生技醫藥生物資訊核心設施

Bioinformatics Core Facility for Biotechnology and Pharmaceuticals



癌症基因體與大數據資料處理分析

【高雄醫學大學場】

美國癌症基因體資料庫 (TCGA)
國際癌症基因體資料庫 (ICGC)

黃柏榕 Ph.D.

長庚大學生物醫學系

pjhuang@gap.cgu.edu.tw

July 9, 2020

不正常人類基因研究室
生物醫學系黃柏榕老師

Home About Me Members Courses Publications Invited Talks

About Me

不正常人類基因研究室
生物醫學系黃柏榕老師

Welcome to PJ's Lab

In our lab, we will never teach you how to become an errand boy/girl.

Superdry + Innovation + Enthusiasm



My Research Interests:

1. Big Data in Precision Medicine
2. Applications of Artificial Intelligence in Genomics
3. Visual Analytics in Bioinformatics

Research Interests



1 癌症基因體突變



2 致病源基因體與模式生物
(生物資訊應用)

Representative Achievements

IF: 4.453
R/C:18%



2013

CPAP: Cancer Panel Analysis Pipeline

Gene Panel

IF: 4.453
R/C:18%



2014

Vanno: a visualization-aided variant annotation tool

Comprehensive Gene Panel

IF:11.147
R/C: 4%



2015

CMPD: Cancer Mutant Proteome Database

Mutant Proteome

IF:11.868
R/C:7%



2017

APOBEC3A is an oral cancer prognostic biomarker

Whole exome

IF:4.688
R/C:18%



2017

circlncRNAnet: an integrated web-based resource for mapping functional networks of long or circular forms of (cancer) non-coding RNAs

Transcriptome

IF:11.147
R/C: 4%



2018

mSignatureDB: A database for deciphering mutational signatures in human cancers

Whole exome & Whole genome

IF:3.51
R/C:29%



2018

VAReporter: variant reporter for cancer research

Whole exome & Whole genome

IF:2.53
R/C:52%



2019

CoMutPlotter: a web tool for visual summary of mutations in cancer cohorts

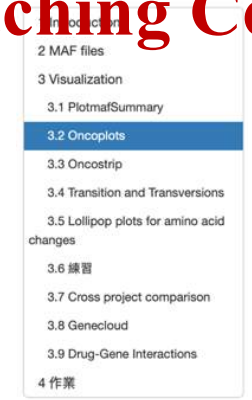
Whole exome & Whole genome

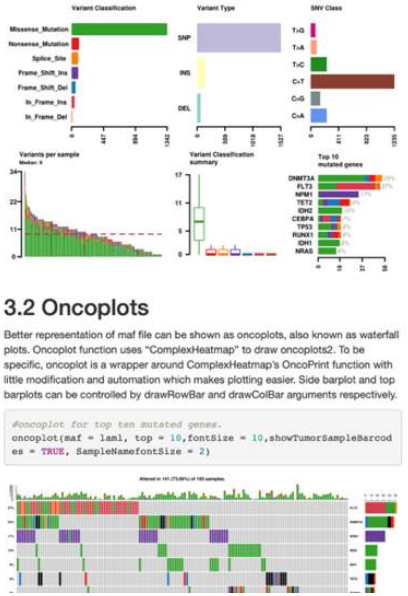
相關論文共8篇

- 自2013 年開發第一個癌症基因檢測自動化分析平台後, 陸續開發 6 個癌症體學相關分析平台
  - Hum Mutat. 2013
- 單基因突變檢測分析擴展到多基因、全外顯子與全基因體
  - Hum Mutat. 2015
- 以突變為連接點，將基因體上的變異連結到轉錄體與蛋白體進而建立癌症突變蛋白體資料庫
  - Nucleic Acids Res. 2015
  - Nat Commun. 2017
  - Gigascience. 2018
- 開發平台從支援單一檢體擴展至研究群體分析，進而支援全世界癌症基因體數據分析，並從中找尋能應用於癌症診斷與治療的新型突變特徵
  - BMC Genomics. 2018
  - Nucleic Acids Res. 2018
  - BMC Med Genomics. 2019

Teaching Courses







3.2 Oncoplots

Better representation of maf file can be shown as oncoplots, also known as waterfall plots. Oncoplot function uses "ComplexHeatmap" to draw oncoplots2. To be specific, oncoplot is a wrapper around ComplexHeatmap's OncoPrint function with little modification and automation which makes plotting easier. Side barplot and top barplots can be controlled by drawRowBar and drawColBar arguments respectively.

```
#oncoplot for top ten mutated genes.
oncoplot(maf = laml, top = 10,fontSize = 10,showTumorSampleBarcodes = TRUE, SampleNamefontSize = 2)
```

任教科目	必選修	開課單位	學分數
網頁程式設計與應用	選修	生醫系	2
生物人的程式設計入門	必修	生醫系	2
量化大數據分析與繪圖	選修	生醫系	2
生物資料庫管理與應用	選修	生醫所	3
生物資訊程式設計與應用	選修	生醫系	3











Maftools Hands-On

pjhuang

20 April, 2020

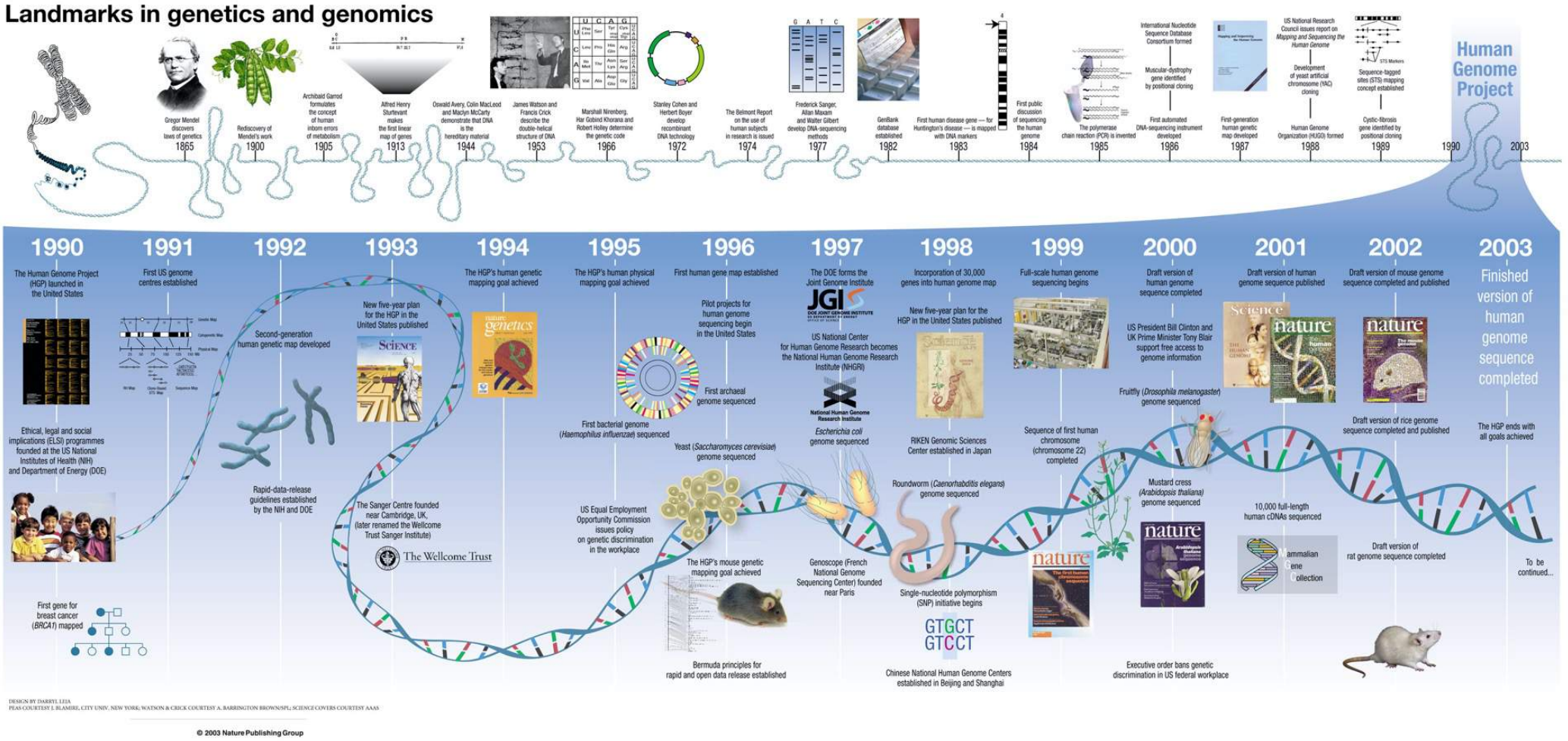
1 Introduction

With advances in Cancer Genomics, Mutation Annotation Format (MAF) is being widely accepted and used to store somatic variants detected. The Cancer Genome Atlas Project has sequenced over 30 different cancers with sample size of each cancer type being over 200. Resulting data consisting of somatic variants are stored in the form of Mutation Annotation Format. This package attempts to summarize, analyze, annotate and visualize MAF files in an efficient manner from either TCGA sources or any in-house studies as long as the data is in MAF format.

Mayakonda A, Lin DC, Assenov Y, Plass C, Koeffler HP. 2018. Maftools:

簡單來說就是癌症基因體計畫所產生的突變資料都存在一種叫MAF格式的檔案裡，maftools就是透過視覺化輔助分析的方法來讓人看得懂MAF的內容。

Landmarks in genetics and genomics



Nature **volume 422**, pages835–847(2003) https://www.nature.com/articles/nature01626

Big biological Data Generator



02. Ion Proton

400+ Genes



04. NextSeq 500

Whole-exome sequencing



01. Ion Torrent

50 Genes



03. MiSeq

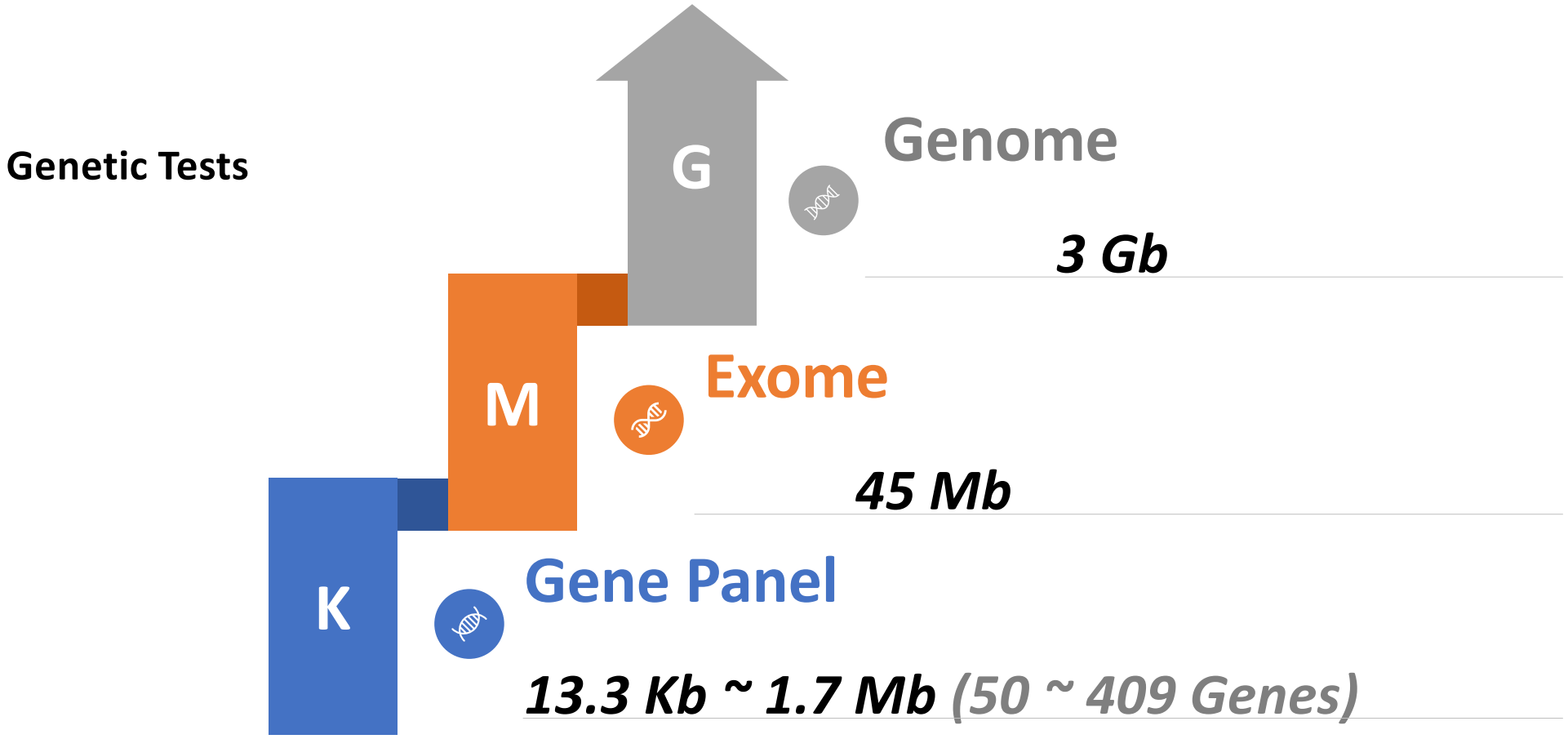
Targeted genes



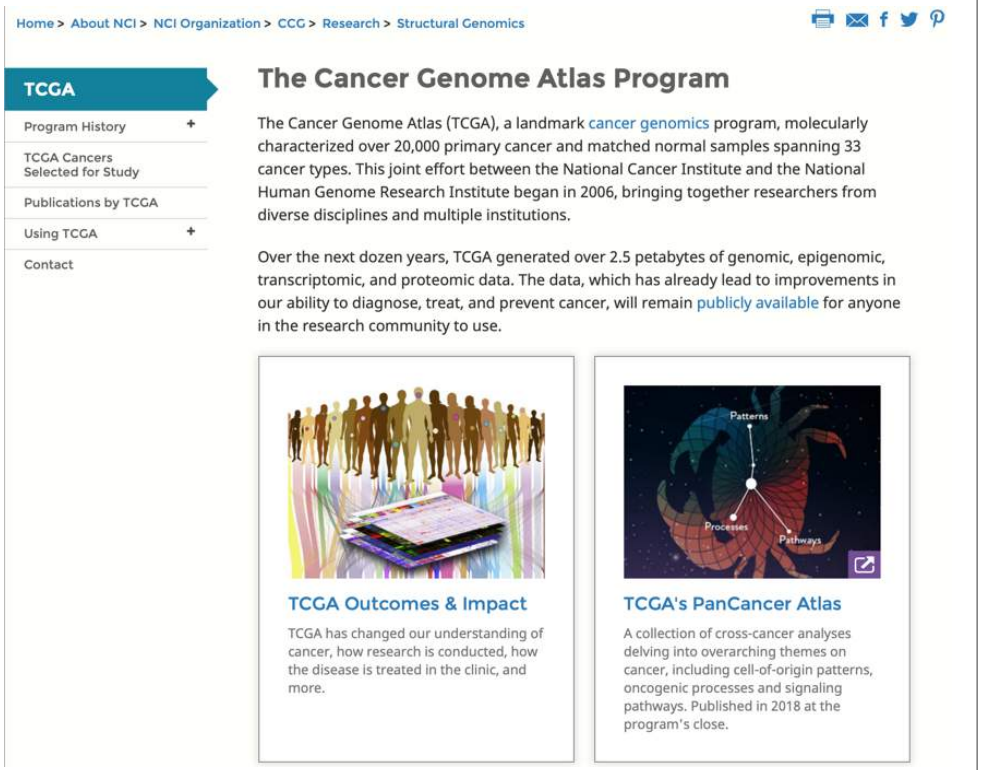
05. HiSeq 2500

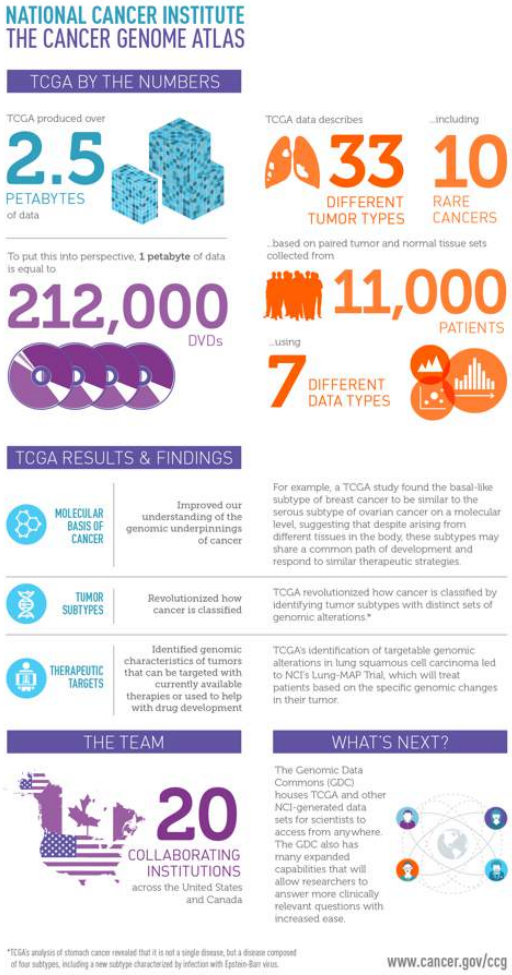
Whole-genome sequencing

Challenges of Genomic Data in the Future



The Cancer Genome Atlas (TCGA)





Began in 2006

https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

TCGA: Timeline



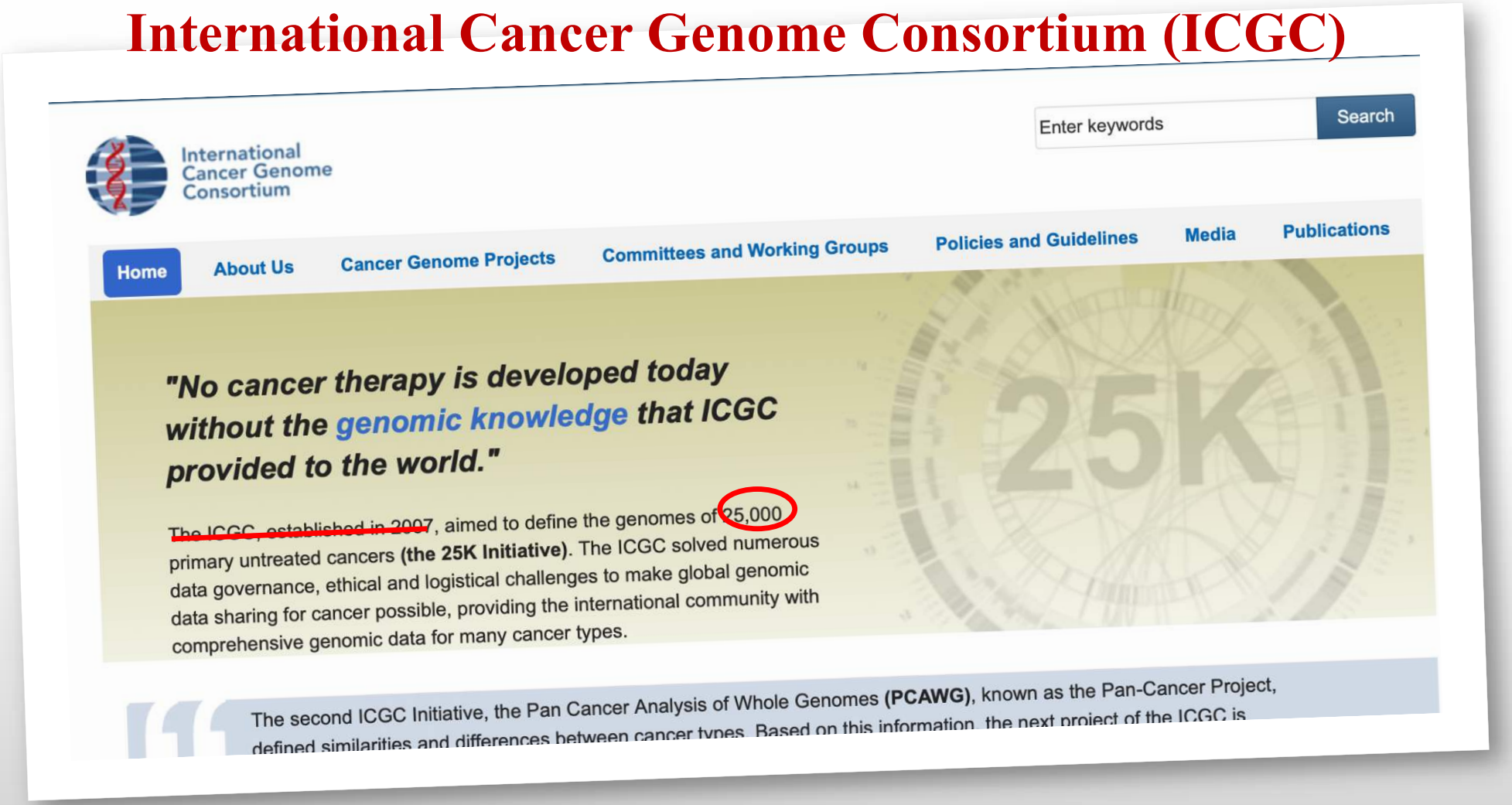
- **Pilot Projects: GBM and Ovarian carcinoma (~500 cases ea.)**
  - Establish infrastructure for effective team science
  - Develop a scalable "pipeline"
  - Demonstrate the feasibility of a large-scale, high throughput approach to identifying the molecular 'parts-list'
  - Make the data publicly and broadly available to the cancer community while protecting patient privacy
- **Expansion 2010 to 2014:**
  - Add 25-35 tumor types
  - Enhancement of sample acquisition & program staff
  - Add Genome Data Analysis Centers
  - Publish "Benchmark Marker Papers"
  - Established FFPE protocols
  - Completely characterize 10,000th case
- **Analysis Completion 2015-2016:**
  - Finish marker papers on rare & "challenging-to-accrue" tumors
  - Complete Pan-Cancer Analysis
  - Broader sharing of tools, analytical methods

