NCI Genomic Enclave: Lowering the Access Barrier with GenAI Chatbot

Zhining Wang, Ph.D.

October 17, 2024

NIH NATIONAL CANCER INSTITUTE

Agenda

- Why Build an NCI Genomic Enclave?
  - Tempus Data Donation
- How the NCI Genomic Enclave Works
- Development Status
- Next Steps

Generous Data Donation from Tempus:
3300 cases of genomic profiling data

xT panel (DNA-Seq)

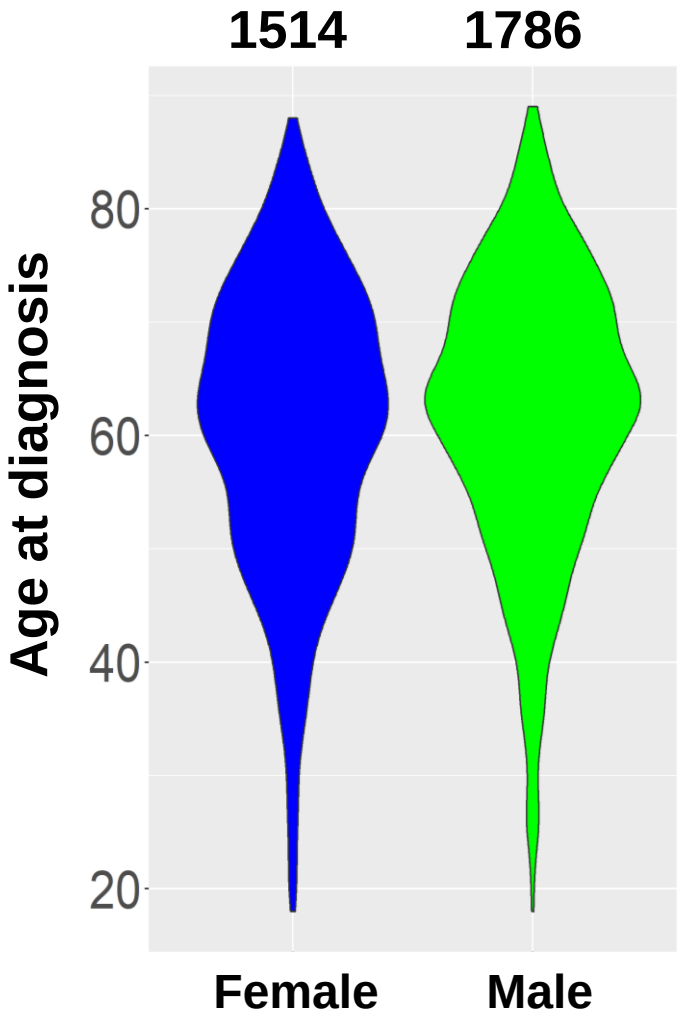
- 648 genes
- Clinically relevant DNA alterations
- Immunotherapy biomarkers such as MSI and TMB
- HLA Class I genotyping

