JASs Journal of Anthropological Sciences
Open Access
Research articles Vol. 99 (2021), pp. 135 - 152
doi.10.4436/jass.99016

The emerging complexity of Open Science: assessing Intelligent Data Openness in Genomic Anthropology and Human Genomics

Paolo Anagnostou[1,2], Marco Capocasa[1,2], Francesca Brisighelli[3], Cinzia Battaggia[1] & Giovanni Destro Bisol[1,2]

1) Department of Environmental Biology, Sapienza University of Rome, Piazzale Aldo Moro 5, Rome, Italy

e-mail: paolo.anagnostou@uniroma1.it, giovanni. destrobisol@uniroma1.it

2) Istituto Italiano di Antropologia, Piazzale Aldo Moro 5, Rome, Italy

3) Forensic Genetics Laboratory, Department of Health Surveillance and Bioethics, F. Policlinico Gemelli IRCCS Roma - Università Cattolica del Sacro Cuore, Italy

Summary - *In recent decades, the scientific community has become aware of the importance of science being effectively open in order to speed up scientific and technological progress. In this context, the achievement of a robust, effective and responsible form of data sharing is now widely acknowledged as a fundamental part of the research process. The production and resolution of human genomic data has steadily increased in recent years, mainly due to technological advances and decreasing costs of DNA genotyping and sequencing. There is, however, a downside to this process due to the huge increase in the complexity of the data and related metadata. This means it is advisable to go beyond traditional forms of sharing analysis, which have focused on data availability only. Here we present a pilot study that aims to complement a survey on the availability of data related to peer-reviewed publications with an analysis of their findability, accessibility, useability and assessability (according to the "intelligent data openness" scheme). Sharing rates in genomic anthropology (73.0%) were found to be higher than human genomics (32.4%), but lower than closely related research fields (from 96.8% to 79.2% for paleogenetics and evolutionary genetics, respectively). We discuss the privacy and methodological issues that could be linked to this finding. Comparisons of sharing rates across a wide range of disciplines has suggested that the idea of human genomics as a forerunner for the open data movement should be questioned. Finally, both in genomic anthropology and human genomics, findability and useability were found to be compliant with the expectations of an intelligent data openness, whereas only a minor part of studies met the need to make the data completely assessable.*

Keywords - *Research lifecycle, Open Data, Findability, Accessibility, Useability, Assessability, Privacy, Personal Identification.*

Defining Open Science (from an anthropological perspective)

What is Open Science? The question can be answered in different ways, mostly depending on whether the emphasis is on: infrastructure, knowledge, impact measurement, social values and collaborative research (Fecher and Friesike 2013). In this article, we will try to combine some of these alternative views, looking at Open Science from an anthropological perspective, as an endeavour to share knowledge not only with the scientific community but also with (and for) society as a whole.

The term combines two complex and multifaceted concepts: open(ness) and science. Lacking a shared definition of Open Science means we must define both words explicitly in order to make our discourse clear. So let's start from the latter. Among the many definitions

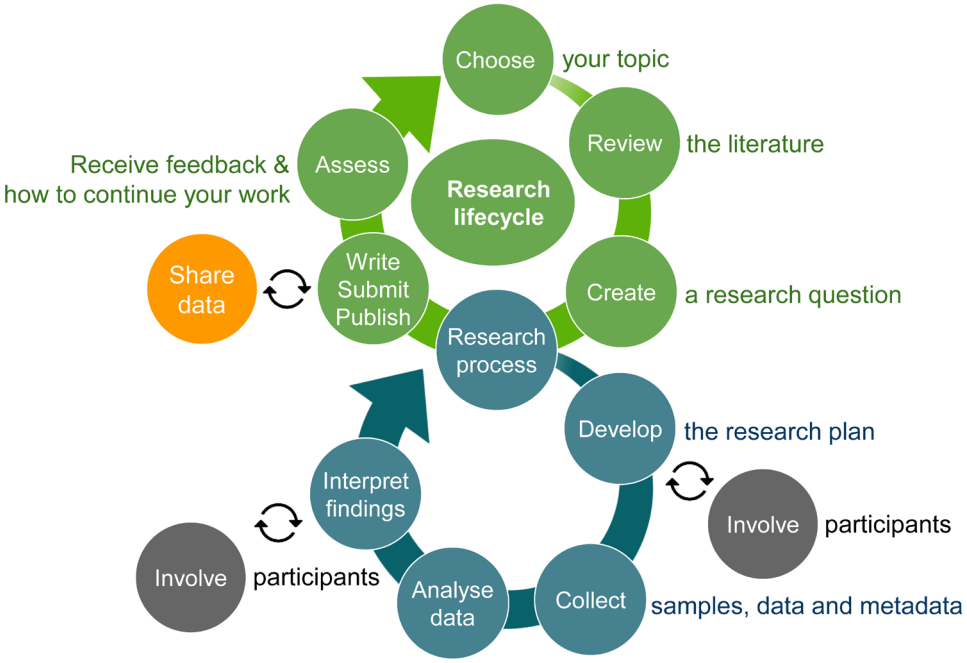

***Fig. 1** - The scheme of a research lifecycle and research process which emphasises the involvement of participants (grey circles) in the development of the research plan and interpretation of findings.*

offered, we can start from Quinn (2009, p. 8): "Science is a process, based on interpretation of experimental or observational data using models and theories, within a strictly constrained logical structure". It aims to know and understand the natural and social world through five pillars: (i) repeatability: investigating the same phenomenon with new approaches in order to challenge its interpretation; (ii) mensuration: describing the properties of physical objects and living organisms unambiguously, using universally accepted methods and scales; (iii) economy: summarising any scientific information (e.g. a series of measurements) in the simplest form, but without any loss of information; (iv) heuristics: by following unexpected avenues of research, science leads to new knowledge which allows further testing of the scientific principles; (v) consilience: the convergence of different lines of evidence towards a comprehensive theory of phenomena (Wilson 1998). Consilience is particularly characteristic in anthropology, due to its intrinsic interdisciplinary nature and epistemological mission to achieve a holistic view of human phenomena.

Coming to openness, its significance goes beyond its immediate meaning of free accessibility of concepts, methods, and data produced by research activities. There is another important implication: guaranteeing that the information is 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 practises, it is not always implemented for a variety of reasons, making the research cycle closed to efficient scrutiny in many cases. Openness may assume an additional meaning in anthropological research, which goes beyond a strict scientific dimension. Due to their background, biological anthropologists should be more aware of how to collaborate and share decisions with community members starting from the early stages of a research project in

order to make it more sustainable, inclusive to all those involved in research, and attentive to their social values (McInnes 2011; Garrison et al. 2018). This helps avoid possible misunderstandings resulting from communication difficulties and adapt scientific practises to the culture and the expectations of communities (Sharp and Foster 2002; Low and Merry 2010; Schensul et al. 2015). Furthermore, prospective participants can better understand the "whys" and "hows" of the study and can make informed choices regarding their degree of involvement (Harry et al. 2000). There are two steps in the research process which are crucial to implement this inclusive approach: the development of the research plan and the interpretation of findings (Fig. 1).

To fully understand the values underlying Open Science, it is necessary to conceptualise science as an organised highly cooperative enterprise in which there is a mutual exchange of information between all components. Cooperative behaviours have been a key factor in the evolution and adaptive success of *Homo sapiens* (Apicella and Silk 2019). Therefore, it is not surprising to find them at the heart of what human beings do to advance their knowledge of the world, with the ultimate goal of bringing new knowledge, progress, and well-being to society. The essay "Science and technology in a democratic order", which was published by the American sociologist Robert King Merton (1942) decades before the surge of the Open Science movement, is an indispensable reference in this regard. The "institutional imperatives that together constitute the ethos of modern science" (universalism, communism, disinterest, and organised scepticism) express an essentially cooperative and cumulative vision of the scientific enterprise. More than 40 years later, in 1985, Chubin introduced and defined the term "Open Science" for the first time in the scientific discourse, basing his idea on Merton's vision of Science.

