



ELSEVIER

Mechanisms of Development 121 (2004) 959–963



www.elsevier.com/locate/modo

Technical report

GSD: a genetic screen database

Thorsten Henrich^{a,b,*}, Mirana Ramialison^{a,b}, Erik Segerdell^c, Monte Westerfield^c,
Makoto Furutani-Seiki^b, Joachim Wittbrodt^a, Hisato Kondoh^b

^aEMBL Heidelberg, Meyerhofstr. 1, D-69117 Heidelberg, Germany

^bJapan Science and Technology Corporation, ERATO Kondoh differentiation signaling project, Kinki-chihou Hatsumei Center Building,
Yoshida-Kawaramachi 14, Sakyo-ku, Kyoto 606-8305, Japan

^cZFIN, University of Oregon, Eugene, OR 97403-5274, USA

Received 22 December 2003; received in revised form 23 February 2004; accepted 25 February 2004

Abstract

The systematic assignment of gene function to a sequenced genome is one of the outstanding challenges in the post-genomic era. Large-scale systematic mutagenesis screens are important tools for reaching this goal. Here we describe GSD, a software package that allows storage and integration of data from genetic screens. GSD was initially developed for a large-scale F3 mutagenesis screen for developmental mutants of medaka (*Oryzias latipes*). The version presented here supports a wide range of different screens (mutagenesis, RNAi, morpholinos, transgenesis and others) using different organisms.

Data are stored in a relational database and can be made accessible through web interfaces. Researchers can enter data describing their screened embryos: They can track statistics, submit images and describe the resulting phenotypes using a phenotype classification ontology. We developed a fish phenotype classification ontology of medaka and zebrafish for this software package and made it available to the public.

In addition, a list of genetic lines resulting from each screen can be generated. These lines (mutant alleles, transgenic lines) can be described and categorized in the same ways as the screened individuals. Raw data from the screen can be integrated to describe these lines. A query module that searches this list can be used to publish the screen results on the Internet.

A test version is available at <http://www.embl.de/wittbrodt/gsd> and the software can be downloaded from this site.

© 2004 Elsevier Ireland Ltd. All rights reserved.

Keywords: Mutant screen; Medaka; Zebrafish; Phenotype; Ontology

1. Introduction

A major goal of the post-genomic era is to elucidate and study the functions of genes and to annotate gene functions in a standardized way. Mutant screens have been very successful in helping discover gene functions and have been conducted with important animal model systems, such as *C. elegans*, *Drosophila*, zebrafish (Haffter et al., 1996), medaka (Loosli et al., 2000) and mouse (Hrabe de Angelis et al., 2000). Databases have been established to provide results from these screens to the scientific community in an organized way (e.g. WormBase for *C. elegans* (Harris et al., 2003), FlyBase for *Drosophila* (FlyBase-Consortium,

2003), ZFIN for zebrafish (Sprague et al., 2003) and MGI for mouse (Blake et al., 2003)). However, a major difficulty for database curators is the annotation of mutant phenotypes resulting from different sources that were analyzed by different researchers.

Here we describe a software package that can be used to describe phenotypes in a standardized way and to organize genetic screen data. We developed a phenotype classification ontology for fish (medaka and zebrafish) and similar ontologies for other model systems can be implemented easily.

The interface can provide screeners with a standardized screening checklist. Researchers can evaluate the screen on a daily basis (e.g. number of screened genomes, frequency of mutations per genome). The huge amount of data gathered during a screen can be used directly to describe, categorize and organize lists of identified lines.

* Corresponding author. EMBL Heidelberg, Meyerhofstr. 1, D-69117 Heidelberg, Germany. Tel.: +49-6221-387-516; fax: +49-6221-387-166.
E-mail address: henrich@embl.de (T. Henrich).

