25 P. C. Bailey *et al.*, *ibid.*, in press.
26. M. D. Gale and K. M. Devos, *Proc. Natl. Acad. Sci.* **95**, 1971 (1998).
27. K. M. Devos *et al.*, *Theor. Appl. Genet.* **85**, 673 (1993).
28. H. Zhang *et al.*, *ibid.* **96**, 69 (1998).
29. J. Peng *et al.*, *Genes Dev.* **11**, 3194 (1997).
30. N. P. Harberd, personal communication.
31. A. H. Paterson *et al.*, *Nature Genet.* **14**, 380 (1996).
32. J. L. Bennetzen, personal communication.
33. K. M. Devos, unpublished data.
34. R. A. Martienssen, L. Parnell, W. R. McCombie, personal communication.
35. A. Kleinhofs, personal communication.
36. S. W. Cartinhour, *Plant Mol. Biol.* **35**, 241 (1997).
37. M. D. Gale *et al.*, recommendations from the BBSRC-USDA Bilateral Plant Bioinformatics Planning and Coordination Meeting, Llangollen, UK, 22 to 24 March 1998.

Databases in Genomic Research

William M. Gelbart

VIEWPOINT

Genome-related databases have already become an invaluable part of the scientific landscape. The role played by these databases will only increase as the volume and complexity of relevant biology data rapidly expand. We are far enough into the genome project and into the development of these databases to assess their attributes and to reexamine some of the conceptual organizations and approaches they are taking. It is clear that there are needs for both highly detailed and simplified database views, the latter being especially needed to make expert domain data more accessible to nonspecialists.

Genomic databases are public windows on the high-throughput genome projects. In a sense, the success or failure of genome projects depends on the availability and utility to the scientific community of the data that are produced. Further, the very thrust of high-throughput science is the creation of large, well-organized, and rigorous sets of data. With this greatly increased biological data set that needs to be traversed, a variety of centralized databases are required to present these data in digestible chunks. Given the nature of biology and of database technology, it is probably impossible to determine in advance the database needs of the biological research community, but periodic retrospective analysis is certainly warranted. In this way, success stories can be identified, systematic problems can be assessed, and important gaps in the range of database coverage can be addressed. Having lived a dual existence as both a provider and a consumer of database information, I would like to offer my perspectives on where the genomic/genetic databases presently are and some of the issues that need to be addressed in the near term.

The Current Database Landscape

It is not my intention to exhaustively review the array of important genome-related databases that abound on the Internet. Rather, I would like to make some general classifications and comments. Genome-related databases can be broken into two major groups: generalized and specialized (or expert domain) databases. Generalized databases include the GenBank/EMBL/DDBJ archives of nucleic acids sequences and the PIR and SwissProt polypeptide sequence databases. Such databases capture and present information on particular classes of molecules, without any phylogenetic or functional exclusions. In contrast, the specialized databases do have more limited purviews, such as those organized around a specific model organism or around a type of biological function, such as protein family databases.

Interestingly, none of these generalized or specialized databases solely contain genome project data, but rather they are a mosaic of data from genome projects intermixed with those from the broader scientific community. This is in fact a recognition that the genome projects do not have exclusive license to produce any particular type of data—they are just much larger scale and frequently more accurate or self-consistent sources of particular types of information. In contrast, although the contributions of the community might lack as much data consistency and breadth of coverage, these possible deficiencies are offset by the greater expertise behind the individual contributions, which often are the culmination of years of focused research. The scientific community is best served by seamless integration of the high-throughput genome project data with the focused contributions of high-expertise groups.

Nothing makes a stronger case for such integration than a consideration of our current ability to decipher the information embedded in genomic DNA. The elucidation of the full genomic DNA sequence of humans, for example, has been referred to as the Rosetta Stone of human biology, which implies that it will allow us to elucidate all of the information encapsulated in this DNA sequence. However, it might be more appropriate to liken the human genomic sequence to the Phaestos Disk: an as yet undeciphered set of glyphs from a Minoan palace on the island of Crete. With regard to understanding how to make sense of the A's, T's, G's, and C's of genomic sequence, by and large we are functional illiterates.

