Crimes against data sharing in functional genomics

Emma Bell, PhD

Princess Margaret Cancer Centre, Toronto

emma.bell@uhnresearch.ca

@emmabell42

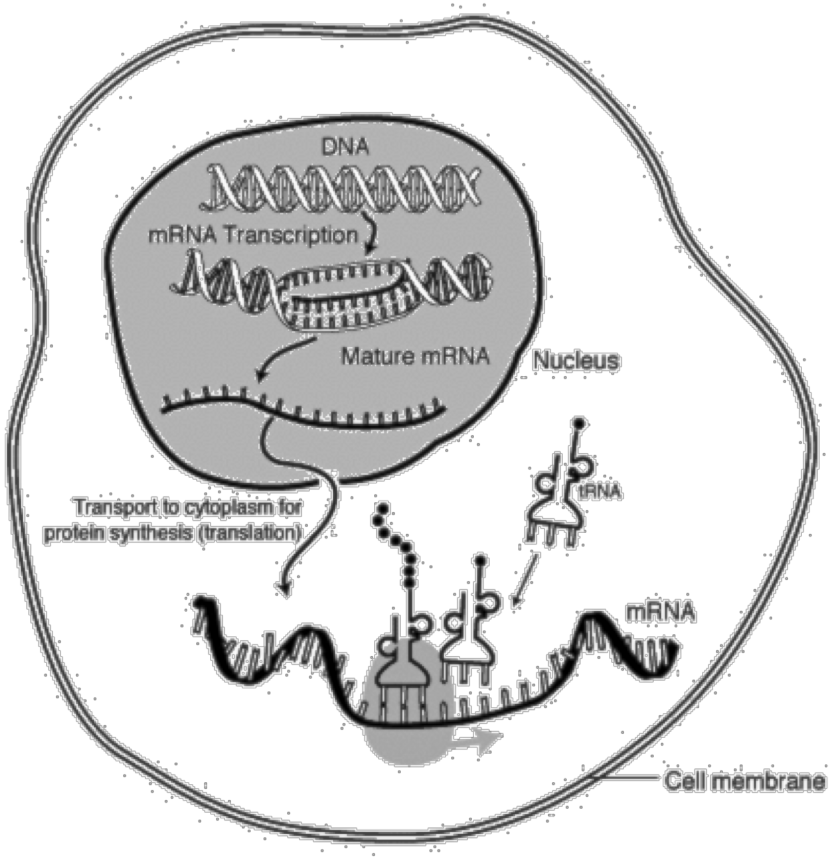
UHN Princess Margaret Cancer Centre

Introduction

- What is functional genomics?
- Why should we share our data?
- What are some common hurdles working with public functional genomics data?

What is functional genomics?

- Functional genomics uses high-throughput biological data to answer questions about dynamic biological processes.
- For example:
  - What predisposes a person to a given type of cancer?
  - Can we identify any biomarkers of a given cancer?
  - Which biological pathways go awry during a given type of cancer?



A recent example

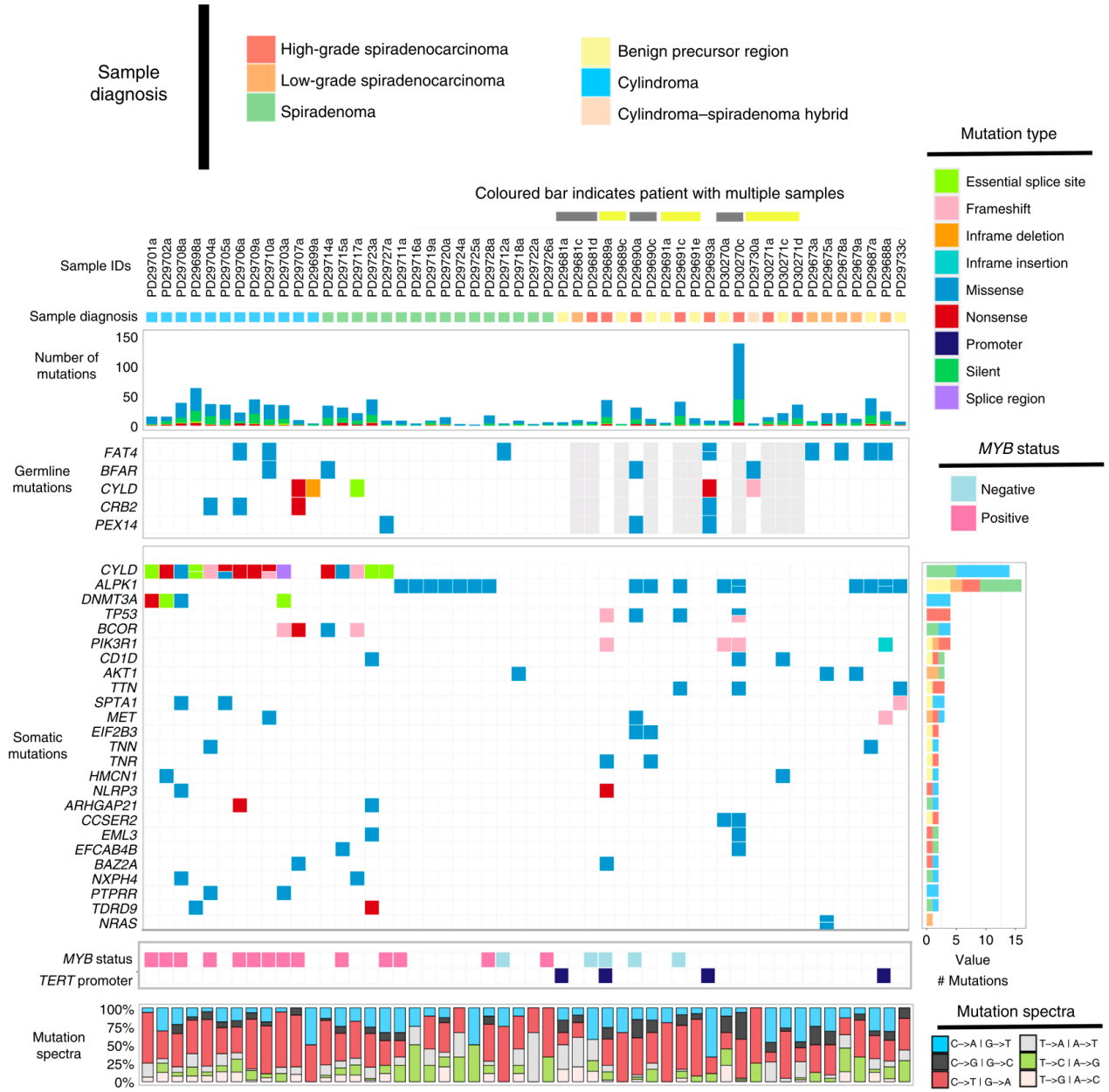
Article | OPEN | Published: 17 May 2019

ALPK1 hotspot mutation as a driver of human spiradenoma and spiradenocarcinoma

Mamunur Rashid, Michiel van der Horst, […] David J. Adams

Nature Communications **10**, Article number: 2213 (2019)

- Whole genome sequencing of 57 patients with a rare and aggressive skin tumour
- 164 sequenced samples in total

