REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
23-01-2017	Final Report	1-Jun-2012 - 30-Apr-2016

4. TITLE AND SUBTITLE
Final Report: Dynamics and Evolution of Associative Memory in Bacterial Populations

5a. CONTRACT NUMBER
W911NF-12-1-0231

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS
Ilias Tagkopoulos

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
University of California - Davis
Sponsored Programs
1850 Research Park Drive, Suite 300
Davis, CA 95618 -6153

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
61745-LS.6

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT
This project addressed two fundamental questions in systems microbiology. The first is, can predict microbial expression and phenotypes in novel conditions if we use past measurements to model their cellular state and behavior? The second is, how cells adapt in the presence of stress combinations that are either sequential or simultaneous. The project has been highly successful providing answers in both, creating the most accurate and integrative model for microbial phenotype prediction and validating findings in our experimental lab.

15. SUBJECT TERMS
microbial evolution, associative memory, multi-omics modeling and simultation, data integration

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Ilias Tagkopoulos
UU	UU	UU	UU		19b. TELEPHONE NUMBER 530-752-4821

Report Title

Final Report: Dynamics and Evolution of Associative Memory in Bacterial Populations

ABSTRACT

This project addressed two fundamental questions in systems microbiology. The first is, can predict microbial expression and phenotypes in novel conditions if we use past measurements to model their cellular state and behavior? The second is, how cells adapt in the presence of stress combinations that are either sequential or simultaneous. The project has been highly successful providing answers in both, creating the most accurate and integrative model for microbial phenotype prediction and validating findings in our experimental lab.

Enter List of papers submitted or published that acknowledge ARO support from the start of the project to the date of this printing. List the papers, including journal references, in the following categories:

(a) Papers published in peer-reviewed journals (N/A for none)

Received		Paper
01/23/2017	4	Minseung Kim, Navneet Rai, Violeta Zorraquino, Ilias Tagkopoulos. Multi-omics integration accurately predicts cellular state in unexplored conditions for Escherichia coli, Nature Communications, ( ): 13090. doi:
01/23/2017	5	Violeta Zorraquino, Minseung Kim, Navneet Rai, Ilias Tagkopoulos. The genetic and transcriptional basis of short and long term adaptation across multiple stresses in, Molecular Biology and Evolution, ( ): . doi:
TOTAL:	**2**	

Number of Papers published in peer-reviewed journals:

(b) Papers published in non-peer-reviewed journals (N/A for none)

Received		Paper
TOTAL:		

Number of Papers published in non peer-reviewed journals:

(c) Presentations

Number of Presentations: 0.00

Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received Paper

TOTAL:

Number of Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received Paper

TOTAL:

Number of Peer-Reviewed Conference Proceeding publications (other than abstracts):

(d) Manuscripts

Received Paper

TOTAL:

Number of Manuscripts:

Books

Received Book

TOTAL:

Received	Book Chapter

TOTAL:

Patents Submitted

Patents Awarded

Awards

National Academy of Engineering, Frontiers of Engineering Symposium, Fellow (83 selected nationwide).

Graduate Students

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Names of Post Doctorates

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Names of Faculty Supported

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Names of Under Graduate students supported

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Student Metrics

This section only applies to graduating undergraduates supported by this agreement in this reporting period

The number of undergraduates funded by this agreement who graduated during this period: 0.00

The number of undergraduates funded by this agreement who graduated during this period with a degree in science, mathematics, engineering, or technology fields:..... 0.00

The number of undergraduates funded by your agreement who graduated during this period and will continue to pursue a graduate or Ph.D. degree in science, mathematics, engineering, or technology fields:..... 0.00

Number of graduating undergraduates who achieved a 3.5 GPA to 4.0 (4.0 max scale):..... 0.00

Number of graduating undergraduates funded by a DoD funded Center of Excellence grant for Education, Research and Engineering:..... 0.00

The number of undergraduates funded by your agreement who graduated during this period and intend to work for the Department of Defense 0.00

The number of undergraduates funded by your agreement who graduated during this period and will receive scholarships or fellowships for further studies in science, mathematics, engineering or technology fields:..... 0.00

Names of Personnel receiving masters degrees

NAME

Total Number:

