Title

Enabling transparent and collaborative computational analysis of 12 tumor types within The Cancer Genome Atlas

Journal

Nature Genetics, 45(10)

ISSN

1061-4036

Authors

Omberg, Larsson
Ellrott, Kyle
Yuan, Yuan
et al.

Publication Date

2013-10-01

DOI

10.1038/ng.2761

Peer reviewed

Published in final edited form as:
Nat Genet. 2013 October ; 45(10): 1121–1126. doi:10.1038/ng.2761.

Enabling transparent and collaborative computational analysis of 12 tumor types within The Cancer Genome Atlas

Larsson Omberg[1,6], **Kyle Ellrott**[2,6], **Yuan Yuan**[3,4], **Cyriac Kandoth**[5], **Chris Wong**[2], **Michael R Kellen**[1], **Stephen H Friend**[1], **Josh Stuart**[2], **Han Liang**[3,4], and **Adam A Margolin**[1]

[1]Sage Bionetworks, Seattle, Washington, USA

[2]Department of Biomolecular Engineering, University of California, Santa Cruz, Santa Cruz, California, USA

[3]Department of Bioinformatics and Computational Biology, University of Texas MD Anderson Cancer Center, Houston, Texas, USA

[4]Graduate Program in Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, Houston, Texas, USA

[5]The Genome Institute, Washington University, St. Louis, Missouri, USA

Abstract

The Cancer Genome Atlas Pan-Cancer Analysis Working Group collaborated on the Synapse software platform to share and evolve data, results and methodologies while performing integrative analysis of molecular profiling data from 12 tumor types. The group's work serves as a pilot case study that provides (i) a template for future large collaborative studies; (ii) a system to support collaborative projects; and (iii) a public resource of highly curated data, results and automated systems for the evaluation of community-developed models.

The Cancer Genome Atlas (TCGA) Pan-Cancer project, consisting of over 250 collaborators spread across almost 30 institutions, required researchers to engage in over 60 different

Correspondence should be addressed to L.O. (larsson.omberg@sagebase.org) or A.A.M. (margolin@sagebase.org).
[6]These authors contributed equally to this work.

URLs. Synapse, https://synapse.org/; Firehose, http://gdac.broadinstitute.org/; GenomeSpace, http://www.genomespace.org/; W3C provenance specification, http://www.w3.org/TR/prov-primer/.

Accession codes. All input data files used in the Pan-Cancer project are available via Synapse DOI syn300013 (to find any Synapse page referenced in this paper, type the Synapse ID or DOI into the search box on any page at http://www.synapse.org/). Synapse is freely available to any registered user, and all source code is released under an open-source license and available through GitHub (https://github.com/Sage-Bionetworks/Synapse-Repository-Services). We invite members of the scientific community to provide feedback (http://support.sagebase.org/sagebase, access the source code and directly contribute to the software, and suggest additional projects through which to explore a collaborative research paradigm.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

AUTHOR CONTRIBUTIONS

L.O., K.E. and A.A.M. wrote the manuscript with assistance from C.K., S.H.F., J.S. and H.L. L.O., K.E. and C.W. created the visuals for the manuscript. L.O. and K.E. coordinated data aggregation and sharing in Synapse. Y.Y. developed data sampling and primary models for survival predictions. C.K. performed MuSiC analysis and created the corresponding Synapse annotations. L.O. developed infrastructure for scoring and evaluations in Synapse. M.R.K. oversaw the development of Synapse. J.S. and A.A.M. conceived of and oversaw the use of Synapse to support the Pan-Cancer project. The TCGA Research Network contributed all of the results in Synapse.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

research projects oriented on the same set of data. As with other team-science efforts, the Pan-Cancer endeavor required researchers around the world to work in distributed teams to generate, share and interpret large amounts of data. Many projects were interdependent, requiring multi-stage analysis and sharing of results, such that results from one group were used as input for the analyses of other groups. Thus, it was essential to standardize the input data used by all researchers. The results in the current collection of papers are based on integrative analysis of 1,930 input data files encompassing 6 different biomolecular technologies, including protein expression, copy number variation, somatic mutation, mRNA expression, DNA methylation, microRNA (miRNA) expression and clinical data for 12 cancer types. The same patients were sampled across most of the platforms, yielding a coherent data set (Fig. 1).