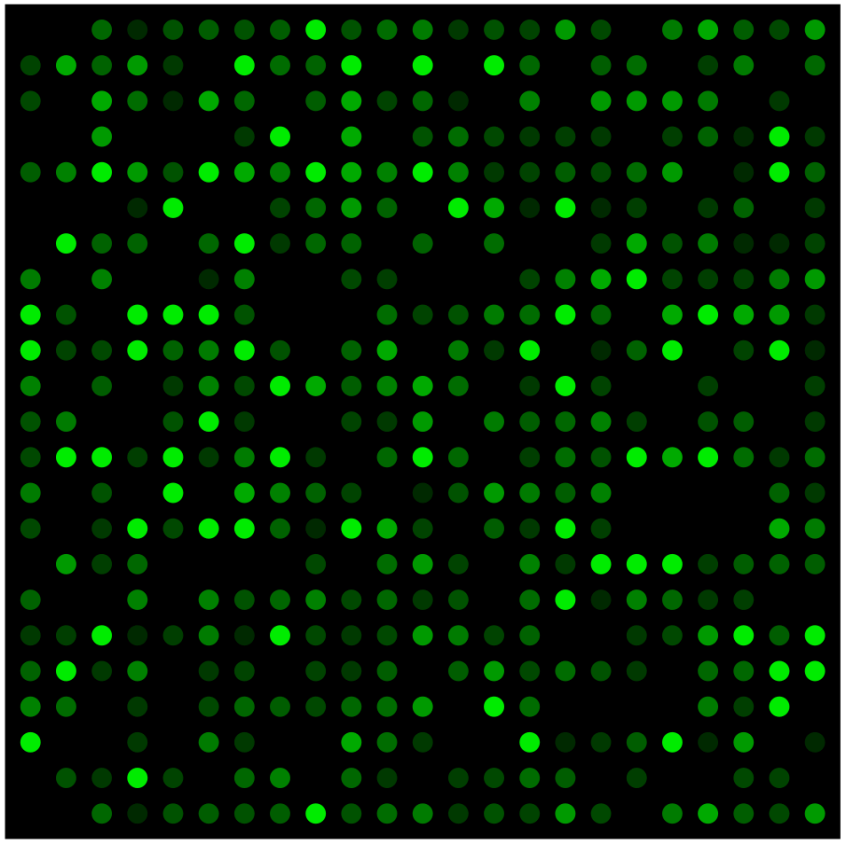


Rashid, M. et al. Nat. Commun. 10, 2213 (2019).

What does functional genomics data look like?

- DNA microarrays
  - Older and cheaper technology
  - Require prior sequence information of the genes or transcripts that you want to assay
  - Output numeric data
    - I.e. a table with ~25,000 rows and at least 1 column per sample

Microarray chip



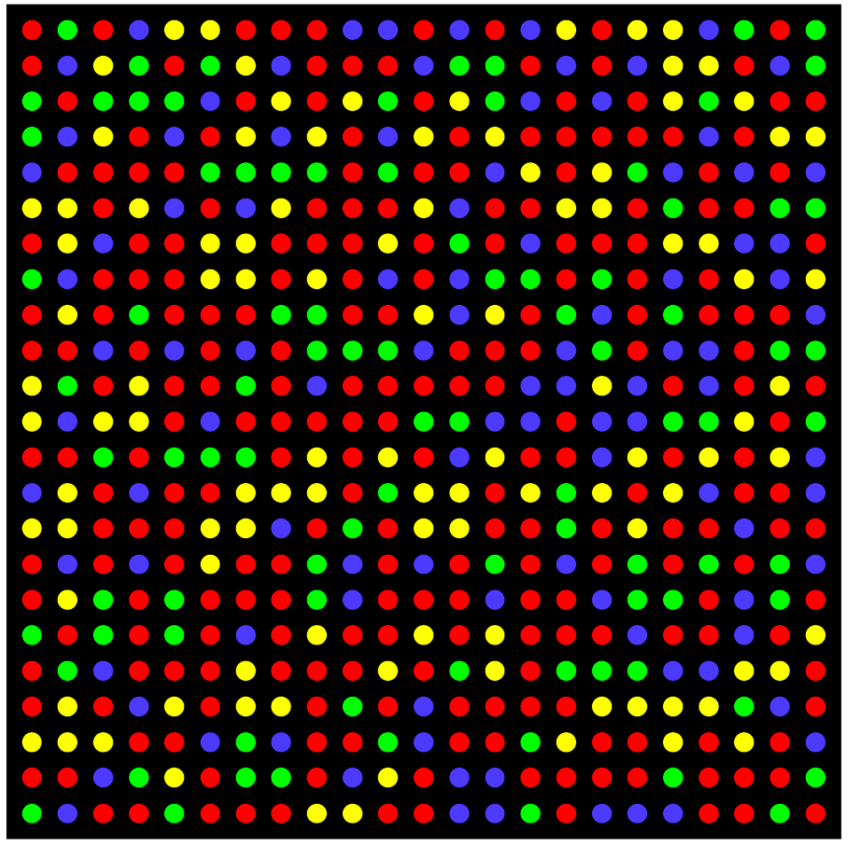
What does functional genomics data look like?

- Next-generation sequencing
  - Newer and more expensive technology
  - Does not necessarily require prior sequence knowledge
  - Outputs data as sequences with quality scores

FASTQ format:

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Sequencing flow cell



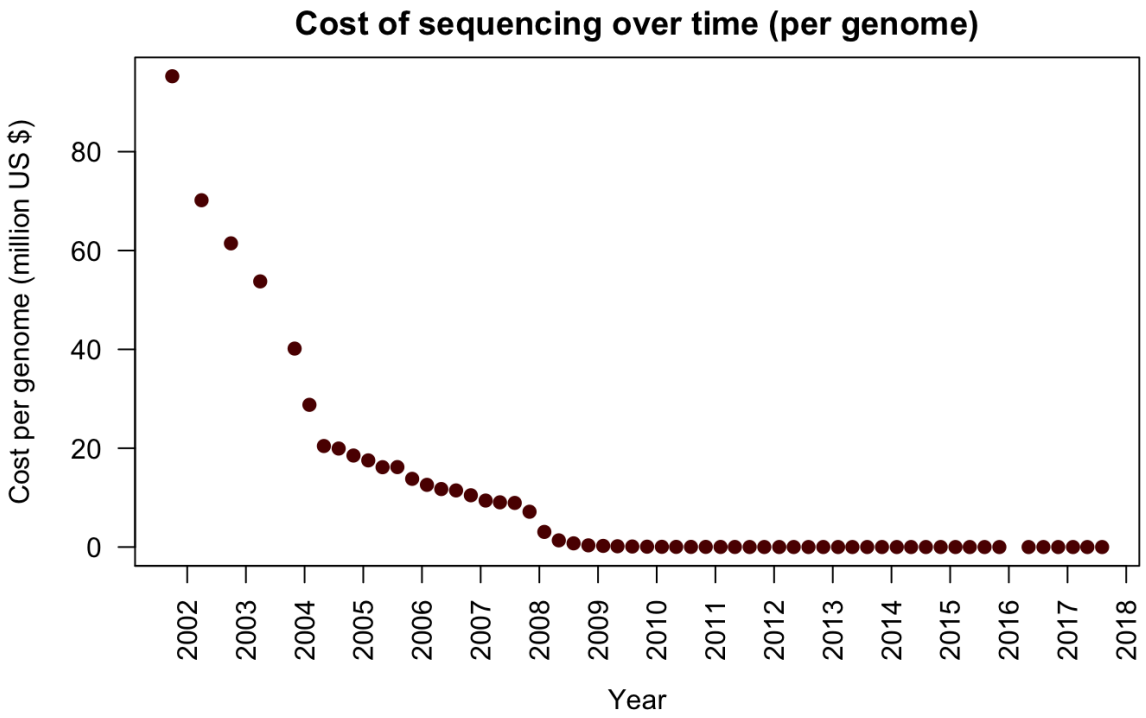
How much does sequencing cost?

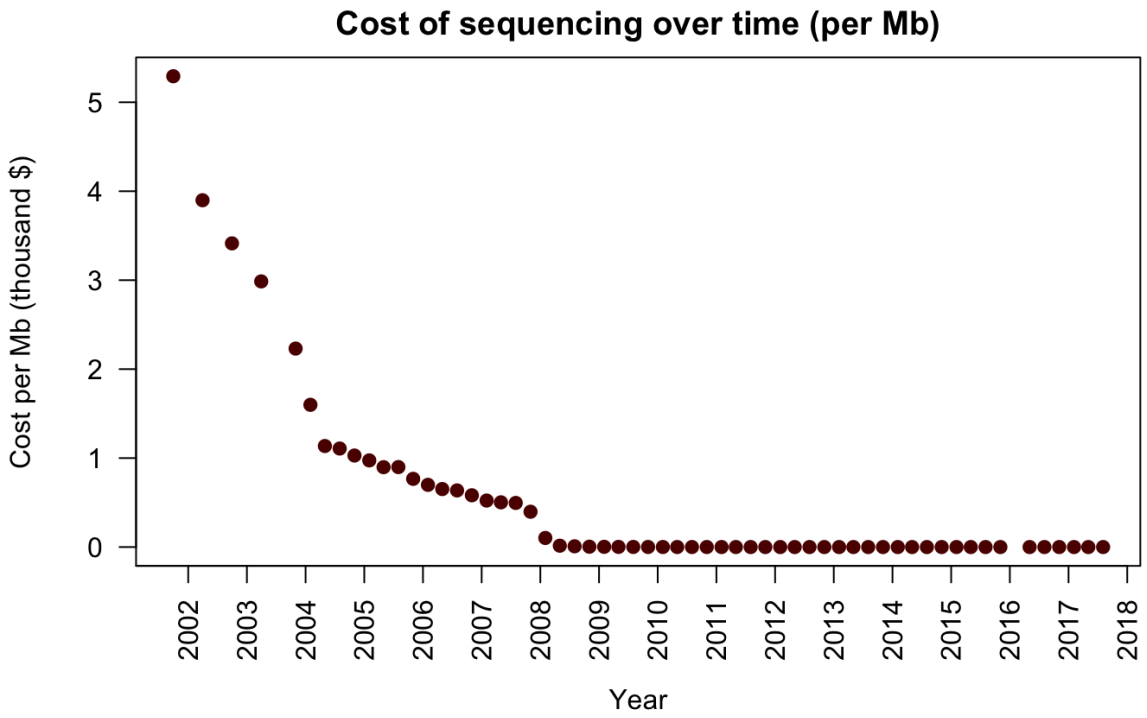
Human Genome Project (2001)

