Community perspective: How do research communities help demonstrate and maximize the utility of a resource and the data it holds.

How can metrics promote usage and utility of a resource, and justification for continued support?

J. Brian Byrd, MD, MS
Assistant Professor of Internal Medicine
University of Michigan

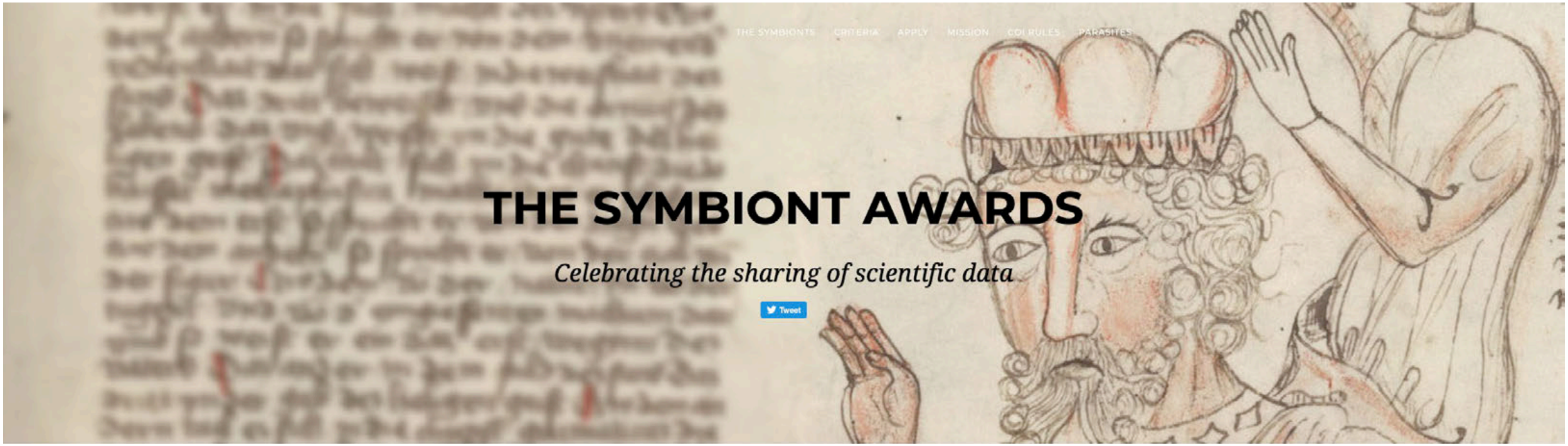
My perspective

Member of the research community

Physician-scientist

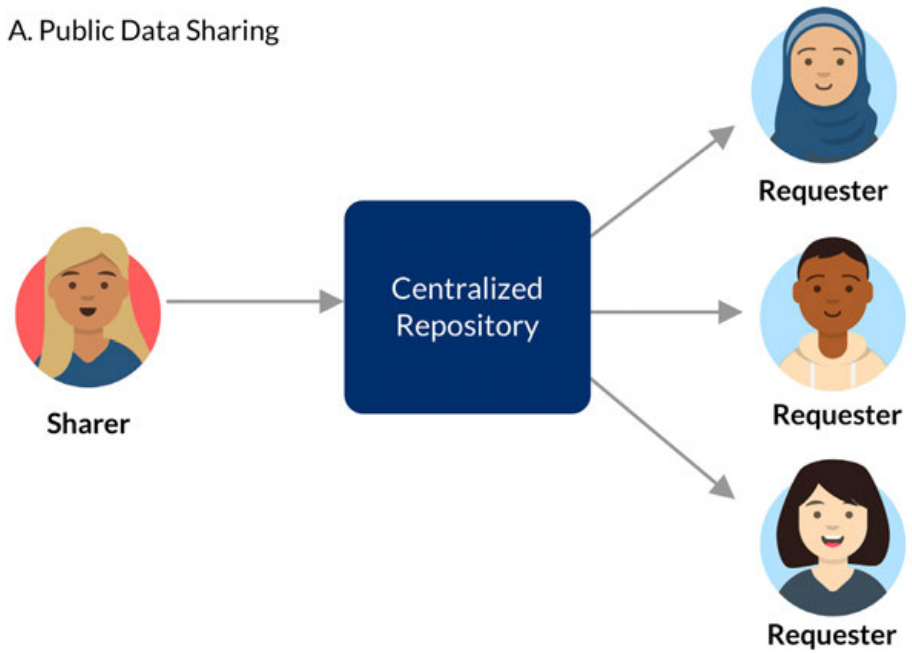
- Phase I-IV clinical trials, with a significant bench component
  - Biomarkers
- Observational studies
- Data re-user

