The Cancer Gene Trust

Background

Cancer is a disease of the genome. Thanks to new genome sequencing technology, we have seen breath-taking advances over the last ten years in our understanding of which genomic mutations accumulate in cancers. However, we also realized how big the sheer number of possible mutations is in the three billion letters of our genome. For the majority of the mutations we do not know if they drive uncontrolled cell growth or are rather mere passengers that have little to no effect. Today, it seems that there are at least thousands of different combinations of genomic mutations that can lead to what looks like identical tumor cells under a microscope and only a few dozen mutations are treatable with drugs, most of them in trials.

Because of the number of possible mutations, it is commonly accepted that in order to better understand cancer, we need bigger cohorts, more data, more sequenced tumors and more medical records describing treatments that were tried. With more data, the hope is that it will be easier to find the “core” of cancer driving mutations. This is why the NCI has funded various projects like The Cancer Genome Atlas to directly sequence tens of thousand of cancer samples from patients and collect the data in the NCI Genomic Data Commons. Most governments in developed countries also have started similar cancer sequencing projects, and pool at least some of their results as part of the International Cancer Genome Consortium. All these projects recruit patients, sequence their tumors and follow their treatment and medical interventions to one extent or another, but sequencing is done in a research setting and the results are more archival and reference in nature. Clinical trials and research projects have dedicated staff for patient consenting, sufficient funds for data management and special access-controlled data centers to comply with data privacy regulations.

At the same time, outside of the research enterprise, cancer genes are increasingly being sequenced for clinical reasons, e.g. to choose between targeted drugs and chemotherapy. The sequencing is done by clinical testing laboratories. Even smaller US hospitals increasingly order these tests. However, only very few of these sequencing results are used in the end, as clinical operations lack the infrastructure mentioned above used by research efforts. For technical reasons, large regions of the genome are usually sequenced in clinical testing, but only the actionable mutations, ones that can be used for immediate diagnostic treatment are reported back to the oncologists.

Problem

A hospital rarely requests the full dataset including non-actionable mutations found in a genome, as these are not relevant yet for diagnosis. Thus, in virtually *every* case, most of the sequencing results are not further analysed or seen by anybody outside the testing lab. Even the data that are reported back to the hospital rarely make their way back to research.

Instead research must depend on explicit data gathering efforts such as those mentioned above. A related problem is the raw genomic data is typically too large to easily move to other research centers for analysis.

If instead all tumor mutations determined from sequencing flowed back to research then a virtuous cycle of data from clinical activity fueling research discoveries leading to more effective clinical activity would emerge. If the cycle could be real-time then communication between clinics with similar patients for the purpose of better diagnosis, and ultimately even research discoveries, could begin to move closer to clinically actionable timescales. Ideally de-identified public data would be globally open, drastically improving the odds of identifying a pattern in rare variants. Currently, it is not clear how the different actors that keep parts of the genomic or clinical data or need access, like diagnostic labs, hospitals, patients, researchers or companies can share these, especially internationally and also - given the privacy requirements in health care - how they can follow up once the data is exchanged.

Solution

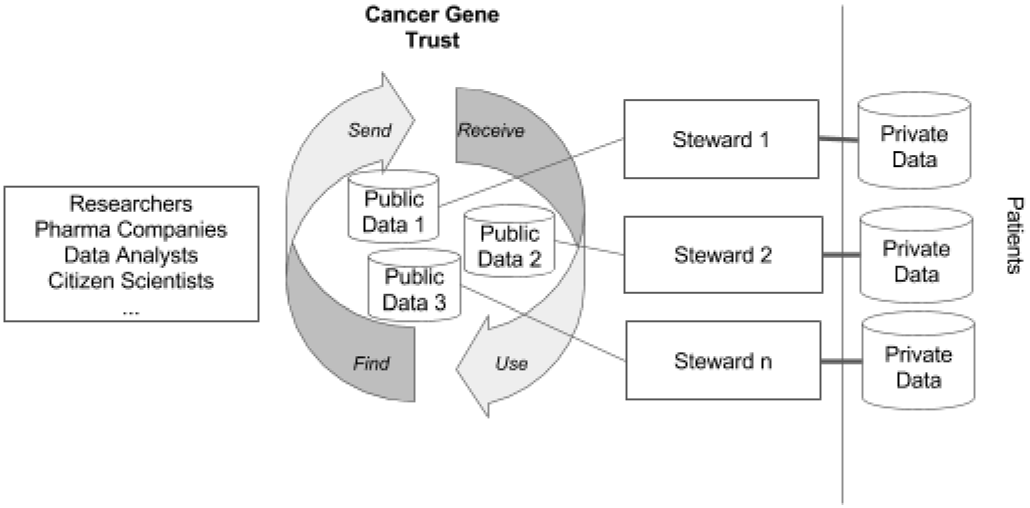
UC Santa Cruz Genomics has significant experience with the management and distribution of genetic data: from the first public assembly of the human genome in 2001 to the UCSC Genome Browser serving genome annotations since then to tens of thousand of users, through providing cancer genomes through the NCI CGHub and currently with the NIH Big Data To Knowledge Center.

