Published online 13 October 2017 *Nucleic Acids Research, 2017, Vol. 45, No. 21* **12113–12124**
doi: 10.1093/nar/gkx910

The Omics Dashboard for interactive exploration of gene-expression data

Suzanne Paley[1], Karen Parker[2], Aaron Spaulding[1], Jean-Francois Tomb[3], Paul O'Maille[4] and Peter D. Karp[1,*]

[1]Bioinformatics Research Group, SRI International, 333 Ravenswood Ave, Menlo Park, CA 94025, USA, [2]Chrysopylae, Los Altos, CA 94024, USA, [3]Department of Chemical and Biomolecular Engineering, University of Delaware, Newark, DE 19716, USA and [4]Biosciences Division, SRI International, 333 Ravenswood Ave, Menlo Park, CA 94025, USA

Received June 08, 2017; Revised August 31, 2017; Editorial Decision September 22, 2017; Accepted September 27, 2017

ABSTRACT

The Omics Dashboard is a software tool for interactive exploration and analysis of gene-expression datasets. The Omics Dashboard is organized as a hierarchy of cellular systems. At the highest level of the hierarchy the Dashboard contains graphical panels depicting systems such as biosynthesis, energy metabolism, regulation and central dogma. Each of those panels contains a series of X–Y plots depicting expression levels of subsystems of that panel, e.g. subsystems within the central dogma panel include transcription, translation and protein maturation and folding. The Dashboard presents a visual read-out of the expression status of cellular systems to facilitate a rapid top-down user survey of how all cellular systems are responding to a given stimulus, and to enable the user to quickly view the responses of genes within specific systems of interest. Although the Dashboard is complementary to traditional statistical methods for analysis of gene-expression data, we show how it can detect changes in gene expression that statistical techniques may overlook. We present the capabilities of the Dashboard using two case studies: the analysis of lipid production for the marine alga *Thalassiosira pseudonana*, and an investigation of a shift from anaerobic to aerobic growth for the bacterium *Escherichia coli*.

INTRODUCTION

The rise of 'omics' technologies has unleashed a flood of complex, high-resolution datasets, yet major barriers block the transformation of these data into new knowledge and biological insights. To address this challenge, we present a novel software tool for interactive exploration and analysis of omics datasets. The Omics Dashboard is based on a dashboard metaphor and presents the user with a visual read-out of the expression status of cellular systems to enable a scientist to quickly ascertain the state of those systems. The Dashboard harnesses the information-processing power of the human visual system in a way that no previous expression-analysis software has done to enable a scientist to couple biological intuition with significance calculations that can be applied to the Dashboard's input data. The Dashboard addresses four gene-expression data analysis tasks that have not been satisfactorily addressed in the past: facilitating a rapid user survey of how all cellular systems are responding to a given stimulus; enabling the user to quickly find and understand the response of genes within one or more specific molecular systems of interest; gauging the relative gene-expression levels of different cellular systems; and comparing the expression levels of a cellular system with those of its known regulators.

The Dashboard is organized as a hierarchy of cellular systems. At its highest level the Dashboard contains *panels* for cellular systems such as biosynthesis, energy metabolism, and non-metabolic functions (see Figure 1). Each of those panels contains a series of X–Y *plots* depicting the expression levels of genes within the subsystems of that panel, e.g. plots within the Central Dogma panel include Transcription, Translation and Protein Metabolism.

In addition to panels and plots, other visual elements available within the Dashboard include pathway diagrams painted with gene expression data, diagrams showing the operon organization of all genes within a given biological system, and plots of all known regulators for the genes within a given system. The Omics Dashboard can also accommodate proteomics, metabolomics and reaction-flux data (e.g. results of executing metabolic models), but in this article we focus on its analysis capabilities for gene-expression data.

*To whom correspondence should be addressed. Tel: +1 650 859 4358; Fax: +1 650 859 3735; Email: pkarp@ai.sri.com
Disclaimer: The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

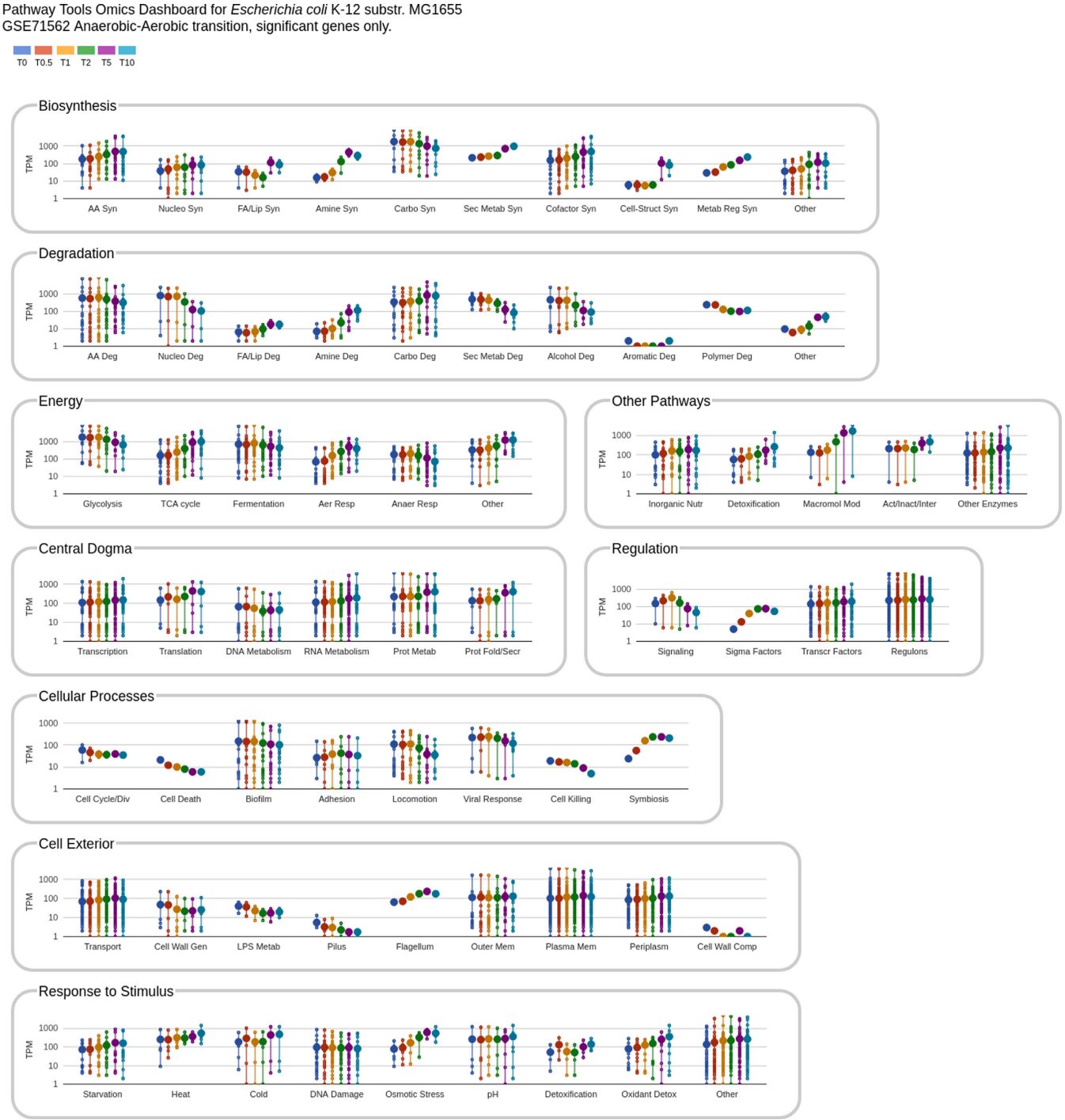

Figure 1. Omics Dashboard showing changes in *Escherichia coli* gene expression in a 10 min time course following a shift to aerobic growth. Normalized average RNA-seq read counts of significantly differentially expressed genes are shown in a log scale on the *Y*-axis. T0 represents samples drawn before the shift to aerobic growth and T0.5, T1, T2, T5 and T10 are samples drawn at 0.5, 1, 2, 5 and 10 min after aeration started at 1 l/min.

