

Data management/ archiving across the biological sciences

GEORGE (PJ) PERRY  @grg_perry

Departments of Anthropology & Biology

Interdisciplinary graduate programs in
Ecology and Bioinformatics & Genomics

Pennsylvania State University

06 June 2016

Opening discussion

Have you had experiences obtaining data from published studies that you needed for your own research?

Have you gone through the effort to deposit data generated for one of your studies available in a repository?

- What were the challenges?
- What are the advantages of having done so?

Reasons some researchers don't share data

Plans for future pubs w/ same data;
Fear of being scooped

Do not want to aid “the competition”

It takes effort

No perceived self-benefit

All science builds on the ideas,
work, and data of others.
Enjoy participating in the process.
Support other researchers.

Plan from the very start of project

Actually benefits impact of work

Required by many journals
(although hard to fully police)

Required by funding agencies
(demonstrate a sharing record)

May benefit (or otherwise limit)
future career opportunities

Reasons some researchers don't share data

THE NEW ENGLAND JOURNAL of MEDICINE

EDITORIALS

21 January 2016



Data Sharing

Dan L. Longo, M.D., and Jeffrey M. Drazen, M.D.

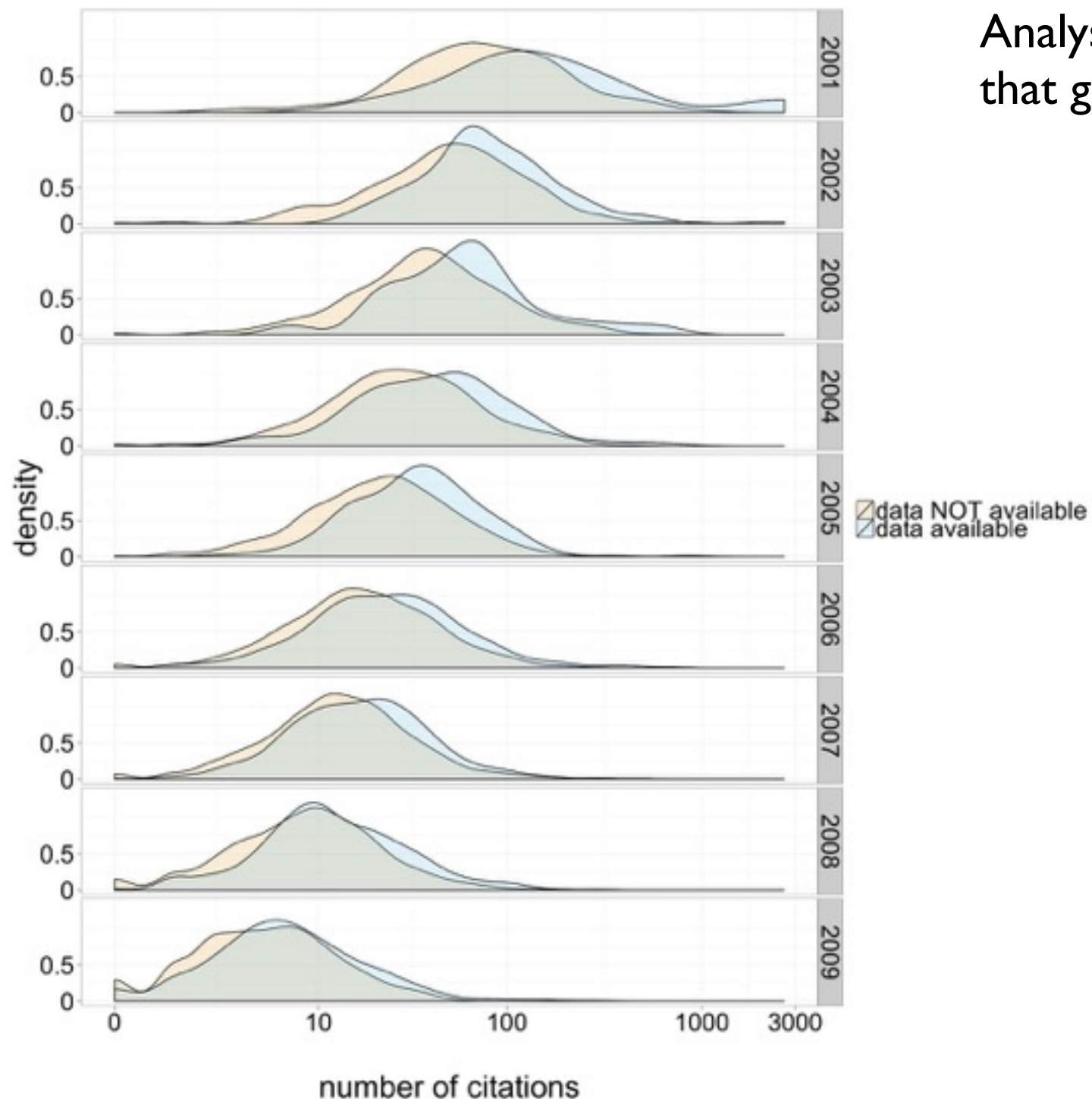
A second concern held by some is that a new class of research person will emerge — people who had nothing to do with the design and execution of the study but use another group's data for their own ends, possibly stealing from the research productivity planned by the data gatherers, or even use the data to try to disprove what the original investigators had posited. There is concern among some front-line researchers that the system will be taken over by what some researchers have characterized as “research parasites.”

What are the advantages of making data accessible?