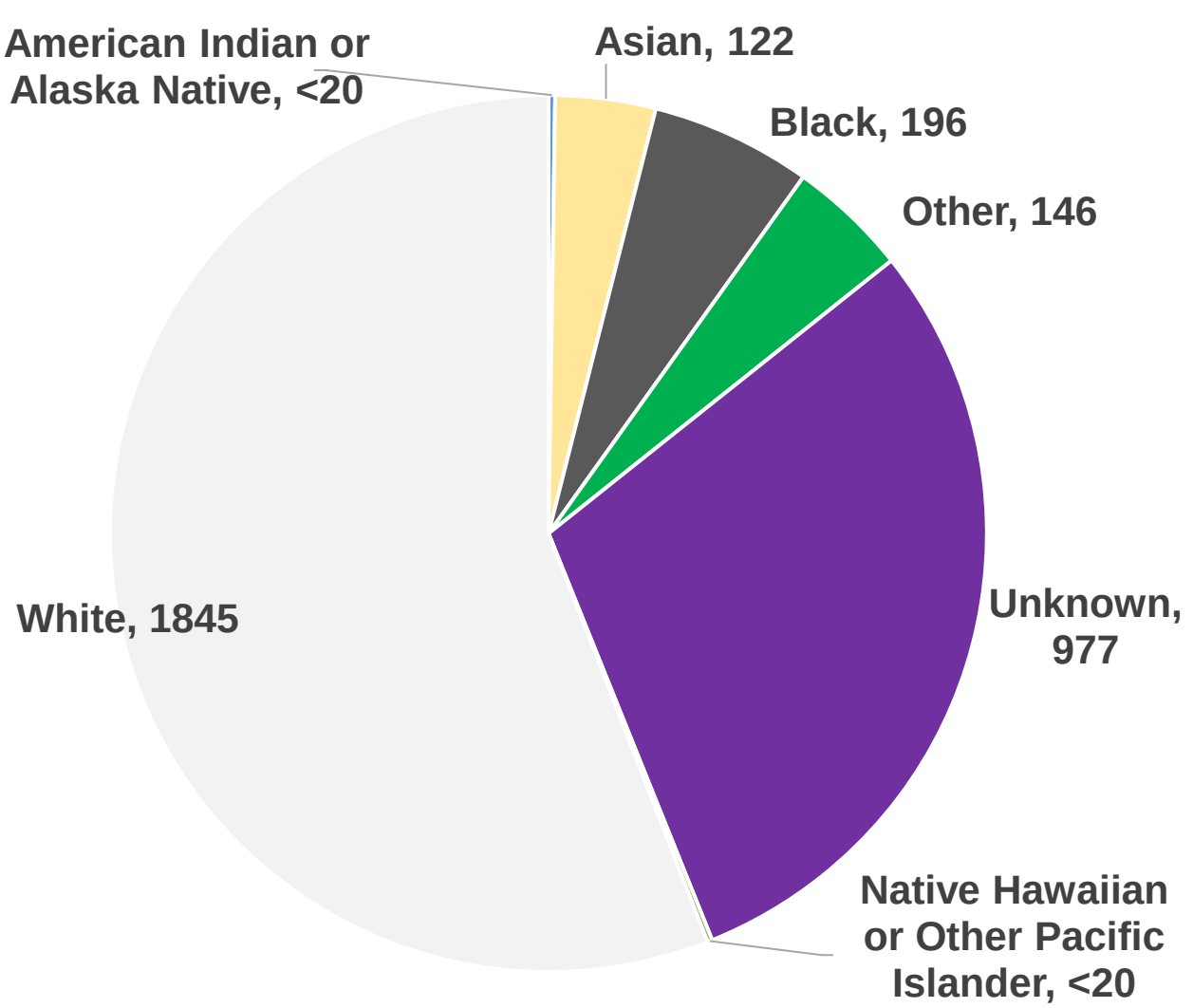
xR panel (RNA-Seq)

- A whole transcriptome RNA-Seq
- Clinically relevant fusions
- Altered splicing events for MET exon 14 and EGFRvIII

