yMGV: a database for visualization and data mining of published genome-wide yeast expression data

Philippe Marc*, Frédéric Devaux and Claude Jacq

Laboratoire de Génétique Moléculaire, CNRS 8541, Ecole Normale Supérieure, 46 Rue d'Ulm, 75005 Paris, France

Received March 26, 2001; Revised and Accepted May 9, 2001

ABSTRACT

The yeast Microarray Global Viewer (yMGV) is an on-line database providing a synthetic view of the transcriptional expression profiles of *Saccharomyces cerevisiae* genes in most of the published expression datasets. yMGV displays a one-screen graphical representation of gene expression variations for each published genome-wide experiment, allowing quick retrieval of experimental conditions affecting expression of this gene. yMGV also provides tools to isolate groups of genes sharing similar transcription profiles in a defined subset of experiments. Additionally, yMGV furnishes a set of statistical tools for critical assessment of published data. We therefore believe that yMGV is an efficient tool that affords a quick and comprehensive overview of microarray data and generates new gene classifications. As of 20 March 2001 the yMGV database contains 6 000 000 measurements, representing genome-wide expression comparisons of 932 experiments from 39 microarray publications. The yMGV interface is available at http://transcriptome.ens.fr/ymgv/.

INTRODUCTION

In the past three years DNA microarrays have evolved from a sophisticated and precious technical advance, available only in a few pioneering laboratories, into a common technology accessible to numerous research groups. The number of published studies using microarrays has grown, generating an ever-increasing mass of data. Data mining constitutes a well-recognized challenge, especially when the data are scattered among numerous websites, when not incomplete or inaccessible. Some efforts dealing with specific aspects of data analyses have been made to construct databases from the published datasets. However, these databases either contain a limited number of published datasets [YPD (1) and Expression connection (2)] or only allow very specific requests [ExpressDB (3) and Webminer (4)].

If one wishes to consider a large variety of genome-wide expression data, the yeast *Saccharomyces cerevisiae* is especially appropriate, since more than 50 studies dealing with a large spectrum of biological conditions have been published and most of the authors have agreed to make their data available for a common database. We have developed a World Wide Web accessible database called the yeast Microarray Global Viewer (yMGV), which contains most of the results of published microarray experiments on the yeast *S.cerevisiae*. For the first time it is possible to view quickly and graphically the expression level of one particular yeast gene across nearly 1000 experiments on a single screen. This immediately indicates the experimental conditions that affect expression of any given *S.cerevisiae* gene. Such information is important in view of recent pioneering studies which have characterized new functional gene clusters, also described as synexpression groups (5–7). For this purpose yMGV provides a user-friendly query system to compare results from different publications, thus allowing users to select genes sharing a particular expression profile.

GENERAL USE OF yMGV

Datasets available in yMGV

The yMGV database contains datasets from most of the publications that have used microarrays to assess genomic expression in yeast. The data stored in yMGV are the ORF identifier and the filtered normalized Cy5/Cy3 ratio provided by the author for each experiment. To date (March 2001) the database contains 6 000 000 records representing 932 *S.cerevisiae* experiments published in 39 articles. All the data used are public data downloaded from the web or directly obtained with the agreement of the corresponding authors. A list of the available publications is shown in Table 1. The publication identifier is composed of the first author's name and a keyword summarizing the major topic of the publication. Complete information concerning each publication is available online. It includes a complete reference (title of the paper, list of authors and Medline reference) and direct links to the publication Medline page, PDF format article (when access is allowed) and website containing the original data (if available).

