Data integration and handling

Building an informatics platform for research integrated biobanks

Tomas Klingström

Faculty of Veterinary Medicine and Animal Science
Department of Animal Breeding & Genetics
Uppsala

Doctoral thesis
Swedish University of Agricultural Sciences
Uppsala 2017

Acta Universitatis agriculturae Sueciae

2017:99

Cover: Photo of the equipment used to perform sequencing as a public display at the 40 year anniversary of the Swedish University of Agricultural Science.

(photo: Erik Bongcam-Rudloff digital editing: Tomas Klingström)

ISSN 1652-6880
ISBN (print version) 978-91-7760-088-6
ISBN (electronic version) 978-91-7760-089-3

Print: SLU Service/Repro, Uppsala 2017

Data integration and handling: Building an informatics platform for research integrated biobanks

Abstract

Modern technology allows researchers to generate data at an ever increasing rate, outpacing the capacity of researchers to analyse it. Developing automated support systems for the collection, management and distribution of information is therefore an important step to reduce error rates and accelerate progress to enable high-quality research based on big data volumes. This thesis encompasses five articles, describing strategies for the creation of technical research platforms, as well as descriptions of the technical platforms themselves.

The key conclusion of the thesis is that technical solutions for many issues have been available for a long time. These technical solutions are however overlooked, or simply ignored, if they fail to recognise the social dimensions of the issues they try to solve. The Molecular Methods database is an example of a technically sound but only partially successful solution in regards to social viability. Thousands of researchers have used the website to access protocols, but only a handful have shared their own work on MolMeth. Experiences from the Molecular Methods database and other projects have provided a foundation for studies supporting the development of the eB3Kit

The eB3Kit is a portable, robust and scalable informatics platform for structured data management. Deploying the platform enables research groups to carry out advanced research projects with very limited means. With the eB3Kit researchers can integrate data from a wide variety of sources, including the local laboratory information management system and analyse it using the Galaksio interface. Galaksio provides user friendly access to the Galaxy workflow management system and provides eB3Kit users with access to tools developed by a far larger user community than the one actively developing the eB3Kit. Using a workflow management system improves reproducibility and enables bioinformaticians to prepare workflows without directly accessing ethically or commercially sensitive data. Therefore, it is especially well-suited for applications where researchers are worried about privacy and during disease outbreaks where persistent storage and analysis capacity must be established quickly.

Keywords: bioinformatics, biobanking, scholarly communication, ethics, genetics, Galaxy, Galaksio, The Molecular Methods database.

Author's address: Tomas Klingström, SLU, Department of animal breeding and
genetics. P.O. Box 7023, 750 07 Uppsala, Sweden

Data integration och hantering: Att utveckla en informatikplattform för forskningsintegrerade biobanker

Abstract

Modern teknik gör det möjligt att generera mer data från experiment än någonsin tidigare. Inom livsvetenskaperna har denna utveckling gått så fort att forskare ofta genererar mer data än de klarar av att analysera. Det är därför viktigt att utveckla automatiserade stödsystem för provinsamling, informationshantering och dataöverföring för att forskare bättre ska kunna hantera stora datamängder. Denna avhandling omfattar fem delarbeten som beskriver strategier för att bygga tekniska informationsplattformar och två exempel på sådana plattformar.

Den övergripande slutsatsen är att tekniska lösningar för många problem har funnits länge. Däremot har dessa lösningar ofta inte applicerats i praktiken då de varit fokuserade på tekniska aspekter utan att hantera de sociala dimensionerna som styr forskningen. The Molecular Methods database är ett sådant exempel. Tusentals forskare har använt hemsidan för att hitta laboratorieprotokoll men endast ett fåtal har valt att själva bidra med protokoll till hemsidan. Erfarenheterna från detta arbete och andra projekt har därefter legat till grund för studier till stöd för utvecklingen av eB3Kit inom ramen för B3Africa-projektet.

eB3Kit är en portabel, robust och skalbar plattform för strukturerad hantering av data. Genom att använda plattformen kan forskargrupper genomföra omfattande forskningsprojekt med väldigt begränsade resurser. Plattformen gör det möjligt att integrera data från många olika källor och analysera dessa med hjälp av Galaksio. Galaksio är ett användargränssnitt utvecklat för att skapa en mer användarvänlig miljö vid hanteringen av bioinformatiska arbetsflöden. Gränssnittet är direkt kopplat till Galaxy workflow management system och innebär att forskare kan dra nytta av arbetsflöden utvecklade av långt fler forskare än de som använder eB3Kit. Genom att använda Galaxy ökar spårbarheten inom dataanalys och det är även möjligt för bioinformatiker att förbereda analysflöden utan att själva komma i kontakt med etiskt eller kommersiellt känsliga data. eB3Kit är därför väl lämpat för situationer där forskare behöver kunna ha kontroll över känsliga data eller snabbt etablera en provsamsamling och analysera data vid exempelvis sjukdomsutbrott.

Keywords: bioinformatics, biobanking, scholarly communication, ethics, genetics, Galaxy, Galaksio, The Molecular Methods database.

Author's address: Tomas Klingström, SLU, Institutionen för Husdjursgenetik. P.O. Box 7023, 750 07 Uppsala, Sweden

Preface

For most researchers, science is a journey of increasing specialisation. As research grows more complex there is a need to build bridges between these specialisations. This thesis is therefore an interdisciplinary endeavour to provide a bioinformatics-based approach to unify knowledge from molecular biology, ethics and bioinformatics, so that researchers may enjoy the fruits of success from other fields while specialising within their own.

In this age of specialization men who thoroughly know one field are often incompetent to discuss another. The great problems of the relations between one and another aspect of human activity have for this reason been discussed less and less in public.

Richard Feynman (2 May 1956) at a Caltech YMCA lunch forum

Dedication

Science is all about dedication and one of the great passions of my life. Thus I dedicate this thesis to my other great passion, my family.

Pa,pa, pa. Ma, ma

- Alvar Klingström (2017), always.

Contents

List of publications

This thesis is based on the work contained in the following papers, referred to by Roman numerals in the text (* Corresponding author):

I Klingström T., Soldatova L., Stevens R., Roos T.E., Swertz M.A., Müller K.M., Kalaš M., Lambrix P., Taussig M.J., Litton J.E., Landegren U., Bongcam-Rudloff E. (2013). Workshop on laboratory protocol standards for the molecular methods database. *New biotechnology*, *30*(2), 109-113.