US $2,700,000,000

23 and Me (2019)

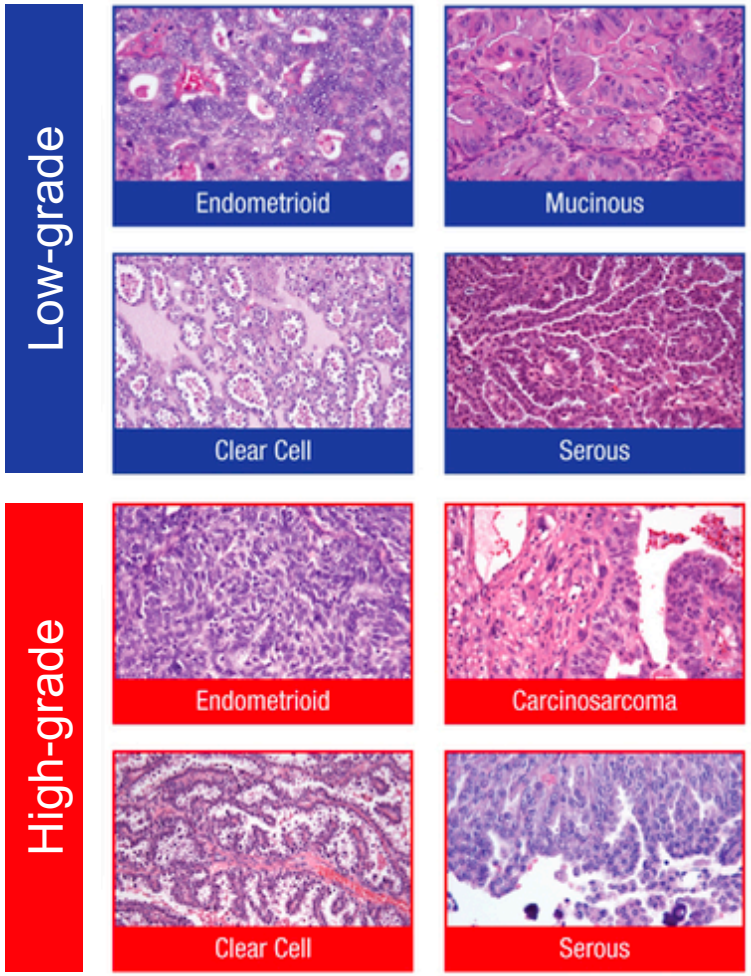
CA $129





How much does a sequencing experiment cost?

- Histological Subtype Specific Epigenetic Risk For Ovarian Cancer Using Whole Genome Bisulphite Sequencing
  - James Flanagan, PhD and Ed Curry, PhD, Imperial College London
- Study design:
  - 891 cases and matched controls
  - Aim to sequence 30-50 million reads
- Costs:
  - DNA sample acquisition = CA $24,270.96
  - DNA sample processing = CA $31,213.09
  - Sequencing = CA $114,018.21
  - Data analysis = CA $25,959.72
  - Data storage = CA $7,286.14
  - Total = CA $202,827.80

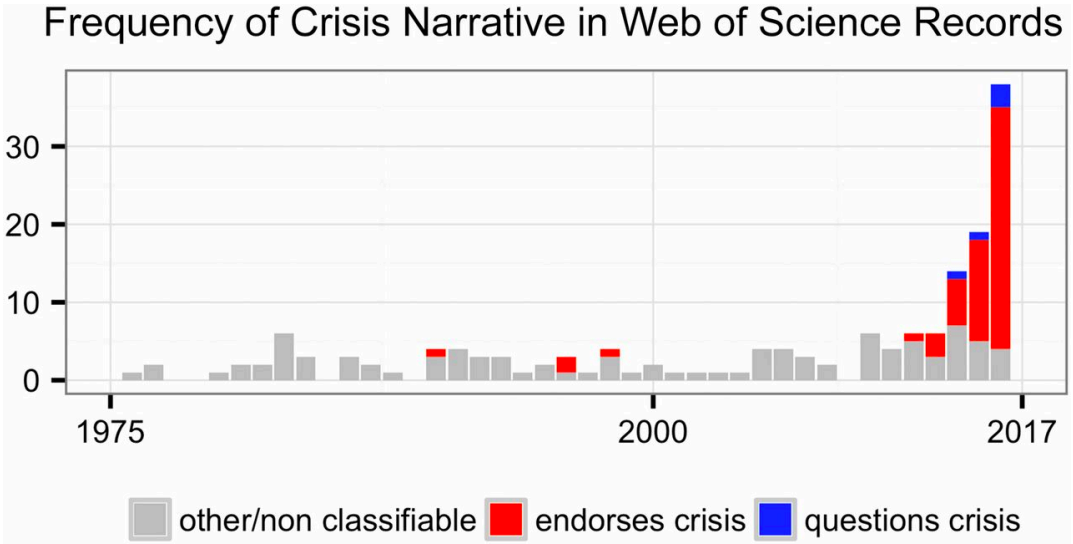


Jones, P. M. & Drapkin, R. *Front. Oncol.* **3**, (2013).

Why should we share our data?