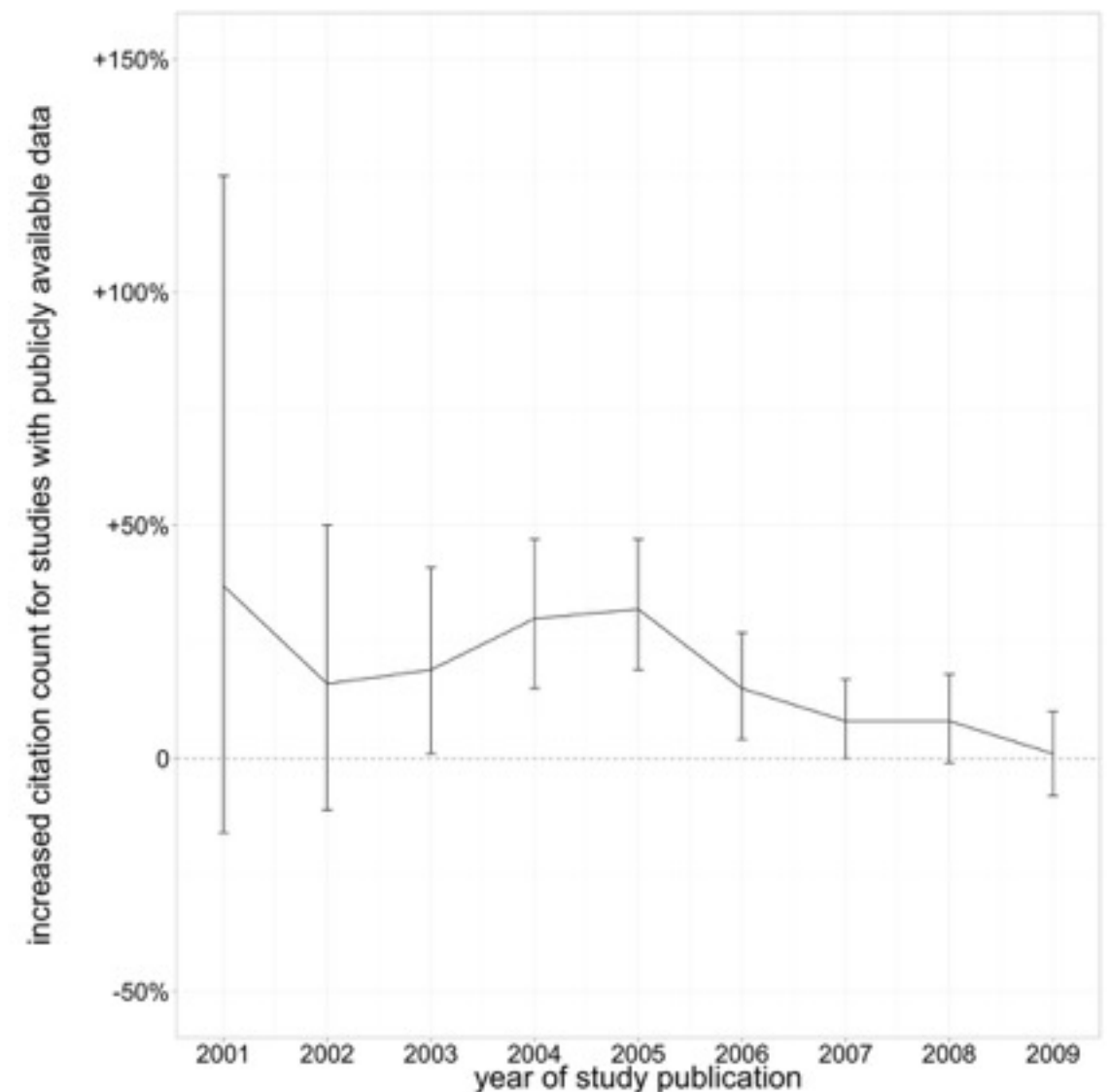
- Ethical obligation for research conducted with taxpayer or foundation funds (and often a requirement)
- Maximize the not only the reproducibility, but also the impact and visibility of your research
- Ensure that your irreplaceable data become part of the permanent record
- Provide opportunities for future generations of scientists; give back to the community that helped provide you this rare opportunity for doing research
- Long-term experience in data sharing community standards and benefits in genetics/ genomics
- Help ensure that scientific research maximizes relevance in an increasingly data-rich, digital, and computational society

Data sharing and scientific impact

Citations, one measure of scientific impact (relative within scientific fields), are typically higher for publications with full data availability



Analysis based on 10,555 studies (from 2001-09) that generated gene expression microarray data



Challenges for data sharing in the biological sciences

Fast-moving technology and methods for data collection/ analysis

- Must maintain knowledge of evolving databases & standards

Large file formats are becoming more and more prominent

- Computational training and skill increasingly needed to use and archive data

In some cases, appropriate community-supported databases do not yet exist

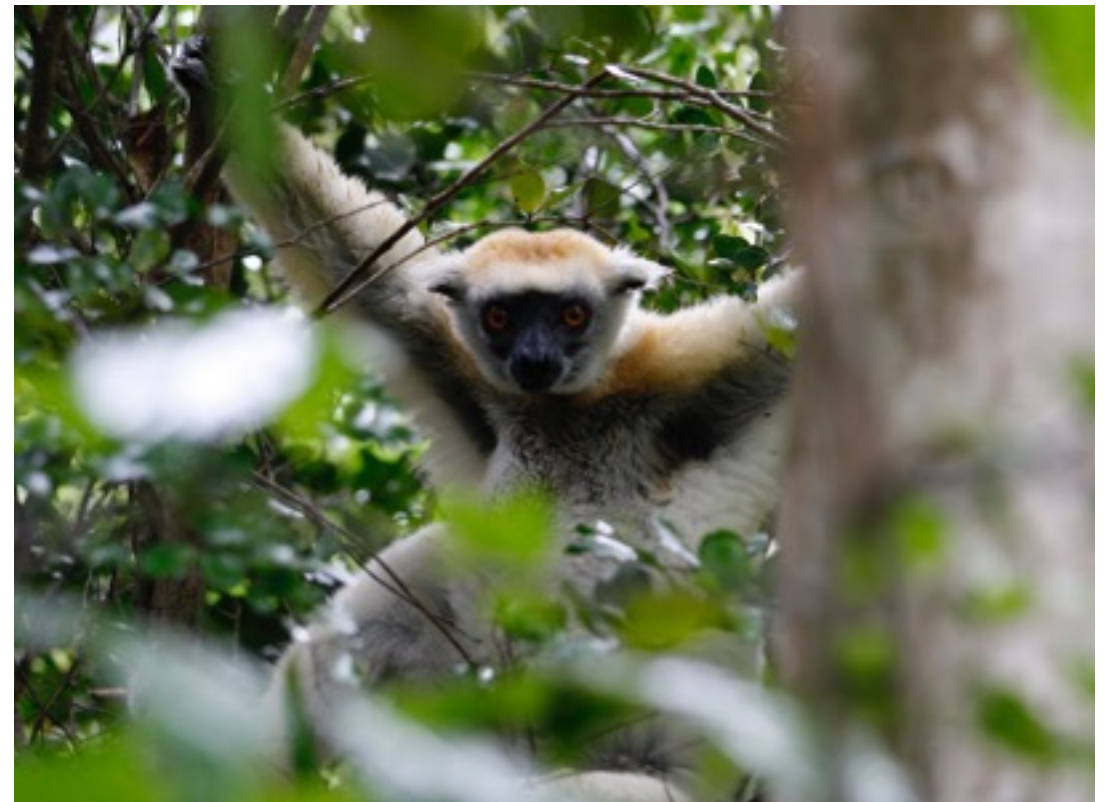
- Require stable funding and management (and “too big to fail” status)

