*Database*, Vol. 2011, Article ID bar018, doi:10.1093/database/bar018

# Original article

# Data mining using the Catalogue of Somatic Mutations in Cancer BioMart

**Rebecca Shepherd, Simon A. Forbes, David Beare, S. Bamford, Charlotte G. Cole, Sari Ward, Nidhi Bindal, Prasad Gunasekaran, Mingming Jia, Chai Yin Kok, Kenric Leung, Andrew Menzies, Adam P. Butler, Jon W. Teague, Peter J. Campbell, Michael R. Stratton and P. Andrew Futreal***

Cancer Genome Project, The Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire CB10 1SA, UK

***Corresponding author:** Tel: 01223 834244; Fax: 01223 494809; Email: paf@sanger.ac.uk

Submitted 17 March 2011; Revised 15 April 2011; Accepted 19 April 2011

Catalogue of Somatic Mutations in Cancer (COSMIC) (http://www.sanger.ac.uk/cosmic) is a publicly available resource providing information on somatic mutations implicated in human cancer. Release v51 (January 2011) includes data from just over 19 000 genes, 161 787 coding mutations and 5573 gene fusions, described in more than 577 000 tumour samples. COSMICMart (COSMIC BioMart) provides a flexible way to mine these data and combine somatic mutations with other biological relevant data sets. This article describes the data available in COSMIC along with examples of how to successfully mine and integrate data sets using COSMICMart.

**Database URL**: http://www.sanger.ac.uk/genetics/CGP/cosmic/biomart/martview/

## Project description

COSMIC is a repository for somatic mutations and associated phenotype/clinical information, combining data from a number of sources. Primarily, data are manually curated from the scientific literature for genes selected from the Cancer Gene Census (http://www.sanger.ac.uk/genetics/CGP/Census/), a listing of genes that are known to be mutated in human cancer. The manual curation initially focused on known cancer genes that had a high proportion of coding point mutations. In recent years, this has been extended to include cancer genes that are also mutated by gene fusion events. With the advent of next-generation sequencing, COSMIC has been adapted to hold complete catalogues of somatic mutations for individual tumour samples.

The COSMIC website has been developed to overview the underlying database in a user-friendly manner. The website can be navigated by gene, cancer sample or tissue/histology type and has a series of graphical and tabular displays to summarize the content of the database.

However, the integration of biological data sets is still a major IT challenge. Although, the data in COSMIC can be viewed and downloaded in a number of useful ways, it is still challenging for users to generate their own custom queries and also to link to related resources. In order to facilitate the integration of the COSMIC data set, we have used the BioMart data mining software. BioMart uses a federated data model to allow integration of biological data from diverse databases (1). There is now a significant number of data resources that have set up their own BioMarts including Ensembl, UniProt and HGNC. The BioMart software also has interfaces that allow users to easily generate custom queries to obtain subsets of data from one or more BioMart data resources. We have successfully set up an instance of a BioMart, COSMICMart, which holds a summary of the somatic mutations and associated phenotype data from the COSMIC database.

## Data content

COSMIC version 51 (January 2011) contains a full curation of the scientific literature for 91 cancer genes, mostly point-mutated, as well as 53 curated fusion gene pairs, with over 11 000 papers having been assessed. Weekly

*Database*, Vol. 2011, Article ID bar018, doi:10.1093/database/bar018

literature searches are carried out for each cancer gene and each abstract/article is manually evaluated. Checks are made to ensure a paper has not been curated already, whether the paper is a review or is reporting original data, if there are any inconsistencies in the data, and whether all the required information for full curation is present, e.g. information on the numbers of samples screened and sufficient mutation detail. Papers which pass the initial scan are submitted for detailed manual curation, while the remaining papers are 'listed' on the COSMIC website. Information is manually recorded on the actual mutation change, the detailed phenotype of the cancer samples and the original publication/study.

Mutation changes are checked for accuracy and recorded at the genome, transcript (to one reference transcript) and protein level using the mutation syntax standards developed by the Human Genome Variation Society (2). Negative results (sample screened does not have a mutation for a particular gene for a particular publication) are also recorded so prevalence statistics can be estimated for each gene. As part of the curation process, cancer sample tissue/histology descriptions from the original publication are recorded and then re-classified to a COSMIC standard tissue/histology ontology (See classification section of the COSMIC additional information page, http://www.sanger.ac.uk/genetics/CGP/cosmic/add_info/). The standardization of the main data types in COSMIC means the utility of the data is significantly enhanced and the data are easier to browse and incorporate into external data sets. When available, clinical and exposure data and therapeutic responses are also recorded. The full functionality of COSMIC has previously been described (3, 4).

