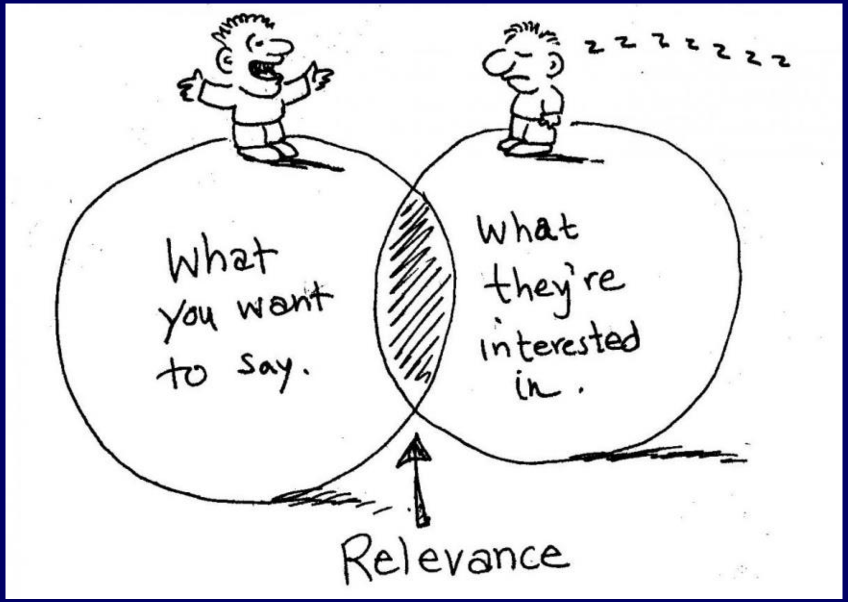
# Informatics II: Data Sharing, Data Provenance, and Downstream…

Jessie Tenenbaum, PhD

@jessiet1023

# Lecture Overview

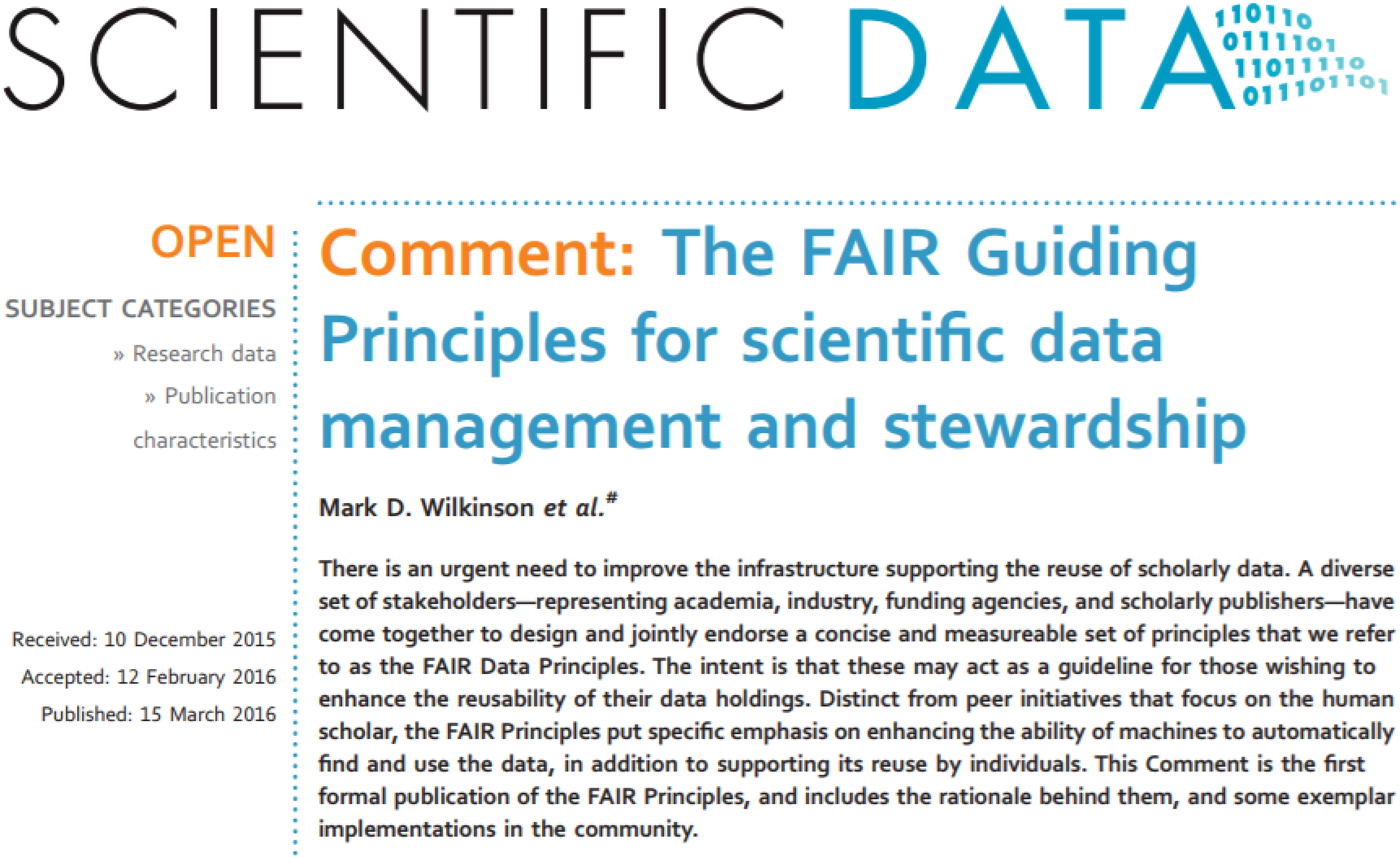
1. Resources for data standards and provenance
2. TCGA
3. Genomics in clinical care
4. DTC Genetic testing



# Data sharing

http://namp.americansforthearts.org/2011/02/16/do-you-have-trust-issues-%E2%80%93-data-sharing-and-the-arts

# FAIR Principles



SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES
» Research data
» Publication
characteristics

## Comment: The FAIR Guiding Principles for scientific data management and stewardship

**Mark D. Wilkinson *et al.*[#]**

Received: 10 December 2015
Accepted: 12 February 2016
Published: 15 March 2016

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measureable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them, and some exemplar implementations in the community.

# Findable

- F1. (meta)data are assigned a globally unique and persistent identifier
- F2. data are described with rich metadata (defined by R1 below)
- F3. metadata clearly and explicitly include the identifier of the data it describes
- F4. (meta)data are registered or indexed in a searchable resource

# Accessible

- A1. (meta)data are retrievable by their identifier using a standardized communications protocol
  - A1.1 the protocol is open, free, and universally implementable
  - A1.2 the protocol allows for an authentication and authorization procedure, where necessary
- A2. metadata are accessible, even when the data are no longer available

# Interoperable

- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles
- I3. (meta)data include qualified references to other (meta)data

# Reusable

- R1. meta(data) are richly described with a plurality of accurate and relevant attributes
  - R1.1. (meta)data are released with a clear and accessible data usage license
  - R1.2. (meta)data are associated with detailed provenance
  - R1.3. (meta)data meet domain-relevant community standards

- *Sometimes, data warehouses resemble landfills more than libraries*
- *Apply the 4 C’s: collect, catalogue, clean and curate.*

Source unknown, with apologies

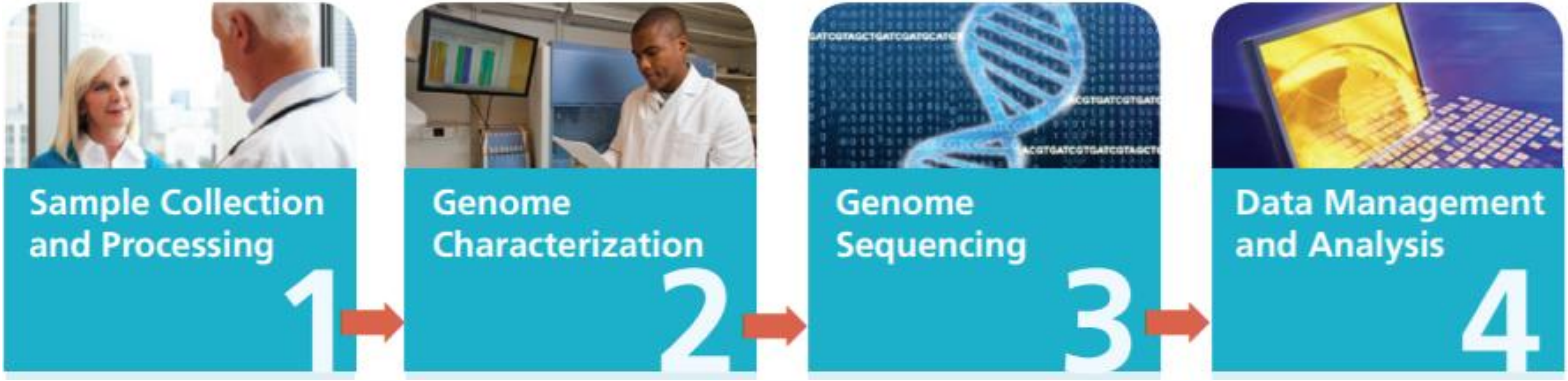
National Cancer Institute
National Human Genome Research Institute

# The Cancer Genome Atlas:

## Charting a New Course for Cancer Prevention, Diagnosis and Treatment

- Create a detailed catalog, or “atlas,” of genomic changes associated with tumors types
- Genome analysis and characterization technologies
- Accelerate understanding of the molecular basis of cancer
- Improve prevention, diagnosis and treatment of cancer

