



COSMIC

Catalogue Of Somatic Mutations In Cancer

What is COSMIC ?

COSMIC - the Catalogue of Somatic Mutations in Cancer - là nguồn dữ liệu lớn nhất trên thế giới về đột biến soma liên quan đến ung thư ở người, được chuyên gia tuyển chọn thủ công.

Dữ liệu này được mô tả về cấu trúc, nội dung và phạm vi. Nhờ đó, người dùng có sử dụng dữ liệu hiệu quả nhất.



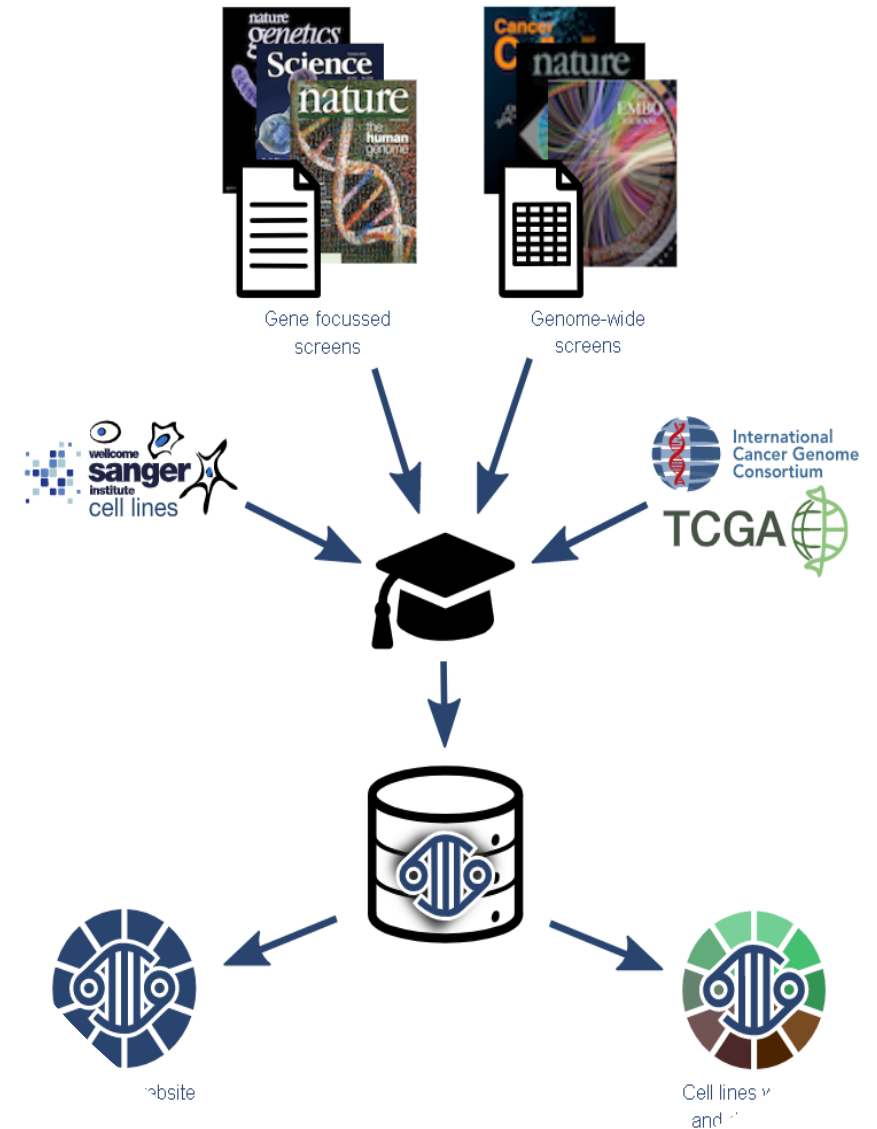
TS. Nguyễn Hồng Quang
Nguyễn Bá Thanh

Tổng quan

COSMIC bao gồm cơ sở dữ liệu COSMIC và Cell Lines Project, hai tài nguyên riêng biệt nhưng có liên quan với nhau.

Cơ sở dữ liệu COSMIC kết hợp hai loại dữ liệu chính:

- Dữ liệu chính xác cao (Targeted screen data): Tập trung vào gen và đột biến đã biết hoặc nghi ngờ là ung thư, được giám tuyển thủ công bởi các chuyên gia
- Dữ liệu sàng lọc rộng (**Genome-wide Screen Data**): hơn 37.000 bộ gen thông qua sàng lọc gen với quy mô lớn và từ các cơ sở dữ liệu sẵn có như TCGA và ICGC.



Hệ thống phân loại (Classification)

- Việc phân loại các loại khối u và các loại phụ có đột biến soma trong các tài liệu đã xuất bản là vô cùng khác nhau.
- COSMIC sử dụng hệ thống phân loại nội bộ của riêng mình để cung cấp tính nhất quán của mô và mô học trong cơ sở dữ liệu và giảm sự dư thừa.

Dưới đây là 1 ví dụ về việc chuyển đổi thông tin trong bài báo thành dữ liệu COSMIC

Type	Paper Definition	COSMIC Definition
Site primary	colon	large intestine
Site subtype 1	descending	colon
Site subtype 2	NS	descending
Histology	carcinoma	carcinoma
Histology subtype 1	polypoid type	adenocarcinoma
Histology subtype 2	with adenoma	NS
Histology subtype 3	NS	NS
Phenotype ID		COS028564826
NCI Code		C4349
EFO link		EFO_1001949

- Mô nguyên phát: Đại tràng => ruột già
- Phân nhóm mô mức 1: Tràng dưới => đại tràng
- Phân nhóm mô mức 2: Không chỉ định => tràng dưới
- Mô học: ung thư biểu mô => ung thư biểu mô
- Phân nhóm mô học mức 1: thể đa bội => ung thư biểu mô tuyến
- Phân nhóm mô học mức 2: u tuyến => không chỉ định

