

# Agenda

1:10 – 1:30 PM

Working together to improve genomic data sharing in 2020 and beyond – *Carolyn Hutter, NHGRI*

1:30 – 1:50 PM

Themes from Journal Participants' Responses – *Chris Gunter, NHGRI*

1:50 – 2:50 PM

Open Discussion – *Moderated by Chris Gunter and Veronique Kiermer, PLOS*

2:50 – 3:00 PM

Summarize Next Steps – *Carolyn Hutter, NHGRI*

# Discussion Instructions

- Please introduce yourself [and your institutional affiliation]
- Use the “Raise Hand” function
  - If you have a direct response/comment to make, you can unmute and speak
- Non-Journal representatives are encouraged to speak up, too
- Feel free to use the chat, but we encourage oral comments

# Working together to improve genomic data sharing in 2020 and beyond

Carolyn Hutter, Ph.D.

Division Director, Division of Genome Sciences

November 30, 2020



National Human Genome  
Research Institute

—  
The **Forefront**  
of **Genomics**  
—



# NHGRI Guiding Principles and Values for Human Genomics Includes Emphasis on Data Sharing

- **Adhere to the highest expectations and requirements related to open science, responsible data sharing, and rigor and reproducibility in genomics research**
- — the genomics enterprise has a well-respected history of leading in these areas, and that commitment must be built upon and continually reaffirmed.

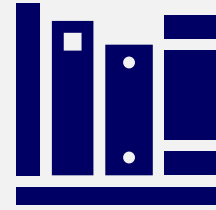
<https://www.genome.gov/2020SV>



Research Institutions



Funders



Journals

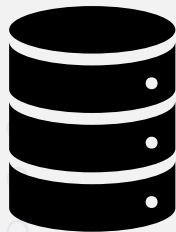


Standards-  
Generating  
Bodies

# Who are the players?



IRBs



Repositories



Data Generators



Data Users



Participants

1999

NIH Research  
Tools Policy

2003

NIH Data Sharing  
Policy

2004

NIH Model Organism  
Sharing Policy

2008

NIH Genome-Wide  
Association Studies  
(GWAS) Policy

2008

NIH Public Access  
Policy

2013

White House Initiative  
(2013 "Holdren Memo")

2014

NIH Genomic Data  
Sharing (GDS) Policy

2013

Big Data to Knowledge  
(BD2K)

2015

NIH Intramural Human  
Data Sharing Policy

2018

Genomic  
Summary  
Results (GSR)  
Update

2017

HHS Rule and NIH  
Policy on Clinical Trial  
Results

2020

NIH Data Management  
and Sharing Policy



2000

2005

2010

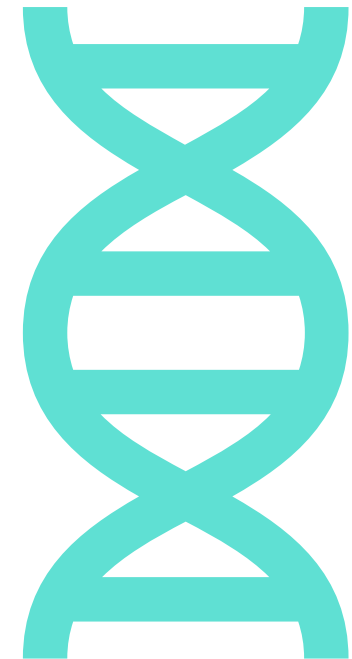
2015

2020

# Evolution of NIH Data Sharing Policies

# The 2014 Genomic Data Sharing (GDS) Policy

- Sets forth expectations that ensure the broad and responsible sharing of genomic research data
- Responsibilities of Investigators Submitting Genomic Data
  - Genomic Data Sharing Plans
  - Non-human Genomic Data
  - Human Genomic Data
- Responsibilities of Investigators Accessing and Using Genomic Data
  - Requests for Controlled-Access Data
  - Terms and Conditions for Research Use of Controlled-Access Data
  - Conditions for Use of Unrestricted-Access Data
- Intellectual Property



