Human Transporter Database: Comprehensive Knowledge and Discovery Tools in the Human Transporter Genes

Adam Y. Ye[1,4], **Qing-Rong Liu**[2], **Chuan-Yun Li**[3], **Min Zhao**[1]*, **Hong Qu**[1]*

1 Center for Bioinformatics, State Key Laboratory of Protein and Plant Gene Research, College of Life Sciences, Peking University, Beijing, China, **2** Behavioral Neuroscience Research Branch, NIH-IRP (NIDA), Baltimore, Maryland, United States of America, **3** Institute of Molecular Medicine, Peking University, Beijing, China, **4** Peking-Tsinghua Center for Life Sciences, College of Life Sciences, Peking University, Beijing, China

Abstract

Transporters are essential in homeostatic exchange of endogenous and exogenous substances at the systematic, organic, cellular, and subcellular levels. Gene mutations of transporters are often related to pharmacogenetics traits. Recent developments in high throughput technologies on genomics, transcriptomics and proteomics allow in depth studies of transporter genes in normal cellular processes and diverse disease conditions. The flood of high throughput data have resulted in urgent need for an updated knowledgebase with curated, organized, and annotated human transporters in an easily accessible way. Using a pipeline with the combination of automated keywords query, sequence similarity search and manual curation on transporters, we collected 1,555 human non-redundant transporter genes to develop the Human Transporter Database (HTD) (http://htd.cbi.pku.edu.cn). Based on the extensive annotations, global properties of the transporter genes were illustrated, such as expression patterns and polymorphisms in relationships with their ligands. We noted that the human transporters were enriched in many fundamental biological processes such as oxidative phosphorylation and cardiac muscle contraction, and significantly associated with Mendelian and complex diseases such as epilepsy and sudden infant death syndrome. Overall, HTD provides a well-organized interface to facilitate research communities to search detailed molecular and genetic information of transporters for development of personalized medicine.

Citation: Ye AY, Liu Q-R, Li C-Y, Zhao M, Qu H (2014) Human Transporter Database: Comprehensive Knowledge and Discovery Tools in the Human Transporter Genes. PLoS ONE 9(2): e88883. doi:10.1371/journal.pone.0088883

Editor: Yu Xue, Huazhong University of Science and Technology, China

Received November 15, 2013; **Accepted** January 12, 2014; **Published** February 18, 2014

Funding: This work was supported by the National Natural Science Foundation of China [grant number 31171270] and the National Basic Research Program of China [grant number 2011CB518000]. QL is supported by the intramural research program of NIDA-NIH. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: zhaomin.2006@gmail.com (MZ); quh@mail.cbi.pku.edu.cn (HQ)

Introduction

Transport proteins (transporters) are membrane channels and molecular pumps to facilitate exchange of ions, small molecules, macromolecules, and drugs across membranes [1]. The movement of biochemical compound through membrane is critical to absorption, distribution, metabolism, and excretion (ADME) of nutrients, neurotransmitters, and drugs [2–5,6]. The dynamic partnerships of transporter with other signaling molecules in subcellular locations are regarded as essential processes for cellular function. Attenuation of transporter gene functions by polymorphisms often contributes to complex human diseases and individual drug responses [4,5–9]. How do transporters cooperate with intracellular signaling systems and metabolic systems to give precise control of transmembrane trafficking? Although crystal structures have shed light on the regulatory mechanisms of a few individual transporters as gateway for metabolites and signals in the past decade [10], the global features of transporter genes are still not clear. Recent advances in high throughput technologies, such as mass spectrometry (MS), genome-wide association study (GWAS), and next-generation sequencing (NGS), provide abundant complementary data to study transporting processes or the effects of transporters on normal cellular processes and various disease states [11,12]. A comprehensive database of human transporters is required to incorporate the most updated high throughput data in an intuitive search engine.

There are two types of previous transporter databases: general transporter collections and gene family specific collections. The earlier general transporter databases include TCDB (Transporter Classification Database), TransportDB, KEGG (Kyoto Encyclopedia of Genes and Genomes), HMTD (Human Membrane Transporter Database), and TSdb (Transporter substrate database). TCDB is dedicated to transporter classification based on functional and phylogenetic information [13], which contains 513 human, 364 mouse, and 165 rat transporters. TransportDB focuses on prediction cytoplasmic membrane transporters for comparative studies with 1,022 human and 1,090 mouse transporters [14]. In KEGG PATHWAY and BRITE database, there are 870 transporter orthology groups in prokaryotes and eukaryotes, which maps to 420 human genes [15]. HMTD (Human Membrane Transporter Database) is specific for drug transport studies and pharmacogenomics with 287 human transporters [16]. TSdb (Transporter substrate database) is constructed to annotate substrates of transporters [17]. Another type of gene family specific transporter databases only focus on specific transporter families including ABCdb (ATPbinding cassette transporters database) [18], MTDB (Medicago truncatula transporter database) [19], and SLCdb (Caenorhabditis elegans SLC homologue database, http://www.wormslc.org/). However, most of the transporter databases were derived from low throughput data, and without integrating high throughput expression and polymorphism data, or without systematically updating for recent pharmacogenetic data.

A lack of integration of these high throughput data across functional, pharmaceutical, and genetic studies hampers our understanding of the molecular mechanisms of transporter related diseases. Some transporters can influence drug efficacy, and their activity can also be affected by some drugs, thus when two or more drugs are coadministered, their dosage may need adjustment [20]. In addition, natural variants such as single-nucleotide polymorphism (SNP) may also affect transporter activity [21], and may sometimes make the protein more sensitive to drug [22]. Data integration will be useful for generating new hypothesis, such as dosage and safety warnings on drug coadministration or population polymorphism, refining our understanding of cellular transporting system in human disease states and development of transporter gene based pharmacogenetics [23]. To provide insight into human transporter systems, we collected 1,555 human non-redundant transporter genes and constructed Human Transporter Database (HTD), a repository for dynamic storage of the ever-increasing bioinformatics on transporter genes in light of personalized medicine.

We extensively annotated human transporter genes (HTGs) from the perspective of sequences, functions, drugs, diseases, pharmacogenetics, genetic variations, interactions, and gene expressions. We noted that the human transporters were enriched in fundamental biological processes and involved in a number of complex human diseases. Overall, HTD provides a publicly accessible resource and a searchable database for communities to explore the human transporters gene families, functional substrates, expressions and polymorphisms in a global way. It is freely available at http://htd.cbi.pku.edu.cn.