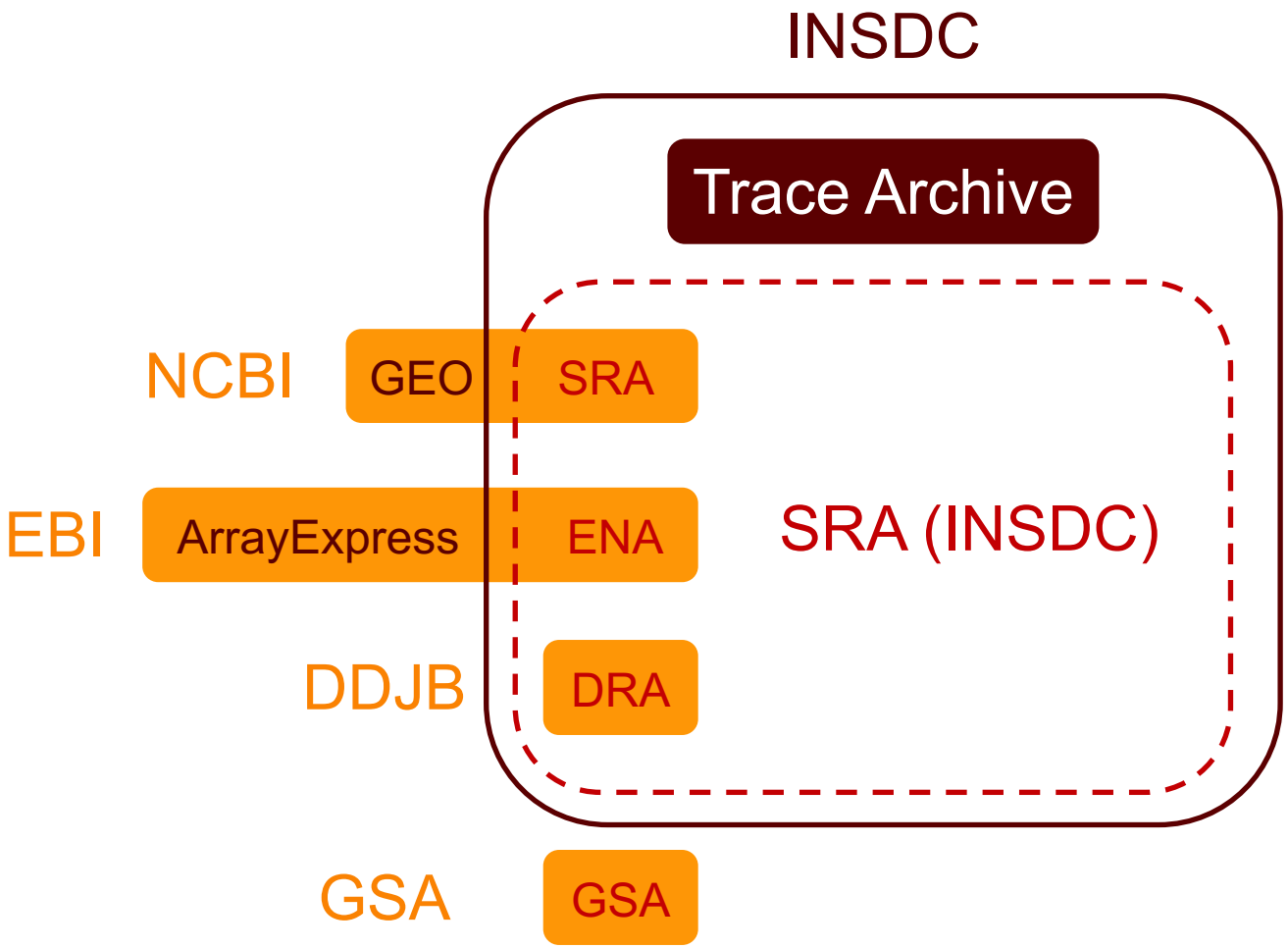
- To ensure our experiments are reproducible
- To allow others to repurpose our data
  - Empowers researchers with more limited resources
  - Gets the most out of precious biological samples (e.g. rare disease samples)
  - Reduces redundant experiments

The research reproducibility crisis (?)



Fanelli, D. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 2628–2631 (2018).

What public repositories exist?



The International Nucleotide Sequence Database Collaboration (INSDC)

National Centre for Biotechnology Information (NCBI)

- Sequence Read Archive (SRA)
- Gene Expression Omnibus (GEO)

European Bioinformatics Institute (EBI)

- European Nucleotide Archive (ENA)
- ArrayExpress

DNA Databank of Japan (DDJB)

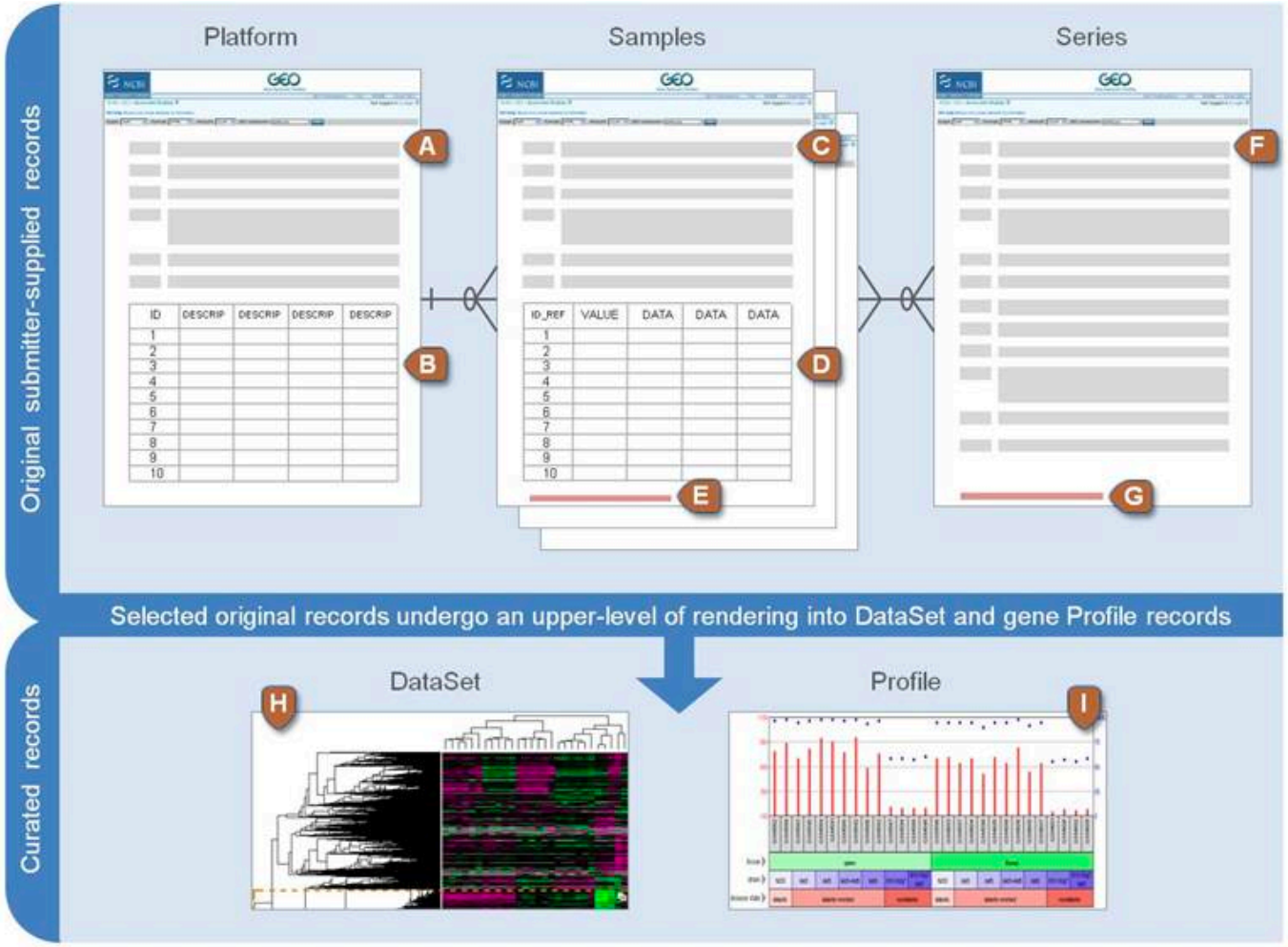
- DDBJ Sequence Read Archive (DRA)

China's Genome Sequence Archive (GSA)

- Genome Sequence Archive (GSA)

Wheeler, K. *Childhood Cancer Data Lab Blog* (2019). Available at: https://www.ccdatalab.org/blog/2019/3/29/gene-expression-repos-explained

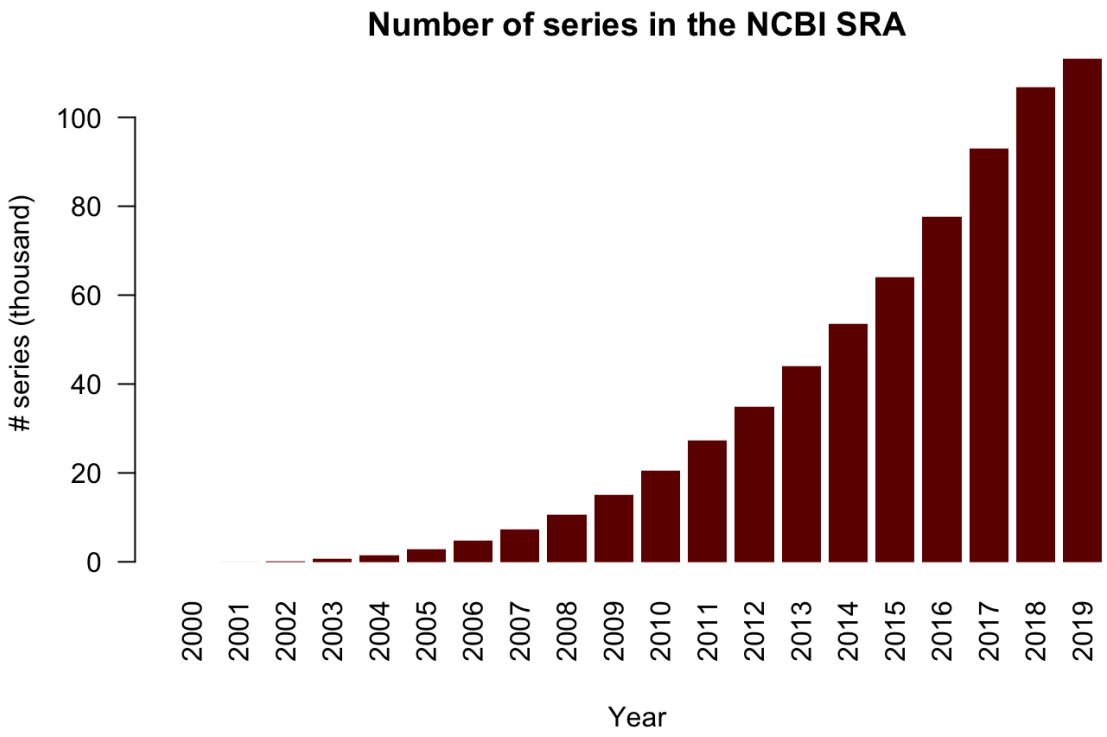
Public genomics databases can be confusing to navigate

