Published online 29 November 2016 *Nucleic Acids Research, 2017, Vol. 45, Database issue* ***D353–D361***
doi: 10.1093/nar/gkw1092

KEGG: new perspectives on genomes, pathways, diseases and drugs

Minoru Kanehisa[1,*], Miho Furumichi[1], Mao Tanabe[1], Yoko Sato[2] and Kanae Morishima[1]

[1]Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan and [2]Healthcare Solutions Department, Fujitsu Kyushu Systems Ltd., Hakata-ku, Fukuoka 812-0007, Japan

Received October 09, 2016; Editorial Decision October 25, 2016; Accepted October 26, 2016

ABSTRACT

KEGG (http://www.kegg.jp/ or http://www.genome.jp/kegg/) is an encyclopedia of genes and genomes. Assigning functional meanings to genes and genomes both at the molecular and higher levels is the primary objective of the KEGG database project. Molecular-level functions are stored in the KO (KEGG Orthology) database, where each KO is defined as a functional ortholog of genes and proteins. Higher-level functions are represented by networks of molecular interactions, reactions and relations in the forms of KEGG pathway maps, BRITE hierarchies and KEGG modules. In the past the KO database was developed for the purpose of defining nodes of molecular networks, but now the content has been expanded and the quality improved irrespective of whether or not the KOs appear in the three molecular network databases. The newly introduced addendum category of the GENES database is a collection of individual proteins whose functions are experimentally characterized and from which an increasing number of KOs are defined. Furthermore, the DISEASE and DRUG databases have been improved by systematic analysis of drug labels for better integration of diseases and drugs with the KEGG molecular networks. KEGG is moving towards becoming a comprehensive knowledge base for both functional interpretation and practical application of genomic information.

INTRODUCTION

In 1995, KEGG (Kyoto Encyclopedia of Genes and Genomes) was originally developed as an integrated database resource for biological interpretation of completely sequenced genomes by KEGG pathway mapping, the procedure to map genes in the genome to manually created pathway maps. At that time, KEGG consisted of only four databases, PATHWAY, GENES, COMPOUND and ENZYME and KEGG pathway mapping was performed through ENZYME because the database contained only metabolic pathway maps. KEGG was later significantly expanded, PATHWAY supplemented by BRITE and MODULE, GENES expanded with GENOME, COMPOUND supplemented by GLYCAN and REACTION, and ENZYME replaced by KO for the role of KEGG pathway mapping. KEGG also became more widely used for analyzing not only genomics data but also transcriptomics, proteomics, glycomics, metabolomics, metagenomics and other high-throughput data.

After more than 20 years, we wish to make KEGG a more comprehensive knowledge base for assisting biological interpretations of large-scale molecular datasets. In the past, our efforts focused on developing the databases for higher-level functions, the PATHWAY, BRITE and MODULE databases, and KOs were defined as network nodes of these databases. Consequently, the content of molecular-level functions in the KO database was incomplete. This is no longer the case. We have started expending major efforts to improve and expand the KO database. First, existing KOs are linked to experimentally characterized protein sequence data with proper reference information. Second, published reports on characterizing protein functions are identified, sequence data are registered in the addendum category of the KEGG GENES database (1), and new KOs are defined accordingly.

The KEGG DRUG and DISEASE databases were released in 2005 and 2008, respectively, and the KEGG MEDICUS resource (2) integrating these databases with drug labels (package inserts) was initiated in 2009. While the content of KEGG is derived mostly from published research articles, drug labels and other regulatory documents used in society are now also examined. For drug labels, the entire content is systematically analyzed, for example, to characterize drug–drug interactions associated with contraindications and to define drug-disease links that are meaningful in practice. The analysis results are used in the development of DISEASE, DRUG, PATHWAY and other databases. This paper describes these new developments of the KEGG database resource.

*To whom correspondence should be addressed. Tel: +81 774 38 4521; Fax: +81 774 38 3269; Email: kanehisa@kuicr.kyoto-u.ac.jp

OVERVIEW AND NEW FEATURES

KEGG databases

KEGG is an integrated database resource consisting of fifteen manually curated databases and a computationally generated database in four categories as shown in Table 1. The databases in the systems information category are PATHWAY, BRITE and MODULE, which constitute the reference knowledge base for understanding higher-level systemic functions of the cell and the organism, including metabolism, other cellular processes, organismal functions and human diseases. The KO database in the genomic information category is another unique database, in which knowledge of molecular-level functions is organized with the concept of functional orthologs. This category also contains GENOME and GENES, which are derived from RefSeq (3), Genbank (4) and NCBI Taxonomy (5) databases and given KEGG original annotations. SSDB is an auxiliary, computationally generated database used for KO-based annotation of the GENES database.