Classification Information

Bảng thông tin phân loại được
lưu trong file classification.csv

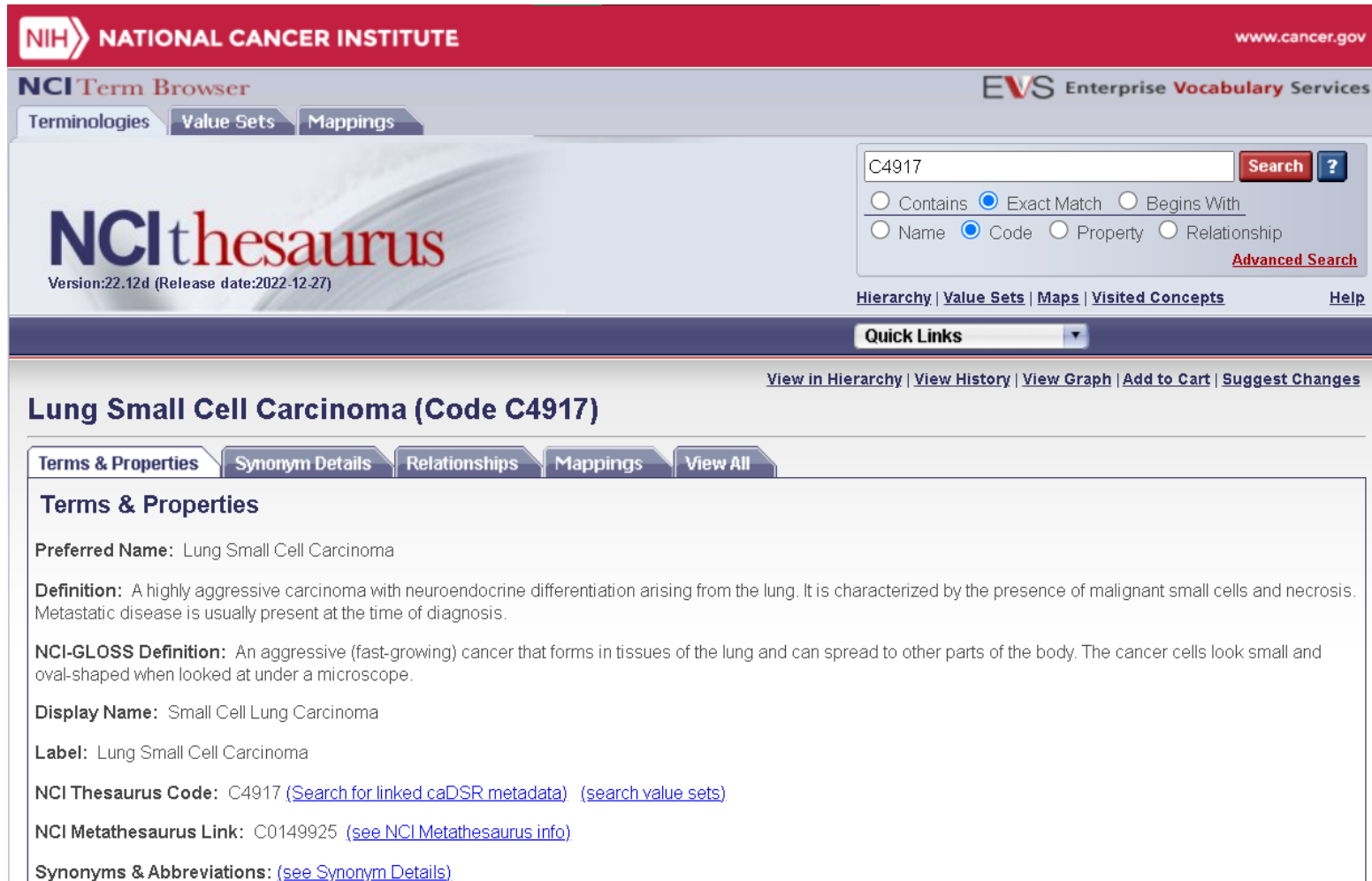
STT	Tên cột	Mô tả	Ví dụ
[1:A]	Cosmic Phenotype id	Mã định danh duy nhất cho phân loại	COSO29935333
[2:B]	Site Primary	Mô nguyên phát được mô tả trong bài báo	lung
[3:C]	Site Subtype	Phân loại mô mức 1 được mô tả trong bài báo	left_lower_lobe
[4:D]	Site Subtype 2	Phân loại mô mức 2 được mô tả trong bài báo	NS
[5:E]	Site Subtype 3	Phân loại mô mức 3 được mô tả trong bài báo	NS
[6:F]	Histology	Mô học gốc được mô tả trong bài báo	carcinoma
[7:G]	Hist Subtype 1	Phân loại mô học mức 1 được mô tả trong bài báo	small-cell
[8:H]	Hist Subtype 2	Phân loại mô học mức 2 được mô tả trong bài báo	NS
[9:I]	Hist Subtype 3	Phân loại mô học mức 3 được mô tả trong bài báo	NS

Classification Information

STT	Tên cột	Mô tả	Ví dụ
[10:J]	Site Primary COSMIC	Mô nguyên phát được chỉ định trong COSMIC	lung
[11:K]	Site_Subtype1_COSMIC	Mô phụ cấp 1 được chỉ định trong COSMIC	left_lower_lobe
[12:L]	Site_Subtype2_COSMIC	Mô phụ cấp 2 được chỉ định trong COSMIC	NS
[13:M]	Site_Subtype3_COSMIC	Mô phụ cấp 3 được chỉ định trong COSMIC	NS
[14:N]	Histology_COSMIC	Mô học nguyên phát được chỉ định trong COSMIC	carcinoma
[15:O]	Hist_Subtype1_COSMIC	Mô học phụ cấp 1 được chỉ định trong COSMIC	small_cell_carcinoma
[16:P]	Hist_Subtype2_COSMIC	Mô học phụ cấp 2 được chỉ định trong COSMIC	NS
[17:Q]	Hist_Subtype3_COSMIC	Mô học phụ cấp 3 được chỉ định trong COSMIC	NS
[18:R]	NCI code	Mã NCI tham chiếu	C4917
[19:S]	EFO code	Mã EFO tham chiếu	EFO_0000702

- <https://ncit.nci.nih.gov/ncitbrowser>
- <http://www.ebi.ac.uk/efo>

NCI và EFO



The screenshot displays the NCI Term Browser interface. At the top, the NIH National Cancer Institute logo and the URL www.cancer.gov are visible. The page title is "NCI Term Browser" with sub-tabs for "Terminologies", "Value Sets", and "Mappings". The "NCI thesaurus" logo is prominently displayed, along with the version information: "Version:22.12d (Release date:2022-12-27)".

A search bar contains the code "C4917". To the right of the search bar are radio buttons for search criteria: "Contains", "Exact Match" (selected), "Begins With", "Name", "Code" (selected), "Property", and "Relationship". A "Search" button and a "Help" icon are also present. Below the search bar is a link to "Advanced Search".

Navigation links include "Hierarchy", "Value Sets", "Maps", "Visited Concepts", and "Help". A "Quick Links" dropdown menu is located below these links.

The main content area is titled "Lung Small Cell Carcinoma (Code C4917)". Above this title are links: "View in Hierarchy", "View History", "View Graph", "Add to Cart", and "Suggest Changes". Below the title is a tabbed interface with "Terms & Properties" (selected), "Synonym Details", "Relationships", "Mappings", and "View All".

The "Terms & Properties" section contains the following information:

- Preferred Name:** Lung Small Cell Carcinoma
- Definition:** A highly aggressive carcinoma with neuroendocrine differentiation arising from the lung. It is characterized by the presence of malignant small cells and necrosis. Metastatic disease is usually present at the time of diagnosis.
- NCI-GLOSS Definition:** An aggressive (fast-growing) cancer that forms in tissues of the lung and can spread to other parts of the body. The cancer cells look small and oval-shaped when looked at under a microscope.
- Display Name:** Small Cell Lung Carcinoma
- Label:** Lung Small Cell Carcinoma
- NCI Thesaurus Code:** C4917 ([Search for linked caDSR metadata](#)) ([search value sets](#))
- NCI Metathesaurus Link:** C0149925 ([see NCI Metathesaurus info](#))
- Synonyms & Abbreviations:** ([see Synonym Details](#))

[Home](#)[Ontologies](#)[Documentation](#)[About](#)

- <https://ncit.nci.nih.gov/ncitbrowser>
- <http://www.ebi.ac.uk/efo>

EFO_0000702

Search

☐ Exact match ☐ Obsolete terms

Term type

Filter by type

class

40

Ontology

× Experimental Factor Ontology (EFO)

EFO

40

Clear all filters

Search results for EFO_0000702

[Previous](#)

Showing 1 to 10 of 40 results

[Next](#)

small cell lung carcinoma EFO:0000702

http://www.ebi.ac.uk/efo/EFO_0000702

Small cell lung cancer (SCLC) is a highly aggressive malignant neoplasm, accounting for 10-15% of lung cancer cases, characterized by rapid growth, and early metastasis. SCLC usually manifests as a large hilar mass with bulky mediastinal lymphadenopathy presenting clinically with chest pain, persistent cough, dyspnea, wheezing, hoarseness, hemoptysis, loss of appetite, weight loss, and neurological and endocrine paraneoplastic syndromes. SCLC is primarily reported in elderly people with a history...

Ontology: [Experimental Factor Ontology](#) EFO

HCC0033 EFO:0006417

http://www.ebi.ac.uk/efo/EFO_0006417

Human small cell lung carcinoma cell line from a 53-year-old Caucasian male patient

