



SageBionetworks

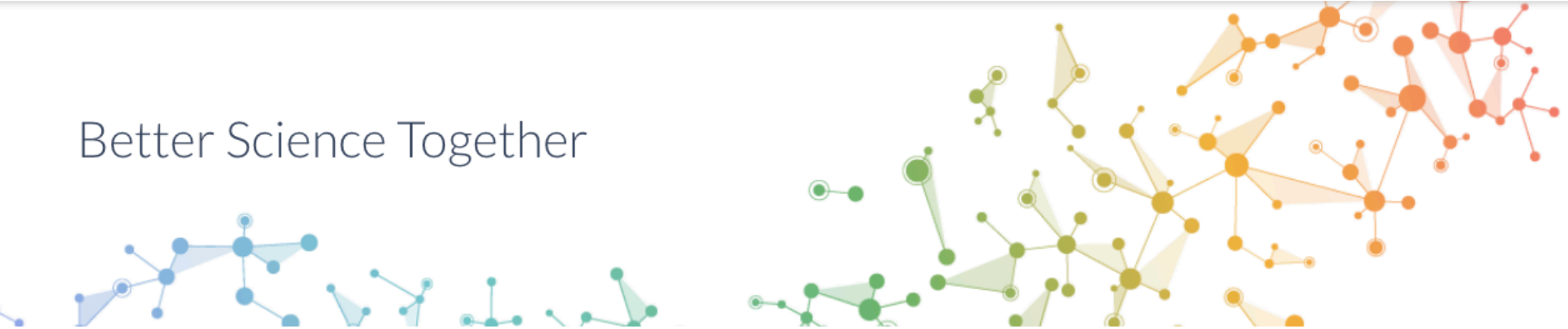
Community management and metadata

Sara Gosline, James Eddy
May 15th, 2018

What is Sage?

A non-profit research institute focused on finding ways to increase *collaboration* and *communication* in science

Better Science Together



Ongoing efforts to support collaborative science

- Mobile apps to connect patients within disease communities
- DREAM challenges to incentivize computational scientists to study novel biological datasets
- Work with granting agencies/foundations to support data sharing (and ultimately community building) across funded scientists

Computational Oncology team

Research scientists:

Sara Gosline, PhD, McGill
systems biology

James Eddy, PhD, UIUC
systems biology

Julie Bletz, PhD, Stanford
genetics

Brian White, PhD, Wash U
computational biology

Kristen Dang, PhD, UNC
computational biology & bioinformatics

Michael Mason, PhD, UCLA
statistics

Robert Allaway, PhD, Dartmouth
RAS signaling

Group leader:

Justin Guinney, PhD, Duke
computational biology & bioinformatics

Research associates:

Thomas Yu, UW

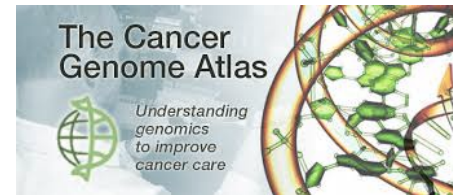
Xindi Guo, UW

Andrew Lamb, Northeastern

Sage supports diverse communities across oncology



PARKER INSTITUTE
for CANCER IMMUNOTHERAPY



... and many others

Some communities have unique metadata needs



PARKER INSTITUTE
for CANCER IMMUNOTHERAPY



CANCER SYSTEMS
BIOLOGY CONSORTIUM

PHYSICAL SCIENCES
in ONCOLOGY

The Cancer
Genome Atlas

Understanding
genomics
to improve
cancer care



AACR
American Association
for Cancer Research
FINDING CURES TOGETHER™

PROJECTGENIE
Genomics Evidence Neoplasia Information Exchange



What is Project GENIE?

- A collaboration of the AACR and cancer centers around the world.
- A data registry of clinical cancer sequencing results accompanied by clinical data.

By the numbers...

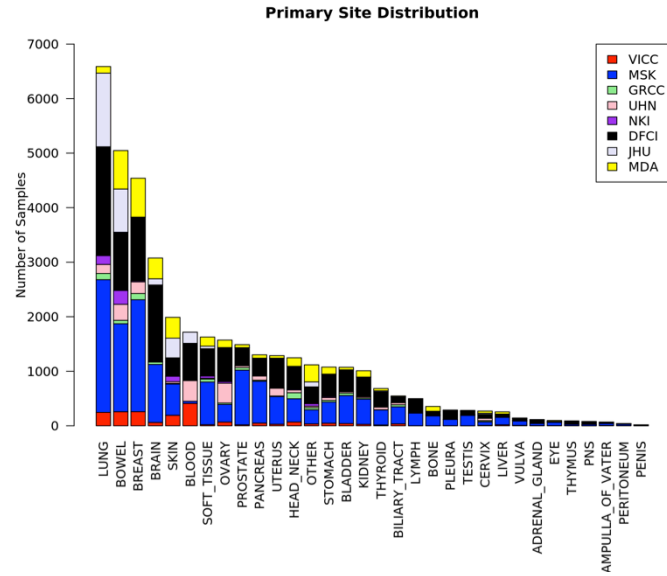
GENIE public release 3.0

39,600 samples

21 sequencing assays

8 cancer centers

15 primary sites with > 1000
samples



Sage serves as data harmonization intermediate

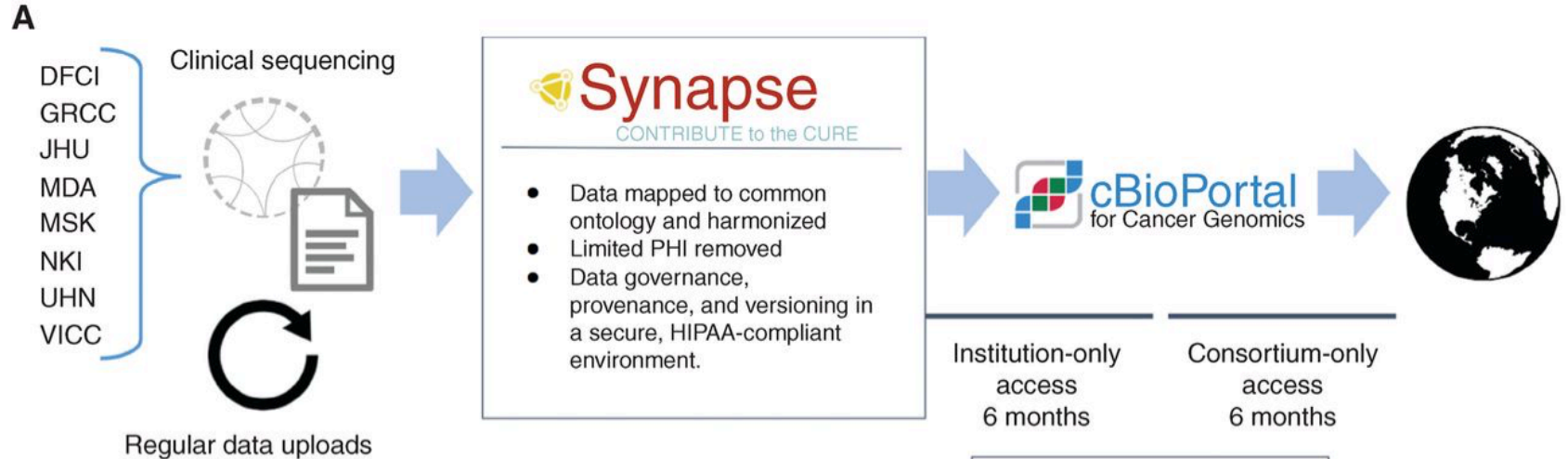


Figure 1. GENIE at a glance.
The AACR Project GENIE Consortium. Cancer Discovery. 2017

GENIE captures *clinical* and *genomic* data

Sample clinical

Sample ID
Oncotree code
Sample type
Seq assay ID
Age at seq report
Patient ID
Center
Seq Quarter

Patient clinical

Patient ID
Sex
Primary race
Secondary race
Tertiary race
Ethnicity
Center
Birth year

Genomic

Position
HUGO Symbol
Seq assay ID
Variant type
Sample ID
+ VEP annotations

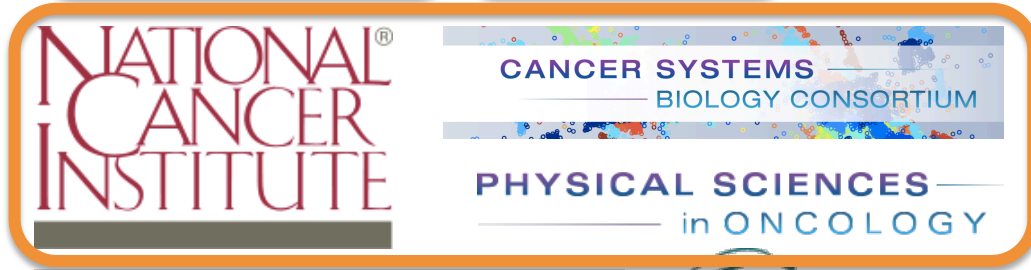
GENIE's clinical data committee establishes data elements and agrees on definitions across all consortium members. External standards used:

- NAACCR (in use)
- NCI Thesaurus (under consideration)
- RxNorm (under consideration)

