D348–D353 *Nucleic Acids Research, 2018, Vol. 46, Database issue* *Published online 25 September 2017*
doi: 10.1093/nar/gkx842

YEASTRACT: an upgraded database for the analysis of transcription regulatory networks in *Saccharomyces cerevisiae*

Miguel C. Teixeira[1,2,*,†], Pedro T. Monteiro[3,4,*,†], Margarida Palma[1,2], Catarina Costa[1,2], Cláudia P. Godinho[1,2], Pedro Pais[1,2], Mafalda Cavalheiro[1,2], Miguel Antunes[1,2], Alexandre Lemos[3,4], Tiago Pedreira[5] and Isabel Sá-Correia[1,2,*]

[1]Department of Bioengineering, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, 1049-001 Lisbon, Portugal, [2]iBB-Institute for BioEngineering and Biosciences, Av. Rovisco Pais, 1049-001 Lisbon, Portugal, [3]Department of Computer Science and Engineering, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, 1049-001 Lisbon, Portugal, [4]INESC-ID, SW Algorithms and Tools for Constraint Solving Group, R. Alves Redol 9, 1000-029 Lisbon, Portugal and [5]Instituto Gulbenkian de Ciência, Rua da Quinta Grande 6, 2780-156 Oeiras, Portugal

Received July 26, 2017; Revised September 03, 2017; Editorial Decision September 04, 2017; Accepted September 18, 2017

ABSTRACT

The YEAst Search for Transcriptional Regulators And Consensus Tracking (YEASTRACT—www.yeastract.com) information system has been, for 11 years, a key tool for the analysis and prediction of transcription regulatory associations at the gene and genomic levels in *Saccharomyces cerevisiae*. Since its last update in June 2017, YEASTRACT includes approximately 163000 regulatory associations between transcription factors (TF) and target genes in *S. cerevisiae*, based on more than 1600 bibliographic references; it also includes 247 specific DNA binding consensus recognized by 113 TFs. This release of the YEASTRACT database provides new visualization tools to visualize each regulatory network in an interactive fashion, enabling the user to select and observe subsets of the network such as: (i) considering only DNA binding evidence or both DNA binding and expression evidence; (ii) considering only either positive or negative regulatory associations; or (iii) considering only one set of related environmental conditions. A further tool to observe TF regulons is also offered, enabling a clear-cut understanding of the exact meaning of the available data. We believe that with this new version, YEASTRACT will improve its role as an open web resource instrumental for Yeast Biologists and Systems Biology researchers.

INTRODUCTION

The YEASTRACT (YEAst Search for Transcriptional Regulators And Consensus Tracking; www.yeastract.com) database was released in 2006 (1) to provide free access to all published information on transcriptional regulation in the model eukaryote *Saccharomyces cerevisiae*, curated by experts in the field. Since then, the database has been maintained up-to-date, and up-graded in terms of the data enclosed in it and of the tools that have been developed to increase the ability to fully exploit the increasingly available data (2–4).

Databases such as MYBS (5), TRANSFAC (6), RSAT (7), YPA (8) and YeTFaSCo (9) focus most of their analysis and predictive power on the understanding of promoter regions, considering information that includes the occurrence of TF binding sites, but also the accessibility of those putative binding sites, determined among other aspects by nucleosome occupancy data. More recently, SGD (10) provided some information on transcription factor – target gene associations based on published information. YEASTRACT, besides providing tools for promoter analysis in yeast, is, to our knowledge, the single information system that offers a complete integration of all the experimentally curated transcriptional regulatory data ever published for *S. cerevisiae*. Furthermore, YEASTRACT provides a number of queries that enable the user to scrutinize the available data, to compare it with his own data and to run predictive analysis on the regulators of a gene or genome-wide response.

*To whom correspondence should be addressed. Tel: +351 218417682; Email: isacorreia@tecnico.ulisboa.pt
Correspondence may also be addressed to Miguel C. Teixeira. Email: mnpct@tecnico.ulisboa.pt
Correspondence may also be addressed to Pedro T. Monteiro. Email: Pedro.Tiago.Monteiro@tecnico.ulisboa.pt
†These authors contributed equally to this work as first authors.