# NHGRI Implementation of the NIH GDS Policy



- NHGRI encourages sharing of all genomic data and data types
- Human and Non-human data submission timelines are the same
- NHGRI encourages human studies to use:
  - sources with consent for general research uses through controlled access
  - sources with consent for unrestricted

<https://www.genome.gov/about-nhgri/Policies-Guidance/Genomic-Data-Sharing>



# **\*New\*** NIH Data Management and Sharing Policy

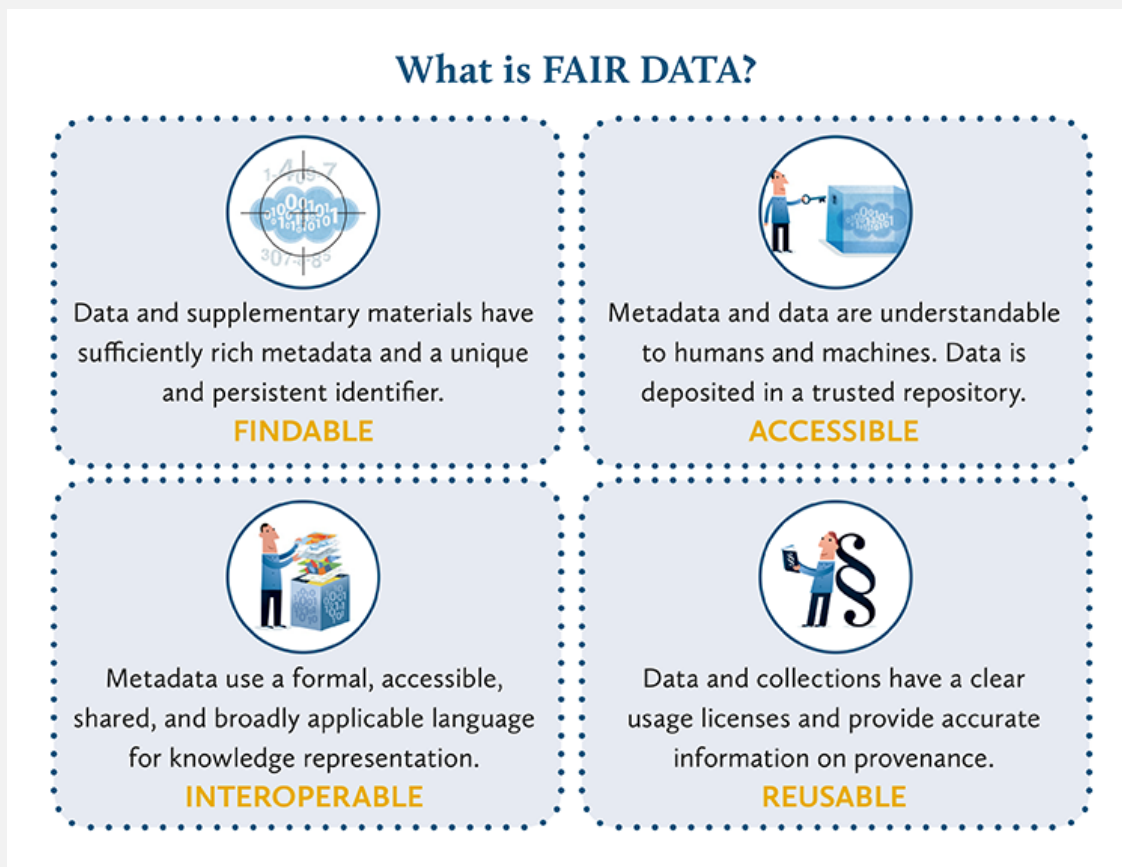
- Applies to all research funded or conducted by NIH that results in the generation of scientific data.
- Effective Date: January 25, 2023
- Two main requirements:
  - Submission of a Data Management and Sharing Plan upon submitting a grant application
  - Compliance with the approved Plan



**'Plans to develop how scientific data generated by research projects will be managed and whether these scientific data accompany research will be shared'**

# The FAIR Guiding Principles

A C G  
C G T  
A C G



Metadata is key!

