doi:10.20944/preprints202002.0146.v1

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

Katharina Frey [1,+,*], Alenka Hafner [+], Boas Pucker [1,2]

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

[2] Molecular Genetics and Physiology of Plants, Faculty of Biology and Biotechnology, Ruhr-University Bochum, Universitätsstraße 150, 44801 Bochum, Germany

[+] contributed equally

[*] corresponding author

email addresses:

KF: kfrey@cebitec.uni-bielefeld.de

AH: ah912@alumni.cam.ac.uk

BP: bpucker@cebitec.uni-bielefeld.de

ORCIDs:

KF: https://orcid.org/0000-0002-4022-8531

AH: https://orcid.org/0000-0003-4262-9176

BP: https://orcid.org/0000-0002-3321-7471

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. Due to the prominence, abundance and wide distribution of sequencing results, we focus on the reuse of publicly available sequence datasets. Through selected examples of successful reuse of different data (genome, transcriptome, proteome, metabolome, phenotype and ecosystem), with their respective limitations and risks, we illustrate the enormous potential of the practice. A checklist to determine the reuse value and potential of particular dataset is also provided.

Introduction

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 worldwide exchange of 'big data', which is common in biology (e.g. genomic sequences). The sharing of datasets leads to statistical robustness and allows re-analysis of existing datasets underlying claims (1) while enabling discovery of novel patterns through meta-analysis (2). Such building on existing knowledge and constant re-examination of prevailing hypotheses is the foundation of the scientific endeavour and is bolstered through ‘open science’.

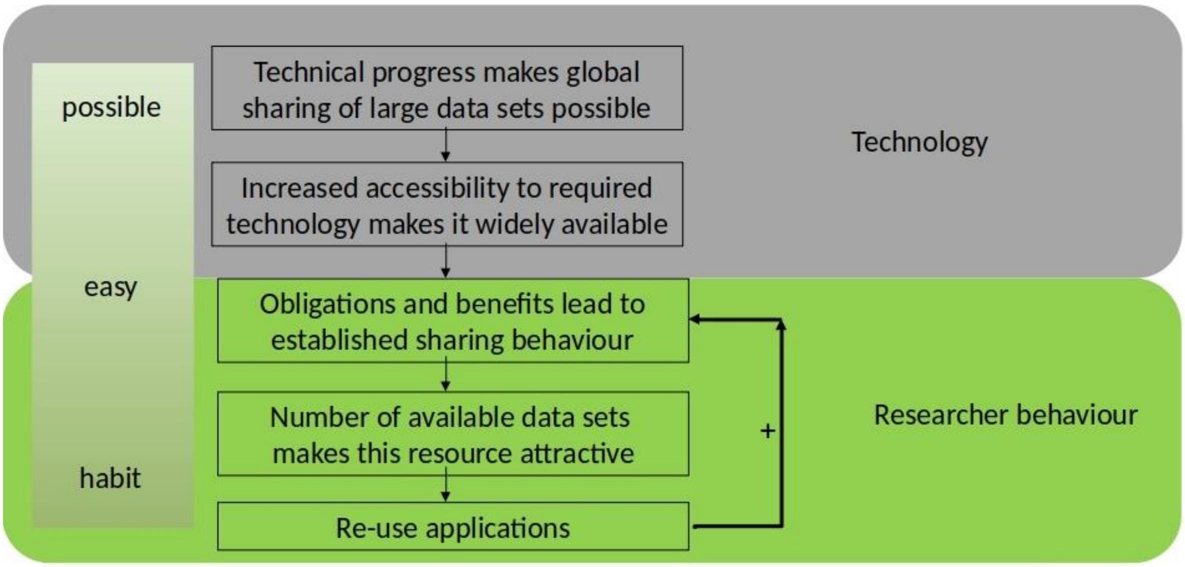

possible
easy
habit
Technical progress makes global sharing of large data sets possible
Increased accessibility to required technology makes it widely available
Technology
Obligations and benefits lead to established sharing behaviour
Number of available data sets makes this resource attractive
Re-use applications
+
Researcher behaviour

Figure 1: (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.

