TRIPLES: a database of gene function in *Saccharomyces cerevisiae*

Anuj Kumar, Kei-Hoi Cheung[1], Petra Ross-Macdonald, Paulo S. R. Coelho, Perry Miller[1] and Michael Snyder*

Department of Molecular, Cellular and Developmental Biology, Yale University, PO Box 208103, New Haven, CT 06520-8103, USA and [1]Center for Medical Informatics, Yale University School of Medicine, 333 Cedar Street, PO Box 208009, New Haven, CT 06520-8009, USA

Received August 31, 1999; Revised September 17, 1999; Accepted October 4, 1999

ABSTRACT

Using a novel multipurpose mini-transposon, we have generated a collection of defined mutant alleles for the analysis of disruption phenotypes, protein localization, and gene expression in *Saccharomyces cerevisiae*. To catalog this unique data set, we have developed TRIPLES, a Web-accessible database of TRansposon-Insertion Phenotypes, Localization and Expression in *Saccharomyces*. Encompassing over 250 000 data points, TRIPLES provides convenient access to information from nearly 7800 transposon-mutagenized yeast strains; within TRIPLES, complete data reports of each strain may be viewed in table format, or if desired, downloaded as tab-delimited text files. Each report contains external links to corresponding entries within the *Saccharomyces* Genome Database and International Nucleic Acid Sequence Data Library (GenBank). Unlike other yeast databases, TRIPLES also provides on-line order forms linked to each clone report; users may immediately request any desired strain free-of-charge by submitting a completed form. In addition to presenting a wealth of information for over 2300 open reading frames, TRIPLES constitutes an important medium for the distribution of useful reagents throughout the yeast scientific community. Maintained by the Yale Genome Analysis Center, TRIPLES may be accessed at http://ycmi.med.yale.edu/ygac/triples.htm

BACKGROUND

With its tractable genetics and fully-sequenced genome, the budding yeast *Saccharomyces cerevisiae* is an ideal model organism for functional genomics; several large-scale studies are already in place to identify cellular functions for each of the 6200 predicted genes within the *Saccharomyces* genome (1–3). These genomic studies, however, possess finite limitations: it is unlikely that any single large-scale project will be completely successful in exhaustively characterizing gene function, as many genes do not generate easily observable phenotypes upon disruption (4,5). Functional genomic approaches, therefore, will need to be augmented with traditional research from individual laboratories studying a single gene or pathway (4). Following this paradigm, we have developed a transposon-based mutagenesis system facilitating both genomic and traditional research approaches: applied to *S.cerevisiae*, our method has yielded an unprecedented quantity of functional data while generating thousands of reagents for further use by researchers throughout the yeast community.

Our transposon-tagging strategy utilizes a multifunctional minitransposon (mTn) to mutagenize a plasmid-based library of yeast genomic DNA in *Escherichia coli* (6,7). Transposon-mutagenized genomic DNA is subsequently transformed into yeast, generating a collection of mutant strains, each carrying a single mTn insertion at a defined site within the *Saccharomyces* genome. This strain collection (available from our web site) can facilitate a wide variety of functional studies. Using mTn-encoded *lacZ* as a reporter, we can determine when each transposon-tagged gene is expressed during the yeast life cycle (e.g., during vegetative growth or sporulation). Additionally, mTn insertion will create a truncation of the mutagenized gene, thereby generating disruption alleles for phenotypic analysis. Finally, the inserted transposon can be modified *in situ*, reducing the mTn to a 93-codon in-frame tag encoding three tandem copies of an epitope from the influenza virus hemagglutinin protein. Epitope-tagged proteins may be localized within the cell by indirect immunofluorescence. A single transposon insertion, therefore, is sufficient to generate three types of data regarding gene function: expression profiles, disruption phenotypes and protein localization.

To catalog these three data sets, we have developed TRIPLES, an on-line database of TRansposon-Insertion Phenotypes, Localization and Expression in *Saccharomyces*. The TRIPLES database has been designed to offer easy access to all data generated from the functional analysis of our strain collection. At present, this collection encompasses nearly 7800 mTn-insertion alleles available in each of three forms: as diploid yeast strains containing *lacZ*-fusions generated from mTn insertion, as diploid yeast strains containing mTn-encoded epitope tags, and as bacterial clones carrying plasmid-borne mTn-mutagenized yeast DNA. Users may request these strains free of charge from order forms linked to TRIPLES. The TRIPLES database also contains external links to the

*To whom correspondence should be addressed. Tel: +1 203 432 6139; Fax: +1 203 432 6161; Email: michael.snyder@yale.edu

Saccharomyces Genome Database (SGD) (8) and GenBank (9), thereby providing full background literature concerning all transposon-tagged genes within our collection. Finally, TRIPLES allows access to a set of strains carrying mTn insertions identifying non-annotated open reading frames (NORFs) within the yeast genome; these putative genes represent a potentially rich source of novel proteins in *Saccharomyces* (10,11). Collectively, the data and reagents made available through TRIPLES constitute a unique information resource designed to promote ongoing research within the yeast community.

