DATABASES

Mutational Data Integration in Gene-Oriented Files of the Hermansky-Pudlak Syndrome Database

Wei Li,[1]* Min He,[1] Helin Zhou,[1] Jonathan W. Bourne,[2] and Ping Liang[3]

[1]*Key Laboratory of Molecular and Developmental Biology, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, China;* [2]*Department of Physiology, Biophysics, and Systems Biology, Weill Medical College of Cornell University, New York, New York;* [3]*Department of Cancer Genetics, Roswell Park Cancer Institute, Buffalo, New York*

Communicated by A. Jamie Cuticchia

Hermansky-Pudlak Syndrome (HPS) is a genetically heterogeneous disorder characterized by oculocutaneous albinism and prolonged bleeding due to abnormal vesicle trafficking to lysosomes and related organelles such as melanosomes and platelet dense granules. This HPS database (HPSD; http://liweilab.genetics.ac.cn/HPSD/) provides integrated, annotatory, and curative data that is distributed in a variety of public databases or predicted by bioinformatics servers for the recently cloned human and mouse HPS genes, as well as for the genes responsible for HPS-related syndromes, such as Chediak-Higashi Syndrome (CHS), Griscelli syndrome (GS), oculocutaneous albinism (OCA), Usher syndrome type 1B (USH1B), and ocular albinism (OA). The HPSD is designed by using a unique Gene-Oriented File (GOF) format. Seven blocks (genomic, transcript, protein, function, mutation, phenotype, and reference) are carefully annotated in each user-friendly GOF entry. The HPSD emphasizes paired human and mouse GOF entries. The genes included in this database (currently 58 in total) are arbitrarily divided into four categories: 1) Human and Mouse HPS, 2) Mouse HPS Only, 3) Putative Mouse or Human HPS, and 4) HPS Related Syndromes. All the mutations in these genes are integrated in the GOFs. We expect that these very informative and peer-reviewed GOFs will be shortcuts to utilize the web-based information for the emerging interdisciplinary studies of HPS. Hum Mutat 0:1–6, 2006. Published 2006 Wiley-Liss, Inc.[†]

KEY WORDS: Hermansky-Pudlak syndrome; HPS; albinism; HPSD; database

INTRODUCTION

Hermansky-Pudlak Syndrome (HPS; MIM# 203300) [Hermansky and Pudlak, 1959] is an autosomal recessive and a genetically heterogeneous disorder characterized by a triad: oculocutaneous albinism, bleeding tendency, and ceroid deposition, which may cause lung fibrosis, colitis, and cardiomyopathy. Patients with HPS often die during the third to fifth decade [Huizing and Gahl, 2002]. The key pathological aspect of both human and mouse HPS is the disrupted biogenesis and/or function of specialized lysosomes (which are now termed as lysosome-related organelles (LROs), such as melanosomes and platelet dense granules) as well as conventional lysosomes [Dell'Angelica et al., 2000; Swank et al., 1998]. To better understand the pathological mechanism of this deleterious disease, researchers initiated projects to identify the genes responsible for HPS by using positional cloning approaches in the early 1990s. These efforts were rewarded by the successful identification of the first HPS gene, *HPS1*, in 1996 [Oh et al., 1996] and the first murine HPS gene, *Hps1/ep*, in 1997 [Gardner et al., 1997]. The cloning of other HPS genes has been accelerated since the completion of the human and mouse genome projects.

Since 1996, seven human HPS genes (HPS1–HPS7) and 14 mouse HPS genes have been identified, as well as a gene (*SLc7a11*) causing a mild form of mouse HPS and a controversial HPS gene, Rab27a [Chintala et al., 2005; Li et al., 2004]. Several complexes formed by these HPS gene products have been demonstrated to be involved in vesicle trafficking pathways. These are the AP-3 complex, the Class C Vps (HOPS) complex, and the biogenesis of lysosome-related organelles complexes-1, -2, and -3 [Di Pietro and Dell'Angelica, 2005; Li et al., 2004]. Better understanding of the biochemical and functional properties of these complexes requires cutting-edge technology and calls for interdisciplinary studies of HPS gene functions.

Retrieval and extraction of the information about HPS genes in a variety of databases are useful for defining their functional aspects at the systems biology level. Tremendous amounts of information have been generated in the post-genomic era. To utilize the existing information, users have to search a variety of

Received 20 October 2005; accepted revised manuscript 5 January 2006

This work was done while Jonathan W. Bourne was a member of the Biophysics Program, SUNY Geneseo, Geneseo, New York.

