtcgccgc ggtgtcgctcg tccaccaaga aggacaaagaa gcaaatgaca gagccggagc
421 tgacagacgt ggtgtctcaag atcaacacggc ggcggccgaa ggcacatgcac gacctaaaca
481 tcgcacatggg tgccctccgg gaggctcatgc cgtacggcaca cggcccttcg gtgcgaagc
541 ccgtccatggat cggccggcgtc gcaactacat cctcatgttc accaaactcgc
601 ttggaggagat gaaggactgt gttagcggaga tctacggggg ccacccacgtt ggttccacc
661 cgtcgccctg cggggggctg ggcacactccg cgcctctggc cggcccccacc ggcaccccg
721 cagcagcggc gcaacccggc catacccggc cggtagccca cccatcttg cggcccccgg
781 cccggccggc tgctggccgc gtgcggccggc cgggtgtgtc cagggcgtct ctggccggat
841 cccggctggc gtccggctggc tccatccggc caccggcggc cctactcaag tctccgtctg
901 ctgcgcggc cgcggccggcgt ggggggggggg ggcggccggcag tggggccggc gggggcttcc
961 agcaactgggg cggatgcggc tgccccctgca gcatgtgcga ggtggccggc cgcacaccacc
1021 acgtgtcgcc tatggggggc ggcacgcttc ggcgcctcaat ctcggacggc aagtggccgg
1081 acgtggccggc ggcggcttcg ggcacggggg agccaggggc cggggggaaag cgaggactgg
1141 cctggcgctgg gtcggggcgt tctgtcgccgaa ggagggggcggc aggaccatgg actgggggtg
1201 gggcaatgtgg gggatccggc catctggcga cccaaaggcaat gggggccccc acagacggat
1261 gggggatgtgg gggatgttct ctccgggacc tgatcgagcg ctgtctggct ttaacctgg
1321 ctggtcaggat agacatcggtt ttagaaaaag gtaccggctgt gtgcattccct cactagaact
```



www.ncbi.nlm.nih.gov

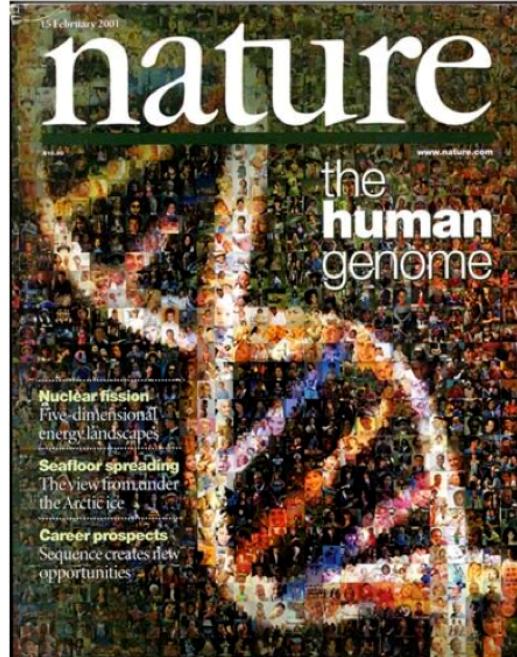
Human Genome Project (HGP)

1st-gen sequence
(Sanger)