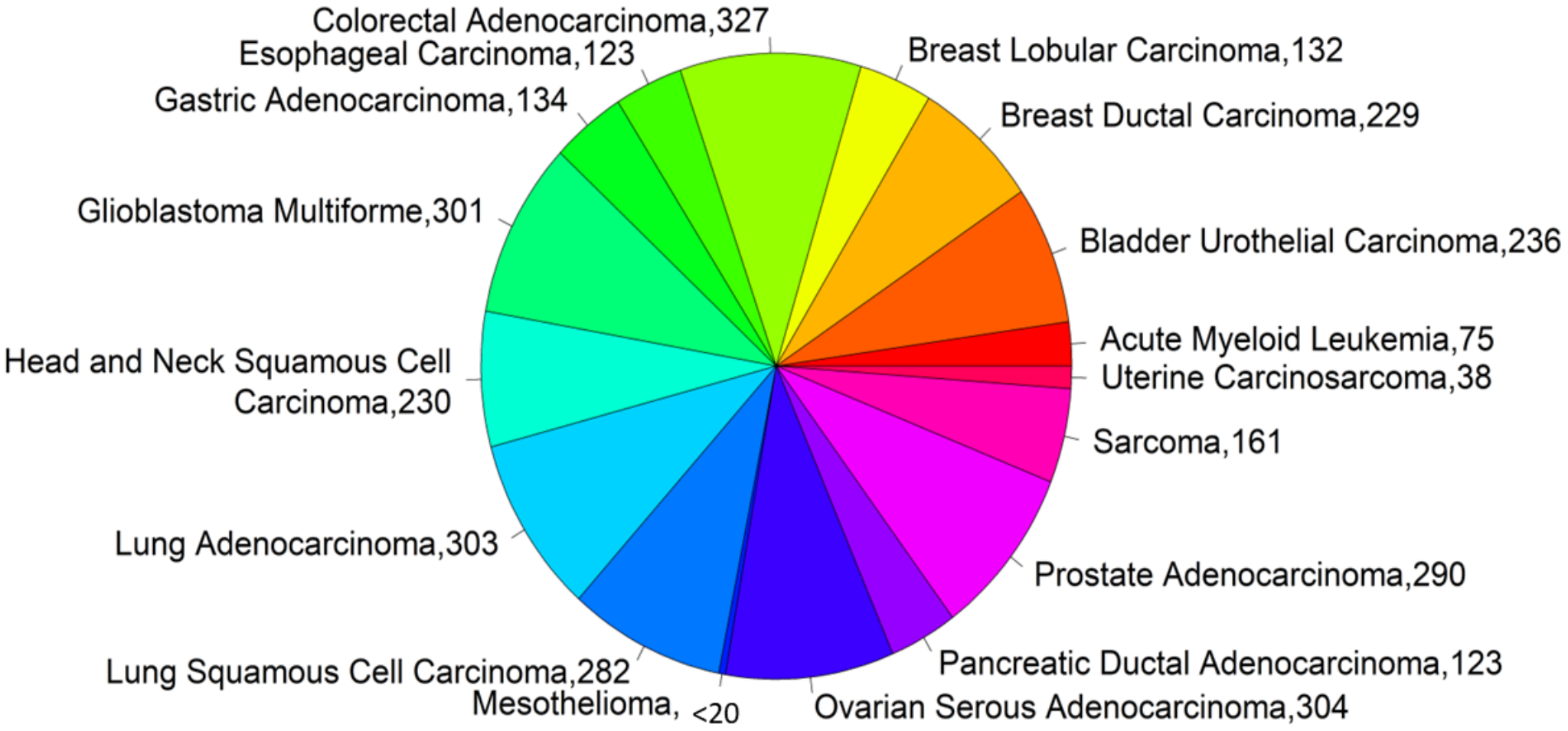
https://www.tempus.com

Tempus Data Represent a Diverse Population

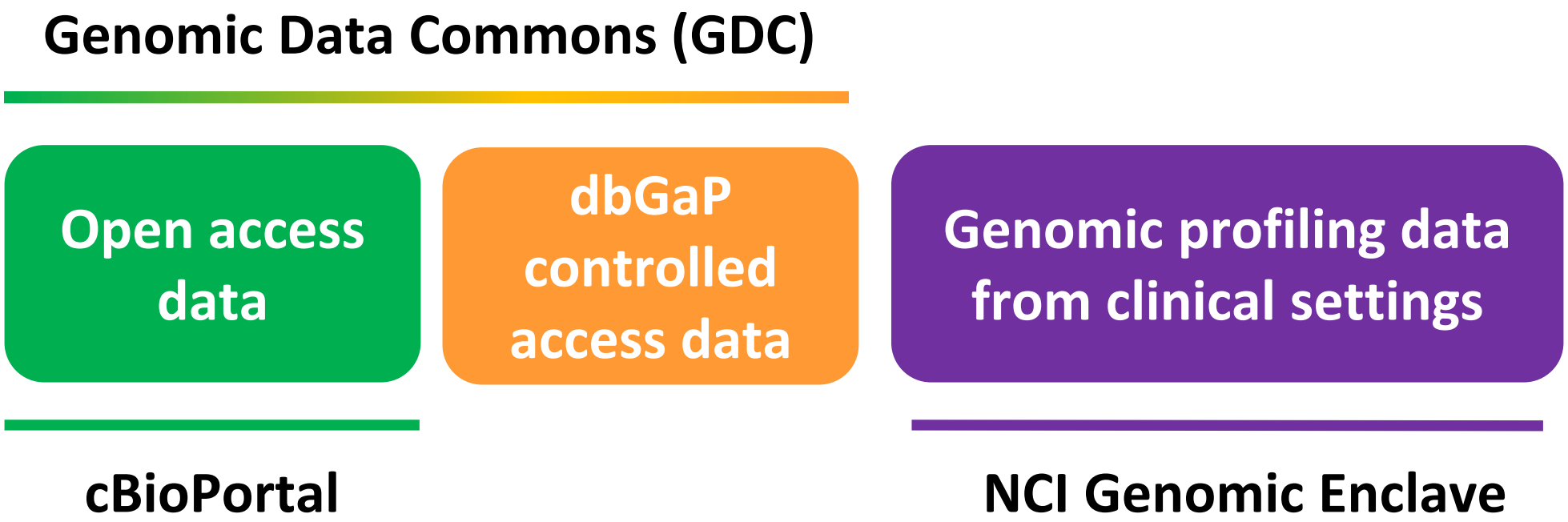




