

Open Access delle pubblicazioni e dei dati della ricerca

Giovanni Destro Bisol
Dipartimento di Biologia Ambientale

giovanni.destrobisol@uniroma1.it

1. Open Science

2. Open Access
...a case-study

3. Open data
...beyond data sharing

What is Science?

an organized systematic enterprise that gathers knowledge about the world and aims to know and understand the natural and social world through five pillars:

1. confirmation of discoveries & support of **hypotheses** through **repetition** by independent investigators, preferably with different tests & analyses;
2. **mensuration**, the quantitative description of the phenomena on universally accepted scales;
3. **economy**, by which the largest amount of information is abstracted into a simple and precise form, which can be unpacked to re-create detail;
4. **heuristics**, the opening of avenues to new discovery and interpretation.
5. and finally, is **consilience**, the interlocking of causal explanations across disciplines.”

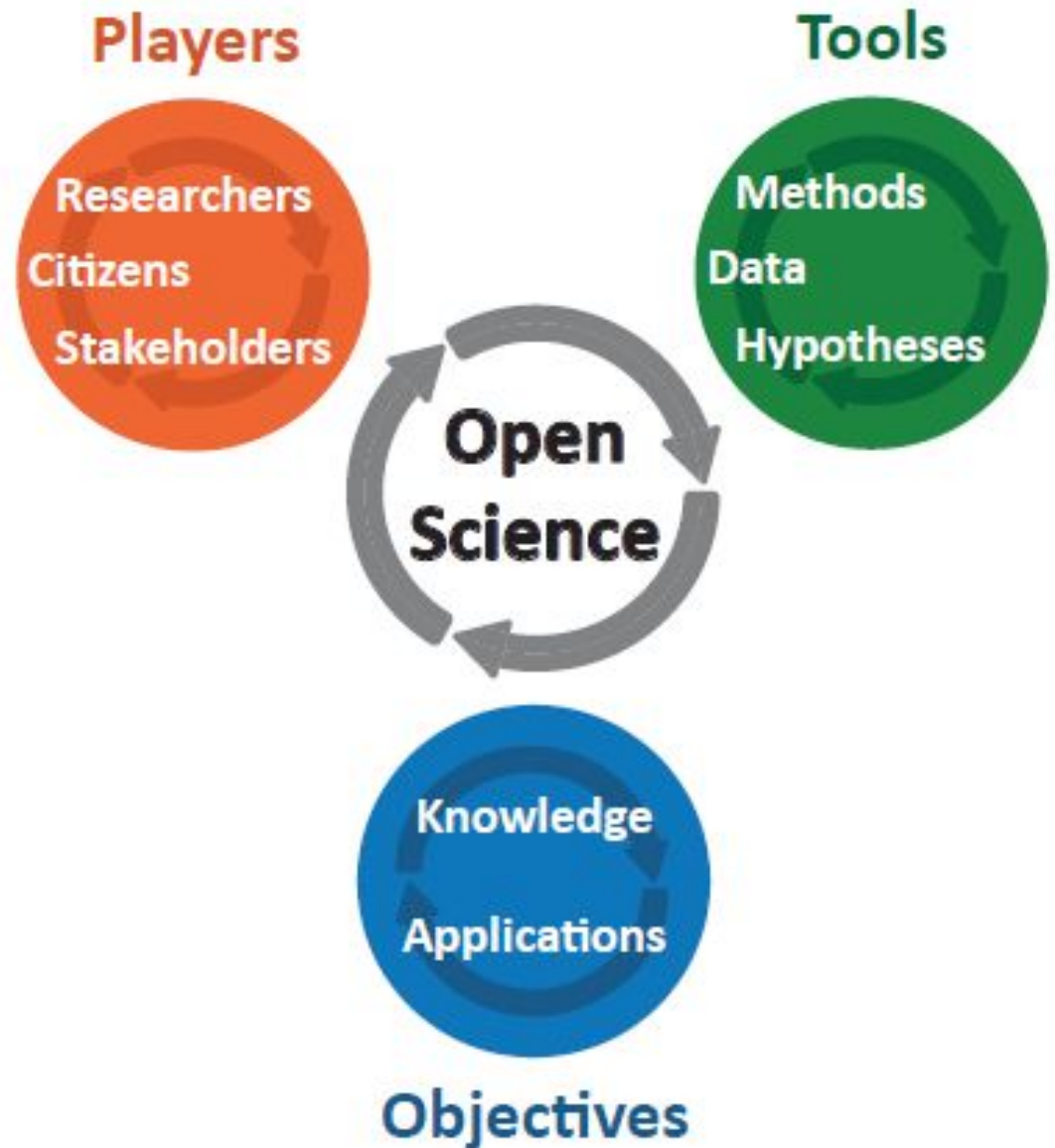
Consilience: “the concurrence of multiple inductions drawn from different data sets”

What is open(ness)?

In a scientific context, the significance of openness goes beyond its immediate meaning of free accessibility of concepts, methods, and data produced by research activities, assuming another implication: making the information shared to an extent that allows others to reproduce a study or experiment in its entirety, what we call **transparency**. While this should be implicit in scientific practices, it is not always implemented for a variety of reasons, making the **research cycle** closed to efficient scrutiny in many cases.

What is Open Science

mutual exchange
cooperation

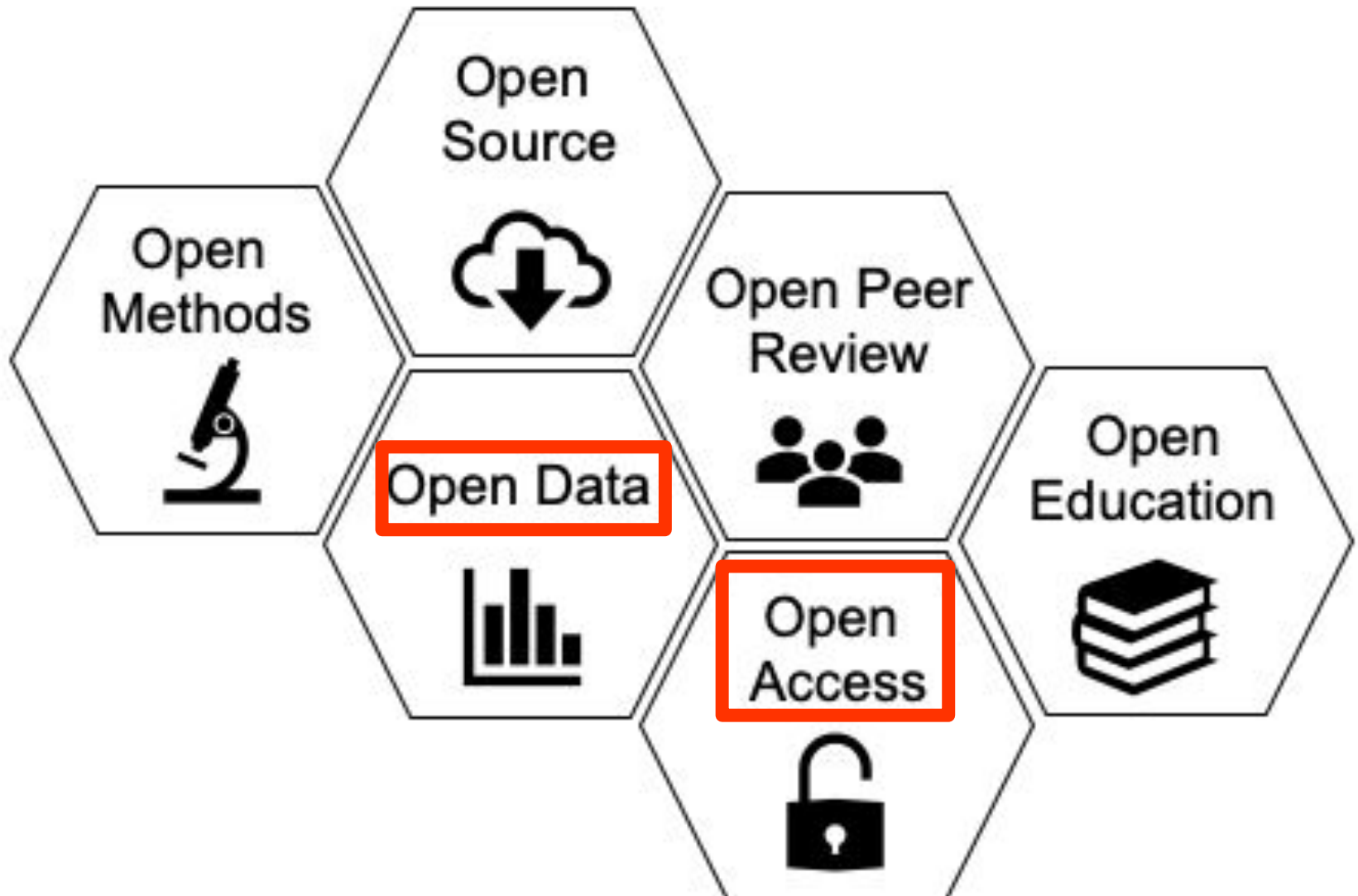


Open Science benefits

- Increase **research efficiency**
- Promote scholarly rigour and enhances **research quality**
- Enhances **visibility** and engagement
- Enables the creation of **new research questions**
- Enhances **collaboration** and community building

but, there are also cons...