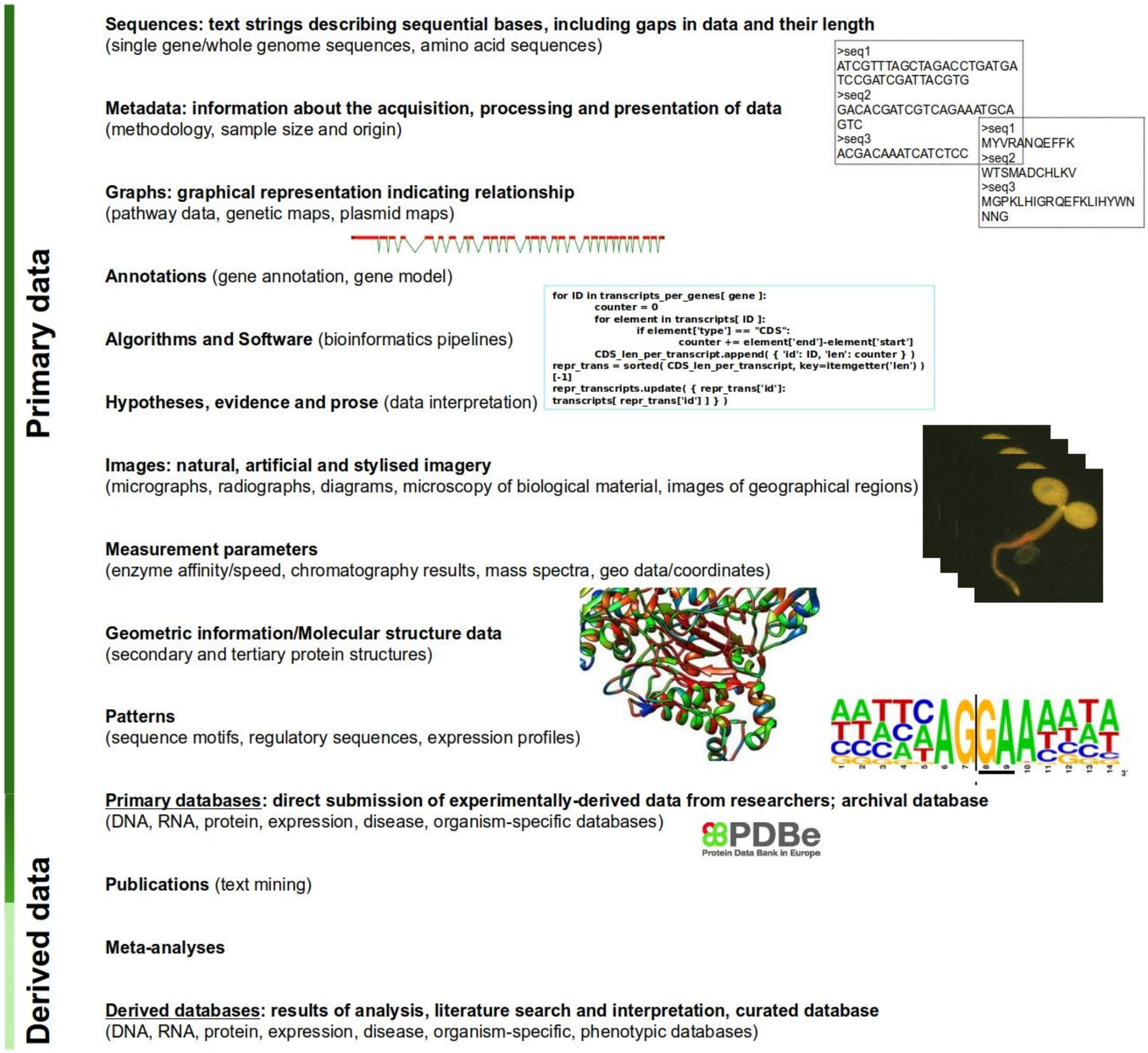
To establish data sharing as a norm it had to be introduced through obligations and promoted through benefits (3) for researchers. Numerous funding agencies and publishers (e.g. Nature: (4), NSF: (5), PLOS: (6)) require all data be made publicly available within a certain time frame, with an indication that this leads to increased transparency in the field (7). Dataset sharing may also increase attention to the associated research and resulting in additional citations, an added encouragement for authors (8). Despite these measures and benefits, an analysis (2017) of 318 biomedical journals revealed that only 11.9% of journals explicitly stated that data sharing was required as a condition of publication (9). Many international data sharing guidelines like FAIR (Findable Accessible Interoperable Reusable) (10), TOP (The Transparency and Openness Promotion) (11), Open Data in a Big Data World (1) and the Beijing Declaration (CODATA) (12) have emerged by necessity of the 'big data revolution'.

However, if the publicly available datasets are not widely re-examined (either checked for quality or re-analysed), enforcement of open science through policy may not be sufficient to harness the full power of global sharing (13). Without a clear distinction between fair and unjust reuse, open science initiatives might not be explicit enough to encourage fair reuse. For this reason, we distinguish 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 reproduction of studies to confirm the result, fair reuse should be considered and encouraged as an option enabled by 'open science'.

Besides this ambiguity, 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 (14). Regardless of the cause, a ‘backlog’ of under-utilised reliable datasets leads to unnecessary experiments (e.g. extensive repetitive sequencing increasing costs) and likely hides useful patterns. Therefore, education about the opportunities, challenges, limitations and techniques of data reuse is a vital task. Here, we highlight hurdles which need to be overcome, point out constraints which must be considered, and give examples of successful data reuse to inform future projects. In addition, we provide a checklist for biologists to aid in determining whether a particular dataset is fit for reuse.

Types of reusable data

There are numerous different types of datasets which harbour reuse potential (Fig. 2) including 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. plasmid maps, sequence motifs e.g. collected in JASPAR (http://jaspar.genereg.net/), 4: chromatography results and mass spectra, 5: information about the structure of proteins, 6: biochemical parameters of enzymes e.g. affinity or speed, 7: geo data e.g. coordinates of observations, 8: images of biological material or geographical regions or 9: phenotypic data e.g. collected in the plant genomics and phenomics research data repository (PGP) using the e!DAL software infrastructure (15).