**Metadata =**  
Data that provide additional information intended to make scientific data interpretable and reusable

Photo Credit: Institute of Mathematical Statistics ([link](#))

# NHGRI Plans to Increase Emphasis on Metadata/Phenotypic Data

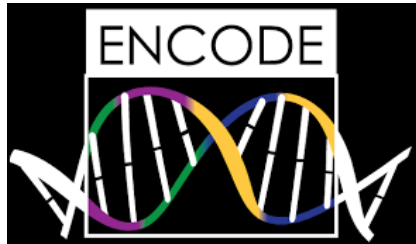
Notice to announce an effort at NHGRI to improve the availability and quality of 'relevant associated data,' as it is referred to in the NIH Genomic Data Sharing (GDS) Policy (e.g., metadata and phenotypic data)

NHGRI-funded and supported researchers will be expected to:

1. Share the metadata and phenotypic data associated with the study.
2. Use standardized data collection protocols and survey instruments for capturing data, as appropriate.
3. Use standardized notation for metadata (e.g., controlled vocabularies or ontologies) to enable the harmonization of datasets for secondary research analyses.

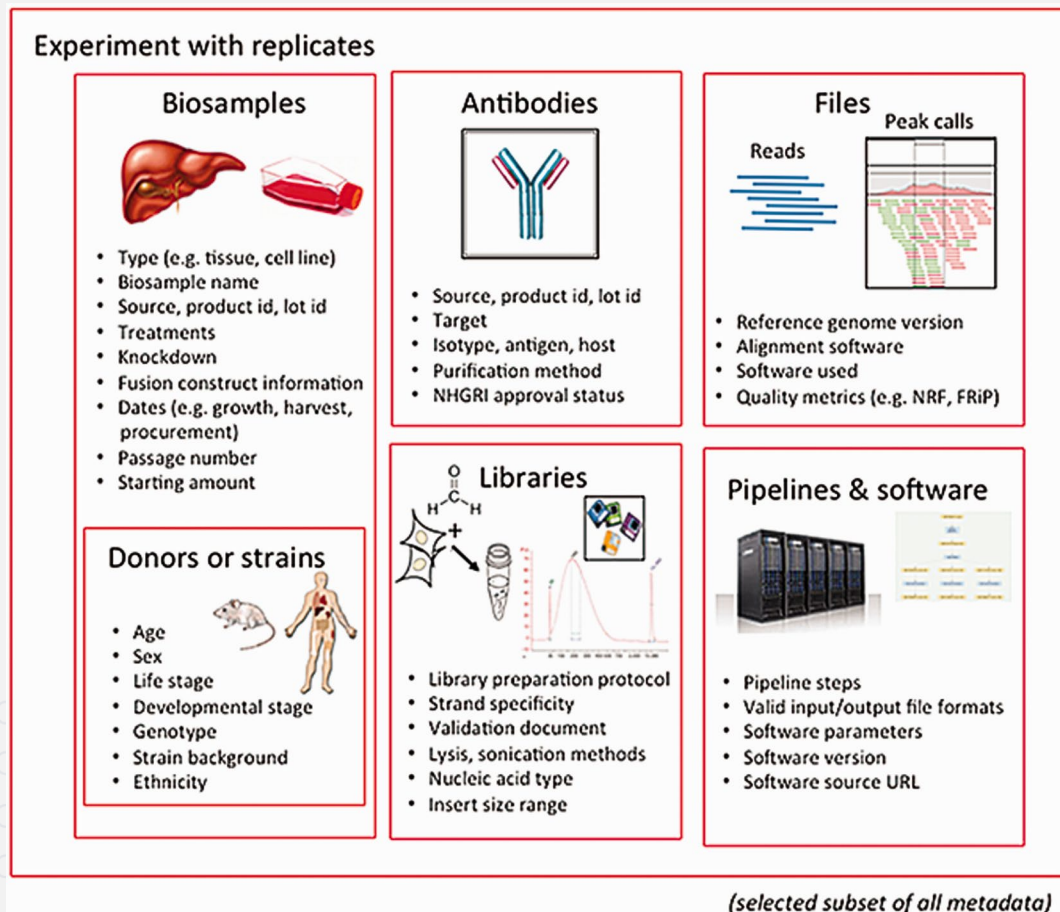


# Moving Policies to Practice

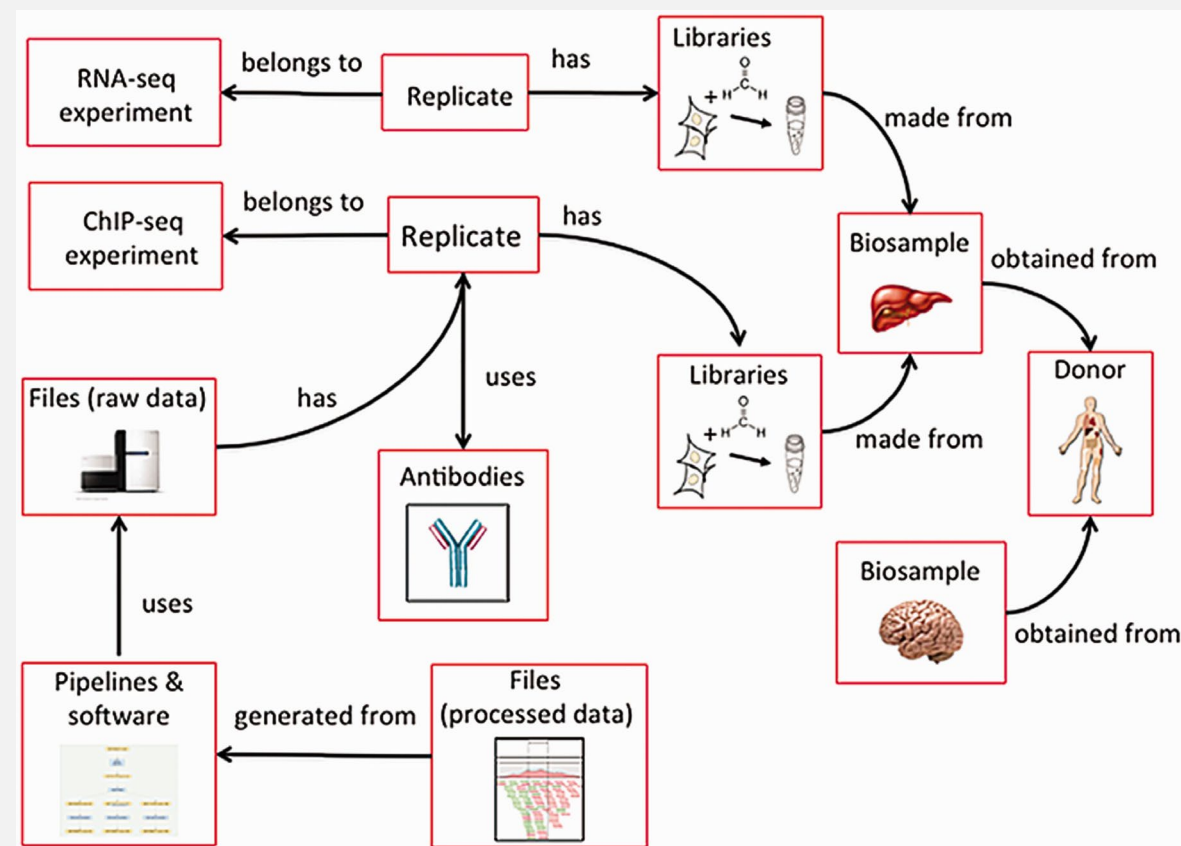


**Large, open genomics resource-building projects:**  
- develop and disseminate standards for metadata

# ENCODE Metadata Categories



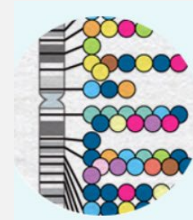
Major categories of metadata.