How many types of Open Science?



1. Open Science

2. Open Access
...a case-study

3. Open data
...beyond data sharing

Open access to scientific literature and the COVID-19 pandemic

Giovanni Destro Bisol

Paolo Anagnostou

Marco Capocasa

Università La Sapienza
Istituto Italiano di Antropologia

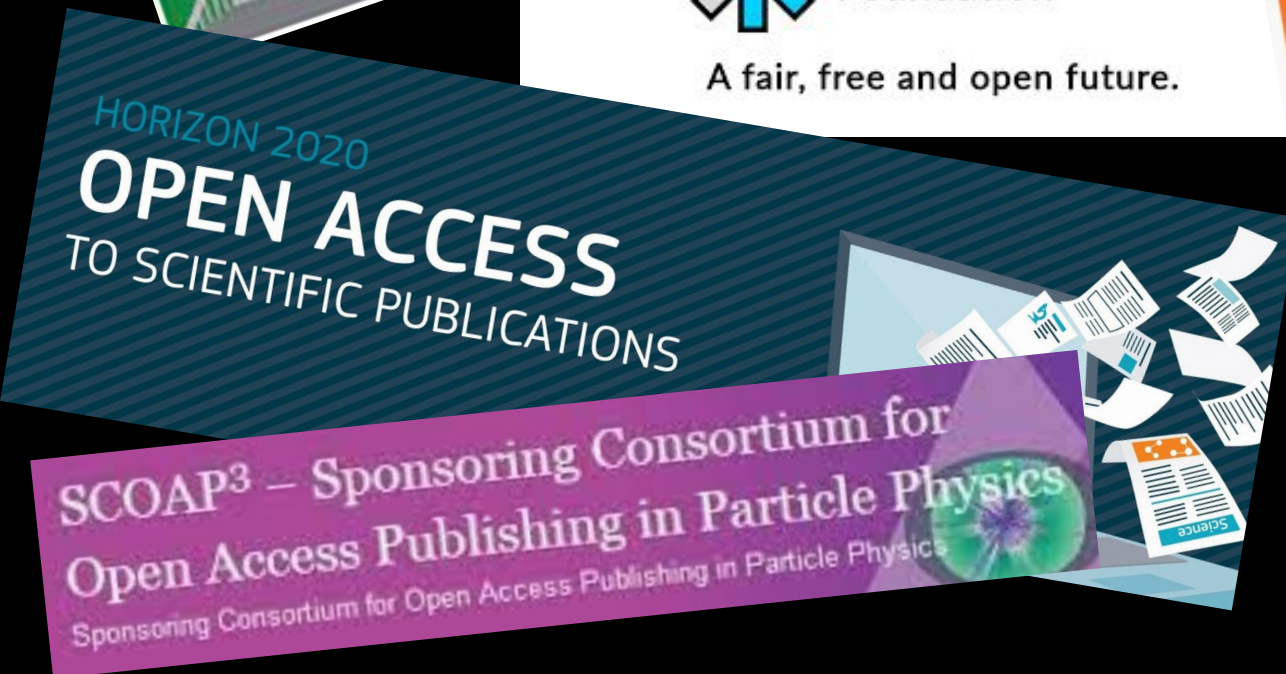
destrobisol@uniroma1.it



Capocasa et al. Information Research, in press

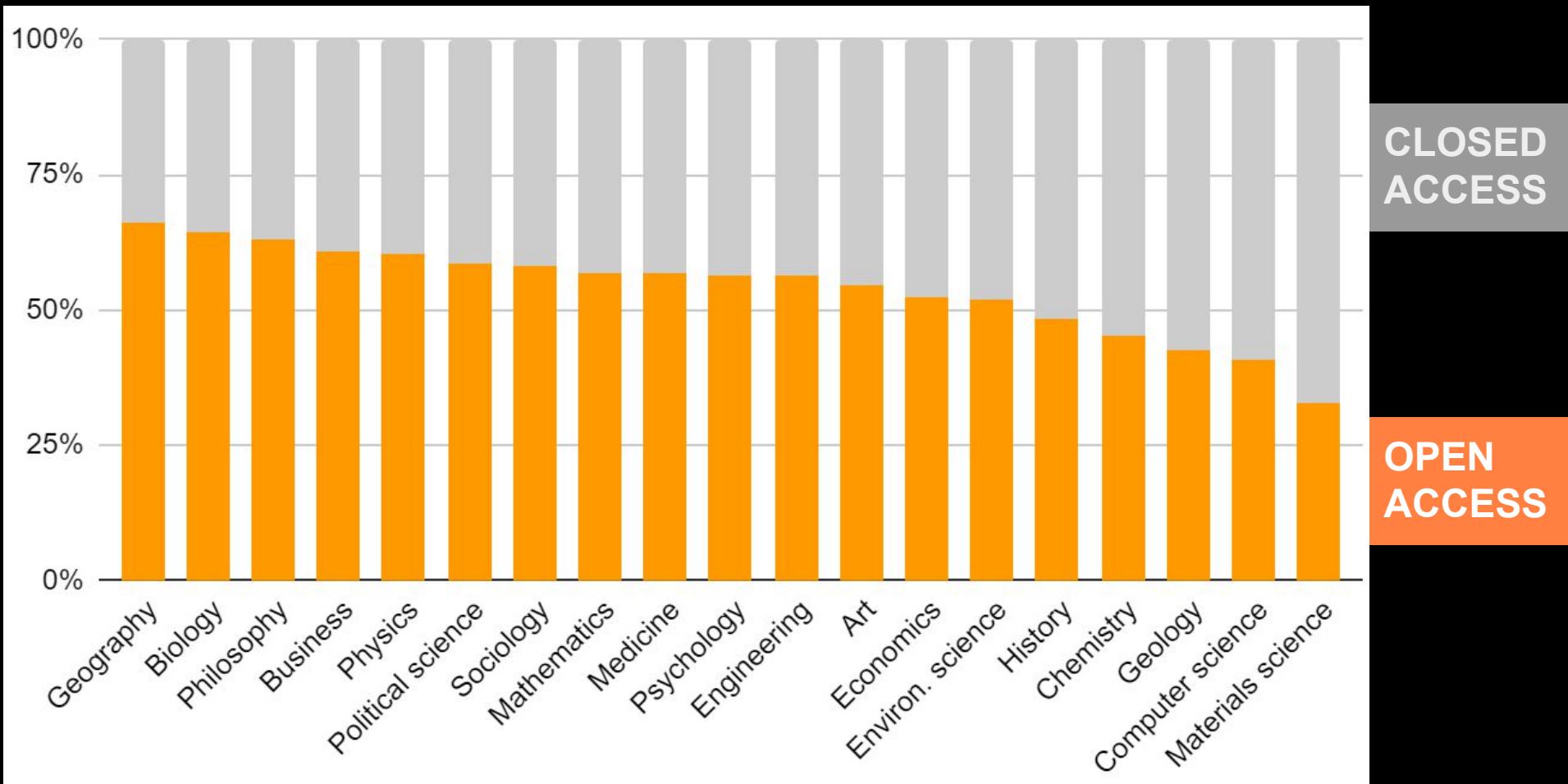
<https://www.medrxiv.org/content/10.1101/2020.07.23.20160481v1>

Despite all this...



We stand here....

Open Access rates, 2019 papers



COVID-19*

Open Access

89.5%

[88.3% – 93.5%]

... a good news

- Web of Science database
- Core collection, acces. July 9th, 2020

13,655 peer-reviewed papers published
from January 1 to July 1, 2020 » almost 10X
of all papers concerning four recent viral
outbreaks

- 95% of OA papers published in journals
with a first quartile Impact Factor

What does this mean?

OK, numerous Journals have made research work on COVID-19 openly accessible

but ...