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

```
>seq1
ATCGTTTAGCTAGACCTGATGA
TCCGATCGATTACGTG
>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 (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)

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)

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)

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: (16). (sources of the pictures: (17–20)).

Generally, datasets can be classified as primary and derived. One example is the famous eGFP browser 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 bioinformatic 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.

Further, it has to be considered that specific fields require the integration of data of various types, formats and abundance (21) which is hard to realise by a single database and therefore requires cooperation to encourage data reuse. Here, we focus on the reuse of primary data, specifically different sequences as the data type characteristic of the life sciences and continuously producing vast amounts of information.

Potential of reusing public datasets

There are numerous advantages of making data publicly available for reuse including reduction of costs, reproduction and accountability of research, enabled discovery of additional scientific knowledge, and detection of novel biological information (13,21,22). While the scientific community and society is profiting most from public datasets, there are additional well documented benefits of open access for authors themselves (e.g. increase in attention and ultimately citations) (3,23,24). Researchers can build a reputation by generating high quality and well documented datasets. Further, the publication of data not only contributes to the advancement of the scientific community but can additionally be helpful for other areas like education or business (13). These benefits for individuals and the scientific community as a whole make a strong case for the obligation to reuse data whenever that is appropriate.

Preventing information loss

Making data publicly available is an elegant way to prevent information loss resulting from underexamined data stored on various servers. 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. An outstanding example 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 (25,26). Further, meta-analyses like the prediction of specific genomic features together with machine learning approaches require large amounts of data which are already available and therefore should not and cannot

doi:10.20944/preprints202002.0146.v1

be produced once again. There are already many public repositories for genomic and gene-expression data like the SRA / European Nucleotide Archive (ENA) and Gene Expression Omnibus (GEO), respectively. However, the low availability of metadata in standardized formats with sufficient additional information leads to a lack of reproducibility (27).

Benefits for databases

The reuse of sequence data is of increasing importance due to the large and still rapidly growing size of the corresponding databases (Fig. 3). The size of the SRA alone increased from 3,092,408 entries to 6,243,265 entries within two years (September 2016 - September 2018) (28,29) and 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 (30). Approximately 120 million sequences and annotations of proteins were available within UniProtKB/TrEMBL in 2018 (31).

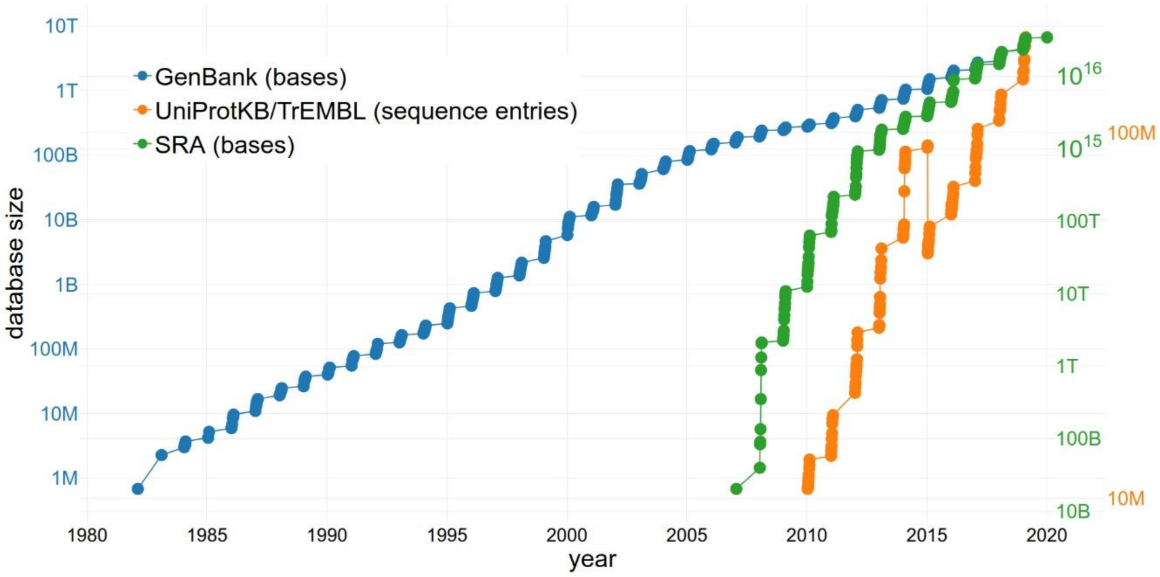


Figure 3: Increasing size of selected databases over time. The number of bases/sequence entries in GenBank, the Sequence Read Archive (SRA) and UniProtKB/TrEMBL is shown, respectively. Note the logarithmic scale of the y-axes. The drop of sequence entries in UniProtKB/TrEMBL (in 2015) can be explained by the removal of duplicates.

The increasing size of scientific data is a challenge for the databases regarding e.g. storage space and data management. However, reusing available data instead of producing new and redundant datasets results in a lower amount of duplicates and keeps databases concise. 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 (32). However, the open access to data would require appropriate quality of the data in order to be reliable for the user which the peer review process can hardly accomplish. This is especially a problem for clinical trials as the results could have a direct impact on the patients involved. Therefore, it is important that the limitations of particular datasets are identified and the associated risks assessed.

Unknown quality

Quality differences are a big issue when reusing public data. Mislabelled or swapped samples alongside intrinsic errors, like missing technical replicates, can be a problem as they are almost impossible to identify. Further, there are quality differences between user-submitted public datasets, very curated databases for specific organisms, ones with inherently small holding size (like PDB or SwissProt) and phenotype databases. 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 always the case that a large number of reads/replicates with low noise means that the emerging results are true. Indeed, one can imagine a large dataset trend with low noise produced when one author/group is responsible for most of the data and a systematic error is present and so there is low deviation. On the other end of the spectrum, the use of a small dataset that is believed to be of “higher quality” and leads to low noise, may hide a novel pattern or even show a non-existing one.

