Sundararaman *et al. BMC Res Notes (2015) 8:771*
DOI 10.1186/s13104-015-1759-7

BMC Research Notes

TECHNICAL NOTE

Open Access

iTAP: integrated transcriptomics and phenotype database for stress response of *Escherichia coli* and *Saccharomyces cerevisiae*

Niveda Sundararaman[1†], Christine Ash[1†], Weihua Guo[1†], Rebecca Button[2], Jugroop Singh[1] and Xueyang Feng[1*]

Abstract

Background: Organisms are subject to various stress conditions, which affect both the organism's gene expression and phenotype. It is critical to understand microbial responses to stress conditions and uncover the underlying molecular mechanisms. To this end, it is necessary to build a database that collects transcriptomics and phenotypic data of microbes growing under various stress factors for in-depth systems biology analysis. Despite of numerous databases that collect gene expression profiles, to our best knowledge, there are few, if any, databases that collect both transcriptomics and phenotype data simultaneously. In light of this, we have developed an open source, web-based database, namely integrated transcriptomics and phenotype (iTAP) database, that records and links the transcriptomics and phenotype data for two model microorganisms, *Escherichia coli* and *Saccharomyces cerevisiae* in response to exposure of various stress conditions.

Results: To collect the data, we chose relevant research papers from the PubMed database containing all the necessary information for data curation including experimental conditions, transcriptomics data, and phenotype data. The transcriptomics data, including the *p* value and fold change, were obtained through the comparison of test strains against control strains using Gene Expression Omnibus's GEO2R analyzer. The phenotype data, including the cell growth rate and the productivity, volumetric rate, and mass-based yield of byproducts, were calculated independently from charts or graphs within the reference papers. Since the phenotype data was never reported in a standardized format, the curation of correlated transcriptomics–phenotype datasets became extremely tedious and time-consuming. Despite the challenges, till now, we successfully correlated 57 and 143 datasets of transcriptomics and phenotype for *E. coli* and *S. cerevisiae*, respectively, and applied a regression model within the iTAP database to accurately predict over 93 and 73 % of the growth rates of *E. coli* and *S. cerevisiae*, respectively, directly from the transcriptomics data.

Conclusion: This is the first time that transcriptomics and phenotype data are categorized and correlated in an open-source database. This allows biologists to access the database and utilize it to predict the phenotype of microorganisms from their transcriptomics data. The iTAP database is freely available at https://sites.google.com/a/vt.edu/biomolecular-engineering-lab/software.

Keywords: Open source, Transcriptomics–phenotype correlation, Yeast, *Escherichia coli*

*Correspondence: xueyang@vt.edu
†Niveda Sundararaman, Christine Ash and Weihua Guo contributed equally to this work
[1] Department of Biological Systems Engineering, Virginia Tech, Blacksburg, VA 24061, USA
Full list of author information is available at the end of the article

Background

Microorganisms face numerous stress conditions [1–3], such as oxidative stress [4–6], weak organic acid stress [7–9], nutrient limitation [10, 11], and environment fluctuation [12]. These stresses, both biotic and abiotic, occur throughout nature and comprise the ecology of the system [1, 13, 14]. Each stress condition elicits a microbial response to adapt to the unfavorable environmental conditions [15–17]. The provoked responses of microorganisms alter the current ecosystem in which they live and affect the other organisms as well [16, 18]. Such microbial responses could be recreated in laboratories and allow for a deeper understanding of the correlation between gene expressions and phenotypes [4, 7, 12]. Of particular interests to systems biologists, uncovering the correlation between transcriptomics and phenotype could identify 'genetic markers' that are primarily responsible for the occurrence of a particular phenotype within a species [6–8]. This would help determine the genetic causations of certain phenotypes across strains. With the genetic markers identified, the phenotype of strains could possibly be predicted from its transcriptomics data directly, which has great potentials in biochemical, ecological, biomedical, and environmental applications [19].

The first step towards uncovering correlations between the transcriptomics and phenotype of various microorganisms is to collect curated and coupled transcriptomics–phenotype datasets for various microorganisms. Currently, there are multiple popular databases such as the Gene Expression Omnibus (GEO) [20–22], the European Bioinformatics Institute (EBI) [23], and Many Microbe Microarrays Database (M3D) [24] that contain gene expression data. Data submitted to these databases is mostly meta-data on transcriptomics analysis, which include experimental conditions and the global gene expressions measured by either microarray [20, 25] or RNAseq analysis [26]. Comprised of over a million samples, these databases allow the analysis of large quantities of transcriptomics data; however, these databases lack the phenotypic data associated with these genotypes. Therefore, although thousands of data series and datasets are enabled for users to query for gene expression analysis, those datasets cannot provide the details about the phenotype such as cell growth rate, and hence, have limited applications in elucidating the correlations between transcriptomics and phenotype of microorganisms.