*Correspondence to: Wei Li, Key Laboratory of Molecular & Developmental Biology, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, P. O. Box 2707, Beijing 100080, China. E-mail: wli@genetics.ac.cn

Grant sponsor: "Bai Ren Ji Hua," Chinese Academy of Sciences; Grant sponsor: National Basic Research Program (973 Program), Ministry of Science and Technology of China; Grant number: 2006CB500700; Grant sponsor: Distinguished Young Scholars Fund, National Natural Science Foundation of China; Grant number: 30525007; Grant sponsor: National Institutes of Health; Grant number: CA101515; Grant sponsor: National Science Foundation; Grant number: 0098082.

DOI 10.1002/humu.20309

Published online in Wiley InterScience (www.interscience.wiley.com).

[†]This article is a US Government work and, as such, is in the public domain in the United States of America.

databases or to predict the results by using different servers or programs. This will be a time-consuming procedure even when focused on only one gene of interest. Although there are several platforms or interfaces such as GDB (www.gdb.org/gdb), UCSC Gene Sorter (http://genome.ucsc.edu/cgi-bin/hgNear), and MGI (www.informatics.jax.org) to generate links to many databases or to do the predictions or annotations automatically, these servers lack curation. One would easily find some errors introduced by automatically annotated information if he/she is working on some specific gene of interest. Manual correction of these errors will be impossible because of the huge number of genes and the great demand of experts on all those genes. A dedicated database (integrated, annotatory, and curative) of a special field such as HPS will be more useful to those researchers who are working in that area.

In this study, we designed a unique Gene-Oriented File (GOF) format in a web-enabled HPS database (HPSD) to compile all the known genes that are responsible for HPS or HPS-related syndromes. We describe both human and mouse orthologs in separate GOFs (58 GOFs in total) as this HPSD emphasizes human and mouse loci. Seven blocks (genomic, transcript, protein, function, mutation, phenotype, and reference) are carefully curated in each user-friendly GOF entry. The text content of each GOF is obtained from 1) searching existing databases; 2) predictions of commonly used bioinformatics servers; 3) reviewing original literature. Each GOF provides not only an interface but also key points of the gene entries, which could be regarded as an updated minireview of the gene. The HPSD is written in HyperText Markup Language (HTML), JavaScript (JS), and Personal Home Page (PHP). Users can access and search the HPSD freely through Internet browsers (http://liweilab.genetics.ac.cn/HPSD).

DATABASE STRUCTURE AND CONTENT

The HPSD has been developed with a model which takes into account various interactions and functions of the database system (Fig. 1).

Source of Data

The data incorporated into the GOFs are gathered from all the available websites are shown in Figure 2. All mutation definitions were unified according to standardized mutation nomenclature [den Dunnen and Antonarakis, 2000]. Numbering of cDNAs is based on the start codon of the reference cDNA sequence (RefSeq).

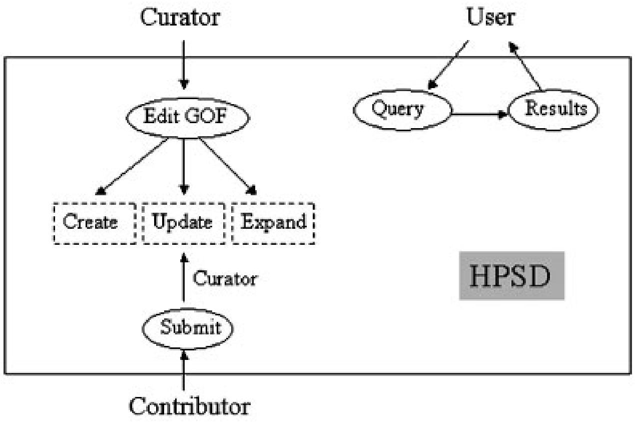


FIGURE 1. **Diagram of HPSD database management system. The solid frame represents the system; the ovals represent the functions of system; the dashed boxes represent the operations of database curator; and arrows show the interactions with the system.**