Names of personnel receiving PHDs

NAME

Total Number:

Names of other research staff

NAME PERCENT_SUPPORTED

FTE Equivalent:
Total Number:

Sub Contractors (DD882)

Inventions (DD882)

Scientific Progress

Executive Summary:
In this project, we investigated (a) how we can build an accurate predictor of E. coli expression and phenotypes in novel environments, (b) how E. coli adapts in combinations of stresses. To achieve the first aim, we have compiled all publicly available omics data in E. coli and created the most accurate predictive model that incorporates transcriptomics, proteomics, metabolomics, fluxomics and phenomics. To address the second challenge, we have performed laboratory evolution of E. coli for 500 and 1000 generations in 5 abiotic stresses and their combinations, we performed RNA-Seq, DNA-Seq, proteomics and competition assays to correlate fitness with adaptive mutations. We have identified 16 cases of cross-stress protection and 1 case of cross-stress vulnerability, while providing dozens of new gene targets for further investigation.

Products:
- 5 papers in top specialized or general interest venues: PLoS Computational Biology, Molecular Systems Biology (x2), Molecular Biology and Evolution, Nature Communications.

- Identification of multiple pathways and genes related to resistance in conditions that are relevant to clinical and industrial environments.

- Design and development of an omics data integration method, a multi-scale modeling approach based on deep learning, a network-based approach to integrate mutations, differential expression and phenotypes.

- The Ecomics compendium: a normalized compendium of all omics data available for E. coli: http://prokaryomics.com/

- The project provided partial support for one postdoc and one graduate student.

Technology Transfer

FINAL REPORT

Title: Dynamics and Evolution of Associative Memory in Bacterial Populations
PI: Ilias Tagkopoulos, UC Davis
Period: 06/01/2012 to 04/30/2016
Contract ID: W911NF1210231

1. Executive Summary

This project addressed two fundamental questions in systems microbiology. The first is, can predict microbial expression and phenotypes in novel conditions if we use past measurements to model their cellular state and behavior? The second is, how cells adapt in the presence of stress combinations that are either sequential or simultaneous. The project has been successful providing answers in both, leading to new discoveries in cross-stress behavior and creating the most accurate integrative model for microbial phenotype prediction so far.

2. Results

2.1 Not all stress combinations are equal. Some make microbes more resilient, others more vulnerable to environmental perturbations.

In the real world, microbes are not exposed to just one stress; instead they have to cope with a myriad of simultaneous or sequential environmental fluctuations in their natural environment. To better understand how their behavior changes across these dimensions, early on in this project we evolved *Escherichia coli* cells over 500 generations in five environments that include four abiotic stressors (acidic, osmotic, butanol, oxidative) [1]. Through growth profiling and competition assays, we identified several cases of positive and negative cross-stress behavior that span all strain–stress combinations. Resequencing the genomes of the evolved strains resulted in the identification of several mutations and gene amplifications, whose fitness effect was further assessed by mutation reversal and competition assays. Transcriptional profiling of all strains under a specific stress, NaCl-induced osmotic stress, and integration with resequencing data further elucidated the regulatory responses and genes that are involved in this phenomenon. What we found out is that cross-stress dependencies are ubiquitous, highly interconnected, and can emerge within short timeframes.

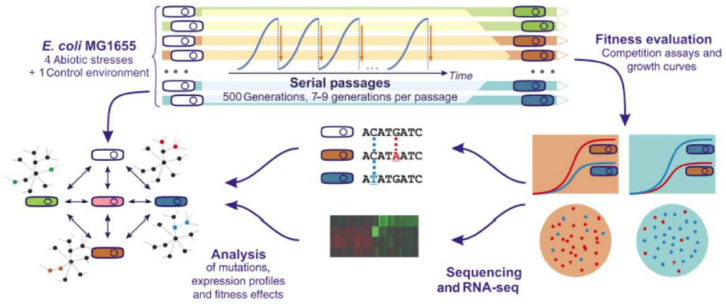


Figure 1 Overview of the experimental setting. *E. coli* MG1655 cell lines were evolved for 500 generations in five environments with minimal M9 salt media and glucose as the sole carbon source. These environments included four abiotic stresses (acidic, osmotic, oxidative, and *n*-butanol stress) and a control medium-only environment. The relative fitness of all evolved strains was measured under all stresses by competition assays and growth curves. Selected clones from adapted populations were sequenced and transcriptional profiles were obtained by RNA-Seq. Individual mutations from the resequenced clones where associated to phenotypic fitness by mutation reversals and competition assays. Identified mutations, expression data and relative fitness of the evolved strains under various stressors were analyzed in a system-level approach.

