DARPA Innovation Fellowship & Advanced Research Concepts

Ms. Ana Saplan, ARC Manager

Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Defense Sciences Office: “DARPA’s DARPA”

- Creates opportunities from scientific discovery
- Invests in multiple scientific disciplines
- Focuses on mission-informed research

DSO: Creating scientific surprise to support national security

Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)

Finding Great DARPA Ideas

Improve access to innovation from a diverse group of organizations

- With support, small technology companies and universities are more likely to be aggressive in pushing capabilities forward. Their products are the ideas they generate that can turn into prototypes. There are lots of ideas in the world, a few are good, while true DARPA ideas are rare. Need to fund as many as possible, and quickly, to find the pearls.
- Need to be efficient. Just spending money will not achieve the desired results.

Connect to new talent pools

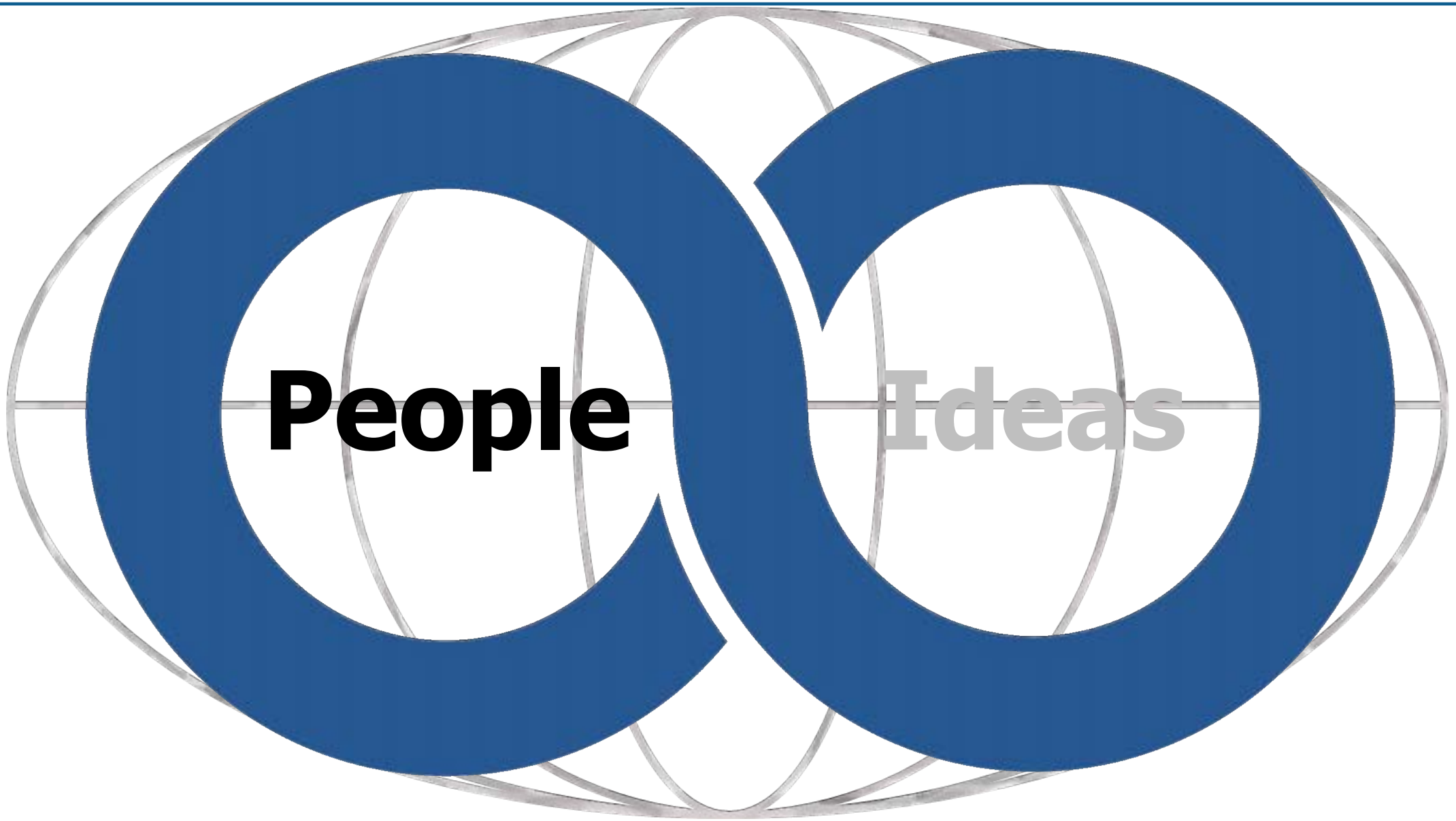
- Paradigm shifts in technology often come from people who are not so deeply indoctrinated in established theories.

Find innovation from VC focused companies

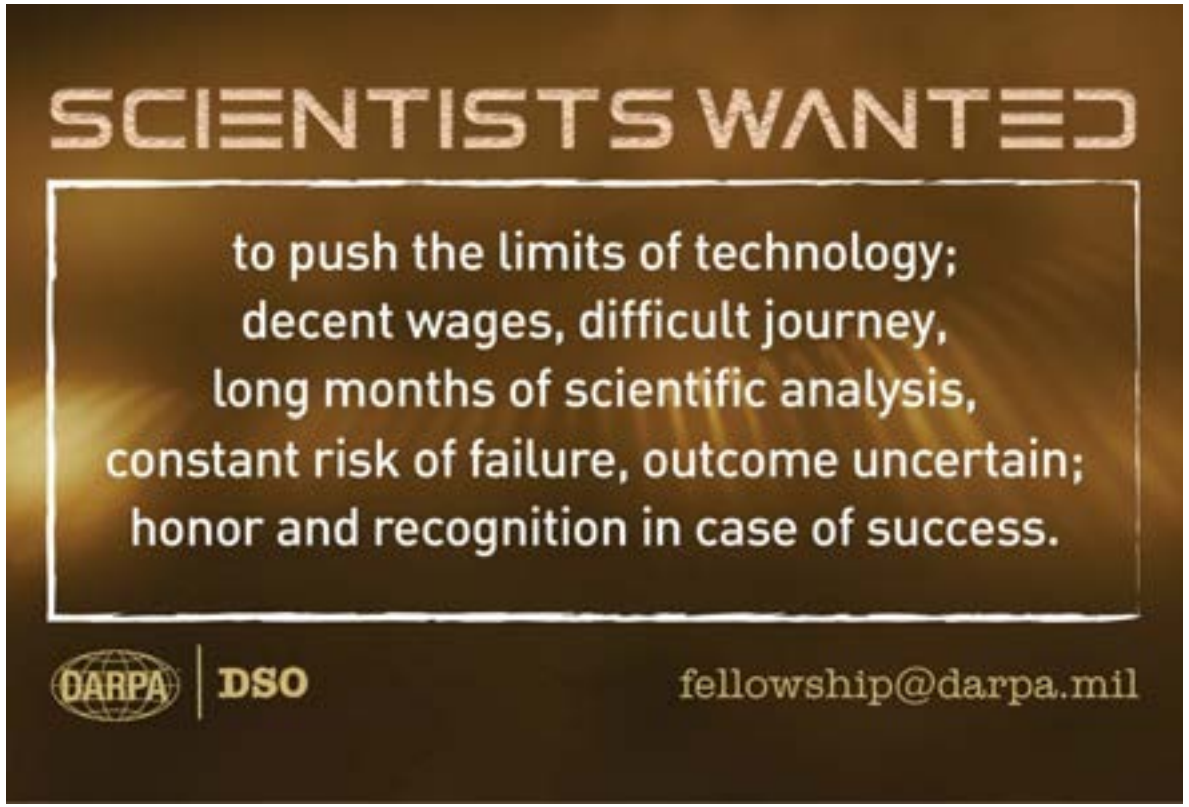
- Forge connections with these small entities at the beginning while they are wide-eyed about changing the world with innovative technologies to advance warfighter needs.

Next breakthrough, paradigm-shifting technology

Finding Great DARPA Ideas



DARPA Innovation Fellowship



- 2-year Fellowship for early career scientists
- 32 recent Ph.D. graduates and 8 active duty military
- Develop and manage a portfolio of high-impact exploratory efforts
- Paradigm shifts in technology often come from those not deeply indoctrinated in established theories
- Build a long-term pool of diverse talent that can focus on national security

Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

DARPA Innovation Fellowship

What is the Innovation Fellowship?

A 2-year Fellowship at DARPA for early career scientists, who received their Ph.D. within the last 5 years. Fellows develop and manage the Advanced Research Concepts (ARC), a portfolio of high-impact exploratory efforts to identify breakthrough technologies for the Department of Defense.

Why become an Innovation Fellow?

Drive technological innovation

Fellows have the opportunity to influence the direction of defense research through developing ARC topics, evaluating proposals, making funding decisions, and assessing the impact of further investment on problems of importance to national security.

Engage with prominent scientists

Fellows travel across the country to visit leading researchers at top university, industry, and government labs and learn about the revolutionary research they are conducting.

Strengthen your transferable skills

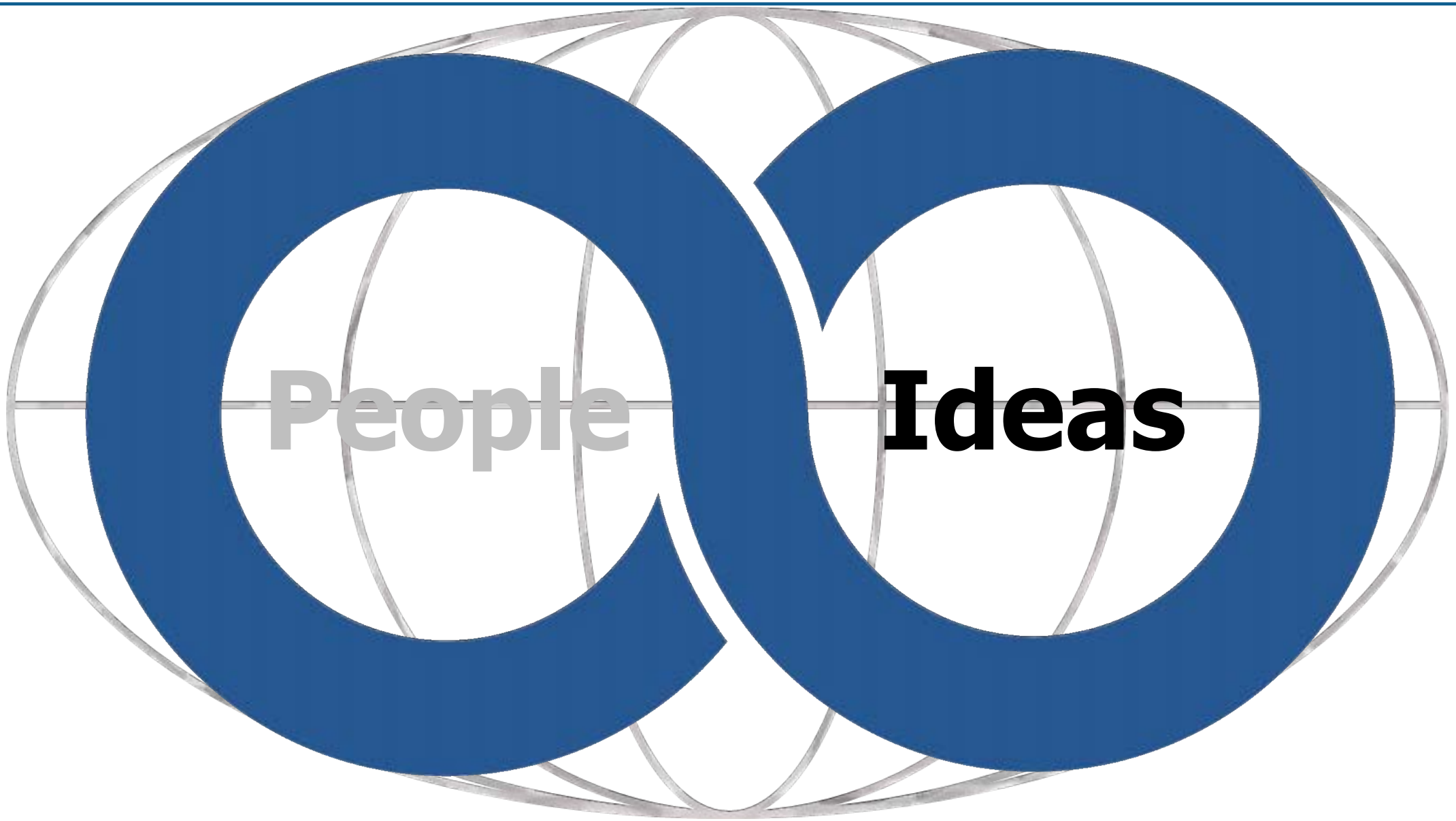
Fellows work across a broad range of scientific fields and gain a deep understanding of the big-picture scope of the state of the art of science and technology.

Advance your career opportunities

Join an extraordinarily rich, technologically-focused network of DARPA Program Managers, military service members, and scientific and technical experts.

Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Finding Great DARPA Ideas



Advanced Research Concepts (ARC) – A DSO Experiment

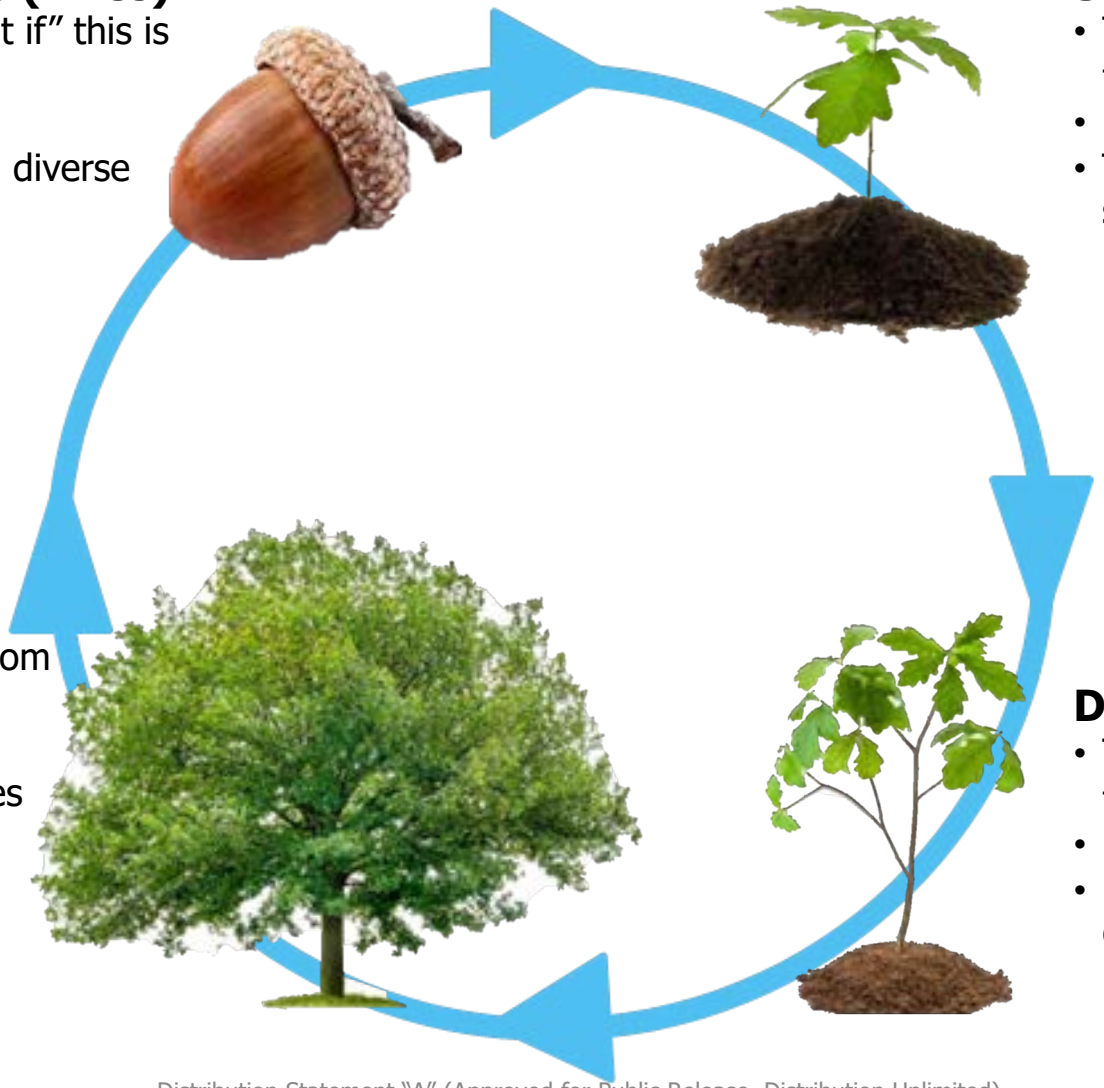
- ARC solicitations will focus on answering high risk/ high-reward “what if?” question
- 8 topics targeted annually
- 30-60 ideas per topic
- One person funded per year per contract
- Streamlined proposal and contracting process

A new process to quickly capture and rigorously evaluate many ideas

Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)

ARCs vs. Seedlings/DOs/Programs

Advanced Research Concepts (ARCs)
- Exploratory efforts to evaluate "what if" this is a possibility
- Effort: 1 year, 1 FTE
- Precise question, broad opportunity, diverse answers



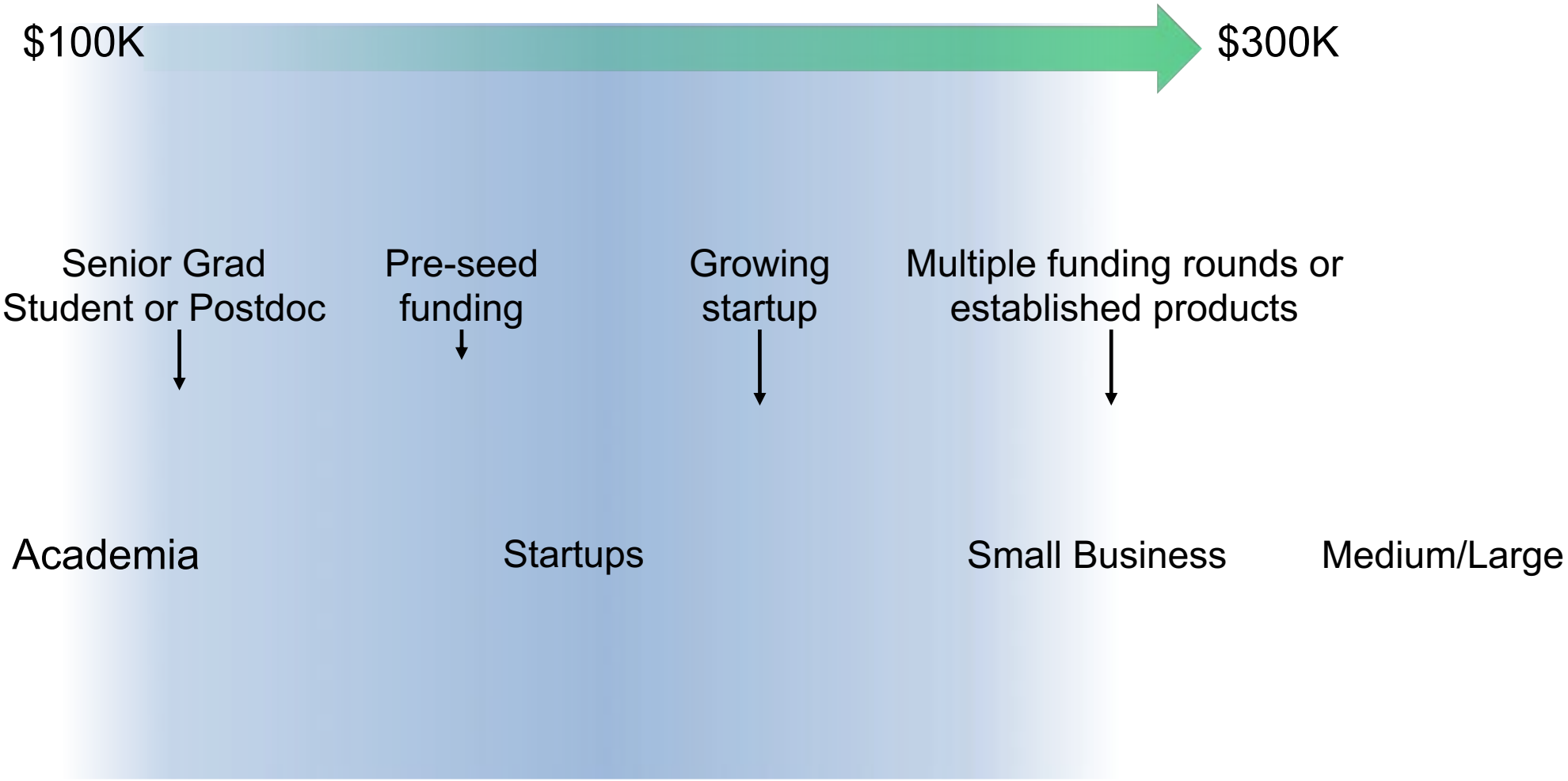
Seedlings
- Technology development to move from "disbelief" to "doubt"
- Effort: 1-2 years, limited personnel
- Target specific problem to enable specific capability

Programs
- Technology development to move from "possibility" to "capability"
- Effort: Multi-year, multi-disciplinary
- Development of capability that scales

Disruptioneerings
- Technology development to move from "disbelief" to "doubt"
- Effort: 2 years, limited personnel
- Expeditated exploration of potential capability development

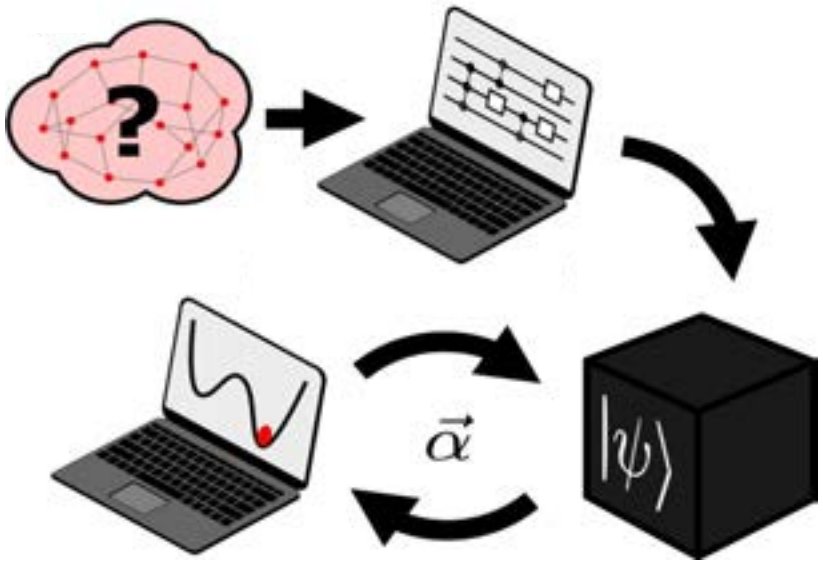
Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

ARC Funding


Funding Range
$100K
$300K
Who
Senior Grad Student or Postdoc
Pre-seed funding
Growing startup
Multiple funding rounds or established products
DARPA Performer
Academia
Startups
Small Business
Medium/Large

Imagining Practical Application for a Quantum Tomorrow (IMPAQT)

- The goal of each ARC is to invest in research that may result in new, game-changing technologies for U.S. national security
- Quantum computing has the potential to bring tremendous advancements to science and could have significant implications for national security
- IMPAQT will explore hybrid classical/quantum computational systems that are expected to be demonstrated within the next several years

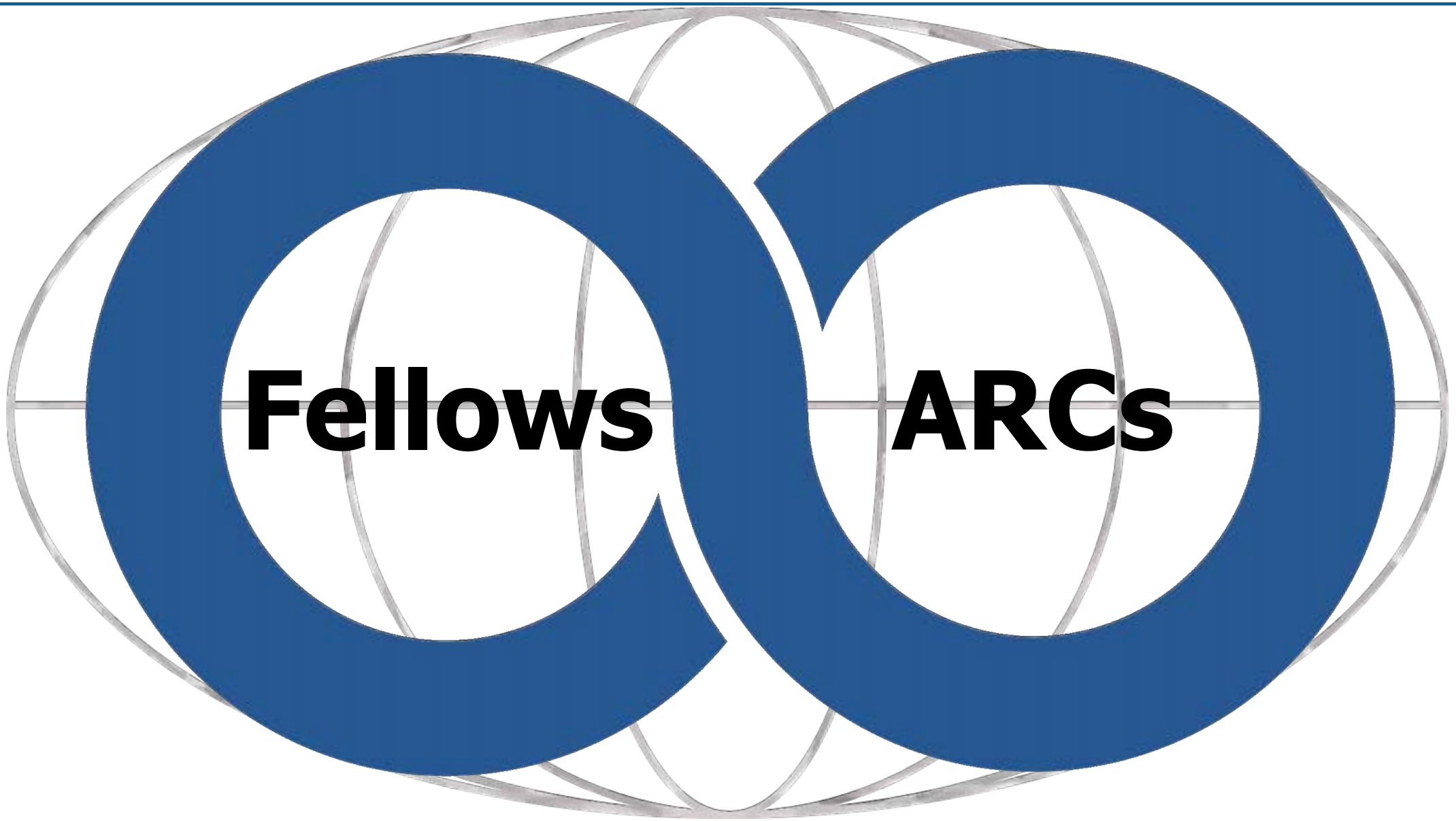


What are the applications for a quantum system with N*q > 10,000, as a co-processor for a classical computational system?

ONISQ: Optimization with Noisy Intermediate-Scale Quantum devices
N*q: Number of Qubits (N) x Circuit Depth (p)

Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Finding Great DARPA Ideas



Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

DARPA Innovation Fellowship

What is the Innovation Fellowship?

A 2-year Fellowship at DARPA for early career scientists, who received their Ph.D. within the last 5 years. Fellows develop and manage the Advanced Research Concepts (ARC), a portfolio of high-impact exploratory efforts to identify breakthrough technologies for the Department of Defense.

Why become an Innovation Fellow?

Drive technological innovation

Fellows have the opportunity to influence the direction of defense research through developing ARC topics, evaluating proposals, making funding decisions, and assessing the impact of further investment on problems of importance to national security.

Engage with prominent scientists

Fellows travel across the country to visit leading researchers at top university, industry, and government labs and learn about the revolutionary research they are conducting.

Strengthen your transferable skills

Fellows work across a broad range of scientific fields and gain a deep understanding of the big-picture scope of the state of the art of science and technology.

Advance your career opportunities

Join an extraordinarily rich, technologically-focused network of DARPA Program Managers, military service members, and scientific and technical experts.

Advanced Research Concepts (ARC)

- Portfolio of fundamental research efforts for assessing the impact of further investment on problems of national security importance.
- Several topics are released per year, each targeting a specific technical area.

www.DARPA.mil/ARC

For more information on the Fellowship visit: https://www.darpa.mil/work-with-us/darpa-innovation-fellowship

To apply submit CV/resume and cover letter to: fellowship@darpa.mil

U.S. citizenship is required

Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Discovering Unknome Function (DUF)

Dr. René Xavier, DARPA Innovation Fellow

Briefing Prepared for DUF Workshop

December 12, 2023

Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

The Importance of High-Confidence Gene Function Annotations

Explains biological phenomenon

Basis for novel disease therapies and biotechnologies

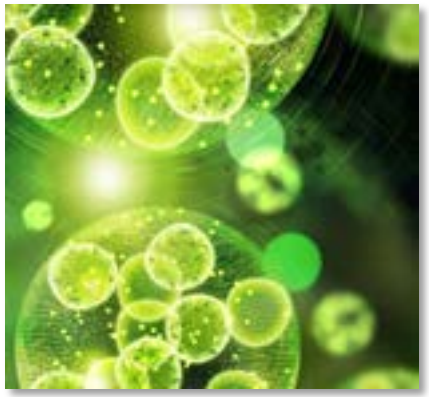
Improved crop yields

Increases sustainability

…. many, many more

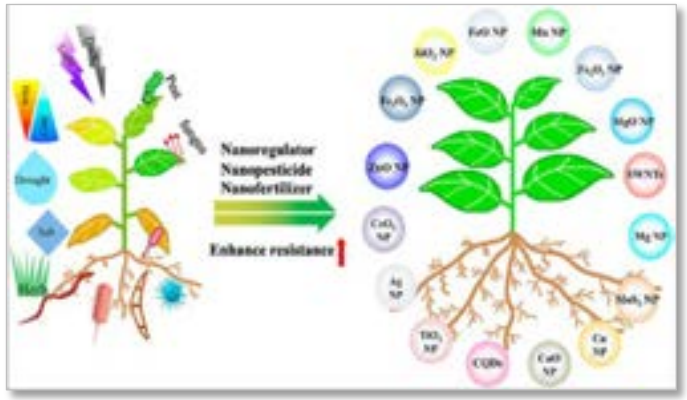
Biomanufacturing

Bioenergy



EzumeImages

Food Security



Reprinted with permission from J Agric Food Chem. 2020;68(7):1935-1947. doi:10.1021/acs.jafc.9b06615. Copyright 2020 American Chemical Society

Whole-Cell Modeling

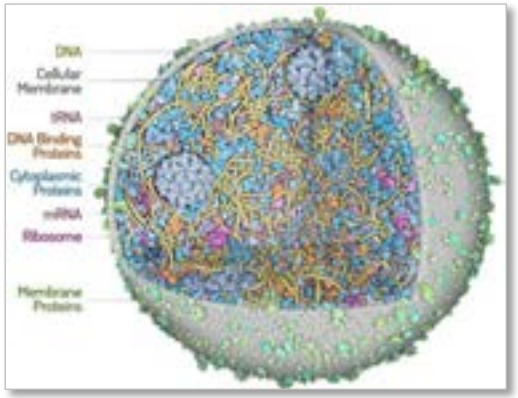
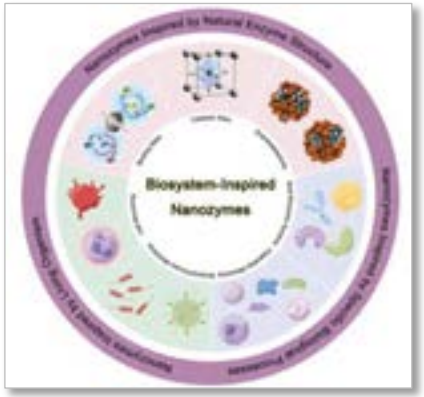


Image by Martina Maritan, Scripps Research

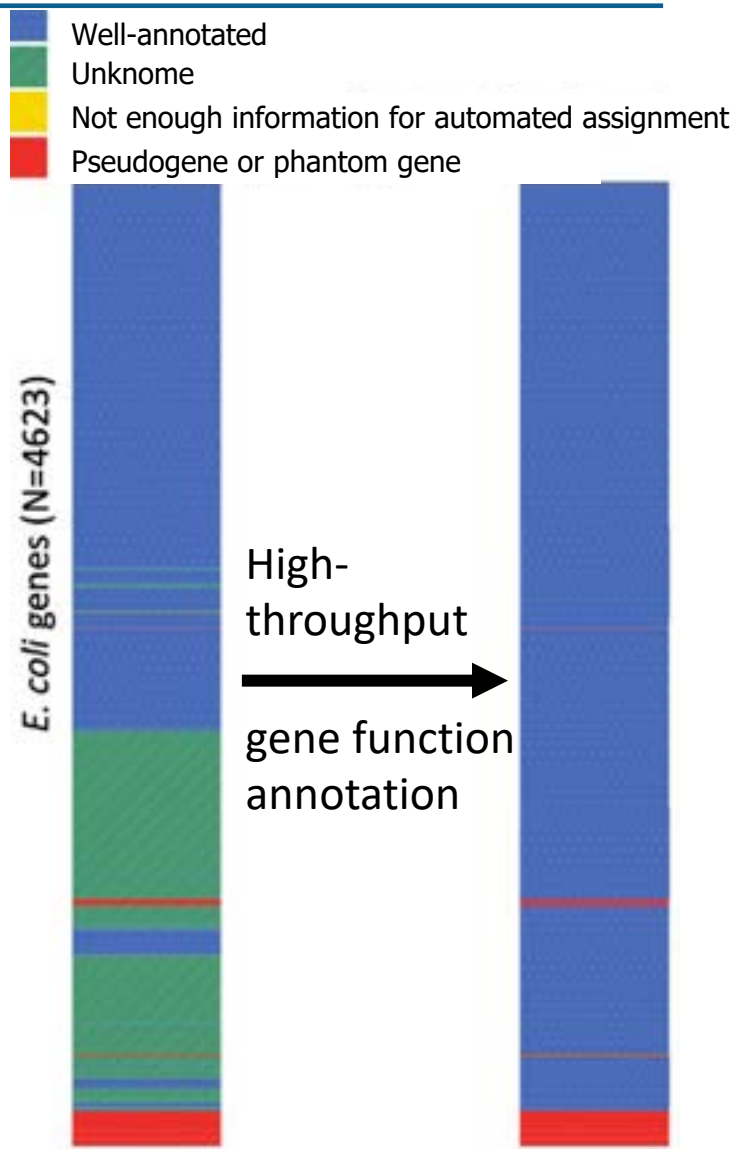
Biomaterial

Biomedical



Adv Mater. 2023;e2211147. doi:10.1002/adma.202211147

Discovering Unknome Functions (DUF)



Depleting the Unknome through reproducible high-throughput gene function annotation methods
Oxford Uni. Press 2019, 47(5), 2446-2454; https://doi.org/10.1093/nar/gkz030

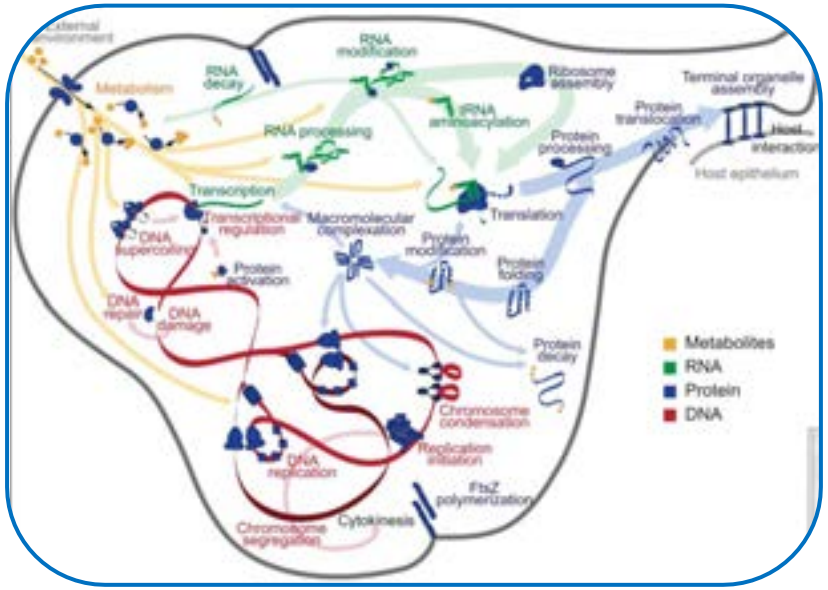
6σ – Six sigma

Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Current Capabilities

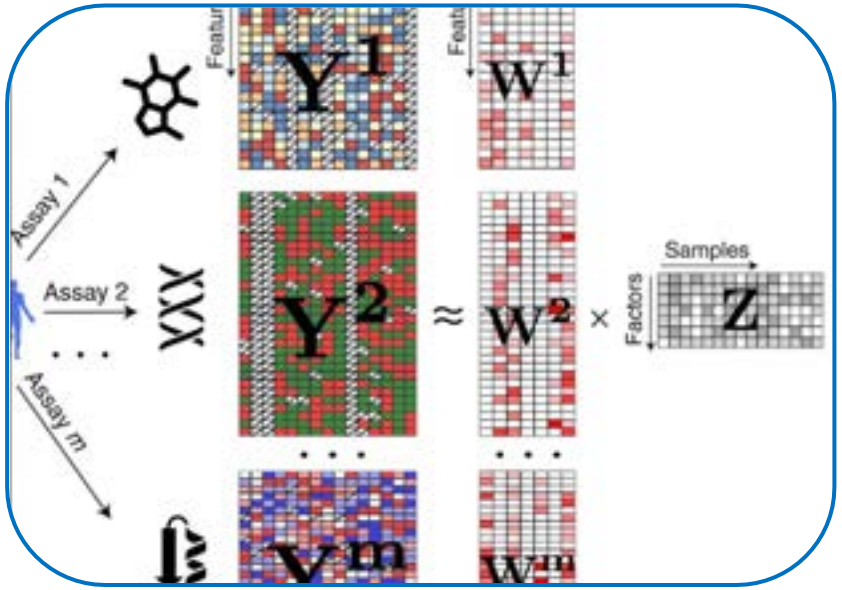
Technological Sweet Spot:

- Automated cultivation techniques, microfluidics, single-cell 'omics, multi-omics, bioinformatics, cloud computing, whole-cell modeling, artificial intelligence, machine learning, computational microscopy, etc…



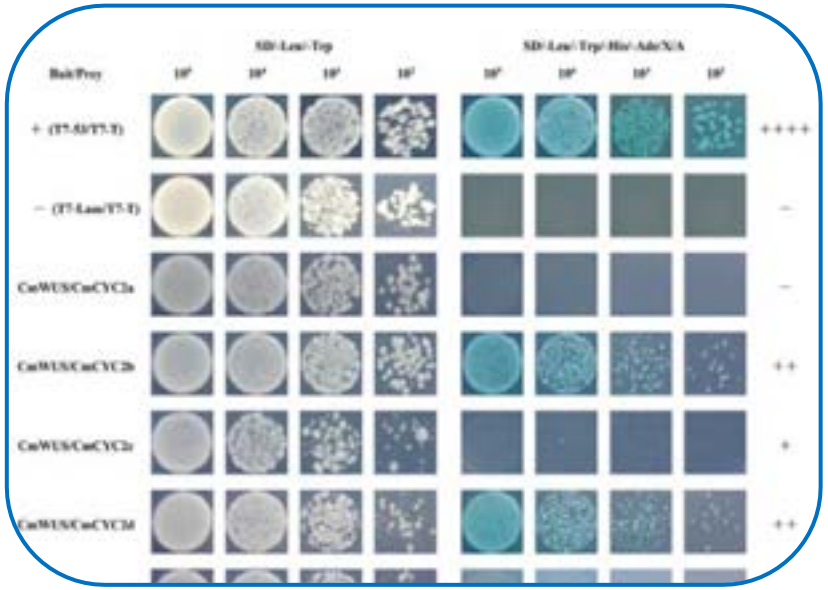
Cell 2012, 150(2), 389-401; https://doi.org/10.1016/j.cell.2012.05.044

multi-omics capabilities



Mol. Sys. Bio. 2018, 14(6), e8124; https://doi.org/10.15252/msb.20178124

big data analysis with bioinformatics



Int. J. Mol. Sci. 2019, 20(6), 1276; https://doi.org/10.3390/ijms20061276

large-scale *in vivo* validation experiments

Technical Challenges

1. Predict gene function

   Current Methods: gene knockouts; database homology; multi-omics data analysis; machine learning (ML); artificial intelligence (AI)

   Challenges: incorporating molecular dynamics and spatiotemporal context; homology creep; versioning; computational power

2. Validate gene function

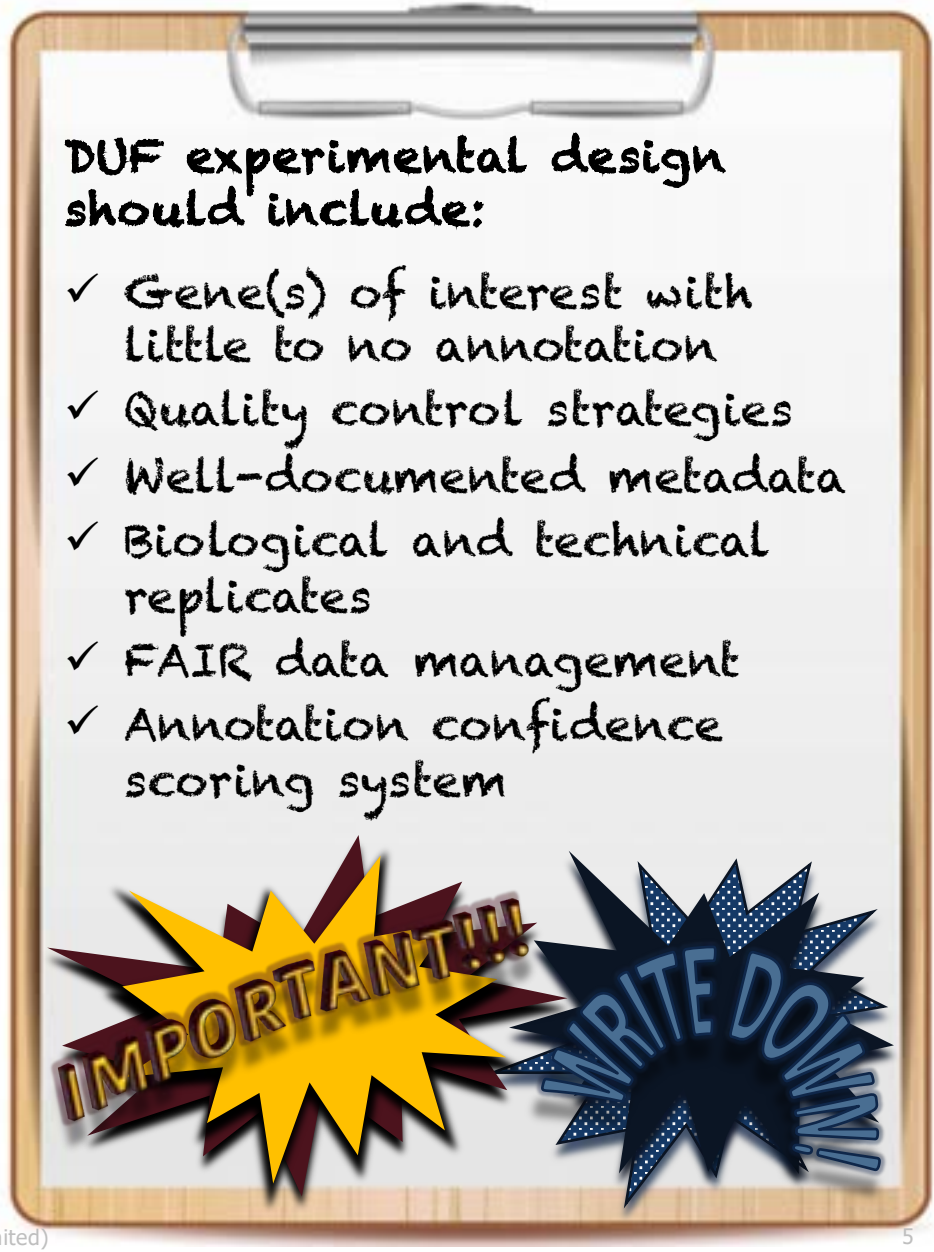
   Current Methods:

   in vitro: protein-to-protein interactions; fluorescent imaging; molecular probes

   biochemically: enzyme kinetics; stochastics; determination of substrates, intermediates, and products

   in vivo: gene overexpression; phenotype rescuing

   Challenges: low throughput; immense biological diversity



Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Successful Abstract

A successful abstract will discuss:

- Innovative high-throughput methods capable of annotating unknown gene function
- A clear technical justification of the method
  - Better than current state-of-the-art
- A clear research plan and experimental design
- The desired goals and output of the study
- The technical ability of the proposer to successfully pursue this research
  - Equipment, facilities, personnel
  - Preliminary data for full-time postdoc

Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

DUF ARC Goals

- **Diversity drives innovation**: Cast wide net to catch innovative ideas for reproducible high-throughput gene function annotation
- High-confidence gene function annotation will benefit multiple research areas

Rapidly generate high-confidence gene function annotations to provide critical knowledge for the advancement of biotechnology in areas vital to the DoD

Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

DUF ARC Agenda Review

2023 Discovering Unknome Function (DUF) Workshop

Dr. René Xavier | December 12, 2023 | Hybrid
Convene at One Boston Place
201 Washington St. | Boston, MA 02108

Workshop Objectives:
Understand the DUF Advanced Research Concept structure and how to apply.
Send all DUF ARC questions to DUF@darpa.mil
Understand the current capabilities for discovering gene function.
Understand the current challenges to discovering unknown gene function (The Unknome).

Time	Topic	Speaker, Organization
0815-0900	Check-in and badging at Convene	
0900-0915	Introduction to Advanced Research Concepts	Ms. Ana Saplan, ARC Manager
0915-0930	Introduction to the DUF Workshop	Dr. René Xavier, DARPA Innovation Fellow
0930-1000	Keynote - Solving the functional puzzle for unknowns: Lessons from 30 years of data mining	Dr. Valerie de Crecy-Lagard, University of Florida
1000-1030	MORNING BREAK	
1030-1200	Lightning Talks (<10 min)	
	The meanings of function in biology	Dr. Anne-Ruxandra Carvunis, University of Pittsburgh
	Approaches to tackling the unknome	Dr. Sean Munro, MRC-LMB, Cambridge
	Systematically discovering and harnessing phenotype-driving	Dr. Gloria Sheynkman, University of Virginia
	Annotation and characterisation of functional noncoding RNA	Dr. Wilfried Haerty, Earlhamm Institute
	Multiscale modeling of intracellular networks and processes	Dr. James Faeder, University of Pittsburgh
	Developing reproducible bioinformatics pipelines	Dr. Olaitan Awe, The Jackson Laboratory for Genomic Medicine
	QC and standards overview	Dr. Samantha Maragh, NIST
1200-1300	LUNCH **Send all DUF ARC questions to DUF@darpa.mil**	
1300-1430	Lightning Talks (<10 min)	
	Beyond the genome: multi-omics across scales	Dr. Kristin Burnum, PNNL
	Characterizing bacterial genes with large-scale genetics	Dr. Adam Deutschbauer, LBNL
	High-throughput culturomics to identify microbial dark matter	Prof. Harris Wang, Columbia University
	Discovery of novel lineages to expand unknome	Dr. Frederik Schulz, DOE Joint Genome Institute
	Identification and prioritization of biosynthetic gene clusters for commercial (meta-)genome mining	Dr. Zachary Charlop-Powers, Ginkgo Bioworks
	Genomics aided host and strain engineering for biotechnology	Dr. Aindrilla Mukhopadhyay, LBNL
	Integrative multi-scale modeling of cellular systems	Dr. Eran Agmon, University of Connecticut Health
	Progress in modeling microbial mechanisms	Dr. Christopher Bettinger, DARPA BTO PM
1430-1500	AFTERNOON BREAK	
1500-1545	Small Group Discussions	
1545-1645	Outbriefs of Small Groups	
1645-1700	DUF ARC Answer Session	
1700	No-host social: Union Oyster House | 41 Union Street | Boston, MA	

Questions

Send all DUF ARC questions to email. Do not put ARC questions in chat or ask speakers.

DUF@darpa.mil

Answers will be given during DUF ARC answer session at 16:45.

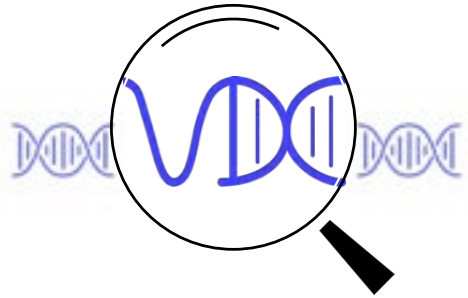
Speaker Q&A: Immediately after talk, if time permits, and during breaks.

Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Solving the functional puzzle for unknowns: Lessons from 30 years of protein function discovery

Valérie de Crécy-Lagard

Dpt of Microbiology and Cell Sciences
& Genetic Institute
University of Florida



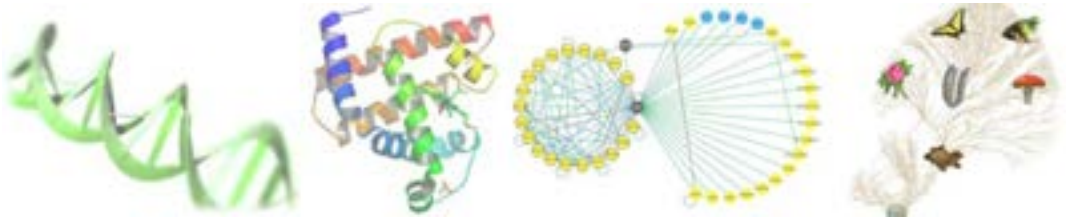
Anne-Ruxandra Carvunis, PhD

Department of Computational and Systems Biology

Pittsburgh Center for Evolutionary Biology and Medicine

University of Pittsburgh School of Medicine

The meanings of "function" in biology



The Carvunis Lab

Change and Innovation in Biological Systems

Pitt Medicine

CEBaM PITTSBURGH

What is biological “function”?

Very complex and debated definitions. Much literature!

The ENCODE controversy

Open access | Published: 05 September 2012

80% of the human genome is functional

An integrated encyclopedia of DNA elements in the human genome

The ENCODE Project Consortium

Nature **489**, 57–74 (2012) | Cite this article

299k Accesses | **11k** Citations | **981** Altmetric | Metrics

The ENCODE controversy

Open access | Published: 05 September 2012

80% of the human genome is functional

An integrated encyclopedia of DNA elements in the human genome

The ENCODE Project Consortium

Nature **489**, 57–74 (2012) | Cite t

299k Accesses | **11k** Citations |

JOURNAL ARTICLE

On the Immortality of Television Sets: "Function" in the Human Genome According to the Evolution-Free Gospel of ENCODE

Dan Graur ✉, Yichen Zheng, Nicholas Price, Ricardo B.R. Azevedo, Rebecca A. Zufall, Eran Elhaik Author Notes

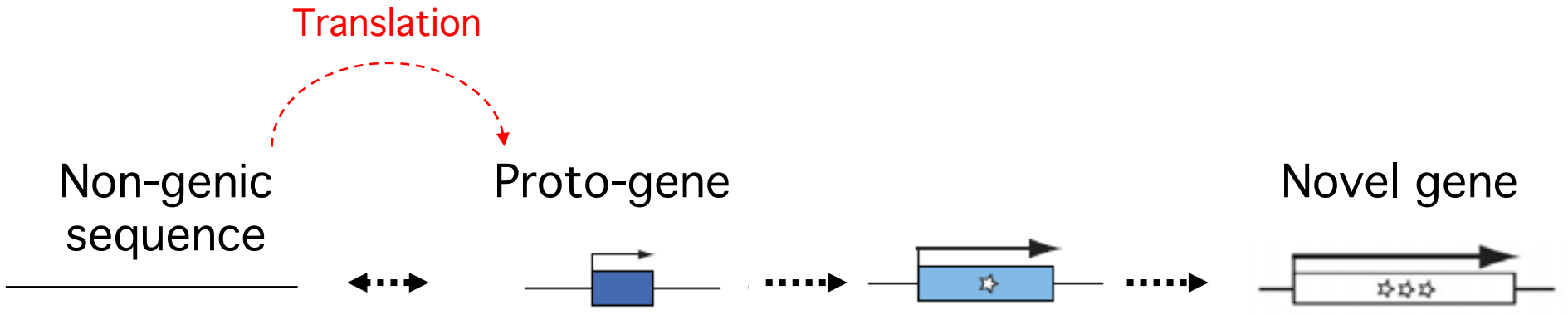
Genome Biology and Evolution, Volume 5, Issue 3, March 2013, Pages 578–590, https://doi.org/10.1093/gbe/evt028

Published: 20 February 2013 **Article history** ▾

the fraction of the genome that is evolutionarily conserved through purifying selection is less than 10%

Evolutionary origins of new genes: *de novo* gene emergence

Carvunis et al., 2012, Van Oss et al 2019



When is a (novel) gene “functional”?

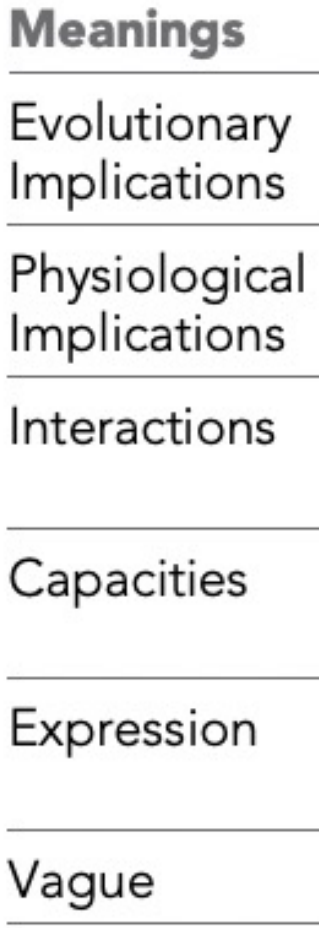
The meanings of function in biology and the problematic case of *de novo* gene emergence

Keeling et al eLife 2019

A philosopher, a biochemist, a rhetoric scholar and me..

.. Analyzed 20 abstracts in the *de novo* field...

.. Found 5 meanings...



The meanings of function in biology and the problematic case of *de novo* gene emergence

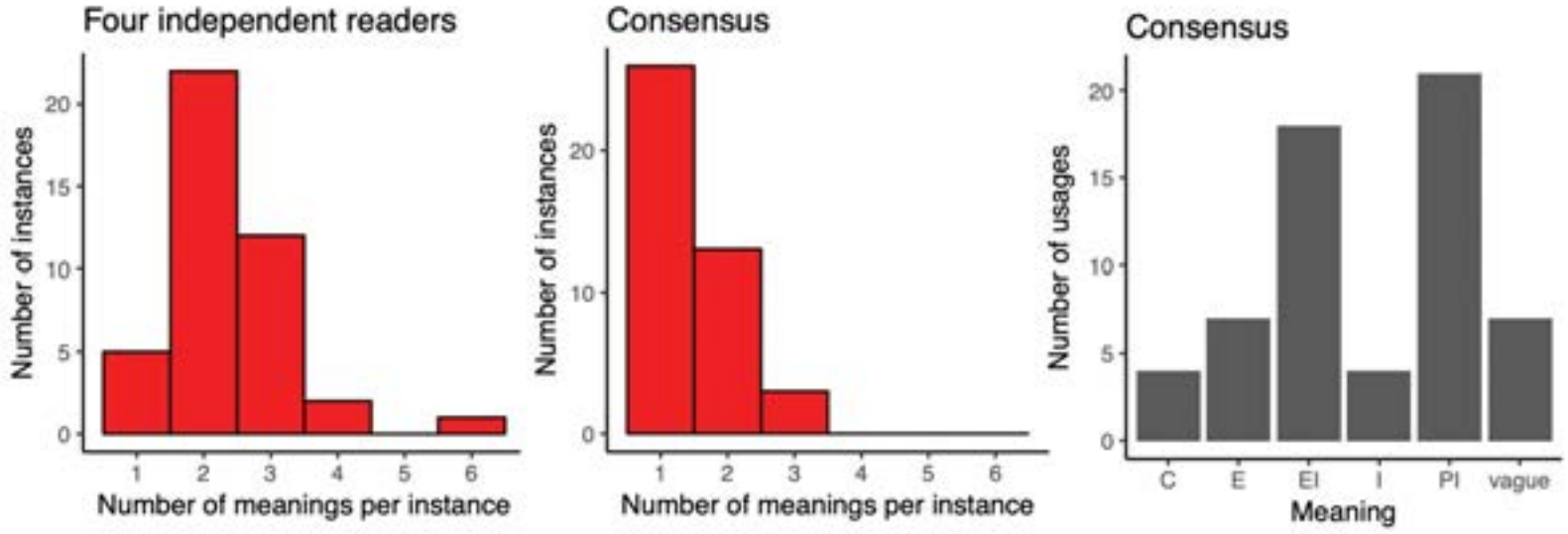
Keeling et al eLife 2019

A philosopher, a biochemist, a rhetoric scholar and me..

.. Analyzed 20 abstracts in the *de novo* field...

.. Found 5 meanings...

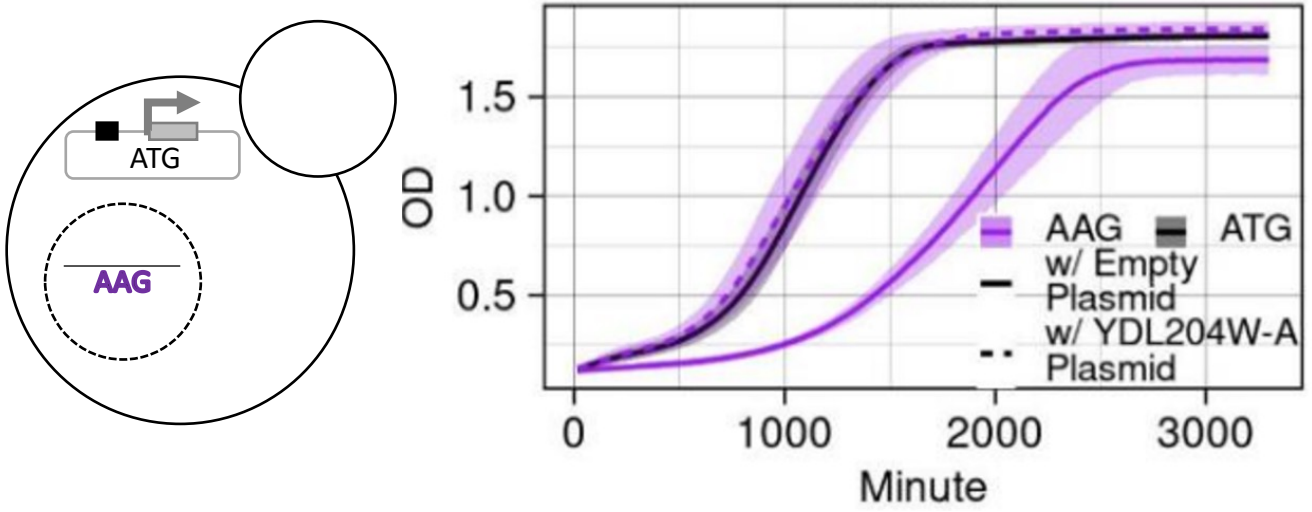
.. and still interpreted sentences differently!



