CSC 483 - Applied Biological Data Sciece
W2D2

CSC 483: Applied Biological Data Science

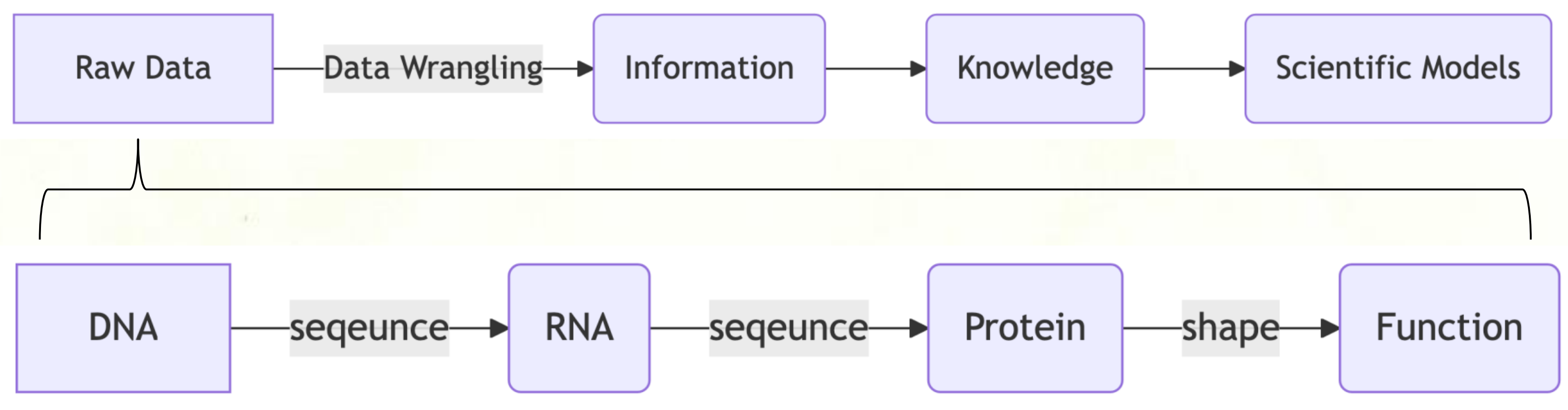
- Georgia Doing, PhD (she/her/hers)
- email: doingg@union.edu
- office: Steinmetz 108B
  - office hours:
    - Wednesday 2:00-3:30 pm
    - Thursday 4:00-5:30 pm

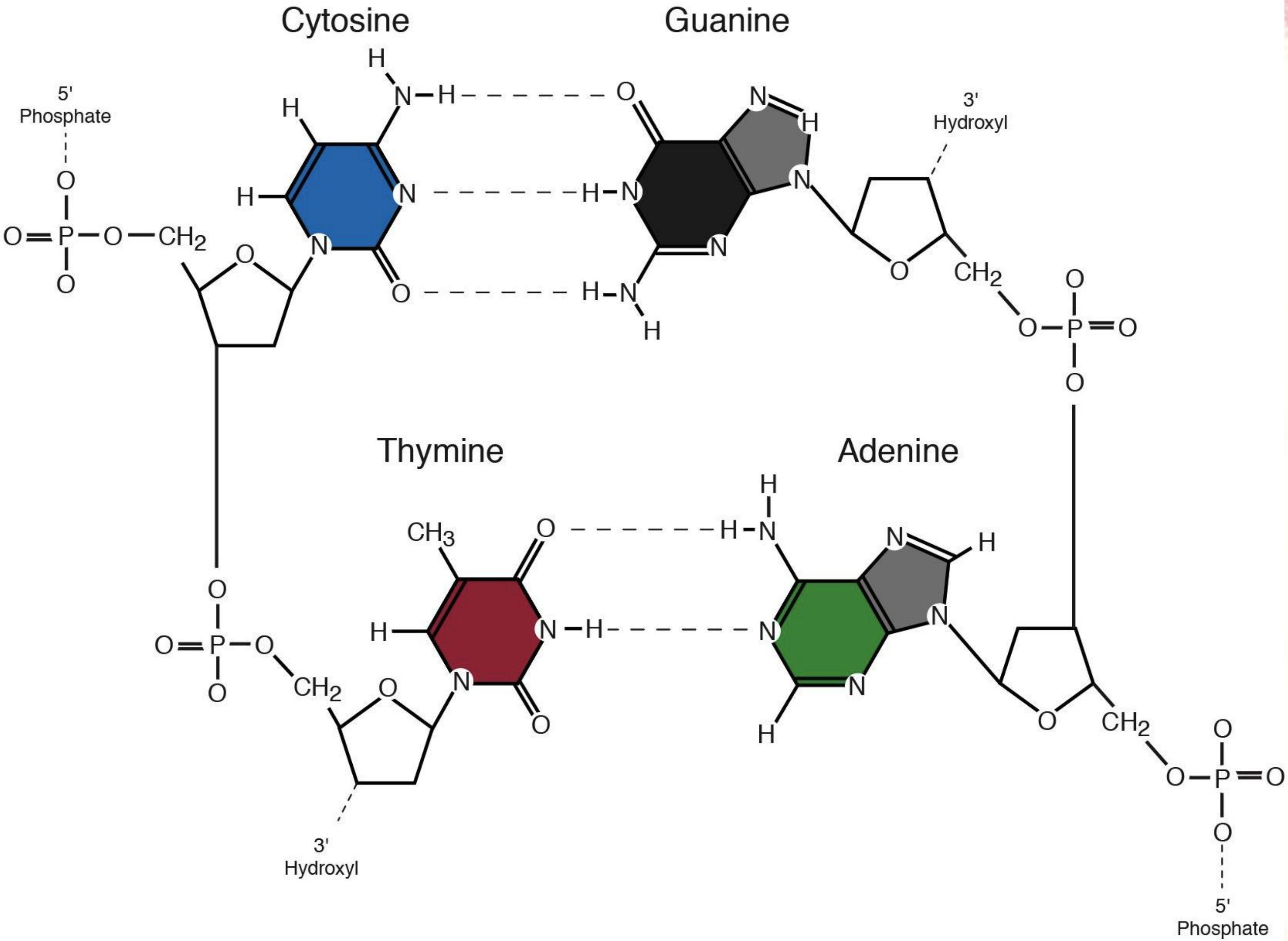
Feel free to drop in whenever my office door is open or schedule an appointment.

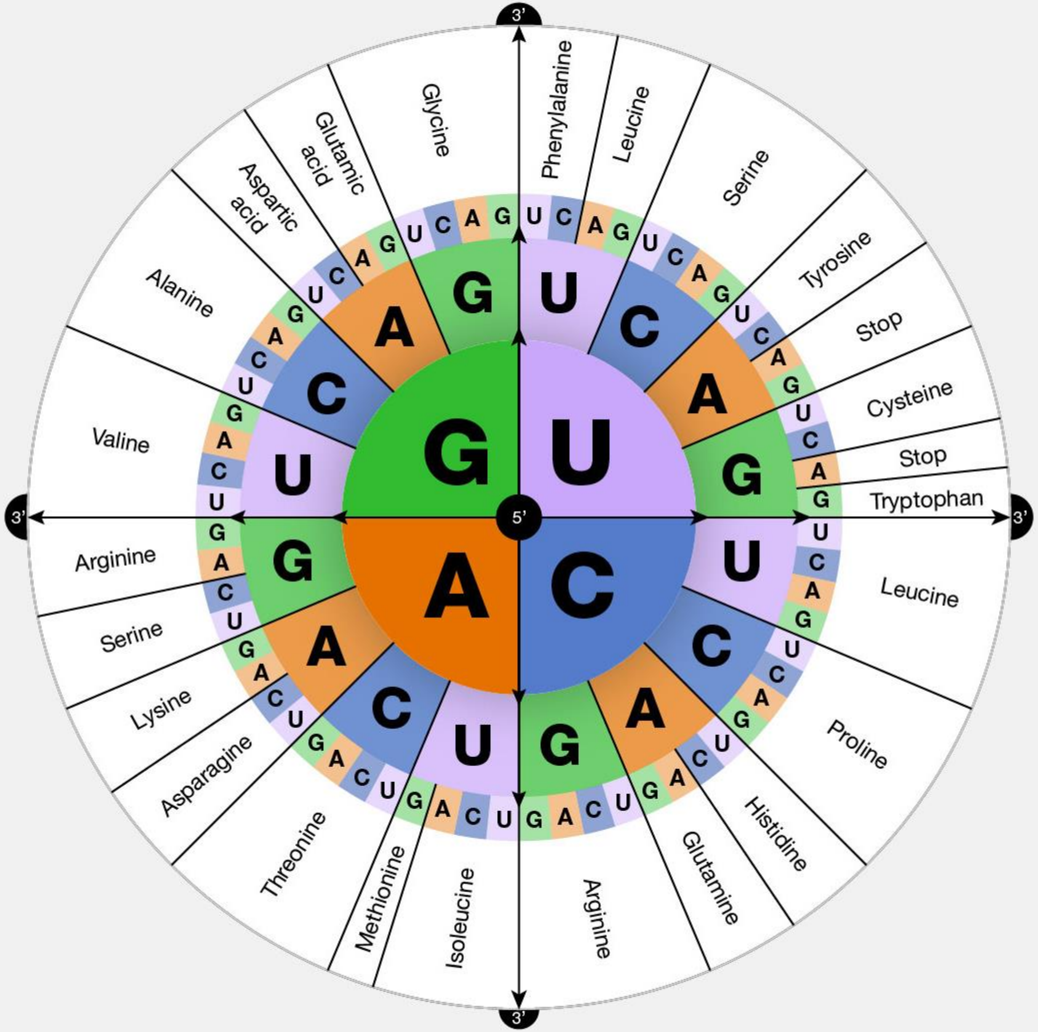
Today (145 min)

- Paper on methods/background
- Metadata summaries
- Explore data ,brainstorm ideas
- HW:
  - rosalind problems on sequence analysis (w/ colab notebooks)
  - due: tues

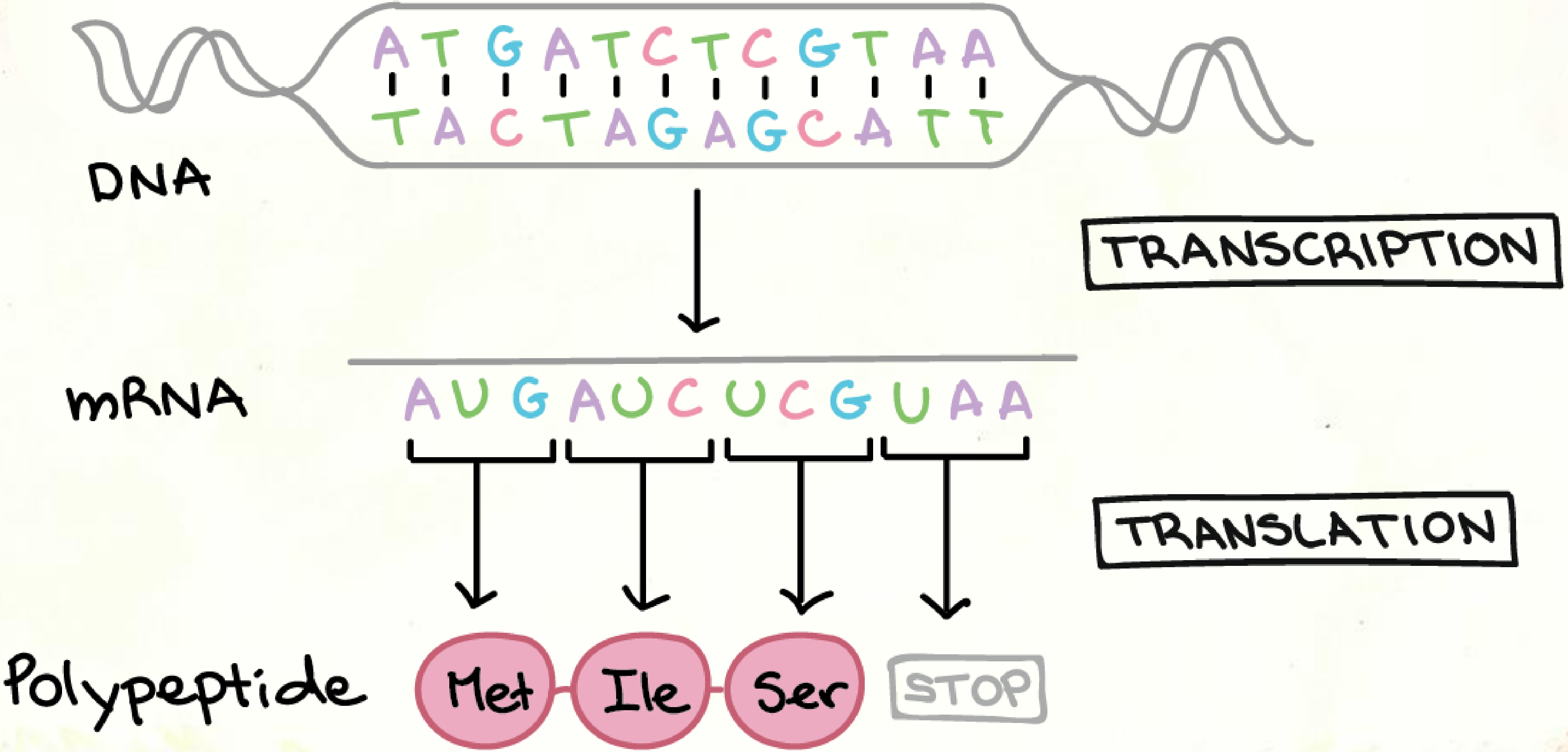
What is Biological Data Science?