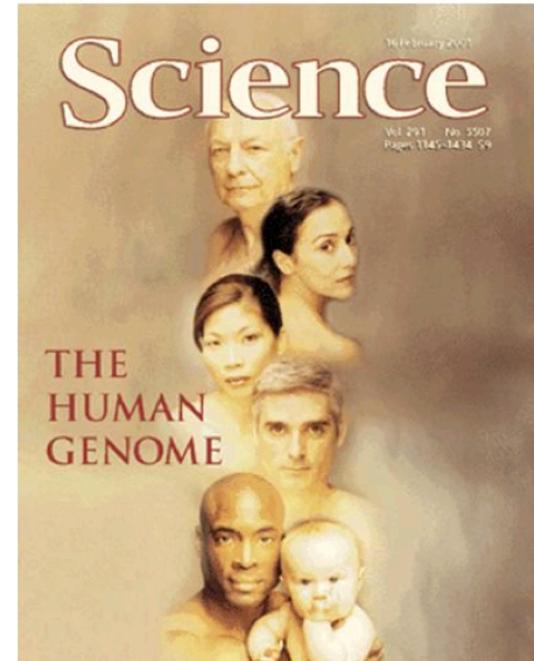
- International Human Genome Sequencing Consortium
Proposed 1985, endorsed in 1988; BAC-by-BAC
- Craig Venter & Celera Genomics:
Founded 1998, finished in 3 years; whole genome shotgun

```
>sample
TGGCGCCTACTAACCGGTTCTGAGAGTCTGAGATG
AGAGAATGCCACTAACCGGTTCTGAGAATGCC
CTAACCGATGCCACTAACCGGTTCTGAGAGTCTG
AGCTGAGATGCCACTAACCGGTTCTGAGAATGCC
ACTAACCGATGCCACTAACCGGTTCTGAGAGTCT
GAGATGAGAGAATGCCACTAACCGGTTCTGAGAA
TGCCACTAACCGATGCCACTAACCGGTTCTGAGA
GTTCTGAGCTATGCCACTAACCGGTTCTGAGAATG
CCTACTAACCGATGCCACTAACCGGTTCTGAGAGT
TCTGAGATGAGAGAATGCCACTAACCGGTTCTGAG
GAATGCCACTAACCGATGCCACTAACCGGTTCTGAG
ATGCCACTAACCGATGCCACTAACCGGTTCTGAG
AGTTCTGAGATGAGAGAATGCCACTAACCGGTTCT
