Project M)

Designing a Cancer Genomics Commons

Kenna Shaw-
Director, The Cancer Genome Atlas (TCGA)

Biology is becoming a data intensive science. Specifically, high-throughput genomic platforms are generating datasets that add petabytes of data each year to the public domain and this rate is likely to increase. To date, most data sharing has been supported through transactional databases that allow for bulk upload by data producers, and data download by individual labs for local compute and analysis. This model is not sustainable and results in extremely high resource costs to the federal government. For a scientist who wants to use this community resource, every byte of interested data must be downloaded to a local data store and they must become intimately acquainted with numerous data models and distinct portals that house data in order to integrate across studies. This requires gigabit level network bandwidth and replication of the storage environment. The result is thousands of hours of download time, hundreds of man hours appropriated, and millions of dollars spent in network and storage costs. Because of these circumstances, access is effectively restricted to researchers with the infrastructure and skills for petabyte level data analysis.

Several principle challenges serve as bottlenecks towards accomplishing a vision where a multi-dimensional, clinically-annotated, computationally robust, widely-available sustainable resource is available to *every* member of the community.

- Data Heterogeneity: Our ability to extract useful information from the integration of molecular and clinical data is predicated on the development of appropriate analytical methods and tools that can cope with the statistical challenges and computational bottlenecks inherent to such heterogeneous, high dimensional data sets. The statistical relationships that exist in such data are highly nonlinear and multivariate (i.e., involving higher-order relationships), while the variables ("features"), which can be continuous, discrete, binary, categorical, etc., can be highly redundant/correlated or have missing values. As new-omics scale technologies become more robust, new data types, formats and approaches will need to be integrated on top of existing data structures, requiring flexibility and portability of data and analytical approaches. Data producers must reach consensus on standards and invest in pipelines that not only generate the highest quality data, but also the accompanying metadata and documentation that will facilitate integration by other users.

- Computational: Carrying out large-scale analyses, which entail enormous search spaces, will require engagement of the high-performance computing experts in this conversation. Practical considerations (e.g., economic, computational/time resources) demand innovations in how HPC can be optimally used for extracting knowledge from data – for example, how can HPC be harnessed with a tight integration of the scientist "in the loop" steering the computation in such a way that answers are obtained more quickly and can be used to ask the next questions,

iteratively. This calls for development of workflows and interactive tools (eg, user interfaces, visualizations) that empower the scientist to work much more effectively with very large heterogeneous data sets and "ask questions" (thereby carrying out large-scale computation on the data). Moreover, developing sustainable models to support both storage and computational infrastructure that supports the broader community must be developed.

- Patient Engagement: As the scientific community works to balance patient expectations related to research-grade data and clinically-actionable tests, more and more patients are still demanding access to data derived from their donated samples. We will be moving into a domain where more patients carry their germline and tumor genomes with them. Bringing these individual expectations and contributions to a broader Cancer Genomics Data Commons will exponentially increase not only our dataset but our need to develop an environment where patients and researchers share the same space and communicate in the same environment. This will require better algorithms, improved consents that enable patients to determine the level of data sharing they desire (e.g. the Portable Consent), and more flexibility around questions/policies related to risk, access and consent groups. Questions about the need for Federal-level policies/laws that could protect against the re-identification of research subjects might also be in play here.

- Scale, Group Science and Value Paradigms: Cancer is complex. Each tumor from each patient represents a unique disease. In order to gain the statistical power, teams from around the world need to be engaged from sample identification through data analysis- often involving dozens, if not hundreds, of individuals. The investment of individual analysts in ‘team science’ projects (e.g. The Cancer Genome Atlas) is equivalent on a single major paper as it is on smaller, more individually directed projects. Yet, these authors are clustered with many others and their contribution unclear. In an era where knowledge best comes from collaborative investigation, the paradigm of counting of ‘first author’/’last author’ on a curriculum vitae must be challenged.

Over to You

M- Designing a Cancer Genomics Commons

G	Organization	Last name	First
M	Columbia	Anastassiou	Dimitris
M	UCSC	Ellrott	Kyle
M	Genentech	Jackson	Peter
M	Genometry	Lamb	Justin
M	Color Genomics	Laraki	Othman
M	NIH/NHLBI	Larkin	Jennie
M	University of Washington	Lee	Su-In
M-Anchor	Sage Bionetworks	Margolin	Adam
M	Cancer Research UK	Markowetz	Florian
M	University of Washington	Olson	Maynard
M-Anchor	Sage Bionetworks	Omberg	Larsson
M-Lead	NIH	Shaw	Kenna
M	UCSC	Stuart	Josh
M	UCSF	Van't Veer	Laura
M	St. Baldrick's	Weaver	Becky
M	University of Chicago	White	Kevin

Designing a Cancer Genomics Commons
Kenna Shaw (TCGA Director) & Larsson Omberg (Sage)

Project Overview:

- **ALL data relevant to cancer ALL in the same place, Accessible to ALL**
- **Patient centric vision, not snapshots but long-term**
- **Leverage the healthcare system (& the patients use it) that is built to track them**
- **Need to build a system that couples top down-driven datasets as well as individual-level uploads**
- **"IT" is unlikely to be limiting today and even into the future; consent and incentives biggest barriers**
  - **Liability- implications of analysis done right and wrong**
  - **Quality- data**
  - **Separate the research and clinical research/data environment**

Potential alignment with existing Commons' approaches

- Moffitt, Vanderbilt.
- Patients like Me.
- Need to drive interoperability
- "Commons" that exist in silos

Unmet needs and issues

- Patients do not easily have access to their own medical information
- Clear incentives to participate
- Clear documentation about how to distribute data and mandate metadata

1-year vision for the future of this project

- Nucleate the model
- Accept standard vocabularies
- Begin requiring clinical trials, other NIH-funded efforts, to participate in Commons
- Part of funding should support the participation
- Minimum set of data that meet a set of standards that the COMMONS generate and adopt
- Structured, open formats needed for genomics and clinical records
  - Pick a specific clinical phenotype (response to treatment; overall survival)
- Empower patients: Consent forms that push towards more complete medical records
- Start the discussion: genetic privacy and legislation
- Show it can work and MAKE IT FUN!
  - Pathology report challenge
  - Mutation caller: benchmarking data sets