The databases in the chemical information category are COMPOUND, GLYCAN, REACTION, RCLASS and ENZYME, which are collectively called KEGG LIGAND. ENZYME is taken from the enzyme nomenclature database ExplorEnz (6) and given additional annotation of sequence data links (1). The health information category consists of DISEASE, DRUG, DGROUP and ENVIRON shown in Table 1 and two outside databases for drug labels: Japanese drug labels obtained from the JAPIC database and FDA drug labels obtained from the DailyMed database. KEGG MEDICUS represents the health information category integrated with these drug labels.

As of 1 October 2016 the following changes were made to KEGG in order to simplify its overall architecture without the loss of content. First, the RPAIR database for reactant pairs (7) was discontinued, but its main content for defining reaction classes (8) remains to be available as pairs of COMPOUND entries. Second, the DGENES database (7) for eukaryotic draft genomes was merged into the GENES database. Third, the plasmids category (1) of the GENES database was discontinued because plasmid genes could be, whenever necessary, incorporated in the addendum category.

KEGG identifiers and WebLinks

KEGG objects are biological entities from molecular to higher levels that are represented as database entries in KEGG. They include concrete objects of genes and protein, small molecules, reactions, pathways, diseases and drugs, organisms and viruses, as well as the more conceptual objects: KOs, modules and BRITE hierarchies. The KEGG object identifier or simply the KEGG identifier is the unique identifier for each KEGG object, which is also the database entry identifier in KEGG (Table 1). In most KEGG databases, the identifier takes the form of a prefix and a five-digit number and is called, for example, map number, K number, C number and D number for PATHWAY, KO, COMPOUND and DRUG, respectively. Exceptions are GENES identifiers and the EC number for ENZYME. Table 2 shows KEGG WebLinks, a convenient way to retrieve from or link to the KEGG website using KEGG identifiers.

Sequence data in KEGG

The GENES database now consists of three categories, KEGG organisms, viruses and addendum, as shown in Table 3. The KEGG organisms category is the main part of GENES consisting of completely or almost completely sequenced genomes taken from RefSeq and GenBank databases. Each organism is identified by the three- or four-letter organism code, such as 'hsa' for *Homo sapiens*. The viruses category is generated from the bi-monthly release of RefSeq, where individual viral genomes are not distinguished and the category is identified by the two-letter code 'vg'. Viruses are distinguished, when necessary, by the NCBI taxonomy identifier, which often corresponds to multiple sequenced genomes. The addendum category is a manually created sequence dataset for functionally characterized proteins. It is a publication-based collection of sequences, mostly author submitted sequences with GenBank/ENA/DDBJ protein accession numbers (ProteinIDs). This category is identified by the two-letter code 'ag'. Using these conventions each GENES entry is identified by the form of 'org:gene' (Table 1) where 'org' is the organism code (or the category code) and 'gene' is GeneID, Locus_tag or ProteinID depending on the data source.

Architecture of KEGG website

The architecture of KEGG website, both at the KEGG main site and the GenomeNet mirror site, was updated on 1 October 2016 and is shown in Table 4. The basic architecture is unchanged. The KEGG home page is directly linked to main databases and software tools and the other top layer pages contain description of and guidance to KEGG. The KEGG2 page in the second layer is the table of contents listing all KEGG databases and software tools including those available as GenomeNet extension databases (MGENOME, MGENES, EXPRESSION and OC) and GenomeNet bioinformatics tools (BLAST/FASTA, KAAS, SIMPCOMP, etc.). The second layer is a collection of entry points to KEGG: data-oriented entry points that correspond to Table 1, newly introduced subject-oriented entry points shown in Table 5, and organism-specific entry points. Since KEGG is a general purpose database covering many different types of data, it is expected that subject-oriented entry points, of which we may add more, will help the users in specific research areas to better understand and utilize KEGG.

MOLECULAR-LEVEL FUNCTIONS

Improvement of GENES and KO databases

In protein sequence databases, such as UniProt (9), functional annotation is included in individual sequence entries. In contrast, the KEGG GENES database is annotated by simply assigning KO identifiers (K numbers) and making links to the KO database. The definition field of each GENES entry contains the data source name in parentheses, such as (RefSeq) and (GenBank), indicating that the

Table 1. The KEGG databases

Category	Database name	Content	KEGG identifier
Systems Information	KEGG PATHWAY	KEGG pathway maps	Map number
	KEGG BRITE	BRITE hierarchies and tables	br/ko number
	KEGG MODULE	KEGG modules	M number
Genomic Information	KEGG ORTHOLOGY (KO)	KO groups for functional orthologs	K number
	KEGG GENOME	KEGG organisms (complete genomes) and selected viruses	org code / T number
	KEGG GENES	Gene catalogs of KEGG organisms, viruses, and addendum category	org:gene
	KEGG SSDB	Sequence similarity among GENES entries (computationally generated)	
Chemical Information (KEGG LIGAND)	KEGG COMPOUND	Metabolites and other small molecules	C number
	KEGG GLYCAN	Glycans	G number
	KEGG REACTION	Biochemical reactions	R number
	KEGG RCLASS	Reaction class	RC number
	KEGG ENZYME	Enzyme nomenclature	EC number
Health Information (KEGG MEDICUS[a])	KEGG DISEASE	Human diseases	H number
	KEGG DRUG	Drugs	D number
	KEGG DGROUP	Drug groups	DG number
	KEGG ENVIRON	Crude drugs and health-related substances	E number

[a]KEGG MEDICUS also includes Japanese drug labels obtained from the JAPIC database (http://www.japic.or.jp) and FDA drug labels obtained from the DailyMed database (http://dailymed.nlm.nih.gov).