II Scholarly publication in the digital age, an investigation into why novel publishing concepts have failed to disrupt the market of scientific journals. Klingström T., Regierer B., Attwood T.K., Bongcam-Rudloff E. *Manuscript.**

III Klingström T., Bongcam-Rudloff E., & Reichel J.. (2017). Legal & ethical compliance when sharing biospecimen. *Briefings in Functional Genomics*, elx008.*

IV Klingstrom, T., Mendy, M., Meunier, D., Berger, A., Reichel, J., Christoffels, A., Bendou H., Swanepoel C., Smit L., Mckellar-Basset,. & Bongcam-Rudloff, E. (2016). Supporting the Development of Biobanks in Low and Medium Income Countries. In *IST-Africa Week Conference, MAY 11-13, 2016, Durban, SOUTH AFRICA*.*

V Klingström T., Hernandez de Diego R., Collard T., Bongcam-Rudloff E. (2017). Galaksio, a user friendly workflow-centric front end for Galaxy. EMBnet.journal. *In press**

Papers I-V are reproduced with the permission of the publishers.

The contribution of Tomas Klingström to the papers included in this thesis was as follows:

I Collected presentations and discussions into a synthesis to guide the development of the Molecular Methods database, wrote the manuscript with feedback from co-authors.

II Conceived the project, carried out the study and wrote the article with feedback from the other co-authors.

III Conceived the project, outlined the problematic areas of ethics and law for geneticists and wrote the manuscript based on the response by co-authors.

IV Participated in the development of the project, co-authored the funding proposal and combined key points of the proposal with results from the first year of the project to produce an initial draft. Work package leaders then added additional information into prepared sections of the manuscript.

V Created the overall design concept of the bioinformatics module and wrote the manuscript in collaboration with co-authors.

List of tables

List of figures

Abbreviations

API	Application Programming Interface
B3Africa	Bridging Biobanking and Biomedical Research across Europe and Africa
BBMRI.se	BioBanking and BioMolecular Resources Research Infrastructure of Sweden
BBMRI-eric	BioBanking and BioMolecular Resources Research Infrastructure European Research Infrastructure Consortium
BBMRI-LPC	BioBanking and BioMolecular Resources Research Infrastructure Large Prospective Cohorts
BCNet	Biobank and Cohort Network
BRISQ	Biospecimen reporting for improved study quality
CEN/TS	European Committee for Standardization/Technical Standard
CMS	Content Management System
CTRNet	Canadian Tissue Repository Network
EMERALD	Enhancing Microarray Data Quality
EU-GDPR	European Union General Data Protection Regulation
EXACT	Experiment ACTions Ontology
Goblet	Global Organisation for Bioinformatics Learning, Education & Training
GUI	Graphical User Interface
H3Africa	The Human Heredity and Health in Africa
H3Bionet	Pan African Bioinformatics Network for H3Africa
HUGO	Human Genome Organisation
IT	Information Technology
LIMS	Laboratory Information Management System
LMIC	Low and Medium Income Country
MIABIS	Minimum Information about Biobank data Sharing
MolMeth	The Molecular Methods database
NGS	Next Generation Sequencing

OBO	Open Biomedical Ontologies
SciLifeLab	Science for Life Laboratory
SOP	Standard Operating Procedure
SPREC	Standard PREanalytical Code
SRF	Sequence Read Format
UMLS	Unified Medical Language System
USD	American dollar
WP4	Work Package 4

1 Introduction

As research projects grow larger and more interdisciplinary they also grow more complex and include more members producing material and co-authoring articles (Park et al., 2015; Papatheodorou et al., 2008; Weeks et al., 2004). This additional complexity makes it harder for researchers to complete a *research cycle* where a hypothesis is formulated, tested, and results presented for review by the research community as it is necessary to exchange a significant amount of information both within research groups and across the wider community to complete the cycle.

As a result, researchers struggle to plan and organise their research in larger and larger groups consisting of experts from several different fields which require information to be shared in a formal and structured way to be meaningful (Fjällbrant, 2012; Fiore, 2008). At the same time, researchers find themselves in an hypercompetitive environment putting substantial pressure on researchers to produce research with limited time and incentives to store data persistently (Alberts et al., 2014; Teitelbaum, 2008) leading to significant loss of data each year (Glasziou et al., 2014). While many researchers have proven reluctant to commit to new and potentially disruptive technologies, the widespread adoption of social networking technologies for dissemination and, demonstrate that researchers are willing to commit to innovations for managing information provided that they provide tangible benefits at an acceptable cost of adoption.

The Molecular Methods database (MolMeth, www.molmeth.org) is one such example of potentially disruptive technology (paper I). MolMeth is built to provide a platform for researchers to communicate experimental procedures based on ontologies within research groups and consortia (Klingström et al., 2013). Upon completion of a project, or, when deemed suitable, protocols can be made publically available for dissemination to the wider research community.

Paper II was devised to support the development of the Molecular Methods database. In 2013 we could conclude that even with the revised web page created based on paper I overall positive feedback did not convert into actual user engagement on the website. Many researchers are frustrated by the peer-review system (Björk and Solomon, 2013; Harley and Acord, 2011; Akerman, 2006) and positive about alternative means for dissemination. Based on user behaviour it was however clear that overall support of the concept was insufficient to commit the time and effort necessary to provide protocols.

Lessons from the MolMeth project and the study of scholarly communication presented in paper II have been applied in the B3Africa project where the author has been head of work package 4 (WP4) which produces the bioinformatics module for the eB3Kit. The eB3Kit is a stand-alone platform which integrates a number of virtual machines into a complete system for managing large sample collections and all research tasks related to them. The software platform can be installed on any server, and when deployed it provides a robust and easy to handle solution for the IT-platform necessary to operate a biobank. To make it useful to the research community it is, however, necessary to combine the technical development with an understanding of the users, their needs, and motivations for using the platform. The B3Africa project and its bioinformatics component have therefore been developed based on valuable experience from other projects such as the eBiokit, H3Africa, BCNet, MolMeth and Goblet (paper IV).

