bioRxiv preprint doi: https://doi.org/10.1101/2021.04.08.439047; this version posted April 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

PRECISE 2.0 - an expanded high-quality RNA-seq compendium for *Escherichia coli* K-12 reveals high-resolution transcriptional regulatory structure

Cameron R. Lamoureux[1†], Katherine T. Decker[1†], Anand V. Sastry[1†], John Luke McConn[1], Ye Gao[1], Bernhard O. Palsson[1,2,*]

1 Department of Bioengineering, University of California San Diego, La Jolla, CA 92093, USA
2 Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, 2800 Lyngby, Denmark
† These authors contributed equally to this work.
* Corresponding author: palsson@ucsd.edu

Abstract

Uncovering the structure of the transcriptional regulatory network (TRN) that modulates gene expression in prokaryotes remains an important challenge. Transcriptomics data is plentiful, necessitating the development of scalable methods for converting this data into useful knowledge about the TRN. Previously, we published the PRECISE dataset for *Escherichia coli* K-12 MG1655, containing 278 RNA-seq datasets created using a standardized protocol. Here, we present PRECISE 2.0, which is nearly three times the size of the original PRECISE dataset and also created using a standardized protocol. We analyze PRECISE 2.0 at multiple scales, demonstrating multiple analytical strategies for extracting knowledge from this dataset. Specifically, we: (1) highlight patterns in gene expression across the dataset; (2) utilize independent component analysis to extract 218 *i*ndependently *modul*ated groups of genes (iModulons) that describe the TRN at the systems level; (3) demonstrate the utility of iModulons over traditional differential expression analysis; and (4) uncover 6 new potential regulons. Thus, PRECISE 2.0 is a large-scale, high-quality transcriptomics dataset which may be analyzed at multiple scales to yield important biological insights.

Introduction

Over the past decade, RNA sequencing (RNA-seq) has emerged as an efficient, high-throughput method to probe the expression state of a cell population. Advances in next-generation sequencing have accelerated the creation of large RNA-seq datasets (1–4), which subsequently enabled the successful development and application of machine learning methods to advance our understanding of transcriptional regulation (5–7).

Previously, we presented the Precision RNA-seq Expression Compendium for Independent Signal Extraction, or PRECISE, containing 278 expression profiles for *Escherichia coli* generated across a five-year period (1). We applied Independent Component Analysis (ICA), a signal deconvolution algorithm, to this dataset to reveal 92 independently-modulated sets of genes, or iModulons, that encoded the structure and dynamics of transcriptional regulation in

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.08.439047; this version posted April 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

this model organism. Two-thirds of iModulons represented the quantitative effects of specific transcriptional regulators on the transcriptome, while most of the remaining iModulons captured interpretable biological signals.

In a comprehensive analysis of module detection methods, ICA outperformed over 40 other algorithms in modeling regulatory networks across multiple organisms (8). Applications of ICA to human transcriptomics datasets include elucidation of cancer pathways (9), prediction of drug responses (10) and gene functions (11), and characterization of tumor subtypes (12, 13). Additionally, ICA has shown promise in extracting functional modules from single cell RNA-seq data (14, 15).

We have also applied ICA to transcriptomics datasets for two additional microbes, *Bacillus subtilis* (16) and *Staphylococcus aureus (17)*, to identify their respective iModulon structures. All microbial iModulons can be explored through the interactive dashboards available at iModulonDB (18).

iModulons are usually associated with a specific transcriptional regulator, and have two major properties: (1) iModulon gene weights, which determine the relative effect of the transcriptional regulator on each gene in an iModulon, and (2) iModulon activities, which represent the overall activity state of the transcriptional regulator under each condition in the dataset. These properties accelerate the interpretation of complex transcriptional changes. For example, application of iModulons to a transcriptomics dataset for 40 heterologous proteins expressed in *E. coli* could explain over half of the gene expression variation through five specific host responses (19). iModulons have also been applied to reconstruct the regulons for two-component systems (20), characterize transcription factor (TF) mutations (21), and interrogate transcriptional changes during naphthoquinone-based aerobic respiration (22).

In the two years following the release of PRECISE 1.0, we have nearly tripled the size of this dataset. Here, we present PRECISE 2.0, a high-quality transcriptomic dataset containing 815 RNA-seq datasets generated using a standardized protocol. We conduct a multi-scale analysis of PRECISE 2.0 to extract knowledge about the TRN. Specifically, we: (1) compare expression patterns amongst groups of genes; (2) apply independent component analysis (ICA) to extract 218 *i*ndependently *modul*ated groups of genes (iModulons) and map them onto the TRN; (3) compare iModulon-based and gene-based methods for differential expression analysis; and (4) discover novel regulons for 6 transcription factors. Thus, PRECISE 2.0 is a large-scale, high-quality transcriptomics dataset which may be analyzed at multiple scales to yield important biological insights.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.08.439047; this version posted April 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Results

Multi-scale analysis of RNA-seq datasets

PRECISE 2.0's size motivates the development of multi-scale analytical methods. **Figure 1** summarizes the distinct scales of analysis that become available upon construction of such a large dataset. RNA-seq experiments yield data at the individual gene scale, with around 4,000 observations per experiment. Classical differential expression of gene (DEG) analysis operates on the scale of thousands of genes. Given the large number of observations , this approach to transcriptomic analysis can quickly become cumbersome.

Recently, iModulons, or *i*ndependently *modul*ated groups of genes discovered through ICA, have been used to summarize and interpret patterns in gene expression (1). iModulons can be viewed as top-down omics-driven analogues to the bottom-up bi-molecular association driven definition of regulon (23, 24). iModulons enable a massive dimensionality reduction in the interpretation of gene expression data, decomposing thousands of variables to around a hundred. Thus, iModulons allow for the description of transcriptome allocation to different cellular functions. This systems viewpoint in turn enables differential iModulon activity (DIMA) analysis, which identifies a set of transcriptional signals (representing transcriptional regulators) that can efficiently explain changes to transcriptome composition.

Furthermore, iModulons may be automatically clustered into tens of groups that correspond to shared cellular processes. Again, this analytical scale has a biological representation, corresponding to the concept of a stimulon, or a set of regulatory and cellular processes induced by a certain environmental stimulus. Certain physiological measurements correspond to the activities of such clusters of iModulons (16, 25).

