D590–D595 *Nucleic Acids Research, 2019, Vol. 47, Database issue* *Published online 13 October 2018*
doi: 10.1093/nar/gky962

New approach for understanding genome variations in KEGG

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

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

Received September 08, 2018; Revised October 01, 2018; Editorial Decision October 03, 2018; Accepted October 05, 2018

ABSTRACT

KEGG (Kyoto Encyclopedia of Genes and Genomes; https://www.kegg.jp/ or https://www.genome.jp/kegg/) is a reference knowledge base for biological interpretation of genome sequences and other high-throughput data. It is an integrated database consisting of three generic categories of systems information, genomic information and chemical information, and an additional human-specific category of health information. KEGG pathway maps, BRITE hierarchies and KEGG modules have been developed as generic molecular networks with KEGG Orthology nodes of functional orthologs so that KEGG pathway mapping and other procedures can be applied to any cellular organism. Unfortunately, however, this generic approach was inadequate for knowledge representation in the health information category, where variations of human genomes, especially disease-related variations, had to be considered. Thus, we have introduced a new approach where human gene variants are explicitly incorporated into what we call 'network variants' in the recently released KEGG NETWORK database. This allows accumulation of knowledge about disease-related perturbed molecular networks caused not only by gene variants, but also by viruses and other pathogens, environmental factors and drugs. We expect that KEGG NETWORK will become another reference knowledge base for the basic understanding of disease mechanisms and practical use in clinical sequencing and drug development.

INTRODUCTION

Conservation and variation are inherent features of biological systems at different levels. The KEGG (Kyoto Encyclopedia of Genes and Genomes) database has been developed for understanding conservation and variation of genes and genomes at the level of cellular organisms. In particular, the KO (KEGG Orthology) system for functional orthologs has been developed for representing conserved features of genes and proteins, and the reference knowledge base of KEGG pathway maps drawn as networks of KO nodes has been developed for representing conserved features of cellular processes and organism behaviors. This generic architecture allows for KEGG mapping, whereby pathways for a specific organism can be automatically reconstructed from the gene set in its genome once genes are annotated with KO identifiers. As a result, KEGG has become a widely used resource for the biological interpretation of different types of data, including genomes, transcriptomes, metabolomes and metagenomes, for many organisms and environmental samples.

A disadvantage of this generic approach is caused by the fact that *Homo sapiens* is treated simply as one of many species stored in KEGG. As the health information category of the KEGG database expands, a new approach has become necessary for better integration of human genomes, human pathways, human diseases and drugs (1). Thus, we now explicitly incorporate conservation and variation within the single species of *H. sapiens*.

Diseases have been viewed in KEGG as perturbed states of molecular networks caused by various perturbants of genetic and environmental factors, and drugs viewed as different types of perturbants (2). Thus far, however, such perturbations were not well represented. For example, known genetic alterations in cancers and other disease-associated genes are marked in red in the disease pathway maps, but since they are drawn as KO nodes, only the reference human gene data are linked from such red-marked nodes. With the new approach described in this paper, the actual dataset of perturbed molecular networks is presented in the new KEGG NETWORK database, which is a collection of network variants, such as variants of signaling networks caused by gene variants, viruses and other factors.

OVERVIEW OF KEGG

In December 1995, the first release of KEGG was made available with just four databases: PATHWAY, GENES, COMPOUND and ENZYME. As shown in Figure 1,