Achieving much greater open access to health information is possible

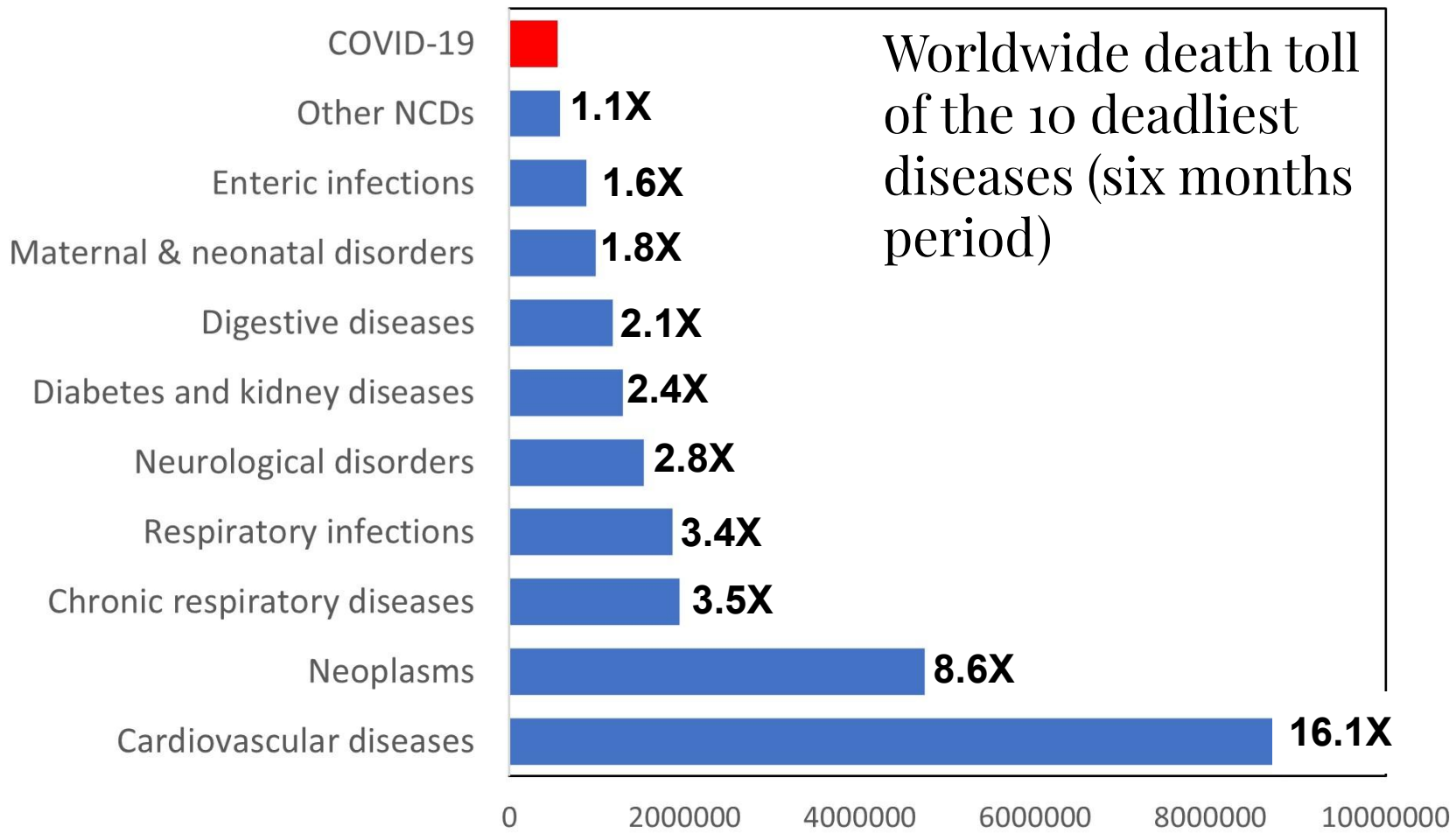
Principle: Scientific knowledge on high impact human diseases should be shared freely and promptly.

So, the next questions...

- How does COVID-19 compare to other diseases?
- How to get closer to COVID-19 OA rates?

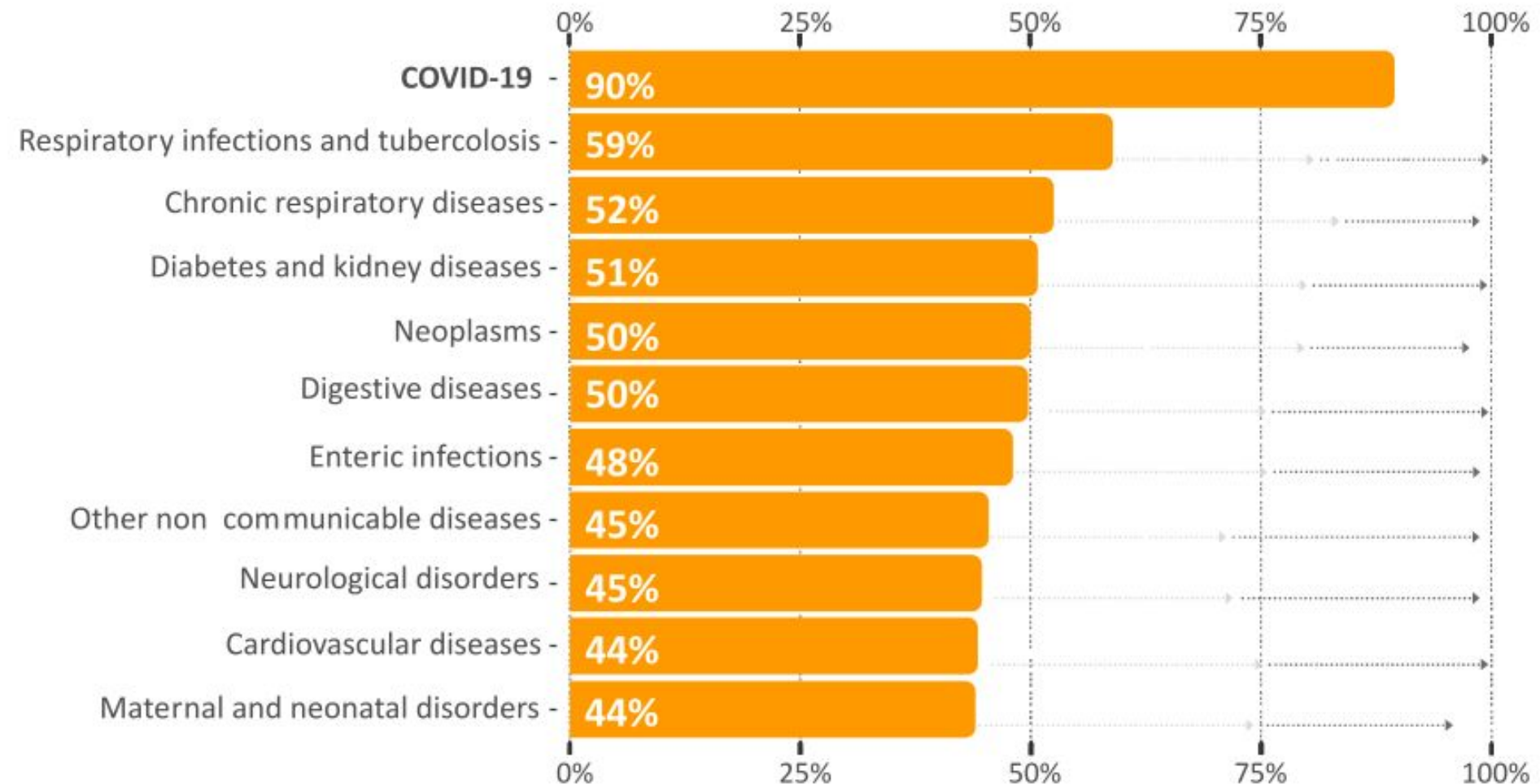


Q1 – How does COVID-19 compare to other diseases?