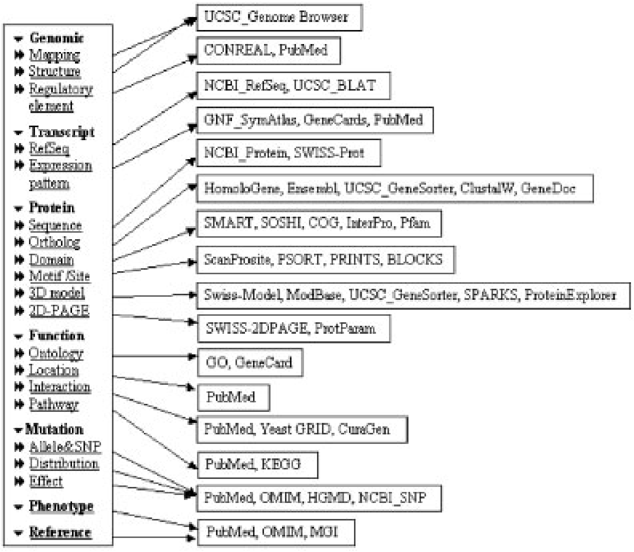


FIGURE 2. **The data structure of Gene-Oriented File (GOF) and data integration of resources in the HPSD. The index of each GOF is shown in the left box. Subtitles of each block are listed in bold in the "Results and Discussion" section. The major websites, programs, or software that are used to obtain the information are shown in the boxes at the right side, which are described in details in the text.**

Database Structure

The HPSD consists of a main page and 58 GOF data pages. The homepage describes the introduction, instructions, comments/feedback, user submission, terms of use, related links, and it also includes a keyword search box and four hyperlinked tables that link to individual GOFs (Fig. 3). Each GOF contains an index frame and a text frame. The index has seven blocks: Genomic, Transcript, Protein, Function, Mutation, Phenotype, and Reference. Except for Phenotype and Reference, there are subtitles under each block (Fig. 2). Each block hyperlinks to the text. The text summarizes the major points in each section. The related information is hyperlinked to relevant websites.

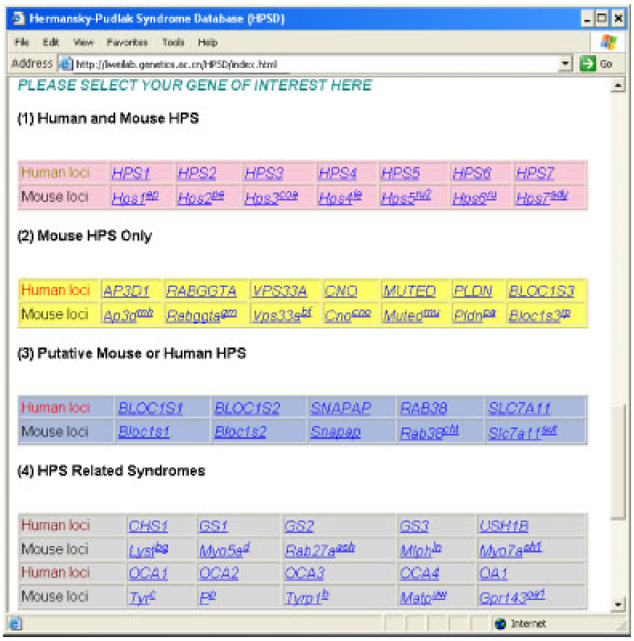
Implementation

The database is hosted on a Windows XP/HP ProLiant ML150 dual Xeon 2.80-GHz server (Hewlett-Packard, Shanghai, China). A backup copy is stored in another Windows XP/HP Pavilion Pentium IV 3.06-GHz workstation (Hewlett-Packard, Shanghai, China), as well as in rewritable CDs whenever a new version is created. The database is implemented with HTML, JS, and PHP. The format of each GOF file is controlled by Cascading Style Sheets (CSS). An Apache server provides a secure, efficient and extensible server for HyperText transfer protocol (HTTP) services.

WEBSITE USE

Web Browsers

Microsoft Internet Explorer (www.microsoft.com/windows/ie/default.mspx) or Netscape Browser (http://browser.netscape.com/ns8/) may be used to view the effects of the Web pages. Browsers that are not compatible with CSS (such as the "gene.css" file) or


Hermansky-Pudlak Syndrome Database (HPSD)
File Edit View Favorites Tools Help
Address http://lwelab.genetics.ac.cn/HPSD/index.html
Go
Internet

PLEASE SELECT YOUR GENE OF INTEREST HERE

(1) Human and Mouse HPS

Human loci	*HPS1*	*HPS2*	*HPS3*	*HPS4*	*HPS5*	*HPS6*	*HPS7*
Mouse loci	$Hps1^{ep}$	$Hps2^{pe}$	$Hps3^{coa}$	$Hps4^{le}$	$Hps5^{ru2}$	$Hps6^{ru}$	$Hps7^{sdy}$