In Figure 2, we have modelled Open Science considering three interacting components: players, tools and objectives. Players include all those involved in research production, application and dissemination: researchers, citizens and (other) stakeholders (e.g. political and institutional

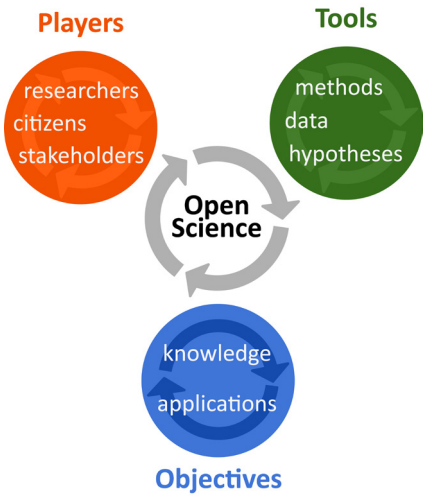

***Fig. 2** - A scheme depicting Open Science as an organised cooperative enterprise. Circular arrows indicate the reciprocal feedback within and among the three components (players, tools and objectives).*

decision-makers, managers, publishers and science communicators) (Gura 2013). Awareness of each other's expectations facilitates and makes interactions among players more productive.

Data, methods and hypotheses (which we have indicated collectively as tools) are at the heart of the scientific method. Advancements in scientific methods increase the quantity and quality of research data, reshape scientific practises, challenge current theories, and promote new ones. This is exemplified by the increasing use of data-based approaches favoured by the increasing availability of big data, with examples (and caveats) also coming from genomic studies of cultural and social phenomena (e.g. Goisauf et al. 2020). The integration of big data, machine learning and artificial intelligence is expected to promote a better understanding of the linkages between genotypes and phenotypes (Ramsay et al. 2019).

As regards the objectives of science, a distinction is usually made between basic (or pure) and applied research. A paradigmatic example of the porosity of this dichotomy is provided by the Human Genome Project which has reshaped the way biological discovery is practised, paving

the way for both the understanding of genetic variation and the advancement of research into cures for genetically related diseases (Hood and Rowen 2013; Gibbs 2020). The growing development of applied anthropology - a younger subfield which aims to solve practical problems outside the academic dimension - shows that the boundary between basic and applied research is becoming increasingly blurred (Van Willigen 2002; Trotter et al. 2014).

The feedback among the three main components creates a virtuous circle: while players use tools to achieve their objectives, new knowledge, stemming from basic or applied research, may generate novel hypotheses to be tested through more powerful data which require innovative methods to be produced.

In summary, Open Science can be defined as a process in which three main components (actors, tools and objectives) act synergistically to produce and share new knowledge in a way that does not only lead to scientific and technological advances, but which is also attentive to societal values and individual rights. Thanks to the systemic sharing and cooperation that occurs within and between components, research carried out in an Open Science perspective is more collaborative, bottom-up, creative and innovative than in scenarios where secrecy prevails over openness (e.g. industrial, military and pharmaceutical research). By encouraging players to share values, expectations, and responsibilities, Open Science helps them become more aware of the fundamental value of reproducibility, heuristics, and consilience.

Open data

No research problem can be framed or any scientific question defined without the support of data. In all scientific disciplines, experimental data are indispensable to test hypotheses, create models and build theories, while their sharing and reuse - what is commonly understood by Open Data - is fundamental to advance knowledge. Therefore, data have a central role in the research life cycle and research process (Fig. 1). The actual usefulness of the data collected to answer the research question depends on the accuracy with which the experimental design is drawn up and, vice versa, the development of an effective research plan requires an accurate identification of all the data to be collected, analysed and compared. Data (and metadata) may be shared either after or even before the publication of the study (Birney et al. 2009; but see Amann et al. 2019). Although the latter practice has been so far limited to great publicly funded projects (Birney et al. 2009), the proliferation of data journals is expected to incentivize this procedure in the near future (Bierer et al. 2017).

Today, making research data available for reuse and scrutiny is an explicit priority for biological and biomedical research (De Silva and Vance 2017), and this has been helped in recent years by the introduction of new computer-assisted technologies and digitization techniques (Sansone et al. 2018). Accordingly, various strategies have been set up by academic institutions and funding bodies to encourage researchers to share their results, including the development of Open Data policies by institutions and journals, community-driven initiatives and, more recently, the creation of data journals (Hrynaszkiewicz et al. 2020; Walters 2020; see also the CODATA, Research Data Alliance and Foster web sites). The *Journal of Anthropological Sciences* was the first anthropological journal to ask authors to deposit the research data of accepted papers (Anagnostou and Destro Bisol 2011). More recently, other journals like the *Journal of Human Evolution* and the *American Journal of Physical Anthropology* (*American Journal of Biological Anthropology* starting from 2022) have adopted an Open Data policy (Turner and Mulligan 2019).

More focus on genomic anthropology

Genomic anthropology is currently a very vital research area. It deals with various aspects of the natural history of our species, including,

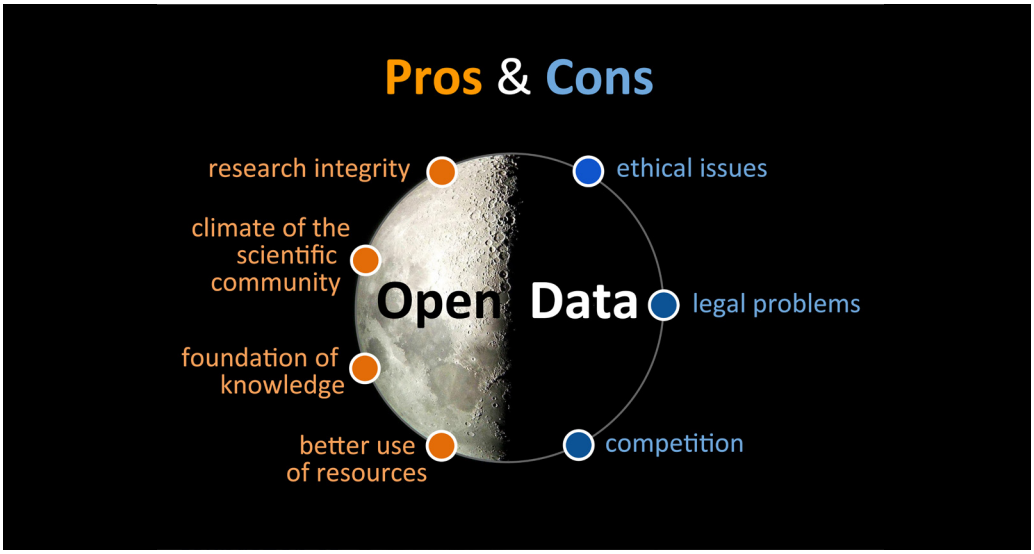

Fig. 3 - Pros and Cons of Open Data in genomic anthropology pictured using the metaphor of the moon.

among others, the characterization of ancient DNA, the history of human populations and population processes, and the role of adaptive processes in shaping the genetic diversity of our species (Jobling et al. 2004; Destro Bisol et al. 2010). As for any anthropological sub-field, the horizon of genomic anthropology embraces cultural aspects (e.g. archaeological, linguistic, social), with the ultimate goal of obtaining consilient explanations of human phenomena. Recently, the exploitation of data has further been developed through the use of complex spatiotemporal models that require sophisticated bioinformatic and statistical tools and their integration with ecological evidence (Pagani and Destro Bisol 2021). However, even recognizing promising methodological and conceptual advances, anthropologists should be aware of the risk that genomics could reintroduce a racial view when data are used assuming simplistic and rigid schemes of human diversity (Lee et al. 2001; Dunklee 2003; Kowal and Llamas 2019).

