The reuse of public datasets in the life sciences: potential risks and rewards

Katharina Sielemann [Corresp., Equal first author, 1, 2], **Alenka Hafner** [Equal first author, 3], **Boas Pucker** [1, 4]

[1] Genetics and Genomics of Plants, Center for Biotechnology (CeBiTec), Bielefeld University, Bielefeld, Germany

[2] Graduate School DILS, Bielefeld Institute for Bioinformatics Infrastructure (BIBI), Bielefeld University, Bielefeld, Germany

[3] independent researcher, Maribor, Slovenia

[4] Evolution and Diversity, Department of Plant Sciences, University of Cambridge, Cambridge, United Kingdom

Corresponding Author: Katharina Sielemann
Email address: kfrey@cebitec.uni-bielefeld.de

The 'big data revolution' has enabled novel types of analyses in the life sciences, facilitated by public sharing and reuse of datasets. Here, we review the prodigious potential of reusing publicly available datasets and the challenges, limitations and risks associated with it. Possible solutions to issues and research integrity considerations are also discussed. Due to the prominence, abundance and wide distribution of sequencing data, we focus on the reuse of publicly available sequence datasets. We define ‘successful reuse’ as the use of previously published data to enable novel scientific findings and use selected examples of such reuse from different disciplines to illustrate the enormous potential of the practice, while acknowledging their respective limitations and risks. A checklist to determine the reuse value and potential of a particular dataset is also provided. The open discussion of data reuse and the establishment of the practice as a norm has the potential to benefit all stakeholders in the life sciences.

The reuse of public datasets in the life sciences: potential risks and rewards

Katharina Sielemann [1,2,+,*], Alenka Hafner [3,+], Boas Pucker [1,4]

[1] Genetics and Genomics of Plants, Center for Biotechnology (CeBiTec), Bielefeld University, Bielefeld, Germany
[2] Graduate School DILS, Bielefeld Institute for Bioinformatics Infrastructure (BIBI), Bielefeld University, Bielefeld, Germany
[3] Independent researcher, Maribor, Slovenia
[4] Evolution and Diversity, Department of Plant Sciences, University of Cambridge, Cambridge, United Kingdom
[+] contributed equally
Corresponding Author: Katharina Sielemann [1,+,*]
Universitätsstraße 25, Bielefeld, Nordrhein-Westfalen, 33602, Germany
Email address: kfrey@cebitec.uni-bielefeld.de

Abstract

The 'big data revolution' has enabled novel types of analyses in the life sciences, facilitated by public sharing and reuse of datasets. Here, we review the prodigious potential of reusing publicly available datasets and the challenges, limitations and risks associated with it. Possible solutions to issues and research integrity considerations are also discussed. Due to the prominence, abundance and wide distribution of sequencing data, we focus on the reuse of publicly available sequence datasets. We define ‘successful reuse’ as the use of previously published data to enable novel scientific findings and use selected examples of such reuse from different disciplines to illustrate the enormous potential of the practice, while acknowledging their respective limitations and risks. A checklist to determine the reuse value and potential of a particular dataset is also provided. The open discussion of data reuse and the establishment of the practice as a norm has the potential to benefit all stakeholders in the life sciences.

Introduction

Data reuse as a part of the ‘big data’ revolution

Data reuse is an essential component of open science and has been facilitated by the ‘big data’ revolution. The transition from (hand) written notes to datasets stored on hard drives can be viewed as the first step on the road to effective data reuse in the life sciences (Fig. 1), allowing the generation of multiple copies at almost no additional cost. The second step was improved connectivity, which was provided by the internet. Together, these technological advances in data storage and transfer enabled a worldwide exchange of ‘big data’, which is common in biology

(e.g. sequence data). This technological basis made data sharing possible (Fig. 1). Next, it needs to become convenient for researchers to share data through increased accessibility. Obligations and benefits lead to established sharing behaviour and more datasets become available which can then be reused in turn. Finally, it becomes common practice and a habit to share all data resulting in a positive feedback loop. Data sharing is already common in several disciplines, including genomics, neuroscience, geoscience and astronomy and an increasing number of studies reuse shared data (Pierce et al., 2019; Tenopir et al., 2020).

For clarity purposes, we distinguish here between fair reuse (for novel purposes, e.g. meta-analysis), reproduction of previous studies with available data (a vital component of open science), and unjust reuse (dual publication and plagiarism). Alongside the reproduction of studies to test findings, only fair reuse should be considered a valid component of open science and is the type of reuse our discussion focuses on.

To establish data sharing as a norm it had to be introduced through obligations and promoted through benefits for researchers (McKiernan et al., 2016). Numerous funding agencies, publishers (e.g. Nature: ("Announcement," 2016), NSF: ("Biological Sciences Guidance on Data Management Plans"), PLOS: ("Plos-One - Data Availability")) and private foundations require all data be made publicly available within a certain time frame, with an indication that this leads to increased transparency in the field (Parker, Nakagawa & Gurevtich, 2016; Tenopir et al., 2020). Many international data sharing guidelines like FAIR (Findable Accessible Interoperable Reusable) (Wilkinson et al., 2016), TOP (The Transparency and Openness Promotion) (Nosek et al., 2015), Open Data in a Big Data World ("Open Data in a Big Data World," 2016) and the Beijing Declaration (CODATA, 2019) have emerged by the necessity of the 'big data revolution'. The sharing of datasets leads to statistical robustness and allows re-analysis of existing datasets underlying claims ("Open Data in a Big Data World," 2016) while enabling the discovery of novel patterns through meta-analysis (Duvallet et al., 2017). Such data reuse leads to cost reduction, reproduction and accountability of research, scientific discovery and detection of novel biological information, and can also be helpful in other areas like education, business and government (Safran, 2017; Pasquetto, Randles & Borgman, 2017; Porto, Pires & Franco, 2017; Leonelli et al., 2017).

Despite these measures and benefits, data reuse is still not ubiquitous, for which there may be different explanations. In 2017, an analysis of 318 biomedical journals revealed that only 11.9% of journals explicitly stated that data sharing was required as a condition of publication (Vasilevsky et al., 2017). Additionally, a survey of 100 datasets associated with ecological and evolutionary research showed that 56% databases were incomplete, and 64% were archived in a way that partially or entirely prevented reuse (Roche et al., 2015). Thus, if the publicly available datasets are not widely re-examined (either checked for quality and/or re-analysed), enforcement of open science through policy may not be sufficient to harness the full power of global sharing

(Pasquetto, Randles & Borgman, 2017). The main causes of researchers refraining from reusing publicly available datasets are (i) concern about the quality and reliability of data (often warranted), (ii) a lack of awareness about the potential in big data or (iii) insufficient bioinformatics knowledge to mine the data (Denk, 2017). Regardless of the cause, the resulting 'backlog' of under-utilised reliable datasets leads to unnecessary experiments (e.g. extensive repetitive sequencing increasing costs) and likely hides useful undiscovered patterns.

Types of reusable data

There are numerous different types of datasets which harbour reuse potential (Fig. 2). These include but are not limited to (1) publications which are accessible to text mining; (2) sequences of genomes, single genes or plasmids or whole sets of sequence reads, (3) annotations of sequences (e.g. sequence motifs collected in JASPAR (Sandelin, 2004)), (4) chromatography results and mass spectra, (5) information about the structure of proteins, (6) biochemical parameters of enzymes e.g. affinity or reaction speed, (7) geodata e.g. coordinates of observations, (8) images of biological material or geographical regions as well as plots and diagrams, (9) algorithms and software (e.g. code), measurements of automatic sensors and global-scaled observatory networks (Soranno et al., 2015), and (11) phenotypic data (Arend et al., 2016).

Generally, datasets can be classified as primary and derived. Primary can be defined as including direct, experimentally obtained data and derived as including meta-analyses and processed results. Statistical data is an example for the fact that data cannot be always classified into 'primary' and 'derived'. Results of a specific experiment can directly be statistically evaluated (primary) whereas e.g. results of multiple studies can be assessed and compared in a statistical manner (derived). Further, it has to be considered that specific fields require the integration of data of various types, formats and abundance (Leonelli et al., 2017) which is hard to achieve by a single database and therefore requires cooperation to encourage data reuse.

Purpose of the review

Education about the opportunities, challenges, limitations and techniques of data reuse is a vital topic that has not been sufficiently addressed in the literature. This review aims to provide an entry point to the discussion of data reuse as part of open science, accessible to a wider audience of life scientist, not only bioinformaticians. The benefits of data sharing and reasons for refraining from it have been extensively reviewed in fields other than bioinformatics, for example ecology (Hampton et al., 2013; LaDeau et al., 2017), medicine (Wade, 2014; Krumholz, 2014; Safran, 2017; Hulsen et al., 2019) and cell biology (Dolinski & Troyanskaya, 2015). However, the existing reviews do not summarise the common benefits and challenges that arise from data reuse in life sciences.

Here, we highlight the potential of data reuse as well as hurdles which need to be overcome. The presented benefits, challenges, risks and potential solutions range cross different fields and aim to illustrate the inherent characteristics of data reuse. We define ‘successful reuse’ as the use of previously published data to enable novel scientific findings usually resulting in a peer reviewed publication. As showing case studies of reuse could reduce the initial barrier to reuse by demonstrating its value in a more practical manner (Curty et al., 2017), we use examples of such successful reuse of different data types (genome, transcriptome, proteome, metabolome, phenotype and ecosystem) to illustrate the enormous potential of the practice. In addition, we provide a checklist of questions for biologists without extensive experience in handling public datasets to consider when determining whether a particular dataset is fit for reuse. We focus on the reuse of primary data, especially different sequences (as the data type characteristic of the life sciences and continuously producing vast amounts of information). The benefits for individuals and the scientific community as a whole make a strong case for reuse of data but only when the risks and limitation have been taken into account.

Survey methodology

A literature review was performed to draw an informed picture of the benefits and challenges associated with data reuse in the life sciences. The publication survey was performed on 14th June 2020, using the PubMed databases to search for relevant peer-reviewed journal articles. The publication year and type of journal were unrestricted. We entered the following search term: "data reuse"[Title/Abstract] OR "dataset reuse"[Title/Abstract]. PubMed produced 188 results (Table S1). Articles from health science, clinical research and medicine were not considered, unless they were reviews. Databases were also excluded, leaving 25 papers (Table S1). From the sources we located in our non-systematic survey, recurring benefits, potential, limitations, challenges and risks of data reuse were identified and categorised, in order to make them more reader friendly. Where we found the issue not sufficiently explained in the article resulting from the initial search, additional resources were sought on that particular topic (on PubMed using the relevant keywords or directly from articles cited in that particular paper). All used references are cited in ‘Potential of reusing public datasets’, ‘Challenges, limitations and risks of data reuse and possible solutions’, ‘Research Integrity considerations’, and ‘Recognizing the value of data reuse’ sections.

Next, we explored published examples from different fields in the life sciences where data reuse was performed. In order to find examples from across a range of biological disciplines, we first identified different types of data that can be reused in the life sciences (genome, transcriptome, proteome, metabolome, phenotype, ecosystem). The relevant literature was selected through the authors' experience in bioinformatics, genetics and genomics of plants, and plant molecular biology. We combined some more specific terms with the keywords mentioned above, namely e.g. ‘genomics’, ‘bioinformatics’, ‘sequencing reads’ and ‘transcriptomics’ as well as search

term keywords specific for each row in the table including the examples for performed data reuse like e.g. ‘coexpression’, ‘pangenomcis’, ‘network analysis’ or ‘metagenomcis’. We acknowledge this is primarily a selection of successful instances of data reuse that focuses on sequences, with only selected examples from other life sciences, however, we deem it illustrative of the potential of data reuse, which was our aim.

To construct the checklist for the selection of datasets appropriate for reuse, we chose the criteria based on the challenges and limitations identified through the literature review. The questions to consider, possible controls and suggestions were identified through our personal experience and backgrounds in bioinformatics.

Potential of reusing public datasets

Reuse of publicly available data is strongly connected to numerous advantages, not only affecting the scientific community and society but also authors themselves and is generally viewed positively among researchers (Tenopir et al., 2020). Sharing data is the prerequisite for easy availability of reusable data leading to positive reuse behaviour and can be key to improved integrity, transparency and reproducibility (Curty et al., 2017; Tenopir et al., 2020). Below is a collation of common benefits of reuse, though we acknowledge there are other advantages of reuse specific to particular fields and/or data types not listed here. Generally, information loss can be prevented, scientific knowledge can be expanded, authors can profit through higher reputation and even databases can benefit from data reuse.