Table 2. KEGG WebLinks

Database	URL form	Example
PATHWAY	www.kegg.jp/pathway/<map number>	www.kegg.jp/pathway/hsa01521
BRITE (hierachies only)	www.kegg.jp/brite/<br/ko number>	www.kegg.jp/brite/ko01504
		www.kegg.jp/brite/br08303_ndc
MODULE	www.kegg.jp/module/<M number>	www.kegg.jp/module/M00810
All databases except BRITE	www.kegg.jp/entry/<KEGG identifer>	www.kegg.jp/entry/K19188
		www.kegg.jp/entry/ag:CAD47941
		www.kegg.jp/entry/3.7.1.19

Table 3. Seqeunce data collection in KEGG GENES

Category	Sequence data	Primary data source	Gene identifier[a]
KEGG organisms	Genes in complete eukaryotic genomes	RefSeq	org:geneid
	Genes in complete prokaryotic genomes	RefSeq (reference genomes), GenBank (other genomes)	org:locus_tag
Viruses	Genes in RefSeq Virus collection	RefSeq	vg:geneid
Addendum	Functionally characterized proteins	PubMed	ag:proteinid

[a]org, three- or four-letter KEGG organism code; vg, Viruses category code; ag, Addendum category code

Table 4. Architecture of KEGG website

Layer	Content
Top pages	KEGG home (www.kegg.jp)
	Release notes, statistics, database/software documents, KEGG API, KGML
DB entry points	KEGG2 page for table of contents (www.kegg.jp/kegg/kegg2.html)
	Data-oriented entry points (corresponding to Table 1)
	Subject-oriented entry points (shown in Table 5)
	Organism-specific entry points (for individual genome, multiple genomes, pangenome, organism group)
DB contents	Database entries (as those shown in Table 2)
Software tools[a]	KEGG Mapper tools (www.kegg.jp/kegg/mapper.html)
	BlastKOALA automatic annotation server (www.kegg.jp/blastkoala/)
	GhostKOALA automatic annotation server (www.kegg.jp/ghostkoala/)

[a]Software tools in the KEGG main site excluding those at the GenomeNet mirror site.

Table 5. Subject-oriented entry points to KEGG

Database name	Subject
KEGG Cancer	Cancer research
KEGG Pathogen	Infectious diseases, pathogens and antimicrobial resistance
KEGG Virus	Virus research
KEGG Plant	Plant research
KEGG Glycan	Glycobiology research
KEGG Annotation	KO annotation of genes and proteins
KEGG RModule	Architecture of metabolic network

definition was given by the original database as shown in Figure 1A. KEGG original annotation is given in the following KO subfield, K19188 in this case. As shown in Figure 2, the KO entry K19188 is based on the experimentally characterized protein sequence CAD47941, which is stored in the addendum category of the GENES database (Figure 1B). Among over 20 000 KO entries, about 80% contain references (PubMed links) and about 60% contain sequence data links (in October 2016). Sequence data links are given to the references when sequences submitted by the authors to GenBank/ENA/DDBJ are identified or when sequences used in the experiments can be retrieved by cited references or sequence accessions.

The addendum category has made it possible to define KOs independently from KEGG organisms (complete genomes) and has filled the gaps of missing sequence links

A

Nocardioides sp. JS614: Noca_0613

Entry	Noca_0613 CDS T00443
Definition	(GenBank) protein of unknown function DUF1100, hydrolase family protein
KO	K19188 2,6-dihydroxypseudooxynicotine hydrolase [EC:3.7.1.19]
Organism	nca Nocardioides sp. JS614

B

Addendum: CAD47941

Help

Entry	CAD47941 CDS T10000
Gene name	pnh
Definition	(KEGG) 2,6-dihydroxypseudooxynicotine hydrolase (EC:3.7.1.19)
KO	K19188 2,6-dihydroxypseudooxynicotine hydrolase [EC:3.7.1.19]
Taxonomy	TAX:29320
Lineage	Bacteria; Actinobacteria; Micrococcales; Micrococcaceae; Paenarthrobacter
Organism	ag Addendum (Paenarthrobacter nicotinovorans)

Figure 1. Definition and KO fields of GENES entries. (**A**) GenBank derived entry nca:Noca_0613 and (**B**) manually created entry ag:CAD47941 in the addendum category.

