

## “Hot Topics” BIOBASE: Proteome and HGMD

Fran Lewitter  
2/25/2009



1

## Our License

- **BIOBASE Knowledge Library (BKL)**
  - HumanPSD (Human Proteome Survey Database; human, rat, mouse)
  - GPCR-PD (G protein-coupled receptors; human, rat, mouse)
  - YPD suite and WormPD (PD stands for Proteome Database)
  - TRANSFAC (transcription factors)
  - TRANSPATH (signal transduction pathways)
  - ExPLAIN (Gene Expression Analysis System)
- **HGMD**
  - Human Gene Mutation Database

2

## ExPlain™ Analysis System

- Facilitates analysis and biological interpretation of large scale data sets generated by DNA microarrays as well as proteomics, ChIP-on-chip, Chip-Seq and other high-throughput experiments.
- Flexible workflows systematically guide the user through the process of identifying the transcription factors and key signal transduction molecules that are unique for a data set

3

## Local Access info

4

## Local Access info

**BIOBASE**  
BIOLOGICAL DATABASES

Home • Contact Form • Sitemap • Imprint • Privacy Policy • Free Trials • Logout

Click the link(s) below to access the product(s) available with your subscription.

Access to the product(s) below is automatically provided via your institution's subscription package. For new users, click [here](#) to create a username to be used for saving personal data. A username is not required for general browsing. If you have a personal subscription outside of the institutional subscription, click [here](#) to login to your personal subscription.

**BIOBASE Knowledge Library**

HumanPSD + GPCR-PD + TRANSFAC Suite + TRANSPATH + YPD Suite + WormPD

**Analysis Tools**

- ExPlain 3.0
- ExPlain 2.4.2
- ExPlain 2.4.1
- ExPlain 2.4
- ExPlain 2.3
- ExPlain 2.0

**3rd Party Databases**

HGMD

Please note that on-line access to ExPlain, 3rd party databases, and downloadable flat files requires a second login with the same username and password.

For help with access or using your products, please contact our [customer support team](#). You are accessing this site from the IP address: 18.4.1.145. Please include this information in any support correspondence.

**Need help getting started?**

Download the following user guides:

- BKL Quick Reference Guide
- BKL Retriever Quick Reference Guide
- BKL Retriever Search Examples
- BKL Plant Quick Reference Guide
- BKL Plant Retriever Quick Reference Guide
- ExPlain Quick Start Guide
- ExPlain for TRANSFAC Only Subscribers
- ExPlain Training Exercises
- ExPlain Training Exercises Excel Table 1
- ExPlain Training Exercises Excel Table 2
- ExPlain Training Exercises Reference

5

## Register for Login

**BIOBASE**  
BIOLOGICAL DATABASES

Please complete this form to initiate your account.

You are subscribing to the account of: **Massachusetts Institute of Technology (MIT) and Whitehead Institute**;  
If you wish to create user for a different subscription, please enter the License Key below.

License Key (Optional)

First Name\*

Middle Initial

Last Name\*

Title

Institution Name\*

Institution Type\*

Role\*

Contact Phone\*

E-mail\*

Please enter a user name and password.

Username\*

Password\*

Password Repeat\*

Please keep me updated about BIOBASE news by adding me to your mailing list

Fields marked with "\*" are mandatory and can not be left blank.

6

## PROTEOME

- Complete Mammalian coverage**
  - Protein characteristics on the full proteomes of human, mouse and rat, including:
    - Biological processes, molecular function, localization, expression
    - Protein modifications and interactions
    - Gene expression regulation
- Yeast**
  - Quickly access experimental details from the literature on *S. cerevisiae*, *S. pombe*, *C. albicans* and 17 other human fungal pathogens
  - Mutant phenotypes and genetic interactions
  - Protein interaction networks
  - Environmental, chemical, and molecular gene regulators
- Worm**
  - Integrates gene and protein interaction data, mutant phenotype, expression, function and protein family data
- Disease**
  - Biomarker associations
  - Drug-target interactions
  - Mouse disease models