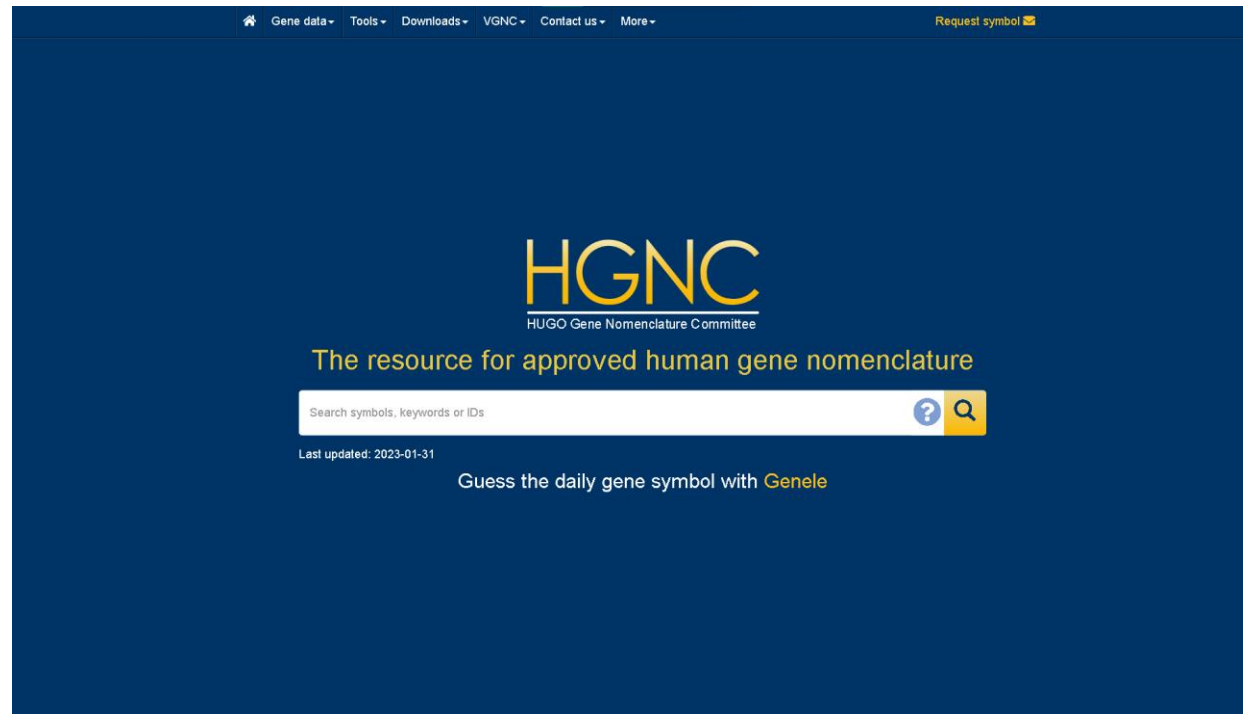
COSMIC Complete Mutation Data (Targeted Screens)

STT	Tên cột	Mô tả	Ví dụ
[1:A]	Gene name	Tên gen (theo HGNC)	CD58
[2:B]	Accession Number	Mã định danh phiên mã của gen	ENST00000369489.9
[3:C]	Gene CDS length	Chiều dài gen (số lượng cặp cơ sở)	753
[4:D]	HGNC id	Mã định danh HGNC	1688
[5:E]	Sample name, Sample id, Id tumour *	Một mẫu là một ví dụ của một phần của khối u đang được kiểm tra đột biến.	P766 2878555 2733070

HGNC

- <https://www.genenames.org/>

HGNC chịu trách nhiệm phê duyệt các ký hiệu và tên duy nhất cho các locus của con người, bao gồm các gen mã hóa protein, gen ncRNA và gen giả, để cho phép giao tiếp khoa học rõ ràng.



Kết quả tìm kiếm

Lọc theo loại

- gen 575
- Nhóm 3
- Địa điểm 184

Lọc theo trạng thái nhập gen

- Tán thành 5
- Rút tiền 570

Lọc theo loại locus gen

- gen mã hóa protein 5
- kiểu hình 569
- Khác 1
 - Khu vực 1

Lọc theo thẻ báo cáo biểu tượng

- ký hiệu giữ chỗ 4

Tải xuống tất cả kết quả



20 items per page ▾

Vật phẩm: 1 đến 20 của 762

Trang 1

của 39

Kế tiếp ▶

Cuối cùng ▶▶

CD58 : phân tử CD58

gen ID HGNC HGNC:1688 **Loại** locus Gene với sản phẩm protein **Trạng thái** Đã phê duyệt
Diêm ID HGNC: **HGNC:1688**
Liên minh tài nguyên gen: **HGNC:1688**

Họ phân tử kết dính tế bào giống Ig

Nhóm **Nhóm HGNC ID** 1410
Diêm ID HGNC: **HGNC:1688**

phân tử CD

Nhóm **Nhóm HGNC ID** 471
Diêm ID HGNC: **HGNC:1688**

ACLS : hội chứng acrocallosal

gen **Mã** HGNC HGNC:114 **Loại** locus Chỉ kiểu hình **Trạng thái** Mục nhập Đã rút tiền
Diêm Ghi chú của người quản lý: **HGNC** không còn duy trì ký hiệu cho kiểu hình.

AFD1 : acrofacial dysostosis 1, loại Nager

gen **Mã** HGNC HGNC:312 **Loại** locus Chỉ kiểu hình **Trạng thái** Mục nhập Đã rút tiền
Diêm Ghi chú của người quản lý: **HGNC** không còn duy trì ký hiệu cho kiểu hình.

AGMX2 : agammaglobulinemia, X-linked 2 (thiếu hormone tăng trưởng)

gen **Mã** HGNC HGNC:323 **Loại** locus Chỉ kiểu hình **Trạng thái** Mục nhập Đã rút tiền
Diêm Ghi chú của người quản lý: **HGNC** không còn duy trì ký hiệu cho kiểu hình.

AIS1 : tính nhạy cảm tự miễn dịch 1

gen ID HGNC HGNC:16304 **Loại** locus Chỉ kiểu hình **Trạng thái** Mục nhập Đã rút tiền
Diêm Ghi chú của người quản lý: **HGNC** không còn duy trì ký hiệu cho kiểu hình.

Sample name
Sample id
Id tumour

Tên mẫu có thể được lấy từ một số nguồn. Trong nhiều trường hợp nó bắt nguồn từ tên dòng tế bào.

Một số mẫu có thể được lấy từ một khối u và một số khối u có thể được lấy từ một cá nhân.

Id mẫu được sử dụng để xác định một mẫu trong cơ sở dữ liệu COSMIC.

Một mẫu có thể có nhiều id nếu được nhập từ nhiều bài báo khác nhau

COSMIC Complete Mutation Data

- Thông tin phân loại mô và mô học

STT	Tên cột	Mô tả	Ví dụ
[8:H]	Primary Site	Mô nguyên phát	haematopoietic_and_lymphoid_tissue
[9:I]	Site Subtype 1	Phân loại mô mức 1	NS
[10:J]	Site Subtype 2	Phân loại mô mức 2	NS
[11:K]	Site Subtype 3	Phân loại mô mức 3	NS
[12:L]	Primary Histology	Mô học gốc	lymphoid_neoplasm
[13:M]	Histology Subtype 1	Phân loại mô học mức 1	diffuse_large_B_cell_lymphoma
[14:N]	Histology Subtype 2	Phân loại mô học mức 2	NS
[15:O]	Histology Subtype 3	Phân loại mô học mức 3	NS
[16:P]	Genome-wide screen	Liệu toàn bộ gen/exome được giải trình tự? (y/n)	n

COSMIC Complete Mutation Data

Thông tin về định danh và
biểu hiện của đột biến

STT	Tên cột	Mô tả	Ví dụ
[17:Q]	GENOMIC_MUTATION_ID	Định danh đột biến gen (COSV) để chỉ vị trí xác định của biến thể trên bộ gen	COSV105299677
[18:R]	LEGACY_MUTATION_ID	Định danh đột biến kế thừa tương ứng (COSM)	COSM9866493
[19:S]	MUTATION_ID	Định danh đột biến nội bộ đại diện duy nhất một đột biến trong một bản phiên mã cụ thể	110346783
[20:T]	Mutation CDS	Thay đổi trong trình tự nucleotit.	c.91C>T
[21:U]	Mutation AA	Thay đổi trong trình tự peptit.	p.Q31*
[22:V]	Mutation Description	Loại đột biến ở mức độ axit amin	Substitution - Nonsense

Định danh đột biến

Legacy Mutation Id: Tất cả các id COSM ở cùng một vị trí bộ gen đã được thu gọn thành một id COSM đại diện. Các id này được duy trì để giúp theo dõi các đột biến hiện có.

COSM9866493

Genomic Mutation Id: Số nhận dạng này có thể theo dõi và ổn định giữa các phiên bản khác nhau của bản phát hành.

COSV105299677

Mutation Id: Đây là các mã định danh nội bộ duy nhất cho một đột biến trên một phiên mã cụ thể và được hiển thị trong URL của các trang đột biến.