Materials and Methods

Collection of Human Transporter Genes

We specifically defined transporters as the membrane proteins facilitating materials (mostly molecules or ions) transporting across membrane. In order to get precise descriptive keywords for transporters, we extensively reviewed 1,178 human transporter genes integrated from NCBI Gene database and four relevant transporter datasets: (i) Transporter classification (TC) systems; (ii) TransportDB; (iii) Transporter family and gene list from HMTD; and (iv) KEGG BRITE transporter (ko02000) and solute carrier family (ko02001). Based on the transporter definition, gene description, and GO annotation in NCBI Gene database, we compiled 54 keywords precisely related to transporter gene names and functions (Table S1). In this process, we excluded some keywords irrelevant with membrane transporting such as "fatty acid binding", which mainly represented apolipoproteins, the proteins bind lipids and transport lipid through circulatory system, and are seldom embedded in cellular membranes for transporting functions.

Using the 54 keywords, we utilized NCBI E-search interface to implement complex query against NCBI Gene database. In total, 1,592 human genes were obtained. In this process, pseudogenes were included, as they may play regulatory roles on transporter related biological processes [24]. Based on gene description, alias, GO annotation, and domain feature, we manually removed those genes irrelevant with transporter function. Further we performed BLAST similarity alignment with these refined genes against all protein sequences in the human genome to include less annotated genes but with high sequence similarities with curated transporter genes. Through additional manually checking, 1,555 human transporter genes with high confidence were stored in our HTD database.

Gene Annotation

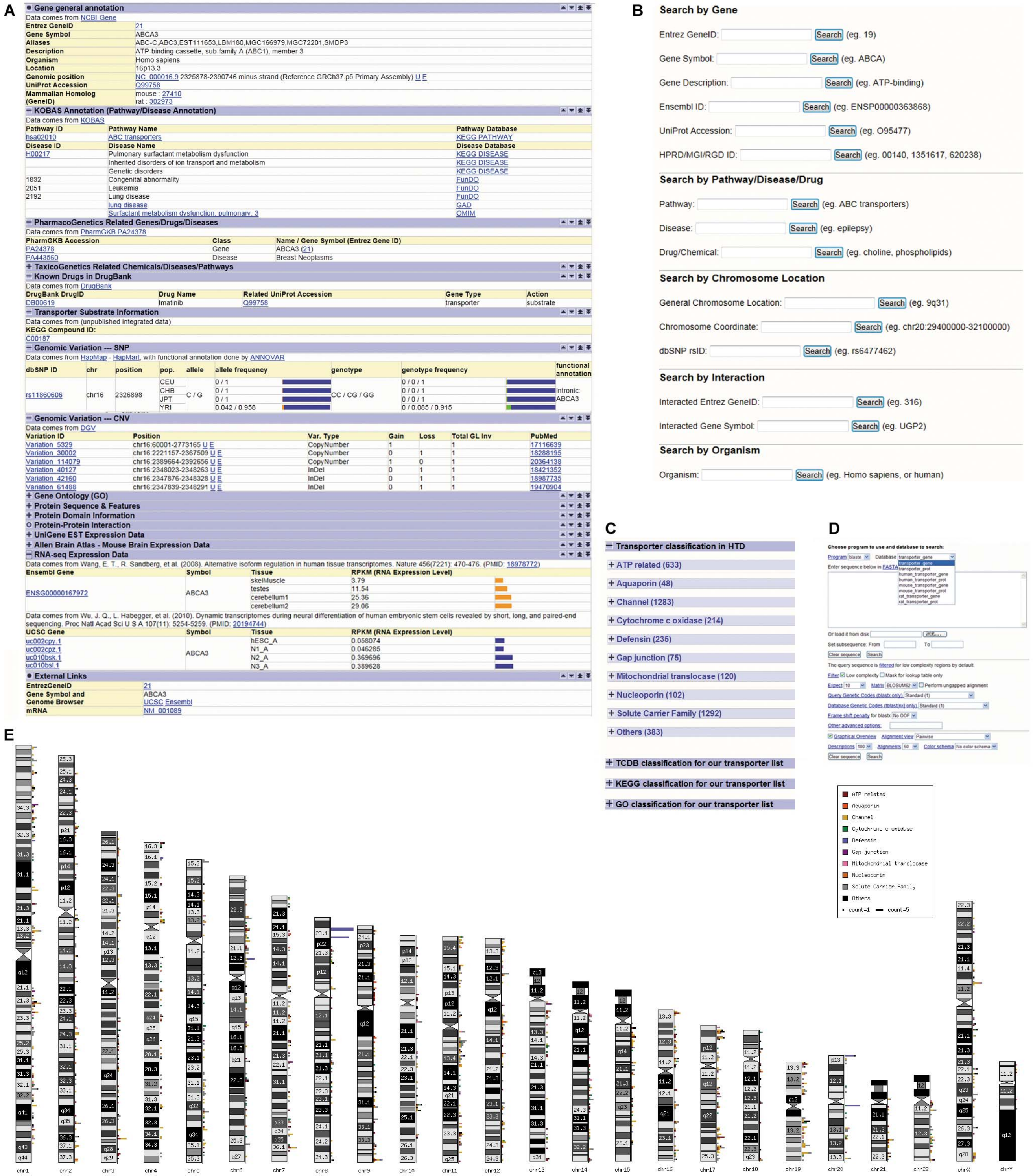
To systematically mine the biological mechanism related to transporter genes, we annotated all transporters in our HTD with extensive functional information. The statistics of those annotation entries in HTD was listed in Table S2. We first extracted basic information including gene symbol, annotation, and function from NCBI Gene database [25], GO annotation, protein sequence and features from UniProt [26], protein domain annotation from InterPro [27], Homolog GeneID mapping data from NCBI HomoloGene [28], SNPs with minor allele frequency or population genetic information from dbSNP [29] and HapMap [30], CNV data from DGV (31), protein-protein interaction information from HPRD [32], expression data from NCBI UniGene [25], Allen Brain Atlas [33] and supplementary materials [34,35], haplotype, epigenetics and regulation information from UCSC and ENCODE [36], and drug information from Pharmacogenomics Knowledge Base (PharmGKB) [37], Comparative Toxicogenomics Database (CTD) [38] and DrugBank [39], and then parsed the data and rearranged it into well-formatted tables. Transporter substrates were integrated from transporter substrate database (TSdb) [17], whose substrates and drugs of a given transporter were mapped to KEGG LIGAND database. In addition, enriched pathways and diseases information was annotated by KOBAS 2.0 [40].