DESIGN AND IMPLEMENTATION

To maximize portability, we have distinguished back end development of the TRIPLES database from its front end development (user interface design). The back end TRIPLES database was designed using the ORACLE relational database system, version 7.3; our front end user interface was implemented using ASP (Active Server Page), an integral part of the Microsoft IIS (Internet Information Server) Web server running on Windows NT. The ASP mechanism has enabled us to embed server-side code written in VBScript (Visual Basic Script) and Javascript within HTML documents. We have minimized client-side scripting to avoid problems resulting from incompatibility between different types and versions of web browsers. To ensure code compatibility with different database platforms, we have used ODBC (Open Database Connectivity) to implement database access.

To facilitate maintenance of necessary background tasks without complicating external use, TRIPLES houses both a private and public area. Members of the Yale Genome Analysis Center may enter a password-protected area of TRIPLES to upload constituent data files using a web-based interface; this same interface can be used to mine and process data for subsequent display in tabular output reports. These publicly accessible reports present data generated specifically from this project in either composite or category-specific formats.

COMPOSITE REPORTS

Users may obtain a composite report of phenotypic, expression and protein localization data for any given strain through queries by gene name or clone ID. Gene names may be entered in either systematic (e.g., YLL021W) or standard form (e.g., *SPA2*); TRIPLES recognizes all SGD-accepted synonyms for any annotated gene. Each strain within our collection may also be identified by its clone ID, a unique alphanumeric tag (e.g., V102A1) assigned to a given strain based upon its position in a 96-well storage plate. Queries using either field return reports listing selected clone IDs in table format, each row specifying the exact genomic site of mTn insertion per clone. Users may ‘click’ upon any clone ID within this table to generate a corresponding composite report; each report presents on a single page all data from this project regarding a given mutant allele. These composite reports feature data defining the precise point of mTn insertion within the yeast genome as well as a listing of each ORF potentially disrupted by this insertion. Additionally, phenotypic, expression and protein localization datasets are conveniently displayed as individual tables accompanied by full on-line help. When available, supplemental background literature may be accessed from direct external links to corresponding entries within the SGD and GenBank. To facilitate distribution of our strain collection, researchers may request any desired reagent by simply completing an order form linked to each composite report page; identical order forms may also be accessed from each category-specific report as well.

CATEGORY-SPECIFIC SEARCHES

Within TRIPLES, individual datasets may be accessed by performing any of five category-specific searches (search demonstrations are offered on the Yale Genome Analysis Center home page at http://ycmi.med.yale.edu/ygac/home.html). Users may search TRIPLES for clones carrying an mTn insertion within a specific gene or NORF, clones exhibiting a specific pattern of *lacZ* expression, clones exhibiting a disruption phenotype of interest, or clones carrying epitope-tagged proteins localized to a given subcellular site. Each search may be launched by supplying a gene name or clone ID. Alternatively, category-specific data may be obtained via queries using any item within a list of controlled vocabulary terms descriptive of each dataset. Output reports may be custom-formatted, eliminating any superfluous data fields. In addition, category-specific output may be sorted by data fields as a means of grouping search results in a logical manner. All search results are returned as tables available for download in tab-delimited format; each clone ID within this table may be clicked upon to generate a corresponding composite report for that clone. To facilitate multi-level searching, the results of a given search may be used to initiate further category-specific searches, each subsequent search generating a data table as described. To clarify available search options within TRIPLES, each specific dataset is discussed in greater detail below.

Identification of ORFs and NORFs potentially disrupted by mTn insertion

Due to the structural complexity of the *S.cerevisiae* genome, it is not always possible to conclusively identify a single gene disrupted by mTn insertion. A single transposon insertion event may disrupt one or more open reading frames (ORFs) within a given region of the genome; for example, two ORFs may overlap each other such that mTn insertion within this region would disrupt both genes. Moreover, a number of potential ORFs within the yeast genome were never annotated (typically due to their small size). Over 100 NORFs have already been identified as expressed sequences (10); our transposon-tagging approach provides an effective means of identifying such NORFs on a large scale (Table 1). Therefore, to characterize transposon insertion events as accurately as possible, we have developed a simple script to identify all annotated as well as NORFs potentially truncated by mTn insertion. Annotated ORFs are displayed in table format within each composite report and ‘insertion point’ category-specific search. All clones identifying putative NORFs may be studied by performing a category-specific search of NORF data. Specific clones carrying a transposon insertion within a potential NORF may be selected through queries by clone ID. Alternatively, NORFs identified within our study may be searched by specifying a nearby annotated gene within the ‘genome region’ field; the resulting search would return any strains carrying mTn insertions in a potential NORF located within 200 bp 5′ or 3′ of that annotated gene.