# 4 Components of TCGA Network



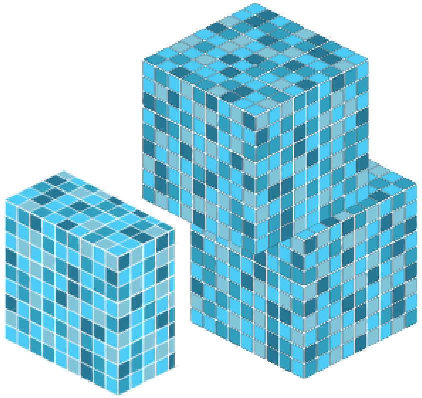
# TCGA by the numbers

TCGA produced over

**2.5**

PETABYTES

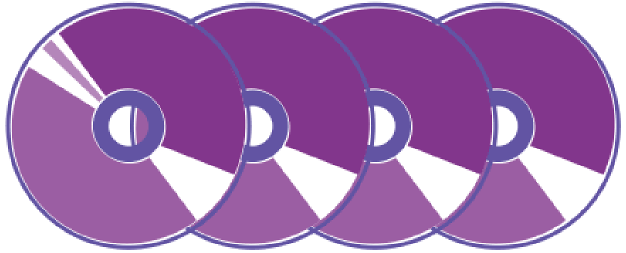
of data



To put this into perspective, **1 petabyte** of data is equal to

**212,000**

DVDs



TCGA data describes

**33**

DIFFERENT TUMOR TYPES

...including

**10**

RARE CANCERS

...based on paired tumor and normal tissue sets collected from

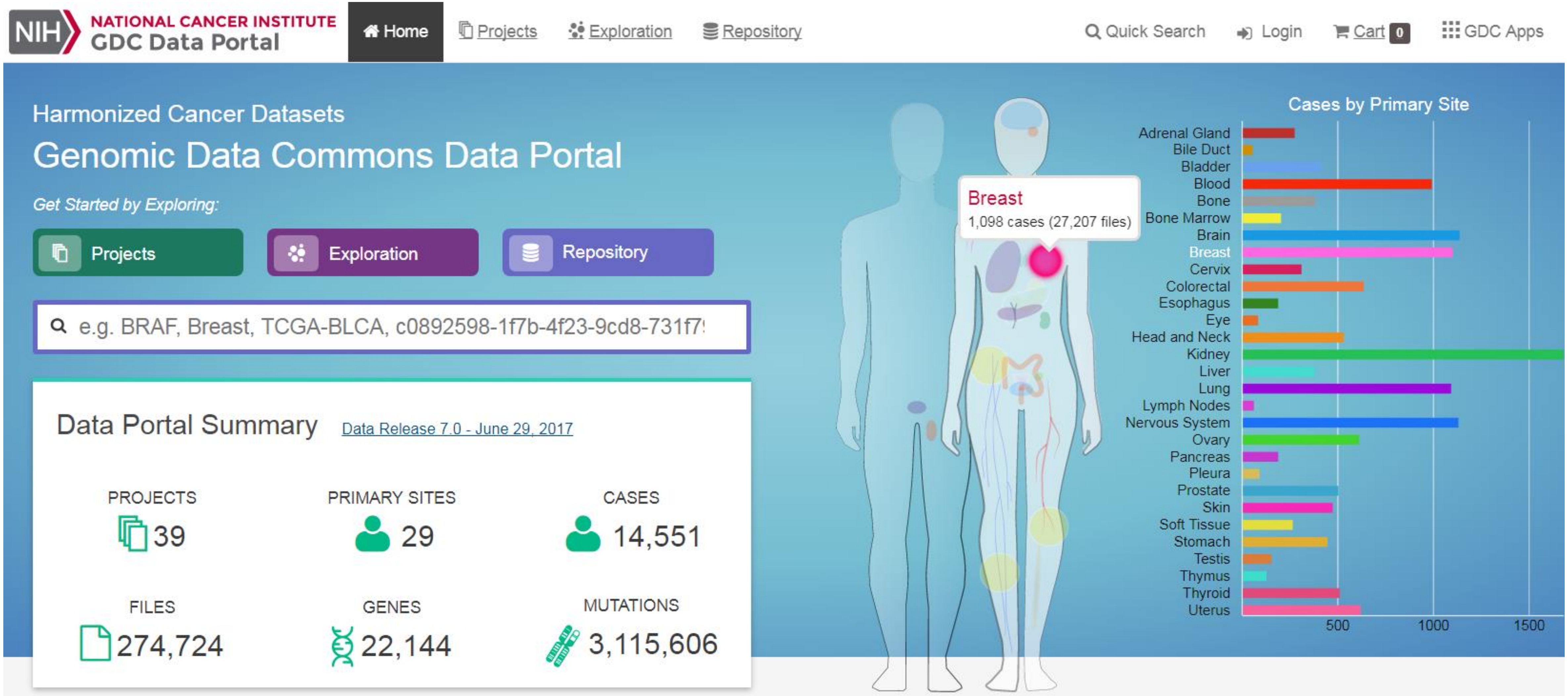
**11,000**

PATIENTS

...using

**7**

DIFFERENT DATA TYPES



Harmonized Cancer Datasets

Genomic Data Commons Data Portal

Data Portal Summary Data Release 7.0 - June 29, 2017

PROJECTS 39

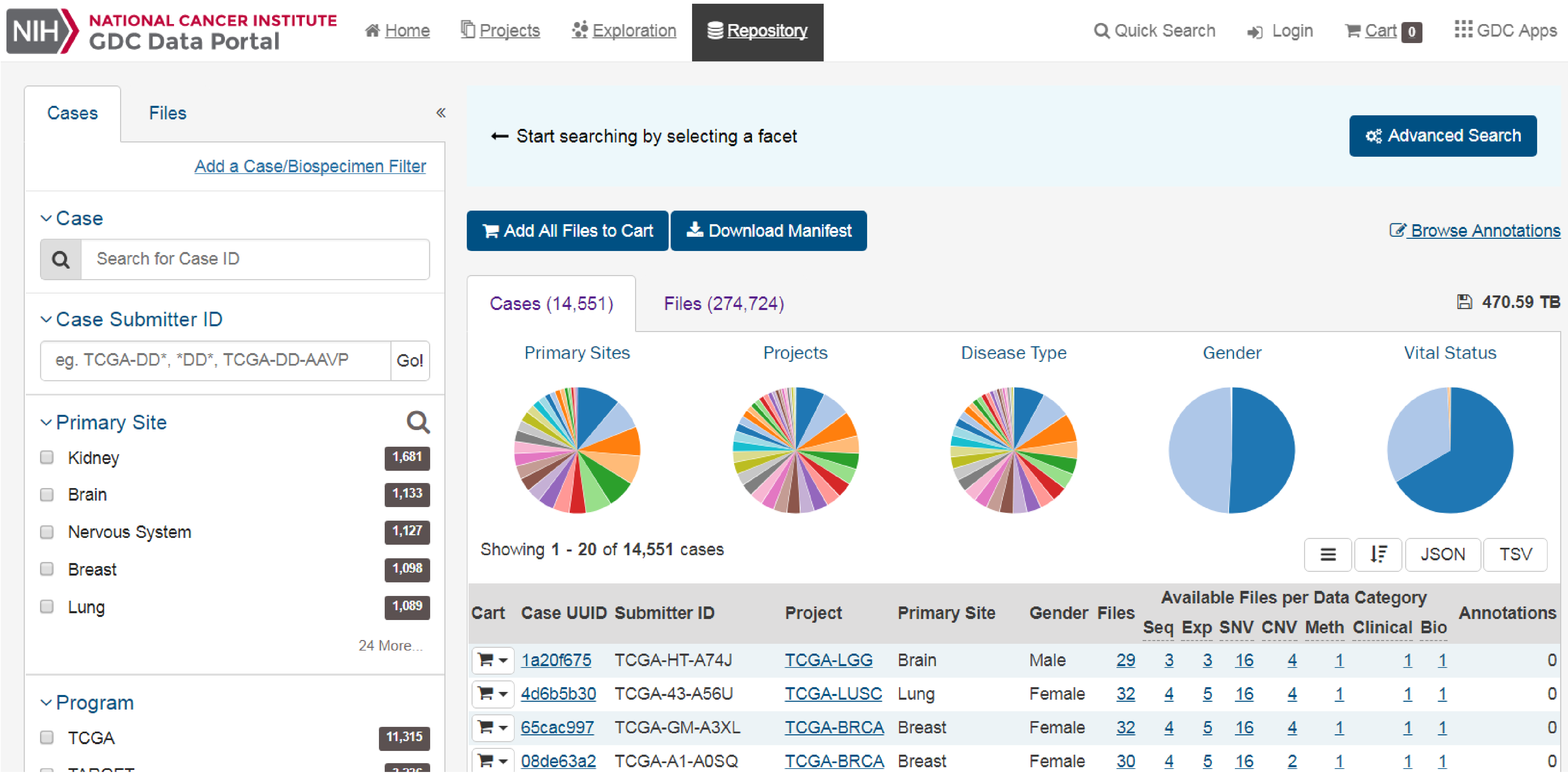
PRIMARY SITES 29

CASES 14,551

FILES 274,724

GENES 22,144

MUTATIONS 3,115,606



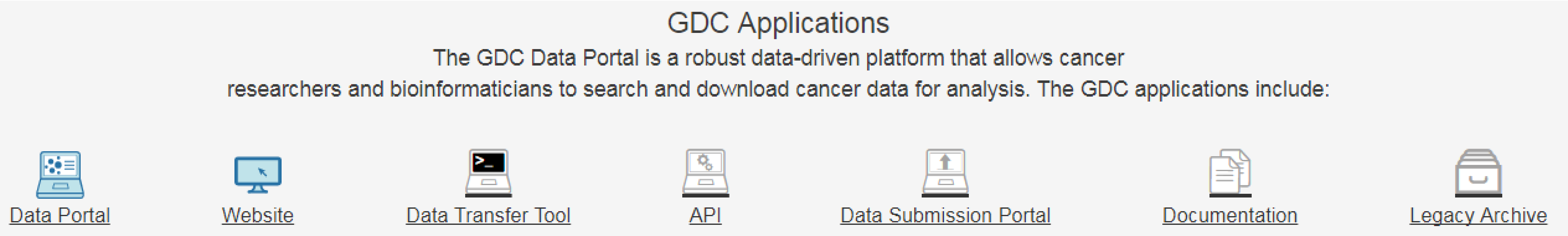
☐ Kidney 1,681
☐ Brain 1,133
☐ Nervous System 1,127
☐ Breast 1,098
☐ Lung 1,089

☐ TCGA 11,315

Cases (14,551) Files (274,724) 470.59 TB

Showing **1 - 20** of **14,551** cases

| Cart | Case UUID | Submitter ID | Project | Primary Site | Gender | Files | Available Files per Data Category | | | | | | | Annotations |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | Seq | Exp | SNV | CNV | Meth | Clinical | Bio | |
| | 1a20f675 | TCGA-HT-A74J | TCGA-LGG | Brain | Male | 29 | 3 | 3 | 16 | 4 | 1 | 1 | 1 | 0 |
| | 4d6b5b30 | TCGA-43-A56U | TCGA-LUSC | Lung | Female | 32 | 4 | 5 | 16 | 4 | 1 | 1 | 1 | 0 |
| | 65cac997 | TCGA-GM-A3XL | TCGA-BRCA | Breast | Female | 32 | 4 | 5 | 16 | 4 | 1 | 1 | 1 | 0 |
| | 08de63a2 | TCGA-A1-A0SQ | TCGA-BRCA | Breast | Female | 30 | 4 | 5 | 16 | 2 | 1 | 1 | 1 | 0 |



GDC Applications

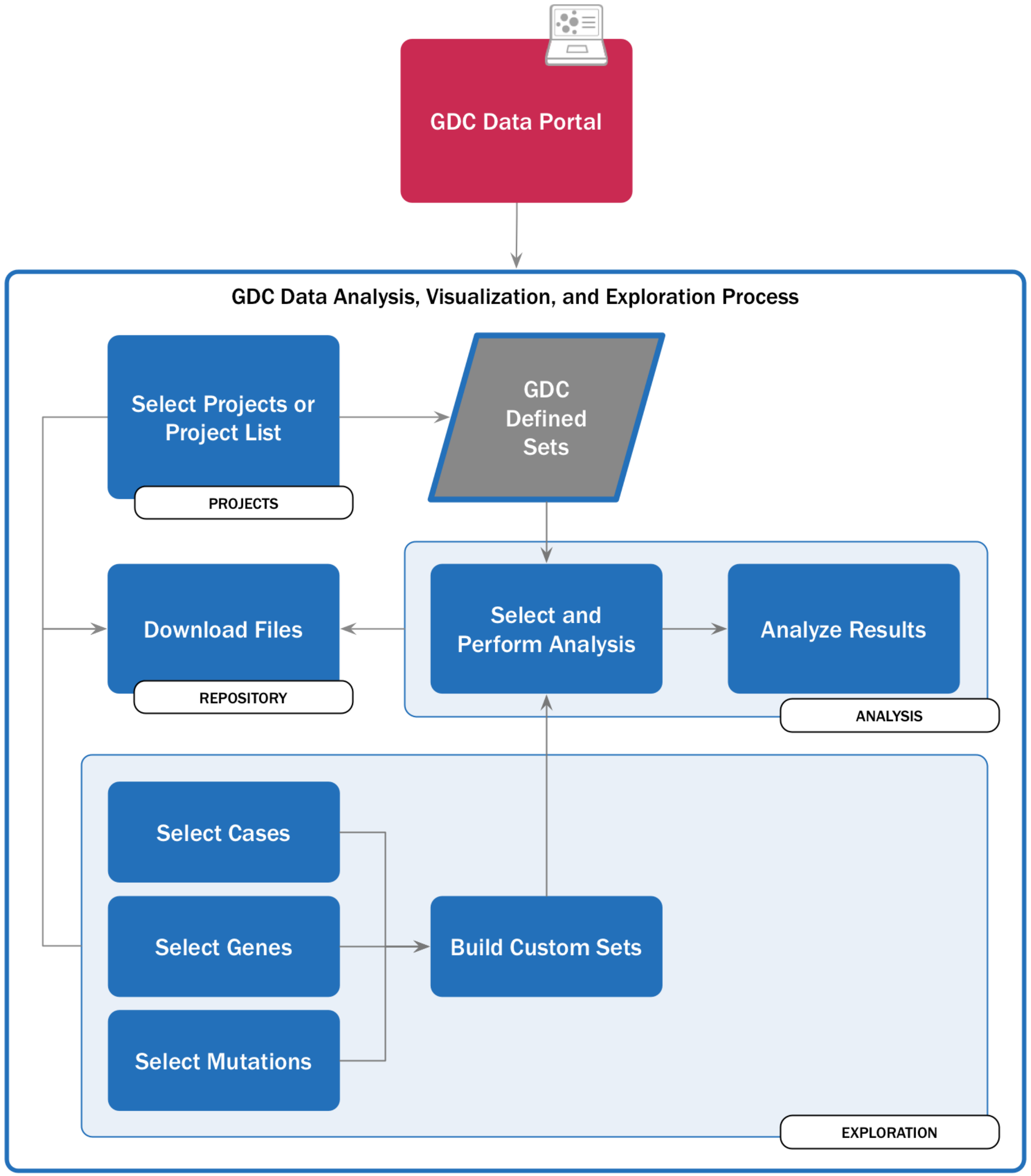
The GDC Data Portal is a robust data-driven platform that allows cancer researchers and bioinformaticians to search and download cancer data for analysis. The GDC applications include:

# TCGA RESULTS & FINDINGS

| | | |
|---|---|---|
| **MOLECULAR BASIS OF CANCER** | Improved our understanding of the genomic underpinnings of cancer | Breast and ovarian cancer similarity |
| **TUMOR SUBTYPES** | Revolutionized how cancer is classified | Tumor type based on genetic alterations |
| **THERAPEUTIC TARGETS** | Identified genomic characteristics of tumors that can be targeted with currently available therapies or used to help with drug development | Targetable alteration in lung squamous cell carcinoma led to trial |

# GDC

# Data Analysis, Visualization, and Exploration (DAVE) Tools

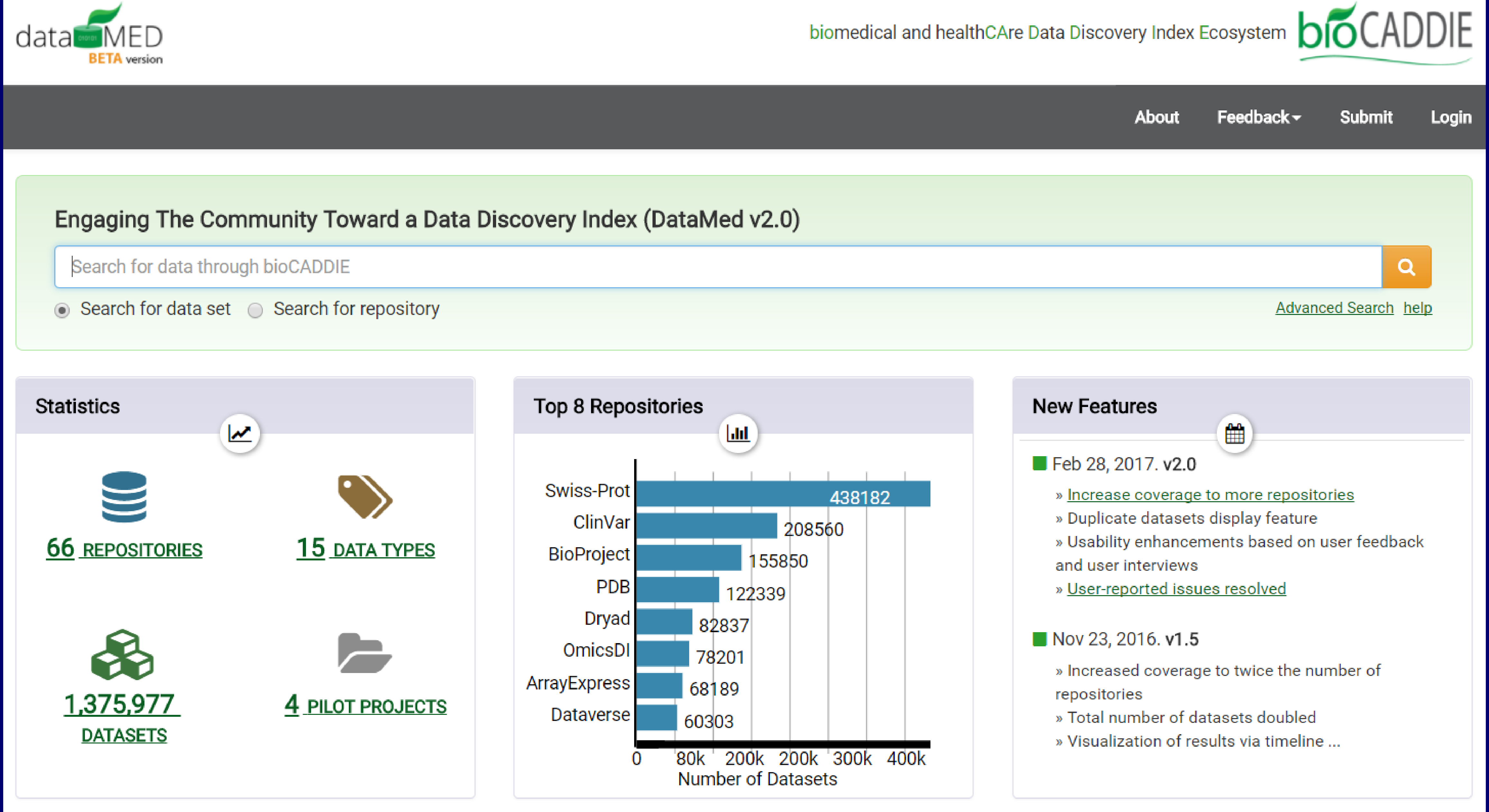


# NCBI Genomic Resources

- ClinVar
  - Like PubMed for clinical variants- database of relationships among human variations and phenotypes, with supporting evidence
  - Content attributable to authors
  - Content from one author may contradict other content
- ClinGen
  - More like a review journal
  - Reviews material from ClinVar and other sources
  - Reports represent combined intellectual effort of contributors and ClinGen staff