Compared to other research fields in biological anthropology, studies on DNA variation, both at genetic and genomic levels, have two characteristics that can make data sharing easier and more effective. Indeed, the encoded nature of DNA, its primary source of information, makes data comparability far superior to, for example, quantitative or quasi-qualitative traits of teeth and bones. However, as in other research fields (Fischer and Zigmond 2010), opening data is not without caveats and has its pros and cons that stem from the inevitable tension between openness and secrecy. Looking at the pros (the light side of the moon, Fig. 3), Open Data can help put into practice the principles of research integrity (honesty, accountability and fairness) by facilitating the detection of errors, falsification and fabrication (OECD 2007). We previously discussed how sharing data in human paleogenomics helped overcome its credibility crisis in the early 1990s caused by the discovery that some important published findings were due to contamination with endogenous DNA and/or Polymerase Chain Reaction artefacts (Anagnostou et al. 2015). It also restored a climate of confidence in ancient human DNA studies, which paved the way for their subsequent development, which was enabled by the application of next-generation sequencing (NGS) techniques in the mid-2000s (Rehm 2017).

Another important advantage is that data on human genomic variation can be archived in and retrieved from large repositories (e.g. Gene Expression Omnibus, Sequence Read Archive, European Genome-phenome Archive) or large international projects (International HapMap 3 Consortium 2010; 1000 Genomes Project Consortium 2015). This means that knowledge can be advanced in two ways: data can be re-examined to shed light on questions that were not answered in the original studies, while larger and more powerful datasets which integrate old and new findings can be assembled (e.g. Zeberg and Pääbo 2020, 2021). Furthermore, large open repositories provide another advantage: when data on specific populations to be studied are already available and can be reused in accordance with informed consent, time and money can be saved, plus donors no longer need to be sampled further.

Looking at the dark side of the moon, the first problem to be considered concerns the possible ethical issues. Dealing with human subjects, anthropological surveys may face risks regarding the leakage of sensitive individual information to unauthorised parties. e.g. when genomic data are combined with other sources of information on individual characteristics. In fact, it has been shown that participants in public sequencing projects can be identified with high probability using free, publicly accessible Internet resources such as online genealogy databases (e.g. see Gymrek et al. 2013; Erlich et al. 2018). Members of small and socially identifiable communities, which are often of particular significance for anthropological research, may undergo even greater risks of privacy violation (McGregor 2007; Tsosie 2007; Clayton et al. 2018) and discrimination based on disease predispositon (Suther and Kiros 2009; Lemke 2013). Being aware of its potential conflicts with the safeguarding of legitimate interests of participants, researchers could embrace the ideal of openness in a more responsible and sustainable way (see Byrd et al. 2020 and related citation therein), starting from the drafting of informed consent, particularly the section regarding the research purposes (Rao 2016). Several authors have pointed out that an improper use of genetic data could expose donors to risks of stigmatisation in the employment environment, in health and life insurance and in education (see Haeusermann et al. 2018; Chapman et al. 2020; Joly et al. 2020).

Improper data sharing practises can also lead to legal problems. Probably, the controversy between the Havasupai of Arizona and the Arizona State University (ASU) has been one of the most discussed (Sterling 2011; Garrison 2013; Van Assche et al. 2013). In this case, an investigation into the genetic causes of high rates of type II diabetes was followed by other studies regarding alcoholism, inbreeding and the historical origin of this community settled in the Grand Canyon. Not seeking consent to accomplish this secondary purpose, the ASU researchers were sued by community members. In 2010, the Arizona Court of Appeals obliged the former to economically compensate the latter for the moral damage caused by the researchers having used genetic data for purposes that had not been specified in the informed consent presented at the time of sampling (Harmon 2010; Mello and Wolf 2010). Other cases of allegations of mere exploitation of indigenous peoples have involved researchers from major scientific companies, such as the Human Genome Diversity Project (HGDP) and the Genographic Project (Lock 2001; TallBear 2007). As a result, particularly in the African context, governments, local institutions and community leaders have reduced their willingness to support the "historic outflow of samples and data from the continent" (Wright 2014:1). At the same time, they have pushed genomic consortia, such as the International HapMap Project and the 1000 Genomes Project, to pay more attention to their relationship with communities and data sharing practises in three ways. Firstly, by setting up ad hoc groups of experts with the task of considering ethical, social and legal issues from the very beginning of the project (Merriman and Molina 2015). Secondly, by defining an approach in which data sharing was conditioned by the legitimacy of the researchers' purposes. Thirdly, by managing data-sharing practises to reduce misunderstanding

and misuse of data, specifying the accountability of data producers, users, and funding agencies in guidelines adapted from the Fort Lauderdale Agreement (Wellcome Trust 2003).

Sharing a complete dataset with a first publication can be problematic especially for groups with limited resources, and therefore with a lower data production capacity, for which it may be important to fully exploit the information contained in the data itself over a longer period of time. Sharing can also be considered disadvantageous as it can be detrimental in academic and grant competition with other groups who may be favoured by not having to collect the data themselves. Both aspects regard "the fear that someone else publishes with my data before I can" (Fecher et al. 2015, p. 16) and the need to control the use of "what's mine is mine". This may lead to data withholding (Blumenthal et al. 2006; Defazio et al. 2020) or requests for co-authorship from data owners (Tenopir et al. 2011; Capocasa et al. 2016).

Aims of the study

In the last decade, the production and resolution of human genomic data has constantly increased, mainly due to technological advancements and decreasing costs of DNA genotyping and sequencing. The downside of this process is a huge increase in the complexity of the data and associated metadata. With NGS, data are no longer stored as individual DNA strings, as was the case with mtDNA data in the 90s, but as millions of raw reads per sequence which need to be accompanied by information concerning filtering and quality control procedures to be carefully reused. As a consequence, simply depositing the raw data into a database or sharing them on a website is not enough. To maximise all the benefits provided by Open Data, they should be opened "intelligently". The concept of "intelligent data openness" (IDO) was formalised by Geoffrey Boulton and the Royal Society working group in 2012 in the Science as an Open Enterprise report. It is based on four fundamental criteria: (i) findability; datasets have to be easily found, (ii) accessibility; datasets must be readily accessed and queried, (iii) useability; datasets have to be in such a form that they can be easily reused and (iv) assessability; the information (metadata) accompanying the data should allow the evaluation of their reliability and provenance.

Unfortunately, despite its usefulness for a deeper understanding of the effectiveness of data sharing practises, no study has so far studied IDO with an empirical approach. Therefore, we thought it would be useful to undertake a pilot study aimed at complementing the simple survey on the availability of data related to peer-reviewed publications in genomic anthropology and human genomics with an analysis of their availability, accessibility, useability and evaluability.

Methods

We focused on three types of human genomic data: Single Nucleotide Polymorphisms (SNPs) genotypes deriving from the processing of microarrays (with a minimum threshold of 300K markers), gene expression data produced through microarrays and DNA sequences produced by NGS technology (limited to whole genome and whole exome data).

Basic definitions

To disentangle the shared/withheld dichotomy we: (i) searched for specific indications of data sharing (e.g. data sharing statement) accession number or link to public archives or web sites; (iii) searched for in the databases of the National Center for Biotechnology Information (NCBI) using the paper titles as a keyword whenever neither accession number nor any other link was provided. Through this procedure, we defined three levels of data sharing:

1) immediate: the data were available and downloadable with no particular pre-conditions;
2) controlled access: the data were stored with an accession number in a public online database, but their downloading depended on the positive outcome of an ad-hoc request procedure;

3) upon request: datasets which could be obtained only by request (not predefined) to the author/s of the relevant paper.