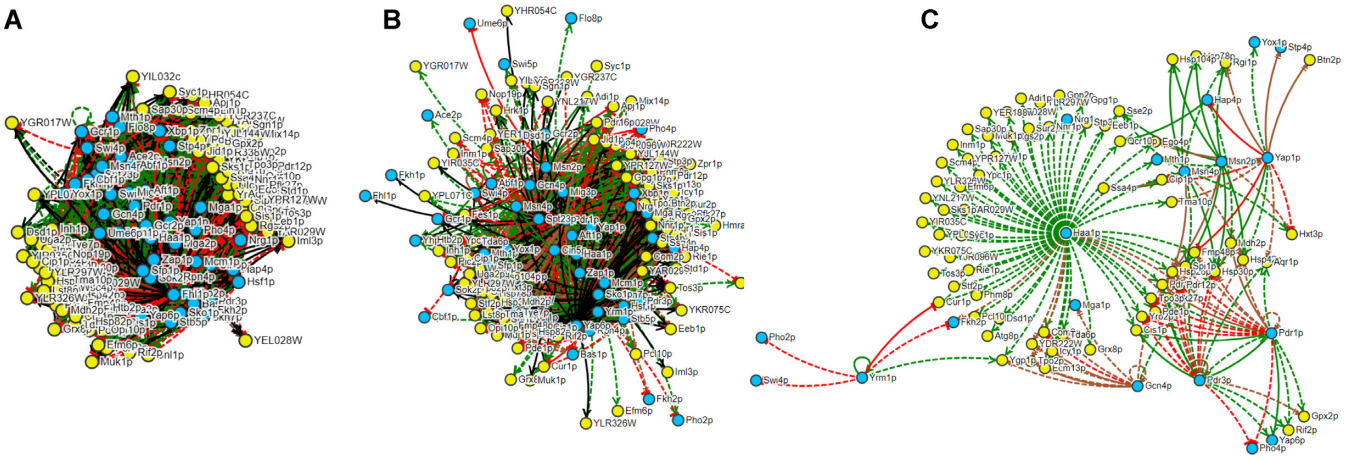

Figure 1. Regulatory network controlling, directly or indirectly, the Haa1 target genes identified in a single transcriptomic experiment (11). In this analysis only top enriched TFs, targeting more than 25% of the gene list were included. The whole TF network predicted to control the expression of these genes is depicted in (**A**), while the subnetwork of stress activated regulatory associations within the network is shown in (**B**) and the sub-subnetwork of weak acid stress activated regulatory associations within the network is shown in (**C**). The experimental evidences underlying each regulatory association (in full lines, in the case of DNA-binding evidence, or in dashed lines in the case of expression evidence), as well as the sign of the interaction—positive (green), negative (red), positive and negative (brown) or undefined (black) are highlighted.

In this new release, YEASTRACT was empowered with new tools, especially focusing on improved and interactive image analysis, as described below. The information in YEASTRACT was further updated and revised, leading to the integration of more reliable data. It currently includes nearly 10-fold more regulatory associations than in its first release, based on the scientific community published work in the field, especially supported by transcriptomics or ChIP-chip or ChIP-seq data, and includes more information associated to each transcriptional association, including the environmental condition and the experimental setup in which the associations were registered, and the direction of the interaction. All the information deposited in the database can now be used to filter query results interactively through a strong visual support, contributing to obtain predictions with an expected higher biological relevance.

Data Upgrade

In YEASTRACT's first release, the database gathered ∼12,000 regulatory associations between transcription factors (TF) and target genes and 257 transcription factor binding sites (TFBS) (1). Since then, a stepwise increase in the number of regulatory associations was registered, following the rate of published data in the field, mostly based on transcriptomic analyses (2–4). With this new release, another 20,000 regulatory associations were included in YEASTRACT, reaching a total of ∼163,000 unique regulatory associations, obtained from >1,600 bibliographic references. With this release 35 new transcription factor binding sites, demonstrated in the literature, were also added to the database, that includes today 268 TFBS.