# More resources

- NCBI
  - dbGaP
  - dbSNP
  - Gene Expression Omnibus
- ArrayExpress (EBI)
- DataMed (NIH BioCADDIE)



dataMED BETA version

biomedical and healthCAre Data Discovery Index Ecosystem bioCADDIE

## Engaging The Community Toward a Data Discovery Index (DataMed v2.0)

Search for data through bioCADDIE

● Search for data set ○ Search for repository

### Statistics

66 REPOSITORIES

15 DATA TYPES

1,375,977 DATASETS

4 PILOT PROJECTS

### Top 8 Repositories

### New Features

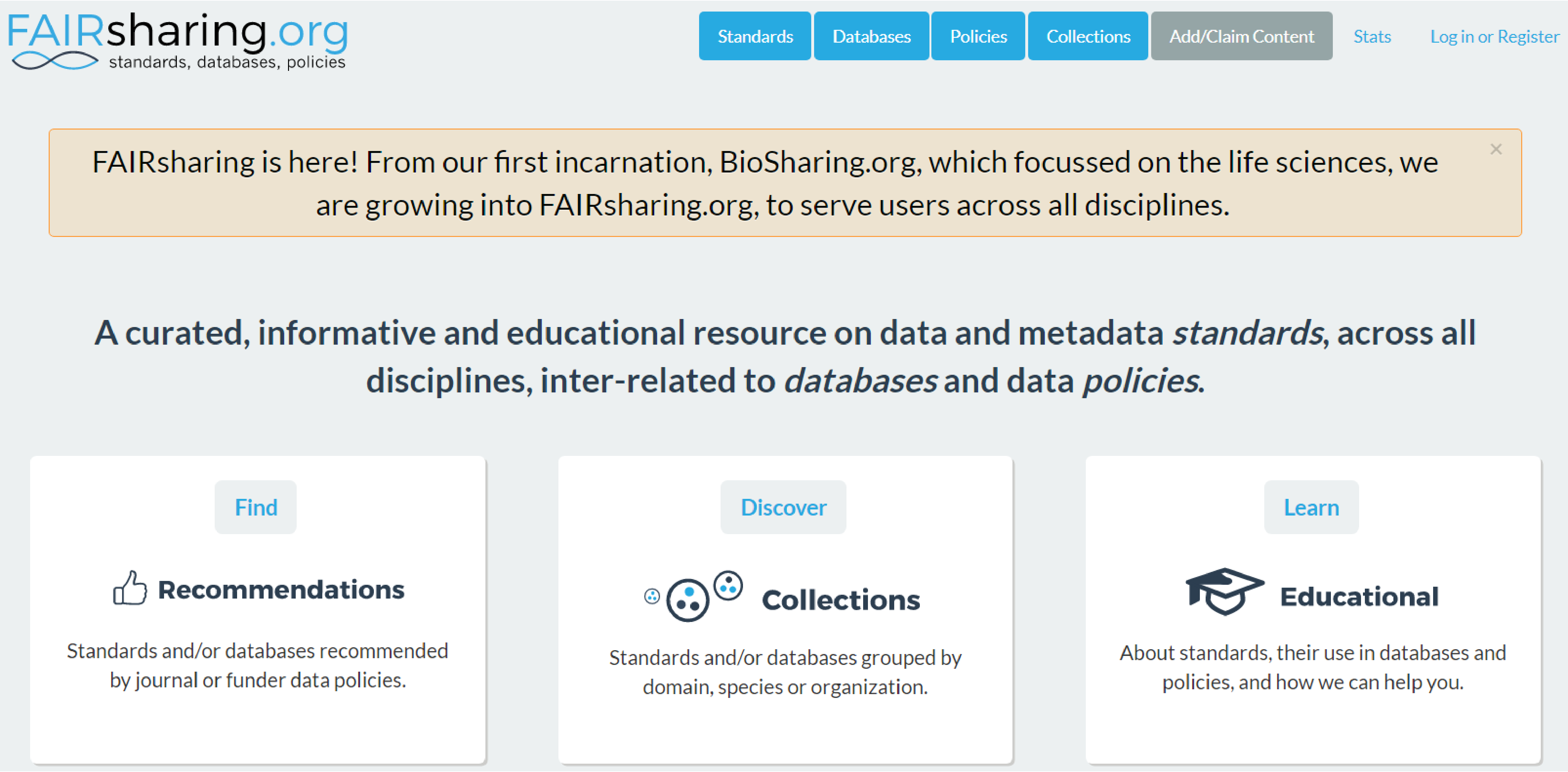
🟩 Feb 28, 2017. **v2.0**

- » Increase coverage to more repositories
- » Duplicate datasets display feature
- » Usability enhancements based on user feedback and user interviews
- » User-reported issues resolved

🟩 Nov 23, 2016. **v1.5**

- » Increased coverage to twice the number of repositories
- » Total number of datasets doubled
- » Visualization of results via timeline ...

Duke Medicine



FAIRsharing is here! From our first incarnation, BioSharing.org, which focussed on the life sciences, we are growing into FAIRsharing.org, to serve users across all disciplines.

A curated, informative and educational resource on data and metadata *standards*, across all disciplines, inter-related to *databases* and data *policies*.

**Recommendations**

Standards and/or databases recommended by journal or funder data policies.

**Collections**

Standards and/or databases grouped by domain, species or organization.

**Educational**

About standards, their use in databases and policies, and how we can help you.

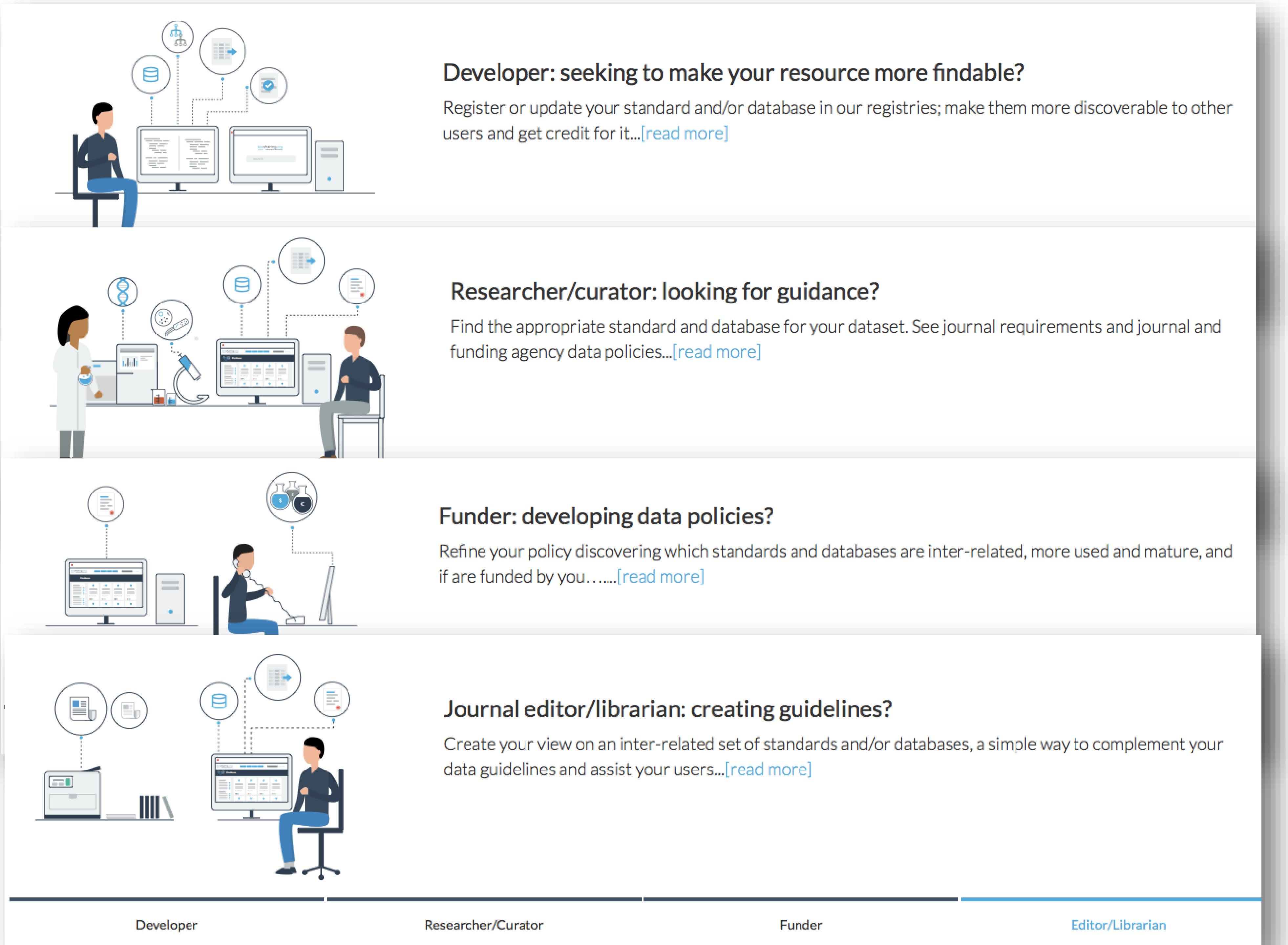
# Helping people make the right decisions

My funder's data policy recommends the use of established standards, but which are widely endorsed and applicable to my **crop** data?

Which are the **mature standards** and **standards-compliant databases** that we should recommend to our authors?

We need a standard for **sharing social science data**, what's out there and who should we talk to?

I have some old **rice genomic data** in format X, which is now **deprecated**; what format has replaced X?



### Developer: seeking to make your resource more findable?

Register or update your standard and/or database in our registries; make them more discoverable to other users and get credit for it...[read more]

### Researcher/curator: looking for guidance?

Find the appropriate standard and database for your dataset. See journal requirements and journal and funding agency data policies...[read more]

### Funder: developing data policies?

Refine your policy discovering which standards and databases are inter-related, more used and mature, and if are funded by you…......[read more]

### Journal editor/librarian: creating guidelines?

Create your view on an inter-related set of standards and/or databases, a simple way to complement your data guidelines and assist your users...[read more]

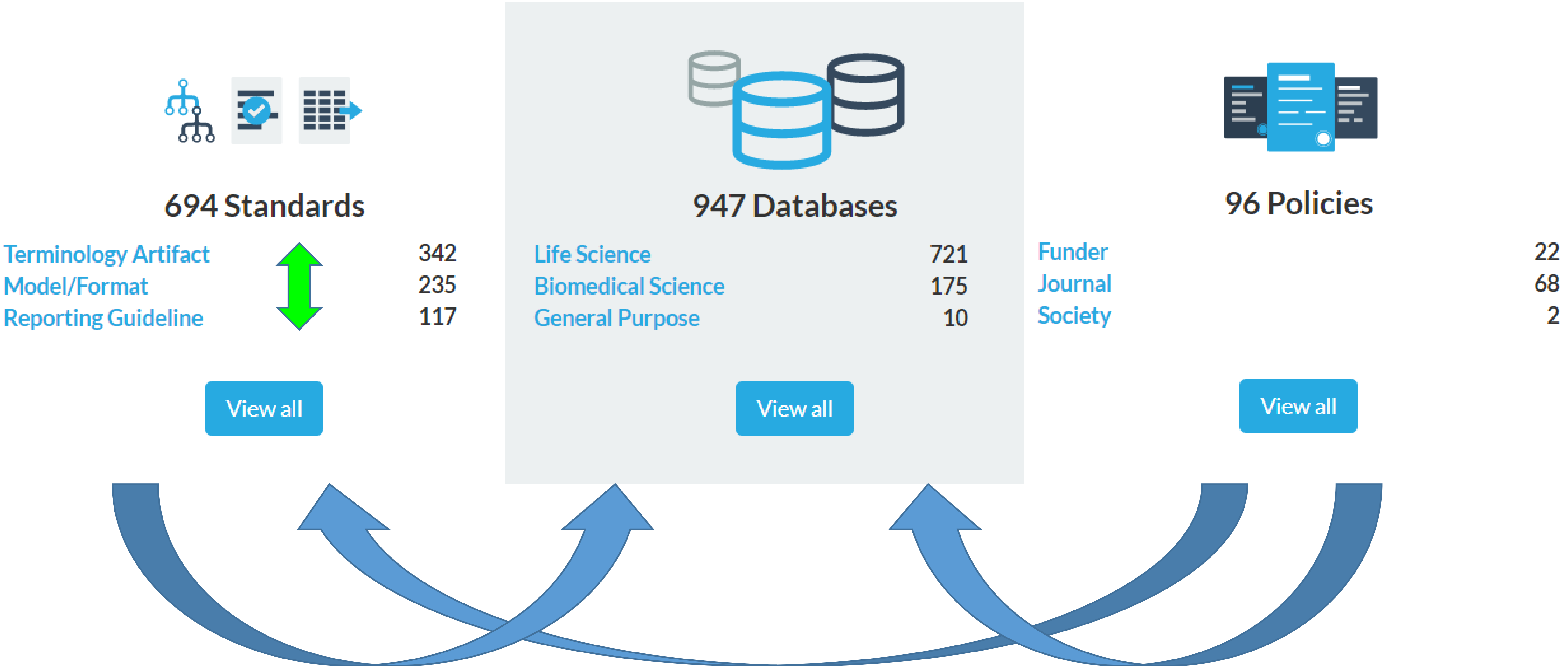
Courtesy of SA Sansone, Univ of Oxford

# Researcher/curator: looking for guidance?

Researcher/Curator

- Find the appropriate standard and database for your dataset
- See journal requirements and journal and funding agency data policies...

# FAIRsharing by the numbers



**694 Standards**

| | |
|---|---|
| Terminology Artifact | 342 |
| Model/Format | 235 |
| Reporting Guideline | 117 |

**947 Databases**

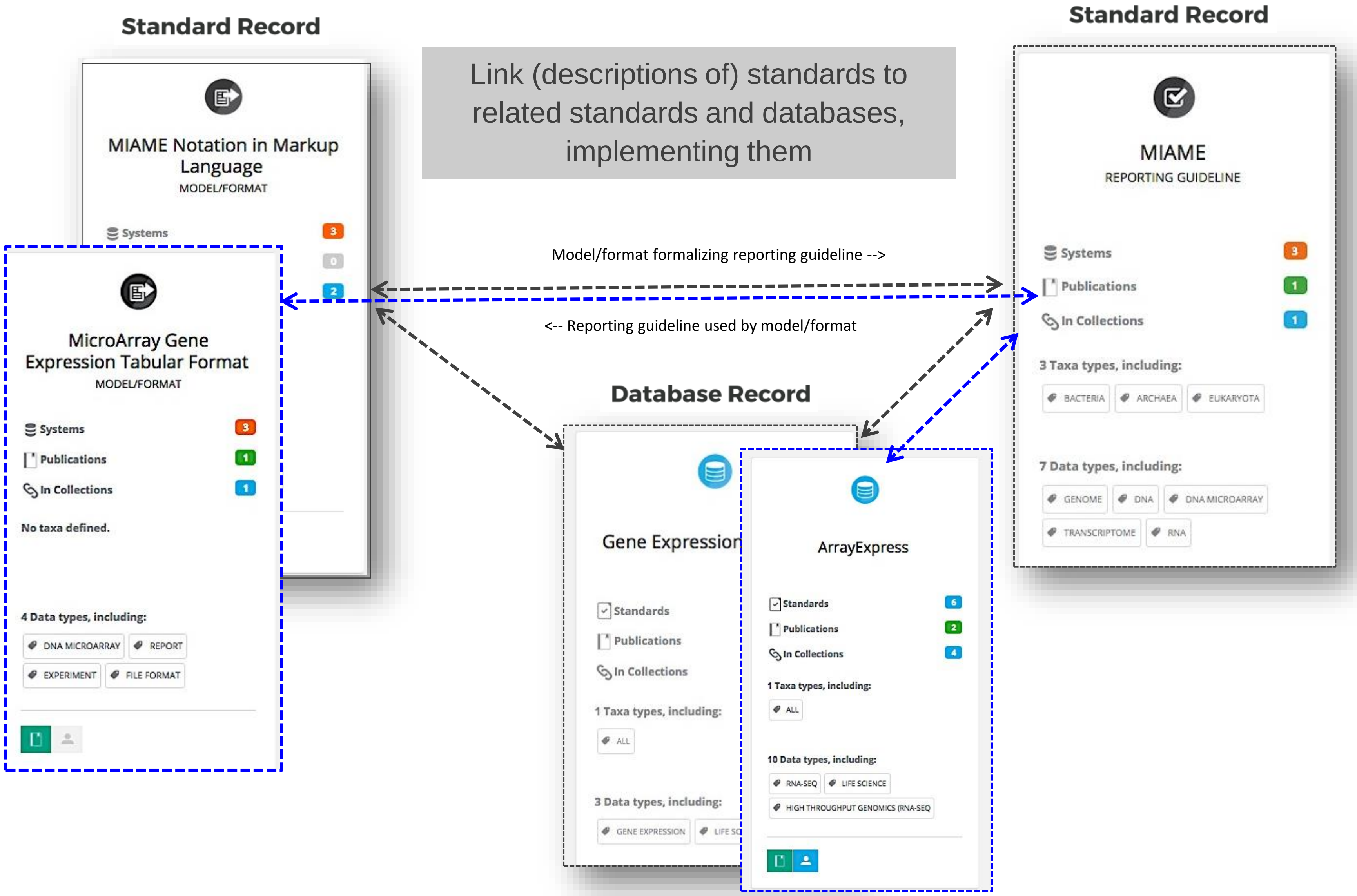
| | |
|---|---|
| Life Science | 721 |
| Biomedical Science | 175 |
| General Purpose | 10 |

**96 Policies**

| | |
|---|---|
| Funder | 22 |
| Journal | 68 |
| Society | 2 |

# Cross-linking standards to standards and databases

Link (descriptions of) standards to related standards and databases, implementing them



**Standard Record**

MIAME Notation in Markup Language
MODEL/FORMAT

Systems 3
0
2

MicroArray Gene Expression Tabular Format
MODEL/FORMAT

Systems 3
Publications 1
In Collections 1

No taxa defined.