Currently, collaborative science and data sharing are supported by a number of tools, including repositories for sharing published data, such as the Gene Expression Omnibus[1] and the database of Genotypes and Phenotypes (dbGaP)[2]; more general solutions for *ad hoc* sharing of data, code or wiki content, such as Sharepoint, GitHub and Confluence; standards for file descriptors, such as ISA-Tab[3]; and software tools for running prepackaged methodologies in analysis pipelines, such as Firehose, Galaxy[3], Taverna[4] and GenomeSpace. However, enabling collaborative scientific projects requires tools that facilitate the evolution of knowledge and resource outputs beyond the sum of the group's individual efforts. By incorporating such collaboration tools throughout all phases of the research cycle, rather than as *post hoc* descriptions added at the time of publication, emergent data and analysis resources are created by the collaboration that may be seamlessly released to the general research community, thereby increasing the resource value of the work.

The Pan-Cancer group explored a collaborative model in which all consortium participants worked through the Synapse software platform to share and evolve data, results and methodologies throughout the full duration of the project[5]. Synapse is composed of a set of shared Representational State Transfer (REST)-based web services that support both a website designed to facilitate collaboration among scientific teams and integration with analysis tools and programming languages to allow computational interactions (Fig. 2). Synapse provides an expanding number of features to enhance collaborative analysis of complex genomic data (Box 1). In the following sections, we highlight how key features of Synapse were used by the Pan-Cancer group through three different examples of collaborative analysis: (i) establishing a canonical data set that required strict use of versions and data freezes; (ii) applying multistep data-processing procedures to infer functionally significant mutations; and (iii) comparative analysis of predictive models of patient survival with real-time evaluation of model performance.

BOX 1

KEY FEATURES OF SYNAPSE

Data versioning

Data change over time. Experiments are rerun and new data are generated. However, new data do not invalidate the previous data, which must be maintained to reproduce previous analyses. Synapse allows for data entities to be updated and keeps track of their previous versions while maintaining a single accession number for each entity. Multiple files and specific versions can be combined into data freezes, corresponding to a collection of specified versions of data entities.

Provenance tracking

Data analysis occurs in stages. One researcher will perform an analysis and produce results that are used by another researcher. This may happen several times. Synapse allows for a series of analysis steps to be recorded and visualized using the provenance system. Every piece of data and analysis in Synapse can be tracked by provenance, including the chronology of ownership and links to data and source code used in the analysis. The graph representing the provenance of analysis is based on the W3C provenance specification proposal (see URLs).

Data annotation

A file by itself is not descriptive. Projects with large amounts of data often devolve into collections of unorganized files with obscure names. Synapse has utilities to attach structured typed annotations to data, such as source species, file type, algorithm used or any other characteristic attribute defined by the user. Optionally, each annotation may be associated with a dictionary of possible values, allowing integration with existing data ontologies.

Query language

Finding relevant data is important. In addition to browsing data using the traditional 'file/folder' hierarchy, all annotations associated with Synapse entities may be queried using the web client or using SQL-like syntax from a variety of programmatic clients. Thus, users can quickly search through thousands of files to find the data that are relevant to their experiments.

Governance

Synapse enables the management of controlled data-access mechanisms to share data while maintaining compliance with human privacy protections and/or legal or ethical restrictions. Data use terms are set by the data contributor and are managed by the Synapse Access and Compliance Team according to institutional review board (IRB)-approved protocols.

Group security

Although open research is the ultimate goal, many groups like to work with a level of controlled access during certain stages of the research process. Access can be limited to individual users or groups and seamlessly transitioned to public access at the appropriate time.

Citation management

Everything stored in Synapse is accessible by unique accession numbers (Synapse IDs) that can also easily be assigned a citable, permanent DOI (digital object identifier). These DOIs can either reference specific versions of data or, by choice, track the most recent version of data.

Clients for R, Python, Java and command line

Synapse provides open APIs and fully functional clients in three popular programming languages as well as command line access. Each client communicates with a common set of Synapse REST services, allowing robust access to Synapse features, including loading data objects into analytical environments, creating and editing Synapse entities and creating provenance records.