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.08.439047; this version posted April 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

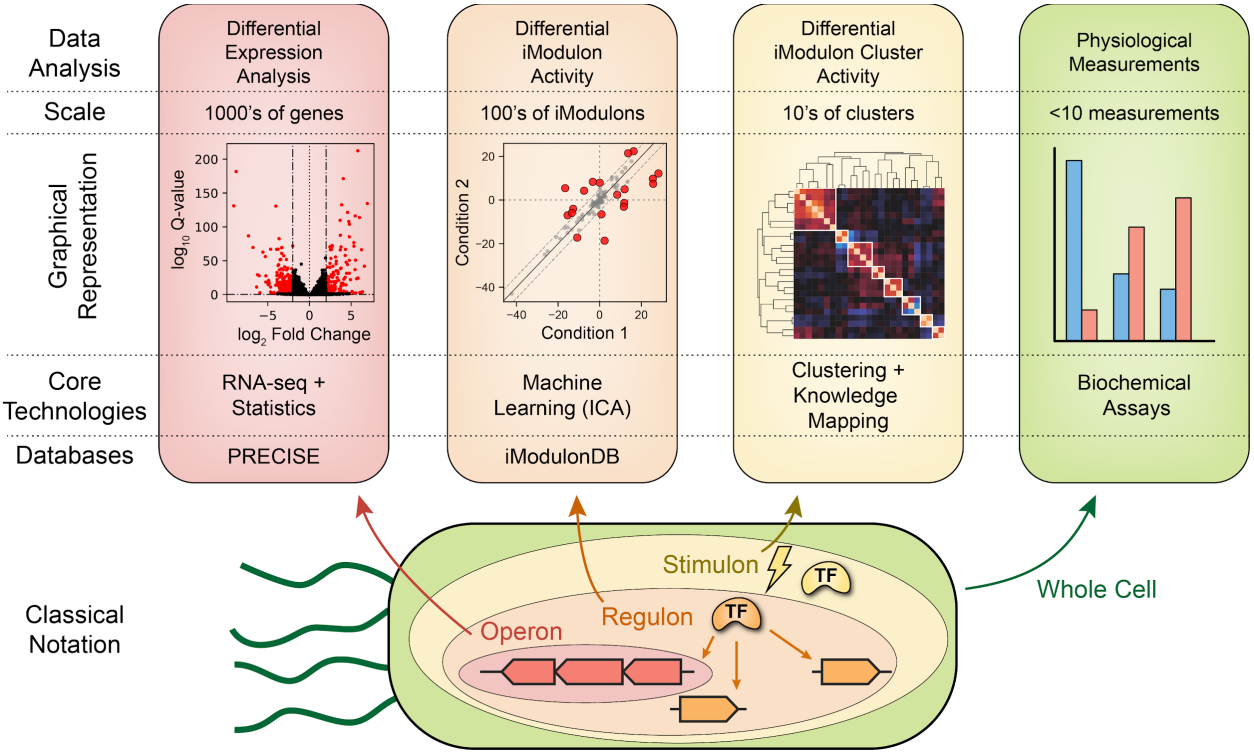

Figure 1: Multi-scale analysis of PRECISE 2.0. The levels of analysis approximately correspond to the definition of an operon and a regulon, and the notion of a stimulon. The 'scale' indicates the reduction of dimensionality over the levels shown.

PRECISE 2.0 contains a wide range of genetic perturbations and environmental conditions

To investigate the properties of RNA-seq data on a large scale, we constructed PRECISE 2.0 to enable multi-scale analysis of the systems biology of *E. coli* K-12 MG1655. PRECISE 2.0 is a large, high-fidelity expression compendium consisting of 815 individual RNA-seq samples generated using a standardized protocol executed in a single lab (cite SI for the Protocol). PRECISE 2.0 constitutes a nearly 3-fold increase in size from the 278-sample PRECISE 1.0 (1), published two years previously (**Figure 2A**).

The samples in PRECISE 2.0 span 32 unique growth conditions, including 167 samples from gene deletion strains and 94 samples of heterologous protein production. 375 samples - almost half of the dataset - are derived from adaptive laboratory evolution (ALE) endpoints. PRECISE 2.0 thus contains changes in nutrients, stresses, genetic parameters, adaptation to new growth conditions (21, 22, 26–28), and forced expression of heterologous (19) and orthologous genes (29). It thus represents a wide range of conditions under which changes in the composition of the *E. coli* transcriptome can be studied.

The first two principal components of the dataset capture 32.3% of the overall variance (**Figure 2B**). While many studies are grouped along the same axis, some projects are outliers in

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.08.439047; this version posted April 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

principal component space. Studies involving ALE endpoints tend to produce more diverse transcriptomes as cells evolve for growth after an environmental or genomic perturbation. Furthermore, projects that use diverse growth media, such as the Two-Component System project (20) or AntibiotICA project (25), likewise result in more distinct gene expression profiles.

To gain further appreciation of the diversity of PRECISE 2.0, we computed the differentially expressed genes (DEGs) across all unique condition pairs in the database. On average, 765 genes were differentially expressed between two conditions, with some comparisons producing over 2000 DEGs (47% of the 4211 genes included in the dataset) (**Figure 2C**). The condition producing the fewest DEGs on average compared to all other conditions was the glucose-fed growth of *rpoB* point mutant E546V, yielding an average of 417 DEGs. Conversely, the condition involving growth in 5% w/v ethanol-supplemented LB media with deletion of two-component system response regulator *baeR* resulted in 1775 DEGs on average.

PRECISE 2.0 also contains 145 expression profiles of strains with knocked-out TFs. This dataset therefore provides an opportunity to compare the number of DEGs resulting from such knockouts. In particular, we find that there is no clear relationship between regulon size and number of DEGs after TF knockout (**Figure 2D**). In particular, TCS response regulators typically have small direct regulons, but knocking them out significantly perturbs the expression of hundreds of genes. This large number of DEGs may be related to the key role of TCSs in instigating broad adaptations to extracellular signals (30).

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.08.439047; this version posted April 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