3

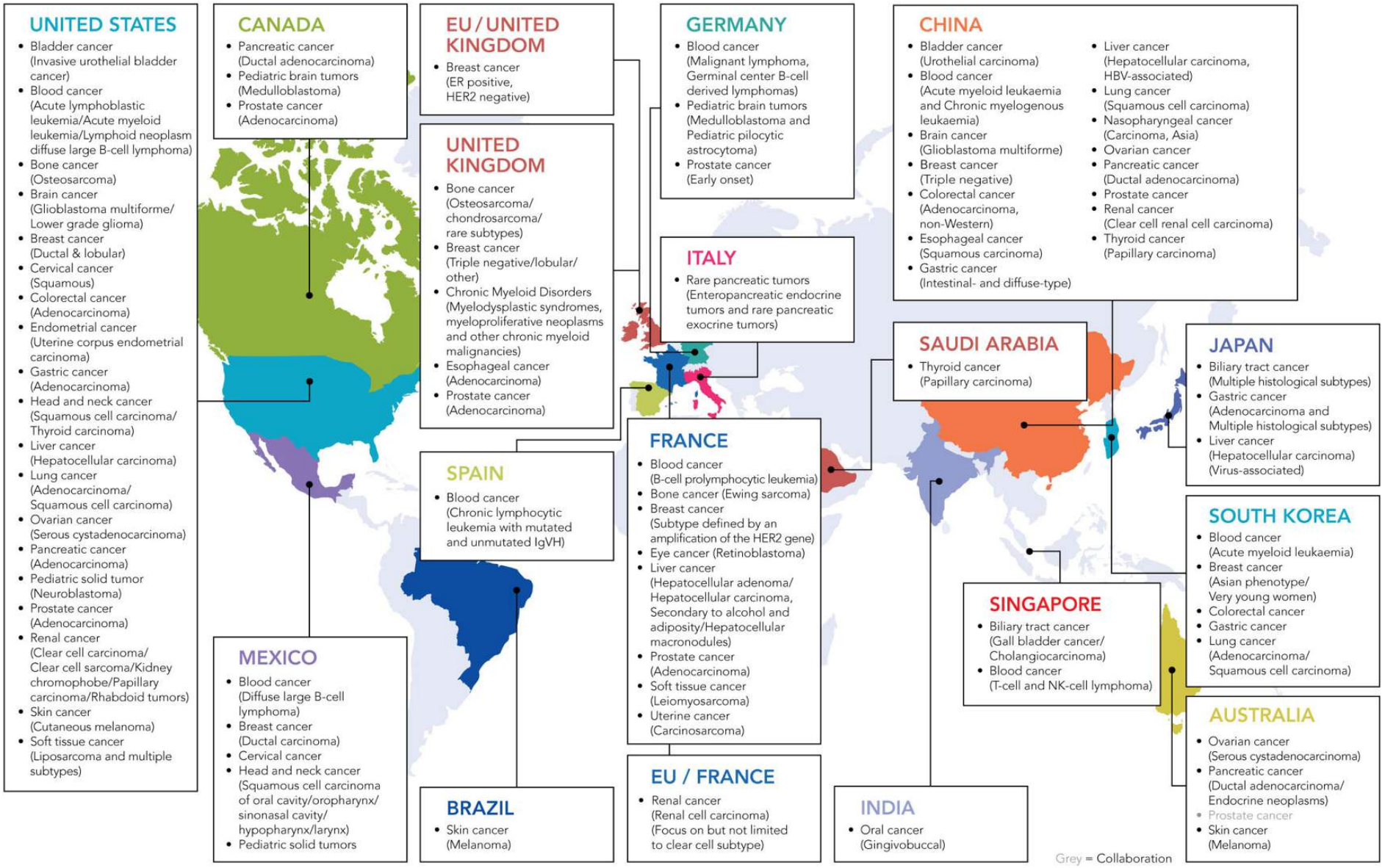
The Cancer Genome Atlas

https://www.slideserve.com/yered/tcga-past-present-future

International Cancer Genome Consortium (ICGC)



https://icgc.org


UNITED STATES
• Bladder cancer (Invasive urothelial bladder cancer)
• Blood cancer (Acute lymphoblastic leukemia/Acute myeloid leukemia/Lymphoid neoplasm diffuse large B-cell lymphoma)
• Bone cancer (Osteosarcoma)
• Brain cancer (Glioblastoma multiforme/ Lower grade glioma)
• Breast cancer (Ductal & lobular)
• Cervical cancer (Squamous)
• Colorectal cancer (Adenocarcinoma)
• Endometrial cancer (Uterine corpus endometrial carcinoma)
• Gastric cancer (Adenocarcinoma)
• Head and neck cancer (Squamous cell carcinoma/ Thyroid carcinoma)
• Liver cancer (Hepatocellular carcinoma)
• Lung cancer (Adenocarcinoma/ Squamous cell carcinoma)
• Ovarian cancer (Serous cystadenocarcinoma)
• Pancreatic cancer (Adenocarcinoma)
• Pediatric solid tumor (Neuroblastoma)
• Prostate cancer (Adenocarcinoma)
• Renal cancer (Clear cell carcinoma/ Clear cell sarcoma/Kidney chromophobe/Papillary carcinoma/Rhabdoid tumors)
• Skin cancer (Cutaneous melanoma)
• Soft tissue cancer (Liposarcoma and multiple subtypes)
CANADA
• Pancreatic cancer (Ductal adenocarcinoma)
• Pediatric brain tumors (Medulloblastoma)
• Prostate cancer (Adenocarcinoma)
EU / UNITED KINGDOM
• Breast cancer (ER positive, HER2 negative)
UNITED KINGDOM
• Bone cancer (Osteosarcoma/ chondrosarcoma/ rare subtypes)
• Breast cancer (Triple negative/lobular/ other)
• Chronic Myeloid Disorders (Myelodysplastic syndromes, myeloproliferative neoplasms and other chronic myeloid malignancies)
• Esophageal cancer (Adenocarcinoma)
• Prostate cancer (Adenocarcinoma)
GERMANY
• Blood cancer (Malignant lymphoma, Germinal center B-cell derived lymphomas)
• Pediatric brain tumors (Medulloblastoma and Pediatric pilocytic astrocytoma)
• Prostate cancer (Early onset)
ITALY
• Rare pancreatic tumors (Enteropancreatic endocrine tumors and rare pancreatic exocrine tumors)
CHINA
• Bladder cancer (Urothelial carcinoma)
• Blood cancer (Acute myeloid leukaemia and Chronic myelogenous leukaemia)
• Brain cancer (Glioblastoma multiforme)
• Breast cancer (Triple negative)
• Colorectal cancer (Adenocarcinoma, non-Western)
• Esophageal cancer (Squamous carcinoma)
• Gastric cancer (Intestinal- and diffuse-type)
• Liver cancer (Hepatocellular carcinoma, HBV-associated)
• Lung cancer (Squamous cell carcinoma)
• Nasopharyngeal cancer (Carcinoma, Asia)
• Ovarian cancer
• Pancreatic cancer (Ductal adenocarcinoma)
• Prostate cancer
• Renal cancer (Clear cell renal cell carcinoma)
• Thyroid cancer (Papillary carcinoma)
SAUDI ARABIA
• Thyroid cancer (Papillary carcinoma)
JAPAN
• Biliary tract cancer (Multiple histological subtypes)
• Gastric cancer (Adenocarcinoma and Multiple histological subtypes)
• Liver cancer (Hepatocellular carcinoma) (Virus-associated)
SPAIN
• Blood cancer (Chronic lymphocytic leukemia with mutated and unmutated IgVH)
FRANCE
• Blood cancer (B-cell prolymphocytic leukemia)
• Bone cancer (Ewing sarcoma)
• Breast cancer (Subtype defined by an amplification of the HER2 gene)
• Eye cancer (Retinoblastoma)
• Liver cancer (Hepatocellular adenoma/ Hepatocellular carcinoma, Secondary to alcohol and adiposity/Hepatocellular macronodules)
• Prostate cancer (Adenocarcinoma)
• Soft tissue cancer (Leiomyosarcoma)
• Uterine cancer (Carcinosarcoma)
SINGAPORE
• Biliary tract cancer (Gall bladder cancer/ Cholangiocarcinoma)
• Blood cancer (T-cell and NK-cell lymphoma)
SOUTH KOREA
• Blood cancer (Acute myeloid leukaemia)
• Breast cancer (Asian phenotype/ Very young women)
• Colorectal cancer
• Gastric cancer
• Lung cancer (Adenocarcinoma/ Squamous cell carcinoma)
MEXICO
• Blood cancer (Diffuse large B-cell lymphoma)
• Breast cancer (Ductal carcinoma)
• Cervical cancer
• Head and neck cancer (Squamous cell carcinoma of oral cavity/oropharynx/ sinonasal cavity/ hypopharynx/larynx)
• Pediatric solid tumors
BRAZIL
• Skin cancer (Melanoma)
EU / FRANCE
• Renal cancer (Renal cell carcinoma) (Focus on but not limited to clear cell subtype)
INDIA
• Oral cancer (Gingivobuccal)
AUSTRALIA
• Ovarian cancer (Serous cystadenocarcinoma)
• Pancreatic cancer (Ductal adenocarcinoma/ Endocrine neoplasms)
• Prostate cancer
• Skin cancer (Melanoma)
Grey = Collaboration

1000 GENOMES

About IGSR and the 1000 Genomes Project

The 1000 Genomes Project ran between 2008 and 2015, creating the largest public catalogue of human variation and genotype data. As the project ended, the Data Coordination Centre at EMBL-EBI has received continued funding from the Wellcome Trust to maintain and expand the resource. The International Genome Sample Resource (IGSR) was set up to do this with the following aims:

1. Ensure the future access to and usability of the 1000 Genomes reference data
2. Incorporate additional published genomic data on the 1000 Genomes samples
3. Expand the data collection to include new populations not represented in the 1000 Genomes Project

The 1000 Genomes Project

Overview of the 1000 Genomes Project

The goal of the 1000 Genomes Project was to find most genetic variants with frequencies of at least 1% in the populations studied.

The 1000 Genomes Project took advantage of developments in sequencing technology, which sharply reduced the cost of sequencing. It was the first project to sequence the genomes of a large number of people, to provide a comprehensive resource on human genetic variation. Data from the 1000 Genomes Project was quickly made available to the worldwide scientific community through freely accessible public databases.

Sequencing remained too expensive to deeply sequence the many samples being studied in the project. However, any particular region of the genome generally contains a limited number of haplotypes. Data was combined across samples to allow efficient detection of most of the variants in a region. The project planned to sequence each sample to 4x genome coverage; at this depth, sequencing can not discover all variants in each sample, but can allow the detection of most variants with frequencies as low as 1%. In the final phase of the project, data from 2,504 samples was combined to allow highly accurate assignment of the genotypes in each sample at all the variant sites the project discovered. The multi-sample approach combined with genotype imputation allowed the project to determine a sample's genotype, even in variants not covered by sequencing reads in that sample.

The contribution of the 1000 Genomes Project to genomics was summarised in *Nature* in the issue containing the final publications from the main project.

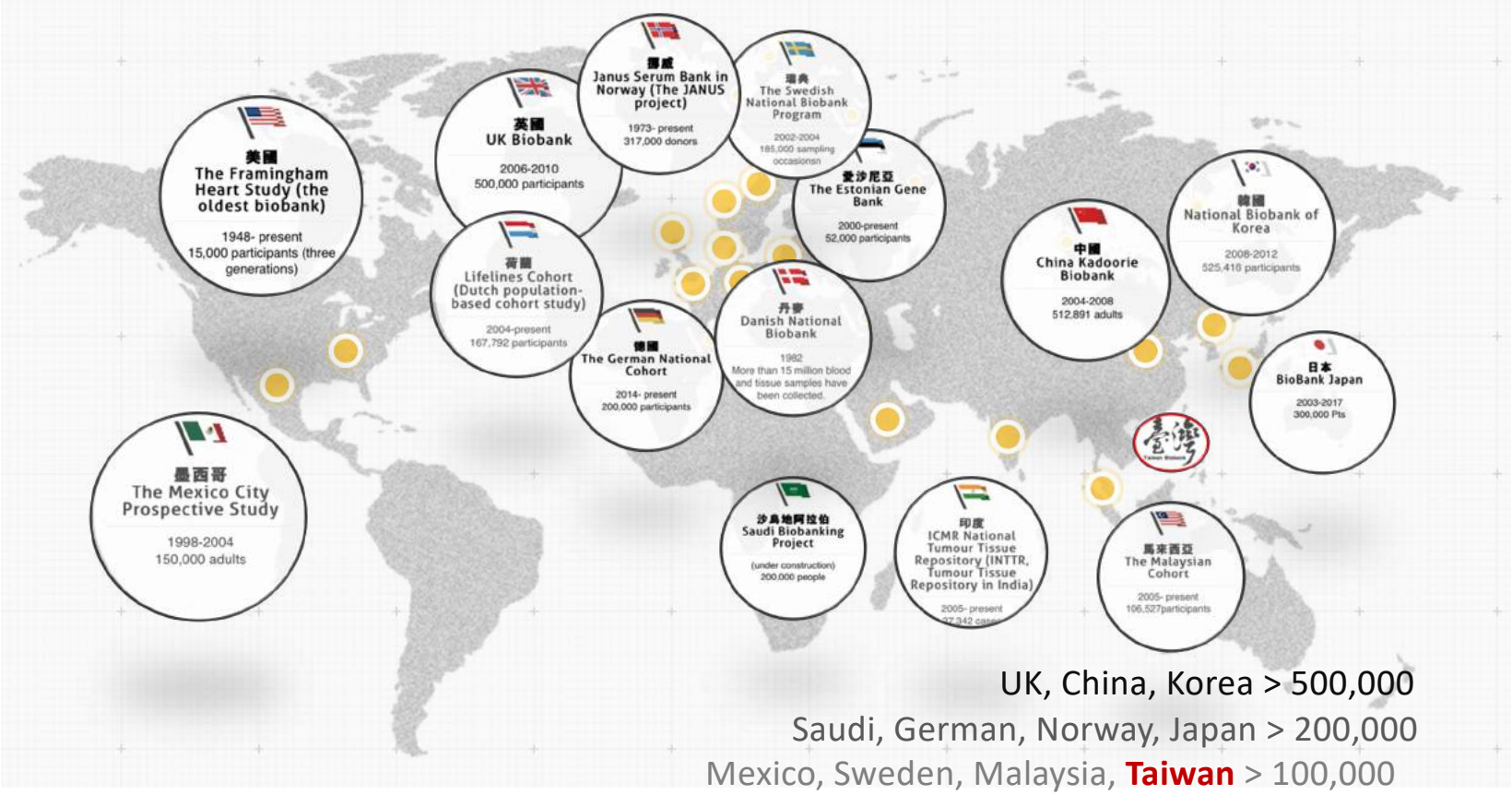
https://www.internationalgenome.org

2008-2015



2,504 samples

BIOBANK



UK, China, Korea > 500,000

Saudi, German, Norway, Japan > 200,000

Mexico, Sweden, Malaysia, **Taiwan** > 100,000

https://www.twbiobank.org.tw/new_web/





https://www.twbiobank.org.tw/new_web_en/

nature

Article | Open Access | Published: 27 May 2020

The mutational constraint spectrum quantified from variation in 141,456 humans

Konrad J. Karczewski, Laurent C. Francioli, [...] Daniel G. MacArthur

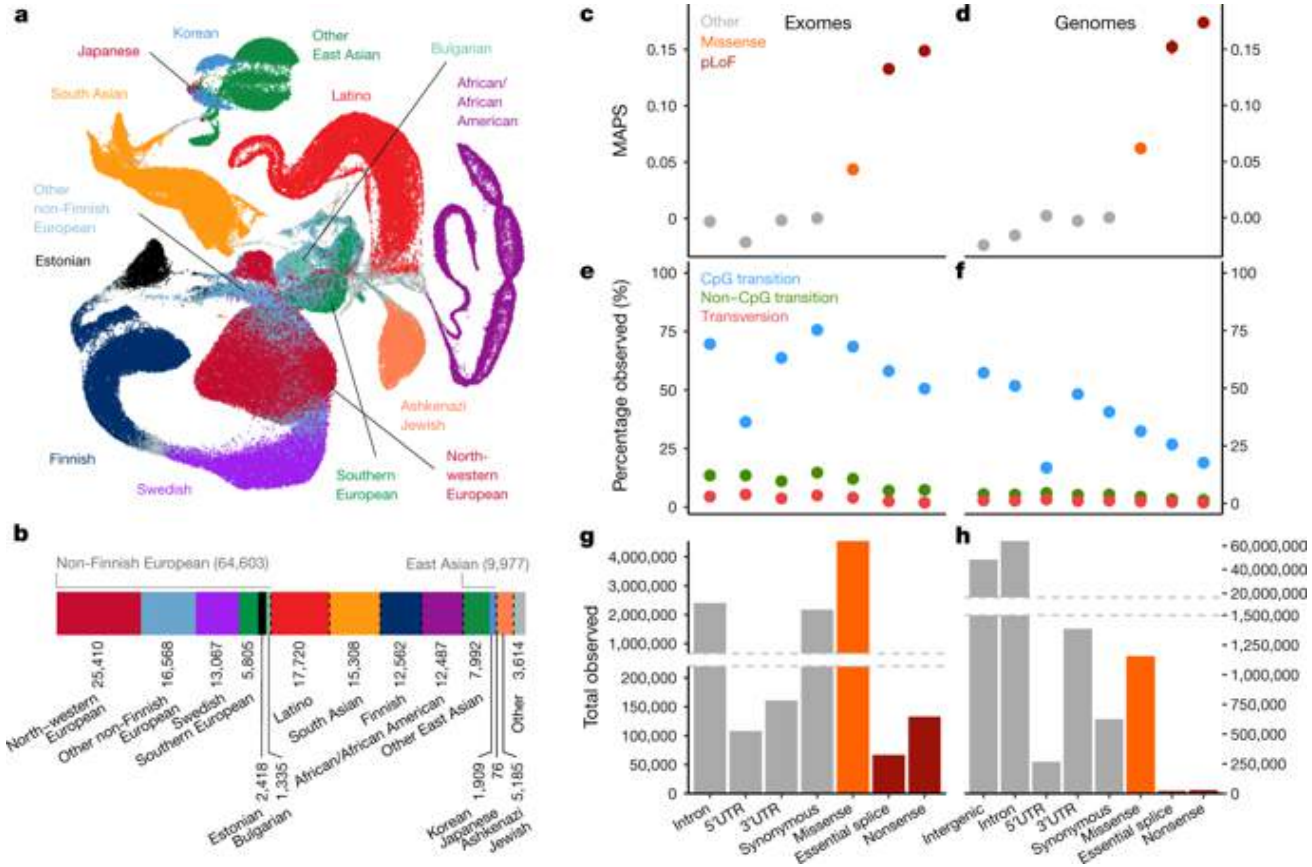
Nature **581**, 434–443(2020) | Cite this article