In the past the analysis of hundreds of interlocking pathways required a team of experts with insight into different parts of the metabolic system and weeks or months of work. In many cases the scope of the investigation was limited significantly by the amount of time available and the expertise of the investigator(s). In our experience, the Omics Dashboard greatly reduces the amount of time and expertise needed to analyze large datasets, based on our experience analyzing the same gene-expression dataset from a marine alga compared to other tools such as Bioconductor. As a result, an investigator can quickly observe and analyze the functioning of the entire metabolic system without analyzing the expression of single genes because the gene and pathway level details are automatically summarized to a higher

functional level (such as the 'lipid biosynthesis' class in the MetaCyc pathway ontology, which includes many genes in multiple pathways). Once a condition of interest is identified, the nested Dashboard panels allow the user to drill down into successive levels of functional detail, bottoming out at the detailed pathway or gene level.

The ability to observe the gene-expression state of all major biological systems in a graphical depiction gives the investigator a comprehensive overview of each system *before* starting a detailed gene-level analysis. Although we acknowledge that tools such as cluster analysis and enrichment analysis have a top-down component, they do not offer the comprehensive survey and drill-down capabilities that the Dashboard provides. The knowledge of the high-level functional state of the organism can be applied as the investigator analyzes the details of the data. As a result of using a top-down Dashboard analysis, the investigation is faster and has the potential to be more insightful and broader in scope. In addition, it allows the investigator to focus on useful datasets and avoid wasting time on datasets that are not representative of the conditions under investigation. For example, in a commercial strain development application in which hundreds of strains are evaluated, the Omics Dashboard can be configured to identify desirable environmental response features and quickly eliminate unfavorable strains.

Related work

We view the Omics Dashboard as complementary to the range of existing methods (1) for analysis of gene expression data. A typical expression-data analysis workflow includes (i) data normalization and detection of differential expression patterns (2); (ii) clustering of gene expression data (3); (iii) finding individual genes whose differential expression is statistically significant; (iv) finding sets of functionally related genes (e.g. pathways) whose differential expression is statistically significant; and (v) visualizing expression data on diagrams of individual pathways (4), on multi-pathway collages (4), and on full pathway maps (4).

The Omics Dashboard replaces steps (iv) and (v) in the preceding workflow with a broader set of tools that support the following functions. (a) The user can perform a rapid visual survey of the state of most cellular systems at a high level (e.g., to see how the levels of all carbohydrate biosynthesis or of all virulence-related genes are changing). (b) The user can drill down to explore the states of subsystems of interest in more detail, such as graphing the responses of genes in a subsystem or pathway. (c) The user can obtain pathway diagrams painted with expression data for pathways and subsystems of interest. (d) The user can obtain operon diagrams for pathways and subsystems of interest. (e) The user can graph the expression levels of regulators for a pathway or subsystem of interest. (f) The user can perform enrichment analysis to determine which pathways and subsystems are enriched for genes whose expression changes are statistically significant.

We show that the Omics Dashboard complements existing statistical techniques in several ways. Because gene expression studies yield large datasets, significance tests are used to filter those datasets to focus the user's attention on a more cognitively manageable number of genes (by significance tests we mean both tests of statistical significance, and calculation of fold-change). We argue that these traditional filters are very useful, but are insufficient, and that the Dashboard provides additional functional filters. We show that using simple significance tests alone the user can miss patterns in gene-expression data that can be detected with the Dashboard. We argue for combining these significance tests with the Dashboard's graphical tools. For example, filtering with significance tests can still return large sets of differentially regulated genes (e.g. for the *Escherichia coli* example in Section 4.2, 639 genes were identified as significantly differentially expressed); the Dashboard enables the user to functionally filter those reduced but still large gene sets. Second, although it is common to provide two thresholds for identifying significantly differentially expressed genes (such as a 2-fold change in expression and a P-value ≤ 0.05), there is likely to be no such pair of thresholds that are meaningful for all genes in all conditions in all organisms. The Dashboard enables the user to examine a functionally filtered set of genes without significance filtering (e.g. the sigma factors of Section 4.2) to visually detect biologically significant patterns.

Computation of Gene Ontology (GO) term enrichment and pathway enrichment are commonly used methods that are implemented in many software tools including AmiGO (5), PANTHER (6), DAVID (7) and Pathway Tools (8). Enrichment analysis is useful but is not a complete solution to the gene-expression analysis problem. For example, running a PANTHER enrichment analysis via http://www.geneontology.org/page/go-enrichment-analysis on a set of five genes all contained in one *E. coli* pathway (tryptophan biosynthesis) produces a list of 31 GO biological-process categories that are significantly enriched (albeit with tryptophan biosynthesis first in the list), which we argue is an excessively complex result for the user to sift through, particularly for such a small starting gene set. Other enrichment tools produce similar results. Furthermore, enrichment analysis tools ignore the gene expression levels themselves, both in the computation of enrichment and in the presentation of results because they consider only whether a significant change occurred—the user cannot tell from the results presentation whether a pathway is upregulated or downregulated, much less the amount of change.

The preceding enrichment example also illustrates a drawback of presenting results according to the GO hierarchies, which were developed based on ontological principles that resulted in deep, complicated, non-intuitive hierarchies that produce complex result displays. Therefore we designed Dashboard-specific system hierarchies that are more concise than the hierarchies within GO, but in many cases are defined based on selected concepts from GO. The GO contains over 29 000 terms in its biological process hierarchy, but the Dashboard contains subsystems (plots) defined from only 74 GO process terms plus 6 GO cellular component terms, 355 MetaCyc (9) pathway classes and a large number of MetaCyc pathway instances (only a fraction of which are likely present for any given organism). We tried to keep those processes/subsystems that would cover critical biological areas and not overwhelming the user with complexity. We solicited feedback from a number of biolo-

gists regarding the Dashboard organization. Users can customize the Dashboard to contain additional subsystems if desired.

MATERIALS AND METHODS

The Dashboard is a component of the Pathway Tools software (8). Pathway Tools powers the BioCyc website, and Pathway Tools is used to construct the organism-specific databases, called Pathway/Genome Databases (PGDBs), that make up the BioCyc database collection. The panel and plot visualizations within the Dashboard are implemented using Google Charts https://developers.google.com/chart/, which in turn is implemented in Javascript. The Dashboard also contains client-side (web browser) components implemented in Javascript, and server-side components implemented in Common Lisp. The pathway and operon diagrams displayed by the Dashboard are generated by existing Pathway Tools algorithms, as is the enrichment analysis operation within the Dashboard.

The Dashboard software defines a mapping from each subsystem (plot) to one or more pathways and/or GO terms. When the Dashboard displays each plot, it dynamically retrieves gene or metabolite lists for each plot from the PGDB for the current organism. For example, it issues PGDB queries to determine what genes (if any) exist in the current organism for the pathway(s) or GO term(s) associated with each plot. More specifically, Dashboard-panel gene groups are obtained from pathways in the PGDB via a Pathway Tools built-in query that returns all genes coding for enzymes catalyzing reactions within a specified metabolic pathway. Similarly, Pathway Tools provides a built-in query for obtaining all genes annotated to a given GO term. When displaying the window of regulators, the Dashboard issues a built-in Pathway Tools query for obtaining a list of all transcriptional regulators of a given gene.

PGDBs within the BioCyc collection are highly variable in terms of the completeness of their GO term annotations and regulatory interactions, but the Dashboard is best suited for use with PGDBs with significant numbers of GO terms and regulatory interactions. Table 1 lists the 16 BioCyc databases containing more than 3000 GO term annotations, and the 10 BioCyc databases containing more than 500 transcriptional regulatory interactions. For our next BioCyc release in 2017 we have downloaded available GO term annotations from UniProt for all of the 42 Tier 2 BioCyc PGDBs (Tier 2 PGDBs have undergone a moderate amount of manual curation). GO annotations will be available for even more organisms in the future.