To create a solution that is integrated into the daily work of the biobank the eB3Kit is designed to provide a platform with structured data which allows researchers to formalise their information management and facilitate interdisciplinary collaboration where they more effectively distribute labour over time between experts. Modules such as the bioinformatics module and STATegra Experimental Management System (Hernández-de-Diego et al., 2014) are therefore not limited to stand alone actions but communicate with each other to enhance communication within the project. In such a setting bioinformatics is not only an important component of analysis towards the end of a project but can also provide information for quality management and identification of pre-analytical variables that may be substantial discoveries themselves or serve as confounding factors in the analysis of samples retrieved from the biobank.

As long as data is appropriately structured and accurately reflects the underlying biological system under investigation, more information is univocally positive for research. It is therefore natural that bioinformatics is a research field with strong connections to open data sharing going back as far in history as the research field itself (Ouzounis and Valencia, 2003).

Unfortunately such open sharing, albeit scientifically beneficial, may interfere with individual interests or rights in regards to individual privacy or ownership rights for livestock, plants and pets. A technology platform supporting bioinformatics should, therefore, provide strong support for data access and sharing, but must do so in a manner that balances the individual rights of the people contributing data. Balancing the benefits of open sharing with individual rights is far from a straightforward task and bioethics has emerged as its field of research (Pellegrino, 1999). To better implement a balance between safeguards and sharing there is close collaboration within the B3Africa project in regards to ethical and legal limitations of data sharing. Paper III provides a primer on the basics of ethical and legal management of data, and the eB3Kit will contain features that help researchers manage the issues while keeping as much data as possible available for open collaboration.

To handle the complexity of projects, it is often necessary to divide research tasks based on the technology or expertise required to resolve them. This division of labour requires that researchers can share information between them and access proven solutions from outside their core area of competency. These proven solutions are often provided in the form of services such as storage in biobanks or bioinformatics support or maintained by dedicated flagship platforms such as the ones hosted at the Science for Life Laboratory. Infrastructures can be developed as external infrastructures open to researchers from a broad range of institutions or internal infrastructures with staff scientists providing specialised services and development of new technologies applicable to research at the institute (Hyman, 2017). The eB3Kit bioinformatics module is based on the Galaxy Workflow management system but with a new layered approach providing a "service layer" to research biologists who mainly seek to apply pre-existing bioinformatics tools. Routine tasks can thereby be automated enabling research institutions to provide a more attractive working environment for bioinformaticians where they can focus on tasks to advance their careers in research which is often a challenge in the core services (Chang, 2015) or develop new tools for other researchers to expand the capacity of the institution.

The bioinformatics module which is called Galaksio (Esperanto for Galaxy) is developed by WP4 but can be installed independently of the eB3Kit and connected to any Galaxy server (paper V). Combined with other Galaxy integrations such as Pulsar and Cloudman this means that bioinformatics can be approached in layers based on the expertise of the user. Researchers who merely wish to apply bioinformatics tools can access the workflows in Galaksio; bioinformaticians can manage or create workflows in Galaxy and workflow tasks can be sent to external computing clusters operated by trained administrators or computer scientists for demanding computations.

The integrated environment with support for ethics, quality management and automated bioinformatics provides an informatics platform that can reduce the workload of researchers faced with the increasing complexity of research. The B3Africa project is focused on supporting the development of biobanks in Low and Medium-Income Countries (LMIC) by producing the eB3Kit (paper IV) but many lessons learned can be applied in other environments as well, and an eB3Kit will also be installed at Karolinska Institutet Biobank for specific research projects.

The modular approach of the eB3Kit also makes it possible to install components of the kit separately in virtual machines. Such an approach is of particular importance for the bioinformatics component as it will allow local installations at larger servers located at the Swedish University of Agricultural Sciences (SLU) and Karolinska Institutet. Allowing researchers to automate routine tasks in biology and engage in international collaboration as workflows and tool setups can easily be exchanged within the network of researchers using Galaksio and Galaxy.

1.1 The importance of biobanks and bioinformatics in Low and Medium-Income Countries

Establishing a biobank allows institutions to create a more efficient supply chain in providing biospecimen for high-throughput experiments and longitudinal studies. By setting up a biobank, an institution can benefit from four major value drivers (Rogers et al., 2011).

- The value of physical capital is optimised as necessary laboratory and storage equipment is handled by a single responsible body. Ensuring that samples are kept in a consistent manner over time, and economic benefits can be leveraged as duplication of work and infrastructure is avoided.
- Human capital is leveraged as dedicated staff can be assigned to specific functions and trained to uphold strict control practices and reduce waste caused by inefficiencies or high defect rates.
- Bioinformatics and information systems can be connected and centralised, allowing researchers to evaluate their data without relying on external stakeholders. The quality of analysis is also improved as the influence of patient data, potential confounding factors and other highly annotated metadata can be evaluated in the study.
- Certification of operating procedures and best practices to improve the credibility of the institution.

Together these four value drivers make it possible for researchers to engage in research projects to collect highly annotated samples and perform credible research attractive for international collaborations and investments. If the biobank is established in a region permissive for investment by the pharmaceutical and medical companies economic benefits can reach tens of millions of USD in reduced clinical trials costs (Rogers et al., 2011) as well as substantial cost savings for medical treatment by enabling the development of personalised medicine (Warnich et al., 2011; Compton, 2007). The economic potential by development in medical care driven by biobank based research has provided motivation for the H3Africa project (The H3Africa Consortium et al., 2014). Providing 21 grants to selected researchers working at leading institutions chosen by the U.S. National Institutes of Health and UK-based Wellcome Trust (The H3Africa Consortium et al., 2014) to develop capacity at flagship institutions collecting and analysing samples from 50 000 to 75 000 human participants.

For researchers engaged in fundamental research or focusing on biological samples the economic benefits are likely to be less dramatic but still sufficient to generate significant returns. To strengthen research capacity in low and medium income countries several stakeholders therefore decided to join together and form the B3Africa consortium (described in paper IV). The consortium will develop the eB3Kit, a low-cost informatics platform providing researchers with the technical infrastructure necessary to manage governance, ethics, infrastructure, bioinformatics and data exchange (see figure 1) and provide researchers with the training to use it.