All the data gathered in YEASTRACT since its first release has been manually curated by experts in the field, to assure its users high reliability. In this release, the data deposited in YEASTRACT was reviewed and a few proteins previously considered TFs, together with its associated regulatory associations, were withdrawn from the database based on the fact that, given a more accurate description, they are not really TFs but rather co-factors or chromatin remodeling proteins. Furthermore, emphasis was given to associate to each regulatory association all available information regarding: (i) the experimental evidences used to define this association; (ii) the environmental conditions in which this association takes place (iii) the directionality of the interaction, e.g. is the transcription factor acting as an activator or as a repressor of the target gene, and (iv) to begin to unveil what could be strain specific differences in terms of transcriptional regulation. Additionally, data on the exact S. cerevisiae strain in which the regulatory association was observed was gathered for the first time. This is so far only possible to a limited extent, given that this information was not yet gathered for all the regulatory associations obtained in previous releases and because it was found that specific strain information is surprisingly missing in some of the published datasets.

The information gathered for each regulatory association has been coupled to existing tools, enabling the user to filter its queries to find regulatory associations that are exclusively based on DNA-binding evidence (including EMSA, DNA footprinting, ChIP, ChIP-chip, ChIP-seq or more recently ChIP-exo data), or exclusively based on expression evidence (including, mostly, that obtained by comparing gene expression in wild-type strains with that observed in TF mutant strains, registered through northern blotting, RT-PCR, DNA microarrays and RNA-seq experiments), or eventually exclusively based on both DNA binding and expression evidence, which increases the degree of confidence of the query predictive analysis. Since TF activation as well as transcriptional activation or repression is highly dependent on environmental cues, the knowledge of the environmental conditions in which every regulatory association was observed enables the possibility to search for the transcription factors that regulate the user's gene or gene-list of inter-

