Computational Methods in Metagenomics and Microbiome Research

0368-3116-01

Prof. Elhanan Borenstein
School of Computer Science
Semester B, 2019

TEL AVIV UNIVERSITY אוניברסיטת תל אביב

Outline

A Few Updates about Seminar Logistics

“Metagenomics”

Genomes, Genes, and the Central Dogma

Real Metagenomics

Computational Challenges in Metagenomics

Outline

A Few Updates about Seminar Logistics

"Metagenomics"

Genomes, Genes, and the Central Dogma

Real Metagenomics

Computational Challenges in Metagenomics

Seminar Format

- ~15 students registered
- 1 paper per student, 1-2 students/papers per class
- Class structure:
  - 1:10 Start talk 1(35 minutes!!)
  - 1:45-2:00 Feedback/Discussion
  - 2:00-2:10 Break
  - 2:10 Start talk 2 (35 minutes!!)
  - 2:45-3:00 Feedback/Discussion
- Paper selection:
  - List posted: Sunday, March 10, 9:00
  - Please select your paper by: Monday, 17:00
  - First talk: March 20

Presentations

- In your presentation:
  - Emphasize the main task of the paper
  - Cover required background
  - Focus more (but not only) on methods (may need to dig)
  - **Choose wisely what to cover and how deeply**
  - Summarize briefly
  - Add something original
- Remember the dos and don’ts of a good talk and of slide design
- Keep us interested

Other Tasks

- As a speaker
  - Come early (or on a different day) to make sure your presentation works
  - Print slides for everybody
  - Send me slides to post online
  - Prepare discussion points (2-3 points, NOT in paper)
- As a listener
  - Read/skim the papers before class
  - Listen, take notes, engage
  - Participate in feedback/discussion
  - Learn and have fun

Outline

A Few Updates about Seminar Logistics

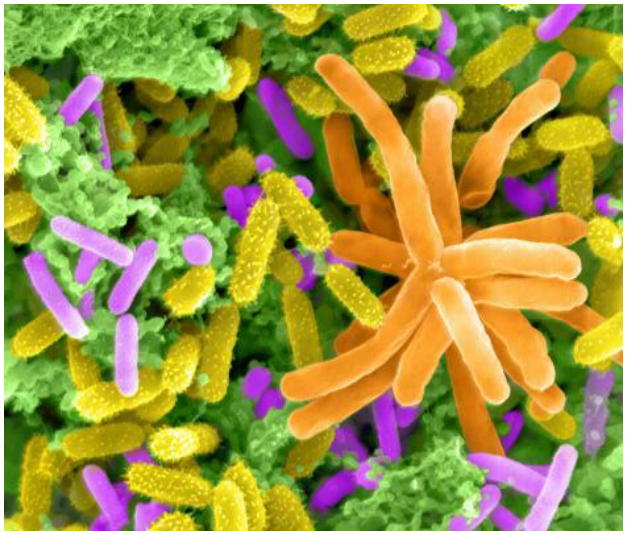
“Metagenomics”

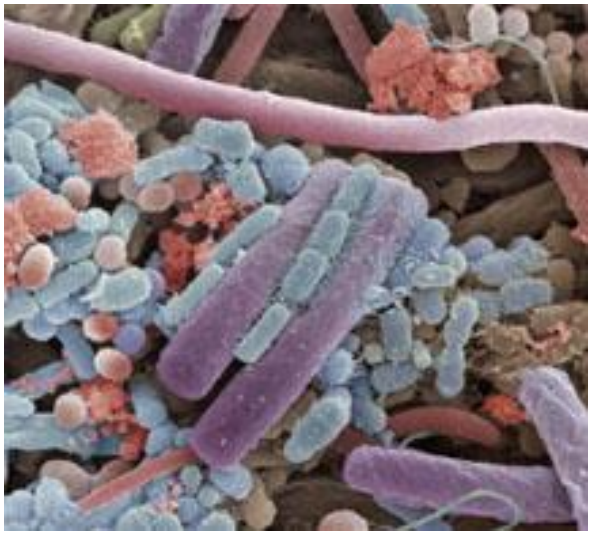
Genomes, Genes, and the Central Dogma

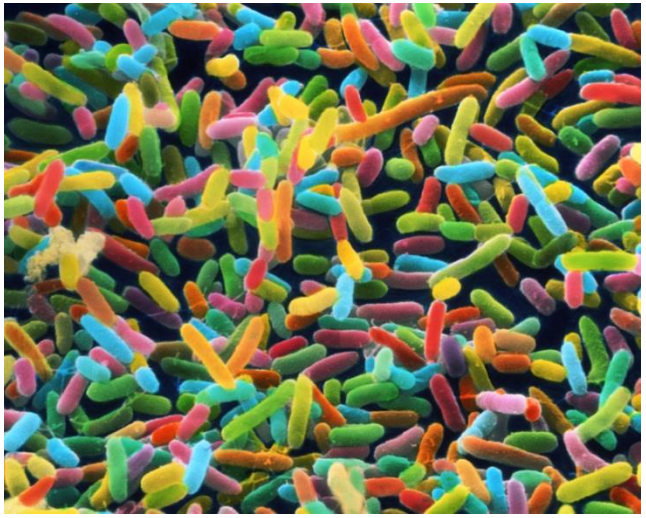
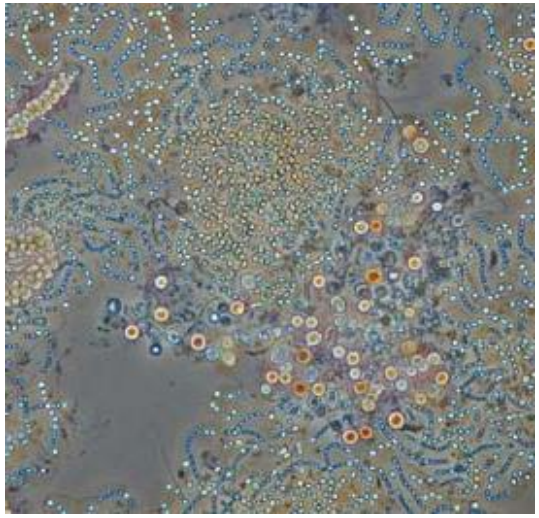
Real Metagenomics

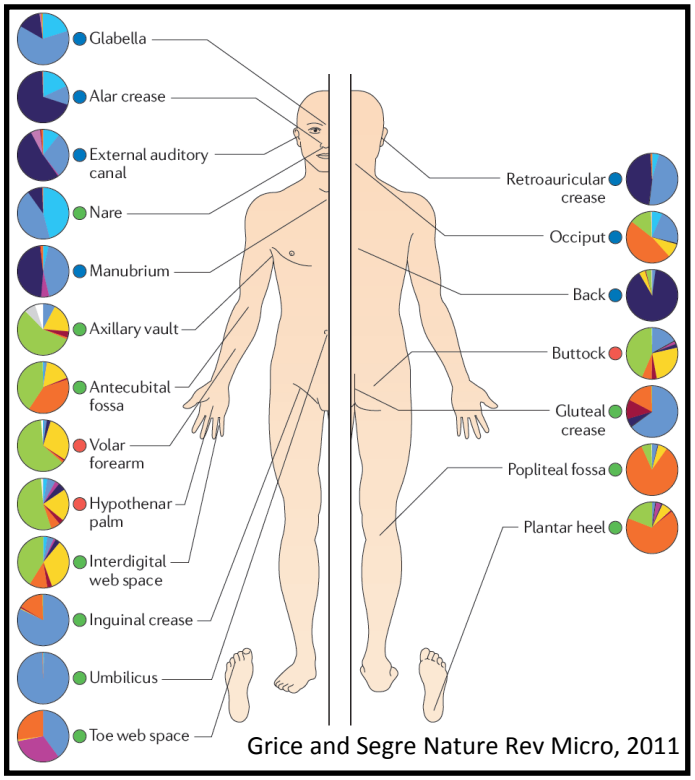
Computational Challenges in Metagenomics