2. Data and data access

2.1. Screen identity

Embryos are grouped in clutches. A clutch of embryos is produced by a single cross at a particular time (approximately 20 eggs/clutch for medaka or 100/clutch for zebrafish) yielding synchronously developing individuals. From the time data are entered, the clutch ‘exists’ in the database and its embryos can be followed and described throughout development.

Clutches are linked to the cross that produces them and crosses are linked to the families from which they originate (Fig. 1). Clutch names are generated automatically from their cross names and cross names are generated from family names.

Data stored for a clutch include: date of fertilization (embryo collection), user name of the collector, number of embryos, number of dead embryos, identifier for the family, generation and cross number. Families are the source of the screen. They could be (depending on the kind of screen) the offspring of a single mutagenized individual, siblings that have been injected with a construct or a clutch of embryos treated with a reagent such as RNAi.

2.2. Screen phenotype

The user can enter a description of a phenotype for a single embryo in a clutch at any particular developmental stage because GSD stores information about phenotypes with a resolution of a single embryo and a single developmental stage (Fig. 1). In addition to the frequency (numbers of wild-type, number of embryos that died during development or mutant embryos), phenotypes can be

specified by images and parameters like screening temperature (important for temperature sensitive mutants), screening method (morphology, mRNA in situ or antibody screen), text comments and descriptive terms defined by a phenotype classification ontology, such as the one we have developed for medaka and zebrafish.

2.3. Line identity

A line represents the result of a screen. It can be a mutant allele or a transgenic line identified in the screen. A line has a name that usually refers to the phenotype (e.g. *eyeless (el)*) as well as a family reference that refers to the family from which the line has been identified. The family reference consists of the family name and an arbitrary letter that makes this reference unique to distinguish it from other lines that may originate from the same family (Fig. 1). It can be left empty to provide a simple list of lines, for example when screening data are not available. Lines that fail to complement one another genetically are assigned the same name but different allele numbers and family references.

2.4. Line phenotype

Established lines or mutant alleles can be described in the same format as screened embryos. Phenotypes are specified by images, an ontology description and other parameters like genomic information (Fig. 1). To simplify data entry, the original screening data can be accessed and reused when describing a line or when selecting representative images. Using the family name, a list of crosses can be selected in which this line was discovered. This list of crosses enables the database to link information about a line to the original screen data.

2.5. Fish gene expression and phenotype ontologies

To enable cross-species comparisons of gene expression patterns and mutant phenotypes between medaka and zebrafish, we developed ontologies for medaka and zebrafish (Table 1) and submitted them to ‘Open Biological Ontologies’ OBO (<http://obo.sourceforge.net/>) where they can be downloaded in DAGedit format. During production of these ontologies, we took great care to use the same terms for corresponding structures in both fish species and, wherever possible, in mouse and *Drosophila*. We have also developed a set of translation tables that define the relationships among terms that differ among these species.

We developed three different ontologies, hierarchical sets of developmental terms, anatomical terms and phenotypic terms. Each term has a unique identifier: MFO (Medaka Fish Ontology), ZDB (Zebrafish Database). We annotate individual database records by defining an intersection of these three orthogonal ontologies.

The first ontology defines the terms and their relationships used to describe **developmental stages**. The medaka

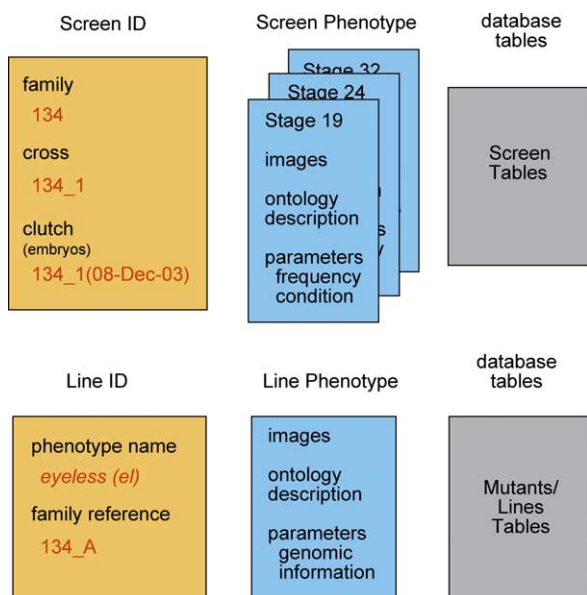


Fig. 1. Identifiers: Human readable identifiers used in GSD in the screen (top) and mutant database (bottom).