Using yMGV to detect conditions that affect the expression of a given gene

The interrogation form allows entry of a gene identifier [ORF name or gene name registered in the *Saccharomyces* Genome Database (8)] and selection of the set of publications one wishes to scan (Fig. 1A). This selection can be done via predefined groups containing related publications. Alternatively, a custom group containing a specific publication selection can be created. A filter option allows the setting of a cut-off (1.5, 2 or 3) for significance of the ratios, thus allowing data exploration according to one's own criteria. The 'aligned transcription

*To whom correspondence should be addressed. Tel: +33 1 44 32 35 46; Fax: +33 1 44 32 37 30; Email: pmarc@biologie.ens.fr

Table 1. Available datasets

Bernstein_deacetylase	Genomewide studies of histone deacetylase function in yeast	Proc Natl Acad Sci U S A 2000 Dec 5;97(25):13708-13
Causton_stress	Remodeling of Yeast Genome Expression in Response to Environmental Changes.	Mol Biol Cell 2001 Feb;12(2):323-337
Cavalieri_vineyard	Manifold anomalies in gene expression in a vineyard isolate of Saccharomyces cerevisiae revealed by DNA microarray analysis.	Proc Natl Acad Sci U S A 2000 Oct 24;97(22):12369-74
Cho_mitotic	A genome-wide transcriptional analysis of the mitotic cell cycle.	Mol Cell 1998 Jul;2(1):65-73
Chu_sporulation	The transcriptional program of sporulation in budding yeast.	Science 1998 Oct 23;282(5389):699-705
DeRisi_metabolic	Exploring the metabolic and genetic control of gene expression on a genomic scale.	Science 1997 Oct 24;278(5338):680-6
DeRisi_PDR	Genome microarray analysis of transcriptional activation in multidrug resistance yeast mutants.	FEBS Lett 2000 Mar 24;470(2):156-60
Diehn_membrane	Large-scale identification of secreted and membrane-associated gene products using DNA microarrays.	Nat Genet 2000 May;25(1):58-62
Eisen_temp_dtt	Cluster analysis and display of genome-wide expression patterns	Proc Natl Acad Sci U S A 1998 Dec 8;95(25):14863-8
Epstein_Mitochondrial	Genome-wide Responses to Mitochondrial Dysfunction	Mol Biol Cell 2001 Feb;12(2):297-308
Ferea_adaptative	Systematic changes in gene expression patterns following adaptive evolution in yeast.	Proc Natl Acad Sci U S A 1999 Aug 17;96(17):9721-6
Foury_YFH1	Mitochondrial control of iron homeostasis. A genome wide analysis of gene expression in a yeast frataxin deficient mutant	J Biol Chem 2000 Dec 8; (epub ahead of print)
Gasch_stress	Genomic expression programs in the response of Yeast cells to environmental changes	Mol Biol Cell 2000 Dec;11(12):4241-4257
Gerton_recomb	Global mapping of meiotic recombination hotspots and coldspots in the yeast Saccharomyces cerevisiae	Proc Natl Acad Sci U S A 2000 Oct 10;97(21):11383-90
Hardwick_rapamycin	Rapamycin-modulated transcription defines the subset of nutrient-sensitive signaling pathways directly controlled by the Tor proteins	Proc Natl Acad Sci U S A 1999 Dec 21;96(26):14866-70
Hughes_aneuploidy	Widespread aneuploidy revealed by DNA microarray expression profiling.	Nat Genet 2000 Jul;25(3):333-7
Hughes_compendium	Functional Discovery via a Compendium of Expression Profiles.	Cell 102, 109-126 (2000).