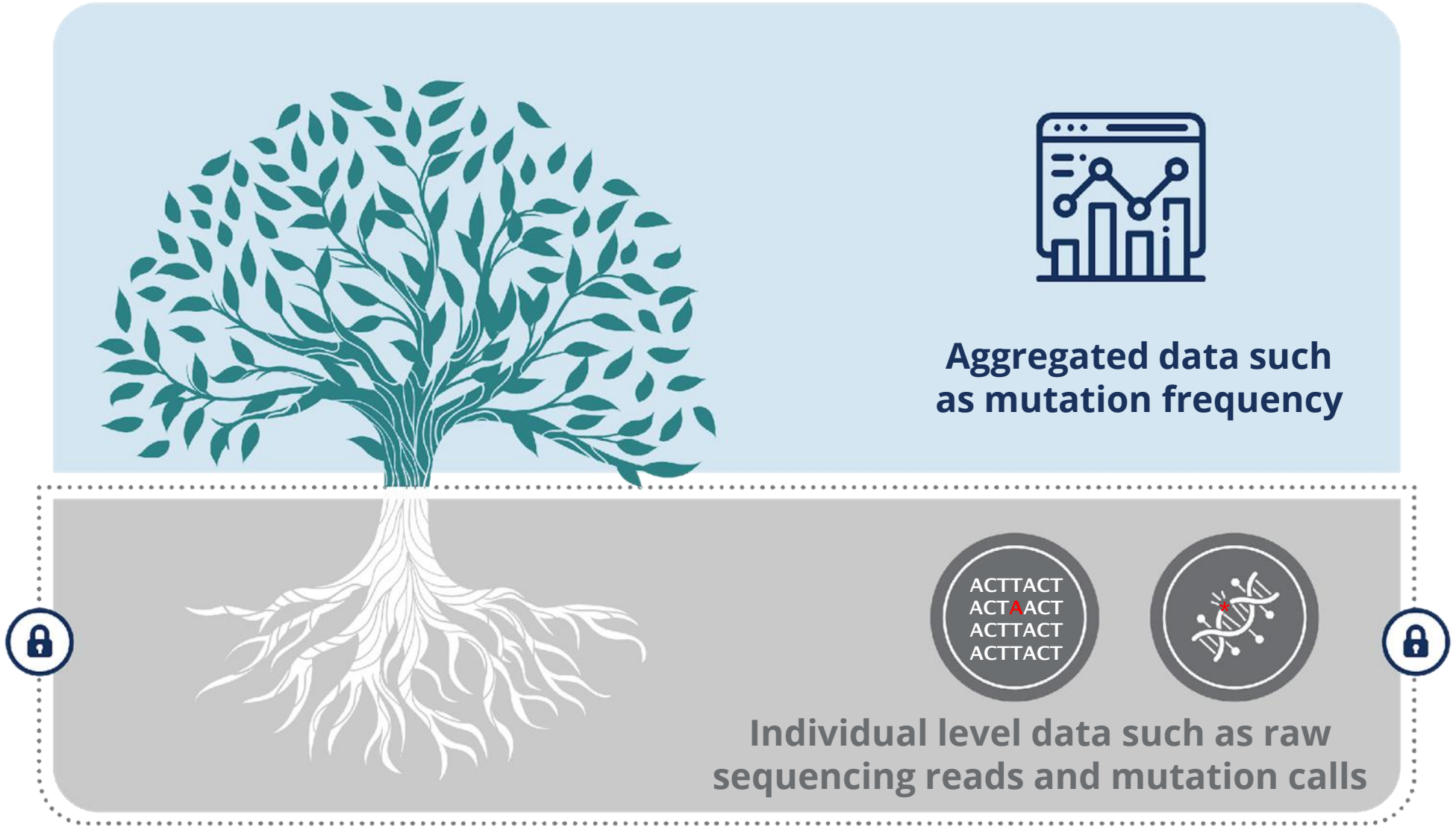
Tempus Data Represent 17 Cancer Types



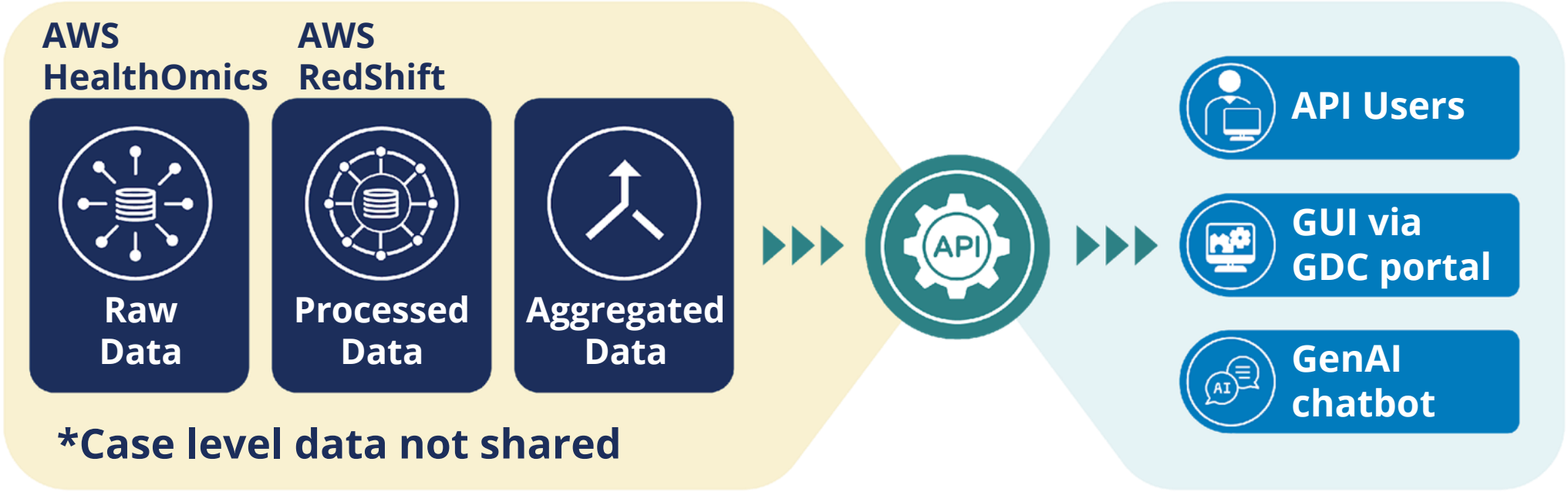