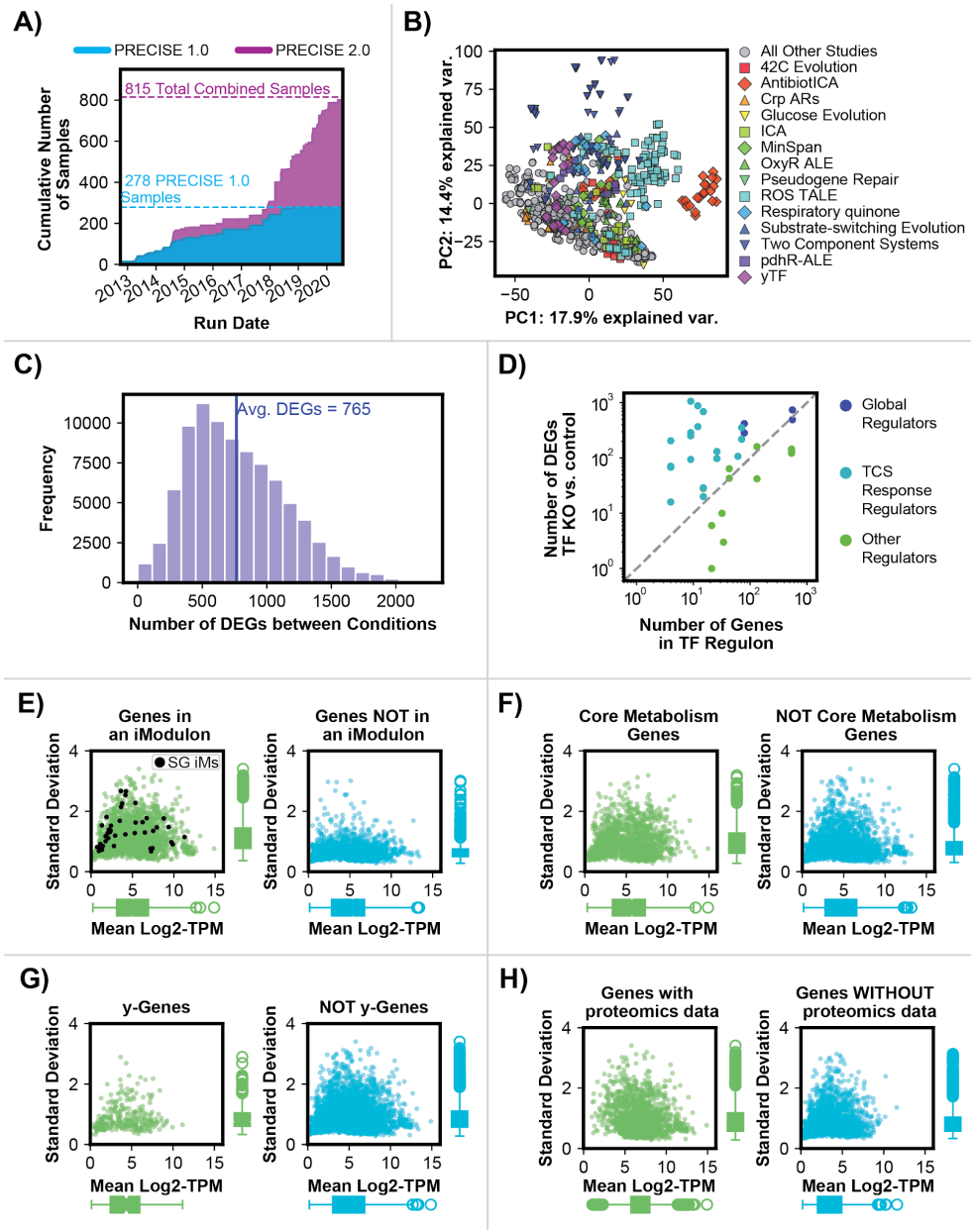

Figure 2: PRECISE 2.0 reveals broad gene expression trends in the *E. coli* transcriptome. **A)** The growth in transcriptomics samples contained in the PRECISE 1.0 to PRECISE 2.0 databases. All transcriptomics samples were generated using the same protocol in the same laboratory. **B)** Principal component plot of PRECISE 2.0. **C)** Numbers of differentially expressed genes (DEGs) determined for all pairs of conditions within PRECISE 2.0 using classical differential gene expression analysis. **D)** Numbers of DEGs for specific pairs of PRECISE 2.0 samples involving transcription factor knockout (TF KO), plotted against the number of genes in the TF's regulon. KO of two-component system (TCS) response regulators tends to produce hundreds of DEGs despite small direct regulon size. **E-H)** Breakdown of gene expression and expression variance by category. SG iMs = single-gene iModulons.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.08.439047; this version posted April 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

PRECISE 2.0 highlights both global and environment-specific gene expression patterns

PRECISE 2.0 provides a high-level view of absolute expression and expression variance across the *E. coli* genome. To gain further insight into the processes that contribute to overall expression and variability, we compared genes based on their presence or absence in four categories.

First, we compared genes that were part of at least one iModulon against genes not in any iModulon, since expression variance is key to detecting independent regulatory signals with ICA. Although genes in iModulons have higher expression variation than genes not in iModulons (P=0.00, Mann-Whitney U test, m=2069, n=2142), average expression itself does not depend significantly on iModulon membership (P = 0.32) (**Figure 2E**). Next, we also found that the expression of metabolic gene s (as defined by the most up-to-date metabolic model, iML1515 (31)) was significantly higher than that of non-metabolic genes (P=3.52E-14, m=1478, n=2733) (**Figure 2F**).

We also compared the expression distribution of uncharacterized genes (referred to as y-genes in *E. coli* (32)) to genes with known functions. Y-genes have significantly lower expression (P=7.31E-8, m=290, n=3921) than non y-genes, highlighting the lack of transcription in many conditions as a potential reason for these genes' relative lack of annotation. However, expression variance is not significantly different based on y-gene status (P=0.47) (**Figure 2F**). Thus, 144 of 290 y-genes (50%) are actually found in iModulons, highlighting the potential for iModulons to uncover putative functions for these uncharacterized genes. Finally, we observed that genes for which proteomics data is available (33, 34) have significantly higher expression (P=0.00, m=1932, n=2279), which is consistent with a known bias towards higher-expressed genes amongst proteomics samples (**Figure 2G**).

iModulons extracted from PRECISE 2.0 summarize systems-level regulatory and biological processes

PRECISE 2.0 can be decomposed into 218 iModulons using ICA. As previously shown, these iModulons map closely onto experimentally-determined regulons, or groups of co-regulated genes (1).

The 218 iModulons extracted from PRECISE 2.0 reconstruct 79% of the total variance in the dataset. 116 of these iModulons are classified as Regulatory, as they are significantly enriched in genes known to belong to the corresponding regulon (**Figure 3A**). These Regulatory iModulons explain 58% of the total variance in PRECISE 2.0 The 34 Genomic and 29 Biological iModulons that lack significant regulator enrichment account for another 17% of the variance (**Figure 3B**). Genomic iModulons are associated with genetic interventions such as knockouts, whereas Biological iModulons are enriched with functionally related groups of genes. An additional 37 iModulons captured single variant genes, although these iModulons only account

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.08.439047; this version posted April 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

for 5% of the variance. Finally, a single uncharacterized iModulon accounts for just 0.4% of the variance in PRECISE 2.0. Thus, functional annotation explains >99% of the variance represented by iModulons.

iModulons related to metabolism and stress responses each account for 22% of the variance in PRECISE 2.0, as well as accounting for 58 and 46 of the 116 Regulatory iModulons by count, respectively (**Figure 3C-D**). The 8 iModulons containing genes encoding transcription and translation-related proteins account for a substantial 13% of the variance, highlighting their central role in the cellular response to changing environmental conditions throughout the dataset. Given the inclusion of a project targeting two-component systems in PRECISE 2.0, 19 two-component system iModulons are extracted from PRECISE 2.0. Thus, ICA with a large dataset like PRECISE 2.0 can capture sensory functions of an organism through transcriptomic changes, in addition to core metabolic and stress functions.

