## A relational database for the FGSC

Kevin McCluskey - Fungal Genetics Stock Center, Department of Microbiology, University of Kansas Medical Center, Kansas City, KS 66160 USA

In our continuing effort to make more information about the materials at the FGSC available, we have developed a new database for the FGSC. This database has information about strains, lesions and a limited amount of information about cloned genes. There are also links to Genbank listings for many genes that are represented in the collection. The new database also manages the FGSC customer list and allows us to track orders.

---

The Fungal Genetics Stock Center has a large amount of information about the strains in the collection. This information includes genotypes, alleles, references for strains, references for genes, strain distribution history, and strain handling data. While not all of this information is relevant to customers, it was imperative that we make a reassessment of the data handling and presentation for the purposes of gaining increased access to the information. Morover, most culture collections are making their databases available on-line.

In the mid 1980's Craig Wilson undertook the herculean task of entering all the information on deposit sheets into an electronic database. Recent advances in database software, however, made the structure of the original FGSC database obsolete. The old database was developed in dBase and its tables were non-relational. This means that while individual tables (columns and rows) held useful information, it was not related to information in other tables. Examples of these tables include strain tables and customer tables (Table 1). With the help of the University of Kansas Medical Center Information Technology Office, we set out to convert to a relational database system using Microsoft Access97. The goals of this conversion were to create a table of lesions (genes) and to link each strain to its lesions, to merge the Neurospora and Aspergillus tables into one strain table, to create a customer table, and to create an order table that linked customers with strains or other materials.

### Format

Using a relational database system offers several advantages. The first is that you have an increased ability to extract information from the data. The second is that the database can be smaller using the relational format. Looking at the genot example, to have a list of markers for each strain, with associated information for each marker, would require that all this information be repeated for each strain. For example, the marker most common in the FGSC collection is the *inl* lesion that appears in 450 strains. *ad-3B* is the second most common marker, occurring in 416 strains. There are 22 markers that occur in over 100 strains and 55 markers occurring in 50 or more strains. For the *inl* marker, there are 34 citations. In a non-relational database, each of the 450 strains carrying *inl* would have to have all 34 references directly associated with them. This would require 15,300 entries in the database. In the relational system, each reference is entered once and assigned a reference number. The table "Lesion to references table" (Figure 1) has each reference number along with each lesion number. While computer speed has increased and storage media become more available, the time to search through these data is profoundly faster in the relational system. None of the information in the preceding examples would have been readily available in either of the previous incarnations of the FGSC database. With the new database, the linking of different tables made extracting this information trivial.

### Challenges

There were a number of problems to overcome in developing the new FGSC database. These are mostly historical. The most serious relates to an old decision to have a separate series of accession numbers for Aspergillus strains. This originally allowed there to be two strains with the FGSC number of, for example, FGSC #4. To simplify this, all Aspergillus strain numbers have long been preceeded with an "A". Hence FGSC # 4 is a *N. crassa* strain with the genotype *lys-5 ylo-1* while the strain FGSC A4 is the *A. nidulans* Glasgow wild-type strain. The most serious challenge presented by this was that we could not use the FGSC number as a "primary key" in the new database. To solve this, each strain now has a new number which is called "OrgStrainCounter" reflecting the fact that this number refers to strains of different organisms. This number is used to track strains through different parts of the database. Since the primary key does not indicate the organism type, each strain now has a strain type designation to indicate whether its strain number needs to be preceded with an "A". Additionally, each strain has genus and species designations.

The second biggest challenge was to generate a list of genes and associate each strain with the lesions it carries. While it would be possible to regenerate the genotype for each strain from such a list, we decided to keep the genotype listing in its current form and use it for generating output. Doing this removes the need to generate the genotype, with appropriate nomenclature, from primary fields. It does, however, require that changes to any strain's genotype be made in more than one location. The advantage of having each strain linked to genes in a listing of genes, however, outweighed any inconvenience it might pose. In orde achieve this, it was necessary to parse the genotype listing for each strain. This was accomplished by first marking each strain's record with a new field to indicate whether to parse it or not. Some strains have other information in their "locus" field (Table 2)

and need to be excepted from the parsing. Other strains have information in these fields that do not conform to established nomenclature. They all needed to be edited by hand. Once the strains that could be parsed were identified, we established rules by which to parse the genotypes. These were mostly minimalist, keeping in mind the rules of nomenclature. Examples of the rules we followed are: "gene names are separated by commas, spaces or semi-colons," or "hyphens are part of a gene's name". Following the automated parsing of the genotypes, additional hand editing was carried out, to "clean-up" the data.

### Genetic information

Having each strain associated with the lesions it carries allows us to easily generate a listing of supplements for each strain and to provide additional information based upon published information for each lesion. Of course not all lesions are as well documented as some. Much of the information for each lesion comes from the 1982 Neurospora compendium and from a listing of Aspergillus genes maintained by Dr. J. Clutterbuck that is available on-line at the FGSC site as well as at Dr. Clutterbuck's site at Glasgow.

Each lesion has been assigned a lesion number. The table "Strain Lesion" consists of two fields, "OrgStrainCounter" and "Lesion ID." In this table each strain with lesions has an entry for each lesion. For example, the strain FGSC 1118 has the "OrgStrainCounter" 1920 and this appears in the "Strain Lesion" table three times, once for each lesion *fz;sg;os-*. In Figure 1, this is the lower right corner. Each marker has an entry in the table "Organism Gene" which includes information such as the organism source, the genetic location, and any cultural requirements or growth characteristics. "Organism Gene" also cross references enzyme names between Aspergillus and Neurospora, where this information is known. There are a number of subordinate tables that provide supporting information for lesions. These include tables for enzyme data such as the EC number, the enzyme name in standard nomenclature, the address of DNA sequence online at Genbank, where available, and published references for many lesions.

