Genomics Proteomics Bioinformatics 18 (2020) 91–103

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

PERSPECTIVE

Quality Matters: Biocuration Experts on the Impact of Duplication and Other Data Quality Issues in Biological Databases

Qingyu Chen [1,*], Ramona Britto [2], Ivan Erill [3], Constance J. Jeffery [4], Arthur Liberzon [5], Michele Magrane [2], Jun-ichi Onami [6,7], Marc Robinson-Rechavi [8,9], Jana Sponarova [10], Justin Zobel [1,*], Karin Verspoor [1,*]

[1] *School of Computing and Information Systems, University of Melbourne, Melbourne, VIC 3010, Australia*
[2] *European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK*
[3] *Department of Biological Sciences, University of Maryland, Baltimore, MD 21250, USA*
[4] *Department of Biological Sciences, University of Illinois at Chicago, Chicago, IL 60607, USA*
[5] *Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA*
[6] *Japan Science and Technology Agency, National Bioscience Database Center, Tokyo 102-8666, Japan*
[7] *National Institute of Health Sciences, Tokyo 158-8501, Japan*
[8] *Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland*
[9] *Department of Ecology and Evolution, University of Lausanne, CH-1015 Lausanne, Switzerland*
[10] *Nebion AG, 8048 Zurich, Switzerland*

Received 8 December 2017; revised 24 October 2018; accepted 14 December 2018
Available online 9 July 2020

Handled by Zhang Zhang

Introduction

Biological databases represent an extraordinary collective volume of work. Diligently built up over decades and comprising many millions of contributions from the biomedical research community, biological databases provide worldwide access to a massive number of records (also known as *entries*) [1]. Starting from individual laboratories, genomes are sequenced, assembled, annotated, and ultimately submitted to primary nucleotide databases such as GenBank [2], European Nucleotide Archive (ENA) [3], and DNA Data Bank of Japan (DDBJ) [4] (collectively known as the International Nucleotide Sequence Database Collaboration, INSDC). Protein records, which are the translations of these nucleotide records, are deposited into central protein databases such as the UniProt KnowledgeBase (UniProtKB) [5] and the Protein Data Bank (PDB) [6]. Sequence records are further accumulated into different databases for more specialized purposes: RFam [7] and PFam [8] for RNA and protein families, respectively; DictyBase [9] and PomBase [10] for model organisms; as well as ArrayExpress [11] and Gene Expression Omnibus (GEO) [12] for gene expression profiles. These databases are selected as

* Corresponding authors.
E-mail: qingyu.chen@unimelb.edu.au (Chen Q), jzobel@unimelb.edu.au (Zobel J), karin.verspoor@unimelb.edu.au (Verspoor K).

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

https://doi.org/10.1016/j.gpb.2018.11.006

examples; the list is not intended to be exhaustive. However, they are representative of biological databases that have been named in the "golden set" of the 24th *Nucleic Acids Research* database issue (in 2016). The introduction of that issue highlights the databases that "*consistently served as authoritative, comprehensive, and convenient data resources widely used by the entire community and offer some lessons on what makes a successful database*" [13]. In addition, the associated information about sequences is also propagated into non-sequence databases, such as PubMed (https://www.ncbi.nlm.nih.gov/pubmed/) for scientific literature or Gene Ontology (GO) [14] for function annotations. These databases in turn benefit individual studies, many of which use these publicly available records as the basis for their own research.

Inevitably, given the scale of these databases, some submitted records are redundant [15], inconsistent [16], inaccurate [17], incomplete [18], or outdated [19]. Such quality issues can be addressed by manual curation, with the support of automatic tools, and by processes such as reporting of the issues by contributors detecting mistakes. Biocuration plays a vital role in biological database curation [20]. It de-duplicates database records [21], resolves inconsistencies [22], fixes errors [17], and resolves incomplete and outdated annotations [23]. Such curated records are typically of high quality and represent the latest scientific and medical knowledge. However, the volume of data prohibits exhaustive curation, and some records with quality issues remain undetected.

In our previous studies, we (Chen, Verspoor, and Zobel) explored a particular form of quality issue, which we characterized as *duplication* [24,25]. As described in these studies, duplicates are characterized in different ways in different contexts, but they can be broadly categorized as *redundancies* or *inconsistencies*. The perception of a pair of records as duplicates depends on the task. As we wrote in a previous study, "*a pragmatic definition for duplication is that a pair of records A and B are duplicates if the presence of A means that B is not required, that is, B is redundant in the context of a specific task or is superseded by A*." [24]. Many such duplicates have been identified through curation, but the prevalence of undetected duplicates remains unknown, as is the accuracy and sensitivity of automated tools for duplicate or redundancy detection. Other studies have explored the detection of duplicates but often under assumptions that limit the impact. For example, some researchers have assumed that similarity of genetic sequence is the sole indicator of redundancy, whereas in practice, some highly similar sequences may represent distinct information and some rather different sequences may in fact represent duplicates [26]. The notion and impacts of duplication are detailed in the next section.

In this study, the primary focus is to explore the characteristics, impacts, and solutions to duplication in biological databases; and the secondary focus is to further investigate other quality issues. We present and consolidate the opinions of more than 20 experts and practitioners on the topic of duplication and other data quality issues via a questionnaire-based survey. To address different quality issues, we introduce biocuration as a key mechanism for ensuring the quality of biological databases. To our knowledge, there is no one-size-fits-all solution even to a single quality issue [27]. We thus explain the complete UniProtKB/Swiss-Prot curation process, via a descriptive report and an interview with its curation team leader, which provides a reference solution to different quality issues. Overall, the observations on duplication and other data quality issues highlight the significance of biocuration in data resources, but a broader community effort is needed to provide adequate support to facilitate thorough biocuration.

The notion and impact of duplication

Our focus is on database records, that is, entries in structured databases, but not on biological processes, such as gene duplication. Superficially, the question of what constitutes an *exact duplicate* in this context can seem obvious: two records that are exactly identical in both data (*e.g.*, sequence) and annotation (*e.g.*, metadata including species and strain of origin) are duplicates. However, the notion of duplication varies. We demonstrate a generic biological data analysis pipeline involving biological databases and illustrate different notions of duplication.

Figure 1 shows the pipeline. We explain the three stages of the pipeline using the databases managed by the UniProt Consortium (http://www.uniprot.org/) as examples.

At "pre-database" stage, records from various sources are submitted to databases. For instance, UniProt protein records come from translations of primary INSDC nucleotide records (directly submitted by researchers), direct protein sequencing, gene prediction, and other sources (http://www.uniprot.org/help/sequence_origin).

The "within database" stage is for database curation, search, and visualization. Records are annotated in this stage, automatically (UniProtKB/Translated European Molecular Biology Laboratory [TrEMBL]) or through curation (UniProtKB/Swiss-Prot). Biocuration plays a vital role at this stage. For instance, UniProtKB/Swiss-Prot manual curation not only merges records and documents the discrepancies of the merged records (*e.g.*, sequence differences), but also annotates the records with biological knowledge drawn from the literature [28]. Additionally, the databases need to manage the records for search and visualization purposes [29]. During this stage, UniProtKB undertakes extensive cross-referencing by linking hundreds of databases to provide centralized knowledge and resolve ambiguities [30].

The "post-database" stage is for record download, analysis, and inference. Records are downloaded and analyzed for different purposes. For instance, both UniProtKB records and services have been extensively used in the research areas of biochemistry, molecular biology, biotechnology, and computational biology, according to citation patterns [31]. The findings of studies may in turn contribute to new sources.