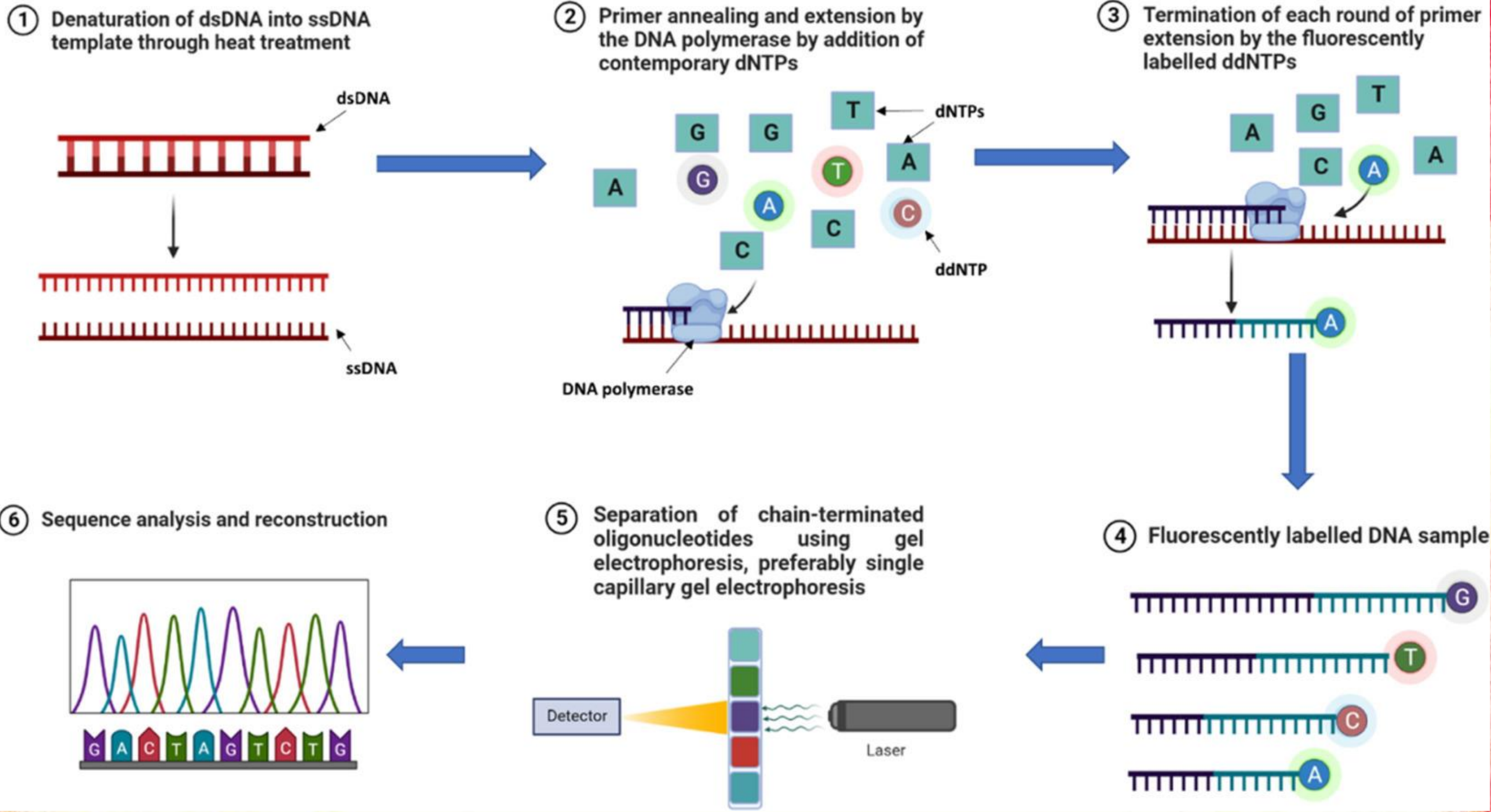



Cytosine
Guanine
Thymine
Adenine
5'
Phosphate
3'
Hydroxyl
3'
Hydroxyl
5'
Phosphate
CH_2
CH_3
O
P
N
H


5'
3'
3'
3'
3'
G
U
A
C
Phenylalanine
Leucine
Serine
Tyrosine
Stop
Cysteine
Stop
Tryptophan
Leucine
Proline
Histidine
Glutamine
Arginine
Isoleucine
Methionine
Threonine
Asparagine
Lysine
Serine
Arginine
Valine
Alanine
Aspartic acid
Glutamic acid
Glycine

THE CENTRAL DOGMA





1. Denaturation of dsDNA into ssDNA template through heat treatment

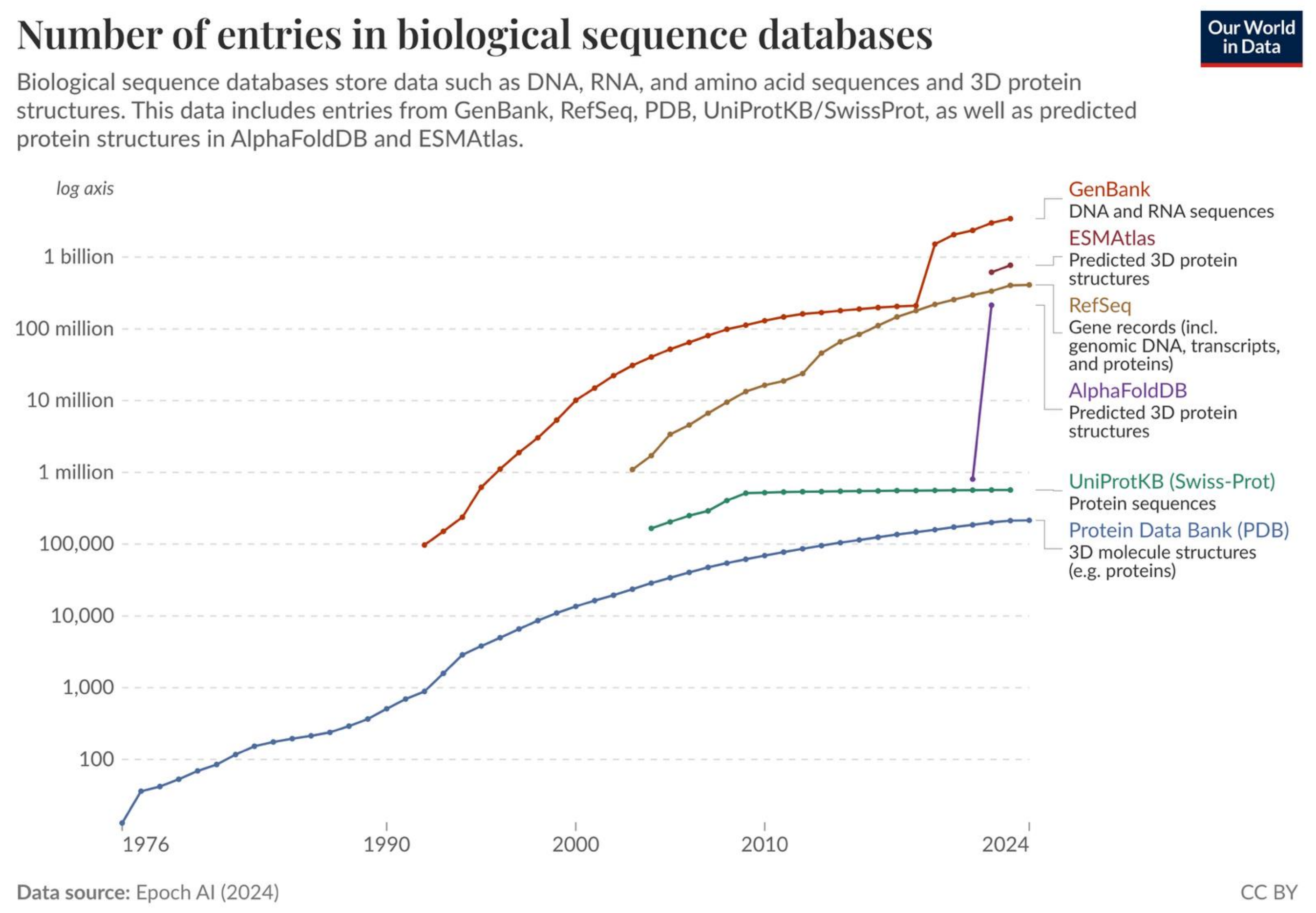
2. Primer annealing and extension by the DNA polymerase by addition of contemporary dNTPs

3. Termination of each round of primer extension by the fluorescently labelled ddNTPs

4. Fluorescently labelled DNA sample

5. Separation of chain-terminated oligonucleotides using gel electrophoresis, preferably single capillary gel electrophoresis

6. Sequence analysis and reconstruction

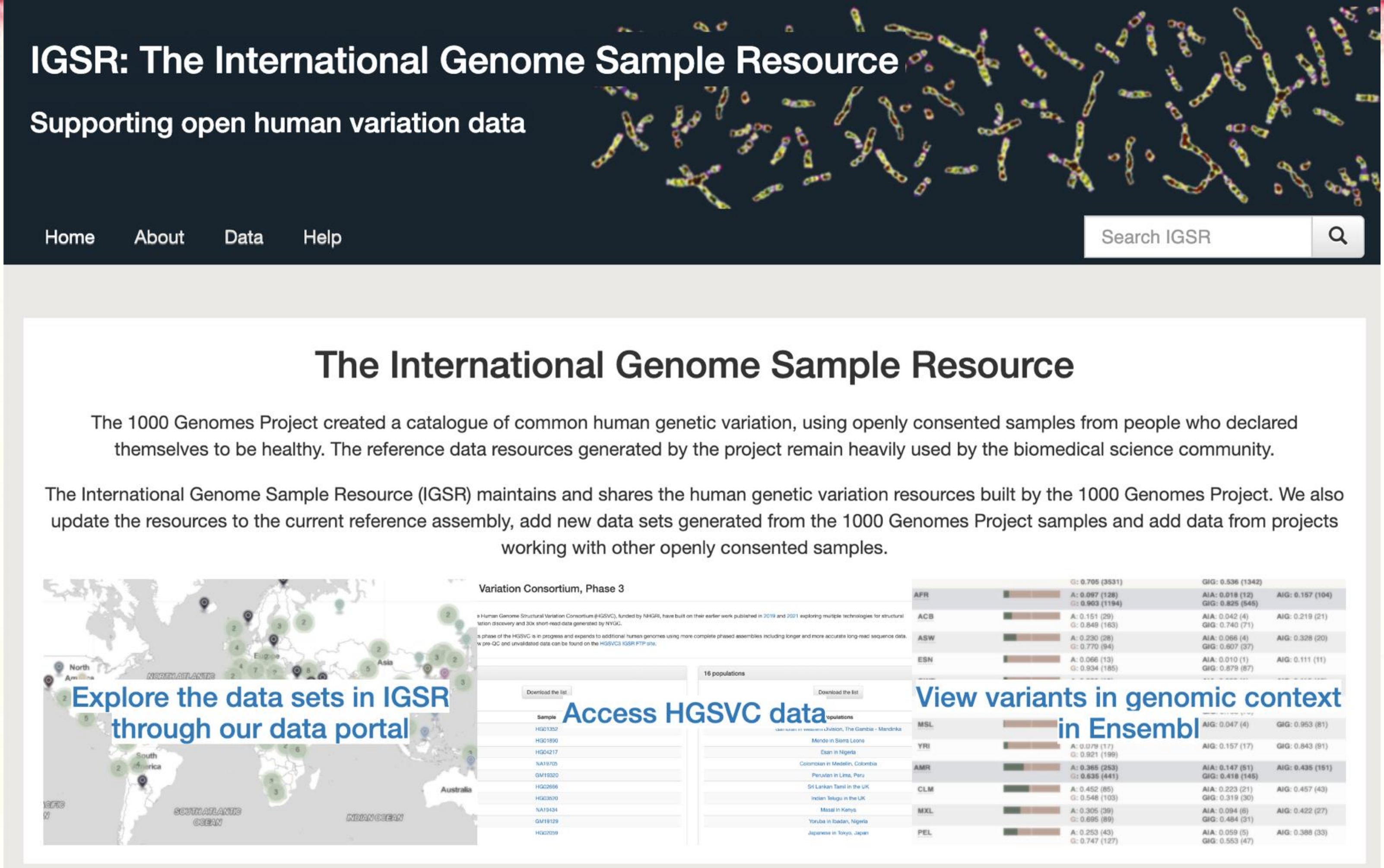


Number of entries in biological sequence databases

Biological sequence databases store data such as DNA, RNA, and amino acid sequences and 3D protein structures. This data includes entries from GenBank, RefSeq, PDB, UniProtKB/SwissProt, as well as predicted protein structures in AlphaFoldDB and ESMAtlas.

Data source: Epoch AI (2024)

https://ourworldindata.org/grapher/number-of-entries-in-biological-sequence-databases



IGSR: The International Genome Sample Resource

Supporting open human variation data

The International Genome Sample Resource

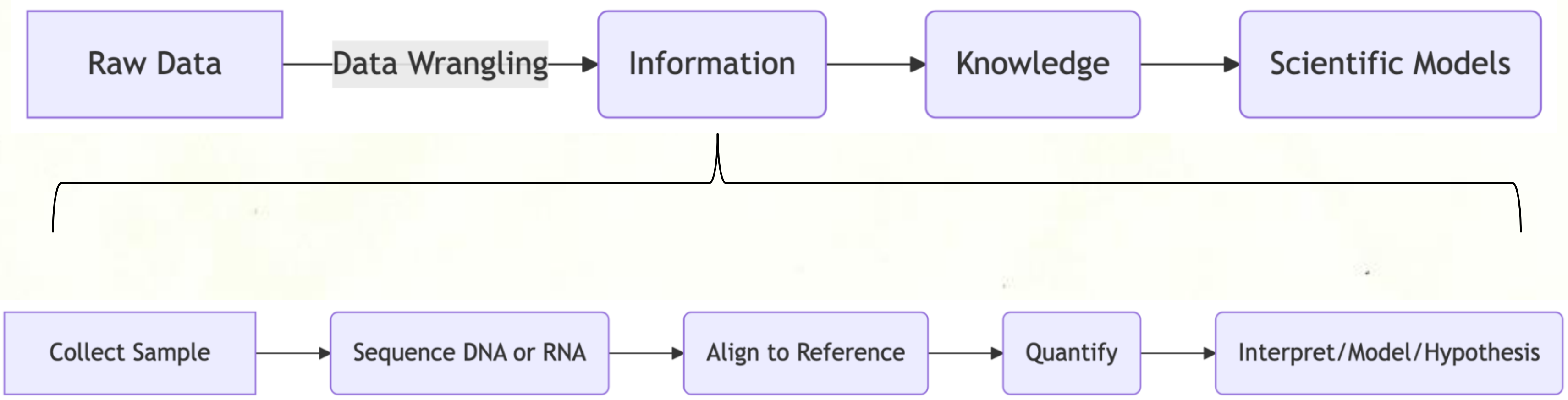
The 1000 Genomes Project created a catalogue of common human genetic variation, using openly consented samples from people who declared themselves to be healthy. The reference data resources generated by the project remain heavily used by the biomedical science community.

The International Genome Sample Resource (IGSR) maintains and shares the human genetic variation resources built by the 1000 Genomes Project. We also update the resources to the current reference assembly, add new data sets generated from the 1000 Genomes Project samples and add data from projects working with other openly consented samples.

Explore the data sets in IGSR through our data portal
Access HGSVC data
View variants in genomic context in Ensembl