Given a set of genes (the user could specify all genes, or a set of genes whose changes are computed to be statistically significant), the Dashboard computes an enrichment p-value for every subsystem using a Lisp implementation of Grossmann's parent–child-union analysis, a variation of the Fisher-exact test in which the enrichment of a given subsystem is determined relative to its parent subsystem rather than to the entire population (10). An optional multiple-hypothesis correction (options are Bonferroni, Benjamini-Hochberg or Benjamini-Yekutieli corrections, with no correction being the default) may be applied. The enrichment p-value is then converted to an enrichment score, $-log(P\text{-value})$.

Experimental designs that the Dashboard should be appropriate for include time-course experiments, dose-response experiments, and experiments that vary growth conditions. The Dashboard performs well up to 20 columns of data, but the display becomes cramped; that effect will be lessened on larger monitors.

This paper uses the analysis of two datasets to illustrate the application of the Dashboard toolset: a genome-wide transcriptome analysis of *Thalassiosira pseudonana*, and an *E. coli* gene-expression analysis of a 10 min time course following a shift from anaerobic to aerobic growth conditions.

Mock *et al*. performed a genome-wide transcriptome analysis on *T. pseudonana* strain CCMP 1335 under five different environmental conditions: low nitrate (low N), low silicic acid (low Si), low iron (Low Fe), low temperature (4°C) and high pH (9.4), with nutrient-replete cultures serving as reference conditions (11). Cultures were maintained in natural seawater that had been autoclaved and supplemented with $2 \times f/2$ nutrients minus one of the limited nutrient (Si, Fe or N) at 20°C and 100 μmol of photons $m^{-2}s^{-1}$. $F/2$ provides the major nutrients including N, Si and P, as well as trace metals and vitamins (12). Alkaline pH condition was obtained by increasing the pH of $2 \times f/2$ seawater to 9.4 by adding 1M NaOH. Temperature limitation was achieved by transferring a culture maintained in nutrient-replete $2 \times f/2$ seawater at 20°C to 4°C for 24 h (11). All limitation experiments were conducted in parallel with nutrient-replete cultures. Cells were harvested for RNA when the growth rate began to decrease significantly relative to the control cultures. Differentially expressed genes include those that have a Bayesian *t*-test *P*-value ≤ 0.05, and a $\geq$2-fold difference in mRNA levels with respect to the control samples. Data are available under GEO accession GSE9697.

Methods from von Wulffen *et al*. (13): *Escherichia coli* K–12 strain W3110 was used in this study. Cells were grown anaerobically in defined medium at pH7 and 37°C in a stirred 3-l bio-reactor until the culture reached an OD (600 nm) of 3. At that point, the first three replicate samples were drawn and aeration was started subsequently at 1 l/min. At 0.5, 1, 2, 5 and 10 min after the onset of aeration additional samples were drawn from the three replicates.

Analysis of von Wulffen *et al*. data performed for this publication: raw gene counts were obtained from the GEO database (accession GSE71562). Replicates were averaged and were next normalized using the TPM (Transcripts per Kilobase Million) approach (14). Genes that had zero counts in more than 15% of the samples were removed from further analysis; in addition, two genes (*ssrA* and *rnpB*) with high expression values that compressed the scales of two panels were removed; see Supplementary File S1. Differentially expressed genes in the samples at 0.5, 1, 2, 5 and 10 minutes were identified with respect to zero time samples by applying a paired T.TEST analysis (computed with Excel). Samples with statistically significant changes (*P*-value ≤ 0.05) and at least a 2-fold increase or decrease in gene expression (for any time point relative to time zero) were retained; see Supplementary File S2. At 0.5 min, 33 genes were found to be differentially expressed versus 487 at 10

Table 1. BioCyc databases containing more than 3000 GO term annotations and more than 500 transcriptional regulatory interactions

BioCyc Database	# GO terms	# Regulatory interactions
Anopheles gambiae	3688	
Bacillus subtilis subtilis 168	3505	788
Burkholderia cepacia complex	3415	
Corynebacterium glutamicum ATCC 13032 (Kalinowski03)		698
Cronobacter turicensis z3032	3611	
Cupriavidus taiwanensis	3259	
E. coli B str. REL606	4802	
E. coli K-12 substr. MG1655 (EcoCyc)	5739	3419
E. coli K-12 substr. W3110	5181	
E. coli O157:H7 str. EDL933		611
E. coli O157:H7 str. EC4115	3438	
Homo sapiens	21 237	
Mycobacterium tuberculosis H37Rv	3838	
Ralstonia solanacearum CMR15	3225	
Saccharomyces cerevisiae S288c	8772	1740
Salmonella enterica enterica serovar Typhi str. CT18		1740
Salmonella enterica enterica serovar Typhimurium str. LT2		1828
Shewanella baltica OS117	3145	
Shewanella oneidensis MR-1		625
Shigella flexneri 2a str. 2457T		561
Shigella flexneri 2a str. 301		1302
Sinorhizobium fredii HH103	3191	
Yersinia enterocolitica palearctica Y11	3490	
Yersinia pestis KIM10+		847

min in aerobic growth. Over the 10-min period 639 genes were identified as significantly differentially expressed; their read counts summed to 12% of the total normalized read counts.

RESULTS

Invoking the Omics Dashboard from BioCyc

The Omics Dashboard is one of many tools within the BioCyc.org website. Follow these steps to invoke the Omics Dashboard from within BioCyc.

i) Open URL https://biocyc.org/ in Chrome (preferred) or Firefox
ii) Select the organism for which omics data will be analyzed with the Dashboard by clicking "change organism database" in the upper right corner (e.g., to analyze human omics data, select organism *Homo sapiens*)
iii) Enter the Dashboard with command Analysis → Omics Dashboard
iv) Follow the instructions on the resulting Omics Dashboard page to load a dataset

Users can interact with the Omics Dashboard online in conjunction with the von Wulffen *et al*. dataset discussed in Section 4.2 at this URL: https://biocyc.org/dashboard/dashboard.html?st=biocyc17-463-3702945483.

The basic dashboard layout

The Dashboard is organized into a set of panels, each of which contains a set of plots (Figure 1). The set of subsystem plots visible within each panel is dependent on the current organism and its molecular machinery. For example, for photosynthetic organisms a Photosynthesis plot will automatically appear in the Energy Metabolism panel (driven by the existence of photosynthetic pathways in the Pathway Tools PGDB (Pathway/Genome Database) for that organism); for humans, plots for Hormone Biosynthesis and Degradation will automatically appear in the Biosynthesis and Degradation panels respectively.

We recommend that normalization and significance calculations be applied before importing data into the Dashboard. A comparison of Figure 1, which contains only the significantly differentially regulated genes from von Wulffen *et al*., with Supplementary Figure S1, which contains all genes from von Wulffen *et al*., indicates that focusing on the significantly differentially regulated genes reveals more patterns.

The user can drill down into a given plot by clicking on it to reveal more details, for example, clicking on 'AA Biosyn' within the Biosynthesis panel produces a new panel, each of whose plots summarizes the biosynthesis of one amino acid (Figure 2A). Continuing to drill down on plots of interest will eventually produce a plot of all genes within a *base* subsystem (a subsystem that contains no subsystems) such as a metabolic pathway. For example, clicking on the Arg plot in Figure 2A plots the expression levels of each individual gene involved in L-arginine biosynthesis (Figure 2B).

The plots within a panel share common X and Y axes. In Figures 1 and 2, six vertical lines are present within each plot (subsystem). For this gene-expression dataset, each of the vertical lines represents a different time point at which a gene-expression measurement was taken. In general, the vertical lines can also represent different experimental conditions.