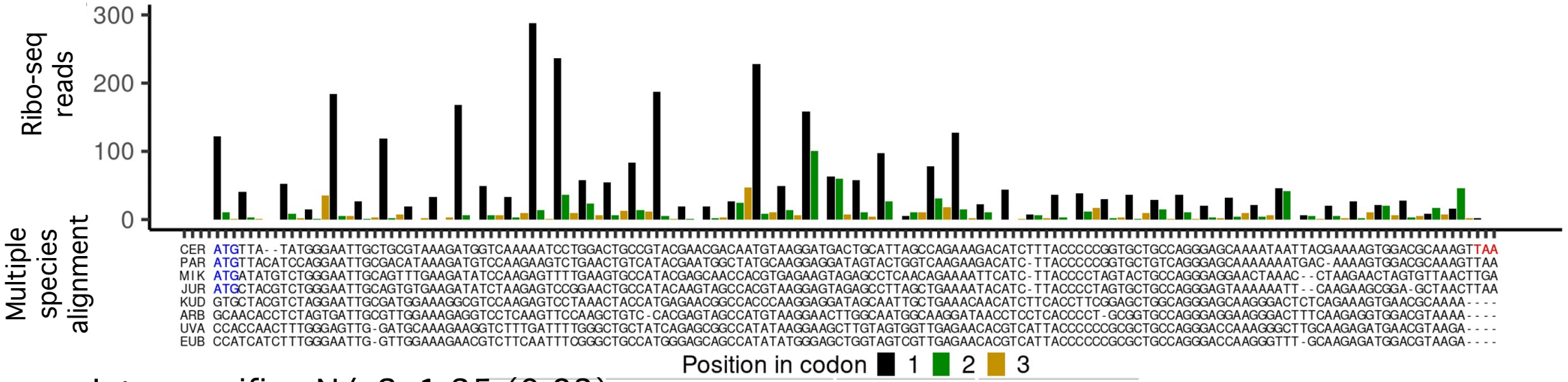
Meanings

- Evolutionary Implications
- Physiological Implications
- Interactions
- Capacities
- Expression
- Vague

Contribute to cellular physiology…



Despite youth and absence of detectable selection:



Intraspecific pN/pS: 1.25 (0.83)

Wacholder et al. 2023

When is a (novel) gene “functional”?

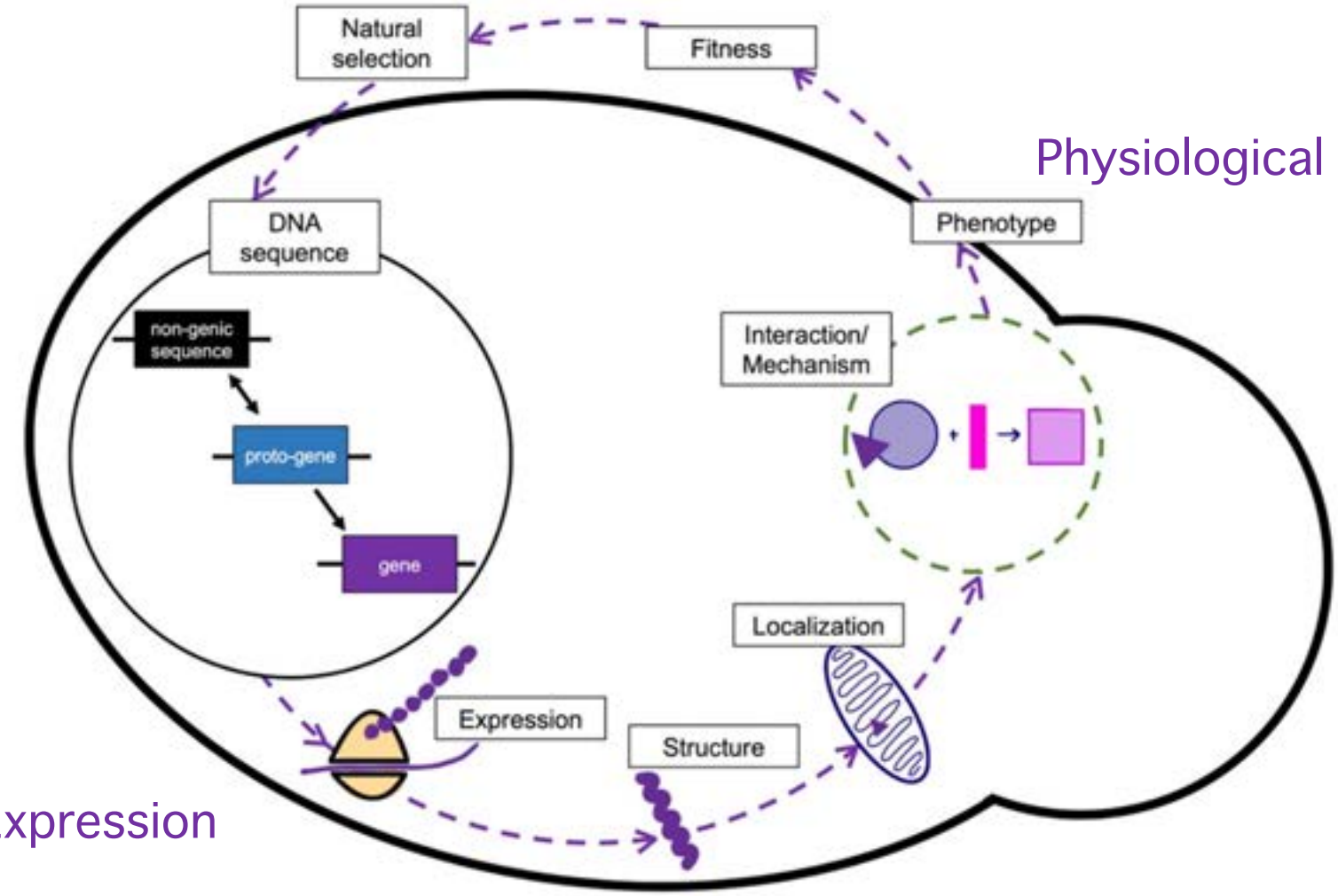
DNA sequence → Protein sequence → Protein function

Mutation

Frameshift

Purifying Selection

My lab's approach (and DUF dream): consider the different components of function independently and their relationship to each other



Physiological implications

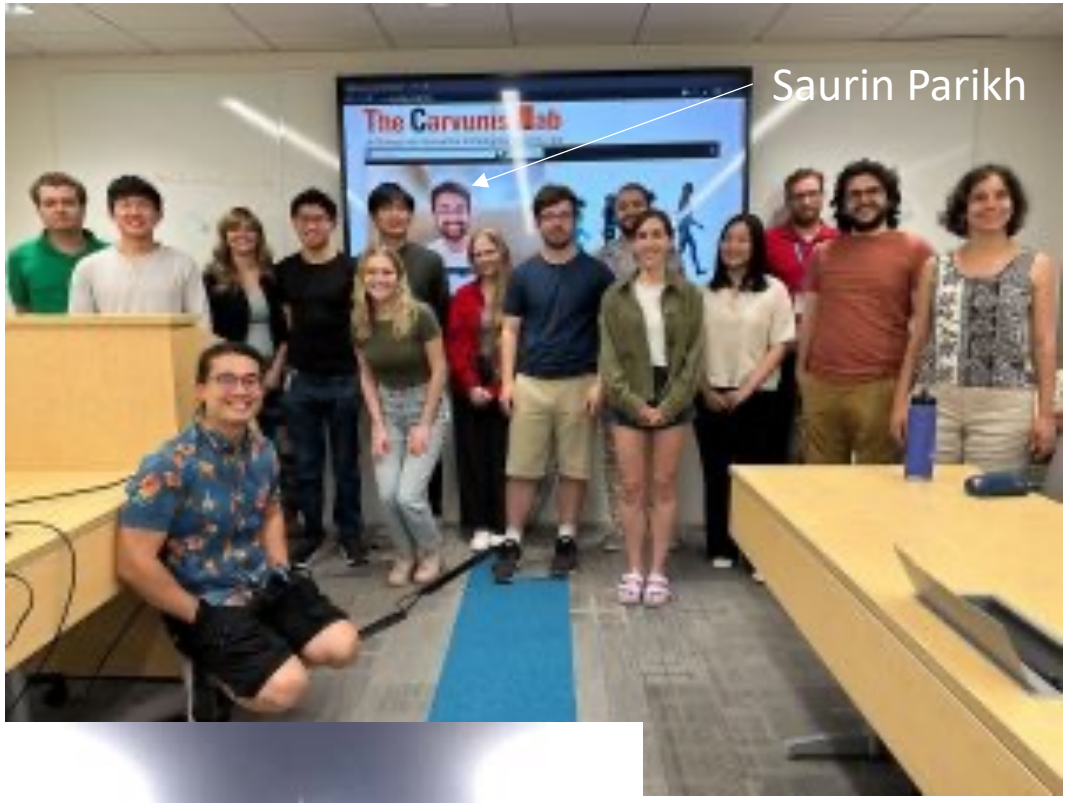
Evolutionary fate: turnover or conservation

Interactions within system

Expression

Intrinsic physical capacities

Keeling et al eLife 2019; Parikh et al Yeast 2022



Drs Nartey, Keeling, Garza

WIKIPEDIA
The Free Encyclopedia

Article Talk

De novo gene birth

From Wikipedia, the free encyclopedia

Thank you!

The Carvunis Lab

Change and Innovation in Biological Systems

NIH DIRECTOR'S TRANSFORMATIVE RESEARCH AWARD

NIH DIRECTOR'S NEW INNOVATOR AWARD

SEARLE SCHOLARS PROGRAM
FUNDING EXCEPTIONAL YOUNG SCIENTISTS

ALFRED P. SLOAN FOUNDATION

NSF National Science Foundation

HFSP

Approaches to tackling the unknome

Sean Munro

Matthew Freeman

Tim Stevens

Human Unknome

Genome sequenced in 2003

19,969 protein-coding genes

~20-30 of these genes have no known molecular function

e.g. 3033 of these genes are not in PubMed

Unknome of life

Complete genome sequences for 34,928 organisms (JGI GOLD)

>600 million proteins from meta genomics (EBI MGnify)

Addressing the human unknome

1) Build an Unknome Database

Quantify "known-ness" by collating experimental evidence from model organisms

2) Select c 200-300 well-conserved but unknown human proteins and examine using *Drosophila* genetics

3) Use machine learning to predict function of unknown human proteins

With Matthew Freeman (Oxford University) and Tim Stevens MRC LMB

1) Constructing an Unknome Database

i) Cluster orthologous proteins from humans and 11 model organisms

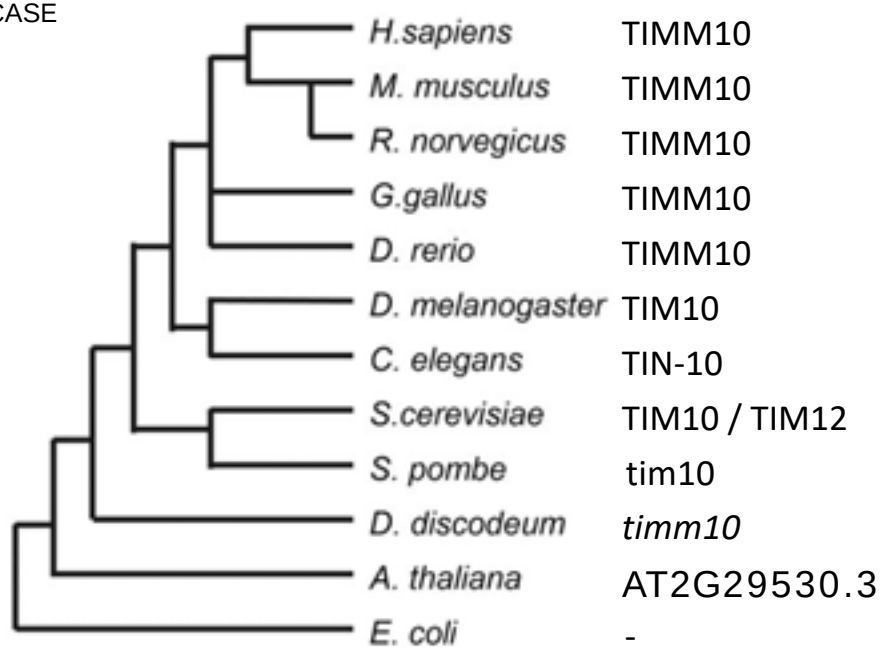
PANTHER Classification System

"a comprehensive, annotated library of gene family phylogenetic trees"

pantherdb.org

Cluster UKP01389

MITOCHONDRIAL IMPORT INNER MEMBRANE TRANSLOCASE SUBUNIT TIM10



H.sapiens	TIMM10
M. musculus	TIMM10
R. norvegicus	TIMM10
G.gallus	TIMM10
D. rerio	TIMM10
D. melanogaster	TIM10
C. elegans	TIN-10
S.cerevisiae	TIM10 / TIM12
S. pombe	tim10
D. discodeum	*timm10*
A. thaliana	AT2G29530.3
E. coli	-

1) Constructing an Unknome Database

ii) Calculate knownness score for cluster from Gene Ontology terms

GO consortium: systematic annotation of genes using a controlled vocabulary

Cluster UKP01389

MITOCHONDRIAL IMPORT INNER MEMBRANE TRANSLOCASE SUBUNIT TIM10

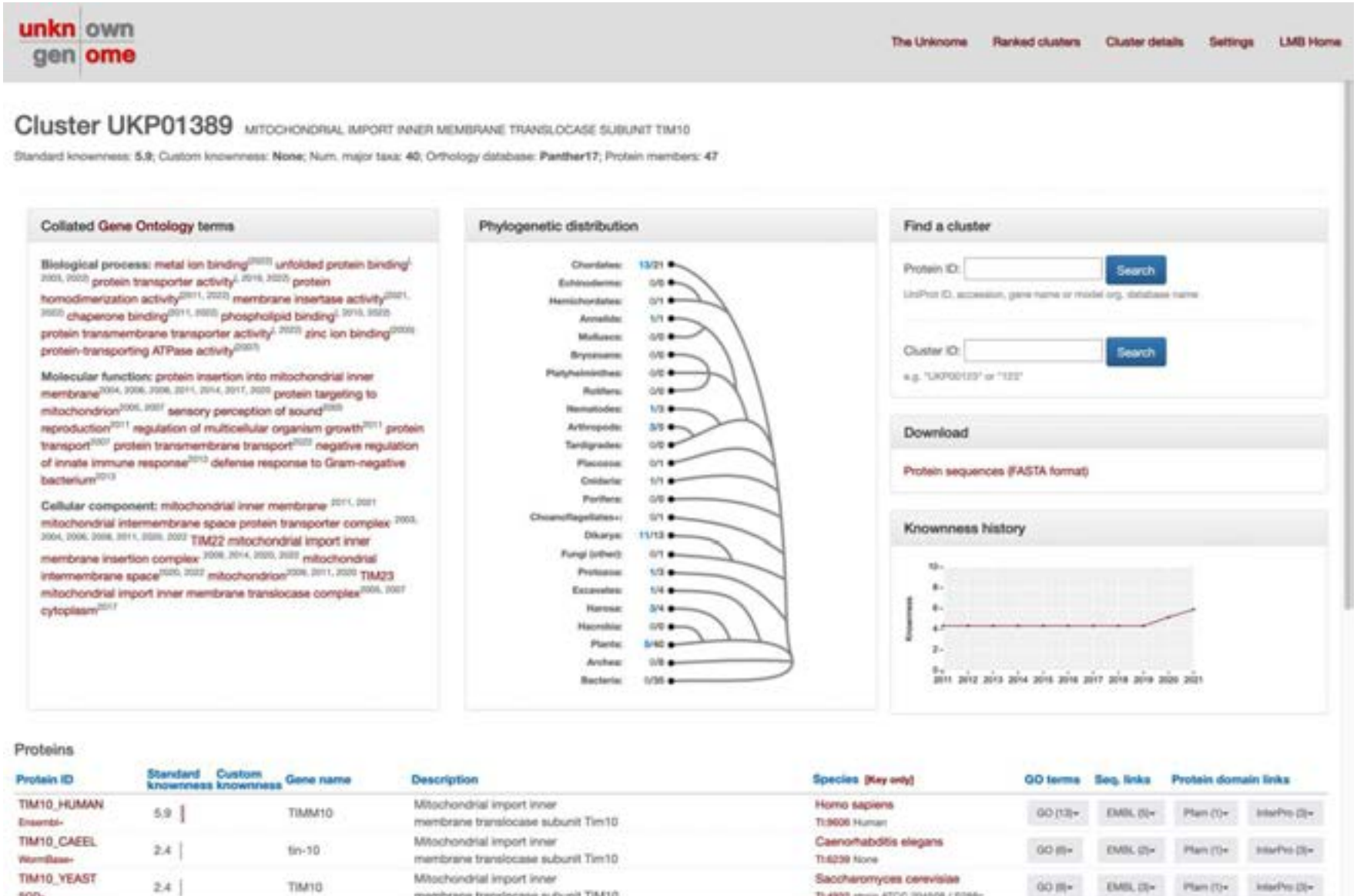
		Weighted score
H.sapiens	TIMM10	5.9
M. musculus	TIMM10	2.8
R. norvegicus	TIMM10	2.5
G.gallus	TIMM10	0.5
D. rerio	TIMM10	0.5
D. melanogaster	TIM10	0.5
C. elegans	TIN-10	2.4
S.cerevisiae	TIM10 / TIM12	2.4/1.6
S. pombe	tim10	0.5
D. discodeum	*timm10*	0.5
A. thaliana	AT2G29530.3	0.5
E. coli	-	-

Knownness

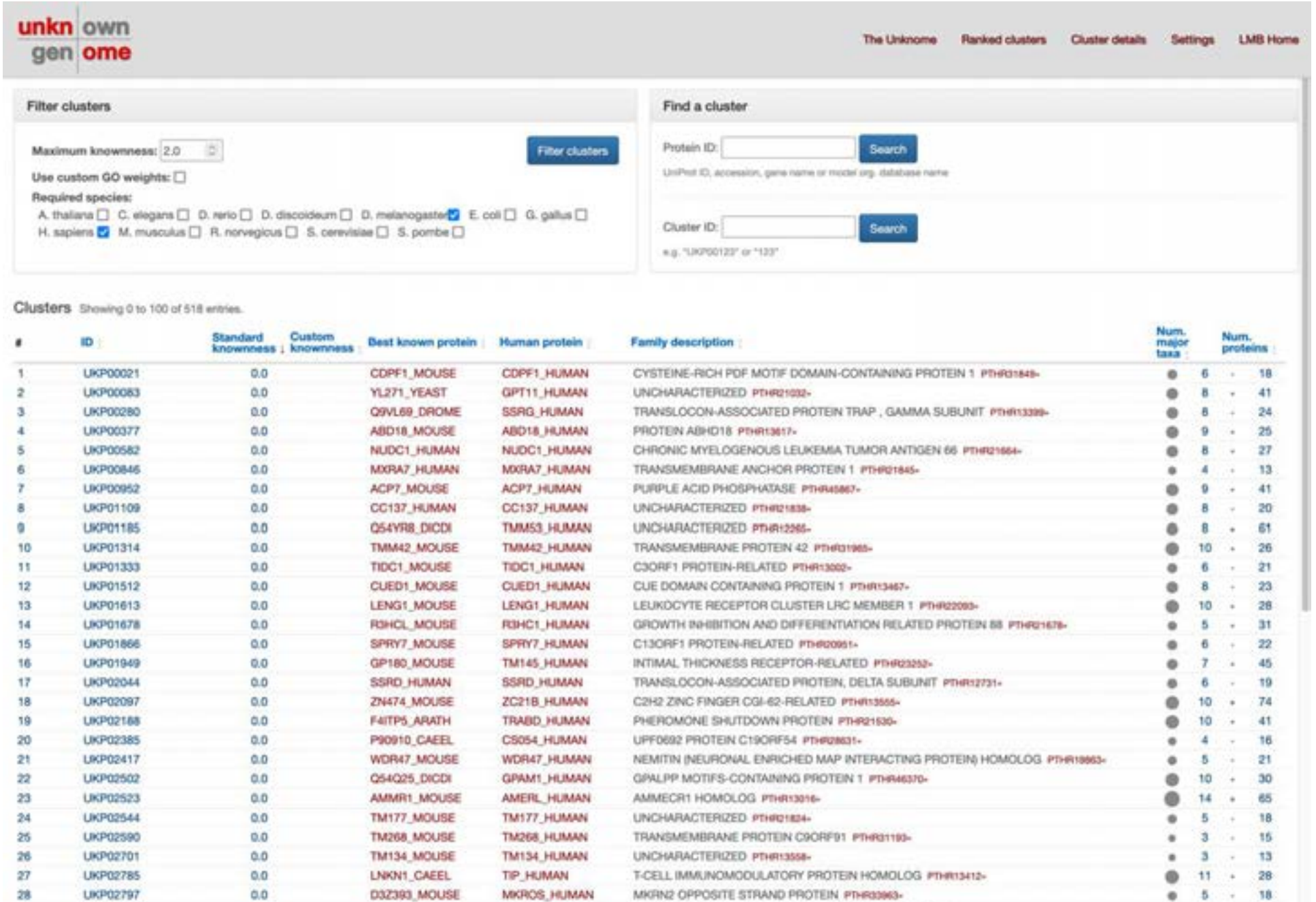
5.9

1) Constructing an Unknome Database

iii) Online: www.unknome.org

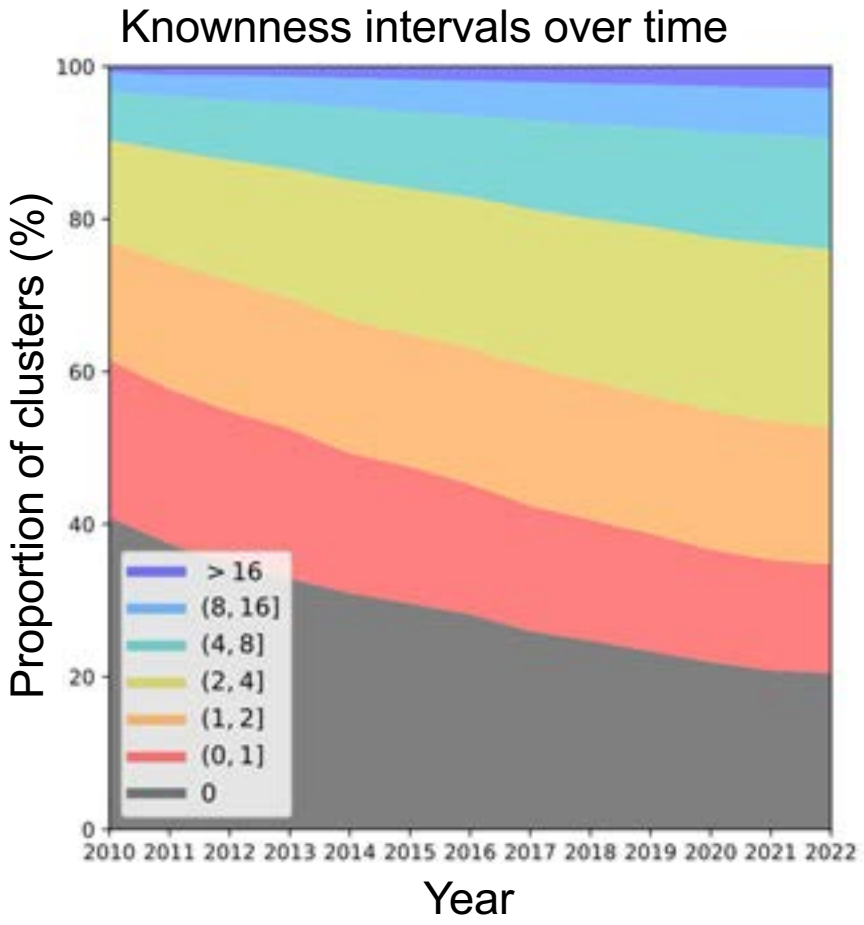


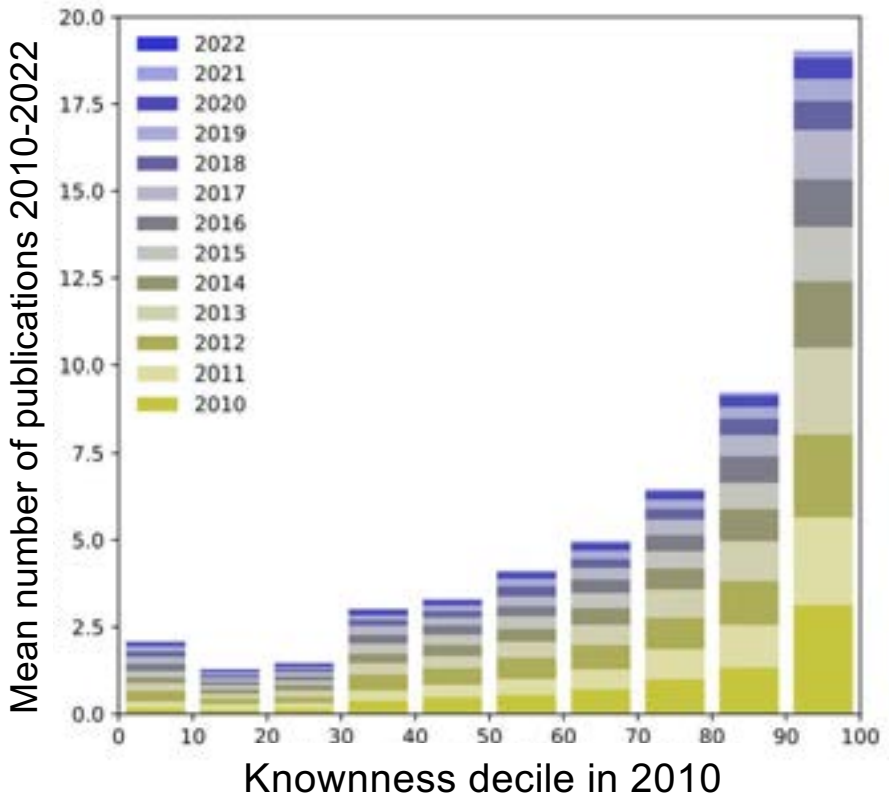
www.unknome.org



#	ID	Standard knowness	Custom knowness	Best known protein	Human protein	Family description	Num. major taxa	Num. proteins
1	UKP00021	0.0		CDPF1_MOUSE	CDPF1_HUMAN	CYSTEINE-RICH PDF MOTIF DOMAIN-CONTAINING PROTEIN 1 PTHR31849-	6	18
2	UKP00083	0.0		YL271_YEAST	GPT11_HUMAN	UNCHARACTERIZED PTHR21032-	8	41
3	UKP00280	0.0		Q9VL69_DROME	SSRG_HUMAN	TRANSLOCON-ASSOCIATED PROTEIN TRAP , GAMMA SUBUNIT PTHR13399-	8	24
4	UKP00377	0.0		ABD18_MOUSE	ABD18_HUMAN	PROTEIN ABHD18 PTHR13617-	9	25
5	UKP00582	0.0		NUDC1_HUMAN	NUDC1_HUMAN	CHRONIC MYELOGENOUS LEUKEMIA TUMOR ANTIGEN 66 PTHR21664-	8	27
6	UKP00846	0.0		MXRA7_HUMAN	MXRA7_HUMAN	TRANSMEMBRANE ANCHOR PROTEIN 1 PTHR21845-	4	13
7	UKP00952	0.0		ACP7_MOUSE	ACP7_HUMAN	PURPLE ACID PHOSPHATASE PTHR45867-	9	41
8	UKP01109	0.0		CC137_HUMAN	CC137_HUMAN	UNCHARACTERIZED PTHR21838-	8	20
9	UKP01185	0.0		Q54YR8_DICDI	TMM53_HUMAN	UNCHARACTERIZED PTHR12565-	8	61
10	UKP01314	0.0		TMM42_MOUSE	TMM42_HUMAN	TRANSMEMBRANE PROTEIN 42 PTHR31965-	10	26
11	UKP01333	0.0		TIDC1_MOUSE	TIDC1_HUMAN	C3ORF1 PROTEIN-RELATED PTHR13002-	6	21
12	UKP01512	0.0		CUED1_MOUSE	CUED1_HUMAN	CUE DOMAIN CONTAINING PROTEIN 1 PTHR13467-	8	23
13	UKP01613	0.0		LENG1_MOUSE	LENG1_HUMAN	LEUKOCYTE RECEPTOR CLUSTER LRC MEMBER 1 PTHR22093-	10	28
14	UKP01678	0.0		R3HCL_MOUSE	R3HC1_HUMAN	GROWTH INHIBITION AND DIFFERENTIATION RELATED PROTEIN 88 PTHR21678-	5	31
15	UKP01866	0.0		SPRY7_MOUSE	SPRY7_HUMAN	C13ORF1 PROTEIN-RELATED PTHR20951-	6	22
16	UKP01949	0.0		GP180_MOUSE	TM145_HUMAN	INTIMAL THICKNESS RECEPTOR-RELATED PTHR23252-	7	45
17	UKP02044	0.0		SSRD_HUMAN	SSRD_HUMAN	TRANSLOCON-ASSOCIATED PROTEIN, DELTA SUBUNIT PTHR12731-	6	19
18	UKP02097	0.0		ZN474_MOUSE	ZC21B_HUMAN	C2H2 ZINC FINGER CGI-62-RELATED PTHR13555-	10	74
19	UKP02168	0.0		F4ITP5_ARATH	TRABD_HUMAN	PHEROMONE SHUTDOWN PROTEIN PTHR21530-	10	41
20	UKP02385	0.0		P90910_CAEEL	CS054_HUMAN	UPF0692 PROTEIN C19ORF54 PTHR28631-	4	16
21	UKP02417	0.0		WDR47_MOUSE	WDR47_HUMAN	NEMITIN (NEURONAL ENRICHED MAP INTERACTING PROTEIN) HOMOLOG PTHR19963-	5	21
22	UKP02502	0.0		Q54Q25_DICDI	GPAM1_HUMAN	GPALPP MOTIFS-CONTAINING PROTEIN 1 PTHR46370-	10	30
23	UKP02523	0.0		AMMR1_MOUSE	AMERL_HUMAN	AMMECR1 HOMOLOG PTHR13016-	14	65
24	UKP02544	0.0		TM177_MOUSE	TM177_HUMAN	UNCHARACTERIZED PTHR21824-	5	18
25	UKP02590	0.0		TM268_MOUSE	TM268_HUMAN	TRANSMEMBRANE PROTEIN C9ORF91 PTHR31193-	3	15
26	UKP02701	0.0		TM134_MOUSE	TM134_HUMAN	UNCHARACTERIZED PTHR13558-	3	13
27	UKP02785	0.0		LNKN1_CAEEL	TIP_HUMAN	T-CELL IMMUNOMODULATORY PROTEIN HOMOLOG PTHR13412-	11	28
28	UKP02797	0.0		D3Z393_MOUSE	MKRN2_HUMAN	MKRN2 OPPOSITE STRAND PROTEIN PTHR33963-	5	18

Unknome is slowly shrinking





Human unknome is shrinking but is still relatively neglected

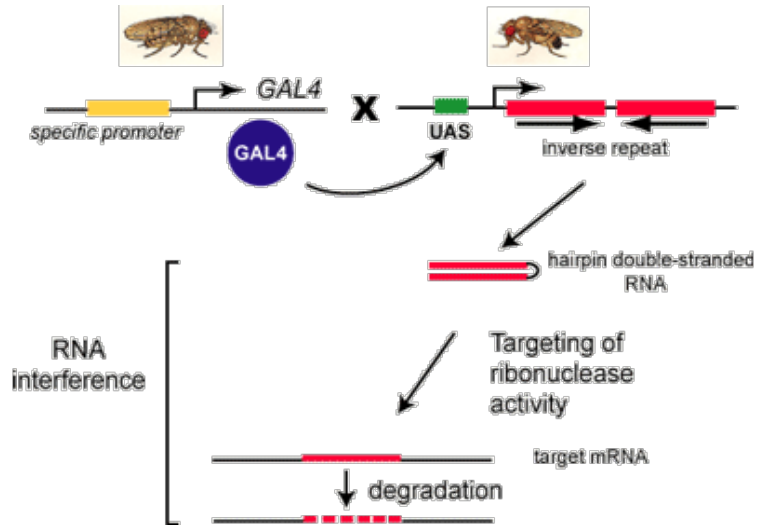
2) RNAi screen for phenotypes using *Drosophila*

RNAi against 260 genes conserved in humans and flies

Gal4 expressed in tissue of choice

gene for dsRNA hairpin

Assays *(quick, visual, quantitative)*



Lethals:	62 / 260 genes (24%)
Specific screens:	59 / 198 genes
Tissue growth	3
Male fertility	11
Female fertility	2
Lifespan -AA	9
Lifespan ROS	13
Proteostasis	6
Locomotion	6

Unknown proteins have key roles even in laboratory conditions

Functional screening in flies is painfully slow

Rocha et al. (2023) PLOS Biology e3002222

3) Machine learning to predict the function of unknown human proteins

Use genome-wide data to group proteins involved in same processes. Start by looking at stable protein complexes

Proteins in the same complex *tend* to have a similar level, phenotype, location and species conservation

Gene expression
RNA-seq; 182 cell lines, 18,730 proteins

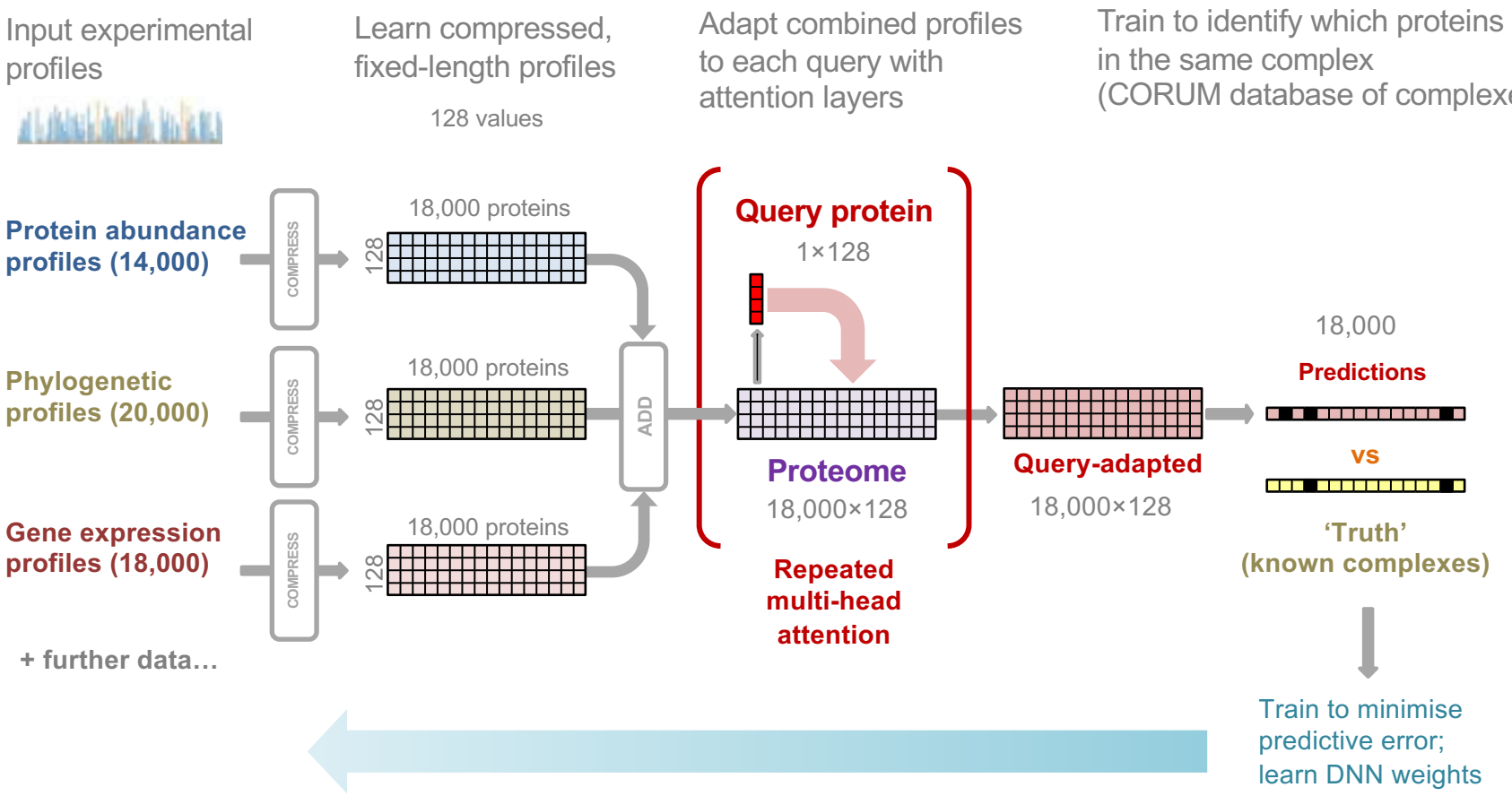
Protein abundance
Shotgun proteomics; 579 cell lines, 14,600 proteins

Subcellular location
Fractionation MS (LOPIT); 40 fractions, 6,575 proteins

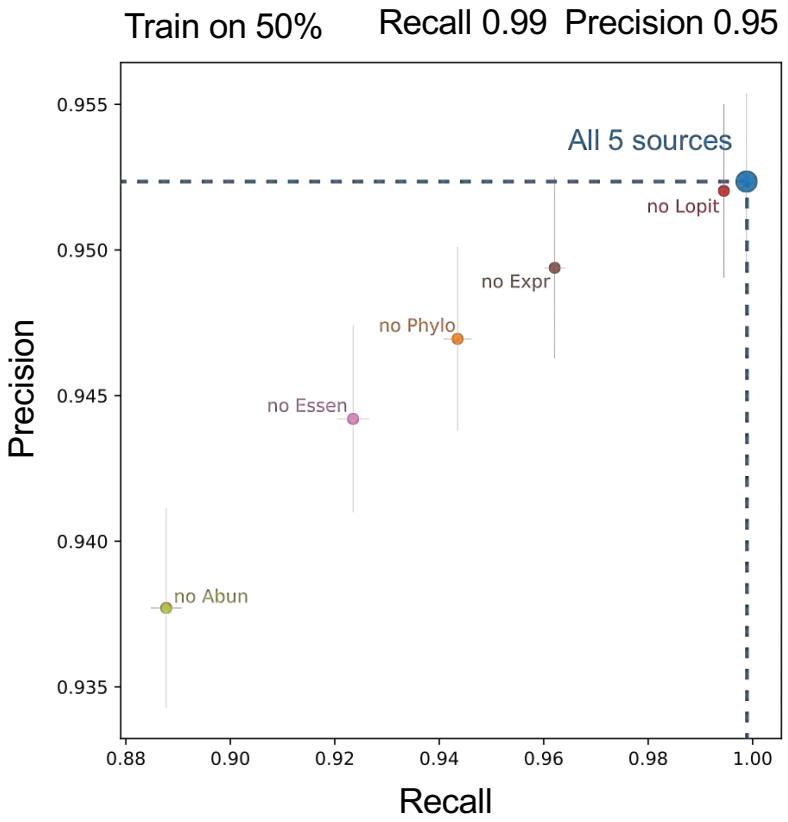
Gene essentiality
CRISPR knock-out (DepMap); 990 cell lines, 17,190 proteins

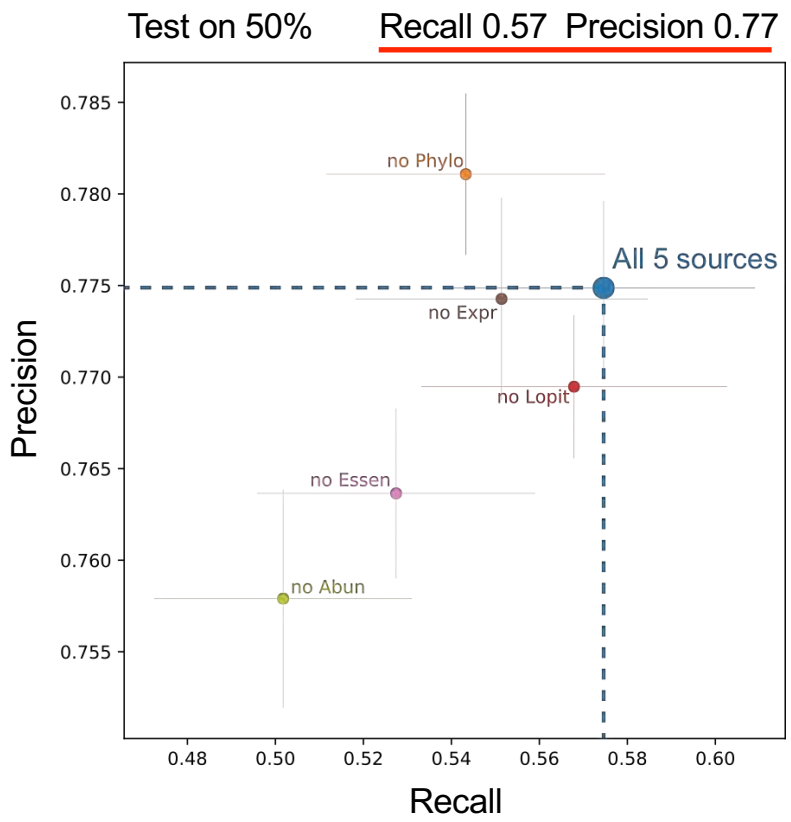
Phylogeny
Orthologue similarity; 246 eukaryote species, 20,250 proteins

Proteome attention deep neural network



Proteome attention deep neural network





Five sources of data:

Abun:	Proteomic abundance, mass spec
Expr:	Expression, RNA-seq
Essen:	Knock down gene essentiality (DepMap)
Lopit:	Sub-cellular fractionation proteomics (LOPIT)
Phylo:	Eukaryote phylogenetic profiles

All sources combine to make better predictors

Proteomic abundance is the best source

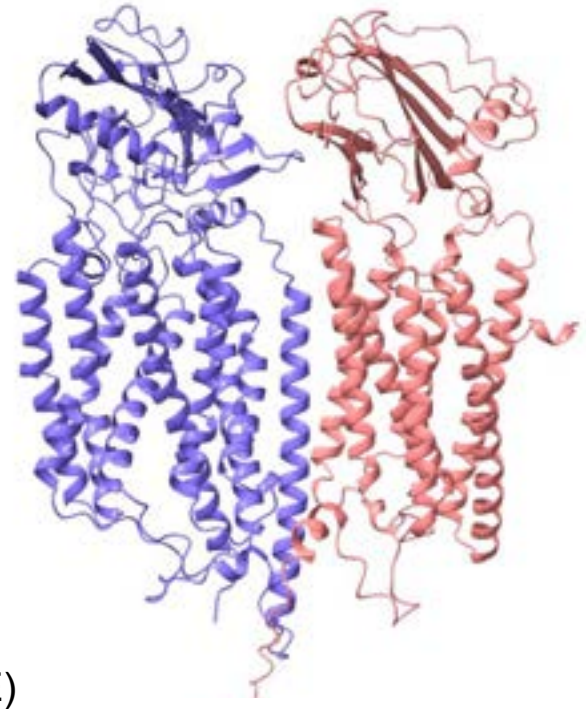
Testing predictions using AlphaFold 2

TM9SF2 - 10 hits from DNN:

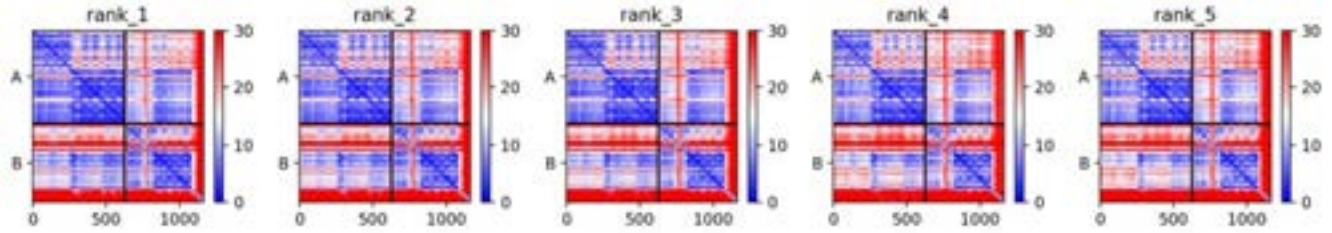
Test by AF2 - one looks real

Recent progress in protein structure/interaction prediction offers great opportunities

TM9SF2 TMEM87A



High confidence prediction (PAE)



Acknowledgments

Unknome database

Machine learning

Tim Stevens

Statistics

Rajen Shah

Centre for Mathematical Sciences
University of Cambridge

Drosophila

Nadine Muschalik

Joao Rocha

Satish Jayaram

Sahar Emran

Cristina Robles

Office workers

Sean Munro

Matthew Freeman

Dunn School of Pathology
University of Oxford

www.unknome.org

MRC Laboratory of Molecular Biology

MRC Medical Research Council

Systematically discovering and harnessing phenotype-driving proteoforms

Dr. Gloria Sheynkman

University of Virginia

gs9yr@virginia.edu

Annotation and characterisation of functional noncoding RNA

Wilfried Haerty

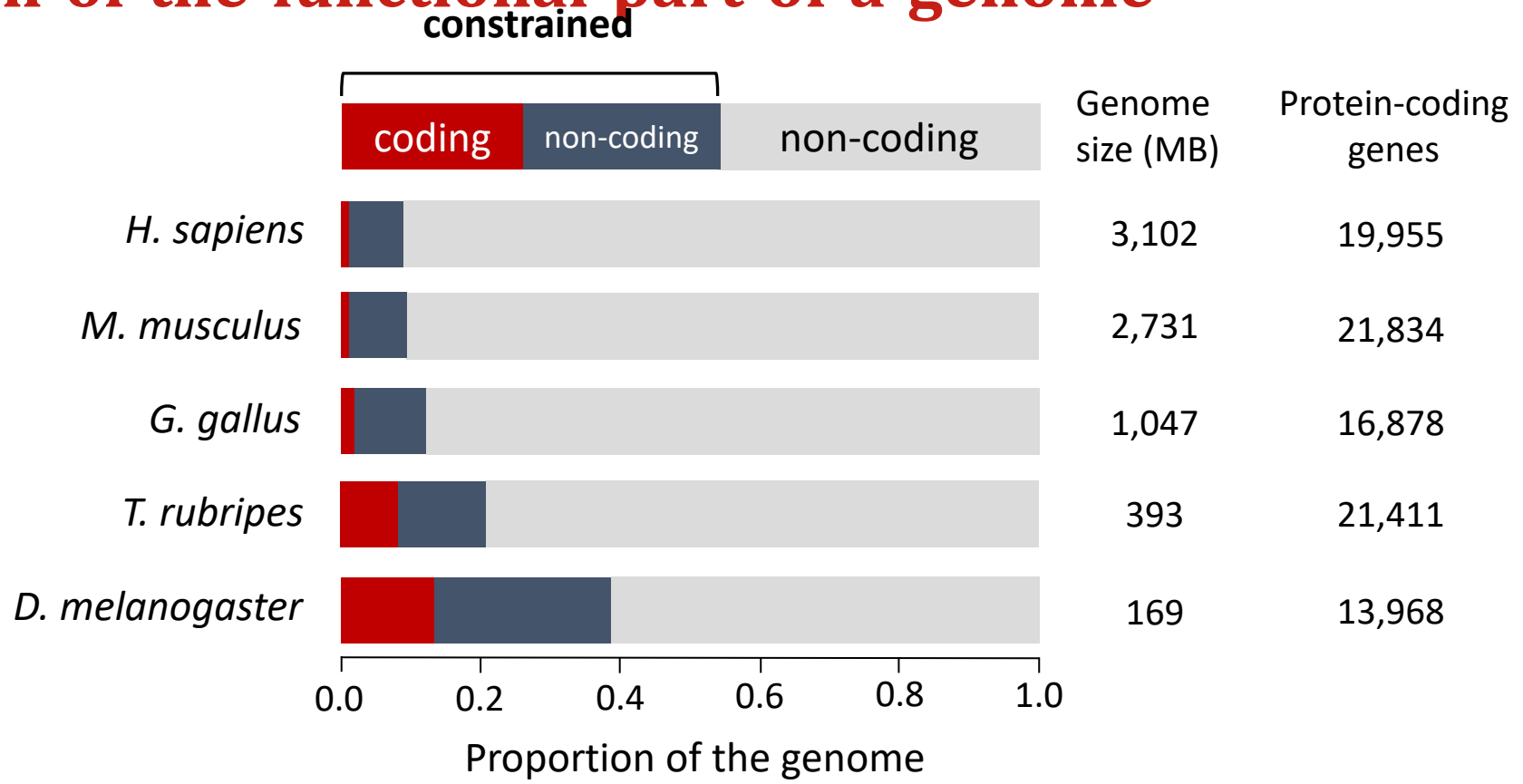
wilfried.haerty@earlham.ac.uk

YouTube

Earlham Institute

Decoding Living Systems

Evolution of the functional part of a genome



Haerty & Ponting. 2014. Annu. Rev. Genomics Hum. Genet.

Impact of variation within the non-coding genome on phenotype

We integrate large data sets to understand the impact of genetic variation on traits

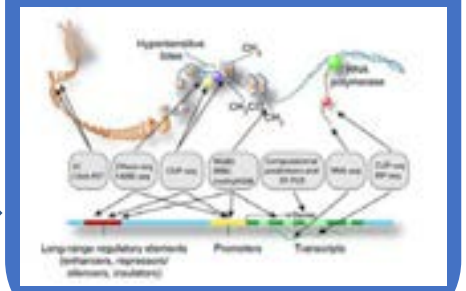
Comparative genomics

Population genomics

Functional genomics

Functional sequence characterization

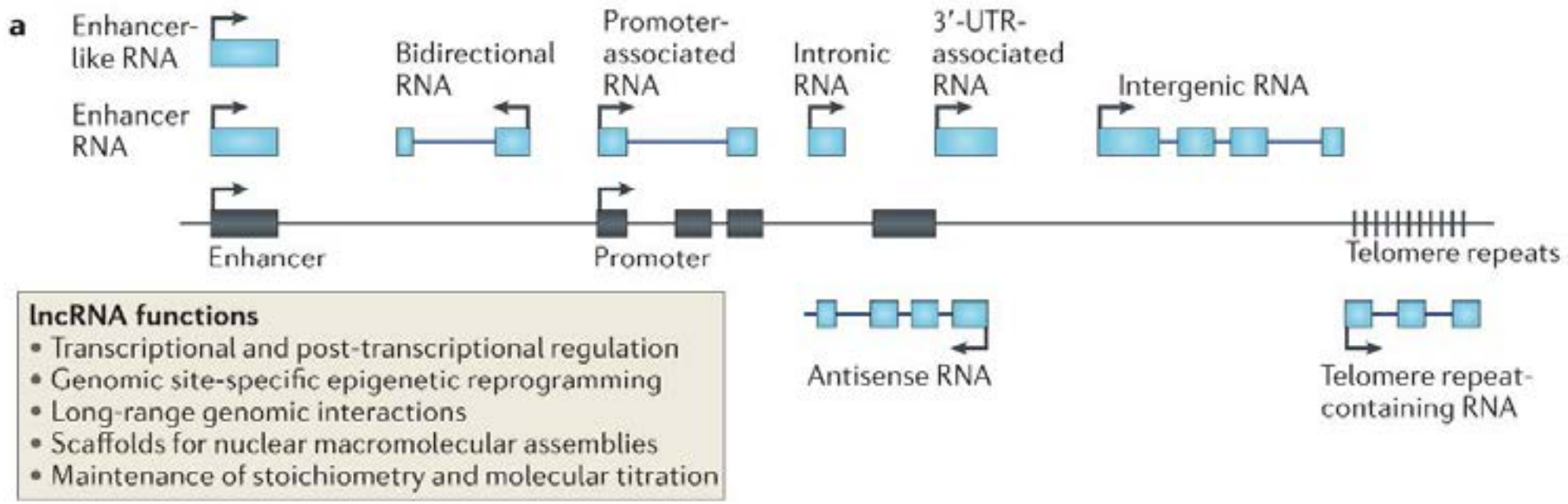
- Novel transcripts
- Regulatory elements



Predict the impact of variation on gene expression and protein representation

Identify variants associated with phenotypes of interest including disease

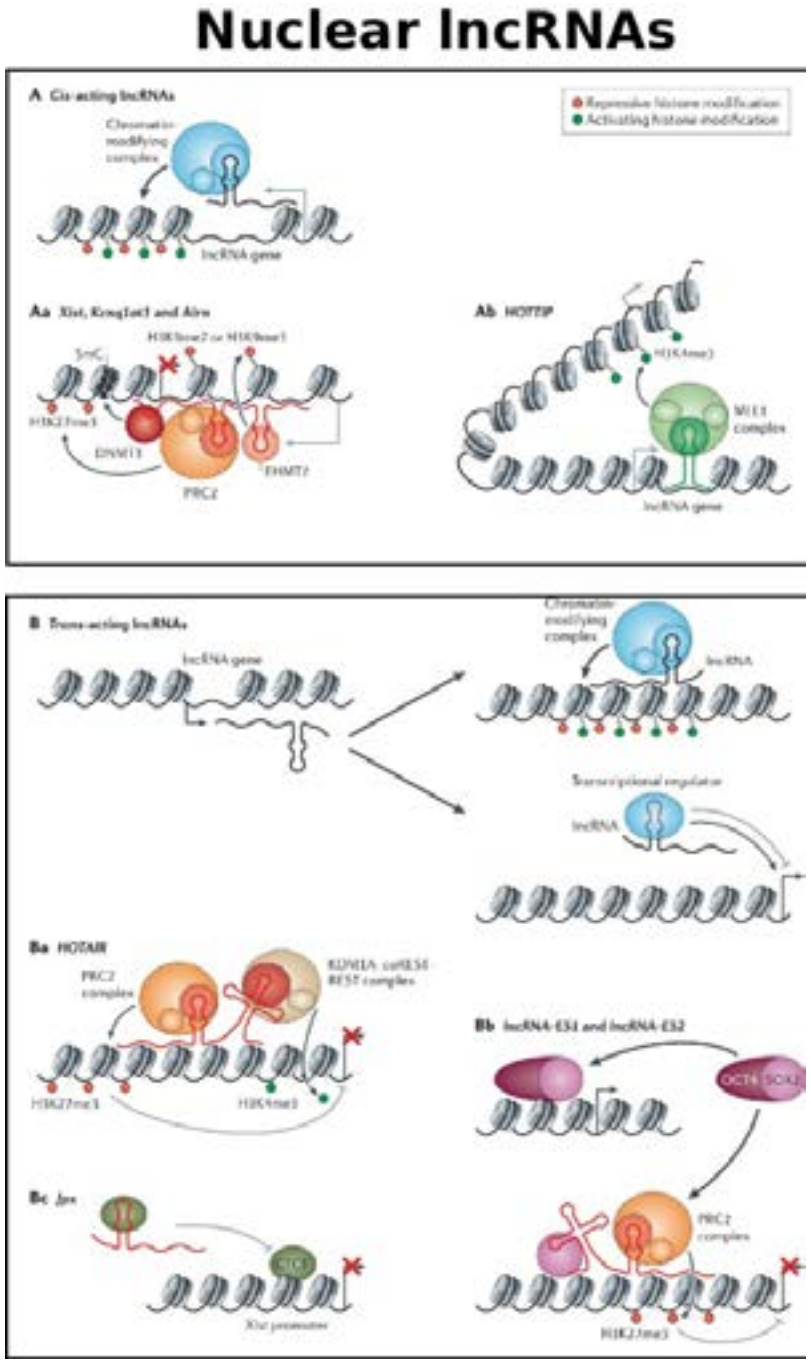
Non-coding RNAs found across kingdoms and in different flavours



Qureshi and Mehler. 2012. Nat. Rev. Neuro.

Non-coding RNAs – functional loci

Name of lncRNA	Mechanism of action	Mutant phenotype
XIST	X chromosome regulation (imprinting and X chromosomal dosage compensation)	*Mus musculus*: females inheriting paternal allele were embryonic lethal; males fully viable
FENDRR	Thought to act by binding to PRC2 and/or TrxG/MLL complexes to promote the methylation of the promoters of target genes, thus reducing their expression; essential for normal development of the heart and body wall	*Mus musculus*: Embryonic lethal
roX1, roX2	Required for sex chromosome dosage compensation in Drosophila (hyper-transcription of X chromosome in males)	*Drosophila melanogaster*: None, except when in combination: male-specific reduction in viability
HOTAIR	The 5' end of HOTAIR interacts with a Polycomb-group protein Polycomb Repressive Complex 2 (PRC2) and as a result regulates chromatin state - required for gene-silencing of the HOXD locus by PRC2. The 3' end of HOTAIR interacts with the histone demethylase LSD1; epigenetic differentiation of skin over the surface of the body	*Mus musculus*: Spine and wrist malformations
COOLAIR	Suggested to function in early cold induced silencing of FLC transcription in *Arabidopsis thaliana*	None reported
COLDAIR	Required to recruit PRC2 to the FLC locus allowing deposition of the repressive H3K27me3 chromatin mark. Binds PRC2 complex protein CURLY LEAF (CLF); required for stable repression of FLC after vernalization	*Arabidopsis thaliana*: Late flowering after vernalization



LncRNAs – Many mechanisms

Chromatin modification in

- cis:
  - recruitment of DNMT3 / PCR2
  - transcriptional interference
- trans:
  - recruitment of chromatin modifying complex
  - transcriptional regulators

LncRNAs can act in :

- competition with mRNAs from miRNAs (ceRNAs)
- miRNA sponges
- modulation of RNA stability

Fatica & Bozzoni. 2014. Nat. Rev. Genet.

Earlham Institute

Annotated but not analysed

- Tens of thousands of loci have been annotated in Eukaryotes genomes
- The function and importance of the vast majority of which remain to be determined
- If biologically relevant the function can be carried out by:
  - The act of transcription over DNA elements
  - The transcript
- A dozen loci have been knocked out and tested in vivo leading to contrasting results:
  - lethality, developmental morphological defects (Xist, Fendrr)
  - phenotypes under specific conditions (BC1)
  - no phenotypes (Visc2)

Tens of thousands of loci – how many are relevant?