When data were not shared at all or only in part (e.g. only data from a subsample of genetic markers or individuals were available), the datasets were considered to have been withheld.

To assess compliance with the four IDO criteria, we proceeded as follows:

- findability: verifying the presence in the article or on the relevant web page of the journal of references (e.g. database links, accession numbers) through which the dataset could be found;
- accessibility: downloading the dataset file/s and checking their actual content;
- useability: checking if the data were in a standard or commonly used format (e.g. VCF, Plink, Geno);
- assessability**:** searching for info which enables the unambiguous interpretation of the results and facilitates their reproduction in body text and supplementary materials of the paper or in the database page where the data were stored. Such items (e.g. enrollment criteria, biomaterials used to extract DNA and data transformation processes) were derived from the MIAME and the MINSEQE (FGED) guidelines (Brazma et al. 2001; Brazma 2009).

A detailed description is provided in Supplementary Material, Table S1.

Data collection

We searched for papers using the WoS database because, differently from others (e.g. Pubmed, Scopus), it makes it possible to filter papers based on research areas. To find papers dealing with genomic anthropology, we initially selected those labelled as "anthropology" which had been published between 2015 and 2017. Next, we did a one-by-one inspection of the papers that had previously been filtered according to the following sequential steps:

step 1: made use of novel human data (studies non pertinent to human species, reviews, meta-analyses, and studies using exclusively already published data were excluded);

step 2: present autosomal genomic data produced by SNP microarrays (SNPs), gene expression microarrays (GE) and next generation sequencing (NGS);

step 3: analyse at least 300K SNPs or whole genome or whole exome sequencing data (no specific limit was considered for gene expression microarrays)

Once the final dataset was assembled, each paper was scrutinised and data were collected by two independent experienced researchers. When conclusions were discordant, a consensus was reached with the help of a third researcher who independently analysed the papers.

We also analysed a "human genomics" dataset (a total of 450 papers) using the previously defined criteria. It was obtained by selecting randomly 50 papers for data type (SNPs, GE and NGS) and year (2015-2017) from a total 16,194 items retrieved using the Pubmed database (see Supplementary Material, Tab. S2 for details).

Results

Data sharing in genomic anthropology

Using the WoS database we retrieved a total of 3177 papers, published between 2015 and 2017. After steps 1 and 2, the size of the dataset was drastically reduced to 57 papers. After step 3, only the SNP dataset reached a sufficient number (37 items) for detailed analysis, while no GE and only 11 NGS datasets survived the selection Supplementary Material, Table S2.

Fifteen SNP datasets (40.5%) were immediately shared, 1.5 times those withheld (10 items, 27.0%). Five datasets (13.5%) were under controlled access; they were all deposited in the database of Genotypes and Phenotypes (dbGaP), a primary database (i.e containing experimentally derived data submitted by authors). The remaining seven datasets (18.9%) were potentially available upon request to the paper's author/s (Fig. 4).

To perform the IDO evaluation, we focused on the 15 immediately shared datasets only. They were all easily findable and downloadable, available either through institutional or private

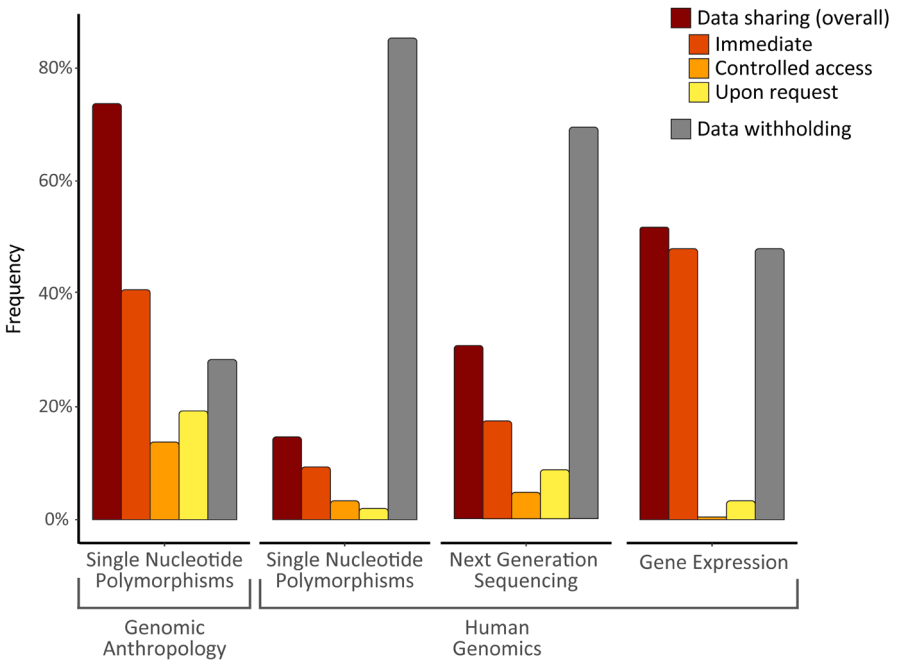

***Fig. 4** - Data sharing and withholding rates in genomic anthropology and human genomics papers (cumulative 2015 - 2017 data; see Supplementary Material, Tab. S3 for absolute and percentage values).*

web sites (eight out of 15; see Supplementary Material, Tab. S2 for details) or a primary database (e.g. Gene Expression Omnibus, European Nucleotide Archive; seven out of 15). In all cases, properly working links and/or accession number resources were found in the papers body text, mainly under the "Data sharing statement" section, or directly in the journal's article web page. Moreover, all the data were available in commonly used standard formats. On the contrary, the assessability criterion was fulfilled only by three out of 15 datasets (20.0%). The most critical aspect concerned the lack of information on the type of biomaterial collected (e.g. blood or saliva) and its transformation processes, which regarded nine datasets (60.0%).

Data sharing in human genomics

Overall, 24.4% (110) of the datasets was found to be immediately shared, whereas 67.6%, (304) was withheld (see Supplementary Table S3 for details). Both controlled access and upon request modalities concerned a small fraction of datasets, 3.8% (17), and 4.2% (19), respectively. Looking at the data sharing rates in a three-year temporal frame, we observed a slight decrease in the immediate data availability (from 30% to 20.7%), as well as the controlled data access (from 5.3% to 2.7%) (see Fig. 5 and Supplementary Material, Tab. S3). In both cases, the decrease was more pronounced from 2015 to 2016. On the contrary, upon request and withheld datasets showed an increasing trend, from 64.0% to 70.7% and from 0.7% to 6.0%, respectively. Turning our attention to data sharing rates among the three types of genomic information considered, we observed a striking difference between GE, on the one side, and both NGS and SNP data, on the other. In fact, while nearly half of former datasets (48.0%) were immediately shared, for the latter, the rates dropped consistently, reaching values of 16.7% (NGS) and

