

Modularization of bacterial transcriptomes leads to the annotation of genes of unknown molecular function to a cellular process

PLANETARY PROTECTION METAGENOMICS IN SPACEFLIGHT WORKSHOP

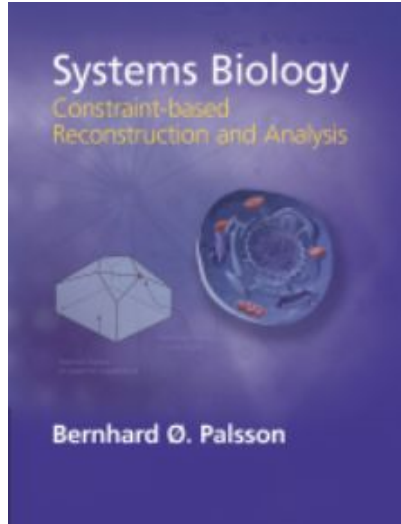
AMES, November 18, 2024

Bernhard Palsson, PhD
University of California, San Diego

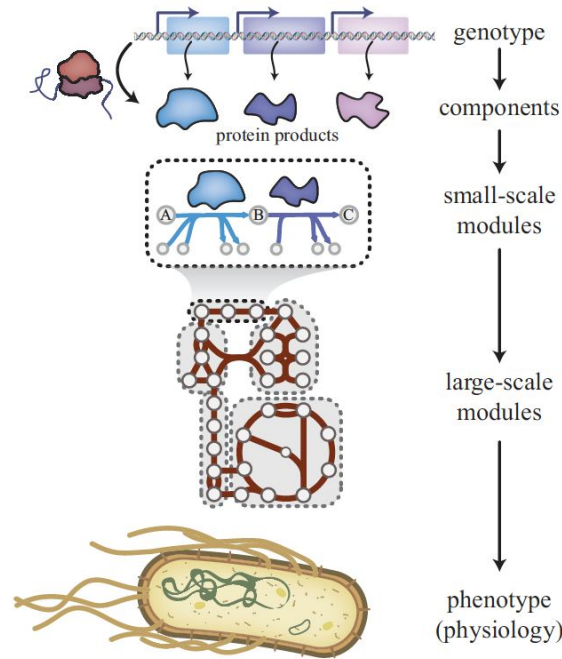
TWO TOPICS FOR DISCUSSION::

1. Modularization of the transcriptome
2. Applications
 - understanding, discovery, design

Levels 2&3: The need for modularization: From many variables to few

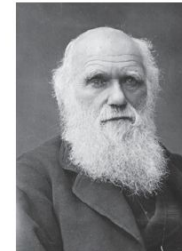


Biology has a
hierarchical character



Ludwig Boltzmann

Systems biology: emphasis on modules and understanding of how coherent physiological functions arise from the totality of molecular components



Charles Darwin

Chemical causation:
Can apply P/C laws
and get causality on
a small scale

Biological causation:
genome-scale changes
and description of 1000s
of variables. Network and
econometric type

Transcriptomics

(expression of all genes)



Anand Sastry



ARTICLE

<https://doi.org/10.1038/s41467-019-13483-w>

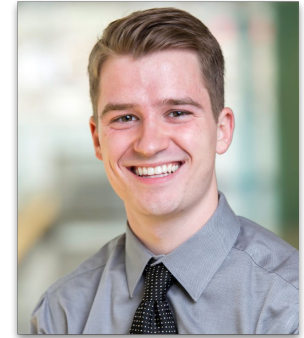
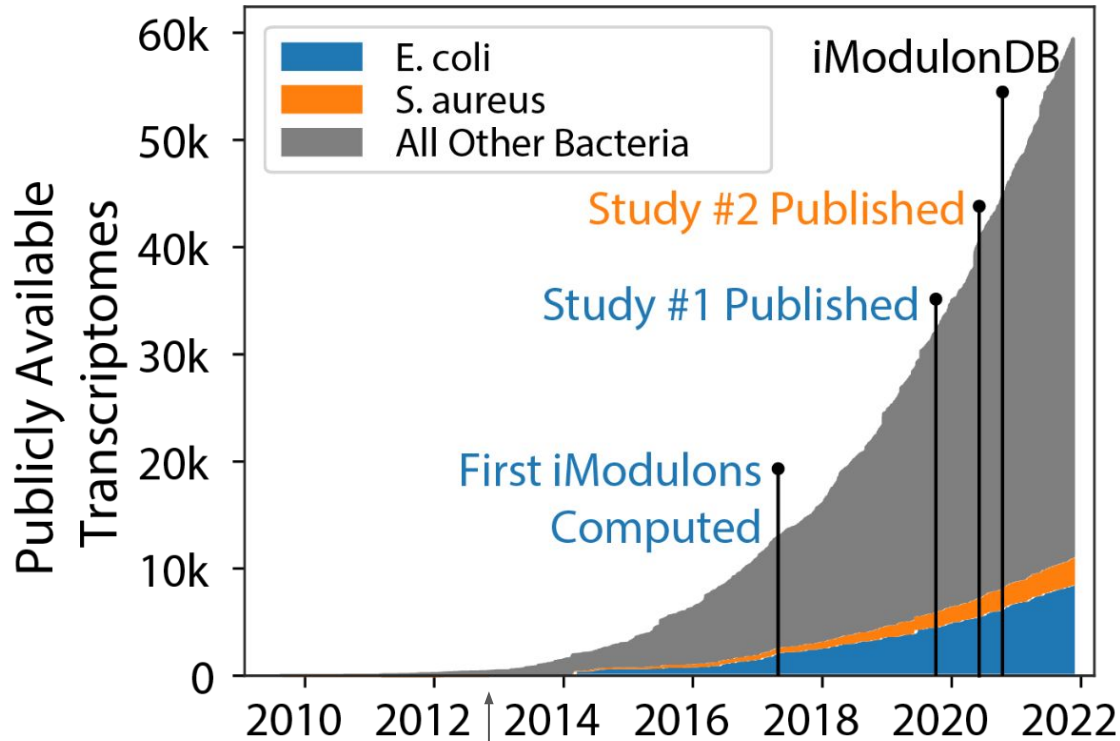
OPEN

The *Escherichia coli* transcriptome mostly consists of independently regulated modules

Anand V. Sastry¹, Ye Gao², Richard Szubin¹, Ying Hefner¹, Sibe Xu¹, Donghyuk Kim^{1,5}, Kumari Sonal Choudhary¹, Laurence Yang^{1,6}, Zachary A. King¹ & Bernhard O. Palsson^{1,3,4*}

Underlying cellular responses is a transcriptional regulatory network (TRN) that modulates gene expression. A useful description of the TRN would decompose the transcriptome into targeted effects of individual transcriptional regulators. Here, we apply unsupervised machine learning to a diverse compendium of over 250 high-quality *Escherichia coli* RNA-seq datasets to identify 92 statistically independent signals that modulate the expression of specific gene sets. We show that 61 of these transcriptomic signals represent the effects of currently characterized transcriptional regulators. Condition-specific activation of signals is validated by exposure of *E. coli* to new environmental conditions. The resulting decomposition of the transcriptome provides: a mechanistic, systems-level, network-based explanation of responses to environmental and genetic perturbations; a guide to gene and regulator function discovery; and a basis for characterizing transcriptomic differences in multiple strains. Taken together, our results show that signal summation describes the composition of a model prokaryotic transcriptome.