Preventing information loss

Making data publicly available for reuse is an elegant way to prevent information loss stemming from different issues. Data can be subject to irretrievable loss in case of storage solely on private computers and servers, which may not convert the data to the currently used format and that are subject to failure. This is avoided when data are shared with the public and stored in adequately funded databases with backup mechanisms. There are already numerous public repositories for genomic and gene-expression data, such as the Sequence Read Archive (SRA)/European Nucleotide Archive (ENA) and Gene Expression Omnibus (GEO), respectively. Recently, GEO has been used in a case study to improve dataset reusability a literature recommendation system (Patra et al., 2020) and it is highly recommended over other databases for the submission of RNA-Seq data sets (Bhandary et al., 2018). In medical research, information loss stems from large amounts of gathered data remaining inaccessible to a wider audience (sometimes even the authors) for reuse after the initial publication (Wade, 2014). Moreover, the development of new tools and methods leads to the possibility of extracting more information from a given dataset than was feasible at the time of publication. A notable example of such extraction of new information, is the basecalling step when working with nanopore sequencing data derived from Oxford Nanopore Technologies (ONT) devices. Enhanced algorithms allow higher accuracy or

even the identification of DNA modifications (Liu et al., 2019b,a). Further, meta-analyses (e.g. the prediction of specific genomic features) and machine learning approaches require large amounts of data which are already available, and therefore should not and cannot be produced again. This is especially important in fields where data is complex, like videos in organismal biology (Brainerd et al., 2017). Such examples illustrate the importance of sharing and maintaining existing datasets, alongside supplementing them with new data, in order to prevent information loss.

Expanding scientific knowledge

Reuse of data from different sources holds immense potential for scientific discovery and therefore generally enhances scientific progress (Curty et al., 2017). Integration of data from different sources, for example, of exRNA metadata, biomedical ontologies and Linked Data technologies can facilitate interpretation and hypothesis generation by providing independent biological context (Subramanian et al., 2015). In medicine, reuse of data collected as a by-product of health care has the potential to transform the practice of medicine and delivery of health care, which is a compelling argument of reuse benefits outweighing the risks (Wade, 2014; Safran, 2017). For example, reuse can help eliminate bottlenecks in biomedical research at all translational levels and data-mining (the hypothesis-free search for patterns in data) can reveal potential starting points of experimental medical research (Wade, 2014). Another method of reuse is meta-analysis that can elucidate new patterns and produce novel hypothesis – inaccessible from the analysis of any individual dataset. In turn, gene expression studies generating large data sets tend to look for general patterns thus an in depth inspection of single genes can provide additional insights (Bhandary et al., 2018). Temporal and spatial limitations of a single experiment can be overcome by new combinations of existing data and applications integrating different disciplines can be made possible (Curty et al., 2017) leading to more interdisciplinary collaboration (Tenopir et al., 2020). When examining communities, metagenomics, metatranscriptomics, metabarcoding, and metaproteomic provide insights into their composition and function (ten Hoopen et al., 2017). Crucially, publicly available data can lead to the development of new experimental designs (Grace et al., 2018) and can be connected with complementary knowledge and reused in novel experiments (Martens & Vizcaíno, 2017) that contribute to expanding scientific knowledge.

Maximising time, labour and cost efficiency

Reuse of data saves time and money, thus is more economical, and can offer the opportunity to overcome the restraints of limited experiments, high costs and technical difficulties (Raju, Tsinoremas & Capobianco, 2016; Curty et al., 2017). In medicine, the reliance on expensive experimental research when a wealth of existing data is available has been criticised and reuse of existing data and information presented as a solution (Wade, 2014). In metagenomics, where datasets tend to be in the gigabyte range, appropriate archiving of workflow intermediates for reuse can decrease the costs of reanalysis (ten Hoopen et al., 2017). Additionally, many datasets

deposited in sequence databases like GEO were collected at enormous effort and used only once and so reusing them greatly increases their utility (Patra et al., 2020). The labour efficiency of reuse is also illustrated by the famous eGFP browser (Winter et al., 2007) which provides the content of RNA-Seq datasets in a simple way to biologists. The alternative would be downloading and analysing raw RNA-Seq datasets from the Sequence Read Archive (SRA) which would require a substantial amount of bioinformatics expertise and computational power during the analysis. Valuable computational resources are provided by international and national cloud computing services like Elixir, CyVerse or the German Network for Bioinformatics Infrastructure (de.NBI) as well as by commercial organizations. Data reuse is a profitable choice for researchers to lessen expenses and to shorten the research process as data collection was already performed by others (Curty et al., 2017). Cutting such costs through reuse (Fell, 2019) enables groups with small budgets to harness extensive datasets thus enhancing equality.

Benefits for authors

While the scientific community and society are profiting most from public datasets, there are additional benefits of open access for authors themselves (McKiernan et al., 2016; Leitner et al., 2016; Ali-Khan, Harris & Gold, 2017). Researchers can build a reputation by generating high quality and well-documented datasets. Although compliance with data standards might be seen as an additional burden in some cases, the chances of reuse occurring are increased by providing data in the proper format (Rocca-Serra et al., 2016). Dataset sharing may also increase visibility of the associated research and results in additional citations, an added encouragement for authors (Piwowar, Day & Fridsma, 2007). There are even reuse examples which are possible without biological context, like benchmarking of bioinformatic tools or the identification of patterns (Bhandary et al., 2018). It has been shown that there is a robust, statistically supported citation benefit from ‘open data’ in comparison to similar studies without publicly available data (Piwowar & Vision, 2013). This especially aids early career researchers, who are outsiders of the scientific establishment and likely experience more barriers to other aspects of open science (National Academies of Sciences, Engineering, and Medicine (U.S.) et al., 2018) yet are highly involved in data collection and analysis (Farnham et al., 2017).

Benefits for databases

The reuse of sequence data is of increasing importance due to the large and rapidly growing size of the databases storing them (Fig. 3). The size of the Sequence Read Archive (SRA) alone increased from 3,092,408 entries to 6,243,265 entries within two years (September 2016 - September 2018) (NCBI Resource Coordinators, 2017; Sayers et al., 2019a). This growth rate continues to increase exponentially. GenBank comprises a total of 3,677,023,810,243 sequences (2018) with an increase of 39,52 % in comparison to 2017 (Sayers et al., 2019b). Approximately 120 million sequences and annotations of proteins were available within UniProtKB/TrEMBL in 2018 (The UniProt Consortium, 2019). Since managing an exponentially growing database has

numerous challenges (Lathe et al., 2008), it is important to consider how they can be addressed through changing practices, including data reuse.

The issue of the rate of nucleotide and proteomics data generation growing faster than storage capacity (Cook et al., 2016) can be partially addresses through data reuse. Reusing available data instead of producing new and redundant datasets (i.e. when a large number of datasets that are in consensus is already available) results in a lower number of duplicates (Grace et al., 2018) and keeps databases concise. Since it takes an enterprise to maintain and upgrade the largest and most-used databases (meaning they must be adequately financed to survive), limiting redundancy through reuse is beneficial. One example of a sustainable business model is The Arabidopsis Information Resource (TAIR) database (Lamesch et al., 2012) which is funded by subscriptions from academic and non-profit institutions but allows a limited number of accesses by individuals. The public availability of datasets also allows the development of effective algorithms to tackle the bottleneck of data processing, all without the need to perform any sequencing (i.e. uncoupling the problem from access to sequence technology and allowing participation of e.g. computational fields). Additionally, not only the storage of a large number of datasets but also the actual reuse of the available data might increase funding of public databases and therefore ensure the long-term existence of these infrastructures.

Challenges, limitations and risks of data reuse and possible solutions

As discussed above, open access to datasets and studies would accelerate science while being cost-efficient (Spertus, 2012), however, it is important that the limitations of particular datasets are identified, and the associated risks assessed. This is of particular concern in clinical trials as the results could have a direct impact on the patients involved, e.g. in the case of invalid secondary analyses which might harm public health (*Sharing Clinical Trial Data*, 2015). Here, we discuss selected common disadvantages and acknowledge that reuse of specific data types possesses some field-specific issues not addressed here (e.g. adding to the burden of paperwork in clinical medicine (Safran, 2017)).

Unknown quality

For successful reuse appropriate data quality is required in order to be reliable for the user. There are inherent quality differences between (1) user-submitted public datasets, (2) carefully curated databases for specific organisms, (3) ones with inherently small holding sizes (like PDB or SwissProt) and (4) phenotype databases. Additionally, information regarding experimental design, methods and conditions is often incomplete and results in datasets unsuitable for reuse. Mislabelled or swapped samples alongside intrinsic errors, like missing technical replicates, can be a problem as they are almost impossible to identify when accessing a public dataset. Precise documentation of all steps is crucial to clearly indicate limitations of the generated data product (Soranno et al., 2015). Despite this, the peer review process rarely reaches the datasets and their

descriptions (Patra et al., 2020). There is also a trade-off between the collection of detailed metadata during submission and high submission numbers. However, additional requirements for data submission should not result in fewer publicly available datasets (Rung & Brazma, 2013). Ultimately, the limitations of each study (and dataset) are best known by the primary investigators and not by the community accessing the data - a trade-off that studies based on reused data must consider.

It is also difficult and time-consuming for the user to discover available data which is relevant and suitable for the analysis and further to ensure sufficient quality of these datasets (Curty et al., 2017). Even if complete metadata are provided, accessing numerous different webpages to collect all information associated with a combined data set can be tedious (Bhandary et al., 2018). This leads to a risk of wasted time and effort on flawed data, thus requiring trust in data producers and their methods and techniques (Curty et al., 2017). Moreover, simply using a large amount of publicly available datasets does not inherently lead to correct patterns. Despite the importance of trends revealed from large datasets, it is not necessarily the case that a large number of reads/replicates, with an associated low noise, means that the emerging trend is true. Conversely, one can imagine a trend with low noise and low deviation, produced from a large dataset, but with the data coming from one author/group that has an undetected systematic error. Equally, low noise and the use of a small dataset that is believed to be of "higher quality" (i.e. has been thoroughly checked for errors), may hide a trend or show a nonvalid one. Ultimately, the low availability of necessary metadata in standardized formats, with insufficient additional information leads to a lack of reproducibility and can result in misuse and wrong assumptions (Rung & Brazma, 2013; Curty et al., 2017).

Denormalization

Of particular concern is the circular reuse of data. It can, for instance, lead to a heavily denormalized annotation in databases, i.e., the same data is stored multiple times in the same database under different names/identifiers (Bell & Lord, 2017). When such data duplication is not recorded and the user is not made aware of it, the data distribution in the database does not reflect the true data distribution. For example, sequencing and annotation errors can be propagated by reuse and not eliminated by additional published sequences that would reveal it to be statistically insignificant. For annotations, it has been shown that it is possible to detect low-quality entries (resulting from this denormalization) by looking for specific patterns of provenance in the database (Bell, Collison & Lord, 2013). With respect to gene models, this problem could be addressed in the future through the integration of RNA-seq datasets in the annotation of new genome sequences. In terms of functional annotations, this issue persists because the experimental characterisation of numerous genes in a diverse set of species cannot be expected in the near future.

Comparison and integration of datasets and databases

The comparison and integration of datasets from different sources remains a challenge of reuse (Pasquetto, Randles & Borgman, 2017). In metagenomics, when communities that have been studied independently are compared several issues arise, including differences in workflow, unrecorded variables, non-unified presentation format, and relevant raw data not being publicly available (ten Hoopen et al., 2017). The same issue is illustrated by the enormous differences in annotations provided by the different databases, for example, on NCBI (Genome), ENSEMBL, and Phytozome for the same species. In plant phenotyping, reuse and meta-analysis is challenging as data comes from different experimental sites, plant species and experimental conditions, while including many different data types (Papoutsoglou et al., 2020). This non-comparability is also repeated in medical research (Wade, 2014). Therefore, for any valid comparison between datasets from different databases or for integration of databases themselves, conditions specific to the data type and field have to be satisfied.

