Exploring how technical and social advances can overcome the hurdles of reproducibility in the life sciences

A thesis submitted to the School of Biology at the University of East Anglia in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Evanthia Kaimaklioti Samota

August 2022

Abstract

Reproducibility is an important element of robust science. Reproducibility is vital for the better understanding, validation, and re-use of research. For the past three decades, literature on the *reproducibility crisis* has increased, highlighting how researchers cannot rerun the analysis of other researchers and reach the same results.

To achieve research reproducibility, there is a need for the raw data, the analysed data (including negative data), the code and complete methodology, the detailed analysis protocol, the standardised data and metadata annotation, to be well documented, and shared so they can be easily accessed by other researchers who wish to reproduce the work. However, with the increasing volume of data, given the advances in life sciences technology, there are still issues with reproducibility, despite the many tools, data and code-sharing mandates created to manage the irreproducibility problem.

This thesis explores how technical and cultural advancements can address irreproducibility issues. The study aims to understand how tools and cultural factors (e.g., training, incentives and rewards for reproducible research practices) promote research reproducibility.

A survey of 251 researchers from various backgrounds reported their knowledge of reproducibility issues in the life sciences, their ability and motivation to reproduce research studies and their opinions on the interactive representation

of research results (in interactive figures instead of static figures) within journal articles.

Additionally, this thesis investigates how interactive figures could reproduce computational experiments presented in the figures and their benefits and limitations in improving research reproducibility.

Lastly, this thesis presents a software prototype, *Deus ex machina*, which automatically annotates articles and their metadata with standardised semantic information (plant ontology terms and IDs). *Deus ex machina* computes a reproducibility metric score that evaluates the reproducibility status of papers, ultimately recognising reproducible research. The software can thereby serve as a means of promoting a more reproducible research culture.

Contents

List of Figures

List of Tables

List of Abbreviations

API Application Programming Interface. 60, 147

BD Big Data. 30, 32

CE Computational Experiments. 26–28, 30, 36–38, 42, 46, 80, 83, 84, 87, 95–98, 100, 109, 113, 197

CO Crop Ontology. 134, 135, 178

DMP Data Management Plan. 31, 32

DOI Digital Object Identifier. 36, 76

DPA Days Post Anthesis. 153

DS Developmental Stage. 98

EBI EMBL European Bioinformatics Institute. 135, 146, 147, 158, 159, 169

EI Earlham Institute. 95

ENA European Nucleotide Archive. 120, 164, 171

ERA Executable Research Article. 107–109, 111, 197

expVIP Expression Visualisation and Integration Platform. 96–100, 102, 103, 105, 106, 115, 121, 122, 127, 128, 133–135, 139, 167

FAIR Findable, Accessible, Interoperable, Reusable (guiding principles). 33, 86, 181, 196, 199

FN False Negatives. 150, 151, 163

FP False Positives. 150, 151, 163

GDPR General Data Policy Regulation. 95

ID Identifier for ontologies. 129, 130, 133, 135, 139, 141–144, 152–155, 159, 162, 166, 167, 170, 176, 197

IT Information Technology. 44, 45

LS Life Science. 20, 22, 24–27, 30, 32, 35, 36, 38–41, 49, 51, 52, 54, 55, 75, 78, 84, 89, 92, 110, 113, 114, 116, 123, 124, 128, 165–167, 172–175, 178, 182, 185–187, 193, 195, 198

NBI Norwich Bioscience Institutes. 9, 55, 56, 59, 61, 63, 64, 71–74, 77, 78, 94, 95, 117

NCBI National Center for Biotechnology Information. 35, 40, 121

NLP Natural Language Processing. 141, 143, 164

OBO Open Biological and Biomedical Ontologies Foundry. 123, 124, 131, 142

OLS Ontology Lookup Service. 124, 153, 154, 170, 172

OS Operating System. 42, 44, 104

OWL Web Ontology Language. 131

PAN ENA Project Accession Number. 121

PO Plant Ontology. 125, 126, 129, 131–135, 139, 141, 143, 145, 146, 150–153, 155, 156, 159, 162–164, 167, 169, 172, 176, 197

POC Plant Ontology Consortium. 129, 130, 167, 176

RDF Resource Description Framework. 131

RMS Reproducibility Metric Score. 10, 22, 23, 119, 148–150, 155–157, 161, 169, 170, 178, 179, 187, 188, 195, 196, 198

RO Research Object. 38

SRA Small Read Archive. 40, 121

TGAC The Genome Analysis Centre. 95, 96

TN True Negatives. 151

TP True Positives. 150, 153, 163

UEA University of East Anglia. 95

URI Universal Resource Identifier. 129

VM Virtual Machine. 36, 42–44, 97, 98, 101, 104, 117

WMS Workflow Management System. 36–38, 40, 198

Dedicated to

My dear children

Acknowledgements

I would like to thank my supervisor, Dr Robert P. Davey, for his insight and support throughout the years.

I am also grateful for the support and guidance of Prof Vincent Moulton, my secondary senior supervisor.

I would like to acknowledge the Journal *eLIFE* and the Biotechnology and Biological Sciences Research Council (BBSR) for funding my PhD.

I must extend my gratitude to all the Davey group colleagues with whom I have had the pleasure to share an office over the years; while there are too many names to list here, I want to record my appreciation to you all for providing a supportive and welcoming space.

My sincere thanks also to the Earlham Institute for creating a friendly working environment. I would like to acknowledge all the staff - the cleaners, the guards, the business support, the other students, the researchers and the executive team - for making this journey pleasant.

The University of East Anglia (UEA) provided me with this opportunity to embark on a PhD and improve my research skills and offered a lovely environment to study. Special thanks to the UEA Student Support office, the Norwich BioScience Institutes' Graduate Student office and, in particular, to Dr Kate Conway for her support and guidance.

Many thanks to Dr Louise Duckling for her invaluable help proofreading my thesis, her writing tips throughout our collaboration and her support and encouragement.

Special thanks to The Enterprise Centre at UEA for their resources which helped me set up and grow my small (for now) business, Piotis Naturals. Spending time with other young entrepreneurs and receiving mentorship from seasoned entrepreneurs and business mentors has been a great pleasure and immense help. I want to especially thank Mr Finbarr Carter for being the heart and soul of The Enterprise Centre. I am indebted to the UEA, The Enterprise Centre and Santander Bank for awarding £5000 in business funding to my business, Piotis Naturals, following my success in winning their business competitions.

I want to thank all of the clients (both online and offline) and followers of Piotis Naturals for their support over the years of my business, for their encouragement for my business and my PhD, and understanding when sometimes I couldn't fulfil their orders as fast as I would have liked.

I want to thank the many friends I met in Norwich and all my friends from back home in Cyprus: you are too numerous to mention, but you have been awesome.

Of course, I want to acknowledge my family for all their support and encouragement all these years. I love you all. I want to especially thank my kids. They have been my motivation to keep working on my PhD, as I want to make them proud and inspire them. I want them to know that they can work to achieve the goals they set for themselves, should they work hard and believe in themselves and in the power that God gives them.

Lastly, I want to thank in advance the examiners of my PhD for agreeing to give their time to review this thesis. I hope you will find it interesting, informative and enjoyable. I am looking forward to meeting with you and discussing my research.

I want to end with the opening stanza of my favourite poem, 'Ithaki', by the Greek

poet, Constantinos P. Kavafis:

Σα βγεις στον πηγαιμό για την Ιθάκη, να εύχεσαι νάναι μακρύς ο δρόμος, γεμάτος περιπέτειες, γεμάτος γνώσεις.

As you set out for Ithaka, hope that your journey is a long one, full of adventure, full of discovery.

(For a beautiful recital of the English version of this poem by the actor Sean Connery to the music of Vangelis, visit this Youtube link).

I believe this poetic stanza summarises my PhD journey. It has been long and challenging but filled with discovery and knowledge. I feel blessed that this opportunity has helped me to grow as a person, in wisdom, experience and strength.

1

Introduction

1.1 Motivation

The Cambridge Crystallographic Data Centre, the main repository for small-molecule crystal data, will soon remove almost 1,000 deposited crystal structures as they seem to have been falsified (Lowe, 2022).

A report from the journal *Science* documented fabricated images in prominent neuroscience papers on Alzheimer's disease and beta-amyloid peptides (Piller, 2022), undermining decades of research based on the falsified findings.

Hundreds of people suffered heart attacks and heart failures over several months whilst taking the diabetes drug Avandia, which was linked to 304 deaths during the third quarter of 2009 before it was removed from the market. This issue arose because the manufacturing pharmaceutical company failed to report safety studies to the FDA (Gardiner, 2010).

In 2011, social psychologist Diederik Stapel admitted having published fraudulent data in 30 peer-reviewed papers (Gardiner, 2012).

Twenty-seven per cent of childhood disease mutations cited in literature were later discovered to be misannotated or common polymorphisms, underpinning the necessity for data to be reproduced before mutation databases are populated (Colhoun et al., 2003).

The US $1.6-million Reproducibility Project for Cancer Biology attempted to reproduce 50 high-impact published cancer research studies. However, given the difficulties, they terminated their efforts at 23 papers (Morrison, 2014).

These are just some examples of how science has experienced issues regarding reproducibility, where reproducibility is believed to be a core aspect of scientific research. Research reproducibility is defined as achieving the same findings using the same data and analysis when the experiment is performed by researchers other than the original researchers (Claerbout and Karrenbach, 1992). Despite the importance of reproducibility, research across many disciplines ranging from the Life Sciences (LSs), to computer science, to economics, is riddled with irreproducible studies (Begley and Ellis, 2012; Christensen and Miguel, 2018; Crick et al., 2017; Fraser et al., 2018; Ioannidis, 2005; Ioannidis et al., 2017).

Given these seemingly wide-ranging problems across all disciplines, many technologies have been developed to facilitate and improve research reproducibility (or to enhance certain aspects). Similarly, policies have been developed explicitly to mandate better research reproducibility from researchers and publishing journals.

Research reproducibility issues have many facets. It is difficult to find a remedy using a single technological solution, hence why most technologies address research irreproducibility one aspect at a time. Neither one technology, one social advance, or one approach can be the panacea to research reproducibility.

We must understand how it is best to overcome the hurdles of reproducibility so that technological solutions can exist to facilitate the work of researchers and publishers to produce and disseminate research reproducibly and so that, importantly, we recognise, reward and incentivise reproducible research. In other words, while we explore technological solutions, we must also consider how these can operate synergistically with social advances to tackle research irreproducibility more efficiently, generating a holistic approach that would be

valuable to the scientific community. This is not a small undertaking, however, thus my research focuses on two main elements: assessing the use of technological solutions to improve reproducibility and; using a proof-of-concept tool to demonstrate the difficulties with solving problems in reproducibility, as many stem from social issues surrounding typical article and data publishing practices. I aim to reinforce the notion that technological and social solutions are required to ensure that research is more reliable, robust, reproducible and reusable.

1.2 Contribution

Within this thesis, I propose that there needs to be an integrated approach to tackling research reproducibility issues, whereby both technical and social solutions need to be implemented in tandem to address the difficulties currently faced with research reproducibility.

With the research survey I conducted (Samota and Davey, 2021), I strive to show how scientists view the reproducibility issue, record their opinions on potential solutions, and quantify and qualify for the first time the knowledge and opinions of researchers on interactive figures (compared to traditional static figures) as a means of enabling reproducibility of experiments within research articles.

With the presentation of prototypes of interactive figures, I endeavour to demonstrate how such technological solutions could help with certain aspects of research reproducibility and highlight the limitations of interactive figures, especially in terms of scaling up and maintaining the technology.

I aim to present a software prototype demonstrating researchers' issues when mining literature for information through automatic semantic annotation of publications and their associated data and metadata files in crop transcriptomics. Using this approach, with the computation of the

Reproducibility Metric Scores (RMSs) and a comparison of how different papers perform, I aim to propose an assessment mechanism for the reproducibility of research papers and related data artefacts in the hope of demonstrating practically how research can become more reproducible. Using this tool prototype, I will exhibit how such technological solutions can help researchers and journals present research more reproducibly and reduce the time researchers need to access information and related data.

Lastly, I aim with the overall discussion components and conclusions of this thesis to inspire the scientific community and associated policymakers to see the value in conducting and presenting research reproducibly and, with the use of the proposed technologies and social factors, to create a more sustainable research culture that has reproducibility at its core, and that rewards and recognises reproducible research.

1.3 Structure of the Thesis

Chapter 2 depicts a more thorough discussion of the field of reproducibility, its importance to LSs, and a review and discussion of existing work on reproducibility and other aspects related to this thesis.

Chapter 3 presents my published survey that assesses researchers' opinions about reproducibility and their perceptions of interactive figures within journal articles as a solution to reproducing computational experiments.

Chapter 4 includes a more thorough explanation of interactive figures and presents a prototype figure framework, as well as a review and discussion of other projects in interactively presenting experiments within journal articles. I discuss the technical and social limitations of interactive representation of figures or experiments within journal articles in being able to address the research reproducibility issue.

Chapter 5 explores the issue of the lack of standardised terminologies in research articles and repositories such as databases and how this negatively impacts reproducibility. Furthermore, I present a software prototype that addresses some of the aforementioned issues, using real-world use cases from researchers in crop transcriptomics. Moreover, I propose the concept of an RMS as a means of assessing how reproducible a study is and how RMS could be used as a new, additional scoring mechanism to assess the value of a published research study beyond the existing traditional published research metrics (such as citation scores).

Chapter 6 concludes this thesis and discusses the field of reproducibility in more broad ways in topics that were not directly investigated here. Aspects of Chapter 6, in the form of a commentary, bring together the topics discussed in this thesis and additional important notions related to research reproducibility, including the limitations and difficulties in achieving reproducibility. Finally, I indicate areas that can be further explored through future research.

2

Reproducible Science

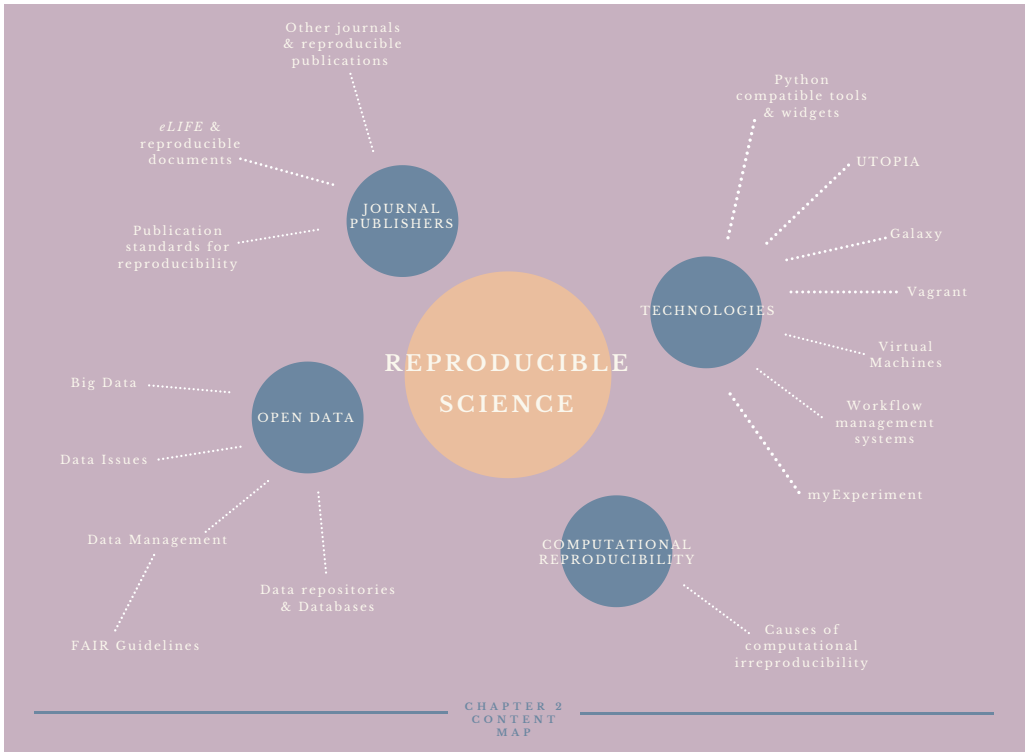

Figure 2.0.1: *Concept Map for Chapter 2: Reproducible Science*

Synopsis

This chapter aims to describe the causes of the irreproducibility of experiments in LSs; delineate the most important technologies addressing the irreproducibility of experiments in LSs and discuss the gaps in the current efforts addressing the reproducibility crisis.

2.1 Introduction

Reproducibility is a key concept in scientific research for producing reusable software and better datasets. It refers to the ability of other researchers to produce the same findings without altering the original data and analyses (Claerbout and Karrenbach, 1992). Research suggests that the irreproducibility of experiments is a significant problem (Baker, 2016; Stupple et al., 2019; Miyakawa, 2020). It is hypothesised that reproducible science leads to robust and quality-oriented scientific work.

Various technologies are available for promoting and enabling reproducible science, with a variable uptake by the LS community. The technologies will be discussed in more detail later in this chapter.

2.1.1 Definitions of Reproducibility

The terms *repeatability*, *replicability* and *reproducibility* are commonly confused (Peng et al., 2006; Liberman, 2015), thus it is imperative to differentiate these terms from each other. Definitions of the terms were included in my published paper of which I was the first author, and I have reproduced these below with permission (Samota and Davey, 2021):

1. **Repeatability** The original researchers using the same data, running precisely the same analysis and getting the same results, on multiple runs (Drummond, 2009).

2. **Replicability** Different teams performing different experimental setups and using independent data, achieving the same results as the original researchers, on multiple trials (Peng et al., 2006; Peng, 2011; Stodden et al., 2013a).

3. **Reproducibility** Different teams re-running the same analysis with the

same data and getting the same results (Claerbout and Karrenbach, 1992; Peng et al., 2006; Peng, 2011; Stodden et al., 2013a).

For the purposes of this chapter, I will present these terms with respect to the concept of computational reproducibility.

2.2 Computational Reproducibility

Computational reproducibility, or reproducible computational research, is the term used to describe the reproducibility of Computational Experiments (CEs). CEs involve steps that require computer(s), and the inputs and outputs are represented digitally. These experiments are performed in disciplines such as bioinformatics, computational biology, systems biology and many others. In this chapter, and in line with the scope of this thesis, the term CEs refers to computational experiments in the LSs. Therefore, *computational reproducibility* is achieved when an independent team creates the same results using the computational methods and data of the original authors (Donoho, 2010; Stodden et al., 2013a; Stodden and Miguez, 2013; Stodden et al., 2018; Leipzig et al., 2020).

2.2.1 Causes of Irreproducibility of Computational Experiments

Research shows that the irreproducibility of Computational Experiments (CEs) is a significant problem that requires appropriate solutions (Schwab et al., 2000; Peng, 2011; Bechhofer et al., 2013). Irreproducible CEs fail to preserve all factors that ensure computational reproducibility, including the independent replication and validation of the experiments (Müller et al., 2003). Statistical analyses, including incorrect statistical tests and an underpowered number of control and test samples, can lead to incorrect or misleading results

and false conclusions that further impede the reproducibility of CEs (Stodden, 2015). Whilst conventions exist to encourage researchers to share their code and data, and documentation standards exist for describing code or analytical methodology, problems still arise as many computational and data-enabled experiments do not have corresponding standards for code and data quality that are fit for purpose. Demarcations such as results' verification, error quantification and validation are not precisely defined in scientific computing analyses (Stodden et al., 2013b).

Science in this field has to be transparent to ensure the reproducibility of CEs (Iqbal et al., 2016). Cherry-picking, fabricating, falsifying and altering data, although rare, could be considered scientific misconduct (Fanelli, 2009) and lead to irreproducibility issues and increased retraction rates (Cokol et al., 2008). Beyond these factors, and given the critical use of computation in modern LSs, the lack of proper data documentation and standards for metadata (the data about the data) annotation or the lack of metadata availability contribute to the irreproducibility of CEs (Teixeira da Silva, 2015; Huang and Gottardo, 2013). The extent to which these problems arise at the publishing stage will be discussed later in this thesis.

2.3 Open Data

Data is a term used inconsistently in science; however, it is important to state that data sharing in science can include raw and processed data (also referred to as computed or simulated data/data after analysis). Computationally reproducible experiments require datasets that reflect the original author's work. Reports have stressed that open, publicly available data are vital for reproducible science (Molloy, 2011). Data publicly available and reproducible allow for an assessment of whether the data are suitable to address a hypothesis or power analysis, thus leading to reproducible science. Published data should

be robust, stable, and distributed in a way that allows the data and experiments to be reproducible (Pulverer, 2015). The persistent identifiers and standardised descriptions of the methods and tools used to produce those datasets lead to more reproducible CEs (Teytelman and Stoliartchouk, 2015; Teytelman et al., 2016). Papers which publish their data and analyses are cited more, independent of the journal's impact factor or other factors such as the date of publication (Piwowar et al., 2007). Initiatives such as DataCite (Brase, 2009), which acts as a centralised resource for generating persistent identifiers and thereby promote consistent and globally recognised citation and attribution of data, can encourage data sharing (Bechhofer et al., 2013). Tracking data provenance can also help find errors or the source of unexpected results within a workflow (Buneman et al., 2000).

Data and Metadata Sharing

By itself, therefore, Open Data is not strictly sufficient to produce reproducible CEs. Associated contextual information must be provided and be well described, i.e. in the form of metadata. Metadata is data that gives information about other data.

Nevertheless, several factors dissuade authors from sharing their data. Annotating a dataset with metadata is time-consuming since the data must be appropriately formatted and annotated. Other complexities include not knowing the most suitable place to publish the metadata, as both the supplementary information cited and the websites of the laboratories may only be temporarily available (Santos et al., 2005). Moreover, the re-analysis of the data by scientists other than the authors could lead to different conclusions from those published by the authors, possibly owing to mistakes in the study (Ioannidis, 2005). For some scientists, publishing their data is perceived as losing competitiveness in the field, or they fear that their intellectual property could be used for profit-making (Pitt and Tang, 2013; Piwowar et al., 2007). Fecher et al. (2015) investigated six categories (and their subcategories) of

factors that can influence academic data sharing through a survey and mining of 630 research articles. They concluded that a range of factors, such as socio-demographic status, degree of control over data sharing, resources needed and returns for sharing, affect data, metadata and code sharing decisions.

Databases as Open Data Repositories

Are databases able to hold and maintain data in ways that are always retrievable? Disciplines such as systems biology, which depends on acquiring data from many heterogeneous molecular biology experiments, involve searching through several databases (Pennisi, 2005), which can be time-consuming, difficult to homogenise due to differences in data structure and metadata, and therefore difficult to analyse. In certain cases, one cannot query all database fields nor fully download the database contents (Philippi and Köhler, 2006). Database information needs to be stored and organised in accessible ways (Attwood et al., 2009) and databases should preferably hold raw and processed data (Anon, 2013) so that it can be better curated and understood.

Data in databases must be associated with good descriptive metadata to be able to understand the data fully. Metadata could be information about the format and the contents of the database, the methods used to produce the data before entry into the database, any recommendations for its curation, the structures in which data are stored, any information about the management of the database once released, and the database version used (Gehani et al., 2011).

Understanding the data in databases is further complicated when non-informative names do not comply with consensus nomenclature and ontology standards (Ashburner et al., 2000). This issue is covered in detail in the Chapter 5. Invalid data and errors can be propagated between databases (Karp, 1998). A solution is to use evidence codes to track how annotations were

created and prevent the automatic inference of annotations unless they were manually curated (Philippi and Köhler, 2006). Arguably, evaluating the quality and correctness of data in databases is difficult.

2.3.1 Big Data

Big Data (BD) is the term used to describe the increasingly large, complex and difficult-to-process datasets generated by the current research (not only in the LSs). However, this thesis will be referring exclusively to BD in the LSs). As a result of the aforementioned characteristics, challenges arise with processing, storing and distributing BD datasets, e.g. excessive costs and the need for appropriate infrastructures are contributing to the CEs irreproducibility (Marx, 2013).

Big Data has become somewhat of a "buzzword" in recent years, with academia and industry using it as a proxy for five key aspects of data life cycles: volume, velocity, variety, veracity and value. These are the five key elements that make Big Data a huge business. Based on these definitions, and given that nowadays, most data is shared or published on the internet, the issues surrounding the scale and complexity of BD provide additional issues for reproducibility.

2.3.2 Data Management

Because BD is now commonplace, and CEs are somewhat universal, the importance of managing data effectively is not underestimated. Data management defines the policy, processes and techniques that determine how data should be generated, stored, analysed and shared. Researchers often need guidance in understanding their data management responsibilities, either from a project or funding perspective, i.e. to determine what a project investigator expects from data generation or analysis experiments or what a funder expects from the projects they fund.

Each grant application should include Data Management Plans (DMPs), as set by the National Science Foundation in January 2011, the Research Councils UK in 2011 and the FAIR Guiding Principles (Wilkinson et al., 2016). Below are some DMP rules, proposed by Michener (2015), that can facilitate reproducibility. DMPs should determine:

The nature of the data to be gathered

- **Type** The type defines whether it is a spreadsheet, software, code, and the experiment parameters.
- **Data Sources** Sources include human observation, a laboratory, and the volume of the data.
- **Data and file formats** The best suggestions are open source, non-proprietary, and Comma Separated Values (CSV) files instead of Excel files.
- **How metadata will be used** Provide information to explain how data were discovered, captured, made sense of, and suitably cited; explain how files were named, constructed and stored; and document the experimental details.
- **Metadata managing software tools** For example, Metavist is a potential management tool.
- **Means that assure data quality** Data Quality Assurance and Quality Control involve computation, appraisal and regulation of a study' s products (e.g., data, software). Quality Assurance and Quality Control might involve calibrating instruments or using statistical and visualisation methods to reveal errors.
- **Data storage plan** A storage plan considers the length of time the data will be available, the storage means, and whether data will be secured and accessible for the future. URLs break, disks fragment, and papers get lost.

> Thus, backing up data using remote storage services such as Amazon or data repositories such as Figshare and Dryad is good practice. I will explain more about databases and data repositories later in the chapter (see Section Databases and Data Repositories).

It is also suggested as good practice for DMPs to be produced by researchers for their projects outside the remit of a funder. This can help those researchers maintain good practice for their data reproducibility and for any staff they employ (Jagadish and Olken, 2004). There are efforts to bring together communities that work on data management practices as part of initiatives such as The Turing Way.

2.3.3 Data Issues in the Life Sciences

The three most important data issues in LS, as identified by Thessen and Patterson (2011), are as follows: 1) lack of standards; 2) lack of incentives for scientists to share data; 3) lack of adequate (technological) infrastructure and support.

Given the matters around BD and Open Data, and with the motives of sharing data reproducibly, for the research results to be shared and reused effectively, the data needs to be structured in a specific manner (Jagadish and Olken, 2004; Hollmann et al., 2018). Data should be created, formatted and deposited in data repositories in line with DMPs and Standard Operating Procedures (General Secretariat of the Council, 2016; Hollmann et al., 2018). For effective and reproducible data sharing, appropriate cyber-infrastructure is necessary, and scientists must be motivated to share data openly and reproducibly. Sociological issues affect the data acquisition, curation, preservation, sharing and reuse of data. To facilitate reproducible data practices, LS researchers and computer programmers must collaborate effectively to produce tools to encourage a high uptake by the community.

The FAIR Guidelines

Scientific data management can benefit from guidelines and standardised processes that attempt to harmonise best practices around implementing reproducible tool development and data sharing policy.

The principles of Findability, Accessibility, Interoperability, and Reusability (FAIR) (Wilkinson et al., 2016) have been developed as a framework for improving the reuse of scholarly data, as they provide measurable standards. They have been adopted (to various degrees) by various key players from academia, industry, funding agencies, and scholarly publishers. Implementing the FAIR principles when developing scientific software and generating data is assumed to improve the retrieval, search and use of the data by machines, in addition to helping data producers and publishers (Wilkinson et al., 2016).

The FAIR Guiding Principles apply to all "scholarly digital research objects from data to analytical pipelines", computational tools and software (Wilkinson et al., 2016). They are summarised in Box 2.1, reproduced with permission from Wilkinson et al. (2016).

The absence of suitable technology means it can take several weeks or months to use specialised tools to collect the necessary data to answer difficult research questions. Instead, a better means of economising time and resources would be to create and preserve digital research objects.

The FAIR guidelines aim to set out a policy for all types of scholarly digital objects to become important elements in scientific publications, where the value of a publication and its impact is determined by how precisely it is found, re-utilised, and cited by all stakeholders, both human and machine.

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

InfoBox 2.1: *The FAIR Guiding Principles, reproduced from (Wilkinson et al., 2016).*

Databases and Data Repositories

Specialised repositories have been produced through many iterations of computational and/or manual curation and therefore form a "gold standard". They tend to hold significant digital objects such as species reference datasets and be easily accessible by humans and computers. In some cases, they seek to integrate or harmonise the uploaded data so that it can be compared directly with other datasets, e.g. the Ensembl database. Such repositories include:

- **Ensembl** is a genome database project at the European Bioinformatics Institute that was started in response to the Human Genome Project nearing completion in 1999. Ensembl intends to be a central resource for

geneticists, molecular biologists, and other academics and researchers working with human genomes and the genomes of other vertebrates and model organisms.

- **Uniprot** is a publicly available protein sequence database with a large number of entries obtained from genome sequencing studies. It provides extensive information about the biological function of proteins collected from the research literature.

- **Genbank** contains all publicly accessible nucleotide sequences and their protein translations in an openly accessible format. The National Center for Biotechnology Information (NCBI) is part of the International Nucleotide Sequence Database Collaboration and provides sequence data for this organisation.

- **Worldwide Protein Data Bank (wwPDB)** is an organisation that archives macromolecular structural data, i.e. information about the 3D structure of nucleic acids, proteins and complex assemblies. The Protein Data Bank archive aims to maintain a single central repository of macromolecular structural data, which is provided to the global science community freely and publicly.

Unfortunately, not all repositories and databases in the LSs hold well-curated, annotated with the correct metadata and properly deposited data, which hinders reproducibility of the data involved (Attwood et al., 2009).

Certain LS databases have been designed to hold specific data types from specific data generation technologies, e.g. DNA sequencing. More general-purpose data repositories attempt to capture the wider scope of data collected within LS experiments. Specifically, some examples of general purpose data repositories are Figshare, Zenodo, Dataverse, Dryad, Mendeley Data and DataHub. Such repositories hold various data types and formats, so they impose minimal constraints on the data descriptors (metadata) because

their core aim is to store data, unstructured or otherwise. Therefore, these repositories help by providing large-scale data storage and extra services such as Digital Object Identifier (DOI) allocation. Still, their lack of structure and standardisation makes them more diverse and less interconnected. Thus they exacerbate the problem with Big Data issues such as data discovery and reuse.

2.4 Technologies and Projects addressing the Reproducibility Problem

Reproducibility is inherently linked with the quality of data analysis methods and their accurate description. Established projects, software and infrastructures are working towards supporting accessible, reproducible, and transparent computational research in LSs (Goecks et al., 2010). The technologies available to group the data, software and dependencies of experiments include Virtual Machines (VMs), Workflow Management Systems (WMSs) and container systems such as Docker and Singularity. Such technologies will be described and compared in the following sections.

Workflows include the chain of actions in a CE. Workflows can be expensive and difficult to design, involving many development hours and skilled personnel to implement and optimise. This difficulty is compounded by scientific software that often undergoes rapid development and, therefore, may require the implementation of workflow changes. That said, workflows are useful as they can be reused, re-purposed and recycled so that other scientists can execute them with their data and possibly extend the workflow to perform their experiments, thus rendering workflows as repeatable and potentially reproducible tools. They typically enable the execution of one or more pieces of scientific software via a user interface. Certain workflows demand local installation of software infrastructures, such as the R statistical tool or Python, which can sometimes be difficult to achieve for scientists with limited

programming skills. Thereafter, it is not enough to simply publish and share workflow specifications (Bechhofer et al., 2013): having good WMSs is vital for supporting CE reproducibility.

2.4.1 Workflow Management Systems

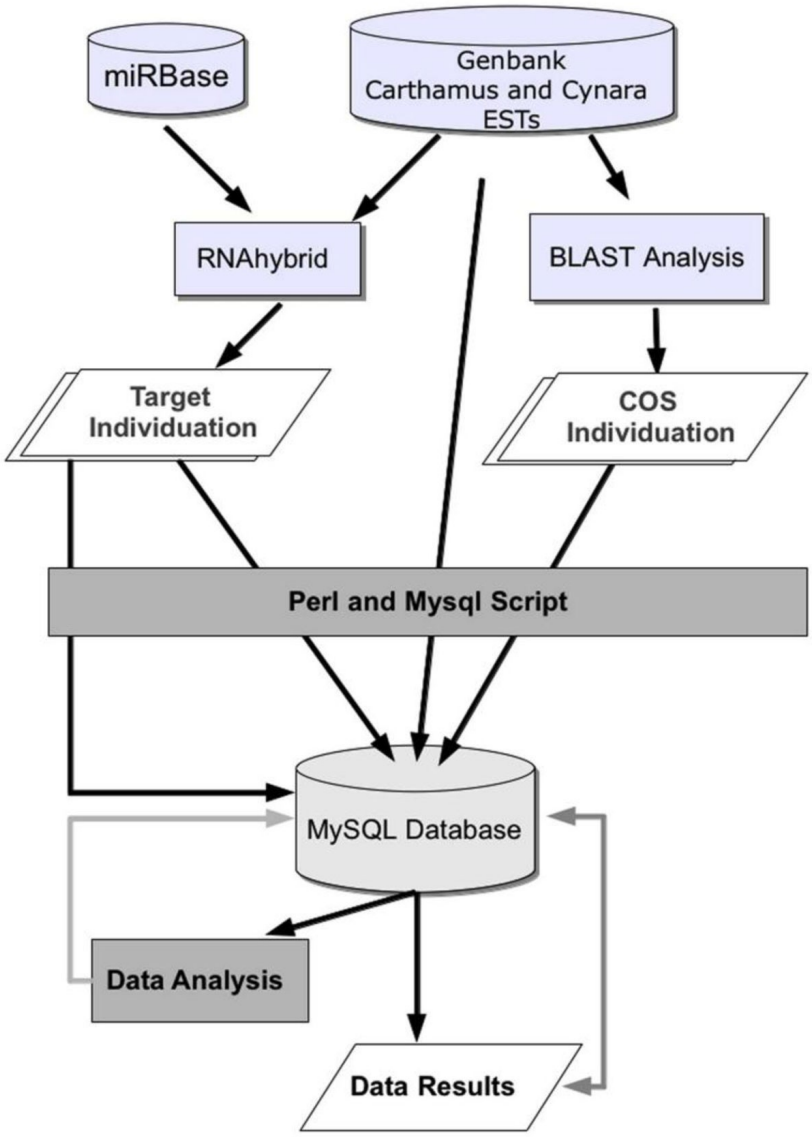

Figure 2.4.1: *Workflow of the computational approach followed to identify miRNAs. The overall procedure for miRNA, target and homologous regions identification.* The figure and its original caption are reproduced verbatim from (Catalano et al., 2012)

Figure 2.4.1 shows an example workflow of the computational approach followed to identify miRNAs. The image reproduced here with permission from (Catalano

et al., 2012).

De Roure et al. (2011) presented a concept of preserving scientific workflows by using Research Objects (ROs) as WMSs. ROs make CEs more reproducible since pieces of data, files or database entries possess individual identifiers that can be utilised in subsequent experiments. The need for ROs was brought about from the observation that workflows and their data description decay over time. ROs include: (1) the "nodes of operation", representing the local or remote analysis steps, and (2) the "edges", referring to the dependencies amongst the operations. Ultimately, with ROs, workflows can be better understood, preserved and re-purposed, and the workflows' additional resources can be more easily accessed. ROs allow for the adding of background metadata information and the scientist's explanation of the results, as well as the tracking of the ROs' evolution with ontologies (Belhajjame et al., 2015).

Running the workflow (in silico) produces final and intermediate results, any provenance information around the services used, and metadata describing the RO structure and the relationships between its containing objects. Ultimately, the creators of ROs admit that this technology cannot guarantee complete reproducibility, but it is another approach to facilitating it.

Galaxy

Galaxy is an open Web-based scientific workflow management system that aims to make computational research in the LSs accessible to researchers without programming or systems administration experience (Giardine et al., 2005). These systems typically provide a graphical user interface that enables users to specify what data to collect and what steps they should take. The Galaxy workflow systems are typically used for multi-step computational analyses (see Figure 2.4.2).

Galaxy is mainly used for genomics research, such as next-generation sequencing, but it is a domain-agnostic system used in many scientific areas,

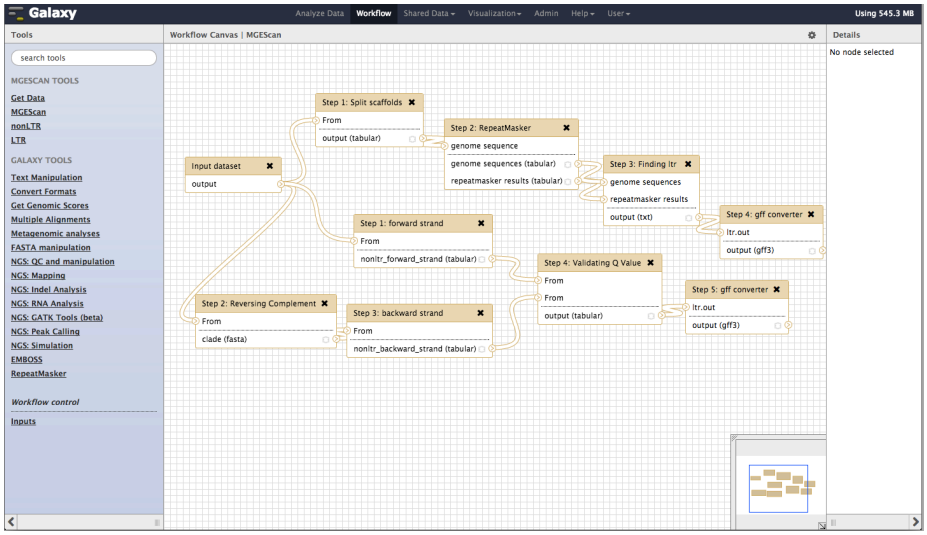

Figure 2.4.2: *Illustration of a MGEScan workflow in the Galaxy workflow editor.* Image source: https://mgescan.readthedocs.io/en/latest/workflow.html

not just in the LSs. It supports data provenance (also referred to as data lineage, i.e. the data's origin), the maintenance of computational workflows and a variety of widely used text manipulation utilities, allowing users to perform their customised interaction with the workflow system.

Workflows can be re-purposed, manipulated, extended and reused by the workflow creators or others. Galaxy is widely accepted within the bioinformatics community, receiving thousands of computational jobs per day alone on the main public server (Goecks et al., 2010), and with many other interactions on public and private installations around the globe. Through the Bioconda software package manager, Galaxy provides access to computational analysis tools, bypassing the common problems experienced when installing tools locally. Galaxy then supports the combining of individual tools within larger analysis workflows and the modification of parameter values.

Through its History system, Galaxy supports reproducibility and enables transparent research as reusable descriptions of data and software can be exported and shared. Galaxy automatically generates metadata for each investigation step, allowing the identical repetition of each workflow (Giardine

et al., 2005). Users can annotate their analysis steps with optional tags to explain their necessity and importance. Galaxy's workflow editor tool determines whether the outputs or tools in a workflow are compatible and allows users to easily see if a tool's output can be used as the input of another through a clear user interface. Galaxy enables LS databases and large datasets held within cloud-based systems to be accessed, such as the NCBI Sequence Read Archive (SRA) data held on Amazon Web Service's servers. Nevertheless, with the way Galaxy currently operates, there are still certain reproducibility traps. For example, the user might not know the details and specifications of the program within the Web service, or the Web services might be unavailable.

Rather than performing the analyses on remote Galaxy servers using the Galaxy Web version, scientists can download and run Galaxy instances and perform their analyses on local machines, which can speed up the time for completing a job, especially for large datasets (Afgan et al., 2011). Nonetheless, local Galaxy instances are often difficult to operate as installing certain tools can be complicated; if used at the institutional level alongside High-Performance Computing environments, their maintenance is expensive as it requires the support of the appropriate infrastructure.

Taverna

Taverna is another freely available bioinformatics WMSs, which acts as an interface for linking molecular biology software, databases and particularly Web services (Hull et al., 2006; Oinn et al., 2004; Wolstencroft et al., 2013).

Specifically, Taverna addresses the lack of communication between databases and bioinformatics applications, which can provide novel insights when integrated. Belonging to the myGrid project (Goble et al., 2003), Taverna allows the user, with or without programming skills, to create and run workflows (see Figure 2.4.3).

Taverna workflows can be deposited and downloaded for reuse from the myGrid

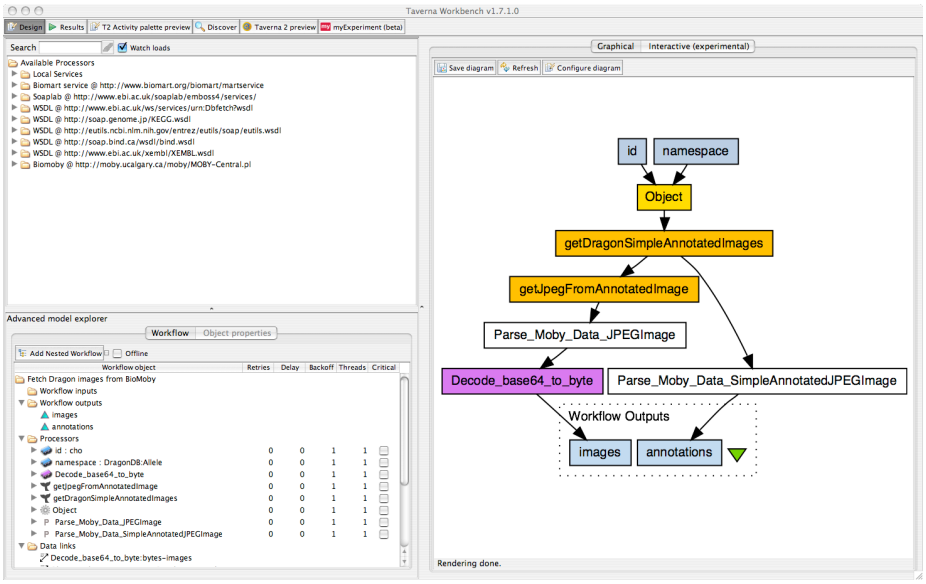

Figure 2.4.3: *Taverna workflow example.* Image taken from: http://www.mygrid.org.uk/files/2008/09/dragon-workflow.png?.

workflow repository (Goble et al., 2003). Taverna bypasses the tedious activity of copying and pasting data between online tools and databases and the need to compose scripts for this action to be performed automatically, where scripts can sometimes be vulnerable and break when websites become obsolete or change format.

A valuable feature of Taverna is that the workflow metadata (the status, start and end time of a task, the kind of processor utilised and descriptions of the analysis) can be saved independently from the workflow and be retrieved using a Life Science Identifier, where the user is obliged to provide metadata to this Life Science Identifier (Oinn et al., 2004).

2.4.2 CyVerse

The iPlant Collaborative launched in 2008 to develop and implement a national US cloud-based cyberinfrastructure for plant biology research (Goff et al., 2011). In 2013 iPlant expanded to cover the whole of the LSs and was renamed CyVerse. CyVerse offers high-performance computing, data storage and cyberinfrastructure, including virtual cloud platforms that integrate software,

databases, hardware and analysis provenance tracking to power virtual collaborations around shared analysis tools and data. Moreover, CyVerse has a mission to train scientists to use the platform for enabling Big Data science (Goff et al., 2011). These characteristics, i.e. the sharing of workflows and experiments being recorded and replayed, support the reproducibility of bioinformatics analyses.

2.4.3 Virtualisation Software

Below is a description of some of the most popular virtualisation technologies that can help with research reproducibility.

Virtual Machines

Virtual Machines (VMs) encapsulate an Operating System (OS) and software with the software dependencies all configured. A VM is built and run by virtualisation software, such as VirtualBox and VMware. VMs of different OSs can be run parallel to the computer's OS. The whole VM can be exported as a binary file (file comprised of a series of sequential bytes, each being eight bits in length). The benefit of this feature is the ability to save time in reproducing the operating environment and most of the computational conditions of any CEs. From the scientist's perspective, they only need to record the installation and configuration steps of an OS, and other scientists only need to install the virtualisation software (not all software components) so analyses can be repeated. This is provided the virtualisation software is compatible with the computer's systems (Smith and Nair, 2005).

Nonetheless, VMs are not always the best solution for helping with computational reproducibility. VM files are typically many gigabytes in size, which means that sharing them becomes difficult. However, cloud computing services (e.g., Amazon) can be more efficient (Piccolo and Frampton, 2016) in sharing snapshots of the VMs (Dudley and Butte, 2010). Depending on the

provider, these cloud services are liable to fee charges. Another downside of VMs is that the software components and the specifics of each analysis may not be documented and distributed adequately by the creators of the VM, making it more challenging to extend the analysis (Piccolo and Frampton, 2016). A solution to this problem would be to have all of the constituents of the VM in a public repository that would then be imported for running the analysis (Piccolo and Frampton, 2016), in addition to using prepacked VMs, such as CloudBioLinux (Krampis et al., 2012).

The building of VMs can be automated using open source VM management tools like Vagrant (explained further in section 2.4.4 below) or Vortex. The advantage of using these tools is that users can compose text-based configuration files that instruct the construction of the VMs and assign computational resources to them. Given that these files are small (a few kilobytes), they can be easily version-tracked and shared, which is advantageous.

2.4.4 Containerisation Software

Docker

Docker was released in 2013, and it is a popular open source technology that enables the reproducibility of experiments (Boettiger, 2015) by encapsulating a whole computing environment within a shareable container.

A Docker container is a running instance of a Docker image, and many Docker containers can be created from the same image, each with its own unique data. Docker images are built using commands from a Dockerfile (Docker Docs, nd) which are lightweight, which means Dockerfiles can be shared easily (via DockerHub) and allow for the tracking of versions in the source control repositories - or for the tracing of only the components in the image that have changed between the different versions (Boettiger, 2015; Vase, 2015; Piccolo

and Frampton, 2016); this is more advantageous to VMs, as for any change made in VMs one has to update the complete machine. Given their above features, Docker containers can support reproducible research (Chamberlain and Schommer, 2014).

Vagrant

Vagrant is a tool for building and managing VM environments (enabling the host's OS to encapsulate and run a guest OS with software dependencies preinstalled and configured) in a single workflow. With an easy-to-use workflow and a focus on automation, Vagrant lowers the development of a reproducible environment setup time and makes the "it works on my machine" excuse a relic of the past (Vagrant, nda).

Vagrant isolates projects' dependencies and configuration in a single environment and can be shared as a Vagrantfile. The Vagrantfiles can be easily shared as they are lightweight. When performing a "vagrant up" command on the Vagrantfile, a guest Vagrant machine can be created and configured according to the Vagrantfile (Vagrant, nda).

Vagrant integrates with Ansible, a tool for "server configuration and automation". More explanation on Ansible is given below. Vagrant, Ansible and Salt technologies are mentioned here as they were used in the project described in Chapter 4 of this thesis (see Section 4.3.3).

Ansible Ansible is a simple and robust Information Technology IT automation tool, which enables the acceleration of managing complex IT environments (Ansible, nd).

Ansible models IT infrastructure not just by managing one system at a time, but by giving an outline of how all systems in a virtual environment inter-relate (Ansible, nd). Ansible is written in YAML. Ansible represents the machines it manages using a simple initiation file. Ansible connects the nodes of the IT

infrastructure used and produces a small program called an "Ansible module", which is then pushed to compute nodes. These modules can serve as a human-readable data serialisation language and are commonly used for configuration files. Ansible modules can also serve as resource modules when executed remotely on a server, supporting the system's preferred configuration.

Salt or SaltStack is a Python-based open-source software for IT infrastructure automation, configuration management and remote task execution (SaltStack, nd).

The advantage of using SaltStack is that it can be configured and run quickly (within minutes). It allows communication between the systems in the network within seconds and can be scaled to tens of thousands of servers (**?**). SaltStack can be used for data-driven computational activities, as it can execute tasks remotely for any infrastructure and provide configuration management for any application type (SaltStack, nd).

According to the site Technopedia, "Virtual provisioning is a virtual storage network-based technology in which storage space is allocated on demand to devices. This process allows virtualised environments to control the allocation and management of physical disk storage connected with VM" (Technopedia, nd). The Vagrant Salt provisioner permits provisioning using Salt states (Vagrant, ndb). Salt states are YAML files which define what packages should be installed, which services should be running and what files should be included in a Vagrant machine (Vagrant, ndb).

There are several other open-source containerisation software applications, such as Linux Containers and lmctfy. Describing the aforementioned technologies further is beyond the scope of this thesis because they were not explored as potential tools in any of the projects described in this thesis. In contrast, Docker and Vagrant have been investigated in more detail, and Vagrant was used as mentioned previously as part of the project described in

Chapter 4 (see Section 4.3.3).

2.4.5 Python Compatible Tools and Widgets

IPython Notebook Interactive Widgets and Jupyter Notebook

IPython was created by Fernando Pérez (Pérez and Granger, 2007) and provides an improved interactive environment above the basic Python interpreter. Its features include data visualisation, which, when implemented appropriately, can aid CE reproducibility. Due to the interactivity that IPython offers, scientists can visualise the data, assess ideas and interpret their results more easily.

Some IPython features are graphical interface tool kits and packages for 3D plotting and visualisation. It runs under Unix, Apple OS X and Microsoft Windows. Apart from IPython, there are other tools, such as myBinder and Shiny by R studio, that offer interactive visualisation.

Jupyter Notebook is "web-based interactive computing platform" (Jupyter Project, nd). With Jupyter Notebook, one can create and share documents with live code (executable code), "equations, visualisations and explanatory text" (Jupyter Project, nd). Jupyter Notebooks support over 100 programming languages and offer several uses, including statistical modelling, numerical simulation, machine learning" and using interactive widget code can produce rich output (videos, images and JavaScript), which can be manipulated and visualised in real-time (**?**). Jupyter Notebooks can be easily shared via email, GitHub, Dropbox and the Jupyter Notebook Viewer.

JupyterLab is the (latest) Next-Generation web-based Notebook interface for Project Jupyter. In addition to the Jupyter Notebook functionalities explained above, JupyterLab allows configuring workflows in data science, machine learning, scientific computing and computational journalism (**?**).

2.4.6 myExperiment

The initiative serves as a public repository of workflows. myExperiment's users can discover, share, re-purpose, curate and comment about workflows and other artefacts, exchange knowledge and collaborate within a virtual research environment (Goble and De Roure, 2007; Goble et al., 2010). The workflows are scientific objects which can be exchanged and reused irrespective of the workflow system (e.g. Taverna and Galaxy (Goble and De Roure, 2007). myExperiment gives credit to the workflow creators by allowing their citation. With all the above activities that can be fostered within the myExperiment environment with a Web 2.0 approach, myExperiment enables reproducible research, or at least repeatable research, incorporating input data, results and logs.

2.4.7 UTOPIA

UTOPIA is a User-friendly Tool for OPerating Informatics Applications composed of interactive and interoperable graphical tools giving access to data, databases and Web services (Pettifer et al., 2004, 2009). This allows users to examine the validity and quality of the data presented in papers, visualise it interactively, thus understanding it better, update data *in situ* and validate the accuracy of tables, all with great real-time performance. UTOPIA links papers and their figures (and underlying data) interactively whilst interacting with remote resources. UTOPIA comprises reusable software components. It provides tools such as the CINEMA sequence aligner and a 3D-structure viewer for large molecules. Given the above, the UTOPIA PDF interactive viewer offers applications that reduce the needless technical hurdles for performing bioinformatics analyses and facilitates the reproducibility of results.

2.4.8 Whole Tale

Whole Tale is a Data Infrastructure Building Block project for designing, publishing and executing Tales i.e. executable research objects. It is funded by the National Science Foundation in the United States. Whole Tale is an open source, Web-based, multi-user research platform that captures data, code, computational workflows, the full research-related software environment and narrative (traditional science story). The system's beta version can be seen at https://dashboard.wholetale.org.

Tales are recorded with metadata in a standardised manner. In the Tale, the data and code for CE are expressly referenced for reproducibility reasons and the citations of the particular versions are utilised for later studies. An external research repository can be used to store or publish the Tale, and the repository can issue a persistent identifier. The Whole Tale platform enables users to create and amend Tales interactively and to re-run a Tale to recreate and verify the outcomes produced by the original maker.

It is important to state that the interaction provided by Whole Tale is not interacting with the presented figures but rather the code behind the figures, which can be amended by interacting with Python Jupyter. It is then possible to run that code again to see how the figure changes. In other words, it is not a user interface interaction where the user clicks buttons, moves sliders, or zooms into the image.

2.5 Discussion

With the advent of faster internet networks and more accessible computational power, more researchers are undertaking CE with larger datasets, either on their systems or within high-performance computing or cloud environments. Still, it comes at a cost in that there is huge variability in data, tools, methods,

computing environments and expertise. Moreover, the current culture in science for career progression is mostly based on the number of publications and projects are undertaken, which can pressure researchers to perform research that is not always focused on being done in a robust and reproducible way.

Regardless, scientists have a responsibility in that the research they produce is reproducible. Otherwise, scientists producing irreproducible research are adding further chaos to the sea of data (and research) produced, given the current low costs of performing research in the LSs. However, research reproducibility is a fundamental element of creating reusable software, better datasets, validating and building upon previous research, and ultimately rigorous and high-quality science that can be trusted by the scientific community and the public.

Even though there are many technological solutions to help improve LSs research reproducibility, each comes with different features and variable uptake by the scientific community. However, all are somewhat detached from the scientific paper because they are not immediately usable from within the publication environment where the consumer of scientific research reads the paper.

The benefit of being able to assess the data, code and other research artefacts within the publication are discussed in the next chapter, as well as the contribution and responsibility of publishing journals in relation to the reproducibility crisis.

The next chapter presents the survey conducted to evaluate the knowledge and attitudes of LSs scientists toward research reproducibility, their opinions around interactive figures versus the traditional static figures in research articles, and their perceived benefit in enabling research reproducibility within journal articles.

3

Evaluating Life Scientists' Knowledge and Attitudes Toward Reproducibility

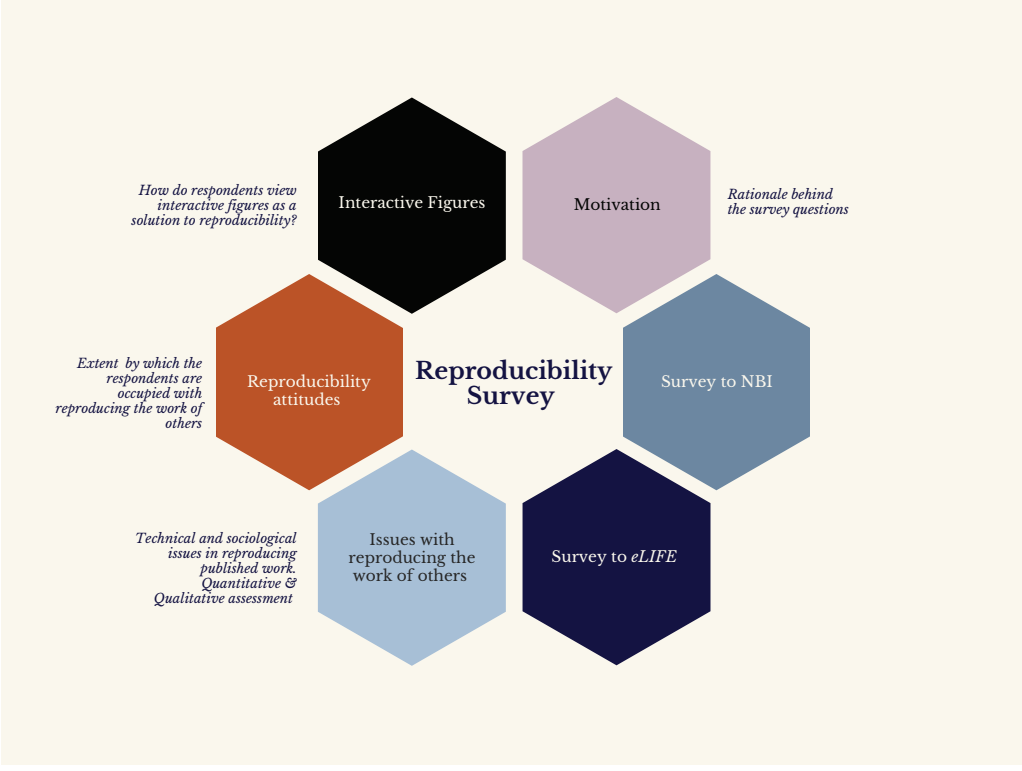

Figure 3.0.1: *Concept Map for Chapter 3: Assessing the knowledge and attitude of life scientists toward reproducibility*

The work contained in this chapter was used towards an open access

publication, reported in *Frontiers in Research Metrics and Analytics* doi:10.3389/frma.2021.678554 of which I am the first author. Moreover, the work in this chapter was carried out as part of a collaboration between my PhD iCASE industrial partner, *eLIFE* Sciences Publications, Ltd, and myself at the Earlham Institute. All findings reported in this chapter are directly derived from my work. A copy of this paper is provided in Appendix A.1, Chapter 3.