39k Accesses | **11** Citations | **340** Altmetric | Metrics

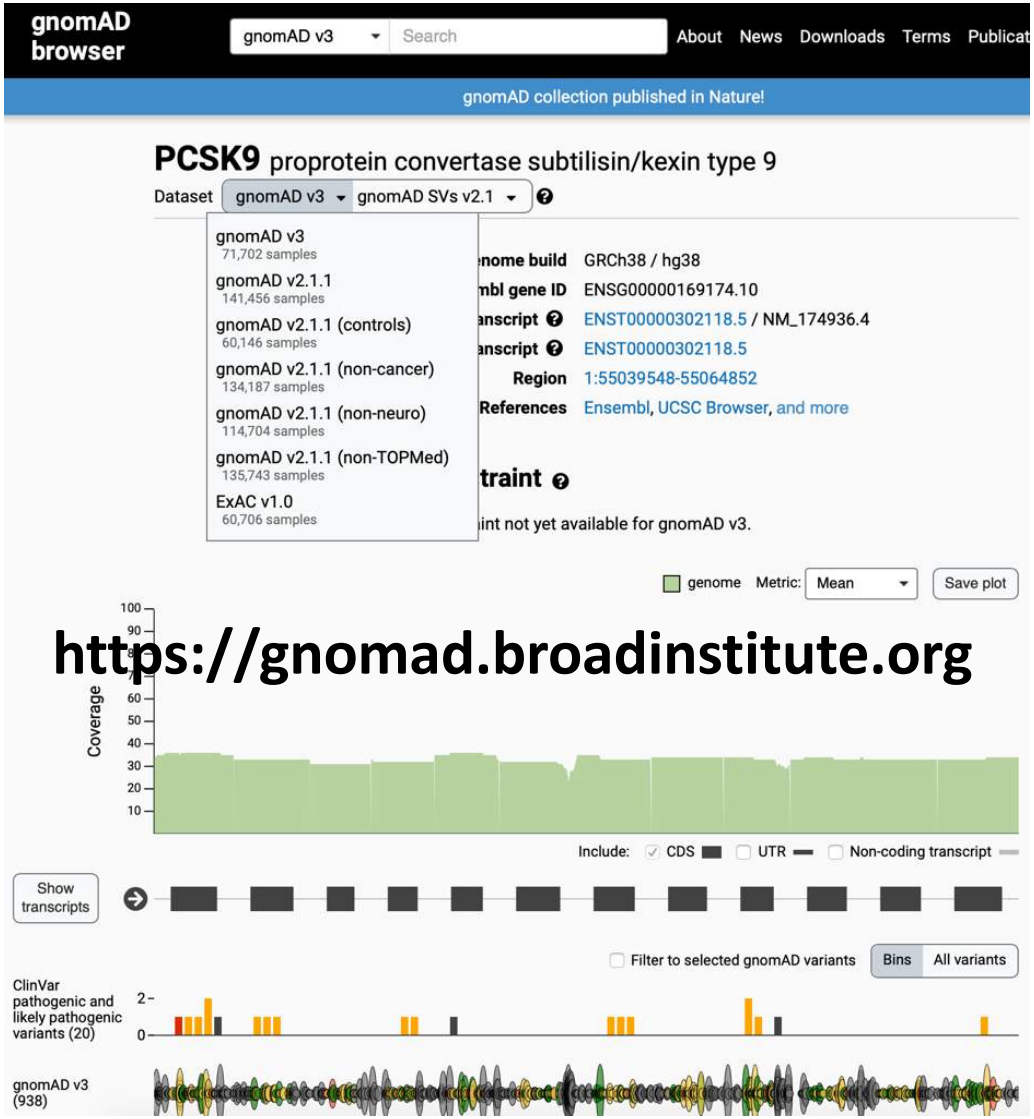
Abstract

Genetic variants that inactivate protein-coding genes are a powerful source of information about the phenotypic consequences of gene disruption: genes that are crucial for the function of an organism will be depleted of such variants in natural populations, whereas non-essential genes will tolerate their accumulation. However, predicted loss-of-function variants are enriched for annotation errors, and tend to be found at extremely low frequencies, so their analysis requires careful variant annotation and very large sample sizes[1]. Here we describe the aggregation of 125,748 exomes and 15,708 genomes from human sequencing studies into the Genome Aggregation Database (gnomAD). We identify 443,769 high-confidence predicted loss-of-function variants in this cohort after filtering for artefacts caused by sequencing and annotation errors. Using an improved model of human mutation rates, we classify human protein-coding genes along a spectrum that represents tolerance to inactivation, validate this classification using data from model organisms and engineered human cells, and show that it can be used to improve the power of gene discovery for both common and rare diseases.

the aggregation of **125,748 exomes** and **15,708 genomes** from human sequencing studies into the **Genome Aggregation Database (gnomAD)**



https://www.nature.com/articles/s41586-020-2308-7



https://gnomad.broadinstitute.org

pLoF only | Missense only | Synonymous only | Other only

Exomes | Genomes | SNVs | Indels | Filtered variants | Search variant table

Export variants to CSV

Note Only variants located in or within 75 base pairs of a coding exon are shown here. To see variants in UTRs or introns, use the region view.

The table below shows the HGVS consequence and VEP annotation for each variant's most severe consequence across all transcripts in this gene. Cases where the most severe consequence occurs in a non-MANE Select transcript (or non-canonical transcript if no MANE Select transcript exists) are denoted with †. To see consequences in a specific transcript, use the transcript view.

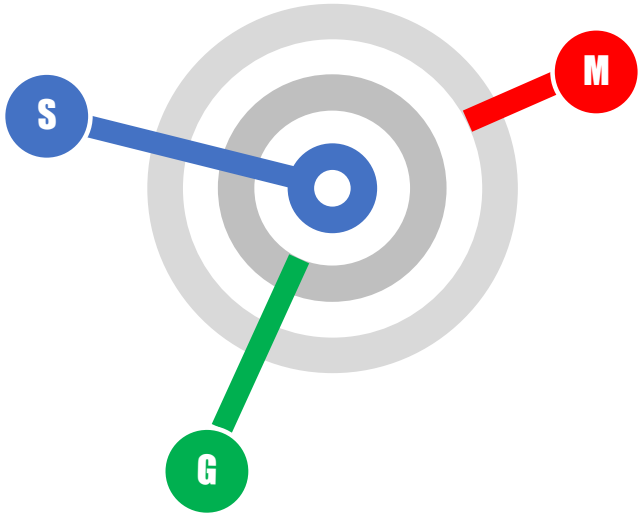
Variant ID	Source	Consequence	Annotation	Flags	Allel
1-55039771-C-T	G	c.-67C>T	5' UTR		1
1-55039774-C-T	G	c.-64C>T	5' UTR		137
1-55039779-T-G	G	c.-59T>G	5' UTR		1
1-55039782-A-T	G	c.-56A>T	5' UTR		4
1-55039784-C-G	G	c.-54C>G	5' UTR		2
1-55039784-C-T	G	c.-54C>T	5' UTR		38
1-55039784-C-A	G	c.-54C>A	5' UTR		1
1-55039785-G-A	G	c.-53G>A	5' UTR		2
1-55039786-C-T	G	c.-52C>T	5' UTR		4
1-55039787-G-T	G	c.-51G>T	5' UTR		1
1-55039791-G-C	G	c.-47G>C	5' UTR		1
1-55039791-G-A	G	c.-47G>A	5' UTR		3
1-55039792-G-A	G	c.-46G>A	5' UTR		1
1-55039793-C-T	G	c.-45C>T	5' UTR		3
1-55039795-T-A	G	c.-43T>A	5' UTR		1
1-55039796-C-T	G	c.-42C>T	5' UTR		1
1-55039800-G-T	G	c.-38G>T	5' UTR		1
1-55039804-C-T	G	c.-34C>T	5' UTR		2
1-55039810-C-G	G	c.-28C>G	5' UTR		1
1-55039812-G-A	G	c.-26G>A	5' UTR		100

Big Data Mania

Somatic Mutation

WES < 50K

- TCGA
  - ✓ 33 Cancer Projects
  - ✓ > **20,000** tumors
- ICGC
  - ✓ 86 Cancer Projects
  - ✓ **24,289** donors



Mutation

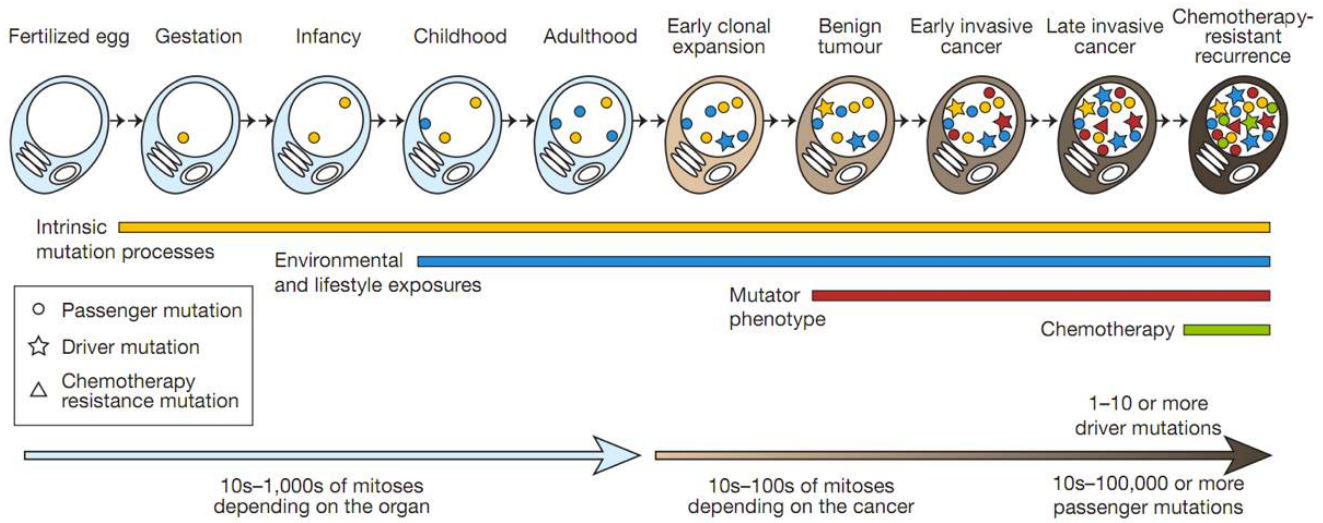
WES 120K
WGS 15K

Germline Mutation

GenomAD: **125,748** Exomes
15,708 Genomes

1000 Genomes: **2,504** Genomes

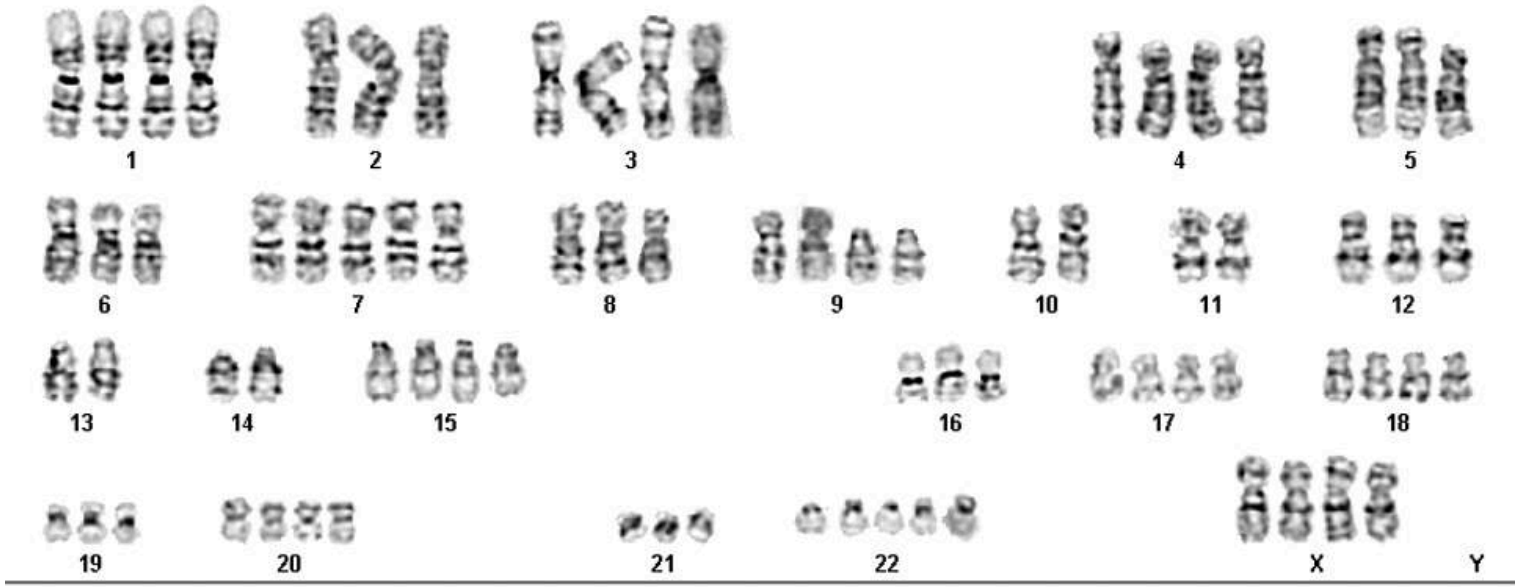
Cancer cells accumulate somatic alternations over time



Mutation frequency depends on cancer type
External forces (e.g. drug treatment) can select for specific clones
(e.g. cells with resistance mutations)

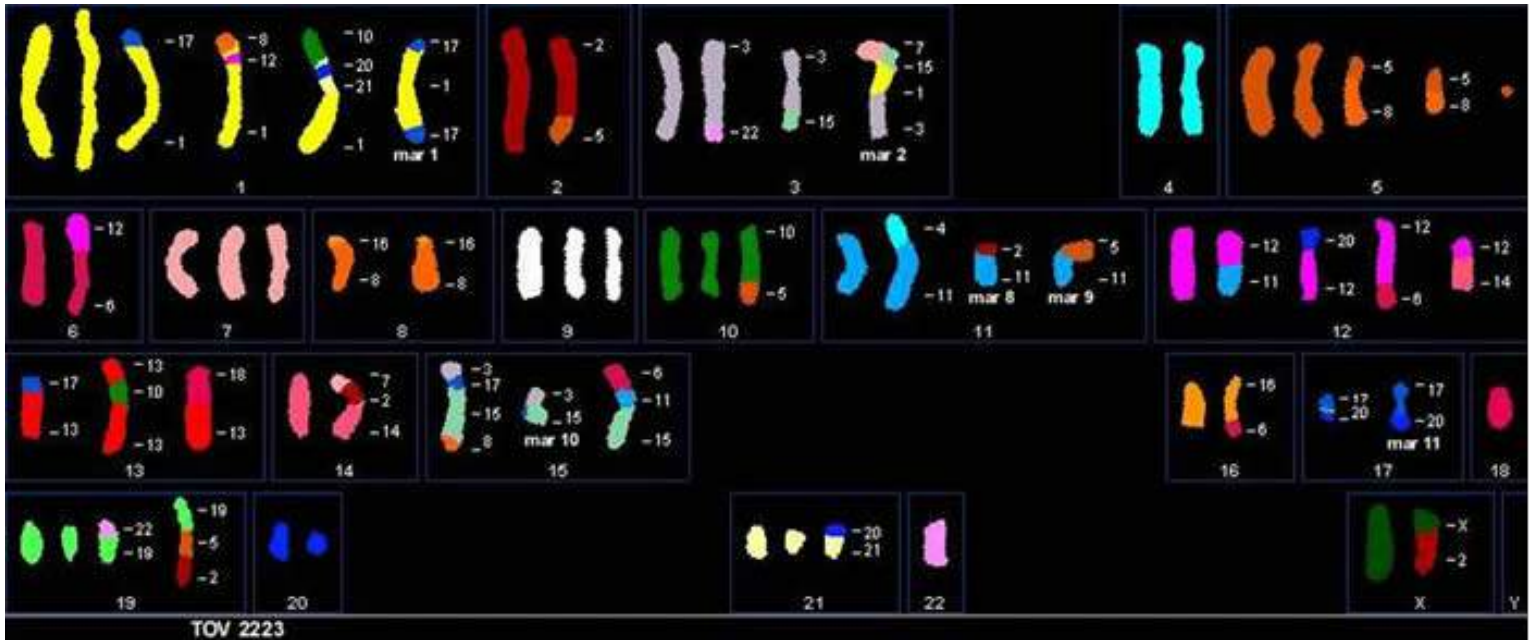
Nature. 2009 Apr 9;458(7239):719-24. Review

Tumors can have multiple genome copies ("ploidy")



Glioma karyogram (GTG-banding). 78,<4n>,XXXX,-2,-5,-6,del(6)(q21q23)x2,+7,-8,del(8)(q22q24.1),del(9)(p10)x2,-10,-10,-11, -11,-12,-13,-13,-14,-14,-16,-19,del(19)(p10),-21,+22

Tumors can have complex structural alterations



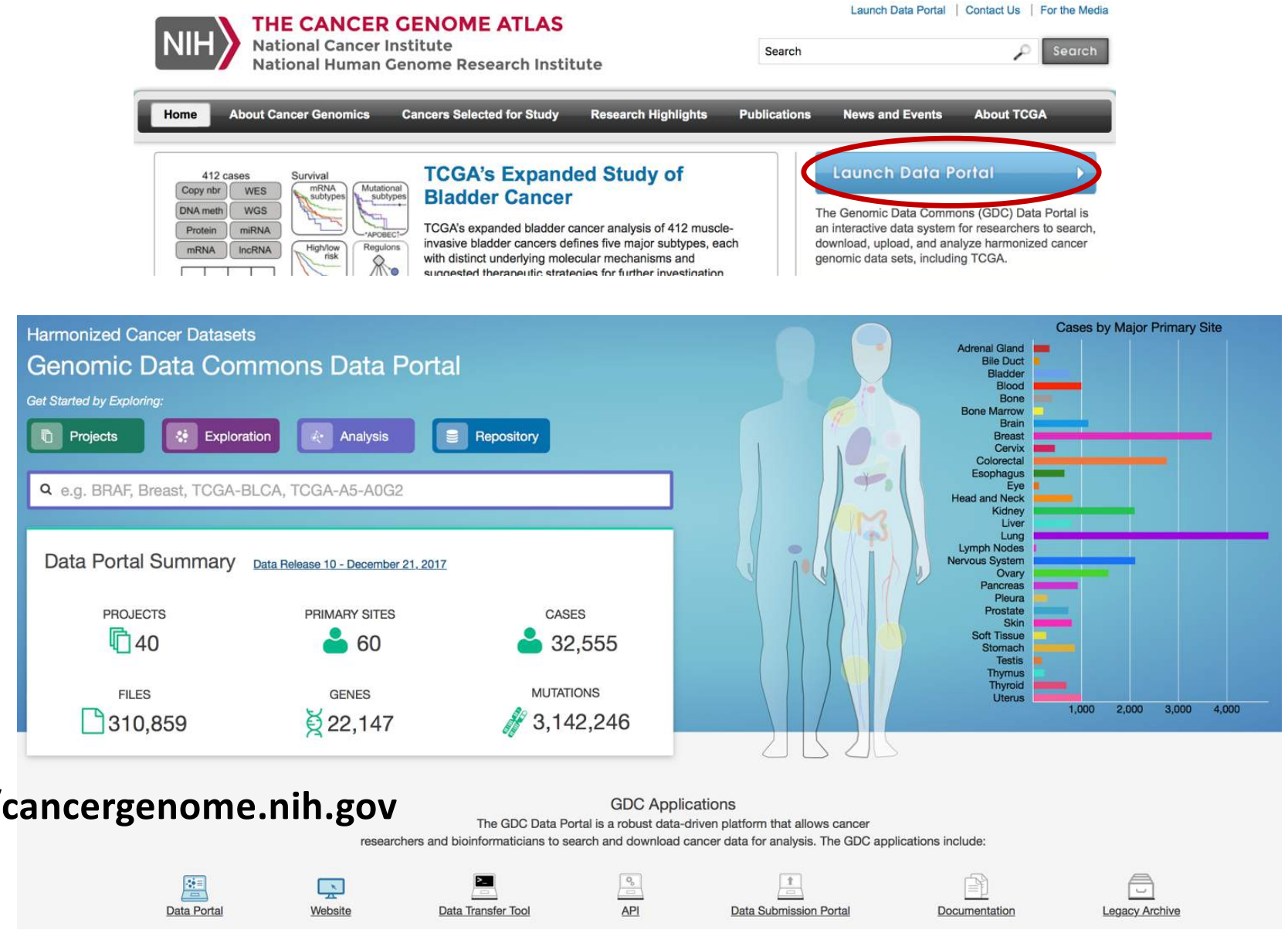
What are the targets in my cancer...

sequence	Mutations, polymorphisms
structure	Copy number variation, translocation, Loss-of-heterozygosity
function	Expression, exon-usage
external	Viruses, bacteria

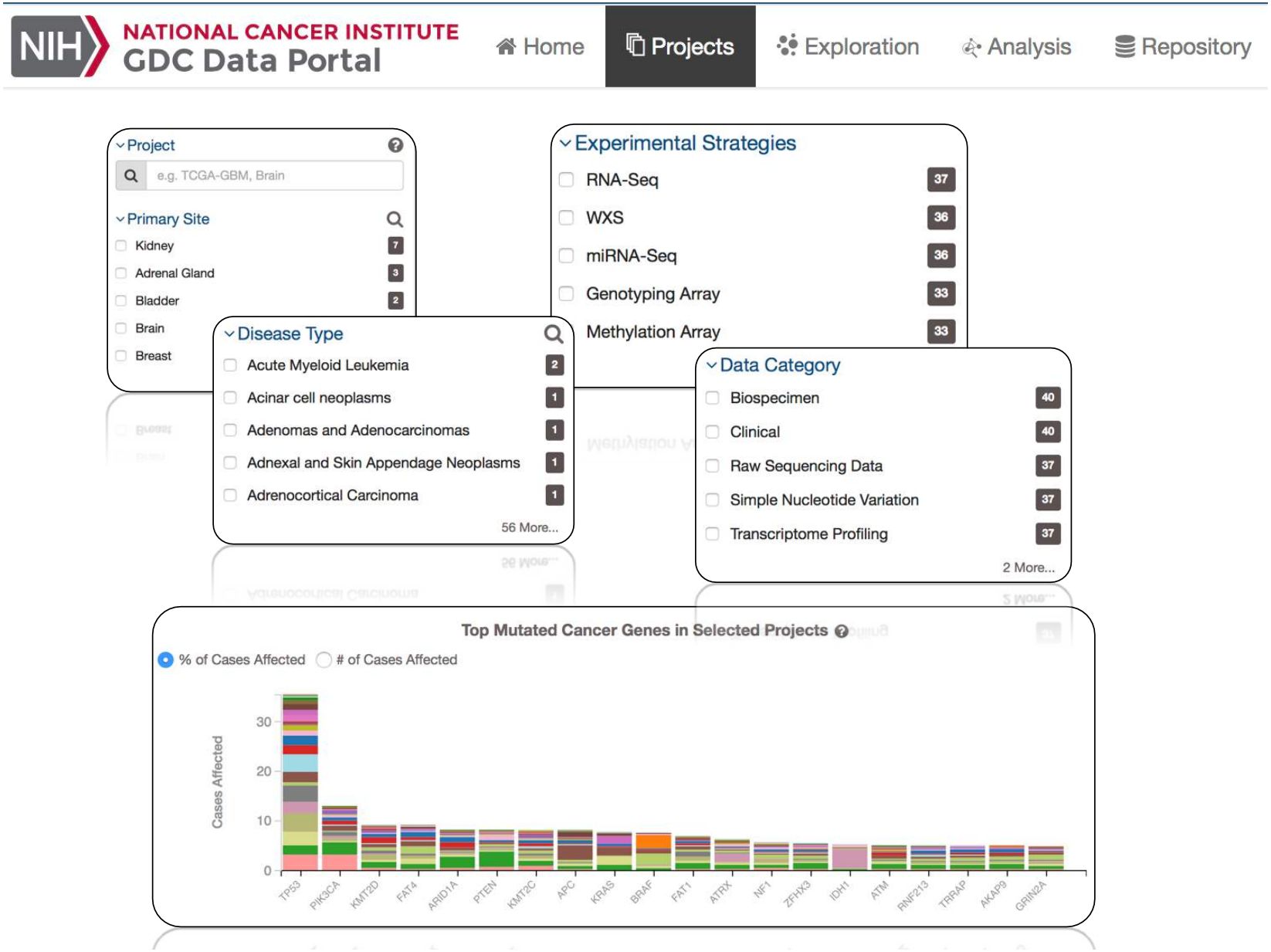
What Information/data do you need about the project and want to put into a cancer database?