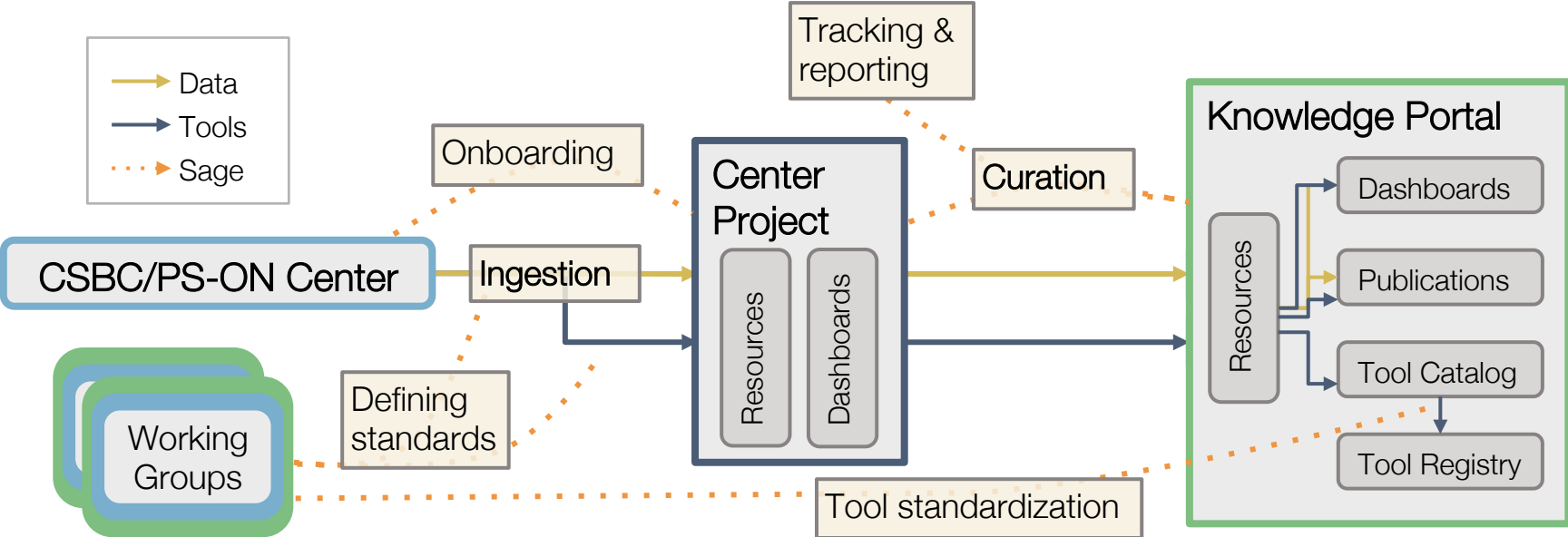
Others communities work with Sage to adopt standards



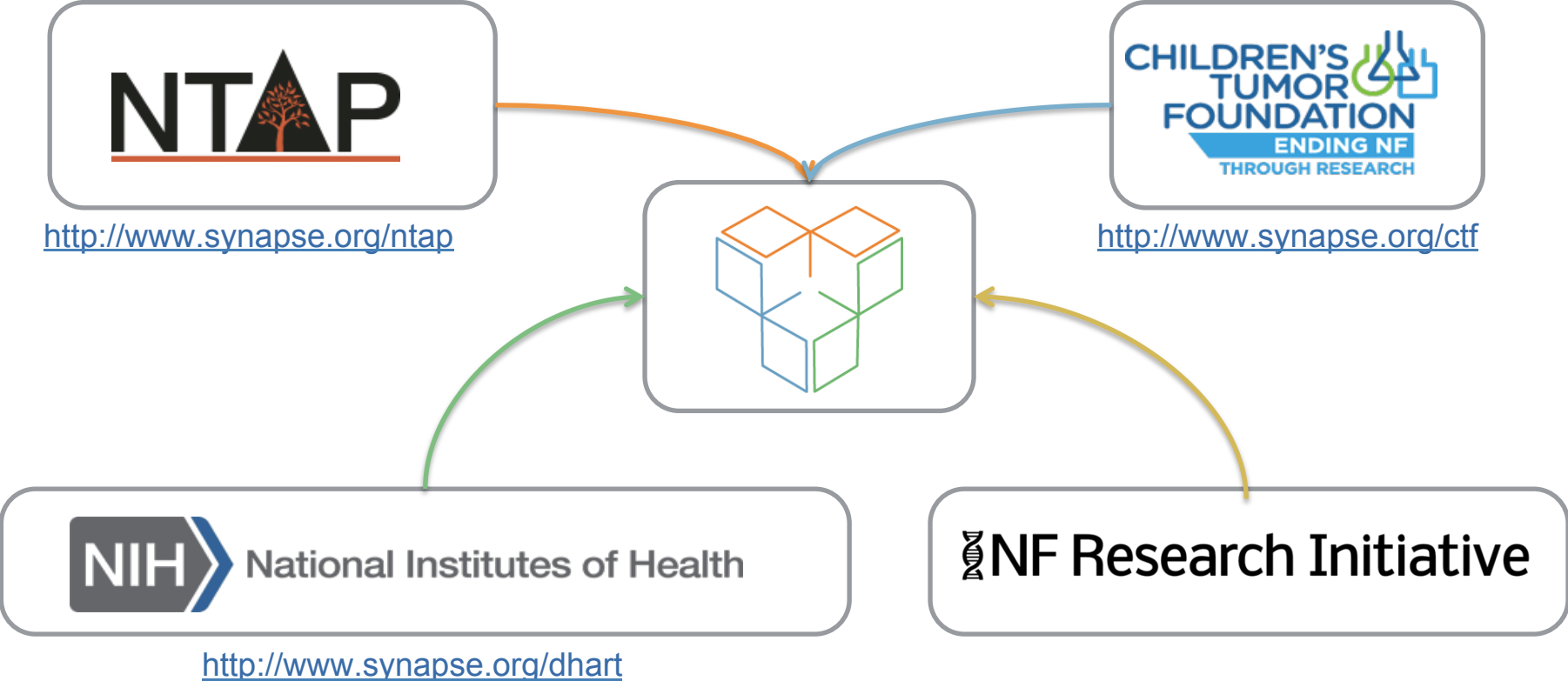
PARKER INSTITUTE
for CANCER IMMUNOTHERAPY



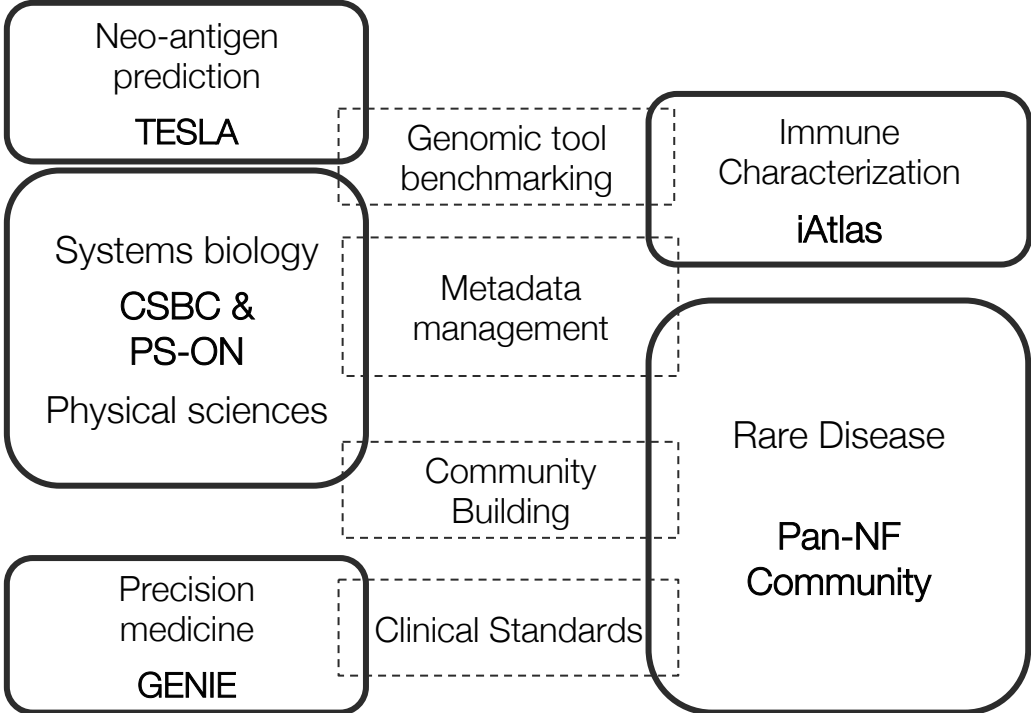
CSBC/PS-ON: thematic organization of publicly shared data



Neurofibromatosis: sharing big data across a small set of patients

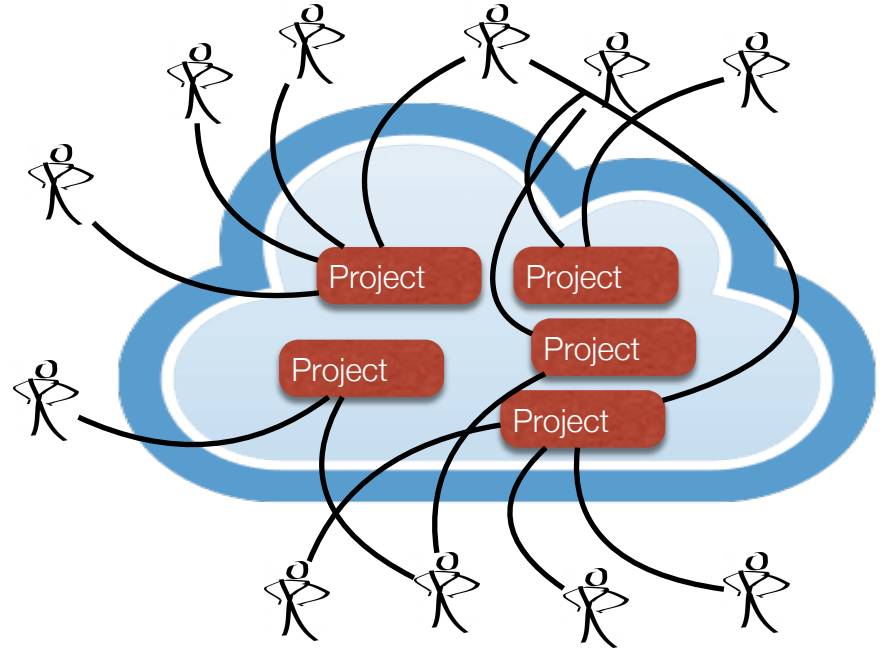


Despite diversity we can solve *common* problems



Projects share common processes & infrastructure: Synapse

- **Typical data ingest**
 - Linked from federated ecosystem site, e.g. GDC, GEO
 - Uploaded to Synapse
- **Resource management in Synapse**
 - Cloud-based resource
 - Centrally located projects
 - Scientists can interact with individual projects



Synapse projects create collaborative space

Center for Modeling Tumor Cell Migration Mechanics ☆

Search [] Sara Gosline (sgosline) ☆ Help []

Synapse ID: syn7349745 Storage Location: Synapse Storage Project Settings [] Tools []

Wiki [?] Files [?] Tables [?] Discussion [?] Docker [?]