### Wildtype strains

In addition to the information about lesions, the FGSC database now includes extensive information about wild-type strains. This information was assembled into a database by Barbara Turner at Stanford University and was shared with the FGSC. This section of the FGSC database is in the top right corner of Figure 1. These tables are called "Perkins WILD," "Perkins PUBS," "Perkins REFS," "Perkins STOCKS," and "Perkins SITECOLD". Facilitating the incorporation of this database into the FGSC database is the inclusion of a value, "FGSCLIST" in the "Perkins WILD" table. This variable is the number of the strain in the FGSC collection, where it exists. 869 of the wild-type strains at the FGSC were listed in this database. The Perkins database includes a treasure-trove of information on the origin of strains and on the ability of strains to mate with species testers.

### Plasmids and gene libraries

Another new aspect of the FGSC database is the incorporation of information about molecular materials available through the FGSC. The existing plasmid table was incorporated directly into the new FGSC database. The table "PLASMID" includes listings of all of the cloned genes and cloning vectors in the FGSC collection. While the information that is deposited with each plasmid varies, we have a minimum of information that we need to curate each plasmid and this information is both in the database and on-line. Gene libraries are listed for the purposes of tracking orders, but are not in a searchable format.

### Order tracking

Another motivation for rebuilding the database was to allow us to merge the order table with the mailing list. This was desirable for a number of reasons. The existing order table kept track of which strains were sent, but there was no easy way to determine whether orders had been paid or track which strains had been sent to which investigators. Moreover, since this table was not related to the strain tables ("Newros" and "Nidulans," Table 1), each strain's identity had to be entered for each order leaving open the possibility for errors or formatting differences. The order table has four subordinate tables, "Order Organism Strain," "Order Additional Item," Order Gene Library," and "Order Plasmid" that keep track of the items for each order.

We used the existing mailing list as the foundation for a new customer database. In December 1999 the mailing list had 793 entries while the entire customer database has 1,114 entries. The difference reflects individuals who have requested materials from the FGSC but who have not indicated an interest in being on the FGSC mailing list. Examples of listings in this category include purchasing agents, students, libraries, and people who have been removed from the mailing list for having not responded to mailings.

### Accessability

Most of the information available in the FGSC databases is accesible on-line at the FGSC web-site via a Sturctured Query Language (SQL) interface implemented in the HTML for the specific search pages. The strain information is available through the Neurospora or Aspergillus collection pages. Plasmid information is available on the "cloned genes and gene libraries" page. The FGSC web-site has strain and plasmid information in text file formats in addition to the searchable database. The mailing list is also on-line as a searchable database with text files from the FGN mailing list available as well.

Other databases, including the Neurospora bibliography and the Neurospora Knockout Registry, are on-line at the FGSC

web-site.

Acknowledgements:
Fritz Achen of the University of Kansas Medical Center Information Technology office facilitated the design of the new datab Chris Kunce of the FGSC worked on importing strain references and on the functionality of the database. Janet Barnet of the University of Kansas Medical Center web development team helped with the on-line searches.

Table 1. Tables from the original FGSC dBase Database

| Table | Newros | Nidulans | Mailist |
|---|---|---|---|
| **Data held** | Strain information for Neurospora | Strain information for Aspergillus | Mailing list |
| **Primary field** | FGSC Number | FGSC Number | Customer number |
| **Data fields** | Genotype | Genotype | Name |
| | Alleles | Alleles | Address |
| | Linkage groups | Linkage groups | Phone |
| | Mating type | Number from another collection | E-mail |
| | Number from another collection | Mutagen | FGN payment |
| | Depositor | Strain of origin | |
| | Genetic background | Depositor | |
| | References | References | |
| | Strain preservation status | Strain preservation status | |
| | Genotype testing status | Genotype testing status | |

Table 2. Examples of information stored in old fields

| Field | Locus | Allele | comment |
|---|---|---|---|
| **Value** | Genotype | Alleles | Growth requirements |
| | Species | Collection country or site | References |
| | Strain trivial name | Genotype | Anomalous testing results |
| | Synonyms | Status | Recipients |
| | Strain category | | Strain history |

Table 3. Examples of Tables in the new database

| Table | Organism Strain | Plasmid | Customer | Order |
|---|---|---|---|---|
| **Primary key** | OrgStrainCounter | Number | Customer Number | Order number |
| **Associated** | Organism type | Name | Name | Customer Number |
| | FGSC Number | Gene | Address | Date ordered |
| | Mating type | Reference | E-mail etc | Order mode |
| | Reporting Genes | Antibiotic | Mail-list status | Date shipped |
| | Alleles | Size | Newsletter payment | Payment date |
| | Linkage groups | Location of Stocks | Invoice address | Payment mode |
| | Stock number in | Notes | Organization code | Payment amount |
| | References | | Research interests | PO Number |
| | Dates | | URL | Comments |
| | Mutagen | | Comments | |
| | Notes | | | |

Figure 1. (following page) Relational structure of the FGSC database.

Tables comprising the FGSC database are connected to one another by different types of relationships. The 32 tables shown all relate to strains, plasmids and other items from the FGSC collection. The tables on the left pertain to customers and orders. The table in the center and those on the right pertain to strains.