The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website

The discovery of mutations in cancer genes has advanced our understanding of cancer. These results are dispersed across the scientific literature and with the availability of the human genome sequence will continue to accrue. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website have been developed to store somatic mutation data in a single location and display the data and other information related to human cancer. To populate this resource, data has currently been extracted from reports in the scientific literature for somatic mutations in four genes, BRAF, HRAS, KRAS2 and NRAS. At present, the database holds information on 66,634 samples and reports a total of 10,647 mutations. Through the web pages, these data can be queried, displayed as figures or tables and exported in a number of formats. COSMIC is an ongoing project that will continue to curate somatic mutation data and release it through the website.

British Journal of Cancer (2004) 91, 355–358. doi:10.1038/sj.bjc.6601894 www.bjcancer.com

© 2004 Cancer Research UK

Keywords: somatic; mutation; database; website

Approximately one in three individuals in Europe and North America develops one of the approximately 200 different classes of cancer and it is the cause of death of one in five (Higginson, 1992). All cancers arise as a result of the acquisition of a series of fixed DNA sequence abnormalities, each of which ultimately confers growth advantage upon the clone of cells in which it has occurred (Vogelstein and Kinzler, 1998). These abnormalities include base substitutions, deletions, amplifications and rearrangements. The extent to which each of these mechanisms contributes to cancer varies markedly between different genes, and probably also between different cancer types. Identification of the genes that are mutated in cancer is a central aim of cancer research. Over the past 25 years, approximately 300 genes have been shown to be somatically mutated in cancer (Futreal et al., 2004). This work forms the foundation for understanding the biological abnormalities within neoplastic cells, provides information on the function of gene products and sheds light on more complex questions such as the relationships between genes and biochemical pathways. Current strategies for the development of new therapeutic and preventive agents in cancer are increasingly dependent upon modulation of these critical molecular targets.

The scientific literature is a rich source of mutation data that, in general, is published in a piecemeal fashion. More comprehensive data sources do exist, such as Online Mendelian Inheritance in Man (OMIM, Wheeler et al., 2004), HGVbase (Fredman et al., 2002) and the Human Gene Mutation Database (HGMD, Stenson et al., 2003). These databases give overviews of the genetics and biology of many genes and associated diseases (OMIM), genome variants and associated genotype–phenotype relationships (HGVbase) or germline mutation data (HGMD). For somatic mutations in cancer, there are many locus-specific web resources, such as those for p53 (Olivier et al., 2002; Béroud and Soussi, 2003), that cover a single gene in depth. The value of these various databases should not be underestimated; however, none of them offer a comprehensive view of all previously reported somatic mutations in cancer. Looking to the future, the volume of somatic mutation data will continue to expand and the scientific community will be better served if this data is provided in a coherent fashion. A public, comprehensive, intuitive, accessible and integrated database is required to maximise the benefit from this rich data set. The Catalogue of Somatic Mutations in Cancer (COSMIC), (http://www.sanger.ac.uk/cosmic) is a database that holds somatic mutation data and associated information, and can be interrogated through a series of web pages to provide a graphical or tabular view of the data along with various export options. To date, the database has been populated with data from four genes: HRAS, KRAS2, NRAS and BRAF.

DATA CURATION

Gene selection

The genes that have been selected for curation are taken from the list of cancer genes assembled in the Cancer Gene Census (Futreal et al., 2004). In the first instance, data was obtained for four genes that are known to be somatically mutated in cancer: HRAS (Reddy et al., 1982), KRAS2 (McCoy et al., 1983), NRAS (Hall et al., 1983) and BRAF (Davies et al., 2002).
Data extraction from the literature

PubMed (Wheeler et al., 2004) is broadly searched for references containing relevant somatic mutation data in cancer (example search: (ras OR genes, ras) AND human AND mutation). In the first instance, the abstract is read to identify, and select for inclusion in the database, papers that are likely to include somatic mutation information relating to cancer or precancerous conditions. Primary research papers are read and information about the samples, mutations and experimental methods (see Table 1) is extracted and entered into the database. Reviews are also selected if thought to be specific to a gene of interest. In order to avoid duplication of data, this source is used to identify the relevant primary literature and not as the source of the mutation data. Any references containing incomplete data (e.g. mutations reported but not fully described) or data of insufficient quality (e.g. errors identified in the data) are not fully curated but are added to a list of additional references containing somatic mutation information. Simple mutations are fed through Mutation Checker (Stajich et al., 2002) before being imported to COSMIC, while more complex alterations are manually annotated.