Interface Development of Database

The data of HTD is stored in a MySQL relational database [41–43]. As shown in Figure S1, to balance the query efficiency and data storage, we integrated multiple dimensional data into some tables, which was easier for update and more flexible to integrate other potential data source. In addition, we mainly used the NCBI Gene ID to crosslink the data tables in our database. The web interface is developed with PHP on Apache server. To achieve a better organization, development and maintenance, we designed the web application into a three-level architecture carefully, including: business layer, which parsed input query and did SQL queries, output function layer, which did sorting and table outputting, and the top web representation layer, which included the frame and styles, and also did some simple redirection. The main interfaces of HTD are user-friendly in browsing the classified transporter genes, querying keywords, and searching sequences for transporter genes (Figure 1).

Results and Discussion

Combining of automated keywords query, sequence similarity search and manual curation on transporters, we collected 1,555 human non-redundant transporter genes to develop the Human Transporter Database (HTD). To provide a reference, we also collected 1,422 and 1,453 transporter genes for mouse and rat, respectively, in a similar way. A quick exploration showed that 383 (80 protein-coding genes and 303 pseudogenes) human transporter genes (HTGs) were only found in human genome in comparison with transporters from the two rodent genomes, which might be either primate or human specific transporters. To better organize



● Gene general annotation

Data comes from NCBI-Gene

Entrez GeneID	21
Gene Symbol	ABCA3
Aliases	ABC-C,ABC3,EST111653,LBM180,MGC166979,MGC72201,SMDP3
Description	ATP-binding cassette, sub-family A (ABC1), member 3
Organism	Homo sapiens
Location	16p13.3
Genomic position	NC_000016.9 2325878-2390746 minus strand (Reference GRCh37.p5 Primary Assembly) U E
UniProt Accession	Q99758
Mammalian Homolog (GeneID)	mouse : 27410 rat : 302973

− KOBAS Annotation (Pathway/Disease Annotation)

Data comes from KOBAS

Pathway ID	Pathway Name	Pathway Database
hsa02010	ABC transporters	KEGG PATHWAY
Disease ID	**Disease Name**	**Disease Database**
H00217	Pulmonary surfactant metabolism dysfunction	KEGG DISEASE
	Inherited disorders of ion transport and metabolism	KEGG DISEASE
	Genetic disorders	KEGG DISEASE
1832	Congenital abnormality	FunDO
2051	Leukemia	FunDO
2192	Lung disease	FunDO
	lung disease	GAD
	Surfactant metabolism dysfunction, pulmonary, 3	OMIM

− PharmacoGenetics Related Genes/Drugs/Diseases

Data comes from PharmGKB PA24378

PharmGKB Accession	Class	Name / Gene Symbol (Entrez Gene ID)
PA24378	Gene	ABCA3 (21)
PA443560	Disease	Breast Neoplasms

+ TaxicoGenetics Related Chemicals/Diseases/Pathways

− Known Drugs in DrugBank

Data comes from DrugBank

DrugBank DrugID	Drug Name	Related UniProt Accession	Gene Type	Action
DB00619	Imatinib	Q99758	transporter	substrate

− Transporter Substrate Information

Data comes from (unpublished integrated data)

KEGG Compound ID:

C00187

− Genomic Variation --- SNP

Data comes from HapMap - HapMart, with functional annotation done by ANNOVAR

dbSNP ID	chr	position	pop.	allele	allele frequency	genotype	genotype frequency	functional annotation
rs11860606	chr16	2326898	CEU	C / G	0 / 1	CC / CG / GG	0 / 0 / 1	intronic: ABCA3
			CHB		0 / 1		0 / 0 / 1	
			JPT		0 / 1		0 / 0 / 1	
			YRI		0.042 / 0.958		0 / 0.085 / 0.915	

− Genomic Variation --- CNV

Data comes from DGV

Variation ID	Position	Var. Type	Gain	Loss	Total GL Inv	PubMed
Variation_5329	chr16:60001-2773165 U E	CopyNumber	1		1	17116639
Variation_30002	chr16:2221157-2367509 U E	CopyNumber	0	1	1	18288195
Variation_114079	chr16:2389664-2392656 U E	CopyNumber	1	0	1	20364138
Variation_40127	chr16:2348023-2348263 U E	InDel	0	1	1	18421352
Variation_42160	chr16:2347876-2348328 U E	InDel	0	1	1	18987735
Variation_61488	chr16:2347839-2348291 U E	InDel	0	1	1	19470904

+ Gene Ontology (GO)

+ Protein Sequence & Features

+ Protein Domain Information

○ Protein-Protein Interaction

+ UniGene EST Expression Data

+ Allen Brain Atlas - Mouse Brain Expression Data

− RNA-seq Expression Data

Data comes from Wang, E. T., R. Sandberg, et al. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature 456(7221): 470-476. (PMID: 18978772)

Ensembl Gene	Symbol	Tissue	RPKM (RNA Expression Level)
ENSG00000167972	ABCA3	skelMuscle	3.79
		testes	11.54
		cerebellum1	25.36
		cerebellum2	29.06

Data comes from Wu, J. Q., L. Habegger, et al. (2010). Dynamic transcriptomes during neural differentiation of human embryonic stem cells revealed by short, long, and paired-end sequencing. Proc Natl Acad Sci U S A 107(11): 5254-5259. (PMID: 20194744)

UCSC Gene	Symbol	Tissue	RPKM (RNA Expression Level)
uc002cpy.1	ABCA3	hESC_A	0.058074
uc002cpz.1		N1_A	0.046285
uc010bsk.1		N2_A	0.369696
uc010bsl.1		N3_A	0.389628

● External Links

EntrezGeneID	21
Gene Symbol and Genome Browser	ABCA3 UCSC Ensembl
mRNA	NM_001089

Search by Gene

Entrez GeneID: Search (eg. 19)

Gene Symbol: Search (eg. ABCA)

Gene Description: Search (eg. ATP-binding)

Ensembl ID: Search (eg. ENSP00000363868)

UniProt Accession: Search (eg. O95477)

HPRD/MGI/RGD ID: Search (eg. 00140, 1351617, 620238)

Search by Pathway/Disease/Drug

Pathway: Search (eg. ABC transporters)

Disease: Search (eg. epilepsy)

Drug/Chemical: Search (eg. choline, phospholipids)

Search by Chromosome Location

General Chromosome Location: Search (eg. 9q31)

Chromosome Coordinate: Search (eg. chr20:29400000-32100000)

dbSNP rsID: Search (eg. rs6477462)