- Up to > 100,000 lnRNAs identified depending on publications

- Most are expressed in a single tissue, cell-type at low level

➢**How do we extract likely functional loci from transcriptional noise?**

From identification to validation

Genomes

Annotations

"Omics" data

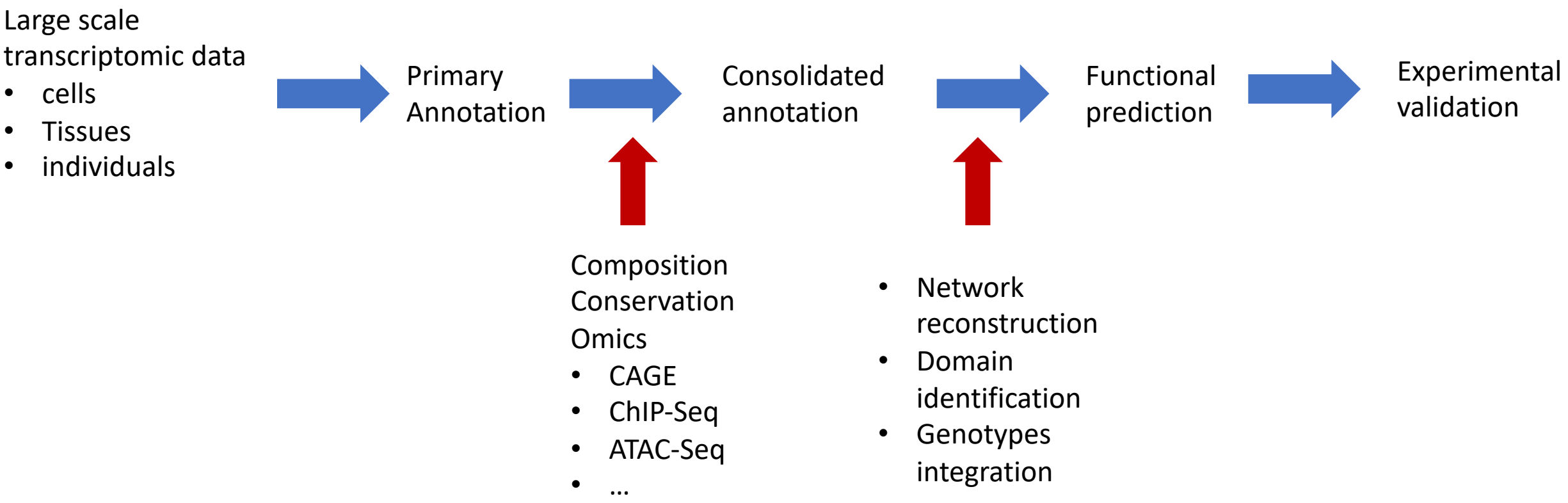
- Transcriptome
- ChIP-Seq
- CAGE

- Individual
  - Cells
  - Tissues
- Population
  - Development
  - Tissues

- Species
  - Shared
  - Specific

- Natural variation
- Knock out / knock down
  - Cellular impact
  - Organismal impact

Omic data integration for functional loci identification

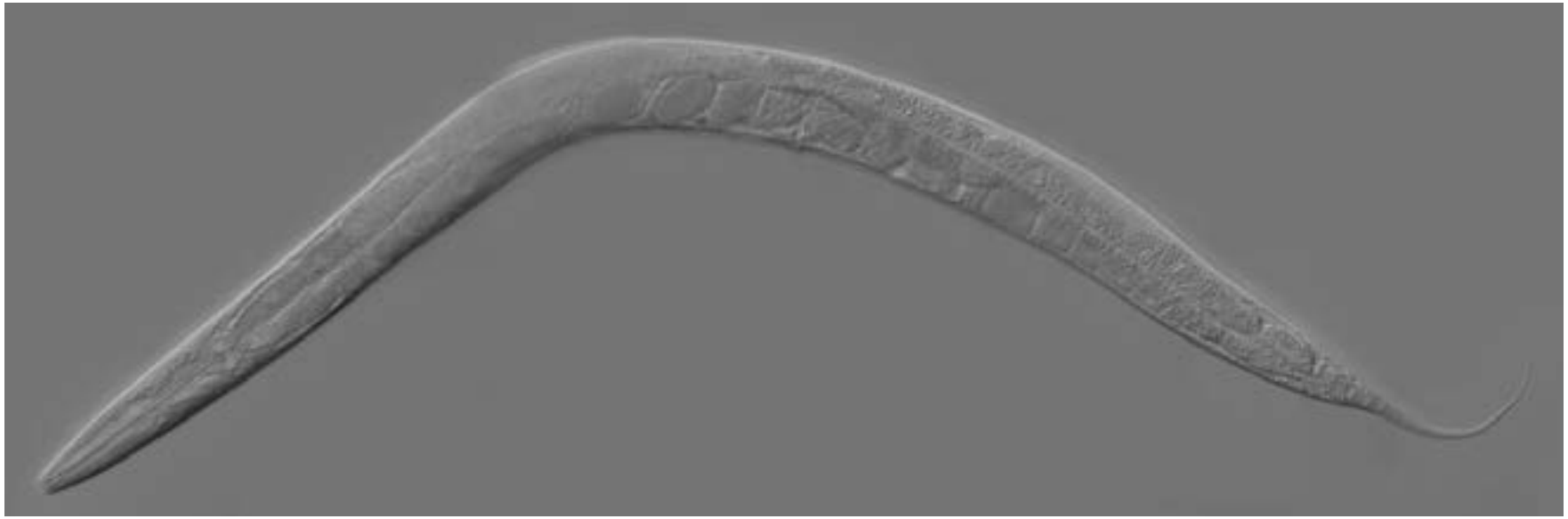


Omic data integration for functional loci identification

- If a lncRNA were to be biologically relevant, one would expect:
  - Reproducible expression between individuals
  - Associated genomic features
  - Phenotype upon disruption

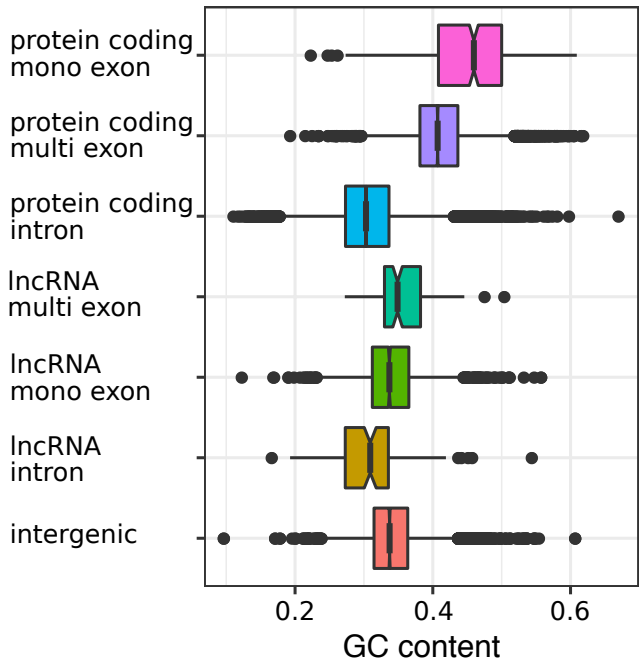
In-vivo phenotyping of knockout / knockdown mutants

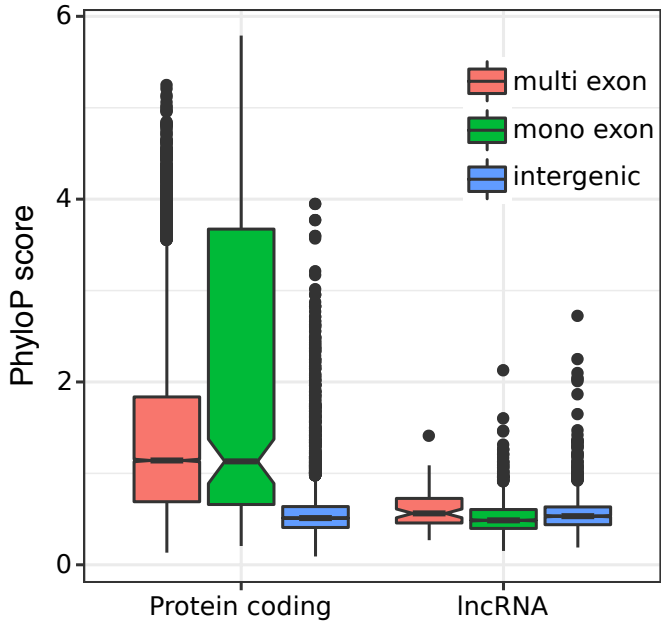
Caenorhabditis elegans



In-vivo phenotyping of knockout / knockdown mutants

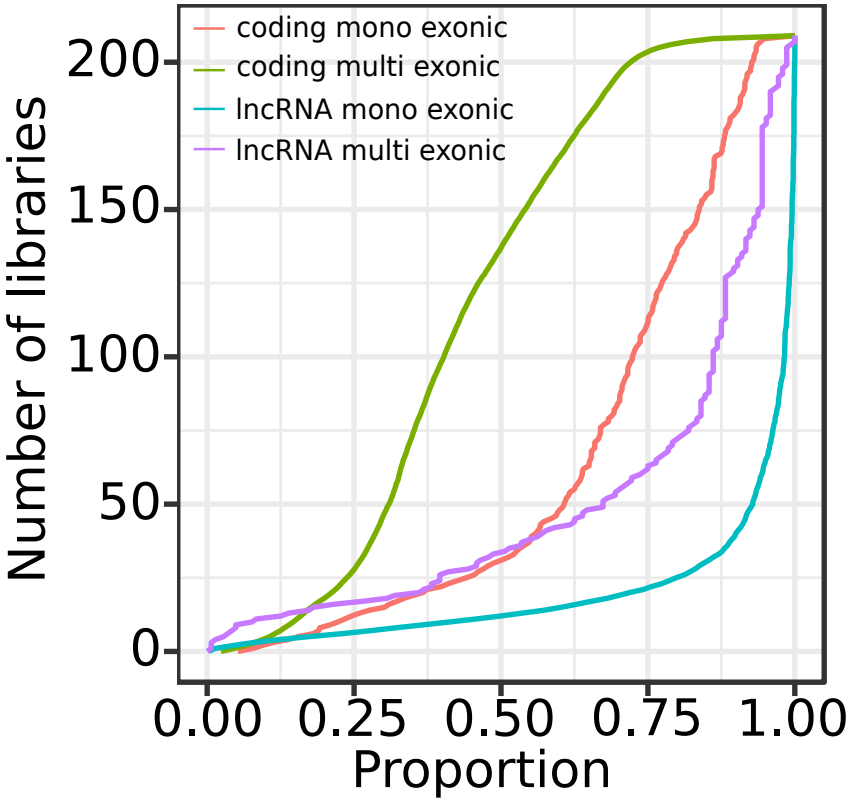
- Annotation of 3,397 lncRNAs using 207 publicly available RNA-Seq libraries
- Integration of all available epigenomic data
  - ChIP-Seq, CAGE-Seq, PAR-CLIP
- Selection intergenic lncRNAs

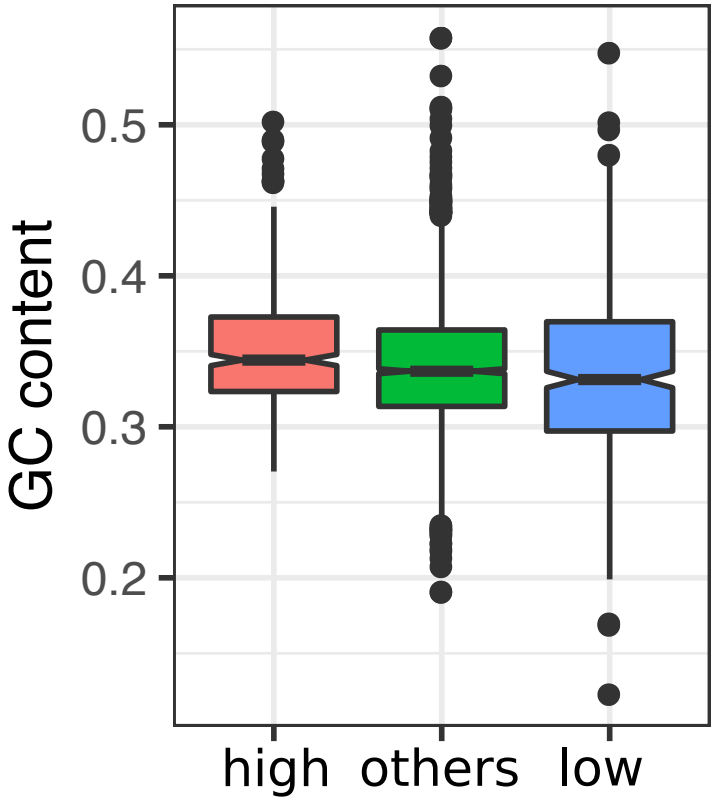


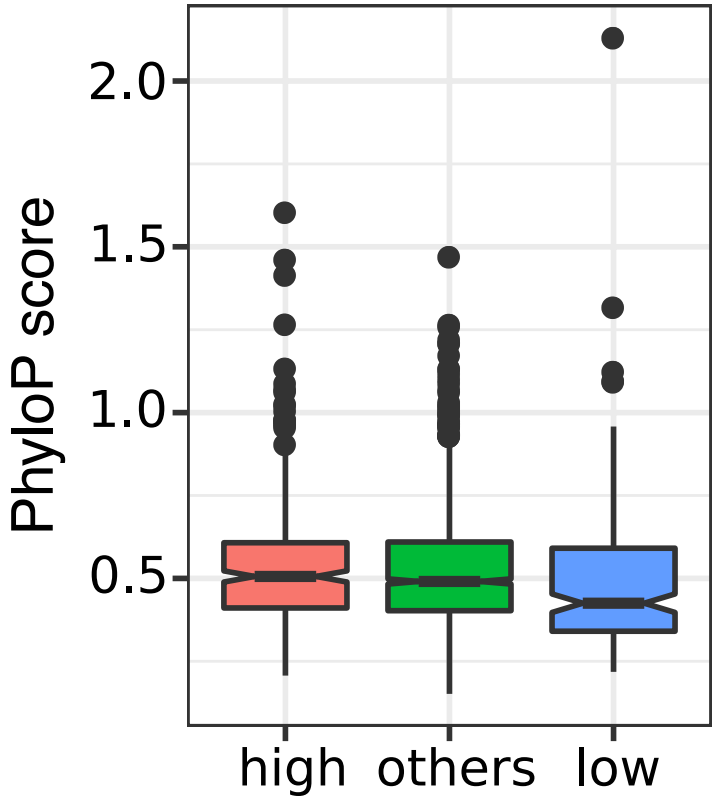


Earlham Institute

In-vivo phenotyping of knockout / knockdown mutants







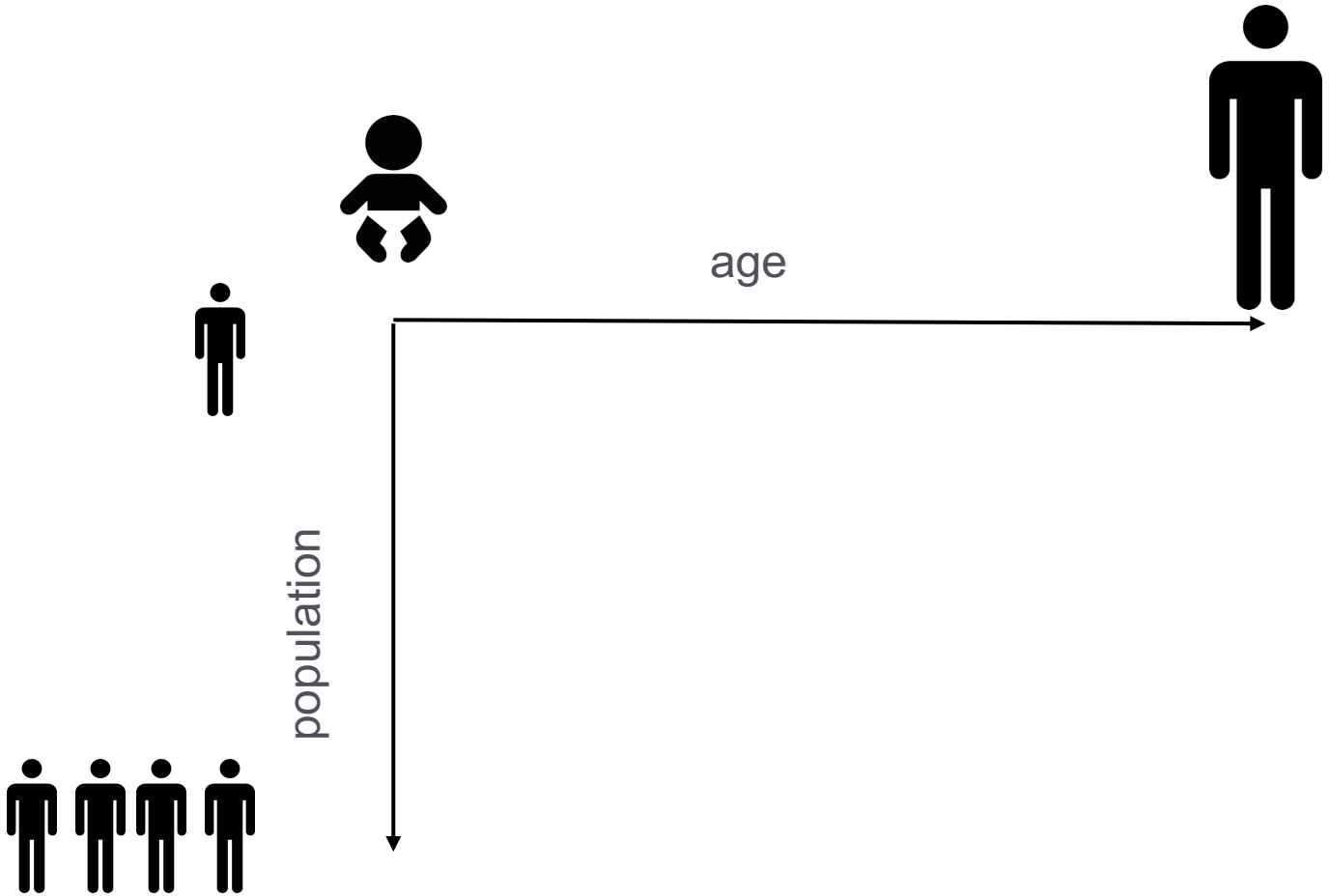
Akay et al. 2020. BMC Biology

In-vivo phenotyping of knockout / knockdown mutants

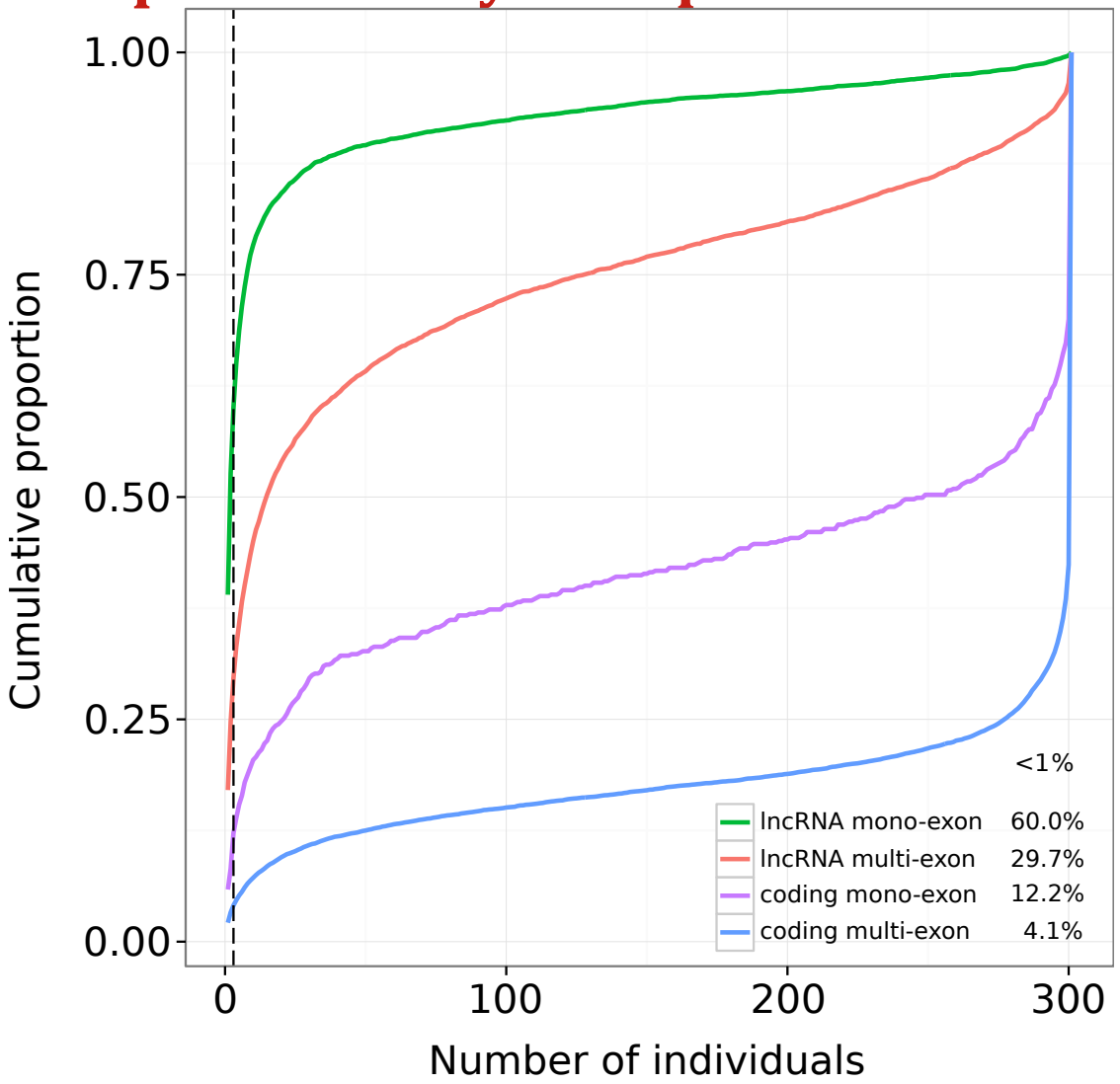
- Novel annotations of lncRNAs in *C. elegans*
- Generation of knockout mutants for 10 multi-exon lncRNAs
  - No evidence for sterility, embryonic lethality or abnormal body development
  - Reduction of brood size for 6 knockouts
  - Reduction of growth rate for 4 mutants
- Phenotypes recapitulated for 2 loci when using knockdown

Akay et al. 2020. BMC Biology

Omic data integration for functional loci identification



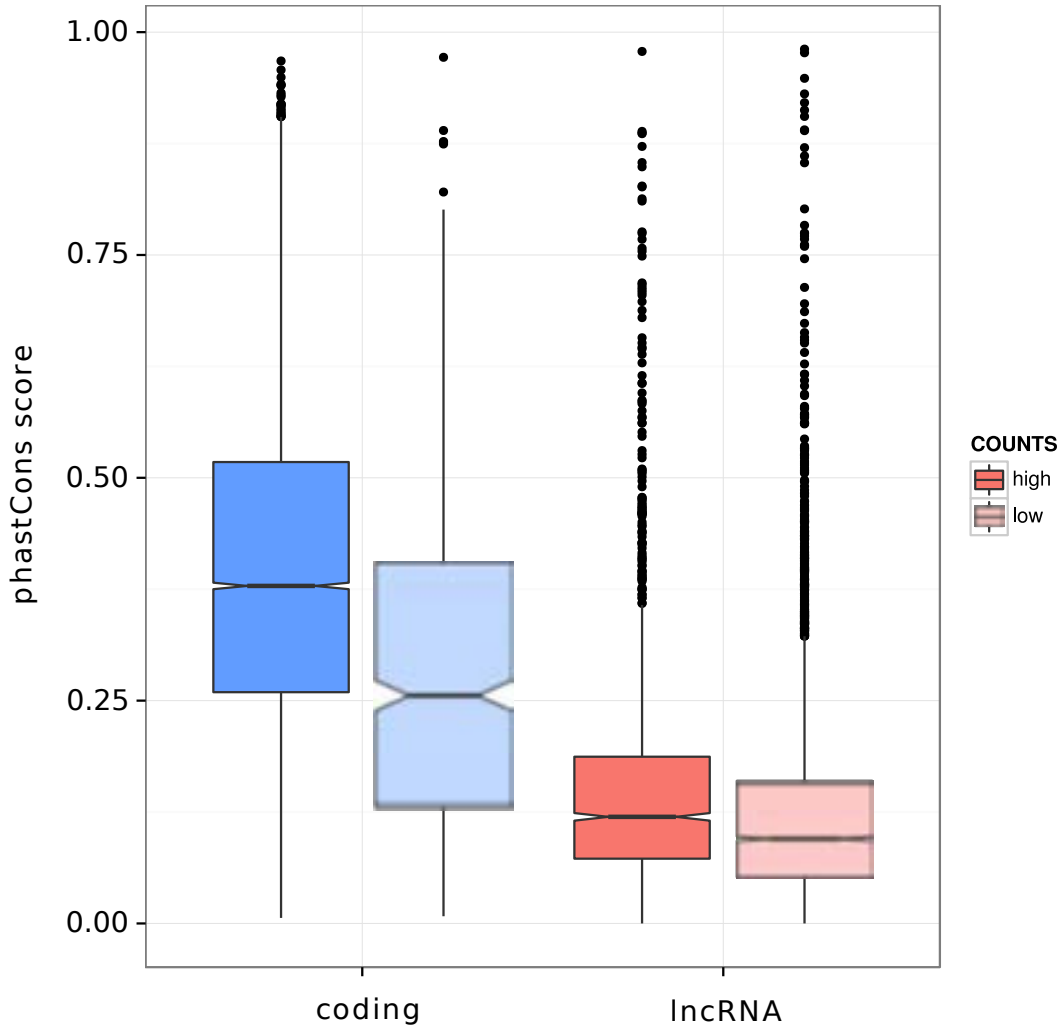
Reproducibility of expression



- 4,232 (21,092) new loci annotated
- up to 65% of lncRNAs found in less than three individuals
- 278 lncRNAs identified in all individuals

Earlham Institute

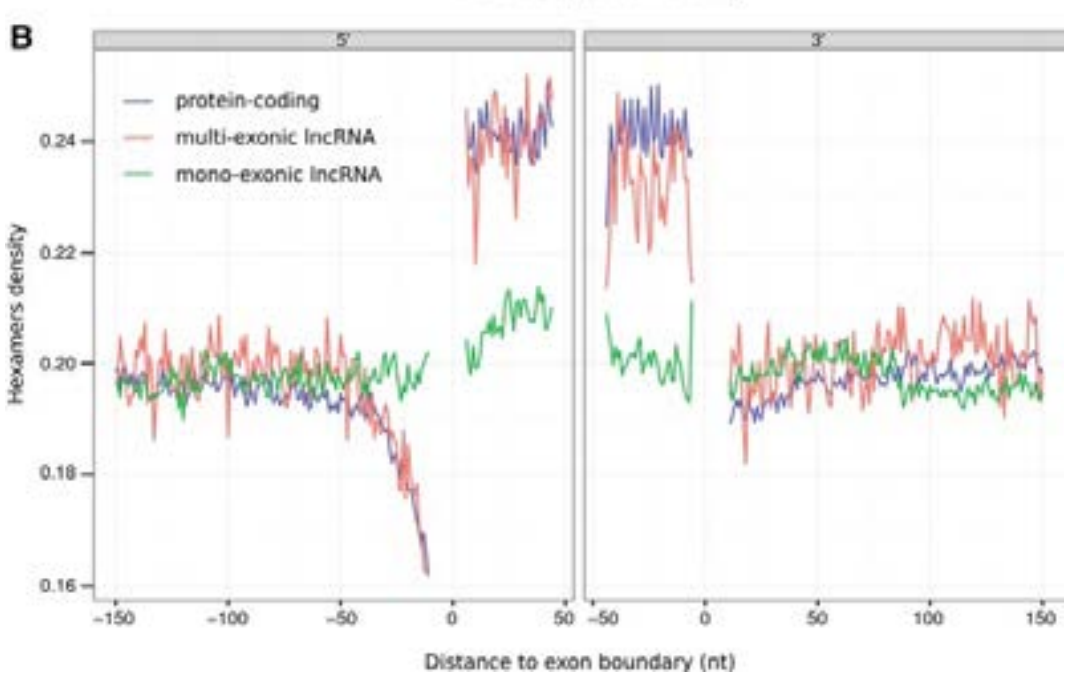
Reproducibility of expression



- Conservation
- Composition
- Epigenetic marks
- eQTLs / GWAS hits

Identification of functional lncRNAs

- Tens of thousands of lncRNAs have been annotated
- Signatures associated with likely functional loci can be detected
  - Expression
  - Reproducibility
  - Nucleotide composition
  - Conservation
  - Chromatin marks
- We have developed approaches to detect motifs (Poddar et al. 2023. arXiv arXiv:2311.12884v1)
- We can predict mechanism (transcript vs transcription)
- Observation of phenotypes upon knockout / knockdown

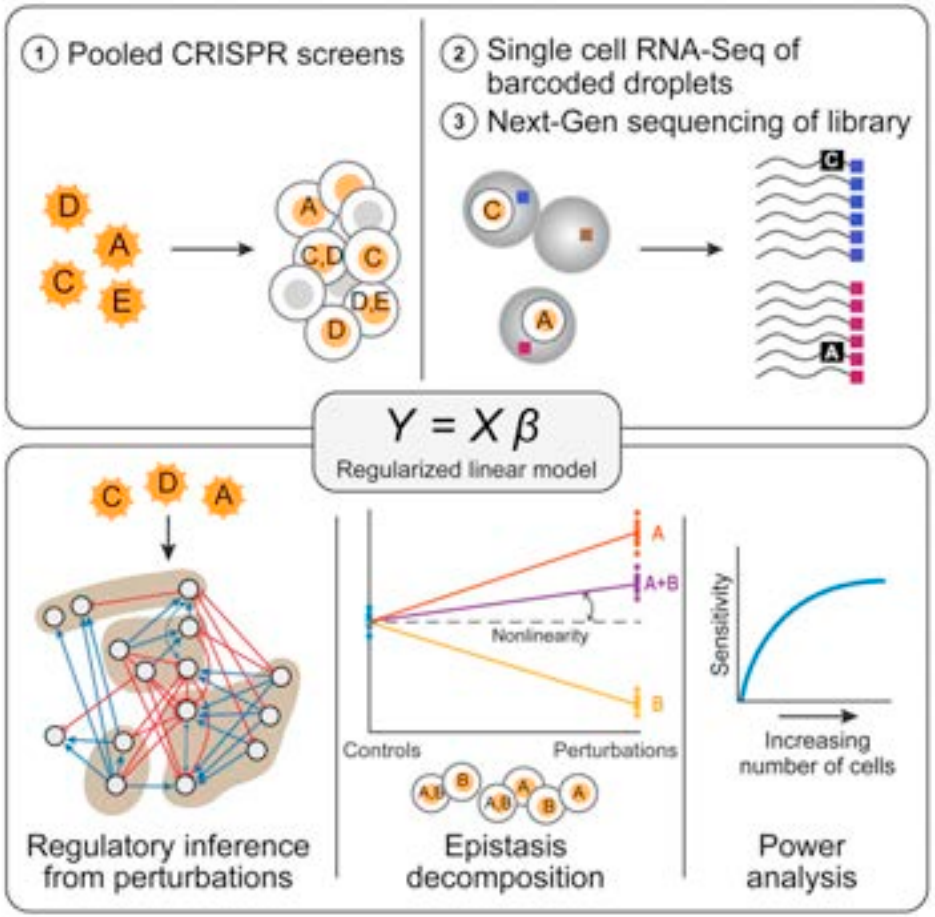


Haerty and Ponting. 2015. RNA

Earlham Institute

From locus identification to function – Need for high-throughput assays

- High-throughput assays using human iPSCs
  - Multimodal Perturb-Seq
    - Dropout assays
    - Positive selection assays
- Use of model organisms for in-vivo phenotyping:
  - Estimation of relative and absolute fitness
  - Effect of interacting genes



Dixit et al. 2016. Cell.

Earlham Institute

Acknowledgments

Earlham Institute

Vladimir Uzun

David Wright

Tomasz Wrzesinski

University of East Anglia

Alper Akay

University of Cambridge

Eric Miska

Edinburgh University

Chris Ponting

Lieber Institute for Brain Development

Daniel Weinberger

UKRI Biotechnology and Biological Sciences Research Council

UKRI Medical Research Council

Earlham Institute, Norwich Research Park, Norwich, Norfolk, NR4 7UZ, UK
www.earlham.ac.uk

Earlham Institute
Decoding Living Systems

Biotechnology and Biological Sciences Research Council

norwich research park

Multiscale modeling of intracellular networks and processes

James R. Faeder

Department of Computational and Systems Biology

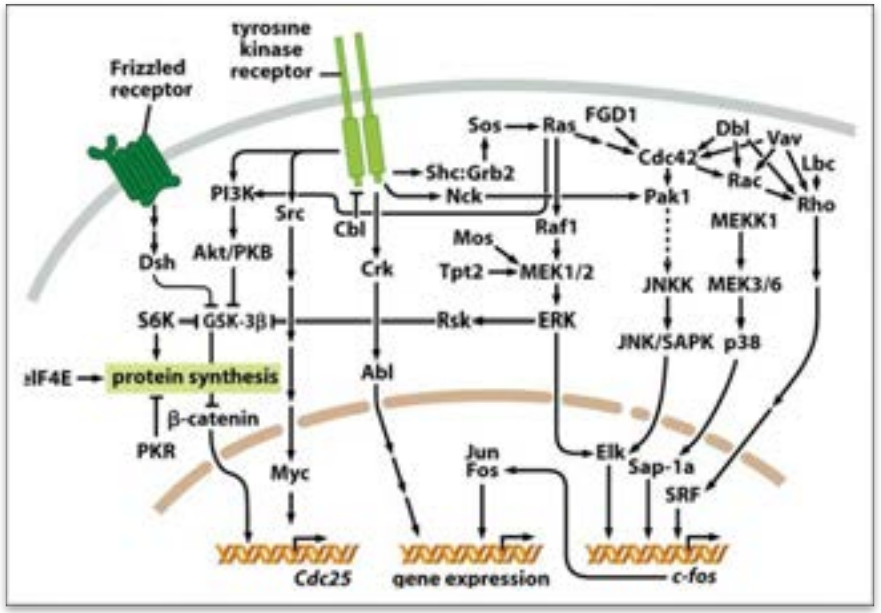
University of Pittsburgh

Discovering Unknome Function (DUF) Workshop
Boston, MA
December 12, 2023

Motivation for studying Cell Decision Processes

Goals

- Develop predictive models of signaling networks
- Understand mechanisms that control outcome
- Exploit understanding to develop new control strategies for medicine and engineering

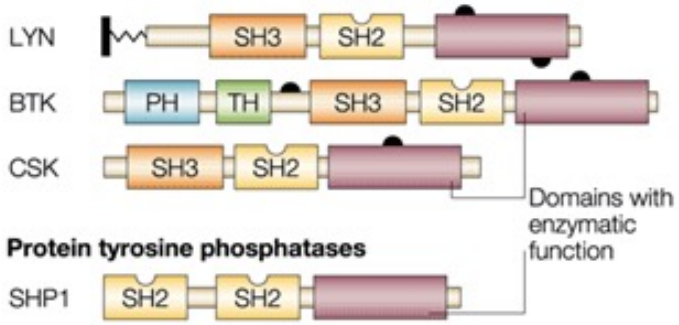


The Biology of Cancer (© Garland Science 2007)

- We use an approach called “rule-based modeling” (RBM) to build and simulate models
- BioNetGen is software for rule-based modeling that our group maintains and develops

Challenges of modeling cell regulatory networks

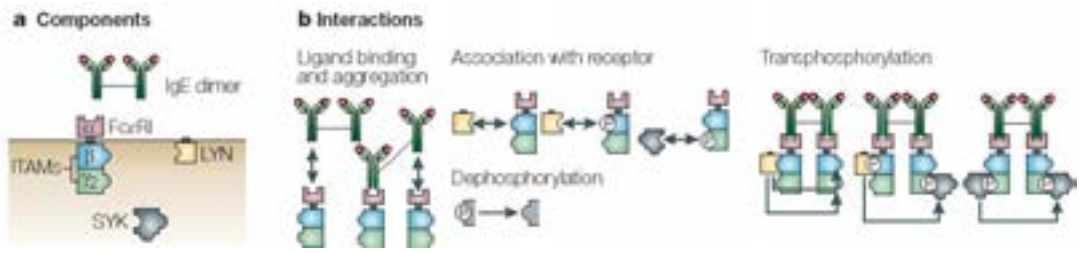
- Proteins are multi-functional

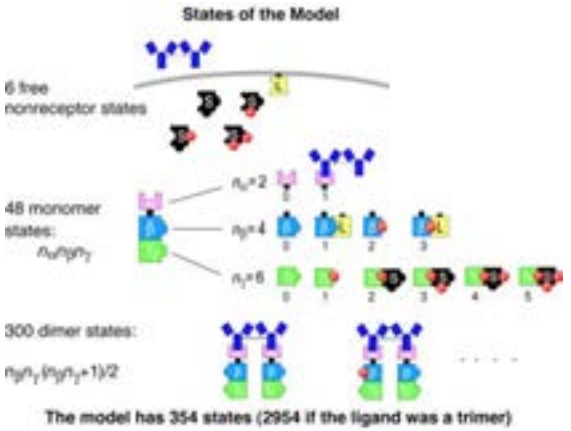


multiple sites of binding

multiple sites of posttranslational modification

- Representing their known interactions requires handling of *combinatorial complexity*



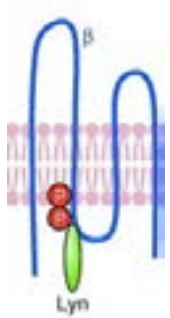


Small number of rules

Small number of components and interactions ➔huge number of possible species and reactions

What is Rule-based Modeling (RBM)?

Rules define the interactions of molecules



"Lyn SH2 domain binds to phosphorylated Tyr 218 on the β subunit of FcεRI "

Reactants **Products** **Rate Law**

Rule: `LYN(SH2) + FceRI(bY_218~P) <-> LYN(SH2!1).FceRI(bY_218~P!1) kpL, kmL`

bond

"Don't write don't care" – elements not mentioned may be in any state

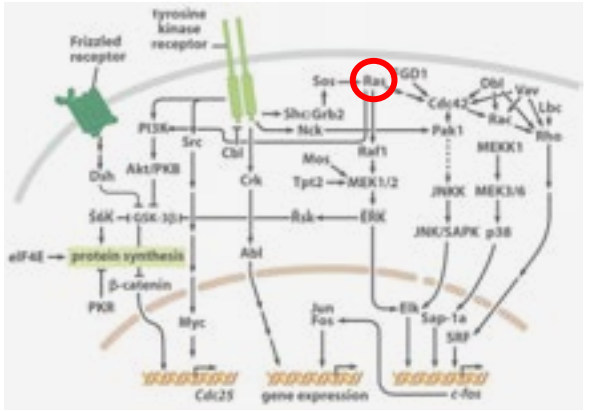
➔ *One rule can generate reactions involving many different species*

Reaction rate determined by **Mass Action kinetics**

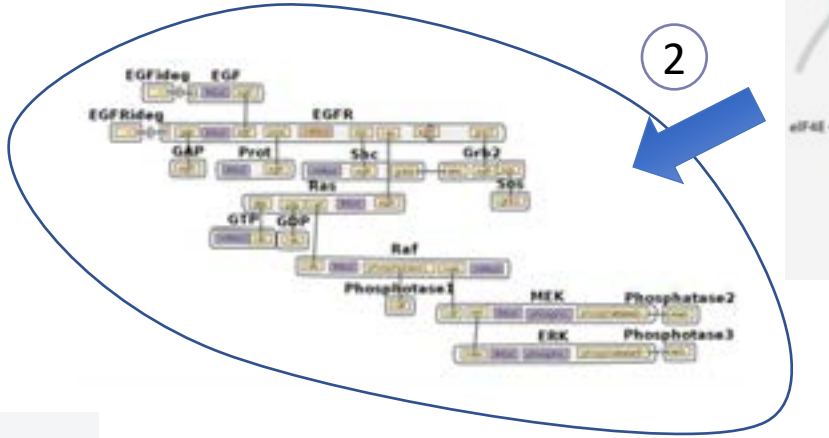
rate forward = `kpL*[Lyn(SH2)]*[FceRI(bY_218~P)]`

Rules bridge between molecular and cellular scales

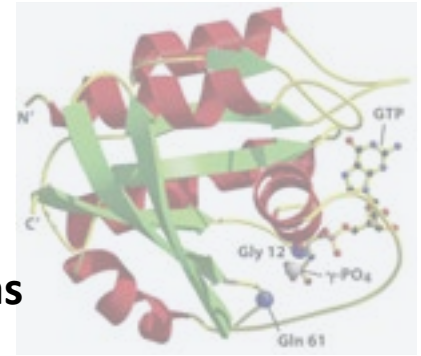
Network scale



Rule-based model



2



Ras

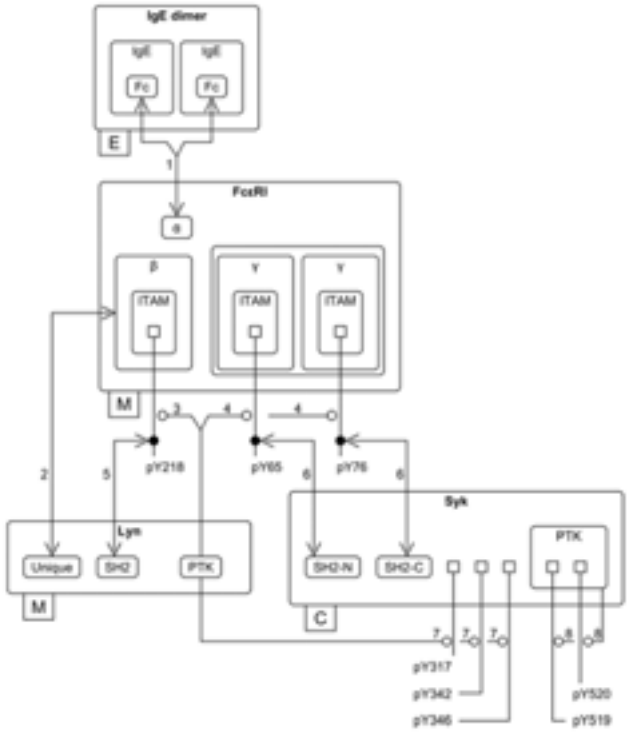
1

2

Molecular scale

Rule-based modeling enables knowledge representation on a large scale

FcεRI model

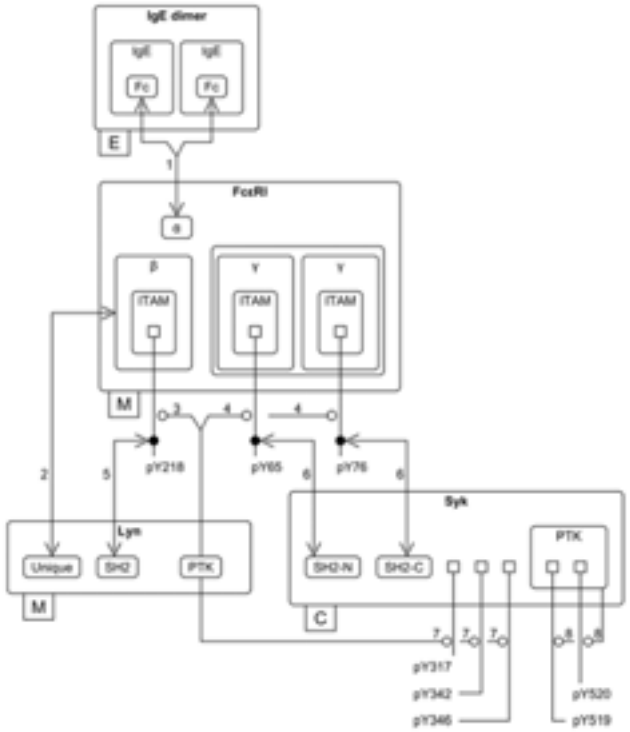


- Precise encoding of modeled structures and interactions
- User avoids combinatorial complexity
- Amenable to visualization
- Extensible as knowledge base grows

Faeder et al., *J. Immunol.* (2003)
Chylek et al., *Mol. BioSys.* (2011)

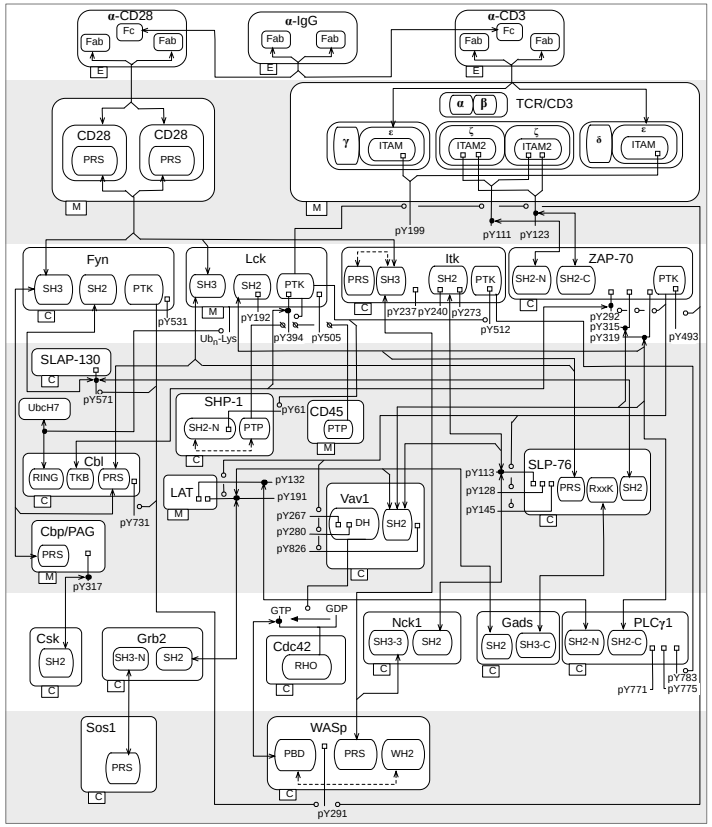
Rule-based modeling enables knowledge representation on a large scale

FcεRI model



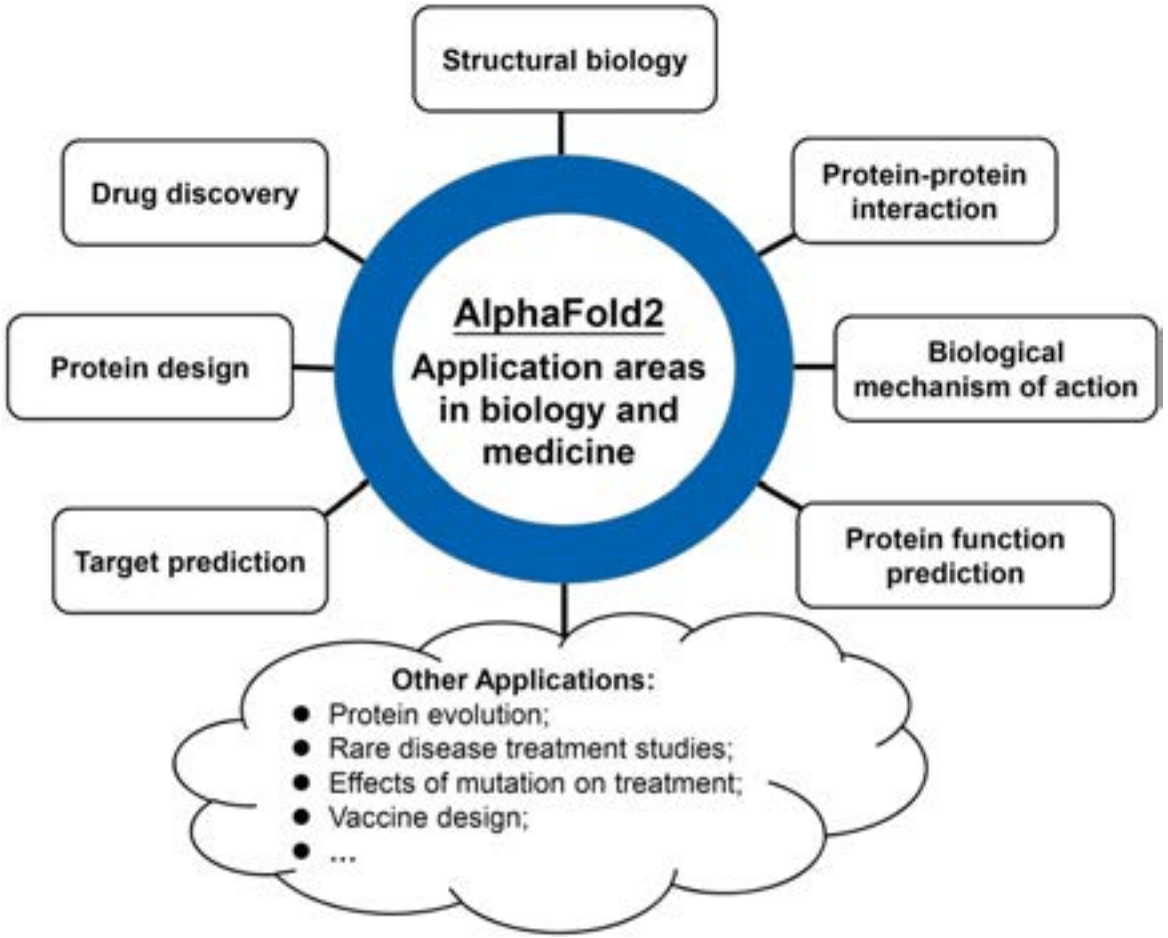
Faeder et al., *J. Immunol.* (2003)

T cell receptor model



Chylek et al., *PLOS One* (2014)

AI Technologies Enabling the Development of Large Scale Models



AI Technologies Enabling the Development of Large Scale Models

IN DEPTH | ARTIFICIAL INTELLIGENCE

DARPA sets out to automate research

Crash program aims to teach computers to read journals and hatch new ideas.

JIA YOU Authors Info & Affiliations

SCIENCE • 30 Jan 2015 • Vol 347, Issue 6221 • p. 465 •

Method

TRANSPARENT PROCESS OPEN ACCESS

molecular systems biology

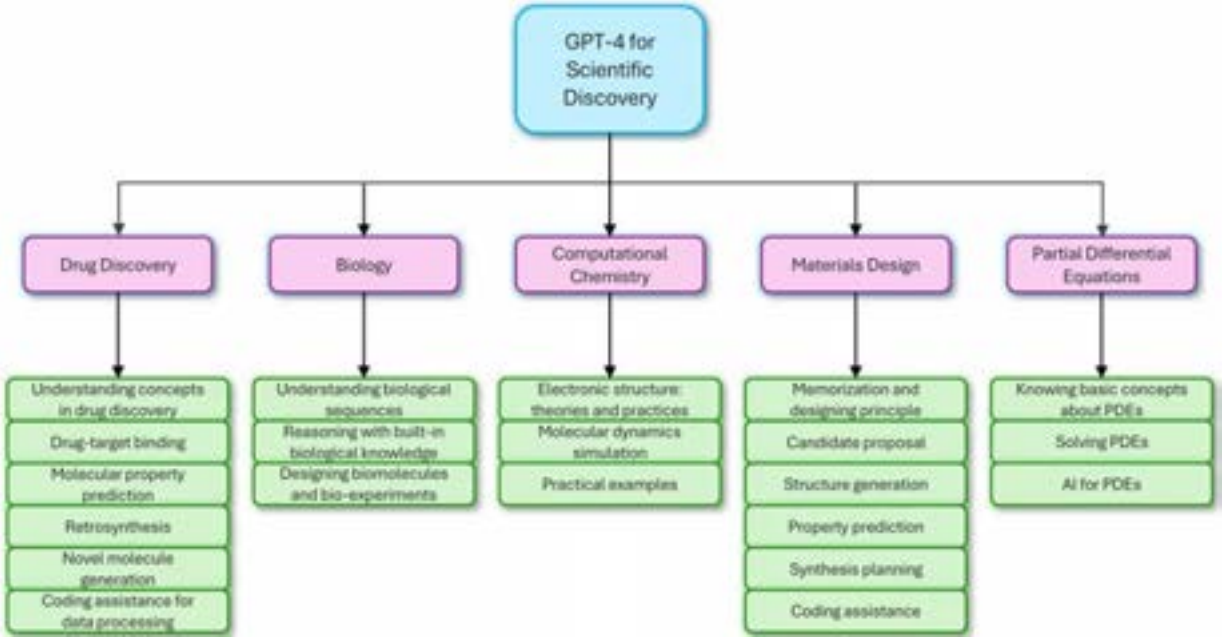
Automated assembly of molecular mechanisms at scale from text mining and curated databases

John A Bachman[1,†], Benjamin M Gyori[1,*,†] & Peter K Sorger[1,2,**]

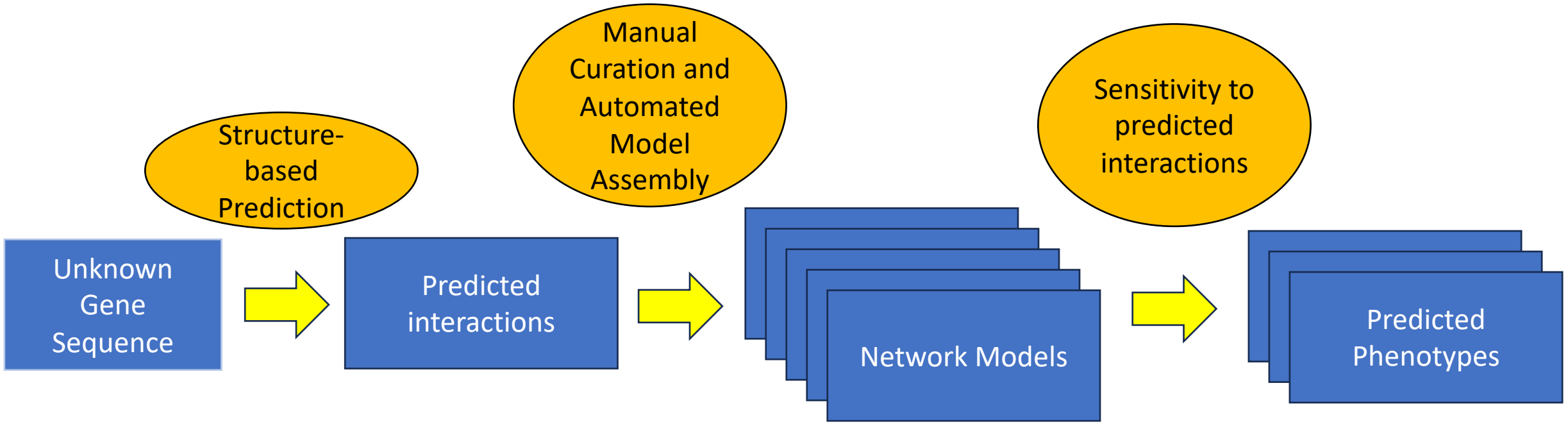
AI Technologies Enabling the Development of Large Scale Models

The Impact of Large Language Models on Scientific Discovery: a Preliminary Study using GPT-4

Microsoft Research AI4Science
Microsoft Azure Quantum
llm4sciencediscovery@microsoft.com
November, 2023



Dream/Vision



Building Reproducible Pipelines

December 12, 2023

Olaitan I. Awe, PhD

@laitanawe

@laitanawe

What is reproducible code?

Code is reproducible if:

- the result of an analysis does not depend on the specific computational environment in which data processing and analysis originally took place
- Workflow will produce the same result when re-run or run on different computing platforms

Framework for reproducible code

1. Collect data
2. Develop the pipeline/codes
3. Generate Output
4. Interpret the Output

Framework (Omic Data Science)

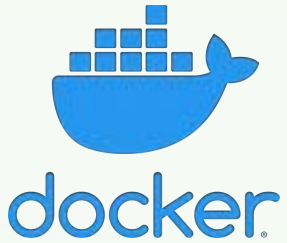
1. Collect data (Biomedical, Omic Sequences etc.)
2. Curate Data
3. Develop the pipeline/code
4. Interpret the Output and Present data (advance our understanding of biology and health)

Write Code and Publish it in a findable Repository (GitHub)

1. Data
2. Accessions (SRA, GEO, ENA, RefSeq, Genbank)
3. Figures
4. Scripts
5. Docs
6. Output
7. Workflow (Step by Step sequence of tasks)
8. Notebooks for Demonstration
9. LICENSE (Open license)
10. README.md

Workflow Management Systems enable Reproducible Coding

1. Nextflow (Interoperability, Component Reuse, Re-entrancy, Parallelisation, Allows use of containers, Reproducibility)



2. Snakemake (Python)
3. Cromwell (WDL/CWL)
4. Galaxy

Automate your Pipelines

Language depends on what you're comfortable with and your application:

1. Bash
2. Python
3. Perl
4. Java
5. C/C++ and others …

Some Life Science Project Categories

1. Bulk Transcriptomics, Metagenomics, Human Genomic Variation, Pipeline Development, Biomarker Discovery, Cheminformatics, Clinical Applications, Drug and Vaccine Design, Antimicrobial Resistance, Population Genomics, Genome Wide Association Studies, Polygenic Risk Scores, Mendelian Randomisation, Structural Bioinformatics, Software Development, Epigenomics, Oncology, Plant Genomics and Machine Learning.

Want to start writing reproducible code?

- You can start practicing by using public data (SRA, GEO, ENA, RefSeq, Genbank)

Research Standard

Open Science:

1. Improve the accessibility, quality and efficiency of science
2. Open Access Articles (*APC can be expensive*)
3. Research data, code and pipelines are FAIR:

(**F**indable, **A**ccessible, **I**nteroperable, **R**eusable)

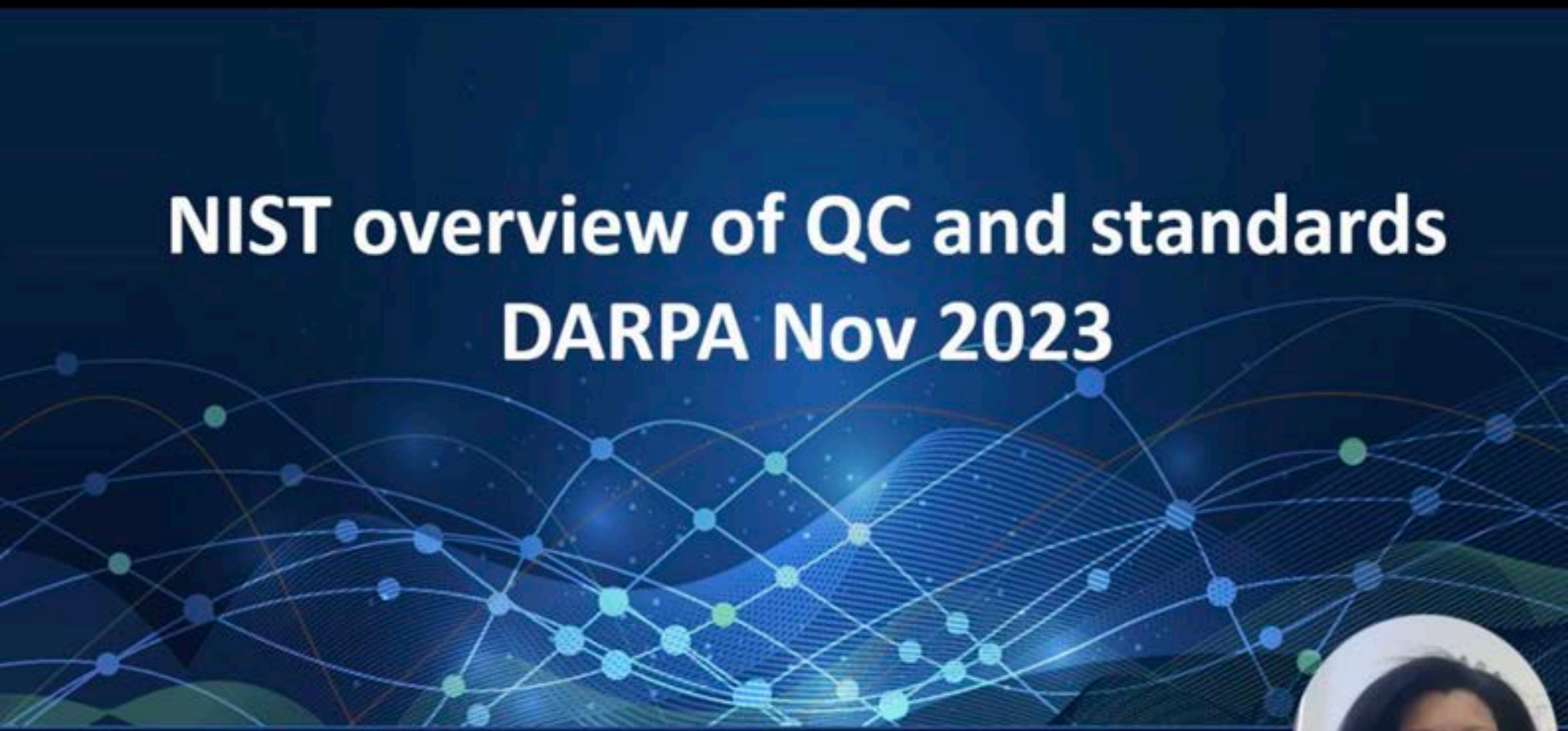
Documentation: add comments to your code

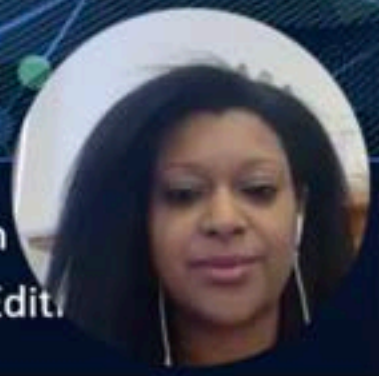
If we're not sharing our data when annotating these unknown genes, it's not helpful.