Growth of the # of transcriptomic data sets



Kevin Rychel

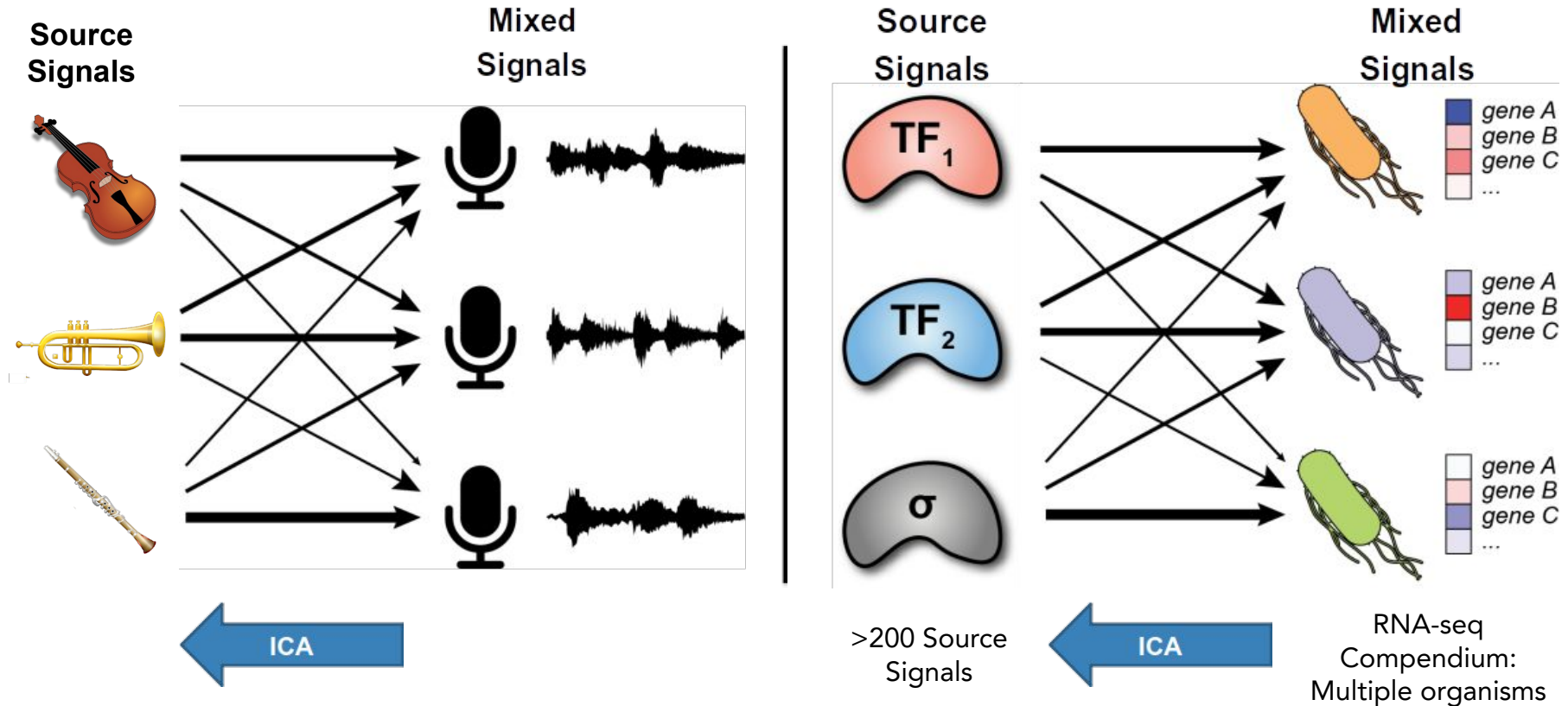
mRNA purification methods ready for deployment

The source signal identification problem:

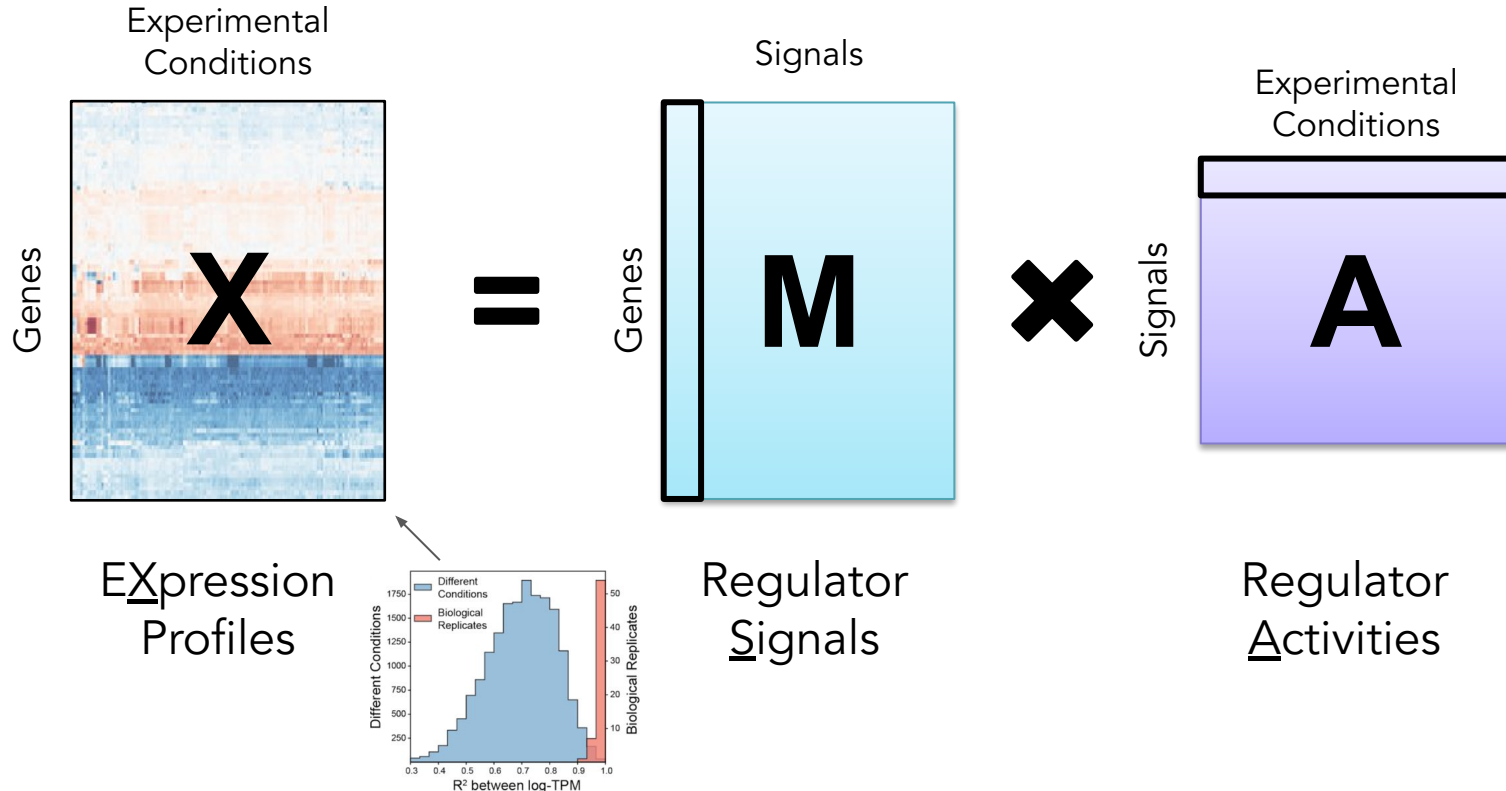
Solved with algorithms like independent component analysis (ICA)