The database coverage has recently been enhanced by the integration of somatic mutations from external data resources. A collaboration with the database curators at International Agency for Research on Cancer (IARC) has allowed the majority of mutations from IARC TP53 database R14 to be integrated into COSMIC (5). Somatic mutations from large-scale systematic cancer screens are curated in COSMIC to include studies on glioblastoma multiforme, breast, colorectal and lung cancer (6–9). Data are also directly curated from the data portals of large-scale projects to include the Cancer Genome Project (CGP) at the Sanger Institute and more recently validated somatic mutations from The Cancer Genome Atlas (TCGA) (6) and International Cancer Genome Consortium (ICGC) (10). COSMIC has been enhanced to accommodate somatic mutations from whole genome and exome screens. This includes all somatic and non-coding mutations, structural rearrangements and gene fusions. Currently, COSMIC has annotated 51 genomes and 332 exomes from a range of cancers to include lung, malignant melanoma, renal, AML, pancreatic and ovarian cancer. The current contents of the database (v51, January 2011) are displayed in Table 1.

**Table 1.** Total contents in v51 of the COSMIC database, January 2011 release

| Curated data type | Curated data count |
|---|---|
| Experiments | 2 946 792 |
| Tumours | 577 304 |
| Mutations | 167 193 |
| References | 11 062 |
| Genes | 19 000 |
| Fusions | 5573 |
| Structural variants | 2729 |
| Whole-cancer genomes | 51 |
| Whole-cancer exomes | 332 |

## Query examples

COSMICMart allows data to be filtered on six different categories (Figure 1): cancer sample, gene, mutation, site of the tumour, histology and other (e.g. Ensembl Gene ID, Swissprot ID, Entrez Gene ID). The interface has a number of pre-selected filters and attributes; mutated samples are selected by default. Users can change these to suit their requirements. Results are displayed in tabulated form and are exportable in various formats for further analysis.

*Query #1*: 'Find all missense substitution mutations for BRAF in cell lines, and display sample, mutation, site, and histology information' (Figure 1, Table 2).

Missense mutations are the most common variant type in COSMIC; over 90% of mutations in BRAF are missense mutations at the p.V600 position. The results are returned as a tabular summary with links back to the COSMIC website. The sample name field links back to the COSMIC sample overview web page, and mutation ID (COSM ID) to the COSMIC mutation summary page (Figure 2). From the COSMIC mutation summary web page, there are links to the Ensembl contig view so the mutation can be viewed in a genomic context. There are also links to the GMOD's GBrowse where COSMIC coding and non-coding mutations, gene footprints, structural rearrangements and copy number variants can be viewed (11).

*Query #2*: 'Find all gene fusion mutations involving the FUS gene with a primary site of bone, and display mutation and sample information' (Table 3).

Gene fusions have been associated with a number of specific tumour types including prostate and blood tumours. These biomarkers can be useful in diagnosis and as targets for drug therapies. COSMIC has annotations, for an increasing number of gene fusion mutations, which are viewable using COSMICMart. The COSMIC fusion mutation ID links to the gene fusion summary pages, which give a graphical view of different fusion structures observed. Many of the papers describing gene fusions have identified