GAGAATGCCACTAACCGATGCCACTAACCGGTTCT
TGAGAGTTCTGAGCTCGATGCCACTAACCGGTTCT
TGAGAGTTCTGAGCTAATACTAACCGGTTATGCC
CTAACCGGTTCTGAGAATGCCACTAACCGGTTCTG
AGAATGCCACTAACCGATGCCACTAACCGGTTCT
GAGAGTTCTGAGATGAGAGAATGCCACTAACCGG
TTCTGAGAATGCCACTAACCGATGCCACTAACCG
GTTCTGAGAGTTCTGAGCTGAGAA
```

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



First draft



DNA sequencing technology

1st-gen sequence
(Sanger)

1980 Nobel Prize

```
>sample
TGCGGCTACTAACGGTCTGAGAGTTCTGAGATG
AGAGAATGCCTACTAACGGTCTGAGAATGCCTA
CTAACCGATGCCTACTAACGGTTCTGAGAGTTCTG
AGCTGAGATGCCTACTAACGGTTCTGAGAATGCCT
ACTAACCGATGCCTACTAACGGTTCTGAGAGTTCTG
GAGATGAGAGAAATGCCTACTAACGGTTCTGAGAA
TGCCTACTAACCGATGCCTACTAACGGTTCTGAGA
GTTCTGAGCTATGCCTACTAACGGTTCTGAGAATG
CCTACTAACCGATGCCTACTAACGGTTCTGAGAGT
TCTGAGATGAGAGAAATGCCTACTAACGGTTCTG
GAATGCTACTAAATGCCTACTAACGGTTCTGAGA
ATGCCCTACTAACCGATGCCTACTAACGGTTCTGAG
AGTTCCTGAGATGAGAGAAATGCCTACTAACGGTTCT
GAGAATGCCTACTAACCGATGCCTACTAACGGTTCT
TGAGAGTTCTGAGCTCGCATGCCTACTAACGGTTCT
TGAGAGTTCTGAGCTAATACTAACGGTTATGCCTA
CTAACGGTTCTGAGAATGCCTACTAACGGTTCTG
AGAATGCCTACTAACCGATGCCTACTAACGGTTCT
GAGAGTTCTGAGATGAGAGAAATGCCTACTAACGG
TTCTGAGAATGCCTACTAACCGATGCCTACTAACGG
GTTCTGAGAGTTCTGAGCTGAGAA
```

next-gen sequence (NGS)



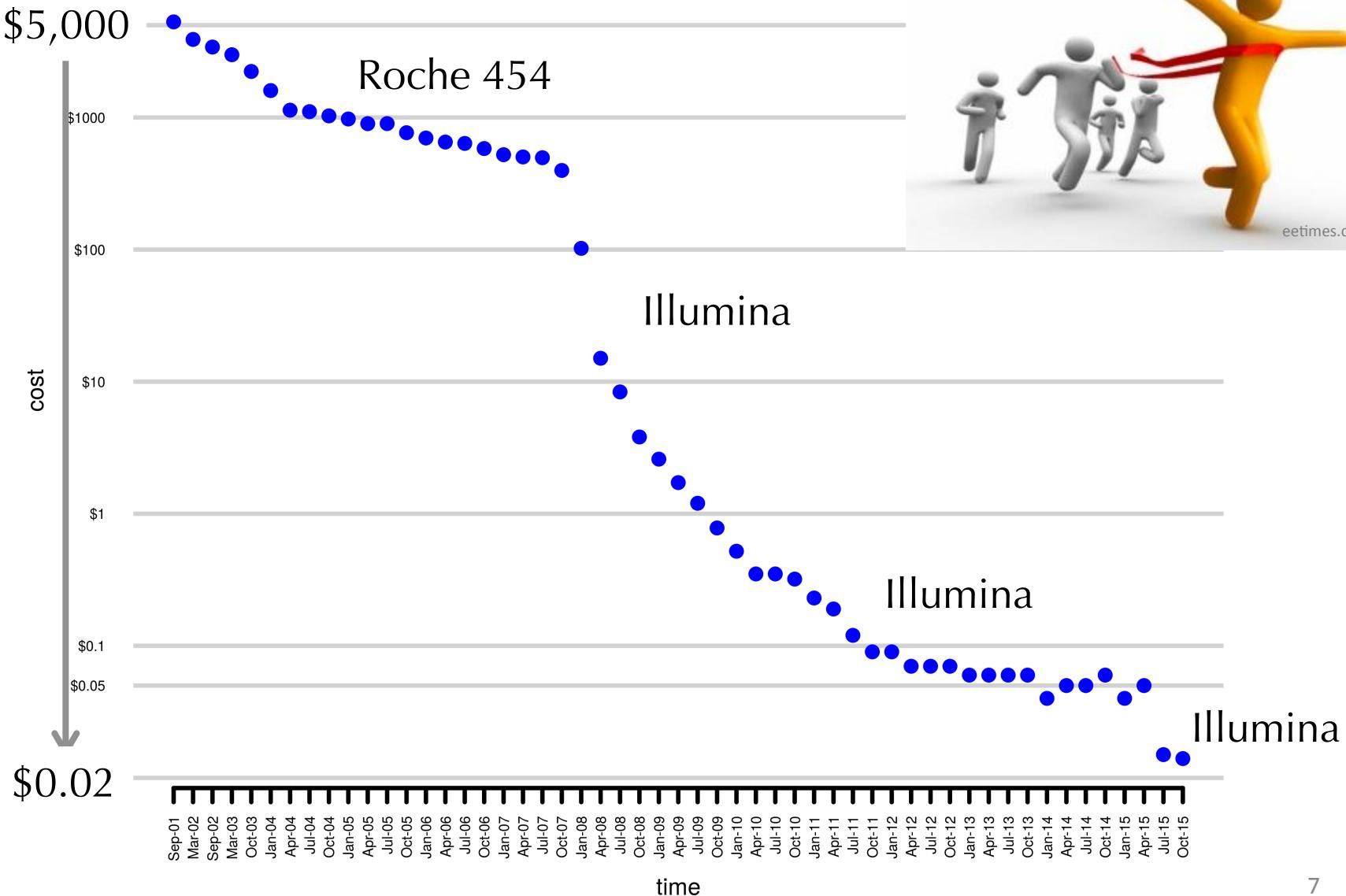
800 letters

billions of letters

Sequencing cost

cost per megabase