Duplication occurs in all of these stages, but its relevance varies. Continuing with the UniProtKB example, the first stage primarily concerns *entity duplicates* (often referred to as *true duplicates*): records that correspond to the same biological entities regardless of whether there are differences in the content of the database records. Merging such records into a single entry is the first step in UniProtKB/Swiss-Prot manual curation [28]. The second stage primarily concerns *near-identical duplicates* (often referred to as *redundant records*): the records may not refer to the same entities, but nevertheless have a high similarity. UniProtKB has found that these records lead to uninformative BLAST search results (http://www.uniprot.org/help/proteome_redundancy). The third stage primarily concerns *study-dependent duplicates*: studies may fur-

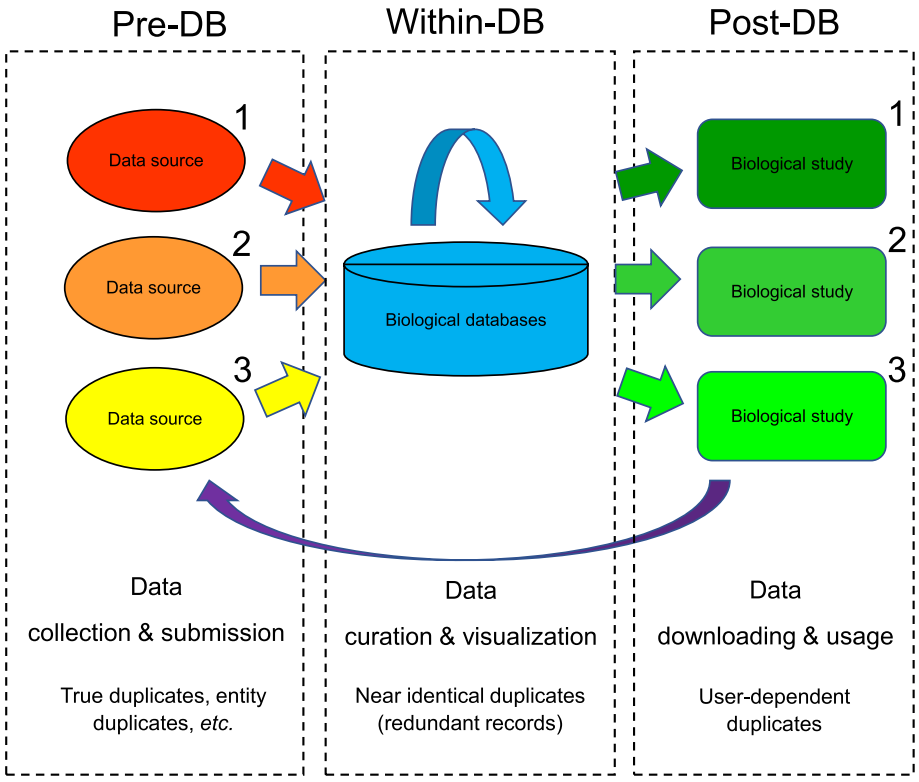

Figure 1 Biological analysis pipeline
Three stages of a biological analysis pipeline, heavily involving biological databases, are presented. Pre-DB: the data collection and submission stage, where entity duplicates often matter. Within-DB: the data curation and visualization stage, where near-identical duplicates often matter. Post-DB: the data downloading and usage stage, where the definition of duplicates is use case dependent. DB: database.

ther de-duplicate sets of records for their own purposes. For instance, studies on secondary protein structure prediction may further remove protein sequences at a 75% sequence similarity threshold [32]. This clearly shows that the notion of duplication varies and in general has two characteristics: *redundancy* and *inconsistency*. Thus, it is critical to understand their characteristics, impacts, and solutions.

Moreover, we found numerous discussions of duplicates in the previous literature. In as early as 1996, Korning et al. [33] observed duplicates from the GenBank *Arabidopsis thaliana* dataset when curating these records. The duplicates were of two main types: the same genes that were submitted twice (either by the same or different submitters) and different genes from the same gene family that were similar enough so that only one was retained. Similar cases were also reported by different groups [21,34–37]. Recently, the most significant case was the duplication in UniProtKB/TrEMBL [15]: in 2016, UniProtKB removed 46.9 million records corresponding to duplicate proteomes (for example, more than 5.9 million of these records belong to 1692 strains of *Mycobacterium tuberculosis*). They identified duplicate proteome records based on three criteria: belonging to the same organisms; sequence identity of greater than 90%; and proteome ranks designed by biocurators (such as whether they are reference proteomes and their annotation level).

As this history shows, investigation of duplication has persisted for at least 20 years. Considering the type of duplicates, as the discussion above illustrates, duplication appears to be richer and more diverse than was originally described (we again note the definition of "duplication" we are following in this paper, which includes the concept of redundancy). This motivates continued investigation of duplication.

An underlying question is: does duplication have positive or negative impact? There has been relatively little investigation of the impact of duplication, but there are some observations in the literature: (1)"*The problem of duplicates is also existent in genome data, but duplicates are less interfering than in other application domains. Duplicates are often accepted and used for validation of data correctness. In conclusion, existing data cleansing techniques do not and cannot consider the intricacies and semantics of genome data, or they address the wrong problem, namely duplicate elimination.*" [38]; (2)"*Biological data duplicates provide hints of the redundancy in biological datasets . . . but rigorous elimination of data may result in loss of critical information.*" [34]; and (3)"*The bioinformatics data is characterized by enormous diversity matched by high redundancy, across both individual and multiple databases. Enabling interoperability of the data from different sources requires resolution of data disparity and transformation in the common form (data integration), and the removal of redundant data, errors, and dis-*

crepancies (data cleaning)." [39]. Thus, the answers to questions on the impact of duplicates remain unclear. The aforementioned views are inconsistent and also outdated. Answering the question of the impact of duplications requires a more comprehensive and rigorous investigation.

From duplication to other data quality issues

Biological sources suffer from data quality issues other than duplication. The diverse biological data quality issues reported in the literature include inconsistencies (such as conflicting results reported in the literature) [22], inaccuracies (such as erroneous sequence records and wrong gene annotations) [40–42], incompleteness (such as missing exons and incomplete annotations) [38,40], and outdatedness (such as outdated sequence records and annotations) [41]. This shows that although duplication is a primary data quality issue, other quality issues are also of concern. Collectively, there are five primary data quality issues: duplication, inconsistency, inaccuracy, incompleteness, and outdatedness identified in general domains [43]. It is thus also critical to understand what quality issues have been observed and how they impact database stakeholders under the context of biological databases.

Practitioner viewpoint

Survey questions

Studies on data quality broadly take one of three approaches: domain expertise, theoretical, or empirical. The first is an opinion-based approach: accumulating views from (typically a small group of) domain experts [44–46]. For example, one book summarizes opinions from domain experts on elements of spatial data quality [44]. The second is a theory-based approach: inference of potential data quality issues from a generic process of data generation, submission, and usage [47–49]. For example, a data quality framework was developed by inferring the data flow of a system (such as input and output for each process) and estimating the possible related quality issues [47]. The third is an empirically-based approach: analysis of data quality issues in a quantitative manner [50–52]. For example, an empirical investigation on what data quality means to stakeholders was performed via a questionnaire [50]. Each approach has its own strengths and weaknesses; for example, opinion-based studies represent high domain expertise, but may be narrow due to the small group size. In contrast, quantitative surveys have a larger number of participants, but the level of expertise may be relatively lower.