7

## Statistics for the Winter 2009 release

HUMAN PSD	Human	Mouse	Rat
Proteins	18,985	21,125	11,410
Diseases/Disease Models	3,018	1,299	-
Disease-Gene Assignments	36,671	5,769	-
GO Assignments	181,359	183,359	99,773
Expression Assignments	149,251	153,560	93,738
Phenotype Assignments	-	66,032	-
References	151,858	77,669	48,086

Other Organisms	YPD	PombePD	MycoPathPD	WormPD
Proteins	6,654	5,001	19,932	20,749
References	30,333	4,237	4,427	5,325
GO Assignments	65,771	25,178	93,290	76,035
Phenotype Assignments	38,632	5,866	2,722	18,845
Annotations	624,654	84,619	70,721	260,568

8

## Ways to search BKL

- Quick Search
  - Locus report
- Ontology Search (Retriever)
  - Upload data
- Advanced Search

9

**BIOBASE**  
BIOLOGICAL DATABASES

BIOBASE Knowledge Library

View statistics for: TRANSFAC® 2009.4 TRANSPATH® 2009.4 Proteome: GPCR-PD™ YPD™ WormPD™ HumanPD™

**QUICK SEARCH**

Quick Search  
Ontology Search  
Advanced Search

**Genome**  
Gene  
Promoter  
Functional Region  
Composite Element  
Site  
ChIP-chip  
– Browse by TF  
Matrix

**Transcriptome**  
RNA

**Proteome**  
Protein  
Family  
TF Classification

**Reactome**  
Map  
Pathway  
Reaction  
Small Molecule  
Drug

**Phenome**  
Disease Biomarker

**Tools**  
Draw Pathways  
MATCH  
ExpPlain  
BLAST

Help  
User data  
Search Results  
Login

Search by  Name  Identifier  Keyword

Use wildcards  
Exact word search  
Fuzzy

Search for disease biomarkers

Gene Set Analysis

Upload your gene set for analysis of over-representation by disease, expression, canonical pathways, etc & create a customized report.

9

## Ontology/Retriever Search

Browse Upload Results Help

View parents / children

Term lookup Results

119933 Characterization  
> 107119 Species and genetic location  
> 100019 Calculated property  
> 79834 Gene ontology (GO)  
> 76516 Protein domains  
> 74777 Regulatory factors and targets  
> 43944 Orthologs related to current set  
> 29618 Expression  
> 26970 Family classification  
> 17989 Interaction with other proteins

52040 Predicted  
138600 Characterized  
129293 Uncharacterized

BP GO biological process  
CC GO cellular component  
CH Characterization  
DI Disease (human,mouse)  
DR Pharmaceuticals  
DO Protein domains  
EX Expression  
GF Family classification  
IN Interactions  
MD Protein modifications  
MF GO molecular function  
OL Orthologs related to current set  
PH Mutant phenotype  
PT Pathways  
RE Regulators of fungal genes (yeast)  
RG Calculated properties  
SP Species and genetic location  
TF Regulatory factors and targets

BIOBASE Knowledge Library

Focus on term  
View all 119933 proteins

11

Browse Upload Results Help

BIOBASE  
BIOLOGICAL DATABASES

Reset

BIOBASE Knowledge Library

Enter a list of genes or proteins to be analyzed

Please specify the type of data:  
Gene or Protein names

Please enter one ID or name per line.

Text files can be loaded here:  
Choose File no file selected

For gene or protein names, please specify species:  
Human

Enter saved text queries:  
Execute

Match synonyms only if necessary

Submit Reset

12

[Browse](#) **Upload** [Results](#) [Help](#)

[Reset](#)

---

BIOBASE Knowledge Library

**Enter a list of genes or proteins to be analyzed**

Please specify the type of data:

- Gene or Protein names
- BioBase accession
- Affymetrix
- AgilentProbe
- CandidaDB
- EMBL/GenBank/DBJ
- Ensembl
- EntrezGene
- FlyBase
- HGMD
- HGNC
- IPI
- InterPro
- MAIZECDB
- MGI
- MIRBASE
- OMIM
- PDB
- PFAM
- PHYTOZOME
- PIR
- PROSITE
- RefSeq
- SGD
- TAIR
- UniGene
- UniProt
- Pubmed ID

Please enter one ID or name per line.