3.1 Introduction

In this chapter, I will explain the motivation behind conducting a survey to canvass life scientists' knowledge and attitudes around reproducibility and assess their opinions on having access to interactive figures as a potential solution for the reproducibility of experiments within articles. The survey was framed by the development of interactive figures as a model to serve a richer and more accessible underlying means to the results of a given study or analysis.

3.1.1 Current Views on Reproducibility

In this section, I describe other studies and surveys that canvassed the opinions of scientists about reproducibility and the issues they experience in attempting to reproduce published research.

A number of recent surveys have assessed the attitude and opinions of researchers about reproducibility in disciplines other than the LSs (Baker, 2016; Feger et al., 2019; Stodden, 2010). However, not many surveys have examined the knowledge and attitude of researchers around reproducibility in LSs. In particular, the LSs community has undertaken only rudimentary investigations (especially in the form of surveys that qualify and quantify the

reproducibility issue) into the incidence of problems encountered with reproducibility, assess researchers' perceptions of the significance of reproducibility, the frequency of difficulties experienced when attempting to reproduce published research and the opinions and preferences concerning possible remedies to research irreproducibility in the LSs.

The current literature (Pulverer, 2015) highlights that issues accessing the data presented in research articles are one of the major reasons leading to the irreproducibility of published studies. For example, Federer et al. (2015) investigated the differences in data practices between clinical and non-clinical scientists and found that the majority of respondents had no prior experience uploading biomedical data to a repository. The majority mentioned various sociological reasons for not doing so, including concerns and motivations about data sharing and the workload required to prepare the data for submission.

Barone et al. (2017) conducted a survey in the US of 3,987 principal investigators from the National Science Foundation's Directorate of Biological Sciences who reported their biggest unmet training requirements by their respective institutions. Specifically, the deficiencies were in the areas of integration of various data sources (89%), data management and metadata (78%), and scaling analysis to the cloud/high-performance computing (71%).

A number of survey studies have explored the attitudes and knowledge of researchers in other disciplines about reproducibility to some extent (Stodden, 2010; Baker, 2016; Feger et al., 2019). Still, few studies have explored the attitudes and knowledge of researchers in the LSs about reproducibility (Baker, 2016). The LSs community, in particular, has undertaken only rudimentary investigations into the incidence of problems encountered with reproducibility, the perception of its significance, and preferences concerning possible remedies in the field.

This chapter presents the results of a survey designed to evaluate researchers' knowledge of the principles of reproducibility and to aid future efforts by

enabling researchers to reproduce published research results. Producing and consuming life science research may be better served by creating technologies, such as interactive figures inside journal articles, which may be used better to fulfil the requirements of both producers and consumers. The survey study described in this chapter does not compare open-access tools for creating reproducible research outputs; instead, the survey is restricted to comparing how open-access tools are viewed for the consumption of research material via interactive methods.

We designed the poll in order to gain a better understanding of how the respondents felt about the following topics:

- Issues with data access, code and technique parameters, and how solutions such as interactive figures might improve reproducibility from inside an article are all technical variables influencing computational reproducibility. **These factors are grouped under the term *technical or technological factors*.** An overview description of interactive figures is depicted in the following section.
- Attitudes toward reproducibility, societal issues that impede reproducibility, and interest in how research results may be consumed through interactive figures and their feature preferences are all explored in this section. **These factors are grouped under the term *cultural or sociological factors*.**

When discussing each factor, the two defining terms - technical and technological, and cultural and sociological - will be used interchangeably in this thesis.

3.1.2 Interactive Figures

The survey study in addition to canvassing the knowledge and attitudes of LS researchers around reproducibility, aimed to assess their opinions around interactive figures. In particular, whether they have previously encountered interactive figures in literature, what features they deem favourable and whether they believe interactive figures can improve the reproducibility of experiments within research articles.

As discussed in Chapter 2, a growing number of tools exist and are being developed to assist researchers in performing reproducible research by making code, data, and analyses accessible to the community for reuse. Static plots and figures play an important role in interpreting the scientific outcomes described in research papers.

Static figures are often presented in line with the text of a publication, but they are images with no interactive elements that can represent underlying information dynamically. Interactive figures provide an alternative to static figures in research articles. Interactive figures are technically more complex, as they are developed using image generation systems to form animated figures to visualise data, code, parameters and other details that can be queried, selected and accessed by the user by various means. This might include ”mousing over” data points to show actual values, resizing or zooming the display, or clicking on individual plot elements to display underlying dataset information.

Consequently, interactive figures within research articles can be designed to incorporate information from various sources, including data, code, and graphics, so that when a user interacts with the figure, for example, by selecting an area of data points within a graph, they are presented with the information underlying those data points. Similar to this, a user may alter the fundamental parameters of the analysis, such as changing a filter threshold, which would result in changes to the visualisation of a figure (Barnes and Fluke, 2008;

Barnes et al., 2013; Grossman et al., 2016; Newe, 2016; Perkel, 2018; Weissgerber et al., 2016). A more detailed description and discussion of interactive figures are presented in Chapter 4.

3.2 Survey Methodology

The methodology employed to conduct and analyse the survey is described in my first author paper (Samota and Davey, 2021) and is included here with permission.

Sample Size and Population

Although the data were analysed anonymously, we obtained ethical approval. This survey research was authorised by the University of East Anglia Computing Sciences Research Ethics Committee (CMPREC/1819/R/13).

Our participants' sample was chosen to represent all segments of the LSs community at various degrees of seniority, discipline, and expertise with the problems we sought to examine. The first poll was performed in November 2016 and sent to 750 researchers at the Norwich BioScience Institutes (NBI) who were at post-doctoral or above job level. We decided to poll scientists with a post-doctoral degree or above since they are more likely to have had the chance to publish in a scientific journal. The NBI is a collaboration between four leading UK research institutions: the Earlham Institute (previously The Genome Analysis Centre), the John Innes Centre, The Sainsbury Laboratory, and the Institute of Food Research (now Quadram Institute Bioscience).

Participants were invited through email and provided with a link to the survey. The second survey, identical to the first but included additional questions, was sent in February 2017 to a random sample of 1,651 academics who had published articles in the *eLIFE* magazine. *eLIFE* employees sent invitations to their authors to join through email. We received a 15% response rate (n=112)

from NBI researchers and an 8% response rate (n=139) from the *eLIFE* survey.

The *eLIFE* sample was randomly selected by an *eLIFE* member of staff that had access to the pool of email addresses of *eLIFE* all authors who had published articles (not only corresponding authors) from 2012 (the year *eLIFE* was founded) to 2017 (the date of the survey distribution). Every author participating in the survey has published at least once between 2012 and 2017. The *eLIFE* sample was randomly selected from the journal's email pool, whereby the authors were contacted at random based on their last names alphabetically. The survey was sent via an email invitation in two waves, with the first wave targeting 852 scientists and the second involving 841 scientists. From the first email, 11 addresses bounced. We cannot say exactly how many *eLIFE* authors the sample was selected from as we did not collect that information at the time. However, a month before the survey was distributed (i.e. January 2017), an author newsletter was sent out to the list of accepted authors, and the list comprised 12,040 authors. The decision not to distribute our survey to the complete cohort of 12,040 authors was made by the *eLIFE* staff member. Under the 2017 General Data Policy Regulations (GDPR), implemented within the EU in May 2018, *eLIFE* would only have been able to contact corresponding authors. However, at the time of the survey (2017), it was still possible for all authors to be contacted.

The questions asked in the surveys were as shown in Table 3.1. The questions were designed to elicit qualitative and quantitative responses on the technological (technical) and sociological (cultural) elements of repeatability. The questionnaire evaluated the frequency with which respondents experienced problems obtaining data, the causes for these issues, and the current method by which respondents acquire the data underlying published papers. They assessed participants' comprehension of what comprises research reproducibility, interactive figures, and computationally reproducible data. Finally, we assessed the perceived value of interactive figures as a potential solution to computational reproducibility and the desired characteristics of

interactive figures.

The survey questions were designed whereby some required respondents to make choices from a set of pre-defined options, some allowed a free response and others required participants to rank their experience or knowledge on a scale. One example of an options question is question 3: "What difficulties have you encountered in accessing the data described in published papers? Select all that apply to you." The answer options for this question were: (a) privacy reason (patients' medical data); (b) commercial sensitivity around the data (e.g. pharmaceutical companies' data that could lead to the production of a drug; (c) data not available at publication; (d) authors cannot be reached or are unresponsive to data provision requests; (e) data is too large to be transferable and (f) not applicable (N/A). The answer options to the survey questions are shown in Appendix A2 for Chapter 3.

Table 3.1: Questions used to survey the knowledge of respondents about research reproducibility and interactive figures within publications.

1	How often do you encounter difficulties working with bioinformatic analysis tools (that are not your own)? (Problems such as: installing, configuring, running, and working with command-line software)?
2	How difficult is it to source (or access) the data presented in published papers?
3	What difficulties have you encountered in accessing the data described in published papers?
4	How are you currently sourcing the data (if applicable)? Select all that apply to you.
5*	What is your current understanding of the reproducibility of experiments? Please select any that apply. Should you wish to add any additional information, please add it to the "Other" box.
6*	Have you ever tried reproducing any published results? Please select the answer that applies best to you.

7*	In your opinion, what could be done to make published research more reproducible? Other, please specify (free text answer).
8	When thinking about interactive figures, what comes to your mind? (Please describe what you understand as an interactive figure, its features, and where you have seen such a feature before, if applicable).
9	An interactive figure is a figure within a paper that is dynamic and becomes "live" when the user interacts with it and where the data displayed changes according to various parameter options. Which of the following features of an interactive figure tool would be good to have? Please rank them in the order of preference, where 1 is the most preferred feature, and 11 is the least preferred feature.
10	What other features could an interactive figure have that were not mentioned in the previous question?
11	Do you perceive a benefit in being able to publish interactive figures?
12	Does the provision or option of an interactive figure in the paper affect your decision in choosing the publishing journal or publisher?
13	Have you heard of the term computationally reproducible data, and do you understand what the term means? If answered yes or unsure, please explain what you understand from the term.
14	Would you benefit from being able to automatically reproduce computational experiments or other analyses (including statistical tests) described within a paper?
15	How often do you work with bioinformatic analysis tools (e.g. assemblers, aligners, structure modelling)?
16	Have you received any of the following training? Training, whether formal or informal (training through a colleague etc.).
17	Which of the following type(s) of data do you work with?

Table 3.1: * Questions indicated with an asterisk were only available to the *eLIFE* survey. Answer options to the questions are shown in Appendix A2 for Chapter 3. This table has also been published in (Samota and Davey, 2021) and appears here with permission.

3.2.1 Statistical Analysis

Typically, results are given as a percentage of those who responded, and stratified by the respondent's field of work, training received, and survey version. Chi-square independence tests were performed to see if there were connections between answers to particular questions or whether responses changed across samples. R (version 3.5.2; R Core Team, 2018) and Microsoft Excel were used for the analysis. The R script is available via Figshare from this link: https://doi.org/10.6084/m9.figshare.11291453.v1 and the survey raw data is available via Figshare from this link: https://doi.org/10.6084/m9.figshare.7855592.v1.

We examined whether there was a significant discrepancy in the ability and willingness to reproduce published studies between the cohort of *eLIFE* respondents who understood the term "computationally reproducible data" and those who did not, as well as the effect of training (bioinformatics, computer science, statistics, or no training). Given the wide range of responses in the "unsure" group regarding their understanding of the term "computationally reproducible data," we excluded the data from those who responded "unsure" from our analysis (see Section *Understanding of reproducibility, training received and achieving reproducibility*). The respondents who selected "yes, I attempted reproducing results, but unsuccessfully," "have not attempted reproducing results," and "it is not necessary to replicate results" were all classified as "unsuccessfully." Free text responses are available in Appendix A3 for Chapter 3.

3.2.2 Validation of the Questionnaire Design

We conducted a two-stage survey, beginning with the NBI participants and concluding with the *eLIFE* cohort survey that included further questions. To evaluate the appropriateness and flow of the first survey, we tested it on a small

cohort of researchers. The qualitative findings from the surveys were presented in line with the Standards for Reporting Qualitative Research (SRQR) (O' Brien et al., 2014).

The survey questions were not developed in accordance with any particular cultural theory but rather with our knowledge of the area of reproducibility, namely human variables and researcher attitudes toward reproducibility, as well as the process of doing science. We presume that these factors affect the reproducibility and robustness of the research and, therefore, the published work. As a result, we use the phrase "culture of reproducibility" to refer to the attitudes of life scientists toward science and reproducibility as they pertain to research publications rather than human demography. The purpose of assessing the reproducibility culture (sociological factors) was to determine how the attitudes of researchers or how they communicate their work in research publications might influence reproducibility.

It is critical to emphasise that no one survey question assessed simply the technological issues influencing reproducibility or just the sociological factors. For example, the reader's ability to obtain data is influenced by cultural and technological factors, e.g., data accessible through permanent identifiers and Application Programming Interface (APIs) versus "data available on request." Whereas when an author is not publishing the underlying data associated with their research is solely a cultural issue, as it may be interpreted that they are not conducting and presenting their research in a reproducible manner or that they lack support or knowledge regarding reproducible and open data publishing practices.

Additionally, we analysed the opinions held around interactive figures as part of our study. They are, in and of themselves, a technological element that we believe may contribute to reproducibility. However, readers' interest in finding interactive figures valuable for improving computational reproducibility, including which characteristics they believe are advantageous, may vary based

on the respondent's social background or demography (for example, the training received, the data they work with, or the discipline they work in).

We recognise that many human variables influence how reproducibility is accomplished, most notably the attitude of life scientists toward reproducibility. The reproducibility of their work is affected by how robust, open-source, and open-access their study is in terms of how it is conducted and shared. In this respect, we may analyse, assess, and quantify the problem by measuring and qualifying how difficult it is to obtain data and code provided in publications and comprehend the techniques described in a study. Our study complements previous surveys that also emphasise the problem of reproducibility.

We obtained consistent answers, indicating that all or most respondents understood the questions similarly, allowing for cross-comparison. The NBI research findings were similar to the *eLIFE* study results, despite the surveys being conducted at separate periods and with distinct survey groups. To develop our construct validity, we mainly relied on published surveys, findings drawn from the existing literature, and the outcomes of conversations with different researchers investigating reproducibility in our local institutions. The questions we posed met our criteria for gaining a deeper understanding of the qualitative character of respondents' responses and being able to conduct empirical analysis (Chi-squared) to demonstrate connections.

We used the same procedure to evaluate content validity. We attempted to include questions that covered the breadth of the area we were evaluating while casting a broad net over possible respondents, who represented a diverse range of skills, disciplines and other demographics. Lastly, our hypothesis of how researchers perceive reproducibility was included in our questionnaires' translation validity, and two practical surveys were created based on our theoretical evaluations from prior literature.

3.2.3 Rationale behind the Survey Questions

The survey questions were designed to assess the survey participants' understanding of the meaning of the concept of reproducibility of experiments and scientific results. The objective was to determine whether they understand the importance, value and worth of the reproducibility of experiments and how devising a process that enables reproducibility would benefit their research and science in general.

A primary goal was to collect information concerning the problems that scientists are currently facing in reproducing experiments. The questions were phrased to discover the problems that users are currently facing. For example, are scientists experiencing difficulty collecting the data presented in the published research? The survey was designed to obtain quantitative measurements of the percentages for each category of hurdle experienced in reproducing experiments. Gathering such information would allow focusing on building features for the interactive figures that would primarily focus on and prioritise the most popular issues surrounding the reproducibility of experiments.

The survey was also conducted to gather from the user which features of an interactive figure would be desirable and useful for them and resonate more with them, and receive further suggestions on other features to be developed. In addition to the above, a qualitative assessment of the perceived value of each of these features was also included to establish how, in the respondents' minds, these added features would solve their current problems with the irreproducibility of experiments within journal articles.

3.3 Results

These results have been published in Samota and Davey (2021) and appear here with permission.

Features of the Samples

Figure 3.3.1 depicts the distribution of respondents' fields of work, which has been stratified based on the survey sample. When comparing the NBI and *eLIFE* populations, the most often involved topic areas of the whole sample were genomics (22%), biochemistry (17% of the whole sample), and computational biology (15%). When asked how often they utilise bioinformatics tools, 25% said they never do so, 39% said they rarely do so, and 36% said they do so often; (number of responses to question 15, n=136). In addition, (43%) obtained statistical training, (31%) bioinformatic training and (20%) computer science training; (number of responses to question 16, n=136).

Data and Bioinformatics Tools

In both groups, 90% of individuals who answered question 2 said they had attempted to gain access to data underlying a previously published scientific paper (see Fig. 3.3.2; the number of respondents to question 2, n=221).

Of those who attempted accessing data described in published papers, only a small percentage found this "easy" (14%) or "very easy" (2%). Whereas 41% described it as "difficult" and 5% described it as "very difficult." The most common reasons for difficulty were sociological, as outlined in Fig. 3.3.2, such as the fact that the data was unavailable with the publication (as discovered by 75% of those who attempted to access data) or that authors could not be contacted or did not respond to requests for data (52%). Only a small number of people discovered that data was inaccessible due to technical factors such as

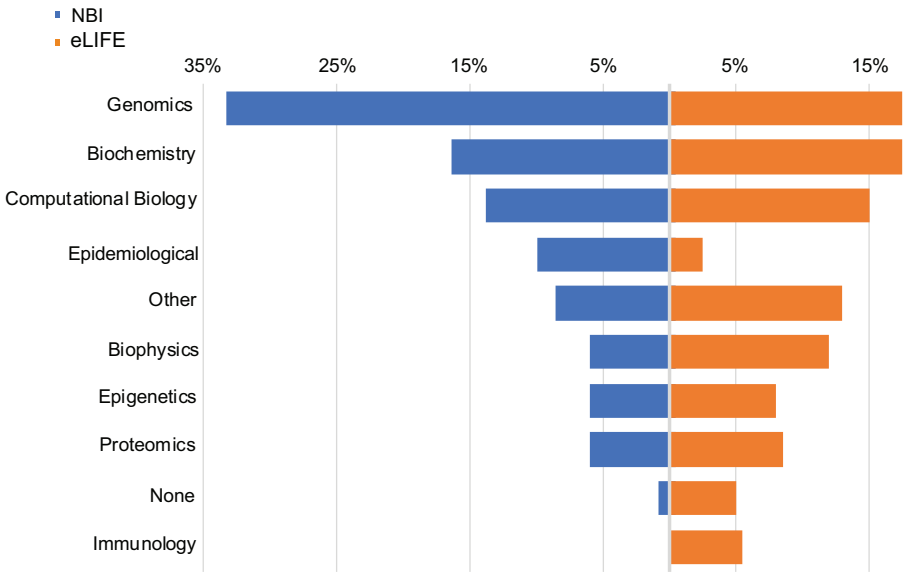

Figure 3.3.1: *Data types NBI and eLIFE respondents work with; (number of respondents to question 17, n=136). Responses were not mutually exclusive. The data type options were the same as those accessible in the eLIFE article categorisation system, providing a familiar experience.* This figure has also been published in Samota and Davey (2021) and appears here with permission.

data size (21%), commercial sensitivity (13%), or confidentiality (12%). The number of respondents to question 3, was 218 (n=218).

For data sources, 57% of the overall sample reported using open public databases, 48% stated that data was accessible via a link in the article, and 47% claimed that they had to contact the paper authors to be provided with the data (number of respondents to question 4, n=219.Responses were not mutually exclusive).

Only a small percentage did not experience difficulties operating, installing, or configuring bioinformatics software. Those who answered either "never" (2%) or "rarely" (8%) encountered issues. Problems with software were reported as occurring "often" (23%) or "very often" (12%), indicating that almost half of those who answered the survey reported regularly encountering technical obstacles to achieving reproducibility. The respondents who answered question 1, N=251.

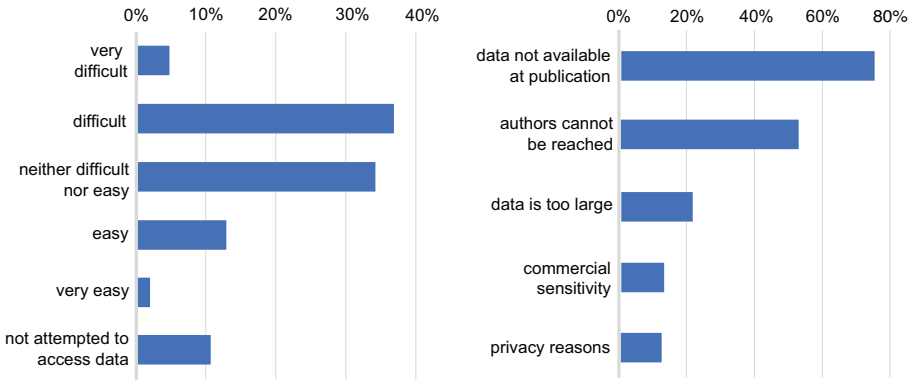

Figure 3.3.2: ***Left panel: Difficulties faced in obtaining data underpinning previously published study***. *When asked if they had tried to gain access to data underlying prior publications, respondents said they had and described the difficulty they usually experienced. Responses to question 2, n=221.* ***Right panel: The reasons for the inability to access data are explained.*** *The reasons cited by respondents for not being able to obtain data (restricted to those who have attempted to access data). Responses to question 3, n=218.*

Understanding of Reproducibility, Training Received and Achieving Reproducibility

The vast majority of respondents (to question 5) said that they understood the phrase "the reproducibility of experiments" and that they chose the explanation for the term as accepted in this thesis and described in Chapter 2 Section 2.1.1 (*different teams re-running the same analysis with the same data and getting the same results*), which conforms to the most widely accepted definitions of reproducibility in scientific research (Claerbout and Karrenbach, 1992; Peng et al., 2006; Peng, 2011; Stodden et al., 2013a). Question 5 was only available to the *eLIFE* cohort with 54 respondents to this question, N=54.

Respondents were able to choose more than one answer from the available options for question 5. The answer options to question 5, were the following:

- (a) If the experiment was performed by another laboratory, the same or similar results are produced
- (b) Using similar materials, reagents, and methods, reaching the same

conclusions

- (c) Running and analysing similar data with the same workflow, and getting similar results
- (d) The original authors or others running the same data with precisely the same workflow and getting the same results
- Other (please specify) as an open comment box.

It is important to note that respondents were able to select more than one answer for this question because we recognised the limitations of the fact that there is no standard and accepted definition for reproducibility, as well as the fact that scientists from different backgrounds may have different levels of familiarity with the term. The first three answer options given to question 5 (see Appendix A2 for Chapter 3), were reasonable definitions of research reproducibility.

We infer from the findings that some of the respondents selected definitions that were both correct and incorrect at the same time. However, the definition of reproducibility as defined in this thesis was included in the vast majority of responses (77%). However, when we looked at the individual answers (n=54), we found that 11.1% (n=6) of respondents selected just option A, indicating that they understood that this corresponded to the concept of reproducibility as stated in Chapter 2 (see Section *Definitions of Reproducibility*). Only option D was selected by 5.5% (n=3), which is a wrong definition for research reproducibility.

The vast majority of people (57%; n=23) chose one of the options A, B, or C rather than D, indicating that they understand that reproducibility is not the same as replicability. However, they were still unsure of the exact definitions, which corresponds to the general lack of consensus on the subject as noted by other literature (Drummond, 2009; Liberman, 2015; Plesser, 2018). The fact that just over a third (37%, n=20) chose one or all of A, B and C, and chose D, suggests that they did not understand or care about the difference between

reproducibility and replicability and that they believed any form of repeating a process could be classified as "the reproducibility of experiments" (see Supplementary Table 4, available on Figshare via this link).

Understanding of the term "computationally reproducible data"

While there is no official definition for "computationally reproducible data," several sources and studies have discussed the concept of data that contributes to computational reproducibility (Baranyi and Greilhuber, 1999; Weinländer et al., 2009; de Ruiter, 2017; Tait, 2017; Perkel, 2017; Pawlik et al., 2019). This thesis defines the term as the result (data outputs) generated when reproducing computational experiments.

The majority of participants (52%) gave a different meaning to the phrase "computationally reproducible data", with just 26% understanding the term and 22% uncertain. The number of respondents to this question (Question 13), was 137 (N=137). We obtained several explanations (from the free text answers) for the term, the vast majority of which were correct (see Appendix A.3, free responses to question 13). We divided the "unsure" answers (n=30) into three categories: those who understood the term (70%, n=21; see Appendix A.3, free responses to question 13), those who did not comprehend the term, and those who did not provide any free text answer. The majority of respondents who selected "unsure" and provided a free text answer (71%, n=15) correctly identified the term "computationally reproducible data." Six people could not grasp the phrase properly in the remaining 29% (n=6).

Willingness and Success in Reproducing Published Research

Some respondents (18%) said that they did not make an effort to reproduce previously published findings. Only a small percentage of the sample (6%) agreed with the statement that "it is not important to reproduce other people's

published results" (see Supplementary Figure 1 in Appendix A.4). Even though the vast majority of respondents (60%) reported having successfully reproduced published findings, almost a quarter of those who responded indicated that their attempts to replicate any results were fruitless (23%). The number of respondents to this question (question 6), was 92, (N = 92). This question was only available in the *eLIFE* cohort of the survey.

Table 3.2, displays how well the respondents were able to reproduce experiments based on their comprehension of the term "computationally reproducible data" and the training they have received (bioinformatics, computer science, statistics). The connection between the capacity to reproduce published experiments and understanding of what is meant by "computationally reproducible data" was investigated using a Chi-square test of independence, whereby no significant relationship was found.

The connection between these factors was statistically significant, with a p-value of $\chi 2(1, N = 75) = 3.90$, $p = .048$. The ability to replicate published experiments was much higher among those who understood the meaning of "computationally reproducible data." Considering that their previous training did not reveal any statistically significant differences. The responses "yes tried reproducing results, but unsuccessfully", "have not tried to reproduce results", and "it is not important to reproduce results" were all grouped under "unsuccessfully" to get an indication of how willingness and success together differed between the training groups. This revealed a statistically significant difference between the two training groups (see Supplementary Table 1 Appendix A.5). The distribution of the training variable differed substantially between individuals who had received computer science training and those who had not (Fisher exact test for independence, $p = .018$). It appeared that individuals who had computer science training were less likely to have attempted to reproduce an experiment but were more likely to have been successful when attempting to reproduce it.

Table 3.2: **Success in reproducing any published results stratified by participants' knowledge of the term "computationally reproducible data" and training received.**
n is different for the two variables as not all participants answered all the questions. [a]Not Successful includes answers: "Yes, I have tried reproducing published results, but I have been unsuccessful in producing any results, or the same results", "No, I have never tried reproducing any published results", and "It is not important to reproduce other peoples' published results'. [b]Statistically significant at the level of $p < 0.005$. [c]Chi-square statistic with Yates correction, applied when expected frequencies were lower than five. This table appears in my first author publication (Samota and Davey, 2021) and is printed here with permission.

	Number (% of the total sample)	**Success in reproducing any published results**		
Variable		**Successful (% within variable)**	**Not Successful[a] (% within variable)**	**P-value**
Knowledge of the term *"computationally reproducible data" (n = 75)*				
Yes	25 (33.3)	18 (72)	7 (28)	0.048[b]
No	50 (66.7)	24 (48)	26 (33)	
Training *(n = 90)*				
Bioinformatics	42 (46.7)	26 (61.9)	16 (38.1)	0.73
Not trained in Bioinformatics	48 (53.3)	28 (58.3)	20 (41.7)	
Computer Science	33 (36.7)	21 (63.6)	12 (36.4)	0.59
Not trained in Computer Science	57 (63.3)	33 (57.9)	24 (42.1)	
Statistics	71 (78.9)	42 (59.2)	29 (40.8)	0.75
Not trained in Statistics	19 (21.1)	12 (63.2)	7 (36.8)	
No training	10 (11.1)	6 (60)	4 (40)	0.73[c]
All other training	80 (88.8)	48 (60)	32 (40)	

There was no evidence of a difference in the ability and willingness to reproduce published results between respondents who used bioinformatics tools frequently and those who used them "rarely" or "never", $\chi 2(3, N = 90) = 0.53$, $p = .91$ (see Supplementary Table 2, Appendix A.6). Those who utilise bioinformatics tools often came from various scientific backgrounds, mostly biophysics, biochemistry, computational biology, and genomics. Those who responded that "reproducibility is not essential" and "haven't attempted reproducing experiments" were mostly scientists from fields that used computational or bioinformatics tools "rarely" or "never" in their research (Supplementary Table 3, Appendix A.7).

Increasing the Reproducibility of Previously Published Research

The vast majority of respondents (91%) agreed that authors detailing all methodology steps, including any formulas and parameters used to analyse the data, may help to make published research more reproducible and thus more valuable. About half (53%) agreed that authors should share the source code of any custom software used to examine the data and that the code should be properly documented. Similarly, 49% said authors should provide a link to the raw data (see Supplementary Figure 2, Appendix A.8). Two respondents stated that improving scientific reproducibility would be simpler if funds were more easily accessible for reproducing the findings of others and if there were more chances to publish the reproduced results (see Appendix A.3 free responses to question 7).

Within the same context, some respondents acknowledged that the current scientific culture does not provide sufficient incentives for publishing reproducible papers (or even negative findings papers). Instead, researchers are rewarded for publishing as many papers as possible in high-impact factor journals (see Appendix A.3 free responses to question 7).

Interactive Figures

The results from question 9 regarding the respondents' preference for interactive figure features are shown in Fig. 3.3.3. Participants were asked to rate their preferences for interactive figure elements included in a research paper. "Simple to manipulate" was the most desired characteristic of interactive figures, followed by "easy to specify parameters" (Fig. 3.3.3). The responses to both the *eLIFE* and NBI questionnaires generally followed a similar pattern.

Additionally, free text answers were gathered, and most respondents indicated

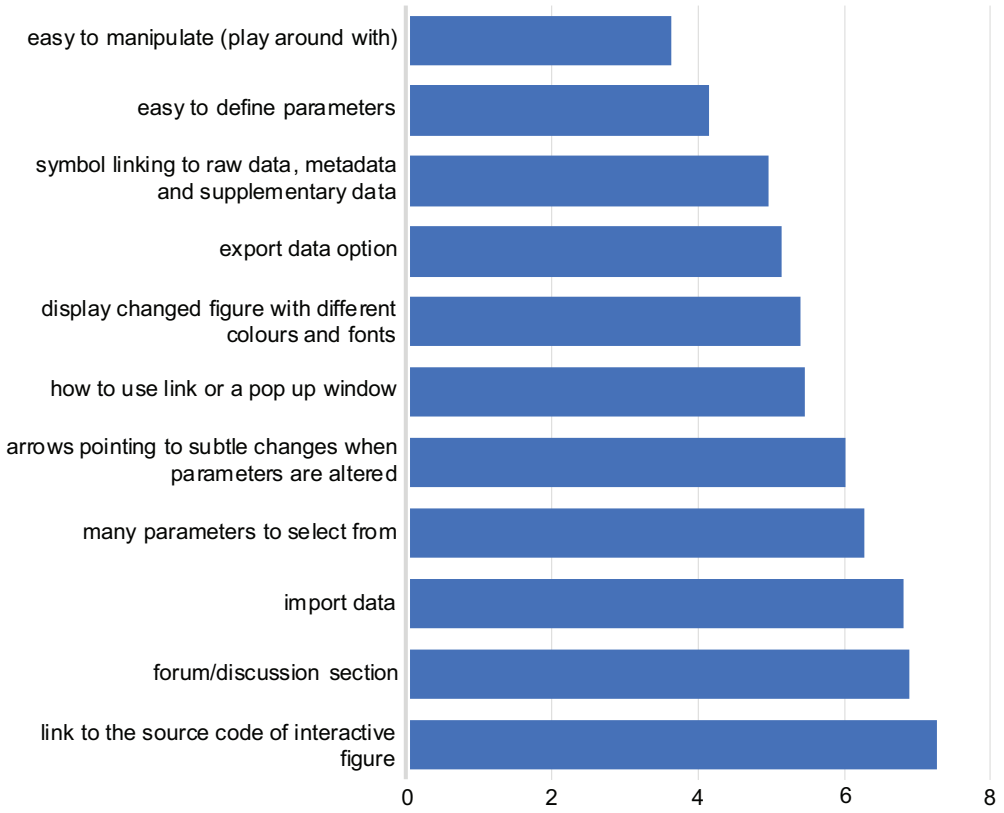

Figure 3.3.3: Preferences for the interactive figure's characteristics. In Question 9, respondents were asked to rate the above characteristics in preference, with 1 being the most desired characteristic and 11 being the least liked characteristic. The average score for each feature was computed in the order in which respondents from the NBI and the *eLIFE* surveys chose their top three choices for each feature. The lower the average score value (on the x-axis), the more desirable the feature is perceived by users (y-axis). The number of responses to this question was 136 (N = 136).

that methods that would aid them in comprehending the data provided in the figure, such as zooming in on data, would be helpful (see Appendix A.3 free responses to question 8).

The overwhelming majority of respondents believed that interactive figures in published papers would help readers and authors (see Fig. 3.3.4). Examples of insights include the following: the interactive figure would enable readers to visualise additional points on the plot from the supplementary section, as well as alter the data presented in the figure; having an interactive figure, such as a movie or displaying protein 3D structures, would benefit readers. The remainder of the answers were classified as software related and included recommendations for tools that might be used to create an interactive figure,

such as R Shiny.

We received 114 free text answers about the respondents' perceptions of interactive figures, and a quarter (25%) said that they had never seen or engaged with one before, with no indication that an interactive figure would aid their work (see Appendix A.3 free responses to question 9).

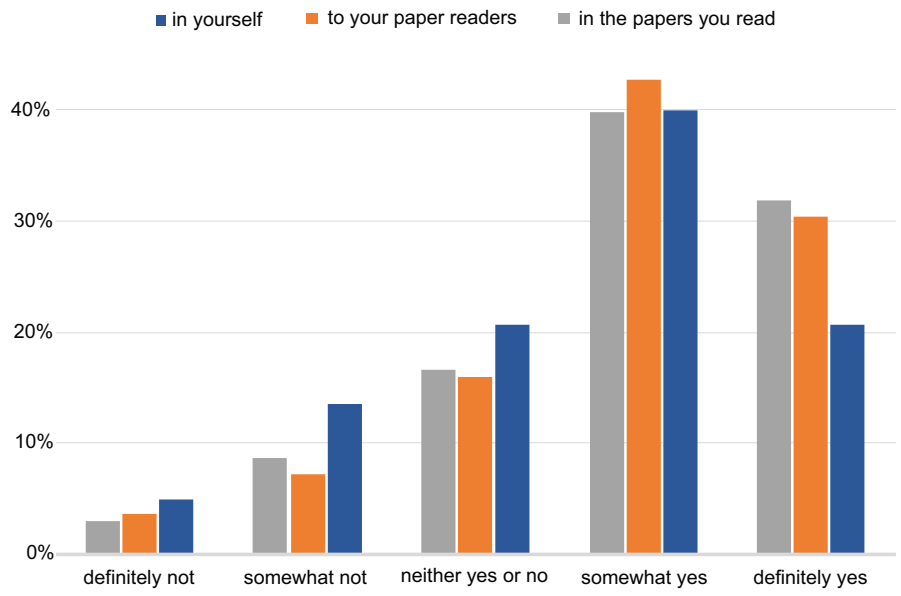

Figure 3.3.4: *Perceptions on the value of being able to publish research papers including interactive figures*. The advantage to the author, the author's readers, and the papers that the author reads. The answers to question 11 come from both the NBI and *eLIFE* surveys.

Additionally, most respondents find value in automatically reproducing computational experiments and modifying and engaging with parameters in computational analytic processes. Equally advantageous was the ability to reproduce statistical analyses computationally (see Fig. 3.3.5).

Despite this apparent advantage, the majority of respondents (61%, n=85; those who answered "just as likely") stated that the option to add an interactive figure would have no impact on their choice of journal for publication. The number of responses to this question (question 12), was 140. The responses that expressed that having interactive figures would impact their choice of a journal "negatively"

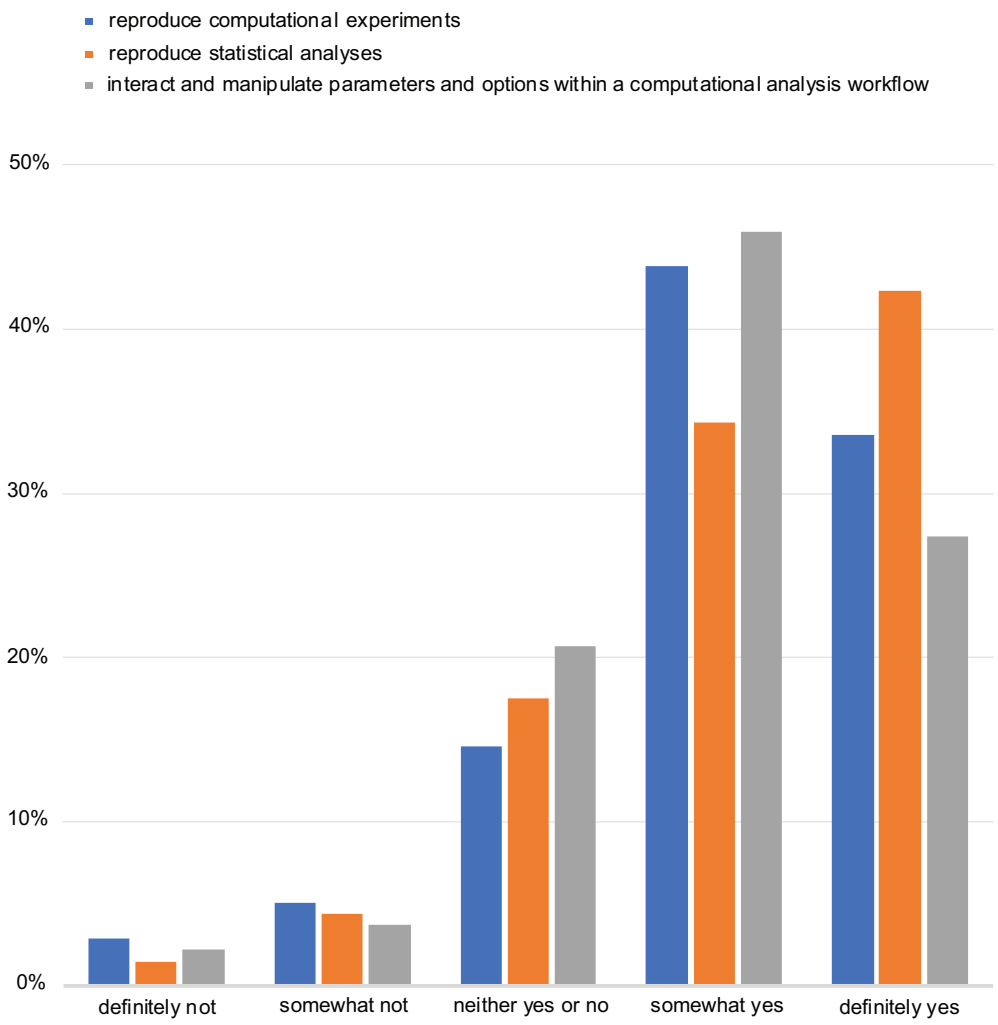

Figure 3.3.5: Perceived usefulness of automatically reproducing computational experiments or other analysis (including statistical tests). Responses from both NBI and *eLIFE*.

were 4% and "less likely" at 11%. Those who perceived the provision of interactive figures would have a positive impact on their choice of the journal to publish with were "likely" 19%, and "positively" 4%.

3.4 Discussion

This research emphasises the challenges that scientists presently face in reproducing experiments, as well as the favourable views that scientists have

about enabling and encouraging the reproducibility of previously published studies, including using interactive figures. Our survey respondents were active researchers in the LSs when the survey was conducted, and our *eLIFE* cohort included researchers who had published at least once in the *eLIFE* journal. Therefore, the opinions collected should be representative of researchers in the LSs who were routinely reading and publishing research when the survey was conducted.

The respondents' views mirror previously reported problems of the publishing processes (Müller et al., 2003; Marx, 2013; Stodden, 2015; Tenopir et al., 2011), although improvement has been achieved in publication standards across all LSs scientific fields. There is a lack of data and code provision; there are no storage standards; and there is no requirement for a detailed description of the methods and code structure (i.e. code scripts and algorithms, full software packages, the programming language used, the versions of any libraries required, the organisation of any modular components, or configuration and deployment options) in the published papers. However, the degree of interest in incentives for reproducing published research is still in its early stages, or it is not a top concern for most academics (Collins and Tabak, 2014; Nosek et al., 2015).

A significant finding of this chapter is that the vast majority of respondents recognised that science becomes implicitly more reproducible if techniques (including data, analysis, and code) are well-described and made accessible to other researchers. Tools that make data, techniques, and code available and automatically replicate computational experiments detailed in the article are seen as advantages by those who responded. The use of interactive figures in publications can be such tools that allow for the automatic reproducibility of computational experiments or other analyses described in a paper. This would include the interaction and manipulation of parameters within the computational analysis workflow, the provision of additional insights, and a detailed view of the data displayed in the figure. Although technologies exist to

assist reproducibility (Crick et al., 2014), and even though writers are aware of their benefits, many scientific papers fail to satisfy the most basic criteria for reproducibility.

The results are consistent with other literature (Pulverer, 2015), which indicate that a lack of access to the publication's associated data is one of the most significant factors contributing to the irreproducibility of published research. Other surveys have corroborated how various factors influence data practices, including fear of being scooped, fear of other researchers using data for their papers, technical difficulties and fear of mistakes being discovered in data or analyses (Stodden, 2010; Tenopir et al., 2011, 2015; Federer et al., 2015).

Through our survey, we found that (57%) of the respondents expressed that data is either not accessible upon publication or that authors cannot be reached/are unresponsive to demands for data supply (44%). However, this is still a cultural artefact associated with referring to methods sections in papers as a description of steps to reproduce analysis rather than a fully reproducible solution that includes easy access to public data repositories, open-source code and comprehensive documentation.

To complicate matters, authors often leave projects and institutions, or they can no longer access the data, meaning that "data accessible on request" no longer serves as a realistic alternative for obtaining data. Requiring paid memberships to view material from a publication, and restricted access to an article may also negatively impact reproducibility. Although there has been a precedence for obtaining specific articles via cross-library loan systems or directly contacting the relevant author(s), this option, as with asking for access to data, is not without its difficulties. In recent years, pre-print servers such as bioRxiv have gained popularity, particularly in the genomics and bioinformatics fields. This development has the potential to reduce publication delays while also providing a "line in the sand" with a Digital Object Identifier (DOI) and meeting the requirements for FAIR data (Abdill and Blekhman, 2019; Figueiredo, 2017;

Hollis, 2016). Some respondents stated that the sensitivity of their data has discouraged them from sharing it, although this was only noted by a tiny percentage of respondents in our survey. Even when there are initiatives that try to apply the FAIR principles to clinical data, such as the OpenTrials database (Chen and Zhang, 2014), they are by no means universally adopted.

In some cases, data in publicly accessible repositories with specific deposition requirements (such as the EMBL-EBI European Nucleotide Archive, https://www.ebi.ac.uk/ena) may not be associated or annotated with standardised metadata that accurately describes it but rather with only the bare minimum for deposition (Attwood et al., 2009). Training scientists to successfully apply data management principles will likely result in more data reuse due to better metadata. The topic of data and metadata annotation standards and their effect on reproducibility will be discussed in Chapter 5.

3.4.1 Limitations

The NBI cohort is composed of researchers solely based in the United Kingdom, with potentially knowledge, attitudes and incentives around reproducibility as well as infrastructure to support reproducibility that varies from the rest of the world. Moreover, it is fair to speculate that researchers working at academic institutions (such as those at NBI) may have divergent incentives and attitudes than government agencies and private sector researchers.

The *eLIFE* cohort were researchers who have published at least once in the *eLIFE* journal. Given how *eLIFE* is an open-access data journal, it is logical to infer that the attitudes of the *eLIFE* researchers were more likely to support and engage in reproducible practices. Nevertheless, the *eLIFE* respondents' demographics are from academia, in public or private scientific institutions all over the world. Moreover, given how NBI researchers come from various countries and research backgrounds and received training from institutions internationally and both cohort samples were randomly selected, we are

confident that the NBI and *eLIFE* respondents' views represent the views of the LSs community.

The participants' responses were gathered by survey self-reporting methods, although this is the most common means by which to conduct surveys and entails many advantages, it may also be subject to certain disadvantages and limitations. Self-reported answers can be exaggerated, various biases may affect the results. For instance, we cannot be certain that respondents have successfully reproduced published results, even if they reported to have done so.

Question 5, asked respondents ("*what is your current understanding of reproducibility of experiments, select any that apply, please add additional information in the free text box"*). Potentially, if respondents were confused about the exact meaning of reproducibility, it could have affected their ability to respond to question 6 ("*have you tried reproducing any published results?"*; respondents were able to select only one answer option) and 7 ("*What could be done to make published research more reproducible, select all that apply, add additional information in the free text box*). However, the responses showed that the questions were sufficiently articulated to give us confidence in the validity of the results, and to divide respondents into 2 groups (successful vs not successful in reproducing published experiments) for our analyses.

Questions 8 ("*When thinking about interactive figures, what comes to your mind? Please describe what you understand as an interactive figure, its features, and where you have seen such a feature before, if applicable"*) and 9 ("*An interactive figure is a figure within a paper that is dynamic and becomes "live" when the user interacts with it and where the data displayed changes according to various parameter options. Which of the following features of an interactive figure tool would be good to have? Please rank them in the order of preference, where 1 is the most preferred feature, and 11 is the least preferred feature"*) were displayed on the same page in the questionnaire. This could

have biased their responses. However, given how question 8 responses included answers which were not available as question 9 answer options; we can deduce (given the lack of evidence to denote otherwise) that displaying questions on the same page did not bias the answers to either question (see Appendix A.3 free responses to question 8).

We did not provide monetary incentives to the survey participants. Should we have provided monetary incentives for our survey and sent email reminders to the same or bigger pool of participants we may have improved our response rate (James and Bolstein, 1990; Shettle and Mooney, 1999; Jobber et al., 2004). Nonetheless, our response rates were typical of or higher than surveys of such nature (Koschke, 2003; Snell and Spencer, 2005; Federer et al., 2015; Schneider et al., 2016; Barone et al., 2017), our survey paper that published the results presented in this Chapter (Samota and Davey, 2021) has already gained traction, received positive feedback from the scientific community and been cited by other publications.

3.4.2 Future Work

Empirical Testing of Knowledge and Attitudes Toward Reproducibility of Life Scientists

A means to have empirically assessed the attitudes of respondents on questions such as question 6 ("*Have you ever tried reproducing any published results?*") would have been insightful, in order to actually have had tests (short preferably) for the respondents to prove that they know how to reproduce or not reproduce published experiments. In our case, we would have focused on computational experiments, as this would have been more efficiently measured. This choice is influenced by the fact that anything needing to conduct bench-level science would take too long or be too expensive to conduct. Hence, the empirical assessment

would be limited to computational experiments.

Our survey questions not only try to assess whether the respondents have the skills to reproduce computational experiments, but also whether the published computational experiments are reproducible. In other words, the success of a computational experiment being reproduced does not rely solely on the person trying to reproduce it and their skills and expertise in a specific scientific domain, but also on whether the scientists (in this case, the paper authors) have published the paper in a reproducible manner. As discussed in the Chapter 2 section the reproducibility elements required are a detailed description of the methodology and provision of the raw and processed data.

What kind of computational test could we give the participants to test their ability to reproduce CEs? This is not a straightforward question. It requires factoring in various elements that can affect the results, which explains why there is no such study of this kind to my knowledge. There are various difficulties in manifesting such an experiment. For example, there is the question of where to find people from the same domain, who are willing to participate in the study, and who is also variable in terms of demographics (gender, age, educational background), so that we are able to adjust for such variable factors, and recruit two groups without introducing any demographic biases. This would be quite difficult and would need quite a large collection of possible participants. It is possible that a collection of possible participants could be found from subscribers in specific domains of research. What we would not normalise between the two groups would be the training received, as we would want to see how training can affect results.

However, the problem is more complex than it is described above. The ideal way of approaching this question would be to ask respondents to

try to reproduce published computational studies. However, this is complex in the sense that we would need to find participants who would be willing to take time away from their own research in order to attempt to reproduce published research. There have been studies investigating the reproducibility rate of published research in particular domains, such as the Reproducibility Project: Cancer Biology run by the Center for Open Science, whose publishing partner is *eLIFE*. However, this was not an observational study of the same sort as the one we are describing. The cancer reproducibility project run by *eLIFE* was conducted by paid personnel whose job was to attempt to reproduce the cancer studies in question.

As explained above, trying to attempt an evidence-based study with volunteers to reproduce published research is not trivial. That is why, for the most part, we see research of this nature to be in the form of self-assessment questions to survey participants.

4

Enabling the Reproducibility of Computational Experiments within Journals using Interactive Figures

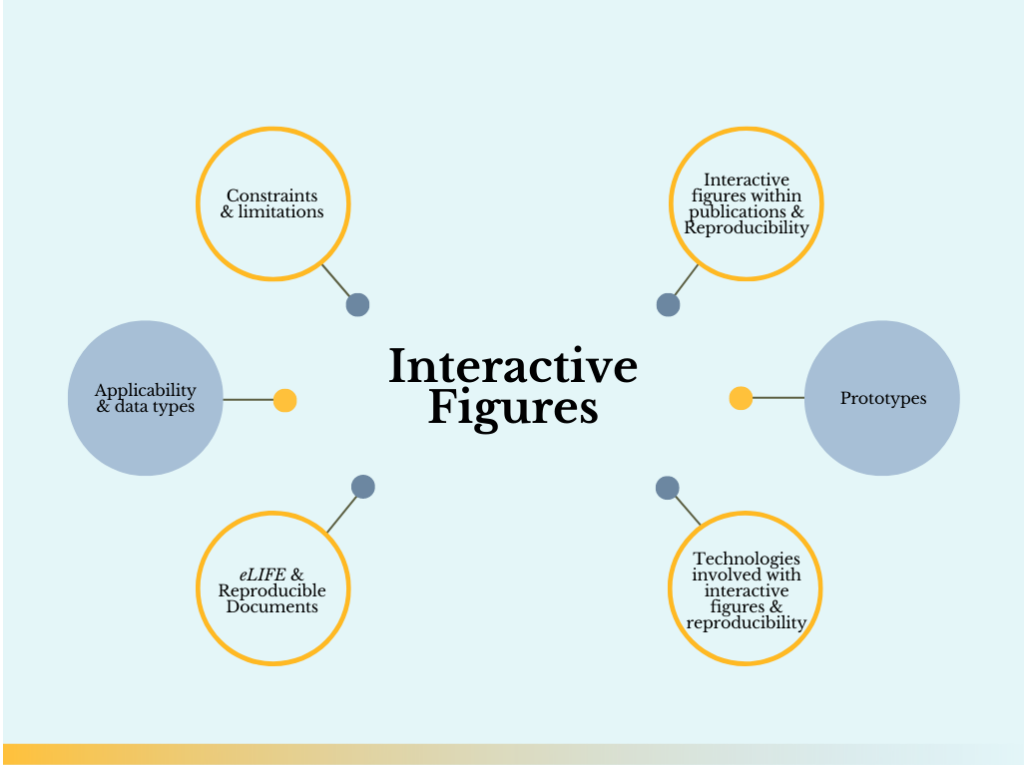

Figure 4.0.1: *Concept Map for Chapter 4: Interactive figures for enabling the reproducibility of computational experiments within journal articles.*

The work contained in this chapter contributed toward an open access article, published in *Frontiers in Research Metrics and Analytics* [doi:10.3389/frma.2021.678554], of which I am the first author. Moreover, the work in this chapter was carried out as part of a collaboration between my PhD iCASE industrial partner, *eLIFE* Sciences Publications, Ltd, and myself at the Earlham Institute. All findings reported in this chapter are directly derived from my work. A copy of this paper is provided in Appendix A.

Synopsis

With many tools available for enabling the reproducibility of CEs, one might think that this issue would be a problem of the past. However, given the integral part that research articles play in the state of science today, where despite the move from paper to electronic journals, they are still the leading route in the dissemination of science, more focus needs to be placed on the role of publications and scientific journals in ensuring and promoting reproducible science.

In this chapter, I will:

- discuss the role of publication standards and reproducibility
- describe the efforts of particular scientific journals working in promoting reproducibility through certain technologies
- argue how interactive figures can promote reproducibility of CEs within journal articles
- present prototypes of interactive figures I have developed

- explain the constraints and limitations of interactive figures with regards to promoting reproducibility within journal articles

4.1 Introduction

"Publish or perish" is a well-used term to describe the academic publishing lifecycle, where researchers need to be producing papers as a measure of success. This phenomenon has also been discussed in my first author paper (Samota and Davey, 2021). Perhaps unsurprisingly, it has been suggested that the majority of authors are most interested in publishing their results rather than preparing data for sharing (Seringhaus and Gerstein, 2007). To what extent is the publishing sector responsible for good quality, well-annotated data and metadata to be provided in articles?

Research papers ideally should deposit data within LS databases and comply with specific standards and prerequisites before publication that relate to the quality and availability of their data. Do journals ensure that any software presented in the papers is shared openly, well-described, reproducible and deposited in appropriate repositories? All these questions relate to the problems contributing to the irreproducibility of CEs that could be addressed at the publishing stage.

Given that the journal article is still the main currency of LSs, arguably, journals should be providing infrastructure and support to researchers so that their work becomes more reproducible. For example, journals can better serve their readership by having improved quality control checks to ensure that the following elements are shared before accepting a paper:

- data (including the raw data)

- well-described metadata (including using standardised terminologies)
- data deposited in appropriate databases and repositories
- well-documented software publicly available in appropriate repositories such as GitHub
- well-described methodology for analysis, including the software used and chosen parameters

For brevity, these elements will be referred to as ***reproducibility elements***.

As the research described in this chapter is in collaboration with the *eLIFE* journal, the following section discusses *eLIFE*'s reproducibility efforts.

4.1.1 *eLIFE* and Reproducible Publications

eLIFE is an open-access journal, a supporter of openly sharing data and concerned with the problems of irreproducible research. *eLIFE* is the publishing partner for the Reproducibility Project: Cancer Biology run by the Center for Open Science, where high-impact cancer studies are assessed to evaluate those which are easily reproducible. Before starting the project, the journal identified that there would be certain limiting factors, namely the expertise, experience, and motivation of the laboratories undertaking investigations specifically to replicate complicated experiments. Ultimately, the project's scope is to define steps promoting robust science, reproducibility and the advancement of this field that would benefit patients.

Moreover, *eLIFE* emphasises the critical role that publishing journals play in reproducibility (Morrison, 2014). False, incorrect or imprecise

conclusions are not only dangerous on their own but also contribute to the perpetuation of false evidence as other laboratories, by the power of suggestion, produce the same results that are false (Wagers et al., 2002). Within its peer-reviewing practices, and review of the papers' statistical analyses, *eLIFE* aims to identify unclear and possibly erroneous claims before publication.

4.1.2 Journals and their Efforts in Reproducible Publications

Scientific journals are responsible for ensuring that what they publish meets specific standards and is reproducible. The minimum requirement is that the methodology and results, including data and software, are shared publicly and openly in public repositories or databases. The exception is where studies produce embargoed data, concerns about commercial and clinical or pharmaceutical benefits, or privacy and sensitivity.

Since the introduction of FAIR data requirements, more journals are requesting data to be provided openly. However, open data is only one aspect of FAIR, and indeed, how much are the FAIR data guidelines followed in the publication cycle of journals? We still see papers stating "data available upon request". This can be an issue, especially when authors move from institutions and cannot be reached via email or other correspondence. Similarly, any code or software should be shared openly in a public repository such as Github. An exception may be code protected with patenting or commercial interest, but sensible software licensing can provide security as well as flexibility for reuse.

Ensuring that code shared in publications is reproducible is more difficult, however. Editors and reviewers cannot guarantee reproducibility unless they devote the time to download, install, and take

the necessary steps to ensure the code works on their system and then run it with the available data, validating any interpretation made in the paper. Even if this were common practice, this approach does not come without pitfalls. The task is inevitably time and labour-intensive, and editors and reviewers cannot usually spare this level of time because they are involved with their research and typically provide their time for free.

There has been a strong movement towards open access and reproducible science in the past 3 decades. However, there are still many problems that hinder research reproducibility. One such problem is the availability of graphical solutions to aid the interpretation of research findings within journal articles. In the following section, I present prototypes of interactive figures residing inside a research article that can serve as a platform for reproducing CEs *in situ* to improve the reproducibility of experiments.

4.2 Interactive Figures as a Potential Solution in Reproducing Computational Experiments within Publications

Conceptually, an interactive figure within journal articles, as an alternative to static figures, is a technical solution to allow access to raw or processed data, data analysis steps (including the open source code underlying the analysis), and allow readers to examine the results depicted in the interactive figure dynamically (interactively). As such, the reader would be able to see the effects of changing the analysis steps in real-time and modifying code and see how the presentation of the results will be altered.