Our approach integrates opinion-based and empirically-based approaches: the study presents opinions from domain experts, but the data was gathered via a questionnaire; the survey questions are provided in File S1. We surveyed 23 practitioners on questions of duplicates and other general data quality issues. These practitioners are from diverse backgrounds (including experimental biology, bioinformatics, and computer science), with a range of affiliation types (such as service providers, universities, or research institutes), but all have domain expertise. These practitioners include senior database staff, project leaders, lab leaders, and biocurators. The publications of the participants are directly relevant to databases, data quality, and curation, as illustrated by some instances [10,15,28,53–69]. They were selected through a personal approach at conferences and in a small number of cases by email; most of the practitioners were not known to the originating authors (Chen, Verspoor, and Zobel) before this study.

The small participant size may mean that we have collected unrepresentative opinions, which is a limitation of the current study. However, the community of biocuration is small and the experience represented by these 23 practitioners is highly relevant. A 2012 survey conducted by the International Society of Biocuration (ISB) included 257 participants [67]. Of these 257 participants, 57% were employed in short-term contracts and only 9% were principal investigators. A similar study initiated by the BioCreative team involved only 30 participants, including all the attendees of the BioCreative conference in 2012 [68]. Therefore, the number of participants in the current study reflects the size of the biocuration community; moreover, the relatively high expertise ensures the validity of the opinions.

The survey asked three primary questions about duplication. (1) *What* are duplicates? We asked practitioners what records they think should be regarded as duplicated. (2) *Why* care about duplicates? We asked practitioners what impact duplicates have. (3) *How* to manage duplicates? We asked practitioners whether and how duplicates should be resolved. The details of questions and their possible responses are provided below.

Defining duplicate records (The "what" question)

We provided five options for experts to select. These include (1) exact duplicate records (two or more records are exactly identical; (2) near-identical duplicates (two or more records are not identical but similar); (3) partial or fragmentary records (one record is a fragment of another); (4) duplicate records with low similarity (records have a relatively low similarity but belong to the same entity); and (5) other types (if practitioners also consider other cases as duplicates).

Respondents were asked to comment on their choices. We also requested them to provide examples to support the choice of options 4 or 5, given that in our review of the literature, we observed that the first three options were prevalent [70,71]. Option 1 refers to exact duplicates; option 2 refers to (highly) similar or redundant records or to some quantitative extent, records share X% similarity; option 3 refers to partial or incomplete records; option 4 refers to entity duplicates that are inconsistent; and the "Other types" option provides capture of remaining types of duplicates.

Quantifying the impacts of duplication (The "why" question)

We asked this question in two steps. The first question is whether respondents believe that duplicates have an impact. The second question is presented only if the answer to the first is yes. This is used to comment on positive and negative impacts. We also ask respondents to explain their opinion or give examples.

Addressing duplication (The "how" question)

We offered three subquestions. (1) Do you believe that duplicate detection is useful or needed? (2) Do you believe that the current duplicate detection methods or software are sufficient to satisfy your requirements? We also ask respondents to

explain what they expect if they select "no." (3) How would you prefer that duplicate records be handled? The suggested options include label and remove duplicates, label and make duplicates obsolete, label but leave duplicates active, and other solutions.

Survey results

The responses are summarized below in the same order as the three primary questions mentioned above. For each question, we detailed the response statistics, summarized the common patterns augmented by detailed responses, and drew conclusions.

Opinion on duplication

The views on *what are duplicates* are summarized in **Figure 2**. Out of 23 practitioners, 21 made a choice by selecting at least one option. Although the other two did not select any options, they think that duplicates have impacts for later questions. Therefore, we do not regard the empty responses as an opinion that duplication does not exist; instead, we simply do not track the response in this case.

The results show that all types of duplicates have been observed by some practitioners, but none is universal. The most common type of duplicates is *similar record*, which was selected by more than half of the respondents, but the other types (*exact duplicates*, *partial records*, and *low similarity duplicates*) were also selected by at least one third of the respondents. We also find that more than 80% of respondents indicated that they observed at least two types of duplicates.

Additionally, recall that existing literature rarely covers the fourth type of duplication, that is, relatively different records that should in fact be considered as duplicates. However, nearly 40% of respondents acknowledge having seen such cases and further point out that identifying them requires considerable manual effort. The following summarizes three primary cases (each identified by a respondent ID, tabulated at the end of this paper).

The first primary case is low similarity duplicates within a single database. Representative comments are "*We have such records in ClinVar* [64]. *We receive independent submissions from groups that define variants with great precision, and groups that define the same variant in the same paper, but describe it imprecisely. Curators have to review the content to determine identity*." [R19] and "*Genomes or proteomes of the same species can often be different enough even they are redundant*." [R24].

The second primary case is low similarity duplicates in databases having cross-references. Representative comments are "*Protein–protein interaction databases: the same publication may be in BioGRID* [72] *annotated at the gene level and in one of the IMEx databases* (*http://www.imexconsortium.org/*) *annotated at the protein level*." [R20] and "*Also secondary databases import data* (*e.g., STRING sticking to the PPI example*) *but will only import a part of what is available*." [R20].

The third primary case is low similarity duplicates in databases having the same kinds of contents. For instance, "*Pathway databases, such as KEGG* (*https://www.genome.jp/kegg/*) *and Reactome* (*https://reactome.org/*), *tend to look at same pathways but are open to curator interpretation and may differ*." [R20].

The views on *why care about duplicates* are summarized in **Figure 3**. All practitioners made a choice. Most (21 out of 23) believe that duplication does matter. Moreover, 19 out of 21 experts weighted the potential impacts of duplicates. Among them, only one respondent believe that the impact is purely positive compared with eight respondents viewing it as solely negative, whereas the remaining 10 respondents think that the impact has both positive and negative sides. We assembled all responses on impacts of duplicates as follows.

Impact on database storage, search, and mapping

Representative comments are (1) "*When duplicates* (*sequence only*) *are in big proportion they will have an impact on sequence search tool like BLAST, when precomputing the database to search against. Then it will affect the statistics on the E-value returned*." [R10]; (2) "*Duplicates in one resource make exact mappings between 2 resources difficult*." [R21] and "*Highly redundant records can result in: increasing bias in statistical analyses; repetitive hits in BLAST searches*." [R24]; and (3) "*Querying datasets with duplicate records impacts the diversity of hits and increase overall noise; we have discussed this in our paper on hallmark signatures*." [56]. [R8].

Impact on meta-analysis in biological studies

Representative comments are (1) "*Duplicate transcriptome records can impact the statistics of metaanalysis*." [R1]; (2) "*Authors often state a fact is correct because it has been*

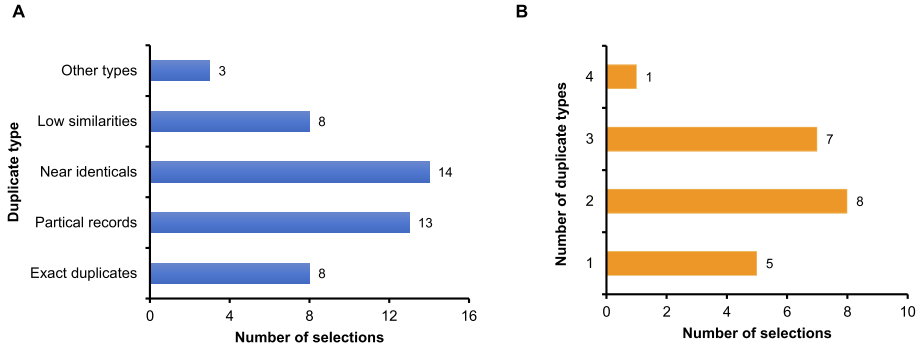

Figure 2 Characteristics of duplicate records

A. Duplicate types and number of participants who selected different duplicate types. **B.** Distribution of participants according to the number of duplicate types they selected. There are 21 participants in total.