Certain individual iModulons account for outsize proportions of the overall variance in the dataset. In particular, the RpoS (stringent response sigma factor) and ppGpp (stringent response alarmone) iModulons - both influencing transcription and translation - account for 5.2% and 2.3% of all dataset variance, respectively (**Figure 3E**). These data highlight the ability of these regulators to mobilize large-scale transcriptomic responses. Flagella-related regulators FlhDC and FliA also combine to explain a full 6.9% of the variance.

The decomposition of PRECISE 2.0 into iModulons enables differential iModulon activity (DIMA) analysis, where the differential expression of hundreds to thousands of genes can be summarized by on average 38 differentially activated iModulons (**Figure 3F**). On average, a comparison between two conditions in PRECISE 2.0 yields almost twenty times fewer significantly different iModulon activities than significant DEGs (**Figure 3G**).

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.08.439047; this version posted April 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

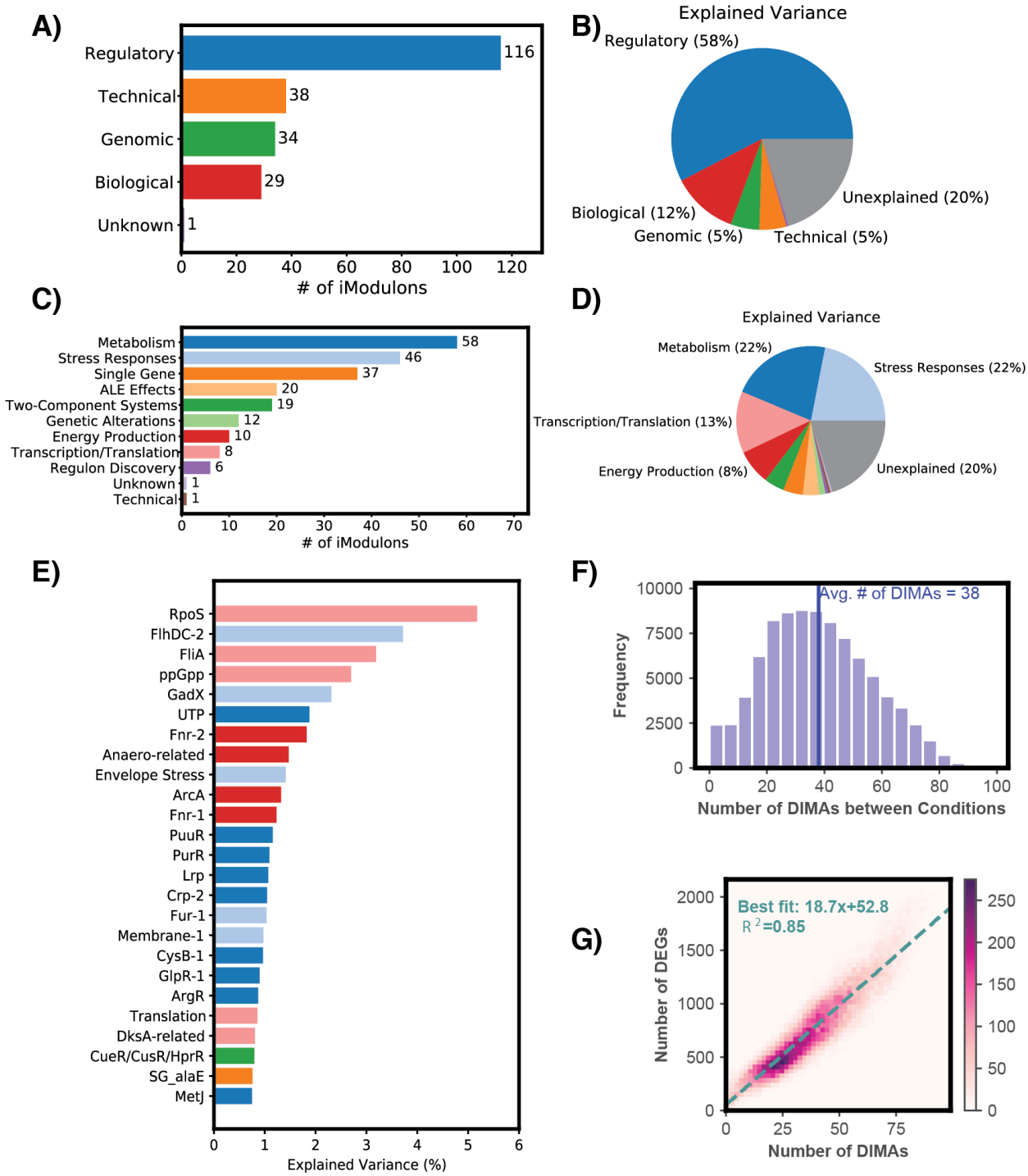

Figure 3: PRECISE 2.0 has 218 iModulons that represent a range of cellular processes. **A)** A breakdown of PRECISE 2.0 iModulons by their annotation category: 'Regulatory' denotes significant enrichment of one or more regulators; 'Technical' includes a single gene or technical artifact iModulon; 'Genomic' includes iModulons related to known genomic interventions (i.e. knockouts or segmental amplifications due to ALE); and 'Biological' includes iModulons containing genes of related function without significant regulator enrichment, or potential new regulons. **B)** iModulon annotation categories by percentage of variance explained. The 218

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.08.439047; this version posted April 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

annotated iModulons together explain 80% of the variance. **C)** More specific functional annotation of the iModulons in PRECISE 2.0. **D)** Explained variances of the functional annotation categories in panel **C**. Color code matches panel **C**. Explanation of transcriptional responses comes through the functional annotation of the iModulons that provides a direct biological explanation of the transcriptomic response. **E)** Top 25 iModulons ranked by % of variance explained. Color code matches panel **C**. **F)** Distribution of differential iModulon activities (DIMAs) for all-to-all condition comparisons. **G)** Relationship between DEGs and DIMAs discovered across all condition comparisons. This graph demonstrates the utility and accuracy of the dimensionality reduction from DEGs to DIMAs, with each condition comparison yielding almost 20 times fewer DIMAs than DEGs.