Different file types and data structures pose a universal challenge for integration which can be overcome by the use of standards if these are *de facto* accepted by the community (Rocca-Serra et al., 2016). For communication between disciplines, an interdisciplinary team with "brokers" has been recommended for the setup of new databases (Soranno et al., 2015). Consistent use of controlled vocabularies and standardized languages like XML (Soranno et al., 2015) also enhances the value of data collections substantially. For example, in meta-analysis of sequences an established and unified file standard is crucial (ten Hoopen et al., 2017). FASTA (Pearson & Lipman, 1988), FASTQ (Cock et al., 2010), and SAM/BAM (Li et al., 2009) are famous examples of file standards that have allowed the effective exchange of information between numerous groups involved in the earliest sequencing projects (Leonard & Littlejohn, 2004; Ondřej & Dvořák, 2012; Zhang, 2016). Any disparities in the sampling method also have to be taken into account when biological material is concerned, so it is essential they are recorded appropriately (ten Hoopen et al., 2017). When appropriate, unified workflow reporting standards (like described by (ten Hoopen et al., 2017)) within a field would largely remove these inhibitions to reuse. It is for these reasons that guidelines like FAIR (Wilkinson et al., 2016) promote universal metadata standards across-the-board are essential to allow comparison and integration of datasets.

Re-analysis as a possible solution

Re-analysis of publicly available data is one way to tackle the issues of unknown quality and denormalization of data, and integration of databases that arise with reuse. This can be achieved through curation and self-correction, with both being difficult to directly enforce. In the same manner that the re-examination of public biodiversity data leads to error correction (Miller et al., 2015; Zizka et al., 2019), so should sequence repositories reflect changes in the field's consensus (e.g. on specific gene annotations). A way to address some of the risks of reuse through re-analysis would be investing in a controlled environment containing extensively peer-reviewed datasets (Spertus, 2012) and manually-curated databases. A defined, suitable environment or

database could also include follow-up data for a detailed understanding of the primary data and the corresponding results. Further, re-analysis and reproducibility might be improved by conventions for standardization, documentation and organization of analysis workflows (Lowndes et al., 2017). This includes detailed records using open science tools like e.g. shared github repositories (Lowndes et al., 2017). Crucially, reuse cannot occur if produced datasets are not widely released to public. Re-examination of databases, like that by (Grechkin, Poon & Howe, 2017) of SRA and GEO, to automatically identify datasets overdue for release are vital in this effort.

Re-analysis has proven to be efficient with some data types but is not practical in all cases. An excellent example of how investing in a manually curated databases eliminates many issues are 'expression atlases' with annotated sequences checked for quality and re-analysed using standardised methods (Kapushesky et al., 2010). However, regarding the enormous and still increasing amount of sequence data, this is hardly an option for all data types. An analysis of re-used data types already indicates that re-using studies often rely on previously identified differentially expressed genes or calculated gene expression values instead of processing raw data again (Wan & Pavlidis, 2007). Therefore, different strategies might work for different data types or different communities. In all cases, specific standards and formats for data reuse should be applied (Pasquetto, Randles & Borgman, 2017), lest "the wealth of data becomes an unmanageable deluge" (Parekh, Armañanzas & Ascoli, 2015).

Metainformation as a possible solution

The trade-off between public access and unknown quality, can be partially resolved by the publication of metadata (the information about the acquisition, processing and presentation) associated with a particular dataset. So far, most researchers (almost 60 %) share their data and metadata using institution specific standards only or even without general metadata standards at all (Tenopir et al., 2020). Further, the metainformation necessary to make data findable (Tenopir et al., 2020) and to enable successful reuse differs between data types and between fields (Parekh, Armañanzas & Ascoli, 2015; Brainerd et al., 2017; ten Hoopen et al., 2017; Papoutsoglou et al., 2020). The amount of metadata that can feasibly be recorded also varies by field, for example, the metadata about a single blood test on a patient includes countless variables (Safran, 2017).

Submitters need to be aware of the importance of providing accurate and complete metadata, but also that a controlled vocabulary is required to facilitate automatic identification of relevant studies/samples (Bhandary et al., 2018). People accessing the dataset also need to be aware of missing information about a dataset - an issue that can be resolved by including a 'completeness of metadata' search criterium in database search engines, like implemented at NeuroMorpho.org (Parekh, Armañanzas & Ascoli, 2015). The MassIVE Knowledge Base aggregates proteomics data including statistical controls and records of the data origin to ensure high quality of the

datasets (Doerr, 2019) and thus is an example for a database providing data fit for reuse. The recently updated MIAPPE metadata standard for plant phenomic databases is another example of reuse facilitation through metadata formatting and was developed to address the shortcomings of the previous guidelines preventing FAIR-complying reuse (Papoutsoglou et al., 2020).

Additionally, methods for reconstructing existing databases are already being investigated. This can be done by curating existing metadata or extracting more of it through natural language processing techniques (Patra et al., 2020) and through metadata predicting frameworks (Posch et al., 2016). Many sequence databases, such as ENA, are handling this elegantly and submitting users can provide a very basic set of meta-information or provide comprehensive details about their study. There are also easy to follow instructions for the submission process (European Nucleotide Archive (ENA)). Despite such incentives, many datasets still lack descriptions that would allow them to be re-sorted according to their metadata (Patra et al., 2020).

Publication of metadata is a practice already routinely implemented in data papers and data journals. Data papers (already common practice in Astronomy (Abolfathi et al., 2018)) have been indicated as a solution to the quality-check problem (Chavan & Penev, 2011) of reuse by providing descriptions of methods for collecting, processing, and verifying data (Pasquetto, Randles & Borgman, 2017). Widespread publication of such metadata in data journals (Figueiredo, 2017) is vital to the construction of high quality, peer-reviewed datasets. Consequently, data-focused journals, like e.g. GigaScience, Scientific Data, and F1000, emerged during the last years. As long-read sequencing became affordable and paved the way for numerous high continuity assemblies, genome announcements describing new genomic or transcriptomic resources became popular. These publication types provide an elegant solution for data reports if a valuable dataset should be shared with the community but do not meet all criteria for publication as a full research article.

Research integrity considerations

Ethical considerations

Ethical considerations of data reuse are important to inspect in all life sciences fields and have been previously discussed in the context of clinical studies (Wade, 2014; Safran, 2017). Informed consent may not have been given with the knowledge that personal data will be utilised in more than the primary study. When sharing medical data, patients must not be identifiable even if advanced methods are applied. This can require modification or masking of the data set parts e.g. defacing of brain images (Milchenko & Marcus, 2013). Similarly, genomic data as a part of modern medicine is not covered by HIPPA and a parent's genomic data could create a privacy risk for their children (Safran, 2017). Thus, there are ethical considerations specific to reuse of data in human research.

Research 'parasitism'

The use of the same dataset in several different studies by the same author could be considered a type of dual publication in some circumstances (Beaufils & Karlsson, 2013). However, such reuse is not contentious to the same extent as plagiarism if it reveals novel findings and is not reused only to boost the number of publications. The trend of publishing from publicly available data ("The parasite awards - Celebrating rigorous secondary data analysis"; Longo & Drazen, 2016; Pucker & Brockington, 2018; Frey & Pucker, 2020) points to the crux of the matter of research integrity reservations about data reuse that some have. At the far end of this spectrum, there are authors exclusively using publicly available data (not generating their own to cross-check the quality), often choosing research topics/systems-of-interest based on the quality of data and not *vice versa*. This practice is associated with numerous advantages, including intensified use of existing datasets which effectively increases the ratio of value drawn from it compared to the costs of generating it in the first place. While multiple studies can benefit from reuse, long term risks might include funding bodies expecting reuse thus rendering the acquisition of financial support for new experiments more challenging. Despite some expressed concerns regarding such 'research parasitism' (Longo & Drazen, 2016), including the fear of exploitation when acquiring the data was particularly expensive or labour-intensive (National Academies of Sciences, Engineering, and Medicine (U.S.) et al., 2018), the practice of reuse seems to prevail in the open science culture. The above-mentioned 'Parasite awards' use this tongue-in-cheek name after it was introduced to dismay such practices. Due to the numerous benefits of reuse for the scientific community, we believe the term 'research parasitism' is unwarranted when fair reuse has been employed.

Recognising the value of data reuse by recognising the data producer

As perceived efficacy and efficiency of data reuse strongly influence reuse behaviour, reuse would be encouraged by demonstrating its value (Curty et al., 2017). So far, there have been extensive efforts to promote and develop standards for data sharing, but less effort to show the real value of data sharing or to recognize, cite or acknowledge the contributions of data sharing (Pierce et al., 2019; Tenopir et al., 2020). Adequate recognition of data producers would simultaneously accelerate data sharing by eliminating the main barrier: the need to publish first (Tenopir et al., 2020).

Both, data providers and creators of databases, deserve recognition for their work. To a certain degree, this issue can be tackled with data publications, which might also prevent the splitting of a coherent dataset over multiple publications. Soranno et al. (2015) recommend specifically to describe the content of new databases in one paper with all contributors listed as co-authors and a description of the methods for the database development in an additional publication with all researchers involved in the process as co-authors. This is a practice that shares all its benefits with data papers discussed above. The preservation and documentation of data provenance is crucial to acknowledge the support of data providers (Soranno et al., 2015). By supplying a

citable source of the dataset, credit is given to the data producer, which eliminates the concern about ownership by providing an official academic record of provenance.

As barriers for data sharing include concerns about loss of credit (Tenopir et al., 2020), the assimilation of sharing and reuse could be positively influenced by recognition (e.g. awards or other compensatory means, like e.g. co-authorships) of the expansion of scientific discovery through studies reusing available data (Piwowar & Vision, 2013; Curty et al., 2017; Tenopir et al., 2020). A way to recognize the producers of reusable data would be the connection of an identifier for each researcher (e.g. ORCID) with an identifier for the dataset (e.g. DOI) which would then have to be cited in each new publication reusing the initial dataset (Piwowar & Vision, 2013; Pierce et al., 2019). In addition, publishers would have to ensure that these citations are available in a searchable system like Crossref to establish a real link between data generator and publication (Pierce et al., 2019). Such a system, which recognizes researchers regularly for generating data, could substantially influence the assessment of the value of scientific data by academic institutions, funders, and society (Pierce et al., 2019).

Establishing a reuse culture

Despite the reservations highlighted above many facets of data reuse provide incentives for the individual to practice it. Through open science initiatives (including but not limited to those listed in the Introduction), modern biologists are encouraged to make use of publicly available sequence repositories and mine data generated by others. Further, not comparing one's dataset to publicly available analogues can be considered akin to ignoring replicated experiments (Denk, 2017). In fields where scientific progress has immediate and measurable positive impacts, such as medicine, the benefits of data reuse to society quickly outweigh the risks (Safran, 2017). Due to these advantages, there is an argument to be made that data reuse is an ethical obligation in the life sciences.

The statistic that the metagenomes of 20% of papers published between 2016 and 2019 are not publicly accessible (Eckert et al., 2020) demonstrates that there is still a long way to go until data sharing becomes routine. Therefore, open science incentives and database contribution guidelines should require the inclusion of metadata in all submissions to public datasets. Gamification via the implementation of fancy statistics about the data connected to a personal profile for each researcher might be another way to encourage researchers to share their data. A reusability score assigned by the community could increase the quality of the provided metadata. Such practices would not only encourage authors to collect data with reuse in mind (Goodman et al., 2014) but enable productive and valid re-analysis. Additional enforcement could be spot-checking of provided metadata by funding agencies (Bhandary et al., 2018) which are providing the money for the data generation and have an interest in the release of high quality data sets.

Only combination of obligation and encouragement from the publishing and educational spheres are likely to ensure a future of successful and fair data "recycling". The prevailing culture of positive publication bias (Mlinarić, Horvat & Šupak Smolčić, 2017) leads to “missing studies” (the desk-drawer effect) and could also introduce a bias into analysis based on existing data sets (Wan & Pavlidis, 2007). For this reason, young researchers should be encouraged to share their data (if it is of sufficient quality) even if the outcome was "negative". Additionally, bioinformatics should be integrated into the education of next generation life scientists to increase data re-use capacities (Soranno et al., 2015). As one learns about a new method of data collection so should one learn about the metadata that must accompany it in publication to render the data useful to others.

