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/PLPTH813Bioinformatis/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

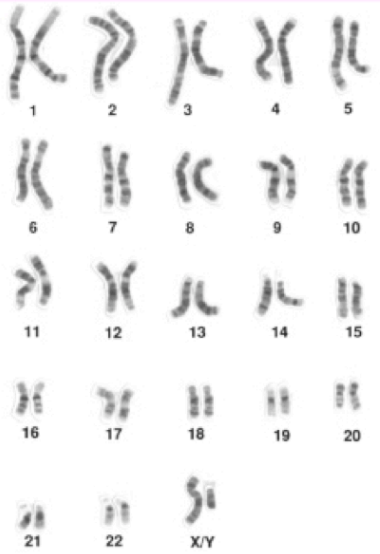
1980 Nobel Prize

```
   1 ccctgaggct tttcggagcg agctcctcaa atcgcatcca gattttcggg tccgagggaa
  61 ggaggaccct gcgaaagctg cgacgactat cttcccctgg ggccatggac tcggacgcca
 121 gcctggtgtc cagccgcccg tcgtcgccag agcccgatga ccttttttctg 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
AGCTGAGATGCCTACTAACCGGTTCTGAGAATGCCT
ACTAACCGATGCCTACTAACCGGTTCTGAGAGTTCT
GAGATGAGAGAATGCCTACTAACCGGTTCTGAGAA
TGCCTACTAACCGATGCCTACTAACCGGTTCTGAGA
GTTCTGAGCTATGCCTACTAACCGGTTCTGAGAATG
CCTACTAACCGATGCCTACTAACCGGTTCTGAGAGT
TCTGAGATGAGAGAATGCCTACTAACCGGTTCTGA
GAATGCCTACTAAATGCCTACTAACCGGTTCTGAGA
ATGCCTACTAACCGATGCCTACTAACCGGTTCTGAG
AGTTCTGAGATGAGAGAATGCCTACTAACCGGTTCT
GAGAATGCCTACTAACCGATGCCTACTAACCGGTTC
TGAGAGTTCTGAGCTCCGATGCCTACTAACCGGTTC
TGAGAGTTCTGAGCTAATACTAACCGGTTATGCCTA
CTAACCGGTTCTGAGAATGCCTACTAACCGGTTCTG
AGAATGCCTACTAACCGATGCCTACTAACCGGTTCT
GAGAGTTCTGAGATGAGAGAATGCCTACTAACCGG
TTCTGAGAATGCCTACTAACCGATGCCTACTAACCG
GTTCTGAGAGTTCTGAGCTGAGAA
```

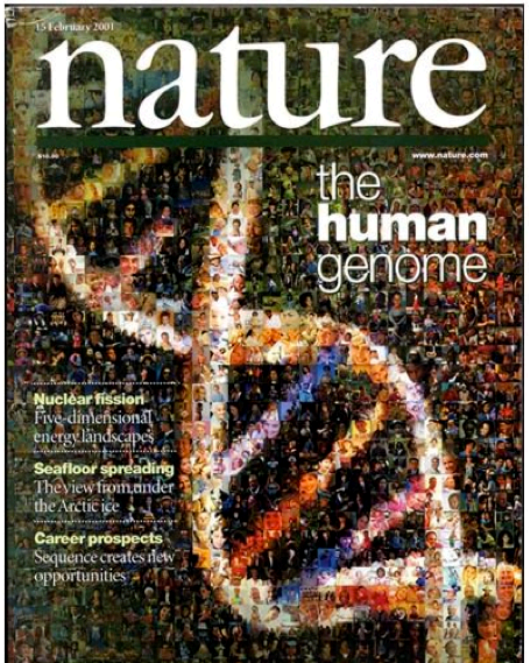
- International Human Genome Sequencing Consortium