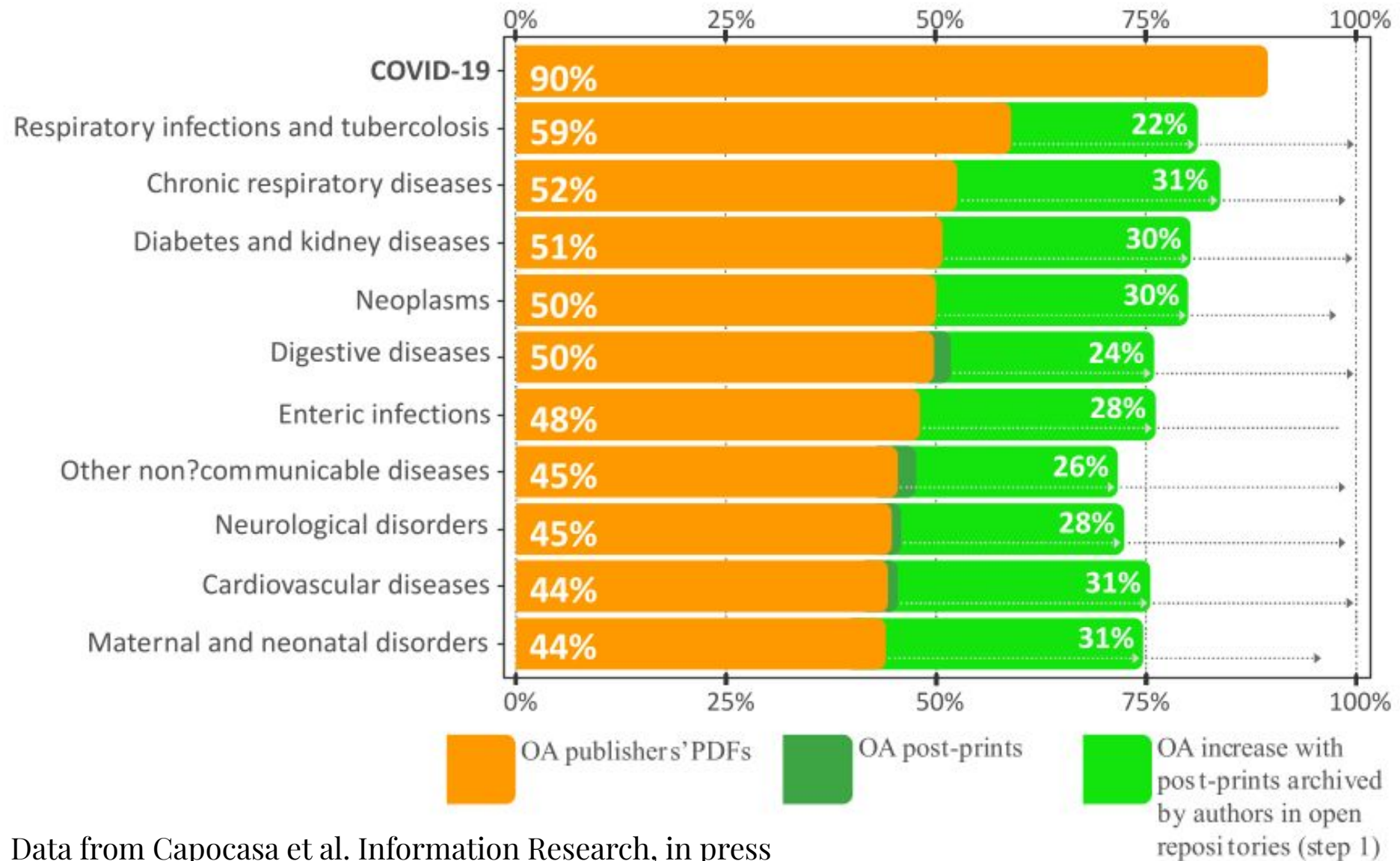


Data from <https://coronavirus.jhu.edu/map.html> (accessed on July 9 and December 21, 00:00 GMT) and GBD 2017 Causes of Death Collaborators (Lancet, 2018)

Q1 – How does COVID-19 compare to other diseases?



Q2 How to get closer to COVID-19 OA rates? ... the green road is the answer



How to make all this more than a wish?

	Step 1	Step 2	Step 3
Level	Individual	Institutions and associations individually	Institutions, associations and some stakeholders, individually or, better, in partnership
Action	Authors archive post-prints online whenever publishers' rules allow it.	Any academic, research and health center, scientific and professional association, and funding agency incentivize the open archiving of post-prints.	Academic, research and health centers, scientific professional and patient associations undertake to make publishers remove restrictions to online post-print archiving.
Time needed	Shorter: can be realized immediately after acceptance of the paper, depending on the authors' willingness.	Intermediate: requires institutional governance deliberation and action	Longer: requires creating synergies and conducting negotiations.
Estimated effects	Increase up to 28.3% in OA rates		Increase up to 21.1% in OA rates

Takeaways

1. By using the green road we could close a substantial portion of the gap in OA between the 10 deadliest diseases and COVID-19.
2. Scientific institutions and associations should more effectively encourage the dissemination of post-print through available online tools.

...and also

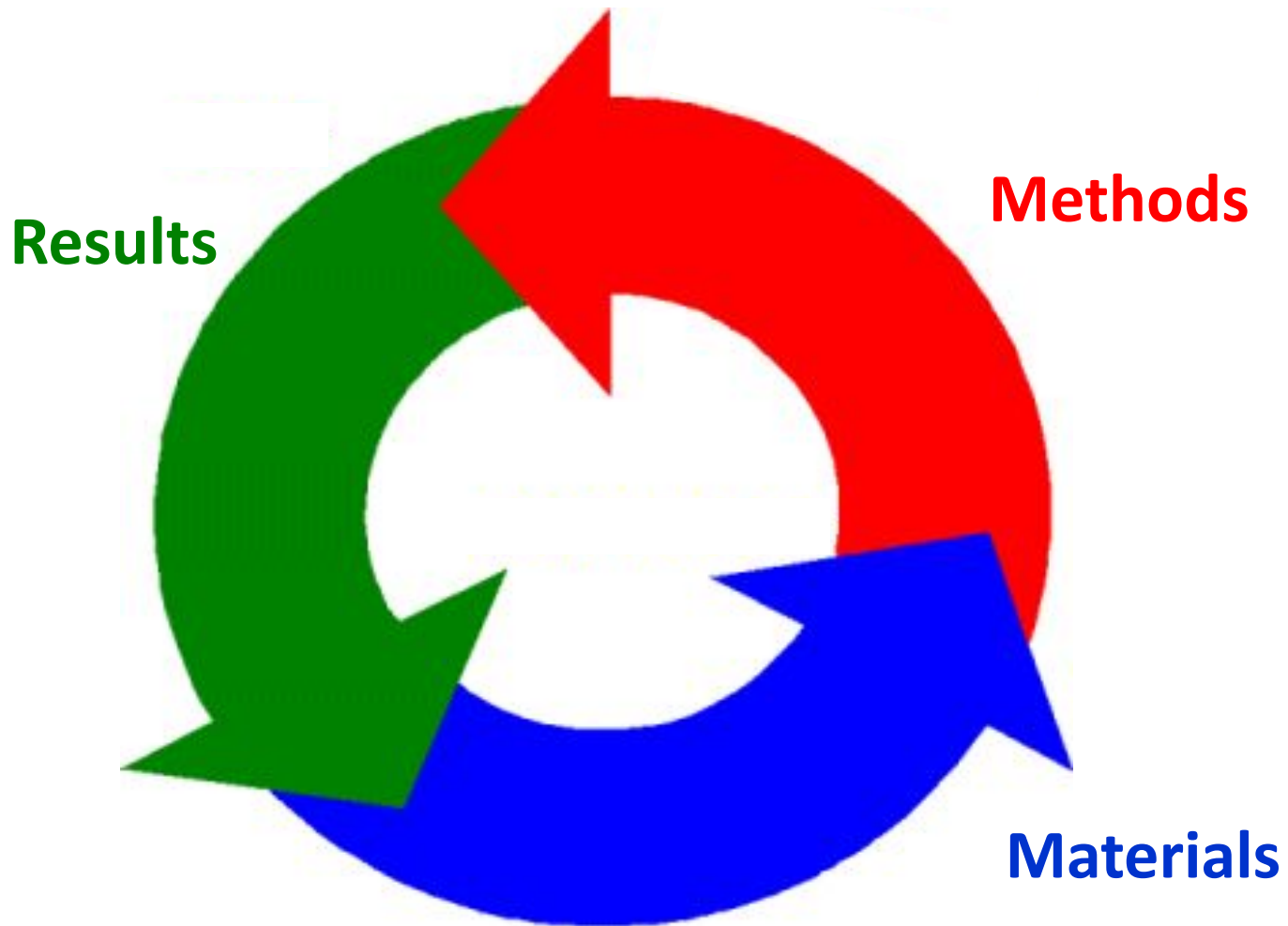
3. Don't let transformative agreements overshadow the green road to Open Access