Hugues_noise	Functional Discovery via a Compendium of Expression Profiles.	Cell 2000 Jul 7;102(1):109-26
Jelinsky_MMS	Global response of Saccharomyces cerevisiae to an alkylating agent	Proc Natl Acad Sci U S A 1999 Feb 16;96(4):1486-91
Jelinski_Rpn4	Regulatory networks revealed by transcriptional profiling of damaged Saccharomyces cerevisiae cells: Rpn4 links base excision repair with	Mol Cell Biol 2000 Nov;20(21):8157-67
Kobor_CTD	An unusual eukaryotic protein phosphatase required for transcription by RNA polymerase II and CTD dephosphorylation in S. cerevisiae.	Mol Cell 1999 Jul;4(1):55-62
Linde_anaerobic	Genome-wide transcriptional analysis of aerobic and anaerobic chemostat cultures of Saccharomyces cerevisiae	J Bacteriol 1999 Dec;181(24):7409-13
Lopez_gcr1	Understanding the growth phenotype of the yeast gcr1 mutant in terms of global genomic expression patterns.	J Bacteriol 2000 Sep;182(17):4970-8
Marton_drug	Drug target validation and identification of secondary drug target effects using DNA microarrays.	Nat Med 1998 Nov;4(11):1293-301
Myers_med2	Mediator protein mutations that selectively abolish activated transcription.	Proc Natl Acad Sci U S A 1999 Jan 5;96(1):67-72
Ogawa_phosphate	New components of a system for phosphate accumulation and polyphosphate metabolism in Saccharomyces cerevisiae revealed by genomic expression analysis.	Mol Biol Cell 2000 Dec;11(12):4309-21
Pan_sok2	Sok2 regulates yeast pseudohyphal differentiation via a transcription factor cascade that regulates cell-cell adhesion	Mol Cell Biol 2000 Nov;20(22):8364-72
Posas_salt_hog1	The transcriptional response of yeast to saline stress.	J Biol Chem 2000 Jun 9;275(23):17249-55
Ren_Gal4Ste12	genome-wide location and function of DNA-binding proteins	Science 2000 290(5500):2306-9
Roberts_MAPK	Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles	Science 2000 Feb 4;287(5454):873-80
Robertson_A_kinase	The yeast A kinases differentially regulate iron uptake and respiratory function.	Proc Natl Acad Sci U S A 2000 May 23;97(11):5984-8
Scherbakova_MLH1	Inactivation of DNA mismatch repair by increased expression of yeast MLH1	Mol Cell Biol 2001 21(3):940-951
Shamji_tor	Partitioning the transcriptional program induced by rapamycin among effectors of the Tor proteins	Curr Biol 2000 Dec 1;10(24):1574-1581
Spellman_cellcycle	Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization.	Mol Biol Cell 1998 Dec;9(12):3273-97
Sudarsanam_snfswi	Whole-genome expression analysis of snf/swi mutants of Saccharomyces cerevisiae.	Proc Natl Acad Sci U S A 2000 Mar 28;97(7):3364-9
Tran_CHD1	The chromo domain protein chd1p from budding yeast is an ATP-dependent chromatin-modifying factor.	EMBO J 2000 May 15;19(10):2323-31
Travers_UPR	Functional and genomic analyses reveal an essential coordination between the unfolded protein response and ER-associated degradation	Cell 2000 Apr 28;101(3):249-58
Yale_salt	Transcript Expression in Saccharomyces cerevisiae at high salinity	JBC papers in press Manuscript M008209200
Zhu_forkheads	Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth.	Nature 2000 Jul 6;406(6791):90-4