(2) Mouse HPS Only

Human loci	*AP3D1*	*RABGGTA*	*VPS33A*	*CNO*	*MUTED*	*PLDN*	*BLOC1S3*
Mouse loci	$Ap3d^{mh}$	$Rabggta^{gm}$	$Vps33a^{bf}$	Cno^{cno}	$Muted^{mu}$	$Pldn^{pa}$	$Bloc1s3^{rp}$

(3) Putative Mouse or Human HPS

Human loci	*BLOC1S1*	*BLOC1S2*	*SNAPAP*	*RAB38*	*SLC7A11*
Mouse loci	*Bloc1s1*	*Bloc1s2*	*Snapap*	$Rab38^{cht}$	$Slc7a11^{sut}$

(4) HPS Related Syndromes

Human loci	*CHS1*	*GS1*	*GS2*	*GS3*	*USH1B*
Mouse loci	$Lyst^{bg}$	$Myo5a^{d}$	$Rab27a^{ash}$	$Mlph^{ln}$	$Myo7a^{sh1}$
Human loci	*OCA1*	*OCA2*	*OCA3*	*OCA4*	*OA1*
Mouse loci	Tyr^{c}	P^{p}	$Tyrp1^{b}$	$Matp^{uw}$	$Gpr143^{oa1}$

FIGURE 3. **Categories of 58 HPS and HPS-related genes. Gene symbols are in italics. All human gene symbols are HUGO-approved. "Human and Mouse HPS" indicates mutations that have been identified in both human and mouse genes. "Mouse HPS Only" refers to those mutations that have been found in mouse HPS genes, but have not yet been identified as mutations in orthologous human genes so far. "Putative Mouse or Human HPS" contains the three genes encoding the subunits (blos1, blos2, and snapin) of BLOC-1, which are suggestive to cause HPS when mutated, the *Rab38* gene that causes HPS in rats, and the *Slc7a11* gene that causes a mild form of HPS in the subtle gray mice. "HPS-Related Syndromes" represents known genes that share common features with HPS, such as Chediak-Higashi Syndrome (CHS), Griscelli syndrome (GS), oculocutaneous albinism (OCA), Usher syndrome type 1B (USH1B), and ocular albinism (OA). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]**

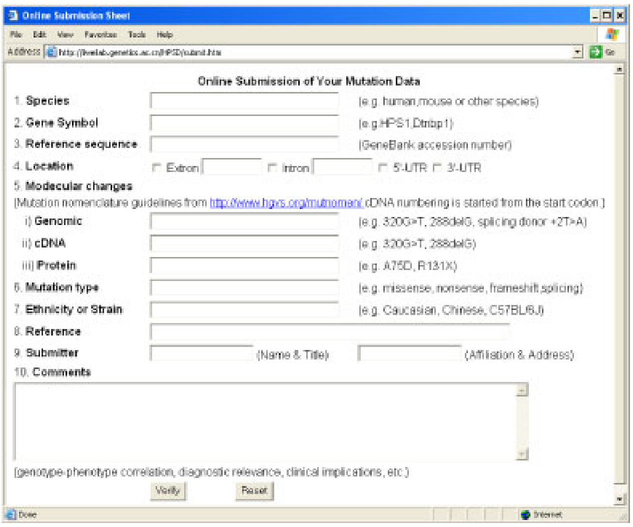
JS may not display the GOF format properly. To browse the database Web pages, a browser must have the plug-ins such as Adobe Acrobat Reader (www.adobe.com), Apple QuickTime PictureViewer or QuickTime Player (www.apple.com).

Query Interface

HPSD has a keyword query interface developed in PHP in the main page. All the data pages in HPSD were converted to structured query language (SQL) files based on the MySQL database system (www.mysql.com). The output of a query links to the specific GOF and highlights the keyword in a found context.

Data Submission

HPSD features an online submission of mutation data. Researchers are encouraged to submit their new mutation data to HPSD regarding all of the current 58 genes by filling out the fields on the online submission form (Fig. 4). This form is developed with PHP. A text file will be automatically generated at the server's side after the submitter verifies and submits the mutation data. The HPSD curator will upload the submitted data and incorporate it into the mutation table based on the guidelines of mutation nomenclature described in the website (www.hgvs.org/mutnomen). The mutation table of each GOF is easily updated through JS codes.