Table 1
Fish ontologies

<i>Medaka fish ontologies (identifier space: MFO)</i>			
Ontology	Terms	Source	Used in DB
Developmental stages	46 terms	From Iwamatsu (1994)	GSD, MEPD
Anatomy	4173 terms	New	MEPD
Phenotype classification	106 structure terms 29 modifier terms	New	GSD, Kyoto Mutant DB
<i>Zebrafish ontologies (identifier space: ZDB)</i>			
Developmental stages	44 terms	From Kimmel (1995)	ZFIN, GSD
Anatomy	1156 terms	ZFIN	ZFIN
Phenotype classification	105 structure terms 29 modifier terms	New	GSD, ZFIN supported

developmental stages ontology contains 46 terms as originally described by Iwamatsu (Iwamatsu, 1994). The zebrafish developmental ontology is based on the staging series from Kimmel (Kimmel et al., 1995) and has been expanded to 44 terms.

The second ontology defines the terms used to describe **anatomy** and has 4173 terms for medaka and 1156 terms for zebrafish. These ontologies are already used to describe endogenous gene expression patterns in ZFIN <http://zfin.org> (Sprague et al., 2003) and in the Medaka Expression Pattern Database MEPD <http://www.embl.de/mepd/> (Henrich et al., 2003), and can be implemented in GSD to describe other expression patterns, for example GFP expression in transgenic lines.

The third ontology defines the terms used to describe mutant **phenotypes**. During an initial screen, it is usually not possible to describe an embryo in detail. Researchers typically record approximate descriptions that allow categorization of phenotypes, thus supporting discovery of the functional relatedness of the mutated genes. For this reason, we developed a relatively high level, **phenotype classification** ontology that contains a subset of terms that have been most useful for large-scale screens. The ‘deeper’ phenotype ontology, using the full anatomy and developmental stages ontologies, is probably more useful for detailed annotation of mutant records in model organism databases, such as ZFIN. The high-level phenotype ontology includes 106 anatomical structure terms (105 for zebrafish) and 29 modifier terms. An *anatomical structure* term can be further specified by a list of appropriate *modifiers*. For example, the anatomical term ‘eyes’ that ‘is-part-of’ the ‘sensory system’ can be modified by the terms ‘abnormal, absent, enlarged, reduced, cyclopic’. We developed this ontology based on our experience in a large-scale medaka mutagenesis (Kyoto screen 2004) screen as well as with the mutants described in ZFIN (Sprague et al., 2003) and the Boston (Driever et al., 1996) and Tübingen (Haffter et al., 1996) zebrafish screens.

We defined common phenotypic terms including shared anatomical terms to enable cross-references between the two databases (GSD and MEPD). The ontologies are

represented directed acyclic graphs (DAG) that are similar to hierarchies, but allow more than one parent for each node. The graphs were developed using DAGedit (<http://www.geneontology.org/doc/GO.tools.html>) and self-written tools. The ontologies implement the three different relationships among terms that are also supported by gene ontology (The-Gene-Ontology-Consortium, 2001): ‘is-part-of’, ‘is-a’ and ‘develops-from’. We made an effort to use orthogonal modifier terms, meaning terms that can be applied to all or many anatomical terms (e.g. abnormal, absent, enlarged) but in a few cases we needed non-orthogonal specific modifiers (e.g. cyclopic).

3. Implementation and results

The software is designed for use in a standard 3-tier model. The client uses a browser to send an HTML request to the web server. Servlets handle the request, query the relational database, compute the result set and compose a response for the client.