Consider all of the structural information required to build a polypeptide chain and all of the regulatory information required to deploy that polypeptide in the correct sets of cells at the proper developmental times and in the requisite quantities. If every set of such information were analogous to one sentence in the instruction manual that we call the genome, a reasonable current assessment is that we have a partial but still quite incomplete knowledge of how to identify and read certain nouns (the structures of the nascent polypeptides and protein-coding exons of mRNAs). Our ability to identify the verbs and adjectives and other components of these genomic sentences (for example, the regulatory elements that drive expression patterns or structural elements within chromosomes) is vanishingly low. Further, we do not understand the grammar at all—how to read a sentence, how to weave the different sentences together to form sensible paragraphs describing how to build multicomponent proteins and other complexes, how to elaborate physiological or developmental pathways, and so on. Finally, we have little knowledge of how to identify and intepret structural information in the genome, such as boundary domains and other punctuation that separate different polypeptide-coding sentences from one another.

Were we to be able to read the genomic instruction manual in the same way we can read a book written in a language we understand, we might not need a huge support system of scientific databases. However, we are nowhere close to being at this point with regard to the genome. For now, the genomic sequence of an organism is written in a language we barely comprehend. However, through the work of the scientific community, we can attach biological meaning to limited regions of the sequence. Until we vastly improve our ability to actually read genomic DNA, we should work toward the goal of attaching all available experimental information as annotations to the framework, or reference, genomic DNA. This should be an important focus for model organism databases, in which substantial genetic information can serve as genomic annotation. Ordinarily, the task of framework sequence annotation should fall to one of the organism-specific expert domain databases. These groups have the specific

The author is in the Department of Molecular and Cellular Biology, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138, USA.

expertise to capture this information and to interpret the literature. Not only will these highly annotated framework sequences be of great immediate value to experimentalists, but in addition, they will be the datasets serving as test beds for the deciphering of the many different codes embedded in genomic DNA.

The creation of these highly annotated framework sequence sets ought to fall to expert groups, but they will be of the greatest value if systematic annotation is carried out in a consistent manner. The best way to accomplish this standardization is for these organism-specific databases to work in close coordination with the GenBank/EMBL/DDBJ collaborators and to employ the features syntax adopted by these nucleic acids databases.

The Gene: A Concept Past Its Time?

For biological research, the 20th century has arguably been the century of the gene. The central importance of the gene as a unit of inheritance and function has been crucial to our present understanding of many biological phenomena. Nonetheless, we may well have come to the point where the use of the term "gene" is of limited value and might in fact be a hindrance to our understanding of the genome. Although this may sound heretical, especially coming from a card-carrying geneticist, it reflects the fact that, unlike chromosomes, genes are not physical objects but are merely concepts that have acquired a great deal of historic baggage over the past decades.

Ultimately, we want to understand the relationships between heritable units, their gene products, and their phenotypes. The classical gene was thought to be the relevant heritable unit for establishing such relationships. However, the realities of genome organization are much more complex than can be accomodated in the classical gene concept. Genes reside within one another, share some of their DNA sequences, are transcribed and spliced in complex patterns, and can overlap in function with other genes of the same sequence families. Consider so-called alternative splicing, in which one or more exons are shared among multiple transcripts. There is a continuum ranging from cases in which two transcripts are almost identical along their entire length to examples in which only a small portion of the two mRNAs is shared. Sometimes these products have very similar biological activities, whereas in other cases their activities are disparate. What are the rules for deciding whether two partially overlapping mRNAs should be declared to be alternative transcripts of the same gene or products of different genes? We have none.