1. Open Science

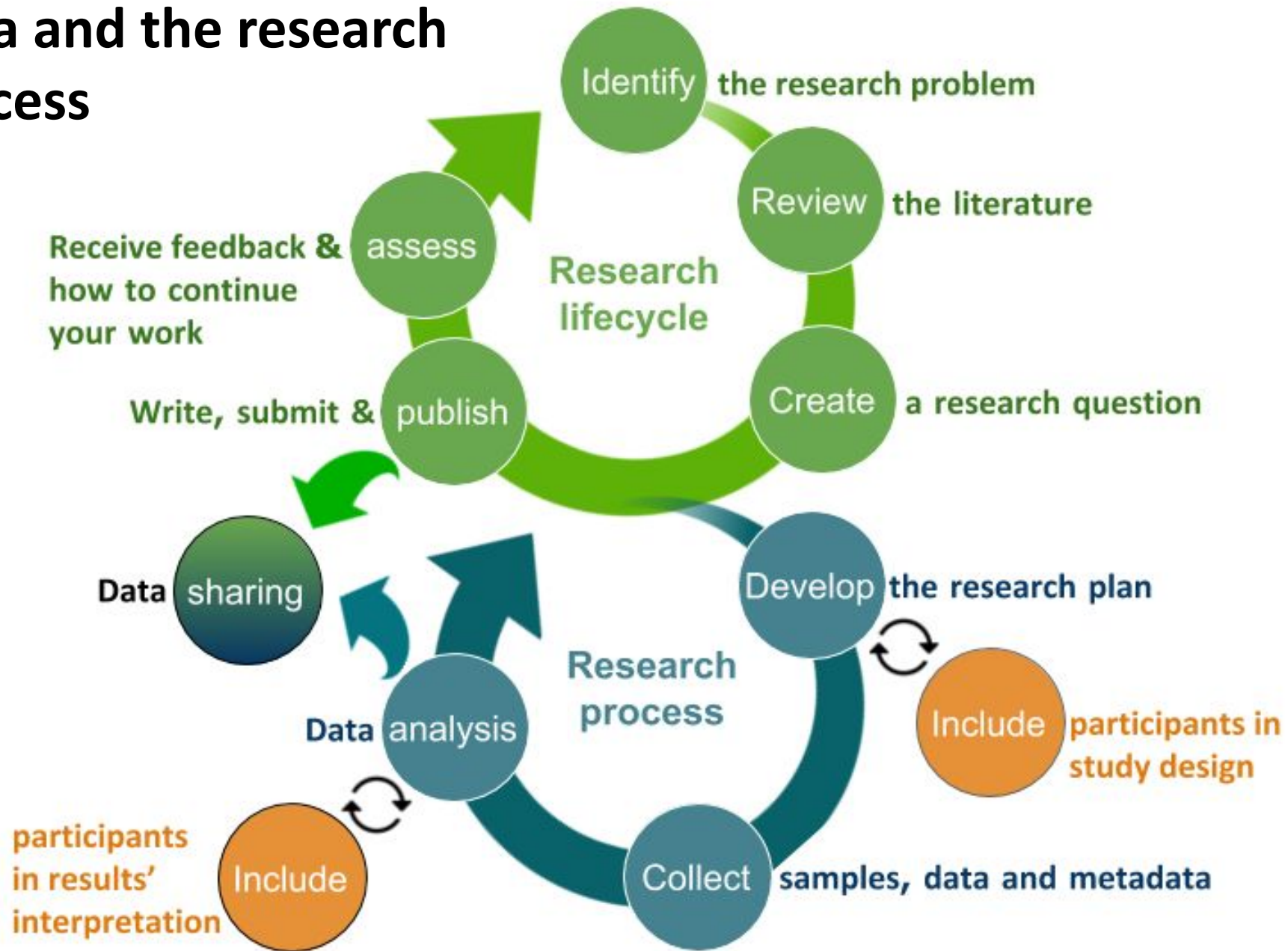
2. Open Access
...a case-study

3. Open data
...beyond data sharing

Open data?



Data and the research process



Beyond data sharing

Intelligent
Openness

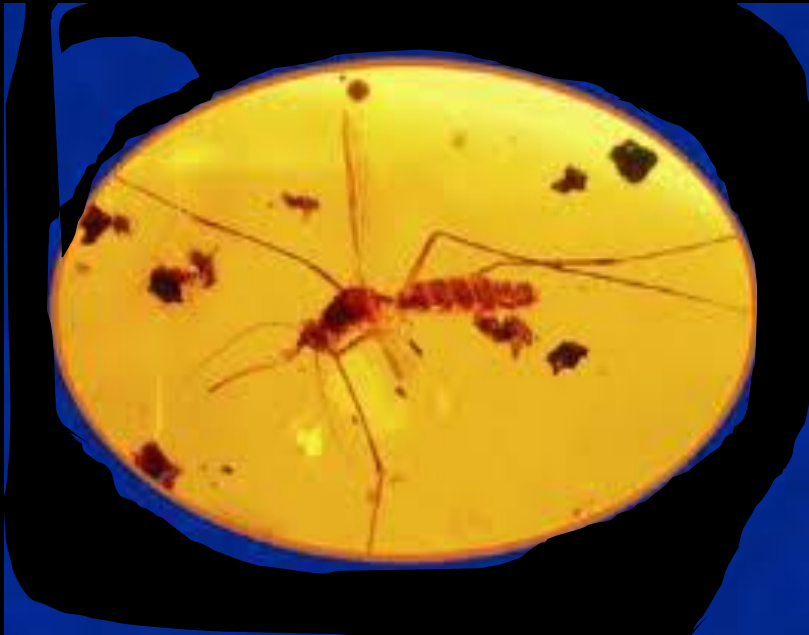
assessability

accessibility
useability

intelligibility

assessability

the data or information's reliability can be evaluated. ...
the results of scientific work must be intelligible
to those wishing to scrutinise them.*



*
Science as an
open enterprise

THE
ROYAL
SOCIETY

When Data Sharing Gets Close to 100 %: What Human Paleogenetics Can Teach the Open Science Movement

Paolo Anagnostou^{1,2*}, Marco Capocasa^{2,3}, Nicola Milia⁴, Emanuele Sanna⁴,
Cinzia Battaglia¹, Daniela Luzi⁵, Giovanni Destro Bisol^{1,2*}

1 Dipartimento di Biologia Ambientale, "Sapienza" Università di Roma, Rome, Italy, **2** Istituto Italiano di Antropologia, Rome, Italy, **3** Dipartimento Biologia e Biotechnologie "Charles Darwin", "Sapienza" Università di Roma, Rome, Italy, **4** Dipartimento di Scienze della Vita e dell'Ambiente, Università di Cagliari, Cagliari, Italy, **5** Istituto di Ricerche sulla Popolazione e le Politiche Sociali, Consiglio Nazionale delle Ricerche, Rome, Italy

* destrobisol@uniroma1.it (GDB); paolo.anagnostou@uniroma1.it (PA)

<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0121409>



OPEN ACCESS

Citation: Anagnostou P, Capocasa M, Milia N, Sanna E, Battaglia C, Luzi D, et al. (2015) When Data Sharing Gets Close to 100%: What Human Paleogenetics Can Teach the Open Science Movement. PLoS ONE 10(3): e0121409. doi:10.1371/journal.pone.0121409

Academic Editor: John Hawks, University of Wisconsin, UNITED STATES

Received: August 1, 2014

Accepted: February 2, 2015

Published: March 23, 2015

Copyright: © 2015 Anagnostou et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are currently provided as Supporting Information files [S1 Table](#) and [S2 Table](#). Data have also been deposited into Zenodo ([Dataset S1 10.5281/zenodo.14804](#), [Dataset S2 10.5281/zenodo.14805](#)).

Funding: This work was supported by the Ministero dell'Istruzione, dell'Università e della Ricerca (PRIN 2009–2011, prot.n. 20097579EW) (<http://www.istruzione.it/>) and the Istituto Italiano di Antropologia (<http://www.ista-roma.org/>). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

This study analyzes data sharing regarding mitochondrial, Y chromosomal and autosomal polymorphisms in a total of 162 papers on ancient human DNA published between 1988 and 2013. The estimated sharing rate was not far from totality ($97.6\% \pm 2.1\%$) and substantially higher than observed in other fields of genetic research (evolutionary, medical and forensic genetics). Both a questionnaire-based survey and the examination of Journals' editorial policies suggest that this high sharing rate cannot be simply explained by the need to comply with stakeholders requests. Most data were made available through body text, but the use of primary databases increased in coincidence with the introduction of complete mitochondrial and next-generation sequencing methods. Our study highlights three important aspects. First, our results imply that researchers' awareness of the importance of openness and transparency for scientific progress may complement stakeholders' policies in achieving very high sharing rates. Second, widespread data sharing does not necessarily coincide with a prevalent use of practices which maximize data findability, accessibility, usability and preservation. A detailed look at the different ways in which data are released can be very useful to detect failures to adopt the best sharing modalities and understand how to correct them. Third and finally, the case of human paleogenetics tells us that a widespread awareness of the importance of Open Science may be important to build reliable scientific practices even in the presence of complex experimental challenges.

Introduction

Making research data openly accessible to the scientific community is one of the main priorities for the global research system. In fact, there is wide consensus that data sharing may help scientific progress allowing a better exploitation of data and an optimized use of resources in a climate of scientific openness and transparency [1–3]. However, there are also considerable barriers to

accessibility

Data must be located in such a manner that it can readily be found and in a form that can be used.*

NCBI Resources How To

PopSet PopSet Limits Advanced

Display Settings: PopSet

Homo sapiens control region, partial sequence; mitochondrial.

PopSet: 375244988

GenBank FASTA

Go to:

Study Details

Evidence of high genetic variation among linguistically diverse populations on a micro-geographic scale: a case study of the Italian Alps.

Coia, V., Boschi, I., Trombetta, F., Cavulli, F., Montinaro, F., Destro-Bisol, G., Grimaldi, S., and Pedrotti, A. (2012) J. Hum. Genet. 57:(4)254-260

PMID: 22418692 Citation

Go to:

Sequences in this data set

JQ623902.1	Homo sapiens isolate 190_VNs control region, partial sequence; mitochondrial
JQ623901.1	Homo sapiens isolate 189_VNs control region, partial sequence; mitochondrial
JQ623900.1	Homo sapiens isolate 188_VNs control region, partial sequence; mitochondrial
JQ623899.1	Homo sapiens isolate 184_VNs control region, partial sequence; mitochondrial
JQ623898.1	Homo sapiens isolate 183_VNs control region, partial sequence; mitochondrial

NCBI GenBank PubMed Entrez BLAST

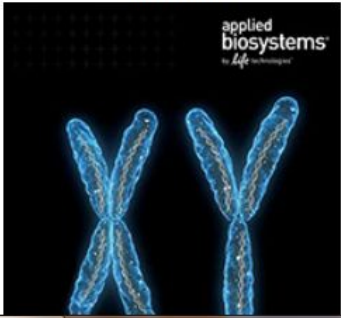
- seminars, oral and poster presentations, informal discussions
- Publications (electronic supplements)
- Online Database

Browser window showing the YHRD (Y Chromosome Haplotype Reference Database) website. The address bar displays <http://www.yhrd.org/>. The page features a world map with red dots representing haplotypes, labeled "R42: 108949 haplotypes". Navigation links include "Search", "Analyse", "Research", "Contribute", and "Meet". A prominent pink button labeled "Download Manual" is visible.