Rich descriptions

Wiki pages describing data or analyses may be optionally associated with any Synapse entity. These wiki pages fully support markdown syntax and can be decorated with rich

content in the form of widgets, such as tables, graphs, videos or even runnable R instances in the form of shiny apps.

Evaluation queues

Entities uploaded to Synapse may be loaded into an evaluation queue, which kicks off an automated procedure to perform a specified analysis on the newly uploaded entity. Results of these downstream analysis procedures may be stored in Synapse, for example, to implement the real-time leader board systems used to evaluate the accuracy of predictive models.

Aggregation standardized data

To coordinate all of the investigators working on the same data, standardized collections of data sets were released in the form of ‘data freezes’, which served as the input for all downstream analysis. Files in the data freeze were intuitively presented to researchers as lists of tab-delimited matrices for each tumor type and experimental platform. As described below, each file in a data freeze was associated with provenance tracking, data versioning, queryable structured meta-data and bindings to multiple analytical clients.

Each processed data set was associated with a provenance record, depicted as a graph of the input data sets and data-processing procedures used to generate the data (Fig. 3). Such input data sets were aggregated from several locations, including the TCGA Data-Coordinating Center (DCC), Broad Firehose and file-sharing platforms used by the individual TCGA analysis working groups. Synapse uses a federated data model such that data stored on any external platform are represented the same way in Synapse. In contrast to commonly used data-sharing solutions that expose lists of files structured in folder hierarchies, Synapse provides graphical provenance records that accurately describe a data resource in terms of dependencies and structured workflows related to its inputs. For example, although aggregated data matrices are the standard input data for all downstream analyses, researchers can also trace the derivation of a processed data set back to its upstream constituents and processing procedures. In addition, researchers can apply alternative data- processing procedures to generate different versions of the aggregated data sets that can be shared with other analysts and fed into downstream analysis.

The use of data versioning allowed the data freezes to evolve as a living resource as new TCGA data were generated during the Pan-Cancer project. The tracking system in Synapse associates a version with each data file that automatically increments with each update, and each data freeze defines a ‘snapshot’ of a collection of specified versions of the constituent data files. Therefore, the data could evolve with new versions while keeping a data freeze constant.

Each data file was associated with structured meta-data annotations, consisting of strongly typed key-value pairs. We employed standardized meta-data schemas and controlled vocabularies for different types of data within the project, allowing queries using SQL-like syntax from a variety of analytical clients (R, Python, Java and command line). For example, using the R interface, the command

```
synQuery(“select name, id
from entity where
freeze= =‘tcga_pancancer_v4’ and
tissueType= =‘breast’ and
dataSubType= =‘geneExp’”)
```

returns the name and Synapse ID for all gene expression data sets for breast cancer in the fourth data freeze. This feature not only allows the relevant data sets to be discoverable for exploratory analysis but also allows downstream data analysis pipelines to be scripted, adapted and reused. For example, with the release of a new data freeze, all downstream analyses can be regenerated by incrementing the 'freeze' parameter in the example query statement. Moreover, because all data sets are stored in the same format, a cross-tissue comparative analysis can be performed by changing the 'tissueType' parameter and applying the same analytical procedure.

Inferring significant mutations

Pan-Cancer researchers rapidly applied and evolved novel analytical techniques and used Synapse to share results. In addition to the input data described above, collaborators have generated over 1,600 data files representing various analysis procedures and results (Table 1). This data resource (summarized in syn1895888) includes results from the most commonly used algorithms in TCGA publications[6–12] and provides the broader community with improved transparency of the results of each methodology. As an illustrative example, we describe how results and analysis procedures of the MuSiC[13] algorithm were represented in Synapse through the use of features such as wiki-based descriptions and runnable source code contained in provenance records.

The 'significantly mutated gene test' in the MuSiC suite of tools uses as input a list of somatic variants detected in tumor samples and identifies functionally significant mutations affecting genes, gene families or protein domains. These variant lists, made available by TCGA, were mirrored in Synapse (syn1695396) with appropriate provenance to the original sources and then standardized, checked for errors and corrected as necessary (syn1710680). Because methods for somatic variant calling are susceptible to errors, these variant lists were further strictly filtered for likely false positive variant calls (syn1729383). All analysis steps are tracked and versioned on Synapse, with appropriate documentation using markdown-formatted wiki pages describing the details of each step and the data formats of associated results files (Fig. 3). Users may also view the provenance record of the multistep data-processing procedure, as well as the intermediate results and processing code used in each analysis step.

