AMS KKU
คณะเทคนิคการแพทย์ มหาวิทยาลัยขอนแก่น
FACULTY OF ASSOCIATED MEDICAL SCIENCES KHONKAEN UNIVERSITY

Cancer Genomics Research in ASIA and Online Databases for Cancer Analysis

Asst. Prof. Apinya Jusakul, Ph.D.
Faculty of Associated Medical Sciences
Khon Kaen University

Wellcome Connecting Science –
Cancer Genome Analysis Asia 2025

From Single-Case Studies to Population-Scale Genomics

- Early genomic studies focused on single cases or small families—identifying causative mutations in Mendelian disorders.

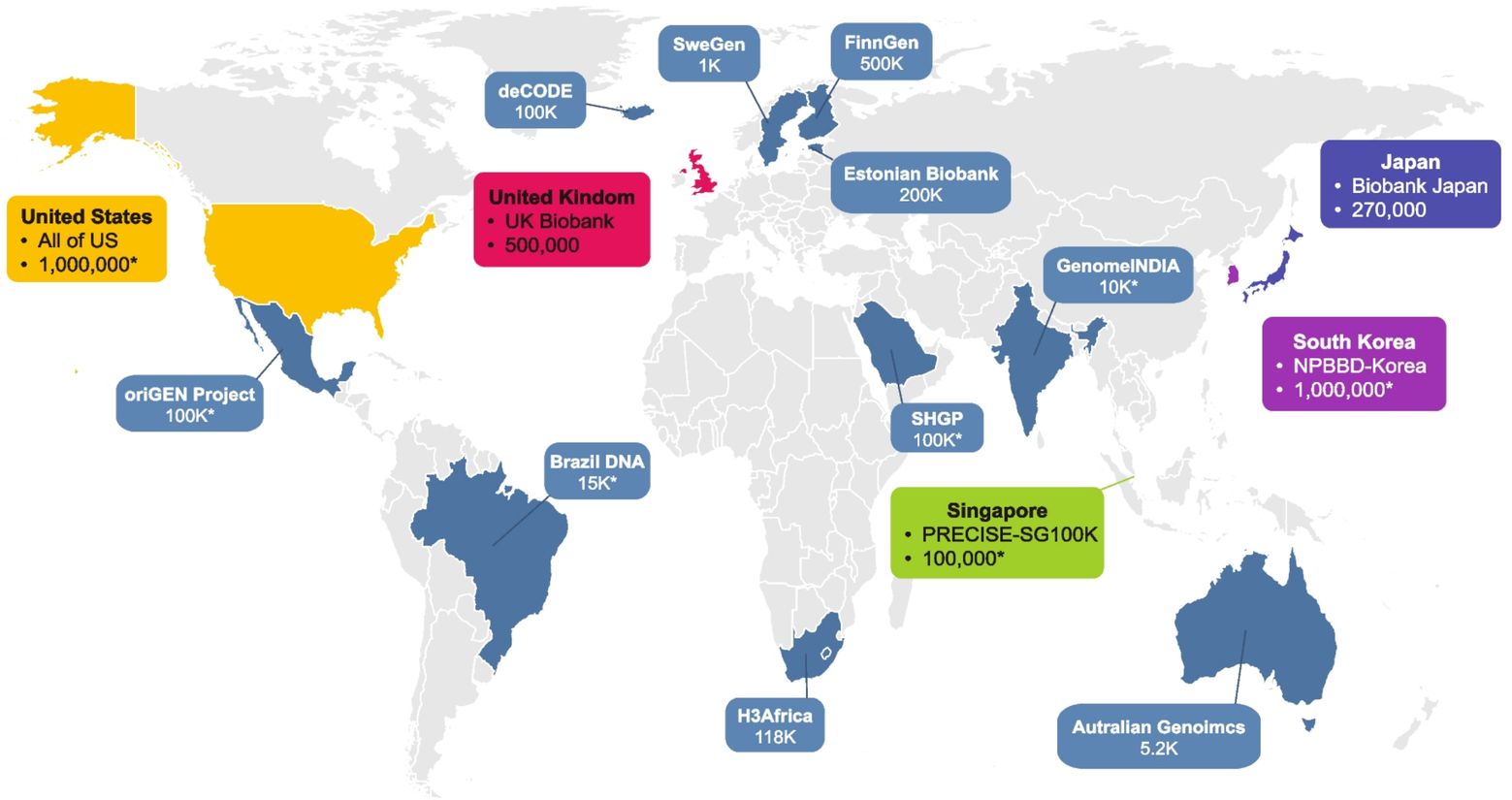
- The next phase expanded to disease-based cohorts (e.g., cancer, diabetes, cardiovascular studies) using exome and genome sequencing to find disease-associated variants.

- Progress in sequencing cost and computational power enabled population-level genomic projects, providing reference datasets for genetic diversity and disease association.

- Today, national biobank initiatives (e.g., UK Biobank, Biobank Japan, Korea Bio-Big Data, China Biobank, Singapore's PRECISE) integrate whole-genome sequencing with phenotypic and clinical data, supporting precision medicine at scale.

Lessons from national biobank projects utilizing WGS for population-scale genomics

a



b

Biobank	Country	Samples (available WGS)	Ancestry	Cohort
All of Us	United States	245,000	Multi-ethnic	Population-based
UK Biobank	United Kingdom	500,000	European	Population-based
PRECISE-SG100K	Singapore	50,000	Chinese, Maly, Indian	Population-based
NPBBD-Korea	South Korea	25,000	Korean	Population-based, Diseased
Biobank Japan	Japan	14,000	Japanese	Population-based, Diseased

Lee, H., Kim, W., Kwon, N. et al. Genom. Inform. 23, 8 (2025)

Key Asian National Genomics Projects

Country	Main Project(s)	Description
Singapore	NPM, SG100k, SGVP	Sequencing tens of thousands of genomes
India	GenomeIndia, GenomeAsia 100K	Nationwide sequencing, rare disease focus
Indonesia	Indonesian Genome Project	Population cohort genetics
Malaysia	Malaysia Genome Project	National biobank, ASEAN participation
Thailand	Thailand Genome Project	National efforts, ASEAN consortium
Vietnam	Vietnam Genome Project	Regional partnership
Uzbekistan	Uzbekistan Genome Project	1,000 Genomes; hereditary mutations research
China, Japan, Pakistan, etc.	GenomeAsia 100K	Regional reference panels

SG100K: Translating insights from 100,000 genomic data sets into improved health strategies

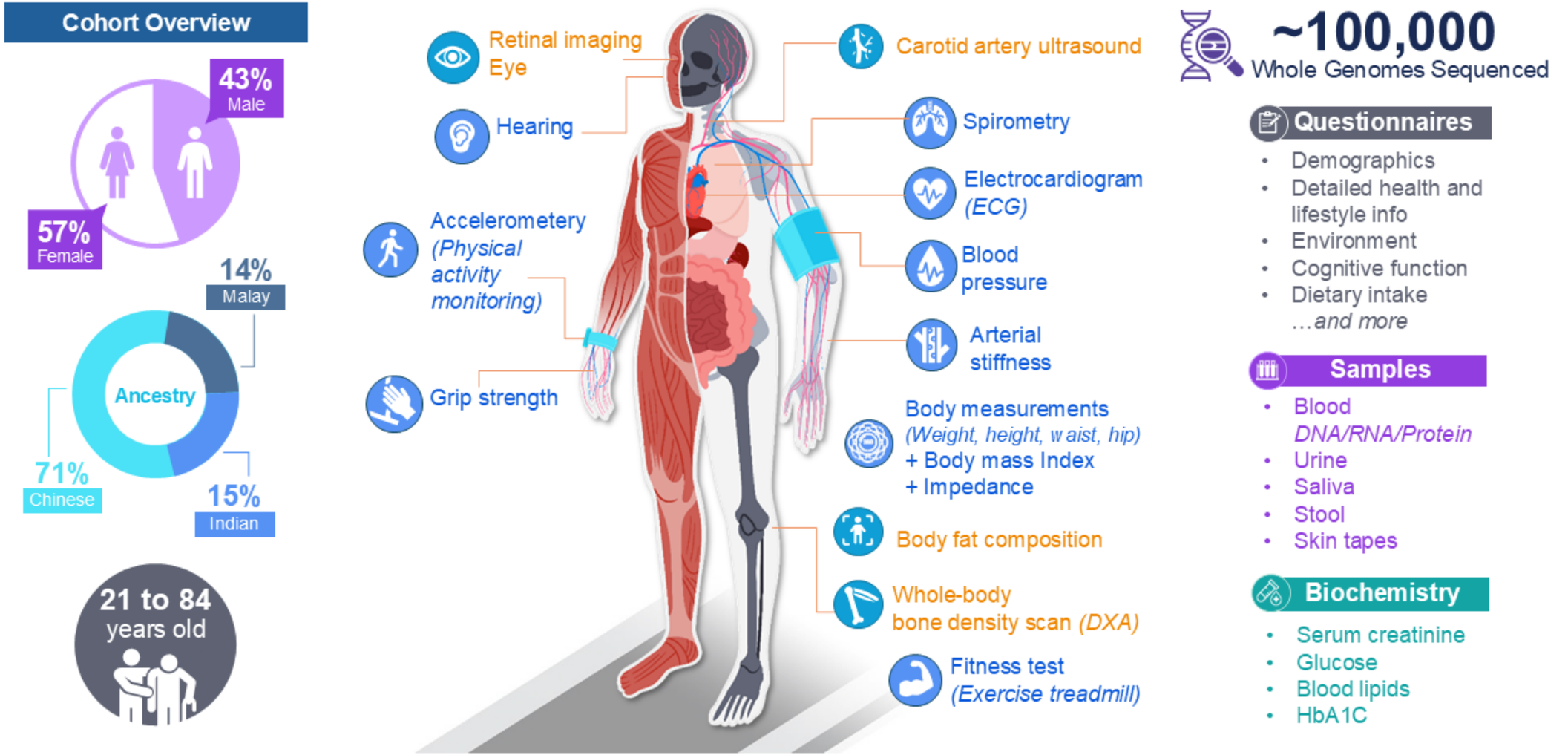


Illustration by Ms Trixie Teo, CRIS Corporate Communications

https://www.npm.sg/sg100k-overview

STRAITS TIMES

Project SG100K: DNA of 100,000 Singaporeans to be mapped to identify new ways to prevent diseases

Together, these rich datasets make SG100K one of the most comprehensive national precision health programs in Asia, integrating genomics with lifestyle and clinical information to support **preventive healthcare and policy planning**."

https://www.straitstimes.com/singapore/health/dna-of-100000-singaporeans-to-be-mapped-to-further-populations-health

Exemplary Research Findings from SG100K

- **Polygenic Risk Scores:** SG100K researchers are generating common variant polygenic risk scores for prevalent cancers, such as breast, colorectal, liver, lung, and prostate cancers. This approach takes into account the diverse ancestral backgrounds within the Singaporean population, improving the accuracy of risk predictions for Asian cohorts.

- **Rare Variant Discovery:** The project identifies rare coding genetic mutations with functional significance in established cancer-related genes, yielding insights into cancer predisposition specifically in Asian ancestries, where such data has traditionally been scarce.

- **Clinical Integration:** By linking genetic datasets with national health registries and medical records, SG100K can determine cancer status and outcomes for participants. This enables the development of genetic risk models directly relevant to real-world clinical contexts for Singapore's multi-ethnic population.

- **Age-Related Biomarkers:** Research teams are also deriving age-related cancer biomarkers (e.g., telomere length) from whole-genome sequencing data and studying their association with cancer risk in various ethnic groups within Singapore.

The variant landscape identified from the first phase of the SG100K Whole-Genome Sequencing project.