Online Submission Sheet
File Edit View Favorites Tools Help
Address http://lwelab.genetics.ac.cn/HPSD/submit.htm
Go
Done
Internet

Online Submission of Your Mutation Data

1. Species (e.g. human,mouse or other species)
2. Gene Symbol (e.g.HPS1,Dtnbp1)
3. Reference sequence (GeneBank accession number)
4. Location □ Extron □ Intron □ 5'-UTR □ 3'-UTR
5. Modecular changes
[Mutation nomenclature guidelines from http://www.hgvs.org/mutnomen/ cDNA numbering is started from the start codon.]
i) Genomic (e.g. 320G>T, 288delG, splicing donor +2T>A)
ii) cDNA (e.g. 320G>T, 288delG)
iii) Protein (e.g. A75D, R131X)
6. Mutation type (e.g. missense, nonsense, frameshift,splicing)
7. Ethnicity or Strain (e.g. Caucasian, Chinese, C57BL/6J)
8. Reference
9. Submitter (Name & Title) (Affiliation & Address)
10. Comments
(genotype-phenotype correlation, diagnostic relevance, clinical implications, etc.)
Verify Reset

FIGURE 4. **Mutation data online submission sheet. A step-by-step mutation data submission form is incorporated into the HPSD homepage. Examples are shown next to boxes to be filled. Both to-be-verified and to-be-submitted forms will be appeared when the submitter clicks the "Verify" button.**

RESULTS AND DISCUSSION

Each GOF is divided into the following seven blocks.

Genomic

In the "Genomic" block, we retrieve data from the UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgGateway) to show the mapping and genomic structure of each gene. The data integrated into the HPSD is based on Human NCBI Build 34 (assembly, July 2003) or Mouse NCBI Build 32 (assembly, Oct. 2003). BAC clones that cover the entire gene are listed in the genomic map. The neighboring genes are also shown in this map. Variants of a gene that have different structures are shown in the genomic structure map.

The 5′ untranslated region (5′-UTR) and 1-kb upstream region are examined for regulatory elements and binding sites of a set of transcription factors from the TransFac database (www.gene regulation.com) by using the CONREAL server (http://conreal.niob.knaw.nl) by comparing the human and mouse genomic sequences. Mutations in these conserved regions may affect the expression of the gene.

Transcript

The "Transcript" block contains two subtitles: RefSeq and Expression Pattern. The NCBI Reference Sequence (RefSeq; www.ncbi.nlm.nih.gov/RefSeq) provides a nonredundant set of sequences, including genomic DNA, transcript (RNA), and protein products, for major model organisms. For each mRNA RefSeq, the alignment to genomic is generated through the BLAT tool at the University of California Santa Cruz (UCSC; http://genome.ucsc.edu/cgi-bin/hgBlat), in which the open reading frame (ORF) and UTRs are shown. The genomic structures are useful resources for mutation screening at both genomic and cDNA levels.

Tissue specificity is a major concern of researchers in experimental design. In HPSD, we summarize the expression profiles from the original publications, as well as provide intuitionistic histograms converted from raw microarray data as featured in GeneCards (http://bioinfo.weizmann.ac.il/cards/index.shtml) for all human genes or in GNF SymAtlas (http://symatlas.gnf.org/SymAtlas/) for all mouse genes.

Protein

In the "Protein" block, six sections are included to describe the proteomic aspects of the gene product.

The protein RefSeq is also from the NCBI protein reference sequence database as described above. The Swiss_Prot entry is also viewed through ExPaSy (http://au.expasy.org/sprot/), which provides a comprehensive interface of the polypeptide.

The ortholog table contains all the known homologous proteins by referring to NCBI HomoloGene (www.ncbi.nlm.nih.gov/entrez/query.fcgi?db = homologene), UCSC GeneSorter, Ensembl Genome Browser (www.ensembl.org). It is difficult to identify the true orthologs in some organisms when the identities are fairly low or when the protein is a member of a superfamily. We have eliminated some examples of apparently non-orthologous proteins that appeared in the above websites. "Gene View" in the table links to the overview of each homologous gene. "Protein" links to NCBI protein sequence entry. Using the BLASTP 2 Sequences (www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi) program, the percentages of identities that span the compared region are calculated. ClustalW (downloaded from ftp.ebi.ac.uk) is used for multiple sequence alignments (MSA). The results are exported to GeneDoc (downloaded from www.psc.edu/biomed/genedoc) for editing and shading the amino acids. The MSA PDF file is produced by Acrobat Distiller from the GeneDoc output. The MSA PDF file is a good resource for domain searching, evolutionary studies, and mutation analysis.