Thank you!

laitanawe@gmail.com

NIST overview of QC and standards
DARPA Nov 2023





NIST National Institute of Standards and Technology
U.S. Department of Commerce

Samantha Maragh
Leader, Genome Edit

Pacific Northwest
NATIONAL LABORATORY

Beyond the genome: multi-omics across scales

Kristin Burnum-Johnson

EMSL
SCIENTIFIC INNOVATION THROUGH INTEGRATION

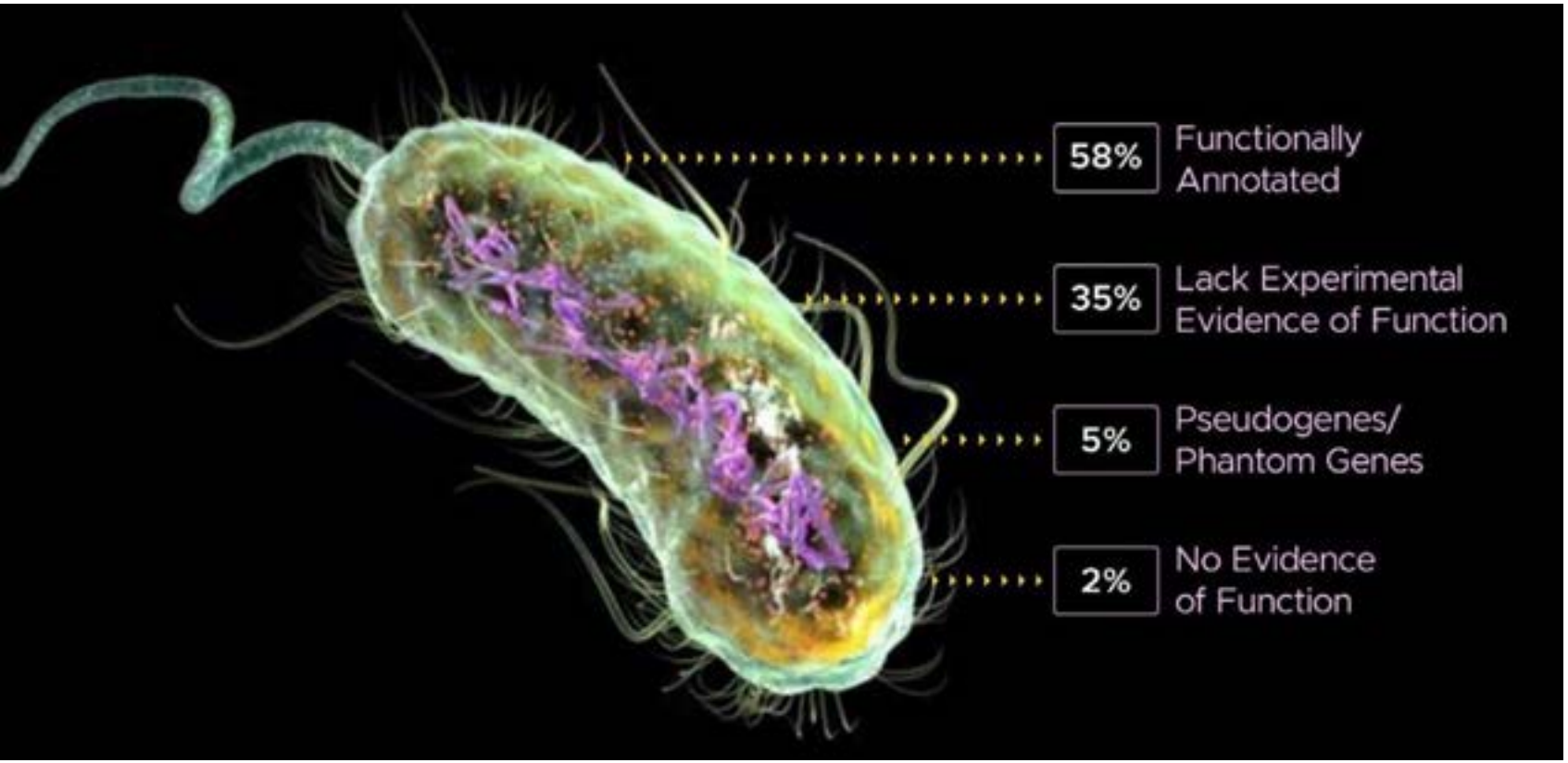
U.S. DEPARTMENT OF ENERGY BATTELLE

PNNL is operated by Battelle for the U.S. Department of Energy

Basics of Biological Function

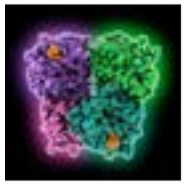
Phenome (n). The set of all phenotypes expressed by a cell, tissue, organ, organism, or species.

Genome-only based strategies only reveal part of the picture



est. prokaryotes on Earth
2,200,000 – 4,300,000

sequenced bacterial genomes
198,640



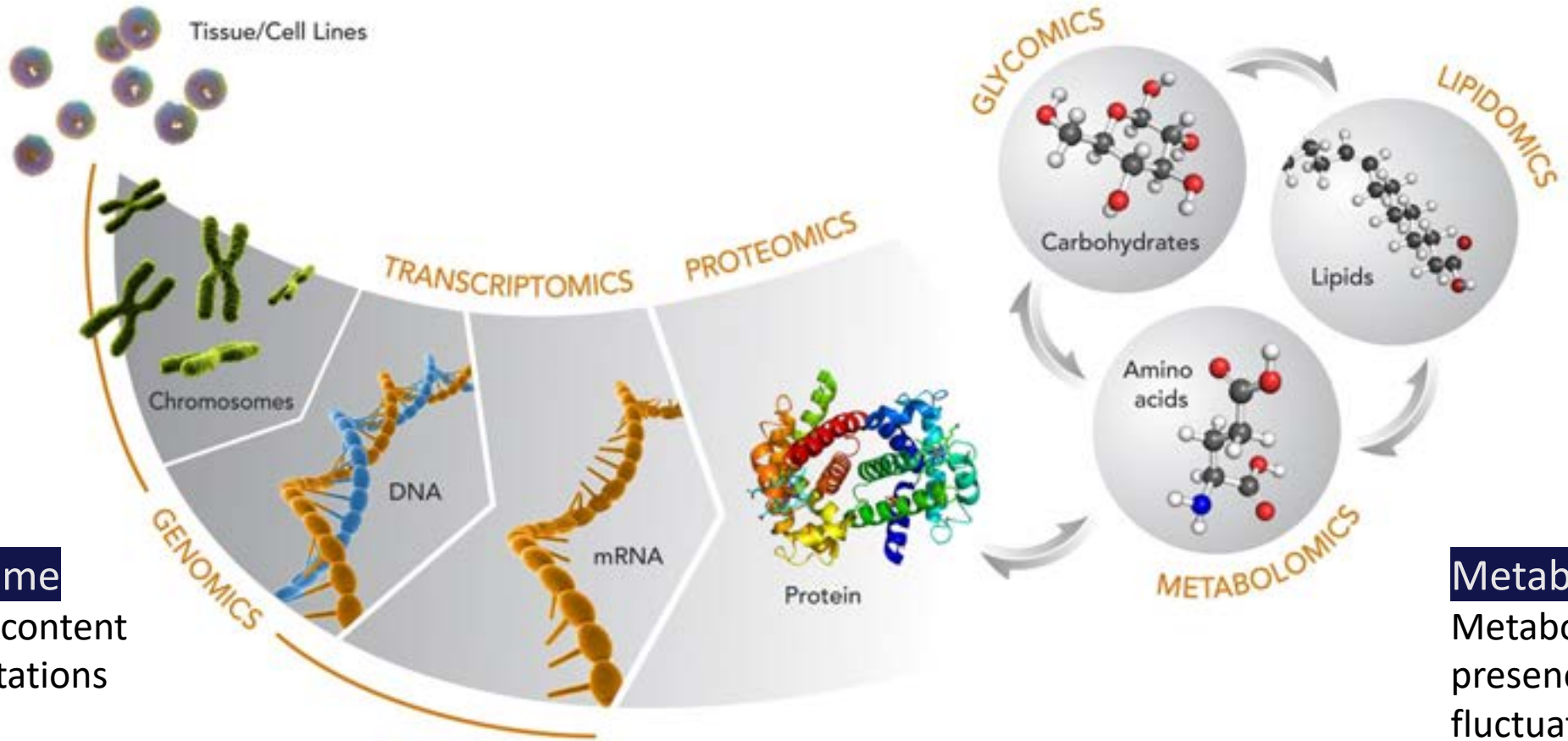
non-redundant proteins identified
154,000,000

new protein sequences added
~3,200,000/month

- **Our ability to READ DNA far surpasses our ability to UNDERSTAND the information it contains**

- **Vast majority of genes have unknown/non-validated functional annotation for proteins encoded**
  - 6,000 of the human genome's ~20,000 genes are still unknown
    - 70% of 154M microbial proteins are unannotated

The flow of molecular information → phenotype



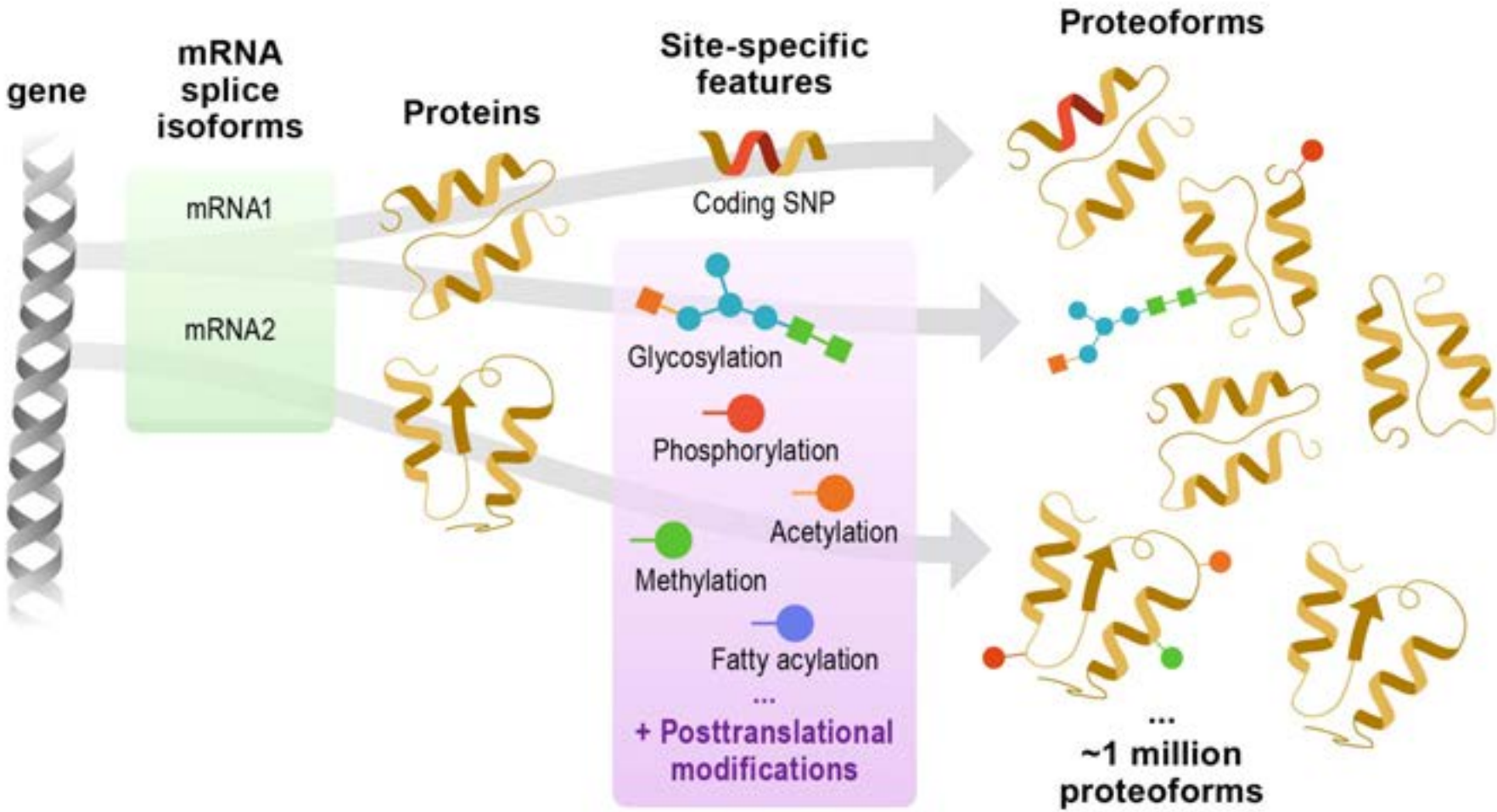
Genome
Gene content & mutations

Transcriptome
Gene regulation & expression

Proteome
Protein content, activity & post-translational modification

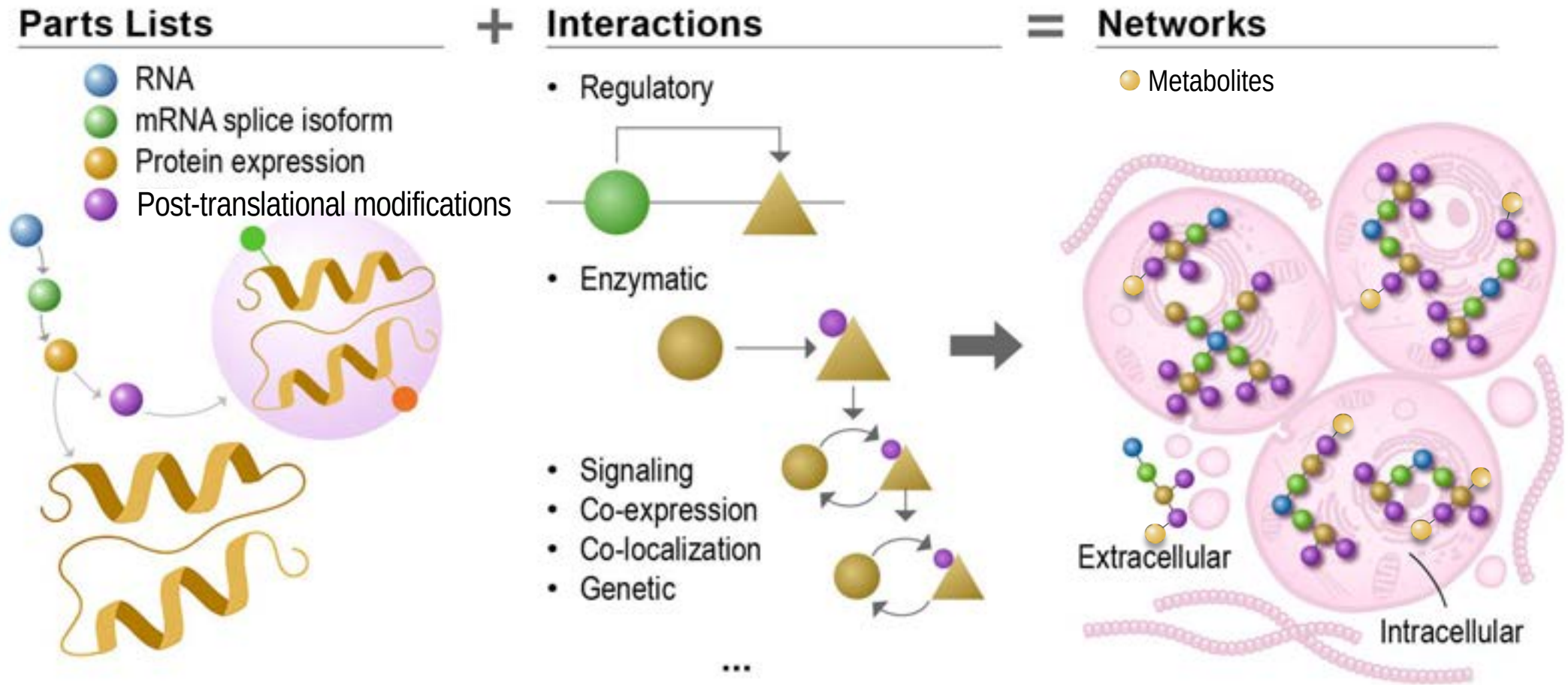
Metabolome
Metabolite presence and fluctuations

The proteome conveys function



Burnum-Johnson, et al., New views of old proteins: clarifying the enigmatic proteome. 2022 21DOI: (10.1016/j.mcpro.2022.100254)

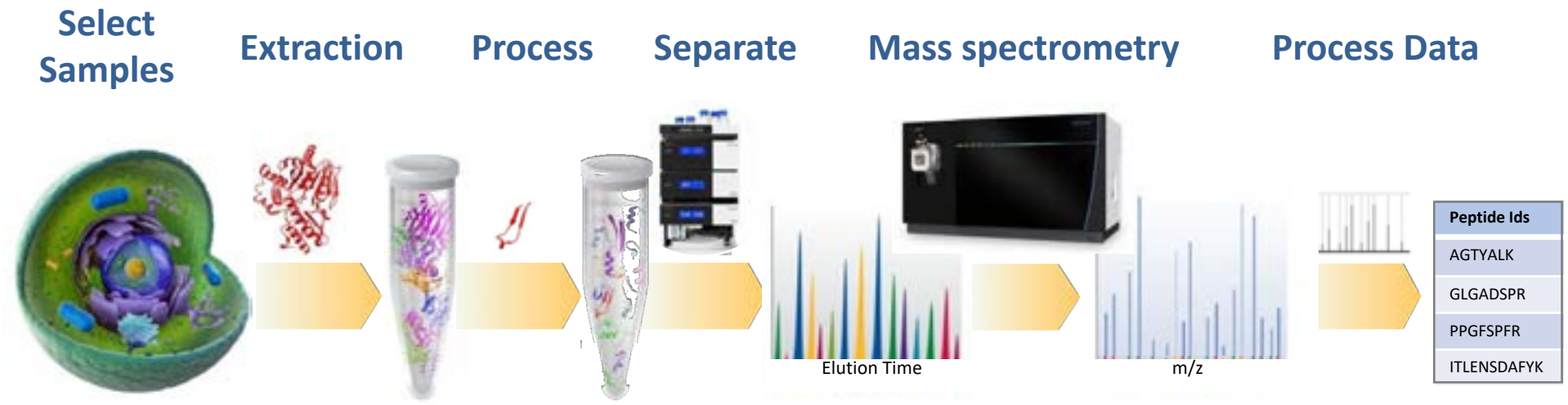
Understanding biological functions through molecular networks



Adapted from Burnum-Johnson, et al., New views of old proteins: clarifying the enigmatic proteome. 2022 21DOI: (10.1016/j.mcpro.2022.100254)

Generalized approach for MS-based omics

Mass Spectrometry



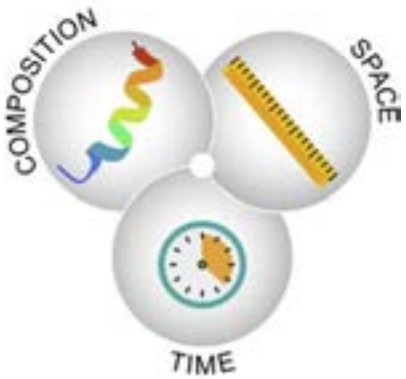
Digest/Leave intact
Isotopically label
Enrich
Chemically block

Fractionation
Long/short gradients
Ion mobility
Orthogonal solid phases

Dissociation types
Label Free/Labeling
Database/Tag
SRM/PRM
DIA/DDA

The biological question drives the approach
Discovery proteomics, targeted proteomics, Post-translational modification (PTM), etc.

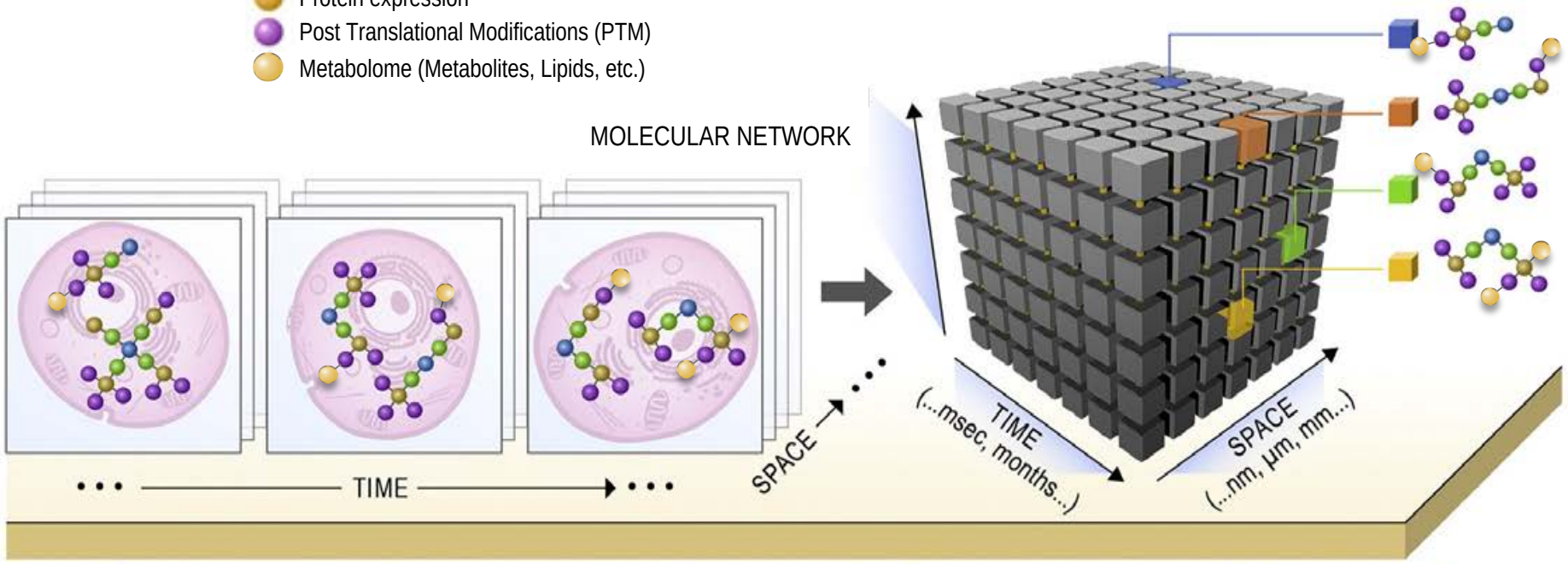
Capturing multidimensional biology



PARTS LIST

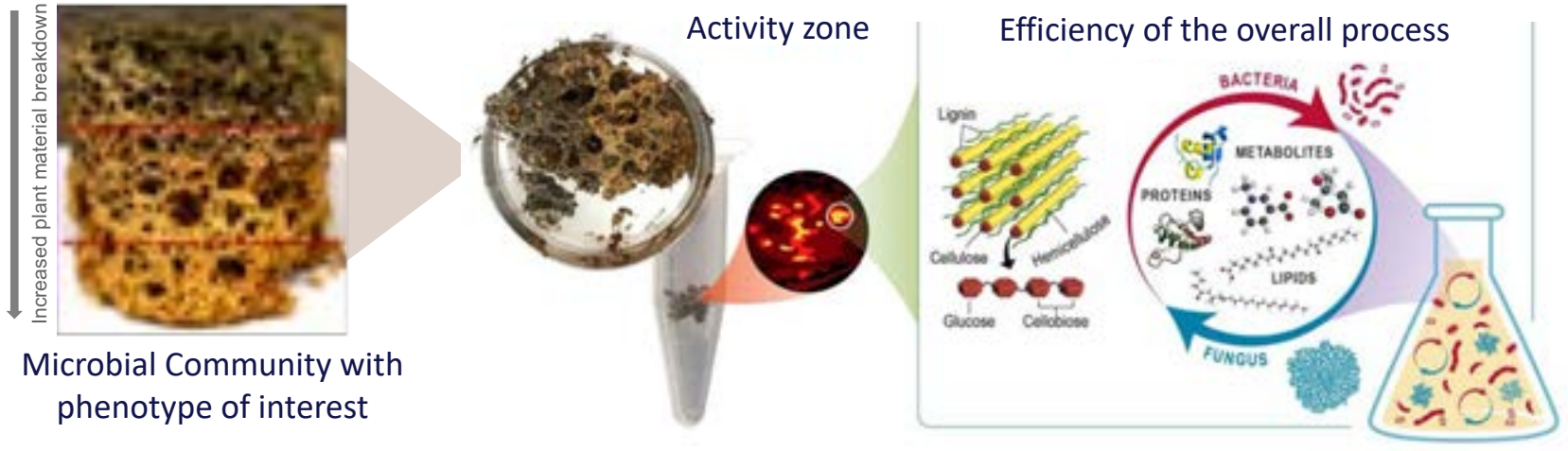
- RNA
- mRNA splice isoform
- Protein expression
- Post Translational Modifications (PTM)
- Metabolome (Metabolites, Lipids, etc.)

MOLECULAR NETWORK

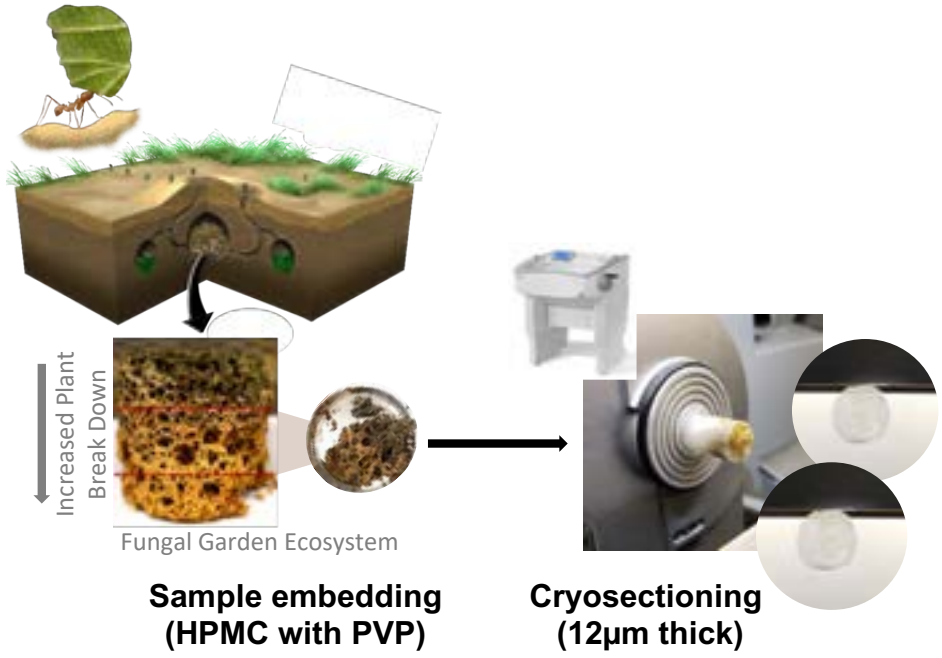


Why study molecular networks in a spatially constrained manner

- Most phenotypes are observed at a global level
- Many cell types or species contribute differentially to the global phenotype.
- Increasing the spatial granularity of the measurements enables the understanding of how each component of a system contributes to the overall phenotype.



Metabolome Informed Proteome Imaging (MIPI)

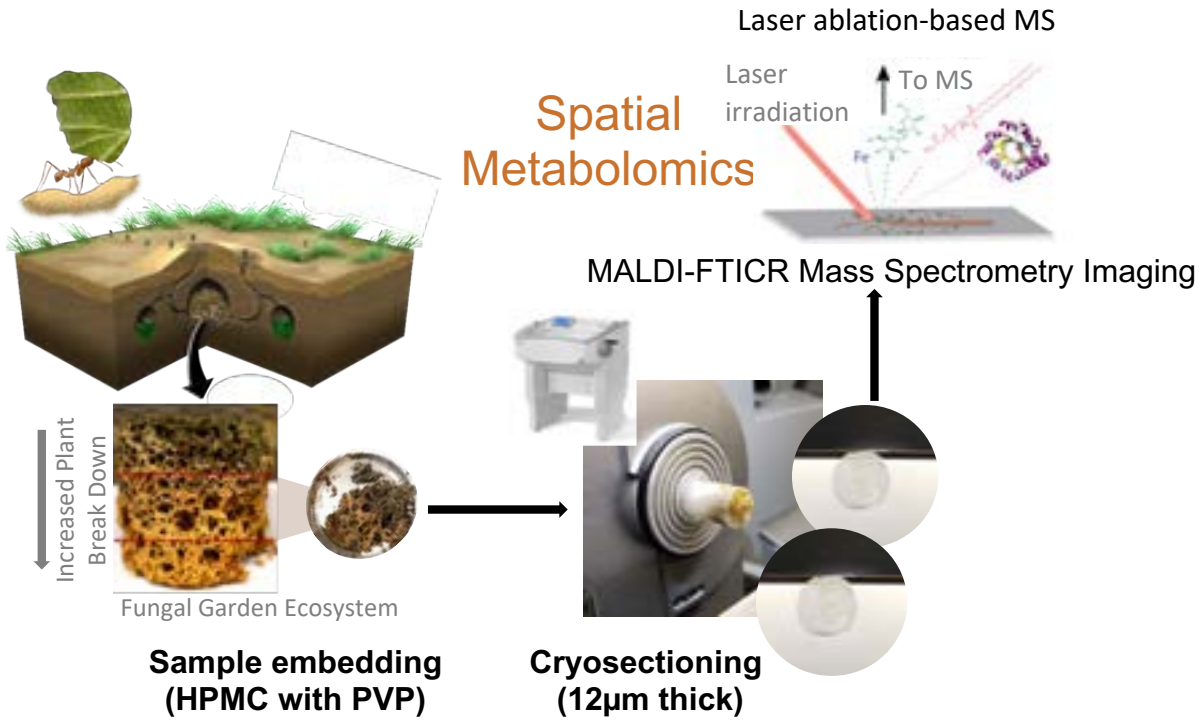


- Map what microbes, enzymes, proteins, lipids, metabolites and activities can be correlated with microscale regions in this ecosystem
- Perform lipidomics, metabolomics, & proteomics on 12-micron thick fungal garden sections
- Obtain mechanistic knowledge on how lignocellulose is degraded in this ecosystem

U.S. DEPARTMENT OF ENERGY

Marija Veličković, Ruonan Wu, Yuqian Gao, M. Thairu, D. Veličković, N. Munoz, C. Clendinen, A. Bilbao, R. Chu, P. Lalli, K. Zemaitis, C. Nicora, J. Kyle, D. Orton, S. Williams, Y. Zhu, R. Zhao, M. Monroe, R. Moore, B.-J. Webb-Robertson, L. Bramer, C. Currie, Paul Piehowski, K. Burnum-Johnson. Mapping Microhabitats of Lignocellulose Decomposition by a Microbial Consortium. *In press* Nature Chemical Biology (2023)

Metabolome Informed Proteome Imaging (MIPI)



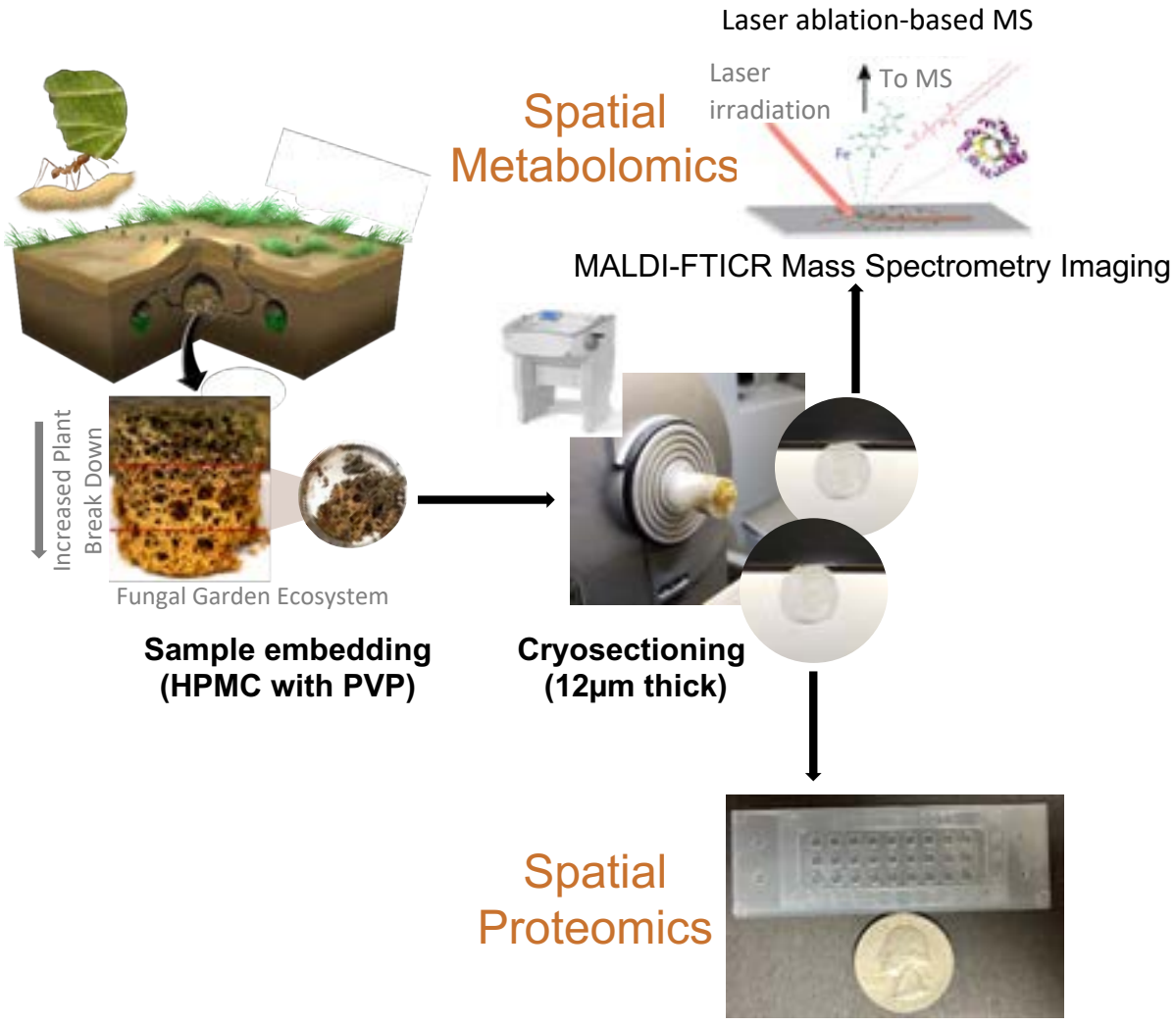
Spatial Metabolomics

- Matrix-assisted laser desorption/ ionization (MALDI) Mass Spectrometry Imaging profiles metabolites with a spatial resolution of 50-microns and correlate morphologically unique features with *metabolome hotspots* of interest

U.S. DEPARTMENT OF ENERGY

Marija Veličković, Ruonan Wu, Yuqian Gao, M. Thairu, D. Veličković, N. Munoz, C. Clendinen, A. Bilbao, R. Chu, P. Lalli, K. Zemaitis, C. Nicora, J. Kyle, D. Orton, S. Williams, Y. Zhu, R. Zhao, M. Monroe, R. Moore, B.-J. Webb-Robertson, L. Bramer, C. Currie, Paul Piehowski, K. Burnum-Johnson. Mapping Microhabitats of Lignocellulose Decomposition by a Microbial Consortium. *In press* Nature Chemical Biology (2023)

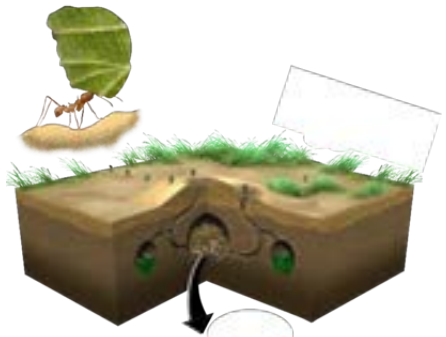
Metabolome Informed Proteome Imaging (MIPI)



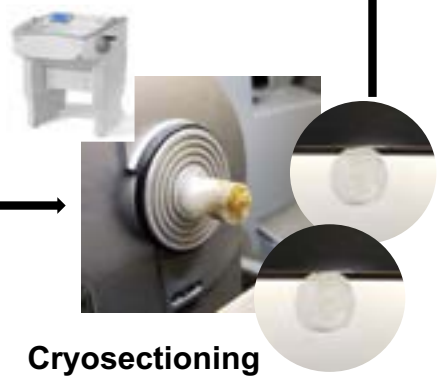
Spatial Proteomics

- Tissue regions containing these activity zones are liberated from the slides with laser capture microdissection and processed in our PNNL developed Microdroplet Processing in One-Pot for Trace Samples (MicroPOTS) chip for high sensitivity mass spectrometry proteomics

Metabolome Informed Proteome Imaging (MIPI)



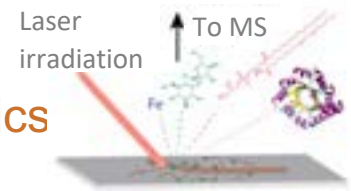
Sample embedding (HPMC with PVP)



Cryosectioning (12µm thick)

Spatial Metabolomics

Laser ablation-based MS



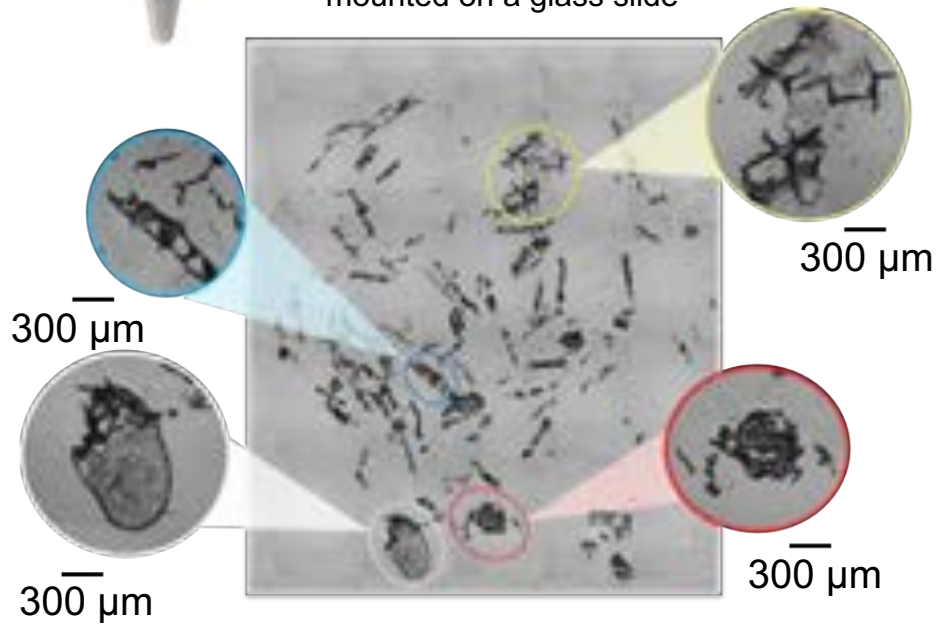
MALDI-FTICR Mass Spectrometry Imaging

Spatial Proteomics

MicroPOTS
Microdroplet Processing in One-Pot for Trace Samples

Microscale multi-omic mapping

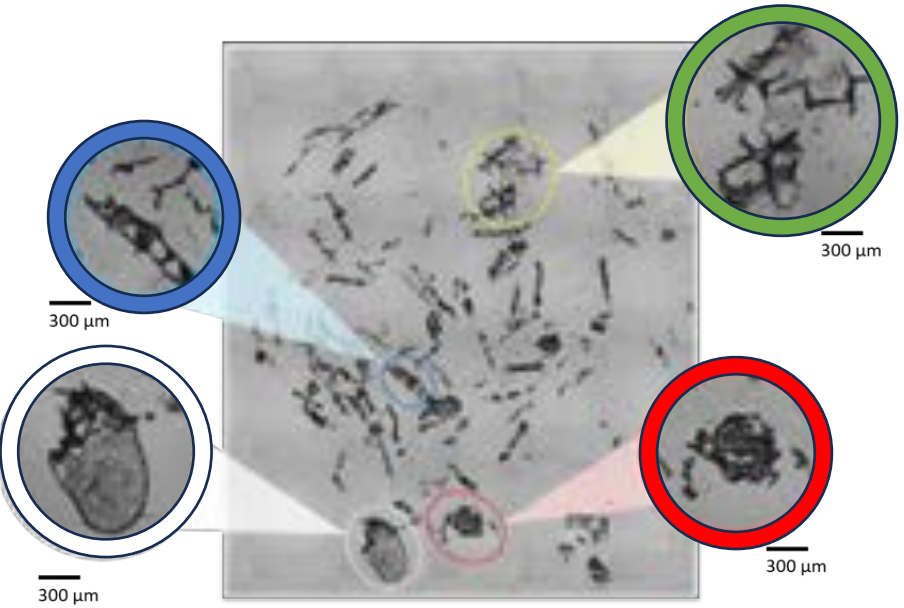
12-µm fungal garden ecosystem section mounted on a glass slide



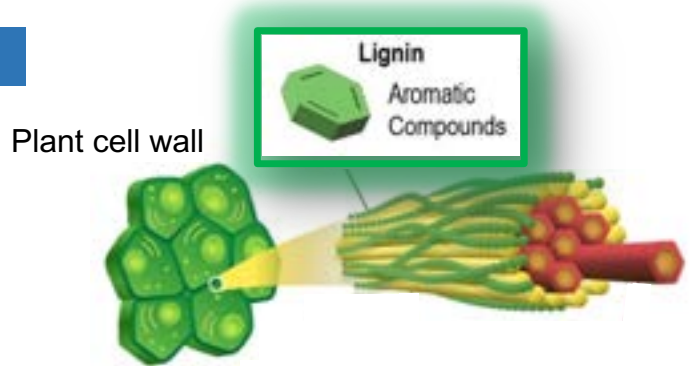
Microscale measurements enable prediction of function

Microscale multi-omic mapping

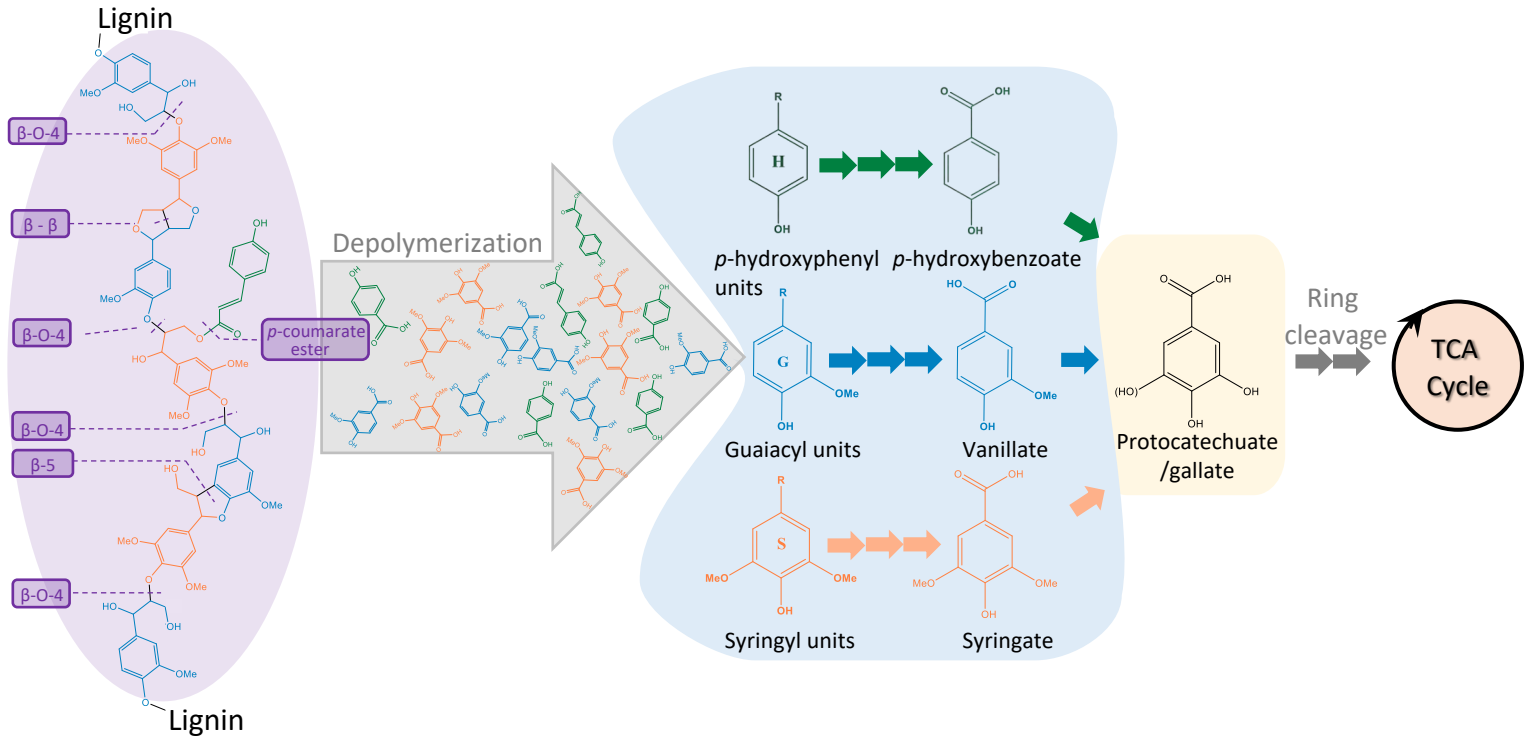
Lignin degradation microscale activity zones **Blue**, **Yellow** and **Red**



Lignin degradation

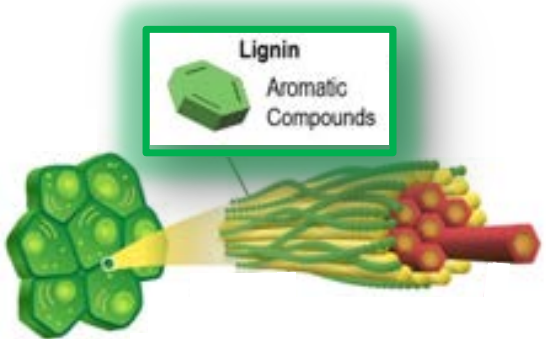


Lignin is a complex organic polymer made up of aromatic compounds in plant cell walls

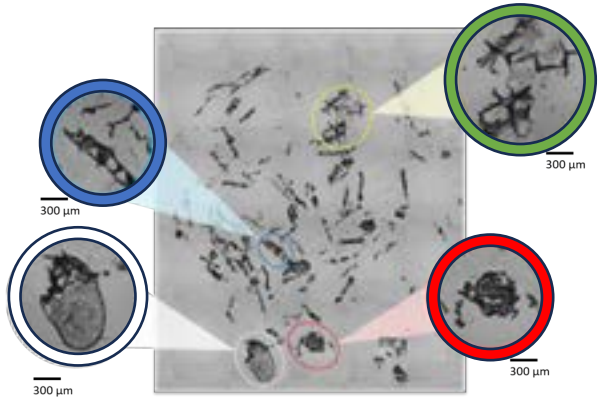


Marija Veličković, et. al., Mapping Microhabitats of Lignocellulose Decomposition by a Microbial Consortium. *In press* Nature Chemical Biology (2023)

Lignin Degradation Pathways

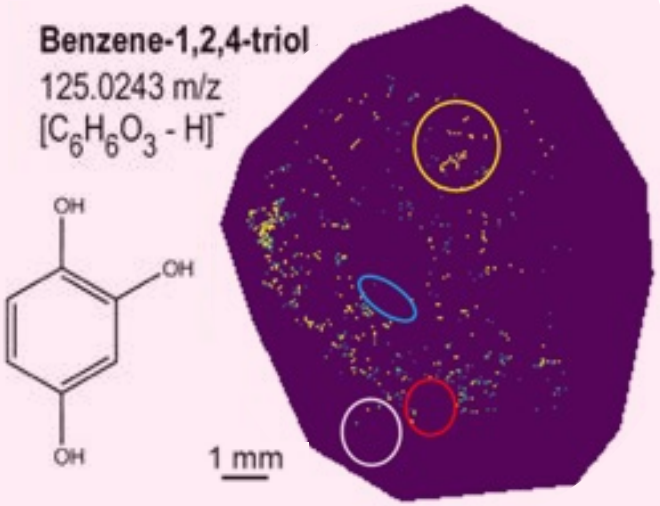


Plant cell wall



Yellow = high metabolite intensity

Substrate



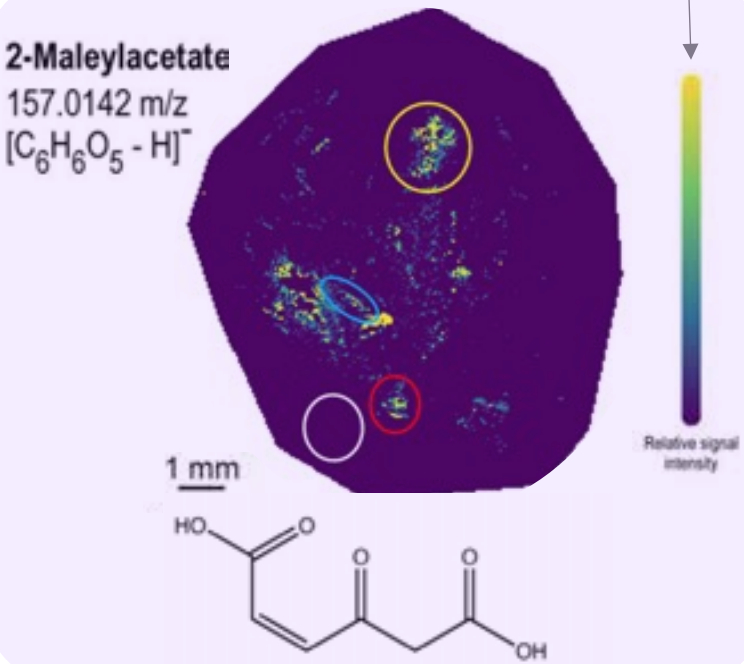
Enzyme

Ring cleavage pathway

Hydroxyquinol 1,2-dioxygenase
[EC 1.13.11.37]
K04098

✅ Detected ; ❌ Not detected

Product



Marija Veličković, et. al., Mapping Microhabitats of Lignocellulose Decomposition by a Microbial Consortium. *In press* Nature Chemical Biology (2023)

Moving biological understanding from phenotype to the phenome

Mapping Phenotypes

Parts List

- DNA
- RNA
- mRNA splice isoform
- Protein expression
- Post Translational Modifications (PTM)
- Metabolome (Metabolites, Lipids, etc.)

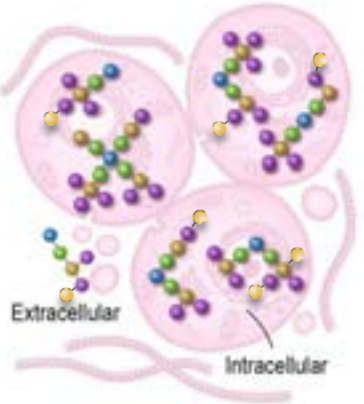
Dynamic measurements

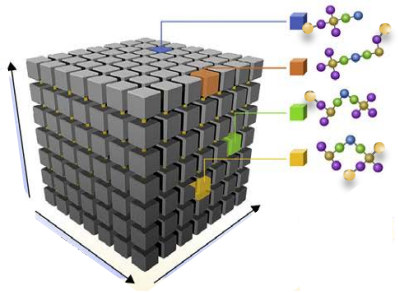
Perturbations, an alteration of the function of a biological system induced by

→ Molecular changes (DNA editing)
→ Environmental changes
→ Temporal changes
→ Spatial changes (across cells, intracellular)

Learn and Predict Molecular Networks

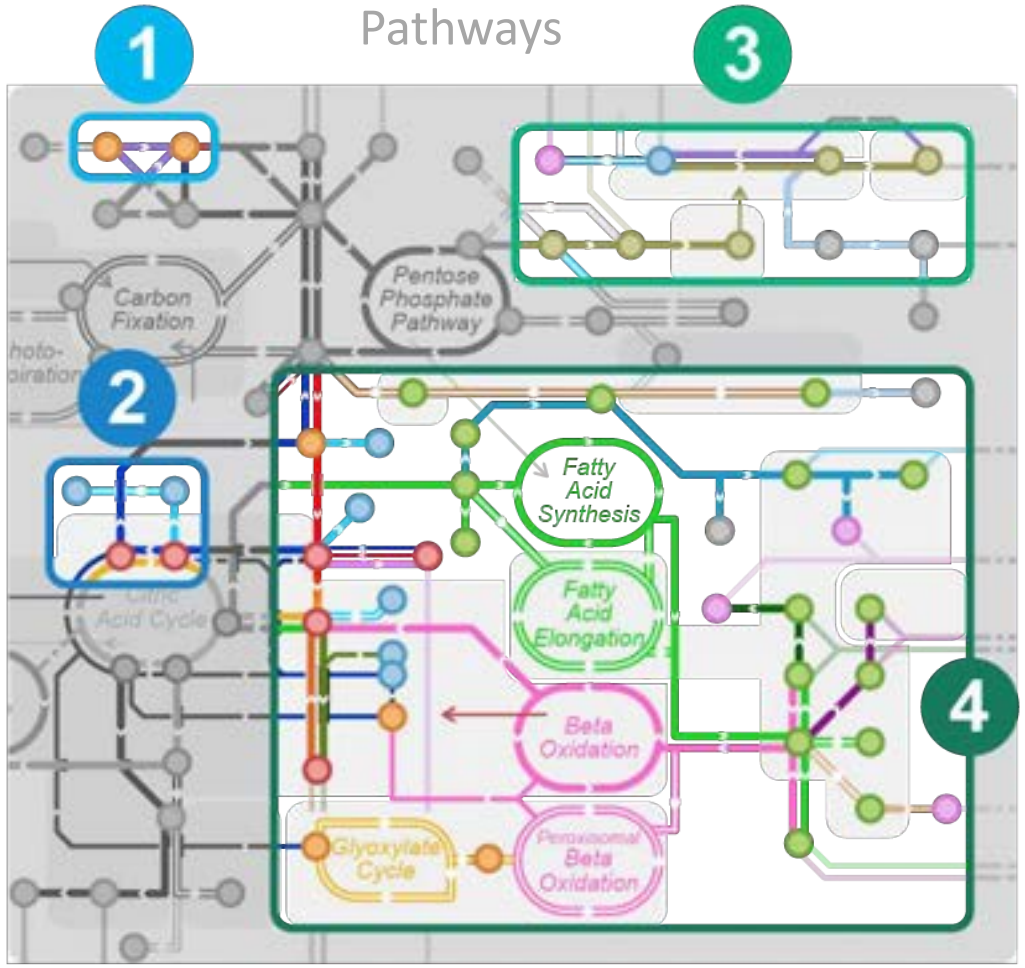
Integrate
Model
Hypothesize



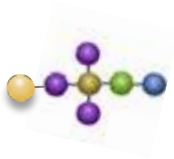


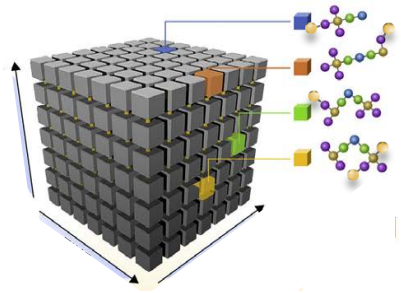
Phenome (n). The set of all phenotypes expressed by a cell, tissue, organ, organism, or species.