Search by Interaction

Interacted Entrez GeneID: Search (eg. 316)

Interacted Gene Symbol: Search (eg. UGP2)

Search by Organism

Organism: Search (eg. Homo sapiens, or human)

− Transporter classification in HTD

+ ATP related (633)

+ Aquaporin (48)

+ Channel (1283)

+ Cytochrome c oxidase (214)

+ Defensin (235)

+ Gap junction (75)

+ Mitochondrial translocase (120)

+ Nucleoporin (102)

+ Solute Carrier Family (1292)

+ Others (383)

+ TCDB classification for our transporter list

+ KEGG classification for our transporter list

+ GO classification for our transporter list

Choose program to use and database to search:

Program blastn Database transporter_gene

transporter_gene
transporter_prot
human_transporter_gene
human_transporter_prot
mouse_transporter_gene
mouse_transporter_prot
rat_transporter_gene
rat_transporter_prot

Enter sequence below in FASTA

Or load it from disk 浏览...

Set subsequence: From To

Clear sequence Search

The query sequence is filtered for low complexity regions by default.

Filter ☑ Low complexity ☐ Mask for lookup table only

Expect 10 Matrix BLOSUM62 ☐ Perform ungapped alignment

Query Genetic Codes (blastx only) Standard (1)

Database Genetic Codes (tblast[nx] only) Standard (1)

Frame shift penalty for blastx No OOF

Other advanced options:

☑ Graphical Overview Alignment view Pairwise

Descriptions 100 Alignments 50 Color schema No color schema

Clear sequence Search

Figure 1. Web interface in HTD. (A) A typical entry for transporter gene (B) Text search; (C) Browsing Human Transporter classification (D) BLAST interface; (E) Chromosome distribution for all the human transporter genes.
doi:10.1371/journal.pone.0088883.g001

our curated transporter genes, we classified HTGs into ten categories such as “ATP related”, “Channel”, and “Solute Carrier Family” with 98 terms specific for Human Transporters. Based on the statistics on the ten categories (Table 1), 70% of HTGs were from “ATP related”, “Channel”, and “Solute Carrier Family”. We conducted a semi-automatic pipeline to map transporter genes to known transporter classification systems, which mainly included BLAST search and manual checking for TC system, and ID mapping for GO system and KEGG BRITE system. As shown in Figure S2, the number of genes in each TC second-level category

Table 1. The statistics for transporter genes in ten HTD categories for Human, Mouse and Rat.

Category	Number of transporter genes		
	Human	Mouse	Rat
ATP-binding cassette	222	197	198
Aquaporin	19	13	13
Channel	427	397	442
Cytochrome C oxidase	75	54	63
Defensin	72	102	60
Gap junction	28	23	24
Mitochondrial translocase	39	36	42
Nucleoporin	39	36	27
Solute carrier family	454	404	430
Others+Unclassified	170	160	154
Total	1555	1422	1453

doi:10.1371/journal.pone.0088883.t001

was counted after mapping to TC system, and it showed that majority (60%) of HTGs were from 2.A (Porters (uniporters, symporters, antiporters)), 1.A (α-type Channels), and 3.A (P-P-bond-hydrolysis-driven transporters). Indeed, these three TC categories are almost corresponding to 'Solute Carrier Family', 'Channel' and 'ATP related' in our classification system.

Information on the Transporter Gene Page

In each gene page, detailed information of general functions, pathways, genetic polymorphisms, phamacogenetics, substrates were listed; and cross-references to their origin databases such as UniProt, PharmGKB, and DrugBank were included (Figure 1A). To better organize information from large scale studies, we collapsed detailed information for protein sequences, features or domain information, protein-protein interactions, and gene expression profiles by default. Clicking on the expanding links '+ ' of each annotation can bring the graphic views of allele frequency, genotype frequency for SNPs and tissue expression profile of each transporter. Users can expand all the annotation in gene page by clicking "Expand all" button in the top. When exploring each type of annotation, users can click the up/down arrow in the right of the table to reach specific annotation quickly (Figure 1A).

For query speed, the full SNP annotation for a transporter gene was shown in another similar page, which allowed users to filter and only leave exonic or nonsynonymous SNPs and to sort those related SNPs by position, minor allele frequency, difference on population allele frequency, heterozygosity, or functional annotation. The functional annotation for each SNP was mainly based on ANNOVAR [44], including intronic/exonic, synonymous/non-synonymous, SIFT score [45] and PolyPhen-2 score [46,47].

Browsing the Classified Transporter Genes

HTD supports a variety of ways to browse transporter genes, including the hierarchical classification and chromosome distribution. The classification page contains four parts of classification systems, including classification in HTD, TC classification, KEGG BRITE and Gene Ontology, which could be easily chosen by users (Figure 1C). The classification in HTD was mainly based on gene name, domain information and GO annotation, which might correspond to categories in the standard TC system. Each transporter under a specific category is linked to detailed information page of the gene. In addition, the genomic distribution of ten categories of HTGs in our HTD was plotted in 24 chromosomes with different colors (Figure 1E). Users can click on each cytoband in chromosome to access all the transporters in the region.