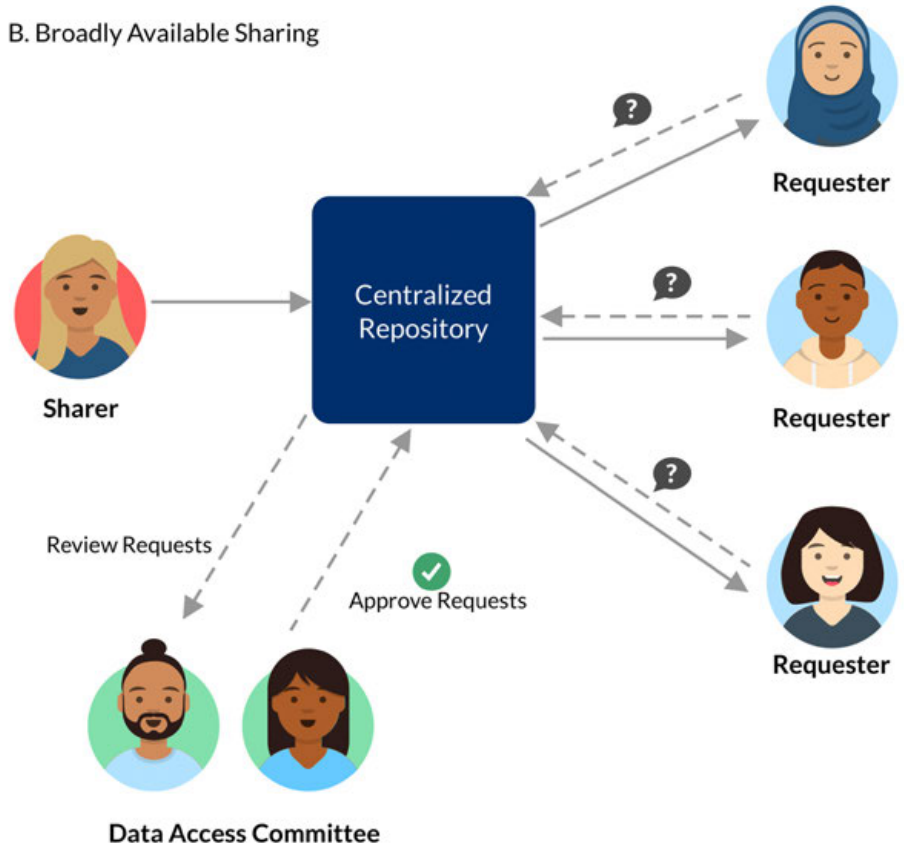
Founder
Research Symbiont Awards for excellence in data sharing



https://researchsymbionts.org/

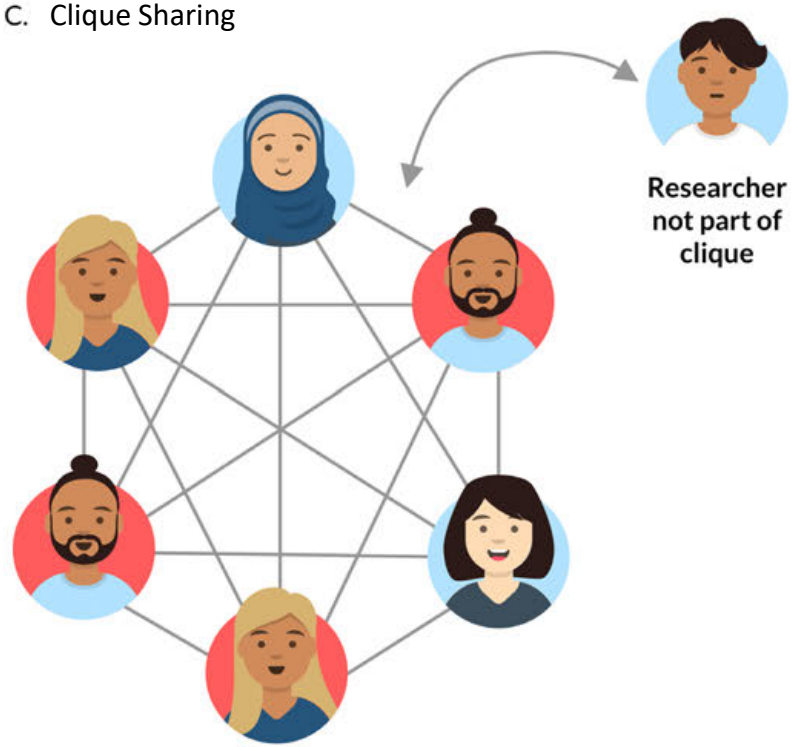
Current situation: diverse sharing arrangements