Gentleman (2005), proposed that the reproducibility of research could be improved by authors publishing the list of activities they followed to produce their figures and plots. This description of methods would usually be in a methods section of an article. Still, interactive figures aim to push this idea further by representing that methodology in the figure itself. In this way, interactive figures aim to make the results described in an article more understandable to the reader, as the reader can interact with the computations and make decisions on the steps. The innovation is that the user could have access to the data, run the workflow and visualise the results interactively within the online article, all whilst reading the paper itself.

4.2.1 A Summary of Projects Representing Efforts in Reproducible Publications

Numerous initiatives have sought to solve some of the technological challenges associated with reproducibility by making it simpler for authors to distribute reproducible data and workflows. Such examples are the F1000 Living Figure (Colomb and Brembs, 2014) and re-executable publications (Ingraham, 2017; Perkel, 2017) using Plotly (plot.ly) and Code Ocean widgets (codeocean.com); the Whole Tale Project (Brinckman et al., 2019); the ReproZip project (Chirigati et al., 2016); and Python-compatible tools and widgets (e.g., interactive widgets for Jupyter Notebooks with Binder).

There seems to be a mismatch between the creation of reproducibility enabling tools and their adoption by the broader scientific and publishing communities. Some examples of open-access repositories for scientific material (including datasets, code, figures, workflows and reports) are Zenodo (zenodo.org); Figshare (figshare.com); CyVerse

(formerly the iPlant Collaborative) (Goff et al., 2011); myExperiment (Goble et al., 2010); Galaxy (Afgan et al., 2018); UTOPIA (Pettifer et al., 2004, 2009); GigaDB (Sneddon et al., 2012); and Taverna (Hull et al., 2006; Oinn et al., 2004; Wolstencroft et al., 2013). Some workflow description projects are the Common Workflow Language (Amstutz et al., 2016), and container systems include Docker (docker.com) and Singularity (sylabs.io) (Kurtzer et al., 2017). While these technologies are readily available and appear to solve many obstacles related to technical reproducibility and reproducibility culture, they have not yet become an integral part of the LSs experimental and publishing life cycle.

Therefore, it may be useful to consider the question, "how do scientists see their position in the creation and consumption of scientific outputs?" A figure or graph is an often-used method for researchers to convey their data, analysis, and conclusions, and scientific figures are frequently presented in publications as static pictures. Static images do not include access to data (including raw and/or processed data), investigation design, code, or a description of how the software used to create the figure was configured (Barnes and Fluke, 2008; Grossman et al., 2016; Newe, 2016; Perkel, 2018; Rao et al., 2017; Weissgerber et al., 2016). For paper readers to examine in detail and reproduce the published results, they must download a complete copy of the data, code, and any associated analysis methodologies (data pre-processing, filtering, and cleaning, for example) and replicate it locally. This is, of course, provided that all of those elements are available and accessible (Stodden et al., 2016).

Increasingly, LS studies involve the execution of specific software, which may require configuration and parameterisation, and operating system and library dependencies. This is a potentially lengthy process, and

reproducibility of computing studies is not always feasible as a result (Kim et al., 2018; Stodden et al., 2016). Thus, systems that recreate computational studies automatically and enable detailed examination of the data and code displayed in an image would be beneficial (Peng, 2011; Perkel, 2017, 2018).

Solutions exist that enable the reproducibility of computational analyses outside of the research paper. These solutions are typically provided as links within the research paper that redirect to a variety of different types of computational systems, such as Galaxy workflows, and Binder interactive workspaces converted to Jupyter notebooks using GitHub repositories (Ragan-Kelley et al., 2018) and myExperiment links (Goble et al., 2010). Often, these technologies include graphical figures or plots, which may be interactive, allowing for adjustment of the plot type, data filtering, and regression lines, among other things. While some figures may be interactive in that the viewer may alter certain aspects of the visualisation, this does not imply that the data or code used to create them is more accessible and therefore more reproducible.

Technologies that enable the public disclosure of code, data, and interactive figures have matured. Jupyter notebooks, for example, are composed of executable "cells" of code that can contain a link to a data file held on a cloud storage service, code to retrieve and analyse the data file, and code to generate an interactive figure illustrating the dataset. Again, this is relatively unrelated to the study article itself. However, as data storage capacity, computational power on the web via cloud services, and the ability of such services to execute code have increased, we are now at a stage where interactive figures inside papers are possible. These interactive figures, which have intrinsic access to the underlying data and analysis, can provide readers with a unique capability enabling the research's reproducibility. This synthesis of data,

code, analysis, interactive image visualisation, and paper constitutes an "executable document" (Ghosh et al., 2017; Maciocci et al., 2019).

Thus, interactive figures embedded in executable documents have combined data, code, and graphics. When the user interacts with the figure, for example, by selecting a cluster of data points within a graph, the user can be presented with the underlying data. Similarly, a user can alter the analysis's underlying parameters, such as a filter threshold, which will affect how the figure or document is shown (Barnes and Fluke, 2008; Grossman et al., 2016; Newe, 2016; Perkel, 2018; Rao et al., 2017; Weissgerber et al., 2016). An executable document, for example, might depict an interactive graphic displaying a heat map of gene expression under various stress situations. In a typical article (with static figures), the user would be tasked with locating and obtaining dataset references, then finding and downloading the code or technique used to analyse the data and retracing the original authors' actions (if the data and code were available). In an executable document, a user can choose a particular gene of interest within an interactive graphic by clicking on the heatmap and reading the gene expression information in a pop-up browser window (as an example). While this is beneficial for broad interpretation, reproducibility requires that this pop-up window includes a button that allows the user to import the sequencing read data used to generate the findings into a computational system and re-run the differential expression analysis. This raises several issues about how this infrastructure will be supplied, the technologies utilised to package all of the components necessary for reproducibility and the costs of deploying and sustaining such infrastructures.

Apart from these caveats, interactive figures embedded within executable documents can benefit the reader by enabling interactive consumption of research outputs, providing easy access to data, and

eliminating the need to install and configure code and parameters to reproduce the computational experiments displayed in the figure (Perkel, 2017). The approaches outlined above would assist paper readers (Tang et al., 2018) and the peer review process (Perkel, 2018).

Within online publications, there have been attempts to integrate the creation and consumption of research products. The Living Figure by Björn Brembs and Julien Colomb was one of the first interactive figures published in a literary LSs magazine (Colomb and Brembs, 2014). It allowed viewers to alter the parameters of a statistical computation behind the figure (Ghosh et al., 2017). F1000Research has published further articles that incorporate Plotly graphs and Code Ocean widgets to enhance interaction and data and code reproducibility within the articles' figures (Ghosh et al., 2017; Ingraham, 2017).

The prototype of *eLIFE*'s computationally reproducible article is designed to convert manuscripts created in a specific format (using the Stencila Desktop, Stenci.la, and saved as a Document Archive file) into interactive documents that allow the reader to "play" with the article and its figures when displayed in a web browser (Maciocci et al., 2019). The Manifold platform (manifoldapp.org) enables researchers to display their research objects alongside their publications in an electronic reader while also incorporating dynamic aspects. The *Cell* journal published an article in 2017 that includes interactive graphics created with Juicebox.js for 3D visualisation of high-contrast data (http://aidenlab.org/juicebox/) (Rao et al., 2017; Robinson et al., 2018).

4.3 Interactive Representation of Results: Approaches and Solutions

I wanted to address a specific question: how would interactive figures (or interactively depicting the presented data of a previously static image) be perceived by the readers of a paper, and how could they be a means of allowing the computational reproducibility of the experiment presented in the figure?

I planned to gain insight into this problem by building interactive figure prototypes and then investigating the best minimum viable product to demonstrate their effectiveness and the researcher's perceptions of interactive figures' benefit to reproducibility with survey studies including empirical testing. To do this, I needed to find suitable datasets and construct a framework in which to present the figure which mimicked an online publication.

4.3.1 Choice of Data for Interactive Figure

This section intends to demonstrate a use case and prototype that could provide initial points of reference for building appropriate future infrastructures to allow the experiment's computational reproduction through an article with an interactive figure.

To find a suitable use case, the following criteria were followed:

1. Data must be mature; researchers have been working with it for at least a year, meaning that they understand it well.
2. Data has to be of good quality, could lead to interesting conclusions and answer interesting biological questions and has already been

published or it is openly available.

3. Data should be genomics-based, as this was entailed by the initial proposal of my project.

4. All of the pieces of software used to analyse the data are available (are in working order, are open access), high quality, well established and in use.

5. Many of the major experimental methodologies would rely heavily on computation.

6. When visualised interactively, the data would have the potential to offer new and interesting insights.

Several papers already deposited in the *eLIFE* journal were investigated, and researchers at the Norwich BioScience Institutes (NBI) were interviewed to find suitable data for the interactive figure prototype. I contacted other on-site scientists who were interested in reproducibility in order to find suitable data for the interactive figure prototype.

To decide on the type of data to be used for the interactive figure prototypes, I wrote a piece of code to find the number of articles per category in the *eLIFE* journal at the time (2015). After familiarisation with the *eLIFE* paper corpus, I wrote a Python script available via this GitHub link https://github.com/code56/LivingFigures/blob/master/testing.py) that when run against the XML data of the total articles in the *eLIFE* repository returns the different categories of articles per tag assigned in descending order. The most popular category would be the suggested category in which to find my first use case data type for the interactive figure. The script produces results in one minute when run with a MacBook with 16GB RAM. The first six article categories in descending

order were:

1. Cell biology: 303
2. Neuroscience: 288
3. Biochemistry: 264
4. Biophysics and structural biology: 204
5. Developmental biology and step cells: 188
6. Genomics and evolutionary biology: 109

The method that *eLIFE* follows for tagging their articles is based on the suggested tag placed by the authors, and then *eLIFE* uses another "higher-order" category by which articles can be grouped in bigger categories.

Finding an appropriate use-case project with suitable data would allow the addressing of biological questions and enable the user to get involved in the data analysis, in addition to finding ways of recreating the CEs.

I interviewed many people within the Earlham Institute (EI) (formerly known as The Genome Analysis Centre TGAC) and NBI to find suitable data and to collaborate on the project. The following are only the most promising candidates whose names have been anonymised to comply with General Data Policy Regulation (GDPR):

Researcher 1, TGAC: Their main research focus is annotating various organisms' genomes. Data generated as part of this work involved a pipeline of 15 tools; on average, running such a workflow takes ten days. For this reason, this workflow is too laborious, impractical and time-consuming to reproduce within the settings of an article.

Researcher 2, TGAC and University of East Anglia (UEA): Their main research involved prostate cancer differential gene expression analysis.

The experiments were intricate, and the datasets were too large to be a good candidate for my project. There was a potential project involving cancer transcriptomics, but Researcher 2 was not directly involved in this study, so could not supply the data. As the primary researchers of the project were not on-site, and potential communication with them would be complicated, I chose not to go forward with this dataset.

Researcher 3, TGAC: The various projects they were currently investigating with their team were discussed, of which there was only one where data was available at the time. This project involved analysis of bifidobacteria differential expression of genes *in vitro*. I familiarised myself with the protocols and analysis workflow of this project, and while it seemed like an interesting dataset, I needed to ensure that the computational analysis parameters, including the experimental steps and tools, were the most appropriate for the dataset. At a later examination, I realised that this dataset was quite new and had not been studied long enough to be an appropriate dataset according to my initial set of requirements. In addition, this project did not have a consolidated analysis workflow at the time.

Of the potential candidates interviewed, the Expression Visualisation and Integration Platform (expVIP) study fulfilled all criteria (Borrill et al., 2016). The expVIP study offered a computational platform that was positively received by the wheat research community (with 300 users a month according to Google Analytics statistics, at the date this PhD project was undertaken in 2015). ExpVIP could have further visualisation assets applied and, importantly, offer the ability to test several hypotheses around the concept of reproducibility of CEs.

The study raised the following initial hypothesis: **By providing standardised workflows (one such example is expVIP) within**

a journal, I can demonstrate that there is an improvement in the reproducibility of the study and a positive contribution to the field of reproducibility of CEs.

Having found a suitable use case for the project, expVIP was studied to see which aspects could be made reproducible. expVIP was developed to integrate a large amount of RNA-seq data in plant studies, starting with wheat, to allow comparative analysis from multiple studies by interactively visualising them (Borrill et al., 2016). ExpVIP has many features, such as the ability to analyse public and user-specified datasets with a VM, without needing particular bioinformatics skills. The VM can then create a customised web browser, where one can sort and filter the RNA-seq data for visualising differential gene expression analyses. This allows expVIP to be applied to various species, facilitating for the comparison of experimental results, in this case, wheat studies.

To validate their methodologies, the authors used 16 wheat RNA-seq studies (sequenced via Illumina) across different tissues, developmental stages, varieties and stress conditions, totalling 418 samples with more than 11 billion reads. From these reads, 7.4 million were mapped to the IWGSC Chinese Spring (CS42) reference gene models stored in EnsemblPlants. Mapping all reads to one reference genome standardises the study's read-mapping process. One outcome of the expVIP analysis was that even though the number of mapped reads between samples varied, it did not hinder comparing studies with each other. As a whole, the expression profiles from the same tissues were the same between the different samples. The mapping in expVIP is carried out by Kallisto v0.42.3 (Bray et al., 2016). Moreover, the authors concluded that their data analyses resulted in the same gene expression patterns as the expected gene expression for plants like wheat, which validates that expVIP is a valuable tool for comparative RNA-seq analyses.

ExpVIP is a robust visualisation and data integration platform. Users can provide metadata for classifying their samples to the provided four categories (tissue, DS, variety, disease/abiotic stress) so their data can be automatically uploaded into the expVIP database to be grouped, filtered and compared against the other data, making expVIP a good research facilitator. On that note, expVIP currently addresses three research questions, providing: (1) function prediction of candidate genes; (2) identification of stably expressed genes that could be used as reference genes in qPCR comparative expression level studies; and (3) identification of genes expressed in "states" or conditions (using the Kallisto companion software Sleuth (Pimentel et al., 2017)), where such analyses can be indicators of the plant transcription responses to the four states.

Hereby, I propose solutions to the lack of reproducibility of CEs in the context of the three questions that expVIP can currently address. Below are the issues around CEs along with suggested solutions that were aimed to be explored via this project:

Availability of data It took the researchers involved in the expVIP study two months to gather the data for the expVIP database. Tasks included downloading the data from databases or asking the original authors to provide it. expVIP hosts 16 wheat RNA-seq studies in one platform. expVIP also allows users to upload their data to perform their investigations, which is only possible at the VM version. Given the above features, expVIP serves as a platform for making data available.

Inconsistent metadata annotation It is common for different researchers to use different terminologies for the same concepts. This makes it difficult for others to understand exactly what authors are referring to and whether the terminology is correct or not. Issues with

diverging terminologies (ontologies) are further discussed in Chapter 5. Therefore, any data not correctly annotated or not annotated with metadata will be less useful for reproducing the study results, let alone be included in further research.

Comparison of data with one reference genome The developers of expVIP took data from 16 different studies carried out over a broad period. Each of those datasets is run against the same tools and the same reference genomes for comparison. At the time of writing, there was only one completed wheat reference genome with a relatively mature set of gene annotations, but new versions of reference genomes are released as data and analysis improve. To make research more reproducible, by comparing the transcriptomics data against a reference genome (or indeed more than one), the mapping of the reads is standardised through the use of consistent file formats and tools that accept those formats, so the effective comparison of results is facilitated.

Analysing the data with a common workflow expVIP maps the reads with Kallisto. All the data are analysed using one workflow, removing any other variability factor that would hinder the experiment's reproducibility.

Interactively visualising the data allows the data to be better understood and enables more biological questions to be posed Posing biological questions at the level of complexity enabled by large-scale comparative RNA-seq studies would be time-consuming and less efficient without an integrated platform such as expVIP, and even more so within a research article. Since this platform would be available through a journal, one can address these questions whilst reading the paper, enabling the validation of results presented in a paper *in situ*.

Therefore, considering all of the five points above, I felt I had a good

starting point for using expVIP as my first use case for testing the incorporation of interactive figures within articles for facilitating the reproducibility of CEs.

4.3.2 A Framework for an Interactive Figure

The figure includes a range of User Experience (UX) features that would make it interactive: the changeable use of colours or other graphical widgets to aid interpretation; being easy to manipulate; the ability to clearly display any changes in the figure resulting from changing any parameters; being easily able to define parameters that would, in turn, change the figure; the inclusion of a "how to use the tool" link, or pop-up window; a forum/comment section for readers and authors to discuss their conclusions from interacting with the figure; import and export data and static figure options. These characteristics for the design of the framework were inspired after seeing the F1000 Living Figures and discerning what improvements could be made to improve their user-friendliness and functionality (Colomb and Brembs, 2014). Moreover, they were rated by the survey participants and deemed favourable as interactive figure features, as discussed in Chapter 3.

Having an interactive figure within the journal article not only allows for the reproducibility of CEs but also benefits the end user as follows:

1. It has the potential to allow the user to better understand the data, workflows, and results by interacting with it. This can allow the user to pose further questions and draw additional conclusions having "played" with the figure.
2. The figure could save a significant amount of time and effort for the scientist as all the data, software and workflow would be there for

them, as well as in a local form, possibly via a VM so that they do not have to worry about installing and configuring the system or dealing with dependencies.

3. The interactivity alone and comparing similar data with data from many different studies is a great way to learn and draw conclusions, as discussed in previous sections.

4.3.3 Prototype of an Interactive Figure in Drupal

I decided to develop the prototype initially in Drupal because *eLIFE*'s back-end infrastructure also uses Drupal. Following further communications with the developer team at *eLIFE*, it was encouraged that Drupal should be used for this reason.

Drupal is a free and open-source content management system written in PHP, released under the GNU General Public License, and based on the Model-View-Controller architectural pattern. It is estimated that approximately 13% of the top 10,000 websites use Drupal, ranging from blogs to governmental sites.

A Drupal site was set up to include the interactive figure and dummy *eLIFE* article. Because I was using a Macintosh, a compatible and easy-to-install solution for a local web server was a MAMP server (an environment incorporating MacOS (operating system), Apache, the web server, MySQL, the database management system, PHP, Perl or Python, and all programming languages used for web development (Wikipedia contributors, 2016)), coupled with a Drupal installation (version 7.43), mySQLite for the backend database, and PHP for the programming language. I initially used this MAMP server configuration but then moved on to using individual VMs (using Drupal 8) for the code to be

more reproducible and distributable. The code for the MAMP server I built, which included a Drupal site of a dummy *eLIFE* paper and an interactive figure, is available on GitHub via this link: https://github.com/code56/MAMP_server.git, the code for the Drupal site is available via this link: https://github.com/code56/drupal8article.git.

Within the Drupal site, I created a dummy web page styled to look like an *eLIFE* article, where I then embedded expVIP. The expVIP datasets were migrated into the mySQLite Database (which I installed in the MAMP server) in order to enable users to display datasets or their subsets based on the users' selection of filters. In the Drupal interactive figure prototype embedded in the dummy *eLIFE* I developed some additional (preliminary) features which intended to enable research reproducibility. One such feature was the incorporation of an autosuggest functionality, to suggest (for example) gene names and species names. With this autocomplete functionality, aiming to be a user-friendly feature; the user can select which genes to display in the expVIP interactive figure. Ultimately, this would have allowed the interactive figure's image to change according to the gene currently selected.

For the layout and styling of the web pages themselves, I used a sample HTML *eLIFE* article page, viewed and downloaded from the source code inspection tools available through a web browser. ExpVIP was then embedded in the main page area using a Javascript script (see Figure 4.3.1). I set up the MySQL Database, which stored the expVIP data. I ran some of the snippets of the expVIP platform with code-sniper (an npm package) to familiarise myself with the code. To make a complete working expVIP platform within the dummy article, I would have had to install the necessary BioJS components, link the data to the

interactive figure, and gain access to the expVIP server. However, at the time when I was working on this project, there were errors in the expVIP platform that the expVIP developer could not attend to and resolve at the time.

Moreover, feedback from *eLIFE* (through their readers) indicated that a more straightforward (bar-plot type) figure would have been more desirable as it could be more versatile for more data types. Therefore, I decided to move on to a different kind of interactive figure as an additional proof of concept interactive figure prototype. Nonetheless, I gained a lot of insights from the efforts in deciding the appropriate platform, tools and software for deploying a working interactive figure depicting expVIP in the dummy *eLIFE* article.

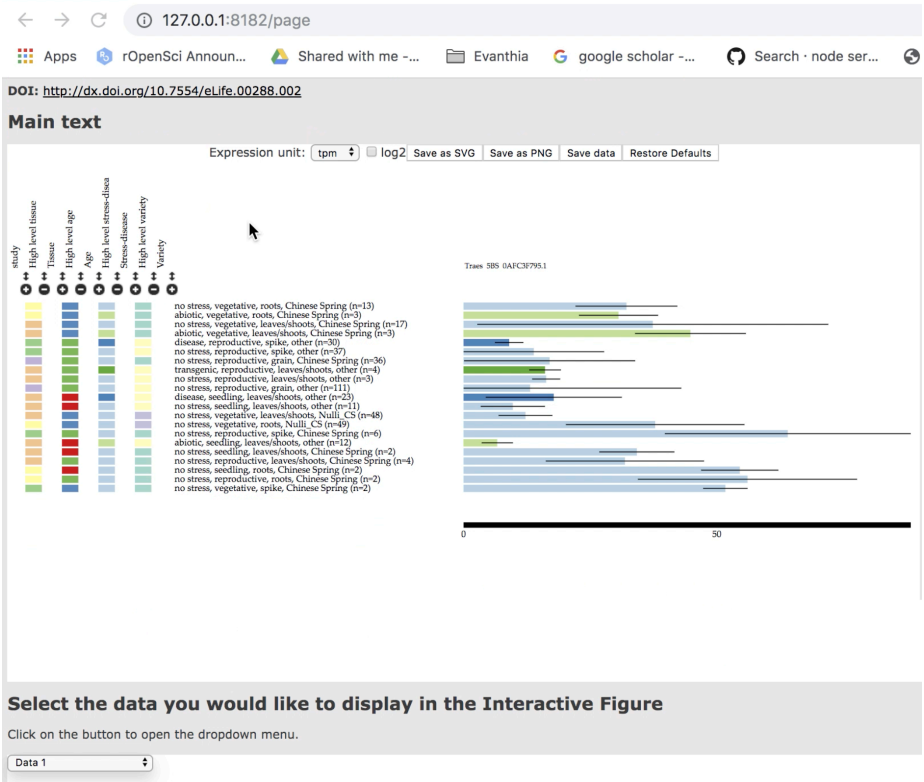

Figure 4.3.1: Interactive figure prototype. The interactive figure is expVIP embedded within a dummy *eLIFE* article within the Node.js server.

Virtual Machines to Contain the Code for the Dummy Article Page and Interactive Figure

Virtual machines (VMs) act as containers for an OS and any other software and/or data that could be used as a complete package for other users, thus improving reproducibility (see the relevant section 2.4.3 in Chapter 2 that explains this better). Installing Drupal within a VM would mean all dependencies would be provided and would remove the need for a user to go through the full installation procedures.

The VM could then be exposed to anyone using Vagrant (tool for building and managing VM environments, see Section 2.4.4. The VirtualBox (virtualisation software to build and run VMs) used was version 5.1.2, and the MAC OS version was 10.10.5. The VM can be downloaded and spun up in the user' s system to reproduce the experiments presented in the figure. This way, the Drupal VM could be made public for receiving feedback from users.

Vagrant was selected for this project as a containerisation software to create, configure and manage this project's VM because its potential as a testing platform had already been discussed with the team at *eLIFE* during an industrial placement there. Effectively, Vagrant acted as a layer of software between the VirtualBox and the VM. Benefits of using Vagrant include the fact that the figure can accept data input from the end user and can be a good tool for checking the quality of the data and the research itself before publication.

After several attempts to develop the interactive figure further in Drupal, it was necessary to move to a different means of making the figure. Drupal is a complex web environment with a steep learning curve, and it is hard to write this type of tool in this environment. Developing even simple components needs in-depth knowledge of Drupal's modular

structure, which was time-consuming and too laborious for the scope of this research.

Therefore, following further investigation of other possible solutions, I decided to develop a more stand-alone approach comprising a Node.js server. The benefits of using Node.js were that there is a more extensive community working in this software library, so it is easier to start with development. There is a lot of documentation and resources online to help develop the tools needed. Drupal could not offer a straightforward minimum viable product approach therefore, given problems with Drupal affecting the project's delivery speed, I decided to move on to Node.js.

4.3.4 Node.js Server Interactive Figure Prototypes

I hereby provide prototypes of the interactive figure architecture, comprising a Node server running an example article page that contains a wheat expression viewer from expVIP. I also provide a second example of an interactive figure comprising a Node server with an interactive BioJS bar plot.

The Node.js server is connected to the expVIP database in the MAMP server, whereby the expVIP database can be queried to select the gene names to display in the expVIP interactive figure for an interactive representation and exploration of the data displayed in the figure, which would otherwise not be feasible with static figures. The Node.js server was developed in Node version 5 ("Node 5") to maintain compatibility with the expVIP BioJS components at the time this project was undertaken.

The Project Repositories are found via the following links: node server: https://github.com/code56/node_web_server.git;

node-web-server-formula: https://github.com/code56/node-web-server-formula.git; node web server with box-plot interactive figure: https://github.com/code56/nodeServerSimpleFig.git. License: MIT [node web server] GPL V3 [box-plot interactive figure]

As mentioned earlier, the expVIP component had certain important bugs that were not addressed while investigating this prototype. Therefore, developing the interactive figure using expVIP as a use-case and adding further functionalities and making it work fully came to a halt. As a complete software package, the relative complexity of expVIP meant that such a solution would be difficult to implement in many scenarios. Therefore, in addition to the expVIP interactive figure, the BioJS boxplot courtesy of Ariane Mora was also added to the Node.js server article dummy page as an example of an interactive figure (Biojsboxplot, http://biojs.io/d/biojs-vis-box-plot. The advantage of such an approach is that the component is simpler to deploy and can be more versatile for use with a broader variety of datasets from different disciplines. BioJS is a library of JavaScript functional elements and guidelines that provide biologically driven user interface components that are easy to reuse, maintain and deploy on the web. BioJS attempts to provide an Open Source Standard for Biological Visualisation.

Figure 4.3.2 shows the BioJS bar box plot interactive figure prototypes within the Node.js server. The figure demonstrates how the visualisation of data is altered dynamically given the user's choice of display options (e.g. logarithmic, raw data, change to a line graph).

Figure 4.3.2: Interactive figure prototype Bar Box Plots. Figures (a), (b) and (c) show how the figure's data can be dynamically displayed depending on the display options chosen by the user. The BioJS Bar Box Plots are courtesy of Arianne Mora. The code behind the Node.js dummy *eLIFE* site with the BioJS Bar Plots is available via this GitHub repository link: https://github.com/code56/nodeServerSimpleFig.git

Since I completed this work, *eLIFE* released their first in collaboration with Stencila (https://stenci.la). In the section below I discuss ERA and its functions, features and limitations.

4.3.5 Executable Research Article

Stencila (https://stenci.la) has been developed by the Substance team (Substance.io) in collaboration with eLIFE to build an ERA. Computational reproducible papers which used to be part of the Reproducible Document Stack Project are currently publicly available as an ERA. This initiative aims to enable reproducibility and transparency and to enrich published work interactively. This is achieved via live code blocks and programmatically generated interactive figures using R

Markdown and Jupyter in combination with Stencila Hub's interface. Future plan development for the ERA project is the creation of a Google Docs plugin to allow authors to insert data blocks and executable code in their manuscripts using the Cloud Service.

Stencila is a document editor intended for research projects. Stencila documents are self-contained, comprising all the text, media, code and data representing the author's research. Stencila removes the need for manual content conversion from source documents (Microsoft Word, Google Docs) to XML and web HTML format for publications (eLife Labs, 2018).

Stencila Desktop (https://stenci.la/use/install) is a prototype of a "researcher's office suite" where researchers can edit their research documents and incorporate their research text, media, data and code. Stencila has its built-in language, Mini (https://stenci.la/learn/languages/mini/), which is a "minimal, functional language focused on data analysis and visualisation". Mini is executed in Javascript with easy-to-understand expressions which are only moderately more complex than those of a spreadsheet cell or a calculator. Stencila's interfaces run in the browser and are available on all major operating systems. Stencila Articles and Sheets offer a spreadsheet-like programming experience (see figure below).

Texture is an open-source manuscript editor, which incorporates all the standard static components of a manuscript (abstract, main body, figures, citations), and is used for visually editing Journal Article Tag Suit (JATS) XML format documents which are extensively used by journals to describe scientific literature. The Stencila article editor offers an extension of the functionality of Texture, with additional code cells that allow for interactively manipulating code and data. The Stencila

code cells can range from simple Excel expressions to executing code in the currently supported languages (Javascript, R, Python and SQL). Datasets can be referenced from within the Stencila publications and can be used, for instance, to create a scatter plot. Such examples can be seen in this Stencila article available via this URL: http://builds.stenci.la/stencila/fix-regressions-2018-06-01-142af5d/example.html?archive=introduction.

Stencila stores projects in an open-source file archive format called Dar, available from https://github.com/substance/dar. Dar is a folder with various files, including the manuscript itself (usually one XML per document) and all relevant media (such as figures) (eLife Labs, 2018).

ERA readers can inspect code, modify and re-execute it, alter plot formats and change the data range of specific analyses, all within individual web browsing sessions, which means that their actions do not affect the publication. Readers can also download the ERA publication with all embedded data and code.

4.4 Discussion

In the survey results described in Chapter 2, respondents to our survey expressed how they would find it beneficial to have their research results interactively presented in journal articles. In this chapter, I have presented the prototypes of interactive figures, including features reflective of the respondents' opinions and preferences. Moreover, different data types and computational experiments were explored and presented, as well as other technologies and projects involved in the field of reproducing CE within journal articles. There are many benefits in presenting CEs interactively within research papers. These include

allowing readers to view the data which contributed to the figure and the ability to zoom into data points that an otherwise static image could not provide.

Will interactive publications become the mainstream means of presenting research in LS? Are interactive figures and reproducible documents the next standard for publications? Although they present many benefits, one must consider that there are many more data and experiment types than those covered by the current Reproducible Document examples (Maciocci et al., 2019; Colomb and Brembs, 2014). Computations take days, weeks or months in HPC setups, meaning complete re-analysis cannot be easily reproduced within journal articles. However, for figures which allow interactivity to assess already computed data points, such as those examples listed on the Stencila Hub website, it is more advantageous for them to be presented interactively instead of as static images.

Nonetheless, one notably challenging aspect of reproducible documents and interactive figures is the question of uptake from the scientific community. Would publishing research results interactively involve a significant amount of effort for the researchers, besides the already time-consuming process of writing and publishing research papers? Stencila, for example, is written in Mini. Even though Mini is advertised as an easy language to code, it is still a new skill that researchers would need to acquire; further, it is not particularly advantageous, given that Mini is not a commonly used language, and would not help them in other aspects of their research (apart from coding for Stencila Documents). The Reproducible Document Stack announced their future development of a Google Docs plugin for converting their manuscripts into Reproducible Documents more easily; such an update could seemingly help researchers by providing a more straightforward

conversion of their research articles to ERAs format. However, many authors write their research articles using other systems, such as LaTex for example. In other words, the provision of a Google Drive plugin will not be a panacea for incentivising researchers to convert their articles into ERA. Thus, if researchers have to do part of the coding of their figures for them to be converted into reproducible document-compatible figures, meaning that the scientists would have to learn to code in Stencila (using Mini) (or other languages compatible with other journals), that would likely be burdensome for the researchers by adding more workload steps and processes before the publication of their work. Surely the researcher would also want to approve how the interactive figure (presenting their work) looks (ensuring it is accurate and free from errors), similarly to how researchers currently approve how the static images will appear in their published articles.

Another key question is whether the journals' decisions on the type of data and experiment type they choose to present as interactive figures would alienate scientists from other communities or disciplines and make them feel left out or unrepresented. At least for the production efficiency of such figures, the journals' developers would need to be focusing on producing one type of data set/computation per project round. For example, the first Reproducible Document examples published by *eLIFE* are bar plots produced with R. Potentially different releases and updates of reproducible documents would represent different types of data from other scientific disciplines. Furthermore, how long do interactive figures stay live and up to date? Surely, as journals change their backend software and configurations, these figures might also need to be updated. This requires further time, labour and costs, so reproducible documents need to be tested long-term before we can understand the full costs (time, monetary, storage costs) of such

functionalities.

What would be the computational cost involved with running such interactive figures, especially with more complicated data and experiment types? Would the journals be willing to endure that extra cost? Would the journals need a higher fee for reproducible documents, and if so, would researchers be willing to pay the extra fee for their publication to be represented interactively? I believe that unless there is an incentive in terms of career progression, such as measurable metrics in terms of an increase in the reach, readability and impact of their research, the majority of researchers will not want to pay an extra fee or even choose one journal over another simply because of their capacity to publish their figures interactively. This belief has been confirmed by our survey (see Chapter 3, Results, figure REF), whereby the majority of respondents expressed that being able to publish interactive figures would not affect their decision in terms of which journal they would choose to publish their articles.

Reproducible documents are a way of developing a level of modularity to conduct an analysis of biological data if it could be possible for frameworks to be built around the reproducible document. From my work, I have found developing these frameworks can be difficult even with existing software and raises key questions about development cost, overall benefits, and user uptake. The options are for you to develop the framework in its entirety or use a ready-made one; it is an expensive process, but then you still need to get the research community to buy into it. Many research articles cite their data and workflow analysis presented in the paper as being available in Galaxy. Whilst this is an effective way of making their work publicly available, and in a reproducible manner, this involves users having to "leave the paper" and visualise the data interactively within a separate web page (in this

case, Galaxy). However, as Galaxy is already a popular web-based platform within the LSs community with many users and an established system for scientists to access the workflows of others and analyse their data, issues such as the costs of rendering interactive figures could be mitigated.

In theory, it is hard to achieve reproducibility when writing code for research purposes in CEs. The reproducibility of research software development (coding) is typically not a priority for researchers. Most will have the primary goal of building a piece of software that will lead them to the biological interpretation of their data. They want to get to the results as fast as possible because again there is the pressure of "publish or perish". Therefore, recording what packages were installed and what version of tools were used is not always the priority of research programmers. Based on my interactions with the *eLIFE* development team, it seems to be the case that, outside of academia, professional developers who are working collaboratively on projects are used to and need to be working in a reproducible manner and pay more attention to provenance. Within commercial/industrial development, full teams of testers, analysts, change managers and designers are also employed. These positions are very rare in academia.

Having said that, there are many projects and infrastructures that support and enable the reproducibility of research, and some with more uptake from the scientific community than others. The complication arises from the fact that there are so many of them. How do researchers know which solution to choose and subsequently know it is the best solution for potentially many to fit their needs? Having to train researchers in how to use all of the available options would also be unfeasible.

Although interactive figures within journal articles represent a useful means of presenting data, they are not a substitute for the importance of performing robust and reproducible science, to begin with. It is the responsibility of scientists to ensure they design their research correctly (e.g. good experimental design including strong statistical power and appropriate sample size), collect and describe their data and metadata appropriately and in standardised formats, write reproducible code and provide all the details in their methodology, upload their data, metadata and code to the appropriate online platforms (e.g. databases) and have those elements available as links in the journal.

As should hopefully be obvious, this requires a large amount of work. In other words, I believe more emphasis should be placed on training researchers in reproducible and robust science rather than prioritising creating interactive figures because interactive figures alone cannot be the mainstream means of disseminating publications for the reasons discussed above. Two examples of projects focusing on training students, their supervisors, funding bodies, publishers and editors in reproducible science are the Turing Way, Data and Software Carpentry. More information on how training can help with the reproducibility crisis in the LSs will be presented in Chapter 6.

4.4.1 Limitations

Drupal has a steep learning curve and limited online educational resources for new Drupal developers to learn from and solve coding problems. For example, getting the interactive figure modularised and functional in the Drupal environment involved certain technical difficulties that were time-consuming to resolve and overcome. Therefore, a significant amount of time was spent attempting to create a

product in Drupal before changing direction and using Node.js.

Due to time constraints, the code I wrote within the first half of my studies has not been revised to ensure that my interactive figure prototypes still work properly. I am assuming that if specific Javascript dependencies are broken, not supported, or obsolete, that might affect the functionality of the interactive figure.

There was a scheduled interruption in my studies, during which time eLIFE took up a collaboration with Stencila for the development of ERA. Whilst this has been an unfortunate situation in terms of the progression of my work with interactive figures, I have learned a lot about building reproducible interactive figures, and the research I completed has been relevant to the field of interactive figures and reproducibility of CE within journal articles. I then spent time exploring a different path for my PhD involving the reproducibility of research in relation to ontologies and automating the process of finding standardised ontological terms within journal articles and their respective metadata. This line of research will be discussed in Chapter 5.

Ultimately, I believe the experiences I gained whilst developing the interactive figures project gave me a good insight into how difficult it is to reliably create a truly reproducible environment within research articles and publisher frameworks.

4.5 Future work

The challenges of this project were in demonstrating that expVIP as a standardised workflow would (a) enable the reproducibility of its containing experiments, (b) be a tool that authors would want to use and (c) improve the reproducibility in the wheat transcriptomic field. To

assess the success of such an effort, surveys are required to canvas the authors' opinions on interactive figures in publications and the use of existing algorithms to compute the popularity of the tool and whether readers found that they could reproduce the results reliably. The aim was then to approach volunteers and have them test the figure and receive user experience feedback. However, due to the time constraints and the interruption to the study as explained above, this testing has not been undertaken. It is important to provide a caveat in this case, as collecting citation rates and other metrics requires the collection of data over an extended time frame and that could not have been completed within the period of my PhD studies. Nevertheless, the outcomes of such a study could provide us with key information such as how many people read or prefer reading articles with interactive figures and how many interacted with the figure. Additionally, it would have been useful to include pop-up questionnaires asking readers whether they benefited from the interactive figure (i.e. it helped them understand the study better) and to record how many people tweeted or blogged about the paper with the interactive figure in order to assess the popularity of interactive figures within journal articles within the LSs scientific community and how many more authors would want to publish their research using interactive figures. However, it would still be difficult to determine whether any improvements in the paper's readability would be due to the interactive figure and not the context of the article itself.

Based on the outcomes of this chapter, I see there are four core areas of potential future work:

(i) Evaluations of the interactive figure prototypes using an empirical survey to gather opinions from authors and consumers of scientific research articles would be a helpful exercise, enabling an assessment of the perceived value of interactive figures within journal articles in

enabling the reproducibility of the research depicted in the interactive figure.

(ii) In the survey we conducted (see Chapter 2), the final question we asked respondents was to include their email addresses should they wish to be contacted to give feedback on the interactive figure prototypes I made. I received 10 responses from local researchers (in the NBI) and conducted a very informal meeting with five respondents who played with and tested the Node.js interactive figure with the expVIP platform. The results were not analysed due to the small size of the sample.

(iii) A more formal survey, with well-thought-out questions, assessing respondents' opinions on interactive figures, having been presented with different types of interactive figures (e.g. different data types, different methodologies), would be a beneficial project in thoroughly understanding the users' opinions on interactive figures. Making the interactive figures publicly available (i.e. making the local server accessible from anywhere) could allow for participants from different countries, universities, and research institutions to give opinions on the interactive figures so that the pool of participants is composed of users from different backgrounds, scientific fields and expertise.

(iv) A beneficial further development of interactive figures would be to allow for the complete compartment of the interactive figure to be downloaded from within the research papers, in the form of VMs, to allow users the ability to add their data into the interactive figure and run their data within the computational workflow presented in the papers' interactive figure. Allowing users to download the interactive figure and upload their data locally can prevent issues with overloading the servers of the respective journals that render the interactive figure, which again would mean increased costs for the journal and the

possibility of cyber-attacks on the journals' servers. This could also mean that researchers can further the work presented in the research paper; in other words, they can build upon the work of others to work collaboratively (even if not officially collaborators to a project) in progressing the scientific field.

5

Ontology Annotation Standards in Publications and Databases and Reproducibility Metric Scores

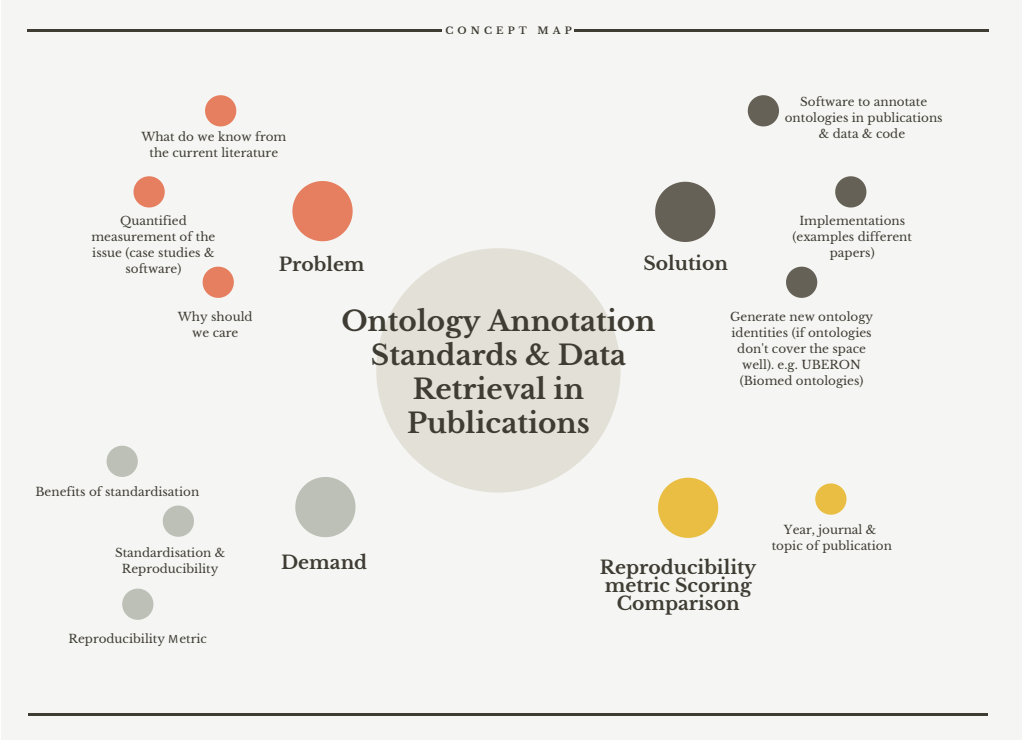

Figure 5.0.1: *Concept Map for Chapter 5: Ontology annotation standards within publications and associated database entries; and Reproducibility Metric Score (RMS).*

Researchers find data in one of two ways: by searching through data repositories or search engines or by reading through research articles

and manually looking for and downloading the data associated with the papers. In many cases, the data is not freely available, and when it is available, it may not be correctly standardised and formatted or described with the required metadata for reanalysis.

A bioinformatician, whom we hereby call Bob, wants to carry out an RNA-seq (transcriptomic) analysis and a comprehensive comparison between different plant species to evaluate the expression of a gene in different developmental stages, plant tissues and stress factors.

To achieve this task, Bob was able to find some datasets by searching the EMBL-EBI European Nucleotide Archive (ENA), but from reading recent literature, he finds that he has to manually search through all papers on the plant species of interest that involve RNA-seq data published between the arbitrary time frame that he has defined. Reading through the methodology of the papers, he finds references to available datasets and notes these down in a spreadsheet, one dataset per row. He then tries to mark each row in his Excel file with extra information, such as the plant species, developmental stage, stress factors, and the study's tissue involved. From there, he looks at the "Data Availability" sections of the papers, trying to find links provided to online databases to download the relevant data and metadata files from the EBI database.

This might seem like an effective way of performing this task. However, in reality, it is not straightforward, primarily due to the lack of standardisation of data and metadata presentation in publications and the lack of integration of scientific databases into the publication lifecycle. As explained in the Introduction of this thesis, lack of standardisation in data and metadata presentation are elements leading to the irreproducibility of experiments.

5.1 Manual Process of Finding Use Cases for Comparison Tools

To find the case studies to include in the expVIP visualisation platform (Borrill et al., 2016), the expVIP authors began their search at the NCBI Sequence Read Archive (SRA) for terms "wheat" or "Triticum aestivum" (the scientific name of wheat). SRA contains raw sequencing and alignment data. The manual process of finding appropriate case studies for expVIP included filtering for "RNA-seq" and manually going through every sample. At the time the expVIP researchers were collecting the case article studies, there were 400 samples; currently, there are over 1500 samples in the NCBI Sequence Read Archive (SRA). Using the SRA number, they conducted a Google search to see if they could find any publications associated with the SRA number. However, this was not a very successful approach. The expVIP researchers found that examining for the SRA or Project Accession Number (PAN), again using Google, was somewhat more successful in finding appropriate case studies for their platform. Then they manually curated the retrieved samples to ensure they were the correct ones by reading the papers (e.g. the number of samples, tissues etc). Following those steps, they similarly used Google with the name of the author or the institution who had deposited the data in SRA and some keywords (e.g. wheat, tissue, or experimental conditions), and they were able to find the associated SRA sample. The researchers found that, sometimes, the person who deposited the paper was not on the authors' list, which made the search more complicated. This often resulted in finding a published paper that seemed likely to be associated with the SRA samples by verifying the tissue and growth conditions. Often it was not that easy to tell, and sometimes people deposit extra samples that are not in the final paper.

Using this approach, the researchers (Borrill et al., 2016) could identify a paper for 95% of samples. Nonetheless, finding appropriate case studies for the expVIP platform described here involved considerable manual curation and time.

Researchers are often tasked with subjectively understanding what authors may mean by a particular terminology, e.g. "30 days post-anthesis" in a plant developmental biology article. Whilst someone might be able to approximate or have an idea of the author's intent, this plant development stage description is not a standardised term which can be universally and unequivocally understood and accepted. When such data and metadata are presented in papers without a deeper understanding of what the author meant, it poses difficulties in using the material for re-purposing analysis and may well affect the ability to carry out cross-species studies where different descriptions ("terms") might be used.

We use ontologies to try and solve issues associated with the lack of standardisation of scientific information presented in papers and datasets. In this chapter, I explain the concept of ontologies, why ontologies are necessary for data standardisation, and the importance and usefulness for reproducibility in standardising the use of ontologies in publications and data. Moreover, I present, qualify and quantify the lack of ontology annotation standards in publications and datasets using real-world case studies and a software prototype I created (*Deus ex machina*) to measure the issue. I also propose a proof-of-concept programmatic solution to the lack of ontology annotation standards and data retrieval in publications which can be used to automate biocuration tasks.

5.2 The Lack of Ontology Annotation Standards in Publications and Databases as an Issue Negatively Affecting the Reproducibility of Experiments

5.2.1 What is an Ontology?

In information and computer science, an ontology is a structured vocabulary of formal naming, allowing for the identification of categories in bodies of knowledge, data and conceptual entities and their inter-relationships (Cooper and Jaiswal, 2016). The purpose of ontologies is to represent the properties that a subject domain encompasses, as defined by the categories or "classes" and "subclasses" to which it belongs (Stevens et al., 2010; Smith et al., 2007).

5.2.2 Ontology Development and Organisation

Many different groups develop ontologies around the world in an attempt to provide consensus within or across scientific domains. Despite this intention to standardise a domain, there are many different ontology technologies and file formats, with varying descriptive complexity and many ontology construction conventions, so there are accompanying efforts to harmonise these into comparable formats to aid consistency when using the terms within them.

One such harmonisation effort is the Open Biological and Biomedical Ontologies (OBO) Foundry (http://www.obofoundry.org/), an organisation constructing and maintaining ontologies in the LSs. At the time of writing, the OBO Foundry includes more than 150 ontologies

submitted by community members. Within this list, ten have been manually reviewed by the OBO Foundry, including the Basic Formal Ontology, Human Disease Ontology, Gene Ontology, Phenotype Ontology and Plant Ontology, amongst others. The mission for the OBO Foundry is to standardise the description of biological data and allow the data to be more easily comprehended and better interconnected amongst distinct research disciplines. This way, we can aggregate and compare plant ontologies amongst different species.

Services have been built around ontologies to make them more generally accessible through aggregation and search functionality. For example, the Ontology Lookup Service (OLS) (http://www.ebi.ac.uk/ols/ontologies) is a search engine to find ontology terms. OLS is part of the ELIXIR infrastructure and includes 257 indexed ontologies at the time of writing. The list includes long-standing ontologies in use by many communities, such as the Agronomy Ontology (AgrO) (https://www.ebi.ac.uk/ols/ontologies/agro) for agronomic practices and agronomic techniques, and EDAM (https://edamontology.org/) for describing bioinformatics operations, data types, formats, identifiers and topics.

Apart from the ontologies listed in the OBO Foundry, other ontologies exist which are utilised in the LSs. These include crop ontologies (https://www.cropontology.org/about). Crop Ontology is part of the Generation Challenge Programme (GCP) (http://www.generationcp.org) storing agriculture-related data and metadata associated with, among other things, phenotypes and germplasm. GCP also developed Trait Dictionaries for breeders for data collection and annotation using ontologies in a standardised way. The crop-related ontologies were developed in 2008, starting with chickpea,

maize, Musa (edible bananas and plantains), potato, rice and wheat, and most recently cassava in 2010.

5.2.3 Standardisation in the use of Data and Metadata Description in Research Papers and Databases

As there are many examples of ontologies and their use in data and metadata description, I will focus on one example of ontology, the Plant Ontology (PO), to describe the most common standardisation issues.

The PO (https://www.plantontology.org) nomenclature was developed as a way of characterising plant anatomy and morphology (Ilic et al., 2007) and the whole-plant growth/developmental stages (Pujar et al., 2006) of all green plants. The PO includes plant genomic annotations corresponding to PO terms (Cooper and Jaiswal, 2016). The PO was first established to describe the growth stages and anatomy of *Arabidopsis thaliana* (thale cress) (Garcia-Hernandez et al., 2002), maize (Vincent et al., 2003) and rice (Jaiswal et al., 2002).

The PO comprises two entities of classification serving as two main branches: the *Plant Anatomical Entity* and the *Plant Structure Development Stage*. Examples of Plant Anatomical Entity are "whole plantome" and "plant cell", and an example of the Plant Structure development stage is the "reproductive shoot system development stage". Ontological concepts/classes organised closest to the root of the tree are the most generic (Cooper et al., 2013) (see Figure 5.2.1).

Ontologies are Important in Standardising Data and Metadata Descriptions

Recent studies have highlighted that metadata is critical to the understanding and reuse of datasets (Löffler et al., 2021), so the

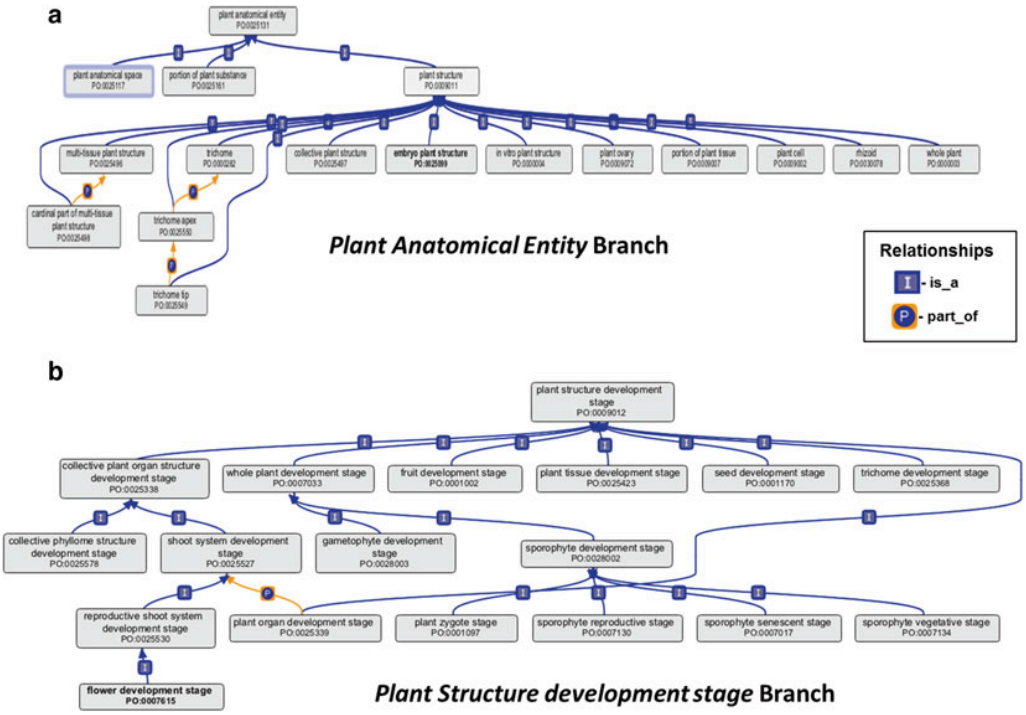

Figure 5.2.1: *Upper-level structure of the two branches of the Plant Ontology. The Plant Ontology (PO) is made up of two interconnected branches organized hierarchically under a root class Plan Anatomical Entity or Plant Structure development stage. The classes are linked by relationships (indicated by the direction of the arrows), with the most general classes positioned towards the top. The examples above are only the "is a" and "part of" relationships. This image is based on PO Version #20 (Released August 2013) and was created using OBO-Edit (Day-Richter et al., 2007). (a) The Plant Anatomical Entity branch has three upper-level subclasses: plant structure, plant anatomical space, and a portion of plant substance, which together describe all the anatomical and morphological parts of a plant. The largest subclass plant structure has 13 direct subclasses, which each have many subclasses (not shown in the figure). (b) The Plant Structure development stage branch describes the developmental stages of all plant structures, including the whole plant. There are six direct subclasses, which describe the developmental stages of the classes of plant structures.*
The figure and its original caption are both reproduced verbatim with permission from Cooper and Jaiswal (2016).

importance of ontologies is apparent when considering the requirement to organise and annotate the plethora of data produced in the current Big Data era, where the number of biological datasets is increasing exponentially (Greene et al., 2014). Mapping ontological descriptions in databases is a vital step for comparative studies of cross-species phenotypic and genotypic studies as well as gene discovery experiments,

ultimately allowing for easier access and integration of data (Shrestha et al., 2010).

Ontologies and their terms and relationships convey a rational structure relative to one another, much like the relationship between object classes. This is particularly beneficial for the computational comprehension of these structured vocabularies and the utilisation of ontologies for efficient study design and execution (Hollmann et al., 2018).

To explain this use of ontologies further, we can consider the example of readers of research papers; where a paper does not use the standardised ontological terms to describe its methodology, readers cannot distinguish at what developmental stage the experiment was performed. The paper's lack of standardised ontological terms ultimately makes the paper less reproducible since readers cannot know the conditions or variables involved in the experiments with certainty. The correct definition of a plant tissue sample, experimental condition, developmental stage, and treatment depends on the proper use of ontologies, which can remove ambiguity and complexity when describing the data.

The Zadoks growth scale was developed by phytopathologist Jan C. Zadoks, classifying the stages of crop development on a scale from 10 to 92 (Zadoks et al., 1974) (see Figure 5.2.2). Using the Zadoks growth scale, researchers can remove ambiguity about crop developmental stages described in their research, which ultimately helps with their research's reproducibility.

Using the wheat transcriptomics examples as a continuation of Chapter 4's expVIP interactive figure, the expVIP authors included the comparison of datasets from different wheat transcriptomics papers. To

Growth Stage	Description of stage
	Seedling growth
GS10	First leaf through coleoptile
GS11	First leaf unfolded
GS13	3 leaves unfolded
GS15	5 leaves unfolded
GS19	9 or more leaves unfolded
	Tillering
GS20	Main shoot only
GS21	Main shoot and 1 tiller
GS23	Main shoot and 3 tillers
GS25	Main shoot and 5 tillers
GS29	Main shoot and 9 or more tillers
	Stem elongation
GS30	Ear at 1cm (pseudostem erect)
GS31	First node detectable
GS33	3rd node detectable
GS35	5th node detectable
GS37	Flag leaf just visible
GS39	Flag leaf blade all visible
	Booting
GS41	Flag leaf sheath extending
GS43	Flag leaf sheath just visibly swollen
GS45	Flag leaf sheath swollen
GS49	First awns visible

Growth Stage	Description of stage
	Ear emergence
GS51	First spikelet of ear just visible
GS55	Half of ear emerged
GS59	Ear completely emerged
	Flowering
GS61*	Start of flowering
GS65	Flowering half-way
GS69	Flowering complete
	Milk development
GS71	Grain watery ripe
GS73	Early milk
GS75	Medium milk
GS77	Late milk
	Dough development
GS83	Early dough
GS85	Soft dough
GS87	Hard dough
	Ripening
GS91	Grain hard (difficult to divide)
GS92	Grain hard (not dented by nail)

*in some spring crops GS61 may precede GS59

Figure 5.2.2: Barley growth stages and benchmarks illustration (using the Zadoks scale system). Source: The Agriculture and Horticulture Development Board (AHDB) website http://cereals.ahdb.org.uk/

curate and categorise the data to different plant development stages, plant tissue, variety, and plant stress factors (e.g. disease, drought, light exposure, etc.) proved problematic since many papers did not include standardised ontological terms to describe their data, metadata and methods. This meant that utilising and re-purposing those papers' data in the expVIP comparison platform could not be done reliably.