ORTHOLOGY: K19188

Entry	K19188 KO
Name	dhponh
Definition	2,6-dihydroxypseudooxynicotine hydrolase [EC:3.7.1.19]
Pathway	ko00760 Nicotinate and nicotinamide metabolism
Module	M00810 Nicotine degradation, pyridine pathway, nicotine => 2,6-dihydroxypyridine/succinate semialdehyde
Brite	KEGG Orthology (KO) [BR:ko00001] Metabolism Metabolism of cofactors and vitamins 00760 Nicotinate and nicotinamide metabolism K19188 dhponh; 2,6-dihydroxypseudooxynicotine hydrolase KEGG modules [BR:ko00002] Pathway module Secondary metabolism Aromatics degradation M00810 Nicotine degradation, pyridine pathway, nicotine => s K19188 E3.7.1.19; 2,6-dihydroxypseudooxynicotine hydrolase Enzymes [BR:ko01000] 3. Hydrolases 3.7 Acting on carbon-carbon bonds 3.7.1 In ketonic substances 3.7.1.19 2,6-dihydroxypseudooxynicotine hydrolase K19188 dhponh; 2,6-dihydroxypseudooxynicotine hydrolase BRITE hierarchy
Other DBs	RN: R07515 GO: 0034948
Genes	ROP: ROP_27470 ROA: Pd630_LPD07472 NCA: Noca_0613 AG: CAD47941(pnh) Taxonomy KOALA UniProt
Reference	PMID:16321959
Authors	Sachelaru P, Schiltz E, Igloi GL, Brandsch R.
Title	An alpha/beta-fold C--C bond hydrolase is involved in a central step of nicotine catabolism by Arthrobacter nicotinovorans.
Journal	J Bacteriol 187:8516-9 (2005)
Sequence	[ag:CAD47941]
LinkDB	All DBs

Figure 2. KO entry K19188 defined from an experimentally characterized protein sequence, ag:CAD47941 for EC:3.7.1.19.

in the KEGG pathway maps. It is now used to associate sequence data to Enzyme Nomenclature and to create various sequence data collections, including antimicrobial resistance genes (1) and cytochrome P450s. The EC number list of Enzyme Nomenclature contains references reporting experimental characterization of enzymatic reactions. Unfortunately, however, it does not contain any enzyme sequence information, and links between sequences and EC numbers are left to interpretation of individual users and database developers. The KEGG ENZYME database is now annotated, whenever possible, with sequence data links given by the same criteria mentioned above.

KO analysis tools

KOs are defined as sequence similarity groups as well as functional orthologs, so that they can be used for sequence similarity based KO assignment. There is no pre-defined threshold of similarity scores, and although the term ortholog is used a KO may consist of a single gene or may contain only genes from closely related species. When defining KOs three factors are considered whenever available, pathways, gene clusters and phylogeny, as illustrated in Figure 3 for K19188.

Figure 3A shows a part of the nicotinate and nicotinamide metabolism pathway (map00760) that corresponds to the nicotine degradation pyridine pathway module (M00810) for Nocardioides sp. JS614 (nca), where the gene Noca_0613 (Figure 1A) for K19188 with EC:3.7.1.19 is marked red. Figure 3B is the ortholog table for M00810, which indicates by coloring that this gene is adjacent on the chromosome to the gene for K19187 for a subunit of EC:1.5.99.14. The fact that the entire set of genes for M00810 is present in the Nocardioides genome and that the genes that are located next to each other on the chromosome are mapped to two consecutive reaction steps on the pathway confirms correctness of assigning K19188 to nca:Noca_0613. The pathway and/or gene cluster information is not always available, but the phylogenetic information can always be considered because the GENES database is organized according to the NCBI taxonomy and similarity scores of all gene pairs are stored in the SSDB database. Figure 3C shows a dendrogram, which was obtained by hierarchical clustering of ag:CAD47941 (Figure 1B), the core sequence of defining K19188, and its neighbors in SSDB, and by coloring of branches according to the assigned K number. The dendrogram tool now linked from the SSDB search result page allows coloring of branches for distinguishing multiple K numbers assigned. It is also used, for example, for viewing phylogenetic relationships and K number assignments of antimicrobial resistance genes in KEGG Pathogen (Table 5).

BlastKOALA tools

BlastKOALA (10) is the web server for automatic annotation (KO assignment) of query amino acid sequences followed by KEGG Mapper analysis for inferring higher-level functions. The server makes full use of the improvements made for the GENES and KO databases, including the addition of the GENES addendum category, the precise taxonomic classification of GENES data and the improvement of KO to sequence links. The taxonomic classification was used to define 'non-redundant' GENES datasets for BlastKOALA. A non-redundant dataset is a collection of pangenomes created at the species or genus level for prokaryotes and at the genus or family level for eukaryotes by removing redundant sequences but retaining functional contents (KO contents) within each pangenome (1,10). The user may select a combination of non-redundant pangenome collections considering the execution time (faster for higher taxonomic ranks) and the annotation quality (better for lower taxonomic ranks).