In this study, we developed an integrated transcriptomics and phenotype (iTAP) database that contained the correlated transcriptomics and phenotype datasets by collecting research articles that reported both types of data during its creation and curating the phenotype data with a standardized format. In general, we collected the transcriptomics data from GEO to provide p values and fold changes for each of the genes in *Escherichia coli* and *Saccharomyces cerevisiae* by comparing the gene expression of various strains against a reference strain as indicated in the corresponding publication. In parallel, we collected phenotype data associated with the transcriptomics data and numerically represented them as growth rates, productivity, volumetric rates, and mass-based yield of byproducts. The iTAP database also contained the experimental conditions and stress factors that the strains were subjected to. So far, we have collected, respectively, 143 and 57 datasets for *S. cerevisiae* and *E. coli*. Additionally, we demonstrated that it was feasible to use the correlated transcriptomics–phenotype datasets within the iTAP database to accurately predict cell growth rates for both *S. cerevisiae* and *E. coli* in a proof-of-concept study. Collecting this data proved to be strenuous and time-consuming, which limited the fast scale-up of the iTAP database. As the first of its kind, the iTAP database was able to identify the genetic markers, and potentially, guide synthetic biologists to rationally modify the microbial phenotypes by suppression or overexpression of genes of interests.

Implementation

Data collection and curation

As shown in Fig. 1, each of the dataseries in the database contained experimental conditions, transcriptomics data, and phenotype values of a microorganism, which were obtained from relevant research papers from the PubMed database. To ensure that both transcriptomics and phenotypic data were available in each of the dataseries, only papers that included all of the necessary details mentioned above were chosen to be included within the iTAP database.

In general, the experimental conditions were obtained directly from the chosen research papers, including information regarding the strain name, carbon source, culture medium, stress factor, and concentration of the stress factor. The transcriptomics data was collected as the gene expression levels of a test strain subjected to a particular stress condition when compared against a reference strain, and was recorded with p values and fold changes. Such data was obtained from the GEO database. The GEO database software tool, GEO2R analyzer [27], was used to compare the gene expression levels of strains facing stress conditions to the reference strain (Fig. 2). After choosing the test and reference strains and running the software, we obtained the corresponding gene expression levels, including p value and fold change. It is worth noticing that GEO, EBI and M3D share a lot of transcriptomics data that are exactly the same. Therefore, the same transcriptomics results can be generated when using EBI or M3D. The phenotype data was collected quantitatively