Examples of successful data reuse

There are already numerous examples of successful studies from various areas of life science which involve intensive reuse of public datasets. Genomic data can, for example, be harnessed for pangenomic analyses (Montenegro et al., 2017) while transcriptomic and ChIP-seq data might be useful for the investigation or construction of regulatory networks (Chow et al., 2019). Phylogenetic analysis of groups from individual gene families (Du et al., 2016) to whole taxonomic groups (Bowles, Bechtold & Paps, 2020) benefits from reuse of genome, transcriptome and proteome data. Further, several tools and techniques have been developed for e.g. mining antimicrobial peptides from public databases (Porto, Pires & Franco, 2017). The taxonomic classification of sequences identified in metagenomic studies is another application which heavily relies on available data as the quality of a study scales with the quality and size of the available data (Breitwieser, Lu & Salzberg, 2019). It can also be expected that machine learning will become ever more ubiquitous in combination with other methods due to its ability to tackle large datasets and reveal novel patterns. Finally, in fields like modelling, the access to reusable data for the generation of models is even required (Curty et al., 2017).

In the life sciences, data reuse spans many data types and fields with substantial overlap in both categories. Table 1 shows selected reuse cases in the life sciences that cover many areas and concepts of data reuse sorted by the type of the analysed data. With every individual case of reuse, one must also consider the specific disadvantages associated with each approach. As highlighted above, there are risks to reuse. In addition to listing successful examples, Table 1 includes the associated limitations and risks associated with that particular method of reuse based on our assessment as data consumers and researchers. They aim illustrate the types of considerations that must be taken into account when reusing a specific type of dataset.

Assessment of reuse suitability for the selection of datasets

We have seen that in the selection of appropriate datasets for reuse, limitations and potential errors must be considered, in order to tap into the full potential of the practice, while avoiding invalid analysis. It has previously been demonstrated that posteori analysis of dataset quality is possible with the quality control metrics for proteomics datasets to assess their suitability for a particular reuse purpose (Foster et al., 2011). Here, we provide a checklist (Table 2) to aid in the selection of datasets suitable for reuse, including suggestions, suitable controls and questions to consider prior to the re-analysis of public data. The questions are inquiries that a life scientist might consider when assessing a dataset of unknown quality and were determined by the authors.

Conclusions

Data reuse is quickly becoming a ubiquitous part of research in the life sciences and scientists increasingly recognize the benefits of open reusable data (Tenopir et al., 2020). There are different steps to achieve and develop actual data-sharing behaviour which complies with 'open data' principles. As stated above (Fig. 1), technological progress together with changing research behaviour make 'open data' and its reuse 1: possible, 2: easy and 3: desirable. Considering the increasing quantity of available public data, *in silico* analyses are starting to supersede classic 'wet lab' experiments in some areas. However, it is still difficult to determine on a case-by-case basis whether the cost-benefit analysis favours data reuse with the associated risks or the increasingly cheaper/faster sequencing.

One of the factors promoting reuse behaviour would be the demonstration of its value (Curty et al., 2017). The provided list of successful examples in Table 1, which is only a narrow selection of studies conducted reusing publicly available datasets, illustrates the high potential of data reuse, thus demonstrating its value. General limitations of these example studies include batch effects, quality issues and incomplete accuracy of predictions due to missing parameters (Fig. 4). Ultimately, the data itself, necessary to gain new scientific knowledge, is already available and only 'waits' to be extensively investigated to answer open scientific questions.

The reuse of publicly available scientific datasets leads to a reduction of costs and saves time, encourages reproducible research, enables the detection of novel information and has benefits for authors themselves (Fig. 4). Across the life sciences, there still remain some outstanding questions and challenges (Fig. 5). Considering all the advantages and taking into account the limitations we highly recommend and encourage data reuse when one is confident in that the reuse can be categorised as fair. We believe that the discussion of responsible data reuse must become more common in the life sciences so everyone can benefit from the largely untapped data resource.

Acknowledgements

We acknowledge support for the Article Processing Charge by the Deutsche Forschungsgemeinschaft and the Open Access Publication Fund of Bielefeld University. We are grateful to Katharina Schiller for helpful comments on the manuscript.

References

Abolfathi B, Aguado D, Aguilar G, Prieto CA, Almeida A, Ananna TT, Anders F, Anderson SF, Andrews BH, Anguiano B. 2018. The fourteenth data release of the Sloan Digital Sky Survey: First spectroscopic data from the extended Baryon Oscillation Spectroscopic Survey and from the second phase of the Apache Point Observatory Galactic Evolution Experiment. *The Astrophysical Journal Supplement Series* 235:42.

Ali-Khan SE, Harris LW, Gold ER. 2017. Motivating participation in open science by examining researcher incentives. *eLife* 6:e29319. DOI: 10.7554/eLife.29319.

Announcement: Where are the data? 2016. *Nature* 537:138–138. DOI: 10.1038/537138a.

Arend D, Junker A, Scholz U, Schüler D, Wylie J, Lange M. 2016. PGP repository: a plant phenomics and genomics data publication infrastructure. *Database* 2016:baw033. DOI: 10.1093/database/baw033.

Beaufils P, Karlsson J. 2013. Legitimate division of large datasets, salami slicing and dual publication. Where does a fraud begin? *Orthopaedics & Traumatology: Surgery & Research* 99:121–122. DOI: 10.1016/j.otsr.2013.01.001.

Bell MJ, Collison M, Lord P. 2013. Can Inferred Provenance and Its Visualisation Be Used to Detect Erroneous Annotation? A Case Study Using UniProtKB. *PLoS ONE* 8:e75541. DOI: 10.1371/journal.pone.0075541.

Bell MJ, Lord P. 2017. On patterns and re-use in bioinformatics databases. *Bioinformatics* 33:2731–2736. DOI: 10.1093/bioinformatics/btx310.

Bhandary P, Seetharam AS, Arendsee ZW, Hur M, Wurtele ES. 2018. Raising orphans from a metadata morass: A researcher's guide to re-use of public 'omics data. *Plant Science* 267:32–47. DOI: 10.1016/j.plantsci.2017.10.014.

Biological Sciences Guidance on Data Management Plans. *Available at https://www.nsf.gov/bio/biodmp.jsp* (accessed February 3, 2020).

Bowles AMC, Bechtold U, Paps J. 2020. The Origin of Land Plants Is Rooted in Two Bursts of Genomic Novelty. *Current Biology*:S0960982219315957. DOI: 10.1016/j.cub.2019.11.090.

Brainerd EL, Blob RW, Hedrick TL, Creamer AT, Müller UK. 2017. Data Management Rubric for Video Data in Organismal Biology. *Integrative and Comparative Biology* 57:33–47. DOI: 10.1093/icb/icx060.

Breitwieser FP, Lu J, Salzberg SL. 2019. A review of methods and databases for metagenomic classification and assembly. *Briefings in Bioinformatics* 20:1125–1136. DOI: 10.1093/bib/bbx120.

Brinkrolf C, Henke NA, Ochel L, Pucker B, Kruse O, Lutter P. 2018. Modeling and Simulating the Aerobic Carbon Metabolism of a Green Microalga Using Petri Nets and New Concepts of VANESA. *Journal of Integrative Bioinformatics* 15. DOI: 10.1515/jib-2018-0018.

Chavan V, Penev L. 2011. The data paper: a mechanism to incentivize data publishing in biodiversity science. *BMC Bioinformatics* 12:S2. DOI: 10.1186/1471-2105-12-S15-S2.

Cheng W-C, Chang C-W, Chen C-R, Tsai M-L, Shu W-Y, Li C-Y, Hsu IC. 2011. Identification of Reference Genes across Physiological States for qRT-PCR through Microarray Meta-Analysis. *PLoS ONE* 6:e17347. DOI: 10.1371/journal.pone.0017347.

Chow C-N, Lee T-Y, Hung Y-C, Li G-Z, Tseng K-C, Liu Y-H, Kuo P-L, Zheng H-Q, Chang W-C. 2019. PlantPAN3.0: a new and updated resource for reconstructing transcriptional regulatory networks from ChIP-seq experiments in plants. *Nucleic Acids Research* 47:D1155–D1163. DOI: 10.1093/nar/gky1081.

688 Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. 2010. The Sanger FASTQ file format for
689 sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids*
690 *Research* 38:1767–1771. DOI: 10.1093/nar/gkp1137.
691 CODATA. 2019.The Beijing Declaration on Research Data. *Available at*
692 *https://zenodo.org/record/3552330#.XgT5eS3MwfE* (accessed February 3, 2020).
693 Cook CE, Bergman MT, Finn RD, Cochrane G, Birney E, Apweiler R. 2016. The European
694 Bioinformatics Institute in 2016: Data growth and integration. *Nucleic Acids Research*
695 44:D20–D26. DOI: 10.1093/nar/gkv1352.
696 Curty RG, Crowston K, Specht A, Grant BW, Dalton ED. 2017. Attitudes and norms affecting
697 scientists' data reuse. *PLOS ONE* 12:e0189288. DOI: 10.1371/journal.pone.0189288.
698 Delmont TO, Eren AM. 2016. Identifying contamination with advanced visualization and analysis
699 practices: metagenomic approaches for eukaryotic genome assemblies. *PeerJ* 4:e1839.
700 DOI: 10.7717/peerj.1839.
701 Denk F. 2017. Don't let useful data go to waste. *Nature* 543:7–7. DOI: 10.1038/543007a.
702 Dierckxsens N, Mardulyn P, Smits G. 2016. NOVOPlasty: *de novo* assembly of organelle
703 genomes from whole genome data. *Nucleic Acids Research*:gkw955. DOI:
704 10.1093/nar/gkw955.
705 Doerr A. 2019. Proteomics data reuse with MassIVE-KB. *Nature Methods* 16:26–26. DOI:
706 10.1038/s41592-018-0283-9.
707 Dolinski K, Troyanskaya OG. 2015. Implications of Big Data for cell biology. *Molecular Biology*
708 *of the Cell* 26:2575–2578. DOI: 10.1091/mbc.E13-12-0756.
709 Du H, Ran F, Dong H-L, Wen J, Li J-N, Liang Z. 2016. Genome-Wide Analysis, Classification,
710 Evolution, and Expression Analysis of the Cytochrome P450 93 Family in Land Plants.
711 *PLOS ONE* 11:e0165020. DOI: 10.1371/journal.pone.0165020.

712 Duvallet C, Gibbons SM, Gurry T, Irizarry RA, Alm EJ. 2017. Meta-analysis of gut microbiome
713 studies identifies disease-specific and shared responses. *Nature Communications*
714 8:1784. DOI: 10.1038/s41467-017-01973-8.
715 Eckert EM, Di Cesare A, Fontaneto D, Berendonk TU, Bürgmann H, Cytryn E, Fatta-Kassinos
716 D, Franzetti A, Larsson DGJ, Manaia CM, Pruden A, Singer AC, Udikovic-Kolic N, Corno
717 G. 2020. Every fifth published metagenome is not available to science. *PLOS Biology*
718 18:e3000698. DOI: 10.1371/journal.pbio.3000698.
719 European Nucleotide Archive (ENA).ENA: Guidelines and Tutorials. *Available at https://ena-*
720 *docs.readthedocs.io/en/latest/* (accessed February 3, 2020).
721 Farnham A, Kurz C, Öztürk MA, Solbiati M, Myllyntaus O, Meekes J, Pham TM, Paz C,
722 Langiewicz M, Andrews S, Kanninen L, Agbemabiese C, Guler AT, Durieux J, Jasim S,
723 Viessmann O, Frattini S, Yembergenova D, Benito CM, Porte M, Grangeray-Vilmint A,
724 Curiel RP, Rehncrona C, Malas T, Esposito F, Hettne K. 2017. Early career researchers
725 want Open Science. *Genome Biology* 18:221. DOI: 10.1186/s13059-017-1351-7.
726 Fell MJ. 2019. The Economic Impacts of Open Science: A Rapid Evidence Assessment.
727 *Publications* 7:46. DOI: 10.3390/publications7030046.
728 Figueiredo AS. 2017. Data Sharing: Convert Challenges into Opportunities. *Frontiers in Public*
729 *Health* 5:327. DOI: 10.3389/fpubh.2017.00327.
730 Foster JM, Degroeve S, Gatto L, Visser M, Wang R, Griss J, Apweiler R, Martens L. 2011. A
731 posteriori quality control for the curation and reuse of public proteomics data.
732 *PROTEOMICS* 11:2182–2194. DOI: 10.1002/pmic.201000602.
733 Frey K, Pucker B. 2020. Animal, Fungi, and Plant Genome Sequences Harbor Different Non-
734 Canonical Splice Sites. *Cells* 9:458. DOI: 10.3390/cells9020458.
735 Goodman A, Pepe A, Blocker AW, Borgman CL, Cranmer K, Crosas M, Di Stefano R, Gil Y,
736 Groth P, Hedstrom M, Hogg DW, Kashyap V, Mahabal A, Siemiginowska A, Slavkovic A.