Text files can be loaded here:

[Choose File](#) no file selected

Enter saved text queries:

[Execute](#)

13

BIOBASE Knowledge Library

View statistics for: TRANSFAC® 2009.4 TRANSPATH™ 2009.4 Proteome: GPCR-PD™ YPD™ WormPD™ HumanPSD™

**QUICK SEARCH**

**Quick Search**

**Ontology Search**

**Advanced Search**

- Genome
  - Gene
  - Promoter
  - Functional Region
  - Composite Element
  - Site
  - ChIP-chip
  - Browse by TF
  - Matrix
- Transcriptome
  - RNA
- Proteome
  - Protein
  - Family
  - TF Classification
- Reactome
  - Map
  - Pathway
  - Reaction
  - Small Molecule
  - Drug
- Phenome
  - Disease Biomarker
- Tools
  - Draw Pathways
  - MATCH
  - ExPlain
  - BLAST

[Help](#)

**User data**

[Search Results](#)

[Login](#)

**Search by**  Name  Identifier  Keyword

Use wildcards

Exact word search

Fuzzy

[Find](#)

**Search for disease biomarkers**

[Find](#)

**Gene Set Analysis** Upload your gene set for analysis of over-representation by disease, expression, canonical pathways, etc & create a customized report.

14

**PROTEIN SEARCH**

Previous search: **None** [show results](#) [Delete](#)

Constraints: **None**

Search across data types with previously saved searches

Enter the search term: **P1k1**

Search Term: **P1k1**

Search Fields: **Name**

Search exact term or use wildcards

Customize search by selecting fields, or blocks, to query

Customize output fields, or blocks, to display in the Results

Output Fields: **Name, Species/Taxon, Type**

Secondary accession: Synonyms, Binding region (ChIP-chip), Binding sites/Regulated genes, Comments/Annotations, Encoding gene, Complexes, Factor class, Gene ontology (GO), Isoelectric point

Add more search conditions, if desired

Submit Query [Reset Query](#)

Submitting a query without search term retrieves the first 1,000 entries for the data type

Field (attribute) description: Search field: Name, Name and synonyms of the protein

View description of a selected search field

15

BIOBASE Knowledge Library

Entity: **Protein** Search criteria: **(Name = P1k1)** Total records: **18**

[Save result](#) [BKL Pathfinder](#) [Export](#) Search for: **ChIP-chip** [FASTA](#)

Mark all on page Hits on page: **50**

18 hits	Primary accession	Name	Species/Taxon	Type
<input type="checkbox"/>	PR000006665	pik1	Human	isogroup
<input type="checkbox"/>	PR000025198	pik1	Human	protein
<input type="checkbox"/>	PR000010860	pik1	Mouse	isogroup
<input type="checkbox"/>	PR0000273471	pik1	Mouse	protein
<input type="checkbox"/>	PR0000443399	pik1	Rat	isogroup
<input type="checkbox"/>	PR0000443400	pik1	Rat	protein
<input type="checkbox"/>	PR000022540	pik1	Mammalia	isogroup
<input type="checkbox"/>	PR000031741	pik1	Vertebrata	protein
<input type="checkbox"/>	PR000019302	pik1	clawed frog, Xenopus laevis	isogroup
<input type="checkbox"/>	PR0000266266	pik1	clawed frog, Xenopus laevis	protein
<input type="checkbox"/>	PR0000226384	pik1		orthogroup
<input type="checkbox"/>	PR000079861	PLK1S1	Human	isogroup
<input type="checkbox"/>	PR0000178707	PLK1S1	Human	protein
<input type="checkbox"/>	PR0000145647	Pik1s1	Mouse	isogroup
<input type="checkbox"/>	PR0000224354	pik1s1	Mouse	protein
<input type="checkbox"/>	PR0000233661	Pik1s1		orthogroup
<input type="checkbox"/>	PR0000239811	PLK1S1		orthogroup
<input type="checkbox"/>	PR0000304395	ERCC6L	Human	isogroup