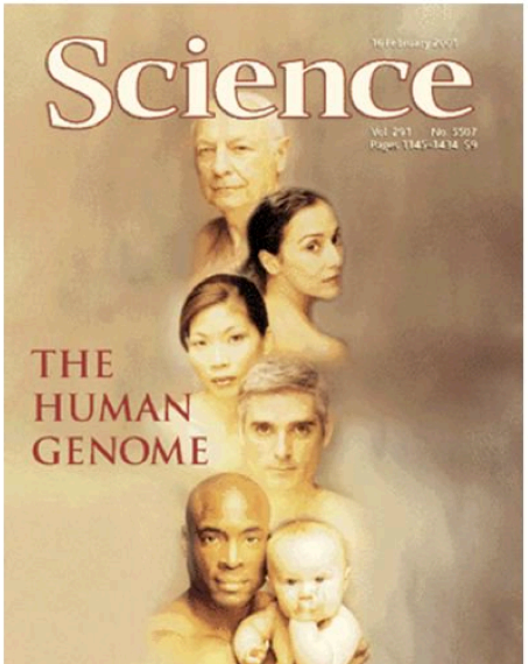
Proposed 1985, endorsed in 1988; BAC-by-BAC

- Craig Venter & Celera Genomics:

Founded 1998, finished in 3 years; whole genome shotgun

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





First draft

DNA sequencing technology

1st-gen sequence (Sanger)

1980 Nobel Prize

```
>sample
TGCCGCCTACTAACCGGTTCTGAGAGTTCTGAGATG
AGAGAATGCCTACTAACCGGTTCTGAGAATGCCTA
CTAACCGATGCCTACTAACCGGTTCTGAGAGTTCTG
AGCTGAGATGCCTACTAACCGGTTCTGAGAATGCCT
ACTAACCGATGCCTACTAACCGGTTCTGAGAGTTCT
GAGATGAGAGAATGCCTACTAACCGGTTCTGAGAA
TGCCTACTAACCGATGCCTACTAACCGGTTCTGAGA
GTTCTGAGCTATGCCTACTAACCGGTTCTGAGAATG
CCTACTAACCGATGCCTACTAACCGGTTCTGAGAGT
TCTGAGATGAGAGAATGCCTACTAACCGGTTCTGA
GAATGCCTACTAAATGCCTACTAACCGGTTCTGAGA
ATGCCTACTAACCGATGCCTACTAACCGGTTCTGAG
AGTTCTGAGATGAGAGAATGCCTACTAACCGGTTCT
GAGAATGCCTACTAACCGATGCCTACTAACCGGTTC
TGAGAGTTCTGAGCTCCGATGCCTACTAACCGGTTC
TGAGAGTTCTGAGCTAATACTAACCGGTTATGCCTA
CTAACCGGTTCTGAGAATGCCTACTAACCGGTTCTG
AGAATGCCTACTAACCGATGCCTACTAACCGGTTCT
GAGAGTTCTGAGATGAGAGAATGCCTACTAACCGG
TTCTGAGAATGCCTACTAACCGATGCCTACTAACCG
GTTCTGAGAGTTCTGAGCTGAGAA
```