Independent of this question is the question of how to relate a mutant phenotype to alterations in multiple overlapping gene products. Suppose that we have a missense mutation that falls within one or more exons that contribute to more than one mRNA and thereby to more than one polypeptide chain. How do we assess the contributions of defects in the different polypeptides to the ultimate phenotype elicited by this mutation?

For reasons such as these, I believe that we are entering a period in which we must shift to the view that the genome largely encodes a series of functional RNAs and polypeptides that are expressed in characteristic spatial, temporal, and quantitative patterns. The classical concept of the gene ultimately forms a barrier to trying to understand phenotypes in terms of encoded functional products.

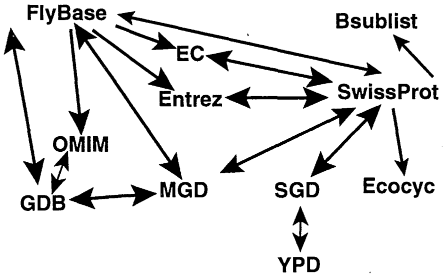


Fig. 1. A schematic example of migration among generalized and expert domain databases, showing the difficulties of establishing and maintaining pairwise links with all other relevant databases. Starting with a FlyBase gene expected to have homologs in many different species (*Pgk*, phosphoglycerate kinase), cross links were used to migrate among some other genetic/genomic databases. Double-headed arrows indicate databases with reciprocal cross links; single-headed arrows indicate unidirectional cross links. This is not intended to be an exhaustive example of migration.

This is not a purely abstract discussion but may well demand that we reexamine how we are organizing data within genome-related databases. In most or all of these databases, much biological data is attached to these suspect units called genes. Although some aspects of these phenotypes might be associated with different subsets of alternative products of these genes, the databases might not support the most rigorous parsing of this phenotypic information.