Each dot shown along a vertical line corresponds to an expression value for a single gene. The set of genes present on a given vertical line corresponds to all genes within the cellular subsystem represented by that plot. For example, for the Ala plot within the Amino Acid Biosynthesis panel (Figure 2), the set of genes shown consists of all genes coding for enzymes within the pathway for L-alanine biosynthesis. A given gene can be assigned to more than one subsystem. The large dot shown on each line is the average of all Y-values present on that line; it enables the eye to easily track the variation in average expression of genes within a subsystem by visually following the position of the large

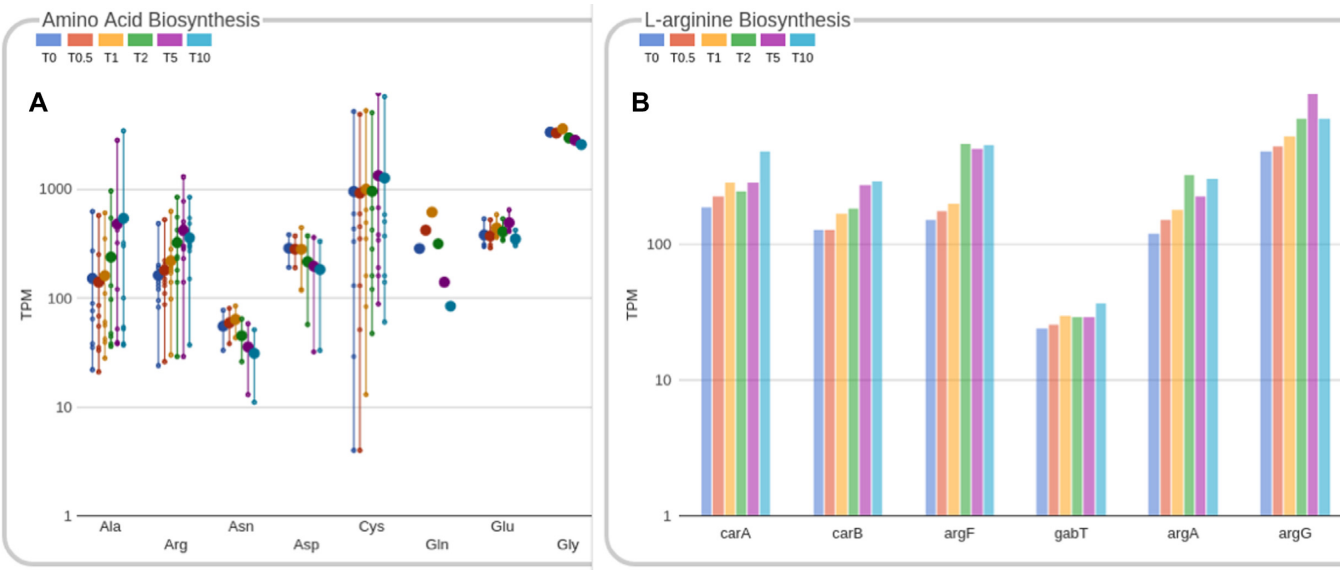

Figure 2. (**A**) The Amino Acid Biosynthesis panel summarizes the expression levels of genes involved in biosynthesis of each amino acid; because of space limitations we omit the portion of the diagram after glycine. (**B**) Clicking on the Arg plot in the window in (A) produces the window shown in (B), which depicts the expression levels of each individual gene involved in biosynthesis of L-arginine; this diagram is truncated to the right of argG.

dot. Note the eye can be misled by placing undue emphasis on the most highly expressed genes in a plot.

Not all genes in an organism are accessible via the Dashboard; the accessible genes are those that are assigned to a system within the Dashboard. For example, 645 *E. coli* genes do not appear in the Dashboard. For the most part these are either genes of unknown function, or genes whose function is only partially known (e.g., those assigned to a generic metabolic or transport process such as 'oxidoreductase'). We chose to disable the display of such genes in the Dashboard because currently, enabling their display slows down most Dashboard operations significantly, but we are working to include these genes in a future software release with faster performance.

Overview of Dashboard usage

Typical usage of the Omics Dashboard involves the following steps. For more details please see the online Omics Dashboard help page https://biocyc.org/dashboard/dashboard-help.html.

i) Prepare data for upload, ensuring it is in correct tab-delimited format, and any desired statistical manipulations have been performed. Note that in some cases, it may be useful to prepare two different versions of a dataset (to be uploaded in different web browser windows): the full dataset, and a filtered dataset containing only the most significant data. Important patterns will likely be most apparent in the filtered dataset, but the Dashboard can also help reveal more subtle patterns in the full dataset that are removed by significance filtering. The Dashboard accepts input gene-expression data from tab-delimited files and from BioCyc SmartTables (15).
ii) Upload the data file.
iii) Organize the display. If the data contain multiple replicate sets, organize the data columns into replicate groups. If the data contains one or more significance columns, they should be hidden from view.
iv) Visually survey the overall behavior of all top-level systems to identify systems whose behavior is changing; adjust display preferences and scaling to make patterns more apparent.
v) Select a subsystem to investigate in more detail, either because it is of intrinsic interest or because the top-level display has indicated it might be of interest, and click on it. If your dataset contains significance columns, you may prefer to use enrichment analysis results to identify interesting subsystems. Continue to identify interesting subsystems and drill down further until you reach a base panel, adjusting scaling and sorting parameters as desired.
vi) From a base panel display, view pathway and/or operon diagrams and, where available, explore regulatory influences. Click on individual genes or other objects to bring up their detail pages in a new tab.

Dashboard advanced capabilities

Figure 3 illustrates several advanced features of the Dashboard. The user may request the display of the pathway diagram(s) for the pathway(s) whose genes are shown in the current panel. The user may also request display of operon diagram(s) for all genes shown. In addition, the Dashboard can produce a graph depicting the expression levels of all genes known to regulate genes within the current panel. The Dashboard also allows the user to request more detailed information on exactly which regulators or combination of regulators regulate which genes within the panel.

If the user's input dataset contains multiple replicates of a given condition or time point, the user can direct the Dash-

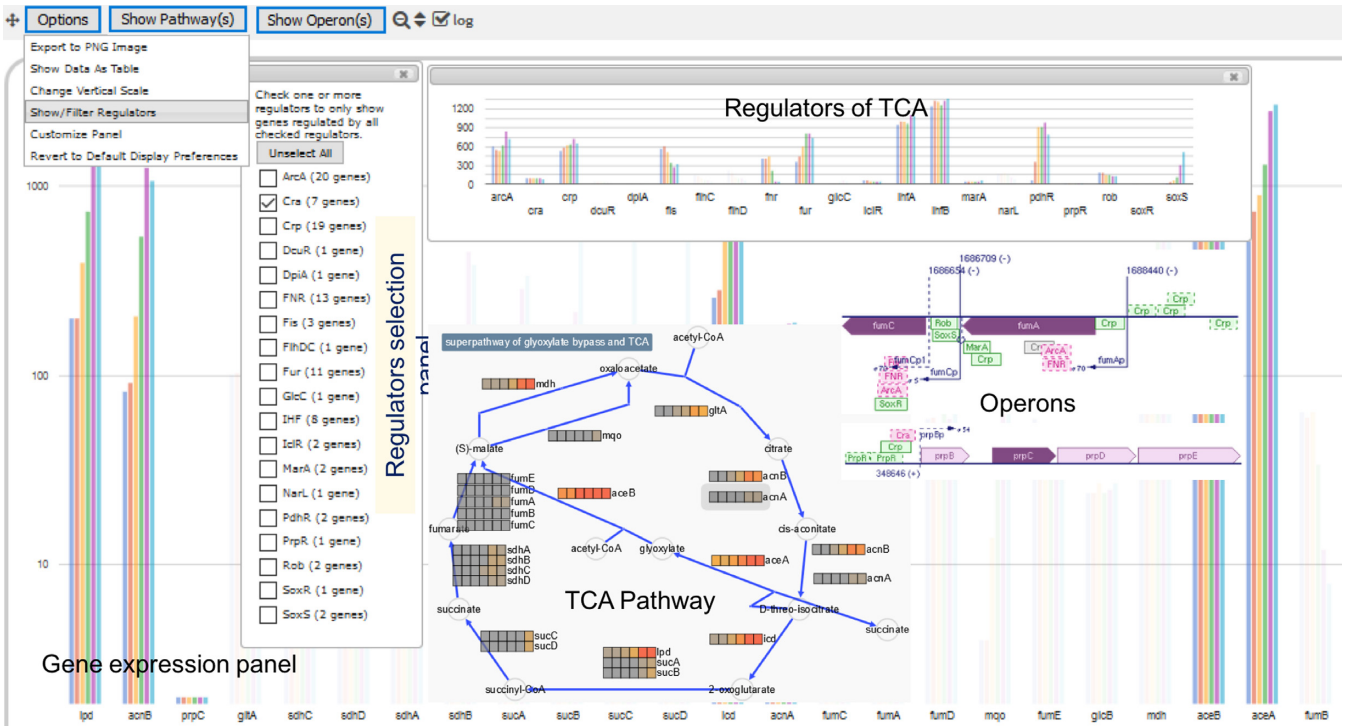

Figure 3. A zoom-in view into the glyoxylate bypass and TCA (tricarboxylic acid) cycle, a pathway that shows significant changes in gene expression. Multiple software windows are superimposed here to provide a compact illustration of the options available. The 'Show/Filter Regulators' option displays a panel (labeled 'Regulators of TCA') with a gene-expression profile of the regulators that control the pathway. Choosing a regulator ('Regulators selection'): the box next to Cra is checked to show in the regulators panel the subset of genes in the pathway that are regulated by Cra. The 'Show Pathway(s)' option displays the pathway with gene-expression heat maps. The 'Show Operons' option displays all the operons in the pathway, and the positions of the regulators and their effects are either experimentally verified or computationally and human inferred. Only a subset of operons is shown here.