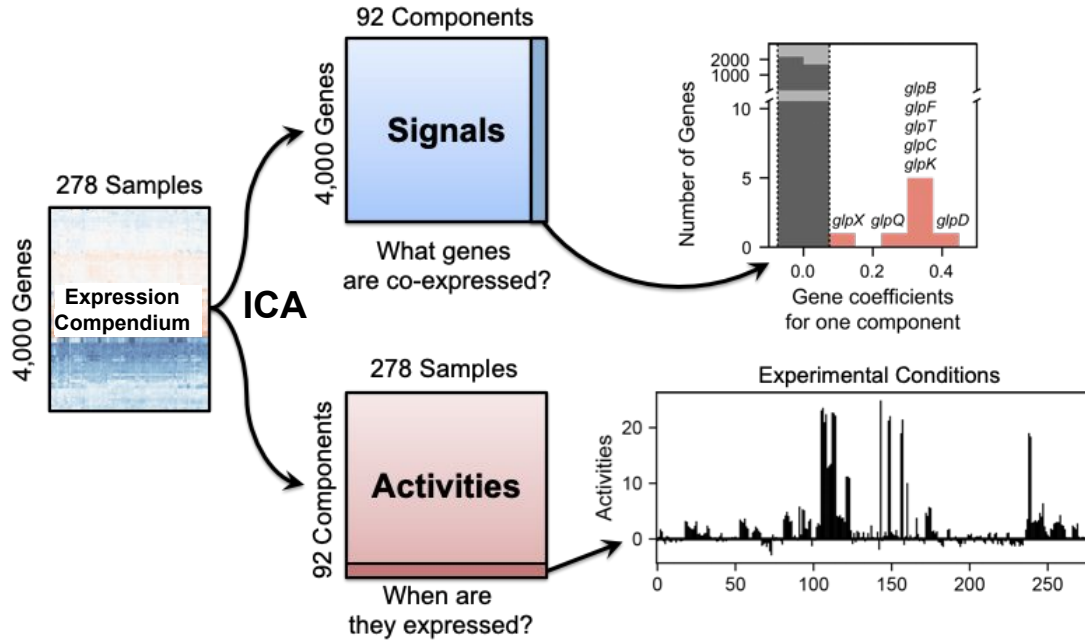
Independent Component Analysis (ICA)



Mathematical Representation



What is an iModulon?

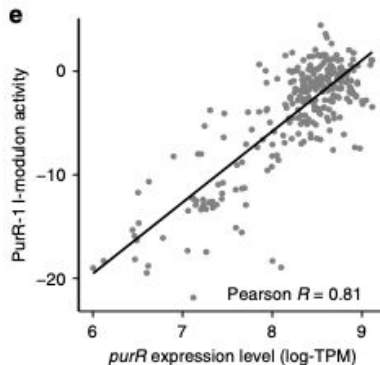
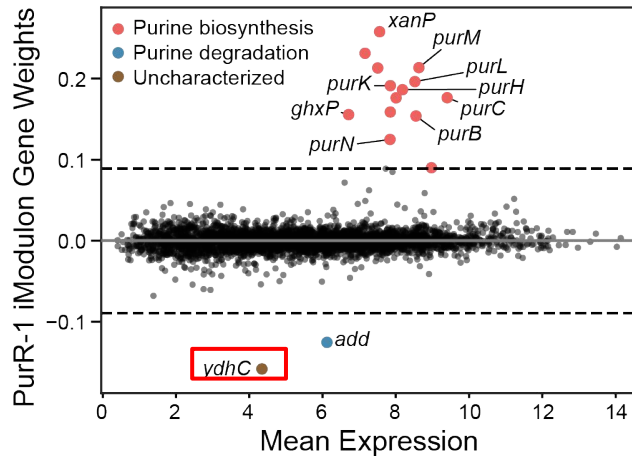


iModulon = independently Modulated group of genes.

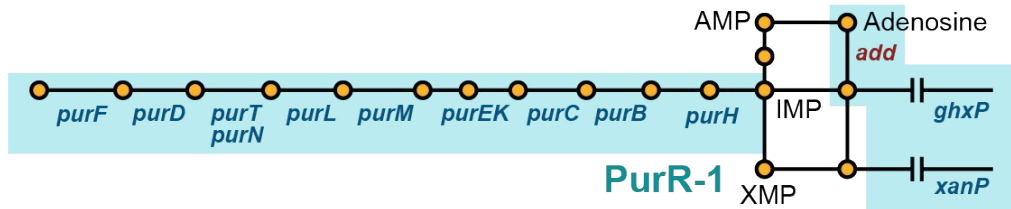
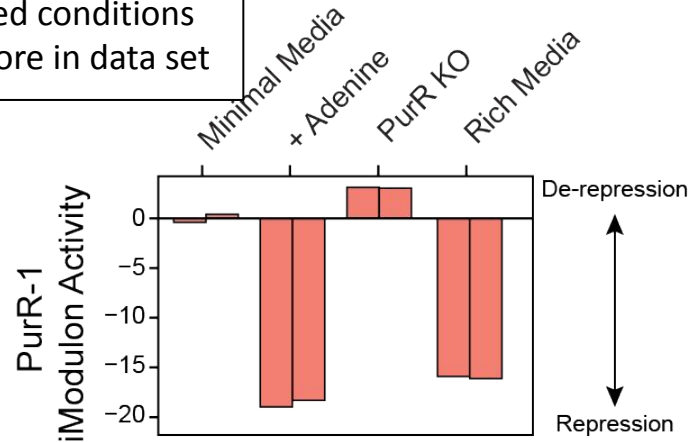
Knowledge-enrichment: every iModulon has a story

Represents an expression "signal," e.g. associated with a transcription factor)

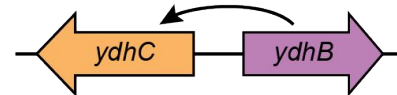
iModulons in action: The *E. coli* PurR iModulon



Selected conditions
100s more in data set



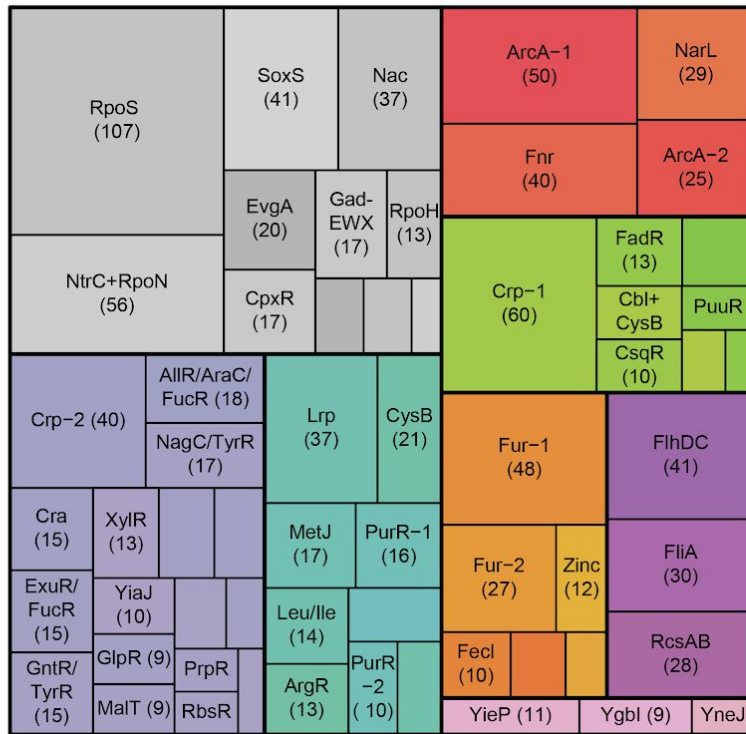
Discovery of a novel
adenosine transporter



Comms biol., 4:991 (2021).

Adapted from Nat Commun. 2019 Dec 4;10(1):5536.