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

Category	Database name	Content	Release
Systems Information	KEGG PATHWAY	KEGG pathway maps	1995
	KEGG BRITE	BRITE functional hierarchies and tables	2005
	KEGG MODULE	KEGG modules	2006
Genomic Information	KEGG ORTHOLOGY (KO)	KO groups for functional orthologs	2002
	KEGG GENOME	KEGG organisms (complete genomes)	2000
	KEGG GENES	Genes and proteins	1995
	KEGG SSDB	Sequence similarity among GENES entries	2001
Chemical Information	KEGG COMPOUND	Metabolites and other small molecules	1995
	KEGG GLYCAN	Glycans	2003
	KEGG REACTION / RCLASS	Biochemical reactions and reaction class	1998/2010
	KEGG ENZYME	Enzyme nomenclature	1995
Health Information	KEGG NETWORK / VARIANT	Disease-related network and gene variants	2017
	KEGG DISEASE	Human diseases	2008
	KEGG DRUG / DGROUP	Drugs and drug groups	2005/2014
	KEGG ENVIRON	Crude drugs and health-related substances	2010
	JAPIC	Japanese drug labels	2007
	DailyMed	FDA drug labels (links only)	2012

Figure 1. KEGG consists of eighteen databases in four categories, which are all manually curated except computationally generated SSDB. The databases in the chemical information category are collectively called KEGG LIGAND. The databases in the health information category together with two outside databases, Japanese drug labels obtained from the JAPIC database (http://www.japic.or.jp) and FDA drug labels linked to the DailyMed database (https://dailymed.nlm.nih.gov), are collectively called KEGG MEDICUS.

KEGG now consists of eighteen databases in four categories, but the basic concept remains the same. The three generic categories of systems, genomic and chemical information have been expanded from the PATHWAY, GENES and COMPOUND databases, respectively, and the additional human-specific category of health information was introduced in order to make KEGG more useful in practice. The idea of pathway mapping was first realized by the EC number system of ENZYME, where KEGG metabolic pathway maps were drawn with EC number nodes and enzyme genes in the genome were assigned EC numbers. However, the EC-based pathway mapping lasted only for 5 years and the EC numbers were replaced by ortholog IDs, which later became KOs. Since 2003 pathway mapping is performed by the KO system, where all KEGG pathway maps are created with KO nodes and genes in the genome are assigned KO identifiers (K numbers). EC numbers are indirectly assigned to individual genes through the KO system. The KO-based mapping is also applied to BRITE hierarchy files and KEGG modules.

In December 2017, the KEGG NETWORK database was released together with the associated database of KEGG VARIANT. Both are human-specific databases and form part of the health information category (Figure 1). KEGG NETWORK is our first attempt to explicitly consider genome variations within a single species. Although we consider only the variations that are relevant to human diseases and drugs, the methodology used in KEGG NETWORK can be applied to any variations in any species. In the following sections, we describe KEGG NETWORK and other developments in the past 2 years. A more complete description of the KEGG database can be found in the previous article of the 2017 Nucleic Acids Research Database Issue (1).

KEGG NETWORK

From gene variants to network variants

Figure 2 illustrates the concept behind KEGG NETWORK. In contrast to the generic KEGG databases for understanding conservation and variation among cellular organisms, KEGG NETWORK is focused on *H. sapiens*, providing a more detailed picture especially for understanding human diseases in terms of network-disease associations (2). In the KEGG PATHWAY database, KO-based reference pathways are manually created from published literature and all instances of organism-specific pathways are computationally generated. The KEGG NETWORK database is a collection of network elements, where both reference and variant network elements are manually created from published literature. Reference network elements are represented by human gene IDs, while variant network elements may contain gene variants, viral proteins, environmental factors and drugs. Thus, the variant network elements, which are also called network variants, can accommodate not only gene variants but also other perturbants for understanding disease-related perturbed molecular networks.

Cancer network variants

The KEGG NETWORK database will contain network variants associated with various diseases, but as of September 2018 it contains network variants for cancers, viral infections and certain types of endocrine and metabolic diseases. Figure 3 shows examples of cancer network variants. Cancer cells acquire characteristic features, termed as hallmarks of cancer by Hanahan and Weinberg (3,4), such as sustaining proliferative signaling and resisting cell death,