Tiers of Genomic Data Sharing



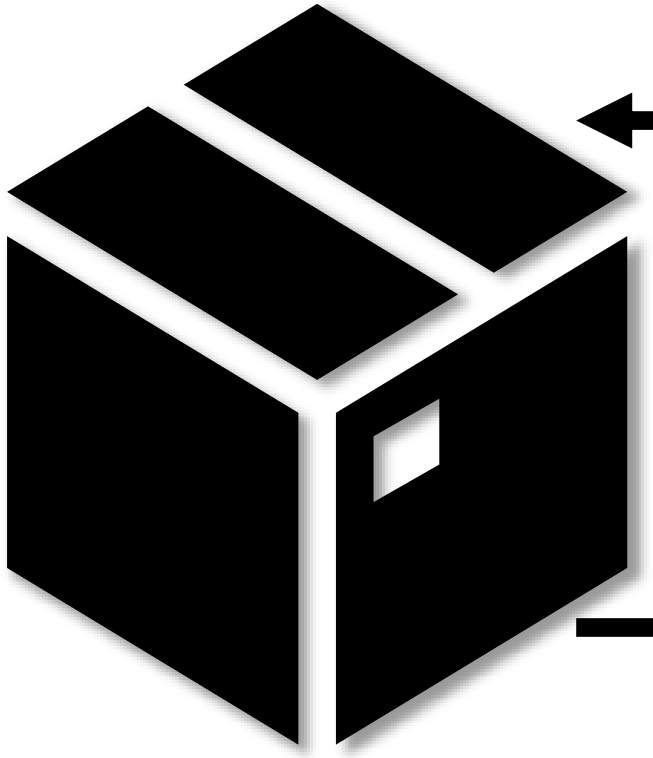
NCI Genomic Enclave Protects Privacy Via Data Aggregation