The importance of using ontology terms in enabling reproducibility in LS is not limited to research papers. Datasets described using

standardised ontology terms can be more easily navigated and shared, enabling a more effective comparison of experiments from different studies and within species (Hollmann et al., 2018).

5.2.4 Ontology Unique Identifiers

To uniquely refer to terms within ontologies, terms have been appointed an Identifier (ID). This ID comprises a unique, seven-digit, zero-padded integer prefixed by an abbreviation of the ontology name, e.g. "PO:" for Plant Ontology, with PO:0009046 referring to *flower*. If a term becomes obsolete, its identifier is removed and never reused. Similarly, when two classes are merged, one of the two IDs is appointed as an alternative identifier (alt_id) (Cooper and Jaiswal, 2016).

A PO ID therefore represents a Universal Resource Identifier (URI) that can be found online if those URIs are mapped to a similarly unique Uniform Resource Locator (URL), e.g. http:/purl.obolibrary.org/obo/PO_XXXXXX). These URLs can be found in ontology indexers, such as the Ontobee website to catalogue and search for ontology terms across many ontologies. Ontobee is a connected data server designed for ontologies, and it enables ontological data exchange, visualisation, querying integration, and analysis. Ontobee dynamically de-references and addresses the individual ontological term URIs to HTML pages, allowing users to browse through ontologies in a web browser.

5.2.5 Development of Plant Ontology

The Plant Ontology Consortium (POC) provides a standard and controlled vocabulary for describing plant anatomy terms,

developmental stages and traits. This is a collaborative process among various databases and projects, aiming to standardise the description of entities and display the hierarchical relationship between the terms (Avraham et al., 2008; Bard and Rhee, 2004; Jaiswal et al., 2005). The ontological relationships are not abstract but can be considered nodes and edges within a network of ontology terms. The relationship can be "directed" (i.e. a parent-child relationship, where the child cannot be the parent of its parent) or "undirected" (i.e. where, if A is next to B, then B is next to A) (Bard and Rhee, 2004).

With this network, data knowledge can be better incorporated into databases, and bio-ontologies can be searched and connected across different databases, as ontologies are machine-readable terminologies (Pérez and Benjamins, 1999). Biological data, which are complicated, can be organised in a hierarchical order, enabling humans and computers to understand them better, ask more meaningful questions, draw conclusions and identify where we lack knowledge. Ontologies do just that: they enunciate the associations between the terms. Ontologies are created to follow certain principles and rules. For example, terms referring to anatomy can be described in the following ways: *as part of*, *has cell type*, *has adhesion points*, *is a* (Bard and Rhee, 2004).

The ontology terms are represented by unique IDs (e.g., PO:0009089) and show their hierarchical relationship with other terms. The IDs connect the biological databases to ontologies and the various databases to each other (Bard and Rhee, 2004). It is important to note that ontologies are not the same as annotations, the latter being the description of data objects. In contrast, ontologies are rules that comprise a "description logic" (Bard and Rhee, 2004). All terms in the POC are manually curated. Some terms become obsolete, and when this is the case, they are labelled as "OBSOLETE". Sometimes, there might

not be a unanimous consensus in agreeing on the term relationships and the facts of the ontologies (Bard and Rhee, 2004); moreover, as knowledge is ever-evolving over time, while some objects can become obsolete, others may have missing links as the object data is unknown.

The PO terms are categorised in classes named according to the OBO Foundry naming conventions. As defined by the authors (Cooper and Jaiswal, 2016), "the PO name (or label in OWL format) is made by a unique noun rather than an adjective e.g., 'embryo' instead of 'embryonic'".

The Web Ontology Language (OWL) is the primary semantic Web language that depicts complicated information about things, groups of things, and relationships between things and was created by the OWL Working Group (World Wide Web Consortium, 2012). The OWL serves as a computational logic language to be processed by computer programs. OWL can be serialised in Resource Description Framework (RDF) and adds ontological capability to RDF. OWL allows for describing more about data models and provides more useful annotations for knowledge representation (larger vocabulary) than RDF.

RDF is a framework for depicting interconnected data from multiple sources on the Web. The semantic Web organises information using RDF (Ontotext, nd). RDF was initially designed (and is often utilised) for describing metadata about Web resources, such as Extensible Markup Language (XML) files (Decker et al., 2000). Other than XML format, RDF is based on other syntaxes such as Terse RDF Triple Language (Turtle), Javascript Object Notation for Linked Data and N-Triples (Ontotext, nd; TechTarget, 2022).

Plant Ontology Term Names and Synonyms

The PO terms have standardised names and synonyms, enabling multiple descriptions of a term to be condensed into a single concept, providing consistency and other benefits such as translations.

Synonyms are typically subdivided into four different categories: *(a) Narrow*: for species-specific names, e.g., "pod" is a narrow synonym of "fruit"; *(b) Broad*: when the synonym can be used for two or more PO classes, e.g., "plant fiber cell" (PO:0025407) has "fiber" and "fiber cell" as broad synonyms; *(c) Exact*: when the same plant class has more than one name e.g., "leaf-derived cultured plant cell" (PO:0000007) has an exact synonym "leaf-derived cultured cell". Moreover, translations into other languages are classed as exact synonyms, e.g., "plant embryo proper" (PO:0000001) has synonyms "embrióforo" (in Spanish) and "胚本体" in Japanese; *(d) Related*: when a word or phrase is used in literature interchangeably, but not strictly accurately used because the synonym has a narrowly different meaning than the primary name (Cooper and Jaiswal, 2016; OBO Foundry, 2012; Walls et al., 2012) (see Figure 5.2.3).

a

```
[Term]
id: PO:0009046
name: flower
namespace: plant_anatomy
alt_id: PO:0004541
def: "A determinate reproductive shoot system (PO:0025082) that has as part at least one
carpel (PO:0009030) or at least one stamen (PO:0009029) and does not contain any other
determinate shoot system (PO:0009006) as a part." [POC:curators]
comment: The characteristic reproductive structure of angiosperms. May have as part one
or more petals, sepals or tepals. May contain one or more pistillode (PO:0009078),
staminode (PO:0009077) or other aborted organs that don't show up in mature form.
subset: Angiosperm
subset: reference
subset: TraitNet
synonym: "basal flower" NARROW [CO_125:0000019]
synonym: "double flower" NARROW []
synonym: "flor (Spanish)" EXACT Spanish [POC:Maria_Alejandra_Gandolfo]
synonym: "floret" RELATED []
synonym: "hermaphrodite flower" NARROW [CO_125:0000028]
synonym: "monoclinous flower" NARROW []
synonym: "perfect flower" NARROW []
synonym: "花 (Japanese)" EXACT Japanese [NIG:Yukiko_Yamazaki]
xref: OBO_SF2_PO:160
xref: OBO_SF2_PO:259
is_a: PO:0025082 ! reproductive shoot system
relationship: develops_from PO:0000229 ! flower meristem
```

Figure 5.2.3: *Elements of an ontology term, for example, flower PO:0009046. The key elements of a PO term are shown for the PO term flower. The identifier (ID) is a unique, seven-digit integer prefixed by "PO". The alternate ID is assigned when two terms are merged. Term definitions are carefully written with the assistance of experts in botany and ontology design and may include references to other ontology terms. The "xref" indicates the related SourceForge tracker, which can be accessed as a link from the term page. The OBO-Edit (Day-Richter et al., 2007) flat file version of the PO term flower shows the key elements in a textual form.*
The image and caption were reproduced here verbatim with permission from Cooper and Jaiswal (2016)

5.3 Methods

For expVIP to be useful for researchers, the conditions and phenotypes recorded against the various RNA-seq experiments needed to be manually curated to ensure they were placed under the appropriate categories for the correct comparison of the RNA-seq results. The categories used in the expVIP application at the time I undertook this project were not all named using the correct ontological terms.

From my correspondence with the authors of expVIP, it became apparent that it would be beneficial to have a system where literature on

the same topic (i.e. genes, transcriptomics data, and information around wheat expression) could be easily gathered automatically. This would allow for a consistent understanding of the development conditions and the part of the plant the authors meant to relate to the expression of the genes of interest, as well as better prepare the data for incorporation into tools like expVIP. In this way, researchers who wanted to compare the results from each of the studies could do so more easily through the expVIP tool. Using the case studies below, I will demonstrate the lack of ontology annotation standards in wheat, rice, and barley papers and how this hinders reproducibility.

With the findings of the examination of the case studies, I will demonstrate the need for a tool that can provide semi- or fully-automated mappings of terms in literature and dataset metadata to standardised ontology terms, thus improving the ability of researchers to integrate biological datasets, reuse results, understand research better and reduce the burden of manual annotation of datasets.

5.3.1 Mapping Plant Ontology Terms in Plant Species: Observations of a Manual Process

In a preliminary evaluation of 16 wheat papers, (Barrero et al., 2015; Cantu et al., 2011, 2013; Choulet et al., 2014; Gillies et al., 2012; Kugler et al., 2013; Leach et al., 2014; Li et al., 2013, 2014; Liu et al., 2015; Oono et al., 2013; Pearce et al., 2015; Pfeifer et al., 2014; Yang et al., 2013, 2015; Zhang et al., 2014) included in the expVIP study (Borrill et al., 2016), the terminologies describing the parameter investigations (disease/stress, tissue and developmental stage) were examined to manually match them to their equivalent PO terms by searching the PO and CO databases. This manual process also involved reading the

paper's methodology and deciding whether the describing terms used were accurate enough to clearly describe what the authors had written in their papers. The findings of this investigation are available in an Excel file which can be accessed via this Dropbox link. I was provided with the Excel file by Dr Philippa Borrill, the principal researcher of the publication (Borrill et al., 2016), who had manually retrieved the case studies for expVIP following the method described in Section 5.1. I was responsible for finding the standardised PO and ID values, as well as a description that was the closest match to the description used by the authors of the studies, by manually searching the PO and CO Databases https://www.ebi.ac.uk/ols/ontologies/po and https://cropontology.org/ respectively.

5.3.2 Manual Data and Metadata Retrieval from Published Articles and Associated Database Entries

Problem Linking data and metadata and mapping ontological terms to the text within papers is a necessary but laborious and time-consuming task. The following section describes the manual steps I undertook to complete the aforementioned task.

Starting with a journal article in wheat transcriptomics, an initial step would be to retrieve the study accession number and subsequently search online for these terms. As an example, we can search for "accession number E-MTAB-1729". This web search returns the following URL: https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-1729 (see Figure 5.3.1 which is a screenshot from the aforementioned EBI search).

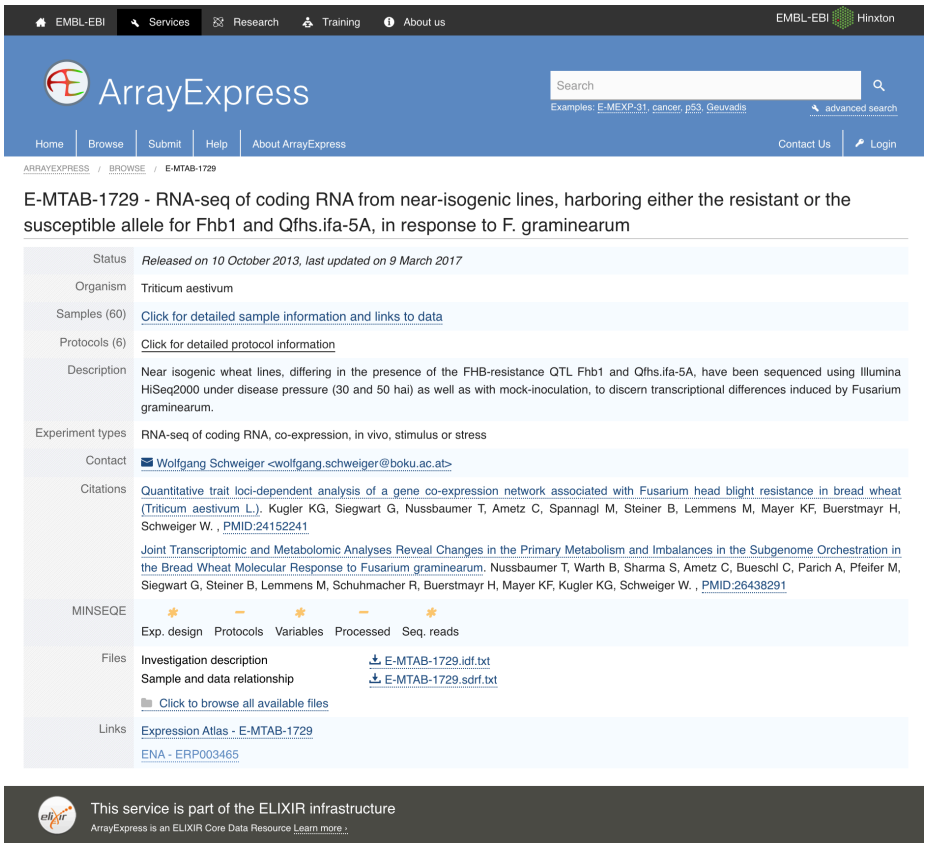

Figure 5.3.1: *Screenshot of the EBI website corresponding to accession number E-MTAB-1729.*

From the "Links" section of the ArrayExpress web result (see Figure 5.3.1), selecting the option for "ENA-ERP003465" redirects the user to the Web page with URL: https://www.ebi.ac.uk/ena/browser/view/PRJEB4202 that shows its associated project, with project accession number PRJEB4202.

As seen in Figure 5.3.2, at the time this search was performed, there were 60 samples with different accession numbers (e.g. SAMEA2151438), experiment accessions (e.g. ERX278684) and run accessions (e.g. ERR305274) associated with the project: PRJEB4202. From the right-hand side of the page (see Figure 5.3.2), we can see the publication linked to this project, which is available via this DOI link: DOI: 10.3390/ijms20235974.

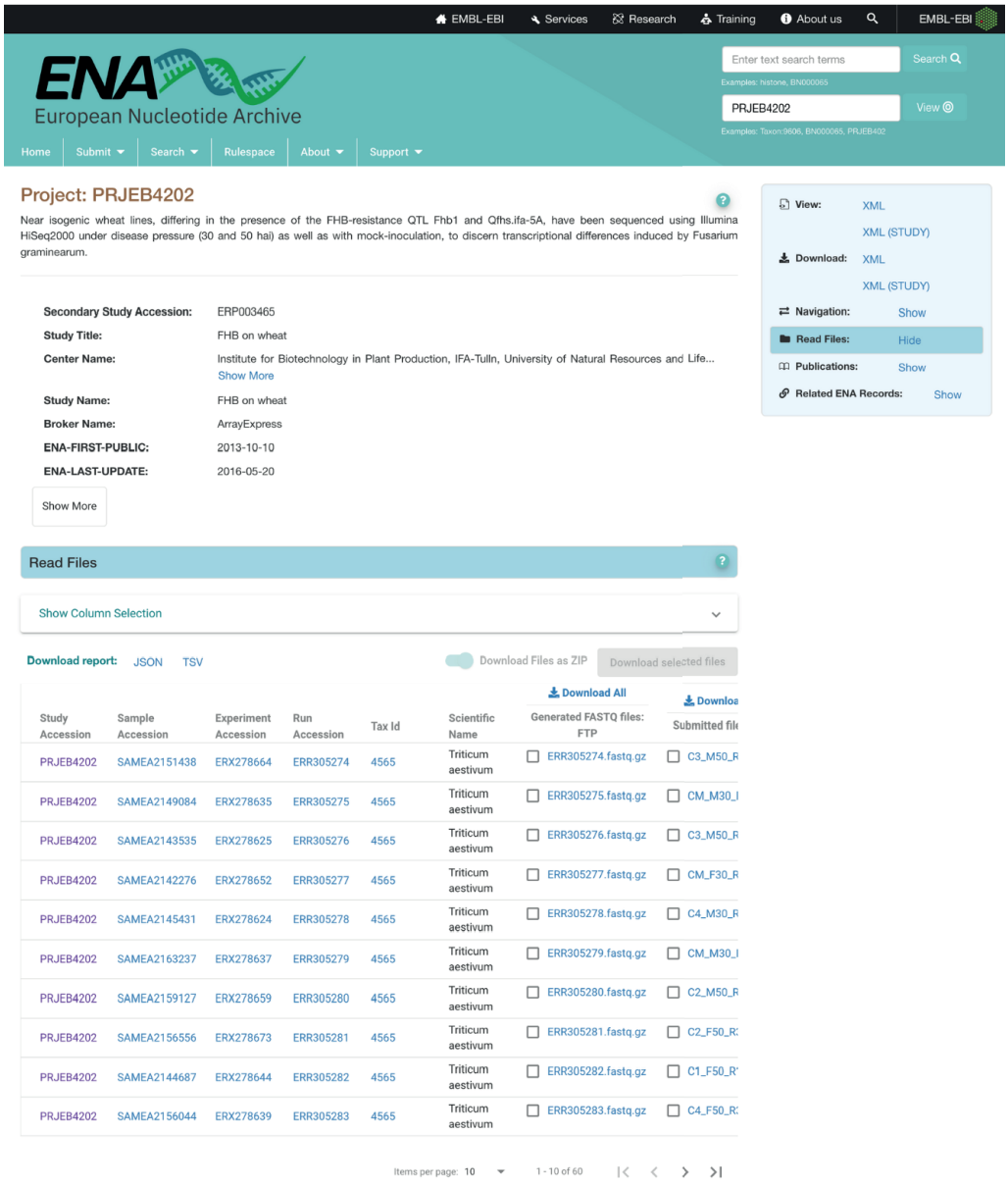

Figure 5.3.2: *Screenshot of the EBI website for the project number PRJEB4202, accessed by the URL: https://www.ebi.ac.uk/ena/browser/view/PRJEB4202.*

One can access the XML file from the EBI project page by specifying the display option in the URL: https://www.ebi.ac.uk/ena/data/view/ERP003465&display=xml. Sometimes XML files can include an <ABSTRACT> tag, which may or may not be an abstract from a published research article associated with the data. For example, see the following XML file at

https://www.ebi.ac.uk/ena/browser/api/xml/ERP003465, which is shown in Figure 5.3.3. In this case, searching the Web using abstract terms did not return a research article.

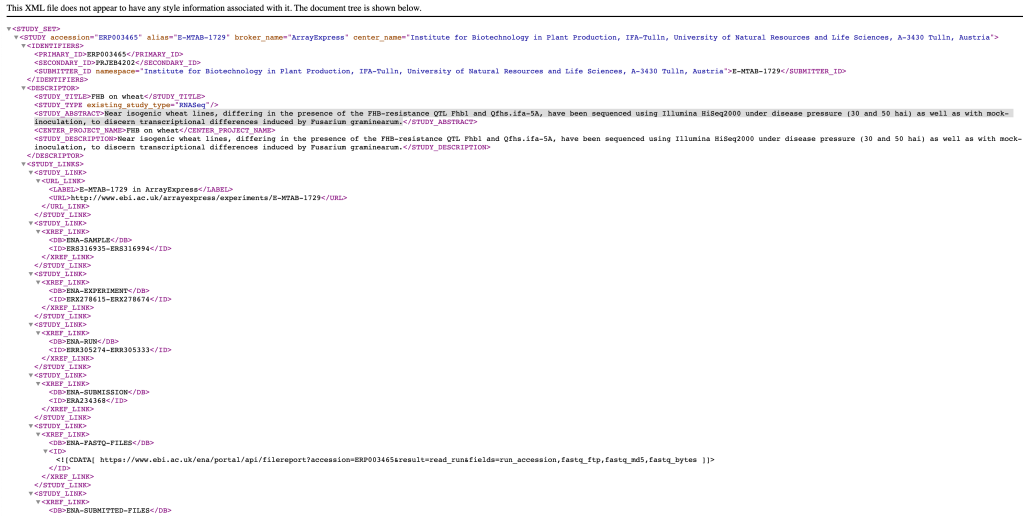

Figure 5.3.3: *Screenshot of the EBI website showing the XML file for ENA ArrayExpress study accession number ERP003465, with the abstract highlighted.*

5.3.3 *Deus ex machina*: Prototype Software for Ontology Mapping and Standardisation of Data and Metadata in Literature and Associated Database Artefacts and Computing a Reproducibility Metric Score

In the previous section (Section 5.3.2), I described the manual process required to retrieve the data and metadata from published articles and associated databases using study and project accession numbers, which can be found (but not always) within research papers. Linking data and metadata and mapping ontological terms to the paper's text is necessary to better understand, replicate and build upon the research conducted behind a publication, but it can be a laborious and time-consuming process.

The *Deus ex machina* software prototype I developed aims to solve the issue of mapping the arbitrary descriptions used in research papers to standardised ontology terms and IDs. Manually mapping these descriptions to standardised ontology terms and IDs is a time-consuming, laborious and often impossible task (where information about a particular study is simply not available) but is necessary for the development of bioinformatics tools such as expVIP.

With the methods undertaken in Section 5.3.1, the results of which are displayed in Sections 5.4.1, albeit a small number of case studies were reviewed, we can deduce that authors do not regularly use the conventional PO terms in their crop transcriptomics research publications. Therefore, it is typically not easy for the readers to understand the data (and metadata) presented in the paper without undertaking manual work. In many cases, whilst this manual curation is often more accurate due to the ability of humans to contextualise phrases into a common domain of knowledge, the accuracy relies on the expertise level of the curator.

For this reason, I created a software prototype named *Deus ex machina* that aims to annotate plant descriptions within an article with standardised POs and IDs. *Deus ex machina* also focuses on automatically retrieving the associated data of the paper by looking up specific keywords and phrases related to data availability. Moreover, it detects the database identifiers and ontology terms found within the full text of research articles. This would enable datasets to be marked with richer metadata than those typically deposited alongside the raw data in the public archives. The case studies used in the development and validation of the *Deus ex machina* prototype were in wheat transcriptomics.

***Deus ex machina* Software Design**

The following sections outline the *Deus ex machina* software design. Figure 5.3.4 shows the *Deus ex machina* flow diagram.

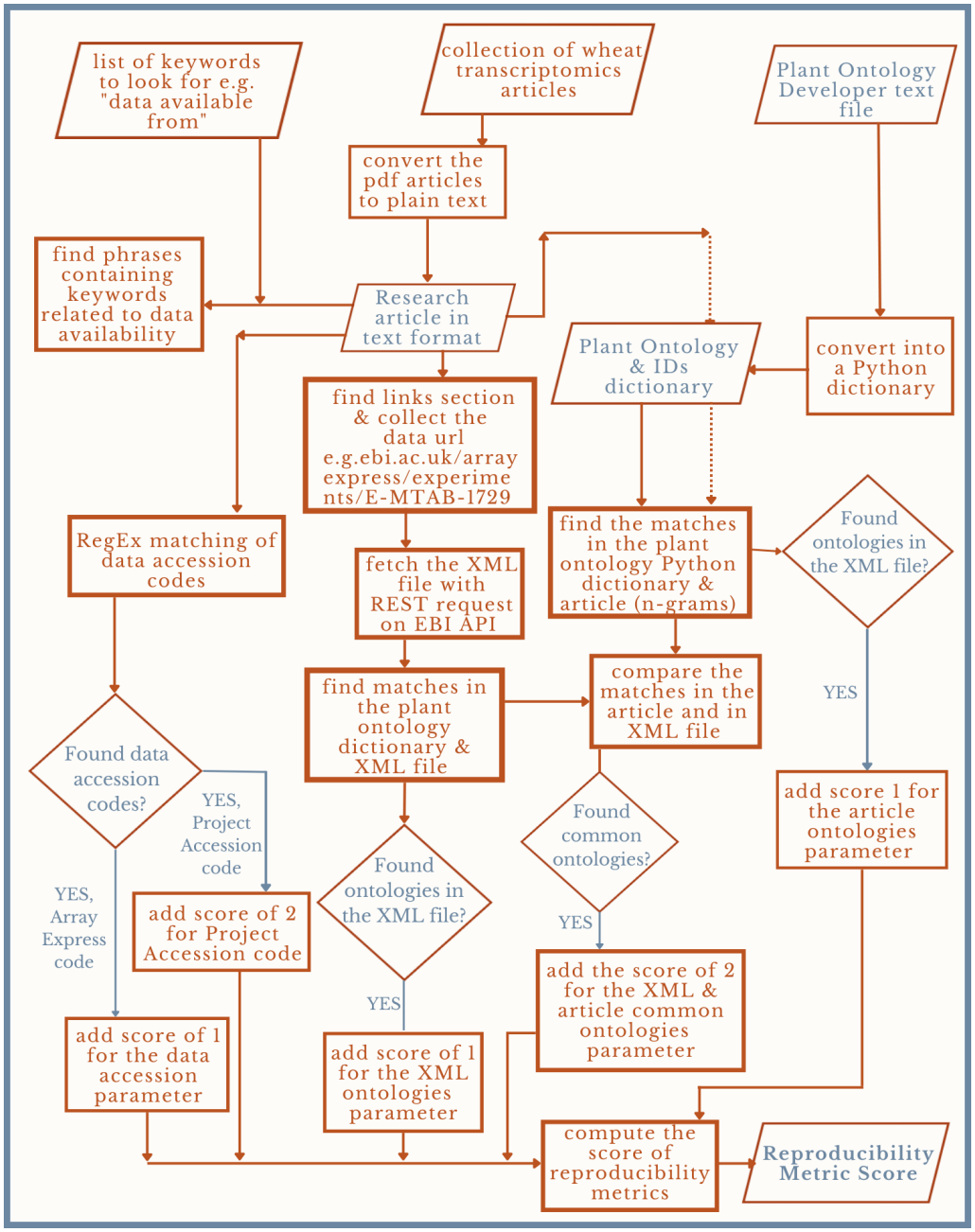

Figure 5.3.4: *Deus ex machina flow diagram.*

Deus ex machina is written in Python version 3.7. The software has been

developed to run on the command line, incorporating different libraries and pre-configured parameters. Of the libraries incorporated in the *Deus ex machina* pipeline, one is the Python pdftotext library, which converts a PDF file into the plain text that is necessary for parsing and extracting data and keywords.

Deus ex machina also incorporates (amongst other libraries) the Python tokenize library to "tokenize" (break strings into separate lexical elements, or "tokens") the text for the subsequent Natural Language Processing (NLP) that searches for PO terms and data access related keywords.

Please refer to the *Deus ex machina* code for the full list of libraries and processes executed by the code. The *Deus ex machina* is available via this GitHub repository.

The inputs for *Deus ex machina* are transcriptomics research papers in PDF format, and the PO description and the ID file (in text format) from the PO database ("plant-ontology-dev.txt", version 401d05a) downloaded from Github via this link: https://github.com/Planteome/plant-ontology/blob/master/plant-ontology.txt, (see Figure 5.3.5 which is a screenshot of the file). This "plant-ontology-dev.txt" file is also available within the GitHub repository of the *Deus ex machina*. The fields included in the "plant-ontology-dev.txt" file are the plant ID, name, definition, synonyms, is_a ID (to signify the parent node by which the entry is classed under), and is_a name (the name of the parent of the entry).

One example is *PO:0000001*, with the name *plant embryo proper*. Its definition is "*An embryonic plant structure (PO:0025099) that is the body of a developing plant embryo (PO:0009009) attached to the maternal tissue in a plant ovule (PO:0020003) by a suspensor*

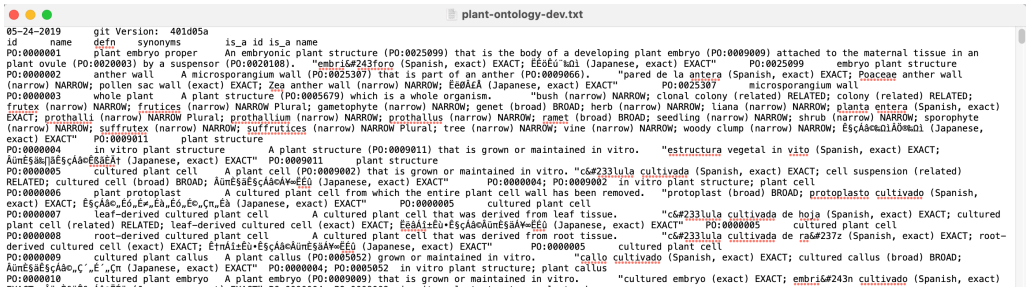

Figure 5.3.5: *Screenshot of the Plant ontology developer text file (version 401d05a), downloaded from Github via https://github.com/Planteome/plant-ontology/blob/master/plant-ontology.txt*

(PO:0020108)". Meaning that the parent of *plant embryo proper* is *embryonic plant structure*. Figure 5.3.6 shows a Tree View of the term plant embryo proper taken from the OBO library, as seen via this link.

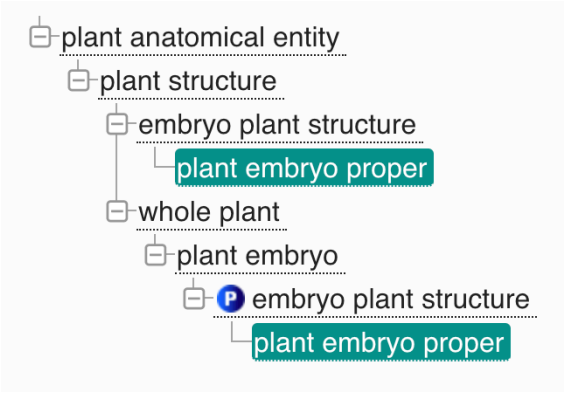

Figure 5.3.6: *Screenshot of the Tree View of an example plant ontology entity, the plant embryo proper (with ID PO:0000001). The screenshot was taken from the OBO library from this link.*

Pdf to Text and Tokenisation

Deus ex machina converts the PDF format of the article into plain text using the Python library pdftotext. Tokenisation is splitting text into individual tokens, whereby tokens can be words, stopwords (commonly used words such as "the", "a", "an", "in"), or special characters (such as _n - newline, _t- horizontal tab, _b- backspace, _"- double quote). Tokens next to each other in the tokenised text are called

N-grams, where N represents several contiguous elements.

For the next stage in the process, *Deus ex machina* converts the tokens into N-grams (N-number of consecutive words that appear in the text; unigrams, bigrams, trigrams, etc.). In other words, it establishes the "N-number of adjacent elements from a string of tokens". For instance, in the sentence "Hello world, have a great day!", the bigram (where N = 2), would be [("hello", "world"), ("world", ","), (",", "have"), ("have", "a"), ("a", "great"), ("great", "day"), ("day", "!")].

The tool tokenises the article text to a word level of 6-grams. The choice of stopping at 6-grams has been determined by the fact that most of the PO terms (according to the PO and ID developer file) comprise 6 or fewer words. The tokenisation of the article is necessary for parsing it in the subsequent pipeline steps. As *Deus ex machina* forms the N-grams, it searches for them against the keywords in a Python dictionary (as described in the next step).

Retrieving Ontology Terms from the Research Article and Associated Database Files

Deus ex machina constructs a Python dictionary (named `po_dict`) from the PO description and IDs text file "plant-ontology-dev.txt". The keys to the `po_dict` are the official name of a PO and the values are its corresponding ID. The code shown in Box 5.1 is a snippet of the code creating the `po_dict` from the "plant-ontology-dev.txt" file.

Deus ex machina searches the text file and finds matches in the `po_dict` dictionary and the article. Using the NLP properties means that strings of n-length, otherwise known as N-grams, can be searched in the text file. The code shown in Box 5.2 is a code snippet that locates ontology matches between the `po_dict` dictionary and the research article. The

function `find_ngrams` is constructed within the class File; see the *Deus ex machina* code for more context, via this GitHub link.

Listing 5.1: *Deus ex machina code snippet that constructs the po_dict Python dictionary with keywords, the names of a plant ontology, and value, its ID.*

```
po_dict = {}
for line in open('plant-ontology-dev.txt'):
    split = line.split("\t")
    if len(split) > 2:
        po_dict[split[1]] = split[0]
```

Listing 5.2: *Python code snippet that finds ngram matches in the research article up to 6-grams.*

```
def find_ngrams(self, data_tokenised, dict_po):
    ngram_matches = []
    n = 6
    for i in range(1, n + 1):
        n_grams = list(ngrams(data_tokenised, i))
        res = [' '.join(tups) for tups in n_grams]
        for query, onto_id in dict_po.items():
            if query in res:
                ngram_matches.append((query,onto_id))
    return ngram_matches


ngram_matches = file1.find_ngrams(textarticlecreation,po_dict)
```

Complexity arises when the authors have not used standardised ontology terms in their papers. At the moment *Deus ex machina* cannot provide "fuzzy" word matching, i.e. terms that have slightly different tokens

from a given input term. For example, the tool can match "plant embryo proper" if this term was explicitly used in the paper. Whereas, if the authors wrote something that is not a standardised ontology description, such as the "embryo of the plant", the software cannot currently match this phrase.

However, to compensate for such cases, the tool provides all phrases that might have at least one word from the standardised ontology descriptors. For example, in the case given above, the tool will return any phrases that contain the single words "embryo" or "plant" and both words "embryo" and "plant". This allows the software user to manually curate which sentences are significant for them and which to ignore. For example, a PO descriptor "plant cell wall" would be irrelevantly matched with a sentence that reads "the wall was painted blue". This is where the user's discretion comes in place to discard and ignore such phrases.

Given how many wheat transcriptomic papers do not use the correct ontology term (as per the "name" column of the PO "plant-ontology-dev.txt", see Figure 5.3.5), but instead use *synonyms* or the *description* (as per the column "defn" of the "plant-ontology-dev.txt", or something similar to the description) for explaining their work, I explored the approach of including the *description* and the *synonym* values in addition to the *name* to see if there were any matches found in the articles. To do so, I created a dictionary of lists, where the key is the "id" and the list of values are the "name", "synonym", and "defn" (i.e. the description of the ontology) of the plant ontology elements as per the "plant-ontology-dev.txt" file.

Therefore, when the code finds matches between the values in the lists and the research article, it can recognise metadata that the researchers

used. This approach can be useful since, despite not using the standardised name for each particular PO, the authors have at least used a synonym. This could usually be the case where common names are used as synonyms rather than official terminologies within the scientific community of specific disciplines because they can be more easily pronounced or remembered or are in popular use. This approach is much like how we use "aspirin" instead of 2-acetoxybenzoic acid, the correct pharmaceutical term.

However, in the current version of the *Deus ex machina*, the functionality to match terms from the synonym field of the "plant-ontology-dev.txt" is unavailable. It can be found in older commits of the code on GitHub. The decision to exclude it in the current code version was taken given the nature of the results retrieved from this approach. The decision to exclude the synonyms-matching functionality, is better explained in the Results of this chapter, see Section 5.4.2.

Keywords Matching and Data Retrieval from the EBI Database. Processing the ArrayExpress XML Output to get Information of Interest

Additionally, the *Deus ex machina* pipeline finds keywords in the articles from a keyword list I composed, having read various transcriptomics papers. From these keywords, the pipeline can then retrieve data accessibility links to fetch the data from an EBI search (see Figure 5.3.4). The list of keywords includes, amongst others, the phrases "accession", "repository", "ArrayExpress", and "Supporting", and they aim to find links to the article's associated data and metadata.

Moreover, the *Deus ex machina* tool uses Python's Regular Expression library (RegEx), where "a RegEx, or Regular Expression, is a sequence of characters that forms a search pattern. RegEx can be used to check

if a string contains the specified search pattern" (W3Schools, n.d.). EBI accession codes for experiments and array projects in ArrayExpress are given unique accession codes in the following format, where "XXXX" represents a four-letter code and n is a number (ArrayExpress, n.d.):

- `E-XXXX-n for experiments`

Using RegEx, the tool captures the EBI ArrayExpress accession number. According to the ArrayExpress help page the URL format needed for REST-style queries to retrieve results in XML format from the ArrayExpress is `https://www.ebi.ac.uk/arrayexpress/xml/v3/experiments/` followed by the ArrayExpress accession number. For example, the URL https://www.ebi.ac.uk/arrayexpress/xml/v3/experiments/E-MTAB-1729 retrieves the XML file associated with the ArrayExpress experiment code E-MTAB-1729. Using the EBI REST Application Programming Interface (API), the software can download XML data files associated with the ArrayExpress code found in the research article.

Finding Ontologies in the XML ArrayExpress file and Comparing them with those found in the Research Article

The code then parses and scrapes the ArrayExpress XML file to find ontologies and then compares the findings with the ontologies in the article (see Section *Retrieving ontology terms from the research article and associated database files*) to establish how many (if any) ontology terms are shared between the research article and the database XML file. In other words, the code finds the common ontologies between the research paper and the XML files and reports those.

The ultimate aim of this part of the code is to verify that the data is publicly

available at the appropriate databases or repositories and to then compare the ontologies found in the article with the ontologies contained in the ArrayExpress XML files of the study. By doing so, we hope to answer the following question: "How well does the metadata found in the research paper match with the metadata of the study records in the database?"

Additionally, the *Deus ex machina* tool performs a similar RegEx search for Project Accession codes of the format (PRJEB12345) and fetches the XML file corresponding to the Project Accession code, by following the URL https://www.ebi.ac.uk/ena/browser/api/xml/ followed by the Project Accession code. One such example is https://www.ebi.ac.uk/ena/browser/api/xml/PRJEB43230 which downloads the XML file saved as a text file for the user to examine. No further processing, in terms of finding ontologies, is done on the XML corresponding to Project Accession codes, but this is a feature that can be added in the future development of the tool.

Computation of a Reproducibility Metric Score

Moreover, the software computes various scores for a Reproducibility Metric Score (RMS). The code snippet (in Box 5.3) also shows how *Deus ex machina* assigns a score of 1 for the research paper if it contained standardised ontologies.

RMS comprises the following scoring components:

- Score of 1 for finding ontology terms within the journal article, irrespective of how many ontologies were found.
- Score of 1 if the journal article quoted an ArrayExpress Experiment accession code because it means the article denotes where the associated data for the study can be found.

- Score of 1 if the ArrayExpress XML file corresponding to the experiment accession code contained ontologies.
- Score of 2 if the journal article quoted a Project Accession code, with a larger score assigned here, as a study's Project Accession code can provide more information and metadata about the study than the experiment ArrayExpress accession code, hence why more weight for the computation of the RMS is given for a study including a Project accession code.
- Score of 2 if there were common ontologies within the journal article and the ArrayExpress XML file associated with the study.

Listing 5.3: *Python code that assigns a score of 1, if a researcher's paper contained any standardised ontologies*

```
ngram_matches = file1.find_ngrams(textarticlecreation, po_dict
    ↪ )
    ngram_matches_score = 0
    if ngram_matches is not None:
        ngram_matches_score = ngram_matches_score + 1
```

The *Deus ex machina* code has been tested and validated using 18 real-world research papers in wheat transcriptomics. The code creates a Common Separated Values (CSV) file, that includes the following fields: the article file name, the ontologies found in the article (if any), the ArrayExpress Experiment Accession code (if any), the Project accession code (if any), the ontologies found in the XML file corresponding to the ArrayExpress Experiment Accession code (if any), common ontologies between the research article and the XML file (if any), the score for finding research ontologies, the score for finding an ArrayExpress experiment accession code, the score for finding ontologies in the XML

file, the score for finding a Project Accession code, the score for finding common ontologies between the research article and the XML file and the total value of the scores which is the RMS.

Confusion Matrix: computing the accuracy of the *Deus ex machina* tool in automatically annotating PO terms

The accuracy of the *Deus ex machina* tool of correctly annotating the PO terms within wheat transcriptomic articles, was assessed by comparing the manually annotated results (assumed as the "actual truth") and the automatically annotated results, produced by *Deus ex machina*, with the help of a confusion matrix. Table 5.1 depicts an example of a typical confusion matrix. The focus of the confusion matrix computation, in this case, is to compare the presence of ontologies, not their absence.

Table 5.1: Example of a typical Confusion matrix. The abbreviations represent the following: True Positives (TP) represent the number of ontologies that are found in both manual and automatic annotation; False Positives (FP) represent the number of ontologies that are found only in the automatic annotation; False Negatives (FN) represent the number of ontologies that are found only in the manual annotation.

	Predicted positive	Predicted negative
Actual positive	True Positive (TP)	False Positive (FP)
Actual negative	False Negative (FN)	

Thus, in such a confusion matrix; where the confusion matrix is used to evaluate a classification algorithm (in this case *Deus ex machina*, only True Positives (TP), False Positives (FP), and False Negatives (FN) are included, as these are the cases where predictions are made about whether the algorithm is correct. These are the cases where the

algorithm makes a prediction, on whether a term is an ontology or not. To include True Negatives (TN), in a confusion matrix, requires a labelled dataset containing both positive and negative examples. This is typically the case with machine learning models. True Positives represent the number of ontologies that are found in both manual and automatic annotation, FP represent the number of ontologies that are found only in the automatic annotation and FN represent the number of ontologies that are found only in the manual annotation.

Using the formulae for *precision, recall and F1 score*, the accuracy of the *Deus ex machina* tool in automatically annotating PO terms can be calculated as follows:

```
Precision = True Positives / (True Positives + False
Positives)
Recall = True Positives / (True Positives + False
Negatives)
F1-score = 2 * (Precision * Recall) / (Precision + Recall)
```

Precision denotes out of all the positive predicted, what percentage that are truly positive. *Recall* defines out of all of the actual positives, what percentage are predicted positive. *F1-score* is the harmonic mean of *precision* and *recall*, indicating the model's ability to capture positive cases (*recall*) and be accurate with the cases it does capture (*precision*). *F1-score* is a robust gauge of model performance, as it takes into account the model's ability over two attributes, rather than just the absolute amount of correct prediction (Allwright, 2022).

The computations and analyses in creating the confusion matrix for assessing the accuracy of the *Deus ex machina* tool (using the 18 real-life wheat transcriptomics papers, are available via this FigShare link:

https://doi.org/10.6084/m9.figshare.22734056.v1.

The following points were taken into account for the computation of the confusion matrix and are listed below:

- The manual annotation of research articles found ontology terms from other ontology databases (i.e. not only from the PO database. However, as the *Deus ex machina* was coded to only find PO for an accurate comparison to take place, only the PO terms from the manual annotation were taken into account in the computation of the confusion matrix.
- Only PO terms of standardised names were taken into account and not the synonyms or definitions as per the "plant-ontology-dev.txt" (version 401d05a) file. The Excel file found on FigShare via this link, includes all the ontology terms found via the manual computation (which included synonyms), but for the accurate computation of the confusion matrix, given how *Deus ex machina* was coded to only retrieve ontologies from the standardised names (as mentioned above) were taken into account for the confusion matrix computation.
- There were a few terms found from the manual annotation whereby it was not possible to assign to a standardised ontology term with accuracy, because of the way they were noted in the articles. This is as mentioned earlier in the chapter, an issue where authors do not use standardised ontology terms and ID values next to the terms. One such example was the word *floret*, but there is a BRENDA Tissue Ontology (BTO) BTO:0000468 for it. There is no exact PO term for floret, but rather options are: "ear floret" PO:0006354; "spikelet floret" PO:0009082; and others. However, in this particular paper, the word *floret* is near the word *spikelet*.

Therefore, the word could be assumed as spikelet floret - PO:0009082, but again without certainty. These small cases of ambiguity are clearly noted in the Excel file found on FigShare, as noted above. These ambiguous terms were not many, and therefore placing them under TP or not was not going to significantly change the confusion matrix analysis results.

5.4 Results

5.4.1 Wheat Case Studies Manual Mapping of Ontology Terms

This section presents the results produced following the methods described in Section 5.3.1. The Excel file showing the results of manual mapping of ontology terms and their respective IDs is available on Figshare via this link: https://doi.org/10.6084/m9.figshare.20673207.v1. The fields that I completed in the Excel file were the fields denoted in the file with blue colour font, which corresponded to the columns named: "Ontology Term ID Tissue", "Comments Tissue", "Ontology Term ID Age", "Comments on the developmental stage", "Stress-disease ontology ID", "High-level tissue ontology ID", "comments high-level tissue", "High-level developmental stage (HLDS) ontology ID", "comments for HLDS ontologies", "high-level stress-disease ontology ID", and "comments for high-level stress-disease". I consulted OLS, Ontobee and sometimes Google searches in order to complete the fields. In general, authors often failed to describe their data and metadata with the appropriate PO terms and IDs and instead used non-standardised descriptions or synonyms.

I will use the term Days Post Anthesis (DPA) to illustrate the hurdles

encountered when trying to reproduce experiments in wheat genomic studies when non-standard terminologies are used. Some of the wheat transcriptomic papers I examined for manually mapping ontology terms and IDs included the term "days post anthesis", e.g. 10 days post anthesis. The term "anthesis stage" is the synonym (narrow) for "flowering stage" (PO:0007616). The description for the flowering stage term is "a flower development stage (PO:0007615) that begins when pollen (PO:0025281) is released by anther dehiscence (GO:0009901) and/or the stigma (PO:0009073) is receptive and ends with the process of pollination (GO:0009856) and/or floral organ senescence (GO:0080187)". However, there is no exact or synonym term for "days post anthesis", currently available under the Plant Ontology (searched via OLS or searching Ontobee, which is a server containing many ontologies).

When studies refer to a specific number of *days* post anthesis, it might not always be easy to reproduce the study accurately because after a certain number of days post anthesis, depending on the weather conditions or other environmental conditions, the plant might develop at different rates and so days post anthesis might not always be an accurate representation of a plant's developmental stage.

5.4.2 Deus ex machina Results

The *Deus ex machina* results are as expected given the current functionalities and features coded into the prototype. Manual validation of the results shows that the tool accurately finds the ontology terms from unigrams to 6-grams within the journal articles and the XML files (corresponding to the ArrayExpress experiment accession codes). There are some minor discrepancies that can be easily accounted for and

resolved with future development of the *Deus ex machina* code.

The computations of the code are tabulated for easy access for the user of the tool. The outputs of *Deus ex machina* on the 18 wheat transcriptomics use-case studies are available in an Excel file available via Figshare from this link: https://doi.org/10.6084/m9.figshare.20614728. The fields of the Excel file are as follows: article file name, the PO and their corresponding ID found in the article, the ArrayExpress experiment accession code found in the research article, the Project accession code found in the research article, the metadata found in the XML file corresponding to the ArrayExpress experiment accession code (the metadata found under the tags <value>, not just the possible ontologies that may be found in the XML file: see Appendix 5, C1), the ontologies found in the XML file, the common ontologies between the research article and the XML file, the score for finding ontologies (named as per the standardised term according to the "plant-ontology-dev.txt" file) thus scoring the article positively for annotating their study with the standardised terminologies, the score for the research article quoting ArrayExpress experiment accession code(s), the score for the research article quoting Project Accession code(s), the score for finding common ontologies between the research article and its associated XML file (corresponding to the ArrayExpress experiment accession code), and the RMS which is the total score of the scores mentioned above.

Figure 5.4.1 is a plot of the distribution of the RMS of the 18 articles assessed by the *Deus ex machina* tool.

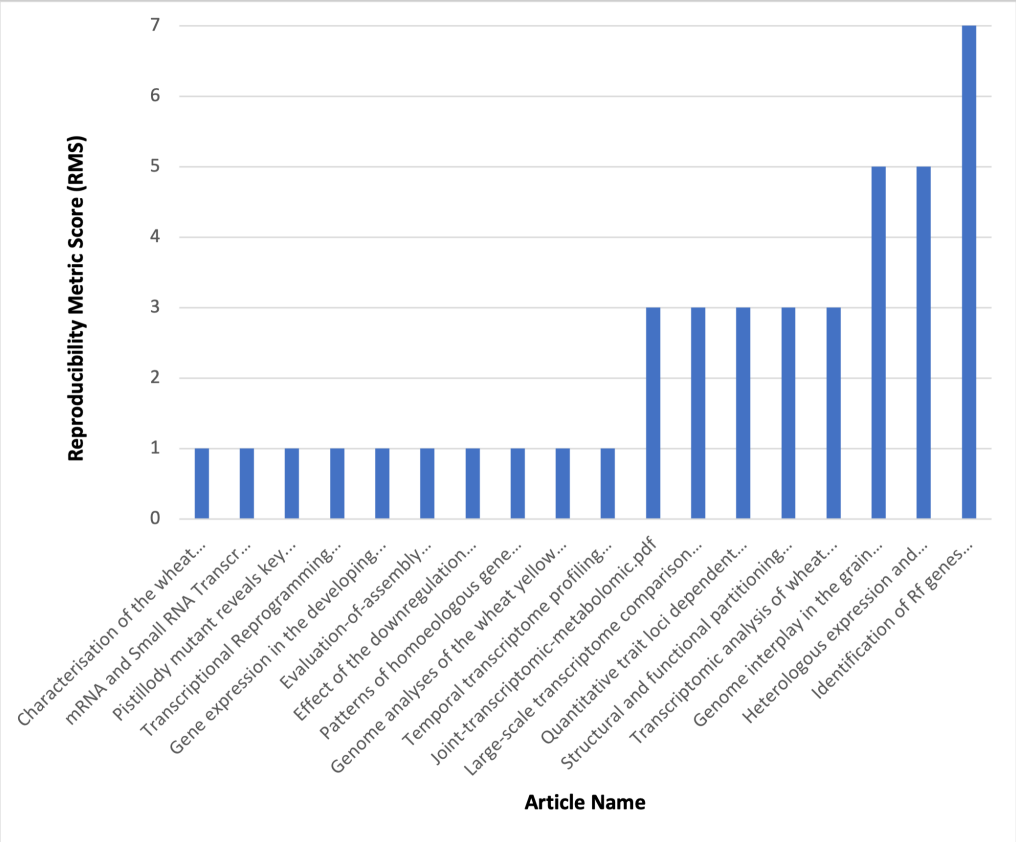

Figure 5.4.1: *Plot distribution of the RMS of the 18 articles assessed by the Deus ex machina tool.*

The majority of the case-study papers assessed scored low (RMS of 1) because they only included terms that were standardised PO terms. The articles did not include any of the other parameters measured and contributing to the RMS namely: ArrayExpress code, Project Accession number, XML metadata, ontologies found in the XML metadata file, common ontologies between the paper article and the XML metadata file).

Some papers achieved an RMS of 3 or 5 because they included PO standard terms, ArrayExpress code and ontologies in the XML metadata files and some had quoted Project accession codes within the article. The article that scored the maximum possible RMS value was the paper Tyrka et al. (2021). Figure 5.4.2 shows the RMS distribution of the case

studies (research articles) assessed by the *Deus ex machina* tool, whereby the RMS results are placed in "frequency bins" to better visualise the distribution of the results. As since in the figure most papers scored low, which means they lacked to include many of the reproducibility metric score parameters assessed by the *Deus ex machina* tool. In other words, those papers that scored lower RMS didn't present their research in a reproducible manner that would have allowed others to successfully reproduce their work.

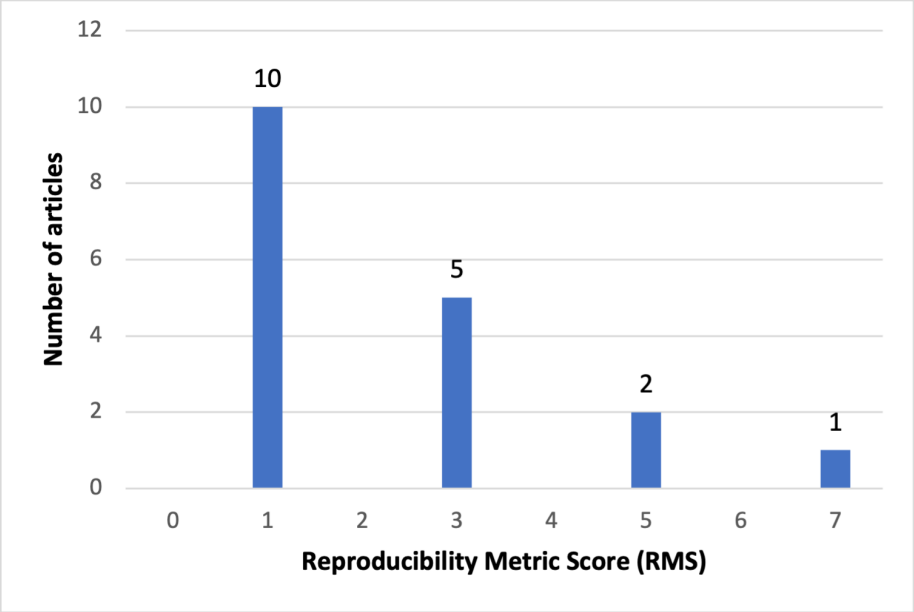

Figure 5.4.2: *Distribution of the RMS of the case studies (research articles) assessed by the Deus ex machina tool, whereby the RMS results are placed in frequency bins to better visualise the distribution of the results.*

Finding Project Accession codes

The *Deus ex machina* finds any Project Accession codes quoted in the research article and downloads the XML file saved as a text file for the user to examine with ease. This saves time for the user, who would have otherwise needed to perform a search via the ArrayExpress using the code and then manually download the XML file. This action is

performed accurately by the code whereby, in all the use cases, which were checked manually to validate the results, the Project Accession codes were detected every time.

Processing the ArrayExpress XML file

Processing of the ArrayExpress XML file fetched from the API-Rest request is conducted using the ArrayExpress experiment accession code. An example XML file corresponding to the EBI ArrayExpress experiment code E-MTAB-1729 is shown in Appendix 5, C1. Examining the XML files retrieved, the elements of interest concerning ontologies were found under the XML tag <value>. So from there capturing all the elements under the XML tag <value> returns a list such as this one:

```
['anthesis', 'CM-82036 resistant parent line', 'NIL1, Fhb1 and
Qfhs.ifa-5A resistance alleles', 'NIL2, Fhb1 resistance
allele', 'NIL3, Qfhs.ifa-5A resistance allele', 'NIL4, non
resistance allele', 'Triticum aestivum', 'spikelet floret',
'Institute for Biotechnology in Plant Production, IFA-Tulln,
University of Natural Resources and Life Sciences, A-3430
Tulln, Austria', 'CM-82036 resistant parent line', 'NIL1, Fhb1
and Qfhs.ifa-5A resistance alleles', 'NIL2, Fhb1 resistance
allele', 'NIL3, Qfhs.ifa-5A resistance allele', 'NIL4, non
resistance allele', 'Fusarium graminearum', 'mock', '30 hour',
'50 hour']
```

The tag <value> was chosen as, after investigating many ArrayExpress XML files, it was determined that ontology terms were included under the <value> tag. See Appendix B for the XML file fetched for the ArrayExpress Experiment accession number. I am unsure of the possible forms the researchers had to complete or the checklists or other fields required in the ArrayExpress database for the XML file to be produced,

but examining other XML files corresponding to ArrayExpress experiment accession codes, the ontologies (if any) were consistently under the <value> tag. Therefore, the assumption that the ontologies would be found under the <value> tag was made with more confidence.

Therefore in the ArrayExpress experiment accession code E-MTAB-1729 example, we can note that the terms `'anthesis'`, `'spikelet floret'` are terms of interest in the search for finding ontologies mentioned in the XML file. Anthesis is a commonly used term to describe the plant developmental stage (as explained above in Section 5.4.1). Spikelet floret is a standardised ontology term name with the ID PO:0009082. Comparing these results with the ontology terms fetched from the research article, we can then deduce the overlap of ontology term annotation between the research article and the XML file.

This is a critical hypothesis to investigate as this relationship is important in the reproducibility of research. The correct access, and availability of data and metadata in published research, both in articles and repositories or databases, are fundamental elements of reproducibility as explained in Chapter 3. As explained earlier, being able to use software to connect information between research articles and database entries means that we can access information that would otherwise be buried in databases or research articles. For the successful understanding, reproducibility, and re-use of research studies, it is vital to have a link of data between research papers and database entries.

Papers commonly use non-standardised ontological terms such as "anthesis". The XML metadata file in the EBI ArrayExpress file, fetched by an API REST request using the experiment accession code found in the paper, demonstrates they have a field for "developmental stage" as "anthesis" (see Appendix 5, C1). The correct "name" for PO ID

PO:0007616 is "lowering stage ", whereas the authors used in the XML metadata file the term "anthesis" (a synonym (narrow) as per the "plant-ontology-dev.txt" file).

The following text is one entry of the `po_dict` of the synonym list constructed by the code that is currently not included in the latest version of *Deus ex machina*, so as to demonstrate how the "synonym" and "description" fields could be queried to find ontologies within the research article or study database data and metadata files:

```
'PO:0030135': ['drepanium inflorescence', '"An inflorescence
(PO:0009049) with a sympodial growth habit, which at flowering
stage (PO:0007616), is composed of pedicellate flowers
(PO:0009046) on determinate inflorescence axes (PO:0020122),
each of which develops at progressively more acute angles from
each higher inflorescence axis (PO:0009081)."', 'drepanium
(exact) EXACT; monochasium (broad) BROAD; scorpioid (broad)
BROAD']
```

However, even if I were to include the functionality to match ontologies in the research papers and database-associated files, using as inputs the synonyms or descriptions of any ontology, which would entail more "fussy" matching principles, this would have still not completely solved the issue of lack of ontology annotation standards in research papers (and database files). To illustrate this better, we can use the same example of the descriptor "anthesis". In the "plant-ontology-dev.txt" file manually searching (using CTRL + F or CMD + F on the keyboard depending on the device used) for the word "anthesis" returns four possible entries, displayed below for more visual clarity:

1. ID: PO:0007016; term name: whole plant flowering stage; synonym: (07-anthesis in barley (related) RELATED; 07-anthesis

in oat (related) RELATED; 07-anthesis in wheat (related) RELATED [...])