Download of input data and upload of results were automated using the Synapse Python application programming interface (API). Intermediate steps performed in house (outside of Synapse) were reduced to a sequence of commands that invoke tools in the MuSiC suite and other minor pre- or post-processing steps that were documented in Synapse's provenance records, thereby enabling any peer with appropriate computational resources to reproduce the results.

Predictive models of patient survival

In addition to creating capabilities for describing and sharing analysis work-flows, the Pan-Cancer group also explored a research model in which independent groups of investigators collaboratively evolved novel analytical methods through the use of automated tools to assess the performance of each approach[14]. Specifically, the Pan-Cancer group used tools to provide real-time automated assessments, based on common performance metrics, of both 'unsupervised' clustering methods (J.S. *et al.*, unpublished data) and 'supervised' molecular prognostic models of patient survival (Y.Y., E.M. Van Allen, L.O., N. Wagle, A. Sokolov *et al.*, unpublished data).

To evaluate the performance of prognostic models, participants submitted predictions of survival times to Synapse along with the executable code that generated the model. Using an

evaluation queue running in Synapse, the performance of each model was assessed on the basis of concordance index scores, and performance results were provided to participants via an online real-time leader board (syn1710282). Specifically, we assessed model performance in the four cancer types with adequate patient survival data and sufficient sample size: kidney renal clear-cell carcinoma, glioblastoma[11], lung squamous cell carcinoma[10] and ovarian serous cystadenocarcinoma[6]. For each tumor type, 100 randomly sampled training and test data set partitions were stored in Synapse, and researchers submitted prediction vectors for the 100 test data sets using models built on the corresponding training data set. Each model was also submitted with publicly available runnable source code, allowing any researcher to reproduce the results or adapt models to apply to additional data sets (Fig. 2).

To facilitate downstream meta-analysis of model results, each model was associated with a structured set of meta-data, corresponding to (i) cancer type; (ii) molecular data type; (iii) clinical features used; (iv) feature preselection method; (v) number of selected features; and (vi) type of algorithm. Ability to query using these annotation fields and retrieve model scores based on specified criteria allowed a controlled evaluation of modeling factors related to predictive accuracy, similar to in the analytical design used by the MAQC-II consortium[15].

The Pan-Cancer evaluation system for the prediction of patient survival is now publicly available as a community resource to evaluate the accuracy of any user-submitted predictive model of patient survival across the four tumor types (syn1710282). Using a crowd-sourced research model similar to the Sage Bionetworks–DREAM Breast Cancer Prognosis Challenge[16], we believe that enabling the entire research community to collaboratively evolve models and providing real-time objective feedback based on predefined metrics will enable the community to more rapidly converge on approaches that are most likely to yield maximal benefit to patients.

A reproducible research commons

The current collection of Pan-Cancer publications documents the innovations and discoveries derived from over 250 researchers analyzing a common set of data. This Commentary describes the group’s attempt to improve the transparency and reproducibility of its research efforts by pioneering a collaborative methodology in which researchers leveraged a common resource to build off each other’s work. The data freezes, analysis results and evaluation framework for survival predictions each correspond to a new publicly available resource released in conjunction with this work. First, the curated Pan-Cancer data freezes are now available (syn300013), allowing researchers to easily access well-curated, analysis-ready data sets from the TCGA Research Network. Data freezes will continue to be maintained and updated in future expansions of the Pan-Cancer project. Updates will be immediately available to the community, allowing any researcher to use data from and contribute to the Pan-Cancer project. Second, we are releasing a resource of Pan-Cancer analysis results (syn1895888), containing the results of applying most commonly used algorithms developed throughout the course of the broader TCGA effort. Compared to previous reports of TCGA analysis procedures via traditional publication mechanisms[6–12], the Synapse resource of analysis results provides improved transparency and reuse of results reported in the current collection of Pan-Cancer papers. In addition, any researcher may contribute content to this central resource so that it may evolve as a broad community effort in the next phase of the Pan-Cancer project. Third, we are launching a ‘collaborative competition’ framework (syn1710282) through which any researcher may submit survival predictions, based on available clinical and molecular data for the four tumor types currently

analyzed, and assess each prediction's accuracy in real time compared to all other submitted models.