How Does the NCI Genomic Enclave Work?



GenAI Chatbot Concept



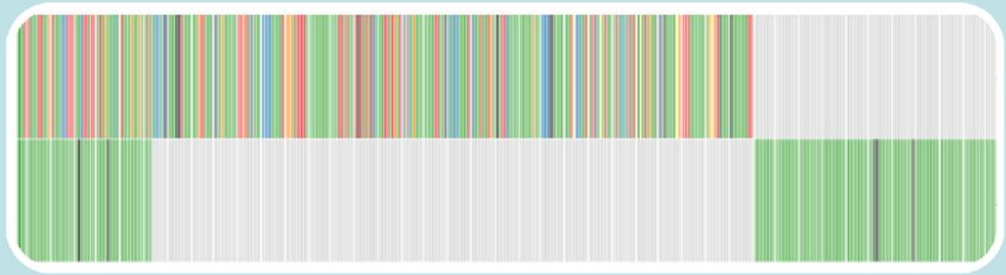
NCI's Genomic Enclave
A Black Box To the User

Users ask questions:

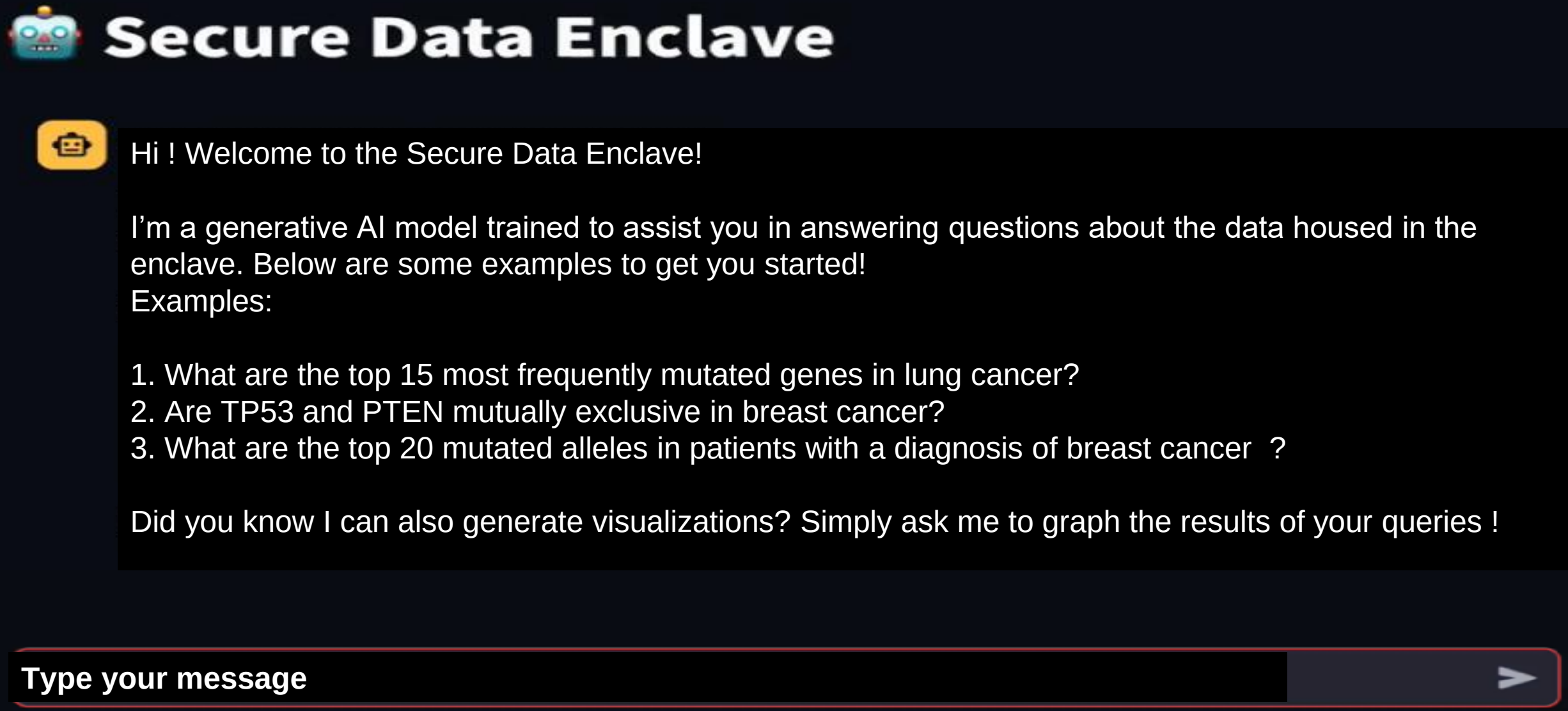
Are **TP53** and **KRAS** mutually exclusively mutated in **lung cancer** patients who are **female** and younger than **50 years old**?