Information about the

- **patient**
- **tumor**
- **samples you collected**
  extractions & how
- **sequences**
- **the analyses**
- **the interpretations**



https://cancergenome.nih.gov


NIH
NATIONAL CANCER INSTITUTE
GDC Data Portal
Home
Projects
Exploration
Analysis
Repository
Project
e.g. TCGA-GBM, Brain
Primary Site
Kidney 7
Adrenal Gland 3
Bladder 2
Brain
Breast
Disease Type
Acute Myeloid Leukemia 2
Acinar cell neoplasms 1
Adenomas and Adenocarcinomas 1
Adnexal and Skin Appendage Neoplasms 1
Adrenocortical Carcinoma 1
56 More...
Experimental Strategies
RNA-Seq 37
WXS 36
miRNA-Seq 36
Genotyping Array 33
Methylation Array 33
Data Category
Biospecimen 40
Clinical 40
Raw Sequencing Data 37
Simple Nucleotide Variation 37
Transcriptome Profiling 37
2 More...
Top Mutated Cancer Genes in Selected Projects
% of Cases Affected
# of Cases Affected
Cases Affected
30
20
10
0
TP53
PIK3CA
KMT2D
FAT4
ARID1A
PTEN
KMT2C
APC
KRAS
BRAF
FAT1
ATRX
NF1
ZFHX3
IDH1
ATM
RNF213
TRRAP
AKAP9
GRIN2A

NATIONAL CANCER INSTITUTE
GDC Data Portal

Home | Projects | Exploration | Analysis | Repository

Table

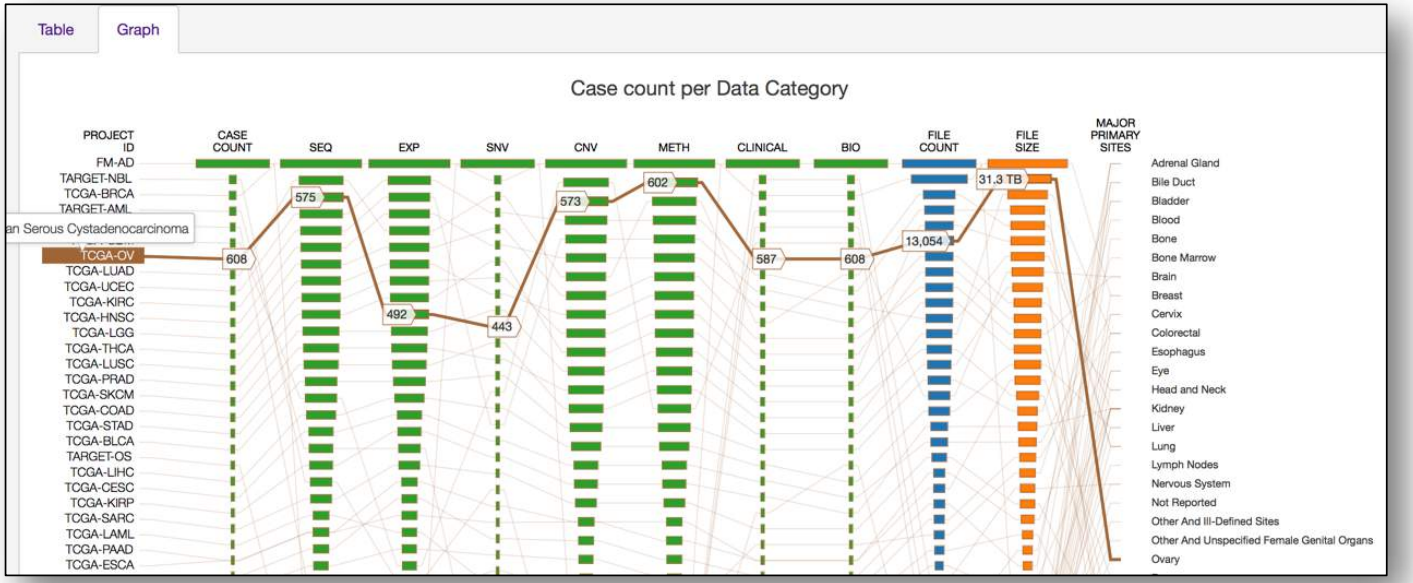
Table | Graph

40 Projects

JSON | TSV

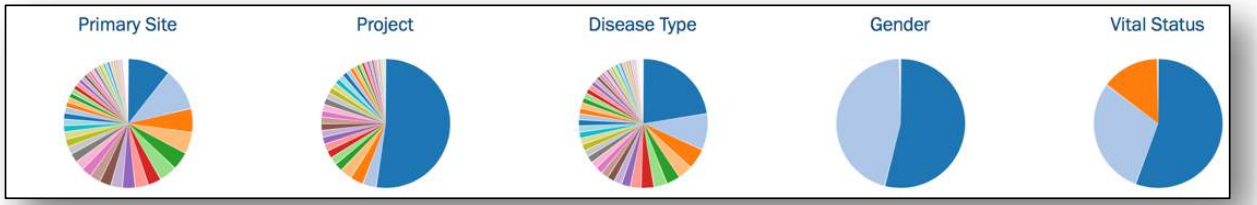
Project	Disease Type	Primary Site	Program	Cases	Available Cases per Data Category							Files
					Seq	Exp	SNV	CNV	Meth	Clinical	Bio	
FM-AD	23 Disease Types	42 Primary Sites	FM	18,004	0	0	18,004	0	0	18,004	18,004	36,134
TARGET-NBL	Neuroblastoma	Nervous System	TARGET	1,127	270	151	216	0	0	1,127	1,127	2,807
TCGA-BRCA	Breast Invasive Carcinoma	Breast	TCGA	1,098	1,098	1,097	1,044	1,096	1,095	1,097	1,098	27,207
TARGET-AML	Acute Myeloid Leukemia	Blood	TARGET	988	299	272	8	0	0	935	988	1,873
TARGET-WT	High-Risk Wilms Tumor	Kidney	TARGET	652	128	128	34	0	0	652	652	1,324
TCGA-GBM	Glioblastoma Multiforme	Brain	TCGA	617	406	166	396	593	423	596	617	9,657
TCGA-OV	Ovarian Serous Cystadenocarcinoma	Ovary	TCGA	608	575	492	443	573	602	587	608	13,054
TCGA-LUAD	Lung Adenocarcinoma	Lung	TCGA	585	582	519	569	518	579	522	585	14,804
TCGA-UCEC	Uterine Corpus Endometrial Carcinoma	Uterus	TCGA	560	559	559	542	547	559	548	560	13,604
TCGA-KIRC	Kidney Renal Clear Cell Carcinoma	Kidney	TCGA	537	535	534	339	532	533	537	537	12,272

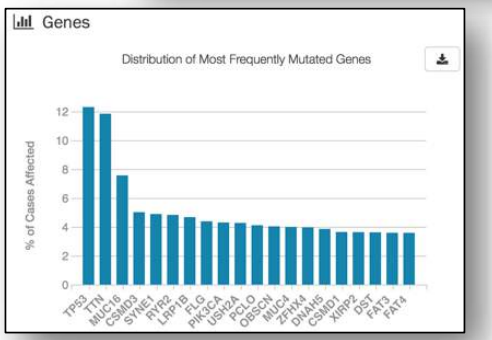
Graph

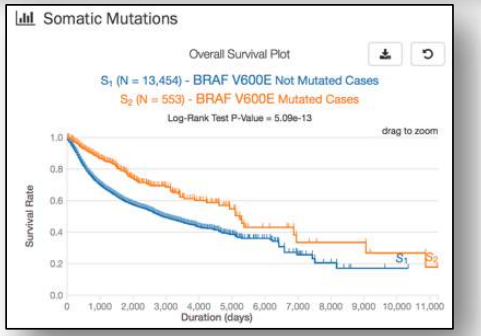


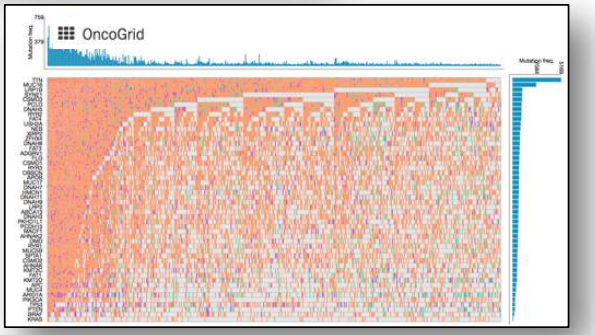
NIH NATIONAL CANCER INSTITUTE
GDC Data Portal

Home | Projects | Exploration | Analysis | Repository

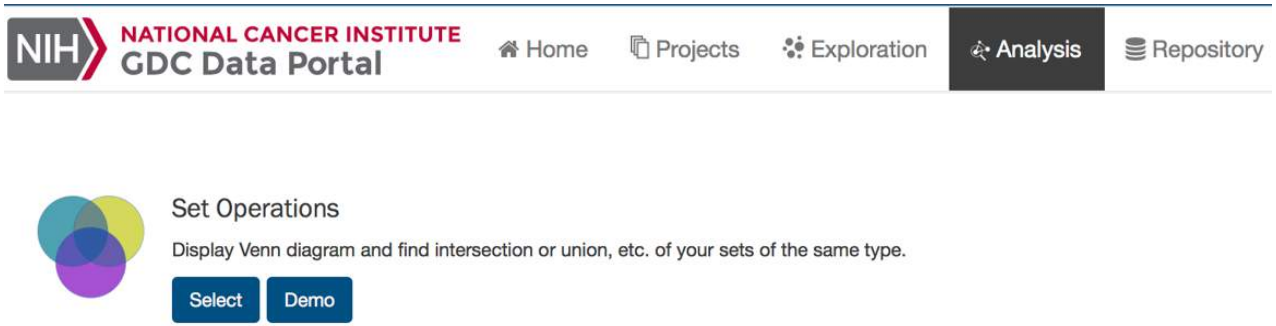


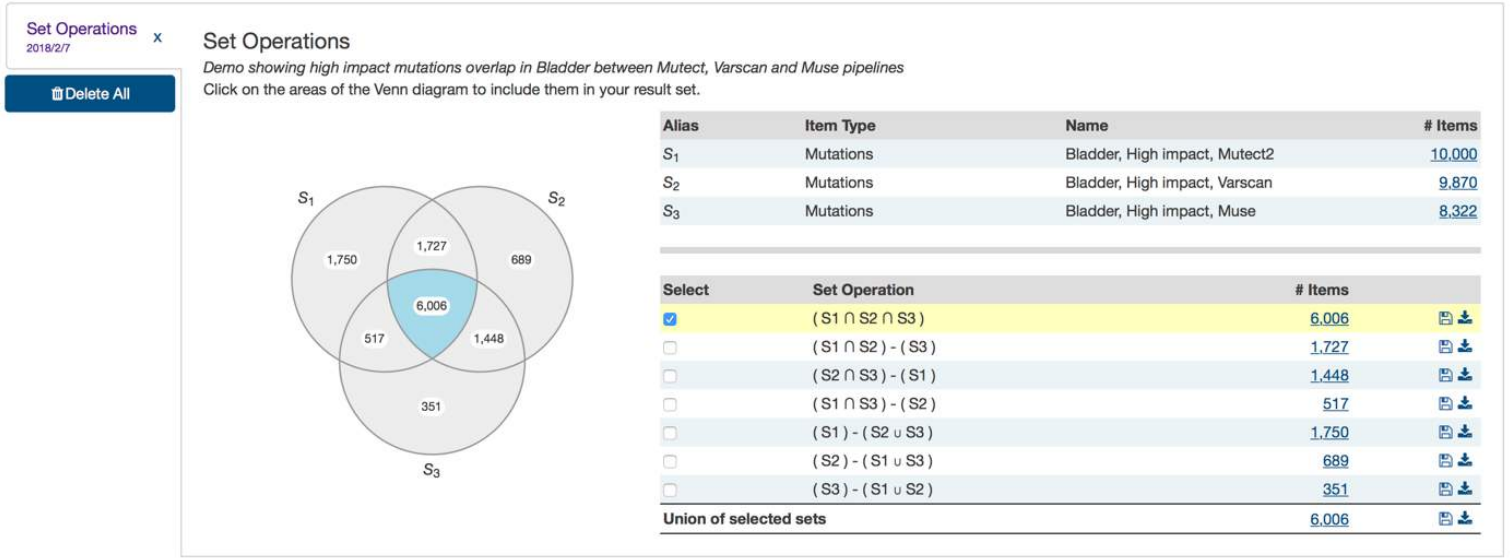






Case ID	Project	Primary Site	Gender	Files	Available Files per Data Category							# Mutations	# Genes
					Seq	Exp	SNV	CNV	Meth	Clinical	Bio		
TCGA-A5-A0G2	TCGA-UCEC	Uterus	Female	32	4	5	16	4	1	1	1	42,051	14,357
TCGA-EO-A22U	TCGA-UCEC	Uterus	Female	32	4	5	16	4	1	1	1	26,998	12,629
TCGA-FI-A2D5	TCGA-UCEC	Uterus	Female	32	4	5	16	4	1	1	1	26,139	12,482
TCGA-AX-A2HC	TCGA-UCEC	Uterus	Female	40	6	10	16	4	2	1	1	24,853	12,205
TCGA-EO-A22R	TCGA-UCEC	Uterus	Female	33	4	5	16	4	2	1	1	24,276	11,920
TCGA-B5-A3FC	TCGA-UCEC	Uterus	Female	32	4	5	16	4	1	1	1	24,584	11,902
TCGA-IB-7651	TCGA-PAAD	Pancreas	Female	32	4	5	16	4	1	1	1	23,084	11,453
TCGA-AP-A1DV	TCGA-UCEC	Uterus	Female	30	4	5	16	2	1	1	1	22,664	11,431
TCGA-E6-A1LX	TCGA-UCEC	Uterus	Female	32	4	5	16	4	1	1	1	23,542	11,397


NIH
NATIONAL CANCER INSTITUTE
GDC Data Portal
Home
Projects
Exploration
Analysis
Repository
Set Operations
Display Venn diagram and find intersection or union, etc. of your sets of the same type.
Select
Demo

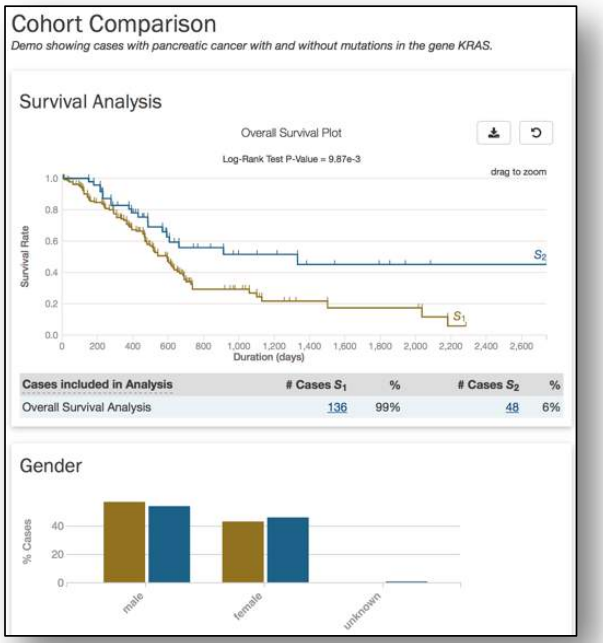

Set Operations
2018/2/7
Delete All
Set Operations
Demo showing high impact mutations overlap in Bladder between Mutect, Varscan and Muse pipelines
Click on the areas of the Venn diagram to include them in your result set.
S1
S2
S3
1,750
1,727
689
6,006
517
1,448
351
Alias | Item Type | Name | # Items
S1 | Mutations | Bladder, High impact, Mutect2 | 10,000
S2 | Mutations | Bladder, High impact, Varscan | 9,870
S3 | Mutations | Bladder, High impact, Muse | 8,322
Select | Set Operation | # Items
(S1 ∩ S2 ∩ S3) | 6,006
(S1 ∩ S2) - (S3) | 1,727
(S2 ∩ S3) - (S1) | 1,448
(S1 ∩ S3) - (S2) | 517
(S1) - (S2 ∪ S3) | 1,750
(S2) - (S1 ∪ S3) | 689
(S3) - (S1 ∪ S2) | 351
Union of selected sets | 6,006

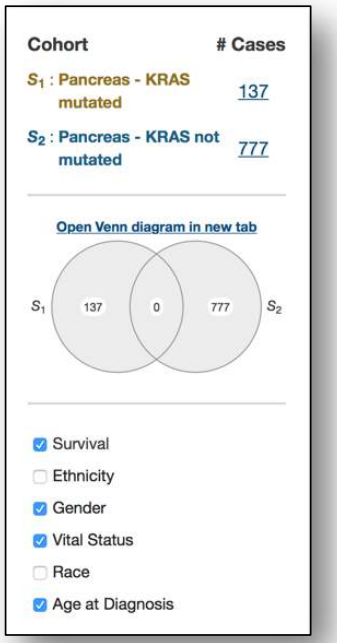


Cohort Comparison

Display the survival analysis of your case sets and compare characteristics such as gender, vital status and age at diagnosis.