Categories in the metadata model are linked to each other

# Data sharing highlights

- Pre-publication availability of GWAS, in response to journals, reviewers, and authors
  - 4,728 unpublished studies (Oct. 2020)
  - On top of 9,406 published studies
- Summary statistics availability
  - Increasing availability over time
  - Increasing proportion that are author-submitted (45% of 2019 and 90% of 2020 studies)

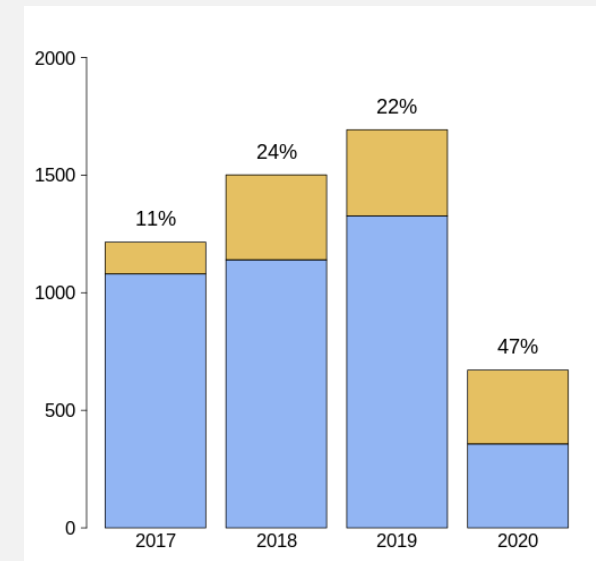
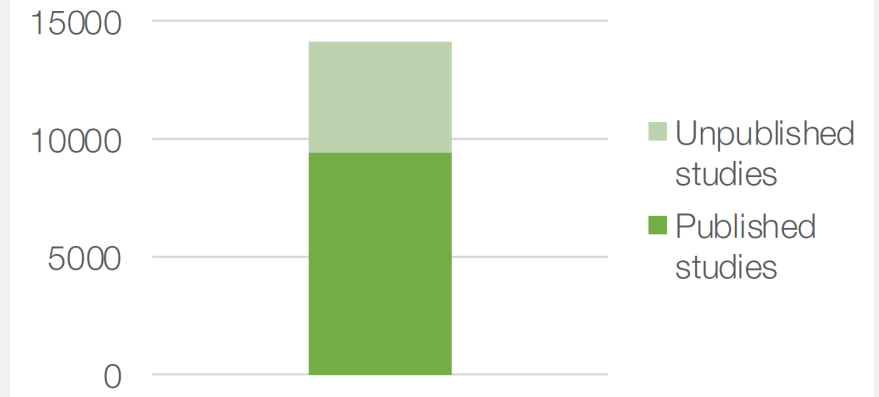


## GWAS Catalog

The NHGRI-EBI Catalog of published genome-wide association studies

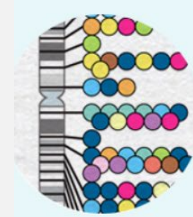
Search the catalog

Examples: breast carcinoma, rs7329174, Yao, 2q37.1, HBS1L, 6:16000000-25000000





# Data sharing highlights



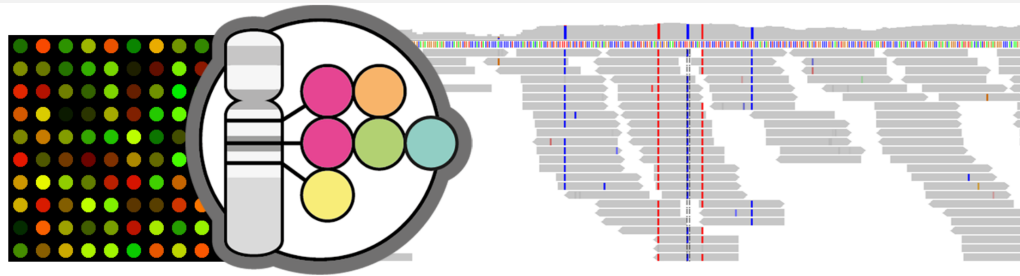
## GWAS Catalog

The NHGRI-EBI Catalog of published genome-wide association studies

Search the catalog

Examples: breast carcinoma, rs7329174, Yao, 2q37.1, HBS1L, 6:16000000-25000000

- Expanding into new data types and formats
  - Sequencing studies
  - Polygenic Score (PGS) Catalog



Current scope

- Array-based genotyping

Expanded scope

- Sequencing-based genotyping
  - More variants (rare)
  - Single variant and multi-variant analysis