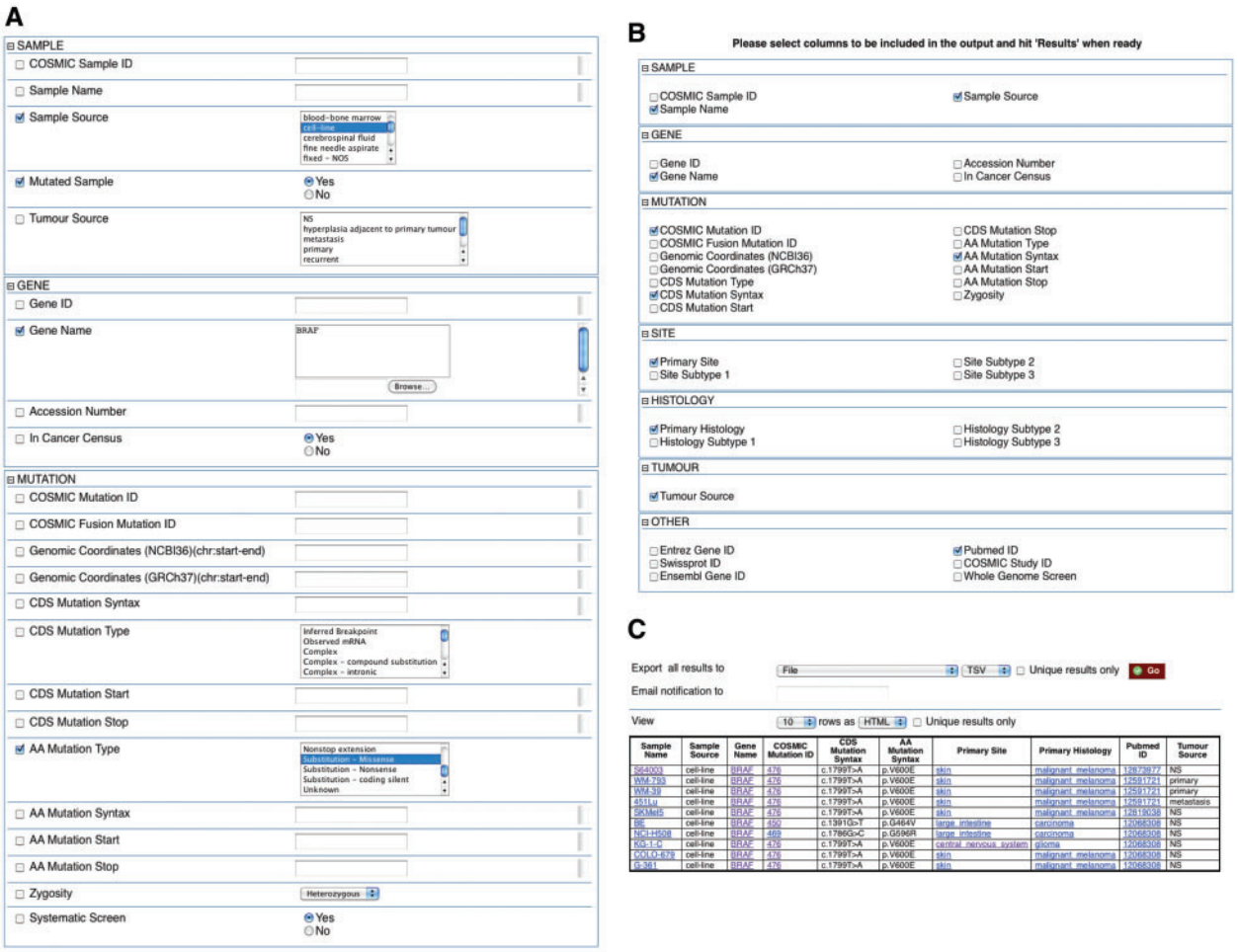


**Figure 1.** Example of how COSMICMart can be queried. This query searches for all cell lines with missense substitution mutations in the BRAF gene (**A**). Attributes can be selected (**B**) to display in the results table (**C**).

**Table 2.** Data sets, filters and attributes selected for query #1

| Data sets | Filters | Attributes |
|---|---|---|
| COSMIC51 | Mutated sample: yes | Sample name |
| | Sample source: cell-line | Sample source |
| | Gene name: BRAF | Gene name |
| | AA mutation type: substitution—missense | Cosmic mutation ID (COSM ID) |
| | | CDS mutation syntax |
| | | AA mutation syntax |
| | | Primary site |
| | | Primary histology |
| | | Tumour source |
| | | Pubmed ID |

**Table 3.** Data sets, filters and attributes selected for query #2

| Data sets | Filters | Attributes |
|---|---|---|
| COSMIC51 | Mutated sample: yes | Cosmic sample ID |
| | Gene name: FUS | Sample name |
| | CDS mutation type: inferred breakpoint, observed mRNA | Sample source |
| | Primary site: bone | Cosmic fusion mutation ID |
| | | CDS mutation syntax |
| | | Pubmed ID |

more than one gene fusion product for the same genes in a single sample. Observed mRNAs are the actual expressed products reported in the results. However, to aid display and website navigation, we have inferred the genomic breakpoint from the experimental data.

*Query #3*: 'Find variation information in Ensembl for all genes from mutated samples with a primary site of breast, and display COSMIC gene, mutation and sample information along with Ensembl variation information' (Table 4).