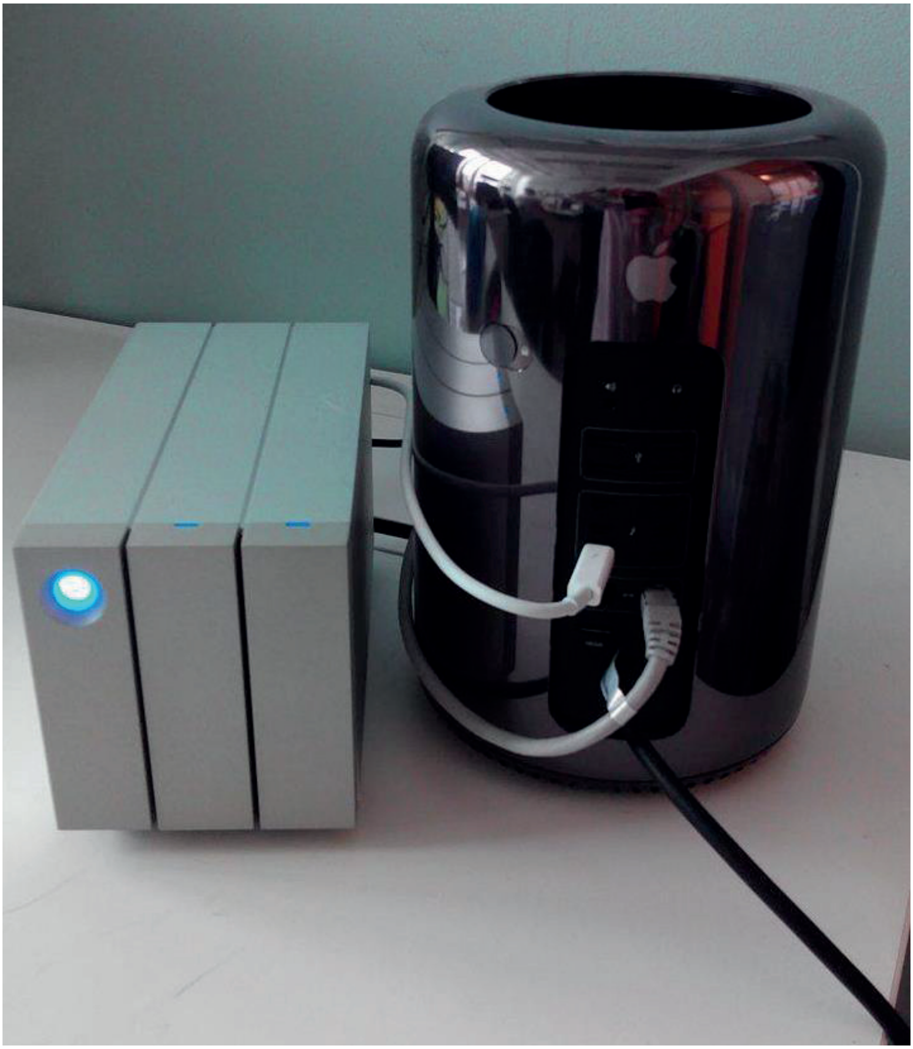


Figure 1. The eB3Kit and the B3Africa project is described in depth in paper IV. The eB3Kit consists of separate software modules installed in a virtual machine. The contribution of the author has been to participate in the writing of the initial proposal and manage work package 4 (bioinformatics) and design the bioinformatics module as well as its interactions with other software modules. As the server runs in a virtual machine, it can be installed on practically any server with sufficient power, but our preferred configuration is the above setup using a Mac Pro server and an external storage device connected by a Thunderbolt cable. This preferred setup is based on previous experience with the eBiokit bioinformatics platform where portability and robustness have been key factors to success by enabling researchers to deal with harsh conditions and manage limitations in internet connectivity and maintenance.

Providing the eB3Kit allows us to capitalise on existing research networks (IARC, 2016; Mulder et al., 2016) and establish an IT-infrastructure that supports the long-term management of samples and the resources necessary to analyse and share data. Such a platform can be rapidly deployed in regions with outbreaks of dangerous pathogens such as the recurring Ebola outbreaks in West Africa (Abayomi, Gevao, et al., 2016). It may also reduce the costs currently restricting the development biobanks for livestock, plant and environmental samples (Groeneveld et al., 2016).

With its genetic diversity and a large number of local breeds adapted to harsh conditions, Africa is a unique region with high potential for the collection of biological (non-human) as well as human samples (Heymann, 2017; Notter, 1999). When engaging in such projects, it is important that collaboration is conducted in a mutually beneficial manner to ensure that stakeholders are motivated to maintain a long-term and sustainable collaboration. In contrast, many previous projects have failed to become sustainable as samples were collected in Africa but contributed little or no tangible benefits to local communities (Abayomi, Gevao, et al., 2016; de Vries et al., 2015) are making future collaborations unlikely between involved partners. For biological (non-human) research the relative lack of support for biological samples collections in Europe (Directorate-General for Research and Innovation, 2016; Groeneveld et al., 2016) and sheer abundance of potential research topics provide additional incentives for supporting the development of strong local research communities. Since the capacity of each research centre is limited, comparative genomics and large-scale projects will need to rely on a large number of relatively small stakeholders collecting samples in each region.

For biological biobanks, comparative genomics and longitudinal studies are central areas, and strong local research communities can collect, annotate and analyse samples in a manner allowing comparisons to be made across different regions and breeds. Thereby strengthening agricultural and environmental research in a wide variety of areas such as genomic selection for breeding (Hayes et al., 2013), conservation efforts (Blackburn, 2012), connecting nutritional values to genomics (Maurice-van Eijndhoven, 2014), interventions against epizootic diseases (Moen et al., 2015). Biological biobanks may also have significant implications for human health as animal models provide valuable insights into human health (Andersson, 2016) and the identification as well as prevention of zoonotic diseases (DiEuliis et al., 2016).

1.2 Managing pre-analytical variables in sample collections

The research community is increasingly worried about issues regarding the validity and replicability of scientific findings (Baker, 2016), and pharmaceutical companies warn that billions are wasted each year on research based on erroneous results (Begley and Ellis, 2012; Prinz et al., 2011). Meta-research to study how to better conduct research and the development of more stringent quality management methods to reduce pre-analytical variation are therefore increasingly important components of research (Kousta et al., 2016). As a result several standards for measuring and reporting sample handling and results such as SPREC (Betsou et al., 2010), BRISQ (Helen M. Moore et al., 2011) and the recently approved Technical Standards on biobank samples from the European Committee for Standardisation(Neururer et al., 2016) have been suggested.