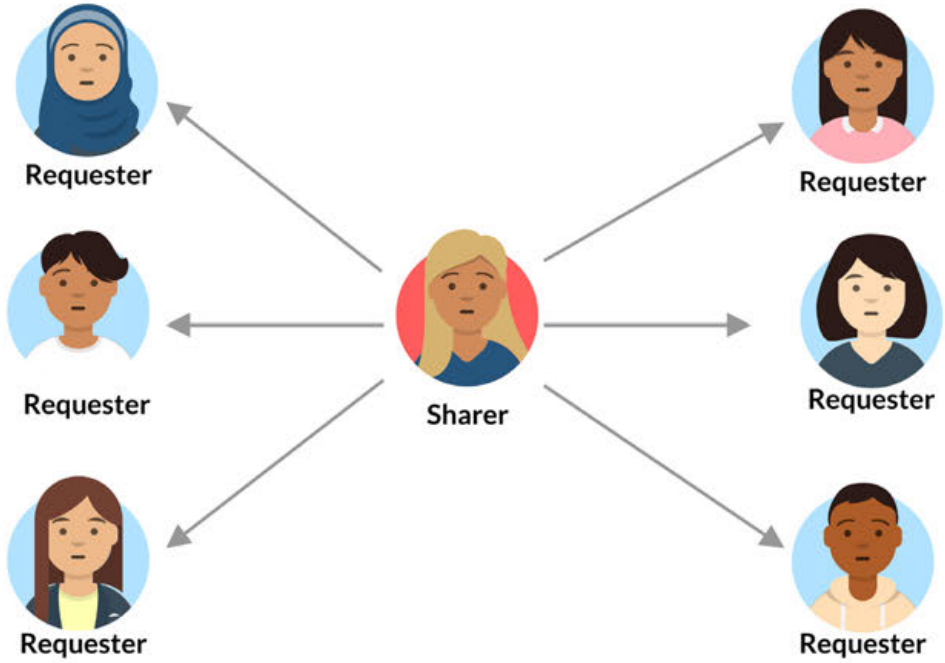


Figures by Deepa Prasad for our manuscript under review

C. Clique Sharing



D. Sharing Upon Request



Figures by Deepa Prasad for our manuscript under review

Limitations of directly reciprocal sharing

Scales poorly since parties' interests must align & both parties must be aware of that alignment

No reason to believe aligned interests are required for excellent science to result from data re-use

> The data might be used to *answer questions outside zone of interest* of the team generating the data