Increasing Access to Genomic Databases: Breaking Down Activation-Energy Barriers

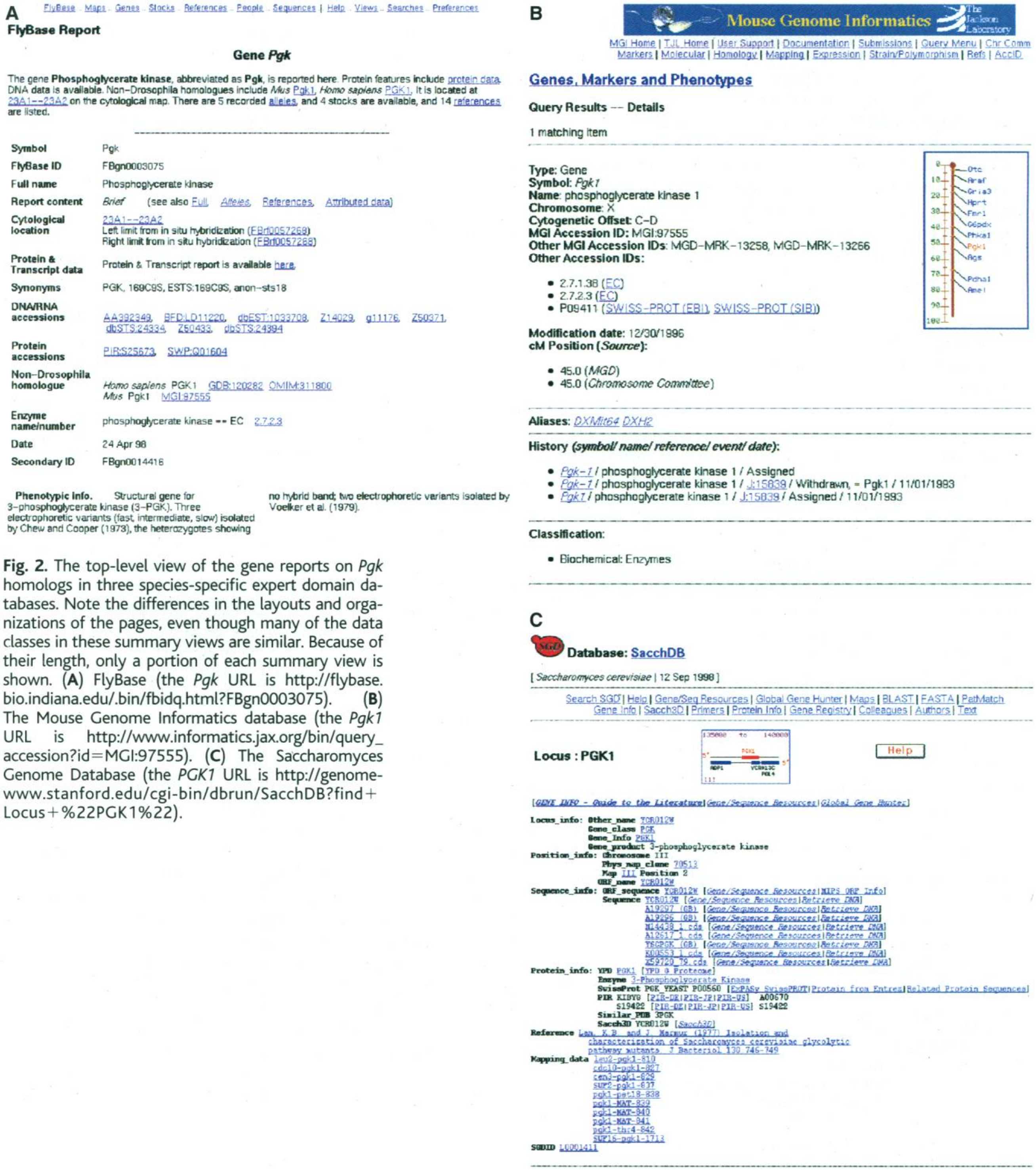
Expert domain databases, most notably the model organism databases, have two major constituencies: the more focused scientific community actively studying that system and the larger scientific community interested in relating this specialized information to data from other systems. At this point, it is probably true that these expert domain databases serve their focused communities better than they serve the broader scientific community. There are several reasons why this is so. Typically, these databases have emerged at the behest of the specialized communities and fill a legitimate need and desire for distribution of vital technical information and reagents. Within each specialized community, there is a shared language with its own jargon and grammar (nomenclature) for describing its research findings. For example, the organism-specific communities have each evolved their own means of describing genetic objects, anatomy, and other aspects of phenotype. The organism-specific community is much more focused in its interests, and it is relatively straightforward for databases to assess the needs of such a focused community. In contrast, the broader scientific community contains numerous orientations and perspectives and has many different reasons for its interests in making connections to data on the model organisms. It is a diffuse and difficult target.

Thus, a major challenge for the organism-specific databases over the next few years is to find a successful formula for meeting the needs of the broader scientific community while not deserting its focused, specialized user groups. Consider how these databases make their information available to the public. The model organism databases are principally accessible through networked servers accessed through World Wide Web browsers. As with any technology, Web access has positives and negatives. On the positive side, the model organism databases are working very hard to incorporate robust and reciprocal links so that users can migrate and meander from one database to another, without prior knowledge of the relationships represented by the links. The richer the links, the more extensive the information that the user will be able to harvest.

There are some problems with this approach as well. For one thing, migrating around the Web is like flipping from one entry in an encyclopedia to another. Migration is incremental, one flip of the pages at a time. The Web thus far has not lent itself to effective querying across many databases, with a compiled set of answers being delivered in response. Further, as the user traverses from an entry in one database to a linked entry in another, the user may need to become educated in the structure of the linked database and in the jargon and grammar of that disparate system (Figs. 1 and 2). Indeed, even a cursory examination of the different model organism databases reveals a daunting diversity of report formats, data organizations, and distinct scientific tongues. These all represent substantial activation-energy barriers to the effective use of these databases by the nonspecialist community.