It is yet too early to evaluate the implementation of the CEN/TS standards but neither SPREC nor BRISQ are widely cited in comparison to the number of biobanks related articles published (Astrin and Betsou, 2016). Suggesting that standards have so far seen limited application in biobanks even if quality management in biobanks is a recurring topic at the annual ESBB and Hands-on Biobanking conferences.

Validation studies performed by the pharmaceutical company Amgen was only able to replicate results in 6 out of 53 high profile papers, despite collaborating with the original authors. It is highly worrisome that such a large number of secondary papers and expensive clinical trials are being based on questionable findings (Begley and Ellis, 2012). Further corroboration that much research and development relies on irreproducible findings is provided by an in-house survey at Bayer Healthcare, estimating that only 20-25 % of the tests were fully reproducible (Prinz et al., 2011) , as well as a wider survey performed by Nature among academic researchers (Baker, 2016). Combined, these findings suggest that the increasing rate of failure among phase II clinical trials in pharmaceutical research may very well be caused by an increasing number of pharmaceutical development projects being based on fundamentally flawed research (Arrowsmith, 2011).

Reducing the influence of independent pre-analytical variables is an obvious step towards increasing the replicability of studies (Freedman et al., 2015). The pre-analytical phase is according to ISO 15189 defined as a process that starts with the clinician’s request and includes the examination request, preparation and identification of the patient, collection of the primary sample(s), transportation to, and within the laboratory, and end when the analytical examination begins (Doucet et al., 2016). Covering the entire pre-analytical phase for a biobank is an extensive undertaking as there is a high number of potentially relevant pre-analytical factors. A workshop sponsored by The College of American Pathologists Diagnostic Intelligence and Health Information Technology Committee suggested a list of 170 pre-analytical variables (Robb et al., 2014) and Neururer et. al. listed over 300 (Neururer et al., 2016). It is impossible for such an extensive list to be covered using informal means or manual record keeping and it is, therefore, necessary to invest in a Laboratory Information Management System (LIMS) suitable for storage as well as processing of samples when establishing a biobank (Hallmans and Vaught, 2011; Dangl et al., 2010). The efficiency of informatics management can be significantly improved by integrating data collection, LIMS, bioinformatics analysis, and ethics support in one package. Biobanks in high-income countries often have access to dedicated systems but with poor integration between the systems used by the biobank and its customers while biobanks in resource-limited environments often completely lack sufficient resources (Soo et al., 2017) and are restricted to record keeping using general software such as Microsoft Access databases and Excel spreadsheets, making sample management and recording a time consuming and error prone process (Mendy et al., 2015).

This is especially problematic as the quality of a sample is dependent on its intended usage, meaning that the quality parameters of relevance to a sample collection may change over time as new technologies become available. Emerging technologies thereby introduce an element of uncertainty and a need for extensive record keeping covering not only variables of known importance but also potentially relevant pre-analytical variables. Current developments in long read sequencing techniques such as the MinION (Jain et al., 2016) and PacBio (Rhoads and Au, 2015) are one such example as DNA fragment length is becoming increasingly important. With next generation sequencing (NGS) dominated by technologies producing reads shorter than 500 bp reads, DNA fragmentation in stored samples was a minor issue in comparison to purity, yield and amplification success. As long read sequencing technologies are now becoming increasingly popular and reaching read lengths above 100 kb the DNA fragmentation is suddenly emerging as a significant issue as current treatment methods in biobanks routinely shear samples to 5-35 kb depending on the methods and procedures used by the biobank (Malentacchi et al., 2015). Being able to track the processing history of samples is therefore critical for biobanks aiming to provide high-quality samples for long-read sequencing techniques as minor alterations to the process may significantly influence the quality of samples. Extensive record keeping and provenance of data are therefore essential for any biobank hoping to provide samples for future technologies as well as currently established analytical techniques. The Janus biobank in Norway provide one compelling example of such administrative foresight with its collection of blood samples from 1973 to 2004 (Langseth et al., 2016) that has been retroactively evaluated as suitable for measurement of microRNA (Rounge et al., 2015), an analyte not discovered until 1993 (Lee et al., 1993; Wightman et al., 1993) and not widely recognised as being of importance until the early 2000s (Lagos-Quintana et al., 2001).

Reproducibility has emerged as a minimum acceptable standard in the minds of most researchers, functioning as a proxy for replicability to determine if the appropriate scientific conduct has been observed (Peng, 2015). From a computational perspective, there are three main elements to a reproducible and replicable study: (i) the raw data from the experiment must be available, (ii) the statistical code and documentation to reproduce the analysis must be available, and (iii) a correct data analysis must be performed (Leek and Peng, 2015).

In a biobank where samples are divided into aliquots and stored properly, reproducibility can be extended to cover most of the pre-analytical phase as well, allowing researchers to reproduce much of the processing and experimental phase of an experiment as well as the computational phase as long as sample aliquots are still available. While reproducibility ensures that a study has been documented appropriately, it does however not guarantee that the results are scientifically relevant. For results to be relevant, they must also be replicable, meaning that the results must be robust enough that other researchers targeting the same scientific question can generate consistent results using a similar setting.

Various reporting standards and quality management standards are being advocated by different stakeholders for improving reproducibility and replicability in research, (Freedman et al., 2015; Freedman and Inglese, 2014; De Souza and Greenspan, 2013; Simera et al., 2010). There is, however, a considerable number of guidelines being authored with limited evaluation regarding whether they are being implemented in research (Simera et al., 2008). The implementation of standards is costly and in a survey among members of the French biobanking organisation Club 3C-R with 48 responders and an 87 % response rate (di Donato, 2014). A majority of the respondents responded that certification had a negative or negligible impact on the financial performance of the biobank (70 % of the respondents) and a majority of the respondents did not see an increase in scientific collaborations for the biobank due to the adoption of standards (69 % of the respondents did not report any gains in this area). Despite the lack of economic gain 96 % of the respondents replied that they perceived that standardisation had improved the functions of the biobank. With improvements in training, documentation and traceability being the most commonly cited benefits. In other words, standardisation does improve the overall quality of resources but does often not provide an immediate economic or socially beneficial impact to the people performing the work to implement them.

This lack of alignment between the overall needs of the research community and incentives for individual stakeholders is a significant obstacle for improving quality in research and particularly pertinent in environments where funding is insufficient in environments lacking the digital platforms necessary to manage the increased administrative overhead. Providing a comprehensive IT platform for research integrated biobanks is therefore of vital importance to support the implementation of quality management tools in low and medium income countries (LMIC), fundamental research and for non-human biological research.