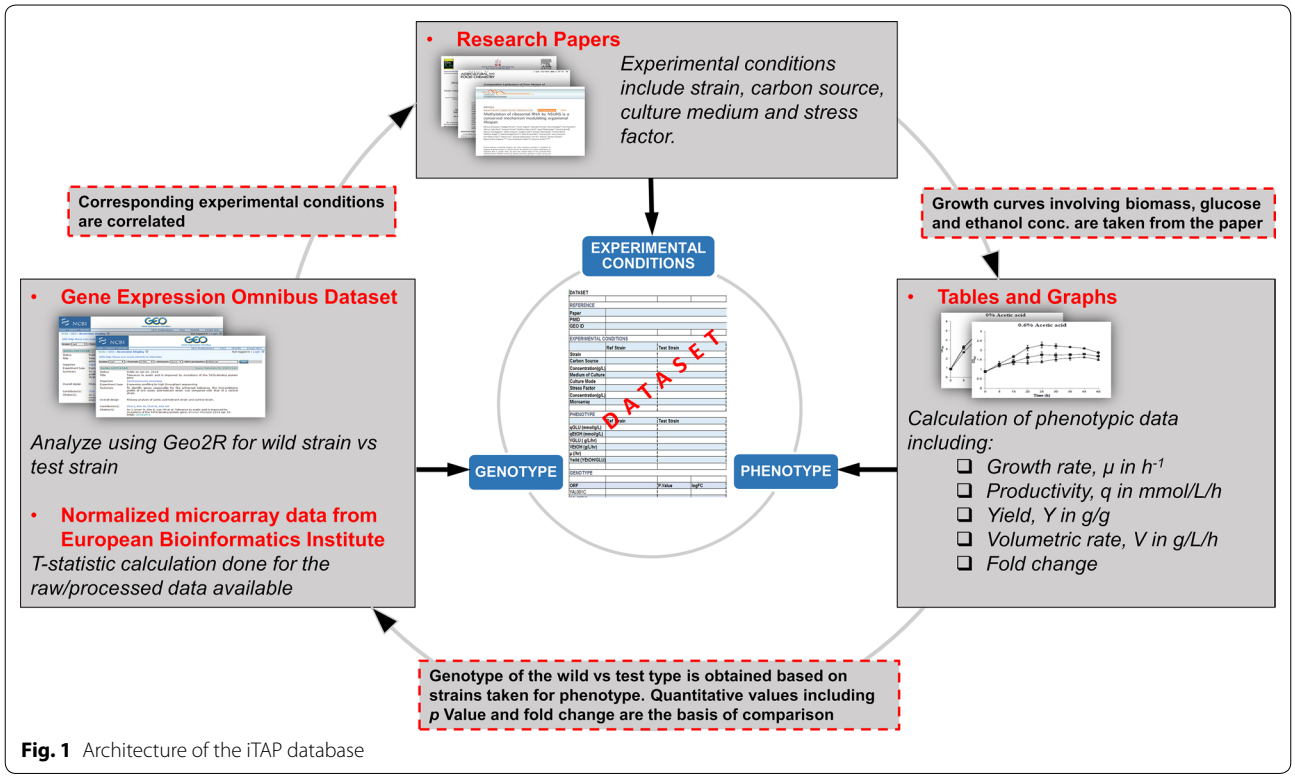


Fig. 1 Architecture of the iTAP database

as growth rates, productivity, volumetric rates, and mass-based yield of byproducts. Such data was obtained from different charts and graphs from the chosen research papers and was calculated independently. For example, all of the growth rates were calculated directly from the biomass data (e.g., dry cell weights at various time points) in the selected publications while the other phenotype data (e.g. the byproduct rates) was not used for calculating growth rates. Specifically, the phenotype data was calculated as:

We directly used the data when the desired phenotype data was reported in the selected publications. Otherwise, we utilized Plot Digitizer software [28] for graphs within the publication to obtain values of specific points depicting the rate of consumption of glucose, rate of production of products, and growth curves of different strains. All of the data, both the raw data collected from the research papers and the standardized data we calculated, were reported in the iTAP database. It is worth

$$\text{Productivity (mmol/g/h)} = \frac{\text{Initial concentration of product} - \text{Final concentration of product } (\text{mmol/L})}{(\text{Final time} - \text{Initial time, h}) \times \text{Initial biomass } (\text{g/L})} \quad (1)$$

where biomass was assumed as: 1 OD = 0.4 g/L;

$$\text{Growth rate } \left(\text{h}^{-1}\right) = \frac{\ln\left(\frac{\text{Final biomass}}{\text{Initial biomass}}\right)}{\text{Final time} - \text{Initial time (h)}} \quad (2)$$

$$\text{Volumetric rate (g/L/h)} = \frac{\text{Initial concentration of product} - \text{Final concentration of product (g/L)}}{\text{Final time} - \text{Initial time (h)}} \quad (3)$$

$$\text{Yield (g/g)} = \frac{\text{Volumetric rate of product (e.g., ethanol)}}{\text{Volumetreic rate of carbon source (e.g., glucose)}} \quad (4)$$