1970's Sanger sequencing

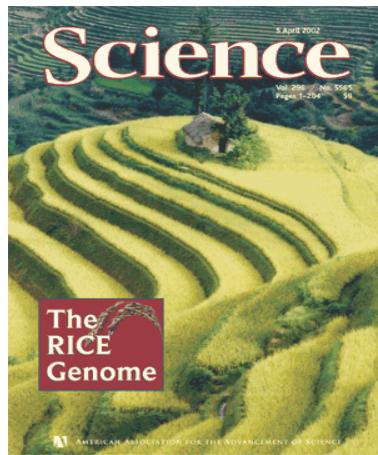


Data source: genome.gov/sequencingcosts

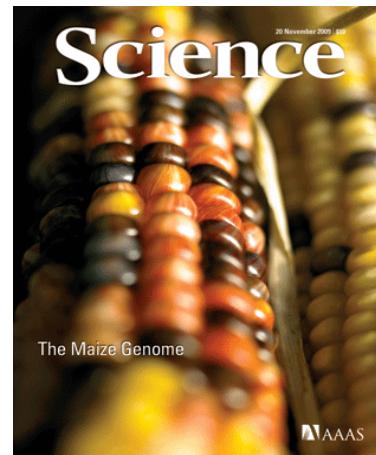
Sequence genomes of model species



2000



2002



2009

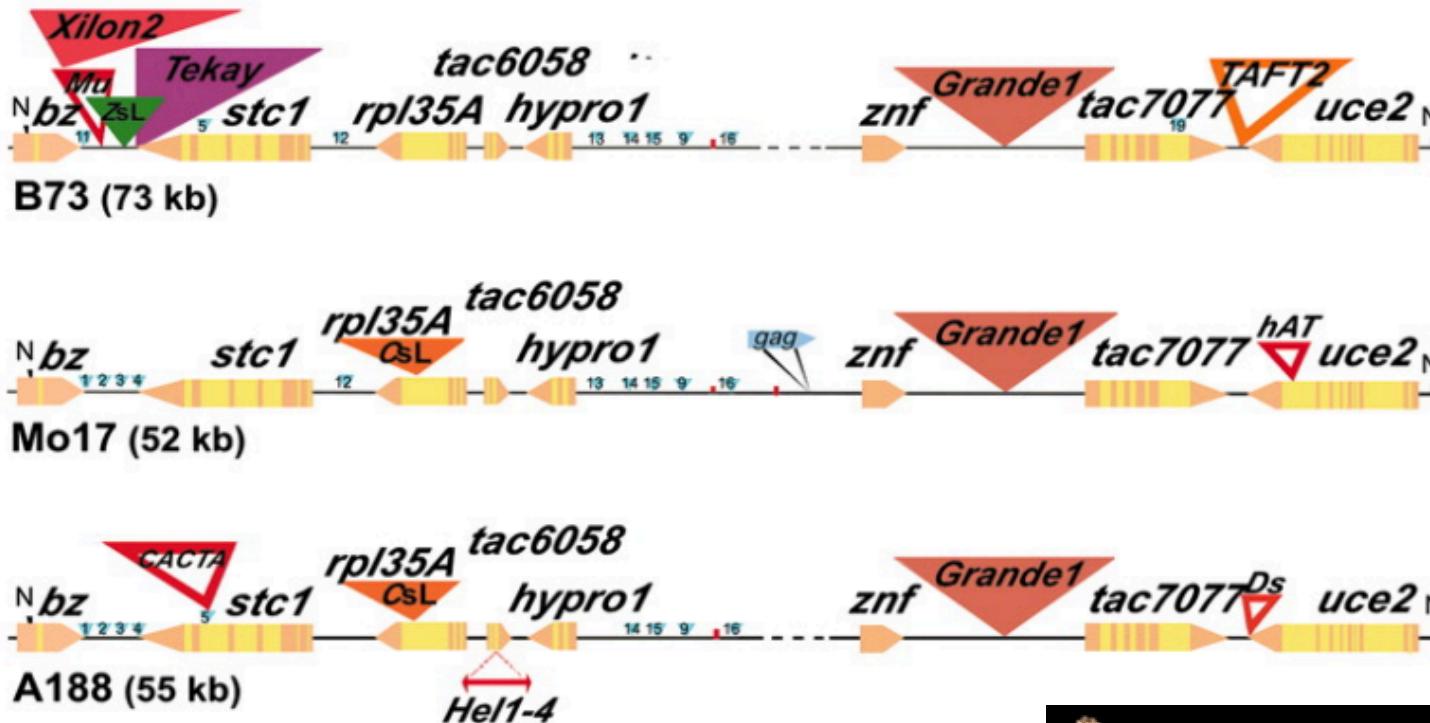


2018



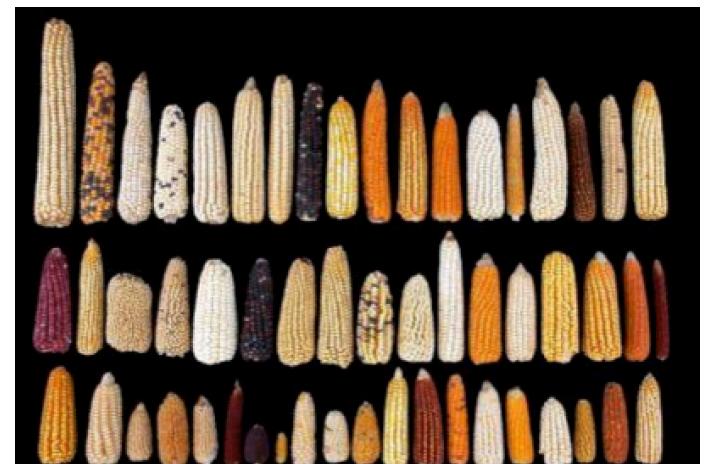
Sequence EVERY species

Comparative genomics

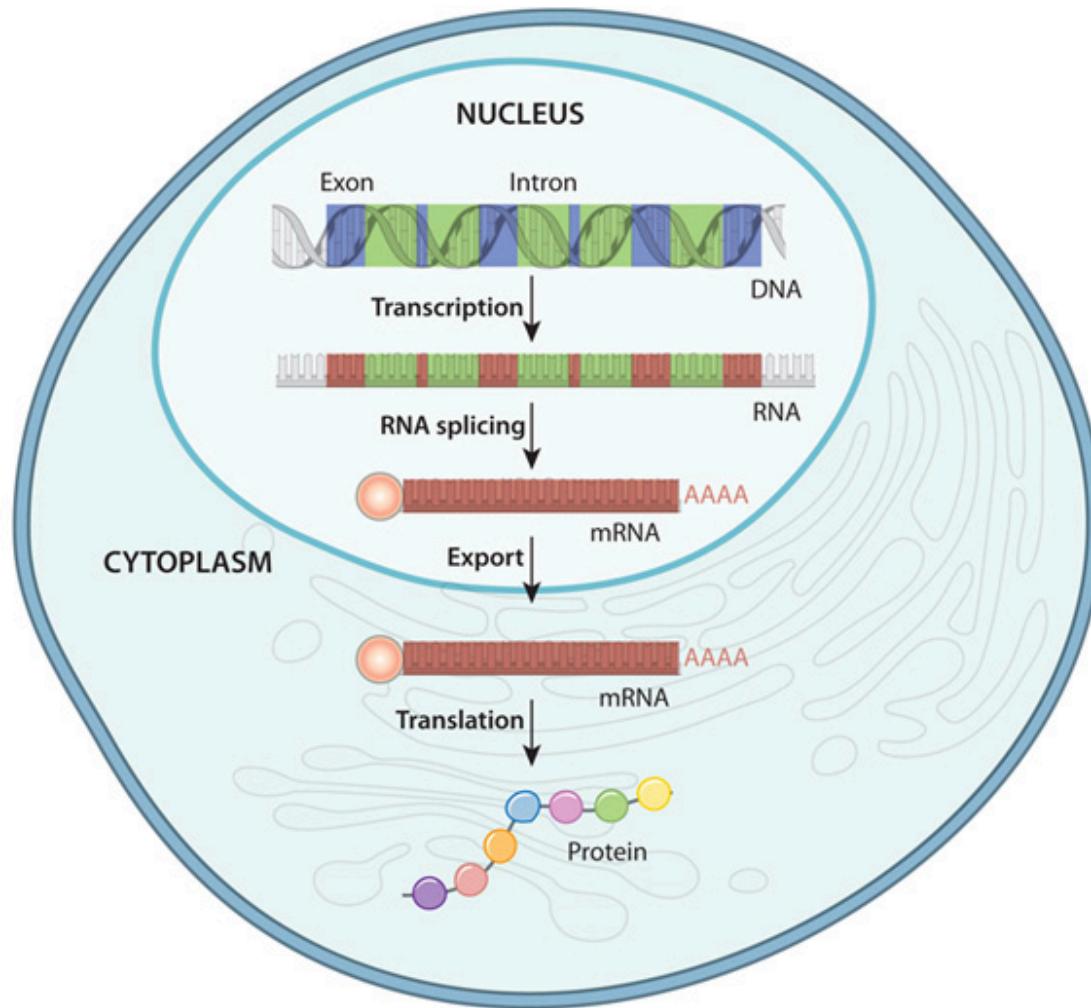


Remarkable variation in maize genome structure at the *bz* locus

PNAS, 2006, 103:17644-49

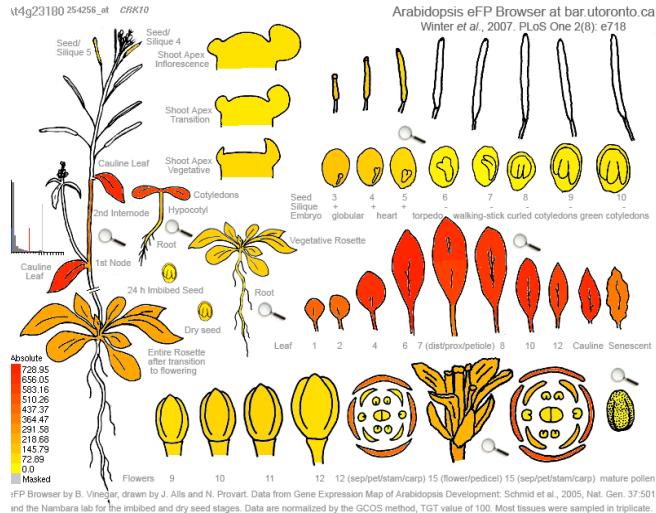


Complexity of transcriptome



DNA to protein in eukaryote

Transcriptome analysis



Expression profiles in different tissues



Response to biotic stress