PGS Catalog

Search...

breast cancer, glaucoma, EFO\_0001645

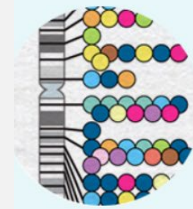
PGS Catalog / Traits / EFO\_0000305

**Trait: breast carcinoma**

Experimental Factor Ontology (EFO) Information	
Identifier	EFO_0000305 <a href="#">↗</a>
Description	A carcinoma that arises from epithelial cells of the breast [MONDO: DesignPattern]
Trait category	<b>Cancer</b>
Synonyms	<a href="#">17 synonyms</a> <a href="#">+</a>
Mapped term(s)	<a href="#">11 mapped terms</a> <a href="#">+</a>
Child trait(s)	<a href="#">7 child traits</a> <a href="#">+</a>



# Data sharing highlights



## GWAS Catalog

The NHGRI-EBI Catalog of published genome-wide association studies

Search the catalog

Examples: breast carcinoma, rs7329174, Yao, 2q37.1, HBS1L, 6:16000000-25000000

- Summary statistics workshop (June 2020)
- Attendees: cohort representatives, summary statistics users, tools developers, resource providers, journal editors, funders

GWAS Catalog recognised as the central resource for all human GWAS



GWAS SumStats and metadata submitted to the GWAS Catalog at the time of submission or sharing



ACCESSION ID

Guidance provided regarding the risks associated with sharing and how these risks can be mitigated



GWAS SumStats and metadata versioned in a way that enables users to identify the most recent dataset



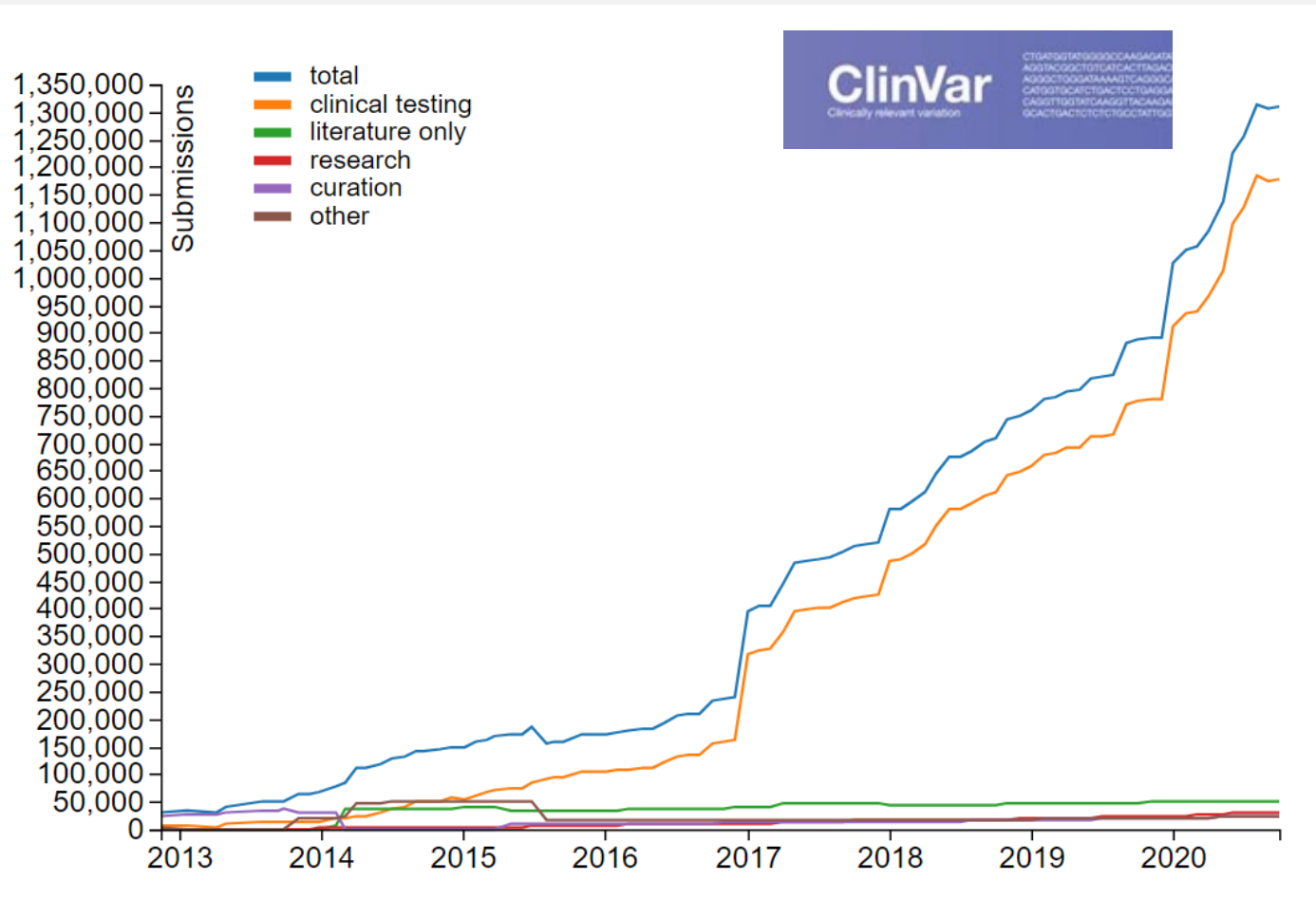
GWAS SumStats and metadata linked to other relevant datasets/resources