Information regarding experimental design, methods and conditions is often insufficient and results in datasets unsuitable for reuse. However, additional requirements for data submission should not result in fewer publicly available datasets (27). There is a trade-off between the collection of detailed metadata during submission and high submission numbers. Many sequence databases like ENA are handling this elegantly. 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 (33).

Denormalisation

Of particular concern is the circular reuse of data which can, for example, lead to a heavily denormalised annotation in databases (34) with the same data being stored multiple times in the same database, without that reflecting the true data distribution (e.g. sequencing and annotation errors propagated by reuse and not eliminated by additional published sequences that would show it to be statistically insignificant). For annotations, it has been shown that it is possible to detect low-quality entries, resulting from this denormalisation, by looking for specific patterns of provenance in the database (35). 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 as the experimental characterization of numerous genes in a diverse set of species cannot be expected in the near future.

Comparison and integration of different databases

Further, the comparison and integration of datasets from different sources remains a challenge (13). Examples are enormous differences in the annotations provided by the different databases e.g. NCBI, ENSEMBL, and phytozome for the same species. For any valid comparison between datasets from different databases or for integration of databases themselves, an established and unified file standard is crucial. FASTA (36), FASTQ (37), and SAM/BAM (38) are famous examples of file standards that allowed effective exchange of information between numerous groups involved in the earliest sequencing projects (39–41).

Re-analysis and metainformation

Re-analysis can be a way to tackle the issues mentioned above. As with re-examination of public biodiversity data to correct errors (42,43), so should sequence repositories reflect changes in the field's consensus (e.g. about specific annotations). This can be achieved through curation and self-correction, with both being difficult to directly re-enforce.

Therefore, a way to reduce some of the risks of reuse would be investing in a controlled environment containing extensively peer-reviewed datasets (32). The epitome of such databases are ‘expression atlases’ with manually curated and annotated sequences checked for quality and re-analysed using standardised methods (44). However, regarding the enormous and still increasing amount of e.g. sequence data, this is hardly an option for all data types. Different strategies might work for different data types or different communities. In all cases, specific standards and formats for data reuse should be applied (13). And a defined, suitable environment or database could also include follow-up data for a detailed understanding of the primary data and the corresponding results. 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.

Indeed, it is the metadata (information about the acquisition, processing and presentation) published that is of critical importance in checking for quality. Data papers (already common practice in Astronomy (45) have been indicated as a solution to the quality-check problem (46) of reuse by providing descriptions of methods for collecting, processing, and verifying data (13). Widespread publication of such metadata in data journals (47) is vital to the construction of high quality, peer-reviewed datasets. Consequently, method-focused or data-focused journals, like e.g. BMC Plant Methods, Nature Methods and GigaScience, emerged during the last years.

A dataset cannot be validly reused, if its metadata is not also assessed. Therefore, 'open science' incentives and database contribution guidelines should require the inclusion of metadata in all submissions to public datasets. This would not only encourage authors to collect data with reuse in mind (48), but enable productive and valid re-analysis. A reusability score assigned by the community could increase the quality of the provided metadata.

Research integrity considerations

The use of the same dataset in several different studies by the same author could be considered as a type of dual publication (49). However, such reuse is not contentious to the same extent as plagiarism, if it reveals novel findings and is not only reused to boost the number of publications. To a certain degree, this issue can be tackled with data publications (as is the case with unknown quality) which might also prevent the splitting of a coherent dataset over multiple publications. By providing 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 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 present prominent examples and an elegant solution for data reports if a valuable dataset should be shared with the community but does not meet all criteria for publication as a full research article.