Table 1. TRIPLES data sets (as of August 1999)

Data sets	Entries	Numbers
mTn insertion point data	Total clones	7749
	Affected genes	2347
NORF data	Total clones	1795
Gene expression data	Total clones	7581
	Induced during vegetative growth	7539
Phenotypic data	Total clones	6611
	Conditions tested per clone	21
	Strains with observed mutant phenotypes	3138
Protein localization data	Total clones	6970
	Subcellular localizations (not background)	1050

The number of ‘total clones’ within each data set indicates the total number of transformants tested for each data type; the number of total records within each data set may exceed the number of total clones, as some clones contain multiple data entries.

Interestingly, a variety of mTn insertion events are responsible for *lacZ* expression within our strain collection. The majority of these strains carry mTn insertions representing in-frame *lacZ* fusions to annotated as well as non-annotated ORFs. We have, however, identified a number of strains carrying productive mTn insertions oriented in the antisense direction or in the +1 or –1 reading frame. Additionally, a significant number of productive *lacZ* fusions result from mTn insertions within genomic sequences not expected to encode protein products (e.g., rDNA genes, sequence 5′ and 3′ of predicted genes). Strains exhibiting these insertion events may be selected by performing a category-specific search of ‘insertion point data’ and/or ‘NORF data’; appropriate queries may be selected from a list of descriptive terms within the ‘insertion’ field. Such category-specific searches of mTn insertion events are a particularly efficient means of identifying strains of interest for further study.

Gene expression data

Expression profiles of transposon-tagged genes may be generated using mTn-encoded *lacZ* as a reporter (6). In our study, strains carrying a transposon insertion were incubated on rich medium and sporulation medium to identify genes induced during vegetative growth and meiosis, respectively (Table 1); corresponding strains were assigned clone IDs beginning with a ‘V’ or ‘M’. To control for starvation-induced genes, clones carrying mTn-tagged genes potentially induced during meiosis were also grown on synthetic medium lacking glucose or nitrogen; corresponding strains carrying genes induced under starvation conditions were assigned clone IDs beginning with a ‘G’, ‘N’ or ‘GN’, respectively. Strains were subsequently assayed for β-gal activity using a filter-based method; intensity of staining has been scored qualitatively as follows: background, faint, medium, strong. These expression levels may be used to launch a category-specific search of gene expression data; expression data searches may also be initiated by specifying a particular gene name, clone ID, or growth condition (either ‘vegetative’, ‘sporulation’ or both). Data output is returned as a table of all selected clones fulfilling specified search criteria—available for download in a tab-delimited format. As described, each highlighted clone ID may be selected to generate a full composite report for that clone; selected clones may be chosen as input for further category-specific searches.

Phenotypic data

TRIPLES houses over 120 000 data points generated from phenotypic analysis of haploid yeast strains carrying mTn-mutagenized genes (Table 1). Transposon-mutagenized strains were scored for 21 phenotypes following growth under appropriate test conditions (described in the supplementary material accompanying this article). Disruption phenotypes were scored by comparison of transformant growth on selective media as compared against growth of the same transformant on standard culturing medium. Differences in growth were scored as follows: ‘ND’ and ‘WT’ indicate identical growth patterns, while ‘weak’, ‘medium’ and ‘strong’ define differing levels of growth abnormality. These scores and assays may be used to initiate category-specific searches of phenotypic data; additionally, phenotypic data may be searched by specifying a particular gene or clone ID as before. Output may be custom-formatted as desired prior to being downloaded in a tab-delimited format. These downloaded tables offer a convenient means of sorting genes for further functional study. All yeast mutants available to the public are derived from the diploid strain Y800 (7); to obtain transformants in a haploid background, corresponding mTn-mutagenized yeast DNA (available from us in plasmid form) may be transformed into any desired *URA3* haploid strain.

Protein localization data

Using monoclonal antibodies directed against the mTn-encoded hemagglutinin epitope, we have analyzed nearly 7000 transposon-tagged proteins by indirect immunofluorescence (Table 1). In particular, TRIPLES contains complete data sets for over 1000 strains carrying epitope-tagged yeast proteins localized to a subcellular site. These and other subsets of localized proteins may be accessed by performing a category-specific search of protein localization data. Searches may be initiated by supplying a gene name, clone ID or subcellular localization; specific localizations may be selected from a list of available localization patterns. All clones carrying epitope-tagged proteins that localized to a discrete cellular site were tested a minimum of two times (indicated in the ‘trials’ data field).