2014. Ten Simple Rules for the Care and Feeding of Scientific Data. *PLoS Computational Biology* 10:e1003542. DOI: 10.1371/journal.pcbi.1003542.

Grace JO, Malik A, Reichman H, Munitz A, Barski A, Fulkerson PC. 2018. Reuse of public, genome-wide, murine eosinophil expression data for hypotheses development. *Journal of Leukocyte Biology* 104:185–193. DOI: 10.1002/JLB.1MA1117-444R.

Grechkin M, Poon H, Howe B. 2017. Wide-Open: Accelerating public data release by automating detection of overdue datasets. *PLOS Biology* 15:e2002477. DOI: 10.1371/journal.pbio.2002477.

Gyawali A, Shrestha V, Guill KE, Flint-Garcia S, Beissinger TM. 2019. Single-plant GWAS coupled with bulk segregant analysis allows rapid identification and corroboration of plant-height candidate SNPs. *BMC Plant Biology* 19:412. DOI: 10.1186/s12870-019-2000-y.

Hampton SE, Strasser CA, Tewksbury JJ, Gram WK, Budden AE, Batcheller AL, Duke CS, Porter JH. 2013. Big data and the future of ecology. *Frontiers in Ecology and the Environment* 11:156–162. DOI: 10.1890/120103.

Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. 2002. A comprehensive review of genetic association studies. *Genetics in Medicine* 4:45–61. DOI: 10.1097/00125817-200203000-00002.

ten Hoopen P, Finn RD, Bongo LA, Corre E, Fosso B, Meyer F, Mitchell A, Pelletier E, Pesole G, Santamaria M, Willassen NP, Cochrane G. 2017. The metagenomic data life-cycle: standards and best practices. *GigaScience* 6. DOI: 10.1093/gigascience/gix047.

Hruz T, Wyss M, Docquier M, Pfaffl MW, Masanetz S, Borghi L, Verbrugghe P, Kalaydjieva L, Bleuler S, Laule O, Descombes P, Gruissem W, Zimmermann P. 2011. RefGenes: identification of reliable and condition specific reference genes for RT-qPCR data normalization. *BMC Genomics* 12:156. DOI: 10.1186/1471-2164-12-156.

762 Hulsen T, Jamuar SS, Moody AR, Karnes JH, Varga O, Hedensted S, Spreafico R, Hafler DA,
763 McKinney EF. 2019. From Big Data to Precision Medicine. *Frontiers in Medicine* 6:34.
764 DOI: 10.3389/fmed.2019.00034.
765 Jetz W, Fine PVA. 2012. Global Gradients in Vertebrate Diversity Predicted by Historical Area-
766 Productivity Dynamics and Contemporary Environment. *PLoS Biology* 10:e1001292.
767 DOI: 10.1371/journal.pbio.1001292.
768 Kapushesky M, Emam I, Holloway E, Kurnosov P, Zorin A, Malone J, Rustici G, Williams E,
769 Parkinson H, Brazma A. 2010. Gene Expression Atlas at the European Bioinformatics
770 Institute. *Nucleic Acids Research* 38:D690–D698. DOI: 10.1093/nar/gkp936.
771 Keilwagen J, Hartung F, Grau J. 2019. GeMoMa: Homology-Based Gene Prediction Utilizing
772 Intron Position Conservation and RNA-seq Data. In: Kollmar M ed. *Gene Prediction*.
773 New York, NY: Springer New York, 161–177. DOI: 10.1007/978-1-4939-9173-0_9.
774 Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and
775 accurate long-read assembly via adaptive *k* -mer weighting and repeat separation.
776 *Genome Research* 27:722–736. DOI: 10.1101/gr.215087.116.
777 Krumholz HM. 2014. Big Data And New Knowledge In Medicine: The Thinking, Training, And
778 Tools Needed For A Learning Health System. *Health Affairs* 33:1163–1170. DOI:
779 10.1377/hlthaff.2014.0053.
780 Kryukov K, Imanishi T. 2016. Human Contamination in Public Genome Assemblies. *PLOS ONE*
781 11:e0162424. DOI: 10.1371/journal.pone.0162424.
782 Kwon MJ, Oh E, Lee S, Roh MR, Kim SE, Lee Y, Choi Y-L, In Y-H, Park T, Koh SS, Shin YK.
783 2009. Identification of Novel Reference Genes Using Multiplatform Expression Data and
784 Their Validation for Quantitative Gene Expression Analysis. *PLoS ONE* 4:e6162. DOI:
785 10.1371/journal.pone.0006162.
786 LaDeau SL, Han BA, Rosi-Marshall EJ, Weathers KC. 2017. The Next Decade of Big Data in
787 Ecosystem Science. *Ecosystems* 20:274–283. DOI: 10.1007/s10021-016-0075-y.

788 Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K,
789 Alexander DL, Garcia-Hernandez M, Karthikeyan AS, Lee CH, Nelson WD, Ploetz L,
790 Singh S, Wensel A, Huala E. 2012. The Arabidopsis Information Resource (TAIR):
791 improved gene annotation and new tools. *Nucleic Acids Research* 40:D1202–D1210.
792 DOI: 10.1093/nar/gkr1090.
793 Lathe W, Williams J, Mangan M, Karolchik D. 2008. Genomic data resources: challenges and
794 promises. *Nature Education* 1:2.
795 Leitner F, Bielza C, Hill SL, Larrañaga P. 2016. Data Publications Correlate with Citation Impact.
796 *Frontiers in Neuroscience* 10. DOI: 10.3389/fnins.2016.00419.
797 Leonard SA, Littlejohn TG. 2004. Common File Formats. *Current Protocols in Bioinformatics* 5.
798 DOI: 10.1002/0471250953.bia01bs05.
799 Leonelli S, Davey RP, Arnaud E, Parry G, Bastow R. 2017. Data management and best practice
800 for plant science. *Nature Plants* 3:17086. DOI: 10.1038/nplants.2017.86.
801 Li H. 2020.auN: a new metric to measure assembly contiguity. *Available at*
802 *http://lh3.github.io/2020/04/08/a-new-metric-on-assembly-contiguity*
803 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R,
804 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map
805 format and SAMtools. *Bioinformatics* 25:2078–2079. DOI:
806 10.1093/bioinformatics/btp352.
807 Liu Q, Fang L, Yu G, Wang D, Xiao C-L, Wang K. 2019a. Detection of DNA base modifications
808 by deep recurrent neural network on Oxford Nanopore sequencing data. *Nature*
809 *Communications* 10:2449. DOI: 10.1038/s41467-019-10168-2.
810 Liu Q, Georgieva DC, Egli D, Wang K. 2019b. NanoMod: a computational tool to detect DNA
811 modifications using Nanopore long-read sequencing data. *BMC Genomics* 20:78. DOI:
812 10.1186/s12864-018-5372-8.

Longo DL, Drazen JM. 2016. Data Sharing. *New England Journal of Medicine* 374:276–277. DOI: 10.1056/NEJMe1516564.

Longo MS, O'Neill MJ, O'Neill RJ. 2011. Abundant Human DNA Contamination Identified in Non-Primate Genome Databases. *PLoS ONE* 6:e16410. DOI: 10.1371/journal.pone.0016410.

Lowndes JSS, Best BD, Scarborough C, Afflerbach JC, Frazier MR, O'Hara CC, Jiang N, Halpern BS. 2017. Our path to better science in less time using open data science tools. *Nature Ecology & Evolution* 1:0160. DOI: 10.1038/s41559-017-0160.

Lu H, Giordano F, Ning Z. 2016. Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics, Proteomics & Bioinformatics* 14:265–279. DOI: 10.1016/j.gpb.2016.05.004.

Ma X, Zhao H, Xu W, You Q, Yan H, Gao Z, Su Z. 2018. Co-expression Gene Network Analysis and Functional Module Identification in Bamboo Growth and Development. *Frontiers in Genetics* 9:574. DOI: 10.3389/fgene.2018.00574.

Marigorta UM, Rodríguez JA, Gibson G, Navarro A. 2018. Replicability and Prediction: Lessons and Challenges from GWAS. *Trends in Genetics* 34:504–517. DOI: 10.1016/j.tig.2018.03.005.

Martens L, Vizcaíno JA. 2017. A Golden Age for Working with Public Proteomics Data. *Trends in Biochemical Sciences* 42:333–341. DOI: 10.1016/j.tibs.2017.01.001.

McKiernan EC, Bourne PE, Brown CT, Buck S, Kenall A, Lin J, McDougall D, Nosek BA, Ram K, Soderberg CK, Spies JR, Thaney K, Updegrove A, Woo KH, Yarkoni T. 2016. How open science helps researchers succeed. *eLife* 5:e16800. DOI: 10.7554/eLife.16800.

Merchant S, Wood DE, Salzberg SL. 2014. Unexpected cross-species contamination in genome sequencing projects. *PeerJ* 2:e675. DOI: 10.7717/peerj.675.

Milchenko M, Marcus D. 2013. Obscuring Surface Anatomy in Volumetric Imaging Data. *Neuroinformatics* 11:65–75. DOI: 10.1007/s12021-012-9160-3.

838 Miller J, Georgiev T, Stoev P, Sautter G, Penev L. 2015. Corrected data re-harvested: curating
839 literature in the era of networked biodiversity informatics. *Biodiversity Data Journal*
840 3:e4552. DOI: 10.3897/BDJ.3.e4552.
841 Mitchell AL, Attwood TK, Babbitt PC, Blum M, Bork P, Bridge A, Brown SD, Chang H-Y, El-
842 Gebali S, Fraser MI, Gough J, Haft DR, Huang H, Letunic I, Lopez R, Luciani A, Madeira
843 F, Marchler-Bauer A, Mi H, Natale DA, Necci M, Nuka G, Orengo C, Pandurangan AP,
844 Paysan-Lafosse T, Pesseat S, Potter SC, Qureshi MA, Rawlings ND, Redaschi N,
845 Richardson LJ, Rivoire C, Salazar GA, Sangrador-Vegas A, Sigrist CJA, Sillitoe I, Sutton
846 GG, Thanki N, Thomas PD, Tosatto SCE, Yong S-Y, Finn RD. 2019. InterPro in 2019:
847 improving coverage, classification and access to protein sequence annotations. *Nucleic*
848 *Acids Research* 47:D351–D360. DOI: 10.1093/nar/gky1100.
849 Mlinarić A, Horvat M, Šupak Smolčić V. 2017. Dealing with the positive publication bias: Why
850 you should really publish your negative results. *Biochemia Medica* 27:030201. DOI:
851 10.11613/BM.2017.030201.
852 Montenegro JD, Golicz AA, Bayer PE, Hurgobin B, Lee H, Chan C-KK, Visendi P, Lai K, Doležel
853 J, Batley J, Edwards D. 2017. The pangenome of hexaploid bread wheat. *The Plant*
854 *Journal* 90:1007–1013. DOI: 10.1111/tpj.13515.
855 Mooij WM, Trolle D, Jeppesen E, Arhonditsis G, Belolipetsky PV, Chitamwebwa DBR,
856 Degermendzhy AG, DeAngelis DL, De Senerpont Domis LN, Downing AS, Elliott JA,
857 Fragoso CR, Gaedke U, Genova SN, Gulati RD, Håkanson L, Hamilton DP, Hipsey MR,
858 ‘t Hoen J, Hülsmann S, Los FH, Makler-Pick V, Petzoldt T, Prokopkin IG, Rinke K,
859 Schep SA, Tominaga K, Van Dam AA, Van Nes EH, Wells SA, Janse JH. 2010.
860 Challenges and opportunities for integrating lake ecosystem modelling approaches.
861 *Aquatic Ecology* 44:633–667. DOI: 10.1007/s10452-010-9339-3.
862 National Academies of Sciences, Engineering, and Medicine (U.S.), National Academies of
863 Sciences, Engineering, and Medicine (U.S.), National Academies of Sciences,

Engineering, and Medicine (U.S.), National Academies of Sciences, Engineering, and Medicine (U.S.) (eds.). 2018. *Open science by design: realizing a vision for 21st century research*. Washington, DC: The National Academies Press.

NCBI Resource Coordinators. 2017. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Research* 45:D12–D17. DOI: 10.1093/nar/gkw1071.

Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Breckler SJ, Buck S, Chambers CD, Chin G, Christensen G, Contestabile M, Dafoe A, Eich E, Freese J, Glennerster R, Goroff D, Green DP, Hesse B, Humphreys M, Ishiyama J, Karlan D, Kraut A, Lupia A, Mabry P, Madon T, Malhotra N, Mayo-Wilson E, McNutt M, Miguel E, Paluck EL, Simonsohn U, Soderberg C, Spellman BA, Turitto J, VandenBos G, Vazire S, Wagenmakers EJ, Wilson R, Yarkoni T. 2015. Promoting an open research culture. *Science* 348:1422–1425. DOI: 10.1126/science.aab2374.

Ondřej V, Dvořák P. 2012. Bioinformatics: a history of evolution *in silico*. *Journal of Biological Education* 46:252–259. DOI: 10.1080/00219266.2012.716776.

Open Data in a Big Data World. 2016. *Chemistry International* 38. DOI: 10.1515/ci-2016-0208.

Papoutsoglou EA, Faria D, Arend D, Arnaud E, Athanasiadis IN, Chaves I, Coppens F, Cornut G, Costa BV, Ćwiek-Kupczyńska H, Droesbeke B, Finkers R, Gruden K, Junker A, King GJ, Krajewski P, Lange M, Laporte M, Michotey C, Oppermann M, Ostler R, Poorter H, Ramírez-Gonzalez R, Ramšak Ž, Reif JC, Rocca-Serra P, Sansone S, Scholz U, Tardieu F, Uauy C, Usadel B, Visser RGF, Weise S, Kersey PJ, Miguel CM, Adam-Blondon A, Pommier C. 2020. Enabling reusability of plant phenomic datasets with MIAPPE 1.1. *New Phytologist* 227:260–273. DOI: 10.1111/nph.16544.

Parekh R, Armañanzas R, Ascoli GA. 2015. The importance of metadata to assess information content in digital reconstructions of neuronal morphology. *Cell and Tissue Research* 360:121–127. DOI: 10.1007/s00441-014-2103-6.

890 Parker TH, Nakagawa S, Gurevtich J. 2016. Open data: towards full transparency. *Nature*
891 538:459–459. DOI: 10.1038/538459d.
892 Pasquetto IV, Randles BM, Borgman CL. 2017. On the Reuse of Scientific Data. *Data Science*
893 *Journal* 16:8. DOI: 10.5334/dsj-2017-008.
894 Patra BG, Maroufy V, Soltanalizadeh B, Deng N, Zheng WJ, Roberts K, Wu H. 2020. A content-
895 based literature recommendation system for datasets to improve data reusability – A
896 case study on Gene Expression Omnibus (GEO) datasets. *Journal of Biomedical*
897 *Informatics* 104:103399. DOI: 10.1016/j.jbi.2020.103399.
898 Pearson WR, Lipman DJ. 1988. Improved tools for biological sequence comparison.
899 *Proceedings of the National Academy of Sciences* 85:2444–2448. DOI:
900 10.1073/pnas.85.8.2444.
901 Peng Y, Leung HCM, Yiu SM, Chin FYL. 2012. IDBA-UD: a de novo assembler for single-cell
902 and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28:1420–
903 1428. DOI: 10.1093/bioinformatics/bts174.
904 Persson S, Wei H, Milne J, Page GP, Somerville CR. 2005. Identification of genes required for
905 cellulose synthesis by regression analysis of public microarray data sets. *Proceedings of*
906 *the National Academy of Sciences* 102:8633–8638. DOI: 10.1073/pnas.0503392102.
907 Pierce HH, Dev A, Statham E, Bierer BE. 2019. Credit data generators for data reuse. *Nature*
908 570:30–32. DOI: 10.1038/d41586-019-01715-4.
909 Piwowar HA, Day RS, Fridsma DB. 2007. Sharing Detailed Research Data Is Associated with
910 Increased Citation Rate. *PLoS ONE* 2:e308. DOI: 10.1371/journal.pone.0000308.
911 Piwowar HA, Vision TJ. 2013. Data reuse and the open data citation advantage. *PeerJ* 1:e175.
912 DOI: 10.7717/peerj.175.
913 Plos-One - Data Availability. *Available at https://journals.plos.org/plosone/s/data-availability*
914 (accessed February 3, 2020).

915 Porto WF, Pires AS, Franco OL. 2017. Computational tools for exploring sequence databases
916 as a resource for antimicrobial peptides. *Biotechnology Advances* 35:337–349. DOI:
917 10.1016/j.biotechadv.2017.02.001.
918 Posch L, Panahiazar M, Dumontier M, Gevaert O. 2016. Predicting structured metadata from
919 unstructured metadata. *Database* 2016:baw080. DOI: 10.1093/database/baw080.
920 Pound MP, Atkinson JA, Townsend AJ, Wilson MH, Griffiths M, Jackson AS, Bulat A,
921 Tzimiropoulos G, Wells DM, Murchie EH, Pridmore TP, French AP. 2017. Deep machine
922 learning provides state-of-the-art performance in image-based plant phenotyping.
923 *GigaScience* 6. DOI: 10.1093/gigascience/gix083.
924 Protein Data Bank in Europe - Logo. *Available at https://www.ebi.ac.uk/pdbe/about/logo*
925 Pucker B, Brockington S. 2018. Genome-wide analyses supported by RNA-Seq reveal non-
926 canonical splice sites in plant genomes. *BMC Genomics* 19:980.
927 Pucker B, Feng T, Brockington SF. 2019. *Next generation sequencing to investigate genomic*
928 *diversity in Caryophyllales*. Genomics. DOI: 10.1101/646133.
929 Pucker B, Holtgräwe D, Weisshaar B. 2017. Consideration of non-canonical splice sites
930 improves gene prediction on the Arabidopsis thaliana Niederzenz-1 genome sequence.
931 *BMC Research Notes* 10:667. DOI: 10.1186/s13104-017-2985-y.
932 Raju HB, Tsinoremas NF, Capobianco E. 2016. Emerging Putative Associations between Non-
933 Coding RNAs and Protein-Coding Genes in Neuropathic Pain: Added Value from
934 Reusing Microarray Data. *Frontiers in Neurology* 7. DOI: 10.3389/fneur.2016.00168.
935 Resnik DB. 2007. Conflicts of Interest in Scientific Research Related to Regulation or Litigation.
936 *The Journal of Philosophy, Science & Law* 7:1. DOI: 10.5840/jpsl2007722.
937 Rocca-Serra P, Salek RM, Arita M, Correa E, Dayalan S, Gonzalez-Beltran A, Ebbels T,
938 Goodacre R, Hastings J, Haug K, Koulman A, Nikolski M, Oresic M, Sansone S-A,
939 Schober D, Smith J, Steinbeck C, Viant MR, Neumann S. 2016. Data standards can

boost metabolomics research, and if there is a will, there is a way. *Metabolomics* 12:14. DOI: 10.1007/s11306-015-0879-3.

Roche DG, Kruuk LEB, Lanfear R, Binning SA. 2015. Public Data Archiving in Ecology and Evolution: How Well Are We Doing? *PLOS Biology* 13:e1002295. DOI: 10.1371/journal.pbio.1002295.

Rung J, Brazma A. 2013. Reuse of public genome-wide gene expression data. *Nature Reviews Genetics* 14:89–99. DOI: 10.1038/nrg3394.

Safran C. 2017. Update on Data Reuse in Health Care. *Yearbook of Medical Informatics* 26:24–27. DOI: 10.15265/IY-2017-013.

Sandelin A. 2004. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research* 32:91D – 94. DOI: 10.1093/nar/gkh012.

Sayers EW, Agarwala R, Bolton EE, Brister JR, Canese K, Clark K, Connor R, Fiorini N, Funk K, Hefferon T, Holmes JB, Kim S, Kimchi A, Kitts PA, Lathrop S, Lu Z, Madden TL, Marchler-Bauer A, Phan L, Schneider VA, Schoch CL, Pruitt KD, Ostell J. 2019a. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* 47:D23–D28. DOI: 10.1093/nar/gky1069.

Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I. 2019b. GenBank. *Nucleic Acids Research* 47:D94–D99. DOI: 10.1093/nar/gky989.

Schaefer RJ, Michno J-M, Jeffers J, Hoekenga O, Dilkes B, Baxter I, Myers CL. 2018. Integrating Coexpression Networks with GWAS to Prioritize Causal Genes in Maize. *The Plant Cell* 30:2922–2942. DOI: 10.1105/tpc.18.00299.

Schilbert HM, Pellegrinelli V, Rodriguez-Cuenca S, Vidal-Puig A, Pucker B. 2018. *Harnessing natural diversity to identify key amino acid residues in prolidase*. Evolutionary Biology. DOI: 10.1101/423475.

Schmidt MH-W, Vogel A, Denton AK, Istace B, Wormit A, van de Geest H, Bolger ME, Alseekh S, Maß J, Pfaff C, Schurr U, Chetelat R, Maumus F, Aury J-M, Koren S, Fernie AR,

Zamir D, Bolger AM, Usadel B. 2017. De Novo Assembly of a New *Solanum pennellii* Accession Using Nanopore Sequencing. *The Plant Cell* 29:2336–2348. DOI: 10.1105/tpc.17.00521.

Schmieder R, Edwards R. 2011. Fast Identification and Removal of Sequence Contamination from Genomic and Metagenomic Datasets. *PLoS ONE* 6:e17288. DOI: 10.1371/journal.pone.0017288.

Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk. 2015. Washington, D.C.: National Academies Press. DOI: 10.17226/18998.

Solares EA, Chakraborty M, Miller DE, Kalsow S, Hall K, Perera AG, Emerson JJ, Hawley RS. 2018. Rapid Low-Cost Assembly of the *Drosophila melanogaster* Reference Genome Using Low-Coverage, Long-Read Sequencing. *G3: Genes|Genomes|Genetics* 8:3143–3154. DOI: 10.1534/g3.118.200162.

Soranno PA, Bissell EG, Cheruvelil KS, Christel ST, Collins SM, Fergus CE, Filstrup CT, Lapierre J-F, Lottig NR, Oliver SK, Scott CE, Smith NJ, Stopyak S, Yuan S, Bremigan MT, Downing JA, Gries C, Henry EN, Skaff NK, Stanley EH, Stow CA, Tan P-N, Wagner T, Webster KE. 2015. Building a multi-scaled geospatial temporal ecology database from disparate data sources: fostering open science and data reuse. *GigaScience* 4:28. DOI: 10.1186/s13742-015-0067-4.

Spertus JA. 2012. The Double-Edged Sword of Open Access to Research Data. *Circulation: Cardiovascular Quality and Outcomes* 5:143–144. DOI: 10.1161/CIRCOUTCOMES.112.965814.

Strong MJ, Xu G, Morici L, Splinter Bon-Durant S, Baddoo M, Lin Z, Fewell C, Taylor CM, Flemington EK. 2014. Microbial Contamination in Next Generation Sequencing: Implications for Sequence-Based Analysis of Clinical Samples. *PLoS Pathogens* 10:e1004437. DOI: 10.1371/journal.ppat.1004437.

Subramanian SL, Kitchen RR, Alexander R, Carter BS, Cheung K-H, Laurent LC, Pico A, Roberts LR, Roth ME, Rozowsky JS, Su AI, Gerstein MB, Milosavljevic A. 2015. Integration of extracellular RNA profiling data using metadata, biomedical ontologies and Linked Data technologies. *Journal of Extracellular Vesicles* 4:27497. DOI: 10.3402/jev.v4.27497.

Tenopir C, Rice NM, Allard S, Baird L, Borycz J, Christian L, Grant B, Olendorf R, Sandusky RJ. 2020. Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide. *PLOS ONE* 15:e0229003. DOI: 10.1371/journal.pone.0229003.

Testa AC, Hane JK, Ellwood SR, Oliver RP. 2015. CodingQuarry: highly accurate hidden Markov model gene prediction in fungal genomes using RNA-seq transcripts. *BMC Genomics* 16:170. DOI: 10.1186/s12864-015-1344-4.

The parasite awards - Celebrating rigorous secondary data analysis. *Available at https://researchparasite.com/*

The UniProt Consortium. 2019. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research* 47:D506–D515. DOI: 10.1093/nar/gky1049.