1000s Genomes Project: https://www.internationalgenome.org/

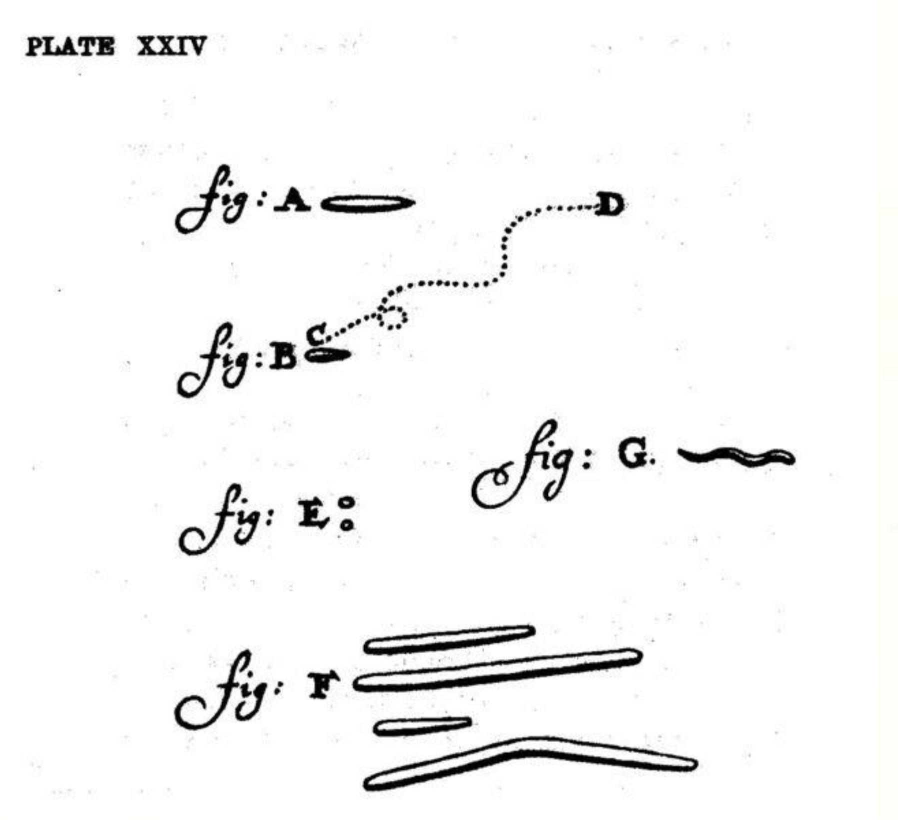
Experimental Workflow



Raw Data → Data Wrangling → Information → Knowledge → Scientific Models

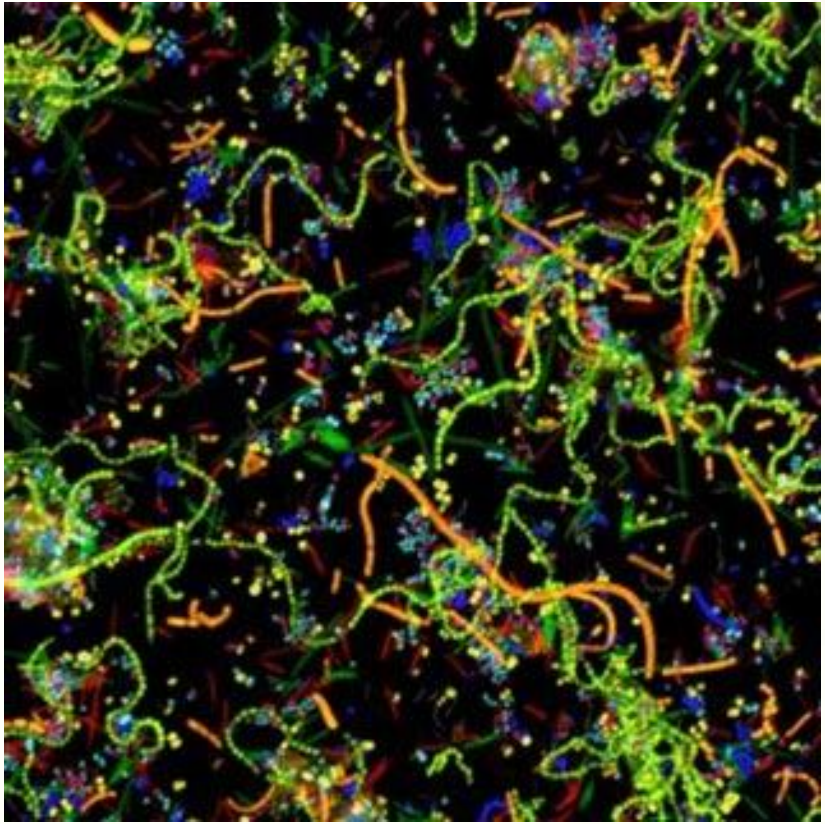
Collect Sample → Sequence DNA or RNA → Align to Reference → Quantify → Interpret/Model/Hypothesis

Oral ‘animacules’ (1676)



A. van Leeuwenhoek

Oral microbiome (2011)