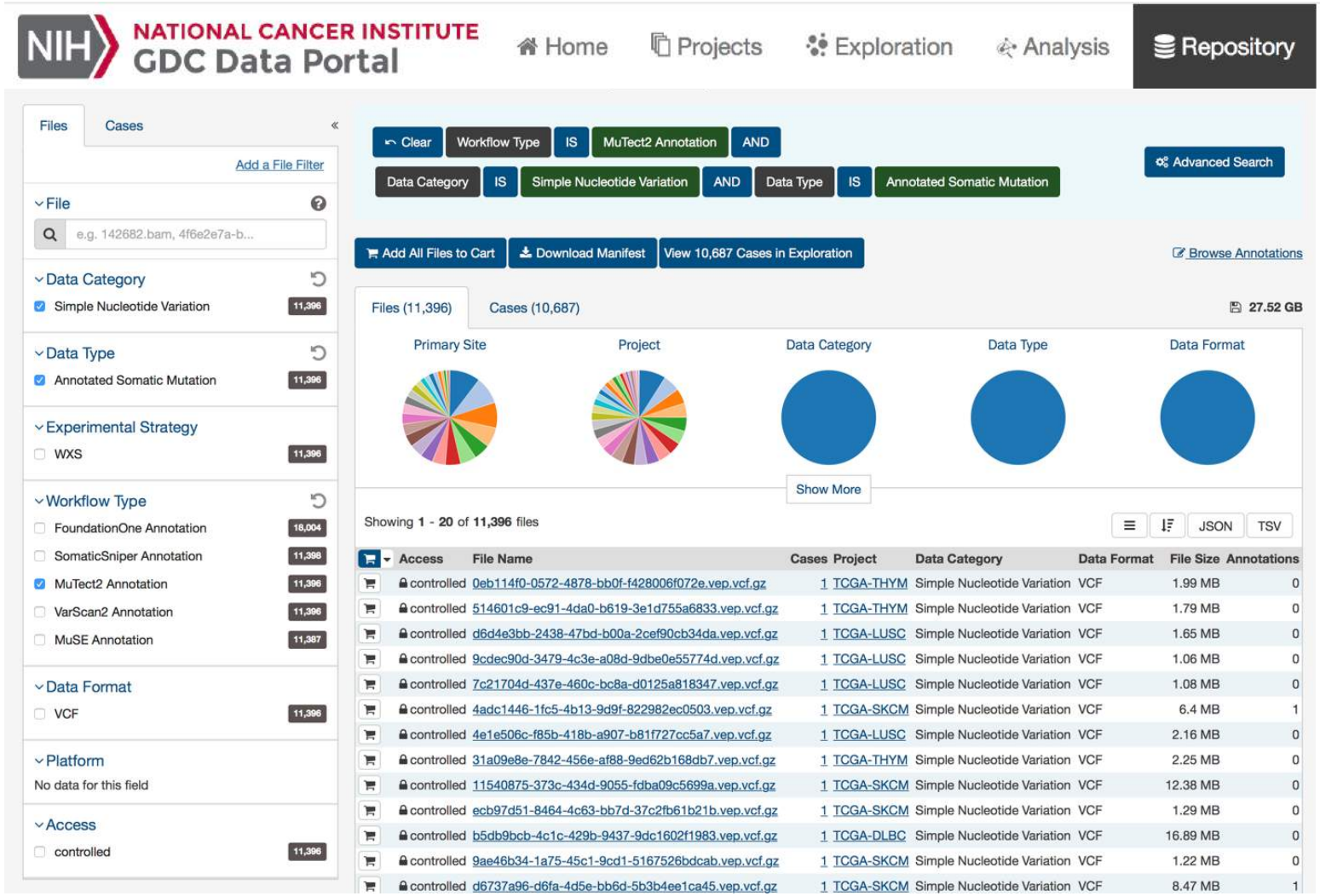
Select Demo

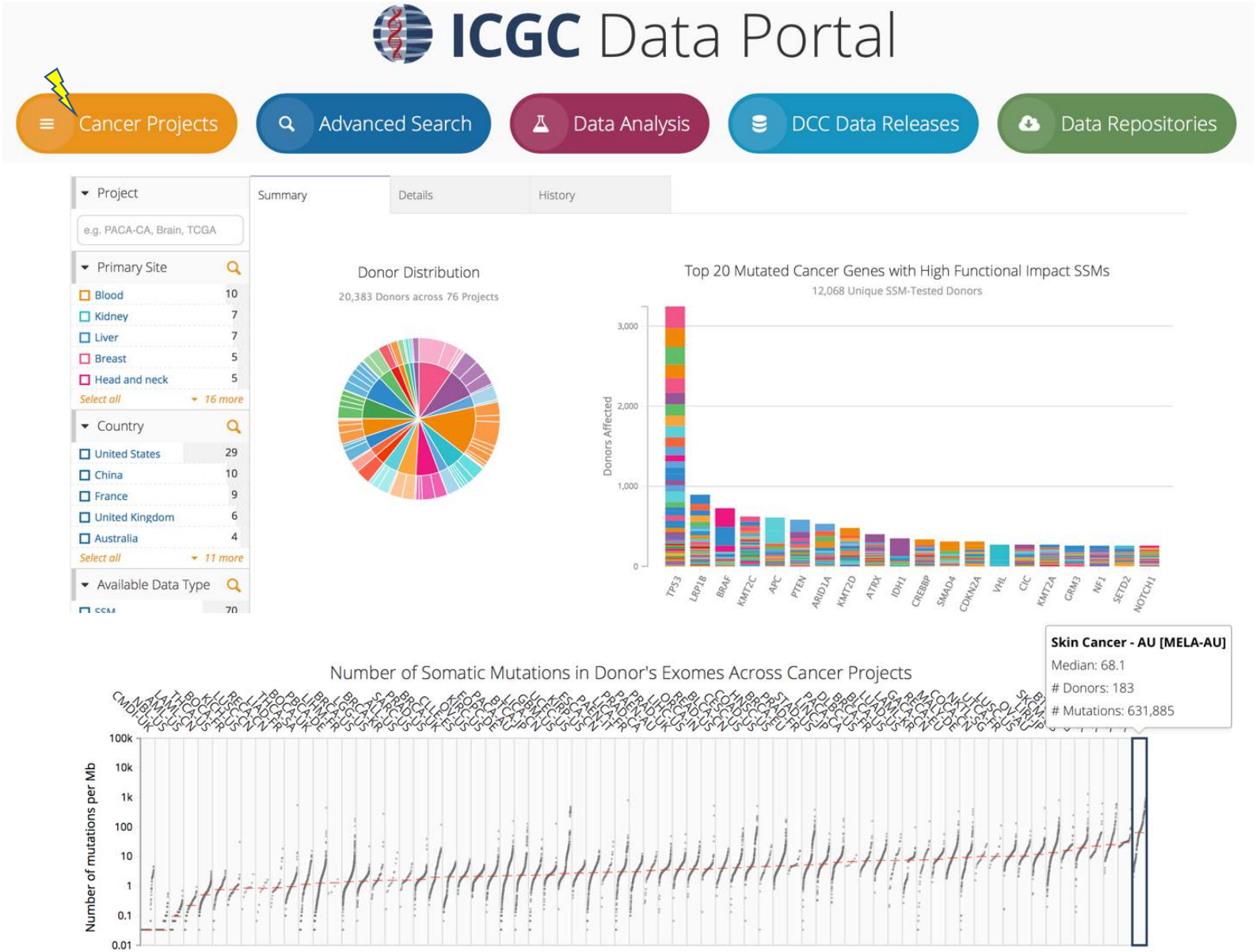


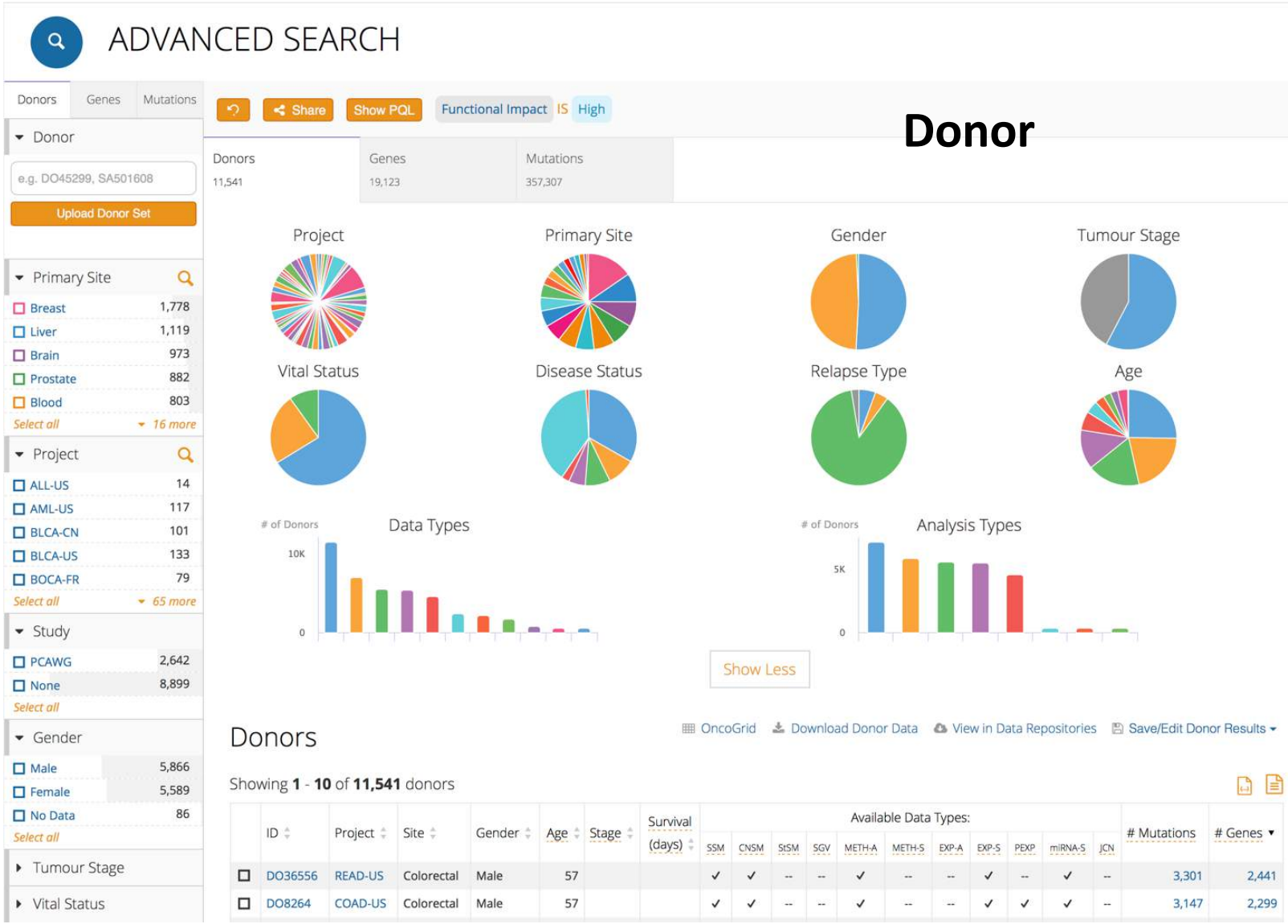


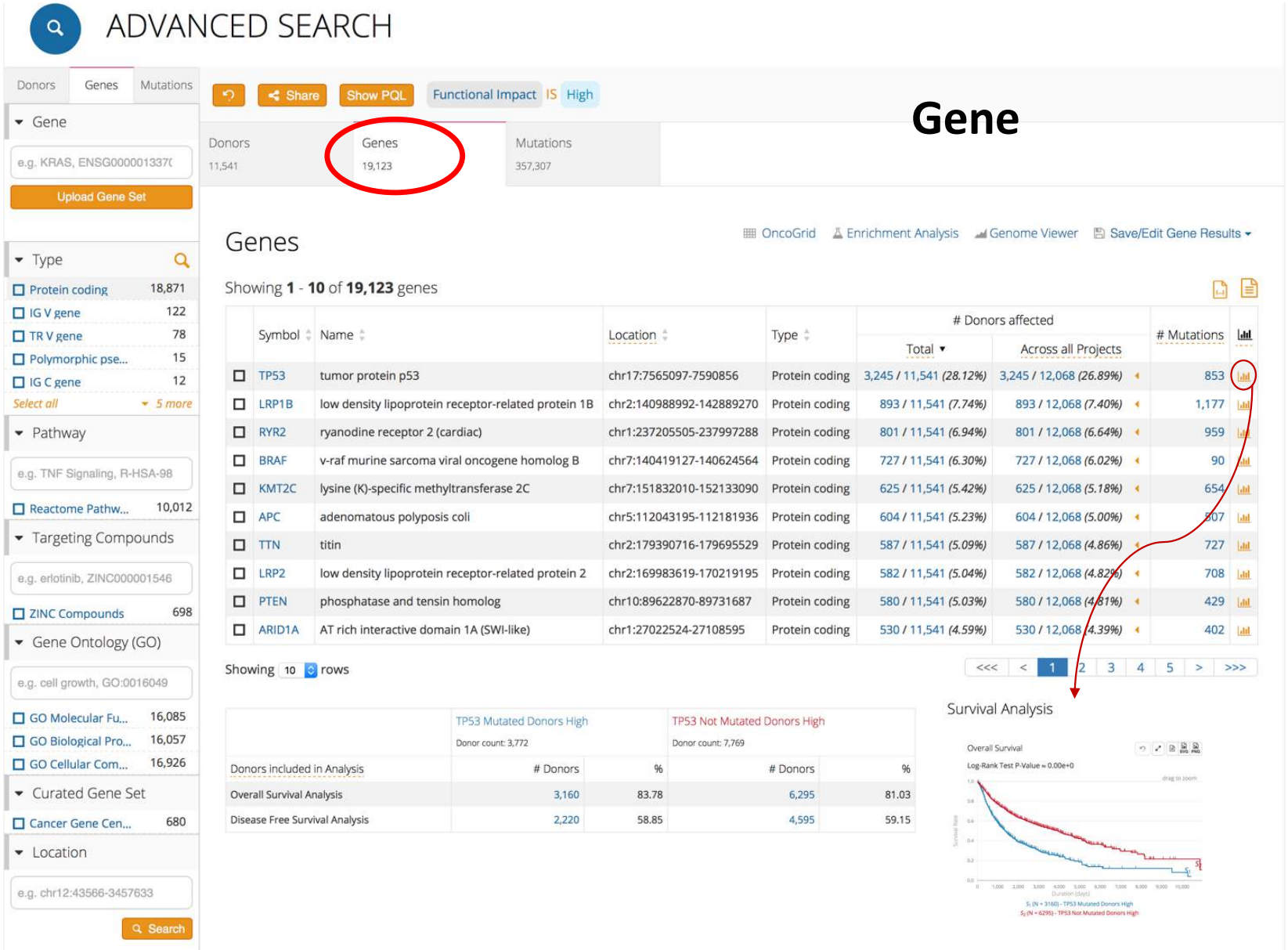
Gender	# Cases S_1	%	# Cases S_2	%
male	78	56.93%	419	53.93%
female	59	43.07%	357	45.95%
unknown	--	0.00%	1	0.13%

P-Value = 5.24e-1

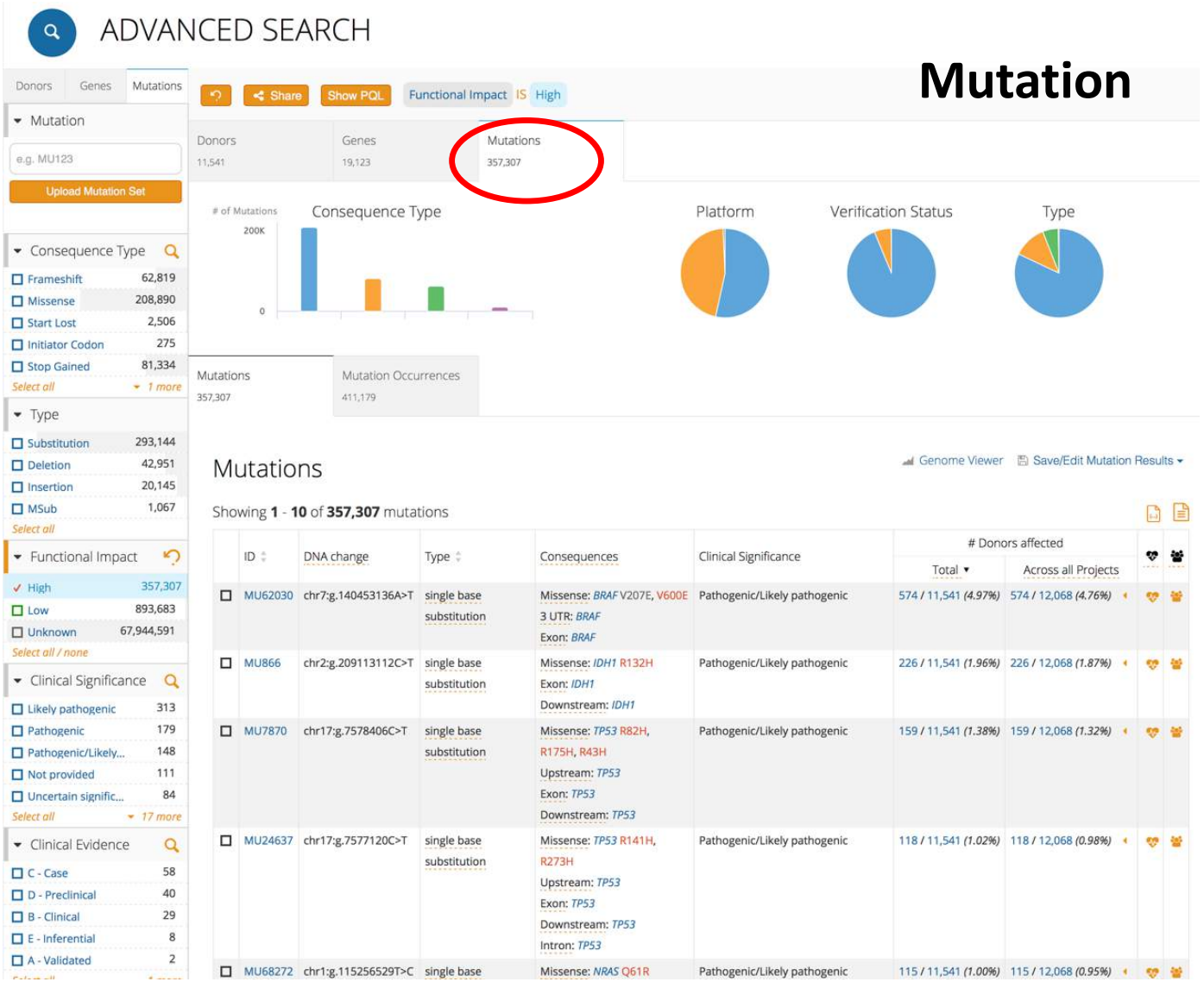

NIH
NATIONAL CANCER INSTITUTE
GDC Data Portal
Home
Projects
Exploration
Analysis
Repository
Files
Cases
Add a File Filter
File
e.g. 142682.bam, 4f6e2e7a-b...
Data Category
Simple Nucleotide Variation 11,396
Data Type
Annotated Somatic Mutation 11,396
Experimental Strategy
WXS 11,396
Workflow Type
FoundationOne Annotation 18,004
SomaticSniper Annotation 11,396
MuTect2 Annotation 11,396
VarScan2 Annotation 11,396
MuSE Annotation 11,387
Data Format
VCF 11,396
Platform
No data for this field
Access
controlled 11,396
Clear
Workflow Type IS MuTect2 Annotation AND
Data Category IS Simple Nucleotide Variation AND Data Type IS Annotated Somatic Mutation
Advanced Search
Add All Files to Cart
Download Manifest
View 10,687 Cases in Exploration
Browse Annotations
Files (11,396)
Cases (10,687)
27.52 GB
Primary Site
Project
Data Category
Data Type
Data Format
Show More
Showing 1 - 20 of 11,396 files
JSON
TSV
Access File Name Cases Project Data Category Data Format File Size Annotations
controlled 0eb114f0-0572-4878-bb0f-f428006f072e.vep.vcf.gz 1 TCGA-THYM Simple Nucleotide Variation VCF 1.99 MB 0
controlled 514601c9-ec91-4da0-b619-3e1d755a6833.vep.vcf.gz 1 TCGA-THYM Simple Nucleotide Variation VCF 1.79 MB 0
controlled d6d4e3bb-2438-47bd-b00a-2cef90cb34da.vep.vcf.gz 1 TCGA-LUSC Simple Nucleotide Variation VCF 1.65 MB 0
controlled 9cdec90d-3479-4c3e-a08d-9dbe0e55774d.vep.vcf.gz 1 TCGA-LUSC Simple Nucleotide Variation VCF 1.06 MB 0
controlled 7c21704d-437e-460c-bc8a-d0125a818347.vep.vcf.gz 1 TCGA-LUSC Simple Nucleotide Variation VCF 1.08 MB 0
controlled 4adc1446-1fc5-4b13-9d9f-822982ec0503.vep.vcf.gz 1 TCGA-SKCM Simple Nucleotide Variation VCF 6.4 MB 1
controlled 4e1e506c-f85b-418b-a907-b81f727cc5a7.vep.vcf.gz 1 TCGA-LUSC Simple Nucleotide Variation VCF 2.16 MB 0
controlled 31a09e8e-7842-456e-af88-9ed62b168db7.vep.vcf.gz 1 TCGA-THYM Simple Nucleotide Variation VCF 2.25 MB 0
controlled 11540875-373c-434d-9055-fdba09c5699a.vep.vcf.gz 1 TCGA-SKCM Simple Nucleotide Variation VCF 12.38 MB 0
controlled ecb97d51-8464-4c63-bb7d-37c2fb61b21b.vep.vcf.gz 1 TCGA-SKCM Simple Nucleotide Variation VCF 1.29 MB 0
controlled b5db9bcb-4c1c-429b-9437-9dc1602f1983.vep.vcf.gz 1 TCGA-DLBC Simple Nucleotide Variation VCF 16.89 MB 0
controlled 9ae46b34-1a75-45c1-9cd1-5167526bdcab.vep.vcf.gz 1 TCGA-SKCM Simple Nucleotide Variation VCF 1.22 MB 0
controlled d6737a96-d6fa-4d5e-bb6d-5b3b4ee1ca45.vep.vcf.gz 1 TCGA-SKCM Simple Nucleotide Variation VCF 8.47 MB 1


ICGC Data Portal
Cancer Projects
Advanced Search
Data Analysis
DCC Data Releases
Data Repositories
Project
e.g. PACA-CA, Brain, TCGA
Primary Site
Blood 10
Kidney 7
Liver 7
Breast 5
Head and neck 5
Select all
16 more
Country
United States 29
China 10
France 9
United Kingdom 6
Australia 4
Select all
11 more
Available Data Type
Summary
Details
History
Donor Distribution
20,383 Donors across 76 Projects
Top 20 Mutated Cancer Genes with High Functional Impact SSMs
12,068 Unique SSM-Tested Donors
Donors Affected
3,000
2,000
1,000
0
TP53
LRP1B
BRAF
KMT2C
APC
PTEN
ARID1A
KMT2D
ATRX
IDH1
CREBBP
SMAD4
CDKN2A
VHL
CIC
KMT2A
GRM3
NF1
SETD2
NOTCH1
Number of Somatic Mutations in Donor's Exomes Across Cancer Projects
Skin Cancer - AU [MELA-AU]
Median: 68.1
Donors: 183
Mutations: 631,885
Number of mutations per Mb
100k
10k
1k
100
10
1
0.1
0.01