Center for Modeling Tumor Cell Migration Mechanics

- Center Investigators
- Research Projects and Core Facilities
- Data and Tools
 - Cell Migration Simulator (CMS)
 - Controlling the Cellular Microenvironment
 - Sleeping Beauty (SB)
 - Transposon

Edit Order [] <<

Physical Sciences Oncology Center
Exploring new Therapies to Inhibit Cancer Cell Migration

The PSOON U54 Research Center @ University of Minnesota

Center Title
Center for Modeling Tumor Cell Migration Mechanics

Overall Project Title
Simulating Tumor Cell Migration

Center Website
<http://psoc.umn.edu/>

Wiki describes project

File browser

Structured data

Discussion

Synapse enables flexible data ingestion & governance

Upload or Link to File

Upload File [Link to URL](#)

URL

Name (Optional)

Sharing Settings **Public**
Everyone can view content in this folder.

Conditions For Use None
Use of the content of this folder does not require agreement to additional terms.

Cancel Save

Upload or Link to File

Upload File [Link to URL](#)

All uploaded files will be stored in **Synapse** storage

Drop files to upload, or [Browse...](#)

Sharing Settings **Public**
Everyone can view content in this folder.

Conditions For Use None
Use of the content of this folder does not require agreement to additional terms.

Cancel Save

Project Settings **Tools**

- Folder Sharing Settings
- Rename Folder
- Annotations
- Upload or Link to a File
- Add New Folder
- Change Folder Storage Location
- Edit Folder Wiki
- Move Folder
- Create DOI for Folder
- Save Link to Folder
- Delete Folder

File Sharing Settings

The sharing settings shown below are currently being inherited from **CSBC PS-ON Knowledge Portal** and cannot be modified here.

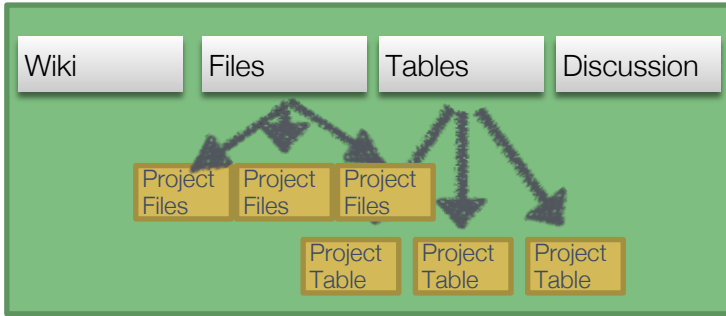
| People | Access |
|------------------------------|---------------|
| Anyone on the web | Can view |
| All registered Synapse users | Can download |
| CSBC PS-ON Administrators | Administrator |
| Andrea Bild (andrea) | Can download |
| Trey Ideker (treideker) | Can download |
| Jonathan Licht (jlicht01) | Can download |
| David Odde (dodde) | Can download |

By default the sharing settings are inherited from the parent folder or project. If you want to have different settings on a specific file, folder, or table you need to create local sharing settings and then modify them.

[+ Create Local Sharing Settings](#)

Cancel Save

Synapse provides protected way to store & share data

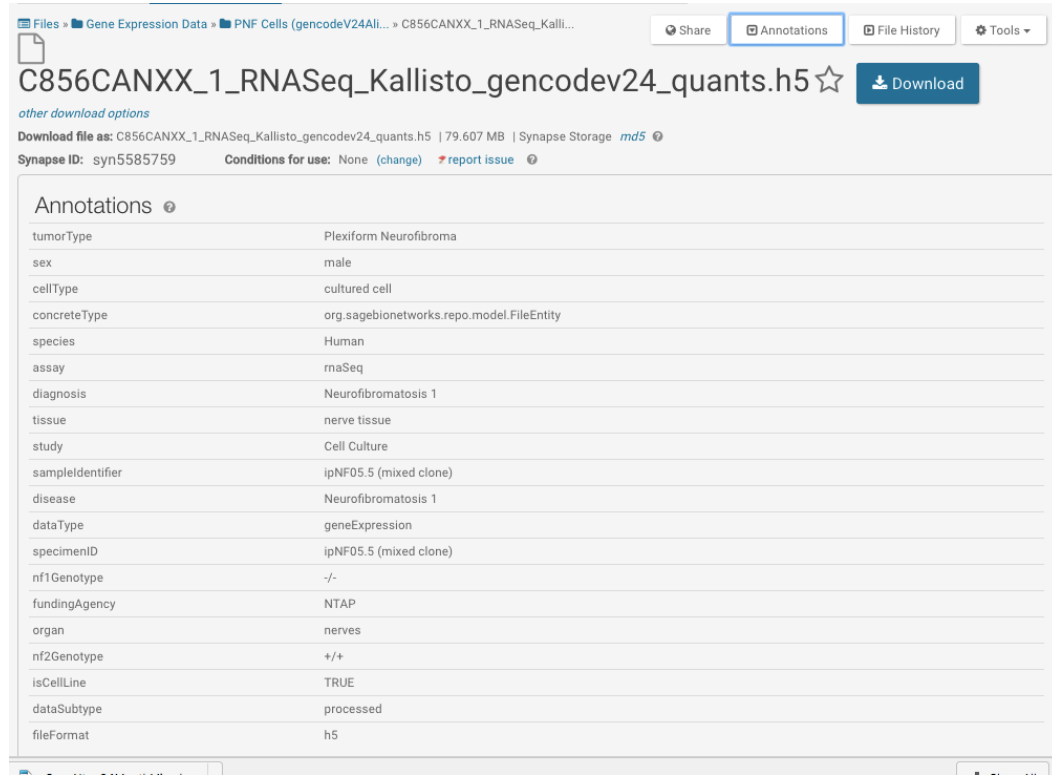


Individual projects allow individual contributors to:

- Upload/annotate files ahead of release
- Analyze files
- Share individual files or results with collaborators
- Contextualize data with wiki figures and text

Every file gets assigned distinct metadata

- Applied to each file
- Standardized descriptions of what is in the file
- Key-value pairs
- Same values used across *all* projects at Sage to facilitate cross-dataset analysis

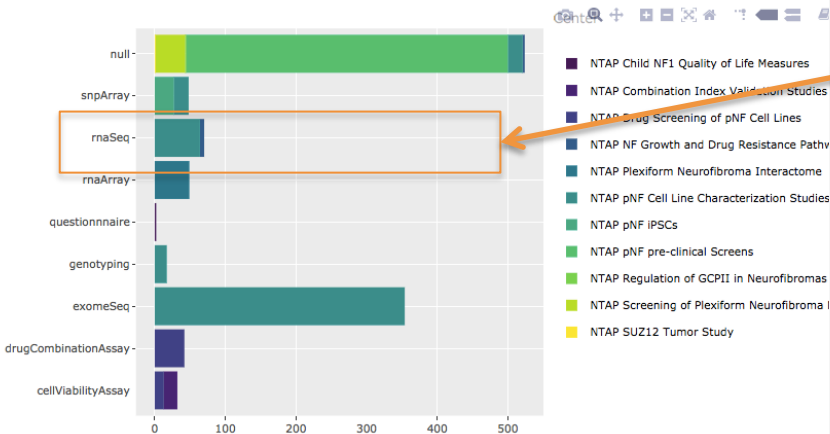


The screenshot shows a file page in the Sage Bionetworks data browser. The file name is `C856CANXX_1_RNASeq_Kallisto_gencodev24_quants.h5`. The file size is 79.607 MB. The Synapse ID is `syn5585759`. The conditions for use are set to "None". There is a "Download" button and a "report issue" link.

Annotations

| | |
|------------------|---|
| tumorType | Plexiform Neurofibroma |
| sex | male |
| cellType | cultured cell |
| concreteType | org.sagebionetworks.repo.model.FileEntity |
| species | Human |
| assay | rnaSeq |
| diagnosis | Neurofibromatosis 1 |
| tissue | nerve tissue |
| study | Cell Culture |
| sampleIdentifier | ipNF05.5 (mixed clone) |
| disease | Neurofibromatosis 1 |
| dataType | geneExpression |
| specimenID | ipNF05.5 (mixed clone) |
| n1Genotype | -/- |
| fundingAgency | NTAP |
| organ | nerves |
| n2Genotype | +/+ |
| isCellLine | TRUE |
| dataSubtype | processed |
| fileFormat | h5 |

These metadata enable summaries across projects



Files > Gene Expression Data > PNF Cells (gencodeV24Alli... > C856CANXX_1_RNASeq_Kalli...