1.3 The role of bioinformatics in research

Bioinformatics is a field of research with roots in the application of computational science to solve issues in biology, medicine and chemistry (Ouzounis and Valencia, 2003). As a consequence, there is an inherent conflict between the perspective of bioinformatics as a service providing the means for other research and bioinformatics as an independent field of research developing theoretical frameworks and new methods on its own.

These two interpretations of bioinformatics are both viable but may cause significant friction unless recognised and actively managed. The Swedish University of Agricultural Sciences (SLU) and Uppsala University provide an informative example of this conflict of interest documented by external reviewers in the Quality and Renewal reports conducted by Uppsala University in 2007 and 2011. In 2007 Uppsala University conducted its first external Quality and renewal review (Nordgren and Uppsala universitet, 2007) combining a qualitative review by external experts with bibliometric data for each department. In the review, the Linnaeus Centre for bioinformatics was ranked as the top department in Life Science for citations per publication compared to the global average of the field in the bibliometric evaluation component.

Due to its interdisciplinary nature, the centre was qualitatively evaluated by the biology panel as well as the information technology panel. Both panels described the centre as an internationally competitive research environment and of central strategic importance to the university, with the Information Technology panel adding that it could demonstrate how Uppsala University deals with the development of a promising new field. Recommendations for further development was to develop an improved IT infrastructure (both panels) and that "the centre should be sensitive to the needs to the local biological community" (the biology panel). In the following survey Quality and Renewal 2011 (Uppsala universitet, 2011), the centre had been shut down and researchers dispersed across Uppsala University and SLU. The biology review panel supported this decision as it ensured that bioinformaticians were available to assist researchers in other fields while it was heavily criticized by the information technology review panel as the field was under an immediate risk of becoming synonymous with software support for the short-term needs of biology projects, leading to a wasted strategic opportunity.

Given the description of conflicting interests in the reports of both panels, it is clear that everyone was aware of the inherent conflict of interest between research and support in an interdisciplinary field like bioinformatics, despite this, it seems that no workable compromise could be reached. Similar concerns are also described in older documents such as the guidelines for the training of biologists for the 21st century provided by the US National Research Council (National Research Council (US) Committee on Undergraduate Biology Education to Prepare Research Scientists for the 21st Century, 2003), suggesting that even as influential stakeholders are well aware of the conflict, it is hard to manage in practice.

Within the field of bioinformatics, there is also a significant difference between researchers trained in bioinformatics from a Life Science perspective and a Computer Science perspective. In Sweden, this is even reflected in the official government statistics as Statistics Sweden assign different codes to the degree depending on if the bioinformatics degree is obtained as a subfield of Computer and Information Sciences or Biological Sciences (Pettersson and Söder, 2016). With this added differentiation between computer-oriented and life science-oriented bioinformatics the field can be characterised in three different ways:

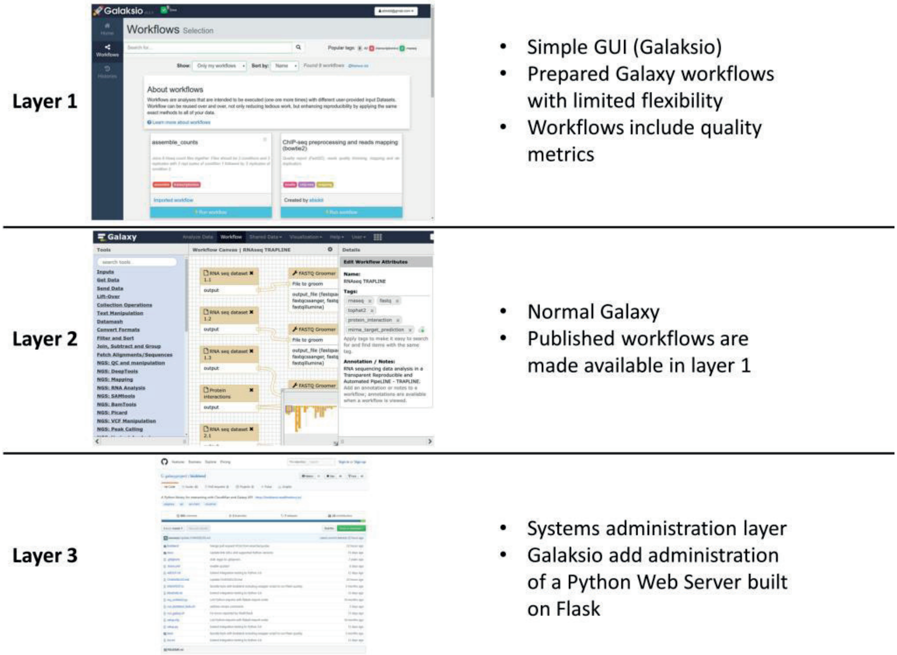
- As an independent research subject pushing the boundaries of human knowledge by the progress of its own.
- As a tool supporting the development of knowledge in other sub-fields of Life Science.
- As an interface to computer science providing access to the computing resources necessary to understand the complex interdependencies, we see in living systems.

When setting up research groups devoted to bioinformatics, it is therefore necessary to have a clear vision of how these different roles will be approached. As needs vary depending on the location, it is necessary for a system like the eB3Kit to provide a flexible approach so that a single technical platform can support a wide variety of use cases. This is achieved by dividing workflow management into several layers where researchers with little bioinformatics training can use a simplified interface centred around workflows covering routine procedures while trained bioinformaticians can focus on more complex support tasks or their own research.

A research centre looking to minimise its investment could, for example, remain largely dependent on prepared workflows available in the Galaksio environment, relying on the eB3Kit and Galaxy communities to provide workflows and obtain advanced bioinformatics support by collaborating with external partners. On the other hand research institutions with access to highly skilled bioinformaticians may use the platform to automate routine applications to maximise the time available for bioinformatics experts to develop new tools or perform their own research in bioinformatics.

To capitalise on the potential networking effect of research groups working on a technical platform where workflows can be exchanged and reused, we decided to build the bioinformatics module on an already established workflow management platform in the form of the Galaxy Workflow management system (Goecks et al., 2010). Thereby not only connecting eB3Kit users with each other but also providing them with an existing user community allowing our users to both download and share their workflows with an already established community.