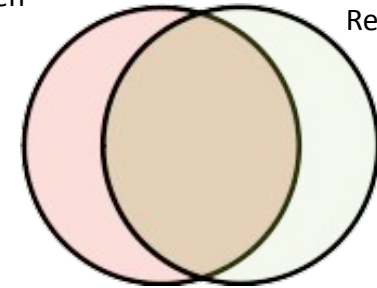
Each iModulon has a story



92 iModulons

- Stress Response (11 i-modulons)
- Carbon Source Utilization (17 i-modulons)
- Amino Acid and Nucleotide Biosynthesis (9 i-modulons)
- Energy Metabolism (4 i-modulons)
- Structural Components (3 i-modulons)
- Metal Homeostasis (6 i-modulons)
- Miscellaneous Metabolism (8 i-modulons)
- Regulator Discovery (3 i-modulons)

Data-driven iModulon



Regulon from literature

Metab. Eng. 2020

Adaptive evolution reveals a tradeoff between growth rate and oxidative stress during naphthoquinone-based aerobic respiration

Independent component analysis of *E. coli*'s transcriptome reveals the cellular processes that respond to heterologous gene expression

PNAS 2020

MBE 2020

OxyR Is a Convergent Target for Mutations Acquired during Adaptation to Oxidative Stress-Prone Metabolic States

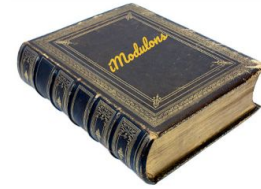
Microbiology 2019

Adaptive laboratory evolution of *Escherichia coli* under acid stress

Elucidation of Regulatory Modes for Five Two-Component Systems in *Escherichia coli* Reveals Novel Relationships

mSystems 2020

The iModulon Bibliome



1. Nat Commun. 2018 March 15; 9:1090. PMID [29545622](#)
2. Nat Commun. 2019 Dec 4;10(1):5536. PMID [31797920](#)
3. Proc Natl Acad Sci U S A. 2020 Jul 21;117(29):17228-17239. PMID [32616573](#)
4. mSystems. 2020 Nov 10;5(6):e00980-20. PMID [33172971](#)
5. Mol Biol Evol. 2020 Mar 1;37(3):660-667. PMID [31651953](#)
6. Metab Eng. 2020 Sep;61:360-368. PMID [32710928](#)
7. Nucleic Acids Res. 2021 Jan 8;49(D1):D112-D120. PMID [33045728](#)
8. PLoS Comput Biol. 2021 Feb 2;17(2):e1008647. PMID [33529205](#)
9. mSphere. 2021 Aug 25;6(4):e0044321. PMID [34431696](#)
10. Commun Biol. 2021 Aug 19;4(1):991. PMID [34413462](#)
11. Front Microbiol. 2021 Oct 27;12:753521. PMID [34777307](#)
12. BMC Bioinformatics. 2021 Dec 8;22(1):584. PMID [34879815](#)
13. Nucleic Acids Res. 2022 Apr 22;50(7):3658-3672. PMID [35357493](#)
14. Nucleic Acids Res. 2022 Sept 23;50(17):9675-9688. PMID [36095122](#)
15. mSystems. 2022 Dec 20;7(6):e0046722. PMID [36317888](#)
16. mSystems. 2022 Dec 20;7(6):e0048022. PMID [36321827](#)
17. mSphere. 2022 Apr 27;7(2):e0003322. PMID [35306876](#)
18. Metab Eng. 2022 Jul;72:297-310. PMID [35489688](#)
19. Proc Natl Acad Sci U S A. 2022 Jul 26;119(30). PMID [35858453](#)
20. PLoS Pathog. 2023 Jan 19;19(1):e1011064. PMID [36656902](#)
21. Cell Rep. 2023 Jun 27;42(6):112619. PMID [37285268](#)
22. mSystems. 2023 Jun 29;8(3):e0024723. PMID [37278526](#)
23. Int J Mol Sci. 2023 Aug 8;24(16):12544. PMID [37628727](#)
24. mSystems. 2023 Oct 26;8(5):e00437-23. PMID [37638727](#)
25. Nucleic Acids Res. 2023 Oct 27;51(19):10176-10193. PMID [37713610](#)
26. Nat Commun. 2023 Nov 22;14(1):7606. PMID [37993418](#)
27. Microbiol Res. 2023 Nov;276:127485. PMID [37683565](#)
28. PLoS Comput Biol. 2024 Jan 22;20(1):e1011824. PMID [38252668](#)
29. mSystems. 2024 Feb 7:e0094223. PMID [38323821](#)
30. <https://www.biorxiv.org/content/10.1101/2021.07.01.450581v1>
31. mSystems. 2024 Feb 13:e0125723. PMID [38349131](#)
32. Nat Commun. 2024 Mar 15;15(1):2356. PMID [38490991](#)
33. Metab Eng. 2024 May 31:S1096-7176(24)00073-9. PMID [38825177](#)
34. Nat Commun. 2024 Jun 19;15(1):5234. PMID [8898010](#)
35. Nucleic Acids Res. 2024 Aug 28:gkae742. doi: 10.1093/nar/gkae742. PMID [39193902](#)
36. Adv Sci (Weinh). 2024 Sep 12:e2403912. PMID [39264300](#)
37. Proc Natl Acad Sci U S A. 2024 Sep 17;121(38):e2410492121. PMID [39269777](#)
38. PLoS Comput Biol. 2024 Oct 23;20(10):e1012546. PMID [39441835](#).
39. PLoS Genet. 2024 Oct 29;20(10):e1011464. PMID [39471211](#)
40. Nucleic Acids Res. 2024 Nov 4:gkae1009. PMID [39494532](#)

iModulonDB.org: Dashboard 2.0 (has built-in analysis tools)

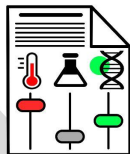
[Projects.html](#)



Eddie Catoiu

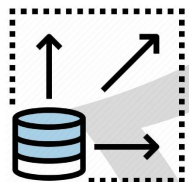
Curated Metadata

100011
010101
111000
011010

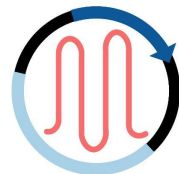


- + Genetic perturbations
- + Experimental conditions
- + Publication text & info

Database Expansion



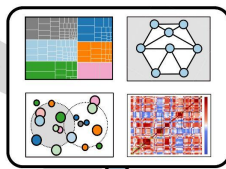
- + 12 organisms
- + 9,000 expression profiles



iModulonDB 2.0

Updated dashboards

Data Visualization

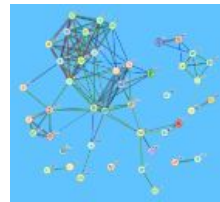


- + New graphs
- + External databases
- + Biological context

Immediate understanding of each experimental condition

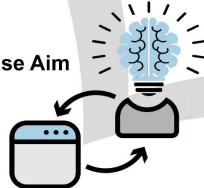
Explained variance, PPI, operon diagrams, regulon recall, gene expression correlations

Fnr-1 PPI



e.g. STRING

Knowledgebase Aim



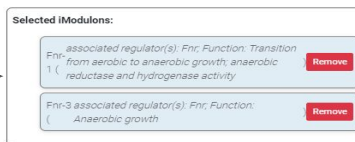
Simplify access to and understanding of transcriptomic datasets