How can the needs of both the specialized and the broader scientific communities be addressed within the constraints of the Web? First, there needs to be a recognition that the same data views will not necessarily serve the interests of both communities. Many of the classes of data that are of interest to specialists are largely irrelevant to the broader community. For example, chromosomal map data and information on mutant strains are typically only of interest to specialists. On the other hand, data on gene products, gene expression patterns, phenotypes, and pathways are of broad general interest. The expert domain databases should continue to support their focused communities through the maintenance of a specialist database. Each community is used to its particular format of presentation and presumably has identified the most important data classes for its needs. Thus, the existing model organism Web sites can continue in their present form.

In addition, the expert domain databases (especially the model organism databases) should work together to develop a nonspecialist Web interface, in which data classes common to many or all systems are presented in a standardized and readily digestible format, with an active effort being made to limit jargon and to identify data items of general interest. Put another way, the expert domain databases need to establish a minimum activation energy representation for the broader community. Exactly how such an interface would look and which data classes are ripe for incorporation are matters for exploration and experimentation.


A
FlyBase . Maps . Genes . Stocks . References . People . Sequences | Help . Views . Searches . Preferences
B
Mouse Genome Informatics
The Jackson Laboratory
MGI Home | TJL Home | User Support | Documentation | Submissions | Query Menu | Chr Comm
Markers | Molecular | Homology | Mapping | Expression | Strain/Polymorphism | Refs | AccID
0
10
20
30
40
50
60
70
80
90
100
Otc
Araf
Gria3
Hprt
Fmr1
G6pdx
Phka1
Pgk1
Ags
Pdha1
Amel
C
SGD
Search SGD | Help | Gene/Seq Resources | Global Gene Hunter | Maps | BLAST | FASTA | PatMatch
Gene Info | Sacch3D | Primers | Protein Info | Gene Registry | Colleagues | Authors | Text
135000 to 140000
5'
PGK1
ADP1
YCR013C
POL4
III
Help

FlyBase Report

Gene *Pgk*

The gene **Phosphoglycerate kinase**, abbreviated as **Pgk**, is reported here. Protein features include protein data. DNA data is available. Non-Drosophila homologues include *Mus* Pgk1, *Homo sapiens* PGK1, It is located at 23A1--23A2 on the cytological map. There are 5 recorded alleles, and 4 stocks are available, and 14 references are listed.

Symbol	Pgk
FlyBase ID	FBgn0003075
Full name	Phosphoglycerate kinase
Report content	*Brief* (see also Full, *Alleles*, References, Attributed data)
Cytological location	23A1--23A2 Left limit from in situ hybridization (FBrf0057268) Right limit from in situ hybridization (FBrf0057268)
Protein & Transcript data	Protein & Transcript report is available here.
Synonyms	PGK, 169C9S, ESTS:169C9S, anon--sts18
DNA/RNA accessions	AA392349, BFD:LD11220, dbEST:1033708, Z14029, g11176, Z50371, dbSTS:24334, Z50433, dbSTS:24394
Protein accessions	PIR:S25673, SWP:Q01604
Non-Drosophila homologue	*Homo sapiens* PGK1 GDB:120282 OMIM:311800 *Mus* Pgk1 MGI:97555
Enzyme name/number	phosphoglycerate kinase -- EC 2.7.2.3
Date	24 Apr 98
Secondary ID	FBgn0014416

Phenotypic info. Structural gene for 3-phosphoglycerate kinase (3-PGK). Three electrophoretic variants (fast, intermediate, slow) isolated by Chew and Cooper (1973), the heterozygotes showing no hybrid band; two electrophoretic variants isolated by Voelker et al. (1979).