The enclave returns publishable, aggregated results:

TP53 and KRAS are mutually exclusively mutated in your cohort with p value: 3.068794e-05



Demo 1: Most frequently Mutated Genes in a Cohort



/

Browser

🤖 Secure Data Enclave

Hi! Welcome to the Secure Data Enclave!

I'm a generative AI model trained to assist you in answering questions about the data housed in the enclave. Below are some examples to get you started!

Examples:

1. What are the top 5 mutated genes in females?
2. Are TP53 and PTEN mutually exclusive in lobular carcinoma patients over the age of 40?
3. What are the top 20 mutated alleles in patients with a diagnosis of breast cancer?

Did you know I can also generate visualizations? simply ask me to graph the results of your queries!

Your message

/

Browser

🤖 Secure Data Enclave

Hi! Welcome to the Secure Data Enclave!

I'm a generative AI model trained to assist you in answering questions about the data housed in the enclave. Below are some examples to get you started!

Examples:

1. What are the top 5 mutated genes in females?
2. Are TP53 and PTEN mutually exclusive in lobular carcinoma patients over the age of 40?
3. What are the top 20 mutated alleles in patients with a diagnosis of breast cancer?

Did you know I can also generate visualizations? simply ask me to graph the results of your queries!

Your message

Demo 2: Mutual Exclusiveness of a Gene Pair

Secure Data Enclave

Hi ! Welcome to the Secure Data Enclave!

I’m a generative AI model trained to assist you in answering questions about the data housed in the enclave. Below are some examples to get you started!
Examples:

1. What are the top 15 most frequently mutated genes in lung cancer?
2. Are TP53 and PTEN mutually exclusive in breast cancer?
3. What are the top 20 mutated alleles in patients with a diagnosis of breast cancer ?

Did you know I can also generate visualizations? Simply ask me to graph the results of your queries !

What is mutual exclusiveness of TP53 and KRAS in lung adenocarcinoma

New Chat

Secure Data Enclave

Hi! Welcome to the Secure Data Enclave!

I'm a generative AI model trained to assist you in answering questions about the data housed in the enclave. Below are some examples to get you started!

Examples:

1. What are the top 5 mutated genes for women?
2. Are TP53 and PTEN mutually exclusive in lobular carcinoma patients over the age of 40?
3. What are the top 20 mutated alleles in patients with a diagnosis of breast cancer?