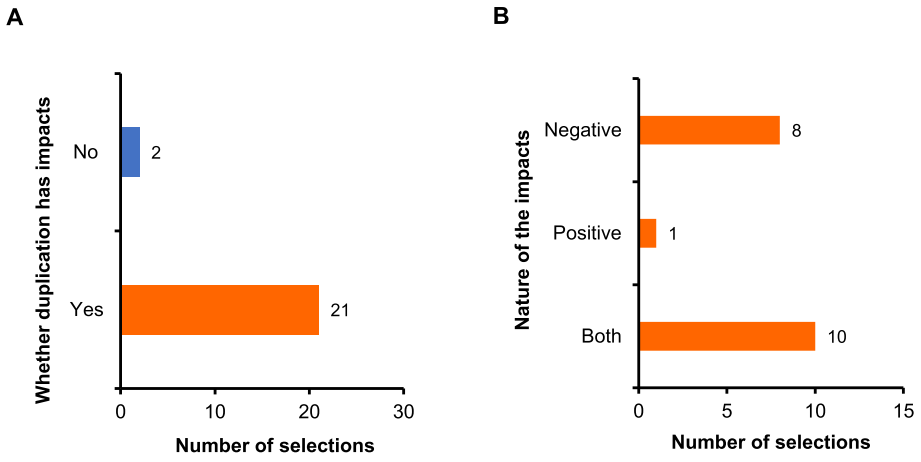

Figure 3 Impacts of duplicate records

A. The number of participants who believed duplication has impacts or not. **B.** A more detailed breakdown by type of impact, for those who believed duplication has impacts.

observed in multiple resources. If the resources are reusing, or recycling the same piece of information,this statement (or statistical measure), is incorrect." [R20] (*note that it has been previously observed that cascading errors may arise due to this type of propagation of information* [73]); and (3) "*Duplicates affect enrichments if duplicate records used in background sets.*" [R21].

Impact on time and resources

Representative comments are (1) "*Archiving and storing duplicated data may just be a waste of resources.*" [R12]; (2) "*Result in time wasted by the researcher.*" [R19]; and (3) "*As a professional curation service; our company suffers from the effects of data duplication daily. Unfortunately there is no prescreening of data done by Biological DBs and thus it is up to us to create methods to identify data duplication before we commit time to curate samples. Unfortunately, with the onset of next generation data, it has become hard to detect duplicate data where the submitter has intentionally rearranged the reads without already committing substantial computational resources in advance.*" [R9].

Impact on users

Representative comments are (1) "*Duplicate records can result in confusion by the novice user. If the duplication is of the 'low similarity' type, information may be misleading.*" [R19] and "*Duplicate gene records may be misinterpreted as species paralogs.*" [R21]; (2) "*When training students, they can get very confused when a protein in a database has multiple entries—which one should they use, for example. Then I would need to compare the different entries and select one for them to use. It would be better if the information in the duplicate entries was combined into one correct and more complete entry.*" [R23]; and (3) "*Near identical duplicate records: two or more records are not strictly identical but very similar and can be considered duplicates; because users don't realize they are the same thing or don't understand the difference between them.*" [R25].

In contrast, practitioners pointed out two primary positive impacts: (1) identified duplicates enrich the information about an entity; for example, "*When you try to look sequence homology across species, it is good to keep duplicates as it allows to build orthologous trees.*" [R10] and "*When they are isoforms of each other so while they are for the same entity, they have distinct biological significance.*" [R25], and (2) identified duplicates verify the correctness as replications; for example, "*On the other hand, if you have many instances of the same data, or near identical data, one could feel more confident on that data point.*" [R12] (note that confidence information ontology can be used to capture "*confidence statement from multiple evidence lines of same type*" [74]) and "*If it is a duplicate record that has arisen from different types of evidence, this could strengthen the claim.*" [R13].

The cases outlined above detail the impact of duplication, and clearly, duplication does matter. The negative impacts are broad, ranging from databases to studies, from research to training, and from curators to students. The potential impacts are severe: valuable search results may be missed, statistical results may be biased, and study interpretations may be misled. Management of duplication is a significant amount of labor.

Our survey respondents identified duplicates as having two main positive impacts: enriching the information and verifying the correctness. This has an implicit yet important prerequisite: the duplicates need to be detected and labeled beforehand. For instance, to achieve information richness, duplicate records must first be accurately identified and cross-references should be explicitly made. Similarly, for confirmation of results, the duplicate records need to be labeled beforehand. Subsequently, researchers can seek labeled duplicates to find additional interesting observations made by other researchers on the same entities, that is, to find out whether their records are consistent with others.

The views on *how to manage duplicates* are summarized in **Figure 4**. None of the practitioners regards duplicate detection as unnecessary. Moreover, 10 practitioners believe that current duplicate detection methods are insufficient. We propose the following suggestions accordingly.

Precision matters

Methods are needed to find duplicates accurately: "*It should correctly remove duplicate records, while leaving legitimate*

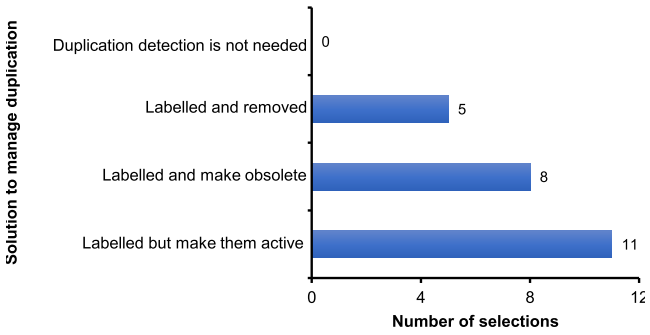

Figure 4 Solutions to duplicate records
The X-axis represents the options to address duplication; the Y-axis represents the corresponding number of participants selecting that option.

similar entries in the database." [R15] and "*Duplicate detection method need to be invariant to small changes (at the file level, or biological sample level); otherwise we would miss the vast majority of these*." [R9].

Automation matters

In some fields, few duplicate detection methods exist: "*We reuse GEO public data sets, to our knowledge there is no systematic duplicate detection*." [R7]; "*Not aware of any software*." [R3]; and "*I do not use any duplicate detection methods, they are often difficult to spot are usually based on a knowledge of the known size of the gene set*." [R21].

Characterization matters

The methods should analyze the characteristics of duplicates: "*A measure of how redundant the database records are would be useful*." [R24].

Robustness and generalization matter

"*All formats of data need to be handled crosswise; it does not help trying to find duplicates only within a single file format for a technology*." [R9].

To our knowledge, there is no universal approach to managing duplication. Similar databases may use different deduplication techniques. For instance, as sequencing databases, Encyclopedia of DNA Elements (ENCODE) uses standardized metadata organization, multiple validation identifiers, and its own merging mechanism for the detection and management of duplicate sequencing reads; the Sequence Read Archive (SRA) uses hash functions; and GEO uses manual curation in addition to hash functions [27]. Likewise, different databases may choose different parameters even when using the same deduplication approach. For instance, protein databases often use clustering methods to handle redundant records. However, the values of chosen similarity thresholds for clustering range from 30% to 100% in different databases [75]. Thus, it is impossible to provide a uniform solution to handling of duplication (as well as other quality issues). We introduce sample solutions used in UniProtKB/Swiss-Prot that demonstrate how quality issues are handled in a single database. The approaches or software used in the UniProtKB/Swiss-Prot curation pipeline may also provide insights into others.