Key Findings on Genetic Variants

1. Identified around **179 million variants**
2. Majority were **single-nucleotide variants (SNVs)**
3. Each genome had **4 to 5 million variants**
4. About **1% of variants were rare**
5. Rare variants had minor allele frequency below 1%

Table 2 | Variants observed in all individuals in NPM Phase I SG10K_Health WGS

	Variants in WGS		Median variants per genome			
	Number of variants	Number of variants with MAF <1%	Number of variants	s.d.	Number of variants with MAF <1%	s.d.
Total variants	179,418,917	166,559,124	4,106,905	207,880.84	393,367	37,222.15
Variant type						
SNVs	158,331,366	148,665,318	3,501,477	129,196.15	346,317	32,594.75
Indels	21,087,551	17,893,806	602,448	84,126.34	47,139	4,794.15
Functional prediction						
Synonymous	595,783	566,740	10,940	405.29	1,084	123.49
Missense	1,147,135	1,114,839	11,147	447.75	1,469	155.03
LoF	104,569	101,917	722	44.33	82	11.64
LoF (high confidence)	63,224	62,324	207	19.61	35	6.23

Median counts and s.d. counts per genome are also presented. See Methods for details on genome mapping, joint variant calling and quality-control filters. MAF, minor allele frequency.

Table 3 | Observed number of clinically relevant variants categorized based on ClinVar clinical relevance values: benign, likely benign, likely pathogenic, pathogenic and variants of unknown relevance

Clinical relevance	Total variants		
	Private	Rare	Common
Benign	10,888	37,397	83,435
Likely benign	19,547	46,116	7,214
Likely pathogenic	1,029	756	3
Pathogenic	1,315	1,072	20
VUS	25,150	41,735	1,352
Other	328	704	654

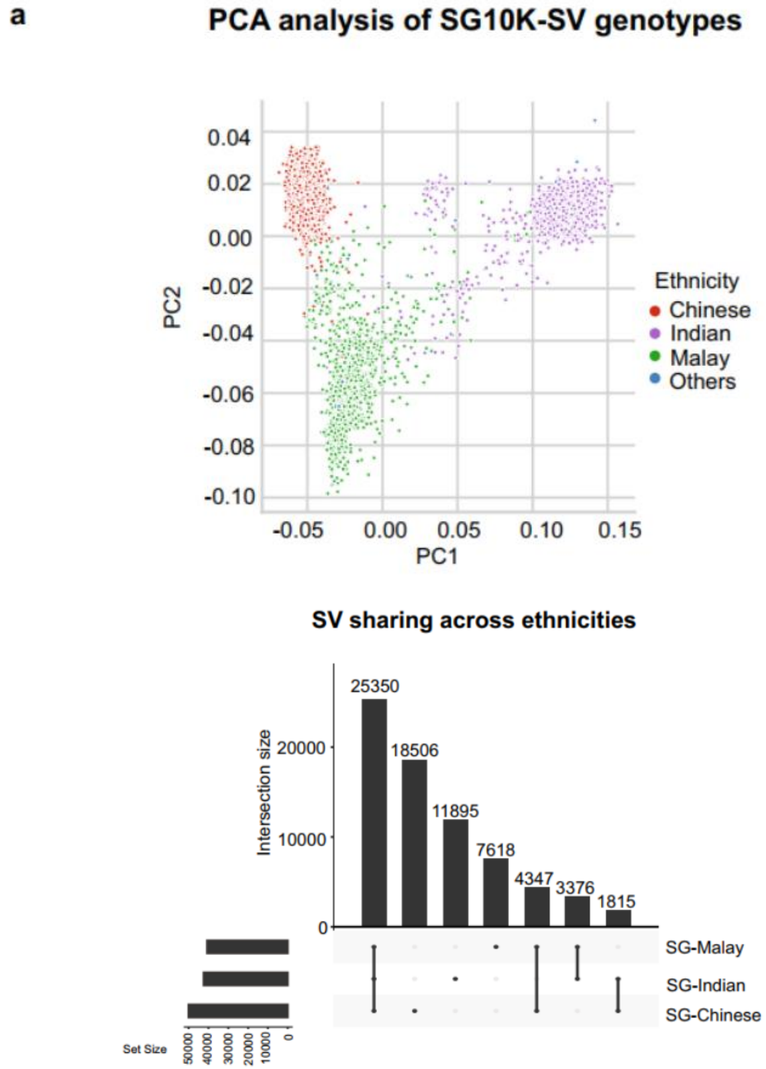
Numbers of private (variant found in only one individual), rare (allele frequency ≤1% and more than one individual) and common (allele frequency >1%) variants are shown. See Methods for details on variant classification. VUS, variants of unknown significance.

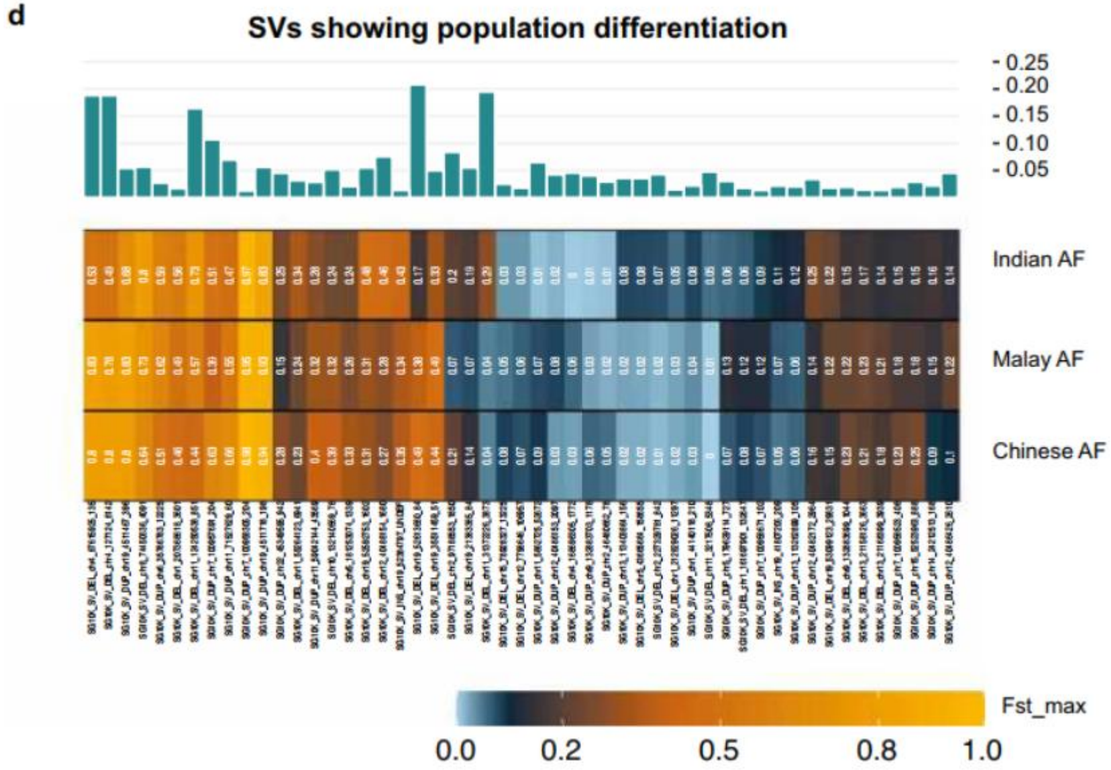
- The study by Wong and colleagues, published in Nature Genetics in 2023, reveals over 25,000 variants of uncertain significance in Singapore's multi-ethnic population.
- This underscores the substantial genetic diversity present and emphasizes the urgent need for additional functional and clinical data to better interpret rare and novel variants, particularly in Asian genomes.

Wong, E., Bertin, N., Hebrard, M. et al. The Singapore National Precision Medicine Strategy. Nat Genet 55, 178–186 (2023).

A Catalogue of Structural Variation across Ancestrally Diverse Asian Genomes

Population specificity of SVs





Tan, J.H.J., Li, Z., Porta, M.G. et al. A Catalogue of Structural Variation across Ancestrally Diverse Asian Genomes. Nat Commun 15, 9507 (2024).

Genomics Thailand

Phase 1: Genomics Thailand Integrated Action Plan (2020–2024)

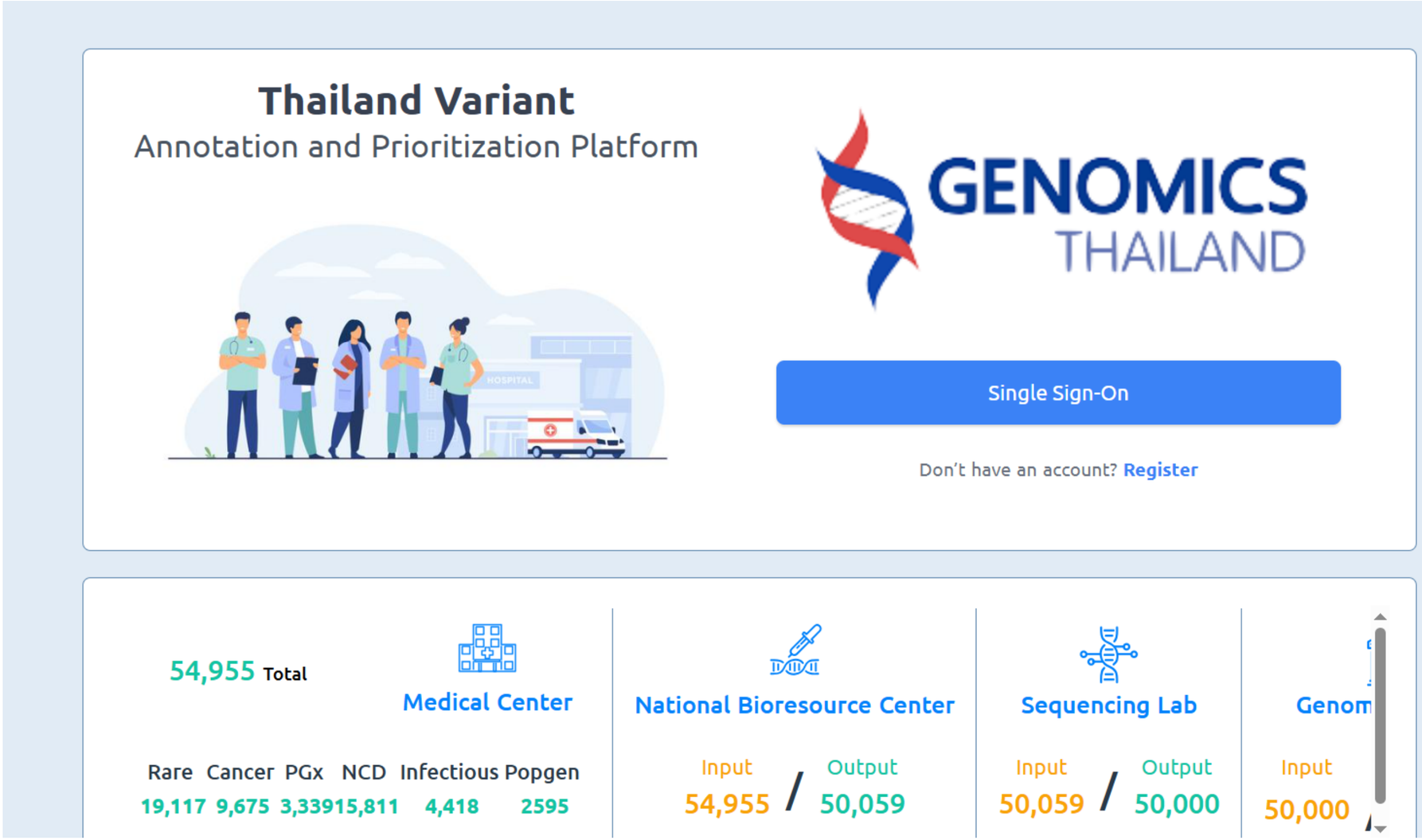
- This five-year strategic program has been covered the research area in genomic epidemiology to estimate genetic factors that influence susceptibility to specific diseases or traits among Thais, thus advancing the genomic-based testing/screening program for clinical and public health implications in Thailand.