2. ID: PO:0007024; term name: FL.04 end of flowering stage; synonym: (6.09 End of flowering in soybean (related) RELATED; 7.03-anthesis completed in barley (related) RELATED [...])

3. ID: PO:0007026; term name: FL.00 first flower(s) open stage; synonym: 6.00 1st flowers on main stem open in soybean (related) RELATED; 7.01-anthesis beginning in barley (related) RELATED [...]

4. ID: PO:0007053; term name: FL.02 1/2 of flowers open stage; synonym: 6.02 1/2 of flowers on main stem open in soybean (related) RELATED; 7.02-anthesis half-way in barley (related) RELATED [...]

The convention [...] is used here to denote that the synonym entry is not complete but truncated for the purpose of brevity.

From the above example, we can deduce that even if the *Deus ex machina* tool incorporated a synonym field, the tool's accuracy in matching a valid ontological term (in the research paper or database entries) would still not always be possible. This is because any synonym term could have multiple meanings; for example, the term anthesis discussed here can serve as a synonym for four different ontology terms. This issue of ontology ambiguity matching could be resolved should authors include the corresponding ontology *ID* for each term they use, even if it is not the standardised term name.

Referring again back to the "anthesis" example, and one of the use case wheat transcriptomic papers (Tyrka et al., 2021), this paper would have scored higher in the RMS (deeming their research more reproducible),

should the authors have incorporated the standardised ontology term and ID corresponding to the term they wanted to include. This would look like the following text: "*Furthermore, the mitochondrial gene orf279, which is important for the cms trait, is highly expressed in the early anthesis stage" (FL.00 first flower(s) open stage; PO:0007026).* This simple addition of the parenthesis with the corresponding term name and ID removes any ambiguity and allows humans to better understand the research and for tools to more efficiently extract information programmatically.

Therefore, in other words, even if I were to apply more language processing, and include the synonyms (in the current final version of *Deus ex machina*, as I performed some tests and given the more complex and ambiguous results) because anthesis appears in the descriptions and synonyms of different PO and IDs, it would still not give us a definite answer on what the authors meant by "anthesis". It is not a straightforward issue to tackle. It would still need some manual human interpretation. This is why it is ever more vital for researchers to be quoting the ID number in parenthesis, to remove ambiguity and ultimately allow for software to parse the information in both papers and database entries more efficiently and produce more accurate results.

5.4.3 Confusion Matrix: computing the accuracy of the *Deus ex machina* tool results

This section presents the results produced following the methods described in Section 5.3.3. The confusion matrix used to compute the precision, recall, and F1-score of *Deus ex machina* tool in automatically annotating the PO terms in the 18 real-life use-cases in wheat transcriptomic articles is shown in Table 5.2.

The computations and analyses in creating the Confusion matrix for assessing the accuracy of the *Deus ex machina* tool are available via this FigShare link: https://doi.org/10.6084/m9.figshare.22734056.v1.

Table 5.2: The confusion matrix for accessing the accuracy of the *Deus ex machina* tool in automatically annotating PO terms. The abbreviations denote the following: TP represent the number of ontologies that are found in both manual and automatic annotation; FP depict the number of ontologies that are found only in the automatic annotation; FN represent the number of ontologies that are found only in the manual annotation.

	Predicted positive	Predicted negative
Actual positive	TP = 112	FP = 6
Actual negative	FN = 12	

The *Deus ex machina precision* is 0.949 (TP=112, FP=6), which is a high precision value. The *Deus ex machina recall* is 0.903 (TP=112, FN=12). The *F1-score* of *Deus ex machina* is 0.926, which is a very good score according to the general conversions of the *F1-score* values (Allwright, 2022).

All of the above scores denote that *Deus ex machina* performs well with high accuracy in correctly annotating the PO terms in wheat transcriptomic papers.

5.4.4 Limitations

The pdftotext Python Library which reads the PDF file line by line and then concatenates it into a text file converts the two-column article PDF files in such a way that, where a sentence is split between the two columns, the text from two different columns is merged wrongly, creating some fragmented sentences.

This limitation, however, is not expected to significantly impact the ability of the code to find the ontology matches, and keywords. It mainly affects the code's functionality which finds the sentences which include the data availability-related keywords. This limitation, however, is not expected to impact the production of reliable results significantly negatively. Given that the *Deus ex machina* code also applies Python RegEx principles to find the ENA Project accession code(s), and ENA ArrayExpress experiment accession code(s), retrieving the study's associated data is not impaired. In other words, the *Deus ex machina* functions which employ Python RegEx principles can compensate for any sentences split and concatenated incorrectly by the pdftotext functionality of the code.

Deus ex machina cannot provide "fuzzy" word matching, i.e. terms that have slightly different tokens from a given input term. Input terms, in this case, are the PO terms. Nonetheless, more complex machine learning principles or NLP principles would still not eliminate ambiguity in certain cases (as explained above with the "anthesis" example), and especially in cases where the authors may have used completely arbitrary language, not related to the name, synonym or definition of a term.

Deus ex machina queries articles and ArrayExpress XML files with currently only the PO database. Some terms in the articles and database-associated files may be missed because they corresponded to, for example, the BRENDA tissue/enzyme source ontology. Incorporating more ontologies within the *Deus ex machina* can be addressed in future developments of the tool.

As noted in the Methods section of this chapter, the manual annotation if taken as the "actual truth" (gold standard), can be inaccurate, especially when undertaken by a single individual. Nevertheless, I have

paid close attention in performing the manual annotation of the 18 use-case studies in finding all of the ontology (or potential) ontology terms and I have clearly denoted them in the FigShare Excel file, as well as a descriptive explanation of the choice of the ontology.

5.5 Discussion

In recent years, technological developments in LSs mean a plethora of data and findings are produced rapidly. However, when these findings are not presented with the appropriate standardised ontological terminologies to contextualise metadata, understanding and interpreting that data (and the research around that data) becomes more complex. Thus, tools are needed to annotate the data in research articles and their associated metadata (in databases and repositories). Where published research or data in repositories and databases is not annotated properly with the correct ontologies, then understanding, re-purposing, and reproducing those study findings becomes an issue.

Apart from tools such as *Deus ex machina* which help automatically annotate the data and metadata from studies with the correct ontology terms, a more fundamental issue needs to be addressed. This issue would be not having the researchers' compliance to publish their study findings using the correct ontological terms. Could this lack of annotation of data and metadata be due to the lack of awareness or training on behalf of the researchers? Could it be because journal publishers or databases and repositories do not have the standardised ontological annotation of studies as a pre-requisite compulsory requirement?

Would pressure from journal publishers, databases, or repositories help

to create a culture or a shift in researchers' attitudes to annotate their studies' data and metadata with the correct ontological terms? Would such pressure help shift the current practices (which do not mandate that authors annotate their work with standardised ontological terms) and ultimately improve the reproducibility of those studies? Is it because not a lot of weight and importance is placed by all parties (researchers, journals, databases and repositories) to publish studies with reproducibility in mind? Could it be because most researchers (and journals) focus on the "Publish or Perish" paradigm and not on performing research that is robust and reproducible, or having their research conducted and published in such a fashion so that others can reproduce it?

All of these questions are fundamental and need to be addressed by the LS community as a whole, and certain mandates put in place so that the reproducibility of studies in LSs can be improved. In this respect, tools such as *Deus ex machina* could be useful to researchers, journals, and databases to help all parties automatically and quickly identify ontological terms and annotate them with the appropriate ontological ID. However, if such tools were made readily available and were at the disposal of researchers, journals, and databases, their ultimate success would depend on the motivation of all parties to use the tools.

Given how *Deus ex machina* is a proof-of-concept tool and does not incorporate complex machine learning principles, it cannot identify words or phrases of interest which could be describing ontological terms. This means that, apart from producing a long list of results, should words of interest be found (e.g. cell) and then left to the discretion of the user to decide whether those results are useful in finding candidate words or phrases to be then annotated with standardised ontological terms, the ultimate call is left to the user to ensure all ontology terms

have been written using the correct standardised ontological terms, including their ID. It is important to state, however, that there are cases where there could be no available standardised ontological terms for a specific word or phrase. One such example, as mentioned in the Results section, is the word "anthesis". In this case, the scientific community can suggest these terms to be included/incorporated into the appropriate databases. The POC encourages users to submit new ontology terms to be reposited and incorporated into their database.

Given how the accurate and effective retrieval and annotation of data and metadata from research are fundamental to reproducibility in the LSs, *Deus ex machina* could be a tool for promoting the reproducibility of research in LSs. For instance, *Deus ex machina* could be run before submitting papers to journals to ensure that manuscripts and their associated data and metadata deposited in public repositories included standardised PO terms and IDs. Furthermore, researchers who want to read and reuse already published data (both in literature and public repositories), similar to the needs of expVIP, can do so more confidently by knowing that the data and metadata associated with the research in question have been marked up with standardised terminologies, thus removing ambiguity in the concepts related to the research in question.

As explained in Section 5.2.3, using the correct ontological terms and IDs to describe research papers and their associated databases can help with the reproducibility of the studies in question and make them more useful both for humans and computers.

Finding the ontology descriptors in the papers and providing the standardised PO and ID can help readers to understand the research, the experiment conducted, the data and the metadata involved and, in turn, be able to use the data to further the research in the field.

HTML pages of research articles are not designed uniformly, so parsing them with algorithms such as those employed in *Deus ex machina* is not a straightforward process. For example, algorithms may need to parse only certain parts of interest in the HTML files, namely: the title, abstract, introduction, methods, results, conclusions, discussion, and data availability. However, not all publishers use the same terminology or structure to present their articles.

Due to this issue, the algorithm could not select sections with particular "div-id" or "section" titles only. These are tags used in HTML files to signify particular parts in an HTML file, and this solution was feasible for HTML files from a single publisher. The complexity comes when publishers use different HTML structures, where the tags or section names differ, which is commonplace. One would have to build separate HTML file scrapers for each journal article structure. As you can imagine, this is a complex and laborious problem. Arguably, if all journal articles had employed a uniform structure for their HTML files, then the job of parsing the HTML files and finding and extracting data from them with the help of machines would have been more straightforward and automated. Tools have been developed to do this, such as ContentMine from Peter Murray-Rust's group in Cambridge (https://github.com/ContentMine). However, as publishers frequently change their website structures and HTML representations, this becomes a game of cat-and-mouse where HTML scrapers and parsers must be continually updated.

Thus, this is an example of where standards in publications and their structure (e.g. HTML-tag uniformity, following the same tagging system) would serve the scientific community to more easily consume and re-use scientific artefacts programmatically. JATS is one such standard, but it is not in widespread use.

Considering these findings, the better practice would be that authors use the correct ontologies when describing their data in the first place and use the PO term in parenthesis. For example, journals could assist with this process by having a drop-down menu with pre-selected terms that authors could select from to describe their data. This way, journals could ensure that authors use standardised terminologies for their metadata, making it easier for other scientists to understand their work, reproduce it, and use it to further the science in the field and potentially increase citations.

Checklists, much as the term suggests, provide a list of fields that scientists are encouraged to use and fill in; some fields are mandatory for describing the correct and standardised terminologies for their data. Adhering to and using the provided checklists when depositing data at EBI makes it easier for researchers, readers of research papers, and researchers searching for the data in EBI databases, thus making the data and the research involved more reproducible.

Purpose and Benefit of the Reproducibility Metric Score of the *Deus ex machina* Software

The Reproducibility Metric Score (RMS) aims to review the availability of data in a standardised presentation format across research papers and their associated database files. Using a quantitative and qualitative approach, the scoring system allows us to evaluate what properties and elements research papers need to possess to be more reproducible and thus more valuable in the scientific community.

Ultimately, the computations of *Deus ex machina* can be further expanded to include more assessment parameters, which can be added with feedback from the crop transcriptomics community who can denote which specific parameters are important to the reproducibility of crop transcriptomics studies. It is important to note that although *Deus ex*

machina has been developed and validated using wheat transcriptomic studies, the concepts around it can be extrapolated and used by other study types.

These transferable concepts include the assessment of standardised ontology annotation in research articles and their associated data and metadata in related databases and the assessment of the availability of data (via links to the data or via the accession codes) which can ultimately enable others to reproduce the studies in question with more ease. Eventually, it is hoped that tools such as *Deus ex machina* may inspire the creation of a generalised means of computationally assessing the reproducibility status of a study utilising a reproducibility metric score. With such RMSs we can then motivate researchers to conduct and publish more reproducible research.

5.6 Further Work

Through the manual mapping efforts from wheat transcriptomics papers, certain terminologies have been identified that were not corresponding to any *PO* terms or other ontology vocabularies. Future efforts to help improve the semantic annotation standards in publications and associated database entries could involve creating a new ontology consortium and making them publicly available and retrievable through OLS and Ontobee, or suggest new terminologies and IDs to the appropriate Ontologies (see the Plant Disease Ontology available on GitHub via this link: https://github.com/Planteome/plant-disease-ontology).

The tool can become more versatile by providing "fuzzy" word matching, i.e. finding terms that have slightly different tokens from a

given input term. This development, however, is a more complex matter that entails more intricate machine learning, natural language processing, and linguistics concepts.

The *Deus ex machina* can be further developed to include assessment parameters that are more generalised and therefore applicable for more types of papers (and research disciplines) as a metric for their reproducibility status. At the present time, it is not possible for the tool to function with all types of papers, as some parameters in the *Deus ex machina* include finding "study accession numbers" and "project accession numbers". These are specific elements that help with the reproducibility of crop transcriptomics papers. In other words, the extension of *Deus ex machina* could include some further generic parameters which can help with assessing the reproducibility of research papers within other disciplines.

Further development of the *Deus ex machina* code can also include searching using Python's RegEx principles for other data accession codes, such as GenBank numbers, and ENA reference numbers for studies such as ERP123456 (as per the ENA guide (guidefromENA:https://ena-docs.readthedocs.io/en/latest/submit/general-guide/accessions.html), and ENA secondary study accession codes such as DRP000768.

However, because I am not involved in the particulars of research in crop transcriptomics I am not sure what is needed, in terms of which of the aforementioned accession codes, or the content of the associated XML files and the elements in the files (other than the ones I already identified), for the reproducibility of a study, beyond the requirement for ontological annotation, which represented the scope of my objectives and aims for the project described in this Chapter.

In other words, we can speculate that other elements, related to data availability, and correct metadata description can be taken into account or are necessary for the successful reproducibility of a study. However, the scope of the project described in this Chapter, and the applicability of the *Deus ex machina*, was centred around the standardised (or not) semantic annotation of research papers and their associated database entries. This is an important distinction to keep in mind when evaluating the applicability of the *Deus ex machina* tool.

Further programmatic features that can assess other reproducibility parameters of crop transcriptomics (and other types of studies) can be added to the tool, or other tools aiming to computationally evaluate the reproducibility score of studies in the effort of recognising, rewarding, and incentivising reproducible research practices and a more sustainable reproducible research culture within the LSs community.

The code can be further improved by allowing the detection of ontologies in the research paper or the associated database files in their plural form, for example, root vs roots. Also, the code can be furthered by including ontologies other than PO. From the manual matching of terms annotating the Excel file (available on Figshare via this link https://doi.org/10.6084/m9.figshare.20673207), consulting OLS and OntoBee platforms, certain terms could be matched with "BRENDA tissue/enzyme source" (BTO) and via https://www.brenda-enzymes.org which is a structured vocabulary for enzyme comprising tissues, cell lines, cell types and cell cultures.

6

Discussion and Conclusion

Research reproducibility has preoccupied the scientific community for many years, with numerous publications appearing on the subject for over three decades. Many articles have highlighted the issues with research reproducibility, ringing the alarm that more must be done to overcome the obstacles hindering research reproducibility. This thesis presented the research exploring technical and cultural solutions for research reproducibility.

This chapter first summarises the research contributions, and next, it remarks on how the findings of this thesis informs discussions in the field of research reproducibility in the LSs. Finally, this chapter discusses the limitations of the research undertaken for this thesis.

6.1 Contributions

The work of this thesis sought to explore how technical and cultural solutions can aid with the issues currently faced in research reproducibility in the LSs. This section provides a summary of the key research findings of this thesis.

6.1.1 Canvassing Researchers' Opinions and Attitudes around Reproducibility

We conducted a survey study that investigated the opinions of 251 researchers working in the LSs about their knowledge and attitudes about research reproducibility and technological solutions for reproducing research within research articles. The findings of this survey study were published in a paper of which I am the first author (Samota and Davey, 2021).

Our survey corroborated other literature results in that many researchers experience issues accessing data from published research, as well as regularly encountering difficulties with installing and running bioinformatic tools. Moreover, our survey highlighted how researchers believe that the most pivotal step towards improving research reproducibility would be for publications to describe methodologies (protocols, parameters, analyses) in detail.

The survey highlighted the confusion regarding definitions of reproducibility versus replication; as in the literature, the terms reproducibility and replicability are often confused or used interchangeably.

Moreover, through our survey result analyses, we were able to determine that training can affect the ability and intent of researchers to reproduce published research. Even though a small sample of respondents participated in this survey section (some skipped this section), it can still give us some insight into how training can affect motivation and the ability to reproduce published research. The responses received can then inspire us to ask further questions to better understand and establish the relationship between training and the ability and intent to reproduce

published research.

Moreover, the survey canvassed the opinions of life scientists around interactive figures, whether they believed they would be useful in reproducing computation experiments within research articles, what interactive figures they would deem favourable and whether the provision of interactive figures within research articles would influence their decision when choosing a journal to publish their research. To my knowledge, this was the first survey study to explore the views of LS researchers around interactive figures and reproducibility.

6.1.2 The Role of Interactive Figures in Supporting Reproducibility within Journal Articles

My PhD is an i-Case collaboration with *eLIFE*, and the initial aim was to investigate the interactive representation of computational, experimental results within research papers to facilitate their reproducibility.

The survey study we conducted to explore the views of respondents around interactive figures as a potential solution for enabling the reproducibility of computational experiments within research articles has given us insights for the development of my interactive figure prototypes as well as *eLIFE's* Reproducible Documents collaboration project with Stencila.

6.1.3 Semantic Annotation of Published Research and Associated Database Artefacts

Additionally, this thesis has explored the question of how we can improve how publications and databases are annotated with standardised ontological terms (semantic information) so that research

papers and their associated metadata (in databases and other repositories) can be presented reproducibly so that others can understand, re-use and build upon the results presented in the research in question.

This contribution aimed to explore the following research question: How can we improve how publications and databases are annotated with the standardised ontological terms (semantic information) so that research papers and their associated metadata (in databases and other repositories) can be presented reproducibly so that others can understand, re-use and build up the results presented in the research in question?

Following the manual mapping of crop transcriptomic studies and along with correspondence with scientists from the field, it was determined that there is an issue with the lack of standardisation with regards to semantic annotation of research depicted in publications and database artefacts. So it was established that there is a need for automatically annotating both research papers and databases, as well as establishing how well the metadata in the articles and the databases overlap.

The solution presented in this thesis is a software prototype, named *Deus ex machina*, which helps to automatically semantically annotate papers (using the PDF format) and their associated database files with the standardised ontology terms and IDs according to POC. The *Deus ex machina* prototype has been developed and tested on real-world use cases of wheat transcriptomics papers and has been demonstrated to work with relatively good accuracy (currently with direct word matches) for PO terms. The *Deus ex machina* can be used on other types of crop papers (or any papers in which one would expect to find PO terms).

Although the specific parameters which are evaluated by *Deus ex*

machina relate to crop transcriptomics; *Deus ex machina* is a tool that researchers can also use when writing research papers in crop transcriptomics to remind themselves to complete the elements that will make their research more reproducible. The *Deus ex machina* tool can be used by publishing journals at the editing stage to check whether researchers have included the standardised semantic annotations for their research papers and include them in the feedback and review process to prompt researchers to edit their manuscripts accordingly and include the standardised ontologies.

For terms that are not able to be captured by the *Deus ex machina*, because it would require more complex machine learning principles, or in the worse case, when the researchers have used some arbitrary terminologies, this editorial stage tool (in the publication cycle), can still serve as a reminder to the authors to ensure that they use standardised terminologies or, where they introduce a new term or description, to specifically denote that term in their manuscript to allow for readers to better understand and reproduce their research.

One option to support institutes is to provide them with an expanded version of the *Deus ex machina* tool to incorporate a user interface, for example with an easier (click button) installation. It is often the case with technologies of this genre that the tools are usually developed to be command line tools. However, the code can be expanded, adapted and adjusted accordingly, depending on what parameters the developer wants to score/assess.

Assuming the *Deus ex machina* is made publicly available through a Web Server, then maintenance and sustainability issues come into place, as staff and resources need to be (in theory perpetually) available for such a tool to be (perpetually) available. The BioLit tool (Fink et al.,

2008; Attwood et al., 2009) was designed to automatically semantically annotate papers with CO terms and connect the papers with their associated database metadata; it was one such example of a tool that was available through the internet. However, the BioLit tool is no longer available; the links from the papers (Fink et al., 2008; Attwood et al., 2009) are no longer functional (http://biolit.ucsd.edu/), and there was also no GitHub link to the tool's code. This obsolescence could be due to issues with maintenance, where the developers of the tool have moved on and are working in new positions, or issues with the lack of funding required to maintain a tool that is publicly available through a Web Server. Given this example, command line tools shared with GitHub links, such as the *Deus ex machina*, can be made available on a more long-term basis, provided GitHub does not archive or delete repositories after many years of inactivity.

Apart from the automatic semantic annotation of research papers and their associated database metadata files (particularly XML files), *Deus ex machina* computes an RMS which assesses the reproducibility status of crop transcriptomics papers by measuring certain parameters which reflect the reproducibility status of a study.

6.1.4 Reproducibility Metric Score for Recognising and Rewarding Reproducible Science

The computation of the RMS as a different means of assessing research reproducibility and as another publication metric can be used as a means of motivating researchers by giving recognition to those who perform and publish reproducible research.

The scientific community in the LSs has many tools and mandates seeking to improve research reproducibility, but we still have issues with

reproducibility. In this thesis, I proposed using a different metric to be applied to research articles to quantify their reproducibility status. The RMS would serve as a metric for assessing the quality of the research based on the reproducibility assessment parameters measured by *Deus ex machina*. We can identify and reward reproducible research using this RMS. If more journals used reproducibility badges with the implementation of RMS systems, then researchers would be motivated to conduct and publish more reproducible research.

6.2 Discussion

6.2.1 Providing all the Reproducibility Elements

To achieve research reproducibility, the research's associated data, metadata, correctly annotated data and metadata, and detailed methodologies (including well-documented code, analysis parameters, and explicit protocols) must be included in the proper knowledge "how-to." According to our survey findings, individuals who said they had prior informatics experience also stated that they are more capable of attempting and reproducing outcomes.

Rather than focusing on particular technologies, practical bioinformatics, and data management training is possibly a more effective method of reinforcing the idea that researchers' contributions to reproducibility are a duty that needs active preparation and execution. This may be particularly beneficial considering the training needs of wet lab and field scientists, who are increasingly accountable for bigger and more complicated computing datasets. It is necessary to complete further studies to better understand how researchers' skills in computational reproducibility may be related to their degree of training

in informatics.

Large and complicated datasets are becoming more common in science, particularly genomics. Storage solutions for Big Data files and referencing them inside the publishing document, particularly those in terabytes, may allow for their broader, more efficient, and appropriate data reusability (Faniel and Zimmerman, 2011; Poldrack and Gorgolewski, 2014).

However, despite the potential benefit these Big Data storage solution services may offer regarding data availability and accessibility, they do not automatically address the issue of data reusability on their own. Most noticeably, issues arise when data is too big to be kept locally or transmitted over internet connections in a reasonable time or when there is no way to attach metadata that defines datasets adequately for reuse or integration with other datasets. The issue of data repository lifespan also arises: who will provide funding for data repositories for decades into the future? Some researchers are now required to pay data egress fees in order to obtain data from cloud service providers (Banditwattanawong et al., 2014; Linthicum, nd).

Using this approach, data providers are supposedly saving money by not having to store big datasets publicly, but the expense of doing so has been transferred to the user to some extent. This presents complicated issues for large-scale data creation initiatives, which must be thoroughly investigated to have a long-term effect, particularly regarding reproducibility within publications.

Moreover, when data deposition to the appropriate repositories or databases is compromised, the issue of reproducibility is compounded further. Federer et al. (2015) investigated the differences between clinical and non-clinical scientists and uncovered that the majority of respondents had no prior experience in uploading biomedical data to a

repository, with the majority citing various social reasons for not doing so, including concerns and motivations about data sharing as well as the work required to prepare the data for submission. The failure to include data alongside publications, despite current policies and guidelines mandating data openness (National Institutes of Health, 2015; Wilkinson et al., 2016), can be attributed to a variety of factors, including technical difficulties, fear of being scooped, fear of mistakes being discovered in data or analyses, and fear of other researchers using the data for their research papers (Stodden, 2010; Tenopir et al., 2011; Federer et al., 2015; Tenopir et al., 2015).

Individual researchers and communities could benefit from the standardisation of FAIR data practices, which could be accomplished through public data deposition and subsequent publication and citation. This would encourage researchers and communities to share and reuse data while taking their individual requirements and needs into consideration (Pawlik et al., 2019). Data accessibility problems are further exacerbated by the fact that data becomes less retrievable with each passing year after the publication of the paper (Vines et al., 2014).

It is important to state that 100% reproducibility cannot always be achieved for certain research fields or published work conducted many years ago. For example, the Reproducibility Project attempted to reproduce high-impact research papers (Morrison, 2014). The team of researchers that were assigned to reproduce the studies was aware that certain factors would be a hindrance to achieving reproducibility. These factors were identified as the inability to find the original reagents; the fact that, at the time of publication of the papers in question, it was not the standard to describe methods and protocols in detail or publish raw data; and the materials not being commercially available (Morrison, 2014; Maher, 2015; Baker and Dolgin, 2017). The author of one paper

that was found to be irreproducible by the Reproducibility Project team noted that their study was reproduced successfully by at least 10 other labs internationally in the past and that the result of the Reproducibility Project was worrisome in that it could be a hindrance in enabling the drug to progress to the clinical study stage and receive further funding (Baker and Dolgin, 2017).

This is an example of where the inability to achieve irreproducibility does not always mean that the research is false or the findings are wrong. It could also mean that the study has been published lacking important reproducibility elements (such as links to data, links to code, and detailed methodology including all the parameters used to analyse the data). Unfortunately, published research can be missing these important reproducibility elements; consequently, good research can fail to achieve the impact and leverage it deserves. In contrast, research published in a reproducible way allows other researchers to successfully reproduce the research findings and build upon them, adding to our knowledge of the topics. More noteworthy results (including negative findings) can be achieved if more research is produced and published reproducibly, enabling other researchers to expand on the outcomes and deepen our understanding of particular research topics.

Although research irreproducibility has negative impacts in all LS domains, for some fields, it is more detrimental and costly. Such LS fields include those related to medicine and biomedicine or biochemistry. For example, as mentioned in the introduction of this thesis, the validity of over two decades of research on Alzheimer's disease and beta-amyloid plaque has been undermined since it was built upon initial studies that have recently been found to have falsified data (Piller, 2022). This example, alongside the misannotations of genetic associations for childhood diseases, demonstrates how important it is to

reproduce data before it is deposited in databases and before we can re-use it and build upon it to further research. However, this process is not always feasible or practical. When a new piece of research is produced and published, it should include the data associated with the research.

Yet, before further research is conducted based on data deposited in databases, that data ought to be deemed reproducible. Similarly to how data is manually curated in certain databases (such as the Swiss-Prot section of the UniProt KnowledgeBase), data that has been reproduced could be labelled with a badge to demonstrate that it is reproducible. In such a manner, researchers who want to build on or use the deposited data in the database can be confident that the data is valid, truthful and correct.

Such efforts, however, require funding, motivation and incentive for research teams (other than the original researchers who deposited that data in the database or repository) to reproduce the study and publish their results to inform the scientific community that they have successfully reproduced the results of others. However, this culture is not presently established, nor is it a phenomenon we generally encounter, as the current research environment mostly rewards "new findings". The Reproducibility Project was undertaken on high-impact cancer studies and cost $1.6 million (Morrison, 2014).

This begs the question, will there ever be a regular practice of making funding available to reproduce studies? It is not a question of assessing how reproducible studies are in hindsight, several years after their publication (such as with the Reproducibility Project (Morrison, 2014)), but rather whether there would ever be a decision-making body for each scientific discipline that would devote funding to various research

groups across the globe to be reproducing research results as and when needed, depending on what "original research" is published. Would research groups be motivated to devote the time and resources (even if funded) to reproduce the results of others instead of competing for funding to perform "original research"? This is what a more sustainable research reproducibility culture would entail, and such an outcome can be achieved if researchers are motivated by recognition and reward to perform studies to assess the reproducibility status of other research.

The costs of reproducing other studies, of course, vary depending on the studies themselves. Computational studies (i.e., not wet lab studies) require far fewer costs to be evaluated for reproducibility, with the main costs being the time and computational resources (if involving high throughput research). However, in the hypothetical scenario, where a group is assigned to reproduce the study findings of a different research group involving animals, or other live organisms, with *in vitro* research equipment, more funding would need to be assigned to perform such reproducibility studies. Assuming that the validation of the methodologies used in the assessed study does not require particularly specific and rare research skills and the investigating research team possesses the necessary laboratory skills, then, in theory, if the original researchers have conducted and published reproducible research, their results could be reproduced with a certain degree of success and accuracy (assuming some variation due to human factors).

6.2.2 Good Research Practices and Reproducibility Training and Institutions' Responsibility to Promote Reproducibility Culture

Is the root cause of the problem the difficulty of reproducibility per se, or is it more the lack of adequate training in how to design and conduct scientific studies and how to present them with reproducibility in mind?

There should be fundamental training in all undergraduate courses in LSs and beyond (in other sciences), providing guidance in correct and rigorous study design (with appropriate control groups); in statistics planning, especially with regards to designing research for statistical power, correct statistical planning, and how to choose the correct statistical tests for analysis; creating complete model systems; and avoiding conceptual flaws, self-bias and human prejudice in favour of our own ideas.

Training in statistics will enable researchers to understand how to design research with appropriate statistical power and choose the correct statistical test for analysing their data. It is commonplace for funders to require that grant applicants supply power calculations when proposing research involving living specimens or experimental assays. Training in computer and data science can allow researchers to write code and analyse data in a reproducible manner, as explained in Crick et al. (2017).

Many of the issues with research irreproducibility, would be rendered obsolete if the correct study design had been executed, and there had been more focus on the correct study design (Collins and Tabak, 2014) and more focus on conducting research reproducibly. I believe that more focus needs to be placed on these two pivotal factors affecting research:

1) training in how to design research (starting from undergraduate studies) and 2) training in how to conduct reproducible research.

In addition to basic training at the undergraduate level, there should also be continuous training throughout a scientist's career (including research institution training) on research reproducibility, computer science and bioinformatics. This claim has been backed up (on a small scale with a small cohort of respondents) by the findings of our research study (Samota and Davey, 2021), which demonstrated that researchers who had training in computer science were more willing and able to reproduce computational research.

Evidently, study design principles and rules will depend on the type of study and the domain. For example, a study in epidemiology needs to be designed differently than a study in computational biology. For this reason, it is not possible to divulge and list all the principles for correct study design for all domains in LSs. However, there are key papers describing the concepts of correct study design and setting a research question (Kuhn and McPartland, 1954).

A paper published in *Nature* (Begley et al., 2015), discussed the responsibility of research institutes in instilling reproducible research culture within their Institutes, naming it Good Institutional Practice (GIP). In a similar way to how many institutes include annual training on fire hazards and diversity and discrimination, reproducible research training should be a yearly requirement that researchers at institutions need to undertake.

It is possible that more institutions need to hire researchers whose responsibility is to be a "second set of eyes" for the research performed by the institute's researchers. Dr Catherine Winchester's job at the Cancer Research UK Beatson Institute is to go over the experimental

design and the data produced by the researchers in the institute to ensure it is valid and reproducible. The outcomes of her work as a research integrity adviser were that her institute produced research published in many high-impact and prestigious journals with no retractions. If funding is dependent on a certified compliance record with GIP, then robust research would get its due recognition (Winchester, 2018).

6.2.3 Measuring Reproducibility, Incentives and Rewards of Reproducible Research

The scientific community in the LSs has many tools and mandates seeking to improve research reproducibility, but we still have issues with reproducibility. In this thesis, I proposed using a different metric to be applied to research articles to quantify their reproducibility status. The RMS would serve as a metric for assessing the quality of the research based on the reproducibility assessment parameters measured by *Deus ex machina*. Using this RMS, we can identify and reward and incentivise reproducible research.

The need for incentivising and rewarding reproducible research has been demonstrated extensively over the past two-plus decades. More and more technologies and more and more data are produced, adding to the data and tool cabinet of the LSs. When researchers produce research (data and analysing tools) and move on to different positions, the technologies and the data can be forgotten and are not easily accessible.

By applying this RMS in research papers, researchers will be more motivated to conduct and publish reproducible research. It has been demonstrated that findings from research that has been validated to be reproducible can be thus deemed robust; we can trust the results are

valid, and we can confidently re-use and build upon the results of the reproducible research, and, importantly, avoid the issues mentioned throughout this thesis with reproducible research (Piller, 2022).

Should such reproducibility assessment metrics be used for promotion and securing more funding and ultimately for furthering the research career of scientists? If so, could the currency of research status be shifted from its present "publish or perish" status towards a more robust, rigorous and reproducible science? This shift would thereby create a need and motivation for researchers to receive more training and acquire more skills (if needed) to design, conduct and publish reproducible research.

With RMS and the real-world use case papers, I have quantitatively demonstrated the lack of semantic annotation of research papers in crop transcriptomics. This issue affects the ability to properly reproduce the papers, understand the research conveyed, build on them and use their findings to further the field.

Reproducibility needs to become normalised in academia and it is difficult to envisage this unless there is fundamental structural change: unless publishing impact factors incorporate reproducibility metric scores; unless reproducible research is taught from the undergraduate level onwards, with continuous mandatory institution training (similar to how equality and fire hazard training is conducted every year); and unless all journals abolish the word limitations on the methods section of articles so that authors can explain all their methodologies and protocols in detail, instead of saying "the protocol published in "XYZ" paper, modified as such [...] was followed".

Until reproducible research is regarded equally or more important than publication frequency, the hiring standards and the mandatory

institution training on reproducibility will not significantly improve. This does not imply that the entire burden of pulling the lever for achieving reproducible science is solely placed on the individual researcher. Instead, with the use of technologies such as interactive figures and *Deus ex machina*, certain aspects of the research reproducibility journey are automated, facilitating the researcher in their quest to perform and publish reproducible research.

We need to demonstrate to researchers that there is a need and a benefit to presenting research reproducibly: to show them "what is in it for them". People tend to first serve their own interests; it is human nature. They do not necessarily care if publishing journals intend to develop a new fancy way of presenting research interactively (but some people might care). This view has been made apparent with our survey research figures see Appendix A.3, with some of the free-text comments making it clear that we "don't need" interactive figures. Encouraging researchers to take up an interest in interactive figures (and publishing their research using interactive figures and reproducible documents) is about showing them how they will change perceptions of themselves and their research; they could be seen as more prestigious, more influential, and win more research grants, more funding, and more promotions.

Now, if we can quantify outcomes, by demonstrating how much per cent their paper readability increases and how the impact of their research is amplified, making comparisons between the papers from the same science domain, then we can convince researchers of the greater effectiveness of papers with interactive figures compared to those without interactive figures.

In other words, it feels like, at the moment, "conducting and presenting research reproducibly" is a "good to have" value/feature, but is not a

"must have" feature. This could be because the research currency is currently measured by the number of papers one researcher produces and how fast they can produce them. The producers of each new tool aiming to improve research reproducibility should spend time and resources to find ways to create the need and incentive for its use. For example, with interactive figures, if we prove that readability increases and the papers' influence increases, we can motivate the research community to create a more robust research reproducibility culture by incorporating interactive figures within publications as the new direction of research publication.

I believe the true need for research reproducibility will arise when experiments are reproduced before their data is deposited. This scenario sounds far away and costly, but we only have to look at the issues with misannotations of polymorphism of genetic diseases or the issue with amyloid studies. Fortunately, there is now technology that aims to detect images that have been modified (falsified). Steps are necessary because studies build upon the results of other studies, and any wrong data - or worse still, any fraudulent data - can significantly derail the course or the direction of a scientific field.

6.2.4 Journals and Reproducibility

It is very positive that more journals are focusing on research reproducibility and spending resources on projects promoting research reproducibility.

I still believe the most vital issues, the most fundamental issues, are not related to the presentation of the figures interactively but relate to the research that cannot be presented interactively as live figures/reproducible documents. In these cases, more rigorous tests

ought to be done to ensure the data submitted are in the correct repositories, and that it is annotated appropriately according to the standardised ontology terms.

What tests can be done to see if the code is reproducible? One option is that papers could have a Discussion section for comments that is open to the public and awarded a badge for reproducibility. However, this opens a potential can of worms, including the possibility for competitors and researchers with conflicting interests or competing projects to sabotage the research and leave fake reviews that the code does not run.

In this scenario, would journals have to employ a review monitoring team, much like those who monitor online marketplace platforms? Of course, this would mean additional costs for the journals whereby, unless the journal can find compensation from funding resources, increased publication costs may arise, which researchers and their institutions would then have to endure.

Collaborations, as opposed to the competitive science field, may enhance reproducibility and generate the opportunity for replication studies to corroborate findings. Imagine a world where all the research groups working in a specific field, e.g. wheat genomics, joined forces. Groups which were previously competitors decided to work together. This has, thankfully, been an increasing phenomenon. Institutions and universities from different countries join forces for specific research causes. For some, this might sound controversial; it might stir some waters, but imagine the depth and breadth of the quality results that could be produced.

With this collaboration model, there would be enough personnel and resources to allocate a sub-team in the project to ensure the research is reproducible; this team would have duties to supervise the research

design and validate the results.

But how realistic would such a research culture be in practice? Given the current state of research, where research groups have to compete for the same funding, then this might seem a naive way of approaching this problem. Nonetheless, there are initiatives, including various consortiums, which aim to bring together researchers and scientists from different countries and institutions.

6.2.5 Literature Challenging whether there is a Research Reproducibility Crisis, or Challenging the Scale and Extent to which it is Perceived by the Scientific Community

It will have been apparent throughout this thesis that I am a passionate advocate for promoting research reproducibility. However, I do not want to be perceived (or for this thesis to be perceived) as dogmatic. The paper by Fanelli (2018) argues that the state of the reproducibility crisis as depicted through the various literature is overly estimated; in particular, the number of claims of misconduct and falsification of data, especially from scientists from Western institutions, is exaggerated, with falsification being a rare event.

Importantly, the premise of the Fanelli (2018) paper is that it is counter-productive to be trying to motivate researchers to perform reproducible research using irreproducible science statistics and "science is in crisis" narratives. Whereas instead, a better approach to promote better science would be to invest and invite greater respect for research and inspire younger generations to do better science.

Others have also argued that the reproducibility crisis and the associated

literature wave discussing the topic could lead to mental health issues for scientists (Clements, 2020). Whereby failing to reproduce published results can cause stress because researchers might be a bad reflection on their research capabilities.

It is important to emphasise to the research community that inability to reproduce published results is not an indication of research incompetence. Initial studies might lack transparency, include biases and have weak methodologies.

6.2.6 Limitations

The interactive figures I created were prototypes and not completed solutions for reproducing computational experiments within research articles. However, they can be used at the front end of a research article. The interactive figure prototypes were developed, and the follow-up step would have been making them public via *eLIFE* to receive feedback from researchers who would have "played" with the figures and provided user feedback on them, including their opinions on whether the interactive figures could help reproduce computational experiments within journal articles; researchers could have also provided feedback for their preferred features for the interactive figures.

The interactive figure prototypes were developed to answer specific scientific questions, particularly in terms of assessing how difficult or easy it would be to have such infrastructures within journal articles for reproducing computational experiments. Moreover, the creation of the prototypes was intended to define what type of data (and thus which experiment types) would have been more popular within the scientific community, to establish targeted data types in specific LS research domains and receive feedback from the LS scientific community. The

prototypes would have needed to be further developed to be more useful: they were not designed or intended to be the framework by which all journals can incorporate interactive figures in their publications. As part of the original project schedule with *eLIFE*, the interactive figures would have gone on to be published through a public server for researchers to provide me with user feedback to address further research questions.

During a scheduled interruption of my PhD studies, *eLIFE* developed a collaboration with the Stencila team and moved the interactive figures project forward in a different direction. As a consequence, I had to change the objectives and aims of my PhD study which regretfully meant having to cease any further development of the interactive figures and the evaluation of their usefulness in improving research reproducibility.

Nonetheless, the work I have completed with the interactive figure prototypes has been productive and has helped to answer some questions about the ability of interactive figures to successfully reproduce computational experiments. Issues explored include which data types would have been useful to investigate for the initial interactive figure prototypes and the limitations and caveats of using interactive figures as a means of reproducing interactive figures within journal articles. The study has also helped to create an understanding that such solutions are difficult to implement, scale up, or be applied to all types of computational experiments, and an understanding of computational costs and maintenance costs.

The shift in the direction of my PhD provided a valuable opportunity to take a more holistic approach to my research. It enabled me to investigate other technological and cultural advances to overcome the hurdles to research reproducibility and to acquire a more overarching understanding of research reproducibility with wider aims and

objectives. As a result of this expansion of focus, the *Deus ex machina* tool was developed.

As described in Chapter 5, *Deus ex machina* is a command line tool, meaning that its usage is currently limited to the cohort of users or researchers for whom it is designed. The researchers/users would need to have some computational programming experience. The *Deus ex machina* tool has been developed to assess specific parameters in crop transcriptomics, and in its present form, it cannot be readily used on all types of papers until there is some code adjustment. Nevertheless, the work depicted in Chapter 5 can demonstrate how this prototype tool, as well as other tools, can be developed to either be specific to certain data types/research domains or to be more generalised to provide a more uniform, applicable comparison as a reproducibility metric for research papers across many disciplines in the LSs.

The parameters included in the RMS are not an exhaustive list. The intention of the code is, following implementation, to test with more use cases to be able to recognise other parameters of interest and include them in the RMS assessment.

6.3 Further Work

Further work concerning interactive figures and reproducible documents and their role in improving research reproducibility would require quantifying results in observational survey studies. We need to assess over the period of years how interactive figures and reproducible documents improve reproducibility in an empirical manner and prove that they can be sustainable and applicable to various types of datasets and research domains. We need to establish again, through

observational empirical studies, the opinions of researchers and research consumers of publications on whether they believe interactive figures and reproducible documents have assisted them in effectively reproducing published experiments. Lastly, we need to establish qualitatively and quantitatively how interactive figures and reproducible documents can affect the research and publishing cultures concerning reproducibility.

Further work about the research presented in Chapter 5 would be to run the tool for at least 100 wheat transcriptomics papers and quantify the reproducibility status of the papers based on the RMS to be computed. This would have given us a quantitative measurement of how reproducibly or not the corpus of the wheat transcriptomics domain presents their research. In line with the above, the *Deus ex machina* can be further developed to extract the date of publication of the research paper to help us assess the following research question: "Can we quantitatively demonstrate that the availability and accessibility of data through publications has increased following the introduction 2016 of FAIR Research Principles (Wilkinson et al., 2016)?" This will allow us to establish quantitatively whether data/code/research openness mandates from policymakers actually translate to the measurable execution of such mandates and, importantly, whether it translates to a more reproducible research culture. The above investigation can also be stratified as a comparison of how different journals compare given the RMS assessment. This way, more journals can emulate the publishing standards of the higher-scoring journals (and improve on the standards), which can help the overall research reproducibility state of the current publishing culture.

6.4 Reflections

The research presented in this thesis contributed to the notion that technological and social solutions should be used holistically to ensure that research is more reliable, robust, reproducible and reusable, and to reward and incentivise researchers for establishing a culture of reproducibility.

Our survey study (Samota and Davey, 2021), corroborated the findings of other studies in the field of reproducibility, in that many researchers experience issues accessing data from published research and that publications should describe methodologies in detail (Baker, 2016; Feger et al., 2019; Stodden, 2010). To my knowledge, this was the first survey to explore the opinions of life scientists around interactive figures, as a potential solution for reproducibility of CEs within journal articles and if the provision of interactive figures would influence their decision when choosing a journal to publish their research. Further insights were gained from developing interactive figure prototypes whereby I explored research questions, which helped the development of the ERA project (Maciocci et al., 2019).

With the manual annotation of ontologies of 18 real-life use case papers, which was a laborious and time-consuming task, I demonstrated how difficult it is to decipher with accuracy the work of others when authors do not use standardised ontology terms and IDs; all of which hinder the reproducibility of the studies (Fink et al., 2008; Shrestha et al., 2010).

The *Deus ex machina* software prototype, amongst other functions, automatically annotates research papers and their associated metadata with standard semantic data (PO terms and IDs). In order to evaluate the reproducibility status of research publications, *Deus ex machina*

calculates a RMS for an article. Thus, the software can be used to encourage and reward a more reproducible research culture, similar to other projects involving badges of reproducibility after manual evaluation of the reproducibility status of papers (Hong, 2021).

The *Deus ex machina* tool can be used by publishing journals at the review stage to check whether researchers have included the standardised semantic annotations for their research papers and associated metadata, as well as for researchers to test their own paper's reproducibility status.

There have been several projects attempting to tackle aspects of the reproducibility issue, with varying features and limitations, many of which have become obsolete when researchers move positions (Colomb and Brembs, 2014; Fink et al., 2008). This demonstrates the importance for technological solutions to be receiving funding to be maintained and promoted for better take up within the LS community; something that has been successfully done, with the Galaxy project (Afgan et al., 2018) as a WMS.

Importantly, for the current status of reproducibility in the LSs to change, there needs to be more emphasis and training in reproducible practices, also presented in our survey study (Samota and Davey, 2021), from the early stages of a scientist's career, as well as offer more incentives for career promotion and recognition for those conducting and sharing reproducible science.

In other words, technological and cultural advancements should be implemented synergistically for better and more long-term improvements in the LSs reproducibility.

6.5 Closing Notes

The ultimate goal of my research in the extensive field of research reproducibility is to demonstrate how synergistic use of technological advances and cultural changes can truly and sustainably shift the reproducibility issue.

I hope to have inspired the readers of this thesis, policymakers, researchers, and publishing journals to think of the reproducibility issue from a more holistic angle. Whereby instead of building more tools and technologies, and instead of creating further mandates for FAIR principles and assigning badges on publications to think of the researcher first.

To think of how the researcher conducts research and how the availability and implementation of technologies can facilitate reproducibility and offer incentives and rewards for receiving training and conducting reproducible research; but also to create a burning need for the researcher to use the tools and a true need for adopting reproducibility research practices and sustainable, reproducible research.

This action is required because, although many researchers acknowledge that there is an issue with research reproducibility and they often experience issues with reproducing the work of others; maybe this is not enough of a motive for researchers to feel the need to conduct and present research reproducibly, because it currently doesn't translate to career development.

Bibliography

R. J. Abdill and R. Blekhman. Meta-research: Tracking the popularity and outcomes of all biorxiv preprints. *eLife*, 8:e45133, 2019.

E. Afgan, D. Baker, N. Coraor, H. Goto, I. M. Paul, K. D. Makova, A. Nekrutenko, and J. Taylor. Harnessing cloud computing with galaxy cloud. *Nature biotechnology*, 29(11):972–974, 2011.

E. Afgan, D. Baker, B. Batut, M. Van Den Beek, D. Bouvier, M. Čech, J. Chilton, D. Clements, N. Coraor, B. A. Grüning, et al. The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic acids research*, 46(W1):W537–W544, 2018.

S. Allwright. What is a good f1 score? how do i interpret it? (2022), Dec 2022. URL https://stephenallwright.com/good-f1-score/. Accessed: 1-2-2023.

P. Amstutz, M. R. Crusoe, N. Tijanić, B. Chapman, J. Chilton, M. Heuer, A. Kartashov, D. Leehr, H. Ménager, M. Nedeljkovich, et al. Common workflow language, v1. 0. 2016.

J. Anon. Announcement: Reducing our irreproducibility. *Nature*, 496 (7446):398, 2013.

Ansible. How Ansible works, nd. URL https://www.ansible.com/overview/how-ansible-works?hsLang=en-us. Accessed: 02-08-2022.

ArrayExpress. Accession Codes, n.d. URL https://www.ebi.ac.uk/arrayexpress/help/accession_codes.html. Accessed: 11-07-2022.

M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, et al. The gene ontology consortium gene ontology: tool for the unification of biology. *Nat genet*, 25(1):25–29, 2000.

T. K. Attwood, D. B. Kell, P. McDermott, J. Marsh, S. R. Pettifer, and D. Thorne. Calling international rescue: knowledge lost in literature and data landslide! *Biochemical Journal*, 424(3):317–333, 2009.

S. Avraham, C.-W. Tung, K. Ilic, P. Jaiswal, E. A. Kellogg, S. McCouch, A. Pujar, L. Reiser, S. Y. Rhee, M. M. Sachs, et al. The plant ontology database: a community resource for plant structure and developmental stages controlled vocabulary and annotations. *Nucleic acids research*, 36(suppl_1):D449–D454, 2008.

M. Baker. 1,500 scientists lift the lid on reproducibility. *Nature News*, 533(7604):452, 2016.

M. Baker and E. Dolgin. Cancer reproducibility project releases first results. *Nature*, 541(7637), 2017.

T. Banditwattanawong, M. Masdisornchote, and P. Uthayopas. Economical and efficient big data sharing with i-cloud. In *2014 International Conference on Big Data and Smart Computing (BIGCOMP)*, pages 105–110. IEEE, 2014.

M. Baranyi and J. Greilhuber. Genome size in allium: in quest of reproducible data. *Annals of Botany*, 83(6):687–695, 1999.

J. B. Bard and S. Y. Rhee. Ontologies in biology: design, applications and future challenges. *Nature Reviews Genetics*, 5(3):213–222, 2004.

D. G. Barnes and C. J. Fluke. Incorporating interactive three-dimensional graphics in astronomy research papers. *New Astronomy*, 13(8):599–605, 2008.

D. G. Barnes, M. Vidiassov, B. Ruthensteiner, C. J. Fluke, M. R. Quayle, and C. R. McHenry. Embedding and publishing interactive, 3-dimensional, scientific figures in portable document format (pdf) files. *PloS one*, 8(9):e69446, 2013.

L. Barone, J. Williams, and D. Micklos. Unmet needs for analyzing biological big data: A survey of 704 nsf principal investigators. *PLoS computational biology*, 13(10):e1005755, 2017.

J. M. Barrero, C. Cavanagh, K. L. Verbyla, J. F. Tibbits, A. P. Verbyla, B. E. Huang, G. M. Rosewarne, S. Stephen, P. Wang, A. Whan, et al. Transcriptomic analysis of wheat near-isogenic lines identifies pm19-a1 and a2 as candidates for a major dormancy qtl. *Genome biology*, 16 (1):1–18, 2015.

S. Bechhofer, I. Buchan, D. De Roure, P. Missier, J. Ainsworth, J. Bhagat, P. Couch, D. Cruickshank, M. Delderfield, I. Dunlop, et al. Why linked data is not enough for scientists. *Future Generation Computer Systems*, 29(2):599–611, 2013.

C. G. Begley and L. M. Ellis. Raise standards for preclinical cancer research. *Nature*, 483(7391):531–533, 2012.

C. G. Begley, A. M. Buchan, and U. Dirnagl. Robust research: Institutions must do their part for reproducibility. *Nature*, 525(7567):25–27, 2015.

K. Belhajjame, J. Zhao, D. Garijo, M. Gamble, K. Hettne, R. Palma, E. Mina, O. Corcho, J. M. Gómez-Pérez, S. Bechhofer, et al. Using a suite of ontologies for preserving workflow-centric research objects. *Journal of Web Semantics*, 32:16–42, 2015.

C. Boettiger. An introduction to docker for reproducible research. *ACM SIGOPS Operating Systems Review*, 49(1):71–79, 2015.

P. Borrill, R. Ramirez-Gonzalez, and C. Uauy. expVIP: a customizable RNA-seq data analysis and visualization platform. *Plant physiology*, 170(4):2172–2186, 2016.

J. Brase. Datacite-a global registration agency for research data. In *2009 fourth international conference on cooperation and promotion of information resources in science and technology*, pages 257–261. IEEE, 2009.

N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter. Near-optimal probabilistic rna-seq quantification. *Nature biotechnology*, 34(5):525–527, 2016.

A. Brinckman, K. Chard, N. Gaffney, M. Hategan, M. B. Jones, K. Kowalik, S. Kulasekaran, B. Ludäscher, B. D. Mecum, J. Nabrzyski, et al. Computing environments for reproducibility: Capturing the "whole tale". *Future Generation Computer Systems*, 94:854–867, 2019.

P. Buneman, S. Khanna, and W.-C. Tan. Data provenance: Some basic issues. In *International Conference on Foundations of Software Technology and Theoretical Computer Science*, pages 87–93. Springer, 2000.

D. Cantu, S. P. Pearce, A. Distelfeld, M. W. Christiansen, C. Uauy, E. Akhunov, T. Fahima, and J. Dubcovsky. Effect of the down-regulation of the high grain protein content (gpc) genes on the wheat transcriptome during monocarpic senescence. *BMC genomics*, 12(1): 1–17, 2011.

D. Cantu, V. Segovia, D. MacLean, R. Bayles, X. Chen, S. Kamoun, J. Dubcovsky, D. G. Saunders, and C. Uauy. Genome analyses of

the wheat yellow (stripe) rust pathogen puccinia striiformis f. sp. triticireveal polymorphic and haustorial expressed secreted proteins as candidate effectors. *BMC genomics*, 14(1):1–18, 2013.

D. Catalano, D. Pignone, G. Sonnante, and M. M. Finetti-Sialer. In-silico and in-vivo analyses of est databases unveil conserved mirnas from carthamus tinctorius and cynara cardunculus. *BMC bioinformatics*, 13 (4):1–11, 2012.

R. Chamberlain and J. Schommer. Using docker to support reproducible research. *DOI: https://doi. org/10.6084/m9. figshare*, 1101910:44, 2014.

C. P. Chen and C.-Y. Zhang. Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information sciences*, 275:314–347, 2014.

F. Chirigati, R. Rampin, D. Shasha, and J. Freire. Reprozip: Computational reproducibility with ease. In *Proceedings of the 2016 international conference on management of data*, pages 2085–2088, 2016.

F. Choulet, A. Alberti, S. Theil, N. Glover, V. Barbe, J. Daron, L. Pingault, P. Sourdille, A. Couloux, E. Paux, et al. Structural and functional partitioning of bread wheat chromosome 3b. *Science*, 345(6194): 1249721, 2014.

G. Christensen and E. Miguel. Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature*, 56 (3):920–80, 2018.

J. F. Claerbout and M. Karrenbach. Electronic documents give reproducible research a new meaning. In *SEG technical program expanded abstracts 1992*, pages 601–604. Society of Exploration Geophysicists, 1992.

J. C. Clements. Is the reproducibility crisis fuelling poor mental health in science? *Nature*, 582(7811):300–301, 2020.

M. Cokol, F. Ozbay, and R. Rodriguez-Esteban. Retraction rates are on the rise. *EMBO reports*, 9(1):2–2, 2008.

H. M. Colhoun, P. M. McKeigue, and G. D. Smith. Problems of reporting genetic associations with complex outcomes. *The Lancet*, 361(9360): 865–872, 2003.

F. S. Collins and L. A. Tabak. Policy: Nih plans to enhance reproducibility. *Nature News*, 505(7485):612, 2014.

J. Colomb and B. Brembs. Sub-strains of drosophila canton-s differ markedly in their locomotor behavior. *F1000Research*, 3, 2014.

L. Cooper and P. Jaiswal. The plant ontology: a tool for plant genomics. In *Plant Bioinformatics*, pages 89–114. Springer, 2016.

L. Cooper, R. L. Walls, J. Elser, M. A. Gandolfo, D. W. Stevenson, B. Smith, J. Preece, B. Athreya, C. J. Mungall, S. Rensing, et al. The plant ontology as a tool for comparative plant anatomy and genomic analyses. *Plant and Cell Physiology*, 54(2):e1–e1, 2013.

T. Crick, B. A. Hall, S. Ishtiaq, and K. Takeda. "Share and Enjoy": Publishing useful and usable scientific models. In *2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing*, pages 957–961. IEEE, 2014.

T. Crick, B. Hall, and S. Ishtiaq. Reproducibility in research: Systems, infrastructure, culture. *Journal of Open Research Software*, 5(1), 2017.