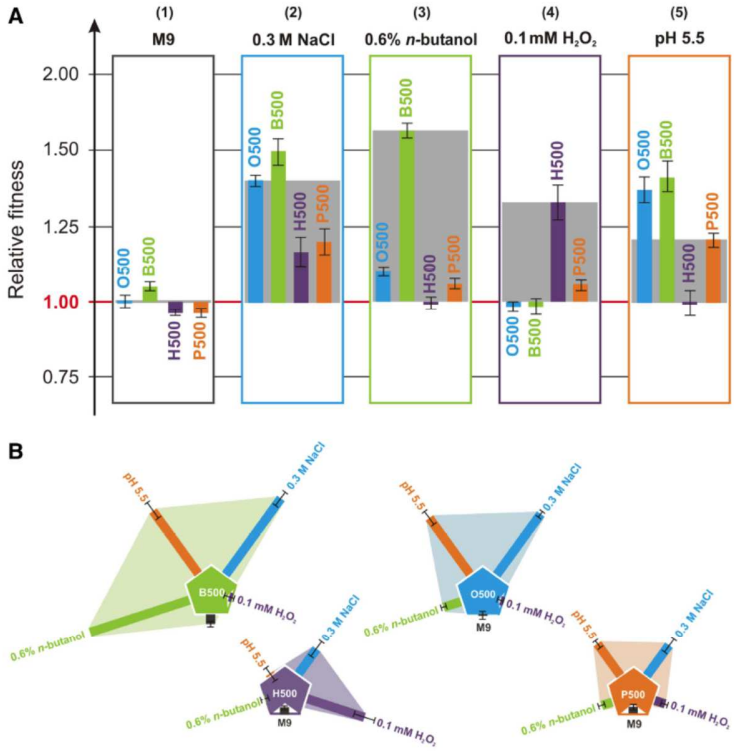


Figure 2 Relative fitness of the evolved populations under all environmental conditions. The relative fitness of populations evolved in all five environments was assessed through competition assays. Fitness refers to Darwinian fitness as measured through competition assays relative to the medium-only adapted population (G500). (**A**) Environment-based representation: (1) In the absence of other stressors, all populations showed small differences in fitness; (2) all populations where significantly more protected in osmotic stress in comparison to the G500 population; (3–4) while the B500 and O500 populations have high fitness under *n*-butanol stress, they showed a small evolutionary trade-off under oxidative stress; (5) Interestingly, O500 and B500 populations outcompeted P500 under acidic stress. Shaded gray areas depict the fitness difference between the strain evolved in the respective environment relative to the G500 population. (**B**) Population-based representation of the relative fitness data shown in (A). Shaded areas depict the degree of cross-stress protection in each respective environment. Error bars show standard error of the mean for eight independent competitions; counts for each competition were averaged over two plates in each experiment. Competition assays were obtained over 48 h of growth with one transfer (1:500 dilution) at 24 h. Supplementary Table S-V summarizes the competition assay results. Source data for this figure is available on the online supplementary information page.

2.2 We identified the genetic basis of cross-stress adaptation. A few genes is all it takes.

We then went on to understand how reproducible evolution is and what is the genetic basis of the cross-stress behavior. We first resequenced the genomes (right) and then repaired the mutations before competing the strains in the corresponding environments, thus linking specific mutations to phenotypic fitness. For more information on the specific genes and how they confer this selective advantage see [1].