Beyond duplication: other data quality issues

We extended the investigation to general quality issues other than duplication to complement the key insights. We asked the respondents for their opinions on general data quality issues. The two primary questions asked are as follows: *what data quality issues have been observed in biological databases*? and *why care about data quality*? The style is the same as the questions above on duplication. The detailed results are summarized in File S2. Overall. It is shown that the quality issues can be widespread; for example, each data quality issue has been observed by at least 80% of the respondents.

Limitations

It is worth noting that although we have carefully phrased the questions in the survey, it may still be the case that different respondents may have different internal definitions of duplicates in mind when responding. For example, some respondents may only consider records with minor differences as redundant records, whereas others may also include records with larger differences, even though they selected the same option. We acknowledge that this diversity of interpretation is inevitable—data is multifaceted; hence, data quality and the associated perspectives on it are also inevitable. The internal definitions of duplicate records depend on more specific context, and indeed, there is no universal agreement [24]. However, we argue that this does not detract from the results of the survey; respondents provided clear examples to support their choices, and these examples demonstrate that the types of duplicates do impact biological studies, regardless of internal variation in specific definitions. Such internal differences are also observed in other data quality studies, such as reviews on general data quality [76] and detection of duplicate videos [77].

It is also noteworthy that some databases primarily serve an archival purpose, such as INSDC and GEO. The records in these databases are directly coordinated by record submitters; therefore, the databases have had relatively little curation compared with databases like UniProtKB/Swiss-Prot. Arguably, data quality issues are not major concerns from an archival perspective. We did not examine the quality issues in archival databases; rather, we suggested that labeling duplicate records or records with other quality issues (without withdrawing or removing the records) could potentially facilitate database usage. The archival purpose does not limit other uses, for example, studies including BLAST searches against GenBank for sequence characterization [78–80]. In such cases, the sequences and annotations would impact the related analyses.

However, data quality issues may be important in archival databases as well. Indeed, in some instances, the database managers have been aware of data quality issues and are working on solutions. A recent work proposed by the ENCODE database team concerns the quality issues, particularly duplication in sequencing repositories such as ENCODE, GEO, and SRA [27]. They acknowledge that although archival databases are responsible for data preservation, duplication affects data storage and could mislead users. As a result, they propose three guidelines to prevent duplication in ENCODE and summarize other deduplication

approaches in GEO and SRA; furthermore, the ENCODE work encourages making a community effort (such as that by archival databases, publishers, and submitters) to handle data quality issues.

Biocuration: a solution to data quality issues in biological databases

In this section, we introduce solutions to data quality issues in biological databases. Biocuration is a general term that refers to addressing data quality issues in biological databases. We provide a concrete case study on the UniProtKB/Swiss-Prot curation pipeline comprising a detailed description on the curation procedure and an interview with the curation team leader. It provides an example of a solution to different quality issues.

The curation pipeline of UniProtKB/Swiss-Prot

UniProtKB has two data sections: UniProtKB/Swiss-Prot and UniProtKB/TrEMBL. Sequence records are first deposited in UniProtKB/TrEMBL, and then, selected records are transferred into UniProtKB/Swiss-Prot. Accordingly, curation in UniProtKB has two stages: (1) automatic curation in UniProtKB/TrEMBL, where records are automatically curated by software without manual review, and (2) expert (or manual) curation in UniProtKB/Swiss-Prot on selected records from UniProtKB/TrEMBL. A major task in automatic curation is to annotate records using annotation systems; for example, UniRules, which contains rules created by biocurators, and external rules from other annotation systems, such as RuleBase [81] and HAMAP [82], are used in this task. Rule UR000031345 is an example of UniRules (http://www.uniprot.org/unirule/UR000031345); Record B1YYB is also a sequence record example that was annotated using the rules during automatic curation. For expert curation, biocurators run a comprehensive set of software, search supporting information from a range of databases, manually review the results, and interpret the evidence level [31]. Table 1 describes representative software and databases used in expert curation [14,83–98]. This expert curation in UniProtKB/Swiss-Prot has 6 dedicated steps as shown in Table 1 and explained below.

Sequence curation

This step focuses on deduplication. It has two components: (1) detection and merging of duplicate records and (2) analysis and documentation of the inconsistencies caused by duplication. In this specific case, "duplicates" are records belonging to the same genes, an example of entity duplicates. Biocurators perform BLAST searches and also search other database resources to confirm whether two records belong to the same genes and merge them if they are. The merged records are explicitly documented in the record's *Cross-reference* section. Sometimes, the merged records do not have the same sequences, mostly owing to errors. Biocurators have to analyze the causes of these differences and document the errors.

Sequence analysis

Biocurators analyze sequence features after addressing duplications and inconsistencies. They run standard prediction tools, review and interpret the results, as well as annotate the records. The complete annotations for sequence features cover 39 annotation fields under 7 categories: molecule processing, regions, sites, amino acid modifications, natural variations, experimental info, and secondary structure (http://www.uniprot.org/help/sequence_annotation). As such, it involves a comprehensive range of software and databases to facilitate sequence analysis, some of which are shown in Table 1.

Literature curation

This step often contains two processes: retrieval of relevant literature and application of text mining tools to the analysis of text data, such as recognizing named entities [99] and identifying critical entity relationships [100]. The annotations are made using controlled vocabularies (the complete list is provided in the UniProtKB keyword documentation via http://www.uniprot.org/docs/keywlist) and are explicitly labeled "*Manual assertion based on experiment in literature*." Record Q24145 is an example that was annotated based on findings published in the literature (http://www.uniprot.org/uniprot/Q24145).

Family-based curation

This step transitions curation from single-record level to family level, finding relationships among records. Biocurators identify putative homologs using BLAST search results and phylogenetic resources, and make annotations accordingly. The tools and databases are the same as those in the *Sequence curation* step.

Evidence attribution

This step standardizes the curations made in the previous steps. Curations are made manually or automatically from different types of sources, such as sequence similarity, animal model results, and clinical study results. This step uses the Evidence and Conclusion Ontology (ECO) to describe evidence in a precise manner; it details the type of evidence and the assertion method (manual or automatic) used to support a curated statement [98]. As such, database users can know how the decision was made and on what basis. For example, ECO_0000269 was used in the literature curation for Record Q24145.

Quality assurance, integration, and update

The curation is complete at this point. This step finally checks everything and integrates curated records to the existing UniProtKB/Swiss-Prot knowledgebase. These records are then available in the new release. In turn, this helps further automatic curation within UniProtKB/Swiss-Prot. The newly made annotations will then be used as the basis for creating automatic annotation rules.

The curation in UniProtKB/Swiss-Prot: an interview

We interviewed UniProtKB/Swiss-Prot annotation team leader Sylvain Poux. The interview questions covered how UniProtKB/Swiss-Prot handles general data quality issues. Some of the responses are also related to specific curation processes in UniProtKB/Swiss-Prot, which show that the solutions are database-dependent as well. The detailed interview is summarized in the File S3. We have edited the questions for clarity and omitted answers where Poux did not offer a view.