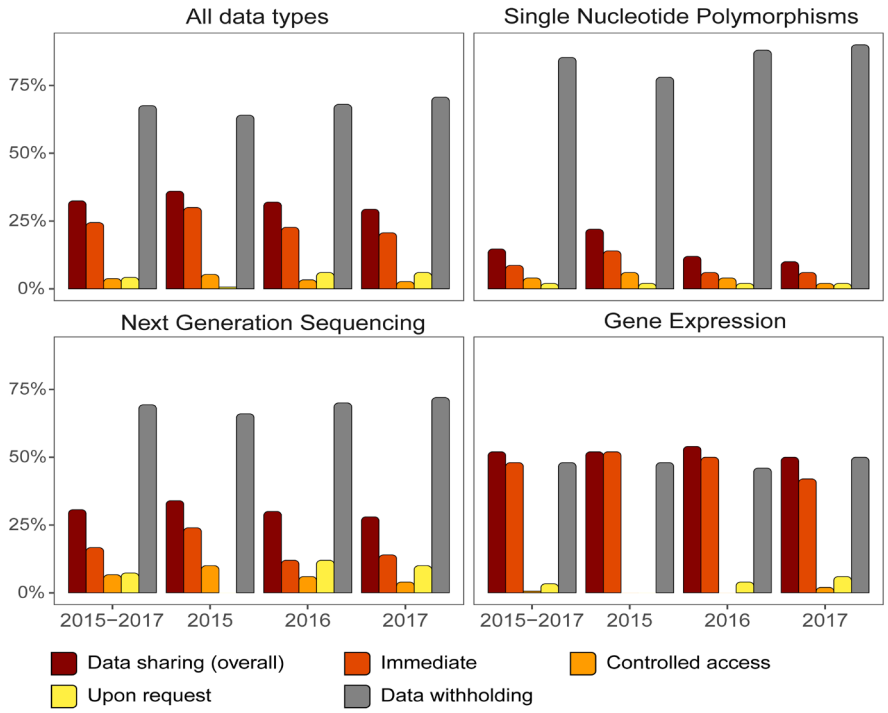

***Fig. 5** - Data sharing rates in human genomics papers by data type and year of publication of papers (see Supplementary Material, Tab. S3 for absolute and percentage values).*

8.7% (SNPs). Consistent with the overall trend, we observed a decrease in immediate sharing for each data type between 2015 and 2017, ranging from eight to ten percentage points.

Concerning the IDO criteria, online primary databases were the vastly preferred way to make data accessible for all three types of immediately shared datasets, with an overall rate of 94.5%. We found only five datasets deposited in other repositories (two for GE and NGS and one for SNP data), while just one GE dataset was made available through the supplementary material accompanying the publication. For all the above-mentioned datasets, the findability criterion was completely fulfilled either by providing accession numbers, or a link to repositories. All the datasets used common standard formats. Similarly to genomic anthropology datasets, the most problematic aspect of IDO was the assessability. Only 39.1% (43 out of 110) of the immediately shared human genomics datasets were fully assessable, with GE and NGS showing rates that were similar (43.1% and 40.0%, respectively) or much higher than SNPs (15.4%). Again, the most failed criterion concerned info on the biomaterial/s used and DNA extraction (for 41 out of 110 datasets, 37.3%), but with noticeable differences across data types (27.8%, 48.8% and 69.2% for GE, NGS and SNP datasets, respectively; see Supplementary Material, Tab. S2).

Discussion

Comparing genomic anthropology and human genomics

The most striking finding of this study is that 73.0% of the datasets produced in genomic anthropological research were found to be

shared, whereas their percentage dropped to 32.4% in human genomics. Similarly, datasets are shared immediately in genomic anthropology almost twice as much as in human genomics. The gap becomes even more striking when only the SNP data are considered, reaching nearly a five fold difference (40.5% vs 8.7%; see Fig. 4). The discrepancy between the two research fields increases even further if we also consider the data that could be potentially shared. Genomic anthropologists seem to be more willing to make their data available than human genomicists (35.1% vs 6.0%) when asked for (controlled and upon request access).

Understanding the reasons behind these patterns is difficult without ad hoc surveys, which could be based on the administration of questionnaires to the studies' authors. Nonetheless, it seems reasonable to think that privacy issues may have played a significant role in the pattern observed. Most of the human genomics papers we scrutinised dealt with biomedical research, where the tension between privacy concerns and potential health benefits may reduce the propensity to share (Knoppers and Thorogood 2017; Bonomi et al. 2020). More in particular, sharing genomic data may pose risks of association between identity and disease susceptibility at individual (Hirschhorn and Daly 2005; McCarthy et al. 2008; Erlich et al. 2018; Von Thenen et al. 2019) and even community level (McGregor 2007; Bonomi et al. 2020). In both cases, the data disclosure can mean that donors risk a loss of confidentiality, stigmatization and discrimination in their social and professional environment (Arias et al. 2015). However, we speculate that privacy issues might not be the only reason. In fact, we believe that methodological aspects should also be taken into account. On the one hand, human genomics SNP data are often used for genome-wide association studies in order to identify genomic regions associated with a specific trait. This implies that the genotypic data are only used in preliminary analyses that are then followed by more in depth genomic investigation (e.g. by sequencing or imputation), which is expected to produce the most significant results. On the other hand, SNPs have a more central role in genomic anthropology since they provide the experimental data needed on which statistical approaches can be developed in order to answer evolutionary questions. Such differences between the two research fields might have an influence on the propensity to share since the lesser importance of SNPs in their experimental design could make researchers in human genomics less aware of the usefulness of making data available to others.

Widening our perspective

With a view to broadening our study of data sharing, we collected results of previous investigations on this subject and grouped the results according to the methodology used (see Fig. 6 and references given therein). On the whole, the range of values is extremely wide, going from 97.6% for Human Palaeogenetics to 0.8% in Addiction research.

Regarding the studies based on the scrutiny of scientific papers, there seem to be three aspects worthy of attention. Firstly, human genetics and genomics show two distinct patterns: the former occupies the first position among the search fields, while the latter is positioned in the medium and low part of the range. A possible explanation could lie in the different quantities of information contained in the data. Usually, it is substantially lower in genetics, which could lead to a quicker exploitation and, therefore, to a more timely release of the data. Conversely, the considerable information contained in genomic data can be split or reused to produce multiple studies, which could slow down its full use and delay its release. The time and effort it takes to share data may represent another, albeit less important, reason. In fact, the size of genomic files largely exceeds genetic ones (from several hundreds of megabases to tens of gigabases to few megabases), making it easier to share the latter data type through journals' supplementary information. Furthermore, the procedures regarding how to submit data to primary databases are simpler for genetic than for genomic data. For example, depositing mtDNA sequences is quite

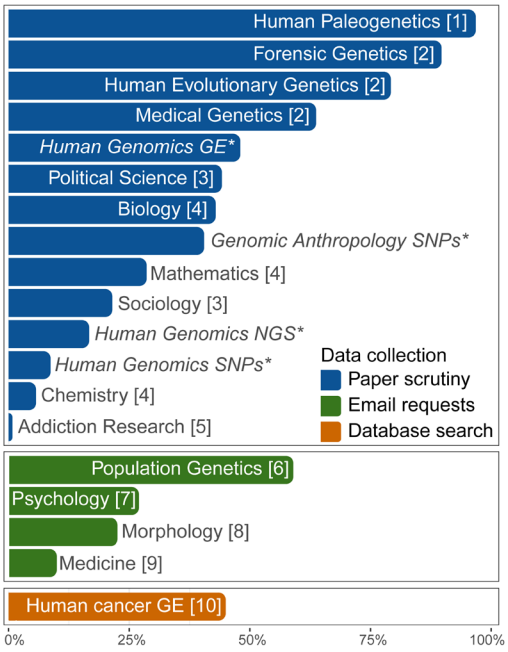

***Fig. 6** - Data sharing rates in different research areas, grouped according to the collection method. References: *this study; [1] Anagnostou et al. 2015; [2] Milia et al. 2012; [3] Zenk-Möltgen et al. 2018; [4] Womack 2015; [5] Gorman 2020; [6] Leberg and Neigel 1999; [7] Wicherts et al. 2006; [8] Vines et al. 2014; [9] Savage and Vickers 2009; [10] Piwowar et al. 2011.*

uncomplicated and requires only a little complementary information, whereas for genomic data, researchers have to prepare several files with metadata and other information in addition to the raw data.