Did you know I can also generate visualizations? simply ask me to graph the results of your queries!

what is the mutual exclusiveness of TP53 and KRAS in lung adenocarcinoma

Use Cases and Key Features

Use cases

- Top N most frequently mutated genes
- Determine mutual exclusiveness
- Screening mutual exclusive gene pairs
- Mutation pattern
- Top N mutated alleles

Key features

- Access to underlying API
- Support for natural language search
- Generate figures, tables and text summaries
- Support synonym
- Uses chat history as context

Next Steps

Data processing

- Process genomic data using GDC pipelines
- Harmonize clinical and demographic data

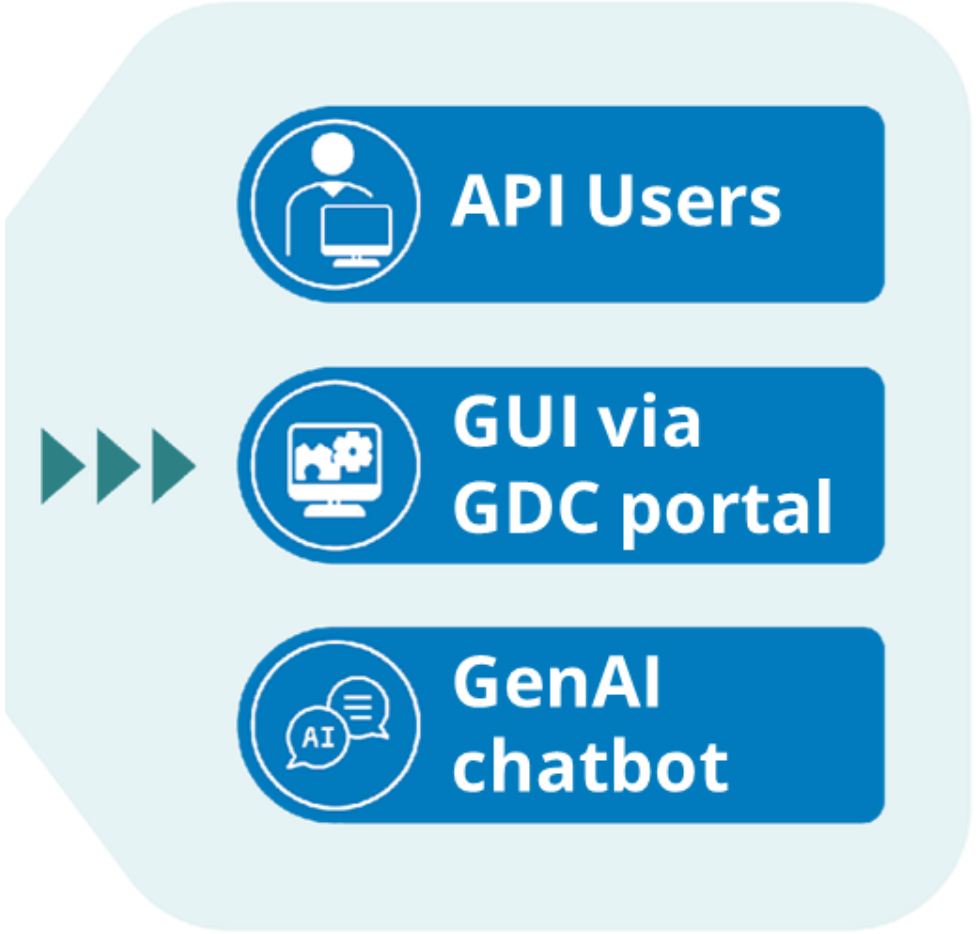
Add new datasets

- The more data, the better

Develop new use cases

- The more use cases, the more useful

Next Steps



Release API to public

Integrate API with GDC portal

Launch GenAI Chatbot

Acknowledgments

NCI

- Jill Barnholtz-Sloan
- Tanja Davidsen
- Allison Dennis
- Jaime Guidry Auvil
- Craig Hayn
- Subhashini Jagu
- Tony Kerlavage
- Sue Pan
- Krish Seshadri
- Jeff Shilling
- Lou Staudt
- David Sturgill
- Jean Claude Zenklusen

AWS ProServe

- Howard Jackson
- Jason Marr
- Tayo Ogunmakin
- Amy Penchuk
- Kevin Sayers
- Jeff Ward
- Elijah Weske
- Tyler Willis

FNL/UChicago

- Sharon Gaheen
- Bob Grossman
- John Otridge
- Jermaine Rogers
- Himanso Sahni
- Mike Warfe
- Zhenyu Zhang

- All CRDC team members
- All partners throughout NCI/NIH and data contributors