Table 1 Representative software and resources used in expert curation

Curation step	Software/database	Purpose	Weblink	Ref.
Sequence curation				
Identify homologs	BLAST	Sequence alignment	https://blast.ncbi.nlm.nih.gov/	[83]
Document inconsistencies	Ensembl	Phylogenetic resources	https://www.ensembl.org/	[84]
	T-Coffee	Sequence difference (*e.g.*, alternative splicing) analysis	https://www.ebi.ac.uk/Tools/msa/tcoffee/	[85]
	MUSCLE		https://www.ebi.ac.uk/Tools/msa/muscle/	[86]
	ClustalW		https://www.ebi.ac.uk/Tools/msa/clustalw2/	[87]
Sequence analysis				
Predict topology	SignalP	Signal peptide prediction	http://www.cbs.dtu.dk/services/SignalP/	[88]
	TMHMM	Transmembrane domain prediction	http://www.cbs.dtu.dk/services/TMHMM/	[89]
Predict PTMs	NetNGlyc	*N*-glycosylation site prediction	http://www.cbs.dtu.dk/services/NetNGlyc/	[90]
	Sulfinator	Tyrosine sulfation site prediction	https://web.expasy.org/sulfinator/	[91]
Identify domains	InterPro	Retrieval of motif matches	https://www.ebi.ac.uk/interpro/	[92]
	REPEAT (REP tool)	Identification of repeats	https://www.ebi.ac.uk/interpro/	[93]
Literature curation				
Identify relevant literature	PubMed	Literature resources	https://pubmed.ncbi.nlm.nih.gov/	[94]
	iHOP		https://bio.tools/ihop	[95]
Extract named entities	PubAnnotation	Information extraction	http://pubannotation.org/	[96]
	PubTator		https://www.ncbi.nlm.nih.gov/research/pubtator/	[97]
Assign GOs	GO	Gene ontology terms	http://geneontology.org/	[14]
Family curation	BLAST	Sequence alignment	https://blast.ncbi.nlm.nih.gov/	[83]
Evidence attribution	ECO	Evidence code ontology	http://www.evidenceontology.org/	[98]

Note: A complete set of the software, including the detailed versions of the software, can be found in UniProt manual curation standard operating procedure documentation (www.uniprot.org/docs/sop_manual_curation.pdf). PTM, post-translational modification.

The aforementioned case study demonstrates that biocuration is an effective solution to diverse quality issues. Indeed, since 2003, when the first regular meeting among biocurators was held [101], the importance of biocuration activities has widely been recognized [20,102–104]. However, the biocuration community still lacks broader support. A survey of 257 former or current biocurators shows that biocurators suffer from a lack of secured funding for primary biological databases, exponential data growth, and underestimation of the importance of biocuration [69]; consistent results have also been demonstrated in other studies [105,106]. According to recent reports, the funding for model-organism databases would be cut by 30–40% and the same threat applies to other databases [107–109].

Conclusion

In this study, we explore the perspectives of both database managers and database users on the issue of data duplication—one of several significant data quality issues. We also extend the investigation to other data quality issues to complement this primary focus. Our survey of individual practitioners shows that duplication in biological databases is of concern: its characteristics are diverse and complex; its impacts cover almost all stages of database creation and analysis; and methods for managing the problem of duplication (manual or automatic) have significant limitations. The overall impacts of duplication are broadly negative, whereas the positive impacts such as enriched entity information and validation of correctness rely on the duplicate records being correctly labeled or cross-referenced. This suggests a need for further developing methods for precisely classifying duplicate records (accuracy), detecting different types of duplicates (characterization), and achieving scalable performance in different data collections (generalization). In some specific domains, duplicate detection software (automation) is a critical need.

The responses relating to general data quality further show that data quality issues go well beyond duplication. As can be inferred from our survey, curation—dedicated efforts to ensure that biological databases represent accurate and up-to-date scientific knowledge—is an effective tool for addressing quality issues. In addition, we provide a concrete case study on the UniProtKB/Swiss-Prot curation pipeline as a sample solution to data quality issues. However, manual curation alone is not sufficient to resolve all data quality issues due to rapidly growing data volumes in a context of limited resources. A broader community effort is required to manage data quality and provide support to facilitate data quality and curation.

Competing interests

The authors have declared no competing interests.

Acknowledgments

The project receives funding from the Australian Research Council through a Discovery Project (Grant No. DP150101550). We thank Sylvain Poux for contributions to the UniProtKB/Swiss-Prot curation case study. We acknowledge the participation of the following people in the survey: Cecilia Arighi (University of Delaware), Ruth C Lovering (University College London), Peter McQuilton (University of Oxford), and Valerie Wood (University of Cambridge).

Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gpb.2018.11.006.

ORCID

0000-0002-6036-1516 (Chen, Q)
0000-0003-1011-5410 (Britto, R)
0000-0002-7280-7191 (Erill, I)
0000-0002-2147-3638 (Jeffery, CJ)
0000-0003-3544-996X (Magrane, M)
0000-0003-0790-8313 (Onami, Ji)
0000-0002-3437-3329 (Robinson-Rechavi, M)
0000-0002-6345-6879 (Sponarova, J)
0000-0001-6622-032X (Zobel, J)
0000-0002-8661-1544 (Verspoor, K)

References

[1] Baxevanis A, Bateman A. The importance of biological databases in biological discovery. Curr Protoc Bioinformatics 2015;50:1–8.
[2] Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, et al. GenBank. Nucleic Acids Res 2017;45:D37.
[3] Toribio AL, Alako B, Amid C, Cerdeño-Tarrága A, Clarke L, Cleland I, et al. European nucleotide archive in 2016. Nucleic Acids Res 2017;45:D32–36.
[4] Cochrane G, Karsch-Mizrachi I, Takagi T. The international nucleotide sequence database collaboration. Nucleic Acids Res 2017;44:D48–51.
[5] The UniProt Consortium. UniProt: the universal protein knowledgebase. Nucleic Acids Res 2017;45:D158–69.
[6] Rose PW, Prlić A, Altunkaya A, Bi C, Bradley AR, Christie CH, et al. The RCSB protein data bank: integrative view of protein, gene and 3D structural information. Nucleic Acids Res 2017;45:D271–81.
[7] Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, et al. Rfam 12.0: updates to the RNA families database. Nucleic Acids Res 2015;43:D130–7.
[8] Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res 2016;44:D279–85.
[9] Basu S, Fey P, Pandit Y, Dodson R, Kibbe WA, Chisholm RL. DictyBase 2013: integrating multiple Dictyostelid species. Nucleic Acids Res 2013;41:D676–83.
[10] McDowall MD, Harris MA, Lock A, Rutherford K, Staines DM, Bähler J, et al. PomBase 2015: updates to the fission yeast database. Nucleic Acids Res 2015;43:D656–61.
[11] Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, et al. ArrayExpress update—simplifying data submissions. Nucleic Acids Res 2015;43:D1113–6.
[12] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res 2013;41:D991–5.
[13] Galperin MY, Fernández-Suárez XM, Rigden DJ. The 24th annual Nucleic Acids Research database issue: a look back and upcoming changes. Nucleic Acids Res 2017;45:D1–11.