Based on this experience, we are building the “Cancer Gene Trust” (CGT), a system to liberate clinical sequencing results from hospitals around the world and make them accessible for research analysis in real time. The immediate sharing of data democratizes the analysis allows more experts to participate and compare results and accelerates the translation of genomic findings towards a clinically useful timescale. Our approach relies on technology to create a lightweight, decentralized network of “stewards” that make de-identified data publicly available and available for re-contact after the initial publication.

A steward can be a hospital, a collection of hospitals or any organisation, that manages protected health data of patients and can anonymize them into a publicly shareable dataset. Stewards can also be research centers or companies, that do not have private data, but download (“mirror”) and analyze / annotate copies of public data provided by other stewards. Public data is anonymized, it includes the DNA mutations that occur only in the tumor (not germline), other molecular tests of gene expression levels and protein activation, and general clinical data such as age, cancer type, year of diagnosis, and as much treatment/drug information as the steward is comfortable sharing. While the genetic data may not reach the quality and reliability of the centralized infrastructure built for the government research projects with dedicated funds, the CGT can have data from an order of magnitude more cases, and ultimately may have better clinical data.

Our data sharing system has no central server but uses a block chain instead. It makes sure that the original steward that submitted a dataset is publicly known and can be recontacted securely. The system makes it easy to give credit to the submitting steward when the dataset is used. When a steward is not available temporarily, the system finds backup copies of the data. For bigger files, data is streamed from multiple stewards in parallel.

In our model, all public data are anonymized, with names, addresses and other personal health information removed. Patient participants are identified with a random number whose association with them is known only to the steward. The protected patient data, including the raw DNA and RNA sequencing reads, always stays with the steward, and to the extent possible is separately archived in the protected data section of the NCI Genome Data Commons so that it too can be made available, but in this case only to specially qualified researchers.



Ingest

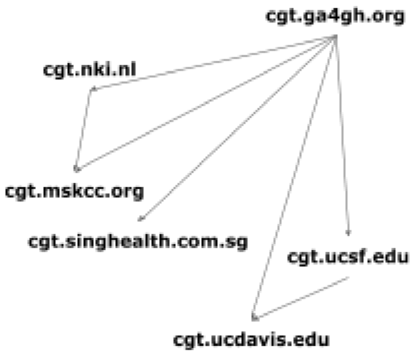
Sharing data is not free, being a steward requires some work. Collecting this information in a hospital setting requires some time. To minimize the cost, we try to reduce manual work to a minimum by writing customized software for the staff from two separate departments:

1. The clinical lab is the department that sends the tissue sample to the testing company and receives the results. We will provide software so that their technicians can extract from the test results only the mutations that occur in the tumor, as well as gene expression levels, protein activations and other clinical test results when available. The software will anonymize and submit these data to the Cancer Gene Trust automatically with just a few clicks.
2. The cancer registrar in a hospital collects clinical data like age, tumor stage, histology etc., summarizes them and reports them to state or federal authorities to aid with epidemiological and related questions, e.g. to identify counties with higher cancer incidence. Our second software module will automatically summarize and share the appropriate de-identified general clinical data that can be made public with the Cancer Gene Trust. Again, our goal is that this should not require more than a few mouse clicks.

Data Storage and Distribution

Many national efforts to share genomic data are organized around a centralized entity with strict standards for content, curation and access. This creates a high barrier for participation as well as a high level of friction slowing the growth of the community as well as the speed by which data is shared. Each individual submission can be addressed by URL as well as the cryptographic hash of its content for compact reference on the blockchain. A cryptographic hash is a unique identifier for the data and changes when the data change.