PRECISE 2.0 enables regulon discovery

Putative TF	Putative Genes in Regulon	Putative Function	TK KO strain included in PRECISE 2.0 experiments?
YmfT	*fur, sulA, intE, xisE, ymfH, ymfJ, ymfT, ymfL, ymfM, ymfN, beeE, jayE, ymfQ, stfE, icdC, recN*	e14 prophage regulator activated by DNA damage	no
YgeV	*ybiY, rcsB, xdhA, xdhB, xdhC, ygeW, ygeX, ygeY, hyuA, ygfK, ssnA, ygfM, xdhD, ygfT, uacT, cpxR*	Nucleoside degradation activated by ethanol treatment	no
YheO	*xisR, hslJ, ldhA, ydfK, yniD, mqo, tusC, tusD, yheO, tnaA, tnaB*	Unknown	yes
YciT	*ybiU, ybiV, ybiW, ybiY, fsaA, yciT*	Unknown	yes
YbaQ	*ybaQ, xisD, ymcF, yohC, yfeC, yfeD*	Unknown	yes
PdeL	*pdeL, gsiB, gsiC, gsiD, pdeI, dgcI*	c-di-GMP control	no

Table 1: Putative regulons discovered from PRECISE 2.0 by ICA

Functional annotation for many putative TFs in *E. coli* still remains elusive (35). Fortunately, iModulons are a powerful tool for the discovery and analysis of new regulons. Our

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.08.439047; this version posted April 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

previous work with PRECISE 1.0 enabled the elucidation of regulons for three previously uncharacterized TFs (YieP, YiaJ/PlaR, and YdhB/AdnB), and expanded the regulons of three known TFs (MetJ, CysB, and KdgR) (1). Many of these regulatory interactions were confirmed through DNA-binding profiles (1, 36, 37). Furthermore, three novel regulons were predicted from iModulons derived from a microarray dataset (24).

iModulons from PRECISE 2.0 have revealed potential regulons for 6 new putative TFs (**Table 1**). Three of these regulons (YheO, YciT, and YbaQ) were identified due to the presence of TF knock-out strains in the compendium. The remaining three regulons (YgeV, YmfT, and PdeL) were identified due to the presence of activating conditions in the database.

The putative YgeV iModulon contains 16 genes, of which 8 are putatively involved in nucleotide transport and metabolism (**Figure 4A**). YgeV is predicted to be a Sigma54-dependent transcriptional regulator, and Sigma54-dependent promoters were previously identified upstream of the *xdhABC* and *ygeWXY* operons, which are in the YgeV iModulon (38). Although the iModulon does not contain the gene *ygeV*, *ygeV* is divergently transcribed from *ygeWXY*. A prior study found that expression of *ygfT* was reduced in a YgeV mutant strain. Since *ygfT* is in the YgeV iModulon, this indicates that YgeV may serve as an activator for the genes in its iModulon.

Although the activity of the YgeV iModulon rarely deviates from the reference condition, it is most active when BaeR or CpxR mutant strains are exposed to ethanol (**Figure 4B**). Therefore, we predict that the TF YgeV responds (directly or indirectly) to ethanol to activate genes related to purine catabolism, and is repressed by the two-component systems BaeRS and CpxAR.

The putative YmfT iModulon contains 14 of the 23 genes in the e14 prophage, including *ymfT* (39) (**Figure 4C**). This iModulon differs from the e14-excision iModulon, whose activities clearly denote strains lacking e14 prophage gene expression (**Figure 4D**). The putative YmfT iModulon is most active in strains lacking the ferric uptake regulator Fur, or in strains challenged by oxidative stress through hydrogen peroxide (**Figure 4E**). Absence of Fur leads to overproduction of iron uptake proteins, oxidative damage, and subsequently mutagenesis (40). Therefore, we predict that YmfT responds to oxidative stress to regulate the genes in the e14 prophage.

The putative PdeL iModulon contains 6 genes across three operons (**Figure 4F**). Five genes (*gsiB*, *gsiC*, *gsiD*, *pdeI*, and *dgcI*) have positive gene weights, whereas *pdeL* has a negative gene weight. PdeL is a dual transcriptional regulator and c-di-GMP phosphodiesterase that activates its own transcription (41). Since the gene weight of *pdeL* is opposite of the other genes in the iModulon, we propose that the remaining genes may be repressed by PdeL.

These three examples illustrate the potential for iModulons to predict new regulons.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.08.439047; this version posted April 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

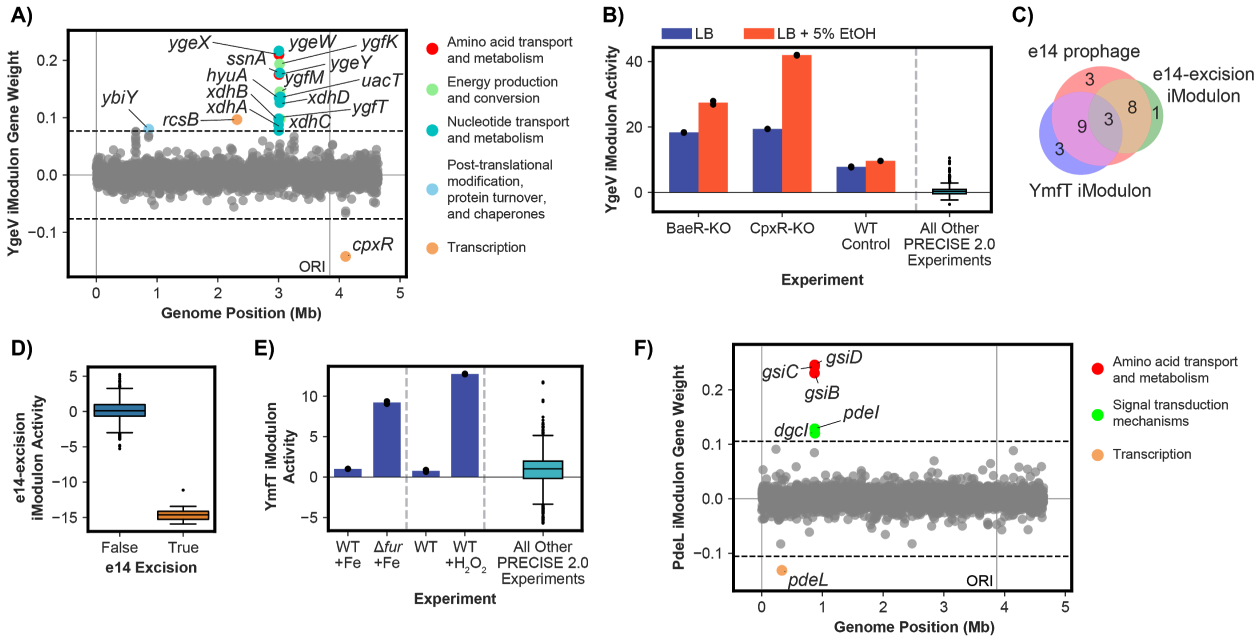

Figure 4: iModulons predict the putative YgeV, YmfT, and PdeL regulons. **A)** iModulon gene weights for the putative YgeV iModulon vs genome position. **B)** Activity of the YgeV iModulon in different media conditions. Each colored bar is the mean of two biological replicates (shown as individual black points). **C)** Venn diagram comparing genes in the e14 prophage, genes in the e14-excision iModulon, and the putative YmfT iModulon **D)** iModulon activities of the e14-excision iModulon, separated between strains with the e14 prophage excised vs. wild-type strains. **E)** Activity of the YmfT iModulon in different media conditions. Each bar is the mean of two biological replicates (shown as black points). **F)** iModulon gene weights for the putative PdeL iModulon vs genome position.