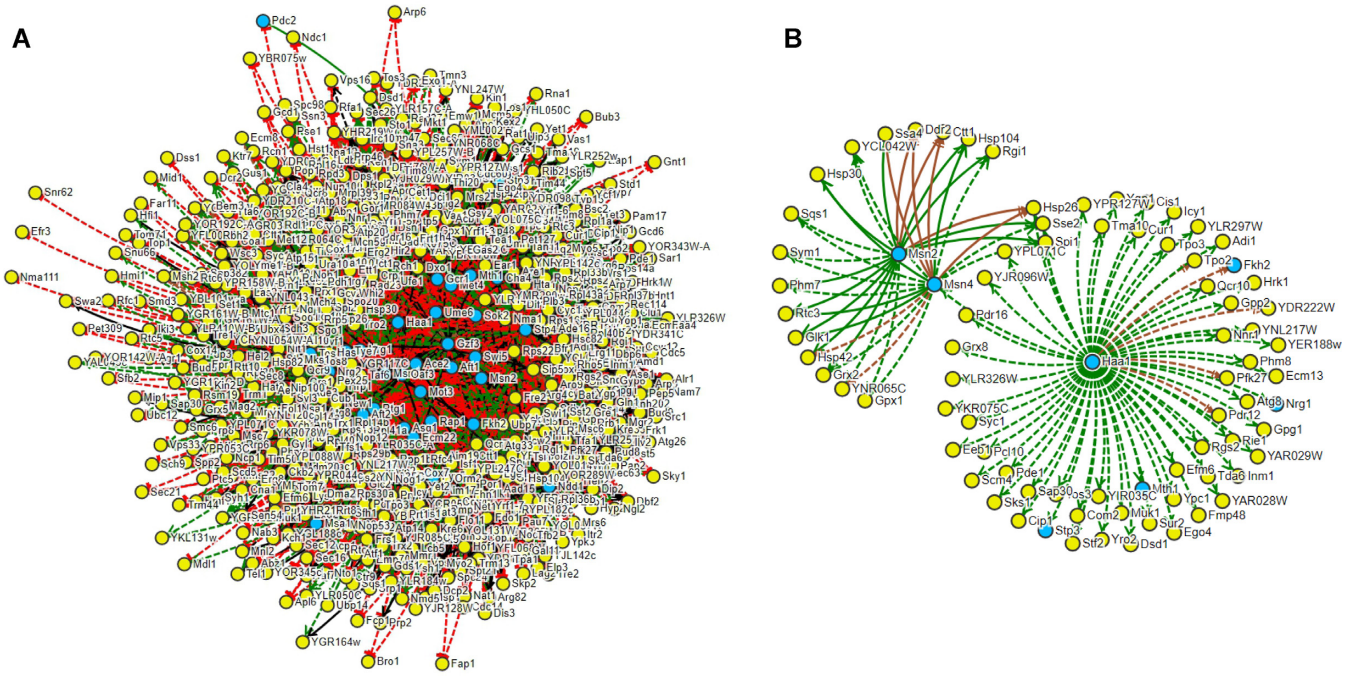

Figure 2. The Haa1 regulon. The whole network of Haa1 target genes is depicted in (**A**), while the subnetwork of Haa1-dependent regulatory associations known to take place under weak acid stress is shown in (**B**). The experimental evidences underlying each regulatory association (in full lines, in the case of DNA-binding evidence, or in dashed lines in the case of expression evidence), as well as the sign of the interaction—positive (green), negative (red), positive and negative (brown) or undefined (black) are highlighted.

est, filtering for only regulatory associations that are known to occur in specific environmental conditions.

The up-to-date high-quality data deposited in YEASTRACT, as done regularly for the past 11 years, is expected to continue to benefit the community working on the Biology of *S. cerevisiae*, a key organism both as cell factory and eukaryotic model, and on the development of Systems Biology, a growing field of research which relies on the existence of massive amounts of reliable data.

Interactive visualization tools

A key feature of this new release of the YEASTRACT database is the significant expansion of the visualization tools offered therein, which we believe follows in line with the requests of many of our users. Visualization tools are indeed crucial for a more intuitive interpretation of results, as exemplified in the following examples, that illustrate how to use these new tools.

The query 'Rank by TF' is particularly useful for the analysis of genome-wide expression data, particularly those coming from transcriptomic approaches, offers now three options for the visualization of the results. The new visualization tool can be selected by pressing the '*Interactive image: force-directed layout*' option in the results page. As an example, Figure 1A depicts the outcome of using the query 'Rank by TF' to search for the regulatory network controlling the genes whose expression is up-regulated in response to acetic acid stress in *Saccharomyces cerevisiae* and obtained through a transcriptomic analysis (11). Only enriched TFs, targeting more than 25% of the dataset were considered, an option made to contemplate only the TFs which apparently play the most important role. The visual depiction of the network, which uses Data Driven Documents (D3.js: https://d3js.org/) to distribute the elements of the network, highlights the enormous complexity of this predicted network, including 110 target genes and 50 TFs. As it is, the network is very complex, which on one hand makes it more difficult to extract meaningful knowledge, and on the other may be misleading as we cannot be sure that all these TFs and corresponding regulatory associations are active under acetic acid stress. To deal with this issue, the user can filter the query to focus only on regulatory associations taking place under a more restricted set of environmental conditions. Selecting the 'Environmental Condition Group' 'Stress' a much more simplified and accurate prediction of the network underlying this transcriptional response is observed (Figure 1B). This simplified network predicts that 42 TFs work together to provide the observed transcriptional response, setting aside eight TFs whose activity has not been registered under stress. Within the 42 TFs remaining, some may be active under stress, but not specifically under acetic acid stress. Thus, an even more specific network can be obtained, by narrowing further the search, using as filter the 'Environmental Condition Sub-Group' 'Weak Acid Stress'. In this case, immediately in the center the TFs which are responsible for the highest number of regulated genes appears Haa1, working as the hub of this system (Figure 1C). This is in agreement with the notion that Haa1 controls the expression of around 70% of the acetic acid response in yeast (11). Interestingly, even in this highly filtered network, it is still possible to detect the contribution of eight other TFs, including the general