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

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

Sequence genomes of model species

2000

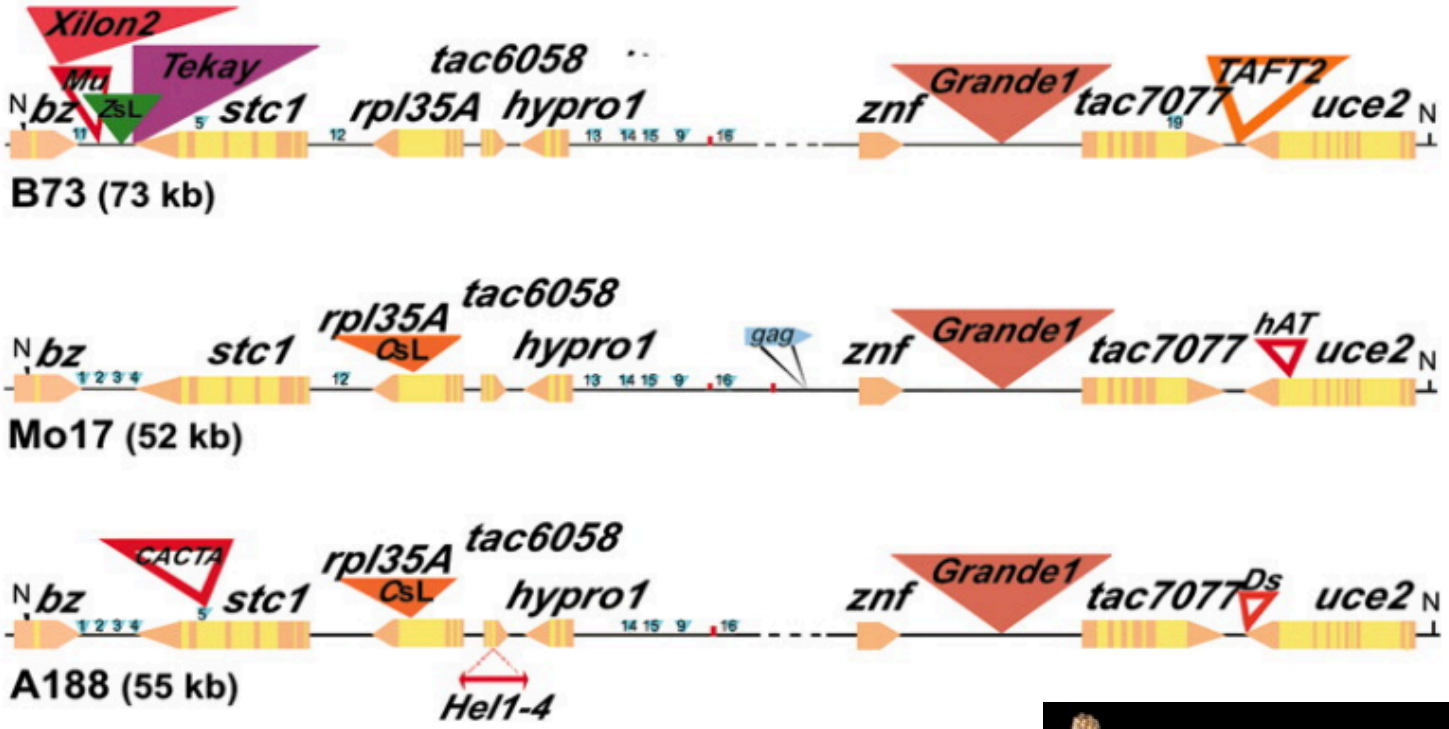
2002

2009

2018

Sequence EVERY species

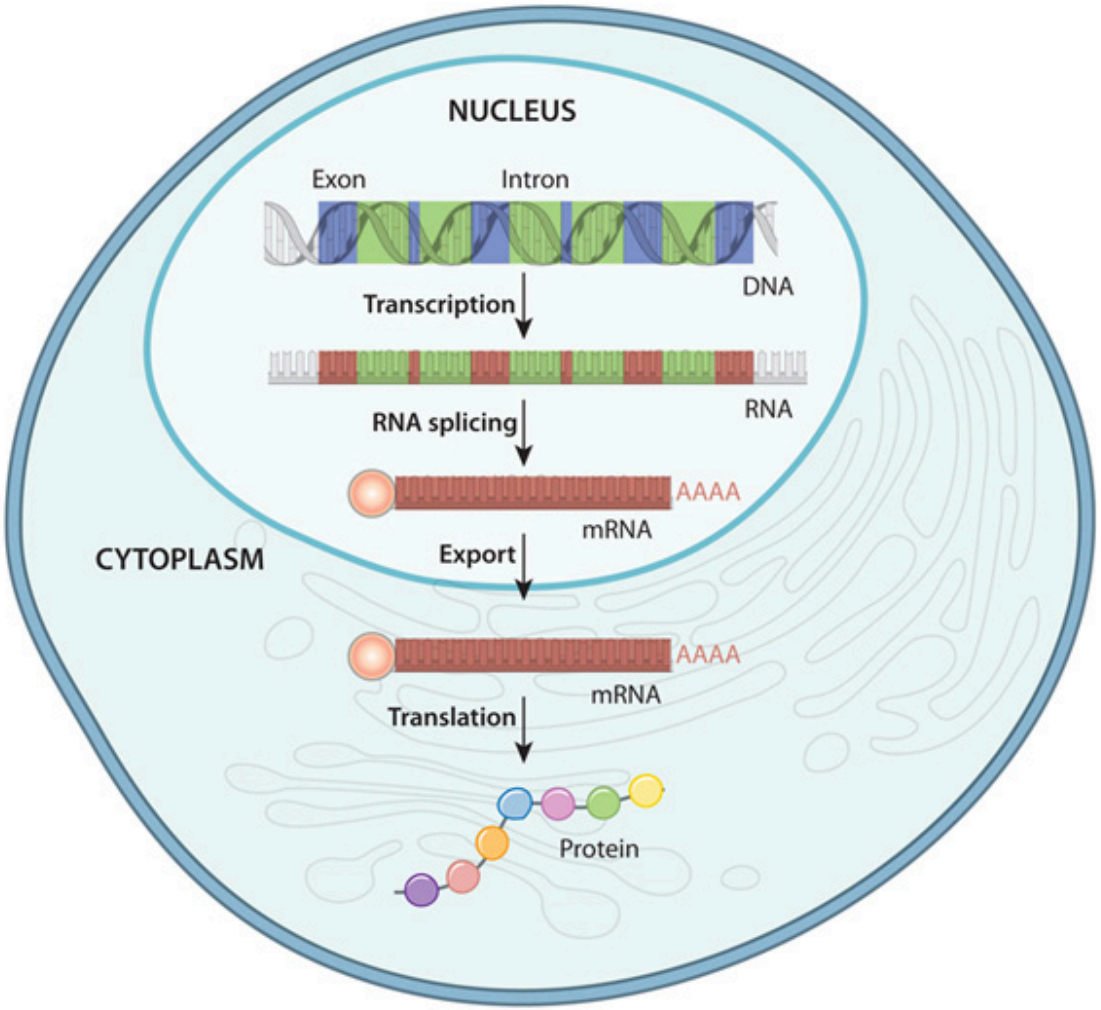
Comparative genomics



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

PNAS, 2006, 103:17644-49

