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
N Bindal, A Webb, S Bamford, C Cole, P Campbell, MR Stratton, PA Futreal
Cancer Genome Project, The Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge ,CB10 1SA, UK

Nature Precedings : doi:10.1038/npre.2010.5106.1 : Posted 25 Oct 2010

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 adapted 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 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: cosmic@sanger.ac.uk

COSMIC v49 Total Statistics (Sep 2010)	
Experiments	2,888,511
Tumours	548,399
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

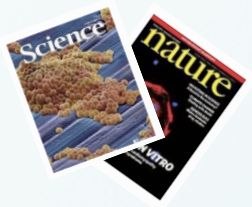
Export custom searches from COSMIC
http://www.sanger.ac.uk/cosmic

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

Download whole datasets from the CGP Archive or EBI EGA Archive (including sequence traces & genotyping data, requires signed agreement)
http://www.sanger.ac.uk/genetics/CGP/Archive
http://www.ebi.ac.uk/ega/page.php?page=studies&name=CGP

CANCER GENOMIC ANNOTATION

Manual Literature Curation



Systematic Screen Publications

Cancer Genome Project (CGP)

High Throughput Sequencing CGP lab in Sanger Institute

Capillary Screen Data
1. Protein Kinase Gene Analysis
2. Pilot Matched Pair Cell Lines
3. Renal Matched Pair Tumour Set
4. Adenoid Cystic Carcinoma Screen

Whole Genome Analysis
1. Breast Tumour Rearrangement Screen
2. Whole Genome screen of Melanoma Cell Line (COLO-829)
3. Whole Genome screen of Lung Cell Line (NCI-H209)
4. Lung Rearrangement Screen (NCI-H2171, NCI-H1770)

Sequencing Experiments	703,872
Tumours	3190
Genes	5,255
Mutations	4,885

Cancer Data from Consortia

International Cancer Genome Consortium (ICGC)

The Cancer Genome Atlas (TCGA)

Sequecing Experiment	513,888
Glioblastoma Samples	424
Genes	1,212
Mutations	1,243

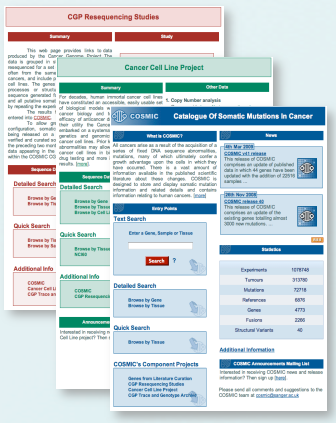
Manual Curation

Semi-auto Curation

Auto Import

Semi-auto Curation

Data Classification



Display in COSMIC

Cancer Data Standardisation

Tumour Classification
Cosmic's Phenotype definition system allows layered and very precise classification for each tumour.

Category	Entry
Primary Site; Subtype1; Subtype 2; Subtype 3	breast; NS; NS; NS
Primary Histology; Subtype1; Subtype 2; Subtype 3	carcinoma; ductal carcinoma; NS; NS

HGVS Nomenclature
The mutation syntax is derived from the HGVS nomenclature recommendations.

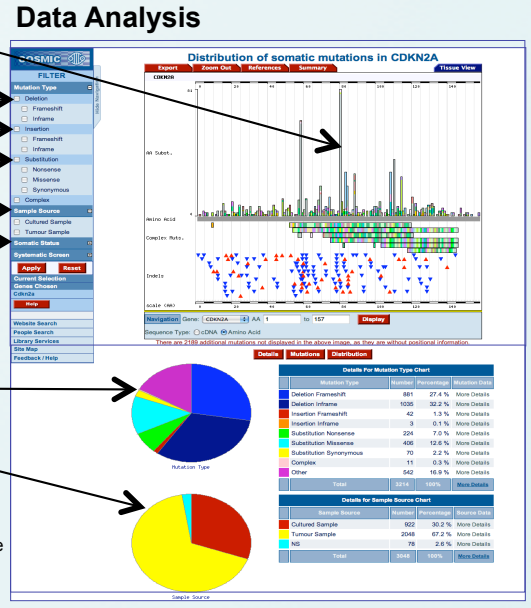
Sample	Gene	AA Mutation	CDS Mutation
1235097	ACCS	p.G221E	c.662G>A
1235097	CUBN	p.I3189V	c.9565A>G

Genomic Reference Sequence
All mutations on COSMIC are currently mapped to human GRCh37