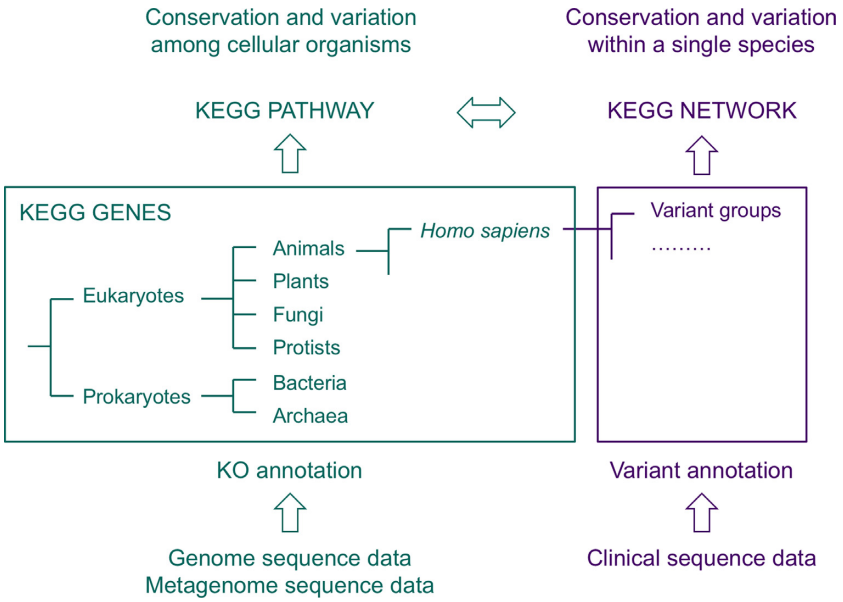

Figure 2. A conceptual diagram of the KEGG NETWORK database. In contrast to the traditional approach where *Homo sapiens* is treated as one of 6000 species in KEGG, the new approach allows variations of human genes and genomes to be explicitly incorporated.