*Database*, Vol. 2011, Article ID bar018, doi:10.1093/database/bar018

**A**

| | |
|---|---|
| **Sample Name** | S64003 |
| **COSMIC id** | 685178 |
| **Source** | Cultured Sample (NS; cell-line) |
| **Tumour Classification** | Category: Primary Site; Subtype1; Subtype 2; Subtype 3 — Entry: skin; NS; NS; NS<br>Category: Primary Histology; Subtype1; Subtype 2; Subtype 3 — Entry: malignant melanoma; NS; NS; NS |
| **Mutations** | Sample: 685178 — Gene: BRAF — AA Mutation: p.V600E — CDS Mutation: c.1799T>A — Somatic Status: Reported in another cancer sample as somatic — Zygosity: Unknown — Validated |
| **References** | **BRAF oncogenic mutations correlate with progression rather than initiation of human melanoma.**<br>*Dong J, Phelps RG, Qiao R, Yao S, Benard O, Ronai Z, Aaronson SA*<br>Cancer Res. 2003;63;3883-5. PMID: 12873977<br>More Details |
| **Individual** | Sample: 685178<br>Age: Unknown<br>Gender: Unknown<br>Parents tested: Unknown<br>Family: Unknown |

**B**

| | |
|---|---|
| **Mutation id** | 476 |
| **AA Mutation** | p.V600E (Substitution - Missense) |
| **CDS Mutation** | c.1799T>A (Substitution) |
| **Mutation Location** | Length (AA) — Amino Acid<br>Length (AA) — Amino Acid Mutation |
| **Genomic View** | NCBI 36 : Ensembl Contig view 7:140099605..140099605<br>GRCh37 : Ensembl Contig view 7:140453136..140453136<br>COSMIC GBrowse: Please click to see this data in COSMIC Cancer Genome Browser |
| **Gene** | BRAF |
| **Tissue Distribution (Top 5)** | thyroid (5750)<br>large intestine (2265)<br>skin (2151)<br>NS (224)<br>ovary (173)<br>Tissues<br>scale (Mutation #) |

| Associated Samples | | | | | |
|---|---|---|---|---|---|
| Sample Name | Primary Tissue | Zygosity | Sample Name | Primary Tissue | Zygosity |
| 10 | skin | Unknown | 1011-mel | NS | Unknown |
| 1020355 | large intestine | Unknown | 1020356 | large intestine | Unknown |
| 1020357 | large intestine | Unknown | 1020428 | thyroid | Unknown |
| 1020429 | thyroid | Unknown | 1020430 | thyroid | Unknown |

**Figure 2.** The COSMIC sample (**A**) and mutation (**B**) summary pages are linked directly from COSMICMart output table.

**Table 4.** Data sets, filters and attributes selected for query #3

| Data sets | Filters | Attributes |
|---|---|---|
| COSMIC51 | Mutated sample: yes | Cosmic sample ID |
| | Primary site: breast | Sample name |
| | | Sample source |
| | | Cosmic mutation ID (COSM ID) |
| | | CDS mutation syntax |
| | | AA mutation syntax |
| Ensembl: *Homo sapiens* genes | | Features: Ensembl gene ID |
| | | Features: Ensembl transcript ID |
| | | Variations: variation source |
| | | Variations: source description |
| | | Variations: reference ID |
| | | Variations: allele |

COSMICMart is federated with Ensembl (12), which allows Biomart queries to return and integrate data from both resources. For instance, the linking of the two resources can allow the retrieval of variation data from both resources (somatic mutations from COSMIC and germline polymorphisms from Ensembl) for a particular gene or set of genes or cancer type. There is an increasing awareness of how genomic variation can affect a tumour's sensitivity or resistance to anti-cancer agents. While this genetic variation can be familial or somatic, an understanding of common genetic variation around known cancer genes can be of much value to investigations searching for loci modifying a tumour's response to drug therapy (13–15). This query is achieved by first selecting the filters/attributes in the COSMIC BioMart and then clicking the 'Dataset' link at the bottom of the left hand margin of the BioMart interface. An additional data set can then be selected from the drop down list, in this instance Ensembl, which allows a federated query between COSMIC and Ensembl. The filters/attributes are then set in the usual way using the Ensembl BioMart to produce an integrated query.