Primary actions:

Whole-genome sequencing was performed on 50,000 Thai individuals within five disease groups:

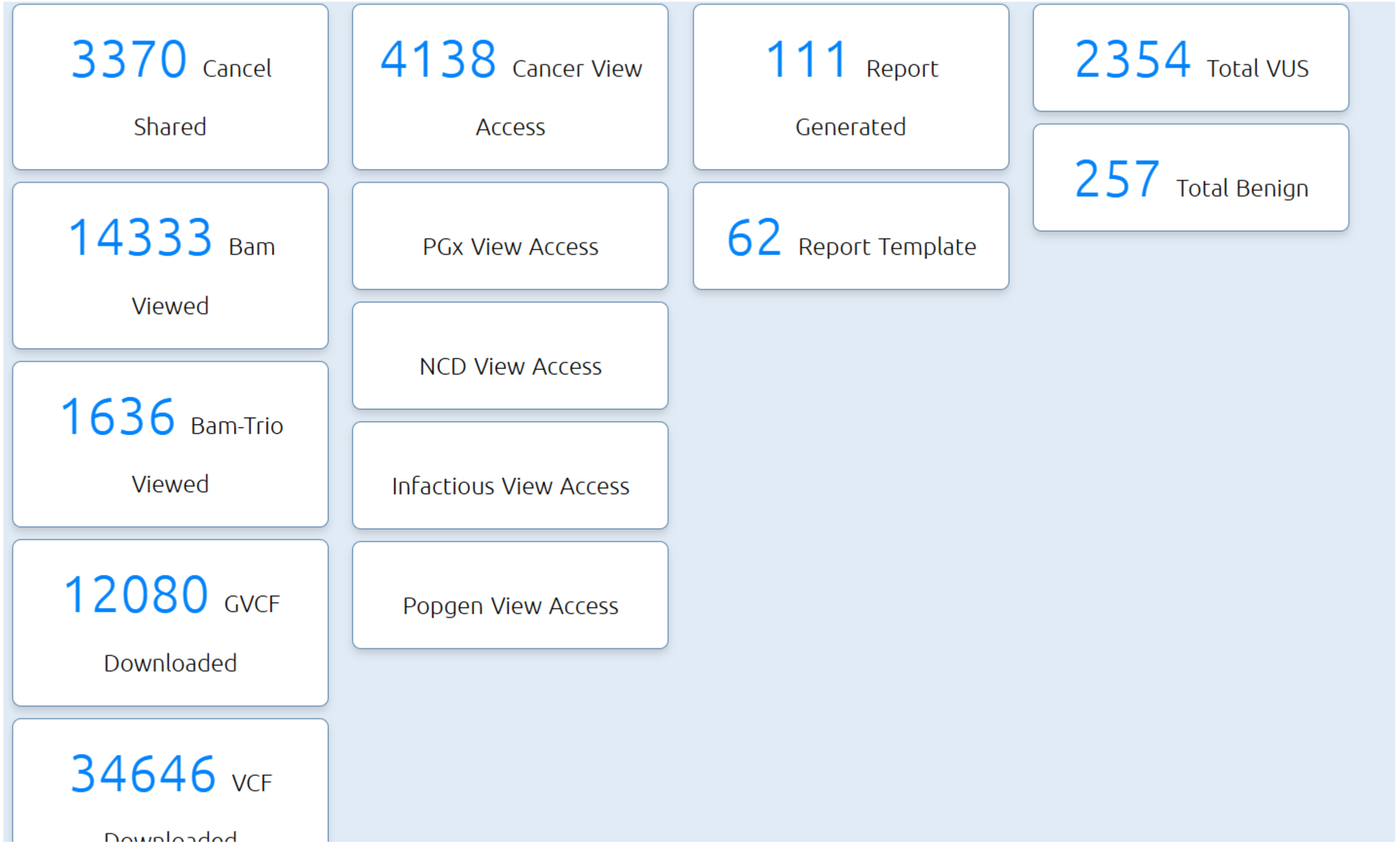
- **Cancers**
- **Rare diseases**
- **Non-communicable diseases (NCDs)**
- **Infectious diseases**
- **Pharmacogenomics**

V@PP (Variant Annotation and Prioritization Platform)
– Access to data and biological samples through a digital platform -



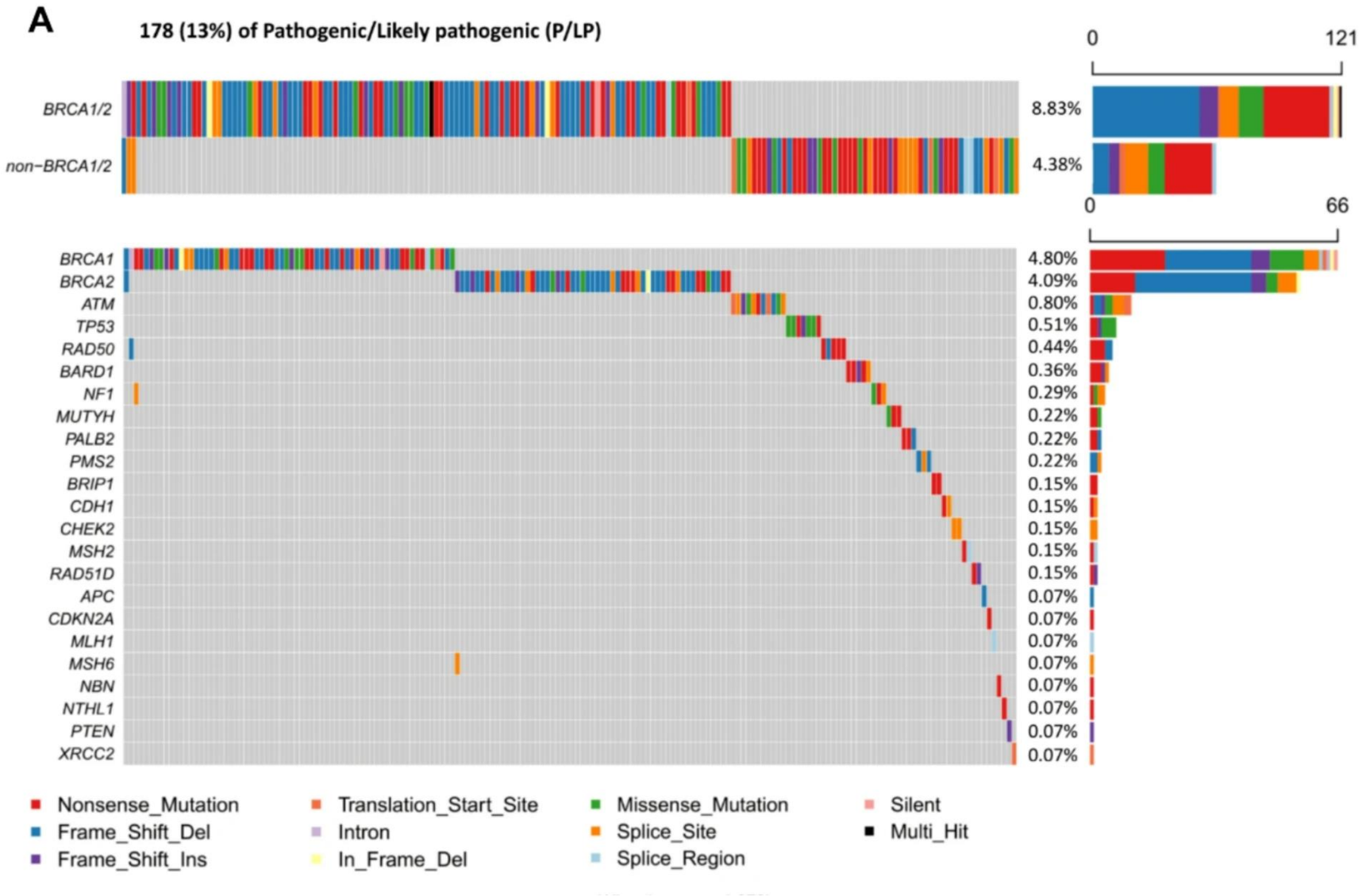
https://vapp.genomicsthailand.com

V@PP – Access to data and biological samples through a digital platform



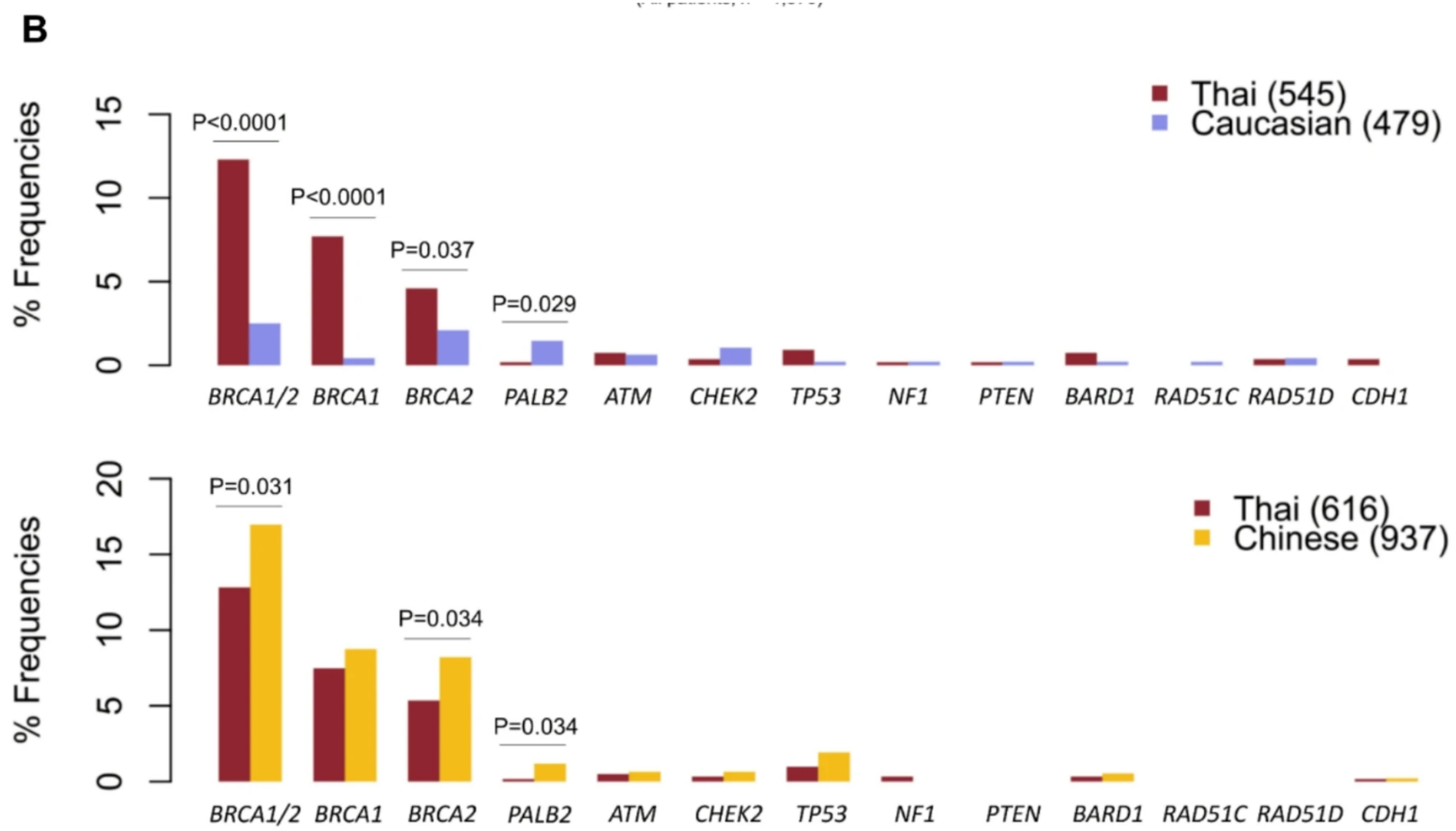
https://vapp.genomicsthailand.com

Germline mutation landscape in Thai breast cancer patients and comparison with other ethnicities.