Challenges for data sharing in the biological sciences

At odds with historical operating procedures in some fields (e.g., paleontology)

- Even upon access to specimens, there may be restrictions placed on use

Concerns from directors of long-term field ecology (e.g., primatology) studies



Challenges for data sharing in the biological sciences

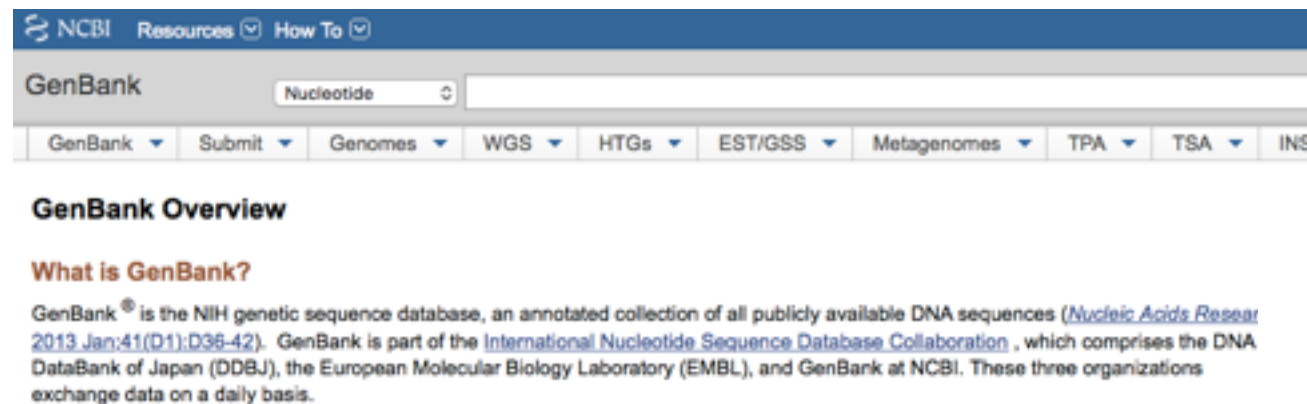
Privacy and ethical considerations with sharing of some forms of human biological data that are potentially identifiable

- Incorporate planning for maximal data sharing, given the privacy risks, into the data management plan and address with participants in the informed consent process from the outset



Data sharing in genetics/ genomics

GenBank - NIH database since 1982 for depositing determined nucleotide sequences of a gene/ genomic region for specified individuals & organisms



- Sequences deposited receive accession numbers and are cross-referenced with associated publications
- Users can search by organism, genetic locus, etc., or query directly against nucleotide sequences via various tools

Growth of GenBank:

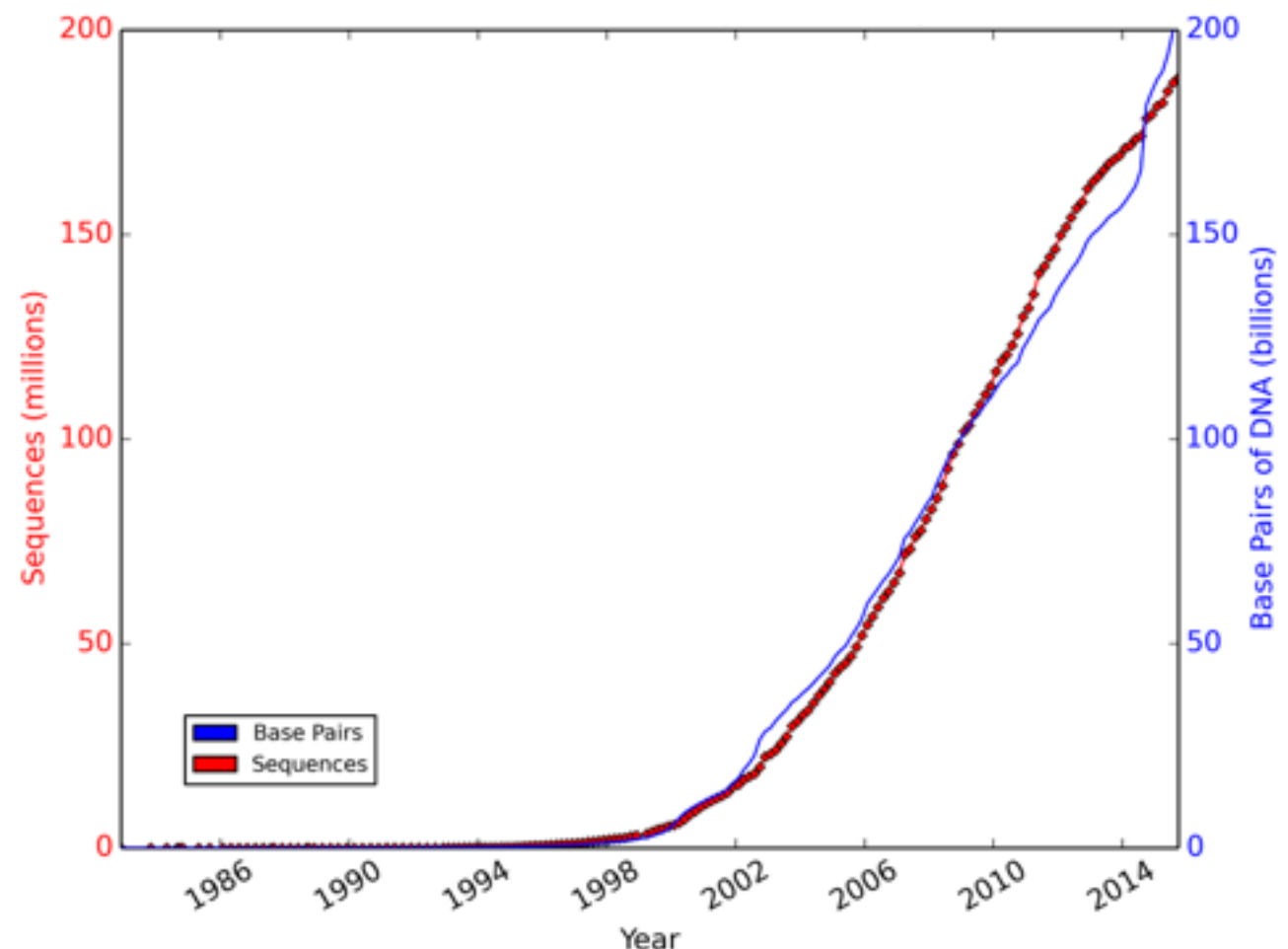
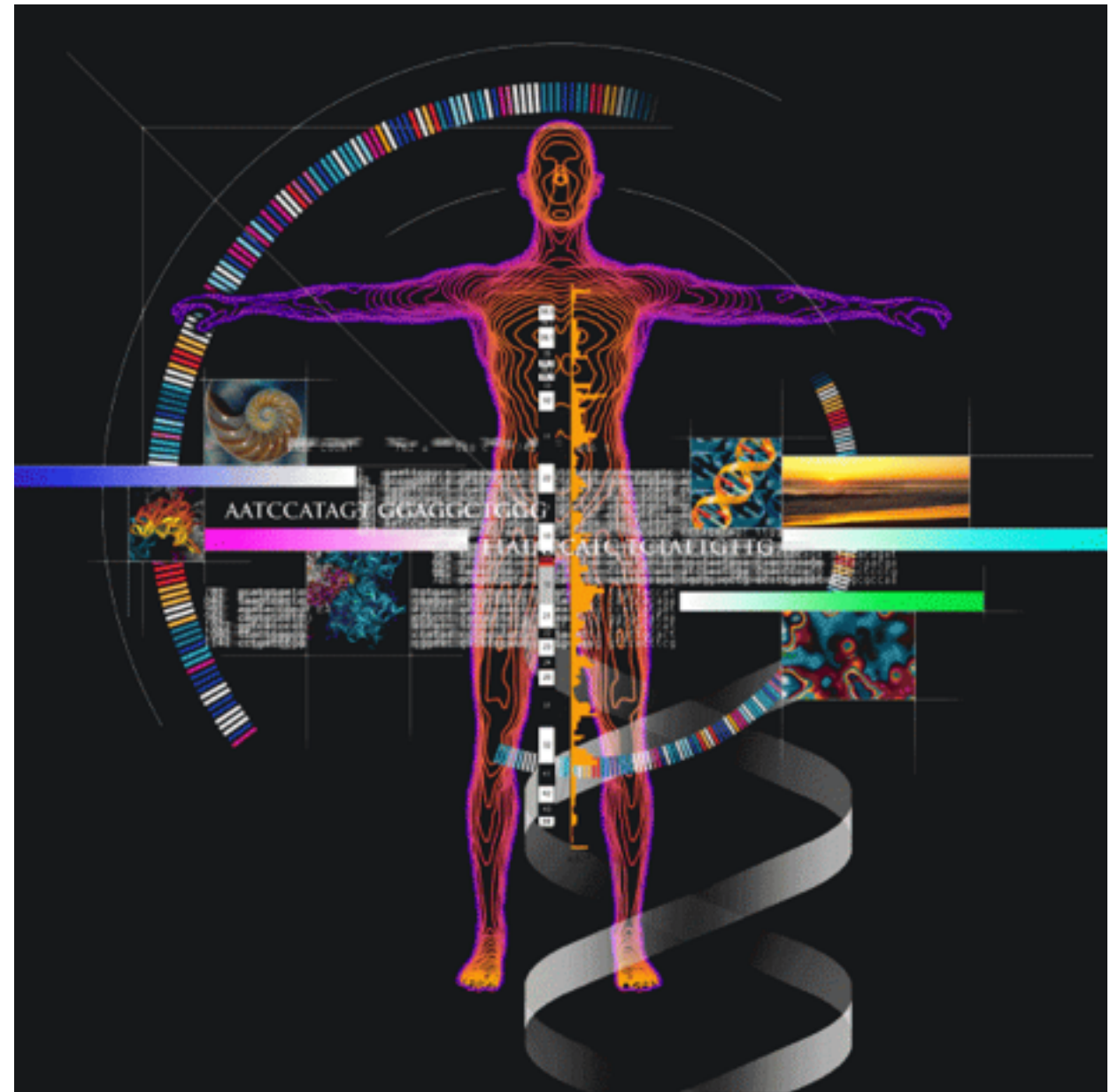
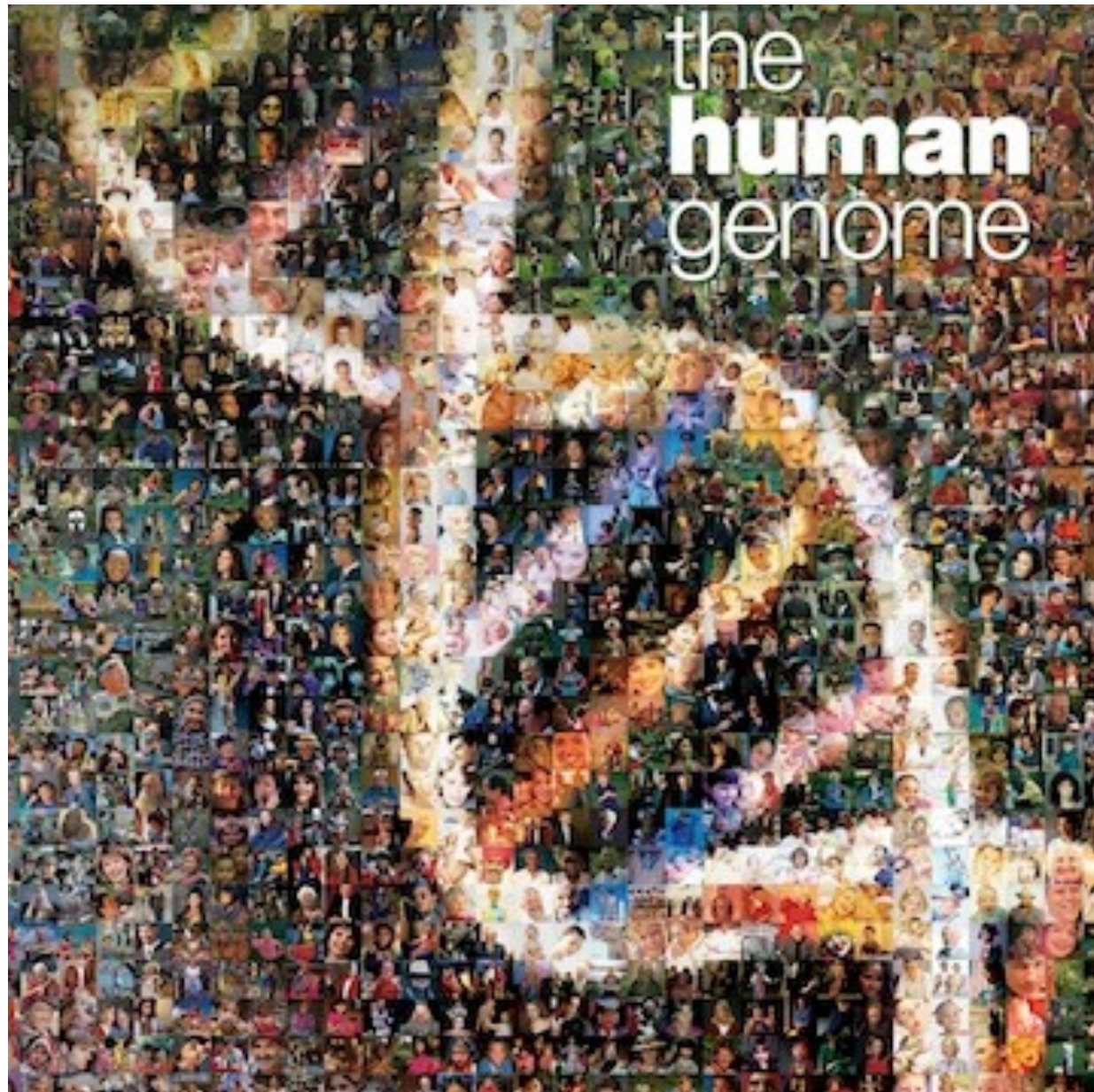