Currently Limited in our ability to Phenotype

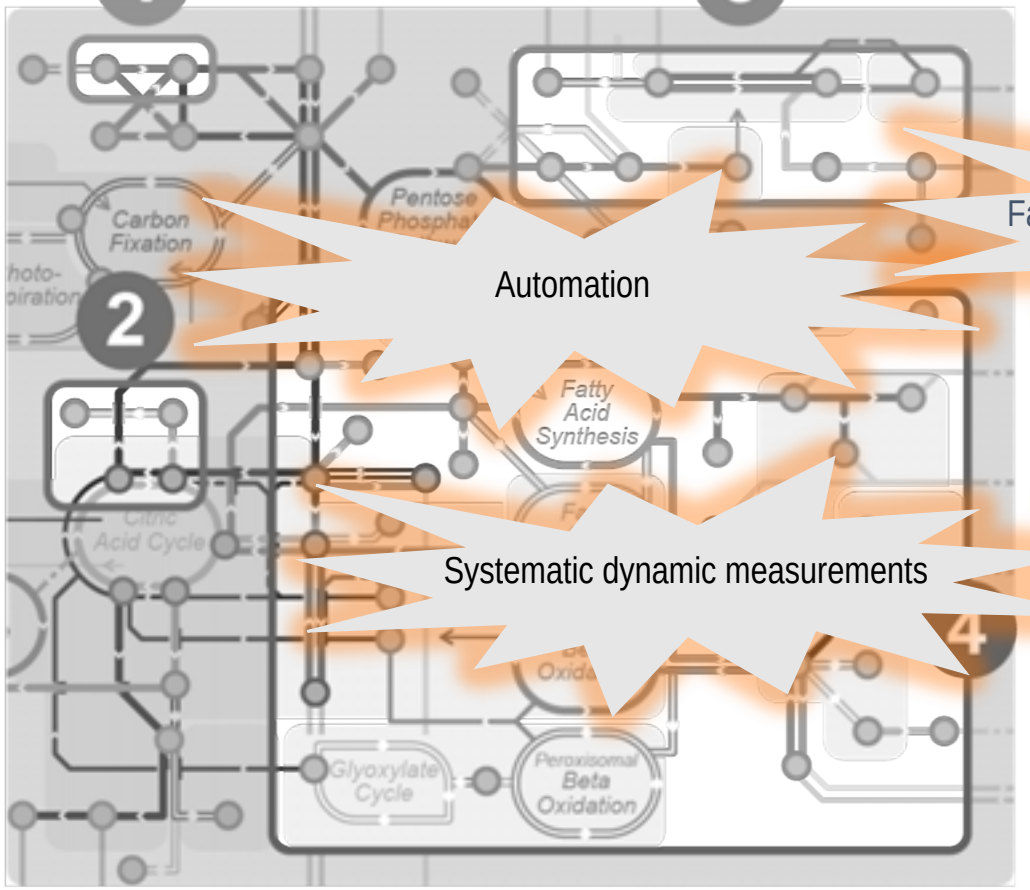


Level of Phenotyping	# Targets	Throughput	Fundamental Sci. Unlocked
Level 1 Singular Network	Small	Small	Low
Level 2 Competing Networks	Small-Moderate	Moderate	Low-Moderate
Level 3 Multiple Interacting Networks	Large	Large	High
Level 4 Dark Phenome	Very Large	Very Large	Very High





Currently Limited in our ability to Phenotype



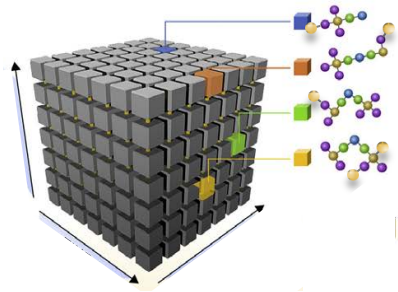
Automation

Systematic dynamic measurements

Fast omics with high depth of coverage

Advanced computation approaches
ML and AI

Level of Phenotyping	# Targets	Throughput	Fundamental Sci. Unlocked
Level 1 Singular Network	Small	Small	Low
Level 2 Competing Networks	Small-Moderate	Moderate	Low-Moderate
Multiple Int. Networks		Large	High
Level 4 Dark Phenome	**Very Large**	**Very Large**	**Very High**



BERKELEY LAB

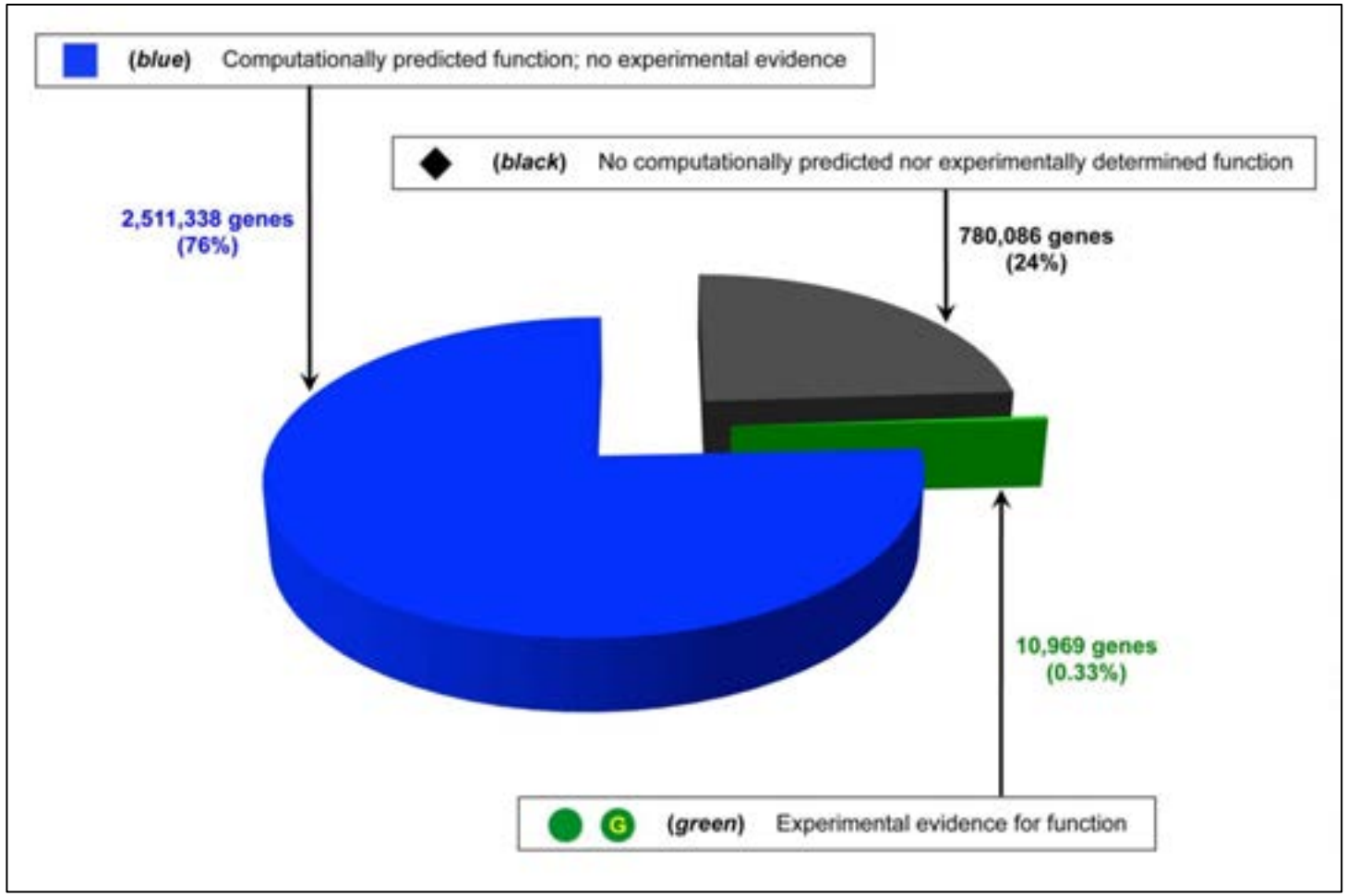
LAWRENCE BERKELEY NATIONAL LABORATORY

U.S. DEPARTMENT OF ENERGY

Characterizing bacterial genes with large-scale genetics

Adam Deutschbauer, LBNL and UC Berkeley
AMDeutschbauer@lbl.gov

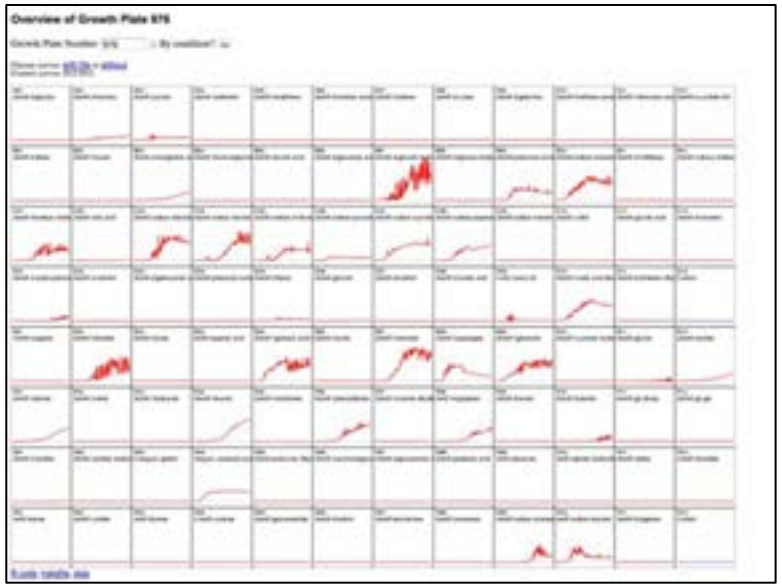
Major Problem: Many genes of unknown function in bacterial genomes



Anton et al. PLoS Biology 2013

Our approach: ”High-throughput” microbiology

- Mostly genetic approaches to infer the function of genes from their phenotypes
- We study many different bacteria
- Miniaturized and multiplexed assays to drive down costs
- Convert different functional assays to a next-generation sequencing readout



Team science

https://enigma.lbl.gov/

https://mcafes.lbl.gov/

https://genomicscience.energy.gov/pnnlbiosystemsdesign/

https://gutworks.stanford.edu/

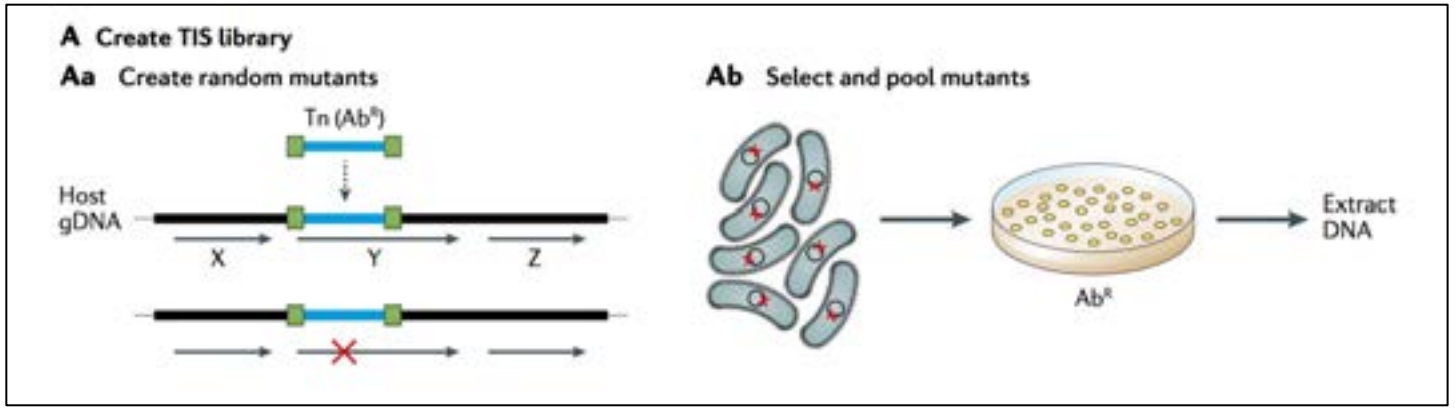
Outline

- RB-TnSeq for characterizing gene function in bacteria
- 6 challenges
- If I funded an effort on gene function discovery in microbes

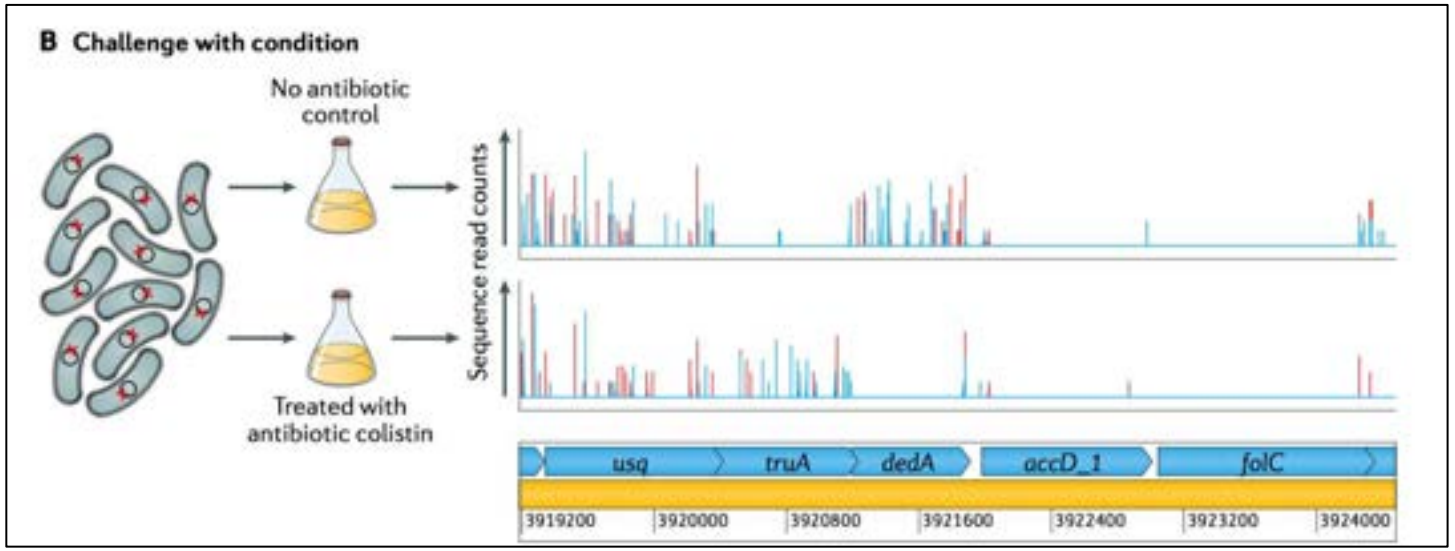
Functional genomics with Tn-seq

A decade of advances in transposon-insertion sequencing

Amy K. Cain, Lars Barquist, Andrew L. Goodman, Ian T. Paulsen, Julian Parkhill and Tim van Opijnen



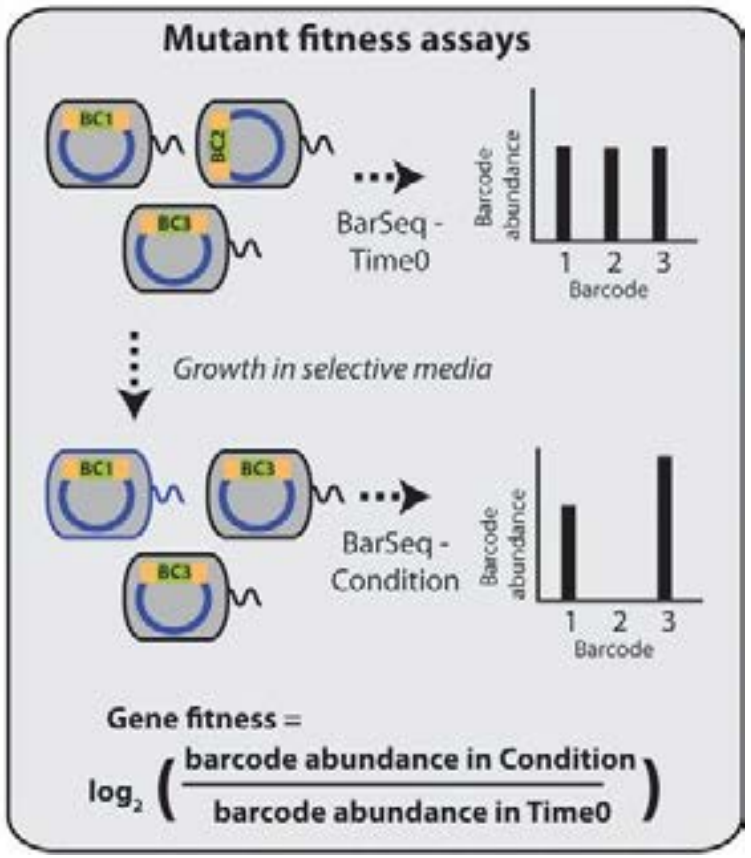
Measure phenotypes of most genes in the genome in parallel.



Random barcode transposon site sequencing (RB-TnSeq)

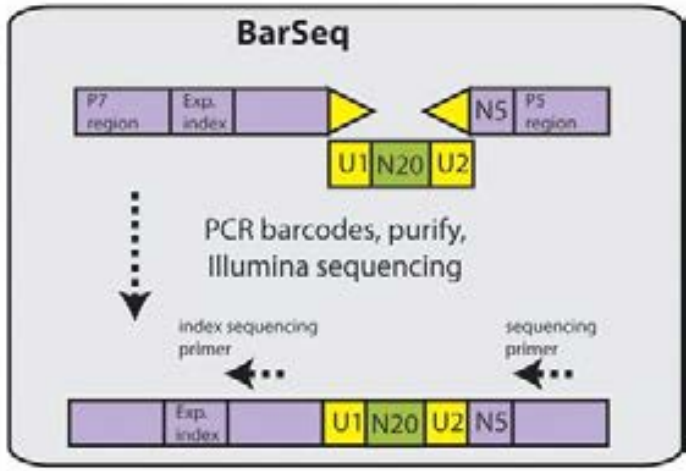
- Incorporate random 20bp DNA tags into the transposons (DNA barcodes)

- Abundance of mutants in the population can be measured by PCR and and sequencing the DNA barcodes (BarSeq)



Rapid Quantification of Mutant Fitness in Diverse Bacteria by Sequencing Randomly Bar-Coded Transposons

Kelly M. Wetmore,[a] Morgan N. Price,[a] Robert J. Waters,[b] Jacob S. Lamson,[a] Jennifer He,[a] Cindi A. Hoover,[b] Matthew J. Blow,[b] James Bristow,[b] Gareth Butland,[c] Adam P. Arkin,[a,d] Adam Deutschbauer[a]



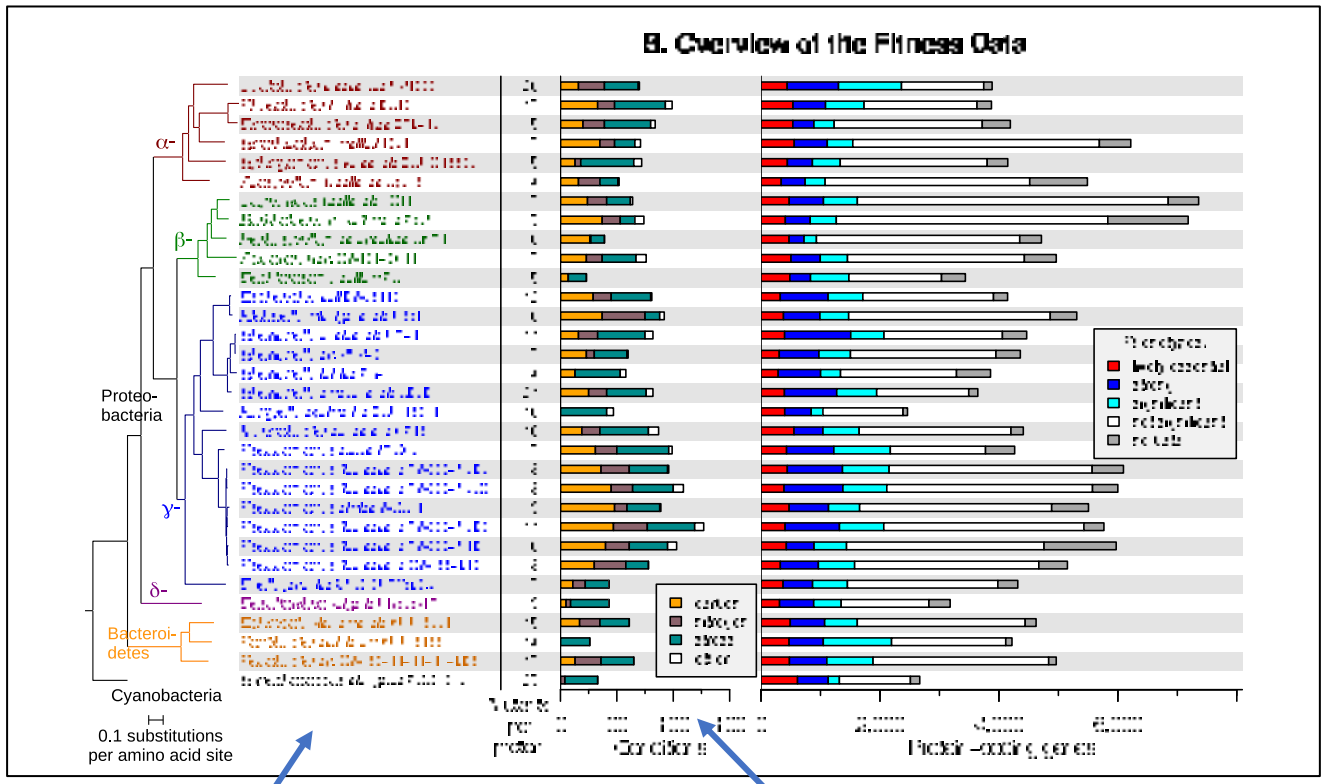
- BarSeq is very easy and scalable. Just mix your amplicons, run over single purification column in 10 minutes, and submit for Illumina sequencing

- We use BarSeq (with same U1 and U2 priming sites for):
  - **RB-TnSeq**
  - **Lineage tracking in evolution studies (Tn7 insertions into neutral location)**
  - **CRISPR interference**
  - **Assessment of genetic systems (magic pools)**
  - **Overexpression studies**
  - **CRISPR-associated transposons**

Genetics data for many bacteria

Mutant phenotypes for thousands of bacterial genes of unknown function

Morgan N. Price[1], Kelly M. Wetmore[1], R. Jordan Waters[2], Mark Callaghan[1], Jayashree Ray[1], Huadan Liu[1], Jennifer V. Kuehl[1], Ryan A. Melnyk[1], Jacob S. Lamson[1], Yumi Suh[1], Hans K. Carlson[1], Zuelma Esquivel[1], Harini Sadeeshkumar[1], Romy Chakraborty[3], Grant M. Zane[4], Benjamin E. Rubin[2], Judy D. Wall[4], Axel Visel[2,5], James Bristow[2], Matthew J. Blow[2]*, Adam P. Arkin[1,3]* & Adam M. Deutschbauer[1,6]*



Expand # of strains

Expand # of conditions

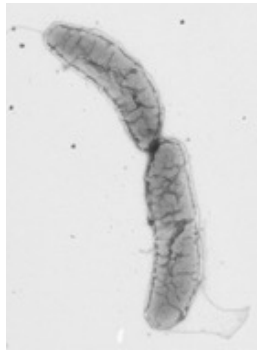
- ~5,000 genome-wide RB-TnSeq assays across 32 bacteria
- Over 20 million gene-phenotype measurements
- Phenotypes for over 10,000 genes without a known function (many are conserved across different bacteria)
- Identify specific functions for hundreds of mis-annotated enzymes and transporters

On a single Illumina X 10B flow cell, we can sequence 3,072 RB-TnSeq (BarSeq) samples (~$3.50 per sample).

Challenge #1: Getting genetics up and running in diverse bacteria

- Gram-positive bacteria are generally more challenging than Gram-negative
- But genetics with current tools often fails for Gram-negative bacteria as well
- Possible solutions: Testing libraries of genetic systems against a target microbe in parallel, overcoming host defense systems, improved DNA delivery methods

Test thousands of different vectors in parallel (magic pool)

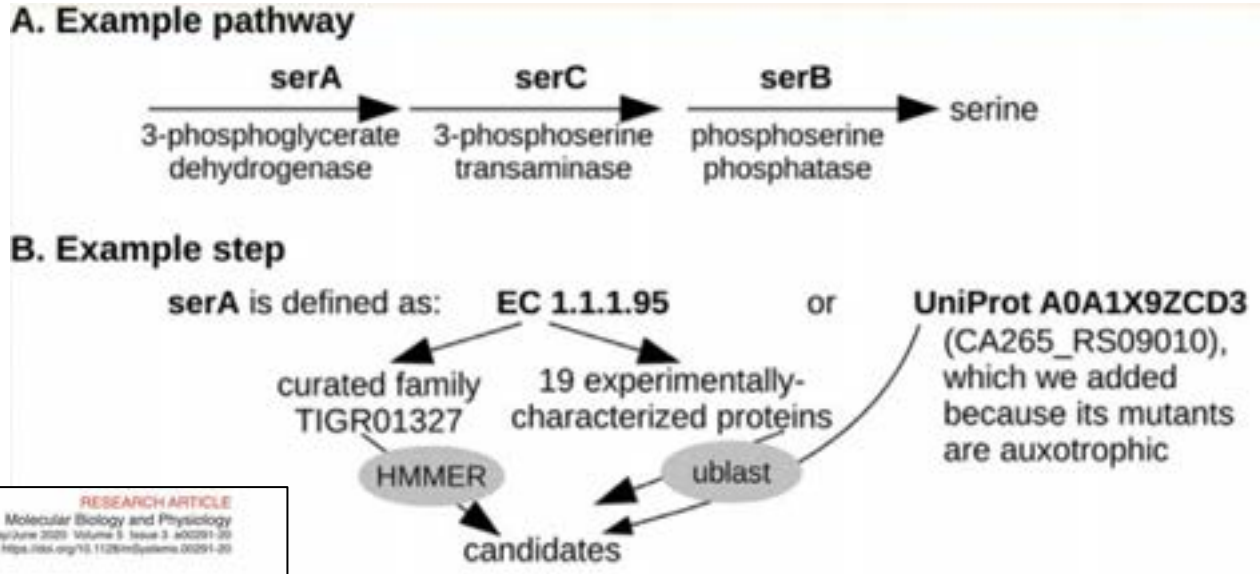


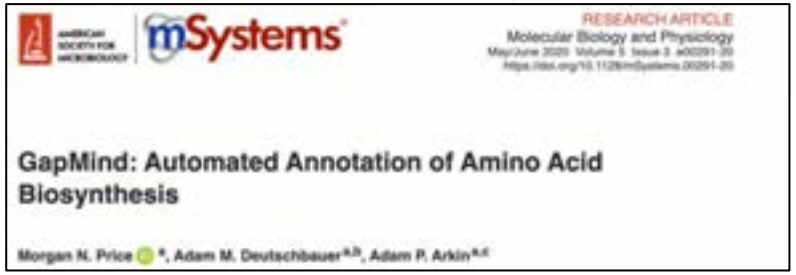
Take advantage of:
- DNA synthesis
- Parts-based cloning
- Long-read DNA sequencing (PacBio and Oxford Nanopore)

Liu et al. *mSystems* 2018

Challenge #2: Propagating inferred gene functions to new genes/genomes

- It's not straightforward getting genetics-based gene annotations into established databases (like UniProt)
- Propagation of updated gene annotations to new genomes also isn't straightforward
- Possible solutions: GapMind, better communication/integration between stakeholders





Challenge #3: Pooled mutant fitness assays aren't ideal for non-growth based phenotypes

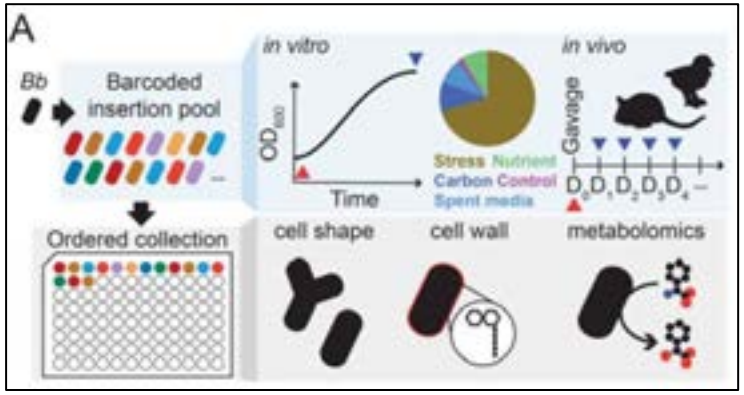
- Pooled fitness assays (like RB-TnSeq) are great for growth-based assays, like nutrient conditions (C, N, S, P sources), stress conditions, etc.
- They're not good for secondary metabolite discovery, secreted factors.
- Most genes do not have a strong phenotype under laboratory conditions

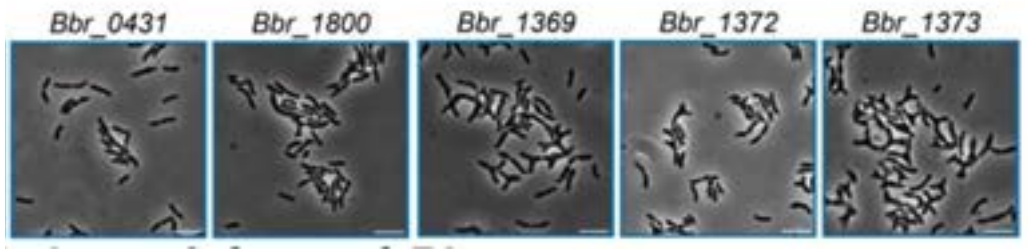
- Possible solutions: Assays using archived collections of individual mutants, new method development to more systematically characterize gene function for other "categories" of genes (like second metabolites)

A mutant fitness compendium in Bifidobacteria reveals molecular determinants of colonization and host-microbe interactions

Anthony L. Shiver, Jiawei Sun, Rebecca Culver, Arvie Violette, Charles Wynter, Marta Nieckarz, Samara Paula Mattiello, Prabhjot Kaur Sekhon, Lisa Friess, Hans K. Carlson, Daniel Wong, Steven Higginbottom, Meredith Weglarz, Weigao Wang, Benjamin D. Knapp, Emma Guiberson, Juan Sanchez, Po-Hsun Huang, Paulo A. Garcia, Cullen R. Buie, Benjamin Good, Brian DeFelice, Felipe Cava, Joy Scaria, Justin Sonnenburg, Douwe Van Sinderen, Adam M. Deutschbauer, Kerwyn Casey Huang

doi: https://doi.org/10.1101/2023.08.29.555234



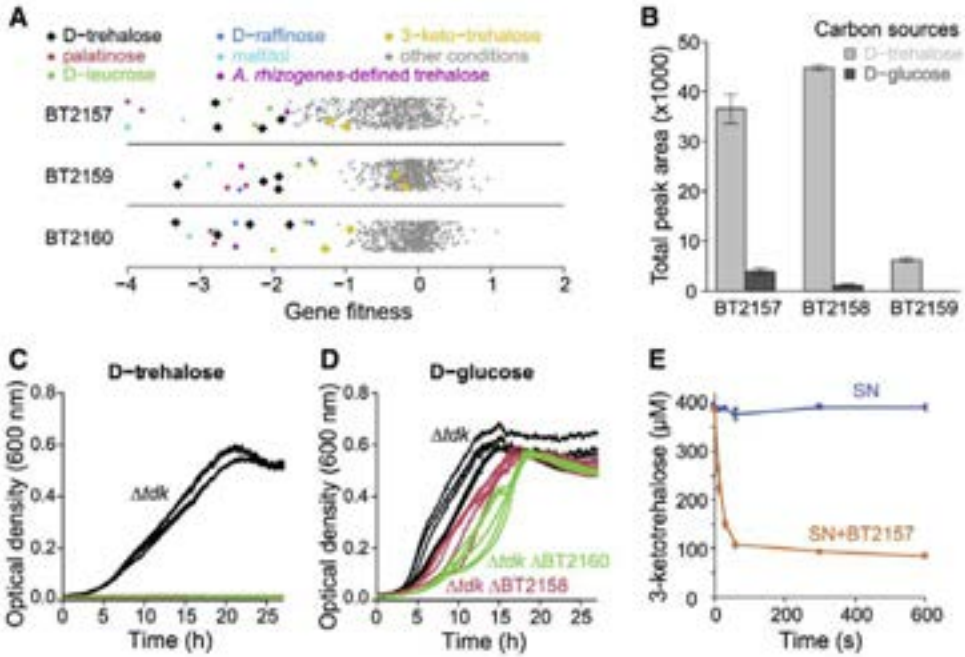


Challenge #4: Availability/cost of compounds for chemical genomic screening

- Compounds of interest are often quite expensive, or not commercially available
- Possible solutions: Spend a lot of money, partner with chemists

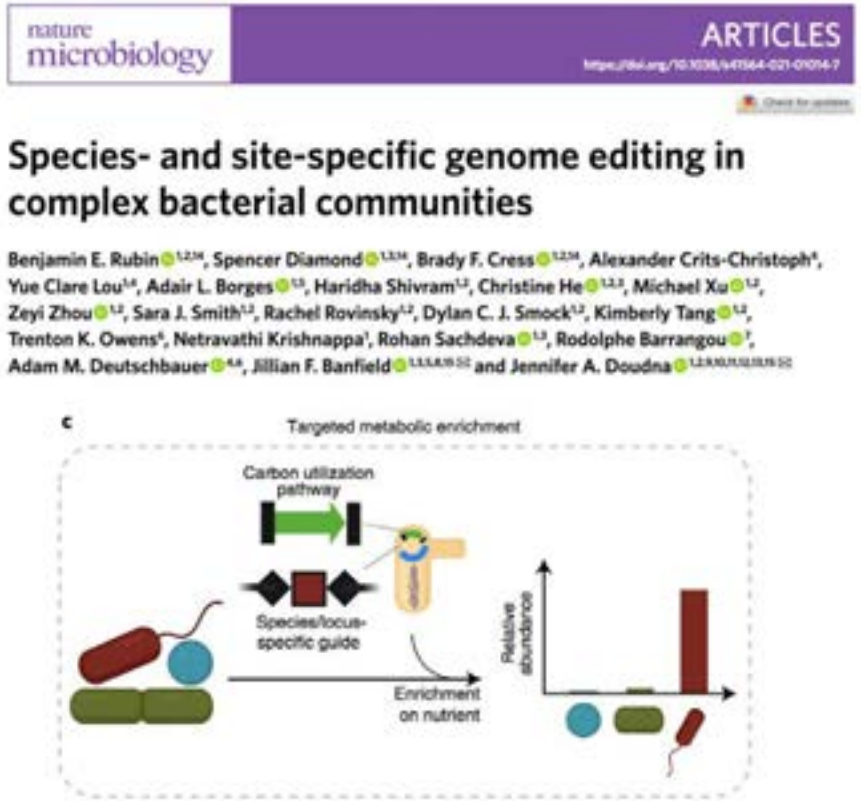
Functional genetics of human gut commensal *Bacteroides thetaiotaomicron* reveals metabolic requirements for growth across environments

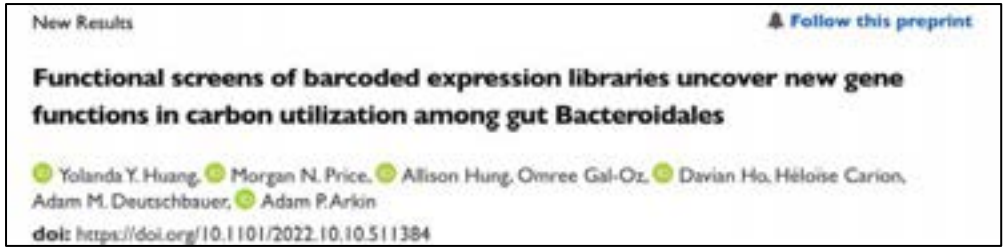
Hualan Liu, Anthony L. Shiver, Morgan N. Price, Hans K. Carlson, Valentine V. Trotter, Yan Chen, Veronica Escalante, Jayashree Ray, Kelsey E. Hern, Christopher J. Petzold, Peter J. Turnbaugh, Kerwyn Casey Huang, Adam P. Arkin, and Adam M. Deutschbauer

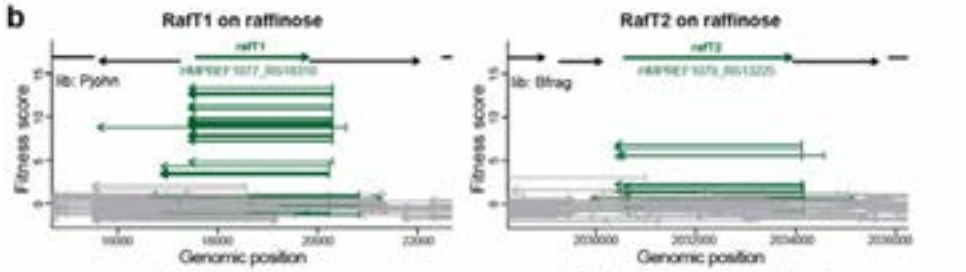


Challenge #5: Large-scale functional genomics typically requires isolates

- Many bacteria are currently uncultivated, so we're currently not assaying a significant fraction of the gene space
- Possible solutions: Get more microbes into cultivation, Microbial community editing, heterologous expression of DNA (random and via DNA synthesis) in diverse hosts







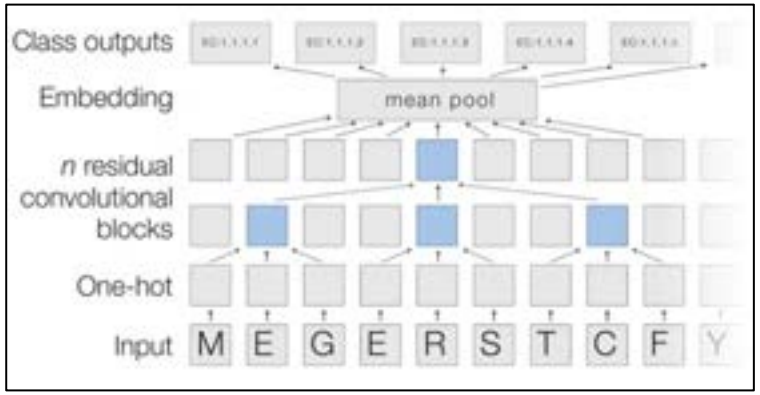
Challenge #6: Manual Inference of gene function from mutant phenotypes

- It's still laborious to manually examine data to make new discoveries
- Possible solutions: GapMind-like tools to quickly identity the "unknowns" in metabolism, AI/machine learning

ProteInfer, deep neural networks for protein functional inference

Theo Sanderson[1*†], Maxwell L Bileschi[2†], David Belanger[2], Lucy J Colwell[2,3*]

[1]The Francis Crick Institute, London, United Kingdom; [2]Google AI, Boston, United States; [3]University of Cambridge, Cambridge, United Kingdom

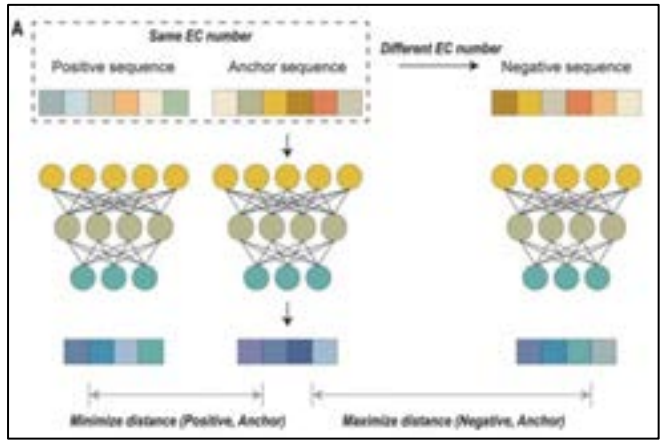


- Neural network-based approach
- SwissProt database used for training model

FUNCTION PREDICTION

Enzyme function prediction using contrastive learning

Tianhao Yu[1,2,3†], Haiyang Cui[1,2,3†], Jianan Canal Li[3,4], Yunan Luo[5], Guangde Jiang[1,2], Huimin Zhao[1,2,3,6*]



- Contrastive learning-based approach
- SwissProt database used for training model

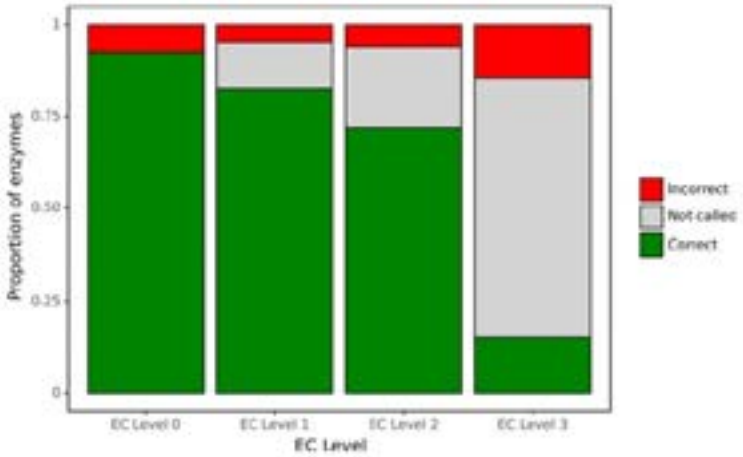
For testing performance, both studies used >100 bacterial enzymes that we annotated using RB-TnSeq data

ML methods work OK, but there's room for improvement

ProteInfer, deep neural networks for protein functional inference

Theo Sanderson[1*†], Maxwell L Bileschi[2†], David Belanger[2], Lucy J Colwell[2,3*]

[1]The Francis Crick Institute, London, United Kingdom; [2]Google AI, Boston, United States; [3]University of Cambridge, Cambridge, United Kingdom

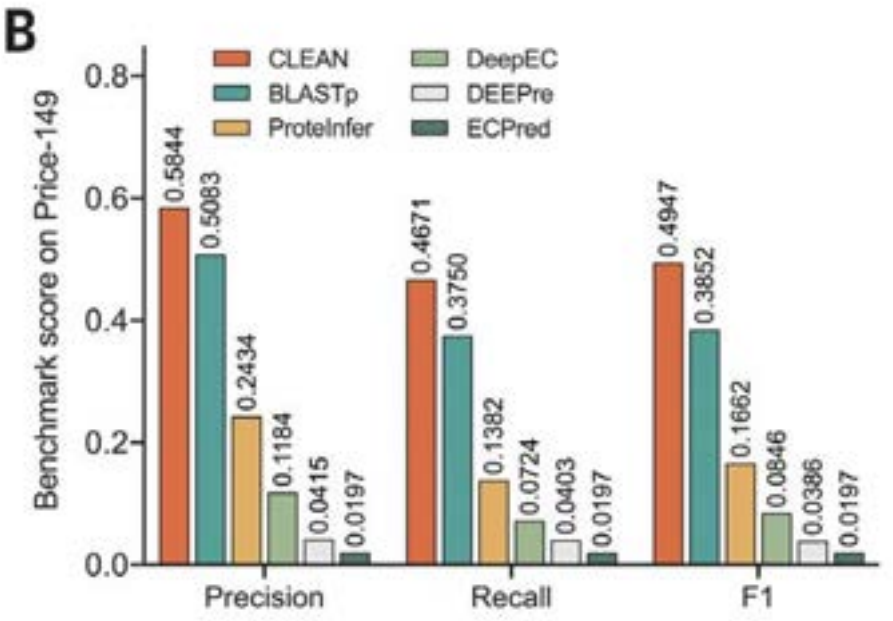


- Accuracy of predictions drops at finer levels of classification
- Network fails to make predictions at higher resolution classifications

FUNCTION PREDICTION

Enzyme function prediction using contrastive learning

Tianhao Yu[1,2,3†], Haiyang Cui[1,2,3†], Jianan Canal Li[3,4], Yunan Luo[5], Guangde Jiang[1,2], Huimin Zhao[1,2,3,6*]



- Performance is only marginally better than BLASTp

If I were funding a large effort to characterize bacterial genes….

- I'd fund a network of researchers to focus on bacterial gene function discovery:
  - Core teams with proven technology (tn-seq, rna-seq, small RNAs, (exo)metabolomics, proteomics, etc.) would apply their methods at scale to thousands of diverse bacteria (would engage the community for their favorite microbes and experimental conditions, and provide all genetic resources and data free of cost and prior to publication)
  - Additional funds would go to high-risk, high-reward technology development projects (Charge could be: "Scale a technology that is informative about gene function in bacteria, such that it could be applied to 1000+ bacteria in 2 years"; perhaps protein-protein interactions, gene regulation, genetic epistasis, structure-function studies, secondary metabolites). The successful tech would be blended into the larger core program.
  - And I wouldn't spend much time mining existing data from literature (like old gene expression data with microarrays), I'd just generate new data at a massive scale linked to accurate metadata, to ensure that it's "machine readable" for the community

COLUMBIA SYSTEMS BIOLOGY

College of Physicians & Surgeons
COLUMBIA UNIVERSITY
Department of Pathology and Cell Biology

wanglab.c2b2.columbia.edu

High-throughput Culturomics & Transcriptomics to Identify The Microbial Dark Matter

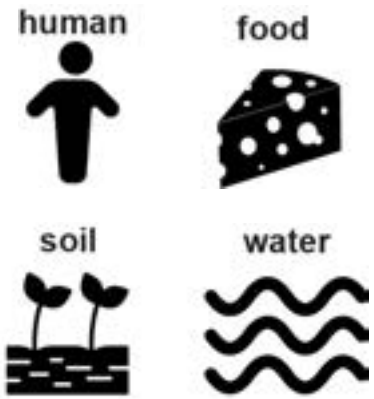
Harris H. Wang, Ph.D.

DARPA DUF Workshop

December 12, 2023

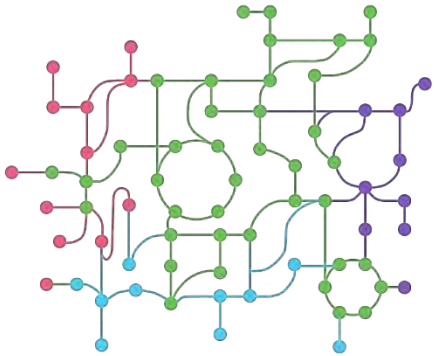
Courtesy Y. Huang

Organism domestication is needed to study function at a mechanistic level

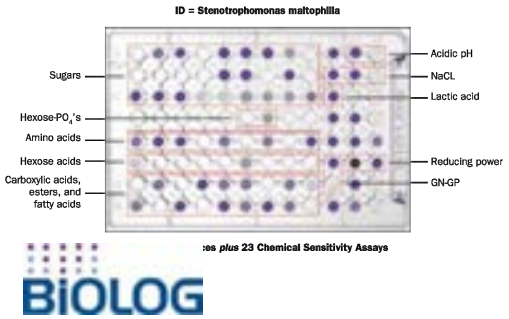


Organisms

Genomics

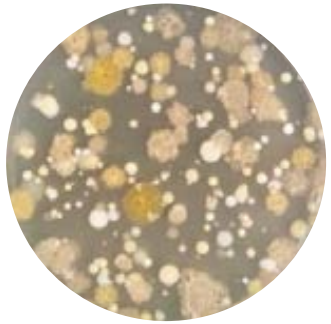


Models



Phenotype

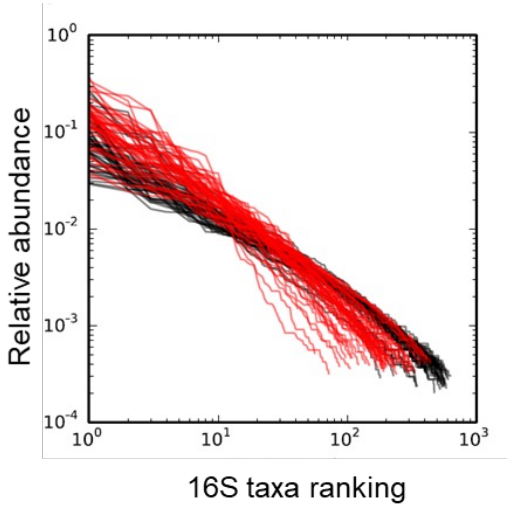
Need strains to do actual experiments!



Culturomics

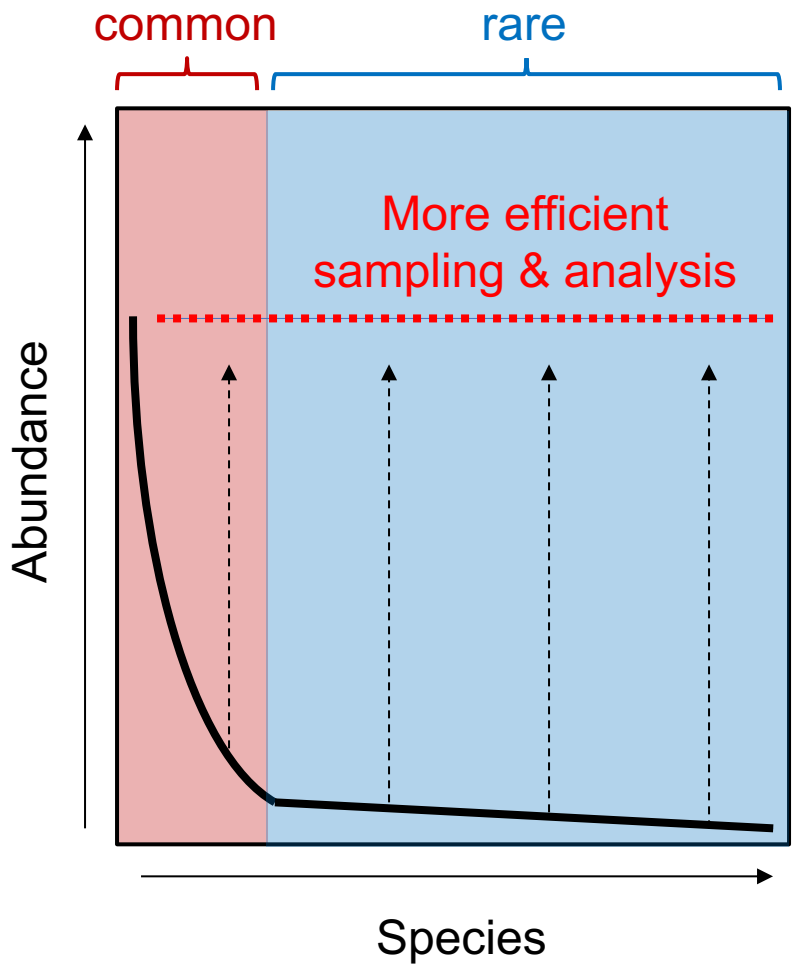
- Systematic: record all info
- Comprehensive: get all strains (hard, but not impossible)
- Cheap: minimizing labor costs/fatigue
- Fast/on-demand: allow iterations

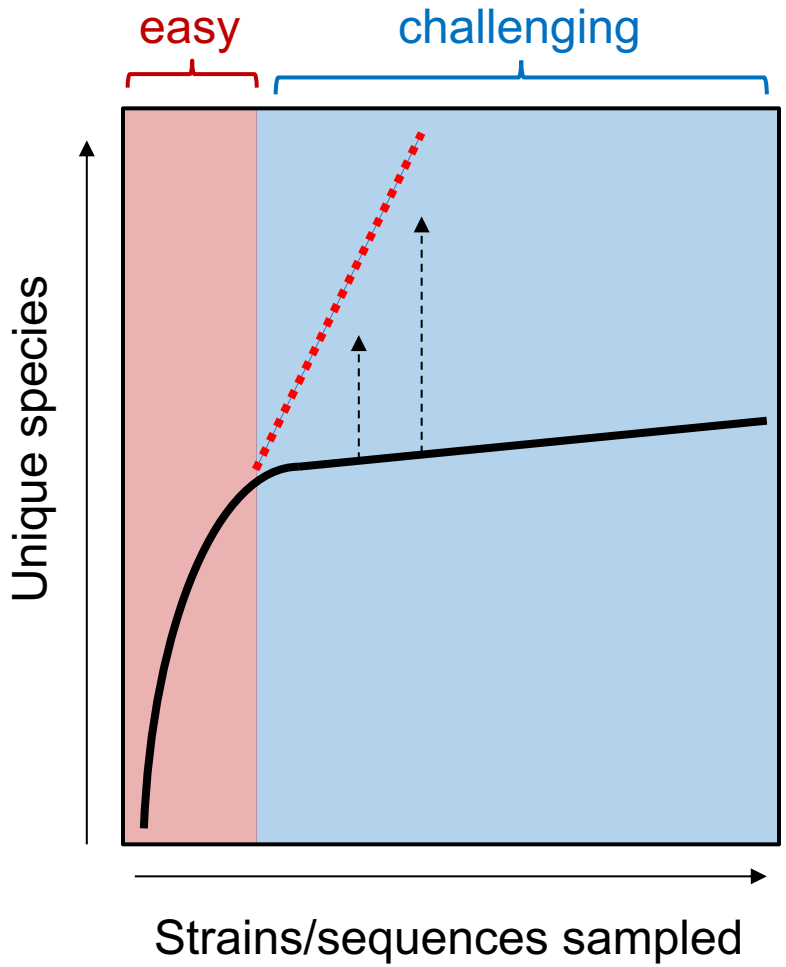
Strain de-duplication through cultivation help fight against the “tragedy of the common” in microbiome research



A universal problem in

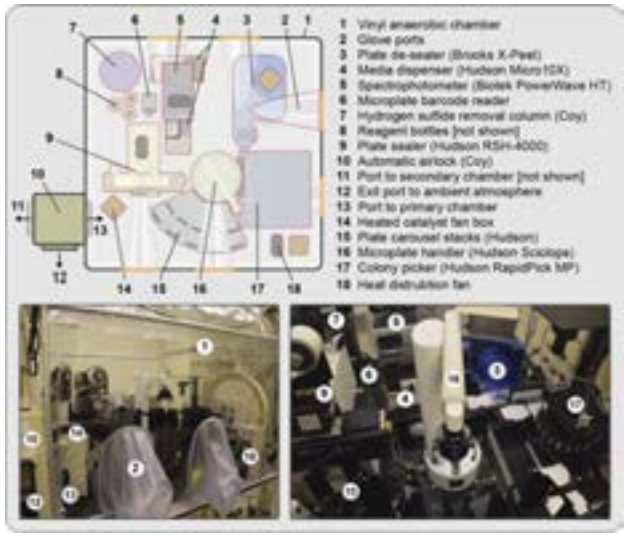
- Metagenomics
- Metatranscriptomics
- Community metabolomics





Culturomics by Automated Microbiome Imaging and Isolation (CAMII) System

Huang et al, *Nature Biotechnology* 41:1424-33 (2023)

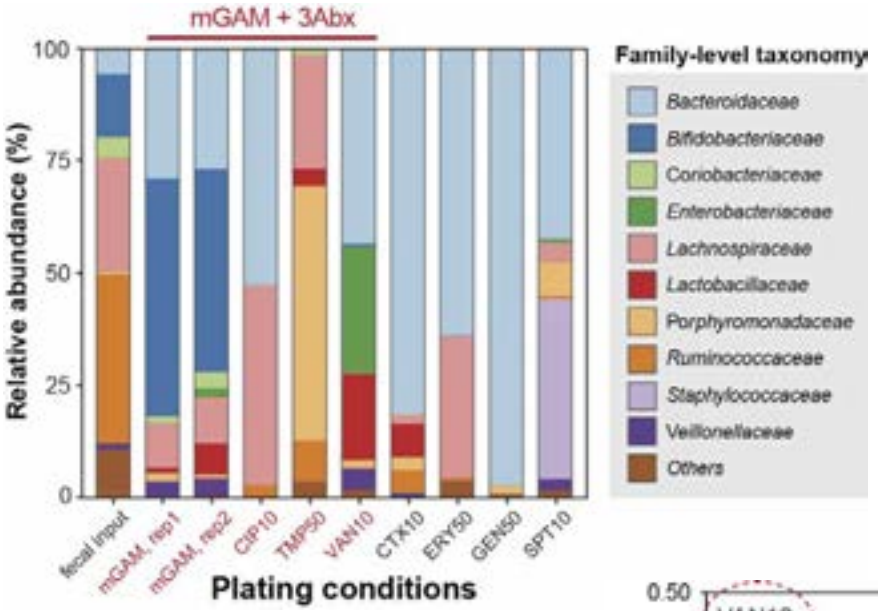


One of the most advanced environmentally controlled microbial cultivation system

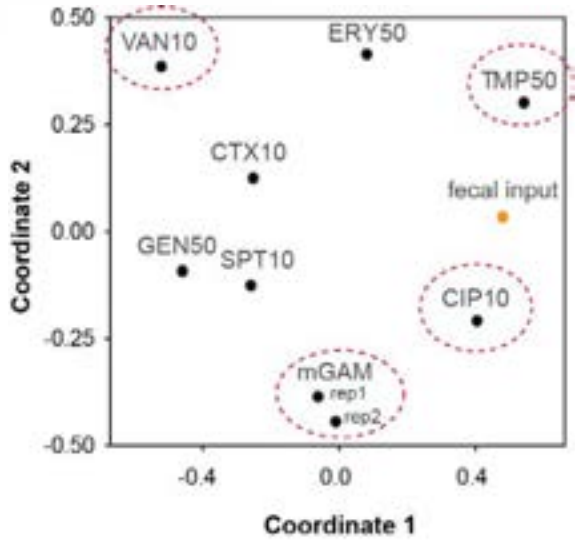
Extensively explored different media formulations and growth conditions

fecal samples

20x plates/sample

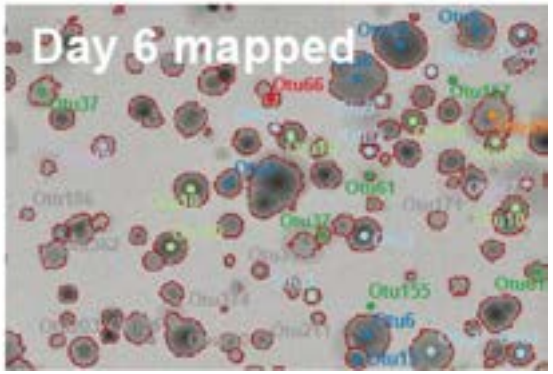


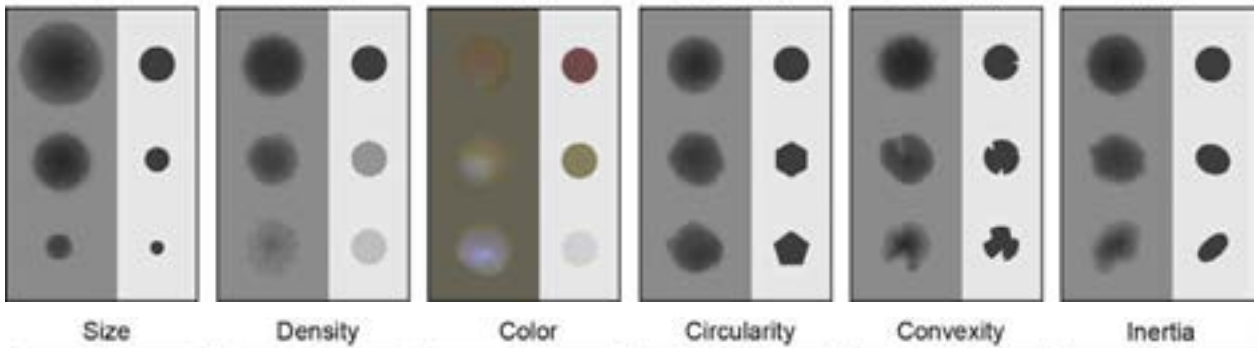
100+ growth conditions:
media, dietary, abx, rumen,
vitamins, menaquinones, etc.