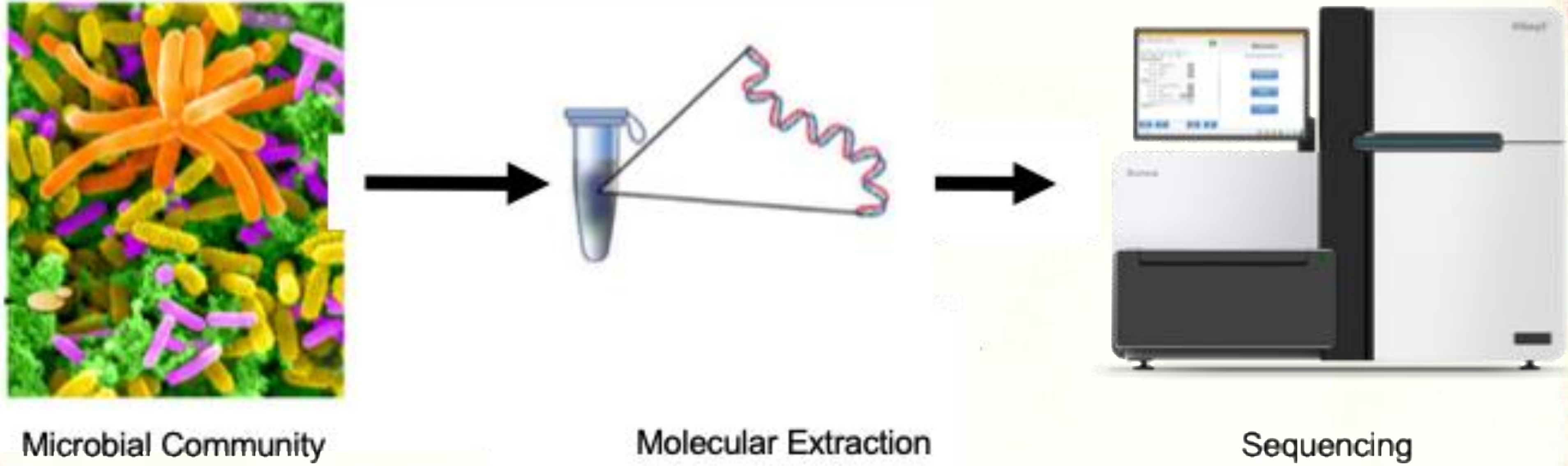
Valm et al., PNAS

Isolate

Genomics / Transcriptomics

Community

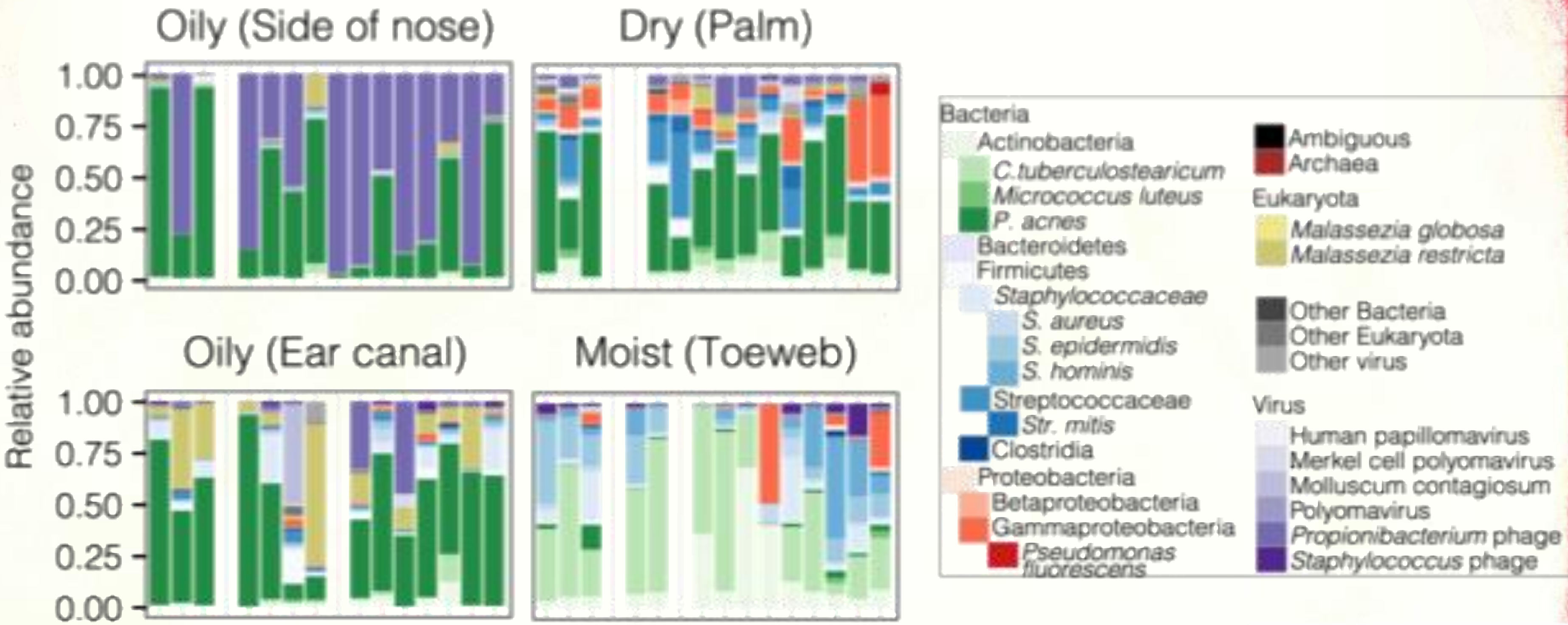
Metagenomics / Metatranscriptomics



Microbial Community

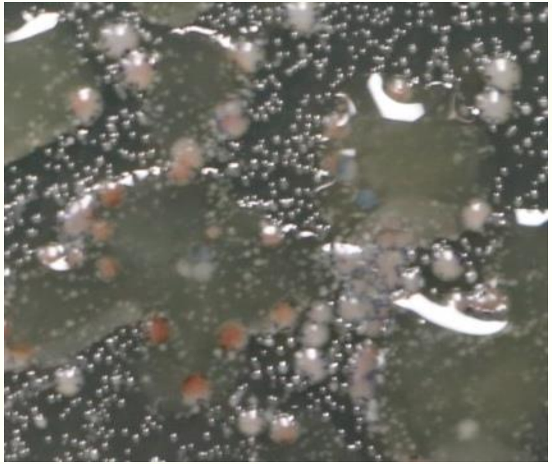
Molecular Extraction

Sequencing

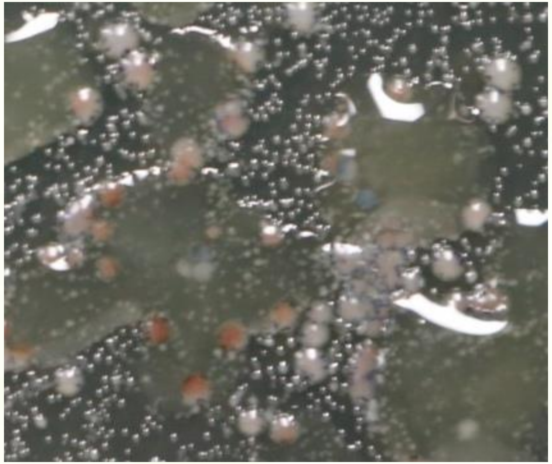


Oh et al., *Nature*

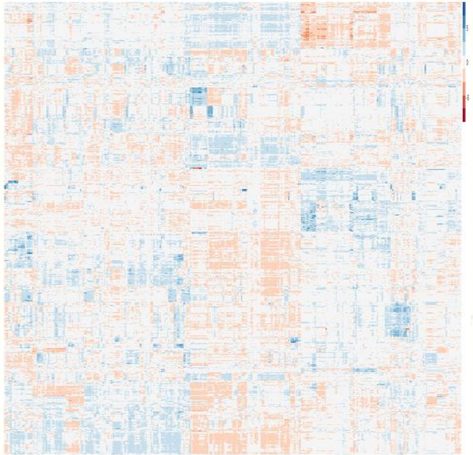
Microbial Culture



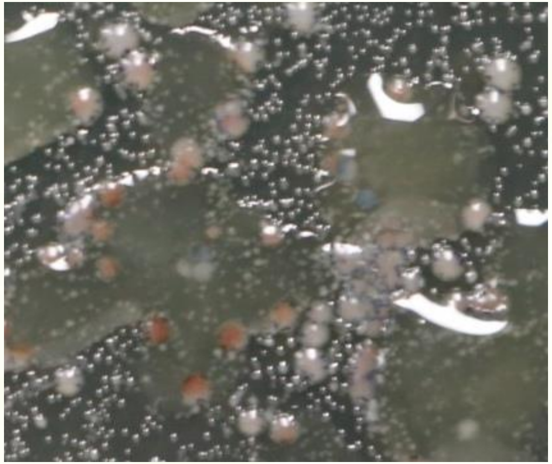
Microbial Culture



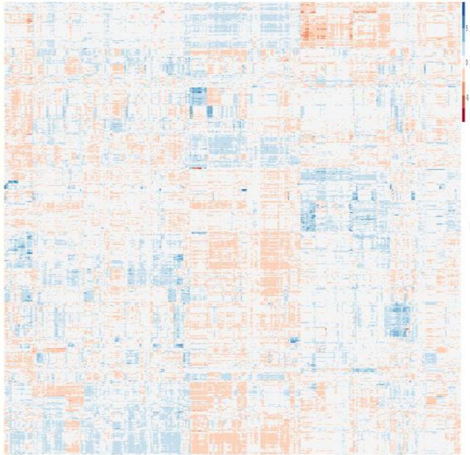
RNA-seq Data



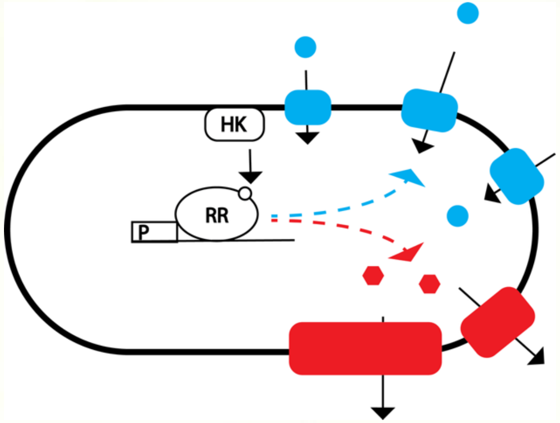
Microbial Culture

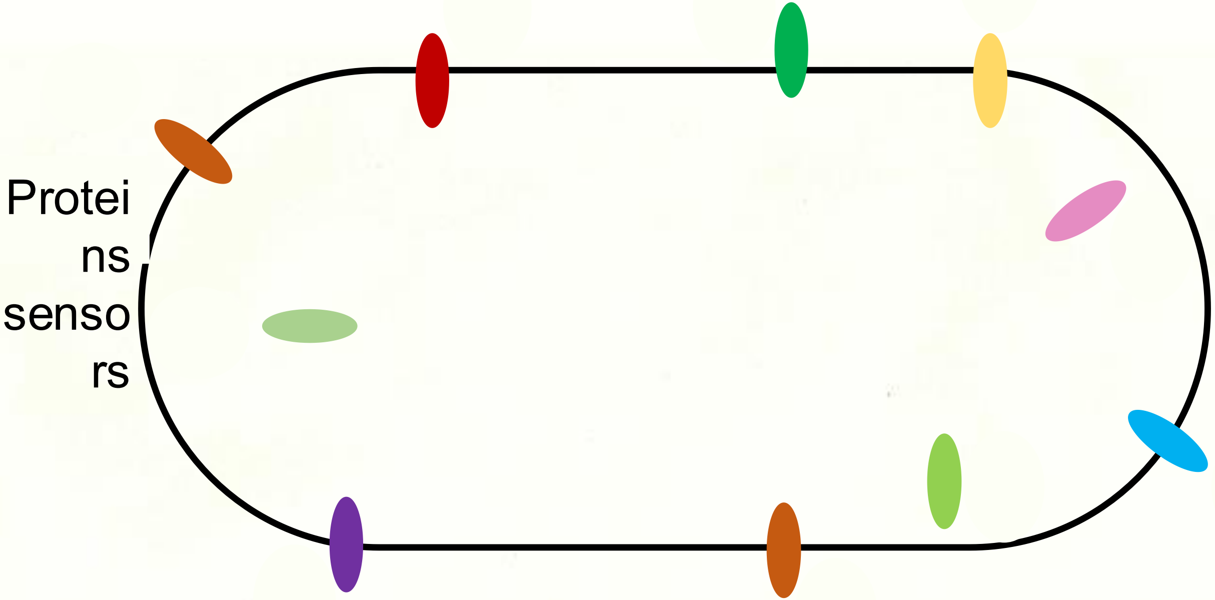


RNA-seq Data

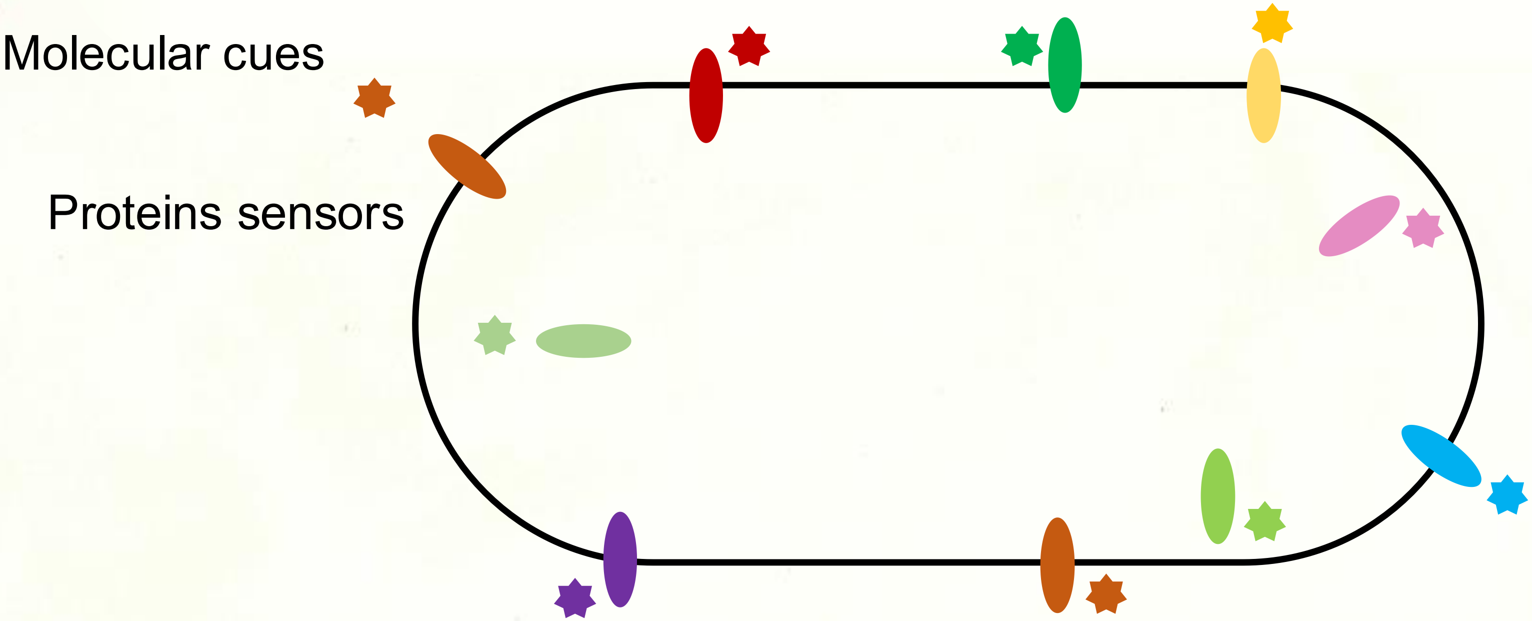


Biological Model



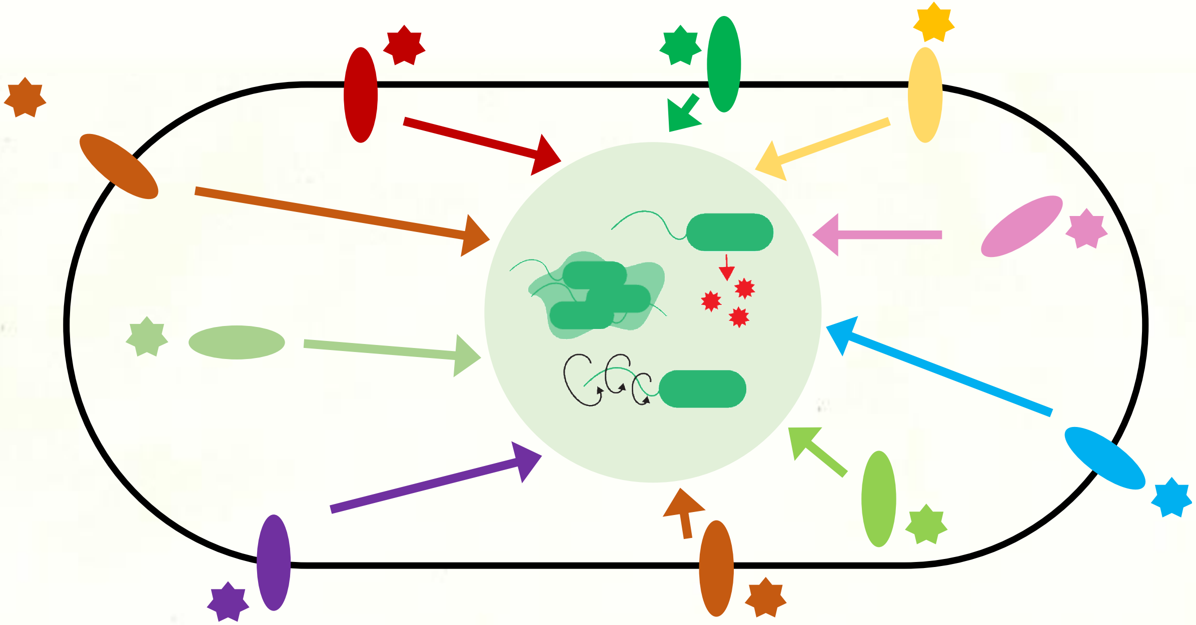


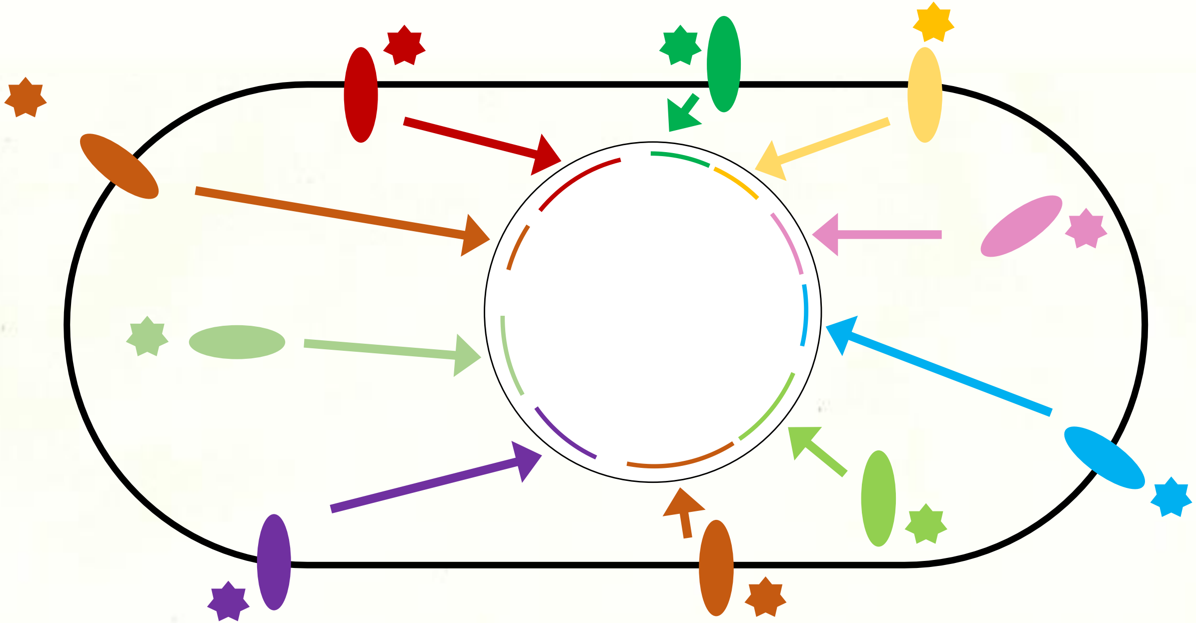
Proteins sensors

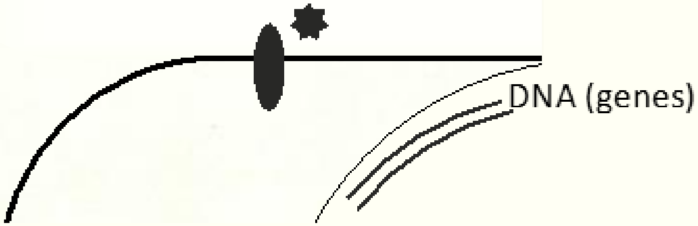


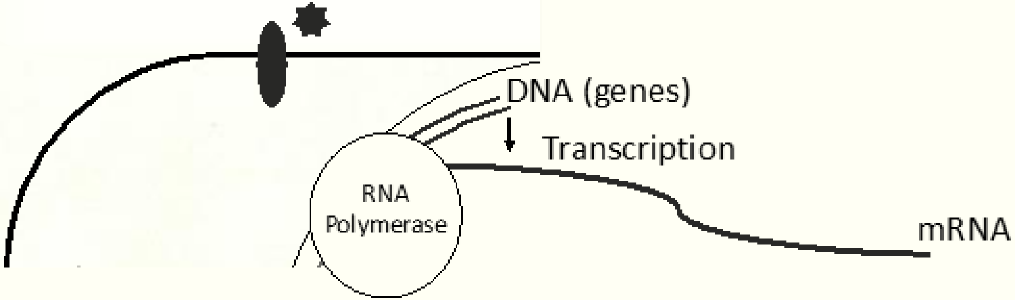
Molecular cues

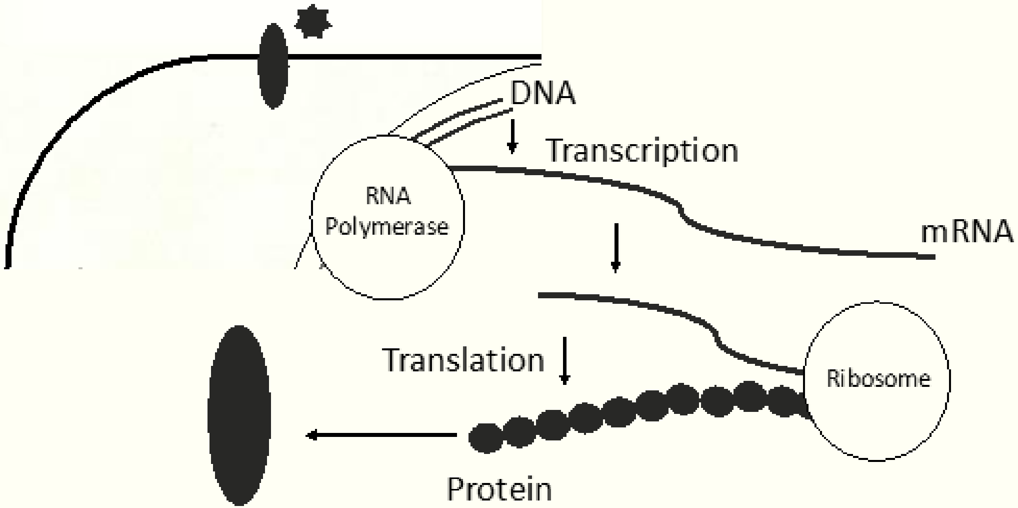
Proteins sensors

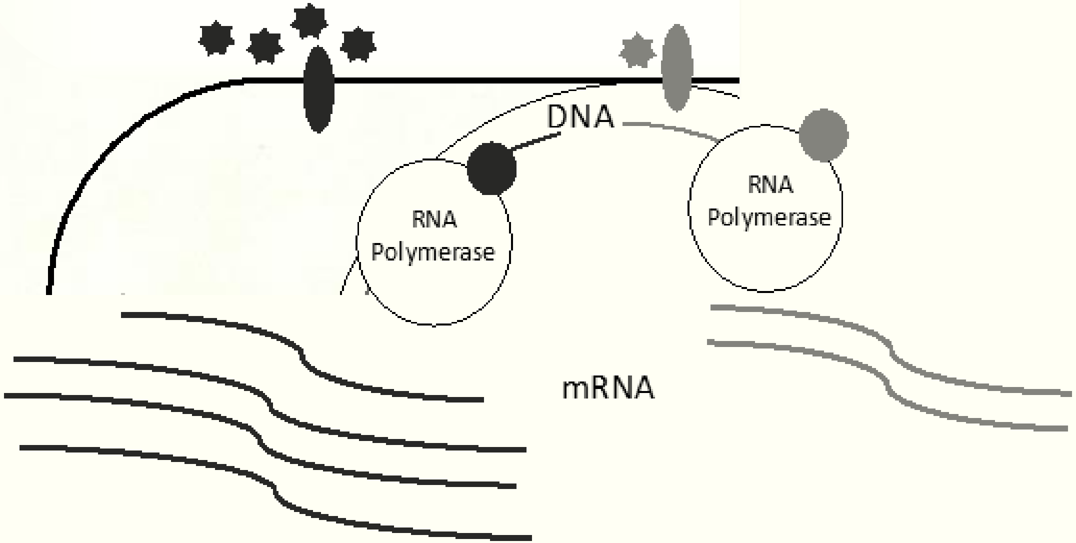


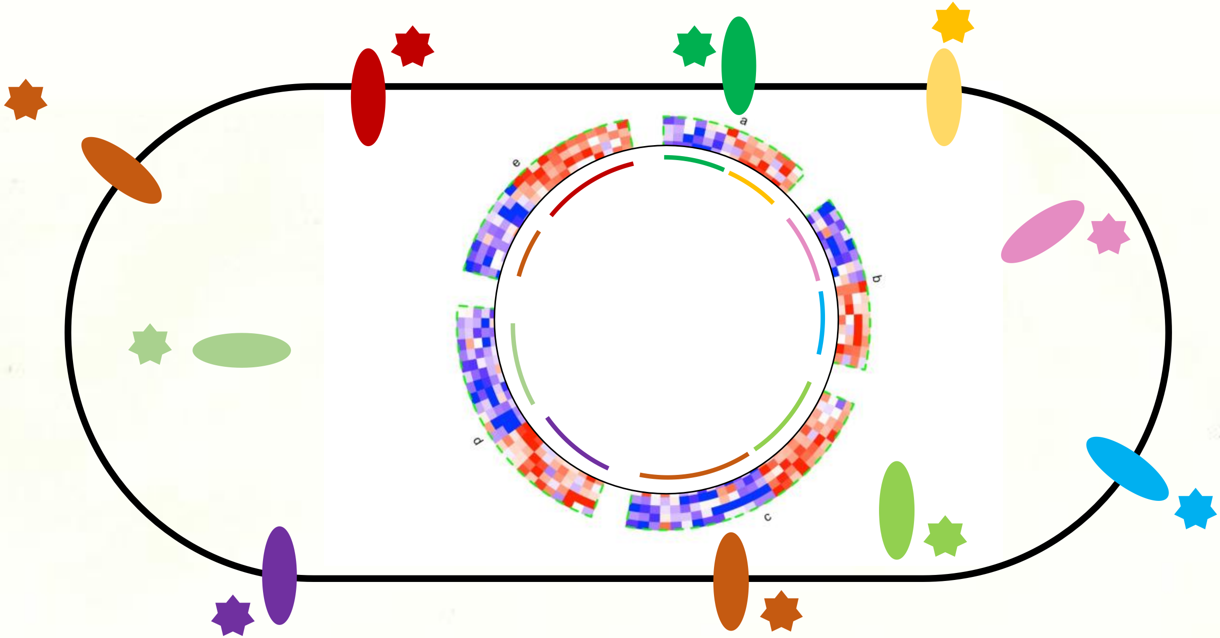


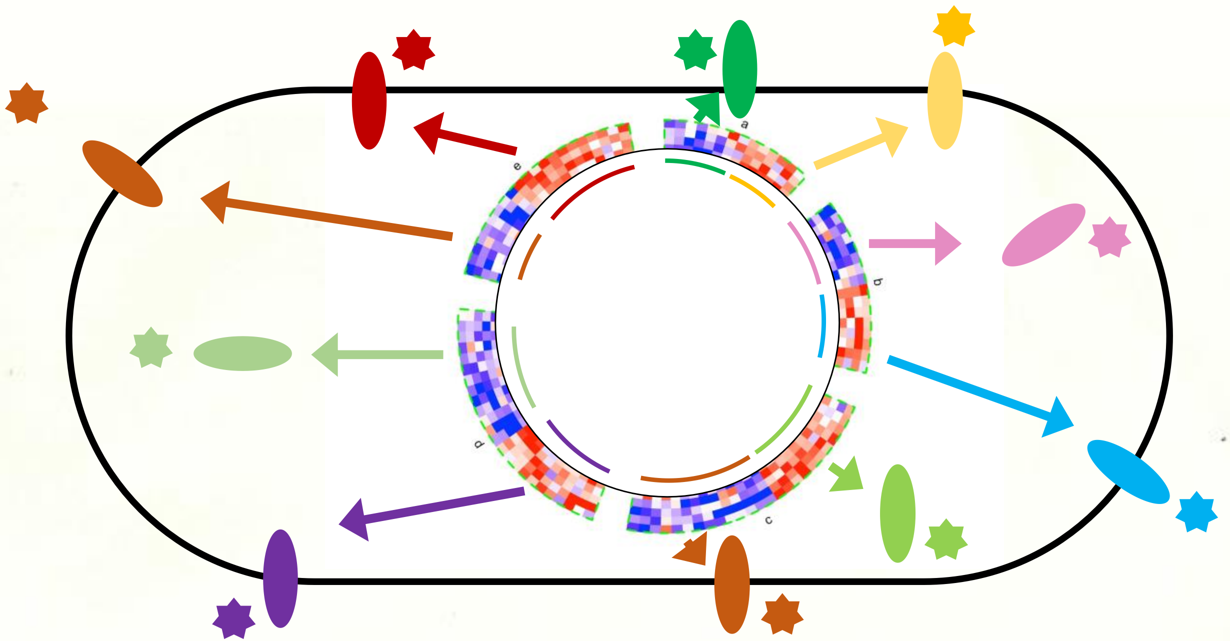

DNA (genes)