ADVANCED SEARCH
Donors
Genes
Mutations
Donor
e.g. DO45299, SA501608
Upload Donor Set
Primary Site
Breast 1,778
Liver 1,119
Brain 973
Prostate 882
Blood 803
Select all
16 more
Project
ALL-US 14
AML-US 117
BLCA-CN 101
BLCA-US 133
BOCA-FR 79
Select all
65 more
Study
PCAWG 2,642
None 8,899
Select all
Gender
Male 5,866
Female 5,589
No Data 86
Select all
Tumour Stage
Vital Status
Share
Show PQL
Functional Impact IS High
Donors 11,541
Genes 19,123
Mutations 357,307
Project
Primary Site
Gender
Tumour Stage
Vital Status
Disease Status
Relapse Type
Age
# of Donors
Data Types
10K
0
# of Donors
Analysis Types
5K
0
Show Less
Donors
OncoGrid
Download Donor Data
View in Data Repositories
Save/Edit Donor Results
Showing 1 - 10 of 11,541 donors
ID
Project
Site
Gender
Age
Stage
Survival (days)
Available Data Types:
SSM
CNSM
StSM
SGV
METH-A
METH-S
EXP-A
EXP-S
PEXP
miRNA-S
JCN
Mutations
Genes
DO36556 READ-US Colorectal Male 57 3,301 2,441
DO8264 COAD-US Colorectal Male 57 3,147 2,299


ADVANCED SEARCH
Donors
Genes
Mutations
Share
Show PQL
Functional Impact IS High
Gene
Donors 11,541
Genes 19,123
Mutations 357,307
Gene
e.g. KRAS, ENSG000001337(
Upload Gene Set
Type
Protein coding 18,871
IG V gene 122
TR V gene 78
Polymorphic pse... 15
IG C gene 12
Select all
5 more
Pathway
e.g. TNF Signaling, R-HSA-98
Reactome Pathw... 10,012
Targeting Compounds
e.g. erlotinib, ZINC000001546
ZINC Compounds 698
Gene Ontology (GO)
e.g. cell growth, GO:0016049
GO Molecular Fu... 16,085
GO Biological Pro... 16,057
GO Cellular Com... 16,926
Curated Gene Set
Cancer Gene Cen... 680
Location
e.g. chr12:43566-3457633
Search
Genes
OncoGrid
Enrichment Analysis
Genome Viewer
Save/Edit Gene Results
Showing 1 - 10 of 19,123 genes
Symbol
Name
Location
Type
Donors affected
Total
Across all Projects
Mutations
TP53 tumor protein p53 chr17:7565097-7590856 Protein coding 3,245 / 11,541 (28.12%) 3,245 / 12,068 (26.89%) 853
LRP1B low density lipoprotein receptor-related protein 1B chr2:140988992-142889270 Protein coding 893 / 11,541 (7.74%) 893 / 12,068 (7.40%) 1,177
RYR2 ryanodine receptor 2 (cardiac) chr1:237205505-237997288 Protein coding 801 / 11,541 (6.94%) 801 / 12,068 (6.64%) 959
BRAF v-raf murine sarcoma viral oncogene homolog B chr7:140419127-140624564 Protein coding 727 / 11,541 (6.30%) 727 / 12,068 (6.02%) 90
KMT2C lysine (K)-specific methyltransferase 2C chr7:151832010-152133090 Protein coding 625 / 11,541 (5.42%) 625 / 12,068 (5.18%) 654
APC adenomatous polyposis coli chr5:112043195-112181936 Protein coding 604 / 11,541 (5.23%) 604 / 12,068 (5.00%) 507
TTN titin chr2:179390716-179695529 Protein coding 587 / 11,541 (5.09%) 587 / 12,068 (4.86%) 727
LRP2 low density lipoprotein receptor-related protein 2 chr2:169983619-170219195 Protein coding 582 / 11,541 (5.04%) 582 / 12,068 (4.82%) 708
PTEN phosphatase and tensin homolog chr10:89622870-89731687 Protein coding 580 / 11,541 (5.03%) 580 / 12,068 (4.81%) 429
ARID1A AT rich interactive domain 1A (SWI-like) chr1:27022524-27108595 Protein coding 530 / 11,541 (4.59%) 530 / 12,068 (4.39%) 402
Showing 10 rows
<<< < 1 2 3 4 5 > >>>
TP53 Mutated Donors High
Donor count: 3,772
TP53 Not Mutated Donors High
Donor count: 7,769
Donors included in Analysis
Donors
%
Donors
%
Overall Survival Analysis 3,160 83.78 6,295 81.03
Disease Free Survival Analysis 2,220 58.85 4,595 59.15
Survival Analysis
Overall Survival
Log-Rank Test P-Value = 0.00e+0

Mutation


ADVANCED SEARCH
Donors
Genes
Mutations
Mutation
e.g. MU123
Upload Mutation Set
Consequence Type
Frameshift 62,819
Missense 208,890
Start Lost 2,506
Initiator Codon 275
Stop Gained 81,334
Select all
1 more
Type
Substitution 293,144
Deletion 42,951
Insertion 20,145
MSub 1,067
Select all
Functional Impact
High 357,307
Low 893,683
Unknown 67,944,591
Select all / none
Clinical Significance
Likely pathogenic 313
Pathogenic 179
Pathogenic/Likely... 148
Not provided 111
Uncertain signific... 84
Select all
17 more
Clinical Evidence
C - Case 58
D - Preclinical 40
B - Clinical 29
E - Inferential 8
A - Validated 2
Share
Show PQL
Functional Impact IS High
Donors 11,541
Genes 19,123
Mutations 357,307
# of Mutations
200K
0
Consequence Type
Platform
Verification Status
Type
Mutations 357,307
Mutation Occurrences 411,179
Mutations
Genome Viewer
Save/Edit Mutation Results
Showing 1 - 10 of 357,307 mutations
ID
DNA change
Type
Consequences
Clinical Significance
Donors affected
Total
Across all Projects
MU62030 chr7:g.140453136A>T single base substitution Missense: BRAF V207E, V600E 3 UTR: BRAF Exon: BRAF Pathogenic/Likely pathogenic 574 / 11,541 (4.97%) 574 / 12,068 (4.76%)
MU866 chr2:g.209113112C>T single base substitution Missense: IDH1 R132H Exon: IDH1 Downstream: IDH1 Pathogenic/Likely pathogenic 226 / 11,541 (1.96%) 226 / 12,068 (1.87%)
MU7870 chr17:g.7578406C>T single base substitution Missense: TP53 R82H, R175H, R43H Upstream: TP53 Exon: TP53 Downstream: TP53 Pathogenic/Likely pathogenic 159 / 11,541 (1.38%) 159 / 12,068 (1.32%)
MU24637 chr17:g.7577120C>T single base substitution Missense: TP53 R141H, R273H Upstream: TP53 Exon: TP53 Downstream: TP53 Intron: TP53 Pathogenic/Likely pathogenic 118 / 11,541 (1.02%) 118 / 12,068 (0.98%)
MU68272 chr1:g.115256529T>C single base Missense: NRAS Q61R Pathogenic/Likely pathogenic 115 / 11,541 (1.00%) 115 / 12,068 (0.95%)

內心的聲音

TCGA、ICGC資料庫真厲害！

資料完整、功能多、介面又漂亮

- **1.**自己手上的定序資料怎麼分析？
- **2.**如何與**TCGA/ICGC**資料一起比較？

Visual summary of mutations in a study cohort (Integration with Clinical Information)

APOBEC3A is an oral cancer prognostic biomarker in Taiwanese carriers of an APOBEC deletion polymorphism (2017)



Mutation Frequencies From TCGA/ICGC

	CGU%	US%	India%
CDKN2A	20	25	4
LCE4A	10	0	1
ORAI1	12	1	0
ZNF717	46	0	0
TP53	72	75	64
PABPC3	26	2	0
SKA3	20	1	0
NOTCH1	28	22	20
ODF1	12	1	0
FAT1	34	27	31
SETD8	20	1	0
CENPV	6	0	0
HRC	8	1	1
DHRS4	12	1	0
LOR	6	0	0
CTDSP2	6	1	1
CASP8	34	13	38
PIK3CA	26	17	10
RASA1	16	5	2

GATK Best Practices workflows

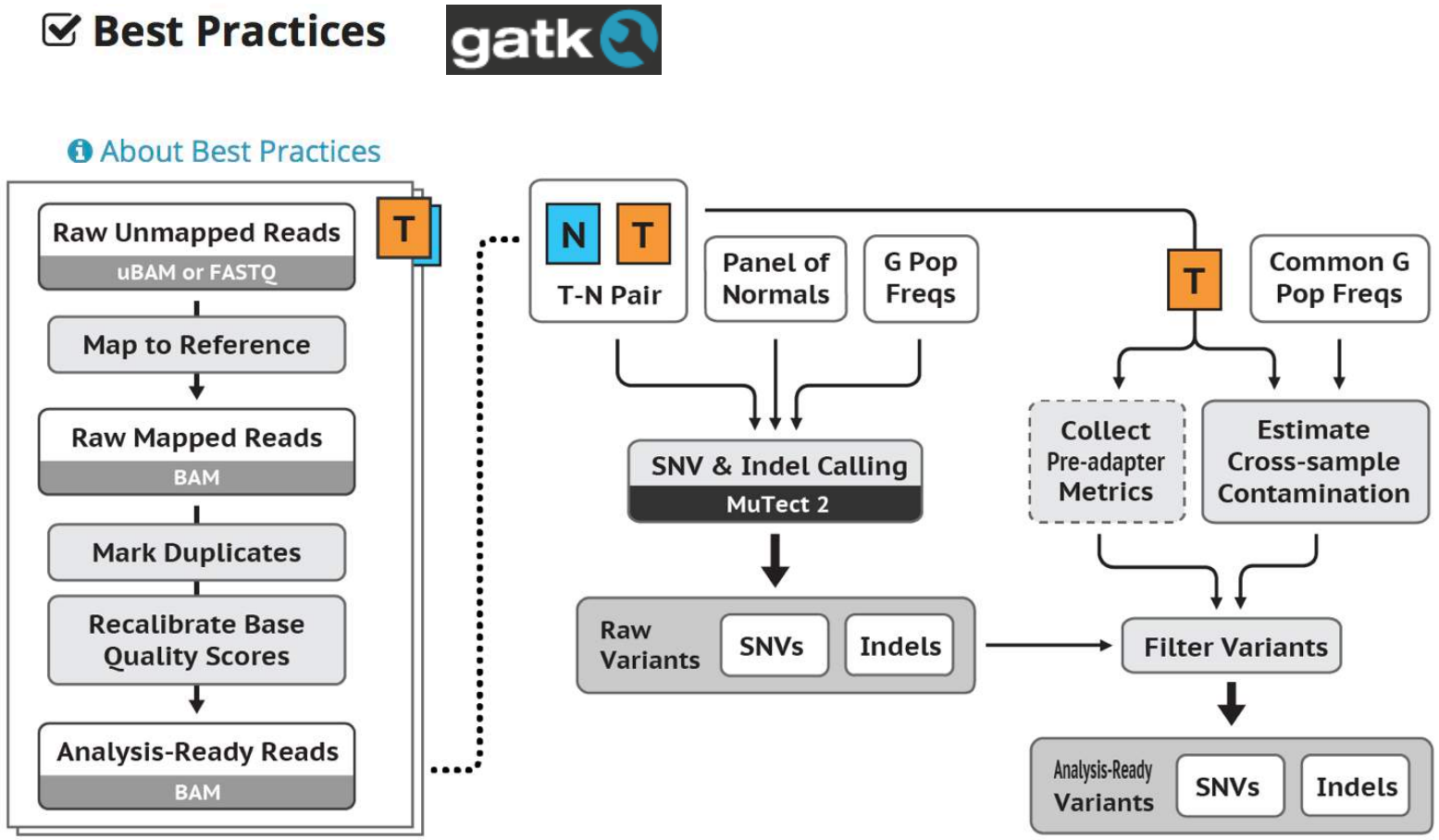


	GERMLINE	SOMATIC
SNPS & INDELS	**HaplotypeCaller GVCF**	**MuTect2** (BETA)
COPY NUMBER VARIANTS	**GATK gCNV** (ALPHA)	**GATK CNV + ACNV**
STRUCTURAL VARIANTS	*in development*	*in development*

Established tools / New tools

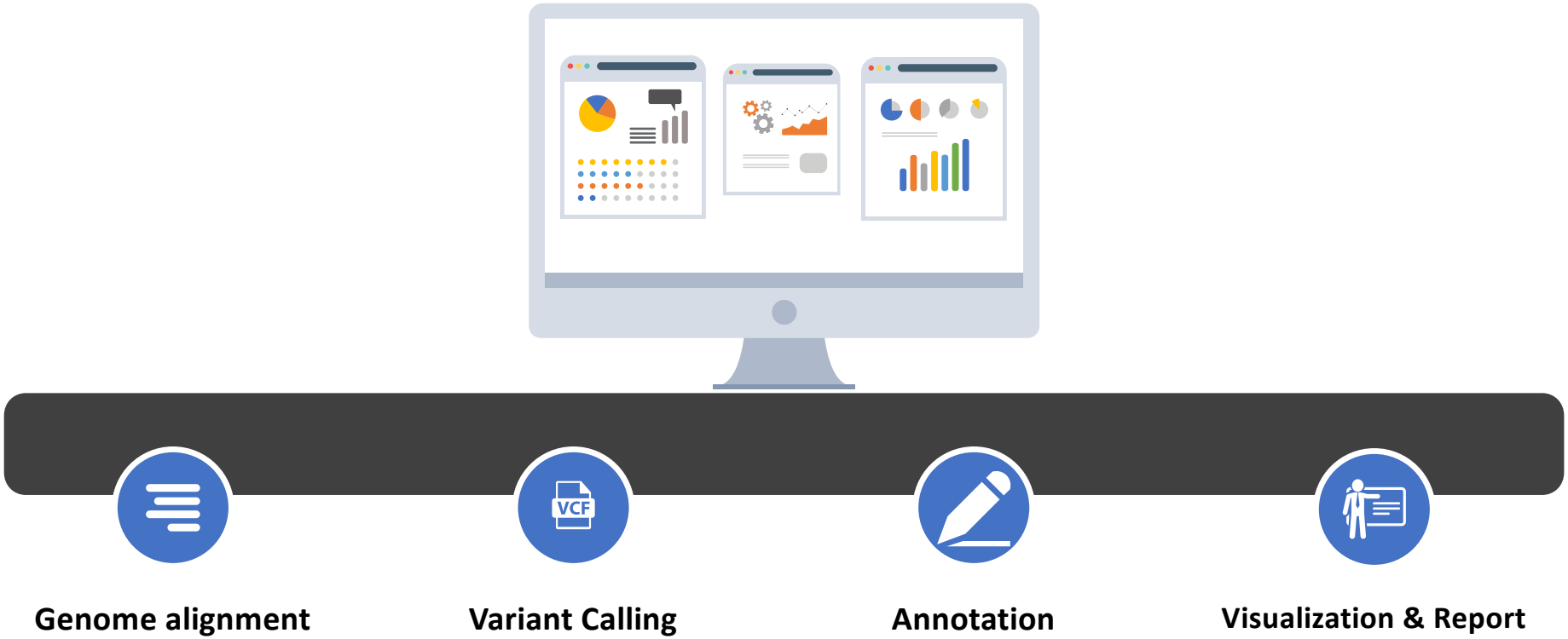
https://gatkforums.broadinstitute.org/gatk/discussion/9644/unboxing-gatk4

Somatic variant discovery (SNVs + Indels)



https://software.broadinstitute.org/gatk/best-practices/

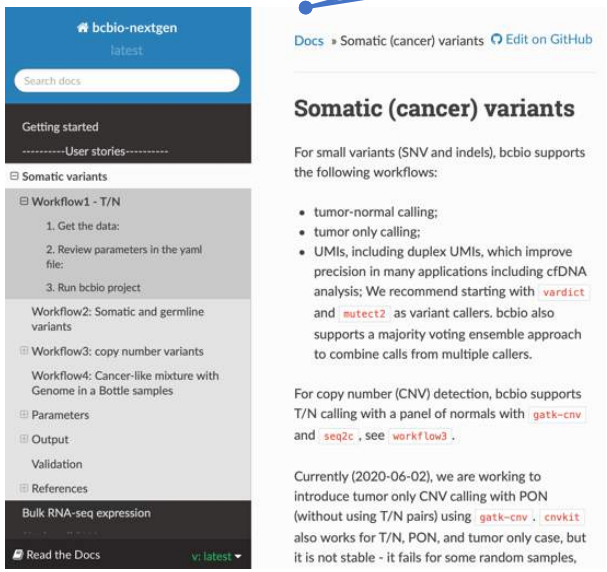
Bioinformatics Analysis Workflow

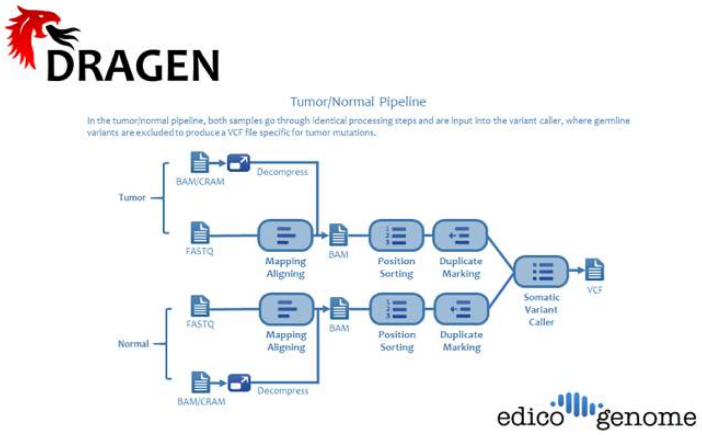


















https://bcbio-nextgen.readthedocs.io/en/latest/

https://basespace.illumina.com

長庚紀念醫院
Chang Gung Memorial Hospital

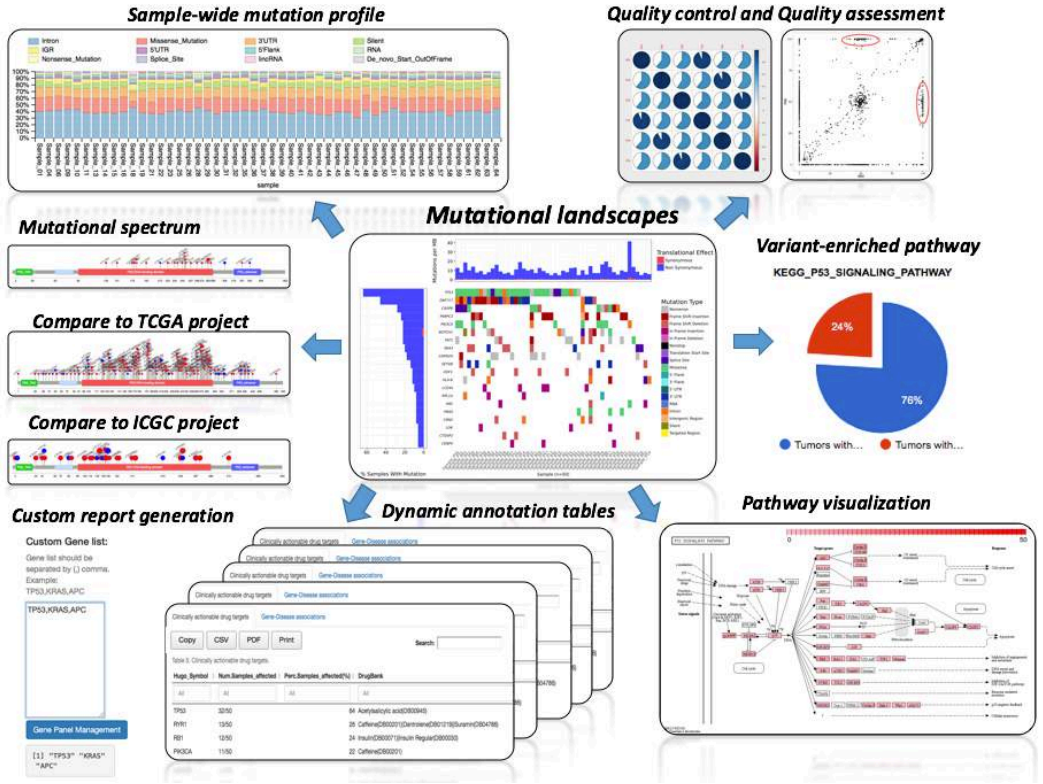
生物資訊分析平台收費標準（2019/10/25修正，2019/12/01生效）

以下分析皆含樣本數據品質檢測、前處理、伺服器核心使用費與資料儲存系統使用空間

服務項 目	單價 (元)	
	□長庚體系	□非長庚體系
RNA-seq標準分析	2,000元/樣本	3,800元/樣本
Small RNA sequencing標準分析	1,500元/樣本	2,000元/樣本
外顯子定序(WES)標準分析	2,000元/樣本	4,200元/樣本
全基因體定序(WGS)標準分析	5,000元/樣本	12,000元/樣本
Panel平台分析	1,000元/樣本	1,500元/樣本
16S metagenome sequencing標準分析	1,500元/樣本	2,000元/樣本
Pathway analysis標準分析	4,500元/次	7,000元/次
生物資訊客製化分析/整合(電洽)	1,000元/小時	1,500元/小時