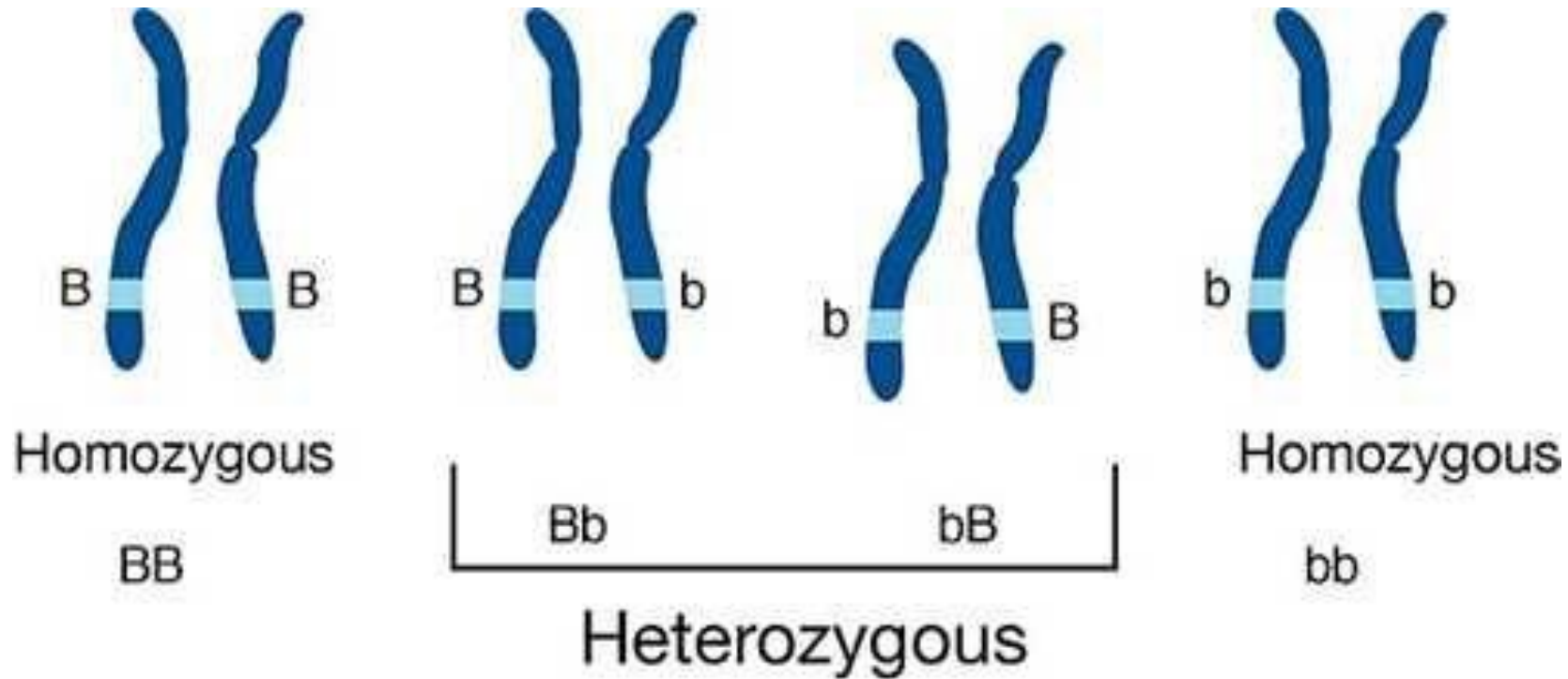
<https://cancer.sanger.ac.uk/cosmic/mutation/overview?id=110346783>

[...?id=110343601](#)

[..?id=134637724](#)

* Một id COSV có thể được liên kết với nhiều id nội bộ và id COSM. Tương tự, một id COSM cũng có thể liên kết tới nhiều id nội bộ.

STT	Tên cột	Mô tả	Ví dụ
[23:W]	Mutation zygosity	Thông tin về việc đột biến được báo cáo là đồng hợp tử , dị hợp tử hay không xác định	
[24:X]	LOH	Thông tin về việc gen được báo cáo là mất tính dị hợp tử trong mẫu: y/n/(unknown)	
[25:Y]	GRCh	Phiên bản gen được sử dụng 38 = GRCh38/Hg38 37 = GRCh37/Hg19	38
[26:Z]	Mutation genome position	Tọa độ xảy ra đột biến	1:116544584-116544584
[27:AA]	Mutation strand	đột biến nằm ở chuỗi + hay -	-
[28:AB]	Resistance Mutation	Đột biến gây ra tình trạng kháng thuốc	-
[29:AC]	Mutation somatic status	Tình trạng soma của đột biến	Variant of unknown origin



Trạng thái hợp tử

- Đồng hợp tử (Homozygous): các alen giống nhau nằm ở các vị trí tương ứng trên cặp nhiễm sắc thể.
- Dị hợp tử (Heterozygous): Các alen khác nhau nằm ở các vị trí tương ứng trên cặp nhiễm sắc thể.
- Mất tính dị hợp tử (Loss of heterozygous - LOH): các biến đổi trong gen làm mất tính dị hợp tử vốn có.

Về tình trạng soma của đột biến

Previously observed: Đã được báo cáo là đột biến soma trước đó trong bài báo (mẫu) khác

Confirmed Somatic: Đã được xác nhận là đột biến soma qua thực nghiệm bằng cách giải trình tự khối u và tế bào bình thường tương ứng

Variant of unknown origin: Biến thể không rõ nguồn gốc, khi đột biến là soma, nhưng khối u được giải trình tự mà không có tế bào bình thường tương ứng.

STT	Tên cột	Mô tả	Ví dụ
[30:AD]	Pubmed_PMID	Mã định danh PUBMED của bài báo mà mẫu được báo cáo	32187361
[31:AE]	Id Study	Danh sách nghiên cứu liên quan đến mẫu	surgery-fixed
[32:AF]	Sample Type, Tumour origin	Mô tả nguồn gốc mẫu, loại khối u	primary
[34:AH]	Age	Tuổi của cá nhân được lấy mẫu	
[35:AI]	HGVSP	Danh pháp HGVS cho trình tự tham chiếu protein	ENSP00000358501.5:p.Gln31Ter
[36:AJ]	HGVSC	Danh pháp HGVS cho trình tự tham chiếu DNA mã hóa	ENST00000369489.9:c.91C>T
[37:AK]	HGVSG	Danh pháp HGVS cho trình tự tham chiếu bộ gen	1:g.116544584G>A

COSMIC Complete Mutation Data

Về danh pháp HGVS:
<https://varnomen.hgvs.org/>

COSMIC Mutation Data (Genome Screens)

Dữ liệu đột biến
điểm từ sàng lọc
toàn bộ gen được
lưu trong bảng

STT	Tên cột	Ví dụ
[1:A]	Gene name	LTA4H
[2:B]	Accession Number	ENST00000228740.6
[3:C]	Gene CDS length	1836
[4:D]	HGNC id	6710
[5:E]	Sample name, Sample id, Id tumour	BN01T 1919132 1806453
[8:H]	Primary Site	liver
[9:I]	Site Subtype 1	NS
[10:J]	Site Subtype 2	NS
[11:K]	Site Subtype 3	NS
[12:L]	Primary Histology	carcinoma
[13:M]	Histology Subtype 1	NS
[14:N]	Histology Subtype 2	NS
[15:O]	Histology Subtype 3	NS
[16:P]	Genome-wide screen	y

COSMIC Mutation Data (Genome Screens)

STT	TÊN CỘT	VÍ DỤ
[17:Q]	MUTATION_ID	100965114
[18:R]	GENOMIC_MUTATION_ID	COSV57383335
[19:S]	LEGACY_MUTATION_ID	COSN1154092
[20:T]	Mutation CDS	c.1379+30C>G
[21:U]	Mutation AA	p.?
[22:V]	Mutation Description	Unknown
[23:W]	Mutation zygosity	het
[24:X]	LOH	
[25:Y]	GRCh	38
[26:Z]	Mutation genome position	12:96013158-96013158
[27:AA]	Mutation strand	-
[29:AC]	Pubmed_PMID	
[30:AD]	Id Study	323
[34:AH]	HGVSP	
[35:AI]	HGVSC	ENST00000228740.6:c.1379+30C>G
[36:AJ]	HGVSG	12:g.96013158G>C