4 Data types, including:
DNA MICROARRAY | REPORT
EXPERIMENT | FILE FORMAT

Model/format formalizing reporting guideline -->

<-- Reporting guideline used by model/format

**Standard Record**

MIAME
REPORTING GUIDELINE

Systems 3
Publications 1
In Collections 1

3 Taxa types, including:
BACTERIA | ARCHAEA | EUKARYOTA

7 Data types, including:
GENOME | DNA | DNA MICROARRAY
TRANSCRIPTOME | RNA

**Database Record**

Gene Expression

Standards
Publications
In Collections

1 Taxa types, including:
ALL

3 Data types, including:
GENE EXPRESSION | LIFE SC

ArrayExpress

Standards 6
Publications 2
In Collections 4

1 Taxa types, including:
ALL

10 Data types, including:
RNA-SEQ | LIFE SCIENCE
HIGH THROUGHPUT GENOMICS (RNA-SEQ

# Indicators describe the ‘status’ of a standard

R Ready for use, implementation, or recommendation

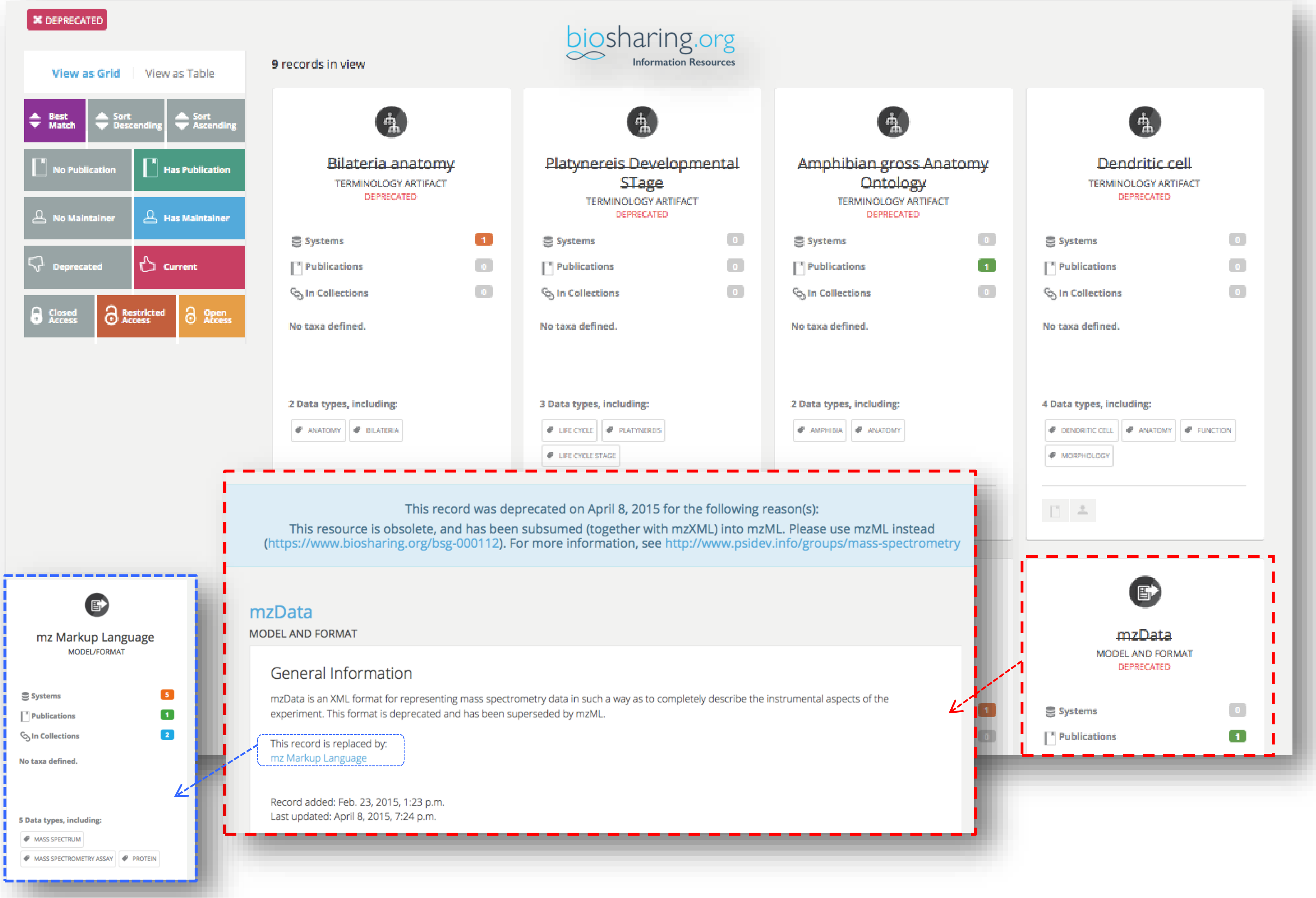
Dev In development

U Status uncertain

D Deprecated as subsumed or superseded

Manually curated, approved by the community

# Deprecations and substitutions are key to track evolution



✖ DEPRECATED

biosharing.org
Information Resources

View as Grid | View as Table

Best Match | Sort Descending | Sort Ascending
No Publication | Has Publication
No Maintainer | Has Maintainer
Deprecated | Current
Closed Access | Restricted Access | Open Access

9 records in view

~~Bilateria anatomy~~
TERMINOLOGY ARTIFACT
DEPRECATED

| | |
|---|---|
| Systems | 1 |
| Publications | 0 |
| In Collections | 0 |

No taxa defined.

2 Data types, including:
ANATOMY | BILATERIA

~~Platynereis Developmental STage~~
TERMINOLOGY ARTIFACT
DEPRECATED

| | |
|---|---|
| Systems | 0 |
| Publications | 0 |
| In Collections | 0 |

No taxa defined.

3 Data types, including:
LIFE CYCLE | PLATYNEREIS | LIFE CYCLE STAGE

~~Amphibian gross Anatomy Ontology~~
TERMINOLOGY ARTIFACT
DEPRECATED

| | |
|---|---|
| Systems | 0 |
| Publications | 1 |
| In Collections | 0 |

No taxa defined.

2 Data types, including:
AMPHIBIA | ANATOMY

~~Dendritic cell~~
TERMINOLOGY ARTIFACT
DEPRECATED

| | |
|---|---|
| Systems | 0 |
| Publications | 0 |
| In Collections | 0 |

No taxa defined.

4 Data types, including:
DENDRITIC CELL | ANATOMY | FUNCTION | MORPHOLOGY

This record was deprecated on April 8, 2015 for the following reason(s):
This resource is obsolete, and has been subsumed (together with mzXML) into mzML. Please use mzML instead (https://www.biosharing.org/bsg-000112). For more information, see http://www.psidev.info/groups/mass-spectrometry

mzData
MODEL AND FORMAT

General Information

mzData is an XML format for representing mass spectrometry data in such a way as to completely describe the instrumental aspects of the experiment. This format is deprecated and has been superseded by mzML.

This record is replaced by:
mz Markup Language

Record added: Feb. 23, 2015, 1:23 p.m.
Last updated: April 8, 2015, 7:24 p.m.

mz Markup Language
MODEL/FORMAT

| | |
|---|---|
| Systems | 5 |
| Publications | 1 |
| In Collections | 2 |

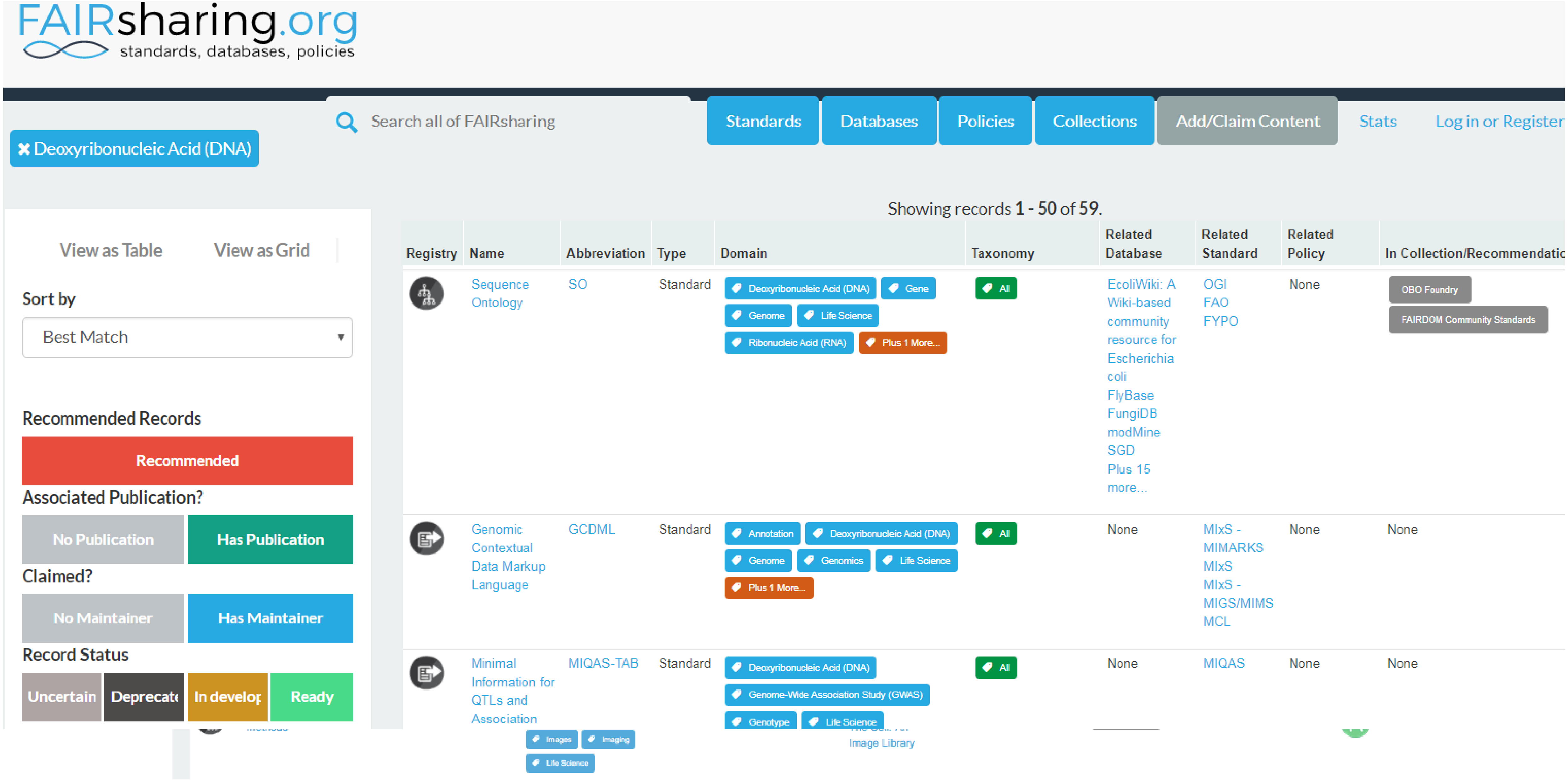
No taxa defined.

5 Data types, including:
MASS SPECTRUM | MASS SPECTROMETRY ASSAY | PROTEIN

~~mzData~~
MODEL AND FORMAT
DEPRECATED

| | |
|---|---|
| Systems | 0 |
| Publications | 1 |

# Discover standards, databases and data policies



FAIRsharing.org
standards, databases, policies

Search all of FAIRsharing

**Sort by**

Best Match ▾

**Recommended Records**

**Associated Publication?**

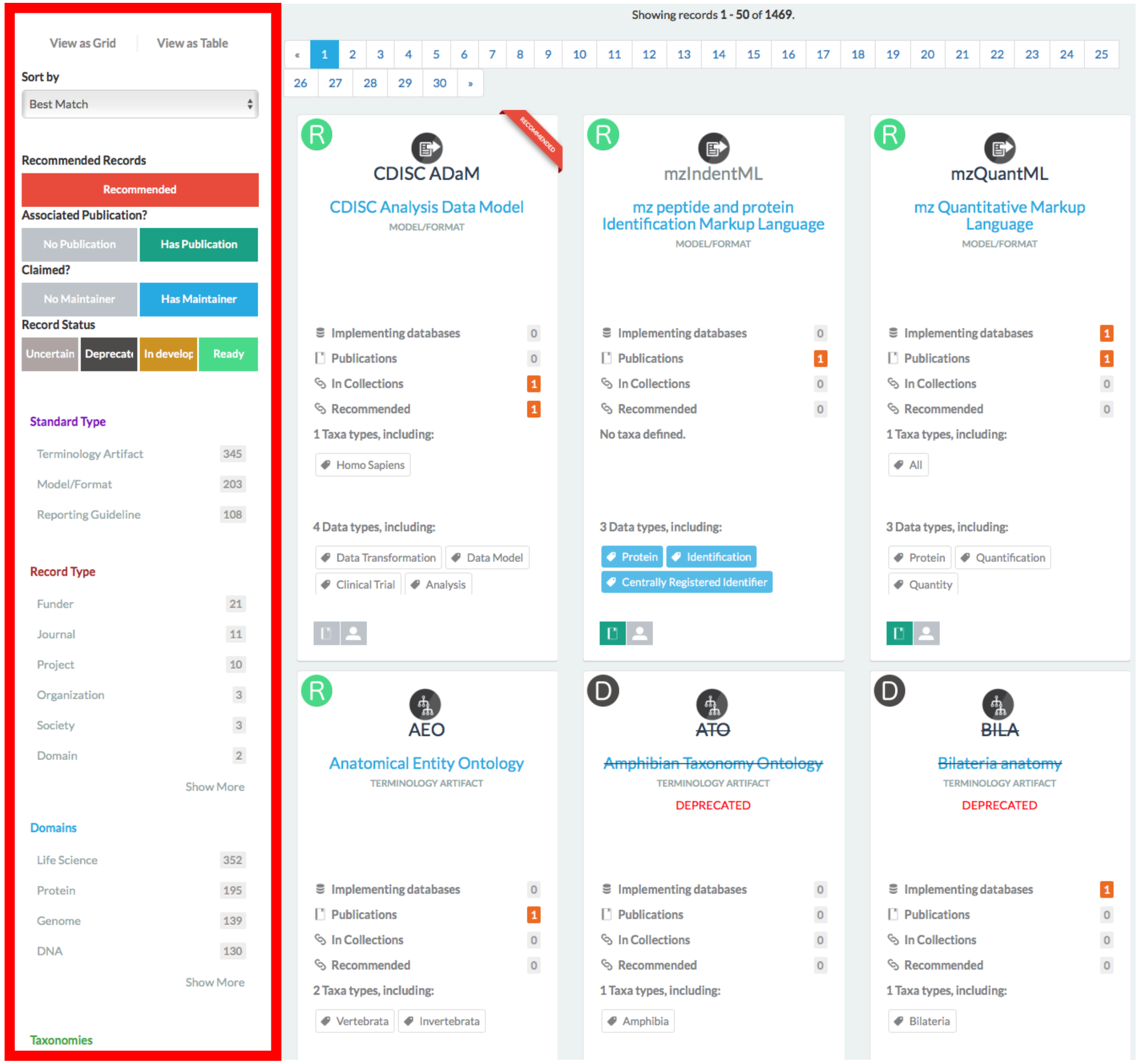
**Claimed?**

**Record Status**

Showing records **1 - 50** of **59**.

| Registry | Name | Abbreviation | Type | Domain | Taxonomy | Related Database | Related Standard | Related Policy | In Collection/Recommendatic |
|---|---|---|---|---|---|---|---|---|---|
| | Sequence Ontology | SO | Standard | Deoxyribonucleic Acid (DNA)<br>Gene<br>Genome<br>Life Science<br>Ribonucleic Acid (RNA)<br>Plus 1 More... | All | EcoliWiki: A Wiki-based community resource for Escherichia coli<br>FlyBase<br>FungiDB<br>modMine<br>SGD<br>Plus 15 more... | OGI<br>FAO<br>FYPO | None | OBO Foundry<br>FAIRDOM Community Standards |
| | Genomic Contextual Data Markup Language | GCDML | Standard | Annotation<br>Deoxyribonucleic Acid (DNA)<br>Genome<br>Genomics<br>Life Science<br>Plus 1 More... | All | None | MIxS - MIMARKS<br>MIxS<br>MIxS - MIGS/MIMS<br>MCL | None | None |
| | Minimal Information for QTLs and Association | MIQAS-TAB | Standard | Deoxyribonucleic Acid (DNA)<br>Genome-Wide Association Study (GWAS)<br>Genotype<br>Life Science | All | None | MIQAS | None | None |

# Filter and refine using the faceted search



View as Grid | View as Table

Sort by: Best Match

Recommended Records

Recommended

Associated Publication?