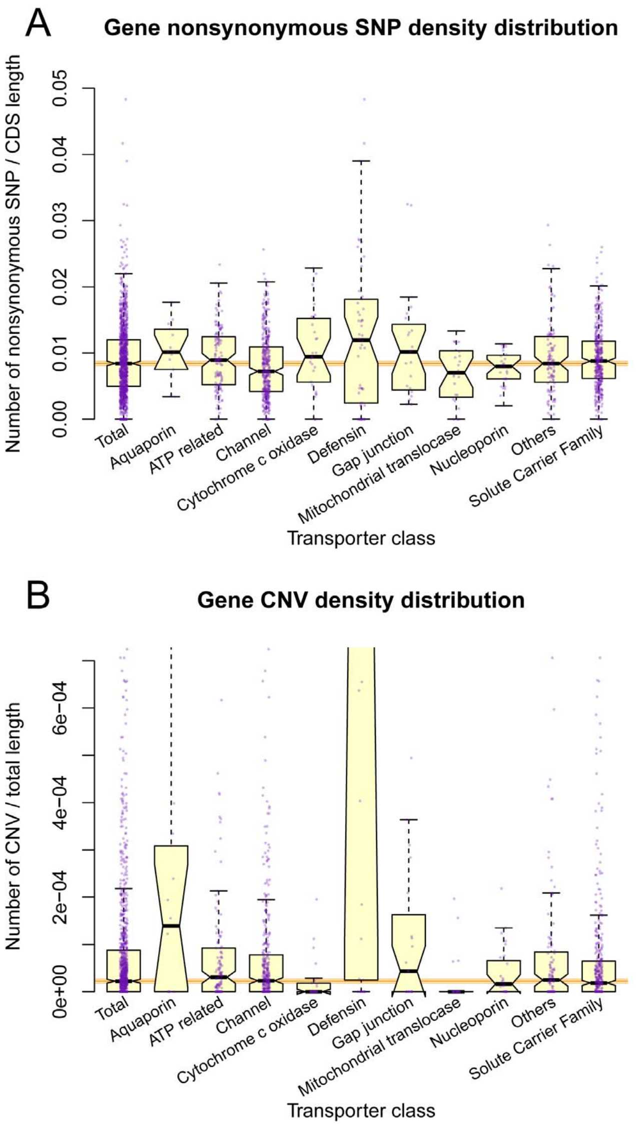


Figure 2. Distribution of nonsynonyous SNP and CNV density on ten categories of transporter genes in HTD. The x-axis shows the ten transporter categories, and y-axis shows the corresponding value: (A) the density of nonsynonymous SNPs, normalized by dividing CDS length, (B) the density of CNVs, normalized by dividing gene total length. Both subfigures are notched boxplot along with scattered real sample points in purple. The thick band inside the box is the median, and the bottom and top of the box are the first quantile (Q1) and the third quantile (Q3). The ends of the whiskers represents data within 1.5 *IQR (= Q3–Q1) from the lower quantile (Q1) or the upper quantile (Q3). The notch is always symmetric around the median, with deviation from median by 1.58 *IQR/sqrt(n), where n is the sample size. The notch approximately shows the confidence interval of median, so that if the notches of two boxes do not overlap, their medians are usually significantly different. Three horizontal orange lines behind the boxes show the median and notch range of the "Total" box.
doi:10.1371/journal.pone.0088883.g002

Table 2. Top ten enriched diseases of human transporter genes.

Disease	Database	P Value	Q Value*
Epilepsy	GAD	1.68E-14	7.06E-13
Sudden infant death syndrome	FunDO	5.26E-10	1.41E-08
Other nervous and sensory system diseases	KEGG DISEASE	2.53E-08	5.20E-07
Long QT syndrome	FunDO	8.40E-08	1.61E-06
Congenital disorders of ion transport and metabolism	KEGG DISEASE	3.09E-07	5.57E-06
Drug abuse	FunDO	3.52E-07	6.11E-06
Atrial fibrillation	KEGG DISEASE	4.23E-07	6.92E-06
Brugada syndrome (BRS)	KEGG DISEASE	4.23E-07	6.92E-06
Serum uric acid	GAD	8.96E-07	1.39E-05
Nervous system diseases	KEGG DISEASE	2.05E-06	2.88E-05

*Benjamini-Hochberg Corrected P Value.

doi:10.1371/journal.pone.0088883.t002

Keyword based Search of Transporter Proteins

A quick search box on the top right of each page was useful to search by transporter names or Entrez Gene IDs quickly. Advanced searches were constructed to query HTD by typing their gene name, accession number from NCBI and EBI gene and protein databases and their functional characteristics including chromosome location, interaction partner, biological process, and disease or drug (Figure 1B).

Sequence based Search of Transporter Proteins

In BLAST page, users can evaluate the transporters with input sequences. The homologs of input sequence are searched among the transporters in HTD using BLAST. The sequence alignment option can be modified with E-value and identity score. This database also provides bulk downloads of all nucleotide and protein sequences in a FASTA format for an advanced local sequence search (Figure 1D).

Comparison to Other Public Transporter Resources

Our HTD shares 944 transporters with the other four databases, which account for 60.7% in HTD and 80.1% for the total in other four databases. We have included 611 unique membrane HTGs (245 protein-coding genes and 366 pseudogenes) that were not found in previous transporter databases (representing 36.3% of our list). Among the 245 coding genes, there are 89 ion channels (including 9 anoctamins, a group of calcium activated chloride channels), 44 defensins, 20 nuclear pore proteins, 26 cytochrome c oxidases, 11 solute carrier family (SLC) proteins, 9 mitochondrial membrane translocases, 7 ATP-synthases, 5 aquaporins, 2 complement components (C8, and C9), 2 blood group-related proteins, and a scavenger receptor. Thus, our collection includes more ion channels, SLC family proteins and other proteins which are not included in other databases.

As shown in Figure S3, 234 genes from the other four databases were not included in our HTD. In detail, these genes contain 32 fatty acid binding proteins (including 8 lipocalins), 20 BCL family proteins, 17 annexins, 16 genes (including motor proteins) which transport other proteins, 15 GPCR proteins, 14 proteins from membrane-spanning 4-domains subfamily, 13 signal transduction kinases (including MAPK), 13 peroxisomal biogenesis factors, 11 apolipoproteins, 9 signal recognition particles, 5 claudins, 3 lysosomal-associated membrane proteins, 2 cells cytochrome b proteins, 2 odorant binding proteins, and 2 nucleoside kinases. The reasons that we do not include these proteins are as following: 1, not transmembrane transporters, but localizing to cytoplasm or plasma, such as apolipoproteins; 2, some proteins such as motor proteins, which are just associated with cytoplasmic vesicle transporting but not transmembrane transporting; 3, signal transduction proteins such as GPCRs and kinases, which do not participate the transmembrane transporting; 4, other proteins whose substrates locate on or in transmembrane.

To compare with TCDB, we downloaded all the human transporters from TCDB (http://www.tcdb.org/hgnc_explore.php) and did one by one gene symbol comparison. We found additional transporters that are not in TCDB, e.g. AQP3 and AQP7. If we include human pseudogene, there are 952 HTD unique entries. If we exclude pseudogene, there are still 579 HTD unique genes not including in TCDB. The complete mapping information between our HTD and TCDB can be found in our web site (http://htd.cbi.pku.edu.cn/download/htd2tcdb.xls). In addition, we also built the phylogenetic trees for all the categories based on our HTD classification system. All the multiple alignment results can be found in our updated web site (http://htd.cbi.pku.edu.cn/download/multiple_alignment_and_ML_tree.zip) that will help users to gain more insight for the evolutionary aspect of each transporter categories. Evolutionarily, HTD is complementary to TCDB.