COSMIC Mutation Data

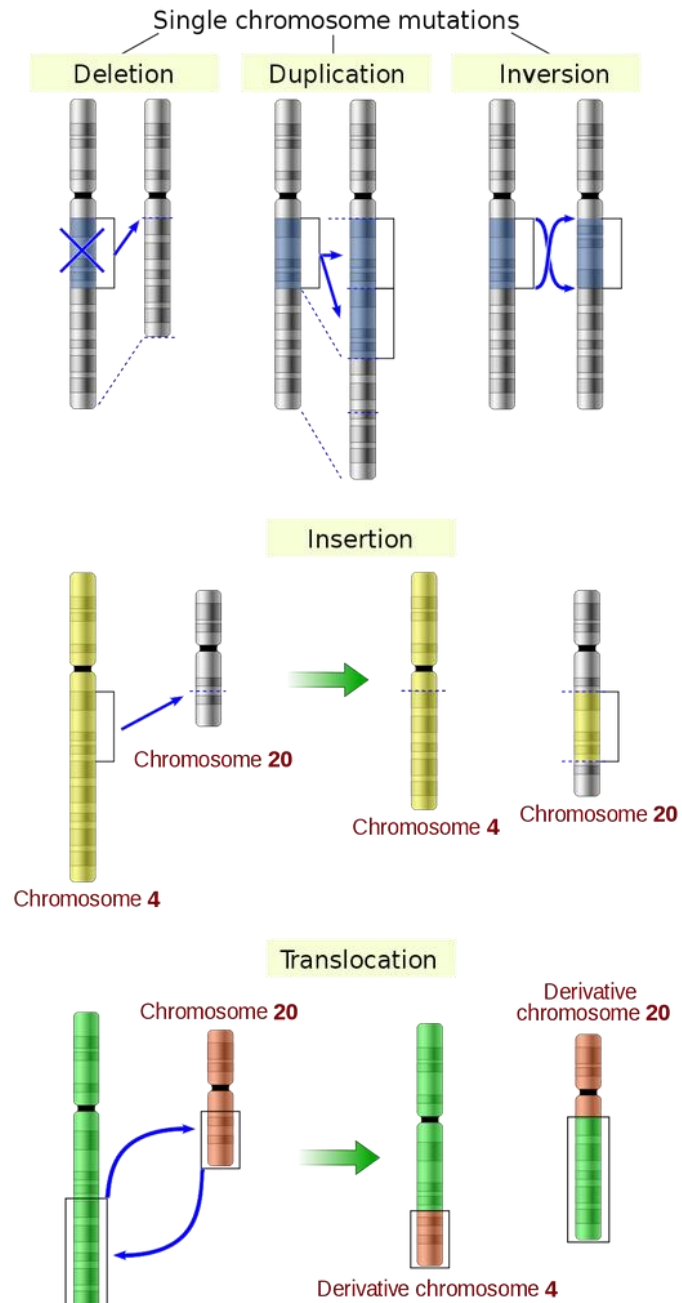
STT	Tên cột	Ví dụ
[1:A]	Gene name	TEX15_ENST00000256246
[2:B]	Accession Number	ENST00000256246.5
[3:C]	Gene CDS length	8370
[4:D]	HGNC id	11738
[5:E]	Sample name, Sample id, Id tumour	TCGA-91-6848-01 1914095 1802377
[8:H]	Primary Site	lung
[9:I]	Site Subtype 1	right_upper_lobe
[10:J]	Site Subtype 2	NS
[11:K]	Site Subtype 3	NS
[12:L]	Primary Histology	carcinoma
[13:M]	Histology Subtype 1	adenocarcinoma
[14:N]	Histology Subtype 2	NS
[15:O]	Histology Subtype 3	NS
[16:P]	Genome-wide screen	y
[17:Q]	GENOMIC_MUTATION_ID	COSV56354944
[18:R]	LEGACY_MUTATION_ID	COSM6180628
[19:S]	MUTATION_ID	100197588

COSMIC Mutation Data

STT	Tên cột	Ví dụ
[20:T]	Mutation CDS	c.6327G>T
[21:U]	Mutation AA	p.M2109I
[22:V]	Mutation Description	Substitution - Missense
[23:W]	Mutation zygosity	
[24:X]	LOH	
[25:Y]	GRCh	38
[26:Z]	Mutation genome position	8:30842691-30842691
[27:AA]	Mutation strand	-
[28:AB]	Resistance Mutation	-
[29:AC]	Mutation somatic status	Variant of unknown origin
[30:AD]	Pubmed_PMID	
[31:AE]	Id Study	417
[32:AF]	Sample Type, Tumour origin	fresh/frozen – NOS primary
[34:AH]	Age	59
[35:AI]	HGVSP	ENSP00000256246.2:p.Met2109Ile
[36:AJ]	HGVSC	ENST00000256246.5:c.6327G>T
[37:AK]	HGVSG	8:g.30842691C>A

Structural Genomic Rearrangements

Genomic Rearrangements là các biến thể trong trình tự bộ gen trên quy mô lớn, thường lớn hơn 1 Kbp (kilo base pair) gây ra bởi sự sao chép, xóa, chèn, đảo ngược hoặc chuyển vị một đoạn gen.



STRUCTURAL VARIANTS

- Mutation id: mã định danh duy nhất xác định đột biến.
- Mutation type: Intra/Inter (chromosomal), tandem duplication, deletion, inversion, complex substitutions, complex amplicons.
- Description: Mô tả về đột biến cấu trúc, theo khuyến nghị của HGVS

https://cancer.sanger.ac.uk/cosmic/help/rearrangement/overview#overview_tab_content

STT	Tên cột	Ví dụ
[1:A]	Sample name, Sample id, Id tumour	0001_CRUK_PC_0001_T1_DNA 2457570 2320407
[4:D]	Primary Site	prostate
[5:E]	Site Subtype 1	NS
[6:F]	Site Subtype 2	NS
[7:G]	Site Subtype 3	NS
[8:H]	Primary Histology	liver
[9:I]	Histology Subtype 1	carcinoma
[10:J]	Histology Subtype 2	adenocarcinoma
[11:K]	Histology Subtype 3	NS
[12:L]	Mutation Id	720948
[13:M]	Mutation Type	intrachromosomal with non-inverted orientation
[14:N]	GRCh	38
[15:O]	Description	chr3:g.76398933_chr3:77955758bkpt
[16:P]	Pubmed_PMID	
[17:Q]	ID_STUDY	538

BREAKPOINTS

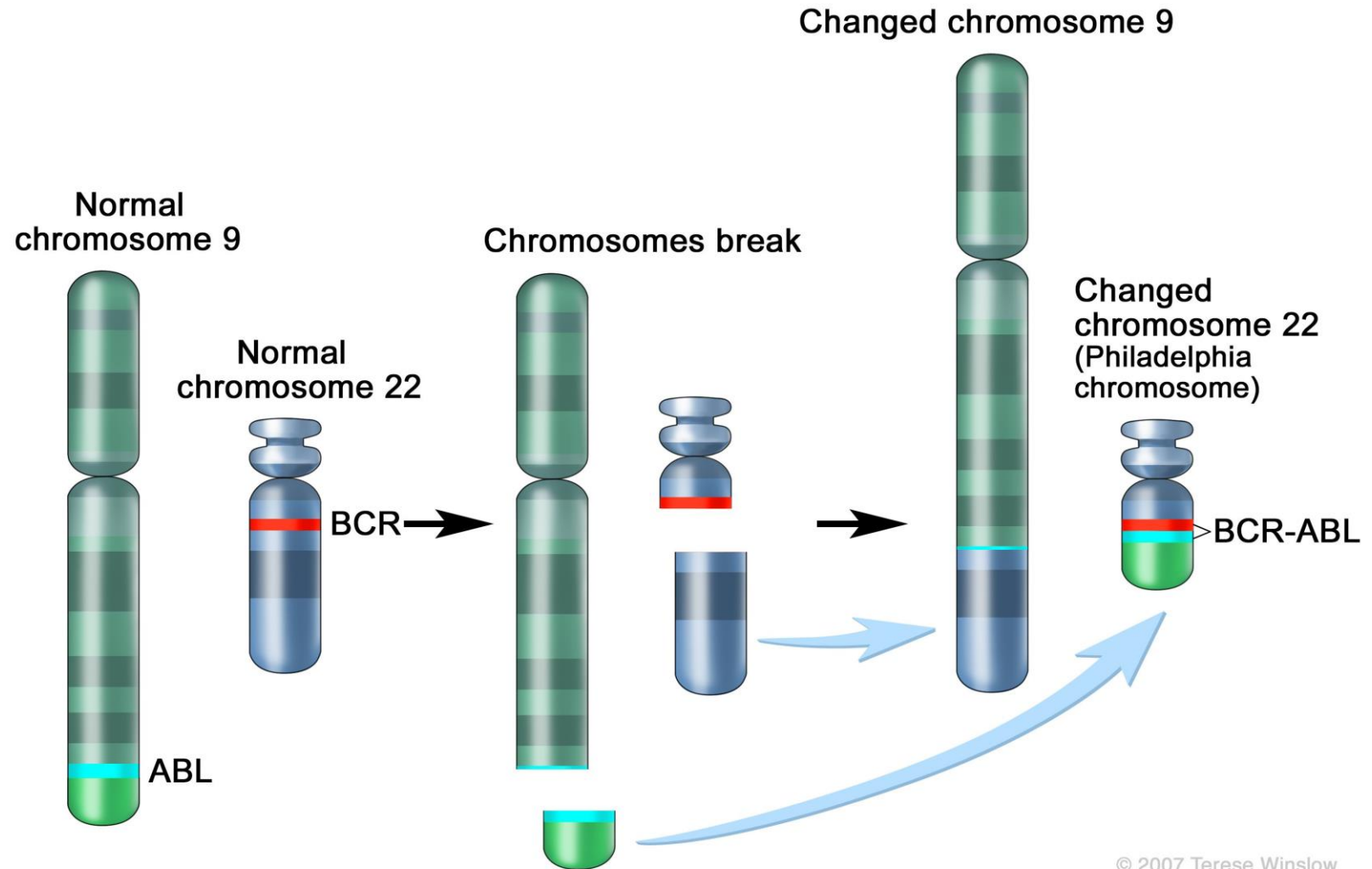
STT	TÊN CỘT	VÍ DỤ
[1:A]	Sample name, Sample id, Id tumour	0001_CRUK_PC_0001_T1_DNA 2457570 2320407
[4:D]	Primary Site	prostate
[5:E]	Site Subtype 1	NS
[6:F]	Site Subtype 2	NS
[7:G]	Site Subtype 3	NS
[8:H]	Primary Histology	liver
[9:I]	Histology Subtype 1	carcinoma
[10:J]	Histology Subtype 2	adenocarcinoma
[11:K]	Histology Subtype 3	NS
[12:L]	Mutation Type	Interchromosomal unknown type
[13:M]	Mutation Id	748120
[14:N]	Breakpoint Order	
[15:O]	GRCh	38