Difficult or unreasonable conditions could be placed on users of data

Sharing data without expectation of direct benefit avoids these problems

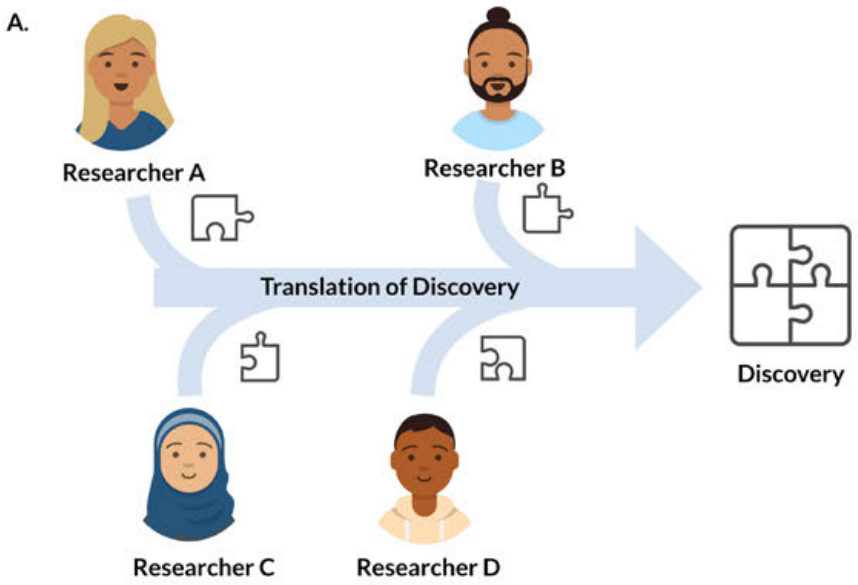
But sharing of this type is likely to stably, frequently occur only if there is an expectation of *indirect benefit*

What is the desired future state?

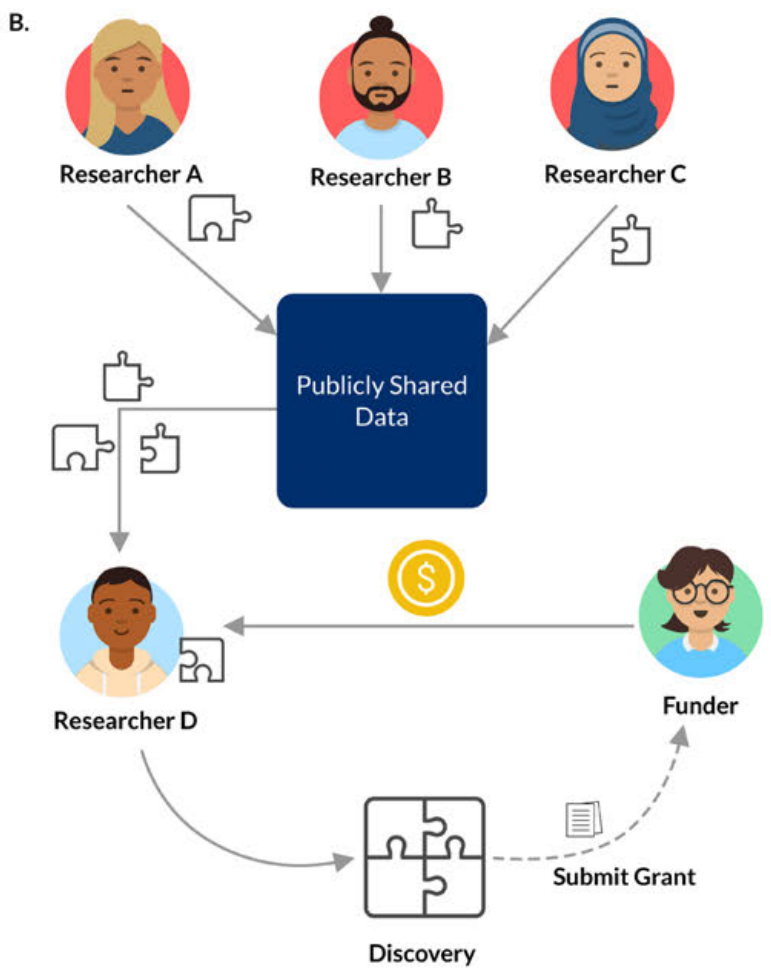
Less clique, more click-to-download

(i.e., more public or broad sharing)

Why do we need a *metric*?