Data Analysis

Distribution of Mutations along the protein sequence of gene CDKN2A

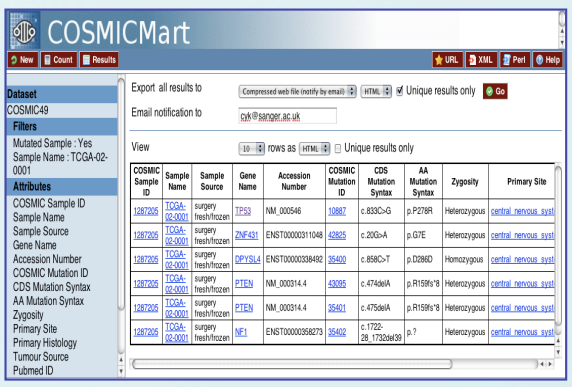
Deletion
Insertion
Substitution
Sample Source
Somatic Status



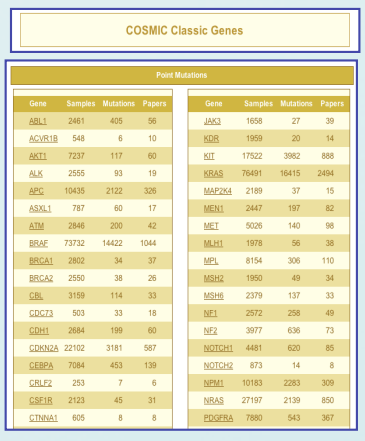
Mutation Type
Sample Source

Pie chart views (together with tables with statistics and links) of the spread of mutation types, together with the relative quantities of cultured and primary tumour samples.

Data Mining : COSMICMart



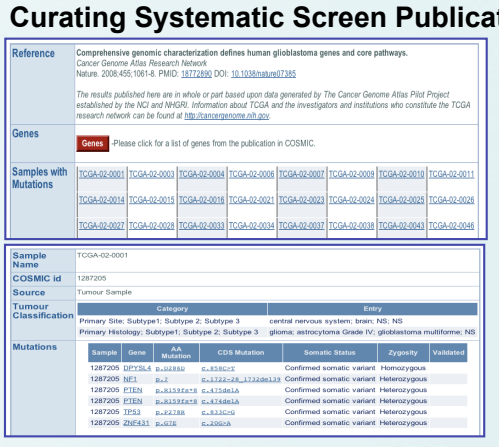
More extensive data mining options are available in the new COSMIC Biomart which, whilst removing the graphical presentation, allows much more specialised queries to be investigated (http://www.sanger.ac.uk/genetics/CGP/cosmic/biomart/martview/).



Manual 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://www.sanger.ac.uk/genetics/CGP/Classic

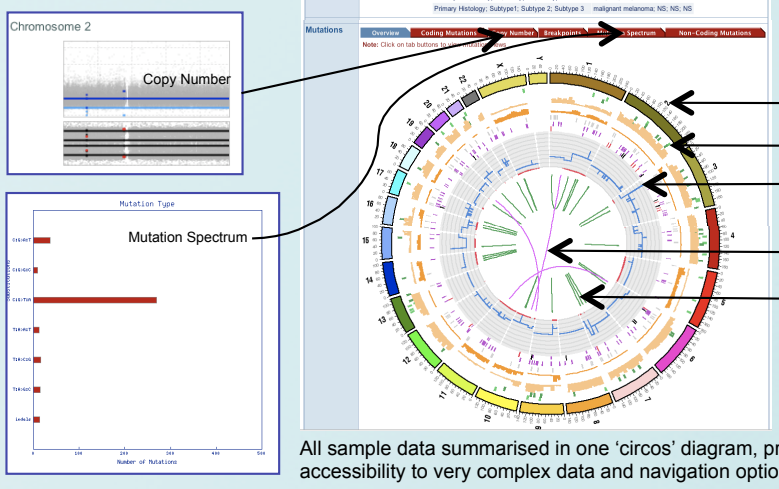
Curating Systematic Screen Publications Semi-Automatically



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 the phase I & II TCGA glioblastoma screen. Whilst in combination, these studies only examine 424 tumour samples, their analysis of up to (approx) 1,212 genes represents almost 513,888 gene sequencing experiments, detailing 1,243 observed mutations.

Whole Genome Sequencing



All sample data summarised in one 'circos' diagram, provides instant accessibility to very complex data and navigation options to further details.