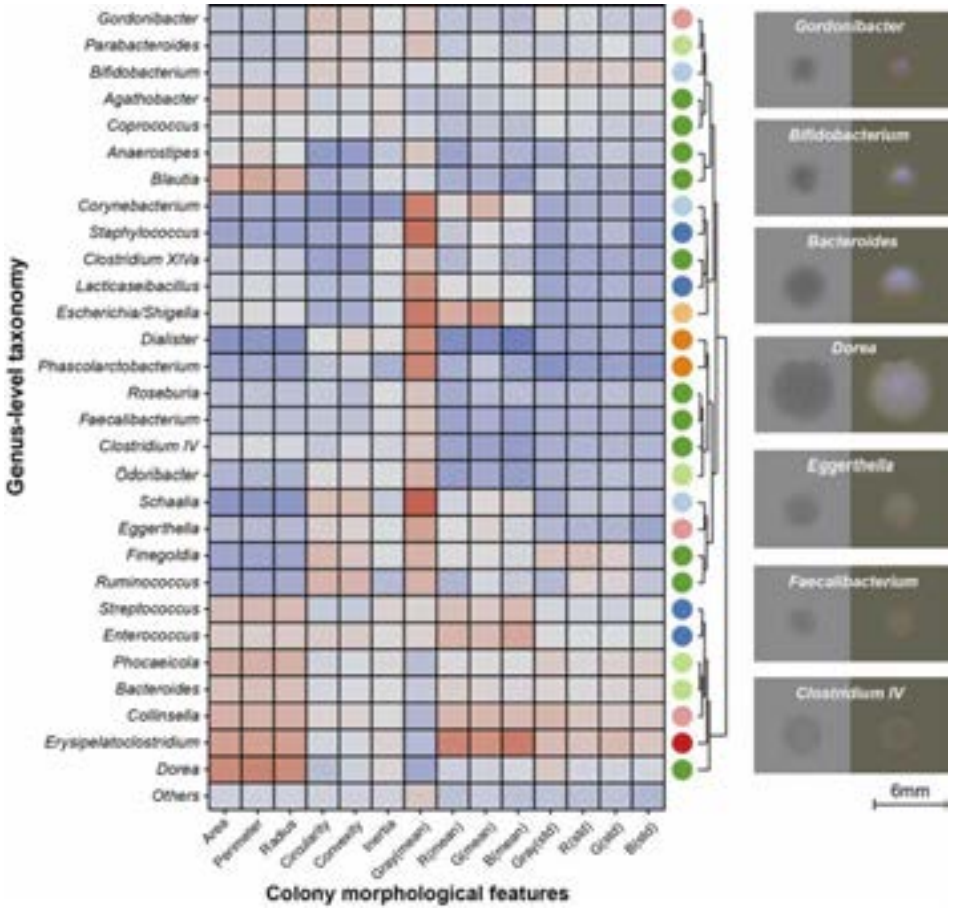


Using AI to predict microbial taxonomy directly from colonies

Colony detection & segmentation



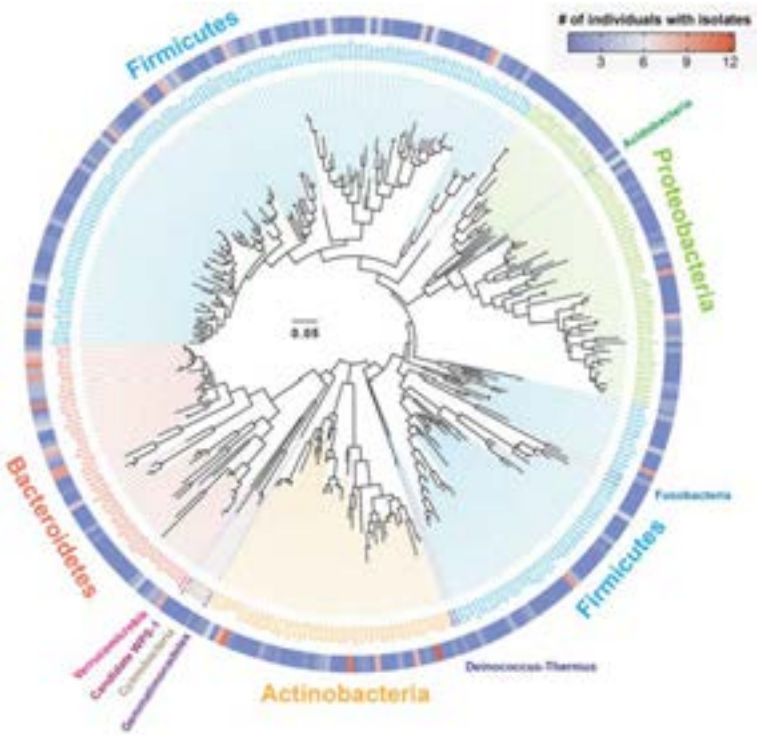




Huang et al, *Nature Biotechnology* doi: 10.1038/s41587-023-01674-2 (2023)

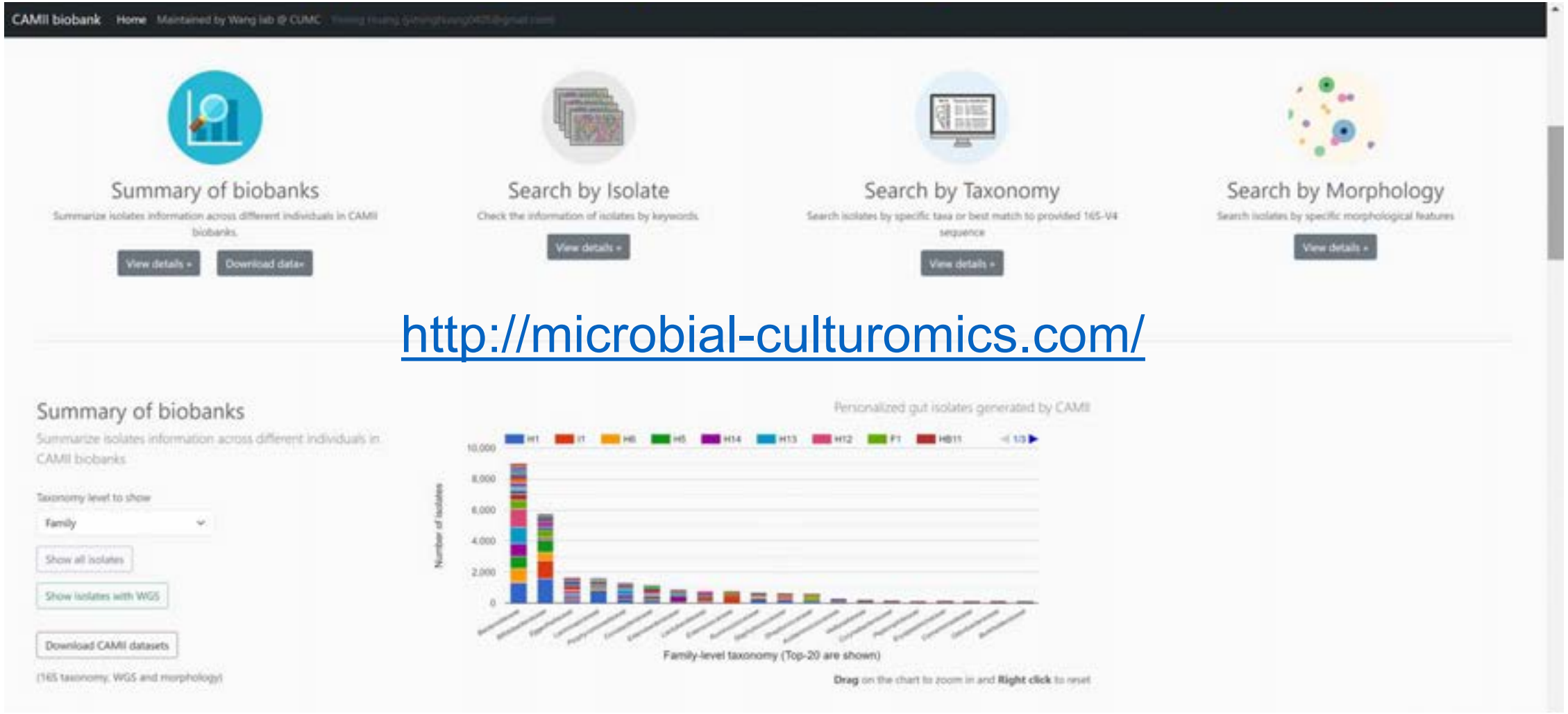
Building the largest microbiome biobanks from unique sources

>32,000 strains in biobank to date



microbial-culturomics.com

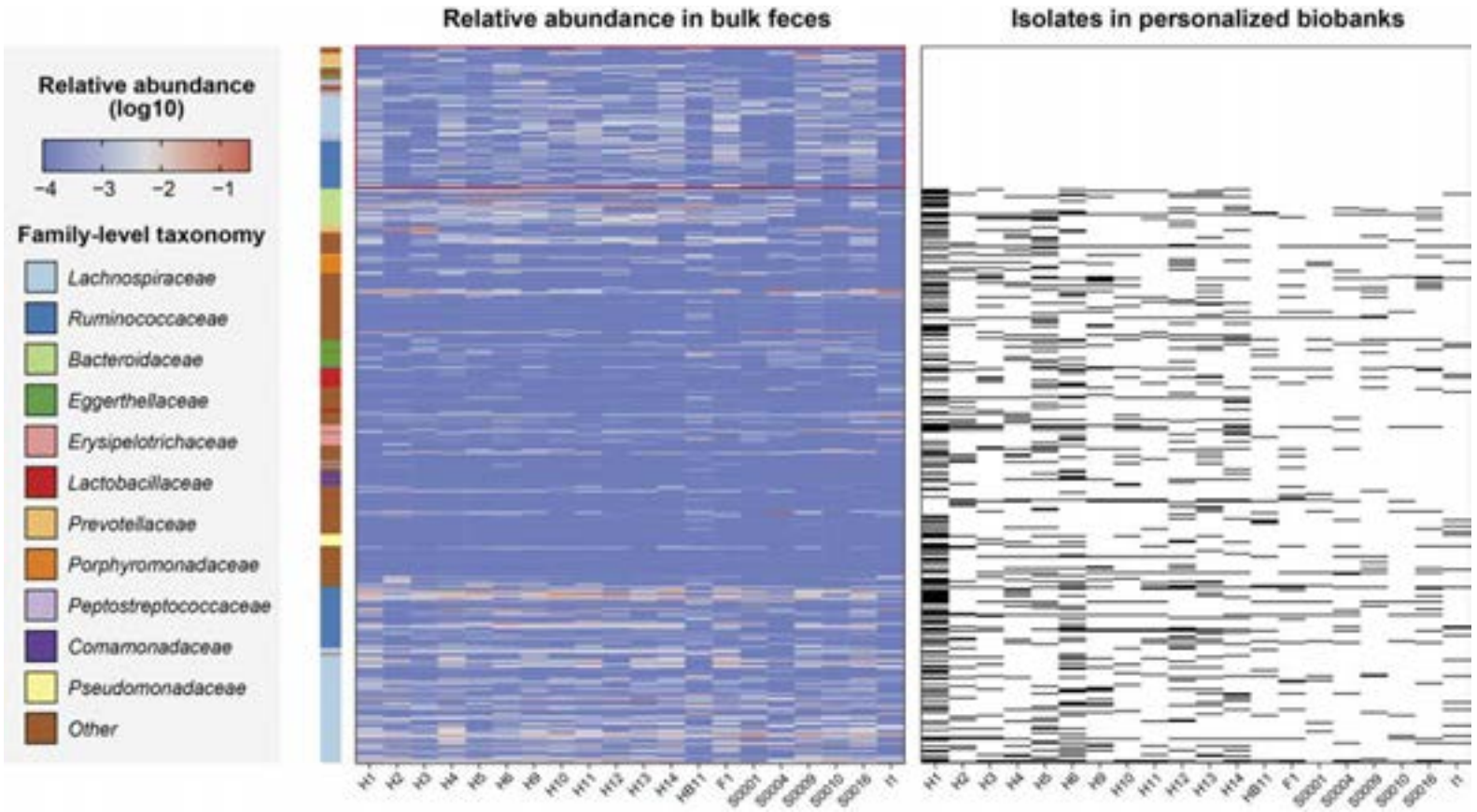
A searchable and open database to share data & biobank



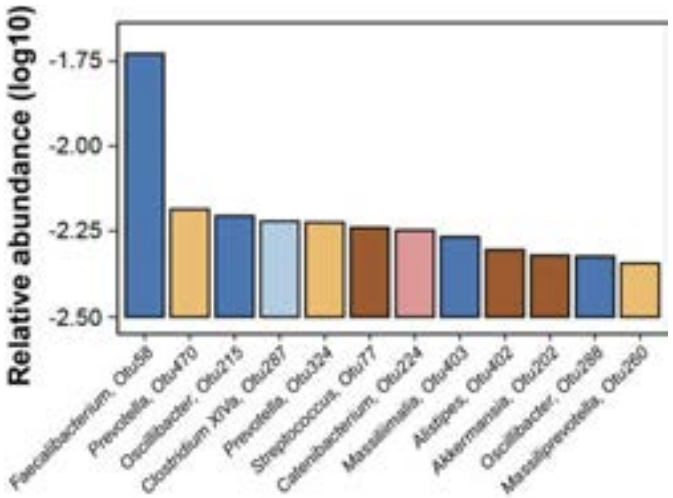
http://microbial-culturomics.com/

Huang et al, *Nature Biotechnology* doi: 10.1038/s41587-023-01674-2 (2023)

Towards illuminating the dark matter of the gut microbiome through systematic culturomics

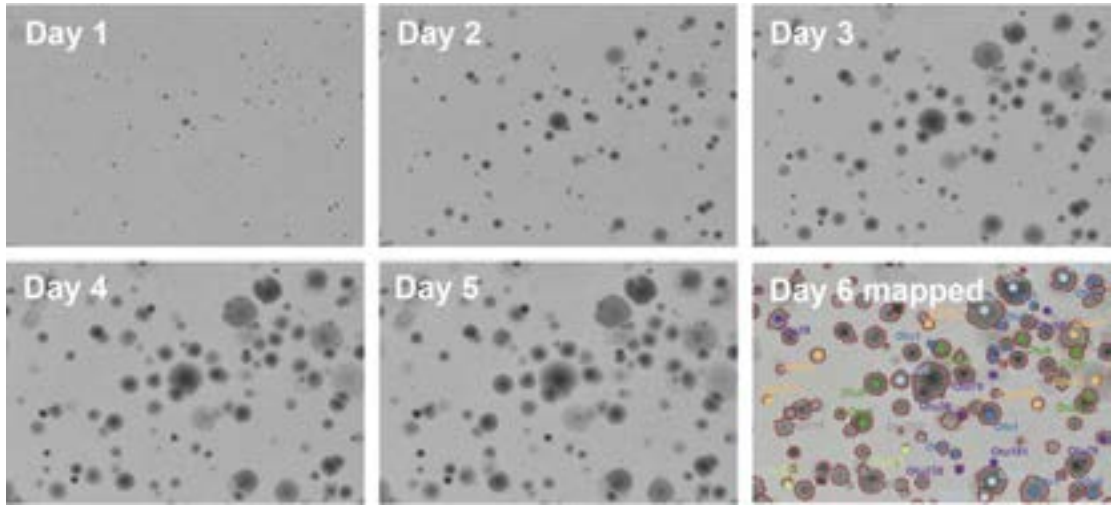


Laura Marshall/NPG

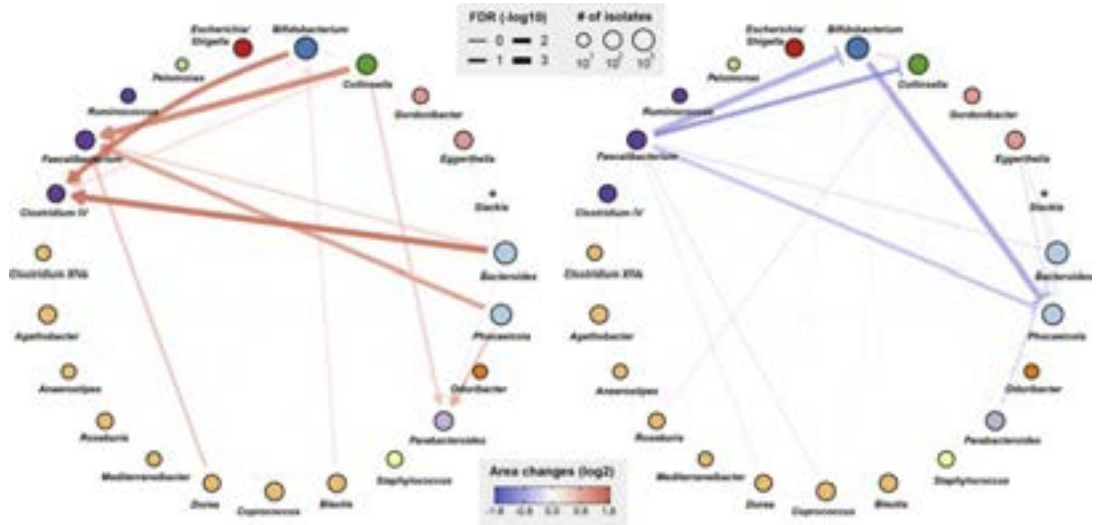


Huang et al, *Nature Biotechnology* doi: 10.1038/s41587-023-01674-2 (2023)

Spatial growth patterns of bacteria on plates provide rich data to delineate species interactions

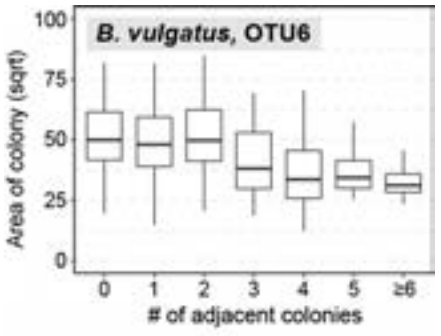


Species interactions

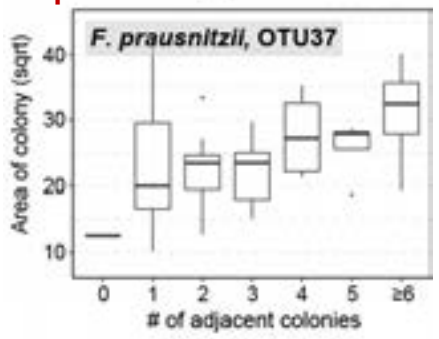


B. adolescentis *C. qucibialis*

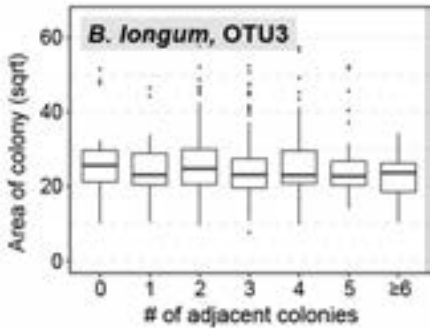
negative interaction

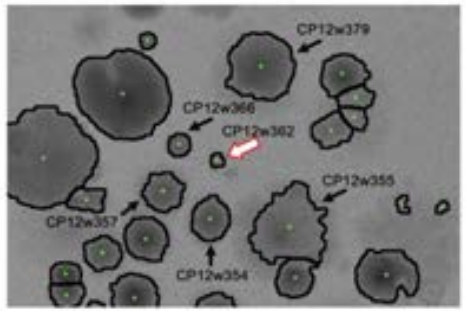


positive interaction



no interaction





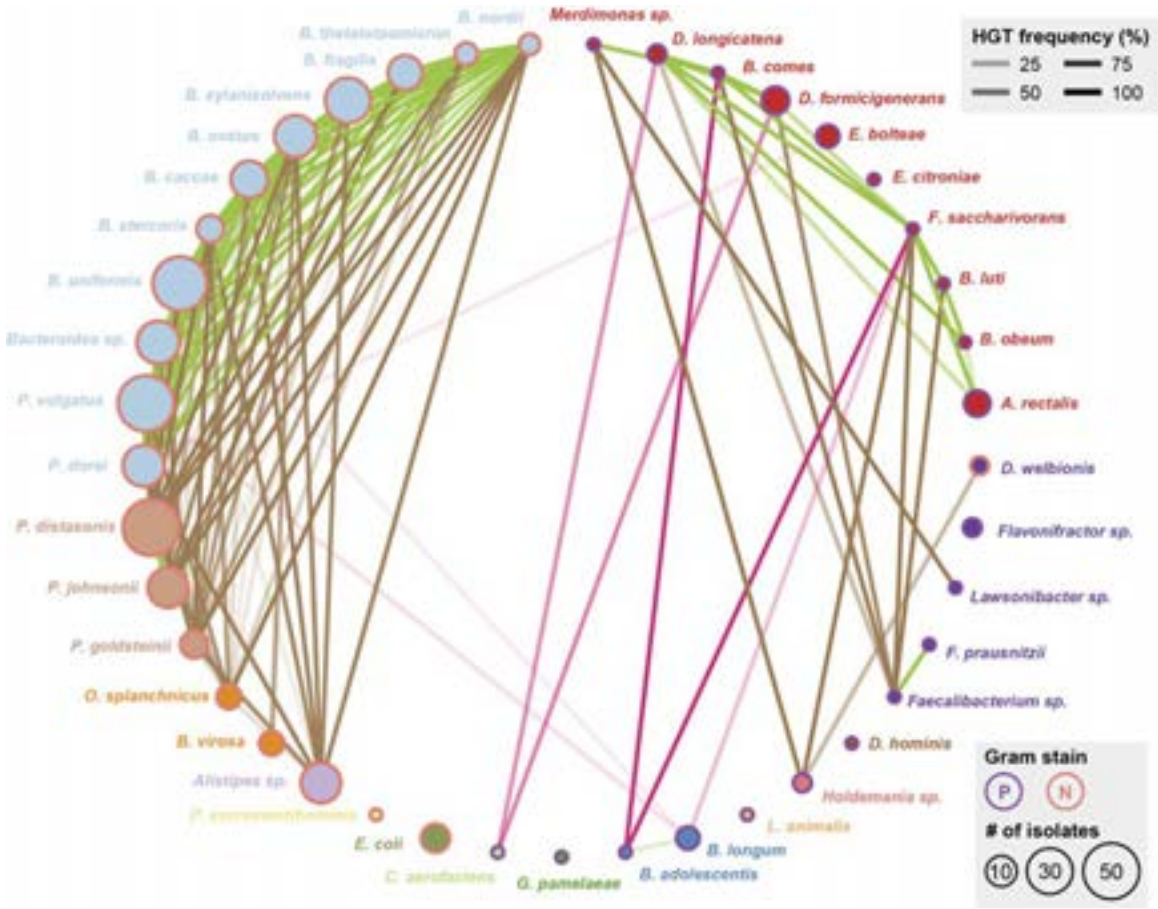
improved growth

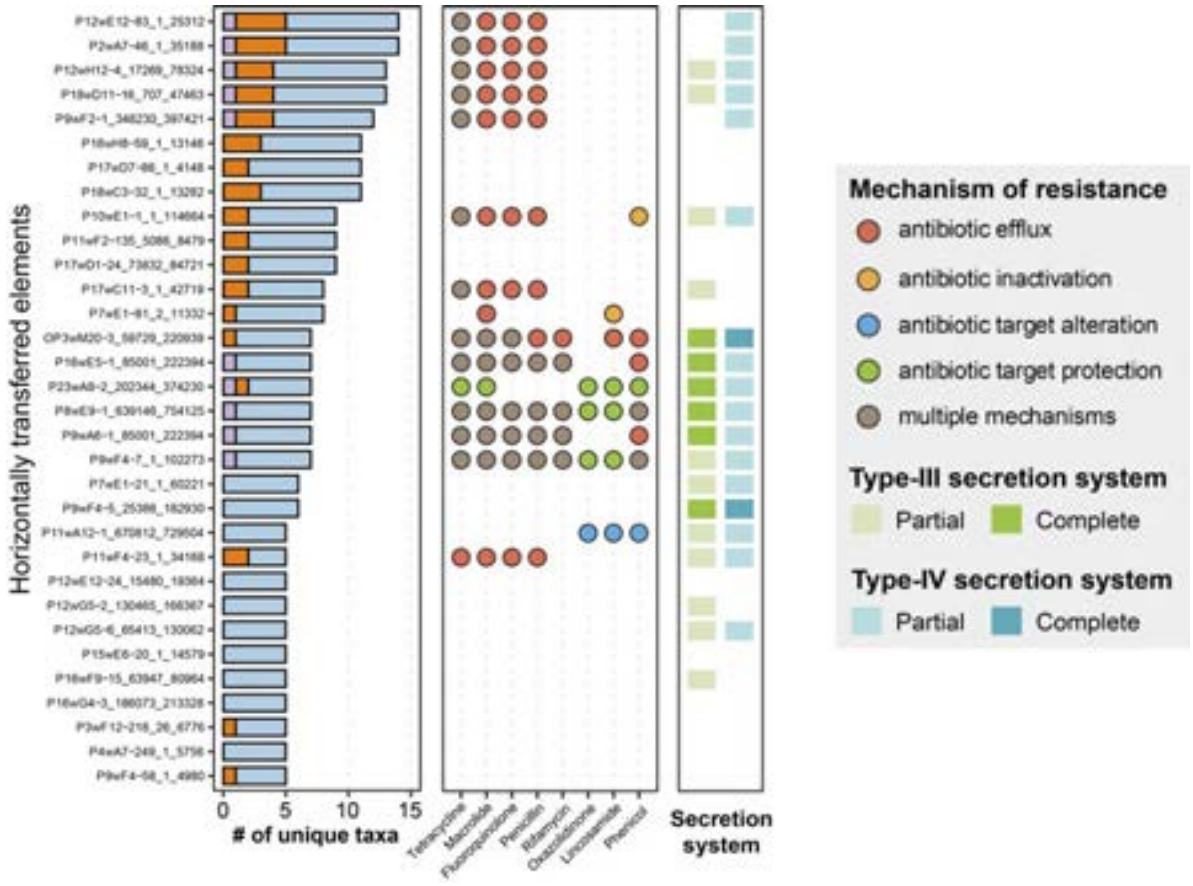
362 = Otu217 *C. qucibialis*

357 = Otu6 *C. aerofaciens*
354 = Otu14 *E. rectale*
366 = Otu2 *B. longum*
379 = Otu19 *P. johnsonii*

Huang et al, *Nature Biotechnology* doi: 10.1038/s41587-023-01674-2 (2023)

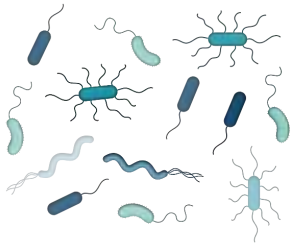
Prevalence and function of most widespread HGT elements





Huang et al, *Nature Biotechnology* doi: 10.1038/s41587-023-01674-2 (2023)

High-throughput transcriptomics to study drug-microbiota interactions



400+ drug-microbe combos

H1 isolates
- *Bacteroides doreii*
- *Collinsella aerofaciens*
- *Dorea longicatena*
- *Alistepes shahii*
- *Bifidobacterium adolescentis*
- *Parabacteroides distatonis*
- *Eubacterium rectale*
- *Bacteroides stercoris*
- *Bacteroides uniformis*
- *Bacteroides fragilis*
- *Bacteroides vulgatus*
- *Bifidobacterium longum*
- *Fuscatinebacter sacchivorans*

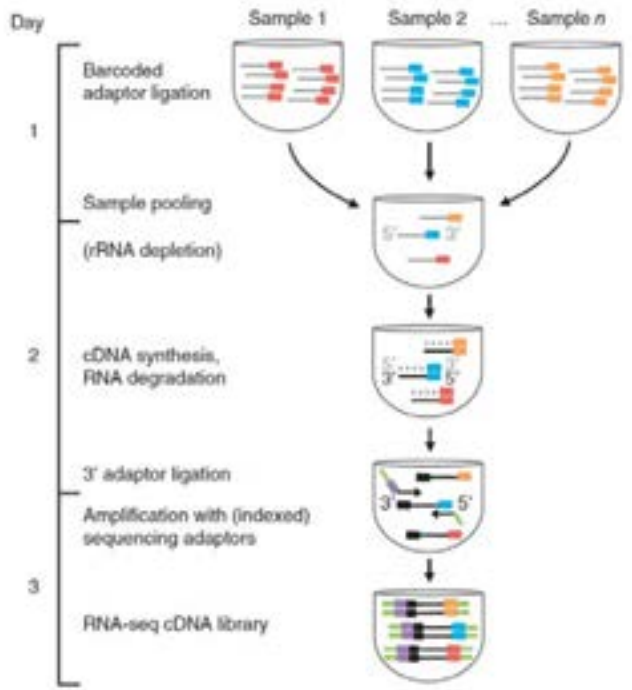
ATCC
- *Escherichia coli*
- *Bacteroides vulgatus*
- *Bacteroides uniformis*
- *Bacteroides fragilis*
- *Fuscatinebacter sacchivorans*
- *Eubacterium rectale*

X

- Lisinopril, ACE inhibitor
- Metoprolol, Beta–blocker
- Omeprazole, PPI
- Lenalidomide, Chemotherapy
- Metformin, Anti–hyperglycemic
- Levothyroxine, Hormone
- Amlodipine, CCI
- Venlafaxine, SSRI, SNRI or NDRI
- Bupropion, SSRI, SNRI or NDRI
- Trazodone, SSRI, SNRI or NDRI
- Escitalopram, SSRI, SNRI or NDRI
- Amitriptyline, SSRI, SNRI or NDRI
- Citalopram, SSRI, SNRI or NDRI
- Sertraline, SSRI, SNRI or NDRI
- Paroxetine, SSRI, SNRI or NDRI
- Duloxetine, SSRI, SNRI or NDRI
- Fluoxetine, SSRI, SNRI or NDRI
- Atorvastatin, Statin
- Simvastatin, Statin

RNAtag-seq

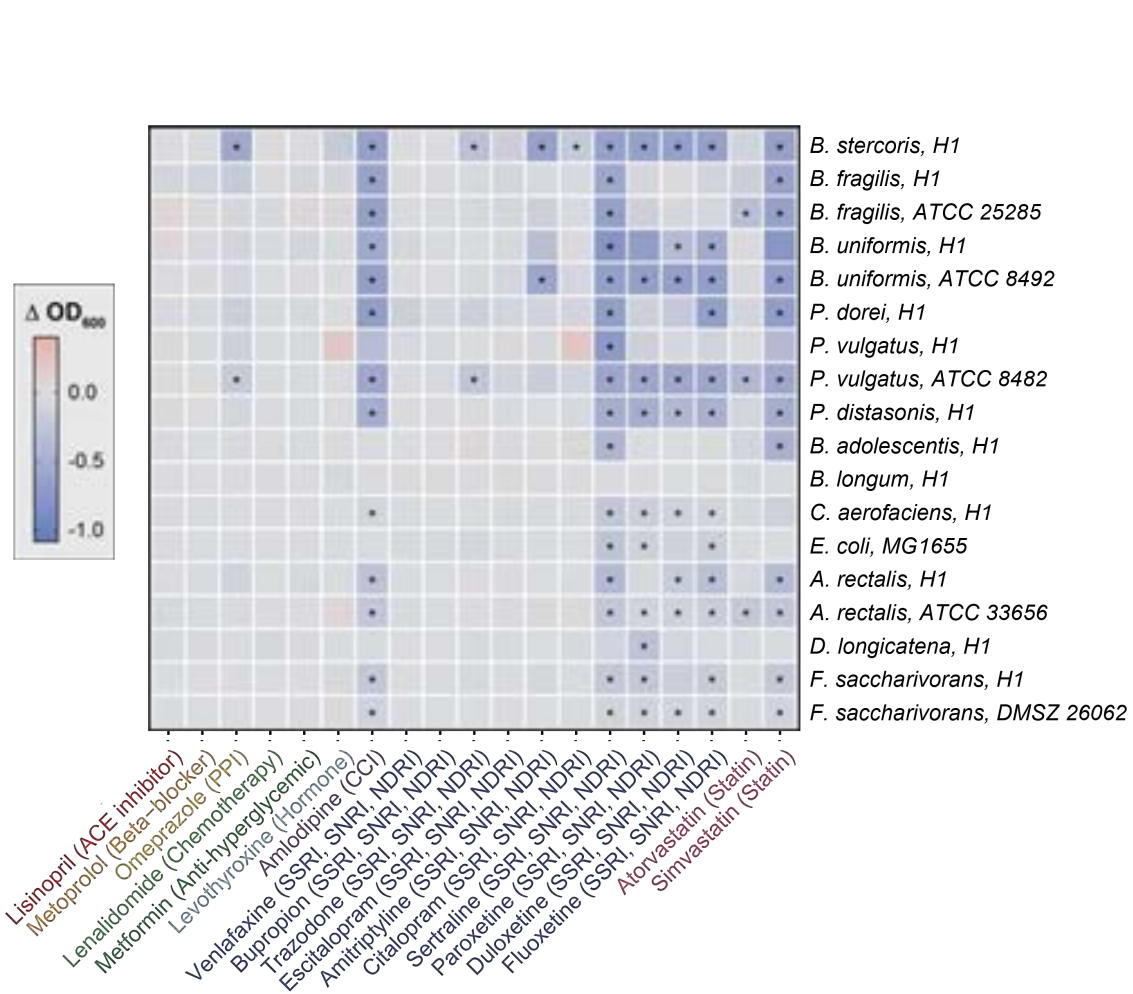
~$20/txome sample of 2M reads

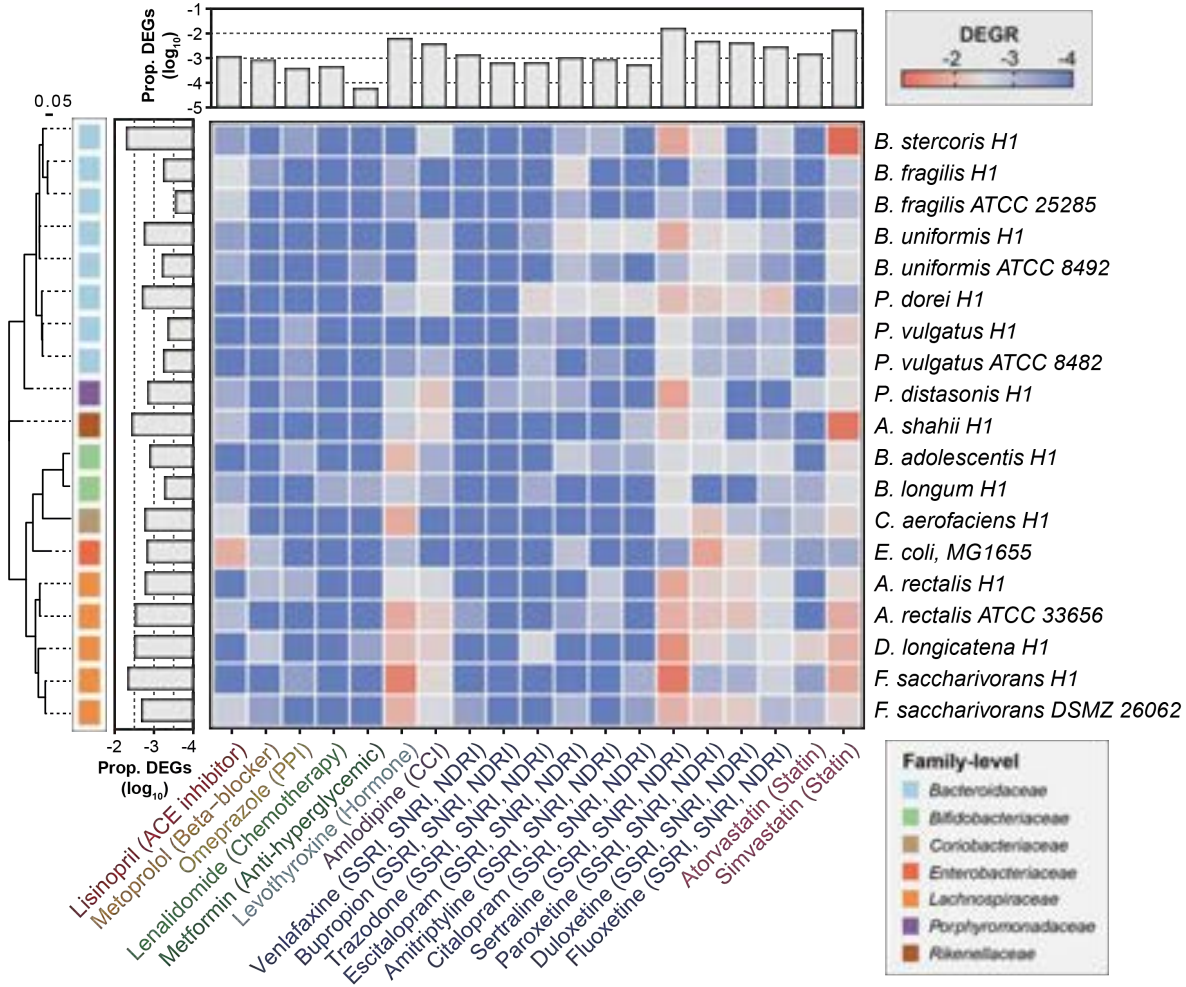


Huang, *NAR,* doi: 10.1093/nar/gkz1169 (2019)
Shishkin, *Nature Methods*, 12(4):323-5 (2015)

Ricaurte, Huang, et al, *Nature Microbiology* (accepted, 2023)

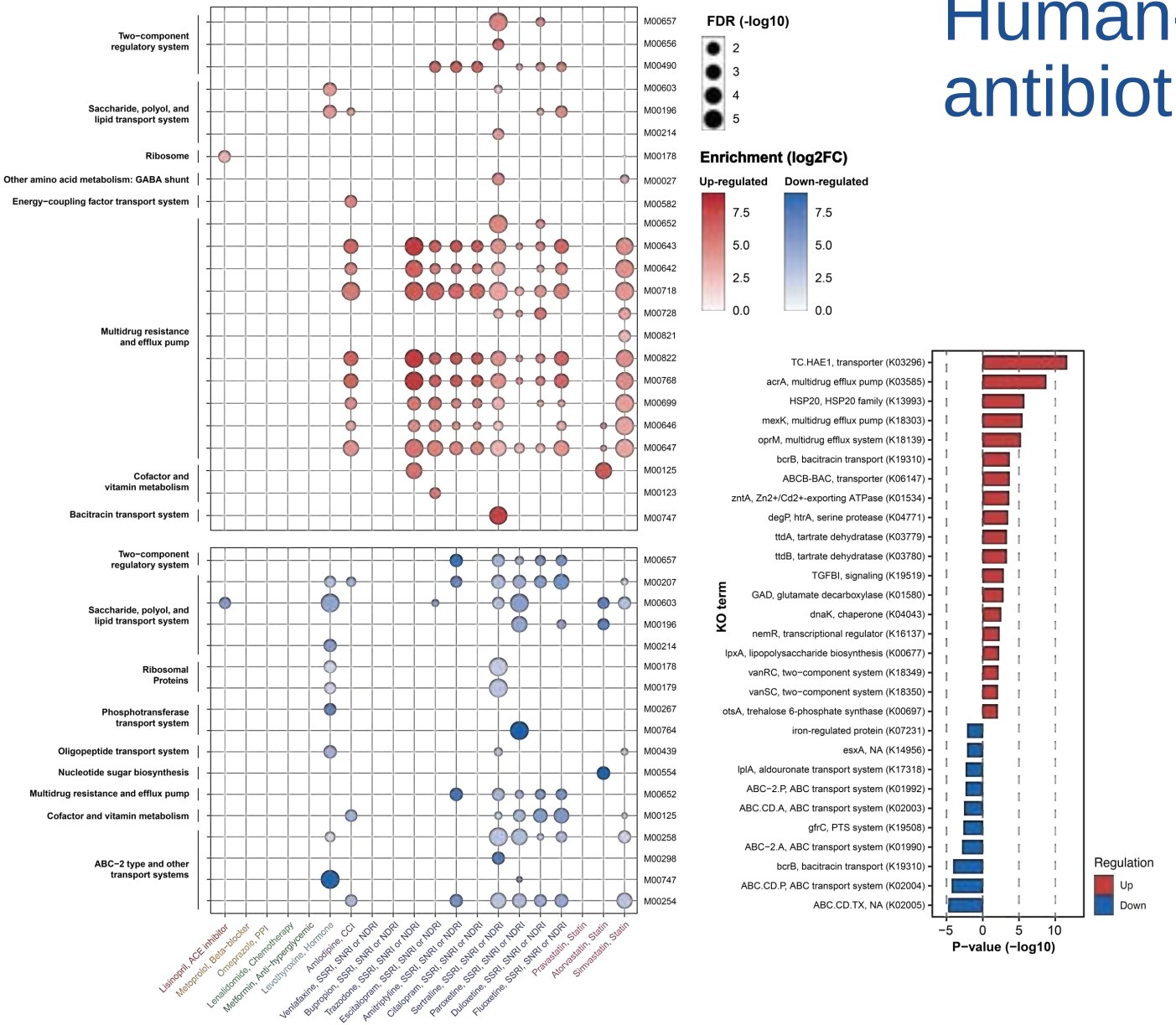
Bacteria produce robust transcriptional responses to top drugs





Ricaurte, Huang, et al, *Nature Microbiology* (accepted, 2023)

Human-targeted drugs promote antibiotic resistance responses

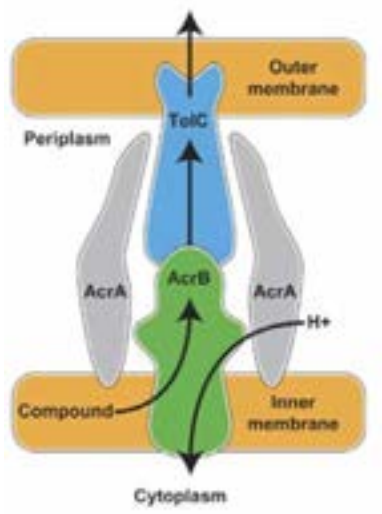


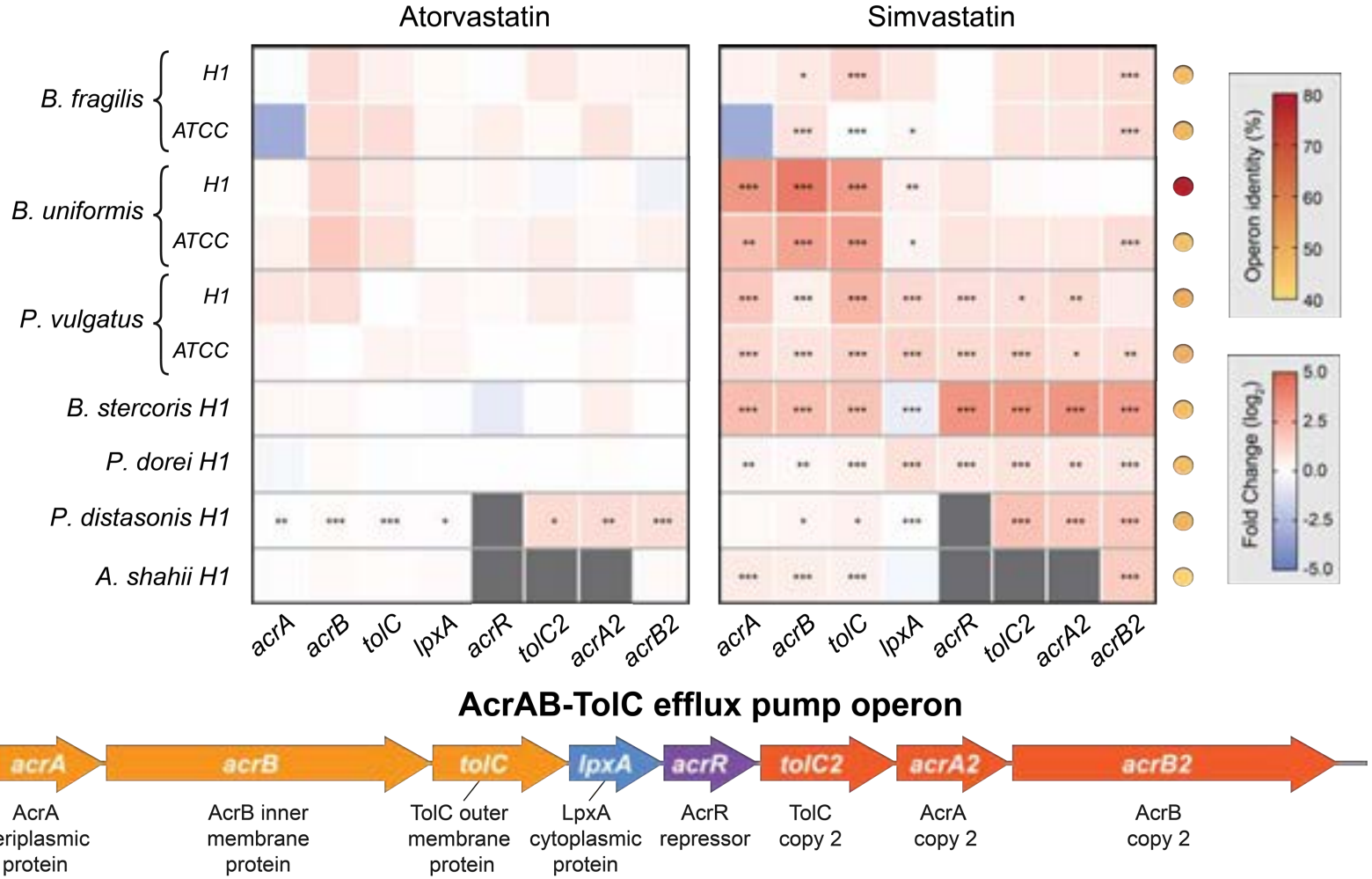
Enriched Pathways

- Transport
- Multi-drug resistance
- Two-component systems

Ricaurte, Huang, et al, *Nature Microbiology* (accepted, 2023)

Example: Statin-induced host-factor toxicity





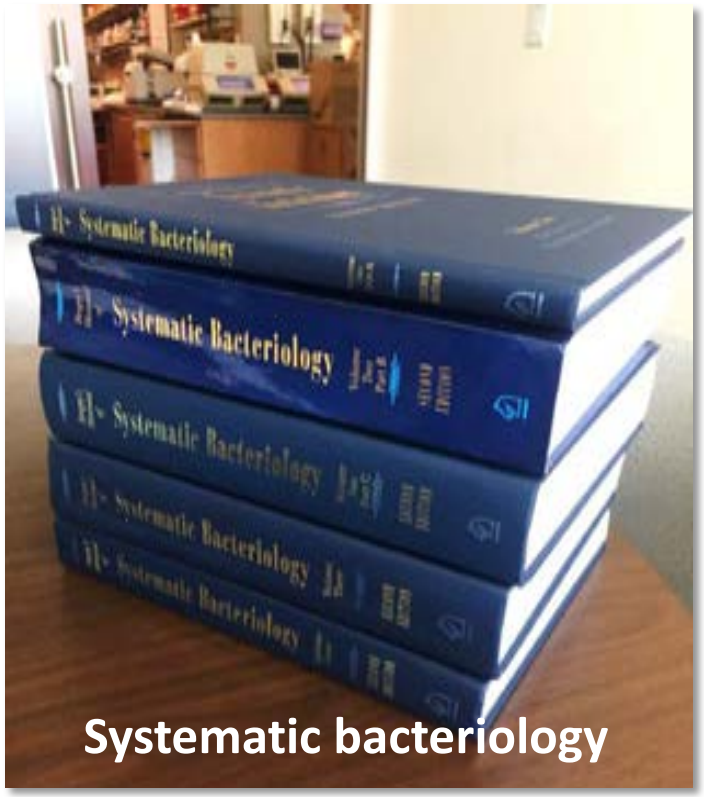
Ricaurte, Huang, et al, *Nature Microbiology* (accepted, 2023)

Need more organized systematic data to train next gen models: transcriptomics, metabolomic, phenomic, imaging

TABLE 191. Characteristics of species of the genus *Ruminococcus*[a]

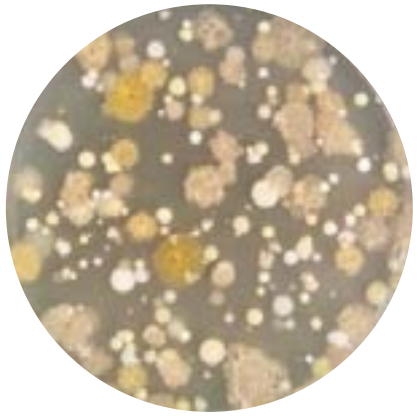
Characteristic	*R. flavefaciens*	*R. albus*	*R. bromii*	*R. callidus*	*R. gnavus*	*R. hansenii*	*R. hydrogenotrophicus*	*R. lactaris*	*R. luti*	*R. obeum*	*R. productus*	*R. schinkii*	*R. torques*
16S rRNA gene accession no.	X83430	X85098	X85099	X85100	D14136	D14155	X95624	L76602	AJ133124	X85101	D14144	X94965	L76604
DNA G+C content (mol%)	39–44	43–46	39–40	42	41	37–38	45	43	43.3	45	44–45	46–47	43
Major PYG product	A, F, S	A, F	A	S, a	A, F	L, a	A	A, F	A	A	L, a	A	L, a
Fermentation of:													
Arabinose	–	–	–	–	+	–	–	–	+	+	+	+	–
Cellobiose	+	+	–	+	–	–	+	–	+	+	+	+	–
Glucose	–	+	+	+	+	+	+	+	+	+	+	+	+
Lactose	+	+	–	+	–	+	ND	+	+	+	+	ND	+
Mannose	–	+	w/–	–	–	–	d	w/–	+	+	+	+	w/–
Maltose	–	–	+	+	+	+	ND	d	+	+	+	+	w
Mannitol	–	–	–	–	–	–	–	+	–	–	+	ND	–
Raffinose	–	–	–	+	+	+	ND	–	+	+	+	+	–
Sucrose	–	+	–	+	–	–	–	–	+	–	+	+	–
Xylose	–	–	–	w/–	+	–	ND	–	+	+	+	+	–

[a]Symbols: +, >85% positive; d, different strains give different reactions; –, 0–15% positive; w, weak reaction; ND, not determined; A, acetate; F, formate; S, succinate; L, lactate.



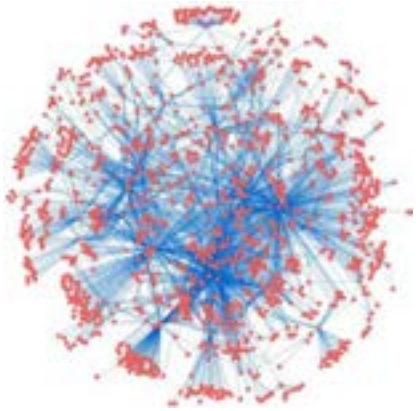
Systematic bacteriology

Dream slide: culturomics + phenotypic/transcriptomic analysis with large-scale perturbations

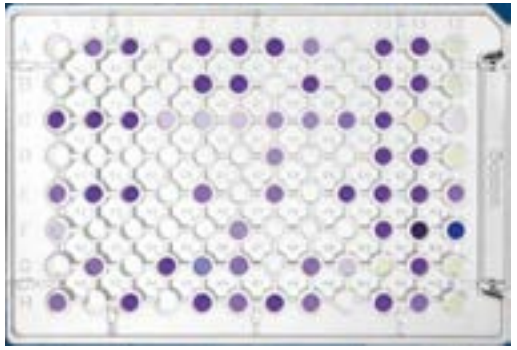


culturomics

genomics

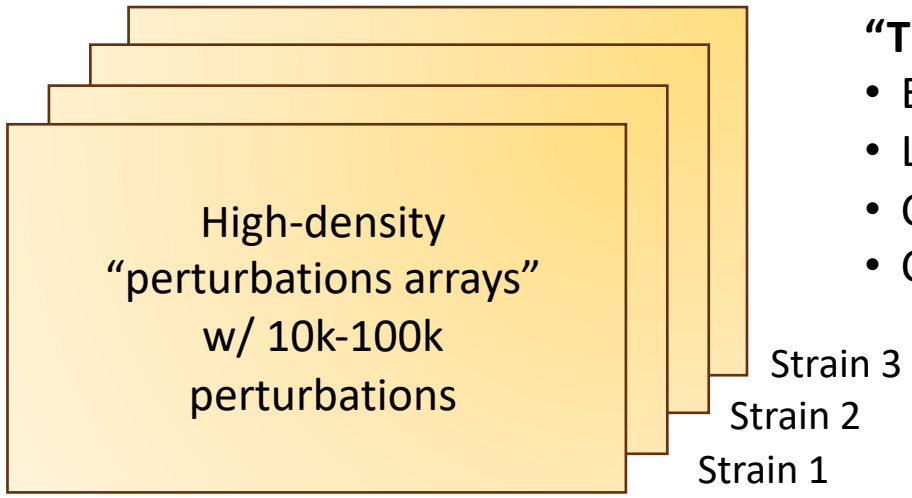


transcriptomics

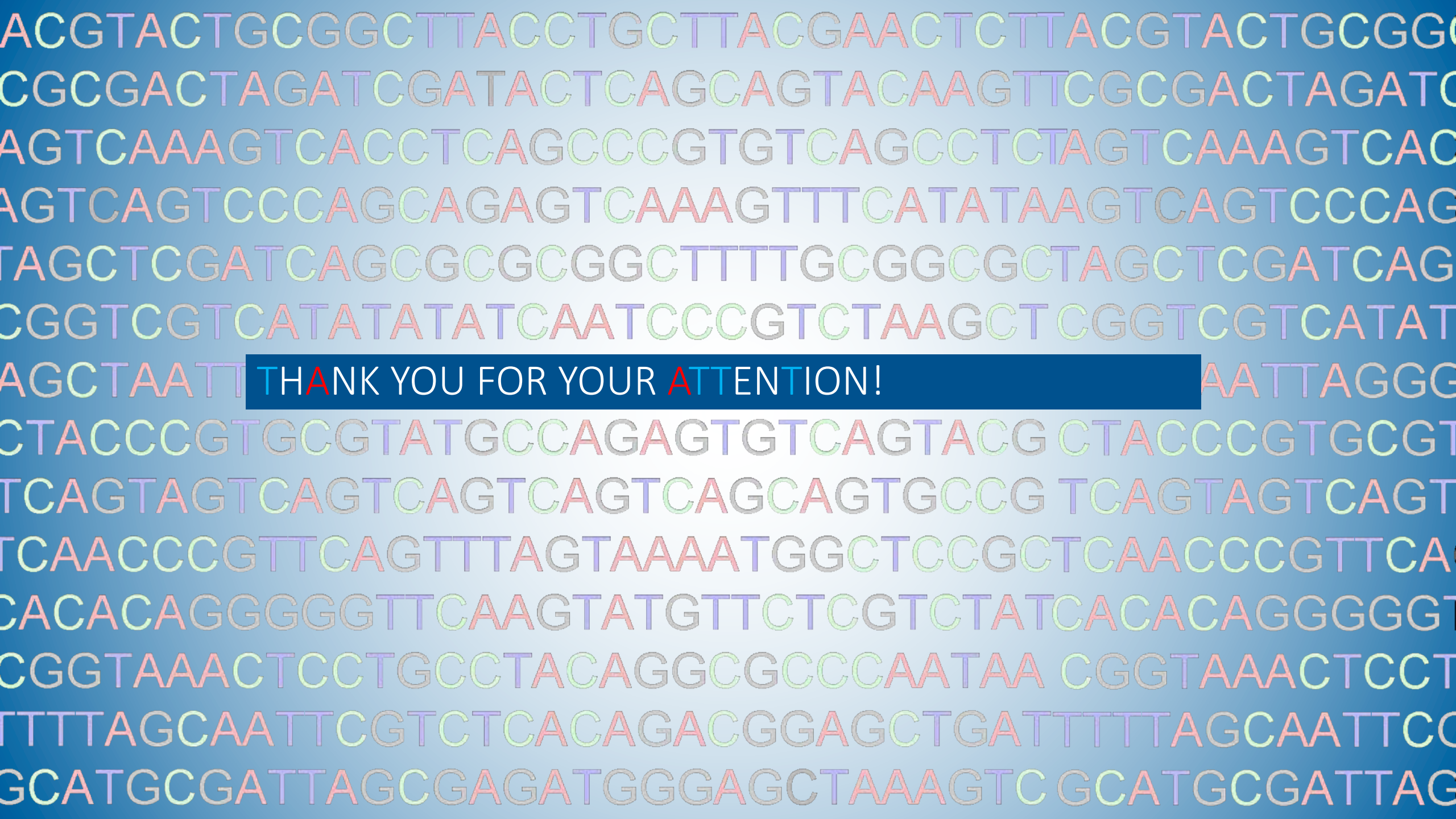


metabolomics/phenomics

X 10,000s of perturbations
- metabolites
- xenobiotics
- other supernatants
- growth conditions
- genetic KO/activation



“The Stimulator”
- Each well is addressable
- Leverage spatial omics
- Capture kinetic data
- Can train large AI models


THANK YOU FOR YOUR ATTENTION!