board to automatically average those replicates. The averaged replicate data produce a new data column depicted in the Dashboard on a single vertical line.

We provide three options to address the case in which a few very highly expressed genes sometimes cause an entire panel to be scaled such that no details are visible for the majority of genes. The height of a panel can be increased to fill the screen, the Dashboard can produce log-scaled plots in addition to linear scaling and the range of values shown in the *Y*-axis can be altered.

For RNAseq data, the user can view the total resources the cell is investing in one or more subsystems. The Dashboard accomplishes that task with an option that draws a bar associated with each vertical line within a plot, where the bar depicts the sum of the data values associated with the vertical line.

For base plots that depict a list of genes, an extensive set of sorting options is available for ordering the genes along the *X*-axis. Genes can be sorted by genome map position, by name, and by increasing or decreasing expression value; in the last case several options are provided to select the expression value to be sorted, such as selecting the minimum, average, or maximum from a time course.

The Omics Dashboard provides a display mode that shows the results of running an enrichment analysis on the expression dataset. In order to perform the analysis, the uploaded dataset must include columns of significance values to associate with each data column. We perform a Fisher-exact hypergeometric test on the set of statistically significant genes, with an optional multiple-hypothesis correction (options are Bonferroni, Benjamini-Hochberg or Benjamini-Yekutieli corrections, with no correction being the default) to generate an enrichment score for each subsystem. The enrichment score for a subsystem is $-log(P\text{-value})$, where the p-value indicates the probability that the subsystem is enriched for significant genes. In this display mode, as shown in Supplementary Figure S2, each plot becomes a bar graph showing the enrichment score for the subsystem as a whole, along with additional line indicators showing the highest enrichment score for any component subsystem whose enrichment score exceeds that of the parent. The presence of the line indicator can suggest to the user when it may be worthwhile to drill down one level to investigate further.

Although the set of panels and subsystems defined within the Dashboard is quite extensive, the user can define new subsystems and add them to Dashboard panels at every level. When defining a new subsystem, the user specifies the name of the subsystem, the panel it is to be added to, and the pathway, GO term, and/or list of genes that comprise that subsystem.

DISCUSSION

Application 1: Lipid accumulation in a marine alga

An application that illustrates the benefits of the Dashboard analytics tool is the investigation into the environmental conditions that increase the accumulation of lipids in the marine alga *Thalassiosira pseudonana* (16) (http://scholarworks.sjsu.edu/etd_theses/4400). The purpose of the investigation was to seek conditions that increase production of lipids for algal biofuels and the commercially valuable omega-3 fatty acids eicosapentaenoic acid (EPA) and docosahexaenoic acid (DHA). In this case a whole genome transcriptome dataset (11) was used in conjunction with an annotated genome database (17). The transcriptome dataset included data for five environmental conditions: low silicate (Si), low iron (Fe), low nitrogen (N), low temperature (T) and high pH (low CO_2), relative to a nutrient replete room temperature environment. Conditions favorable for lipid accumulation would show increasing lipid biosynthesis, decreasing lipid degradation and decreasing carbohydrate synthesis while maintaining or increasing the level of carbon fixation through photosynthesis.

A previous investigation by one of us (16) used a traditional bottom-up, gene to pathway to functional analysis and took weeks to complete. Using the Dashboard, it is reasonable to expect that a person with a basic understanding of metabolism could inspect the Dashboard in Figure 4 in well under one hour and deduce that the low temperature (T) condition appeared most favorable for lipid accumulation: fatty acid and lipid biosynthesis increased, fatty acid degradation decreased and carbohydrate synthesis decreased.

The natural flow of inquiry is to investigate what kinds of lipids pathways are upregulated. The high-value EPA and DHA are very long chain unsaturated fatty acids that are synthesized by elongation of the carbon chain of palmitic acid. The nested dashboard for fatty acid synthesis in Figure 5 shows that in the low temperature (shown as a green bar) condition, palmitic acid and the elongation of palmitic acid synthesis are upregulated. This is a good indicator that the low-temperature condition is favorable for EPA and DHA production.

The functional analysis suggests the low temperature condition is favorable for lipid accumulation. The next logical inquiry is to investigate the energy systems that produce the precursors of fatty acid synthesis: photosynthesis, carbon fixation and pyruvate synthesis. Carbon is fixed during the dark cycle of photosynthesis and pyruvate is formed from glucose during glycolysis. Pyruvate is combined with Coenzyme A to produce the fatty acid metabolite acetyl CoA. The energy dashboard in Figure 6 illustrates that carbon fixation is the same as in the nutrient replete condition and glycolysis and photosynthesis are upregulated, suggesting there is an influx of carbon to the fatty acid synthesis pathway.

These intuitive nested dashboards inform the investigator as to the state of the organism *before* a detailed gene-level analysis. As a result, the gene-level analysis is more informed, focused and productive compared to other tools. The nested functional dashboards facilitate a natural inquiry flow from big-picture functionality to detailed gene product interactions. The analysis of the gene product interactions occur with the knowledge and context of the larger functional state of the organism. It is analogous to surveying a forest by flying in a helicopter over the forest versus walking through the trees. The result of using these nested functional dashboards is to not only save time, but also to provide for a more insightful, informed top-down analysis that ultimately improves the rigor and scope of the analysis. It also guides the novice investigator through the inquiry process with the agility of an expert. Thus, an investigator needs less experience and less time to process large and complex datasets. The customizable plots can also be used to concisely convey key points in presentations and publications.

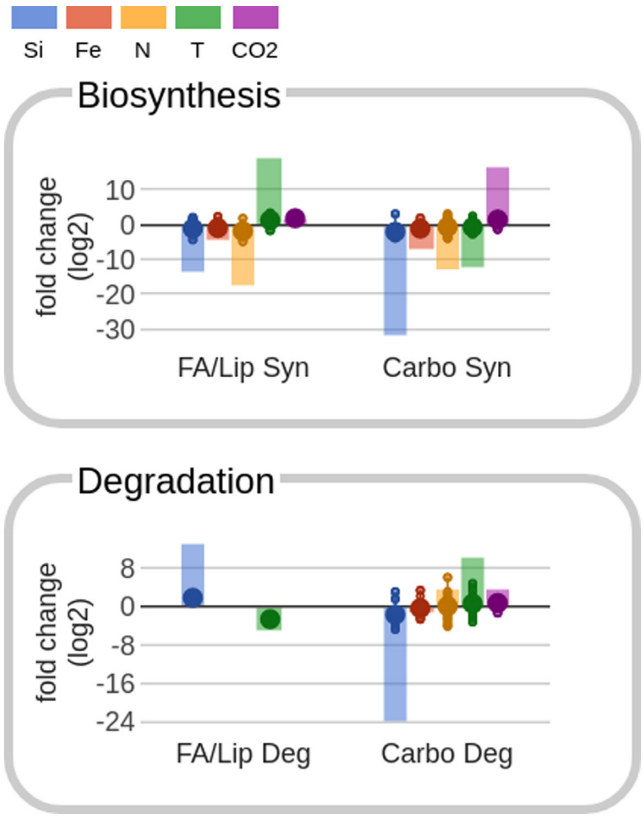

Figure 4. The Omics Dashboard analytics tool provides a visual comparison of the gene-expression data from five experimental conditions. It is quickly apparent that the low-temperature condition (T, green) indicates a transition into a more optimal lipid accumulation environment. It is expressing genes to increase lipid synthesis, decrease lipid degradation and decrease carbohydrate synthesis. Note that the default display for these panels has been customized to remove subsystems not relevant to the analysis. The small dots represent the fold differences in expression between a treatment sample set and the nutrient-replete 20°C control samples using a base 2-logarithmic scale. The large dot represents the average of the small-dot values and the bar represents the sum of the small dot values.