COSMIC DATABASE

The COSMIC database is implemented in an Oracle relational database and has five sections each containing multiple tables.

Gene information

A static version of each gene is maintained in COSMIC. The genomic structure of each gene and chromosome location is derived from Ensembl (Birney et al., 2004) and cDNA sequence and protein sequence from the RefSeq project (Wheeler et al., 2004). Other information is held to provide links to web resources such as Ensembl (Birney et al., 2004), Pfam (Bateman et al., 2004), InterPro (Mulder et al., 2003) and OMIM (Wheeler et al., 2004).
Mutation information

COSMIC can accommodate information on base substitutions, insertions and deletions, translocations and changes in copy number. For the four genes presently in COSMIC, there are 147 unique mutations (36 for BRAF, 27 for HRAS, 52 for KRAS2 and 32 for NRAS). In the tumours that have been analysed, there are a total of 10,647 mutations, 736 in BRAF, 477 in HRAS, 8,302 in KRAS2 and 1,132 in NRAS.

Tumour classification system

The tissue site and histology data is taken from the curated papers and entered into COSMIC (this forms the ‘paper definition’). Tumour classification is a continually evolving field and there is no standard nomenclature adhered to for the purposes of publication in the various journals. Identical tissues and histologies can have different labels depending on the origin and age of the study. To overcome difficulties caused by these alternate nomenclatures, a standardised system of definitions has been developed (the ‘COSMIC definitions’) through consultation with experts in the field. This groups data from the same tissue types and histologies and can be used to translate the ‘paper definitions’ to ‘COSMIC definitions’. Every sample has up to eight definitions: primary tissue, tissue subtype 1, 2 and 3, primary histology and histology subtypes 1, 2 and 3. If there is no data for any of these definitions, COSMIC records an entry of NS, not specified. A total of 513 tissue definitions have been noted in the papers in COSMIC and have been translated to 372 COSMIC tissue definitions. Likewise, a total of 1,150 histology definitions were found in the papers in COSMIC that were translated to 425 COSMIC histology definitions. This unified classification system is presented through the web pages to present a normalised browsing tool.

Individual/tumour/sample data

The sample data is taken from the curated papers and linked to the appropriate gene, paper, classification and when present a mutation. This forms the core of the COSMIC database. An individual can have many tumours and each tumour can have many samples. However in the COSMIC scheme, each sample is unique and could be considered as a single experiment. There are 66,634 sample records in COSMIC (5,158, 11,876, 35,716 and 13,884 for BRAF, HRAS, KRAS2 and NRAS, respectively). These samples are derived from 57,444 tumours of which 51,988 were analysed in one gene, 2,353 in two genes, 2,950 in three genes and 173 in all four genes.

Figure 1 The initial output from COSMIC is a graphical view of the mutations distributed along the linear amino-acid sequence of the gene. The scale bar incorporates a zoom function to generate a more detailed view of the protein to the point where individual amino acids are named (when there are fewer than 31 amino acids displayed). When a Pfam or Interpro domain is present, a link is provided to these resources (adjacent to the Domain label) while links to the papers that were curated are positioned beneath the mutations (in red) with an option of either viewing the papers that have data for a particular location in the protein or all of the papers for the selected gene.
then one or more histologies, and finally one or more genes. If only one tissue or histology is selected, it is possible to select one or more tissue or histology subtypes before making a gene selection. All of the tissues present in the COSMIC classification scheme are available from the first page; however, subsequent pages only show the relevant options and not the entire list of options, for example having selected eye, the tissue subtype options are retina and uveal tract.