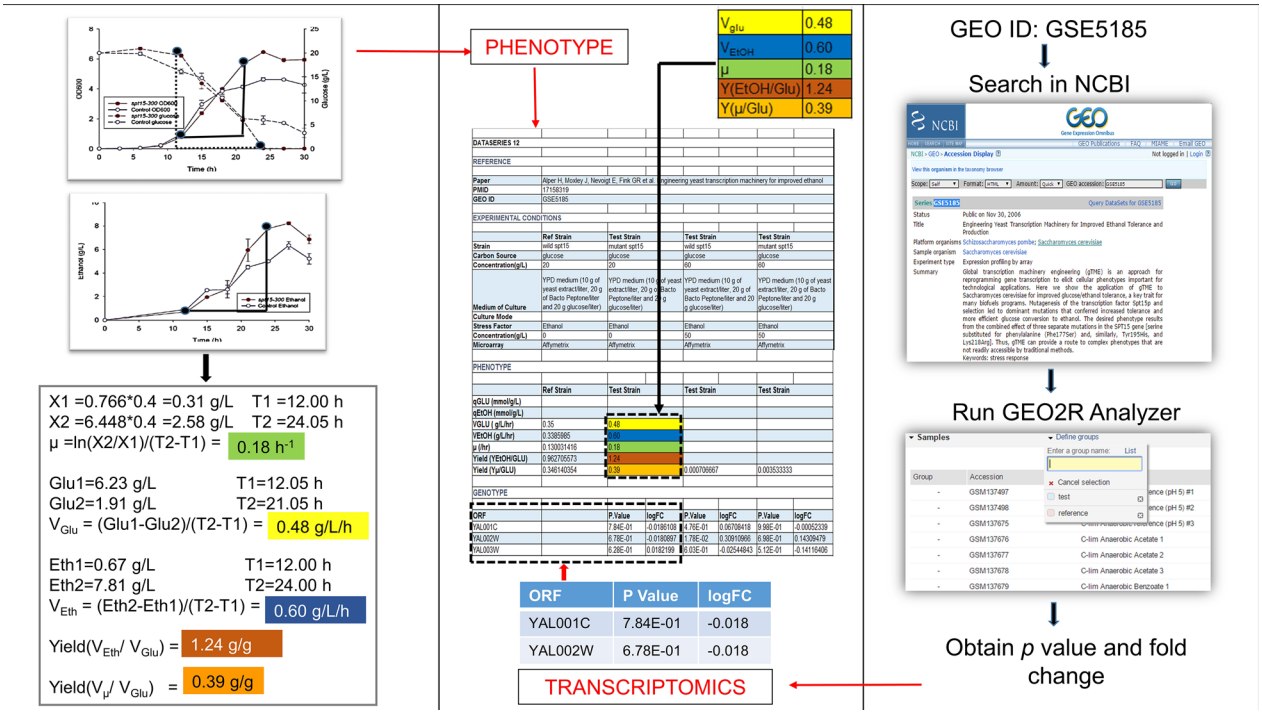


Fig. 2 Example of collecting the correlated transcriptomics–phenotype datasets

noticing that the majority of the phenotype data collected in the iTAP database were growth related, which is one aspect of the composite of observable characteristics of *E. coli* or *S. cerevisiae*.