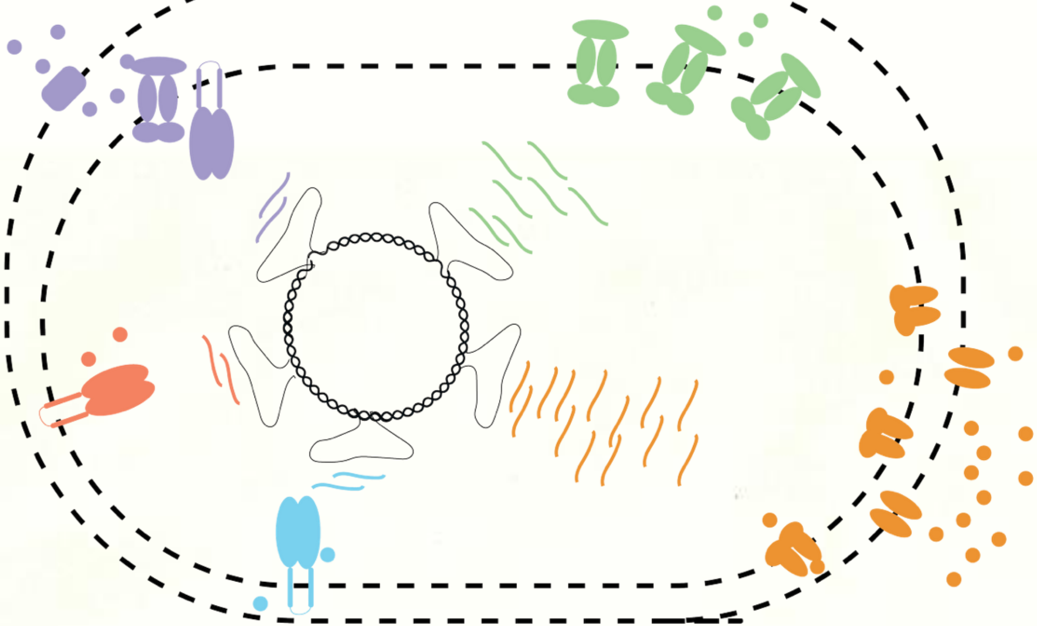

DNA (genes)
Transcription
RNA Polymerase
mRNA

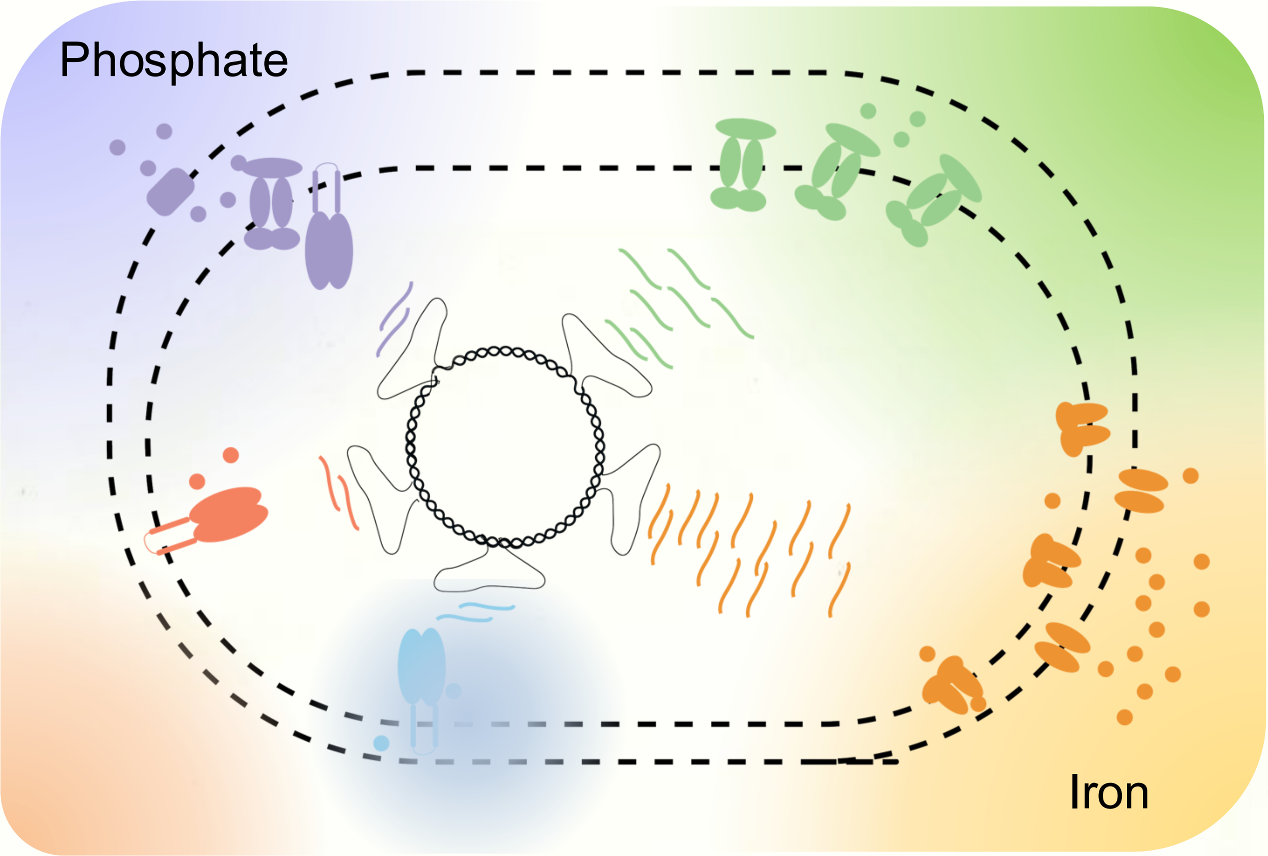

DNA
Transcription
RNA Polymerase
mRNA
Translation
Ribosome
Protein

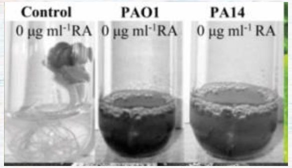

DNA
RNA Polymerase
RNA Polymerase
mRNA

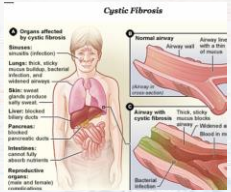

a
b
c
d
e

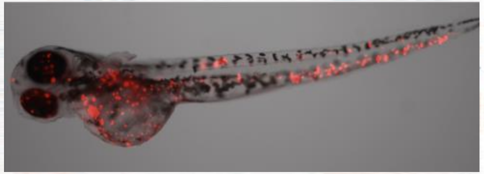

a
b
c
d
e

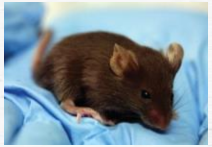


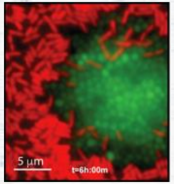

Phosphate
Iron

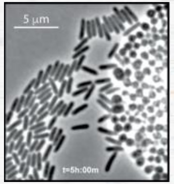

Control
PAO1
PA14
0 μg ml-1 RA
0 μg ml-1 RA
0 μg ml-1 RA

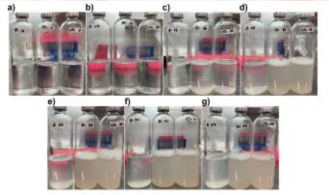

Cystic Fibrosis
A Organs affected by cystic fibrosis
Sinuses: sinusitis (infection)
Lungs: thick, sticky mucus buildup, bacterial infection, and widened airways
Skin: sweat glands produce salty sweat.
Liver: blocked biliary ducts
Pancreas: blocked pancreatic ducts
Intestines: cannot fully absorb nutrients
Reproductive organs: (male and female) complications
B Normal airway
Airway wall
Airway line with a thin of mucus
(Airway in cross-section)
C Airway with cystic fibrosis
Thick, sticky mucus blocks airway
Widened a
Blood in m
Bacterial infection

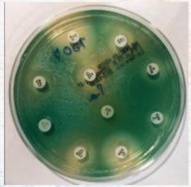


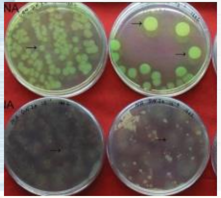


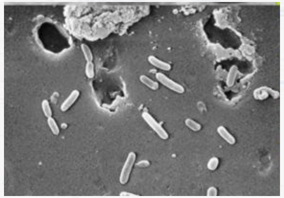

5 μm
t=6h:00m


5 μm
t=5h:00m

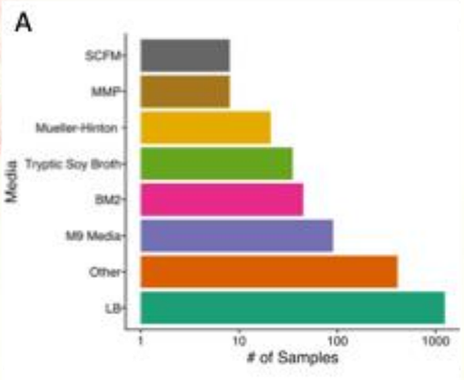

a)
b)
c)
d)
e)
f)
g)