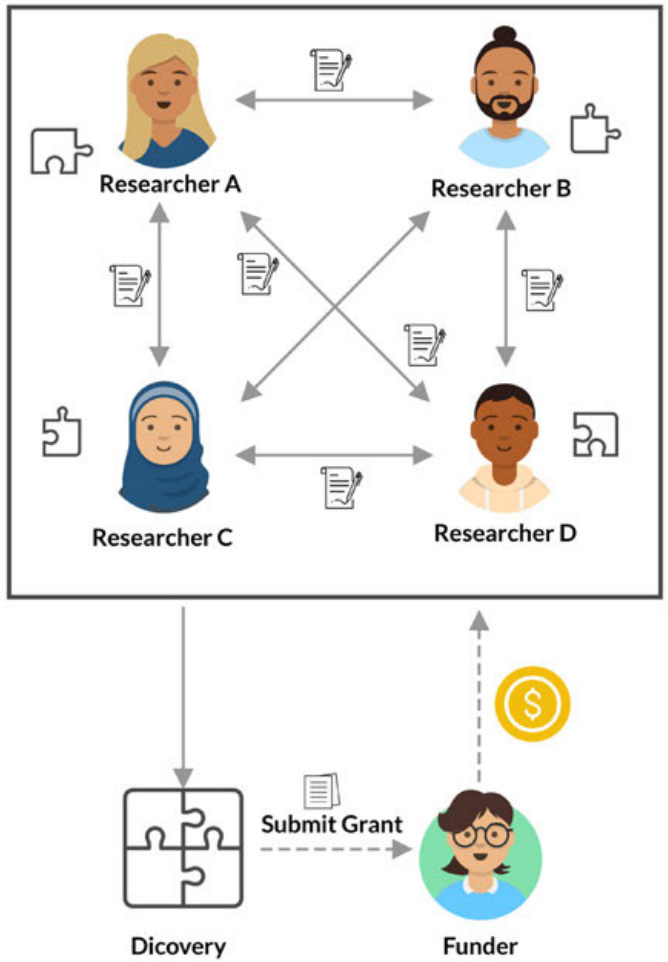
Researcher work products from multiple groups need to be combined to produce a discovery that improves human health.



If funders allocate credit without considering sharing behavior in a system of open sharing, much of the credit and funding can accrue to the researcher who brings the final component that enables translation.

Figures by Deepa Prasad for our manuscript under review

C.

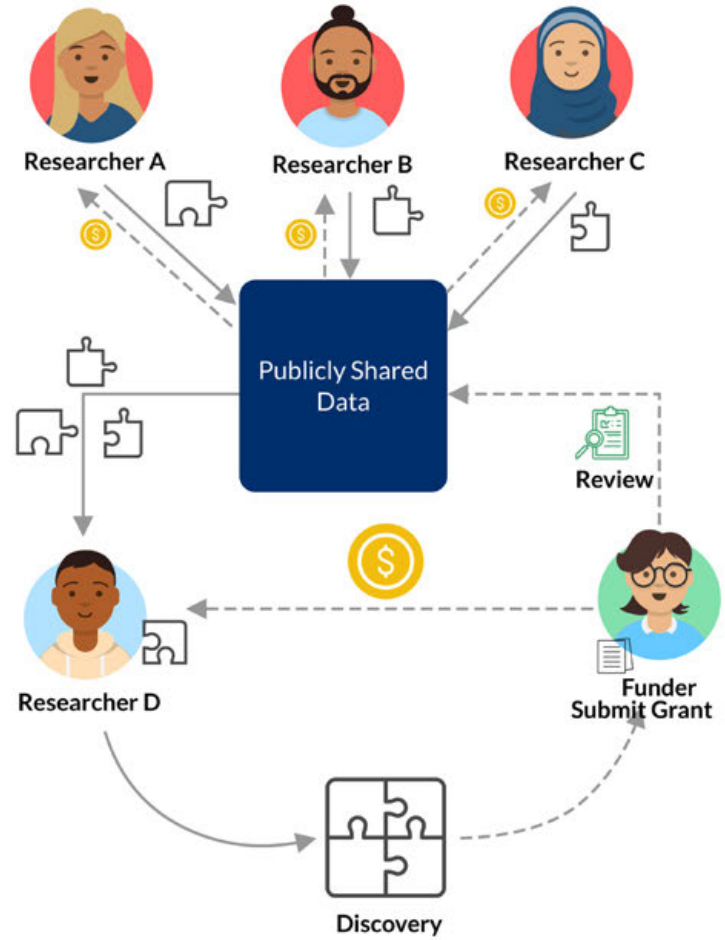


Researchers can restrict sharing and negotiate agreements through consortia to enhance the equity of credit distribution, but negotiating agreements is time consuming and may delay or prevent advances.