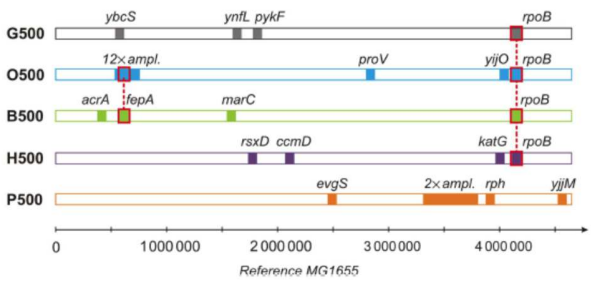
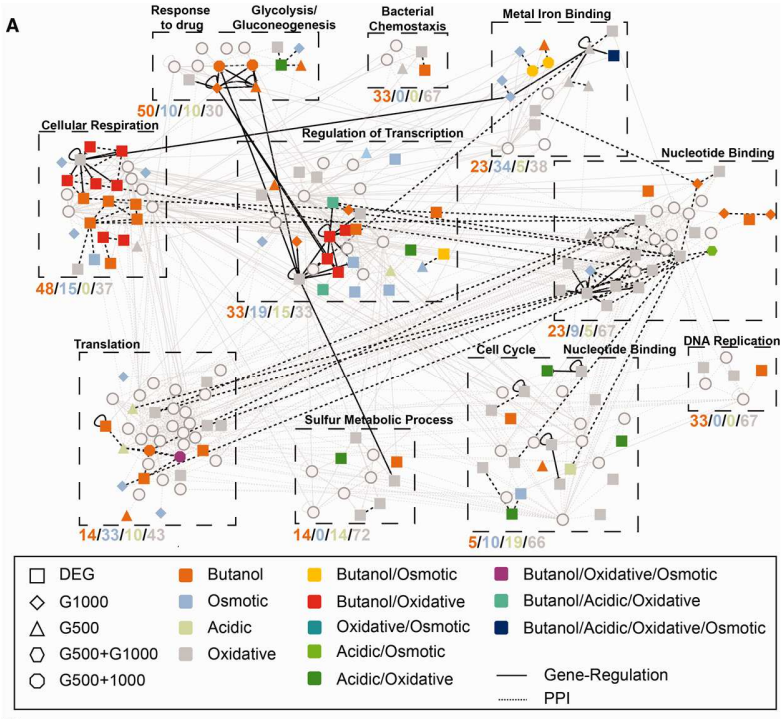


Figure 3 Mutation map of fixed genetic changes in the evolved strains. The map of mutations identified in *E. coli* strains evolved for 500 generations. Coordinates are relative to the reference MG1655 genome. The seven identified mutations in the ancestral genome are not shown here (Supplementary Table S-IX).

2.3 We can infer the impact of the mutation when putting everything together

In a more recent work [2], we created a method to integrate mutations, differential expression and GO/pathway annotation to understand what are the key players during cross-stress, short and long term adaptation. What we found out it that we can predict fairly accurately the fitness impact of a mutation from the network structure and following a "guilt-by-association" methodology.



