**Annotating Whole Genomes Sequencing in COSMIC**

*(The Catalogue Of Somatic Mutations In Cancer)*

http://www.sanger.ac.uk/cosmic/

**CY Kok, SA Forbes, A Menzies, J Teague, A Butler, R Shepherd, M Jia, K Leung**

**Cancer Genome Project, The Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK**

COSMIC, the Catalogue Of Somatic Mutations In Cancer is designed to store and display somatic mutation information relating to human cancers, combining detailed information on publications, samples and mutation types. The information is curated both from the primary literature and the laboratories at the Cancer Genome Project, Sanger Institute, UK, and then semi-automatically entered into the COSMIC database. In order to provide consistent annotation of the data, COSMIC has developed a classification system for cancer histology and tissue ontology, and adopted HGVS mutation nomenclature recommendations to describe the multiple mutation types involved in cancer. Cancer genetics is moving from systematic screens of candidate gene sets to whole genome sequencing analyses, and COSMIC displays and navigates this new data; we have recently included systematic gene screens and whole genome sequencing studies. COSMIC will annotate and display somatic mutation data that will be emerging from the International Cancer Genome Consortium (ICGC) and the Cancer Genome Atlas (TCGA) projects. New tools are being developed to interpret this new genomic data with coding mutation annotations. In addition COSMIC will be expanded to curate and display data from mouse insertional mutagenesis screening and mouse cancer model exome/genome sequencing in the future.

Email: cosmos@sanger.ac.uk

---

**COSMIC v49 Total Statistics (Sep 2010)**

- **Experiments**: 2,888,511
- **Tumours**: 546,399
- **Samples**: 551,325
- **Mutant Samples**: 138,836
- **Mutations**: 143,772
- **Unique Mutations**: 25,079
- **Papers curated**: 10,578
- **Genes**: 18,647
- **Fusions**: 5,050
- **Structural Variants**: 2,306
- **Whole Genomes**: 28

The data is available without restriction via a website and in datasheets on the FTP site, and through the COSMIC Biomart.

**Examining Data Offline**

Export custom searches from COSMIC
http://www.sanger.ac.uk/cosmic

Download prepared datasheets from COSMIC’s FTP site
ftp://ftp.sanger.ac.uk/pub/COSMIC/

Download whole datasets from the CGP Archive or EBI EGA Archive (including sequence traces & genotyping data, requires signed agreement)

http://www.ebi.ac.uk/ega/page.php?page=studies&name=CGP

**Whole Genome Sequencing**

This literature component is overviewed on a separate web page:
http://Genes.sanger.ac.uk/genetics/CGP/Classic

**Manuel Literature Curation**

Aiming to curate the world’s literature on somatic cancer mutations, COSMIC now holds full and up-to-date information point-mutated genes and gene fusion pairs, resulting from curation of 10,383 individual papers (COSMIC v49). This data is combined with the output of the CGP in the standard blue COSMIC web pages. Over the last 10 years, the information has been entered by dedicated curators with software checks at each datapoint for accuracy and integrity. This literature component is overviewed on a separate web page:
http://Genes.sanger.ac.uk/genetics/CGP/Classic

---

**Cancer Genomic Annotation**

**Manual Literature Curation**

Systematic Screen Publications

- Manual Curation
- Semi-auto Curation
- Auto Import

Data Classification

Display in COSMIC

**Curating Systematic Screen Publications Semi-Automatically**

In the last few years, the improvement of sequencing technology has allowed the examination of thousands of candidate genes for potential involvement in cancer. Whilst such large screens by the CGP have been available in COSMIC for years, the global growth of publications documenting such analyses has encouraged the extension of COSMIC’s curation systems to encompass other publications of these large studies.

Currently 7 such systematic screens are represented in COSMIC, including (approx) 1,212 genes. Each of these studies are mainly focused on cancer tissue samples, and analysis shows a mix of genetic studies.


- **1. Breast Tumour Rearrangement Screen**: 1,212 genes
- **2. Whole Genome screen of Lung Cell Line (NCI-H209)**: 1,243 genes
- **3. Whole Genome screen of Melanoma Cell Line (COLO-829)**: 1,243 genes
- **4. Lung Rearrangement Screen (NCI-H2171, NCI-H1770)**: 1,243 genes
- **5. Structural Variant Screen**: 1,243 genes
- **6. Exome sequencing experiments**: 1,243 genes

**Data Analysis**

- **Distribution of Mutations along the protein sequence of gene CDKN2A**
- **Gene**: CDKN2A
- **Sample Source**: Tumours
- **Channels**: Mutations
- **Relative Type**: All mutations
- **Pie chart views**: Together with tables with statistics and links (approx) 1,212 genes represents almost 513,888 gene sequencing experiments, detailing 1,243 observed mutations.