Secondly, data are shared much less in genomic anthropology than in its closely related research fields, that is human evolutionary genetics and human paleogenomics (40.5% vs 79.2% and 97.6%, respectively). Privacy issues may contribute to this difference. In fact, the risk of personal identification is much higher for genomic rather than genetic (e.g. mitochondrial DNA sequences) and, obviously, paleogenetic data. The fact that GE data, which do not pose risks of personal identification, are shared to a substantially larger extent than SNPs and NGS data seems to support the above-mentioned interpretation.

Thirdly, to broaden our view, we have also taken into consideration more distant fields of research (Fig. 6). Obviously the results of these comparisons must be interpreted with caution as each research field has its own idea of what the term data means, a different culture of sharing and variable constraints and incentives to make the data available to the public. (Tenopir et al. 2011; Tedersoo et al. 2021). However, the evidence that sharing rates are reported to be higher in disciplines (e.g. political science or sociology) where data should be less encoded and reproducible than genomic anthropology and human genomics seems to challenge the widespread, but perhaps simplistic, idea of genomics as a forerunner for the Open Data movement.

Data sharing rates obtained with different methods are in line with the findings we have previously described. Studies using the "email to authors" approach found the highest values for genetic data, while that for GE, which was the only one based on text mining, found a similar data availability rate to ours for the same type of data (45.0% and 48.0%, respectively).

Intelligent data openness

As a final step in our study, we assessed compliance with the IDO criteria. This was limited to the analysis of the "immediately available datasets" since these, being completely accessible, were the only ones for which it was possible to evaluate their availability, useability and evaluability.

Overall, we found no substantial difference in IDO between genomic anthropology and human genomics. In both research fields, the largely preferred way of making data accessible was through online repositories (institutional or primary databases) or private web pages. We could only find one GE dataset out of a total of 110 items in human genomics that was shared via supplemental material. However, it should be noted that genomic anthropologists make less use of primary online databases than human genomics (46.7% vs 94.5%, respectively see

Supplementary Table S2) and, consequently, more frequent use of personal or institutional web pages. The use of online databases is regarded as the best way to store data and protect against accidental data loss or corruption since they guarantee long term archiving preservation (Uzwyshyn 2016), and publicly funded primary databases (e.g. those maintained by NCBI) are probably the best choice (Tellam et al. 2015). Furthermore, online databases improve findability, by providing each dataset with a unique identifier that is searchable through the web, and useability by requiring that the data be submitted in commonly used and standardised data formats (Sim 2020; Wilson 2021). The only expectation of the IDO that we found to be still lacking in both fields of research was "assessability". Only 20% of the genomic anthropology datasets were accompanied by sufficient information to assess the quality of the data. In human genomics, the rate was slightly lower for SNPs (15.4%), but double for GE and NGS (43.1% and 40.0%, respectively). The most frequent drawback concerned the biomaterials and, to a lesser extent, the protocols applied for DNA extraction. Knowledge regarding the biological tissue used to extract DNA is not a trivial issue. Recent studies have shown that the quality of data from high-throughput sequencing can be largely influenced by the biological sample used to extract the DNA (Bruinsma et al. 2018; Yao et al. 2020). It is worth noting that in the field of genomic anthropology, we have found greater use of repositories with controlled access, such as dbGaP and the European Genome-phenome Archive (EGA) (see Supplementary Material, Tab. S2) This may derive from the fact that genomic anthropologists are more careful regarding the need to regulate the reuse of data by third parties (Mailman et al. 2007; Freeberg et al. 2021). Data access commissions check whether applicants' requests for reuse comply with the original research purposes (Alpaslan-Roodenberg et al. 2021). Several attempts to go a step further to reduce the risks of misuse, which meet important needs for anthropologists, have recently been conducted in genomics research. We are referring to the so-called indigenous genomic databases (D'Angelo et al. 2020) and their contribution to the development of controlled data access systems based on community approval for data reuse (e.g. the National Center For Indigenous Genomics in Australia, the Genomics Aotearoa project in New Zealand and the Silent Genomes project in Canada; see Easteal 2018; Robertson et al. 2018; Caron et al. 2020). These participative repositories aim "to enable researchers to position genomic data and science within culturally appropriate overarching research and oversight frameworks that maximise benefits and minimise risks for participating communities" (Caron et al. 2020:4). We argue that this could be a sustainable approach also for data management in genomic anthropological studies involving socially identifiable groups.

Acknowledgements

This work was funded by the University of Rome "La Sapienza" (Italy), through the project "Setting up a network for Responsible Research and Innovation - RRI-Net" (PI116154C845CA8D) and received support from the Istituto Italiano di Antropologia (project "Open Access and Open Data"). We thank Valentina Dominici and Nicola Milia for their help in collecting the data.

References

1000 Genomes Project Consortium (2015) A global reference for human genetic variation. Nature 526:68-74. https://doi.org/10.1038/nature15393

Alpaslan-Roodenberg S, Anthony D, Babiker H, et al. (2021) Ethics of DNA research on human remains: five globally applicable guidelines. Nature 599:41-46. https://doi.org/10.1038/s41586-021-04008-x

Amann RI, Baichoo S, Blencowe BJ, et al (2019) Toward unrestricted use of public genomic data. Science 363:350-352. https://doi.org/10.1126/science.aaw1280

Anagnostou P, Capocasa M, Milia N, et al (2015) When data sharing gets close to 100%: what human paleogenetics can teach the open science movement. PLoS One 10:e0121409. https://doi.org/10.1371/journal.pone.0121409

Anagnostou P, Destro Bisol G (2011) Anthro-Digitdata, an online resource for anthropological data sharing. J Anthropol Sci 89:221-222.

Apicella CL, Silk JB (2019) The evolution of human cooperation. Curr Biol 29:R447-R450. https://doi.org/10.1016/j.cub.2019.03.036

Arias JJ, Pham-Kanter G, Gonzalez R, et al. (2015) Trust, vulnerable populations, and genetic data sharing. J Law Biosci 2:747-753. https://doi.org/10.1093/jlb/lsv044

Bierer BE, Crosas M, Pierce HH (2017) Data authorship as an incentive to data sharing. N Engl J Med 376:1684-1687. https://doi.org/10.1056/NEJMsb1616595

Birney E, Hudson TJ, Green ED, et al (2009). Prepublication data sharing. Nature 461:168-170. https://doi.org/10.1038/461168a

Blumenthal D, Campbell EG, Gokhale M, et al. (2006). Data withholding in genetics and other life sciences: prevalences and predictors. Acad Med 81:137–145. https://doi.org/10.1097/00001888-200602000-00006

Bonomi L, Huang Y, Ohno-Machado L (2020) Privacy challenges and research opportunities for genomic data sharing. Nat Genet 52:646-654. https://doi.org/10.1038/s41588-020-0651-0

Boulton G, Campbell P, Collins B, et al. (2012) Science as an open enterprise, The Royal Society, London.

Brazma A (2009) Minimum information about a microarray experiment (MIAME)–successes, failures, challenges. Sci World J 9:420-423. https://doi.org/10.1100/tsw.2009.57

Brazma A, Hingamp P, Quackenbush J, et al. (2001) Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. Nat Genet 29:365–371. https://doi.org/10.1038/ng1201-365

Bruinsma FJ, Joo JE, Wong EM, et al. (2018). The utility of DNA extracted from saliva for genome-wide molecular research platforms. BMC Res Notes 11:8. https://doi.org/10.1186/s13104-017-3110-y

Byrd JB, Greene AC, Prasad DV, et al (2020) Responsible, practical genomic data sharing that accelerates research. Nat Rev Genet 21:615-629. https://doi.org/10.1038/s41576-020-0257-5

Capocasa M, Anagnostou P, D'Abramo F, et al. (2016) Samples and data accessibility in research biobanks: an explorative survey. PeerJ 4:e1613. https://doi.org/10.7717/peerj.1613

Caron NR, Chongo M, Hudson M, et al. (2020) Indigenous genomic databases: pragmatic considerations and cultural contexts. Front Public Health 8:111. https://doi.org/10.3389/fpubh.2020.00111

Chapman CR, Mehta KS, Parent B, et al. (2020) Genetic discrimination: emerging ethical challenges in the context of advancing technology. J Law Biosci 7:lsz016. https://doi.org/10.1093/jlb/lsz016

Chubin DE (1985) Open science and closed science: tradeoffs in a democracy. Sci Technol Hum Values 10:73–80. https://doi.org/10.1177/016224398501000211

Clayton EW, Halverson CM, Sathe NA, et al (2018). A systematic literature review of individuals' perspectives on privacy and genetic information in the United States. PLoS One 13: e0204417. https://doi.org/10.1371/journal.pone.0204417

D'Angelo CS, Hermes A, McMaster CR, et al. (2020) Barriers and considerations for diagnosing rare diseases in Indigenous populations. Front Pediatr 8: 579924. https://doi.org/10.3389/fped.2020.579924

De Silva PUK, Vance CK (2017) Scientific scholarly communication. The changing landscape, Springer, Cham.