Statistical Analyses on Expression, Variation, Function, Disease Profiles

Based on our collected heterogeneous data, we conducted systems biology data integration which might remove bias resulting from any single technology platform and provide additional insight into the genetic etiology not observed by any individual study [48,49]. The expression level changes of transporters could cause wide effects on compound and drug metabolism. In helping users to gain an overview for the gene expression pattern of a given transporter, we integrated publicly available gene expression profiling data of the transporters. Overall, the expression data integration was mainly based on ID mapping. The EST expression levels in different tissues were integrated from NCBI UniGene, which could be directly linked to NCBI Entrez Gene ID. Mouse brain region expression profiles were from Allen Brain Atlas [33], which were mapped to human Gene ID based on homology information from NCBI HomoloGene. The RNA-seq expression data was extracted from supplementary materials [34,35], which were followed by ID mapping from Ensembl Gene ID or UCSC Gene ID to Entrez Gene ID. Transporters with drug targets were reported more commonly to express in many tissues such as intestine, liver, kidney, and brain for drug absorption and excretion. Based on the brain gene expression data from Allen Brain Atlas, we compared the expression levels of transporters with non-transporter genes. We applied Fisher's exact test on a 2-by-2 contingency table counting the gene number of transporter or non-transporter genes with low (< = 5 in 100 scale) or high (>5 in 100 scale) expression level. In almost all brain regions, the proportion of mouse transporter genes with low expression level is significantly smaller than non-transporter genes. This indicated that transporter genes overall express higher than other genes in brain regions (Figure S4). Additionally, based on RNA-seq data for human tissues, we observed similar expression pattern in various brain regions when comparing to other tissues or cell lines (Figure S5) [34].

The genetic polymorphisms in transporters often have direct or adverse effects on the pharmacokinetics, drug-drug interactions, and personalized drug treatments [50]. The integration of genetics, disease, and drug information related to transporters provides an overview for the therapeutic safety and efficacy of drugs in various diseases. Based on population SNP information from dbSNP and HapMap, 1,279 genes (82.3%) from 1,555 human transporters overlapped 1,201,561 SNPs, in which 35,358 SNPs are exonic and 19,183 are nonsynonymous. When focusing on nonsynonymous SNPs, the HTGs from "Cytochrome c oxidase", "Defensin", and "Mitochondrial translocase" contained significantly less nonsynonymous SNPs in comparison with other transporter genes (Figure S6A). To control the potential influence of CDS length, which was shown different between categories (Figure S6B), we calculated the SNP density by dividing gene CDS length. After normalization, the average nonsynonymous SNP density for "Defensin" was marginally significantly higher than others (Wilcoxon rank sum test, *p*-value = 0.078), and "Channel" has lower SNP density (*p*-value = 2.5e-5) (Figure 2A). Copy-number variations (CNVs) refer a structure variation resulting gain or loss of copies of one or more sections of chromosome. Based on the integrated CNV data from DGV database, 855 genes (55.0%) from 1,555 human transporters were overlapped with known CNV regions. With the same analysis approach, after controlling gene total length (Figure S6C, D), CNV density was found significantly higher in "Defensin" (*p*-value = 7.0e-12), and lower in "Cytochrome c oxidase" (*p*-value = 3.1e-03) and "Mitochondrial translocase" (*p*-value = 1.5e-03) (Figure 2B). These results might suggest that "Defensin" genes were subjected to weaker negative selection than other transporter genes.

Further functional enrichment analyses showed that 1,555 HTGs were enriched in various cellular processes. Some of the highlights include oxidative phosphorylation, cardiac muscle contraction, Parkinson's disease, vibrio cholerae infection, mineral absorption, collecting duct acid secretion, synaptic vesicle cycle, ABC transporter, Alzheimer's disease, and bile secretion (all FDR corrected *p*-value <1e-16). Furthermore, we found that HTGs were mostly enriched in neural disease, drug abuse, and other metabolic disorders such as epilepsy, sudden infant death syndrome, long QT syndrome, and congenital disorders of ion transport and metabolism (Table 2). With manually integrated information according to OMIM, GAD, and MeSH, 215 HTGs were related to 21 diseases categories. There were 101 HTGs that are related to "nervous system diseases", 79 HTGs related to "congenital, hereditary, and neonatal diseases and abnormalities", 58 HTGs related to "nutritional and metabolic diseases", and 43 HTGs related to "cardiovascular diseases" (Figure S7A). To get an overview of natural polymorphisms on disease associated transporters we counted the number of non-synonymous mutations and CNVs on the HTGs and normalized by length. When comparing to all HTGs, most disease-related HTGs tended to have longer CDS length, among which HTGs related to "congenital, hereditary, and neonatal diseases and abnormalities" and "nutritional and metabolic diseases" were found to have significantly longer CDS (Wilcoxon rank sum test, *p*-value = 8.5e-9 and 2.6e-03) (Figure S7B). HTGs related to "bacterial infections and mycoses" and "respiratory tract diseases" were found to have significantly higher nonsynonymous SNP density (*p*-value = 7.3e-03 and 8.6e-03), while HTGs related to "mental disorders" tended to have lower nonsynonymous SNP density (*p*-value = 0.086) (Figure S7C). Most of 21 disease categories showed similar distribution on the density of CNVs involved with HTGs (Figure S7D).

Integrated Analyses on Variations and Drugs of Human Transporters

As HTGs may play the important roles in drug metabolism, we integrated pharmacogenetics and drug information from PharmGKB, CTD, and DrugBank, which was also mainly based on NCBI Gene ID mapping. These databases told about the relationship between a drug or chemical and a gene. Due to the statistical power on the number of related genes for a chemical, here we only showed the analysis results based on CTD annotation data. In Figure S8A, the numbers of related HTGs on top 30 types of chemicals with at least 60 related HTGs were plotted. The result indicated that HTGs are highly related to the drugs such as Acetaminophen (pain reliever), Phenobarbital (anticonvulsant), and Valproic Acid (anticonvulsant). We further calculated the nonsynonymous SNP density and CNV density for these groups of HTGs, and found that "Ozone"-related transporter genes had significantly longer CDS than other transporter genes (Wilcoxon rank sum test, *p*-value = 1.1e-03) (Figure S8B), and "Aflatoxin B1"-related transporter genes had significantly higher nonsynonymous SNP density than others (*p*-value = 7.8e-03) (Figure S8C). When checking about CNV, "Ozone"-related and "bisphenol A"-related HTGs tended to have higher CNV density, but not significantly (*p*-value = 0.124 and 0.314), while "Sodium Selenite"-related HTGs had lower CNV density, which was marginally significant (*p*-value = 0.057) (Figure S8D).