Microbial communities

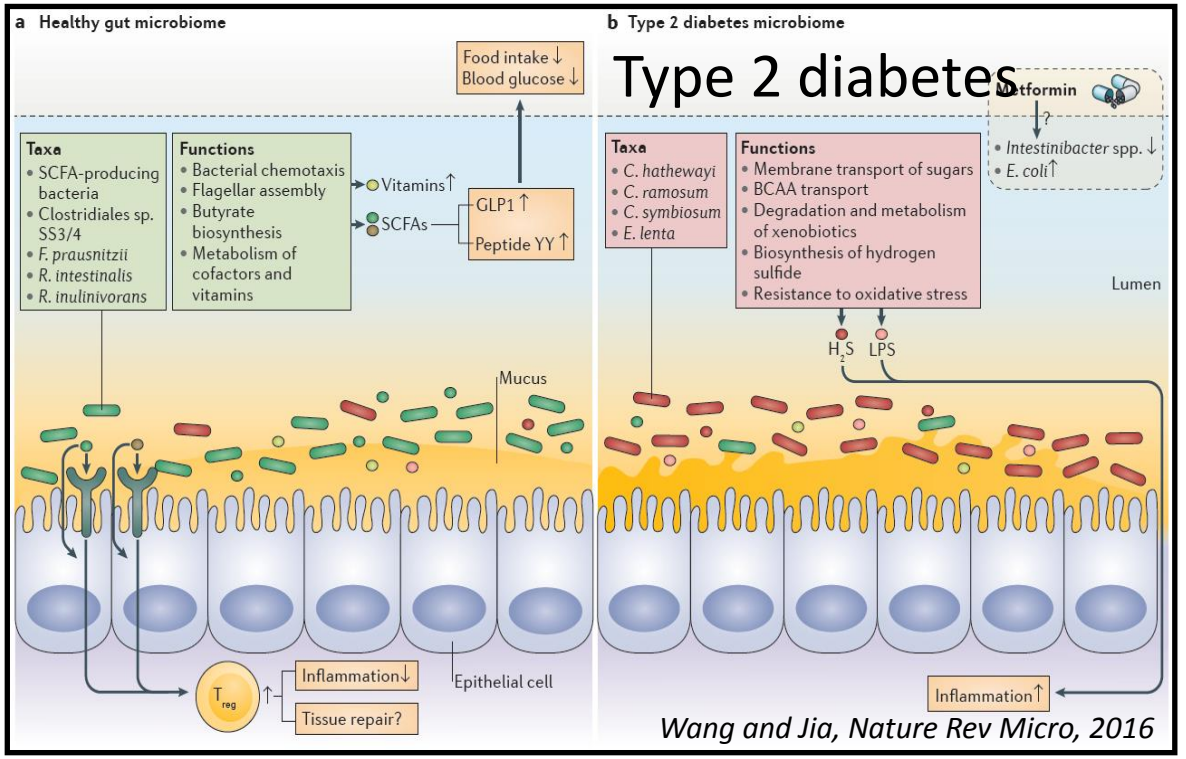




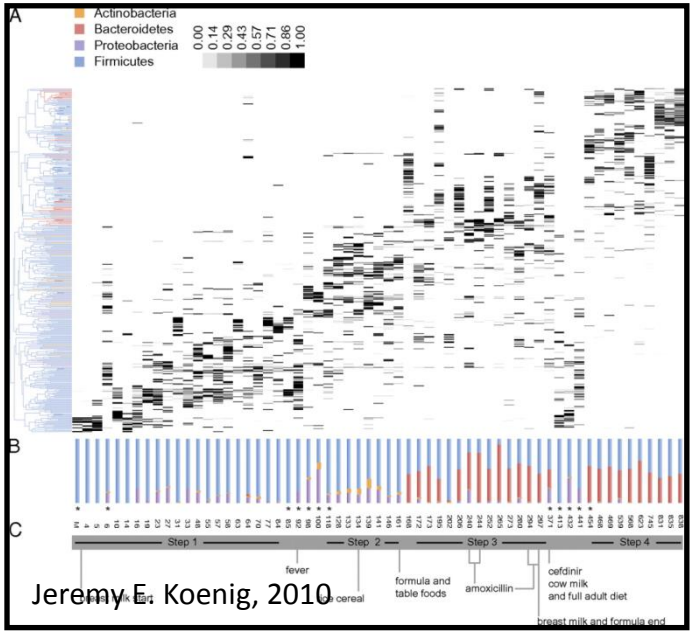


Grice and Segre Nature Rev Micro, 2011



Wang and Jia, Nature Rev Micro, 2016

- **Hundreds of species!**
- **100 trillion microbes!**
  (weighing ~3-4lb)
- **150x more genes**
  (~3,300,000)