On the one hand, modern biologists are encouraged to make use of publicly available sequence repositories and mine data generated by others. Furthermore, there is an argument to be made that data reuse is an obligation; not comparing one’s dataset to publicly available analogues is akin to ignoring replicated experiments (14). This reduces the rate of redundant sequencing and unveils new correlations through meta-analysis. Cutting such costs (50) enables groups with small budgets to harness extensive datasets thus enhancing equality. This especially aids early career researchers, who are outsiders of the scientific establishment and likely experience more barriers to other aspects of open science (51), yet are highly involved in data collection and analysis (52).

On the other hand, there appears to be a trend for people to not only supplement their wet lab and sequencing results with publicly available data but to publish almost exclusively from the latter (19,53–55). At the far end of the 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. ‘Research parasitism’ is associated with numerous advantages like 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. Despite some expressed concerns regarding such ‘parasitism’ (53), including the fear of exploitation when acquiring the data was particularly expensive or labour-intensive (51), the practice of reuse seems to prevail in the open science culture. 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.

Examples of successful data reuse

There are already numerous examples of successful studies which involved intensive reuse of public datasets. Table 1 shows selected reuse cases to cover many areas and concepts of data reuse sorted by the type of the analysed data. Genomic data can, for example, be harnessed for pangenomic analyses (56) while transcriptomic and ChIP-seq data might be useful for the investigation or construction of regulatory networks (57). Phylogenetic analysis of groups from individual gene families (58) to whole taxonomic groups (59) 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 (22). The taxonomic classification of sequences identified in metagenomic studies is another application which heavily relies on available data as the quality scales with quality and size of the available data (60). It can 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.

Table 1: Examples of dataset reuse for a novel purpose.

Examples	**Limitations / risks**
Genome	
Assembly of new genome sequences, e.g. organelle genomes, based on public datasets (61)	Potential contaminations are unknown, only submitter of original reads can submit assembly
Motif identification, e.g. Deep-learning method for identifying Poly(A) signals (62)	A large and suitable training set is required; can still not predict 100 % of the motifs correctly
Pangenomic analysis, e.g. for bread wheat (56)	Assembly quality might differ between different studies
GWAS to associate variants (QTLs, SNPs) with traits, e.g. single-plant GWAS for identification of plant-height candidate SNPs (63)	Large number of false positives requires large datasets, their sharing and compulsory replication (64)
Transcriptome	
Co-expression analysis to find connected genes, e.g. identification of long non-coding RNAs associated with atherosclerosis progression (65); Co-expression networks, e.g. related to bamboo development using public RNA-Seq data (66) or related to cellulose synthesis using public microarray data (67); 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 (68)	Batch effects if large sample groups come from the same source
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 (69)	Batch effects if large sample groups come from the same source, success depends on the gene expression data context
Identification of qRT-PCR reference genes (70–72)	Batch effects if large sample groups come from the same source
Gene prediction via analysis of RNA-Seq data (73) and e.g. GeMoMa is using this heavily (74)	Batch effects if large sample groups come from the same source
Gene expression web sites, e.g. the eGFP browser (75)	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 (19,55)	Batch effects if large sample groups come from the same source and annotation errors will impact analysis results
Extraction of new sequences for phylogenetic analysis (18,76)	Reliability of source is crucial, transcriptome assemblies are inherently incomplete as not all genes are expressed at the same time
Proteome	
Identification of antimicrobial peptides (22)	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 (77)	Meta-analysis allows extrapolation only for highly specific conditions due to numerous different experimental conditions in the used studies
Metabolome	
Metabolic modelling (78)	Precise conditions of experiments are different between labs, measurement biases possible
Combining network analysis and machine learning to predict metabolic pathways (79)	Cannot be used to predict catalytic activity, only predict pathways