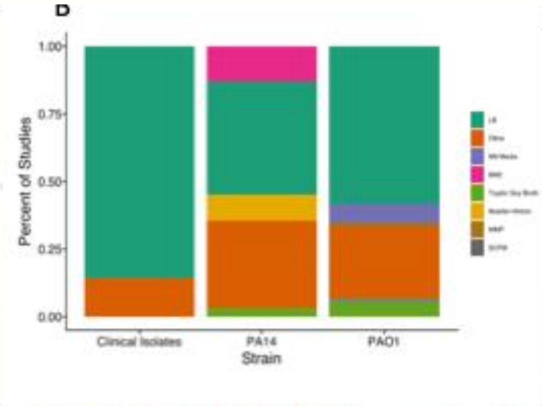


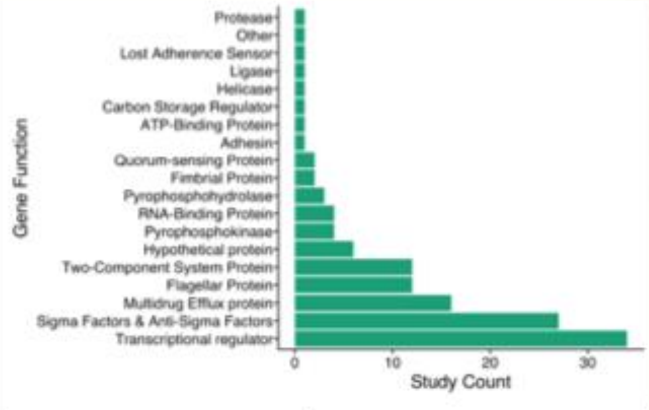

NA
A

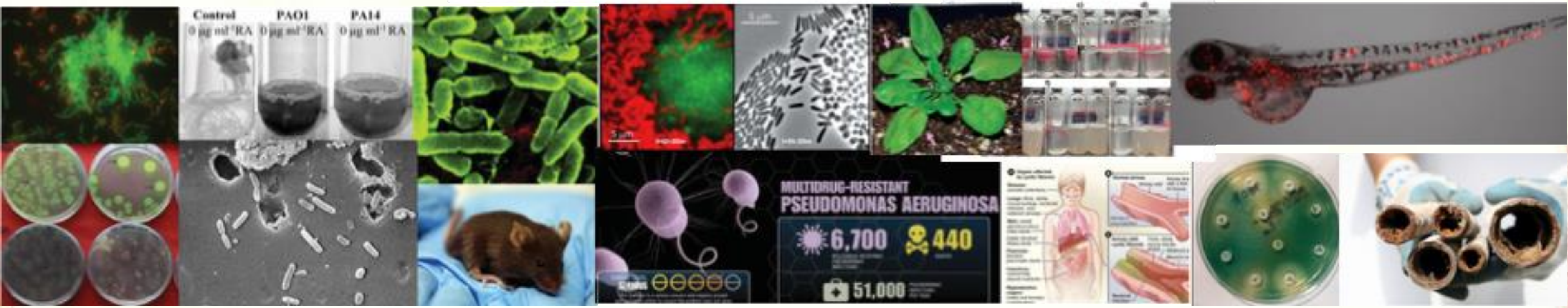


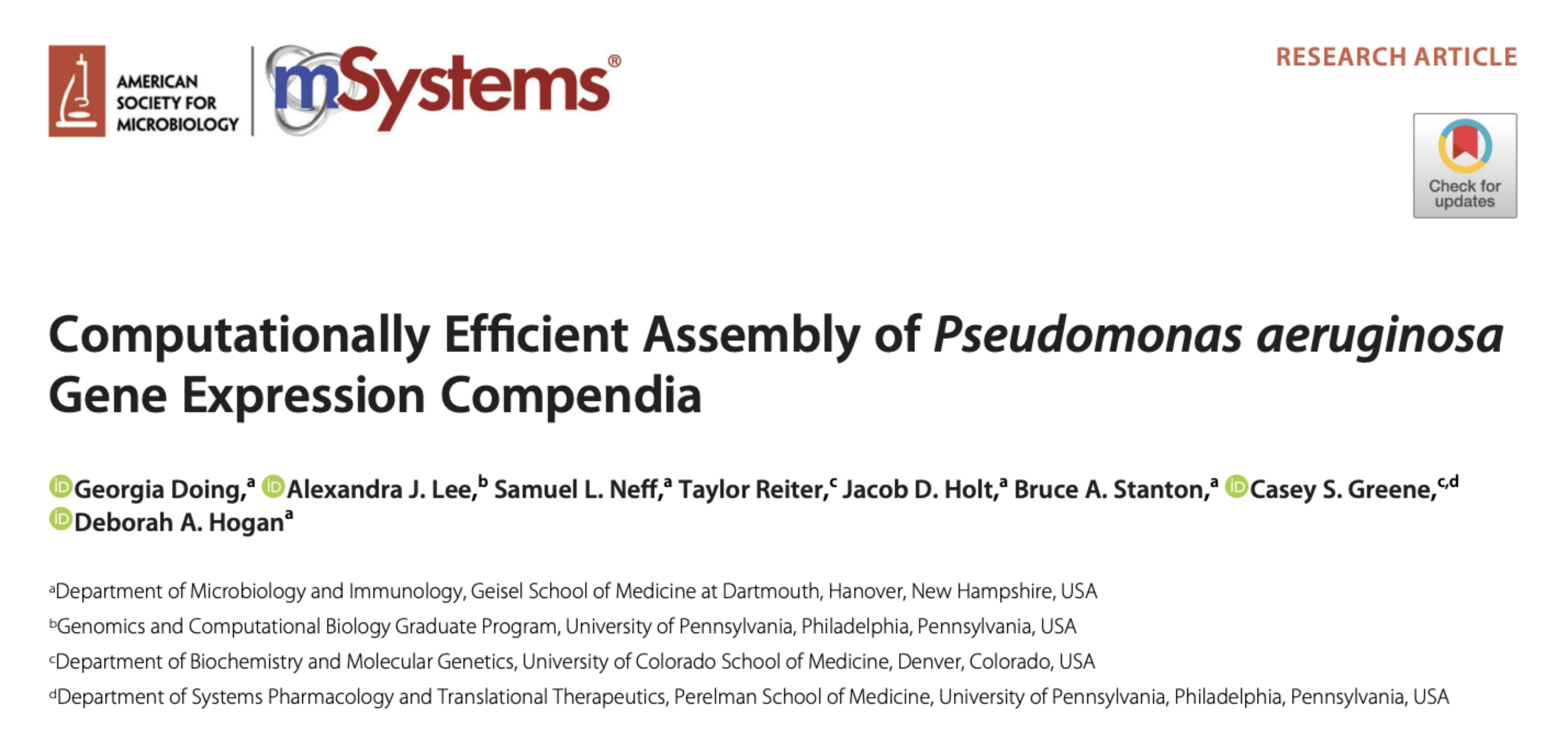
MULTIDRUG-RESISTANT
PSEUDOMONAS AERUGINOSA
6,700
MULTIDRUG-RESISTANT PSEUDOMONAS INFECTIONS
440
DEATHS
51,000
PSEUDOMONAS INFECTIONS PER YEAR
THREAT LEVEL
SERIOUS
This bacteria is a serious concern and requires prompt and sustained action to ensure the problem does not grow.


A
SCFM
MMP
Mueller-Hinton
Tryptic Soy Broth
BM2
M9 Media
Other
LB
Media
1
10
100
1000
# of Samples


B
1.00
0.75
0.50
0.25
0.00
Percent of Studies
Clinical Isolates
PA14
PAO1
Strain
LB
Other
M9 Media
BM2
Tryptic Soy Broth
Mueller-Hinton
MMP
SCFM


Protease
Other
Lost Adherence Sensor
Ligase
Helicase
Carbon Storage Regulator
ATP-Binding Protein
Adhesin
Quorum-sensing Protein
Fimbrial Protein
Pyrophosphohydrolase
RNA-Binding Protein
Pyrophosphokinase
Hypothetical protein
Two-Component System Protein
Flagellar Protein
Multidrug Efflux protein
Sigma Factors & Anti-Sigma Factors
Transcriptional regulator
Gene Function
0
10
20
30
Study Count


Control
PAO1
PA14
0 μg ml⁻¹ RA
0 μg ml⁻¹ RA
0 μg ml⁻¹ RA
5 μm
MULTIDRUG-RESISTANT
PSEUDOMONAS AERUGINOSA
6,700
440
51,000



AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

RESEARCH ARTICLE

Check for updates

Computationally Efficient Assembly of *Pseudomonas aeruginosa* Gene Expression Compendia

Georgia Doing,[a] Alexandra J. Lee,[b] Samuel L. Neff,[a] Taylor Reiter,[c] Jacob D. Holt,[a] Bruce A. Stanton,[a] Casey S. Greene,[c,d] Deborah A. Hogan[a]

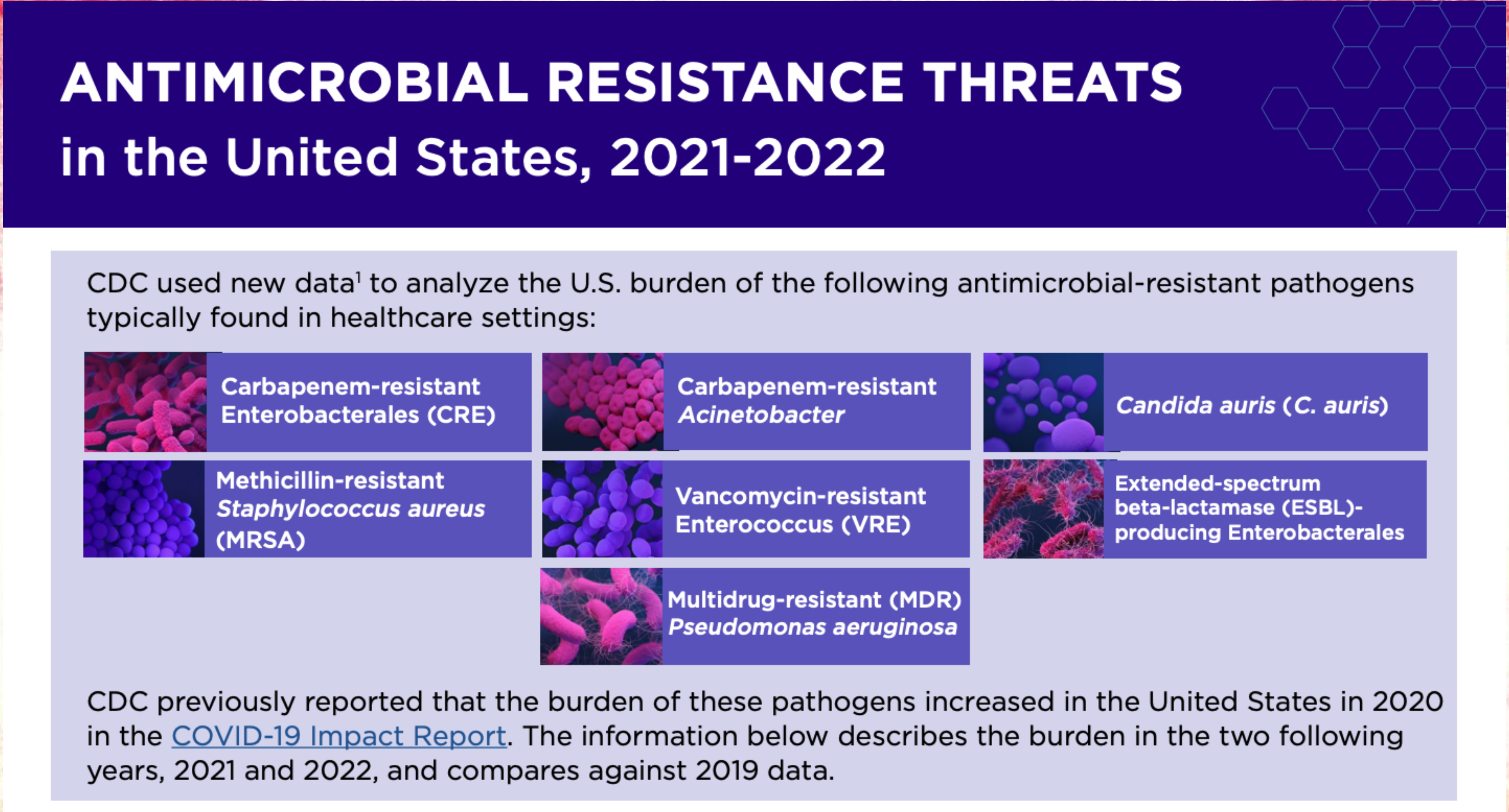
[a]Department of Microbiology and Immunology, Geisel School of Medicine at Dartmouth, Hanover, New Hampshire, USA

[b]Genomics and Computational Biology Graduate Program, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[c]Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine, Denver, Colorado, USA

[d]Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA

https://pubmed.ncbi.nlm.nih.gov/36541761/



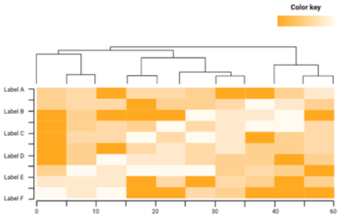
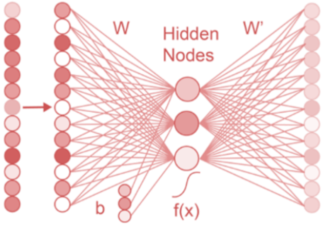
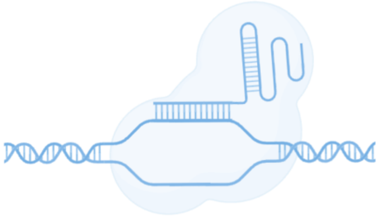
ANTIMICROBIAL RESISTANCE THREATS in the United States, 2021-2022

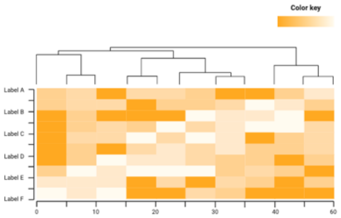
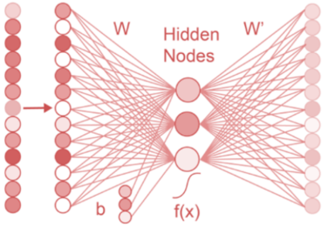
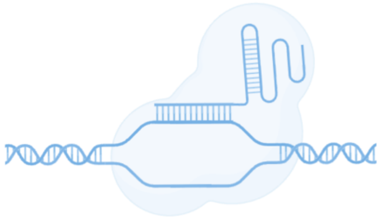
CDC used new data[1] to analyze the U.S. burden of the following antimicrobial-resistant pathogens typically found in healthcare settings:

- Carbapenem-resistant Enterobacterales (CRE)
- Carbapenem-resistant *Acinetobacter*
- *Candida auris (C. auris)*
- Methicillin-resistant *Staphylococcus aureus* (MRSA)
- Vancomycin-resistant Enterococcus (VRE)
- Extended-spectrum beta-lactamase (ESBL)-producing Enterobacterales
- Multidrug-resistant (MDR) *Pseudomonas aeruginosa*

CDC previously reported that the burden of these pathogens increased in the United States in 2020 in the COVID-19 Impact Report. The information below describes the burden in the two following years, 2021 and 2022, and compares against 2019 data.

CDC https://www.cdc.gov/ecoli/about/index.html

SRA Search	Compendia	DAE Models	Bi-partite Graphs	Comparative Genetics
				
• organism: *E. coli* • molecule: mRNA • pull data from Short Read Archive (SRA) and Gene Expression Omnibus (GEO)	• metadata • *E. coli* (20 strains) reference pan-genomes • salmon k-mer mapping (k=15)	• denoising autoencoder (tied) • 1 layer of 50 hidden nodes • fully connected • objective): reconstruction error	• G(E,V) where E ∈ $\text{Corr}_{\text{gene, gene}}$ and V ∈ $\text{Genes}_{\text{S.e.}} \cup \text{Genes}_{\text{S.a.}}$ • Weight edges by correlation strength • Color genes by homology	• knock-down genes of interest under function-relevant conditions • look for fitness effects and/or phenotypes

SRA Search	Compendia	DAE Models	Bi-partite Graphs	Comparative Genetics
	 	 		
- organism: *E. coli* - molecule: DNA, RNA or text - pull data from Short Read Archive (SRA) and Gene Expression Omnibus (GEO)	- metadata - subset of strains and samples reference pan-genomes - salmon k-mer mapping (k=15)	Dimensionality reduction: PCA ICA NMD DAE	Pattern detection: correlation regression clustering classification	Hypothesis testing: CRISPR data phenotypes (future directions)