BREAKPOINTS

STT	Tên cột	Ví dụ
[16:p]	Chrom From	3
[17:Q]	Location From min	165212802
[18:R]	Location From max	165212802
[19:S]	Strand From	-
[20:T]	Chrom To	4
[21:U]	Location To min	64685047
[22:V]	Location To max	64685047
[23:W]	Strand To	-
[24:X]	Non-templated ins seq	
[25:Y]	Pubmed_PMID	
[26:Z]	Id Study	538

Complete Fusion Export

- Gen tổng hợp (Fusion gene) là gen được nối từ các phần của 2 gen khác nhau. Gen tổng hợp được hình thành khi một phần DNA của một nhiễm sắc thể di chuyển tới một nhiễm sắc thể khác.
- Gen nằm gần đầu 5' gọi là gen 5', gen nằm gần đầu 3' gọi là gen 3'.
- Ví dụ: Gen tổng hợp BCR-ABL được hình thành khi các đoạn nhiễm sắc thể số 9 và 22 bị đứt ra và đổi chỗ cho nhau.



STT	Tên cột	Ví dụ
[1:A]	Sample name, Sample id	749711 HCC1187
[3:C]	Primary Site	breast
[4:D]	Site Subtype 1	NS
[5:E]	Site Subtype 2	NS
[6:F]	Site Subtype 3	NS
[7:G]	Primary Histology	carcinoma
[8:H]	Histology Subtype 1	ductal_carcinoma
[9:I]	Histology Subtype 2	NS
[10:J]	Histology Subtype 3	NS
[11:K]	Fusion Id	665
[12:L]	Translocation name	ENST00000360863.10(RGS22):r.1_3555::ENST00000369518.1(SYCP1):r.2100_3452
[13:M]	5'_CHROMOSOME	8
[14:N]	5'_STRAND	-
[15:O]	5'_GENE_ID	197199
[16:P]	5'_GENE_NAME	RGS22
[17:Q]	5'_LAST_OBSERVED_EXON	22

Complete Fusion Export

- Dữ liệu về đột biến gen tổng hợp được lưu trong file CosmicFusionExport.tsv.gz

Complete Fusion Export

- Tổng quan:
</cosmic/fusion>
- Ví dụ về gen id:
</cosmic/fusion/overview?fid=212470&gid=197199>

STT	Tên cột	Ví dụ
[18:R]	5'_GENOME_START_FROM	99981937
[19:S]	5'_GENOME_START_TO	99981937
[20:T]	5'_GENOME_STOP_FROM	100106116
[21:U]	5'_GENOME_STOP_TO	100106116
[22:V]	3'_CHROMOSOME	1
[23:W]	3'_STRAND	+
[24:X]	3'_GENE_ID	212470
[25:Y]	3'_GENE_NAME	SYCP1_ENST00000369518
[26:Z]	3'_FIRST_OBSERVED_EXON	24
[27:AA]	3'_GENOME_START_FROM	114944339
[28:AB]	3'_GENOME_START_TO	114944339
[29:AC]	3'_GENOME_STOP_FROM	114995367
[30:AD]	3'_GENOME_STOP_TO	114995367
[31:AE]	Fusion type	Inferred Breakpoint
[32:AF]	Pubmed_PMID	20033038

All Mutations in Census Genes

- Dữ liệu đột biến gen trong danh sách Cancer Gene Census được lưu trong file CosmicMutantExportCensus.tsv.gz
- Điều tra gen ung thư (the Cancer Gene Census - CGC) là một dự án nhằm lập danh mục những gen chứa đột biến dẫn đến ung thư và giải thích hoạt động ở những gen này.
- Về dự án Cancer Gene Census: [/cosmic/census](https://cancer.sanger.ac.uk/cgc/)

STT	Tên cột	Mô tả	Ví dụ
[1:A]	Gene name	Tên gen(theo HGNC)	ERBB4
[2:B]	Accession Number	Mã định danh phiên mã của gen	ENST00000342788.8
[3:C]	Gene CDS length	Chiều dài gen(số lượng cặp cơ sở)	12136
[4:D]	HGNC id	Mã định danh HGNC	3432
[5:E]	Sample name, Sample id, Id tumour	Một mẫu là một ví dụ của một phần của khối u đang được kiểm tra đột biến.*	PD23569a 2213156 2081286

All Mutations in Census Genes

Dữ liệu về đột biến gen
trong danh sách Cancer
Gene Census được lưu
trong file
CosmicMutantExportCensus.
tsv.gz

STT	Tên cột	Ví dụ
[8:H]	Primary Site	breast
[9:I]	Site Subtype 1	NS
[10:J]	Site Subtype 2	NS
[11:K]	Site Subtype 3	NS
[12:L]	Primary Histology	carcinoma
[13:M]	Histology Subtype 1	ductal_carcinoma
[14:N]	Histology Subtype 2	NS
[15:O]	Histology Subtype 3	NS
[16:P]	Genome-wide screen	Y
[17:Q]	GENOMIC_MUTATION_ID	COSV53530422
[18:R]	LEGACY_MUTATION_ID	COSN20636712
[19:S]	MUTATION_ID	106295671
[20:T]	Mutation CDS	c.82+137463T>G
[21:U]	Mutation AA	p.?
[22:V]	Mutation Description	Unknown

Cancer Gene Cencus

Hai cấp độ của dự án:

- Cấp 1: gen được ghi nhận là liên quan đến ung thư, cùng với bằng chứng về đột biến làm thúc đẩy quá trình gây ung thư
- Cấp 2: gen mới được thêm vào dự án, có dấu hiệu rõ ràng nhưng có ít bằng chứng

STT	Tên cột	Ví dụ
[23:W]	Mutation zygosity	
[24:X]	LOH	
[25:Y]	GRCh	38
[26:Z]	Mutation genome position	1:116544584-116544584
[27:AA]	Mutation strand	-
[28:AB]	Resistance Mutation	-
[29:AC]	Mutation somatic status	Confirmed somatic variant
[30:AD]	Pubmed_PMID	
[31:AE]	Id Study	652
[32:AF]	Sample Type,Tumour origin	NS NS
[34:AH]	Age	
[35:AI]	Tier	1
[36:AJ]	HGVSP	
[37:AK]	HGVSC	ENST00000342788.8:c.82+137463T>G
[38:AL]	HGVSG	2:g.212400986A>C

Non-coding variants

- Dữ liệu về đột biến không mã hóa được lưu trong file CosmicNCV.tsv.gz
- Nhiều vùng của DNA không mã hóa đóng vai trò kiểm soát hoạt động của gen. Sự thay đổi một trong những vùng này được gọi là đột biến không mã hóa (non coding variant).