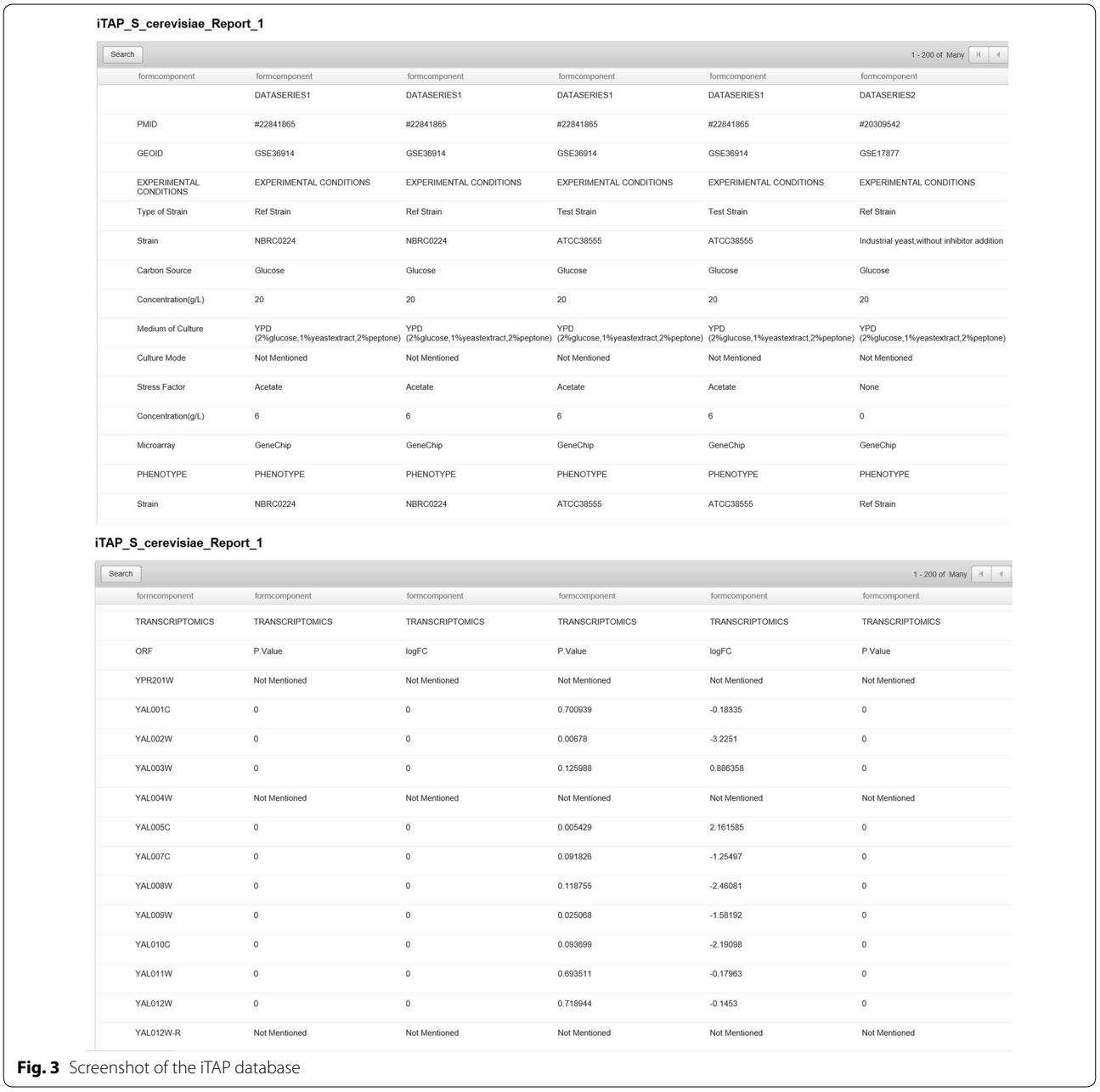
Data distribution

The iTAP database is an open source, web-based database that is freely available for use (https://sites.google.com/a/vt.edu/biomolecular-engineering-lab/software). It was developed based on Zoho Creator, an online database software that offers the data collection, cloud storage, data backup, and basic data analysis to present all the information in an efficient, user-friendly manner, as shown in Fig. 3. Advanced search with various logic symbols was available for each dataset, allowing users to find their required information efficiently. In addition, user could sort and group one or multiple dataset(s), and browse and print each dataset with different kinds of information by using the "Print" or "View Record" option to output the database. Users were also able to access the real-time data in mobile apps and download any datasets within iTAP as .csv files.

Results and discussion

To explore the possibility of using iTAP to predict cell phenotype, we first calculated the Pearson's correlation coefficient [29] of the expression levels of each gene and the corresponding cell growth rate in the entire iTAP database for *S. cerevisiae* and *E. coli*, respectively, then picked the top five genes whose expression levels were highly correlated to cell growth rate as the genetic markers for *S. cerevisiae* and *E. coli* respectively, and applied multi-variant linear regression model in MATLAB to correlate the expression levels of the genetic markers and the cell growth rates (Fig. 4). The genes whose expression levels in the test strain were not significantly different from those in the reference strain (i.e., $p > 0.25$) were set to have a fold change as zero. We found that the growth rates of both *S. cerevisiae* and *E. coli* could be accurately predicted, with R^2 reaching 0.73 and 0.86, respectively. Also, the genetic markers we identified had high coverage of all the case studies collected in the iTAP database, reaching 72.7 and 93.0 % for *S. cerevisiae* and *E. coli* respectively. This indicated that in most of the transcriptomics studies on stress responses of *S. cerevisiae* and *E. coli*, the selected genetic markers were significantly regulated and could be generally used to predict cell phenotypes such as growth rate.

We next analyzed the effect of the *p* value, which was used to judge whether or not a gene expression level in the test strain was significantly different from that in the reference strain, on prediction accuracy of cell growth rates and the coverage of case studies in the iTAP database. We found that the prediction could



iTAP_S_cerevisiae_Report_1

Search 1 - 200 of Many

formcomponent	formcomponent	formcomponent	formcomponent	formcomponent	formcomponent
	DATASERIES1	DATASERIES1	DATASERIES1	DATASERIES1	DATASERIES2
PMID	#22841865	#22841865	#22841865	#22841865	#20309542
GEOID	GSE36914	GSE36914	GSE36914	GSE36914	GSE17877
EXPERIMENTAL CONDITIONS	EXPERIMENTAL CONDITIONS	EXPERIMENTAL CONDITIONS	EXPERIMENTAL CONDITIONS	EXPERIMENTAL CONDITIONS	EXPERIMENTAL CONDITIONS
Type of Strain	Ref Strain	Ref Strain	Test Strain	Test Strain	Ref Strain
Strain	NBRC0224	NBRC0224	ATCC38555	ATCC38555	Industrial yeast,without inhibitor addition
Carbon Source	Glucose	Glucose	Glucose	Glucose	Glucose
Concentration(g/L)	20	20	20	20	20
Medium of Culture	YPD (2%glucose,1%yeastextract,2%peptone)	YPD (2%glucose,1%yeastextract,2%peptone)	YPD (2%glucose,1%yeastextract,2%peptone)	YPD (2%glucose,1%yeastextract,2%peptone)	YPD (2%glucose,1%yeastextract,2%peptone)
Culture Mode	Not Mentioned	Not Mentioned	Not Mentioned	Not Mentioned	Not Mentioned
Stress Factor	Acetate	Acetate	Acetate	Acetate	None
Concentration(g/L)	6	6	6	6	0
Microarray	GeneChip	GeneChip	GeneChip	GeneChip	GeneChip
PHENOTYPE	PHENOTYPE	PHENOTYPE	PHENOTYPE	PHENOTYPE	PHENOTYPE
Strain	NBRC0224	NBRC0224	ATCC38555	ATCC38555	Ref Strain