Application 2: Adaptation of *Escherichia coli* from anaerobic to aerobic growth

Von Wulffen *et al.* (13) studied the changes in gene expression that occur in the metabolically versatile organism *E.*

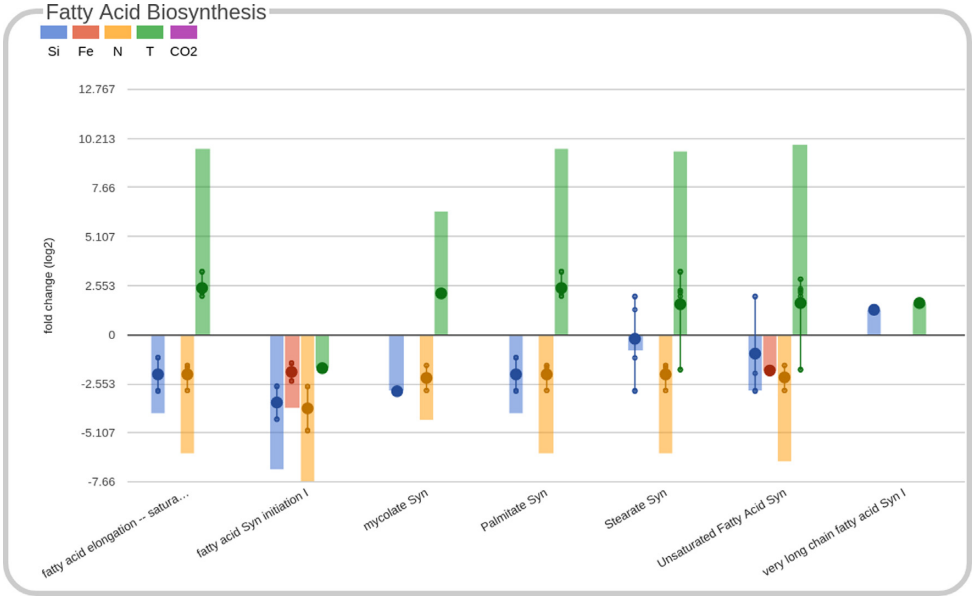

Figure 5. A drill down into the pathways that contribute to biosynthesis gives insight into what kinds of lipids are being synthesized. In the case of low temperature (green), palmitic acid and elongation have increased significantly, as well as very long chain fatty acid synthesis and unsaturated fatty acid synthesis. This is an indication that the metabolic system is in a state suitable for the production of omega 3 fatty acids, EPA and DHA. The small dots represent the fold differences in expression between a treatment sample set and the nutrient-replete 20°C control samples using a base 2-logarithmic scale. The large dot represents the average of the small-dot values and the bar represents the sum of the small dot values.

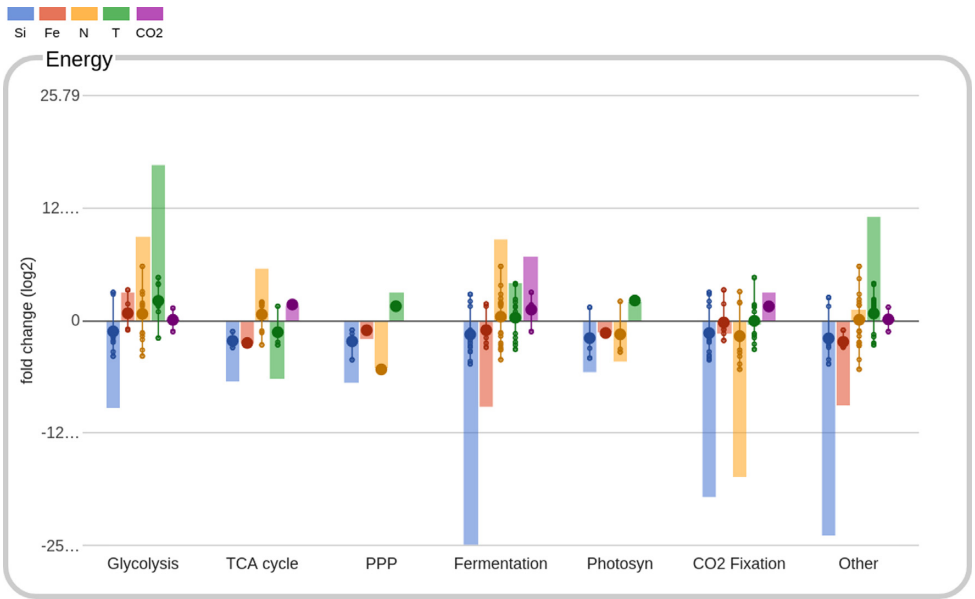

Figure 6. The energy panel indicates that in the low-temperature (green) condition, glycolysis is upregulated to increase pyruvate production, photosynthesis has increased and carbon fixation is similar to the room temperature, nutrient-replete condition. These indicators are favorable for lipid accumulation. The small dots represent the fold differences in expression between a treatment sample set and the nutrient-replete 20°C control samples using a base 2-logarithmic scale. The large dot represents the average of the small-dot values and the bar represents the sum of the small dot values.

coli when cells grown anaerobically in batch culture mode are subjected to aeration. The switch from anaerobic to aerobic growth is known to lead to profound changes in *E. coli* physiology and is expected to trigger multiple changes in gene expression. From previous studies it is well established that FNR (DNA-binding transcriptional dual regulator) and ArcAB (a two-component signal transduction system) are major regulators of gene expression in this transition. Here we illustrate how the Omics Dashboard helps in elucidating changes associated with the switch and focus on the pathways related to energy production and central carbon metabolism and their regulation.

Upon a quick inspection of the Dashboard following the upload of the average and normalized read counts of all expressed genes, it is immediately obvious that major changes in gene expression occur over the 10 minute time

course following a switch to aerobic growth. For example, gene expression in glycolysis and fermentation is decreased whereas for tricarboxylic acid (TCA) and aerobic respiration gene expression is increased. Other noticeable changes are in amino acid and carbohydrate synthesis as well as in nucleotide and carbohydrate degradation, detoxification and activation systems (Supplementary Figure S1). These changes and others are more noticeable when a table of normalized and averaged read counts for differentially expressed genes is uploaded. For example, many changes in gene expression in non-metabolic panels are brought into focus (Figure 1).

To further determine which genes and pathways are subject to transcriptional changes, the Dashboard supports quick navigation to a next level of informational detail. For example, clicking on 'TCA cycle' provides graphs of expression of all genes in the pathway (Figure 3); clicking on any gene brings up the an EcoCyc gene page for the gene, which often contains extensive information about gene function. From this window, additional windows can be created (Figure 3): a graph of expression of all regulators of this pathway; a pathway diagram for the TCA cycle; and diagrams of the operons containing TCA cycle genes. Additional options enable the user to change the diagram scale, sort the diagram and export data. Both the gene-expression panel and the pathway graphics show that most of the genes in TCA are upregulated during the first 10 min of the shift. However, expression of malate synthase G (*glcB*) and that of fumarase D and E (*fumDE*) change by <2-fold, while the expression of fumarase B (*fumB*) is decreased (the EcoCyc gene page points out that malate synthase G functions primarily during growth on glycolate). The operon panel informs that in addition to ArcA and FNR the expression of the TCA genes is potentially controlled by several regulators including Cra, Crp, Fis, Fur, GlcC, IHF, MarA, PdhR, PrpR, Rob and SoxS (Figure 3).