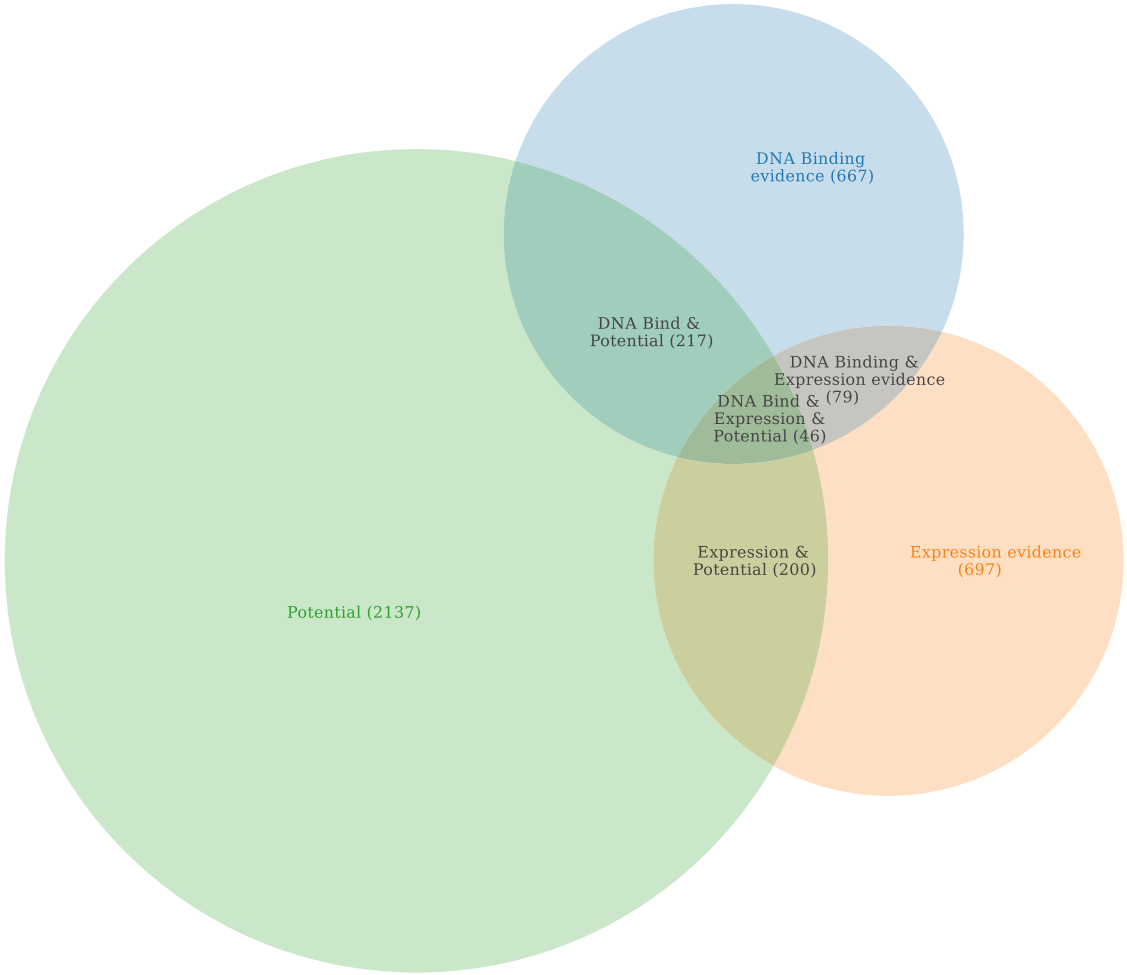

Figure 3. The Yrr1 regulon. The distribution of Yrr1 target genes in the dependency of the underlying experimental evidence or computational prediction is highlighted as a Venn diagram.