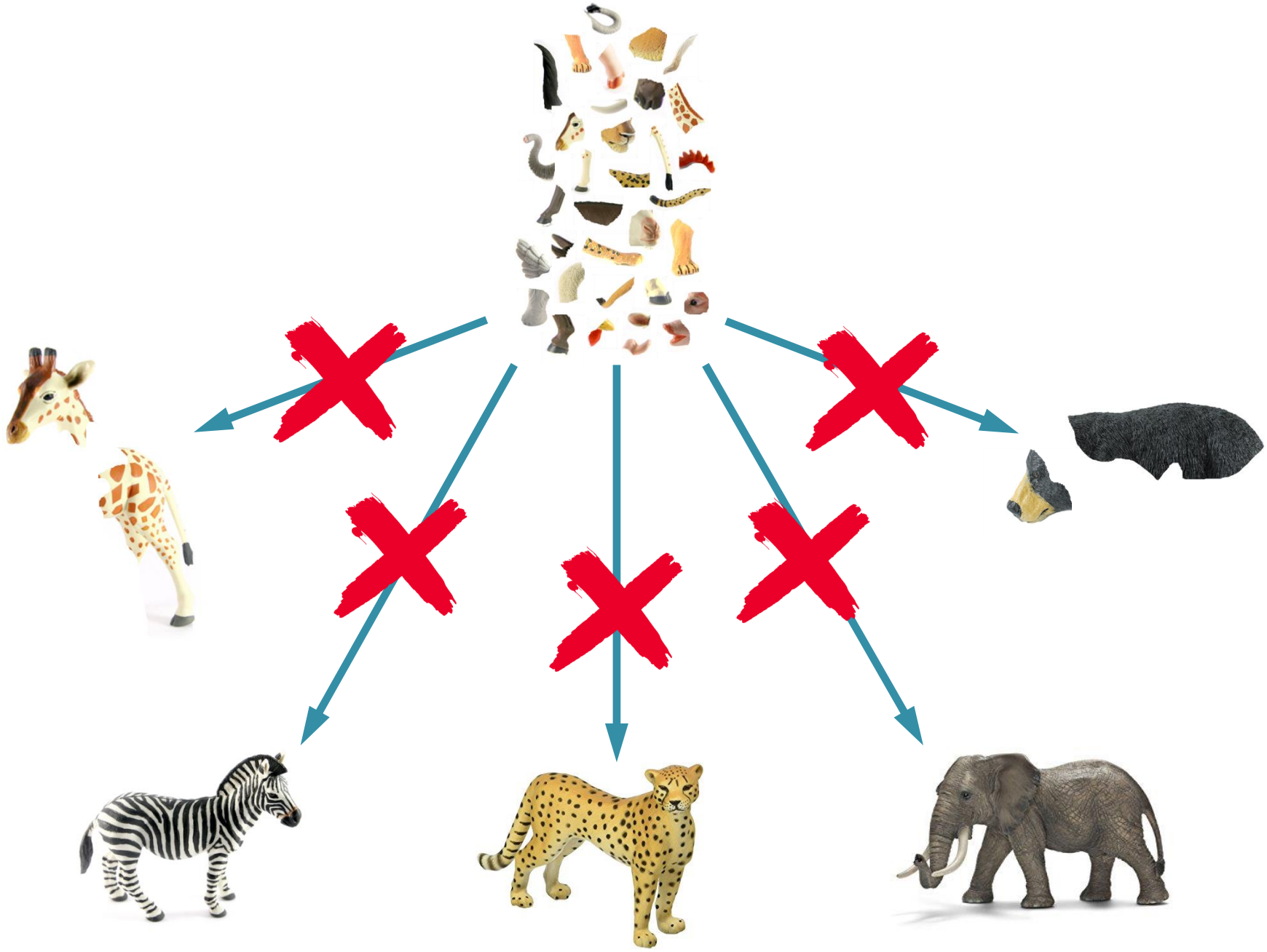


Jeremy E. Koenig, 2010

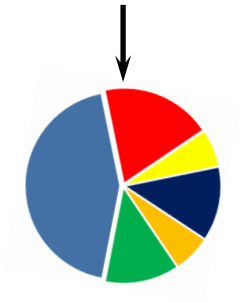
metagenomics

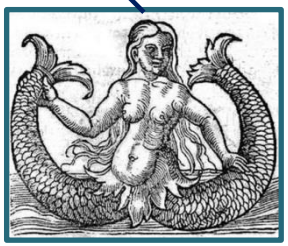
"The study of genetic material recovered directly from environmental samples"

(the term was first used by Jo Handelsman in 1998)



Who's there?

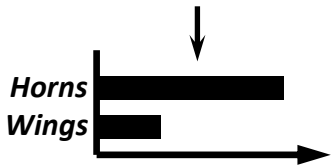




What are they doing?

Horns

Wings



Outline

A Few Updates about Seminar Logistics

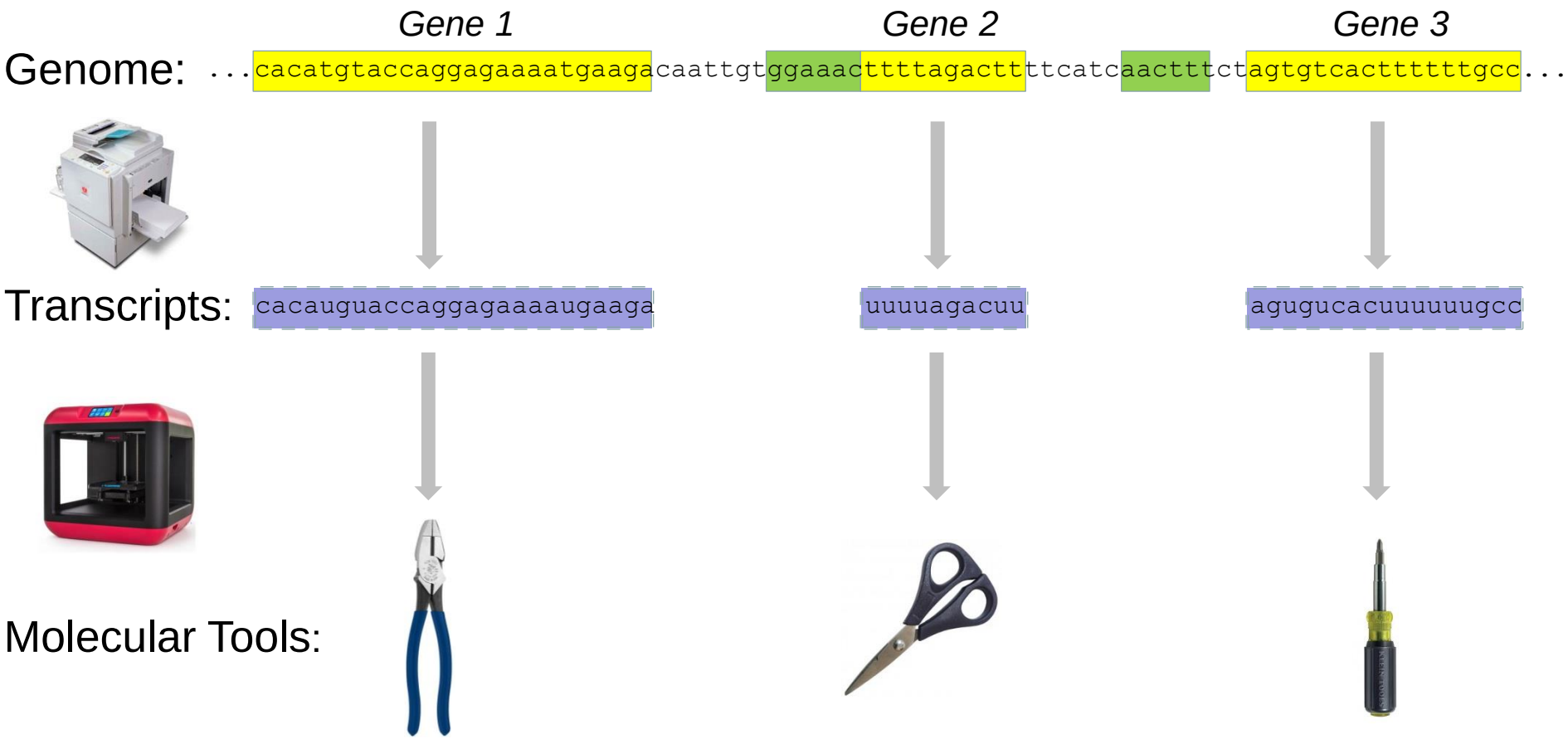
"Metagenomics"

Genomes, Genes, and the Central Dogma

Real Metagenomics

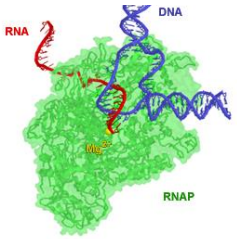
Computational Challenges in Metagenomics

Genomes and Genes:
A two minute conceptual view

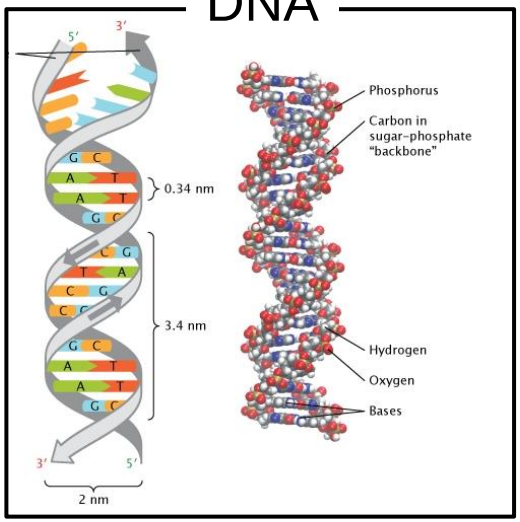


Genomes and Genes: In Real Life

RNA polymerase



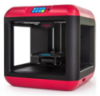
DNA



Genome: ...cacatgtaccaggagaaaatgaagacaattgtggaaactttagactttcatcaactttctagtgtcacttttttgcc...