iTAP_S_cerevisiae_Report_1

Search 1 - 200 of Many

formcomponent	formcomponent	formcomponent	formcomponent	formcomponent	formcomponent
TRANSCRIPTOMICS	TRANSCRIPTOMICS	TRANSCRIPTOMICS	TRANSCRIPTOMICS	TRANSCRIPTOMICS	TRANSCRIPTOMICS
ORF	P.Value	logFC	P.Value	logFC	P.Value
YPR201W	Not Mentioned	Not Mentioned	Not Mentioned	Not Mentioned	Not Mentioned
YAL001C	0	0	0.700939	-0.18335	0
YAL002W	0	0	0.00678	-3.2251	0
YAL003W	0	0	0.125988	0.886358	0
YAL004W	Not Mentioned	Not Mentioned	Not Mentioned	Not Mentioned	Not Mentioned
YAL005C	0	0	0.005429	2.161585	0
YAL007C	0	0	0.091826	-1.25497	0
YAL008W	0	0	0.118755	-2.46081	0
YAL009W	0	0	0.025068	-1.58192	0
YAL010C	0	0	0.093699	-2.19098	0
YAL011W	0	0	0.693511	-0.17963	0
YAL012W	0	0	0.718944	-0.1453	0
YAL012W-R	Not Mentioned	Not Mentioned	Not Mentioned	Not Mentioned	Not Mentioned

Fig. 3 Screenshot of the iTAP database

maintain a high accuracy, with R^2 ranging from 0.61 to 0.74 for *S. cerevisiae* and 0.82–0.86 for *E. coli*. With the increase of the *p*-value, the coverage of case studies in the iTAP strain increased accordingly, since the expression levels of a gene in the test strain would be more frequently recognized as significantly different from that in the reference strain with a loose threshold of the *p*-value. Overall, by using iTAP database, we successfully proved that it was indeed possible to predict cell phenotypes from the characterization of global gene expressions.

Conclusions

In this study, the iTAP database was constructed by utilizing research papers involving stress responses for two model organisms, *E. coli* and *S. cerevisiae.* To develop iTAP database, gene expression data, specifically the *p* values and fold changes, were obtained from the GEO database, while the phenotype data was calculated from numerical information provided in multiple research papers and standardized to a defined form of representation to ensure the uniformity of the data. Till now, we have successfully curated 57 and 143 datasets for *E. coli*