A similar analysis of mixed-acid fermentation, glycolysis and respiration plots reveals the following:

i) A downregulation of pathways leading to the production of H_2, CO_2, ethanol, acetate and succinate, an upregulation of steps leading to 2-oxoglutarate (TCA branch) (Supplementary Figure S3)
ii) A glycolytic pathway that is in flux, where gene expression is either unchanged, downregulated or upregulated (Supplementary Figure S4)
iii) A downregulation of systems involved in anaerobic respiration (see https://biocyc.org/ECOLI/NEW-IMAGE?type=PATHWAY&object=ANAEROBIC-RESPIRATION), represented here by the substrates involved in the electron transfer chain:
   – Formate to di-methyl sulfoxide and nitrite
   – H_2 to di-methyl sulfoxide and fumarate
   – NADH to di-methyl sulfoxide and fumarate, though genes of the multi-subunits NADH:quinone oxidoreductase I enzyme are slightly upregulated from 1.5- to 2.0-fold
   – Nitrate reduction
iv) An upregulation of systems involved in aerobic respiration (see https://biocyc.org/ECOLI/NEW-IMAGE?type=PATHWAY&object=AEROBIC-RESPIRATION), represented here by the substrates involved in the electron transfer chain:
   – D-lactate and glycerol-3-phosphate and proline to cytochrome bo oxidase
   – NADH to cytochrome bd and cytochrome bo oxidase mainly through the NADH dehydrogenase II enzyme whose transcript is increased by about 50-fold
   – Pyruvate and succinate to cytochrome bd and cytochrome bo oxidase

To gain a better understanding of the transcriptional regulation of genes involved in energy production and central carbon metabolism, a table containing normalized read counts for 133 genes present in those systems was uploaded to the Dashboard. A list of regulators and gene membership in the regulons they control can be directly accessed on the Dashboard via the sub-category Regulons within the Non-Metabolic Functions panel. Figure 7 shows a time series of the sum of read counts for all the genes present in each of 10 regulons. Data for the first 10 regulons is displayed. A total of 44 regulons potentially control the gene expression of genes involved in energy and central carbon metabolism. Although FNR and ArcA are reported to control the largest number of genes, 64 and 53, respectively, the sums of the normalized read counts, over the 10 min sampling regime, are the largest for genes in the Cra and Crp regulons (data not shown).

The Sigma Factors panel (see Figure 8) illustrates the ability of the Dashboard to facilitate the detection of expression patterns that were missed by the significance analysis applied here. Figure 8 shows a Dashboard mode that depicts the three replicates at each time point in addition to the average. (To generate a similar display, see https://biocyc.org/dashboard/dashboard.html?st=samo-463-3699298732 and click the Sigma Factors plot.) Genes *rpoE, rpoS* and *fecI* all show clear step-wise increases in expression, suggesting that these changes in expression are significant. Furthermore, the magnitudes of the changes in expression levels of these genes over time is clearly greater than the variation among replicates. Although *rpoE* and *rpoS* are not calculated to be significantly differentially expressed (neither gene exhibits a 2-fold change and a *P*-value ≤ 0.05), given the preceding visual patterns in the Dashboard it is quite plausible that the observed changes are significant. Note that *fecI* is calculated to be significantly differentially expressed.

Both RpoE and RpoS are stress-related sigma factors. RpoE is a minor sigma factor that is mostly related to heat-shock and mis-folded proteins and is also induced by other stresses including hyperosmotic shock, metal ion exposure, and by the starvation signal ppGpp upon entry into stationary phase https://biocyc.org/gene?orgid=ECOLI&id=RPOS-MONOMER. RpoS is a master regulator of the general stress response in *E. coli* https://biocyc.org/gene?orgid=ECOLI&id=RPOS-MONOMER. A modest increase in their gene expression may be associated with the stress *E. coli* cells experience upon entry to aerobic growth. SoxS, a dual transcriptional regulator that participates in the removal of superoxide and nitric oxide and protection from organic solvents and antibiotics, is indeed induced 37-

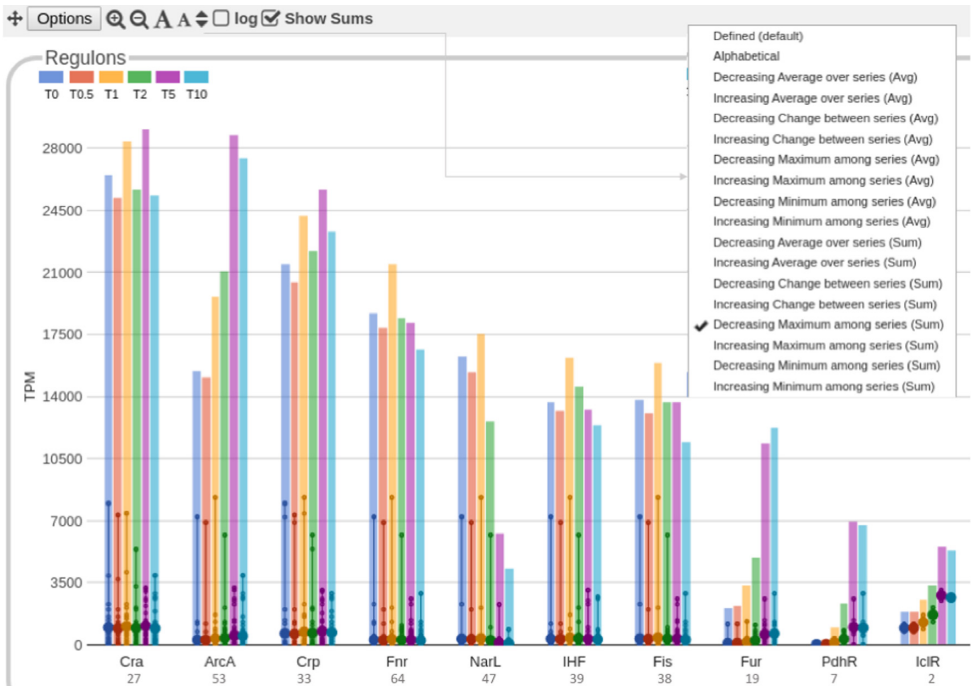

Figure 7. Gene-expression profiles at the regulon level of 133 *Escherichia coli* genes involved in fermentation, respiration, glycolysis, TCA and in the glyoxylate bypass. Each vertical line depicts the expression levels at one time point of all genes controlled by the indicated transcriptional regulator. The bars represent the total counts for all genes on the vertical line within the bar. Regulons are sorted by the 'Decreasing Maximum among series (Sum)' option. Underneath a regulon name is the number of genes reported present in that regulon.

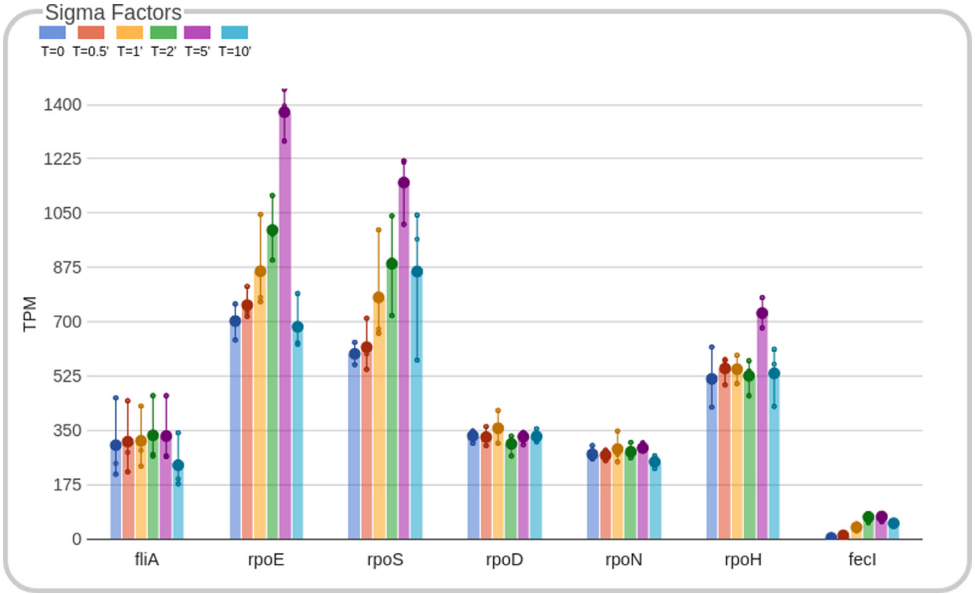

Figure 8. Sigma factors panel from the Dashboard for *Escherichia coli* anaerobic to aerobic shift (linear scale).

fold. Similarly, gene expression of *sodA,* the manganese superoxide dismutase (SOD), is increased by 158-fold.