## Future directions

COSMIC will continue to curate newly discovered cancer genes and is committed to update existing cancer genes with a data release every 2 months. This will ensure that the scientific community has an up-to-date catalogue of somatic mutations implicated in human cancer. COSMICMart is also automatically updated with each new COSMIC release, which allows the data set to be easily mined and integrated with other resources. COSMIC has been successfully adapted to hold complete catalogues of somatic mutations for individual cancer samples. Currently COSMIC holds genome-wide data on 383 tumour samples and we expect this to increase in the near future.

It is intended to federate COSMICMart with further BioMart-driven data resources in addition to the current link with Ensembl. Linking our data to PRIDE (16), UniProt (17) and InterPro (18) will allow COSMIC somatic mutation data to be linked to protein and peptide annotation, while the addition of the Reactome (19) database will allow the incorporation of pathway data. We also intend to create direct links between COSMIC and the ICGC Data Portal (http://dcc.icgc.org/) so somatic mutation data can be integrated between the two data resources.

## Funding

Funding for open access charge: Wellcome Trust (grant reference 077012/Z/05/Z).

*Conflict of interest*. None declared.

## References

1. Haider,S., Ballester,B., Smedley,D. *et al*. (2009) BioMart Central Portal–unified access to biological data. Nucleic Acids Res., **1**, 37.
2. Den Dunnen,J.T. and Antonarakis,S.E. (2000) Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. Hum. Mutat., **15**, 7–12.
3. Forbes,S.A., Tang,G., Bindal,N. *et al*. (2010) COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. Nucleic Acids Res., **38**, D652–D657.
4. Forbes,S.A., Bhamra,G., Bamford,S. *et al*. (2008) The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr. Protoc. Hum. Genet.*, Chapter 10, 11.
5. Petitjean,A., Mathe,E., Kato,S. *et al*. (2007) Impact of mutant p53 functional properties on TP53 mutation patterns and tumor phenotype: lessons from recent developments in the IARC TP53 database. Hum. Mutat., **28**, 622–629.
6. Cancer Genome Atlas Research Network. (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature, **455**, 1061–1068.
7. Sjöblom,T., Jones,S., Wood,L.D. *et al*. (2006) The consensus coding sequences of human breast and colorectal cancers. *Science*, **314**, 268–274.
8. Parsons,D.W., Jones,S., Zhang,X. *et al*. (2008) An integrated genomic analysis of human glioblastoma multiforme. *Science*, **321**, 1807–1812.
9. Ding,L., Getz,G., Wheeler,D.A. *et al*. (2008) Somatic mutations affect key pathways in lung adenocarcinoma. *Nature*, **455**, 1069–1075.
10. International Cancer Genome Consortium. Hudson,T.J., Anderson,W. *et al*. (2010) International network of cancer genome projects. *Nature*, **464**, 993–998.
11. Stein,L.D., Mungall,C., Shu,S. *et al*. (2002) The generic genome browser: a building block for a model organism system database. Genome Res., **12**, 1599–1610.

*Database*, Vol. 2011, Article ID bar018, doi:10.1093/database/bar018

12. Flicek,P., Aken,B.L., Ballester,B. *et al*. (2010) Ensembl's 10th year. Nucleic Acids Res., **38**, D557–D562.
13. Sharma,S.V., Haber,D.A. and Settleman,J. (2010) Cell line-based platforms to evaluate the therapeutic efficacy of candidate anticancer agents. *Nat. Rev. Cancer*, **10**, 241–253.
14. Jänne,P.A., Gray,N. and Settleman,J. (2009) Factors underlying sensitivity of cancers to small-molecule kinase inhibitors. Nat. Rev. Drug Discov., **8**, 709–723.
15. McDermott,U., Sharma,S.V. and Settleman,J. (2008) High-throughput lung cancer cell line screening for genotype-correlated sensitivity to an EGFR kinase inhibitor. Methods Enzymol., **438**, 331–341.
16. Vizcaíno,J.A., Reisinger,F., Côté,R. *et al*. (2011) PRIDE and "Database on Demand" as valuable tools for computational proteomics. Methods Mol. Biol., **696**, 93–105.
17. UniProt Consortium. (2010) The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res., **38**, D142–D148.
18. Hunter,S., Apweiler,R., Attwood,T.K. *et al*. (2009) InterPro: the integrative protein signature database. Nucleic Acids Res., **37**, D211–D215.
19. Croft,D., O'Kelly,G., Wu,G. *et al*. (2011) Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res., **39**, D691–D697.