Kamdee, K., Roothumnong, E., Thongnoppakhun, W. et al. Comprehensive germline and somatic profiling of high-risk Thai breast cancer via next-generation sequencing. Sci Rep 15, 11427 (2025)

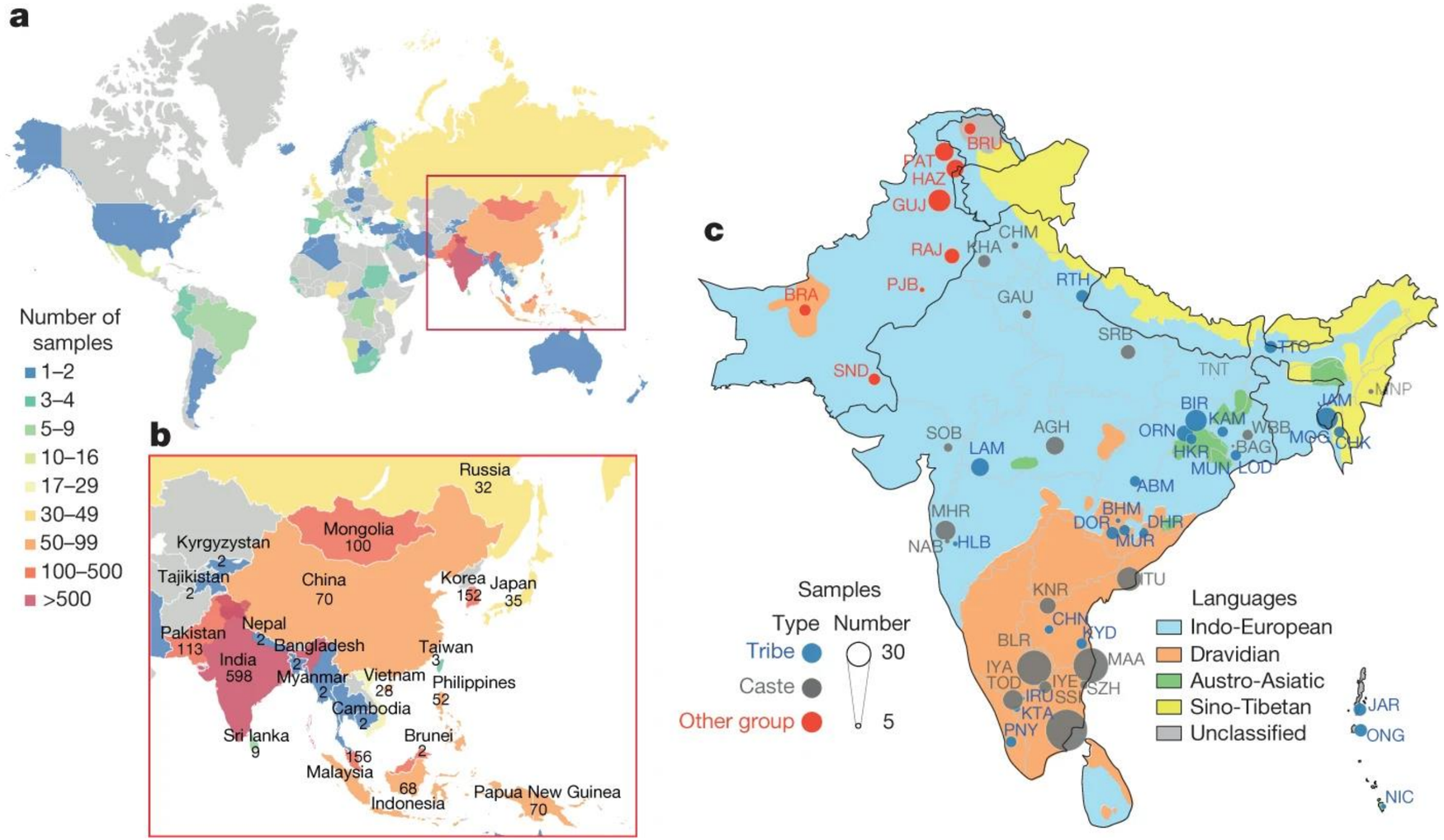
Comparison of the 10 most frequently mutated genes (high- and moderate-penetrance genes) in Thai patients with those in high-risk Caucasian and Chinese cohorts



Kamdee, K., Roothumnong, E., Thongnoppakhun, W. et al. Comprehensive germline and somatic profiling of high-risk Thai breast cancer via next-generation sequencing. Sci Rep 15, 11427 (2025)

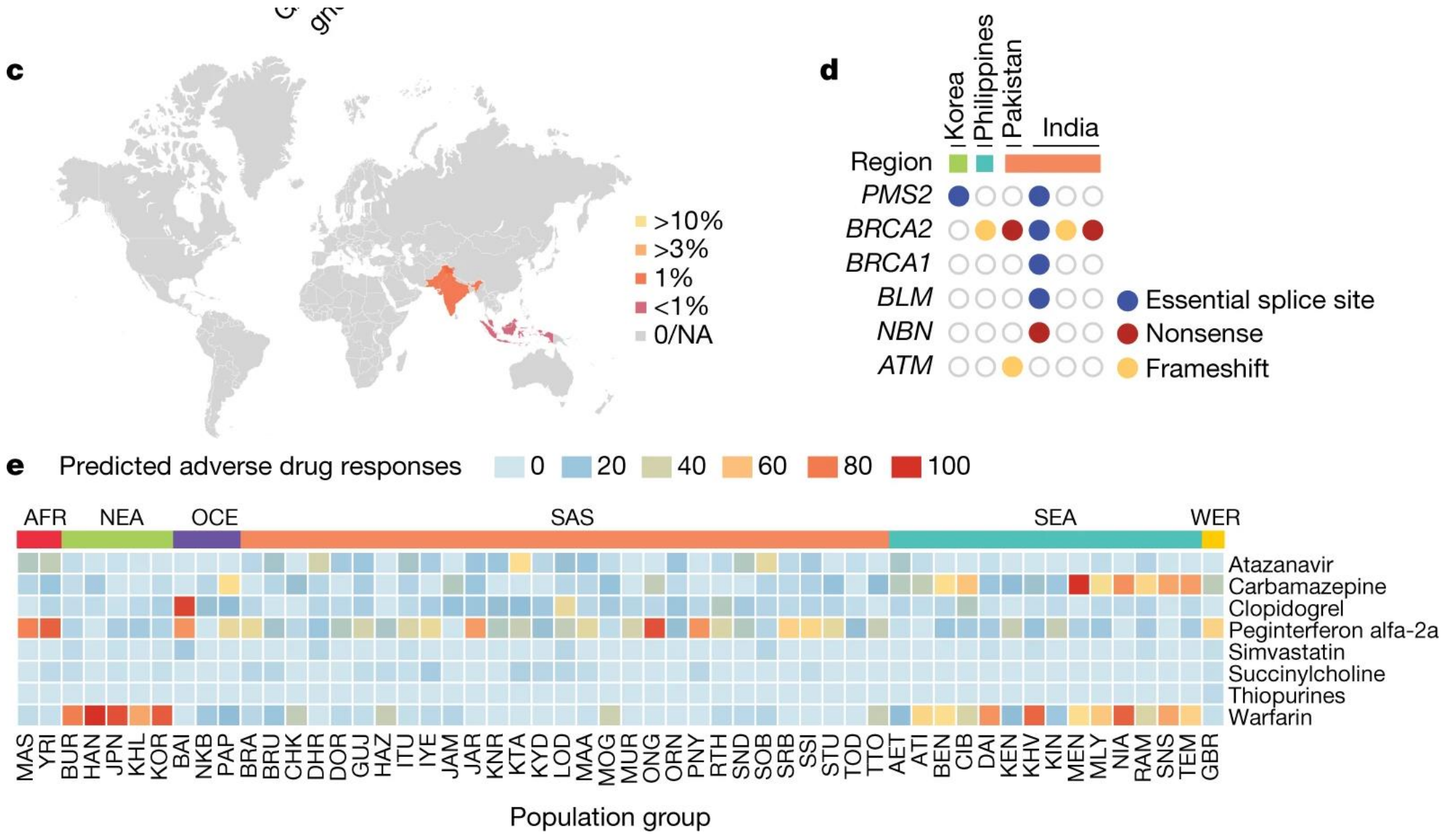
Pan-Asian consortium project: GenomeAsia 100K

WGS reference dataset from 1,739 individuals of 219 population groups and 64 countries across Asia



GenomeAsia100K Consortium. The GenomeAsia 100K Project enables genetic discoveries across Asia. Nature 576, 106–111 (2019)

Novel cancer-predisposing variants in Asian cohorts



- GAsP identified 13 unique variants affecting 6 cancer-risk genes in 17 samples: BRCA2 (n = 9), BRCA1 (n = 1), ATM (n = 2), BLM (n = 1), NBN (n = 2), PMS2 (n = 2)—including frameshift, stop-gained, and essential splice-site variants.

- **These findings highlight Asian-specific cancer predisposition alleles that would likely be missed or mis-prioritized using non-Asian reference panels.**

Predicted adverse drug response (ADR) risks vary widely across Asian populations for several drugs (e.g., carbamazepine, clopidogrel, peg-interferon, warfarin), with cohorts ranging from 0% to 100% predicted ADR risk depending on allele frequencies.

GenomeAsia100K Consortium. The GenomeAsia 100K Project enables genetic discoveries across Asia. Nature 576, 106–111 (2019)

Online Databases for Cancer Analysis

Online Databases Empowering Cancer Genomics Research

TCGA, **GDC**, **ICGC**, **cBioPortal**, **COSMIC**, and **OncoDB**,

- Open-access databases accelerate discovery in cancer genomics.
- Provide integrated multi-omics and clinical data for research and translation.
- Enable comparative analysis, validation, and biomarker discovery.