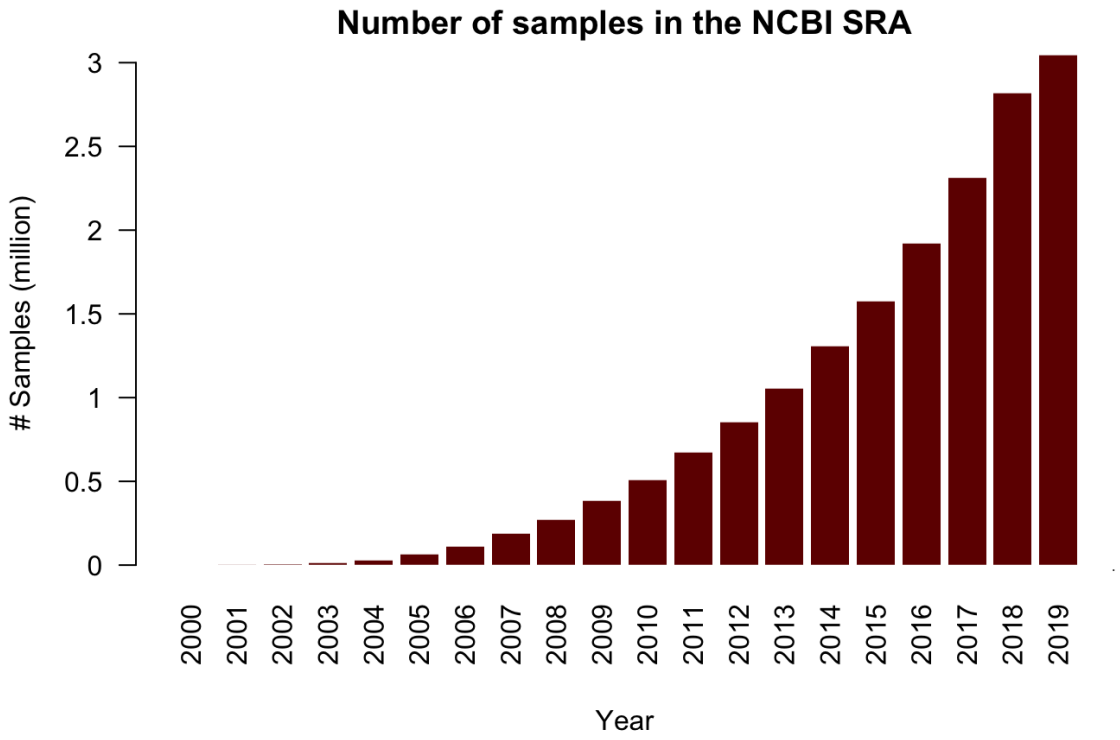


NCBI GEO accessions

- Original data
  - GSM – sample
  - GSE – series
  - GPL – platform
- Lightly curated data
  - GDS – data set

How much data is in those repositories?





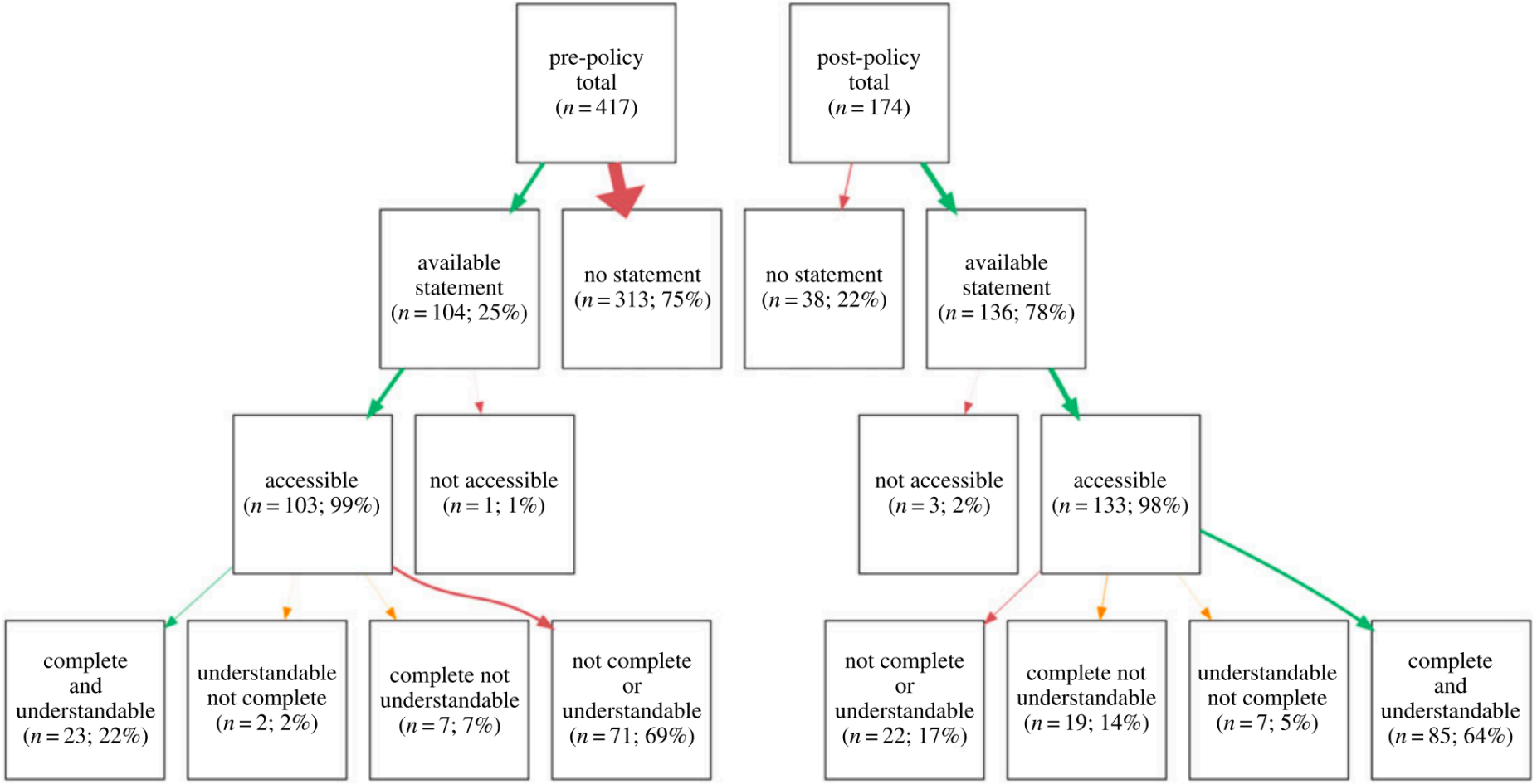
What are some common hurdles in using publicly available functional genomics data?

- Databases can be confusing to navigate
- The accompanying metadata can be incoherent, inconsistent, and/or incomplete
- Researchers withhold data

Incoherent, inconsistent, and incomplete metadata

- What metadata are we talking about?
  - What is the sample?
  - Where did it come from?
  - How was the experiment conducted?