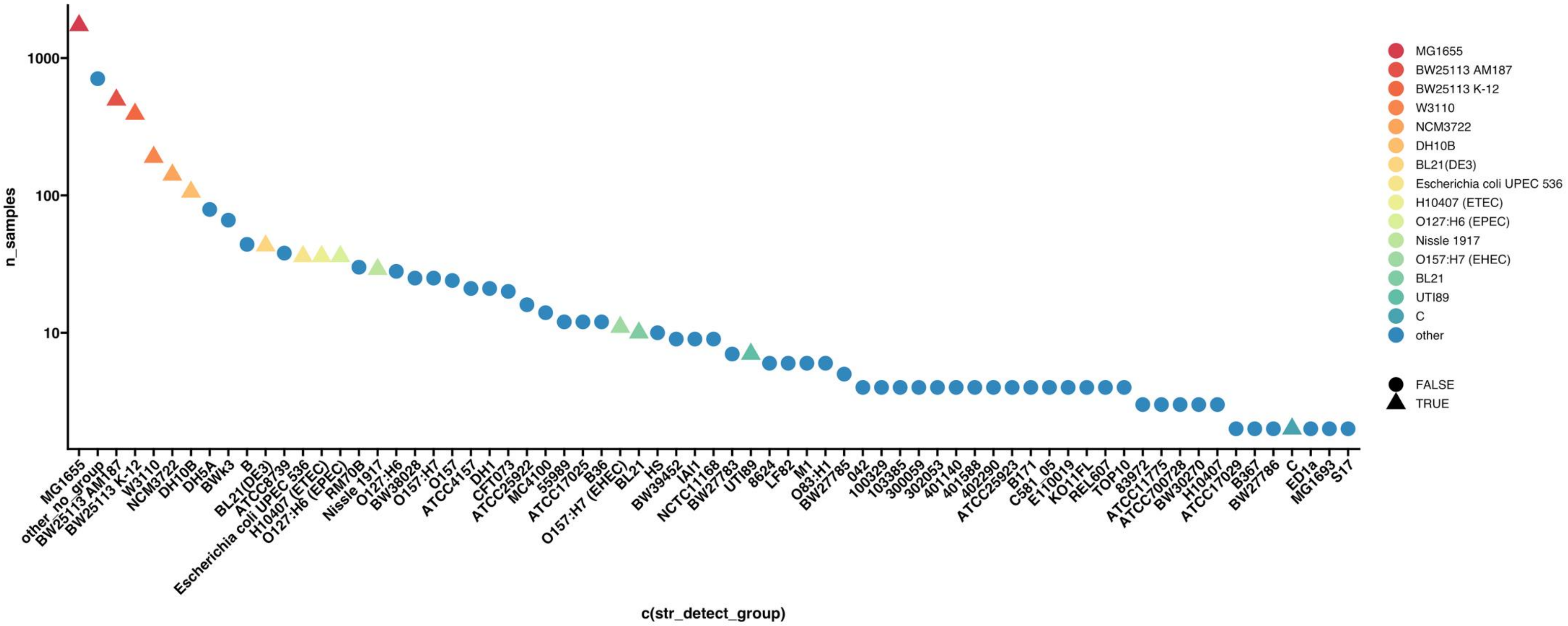
Ecoli Pangenome Compendium

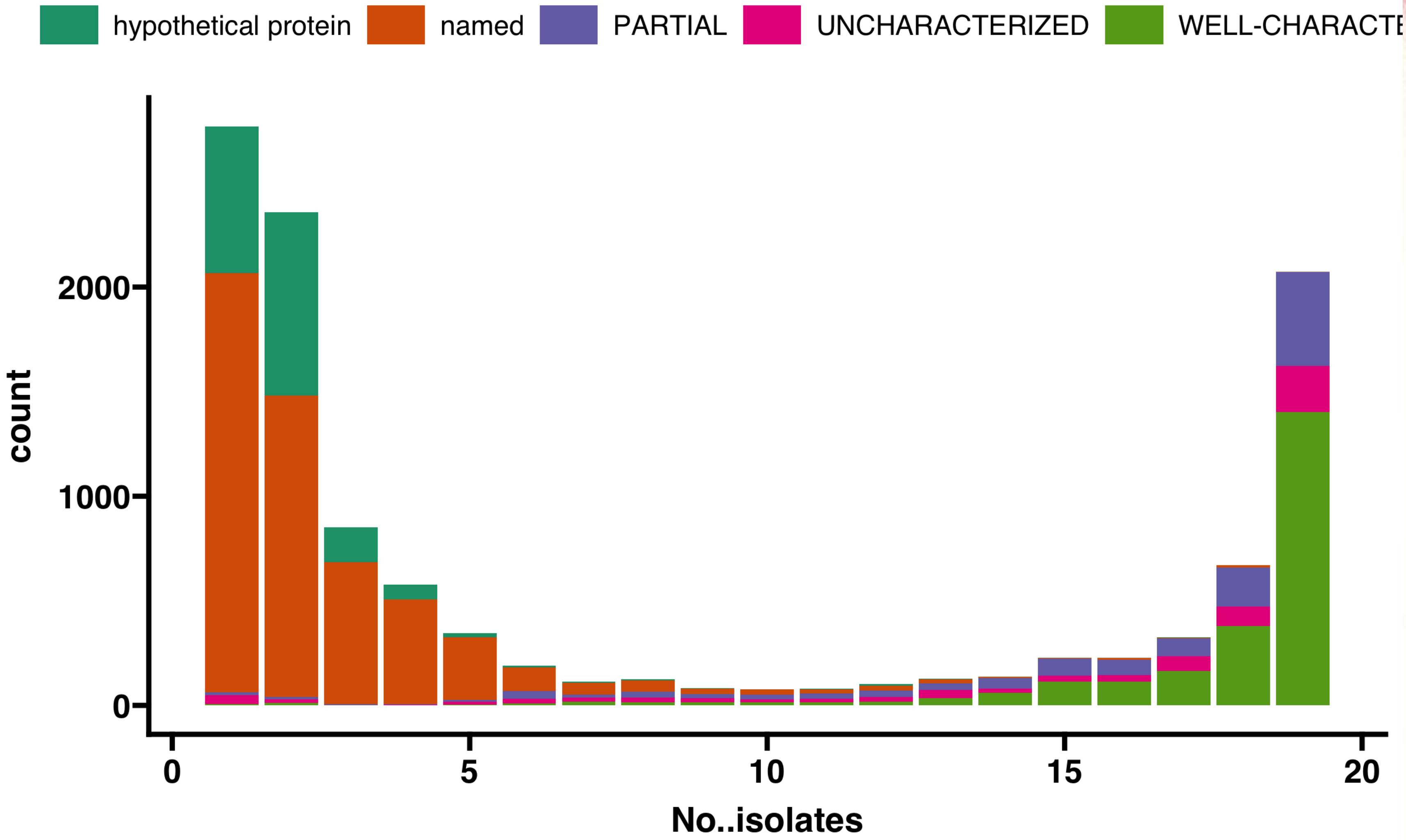
samples: 20,292

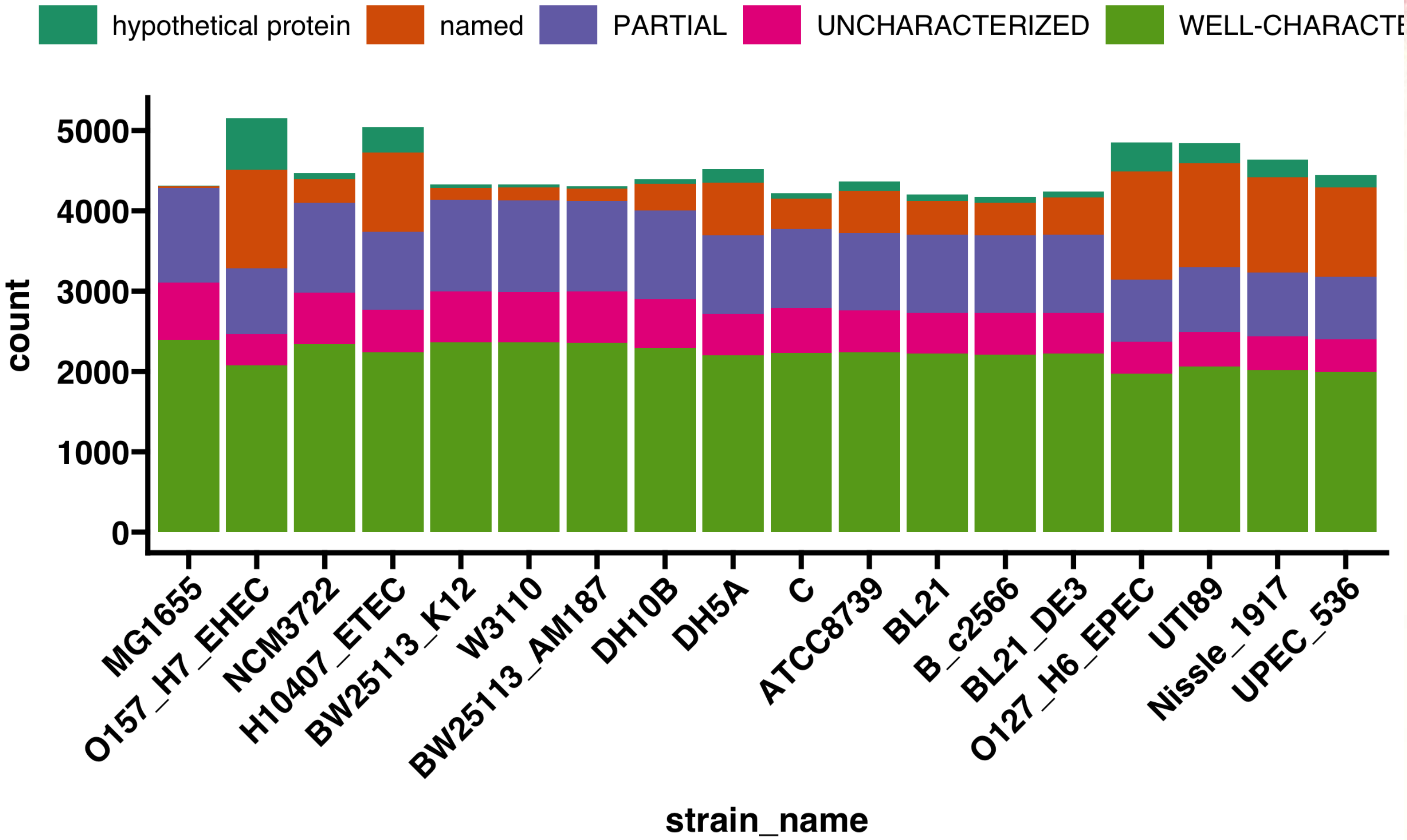
unique studies: 1,365

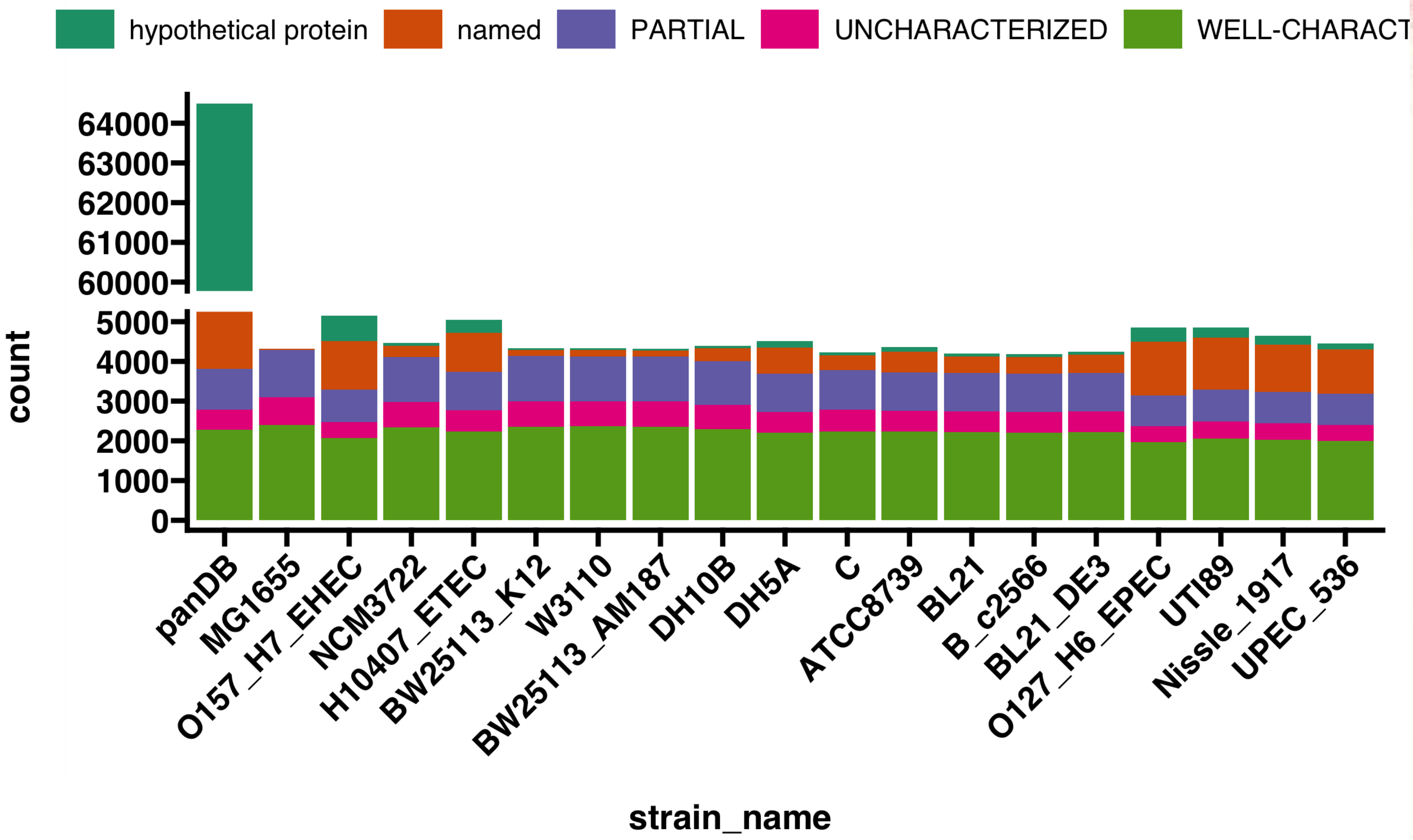
samples with high confidence strain: 4,260

samples with putative strain: 11,859


n_samples
1000
100
10
c(str_detect_group)
MG1655
other_no_group
BW25113 AM187
BW25113 K-12
W3110
NCM3722
DH10B
DH5A
BWk3
B
BL21(DE3)
ATCC8739
Escherichia coli UPEC 536
H10407 (ETEC)
O127:H6 (EPEC)
RM70B
Nissle 1917
O127:H6
BW38028
O157:H7
O157
ATCC4157
DH1
CFT073
ATCC25922
MC4100
55989
ATCC17025
B36
O157:H7 (EHEC)
BL21
HS
BW39452
IAI1
NCTC11168
BW27783
UTI89
8624
LF82
M1
O83:H1
BW27785
042
100329
103385
300059
302053
401140
401588
402290
ATCC25923
B171
C581_05
E110019
KO11FL
REL607
TOP10
83972
ATCC11775
ATCC700728
BW30270
H10407
ATCC17029
B367
BW27786
C
ED1a
MG1693
S17
● MG1655
● BW25113 AM187
● BW25113 K-12
● W3110
● NCM3722
● DH10B
● BL21(DE3)
● Escherichia coli UPEC 536
● H10407 (ETEC)
● O127:H6 (EPEC)
● Nissle 1917
● O157:H7 (EHEC)
● BL21
● UTI89
● C
● other
● FALSE
▲ TRUE

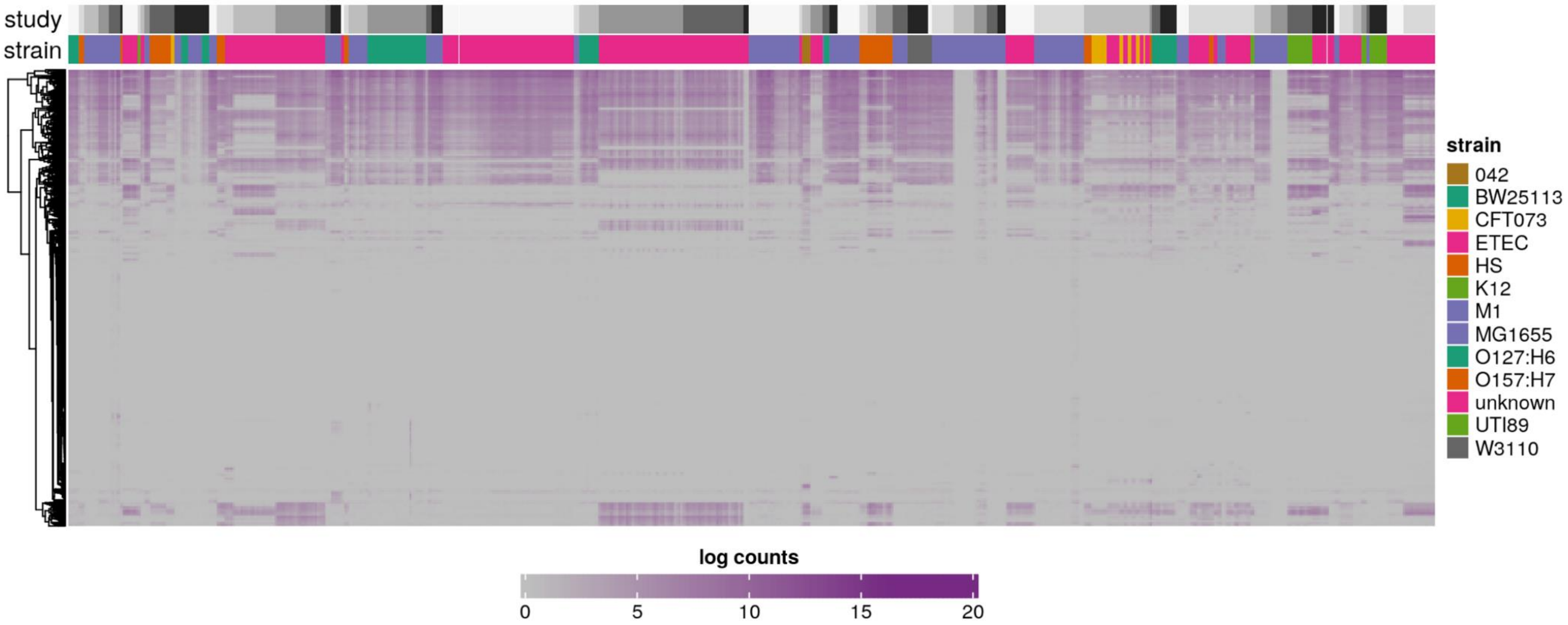

hypothetical protein
named
PARTIAL
UNCHARACTERIZED
WELL-CHARACTE
count
2000
1000
0
0
5
10
15
20
No..isolates