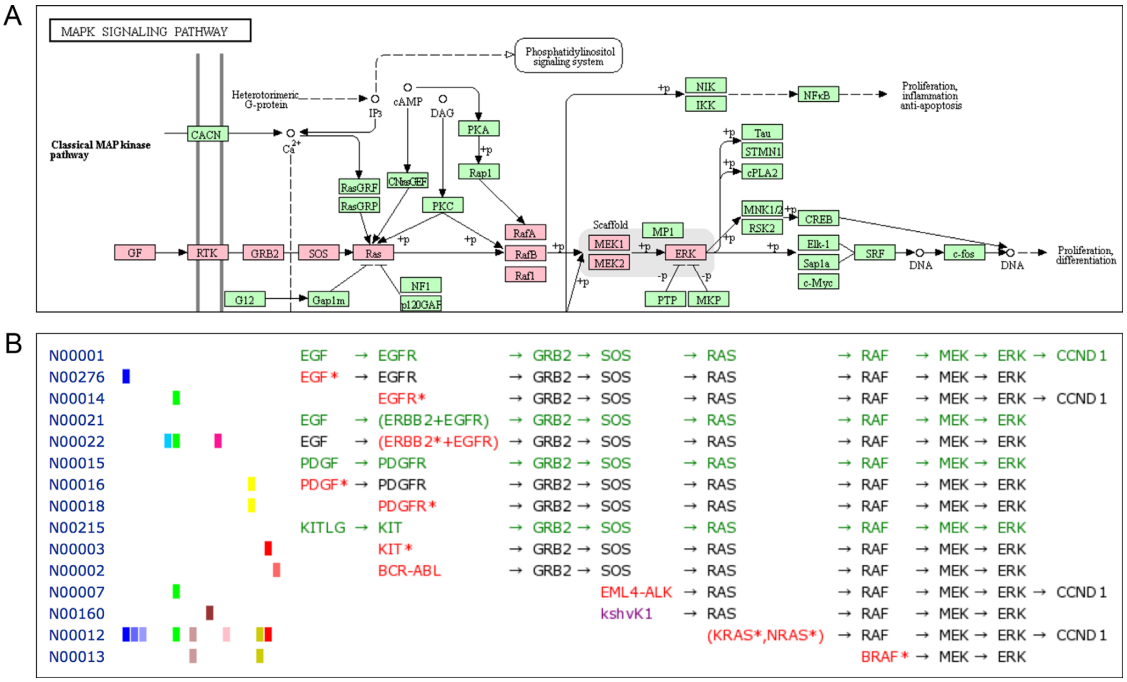

Figure 3. (**A**) The MAPK (ERK) signaling pathway in the KEGG pathway map (hsa04010) where the main path from growth factor to ERK kinase is marked in pink. (**B**) An example of the network variation map (nt06201) as a collection of network elements that correspond to the main path in (A). Coloring of text indicates: green for reference network element, red for gene variant and purple for viral protein.

which are tightly related to perturbations to signaling pathways such as MAPK signaling and PI3K–AKT signaling pathways. Figure 3A shows a part of the MAPK signaling pathway map (KEGG ID: hsa04010) with its main path from growth factor to ERK kinase marked in pink. Figure 3B shows selected data for this path in the NETWORK database, where each network element is identified by the N number and the collection of network elements can be viewed in, what we call, the network variation map.

The network variation map is color-coded: reference network elements in green, gene variants in red, viral proteins in purple and drugs in blue. As in Figure 3B, the map may be shown in alignment mode, meaning that variant network elements are aligned with reference network elements, and

Node	Identifier	Coloring
Human reference gene/protein	hsa ID	None
Human gene variant	variant ID	Red
Viral or other pathogen gene/protein	K number	Purple
Metabolite and other chemical compound	C number	None
Drug	D number	Blue

Edge			
Signaling network	Metabolic network	Text	Symbol
Activation (single- or multi-step)	Enzymatic reaction	–>	→
Inhibition (single- or multi-step)		–|	⊣
Complex formation	Substrate binding	--	—
Missing interaction	Missing reaction	//	⊁
Expression (single-step)	Enhanced reaction	=>	⇒
Repression (single-step)		=|	⇉

Figure 4. Nodes and edges of KEGG network elements.

may be associated with cancer types indicated by another color-coding on the left, such as light green for non-small cell lung cancer and brown for melanoma. Figure 3 contains well-known examples of constitutive activation of the MAPK signaling pathway leading to sustained proliferative signaling, such as by EGFR mutation (N00014) and EML4–ALK fusion (N00007) in non-small cell lung cancer, BCR–ABL fusion (N00002) in chronic myeloid leukemia and BRAF mutation (N00013) in melanoma.

Nodes and edges of network element

Currently, each network element is a linear sequence of nodes connected by different types of edges as summarized in Figure 4. Nodes include human reference genes (identified by hsa IDs in KEGG GENES), human gene variants, viral proteins, metabolites and drugs. The two-character text representation of edges is used in the definition field of the N number entry, and the symbol representation of edges is used in the network variation map.

Gene variants may include (i) gain of function by activating mutation, amplification and fusion, (ii) loss of function by inactivating mutation and deletion and (iii) gene/protein overexpression. They are stored in the KEGG VARIANT database and identified by variant IDs, such as hsa_var:1956v1 and hsa_var:1956v2, representing EGFR (hsa:1956) amplification and mutation, respectively. Each variant entry may contain multiple instances of known mutations and other genetic alterations whenever functional consequences are considered equivalent. For example, 1956v2 consists of exon 19 deletion and L858R mutation. The KEGG VARIANT database also contains links to outside databases such as ClinVar (5), dbSNP (6) and COSMIC (7).

Figure 3B contains a network variant (N00160) caused by a viral protein, the K1 protein of Kaposi sarcoma-associated herpesvirus (KSHV). This viral oncoprotein activates multiple pathways for sustaining proliferative signaling and resisting cell death, which is similar to human oncogenes, such as EML4–ALK fusion gene as shown in Figure 5. Another important aspect of viral protein perturbations involves evading immune destruction, which is observed in both cancer-causing viruses and non-cancer-causing viruses. One of the evasion strategies is called viral mimicry (8), which is to encode homologs of proteins that regulate immune responses such as cytokines and cytokine receptors. These and other perturbations of viral proteins are being organized in KEGG NETWORK, together with enhanced versions of pathway maps for viral infections in KEGG PATHWAY.