J. Day-Richter, M. A. Harris, M. Haendel, G. O. O.-E. W. Group, and S. Lewis. Obo-edit—an ontology editor for biologists. *Bioinformatics*, 23(16):2198–2200, 2007.

D. De Roure, C. A. Goble, G. Klyne, M. Roos, K. M. Hettne, J. E. Ruiz, R. Palma, J. M. Gómez-Pérez, P. Missier, and K. Belhajjame. Towards the preservation of scientific workflows. In *iPRES*, 2011.

P. de Ruiter. The significance of reproducible data, June 2017. URL https://www.labfolder.com/the-significance-of-reproducible-data/. Accessed: 1-9-2019.

S. Decker, P. Mitra, and S. Melnik. Framework for the semantic web: an rdf tutorial. *IEEE Internet Computing*, 4(6):68–73, 2000.

Docker Docs. Best practices for writing Dockerfiles, nd. URL https://docs.docker.com/develop/develop-images/dockerfile_best-practices/. Accessed: 02-08-2022.

D. L. Donoho. An invitation to reproducible computational research. *Biostatistics*, 11(3):385–388, 2010.

C. Drummond. Replicability is not reproducibility: nor is it good science. Proceedings of the Evaluation Methods Learning Workshop at the 26th ICML, 2009. Available from: http://cogprints.org/7691/7/ICMLws09.pdf.

J. T. Dudley and A. J. Butte. In silico research in the era of cloud computing. *Nature biotechnology*, 28(11):1181–1185, 2010.

eLife Labs. Stencila ‑an office suite for reproducible research, 2018. URL https://elifesciences.org/labs/c496b8bb/stencila-an-office-suite-for-reproducible-research. Accessed: 11-02-2019.

D. Fanelli. How many scientists fabricate and falsify research? a systematic review and meta-analysis of survey data. *PloS one*, 4(5): e5738, 2009.

D. Fanelli. Is science really facing a reproducibility crisis, and do we need it to? *Proceedings of the National Academy of Sciences*, 115(11):2628–2631, 2018.

I. M. Faniel and A. Zimmerman. Beyond the data deluge: a research agenda for large-scale data sharing and reuse. *International Journal of Digital Curation*, 6(1):58–69, 2011.

B. Fecher, S. Friesike, and M. Hebing. What drives academic data sharing? *PloS one*, 10(2):e0118053, 2015.

L. M. Federer, Y.-L. Lu, D. J. Joubert, J. Welsh, and B. Brandys. Biomedical data sharing and reuse: Attitudes and practices of clinical and scientific research staff. *PloS one*, 10(6):e0129506, 2015.

S. S. Feger, S. Dallmeier-Tiessen, P. W. Woundefinedniak, and A. Schmidt. *Gamification in Science: A Study of Requirements in the Context of Reproducible Research*, page 1–14. Association for Computing Machinery, New York, NY, USA, 2019. ISBN 9781450359702. URL https://doi.org/10.1145/3290605.3300690.

A. S. Figueiredo. Data sharing: convert challenges into opportunities. *Frontiers in public health*, 5:327, 2017.

J. L. Fink, S. Kushch, P. R. Williams, and P. E. Bourne. Biolit: integrating biological literature with databases. *Nucleic acids research*, 36(suppl_2):W385–W389, 2008.

H. Fraser, T. Parker, S. Nakagawa, A. Barnett, and F. Fidler. Questionable research practices in ecology and evolution. *PloS one*, 13(7):e0200303, 2018.

M. Garcia-Hernandez, T. Berardini, G. Chen, D. Crist, A. Doyle, E. Huala, E. Knee, M. Lambrecht, N. Miller, L. A. Mueller, et al. Tair: a resource

for integrated arabidopsis data. *Functional & integrative genomics*, 2 (6):239–253, 2002.

H. Gardiner. Research ties diabetes drug to heart woes, 2010. URL https://www.nytimes.com/2010/02/20/health/policy/20avandia.html. Accessed: 05-08-2022.

H. Gardiner. False positives: fraud and misconduct are threatening scientific research, 2012. URL https://www.theguardian.com/science/2012/sep/13/scientific-research-fraud-bad-practice. Accessed: 05-08-2022.

A. Gehani, D. Tariq, B. Baig, and T. Malik. Policy-based integration of provenance metadata. In *2011 IEEE International Symposium on Policies for Distributed Systems and Networks*, pages 149–152. IEEE, 2011.

General Secretariat of the Council. Council conclusions on the transition towards an open science system, 2016.

R. Gentleman. Reproducible research: a bioinformatics case study. *Statistical applications in genetics and molecular biology*, 4(1), 2005.

S. S. Ghosh, J.-B. Poline, D. B. Keator, Y. O. Halchenko, A. G. Thomas, D. A. Kessler, and D. N. Kennedy. A very simple, re-executable neuroimaging publication. *F1000Research*, 6, 2017.

B. Giardine, C. Riemer, R. C. Hardison, R. Burhans, L. Elnitski, P. Shah, Y. Zhang, D. Blankenberg, I. Albert, J. Taylor, et al. Galaxy: a platform for interactive large-scale genome analysis. *Genome research*, 15(10): 1451–1455, 2005.

S. A. Gillies, A. Futardo, and R. J. Henry. Gene expression in the developing aleurone and starchy endosperm of wheat. *Plant biotechnology journal*, 10(6):668–679, 2012.

C. Goble, C. Wroe, R. Stevens, myGrid Consortium, et al. The mygrid project: services, architecture and demonstrator. In *Proceedings of the UK e-Science Programme All Hands Conference*, pages 595–603. Engineering and Physical Sciences Research Council, 2003.

C. A. Goble and D. C. De Roure. myexperiment: social networking for workflow-using e-scientists. In *Proceedings of the 2nd workshop on Workflows in support of large-scale science*, pages 1–2, 2007.

C. A. Goble, J. Bhagat, S. Aleksejevs, D. Cruickshank, D. Michaelides, D. Newman, M. Borkum, S. Bechhofer, M. Roos, P. Li, et al. myexperiment: a repository and social network for the sharing of bioinformatics workflows. *Nucleic acids research*, 38(suppl_2):W677–W682, 2010.

J. Goecks, A. Nekrutenko, and J. Taylor. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome biology*, 11(8):1–13, 2010.

S. A. Goff, M. Vaughn, S. McKay, E. Lyons, A. E. Stapleton, D. Gessler, N. Matasci, L. Wang, M. Hanlon, A. Lenards, et al. The iplant collaborative: cyberinfrastructure for plant biology. *Frontiers in plant science*, 2:34, 2011.

C. S. Greene, J. Tan, M. Ung, J. H. Moore, and C. Cheng. Big data bioinformatics. *Journal of cellular physiology*, 229(12):1896–1900, 2014.

T. Grossman, F. Chevalier, and R. H. Kazi. Bringing research articles to life with animated figures. *interactions*, 23(4):52–57, 2016.

K. F. Hollis. To share or not to share: ethical acquisition and use of medical data. *AMIA Summits on Translational Science Proceedings*, 2016:420, 2016.

S. Hollmann, T. K. Attwood, E. Bongcam-Rudloff, D. Duca, D. D' Elia, C. Endrullat, M. Frohme, K. Messerschmidt, and B. Regierer. Standardization and quality assurance in life-science research-crucially needed or unnecessary and annoying regulation? In *International Conference on Telecommunications*, pages 13–20. Springer, 2018.

N. P. C. Hong. Reproducibility badging and definitions: a recommended practice of the national information standards organization. 2021.

Y. Huang and R. Gottardo. Comparability and reproducibility of biomedical data. *Briefings in bioinformatics*, 14(4):391–401, 2013.

D. Hull, K. Wolstencroft, R. Stevens, C. Goble, M. R. Pocock, P. Li, and T. Oinn. Taverna: a tool for building and running workflows of services. *Nucleic acids research*, 34(suppl_2):W729–W732, 2006.

K. Ilic, E. A. Kellogg, P. Jaiswal, F. Zapata, P. F. Stevens, L. P. Vincent, S. Avraham, L. Reiser, A. Pujar, M. M. Sachs, et al. The plant structure ontology, a unified vocabulary of anatomy and morphology of a flowering plant. *Plant physiology*, 143(2):587–599, 2007.

T. Ingraham. Researchers can now publish interactive plotly figures in F1000, 2017. URL https://medium.com/plotly/researchers-can-now-publish-interactive-plotly-figures-in-f1000-87827a1b5d94.

J. P. Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8):e124, 2005.

J. P. Ioannidis, T. D. Stanley, and H. Doucouliagos. The power of bias in economics research, 2017.

S. A. Iqbal, J. D. Wallach, M. J. Khoury, S. D. Schully, and J. P. Ioannidis. Reproducible research practices and transparency across the biomedical literature. *PLoS biology*, 14(1):e1002333, 2016.

H. V. Jagadish and F. Olken. Database management for life sciences research. *SIGMOD Rec.*, 33(2):15–20, June 2004. ISSN 0163-5808. doi: 10.1145/1024694.1024697. URL https://doi.org/10.1145/1024694.1024697.

P. Jaiswal, D. Ware, J. Ni, K. Chang, W. Zhao, S. Schmidt, X. Pan, K. Clark, L. Teytelman, S. Cartinhour, et al. Gramene: development and integration of trait and gene ontologies for rice. *Comparative and functional genomics*, 3(2):132–136, 2002.

P. Jaiswal, S. Avraham, K. Ilic, E. A. Kellogg, S. McCouch, A. Pujar, L. Reiser, S. Y. Rhee, M. M. Sachs, M. Schaeffer, et al. Plant ontology (po): a controlled vocabulary of plant structures and growth stages. *Comparative and functional genomics*, 6(7-8):388–397, 2005.

J. M. James and R. Bolstein. The effect of monetary incentives and follow-up mailings on the response rate and response quality in mail surveys. *Public opinion quarterly*, 54(3):346–361, 1990.

D. Jobber, J. Saunders, and V.-W. Mitchell. Prepaid monetary incentive effects on mail survey response. *Journal of Business Research*, 57(1): 21–25, 2004.

Jupyter Project. Jupyter Project Documentation, nd. URL https://docs.jupyter.org/en/latest/. Accessed: 02-08-2022.

P. D. Karp. What we do not know about sequence analysis and sequence databases. *Bioinformatics (Oxford, England)*, 14(9):753–754, 1998.

Y.-M. Kim, J.-B. Poline, and G. Dumas. Experimenting with reproducibility: a case study of robustness in bioinformatics. *GigaScience*, 7(7):giy077, 2018.

R. Koschke. Software visualization in software maintenance, reverse engineering, and re-engineering: a research survey. *Journal of*

Software Maintenance and Evolution: Research and Practice, 15(2): 87–109, 2003.

K. Krampis, T. Booth, B. Chapman, B. Tiwari, M. Bicak, D. Field, and K. E. Nelson. Cloud biolinux: pre-configured and on-demand bioinformatics computing for the genomics community. *BMC bioinformatics*, 13(1):1–8, 2012.

K. G. Kugler, G. Siegwart, T. Nussbaumer, C. Ametz, M. Spannagl, B. Steiner, M. Lemmens, K. F. Mayer, H. Buerstmayr, and W. Schweiger. Quantitative trait loci-dependent analysis of a gene co-expression network associated with fusarium head blight resistance in bread wheat (triticum aestivuml.). *BMC genomics*, 14(1):1–15, 2013.

M. H. Kuhn and T. S. McPartland. An empirical investigation of self-attitudes. *American sociological review*, 19(1):68–76, 1954.

G. M. Kurtzer, V. Sochat, and M. W. Bauer. Singularity: Scientific containers for mobility of compute. *PloS one*, 12(5):e0177459, 2017.

L. J. Leach, E. J. Belfield, C. Jiang, C. Brown, A. Mithani, and N. P. Harberd. Patterns of homoeologous gene expression shown by rna sequencing in hexaploid bread wheat. *BMC genomics*, 15(1):1–19, 2014.

J. Leipzig, D. Nüst, C. T. Hoyt, S. Soiland-Reyes, K. Ram, and J. Greenberg. The role of metadata in reproducible computational research. *arXiv preprint arXiv:2006.08589*, 2020.

A. Li, D. Liu, J. Wu, X. Zhao, M. Hao, S. Geng, J. Yan, X. Jiang, L. Zhang, J. Wu, et al. mrna and small rna transcriptomes reveal insights into dynamic homoeolog regulation of allopolyploid heterosis in nascent hexaploid wheat. *The Plant Cell*, 26(5):1878–1900, 2014.

H.-Z. Li, X. Gao, X.-Y. Li, Q.-J. Chen, J. Dong, and W.-C. Zhao. Evaluation of assembly strategies using rna-seq data associated with

grain development of wheat (triticum aestivum l.). *PlOS one*, 8(12): e83530, 2013.

M. Liberman. Replicability vs. reproducibility—or is it the other way around. *The Language Log*, 2015. URL https://languagelog.ldc.upenn.edu/nll/?p=21956.

D. Linthicum. Don't get surprised by the cloud's data-egress fees, nd. URL https://www.infoworld.com/article/3266676/dont-get-surprised-by-the-clouds-data-egress-fees.html. Accessed: 06-09-2019.

Z. Liu, M. Xin, J. Qin, H. Peng, Z. Ni, Y. Yao, and Q. Sun. Temporal transcriptome profiling reveals expression partitioning of homeologous genes contributing to heat and drought acclimation in wheat (triticum aestivum l.). *BMC plant biology*, 15(1):1–20, 2015.

F. Löffler, V. Wesp, B. König-Ries, and F. Klan. Dataset search in biodiversity research: Do metadata in data repositories reflect scholarly information needs? *PloS one*, 16(3):e0246099, 2021.

D. Lowe. Faked crystallography, 2022. URL https://www.science.org/content/blog-post/faked-crystallography. Accessed: 05-08-2022.

G. Maciocci, M. Aufreiter, and N. Bentley. Introducing elife's first computationally reproducible article. *eLife Labs [Internet]*, 20, 2019.

B. Maher. Cancer reproducibility project scales back ambitions. *Nature*, 2, 2015.

V. Marx. The big challenges of big data. *Nature*, 498(7453):255–260, 2013.

W. K. Michener. Ten simple rules for creating a good data management plan. *PLoS computational biology*, 11(10):e1004525, 2015.

T. Miyakawa. No raw data, no science: another possible source of the reproducibility crisis, 2020.

J. C. Molloy. The open knowledge foundation: open data means better science. *PLoS Biol*, 9(12):e1001195, 2011.

S. J. Morrison. Reproducibility project: cancer biology: time to do something about reproducibility. *Elife*, 3:e03981, 2014.

H. Müller, F. Naumann, and J.-C. Freytag. Data quality in genome databases. 2003.

National Institutes of Health. Plan for increasing access to scientific publications and digital scientific data from NIH funded scientific research., 2015. https://grants.nih.gov/grants/nih-public-access-plan.pdf.

A. Newe. Enriching scientific publications with interactive 3d pdf: an integrated toolbox for creating ready-to-publish figures. *PeerJ Computer Science*, 2:e64, 2016.

B. A. Nosek, G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen, et al. Promoting an open research culture. *Science*, 348(6242):1422–1425, 2015.

OBO Foundry. Ontology Development Kit, 2012. URL https://ontology-development-kit.readthedocs.io/en/latest/Synonyms.html.

T. Oinn, M. Addis, J. Ferris, D. Marvin, M. Senger, M. Greenwood, T. Carver, K. Glover, M. R. Pocock, A. Wipat, et al. Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, 20(17):3045–3054, 2004.

Ontotext. What is RDF?, nd. URL https://www.ontotext.com/knowledgehub/fundamentals/what-is-rdf/. Accessed: 11-05-2022.

Y. Oono, F. Kobayashi, Y. Kawahara, T. Yazawa, H. Handa, T. Itoh, and T. Matsumoto. Characterisation of the wheat (triticum aestivum l.) transcriptome by de novo assembly for the discovery of phosphate starvation-responsive genes: gene expression in pi-stressed wheat. *BMC genomics*, 14(1):1–14, 2013.

B. C. O' Brien, I. B. Harris, T. J. Beckman, D. A. Reed, and D. A. Cook. Standards for reporting qualitative research: a synthesis of recommendations. *Academic Medicine*, 89(9):1245–1251, 2014.

M. Pawlik, T. Hütter, D. Kocher, W. Mann, and N. Augsten. A link is not enough–reproducibility of data. *Datenbank-Spektrum*, 19(2):107–115, 2019.

S. Pearce, A. K. Huttly, I. M. Prosser, Y.-d. Li, S. P. Vaughan, B. Gallova, A. Patil, J. A. Coghill, J. Dubcovsky, P. Hedden, et al. Heterologous expression and transcript analysis of gibberellin biosynthetic genes of grasses reveals novel functionality in the ga3ox family. *BMC plant biology*, 15(1):1–19, 2015.

R. D. Peng. Reproducible research in computational science. *Science*, 334 (6060):1226–1227, 2011.

R. D. Peng, F. Dominici, and S. L. Zeger. Reproducible epidemiologic research. *American journal of epidemiology*, 163(9):783–789, 2006.

E. Pennisi. How will big pictures emerge from a sea of biological data? *Science*, 309(5731):94–95, 2005.

A. G. Pérez and R. Benjamins. Overview of knowledge sharing and reuse components: Ontologies and problem-solving methods. In *IJCAI 1999*, 1999.

F. Pérez and B. E. Granger. Ipython: a system for interactive scientific computing. *Computing in science & engineering*, 9(3):21–29, 2007.

J. Perkel. Techblog: Interactive figures address data reproducibility, October 2017. URL http://blogs.nature.com/naturejobs/2017/10/20/techblog-interactive-figures-address-data-reproducibility/.

J. M. Perkel. Data visualization tools drive interactivity and reproducibility in online publishing. *Nature*, 554(7690):133–134, 2018.

S. Pettifer, J. Sinnott, and T. Attwood. Utopia—u ser-friendly t ools for op erating i nformatics a pplications. *Comparative and Functional Genomics*, 5(1):56–60, 2004.

S. Pettifer, D. Thorne, P. McDermott, J. Marsh, A. Villéger, D. B. Kell, and T. K. Attwood. Visualising biological data: a semantic approach to tool and database integration. In *BMC bioinformatics*, volume 10, pages 1–12. Springer, 2009.

M. Pfeifer, K. G. Kugler, S. R. Sandve, B. Zhan, H. Rudi, T. R. Hvidsten, I. W. G. S. Consortium, K. F. Mayer, and O.-A. Olsen. Genome interplay in the grain transcriptome of hexaploid bread wheat. *Science*, 345 (6194):1250091, 2014.

S. Philippi and J. Köhler. Addressing the problems with life-science databases for traditional uses and systems biology. *Nature Reviews Genetics*, 7(6):482–488, 2006.

S. R. Piccolo and M. B. Frampton. Tools and techniques for computational reproducibility. *GigaScience*, 5(1):s13742–016, 2016.

C. Piller. Blots on a field, 2022. URL https://www.science.org/content/article/potential-fabrication-research-images-threatens-key-theory-alzheimers-disease. Accessed: 05-08-2022.

H. Pimentel, N. L. Bray, S. Puente, P. Melsted, and L. Pachter. Differential analysis of rna-seq incorporating quantification uncertainty. *Nature methods*, 14(7):687–690, 2017.

M. A. Pitt and Y. Tang. What should be the data sharing policy of cognitive science? *Topics in Cognitive Science*, 5(1):214–221, 2013.

H. A. Piwowar, R. S. Day, and D. B. Fridsma. Sharing detailed research data is associated with increased citation rate. *PloS one*, 2(3):e308, 2007.

H. E. Plesser. Reproducibility vs. replicability: a brief history of a confused terminology. *Frontiers in neuroinformatics*, 11:76, 2018.

R. A. Poldrack and K. J. Gorgolewski. Making big data open: data sharing in neuroimaging. *Nature neuroscience*, 17(11):1510–1517, 2014.

A. Pujar, P. Jaiswal, E. A. Kellogg, K. Ilic, L. Vincent, S. Avraham, P. Stevens, F. Zapata, L. Reiser, S. Y. Rhee, et al. Whole-plant growth stage ontology for angiosperms and its application in plant biology. *Plant Physiology*, 142(2):414–428, 2006.

B. Pulverer. Reproducibility blues. *The EMBO Journal*, 34(22):2721, 2015.

B. Ragan-Kelley, C. Willing, F. Akici, D. Lippa, D. Niederhut, and M. Pacer. Binder 2.0-reproducible, interactive, sharable environments for science at scale. In *Proceedings of the 17th python in science conference*, pages 113–120. F. Akici, D. Lippa, D. Niederhut, and M. Pacer, eds., 2018.

S. S. Rao, S.-C. Huang, B. G. St Hilaire, J. M. Engreitz, E. M. Perez, K.-R. Kieffer-Kwon, A. L. Sanborn, S. E. Johnstone, G. D. Bascom, I. D. Bochkov, et al. Cohesin loss eliminates all loop domains. *Cell*, 171(2): 305–320, 2017.

J. T. Robinson, D. Turner, N. C. Durand, H. Thorvaldsdóttir, J. P. Mesirov, and E. L. Aiden. Juicebox. js provides a cloud-based visualization system for hi-c data. *Cell systems*, 6(2):256–258, 2018.

SaltStack. Introduction to salt, nd. URL https://docs.saltproject.io/en/latest/topics/index.html. Accessed: 02-08-2022.

E. Samota and R. P. Davey. Knowledge and attitudes amongst life scientists towards reproducibility within journal articles: a research survey. *Frontiers in Research Metrics and Analytics*, 6:35, 2021.

C. Santos, J. Blake, et al. Supplementary data need to be kept in public repositories. 2005.

M. Schneider, M. Flannery, and P. Griffin. Survey of bioinformatics and computational needs in australia 2016. pdf. *Figshare*, 10:m9, 2016.

M. Schwab, N. Karrenbach, and J. Claerbout. Making scientific computations reproducible. *Computing in Science & Engineering*, 2 (6):61–67, 2000.

M. R. Seringhaus and M. B. Gerstein. Publishing perishing? towards tomorrow's information architecture. *BMC bioinformatics*, 8(1):1–5, 2007.

C. Shettle and G. Mooney. Monetary incentives in us government surveys. *Journal of Official Statistics*, 15(2):231, 1999.

R. Shrestha, E. Arnaud, R. Mauleon, M. Senger, G. F. Davenport, D. Hancock, N. Morrison, R. Bruskiewich, and G. McLaren. Multifunctional crop trait ontology for breeders' data: field book, annotation, data discovery and semantic enrichment of the literature. *AoB plants*, 2010, 2010.

B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, et al. The obo

foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25(11):1251–1255, 2007.

J. Smith and R. Nair. *Virtual machines: versatile platforms for systems and processes*. Elsevier, 2005.

T. P. Sneddon, P. Li, and S. C. Edmunds. Gigadb: announcing the gigascience database. *GigaScience*, 1(1):2047–217X, 2012.

L. Snell and J. Spencer. Reviewers' perceptions of the peer review process for a medical education journal. *Medical education*, 39(1):90–97, 2005.

R. Stevens, A. Rector, and D. Hull. What is an ontology? *Ontogenesis*, 2010.

V. Stodden. The scientific method in practice: Reproducibility in the computational sciences. 2010.

V. Stodden. Reproducing statistical results. *Annual Review of Statistics and Its Application*, 2:1–19, 2015.

V. Stodden and S. Miguez. Best practices for computational science: Software infrastructure and environments for reproducible and extensible research. *Available at SSRN 2322276*, 2013.

V. Stodden, J. Borwein, and D. H. Bailey. Setting the default to reproducible. *computational science research. SIAM News*, 46(5):4–6, 2013a.

V. Stodden, P. Guo, and Z. Ma. Toward reproducible computational research: an empirical analysis of data and code policy adoption by journals. *PloS one*, 8(6):e67111, 2013b.

V. Stodden, M. McNutt, D. H. Bailey, E. Deelman, Y. Gil, B. Hanson, M. A. Heroux, J. P. Ioannidis, and M. Taufer. Enhancing reproducibility for computational methods. *Science*, 354(6317):1240–1241, 2016.

V. Stodden, J. Seiler, and Z. Ma. An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences*, 115(11):2584–2589, 2018.

A. Stupple, D. Singerman, and L. A. Celi. The reproducibility crisis in the age of digital medicine. *NPJ digital medicine*, 2(1):1–3, 2019.

A. Tait. 10 rules for creating reproducible results in data science, July 2017. URL https://dataconomy.com/2017/07/10-rules-results-data-science/.

B. Tang, F. Li, J. Li, W. Zhao, and Z. Zhang. Delta: a new web-based 3d genome visualization and analysis platform. *Bioinformatics*, 34(8): 1409–1410, 2018.

Technopedia. Virtual Provisioning, nd. URL https://www.techopedia.com/definition/16785/virtual-provisioning#:~:text=Explains%20Virtual%20Provisioning-,What%20Does%20Virtual%20Provisioning%20Mean%3F,with%20virtual%20machines%20(VM). Accessed: 02-08-2022.

TechTarget. Resource Description Framework (RDF), 2022. URL https://www.techtarget.com/searchapparchitecture/definition/Resource-Description-Framework-RDF. Accessed: 31-07-2022.

J. A. Teixeira da Silva. Negative results: negative perceptions limit their potential for increasing reproducibility. *Journal of negative results in biomedicine*, 14(1):1–4, 2015.

C. Tenopir, S. Allard, K. Douglass, A. U. Aydinoglu, L. Wu, E. Read, M. Manoff, and M. Frame. Data sharing by scientists: practices and perceptions. *PloS one*, 6(6):e21101, 2011.

C. Tenopir, E. D. Dalton, S. Allard, M. Frame, I. Pjesivac, B. Birch, D. Pollock, and K. Dorsett. Changes in data sharing and data reuse

practices and perceptions among scientists worldwide. *PloS one*, 10(8): e0134826, 2015.

L. Teytelman and A. Stoliartchouk. Protocols. io: Reducing the knowledge that perishes because we do not publish it. *Information Services & Use*, 35(1-2):109–115, 2015.

L. Teytelman, A. Stoliartchouk, L. Kindler, and B. L. Hurwitz. Protocols. io: virtual communities for protocol development and discussion. *PLoS Biology*, 14(8):e1002538, 2016.

A. E. Thessen and D. J. Patterson. Data issues in the life sciences. *ZooKeys*, (150):15, 2011.

M. Tyrka, B. Bakera, M. Szeliga, M. Święcicka, P. Krajewski, M. Mokrzycka, and M. Rakoczy-Trojanowska. Identification of rf genes in hexaploid wheat (triticum aestivum l.) by rna-seq and paralog analyses. *International journal of molecular sciences*, 22(17):9146, 2021.

Vagrant. Introduction to Vagrant, nda. URL https://www.vagrantup.com/intro. Accessed: 02-08-2022.

Vagrant. Salt Provisioner, ndb. URL https://www.vagrantup.com/docs/provisioning/salt. Accessed: 02-08-2022.

T. Vase. Advantages of docker. 2015.

P. L. D. Vincent, E. H. Coe Jr, and M. L. Polacco. Zea mays ontology–a database of international terms. *Trends in Plant Science*, 8(11):517–520, 2003.

T. H. Vines, A. Y. Albert, R. L. Andrew, F. Débarre, D. G. Bock, M. T. Franklin, K. J. Gilbert, J.-S. Moore, S. Renaut, and D. J. Rennison. The availability of research data declines rapidly with article age. *Current biology*, 24(1):94–97, 2014.

W3Schools. Python RegEx, n.d. URL https://www.w3schools.com/python/python_regex.asp. Accessed: 11-07-2022.

A. J. Wagers, R. I. Sherwood, J. L. Christensen, and I. L. Weissman. Little evidence for developmental plasticity of adult hematopoietic stem cells. *Science*, 297(5590):2256–2259, 2002.

R. L. Walls, B. Athreya, L. Cooper, J. Elser, M. A. Gandolfo, P. Jaiswal, C. J. Mungall, J. Preece, S. Rensing, B. Smith, et al. Ontologies as integrative tools for plant science. *American journal of botany*, 99(8): 1263–1275, 2012.

M. Weinländer, V. Lekovic, S. Spadijer-Gostovic, B. Milicic, G. Krennmair, and H. Plenk Jr. Gingivomorphometry–esthetic evaluation of the crown–mucogingival complex: a new method for collection and measurement of standardized and reproducible data in oral photography. *Clinical oral implants research*, 20(5):526–530, 2009.

T. L. Weissgerber, V. D. Garovic, M. Savic, S. J. Winham, and N. M. Milic. From static to interactive: transforming data visualization to improve transparency. *PLoS biology*, 14(6):e1002484, 2016.

M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016.

C. Winchester. Give every paper a read for reproducibility. *Nature*, 557 (7706):281–282, 2018.

K. Wolstencroft, R. Haines, D. Fellows, A. Williams, D. Withers, S. Owen, S. Soiland-Reyes, I. Dunlop, A. Nenadic, P. Fisher, et al. The taverna workflow suite: designing and executing workflows of web services on

the desktop, web or in the cloud. *Nucleic acids research*, 41(W1):W557–W561, 2013.

World Wide Web Consortium. Web Ontology Language (OWL), 2012. URL https://www.w3.org/2001/sw/wiki/OWL.

F. Yang, W. Li, and H. J. Jørgensen. Transcriptional reprogramming of wheat and the hemibiotrophic pathogen septoria tritici during two phases of the compatible interaction. *PLoS One*, 8(11):e81606, 2013.

Z. Yang, Z. Peng, S. Wei, M. Liao, Y. Yu, and Z. Jang. Pistillody mutant reveals key insights into stamen and pistil development in wheat (triticum aestivum l.). *BMC genomics*, 16(1):1–10, 2015.

J. C. Zadoks, T. T. Chang, C. F. Konzak, et al. A decimal code for the growth stages of cereals. *Weed research*, 14(6):415–421, 1974.

H. Zhang, Y. Yang, C. Wang, M. Liu, H. Li, Y. Fu, Y. Wang, Y. Nie, X. Liu, and W. Ji. Large-scale transcriptome comparison reveals distinct gene activations in wheat responding to stripe rust and powdery mildew. *BMC genomics*, 15(1):1–14, 2014.

A

Appendix for Chapter 3

A.1 Research paper: Samota EK and Davey RP (2021). Knowledge and Attitudes Among Life Scientists Toward Reproducibility Within Journal Articles: A Research Survey. Front. Res. Metr. Anal. 6:678554. doi: 10.3389/frma.2021.6785

frontiers in Research Metrics and Analytics

ORIGINAL RESEARCH
published: 29 June 2021
doi: 10.3389/frma.2021.678554

Knowledge and Attitudes Among Life Scientists Toward Reproducibility Within Journal Articles: A Research Survey

Evanthia Kaimaklioti Samota[1,2]* *and Robert P. Davey*[1]

[1]Earlham Institute, Norwich, United Kingdom, [2]School of Biological Sciences, University of East Anglia, Norwich, United Kingdom

We constructed a survey to understand how authors and scientists view the issues around reproducibility, focusing on interactive elements such as interactive figures embedded within online publications, as a solution for enabling the reproducibility of experiments. We report the views of 251 researchers, comprising authors who have published in *eLIFE* Sciences, and those who work at the Norwich Biosciences Institutes (NBI). The survey also outlines to what extent researchers are occupied with reproducing experiments themselves. Currently, there is an increasing range of tools that attempt to address the production of reproducible research by making code, data, and analyses available to the community for reuse. We wanted to collect information about attitudes around the consumer end of the spectrum, where life scientists interact with research outputs to interpret scientific results. Static plots and figures within articles are a central part of this interpretation, and therefore we asked respondents to consider various features for an interactive figure within a research article that would allow them to better understand and reproduce a published analysis. The majority (91%) of respondents reported that when authors describe their research methodology (methods and analyses) in detail, published research can become more reproducible. The respondents believe that having interactive figures in published papers is a beneficial element to themselves, the papers they read as well as to their readers. Whilst interactive figures are one potential solution for consuming the results of research more effectively to enable reproducibility, we also review the equally pressing technical and cultural demands on researchers that need to be addressed to achieve greater success in reproducibility in the life sciences.

Keywords: reproducibility in life sciences, replication of experiments, reproducibility of computational experiments, interactive figures, reproducibility, reproducibility metrics, open science, reproducibility survey in life sciences

OPEN ACCESS

Edited by:
Hamid R. Jamali,
Charles Sturt University, Australia

Reviewed by:
Tom Crick,
Swansea University, United Kingdom
Iman Tahamtan,
The University of Tennessee, United States

***Correspondence:**
Evanthia Kaimaklioti Samota
evanthia.samota@earlham.ac.uk

Specialty section:
This article was submitted to Scholarly Communication, a section of the journal Frontiers in Research Metrics and Analytics

Received: *09 March 2021*
Accepted: *18 May 2021*
Published: *29 June 2021*

Citation:
Samota EK and Davey RP (2021) Knowledge and Attitudes Among Life Scientists Toward Reproducibility Within Journal Articles: A Research Survey. Front. Res. Metr. Anal. 6:678554. doi: 10.3389/frma.2021.678554

INTRODUCTION

Reproducibility is a defining principle of scientific research, and broadly refers to the ability of researchers, other than the original researchers, to achieve the same findings using the same data and analysis (Claerbout and Karrenbach, 1992). However, irreproducible experiments are common across all disciplines of life sciences (Begley and Ellis, 2012) and many other disciplines (Ioannidis, 2005), such as psychology (Open Science Collaboration, 2015), computer science (Crick et al., 2017), economics (Ioannidis et al., 2017; Christensen and Miguel, 2018) and ecology (Fraser et al., 2018). A

2012 study showed that 88% of drug-discovery experiments could not be reproduced even by the original authors, in some cases forcing retraction of the original work (Begley and Ellis, 2012). Irreproducible genetic experiments with weak or wrong evidence can have negative implications for our healthcare (Yong, 2015). For example, 27% of mutations linked to childhood genetic diseases cited in literature have later been discovered to be common polymorphisms or misannotations (Bell et al., 2011). While irreproducibility is not confined to biology and medical sciences (Ioannidis and Doucouliagos, 2013), irreproducible biomedical experiments pose a strong financial burden on society; an estimated $28 billion was spent on irreproducible biomedical science in 2015 in the United States alone (Freedman et al., 2015).

Reproducibility should inevitably lead to robust science, relating to the way in which conclusions rely on specific analyses or procedures undertaken on experimental systems. Unfortunately, the community has yet to reach consensus on how we traverse the space of re-use, re-analysis and re-interpretation of scientific research to try to define suitable overarching definitions for reproducibility. Thus, there are different definitions of *reproducibility* used in the literature (Drummond, 2009; Plesser, 2018), some of which contradict one another. A recent exhaustive review has also documented this problem (Leipzig et al., 2020), so our survey and results do need to be contextualised somewhat by this lack of consensus. The terms *repeatability, replicability* and *reproducibility* are also occasionally confused (Peng et al., 2006; Liberman, 2015), therefore it is important to differentiate these terms from each other.

1. **Repeatability** The original researchers using the same data, running precisely the same analysis and getting the same results, on multiple runs (Drummond, 2009).

2. **Replicability** Different teams performing different experimental setups and using independent data, achieving the same result as the original researchers, on multiple trials (Peng et al., 2006; Peng, 2011; Stodden et al., 2013a).

3. **Reproducibility** Different teams re-running the same analysis with the same data and getting the same result (Claerbout and Karrenbach, 1992; Peng et al., 2006; Peng, 2011; Stodden et al., 2013a).

It is argued that in many science disciplines replicability is more desirable than reproducibility because a result needs to be corroborated independently before it can be generally accepted by the scientific community (Peng, 2011; Nuijten et al., 2018). However, reproducibility can serve as a cost-effective way of verifying results prior to replicating results (Nuijten et al., 2018).

Computational reproducibility, or reproducible computational research, refers to the reproducibility of computational experiments, where an independent team can produce the same result utilising the data and computational methods (code and workflow) provided by the original authors (Donoho, 2010; Stodden et al., 2013a; Stodden and Miguez, 2013; Stodden et al., 2018; Leipzig et al., 2020). Computational reproducibility is influenced by both technical and cultural (social) factors (LeVeque et al., 2012; Stodden et al., 2013a; Stodden et al., 2018). Technical challenges to computational reproducibility include poorly written, incorrect, or unmaintained software, changes in software libraries on which tools are dependent, or incompatibility between older software and newer operating systems (Cataldo et al., 2009). Cultural factors that challenge computational reproducibility include the attitudes and behaviors of authors when performing and reporting research. Examples include authors not providing sufficient descriptions of methods and being reluctant to publish original data and code under FAIR (Findable, Accessible, Interoperable, and Reusable) principles (Stodden et al., 2013b; Baker, 2016; Munafo et al., 2017). Other cultural factors include favoring of high prestige or high impact scientific publications over performing rigorous and reproducible science (which tends to be improved by open access policies) (Eisner, 2018; Hardwicke et al., 2018). We refer to the cultural factors affecting computational reproducibility as the *culture of reproducibility* (Peng, 2011).

Several projects have attempted to address some of the technical aspects of reproducibility by making it easier for authors to disseminate fully reproducible workflows and data, and for readers to perform computations. For example F1000 Living Figure (Colomb and Brembs, 2014) and re-executable publications (Ingraham, 2017; Perkel, 2017; Ingraham, 2017) using Plotly (plot.ly) and Code Ocean widgets (codeocean.com); Whole Tale Project (Brinckman et al., 2018); ReproZip project (Chirigati et al., 2016); Python-compatible tools and widgets (interactive widgets for Jupyter Notebooks with Binder); Zenodo (zenodo.org) and FigShare (figshare.com) as examples of open access repositories for scientific content (including datasets, code, figures, reports); Galaxy (Afgan et al., 2018); CyVerse (formerly iPlant Collaborative) (Goff et al., 2011); myExperiment (Goble et al., 2010); UTOPIA (Pettifer et al., 2009; Pettifer et al., 2004); GigaDB (Sneddon et al., 2012); Taverna (Hull et al., 2006; Oinn et al., 2004; Wolstencroft et al., 2013); workflow description efforts such as the Common Workflow Language (Amstutz et al., 2016); and Docker (docker.com), Singularity (sylabs.io) (Kurtzer et al., 2017) and other container systems. Even though these tools are widely available and seem to address many of the issues of *technical reproducibility* and the *culture of reproducibility*, they have not yet become a core part of the life sciences experimental and publication lifecycle. There is an apparent disconnection between the development of tools addressing reproducibility and their use by the wider scientific and publishing communities who might benefit from them.

This raises the question of "how do researchers view their role in the production and consumption of scientific outputs?" A common way for researchers to quickly provide information about their data, analysis and results is through a figure or graph. Scientific figures in publications are commonly presented as static images. Access to the data (including the raw, processed and/or aggregated data), analysis, code or

description of how the software was used that produced the figure are not available within the static images (Barnes and Fluke, 2008; Barnes et al., 2013; Grossman et al., 2016; Newe, 2016; Weissgerber et al., 2016; Rao et al., 2017; Perkel, 2018). This can be especially pertinent to figures that have thousands or millions of points of data to convey (Perkel, 2018). In order for readers to interrogate published results in more detail, examine the transparency and reproducibility of the data and research, they would need to download a complete copy of the data, code, and any associated analysis methodology (data pre-processing, filtering, cleaning, etc) and reproduce this locally, provided all those elements are available and accessible (Stodden et al., 2016). Computational analyses often require running particular software which might require configuration and parameterisation, as well as library dependencies and operating system prerequisites. This is a time-consuming task and achieving reproducibility of computational experiments is not always possible (Stodden et al., 2016; Kim et al., 2018). Thus, solutions that automatically reproduce computational analyses and allow the investigation of the data and code presented in the figure in detail would be advantageous (Peng, 2011; Perkel, 2017; Perkel, 2018).

Many solutions now exist that allow for the reproducibility of computational analyses outside the research paper and are typically supplied as links within the research paper or journal website redirecting to many different types of computational systems, such as Galaxy workflows, Binder interactive workspaces converted by GitHub repositories with Jupyter notebooks (Jupyter et al., 2018), and myExperiment links (Goble et al., 2010). The endpoint of these analyses are often graphical figures or plots, and these may well be interactive, thus allowing modification of plot type, axes, data filtering, regression lines, etc. Whilst these figures may well be interactive in that a user can modify some part of the visualisation, this does not implicitly make the data or code that produced that figure more available, and hence more reproducible.

Technologies that can expose code, data and interactive figures are now mature. For example, Jupyter notebooks are built up of executable "cells" of code which can encapsulate a link to a data file hosted on a cloud service, code to get and analyze this data file, and then produce an interactive figure to interpret the dataset. Again, this is somewhat disconnected from the actual research publication. However, as technology has progressed in terms of available storage for data, computational power on the web through cloud services, and the ability of these services to run research code, we are now coming to the point where the production of interactive figures within publications themselves is achievable. These interactive figures which would inherently have access to the underlying data and analytical process can provide users with unique functionality that can help increase the reproducible nature of the research. This combination of code, data, analysis, visualisation and paper are examples of "executable documents" (Ghosh et al., 2017; Maciocci et al., 2019).

Interactive figures within executable documents, therefore, have incorporated data, code and graphics so that when the user interacts with the figure, perhaps by selecting a cluster of data points within a graph, the user could then be presented with the data that underlies those data points. Similarly, a user could make changes to the underlying parameters of the analysis, for example modifying a filter threshold, which would ultimately make changes to the visualisation of the figure or the document itself (Barnes and Fluke, 2008; Barnes et al., 2013; Grossman et al., 2016; Newe, 2016; Weissgerber et al., 2016; Rao et al., 2017; Perkel, 2018). By means of an example, an executable document could represent an interactive figure showing a heat map of gene expression under different stress conditions. In a traditional article, the user would be tasked with finding references to the datasets and downloading them, and subsequently finding the code or methodology used to analyze the data and retrace the original authors' steps (if the code and data were available at all). Within an interactive figure in an executable document, a user could select a particular gene of interest by clicking on the heatmap and viewing the gene expression information within a pop-up browser window. Whilst this is useful for general interpretation, to achieve reproducibility this pop-up window would provide a button that allows the user to pull the sequencing read data that was the basis for the results into a computational system in order to re-run the differential expression analysis. This raises many questions around how this infrastructure is provided, what technologies would be used to package up all elements needed for reproducibility, the subsequent costs of running the analysis, and so on.

These caveats aside, interactive figures within executable documents can benefit the reader for the consumption of the research outputs in an interactive way, with easy access to the data and removing the need for installing and configuring code and parameters for reproducing the computational experiments presented in the figure within the publication (Perkel, 2017). The aforementioned solutions would not only be helpful to the readers of papers (Tang et al., 2018) but benefit the peer review process (Perkel, 2018).

There have been efforts to make the connection between production and consumption of research outputs within online publications. One of the first interactive figures to have been published in a scholarly life sciences journal is the *Living Figure* by Björn Brembs and Julien Colomb which allowed readers to change parameters of a statistical computation underlying a figure (Ghosh et al., 2017). *F1000Research* has now published more papers that include Plotly graphs and Code Ocean widgets in order to provide interactivity and data and code reproducibility from within the article figures (Ghosh et al., 2017; Ingraham, 2017). The first prototype of *eLIFE's* computationally reproducible article aims to convert manuscripts created in a specific format (using the Stencila Desktop, stenci.la, and saved as a Document Archive file) into interactive documents allowing the reader to "play" with the article and its figures when viewed in a web browser (Maciocci et al., 2019). The Manifold platform (manifoldapp.org) allows researchers to show their research objects alongside their publication in an electronic reader, whilst including some dynamic elements. The *Cell* journal included interactive figures in a paper using Juicebox js for 3D visualisation of Hi-C data (http://aidenlab.org/juicebox/) (Rao et al., 2017; Robinson et al., 2018). Whilst there are few incentives to promote the culture of

reproducibility (Pusztai et al., 2013; Higginson and Munafò, 2016), efforts in most science domains are being made to establish a culture where there is an expectation to share data for all publications according to the FAIR principles. The implementation of these principles is grounded in the assumption that better reproducibility will benefit the scientific community and the general public (National Institutes of Health, 2015; Wilkinson et al., 2016). Studies have suggested that reproducibility in science is a serious issue with costly repercussions to science and the public (Stodden et al., 2013b; Pulverer, 2015). Whilst there have been survey studies canvassing the attitudes of researchers around reproducibility in other disciplines to some extent (Baker, 2016; Feger et al., 2019; Stodden, 2010), fewer studies have investigated the attitudes and knowledge of researchers around reproducibility in the life sciences (Baker, 2016). In particular, minimal research has been conducted into the frequency of difficulties experienced with reproducibility, the perception of its importance, and preferences with respect to potential solutions among the life sciences community.

This paper presents a survey that was designed to assess researchers' understanding of the concepts of reproducibility and to inform future efforts in one specific area: to help researchers be able to reproduce research outputs in publications. The development of tools, one example of which are interactive figures within journal publications, may better meet the needs of producers and consumers of life science research. Our survey is limited in that we do not assess how open-access tools for the production of reproducible research outputs compare, but how the consumption of research information through interactive means is regarded. We constructed the survey in order to understand how the following are experienced by the respondents:

- *Technical factors affecting computational reproducibility*: issues with accessing data, code and methodology parameters, and how solutions such as interactive figures could promote reproducibility from within an article.
- *Culture of reproducibility*: attitudes toward reproducibility, the social factors hindering reproducibility, and interest in how research outputs can be consumed *via* interactive figures and their feature preferences.

METHODS

Population and Sample

The data were analyzed anonymously, nonetheless, we sought ethical approval. The University of East Anglia Computing Sciences Research Ethics Committee approved this study (CMPREC/1819/R/13). Our sample populations were selected to include all life sciences communities across levels of seniority, discipline and level of experience with the issues we wished to survey. The first survey was conducted in November 2016 and sent out to 750 researchers working in the Norwich Biosciences Institutes (NBI) at a post-doctoral level or above. We chose to survey scientists of post-doctoral level or above, as these scientists are more likely to have had at least one interaction with publishing in scientific journals. The NBI is a partnership of four United Kingdom research institutions: the Earlham Institute (formerly known as The Genome Analysis Center), the John Innes Center, the Sainsbury Center, and the Institute of Food Research (now Quadram Institute Bioscience). Invitations to participate were distributed *via* email, with a link to the survey. The second survey, similar to the first but with amendments and additions, was distributed in February 2017 to a random sample of 1,651 researchers who had published papers in the *eLIFE* journal. Further information about the *eLIFE* sample is found in Supplementary section 3. Invitations to participate were sent using email by *eLIFE* staff. We achieved a 15% ($n = 112$) response rate from the NBI researchers and an 8% response rate from the *eLIFE* survey ($n = 139$). **Table 1** shows the survey questions. Questions were designed to give qualitative and quantitative answers on technical and cultural aspects of reproducibility. Questions assessed the frequency of difficulties encountered in accessing data, the reasons for these difficulties, and how respondents currently obtain data underlying published articles. They measured understanding of what constitutes reproducibility of experiments, interactive figures, and computationally reproducible data. Finally, we evaluated the perceived benefit of interactive figures and of reproducing computational experiments, and which features of interactive figures would be most desirable.

Validation of the Survey Design

We undertook a two-step survey: firstly NBI, then *eLIFE* interactions leading to additional questions. We tested the initial survey on a small cohort of researchers local to the authors to determine question suitability and flow. We reported the qualitative results of the surveys in accordance with the Standards for Reporting Qualitative Research (SRQR) (O'Brien et al., 2014).

The survey questions were not designed based on specific culture theory, but rather on our understanding of the field of reproducibility, that is the human factors and researcher attitudes toward reproducibility, as well as the mode of conducting science. We assume that these factors affect how reproducible and robust the science, and therefore the published work, will be. Therefore, we adopt the term "culture of reproducibility" to encompass the attitudes of life scientists toward science and reproducibility directly related to research articles, and not referring to human demographics. The rationale behind evaluating the culture of reproducibility was to examine how the attitudes or means by which researchers present their work in research papers can affect reproducibility.

It is important to state that not any one survey question was assessing solely the technical factors affecting reproducibility or solely the culture of reproducibility. For example, accessing data for the reader is both a cultural and technical factor, i.e., data available from public repositories *via* persistent identifiers and APIs vs. "data available on request". For the author of the paper, not publishing the data is solely a cultural/social factor as it could be seen that they are not conducting and presenting their research in an open reproducible manner, or they do not have the support or knowledge around the best practice for reproducible data publishing in their domain.

TABLE 1 | Questions used to survey the knowledge of respondents about research reproducibility.

	Survey questions
1	How often do you encounter difficulties with working with bioinformatic analysis tools (that are not your own)? (Problems such as: installing, configuring, running the software, working with command-line software)?
2	How difficult is it to source (or access) the data presented in published papers?
3	What difficulties have you encountered in accessing the data described in published papers?
4	How are you currently sourcing the data (if applicable)? Select all that apply to you.
5[a]	What is your current understanding of the reproducibility of experiments? Please select any that apply. Should you wish to add any additional information, please add it to the "Other" box.
6[a]	Have you ever tried reproducing any published results? Please select the answer that applies best for you.
7[a]	In your opinion, what could be done to make published research more reproducible? Other please specify (free text answer).
8	When thinking about interactive figures, what comes to your mind? (please describe what you understand of what an interactive figure to be, its features, and where you have seen such a feature before if applicable).
9	An interactive figure is a figure within a paper that is dynamic and becomes "live" when the user interacts with it and where the data displayed changes according to various parameter options. Which of the following features of an interactive figure tool would be good to have? Please rank them in the order of preference, where 1 is the most preferred feature, and 11 the least preferred feature.
10	What other features an interactive figure could have that were not mentioned in the previous question?
11	Do you perceive a benefit in being able to publish interactive figures?
12	Does the provision or option of an interactive figure in the paper affect your decision in choosing the publishing journal or publisher?
13	Have you heard of the term computationally reproducible data, and do you understand what the term means? If answered yes or unsure, please explain what you understand from the term.
14	Would you benefit from being able to automatically reproduce computational experiments or other analyses (including statistical tests) described within a paper?
15	How often do you work with bioinformatic analysis tools (e.g., assemblers, aligners, structure modeling)?
16	Have you received any of the following training? Training whether formal or informal (training through a colleague etc.).
17	Which of the following type(s) of data do you work with?

[a]*Questions indicated with an asterisk were only available to the eLIFE survey. Answer options to the questions are shown in Supplementary section 1.*

We also evaluated the sentiment around interactive figures in our research. In themselves, they are a technical factor that we suggest can promote reproducibility. However, the interest of the readers in finding interactive figures desirable, including what features they think are favourable, can be a variable factor depending on the social background or demographic of the respondent (for example, training received, data they work with, discipline they work in).

We understand that there are a lot of human factors in the way reproducibility is achieved, which are mainly centered around the attitude of life scientists toward reproducibility. How robust, open-source, open-access they conduct and share their research affects the reproducibility of their work. In this way we wanted to evaluate, assess and see the extent of the issue in quantifying and qualifying how difficult it is to access data and code presented in papers, and how difficult it is to understand the methods presented in a paper. Our work adds to existing surveys that also highlight reproducibility as an issue.

We received consistent responses where all or most respondents interpreted the questions in a similar manner suitable for cross-comparison. The NRP study results produced consistent results with the *eLIFE* study results, which were sent out at different times and to the different survey cohorts. We primarily took surveys and conclusions raised in the existing literature and the results of discussions with various researchers who are looking into reproducibility in our local institutions to form our construct validity. The questions we asked indicated they fit our requirements to better understand the qualitative nature of the respondents' answers and perform empirical analyses (Chi-squared) to show relationships. We used the same process to determine content validity, where we tried to provide questions that would cover the breadth of the domain we were assessing as we "cast a wide net" over potential respondents that would comprise people from a wide variety of domains, expertise, and other demographics. Finally, our theory of how researchers view reproducibility fed into our questions to provide translation validity where we formed two practical surveys based on our theory assessments from previous literature.

Statistical Analysis

Results are typically presented as proportions of those responding, stratified by the respondent's area of work, training received, and version of the survey as appropriate. Chi-square tests for independence were used to test for relationships between responses to specific questions, or whether responses varied between samples. The analysis was conducted using R (version 3.5.2; R Core Team, 2018) and Microsoft Excel. All supplementary figures and data are available on Figshare (see Data Availability).

We assessed if there was a significant difference in the ability and willingness to reproduce published results between the cohort of *eLIFE* respondents who understand the term "computationally reproducible data" and those who do not and whether training received (bioinformatics, computer science, statistics, or no training) had an effect. Given the free-text responses within the "unsure" group as to the understanding of the term "computationally reproducible data", where many understood the term, we did not

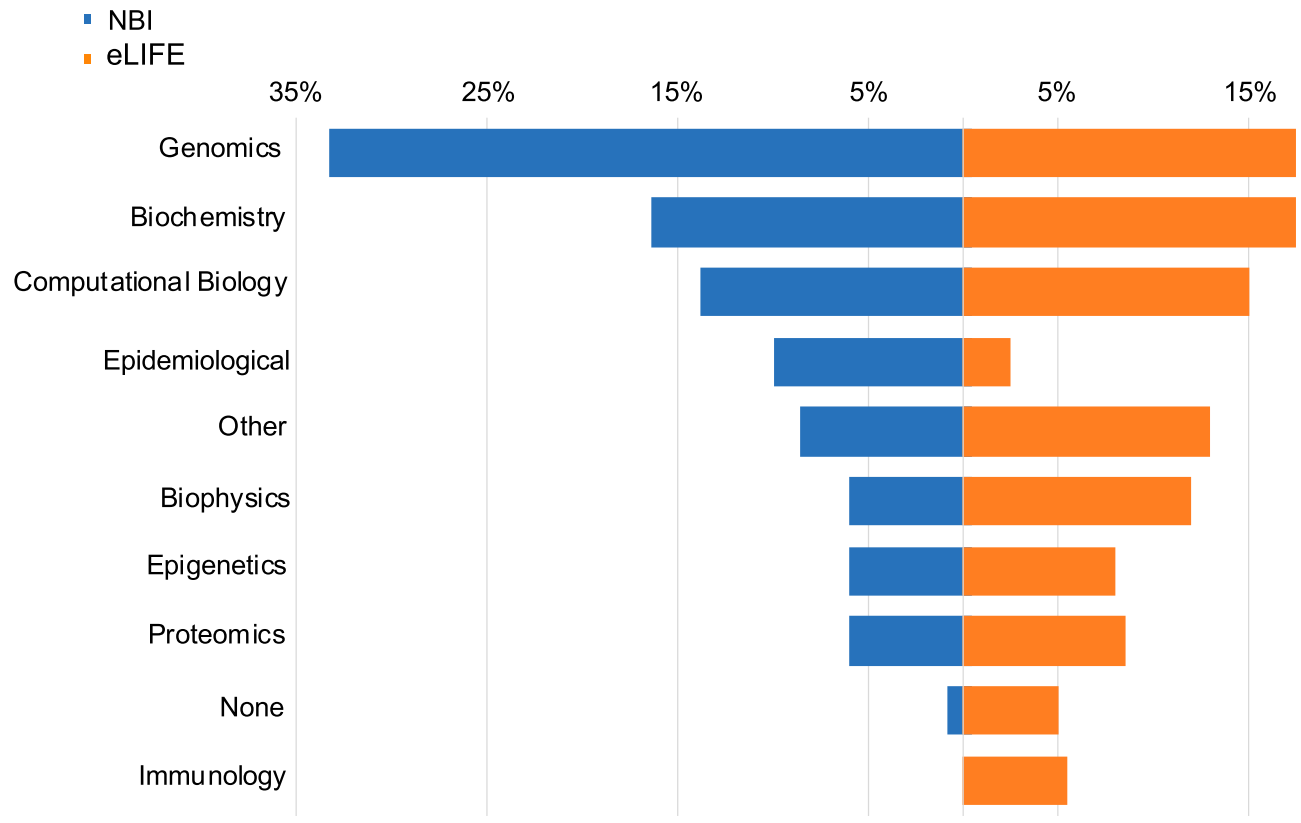

FIGURE 1 | Data types used by NBI and *eLIFE* respondents. Responses were not mutually exclusive. Data type choices were the same as the article data types available in the *eLIFE* article categorisation system.