In-page Analysis Tools

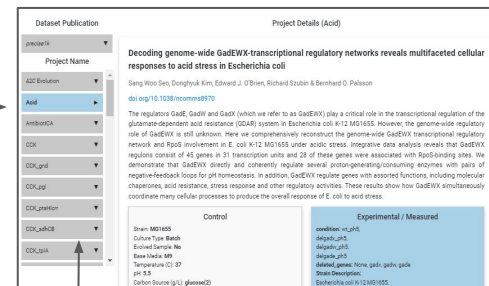


- + Search-based
- + Species-wide
- + Promote user-engagement

Search.html



Any pair of iModulons and/or genes

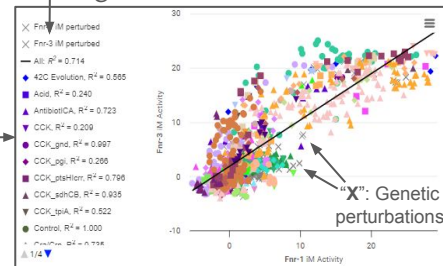


Abstract, experimental & control variables

Conditions

e.g. iModulon Phase Plane

Fnr-1 vs Fnr-3



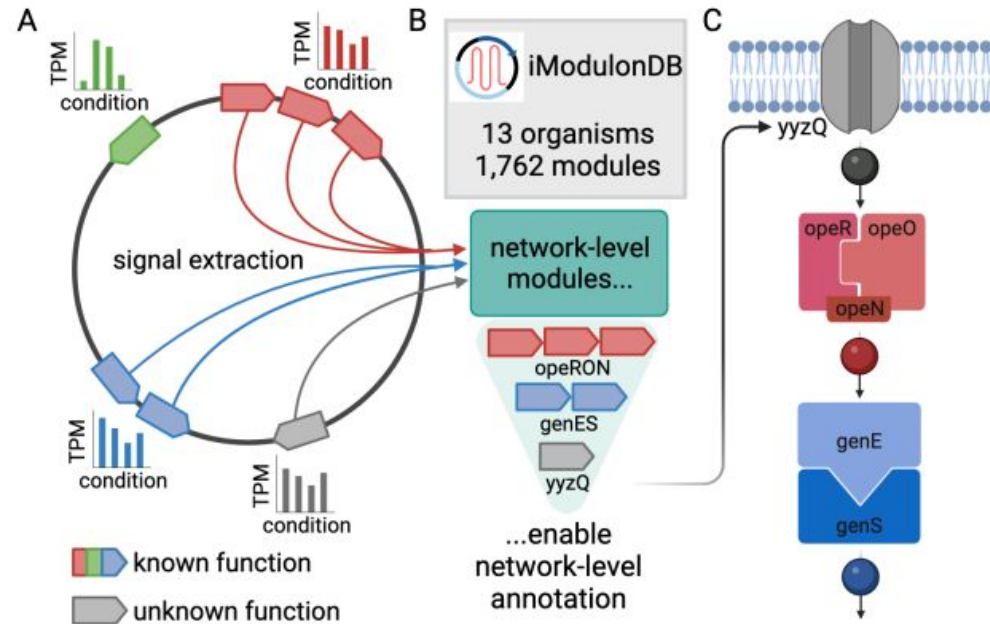
Condition-specific and genetic perturbations displayed

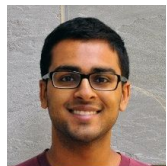
What are iModulons good for?

1. Synthetic biology on a new scale: cross-species iM transfer (1).
2. First-principles based media formulation (2)
3. Improving heterologous protein expression (3)
4. Understanding cell physiology
- 5. Deorphanization of the y-ome
6. DNA transfer processes
 - a. Competence (3)
 - b. We just found a conjugation iModulon
7. Carbon concentrating mechanisms (CCM) in cyanobacteria (4)
8. Soil microbiomes (5)
9. Finding BGCs (6)
10. Phylogeny: comparative iModulon structures

Gene co-regulation predicts co-function

Co-expressed genes often are related to the same biological process (and thus share regulators)

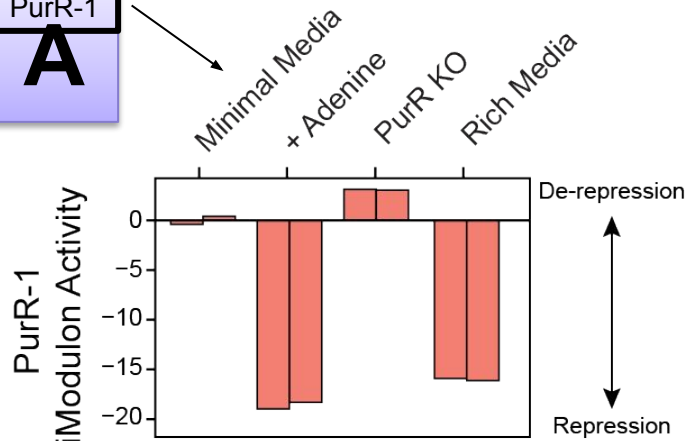
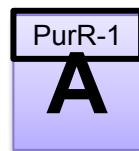
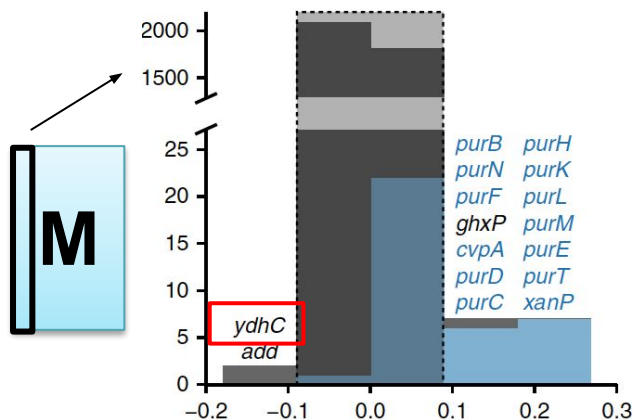




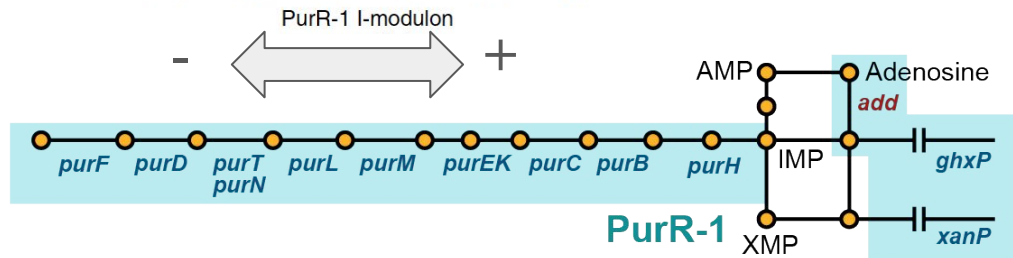
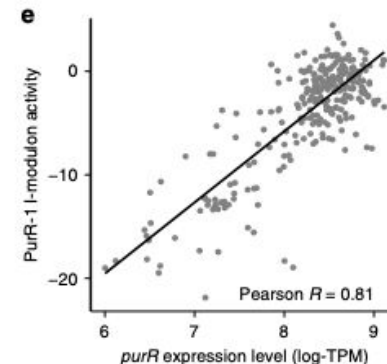
Anand Sastry

Small scale workflow:

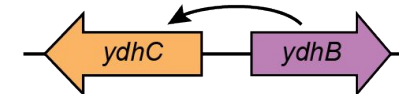
y-Gene mechanistic function revealed within an iModulon



Activity correlates with PurR expression



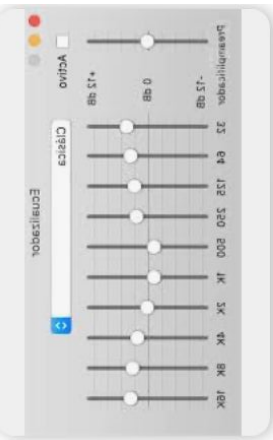
Discovery of a novel adenosine transporter



Comms biol., 4:991 (2021).

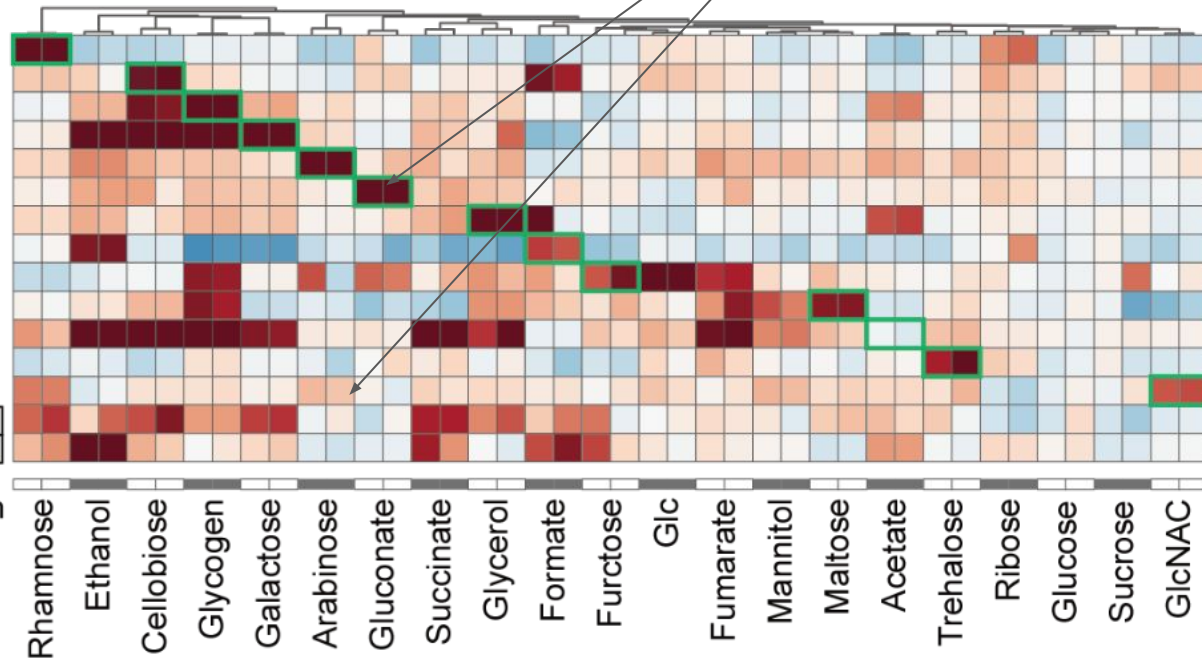
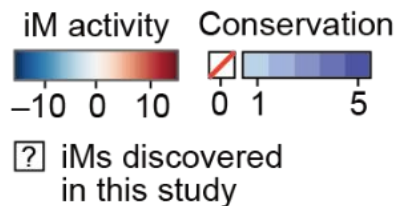
Carbon utilization iModulons in *V. natriegens*