B

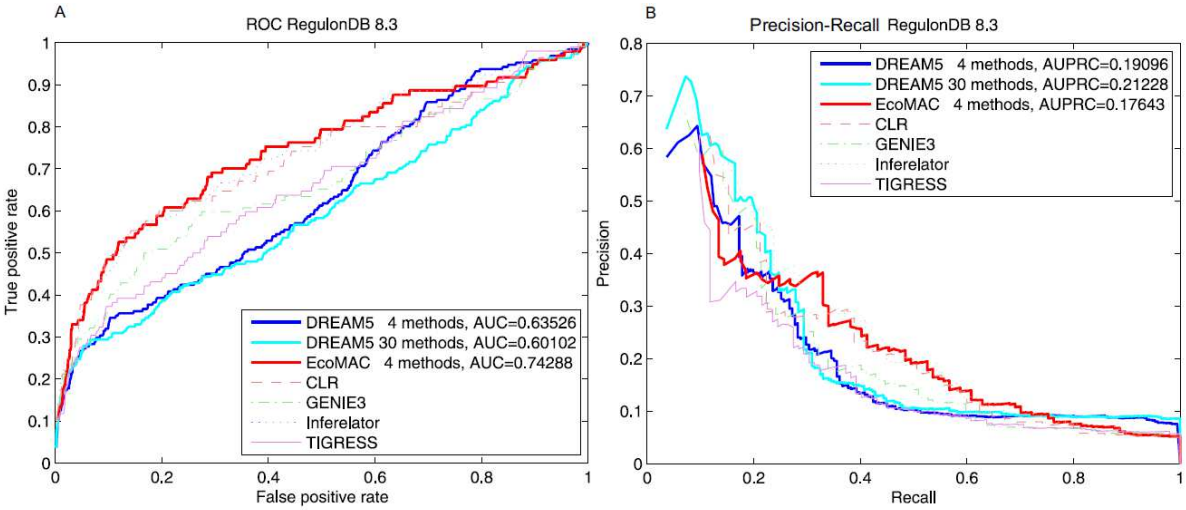
Stress	Pathway	Interactors
n-butanol	Response to Drug	*marA, acrB, acrA, acrD*
	Cellular Respiration	*ubiC, fnr, hyaD, hyaB, hyaB, hyaE, hyaF, hyaA, narL hyaC, narG, narH, narJ*
	Glycolysis/Gluconeogenesis	*eno, pykF, pfkA, pgi*
Osmotic	Metal Iron Binding	*entD, katG, pyrC, fes, fur, fepA, yhjA*
	Translation	*rpsM, spsD, rplA, rplM, rplB, rplV, rpmB, rplC, pth, rplC*
	Regulation of Transcription	*crp, hns, flhC, gadX, evgA, evgS, gadE, phoP, yjM, uxuR*
Acidic	Regulation of Transcription	*crp, hns, flhC, gadX, evgA, evgS, gadE, phoP, yjM, uxuR*
	Cell cycle	*groL, ftsE, ftsA, ftsN, ftsY, murB*
Oxidative	Cellular Respiration	*ubiC, fnr, hyaD, hyaB, hyaE, hyaF, hyaA, narL hyaC, narG, narH, narJ*
	Metal Iron Binding	*entD, katG, pyrC, fes, fur, fepA, yhjA*
	Bacterial Chemostaxis	*cheY, cheR, motB, tsr, trg*
	Nucleotide Binding	*groL, atpA, ftsE, ftsA, lpdA, tyrR, murB, ydiA*
	Sulfur Metabolic Process	*clb, cysB, cysU, cysW, tauB, tauC, cysP*

What you see in the figure to the right: Network analysis and implicated pathways in stress resistance. (A) A functional network was constructed from PPI and TF-DNA data, superimposed with the re-sequencing and transcriptional profiling results of our analyses. Genomic data from cells evolved for 1000 generations was added to the network Modularity-based algorithms were used to identify communities within the network, which were further analyzed for enriched clusters. The name of the most statistically significant cluster and the profile of the mutants/DEGs in terms of the corresponding stress (Butanol, Osmotic, Acidic, Oxidative, in that order) for each community is shown. For example, in the case of the first community, the Glycolysis GO term is the most over-represented and 50% of the observed mutations/DEGs in the community were identified in cell lines exposed/evolved in n-butanol. Light pink nodes are genes that are not mutated or DEGs, but connect two or more mutated/DEG genes in a path with a length shorter than three. (B) Highly enriched pathways and their members that are implicated in each stress.

2.4 When it comes to prediction, data size does matter.

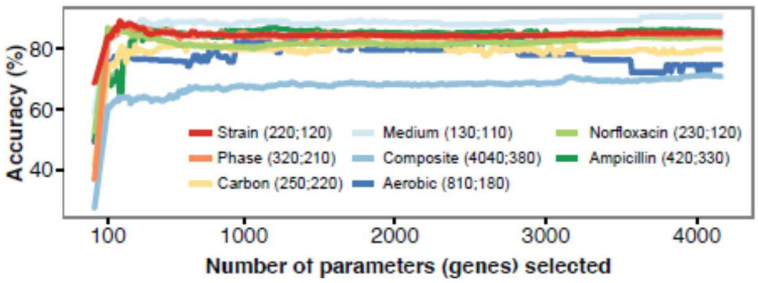
In [3] we created an integrative Metabolism-Expression model of all microarray data so far. Although that model has now superceded by the MoMA model in [6], one key observation was that the availability of high quality normalized and corrected data is the key component on making cellular predictions. In Marbach et al., Nat. Meth. 2012, the authors used ~800 arrays to

train 30 methods to predict transcription factor – DNA interactions as part of the DREAM challenge. In our work in [3], we use 4 of those methods only but we 4x the number of data and after normalization. The result was a ~23% better performance in TF-DNA prediction (interactions validated experimentally), arguing that data quality and quantity is more important than algorithms when it comes to prediction. This led us to spend the next 2 years creating the best omics compendium ever constructed for a microbe.