Figure: Mark Pauley

Data sharing in genetics/ genomics

Combined with the availability of the human genome reference sequence (and for other organisms), sequence depositions from individual labs have collectively facilitated an otherwise impossible level of scientific advance.



Data sharing in genetics/ genomics

With new, massively-parallel sequencing technology, genomic data are now more frequently being deposited into other databases - e.g., the Sequence Read Archive.



```
.CAGATGTGGATAACTTGGGTAGAATGGCGACCCCTTCTCATCAGGAAGGGTTAATCTTTAAATGATTT  
ATTTAAAACGCAGACATAGGGGATACACATGCTTTGGACAGACTGCTTAACTCGCTTGCGACAAGAGC  
TCTGATAACGTCTTTGCGATGTGGATTGCGCCCTTTAGTAGCTGAAGAAGTAGAGGGGATACTACGTCT  
ATGCTCCTAATCCTTATTGGACGCGTTATATTCAAGAGAATCATTTAGAGTTAATTTCTATATTGGCT  
ACAATTGTCGGAAGGGCGGGTGCCTCAGGTTGAAATCTTGGTAGATTCTCGTCCTGGTAGTATTTTGT  
TCTAGTGAACAGCCTGCAACAACACTACAGCAGCTTTACAACTGCCCCCTATACCTCAACCTGCTAAGGT  
AAAGAGAACCGGAACCTGTTGCTAATACTGCAGTTAGTTCTAAGAGTTCAAAAAAGAACTATTAAAT  
ACAATTTACTTTTTCACTATTTGTTGAAGGCCGTTCTAACCAAATGGCAGCAGAAACCTGTAGAAAAG  
TTAACACAGTTAGGTGCTTCTCAACATAACCCTTTGTTTTTATATGGCCCGACAGGTCTTGGTAAGAC  
ACTTAATGCAAGCAGTTGGTAATGCCTTACTGCAAGCGAAGCCGAATGCAAGAGTCATGTATATGACT  
AGAAAGTTTTGTACAAGATTTTGTGAGCTCATTACAAAAAGGAAAGGTAGAAGAGTTTAAGAAAAATT  
CGTTCTTTAGACTTGTTATTAGTAGATGATATTCATCTTTTGGCAGGAAAAGAAGCAAGCCTCGTTGA  
TCTTCTATACATTTAATGCCTTACTTGATGAATCCAAACAAATTATTTTAACGTCAGATCGATATCCT  
AGAATTAACAGAACTTGATCCTCGATTGGTTTCTCGTTTTTCTGGGGACTGTCAGTAGGTGTTGAAC  
CCTGATATTGAAACACGAATCGAAATTCTGCTTAAAAAAGCTGAAAATAGTGGCGTTGATTTACCTAG
```

NCBI Resources How To Sign in to NCBI

SRA SRA Search Advanced Help



SRA

Sequence Read Archive (SRA) makes biological sequence data available to the research community to enhance reproducibility and allow for new discoveries by comparing data sets. The SRA stores raw sequencing data and alignment information from high-throughput sequencing platforms, including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLID System®, Helicos Heliscope®, Complete Genomics®, and Pacific Biosciences SMRT®.

Paleontology/ Skeletal biology

For some analyses, it is important to work with original fossil and skeletal material

- Data, e.g., measurements of individual specimens, should be made available
- Limited access to materials, at minimum due to travel expenses



Paleontology/ Skeletal biology

However, with recent advances in digital imaging technologies:

- Digital measurements can now be more precise than by hand
- Opportunity for automated, scalable, highly reproducible measurements
- Analyses of internal structure is only possible with imaging technology (e.g., CT)