Mark all on page

**Quick Search**

**Ontology Search**

**Advanced Search**

- Genome
  - Gene
  - Promoter
  - Functional Region
  - Composite Element
  - Site
  - ChIP-chip
  - Browse by TF
  - Matrix
- Transcriptome
  - RNA
- Proteome
  - Protein
  - Family
  - TF Classification
- Reactome
  - Map
  - Pathway
  - Reaction
  - Small Molecule
  - Drug
- Phenome
  - Disease Biomarker
- Tools
  - Draw Pathways
  - MATCH
  - ExPlain
  - BLAST

[Help](#)

**User data**

[Search Results](#)

User: [Logout](#)

16

## HGMD Professional Winter 2009

- Contains **96,631** mutations and disease associated/functional polymorphisms in **3,526** genes taken from **28,466** journal articles and provides **3,456** reference cDNA sequences. This represents a growth of **2,214** mutations over the previous release.
- Micro-lesions – missense, splicing, etc.
- Gross lesions – repeats, indels, rearrangements

17

## Access to HGMD data

- Gene
  - Browse or Search
- Mutation
  - Search
- Reference
  - Search
- Advanced
  - Browse by chromosome or complex search

18

## HGMD – Winter 2009

Type of Mutation	Number
Missense/nonsense	54,422
Splicing	9,267
Regulatory	1,700
Small deletions	15,231
Small insertions	6,273
Small indels	1,413
Gross deletions	5,912
Gross insertions	1,210
Complex rearrangements	912
Repeat variations	29
	96,631

19

**BIOBASE**
HGMD® professional release 2009.4 (2009-12-11)

To start a search choose the search option in the menu to the left.

This release comprises the following tables:

Table:	Description:	Entries:
Missense/nonsense	Single base-pair substitutions in coding regions are presented in terms of a triplet change with an additional flanking base included if the mutated base lies in either the first or third position in the triplet.	54422
Splicing	Mutations with consequences for mRNA splicing are presented in brief with information specifying the relative position of the lesion with respect to a numbered intron donor or acceptor splice site. Positions given as positive integers refer to a 3' (downstream) location, negative integers refer to a 5' (upstream) location.	9267
Regulatory	Substitutions causing regulatory abnormalities are logged in with thirty nucleotides flanking the site of the mutation on both sides. The location of the mutation relative to the transcriptional initiation site, initiation codon, polyadenylation site or termination codon is given.	1700
Small deletions	Micro-deletions (20 bp or less) are presented in terms of the deleted bases in lower case plus, in upper case, 10 bp DNA sequence flanking both sides of the lesion. The numbered codon is preceded in the given sequence by the caret character (^).	15231
Small insertions	Micro-insertions (20 bp or less) are presented in terms of the inserted bases in lower case plus, in upper case, 10 bp DNA sequence flanking both sides of the lesion. The numbered codon is preceded in the given sequence by the caret character (^).	6273
Small indels	Micro-indels (20 bp or less) are presented in terms of the deleted/inserted bases in lower case plus, in upper case, 10 bp DNA sequence flanking both sides of the lesion. The numbered codon is preceded in the given sequence by the caret character (^).	1413
Gross deletions	Information regarding the nature and location of each lesion is logged in narrative form because of the extremely variable quality of the original data reported.	5912
Gross insertions	Information regarding the nature and location of each lesion is logged in narrative form because of the extremely variable quality of the original data reported.	1210
Complex rearrangements	Information regarding the nature and location of each lesion is logged in narrative form because of the extremely variable quality of the original data reported.	912
Repeat variations	Information regarding the nature and location of each lesion is logged in narrative form because of the extremely variable quality of the original data reported.	291
<b>Mutation total (HGMD release 2009.4)</b>		<b>96631</b>

HGMD  
Start  
  
 Gene  
Mutation  
Reference  
Advanced  
  
 Information  
Contact us

20