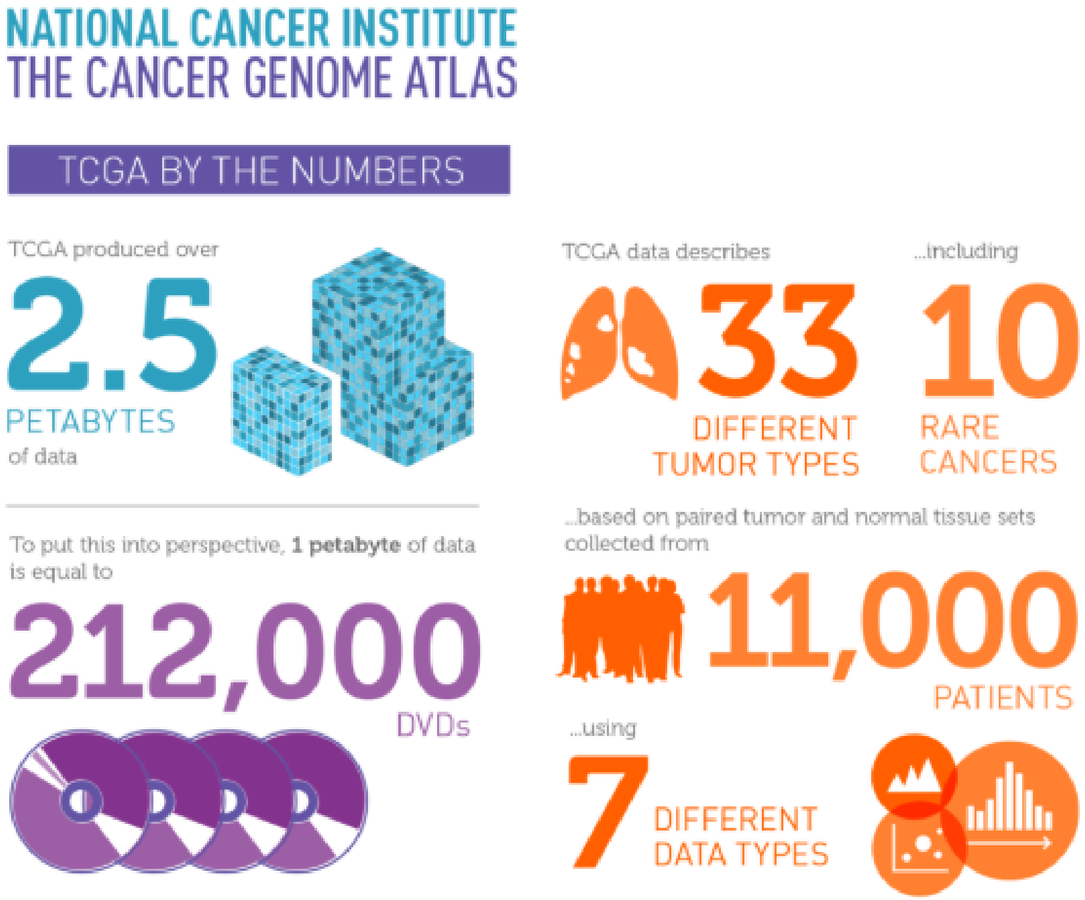
The Cancer Genome Atlas (TCGA)

- The Cancer Genome Atlas (TCGA), a landmark cancer genomics program, molecularly characterized over 20,000 primary cancer and matched normal samples spanning 33 cancer types.

- This joint effort between NCI and the National Human Genome Research Institute began in 2006

- TCGA generated over 2.5 petabytes of genomic, epigenomic, transcriptomic, and proteomic data.

- Publicly available for anyone in the research community to use.



Genomic Data Commons (GDC)

- *The GDC is a repository and computational platform for cancer researchers who need to understand cancer, its clinical progression, and response to therapy.*
- Link: https://portal.gdc.cancer.gov
- Unified data platform for TCGA, TARGET, and other projects.
- **Provides both raw and processed data with controlled access.**
- Integrated pipelines ensure consistency in variant calling and data harmonization.
- Offers tools for visualization, analysis, and data download.

Building Custom Cohorts

Creating custom groups, or cohorts, for analysis is simple with the GDC. Researchers can choose from various clinical, genomic, and other features to find the specific cases they want to study. Once a cohort is built, there are multiple options to save it throughout the data portal.

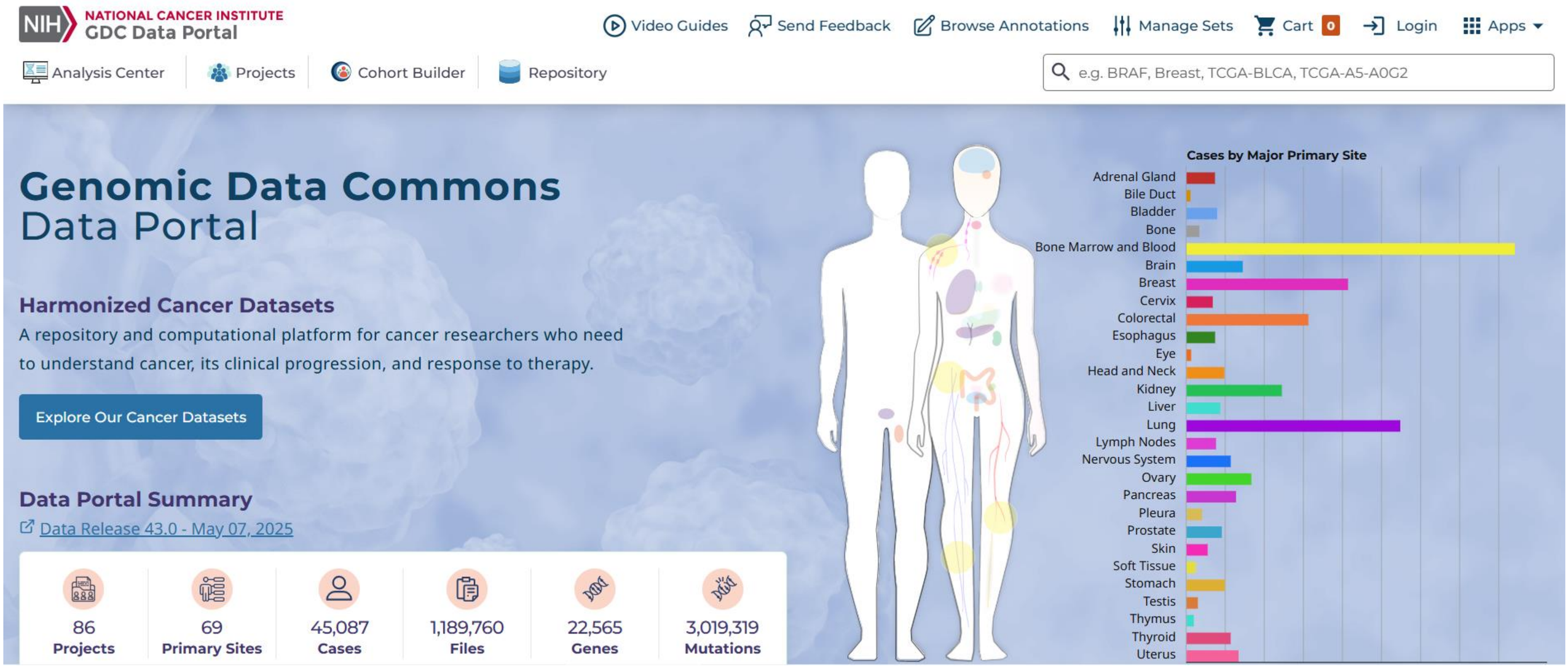
Interactive Analysis Tools

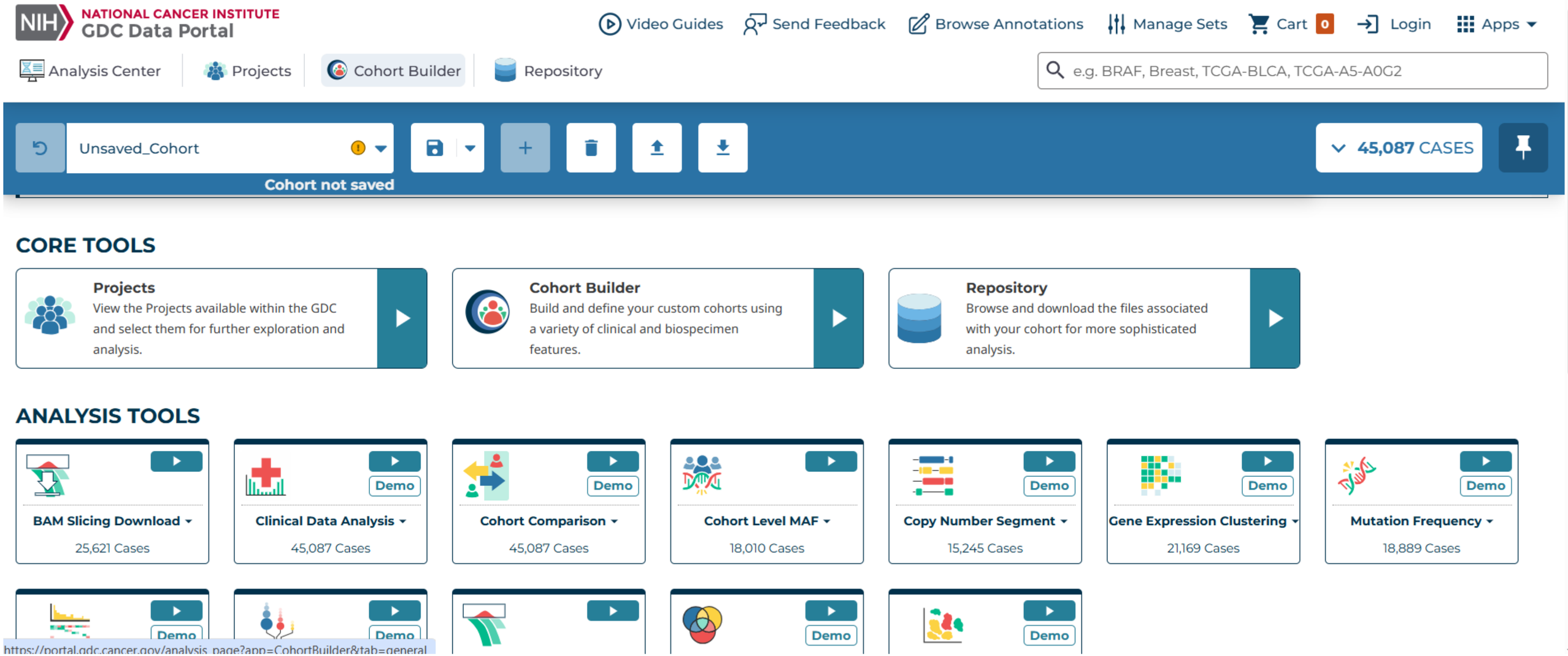
After forming a cohort, the GDC provides several interactive tools for deeper analysis. Some of these tools include:

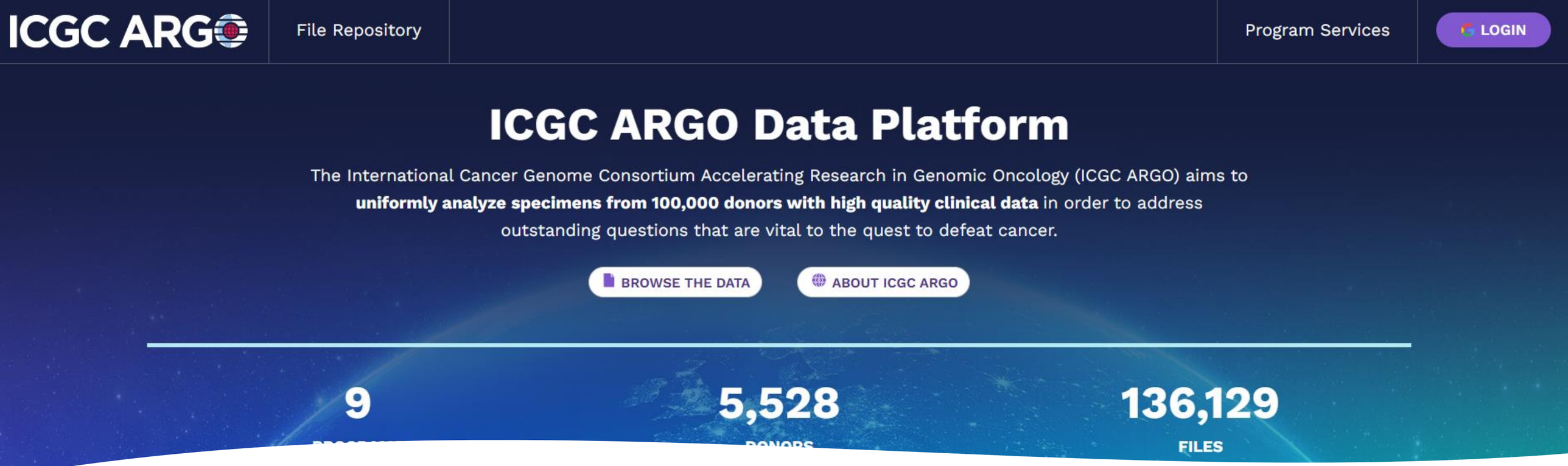
- **Clinical Data Analysis**: Researchers can create bar charts to explore clinical variables and compare survival rates based on these factors.
- **Gene Expression Clustering**: This tool allows users to visualize gene expression patterns through heatmaps and cluster diagrams.
- **Mutation Frequency**: Users can identify the most frequently mutated genes connected to specific somatic mutations.
- **OncoMatrix**: Researchers can visualize combinations of mutations to discover common co-occurrences or mutual exclusivity. This feature includes stylish updates and many customization options.
- **ProteinPaint**: This tool visually displays where mutations occur on proteins, how often they happen, and their potential effects.