Phenotype	
Deep learning methods for image-based phenotyping, e.g. leaf counting (80) or root and shoot feature identification (81)	Large datasets are required
Ecosystem	
Ecosystem modelling, e.g. reuse of model code/reuse of eutrophication models for studying climate change (82)	Partly overly simplified models; validity of outcomes must be tested; observations of species are sometimes placed at institutes of districts/regions

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. 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.

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

Criteria	**Questions to consider**	**Suggestions/Suitable controls**
Integrity of the source	Is the source/submitter associated with data fabrication / plagiarism?	Check potential conflicts of interests/funding
Biases	How was the data generated? Are there batch effects?	Selection in comparison of random subsets to avoid biases
Missing metainformation (sparsity)	Do you have all relevant information (e.g., information about the biological material)?	Possibility to contact the authors
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: (NGS) technologies - For assemblies: type/version of bioinformatic tools and 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
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

Conclusion

In this review, challenges, limitations and risks as well as the potential of reusing public datasets were shown. Further, successful examples of data reuse were provided.

There are different steps to achieve a data sharing behaviour which complies with open data principles. As already 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.

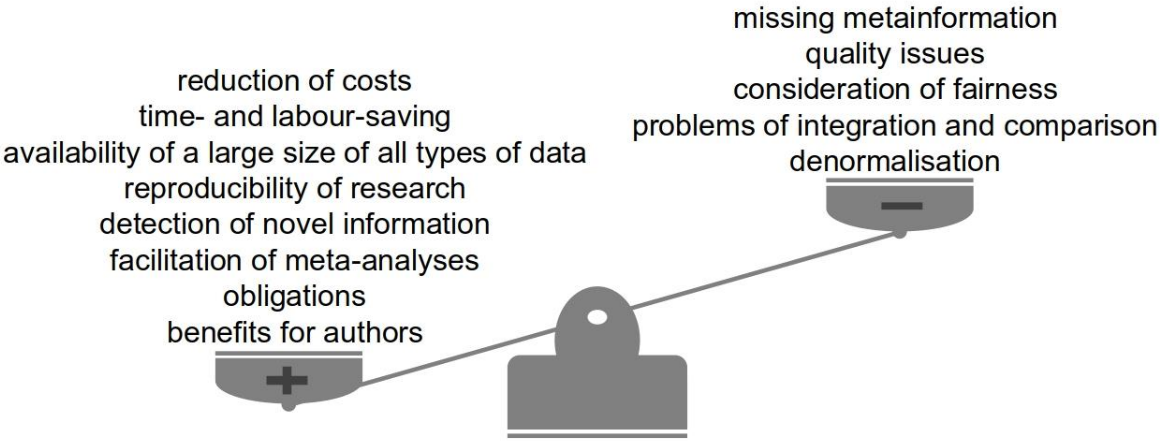

missing metainformation
quality issues
consideration of fairness
problems of integration and comparison
denormalisation
−
reduction of costs
time- and labour-saving
availability of a large size of all types of data
reproducibility of research
detection of novel information
facilitation of meta-analyses
obligations
benefits for authors
+

Figure 4: Advantages and limitations of data reuse.

Missing metainformation, quality issues and denormalisation can prevent successful reuse of public datasets (Fig. 4). Fairness must be considered in both cases: ownership of the data must be acknowledged while e.g. unnecessary additional animal experiments have to be avoided if the corresponding data already exists. Next, a large number of all sorts of data is already available and enables meta-analyses. The reuse of public scientific datasets further leads to a reduction of costs and time, encourages reproducible research, enables the detection of novel information and has benefits for authors themselves. There are still some outstanding questions and challenges (Fig. 5) but considering all the advantages and taking into account the limitations we still highly recommend and encourage data reuse.

- 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?

Figure 5: Summary of outstanding questions and challenges.

Funding

Not applicable.

Acknowledgements

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

doi:10.20944/preprints202002.0146.v1

References

1. Open Data in a Big Data World. (2016) *Chemistry International*, 38.
2. Duvallet, C., Gibbons, S. M., Gurry, T., et al. (2017) Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nat Commun*, 8, 1784.
3. McKiernan, E. C., Bourne, P. E., Brown, C. T., et al. (2016) How open science helps researchers succeed. *eLife*, 5, e16800.
4. Announcement: Where are the data? (2016) *Nature*, 537, 138–138.
5. Biological Sciences Guidance on Data Management Plans. Biological Sciences Guidance on Data Management Plans.
6. Plos-One - Data Availability. Plos-One - Data Availability.
7. Parker, T. H., Nakagawa, S. and Gurevtich, J. (2016) Open data: towards full transparency. *Nature*, 538, 459–459.
8. Piwowar, H. A., Day, R. S. and Fridsma, D. B. (2007) Sharing Detailed Research Data Is Associated with Increased Citation Rate. *PLoS ONE*, 2, e308.
9. Vasilevsky, N. A., Minnier, J., Haendel, M. A., et al. (2017) Reproducible and reusable research: are journal data sharing policies meeting the mark? *PeerJ*, 5, e3208.
10. Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*, 3, 160018.
11. Nosek, B. A., Alter, G., Banks, G. C., et al. (2015) Promoting an open research culture. *Science*, 348, 1422–1425.
12. The Beijing Declaration on Research Data. .
13. Pasquetto, I. V., Randles, B. M. and Borgman, C. L. (2017) On the Reuse of Scientific Data. *Data Science Journal*, 16, 8.
14. Denk, F. (2017) Don’t let useful data go to waste. *Nature*, 543, 7–7.
15. Arend, D., Junker, A., Scholz, U., et al. (2016) PGP repository: a plant phenomics and genomics data publication infrastructure. *Database*, 2016, baw033.
16. Wooley, J. C., Lin, H., National Research Council (U.S.), et al. (2005) Catalyzing inquiry at the interface of computing and biology. *Catalyzing inquiry at the interface of computing and biology*; National Academies Press, Washington, D.C., (2005) .
17. Pucker, B., Holtgräwe, D. and Weisshaar, B. (2017) Consideration of non-canonical splice sites improves gene prediction on the Arabidopsis thaliana Niederzenz-1 genome sequence. *BMC Res Notes*, 10, 667.