**Auto Import**

**Data Mining : COSMICmart**

More extensive data mining options are available in the new COSMIC Biowidget which, whilst removing the graphical presentation, allows much more specialised queries to be investigated (http://www.sanger.ac.uk/genetics/CGP/cosmicBiowidget.html).

**Examining Data Offline**

Export custom searches from COSMIC
http://www.sanger.ac.uk/cosmic

Download prepared datasheets from COSMIC’s FTP site
ftp://ftp.sanger.ac.uk/pub/COSMIC/

Download whole datasets from the CGP Archive or EBI EGA Archive (including sequence traces & genotyping data, requires signed agreement)

http://www.ebi.ac.uk/ega/page.php?page=studies&name=CGP

**Whole Genome Sequencing**

This literature component is overviewed on a separate web page:
http://Genes.sanger.ac.uk/genetics/CGP/Classic

**Manuel Literature Curation**

Aiming to curate the world’s literature on somatic cancer mutations, COSMIC now holds full and up-to-date information point-mutated genes and gene fusion pairs, resulting from curation of 10,383 individual papers (COSMIC v49). This data is combined with the output of the CGP in the standard blue COSMIC web pages. Over the last 10 years, the information has been entered by dedicated curators with software checks at each datapoint for accuracy and integrity. This literature component is overviewed on a separate web page:
http://Genes.sanger.ac.uk/genetics/CGP/Classic

**Cancer Data from Consortium**

International Cancer Genome Consortium (ICGC) and the Cancer Genome Atlas (TCGA) projects. New tools are being developed to interpret this new genomic data with coding mutation annotations. In addition COSMIC will be expanded to curate and display data from mouse insertional mutagenesis screening and mouse cancer model exome/genome sequencing in the future.

Email: cosmos@sanger.ac.uk

---

**Cancer Genomic Annotation**

**Manual Literature Curation**

Systematic Screen Publications

- Manual Curation
- Semi-auto Curation
- Auto Import

Data Classification

Display in COSMIC

**Curating Systematic Screen Publications Semi-Automatically**

In the last few years, the improvement of sequencing technology has allowed the examination of thousands of candidate genes for potential involvement in cancer. Whilst such large screens by the CGP have been available in COSMIC for years, the global growth of publications documenting such analyses has encouraged the extension of COSMIC’s curation systems to encompass other publications of these large studies.


- **1. Breast Tumour Rearrangement Screen**: 1,212 genes
- **2. Whole Genome screen of Lung Cell Line (NCI-H209)**: 1,243 genes
- **3. Whole Genome screen of Melanoma Cell Line (COLO-829)**: 1,243 genes
- **4. Lung Rearrangement Screen (NCI-H2171, NCI-H1770)**: 1,243 genes
- **5. Structural Variant Screen**: 1,243 genes
- **6. Exome sequencing experiments**: 1,243 genes

**Data Analysis**

- **Distribution of Mutations along the protein sequence of gene CDKN2A**
- **Gene**: CDKN2A
- **Sample Source**: Tumours
- **Channels**: Mutations
- **Relative Type**: All mutations
- **Pie chart views**: Together with tables with statistics and links (approx) 1,212 genes represents almost 513,888 gene sequencing experiments, detailing 1,243 observed mutations.

**Auto Import**

**Data Mining : COSMICmart**

More extensive data mining options are available in the new COSMIC Biowidget which, whilst removing the graphical presentation, allows much more specialised queries to be investigated (http://www.sanger.ac.uk/genetics/CGP/cosmicBiowidget.html).

**Examining Data Offline**

Export custom searches from COSMIC
http://www.sanger.ac.uk/cosmic

Download prepared datasheets from COSMIC’s FTP site
ftp://ftp.sanger.ac.uk/pub/COSMIC/

Download whole datasets from the CGP Archive or EBI EGA Archive (including sequence traces & genotyping data, requires signed agreement)

http://www.ebi.ac.uk/ega/page.php?page=studies&name=CGP

**Whole Genome Sequencing**

This literature component is overviewed on a separate web page:
http://Genes.sanger.ac.uk/genetics/CGP/Classic