Discussion

In this work, we establish PRECISE 2.0, a high-fidelity *E. coli* expression compendium that enables the discovery of independently-modulated regulatory signals. We use PRECISE 2.0's unprecedented quality and scale to deliver meaningful insights into the regulatory dynamics of *E. coli* at multiple scales. First, we find that gene expression levels and variance across the dataset are differentiated by factors such as measurability in the proteome, functional characterization, and membership in an iModulon. Moreover, we present 218 iModulons that explain nearly 80% of the total variance in our dataset. Over half of these iModulons correspond to known regulons. iModulons derived from PRECISE 2.0 cover the full range of cellular processes, from sensory two-component systems to core metabolic pathways to translation. Differential iModulon activity analysis also greatly simplifies differential expression analysis; with an average of twenty times fewer significantly differential expressions to analyze, DIMA analysis simplifies systems-level analysis of transcriptomic changes.

Perhaps most importantly, PRECISE 2.0 enables us to discover and partially characterize putative regulons for predicted TFs. We demonstrate this capability by assigning a

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.08.439047; this version posted April 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

putative function in either ethanol stress tolerance or nucleotide metabolism to the YgeV regulon, based on the YgeV iModulon activation pattern and backed by ChIP-exo binding data. In particular, this activation coincides with knockouts of two-component system response regulators BaeR and CpxR; thus, YgeV's role in nucleotide metabolism upon ethanol stress response may arise as a compensatory mechanism following inactivation of these more prominent TCS regulators. The specificity of this activating condition may play a role explaining why the functions of this regulator and the genes in its regulon remain unknown.

PRECISE 2.0's success at uncovering *E. coli*'s transcriptome demonstrates the power of ICA and iModulon analysis for the systems-level analysis of the transcriptional regulatory network. We find that by increasing the size of our dataset 3-fold, we have more than doubled the number of discovered iModulons. In addition, we have retained nearly all of the iModulons extracted from PRECISE 1.0, indicating that the iModulons represent fundamental regulatory modes, and not dataset-specific artifacts (**Figure S1**). In some cases, we have discovered multiple iModulons from PRECISE 2.0 that correspond to a single PRECISE 1.0 iModulon. Here, we see an opportunity to refine the sometimes broad definition of a regulon to its most incisive and functionally-relevant form, especially for global regulators that can have hundreds of regulon members.

PRECISE 2.0 adds to a series of successes using ICA and iModulons to characterize bacterial transcriptional regulatory networks. While this dataset was generated entirely from a single laboratory, many more transcriptomics datasets are publicly available for *E. coli* and other common microorganisms from the NCBI's Sequence Read Archive. We have previously shown that ICA can discover stable iModulon cohorts from combined datasets (24). Thus, the time is right to apply these mature analytical techniques to a wider range of organisms for which a critical quantity of transcriptional data is available. We believe that our analytical pipeline for generating, curating, decomposing, and analyzing PRECISE 2.0 can be readily applied to many other microorganisms, with the potential to yield equally impactful insights into those organisms' regulatory network structure.

Methods

Compiling PRECISE 2.0

PRECISE 2.0 consists of all data in PRECISE 1.0 (1), along with data from the following BioProject accession numbers: PRJNA546062, PRJNA559161, PRJNA560374, PRJNA689602, PRJNA704556 and PRJNA601908.

Data was processed using a Nextflow (42) pipeline designed for processing microbial RNA-seq datasets (https://github.com/avsastry/modulome-workflow), and run on Amazon Web Services (AWS) Batch.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.08.439047; this version posted April 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

First, raw read trimming was performed using Trim Galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) with the default options, followed by FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) on the trimmed reads. Next, reads were aligned to the genome using Bowtie (43). The read direction was inferred using RSEQC (44) before generating read counts using featureCounts (45). Finally, all quality control metrics were compiled using MultiQC (46) and the final expression dataset was reported in units of log-transformed Transcripts per Million (log-TPM).

PRECISE 2.0 and associated data files can be found at https://github.com/SBRG/precise2.

Computing the optimal number of robust Independent Components

To compute the optimal independent components, an extension of ICA was performed on the RNA-seq dataset as described in McConn et al. (unpublished).

Briefly, the scikit-learn (v0.23.2) (47) implementation of FastICA (48) was executed 100 times with random seeds and a convergence tolerance of 10^{-7}. The resulting independent components (ICs) were clustered using DBSCAN (49) to identify robust ICs, using an epsilon of 0.1 and minimum cluster seed size of 50. To account for identical with opposite signs, the following distance metric was used for computing the distance matrix:

$$d_{x,y} = 1 - ||\rho_{x,y}||$$

where $\rho_{x,y}$ is the Pearson correlation between components x and y. The final robust ICs were defined as the centroids of the cluster.

Since the number of dimensions selected in ICA can alter the results, we applied the above procedure to the *B subtilis* dataset multiple times, ranging the number of dimensions from 10 to 260 (i.e. the approximate size of the dataset) with a step size of 10. To identify the optimal dimensionality, we compared the number of ICs with single genes to the number of ICs that were correlated (Pearson R > 0.7) with the ICs in the largest dimension (called “final components”). We selected the number of dimensions where the number of non-single gene ICs was equal to the number of final components in that dimension.

Identifying differentially expressed genes (DEGs)

Differentially expressed genes (DEGs) were identified using the *DESeq2* package (50) on the PRECISE 2.0 RNA-seq dataset. Genes with a $\log_2$ fold change greater than 1.5 and a false discovery rate (FDR) value less than 0.05 were considered to be differentially expressed genes. Genes with p-values assigned “NA” based on extreme count outlier detection were not considered as potential DEGs. The number of DEGs was computed for each unique pair of conditions in the PRECISE 2.0 compendium.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.08.439047; this version posted April 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Identifying differential iModulon activities (DIMAs)

Differentially activated iModulons were computed with a similar process as previously detailed (1). For each iModulon, the average activity of the iModulon between biological replicates, if available, was computed. Then, the absolute value of the difference in iModulon activities between the two conditions was compared to the fitted log-normal distribution of all differences in activity for the iModulon. iModulons that had an absolute value of activity greater than 5, and an FDR below 0.05 were considered to be significant. The number of DIMAs was computed for each unique pair of conditions in the PRECISE 2.0 compendium.