Datasets currently accessible from yMGV (20 March 2001). An updated version is available online.

profiles' option applies the same search criteria to several genes at the same time (Fig. 1B). Once again, a one-screen representation allows direct comparison of the requested transcription profiles (Fig. 1D).

yMGV output histograms represent the expression profile of the requested gene in each selected publication. Base 2 logarithms have been used to make induction and repression effects directly comparable. One histogram is drawn for each

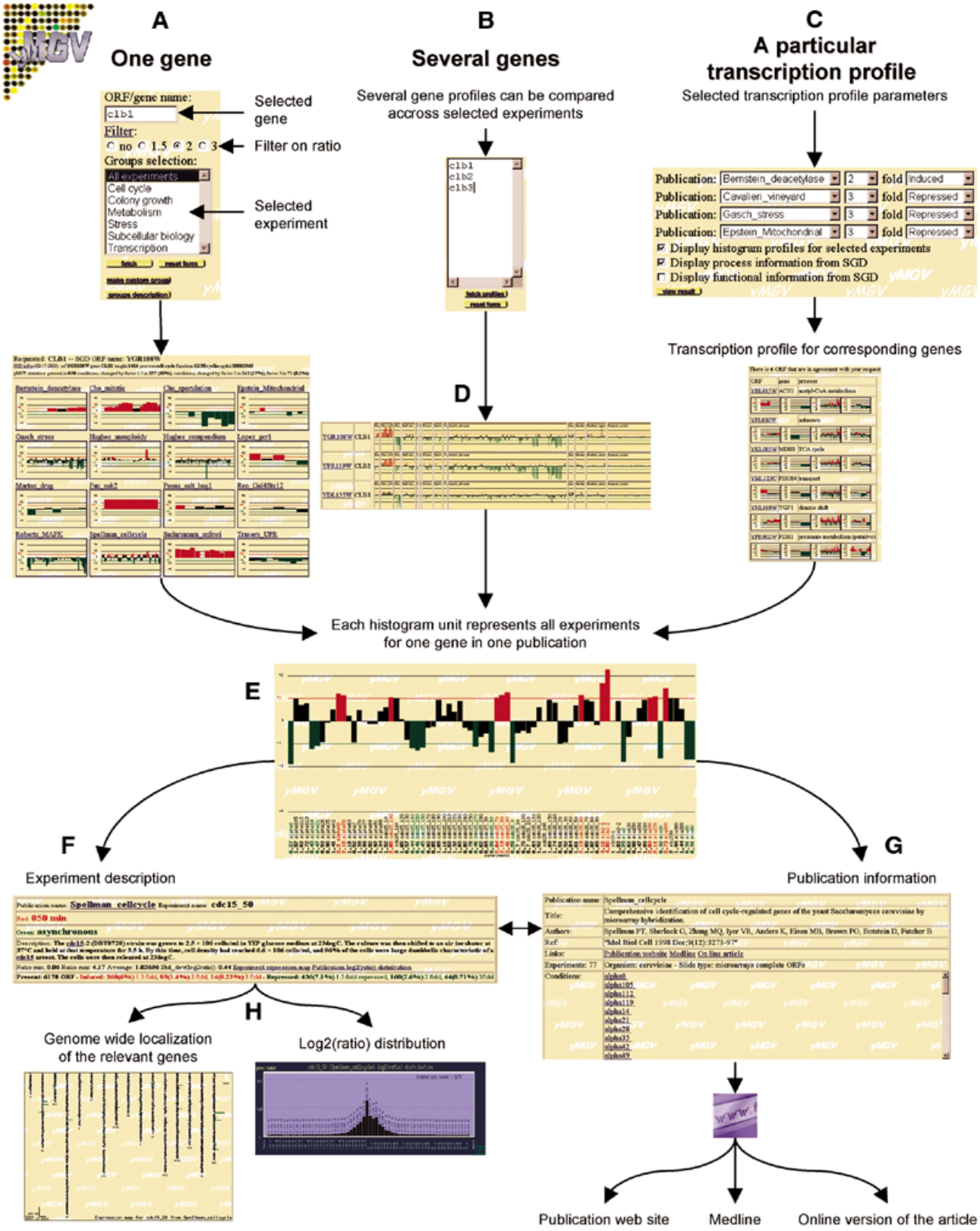

Figure 1. Main features of yMGV. The upper part of this figure indicates three major ways to use yMGV. (**A**) One gene can be selected and the expression profiles corresponding to the selected experiments appear in the second line. The red and green bars correspond, respectively, to a 2-fold increase or decrease in the ratio. (**B**) Several genes can be selected at the same time, thus allowing direct comparison of their expression profiles in the different microarray experiments. (**C**) One can impose specific parameters, for instance a 2-fold increase in ratio, select a subset of experiments and obtain the expression profiles of the corresponding genes. The histogram (**E**) corresponds to one microarray experiment. Each bar corresponds to one specific experiment for which details can be obtained directly. These details include not only a description of the exact conditions, the minimum, maximum and average ratio (**F**), but also a graphical representation of the log2 (ratio) distribution, the standard deviation and the genome-wide expression map for the experiment (**H**). Details on publications are directly available (**G**).

experiment in the selected publication. The histogram bars are red when the ratio is greater than 2 ($\log_2 > 1$) and green when the ratio is below 0.5 ($\log_2 < -1$) (these limits can be manually changed in zoom mode). Direct links to the publication Pubmed page, the article and the website containing the original data can be obtained by clicking on the publication identifier (Fig. 1G). A click on a histogram set reveals a full-screen version, including an experimental description and ratios (Fig. 1E). The experimental conditions are those reported in the original publication. A short comment on each experiment can be obtained by clicking on it (Fig. 1F). This comment includes strains and exact conditions used for each fluorochrome as well as statistics on the distribution of the ratios in the experiment.