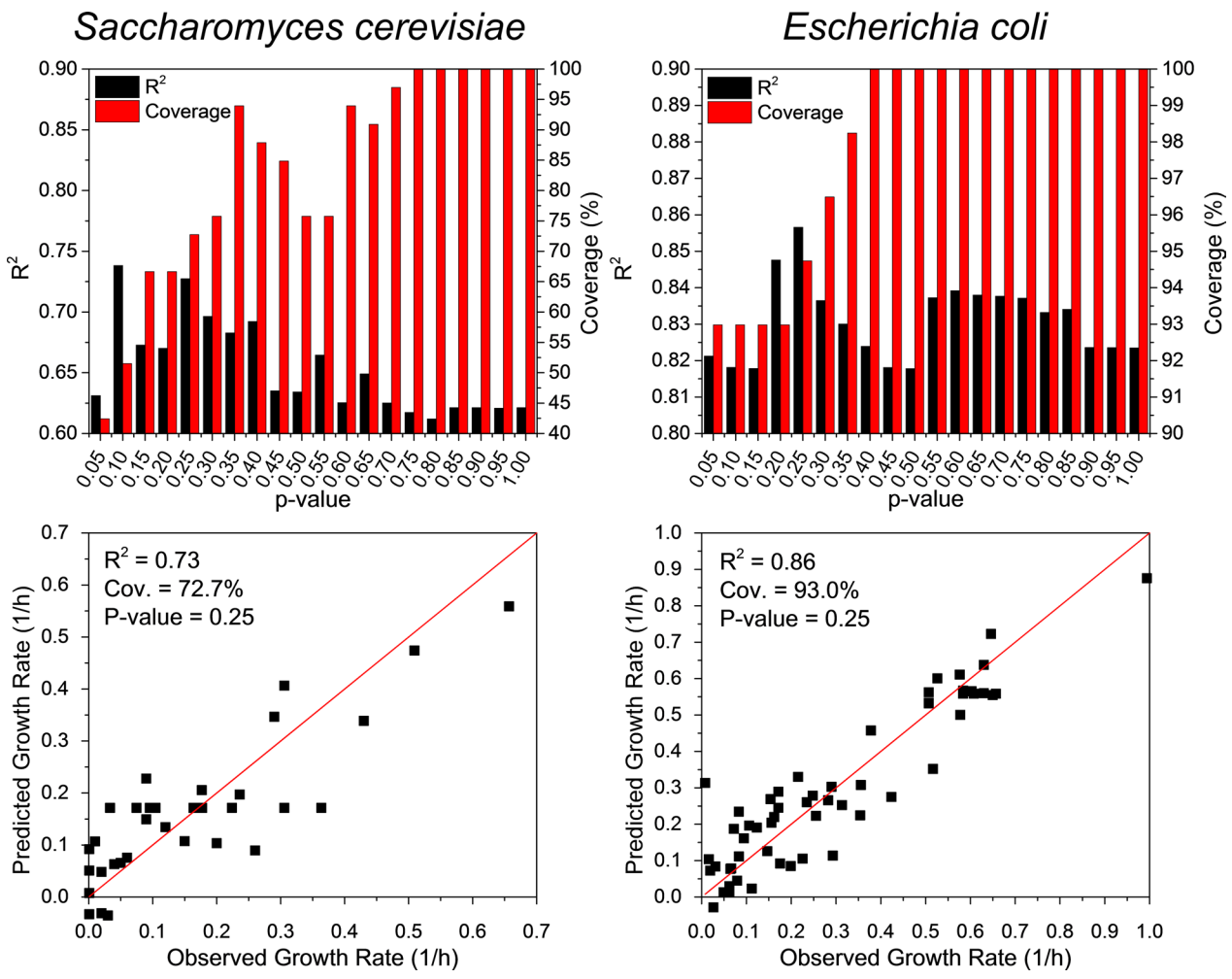


Fig. 4 Application of the iTAP database to accurately predict growth rates of *S. cerevisiae* and *E. coli*

and *S. cerevisiae*, respectively. This study also proved that with the "big data" of coupled transcriptomics–phenotype datasets, we could achieve accurate predictions of cell phenotypes, such as growth rates, directly from transcriptomic readouts and identify the genes that affect the occurrence of the phenotype most significantly. It is intended that this open-source, web-based database will be expanded to include not only more dataseries for the existing microorganisms by considering other stress conditions, but also to increase the number of microorganisms studied and include multi-omics data in future.

Availability and requirements

Project name: Integrated Transcriptomics and Phenotype Database (iTAP)
Project homepage: https://sites.google.com/a/vt.edu/biomolecular-engineering-lab/software
Operating systems: Platform independent
Programming language: Zoho Creator
License: iTAP is freely available for noncommercial purposes
Any restrictions to use by non-academics: none

Authors' contributions

NS and XF initiated this study. NS, CA, WG, and JS collected the datasets. RB assisted with the development of iTAP database using Zoho Creator. NS, CA, and WG performed the computational analysis. NS, CA, WG and XF wrote the manuscript. All authors read and approved the final manuscript.

Author details

[1] Department of Biological Systems Engineering, Virginia Tech, Blacksburg, VA 24061, USA. [2] Commonwealth Governor's School, Fredericksburg, VA 22407, USA.

Acknowledgements

We thank the writing center in Virginia Tech for improving the language of the paper. This study was supported by start-up fund (#175323) from Virginia Tech and NSF (DBI 1356669).

Competing interests
The authors declare that they have no competing interests.

Received: 8 October 2015 Accepted: 26 November 2015
Published online: 12 December 2015