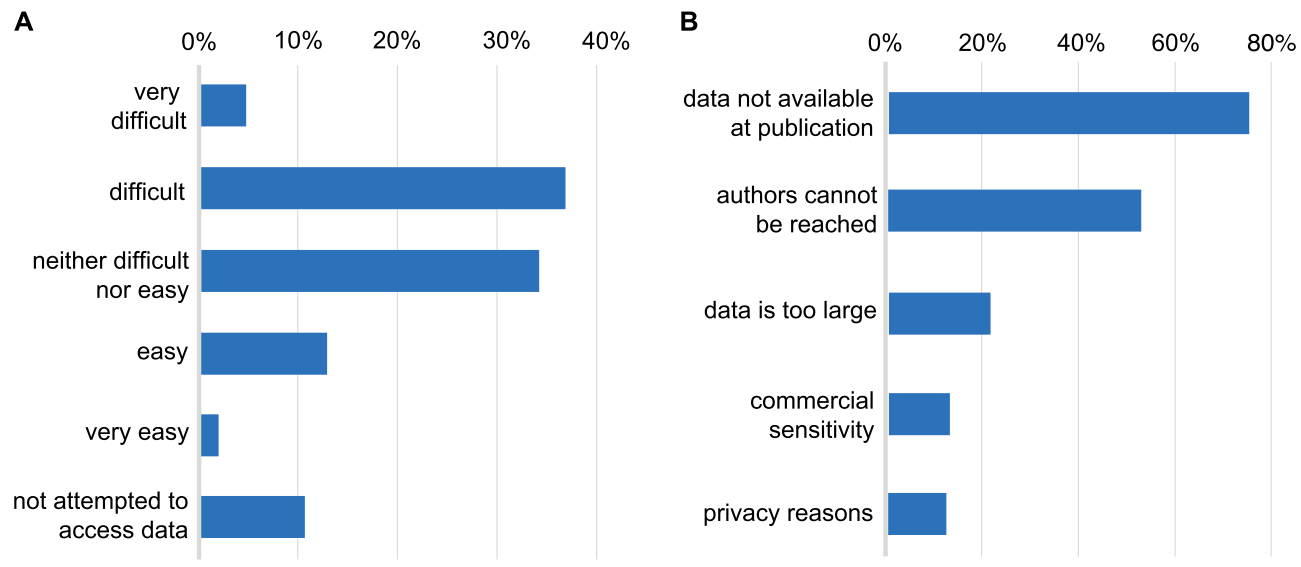

FIGURE 2 | (A): Difficulty encountered accessing data underlying published research. Whether respondents have attempted to access data underlying previous publications and the level of difficulty typically encountered in doing so. **(B)**: Reasons given for difficulty accessing data. The reasons given by respondents for being unable to access data (restricted to those who have attempted to access data).

include in our analysis the data from those who replied, "unsure" (see Section "Understanding of reproducibility, training and successful replication" below). The respondents who chose "yes tried reproducing results, but unsuccessfully", "have not tried to reproduce results" and "it is not important to reproduce results" were grouped under "unsuccessfully".

RESULTS

Characteristics of the Sample

Figure 1 shows the distribution of areas of work of our respondents, stratified by survey sample. Genomics (proportion in the whole sample = 22%), biochemistry (17%), and computational biology (15%) were the most common subject areas endorsed in both NBI and *eLIFE* samples. With regard to how often respondents use bioinformatics tools, 25% replied "never", 39% "rarely", and 36% "often". Many (43%) received statistical training, (31%) bioinformatic training, (20%) computer science training.

Access to Data and Bioinformatics Tools

In both samples, 90% of those who responded, reported having tried to access data underlying a published research article (**Figure 2**). Of those who had tried, few had found this "easy" (14%) or "very easy" (2%) with 41% reporting that the process was "difficult" and 5% "very difficult". Reasons for difficulty were chiefly cultural (**Figure 2**), in that the data was not made available alongside the publication (found by 75% of those who had tried to access data), or authors could not be contacted or did not respond to data requests (52%). Relatively few found data unavailable for technical reasons of data size (21%), commercial sensitivity (13%) or confidentiality (12%). With respect to data sources, 57% of the total sample have used open public databases, 48% reported data was available with a link in the paper, and 47% had needed to contact authors.

Very few of the respondents either "never" (2%) or "rarely" (8%) had problems with running, installing, configuring bioinformatics software. Problems with software were encountered "often" (29%) or "very often" (15%) suggesting that nearly half of respondents regularly encountered technical barriers to computational reproducibility.

Understanding of Reproducibility, Training and Successful Replication

Most respondents reported that they understood the term "reproducibility of experiments" and selected the explanation for the term as defined in the introduction above, which corresponds to the most established definitions of reproducibility (Peng et al., 2006; Peng, 2011; Stodden et al., 2013b). It is important to state that for this question, we allowed for respondents to choose more than one answer, as we recognise the limitation that there is no standard and accepted definition for reproducibility, as well as the familiarity of the term between scientists from different backgrounds, can differ. The first three definitions are plausible definitions for reproducibility. Given the results, we can assume that some of the respondents chose both correct and wrong definitions. The majority of the answers (77%) included the definition of reproducibility as we define it in the manuscript. However, by looking into the individual responses (n = 54), 11.1% (n = 6) of respondents chose only option A thus appeared to understand that this matched the definition of reproducibility, as we state in the manuscript. 5.5% (n = 3) chose only option D, which is incorrect. The majority of people (57%, n = 23) picked any of A, B, or C and did *not* pick D, which seems to suggest that they understand that replicability is not reproducibility, but they are still not clear on exact definitions, which matches the general lack of consensus (Drummond, 2009; Liberman, 2015; Plesser, 2018). Just over a third (37%, n = 20) picked one or all of A, B and C, *and* picked D, which seems to suggest that they didn't understand the difference between reproducibility and replicability at all and considered any form of repeating a process could be classed as reproducibility of experiments (see Supplementary Table 4).

Most (52%) participants provided a different interpretation of the term "computationally reproducible data" to our interpretation, while 26% did know and 22% were unsure. We received several explanations (free text responses) of the term of which the majority were accurate (Supplementary section 2, free responses to question 13). We assign meaning to the term as data as an output (result) in a computational context, which was generated when reproducing computational experiments. Although the term "computationally reproducible data" is not officially defined, other sources and studies have referred to the concept of data that contributes to computational reproducibility (Baranyi and Greilhuber, 1999; Weinländer et al., 2009; de Ruiter, 2017; Perkel, 2017; Tait, 2017; Pawlik et al., 2019). From the unsure responses (n = 30), we categorised those that gave free-text responses (70%, n = 21, see Supplementary section 2, free responses) into whether they did actually understand the term, those that did not understand the term, and those that did not give any free text. The majority of respondents that chose "unsure" and gave a free text response (71%, n = 15) did understand the term "computationally reproducible data". The remaining 29% (n = 6) did not understand the term correctly.

Some (18%) reported not attempting to reproduce published research. Very few (n = 5; 6%) of the sample endorsed the option that "it is not important to reproduce other people's published results" (Supplementary figure 1). Even though the majority (60%) reported successfully reproducing published results, almost a quarter of the respondents found that their efforts to reproduce any results were unsuccessful (23%). **Table 2** shows respondents' ability to reproduce experiments, stratified by their understanding of the term "computationally reproducible data" and the training received (bioinformatics, computer science, statistics). A chi-square test of independence was performed to examine the relationship between the ability to reproduce published experiments and knowing the meaning of the term "computationally reproducible data". The relationship between these variables was significant, $\chi^2(1, n = 75) = 3.90$, p = .048. Those who knew the meaning of the term "computationally reproducible data" were more likely to be

TABLE 2 | Success in reproducing any published results stratified by their knowledge of the term "computationally reproducible data" and training received.

Variable	Number (% of the total sample)	Success in reproducing any published results		
		Successful (% within variable)	Not Successful[a] (% within variable)	*p*-value
Knowledge of the term "computationally reproducible data" (*n* = 75)				
Yes	25 (33.3)	18 (72)	7 (28)	0.048[b]
No	50 (66.7)	24 (48)	26 (33)	
Training (*n* = 90)				
Bioinformatics	42 (46.7)	26 (61.9)	16 (38.1)	0.73
Not trained in Bioinformatics	48 (53.3)	28 (58.3)	20 (41.7)	
Computer Science	33 (36.7)	21 (63.6)	12 (36.4)	0.59
Not trained in Computer Science	57 (63.3)	33 (57.9)	24 (42.1)	
Statistics	71 (78.9)	42 (59.2)	29 (40.8)	0.75
Not trained in Statistics	19 (21.1)	12 (63.2)	7 (36.8)	
No training	10 (11.1)	6 (60)	4 (40)	0.73[c]
All other training	80 (88.8)	48 (60)	32 (40)	

n is different for the two variables as not all participants answered all the questions.

[a]*Unsuccessful includes answers: "Yes, I have tried reproducing published results, but I have been unsuccessful in producing any results, or the same results", "No, I have never tried reproducing any published results" and "It is not important to reproduce other people's published results".*

[b]*Statistically significant at the level of* $p < 0.05$.

[c]*Chi-square statistic with Yates correction, applied when expected frequencies were lower than five.*

able to reproduce published experiments. Taking their training background into account did not show any significant difference. However, when testing with the responses "yes tried reproducing results, but unsuccessfully", "have not tried to reproduce results" and "it is not important to reproduce results" (not grouped under "unsuccessfully" in order to get an indication of how willingness and success together differed between the training groups), we found a significant difference (see Supplementary Table 1). The distribution of the training variable with those who received computer science training and those without was significantly different (Fisher exact test for independence, $p = 0.018$). It appears that respondents with computer science training are less likely to have tried to reproduce an experiment but be more likely to succeed when they did try.

There was no evidence for a difference in the ability and willingness to reproduce published results between the respondents who use bioinformatics tools often, and those who use them rarely or never, $\chi^2(3, n = 90) = 0.53$, $p = 0.91$ (Supplementary Table 2). The majority of the respondents who use bioinformatics tools often were coming from the scientific backgrounds of Biophysics, Biochemistry, Computational Biology and Genomics. Most of the respondents who answered "reproducibility is not important" and "haven't tried reproducing experiments" were scientists coming from disciplines using computational or bioinformatics tools "rarely" or "never" (Supplementary Table 3).

Improving Reproducibility of Published Research

The majority (91%) of respondents stated that authors describing all methodology steps in detail, including any formulae analysing the data can make published science more reproducible. Around half (53%) endorsed the view that "authors should provide the source code of any custom software used to analyze the data and that the software code is well documented", and that authors provide a link to the raw data (49%) (Supplementary figure 2). Two respondents suggested that achieving better science reproducibility would be easier if funding was more readily available for reproducing the results of others and if there were opportunities to publish the reproduced results (Supplementary section, free responses). Within the same context, some respondents recognised the current culture in science that there are not sufficient incentives in publishing reproducible (or indeed negative findings) papers, but rather being rewarded in publishing as many papers as possible in high impact factor journals (Supplementary section, free responses).

Interactive Figures

Participants ranked their preferences for interactive figure features within a research article. The most preferred interactive figure feature was "easy to manipulate", followed by "easy to define parameters" (**Figure 3**). Generally, the answers from both the *eLIFE* and NBI surveys followed similar trends. Furthermore, free-text responses were collected, and most respondents stated that mechanisms to allow them to better understand the data presented in the figure would be beneficial, e.g., by zooming in on data (Supplementary section, free responses).

The majority of the respondents perceive a benefit in having interactive figures in published papers for both readers and authors (**Figure 4**). Examples of insights included: the interactive figure would allow visualising further points on the plot from data in the supplementary section, as well as be able to alter the data that is presented in the figure; having an interactive figure like a movie or to display protein 3D structures, would be beneficial to readers. The remaining responses we categorised as software related, which

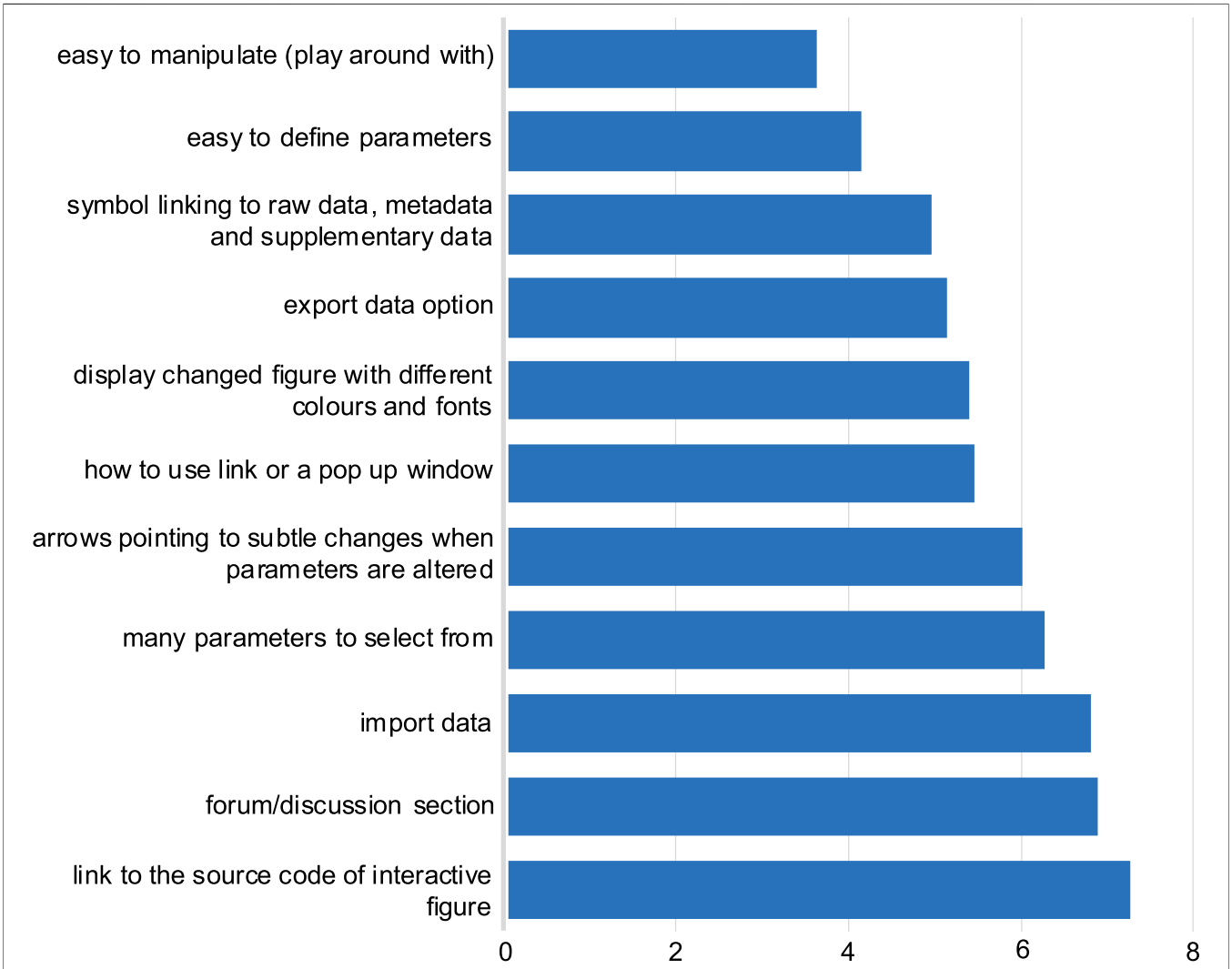

FIGURE 3 | Preferred features for the interactive figure. Responses to question 9: Respondents were asked to rank in order of preference the above features, with one most preferred feature, to 11 the least preferred feature. The average score for each feature was calculated in order of preference as selected by the respondents from both NBI and *eLIFE* surveys. The lower the average score value (*x*-axis), the more preferred the feature (*y*-axis).

included suggestions of software that could be used to produce a figure that can be interactive, such as R Shiny (shiny.studio.com). We received a total of 114 free-text responses about the respondents' opinions on what interactive figures are and a proportion of those (25%) suggested that they had never seen or interacted with such a figure before, and no indication was given that an interactive figure would help their work (see Supplementary section, free responses).

The majority of the respondents also said that they see a benefit in automatically reproducing computational experiments and manipulating and interacting with parameters in computational analysis workflows. Equally favourable was to be able to computationally reproduce statistical analyses (**Figure 5**). Despite this perceived benefit, most respondents (61%) indicated that the ability to include an interactive figure would not affect their choice of a journal when seeking to publish their research.

Limitations

The findings were collected using the self-reporting method which can be limited in certain ways, especially with regards to the reported reproducibility success or lack of success of the respondents. We do not know categorically that someone reproduced experiments successfully because they checked the box. Despite the potential for confusing the exact meaning of reproducibility, which could affect the answers to questions five, six and seven, the general consensus among respondents showed that the questions were sufficiently phrased to help us divide people into two groups of assessment (successful vs not successful) for subsequent analysis.

Part of our survey sample were researchers from the NBI, and this population might not be representative of the life sciences research community. Researchers working in academic institutions may have attitudes, incentives, or infrastructure to support reproducibility that may be different from those

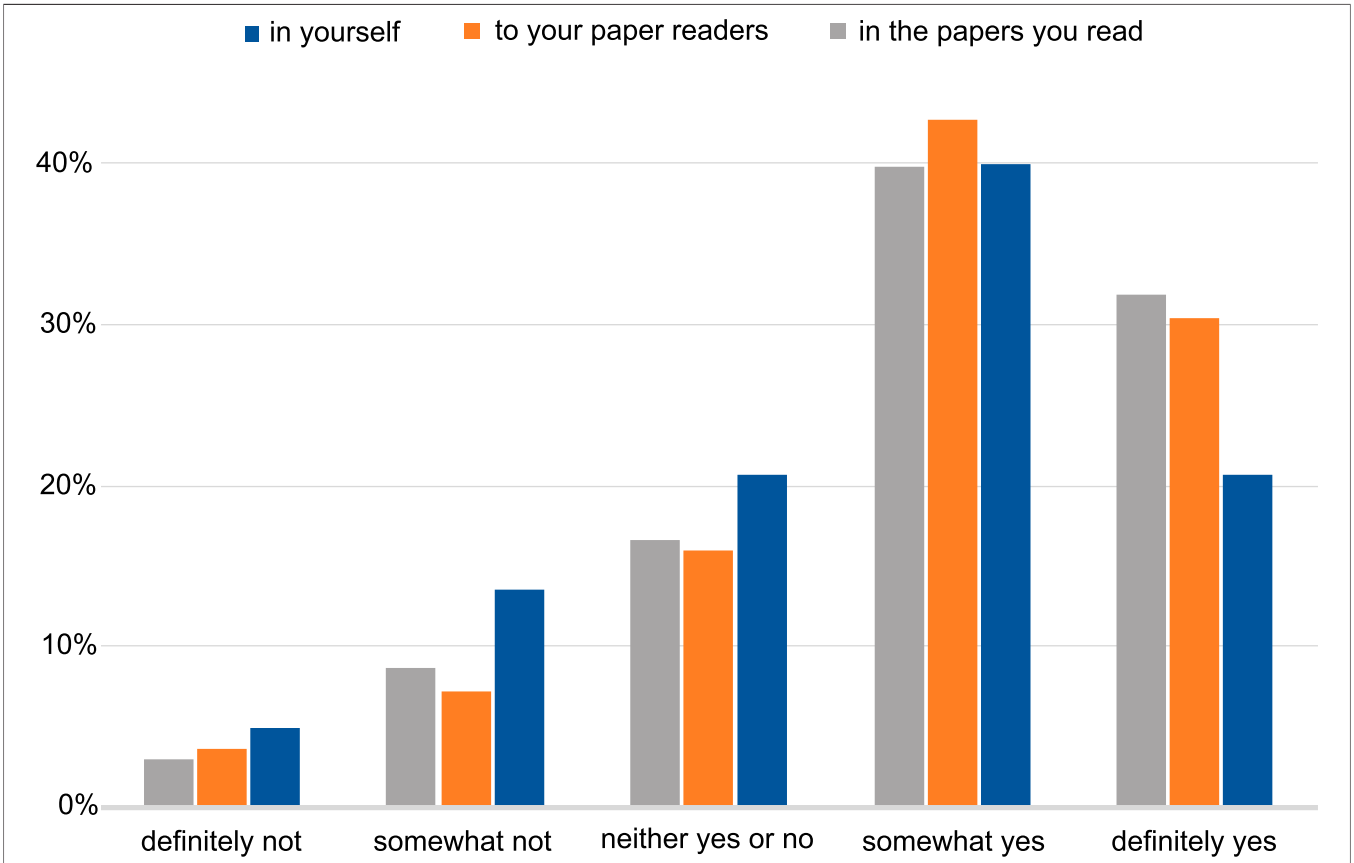

FIGURE 4 | The level of perception of benefit to having the ability to publish papers with interactive figures. The benefit to the author, to the readers of the author's papers and to the papers the author reads. Answers include the responses from both NBI and *eLIFE* surveys for question 11.

who work in the private sector or government agencies. In addition, as the population of the NBI researchers was solely United Kingdom based, the attitudes of these researchers might differ from those in the rest of the world, even though the NBI comprises scientists who are from multiple countries and have trained and worked in global institutions. *eLIFE* authors work across the breadth of scientific institutions, both private and public, from the international stage, thus we believe that both *eLIFE* and NBI participants to be sufficiently representative for the purposes of our survey study.

Although we do not have distinct evidence that the *eLIFE* authors' cohort had a predisposition to reproducibility, and the authors we surveyed were randomly selected, we acknowledge that as *eLIFE* is a journal that requires data sharing and is also heavily involved in reproducibility efforts, such as the Reproducibility Project: Cancer Biology. In the absence of data to the contrary, thus it is reasonable to assume that some factors might have influenced the *eLIFE* respondents' opinions about reproducibility. We do not think that this fact undermines our conclusion, but it is a factor that future studies should be aware of when drawing comparisons that can shed further light on this issue.

We acknowledge that questions eight and nine were on the same page when the participants were taking the survey and seeing the two questions together might have introduced bias into their answers. Nonetheless, free text answers to question eight included answers which were not presented as options for question nine. Some respondents also declared that they were not aware of, or have not previously encountered, interactive figures (see Supplementary section free-text responses to question eight).

We have found that the response rate for studies of this nature is fairly typical and indeed, other studies (Koschke, 2003; Snell and Spencer, 2005; Federer et al., 2015; Schneider et al., 2016; Barone et al., 2017) have experienced comparable or lower rates. Ideally, we would want to aim for a higher response rate for future studies, which could be achieved by providing monetary incentives, as well as sending email reminders to the same or bigger cohort of invited people to participate in the study (James and Bolstein, 1990; Shettle and Mooney, 1999; Jobber et al., 2004).

DISCUSSION

This study highlights the difficulties currently experienced in reproducing experiments and conveys positive attitudes of scientists toward enabling and promoting reproducibility of

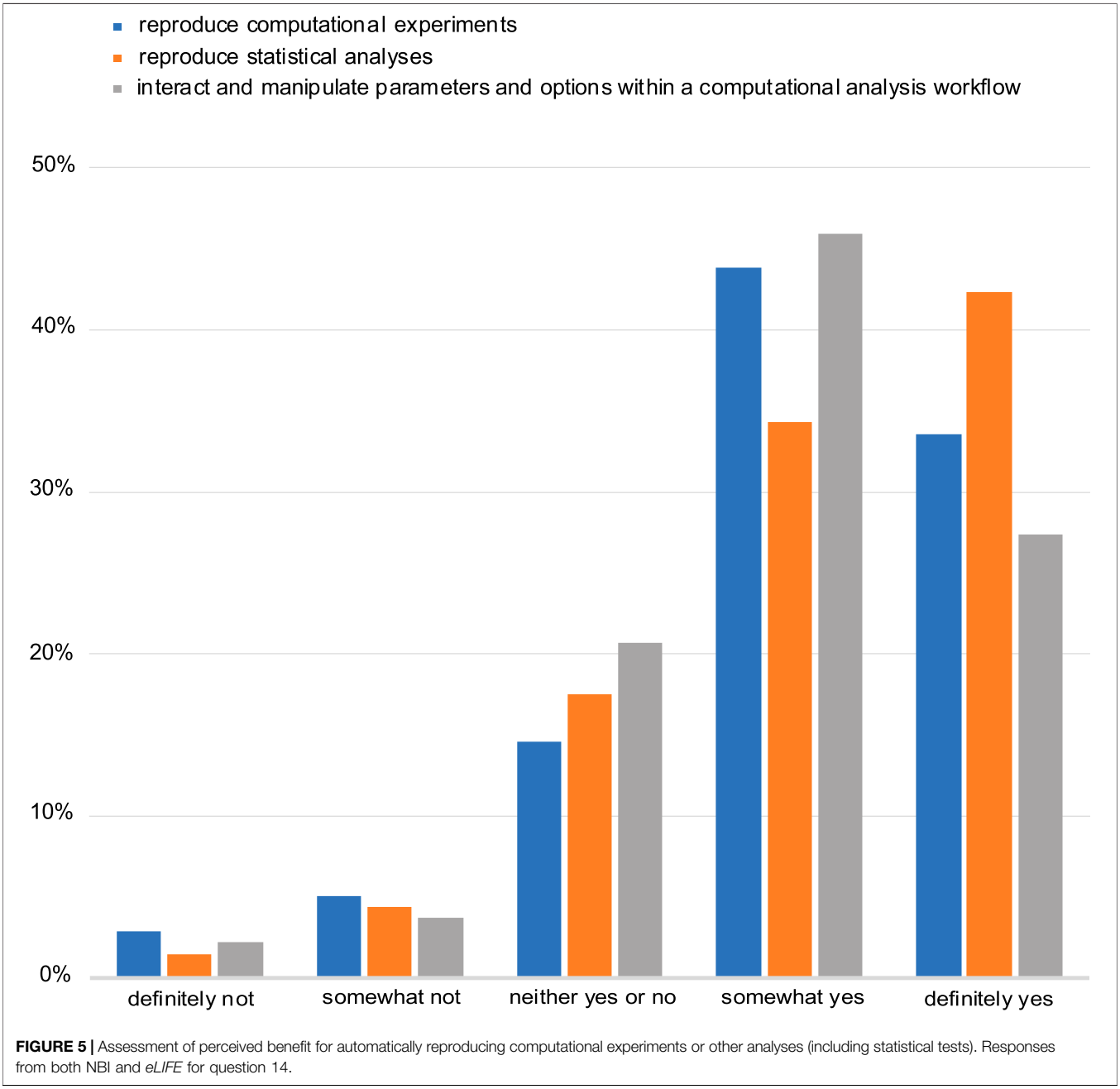

FIGURE 5 | Assessment of perceived benefit for automatically reproducing computational experiments or other analyses (including statistical tests). Responses from both NBI and *eLIFE* for question 14.

published experiments through interactive elements in online publications. The NBI cohort of respondents were active life sciences researchers at the time the survey was conducted, and the *eLIFE* cohort were researchers that have published at least once in the *eLIFE* journal; therefore we believe the opinions collected are representative of researchers in life sciences who are routinely reading and publishing research.

While progress has been made in publishing standards across all life science disciplines, the opinions of the respondents reflect previously published shortcomings of the publishing procedures (Müller et al., 2003; Tenopir et al., 2011; Marx, 2013; Stodden, 2015): lack of data and code provision; storage standards; not including or requiring a detailed description of the methods and code structure (i.e., code scripts, algorithms, full software packages, language used, versions of any libraries required, organisation of any modular components, configuration and deployment options) in the published papers. However, the level of interest and incentives in reproducing published research is in its infancy, or it is not the researchers' priority (Collins and Tabak, 2014; Nosek et al., 2015). A key outcome of our survey is the acknowledgment of the large majority who understand that science becomes implicitly more reproducible if

methods (including data, analysis, and code) are well-described and available. Respondents also perceive the benefit of having tools that enable the availability of data, methods and code and being able to automatically reproduce computational experiments described within the paper. Interactive figures within publications and executable documents can be such tools that allow the automatic reproducibility of computational experiments, or other analyses described within the paper, interact and manipulate parameters within the computational analysis workflow and give further insights and detailed view of the data in the figure. Despite technologies existing to aid reproducibility (Crick et al., 2014) and authors knowing they are beneficial, many scientific publications do not meet basic standards of reproducibility.

Our findings are in accordance with the current literature (Pulverer, 2015; Berg, 2018) that highlight that the lack of access to the data presented and described in research articles is one of the major reasons leading to the irreproducibility of published studies. When data is difficult to obtain, the reproducibility problem is exacerbated. A study that examined the differences between clinical and non-clinical scientists, showed that the majority of respondents did not have experience with uploading biomedical data to a repository, stemming from different social reasons not to do so: concerns and motivation around data sharing; work necessary to prepare the data (Federer et al., 2015). Even with current policies mandating data openness (National Institutes of Health, 2015; Wilkinson et al., 2016), authors still fail to include their data alongside their publication, and this can not only be attributed to technical complications, but also fear of being scooped, fear of mistakes being found in data or analyses, and fear of others using their data for their own research papers (Federer et al., 2015; Stodden, 2010; Tenopir et al., 2011, Tenopir et al., 2015). Making FAIR data practices standard, through public data deposition and subsequent publication and citation, could encourage individual researchers and communities to share and reuse data considering their individual requirements and needs (Pawlik et al., 2019). Data accessibility issues are also compounded by data becoming less retrievable with every year passing after the publication (Vines et al., 2014). This is supported by our findings where data is either not available upon publication (57%) or authors cannot be reached/are unresponsive to data provision requests (44%). This continues to be a cultural artifact of using a paper's methods section as a description of steps to reproduce analysis, rather than a fully reproducible solution involving easy access to public data repositories, open-source code, and comprehensive documentation.

As evidenced by the respondents, the lack of data availability is a common hurdle for researchers to encounter that prevents the reproducibility of published work. Thus, the reproducibility of experiments could be improved by increasing the availability of data. Datasets are becoming larger and more complex, especially in genomics. Storage solutions for large data files and citing them within the publication document, especially those in the order of terabytes, can allow for their wider, more efficient and proper data reusability (Faniel and Zimmerman, 2011; Poldrack and Gorgolewski, 2014). Despite the potential advantage, these services can provide for data availability and accessibility, they do not implicitly solve the problem of data *reusability*. This is most apparent when data is too large to be stored locally or transferred *via* slow internet connections, or there is no route to attach metadata that describes the datasets sufficiently for reuse or integration with other datasets. There is also the question of data repository *longevity* - who funds the repositories for decades into the future? Currently, some researchers now have to pay data egress charges for downloading data from cloud providers (Banditwattanawong et al., 2014; Linthicum, 2018). This method presumably saves the data producers money in terms of storing large datasets publicly, but the cost is somewhat now presented to the consumer. This raises complex questions around large data generation projects that also need to be studied extensively for future impact, especially with respect to reproducibility within publications. Moreover, access to the raw data might not be enough, if the steps and other artifacts involved in producing the processed data that was used in the analysis are not provided (Pawlik et al., 2019). In addition, corresponding authors often move on from projects and institutions or the authors themselves can no longer access the data, meaning "data available on request" ceases to be a viable option to source data or explanations of methods. Restricted access to an article can also affect reproducibility by requiring paid subscriptions to read content from a publisher. Although there is precedent for requesting single articles within cross-library loan systems or contacting the corresponding author(s) directly, this, much like requesting access to data, is not without issues. Pre-print servers such as bioRxiv have been taken up rapidly (Abdill and Blekhman, 2019), especially in the genomics and bioinformatics domains, and this has the potential to remove delays in publication whilst simultaneously providing a "line in the sand" with a Digital Object Identifier (DOI) and maintaining the requirements for FAIR data. In some cases, the sensitivity of data might discourage authors from data sharing (Hollis, 2016; Figueiredo, 2017), but this reason was only reported by a small proportion of our respondents. Whilst there are efforts that attempt to apply the FAIR principles to clinical data, such as in the case of the OpenTrials database (Chen and Zhang, 2014), they are by no means ubiquitous.

Data within public repositories with specific deposition requirements (such as the EMBL-EBI European Nucleotide Archive, ebi.ac.uk/ena), might not be associated or annotated with standardised metadata that describes it accurately (Attwood et al., 2009), rather the bare minimum for deposition. Training scientists to implement data management policies effectively is likely to increase data reuse through improved metadata. In a 2016 survey of 3,987 National Science Foundation Directorate of Biological Sciences principal investigators (BIO PIs), expressed their greatest unmet training needs by their institutions (Collins and Tabak, 2014). These were in the areas of integration of multiple data (89%), data management and metadata (78%) and scaling analysis to cloud/high-performance computing (71%). The aforementioned data and computing elements are integral to the correct knowledge "how-to" for research reproducibility. Our findings indicated that those who stated they had experience

in informatics also stated they are better able to attempt and reproduce results. Practical bioinformatics and data management training, rather than in specific tools, may be an effective way of reinforcing the notion that researchers' contributions toward reproducibility are a responsibility that requires active planning and execution. This may be especially effective when considering the training requirements of wet-lab and field scientists, who are becoming increasingly responsible for larger and more complex computational datasets. Further research needs to be undertaken to better understand how researchers' competence in computational reproducibility may be linked to their level of informatics training.

Furthermore, for transparent and reproducible science, both negative (or null) and positive results should be reported for others to examine the evidence (Franco et al., 2014; Prager et al., 2019; Miyakawa, 2020). However, there remains a perception that researchers do not get credit for reproducing the work of others or publishing negative or null results (Franco et al., 2014; Teixeira da Silva, 2015). Whilst some journals explicitly state that they welcome negative results articles (e.g., PLOS ONE "Missing Pieces" collection), this is by no means the norm in life science publishing as evidenced by low, and dropping publication rates of negative findings (Fanelli, 2012; Franco et al., 2014; Teixeira da Silva, 2015). In addition, the perception that mostly positive results are publication-worthy might discourage researchers from providing enough details on their research methodology, such as reporting any negative findings. Ideally, the publication system would enable checking of reproducibility by reviewers and editors at the peer-review stage, with authors providing all data (including raw data), a full description of methods including statistical analysis parameters, any negative findings based on previous work and open source software code (Iqbal et al., 2016). These elements can all be included within the interactive figure, such as by zooming in on over data points to reveal more information on the data, pop-up windows to give details on negative results and parameters and the figure offering re-running of the computational experiment in the case of executable documents. Peer reviewers would then be better able to check for anomalies, and editors could perform the final check to ensure that the scientific paper to be published is presenting true, valid, and reproducible research. Some respondents have suggested that if reviewers and/or editors were monetarily compensated, spending time to reproduce the computational experiments in manuscripts would become more feasible and would aid the irreproducibility issue. However, paying reviewers does not necessarily ensure that they would be more diligent in checking or trying to reproduce results (Hershey, 1992) and there must be optimal ways to ensure effective pressure is placed upon the authors and publishing journals to have better publication standards (Anon, 2013; Pusztai et al., 2013). The increasing adoption by journals of reporting standards for experimental design and results, provide a framework for harmonising the description of scientific processes to enable reproducibility. However, these standards are not universally enforced (Moher, 2018). Similarly, concrete funding within research grants for implementing reproducibility itself manifested as actionable Data Management Plans (Digital Curation Center), rather than what is currently a by-product of the publishing process, could give a level of confidence to researchers who would want to reproduce previous work and incorporate that data in their own projects.

Respondents mentioned that there are word count restrictions in papers, and journals often ask authors to shorten methods sections and perhaps move some text to supplementary information, many times placed in an unorganised fashion or having to remove it altogether. This is a legacy product of the hard-copy publishing era and readability aside; word limits are not consequential for most internet journals. Even so, if the word count limit was only applicable to the introduction, results and discussion sections, then the authors could describe methods in more detail within the paper, without having to move that valuable information in the supplementary section. When methods are citing methodology techniques as described in other papers, where those original references are hard to obtain, typically through closed access practices or by request mechanisms as noted above, then this can be an additional barrier to the reproducibility of the experiment. This suggests that there are benefits to describing the methods in detail and stating that they are similar to certain (cited) references as well as document the laboratory's expertise in a particular method (Moher et al., 2015). However, multi-institutional or consortium papers are becoming more common with ever-increasing numbers of authors on papers, which adds complexity to how authors should describe every previous method available that underpins their research (Gonsalves, 2014). There is no obvious solution to this issue. Highly specialised methods (e.g., electrophysiology expertise, requirements for large computational resources or knowledge of complex bioinformatics algorithms) and specific reagents (e.g., different animal strains), might not be readily available to other research groups (Collins and Tabak, 2014). As stated by some respondents, in certain cases the effective reproducibility of experiments is obstructed by numerical issues with very small or very large matrices or datasets, or different versions of analysis software used, perhaps to address bugs in analytical code, will cause a variation in the reproduced results.

Effects on Technical Developments

Previous studies have provided strong evidence that there is a need for better technical systems and platforms to enable and promote the reproducibility of experiments. We provide additional evidence that paper authors and readers perceive a benefit from having an interactive figure that would allow for the reproducibility of the experiment shown in the figure. An article that gives access to the data, code and detailed data analysis steps would allow for *in situ* reproduction of computational experiments by re-running code including statistical analyses "live" within the paper (Perkel, 2017). Whilst our study did not concentrate on how these "executable papers" may be constructed, this is an active area of development and some examples of how this may be achieved have been provided (Jupyter et al., 2018; Somers, 2018). We provide additional evidence that paper authors and readers perceive a benefit from having publication infrastructure available that would allow for the reproducibility of an experiment. As such, the

findings of this survey helped the development of two prototypes of interactive figures (see Data and Code availability) and subsequently the creation of *eLIFE*'s first computationally reproducible document (Ghosh et al., 2017).

We also asked whether presenting published experiments through interactive figures elements in online publications might be beneficial to researchers, in order to better consume research outputs. Respondents stated they could see the benefit in having interactive figures for the readers of their papers and the papers they read and being able as authors to present their experiment analysis and data as interactive figures. Respondents endorsed articles that include interactive elements, where access to the processed and raw data, metadata, code, and detailed analysis steps, in the form of an interactive figure, would help article readers better understand the paper and the experimental design and methodology. This would, in turn, improve the reproducibility of the experiment presented in the interactive figure, especially computational experiments. The notion of data visualisation tools promoting interactivity and reproducibility in online publishing has also been discussed in the literature (Perkel, 2018). Other efforts have been exploring the availability of interactive figures for driving reproducibility in publishing in the form of executable documents (Ghosh et al., 2017; Ingraham, 2017; Rao et al., 2017; Jupyter et al., 2018). Moreover, technologies such as Jupyter Notebooks, Binder, myExperiment, CodeOcean enable the reproducibility of computational experiments associated with publications, provided by the authors as links from the paper. However, the benefit of having the interactivity and availability of reproducing experiments from within the article itself in the form of interactive figures, is that the reader can stay within the article itself and explore all the details of the data presented in the figure, download the data, play with the code or analysis that produced the figure, interact with parameters in the computational analysis workflows and computationally reproduce the experiments presented in the figure. This can enable the reader to better understand the research done presented in the interactive figure. Despite the self-reported perceived benefits of including interactive figures, the availability of this facility would not affect the respondents' decisions on where to publish. This contradiction suggests that cultural factors (incentives, concerns authors have with sharing their data, attitudes toward open research) (Stodden, 2010; Federer et al., 2015) play an underestimated role in reproducibility.

Despite the benefits, the interactive documents and figures can provide to the publishing system for improved consumption of research outputs, and that those benefits are in demand by the scientific community, work is needed in order to promote and support their use. Given the diversity of biological datasets and ever-evolving methods for data generation and analysis, it is unlikely that a single interactive infrastructure type can support all types of data and analysis. More research into how different types of data can be supported and presented in papers with interactivity needs to be undertaken. Yet problems with data availability and data sizes will persist - many studies comprise datasets that are too large to upload and render within web browsers in a reasonable timescale. Even if the data are available through well-funded repositories with fast data transfers, e.g., the INSDC databases (insdc.org), are publishers ready to bear the extra costs of supporting the infrastructure and people required to develop or maintain such interactive systems in the long run? These are questions that need to be further investigated, particularly when considering any form of industry standardisation of such interactivity in the publishing system. Publishing online journal papers with embedded interactive figures requires alterations to infrastructure, authoring tools and editorial processes (Perkel, 2018). In some cases, the data underpinning the figures might need to be stored and managed by third parties and this means the data, as well as the figures, may not be persistent. The same argument is relevant to software availability and reuse - publishers would need to verify that any links to data and software were available and contained original unmodified datasets. As datasets become larger and more complex, and more software and infrastructure is needed to re-analyse published datasets, this will affect how infrastructure will need to be developed to underpin reproducible research. Incentives will need to be put in place to motivate investment in these efforts.

Effects on Research Policy and Practice

We show that providing tools to scientists who are not computationally aware also requires a change in research culture, as many aspects of computational reproducibility require a change in publishing behavior and competence in the informatics domain. Encouraging and incentivising scientists to conduct robust, transparent, reproducible and replicable research, such as with badges to recognise open practices should be prioritised to help solve the irreproducibility issue (Kidwell et al., 2016). Implementing hiring practices with open science at the core of research roles (Schönbrodt, 2016) will encourage attitudes to change across faculty departments and institutions. In general, as journal articles are still the dominant currency of research in terms of career development, measures of reproducibility and openness may well become more important to hiring institutions when considering candidates rather than publication placement and impact. Indeed, DORA (sfdora.org) now has many signatories, showing that research institutions are taking their role seriously in changing the previous cultural practices of closed "prestigious" science.

We believe that the attitudes highlighted in this survey reflect the growing acceptance of open publishing of code and data, in at least some disciplines. Some publishers are acknowledging that they have a part to play in the improvement of reproducibility through their publishing requirements, e.g., PLOS Computational Biology recently announced that the journal is implementing a "more-rigorous code policy that is intended to increase code sharing on publication of articles" (Cadwallader et al., 2021). Google Scholar now includes a measure of the number of publications in a researcher's profile that meet funder mandates for open access. Whilst, not a perfect system (institutional repositories do not seem to be well covered currently), this shows that even search engines that are heavily in use by researchers to find and consume research outputs are trying to both adapt to cultural changes and automate the presentation of open reproducible science as a goal for researchers. Our survey reflects movements toward open scholarly communications and reproducible academic publishing that are being put into practice.

Further work in this area should include surveys to quantitatively and qualitatively assess how these changes and developments in policy and practice are having an effect on the research culture of reproducibility in the life sciences.

Another potential solution to the reproducibility crisis is to identify quantifiable metrics of research reproducibility and its scientific impact, thus giving researchers a better understanding of how their work stands on a scale of measurable reproducibility. The current assessment of the impact of research articles is a set of quantifiable metrics that do not evaluate research reproducibility, but stakeholders are starting to request that checklists and tools are provided to improve these assessments (Wellcome Trust, 2018). It is harder to find a better approach that is based on a thoroughly informed analysis by unbiased experts in the field that would quantify the reproducibility level of the research article (Flier, 2017). That said, top-down requirements from journals and funders to release reproducible data and code may go some way to improving computational reproducibility within the life sciences, but this will also rely on the availability of technical solutions that are accessible and useful to most scientists.

Opinions are mixed regarding the extent and severity of the reproducibility crisis. Our study and previous studies are highlighting the need to find effective solutions toward solving the reproducibility issue. Steps toward modernising the publishing system by incorporating interactivity with interactive figures and by automatically reproducing computational experiments described within a paper are deemed desirable. This may be a good starting point for improving research reproducibility by reproducing experiments within research articles. This, however, does not come without its caveats, as we described above. From our findings and given the ongoing release of tools and platforms for technical reproducibility, future efforts should be spent in tackling the cultural behavior of scientists, especially when faced with the need to publish for career progression.

DATA AVAILABILITY STATEMENT

All datasets presented in this study can be found below: https://doi.org/10.6084/m9.figshare.c.4436912. Prototypes of interactive figures developed by the corresponding author are available via these GitHub repositories: https://github.com/code56/nodeServerSimpleFig and https://github.com/code56/prototype_article_interactive_figure.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by UEA Computing Sciences Research Ethics Committee Approval reference: CMPREC/1819/R/13. Data were analysed anonymously, however, we did seek and received ethics approval. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

ES and RD contributed to the conception and design of the study. ES conducted the survey, collected the data and organised the database. ES and RD performed the statistical analysis. ES wrote the first draft of the manuscript. Both authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This work was supported by the Biotechnology and Biological Sciences Research Council (BBSRC) through the Core Capability Grant BB/CCG1720/1 at the Earlham Institute. ES was supported by a BBSRC iCASE PhD training grant (project reference: BB/M017176/1) and *eLIFE*. RD was supported by the BBSRC grants BBS/E/T/000PR9817 and BBS/E/T/000PR9783. Open Access fee support was provided by the BBSRC.

ACKNOWLEDGMENTS

We would like to thank all the respondents of the surveys for their time. We would also like to thank George Savva from the Quadram Institute (QIB, United Kingdom) for comments and suggestions for this manuscript; *eLIFE* Sciences Publications Ltd., with whom the corresponding author collaborates as an iCASE student; as well as Ian Mulvany, former *eLIFE* Head of Development, for his help in developing the survey questionnaire.

REFERENCES

Abdill, R. J., and Blekhman, R. (2019). Meta-Research: Tracking the Popularity and Outcomes of All bioRxiv Preprints. *eLife* 8, e45133. doi:10.7554/elife.45133

Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Čech, M., et al. (2018). The Galaxy Platform for Accessible, Reproducible and Collaborative Biomedical Analyses: 2018 Update. *Nucleic Acids Res.* 46 (W1), W537–W544. doi:10.1093/nar/gky379

Amstutz, P., Crusoe, M. R., Tijanić, N., Chapman, B., Chilton, J., and Heuer, M. (2016). Common Workflow Language Specifications, v1.0.2. Retrieved from: www.commonwl.org website: https://w3id.org/cwl/v1.0/.

Anon, J. (2013). Announcement: Reducing Our Irreproducibility. *Nature* 496 (7446), 398. doi:10.1038/496398a

Attwood, T. K., Kell, D. B., McDermott, P., Marsh, J., Pettifer, S. R., and Thorne, D. (2009). Calling International rescue: Knowledge Lost in Literature and Data Landslide!. *Biochem. J.* 424 (3), 317–333. doi:10.1042/bj20091474

Baker, M. (2016). Is There a Reproducibility Crisis? A Nature Survey Lifts the Lid on How Researchers View the Ciris Rocking Science and what They Think Will Help. *Nature* 533 (26), 353–366. doi:10.1038/533452a

Banditwattanawong, T., Masdisornchote, M., and Uthayopas, P. (2014). "Economical and Efficient Big Data Sharing with I-Cloud," in 2014 International Conference on Big Data and Smart Computing (BIGCOMP), Bangkok, Thailand, 15–17 Janurary 2014, (IEEE), 105–110.

Baranyi, M., and Greilhuber, J. (1999). Genome Size inAllium: in Quest of Reproducible Data. *Ann. Bot.* 83 (6), 687–695. doi:10.1006/anbo.1999.0871

Barnes, D. G., and Fluke, C. J. (2008). Incorporating Interactive Three-Dimensional Graphics in Astronomy Research Papers. *New Astron.* 13 (8), 599–605. doi:10.1016/j.newast.2008.03.008

Barnes, D. G., Vidiassov, M., Ruthensteiner, B., Fluke, C. J., Quayle, M. R., and McHenry, C. R. (2013). Embedding and Publishing Interactive, 3-dimensional, Scientific Figures in Portable Document Format (Pdf) Files. *PloS one* 8 (9), e69446. doi:10.1371/journal.pone.0069446

Barone, L., Williams, J., and Micklos, D. (2017). Unmet Needs for Analyzing Biological Big Data: A Survey of 704 NSF Principal Investigators. *PLoS Comput. Biol.* 13 (10), e1005755. doi:10.1371/journal.pcbi.1005755

Begley, C. G., and Ellis, L. M. (2012). Raise Standards for Preclinical Cancer Research. *Nature* 483 (7391), 531–533. doi:10.1038/483531a

Bell, C. J., Dinwiddie, D. L., Miller, N. A., Hateley, S. L., Ganusova, E. E., Mudge, J., et al. (2011). Carrier Testing for Severe Childhood Recessive Diseases by Next-Generation Sequencing. *Sci. translational Med.* 3 (65), 65ra4. doi:10.1126/scitranslmed.3001756

Berg, J. (2018). Progress on Reproducibility. *Science* 359 (6371), 9. doi:10.1126/science.aar8654

Brinckman, A., Chard, K., Gaffney, N., Hategan, M., Jones, M. B., Kowalik, K., et al. (2018). Computing Environments for Reproducibility: Capturing the "Whole Tale". *Future Generation Computer Syst.* 94, 854–867.

Cadwallader, L., Papin, J. A., Mac Gabhann, F., and Kirk, R. (2021). Collaborating with Our Community to Increase Code Sharing. *Plos Comput. Biol.* 17 (3), e1008867. doi:10.1371/journal.pcbi.1008867

Cataldo, M., Mockus, A., Roberts, J. A., and Herbsleb, J. D. (2009). Software Dependencies, Work Dependencies, and Their Impact on Failures. *IIEEE Trans. Softw. Eng.* 35 (6), 864–878. doi:10.1109/tse.2009.42

Chardon, M., and Vandewalle, P. (1991). *Acoustico-lateralis System Cyprinid Fishes.*

Chen, C. P., and Zhang, C.-Y. (2014). Data-intensive Applications, Challenges, Techniques and Technologies: A Survey on Big Data. *Inf. Sci.* 275, 314–347.

Chirigati, F., Rampin, R., Shasha, D., and Freire, J. (2016). "Reprozip: Computational Reproducibility with Ease," in Proceedings of the 2016 International Conference on Management of Data, San Francisco, CA, United States, 26 June 2016–July 1 2016, (ACM (Association for Computing Machinery), 2085–2088. doi:10.1145/2882903.2899401

Christensen, G., and Miguel, E. (2018). Transparency, Reproducibility, and the Credibility of Economics Research. *J. Econ. Lit.* 56 (3), 920–980. doi:10.1257/jel.20171350

Claerbout, J. F., and Karrenbach, M. (1992). "Electronic Documents Give Reproducible Research a New Meaning," in *Seg Technical Program Expanded Abstracts 1992* (Stanford, CA, United States: Standford Exploration Project), 601–604.

Collins, F. S., and Tabak, L. A. (2014). Policy: Nih Plans to Enhance Reproducibility. *Nature* 505 (7485), 612–613. doi:10.1038/505612a

Colomb, J., and Brembs, B. (2014). Sub-strains of drosophila Canton-s Differ Markedly in Their Locomotor Behavior. *F1000Res* 3, 176. doi:10.12688/f1000research.4263.1

Crick, T., Hall, B. A., and Ishtiaq, S. (2017). Reproducibility in Research: Systems, Infrastructure, Culture. *J. Open Res. Softw.* 5. doi:10.5334/jors.73

Crick, T., Hall, B. A., Ishtiaq, S., and Takeda, K. (2014). ""Share and Enjoy": Publishing Useful and Usable Scientific Models," in In 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing, London, UK, 8–11 December 2014, (IEEE), 957–961. doi:10.1109/UCC.2014.156

de Ruiter, P. (2017). The Significance of Reproducible Data. Retrieved from: https://www.labfolder.com/blog/the-significance-of-reproducible-data/ (Accessed September 1, 2019).

Donoho, D. L. (2010). An Invitation to Reproducible Computational Research. *Biostatistics* 11 (3), 385–388. doi:10.1093/biostatistics/kxq028

Drummond, C. (2009). "Replicability Is Not Reproducibility: Nor Is it Good Science," in Proceedings of the Evaluation Methods for Machine Learning Workshop at the 26th ICML., 2009, Montreal, Canada, 14–18 June 2009, (New York, United States: ACM (Association for Computing Machinery)). doi:10.1145/1553374.1553546

Eisner, D. (2018). Reproducibility of Science: Fraud, Impact Factors and Carelessness. *J. Mol. Cell. Cardiol.* 114, 364–368. doi:10.1287/865b2258-d578-4194-9895-d5d65d7a739c

Fanelli, D. (2012). Negative Results Are Disappearing from Most Disciplines and Countries. *Scientometrics* 90 (3), 891–904. doi:10.1007/s11192-011-0494-7

Faniel, I. M., and Zimmerman, A. (2011). Beyond the Data Deluge: A Research Agenda for Large-Scale Data Sharing and Reuse. *Ijdc* 6 (1), 58–69. doi:10.2218/ijdc.v6i1.172

Federer, L. M., Lu, Y.-L., Joubert, D. J., Welsh, J., and Brandys, B. (2015). Biomedical Data Sharing and Reuse: Attitudes and Practices of Clinical and Scientific Research Staff. *PloS one* 10 (6), e0129506. doi:10.1371/journal.pone.0129506

Feger, S. S., Dallmeier-Tiessen, S., Woźniak, P. W., and Schmidt, A. (2019). "Gamification in Science: A Study of Requirements in the Context of Reproducible Research," in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow Scotland, United Kingdom, May 4–9, 2019, (New York, United States: ACM (Association for Computing Machinery)), 1–14.

Figueiredo, A. S. (2017). Data Sharing: Convert Challenges into Opportunities. *Front. Public Health* 5, 327. doi:10.3389/fpubh.2017.00327

Flier, J. S. (2017). Irreproducibility of Published Bioscience Research: Diagnosis, Pathogenesis and Therapy. *Mol. Metab.* 6 (1), 2–9. doi:10.1016/j.molmet.2016.11.006

Franco, A., Malhotra, N., and Simonovits, G. (2014). Publication Bias in the Social Sciences: Unlocking the File Drawer. *Science* 345 (6203), 1502–1505. doi:10.1126/science.1255484

Fraser, H., Parker, T., Nakagawa, S., Barnett, A., and Fidler, F. (2018). Questionable Research Practices in Ecology and Evolution. *PLOS ONE* 13 (7), e0200303, e0200303. doi:10.1371/journal.pone.0200303

Freedman, L. P., Cockburn, I. M., and Simcoe, T. S. (2015). The Economics of Reproducibility in Preclinical Research. *Plos Biol.* 13 (6), e1002165. doi:10.1371/journal.pbio.1002165

Ghosh, S. S., Poline, J.-B., Keator, D. B., Halchenko, Y. O., Thomas, A. G., Kessler, D. A., et al. (2017). A Very Simple, Re-executable Neuroimaging Publication. *F1000Research* 6, 124. doi:10.12688/f1000research.10783.2

Goble, C. A., Bhagat, J., Aleksejevs, S., Cruickshank, D., Michaelides, D., and Newman, D. (2010). myExperiment: a Repository and Social Network for the Sharing of Bioinformatics Workflows. *Nucleic Acids Res.* 38 (Suppl. 2), W677–W682. doi:10.1093/nar/gkq429

Goff, S. A., Vaughn, M., McKay, S., Lyons, E., Stapleton, A. E., Gessler, D., et al. (2011). The iPlant Collaborative: Cyberinfrastructure for Plant Biology. *Front. Plant Sci.* 2, 34. doi:10.3389/fpls.2011.00034

Gonsalves, A. (2014). *Lessons Learned on Consortium-Based Research in Climate Change and Development*. (Ottawa, Canada; London, United Kingdom: International Development Research Centre; UK Aid), 1. Available at: https://idl-bnc-idrc.dspacedirect.org/bitstream/handle/10625/52501/1/IDL-52501.pdf.

Grossman, T., Chevalier, F., and Kazi, R. H. (2016). Bringing Research Articles to Life with Animated Figures. *Interactions* 23 (4), 52–57. doi:10.1145/2949762

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., et al. (2018). Data Availability, Reusability, and Analytic Reproducibility: Evaluating the Impact of a Mandatory Open Data Policy at the Journal Cognition. *R. Soc. Open Sci.* 5 (8), 180448. doi:10.1098/rsos.180448

Hershey, N. (1992). Compensation and Accountability: The Way to Improve Peer Review. *Qual. Assur. Utilization Rev.* 7 (1), 23–29. doi:10.1177/106286069200700104

Higginson, A. D., and Munafò, M. R. (2016). Current Incentives for Scientists lead to Underpowered Studies with Erroneous Conclusions. *PLoS Biol.* 14 (11), e2000995. doi:10.1371/journal.pbio.2000995

Hollis, K. F. (2016). To Share or Not to Share: Ethical Acquisition and Use of Medical Data. *AMIA Summits Translational Sci. Proc.* 2016, 420.

Hull, D., Wolstencroft, K., Stevens, R., Goble, C., Pocock, M. R., Li, P., et al. (2006). Taverna: a Tool for Building and Running Workflows of Services. *Nucleic Acids Res.* 34 (Suppl. 2), W729–W732. doi:10.1093/nar/gkl320

Ingraham, T. (2017). *Reanalyse(a)s: Making Reproducibility Easier with Code Ocean Widgets*. Retrieved from. doi:10.2514/6.2017-0456https://blog.f1000.com/2017/04/20/reanaly-seas-making-reproducibility-easier-with-code-ocean-widgets/(Accessed September 6, 2019).

Ingraham, T. (2017). Researchers Can Now Publish Interactive Plotly Figures in F1000. Plotly. Retrieved from: https://medium.com/plotly/researchers-can-

now-publish-interactive-plotly-figures-in-f1000-87827a1b5d94 (Accessed August 7, 2019).

Ioannidis, J., and Doucouliagos, C. (2013). What's to Know about the Credibility of Empirical Economics? *J. Econ. Surv.* 27 (5), 997–1004. doi:10.1111/joes.12032

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *Plos Med.* 2 (8), e124. doi:10.1371/journal.pmed.0020124

Ioannidis, J. P., Stanley, T. D., and Doucouliagos, H. (2017). *The Power of Bias in Economics Research*. Oxford, UK: Oxford University Press.