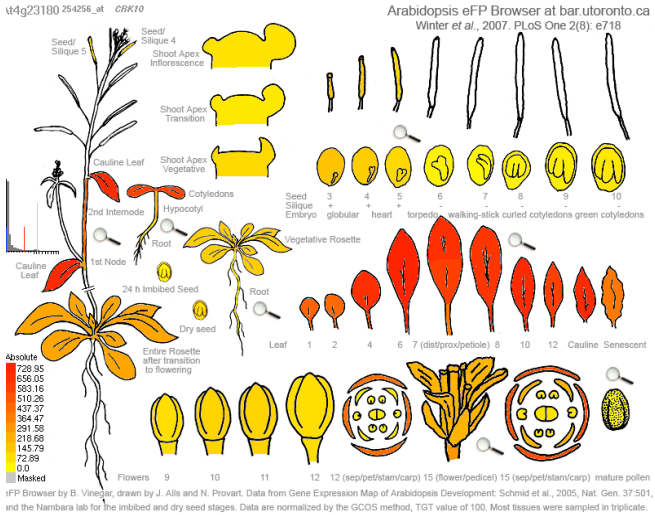
Complexity of transcriptome



DNA to protein in eukaryote

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

NGS is changing the way to discover genetic variants

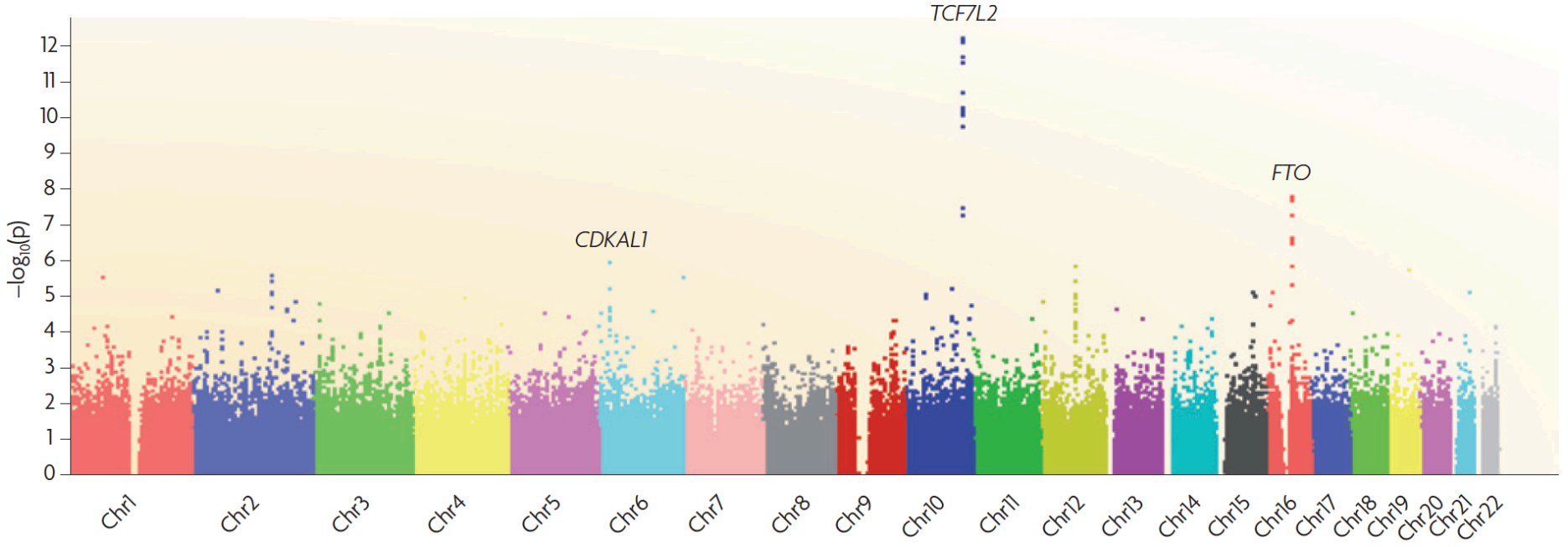
Reference ATCGCTGCCGATCTGCGTCATACGGAATCGTCGGCTTCAG

Sequences

ATCGCTGCCGATCTGCGTCATACGGAATCGTCGGCTTCAG
ATCGCTGCCGATCTGCGTGATACGGAATCGTCGGCTTCAG
ATCGCTGCCGATCTGCGTCATACGGAATCGTCGGCTTCAG
ATCGCTGCCGATCTGCGTGATACGGAATCGTCGGCTTCAG
ATCGCTGCCGATCTGCGTGATACGGAATCGTCGGCTTCAG
ATCGCTGCCGATCTGCGTGATACGGAATCGTCGGCTTCAG
ATCGCTGCCGATCTGCGTCATACGGAATCGTCGGCTTCAG
ATCGCTGCCGATCTGCGTCATACGGAATCGTCGGCTTCAG
ATCGCTGCCGATCTGCGTGATACGGAATCGTCGGCTTCAG
ATCGCTGCCGATCTGCGTCATACGGAATCGTCGGCTTCAG