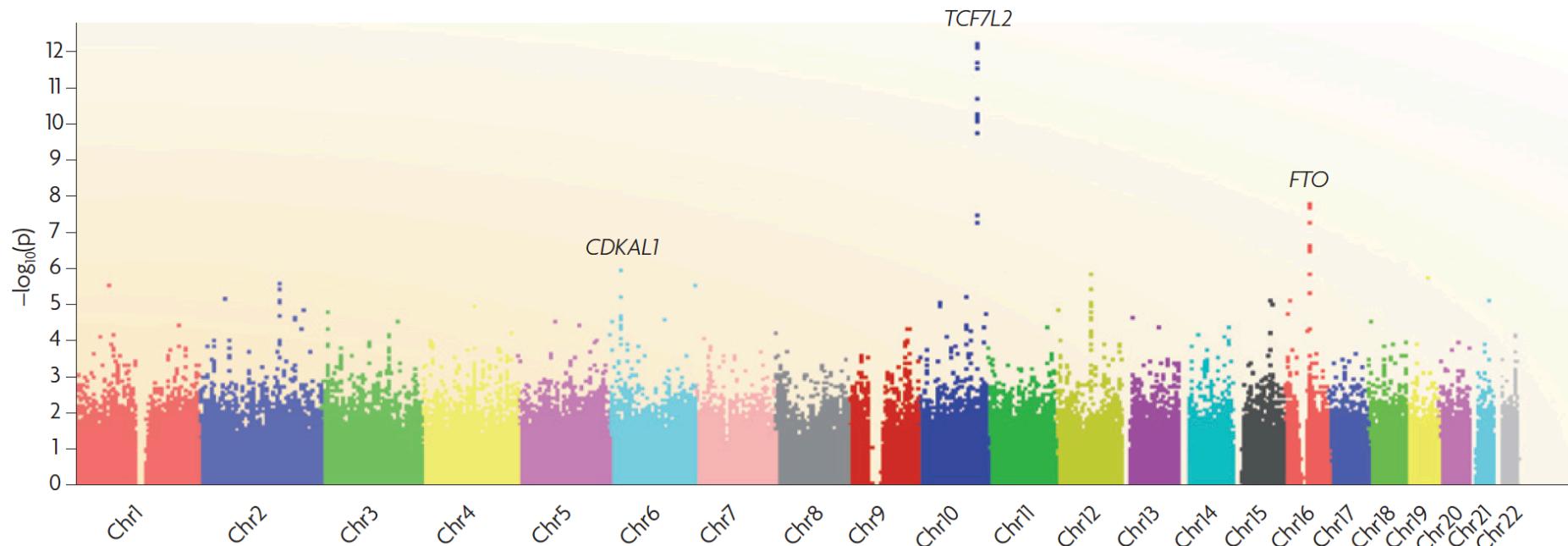
1. What are sequences of transcripts?
2. What is the expression level of each transcript?

RNA-Seq addresses both questions pretty well

NGS is changing the way to discover genetic variants

Reference	ATCGCTGCCGATCTGCGTCATACGGAATCGTCGGCTTCAG
Sequences	ATCGCTGCCGATCTGCGTCATACGGAATCGTCGGCTTCAG
	ATCGCTGCCGATCTGCGTGATACGGAATCGTCGGCTTCAG
	ATCGCTGCCGATCTGCGTCATACGGAATCGTCGGCTTCAG
	ATCGCTGCCGATCTGCGTGATACGGAATCGTCGGCTTCAG
	ATCGCTGCCGATCTGCGTGATACGGAATCGTCGGCTTCAG
	ATCGCTGCCGATCTGCGTGATACGGAATCGTCGGCTTCAG
	ATCGCTGCCGATCTGCGTCATACGGAATCGTCGGCTTCAG
Genotype	-----C/G-----

Genome-wide association mapping



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

Environmental microbiomes



Water



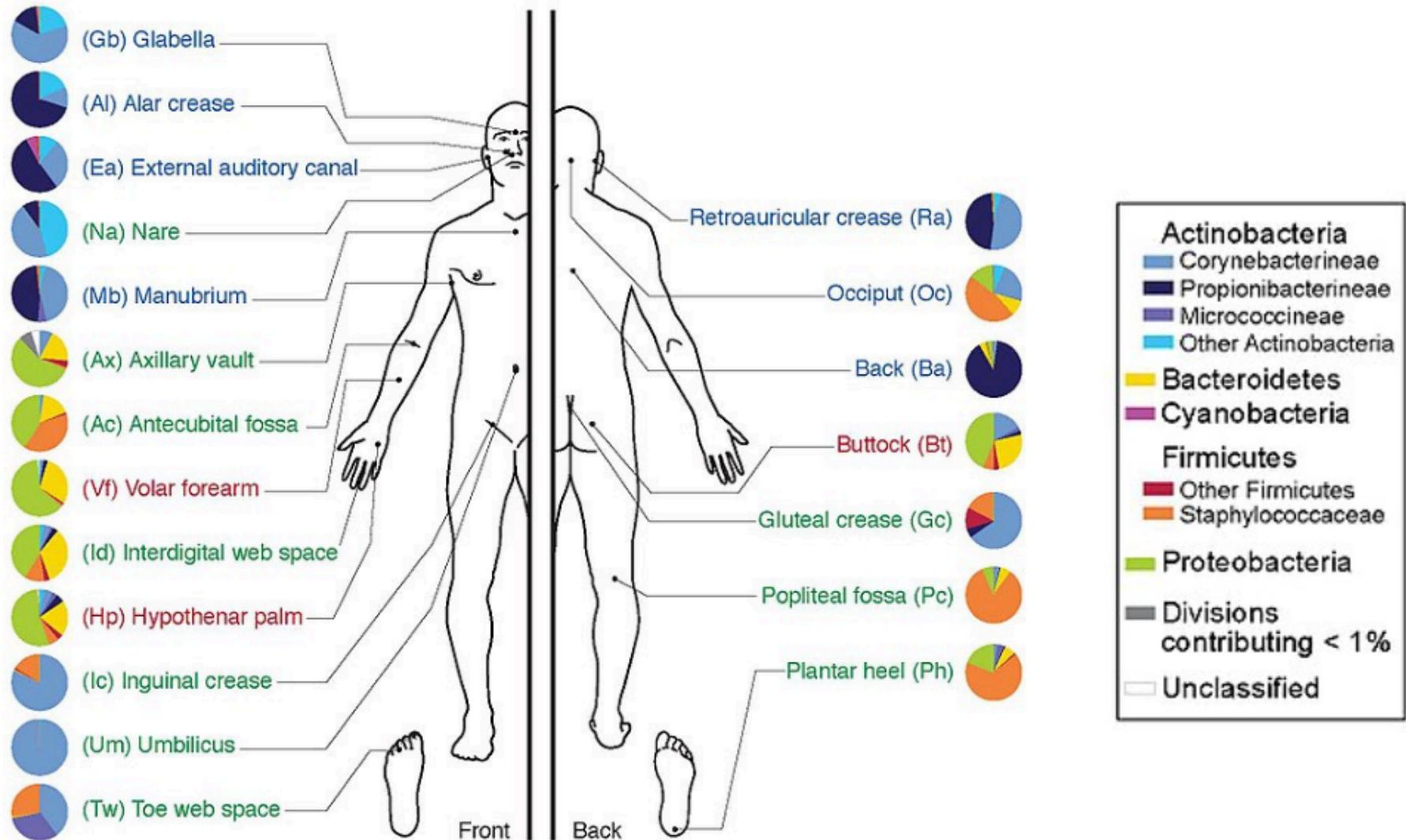
Plants



Soils



Human microbiomes



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)

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. Construction of a genetic map