The Pan-Cancer project provides only a starting point to guide aspects of future large-scale collaborative studies, and many improvements must be made in future work. Synapse is currently in beta release, and the feature set (Box 1) evolved dynamically according to the needs of consortium researchers.

To maximize user participation, Synapse was designed to favor simplicity of use at the possible expense of the consistency that is favored by more strict standardization. For example, each data object was associated with structured meta-data represented as key-value pairs. However, the set of keys associated with each object was standardized by convention rather than by enforcing a strict schema. Moreover, the set of values associated with each key was strongly typed but not restricted to dictionaries of possible terms. In future implementations, we will allow project owners to define stricter schemas, including associations with ontologies and integration with tools such as ISA-Tab[3] for meta-data standardization.

Similarly, we provided tools for users to document data dependencies in the form of provenance records; however, to minimize barriers to participation, we did not enforce the use of such tools. An alternative strategy could be to define a minimal standard of provenance associated with each Synapse entity, verifying compliance of uploaded entities or providing tools for the community to flag entities that may require additional documentation. A more technical solution that we have recently implemented allows users to upload executable code specifying a function such that Synapse automatically executes the function, stores the output in a Synapse entity and creates a provenance record corresponding to the input arguments. We adapted this strategy in a previous project[16] in which we enforced adherence to predefined APIs and verified code execution; however, in the Pan-Cancer project, we allowed users to upload analysis code without enforcing such constraints.

Consistent with the strategy of favoring flexibility over standardization, Synapse is intended to provide the connection (the metaphorical 'synapse') between analyses performed in any analytical environment using any computational infrastructure and data stored in any distributed location. Although tools exist for many (perhaps all) individual features supported by Synapse, we believe that the integrated system provides a unique framework for enabling large-scale collaborative analysis projects, and integration with additional tools is a priority for future development. Initial integration has focused on popular programming languages, such as R and Python. Recent work has focused on integration with popular tools, such as Galaxy for the design of analysis pipelines, Cytoscape[17] and WikiPathways[18] for representing and visualizing networks and pathway data, and GenomeSpace for facilitating data conversions across multiple tools. We hope to accentuate and enable the strengths of each tool through integration with Synapse capabilities to share and organize data and results, as well as other features designed to facilitate large-scale collaborative projects (Box 1).

The federated data model employed by Synapse is a particular choice designed to facilitate flexibility. This design allows for the distributed analysis of large data sets in which data and analysis servers must be colocalized to minimize bottlenecks due to data transfer. This design was not required for the Pan-Cancer group, as data consisted mainly of gene-level summarizations, and data transfer to collaborators' local analysis environments did not cause a bottleneck. To facilitate more data-intensive projects, we are exploring the ability to

provision computational environments hosted by Synapse along with colocalized copies of data.

Our work represents a pilot project designed to demonstrate the ability to facilitate a large-scale, distributed collaboration, including design choices intended to minimize barriers to adoption and achieve engagement from all collaborators. As we improve and expand the capabilities of Synapse in the context of future phases of the Pan-Cancer project and related collaborative projects, it will be interesting to explore the tradeoffs between enforcing constraints on how users may perform analyses and represent results versus providing a flexible set of tools and allowing standards and protocols to emerge organically from users.

We believe that the open research strategy of the Pan-Cancer project both increased the resource value of the group's work and accelerated the group's pace of scientific progress. Moreover, by exposing the entire research process through an open resource, any stage of analysis may serve as a starting point for additional scientific projects throughout the community. Thus, we believe that Synapse will enable an acceleration in the rate of discoveries building off our work. As recent studies have suggested[19,20], the benefits of open projects are often most dramatically observed over time. Biomedical research—characterized by worldwide efforts to harness massive data sets and collaboratively evolve understanding of complex systems over long periods of time—may be the field of study poised to reap maximal benefits through an open-research paradigm.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We gratefully acknowledge contributions from the TCGA Research Network and its TCGA Pan-Cancer Analysis Working Group (contributing consortium members are listed in the **Supplementary Note**). We thank J. Wilbanks for discussions on the benefits of open systems. We thank the Synapse development team for building Synapse and providing technical support. We thank the anonymous reviewers for helping us improve the quality of this work through many insightful comments. This work was funded by US National Institutes of Health/National Cancer Institute grant 5U54CA149237.