A prototype version for testing the software has been installed at: <http://www.embl.de/wittbrodt/gsd>. The GSD software is freely available for academic use and a commercial license can be obtained for commercial users. A link to this site will be added to the new update of the Molecular Biological Database Collection (Galperin, 2004) to direct potential users to the GSD package.

3.1. Database

We provide SQL scripts to create the tables through SQL commands and scripts to insert a) a species specific *phenotype classification ontology* (medaka, zebrafish) b) a species specific *developmental stages ontology* (medaka, zebrafish) and c) data to describe users and their access privileges.

The tables describing the phenotypes of individual embryos (Fig. 2, green tables) and lines (Fig. 2, orange tables) have similar relationships. Phenotypes for both embryos and lines are described using the same reference

GSD: Entity Diagram

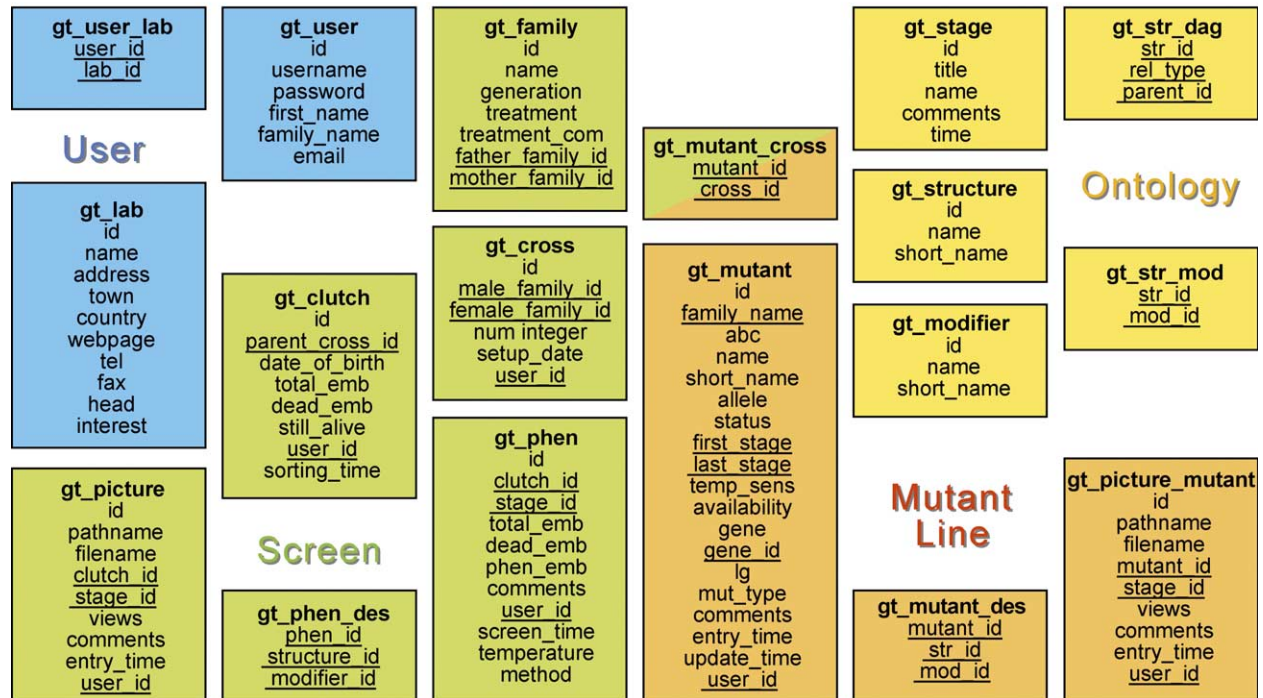


Fig. 2. Entity diagram of GSD. Columns representing foreign keys in the tables are underlined; relationships are not shown.

tables including images and the ontologies for the developmental stages and the phenotype classification (Fig. 2, yellow tables). This makes it easy to use screen data for mutant descriptions. These two domains are linked through the table *gt_mutant_cross* (Fig. 2) that contains a list of crosses for a mutant that were used to establish the mutant line.