Key

Material Transfer Agreement

D.



Funders who consider the value of shared resources when assessing impact provide a benefit not only to the researcher bringing the final component but also all others on the value chain.

Figures by Deepa Prasad for our manuscript under review

A *reputation* for sharing must improve one's lot in life for sharing to be frequent and stable

For researchers, this can be reduced in practice to an improved chance of funding

The researchers who judge funding applications may not know each applicant's personal reputation for sharing

Thus, a metric or judging rubric is required

Criteria can be devised to identify and reward great sharers

Case study 1: S.K. Morgan Ernest, PhD

Associate Professor, University of Florida

Openly sharing data in ecology, organismal traits, and life history for over decade

During grad school & post-doc, assembled a dataset shared as a data paper

- Cited >120 times, mostly for data re-use
- Re-use of data in papers in *Science*, *Nature*, *PNAS*

Additional sharing of subsequent datasets

This type of sharing should influence chance of funding since it *amplifies the impact* of the research funding Dr. Ernest received

Case study 2: Fabio Zanini, PhD

Post-doctoral fellow, Stanford University

At Max Planck, studied evolution of the HIV genome in patients over time spans up to 15 years

His group deep sequenced the virus

Uploaded to SRA, but felt more needed to be done to make the data understandable

https://hiv.biozentrum.unibas.ch/

HiV EVO

Intrapatient HIV evolution

Deep longitudinal data

- 10 patients
- more than 80 time points
- defined time of infection
- 4.5 - 16 years follow-up without therapy
- 6 - 12 samples per patient
- whole genome
- coverage >1000 with quantified template input

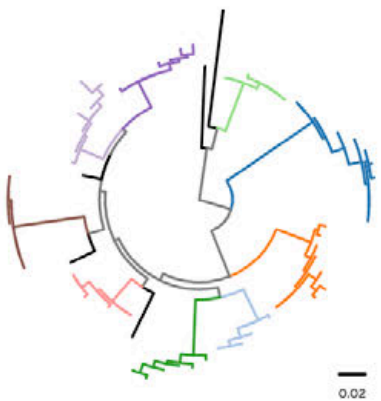
How to access and explore

- browse data by patient
- compare patients and genomic regions
- the sequencing methods and quality controls
- download the cleaned reads

Table: Overview of patient specific data.

Patient	Subtype	# samples	1st sample	Last sample
			[days since infection]	
p1	AE	12	122	2996
p2	B	6	74	2018
p3	B	10	146	3079
p4	B	8	78	3054
p5	B	7	134	2149
p6	C	7	62	2556
p7	B	11	1905	5811
p8	B	7	87	2208
p9	B	8	106	2955
p10	B	9	33	2256
p11	B	7	209	2043

Example: Phylogeny of samples in the p17 region, colored by patient.



Reference sequences are colored black.

Case study 3: Leonardo Collado-Torres, PhD

Staff scientist, Johns Hopkins

Lead R developer for recount2, which synthesized, uniformly processed, and made available over 70,000 public human RNA-seq samples

Over 8 TB of data

46 publications had cited the paper describing this R package

https://jhubiostatistics.shinyapps.io/recount/

Case study 4: Brian Bot

Curator of the mPower Public Researcher Portal, Sage Bionetworks

One of the first large-scale attempts to assess the feasibility of quantifying Parkinson disease symptoms and their changes in a ‘real world setting’

First six months of data made available quickly

Years before the manuscript analyzing these data was submitted

Data were collected with an informed consent process that allowed participants the choice to determine whether their data was (1) shared only with the study team; or (2) shared broadly with qualified researchers worldwide

229 researchers had gone through qualified researcher process, gaining access

Case study 5: Alexander LeNail

At time of nomination:

PhD student, MIT

Built a data portal to share data from 1000 ALS patients

Collected, identically pre-processed, and systematically harmonized approximately 400TB of diverse biomolecular data

http://data.answerals.org/

Each case study was selected using unified criteria: a potential starting point for a metric

Did this person create an openly shared scientific resource or dataset beyond typical standards of their field?