References

1. Edgar R, Domrachev M, Lash AE. Nucleic Acids Res. 2002; 30:207–210. [PubMed: 11752295]
2. Mailman MD, et al. Nat Genet. 2007; 39:1181–1186. [PubMed: 17898773]
3. Goecks J, Nekrutenko A, Taylor J. Genome Biol. 2010; 11:R86. [PubMed: 20738864]
4. Wolstencroft K, et al. Nucleic Acids Res. 2013; 41:W557–W561. [PubMed: 23640334]
5. Derry JMJ, et al. Nat Genet. 2012; 44:127–130. [PubMed: 22281773]
6. Cancer Genome Atlas Research Network. Nature. 2011; 474:609–615. [PubMed: 21720365]
7. Cancer Genome Atlas Research Network. N Engl J Med. 2013; 368:2059–2074. [PubMed: 23634996]
8. Cancer Genome Atlas Research Network. Nature. 2013; 497:67–73. [PubMed: 23636398]
9. Cancer Genome Atlas Network. Nature. 2012; 490:61–70. [PubMed: 23000897]
10. Cancer Genome Atlas Research Network. Nature. 2012; 489:519–525. [PubMed: 22960745]
11. Cancer Genome Atlas Research Network. Nature. 2008; 455:1061–1068. [PubMed: 18772890]
12. Cancer Genome Atlas Network. Nature. 2012; 487:330–337. [PubMed: 22810696]
13. Dees ND, et al. Genome Res. 2012; 22:1589–1598. [PubMed: 22759861]
14. Bilal E, et al. PLOS Comput Biol. 2013; 9:e1003047. [PubMed: 23671412]
15. Shi L, et al. Nat Biotechnol. 2010; 28:827–838. [PubMed: 20676074]

16. Margolin AA, et al. Sci Transl Med. 2013; 5:181re1.
17. Saito R, et al. Nat Methods. 2012; 9:1069–1076. [PubMed: 23132118]
18. Kelder T, et al. PLoS ONE. 2009; 4:e6447. [PubMed: 19649250]
19. Benkler Y. Yale Law J. 2005; 114:273–358.
20. Kolata G. New York Times. Aug 12.2010
21. Zack TI, et al. Nat Genet. Oct 26.2013 10.1038/ng.2760
22. Lawrence MS, et al. Nature. 2013; 499:214–218. [PubMed: 23770567]
23. Li J, et al. Nat Meth. Oct 15.2013 10.1038/nmeth.2650

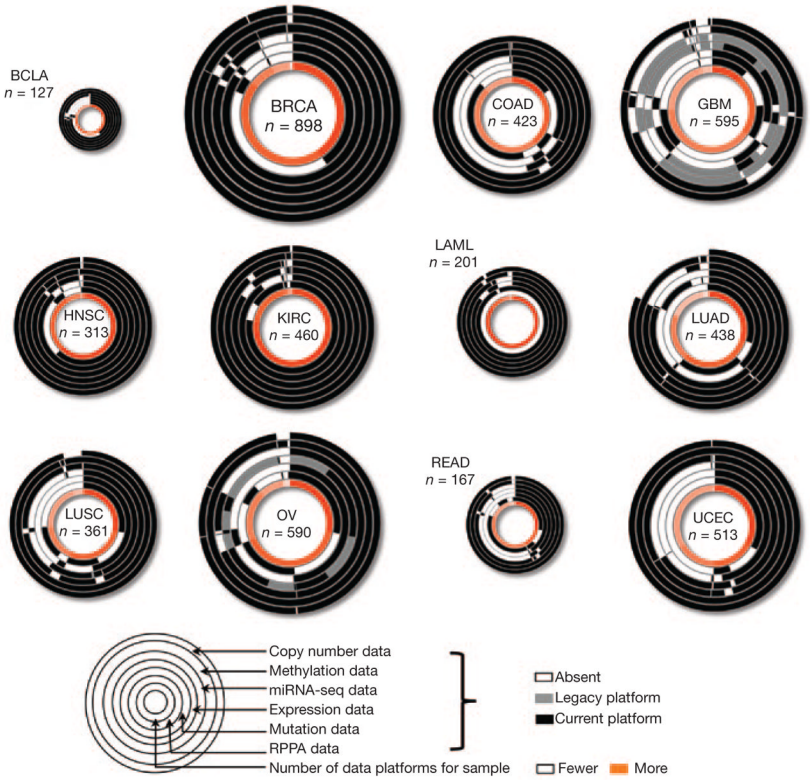

Figure 1.