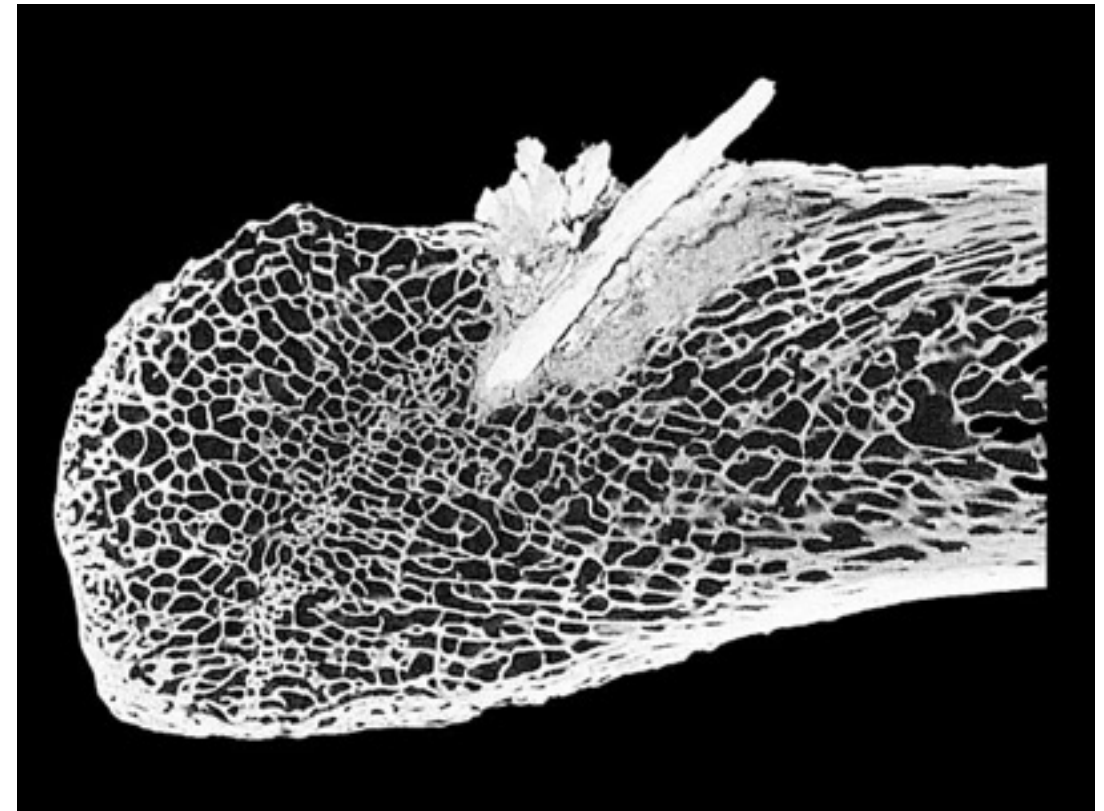
Stored image data facilitate subsequent, and not originally envisioned, analyses

Digital data can be shared!

Paleontology/ Skeletal biology

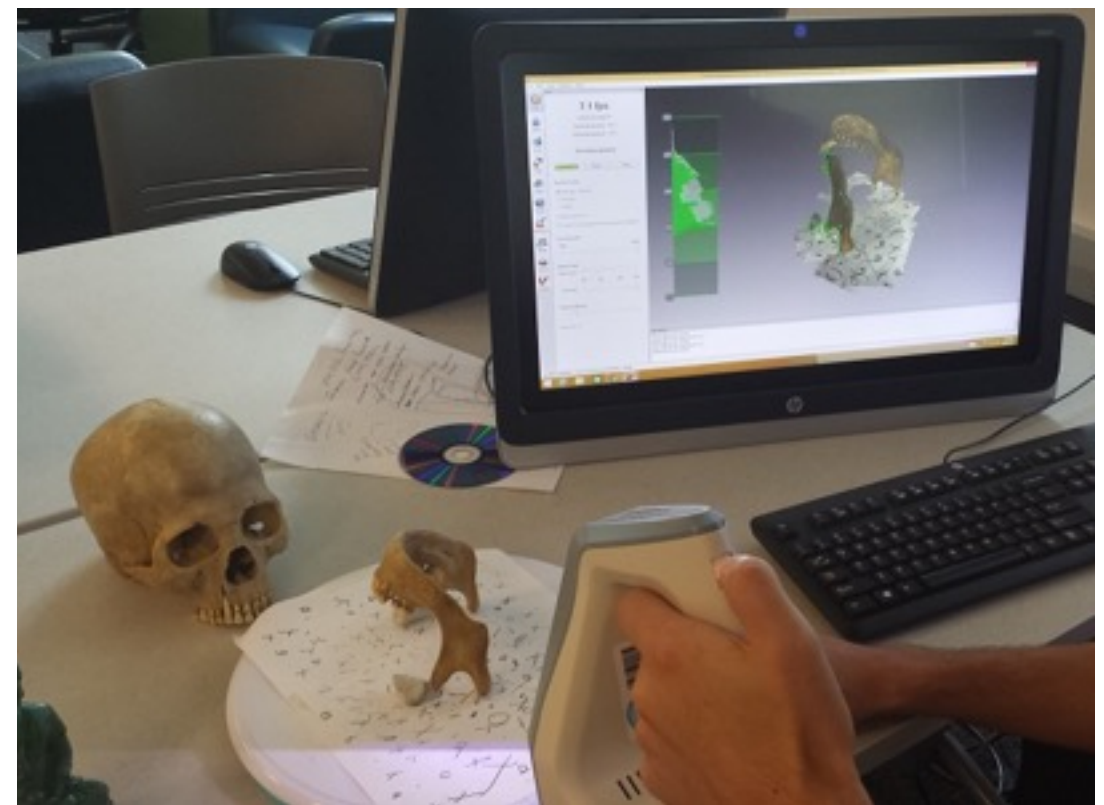
CT scanning - both external and internal surfaces

- Varying resolutions facilitate analyses of different scales of structure




External laser scanning - now sufficient for high quality measurements and analyses of fine-scale shape


- Not all external laser scanning methods provide fine-scale resolution sufficient for most external surface research purposes (although some now do!)
- Relatively inexpensive to collect



Paleontology/ Skeletal biology

Morphosource - NSF funded, cost-free database for open source storage and retrieval of imaging data files