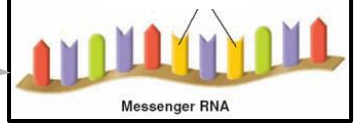
Gene 1 Gene 2 Gene 3

Transcripts: cacauguaccaggagaaaaugaaga uuuuagacuu agugucacuuuuuugcc

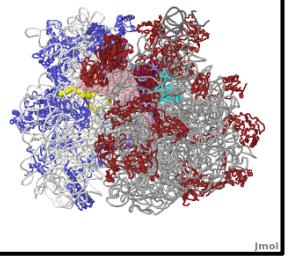


Molecular Tool:

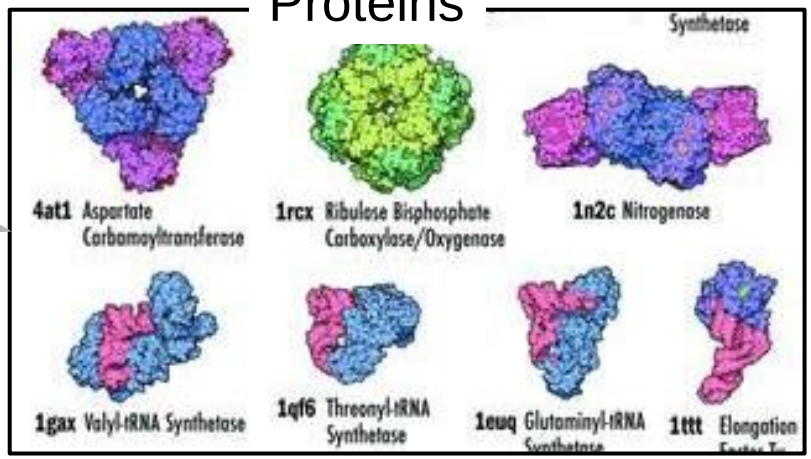
RNA

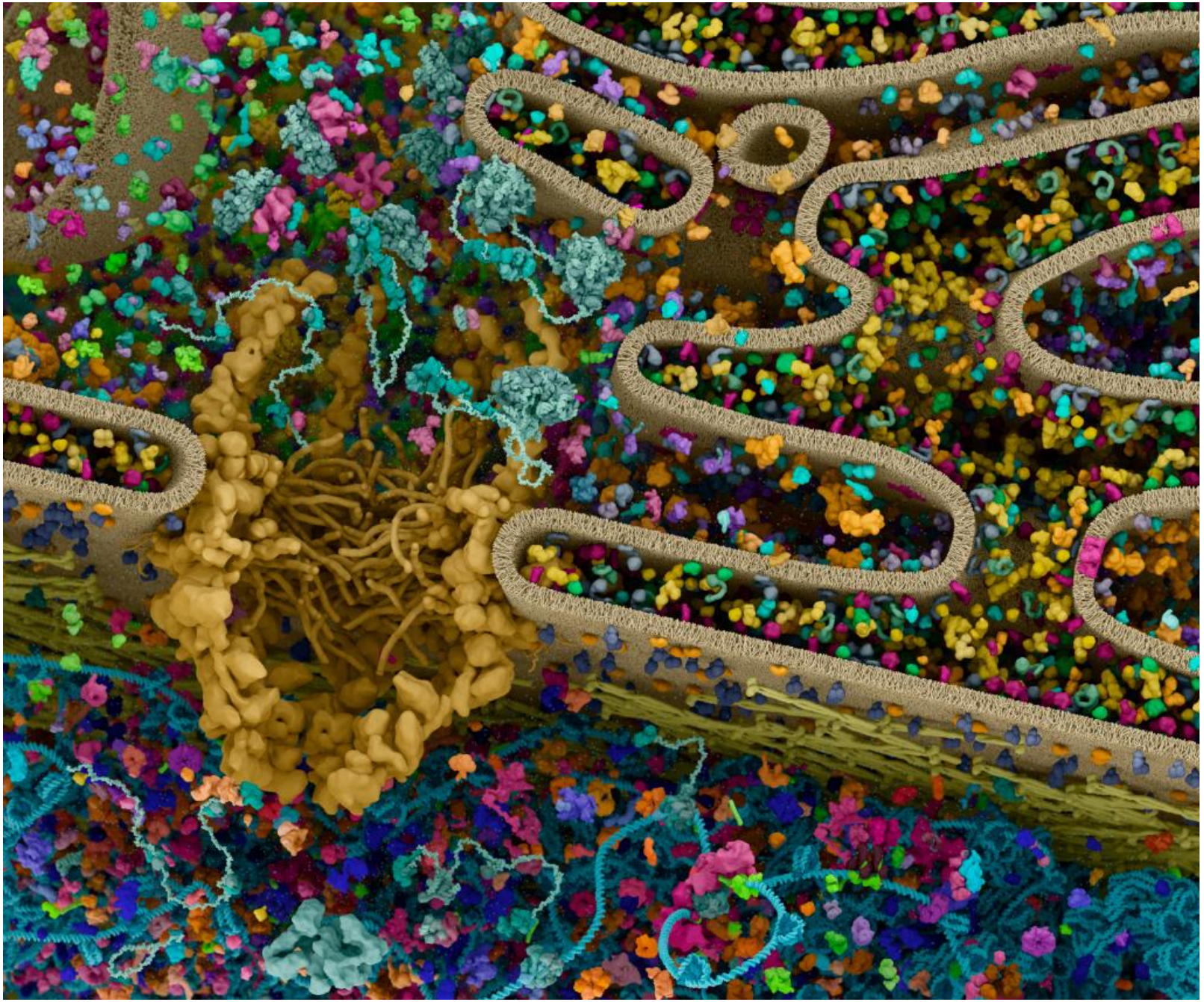


Ribosome

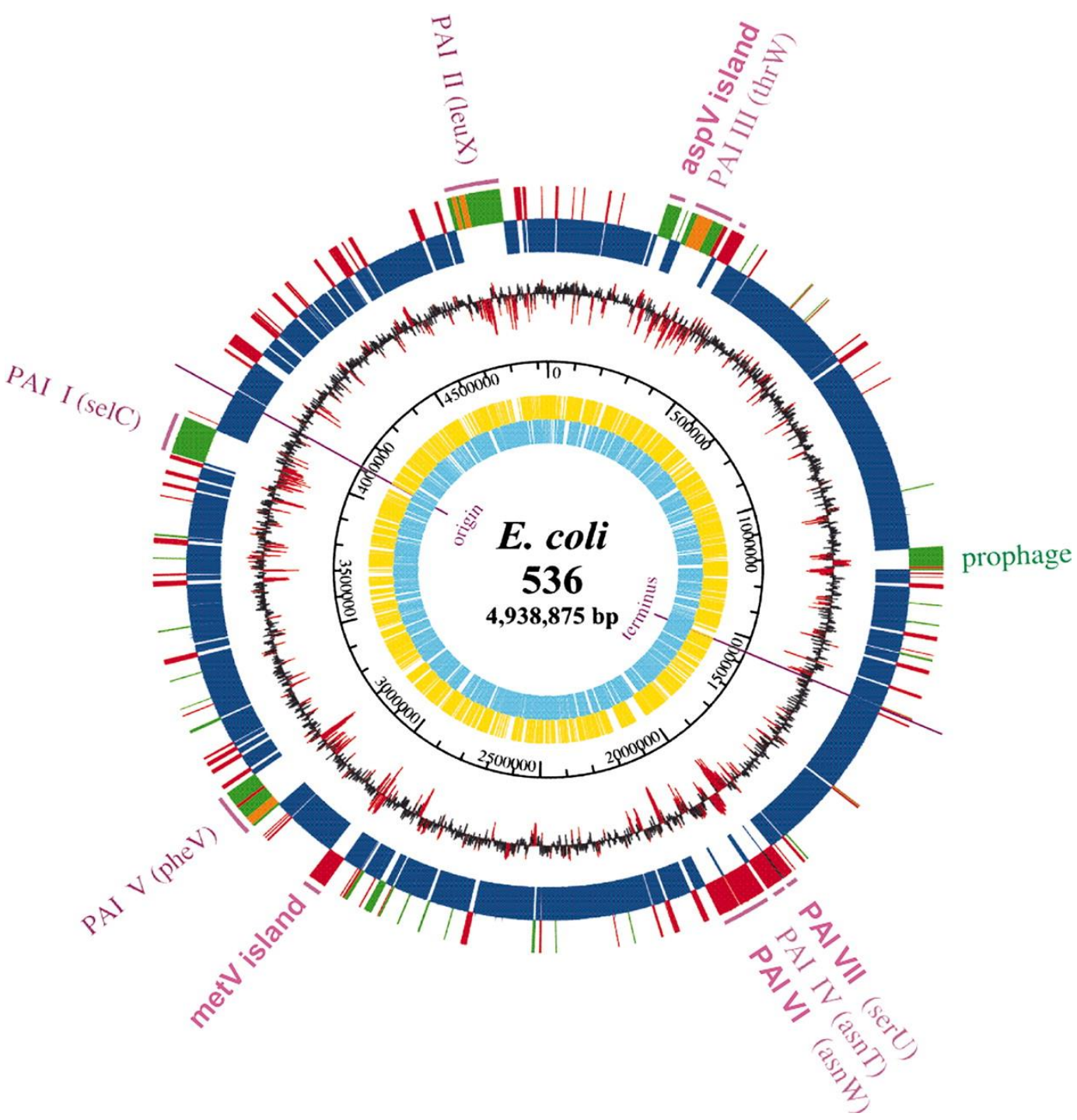


Proteins

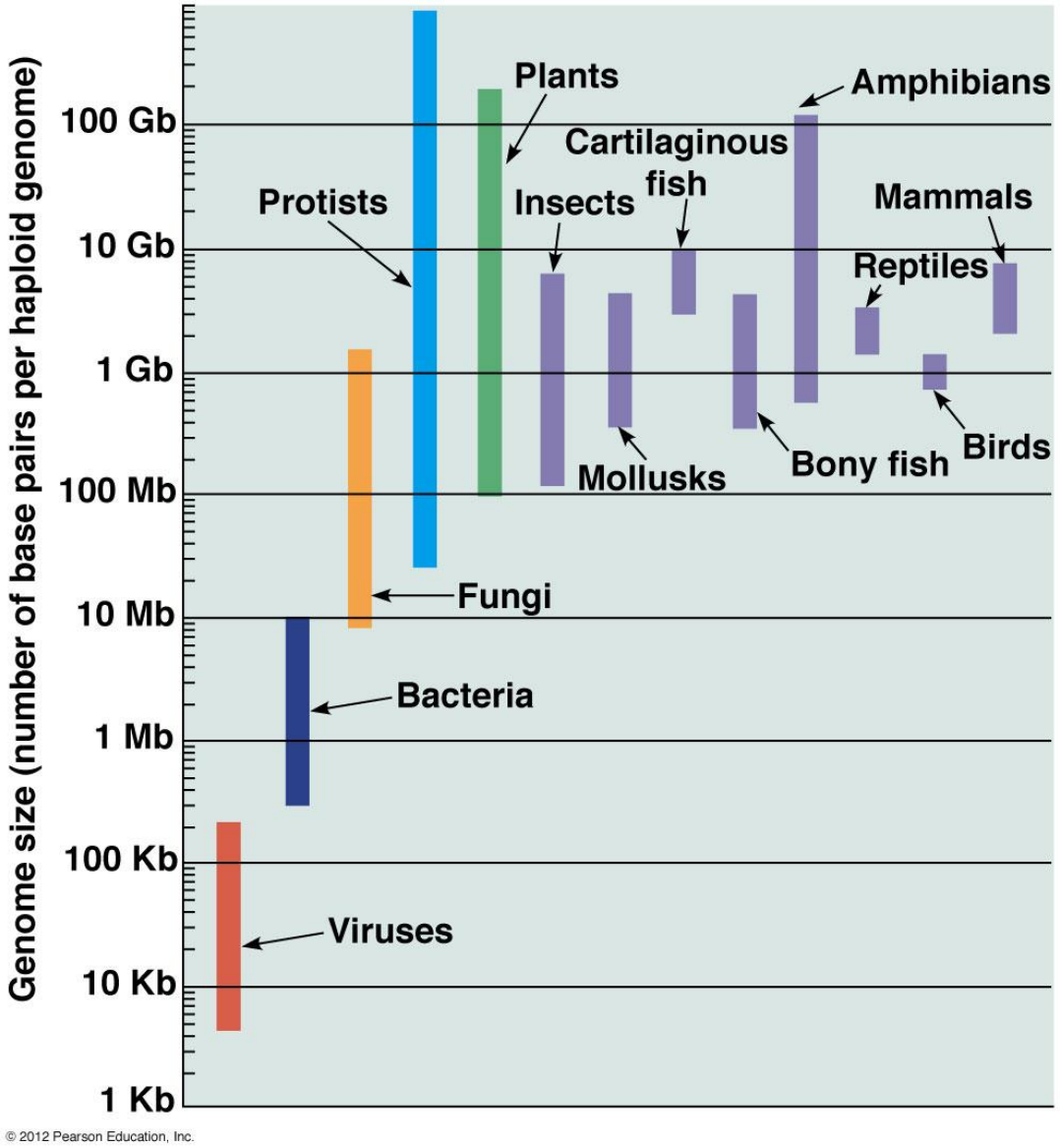




https://media.cellsignal.com/www/html/science/landscapes/protein-synthesis/protein-synthesis.html


PAI II (leuX)
aspV island
PAI III (thrW)
PAI I (selC)
prophage
PAI V (pheV)
metV island
PAI VII (serU)
PAI IV (asnT)
PAI VI (asnW)
E. coli
536
4,938,875 bp
origin
terminus
0
500000
1000000
1500000
2000000
2500000
3000000
3500000
4000000
4500000