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

Genome-wide association mapping



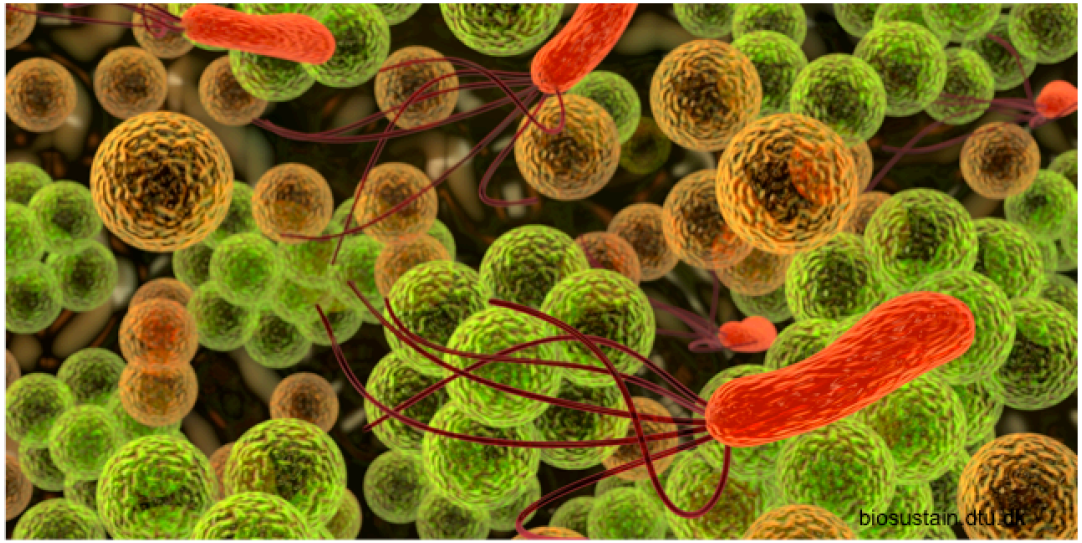
McCarthy et al., Nature Review Genetics, 2008: 9:356-369