[14] The Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Res 2017;45:D331–8.
[15] Bursteinas B, Britto R, Bely B, Auchincloss A, Rivoire C, Redaschi N, et al. Minimizing proteome redundancy in the UniProt Knowledgebase. Database (Oxford) 2016;2016:baw139.
[16] Bouadjenek MR, Verspoor K, Zobel J. Literature consistency of bioinformatics sequence databases is effective for assessing record quality. Database (Oxford) 2017;2017:bax021.
[17] Poux S, Arighi CN, Magrane M, Bateman A, Wei CH, Lu Z, et al. On expert curation and sustainability: UniProtKB/Swiss-Prot as a case study. Bioinformatics 2017:3454–60.
[18] Nellore A, Jaffe AE, Fortin JP, Alquicira-Hernández J, Collado-Torres L, Wang S, et al. Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the Sequence Read Archive. Genome Biol 2016;17:266.
[19] Huntley RP, Sitnikov D, Orlic-Milacic M, Balakrishnan R, D'Eustachio P, Gillespie ME, et al. Guidelines for the functional annotation of microRNAs using the Gene Ontology. RNA 2016;22:667–76.
[20] Howe D, Costanzo M, Fey P, Gojobori T, Hannick L, Hide W, et al. Big data: the future of biocuration. Nature 2008;455:47–50.
[21] Rosikiewicz M, Comte A, Niknejad A, Robinson-Rechavi M, Bastian FB. Uncovering hidden duplicated content in public transcriptomics data. Database (Oxford) 2013;2013:bat010.
[22] Poux S, Magrane M, Arighi CN, Bridge A, O'Donovan C, Laiho K. Expert curation in UniProtKB: a case study on dealing with conflicting and erroneous data. Database (Oxford) 2014;2014:bau016.
[23] Pfeiffer F, Oesterhelt D. A manual curation strategy to improve genome annotation: application to a set of haloarchael genomes. Life 2015;5:1427–44.
[24] Chen Q, Zobel J, Verspoor K. Duplicates, redundancies and inconsistencies in the primary nucleotide databases: a descriptive study. Database (Oxford) 2017;2017:baw163.
[25] Chen Q, Zobel J, Verspoor K. Benchmarks for measurement of duplicate detection methods in nucleotide databases. Database (Oxford) 2017. https://doi.org/10.1093/database/baw164.
[26] Chen Q, Zobel J, Zhang X, Verspoor K. Supervised learning for detection of duplicates in genomic sequence databases. PLoS One 2016;11:e0159644.
[27] Gabdank I, Chan ET, Davidson JM, Hilton JA, Davis CA, Baymuradov UK, et al. Prevention of data duplication for high throughput sequencing repositories. Database (Oxford) 2018;2018:bay008.
[28] The UniProt Consortium. Activities at the universal protein resource (UniProt). Nucleic Acids Res 2014;42:D191–8.
[29] Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 2015;31:926–32.
[30] Gasteiger E, Jung E, Bairoch AM. SWISS-PROT: connecting biomolecular knowledge via a protein database. Curr Issues Mol Biol 2001;3:47–55.
[31] The UniProt Consortium. UniProt: a hub for protein information. Nucleic Acids Res 2015;43:D204–12.
[32] Cole C, Barber JD, Barton GJ. The Jpred 3 secondary structure prediction server. Nucleic Acids Res 2008;36:W197–201.
[33] Korning PG, Hebsgaard SM, Rouzé P, Brunak S. Cleaning the GenBank *Arabidopsis thaliana* data set. Nucleic Acids Res 1996;24:316–20.
[34] Koh J, Lee ML, Khan AM, Tan P, Brusic V. Duplicate detection in biological data using association rule mining. Proceedings of the Second European Workshop on Data Mining and Text Mining in Bioinformatics. Pisa, Italy. September 20–24, 2004;501:S22388.
[35] Salgado H, Santos-Zavaleta A, Gama-Castro S, Peralta-Gil M, Peñaloza-Spínola MI, Martínez-Antonio A, et al. The comprehensive updated regulatory network of *Escherichia coli* K-12. BMC Bioinformatics 2006;7:5.
[36] Bouffard M, Phillips MS, Brown AM, Marsh S, Tardif JC, van Rooij T. Damming the genomic data flood using a comprehensive analysis and storage data structure. Database (Oxford) 2010;2010:baq029.
[37] Bastian F, Parmentier G, Roux J, Moretti S, Laudet V, Robinson-Rechavi M. Bgee: integrating and comparing heterogeneous transcriptome data among species. International Workshop on Data Integration in the Life Sciences. Evry, France. June 25–27, 2008:124–31.
[38] Müller H, Naumann F, Freytag JC. Data quality in genome databases. Proceedings of the Conference on Information Quality. Cambridge, USA. November 7–9, 2003.
[39] Chellamuthu S, Punithavalli DM. Detecting redundancy in biological databases? An efficient approach. Global J Comput Sci Technol 2009;9.
[40] Bork P, Bairoch A. Go hunting in sequence databases but watch out for the traps. Trends Genet 1996;12:425–7.
[41] Pennisi E. Keeping genome databases clean and up to date. Science 1999;286:447–50.
[42] Schnoes AM, Brown SD, Dodevski I, Babbitt PC. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput Biol 2009;5:e1000605.
[43] Fan W. Data quality: from theory to practice. Proc ACM SIGMOD Int Conf Manag Data. Melbourne, Australia. May 2015;44:7–18.
[44] Guptill SC, Morrison JL. Elements of spatial data quality. Amsterdam: Elsevier B.V; 2013.
[45] Abiteboul S, Dong L, Etzioni O, Srivastava D, Weikum G, Stoyanovich J, et al. The elephant in the room: getting value from Big Data. In: Proceedings of the 18th International Workshop on Web and Databases. Melbourne, Australia. May; 2015:1–5.
[46] Sadiq S, Papotti P. Big data quality-whose problem is it? In: IEEE 32nd International Conference on Data Engineering (ICDE). Helsinki, Finland. May; 2016:1446–7.
[47] Ballou DP, Pazer HL. Modeling data and process quality in multi-input, multi-output information systems. Manage Sci 1985;31:150–62.
[48] Wang RY, Storey VC, Firth CP. A framework for analysis of data quality research. IEEE Trans Knowl Data Eng 1995;7:623–40.
[49] Yeganeh NK, Sadiq S, Sharaf MA. A framework for data quality aware query systems. Inf Syst 2014;46:24–44.
[50] Wang RY, Strong DM. Beyond accuracy: what data quality means to data consumers. J Manag Inf Syst 1996;12:5–33.
[51] Wixom BH, Watson HJ. An empirical investigation of the factors affecting data warehousing success. MIS Quarterly 2001:17–41.
[52] Coussement K, Van den Bossche FA, De Bock KW. Data accuracy's impact on segmentation performance: benchmarking RFM analysis, logistic regression, and decision trees. J Bus Res 2014;67:2751–8.
[53] Bultet LA, Aguilar Rodriguez J, Ahrens CH, Ahrne EL, Ai N, Aimo L, et al. The SIB Swiss Institute of Bioinformatics' resources: focus on curated databases. Nucleic Acids Res 2016;44:D27–37.
[54] Magrane M. The UniProt Consortium. UniProt Knowledgebase: a hub of integrated protein data. Database (Oxford) 2011;2011:bar009.
[55] Mani M, Chen C, Amblee V, Liu H, Mathur T, Zwicke G, et al. MoonProt: a database for proteins that are known to moonlight. Nucleic Acids Res 2014;43:D277–82.
[56] Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The molecular signatures database hallmark gene set collection. Cell Syst 2015;1:417–25.
[57] Kılıç S, White ER, Sagitova DM, Cornish JP, Erill I. CollecTF: a database of experimentally validated transcription factor-binding sites in bacteria. Nucleic Acids Res 2014;42:D156–60.