Share Annotations File History Tools

C856CANXX_1_RNASeq_Kallisto_gencodev24_quants.h5

Download

other download options

Download file as: C856CANXX_1_RNASeq_Kallisto_gencodev24_quants.h5 | 79.607 MB | Synapse Storage md5

Synapse ID: syn5585759 Conditions for use: None (change) report issue

Annotations

| | |
|------------------|---|
| tumorType | Plexiform Neurofibroma |
| sex | male |
| cellType | cultured cell |
| concreteType | org.sagebionetworks.repo.model.FileEntity |
| species | Human |
| assay | rnaSeq |
| diagnosis | Neurofibromatosis 1 |
| tissue | nerve tissue |
| study | Cell Culture |
| sampleIdentifier | ipNF05.5 (mixed clone) |
| disease | Neurofibromatosis 1 |
| dataType | geneExpression |
| specimenID | ipNF05.5 (mixed clone) |
| n1Genotype | -/- |
| fundingAgency | NTAP |
| organ | nerves |
| n2Genotype | +/+ |
| isCellLine | TRUE |
| dataSubtype | processed |
| fileFormat | h5 |

Diverse metadata needs require collective efforts

- Standardized terms and vocabulary for community resources
 - Pulled from known ontologies
 - Updated with missing terms
- Tooling & processes to ascribe resource annotations
- Need to provide standards across Sage-led communities but also flexibility

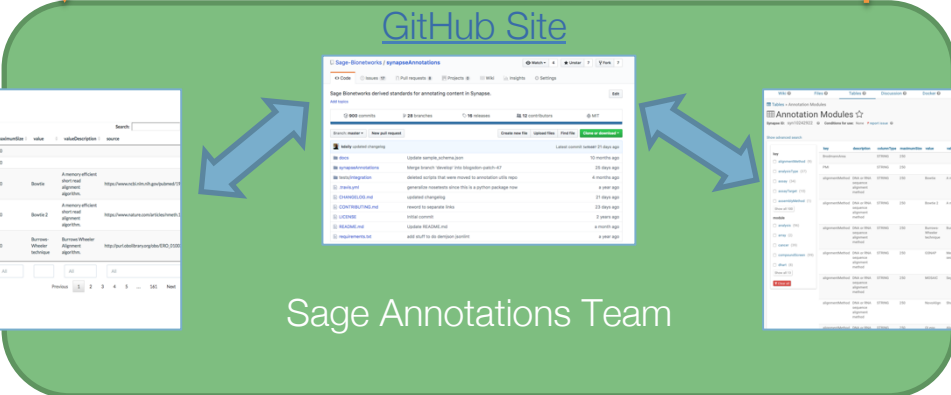
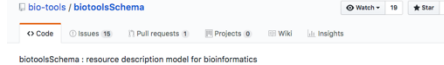
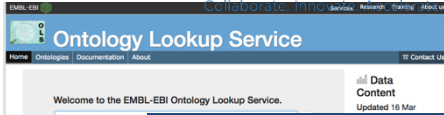
Organically evolving metadata and SOPs



Define Use Cases

Users, Working Groups

Search Standards



Annotation Browser

Synapse Table

Reviewed by Sage-wide team via GitHub

Sage-Bionetworks / **synapseAnnotations**

Code Issues 19 Pull requests 5 ZenHub Projects 0

Sage Bionetworks derived standards for annotating content in Synapse.
Add topics

900 commits 30 branches 16 releases

Branch: master New pull request

kdaily updated changelog

- docs Update sample_schema.json
- synapseAnnotations Merge branch 'develop' into blogsdon-patch-47
- tests/integration deleted scripts that were moved to annotation util
- .travis.yml generalize nosetests since this is a python package
- CHANGELOG.md updated changelog
- CONTRIBUTING.md reword to separate links
- LICENSE Initial commit
- README.md Update README.md
- requirements.txt add stuff to do demjson jsonlint

Filters is:issue is:open Labels Milestones New issue

19 Open 88 Closed Author Labels Projects Milestones Assignee Sort

- Need to find solution to cancer subtype #366 opened 22 days ago by sgosline 4
- Add multi-sample/individual SOP to contributing.md #365 opened on Mar 21 by sgosline New Issues
- Get serious about tool/model annotations #364 opened on Mar 20 by jaeddy 2
- Mouse Strain key question #362 opened on Mar 9 by amapeters New Issues
- Think more and come to some conclusion about JSON-LD #351 opened on Feb 27 by kdaily 2018-05-08 8
- Discussion about mouse behavioral analysis annotations #341 opened on Feb 21 by ychae 2018-02-28 3
- should keys have sources? question #301 opened on Dec 12, 2017 by kdaily New Issues
- Recommended annotations for QC files help wanted #292 opened on Nov 28, 2017 by xindiguo New Issues
- Synonyms/aliases for annotation values enhancement #270 opened on Oct 24, 2017 by jaeddy New Issues
- add cellType: monocyte derived microglia AMP-AD sprint create value #260 opened on Oct 19, 2017 by ychae New Issues
- add cellType: iPSC-derived neurons AMP-AD sprint create value #258 opened on Oct 19, 2017 by ychae New Issues

<https://github.com/Sage-Bionetworks/synapseAnnotations>

Metadata dictionary is always evolving

- Want to get as many terms needed for analysis, but no burden researchers
- Adding weekly to satisfy specific use cases

Challenges

Synapse-based limitations

- Metadata requires key-value pairs, no hierarchy
 - Metadata are tied to file
 - Very flexible, can annotate with *anything*
-
- We're working on these things!

Diverse requirements within and across projects

- Most projects are not *data generation* projects, require on-the-fly identification of metadata terms
- Some projects require more depth than others, for example:
 - Is this cancer or not?
 - What is the organ of origin?
 - What subtype of cancer?
- Not just data:
 - Tools
 - Analysis

In the community: lack of single standard

- Ontologies are limited to a single domain
 - Defining disease and clinical parameters
 - Defining computational analysis or experimental protocols
 - Defining tools
- We like to balance using existing terms with what is provided
 - For example: ChIP-Seq can be applied to transcription factor or histone mark, do we:
 - Create TF ChIP-Seq and histone ChIP-Seq?
 - Create a new term called assayTarget and define there?

Dearth of tools in the field

- How can we validate that metadata is correct?
- Schema are still very hard to visualize, select in piecemeal
- Semi-automated metadata assignment
- Mapping between ontologies

Summary

- Sage has a diverse set of projects/communities
- We work to standardize metadata and tools across these communities
- Adhere to standards whenever possible

Questions?

Extra slides

Select metadata dictionary

- Online tool enables download of manifest or JSON file

Annotation UI v8.0.0

Annotation Modules

- neuro
- analysis
- experimentalData
- genie
- dhart
- cancer
- array
- compoundScreen
- sageCommunity
- toolExtended
- tool
- ngs
- neurofibromatosis
- All/None

Upload Your Annotation's Module

Module Name

Module CSV File

No file selected

| key | description | columnType | maximumSize | value | valueDescription | source |
|-------|---|------------|-------------|----------|---|---|
| assay | The technology used to generate the data in this file | STRING | 250 | ATACSeq | Open chromatin regions measured by sequencing DNA after assay for transposase-accessible chromatin (ATAC) treatment | http://purl.obolibrary.org/obo/OBI_0002039 |
| assay | The technology used to generate the data in this file | STRING | 250 | ChIPSeq | Chromatin immunoprecipitation followed by sequencing | http://purl.obolibrary.org/obo/OBI_0000716 |
| assay | The technology used to generate the data in this file | STRING | 250 | FIA-MSMS | Flow injection analysis - tandem mass spectrometer | https://www.ncbi.nlm.nih.gov/pubmed/28667829 |
| assay | The technology used to generate the data in this file | STRING | 250 | Hi-C | Chromatin interactions detected by Hi-C protocol | http://www.ebi.ac.uk/efo/EFO_0007693 |
| assay | The technology used to generate the data in this file | STRING | 250 | ISOSeq | Full isoform sequencing | |
| assay | The technology used to generate the data in this file | STRING | 250 | LC-MS | A method where a sample mixture is first separated by liquid chromatography before being converted into ions which are characterised by their mass-to-charge ratio and relative abundance | http://purl.obolibrary.org/obo/CHMO_0000524 |

Select metadata dictionary

- Select module of interest
 - experimentalData
 - cancer
 - sageCommunity
 - ngs

Annotation UI v8.0.0

Annotation Modules

- neuro
- analysis
- experimentalData
- genie
- dhart
- cancer
- array
- compoundScreen
- sageCommunity
- toolExtended
- tool
- ngs
- neurofibromatosis
- All/None