Iqbal, S. A., Wallach, J. D., Khoury, M. J., Schully, S. D., and Ioannidis, J. P. (2016). Reproducible Research Practices and Transparency across the Biomedical Literature. *PLoS Biol.* 14 (1), e1002333. doi:10.1371/journal.pbio.1002333

James, J. M., and Bolstein, R. (1990). The Effect of Monetary Incentives and Follow-Up Mailings on the Response Rate and Response Quality in Mail Surveys. *Public Opin. Q.* 54 (3), 346–361. doi:10.1086/269211

Jobber, D., Saunders, J., and Mitchell, V.-W. (2004). Prepaid Monetary Incentive Effects on Mail Survey Response. *J. Business Res.* 57 (1), 21–25. doi:10.1016/s0148-2963(02)00280-1

Jupyter, P., Bussonnier, M., Forde, J., Freeman, J., Granger, B., and Head, T. (2018). "Binder 2.0-reproducible, Interactive, Sharable Environments for Science at Scale," in Proceedings of the 17th python in science conference, Austin, TX, United States, July 9–July 15, 2018, 113, 120. doi:10.25080/majora-4af1f417-011

Kidwell, M. C., Lazarevic, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., and Falkenberg, L.-S. and . . . others (2016). Badges to Acknowledge Open Practices: A Simple, Low-Cost, Effective Method for Increasing Transparency. *PLoS Biol.* 14 (5), e1002456. doi:10.1371/journal.pbio.1002456

Kim, Y. M., Poline, J. B., and Dumas, G. (2018). Experimenting with reproducibility: a case study of robustness in bioinformatics. *GigaScience* 7 (7), giy077. doi:10.1093/gigascience/giy077

Koschke, R. (2003). Software Visualization in Software Maintenance, Reverse Engineering, and Re-engineering: a Research Survey. *J. Softw. Maint. Evol. Res. Pract.* 15 (2), 87–109. doi:10.1002/smr.270

Kurtzer, G. M., Sochat, V., and Bauer, M. W. (2017). Singularity: Scientific Containers for Mobility of Computer. *PloS one* 12 (5), e0177459. doi:10.1371/journal.pone.0177459

Leipzig, J., Nust, D., Hoyt, C. T., Soiland-Reyes, S., Ram, K., and Greenberg, J. (2020). *The Role of Metadata in Reproducible Computational Research*. preprint arXiv:2006.08589.

LeVeque, R. J., Mitchell, I. M., and Stodden, V. (2012). Reproducible Research for Scientific Computing: Tools and Strategies for Changing the Culture. *Comput. Sci. Eng.* 14 (4), 13–17. doi:10.1109/mcse.2012.38

Liberman, M. (2015). Replicability vs. Reproducibility — or Is it the Other Way Around? Retrieved from: https://languagelog.ldc.upenn.edu/nll/?p=21956 (Accessed October 5, 2019).

Linthicum, D. (2018). Don't Get Surprised by the Cloud's Data-Egress Fees. Retrieved from: https://www.infoworld.com/article/3266676/dont-get-surprised-by-the-clouds-data-egress-fees.html (Accessed September 6, 2019).

Maciocci, G., Aufreiter, M., and Bentley, N. (2019). *Introducing* eLIFE's *First Computationally Reproducible Article*, Cambridge, United Kingdom: eLife Labs [Internet].

Marx, V. (2013). The Big Challenges of Big Data. *Nature* 498 (7453), 255–260. doi:10.1038/498255a

Miyakawa, T. (2020). *No Raw Data, No Science: Another Possible Source of the Reproducibility Crisis*. BioMed Central. doi:10.1201/9780367811044

Moher, D., Avey, M., Antes, G., and Altman, D. G. (2015). The National Institutes of Health and Guidance for Reporting Preclinical Research. *BMC Med.* 13 (1), 1–4. doi:10.1186/s12916-015-0321-8

Moher, D. (2018). Reporting Guidelines: Doing Better for Readers. *BMC Med.* 16 (1). doi:10.1186/s12916-018-1226-0

Muller, H., Naumann, F., and Freytag, J.-C. (2003). *Data Quality in Genome Databases*.

Munafo, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Du Sert, N. P., et al. (2017). A Manifesto for Reproducible Science. *Nat. Hum. Behav.* 1 (1), 1–9. doi:10.1038/s41562-016-0021

National Institutes of Health (2015). Plan for Increasing Access to Scientific Publications and Digital Scientific Data from NIH Funded Scientific Research. Retrieved from https://grants.nih.gov/grants/NIH-Public-Access-Plan.pdf.

Newe, A. (2016). Enriching Scientific Publications with Interactive 3D Pdf: an Integrated Toolbox for Creating Ready-To-Publish Figures. *PeerJ Computer Sci.* 2, e64. doi:10.7717/peerj-cs.64

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., et al. (2015). Promoting an Open Research Culture. *Science* 348 (6242), 1422–1425. doi:10.1126/science.aab2374

Nuijten, M. B., Bakker, M., Maassen, E., and Wicherts, J. M. (2018). Verify Original Results through Reanalysis before Replicating. *Behav. Brain Sci.* 41 (e143). doi:10.1017/s0140525x18000791

O'Brien, B. C., Harris, I. B., Beckman, T. J., Reed, D. A., and Cook, D. A. (2014). Standards for Reporting Qualitative Research: a Synthesis of Recommendations. *Acad. Med.* 89 (9), 1245–1251. doi:10.1097/ACM.0000000000000388

Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., et al. (2004). Taverna: a Tool for the Composition and Enactment of Bioinformatics Workflows. *Bioinformatics* 20 (17), 3045–3054. doi:10.1093/bioinformatics/bth361

Open Science Collaboration (2015). PSYCHOLOGY. Estimating the Reproducibility of Psychological Science. *Science* 349 (6251), aac4716. doi:10.1126/science.aac4716

Pawlik, M., Hütter, T., Kocher, D., Mann, W., and Augsten, N. (2019). A Link Is Not Enough - Reproducibility of Data. *Datenbank Spektrum* 19 (2), 107–115. doi:10.1007/s13222-019-00317-8

Peng, R. D., Dominici, F., and Zeger, S. L. (2006). Reproducible Epidemiologic Research. *Am. J. Epidemiol.* 163 (9), 783–789. doi:10.1093/aje/kwj093

Peng, R. D. (2011). Reproducible Research in Computational Science. *Science* 334 (6060), 1226–1227. doi:10.1126/science.1213847

Perkel, J. M. (2018). Data Visualization Tools Drive Interactivity and Reproducibility in Online Publishing. *Nature* 554 (7690), 133–134. doi:10.1038/d41586-018-01322-9

Perkel, J. (2017). TechBlog: Interactive Figures Address Data Reproducibility: Naturejobs Blog. Retrieved from: Accessed http://blogs.nature.com/naturejobs/2017/10/20/techblog-interactive-figures-address-data-reproducibilityAugust 6, 2019).

Pettifer, S. R., Sinnott, J. R., and Attwood, T. K. (2004). UTOPIA-User-Friendly Tools for Operating Informatics Applications. *Comp. Funct. Genomics* 5 (1), 56–60. doi:10.1002/cfg.359

Pettifer, S., Thorne, D., McDermott, P., Marsh, J., Villeger, A., Kell, D. B., et al. (2009). Visualising Biological Data: a Semantic Approach to Tool and Database Integration. *Bmc bioinformatics* 10, 1–12. doi:10.1186/1471-2105-10-s6-s19

Plesser, H. E. (2018). Reproducibility vs. Replicability: a Brief History of a Confused Terminology. *Front. neuroinformatics* 11, 76. doi:10.3389/fninf.2017.00076

Poldrack, R. A., and Gorgolewski, K. J. (2014). Making Big Data Open: Data Sharing in Neuroimaging. *Nat. Neurosci.* 17 (11), 1510–1517. doi:10.1038/nn.3818

Prager, E. M., Chambers, K. E., Plotkin, J. L., McArthur, D. L., Bandrowski, A. E., Bansal, N., et al. (2019). *Improving Transparency and Scientific Rigor in Academic Publishing*. Wiley Online Library.

Pulverer, B. (2015). Reproducibility Blues. *Embo J.* 34 (22), 2721–2724. doi:10.15252/embj.201570090

Pusztai, L., Hatzis, C., and Andre, F. (2013). Reproducibility of Research and Preclinical Validation: Problems and Solutions. *Nat. Rev. Clin. Oncol.* 10 (12), 720–724. doi:10.1038/nrclinonc.2013.171

Rao, S. S. P., Huang, S.-C., Glenn St Hilaire, B., Engreitz, J. M., Perez, E. M., Kieffer-Kwon, K.-R., et al. (2017). Cohesin Loss Eliminates All Loop Domains. *Cell* 171 (2), 305–320. doi:10.1016/j.cell.2017.09.026

Robinson, J. T., Turner, D., Durand, N. C., Thorvaldsdóttir, H., Mesirov, J. P., and Aiden, E. L. (2018). Juicebox.js Provides a Cloud-Based Visualization System for Hi-C Data. *Cel Syst.* 6 (2), 256–258. doi:10.1016/j.cels.2018.01.001

Schneider, M. V., Flannery, M., and Griffin, P. (2016). Survey of Bioinformatics and Computational Needs in Australia 2016. Figshare. Available from: https://figshare.com/articles/Survey_of_Bioinformatics_and_Computational_Needs_in_Australi a_2016_pdf/4307768/1.

Schonbrodt, F. (2016). Changing Hiring Practices towards Research Transparency: The First Open Science Statement in a Professorship Advertisement. Felix Schonbrodt Blog. Available from: https://www.nicebread.de/open-science-hiring-practices.

Shettle, C., and Mooney, G. (1999). Monetary Incentives in Us Government Surveys. *J. Official Stat.* 15 (2), 231.

Sneddon, T. P., Li, P., and Edmunds, S. C. (2012). GigaDB: Announcing the GigaScience Database. *GigaScience* 1 (1), 2047-217X-221X-11. doi:10.1186/2047-217x-1-11

Snell, L., and Spencer, J. (2005). Reviewers' Perceptions of the Peer Review Process for a Medical Education Journal. *Med. Educ.* 39 (1), 90–97. doi:10.1111/j.1365-2929.2004.02026.x

Somers, S. (2018). The Scientific Paper Is Obsolete. Retrieved from: https://www.theatlantic.com/science/archive/2018/04/the-scientific-paper-is-obsolete/556676/(Accessed September 1, 2019).

Stodden, V., Borwein, J., Bailey, D., Malm, H., Novitzky, P., and Gordijn, B. (2013a). Setting the Default to Reproducible: Reproducibility in Computational and Experimental Mathematics. *Comput. Sci. ResearchSIAM News* 46 (1).

Stodden, V., Guo, P., and Ma, Z. (2013b). Toward Reproducible Computational Research: an Empirical Analysis of Data and Code Policy Adoption by Journals. *PloS ONE* 8 (6), e67111. doi:10.1371/journal.pone.0067111

Stodden, V., McNutt, M., Bailey, D. H., Deelman, E., Gil, Y., Hanson, B., et al. (2016). Enhancing Reproducibility for Computational Methods. *Science* 354 (6317), 1240–1241. doi:10.1126/science.aah6168

Stodden, V., and Miguez, S. (2013). *Best Practices for Computational Science: Software Infrastructure and Environments for Reproducible and Extensible Research*.Available at SSRN 2322276

Stodden, V. (2015). Reproducing Statistical Results. *Annu. Rev. Stat. Its Appl.* 2, 1–19. doi:10.1145/2807591.2897788

Stodden, V., Seiler, J., and Ma, Z. (2018). An Empirical Analysis of Journal Policy Effectiveness for Computational Reproducibility. *Proc. Natl. Acad. Sci. USA* 115 (11), 2584–2589. doi:10.1073/pnas.1708290115

Stodden, V. (2010). "The Scientific Method in Practice: Reproducibility in the Computational Sciences,". Paper No. 4773-10 (MIT Sloan Research).

Tait, A. (2017). 10 Rules for Creating Reproducible Results in Data Science - Dataconomy. Retrieved from: https://dataconomy.com/2017/07/10-rules-results-data-science (Accessed August 1, 2019).

Tang, B., Li, F., Li, J., Zhao, W., and Zhang, Z. (2018). Delta: a new web-based 3D genome visualization and analysis platform. *Bioinformatics* 34 (8), 1409–1410.

Teixeira da Silva, J. A. (2015). Negative Results: Negative Perceptions Limit Their Potential for Increasing Reproducibility. *J. negative results Biomed.* 14 (1), 1–4. doi:10.1186/s12952-015-0033-9

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., and Frame, M. (2011). Data Sharing by Scientists: Practices and Perceptions. *PloS one* 6 (6), e21101. doi:10.1371/journal.pone.0021101

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., and Dorsett, K. (2015). Changes in Data Sharing and Data Reuse Practices and Perceptions Among Scientists Worldwide. *PloS one* 10 (8), e0134826. doi:10.1371/journal.pone.0134826

Vines, T. H., Albert, A. Y. K., Andrew, R. L., Débarre, F., Bock, D. G., Franklin, M. T., et al. (2014). The Availability of Research Data Declines Rapidly with Article Age. *Curr. Biol.* 24 (1), 94–97. doi:10.1016/j.cub.2013.11.014

Weinländer, M., Lekovic, V., Spadijer-Gostovic, S., Milicic, B., Krennmair, G., and Plenk Jr, H., Jr (2009). Gingivomorphometry - Esthetic Evaluation of the crown-mucogingival Complex: a New Method for Collection and Measurement of Standardized and Reproducible Data in Oral Photography. *Clin. Oral Implants Res.* 20 (5), 526–530. doi:10.1111/j.1600-0501.2008.01685.x

Weissgerber, T. L., Garovic, V. D., Savic, M., Winham, S. J., and Milic, N. M. (2016). From Static to Interactive: Transforming Data Visualization to Improve Transparency. *PLoS Biol.* 14 (6), e1002484. doi:10.1371/journal.pbio.1002484

Wellcome Trust (2018). Request for Information (RFI) A Software Tool to Assess the FAIRness of Research Outputs against a Structured Checklist of Requirements [FAIRWare]. In Wellcome.Org. Retrieved from: https://wellcome.ac.uk/sites/default/files/FAIR-checking-software-request-for-information.pdf.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., and Baak, A. and . . . others (2016). The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific data* 3 (1), 1–9. doi:10.1038/sdata.2016.18

Wolstencroft, K., Haines, R., Fellows, D., Williams, A., Withers, D., Owen, S., et al. (2013). The Taverna Workflow Suite: Designing and Executing Workflows of Web Services on the Desktop, Web or in the Cloud. *Nucleic Acids Res.* 41 (W1), W557–W561. doi:10.1093/nar/gkt328

Yong, E. (2015). Reproducibility Problems in Genetics Research May Be Jeopardizing Lives. Retrieved from https://geneticliteracyproject.org/2015/12/17/reproducibility-problems-genetics-research-may-costing-lives.(Accesssed February 12, 2015).

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

A.2 Survey Questions and Answer Options

Knowledge and attitudes among life scientists towards reproducibility within journal articles. Survey Questions and answer options. Developed by Evanthia Samota, Dr Robert P. Davey. Earlham Institute, University of East Anglia, United Kingdom.

1. How often do you encounter difficulties with working with bioinformatic analysis tools (that are not your own)? (Problems such as: installing, configuring, running the software, and working with command line software).
    - Never
    - Rarely
    - Sometimes
    - Often
    - Very often
    - N/A

2. How difficult is it to source (or access) the data presented in published papers?
    - Very difficult
    - Difficult
    - Neither difficult nor difficult
    - Easy
    - Very easy

- I do not seek to source or access the data presented in published papers

3. What difficulties have you encountered in accessing the data described in published papers? Select all that apply to you.
    - Privacy reason (patients' medical data)
    - Commercial sensitivity around the data (e.g. pharmaceutical companies' data that could lead to the production of a drug)
    - Data not available at publication
    - Authors cannot be reached or are unresponsive to data provision requests
    - Data is too large to be transferable
    - N/A

4. How are you currently sourcing the data (if applicable)? Select all that apply to you.
    - Data is readily available through a link from the paper
    - Data is available in a public database
    - I contact the author(s)
    - N/A
    - Other (please specify): *free response*

5. What is your current understanding of the reproducibility of experiments? *Please select any that apply. Should you wish to add any additional information, please add it to the "Other" box (only available in the eLIFE survey).*

- If the experiment was performed by another laboratory, the same or similar results are produced
- Using similar materials, reagents, and methods, reaching the same conclusions
- Running and analysing similar data with the same workflow and getting similar results
- The original authors or others running the same data with precisely the same workflow and getting the same results
- Other (please specify) as an open comment box.

6. Have you ever tried reproducing any published results? Please select the answer that applies best to you (only available in the eLIFE survey).

   - Yes, I have tried reproducing published results, and I have been successful in producing exact or similar results
   - Yes, I have tried reproducing published results, but I have been unsuccessful in producing any results or similar results
   - No, I have never tried reproducing any published results
   - It is not important to reproduce other people' s published results.

7. In your opinion, what could be done to make published research more reproducible? *Select all that apply to you (only available in the eLIFE survey).*

   - The authors provide the raw data and a link to it through the paper

- The authors provide the source code of any custom software used to analyse the data and the code of the software is well documented
- The authors describe all methodology steps in detail, including any formulae analysing the data
- Other (please specify)

8. When thinking about interactive figures, what comes to your mind? *Please describe what you understand of what an interactive figure to be, its features, and where you have seen such a feature before, if applicable.*

    - Free response

9. An interactive figure is a figure within a paper that is dynamic and becomes "live" when the user interacts with it and where the data displayed changes according to various parameter options. Which of the following features of an interactive figure tool would be good to have? *Please rank them in the order of preference, where 1 is the most preferred feature and 11 is the least preferred feature.*

    - Display the altered figure (after having chosen different parameters) with different colours, fonts etc.
    - Display arrows that point out any subtle changes in the figure (after changing some parameters)
    - Easy to manipulate (play around with)
    - Easy to define parameters (that would change the image of the figure)
    - Have many parameters to select from a "how to use the tool" link, pop-up window

- A forum/comment section for readers and authors to discuss their conclusions from interacting with the figure
- A symbol that would link to the raw data, metadata and supplementary data related to the figure
- Import data option
- Export data option
- Link to the source code that builds the interactive figure

10. What other features an interactive figure could have that were not mentioned in the previous question?
    - Free response

11. Do you perceive a benefit in being able to publish *interactive figures*? Please answer the sub-questions below.
    - Do you see a benefit to yourself in publishing Interactive figures?
    - Do you see a benefit to the readers of your papers?
    - Do you see a benefit to yourself in having Interactive figures available in papers that you read?
      - Definitely, not
      - Somewhat not, neither yes nor no, somewhat yes, definitely yes

12. Does the provision or option of an interactive figure in the paper affect your decision in choosing the publishing journal or publisher?
    - Negatively

- Less likely
- Just as likely
- Likely
- Positively

13. Have you heard of the term computationally reproducible data, and do you understand what the term means? *If answered yes or unsure, please explain what you understand from the term.*
    - Yes
    - No
    - Unsure

14. Would you benefit from being able to automatically reproduce computational experiments, or other analyses (including statistical tests) described within a paper?
    - Reproduce the computational experiments (i.e. follow the workflow of the authors)
    - Reproduce the statistical analyses
    - Be able to interact and manipulate parameters and options within the computational analysis workflow
        - Definitely, not
        - Somewhat not
        - Neither yes nor no
        - Somewhat yes
        - Definitely yes

15. How often do you work with *bioinformatic analysis tools* (e.g. assemblers, aligners, structure modelling)?

    - Often
    - Rarely
    - Never
    - Free Response, if it applies: *Please specify the nature of the tools you use (e.g. assemblers, RNA-Seq analysis tools, stochastic modelling tools).*

16. Have you received any of the following training? Training whether formal or informal (training through a colleague etc.).

    - Bioinformatics
    - Computer Science
    - Statistical
    - N/A

17. Which of the following type(s) of data do you work with? *Select all that apply to you.*

    - Biochemistry
    - Biophysics
    - Computational biology (e.g. stochastic modelling; machine learning)
    - Epigenetics
    - Genomics
    - Immunology
    - Population/epidemiological

- Proteomics
- None
- Other (please specify)

A.3 Free Text Answers to Survey Questions

Knowledge and attitudes among life scientists towards reproducibility within journal articles. Survey Questions. Developed by Evanthia Samota, Dr Robert P. Davey. Earlham Institute, University of East Anglia, United Kingdom.

Free text responses from question 5: What is your current understanding of reproducibility of experiments?

- Raw materials (testing different mice, testing different DNA samples from humans) with the same methodology and coming to the same conclusion.
- Most of the time you get similar results, unless the original paper obviously lacked controls or was in a [perceived low quality] journal.
- There is technical reproducibility and biological variability to define. At least three biological replicates of the whole protocol is required to derive statistical significance.
- 1. In my opinion, the reproducibility of an experiment starts by being able to get the same results in a lab by exactly repeating the same method and using the same materials, etc; but it should also be possible to reproduce it in a different laboratory even when small changes are introduced in terms of reagents, materials or

equipment. 2. ”The original authors or others running the same data getting the same results” as a statement that describes experiment reproducibility because although I would expect that option to be valid for a reproducible experiment, I do not think it is conclusive by itself, not only the analysis of the data should be reproducible.

Free text responses from question 7: In your opinion, what could be done to make published research more reproducible?

- Ease getting funding to reproduce published results and opportunity for publication.
- Raw data may be in the form of large files in a proprietary format (e.g. data from mass spectrometers) and can require significant pre-processing before they are reported in a tabularised format. For a variety of reasons, I think it’s impractical to require links to true raw data.
- Strict word counts in publications limit the amount of detail in methods sections. Sometimes moving methods to “supplement” can help with this. But in general I think people don’t want to waste discussion and results space with over describing methods. Word counts should count only for Intro-Results-Discussion, and authors should be encouraged to put in as much detail as possible in the methods section.
- A main problem for open sharing of data is storage space! Who will pay for it when you are dealing with very large amounts of data? E.g. terabytes.
- Authors must publish result with significant intrinsic reproducibility

(n sample high enough and independent experiments).

- Authors provide better descriptions of what the data is. For example, when depositing RNA-seq samples in NCBI they described in detail what each sample is and this description matches to the samples mentioned in the manuscript. This is frequently not the case!
- Methods are usually too brief in papers e.g. experiments performed as per (8), which in turns references further back. Experiments change, but methods sections do not. We've had journals tell us to make our methods briefer.
- Source code should definitely be available. Whether one can understand how it was written or how to use it is another matter though.
- Authors provide data analysis files and output.
- I have had some successful and some unsuccessful attempts to reproduce published results in plant biology. I think part of the "problem" is that we are investigating fine details of very complex living systems, and I suspect that seemingly minor differences in environmental conditions between experiments can have big impacts on some of the components of the system, even if there are no obvious visible effects at the whole plant level.
- Due to word limits, methods may need to be in supplemental materials, but still should be presented in detail. Methods citing a prior paper can be really aggravating as sometimes in my field I find the prior paper was done on a different animal (slice electrophysiology in mice vs rats) or has other differences that make it unclear if methods were adjusted for changes in preparation or not. Additionally, sometimes the original references

can be hard to obtain, especially when the cited paper cites another paper in its methods and so forth. It would be more useful to describe methods in detail, and then say these methods are similar to the methods in the following references, to document the lab's expertise in a particular method.

- Depending on the type of analysis, a more detailed description of the methodology might help. But peer review is key IMAO.
- Authors only publish results when they are confident that anyone following their methods will get the same results: like a good cook book.•Generally, problems with reproducibility arise from highly specialized methods and reagents that are not readily available. If only one lab has access to cell lines, antibodies, or the electrophysiology expertise (for instance) to do the experiments, it can become very burdensome and difficult to independently reproduce results. Ideally, key evidence for a proposed model from a paper can be validated independently through the use of widely available and standardised techniques, but this is the exception rather than the rule. More importantly, as a separate cultural issue, the pressures for graduate students, postdocs and faculty to churn out large quantities of "high impact" papers also creates awful incentives to take ill-advised shortcuts or outright fabricate and obfuscate. Simply put: the competition for limited jobs creates a cut-throat environment that is toxic to intellectual honesty and academic integrity. This problem has no easy solution and will continue to get worse.
- In some cases, computer hardware might have an influence. For example, numerical issues with very small or very large numbers will vary.
- I think raw images would be especially helpful when reproducing co-

localization with fluorescence microscopy.

- A small percentage of funding is set aside for testing the reproducibility of data, to be awarded to grantees who need to know that the data are reliable before going into their own project.
- Provide more specific guidelines about how to describe procedures. Providing code gives a technical description, but we also need a less technical summary of the intention of each step in the analysis so that we can judge whether the operations performed actually did what they were supposed to. Having some recommended structure for how analysis should be described would be a way to potentially achieve both of these aims.

Free text responses from question 8: When thinking about interactive figures, what comes to your mind? (please describe what you understand of what an interactive figure is, its features, and where you have seen such a feature before, if applicable).

***eLIFE* authors**

- Include plots with individual data points and the ability to have the option to add additional data coming from the supplementary data of the paper.
- The user can change the information being plotted or mouse over specific elements to gain more information.
- Interactive is a figure in which you can either zoom in on parts of an image to get additional information about the data and/or to change the appearance of a figure by changing parameters.

- Figure, which can be used to show different parameters and options.
- Options to set different axis and scales and choose which data to display together. A pop-up or navigation through figure segments to better see the variation in "noisy" or dense data.
- Where one can click to expand certain aspects of the figure.
- Allows you to select subsets and zoom in and out additional info appearing when selecting specific points, for example.
- Where you can change the figure when viewing it.
- Data values used to make graphs and plots should be available so new analyses can be performed, or data can be combined.
- A figure with links to the original data?
- I would think about a figure in which by "clicking" (or something similar) one could get raw data (like raw scan image of a gel or a table with all the data from an experiment).
- Perhaps a figure that allows to zoom in particular aspects of it and give information on such detail. In general, a figure that personalises the information given based on the reader's interest.
- A figure that can be accessed dynamically going through the data and the way they were analysed, with the possibility of accessing raw data and re-analyse them in alternative ways.
- Interactive figures should allow you to navigate to the raw data that lies behind any data point in a plot.
- GIFs of animations or slideshows that point to where readers should focus on step by step.
- Clicking on a figure brings up more data.

- Figure with a feature that allows additional layer of information superimposed on the image (upon mouse hover, e.g.).
- I don't really know. Perhaps hovering over a graph and having the actual values appear?
- Interactive figures mean access to the source data to include/exclude the desired information by the user.
- Links to primary data.
- Move mouse over figure and pop-up boxes appear to highlight specific information.
- I think it is a figure with which the reader can interact; maybe there are links in the figure that can bring you to specific parts of the paper or provide you with further information about the data.
- Interactive figure = image/plot can be manipulated by the reader to change some variables/parameters (e.g. threshold value).
- Ability to change the display (e.g. of the axes); click on bars or data points and have numerical data pop up; ability to click on numerical data and see the underlying raw data; click on axes for a longer explanation of what is shown, including methodology.
- A figure in an electronic journal or book allows the reader to manipulate that figure and gain a better understanding of the trend or concept defined by the figure.
- It would be a figure where you can select or deselect either cases/subjects or conditions.
- Structures that have immediate zoom in-out-rotate functions or plots that you can play with the axes.

- I have never seen interactive figures in an academic paper, but I can imagine being able to remove groups from plots to more easily compare relevant data. Or, in a qPCR experiment, clicking on a bar graph might load a spreadsheet document with raw data, primers use and analysis. I have seen inappropriate primers used in qRT-PCR figures in high-profile journals. These things are seldom checked by reviewers who are (correctly) more interested in checking the validity of the reasoning and approaches in the overall manuscript.
- That would be great to interactively increase different parts of the figures (plots, curves…etc.) and the possibility to pick up data points and see raw data behind them.
- Figure that gives all information to reproduce the data and understand what the result means.
- The possibility to go through the different parts of a 3D figure (a brain with activations for example). Also, the possibility to click on the figure and access the data that have produced the result. I do not remember where I have seen it.
- An interactive figure is a figure that is dynamic, and the reader interacts with it.
- For structures, 3d representation.
- A figure where one can check boxes to include/exclude certain reps/results with the plots/stats updated in real-time.
- An interactive figure allows users to set certain parameters (e.g. a range of years, particular countries/states/districts) and view the results that are of interest to them. These figures might be either static or dynamic. Our group has produced a website that allows

this for our results, and this has increased engagement with the media, who have embedded our dynamic visualisations in some of their stories.

- Images that can be adjusted or overlaid on the web page, or graphs that can be rearranged or provide more data when clicked on. I don't think I have encountered an interactive figure apart from movies.
- An interactive figure is something that you will be able to change. Most usefully, the interactive figure will let you see and work on the raw data on which the figure is based.
- Someone can pick an area of the figure getting extra information.
- A figure that can be modified directly in the online version of the publication using the data provided by the authors.
- Figure with direct links to methodology and raw and supplementary data.
- When/where applicable (a proposed model, or large dataset with outliers), the ability to assess the effect of input parameters and inclusion of outliers into the model behaviour.
- An interactive figure is a figure within a paper that is dynamic and becomes live when the user interacts with it and where the data displayed changes according to various parameter options.
- An interactive figure should provide more detailed information about the different figure parts if needed. It should give the reader the possibility to visualize the raw data submitted by the authors in an unbiased but still easy-to-understand manner.

Which software or platform to use to display the interactive figure

- Shiny in R.
- Stuff like Shiny in R where you can play around with a limited set of parameters. It's an interesting toy, but ultimately a real examination of the data is going to need access to the full raw data to put it into a more powerful environment.

Responses of the NBI participants

No Opinion

- n/a
- n/a
- I don' t know what is an Interactive Figure.
- I don't know what that is.
- I have never made nor used one.
- No idea.
- I have never come across an interactive figure and I'm not sure what that even means.
- I have no prior knowledge concerning interactive figures.
- I have no idea what this is or how it relates to experimental variability.
- Did not see any in the publications I am looking at.
- I have never seen such a thing.
- I don't need them.

Data and parameters related

- Being able to access the data underlying the figure.
- I have not come across this before, but being able to click on figures to gain more data comes to mind.
- It is not quite usual to have such figures in biology, but to me, it is mainly associated with "clickable models" such as this one: http://labs.biology.ucsd.edu/schroeder/clickablegc/pages/closure.html
- I have not seen such a figure. I imagine it to be a figure where you can select the different panels to obtain detailed figure legends and links to the relevant text section or other figures as well as to the original data.
- I think of interactive figures mainly in the context of online news articles, where additional features of a dataset become visible when interacting with them (e.g. given a mouse-over event).
- Roll-over information and clickable element.
- Figures whose format can be modified by readers and whose raw data can be downloaded by the readers.
- A graph where when you hover on a point, the raw value is displayed. A chromatogram where if you click on a peak, the underlying spectral data might be displayed. A metabolic pathway where when you click on an arrow, data about the enzyme or metabolite are displayed, or further details of measurements on that piece of metabolism are displayed. A result of a mathematical model, where there might be sliders and text-boxes where parameters can be specified whose results are then displayed in the figure; this is often done in learning tools/interactive experiments in websites.

- Figures you can click on to get more details.
- Only ever seen molecular structures that can be moved around an axis to see the structure better. Not sure what else might be wanted, except perhaps a model of a system. However, being able to pick out a particular gene in a big dataset (e.g. from expression cluster in RNA seq data) by just typing in a would be the only thing I could imagine really wanting.
- See how figures change if you alter the parameters that were made to produce them, or by leaving certain data out.
- I understand there is a piece of code and piece of data "behind" the figure, which I can see, and I can adjust the data (parameters, inputs) to see the difference rendered in the output. Ideally, I would also be able to play with the code itself.
- Figures that allow user input to access more details.
- Changing axis, datasets, plotted functions, dimensionality, rotations, colours, scales etc.
- For me an interactive figure is a figure that automatically changes when you change an option, the data or anything. You don't have to reload it every time.
- Possibility to plot the same data in different ways.
- In the context of this survey, I think that a useful interactive figure would be one that could be linked online to the raw dataset that it represents. I have never come across this type of figure.
- I'd guess an online figure, where the data can be viewed interactively. I have not seen one before.
- Javascript, plots in which parameters can be changed and the view of the graph changes.

- I have not used interactive figures, but I would imagine a Circos-style plot that presents data at each level at which the user zooms into the figure. This would require re-plotting sliding windows to increase resolution.
- An interactive figure, for me, is a way of presenting huge datasets by exposing dynamically only a user-requested fraction. An in-house example could very well be the ExpVip website by Ricardo Ramirez, or related the sleuth interface to kallisto.
- Point at a graph read off a value.
- A dynamic figure where the user can select different parameters and the figure changes depending on what you want to see.

Which software or platform to use to display the interactive figure

- D3.js, An interactive I think allows you to click on it with a mouse, pan around (either to see other views of a 3D surface, or to focus more on a certain section), zoom in/out, subset. These can be created in D3-like applications such as plotly, and there are numerous examples on data science blogs etc.
- D3.js kind of figure, where you can interact with the data points to make them more relevant to your interests.

Miscellaneous

- The same comes to mind as in the description of question 7. Seen it before in a paper I reviewed and I am currently producing something similar.

- Animation.
- A figure that is not static, but can be viewed in many dimensions.
- Figure legend will be enough to explain the whole story in figure and I don't have to go and read the article to understand it.
- Generally annoying poorly used graphics that don't enhance my understanding. Want: intuitive visual interfaces with complex datasets.
- Videos, only ever seen them in *eLIFE*. There are others suggested as produced by the R package Shiny, where axis can be changed, parameters of models altered. However, I've never seen them used.
- I don't know. A moving or rotating 3D protein structure?
- I am not sure what they are, maybe figures available online only.

Free text responses from question 10: What other features could an interactive figure have that were not mentioned in the previous question?

***eLIFE* authors**

Manipulate the view of the figure

- Ability to alter figures to show different cohorts (colours, line/bar styles, etc.).
- View of the original picture with no adjustment.
- Change the format e.g. Bar chart to a line graph.
- The option to customize the figure in terms of graphic design.

- Molecule or protein structures rotate zoom functions.
- When selecting a part of a figure (e.g. a diagram or a fluorescence picture) or "hovering" over it with the pointer, show the applicable part of the figure legend or at least some supporting information (e.g. how often was the experiment repeated; abbreviations etc.) in a pop-up window.

Analysis related

- Ability to test the results displayed in the figure.
- Statistical toolbox for verifying results presented and/or being able to run different statistical testing procedures on the same data.
- A log of what parameters were changed by the user.
- Possibility of zooming an area of the picture.
- The ability to view statistical "cleaning" or treatment of the raw data (removal of outliers), with formulae available for any data transformations that occurred.
- Link to methodology.
- Link or mouse-over to the methodology used to obtain raw data and process the data to the form presented.
- Usually with large parameters the models become quite complex and difficult to interpret. So always our aim should be to have a simplistic model which a beginner can understand and adapt to his research quite easily.
- If it's a graph, the option to change for example the "Y" axis by clicking on it to plot the data according to different parameters.

Comments section to provide more insights

- If interactive figures are included, there should be an opportunity for authors to describe/recommend alternative analyses and plotting in the published text (i.e., not just in an online forum post publication). Also, the forum/comment section option needs to be carefully thought out since comments will not be peer-reviewed to the same extent that the manuscript has. There may be conflict of interest issues, unqualified negative comments, and so forth.
- In a perfect world, a link to other papers with similar experiments and data.
- A section for the authors to explain why they initially used preferential sets of parameters to analyse and/or display their results.

Data related

- Possibility to import data or include track/view from public repositories, e.g. ucsc or washu genome browser.
- To me the most useful aspect of an interactive figure would be to show the underlying distribution of the data, and/or raw traces of the data. This helps provide more insight that is often hidden in the averaged final figures.
- Ways to access the data behind parts of the figure (identify a gene on mouse over, show specific data values for selected points etc).
- A figure that allowed visualization of each series in a dataset independent of the other series, allowed the bar + SEM to be replaced by scatterplots, etc.

- Labels for data points that appear when the user hovers over the point.

Unclassified or cannot recommend

- What above seems comprehensive
- The options in q9 didn't work in my browser.
- 24 responded either as no extra features than those mentioned in the previous questions, NA, didn't know, or unsure, or didn't know what is the point of having extra features for a pop-up reference literature for example.
- Interactive figure.
- I'm afraid I can't imagine how informative any of this would be. It just seems likely to add confusion.
- None.
- I am concerned that this would complicate further the publication process: it should be optional but strongly recommended.
- I ordered last question points, but consider my ordering meaningless: all points are equally important.
- Ability to export the figure as a GIF for presentations.

Free text responses from question 13, 137 answered question 13 of which 47 provided free text responses explaining their understanding of the term "computationally reproducible data"

eLIFE responses, who answered, "yes that they know the term's meaning"

- Data generated by computational models.
- Within the FAIR principle.
- Computational workflows should be easy to reproduce, meaning the same data produces the same output. But software differences can make this not the case.
- "Computationally reproducible data" means that the computational model is specified in sufficient detail that someone else building the model will get identical results. "Reproducible data" with the adjective "computationally" prepended is less clear to me.
- Rerunning the same analysis pipeline yields the same results.
- Using the same starting data and the same software (original version + any custom scripts) you arrive at the exact same results.
- Being able to reproduce the date [data] in your own computer.
- To be able to reproduce the results obtained by others when using the same computational tools.
- Where you get the same results when a different person analyses it.
- We deal with cell signalling regulatory mathematical models and have adapted a couple of previously published models and reproduced computational data although at certain times it needs curation/parameter fitting of the model based on the additional parameters you modify.
- For example -if I take a RNAseq dataset (for instance) and run it through an analysis pipeline (say DESeq) I should be able to draw

the same conclusions as other people who have used the same data and run it through the same pipeline. Problems with reproducibility might be due to software version numbers, exact parameters used, or software bugs.

- Data or estimates should be the same, whatever system/software is used to analyse them.
- Computational data performed with similar conditions should result reasonably similar and lead to similar conclusions.
- Answered in a comment to the previous question. [previous question free-text response of the participant: Ways to access the data behind parts of the figure (identify a gene on mouse over, show specific data values for selected points etc).
- Computational data is reproducible when multiple researchers, on different or similar computational platforms, can arrive at the same analysis/conclusions and said analysis is not affected by software packaging, implementation, researcher documentation, etc.

NBI responses, who answered "yes that they know the term's meaning"

- Reproducing the data results of a computational experiment (given that this is achievable as some software will, of course, use heuristics).
- If I understand this rather badly worded question, you're asking what reproducible data is from a computational point. To me, it means data that would be identical every time you produce it.

- Being able to use the same sets of workflows and arrive at a close enough result with the same data.
- A study is reproducible if all of the code and data used to generate the numbers and figures in the paper are available and exactly produce the published results.
- A result which I can reproduce by myself through using the original author's raw data and code.
- Running a process on a dataset and getting the same answer twice.
- It means that the paper provides ALL the information needed for reproducing the analysis. It includes the workflow, software (with version and options), the code, statistical analysis, figure making.
- Providing some means of automatically re-running all computational analyses i.e. Jupyter notebooks, or the R equivalent (forgot the name), or even a very comprehensive bash script.
- Able to reproduce the same results using the same input data and workflow.
- Same data, same pipeline, same results. Docker containers, versioned workflows, VMs.

***eLIFE* free-text responses of participants who answered "unsure" of the meaning of the term "computationally reproducible data". Those denoted with an asterisk * are the answers we deemed as correct knowledge of the term**

- Modelling-based analysis, using certain measurements as input and predicting system behaviour based on parameters assessed

from those measurements. Or statistics-based analysis where certain thresholds govern the selection of significant results.

- *Given the exact same raw data, you are able to perfectly reproduce the results of the analysis performed?
- *Should "computationally reproducible data" all be italicized? Am I hearing the term "reproducible data" in a computational way?
- *I think it means that running the same or similar dataset through some computational paradigm produces the same or similar results.
- Concern result that is processed and generated by algorithms?
- *Mathematical models that can be reproduced by using the same parameters.
- *Having the same raw data, slightly different computational methods of analysis giving similar results.
- *I think this refers to the reproducibility of all the analyses that are done with software and the data that arises from those analyses.
- *Just as for non-computational data, reproducibility should be when using the same set of data, and the same code, by a different group, results are similar. One can also like that, in the case of homemade software, that similar results could be obtained when using publicly available software, provided that any parameters used in the analysis are the same.
- *I'd guess it means that given the raw data I'd be able to compute results identical or very similar to those provided by the authors. Well, this is ambiguous as "reproducible data" is italicized, making you think this is the key term, but the adjective "computationally" must be critical too. Recommend italicizing "computationally

reproducible data." I'm not sure what to make of this.

NBI free-text responses of participants who answered "unsure" of the meaning of the term "computationally reproducible data". Those denoted with an asterisk * are the answers we deemed as correct knowledge of the term

- *Getting similar results when rerunning the analysis. Getting similar results when using different approaches for the data analysis.
- Meaning repeating an in-Silico analysis of the data?
- *I believe it is providing complete and coherent data sets and data treatment protocols that are clear and provide consistent results.
- *To me it means that I have enough information about the data-handling applied to a data-set, and enough access to any tools used to handle the data, to reproduce exactly the output that someone (possibly me!) has already created from that data-set (example: if someone tells me they've carried out a principal component analysis on a data-set, and they tell me what sort of scaling and mean-centring they've done, I should be able to generate a PCA plot that is identical to theirs, ignoring, in this case, the fact that the sign of the loadings is arbitrary between algorithms, so axes might get flipped). I'm a chromatographer, so a lot of the data handling is using proprietary software from the instrument manufacturer (which is highly validated; open-source chromatography is rarer because regulated labs are never going to adopt something that isn't highly validated). This means that question 12 is irrelevant to me because I won't be able to reproduce others' results unless I use the same software as them, but it is

important to me that I can reproduce results we generated in our own lab, using the software we have. There is no way we'd be able to do this with data in a paper (until papers come with an embedded chromatography analysis system...probably not in my lifetime sadly).

- I don't see what you mean by "computationally". For me "reproducible data" is data that is likely to be gathered again by repeating the experiment.
- Data that is easy to access and use reproducibly in line with a specific methodology or pipeline.
- *The ability to come up with the same results as a published study by using the same data and workflow that is described in the paper.
- *Being able to have access to a raw input dataset and be able to put it into an open access analysis pipeline to recreate a figure?
- *Running the same program and parameters will give similar results.
- *The data and process presented in a paper should be reproducible by any interested reader; so, the data should come with: -a description of the system setup -a description of each step of the analysis where "description", above, encompasses also ancillary data such as configuration files, glue scripts, and hopefully a pipeline manager script that allows going from point 0 to the end of the analysis in an automatic way.

NBI free-text response of the respondent that replied "No, don't know the meaning of the term computationally reproducible data. "No" responses were not asked to provide a free-text explanation

- But it is guessable: scripts which can be run again to generate data which recapitulate the results.

Free text responses from question 15: What type of bioinformatic tools were used by the participants of the survey (36 responses)

- DNA Sequences, RNA-Seq analysis.
- A variety of tools to analyse metabolomics data.
- Transcriptomics (filtering, alignment, differential expression), operation on genomic intervals, data integration.
- Spades, STAR.
- RNA-Seq alignment, False Discovery Rate Analysis, Monte Carlo Simulations.
- Sometimes RNA-Seq tools.
- RNAseq, domain predictions, coding predictions, aligners, sometimes modellers.
- Structural bioinformatics and modelling tools, docking, molecular simulations.
- RNA-seq analysis, ChIP-seq analysis, statistical analysis with R.
- Protein modelling.
- Proteomics software.
- Assemblers, 3D protein viewers.
- RNA-Seq, ChIP-Seq and Hi-C.

- Neural networks modelling tools.
- Sequence and NGS analysis.
- Genome browsers and databases, RNA-seq, ChIP-seq.
- Proteomics processing tools, GO terms, KEGG< String etc.
- Sequence alignment.
- ChIP-seq and RNA-seq tools.
- Metagenomic tools (Metaphlan, HUMANN, qiime, mothur, PICRUSt, etc.).
- DNA/RNA/protein sequence analysis tools, statistical programs.
- Aligners, homology BLAST.
- Network analysis tools, statistical tools.
- Assemblers.
- Ensemble, NCBI BLAST.
- Perseus -mass spec analysis tool.
- Alignment protein, RNA, DNA sequences; image analysis; tracking, etc.
- Structure modelling tools.
- RNA-seq read assembly, aligner, read count quantification, differential expression.
- Aligners, mostly. I do very basic bioinformatics on a fairly regular basis and have done some more demanding work on occasion (microarray and RNAseq analysis).
- Proteomics tools.

- I did in the past -my experience must be way out of date by now!
- RNA-seq, mutation analysis, models.
- Structural modelling (e.g. Rosetta, Modeller) Molecular dynamics Kinetic Monte Carlo.
- All sorts of stuff. All of the above and more.
- RNAseq analysis; microbial community analysis; (meta)genomic analysis, phylogenetic tools.

A.4 Supplementary Figure 1

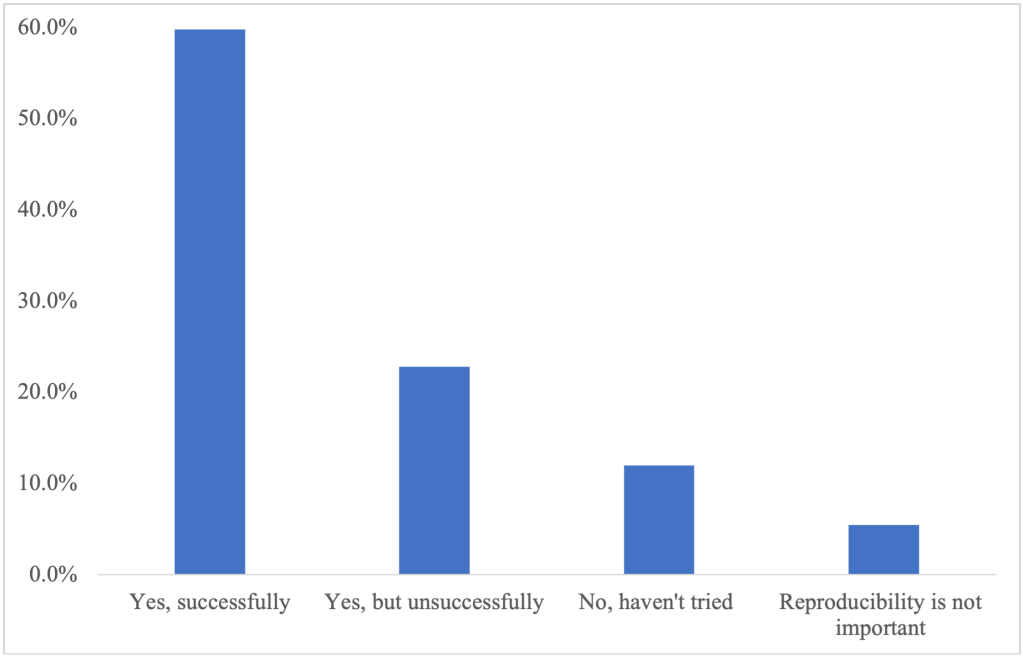

Figure A.4.1: Responses to question 6 (which was only available for the *eLIFE* authors' survey as it was added after the NBI study was distributed). The percentage of respondents that attempted to reproduce published results and to what extent they have been successful or not. The figure has been reproduced with permission from Samota and Davey (2021)

.

A.5 Supplementary Table 1

	Number (% of total sample)	Success and willingness in reproducing any published results				
Variable		**Successful (% within variable)**	**Attempted but unsuccessfully (% within variable)**	**Never tried (% within variable)**	**Not important to reproduce (% within variable)**	**P-value**
Training *(n = 90)*						
Bioinformatics	42 (46.7)	26 (61.9)	10 (23.8)	5 (11.9)	1 (2.4)	0.758*
Not trained in Bioinformatics	48 (53.3)	28 (58.3)	11 (22.9)	5 (10.4)	4 (8.3)	
Computer Science	33 (36.7)	21 (63.6)	3 (9.1)	7 (21.2)	2 (6.1)	0.018*
Not trained in Computer Science	57 (63.3)	33 (57.9)	18 (31.8)	3 (5.3)	3 (5.3)	
Statistics	71 (78.9)	42 (59.2)	16 (22.5)	9 (12.7)	4 (5.6)	0.937*
Not trained in Statistics	19 (21.1)	12 (63.2)	5 (26.3)	1 (5.3)	1 (5.3)	

*Fisher Exact test, more than 20% of cells had expected frequencies lower than 5.

This table appears in my first author publication (Samota and Davey, 2021) and is printed here with permission.

A.6 Supplementary Table 2

Frequency using bioinformatics tools and ability and willingness to reproduce experiments: a research survey.

	Frequency of using bioinformatic tools	
Success and willingness in reproducing any published results (N=90)	Often	Rarely/Never
Successful	19	35
Attempted but unsuccessful	7	14
Never tried	3	7
Not important to reproduce	1	4

This table appears in my first author publication (Samota and Davey, 2021) and is printed here with permission.

A.7 Supplementary Table 3

The profiles of individuals in the *eLIFE* survey that replied "No, have never tried reproducing any published results" in Question 6 and how the respondents replied in each of the questions 13, 15, 16, 17. From the manuscript "Knowledge and attitudes among life scientists towards reproducibility within journal articles: a research survey." (Samota and Davey, 2021)

Have you heard the term "computationally reproducible data", and do you understand what does the term mean? (If answered yes, or unsure, explanation of term means) (q13)	How often do you work with bioinformatics analysis tools (please specify the nature of the tools you use)? (q15)	Have you received any of the following training? Formal or informal (through a colleague etc.).	Types of data they work with
No	Rarely	Bioinformatics, Statistical	Biochemistry, Genomics
Yes (did not give any comments to explain what the term means)	Rarely	Bioinformatics, Statistical	Biochemistry, Computational Biology, Genomics
Yes (did not give any comments to explain what the term means)	Often	Bioinformatics, Statistical	Computational Biology, Genomics
No	Never	Computer Science, Statistical	Immunology, Population/Epidemiology
No	Never	Statistical	Biophysics, Computational Biology, Other (neural potential data - single cell, LFP, EEG)
No	Often (RNA-seq read assembly, aligner, read count quantification, differential expression)	Bioinformatics	Genomics
No	Rarely	Statistical	Other (Brain imaging, eye tracking
No	Rarely	Statistical	Biophysics, Imaging and electrophysiology
No	Never	Computer Science, Statistical	Biochemistry
Yes (Computational data is reproducible when multiple researchers on different or similar computational platforms, can arrive at the same analysis/conclusions and said analysis is not affected by software packaging, implementation, researcher documentation, etc.)	Often	RNAseq analysis, microbial community analysis, metagenomic analysis, phylogenetic tools	Genomics, Population/epidemiological
No	Did not answer this question	Computer Science	Did not answer this question

A.8 Supplementary Figure 2

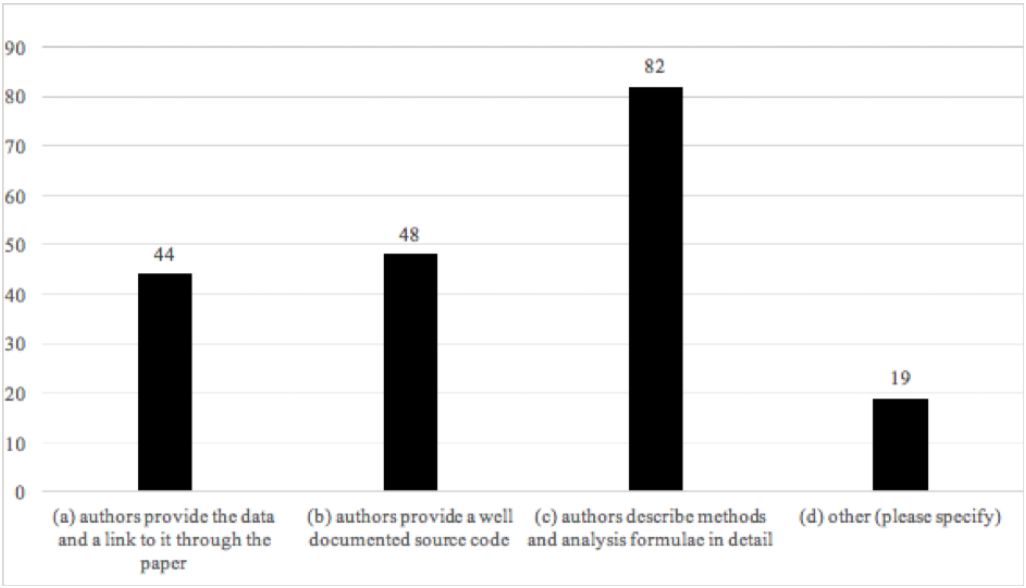

Figure A.8.1: Responses to question 7 (which was only available for the *eLIFE* authors survey as it was added after the NBI study was distributed): The opinions of the respondents on what could be done to make published research more reproducible. We received 19 comments from the "other" comment. The numbers (shown on the y-axis) correspond to the number of participants. For the free-text (other) responses, see Supplementary Section 2. The figure has been reproduced with permission from Samota and Davey (2021)

.

B

Appendix for Chapter 5

B.1 Example XML file corresponding to the EBI ArrayExpress Experiment Accession code E-MTAB-1729. The XML file is fetched by the *Deus ex machina* code by executing REST-Requests to the ArrayExpress API.

<?xml version="1.0" encoding="UTF-8"?><experiments api-version="3" api-revision="091015" version="3.0" revision="091015" total="1" total-samples="60" total-assays="60"><experiment><id>554620</id><accession>E-MTAB-1729</accession><secondaryaccession>ERP003465</secondaryaccession><name>RNA-seq of coding RNA from near-isogenic lines, harboring either the resistant or the susceptible allele for Fhb1 and Qfhs.ifa-5A, in response to F. graminearum</name><releasedate>2013-10-10</releasedate><lastupdatedate>2017-03-09</lastupdatedate><organism>Triticum aestivum</organism><experimenttype>RNA-seq of coding RNA</experimenttype><experimentdesign>co-expression</experimentdesign><experimentdesign>in vivo</experimentdesign><experimentdesign>stimulus or stress</experimentdesign><description><id/><text>Near isogenic wheat lines, differing in the presence of the FHB-resistance QTL Fhb1 and Qfhs.ifa-5A, have been sequenced using Illumina HiSeq2000 under disease pressure (30 and 50 hai) as well as with mock-inoculation, to discern transcriptional differences induced by Fusarium graminearum.</text></description><provider><contact>Wolfgang Schweiger</contact><role>data generator submitter</role><email>wolfgang.schweiger@boku.ac.at</email></provider><bibliography><accession>24152241</accession><authors>Kugler KG, Siegwart G, Nussbaumer T, Ametz C, Spannagl M, Steiner B, Lemmens M, Mayer KF, Buerstmayr H, Schweiger W</authors><title>Quantitative trait loci-dependent analysis of a gene co-expression network associated with Fusarium head blight resistance in bread wheat (Triticum aestivum L.).</title><doi>10.1186/1471-2164-14-728</doi></bibliography><bibliography><accession>26438291</accession><authors>Nussbaumer T, Warth B, Sharma S, Ametz C, Bueschl C, Parich A, Pfeifer M, Siegwart G, Steiner B, Lemmens M, Schuhmacher R, Buerstmayr H, Mayer KF, Kugler KG, Schweiger W</authors><title>Joint Transcriptomic and Metabolomic Analyses Reveal Changes in the Primary Metabolism and Imbalances in the Subgenome Orchestration in the Bread Wheat Molecular Response to Fusarium graminearum</title><doi>10.1534/g3.115.021550</doi></bibliography><samplecharacteristic><category>developmental stage</category><value>anthesis</value></samplecharacteristic><samplecharacteristic><category>genotype</category><value>CM-82036 resistant parent line</value><value>NIL1, Fhb1 and Qfhs.ifa-5A resistance alleles</value><value>NIL2, Fhb1 resistance allele</value><value>NIL3, Qfhs.ifa-5A resistance allele</value><value>NIL4, non resistance allele</value></samplecharacteristic><samplecharacteristic><category>organism</category><value>Triticum aestivum</value></samplecharacteristic><samplecharacteristic><category>organism part</category><value>spikelet floret</value></samplecharacteristic><samplecharacteristic><category>provider</category><value>Institute for Biotechnology in Plant Production, IFA-Tulln, University of Natural Resources and Life Sciences, A-3430 Tulln, Austria</value></samplecharacteristic><experimentalvariable><name>genotype</name><value>CM-82036 resistant parent line</value><value>NIL1, Fhb1

and Qfhs.ifa-5A resistance alleles</value><value>NIL2, Fhb1
resistance allele</value><value>NIL3, Qfhs.ifa-5A resistance
allele</value><value>NIL4, non resistance allele</value></
experimentalvariable><experimentalvariable><name>infect</
name><value>Fusarium graminearum</value><value>mock</value></
experimentalvariable><experimentalvariable><name>time</
name><value>30 hour</value><value>50 hour</value></
experimentalvariable><protocol><id>1229079</id><accession>P-
MTAB-33418</accession></protocol><protocol><id>1229077</
id><accession>P-MTAB-33419</accession></
protocol><protocol><id>1229076</id><accession>P-MTAB-33416</
accession></protocol><protocol><id>1229075</id><accession>P-
MTAB-33417</accession></protocol><protocol><id>1229078</
id><accession>P-MTAB-33415</accession></
protocol><protocol><id>1229074</id><accession>P-MTAB-33420</
accession></protocol><bioassaydatagroup><id/><name>scan</
name><bioassaydatacubes>60</bioassaydatacubes><arraydesignprovider/
><dataformat>scan</dataformat><bioassays>60</
bioassays><isderived>0</isderived></bioassaydatagroup></
experiment></experiments>