Since the yMGV database contains the final filtered ratio, it is especially important to be able to rapidly access the relevant publication for assessment of the absolute levels of expression, together with the methods used to filter the data.

A FEW EXAMPLES OF SPECIFIC REQUESTS ADDRESSED TO yMGV

Search for genes sharing similar transcription profiles

Several recent genome-wide studies have highlighted the biological importance of group effects in transcription profile patterns. Genes that have similar expression profiles are called gene clusters (5) or synexpression groups (7). The premise of this guilt-by-association approach is that clustered genes may be co-regulated and therefore involved in similar functions. Versatile access to such analyses covering the available published data is important since there are many ways to consider the microarray data, depending on one's interests. yMGV allows the user to set up his or her own search criteria (selected set of experiments and significance ratio), for instance for all the ORFs exhibiting a 2-fold increase in expression in a first publication and a 3-fold decrease in two others (Fig. 1C). Such overlaps between the expression patterns of different experiments are likely to be of biological significance and give important clues to guide new research.

Statistical analyses of the published data

Few transcriptome analyses indicate the statistical significance of their results. yMGV provides simple statistical tools which allow the user to critically assess the expression profile changes between two tested conditions in any publication (Fig. 1H). For each experiment it is possible: (i) to see the number of genes whose expression level has been increased (or decreased) 1.5-, 2- or 3-fold; (ii) to draw a graphical representation of the $\log_2$ (ratio) distribution and to obtain the standard deviation of the distribution; (iii) to access maxima, minima and the average ratio. yMGV can also display a graphical representation of the genomic localization of the activated (or repressed) genes on the whole set of chromosomes. This representation of more or less transcriptionally active genomic domains is also available using all the datasets included in yMGV.

Search for marginally fluctuating gene expression

yMGV allows the user to select all the yeast genes that, for instance, have never been induced >1.5-fold in all the experiments considered. The whole dataset naturally does not cover all possible conditions; some genes are not represented on currently used microarrays (the corresponding list is available in yMGV) and poorly expressed genes are not faithfully analyzed in microarray experiments. Bearing these restrictions in mind, such non-fluctuating genes might be good candidates as artificial ORFs devoid of biological significance. The list of these genes will become increasingly accurate as the number of microarray experiments increases.

Analysis of highly fluctuating genes can also be interesting. Table 2 contains the 50 most induced genes across the 932 experiments in yMGV. A precise knowledge of this biological noise is, of course, important in the analysis of microarray data. Furthermore, this 'fluctuating effect' can be biologically relevant and a better definition of the genes involved is worth considering.

IMPLEMENTATION, LIMITATIONS AND FUTURE DEVELOPMENTS

Implementation

The yMGV interface is freely available via the internet at www.transcriptome.ens.fr/ymgv/. All the software used to power yMGV is distributed under an open source licence, i.e. anybody can download it totally free of charge from the World Wide Web (see www.opensource.org for more information). We plan to distribute yMGV database schema and scripts (including graphical administration) so that any laboratory can create its own local database containing private datasets.

Limitations

The first limitation of yMGV is that, despite all our efforts, a number of published datasets remain unreachable because the corresponding authors did not answer our request or refused to make their data publicly available. The existence of a public repository coupled to a well-defined publication policy should solve this problem. Such a public repository imposing a universal transfer format is urgently required for comparative approaches to microarray data. The establishment of universal standards for DNA array experiment annotation, data representation and universal controls is presently being elaborated (see the MGED group at www.mged.org). Prototypes using the first draft of the MGED group have already been constructed, but are not yet functional (ArrayExpress at www.ebi.ac.uk/arrayexpress/, GeneX at www.ncgr.org/research/genex/ and GEO at www.ncbi.nlm.nih.gov/geo/).

Future developments

The next step in yMGV development will be to pre-calculate cluster groups in each biological system as previously defined (cell cycle, metabolism, stress, etc.). This should help the user to establish the complete list of genes that share similar transcription profiles with the requested gene. Notably, this should help to suggest functions for orphan genes.