No Publication | Has Publication

Claimed?

No Maintainer | Has Maintainer

Record Status

Uncertain | Deprecat | In develop | Ready

**Standard Type**

| | |
|---|---|
| Terminology Artifact | 345 |
| Model/Format | 203 |
| Reporting Guideline | 108 |

**Record Type**

| | |
|---|---|
| Funder | 21 |
| Journal | 11 |
| Project | 10 |
| Organization | 3 |
| Society | 3 |
| Domain | 2 |

Show More

**Domains**

| | |
|---|---|
| Life Science | 352 |
| Protein | 195 |
| Genome | 139 |
| DNA | 130 |

Show More

**Taxonomies**

Showing records 1 - 50 of 1469.

« 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 »

R — RECOMMENDED

CDISC ADaM

CDISC Analysis Data Model

MODEL/FORMAT

| | |
|---|---|
| Implementing databases | 0 |
| Publications | 0 |
| In Collections | 1 |
| Recommended | 1 |

1 Taxa types, including: Homo Sapiens

4 Data types, including: Data Transformation, Data Model, Clinical Trial, Analysis

R

mzIndentML

mz peptide and protein Identification Markup Language

MODEL/FORMAT

| | |
|---|---|
| Implementing databases | 0 |
| Publications | 1 |
| In Collections | 0 |
| Recommended | 0 |

No taxa defined.

3 Data types, including: Protein, Identification, Centrally Registered Identifier

R

mzQuantML

mz Quantitative Markup Language

MODEL/FORMAT

| | |
|---|---|
| Implementing databases | 1 |
| Publications | 1 |
| In Collections | 0 |
| Recommended | 0 |

1 Taxa types, including: All

3 Data types, including: Protein, Quantification, Quantity

R

AEO

Anatomical Entity Ontology

TERMINOLOGY ARTIFACT

| | |
|---|---|
| Implementing databases | 0 |
| Publications | 1 |
| In Collections | 0 |
| Recommended | 0 |

2 Taxa types, including: Vertebrata, Invertebrata

D

~~ATO~~

~~Amphibian Taxonomy Ontology~~

TERMINOLOGY ARTIFACT

DEPRECATED

| | |
|---|---|
| Implementing databases | 0 |
| Publications | 0 |
| In Collections | 0 |
| Recommended | 0 |

1 Taxa types, including: Amphibia

D

~~BILA~~

~~Bilateria anatomy~~

TERMINOLOGY ARTIFACT

DEPRECATED

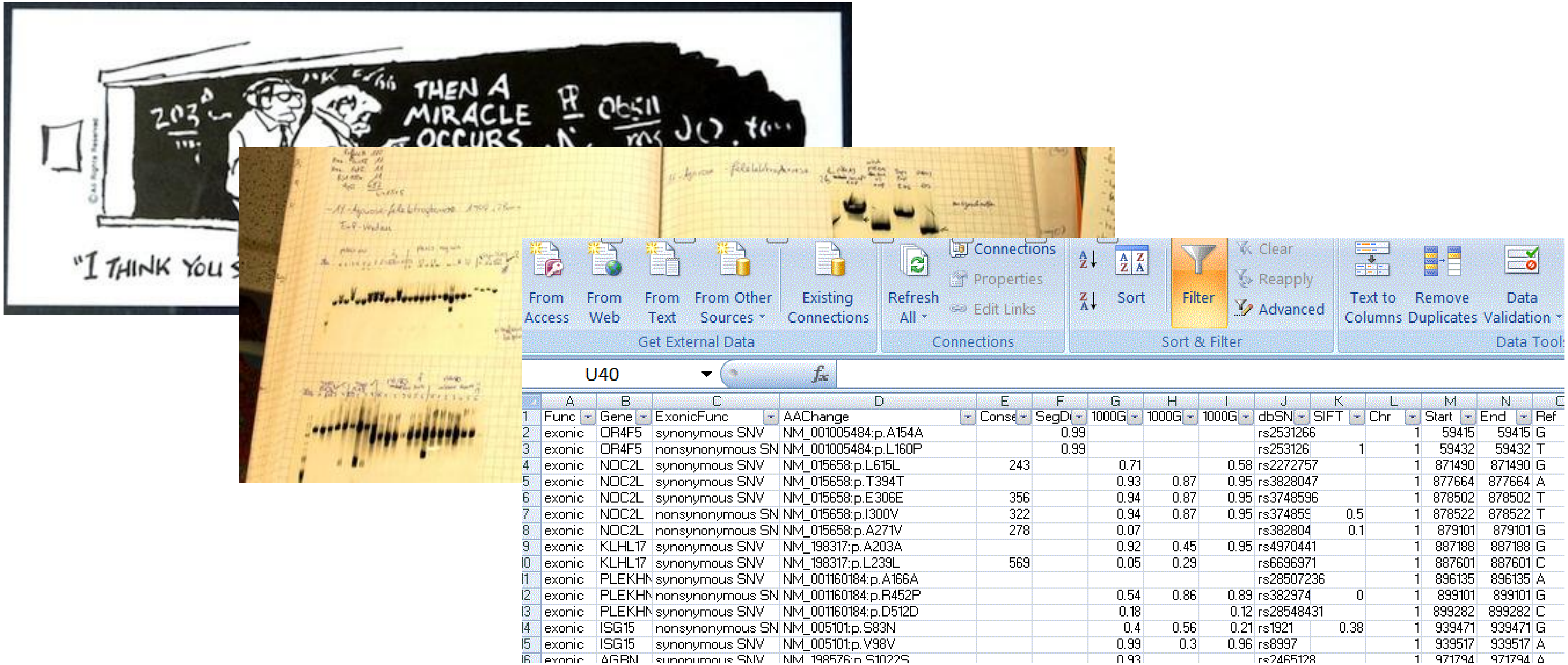
| | |
|---|---|
| Implementing databases | 1 |
| Publications | 0 |
| In Collections | 0 |
| Recommended | 0 |

1 Taxa types, including: Bilateria

FAIRsharing.org
standards, databases, policies

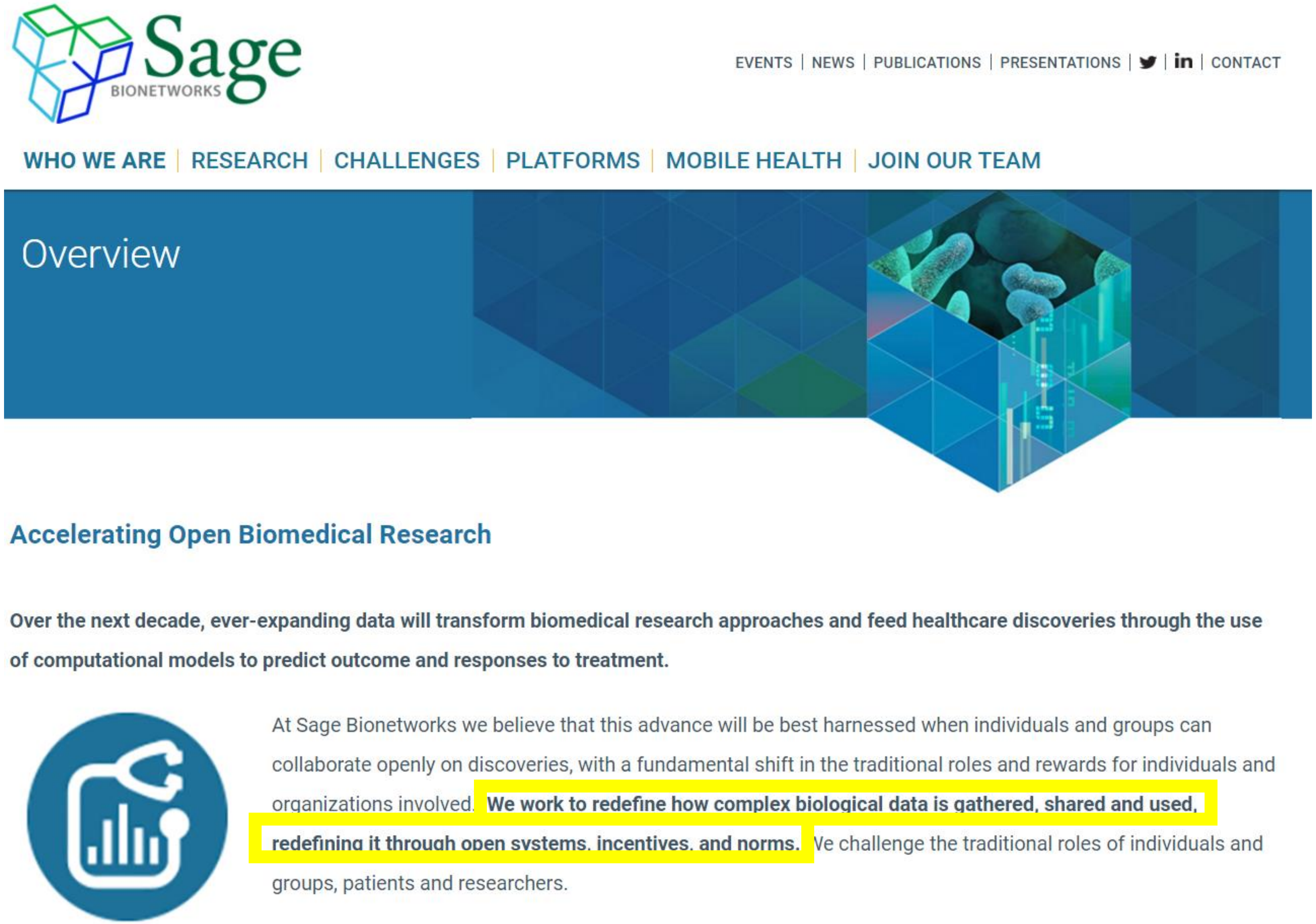
# FAIRsharing DEMO

# Data Provenance



| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Func | Gene | ExonicFunc | AAChange | Conse | SegD | 1000G | 1000G | 1000G | dbSN | SIFT | Chr | Start | End | Ref |
| 2 | exonic | OR4F5 | synonymous SNV | NM_001005484:p.A154A | | 0.99 | | | | rs2531266 | | 1 | 59415 | 59415 | G |
| 3 | exonic | OR4F5 | nonsynonymous SN | NM_001005484:p.L160P | | 0.99 | | | | rs253126 | 1 | 1 | 59432 | 59432 | T |
| 4 | exonic | NOC2L | synonymous SNV | NM_015658:p.L615L | 243 | | 0.71 | | 0.58 | rs2272757 | | 1 | 871490 | 871490 | G |
| 5 | exonic | NOC2L | synonymous SNV | NM_015658:p.T394T | | | 0.93 | 0.87 | 0.95 | rs3828047 | | 1 | 877664 | 877664 | A |
| 6 | exonic | NOC2L | synonymous SNV | NM_015658:p.E306E | 356 | | 0.94 | 0.87 | 0.95 | rs3748596 | | 1 | 878502 | 878502 | T |
| 7 | exonic | NOC2L | nonsynonymous SN | NM_015658:p.I300V | 322 | | 0.94 | 0.87 | 0.95 | rs374859 | 0.5 | 1 | 878522 | 878522 | T |
| 8 | exonic | NOC2L | nonsynonymous SN | NM_015658:p.A271V | 278 | | 0.07 | | | rs382804 | 0.1 | 1 | 879101 | 879101 | G |
| 9 | exonic | KLHL17 | synonymous SNV | NM_198317:p.A203A | | | 0.92 | 0.45 | 0.95 | rs4970441 | | 1 | 887188 | 887188 | G |
| 10 | exonic | KLHL17 | synonymous SNV | NM_198317:p.L239L | 569 | | 0.05 | 0.29 | | rs6696971 | | 1 | 887601 | 887601 | C |
| 11 | exonic | PLEKHN | synonymous SNV | NM_001160184:p.A166A | | | | | | rs28507236 | | 1 | 896135 | 896135 | A |
| 12 | exonic | PLEKHN | nonsynonymous SN | NM_001160184:p.R452P | | | 0.54 | 0.86 | 0.89 | rs382974 | 0 | 1 | 899101 | 899101 | G |
| 13 | exonic | PLEKHN | synonymous SNV | NM_001160184:p.D512D | | | 0.18 | | 0.12 | rs28548431 | | 1 | 899282 | 899282 | C |
| 14 | exonic | ISG15 | nonsynonymous SN | NM_005101:p.S83N | | | 0.4 | 0.56 | 0.21 | rs1921 | 0.38 | 1 | 939471 | 939471 | G |
| 15 | exonic | ISG15 | synonymous SNV | NM_005101:p.V98V | | | 0.99 | 0.3 | 0.96 | rs8997 | | 1 | 939517 | 939517 | A |
| 16 | exonic | AGRN | synonymous SNV | NM_198576:p.S1022S | | | 0.93 | | | rs2465128 | | 1 | 971794 | 971794 | A |

# Synapse- from Sage Bionetworks

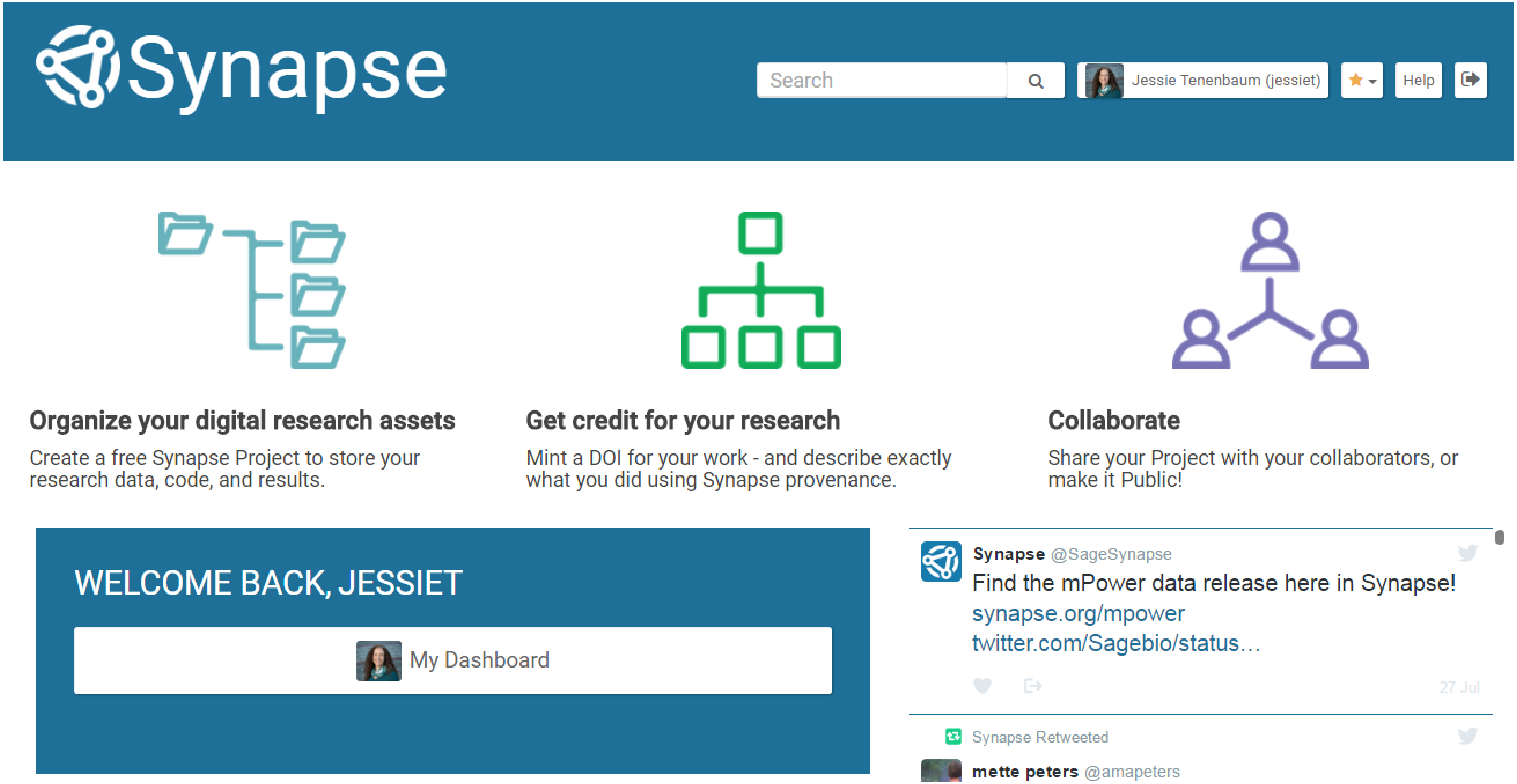


Overview

## Accelerating Open Biomedical Research

**Over the next decade, ever-expanding data will transform biomedical research approaches and feed healthcare discoveries through the use of computational models to predict outcome and responses to treatment.**

At Sage Bionetworks we believe that this advance will be best harnessed when individuals and groups can collaborate openly on discoveries, with a fundamental shift in the traditional roles and rewards for individuals and organizations involved. **We work to redefine how complex biological data is gathered, shared and used, redefining it through open systems, incentives, and norms.** We challenge the traditional roles of individuals and groups, patients and researchers.