2.5 Microbial forensics: Predicting cellular characteristics and behavior from transcriptional profiles.

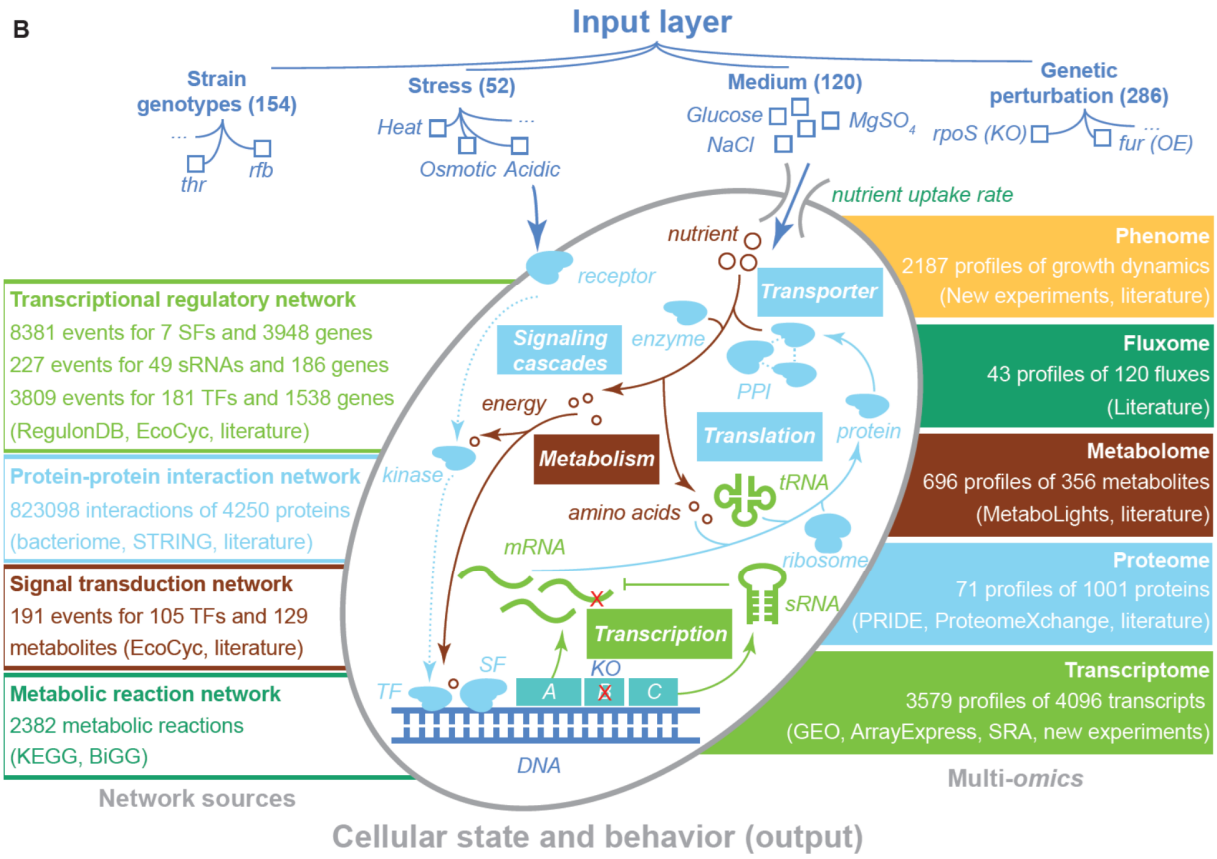
We used the dataset from [3] to train classifiers and answer the following two questions: can we predict microbial traits like growth, environment that the microbe lives in, functionality that it has, etc. from purely transcriptional profiling information? And if yes, the expression of how many genes is necessary.