Environmental microbiomes

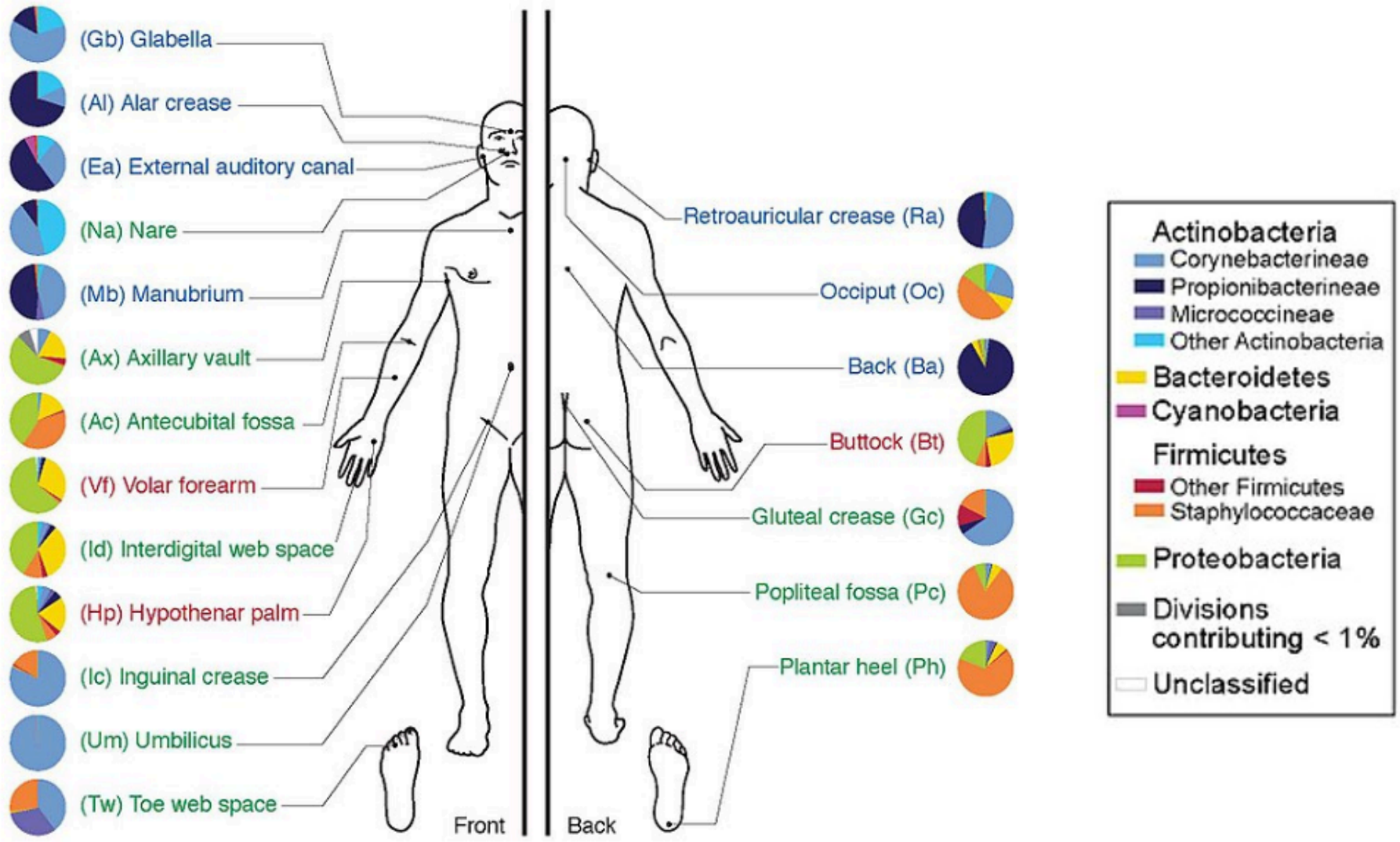
Water

Plants

Soils



Human microbiomes



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

Reason for command-lines analyses

- To perform **efficient** and **reproducible** data analyses
- To use advanced tools in research projects (most genomic software 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

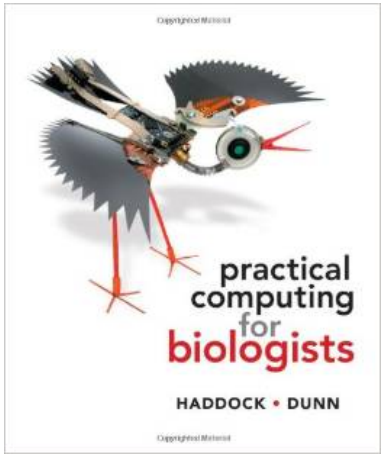
- **Grading**

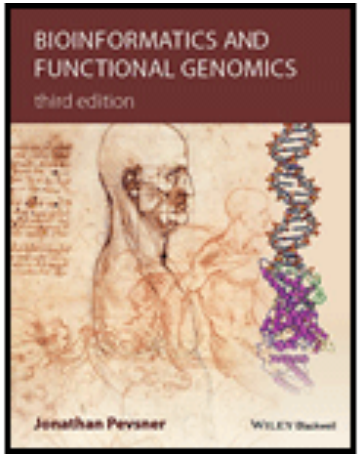
Class participation 10%, Homework 35%, Midterm Exam 20%, Oral presentation 10%, Final Exam 25%

- 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: 11:00am-11:50pm, Tuesday, Thursday
- Lab: 1:00-2:30pm, Thursday

(typically each lab finished in 1:30 hours)

- Office hours: 12:00-1:00pm Tuesday