**Organize your digital research assets**

Create a free Synapse Project to store your research data, code, and results.

**Get credit for your research**

Mint a DOI for your work - and describe exactly what you did using Synapse provenance.

**Collaborate**

Share your Project with your collaborators, or make it Public!

WELCOME BACK, JESSIET

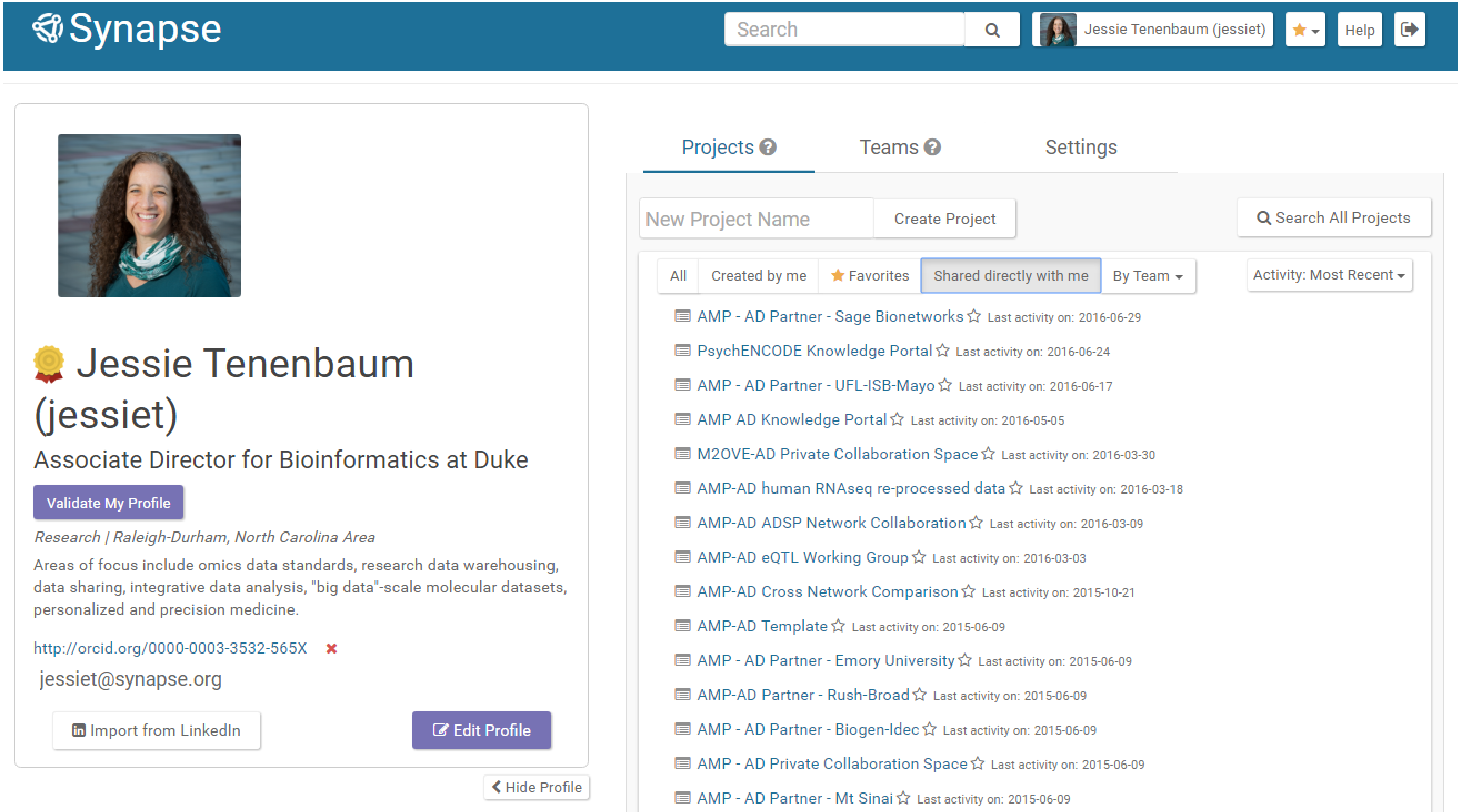
**Synapse** @SageSynapse

Find the mPower data release here in Synapse! synapse.org/mpower twitter.com/Sagebio/status…

27 Jul

Synapse Retweeted

**mette peters** @amapeters



# Jessie Tenenbaum (jessiet)

Associate Director for Bioinformatics at Duke

*Research | Raleigh-Durham, North Carolina Area*

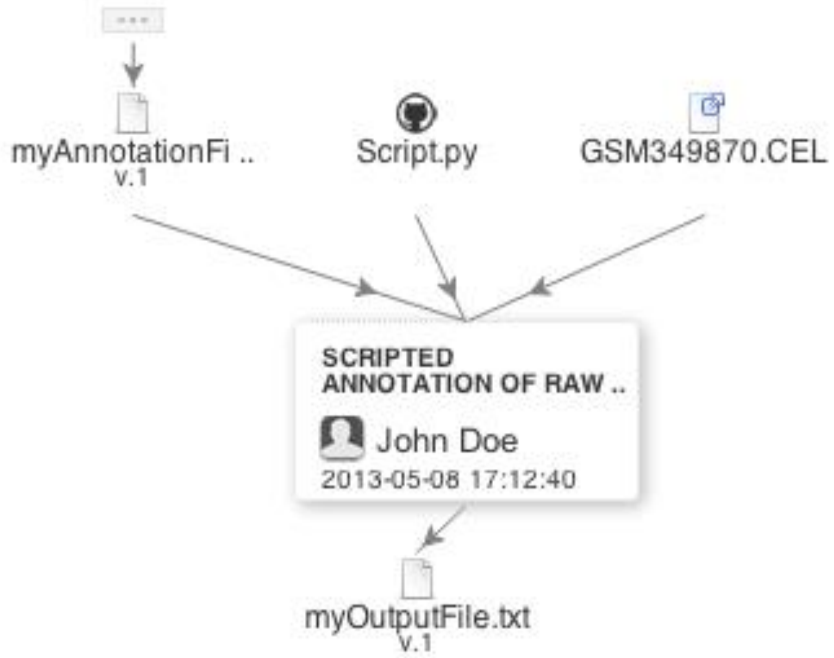
Areas of focus include omics data standards, research data warehousing, data sharing, integrative data analysis, "big data"-scale molecular datasets, personalized and precision medicine.

http://orcid.org/0000-0003-3532-565X ✖

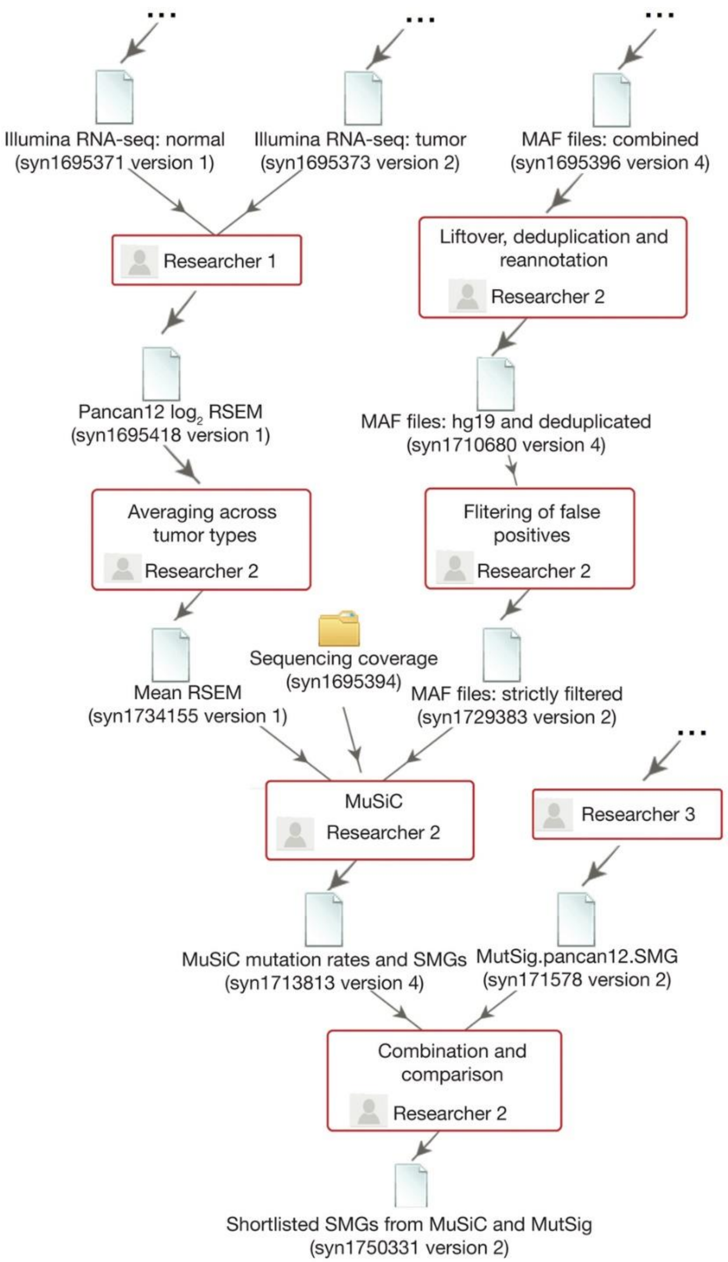
jessiet@synapse.org

- AMP - AD Partner - Sage Bionetworks ☆ Last activity on: 2016-06-29
- PsychENCODE Knowledge Portal ☆ Last activity on: 2016-06-24
- AMP - AD Partner - UFL-ISB-Mayo ☆ Last activity on: 2016-06-17
- AMP AD Knowledge Portal ☆ Last activity on: 2016-05-05
- M2OVE-AD Private Collaboration Space ☆ Last activity on: 2016-03-30
- AMP-AD human RNAseq re-processed data ☆ Last activity on: 2016-03-18
- AMP-AD ADSP Network Collaboration ☆ Last activity on: 2016-03-09
- AMP-AD eQTL Working Group ☆ Last activity on: 2016-03-03
- AMP-AD Cross Network Comparison ☆ Last activity on: 2015-10-21
- AMP-AD Template ☆ Last activity on: 2015-06-09
- AMP - AD Partner - Emory University ☆ Last activity on: 2015-06-09
- AMP-AD Partner - Rush-Broad ☆ Last activity on: 2015-06-09
- AMP - AD Partner - Biogen-Idec ☆ Last activity on: 2015-06-09
- AMP - AD Private Collaboration Space ☆ Last activity on: 2015-06-09
- AMP - AD Partner - Mt Sinai ☆ Last activity on: 2015-06-09

# Provenance graphs in Synapse



https://sagebionetworks.jira.com/wiki/display/PLFM/Analysis+Provenance+in+Synapse



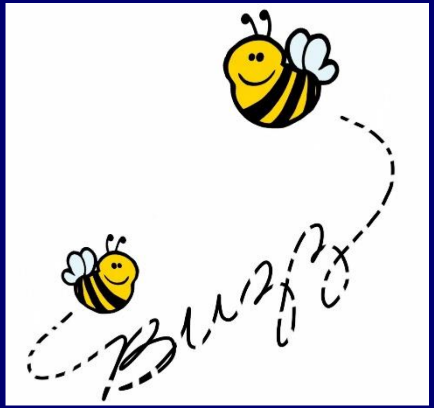
Omberg *et al*. Nat Gen. 2013

# Reproducible research: a cautionary tale

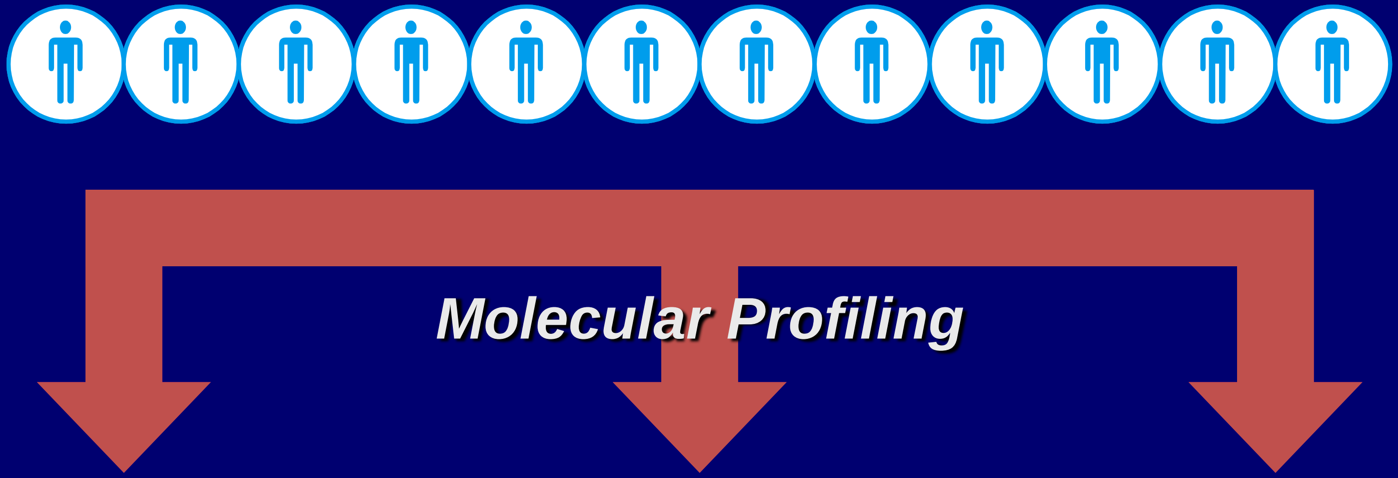
# WHAT ALL THIS ENABLES…

# P* Medicine?

- Personalized medicine
- P4: predictive, preventive, personalized, and participatory (Hood)
- Aka stratified, genomic, individualized…
- Precision medicine
- Each a different emphasis, but fundamentally: use more data to deliver
  - – the right intervention
  - – for the right person
  - – at the right time