Defazio D, Kolympiris C, Perkmann M, et al. (2020) Busy academics share less: the impact of professional and family roles on academic withholding behaviour. Stud High Educ [e-pub ahead of print]. https://doi.org/10.1080/03075079.2020.1793931

Destro Bisol G, Jobling MA, Rocha J, et al. (2010) Molecular anthropology in the genomic era. J Anthropol Sci 88:93-112.

Dunklee B (2003) Sequencing the trellis: the production of race in the new human genomics.

Undergraduate honors thesis, Brown University, Providence.
Easteal S (2018) NCIG—National Centre for Indigenous Genomics. Impact 2018:72-74. https://doi.org/10.21820/23987073.2018.10.72
Erlich Y, Shor T, Pe'er I, et al. (2018) Identity inference of genomic data using long-range familial searches. Science 362:690-694. https://doi.org/10.1126/science.aau4832
Fecher B, Friesike S (2013) Open Science: One Term, Five Schools of Thought. In: Bartling S, Friesike S (eds) Opening Science, Springer, Cham, p. 17-47.
Fecher B, Friesike S, Hebing M (2015) What drives academic data sharing? PLoS One 10:e0118053. https://doi.org/10.1371/journal.pone.0118053
Fischer BA, Zigmond MJ (2010) The essential nature of sharing in science. Sci Eng Ethics 16:783–799. https://doi.org/10.1007/s11948-010-9239-x
Freeberg MA, Fromont LA, D'Altri T, et al. (2021) The European Genome-phenome Archive in 2021. Nucleic Acids Res gkab1059. https://doi.org/10.1093/nar/gkab1059
Garrison N (2013) Genomic justice for Native Americans: impact of the Havasupai case on genetic research. Sci Technol Hum Values 38:201-223. https://doi.org/10.1177/0162243912470009
Garrison N, Hudson M, Ballantyne LL, et al. (2019) Genomic research through an indigenous lens: understanding the expectations. Annu Rev Genom Hum Genet 20:495-517. https://doi.org/10.1146/annurev-genom-083118-015434
Gibbs RA (2020) The Human Genome Project changed everything. Nat Rev Genet 21:575–576. https://doi.org/10.1038/s41576-020-0275-3
Goisauf M, Akyüz K, Martin GM (2020) Moving back to the future of big data-driven research: reflecting on the social in genomics. Humanit Soc Sci Commun 7:55. https://doi.org/10.1057/s41599-020-00544-5
Gorman DM (2020) Availability of research data in high-impact addiction journals with data sharing policies. Sci Eng Ethics 26:1625-1632. https://doi.org/10.1007/s11948-020-00203-7
Gura T (2013) Citizen science: amateur experts. Nature 496:259–261. https://doi.org/10.1038/nj7444-259a
Gymrek M, McGuire AL, Golan D, et al. (2013) Identifying personal genomes by surname inference. Science 339:321-324. https://doi.org/10.1126/science.1229566
Haeusermann T, Fadda M, Blasimme A, et al. (2018) Genes wide open: data sharing and the social gradient of genomic privacy. AJOB Empir Bioeth 9:207-221. https://doi.org/10.1080/23294515.2018.1550123
Harmon A (2010) Indian tribe wins fight to limit research of its DNA, New York Times, April 21. http://www.nytimes.com/2010/04/22/us/22dna.html?pagewanted=all&_r=0
Harry D, Howard S, Shelton BL (2000) Indigenous peoples, genes and genetics: what indigenous people should know about biocolonialism, Indigenous People Council on Biocolonialism, Wadsworth.
Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nature Rev Genet 6:95-108. https://doi.org/10.1038/nrg1521
Hood L, Rowen L (2013) The Human Genome Project: big science transforms biology and medicine. Genome Med 5:79. https://doi.org/10.1186/gm483
Hrynaszkiewicz I, Simons N, Hussain A, et al. (2020) Developing a research data policy framework for all journals and publishers. Data Sci J 19:5. http://doi.org/10.5334/dsj-2020-005
International HapMap 3 Consortium (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467:52-58. https://doi.org/10.1038/nature09298
Jobling M, Hurles M, Tyler Smith C (2004) Human evolutionary genetics: origins, peoples and disease, Garland Science Publishing, New York.
Joly Y, Dalpé G, Dupras C, et al. (2020) Establishing the International Genetic Discrimination Observatory. Nat Genet 52:466-468. https://doi.org/10.1038/s41588-020-0606-5

Knoppers BM, Thorogood AM (2017) Ethics and big data in health. Curr Opin Syst Biol 4:53-57. https://doi.org/10.1016/j.coisb.2017.07.001

Kowal E, Llamas B (2019) Race in a genome: long read sequencing, ethnicity-specific reference genomes and the shifting horizon of race. J Anthropol Sci 97:91-106. https://doi.org/0.4436/jass.97004

Leberg PL, Neigel JE (1999) Enhancing the retrievability of population genetic survey data? An assessment of animal mitochondrial DNA studies. Evolution 53:1961-1965. https://doi.org/10.1111/j.1558-5646.1999.tb04576.x

Lee SSJ, Mountain J, Koenig BA (2001) The meanings of race in the new genomics: implications for health disparities research. Yale J Health Pol'y L & Ethics 1:3.

Lemke T (2013) Perspectives on genetic discrimination, Routledge, New York.

Lock M (2001) The alienation of body tissue and the biopolitics of immortalized cell lines. Body Soc 7:63–91. https://doi.org/10.1177/1357034X0100700204

Low SM, Merry SE (2010) Engaged anthropology: diversity and dilemmas: an introduction to supplement 2. Curr Anthropol 51:S203-S226. https://doi.org/10.1086/653837

Mailman MD, Feolo M, Jin Y, et al. (2007) The NCBI dbGaP database of genotypes and phenotypes. Nat Genet 39:1181-1186. https://doi.org/10.1038/ng1007-1181

McCarthy MI, Abecasis GR, Cardon LR, et al. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9:356-369. https://doi.org/10.1038/nrg2344

McGregor JL (2007) Population genomics and research ethics with socially identifiable groups. J Law Med Ethics 35:356-370. https://doi.org/10.1111/j.1748-720X.2007.00160.x

McInnes R (2011) 2010 Presidential address. Culture: the silent language geneticists must learn-genetic research with Indigenous populations. Am J Hum Genet 88:254-261. https://doi.org/10.1016/j.ajhg.2011.02.014

Mello MM, Wolf LE (2010) The Havasupai Indian tribe case. Lessons for research involving stored biologic samples. N Engl J Med 363:204–207. https://doi.org/10.1056/NEJMp1005203

Merriman B, Molina SJ (2015) Variable conceptions of population in community resource genetic projects: a challenge for governance. New Genet Soc 34:294-318. https://doi.org/10.1080/14636778.2015.1060115

Merton RK (1942) Science and technology in a democratic order. J Legal Pol Soc 1:115-126.