Incoherent, inconsistent, and incomplete metadata



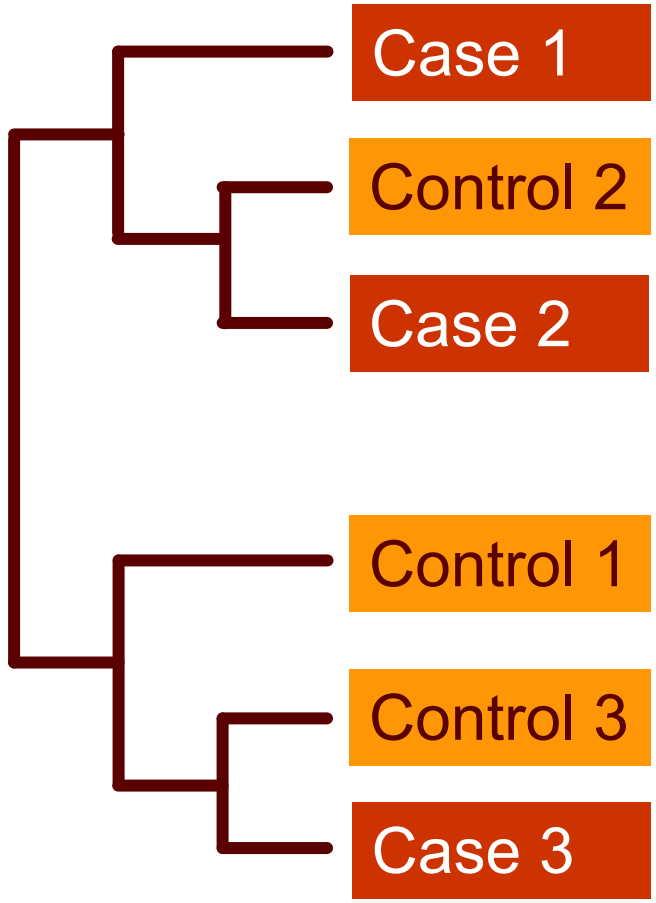
Hardwicke TE et al. 2018 Data availability, reusability, and analytic reproducibility: evaluating the impact of a mandatory open data policy at the journal Cognition. R. Soc. open sci. 5: 180448. http://dx.doi.org/10.1098/rsos.180448

Incoherent, inconsistent, and incomplete metadata

- age
- Age
- age (yrs)
- age (years)
- age (y)
- age in years
- age_years
- AGE
- age(years)
- age (year)
- age (yr)
- Age (years)
- age (years)
- age [y]
- age [years]
- Age(yrs.)
- age.year
- age (yr-old)
- age(yrs)
- age of patient
- Age, year
- Age (yrs)
- Age of patient
- age, years
- `Age
- Age (Years)
- age (after birth)
- age, yrs
- age of subjects

Prof Michel Dumontier (@micheldumontier), Maastricht University, via @neuronJoy. Available: https://twitter.com/neuronJoy/status/1027616795279028225

Incoherent, inconsistent, and incomplete metadata



Researchers withhold data

Original Contribution

January 23/30, 2002

Data Withholding in Academic Genetics

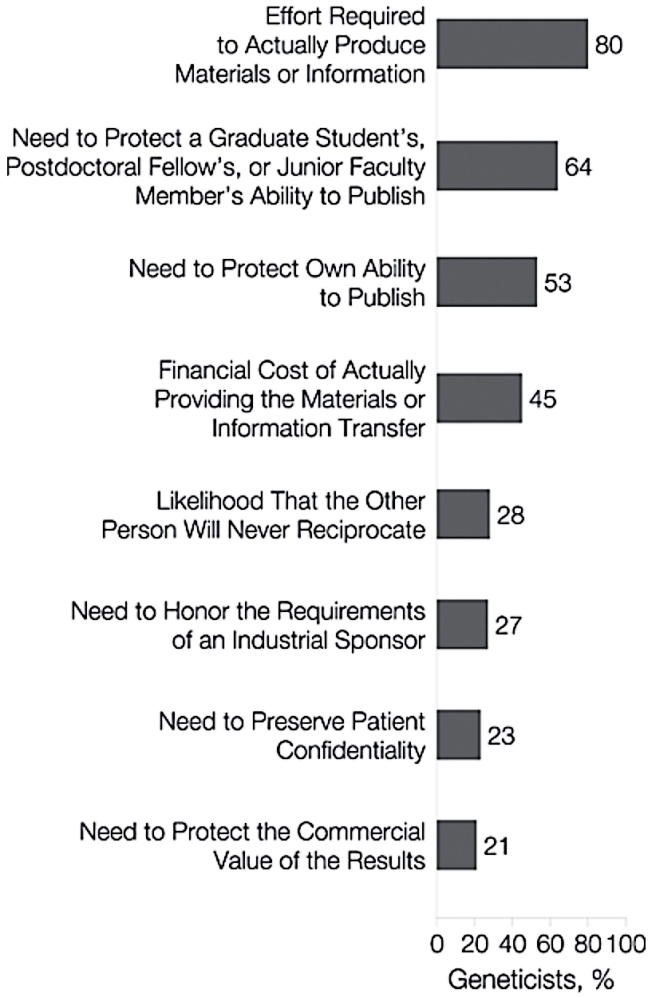
Evidence From a National Survey

Eric G. Campbell, PhD; Brian R. Clarridge, PhD; Manjusha Gokhale, MA; et al

JAMA. 2002;287(4):473-480. doi:10.1001/jama.287.4.473

- Survey of 2,893 geneticists
  - Including all members of the Human Genome Project
- 1,849 responded (64%)
- 84% of respondents requested data. Of those 47% reported data withholding.

Figure. Geneticists' Reasons for Withholding Postpublication Information, Data, or Materials



Respondents who considered these motivating factors very important or important.

Campbell, E. G. et al. JAMA 287, 473 (2002).

Researchers *still* withhold data

LETTERS

Edited by **Jennifer Sills**

Reminder to deposit DNA sequences

AS MEMBERS OF the Advisory Committee to the International Nucleotide Sequence Database Collaboration (INSDC), which includes the DNA Data Bank of Japan (DDBJ), ENA, and GenBank databases, we wish to remind the research community of the importance of depositing complete DNA-sequence data in these databases upon publication of their results [see also S. L. Salzberg *et al.*, *Nature*, http://dx.doi.org/10.1038/533179a (2016)]. Indeed, most journals demand a database accession number as a condition of publication.

Reminder to deposit DNA sequences By Mark Blaxter, Antoine Danchin, Babis Savakis, Kaoru Fukami-Kobayashi, Ken Kurokawa, Sumio Sugano, Richard J. Roberts, Steven L. Salzberg, Chung-I Wu Nature 12 May 2016 533 : 179 | Science 13 May 2016 : 780

