

What is the Cancer Gene Trust (CGT)?

The [Cancer Gene Trust](#) is a Demonstration Project of the [Global Alliance for Genomics and Health](#) (GA4GH). It is an online resource for sharing somatic cancer genomic and clinical data, consistent with appropriate patient consent. The CGT contains some fully public data, specifically somatic variants, from consented individuals, which are freely accessible to anyone with an internet connection in real time. Additional, restricted-access data remain under the jurisdiction of the data depositor, called a steward, and more information on specific cases may be made available to individual users directly from the steward.

What happened to the Actionable Cancer Genome Initiative (ACGI)?

The CGT evolved from the Actionable Cancer Genome Initiative, which had a more clinical focus on actionability in cancer variant data, due to a few observations:

- It is straightforward to share all somatic variants in a VCF from an informatics perspective, and relatively straightforward in some jurisdictions from a regulatory and ethics perspective
- Variants in a VCF cannot *a priori* be disqualified in the long-term from potentially contributing to clinically relevant observations, and therefore should be aggregated when possible
- Therefore, a foundational repository of **all** somatic variants, along with some clinical information, provides a practical and powerful way to achieve comprehensive clinically useful outputs
 - Instead of an initial focus on actionability, we will support development of annotation applications on top of this foundational repository - thus, we will focus on developing the foundational CGT resource as a necessary precursor to any focus on actionability

Why is data sharing important for scientific and medical advances?

No one institution houses enough data to understand every individual's cancer; however, if many institutions pool their datasets, many more discoveries will be made that improve our research understanding of cancer as well as potentially enabling improved diagnosis, treatment, and prognosis of cancer patients in real time. Making data available publicly, thereby giving researchers and clinicians from all over the world access to the data, allows these advances to be made as quickly as possible.

What data are publicly available in the CGT?

The CGT publicly displays somatic cancer variants (the genetic mutations that occur only in the tumor and **differ** from an individual's regular genome, which was inherited from his or her parents) as well as some clinical information that is not identifiable. The CGT may add other somatic tumor data types in the future, such as gene expression levels.

Who can access the CGT?

Anyone who has access to the internet can access the public CGT at cgt.ga4gh.org.

What is a steward?

A steward is an organization that contributes data to the CGT. The steward maintains ownership of and jurisdiction over its complete dataset, including patient consents and appropriate approvals for use. Stewards deposit somatic tumor mutations and some clinical data to the publicly available CGT, and can be contacted by users of the CGT if the users are interested in learning more information about individual cases they have deposited. The steward alone controls who gets access to its non-public data, to ensure consistency with its own patient consents.

Who can contact a steward through the CGT?

Any user of the CGT can contact a steward.

How can organizations with cancer data share their data through the CGT?

By becoming a steward themselves and submitting data through their own CGT server, or by submitting data to an existing CGT steward.

How does a steward give a CGT user access to case-level (i.e., protected) data?

Initially, each steward defines user access individually. Future projects related to the CGT may define more standardized processes to access and compute on case-level data for qualified users.

Does the CGT Public Record pose a risk to individual privacy?

It is our belief and understanding that the CGT public record containing limited somatic mutation data, limited clinical data (e.g., tumor type, location and age at diagnosis), and gene expression levels about an individual patient pose *no reasonable likelihood of re-identification*. It is commonly understood and accepted that somatic genomic data pose a much lower risk of re-identification than germline genomic data, due to the absence of a hereditary component to the genomic information (i.e., even if an individual's family members or ancestors have genome sequencing performed, those relatives reveal no somatic information about the individual; an attacker would require access to the individual's tumor sample sequence to re-identify). We also believe that sharing a limited number of somatic variants, in turn, poses a much lower risk of revealing germline genomic data than tumor whole genome data. Because of international regulatory variation, however, it remains the responsibility of the local steward to determine whether or not the limited data included in a CGT public record are identifiable under applicable laws and guidelines, and to obtain the appropriate patient consents and/or authorizations for sharing.

How is the random number in each submission generated?

Each submission may contain a random number to enable access requests or re-contact. This number is not derived from a patient's personal information in any way. In particular it is not derived by 1-way hashing the patient's personal information. It is completely random. The

random number maintains a link between a public record and a patient's more detailed genomic and clinical data held securely by the trusted steward.

How can data scientists participate in the CGT?

Data scientists can participate in the development of the CGT software via [GitHub](#). Once the network has been established, they can participate by developing applications to search and curate the publicly available CGT to improve cancer research and clinical care.

How can cancer researchers and clinicians participate in the CGT?

Using applications developed to query the CGT network, cancer researchers and clinicians can search cases in the CGT to identify data that are useful for their research and patients. They can also encourage and participate in data deposits to the CGT as members of a steward.

How can cancer patients, families, and patient advocates participate in the CGT?

Patients, families, and advocates can donate their data to the CGT through an existing steward, or by setting up a new steward. Until the network of stewards grows, the choice of stewards will be limited. However, we will help patients, families and advocates encourage their care providers' organizations and advocacy groups to set up new stewards. Patients, families, and advocates will also be able to use applications developed over the CGT network.

How does the CGT interact with other data sharing projects, like the Genomic Data Commons, ICGCmed, and GENIE?

The CGT has active collaborations with many other cancer data sharing initiatives, including the Genomic Data Commons, ICGCmed, and AACR's GENIE. It is the intent of the CGT that these organizations will either become stewards or support their contributors as stewards.

How will the CGT help cancer patients? Clinicians? Researchers?

The first goal of the CGT is to build a foundational online data network that will share cancer data simply and efficiently in real time. When this network is fully operational, it will be possible to build web applications that use the CGT's data to help locate cases that may inform research as well as clinical care of individual patients. In this way, the CGT will help researchers identify the best and most complete data for their research, it will help clinicians make observations that lead to new translational research and improved clinical trials, and ultimately, it will create the ecosystem in which an individual cancer patient could ensure that the data generated by their struggle is used to help others, and receive personalized insight into his or her disease in real time from other cases that came before, thus enabling discoveries of better therapeutic options and improved long-term outcomes.

What types of web applications can be developed using the CGT network?

Many types of applications will be possible using the CGT data network. For example, ongoing work in the GA4GH including the [Variant Interpretation for Cancer Consortium](#) and RNA Task Team will be able to use the network for improved cancer variant clinical curation and RNASeq analysis. Additional applications could allow users to watch for mentions of a specific cancer

variant, drug dependency, or tumor site of origin, sending real-time updates when new data are available that meet specific criteria. We also expect that researchers will think of new uses for the data that will benefit patients, that the original developers of the CGT have never considered - the public, real-time nature of the data will enable new cancer discoveries and collaborations in a way not previously possible.