https://gdc.cancer.gov/about-gdc/gdc-overview

https://portal.gdc.cancer.gov/

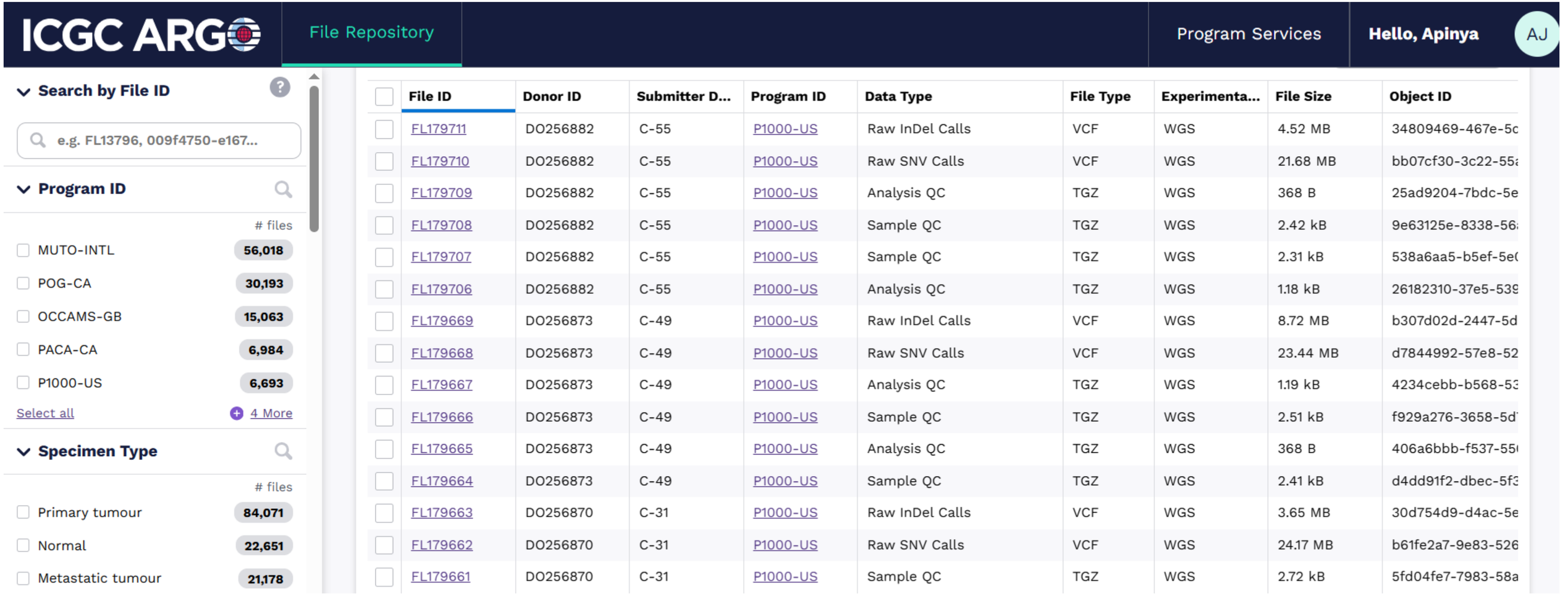



NIH
NATIONAL CANCER INSTITUTE
GDC Data Portal
Video Guides
Send Feedback
Browse Annotations
Manage Sets
Cart 0
Login
Apps
Analysis Center
Projects
Cohort Builder
Repository
e.g. BRAF, Breast, TCGA-BLCA, TCGA-A5-A0G2
Unsaved_Cohort
Cohort not saved
45,087 CASES
CORE TOOLS
Projects
View the Projects available within the GDC and select them for further exploration and analysis.
Cohort Builder
Build and define your custom cohorts using a variety of clinical and biospecimen features.
Repository
Browse and download the files associated with your cohort for more sophisticated analysis.
ANALYSIS TOOLS
BAM Slicing Download
25,621 Cases
Clinical Data Analysis
45,087 Cases
Cohort Comparison
45,087 Cases
Cohort Level MAF
18,010 Cases
Copy Number Segment
15,245 Cases
Gene Expression Clustering
21,169 Cases
Mutation Frequency
18,889 Cases
Demo
https://portal.gdc.cancer.gov/analysis_page?app=CohortBuilder&tab=general



International Cancer Genome Consortium (ICGC)

- Link: https://icgc.org/
- Global collaboration involving 22 countries and 90+ cancer projects.
- Complements TCGA by expanding population diversity — including Asian cohorts.
- Provides harmonized WGS/WES and transcriptomic data with metadata.
- ICGC-ARGO continues the mission for clinically-linked cancer data.

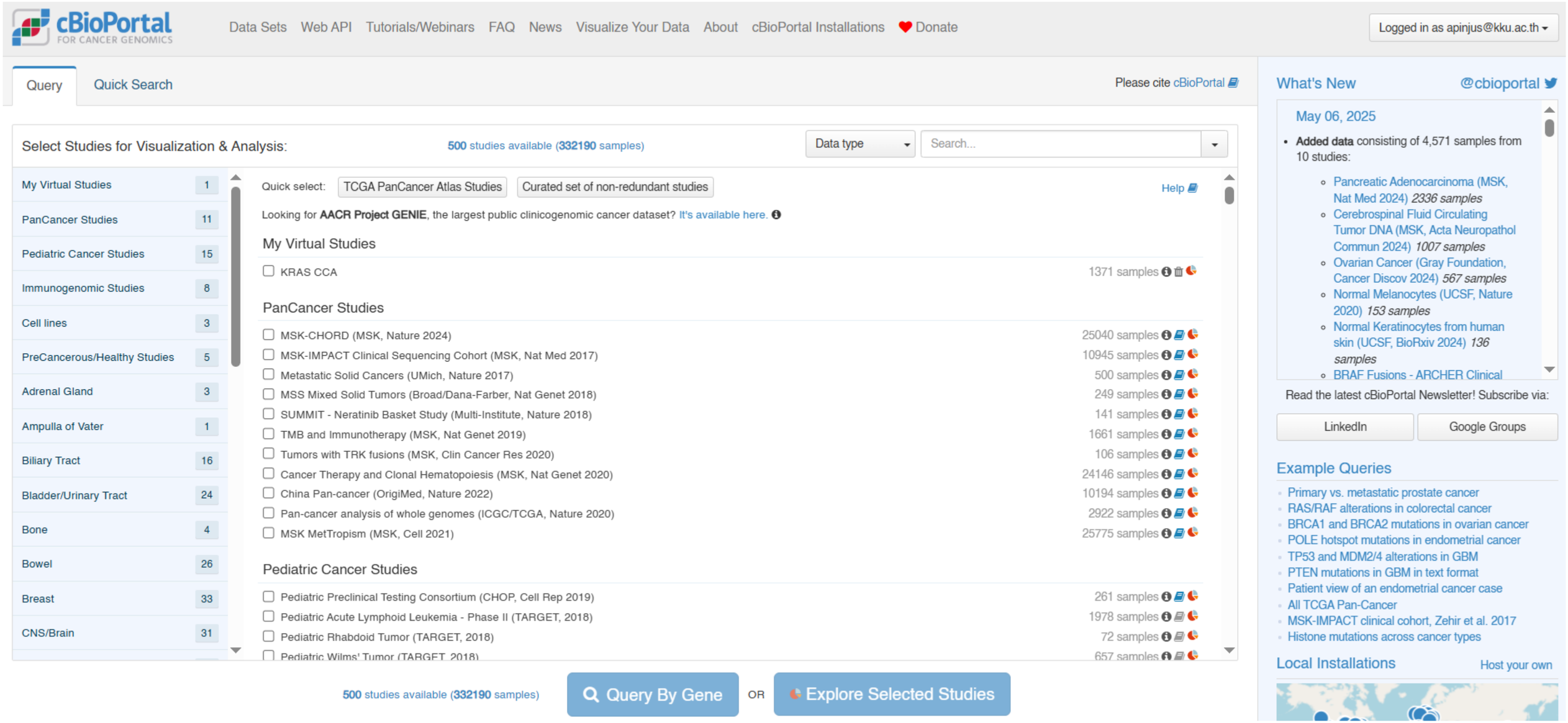


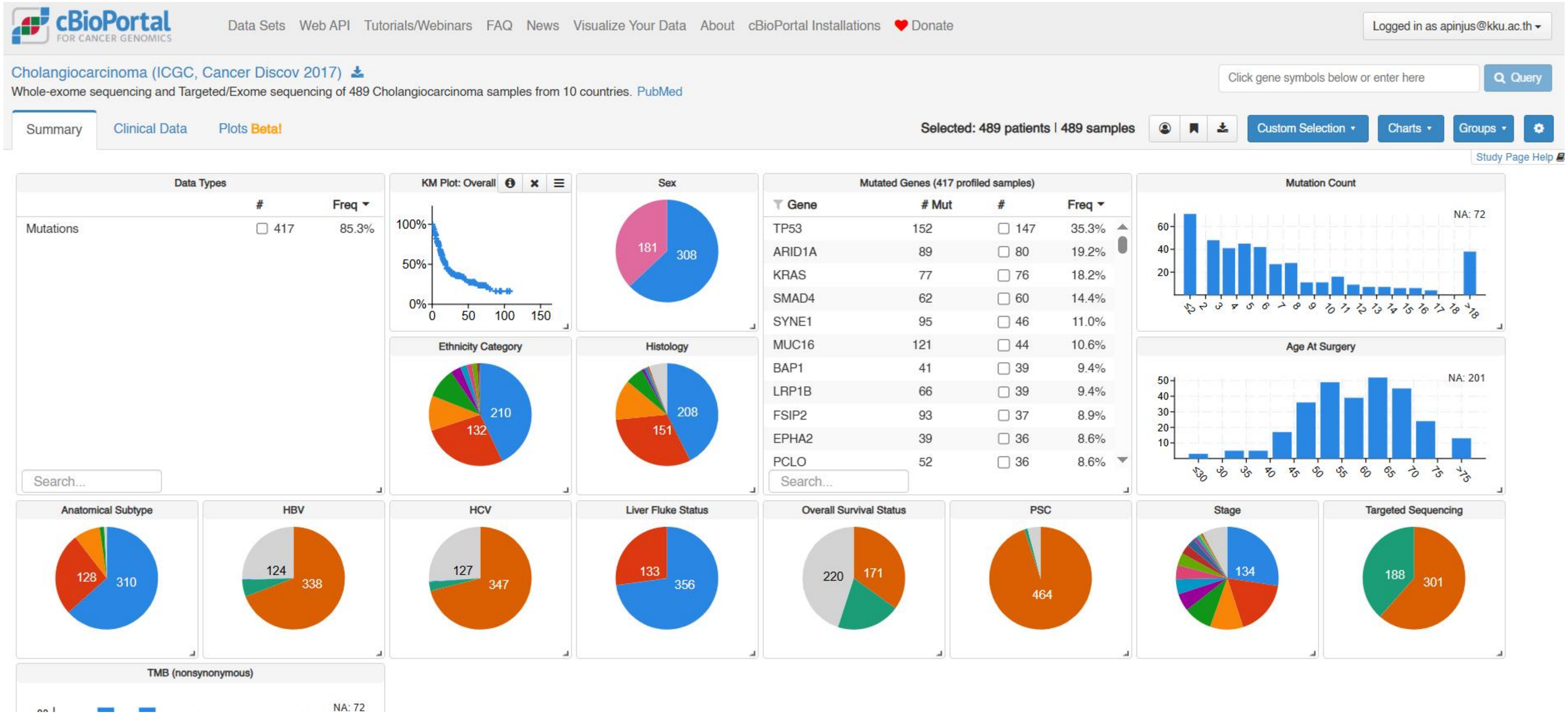
https://platform.icgc-argo.org/repository?filters=%7B%22content%22%3A%5B%5D%2C%22op%22%3A%22and%22%7D

cBioPortal for Cancer Genomics

- Link: https://www.cbioportal.org
- User-friendly interface for exploring multidimensional cancer genomics data.
- Visualization tools: oncoplots, co-mutation plots, survival curves, correlations.
- Integrates datasets from TCGA, ICGC, and institutional studies.
- Ideal for researchers without coding expertise.
- **Data Exploration**
  - Integrates **multi-omics cancer data**: somatic mutations, CNVs, gene expression, methylation, and clinical attributes.
  - Supports > 50,000 tumor samples from **TCGA**, **ICGC**, and institutional studies.
  - Query by **gene**, **patient**, or **cancer type**.
- **Visualization Tools**
  - **Oncoprint:** overview of mutations, CNVs, and expression in a cohort.
  - **Mutations tab:** displays variant types, frequencies, and protein-domain mapping.
  - **Plots:** gene-to-gene correlation and scatter plots for quantitative relationships.
  - **Survival curves (Kaplan–Meier):** survival analysis by mutation or expression status.