18. Schilbert, H. M., Pellegrinelli, V., Rodriguez-Cuenca, S., et al. (2018) Harnessing natural diversity to identify key amino acid residues in prolidase. *Harnessing natural diversity to identify key amino acid residues in prolidase*; preprint; Evolutionary Biology, (2018) .
19. Frey, K. and Pucker, B. (2019) Animal, fungi, and plant genome sequences harbour different non-canonical splice sites. *Animal, fungi, and plant genome sequences harbour different non-canonical splice sites*; preprint; Genomics, (2019) .
20. Protein Data Bank in Europe - Logo. Protein Data Bank in Europe - Logo.
21. Leonelli, S., Davey, R. P., Arnaud, E., et al. (2017) Data management and best practice for plant science. *Nature Plants*, 3, 17086.
22. Porto, W. F., Pires, A. S. and Franco, O. L. (2017) Computational tools for exploring sequence databases as a resource for antimicrobial peptides. *Biotechnology Advances*, 35, 337–349.

23. Leitner, F., Bielza, C., Hill, S. L., et al. (2016) Data Publications Correlate with Citation Impact. *Front. Neurosci.*, 10.
24. Ali-Khan, S. E., Harris, L. W. and Gold, E. R. (2017) Motivating participation in open science by examining researcher incentives. *eLife*, 6, e29319.
25. Liu, Q., Georgieva, D. C., Egli, D., et al. (2019) NanoMod: a computational tool to detect DNA modifications using Nanopore long-read sequencing data. *BMC Genomics*, 20, 78.
26. Liu, Q., Fang, L., Yu, G., et al. (2019) Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data. *Nat Commun*, 10, 2449.
27. Rung, J. and Brazma, A. (2013) Reuse of public genome-wide gene expression data. *Nat Rev Genet*, 14, 89–99.
28. NCBI Resource Coordinators (2017) Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, 45, D12–D17.
29. Sayers, E. W., Agarwala, R., Bolton, E. E., et al. (2019) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 47, D23–D28.
30. Sayers, E. W., Cavanaugh, M., Clark, K., et al. (2019) GenBank. *Nucleic Acids Research*, 47, D94–D99.
31. The UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47, D506–D515.
32. Spertus, J. A. (2012) The Double-Edged Sword of Open Access to Research Data. *Circ Cardiovasc Qual Outcomes*, 5, 143–144.
33. European Nucleotide Archive (ENA) ENA: Guidelines and Tutorials. ENA: Guidelines and Tutorials.
34. Bell, M. J. and Lord, P. (2017) On patterns and re-use in bioinformatics databases. *Bioinformatics*, 33, 2731–2736.
35. Bell, M. J., Collison, M. and Lord, P. (2013) Can Inferred Provenance and Its Visualisation Be Used to Detect Erroneous Annotation? A Case Study Using UniProtKB. *PLoS ONE*, 8, e75541.
36. Pearson, W. R. and Lipman, D. J. (1988) Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences*, 85, 2444–2448.
37. Cock, P. J. A., Fields, C. J., Goto, N., et al. (2010) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, 38, 1767–1771.
38. Li, H., Handsaker, B., Wysoker, A., et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25, 2078–2079.
39. Leonard, S. A. and Littlejohn, T. G. (2004) Common File Formats. *Current Protocols in Bioinformatics*, 5.
40. Ondřej, V. and Dvořák, P. (2012) Bioinformatics: a history of evolution *in silico*. *Journal of Biological Education*, 46, 252–259.
41. Zhang, H. (2016) Overview of Sequence Data Formats. In Mathé, E., Davis, S. (eds.), *Statistical Genomics*, Springer New York, New York, NY, Vol. 1418, pp. 3–17.
42. Miller, J., Georgiev, T., Stoev, P., et al. (2015) Corrected data re-harvested: curating literature in the era of networked biodiversity informatics. *BDJ*, 3, e4552.
43. Zizka, A., Silvestro, D., Andermann, T., et al. (2019) CoordinateCleaner : Standardized cleaning of occurrence records from biological collection databases. *Methods Ecol Evol*, 10, 744–751.

44. Kapushesky, M., Emam, I., Holloway, E., et al. (2010) Gene Expression Atlas at the European Bioinformatics Institute. *Nucleic Acids Research*, 38, D690–D698.
45. Abolfathi, B., Aguado, D. S., Aguilar, G., et al. (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. *arXiv:1707.09322 [astro-ph]*.
46. Chavan, V. and Penev, L. (2011) The data paper: a mechanism to incentivize data publishing in biodiversity science. *BMC Bioinformatics*, 12, S2.
47. Figueiredo, A. S. (2017) Data Sharing: Convert Challenges into Opportunities. *Front. Public Health*, 5, 327.
48. Goodman, A., Pepe, A., Blocker, A. W., et al. (2014) Ten Simple Rules for the Care and Feeding of Scientific Data. *PLoS Comput Biol*, 10, e1003542.
49. Beaufils, P. and 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.
50. Fell, M. J. (2019) The Economic Impacts of Open Science: A Rapid Evidence Assessment. *Publications*, 7, 46.
51. Open science by design: realizing a vision for 21st century research. *Open science by design: realizing a vision for 21st century research*; National Academies of Sciences, Engineering, and Medicine (U.S.), National Academies of Sciences, Engineering, and Medicine (U.S.), National Academies of Sciences, Engineering, and Medicine (U.S.), et al. (eds.); A consensus study report; The National Academies Press, Washington, DC, (2018) .
52. Farnham, A., Kurz, C., Öztürk, M. A., et al. (2017) Early career researchers want Open Science. *Genome Biol*, 18, 221.