Based on the gene family and multiple sequence alignment, domains of a protein are deposited in databases such as Pfam (www.sanger.ac.uk/Software/Pfam), NCBI Conserved Domain Database (CDD; www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml) and SMART (http://smart.embl-heidelberg.de). InterPro (www.ebi.ac.uk/interpro) is a widely used composite server to search the domains deposited in several databases. The transmembrane (TM) domains are often predicted by several reliable programs. SOSUI (http://sosui.proteome.bio.tuat.ac.jp/sosuiframe0.html) is one of the most commonly used servers for searching TMs. To search the domains of the protein of interest in the above websites, we cut and paste the primary sequence or input the Swiss-Prot ID.

The motifs or binding sites of a protein are collected in databases such as PRINTS (http://umber.sbs.man.ac.uk/dbbrowser/PRINTS), SCOP (http://scop.mrc-lmb.cam.ac.uk/scop), CATH (www.biochem.ucl.ac.uk/bsm/cath) and PROSITE (http://us.expasy.org/prosite). ScanProsite (http://us.expasy.org/tools/scanprosite) and PSORT II (http://psort.nibb.ac.jp/form2.html) are the two commonly used composite platforms to search these databases. These predictions of the secondary structures of a protein, together with the domain searching results will give clues to protein location, posttranslational modifications, protein sorting signals, and protein interactions.

Three-dimensional (3D) structures are expanding dramatically by X-ray crystallography or nuclear magnetic resonance (NMR), though there are limitations in analyses of insoluble membrane protein and large proteins. Computational molecular modeling provides an alternative method to view the 3D structures of the proteins without experimentally determined structures. Currently, three major molecular modeling predictions are used: homology modeling (sequence-based), fold recognition (structure-based), and *ab initio* prediction [Banerjee-Basu and Baxevanis, 2001].

Homology modeling is based on homology analysis (usually more than 30% identities) and template structures. ModBase (http://alto.compbio.ucsf.edu/modbase-cgi/index.cgi) is a database of collecting the 3D protein models. For those proteins in HPSD with match(es) in the ModBase, the 3D models with the front, top, and side view are downloaded from UCSC GeneSorter directly.

Fold recognition is based on structure recognition by threading the folds to the protein sequence. There are many protein folding predictors such as SPARKS (http://theory.med.buffalo.edu/hzhou/anonymous-fold-sparks2.html) and 3D-PSSM (http://www.sbg.bio.ic.ac.uk/~3dpssm). Currently, the best fold recognition algorithms are based on the combined methods of homology modeling and genuine fold recognition, which are called profile–profile based methods [Friedberg et al., 2004]. SPARKS stands out for its sensitivities and its accuracy by using a profile–profile based method [Zhou and Zhou, 2004]. To view the 3D model, a variety of visualization tools are in use. The freely accessed, commonly used visualization tools include Protein Explorer (http://molvis.sdsc.edu/protexpl/frntdoor.htm), which requires the MDL Chime (www.mdlchime.com) plug-in, PDB2MGIF (www.dkfz-heidelberg.de/spec/pdb2mgif). Most of the visualization tools utilize PDB file format (www.umass.edu/microbio/rasmol/pdb.htm) and share the RasMol (www.umass.edu/microbio/rasmol) commands. In HPSD we use PDB2MGIF for displaying the predicted 3D models. The RasMol codes for displaying the pictures are as follows:

```
wireframe off
centre
background black
select protein
color structure
cartoon 400
backbone 200
```

SWISS-2DPAGE (http://us.expasy.org/ch2d) contains data on proteins identified on various 2D PAGE and SDS-PAGE reference maps. None of the proteins in HPSD has been found in the SWISS-2DPAGE database. However, the molecular weight (MW) and pI are computed with ProtParam (http://us.expasy.org/tools/protparam.html). The estimated size in Western blot analysis may differ from the theoretical MW because of post-translational modifications or electrophoresis conditions.

Function

To get the basic overview of the function of a gene, the Gene Ontology (www.geneontology.org) provides three structured, controlled vocabularies (ontologies) that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner.

Localization of a protein is mainly based on reviewing the published data of the PubMed database (www.ncbi.nlm.nih.gov/entrez/query.fcgi).