SIGNIFICANCE

The strain collection presented in TRIPLES will contribute to our current understanding of yeast genetics in two ways. At present, TRIPLES houses sufficient data to help infer cellular functions associated with hundreds of previously uncharacterized genes—a significant fraction of the yeast genome. TRIPLES, therefore, offers a sizable contribution to the current knowledge base of gene function in *Saccharomyces*. Secondly, our strain collection is an important source of reagents fostering both genomic as well as traditional studies. Researchers may easily apply genomic approaches to screen our strain collection for mutants exhibiting a desired disruption phenotype; alternatively, labs studying a biochemical process may benefit from intensive study of a few mutant strains carrying mTn insertions within a gene of interest. To facilitate either mode of research, however, our strain collection must be freely available to any interested laboratory. To date, we have already distributed over 400 mTn-mutagenized strains to researchers throughout the yeast community; the TRIPLES database should serve as an ideal channel for the further dissemination of these resources and information in the immediate future.

ACCESS

TRIPLES may be accessed from the Yale Genome Analysis Center (YGAC) home page (http://ycmi.med.yale.edu/ygac/home.html) or directly at http://ycmi.med.yale.edu/ygac/triples.htm . User support may be obtained from the YGAC staff at either ygac@yale.edu or anuj.kumar@yale.edu ; please direct all technical concerns and questions to these addresses as well. When referencing the TRIPLES database, please cite this article.

INDEX OF SUPPLEMENTARY MATERIAL

- Summary of TRIPLES data sets

mTn insertion point data	Summary and description
NORF data	Summary and description
Gene expression data	Summary and description
Phenotypic data	Summary and description
Protein localization data	Summary and description

ACKNOWLEDGEMENTS

This work was supported by grant CA77808 from the National Institutes of Health to M.S and NIH grant G08 LM 05583 from the National Library of Medicine. P.S.R.C. is supported by the Fundacao de Amparo a Pesquisa do Estado de San Paulo, Brazil.

REFERENCES

1. Winzeler,E.A., Shoemaker,D.D., Astromoff,A., Liang,H., Anderson,K., Andre,B., Bangham,R., Benito,R., Boeke,J.D., Bussey,H. *et al.* (1999) *Science*, **285**, 901–906.
2. Smith,V., Chou,K.N., Lashkari,D., Botstein,D. and Brown,P.O. (1996) *Science*, **274**, 2069–2074.
3. Fromont-Racine,M., Rain,J.C. and Legrain,P. (1997) *Nature Genet.*, **16**, 277–282.
4. Johnston,M. (1996) *Trends Genet.*, **12**, 242–243.
5. Hodges,P.E, McKee,A.H.Z., Davis,B.P., Payne,W.E. and Garrels,J.I. (1999) *Nucleic Acids Res.*, **27**, 69–73. Updated article in this issue: *Nucleic Acids Res.* (2000), **28**, 73–76.
6. Burns,N., Grimwade,B., Ross-Macdonald,P.B., Choi,E.-Y., Finberg,K., Roeder,G.S. and Snyder,M. (1994) *Genes Dev.*, **8**, 1087–1105.
7. Ross-Macdonald,P., Sheehan,A., Roeder,G.S. and Snyder,M. (1997) *Proc. Natl Acad. Sci USA*, **94**, 190–195.
8. Chervitz,S.A., Hester,E.T., Ball,C.A., Dolinski,K., Dwight,S.S., Harris,M.A., Juvik,G., Malekian,A., Roberts,S., Roe,T.Y. *et al.* (1999) *Nucleic Acids Res.*, **27**, 74–78. Updated article in this issue: *Nucleic Acids Res.* (2000), **28**, 77–80.
9. Benson,D.A., Boguski,M.S., Lipman,D.J., Ostell,J., Ouellette,F., Rapp,B.A. and Wheeler,D.L. (1999) *Nucleic Acids Res.*, **27**, 12–17. Updated article in this issue: *Nucleic Acids Res*. (2000), **28**, 15–18.
10. Velculescu,V.E., Zhang,L., Zhou,W., Vogelstein,J., Basrai,M.A., Bassett,D.E.,Jr, Hieter,P., Vogelstein,B. and Kinzler,K.W. (1997) *Cell*, **88**, 243–251.
11. Olivas,W.M., Muhlrad,D. and Parker,R. (1997) *Nucleic Acids Res.*, **25**, 4619–4625.