stress response regulators Msn2 and Msn4, and the major regulators of multidrug resistance Pdr1, Pdr3. Interestingly, both Msn2 and Msn4 have been reported to be Haa1 targets and participate in the control of the yeast response to weak acids, both in the presence of short chain (12) and long chain weak acids (13,14). For Pdr1 and Pdr3, on the other hand, there is only evidence to support its involvement in the control of the highly lipophilic weak acid herbicide 2–4-dichlorophenoxyacetic acid (2,4-D) (15) and the antimalarial drug artesunate (16). This case-study exemplifies the potential of using the new interactive visualization tools offered in YEASTRACT as an approach to conduct higher-confidence predictions of what are the regulators of a given gene or a given transcriptome-wide response.

In the specific case of analyzing a TF regulon (e.g. the complete set of targets of a transcription factor, resulting from all available published data on this TF), two new interactive visualization tools are further offered in this new release of the YEASTRACT database. First, the user may select the 'Search for genes' query to find the full network of genes controlled by a single transcription factor, and visualize it in the '*Interactive image: force-directed layout*' option. This option is now available in every query offered in YEASTRACT, by selecting the 'MENU' icon that appears in the top of each generated table of results. As an example, Figure 2A displays the image obtained for the (directly or indirectly) Haa1-regulated genes, showing the TF in the centre of those 650 target genes, together with many other TFs that are regulated by Haa1, and that also regulate its targets. Since the Haa1 transcription factor is activated under acetic acid, but this is not the case, as far as current knowledge goes, for many of its targeted TF encoding genes, it is possible to get a more accurate depiction of the Haa1 regulon by filtering by 'Environmental Condition Sub-Group' 'Weak Acid Stress'. The result shows a smaller network of Haa1 targets, including 76 genes which are regulated by Haa1 (Figure 2B), 14 of which are indirectly regulated via Msn4, and are more likely to represent the core of the Haa1 regulon relevant for the weak acid stress response.

To further provide a clear idea on what are the most reliable targets of a given transcription factor, based on the existing information, a second format of regulon depiction is also offered. This can be accessed by searching (through the 'Quick Search' option) for a specific gene and selecting the corresponding 'Protein info' tab. Then selecting the 'See regulon' option, a Venn diagram of the TF regulon is obtained.

As a case-study, the Yrr1 regulon is shown in Figure 3. According to the information deposited in YEASTRACT, there are a total of 1443 documented targets of Yrr1, an important regulator of multidrug resistance in yeast (17). If the user considers the potential targets of Yrr1, that are those genes in whose promoter region it is possible to find the Yrr1 recognition sequence WCCGYKKWW (17), the number increases a lot. Among these, there is a surprising lack of overlap between the genes whose promoters are bound by Yrr1, the promoters which have a perfect Yrr1 binding site and the genes whose expression is indeed controlled by Yrr1. Only 79 genes meet both criteria. Surprisingly, this lack of overlap is observed for many of the TF regulons deposited in YEASTRACT. As such, this visualization tool enables a more intuitive analysis of the TF regulons, highlighting the regulatory subnetwork for which there is the highest degree of confidence. It is not surprising to see that within the 46 most verified targets of Yrr1 are a number of genes encoding multidrug transcription factors (e.g. *PDR3* and *YRR1*), drug:H+ antiporters (*FLR1*, *AZR1* and *TPO1*) and ABC drug efflux pumps (*PDR10*, *PDR15*, *SNQ2* and *YOR1*). A few questions arise from the analysis of this figure as well: are the genes whose promoter is bound by Yrr1, but whose expression was not found to be controlled by Yrr1, controlled by it in yet unknown environmental conditions? And are the genes whose expression is affected by Yrr1, but for which there is no evidence of direct promoter targeting by the TF, indirect targets of Yrr1? Is the search for TF consensus sequences in a promoter region a reliable way of finding a putative target gene?