Toubiana D, Puzis R, Wen L, Sikron N, Kurmanbayeva A, Soltabayeva A, del Mar Rubio Wilhelmi M, Sade N, Fait A, Sagi M, Blumwald E, Elovici Y. 2019. Combined network analysis and machine learning allows the prediction of metabolic pathways from tomato metabolomics data. *Communications Biology* 2:214. DOI: 10.1038/s42003-019-0440-4.

Ubbens J, Cieslak M, Prusinkiewicz P, Stavness I. 2018. The use of plant models in deep learning: an application to leaf counting in rosette plants. *Plant Methods* 14:6. DOI: 10.1186/s13007-018-0273-z.

Vasilevsky NA, Minnier J, Haendel MA, Champieux RE. 2017. Reproducible and reusable research: are journal data sharing policies meeting the mark? *PeerJ* 5:e3208. DOI: 10.7717/peerj.3208.

Wade TD. 2014. Refining gold from existing data: *Current Opinion in Allergy and Clinical Immunology* 14:181–185. DOI: 10.1097/ACI.0000000000000051.

Wan X, Pavlidis P. 2007. Sharing and Reusing Gene Expression Profiling Data in Neuroscience. *Neuroinformatics* 5:161–175. DOI: 10.1007/s12021-007-0012-5.

Wang C, Shi H, Chen L, Li X, Cao G, Hu X. 2019. Identification of Key lncRNAs Associated With Atherosclerosis Progression Based on Public Datasets. *Frontiers in Genetics* 10:123. DOI: 10.3389/fgene.2019.00123.

van Wijk KJ, Friso G, Walther D, Schulze WX. 2014. Meta-Analysis of *Arabidopsis thaliana* Phospho-Proteomics Data Reveals Compartmentalization of Phosphorylation Motifs. *The Plant Cell* 26:2367–2389. DOI: 10.1105/tpc.114.125815.

Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3:160018. DOI: 10.1038/sdata.2016.18.

Winter D, Vinegar B, Nahal H, Ammar R, Wilson GV, Provart NJ. 2007. An "Electronic Fluorescent Pictograph" Browser for Exploring and Analyzing Large-Scale Biological Data Sets. *PLoS ONE* 2:e718. DOI: 10.1371/journal.pone.0000718.

Wooley JC, Lin H, National Research Council (U.S.), Committee on Frontiers at the Interface of Computing and Biology. 2005. *Catalyzing inquiry at the interface of computing and biology*. Washington, D.C.: National Academies Press.

Yu H, Dai Z. 2020. SANPolyA: a deep learning method for identifying Poly(A) signals. *Bioinformatics*:btz970. DOI: 10.1093/bioinformatics/btz970.

Zhang H. 2016. Overview of Sequence Data Formats. In: Mathé E, Davis S eds. *Statistical Genomics*. New York, NY: Springer New York, 3–17. DOI: 10.1007/978-1-4939-3578-9_1.

Zhang N, Zhao B, Fan Z, Yang D, Guo X, Wu Q, Yu B, Zhou S, Wang H. 2020. Systematic identification of genes associated with plant growth–defense tradeoffs under JA signaling in Arabidopsis. *Planta* 251:43. DOI: 10.1007/s00425-019-03335-8.

Zizka A, Silvestro D, Andermann T, Azevedo J, Duarte Ritter C, Edler D, Farooq H, Herdean A, Ariza M, Scharn R, Svantesson S, Wengström N, Zizka V, Antonelli A. 2019. CoordinateCleaner : Standardized cleaning of occurrence records from biological collection databases. *Methods in Ecology and Evolution* 10:744–751. DOI: 10.1111/2041-210X.13152.

Table 1(on next page)

Examples of dataset reuse for a novel purpose with the limitations/risks associated with each method.

Table 1: **Examples of dataset reuse for a novel purpose with the limitations/risks associated with each method.**

Examples	Limitations / risks
Genome	
Assembly of new genome sequences, e.g. organellar genome sequences, based on public datasets (Dierckxsens, Mardulyn & Smits, 2016)	Potential contaminations, e.g. of non-target organisms, are unknown. Only the submitter of the original reads can submit the assembly. There are several cases of contamination in published datasets as well as methods for the identification of such contaminations (Longo, O'Neill & O'Neill, 2011; Merchant, Wood & Salzberg, 2014; Strong et al., 2014; Delmont & Eren, 2016; Kryukov & Imanishi, 2016). One study found possible human DNA contamination in 72 % of the analysed (n=202) previously published metagenomes (Schmieder & Edwards, 2011).
Motif identification, e.g. deep learning method for identifying Poly(A) signals (Yu & Dai, 2020)	A large and suitable training set is required. A prediction accuracy of more than 90% can be achieved, but this highly depends on the context of the respective analysis.
Pangenomic analysis, e.g. for bread wheat (Montenegro et al., 2017)	Assembly quality might differ between different studies due to factors like e.g. coverage. Samples with 10x coverage have an assembly efficiency of 81% using the IDBA-UD assembler (Peng et al., 2012; Montenegro et al., 2017), while high continuity long-read assemblies require at least 20-50x coverage (Lu, Giordano & Ning, 2016; Koren et al., 2017; Schmidt et al., 2017; Solares et al., 2018).
GWAS to associate variants (QTLs, SNPs) with traits, e.g. single-plant GWAS for identification of plant height candidate SNPs (Gyawali et al., 2019)	A large number of false positives requires large datasets, their sharing and compulsory replication (Marigorta et al., 2018). One possibility to check for sufficient sample size in e.g. genetic association studies is the random division of the study population by 2 and the requirement that any results have to be detected in both subsets (Hirschhorn et al., 2002).
Transcriptome	
Co-expression analysis to find connected genes, e.g. identification of long non-coding RNAs associated with atherosclerosis progression (Wang et al., 2019); Co-expression networks, e.g. related to bamboo development using public RNA-Seq data (Ma et al., 2018) or related to cellulose synthesis using public microarray data (Persson et al., 2005); Construction of regulatory networks using co-expression data, e.g. co-expression network analysis to reveal genes in growth-defence trade-offs under JA signalling (Zhang et al., 2020)	Batch effects might be possible if large sample groups come from the same source. Ideally, networks for different samples should be incorporated as there is high variation between co-expression networks with different samples (Ma et al., 2018).

Gene expression analysis to find/identify best gene candidate for cloning (and select the right tissue), e.g. integration with GWAS to identify causal genes in maize (Schaefer et al., 2018)	Batch effects if large sample groups come from the same source. The success depends on the gene expression data context.
Identification of qRT-PCR reference genes (Kwon et al., 2009; Cheng et al., 2011; Hruz et al., 2011)	Batch effects if large sample groups come from the same source. E.g. for formalin-fixed paraffin-embedded tissues, accurate normalization requires two to four endogenous reference genes (Kwon et al., 2009). Further, for the normalization of RT-qPCR data condition-specific reference genes should be used (Hruz et al., 2011).
Gene prediction via analysis of RNASeq data (Pucker, Feng & Brockington, 2019) and e.g. GeMoMa is using this heavily (Keilwagen, Hartung & Grau, 2019)	Batch effects if large sample groups come from the same source. In regions without RNA-Seq data, *ab initio* prediction is required (Testa et al., 2015).
Gene expression web sites, e.g. the eGFP browser (Winter et al., 2007)	Only genes in the annotation included. Only based on the available structural annotation thus alternative transcripts would be missed.
Analysis of non-canonical splice sites based on genome sequences, annotations, and RNA-Seq datasets (Pucker & Brockington, 2018; Frey & Pucker, 2020)	Batch effects if large sample groups come from the same source and annotation errors will impact analysis results. One example is the high number of annotated CT-AC splice site combinations in fungal genome sequences which are probably caused by a systematic error in the assignment of RNA-Seq reads to DNA strands (Frey & Pucker, 2020).
Extraction of new sequences for phylogenetic analysis (Schilbert et al., 2018)	Reliability of source is crucial; transcriptome assemblies are inherently incomplete as not all genes are expressed at the same time.
Reuse of microarray data for meta-analyses including the investigation of non-coding RNAs (Raju, Tsinoremas & Capobianco, 2016)	Microarray technology is not comprehensive e.g. in comparison to RNA-Seq; the study is limited to known non-coding RNAs and is not suitable for the detection of new non-coding RNAs (Raju, Tsinoremas & Capobianco, 2016).
Gene expression analysis based on microarray data (Wan & Pavlidis, 2007)	Submitters might fail to indicate technical replicates. Sequences of probes are sometimes unknown.
Investigation of the underlying mechanisms of homeostatic eosinophil gene expression (Grace et al., 2018)	Variation in the methods for RNA-Seq library construction likely contributes to part of the detected differential expression (Grace et al., 2018).
Proteome	
Identification of antimicrobial peptides (Porto, Pires & Franco, 2017)	Prediction, correct modelling and structural analysis are not completely accurate due to e.g. the presence of precursors. Validation is required.
Phospho-proteomics, e.g. compartmentalisation of phosphorylation motifs (van Wijk et al., 2014)	Meta-analysis allows extrapolation only for highly specific conditions due to numerous different experimental conditions in the used studies. E.g. in seedling/rosette samples, plastid proteins might be (50%) overrepresented /mitochondrial and secretory proteins might be (10%) underrepresented in comparison to cell cultures/root/pollen/seed samples (van Wijk

	et al., 2014).
Metabolome	
Metabolic modelling (Brinkrolf et al., 2018)	Precise conditions of experiments are different between labs and measurement biases possible.
Combining network analysis and machine learning to predict metabolic pathways (Toubiana et al., 2019)	Cannot be used to predict catalytic activity, but only to predict pathways. Stabilized correlation and reduced error rate can be achieved using a large sample size. Large sample sizes can be exploitation of the natural variability of mapping populations or collections of different varieties or cultivars (Toubiana et al., 2019).
Phenotype	
Deep learning methods for image-based phenotyping, e.g. leaf counting (Ubbens et al., 2018) or root and shoot feature identification (Pound et al., 2017)	Large datasets are required. For vision-based deep learning analyses, tens of thousands to tens of millions of images might be required (Ubbens et al., 2018).
Ecology	
Modelling and prediction of the variability of biodiversity to explain ecological and evolutionary mechanisms (Jetz & Fine, 2012)	Choice and accuracy of predictor variables are crucial for the model: Challenges regarding the definition of exact region boundaries, climate reconstruction and comparability across clades remain; (Jetz & Fine, 2012).
Ecosystem modelling, e.g. reuse of model code/reuse of eutrophication models for studying climate change (Mooij et al., 2010)	Partly overly simplified models: the validity of outcomes must be tested. Observations of species are sometimes placed at institutes of districts/regions.
Database of lake water quality (Soranno et al., 2015)	Underlying data sets can be incomplete (e.g. missing lake coordinates)

3
4

Table 2(on next page)

Checklist for the selection of appropriate datasets.

For each possible criterium, several questions to consider and suggestions for the reuse of public data are mentioned.

Table 2: **Checklist for the selection of appropriate datasets.** For each possible criterium, several questions to consider and suggestions for the reuse of public data are mentioned.

Criteria	Question(s) to consider	Suggestions/Suitable controls
Integrity of the source	Is the source/submitter associated with data fabrication/plagiarism?	Check potential conflicts of interests/funding (useful resources: NSF conflict of interest guidelines (https://www.nsf.gov/pubs/policydocs/pappguide/nsf16001/aag_4.jsp), examples and strategies of dealing with conflicts of interest (Resnik, 2007)
Biases	How was the data generated? Are there batch effects?	Comparison of random samples from the dataset with replacement (bootstrapping) to reveal any bias/errors; Principal component analyses
Missing meta-information (sparsity)	Do you have all relevant information (e.g., information about the biological material)?	Possibility to contact the authors; infer metadata from data sets e.g. identify RNA-Seq tissue based on gene expression patterns of marker genes
Integration of datasets from different sources	Is the data comparable? Are the methods used for data collection/generation comparable?	Check relevant parameters: For sequencing reads: same (NGS) technology/platform and same sequencing chemistry (differences between versions of sequencing chemistry are possible) For assemblies: same type/version of bioinformatic tools and a full list of parameters
Quality issues	Is the quality high enough to reach your goals (e.g. looking at gene expression differences between strains or making evolutionary trees)? Are there any scores/hints available to check the quality of the dataset?	Check relevant parameters: For sequencing reads: Phred scores, length, paired-end status For assemblies: continuity, contig/scaffold N50, auN (Li, 2020)
Copyright/ Legal issues	Are there any restrictions for reuse and publication of the data, especially due to the Nagoya protocol?	Check copyright information/licenses when selecting data prior to the actual reuse

Figure 1

The evolution of data sharing behaviour.