A similarly layered approach is available in the intersection between bioinformatics and computer science. Most bioinformaticians are comfortable with high-level languages and have been slow to adopt new technology such as Hadoop based distributed computing systems (Oliphant, 2016). The Galaxy workflow management system can be augmented by several modules allowing researchers to run Galaxy workflows but send computation-intensive tasks to external computation resources using Cloudman (Afgan et al., 2010), Bioblend (Sloggett et al., 2013) and Pulsar (Afgan et al., 2015). A research group with trained bioinformaticians but limited skills in computer sciences can thereby develop new tools but only deal with advanced computing solutions when necessary and by connecting to systems hosted by external experts (figure 2).



Layer 1

- Simple GUI (Galaksio)
- Prepared Galaxy workflows with limited flexibility
- Workflows include quality metrics

Layer 2

- Normal Galaxy
- Published workflows are made available in layer 1

Layer 3

- Systems administration layer
- Galaksio add administration of a Python Web Server built on Flask

Figure 2. The layered approach enabled by Galaksio and Galaxy.

Resolving the question of if a specific research institution should treat bioinformatics as a service or as a research field is beyond the aims of this thesis. Instead, we have devised a technical solution based on available technology to accommodate research groups with expertise spanning from a lack of dedicated bioinformaticians to advanced institutions with qualified staff scientists or complete research centres dedicated to bioinformatics with a limited amount of time available to support other researchers.

1.4 Approaching ELSI limitations in data sharing

From an informatics perspective, ethics and law are limitations to the efficient transfer of information. The extent of these constraints are under constant debate and revisions (Litton, 2017; Anon, 2015), but researchers must consider these limitations just like they do with limitations in connectivity, computing power or storage persistence.

As ethical and legal considerations are an important limiting factor in data sharing (van Panhuis et al., 2014) there has been extensive collaboration between work package 1 (ethics) and work package 4 in the B3Africa project. This close collaboration is motivated by a number of factors regarding the implications of bioethics on the functions of the eB3Kit and its bioinformatics component:

- There is no evidence that dedicated ethics experts provide an efficient and consistent safeguard for the public (Scherzinger and Bobbert, 2017; Nicholls et al., 2015; Abbott and Grady, 2011; Angell et al., 2006).
- Projects may be hijacked or mismanaged as it hard for people with little own knowledge to evaluate the relative competence among researchers in a field. Especially in a value-based research field like bioethics it is hard to find the appropriate authority on relevant topics (Satel, 2015; Powers, 2005; Crosthwaite, 1995).
- Fines and uncertain legal conditions make researchers unwilling to share data without extensive reviews and legal negotiations (van Panhuis et al., 2014).
- By exclusively relying on dedicated ELSI-experts it becomes hard, if not impossible, to create innovative solutions supporting the management of ELSI-questions (personal observation).

It is therefore important that at least a minority of practitioners of natural science are making an effort to engage in developing methods to deal with ethical and legal questions. This is especially important in data-intensive research (Dove et al., 2016) and it is evident from studies in the UK that academic chief investigators are prone to submit erroneous applications for ethical review (Angell and Dixon-Woods, 2009) and that dedicated ethics officers do not substantially improve the quality of submissions (Dixon-Woods et al., 2016).

From the B3Africa project perspective there are two important objectives to achieve in regards to ethics and the management of personal data:

- To promote sharing and collaboration to encourage efficient utilisation of samples and data.
- To do this in an ethical, legal, and above all responsible manner that maintains public trust and respect for the individual as well as communal rights.

The article Legal & ethical compliance when sharing biospecimen (Klingstrom et al., 2017) provide an overview of key concepts on how extra-legal means bridge the gap between the national legal systems and advocate a bottom up-approach where continuous collaboration creates a basis for mutual understanding and minimises the waste created when protective measures turn into unnecessary administration (Salman et al., 2014). These conclusions have contributed to the development of the B3Africa project and enable it to take a complementary approach to the establishment of cutting-edge biobanks funded by British and American interests in the H3Africa project(Klingstrom et al., 2016).

The H3Africa project has generated significant documentation on ethical and legal frameworks in Africa applicable to major research centres funded, generating significant support at the locations supported by the project (de Vries et al., 2015; Adoga et al., 2014; The H3Africa Consortium et al., 2014). The B3Africa project follows a different approach as it targets smaller biobanks providing the IT infrastructure necessary to turn sample collections into operational biobanks with controlled storage of samples. Smaller research groups with lesser resources will, therefore, be exposed to the challenges of bridging national regulations of ethical and legal questions by the B3Africa project which must be reflected in the technical infrastructure as well as the capacity building components of the project.

The European Union General Data Protection Regulation (EU-GDPR) will significantly alter how digital data may be stored or processed within Europe and may also be of help to clarify regulations abroad. It provides, for the first time, a homogeneous implementation of binding law over a large number of countries. It should, however, be noted that despite the similarities between programming code (binary code) and DNA which is written in four letter code (quaternary code), the EU-GDPR only covers digital data as the European Union does not possess legal competence over personal information in other forms than digital format.

The EU-GDPR means that for data being transferred for analysis or storage we now, for the first time, have a resource with standardised terminology accepted by a large number of countries that will be applicable for all data handled within the EU. By connecting the technical solutions implemented in the eB3Kit to the legal concepts mentioned in the EU-GDPR it becomes easier for legal and ethics experts to evaluate applications within the EU and by comparison to the EU with other jurisdictions as well.

1.5 Ontologies to manage data

Formal ontologies became popular in the 1980s as a way of specifying content-specific agreements for the sharing and reuse of knowledge in computer science (Gruber, 1995) and its philosophical roots go even further back to the 17th century with philosophers investigating the concept of being (Lawson, 2004). The application of ontologies in computer science, and later bioinformatics, does however tend to be more pragmatic in its approach, using it to convert real-world knowledge into formal specifications that allow us to share conceptualisations that define objects, properties of objects, and relationships between them in a manner that can be interpreted by computers (Smith, 2004; Chandrasekaran et al., 1999). Alternatively, using a less formal language, it provides us with the means to convert the ambiguous language we use in our normal life into strict definitions suitable for computations and data modelling. For a comprehensive practical review of different kinds of ontologies and their role in biomedical research Bogumil M. Konopka has produced a thorough review of the practical application of ontologies in Life Science (Konopka, 2015) along with a guide to different types of ontologies based on a classification scheme developed by Gómez-Pérez et al (Gomez-Cabrero et al., 2014).