An individual submission consists of de-identified clinical data, a list of somatic mutations and optionally expression data. If a user asks any node for a submission (by hash), the node will send the file if it has it, and if not requests the file from other nodes. Each node may also list the address of peers in its index and thereby provide a means to add and find other nodes in a decentralized manner. Given the cryptographic hash of a data submission, the system will automatically find the closest stewards that have the file and stream the file in parallel to the requester.



CGT peering connections between domestic and international institutions

Blockchain Public Ledger

Our data storage system does not explicitly track who submitted which dataset, when they submitted it, where backups are kept or who used which dataset. These functions are provided by the Cancer Gene Trust (CGT) Public Ledger. The CGT Public Ledger is a Blockchain, instead of a central server.

The CGT Public Ledger supports the following operations:

1. **Submit:** When a dataset is uploaded to a webserver on the internet, the steward adds an entry to the block chain that includes the cryptographic hash of the

submission and its location on the internet. The hash publicly shows that this steward and no one else had this data and when. The location shows where the file can be downloaded. The submission hash is used in all subsequent block chain entries that refer to this submission.

2. **Mirror:** Some research or commercial institutions (e.g. NIH's data centers like the NCI Data Commons) may not generate a lot of data but focus on data analysis and run bigger IT systems than a typical hospital. They have an interest in keeping a complete copy of the data for their own analyses. With this blockchain entry, the "analysis" stewards can publish the fact that they keep a backup copy of a given dataset which can be used when the main steward's copy is unavailable. CGT will automatically serve the data download from the closest mirroring steward. Ideally, a smart contract organizes mirrors in a way that ensures that a certain number of copies of each file are always available in the system.

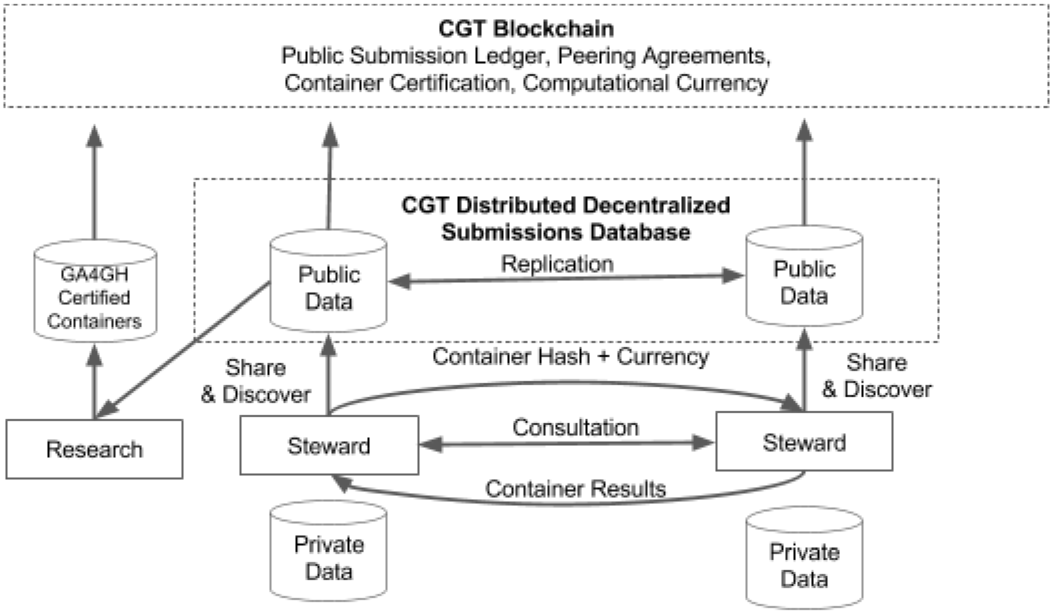
3. **Download:** Users can publicly declare that they have obtained a dataset identified by a hash. Because data storage is separate in our system from data tracking, the system has no way of enforcing that this transaction is indeed used. However, if this is part of the download client software used to obtain files, and possibly checked later when stewards are recontacted about a dataset or by journals when analyses are published, we hope that at least partial external enforcement of this transaction leads to credit given back to the original submitting steward that is independent of scientific publications or patent applications. A smart contract can sum up the data reuse transactions into a "reliability score" for stewards that could be used to rank stewards by how much their data is used by third parties.