# ‘-omics’ Technologies Can Help stratify a seemingly homogeneous population.



**Diabetes**

***Molecular Profiling***

**Exercise +**
**Diet A**

**Responder**

**Exercise +**
**Diet B**

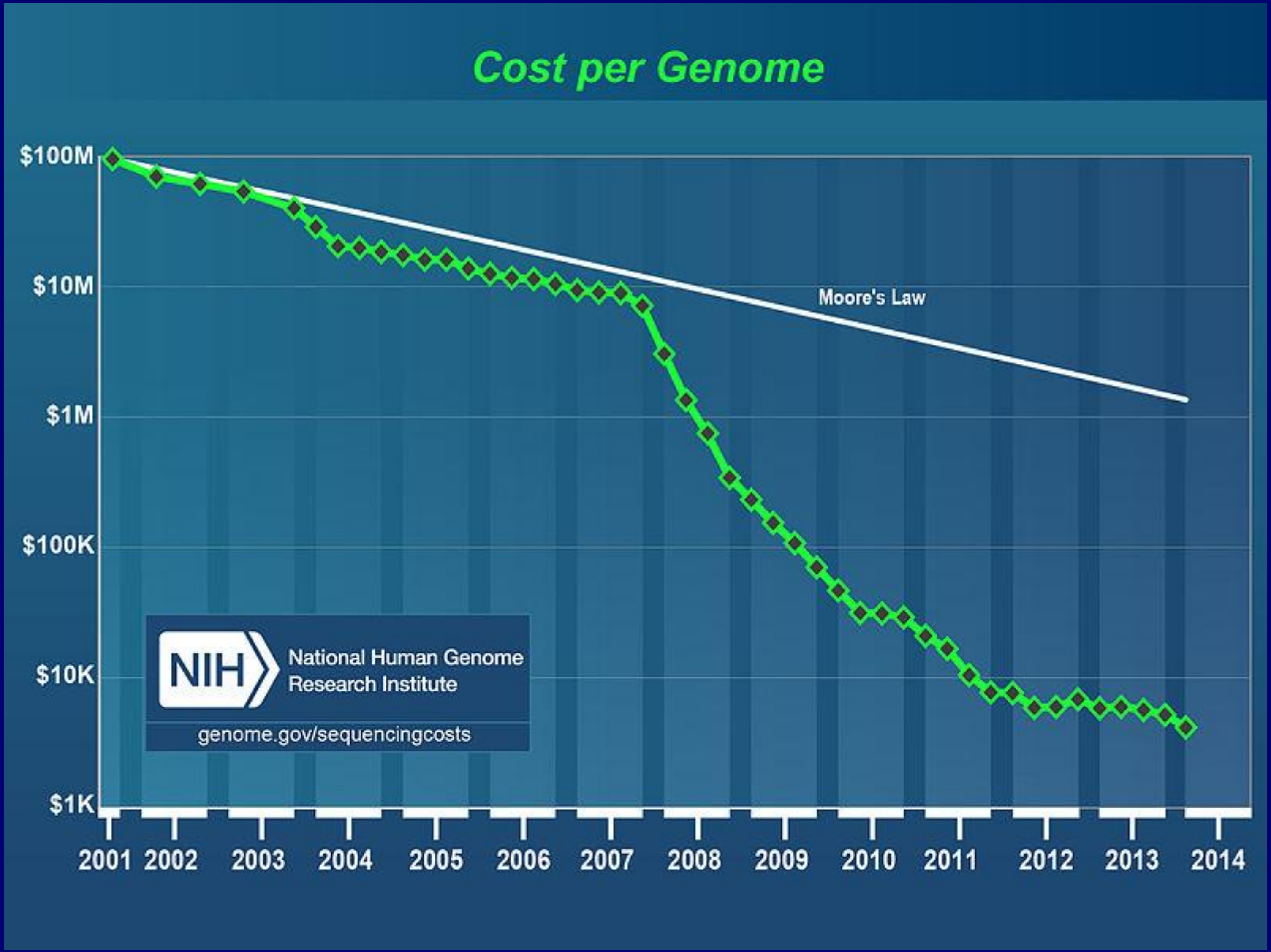
**Adverse event**

**Exercise +**
**Diet +**
**Medication**

**Non-responder**

Courtesy of G Ginsburg, Duke University

# Decreasing costs of technology

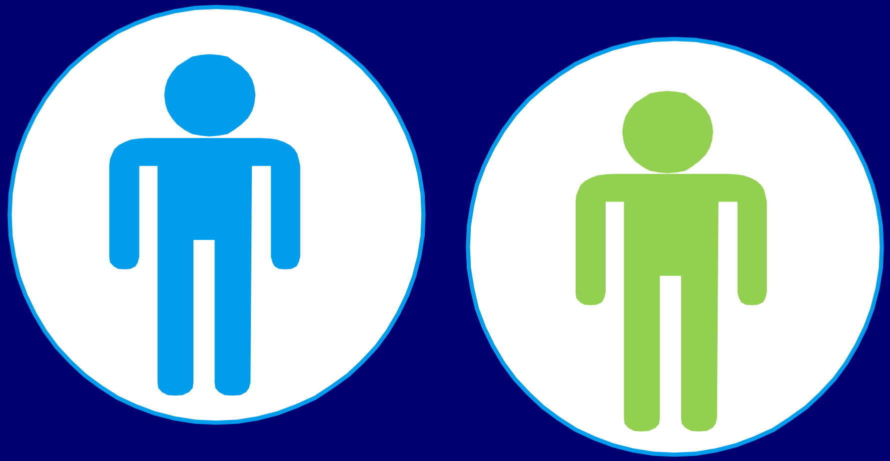


Duke Medicine

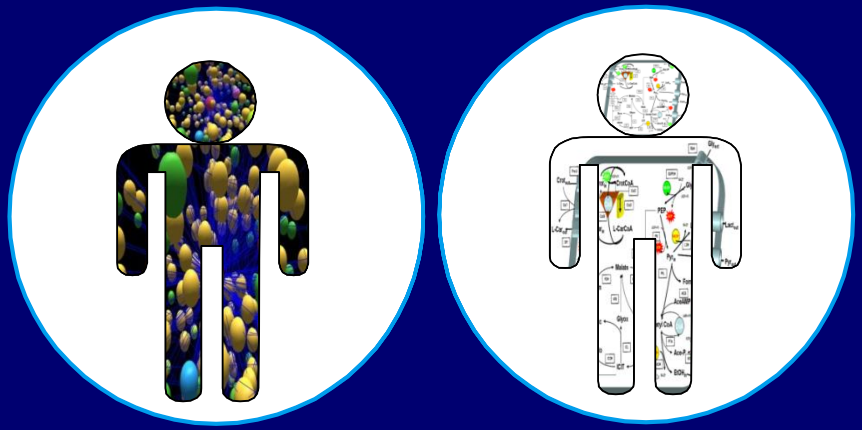
# A New Taxonomy for disease

From macroscopic observation to underlying molecular basis

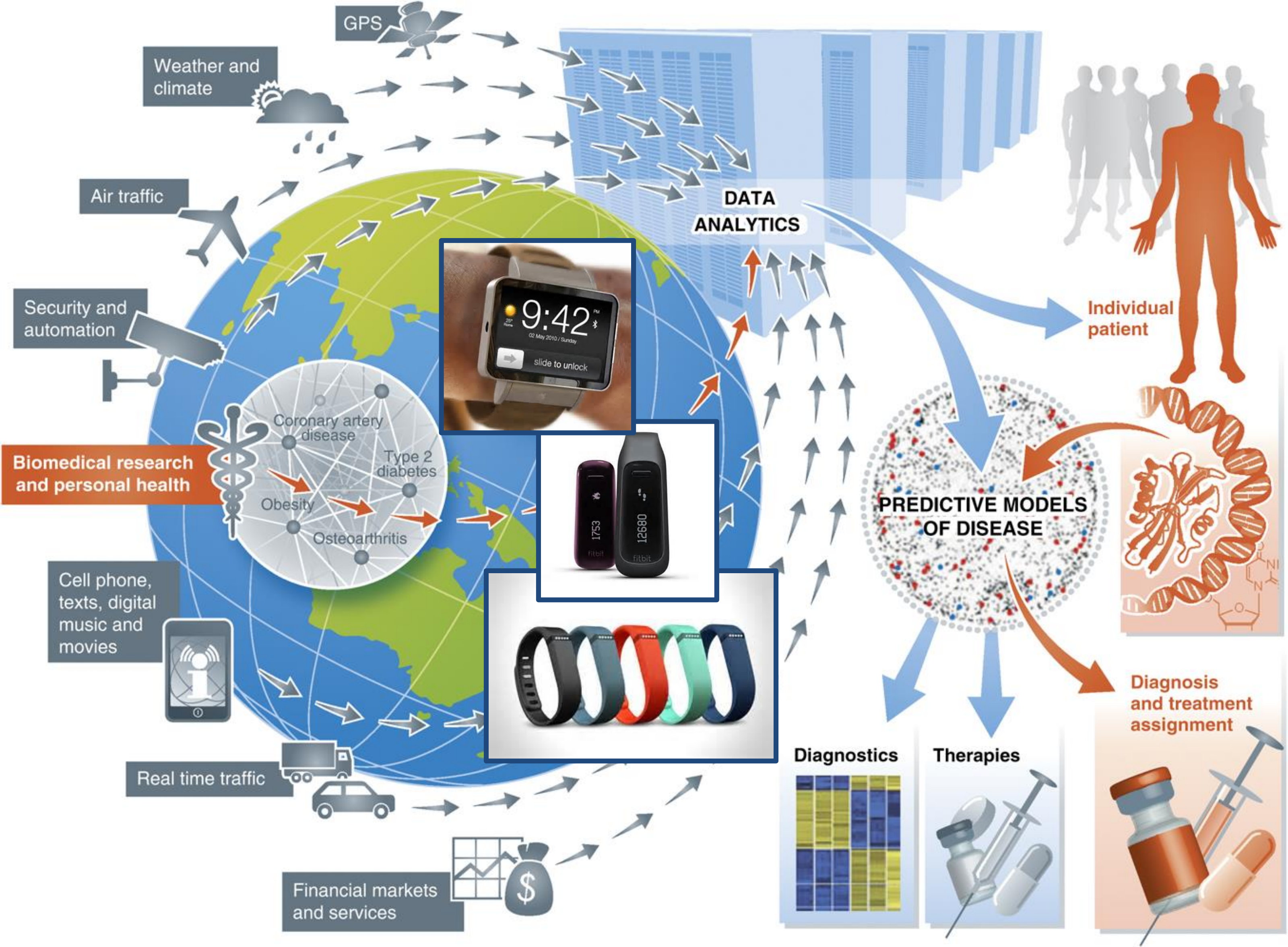
From this…



To this



# Non-traditional data sources



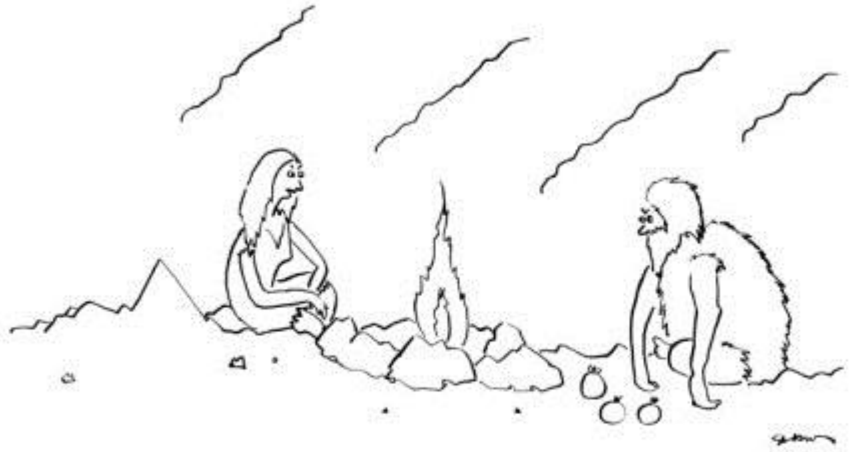
Schadt, *Mol Sys Bio*, 2012

# GENOMICS IN CLINICAL CARE

http://researcher.watson.ibm.com/

# Pharmacogenomics:
# “Low hanging fruit”

- Good tools for detection
- Genome is relatively stable
- Mutations not selected against



*"Enough with the low-hanging fruit. How about some slow-moving meat?"*

# PGX by the numbers

- 241 FDA recognized pharmacogenomic biomarkers
- 55 require genetic testing
- 4 recommend genetic testing
- 103 are “actionable”
- 61 are “informative”

Source: PharmGKB, as of July 20, 2017

# Nicholas Volker

Worthey et al. *Genet Med.* 2011

- Mysterious bowel condition
- Rule out numerous diseases
- Resort to exome sequencing: 16,000 mutations
- Causal mutation discovered, verified *in vitro*
  - Gene: XIAP
  - Diseases: XLP *and* new one
- Bone marrow transplant treatment for XLP

- Pulitzer Prize for explanatory reporting: “One in a Billion”

# Actionable infection diagnosis by NGS

Wilson et al. *NEJM*, 2014

- 14-year-old boy with severe combined immunodeficiency (SCID)
- Repeat hospital admissions- headache, fever, etc.→ bad stuff→ medically induced coma
- Diagnostic workup "unrevealing"
- NGS of CSF yielded 475 leptospira reads
- Targeted antibiotics→ recovery
- Note: standard Dx assay for leptospira depends on host response

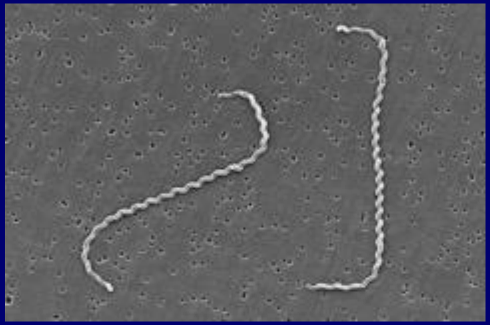


Image source: Wikipedia.org

# Clinical Actionability of FoundationOne testing

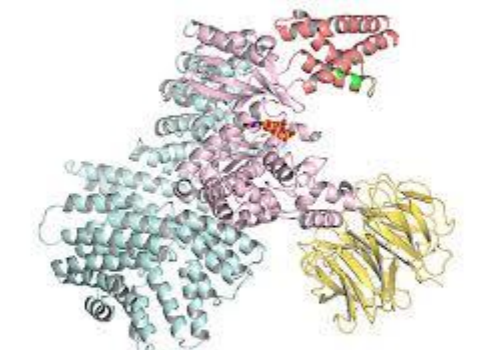
Johnson et al. *The Oncologist*, 2014

- Retrospectively assessed demographics, genomic test results, therapies received (N = 103)
- Co-primary endpoints: % of patients with actionable results, % receiving genotype-directed therapy.
- 83% had potentially actionable genotypes
- 21% received genotype-directed treatment
- Relatedly: Priority Health first US health insurance plan to cover FoundationOne test, Oct 2014

# Repeated tumor WGS and drug sensitivity, resistance

Wagle et al. *NEJM* 2014

- 57-year-old woman with Anaplastic Thyroid Cancer
- Everolimus inhibits mTOR, effective for tumors with mTOR mutation.
- Sequencing of tumor before resistance revealed mutation in TSC2, a negative regulator of mTOR- could explain her response to everolimus, an allosteric inhibitor of mTOR.
- Drug-resistant tumor- mTOR mutation that leads to resistance to allosteric mTOR inhibitors
  - though not to mTOR kinase inhibitors, suggesting an avenue for further treatment.