註：此服務收費標準為未稅價，欲開發票者須外加5%稅金

VAReporter: Variant Reporter for clinical applications (Whole-exome analysis Pipeline)

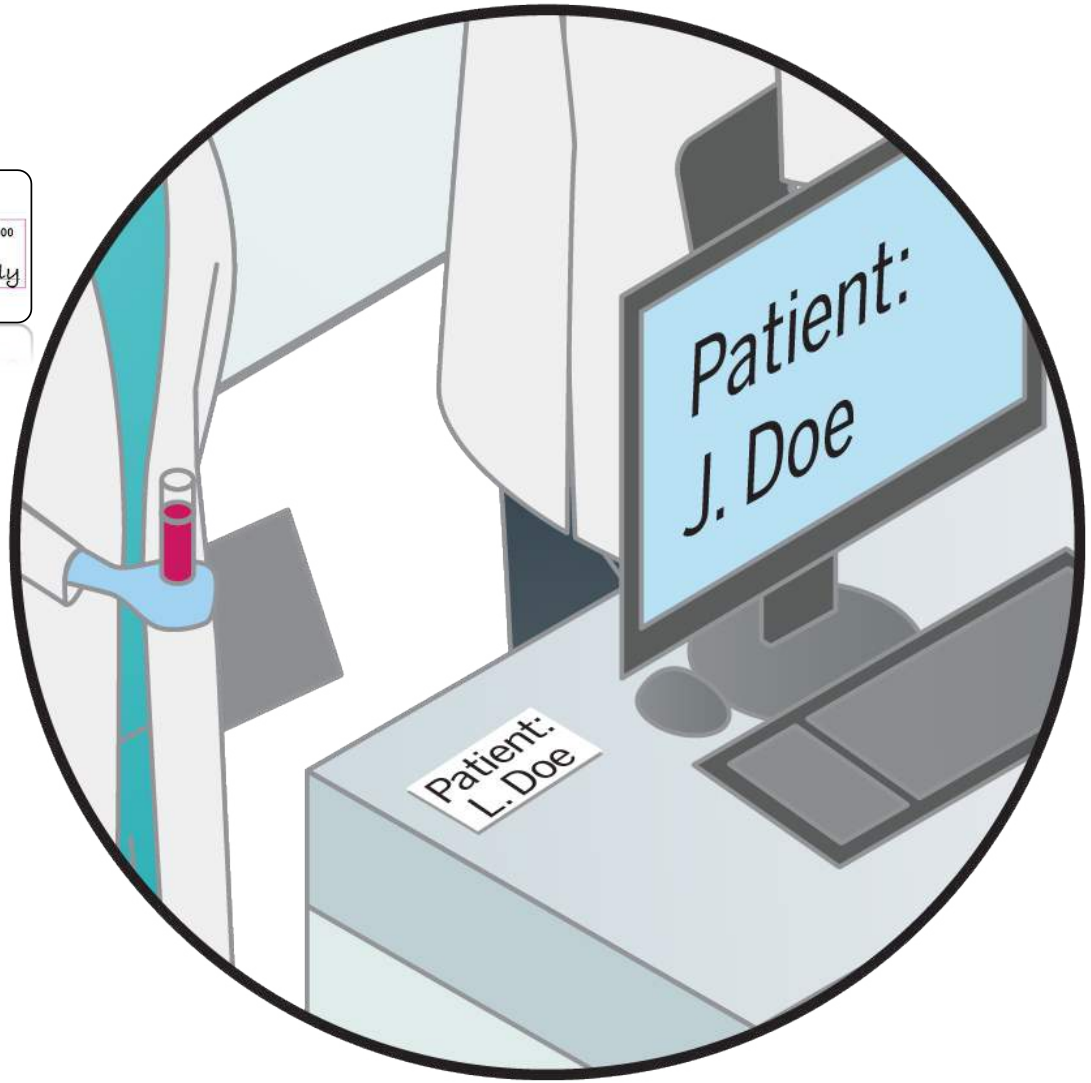


(http://rnd.cgu.edu.tw/vareporter)

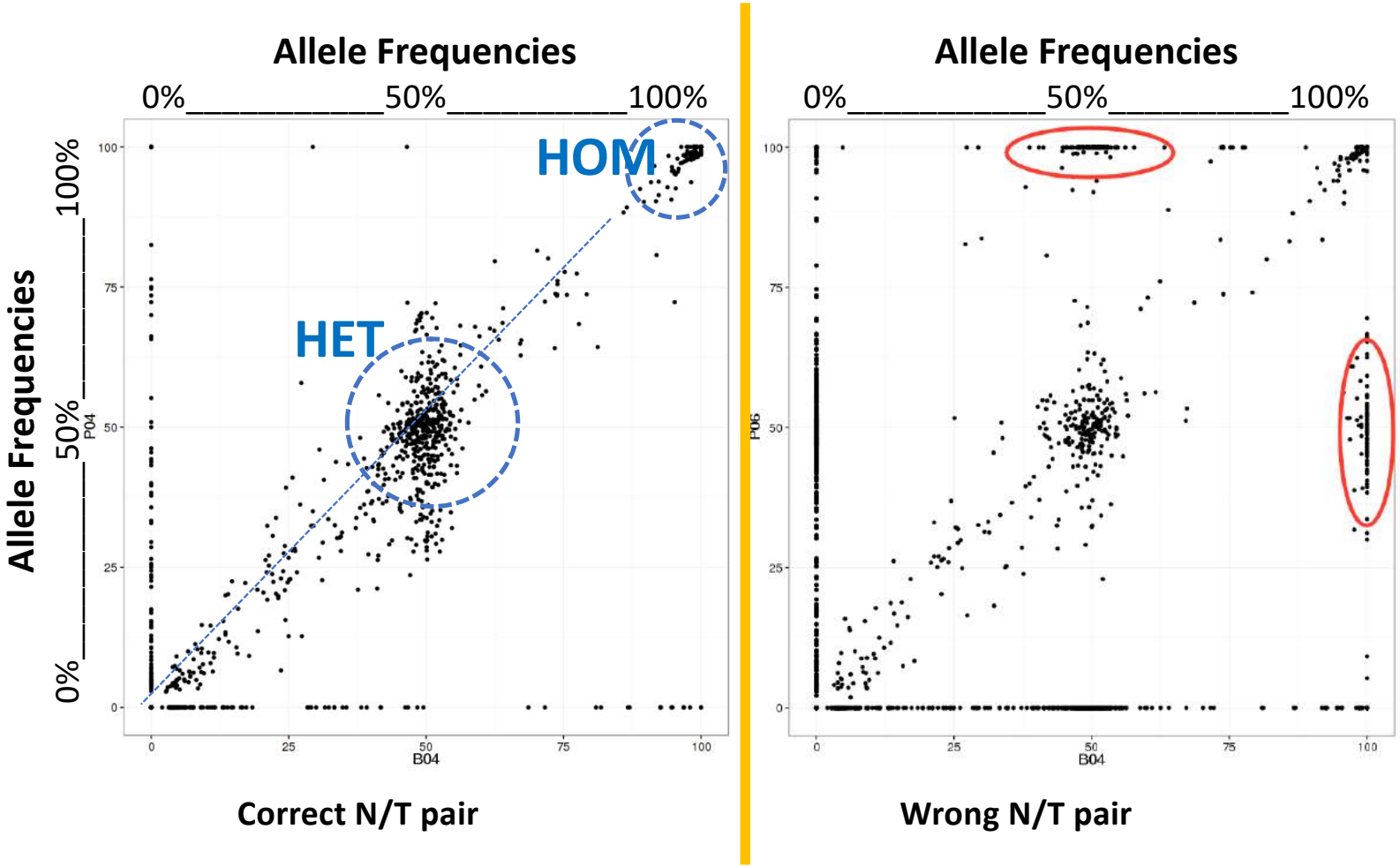
Mislabeled specimens





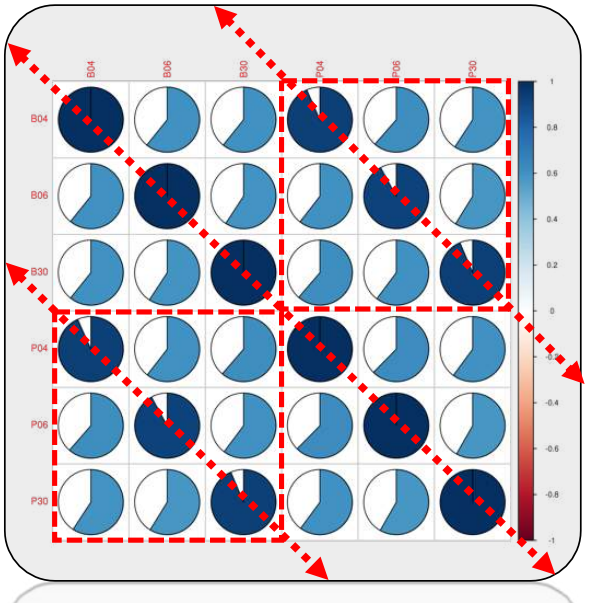


Identification of mislabeling errors

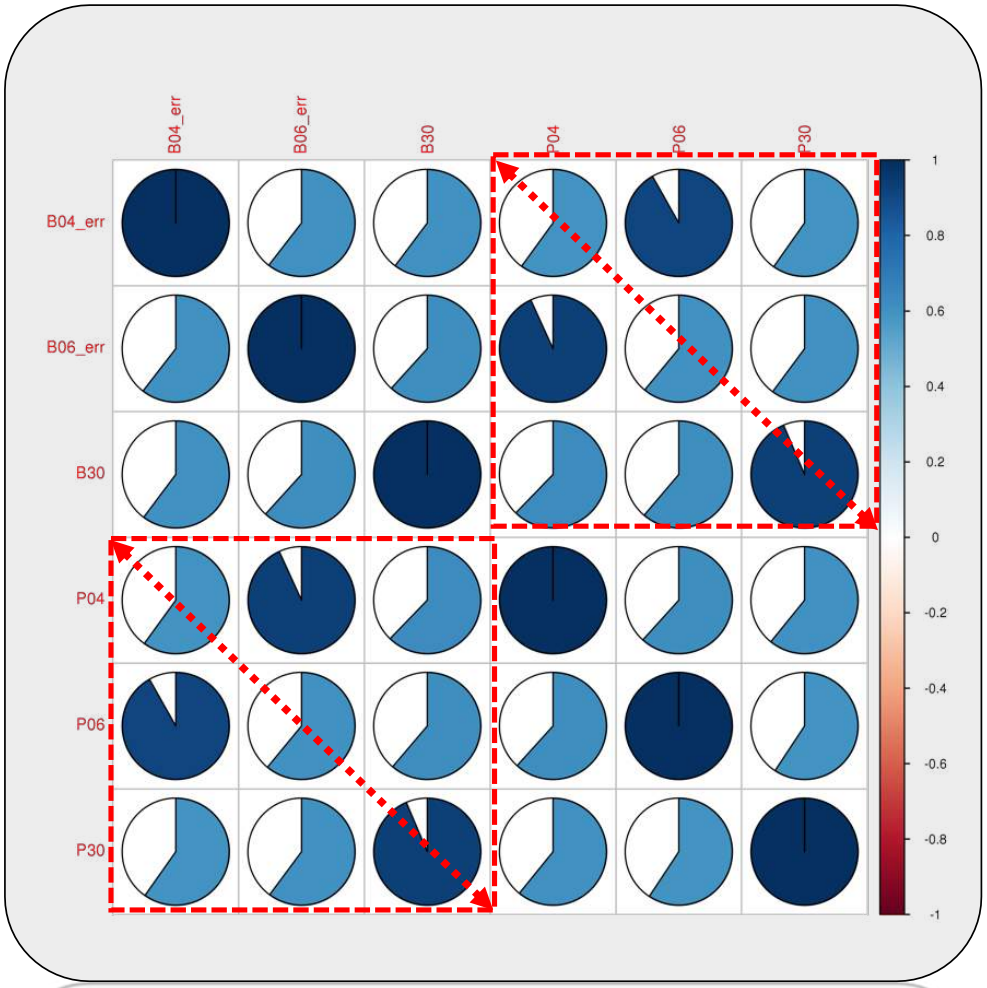


Visual analytics

Correct



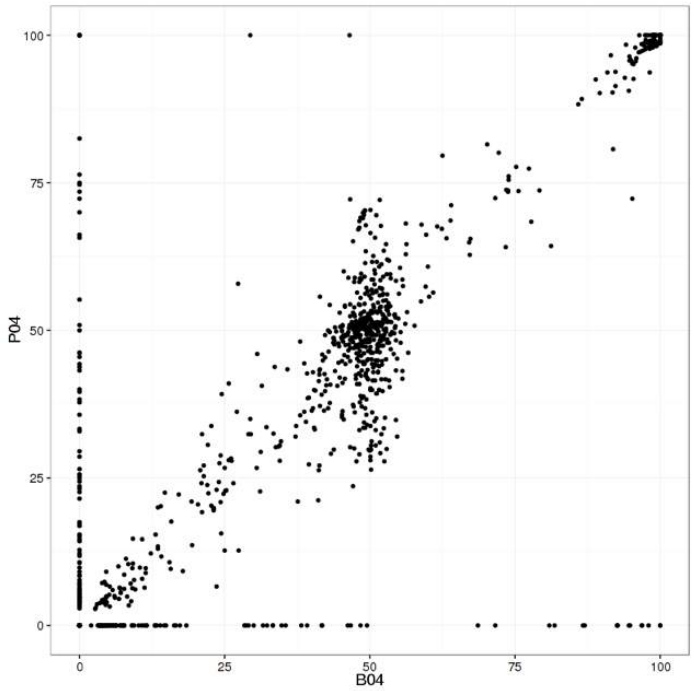
Error-prone



Real case

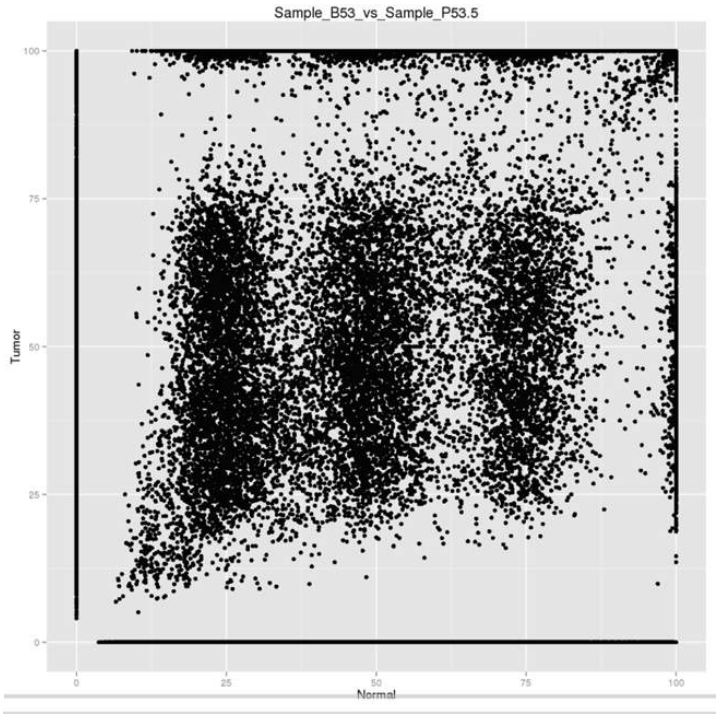
Allele Frequencies

0%.....50%.....100%



Allele Frequencies

0%.....50%.....100%



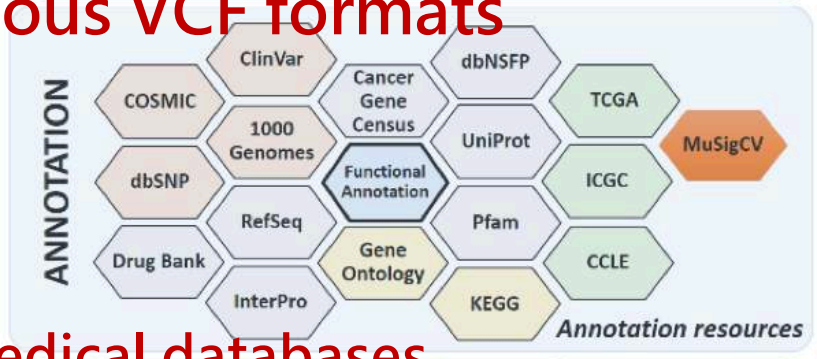
VAReporter Analysis Framework

VAReporter takes genetic variants from **gene panel** or **whole-exome sequencing** as input materials, supporting heterogeneous VCF formats such as GATK, VarScan, MuTect and VarDict.



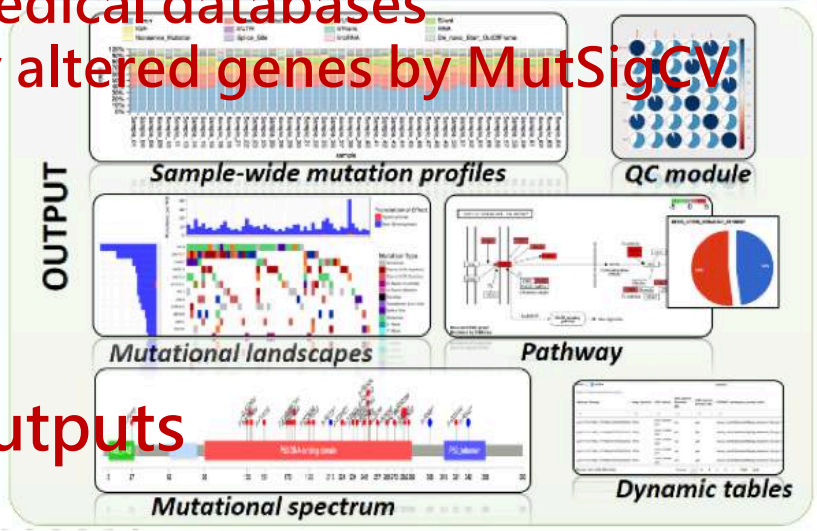
1.Support heterogeneous VCF formats

After successfully uploading the **VCF files** to VAReporter server, the identified variants of each sample are compare to a wide variety of **biomedical databases** that have been compiled as local annotation resources, to facilitate the interpretation of biological effects introduced by genetic alterations. The **MutSigCV algorithm** are adopted to detect **significantly altered genes** in study cohorts.



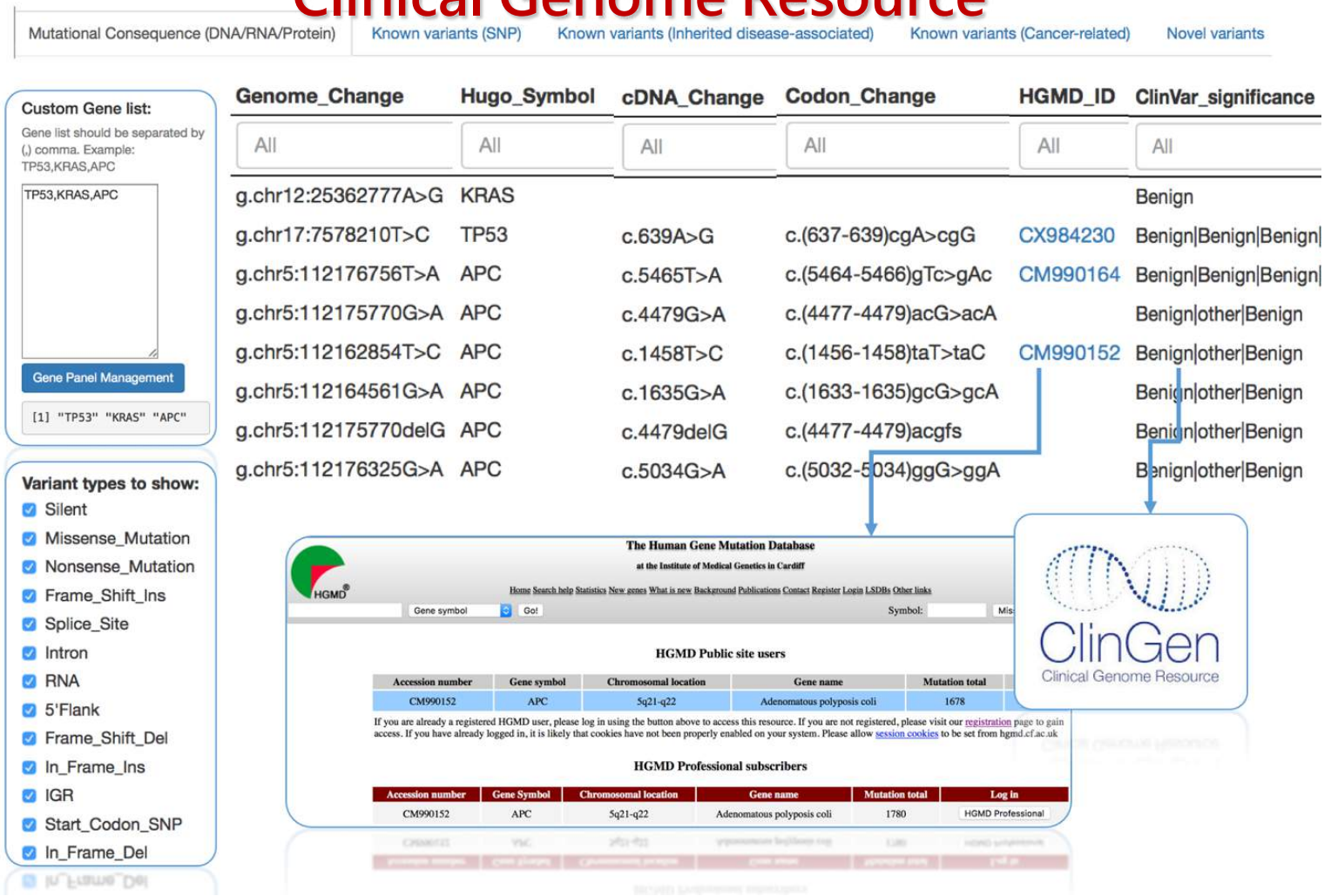
2.A Wide variety of biomedical databases
Detection of significantly altered genes by MutSigCV

Visualization modules are widely used for displaying sample-wide mutation profiles, landscapes, spectra and affected pathways, which are developed based on **ggplot2**, **corrplot**, **GenVisR**, **pathview** and **lollipop** packages. Dynamic tables with filtering and sorting functionalities are provided to facilitate the prioritization of clinically actionable targets, which are based on **R Shiny DT package**.