WELCOME TO THE Y CHROMOSOME HAPLOTYPE REFERENCE DATABASE (YHRD)

The ability to identify male-specific DNA renders polymorphic Y-chromosomal sequences an invaluable addition to the standard panel of autosomal loci used in forensic genetics (REF Roewer 2009). Y-STR haplotyping is particularly important for sensitive typing of male DNA in mixed stains as well as for rapid assortment of biological crime scene evidence. Moreover, Y chromosomal profiling can trace back paternal lineages into the past and has thus been proven a useful tool in genealogical and kinship testing. The individuality of the male-specific part of the Y chromosome can be optimally explored by the Y-STR haplotype analysis using a set of highly variable short tandem repeat markers approved by the forensic and scientific community. An extremely informative Y-STR core set or minimal haplotype (minHt) amplifiable in a multiplex reaction has been recommended for court use : DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS385ab (REF Kayser et al. 1997 and REF Pascali et al. 1999). This core haplotype can be extended by other hypervariable Y-STR loci (DYS438, DYS439, DYS437).

advertisement



SEARCH YHRD

LATEST NEWS

Release 42
We have updated the YHRD with 3,452 new haplotypes from 14 populations submitted by 16 research groups. The YHRD has now 108,949 haplotypes [...]
Posted 1 month ago by Lutz Roewer

Release 41
We have updated the YHRD with 1,324 new haplotypes from 13 populations submitted by 7 research groups. The YHRD has now 105,498 haplotypes [...]
Posted 4 months ago by Lutz Roewer

Release 40 with new features

16:49
18/02/2013

http://empop.org/

Posta in arrivo (8) - giovanni.d... Publications | Mappa Submission form YHRD - Contribute Home :: EMPOP.org - Mito...

File Modifica Visualizza Preferiti Strumenti ?

Google empop Effettua la ricerca Segnalibri Controllo Traduci Altro

Il F... Post... .ITA... Corr... La R... Home... JASs Isita Goog... Acce... Prev... PLoS... open...

Trova: popset Precedente Successivo Opzioni

Version: 2.2, Release: 9 Username/Email Login Register

EMPOP

Home
Contribute
Help
Imprint
Contact
Terms of use

Query
Tools
Account
endorsed by

ISFG
© 1999-2013
IMI

News Introduction Concept Alignment Users Collaborators Support

Sample update (2013-02-14)
We have become aware of an alignment variant that does not follow the phylogenetic principle. Although this does not affect search results due to the alignment-free query engine SAM (Röck et al., 2011) we decided to update the sample. Please note that the release status of EMPOP remains unchanged. We thank Kimberly Andreaggi (AFDIL) for pinpointing this haplotype. ...

Release 9 is online, N=29444 (2013-01-17)
We have updated the database with 3.372 new haplotypes from 15 different countries (Angola, Bahrain, China, Croatia, Dominican Republic, Germany, Haiti, Honduras, Jamaica, Japan, Mexico, Philippines, Portugal, Somalia, Spain).

Sample update (2013-01-17)
One of our collaborators notified us of two maternally closely related samples. We therefore deleted one of the samples (Argentina). Thorough review and feedback from our collaborators brought changes in polymorphisms in five samples. New data of release 9 led to changes in the alignment according to the recent phylogeny for 13 samples. Please recall that alignment changes do not affect the search result due to the alignment-free query engine SAM (Röck et al., 2011). ...

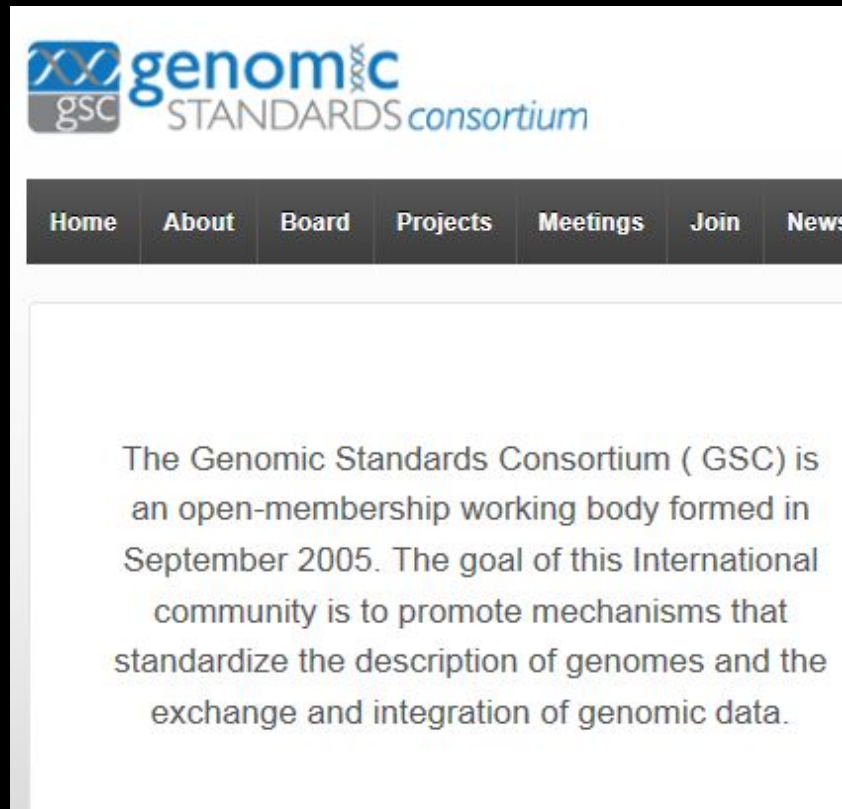
DNA in Forensics 2014 (2013-01-10)
The Belgian National Institute of Criminalistics and Criminology (NICC) will host DNA in Forensics 2014 (9th International Y Chromosome User Workshop and 6th International EMPOP Meeting) in Brussels, Belgium in spring of 2014. ...

Release 8 is online, N=26073 (2012-09-05)
Thanks to the outstanding efforts of our collaborators, especially the Armed Forces DNA Identification Laboratory (AFDIL), release 8 presents 8,752 new haplotypes from Argentina, Brazil, Colombia, Iraq, Spain, USA, Uganda. The majority of the new data comes from ...

16:51
18/02/2013

useability

.in a format where others can use the data or information..proper background information and metadata.



What to share?

Data and results

Metadata

structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource

- **WHO** created the data?
- **WHAT** is the content of the data?
- **WHEN** were the data created?
- **WHERE** is it geographically?
- **HOW** were the data developed?
- **WHY** were the data developed?



POPULATION METADATA

* Name (Population/ethnic group/tribe) Fare clic qui per immettere testo.

* Continent

* Region Fare clic qui per immettere testo.

* Nation Fare clic qui per immettere testo.

* Sampling location (Name, latitude, longitude) Fare clic qui per immettere testo.

Language (Ethnologue classification) Fare clic qui per immettere testo.

Marital behaviour Fare clic qui per immettere testo.

Other features Fare clic qui per immettere testo.

Other features Fare clic qui per immettere testo.

Other features Fare clic qui per immettere testo.

Comments

Fare clic qui per immettere testo.

INDIVIDUAL METADATA

- * ID Fare clic qui per immettere testo.
- * Birth date Fare clic qui per immettere testo.
- * Birth location Fare clic qui per immettere testo.
- * Gender Fare clic qui per immettere testo.
- Other languages spoken Fare clic qui per immettere testo.
- * Relationship with other donors:
 - Degree of relationship
 - ID of related individual Fare clic qui per immettere testo.
 - Sampling location Fare clic qui per immettere testo.
 - Population subunit (Clan, Tribe, Cast) Fare clic qui per immettere testo.
 - Religion Fare clic qui per immettere testo.
 - Father birth location Fare clic qui per immettere testo.
 - Mother birth location Fare clic qui per immettere testo.
 - Paternal grandfather birth location Fare clic qui per immettere testo.
 - Paternal grandmother birth location Fare clic qui per immettere testo.
 - Maternal grandfather birth location Fare clic qui per immettere testo.
 - Maternal grandmother birth location Fare clic qui per immettere testo.
- Comments

What is a Metadata Standard?

- A Standard provides a structure to **describe** data with:
 - Common terms to allow consistency between records
 - Common definitions for easier interpretation
 - Common language for ease of communication
 - Common structure to quickly locate information
- In **search** and **retrieval**, standards provide:
 - Documentation structure in a reliable and predictable format for computer interpretation
 - A uniform summary description of the dataset



Multiple Metadata Standards Exist: Examples

NCBI GenBank Overview

PubMed Entrez BLAST OMIM Books Taxonomy Structure

Search for

NCBI
SITE MAP
Submit to GenBank
BankIt
Sequin

What is GenBank?