Variation in Genome Length and # Genes



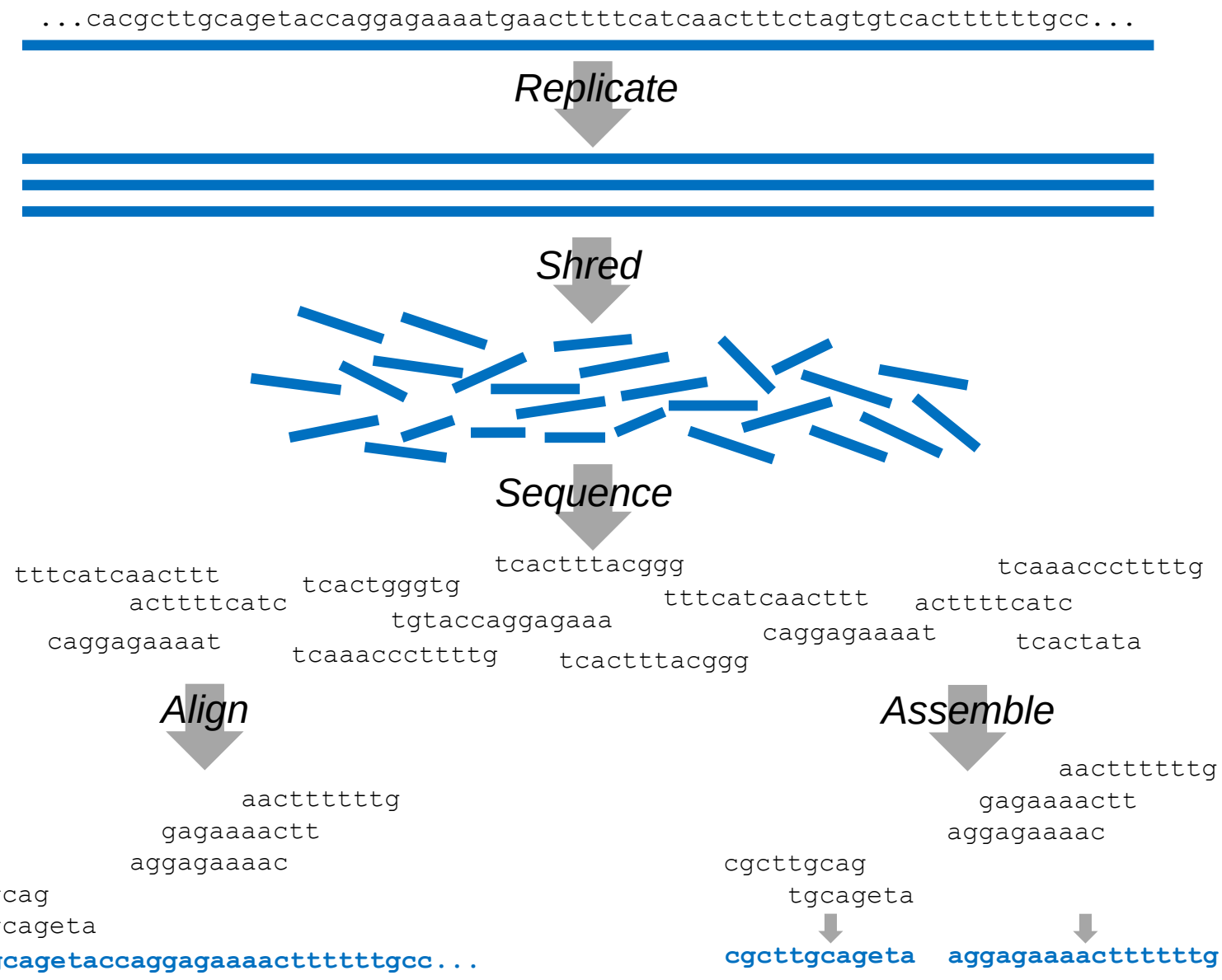
Variation in Genome Length and # Genes

organism	genome size (base pairs)	protein coding genes	number of chromosomes
model organisms			
model bacteria *E. coli*	4.6 Mbp	4,300	1
budding yeast *S. cerevisiae*	12 Mbp	6,600	16
fission yeast *S. pombe*	13 Mbp	4,800	3
amoeba *D. discoideum*	34 Mbp	13,000	6
nematode *C. elegans*	100 Mbp	20,000	12 (2n)
fruit fly *D. melanogaster*	140 Mbp	14,000	8 (2n)
model plant *A. thaliana*	140 Mbp	27,000	10 (2n)
moss *P. patens*	510 Mbp	28,000	27
mouse *M. musculus*	2.8 Gbp	20,000	40 (2n)
human *H. sapiens*	3.2 Gbp	21,000	46 (2n)
bacteria			
C. ruddii (smallest genome of an endosymbiont bacteria)	160 kbp	182	1
M. genitalium (smallest genome of a free living bacteria)	580 kbp	470	1
H. pylori	1.7 Mbp	1,600	1
Cyanobacteria S. elongatus	2.7 Mbp	3,000	1
methicillin-resistant *S. aureus* (MRSA)	2.9 Mbp	2,700	1
B. subtilis	4.3 Mbp	4,100	1
S. cellulosum (largest known bacterial genome)	13 Mbp	9,400	1
eukaryotes - multicellular			
pufferfish *Fugu rubripes* (smallest known vertebrate genome)	400 Mbp	19,000	22
poplar *P. trichocarpa* (first tree genome sequenced)	500 Mbp	46,000	19
corn *Z. mays*	2.3 Gbp	33,000	20 (2n)
dog *C. familiaris*	2.4 Gbp	19,000	40
chimpanzee *P. troglodytes*	3.3 Gbp	19,000	48 (2n)
wheat *T. aestivum* (hexaploid)	16.8 Gbp	95,000	42 (2n=6x)
marbled lungfish *P. aethiopicus* (largest known animal genome)	130 Gbp	unknown	34 (2n)
herb plant *Paris japonica* (largest known genome)	150 Gbp	unknown	40 (2n)

A Bit More About Genes (≈1 course)

- Different species differ in the set of genes they encode and in the exact sequence of each gene
- Homology:
  - Homologues genes are genes that derive from a common ancestor
  - Orthologues genes are homologous in different species (arise via speciation)
  - Paralogues are homologous genes in the same species (arise via gene duplication)
- As species evolve, their genomes (and genes) gradually diverge in sequence
- The closer the sequences of two genes are, the more likely it is the proteins they encode function similarly

Sequencing, Alignment, Assembly (≈1 course)



Outline

A Few Updates about Seminar Logistics

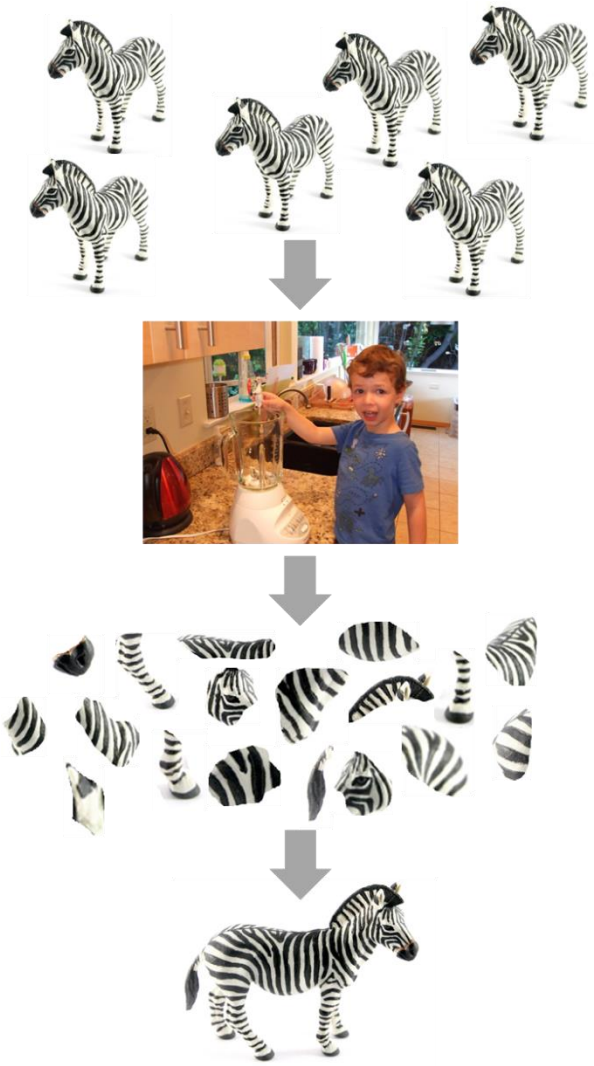
"Metagenomics"