Drug–target relationship

The KEGG NETWORK database also contains drug–target relationships, especially for those drugs with variant proteins as targets. Figure 6 shows anticancer drugs against gene variants in the MAPK signaling pathway shown in Figure 3B. Since cancer cells can develop resistance to molecular targeted drugs by secondary mutations, the drug–target relationships and the corresponding variant data are distinguished. For example, the first-generation tyrosine kinase inhibitors of imatinib, crizotinib and gefitinib are distinguished from the later-generation counterparts. These data are accumulated mostly from the FDA drug labels in the DailyMed database.

OTHER DEVELOPMENTS IN KEGG

KO system updates

The addendum category of the KEGG GENES database was introduced in 2015 as a collection of published protein sequence data with experimentally verified functional information (9). Although the number of sequences is very small (<5000 proteins) compared to the main category of complete genomes (27 million genes), the addendum category is extremely useful for defining KO groups of functional orthologs. As of September 2018, the KO database contains over 22 000 KO entries, in which 85% are linked to publications and 68% are further linked to sequence data, which may be considered as core sequence data for defining KOs. Ten percent of linked sequence data are in the addendum category. The annotation (KO assignment) rate of the KEGG GENES database is continuously improving, currently at 48%, as the KO database increases by 5–7% every year.

The KO system is a hierarchical classification of KO entries representing functional classification of genes and proteins. The KO system was originally developed as a pathway-based classification, but due to inclusion of other datasets, there were discrepancies among the KO system (KEGG ID: ko00001), the PATHWAY classification (br08901) and the BRITE classification (br08902). This has been corrected and the new KO system consists of eight top categories: six for PATHWAY (Metabolism, Genetic Information Processing, Environmental Information Processing, Cellular Processes, Organismal Systems and Human Diseases), one for BRITE (Brite Hierarchies) and one for the rest (Not Included in Pathway or Brite).

Sequence data for Enzyme Nomenclature

Since 1961 the Enzyme Commission, currently the IUBMB/IUPAC Biochemical Nomenclature Committee,

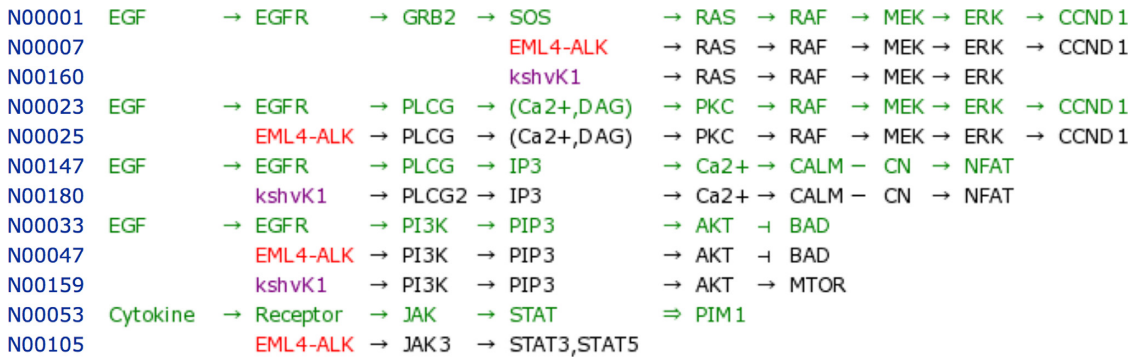

Figure 5. A comparison of signaling pathways activated by the oncoprotein K1 of KSHV and the oncogene EML4–ALK in non-small cell lung cancer. The pathways are involved in sustaining proliferative signaling and resisting cell death.