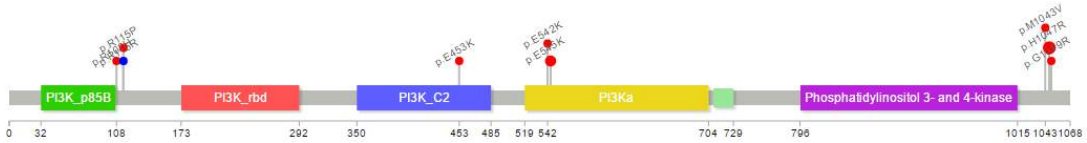
3.Rich visualization outputs

Custom Gene Panel, The Human Gene Disease Database & Clinical Genome Resource

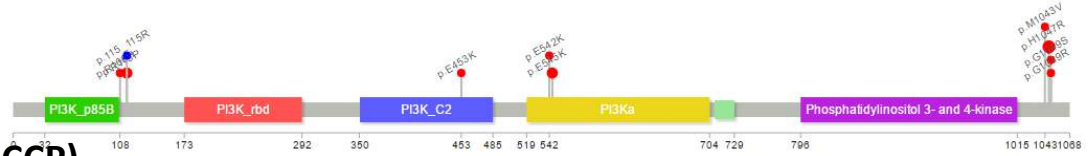


Experimental Validation & Cross Project Comparison

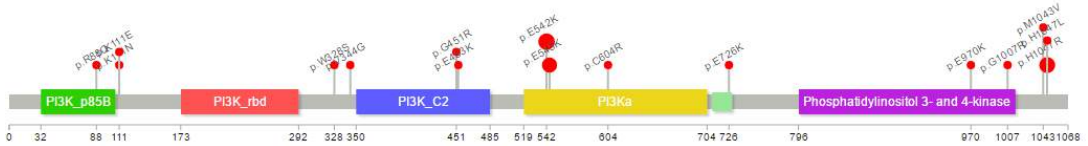
CGU (WES)



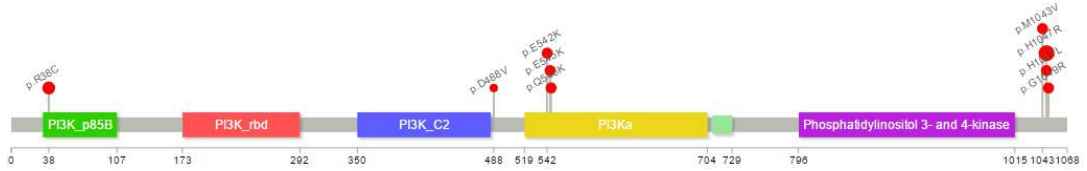
CGU
(Validation by CCP)



TCGA
(oral cavity)

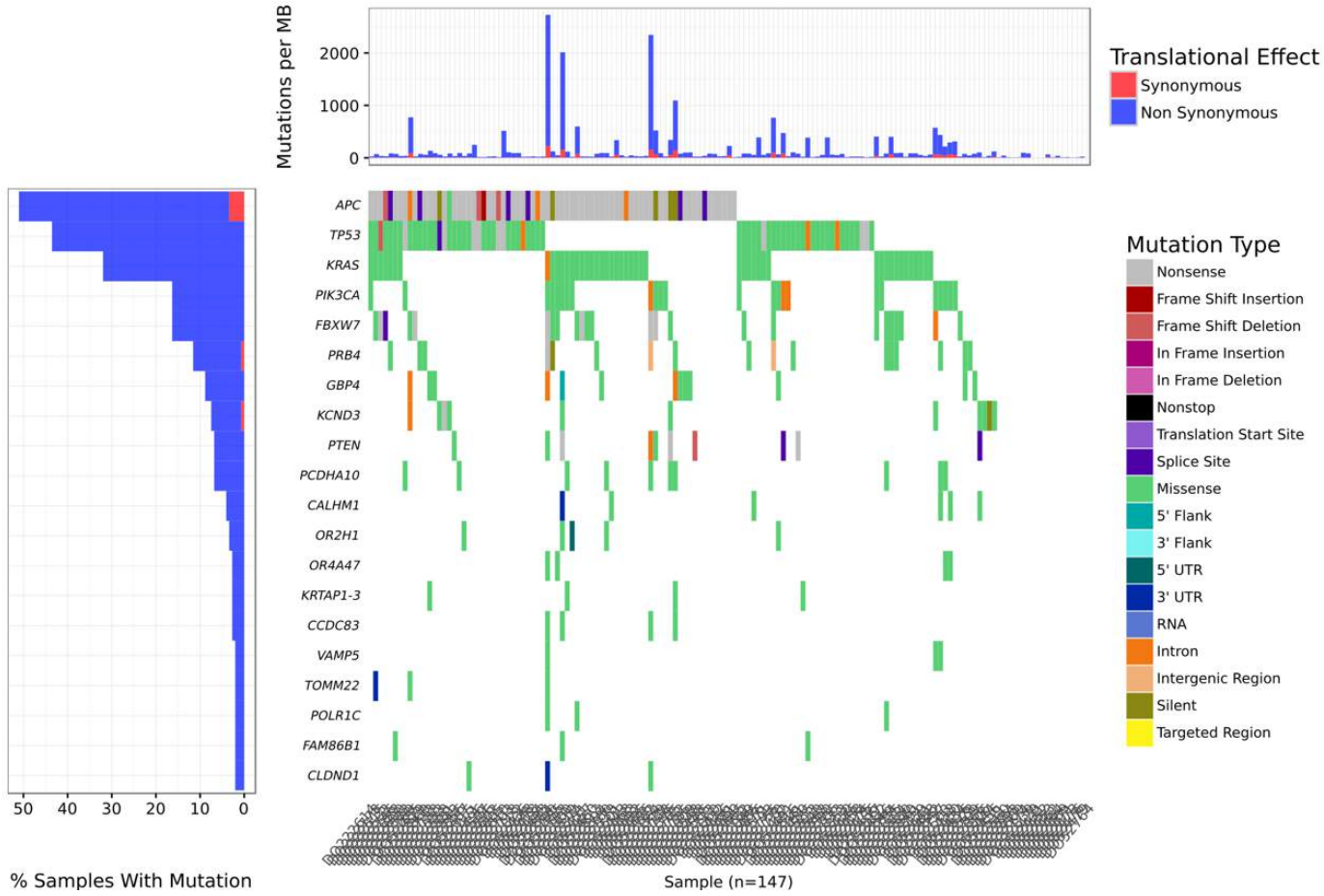


ICGC-India
(oral cancer)

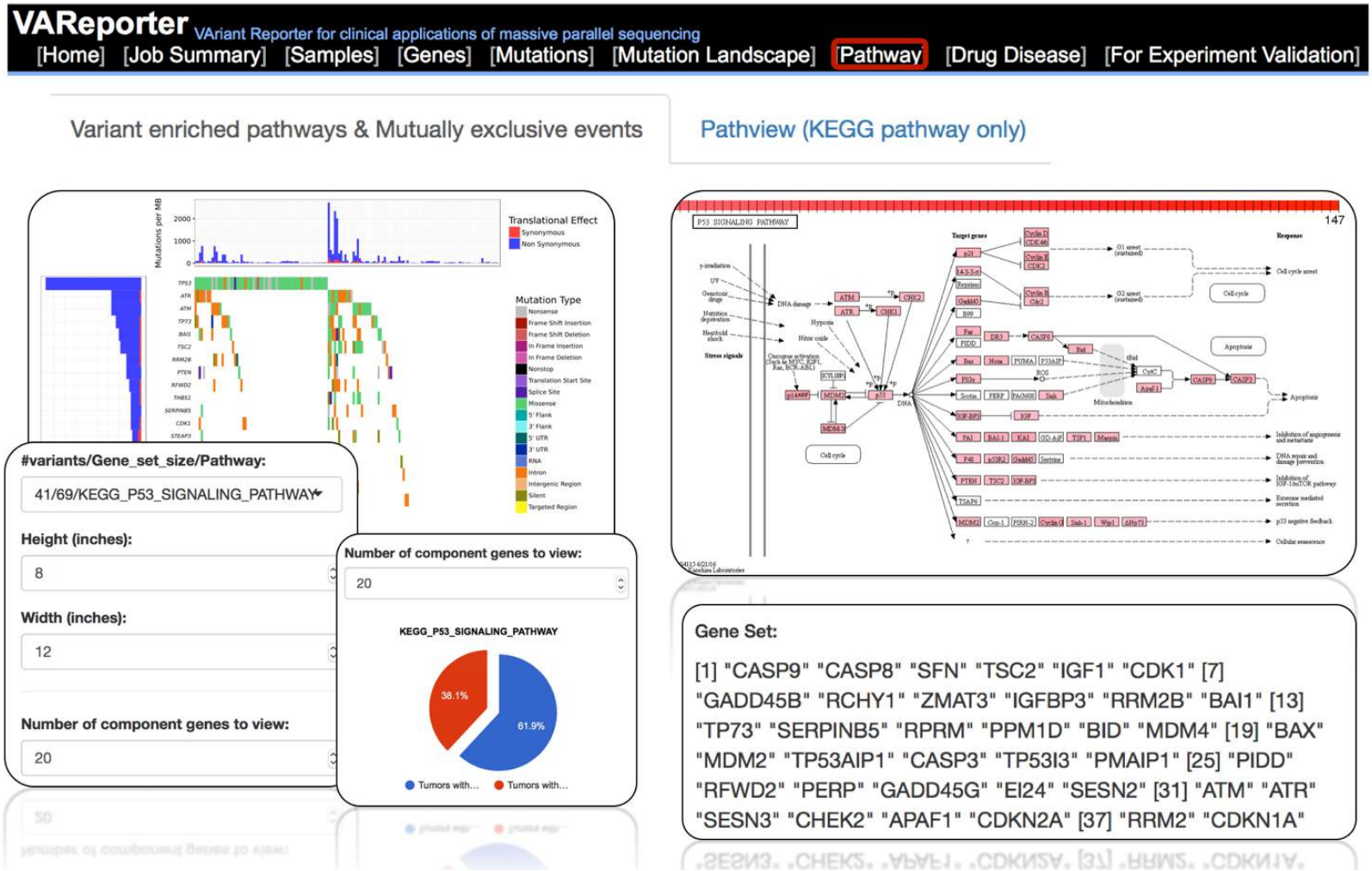


PIK3CA

Mutation Landscape (CoMut plot) + Significant altered genes identification (MuSigCV)



Identification of mutation enriched pathways



Clinically Actionable Drug Targets

Hugo_Symbol	Num.Samples_affected	Perc.Samples_affected(%)	Drug	Score	#PMIDs	#Sources	CIViC	CancerCommons
EGFR	All	All	All	A	A	Al	A	All
EGFR	8/216	3.7	GEFITINIB	26	13	13	v	
EGFR	8/216	3.7	ERLOTINIB	24	11	13	v	v
EGFR	/216	3.7	AFATINIB	12	4	8	v	
EGFR	8/216		ANDETANIE		0	8		
EGFR	8/216		ACO		1	6	v	

Showing 1 to 5 of (filtered from 2 to tri

Previous 1 2 3 4 5 ...

DGIdb

THE DRUG GENE INTERACTION DATABASE

Search Term	Drug	Interaction Type	Source	Human Curated	Full Gene Name	Source Version	PubMed IDs
EGFR	All	All	All	All	All	All	All
EGFR	EGFR - epidermal growth factor receptor			EGFR	epidermal growth factor receptor	4.3	11752352, 12840797, 12517254, 12814826, 12820772, 12498017
EGFR	EGFR - epiderm growth factor receptor			EGFR	epidermal growth factor receptor	4.3	11673690, 11752352, 11585753, 11522647, 10815932, 11566608
EGFR	EGFR - epiderm growth factor receptor			EGFR	epidermal growth factor receptor	4.3	16551906
EGFR	EGFR - epiderm growth factor receptor		Genome	EGFR	epidermal growth factor receptor	13-March-2013	
EGFR	EGFR - epiderm growth factor receptor		Genome	EGFR	epidermal growth factor receptor	13-March-2013	

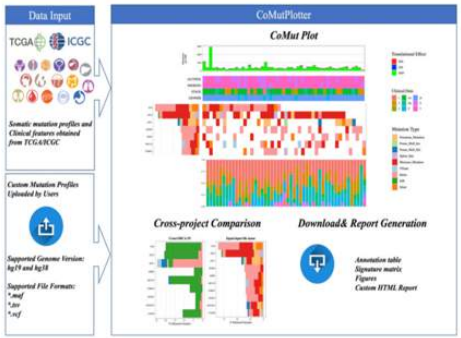
HGMD®

News Letter

TBI Core Facility

108年08月15日

介紹癌症基因體變異視覺化整合分析工具CoMutPlotter

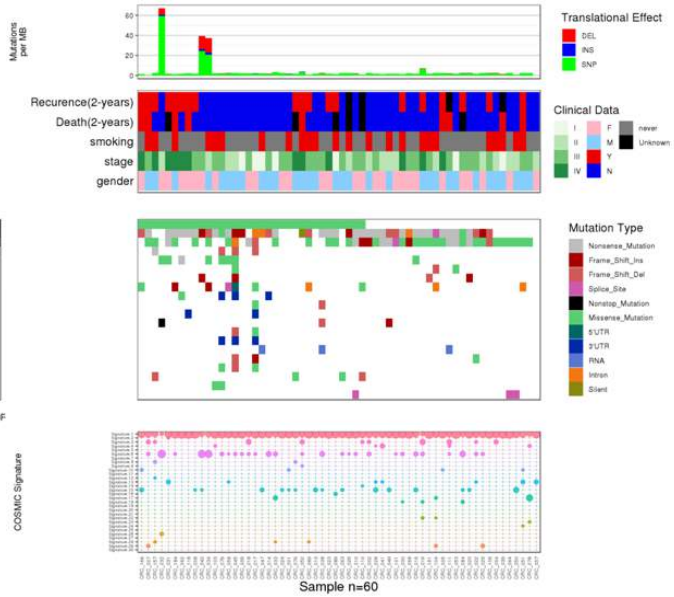


CoMut plot是廣泛應用於癌症基因體研究，利用條狀圖將群體間突變頻率最高的基因由高到低排列出來，同時也可以清楚找出基因體突變最多的個體。此外，也用熱點圖來呈現個體的每個特定基因上突變的程度與變異的種類。最後利用程式語言將這些圖縫合在一起，僅用一張圖就能呈現研究群體中個體間的基因體變異圖譜。看似簡單，但需要經過複雜的檔案格式轉換、變體位點註釋、顯著變異基因預測、變異種類統計分析、突變特徵分析等繁鎖的分析過程。雖然目前已有少數軟體可以用來進行CoMut Plot的製作，但存在著幾個缺點：1.不支援主流的檔案格式(如：VCF格式) 2.缺乏預測顯著變異基因與突變特徵的功能 3.缺乏跨癌症群體比較功能。

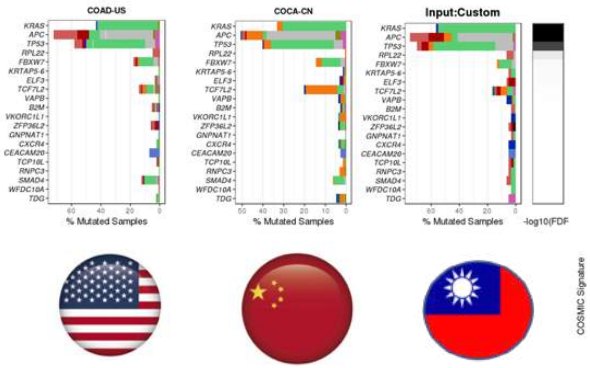
有鑑於此，本團隊長庚大學黃柏榕老師實驗室開發了一個網頁工具: CoMutPlotter (http://tardis.cgu.edu.tw/comutplotter/)。使用者不需生物資訊背景就能夠自行操作，上傳研究群體的癌症基因體突變資料，進行全自動分析與產生CoMut plot圖表。CoMutPlotter支援多種基因體變異資料格式(TSV, MAF,VCF)，變體位點經過基因功能註解、癌症驅動基因找尋和突變特徵辨認等分析流程，最後將所有結果整合繪製成CoMut plot。除此之外，使用者將自己的研究資料與現有癌症基因體資料庫(TCGA/ICGC)的資料進行比較分析並合併繪圖。使用者能輕易的使用CoMutPlotter比較不同國家的癌症資料差異性。最重要的是，所有比較分析圖表都能讓使用者自由下載。

此工具已於2019年中發表於BMC Medical Genomics期刊。

1st web tool provides the **most-wanted** function for Comparing mutation patterns between **custom cohorts** and **TCGA/ICGC projects**



Cross-project Comparison

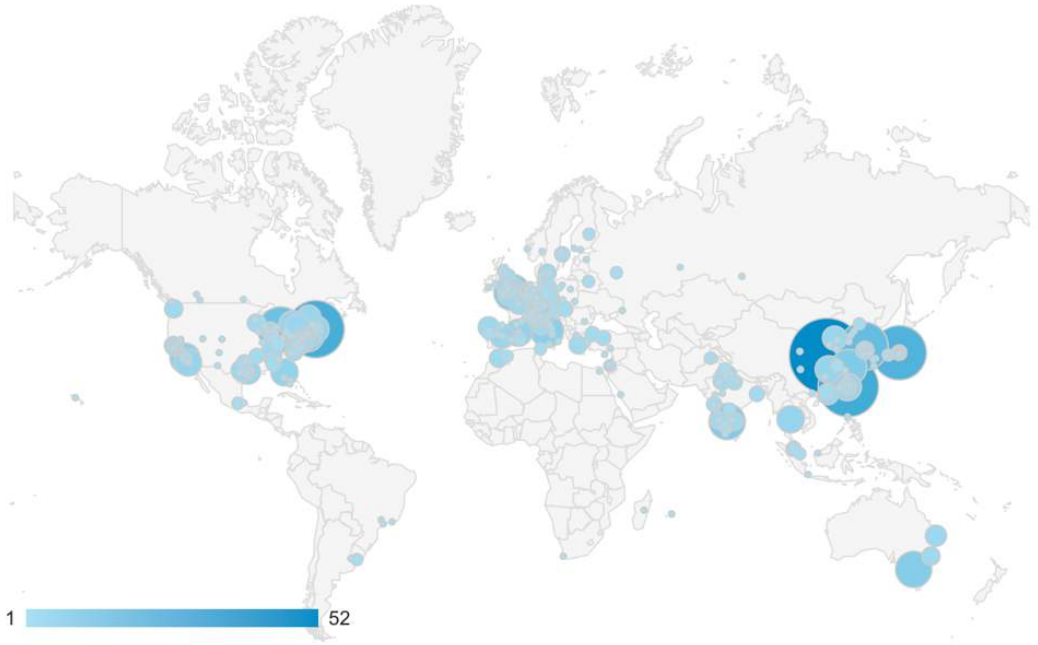


- **22,961 Exomes**
- **84 TCGA/ICGC Cancer Projects**

http://tardis.cgu.edu.tw/comutplotter/

＞1,000 Users Since 2019.7.1 (CoMutPlotter)

Country	Users	New Users	Sessions	Bounce Rate
	1,001 % of Total: 100.00% (1,001)	997 % of Total: 100.30% (994)	1,791 % of Total: 100.00% (1,791)	36.18% Avg for View: 36.18% (0.00%)
1. United States	264 (26.11%)	265 (26.58%)	513 (28.64%)	38.79%
2. China	107 (10.58%)	102 (10.23%)	126 (7.04%)	63.49%
3. Taiwan	86 (8.51%)	79 (7.92%)	264 (14.74%)	34.47%
4. India	57 (5.64%)	57 (5.72%)	81 (4.52%)	23.46%
5. Italy	56 (5.54%)	56 (5.62%)	138 (7.71%)	28.99%
6. Japan	50 (4.95%)	50 (5.02%)	61 (3.41%)	65.57%
7. United Kingdom	48 (4.75%)	48 (4.81%)	77 (4.30%)	25.97%
8. Germany	33 (3.26%)	33 (3.31%)	52 (2.90%)	25.00%
9. South Korea	33 (3.26%)	33 (3.31%)	74 (4.13%)	20.27%
10. Spain	30 (2.97%)	30 (3.01%)	50 (2.79%)	24.00%



BMC Medical Genomics, 12(Suppl 5):99.