Molecular profiling data sets in the Pan-Cancer project. Each circular plot displays the total number of samples analyzed across each of the 12 tumor types in the Pan-Cancer project. Samples are arranged in the same order in each concentric circle for each tumor type. Different circles are colored according to whether the sample was profiled using the most current platform, was profiled using a legacy platform or was not profiled. Each data set, including older versions, is available in Synapse (syn300013).

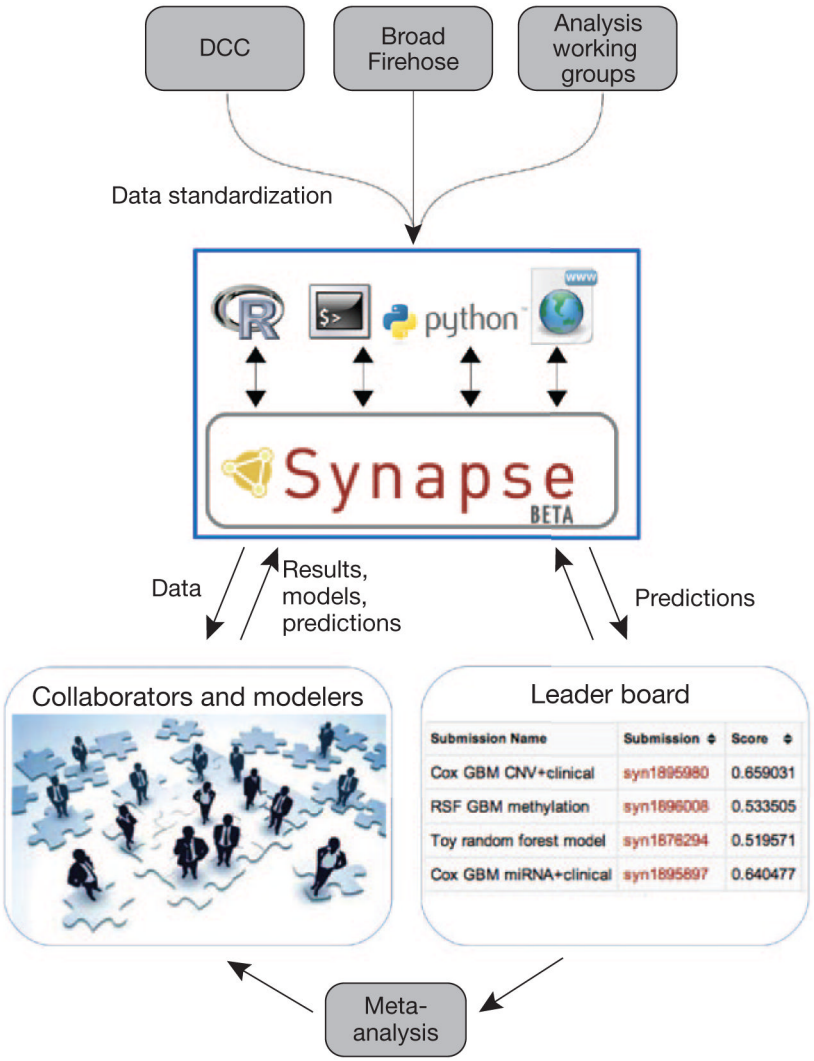

Figure 2.

Schematic of the Pan-Cancer analysis workflow. Data were aggregated and standardized from the TCGA DCC, Broad Firehose and individual analysis working groups and processed into easy-to-use tab-delimited files. Collaborators used a variety of analytic tools, such as R, Python, Unix shell and the web client, to interact with data in Synapse while also storing results, provenance records, analysis descriptions and source code. For a subset of these results (for example, patient survival predictions), Synapse carried out automated performance evaluations and displayed results on a real-time leader board, which were available to collaborators to perform comparative meta-analysis or adapt model source code to additional applications.