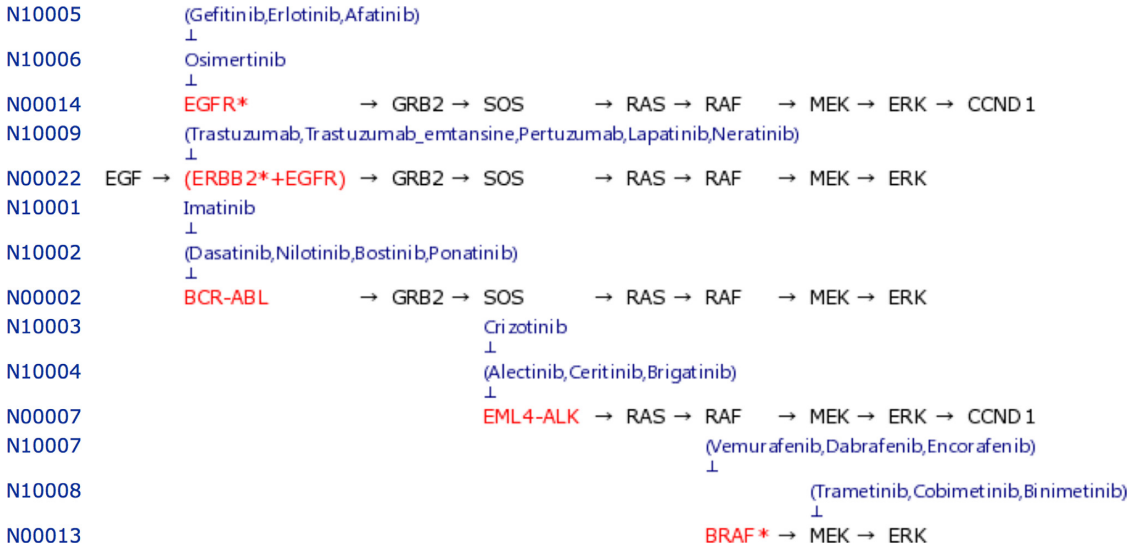

Figure 6. Examples of drug–target relationships for anticancer drugs that inhibit signaling pathways shown in Figure 3B.

has produced the Enzyme Nomenclature list consisting of hierarchically classified EC (Enzyme Commission) numbers given to experimentally observed and published enzymatic reactions. KEGG ENZYME is an implementation of the Enzyme Nomenclature taken from the ExplorEnz database (10) with additional information about sequence data for the enzymes used in the original experiments. The Enzyme Nomenclature list is constantly expanded, and it is the most important source of finding new publications on protein functions. These publications are manually examined to identify sequence data, which are incorporated in KEGG GENES usually as addendum entries. Then new KO entries are defined, whenever appropriate, with association of EC numbers. The relationships between KOs and EC numbers are many to many. One KO may be associated with multiple EC numbers, and one EC number may be given to multiple KOs.

As of September 2018 about half of over 6000 EC entries are linked to sequence data. Figure 7 shows the number of EC entries versus the created year with known sequences in blue and without known sequences in gray. Most of the recently added EC entries are linked to sequence data, but the old entries originally appeared in the printed versions of the Enzyme list are more problematic, because it is often the case that the enzymes were isolated and the experiments were performed without knowledge of sequence data.