The answer is yes, we can and tested our predictors against growth, phase, carbon source, medium, aerobic/anaerobic respitration, antibiotic resistance, etc. In all cases the predictive performance is more than 80% [4]. The answer to the second question is 50-300 genes, depending on the environmental and genetic background. The figure on the right shows that the performance (accurracy here) of the predictor converges when up to the top 300 most informative genes are integrated/

2.6 We figured out the optimal integration pipeline for normalized omics data

Following our work on [3], we spend 2 years to create Ecomics, the first cohesive, normalized compendium of the E. coli universe to be used for machine learning. This created a resource with ~4000 transcriptomics, proteomics and metabolomics profiles that have been re-processed from raw data under the same pipeline and connected to a metadata and ontology schemantic fabric.



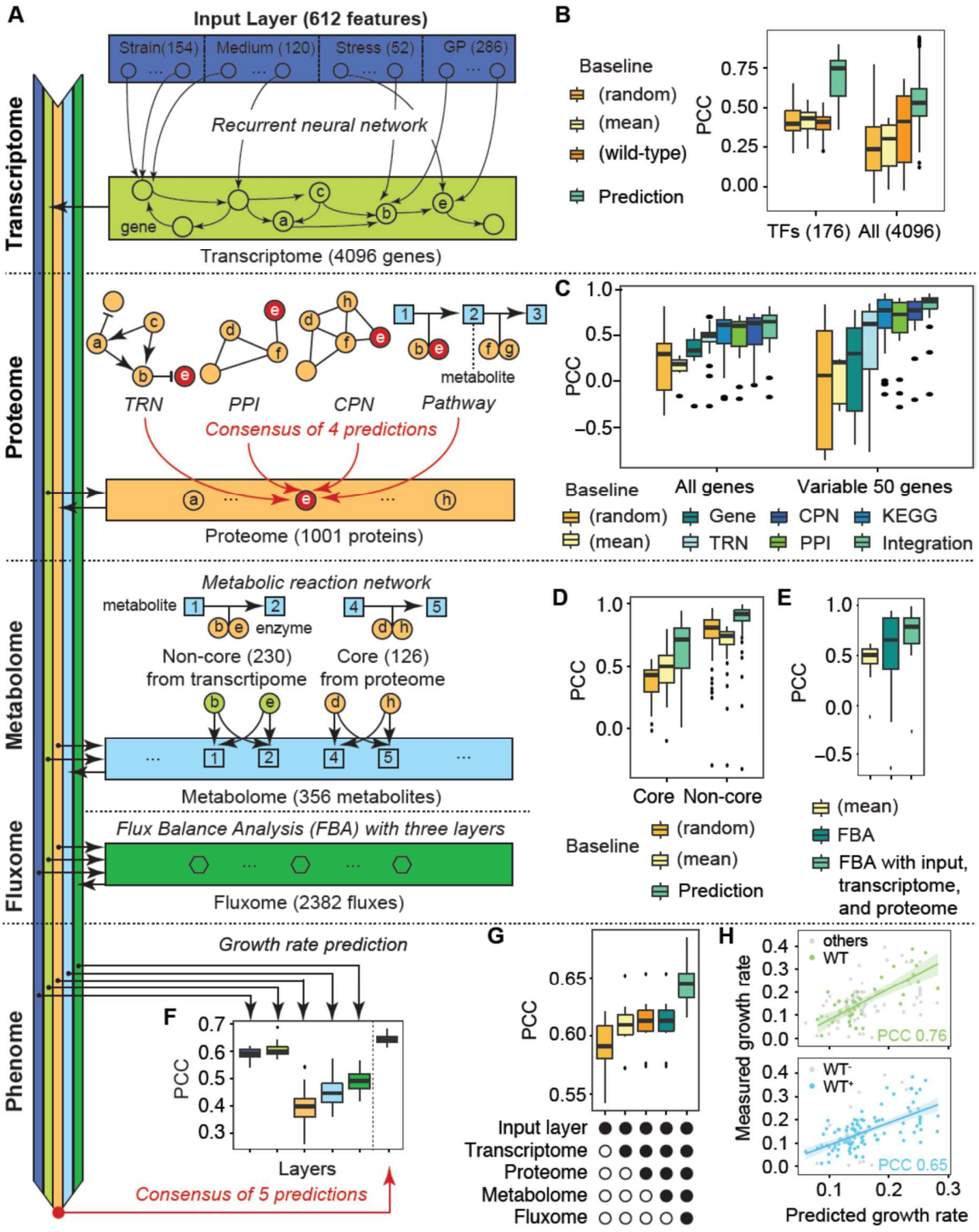
2.7 We can predict ~ 0.65 PCC the traits and gene expression of *E. coli* in new conditions