GWAS SumStats made available following the FAIR indicators and include standard elements



# Quality of variant interpretations improved by data sharing



- ClinGen-ClinVar partnership led to > 1.35M sequence variant interpretations shared by ClinVar
- Making interpretations available enables resolution of conflict discrepancies among labs and better care for patients

**Human Mutation**  
Variation, Informatics, and Disease



RESEARCH ARTICLE














## Scaling resolution of variant classification differences in ClinVar between 41 clinical laboratories through an outlier approach

Steven M. Harrison, Jill S. Dolinsky, Wenjie Chen, Christin D. Collins, Soma Das, Joshua L. Deignan, Kathryn B. Garber, John Garcia, Olga Jarinova, Amy E. Knight Johnson ... [See all authors](#)

First published: 11 October 2018 | <https://doi.org/10.1002/humu.23643> | Citations: 19

# Clinical Labs Recognized for Meeting Data Sharing Requirements to Support QA



Laboratory	Meets Requirements	Additional Achievements		
		>95% from past 5 years <sup>1</sup>	Discrepancy resolution <sup>2</sup>	Consenting mechanism <sup>3</sup>
Ambry	✓			
ARUP	✓			
Athena Diagnostics Inc.	✓			
Center for Pediatric Genomic Medicine, Children's Mercy Hospital and Clinics	✓			
Color Genomics, Inc.	✓			
GeneDx	✓			

Recognized labs have:

- Been CLIA certified
- Submitted to ClinVar as 'Single Submitter'
- Submitted at least 100 variants
- Submitted new variants at least once a year
- Submitting at least 95% of all sequence and/or CNV variants reported in the past two years

<https://www.clinicalgenome.org/tools/clinical-lab-data-sharing-list/>

# Areas for improvement



A C G  
C G T  
A C G

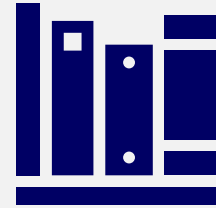
- Where are all the data going? Are people supplying sufficient metadata?
- Consistent metadata, truly FAIR data
- Standardization and dissemination of code, pipelines, and analytical tools
- Accounting for changing landscapes (e.g., cloud-computing, addressing emerging technologies)
- Modernizing approaches to data management and sharing
- Others?



Research Institutions



Funders



Journals



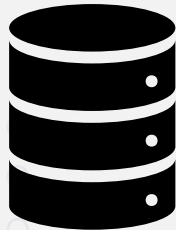
Standards-  
Generating  
Bodies

A C G  
C G T  
A C G

# Working together



IRBs



Repositories



Data Generators



Data Users



Participants

# Forefront of Genomic Data Sharing