The second step in the yMGV project will be to add datasets from other organisms to the database. Connections between genes from different organisms via sequence similarities or via function assignment should permit the use of the vast amount of knowledge accumulated on yeast to find useful information on related genes in other model organisms.

Table 2. Highly fluctuating genes list

	Score	Present	Percent	Orf	Gene	Process	Function
1	436	921	47%	YFL014W	HSP12	glucose and lipid utilization	heat shock protein
2	204	514	40%	YMR173W-A		unknown	unknown
3	347	932	37%	YNL160W	YGP1	diauxic shift	unknown; response to nutrient limitation
4	318	928	34%	YML128C	MSC1	unknown	unknown
5	314	928	34%	YBR072W	HSP26	diauxic shift	stress-induced protein
6	296	906	33%	YML100W	TSL1	trehalose metabolism	trehalose-6-phosphate synthase/phosphatase complex regulatory subunit
7	306	930	33%	YMR250W	GAD1	glutamate metabolism	glutamate decarboxylase
8	296	926	32%	YMR105C	PGM2	glycolysis	phosphoglucomutase
9	296	927	32%	YNL036W	NCE103	secretion, non-classical	unknown
10	280	915	31%	YMR173W	DDR48	unknown	induced by DNA damage, heat shock, or osmotic stress
11	284	923	31%	YCL040W	GLK1	glycolysis	glucokinase
12	284	927	31%	YLR178C	TFS1	cell cycle	suppresses cdc25 mutations
13	287	928	31%	YER150W	SPI1	unknown	unknown; similar to Sed1p; induced in stationary phase
14	288	932	31%	YKL096W	CWP1	cell wall protein	beta-1,6-glucan acceptor
15	286	937	31%	YFR053C	HXK1	glycolysis	hexokinase I
16	280	919	30%	YBR054W	YRO2	unknown	putative heat shock protein
17	278	922	30%	YLR327C		unknown	unknown; similar to Stf2p
18	283	933	30%	YGR043C		unknown	unknown; similar to Tal1p
19	261	886	29%	YGR248W	SOL4	unknown	unknown; similar to Sol3p
20	260	900	29%	YHR215W	PHO12	phosphate metabolism	secreted acid phosphatase
21	269	929	29%	YBL064C		unknown	unknown; similar to Tsa1p
22	272	941	29%	YAR071W	PHO11	phosphate metabolism	secreted acid phosphatase
23	257	909	28%	YLR258W	GSY2	glycogen metabolism	glycogen synthase
24	257	912	28%	YGL255W	ZRT1	transport	high-affinity zinc transporter
25	262	926	28%	YHR087W		unknown	unknown
26	257	927	28%	YKL103C	LAP4	protein degradation	vacuolar aminopeptidase ysc1
27	259	927	28%	YML123C	PHO84	transport	inorganic phosphate permease
28	257	928	28%	YMR090W		unknown	unknown; similar to malate dehydrogenases
29	260	928	28%	YGL037C	PNC1	pyridine nucleotide cycle	pyrazinamidase and nicotinamidase
30	246	924	27%	YCL042W		unknown	unknown
31	246	927	27%	YOR289W		unknown	unknown
32	250	928	27%	YHL021C		unknown	unknown
33	254	929	27%	YDL124W		unknown	unknown
34	248	932	27%	YOR382W		unknown	unknown
35	240	913	26%	YOR374W	ALD4	ethanol utilization	mitochondrial aldehyde dehydrogenase
36	236	921	26%	YGL156W	AMS1	cell wall catabolism	vacuolar alpha-mannosidase
37	236	921	26%	YGR008C	STF2	ATP synthesis	ATPase stabilizing factor
38	240	926	26%	YHR137W	ARO9	aromatic amino acid metabolism	aromatic amino acid aminotransferase II
39	238	927	26%	YLL026W	HSP104	heat shock response	heat shock protein
40	239	928	26%	YDL204W		unknown	unknown
41	245	930	26%	YDR171W	HSP42	cytoskeleton assembly	heat shock protein, similar to HSP26
42	226	907	25%	YNL015W	PBI2	unknown	protease inhibitor
43	224	909	25%	YJL153C	INO1	inositol biosynthesis	L-myo-inositol-1-phosphate synthase
44	232	913	25%	YOR120W	GCY1	unknown	unknown; similar to mammalian aldo-keto reductases
45	235	923	25%	YER067W		unknown	unknown
46	235	925	25%	YDR070C		unknown	unknown
47	231	925	25%	YHR136C	SPL2	cell cycle	protein kinase inhibitor
48	234	927	25%	YDL021W	GPM2	glycolysis	phosphoglycerate mutase
49	230	927	25%	YGL121C		unknown	unknown
50	234	928	25%	YNL134C		unknown	unknown; similar to C. carbonum toxD gene