Was the sharing mechanism clearly permissible per all applicable ethical or legal restrictions, e.g., informed consent document?

Was the sharing mechanism as *easy for people who wish to use the data as is feasible* within ethical and legal constraints?

Additional suggested criteria for evaluating data sharing

Was the dataset remarkable for its richness, granularity, and quality, such that it is inviting to people who wish to use the data?

Is there evidence that a conflict of interest limits the data sharing?

Were the data effectively re-used to answer questions not addressed in an initial publication reporting the dataset or data notification?

How clear is the publicly available audit trail of decisions potentially affecting people who wish to use the data?

These criteria have been adapted for use by a foundation

Rubric for Reviewers:

Please use the full range of scores (1-9) for this criterion. We expect that very few applications will receive a perfect score in this area.

General Track Record:

- Do the applicants have a track record of sharing resources that are remarkable for their richness, granularity, or quality such that those resources are particularly inviting to people who wish to use them.
- Do the applicants have a track record of sharing resources in a manner that is as easy as possible for people to re-use within ethical and legal constraints.
- Have the applicants shared resources that have *already been reused* by other investigators to answer a new question?
- Early Career Grants: Young Investigator, 'A' or Psychosocial Launch. Applicants are encouraged to describe past experience; however, it is understood they may not have a track record. The reviewer should focus on the Sharing Plan.

General Resource Sharing Plan:

- Do the authors use an established repository for the resource? (See AHA guidelines on repositories for questions https://goo.gl/2UCZ43). A lab website is not acceptable.)
- Is the resource distributed in a way that maximally facilitates reuse?
- Will the resource as described have sufficient metadata available to promote reuse?
- For resources that must be maintained, is there a plan in place to maintain the resource?

Data Sharing:

- Public, widely-used repositories should be used if possible (e.g., GEO or ArrayExpress for gene expression data, SRA for RNA-Seq data, etc.).
- If no public, widely-used repository is available for the data type in question, a general purpose archival repository (e.g., FigShare, Zenodo) should be used.
- For more detailed discussion, the guidelines provided by F1000 research for authors are an excellent resource:

https://www.alexslemonade.org/researchers-reviewers/applicants
https://www.alexslemonade.org/sites/default/files/resource_sharing_form_all_grants_final_11.25.19.docx

ALSF asks applicants to provide information

FORM (1-page maximum)

Data Sharing:

- *Highlight how you have shared data publicly – i.e., not upon request – and how those data have been reused. Illustrate with reuse metrics such as citation counts, downloads, or other such data if available.*
- *Discuss how you plan to share the outputs from this proposal and how the data will be archived (via the recognized repository for the type of data or, for data without such a repository, via Zenodo, FigShare, or similar archival services). How will data be licensed (i.e., CC0 or another license). You must discuss how and when data that you generate during the course of this project will be shared. If access will be controlled via a data access committee or other such structure, describe the conditions under which data will be shared and specify how relevant metrics (number of requests made, number of requests approved, time to respond to requests) will be stored and reported to us and the scientific community.*

Protocol Sharing:

- *Highlight how you have shared protocols openly – i.e., not upon request – and how those protocols have been used by others. For example, you may have posted them to protocols.io or a similar service.*
- *Discuss how and when you plan to share the outputs from this proposal. Not all projects will result in protocols. If yours does not, this section can be deleted.*

Material and Reagent Sharing:

- *Highlight how you have shared materials and reagents and how those reagents have been reused.*
- *Discuss how and when you plan to share the reagents and materials developed in your group as part of the proposal*

More characteristics of a good sharing metric

Would not be limited to a particular type of artifact

- Data
- Derivative models (e.g., machine learning models)
- Code
- Transgenic animals
- Cell lines
- Other unique reagents

More characteristics of a good sharing metric

Not easily evaded

> If a history of failure to cooperate rather than cooperating can be hidden, then the metric will create problems

Persistent

As objective as possible

More characteristics of a good sharing metric

Low burden for research applicants

Low burden for study section members

Goodhart's Law 'attack surface' is well understood

"When a metric becomes a target, it ceases to be a good metric."

Challenge the community to help uncover the problems likely to arise

Good *use* of a sharing metric

Influence the probability of future funding