Researchers *still* withhold data

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

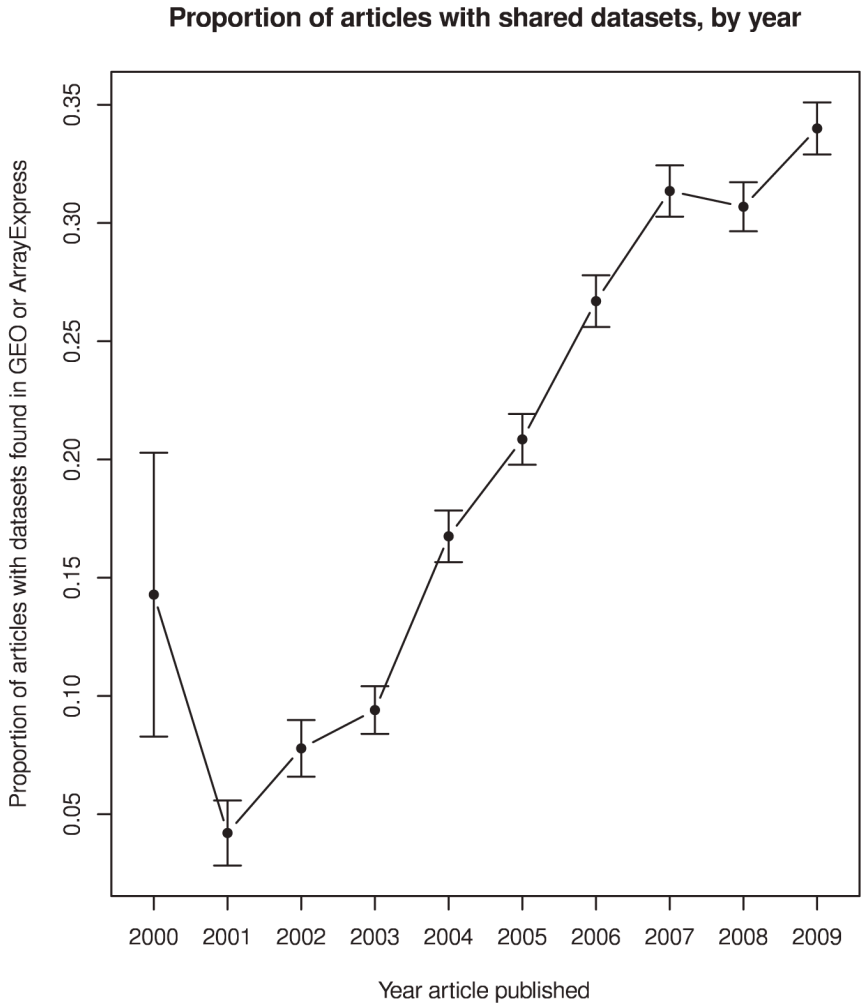
PLOS ONE

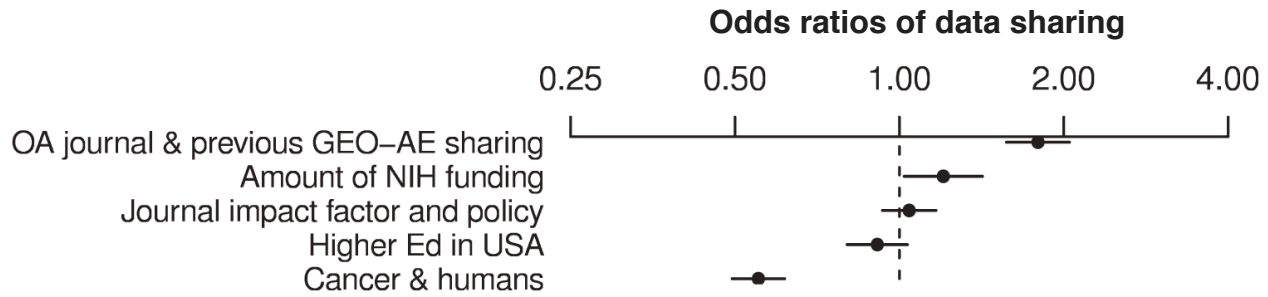
Who Shares? Who Doesn't? Factors Associated with Openly Archiving Raw Research Data

Heather A. Piwowar

Published: July 13, 2011 • https://doi.org/10.1371/journal.pone.0018657

- Identified 11,603 papers published between 2000 and 2009 containing microarray data
- Data sharing has increased over time
- The best predictor of data sharing is previous data sharing
- The worst predictor is cancer research on human samples





Piwowar, H. A. *PLoS One* **6**, e18657 (2011).

How do we encourage researchers to share data?

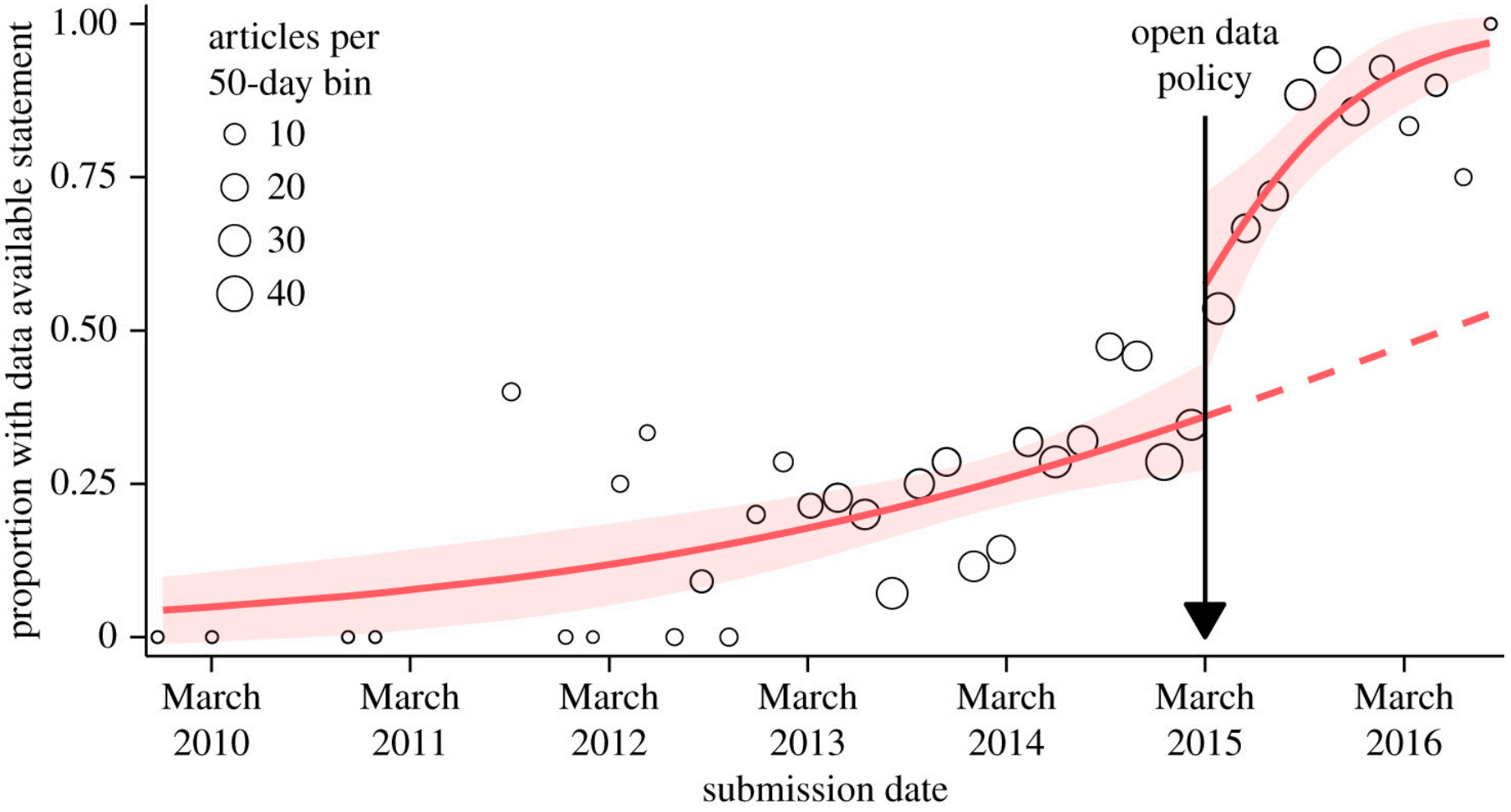
- We make data sharing a prerequisite for:
  - Funding applications
  - Publication

CANCER RESEARCH UK

We regard it good research practice for all researchers to consider at the research proposal stage how they will manage and share the data they will generate. Therefore, we require that applicants applying for funding provide a data management and sharing plan as part of their application. This plan will be reviewed as part of the funding decision.

Cancer Research UK. Submission of a data sharing and preservation strategy (2019). Available at: https://www.cancerresearchuk.org/funding-for researchers/applying-for-funding/policies-that-affect-your-grant/submission-of-a-data-sharing-and-preservation-strategy

Does marrying data sharing to money encourage researchers to release their data?



Hardwicke TE et al. 2018 Data availability, reusability, and analytic reproducibility: evaluating the impact of a mandatory open data policy at the journal Cognition. R. Soc. open sci. 5: 180448. http://dx.doi.org/10.1098/rsos.180448

Does marrying data sharing to money encourage researchers to release their data?

OPEN ACCESS Freely available online

PLoS BIOLOGY

Correspondence

Data Sharing: How Much Doesn't Get Submitted to GenBank?

Mohamed A. F. Noor, Katherine J. Zimmerman, and Katherine C. Teeter

reported in published work into GenBank. We know from personal experience that authors of published papers reporting DNA sequences sometimes intentionally fail to deposit their sequences to GenBank and refuse to release them upon request. Is this a rare exception, or do many papers make it past coauthors, associate editors, editors, reviewers, and journal staff without providing the purportedly required data accession numbers?