53. Longo, D. L. and Drazen, J. M. (2016) Data Sharing. *N Engl J Med*, 374, 276–277.
54. The parasite awards - Celebrating rigorous secondary data analysis. The parasite awards - Celebrating rigorous secondary data analysis.
55. Pucker, Boas and Brockington, Samuel F (2018) Genome-wide analyses supported by RNA-Seq reveal non-canonical splice sites in plant genomes. *BMC Genomics*, 19, 980.
56. Montenegro, J. D., Golicz, A. A., Bayer, P. E., et al. (2017) The pangenome of hexaploid bread wheat. *Plant J*, 90, 1007–1013.
57. Chow, C.-N., Lee, T.-Y., Hung, Y.-C., et al. (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.
58. Du, H., Ran, F., Dong, H.-L., et al. (2016) Genome-Wide Analysis, Classification, Evolution, and Expression Analysis of the Cytochrome P450 93 Family in Land Plants. *PLoS ONE*, 11, e0165020.
59. Bowles, A. M. C., Bechtold, U. and Paps, J. (2020) The Origin of Land Plants Is Rooted in Two Bursts of Genomic Novelty. *Current Biology*, S0960982219315957.
60. Breitwieser, F. P., Lu, J. and Salzberg, S. L. (2019) A review of methods and databases for metagenomic classification and assembly. *Briefings in Bioinformatics*, 20, 1125–1136.
61. Dierckxsens, N., Mardulyn, P. and Smits, G. (2016) NOVOPlasty: *de novo* assembly of organelle genomes from whole genome data. *Nucleic Acids Res*, gkw955.
62. Yu, H. and Dai, Z. (2020) SANPolyA: a deep learning method for identifying Poly(A) signals. *Bioinformatics*, btz970.

63. Gyawali, A., Shrestha, V., Guill, K. E., et al. (2019) Single-plant GWAS coupled with bulk segregant analysis allows rapid identification and corroboration of plant-height candidate SNPs. *BMC Plant Biol*, 19, 412.
64. Marigorta, U. M., Rodríguez, J. A., Gibson, G., et al. (2018) Replicability and Prediction: Lessons and Challenges from GWAS. *Trends in Genetics*, 34, 504–517.
65. Wang, C., Shi, H., Chen, L., et al. (2019) Identification of Key lncRNAs Associated With Atherosclerosis Progression Based on Public Datasets. *Front. Genet.*, 10, 123.
66. Ma, X., Zhao, H., Xu, W., et al. (2018) Co-expression Gene Network Analysis and Functional Module Identification in Bamboo Growth and Development. *Front. Genet.*, 9, 574.
67. Persson, S., Wei, H., Milne, J., et al. (2005) Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. *Proceedings of the National Academy of Sciences*, 102, 8633–8638.
68. Zhang, N., Zhao, B., Fan, Z., et al. (2020) Systematic identification of genes associated with plant growth–defense tradeoffs under JA signaling in Arabidopsis. *Planta*, 251, 43.
69. Schaefer, R. J., Michno, J.-M., Jeffers, J., et al. (2018) Integrating Coexpression Networks with GWAS to Prioritize Causal Genes in Maize. *Plant Cell*, 30, 2922–2942.
70. Kwon, M. J., Oh, E., Lee, S., et al. (2009) Identification of Novel Reference Genes Using Multiplatform Expression Data and Their Validation for Quantitative Gene Expression Analysis. *PLoS ONE*, 4, e6162.
71. Cheng, W.-C., Chang, C.-W., Chen, C.-R., et al. (2011) Identification of Reference Genes across Physiological States for qRT-PCR through Microarray Meta-Analysis. *PLoS ONE*, 6, e17347.
72. Hruz, T., Wyss, M., Docquier, M., et al. (2011) RefGenes: identification of reliable and condition specific reference genes for RT-qPCR data normalization. *BMC Genomics*, 12, 156.
73. Pucker, B., Feng, T. and Brockington, S. F. (2019) Next generation sequencing to investigate genomic diversity in Caryophyllales. *Next generation sequencing to investigate genomic diversity in Caryophyllales*; preprint; Genomics, (2019) .
74. Keilwagen, J., Hartung, F. and Grau, J. (2019) GeMoMa: Homology-Based Gene Prediction Utilizing Intron Position Conservation and RNA-seq Data. In Kollmar, M. (ed.), *Gene Prediction*, Springer New York, New York, NY, Vol. 1962, pp. 161–177.
75. Winter, D., Vinegar, B., Nahal, H., et al. (2007) An "Electronic Fluorescent Pictograph" Browser for Exploring and Analyzing Large-Scale Biological Data Sets. *PLoS ONE*, 2, e718.
76. Sheehan, H., Feng, T., Walker-Hale, N., et al. (2019) Evolution of L - DOPA 4,5-dioxygenase activity allows for recurrent specialisation to betalain pigmentation in Caryophyllales. *New Phytol*, nph.16089.
77. van Wijk, K. J., Friso, G., Walther, D., et al. (2014) Meta-Analysis of *Arabidopsis thaliana* Phospho-Proteomics Data Reveals Compartmentalization of Phosphorylation Motifs. *Plant Cell*, 26, 2367–2389.
78. Brinkrolf, C., Henke, N. A., Ochel, L., et al. (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.
79. Toubiana, D., Puzis, R., Wen, L., et al. (2019) Combined network analysis and machine learning allows the prediction of metabolic pathways from tomato metabolomics data. *Commun Biol*, 2, 214.

80. Ubbens, J., Cieslak, M., Prusinkiewicz, P., et al. (2018) The use of plant models in deep learning: an application to leaf counting in rosette plants. *Plant Methods*, 14, 6.
81. Pound, M. P., Atkinson, J. A., Townsend, A. J., et al. (2017) Deep machine learning provides state-of-the-art performance in image-based plant phenotyping. *GigaScience*, 6.
82. Mooij, W. M., Trolle, D., Jeppesen, E., et al. (2010) Challenges and opportunities for integrating lake ecosystem modelling approaches. *Aquat Ecol*, 44, 633–667.