Further available resources

In this release of the YEASTRACT database, a new data submission tool is offered. This tool comes as the answer to many users´ requests, turning it possible for individual researchers to upload their recently published data in the database. To do that, the user has to select the 'Data Submission Form', where an excel document can be downloaded, with filling instructions. Afterwards, the document can be easily uploaded in our portal, together with a few details on author information. For each submitted regulatory association, additional information is requested, so that the data can fit the standards of the YEASTRACT curation process, including the strain and the environmental conditions in which the association was found to take place, as well as the nature of the experimental setup used to uncover it. Once the users´ data is submitted, it will be made available in a short term, after verification by our curation staff. As always, only published data will be considered.

This new tool is expected to increase the interaction between the large community generating regulatory data in *S. cerevisiae* and the YEASTRACT curation team, which will inevitably increase the rate of database updating and the quality of the collected information.

FUTURE DIRECTIONS

The YEASTRACT team is committed to continue to offer updated, reliable and complete information on the field of transcriptional regulation in yeast to the international community of yeast and systems biologists. Furthermore, continuous improvements of the provided tools will be made available, in response to the requests and needs of its users. Focus will be given in the future to the extension of the database to the *S. cerevisiae* pan-genome and to other yeasts of biotechnological interest, together with a close interconnection with the PathoYeastract database (18), in a comparative genomics approach.

ACKNOWLEDGEMENTS

The information about yeast genes other than documented regulations, potential regulations and the transcription factor binding sites contained in YEASTRACT was gathered from SGD and the GO Consortium. We acknowledge all those who have, over the years, contributed to YEASTRACT (http://yeastract.com/credits.php). We are also grateful to colleagues and friends from the yeast community for their encouragement and suggestions.

FUNDING

YEASTRACT is currently supported by national funds through Programa Operacional Regional de Lisboa 2020 [LISBOA-01-0145-FEDER-022231—the BioData.pt Research Infrastructure]; FCT—Fundação para a Ciência e a Tecnologia [PTDC/BBB-BIO/4004/2014] grants (to A.L, T.P); post-doc and PhD grants (to M.P., C.C., C.G., P.P., M.C.). iBB-Institute for Bioengineering and Biosciences from Programa Operacional Regional de Lisboa 2020 [007317]; iBB and INESC-ID from FCT [UID/BIO/04565/2013, UID/CEC/50021/2013, respectively]. Funding for open access charge: Instituto Superior Técnico.

Conflict of interest statement. None declared.

REFERENCES

1. Teixeira,M.C., Monteiro,P.T., Jain,P., Tenreiro,S., Fernandes,A.R., Mira,N.P., Alenquer,A., Freitas,A.T., Oliveira,A.L. and Sá-Correia,I. (2006) The YEASTRACT database: a tool for the analysis of transcription regulatory associations in Saccharomyces cerevisiae. *Nucleic Acids Res.*, **34**, D446–D451.
2. Monteiro,P.T., Mendes,N.D., Teixeira,M.C., D'orey,S., Tenreiro,S., Mira,N.P., Pais,H., Francisco,A.P., Carvalho,A.M., Lourenço,A.B. *et al.* (2008) YEASTRACT-DISCOVERER: new tools to improve the analysis of transcriptional regulatory associations in Saccharomyces cerevisiae. *Nucleic Acids Res.*, **36**, 132–136.
3. Abdulrehman,D., Monteiro,P.T., Teixeira,M.C., Mira,N.P., Lourenço,A.B., Dos Santos,S.C., Cabrito,T.R., Francisco,A.P., Madeira,S.C., Aires,R.S. *et al.* (2011) YEASTRACT: providing a programmatic access to curated transcriptional regulatory associations in Saccharomyces cerevisiae through a web services interface. *Nucleic Acids Res.*, **39**, 136–140.
4. Teixeira,M.C., Monteiro,P.T., Guerreiro,J.F., Gonçalves,J.P., Mira,N.P., Dos Santos,S.C., Cabrito,T.R., Palma,M., Costa,C., Francisco,A.P. *et al.* (2014) The YEASTRACT database: an upgraded information system for the analysis of gene and genomic transcription regulation in Saccharomyces cerevisiae. *Nucleic Acids Res.*, **42**, 161–166.
5. Tsai,H.K., Chou,M.Y., Shih,C.H., Huang,G.T.W., Chang,T.H. and Li,W.H. (2007) MYBS: a comprehensive web server for mining transcription factor binding sites in yeast. *Nucleic Acids Res.*, **35**, 221–226.
6. Wingender,E. (2000) TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.*, **28**, 316–319.