Computing iModulon enrichments

The TRN was taken from RegulonDB (51). iModulon enrichments against known regulons were computed using Fisher's Exact Test, with the FDR controlled at 10^{-5} using the Benjamini-Hochberg correction. Fisher's Exact Test was used to identify GO and KEGG annotations as well, with an FDR < 0.01. By default, iModulons were compared to all possible single regulons and all possible combinations of two regulons (both union and intersection) to yield significant enrichments. The regulons used by default consisted of only strong and confirmed evidence regulatory interactions, per RegulonDB. When multiple significant enrichments were available, the enrichment with the lowest adjusted *P* value was used for annotation. If no significant enrichments were available, the following adjustments were used, in this order: relax evidence requirement to include weak evidence regulatory interactions; search only for single regulon enrichments; allow up to 3 regulons to be combined for enrichment. If none of these adjustments yielded a significant enrichment, the iModulon was annotated as non-regulatory. iModulons annotated not using the default enrichment setup are noted in the iModulon table available as part of the included dataset.

Code Availability

Code and Jupyter notebooks are available at: https://github.com/SBRG/precise2

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.08.439047; this version posted April 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Figures

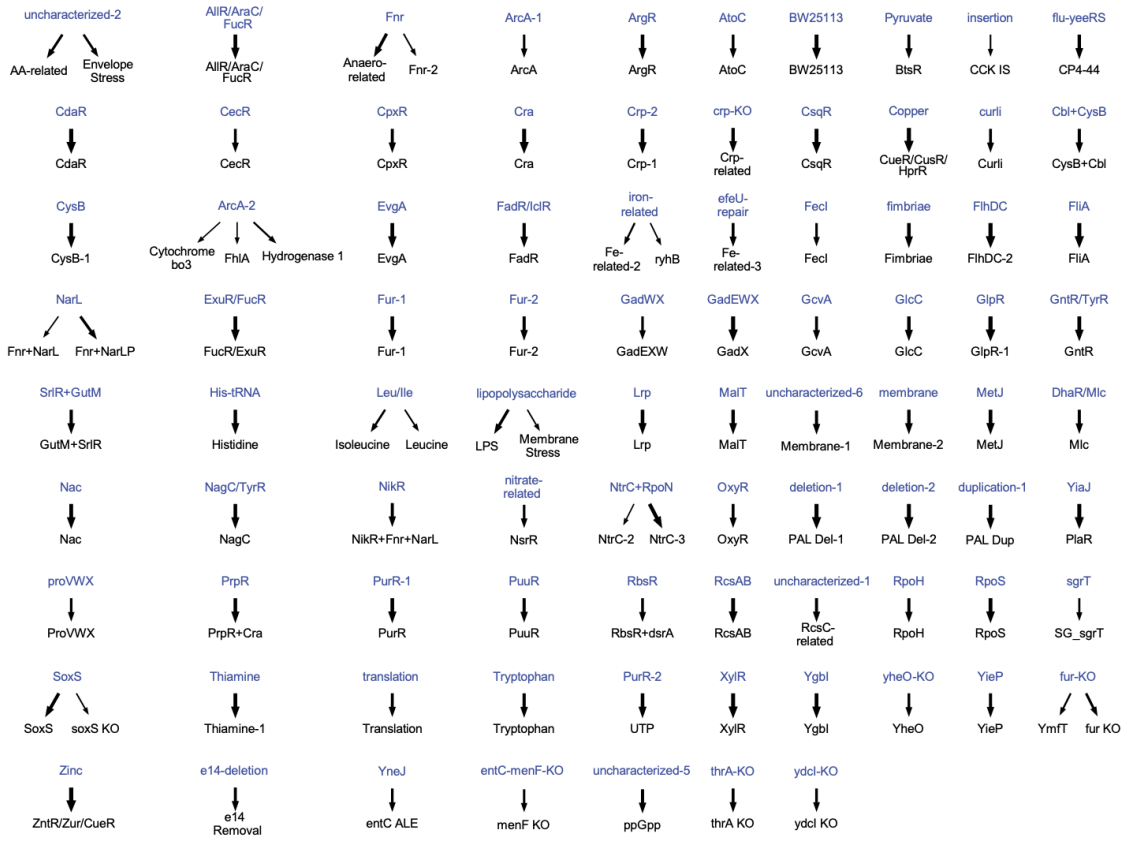

Figure S1: Comparison of PRECISE 1.0 iModulons to PRECISE 2.0 iModulons

Bibliography

1. A. V. Sastry, *et al.*, The Escherichia coli transcriptome mostly consists of independently regulated modules. *Nat. Commun.* **10**, 5536 (2019).

2. M. Ziemann, A. Kaspi, A. El-Osta, Digital expression explorer 2: a repository of uniformly processed RNA sequencing data. *Gigascience* **8** (2019).

3. D. P. Leader, S. A. Krause, A. Pandit, S. A. Davies, J. A. T. Dow, FlyAtlas 2: a new version of the Drosophila melanogaster expression atlas with RNA-Seq, miRNA-Seq and sex-specific data. *Nucleic Acids Res.* **46**, D809–D815 (2018).

4. ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

5. J. Zrimec, *et al.*, Deep learning suggests that gene expression is encoded in all parts of a

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.08.439047; this version posted April 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

co-evolving interacting gene regulatory structure. *Nat. Commun.* **11**, 6141 (2020).

6. Z. Zhang, *et al.*, Deep-learning augmented RNA-seq analysis of transcript splicing. *Nat. Methods* **16**, 307–310 (2019).

7. M. S. Kwon, B. T. Lee, S. Y. Lee, H. U. Kim, Modeling regulatory networks using machine learning for systems metabolic engineering. *Curr. Opin. Biotechnol.* **65**, 163–170 (2020).

8. W. Saelens, R. Cannoodt, Y. Saeys, A comprehensive evaluation of module detection methods for gene expression data. *Nat. Commun.* **9**, 1090 (2018).

9. A. E. Teschendorff, M. Journée, P. A. Absil, R. Sepulchre, C. Caldas, Elucidating the altered transcriptional programs in breast cancer using independent component analysis. *PLoS Comput. Biol.* **3**, e161 (2007).

10. J. M. Engreitz, *et al.*, Content-based microarray search using differential expression profiles. *BMC Bioinformatics* **11**, 603 (2010).

11. C. G. Urzúa-Traslaviña, *et al.*, Improving gene function predictions using independent transcriptional components. *Nat. Commun.* **12**, 1464 (2021).

12. A. Biton, *et al.*, Independent component analysis uncovers the landscape of the bladder tumor transcriptome and reveals insights into luminal and basal subtypes. *Cell Rep.* **9**, 1235–1245 (2014).