GenBank[®] is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences ([Nucleic Acids Research 2004 Jan 1;32\(1\):23-6](#)). There are approximately 37,893,844,733 bases in 32,549,400 sequence records as of February 2004 (see [GenBank growth statistics](#)). As an example, you may view the [record](#) for a *Saccharomyces cerevisiae* gene. The complete [release notes](#) for the current version of GenBank are available. A new release is made every

Homo sapiens isolate TAL966 D-loop, partial sequence; mitochondrial

GenBank: HQ651672.1

[FASTA](#) [Graphics](#) [PopSet](#)

[Go to:](#) ☐

LOCUS HQ651672 360 bp DNA linear PRI 12-JUN-2011

DEFINITION Homo sapiens isolate TAL966 D-loop, partial sequence;
mitochondrial.

ACCESSION HQ651672

VERSION HQ651672.1 GI:334866457

KEYWORDS .

SOURCE mitochondrion Homo sapiens (human)

ORGANISM [Homo sapiens](#)

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
Catarrhini; Hominidae; Homo.

REFERENCE 1 (bases 1 to 360)

AUTHORS Coia,V., Destro-Bisol,G., Verginelli,F., Battaggia,C., Boschi,I.,
Cruciani,F., Spedini,G., Comas,D. and Calafell,F.

TITLE Brief communication: mtDNA variation in North Cameroon: lack of
Asian lineages and implications for back migration from Asia to
sub-Saharan Africa

JOURNAL Am. J. Phys. Anthropol. 128 (3), 678-681 (2005)

PUBMED [15895434](#)

REFERENCE 2 (bases 1 to 360)

AUTHORS Coia,V., Destro-Bisol,G., Verginelli,F., Battaggia,C., Boschi,I.,
Cruciani,F., Spedini,G., Comas,D. and Calafell,F.

TITLE Direct Submission

JOURNAL Submitted (24-NOV-2010) Department of Animal and Human Biology,
University La Sapienza, Rome, Italy

FEATURES Location/Qualifiers

source 1..360

Welcome to Sequin

Misc

Sequin

Sequin Application Version 12.30
Standard Release [Nov 13 2012]

National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
(301) 496-2475
info@ncbi.nlm.nih.gov

Database for submission ☐ GenBank ☒ EMBL

[Start New Submission](#)

[Read Existing Record](#)

[Network Configure](#)

[Show Help](#)

[Quit Program](#)

What is the Value to Data Users?

- Metadata gives a user the ability to:
 - Search, retrieve, and evaluate data set information from both inside and outside an organization
 - Find data: Determine what data exists for a geographic location and/or topic
 - Determine applicability: Decide if a data set meets a particular need
 - Discover how to acquire the dataset you identified; process and use the dataset



Intelligibility

Audiences need to be able to make some judgment of what is communicated. ..to judge the nature of the claims...



The image is a screenshot of a web page from PLOS Genetics. At the top, the PLOS logo is on the left, and the words "BROWSE", "PUBLISH", and "ABOUT" are on the right. Below the logo, the text "GENETICS" is visible. Further down, there are icons for "OPEN ACCESS" and "PEER-REVIEWED". Below that, it says "RESEARCH ARTICLE". The main title of the article is "Genetic Variation and Population Structure in Native Americans". Below the title, the authors are listed: Sijia Wang, Cecil M Lewis Jr., Mattias Jakobsson, Sohini Ramachandran, Nicolas Ray, Gabriel Bedoya, Winston Rojas, Maria V Parra, Julio A Molina, Carla Gallo, Guido Mazzotti, Giovanni Poletti, Kim Hill, and Andrés Ruiz-Linares. There is a link to "view all" authors. At the bottom, it says "Published: November 23, 2007" and provides a DOI link: "https://doi.org/10.1371/journal.pgen.0030185".

PLOS | GENETICS

BROWSE PUBLISH ABOUT

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Genetic Variation and Population Structure in Native Americans

Sijia Wang , Cecil M Lewis Jr. , Mattias Jakobsson , Sohini Ramachandran, Nicolas Ray, Gabriel Bedoya, Winston Rojas, Maria V Parra, Julio A Molina, Carla Gallo, Guido Mazzotti, Giovanni Poletti, Kim Hill, [...], Andrés Ruiz-Linares [view all]

Published: November 23, 2007 • <https://doi.org/10.1371/journal.pgen.0030185>

Abstract

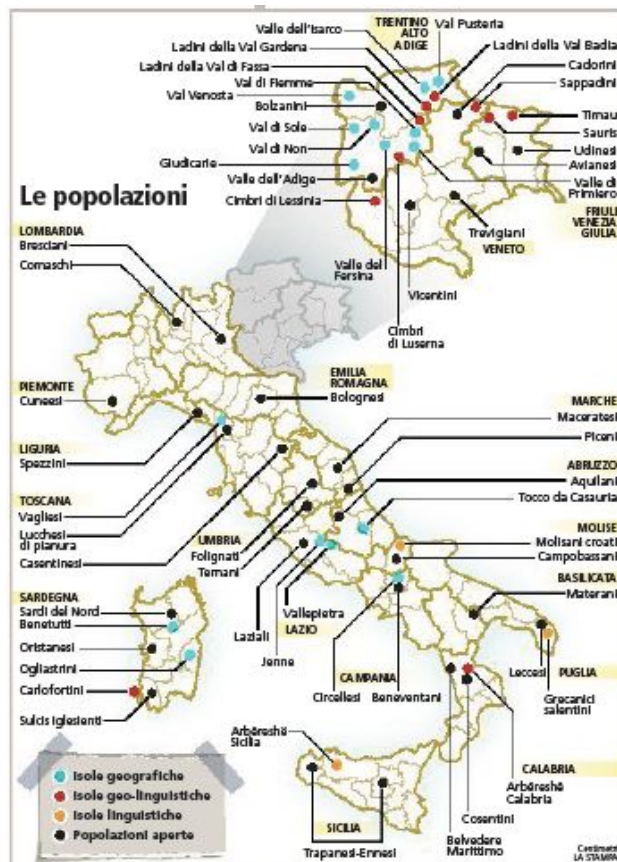
We examined genetic diversity and population structure in the American landmass using 678 autosomal microsatellite markers genotyped in 422 individuals representing 24 Native American populations sampled from North, Central, and South America. These data were analyzed jointly with similar data available in 54 other indigenous populations worldwide, including an additional five Native American groups. The Native American populations have lower genetic diversity and greater differentiation than populations from other continental regions. We observe gradients both of decreasing genetic diversity as a function of geographic distance from the Bering Strait and of decreasing genetic similarity to Siberians—signals of the southward dispersal of human populations from the northwestern tip of the Americas. We also observe evidence of: (1) a higher level of diversity and lower level of population structure in western South America compared to eastern South America, (2) a relative lack of differentiation between Mesoamerican and Andean populations, (3) a scenario in which coastal routes were easier for migrating peoples to traverse in comparison with inland routes, and (4) a partial agreement on a local scale between genetic similarity and the linguistic classification of populations. These findings offer new insights into the process of population dispersal and differentiation during the peopling of the Americas.

...make your findings accessible to a wide audience that includes both scientists and non-scientists.

Author Summary

Studies of genetic variation have the potential to provide information about the initial peopling of the Americas and the more recent history of Native American populations. To investigate genetic diversity and population relationships in the Americas, we analyzed genetic variation at 678 genome-wide markers genotyped in 29 Native American populations. Comparing Native Americans to Siberian populations, both genetic diversity and similarity to Siberians decrease with geographic distance from the Bering Strait. The widespread distribution of a particular allele private to the Americas supports a view that much of Native American genetic ancestry may derive from a single wave of migration. The pattern of genetic diversity across populations suggests that coastal routes might have been important during ancient migrations of Native American populations. These and other observations from our study will be useful alongside archaeological, geological, and linguistic data for piecing together a more detailed description of the settlement history of the Americas.