https://www.cbioportal.org/





https://www.cbioportal.org/

Visualize Your Data

1. Download and install a local version of cBioPortal

- The source code of cBioPortal is available on GitHub under the terms of Affero GPL V3.
- Please note that, installing a local version requires system administration skills, for example, installing and configuring Tomcat and MySQL. With limited resources, we cannot provide technical support on system administration.

2. We host data for you (academic use)

- Public data will be available to everyone. Suggestions on data sets are welcome.
- Please contact us for details.

3. Commercial support

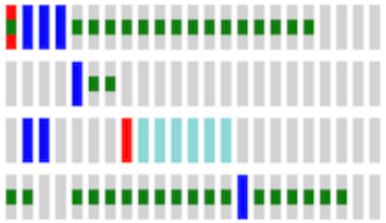
- The Hyve is an open source software company that provides commercial support for cBioPortal. They can help with deployment, data loading, development, consulting and training. Please contact The Hyve for details.
- SE4BIO provides software engineering and consultancy services, covering software architecture, business analysis, and the design of end-to-end solutions. SE4BIO also supports organizations in aligning cBioPortal with their broader data infrastructure, improving interoperability, streamlining workflows, and implementing custom features to meet specific research and clinical needs. Please contact SE4BIO for details.

The following tools are for visualization and analysis of custom datasets

When using these tools in your publication, **please cite** cBioPortal.

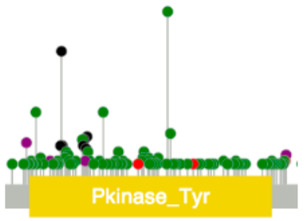
OncoPrinter

Generates oncoprints from your own data. Try it!

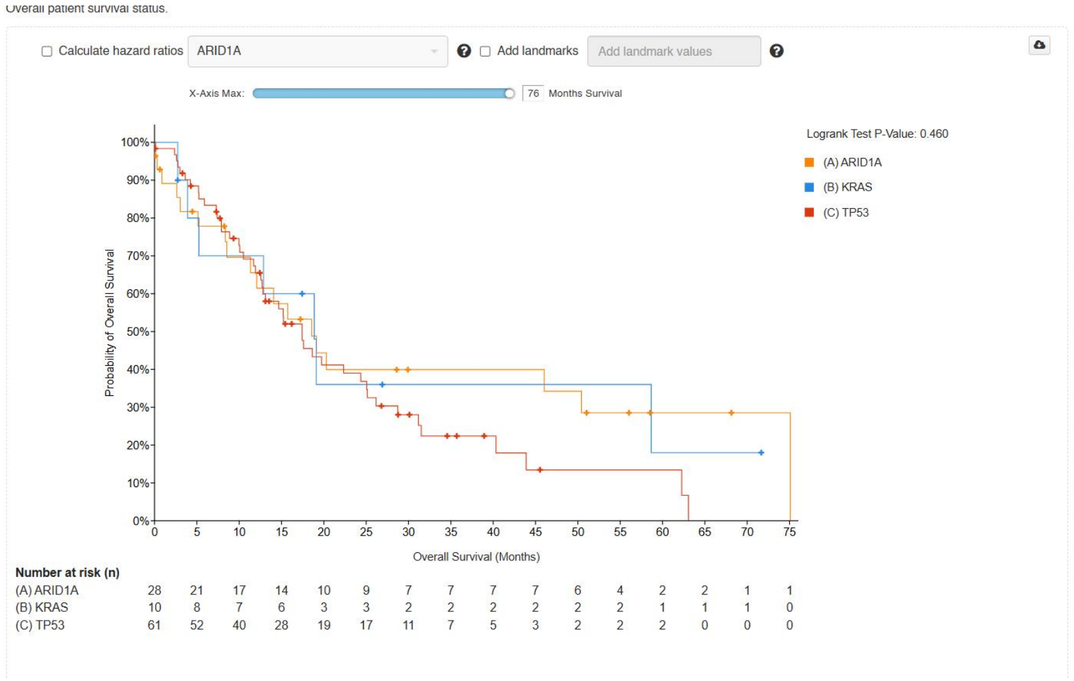


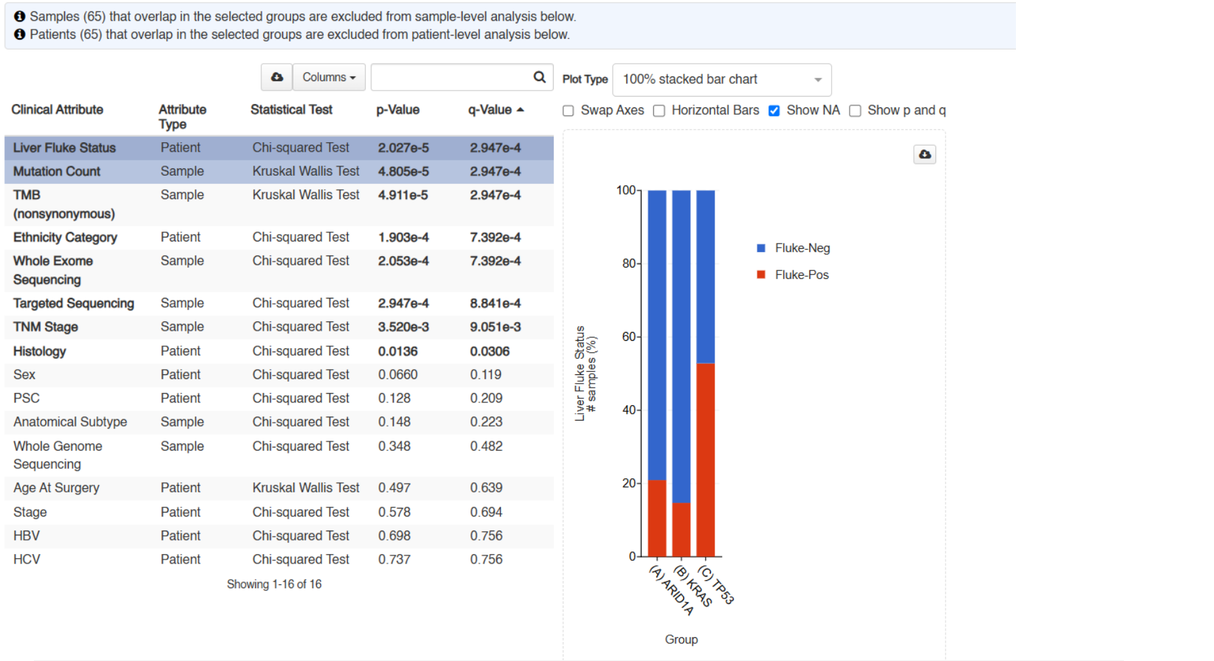
MutationMapper

Maps mutations on a linear protein and its domains (lollipop plots). Try it!



https://www.cbioportal.org/



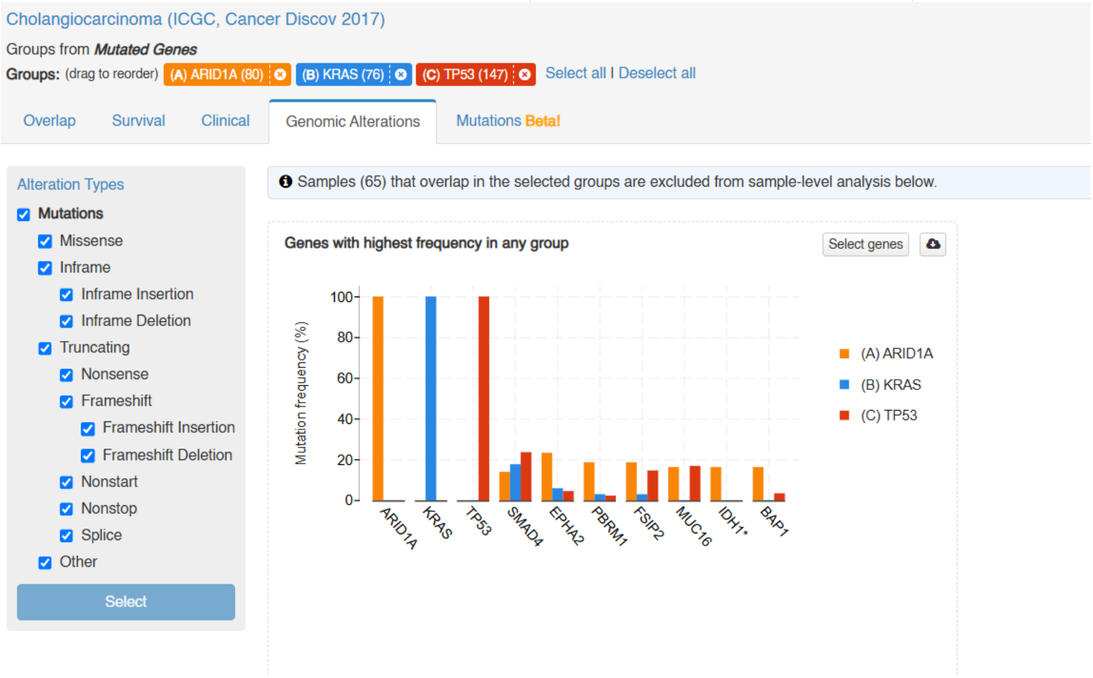


Genomic Correlation & Enrichment

- Analyze **co-occurrence or mutual exclusivity** between genomic alterations.
- Perform **correlation** between mRNA expression, copy number, and methylation.
- Enrichment analysis for altered pathways or gene sets.

Clinical Integration

- Link molecular profiles with **clinical parameters** (e.g., stage, subtype, therapy).
- Identify biomarkers associated with **prognosis or treatment response**.



https://www.cbioportal.org/comparison/alterations?comparisonId=68fc72937783fa02c7f3136f

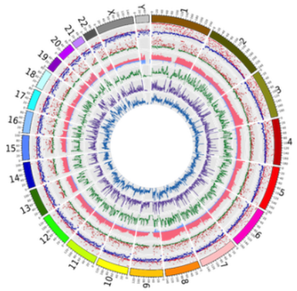
COSMIC :Catalogue of Somatic Mutations in Cancer

- Link: https://cancer.sanger.ac.uk/cosmic
- World's largest expert-curated database of somatic mutations.
- Includes driver mutations, drug-resistance variants, and mutational signatures.
- Integrates data from literature, sequencing studies, and clinical trials.
- Crucial for interpreting mutation pathogenicity and drug response.