Giant Viruses: A Treasure Trove of Unknome Function

Frederik Schulz, DOE Joint Genome Institute, Lawrence Berkeley National Laboratory

December 12th, 2023

JGI

BERKELEY LAB

U.S. DEPARTMENT OF ENERGY

Office of Science

NeLLi New Lineages of Life

Identification and prioritization of biosynthetic gene clusters for commercial (meta-)genome mining

Zachary Charlop-Powers
R&D Director, Ginkgo Bioworks
December 12, 2023

Property of Ginkgo Bioworks

Genomics aided host and strain engineering for biotechnology

Aindrila Mukhopadhyay

Senior Scientist

Biological Systems and Engineering Division

Lawrence Berkeley National Laboratory

Dec 12th 2023

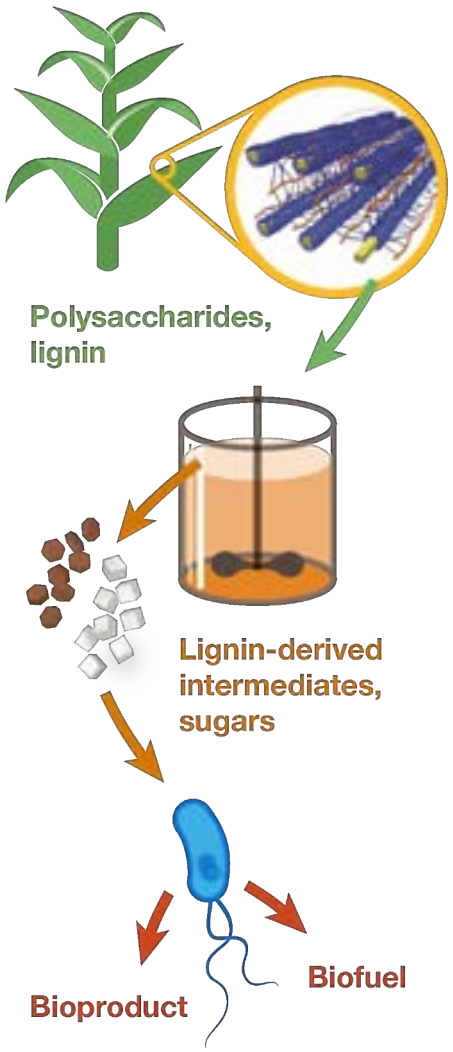
Large multi-team projects at LBNL

BioSciences

ENIGMA
Ecosystems & Networks Integrated with Genes & Molecular Assemblies

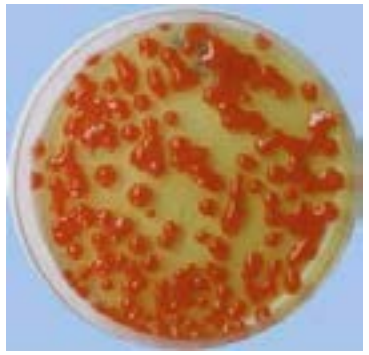
JBEI
Joint BioEnergy Institute

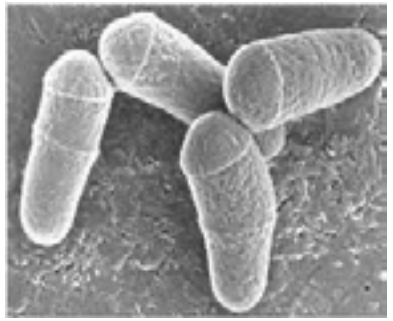
m-CAFEs



Many carbon sources can be used across a range of host systems form conversion to many targets

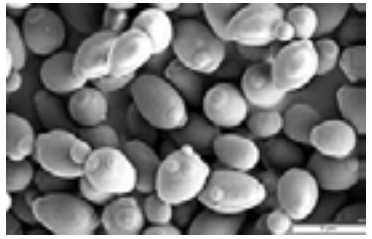
Martinez Garcia et al 2016 Env Mic



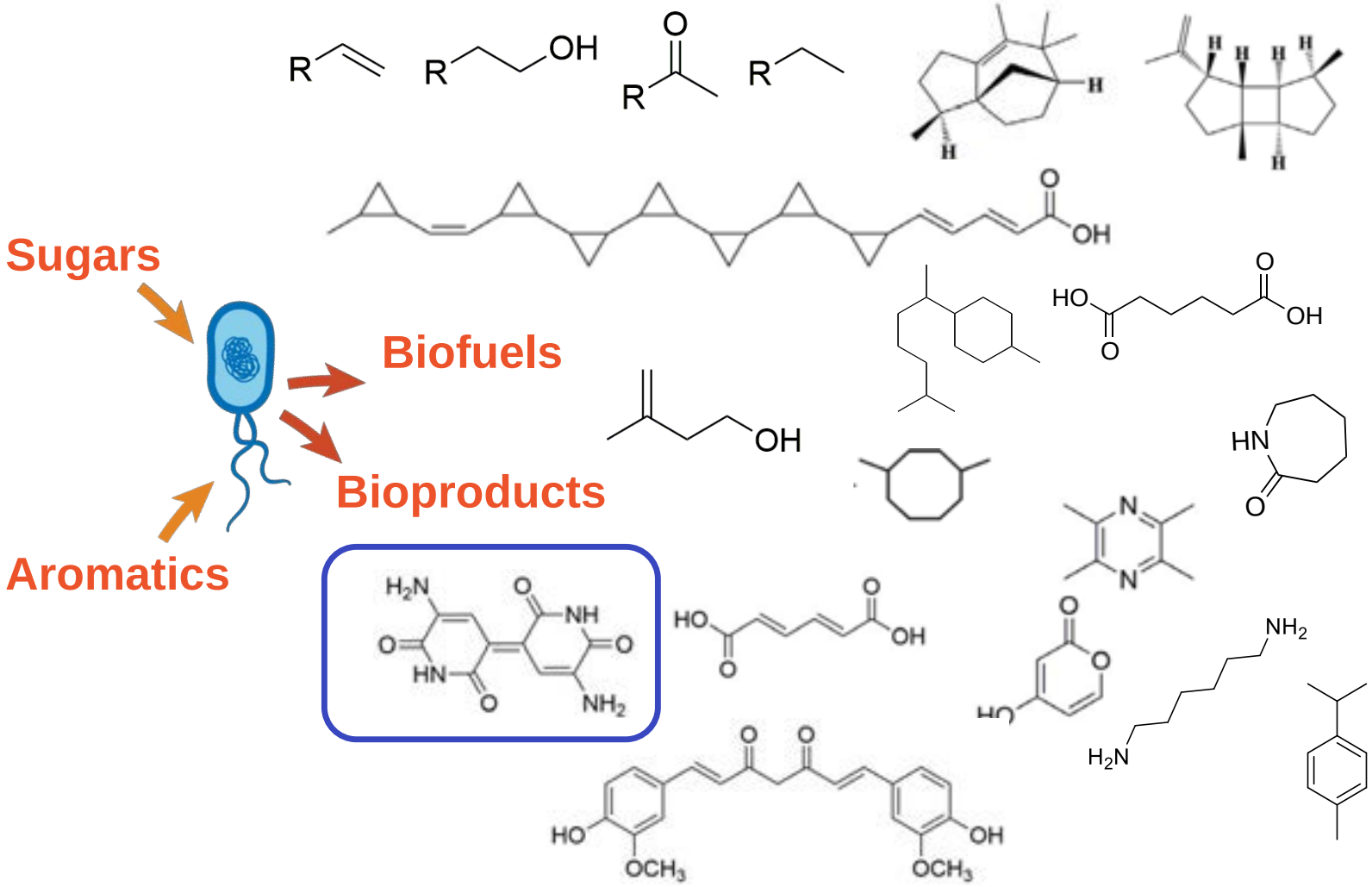


www.profacgen.com

education.nationalgeographic.org



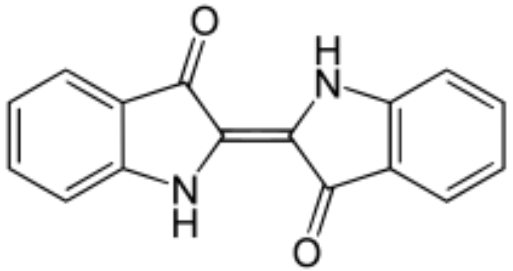
Mogana Das Murtey, Patchamuthu Ramasamy



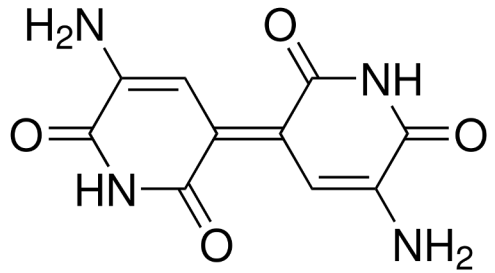
Bioproduct case study: sustainable materials for dyes and pigments

https://aecom.com/blog/la-denim-city-2

http://www.tejidosroyo.com/

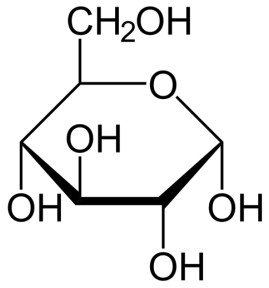


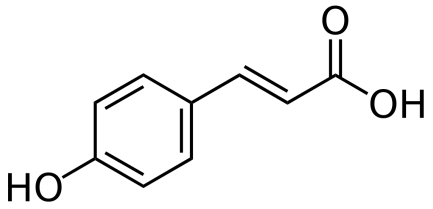
Indigo

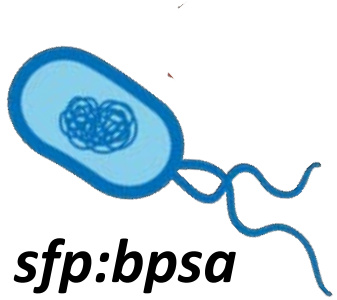


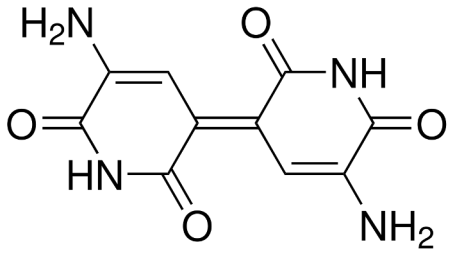
Indigoidine

Microbes with versatile catabolism can be engineered for such final products but scale up is challenging

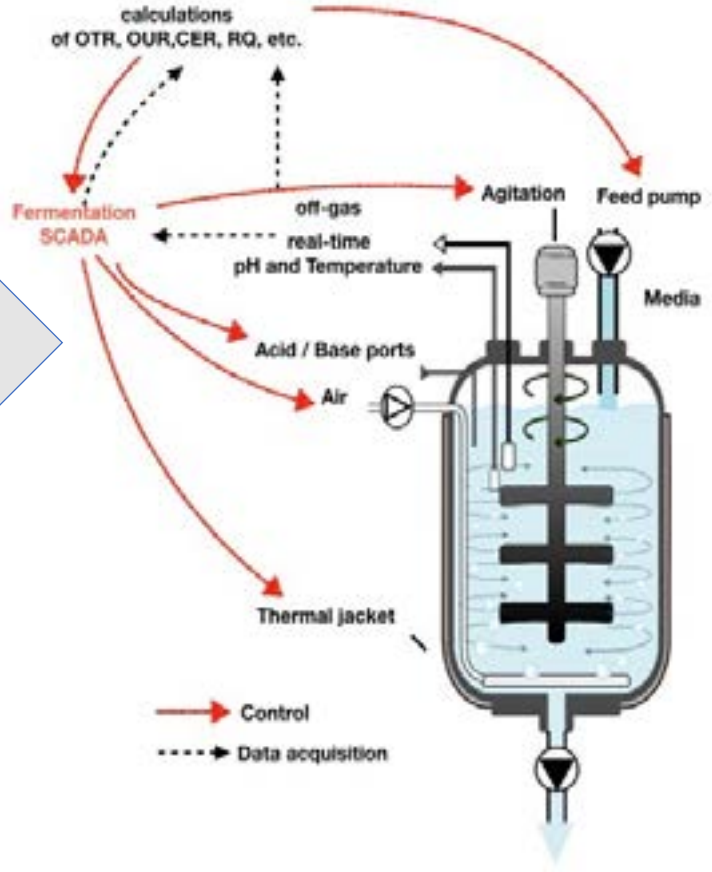








?



Scaled-up production using hydrolysate

Milled biomass

Mixing before pretreatment

After pretreatment

Enzymatic hydrolysis, 0 h

10% Cholinium lysinate

Enzymatic hydrolysis, 3 h

Sundstrom et al 2017 Green Chem

Development of the optimal host..

consume many biomass components

Engineer pathway to final products or precursors

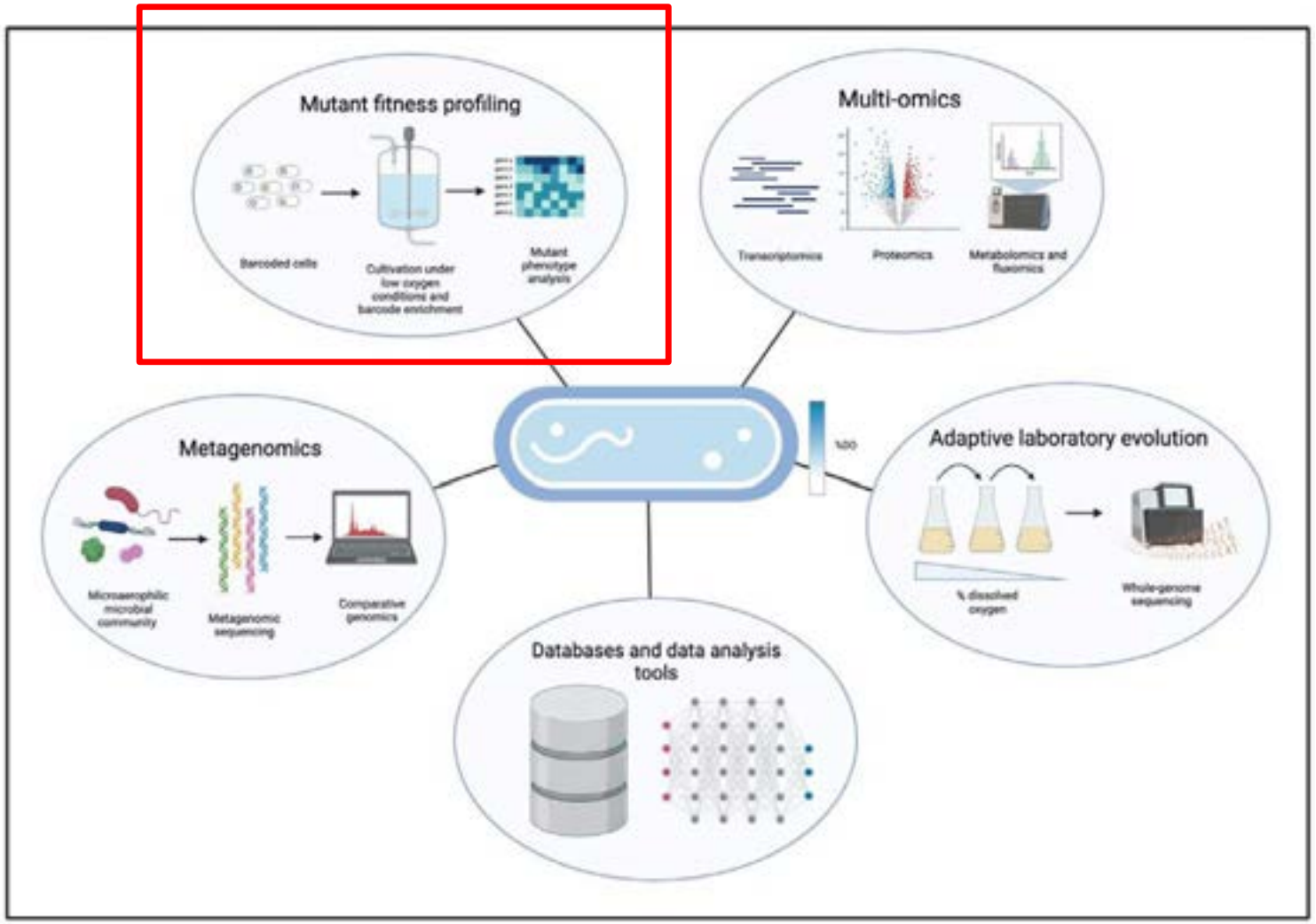
tolerate biomass hydrolysates

Flux to key intermediate or product

tolerate intermediates and final products

maintain the right metabolism and fitness

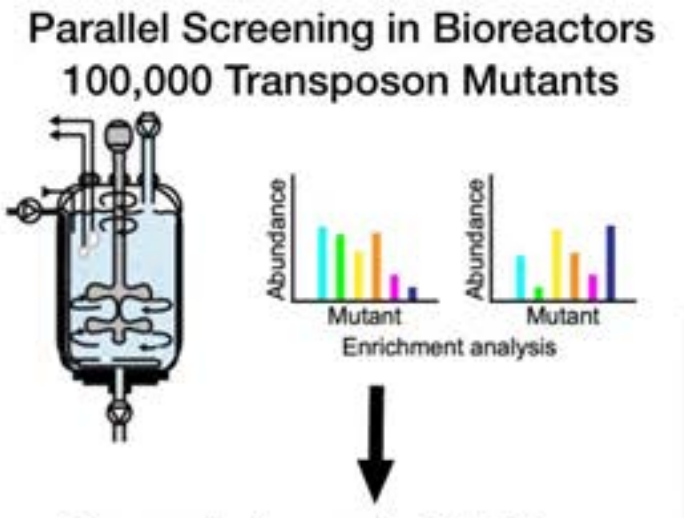
Functional genomics approaches can reveal many non-obvious targets

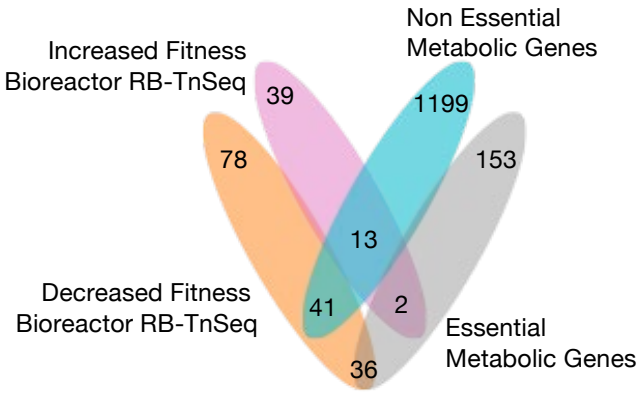


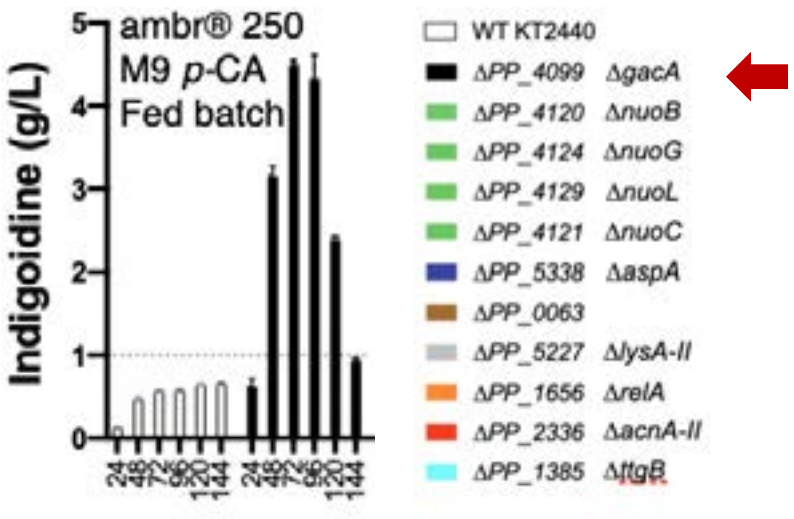
- Identification of key genes with known functions
- Role of non-metabolic genes and proteins
- New roles for known genes and proteins
- Genes with unknown functions
- Roles of regulators and signaling systems

Kulakowski et al 2023 COBIOT

Functional genomics and systems biology as approaches to identify new gene targets







Eng, Banerjee, et al., (2020) ***Met Eng.***

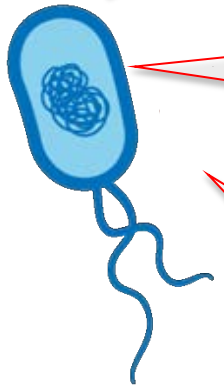
Development of the optimal host..

consume many biomass components

Engineer pathway to final products or precursors

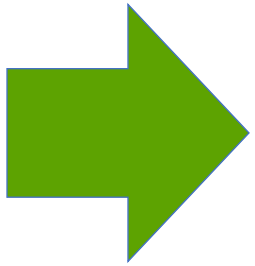
tolerate biomass hydrolysates

Flux to key intermediate or product

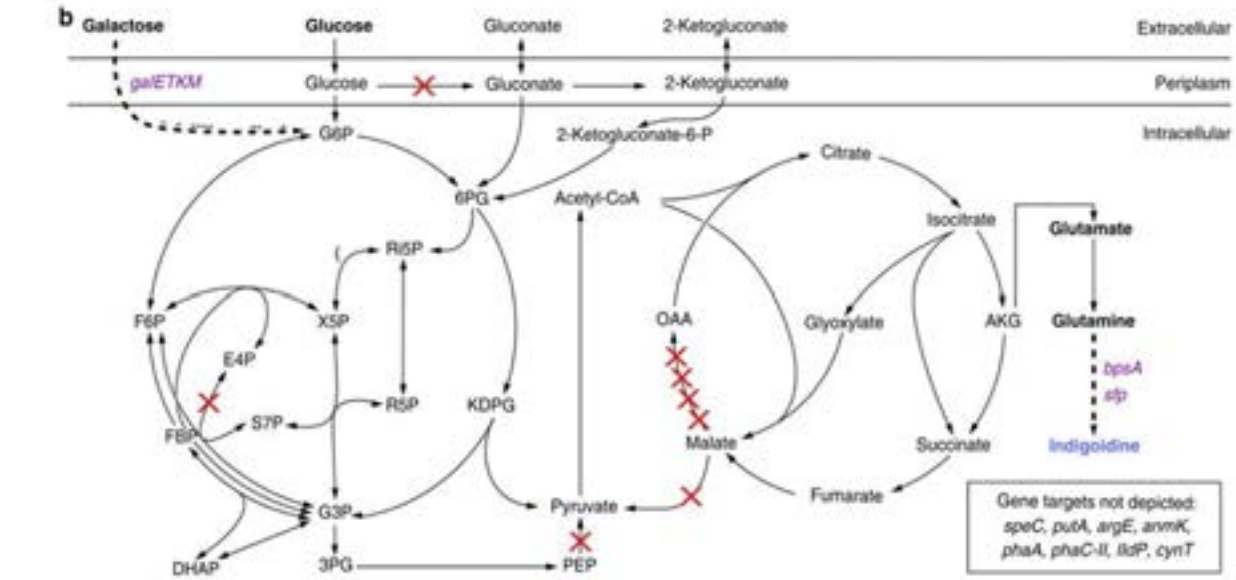


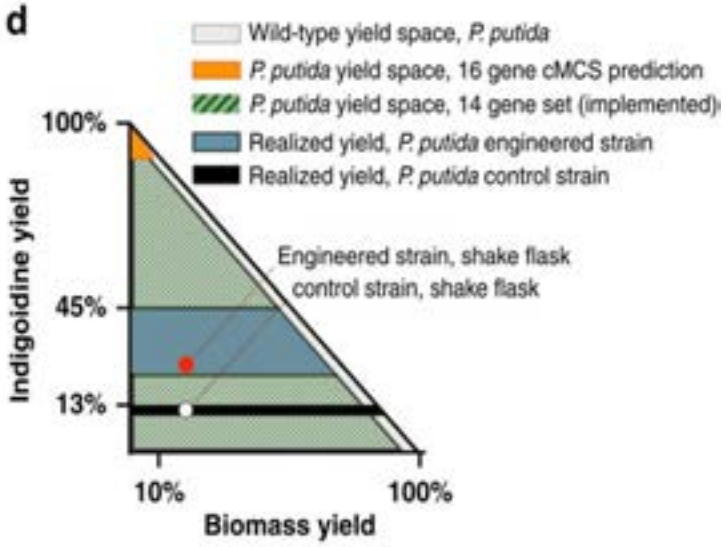
tolerate intermediates and final products

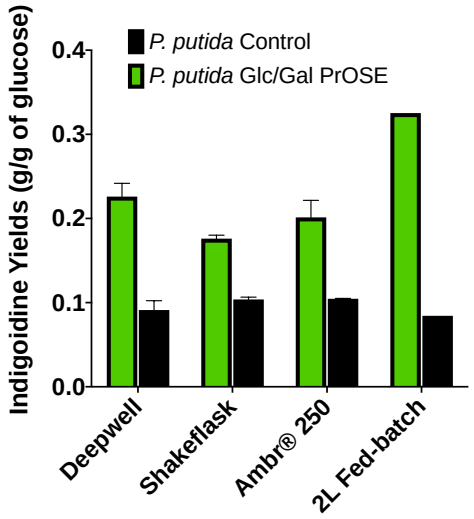
maintain the right metabolism and fitness



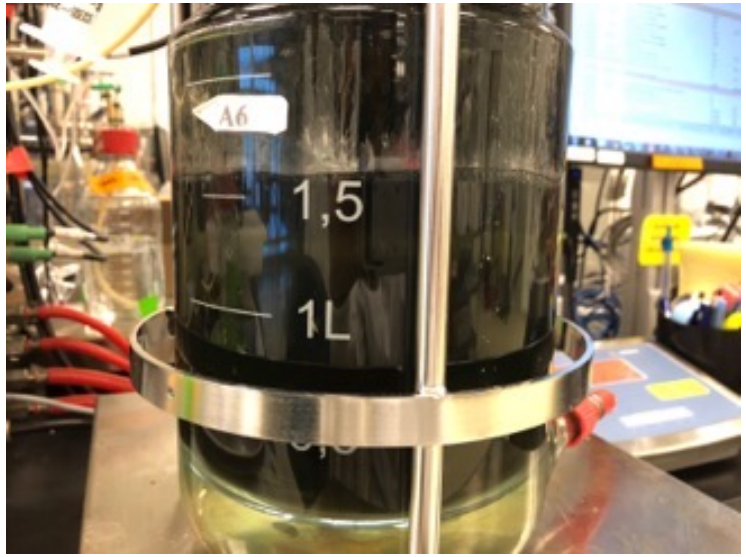
Systems biology driven metabolic rewiring for growth coupling







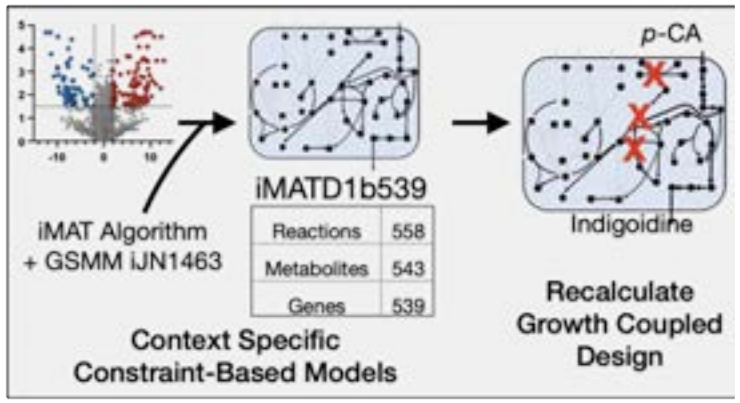
- Genome Scale model driven designs
- 14 independent genes simultaneously deleted
- Product substrate pairing results in high production

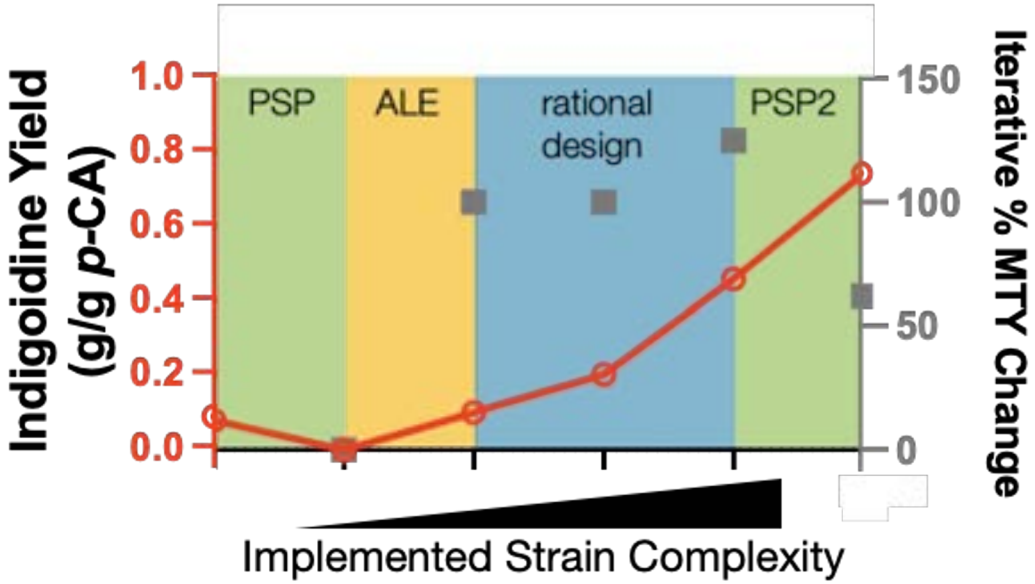


Banerjee, Eng, et al., (2020) ***Nat Comm.***

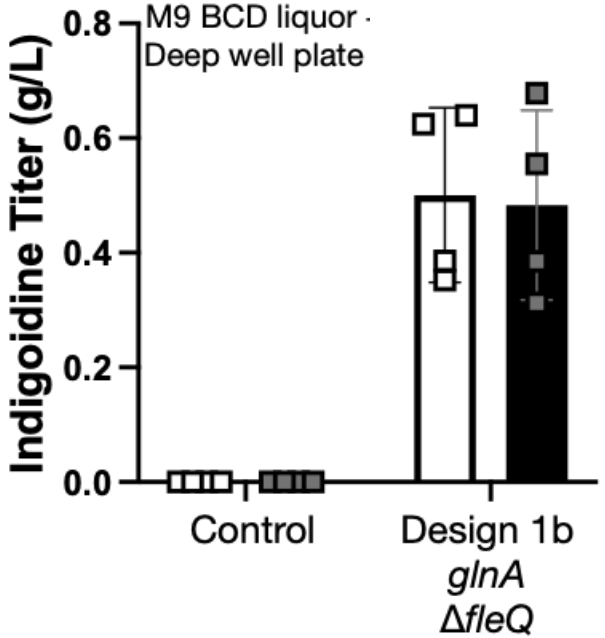
Iterative approaches using systems biology and functional genomics

Proteomics guided models





Conversion from hydrolysate



Products Substrate Pairing for aromatic carbon sources to bioproducts using Genome Scale Metabolic Models (GSMM)

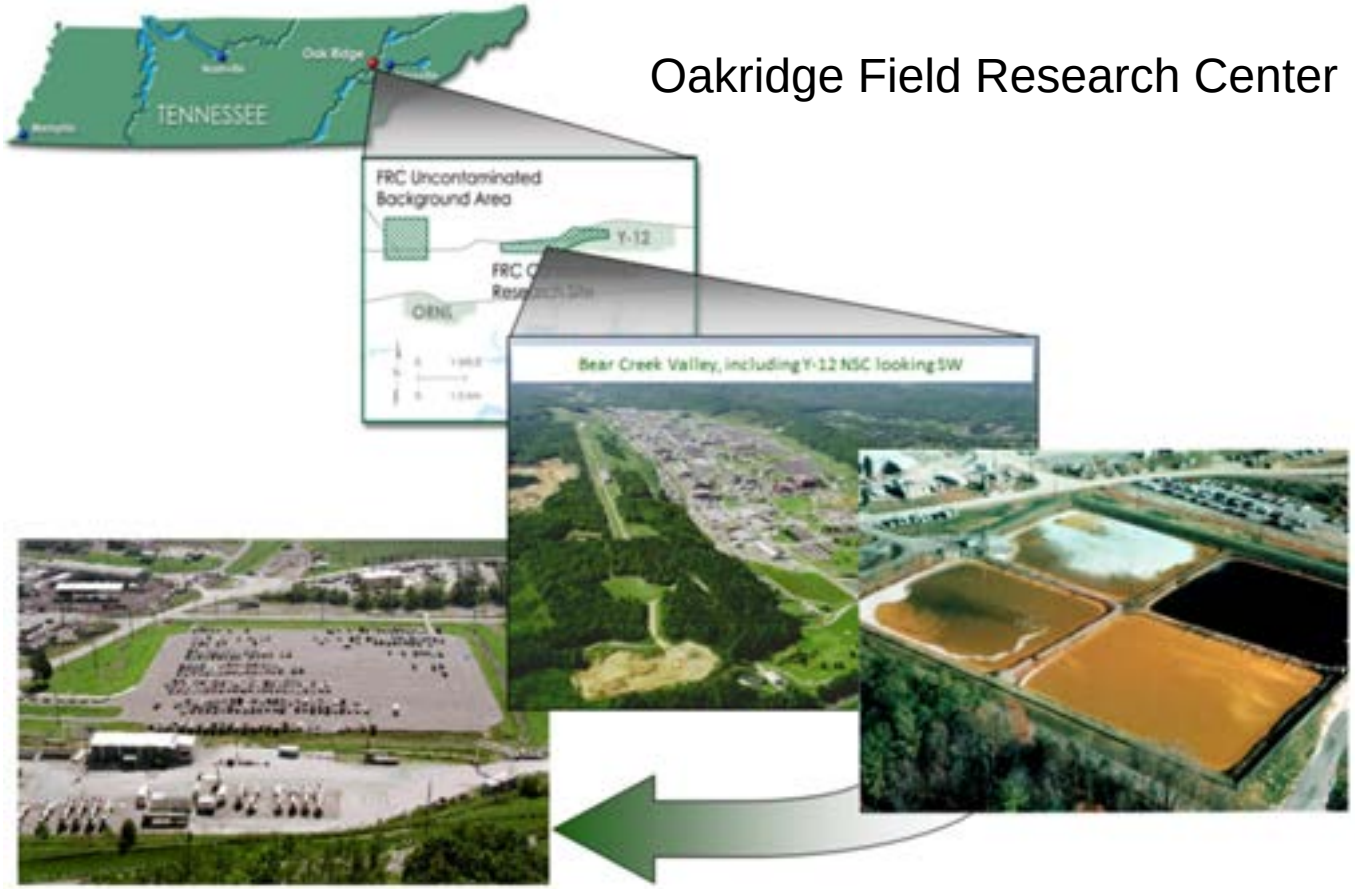
Omics reveals the roles of many metabolic and non-metabolic genes.

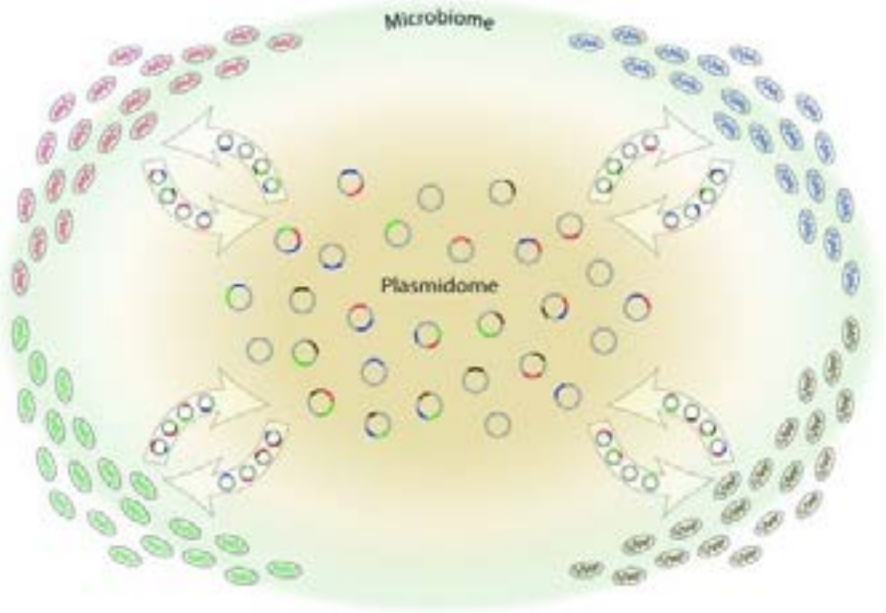
Complete implementation involved use of curated models, fitness data, 7 gene deletions, 2 modifications from rational engineering guided by proteomics, and Adaptive Lab Evolution

Eng, Banerjee, et al., (2023) ***Cell Rep.***

Discovery of new parts to enable genetic tractability

Oakridge Field Research Center

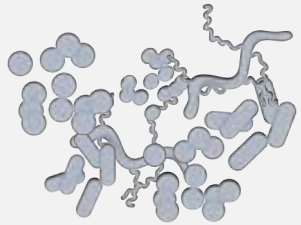




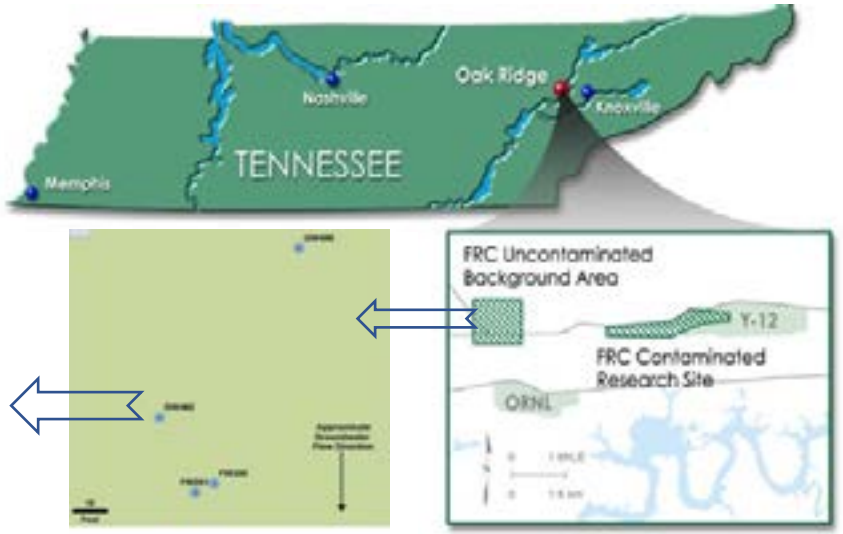
Mizrahi I - Mob Genet Elements (2012)

Isolation of mobile genetic elements from ground water samples

Discovery of new parts to enable genetic tractability



Environmental Bacteria

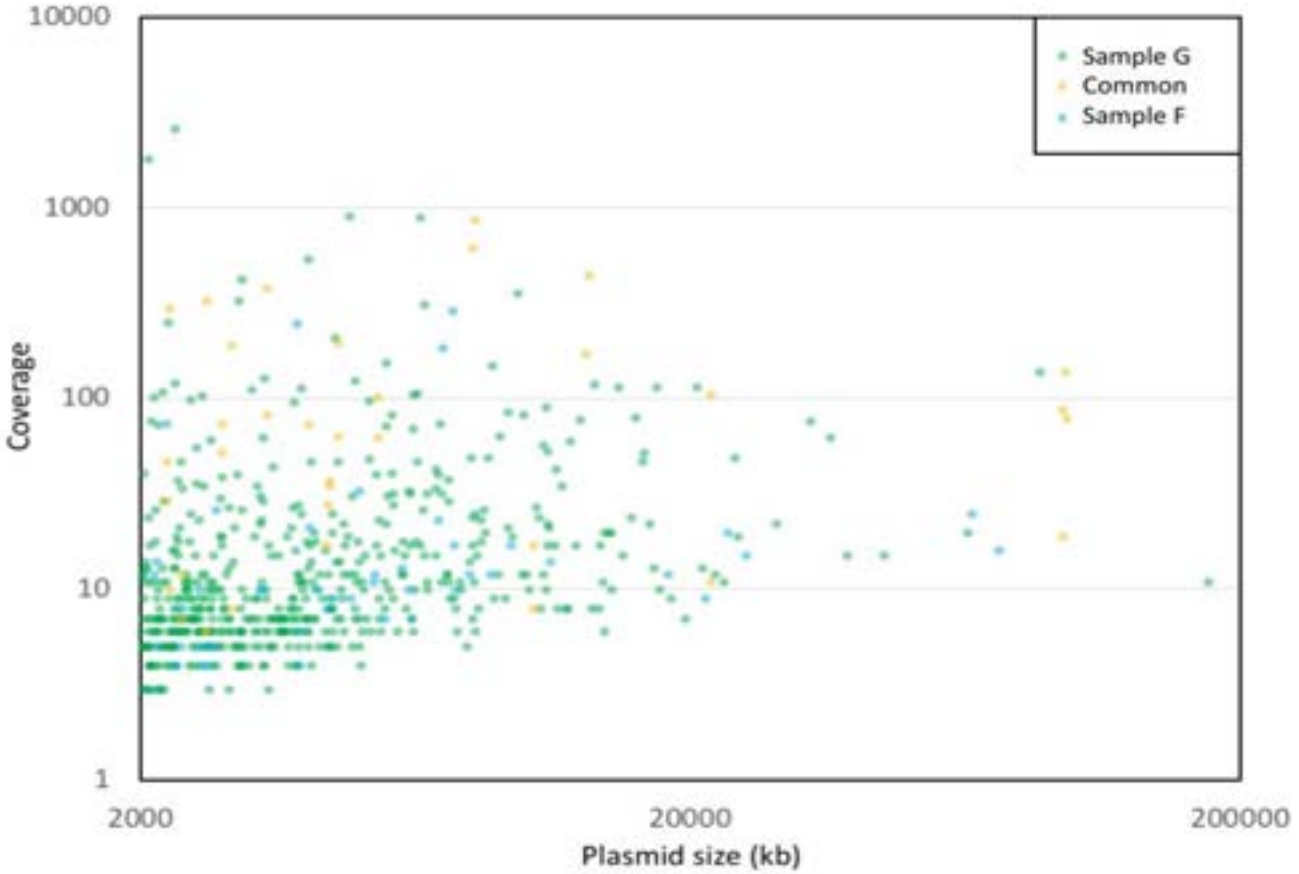


Sample Name	Well	Sampling Date	Filter Size (µm)
Sample_A	FW301	12/1/14	0.2
Sample_B	FW300	11/11/14	10
Sample_C	FW300	12/1/14	10
Sample_D	GW460	12/1/14	0.2
Sample_E	FW301	12/1/14	10
Sample_F	GW456	11/11/14	10
Sample_G	GW460	12/1/14	10

4l water filtered through 0.2 and 10µM filters

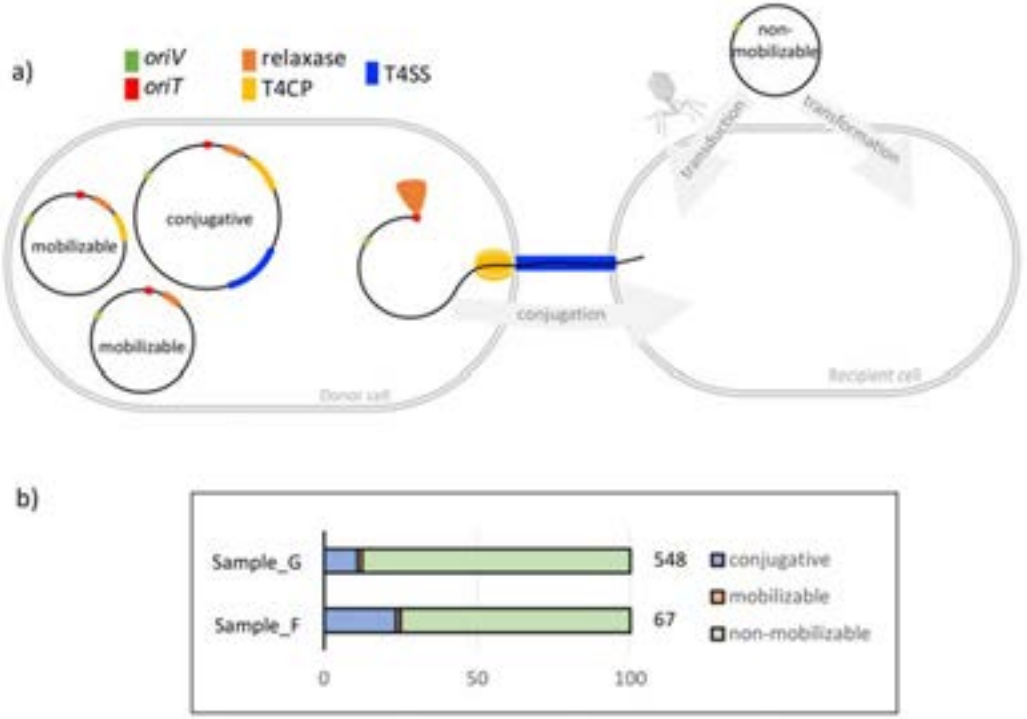
Kothari et al (2019) ***mBio***

Discovery of plasmid from the Oakridge FRC



- 1.7Mb plasmid, among largest ever found in a plasmidome studies
- 11 plasmids more than 50 kb in size

Plasmid distribution based on size and types

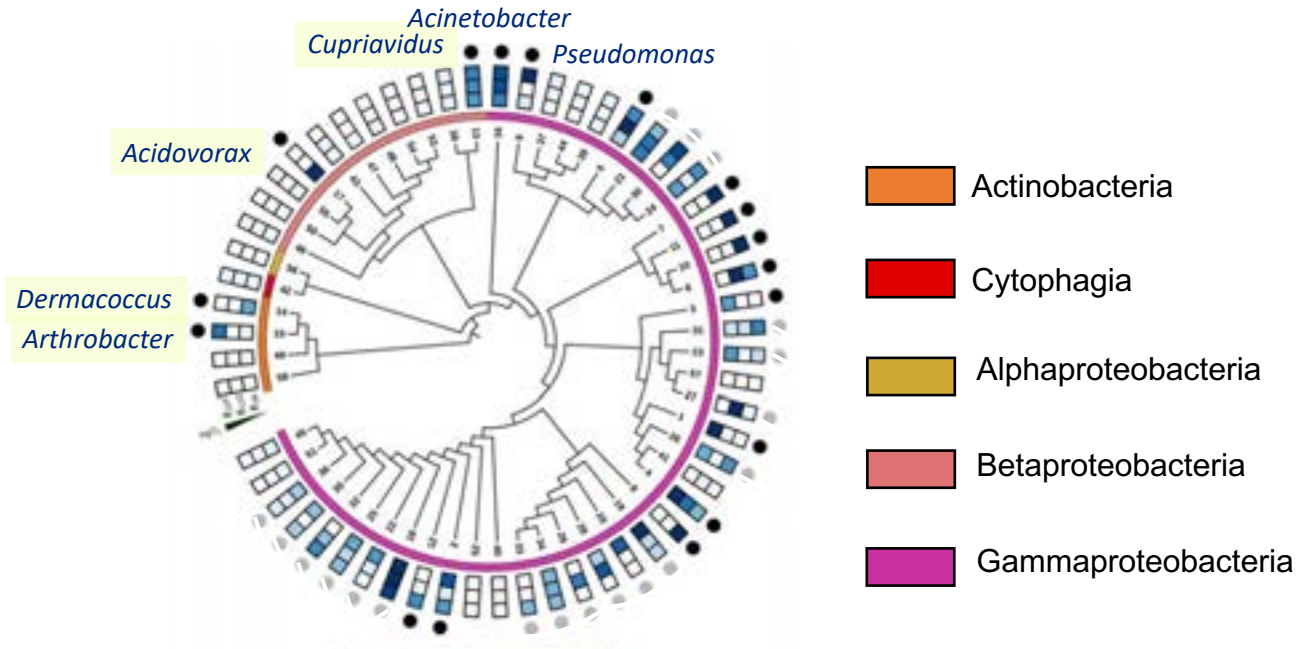


- Seven different incompatibility groups were identified

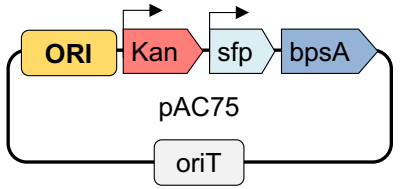
Kothari et al (2019) ***mBio***

Plasmids provide the first step to transformation and genetics

The most ubiquitous plasmid was tested across several isolates



Discovery of new origins to create new Synbio parts



612 plasmids discovered

19 novel predicted origins

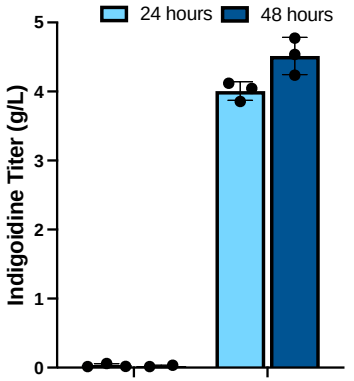
21 literature origins

9 promoters

2 kanamycin resistance markers

Magic pools library

Transform non-model microbes

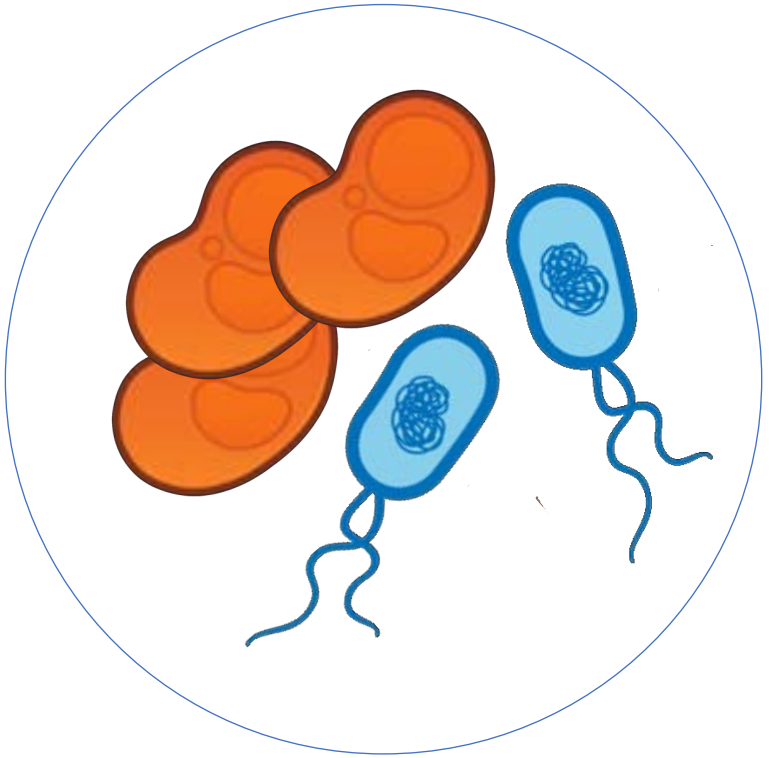


Origin validation

Kothari et al (2019) ***mBio***

Codik et al (2023) ***in prep***

Applications of new gene functions



Degradation of harmful substrates – toxins, explosives, biocidal agents

Conversion to valuable materials – biomanufacturing, therapeutics, chemicals, materials, fuels

Biosensor development – dynamic systems, diagnostics, tracking and measuring

Discover fitness targets – therapeutics
Precision synthetic communities, Ag application, probiotics, complex manufacturing

Our group at JBEI and LBNL

m-group.lbl.gov
www.jbei.org

Thanks to DOE BER for funding!!

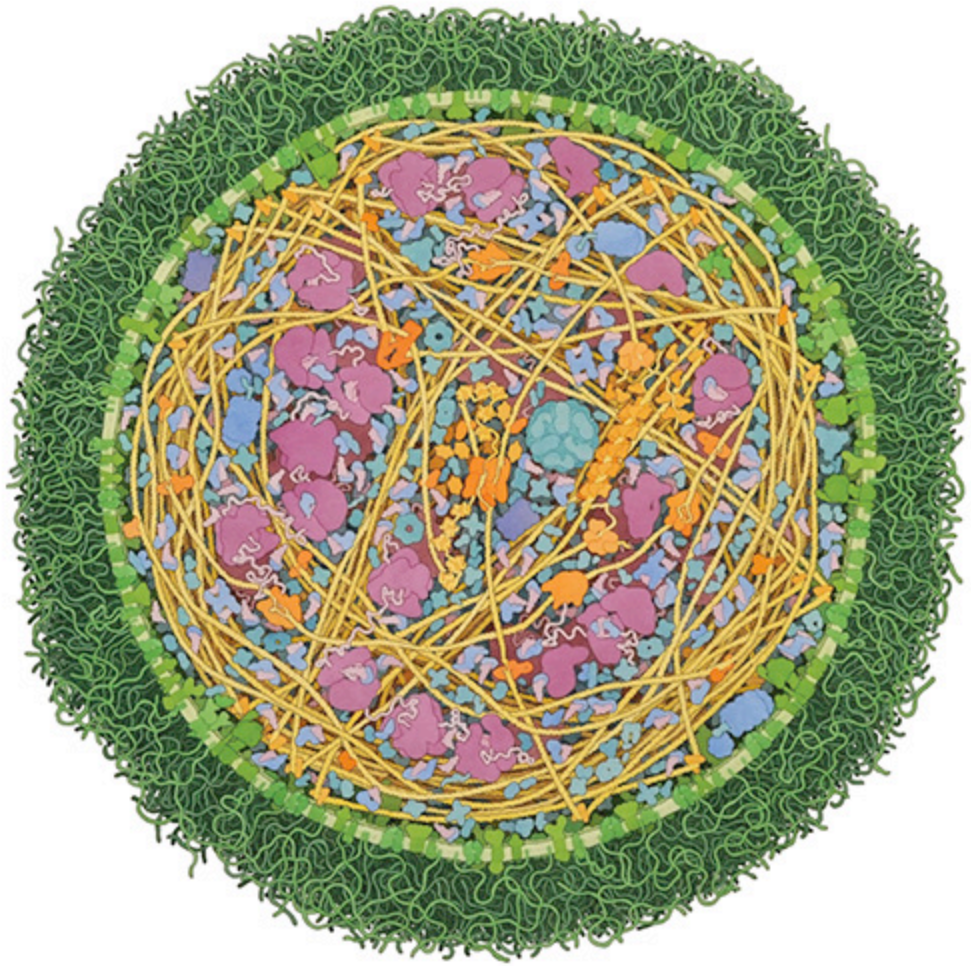
m-CAFEs ENIGMA Ecosystems & Networks Integrated with Genes & Molecular Assemblies JBEI

Integrative, multiscale modeling of cellular systems

Eran Agmon, PhD

Assistant Professor | University of Connecticut Health

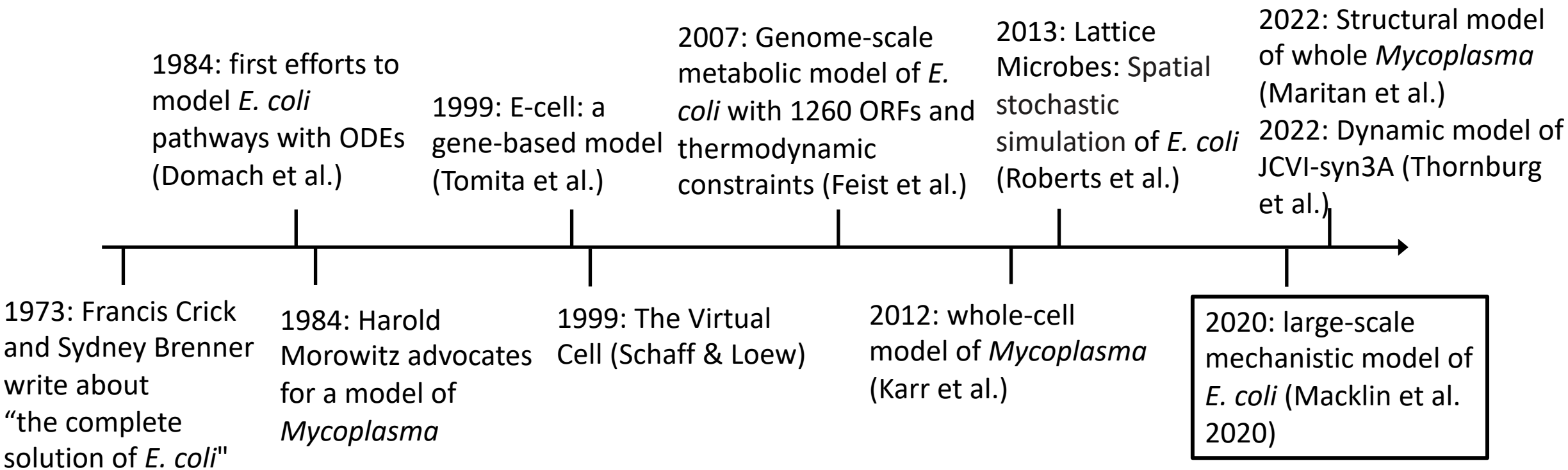
DARPA Discovering Unknown Function Workshop | 12/12/2023

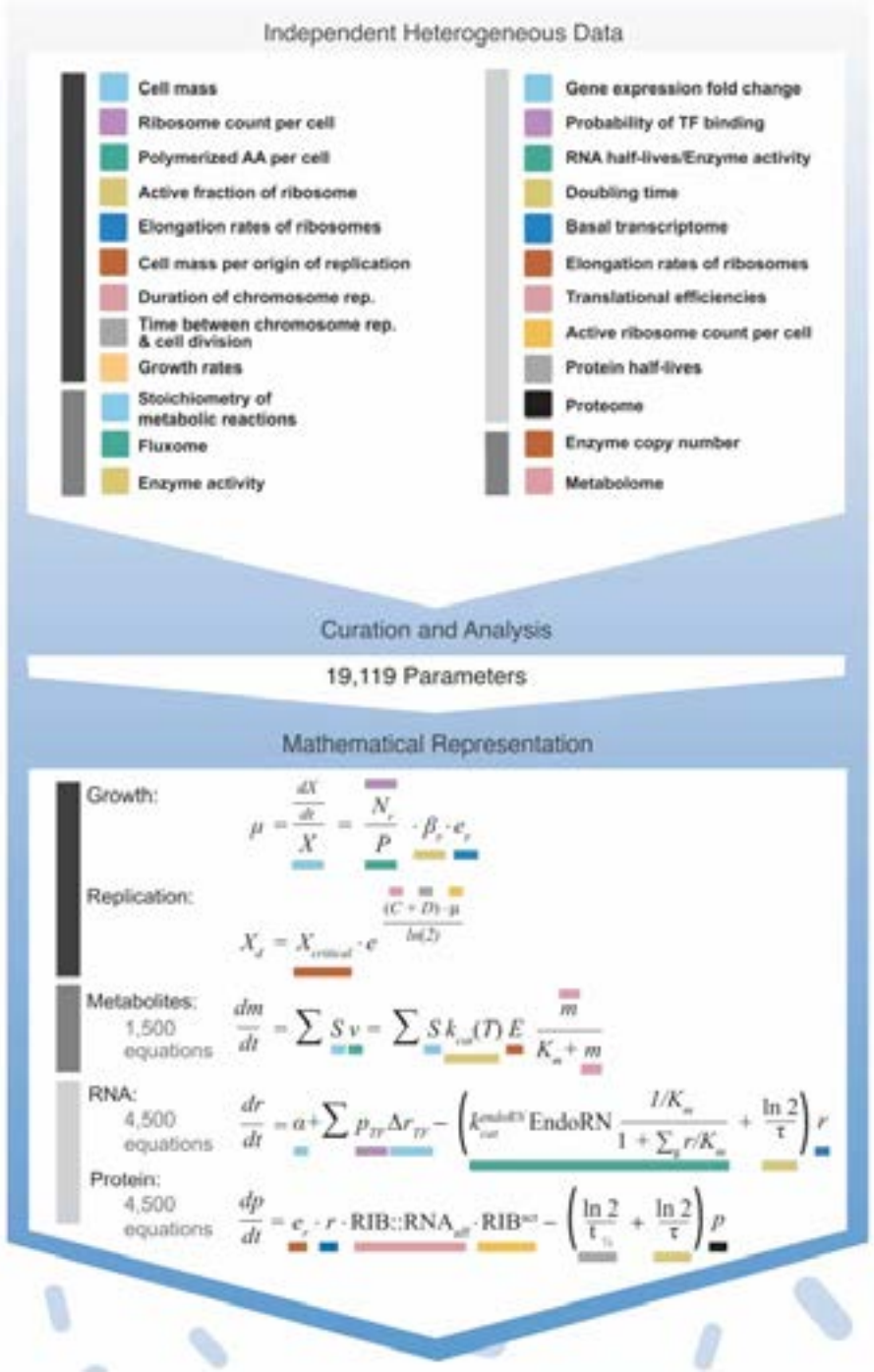


“a computer model is feasible, and every experiment that can be carried out in the laboratory can also be carried out on the computer. The extent to which these match measures the completeness of the paradigm of molecular biology.”

– Harold Morowitz 1984

A short history of whole-cell modeling





"Whole-cell" model of E. coli

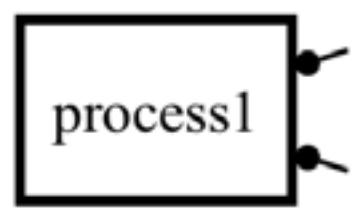
- Combines a massive, heterogeneous set of measurements reported in *E. coli* in thousands of studies across hundred of laboratories over the past decades. >19,000 parameter values curated from this set.
- Linking these data via mechanistic models provides the most natural interpretation of the integrated dataset. >10,000 equations.
- While all genes are expressed in the model, only 1214 of them are given a function (43% of annotated genes in EcoCyc). Mostly metabolic genes.
- Simulated in three environments: minimal medium (M9 salts plus glucose), rich medium (with added amino acids), and a minimal anaerobic medium.

Macklin, D.N., Ahn-Horst, T.A., Choi, H., Ruggero, N.A., Carrera, J., Mason, J.C., Agmon, E., ... & Covert, M.W. (2020). Simultaneous cross-evaluation of heterogeneous E. coli datasets via mechanistic simulation. *Science*, *369*(6502)

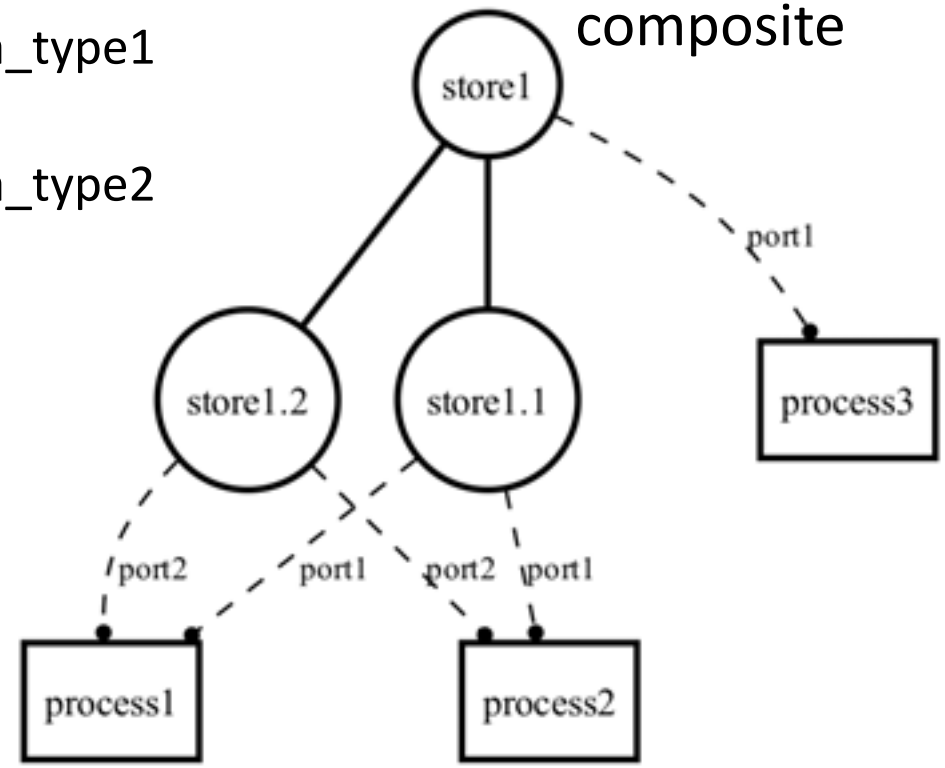
Can we leverage modular software design to integrate heterogeneous data types and models of cellular/molecular functions?

Vivarium: an "interface protocol" for connecting heterogeneous models, algorithms, and data into a hierarchical network that represents distributed, interacting processes.