- Diagonal – expected
- Off-diagonal – ‘side effects’



Carbon utilization iModulons

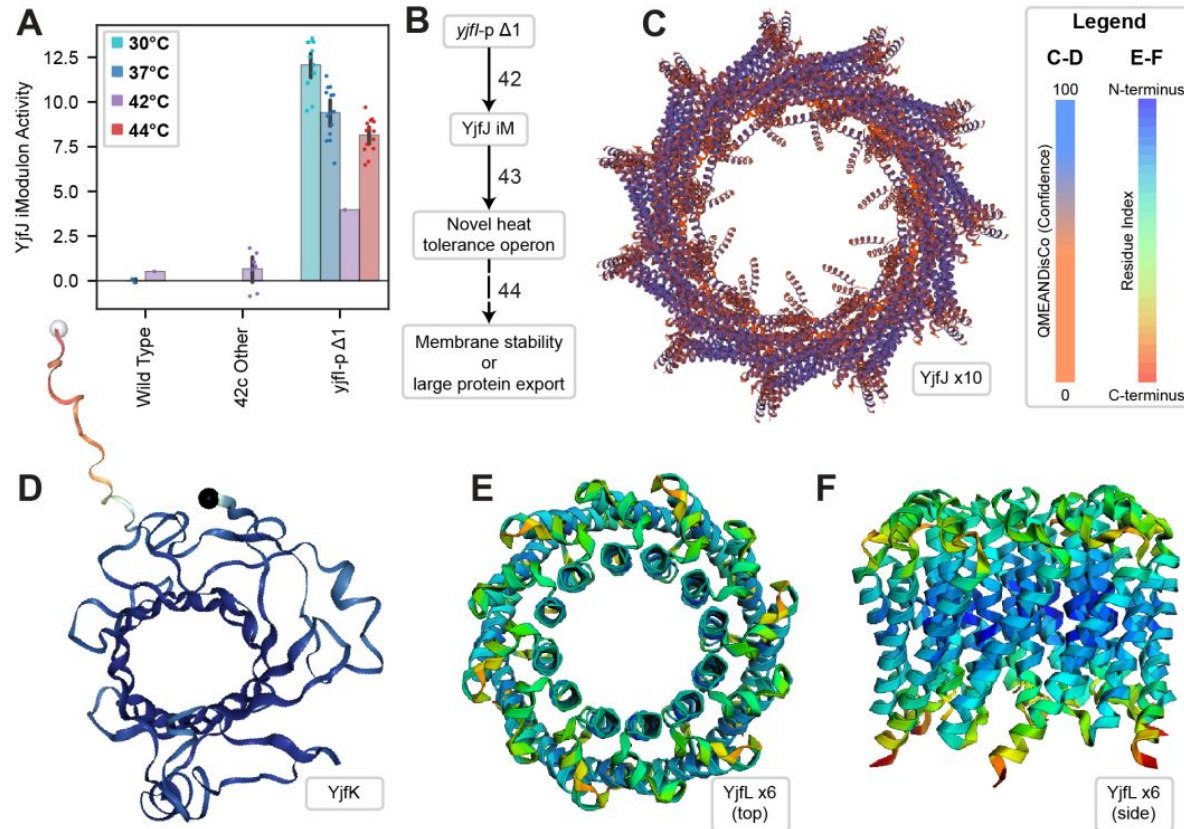
Rhamnose iM	1
Cellobiose iM	1
AGL iM	1
GalR iM	2
Arabinose iM	2
GntR iM	1
Glycerol iM	4
Formate iM	1
FruR iM	5
Maltose iM	1
Acetate iM	1
TreR iM	1
NagC iM	1
TTT iM	?
TRAP iM	?



TRAP: Tripartite ATP-Independent Periplasmic Transporters
TTT: Tripartite Tricarboxylate Transporters

iModulons, TTT and TRAP, are activated under several carbon sources, suggesting they could be involved in the uptake of broad range of carbons

The newly discovered *yjflJKL* operon (contains only y-genes) may be a new heat tolerance operon in *E. coli*



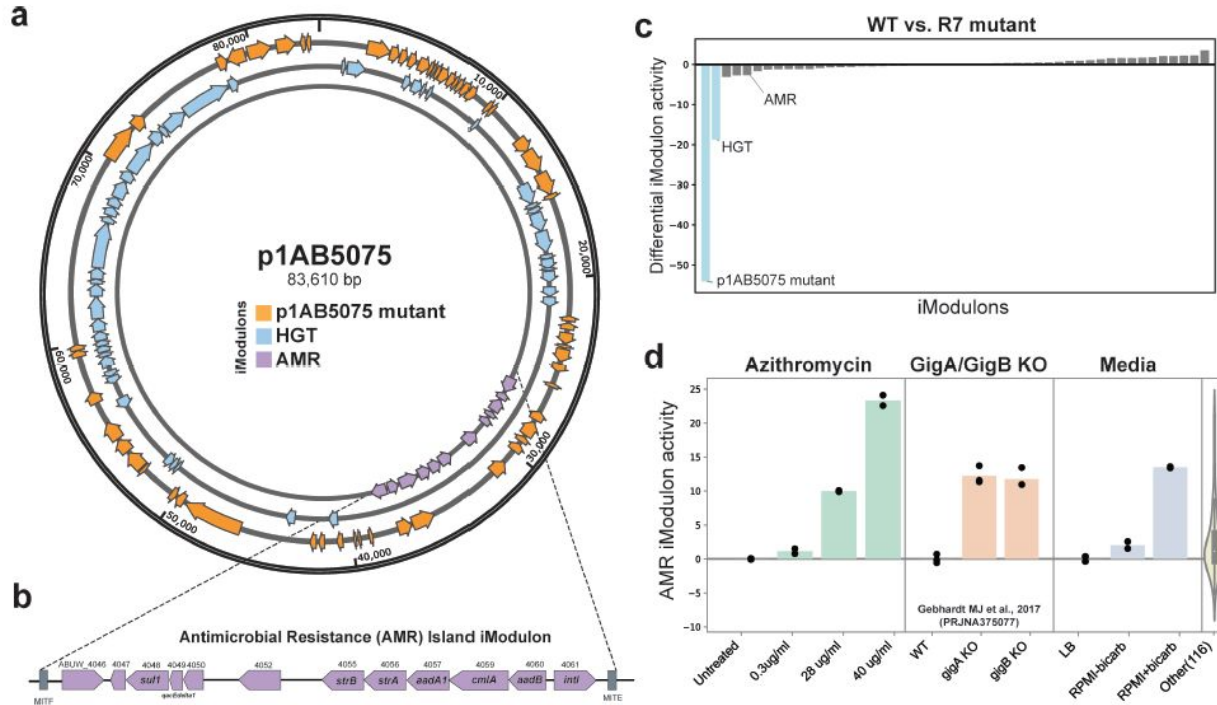
Using computations of protein structure to generate hypotheses about y-gene function (temperature stabilization of membrane, export of misfolded protein, ...)