[58] Kılıç S, Sagitova DM, Wolfish S, Bely B, Courtot M, Ciufo S, et al. From data repositories to submission portals: rethinking the role of domain-specific databases in CollecTF. Database (Oxford) 2016;2016:baw055.
[59] Rutherford KM, Harris MA, Lock A, Oliver SG, Wood V. Canto: an online tool for community literature curation. Bioinformatics 2014;30:1791–2.
[60] Arighi CN, Drabkin H, Christie KR, Ross KE, Natale DA. Tutorial on protein ontology resources. In: Wu C, Arighi CN, Ross KE, editors. Protein bioinformatics: from protein modifications and networks to proteomics. New York: Humana Press; 2017. p. 57–78.
[61] Poux S, Arighi CN, Magrane M, Bateman A, Wei CH, Lu Z, et al. On expert curation and sustainability: UniProtKB/Swiss-Prot as a case study. Bioinformatics 2017;33:3454–60.
[62] Gaudet P, Michel PA, Zahn-Zabal M, Britan A, Cusin I, Domagalski M, et al. The neXtProt knowledgebase on human proteins: 2017 update. Nucleic Acids Res 2017;45:D177–82.
[63] Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, et al. CDD: NCBI's conserved domain database. Nucleic Acids Res 2014;43:D222–6.
[64] Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res 2016;44:D862–8.
[65] Orchard S. Data standardization and sharing—the work of the HUPO-PSI. Biochim Biophys Acta 2014;1844:82–7.
[66] Poux S, Gaudet P. Best practices in manual annotation with the gene ontology. In: Dessimoz C, Škunca N, editors. The gene ontology handbook. New York: Humana Press; 2017. p. 41–54.
[67] Burge S, Attwood TK, Bateman A, Berardini TZ, Cherry M, O'Donovan C, et al. Biocurators and biocuration: surveying the 21st century challenges. Database (Oxford) 2012;2012:bar059.
[68] Hirschman L, Burns GAC, Krallinger M, Arighi C, Cohen KB, Valencia A, et al. Text mining for the biocuration workflow. Database 2012;2012:bas020.
[69] Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics 2011;27:1739–40.
[70] Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 2012;28:3150–2.
[71] Song M, Rudniy A. Detecting duplicate biological entities using Markov random field-based edit distance. Knowl Inf Syst 2010;25:371–87.
[72] Chatr-aryamontri A, Oughtred R, Boucher L, Rust J, Chang C, Kolas NK, et al. The BioGRID interaction database: 2017 update. Nucleic Acids Res 2017;45:D369–79.
[73] Gilks WR, Audit B, De Angelis D, Tsoka S, Ouzounis CA. Modeling the percolation of annotation errors in a database of protein sequences. Bioinformatics 2002;18:1641–9.
[74] Bastian FB, Chibucos MC, Gaudet P, Giglio M, Holliday GL, Huang H, et al. The Confidence Information Ontology: a step towards a standard for asserting confidence in annotations. Database (Oxford) 2015;2015:bav043.
[75] Chen Q, Wan Y, Zhang X, Lei Y, Zobel J, Verspoor K. Comparative analysis of sequence clustering methods for deduplication of biological databases. ACM J Data Inf Qual 2018;9:17.
[76] Batini C, Scannapieco M. Data and information quality: dimensions, principles and techniques. Berlin: Springer; 2016.
[77] Liu J, Huang Z, Cai H, Shen HT, Ngo CW, Wang W. Near-duplicate video retrieval: current research and future trends. ACM Comput Surv 2013;45:44.
[78] Chowdhary A, Kathuria S, Singh PK, Sharma B, Dolatabadi S, Hagen F, et al. Molecular characterization and in vitro antifungal susceptibility of 80 clinical isolates of mucormycetes in Delhi, India. Mycoses 2014;57:97–107.
[79] Qiao Y, Xu D, Yuan H, Wu B, Chen B, Tan Y, et al. Investigation on the association of soil microbial populations with ecological and environmental factors in the Pearl River Estuary. J Geosci Environ Protect 2018;6:8.
[80] Persson S, Al-Shuweli S, Yapici S, Jensen JN, Olsen KEP. Identification of clinical aeromonas species by rpoB and gyrB sequencing and development of a multiplex PCR method for detection of *Aeromonas hydrophila*, *A. caviae*, *A. veronii*, and *A. media*. J Clin Microbiol 2015;53:653–6.
[81] Fleischmann W, Gateau A, Apweiler R. A novel method for automatic functional annotation of proteins. Bioinformatics 1999;15:228–33.
[82] Pedruzzi I, Rivoire C, Auchincloss AH, Coudert E, Keller G, De Castro E, et al. HAMAP in 2015: updates to the protein family classification and annotation system. Nucleic Acids Res 2015;43:D1064–70.
[83] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–402.
[84] Herrero J, Muffato M, Beal K, Fitzgerald S, Gordon L, Pignatelli M, et al. Ensembl comparative genomics resources. Database (Oxford) 2016;2016:bav096.
[85] Notredame C, Higgins DG, Heringa J. T-coffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol 2000;302:205–17.
[86] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004;32:1792–7.
[87] Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:4673–80.
[88] Emanuelsson O, Brunak S, Von Heijne G, Nielsen H. Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc 2007;2:953.
[89] Krogh A, Larsson B, Von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. J Mol Biol 2001;305:567–80.
[90] Julenius K, Mølgaard A, Gupta R, Brunak S. Prediction, conservation analysis, and structural characterization of mammalian mucin-type *O*-glycosylation sites. Glycobiology 2005;15:153–64.
[91] Monigatti F, Gasteiger E, Bairoch A, Jung E. The Sulfinator: predicting tyrosine sulfation sites in protein sequences. Bioinformatics 2002;18:769–70.
[92] Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, et al. InterPro in 2017—beyond protein family and domain annotations. Nucleic Acids Res 2016;45:D190–9.
[93] Andrade MA, Ponting CP, Gibson TJ, Bork P. Homology-based method for identification of protein repeats using statistical significance estimates. J Mol Biol 2000;298:521–37.
[94] NCBI RC. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 2016;44:D7.
[95] Müller HM, Kenny EE, Sternberg PW. Textpresso: an ontology-based information retrieval and extraction system for biological literature. PLoS Biol 2004;2:e309.
[96] Kim JD, Wang Y, Fujiwara T, Okuda S, Callahan TJ, Cohen KB. Open Agile text mining for bioinformatics: the PubAnnotation ecosystem. Bioinformatics 2019;35:4372–80.
[97] Wei CH, Kao HY, Lu Z. PubTator: a web-based text mining tool for assisting biocuration. Nucleic Acids Res 2013;41:W518–22.
[98] Chibucos MC, Mungall CJ, Balakrishnan R, Christie KR, Huntley RP, White O, et al. Standardized description of scientific

evidence using the Evidence Ontology (ECO). Database (Oxford) 2014;2014:bau075.
[99] Choi M, Liu H, Baumgartner W, Zobel J, Verspoor K. Coreference resolution improves extraction of Biological Expression Language statements from texts. Database (Oxford) 2016;2016:baw076.
[100] Peng Y, Wei CH, Lu Z. Improving chemical disease relation extraction with rich features and weakly labeled data. J Cheminform 2016;8:53.
[101] Harding A. Rise of the Bio-librarian: the field of biocuration expands as the data grows. Scientist 2006;20:82–4.
[102] Bourne PE, McEntyre J. Biocurators: contributors to the world of science. PLoS Comput Biol 2006;2:e142.
[103] Bateman A. Curators of the world unite: the International Society of Biocuration. Bioinformatics 2010;26:991.
[104] Mitchell CS, Cates A, Kim RB, Hollinger SK. Undergraduate biocuration: developing tomorrow's researchers while mining today's data. J Undergrad Neurosci Educ 2015;14:A56–65.
[105] Reiser L, Berardini TZ, Li D, Muller R, Strait EM, Li Q, et al. Sustainable funding for biocuration: The Arabidopsis Information Resource (TAIR) as a case study of a subscription-based funding model. Database (Oxford) 2016;2016:baw018.
[106] Karp PD. How much does curation cost?. Database (Oxford) 2016;2016:baw110.
[107] Hayden EC. Funding for model-organism databases in trouble. Nature 2016. https://doi.org/10.1038/nature.2016.20134.
[108] Kaiser J. Funding for key data resources in jeopardy. Science 2016;351:14.
[109] Bourne PE, Lorsch JR, Green ED. Perspective: sustaining the big-data ecosystem. Nature 2015;527:S16–7.