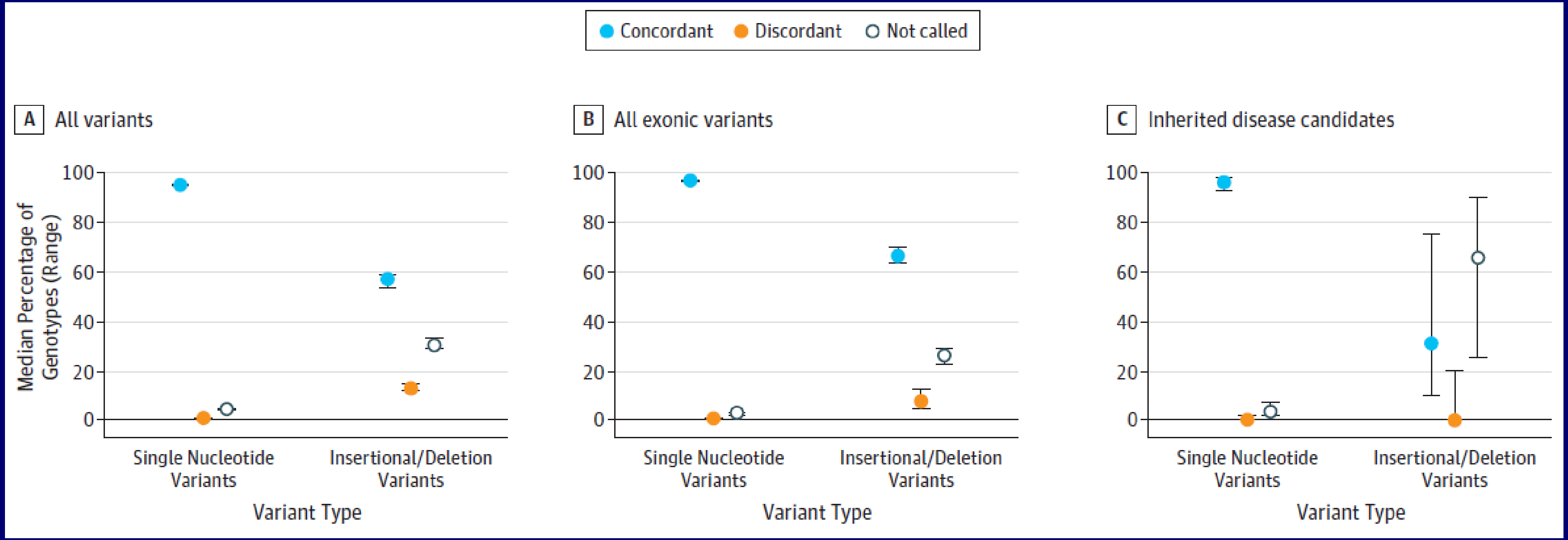


http://www.mskcc.org

# Clinical Interpretation and Implications of Whole-Genome Sequencing

Dewey et al. *JAMA* 2014

- WGS on 12 participants, including 9 by multiple technologies
- 10-19% of inherited disease genes not covered to accepted standards for SNP discovery



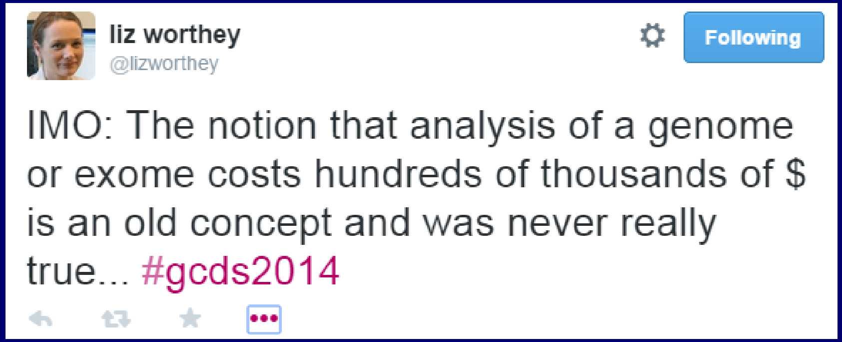
Duke Medicine

# Dewey et al. cont.

- Consideration of a median of 1 to 3 initial diagnostic tests and referrals per participant
- “fair” inter-rater agreement (Fleiss κ = 0.24) about suitability of findings for clinical follow-up across all findings
- Inter-rater agreement worse than random for cardiometabolic disease risk scores (Fleiss κ = -0.03)
- Estimated median cost for sequencing and variant interpretation $14,815 plus computing infrastructure and data storage.

# About that $100k interpretation…

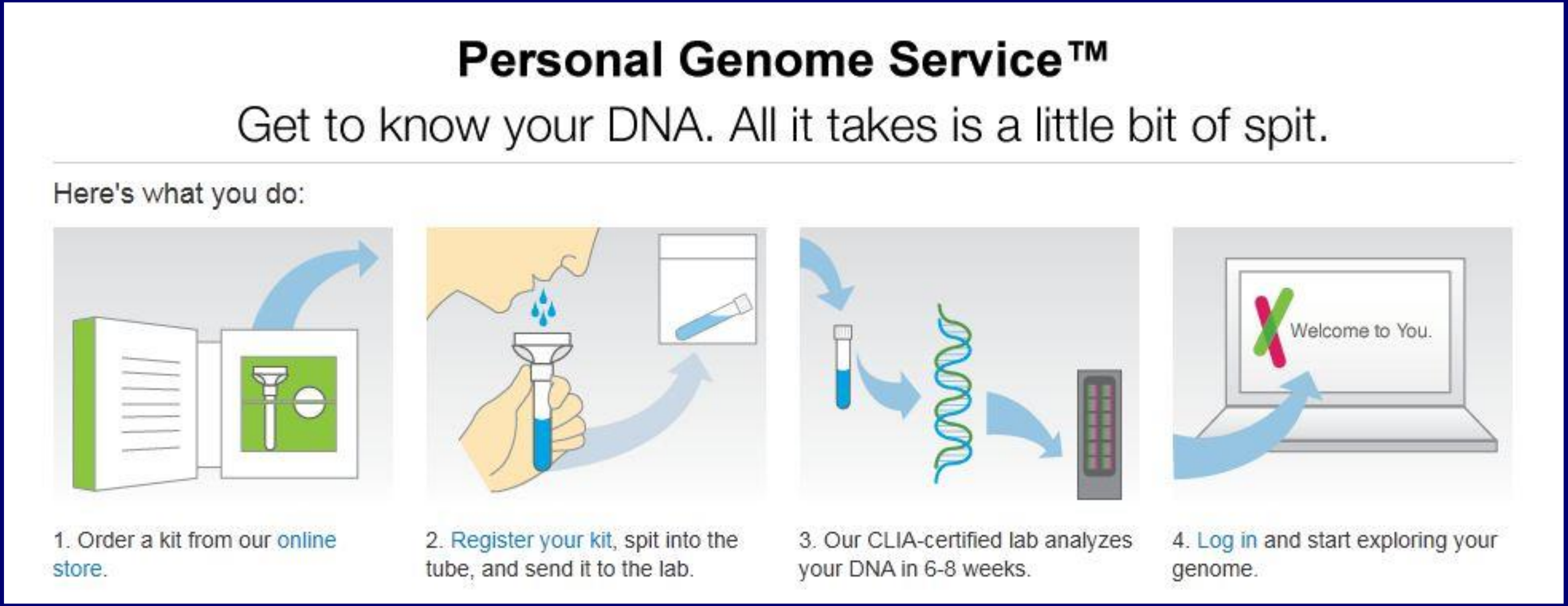
- Ashley et al. *Lancet* 2010: Quake genome
  - – Hundreds of PhD level person-hours!
  - – Personalis.com
- Nic Volker lead author



IMO: The notion that analysis of a genome or exome costs hundreds of thousands of $ is an old concept and was never really true... #gcds2014

- “Artisanal” to “factory”
  - – Google’s David Glazer via GenomeWeb

# Direct to Consumer (DTC) Genetic Testing


**Personal Genome Service™**

Get to know your DNA. All it takes is a little bit of spit.

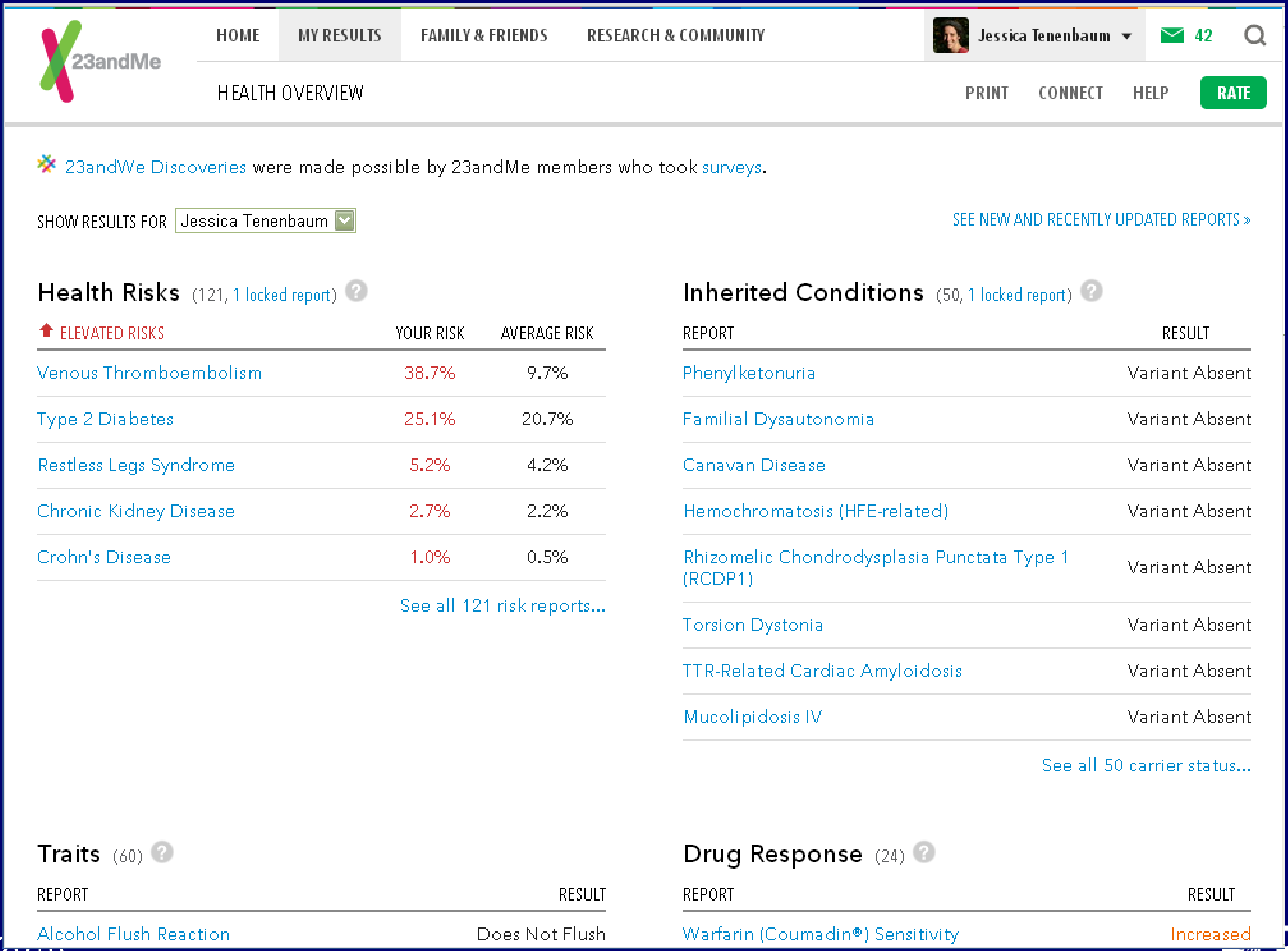
Here's what you do:

1. Order a kit from our online store.
2. Register your kit, spit into the tube, and send it to the lab.
3. Our CLIA-certified lab analyzes your DNA in 6-8 weeks.
4. Log in and start exploring your genome.

http://www.singularityweblog.com/23andme-dna-test-review-its-right-for-me-but-is-it-right-for-you/

# DTC Genetic Testing

- 2013- 23andMe was last one standing in US
- FDA ordered them to stop advertising and offering health-related information (ancestry ok)
- Started bringing back one test at a time with FDA approval
- Now approved to provide specific carrier status results
- ☑ Ethical, Legal, Social Issues (ELSI)
  - – Regulated by government?
  - – Are consumers ready?
  - – Are providers ready?



23andWe Discoveries were made possible by 23andMe members who took surveys.

SHOW RESULTS FOR Jessica Tenenbaum

SEE NEW AND RECENTLY UPDATED REPORTS »

## Health Risks (121, 1 locked report)

| ELEVATED RISKS | YOUR RISK | AVERAGE RISK |
|---|---|---|
| Venous Thromboembolism | 38.7% | 9.7% |
| Type 2 Diabetes | 25.1% | 20.7% |
| Restless Legs Syndrome | 5.2% | 4.2% |
| Chronic Kidney Disease | 2.7% | 2.2% |
| Crohn's Disease | 1.0% | 0.5% |

See all 121 risk reports...

## Inherited Conditions (50, 1 locked report)

| REPORT | RESULT |
|---|---|
| Phenylketonuria | Variant Absent |
| Familial Dysautonomia | Variant Absent |
| Canavan Disease | Variant Absent |
| Hemochromatosis (HFE-related) | Variant Absent |
| Rhizomelic Chondrodysplasia Punctata Type 1 (RCDP1) | Variant Absent |
| Torsion Dystonia | Variant Absent |
| TTR-Related Cardiac Amyloidosis | Variant Absent |
| Mucolipidosis IV | Variant Absent |

See all 50 carrier status...

## Traits (60)

| REPORT | RESULT |
|---|---|
| Alcohol Flush Reaction | Does Not Flush |

## Drug Response (24)

| REPORT | RESULT |
|---|---|
| Warfarin (Coumadin®) Sensitivity | Increased |

© J. Tenenbaum 2016

Duke Medicine

# User-friendly graphical results

**Jessica Tenenbaum**

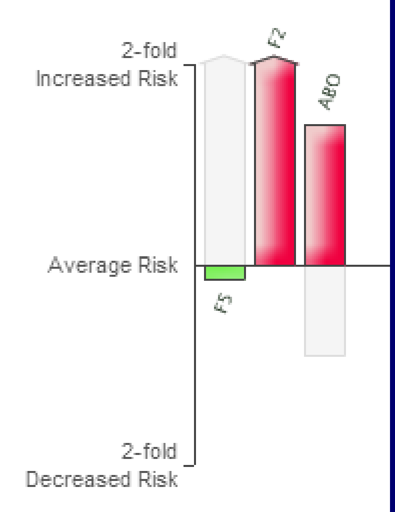
**38.7 out of 100**

women of European ethnicity who share Jessica Tenenbaum's genotype will develop Venous Thromboembolism between the ages of 0 and 79.

**Average**

**9.7 out of 100**

women of European ethnicity will develop Venous Thromboembolism between the ages of 0 and 79.



# Precision medicine from the pin cushion perspective

Tenenbaum et al., *J Pers Med*. 2012

- 35 years old female
- Heterozygous prothrombin gene mutation (rs1799963) aka factor 2
  - Present in ~2% of population
- Pregnant with twins
- Mutation validated through doctor- part of medical record

- Recommendation: anti-coagulant throughout pregnancy


NDC 0075-0621-60
LOVENOX®
(enoxaparin sodium injection)
60mg/0.6mL
[100mg/mL]
Rx ONLY
Lovenox

# Other factors for consideration

- No pe
- Infertil
- “Advar
- Econo
- Profes
- Fear o

© J. Tenenbaum 2016

# 23andMe cont.

- Patient applied for life insurance
- Annual rate >2x what it would have been without "downstream effects" from DTC testing
- GINA- Genetic Information Nondiscrimination Act
  - covers employment and health insurance, NOT life, disability, or long term care insurance

- Is that wrong?

# QUESTIONS?

jessie.tenenbaum@duke.edu