Select metadata dictionary

- Learn about keys/values

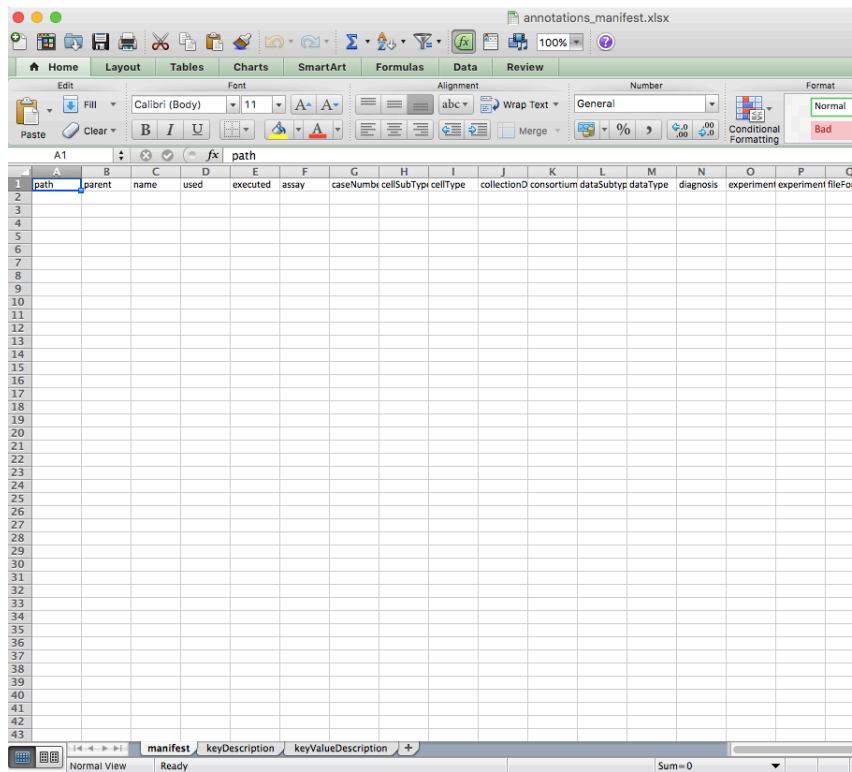
Table Key Description Value Description

Show 50 entries Search:

| key | description | columnType | maximumSize | value | valueDescription | source |
|-------|---|------------|-------------|---------|---|---|
| assay | The technology used to generate the data in this file | STRING | 250 | ATACseq | Open chromatin regions measured by sequencing DNA after assay for transposase-accessible chromatin (ATAC) treatment | http://purl.obolibrary.org/obo/OBI_0002039 |
| assay | The technology used to generate the data in this file | STRING | 250 | ChIPseq | Chromatin immunoprecipitation followed by sequencing | http://purl.obolibrary.org/obo/OBI_0000716 |

Download terms as manifest, fill out

- Identify files of interest
- Browse keys/values
- Upload to file view programmatically or via web UI

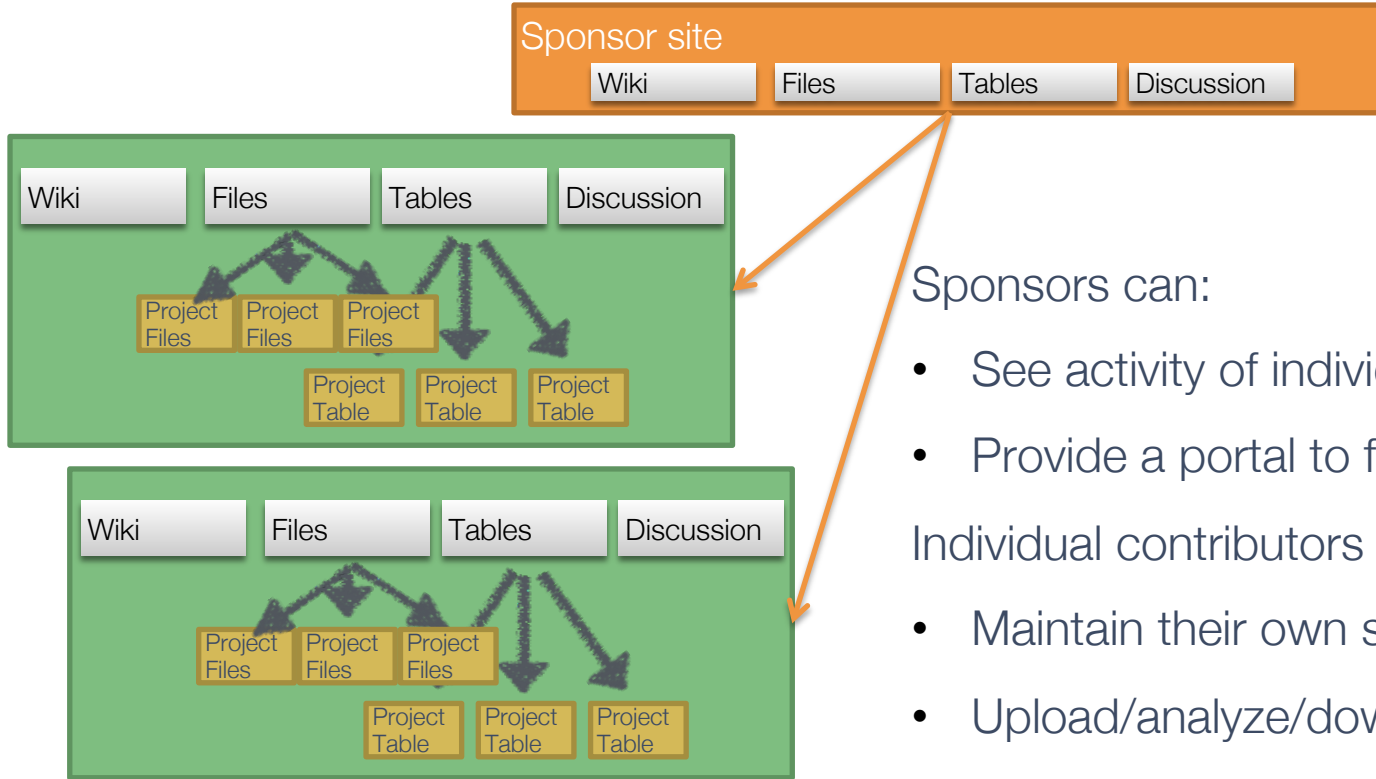


Can be applied via command line or file view

- Results are viewed on web
- Can be edited/updated

| diagnosis | egfrStatus | fileFormat | fundingAgency | id1Status | individualID | isMultiIndividual | isMultiSpecimen | location | organ | platform | resourceType | sex | species | specimenID | study | tissue | tumorType | geoAccession | librarySource | molecule |
|---------------|------------|------------|---------------|-----------|--------------|-------------------|-----------------|----------|--------|------------|------------------|-------|---------|------------|---------------|---------------|-----------|--------------|----------------|-----------|
| Brain Cancer | | csv | | | | False | False | brain | | NextSeq500 | experimentalData | | Human | | | | | GSE84465 | | |
| Brain Cancer | | csv | NIH-NCI | | | True | True | brain | | NextSeq500 | experimentalData | | | | scRNA | Glioblastoma | | GSE84465 | transcriptomic | polyA RNA |
| Breast Cancer | | html | | | | | | | breast | | | | | | | | | | | |
| Breast Cancer | | SRA | | | BC11 | False | False | | | HiSeq2500 | experimentalData | Human | BC11_88 | scRNA | primary tumor | breast cancer | | GSM2392141 | transcriptomic | total RNA |
| Breast Cancer | | SRA | | | BC11 | False | False | | | HiSeq2500 | experimentalData | Human | BC11_81 | scRNA | primary tumor | breast cancer | | GSM2392140 | transcriptomic | total RNA |
| Breast Cancer | | SRA | | | BC11 | False | False | | | HiSeq2500 | experimentalData | Human | BC11_78 | scRNA | primary tumor | breast cancer | | GSM2392139 | transcriptomic | total RNA |
| Breast Cancer | | SRA | | | BC11 | False | False | | | HiSeq2500 | experimentalData | Human | BC11_70 | scRNA | primary tumor | breast cancer | | GSM2392138 | transcriptomic | total RNA |
| Breast Cancer | | SRA | | | BC11 | False | False | | | HiSeq2500 | experimentalData | Human | BC11_69 | scRNA | primary tumor | breast cancer | | GSM2392137 | transcriptomic | total RNA |
| Breast Cancer | | SRA | | | BC11 | False | False | | | HiSeq2500 | experimentalData | Human | BC11_56 | scRNA | primary tumor | breast cancer | | GSM2392136 | transcriptomic | total RNA |

Summaries can be aggregated across projects workspaces



Sponsors can:

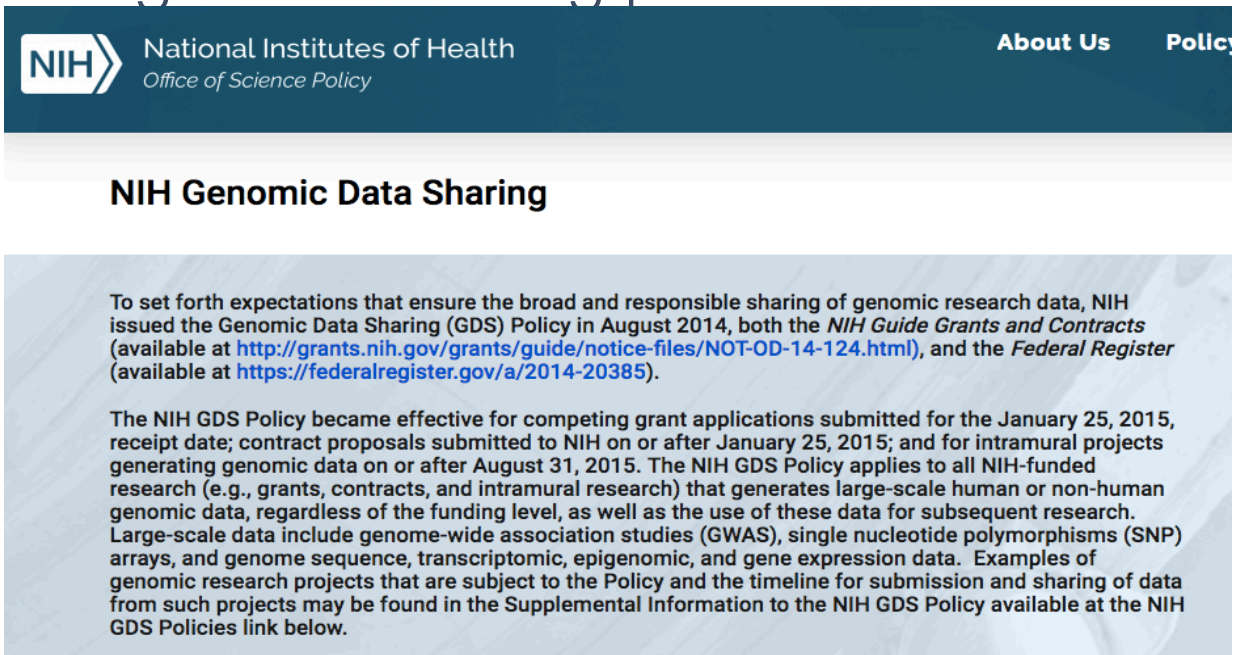
- See activity of individual projects
- Provide a portal to funded research

Individual contributors can:

- Maintain their own site
- Upload/analyze/download files as needed

Building data repositories: a way to encourage data sharing

- Journals and the NIH are more actively monitoring - and enforcing - data sharing policies.