7. van Helden,J. (2003) Regulatory sequence analysis tools. *Nucleic Acids Res.*, **31**, 3593–3596.
8. Chang,D.T.H., Huang,C.Y., Wu,C.Y. and Wu,W.S. (2011) YPA: An integrated repository of promoter features in Saccharomyces cerevisiae. *Nucleic Acids Res.*, **39**, 647–652.
9. De Boer,C.G. and Hughes,T.R. (2012) YeTFaSCo: a database of evaluated yeast transcription factor sequence specificities. *Nucleic Acids Res.*, **40**, 169–179.
10. Cherry,J.M., Hong,E.L., Amundsen,C., Balakrishnan,R., Binkley,G., Chan,E.T., Christie,K.R., Costanzo,M.C., Dwight,S.S., Engel,S.R. *et al.* (2012) Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.*, **40**, 700–705.
11. Mira,N.P., Becker,J.D. and Sá-Correia,I. (2010) Genomic expression program involving the Haa1p-Regulon in Saccharomyces cerevisiae response to acetic acid. *Omics: A J. Integr. Biol.*, **14**, 587–601.
12. Simões,T., Mira,N.P., Fernandes,A.R. and Sá-Correia,I. (2006) The SPI1 gene, encoding a glycosylphosphatidylinositol-anchored cell wall protein, plays a prominent role in the development of yeast resistance to lipophilic weak-acid food preservatives. *Appl. Environ. Microbiol.*, **72**, 7168–7175.
13. Schuller,C., Mamnun,Y.M., Mollapour,M., Krapf,G., Schuster,M., Bauer,B.E., Piper,P.W. and Kuchler,K. (2003) Global phenotypic analysis and transcriptional profiling defines the weak acid stress response regulon in Saccharomyces cerevisiae. *Mol. Biol. Cell*, **15**, 706–720.
14. Simões,T., Teixeira,M., Fernandes,A. and Sá-Correia,I. (2003) Adaptation of Saccharomyces cerevisiae to the herbicide 2, 4-dichlorophenoxyacetic acid, mediated by Msn2p-and Msn4p-regulated genes: important role of SPI1. *Appl. Environ. Microbiol.*, **69**, 4019.
15. Teixeira,M.C. and Sá-Correia,I. (2002) Saccharomyces cerevisiae resistance to chlorinated phenoxyacetic acid herbicides involves Pdr1p-mediated transcriptional activation of TPO1 and PDR5 genes. *Biochem. Biophys. Res. Commun.*, **292**, 530–537.
16. Alenquer,M., Tenreiro,S. and Sá-Correia,I. (2006) Adaptive response to the antimalarial drug artesunate in yeast involves Pdr1p/Pdr3p-mediated transcriptional activation of the resistance determinants *TPO1* and *PDR5*. *FEMS Yeast Res.*, **6**, 1130–1139.
17. Le Crom,S., Devaux,F., Marc,P., Zhang,X., Moye-Rowley,W.S. and Jacq,C. (2002) New insights into the pleiotropic drug resistance network from genome-wide characterization of the YRR1 transcription factor regulation system. *Mol. Cell. Biol.*, **22**, 2642–2649.
18. Monteiro,P.T., Pais,P., Costa,C., Manna,S., Sá-Correia,I. and Teixeira,M.C. (2017) The PathoYeastract database: an information system for the analysis of gene and genomic transcription regulation in pathogenic yeasts. *Nucleic Acids Res.*, **45**, D597–D603.