Genes, Markers and Phenotypes

Query Results --- Details

1 matching item

Type: Gene
Symbol: *Pgk1*
Name: phosphoglycerate kinase 1
Chromosome: X
Cytogenetic Offset: C-D
MGI Accession ID: MGI:97555
Other MGI Accession IDs: MGD-MRK-13258, MGD-MRK-13266
Other Accession IDs:

- 2.7.1.38 (EC)
- 2.7.2.3 (EC)
- P09411 (SWISS-PROT (EBI), SWISS-PROT (SIB))

Modification date: 12/30/1996
cM Position (*Source*):

- 45.0 (*MGD*)
- 45.0 (*Chromosome Committee*)

Aliases: *DXMit64* *DXH2*

History *(symbol/ name/ reference/ event/ date)*:

- *Pgk-1* / phosphoglycerate kinase 1 / Assigned
- *Pgk-1* / phosphoglycerate kinase 1 / J:15839 / Withdrawn, = Pgk1 / 11/01/1993
- *Pgk1* / phosphoglycerate kinase 1 / J:15839 / Assigned / 11/01/1993

Classification:

- Biochemical: Enzymes

Database: SacchDB

[*Saccharomyces cerevisiae* | 12 Sep 1998]

Locus : PGK1

```
[GENE INFO - Guide to the Literature|Gene/Sequence Resources|Global Gene Hunter]

Locus_info: Other_name YCR012W
            Gene_class PGK
            Gene_Info PGK1
            Gene_product 3-phosphoglycerate kinase
Position_info: Chromosome III
               Phys_map_clone 70513
               Map III Position 2
              ORF_name YCR012W
Sequence_info: ORF_sequence YCR012W [Gene/Sequence Resources|MIPS ORF Info]
               Sequence YCR012W [Gene/Sequence Resources|Retrieve DNA]
                        A19297 (GB) [Gene/Sequence Resources|Retrieve DNA]
                        A19295 (GB) [Gene/Sequence Resources|Retrieve DNA]
                        M14438_1.cds [Gene/Sequence Resources|Retrieve DNA]
                        A12617_1.cds [Gene/Sequence Resources|Retrieve DNA]
                        YSCPGK (GB) [Gene/Sequence Resources|Retrieve DNA]
                        K00553_1.cds [Gene/Sequence Resources|Retrieve DNA]
                        X59720_79.cds [Gene/Sequence Resources|Retrieve DNA]
Protein_info: YPD PGK1 [YPD @ Proteome]
              Enzyme 3-Phosphoglycerate Kinase
              SwissProt PGK_YEAST P00560 [ExPASy SwissPROT|Protein from Entrez|Related Protein Sequences]
              PIR KIBYG [PIR-DE|PIR-JP|PIR-US]  A00670
                  S19422 [PIR-DE|PIR-JP|PIR-US] S19422
              Similar_PDB 3PGK
              Sacch3D YCR012W [Sacch3D]
Reference Lam, K.B. and J. Marmur (1977) Isolation and
          characterization of Saccharomyces cerevisiae glycolytic
          pathway mutants. J Bacteriol 130:746-749
Mapping_data leu2-pgk1-810
             cdc10-pgk1-827
             cen3-pgk1-829
             SUF2-pgk1-837
             pgk1-pet18-838
             pgk1-MAT-839
             pgk1-MAT-840
             pgk1-MAT-841
             pgk1-thr4-842
             SUF16-pgk1-1713
SGDID L0001411
```

Fig. 2. The top-level view of the gene reports on *Pgk* homologs in three species-specific expert domain databases. Note the differences in the layouts and organizations of the pages, even though many of the data classes in these summary views are similar. Because of their length, only a portion of each summary view is shown. (**A**) FlyBase (the *Pgk* URL is http://flybase.bio.indiana.edu/.bin/fbidq.html?FBgn0003075). (**B**) The Mouse Genome Informatics database (the *Pgk1* URL is http://www.informatics.jax.org/bin/query_accession?id=MGI:97555). (**C**) The Saccharomyces Genome Database (the *PGK1* URL is http://genome-www.stanford.edu/cgi-bin/dbrun/SacchDB?find+Locus+%22PGK1%22).