Genomes, Genes, and the Central Dogma

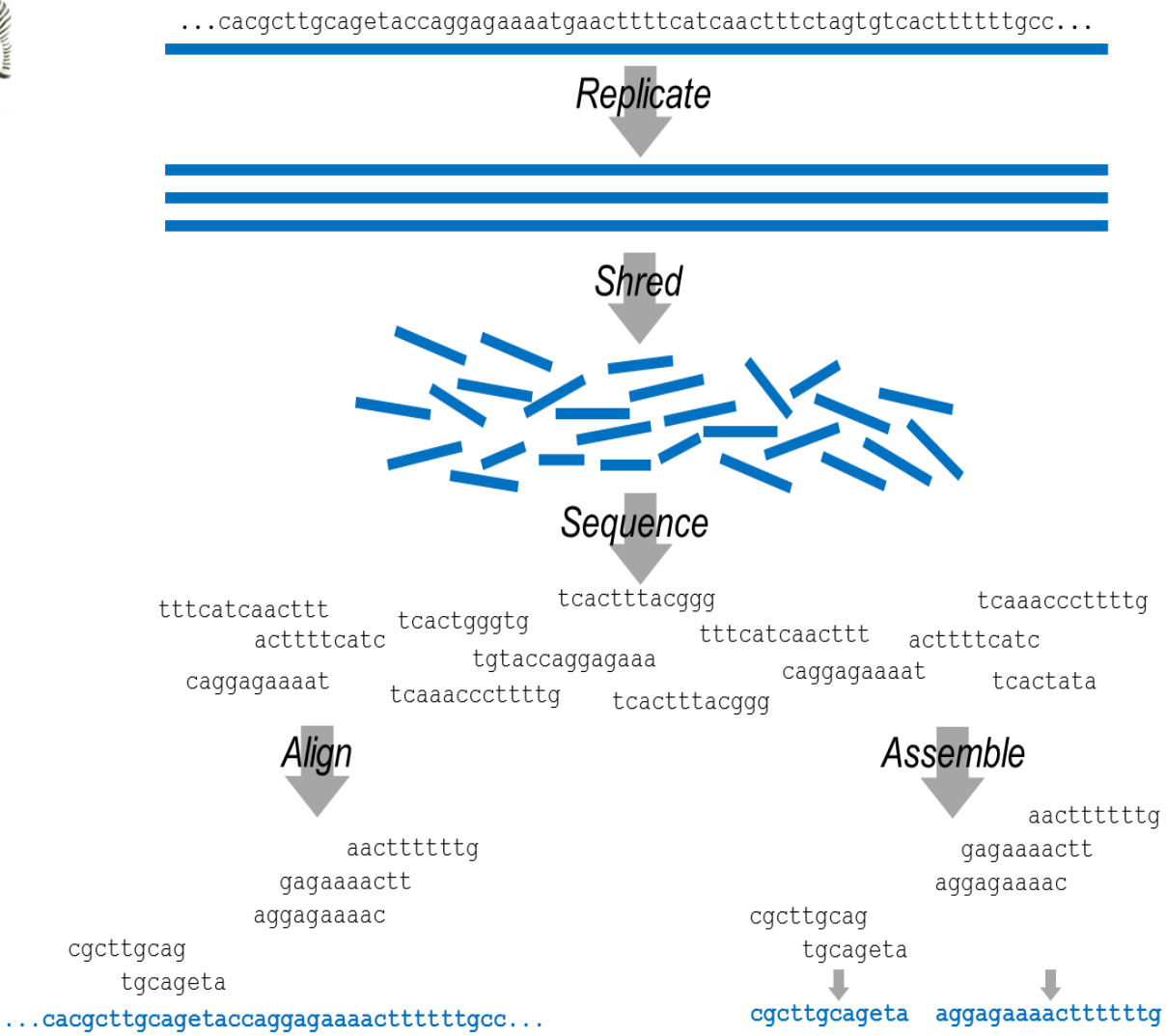
Real Metagenomics

Computational Challenges in Metagenomics

“Genomics”

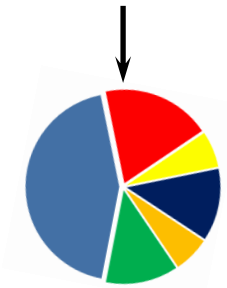


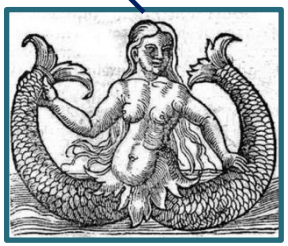
Real Genomics



“Metagenomics”

Who’s there?

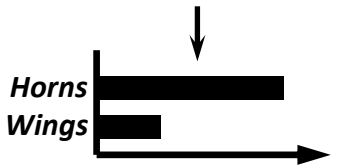




What are they doing?

Horns

Wings

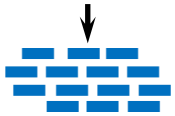


‘Real’ Metagenomics

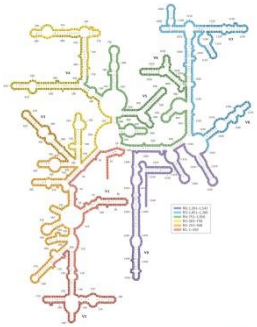
Microbiome sample

Extract DNA

Targeted sequencing (16S)



AG-AGCTGCTCA
AG-AGCTGCTCA
AG-AGCTGCTCA
AGGAGGTGCA-A
AGGAGGTGCA-A
AAA--CTGCTCA
AGGAGCT--TCA



Cluster into species or OTUs

Species/Taxonomic composition

Who’s there?

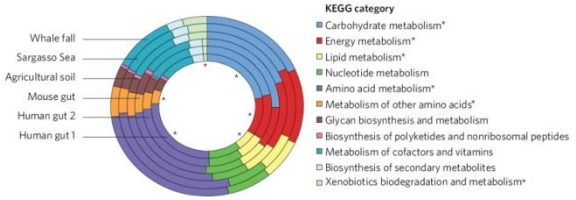
Shotgun sequencing

AGGAGCTGCTCA AGCGCGATCTGAA GAACCTGGTGA
ATTGAGAACCCGGCTG CCATTGCCGAGC TGAAGGCAATCA
TTCTACCCC TATGATGTGACCGT GCCCATTGCCGAG ACCACCAAG
TTCGCCCAGCTGAAG ACCGTCGATGTCC TGATCAAC
CAGATCGAG CGCACCATTGCCG TCAACTAC ACTGGCCTGGTC
CTGGACTTCTGG GACAAGC GCAAGGGCGTCCC GGTGGTATCATCTG
GTCACTGGA TTCAATGCCAT CTACCAGGTG CCCGTCTACTCC
GTCAACTTCAC CAGCTCCCTGGCG AAACTGGC

Map to a gene database (annotate)

BLASTx

Functional composition & comparative analysis



What are they doing?