(1) Technical progress makes global sharing of large data-sets possible, (2) increased accessibility to required technology makes it widely available, (3) obligations and benefits for researchers establish sharing behaviour, (4) the size of datasets increases and makes them attractive, (5) reuse develops over time – which results in a positive feedback loop and a habit to share all data.

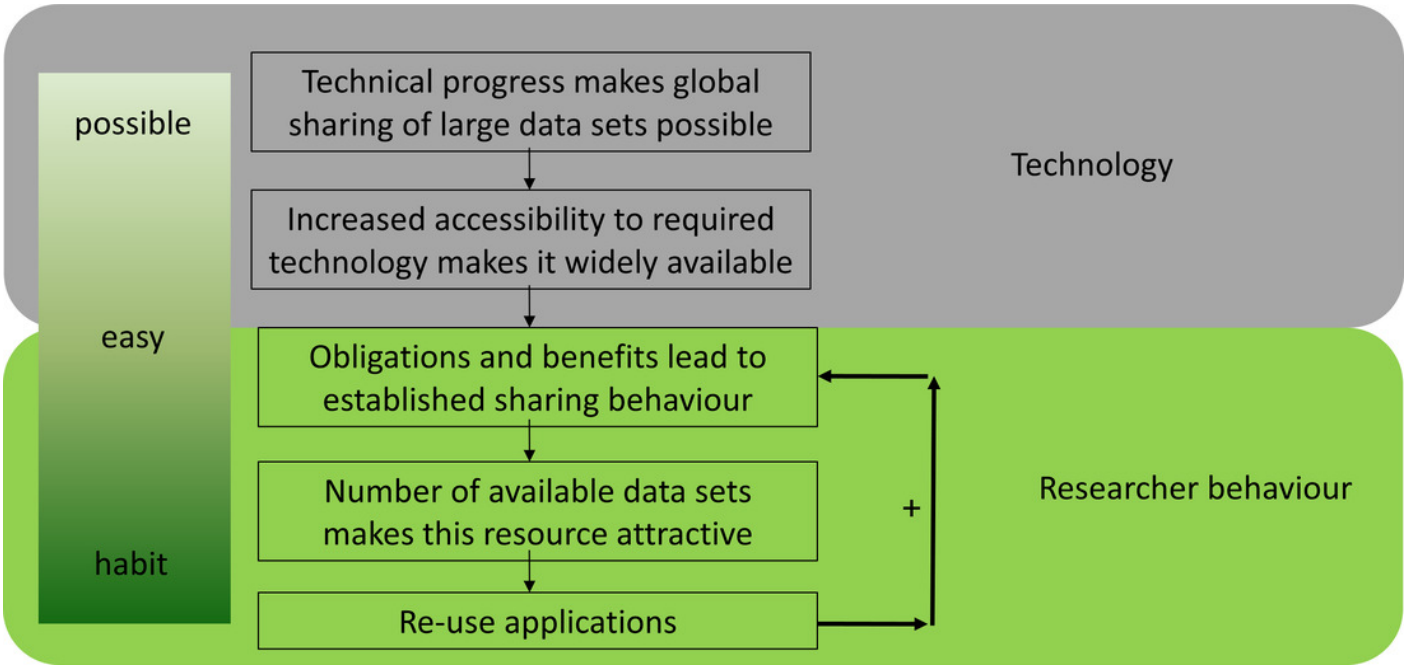


Figure 2

Types of reusable data classified into primary and derived/secondary data.

Specific examples for each data type are provided in parentheses. The data classification is based on: (Wooley et al., 2005). (Sources of the pictures: ("Protein Data Bank in Europe - Logo"; Pucker, Holtgräwe & Weisshaar, 2017; Schilbert et al., 2018; Mitchell et al., 2019; Frey & Pucker, 2020)).

Primary data

Sequences: text strings describing sequential bases, including gaps in data and their length
(single gene/whole genome sequences, amino acid sequences)

```
>seq1
ATCGTTTAGCTAGACCTGATG
ATCCGATCGATTACGTG
>seq2
GACACGATCGTCAGAAATGCA
GTC
>seq3
ACGACAAATCATCTCC
```

```
>seq1
MYVRANQEFFK
>seq2
WTSMADCHLKV
>seq3
MGPKLHIGRQEFKLIHYWN
NNG
```

Metadata: information about the acquisition, processing and presentation of data
(methodology, sample size and origin)

Graphs: graphical representation indicating relationship
(pathway data, genetic maps, plasmid maps)

Annotations (gene annotation, gene model)

Algorithms and Software (code, bioinformatics pipelines)

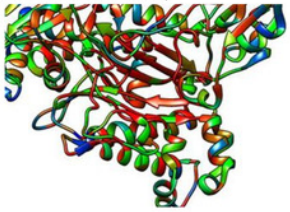
```
for ID in transcripts_per_genes[ gene ]:
        counter = 0
        for element in transcripts[ ID ]:
                if element['type'] == "CDS":
                        counter += element['end']-
element['start']
        CDS_len_per_transcript.append( { 'id': ID, 'len': counter } )
repr_trans = sorted( CDS_len_per_transcript, key=itemgetter('len') )[-1]
repr_transcripts.update( { repr_trans['id']: transcripts[ repr_trans['id'] ] } )
```

Hypotheses, evidence and prose (data interpretation)

Plots and Images: natural, artificial and stylised imagery
(micrographs, radiographs, diagrams, microscopy of biological material, images of geographical regions)

Measurement parameters
(enzyme affinity/speed, chromatography results, mass spectra, geo data/coordinates)

Geometric information/Molecular structure data
(secondary and tertiary protein structures)



Patterns
(sequence motifs, regulatory sequences, expression profiles)

Primary databases: direct submission of experimentally-derived data from researchers; archival database
(DNA, RNA, protein, expression, disease, organism-specific databases)

PDBe Protein Data Bank in Europe

Derived data

Publications (text mining)

Meta-analyses

Derived databases: results of analysis, literature search and interpretation, curated database
(DNA, RNA, protein, expression, disease, organism-specific, phenotypic databases)

InterPro

Figure 3

The increasing size of selected databases over time.

The number of bases/sequence entries in GenBank, the Sequence Read Archive (SRA) and UniProtKB/TrEMBL are shown, respectively. Note the logarithmic scale of the y-axes. The drop of sequence entries in UniProtKB/TrEMBL (in 2015) can be explained by the removal of duplicates.

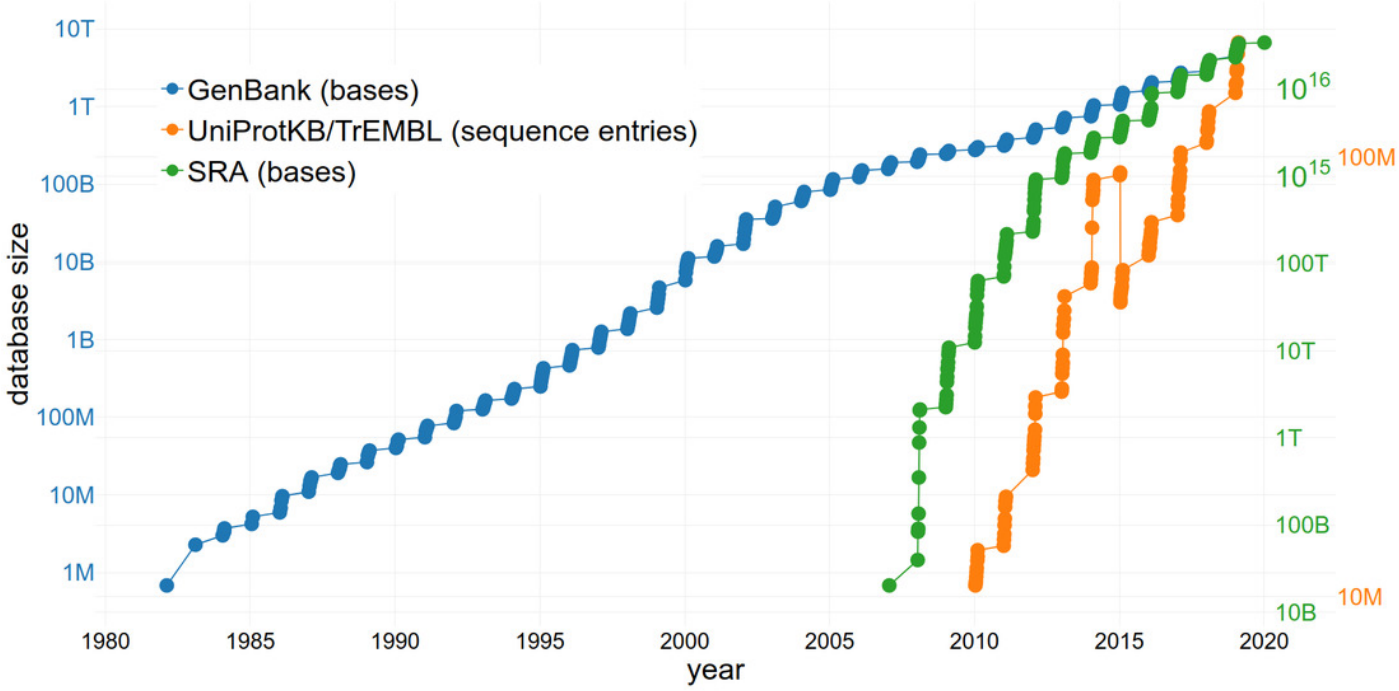


Figure 4

Advantages and limitations of data reuse.

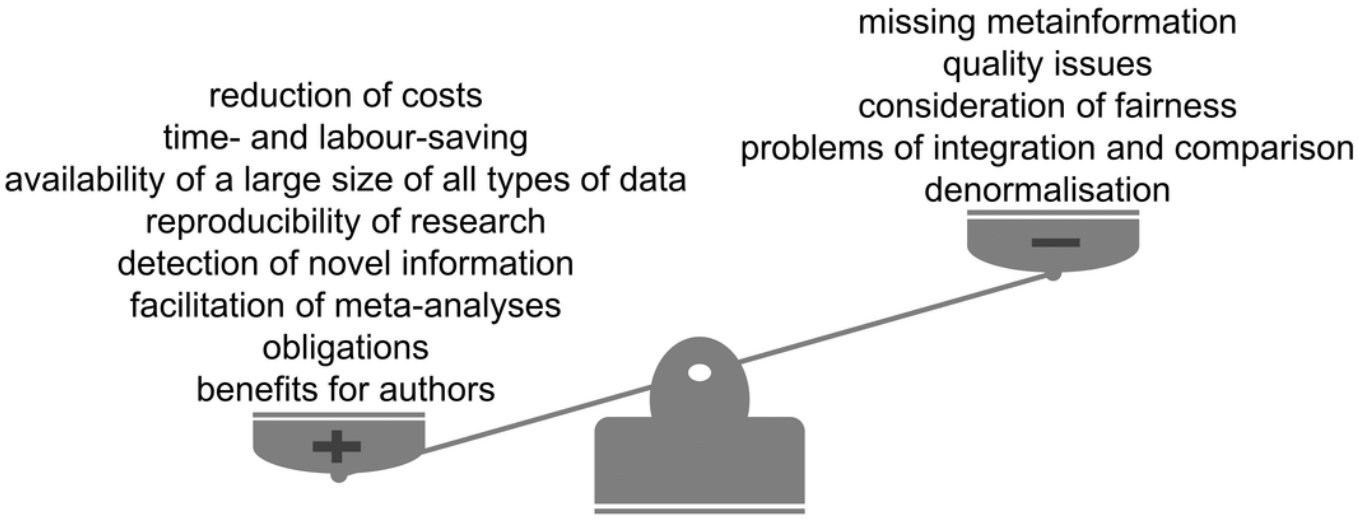


Figure 5

Summary of outstanding questions and challenges.

- Should there be an obligation/encouragement to reuse public datasets instead of producing new ones? At which level could this be reinforced?
- Should journals require the release of all acquired data (except patient information) as a prerequisite for publication?
- How many datasets are just lost because scientists/students are moving on to other projects without publishing?
- How can the quality of the datasets be ensured?
- Who is responsible for the management of the rapidly growing databases and how can sufficient storage space/funding be realized to ensure long-term sustainability?
- Is there a suitable way for both, scientists and databases, to provide the metainformation needed for efficient and correct data reuse?