The image is a screenshot of the NIH Genomic Data Sharing page. At the top, there is a dark blue header with the NIH logo on the left, the text "National Institutes of Health" and "Office of Science Policy" in the center, and "About Us" and "Policy" on the right. Below the header is a white section with the title "NIH Genomic Data Sharing" in bold. The main content area has a light blue background with a grid pattern. It contains two paragraphs of text. The first paragraph states that NIH issued the Genomic Data Sharing (GDS) Policy in August 2014, both the *NIH Guide Grants and Contracts* (available at <http://grants.nih.gov/grants/guide/notice-files/NOT-OD-14-124.html>), and the *Federal Register* (available at <https://federalregister.gov/a/2014-20385>).

The NIH GDS Policy became effective for competing grant applications submitted for the January 25, 2015, receipt date; contract proposals submitted to NIH on or after January 25, 2015; and for intramural projects generating genomic data on or after August 31, 2015. The NIH GDS Policy applies to all NIH-funded research (e.g., grants, contracts, and intramural research) that generates large-scale human or non-human genomic data, regardless of the funding level, as well as the use of these data for subsequent research. Large-scale data include genome-wide association studies (GWAS), single nucleotide polymorphisms (SNP) arrays, and genome sequence, transcriptomic, epigenomic, and gene expression data. Examples of genomic research projects that are subject to the Policy and the timeline for submission and sharing of data from such projects may be found in the Supplemental Information to the NIH GDS Policy available at the NIH GDS Policies link below.

New terms are gleaned from existing ontologies


The screenshot shows the EMBL-EBI Ontology Lookup Service (OLS) website. The header includes the EMBL-EBI logo and navigation links for Services, Research, Training, and About us. Below the header is the OLS logo and the main title 'Ontology Lookup Service'. A secondary navigation bar contains links for Home, Ontologies, Documentation, About, and Contact Us. The main content area features a search bar with the placeholder text 'Search OLS...' and a search icon. Below the search bar, there are examples: 'diabetes, GO:0098743' and a link 'Looking for a particular ontology?'. To the right, there is a 'Data Content' section with a bar chart icon, updated on '25 Apr 2018 03:03', and a list of statistics: 207 ontologies, 5,376,937 terms, 17,811 properties, and 479,385 individuals. Below this is a 'Tweets by @EBIOLS' section showing a tweet from EBISPOT OLS (@EBIOLS) dated Dec 18, 2017, about the integration of the Oxo service into OLS. At the bottom of the main content area, there are three columns: 'About OLS' (describing the service as a repository for biomedical ontologies), 'Related Tools' (listing Oxo, Zooma, and Webulous), and 'Contact Us' (providing contact information for support and reporting issues).

EMBL-EBI  Services Research Training About us

 **Ontology Lookup Service**

Home Ontologies Documentation About Contact Us

Welcome to the EMBL-EBI Ontology Lookup Service.

Search OLS... 

Examples: [diabetes](#), [GO:0098743](#) [Looking for a particular ontology?](#)

About OLS

The Ontology Lookup Service (OLS) is a repository for biomedical ontologies that aims to provide a single point of access to the latest ontology versions. You can browse the ontologies through the website as well as programmatically via the OLS API. OLS is developed and maintained by the [Samples, Phenotypes and Ontologies Team](#) (SPOT) at EMBL-EBI.

Related Tools

In addition to OLS the SPOT team also provides the Oxo, Zooma and Webulous services. [Oxo](#) provides cross-ontology mappings between terms from different ontologies. [Zooma](#) is a service to assist in mapping data to ontologies in OLS and [Webulous](#) is a tool for building ontologies from spreadsheets.

Contact Us



For feedback, enquiries or suggestion about OLS or to request a new ontology please contact [ols-support@ebi.ac.uk](#). For bugs or problems with the code or API please report on [GitHub issue](#) For announcements relating to OLS, such as new releases and new features sign up to the [OLS announce mailing list](#)

Data Content



Updated 25 Apr 2018 03:03



- 207 ontologies
- 5,376,937 terms
- 17,811 properties
- 479,385 individuals

Tweets by @EBIOLS

 **EBISPOT OLS** @EBIOLS 

Our new Ontology mapping service (Oxo) will be integrated into OLS in 2018 [ebi.ac.uk/spot/oxo/](#)

  Dec 18, 2017

 **EBISPOT OLS** @EBIOLS 

OLS hits 200 #ontologies!

Sage approach: build community, not just a repository

- Platform is the foundation of a community



Sage approach: build community, not just a repository

- **Platform** is the foundation of a community
- Engaging the **people** in the community



Sage approach: build community, not just a repository

- **Platform** is the foundation of a community
- Engaging the **people** in the community
- Shared **principles** build trust across community