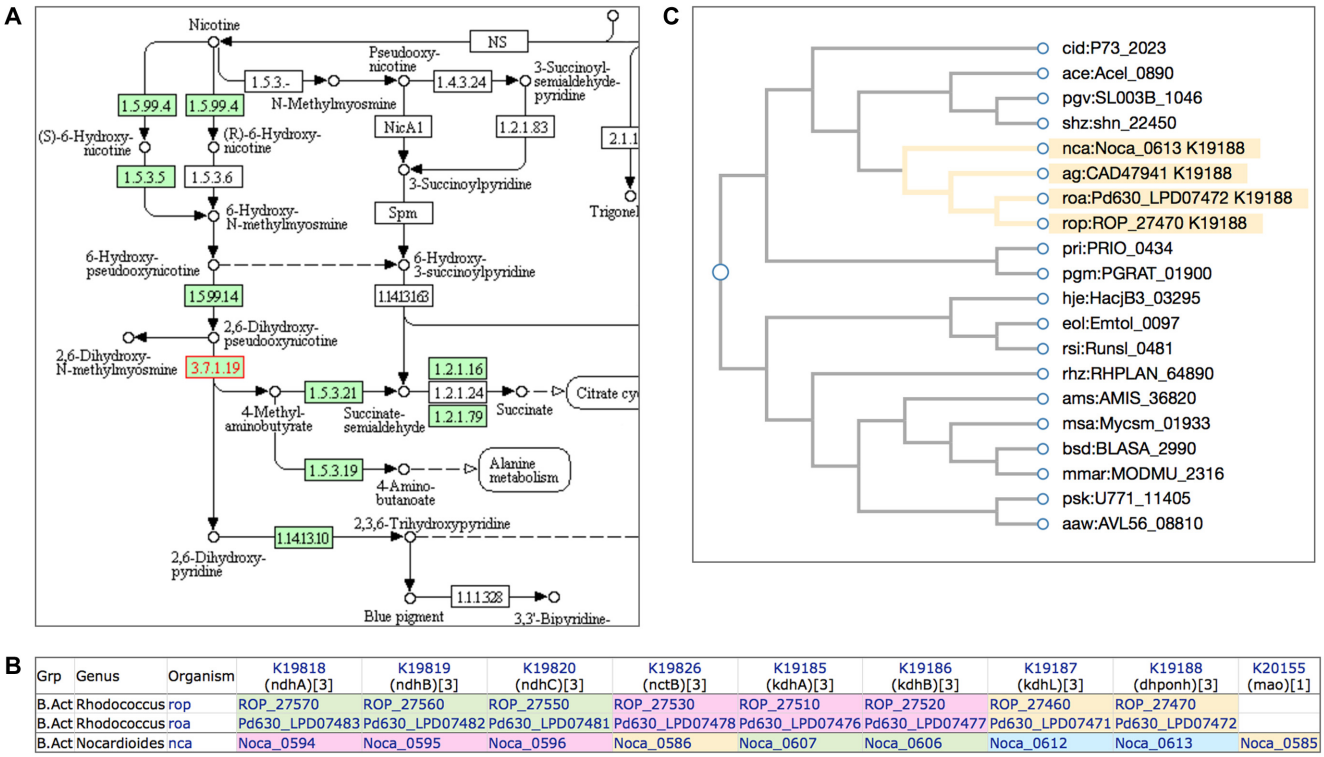
There are three variants of BlastKOALA. The Annotate Sequence tool in KEGG Mapper is an interactive version of BlastKOALA, which is suitable when closely related genomes are already present in KEGG because the database to be searched is limited to a single pangenome. The Pathogen Checker tool in KEGG Pathogen allows search against the dataset of antimicrobial resistance genes in KEGG and interpretation of drugs or drug groups to be affected. The third one is the GhostKOALA server using more rapid GHOSTX (11) rather than BLAST and is suitable for analysis of large-scale datasets such as metagenomes (10).

HIGHER-LEVEL FUNCTIONS

Improvement of PATHWAY database

The KEGG PATHWAY database has been and will continue to be the main database in KEGG. It consists of manually drawn reference pathway maps together with organism-specific pathway maps that are computationally generated by matching KO assignments in the genome with reference pathway maps. In regular maps rectangular nodes are linked to KOs and the matching is shown by coloring. For example, the organism-specific glycolysis pathway hsa00010 with green coloring is generated from the hsa (*Homo sapiens*) genome and the reference pathway map00010. Table 6 shows the list of manually drawn pathway maps by category. Chemical structure transformation maps and drug structure maps are special maps showing relationships between chemical/drug structures and are not subject to genome-based expansion. Global maps (map numbers 01100s) and overview maps (map numbers 01200s) are intended to show global and overall features of metabolism. No rectangles are used and lines connecting circles (chemical compounds) are linked to KOs in the reference pathway maps and colored in the organism-specific pathway maps. An additional feature of these maps is that KEGG modules and reaction modules (12) are embedded in the pathway diagrams.