STT	Tên cột	Ví dụ
[1:A]	Sample name, Sample id, Id tumour	TCGA-4P-AA8J-01 2339267 2204250
[4:D]	Primary Site	upper_aerodigestive_tract
[5:E]	Site Subtype 1	head_neck
[6:F]	Site Subtype 2	NS
[7:G]	Site Subtype 3	NS
[8:H]	Primary Histology	carcinoma
[9:I]	Histology Subtype 1	squamous_cell_carcinoma
[10:J]	Histology Subtype 2	NS
[11:K]	Histology Subtype 3	NS

Non-coding variants

STT	Tên cột	Ví dụ
[12:L]	GENOMIC_MUTATION_ID	COSV53729554
[13:M]	LEGACY_MUTATION_ID	COSM5185381
[14:N]	Zygotity	Unknown
[15:O]	GRCh	38
[16:P]	Genome position	3:99795824-99795824
[17:Q]	Mutation somatic status	Confirmed somatic variant
[18:R]	WT SEQ	C
[19:S]	MUT SEQ	T
[20:T]	Whole Genome Reseq	n
[21:U]	Whole_Exome	y
[22:V]	Id Study	627
[23:W]	Pubmed_PMID	
[24:X]	HGVSG	3:g.99795824C>T

Copy Number Variants

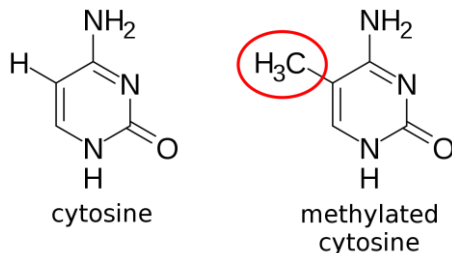
STT	Tên cột	Ví dụ
[1:A]	CNV_ID	590627
[2:B]	Id gene, Gene name	349541 NUMA1_ENST00000616538
[4:D]	Sample id, Id tumour	687432 615592
[6:F]	Primary Site	skin
[7:G]	Site Subtype 1	NS
[8:H]	Site Subtype 2	NS
[9:I]	Site Subtype 3	malignant_melanoma
[10:J]	Primary Histology	squamous_cell_carcinoma
[11:K]	Histology Subtype 1	NS
[12:L]	Histology Subtype 2	NS
[13:M]	Histology Subtype 3	NS
[14:N]	Sample Name	C32
[15:O]	Total_CN	3
[16:P]	Minor Allele	
[17:Q]	Mut Type	gain
[18:R]	Id Study	646
[19:S]	GRCh	38
[20:T]	Chromosome:G_Start..G_Stop	11:67483117..73661653

Gene Expression

- [/cosmic/grch38/cosmic/v97/CosmicCompleteGeneExpression.tsv.gz](#)

STT	Tên cột	Ví dụ
[1:A]	Sample id, Sample name	2194598 TCGA-DD-A114-01
[3:C]	Gene name	TBX20
[4:D]	Regulation	Normal
[5:E]	Z-score	0.745
[6:F]	Id Study	628

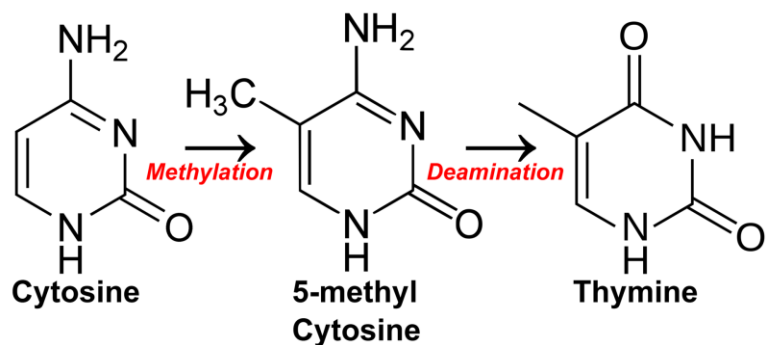
Methylation



- Dữ liệu về mức độ methyl hóa được lưu trong file CosmicCompleteDifferentialMethylation.tsv.gz
- DNA Methylation là quá trình hóa học trong đó nhóm methyl được thêm vào DNA. Việc bổ sung các nhóm methyl có thể ảnh hưởng đến cách một số phân tử hoạt động trong cơ thể. Những thay đổi trong tình trạng methyl hóa của gen có thể dẫn đến ung thư.

STT	Tên cột	Ví dụ
[1:A]	Study_ID	417
[2:B]	Id Sample, Sample name, Id tumour	1780195 TCGA-64-5779-01 1684194
[5:E]	Primary Site	lung
[6:F]	Site Subtype 1	NS
[7:G]	Site Subtype 2	NS
[8:H]	Site Subtype 3	NS
[9:I]	Primary Histology	carcinoma
[10:J]	Histology Subtype 1	adenocarcinoma
[11:K]	Histology Subtype 2	NS
[12:L]	Histology Subtype 3	NS

Methylation



- Điểm CpG (5'—C—phosphate—G—3') là vị trí mà C được theo sau bởi G trong trình tự nucleotit theo chiều 5' – 3'.
- Điểm CpG được chú ý trong quá trình sàng lọc vì C trong CpG dễ bị methyl hóa và khử amino thành T (cao hơn từ 10 tới 50 lần so với những biến đổi khác).

STT	Tên cột	Ví dụ
[13:M]	Fragment Id	cg26246880
[14:N]	Genome Version	38
[15:O]	Chromosome	1
[16:P]	Position	158930759
[17:Q]	Strand	-1
[18:R]	Gene Name	
[19:S]	Methylation	L
[20:T]	Avg Beta Value Normal	0.710
[21:U]	Beta Value	0.171
[22:V]	Two Sided P-Value	0.0000000021105207582

Giá trị beta

Giá trị beta (β) tại vị trí CpG thứ i được tính:

$$Beta_i = \frac{\max(y_{i,methy}, 0)}{\max(y_{i,unmethy}, 0) + \max(y_{i,methy}, 0) + \alpha}$$

Trong đó,

- $y_{i,menty}$ và $y_{i,unmenty}$ lần lượt là cường độ methyl hóa và không methyl hóa. Nếu các giá trị này nhỏ hơn 0 thì lấy giá trị 0 để tính toán
- α là giá trị bù, để chuẩn hóa giá trị β trong trường hợp $y_{i,menty}$ và $y_{i,unmenty}$ quá nhỏ

Giá trị beta là một số từ 0 đến 1, giá trị bằng 0 cho biết tại vị trí CpG đang dò của mẫu hoàn toàn không bị methyl hóa, và giá trị bằng 1 cho biết đã bị methyl hóa hoàn toàn

COSMIC Sample Features

Dữ liệu về thông tin
mẫu được lưu trữ
trong file

CosmicSample.tsv.gz

Id Individual : id cá
nhân được lấy mẫu

STT	Tên cột	Ví dụ
[1:A]	Sample id, Sample name, Id tumour, Id Individual	1044159 1044159 960500 937534
[5:E]	Primary Site	oesophagus
[6:F]	Site Subtype 1	NS
[7:G]	Site Subtype 2	NS
[8:H]	Site Subtype 3	NS
[9:I]	Primary Histology	adenocarcinoma
[10:J]	Histology Subtype 1	NS
[11:K]	Histology Subtype 2	NS
[12:L]	Histology Subtype 3	NS