The top 50 most often 2-fold changed (up or down) ORFs in yMGV (as of 20 March 2001). An updated version is available online. Functional information is from the SGD. Score is the number of conditions where the ratio falls outside the range 0.5–2. Present is the number of conditions where the ORF was present in the array. Percent is score/present.

CONCLUSION

It is clear that the extensive scientific endeavor which aims at characterizing the transcriptome properties of several model organisms will lead to important new biological concepts only if reliable data are shared. yMGV is the first database offering a global view of existing yeast transcriptome data coupled to a simple interface. yMGV also provides direct access to yeast microarray data, allowing the user to elaborate his or her own interpretation of the published data. Moreover, with this large overview of the data it is possible to address global questions which overlap in several distinct experiments. In this respect new tools will be introduced into yMGV for better characterization of gene clusters. Furthermore, this database has been designed to store microarray data from different organisms. Such cross-talk between organism data will allow direct assessment of the progress and functional meaning of transcription regulation networks through evolution.

ACKNOWLEDGEMENTS

This work would not have been possible without the help of all the authors who made their data freely accessible. We thank GNU projects for providing such good free software. This work was supported by CNRS and by grants from ARC (no. 5691) and from Genopole Ile de France.

REFERENCES

1. Costanzo,M.C., Crawford,M.E., Hirschman,J.E., Kranz,J.E., Olsen,P., Robertson,L.S., Skrzypek,M.S., Braun,B.R., Hopkins,K.L., Kondu,P., Lengieza,C., Lew-Smith,J.E., Tillberg,M. and Garrels,J.I. (2001) YPD™,

PombePD™ and WormPD™: model organism volumes of the BioKnowledge™ library, an integrated resource for protein information. *Nucleic Acids Res.*, **29**, 75–79.

2. Ball,C.A., Jin,H., Sherlock,G., Weng,S., Matese,J.C., Andrada,R., Binkley,G., Dolinski,K., Dwight,S.S., Harris,M.A., Issel-Tarver,L., Schroeder,M., Botstein,D. and Cherry,J.M. (2001) *Saccharomyces* genome database provides tools to survey gene expression and functional analysis data. *Nucleic Acids Res.*, **29**, 80–81.
3. Aach,J., Rindone,W. and Church,G.M. (2000) Systematic management and analysis of yeast gene expression data. *Genome Res.*, **10**, 431–445.
4. Heiman,M.G. and Walter,P. (2000) Prm1p, a pheromone-regulated multispanning membrane protein, facilitates plasma membrane fusion during yeast mating. *J. Cell Biol.*, **151**, 719–730.
5. Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.
6. Hughes,T.R., Marton,M.J., Jones,A.R., Roberts,C.J., Stoughton,R., Armour,C.D., Bennett,H.A., Coffey,E., Dai,H., He,Y.D., Kidd,M.J., King,A.M., Meyer,M.R., Slade,D., Lum,P.Y., Stepaniants,S.B., Shoemaker,D.D., Gachotte,D., Chakraburtty,K., Simon,J., Bard,M. and Friend,S.H. (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.
7. Niehrs,C. and Pollet,N. (1999) Synexpression groups in eukaryotes. *Nature*, **402**, 483–487.
8. Cherry,J.M., Adler,C., Ball,C., Chervitz,S.A., Dwight,S.S., Hester,E.T., Jia,Y., Juvik,G., Roe,T., Schroeder,M., Weng,S. and Botstein,D. (1998) SGD: *Saccharomyces* Genome Database. *Nucleic Acids Res.*, **26**, 73–79.