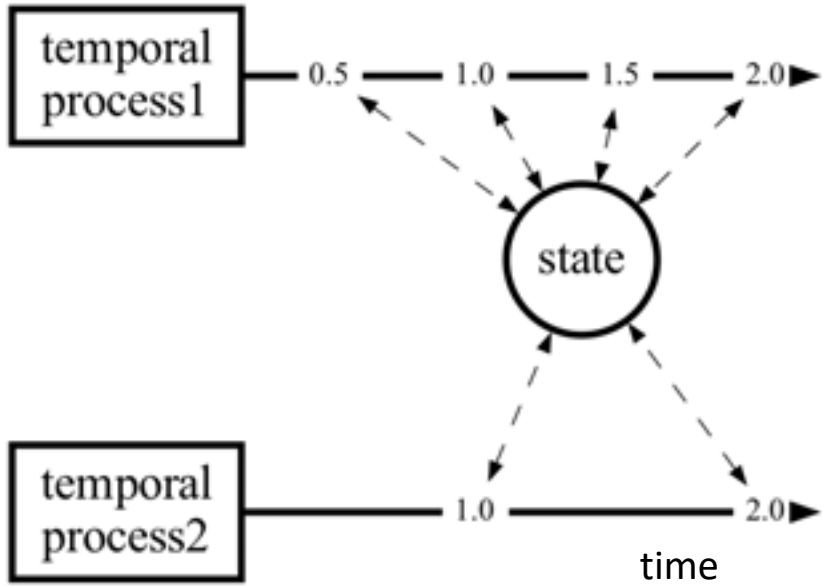
- **Processes:** consist of parameters, ports, and an update function.
- **Stores:** hold the state variables, map the variable names to their values, and apply the updates.
- **Composites:** bundles of processes and stores, wired together by their ports, and run together in time.



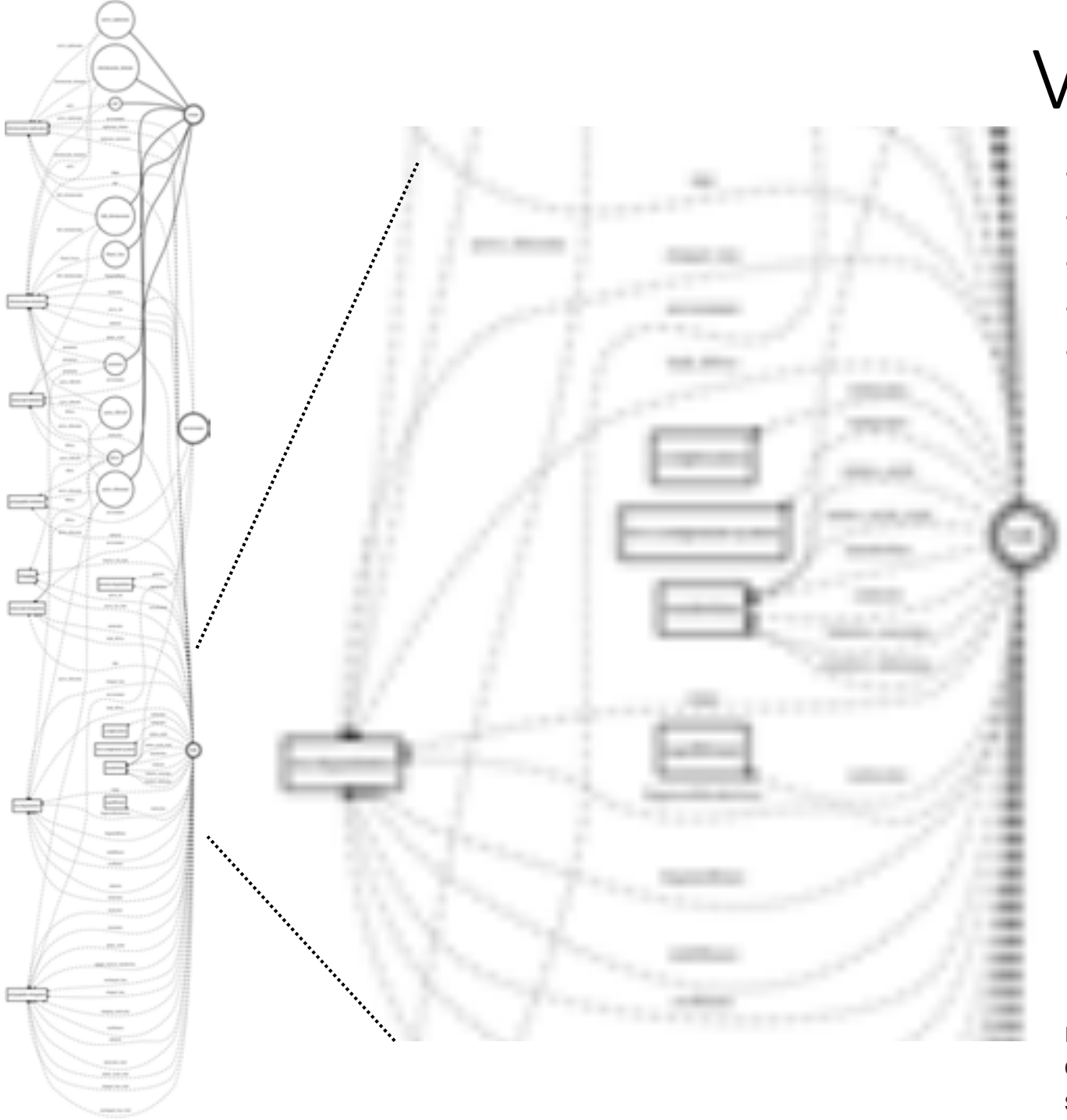
composite



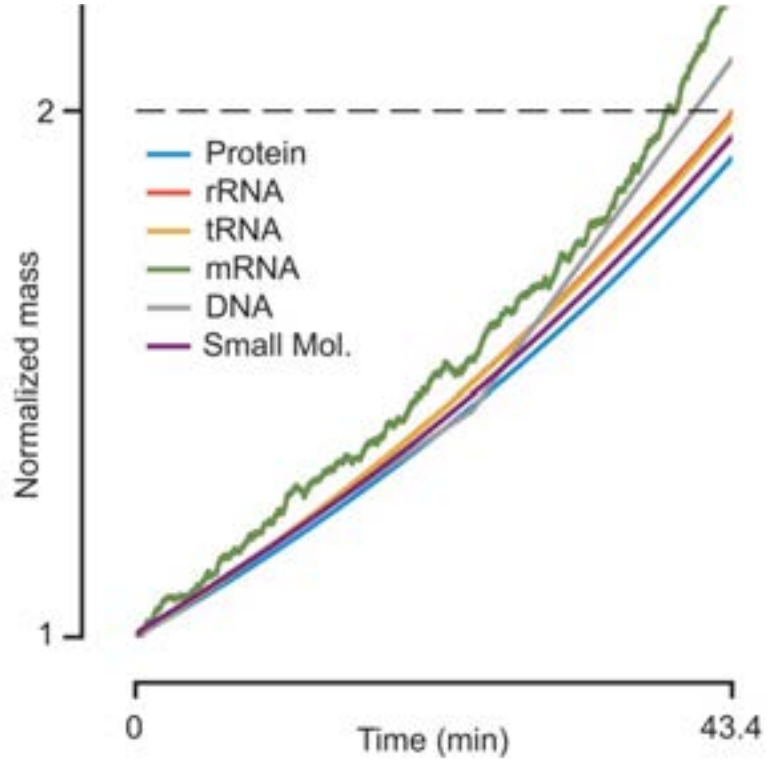
orchestration



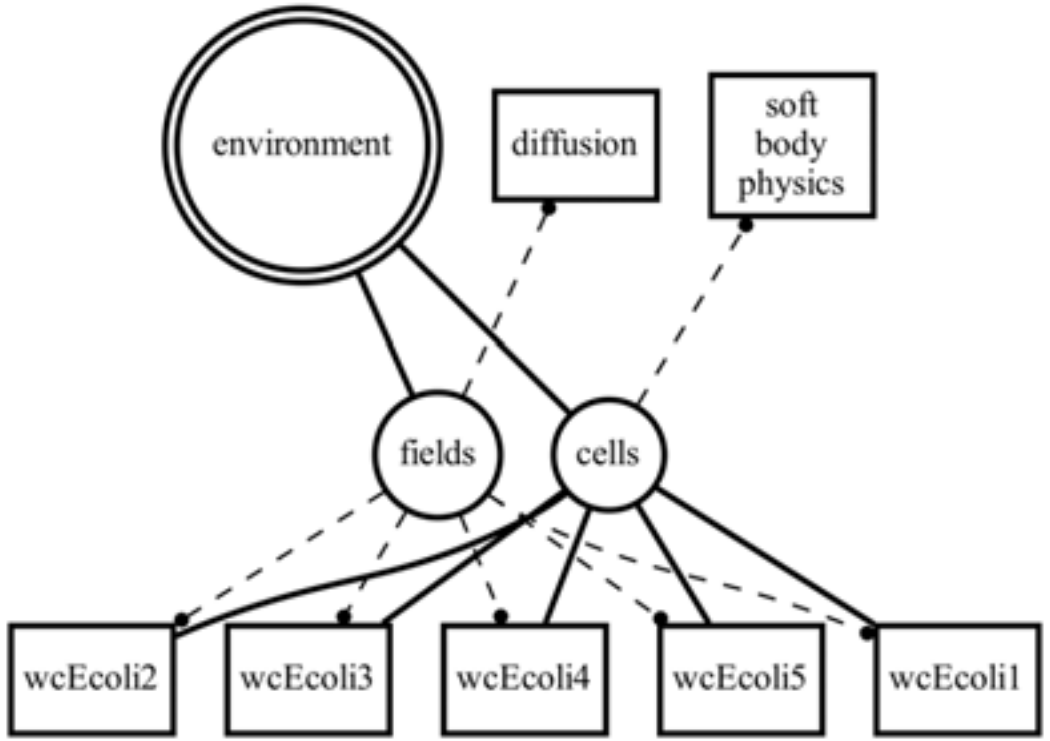
Vivarium-Ecoli

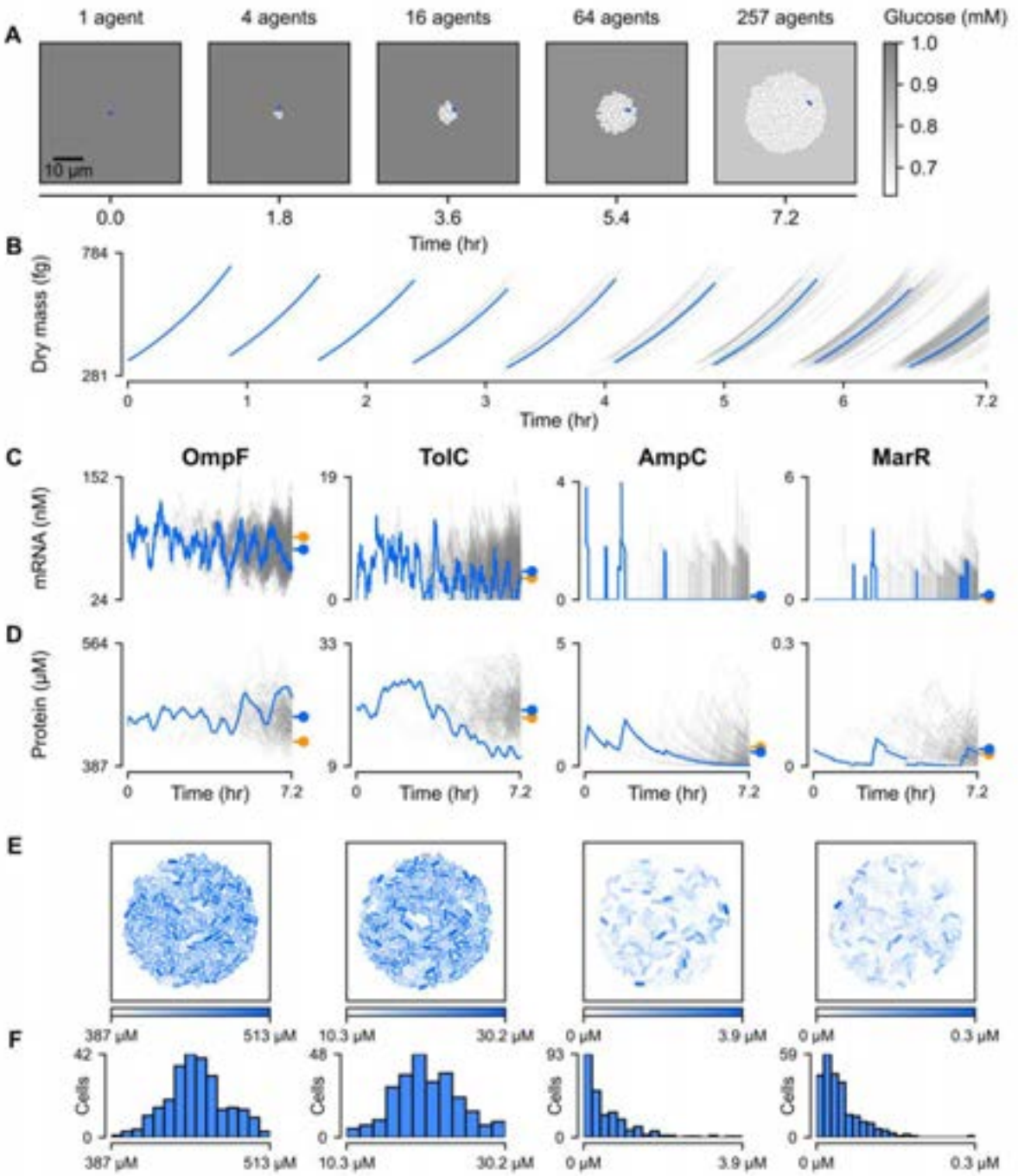


- Re-created as 12 composable processes
- functions for 1214 (43%) of well-characterized genes
- >19,000 parameter values
- >10,000 mathematical equations
- https://github.com/CovertLab/vivarium-ecoli



reproduces model from Macklin, et al. "Simultaneous cross-evaluation of heterogeneous E. coli datasets via mechanistic simulation." *Science* (2020)

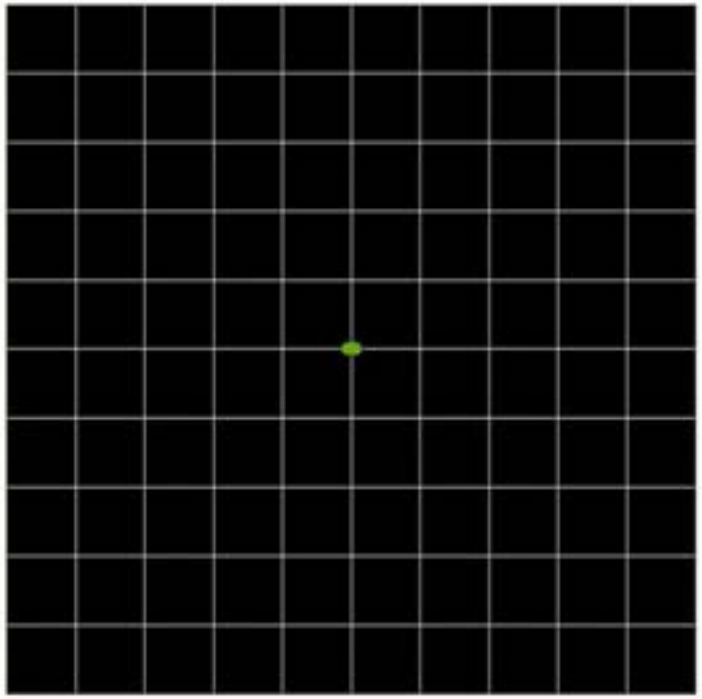




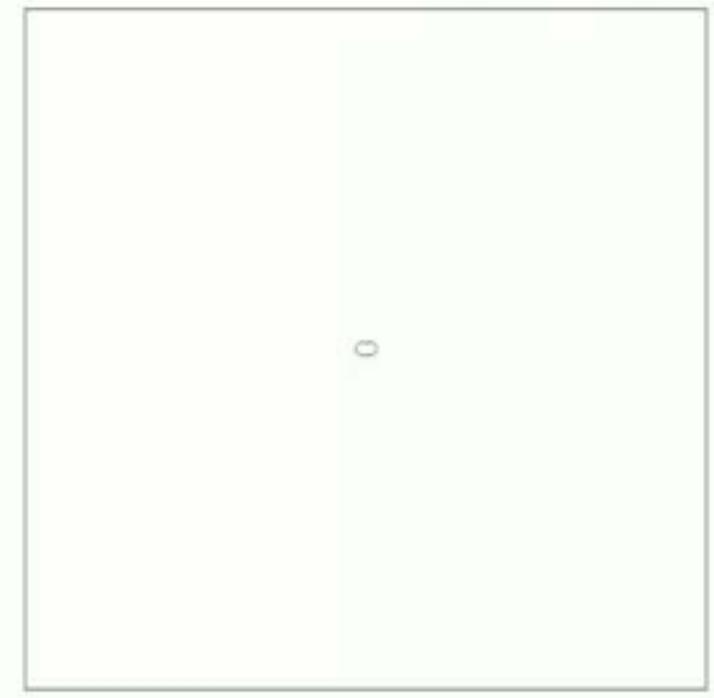
Skalnik, Cheah, Yang, et al. Whole-cell modeling of *E. coli* colonies enables quantification of single-cell heterogeneity in antibiotic responses. *PLoS Computational Biology.* (2023)

heterogeneous gene expression

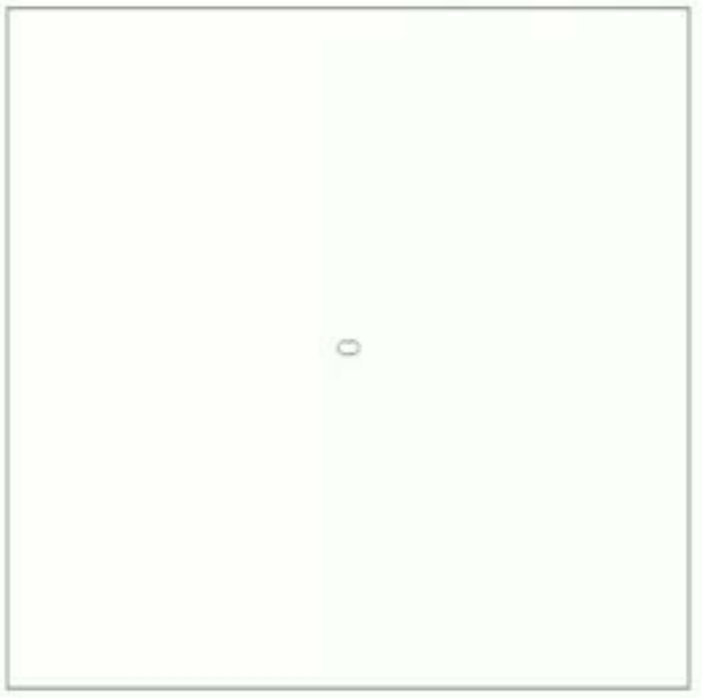
growth on glucose



AmpC

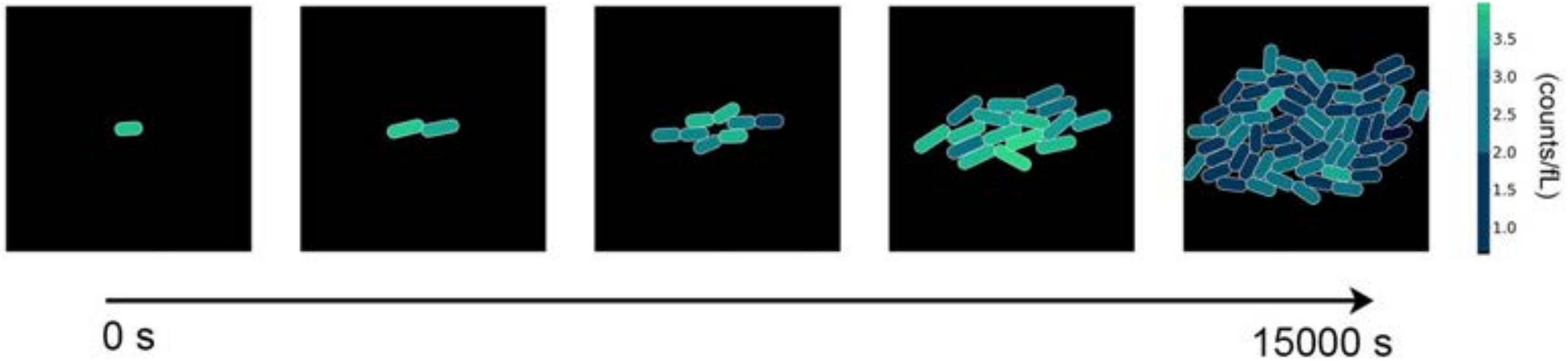


AcrAB-TolC



Skalnik, Cheah, Yang, et al. Whole-cell modeling of *E. coli* colonies enables quantification of single-cell heterogeneity in antibiotic responses. *PLoS Computational Biology.* (2023)

Adding function:
from flagella expression to behavior

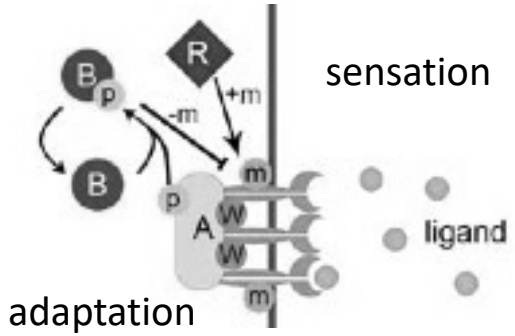


Flagella

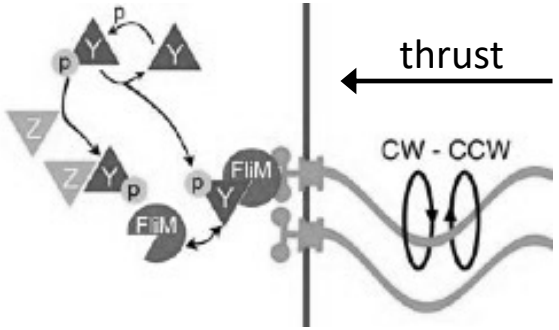
Agmon, E., & Spangler, R. K. (2020). A multi-scale approach to modeling E. coli chemotaxis. *Entropy*.

Adding function: from flagella expression to behavior

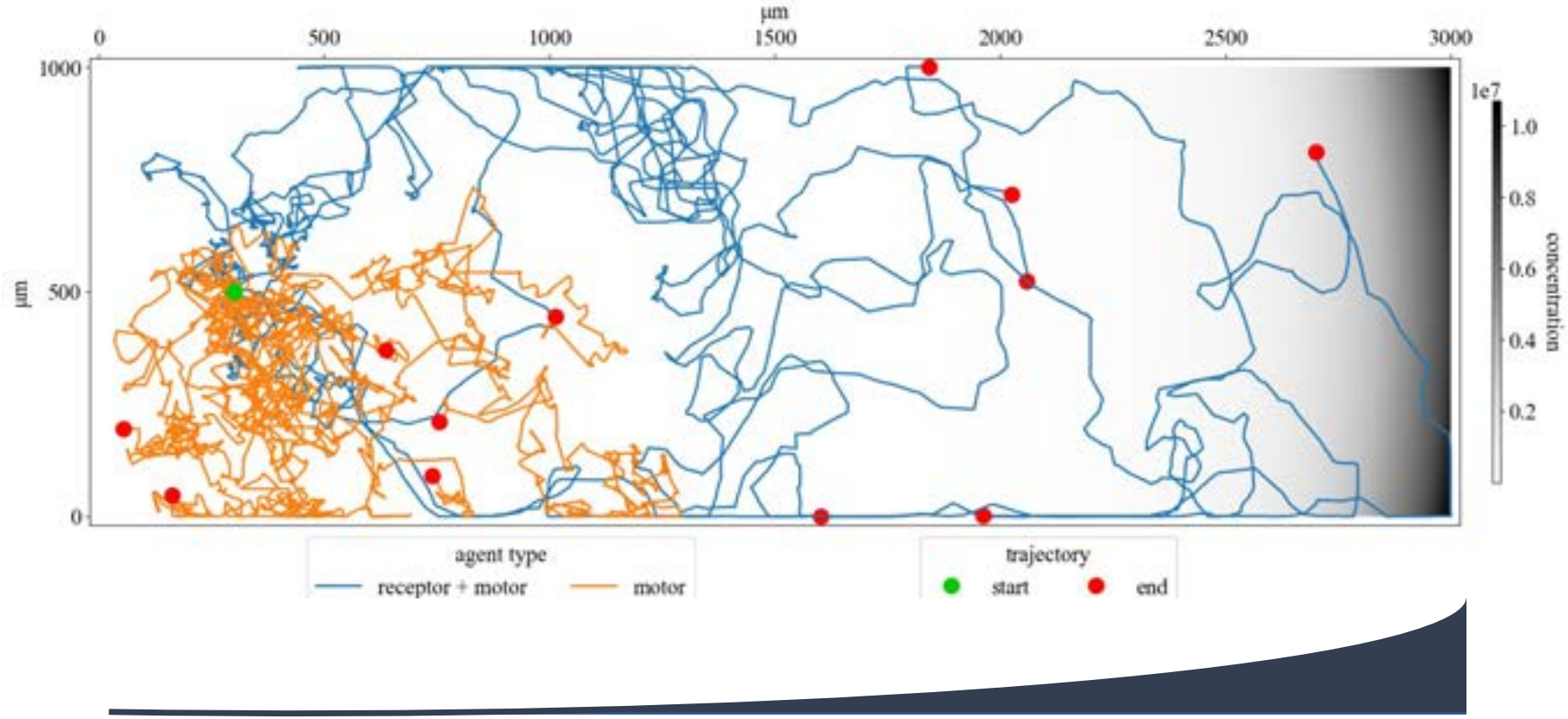
Chemoreceptors (Monod-Wyman-Changeux model)



Flagella activity (stochastic switching model)



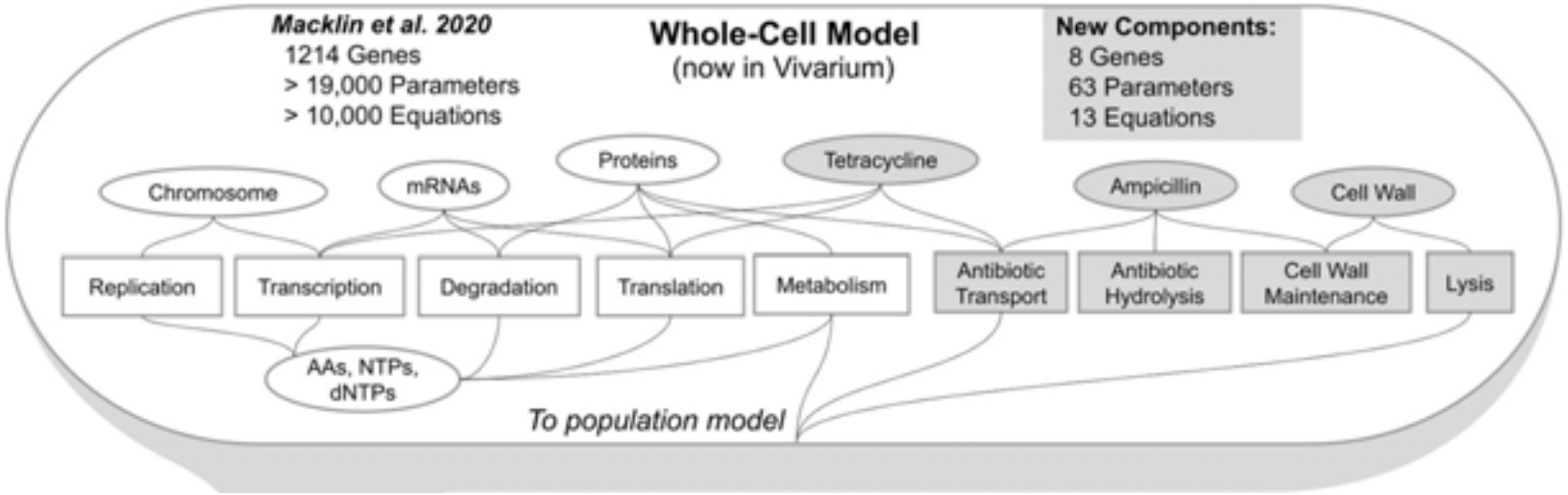
Motility and chemotaxis in a small population of E. coli WCMs



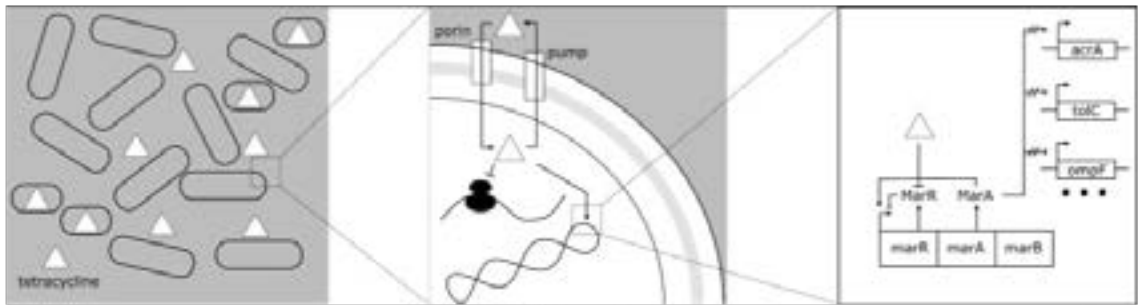
ligand gradient

Agmon, E., & Spangler, R. K. (2020). A multi-scale approach to modeling E. coli chemotaxis. *Entropy*.

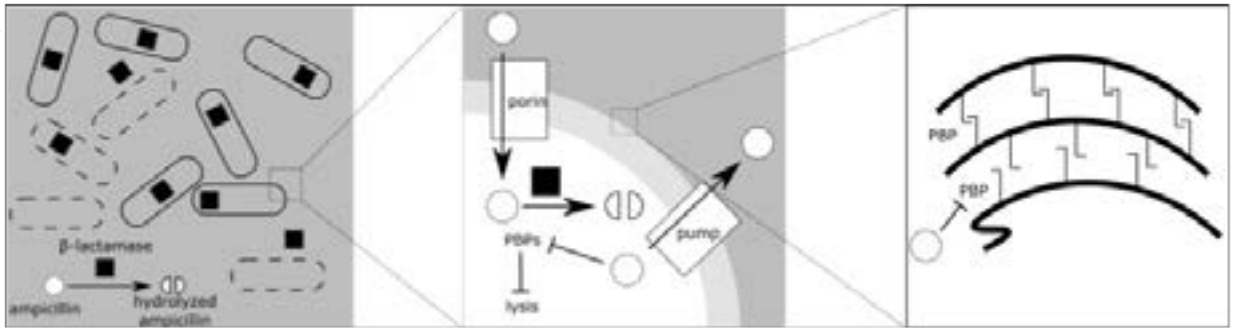
Adding function: response to antibiotics



Response to Tetracycline

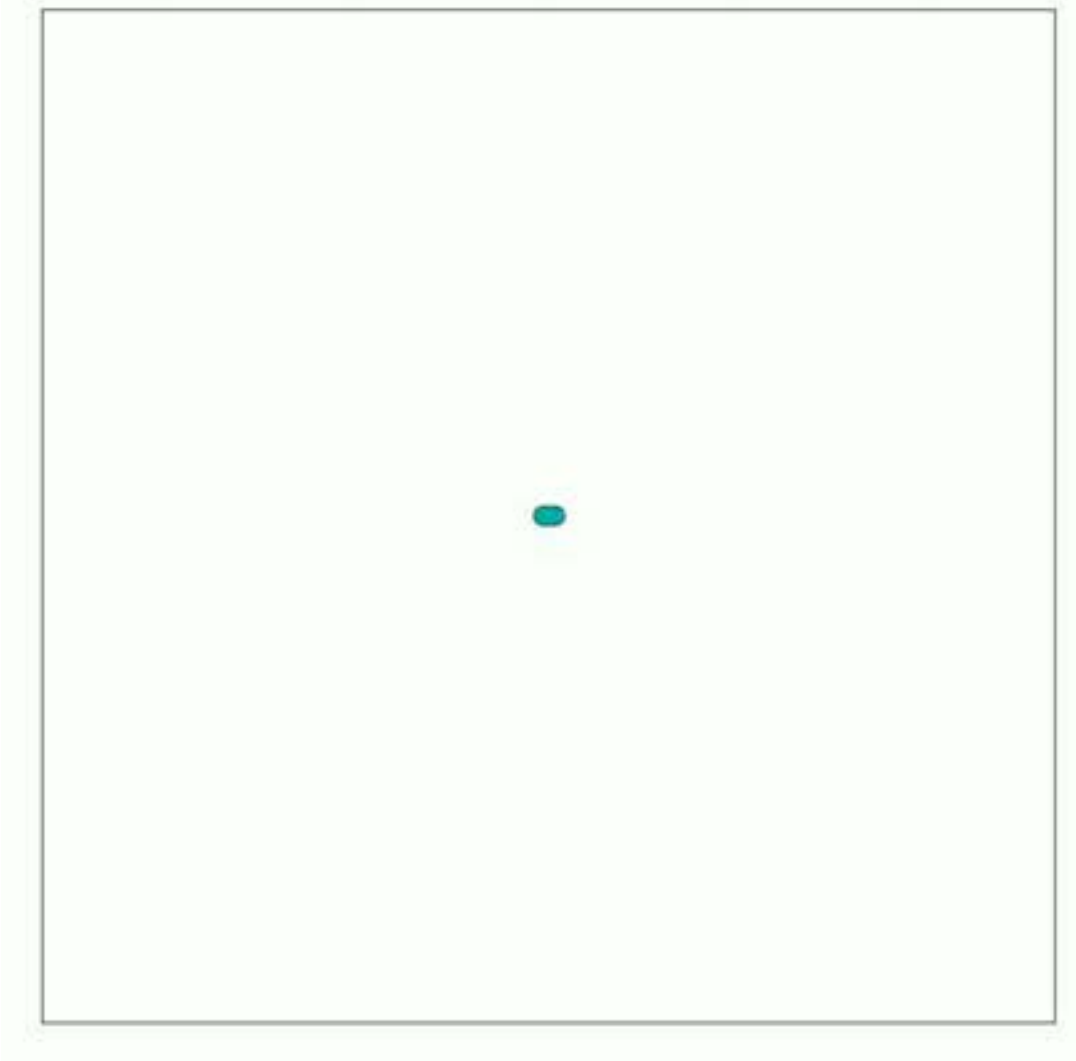


Response to Ampicillin



Skalnik, Cheah, Yang, et al. Whole-cell modeling of *E. coli* colonies enables quantification of single-cell heterogeneity in antibiotic responses. *PLoS Computational Biology.* (2023)

Simulating colony response to antibiotics



Skalnik, Cheah, Yang, et al. Whole-cell modeling of *E. coli* colonies enables quantification of single-cell heterogeneity in antibiotic responses. *PLoS Computational Biology.* (2023)

What needs to happen next

To represent all the molecular processes in a cell we need to integrate heterogeneous data. Ideally, each of the following can be experimentally determined, but may require inference algorithms to fill missing knowledge:

- sequence of each chromosome, RNA, and protein; the location of each chromosomal feature including each gene, operon, promoter, and terminator; and the location of each site on each RNA and protein.
- structure of each molecule, the domains and sites of macromolecules, and the subunit composition of complexes.
- subcellular organization of cells into organelles and microdomains.
- participants and effect of each molecular interaction, including the molecules that are consumed, produced, and transported, the molecular sites that are modified, and the bonds that are broken and formed.
- kinetic parameters of each interaction.
- concentration of each species in each organelle and microdomain.
- concentration of each species in the extracellular environment.

To connect a cell's molecular composition with its behavior and function, we need:

- function/process curation pipelines, expanding upon the processes developed for vivarium-ecoli. This include modules for metabolism, TF binding, transcription, translation, chromosome replication, degradation, signal transduction, and more are required.
- whole-cell models made of these processes need to be calibrated with molecular data acquired across heterogeneous cell populations, in different environments, and with different experimental perturbations.

Thank You!

Vivarium-Lab: Amin Boroomand (UConn Health, WHOI), Isha Mendiratta (UConn Storrs), Edwin Appiah (UConn Health), Jayde Schlesener (UConn Health, WHOI) **Vivarium-Core:** Ryan Spangler (Altos Labs), Chris Skalnik (MIT), William Poole (Altos Labs), Jerry Morrison (Stanford), Shayn Peirce-Cottler (UVA), Markus Covert (Stanford). **Vivarium-Ecoli:** Chris Skalnik (Stanford), Michael Yang (Stanford), Sean Cheah (Stanford), Matt Wolff (Stanford). **BioSimulators:** Jonathan Karr (Formic Labs), Ion Moraru (UConn Health), Alex Patrie (UConn Health), Logan Drescher (UConn Health), Jim Schaff (UConn Health), Herbert Sauro (University of Washington). **Vivarium-Mechanobiology:** Blair Lyons (Allen Institute for Cell Science), Jessica Yu (Allen Institute for Cell Science), Saurabh Mogre (Allen Institute for Cell Science), Karthik Vegesna (Allen Institute for Cell Science), Matt Akamatsu (University of Washington)

Contact: agmon@uchc.edu

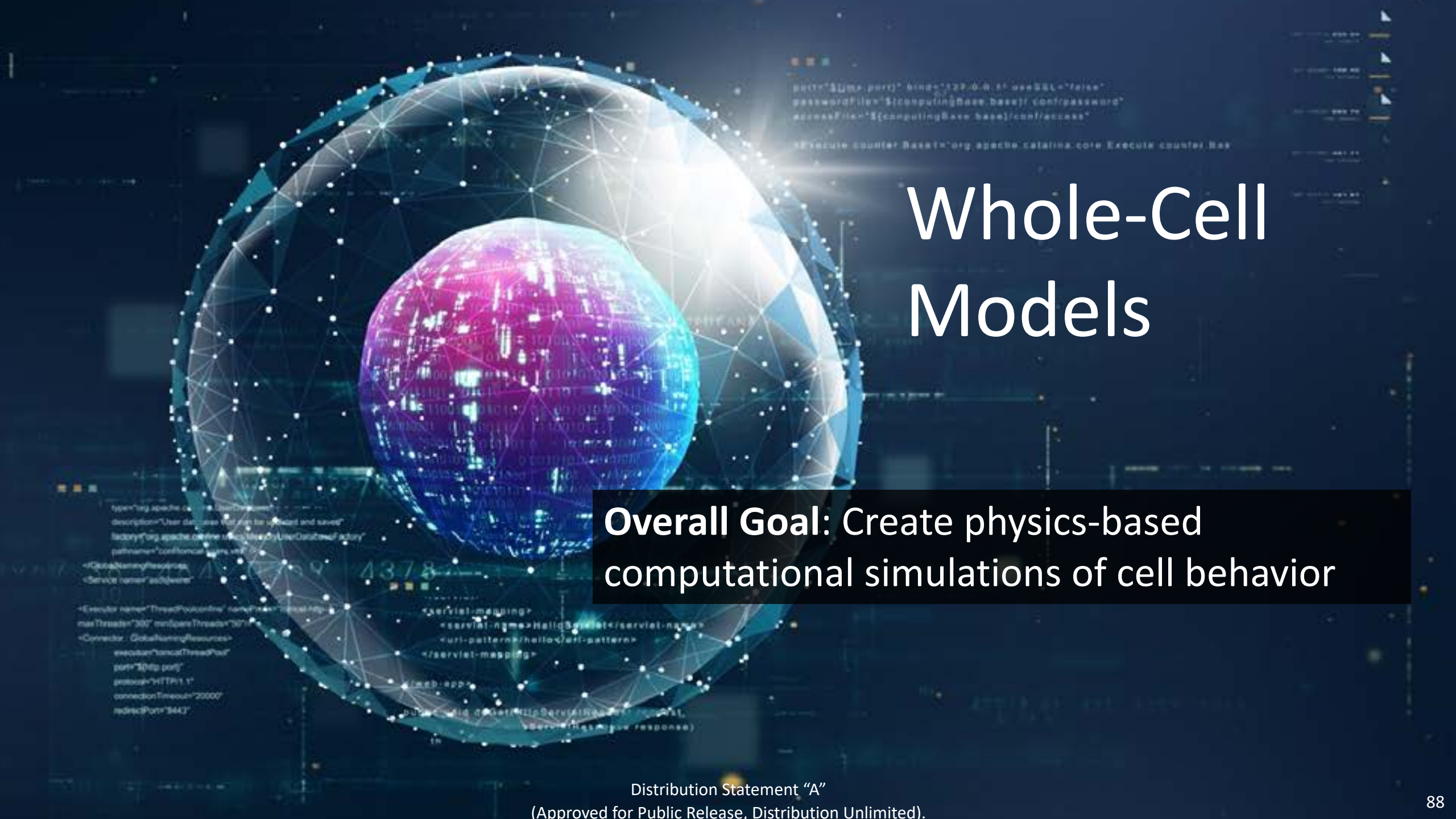
Unknown Protein Function in Whole-Cell Modeling

Christopher J. Bettinger, Ph.D.
Biological Technologies Office (BTO)

Discovering Unknome Function (DUF)

12 Dec 2023

Distribution Statement “A”
(Approved for Public Release, Distribution Unlimited).



Whole-Cell Models

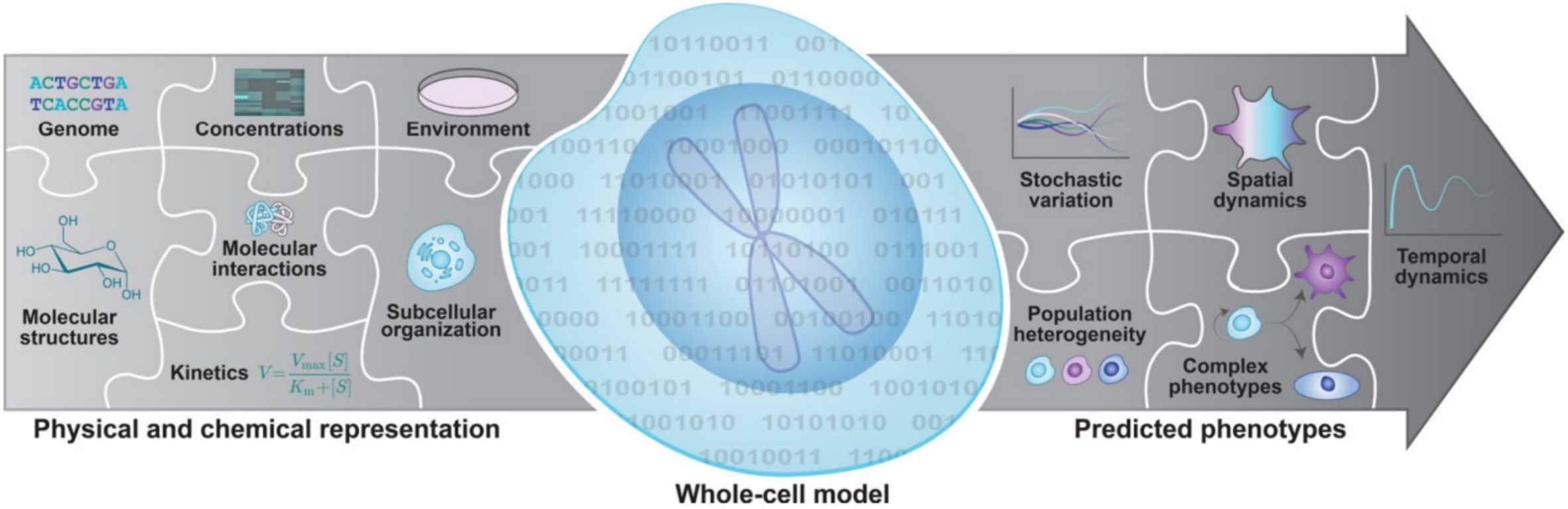
Overall Goal: Create physics-based computational simulations of cell behavior

Distribution Statement “A”
(Approved for Public Release, Distribution Unlimited).

What is Whole-Cell Modeling?

***WCM** – Whole-cell model*

WCM: A computer simulation that predicts phenotypes from genotype, including all molecular species and each molecular interaction.



Impact: A practical whole-cell model uses genotype to (a) predict disease, (b) anticipate pathogenicity, (c) accelerate design-build-test-learn cycles in synthetic biology.

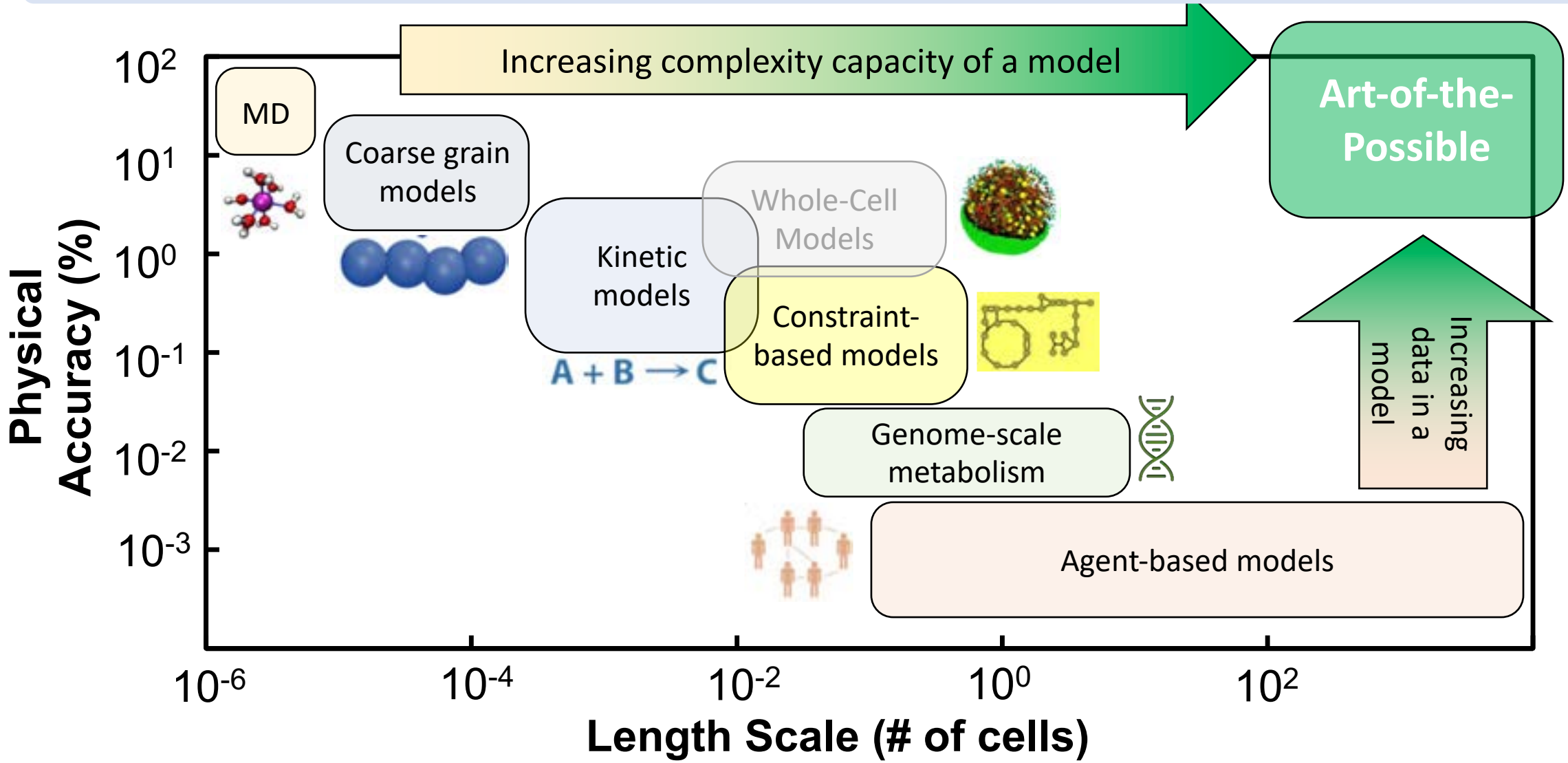
Goldberg, et al. *Curr Opin Biotechnol*. 2018

Distribution Statement "A"
(Approved for Public Release, Distribution Unlimited).

Overall landscape of models to describe cell behavior

***MD** – Molecular dynamics*

SoA: Models of cells are either: (a) physically accurate; (b) scalable, **but not both!**



Whole-cell modeling provides a useful conceptual frame

Whole-cell modeling SoA: Solve large systems of ODE across many cell "modules" to gain a comprehensive chemical-physical representation of the cell.

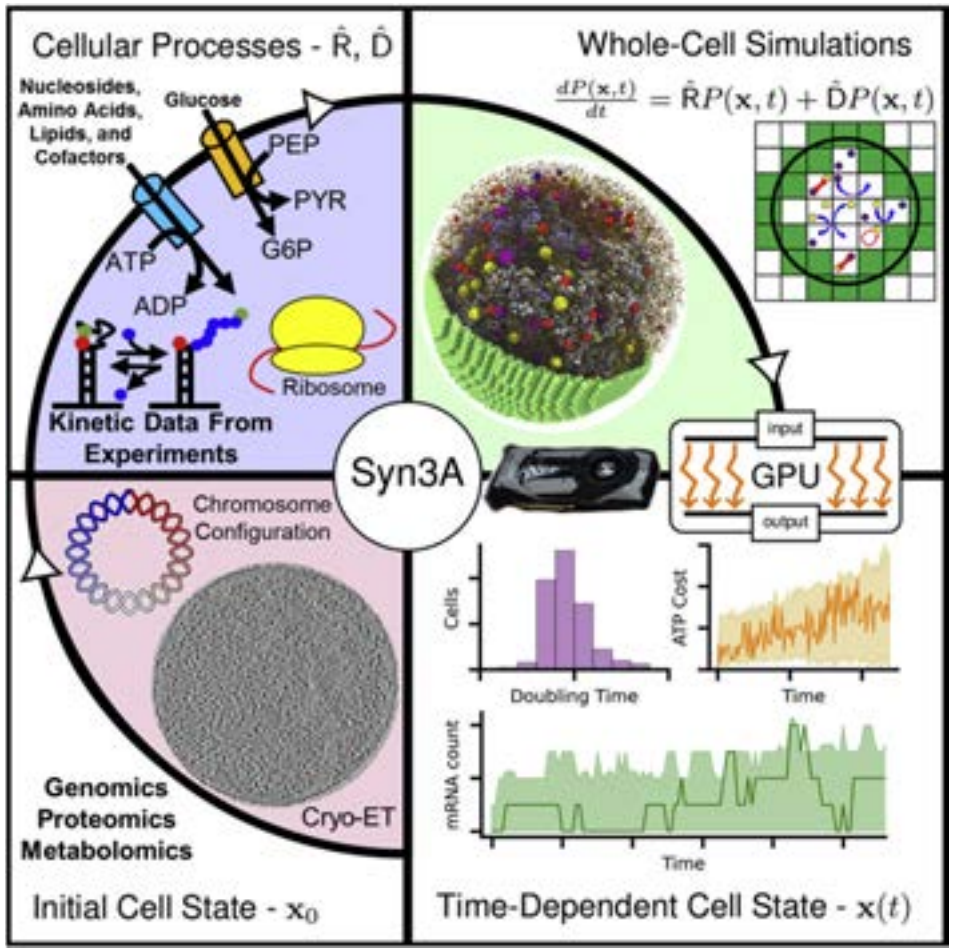
WCM Tools

1. Cryo-electron microscopy to image proteins
2. Experimental –omics data
3. High-performance computing

WCM Demo

WCM simulate doubling time & metabolism for **one division of a synthetic cell**[a]

Capability Gap: Unable to handle large complexity
Knowledge Gap: Unannotated proteins & sparse data



Thornberg, et al. *Cell*. 2022

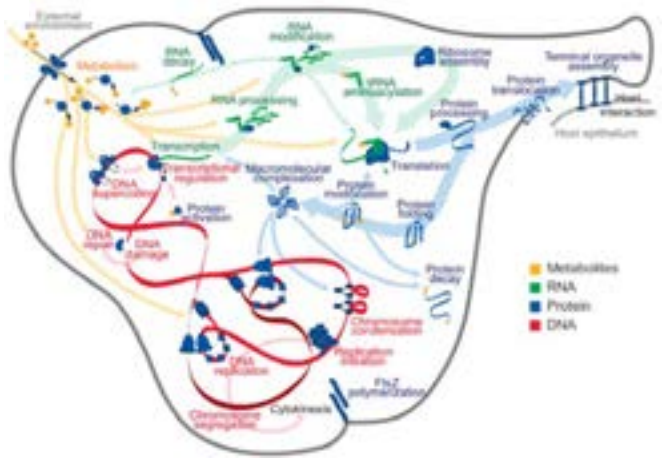
[a]Synthetic cell composed of 493 genes (543 kbp genome)

Distribution Statement "A"
(Approved for Public Release, Distribution Unlimited).

What are "DARPA"-hard roadblocks to biological modeling of cells?

Impact: An interpretable physics-based model of *E. coli* could predict evolution and accelerate synthetic biology research but there are challenges…

Complexity



493 genes (mycoplasma)

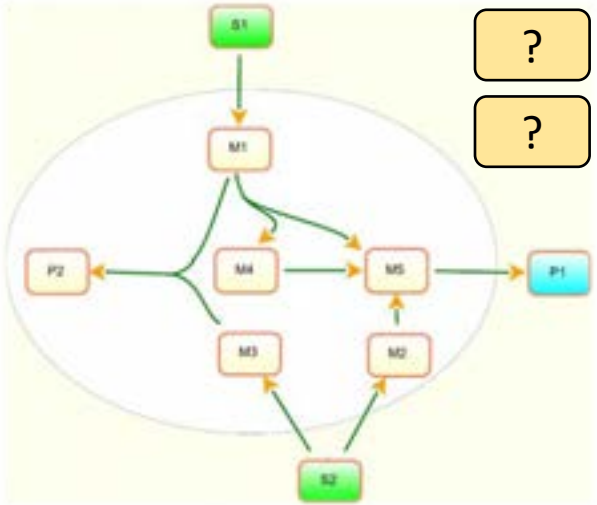
10X

4000 genes (*E. coli*)

Karr, et al. *Cell*. 2012

Sparse Data

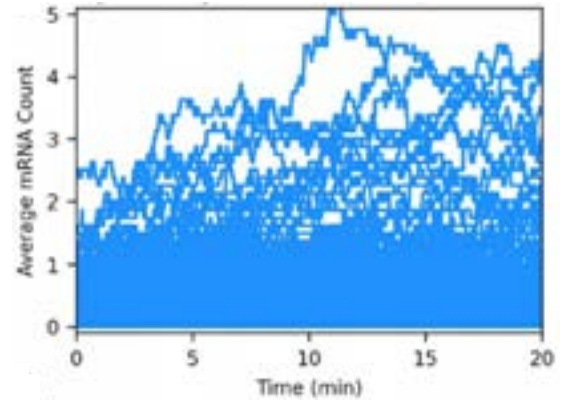
~35% of proteins in *E. coli* have unknown function

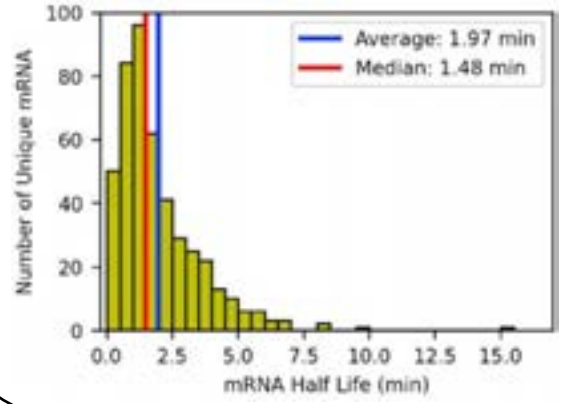


Metabolic network

Noise

RNA expression in 8 runs



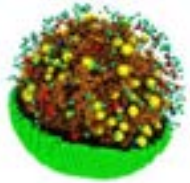


Thornberg, et al. *Cell*. 2022

Distribution Statement "A"
(Approved for Public Release, Distribution Unlimited).

Leveraging modeling and automated "cloud labs" for WCM

Model & Predict

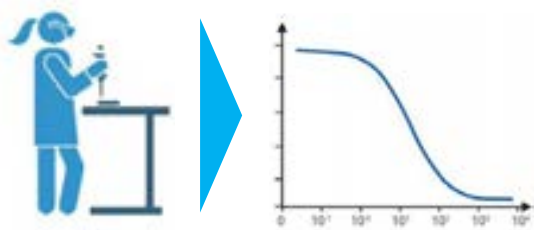


Develop WCMs for complex organisms & multi-organism communities

Technical Challenges

- **Complexity**: Solve high-dimensional systems of equations; Incorporating features to model cell-cell interactions
- **Sparse Data**: Getting values for initial state and parameters
- **Noise**: Models need to be robust and tolerate noise

Measure & Verify



Collect ground-truth data to inform & validate models

Technical Challenges

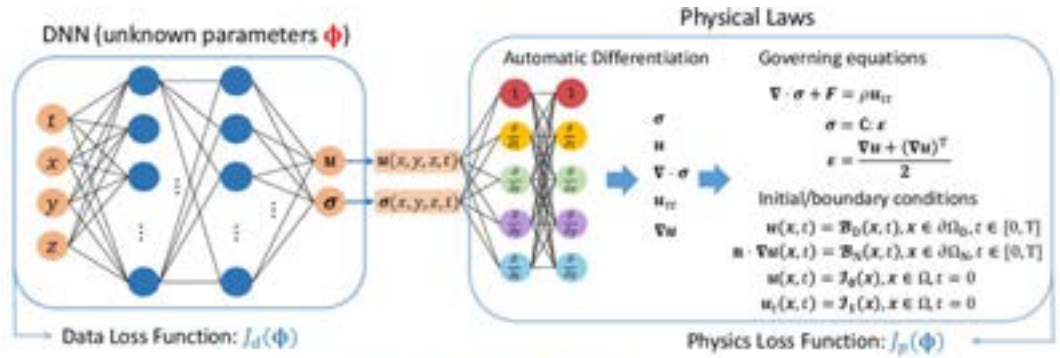
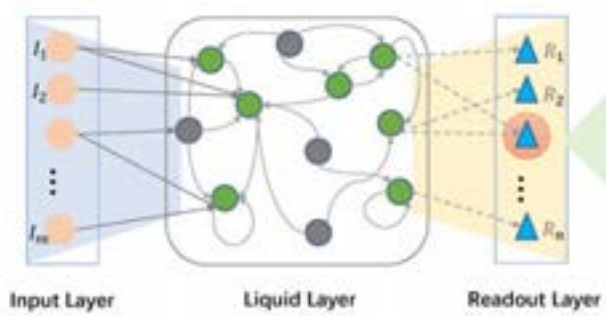
- Automating large-scale experiments
- Handling of noisy and stochastic data from small sample sizes
- Human "out-of-the-loop" experimentation
- Being able to handle and curate heterogeneous data

Experiments and modeling exercises run concurrently – models & data help inform each other.
Interest: WCM software that can predict the behavior of microbial communities.

What are the computational tools available for WCM?

***PDE** – Partial differential equations*
***NN** – Neural network*

Recent innovations in neural networks allow: (a) handling of sparse datasets; (b) descriptions of more complex systems with fewer "neurons"

Innovation	Impact	Implication for WCM
Physics-informed Neural Networks (2019) Raissi, et al. *J Comp Phys*. 2019	Can solve large systems of PDE with sparse data by using governing equations to connect "neurons" 	❖ Can solve high-dimensional systems of PDE using sparse data sets ❖ Ideal for modeling biological systems where data is limited
"Liquid" Neural Networks (2022) Hasani, et al. *Nat Art Intelligence*. 2012	Neural networks are simplified by connecting fewer "neurons" w/ **non-linear equations** (opposed to scalars) 	❖ Improves scaling **~10,000x** ❖ Standard: 100,000 "neurons" ❖ Liquid: 19 "neurons" ❖ Fewer "neurons" makes networks more interpretable

Distribution Statement "A"
(Approved for Public Release, Distribution Unlimited).

www.darpa.mil

Distribution Statement “A”
(Approved for Public Release, Distribution Unlimited).