Linguistic, geographic and genetic isolation: a collaborative study of Italian populations

27 DICEMBRE 2013 **il venerdì**

re sarde.
al centro
genetici
lamento
ografico



SPAGNOLI E RUMENI PER **DNA** RISULTANO PIÙ VICINI
DI CERTE NOSTRE COMUNITÀ NELLA STESSA REGIONE

DALLE ALPI ALLA SARDEGNA NOI ITALIANI SIAMO UN POPOLO DI BIODIVERSI

di **Giuliano Aluffi**

bacino del Mediterraneo, uno ad aver favorito la varietà del



Presentazione del
progetto
alle comunità

Continuazione della
ricerca

Consenso
Informato

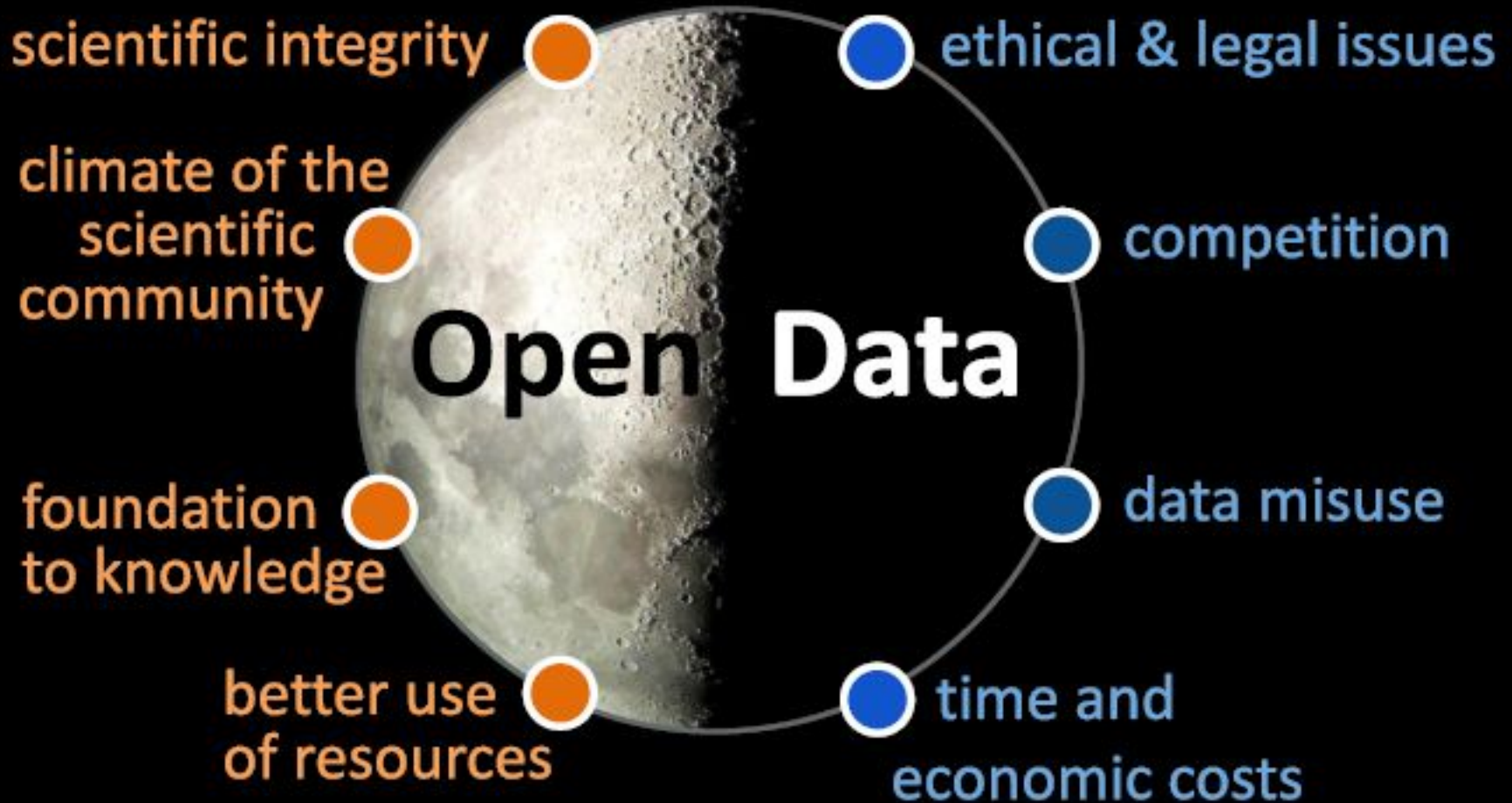
Restituzione dei risultati

- forma orale
- forma scritta

Questions

- A. How data sharing may foster research progress?
- B. Are there cons for data sharing
- C. What data sharing tell us about the relation between ethics and science?

Pros & Cons



A. How data sharing may foster research progress?

Sharing broadens the scope and perspective of research

- Broadens Scope of Research

Complexity of Science: when scientists share their expertise and/or the fruits of their research, the overall effort moves more quickly and to greater depths

- Diversifies Perspective

enlarge the pool of researchers who can work on a topic, it also is likely to increase the diversity of that pool

- Contributions from institutions having limited resources

A. How data sharing may foster research progress?

Sharing Allows Resources to Be Used More Efficiently

- Reduces Costs—Both Money and Effort

Risk of “excessive” duplication

- Maximizes Use of Data

data sets too large to be fully explored

- Reduces Subject Burden

Human and animal...

A. How data sharing may foster research progress?

Sharing Enhances the Climate of the Scientific Community

- Corrects Error of Analysis

misjudge the specificity of their assays, use inappropriate statistics, or fail to recognize a bias in their subject population: University Group Diabetes Program

- Discourages Fraud and Enhances Confidence

misconduct are less likely to occur

Withholding suggests that one is trying to hide something

- Promotes creativity

Favours a climate of openness and transparency

cross-pollination of thoughts

B. Are there cons for data sharing?

Negative Career Impact

Need for Publications

Authors can expect to realize multiple publications from a resource, dataset...

Potential for Non-Replication

the recipient might not have the requisite skills and qualifications necessary to use the materials appropriately

Lack of Recognition and Increased Competition

scientists are not recognized for sharing their resources and datasets (but they are for papers)

B. Are there cons for data sharing?

Limited Resources

Time and financial Cost

extra effort or money necessary for converting the data set or learning the software necessary for depositing data to a central repository

.....to successfully deposit experimental data from approximately 30 microarrays into a data repository may take several weeks for a novice.....

Personnel Required

Experienced personnel may be needed to prepare the material for distribution

Availability

Once investigators complete their work with a given resource there may be little reason to maintain the resource

Infrastructures

Are reliable and maintained infrastructures (archives, repositories) available?

B. Are there cons for data sharing?

Data misuse

- Misappropriation

...a geologist had experiences in which other scientists published data he had shared while he was still working on his own analysis

- Misinterpretation

. . . and if I don't have some relationship of trust then I don't know whether they're going to, you know, just go off and do something and never check with me to see, well, was this a good interpretation. . . .

- Disregard of good faith practices

...incidents where deposited data had been 'cherry-picked' to make claims about the efficacy of certain products in marketing materials...

C. What data sharing tell us about the relation between ethics and science?

Alarming shift away from sharing results

There was the review of two H5N1 avian influenza virus studies in ferrets by the US National Science Advisory Board for Biosecurity in December 2011. The board

In December 2011 a bill was proposed to the US Congress to reverse the National Institutes of Health policy that all taxpayer-funded research should be freely accessible online (see go.nature.com/uvj68l). The bill's

Science should be available for evaluation by other scientists and for public scrutiny, just as it has been since Galileo's time. It should not be heading for epistemological suicide as a result of vested interests or a creeping loss of awareness of the theory of knowledge.

Benefits and Risks of Influenza Research: Lessons Learned

Anthony S. Fauci* and Francis S. Collins

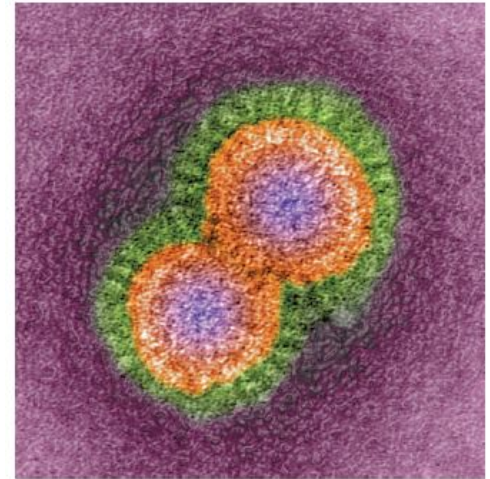


Fig. 1. H5N1 avian influenza virus particles, colored transmission electron micrograph. Magnification: $\times 670,000$ when printed 10 cm wide.

two NIH-funded studies of H5N1 transmissibility and pathogenesis in ferrets. In those studies, H5N1 viruses were made transmissible via respiratory droplets among ferrets by engineering the virus; well-described and published protocols including reverse genetics, reassortment, and passaging of viruses in mammals were used.

the research results could be used by bioterrorists to intentionally cause harm, or that an accidental release of a pathogen from a laboratory could inadvertently cause harm.

3 Ethical and Legal Issues

National Security

Government regulations may prohibit researchers from sharing some types of information or materials (e.g., pathogens, missile technology, encryption Software and even human genomes) because of security concerns

3 Ethical and Legal Issues

Subject Confidentiality

Privacy issues: it may not be possible to maintain the value of the data set while stripping the data of all personal identifiers

Identifying Personal Genomes by Surname Inference

Melissa Gymrek,^{1,2,3,4} Amy L. McGuire,⁵ David Golan,⁶ Eran Halperin,^{7,8,9} Yaniv Erlich^{1*}

Sharing sequencing data sets without identifiers has become a common practice in genomics. Here, we report that surnames can be recovered from personal genomes by profiling short tandem repeats on the Y chromosome (Y-STRs) and querying recreational genetic genealogy databases. We show that a combination of a surname with other types of metadata, such as age and state, can be used to triangulate the identity of the target. A key feature of this technique is that it entirely relies on free, publicly accessible Internet resources. We quantitatively analyze the probability of identification for U.S. males. We further demonstrate the feasibility of this technique by tracing back with high probability the identities of multiple participants in public sequencing projects.

SCIENCE VOL 339 18 JANUARY 2013