Data entries of protein–protein interactions or protein binding partners are based on the experimental data in PubMed. Although there are several interaction databases such as DIP, BIND and MIPS, users will find no hits if the original authors have not submitted their findings to the databases or if those papers have not been reviewed or updated by the database curators. When a protein has a Drosophila or Yeast homolog, 1 yeast-two-hybrid interactions are linked to the CuraGen Drosophila interaction database (http://portal.curagen.com/cgi-bin/interaction/flyHome.pl) or Yeast GRID database (http://biodata.mshri.on.ca:80/yeast_grid/servlet/SearchPage). Users should be aware that some of these interactions are false-positives. We do not provide prediction of protein–protein interaction in the current version of HPSD.

The protein network/pathway is defined upon the function and interaction of a protein. KEGG Pathway Database (www.genome.ad.jp/kegg/pathway.html) is a collection of metabolic pathways. For example, the role of tyrosinase in tyrosine metabolism is well defined in the KEGG database. We also draw several diagrams to depict the vesicle trafficking pathways that will be helpful to understand the molecular and cellular mechanisms of HPS.

Mutation

In this block, we linked all the mutational alleles or SNPs to the public databases such as HGMD (http://archive.uwcm.ac.uk/uwcm/mg/hgmd0.html), dbSNP (www.ncbi.nlm.nih.gov/SNP/), OMIM (www.ncbi.nlm.nih.gov/entrez/query.fcgi?db = OMIM), and MGI.

For the distribution of the identified mutations, we generated a table to define the details of the mutations, including Location, Genomic, cDNA, Protein, Type, Ethnicity (human)/Strain (mouse), and Reference. Our focuses are on the HPS genes, while the distribution of the genes of HPS-related syndromes are better described in other databases such as the Albinism Database (http://albinismdb.med.umn.edu) and the Retina International Scientific Newsletter Databases (www.retina-international.com/sci-news/database.htm). Links to these mutation databases are available in HPSD. All mutations collected in HPSD are carefully curated and errors are corrected. The discrepancies arising from different cDNA numbering systems are resolved by using the first base pair of start codon as number 1. It is easy to identify the hotspots of mutation by examining the distribution table and the mutation frequencies.

The effects of mutations are defined from the published papers. It is better to check the MSA PDF file to see if a missense mutation is conserved or not. Advanced users may use the PDB file to model the mutation by comparing the 3D structures (an example is shown in http://liweilab.genetics.ac.cn/HPSD/sut.htm).

The description of genotype–phenotype relationship is based on observations of case reports. The profound effect of nonsense mutation is nonsense-mediated decay (NMD). The mechanisms by which NMD operates have received much attention [Holbrook et al., 2004; Maquat, 2004]. NMD eliminates mRNAs containing premature termination codons (PTCs) and thus helps limit the synthesis of abnormal proteins. It protects many heterozygous carriers of genes with PTC mutations from manifesting disease phenotypes that would result from expression of truncated proteins. On the other hand, the NMD-incompetent PTCs give rise to the production of truncated proteins that may overwhelm the cells as in beta-thalassemia [Kugler et al., 1995]. Different patterns of inheritance may occur in the heterozygotes due to the position of PTCs. It also regulates the alternative splice forms through degradation of transcripts containing PTC. Although the targets of NMD can be predicted as the transcripts containing PTCs 5' to the last 50 bp of the penultimate exon, the exceptions of NMD suggest that the functional consequences of a PTC mutation must be established by experiments. Hence, we include the NMD results in HPSD by reviewing the experimental data.

Phenotype

Phenotypes of human patients are from the original papers or OMIM. The description of mouse mutants is based on the original papers or MGI entries. To better understand the mouse phenotypes, the link to Mouse Locus Catalog in MGI and JAX MICE (http://jaxmice.jax.org/) are two links that cannot be overlooked.

Reference

The key references of each GOF are listed. Each reference is linked to the PubMed database when an abstract is available.

In summary, we here provide comprehensive data integration to all of the HPS and HPS-related genes (so far 58 genes in HPSD). The GOF format is unique. It may be applied to similar dedicated databases in some special fields in clinical bioinformatics. All the information in the HPSD is peer-reviewed. The links of this database dramatically expand the contents of the HPSD. The user-friendly webpages will be impressive to the users who browse this database.

A web-enabled database in a file system is accessible to any user in the world through the Internet. The easy-to-learn HTML programming makes the coding of the database simple. It is suitable for maintenance or updating at the server-side. We will expand the entries of HPSD whenever a new HPS or HPS-related gene is identified and will keep updating the HPSD in a timely manner.