Conclusion and Future Direction

HTD is a comprehensive knowledge-base of Human Transporter resource with extensive pharmacogenetic and genomic annotations. HTD will aid personalized drug development in keeping pace with high-throughput NGS data related to transporters and be updated periodically. Additionally to those integration issues, new tools like literature mining on transporter substrate relationship will be developed to enhance specificity in Human Transporter annotations, and more convenient online analytic tools will be developed to assist online data visualization.

Supporting Information

Figure S1 Diagram of Internal Table Structure of HTD. The table names are shown in the cells with blue background. Those cells filled in red are the major cross-link fields between tables; the two filled in yellow (HomoloGene table) are cross-links for multiple species.
(PDF)

Figure S2 Data statistics based on TC system. The transporter genes were classified into TC system by BLAST and manually checked for three species. The amounts in TC system level 2 categories are shown in the graph (The level 2 categories are: 1.A: α-Type Channels; 1.B: β-Barrel Porins; 1.C: Pore-Forming Toxins (Proteins and Peptides); 1.F: Vesicle Fusion Pores; 1.H: Paracellular Channels; 2.A: Porters (uniporters, symporters, antiporters); 3.A: P-P-bond-hydrolysis-driven transporters; 3.D: Oxidoreduction-driven transporters; 4.C: Acyl CoA ligase-coupled transporters; 5.B: Transmembrane 1-electron transfer carriers; 8.A: Auxiliary transport proteins; 9.A: Recognized transporters of unknown biochemical mechanism; 9.B: Putative transport proteins).
(PDF)

Figure S3 A venn diagram comparison of human transporter genes in HTD with other four popular transporter databases.
(PDF)

Figure S4 Gene expression patterns in different brain regions. The expression patterns of human transporter genes (in red) were shown in different mouse brain areas along with all genes (in black) as background based on mouse brain region expression profiles described in Allen Brain Atlas data. The *p*-values from Fisher's exact tests demonstrate the decreased proportion of low expression level (scaled expression level 0~5) of transporter genes compared with all background genes.
(PDF)

Figure S5 The Gene expression patterns in different tissues. The expression patterns of human transporter genes (in red) were shown in different tissues along with all genes (in black) as background based on the data from one RNA-seq paper [34]. The *p*-values from Fisher's exact tests demonstrate the decreased proportion of low expression level (RPKM 0~5) of transporter genes compared with all background genes.
(PDF)

Figure S6 Distribution of SNP, CNV count and gene length on ten categories of transporter genes in HTD. The x-axis shows the ten transporter categories, and y-axis shows the corresponding value: (A) the number of nonsynonymous SNPs on gene CDS region, (B) gene CDS length, (C) the number of CNVs overlapping the total-length gene, (D) gene total length. All four subfigures are standard notched boxplot with scattered real sample points in purple. The thick band inside the box is the median, and the bottom and top of the box are the first quantile (Q1) and the third quantile (Q3). The ends of the whiskers represents data within 1.5 *IQR (= Q3–Q1) from the lower quantile (Q1) or the upper quantile (Q3). The notch is always symmetric around the median, with deviation from median by 1.58 *IQR/sqrt(n), where n is the sample size. The notch approximately shows the confidence interval of median, so that if the notches of two boxes do not overlap, their medians are usually significantly different. Three horizontal orange lines show the median and notch range of the "Total" box.
(PDF)

Figure S7 Distribution of gene count, gene length, SNP and CNV density on transporters related to different disease categories. The x-axis shows 21 disease categories, and y-axis shows the corresponding value: (A) the number of related genes for each disease category, (B) gene CDS length, (C) the density of nonsynonymous SNPs on gene CDS length, (D) the density of CNVs on gene total length. Except the first barplot shows the number of genes related to each disease category, the other three subfigures are standard notched boxplot with scattered real sample points in purple. Three horizontal orange lines show the median and notch range of the "Total" box. The meaning of notched boxplot representation is described in figure legends for Figure S6.
(PDF)

Figure S8 Distribution of gene count, gene length, SNP and CNV density on transporters related to some chemicals. The x-axis shows top 30 chemicals related with most transporter genes, and y-axis shows the corresponding value: (A) the number of related genes for each disease category, (B) gene CDS length, (C) the density of nonsynonymous SNPs on gene CDS length, (D) the density of CNVs on gene total length. Except the first barplot shows the number of genes related to a chemical, the other three subfigures are standard notched boxplot with scattered real sample points in purple. Three horizontal orange lines show the median and notch range of the "Total" box. The meaning of notched boxplot representation is described in figure legends for Figure S6.
(PDF)

Table S1 Curated keywords for literature searching and corresponding category in HTD.
(DOCX)

Table S2. Annotation entry statistics for 1555 human transporter genes.
(DOC)

Author Contributions

Conceived and designed the experiments: MZ HQ. Performed the experiments: AYY QL CL HQ. Analyzed the data: AYY HQ. Wrote the paper: AYY QL MZ HQ.

References

1. Nelson DL, Lehninger AL, Cox MM (2008) Lehninger principles of biochemistry. New York: W.H. Freeman.
2. Giacomini KM, Huang SM, Tweedie DJ, Benet LZ, Brouwer KL, et al. (2010) Membrane transporters in drug development. Nat Rev Drug Discov 9: 215–236.
3. Overington JP, Al-Lazikani B, Hopkins AL (2006) How many drug targets are there? Nat Rev Drug Discov 5: 993–996.
4. Rask-Andersen M, Almen MS, Schioth HB (2011) Trends in the exploitation of novel drug targets. Nat Rev Drug Discov 10: 579–590.
5. Ware JA (2006) Membrane transporters in drug discovery and development: A new mechanistic ADME era. Mol Pharm 3: 1–2.
6. Yan Q (2010) Membrane transporters and drug development: relevance to pharmacogenomics, nutrigenomics, epigenetics, and systems biology. Methods Mol Biol 637: 1–21.
7. Cardoso AR, Queliconi BB, Kowaltowski AJ (2010) Mitochondrial ion transport pathways: role in metabolic diseases. Biochim Biophys Acta 1797: 832–838.
8. Longo N, Ardon O, Vanzo R, Schwartz E, Pasquali M (2011) Disorders of creatine transport and metabolism. Am J Med Genet C Semin Med Genet 157: 72–78.
9. Ueda K (2011) ABC proteins protect the human body and maintain optimal health. Biosci Biotechnol Biochem 75: 401–409.
10. van Veen HW (2010) Structural biology: Last of the multidrug transporters. Nature 467: 926–927.
11. Zhao FQ, Keating AF (2007) Functional properties and genomics of glucose transporters. Curr Genomics 8: 113–128.
12. Huang Y (2007) Pharmacogenetics/genomics of membrane transporters in cancer chemotherapy. Cancer Metastasis Rev 26: 183–201.
13. Saier MH Jr, Tran CV, Barabote RD (2006) TCDB: the Transporter Classification Database for membrane transport protein analyses and information. Nucleic Acids Res 34: D181–186.
14. Ren Q, Chen K, Paulsen IT (2007) TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. Nucleic Acids Res 35: D274–279.

15. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res 38: D355–360.
16. Yan Q, Sadee W (2000) Human membrane transporter database: a Web-accessible relational database for drug transport studies and pharmacogenomics. AAPS PharmSci 2: E20.
17. Zhao M, Chen Y, Qu D, Qu H (2011) TSdb: a database of transporter substrates linking metabolic pathways and transporter systems on a genome scale via their shared substrates. Science China-Life Sciences 54: 60–64.
18. Fichant G, Basse MJ, Quentin Y (2006) ABCdb: an online resource for ABC transporter repertories from sequenced archaeal and bacterial genomes. FEMS Microbiol Lett 256: 333–339.
19. Miao Z, Li D, Zhang Z, Dong J, Su Z, Wang T (2012) Medicago truncatula transporter database: a comprehensive database resource for M. truncatula transporters. BMC Genomics. 13: 60.
20. Hodges LM, Markova SM, Chinn LW, Gow JM, Kroetz DL, et al. (2011) Very important pharmacogene summary: ABCB1 (MDR1, P-glycoprotein). Pharmacogenet Genomics 21: 152–161.
21. Oshiro C, Mangravite L, Klein T, Altman R (2010) PharmGKB very important pharmacogene: SLCO1B1. Pharmacogenetics and Genomics 20: 211–216.
22. Li CY, Yu Q, Ye ZQ, Sun Y, He Q, et al. (2007) A nonsynonymous SNP in human cytosolic sialidase in a small Asian population results in reduced enzyme activity: potential link with severe adverse reactions to oseltamivir. Cell Res 17: 357–362.
23. Hwang D, Rust AG, Ramsey S, Smith JJ, Leslie DM, et al. (2005) A data integration methodology for systems biology. Proc Natl Acad Sci U S A 102: 17296–17301.
24. Kalyana-Sundaram S, Kumar-Sinha C, Shankar S, Robinson DR, Wu YM, et al. (2012) Expressed pseudogenes in the transcriptional landscape of human cancers. Cell 149: 1622–1634.
25. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, et al. (2011) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 39: D38–51.
26. UniProt_Consortium (2010) The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res 38: D142–148.
27. Burge S, Kelly E, Lonsdale D, Mutowo-Muellenet P, McAnulla C, et al. (2012) Manual GO annotation of predictive protein signatures: the InterPro approach to GO curation. Database (Oxford) 2012: bar068.
28. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, et al. (2012) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 40: D13–25.
29. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, et al. (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29: 308–311.
30. International HapMap Consortium. (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467(7311): 52–58.
31. Zhang J, Feuk L, Duggan GE, Khaja R, Scherer SW (2006) Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. Cytogenet Genome Res, 115, 205–214.
32. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, et al. (2009) Human Protein Reference Database–2009 update. Nucleic Acids Res, 37, D767–772.
33. Jones AR, Overly CC, Sunkin SM (2009) The Allen Brain Atlas: 5 years and beyond. Nat Rev Neurosci 10: 821–828.
34. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, et al. (2008) Alternative isoform regulation in human tissue transcriptomes. Nature 456: 470–476.
35. Wu JQ, Habegger L, Noisa P, Szekely A, Qiu C, et al. (2010) Dynamic transcriptomes during neural differentiation of human embryonic stem cells revealed by short, long, and paired-end sequencing. Proc Natl Acad Sci U S A 107: 5254–5259.
36. Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, et al. (2014) The UCSC Genome Browser database: 2014 update. Nucleic Acids Res 42: D764–770.
37. Relling MV, Gardner EE, Sandborn WJ, Schmiegelow K, Pui CH, et al. (2011) Clinical Pharmacogenetics Implementation Consortium guidelines for thiopurine methyltransferase genotype and thiopurine dosing. Clin Pharmacol Ther 89: 387–391.
38. Davis AP, King BL, Mockus S, Murphy CG, Saraceni-Richards C, et al. (2011) The Comparative Toxicogenomics Database: update 2011. Nucleic Acids Res 39: D1067–1072.
39. Knox C, Law V, Jewison T, Liu P, Ly S, et al. (2011) DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. Nucleic Acids Res 39: D1035–1041.
40. Xie C, Mao X, Huang J, Ding Y, Wu J, et al. (2011) KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. Nucleic Acids Res, 39, W316–322.
41. Zhao M, Qu H (2010) PathLocdb: a comprehensive database for the subcellular localization of metabolic pathways and its application to multiple localization analysis. BMC Genomics 11 Suppl 4: S13.
42. Zhao M, Chen X, Gao G, Tao L, Wei L (2009) RLEdb: a database of rate-limiting enzymes and their regulation in human, rat, mouse, yeast and E. coli. Cell Res 19: 793–795.
43. Kong L, Cheng L, Fan LY, Zhao M, Qu H (2013) IQdb: an intelligence quotient score-associated gene resource for human intelligence. Database (Oxford) 2013: bat063.
44. Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38: e164.
45. Ng PC, Henikoff S (2003) SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 31: 3812–3814.
46. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, et al. (2010) A method and server for predicting damaging missense mutations. Nat Methods 7: 248–249.
47. Liu X, Jian X, Boerwinkle E (2011) dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Hum Mutat 32: 894–899.
48. Zhao M, Qu H (2011) High similarity of phylogenetic profiles of rate-limiting enzymes with inhibitory relation in Human, Mouse, Rat, budding Yeast and E. coli. BMC Genomics 12 Suppl 3: S10.
49. Zhao M, Qu H (2009) Human liver rate-limiting enzymes influence metabolic flux via branch points and inhibitors. BMC Genomics 10 Suppl 3: S31.
50. Takahashi N, Miura M, Scott SA, Kagaya H, Kameoka Y, et al. (2010) Influence of CYP3A5 and drug transporter polymorphisms on imatinib trough concentration and clinical response among patients with chronic phase chronic myeloid leukemia. Journal of Human Genetics 55: 731–737.