ABOUT | BROWSE | DASHBOARD | 


LOGIN/REGISTER

Getting Started

Find & Download Datasets

BROWSE

 or

enter search terms 

LOGIN OR REGISTER

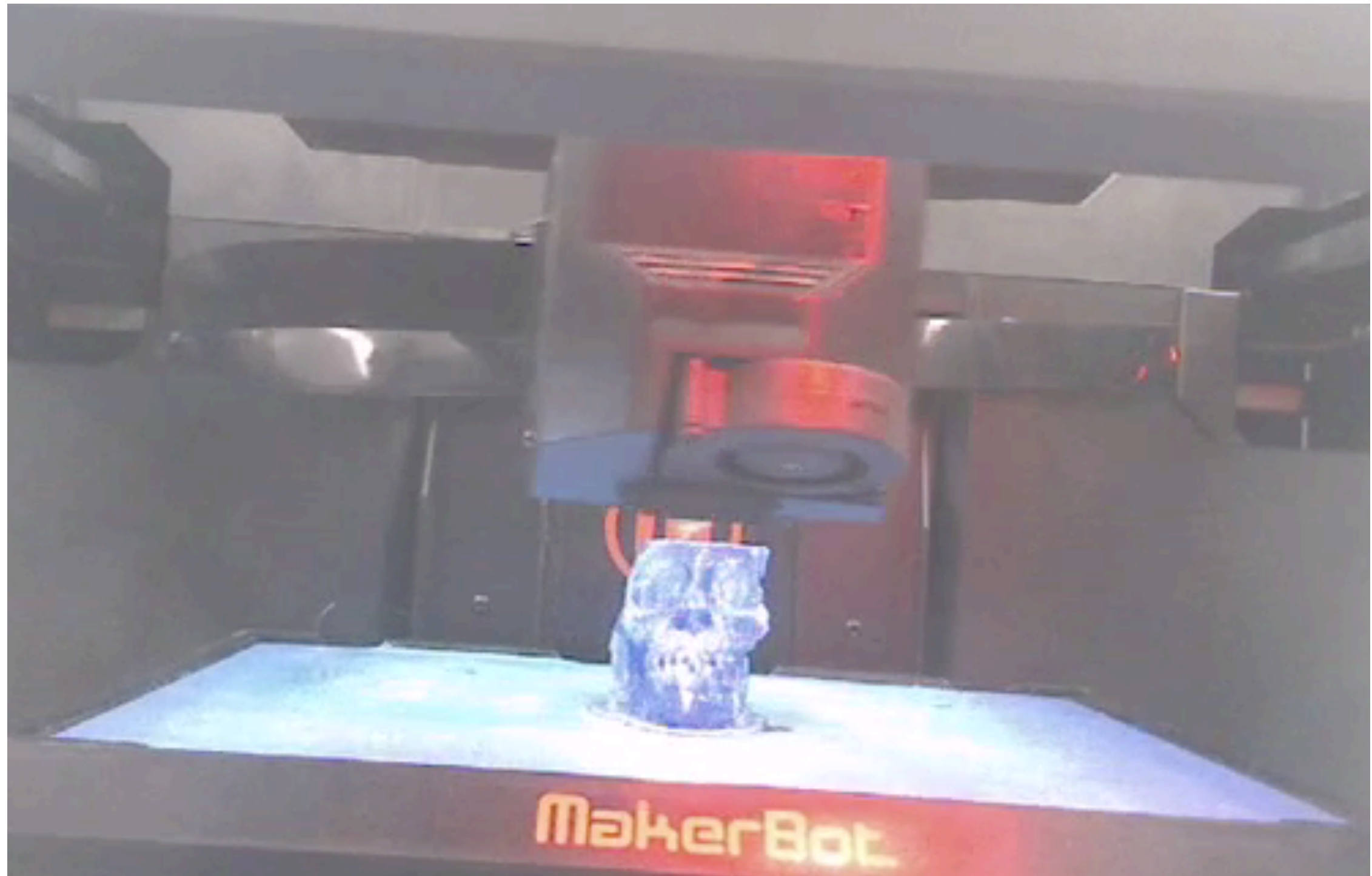
Useful Info

- [Information for Users](#)
- [Information for Contributors](#)
- [Terms](#)
- [User Guide](#)



foot of Daubentonia madagascariensis scanned at 38micron resolution at Duke Evolutionary Anthropology department's new high resolution microCt facility. [Click here if you are interested in details on the facility](#)

Research AND educational benefits



Behavioral ecology

Publications often incorporate analyses of long-term, high-investment, ongoing field data that are expected to be the basis of many subsequent publications



Understandable apprehension with open sharing of all data underlying each paper

- At odds with increasing expectations for publication and funding of data sharing
- Some risk of reduced incentive for initiating or continuing long-term studies

Absence of a current solution; some recent suggestions include:

- Willingness from journals and funders for relatively long data embargos, e.g., 5 yrs
- Increased data tracking processes and communication with data generators

Mills et al. 2015 *Trends in Ecology and Evolution*, Archiving primary data: Solutions for long-term studies.

Whitlock et al. 2016 *Trends in Ecology and Evolution*, A balanced data archiving policy for long-term studies.

What is data management?

A purposeful approach to data across the project lifecycle and beyond.

- Preparation to collect data with materials and permissions conducive to downstream analysis and access goals
- Data management and archiving begin at the research design phase!
- Backup of data as regularly and securely as feasible
- Collection or conversion of data in/ into durable (i.e., digital) and reusable formats
- Record analysis/ processing steps in sufficient detail to ensure reproducibility and consider this record as part of your data
- Permanent data archiving on dedicated public repositories
- ★ With appropriate confidentiality and privacy considerations for data on human subjects, but with planning from the outset (see above) to maximize data accessibility and reproducibility, rather than use as a reason for not sharing data

What is a data management plan?

For a given project or research program, explicit documentation of the approach to data management across the full project lifecycle.

- Increasingly required by funding agencies and universities
- Creates a “contract” with the agency and your scholarly community
- Allows evaluation of appropriate resource availability for execution of the plan
- Facilitates the request for any funding resources necessary to implement the plan’s data management steps
- Permanent data archiving through a permanent repository rather than a personal, laboratory, departmental, or other temporary website or server