COSMIC- https://cancer.sanger.ac.uk - holds details on millions of mutations across thousands of cancer types. It is constantly growing in both content and scope.

About

Hand-curation of key cancer genes (selected from the Cancer Gene Census) provide in-depth detail on mutation distributions and effects, whilst semi-automated curation of cancer genomes provides broad somatic annotations toward target discovery and identification of patterns and signatures. This information is fully available via website or download, updated every three months.



Curation

Cancer Gene Census: This is a list of hundreds of genes with substantial published evidence in oncology. Necessarily conservative, this is a very high-confidence list based on good-quality publications. Selection of high-impact genes from this list for curation drives COSMIC.

https://cancer.sanger.ac.uk/cosmic/

COSMIC v102, released 21-MAY-25

COSMIC, the Catalogue Of Somatic Mutations In Cancer, is the world's largest and most comprehensive resource for exploring the impact of somatic mutations in human cancer.

Start using COSMIC by searching for a gene, cancer type, mutation, etc. below.

eg Braf, COLO-829, Carcinoma, V600E, BRCA-UK, Campbell SEARCH

Projects

COSMIC is divided into several distinct projects, each presenting a separate dataset or view of our data:

COSMIC
The core of COSMIC, an expert-curated database of somatic mutations

Cell Lines Project
Mutation profiles of over 1,000 cell lines used in cancer research

COSMIC-3D
An interactive view of cancer mutations in the context of 3D structures

Cancer Gene Census
A catalogue of genes with mutations that are causally implicated in cancer

Release Notes

v102 - 21st May 2025

Summary

Cancer remains one of the leading causes of death among children and adolescents worldwide. I receive only a fraction of global cancer research funding, approximately 4%. It is with disparit have chosen to focus curation on pediatric tumour data. As part of this release, we have curated

- 29 whole exome/ genome studies, 21 of which cover pediatric cancers. These studies cover a to new mutations
- The full somatic mutation landscape for 14 cancer genes which have recently been added to the
- 75 new anatomical site-histology pairs
- Further updates to all tumour types, and all classic cancer gene mutation profiles

Key Updates

New genes curated: RAD50, RAD51C, PMS1, PMS2, PLEC, PIK3R2, PIK3R3, RXRA, MYB (No cu NR2F2,SOS1, TNRC18, YAP1 and GTF2I)

Histologies: 75 new site-histology pairs added including a new tumour type entry added to Na

Tools

Cookie settings

OncoDB-An Integrated Multi-Omic Cancer Database

- OncoDB: https://oncodb.org
- Comprehensive database for cancer genomics and systems biology research.
- Integrates **multi-omic data** for ~10,000 patients across **33 cancer types**.
- Sources include **TCGA**, **GTEx**, and **CPTAC** datasets.
- **Data Types**
  - Gene expression (RNA-seq)
  - DNA methylation
  - Somatic mutations
  - Proteomic profiles
  - Chromatin accessibility (ATAC-seq)
- **Core Analytical Functions**
  - Compare **tumor vs. normal** for gene expression, methylation, and protein levels.
  - Identify **differentially expressed genes (DEGs)** and **proteins**.
  - Perform **survival analysis** based on gene expression or methylation linked to clinical data.
  - Explore **gene-to-gene correlations** and **mutation-associated variations**.
  - Visualize **oncogene mutation profiles** and their clinical significance

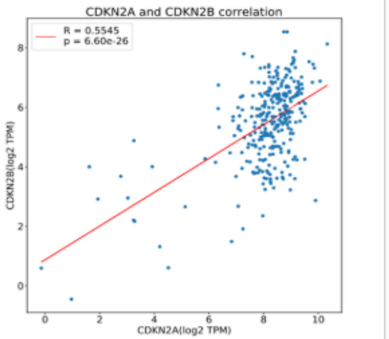
https://oncodb.org/

OncoDB

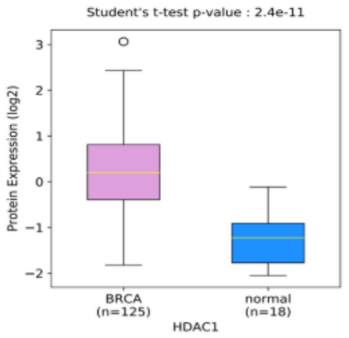
- RNA Expression
- Proteomics
- DNA Methylation
- Somatic Mutation
- Chromatin Accessibility
- Multi-Omics Analysis
- Oncovirus
- Clinical Analysis
- Data Download
- Tutorial
- Contact

Welcome to OncoDB!

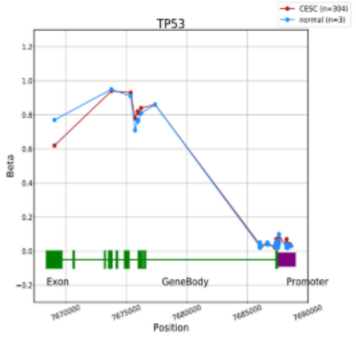
OncoDB2.0 is a robust database designed for cancer researchers, offering integrated multi-omic data for approximately 10,000 patients across 33 cancer types. It encompasses gene expression, DNA methylation, somatic mutations, proteomic profiles, and chromatin accessibility, drawing from TCGA, GTEx, and CPTAC projects. Users can compare gene expression, DNA methylation, and protein levels between tumor and normal tissues, identifying differentially expressed genes and proteins, and examining gene-to-gene correlations. The platform also provides oncogene mutation profiles and allows for survival analysis based on gene expression and methylation, linked to clinical parameters. Furthermore, OncoDB2.0 facilitates the exploration of multi-omic correlations, such as gene expression with DNA methylation, and their variations with mutation status. Beyond cancer, the database extends its analytical capabilities to include six major oncoviruses, offering insights into their impact on gene expression, methylation, and patient survival.



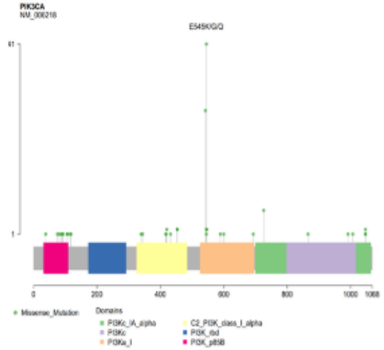
RNA Expression



Proteomics

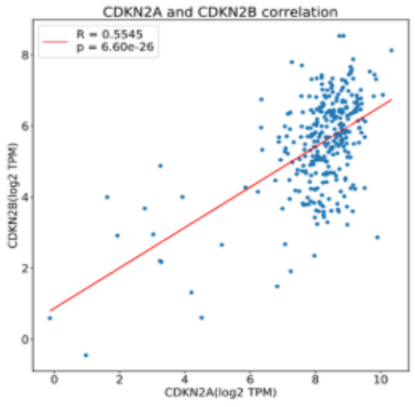


DNA Methylation

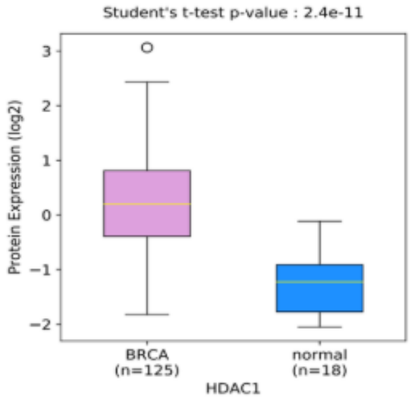


Somatic Mutation

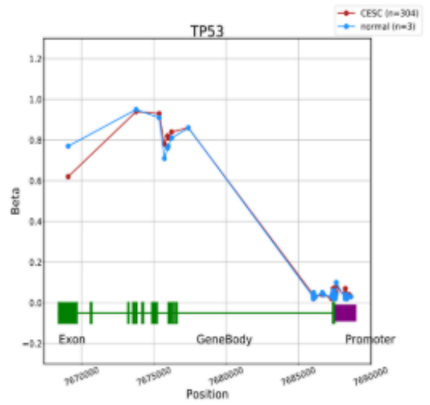
https://oncodb.org/



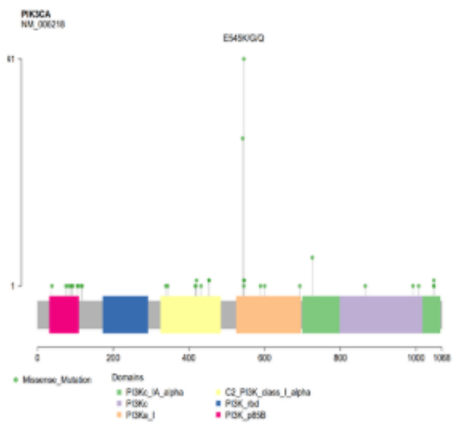
RNA Expression



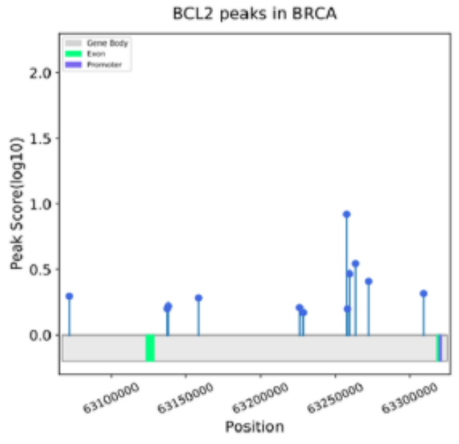
Proteomics



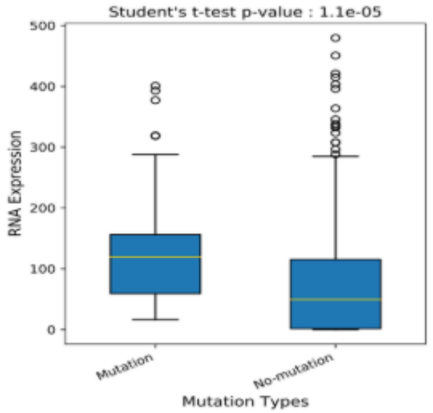
DNA Methylation



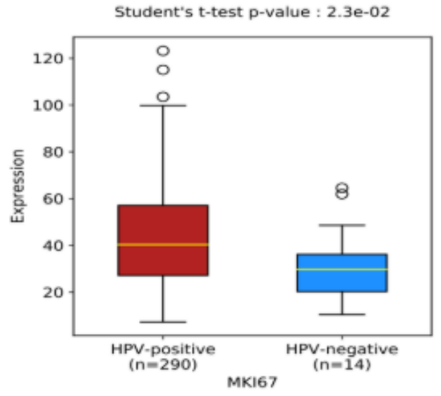
Somatic Mutation



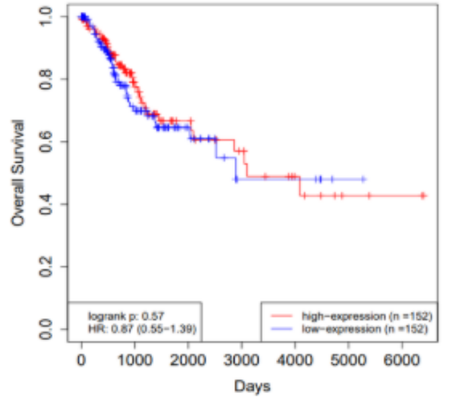
Chromatin Accessibility



Multi-Omics Analysis



Oncovirus Analysis



Clinical Analysis

Connecting Data to Discovery

- Together, these databases create a continuum from data generation ⟶ curation ⟶ clinical application.
- Support reproducible, large-scale cancer research globally.
- Essential tools for precision oncology and bioinformatics training.

THANK YOU