4. **Remove:** Legally, patients have to be able to withdraw their data. In a health data setting, one disadvantage of the block chain is that it is unmodifiable. In our system, the data is not stored on the block chain so data can be removed. This entry publicly requests that all stewards remove a given dataset from all storage devices. Compliance (within the network) can be enforced by checking the data storage system.

5. **Update:** In a distributed system, no central curation staff quality-check submissions, so updates may often be necessary. When a submitting steward or any other party finds errors in a submission or if they want to enrich a dataset with their own annotations (e.g. flag suspicious data or provide mutation coordinates on a new human genome version), they can point to an updated submission. This record includes the original submission hash, the new hash and the web address of the update. The identity of the updating steward is guaranteed through the blockchain and stewards can use the history of transactions of this steward in the blockchain to determine if they trust the steward enough to use the updated data version (see reliability score, under "Download").

6. **Recontact:** Data users can declare that they are interested in contacting the submitting steward, optionally including a given patient ID. With the same type of blockchain transaction, stewards can declare that they followed up on a request. A possible scenario could be that a researcher or company found a patient that qualifies for a clinical trial. The source of the "submission" transaction for the dataset is the receiver of this transaction and its public key can be used for encrypted out-of-band (e.g. email) messages. Like the "download" entry, this transaction can be used by a smart contract to give credit to stewards for successful recontact requests, e.g. calculate a "reliability score".

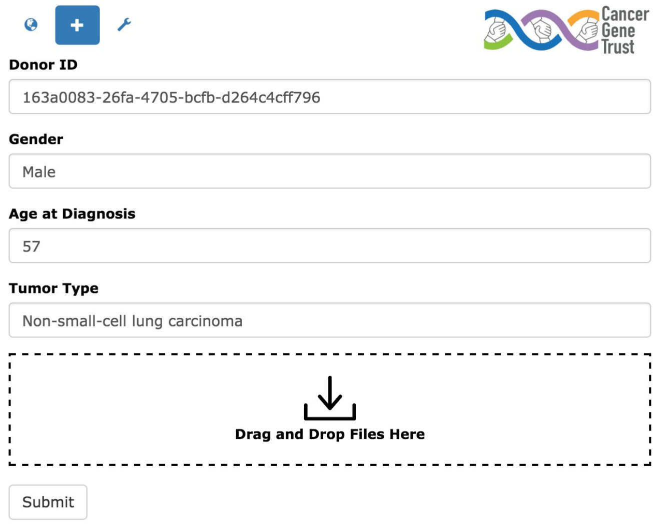
7. **Recompute:** As the full genomic data is private, relatively large and only available to the steward, a third party cannot analyze the full genome. Instead a steward may send a transaction to another steward that includes the hash of a "docker container" that it wants to run on the receiving steward's private data. A "docker container" is a complete analysis pipeline that can be run on any computer system. The Global Alliance for Global Health (GA4GH, https://genomicsandhealth.org/) has begun to curate Docker containers as well as an API. These containers can be used to interact with private genomic data (http://ga4gh.org/#/cwf-team). If the transaction is accepted by the receiving steward, the results of the analysis can be sent back encrypted using the public key of the requesting steward. Optionally the results of the computation may also include a fee charged against the requesting steward's account on the blockchain. This both provides an incentive to maintain private data and allow compute on it as well as prevent abusing the system. Much like the organic growth of the underlying network this will allow an economy to evolve whereby there may be stewards that primarily consume, or those that primarily provide, all mediated via a genomic currency on the blockchain.



Implementation

The solution presented above is not theoretical, we have started to implement most parts with the exception of a reliability score smart contract and the private data compute (docker integration). Our data storage layer is based on IPFS (interplanetary file system, ipfs.org) which handles decentralized storage and parallel streaming. Our IPFS upload module has a straightforward user interface, where files can be dragged onto a webpage to publish them.

Our public ledger module does not implement a new block chain but uses existing frameworks. The Cancer Gene Trust block chain can be either public or consortium-based, where existing stewards are allowed to add their peers to the network. A public blockchain data submission command line interface is available from github and uses the Ethereum network. Ethereum is a publicly accessible block chain with miners and around 9000 active compute nodes, which means that scalability is not a problem with this software. We are now working towards a solution built on the software IBM Open Blockchain / Hyperledger that allows a consortium-based network without mining. All stewards as consortium members would then provide the data in the block chain for outside parties as read-only tables and someone that is not known to at least one steward could not write to the block chain.