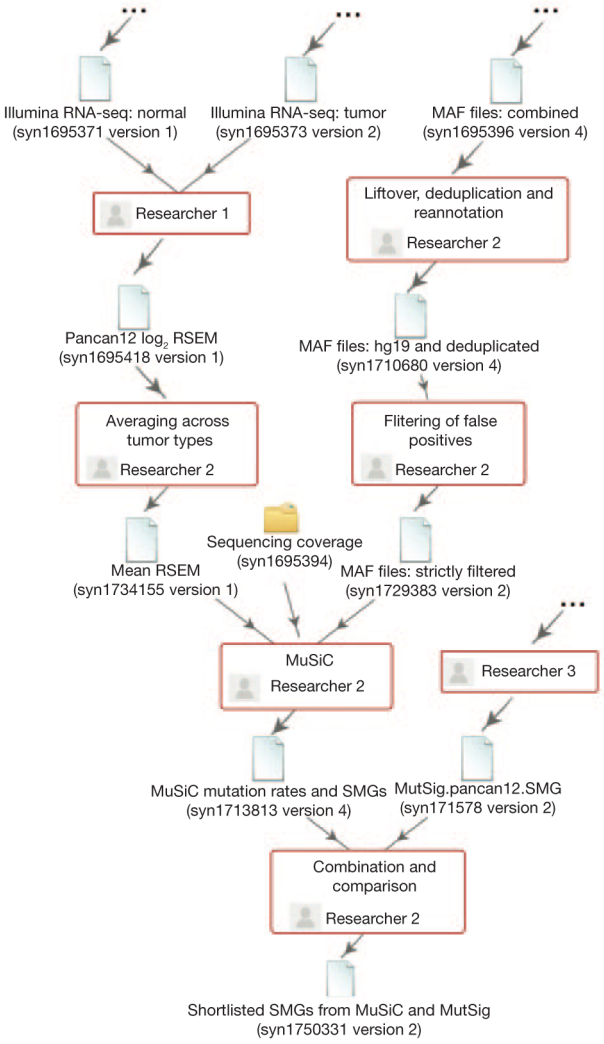

Figure 3.

Example provenance graph of a multistep workflow showing interaction between the analysis of three researchers. The provenance record consists of two types of nodes—activities (shown as red boxes above) performed by a researcher and input and output files of these actions (shown as file and folder icons and identified by their name and Synapse ID). In addition, every activity has metadata associated with it to further describe the details of the actions performed. This specific graph shows the workflow used to perform comparative analysis of two mutation-calling algorithms—MuSiC and MutSig. For MuSiC, the provenance of analysis is displayed from input data to derivation of mutation calls. Provenance records may be further expanded (ellipses) to trace the origin of input files to their original data source in Firehose, DCC or personal communications with AWG members. For brevity, the MutSig graph is not expanded. This graph was produced from version 2 of the data in doi:10.7303/ syn1750331.

Table 1

Publications and Synapse DOIs containing output results

Project title	Synapse DOIs
Dissecting the clinical prognostic and predictive utility of cancer genomic data across tumor types (Y.Y., E.M. Van Allen, L.O., N. Wagle, A. Sokolov *et al.*, unpublished data)	doi:10.7303/syn1710282
Pan-cancer patterns of somatic copy number alteration[21]	doi:10.7303/syn1703335
Analysis of somatic mutations across many tumor types[22]	doi:10.7303/syn1715784
Integrated genomics analysis of 12 tumor types reveals new cell-of-origin cancer subtypes (J.S. *et al.*, unpublished data)	doi:10.7303/syn1868708[a]
1,000 tumor-normal pairs across 5 cancers by whole-genome sequencing	doi:10.7303/syn1709899[a]
Identification of pan-cancer oncogenic miRNA superfamily anchored by a central core seed motif	doi:10.7303/syn1703131
Estimating the presence of tumor-associated normal cells using gene expression signatures predicts tumor purity	doi:10.7303/syn1809223[a], doi:10.7303/syn1901044
The Cancer Proteome Atlas: a resource for cancer proteomic data[23]	doi:10.7303/syn1750330
Multi-cancer molecular signatures and their interrelationships	doi:10.7303/syn1686966

[a]The data for these projects will be made public when the projects publish. Requests for access should be made to the corresponding author.