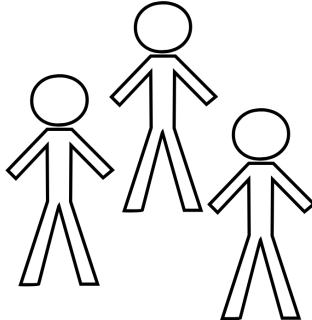


Sage is built on these three pillars

Platform



People



Principles

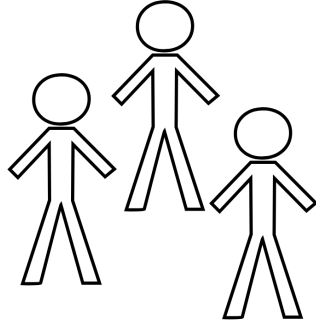


Sage is built on these three pillars

Platform



People



Principles



 Synapse

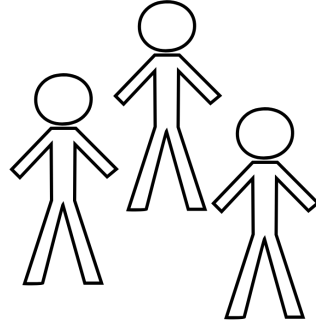
 Bridge

Sage is built on these three pillars

Platform



People



Principles

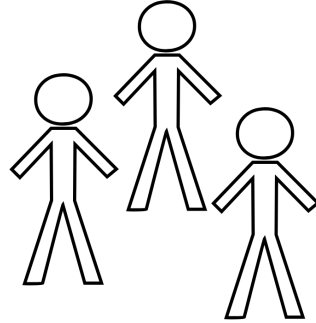


Sage is built on these three pillars

Platform



People



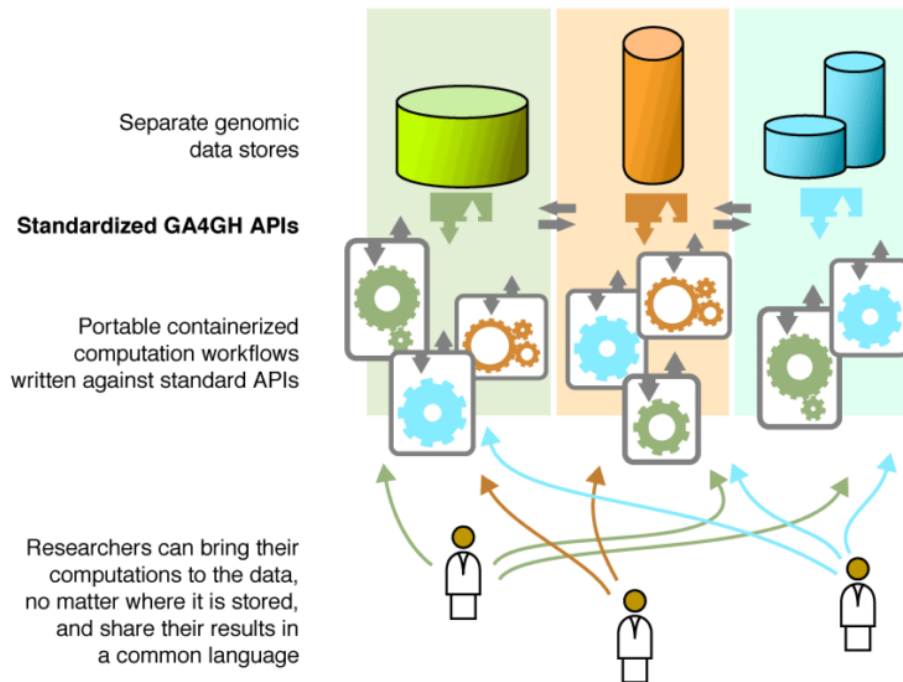
Principles



Findable Interoperable
Accessible Reusable

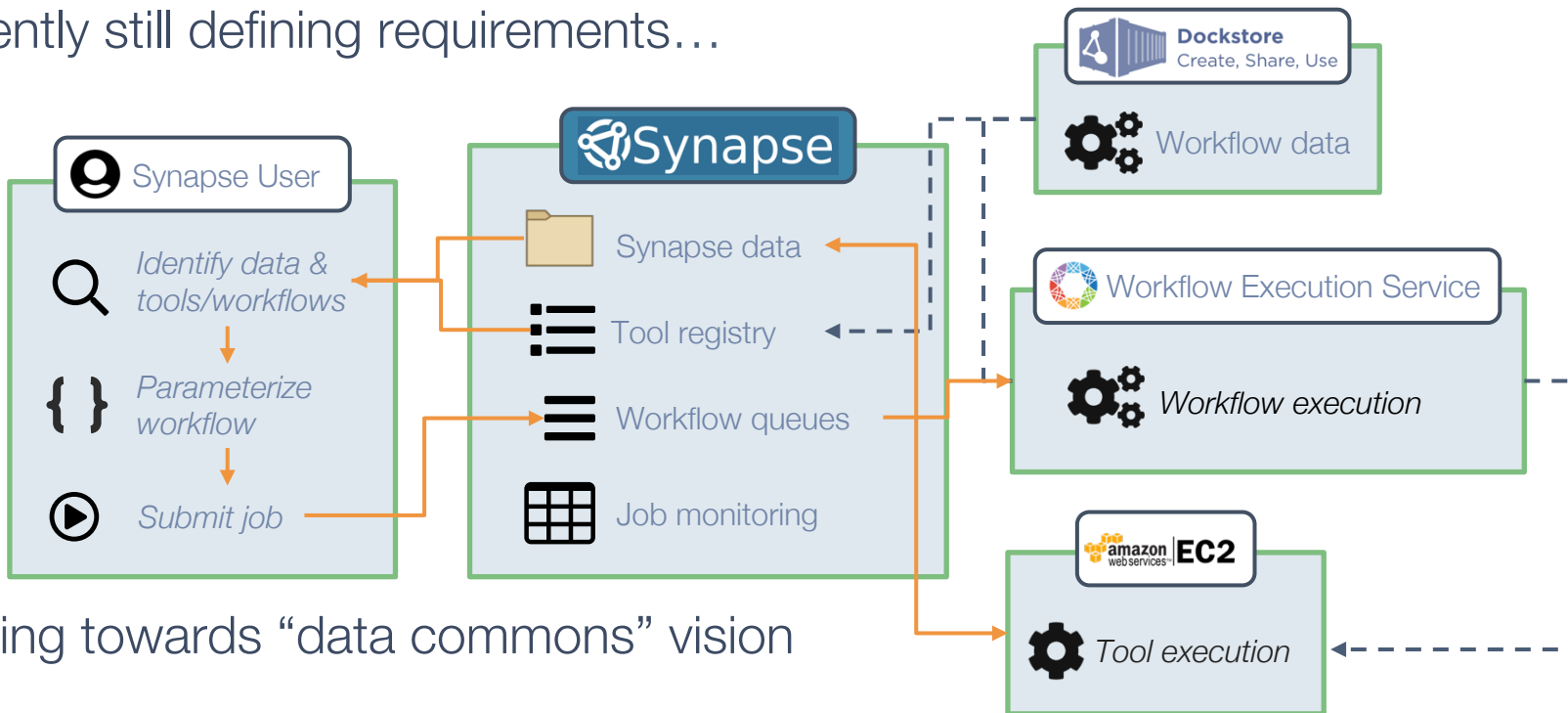
Workflows on Synapse: Emerging Requirements

- Workflow use cases for Sage and Synapse communities:
 - Challenges, benchmarking
 - Data coordination, curation, & validation
 - Mobile data processing
 - Scientific computation
- Solutions and standards developed by communities like GA4GH, BD2K aim to make writing, sharing, and running workflows easier



Reproducible analysis ecosystem through Synapse

- Currently still defining requirements...



- Building towards “data commons” vision



Future: “Communities” and Real-Time Insights

- Vision: Synapse as an interoperable platform/commons
- Native support for concept of a “community” comprised of multiple projects/teams
- **Activity feeds** for project/community insight
- Platform-driven **notifications** of key activity
- **Dashboards** for key community metrics

Share Annotated, Standardized Tools through Unified Portal

Tool Catalog:
retrospective, in development

| methodName | inputDataTypes | outputDataTypes | softwareLanguage |
|---|---|---|------------------|
| Center for Cancer Systems Therapeutics (CaST) | • geneExpression • network • cellularPhysiology | • cellularPhysiology • geneExpression • network | Matlab, R |
| Center for Modeling Tumor Cell Migration Mechanics | • geneExpression • pathway or network | • gene list | Matlab, R |
| Embryonal Brain Tumor Networks | • geneExpression • network | • network | Python |
| Speedy Reimaging to Immunofluorescent Translation of whole slide images using conditional generative adversarial networks | • image | • image | Matlab, Python |

Tools on Synapse:
in development, published

| Center | Files | Software Types | Software Languages | Studies | View Files |
|---|-------|----------------|--------------------|--|------------|
| Center for Cancer Systems Therapeutics (CaST) | 1 | 1 | 1 | • metaViewer | View |
| Center for Modeling Tumor Cell Migration Mechanics | 7 | 1 | 1 | • Bead Track • Flow Track • Model • Compulsion • Cell Tracks • Motor Clutch Model • Cell Migration Simulator | View |
| Embryonal Brain Tumor Networks | 1 | 1 | 1 | • OncoIntegrator | View |
| Modeling and targeting stroma-tumor crosstalk in non small cell lung cancer | 4 | 1 | 1 | • CCCExplorer | View |

Tool Registry:
published, searchable

| Name | Author | Project Links | Docker Pull |
|------|--------------|---------------|-------------------------|
| CaST | Andy Hatcher | GitHub | docker pull castry/cast |
| CaST | Andy Hatcher | GitHub | docker pull castry/cast |
| CaST | Andy Hatcher | GitHub | docker pull castry/cast |
| CaST | Andy Hatcher | GitHub | docker pull castry/cast |
| CaST | Andy Hatcher | GitHub | docker pull castry/cast |
| CaST | Andy Hatcher | GitHub | docker pull castry/cast |
| CaST | Andy Hatcher | GitHub | docker pull castry/cast |
| CaST | Andy Hatcher | GitHub | docker pull castry/cast |
| CaST | Andy Hatcher | GitHub | docker pull castry/cast |
| CaST | Andy Hatcher | GitHub | docker pull castry/cast |

Tool Portal

Future tool catalog will link to all tool resources

Precision Immunology: Tumor Neoantigen Selection Alliance (TESLA)

- Began in 2017 with Parker Institute and CRI
- A benchmarking exercise in predicting immunogenic (neo)epitopes.
- Multiple validation assays (MHC binding, TCR binding by flow and nanoparticles, T cell reactivity)
- >25 teams participating
- Multiple cancer types: melanoma, breast cancer, NSCLC, CRC (MSI-high, MSS)

Sample procurement
and sequencing
(WES, RNA)

Team predictions

Validation assays

Analysis of predictive
ability and features of
immunogenicity



Tool Catalog Collates Submissions from Centers

Individual centers submit data to survey



The screenshot shows a web-based survey form titled "CSBC/PSO Tool Catalog". At the top, there are tabs for "QUESTIONS" and "RESPONSES" with a count of "16". The form includes a header, a description, and three input fields: "Email address" (with a validation message), "Name of method", and "Your Synapse user ID (if registered)". A sidebar on the right contains navigation icons.

QUESTIONS RESPONSES **16**

CSBC/PSO Tool Catalog

This form provides a way to enter specific information about your tool. Please fill out as much as possible so we can document it in the larger CSBC/PSO tool catalog.

Email address *

Valid email address

This form is collecting email addresses. [Change settings](#)

Name of method *

Short answer text

Your Synapse user ID (if registered)

Short answer text

Single Interface Combines All Tools and Provides Links

CSBC Tool Catalog

Input Data Type

- proteomics
- cellularPhysiology
- drugScreen
- geneExpression
- network
- genomicVariants
- isoformExpression
- pathway or network
- chromatinActivity
- gene list
- functional screen
- image
- N/A

Output Data Type

- proteomics
- image
- inputted drug response
- geneExpression
- isoformExpression
- genomicVariants
- network
- cellularPhysiology
- chromatinActivity
- drugScreen
- N/A

Software/Tool Types

Script Package Binary Package Library Web Application Other

| methodName | centerName | inputDataType | outputDataType | softwareLanguage |
|---|--|---|---|------------------|
| CIPM | Cancer Systems Biology Center of HoPE (Heterogeneity of Phenotypic Evolution) | <ul style="list-style-type: none">geneExpressionnetworkcellularPhysiology | <ul style="list-style-type: none">cellularPhysiologygeneExpressionnetwork | Matlab, R |
| Master Regulator Inference Algorithm | Center for Cancer Systems Therapeutics (CaST) | <ul style="list-style-type: none">geneExpressionpathway or network | <ul style="list-style-type: none">gene list | Matlab, R |
| Omics Integrator | Embryonal Brain Tumor Networks | <ul style="list-style-type: none">geneExpression | <ul style="list-style-type: none">network | Python |
| Speedy Histopathological-to-ImmunoFluorescent Translation of whole slide images using conditional generative adversarial networks | N/A | <ul style="list-style-type: none">image | <ul style="list-style-type: none">image | Matlab, Python |

Single Interface Combines All Tools and Provides Links

Filter by
data type

The screenshot displays the CSBC Tool Catalog interface. On the left, a dark sidebar contains a filter menu with two sections: 'Input Data Type' and 'Output Data Type'. Each section lists various data types with a checked checkbox. An arrow points from the text 'Filter by data type' to this sidebar. The main content area is titled 'Software/Tool Types' and features a tabbed interface with 'Script' selected. Below the tabs is a table with columns for 'methodName', 'centerName', 'inputDataType', 'outputDataType', and 'softwareLanguage'. The table lists several tools, including CIPM, Master Regulator Inference Algorithm, Omics Integrator, and Speedy Histopathological-to-ImmunoFluorescent Translation of whole slide images using conditional generative adversarial networks.