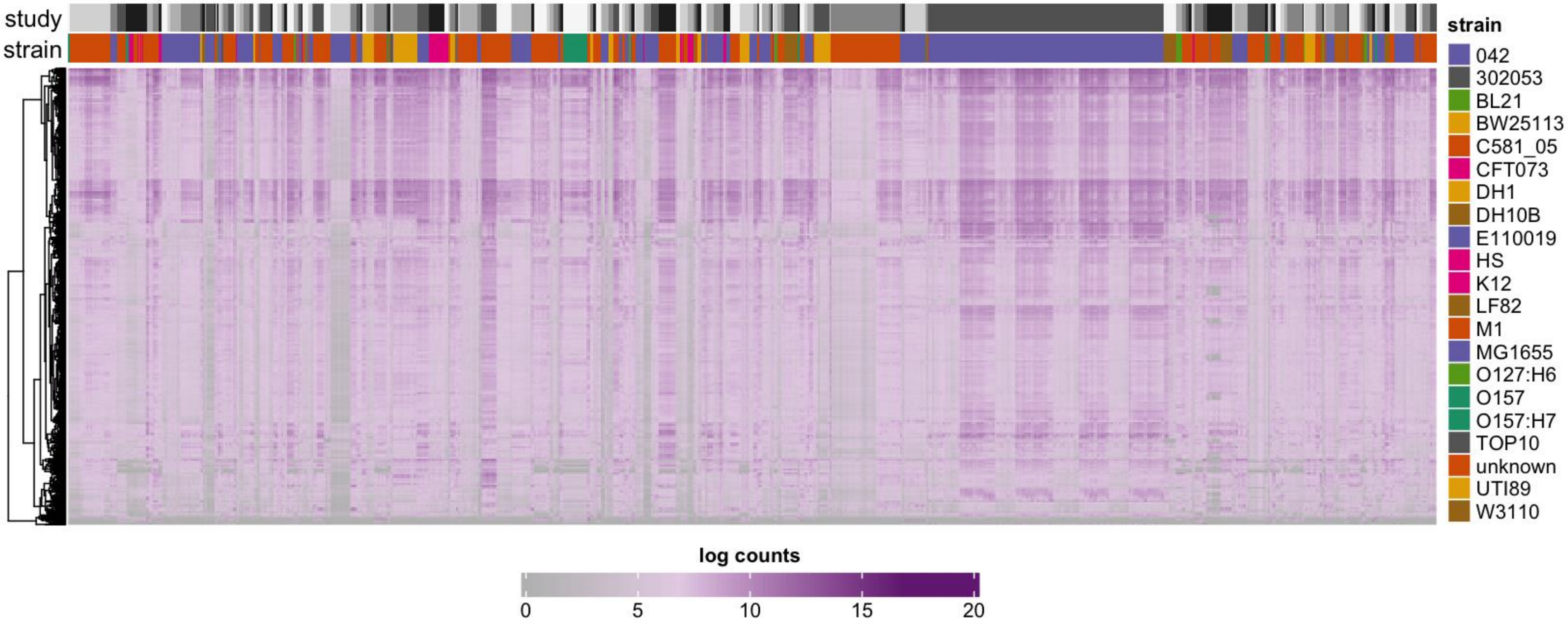

hypothetical protein
named
PARTIAL
UNCHARACTERIZED
WELL-CHARACTE
count
5000
4000
3000
2000
1000
0
MG1655
O157_H7_EHEC
NCM3722
H10407_ETEC
BW25113_K12
W3110
BW25113_AM187
DH10B
DH5A
C
ATCC8739
BL21
B_c2566
BL21_DE3
O127_H6_EPEC
UTI89
Nissle_1917
UPEC_536
strain_name


hypothetical protein
named
PARTIAL
UNCHARACTERIZED
WELL-CHARACT
count
64000
63000
62000
61000
60000
5000
4000
3000
2000
1000
0
panDB
MG1655
O157_H7_EHEC
NCM3722
H10407_ETEC
BW25113_K12
W3110
BW25113_AM187
DH10B
DH5A
C
ATCC8739
BL21
B_c2566
BL21_DE3
O127_H6_EPEC
UTI89
Nissle_1917
UPEC_536
strain_name

Head of “count” table

X	Name	ERX1805690	ERX1805691	ERX1805692	ERX2279850	ERX2279851	ERX2279852	ERX2279853	ERX2279854
0ed86576d343ce011a61773e0620e335_1045	character(0)	1.000	0.000	0.00	0.000	0.000	0.000	0.000	0.000
0ed86576d343ce011a61773e0620e335_1047	character(0)	0.000	0.000	0.00	0.000	0.000	0.000	0.000	0.000
0ed86576d343ce011a61773e0620e335_1120	character(0)	0.000	0.000	0.00	6.476	6.236	0.000	0.000	7.554
0ed86576d343ce011a61773e0620e335_1175	character(0)	0.000	0.000	0.00	2.000	2.000	3.000	2.000	0.000
0ed86576d343ce011a61773e0620e335_1179	character(0)	0.000	0.000	0.00	2.000	4.000	0.000	4.000	2.000
0ed86576d343ce011a61773e0620e335_1311	character(0)	0.000	0.000	0.00	0.000	0.000	0.000	0.000	0.000
0ed86576d343ce011a61773e0620e335_1450	character(0)	0.000	0.000	0.00	3.000	0.000	0.000	0.000	2.000
0ed86576d343ce011a61773e0620e335_1476	character(0)	5.403	2.351	0.00	15.370	11.212	23.317	5.232	12.820
0ed86576d343ce011a61773e0620e335_1484	character(0)	0.000	9.229	10.55	2.000	2.000	0.000	0.000	3.000
0ed86576d343ce011a61773e0620e335_2071	character(0)	0.000	0.000	0.00	0.000	0.000	36.005	0.000	0.000


study
strain
strain
042
BW25113
CFT073
ETEC
HS
K12
M1
MG1655
O127:H6
O157:H7
unknown
UTI89
W3110
log counts
0
5
10
15
20


study
strain
strain
042
302053
BL21
BW25113
C581_05
CFT073
DH1
DH10B
E110019
HS
K12
LF82
M1
MG1655
O127:H6
O157
O157:H7
TOP10
unknown
UTI89
W3110
log counts
0
5
10
15
20

From DEseq2, median of ratios normalization:

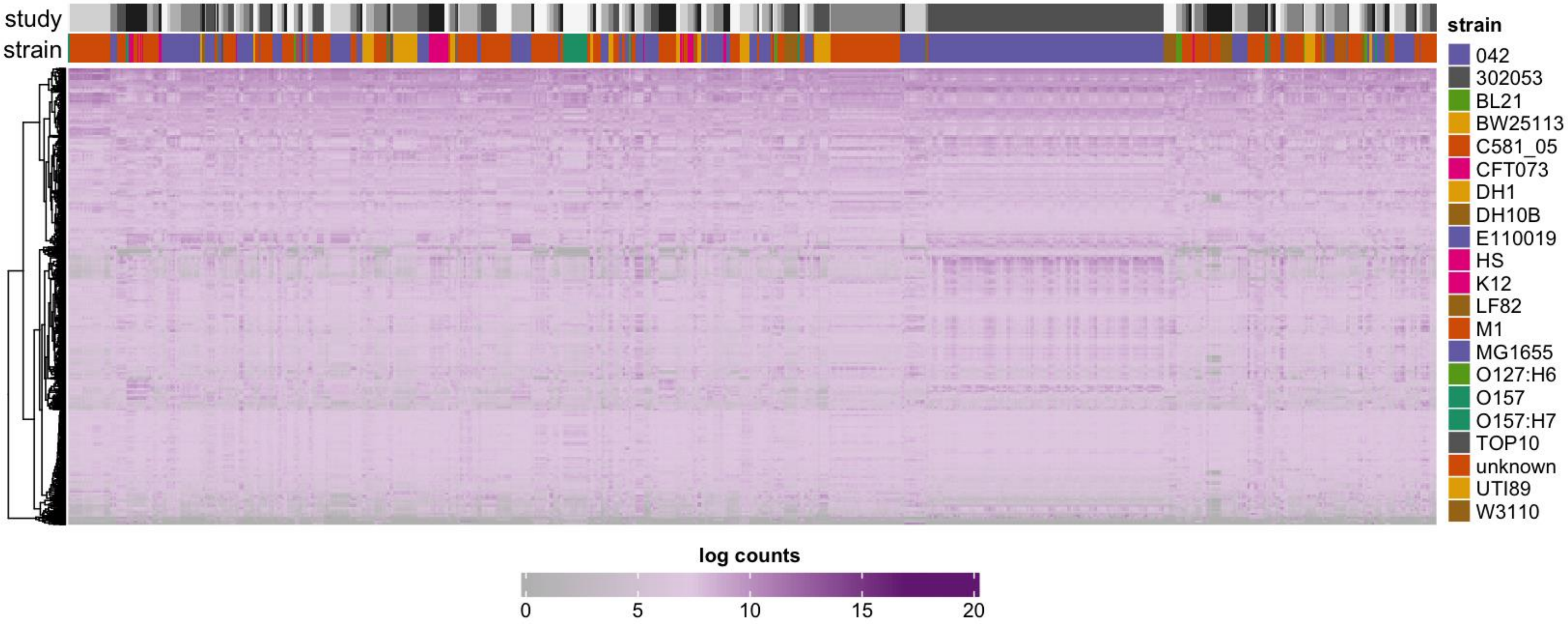
(Y_{ij}): The raw count for gene (i) in sample (j).
(g_i) :The geometric mean of counts for gene (i) across all samples.
The size factor for sample (j) ,denoted as (S_j) ,is calculated as:

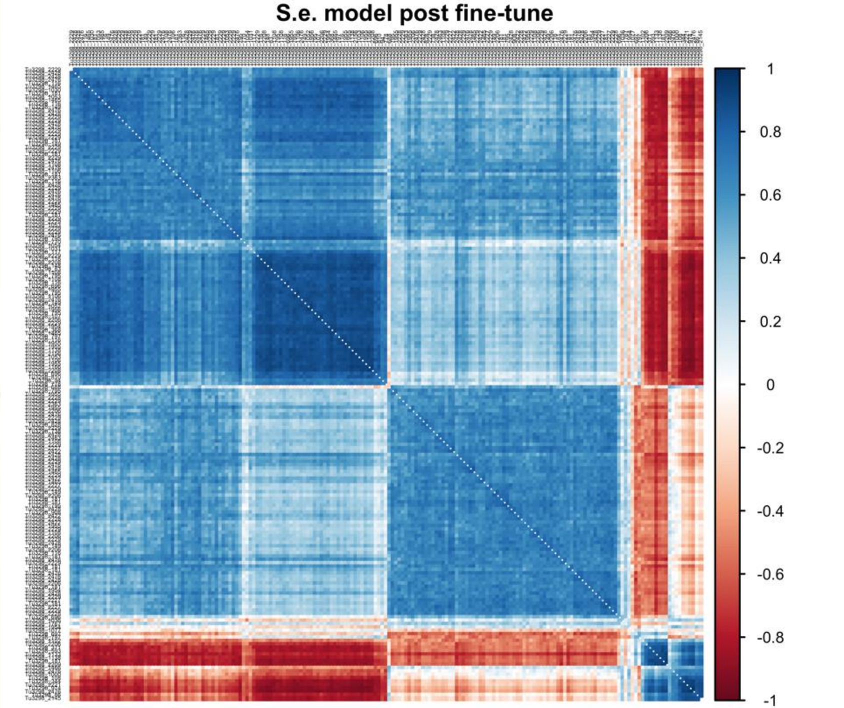
$$\left[S_j = \text{median}\left(\frac{Y_{ij}}{g_i}\right)_{i\in 1,\ldots,m}\right]$$

Where:

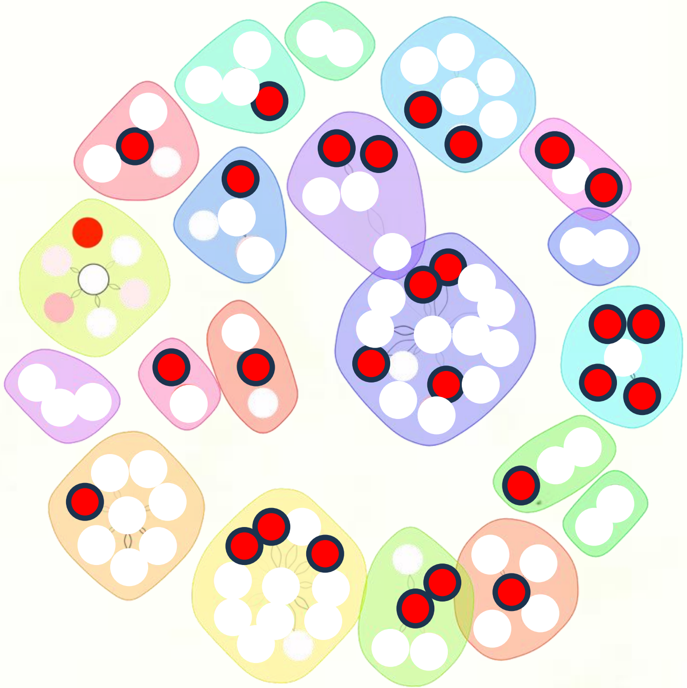
$\left(g_i = \left(\prod_{j=1}^{n} Y\, ij\right)^{1/n}\right)$,the geometric mean of counts for gene (i) across (n) samples.

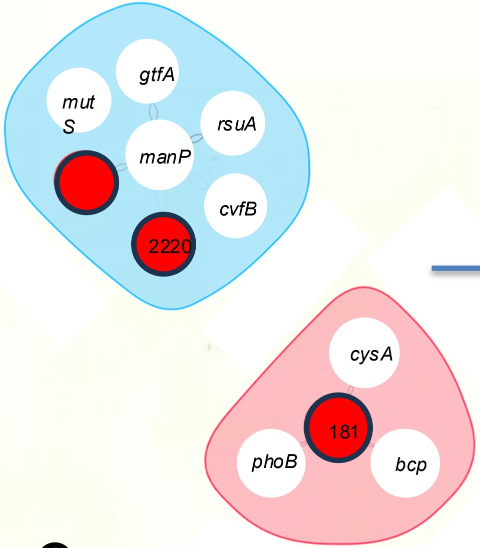
The ratio $\left(\frac{Y_{ij}}{g_i}\right)$ is computed for each gene (i) in sample (j) .The median of these ratios is taken as the size factor (S_j) for sample (j).


study
strain
strain
042
302053
BL21
BW25113
C581_05
CFT073
DH1
DH10B
E110019
HS
K12
LF82
M1
MG1655
O127:H6
O157
O157:H7
TOP10
unknown
UTI89
W3110
log counts
0
5
10
15
20



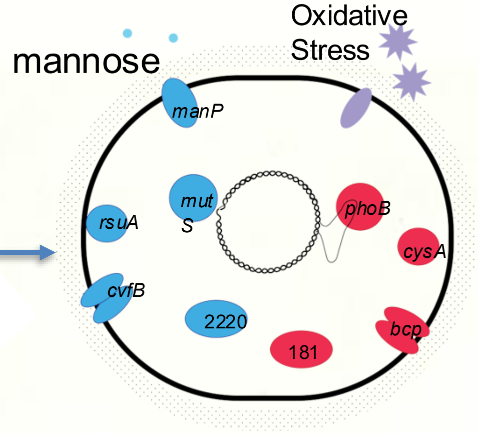
S.e. model post fine-tune

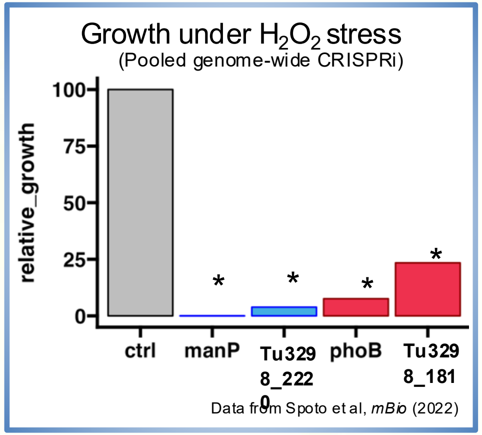




● Uncharacterized gene

○ Gene with known function




Growth under H_2O_2 stress
(Pooled genome-wide CRISPRi)

Data from Spoto et al, *mBio* (2022)

Lab machines: getting started

- Log on:
  - Your Union usernames
  - Default password: the word union followed by your ID number; e.g., union12345

Always log off before leaving the lab