The purpose of developing the HPSD is not only to provide a database in clinical bioinformatics to be conveniently accessed by the scientists who are interested in the field of HPS and related research, but also to provide a format which may be unified as large-scale data entries at the systems biology level. To edit or update the GOF more efficiently, algorithms that automatically edit a block such as Function or Phenotype are under development.

ACKNOWLEDGMENTS

We thank Dr. Richard Swank and Dr. Edward Novak for their help in peer-reviewing the entire database and in providing revisions to this manuscript. We thank Dr. Rosemary Elliott and the other members at Dr. Richard Swank's laboratory for their invaluable suggestions to this database. This work was supported in part by grants from the Distinguished Young Scholars Fund of the National Natural Science Foundation of China (30525007; to W.L.), the National Institutes of Health Grant (CA101515; to P.L.), and the National Science Foundation (0098082; to J.W.B.).

REFERENCES

Banerjee-Basu S, Baxevanis AD. 2001. Predictive methods using protein sequences. In: Baxevanis AD, Ouellette BFF, editors. Bioinformatics: a practical guide to the analysis of genes and proteins, 2nd edition. New York: Wiley-InterScience. p 253–282.

Chintala S, Li W, Lamoreux ML, Ito S, Wakamatsu K, Sviderskaya EV, Bennett DC, Park Y-M, Gahl WA, Huizing M, Spritz RA, Ben S, Novak EK, Tan J, Swank RT. 2005. Slc7a11 controls the production of pheomelanin pigment and the proliferation of cultured cells. Proc Natl Acad Sci USA 102:10964–10969.

Dell 'Angelica EC, Mullins C, Caplan S, Bonifacino JS. 2000. Lysosome-related organelles. FASEB J 14:1265–1278.

den Dunnen JT, Antonarakis SE. 2000. Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. Hum Mutat 15:7–12.

Di Pietro SM, Dell 'Angelica EC. 2005. The cell biology of Hermansky-Pudlak syndrome: recent advances. Traffic 6: 525–533.

Friedberg I, Jaroszewski L, Ye Y, Godzik A. 2004. The interplay of fold recognition and experimental structure determination in structural genomics. Curr Opin Struct Biol 14:307–312.

Gardner JM, Wildenberg SC, Keiper NM, Novak EK, Rusiniak ME, Swank RT, Puri N, Finger JN, Hagiwara N, Lehman AL, Gales TL, Bayer ME, King RA, Brilliant MH. 1997. The mouse pale ear (ep) mutation is the homologue of human Hermansky-Pudlak syndrome. Proc Natl Acad Sci USA 94: 9238–9243.

Hermansky F, Pudlak P. 1959. Albinism associated with hemorrhagic diathesis and unusual pigmented reticular cells in the bone marrow: report of two cases with histochemical studies. Blood 14:162–169.

Holbrook JA, Neu-Yilik G, Hentze MW, Kulozik AE. 2004. Nonsense-mediated decay approaches the clinic. Nat Genet 36: 801–808.

Huizing M, Gahl WA. 2002. Disorders of vesicles of lysosomal lineage: the Hermansky-Pudlak syndromes. Curr Mol Med 2: 451–467.

Kugler W, Enssle J, Hentze MW, Kulozik AE. 1995. Nuclear degradation of nonsense mutated beta-globin mRNA: a post-transcriptional mechanism to protect heterozygotes from severe clinical manifestations of beta-thalassemia? Nucleic Acids Res 23:413–418.

Li W, Rusiniak ME, Chintala S, Gautam R, Novak EK, Swank RT. 2004. Murine Hermansky-Pudlak syndrome genes: regulators of lysosome-related organelles. BioEssays 26:616–628.

Maquat LE. 2004. Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. Nat Rev Mol Cell Biol 5: 89–99.

Oh J, Bailin T, Fukai K, Feng GH, Ho L, Mao J, Frenk E, Tamura N, Spritz RA. 1996. Positional cloning of a gene for Hermansky-Pudlak syndrome, a disorder of cytoplasmic organelles. Nat Genet 14:300–306.

Swank RT, Novak EK, McGarry MP, Rusiniak ME, Feng L. 1998. Mouse models of Hermansky Pudlak syndrome: a review. Pigment Cell Res 11:60–80.

Zhou H, Zhou Y. 2004. Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. Proteins 55: 1005–1013.