| methodName | centerName | inputDataType | outputDataType | softwareLanguage |
|---|---|---|---|------------------|
| CIPM | Cancer Systems Biology Center of HoPE (Heterogeneity of Phenotypic Evolution) | <ul style="list-style-type: none">geneExpressionnetworkcellularPhysiology | <ul style="list-style-type: none">cellularPhysiologygeneExpressionnetwork | Matlab, R |
| Master Regulator Inference Algorithm | Center for Cancer Systems Therapeutics (CaST) | <ul style="list-style-type: none">geneExpressionpathway or network | <ul style="list-style-type: none">gene list | Matlab, R |
| Omics Integrator | Embryonal Brain Tumor Networks | <ul style="list-style-type: none">geneExpression | <ul style="list-style-type: none">network | Python |
| Speedy Histopathological-to-ImmunoFluorescent Translation of whole slide images using conditional generative adversarial networks | N/A | <ul style="list-style-type: none">image | <ul style="list-style-type: none">image | Matlab, Python |

Single Interface Combines All Tools and Provides Links

Broken down by type of tools

Filter by data type

The screenshot shows the CSBC Tool Catalog interface. On the left, there is a sidebar with two sections: 'Input Data Type' and 'Output Data Type'. Each section contains a list of data types with checkboxes, all of which are checked. The 'Input Data Type' list includes: proteomics, cellularPhysiology, drugScreen, geneExpression, network, genomicVariants, isoformExpression, pathway or network, chromatinActivity, gene list, functional screen, image, and N/A. The 'Output Data Type' list includes: proteomics, image, inputted drug response, geneExpression, isoformExpression, genomicVariants, network, cellularPhysiology, chromatinActivity, drugScreen, and N/A. The main content area is titled 'Software/Tool Types' and has a tabbed interface with five tabs: 'Script', 'Package Binary', 'Package Library', 'Web Application', and 'Other'. The 'Script' tab is selected. Below the tabs is a table with the following columns: 'methodName', 'centerName', 'inputDataType', 'outputDataType', and 'softwareLanguage'. The table contains four rows of tool information.

| methodName | centerName | inputDataType | outputDataType | softwareLanguage |
|---|---|---|---|------------------|
| CIPM | Cancer Systems Biology Center of HoPE (Heterogeneity of Phenotypic Evolution) | <ul style="list-style-type: none">geneExpressionnetworkcellularPhysiology | <ul style="list-style-type: none">cellularPhysiologygeneExpressionnetwork | Matlab, R |
| Master Regulator Inference Algorithm | Center for Cancer Systems Therapeutics (CaST) | <ul style="list-style-type: none">geneExpressionpathway or network | <ul style="list-style-type: none">gene list | Matlab, R |
| Omics Integrator | Embryonal Brain Tumor Networks | <ul style="list-style-type: none">geneExpression | <ul style="list-style-type: none">network | Python |
| Speedy Histopathological-to-ImmunoFluorescent Translation of whole slide images using conditional generative adversarial networks | N/A | <ul style="list-style-type: none">image | <ul style="list-style-type: none">image | Matlab, Python |

Single Interface Combines All Tools and Provides Links

Abstract and information

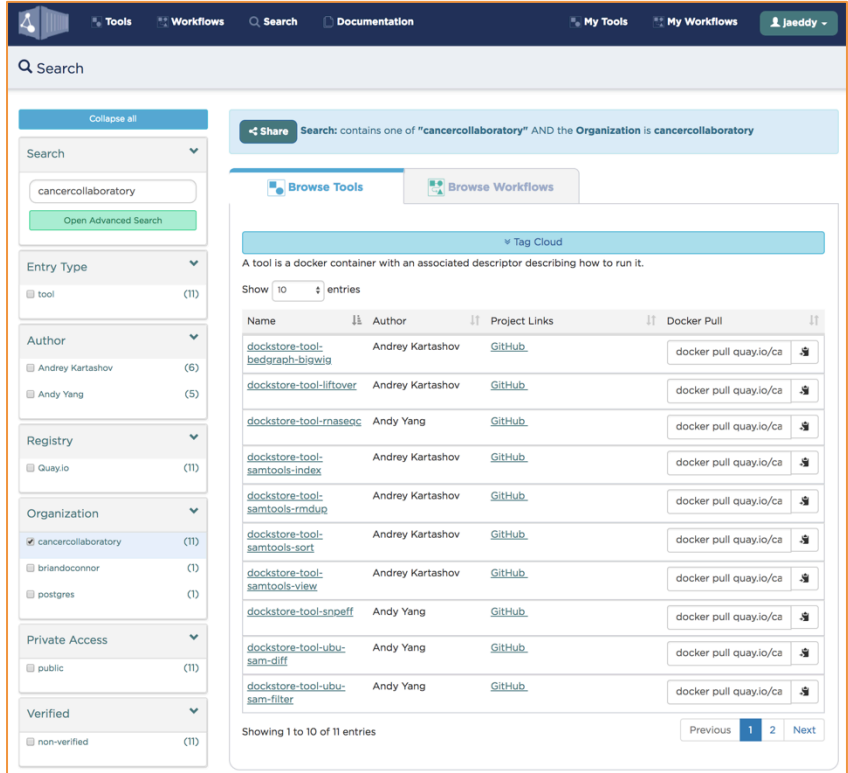
The screenshot displays the CSBC Tool Catalog interface. On the left, there is a sidebar with filter options for 'Input Data Type' and 'Output Data Type'. The main content area shows a modal window for the 'Omics Integrator' tool. The modal window contains the following information:

- Center:** Embryonal Brain Tumor Networks
- Synapse Site:** syn9898746
- Website:** <http://fraenkel-nsf.csbi.mit.edu/omicsintegrator/>
- Abstract:** High-throughput, 'omic' methods provide sensitive measures of biological responses to perturbations. However, inherent biases in high-throughput assays make it difficult to interpret experiments in which more than one type of data is collected. In this work, we introduce Omics Integrator, a software package that takes a variety of 'omic' data as input and identifies putative underlying molecular pathways. The approach applies advanced network optimization algorithms to a network of thousands of molecular interactions to find high-confidence, interpretable subnetworks that best explain the data. These subnetworks connect changes observed in gene expression, protein abundance or other global assays to proteins that may not have been measured in the screens due to inherent bias or noise in measurement. This approach reveals unannotated molecular pathways that would not be detectable by searching pathway databases. Omics Integrator also provides an elegant framework to incorporate not only positive data, but also negative evidence. Incorporating negative evidence allows Omics Integrator to avoid unexpressed genes and avoid being biased toward highly-studied hub proteins, except when they are strongly implicated by the data. The software is comprised of two individual tools, Garnet and Forest, that can be run together or independently to allow a user to perform advanced integration of multiple types of high-throughput data as well as create condition-specific subnetworks of protein interactions that best connect the observed changes in various datasets.

A 'Close' button is located at the bottom right of the modal window.

CSBC/PS-ON Tool Registry: Vision

- Dockerize/standardize tools and workflows in the CSBC/PS-ON Tool Catalog
- Register tools and workflows in Dockstore to enable findability and sharing
- Proposed Dockstore integration in Synapse would provide a way to view and access curated computational tools from CSBC/PS-ON investigators



The screenshot displays the Dockstore web interface. The top navigation bar includes 'Tools', 'Workflows', 'Search', 'Documentation', 'My Tools', and 'My Workflows'. A search bar at the top left contains the text 'cancerlaboratory'. Below the search bar, there are filters for 'Entry Type' (tool), 'Author' (Andrey Kartashov, Andy Yang), 'Registry' (Quay.io), 'Organization' (cancerlaboratory), 'Private Access' (public), and 'Verified' (non-verified). The main content area shows a list of tools with columns for Name, Author, Project Links, and Docker Pull. The tools listed include 'dockstore-tool-bedgraph-biowig', 'dockstore-tool-liftover', 'dockstore-tool-rnaseq', 'dockstore-tool-samtools-index', 'dockstore-tool-samtools-rmdup', 'dockstore-tool-samtools-sort', 'dockstore-tool-samtools-view', 'dockstore-tool-samtools-diff', 'dockstore-tool-snpEff', 'dockstore-tool-ubu-sam-diff', and 'dockstore-tool-ubu-sam-filter'. Each tool entry includes a 'Docker Pull' button and a 'Share' button. The interface also features a 'Tag Cloud' and a 'Showing 1 to 10 of 11 entries' indicator.

| Name | Author | Project Links | Docker Pull |
|--|------------------|------------------------|--|
| dockstore-tool-bedgraph-biowig | Andrey Kartashov | GitHub | docker pull quay.io/ca |
| dockstore-tool-liftover | Andrey Kartashov | GitHub | docker pull quay.io/ca |
| dockstore-tool-rnaseq | Andy Yang | GitHub | docker pull quay.io/ca |
| dockstore-tool-samtools-index | Andrey Kartashov | GitHub | docker pull quay.io/ca |
| dockstore-tool-samtools-rmdup | Andrey Kartashov | GitHub | docker pull quay.io/ca |
| dockstore-tool-samtools-sort | Andrey Kartashov | GitHub | docker pull quay.io/ca |
| dockstore-tool-samtools-view | Andrey Kartashov | GitHub | docker pull quay.io/ca |
| dockstore-tool-snpEff | Andy Yang | GitHub | docker pull quay.io/ca |
| dockstore-tool-ubu-sam-diff | Andy Yang | GitHub | docker pull quay.io/ca |
| dockstore-tool-ubu-sam-filter | Andy Yang | GitHub | docker pull quay.io/ca |