COSMIC Sample Features

STT	Tên cột	Mô tả	Ví dụ
[13:M]	Therapy Relationship	Liên quan đến thời điểm lấy mẫu mô với liệu pháp dùng thuốc được sử dụng để điều trị khối u.	
[14:N]	Sample Differentiator	Cung cấp thông tin bổ sung nếu có nhiều hơn một mẫu từ một khối u	
[15:O]	Mutation Allele Specification	Liệu đột biến xảy ra trên cặp nhiễm sắc thể giống hay khác nhau khi nhiều đột biến được phát hiện trên một gen với một mẫu trong báo cáo	
[16:P]	Msi		Unknown
[17:Q]	Average Ploidy	Số bội thể trung bình của mẫu, từ dữ liệu số lượng bản sao (nếu có).	
[18:R]	Whole Genome Screen	Liệu toàn bộ bộ gen đã được sàng lọc	n
[19:S]	Whole Exome Screen	Liệu toàn bộ exom đã được sàng lọc	n
[20:T]	Sample Remark	Thông tin bổ sung về mẫu	
[21:U]	Drug Response	Phản ứng với thuốc	
[22:V]	Grade	Cấp độ khối u, tuân theo hệ thống phân cấp cho khối u.	Some Grade data are given in publication
[23:W]	Age at tumour recurrence	Tuổi cá nhân tại thời tái phát	
[24:X]	Stage	Giai đoạn của khối u	Some Stage data are given in publication
[25:Y]	Cytogenetics	Loại nhân	

COSMIC Sample Features

STT	Tên cột	Mô tả	Ví dụ
[26:Z]	Metastatic Site	Vị trí mô di căn được xác định	
[27:AA]	Tumour Source	Nguồn của khối u (Nguyên phát, di căn, ...)	NS
[29:AC]	Age	Tuổi tại thời điểm được chẩn đoán	
[30:AD]	Ethnicity	Sắc tộc (Da trắng, ...)	
[31:AE]	Environmental Variables	Biến môi trường (virus, thuốc lá, ...)	
[32:AF]	Germline Mutation	Tên gen/đột biến nếu đột biến dòng mầm và soma được phát hiện trên cùng một gen với một mẫu	
[33:AG]	Therapy	Trị liệu trước khi sàng lọc	
[34:AH]	Family	Tiền sử đột biến trong gia đình của cá nhân được sàng lọc	
[35:AI]	Normal tissue tested	Liệu mô bình thường đã được sàng lọc	
[36:AJ]	Gender	Giới tính (u: unknown, m: male, f: female)	u
[37:AK]	Individual Remark	Thông tin bổ sung về cá nhân	
[38:AL]	NCI code	Mã NCI	C4025
[39:AM]	SAMPLE_TYPE	Mô tả nguồn gốc mẫu	surgery-fixed
[40:AN]	COSMIC_PHENOTYPE_ID	Mã định danh duy nhất cho phân loại mô và mô học	31394826

COSMIC HGNC

/cosmic/grch38/cosmic/v97/CosmicHGNC.tsv.gz

STT	Tên cột	Mô tả	Ví dụ
[1:A]	COSMIC_ID	Id gen trong COSMIC	10
[2:B]	COSMIC_GENE_NAME	Tên gen trong COSMIC	FLT3
[3:C]	Entrez_id	ID entrez	2322
[4:D]	HGNC_ID	ID HGNC	3765
[5:E]	Mutated?	Liệu gen có đột biến mã hóa?	c
[6:F]	Cancer_census?	Liệu gen có thuộc dự án CGC	y
[7:G]	Expert Curated?	Liệu gen có được giám tuyển bởi chuyên gia	y

COSMIC Resistance Mutations

- `/cosmic/grch38/cosmic/v97/CosmicResistanceMutations.tsv.gz`

STT	Tên cột	Ví dụ
[1:A]	Sample name, Sample id	MGH913 2831769
[3:C]	Gene Name	ALK
[4:D]	Transcript	ENST0000038904 8.7
[5:E]	Census Gene	Yes
[6:F]	Drug Name	Alectinib
[7:G]	MUTATION_ID	178878704
[8:H]	GENOMIC_MUTATION_ID	
[9:I]	LEGACY_MUTATION_ID	COSM5352215
[10:J]	AA Mutation	p.I1171S
[11:K]	CDS Mutation	c.?

COSMIC Resistance Mutations

Tổng quan:
[/cosmic/drug_resistance](#)

STT	TÊN CỘT	VÍ DỤ
[12:L]	Primary Tissue	lung
[13:M]	Tissue Subtype 1	NS
[14:N]	Tissue Subtype 2	NS
[15:O]	Histology	carcinoma
[16:P]	Histology Subtype 1	non_small_cell_carcinoma
[17:Q]	Histology Subtype 2	NS
[18:R]	Pubmed ID	27432227
[19:S]	CGP Study	
[20:T]	Somatic Status	Variant of unknown origin
[21:U]	Sample Type	Tumour Sample
[22:V]	Zygosity	Unknown
[23:W]	Genome Coordinates (GRCh37/38)	
[24:X]	Tier	1
[25:Y]	HGVSP	ENSP00000373700.3:p.Ile1171Ser
[26:Z]	HGVSC	
[27:AA]	HGVSG	

COSMIC Mutation Tracking

- [/cosmic/grch38/cosmic/v97/CosmicMutationTracking.ts](https://cosmic.grch38.cosmic.v97/CosmicMutationTracking.ts)
v.gz
- Is canonical: liệu phiên mã có là chính tắc.

Bản phiên mã chính tắc được định nghĩa là CDS dài nhất nếu gen đã dịch mã hoặc cDNA dài nhất.

STT	TÊN CỘT	VÍ DỤ
[1:A]	Gene Name	EPB41L5_ENST00000443124
[2:B]	Accession Number	ENST00000443124.5
[3:C]	GENOMIC_MUTATION_ID	COSV55354772
[4:D]	LEGACY_MUTATION_ID	COSN21896666
[5:E]	MUTATION_ID	128026226
[6:F]	GRCH	38
[7:G]	Mutation_type	coding
[8:H]	Is_canonical	n

STT	Tên cột	Mô tả	Ví dụ
[1:A]	GENE_NAME	Tên gen	ABI1
[2:B]	CELL_TYPE	Loại mô hay ung thư	colorectal cancer
[3:C]	PUBMED_PMI D	Bài báo mà dấu hiệu này được đề cập	24913355
[4:D]	HALLMARK	Loại quá trình sinh học dấu hiệu	invasion and metastasis
[5:E]	IMPACT	Tác động của gen (thúc đẩy, ức chế, gây ung thư, ...)	promotes
[6:F]	DESCRIPTION	Mô tả tóm tắt tác động của gen tới dấu hiệu ung thư	promotes tumour cell adhesion, extracellular matrix degradation and invasion
[7:G]	CELL_LINE	Tên Cell Line khi bằng chứng thu được từ dự án Cell Line	

Census Hallmarks

Dữ liệu về các dấu hiệu đặc trưng của ung thư ở một số gen thuộc dự án CGC được lưu trong file Cancer_Gene_Census_Hallmarks_Of_Cancer.tsv.gz

ASCAT Ploidy and Purity Estimates

STT	Tên cột	Ví dụ
[1:A]	Cancer_Type_Code	ABI1
[2:B]	Sample	colorectal cancer
[3:C]	Aberrant_Cell_Fraction(Purity)	24913355
[4:D]	Ploidy	invasion and metastasis

- `/cosmic/grch38/cosmic/v97/ascat_acf_ploidy.tsv`

COSMIC Transcripts

STT	Tên cột	Ví dụ
[1:A]	Gene ID	65328
[2:B]	Gene_NAME	A1BG
[3:C]	Transcript ID	ENST00000263100.7
[4:D]	Strand	-1

/cosmic/grch38/cosmic/v97
/CosmicTranscripts.tsv.gz

NCV CDS Syntax Mapping

Dữ liệu về ánh xạ
các đột biến không
mã hóa quan trọng
được lưu trong file
NCV_CDS_syntax_
mappin.tsv

STT	Tên cột	Ví dụ
[1:A]	GENE	TERT
[2:B]	GENOMIC_MUTATION_ID	COSV57206533
[3:C]	LEGACY_MUTATION_ID	COSM1717386
[4:D]	ALT_ID	
[5:E]	CHR	5
[6:F]	POS_GRCh37	1295171
[7:G]	POS_GRCh38	1295056
[8:H]	WT_ALLELE	G
[9:I]	MUT_ALLELE	A
[10:J]	TRANSCRIPT_ID	ENST00000310581.5
[11:K]	CURATED_CDS_SYNTAX	c.1-67C>T

- COSS# sample id
- COSF# fusion id
- COSV# mutation id
- COSN# non-coding mutation id
- COSM# legacy mutation id
- COST# structural mutation id
- COSU# study id
- COSP# pubmed id
- COSO# classification id