Donor ID
163a0083-26fa-4705-bcfb-d264c4cff796

Gender
Male

Age at Diagnosis
57

Tumor Type
Non-small-cell lung carcinoma

Limitations of and Alternatives to the Block Chain

Fully centralized systems for public and private genomics data sharing already exist. Databases like NIH's dbGaP host private and public data and allow access to them for researchers that apply and can show that they have a legitimate interest in the data. Once data is submitted, dbGaP takes all burden away from the data submitter in that the dbGaP review board decides who is allowed access private data. However, patients have to be consented specifically for this type of data sharing, hospitals have to submit the data in a special way and the dbGaP infrastructure is mostly meant for NIH-funded projects. It is currently not clear who on the hospital side will take care of recontact or recompute requests that come in through the Cancer Gene Trust. As there is no existing system, there is no staff to handle these questions. We hope that with increasing demand, contact points will be created. A light-weight system that shares only public data from various sources can never achieve the consistency of a centralized infrastructure, but it can collect much more data and with a faster turnaround. Some stewards may never reply to recontact requests, this will be publicly visible in the block chain, but still, at least the somatic mutation data has been made public. In the larger sense, databases like dbGaP are not incompatible at all with the distributed system presented here, but rather would become data stewards, as they manage private and public data access.

We are unaware of an existing centralized sharing system for only public data, where private data stays with the hospital. Currently, cancer mutations are collected and shared by some institutions (COSMIC, Sanger Center, UK) but with little clinical data and only for datasets tied to research studies, without a formal recontact procedure. As an alternative to the Cancer Gene Trust Network, a new centralized database could keep copies of all public steward data, the public keys of submitting stewards and could keep track of data download, reuse and recompute requests. This central database would require funding at the least for its IT infrastructure and to gain trust in its long-term survival, it would have to be funded in some form by the US government, and to be international, with matching funds from other countries. Such international funding agreements may be hard to achieve.

Thus, in theory, a decentralized system could be built that uses existing database federation technology to enable shared write access to a global table of data submissions and use tracking. However, we are unaware of an existing non-blockchain database system that guarantees database consistency, scales to thousands of nodes, manages cryptographic keys and implements this without using a central server.

Conclusions

More genetic data sharing will allow to make better informed health decisions (a PCORI aim) and follows the Nationwide Interoperability Roadmap Principles, as it builds upon existing systems, protects privacy, scales to thousands of institutions and is modular and cost-effective.

Our proposed system for clinical genomic data sharing uses the blockchain as a public ledger to organize a decentralized data distribution network that connects data providers and data users without relying on a central database. It addresses two increasingly important problems in the life sciences, sharing of data (see Siu et al, Nature Medicine 2016 22:464) and tracking of data reuse independent of publications (Bourne et al, Nature 2015 527:S16). It establishes trust between data submitters and users, by tracking which public key holder submitted what, when and which third parties downloaded the data. By separating the distributed data storage from the public ledger, the system allows deletions and more importantly combines “big data” with a block chain.

Our block chain solution has no centralized database that is located in the US. This means it does not require permanent central funding for the data storage and makes it a lot easier to sell the idea abroad; international collaboration is needed to maximize the network size. The decentralized nature gives the hospitals/healthcare organisations a sense of ownership and participation which probably fits the structure of the US healthcare system better than a new centralized system. A decentralized network also reduces the tendency of certain institutions to “own” the platform and restrict or filter submissions. We believe that by distributing the storage and submission work over the network means that our system is cost-effective, in the sense that the cost is spread over many actors. The main application of the blockchain is public information, but the Cancer Gene Trust system still enables computation on private data as it establishes the trust relationships needed to permit running certified software containers by stewards on private data for third parties. These can potentially reimburse stewards for the computation cost through a future form of currency stored in the block chain.

Our group played a big role in the Human Genome Project, which brought the world together to discover our shared genetic heritage and lay the groundwork for a revolution in medicine. We believe that by using the blockchain, the Cancer Gene Trust system will enable patients through their stewards to make their precious data readily available for research and treatment. At scale, such a system could reveal the nearly complete molecular character of cancer, and engender a revolution in cancer understanding and treatment.