On average one or two new maps are released every month, and existing maps are constantly updated. As indicated in Table 6 human diseases have become one of the major categories of the KEGG PATHWAY database. Recent additions include the subcategories of drug resistance for both antimicrobial and antineoplastic agents. For example, Figure 4A is a disease pathway map for non-small cell lung cancer (hsa05223), where known genetic alterations are marked in red. Targeted drug names are given on this map: Gefitinib (D01977) and Erlotinib (D04023) for EGFR mutation and Crizotinib (D09731) for ALK fusion gene.



B

Grp	Genus	Organism	K19818 (ndhA)[3]	K19819 (ndhB)[3]	K19820 (ndhC)[3]	K19826 (nctB)[3]	K19185 (kdhA)[3]	K19186 (kdhB)[3]	K19187 (kdhL)[3]	K19188 (dhponh)[3]	K20155 (mao)[1]
B.Act	Rhodococcus	rop	ROP_27570	ROP_27560	ROP_27550	ROP_27530	ROP_27510	ROP_27520	ROP_27460	ROP_27470	
B.Act	Rhodococcus	roa	Pd630_LPD07483	Pd630_LPD07482	Pd630_LPD07481	Pd630_LPD07478	Pd630_LPD07476	Pd630_LPD07477	Pd630_LPD07471	Pd630_LPD07472	
B.Act	Nocardioides	nca	Noca_0594	Noca_0595	Noca_0596	Noca_0586	Noca_0607	Noca_0606	Noca_0612	Noca_0613	Noca_0585

Figure 3. (**A**) A portion of the KEGG pathway map of nicotinate and nicotinamide metabolism (map00760) for Nocardioides sp. JS614 (nca), where Noca_0613 for EC:3.7.1.19 is marked red. (**B**) Ortholog table for the KEGG module of nicotine degradation pyridine pathway (M00810), where coloring of positional correlation indicates that Noca_0613 for EC:3.7.1.19 is adjacent to Noca_0612 for EC:1.5.99.14. (**C**) Dendrogram obtained by hierarchical clustering of ag:CAD47941 and its neighbors in SSDB, where genes annotated with K19188 are distinguished by coloring.

Table 6. Manually drawn KEGG reference pathway maps

Category	Type	Number of maps[a]
Metabolism	Global map	4
	Overview map	5
	Regular map	160
	Chemical structure transformation map	9
Genetic information processing	Regular map	22
Environmental information processing	Regular map	38
Cellular processes	Regular map	24
Organismal systems	Regular map	78
Human diseases	Regular map	81
Drug development	Drug structure map	75

[a]As of 1 October 2016.

Figure 4B is a drug resistance map for EGFR tyrosine kinase inhibitor resistance (hsa01521) caused by a secondary mutation in EGFR or activation of alternative pathways such as involving MET. A newer generation of EGFR inhibitor, Osimertinib (D10766), is for the former type of resistance.

KEGG mapper tools

KEGG Mapper is a collection of KEGG mapping tools for linking molecular objects (genes, proteins, metabolites and glycans) to higher-level objects (pathways, modules, hierarchies, taxonomy and diseases). Table 7 shows the current list of KEGG mapping tools including the new Search Disease tool. Two pathway mapping tools, Search Pathway and Search&Color Pathway, were made available from the beginning of the KEGG project, and they are still the most widely used. The mapping of genes and proteins can be made either in reference mode using KOs or in organism-specific mode using gene identifiers, such as human gene identifiers used to search against human pathway maps. The three tools, Reconstruct Pathway, Reconstruct Brite and Reconstruct Module, accept KOs only. They are linked from the BlastKOALA and GhostKOALA servers for interpretation of the KO assignment results.

The tools for characterizing molecular-level functions (BlastKOALA and GhostKOALA) and higher-level functions (KEGG Mapper tools) have been developed separately as general-purpose tools. In contrast the Pathogen Checker tool, a BlastKOALA variant, contains a focused KEGG mapping feature against a specific BRITE table file (br08451) for interpretation of antimicrobial resistance. We plan to make it easier to select and integrate appropriate