8. QTL and GWAS
9. Genome assembly
10. Comparative genomics
11. Metagenomics
12. RNA-Seq

Grading and schedule

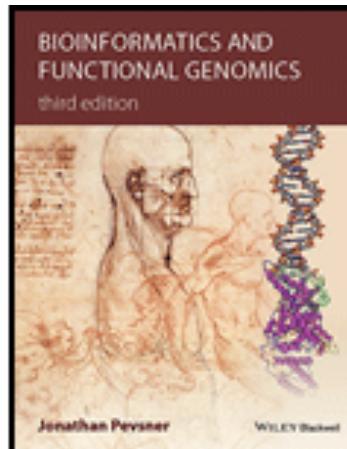
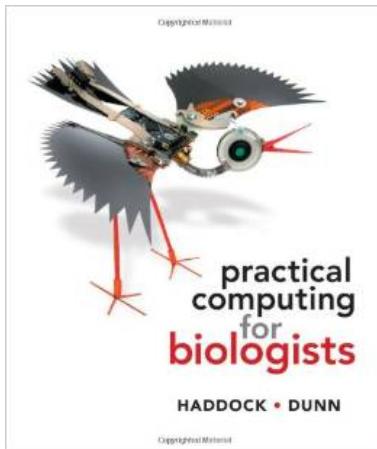
- **Grading**

Class participation 10%, Homework 35%, Midterm Exam 20%,
Oral presentation 10%, Final Exam 25%

- Homework: 6+ times
- Oral presentation
 - 1. Genome assembly and annotation
 - 2. Comparative genomics and pan-genomics
 - 3. Genetic mapping (QTL, GWAS, and post-GWAS)
 - 4. RNA-Seq (analysis methods, small RNA, non-coding RNA, circle RNA)
 - 5. metagenomics
 - 6. new genome technologies and new bioinformatics tools
- Two exams (midterm and final)

References

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



Schedule

- Lecture: 10:30am-11:20pm, Tuesday, Thursday
- Lab: 12:30-2:00pm, Thursday
(typically each lab finished in 1:30 hours)
- Office hours: 12:00-1:00pm Tuesday