3.2. Web interfaces

We provide the source code of the JAVA servlet web modules: three for entering data and five to retrieve data. For each entry to the database, information about the update time and the user who entered the data is recorded.

3.2.1. Data submission

- The **Register Family** module enables the user to enter general information about a clutch such as the origin (family name, parent cross, embryo collector, date of fertilization) and numbers of total and dead embryos.
- The **Screen DB** module can be used to describe the phenotypes of the screened embryos using a species dependent *phenotype classification ontology* and comments. The conditions and screening methods are recorded and images can be submitted.
- The **Mutant DB** module is used to describe a mutant allele or other line. Data (descriptions and images) gathered during the screen can be used to complete the description of the mutant phenotype.

3.2.2. Data query

- The **Search Screen DB** module is used to search data entered with the 'Register Family' and 'Screen DB' modules.
- The **Search Mutant DB** module is used to search data entered with the 'Mutant DB' module. A copy of this module **Search Public Mutant DB** does not have password protection and displays information about mutants that have been labeled 'public'. It can be used by a research group to publish their list of mutants to the scientific community via the Internet.
- The **Developmental Stages** module displays the ontology terms that describe the developmental stages of the organism.
- The **Phenotype Classification** module displays the ontology terms used to classify the observed phenotypes.

4. Discussion

The software package presented here originates from a database that was specifically designed for a large-scale medaka ENU F3 mutagenesis screen. During development of this database, we realized that with a few changes it could be easily adapted for many other applications. Hence, we modified it to be species independent, screen method independent and platform independent.

We achieved **species independence** by implementing the database in a modular way. By substituting the species

dependent ontologies for the developmental stages and the phenotype classification, this software package can be used for other organisms. Here we developed an ontology for fish. Because similar ontologies exist for other model organisms such as *Drosophila*, *C. elegans* and mouse, GSD can easily be adapted for use during genetic screens in these organisms. Once the proper ontologies have been entered into the database tables, the web interfaces dynamically display them when needed for describing phenotypes. However, GSD requires that the individuals described in the same clutch develop synchronously.

We developed the software for a large-scale ENU mutagenesis screen, but the current version can be applied as well to other screens (**screen method independence**) including phenotypes induced by knockout, P-element insertion or morpholino injection. For any particular application, a unique identifier is required for each family with a similar genetic origin. For specific screens (e.g. eye development), the ontologies can be extended to provide higher resolution in the system of interest while still allowing classification of general defects. GSD can even be used independently of a screen, for example to maintain and track mutant or transgenic lines.

We used the JAVA programming language in combination with SQL and HTML forms to support **platform independence**. Our JAVA servlets were developed on a Tomcat server and were tested on an IBM WebSphere server and JDeveloper. The original version of the Kyoto Mutant Database was developed using IBM DB2 as a relational database. For GSD we used Oracle 8i and tested it on PostgreSQL. Because we used standard SQL, GSD will work with many other relational database management systems.

A genetic screen can yield a very large amount of data. In the Kyoto medaka screen we have recorded more than 10,000 images (3 GB) and we have screened more than 250,000 embryos in approximately 25,000 clutches of 1250 families. To centralize such all these data in a single database, instead of recording information on a variety of different media on different computers in different laboratories, is a great advantage and will facilitate analysis and publication of information from the screen. In addition, the database can be analyzed by scripts to generate daily reports and summaries of the screen (e.g. number of screened families and genomes, estimated mutation rate per genome). The use of standardized ontologies facilitates classification and comparisons of mutants because all screeners use the same set of standard terms.

The GSD software package provides researchers with the opportunity to perform a *distributed screen*. Using web interfaces enables research groups in different parts of the world to collaborate, thus facilitating saturating screens.