References

1. Storz G, Hengge R. Bacterial stress responses. Washington, DC: American Society for Microbiology Press; 2010.
2. Hecker M, Völker U. General stress response of *Bacillus subtilis* and other bacteria. In: Advances in microbial physiology. vol 44. New York: Academic Press; 2001. p. 35–91.
3. Beales N. Adaptation of microorganisms to cold temperatures, weak acid preservatives, low pH, and osmotic stress: a review. Compr Rev Food Sci Food Saf. 2004;3(1):1–20.
4. Jozefczuk S, Klie S, Catchpole G, Szymanski J, Cuadros-Inostroza A, Steinhauser D, Selbig J, Willmitzer L. Metabolomic and transcriptomic stress response of *Escherichia coli*. Mol Syst Biol. 2010;6:364.
5. Lushchak VI. Adaptive response to oxidative stress: bacteria, fungi, plants and animals. Comp Biochem Physiol C: Toxicol Pharmacol. 2011;153(2):175–90.
6. Ma M, Liu ZL. Comparative transcriptome profiling analyses during the lag phase uncover YAP1, PDR1, PDR3, RPN4, and HSF1 as key regulatory genes in genomic adaptation to the lignocellulose derived inhibitor HMF for *Saccharomyces cerevisiae*. BMC Genom. 2010;11:660.
7. Abbott DA, Knijnenburg TA, de Poorter LMI, Reinders MJT, Pronk JT, van Maris AJA. Generic and specific transcriptional responses to different weak organic acids in anaerobic chemostat cultures of *Saccharomyces cerevisiae*. FEMS Yeast Res. 2007;7:819–33.
8. An J, Kwon H, Kim E, Lee YM, Ko HJ, Park H, Choi I-G, Kim S, Kim KH, Kim W, et al. Tolerance to acetic acid is improved by mutations of the TATA-binding protein gene. Environ Microbiol. 2015;17(3):656–69.
9. Xia J-M, Yuan Y-J. Comparative lipidomics of four strains of *Saccharomyces cerevisiae* reveals different responses to furfural, phenol, and acetic acid. J Agric Food Chem. 2009;57(1):99–108.
10. Tai SL, Boer VM, Daran-Lapujade P, Walsh MC, de Winde JH, Daran J-M, Pronk JT. Two-dimensional transcriptome analysis in chemostat cultures: combinatorial effects of oxygen availability and macronutrient limitation in *Saccharomyces cerevisiae*. J Biol Chem. 2005;280(1):437–47.
11. Brauer MJ, Huttenhower C, Airoldi EM, Rosenstein R, Matese JC, Gresham D, Boer VM, Troyanskaya OG, Botstein D. coordination of growth rate, cell cycle, stress response, and metabolic activity in yeast. Mol Biol Cell. 2008;19(1):352–67.
12. Singh J, Kumar D, Ramakrishnan N, Singhal V, Jervis J, Garst JF, Slaughter SM, DeSantis AM, Potts M, Helm RF. Transcriptional response of *Saccharomyces cerevisiae* to desiccation and rehydration. Appl Environ Microbiol. 2005;71(12):8752–63.
13. Schimel J, Balser TC, Wallenstein M. Microbial stress-response physiology and its implications for ecosystem function. Ecology. 2007;88(6):1386–94.
14. Çakar ZP, Seker UOS, Tamerler C, Sonderegger M, Sauer U. Evolutionary engineering of multiple-stress resistant *Saccharomyces cerevisiae*. FEMS Yeast Res. 2005;5(6–7):569–78.
15. Kültz D. Molecular and evolutionary basis of the cellular stress response. Annu Rev Physiol. 2005;67(1):225–57.
16. Abee T, Wouters JA. Microbial stress response in minimal processing. Int J Food Microbiol. 1999;50(1–2):65–91.
17. Ramos JL, Gallegos Ma-T, Marqués S, Ramos-González M-I, Espinosa-Urgel M, Segura A. Responses of gram-negative bacteria to certain environmental stressors. Curr Opin Microbiol. 2001;4(2):166–71.
18. Love N, Bott C. Evaluating the role of microbial stress response mechanisms in causing biological treatment system upset. Water Sci Technol. 2002;46(1–2):11–8.
19. Heer D, Heine D, Sauer U. Resistance of *Saccharomyces cerevisiae* to high concentrations of furfural is based on NADPH-dependent reduction by at least two oxireductases. Appl Environ Microbiol. 2009;75(24):7631–8.
20. Edgar R, Barrett T. NCBI GEO standards and services for microarray data. Nat Biotechnol. 2006;24(12):1471–2.
21. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Marshall KA, et al. NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Res. 2009;37(suppl 1):D885–90.
22. Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, et al. NCBI GEO: archive for functional genomics data sets—10 years on. Nucleic Acids Res. 2011;39(suppl 1):D1005–10.
23. Stoesser G, Tuli MA, Lopez R, Sterk P. The EMBL nucleotide sequence database. Nucleic Acids Res. 1999;27(1):18–24.
24. Faith JJ, Driscoll ME, Fusaro VA, Cosgrove EJ, Hayete B, Juhn FS, Schneider SJ, Gardner TS. Many microbe microarrays database: uniformly normalized affymetrix compendia with structured experimental metadata. Nucleic Acids Res. 2008;36:D866–70.
25. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 1995;270(5235):467–70.
26. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10(1):57–63.
27. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. 2013;41(D1):D991–5.
28. Huwaldt JA. Plot Digitizer: GPL. Available: http://plotdigitizer.sourceforge.net/. Accessed 2013 May 6.
29. Stigler SM. Francis Galton's account of the invention of correlation. Stat Sci. 1989;4(2):73–9.