Structures generated by QSPACE

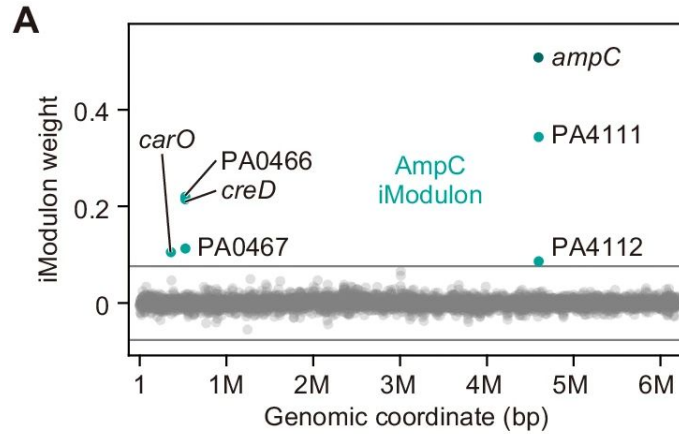


Kevin Rychel

Antibiotic-responsive iModulons reveal AMR mechanisms with y-genes (new targets?) in *Acinetobacter baumannii*

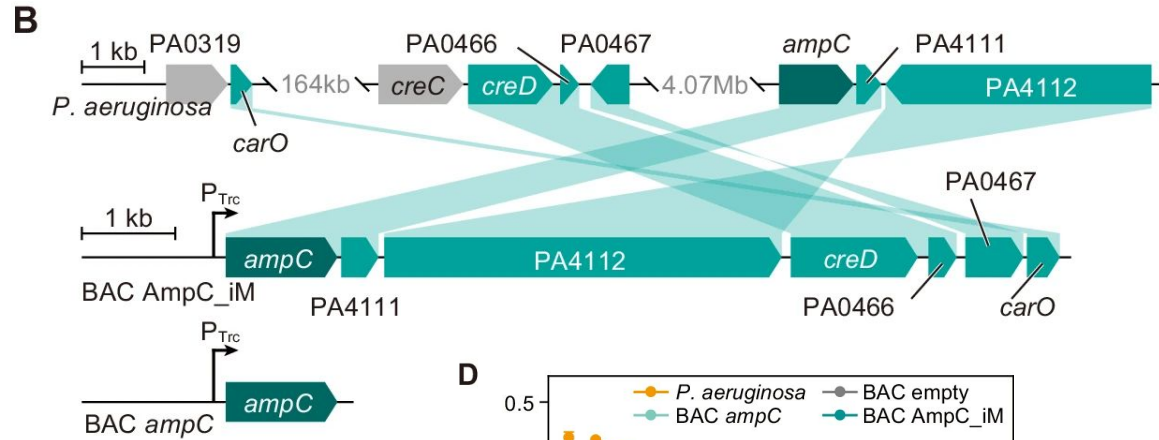


'Auxiliary' genes in AMR iModulons can infer improved phenotypes

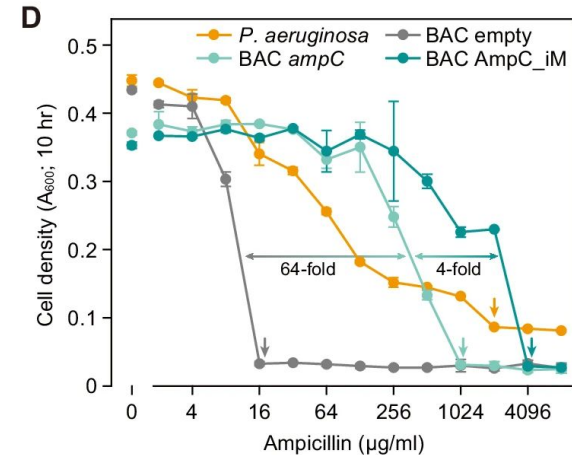


CROSS-SPECIES iMODULON TRANSFER

E. coli carrying the *Pseudomonas aeruginosa* AmpC iModulon (7 genes) confers better ampicillin resistance than cells expressing beta-lactamase alone.

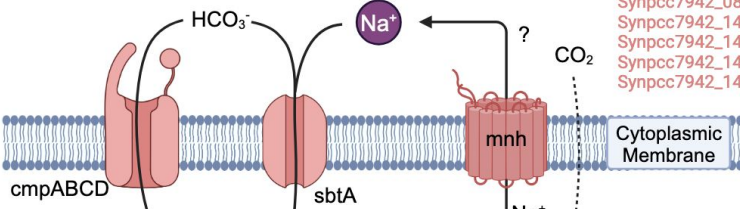
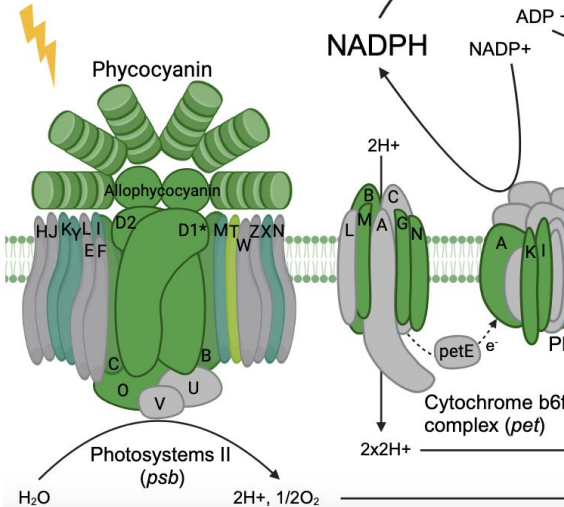
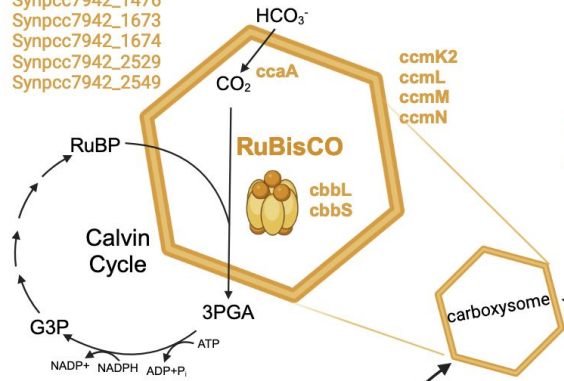


Donghui Choe



y-genes(6):

Synpcc7942_0900
Synpcc7942_1476
Synpcc7942_1673
Synpcc7942_1674
Synpcc7942_2529
Synpcc7942_2549



- Photosystems iModulon
- PSII iModulon
- Photosystems & PSII iModulon
- ccm-1 iModulon
- ccm-2 iModulon

y-genes(8):

Synpcc7942_0259
Synpcc7942_0551
Synpcc7942_0797
Synpcc7942_0857
Synpcc7942_1090
Synpcc7942_2161
Synpcc7942_2413
Synpcc7942_2414

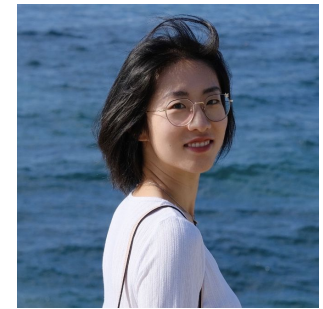
y-genes(9):

Synpcc7942_0113
Synpcc7942_0911
Synpcc7942_1279
Synpcc7942_1483
Synpcc7942_1702
Synpcc7942_1712
Synpcc7942_2233
Synpcc7942_2523
Synpcc7942_2060

iModulons reveal basic cellular functions in *Synechococcus*:

- 1) Carboxysome
- 2) Carbon concentrating mechanisms (CCMs)
- 3) photosynthesis

They include the identification of 29 y-genes that are involved in these basic cellular functions

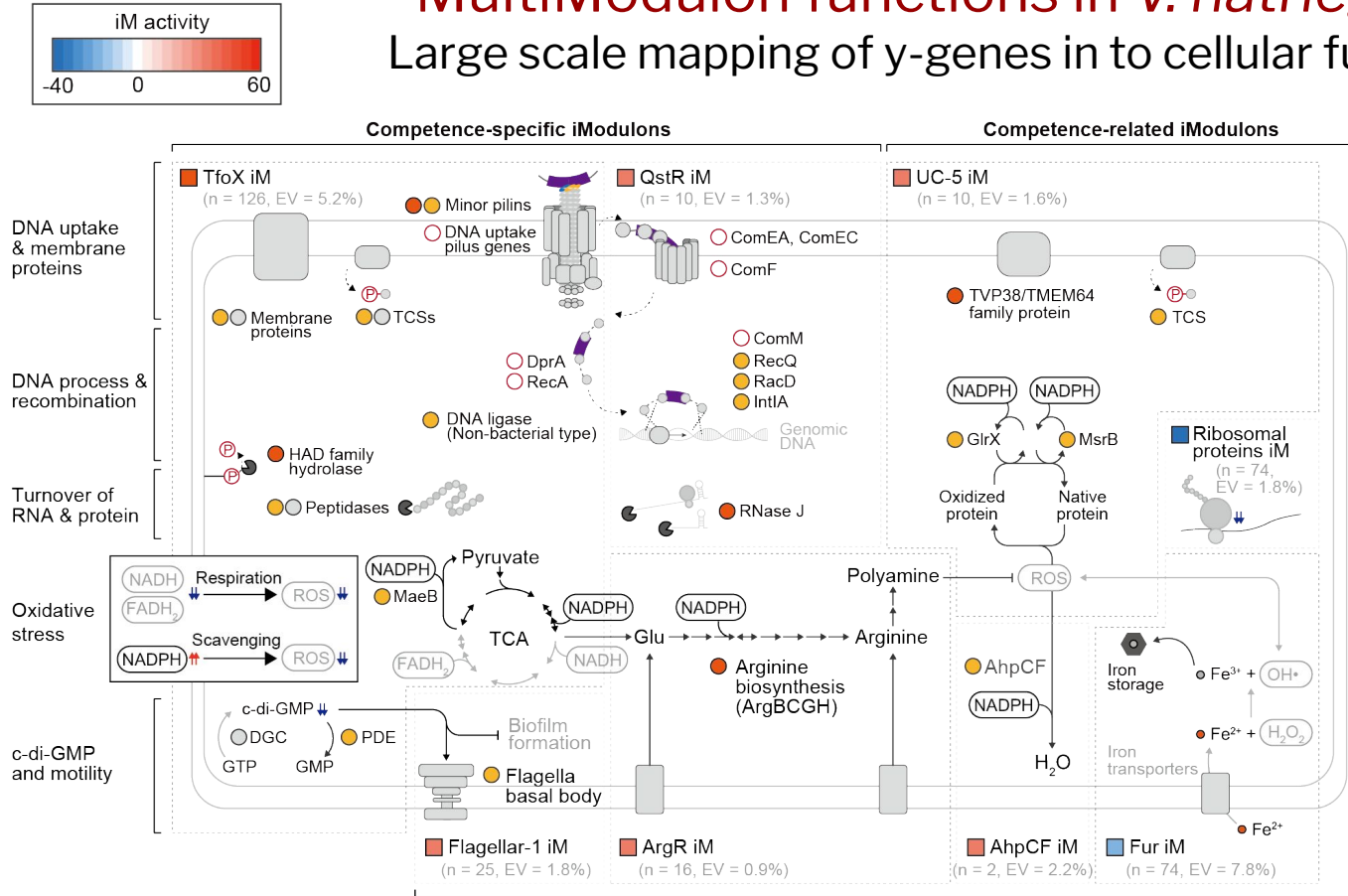


Yuan Yuan

MultiModulon functions in *V. natriegens*: Large scale mapping of γ -genes in to cellular functions



Jongoh Shin



189 genes identified in three iModulons (two known, one γ -iModulon) as competency-associated;

102 of them were γ -genes.

γ -genes knocked out and classified as essential, quasi-essential, or non-essential to a functional competency state.

Large-Scale Discovery Opportunity: Substantial # of y-gene targets from iModulon-based annotation

iM-based annotation based on 1035 RNAseq samples
in *E. coli*:

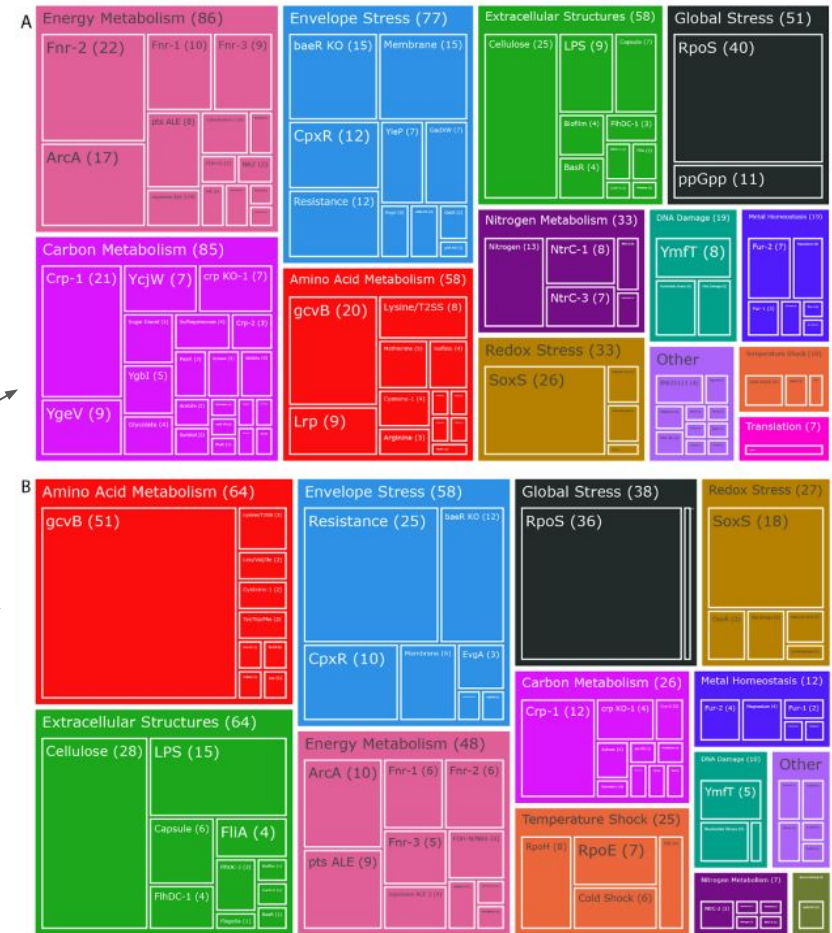
961 of 1693 under-annotated genes in iModulons

530 partially-characterized genes (panel A)

431 uncharacterized genes (panel B)

We have a road map for y-gene deorphanization

Many y-TFs (yneJ, ygeV, ymfT, punR, other LysR-like
TFs) with putative regulons being systematically
validated in several studies from our lab



Summary

- The relationship between the molecular makeup of cells and their phenotypic functions is hierarchical
- Bacterial transcriptomes have been modularized using source signal extraction algorithms
- These iModulons (sets of genes that are independently modulated) can be knowledge-enriched and thus assigned to a cellular function
- The majority of iModulons identified have y-genes
- Thus, we can now annotate y-genes en mass to cellular functions, that makes molecular function much easier to elucidate
- *"Presented in a logical order, not necessarily the chronological order"*, Jared Brodrick