Improvements of DISEASE and DRUG databases

As the NETWORK and VARIANT databases were introduced in the health information category, the DISEASE and DRUG databases have undergone some changes. First, the Target field of the DRUG database now contains variant IDs in accordance with the network data for drug–target relationship (Figure 6). Second, the links between disease entries (identified by H numbers) and drug entries (identified by D numbers) are based solely on drug labels. The Disease field of the DRUG database contains diseases indicated in the drug labels, and reverse links are automatically generated for the Drug field of the DISEASE database. Consequently, there are some differences in the drug-disease links between the English version based on FDA drug labels and the Japanese version based on Japanese drug labels. Third, the relationships among disease entries are being reorganized by introducing the subgroup and supergroup names. Fourth, the disease entries are given ICD-11 codes released

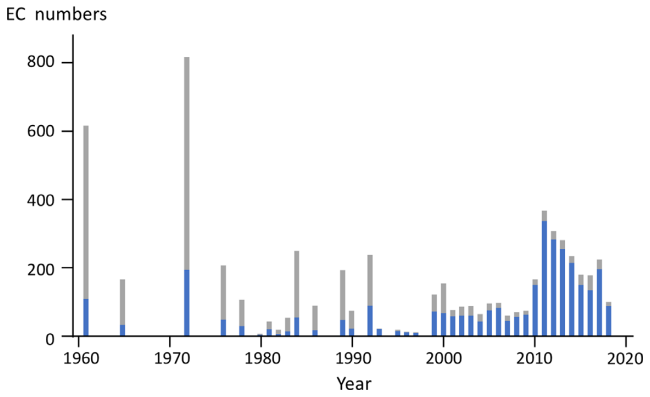

Figure 7. The EC numbers assigned each year. Blue coloring indicates the fraction of EC numbers in which sequence data for the enzymes used in the original experiments could be identified.

by WHO in June 2018. The ICD-11 codes for diseases and the ATC codes for drugs play roles of linking to/from many outside resources.

In addition to the human genome to disease relationship in the Gene field of the DISEASE database, the pathogen genome to infectious disease relationship is being reorganized in the Pathogen field of the DISEASE database, which now contains the Module subfield for signature modules of pathogenicity and antimicrobial resistance. A tool to identify antimicrobial resistance in pathogen genomes is available (11) as part of the BlastKOALA server (12,13).

Accessing KEGG

KEGG is made available at both the KEGG main site (https://www.kegg.jp/) and the GenomeNet mirror site (https://www.genome.jp/kegg/). Direct queries against the KEGG relational databases and some tools such as BlastKOALA and GhostKOALA (12,13) are available only at the main site, while metagenome data (MGENOME and MGENES) and various analysis tools are maintained at the GenomeNet site. The content of KEGG IDs mentioned in this paper, such as hsa04010, nt06201, N00014 and hsa_var:1956v2, can be retrieved by entering an ID into the search box at the top page of either 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., Furumichi,M., Tanabe,M., Sato,Y. and Morishima,K. (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, **45**, D353–D361.
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. Hanahan,D. and Weinberg,R.A. (2000) The hallmarks of cancer. *Cell*, **100**, 57–70.
4. Hanahan,D. and Weinberg,R.A. (2011) The hallmarks of cancer: the next generation. *Cell*, **144**, 646–674.
5. Landrum,M.J., Lee,J.M., Benson,M., Brown,G., Chao,C., Chitipiralla,S., Gu,B., Hart,J., Hoffman,D., Hoover,J. *et al.* (2016) ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.*, **44**, D862–D868.
6. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielsk,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
7. Forbes,S.A., Beare,D., Boutselakis,H., Bamford,S., Bindal,N., Tate,J., Cole,C.G., Ward,S., Dawson,E., Ponting,L. *et al.* (2017) COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.*, **45**, D777–D783.
8. Alcami,A. (2003) Viral mimicry of cytokines, chemokines and their receptors. *Nat. Rev. Immunol.*, **3**, 36–50.
9. 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.
10. McDonald,A.G. and Tipton,K.F. (2014) Fifty-five years of enzyme classification: advances and difficulties. *FEBS J.*, **281**, 583–592.
11. Kanehisa,M. (2018) Inferring antimicrobial resistance from pathogen genomes in KEGG. *Methods Mol. Biol.*, **1807**, 225–239.
12. 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.
13. Kanehisa,M. (2017) Enzyme annotation and metabolic reconstruction using KEGG. *Methods Mol. Biol.*, **1611**, 135–145.