Outline

A Few Updates about Seminar Logistics

"Metagenomics"

Genomes, Genes, and the Central Dogma

Real Metagenomics

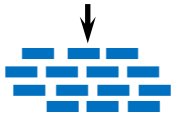
Computational Challenges in Metagenomics

Key Challenge 1

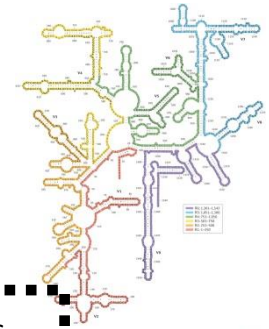
Microbiome sample

Extract DNA

Targeted sequencing (16S)



```
AG-AGCTGCTCA
AG-AGCTGCTCA
AG-AGCTGCTCA
AGGAGGTGCA-A
AGGAGGTGCA-A
AAA--CTGCTCA
AGGAGCT--TCA
```



Cluster into species or OTUs

Species/Taxonomic composition

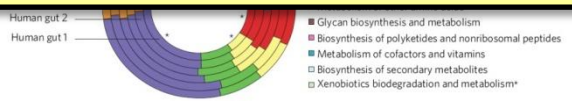
Who's there?

Shotgun sequencing

```
AGGAGCTGCTCA  AGCGCGATCTGAA   GAACCTGGTGA
  ATTGAGAACCCGGCTG  CCATTGCCGAGC   TGAAGGCAATCA
TTCTACCCC  TATGATGTGACCGT  GCCCATTGCCGAG  ACCACCAAG
    TTCGCCCAGCTGAAG   ACCGTCGATGTCC   TGATCAAC
```

From 16S Sequences to Taxa Composition

- Clear clusters are not always feasible
- True variation vs. sequencing errors?
- Low resolution
- A fixed threshold doesn't always reflect the same phylogenetic closeness
  - Different species clustered together
  - Strains clustered separately



What are they doing?

Key Challenge 2

Microbiome sample

Extract DNA

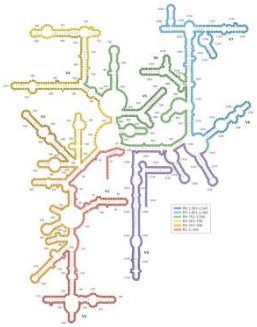
Targeted sequencing (16S)

```
AG-AGCTGCTCA
AG-AGCTGCTCA
AG-AGCTGCTCA
AGGAGGTGCA-A
AGGAGGTGCA-A
AAA--CTGCTCA
AGGAGCT--TCA
```

Cluster into species or OTUs

Species/Taxonomic composition

Who's there?

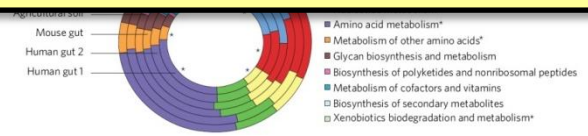


Shotgun sequencing

```
AGGAGCTGCTCA AGCGCGATCTGAA GAACCTGGTGA
ATTGAGAACCCGGCTG CCATTGCCGAGC TGAAGGCAATCA
TTCTACCCC TATGATGTGACCGT GCCCATTGCCGAG ACCACCAAG
TTCGCCCAGCTGAAG ACCGTCGATGTCC TGATCAAC
CAGATCGAG CGCACCATTGCCG TCAACTAC ACTGGCCTGGTC
CTGGACTTCTGG GACAAGC GCAAGGGCGTCCC GGTGGTATCATCTG
GTCACTGGA TTCAATGCCAT CTACCAGGTG CCCGTCTACTCC
GTCAACTTCAC CAGCTCCCTGGCG AAACTGGC
```

From Shotgun Sequencing to Species- and Strain-Level Profile

- Different species encode different genes
- Low abundance strains
- Shared variation



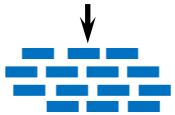
What are they doing?

Key Challenge 3

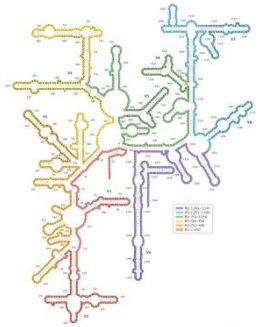
Microbiome sample

Extract DNA

Targeted sequencing (16S)



AG-AGCTGCTCA
AG-AGCTGCTCA
AG-AGCTGCTCA
AGGAGGTGCA-A
AGGAGGTGCA-A
AAA--CTGCTCA
AGGAGCT--TCA



Cluster into species or OTUs

Shotgun sequencing

AGGAGCTGCTCA AGCGCGATCTGAA GAACCTGGTGA
ATTGAGAACCCGGCTG CCATTGCCGAGC TGAAGGCAATCA
TTCTACCCC TATGATGTGACCGT GCCCATTGCCGAG ACCACCAAG
TTCGCCCAGCTGAAG ACCGTCGATGTCC TGATCAAC
CAGATCGAG CGCACCATTGCCG TCAACTAC ACTGGCCTGGTC
CTGGACTTCTGG GACAAGC GCAAGGGCGTCCC GGTGGTATCATCTG
GTCACTGGA TTCAATGCCAT CTACCAGGTG CCCGTCTACTCC
GTCAACTTCAC CAGCTCCCTGGCG AAACTGGC

Map to a gene database (annotate)

BLASTx

nparative analysis

Comperting Taxonomic Profiles

- **Compositionality**
- **Phylogeny**
- Distance metrics

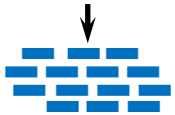
doing?

Key Challenge 4

Microbiome sample

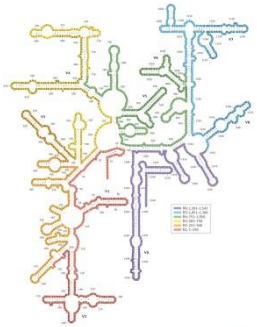
Extract DNA

Targeted sequencing (16S)



AG-AGCTGCTCA
AG-AGCTGCTCA
AG-AGCTGCTCA
AGGAGGTGCA-A
AGGAGGTGCA-A
AAA--CTGCTCA
AGGAGCT--TCA

Cluster into species or OTUs



Species/Taxonomic composition

Who's there?

Shotgun sequencing

AGGAGCTGCTCA AGCGCGATCTGAA GAACCTGGTGA
ATTGAGAACCCGGCTG CCATTGCCGAGC TGAAGGCAATCA
TTCTACCCC TATGATGTGACCGT GCCCATTGCCGAG ACCACCAAG
TTCGCCCAGCTGAAG ACCGTCGATGTCC TGATCAAC
CAGATCGAG CGCACCATTGCCG TCAACTAC ACTGGCCTGGTC
CTGGACTTCTGG GACAAGC GCAAGGGCGTCCC GGTGGTATCATCTG
GTCACTGGA TTCAATGCCAT CTACCAGGTG CCCGTCTACTCC
GTCAACTTCAC CAGCTCCCTGGCG AAACTGGC

Map to a gene database (annotate)

BLASTx

Inferring Species Interactions

- Not phenotypic data (uncaltured species)
- Low abundance species
- Sparsity
- Complex dynamics
- **Compositionality**