The following additional analysis supported the notion that the observed changes in *rpoS* gene expression may be biologically meaningful. Genes that are reported in EcoCyc to be regulated solely by RpoS, or by RpoS and other regulators (extracted from an EcoCyc SmartTable) were analyzed. A total of 312 such genes were identified and 89 were found to be differentially expressed (*P*-value of $\leq$0.05 and 2-fold or greater change), which is 28%—almost a 2-fold enrichment over the general pool (15% of all genes were identified as differentially expressed).

Additional genes with visually significant changes that do not satisfy the test for significance (exhibits a 2-fold change and a *P*-value $\leq$ 0.05) are flagellar genes *flgBCDEFGHI* (see Supplementary Figure S5), *flgKLMN* and *fliKLMNOP*

(data not shown). All of these genes show step-wise changes in expression during the shift to aerobic growth. Although the gene expression data for the flagellar subsystem has high variability (based on replicates), the fact that multiple genes within the same subsystem show such similar changes more supports the notion that these changes are real. Most of the flagellar genes are activated by the FlhDC dual regulator and their promoters are driven by either the sigma70 or the sigma28 factor. The *flhDC* genes are subject to complex regulation and are down regulated by about 3-fold during the shift. Therefore, it seems plausible that a decrease in the activator leads to a decrease in the transcription of the flagellar genes and the changes observed are biologically significant. Thus, the Dashboard enables the eye to detect changes in expression that are biologically meaningful and that are overlooked by significance analysis.

CONCLUSION

The Dashboard tool enhances the interpretation of omics data to reduce analysis time and to make omics data more accessible. This unique and customizable visualization tool enables the user to quickly survey cellular activity. The Dashboard complements previous methods for significance analysis (computing statistical significance and fold change), enabling the user to quickly find and understand the responses of genes within cellular subsystems of interest; gauge the relative activity levels of different cellular systems; and compare the expression levels of a cellular system with those of its known regulators. We have shown that the Dashboard enables visual detection of expression patterns overlooked by significance analyses. The user can produce unique diagrams with the Dashboard that will facilitate novel analyses as well as enhance presentations and publications. The Omics Dashboard can be used with a variety of omics datasets including RNAseq, proteomics, metabolomics and reaction flux datasets. It can be used by novices and experts alike to efficiently and intuitively navigate and communicate complex cell responses.

AVAILABILITY

The Omics Dashboard is available as both, a web application and as part of the downloadable Pathway Tools software. The web version of the Omics Dashboard is freely available in conjunction with the EcoCyc *E. coli* database at BioCyc.org; a subscription is required for its use with other BioCyc databases. In conjunction with the downloadable Pathway Tools (see https://biocyc.org/download.shtml for download instructions) the Omics Dashboard is freely available for academic research purposes for application to any organism; a fee applies to other types of use.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Kali Pruss, Andrew Hryckowian, Justin Sonnenburg, Peter Midford, Robert Landick and Robert Gunsalus for helpful feedback regarding the Dashboard.

FUNDING

National Institute of General Medical Sciences of the National Institutes of Health (NIH) [R01GM075742]. Funding for open access charge: Referenced NIH Grant.
Conflict of interest statement. None declared.

REFERENCES

1. Slonim,D.K. and Yanai,I. (2009) Getting started in gene expression microarray analysis. *PLoS Comput. Biol.*, **5**, e1000543.
2. Seyednasrollah,F., Laiho,A. and Elo,L.L. (2015) Comparison of software packages for detecting differential expression in RNA-seq studies. *Brief. Bioinform.*, **16**, 59–70.
3. Eren,K., Deveci,M., Kucuktunc,O. and Catalyurek,U.V. (2013) A comparative analysis of biclustering algorithms for gene expression data. *Brief. Bioinform.*, **14**, 279–292.
4. Paley,S., O'Maille,P. E., Weaver,D. and Karp,P.D. (2016) Pathway collages: personalized multi-pathway diagrams. *BMC Bioinform.*, **17**, 529–538.
5. Carbon,S., Ireland,A., Mungall,C.J., Shu,S., Marshall,B. and Lewis,S. (2009) Amigo: online access to ontology and annotation data. *Bioinformatics*, **25**, 288–289.
6. Mi,H., Huang,X., Muruganujan,A., Tang,H., Mills,C., Kang,D. and Thomas,P.D. (2017) PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res.*, **45**, D183–D189.
7. Huang,D.W., Sherman,B.T., Tan,Q., Collins,J.R., Alvord,W.G., Roayaei,J., Stephens,R., Baseler,M.W., Lane,H.C. and Lempicki,R.A. (2007) The DAVID gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol.*, **8**, R183–R198.
8. Karp,P.D., Latendresse,M., Paley,S.M., Krummenacker,M., Ong,Q.D., Billington,R., Kothari,A., Weaver,D., Lee,T., Subhraveti,P. *et al.* (2015) Pathway Tools version 19.0 update: Software for pathway/genome informatics and systems biology. *Brief. Bioinform.*, **17**, 877–890.
9. Caspi,R., Billington,R., Ferrer,L., Foerster,H., Fulcher,C.A., Keseler,I.M., Kothari,A., Krummenacker,M., Latendresse,M., Mueller,L.A. *et al.* (2016) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.*, **44**, D471–D480.
10. Grossmann,S., Bauer,S., Robinson,P.N. and Vingron,M. (2007) Improved detection of overrepresentation of gene-ontology annotations with parent child analysis. *Bioinformatics*, **23**, 3024–3031.
11. Mock,T., Samanta,M.P., Iverson,V., Berthiaume,C., Robison,M., Holtermann,K., Durkin,C., Bondurant,S.S., Richmond,K., Rodesch,M. *et al.* (2008) Whole-genome expression profiling of the marine diatom *Thalassiosira pseudonana* identifies genes involved in silicon bioprocesses. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 1579–1584.
12. Guillard,R.R.L. and Ryther,J.H. (1962) Studies of marine planktonic diatoms: I. cyclotella nana hustedt, and detonula confervacea (cleve) gran. *Can. J. Microbiol.*, **8**, 229–239.
13. vonWulffen,J., Ulmer,A., Jager,G., Sawodny,O. and Feuer,R. (2017) Rapid sampling of *Escherichia coli* after changing oxygen conditions reveals transcriptional dynamics. *Genes (Basel)*, **8**, 90–114.
14. Li,B., Ruotti,V., Stewart,R.M., Thomson,J.A. and Dewey,C.N. (2010) RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, **26**, 493–500.
15. Travers,M., Paley,S.M., Shrager,J., Holland,T.A. and Karp,P.D. (2013) Groups: knowledge spreadsheets for symbolic biocomputing. *Database*, **2013**, 1–12.
16. Parker,K. (2013) *Metabolic Network Construction Based on the Genome of the Marine Diatom Thalassiosira pseudonana and the Analysis of Genome-wide Transcriptome Data to Investigate Triacylglyceride Accumulation*. Master's Thesis 4400, San Jose State University Department of Biological Sciences, http://scholarworks.sjsu.edu/etd_theses/4400.
17. Armbrust,E.V., Berges,J.A., Bowler,C., Green,B.R., Martinez,D., Putnam,N.H., Zhou,S., Allen,A.E., Apt,K.E., Bechner,M. *et al.* (2004) The genome of the diatom *Thalassiosira pseudonana:* ecology, evolution, and metabolism. *Science*, **306**, 79–86.