Data can be shared among the groups or limited to access by only particular groups or individuals.

Acknowledgements

We thank our collaborators of J. Wittbrodt's group Felix Loosli, Rebecca Quiring, Matthias Carl, Clemens Grabher, Filippo del Bene and Sylke Winkler. The authors gratefully acknowledge Hiroki Yoda, Yukihiro Hirose, Takao Sasado, Akihiro Yasuoka, Tomonori Deguchi, Akihiro Momoi, Masakazu Osakada, Chikako Morinaga, Katsutosi Niwa, Hiroshi Suwa and all the other screeners and technicians for their constant feedback to the database interfaces. This work was supported by NIH DC04186, HG002659 and HD22486 (M.W) and by a grant of the European commission (J.W.) and the HFSP (J.W., H.K.).

References

- Blake, J.A., Richardson, J.E., Bult, C.J., Kadin, J.A., Eppig, J.T., 2003. MGD: the Mouse Genome Database. *Nucleic Acids Res.* 31, 193–195.
- Driever, W., Solnica-Krezel, L., Schier, A.F., Neuhauss, S.C., Malicki, J., Stemple, D.L., et al., 1996. A genetic screen for mutations affecting embryogenesis in zebrafish. *Development* 123, 37–46.
- FlyBase-Consortium (2003), 2003. The FlyBase database of the *Drosophila* genome projects and community literature. *Nucleic Acids Res.* 31, 172–175.
- Galperin, M.Y., 2004. The Molecular Biology Database Collection: 2004 update. *Nucleic Acids Res.* 32, D3–22.
- Haffter, P., Granato, M., Brand, M., Mullins, M.C., Hammerschmidt, M., Kane, D.A., Odenthal, J., van Eeden, F.J., Jiang, Y.J., et al., 1996. The identification of genes with unique and essential functions in the development of the zebrafish *Danio rerio*. *Development* 123, 1–36.
- Harris, T.W., Lee, R., Schwarz, E., Bradnam, K., Lawson, D., Chen, W., et al., 2003. WormBase: a cross-species database for comparative genomics. *Nucleic Acids Res.* 31, 133–137.
- Henrich, T., Ramialison, M., Quiring, R., Wittbrodt, B., Furutani-Seiki, M., Wittbrodt, J., Kondoh, H., 2003. MEPD: a Medaka gene expression pattern database. *Nucleic Acids Res.* 31, 72–74.
- Hrabe de Angelis, M.H., Flaswinkel, H., Fuchs, H., Rathkolb, B., Soewarto, D., Marschall, S., et al., 2000. Genome-wide, large-scale production of mutant mice by ENU mutagenesis. *Nat. Genet.* 25, 444–447.
- Iwamatsu, T., 1994. Stages of Normal Development in the Medaka *Oryzias latipes*. *Zoolog. Sci.* 11, 825–839.
- Kimmel, C.B., Ballard, W.W., Kimmel, S.R., Ullmann, B., Schilling, T.F., 1995. Stages of embryonic development of the zebrafish. *Dev. Dyn.* 203, 253–310.
- Loosli, F., Koster, R.W., Carl, M., Kuhnlein, R., Henrich, T., Mucke, M., Krone, A., Wittbrodt, J., 2000. A genetic screen for mutations affecting embryonic development in medaka fish (*Oryzias latipes*). *Mech. Dev.* 97, 133–139.
- Sprague, J., Clements, D., Conlin, T., Edwards, P., Frazer, K., Schaper, K., Segerdell, E., Song, P., Sprunger, B., Westerfield, M., 2003. The Zebrafish Information Network (ZFIN): the zebrafish model organism database. *Nucleic Acids Res.* 31, 241–243.
- The-Gene-Ontology-Consortium (2001), 2001. Creating the gene ontology resource: design and implementation. *Genome Res.* 11, 1425–1433.