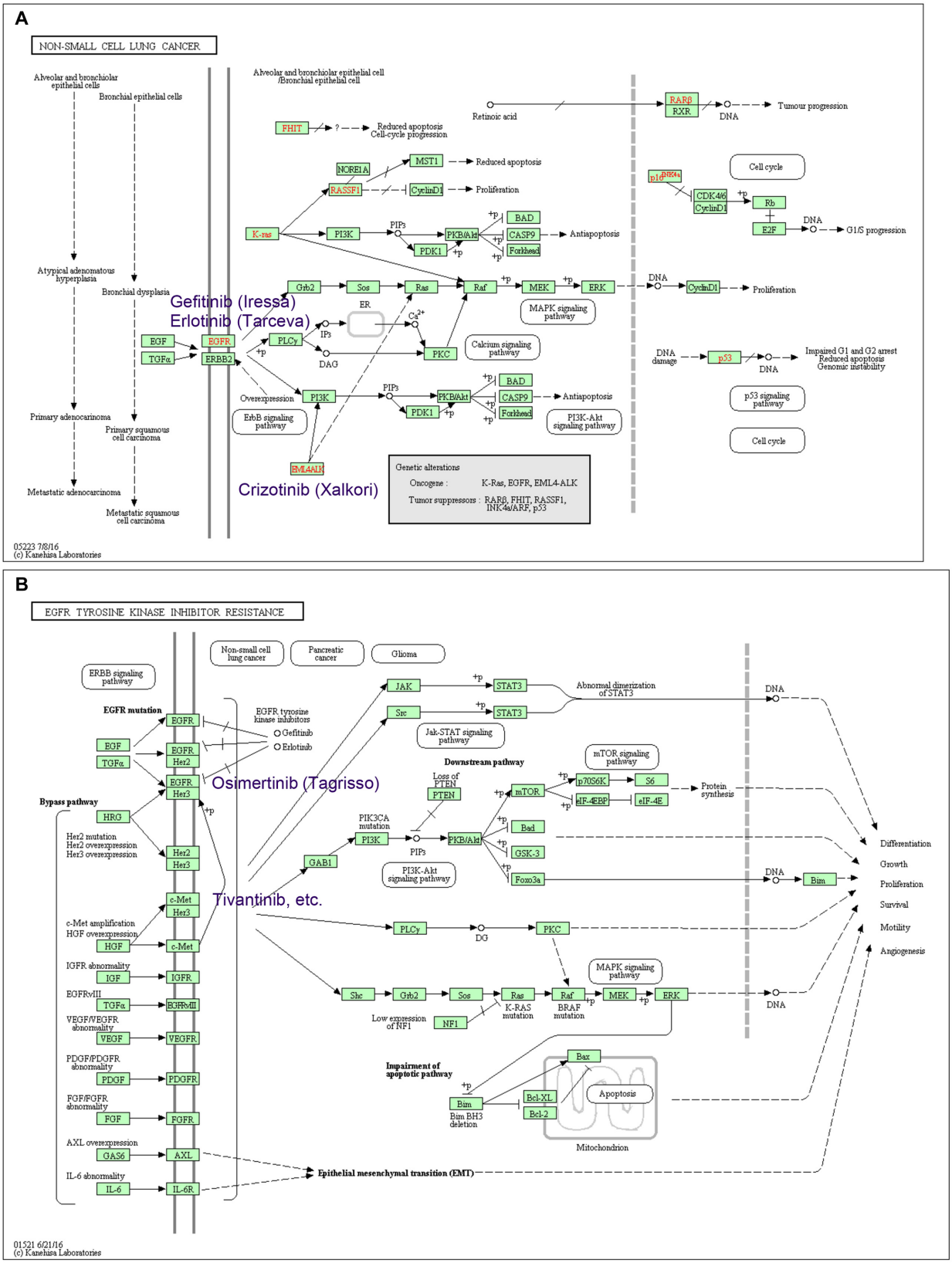


Figure 4. KEGG pathway maps for (**A**) non-small cell lung cancer (hsa05223) and (**B**) EGFR tyrosine kinase inhibitor resistance (hsa01521). Some of the well-known molecular target drugs are included in these maps.

Table 7. KEGG Mapper tools

Tool	Query dataset	Database
Search Pathway	KOs, gene identifiers, C numbers, etc.	PATHWAY
Search&Color Pathway	KOs, gene identifiers, C numbers, etc.	PATHWAY
Color Pathway	KOs, gene identifiers	single KEGG pathway map
Color Pathway WebGL	KOs, gene identifiers	single KEGG pathway map
Search Brite	KOs, gene identifiers, C numbers, etc.	BRITE
Search&Color Brite	KOs, gene identifiers, C numbers, etc.	BRITE
Join Brite	KOs, D numbers, etc.	single BRITE hierarchy
Join Brite Table	KOs, D numbers, etc.	single BRITE table
Search Module	KOs, gene identifiers, C numbers, etc.	MODULE
Search&Color Module	KOs, gene identifiers, C numbers, etc.	MODULE
Search Disease	KOs, human gene identifiers	DISEASE
Reconstruct Pathway	KOs	PATHWAY
Reconstruct Brite	KOs	BRITE
Reconstruct Module	KOs	MODULE
Map Taxonomy	Organism codes, NCBI taxonomy IDs	Taxonomy file

tools and datasets for such focused mapping and functional characterization.