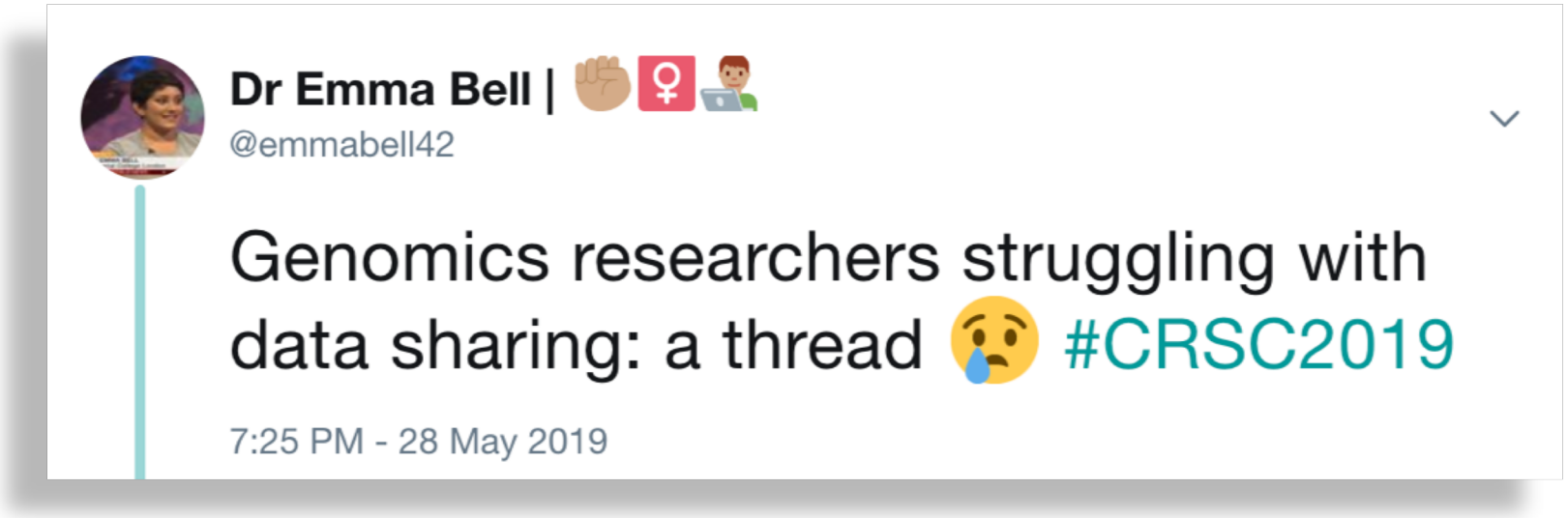
Journal	Number of Papers Examined	Number Missing Accession Numbers	Number with Sequences Never Submitted to GenBank	Number of Special Cases[a]
Evolution	39	8	6	0
Molecular Biology and Evolution	109	7	4	0
Nature	42	3	3	1
PLoS Biology	30	3	3	2
Proceedings of the National Academy of Sciences USA	30	1	1	0
Science	30	4	2	0

[a]These papers provided their DNA sequences as supplementary materials or by noting identity to other published sequences, but they were not submitted to GenBank. They are also included in the totals in the first three columns.

DOI: 10.1371/journal.pbio.0040228.t001

Noor, M. A. F., Zimmerman, K. J. & Teeter, K. C. *PLoS Biol.* **4**, e228 (2006).

Does that work?



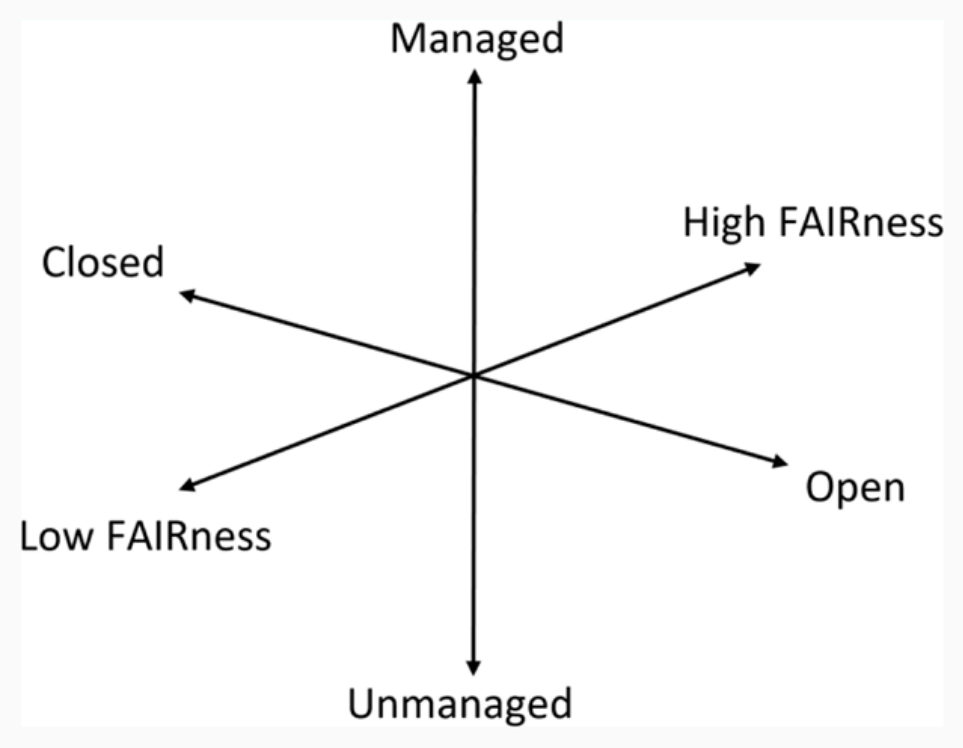
Click here → https://twitter.com/emmabell42/status/1133439189519876097

What are some common hurdles in using publicly available functional genomics data?

- Databases can be confusing to navigate
- The accompanying metadata can be incoherent, inconsistent, and/or incomplete
- Researchers withhold data

Why is data sharing in functional genomics such a big problem?

- Poor widespread understanding of:
  - Research data management
  - Open data
  - Findable Accessible Interoperable Reusable (FAIR) data



Higman, R., Bangert, D. & Jones, S. *Insights UKSG J.* **32**, (2019).

Why is data sharing in functional genomics such a big problem?

A guide for the lonely bioinformatician

23RD APRIL 2013 / BIOMICKWATSON / 60 COMMENTS

pet bio·in·for·mat·ician

/pet ˌˈbī-ō-in-fər-mə-ˈti-shən/

noun (computing)

“a single bioinformatician employed within a laboratory based group”

“Don’t worry”, the lead PI said, “we’ve put money on the application to fund a [pet] bioinformatician.”

Watson, M. *Opiniomics* (2013). Available at: http://biomickwatson.wordpress.com/2013/04/23/a-guide-for-the-lonely-bioinformatician/

References

- Rashid, M. et al. Nat. Commun. 10, 2213 (2019).
- Wheeler, K. *Childhood Cancer Data Lab Blog* (2019). Available at: https://www.ccdatalab.org/blog/2019/3/29/gene-expression-repos-explained
- Fanelli, D. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 2628–2631 (2018).
- Hardwicke, T. E. et al. R. Soc. Open Sci. 5, 180448 (2018).
- Campbell, E. G. *et al. JAMA* **287**, 473 (2002).
- Salzberg, S. L. *Nature* **533**, 179–179 (2016).
- Cancer Research UK. Submission of a data sharing and preservation strategy (2019). Available at: https://www.cancerresearchuk.org/funding-for researchers/applying-for-funding/policies-that-affect-your-grant/submission-of-a-data-sharing-and-preservation-strategy
- Noor, M. A. F., Zimmerman, K. J. & Teeter, K. C. *PLoS Biol.* **4**, e228 (2006).
- Higman, R., Bangert, D. & Jones, S. *Insights UKSG J.* **32**, (2019).
- Watson, M. *Opiniomics* (2013). Available at: http://biomickwatson.wordpress.com/2013/04/23/a-guide-for-the-lonely-bioinformatician/