Data display

After querying the database, the results are displayed as a figure (Figure 1) and as a series of tables (Table 2) for each gene that was selected. The figure shows the linear amino-acid sequence derived from the gene with the mutations positioned along its length. Further information and links are provided as appropriate to the protein sequence. The table gives a summary of the mutations stratified by tissue and histology. The depth of the stratification relates to the depth of the original query. If only tissue was selected, the data will be stratified by tissue; however, if tissue, subtissue, histology and subhistology are selected, the data will be broken down further. Links from this table reload the figure to display a subset of the data and provide more details of the specific mutations. Two other tables provide a summary of the statistics in COSMIC for the selected gene and a summary of the mutations shown in the figure.

Exports and downloads

Having displayed the results from a query, the data can be formatted in simple text, Excel or HTML that can be downloaded from the COSMIC site. The cDNA and protein sequences are available through the Additional Info link on the COSMIC home page as is the Classification Scheme.

FUTURE DIRECTIONS

There is a continuing effort to enter additional somatic mutation data into COSMIC. In order to keep the data in COSMIC up-to-date, we regularly monitor the literature for new reports of mutations in the genes that exist in COSMIC. In addition, further cancer genes will be taken from the Cancer Gene Census (Futreal et al, 2004) and curated. The COSMIC website will be developed further to make use of the underlying data. This will include a DNA view of the mutations and methods to display insertions and deletions. In addition, we will display other data that has already been captured such as the patient sex and age for the samples and the experimental methods used to screen for the mutations. There are however limitations to this data as we can only collect data that is described in the original work. Even with this caveat the data provides a direct summary of the somatic mutation literature. Considering the data set as a whole it will be possible to analyse, in greater detail, the wider aspects of the biology underlying the genetic changes that take place in cancer.

ACKNOWLEDGEMENTS

We thank Frances Martin and the Institute of Cancer Research and The Wellcome Trust for funding this work.

REFERENCES

Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR (2004) The Pfam protein families database. Nucleic Acids Res 32: D138 – D141
Béroud C, Soussi T (2003) The UMD-p53 database: new mutations and quality and a broad spectrum of data sources. Nucleic Acids Res 31: 455 – 461

Higginson J (1992) Human cancer: epidemiology and environmental causes. In: Higginson, Muis, Munoz (eds). Cambridge Monographs on Cancer Research. Cambridge, UK: Cambridge University Press

Mulder NJ, Apteve R, Atwood TK, Bairroch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P, Boucher P, Copley RR, Coutelle C, Das U, Durbin R, Falquet L, Fleischmann W, Griffiths-Jones S, Haft D, Harte N, Huso N, Kahn D, Kanapin A, Krestyaninova M, Lopez R, Letunic I, Lonsdale D, Silventoinen V, Orchard SE, Pagni M, Peyruc D, Ponting CP, Selengut JD, Servent F, Sigrist CJA, Vaughan R, Zdobnov EM (2003) The InterPro Database, 2003 brings increased coverage and new features. Nucleic Acids Res 31: 315 – 318

Olivier M, Eeles R, Hollstein M, Khan MA, Harris C.C, Hainaut P (2002) The IARC TP53 Database: new online mutation analysis and recommendations to users. Hum Mutat 19: 607 – 614

Reddy EP, Reynolds RK, Santos E, Barbadic M (1982) A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene. Nature 300:149 – 152

Stajich JE, Block D, Boule K, Brenner SE, Chervitz SA, Dagdignian C, Fuellen G, Gilbert JG, Korf I, Lapp H, Lehvakslaiho H, Matsall C, Mungall CJ, Oregon BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E (2002) The BioPerl toolkit: Perl modules for the life sciences. Genome Res 12: 1611 – 1618

Vogelstein B, Kinzler K (1998) The Genetic Basis of Human Cancer. New York: McGraw Hill

Wheeler DL, Church DM, Edgar R, Federhen S, Helmbright W, Madden TL, Pontius JU, Schuldt GD, Schriml LM, Sequeira E, Suzek TO, Tatusova TA, Wagner L (2004) Database resources of the National Center for Biotechnology Information: update. Nucleic Acids Res 32: D35 – D40