DISEASE AND DRUG INFORMATION

Perturbed KEGG molecular networks

The KEGG pathway maps describe, for example, how beta-lactam antibiotics are synthesized as natural products (map00311 for penicillin and cephalosporin, map00332 for carbapenem and map00261 for monobactam). However, the pathway map of peptidoglycan biosynthesis (map00550) does not describe how beta-lactams interact with penicillin binding proteins and affect bacterial cell wall biosynthesis, because this type of interaction is considered as perturbation to the normal pathway. Similarly, drugs as well as genetic and environmental factors of diseases are considered as perturbants to the KEGG molecular networks, and they are not explicitly shown in the KEGG pathway maps except for a small number of disease associated genes marked in red (Figure 4). Instead, perturbations are described in the DISEASE and DRUG databases.

The DISEASE database is a collection of disease entries, each consisting of a list of disease genes, carcinogens (for cancers), pathogens (for infectious diseases) and other environmental factors. The DRUG database, which is a comprehensive collection of approved drugs, contains drug target information and drug metabolism information, the latter further divided by metabolizing enzymes and transporters, as well as by substrates, inhibitors and inducers. Thus, the DISEASE and DRUG databases may be integrated with PATHWAY and other molecular network databases to understand perturbed molecular networks. Computationally generated pathway maps are available with the special five-letter organism code 'hsadd' for disease genes and drug targets shown on human pathway maps with coloring of pink and light blue, respectively.

Systematic analysis of drug labels

One problem of developing the DISEASE database was how to define a disease entry because, for example, a broad disease name may represent multiple specific diseases or a disease may consist of multiple types corresponding to different gene mutations. KEGG DISEASE entries contain this inherent hierarchy, usually multiple entries created for a broad name and a specific name, but a single entry covering multiple types of minor variations. We have recently started using all Japanese drug labels and selected FDA drug labels to improve the quality of the DISEASE database. Indications described in drug labels are used to add or modify disease entries. Many drugs, especially antineoplastic agents, are often indicated for narrowly defined diseases with genomic features, such as ALK-positive non-small cell lung cancer (Figure 4). Such features are usually treated as different types of a single disease and the information about pharmacogenomic biomarkers is stored in the DRUG database.

As part of the KEGG MEDICUS resource, drug-drug interactions designated as contraindications or precautions are extracted from all the Japanese drug labels and standardized with KEGG identifiers to maintain the drug-drug interaction database. In drug labels interactions are described not only by individual drug names but also by drug group or drug class names, such as nonsteroidal anti-inflammatory drug (NSAID). In a similar way as KOs are defined as network nodes, appropriate grouping is necessary for incorporating drugs into the KEGG molecular networks (1). Drug groups in the KEGG DGROUP database have been developed for this purpose, such as DG01504 for NSAID, and they have been systematically compared with all the Japanese drug labels to improve the content.

Drug labels and other documents regulated by government agencies contain rich information with practical relevance that may not be available in research articles. KEGG is now expanding to incorporate such regulatory information in order to make better links between research findings and practical values.

Accessing KEGG

KEGG is made available at both the KEGG main site (http://www.kegg.jp/) and the GenomeNet mirror site (http://www.genome.jp/kegg/). DBGET queries are available at both sites, but direct queries against the KEGG relational databases, as well as some tools, are available only at the main site. The difference may not be noticeable by the users because mutual links are made as if they belong to a single site.

ACKNOWLEDGEMENTS

Computational resources were provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University.

FUNDING

National Bioscience Database Center of the Japan Science and Technology Agency (in part). Funding for open access charge: National Bioscience Database Center of the Japan Science and Technology Agency.
Conflict of interest statement. None declared.

REFERENCES

1. Kanehisa,M., Sato,Y., Kawashima,M., Furumichi,M. and Tanabe,M. (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.*, **44**, D457–D462.
2. Kanehisa,M., Goto,S., Furumichi,M., Tanabe,M. and Hirakawa,M. (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.*, **38**, D355–D360.
3. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D. *et al.* . (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
4. Clark,K., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2016) GenBank. *Nucleic Acids Res.*, **44**, D67–D72.
5. Federhen,S. (2012) The NCBI Taxonomy database. *Nucleic Acids Res.*, **40**, D136–D143.
6. McDonald,A.G. and Tipton,K.F. (2014) Fifty-five years of enzyme classification: advances and difficulties., *FEBS J.* **281**, 583–592.
7. Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K.F., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–D357.
8. Muto,A., Kotera,M., Tokimatsu,T., Nakagawa,Z., Goto,S. and Kanehisa,M. (2013) Modular architecture of metabolic pathways revealed by conserved sequences of reactions. *J. Chem. Inf. Model.*, **53**, 613–622.
9. UniProt Consortium (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.
10. Kanehisa,M., Sato,Y. and Morishima,K. (2016) BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *J. Mol. Biol.*, **428**, 726–731.
11. Suzuki,S., Kakuta,M., Ishida,T. and Akiyama,Y. (2014) GHOSTX: an improved sequence homology search algorithm using a query suffix array and a database suffix array. *PLoS One*, **9**, e103833.
12. Kanehisa,M. (2013) Chemical and genomic evolution of enzyme-catalyzed reaction networks. *FEBS Lett.*, **587**, 2731–2737.