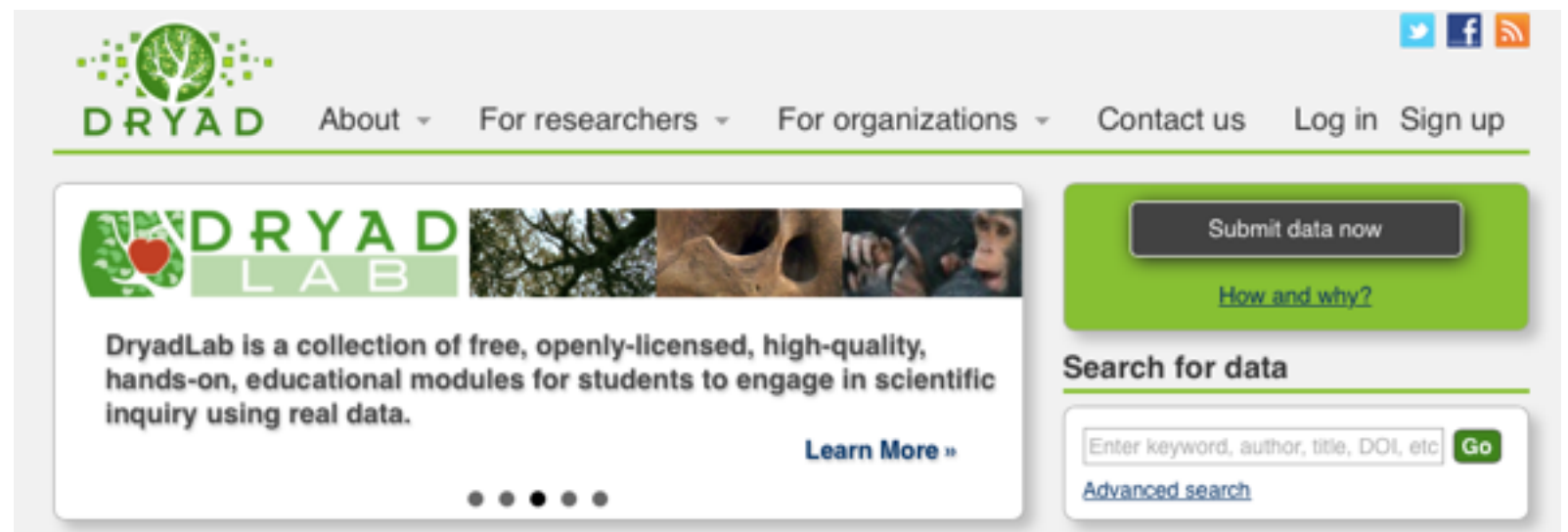
Primary vs. processed data

Consider depositing beyond the minimum required by journals, funders, etc.

Processed data files, rather than only the raw data, and other information can greatly aid reproducibility of the work and maximize impact/ usefulness

- e.g., sequence alignments rather than only raw reads
- e.g., both raw and processed image data
- Code used for analyses

Options include the Dryad Digital Repository, Figshare, GitHub (for code)



Personal or department websites are not acceptable options

- Need reliability of permanent hosting commitment, “too big to fail” status
- Data generators, users both benefit from functionality of community archives

Availability of scripts and integration into the data management process

EDUCATION

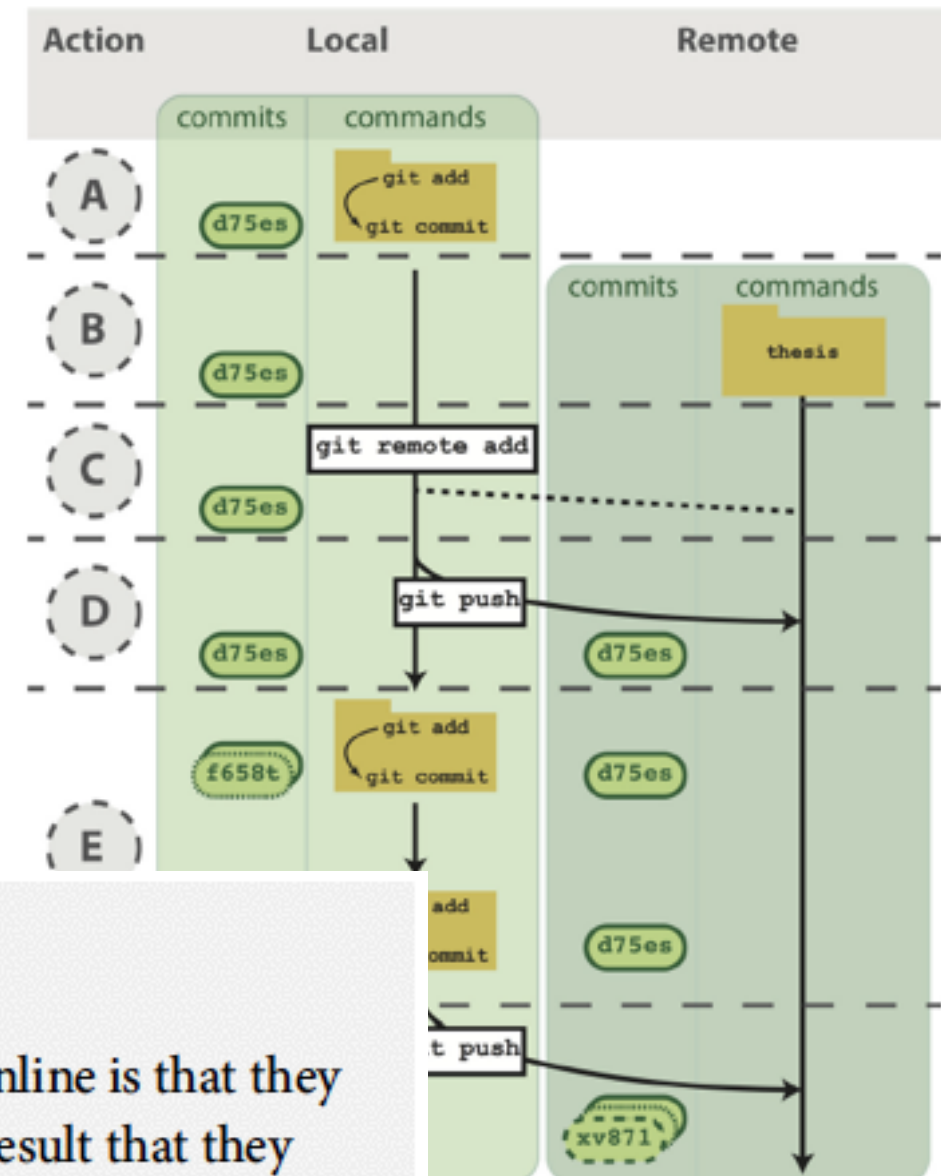
A Quick Introduction to Version Control with Git and GitHub

John D. Blischak^{1*}, Emily R. Davenport², Greg Wilson³

¹ Committee on Genetics, Genomics, and Systems Biology, University of Chicago, Chicago, Illinois, United States of America, ² Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York, United States of America, ³ Software Carpentry Foundation, Toronto, Ontario, Canada

* jdblichak@gmail.com

"This is part of the PLOS Computational Biology Education collection."



Box 5. Being Scooped

One concern scientists frequently have about putting work in progress online is that they will be scooped, e.g., that someone will analyze their data and publish a result that they themselves would have, but hadn't yet. In practice, though, this happens rarely, if at all: in fact, the authors are not aware of a single case in which this has actually happened, and would welcome pointers to specific instances. In practice, it seems more likely that making work public early in something like a version control repository, which automatically adds timestamps to content, will help researchers establish their priority.