Milia N, Congiu A, Anagnostou P, et al. (2012). Mine, yours, ours? Sharing data on human genetic variation. PLoS One 7:e37552. https://doi.org/10.1371/journal.pone.0037552

OECD (2007) Best Practices for Ensuring Scientific Integrity and Preventing Misconduct, Organisation for Economic Co-operation and Development, Paris.

Pagani L, Destro Bisol G (2021) Next questions in molecular anthropology. J Anthropol Sci [epub ahead of print]. https://doi.org/10.4436/jass99013

Piwowar HA (2011) Who shares? Who doesn't? Factors associated with openly archiving raw research data. PLoS One 6:e18657. https://doi.org/10.1371/journal.pone.0018657

Quinn H (2009) What is science? Physics Today 62:8-9. https://doi.org/10.1063/1.3177240

Ramsay M, Brunner HG, Djikeng A (2019) Leveraging genomic diversity to promote human and animal health. Commun Biol 2:463. https://doi.org/10.1038/s42003-019-0708-8

Rao R (2016) Informed consent, body property, and self-sovereignty. J Law Med Ethics 44:437-444. https://doi.org/10.1177/1073110516667940

Rehm H (2017) A new era in the interpretation of human genomic variation. Genet Med 19:1092–1095. https://doi.org/10.1038/gim.2017.90

Robertson SP, Hindmarsh JH, Berry S, et al. (2018) Genomic medicine must reduce, not compound, health inequities: the case for hauora-enhancing genomic resources for New Zealand. N Z Med J 131:81-89.

Sansone SA, Cruse P, Thorley M (2018) High-quality science requires high-quality open data infrastructure. Sci Data 5:180027. https://doi.org/10.1038/sdata.2018.27

Savage CJ, Vickers AJ (2009) Empirical study of data sharing by authors publishing in PLoS journals. PLoS One 4:e7078. https://doi.org/10.1371/journal.pone.0007078

Schensul SL, Schensul JJ, Singer M, et al. (2015) Participatory Methods and Community-Based Collaborations. In: Bernard HR, Gravlee CC (eds) Handbook of Methods in Cultural Anthropology, 2nd Edition, Rowman & Littlefield, Lanham, p. 185-212.

Sharp RR, Foster MW (2002) Community involvement in the ethical review of genetic research: lessons from American Indian and Alaska native populations. Env Health Perspect 110:145-148. https://doi.org/10.1289/ehp.02110s2145

Sim I (2020) Data Sharing and Reuse. In: Piantadosi S, Meinert C (eds) Principles and Practice of Clinical Trials, Springer, Cham. https://doi.org/10.1007/978-3-319-52677-5_190-1

Sterling RL (2011). Genetic research among the Havasupai: a cautionary tale. AMA J Ethics 13:113-117. https://doi.org/10.1001/virtualmentor.2011.13.2.hlaw1-1102

Suther S, Kiros GE (2009) Barriers to the use of genetic testing: a study of racial and ethnic disparities. Genet Med 11:655-662. https://doi.org/10.1097/GIM.0b013e3181ab22aa

TallBear K (2007) Narratives of race and indigeneity in the Genographic Project. J Law Med Ethics 35:412-424. https://doi.org/10.1111/j.1748-720X.2007.00164.x

Tedersoo L, Küngas R, Oras E, et al. (2021) Data sharing practices and data availability upon request differ across scientific disciplines. Sci Data 8:192. https://doi.org/10.1038/s41597-021-00981-0

Tellam RL, Rushton P, Schuerman P, et al. (2015) The primary reasons behind data sharing, its wider benefits and how to cope with the realities of commercial data. BMC Genom 16:626 https://doi.org/10.1186/s12864-015-1789-5

Tenopir C, Allard S, Douglass K, et al. (2011) Data sharing by scientists: practices and perceptions. PLoS One 6:e21101. https://doi.org/10.1371/journal.pone.0021101

Trotter RI, Schensul JJ, Kostick K (2014) Theories and Methods in Applied Anthropology. In: Bernard HR, Gravlee CC (eds) Handbook of Methods in Cultural Anthropology, 2nd Edition, Rowman & Littlefield, Lanham, p. 661-693.

Tsosie R (2007) Cultural challenges to biotechnology: Native American genetic resources and concept of cultural harm. J Law Med Ethics 35:396-411. https://doi.org/10.1111/j.1748-720X.2007.00163.x

Turner TR, Mulligan CJ (2019) Data sharing in biological anthropology: guiding principles and best practices. Am J Phys Anthropol 170:3-4. https://doi.org/10.1002/ajpa.23909

Uzwyshyn R (2016) Research data repositories: the what, when, why and how. Comput Libr 36:18-21.

Van Assche K, Gutwirth S, Sterckx S (2013) Protecting dignitary interests of biobank research participants: lessons from Havasupai Tribe v Arizona Board of Regents. Law Innov Technol 5:54-84. https://doi.org/10.5235/17579961.5.1.54

Van Willigen J (2002) Applied anthropology: An introduction, Greenwood Publishing Group, Westport.

Vines TH, Albert AY, Andrew RL, et al. (2014) The availability of research data declines rapidly with article age. Curr Biol 24:94-97. https://doi.org/10.1016/j.cub.2013.11.014

Von Thenen N, Ayday E, Cicek AE (2019) Re-identification of individuals in genomic data-sharing beacons via allele inference. Bioinformatics 35:365-371. https://doi.org/10.1093/bioinformatics/bty643

Walters WH (2020) Data journals: incentivizing data access and documentation within the scholarly communication system. Insights 33:18. http://doi.org/10.1629/uksg.510

Wellcome Trust (2003). Sharing data from large-scale biological research projects: a system of tripartite responsibility. Report of a meeting organized by the Wellcome Trust and held on 14–15 January 2003 at Fort Lauderdale, USA, Wellcome Trust, London.

Wicherts JM, Borsboom D, Kats J, et al. (2006) The poor availability of psychological research data for reanalysis. Am Psychol 61:726-728. https://doi.org/10.1037/0003-066X.61.7.726

Wilson EO (1998) The unity of knowledge, Knopf, New York.

Wilson SL, Way GP, Bittremieux W, et al. (2021) Sharing biological data: why, when, and how. FEBS Lett 595:847-863. https://doi.org/10.1002/1873-3468.14067

Womack RP (2015) Research data in core journals in biology, chemistry, mathematics, and physics. PLoS One 10:e0143460. https://doi.org/10.1371/journal.pone.0143460

Wright GE, Adeyemo AA, Tiffin N (2014) Informed consent and ethical re-use of African genomic data. Hum Genomics 8:18. https://doi.org/10.1186/s40246-014-0018-7

Yao RA, Akinrinade O, Chaix M, et al. (2020) Quality of whole genome sequencing from blood versus saliva derived DNA in cardiac patients. BMC Med Genomics 13:11. https://doi.org/10.1186/s12920-020-0664-7

Zeberg H, Pääbo S (2020) The major genetic risk factor for severe COVID-19 is inherited from Neanderthals. Nature 587:610–612. https://doi.org/10.1038/s41586-020-2818-3

Zeberg H, Pääbo S (2021) A genomic region associated with protection against severe COVID-19 is inherited from Neandertals. Proc Natl Acad Sci USA 118: e2026309118. https://doi.org/10.1073/pnas.2026309118

Zenk-Möltgen W, Akdeniz E, Katsanidou A, et al. (2018) Factors influencing the data sharing behavior of researchers in sociology and political science. J Doc 74:1053-1073. https://doi.org/10.1108/JD-09-2017-0126

Associate Editor, Fabio di Vincenzo