We created the most accurate and advanced predicted model that exist today. We tested it against all data that humanity has about E. coli and it can predict gene expression and growth new conditions with ~0.65 - 0.8 PCC accuracy (next best is ~0.2 PCC), depending how close the new condition is to one that is already in the compendium. To train the predictor we used ~10,000 cores of the Blue Waters supercomputer (NCSA). We validated findings with forward predictions [6].

For a news alert on this, check this article:
https://phys.org/news/2016-10-crystal-ball-coli-bacteria.html

or here for a podcast:
http://blogs.ucdavis.edu/egghead/2016/10/31/podcast-computer-model-crystal-ball-e-coli/

The compendium is currently available here: http://www.prokaryomics.com . In a related work, we used protein modeling features to predict enzymatic activities and understand which features are informative for forward predictions [5]. This work was in collaboration with the Siegel Lab in UC Davis.



Model architecture and prediction performance. (a) The data and work flow over the five modules, one for each layer. (b) Prediction performance of the transcriptome module ($P < 10^{-13}$), (c) Proteome module ($P < 10^{-5}$), (d) metabolite concentrations ($P < 10^{-13}$), (e) fluxes by all three layers, (f) growth rate from each layer. (g) Additive effect of additional layers in growth rate prediction. (h) Comparison of predicted and measured growth rate.

3. Published Work

The following are the published manuscripts that stem from this work and have acknowledged funding from this grant. Bold indicates the people funded partially by this grant.

A. Cross-stress behavior and stress combinations

[1] Dragosits M, Mozhayskiy V, Quinones-Soto S, Park J, Tagkopoulos I. Evolutionary potential, cross-stress behavior and the genetic basis of acquired stress resistance in Escherichia coli. *Mol Syst Biol.* 2013;9:643. PubMed PMID: 23385483; PubMed Central PMCID: PMC3588905.

[2] Zorraquino V., Kim M., Rai N., Tagkopoulos I. The genetic and transcriptional basis of short and long term adaptation across multiple stresses in *Escherichia coli*. Mol. Bio. Evol. 2016; 34(3):707-717

B. Omics integration and predictive modeling

[3] Carrera J, Estrela R, Luo J, Rai N, Tsoukalas A, Tagkopoulos I. An integrative, multi-scale, genome-wide model reveals the phenotypic landscape of Escherichia coli. *Mol Syst Biol.* 2014 Jul 1;10:735. PubMed PMID: 24987114.

[4] Kim M, Zorraquino V, Tagkopoulos I. Microbial forensics: predicting phenotypic characteristics and environmental conditions from large-scale gene expression profiles. *PLoS Comput Biol.* 2015 Mar;11 (3):e1004127. PubMed PMID: 25774498; PubMed Central PMCID: PMC4361189.

[5] Carlin, D. A., Caster, R. **W., Wang**, X., Betzenderfer, S. A., Chen, C. X., Duong, V. M., Ryklansky C.V., Alpekin A., Beaumont N., Kim N., Mohabbot, H., Pang, B., Teel, R., Whithaus, L., **Tagkopoulos, I**., Siegel J.B. (2016). Kinetic Characterization of 100 Glycoside Hydrolase Mutants Enables the Discovery of Structural Features Correlated with Kinetic Constants. *PloS One*, 11(1), e0147596

[6] Kim M., Rai N., Zorraquino V., Tagkopoulos I. Multi-omics integration accurately predicts cellular state in unexplored conditions for *Escherichia coli. Nat. Comm.* 2016; 7:13090