13. P. V. Nazarov, *et al.*, Deconvolution of transcriptomes and miRNomes by independent component analysis provides insights into biological processes and clinical outcomes of melanoma patients. *BMC Med. Genomics* **12**, 132 (2019).

14. W. Wang, *et al.*, Independent component analysis based gene co-expression network inference (ICAnet) to decipher functional modules for better single-cell clustering and batch integration. *Nucleic Acids Res.* (2021) https:/doi.org/10.1093/nar/gkab089.

15. C. Trapnell, *et al.*, The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).

16. K. Rychel, A. V. Sastry, B. O. Palsson, Machine learning uncovers independently regulated modules in the Bacillus subtilis transcriptome. *Nat. Commun.* **11**, 6338 (2020).

17. S. Poudel, *et al.*, Revealing 29 sets of independently modulated genes in Staphylococcus aureus, their regulators, and role in key physiological response. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 17228–17239 (2020).

18. K. Rychel, *et al.*, iModulonDB: a knowledgebase of microbial transcriptional regulation derived from machine learning. *Nucleic Acids Res.* **49**, D112–D120 (2021).

19. J. Tan, *et al.*, Independent component analysis of E. coli's transcriptome reveals the cellular processes that respond to heterologous gene expression. *Metab. Eng.* **61**, 360–368 (2020).

20. K. S. Choudhary, *et al.*, Elucidation of Regulatory Modes for Five Two-Component Systems

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.08.439047; this version posted April 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

in Escherichia coli Reveals Novel Relationships. *mSystems* **5** (2020).

21. A. Anand, *et al.*, OxyR Is a Convergent Target for Mutations Acquired during Adaptation to Oxidative Stress-Prone Metabolic States. *Mol. Biol. Evol.* **37**, 660–667 (2020).

22. A. Anand, *et al.*, Adaptive evolution reveals a tradeoff between growth rate and oxidative stress during naphthoquinone-based aerobic respiration. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 25287–25292 (2019).

23. C. Mejía-Almonte, *et al.*, Redefining fundamental concepts of transcription initiation in bacteria. *Nat. Rev. Genet.* **21**, 699–714 (2020).

24. A. V. Sastry, *et al.*, Independent component analysis recovers consistent regulatory signals from disparate datasets. *PLoS Comput. Biol.* **17**, e1008647 (2021).

25. A. Sastry, *et al.*, Decomposition of transcriptional responses provides insights into differential antibiotic susceptibility. *bioRxiv*, 2020.05.04.077271 (2020).

26. B. Du, *et al.*, Adaptive laboratory evolution of Escherichia coli under acid stress. *Microbiology* **166**, 141–148 (2020).

27. K. Chen, *et al.*, Bacterial fitness landscapes stratify based on proteome allocation associated with discrete aero-types. *PLoS Comput. Biol.* **17**, e1008596 (2021).

28. D. McCloskey, *et al.*, Evolution of gene knockout strains of E. coli reveal regulatory architectures governed by metabolism. *Nat. Commun.* **9**, 3796 (2018).

29. T. E. Sandberg, R. Szubin, P. V. Phaneuf, B. O. Palsson, Synthetic cross-phyla gene replacement and evolutionary assimilation of major enzymes. *Nat Ecol Evol* **4**, 1402–1409 (2020).

30. J. M. Skerker, M. S. Prasol, B. S. Perchuk, E. G. Biondi, M. T. Laub, Two-component signal transduction pathways regulating growth and cell cycle progression in a bacterium: a system-level analysis. *PLoS Biol.* **3**, e334 (2005).

31. J. M. Monk, *et al.*, iML1515, a knowledgebase that computes Escherichia coli traits. *Nat. Biotechnol.* **35**, 904–908 (2017).

32. S. Ghatak, Z. A. King, A. Sastry, B. O. Palsson, The y-ome defines the 35% of Escherichia coli genes that lack experimental evidence of function. *Nucleic Acids Res.* **47**, 2446–2454 (2019).

33. A. Schmidt, *et al.*, The quantitative and condition-dependent Escherichia coli proteome. *Nat. Biotechnol.* **34**, 104–110 (2016).

34. D. Heckmann, *et al.*, Kinetic profiling of metabolic specialists demonstrates stability and consistency of in vivo enzyme turnover numbers. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 23182–23190 (2020).

35. Y. Gao, *et al.*, Systematic discovery of uncharacterized transcription factors in Escherichia coli K-12 MG1655. *Nucleic Acids Res.* (2018) https:/doi.org/10.1093/nar/gky752.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.08.439047; this version posted April 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

36. I. A. Rodionova, *et al.*, Synthesis of the novel transporter YdhC, is regulated by the YdhB transcription factor controlling adenosine and adenine uptake. *bioRxiv*, 2020.05.03.074617 (2020).

37. I. A. Rodionova, Y. Gao, A. V. Sastry, J. Monk, R. Szubin, PtrR (YneJ) is a novel E. coli transcription factor regulating the putrescine stress response and glutamate utilization. *bioRxiv* (2020).

38. L. Reitzer, B. L. Schneider, Metabolic Context and Possible Physiological Themes of ς54-Dependent Genes in Escherichia coli. *Microbiol. Mol. Biol. Rev.* **65**, 422–444 (2001).

39. P. Mehta, S. Casjens, S. Krishnaswamy, Analysis of the lambdoid prophage element e14 in the E. coli K-12 genome. *BMC Microbiol.* **4**, 4 (2004).

40. D. Touati, M. Jacques, B. Tardat, L. Bouchard, S. Despied, Lethal oxidative damage and mutagenesis are generated by iron in delta fur mutants of Escherichia coli: protective role of superoxide dismutase. *J. Bacteriol.* **177**, 2305–2314 (1995).

41. A. Reinders, *et al.*, Expression and Genetic Activation of Cyclic Di-GMP-Specific Phosphodiesterases in Escherichia coli. *J. Bacteriol.* **198**, 448–462 (2016).

42. P. Di Tommaso, *et al.*, Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).

43. B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).

44. L. Wang, S. Wang, W. Li, RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184–2185 (2012).

45. Y. Liao, G. K. Smyth, W. Shi, featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).

46. P. Ewels, M. Magnusson, S. Lundin, M. Käller, MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).

47. F. Pedregosa, *et al.*, Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

48. A. Hyvärinen, Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Netw.* **10**, 626–634 (1999).

49. M. Ester, H.-P. Kriegel, J. Sander, X. Xu, Others, A density-based algorithm for discovering clusters in large spatial databases with noise in *Kdd*, (1996), pp. 226–231.

50. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

51. A. Santos-Zavaleta, *et al.*, RegulonDB v 10.5: Tackling challenges to unify classic and high throughput knowledge of gene regulation in E. coli K-12. *Nucleic Acids Res.* **47**, D212–D220 (2019).

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.08.439047; this version posted April 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.