It should, however, be noted that despite the formalistic ambitions of the field. Many ontology-related terms may possess different meanings in different domains (Smith et al., 2006) and even highly successful ontologies such as Gene Ontology (Ashburner et al., 2000) and The Unified Medical Language System (UMLS) (Bodenreider, 2004) suffer from issues regarding internal inconsistencies and formal integrity (Geller et al., 2009; Ceusters et al., 2005; Smith et al., 2004). Given the difficulties of appropriately describing ontologies outside of their own formally defined domain a comprehensive description of the field is beyond the scope of the thesis, and we will instead focus on the three kinds of ontologies relevant to the MolMeth and eB3Kit projects.

1.5.1 Formal *is-a* hierarchies

Formal *is-a* hierarchies are familiar to most researchers in the form of taxonomies. A formal *is-a* hierarchy forms a tree where all instances of a subclass are also a member of the parent(s) of the subclass. Using taxonomies as an example, this means that if we know that an animal belongs to one of the 13 subspecies of *Canis lupus.* Then we can also automatically infer that the animal belongs to the genus *Canis*, the family Candiae and all other higher levels associated with the class in its scientific classification.

In molecular biology and bioinformatics, the Gene Ontology (GO) (Ashburner et al., 2000) is perhaps the most famous implementation of a formal *is-a* hierarchy with three separate ontologies that with increasing granularity describe the biological process, molecular function and cellular component of genes and their products.

1.5.2 Vocabularies, glossaries and thesauruses

A vocabulary is a collection of words available within a domain. Such collections may seem primitive but can still provide value such as in the case of the 300-400 words long list that the director of Houghton Mifflins education division, William Spaulding, believed every American six-year-old should know. This vocabulary was then used to define the words Dr Seuss could use for the now famous children story about the Cat in the Hat (Nel, 2004).

A glossary is a vocabulary where natural language definitions have been added to each term providing entries such as:

Glossary – The ontological classification of a dictionary.

A Thesaurus is similar to a glossary but can include any number of semantic relations linking it to other resources, thus making it significantly larger and more complex to maintain than vocabularies and glossaries. For example, The Unified Medical Language System (UMLS) Metathesaurus published by the US National Institute of Health contain over two million concepts based on 150 electronic versions of classifications, code sets, thesauri, and lists of controlled terms in the biomedical domain according to its factsheet (National Institutes of Health, n.d.). Table 1 contains examples of the different kinds of non-hierarchical ontologies describe above.

Table 1. *An example of how the term "Cancer Biobank" can be described in different kinds of ontologies.*

Ontology class	Vocabulary	Glossary	Thesaurus
Term	Biobank	Biobank	Biobank
Natural language definition		A repository of biospecimen - including tissue samples, fluids ...	A repository of biospecimen - including tissue samples, fluids ...
Synonyms & abbreviations			Cancer Biorepository, Tumour Bank
Concept Unique Identifier			C1519671
NCI Thesaurus Code			C15863
Semantic Type			Health Care Related Organization
External Source Codes:			NCI Thesaurus Code
Other properties			(none)
URL to Bookmark			https://ncim.nci.nih.gov/ncimbrowser/ConceptReport.jsp?dictionary=NCI Metathesaurus&code=C1519671

1.5.3 The EXACT definition of laboratory protocols

Experiment ACTions (EXACT)(Soldatova et al., 2008) is a hierarchical ontology with additional non-hierarchical relations, thus possessing an overall formal *is-a* hierarchical structure but with additional relations attached to classes like in a Thesaurus. It aims to provide an exact representation of laboratory actions to support reproducible protocols using the EXACT/OBI ontology and the full automation of research using EXACT/EXPO in projects such as the robot scientist (King et al., 2009). In its second iteration the ontology depends on three top-level classes based on other existing ontologies:

- **EXACT2: descriptor of experimental actions** which describes equipment and experimental conditions.
- **OBI:Process** describing experimental actions, procedures and protocols.
- **IAO: information content entity** which provides meta-information about the document title, author and license.

In practice this allows the researcher (or computer) to read a protocol describing the exact actions taken by a researcher when performing a specific protocol such as in the example below. The first nine lines describe metadata and relevant starting conditions. After that, each experiment action (a subclass of OBI: a process called EXACT2: experimental action) is described together with all information necessary to describe the action. Creating an exact but very cumbersome representation of the experimental actions performed when executing a laboratory protocol in the following manner:

```
Protocol
DC title: partial protocol for the preparation of
Saccharomyces cerevisiae competent cells
DC author: Wayne Aubrey
DC organisation: Aberystwth University
status: draft
DC submission date: 15 January 2008

operating procedure: grow yeast culture
pre-condition: sealed yeast colonies plate in cold room
pre-condition: YPD media bottle in cold room
experiment action: move
object: YPD media bottle
start location: in store
end location: in laminar flow hood

experiment action: move
object: conical flask
start location: in store
end location: in laminar flow hood
```

```
experiment action: move
object: sealed yeast colonies plate
start location: in cold room
end location : in laminar flow hood

experiment action: add
component 1: YPD medium
volume: 50ml
start container: YPD media bottle
end container: 500ml conical flask
equipment: pipette

experiment action: rename
old name: 500ml conical flask
new name: YPD conical flask

experiment action: add
component 1: single yeast colony
volume: small volume
precision: N/A
start container: yeast single colonies plate
end container: YPD conical flask
equipment: inoculating loop

experiment action: rename
old name:YPD conical flask
new name:yeast culture flask

experiment action: incubate
object: yeast culture flask
equipment: shaking incubator
rpm: 200
temp: 30C
time interval: 12-24h
```

In a normal materials & methods article, the above text would perhaps, have been described as ”Yeast cultures were taken from cold storage and a single yeast colony was added to a 500 ml conical flask and incubated for 12-24 h in a shaking incubator (30 C, 200 rpm)”. For a website like the Molecular Methods database, an ontology like EXACT, therefore, provides a potentially valuable back-end to categorise experimental actions and objects but requires an adapted front-end to make the format more readable and convenient for human users.

1.5.4 Coordination among ontologies

Ontologies suffer from the same weakness as standards in the sense that they depend on the acceptance of the community to become useful. To combat fragmentation of knowledge represented in ontologies two major efforts are ongoing in Life Science (Konopka, 2015).

The Unified Medical Language System (UMLS) unifies a large number of different source vocabularies using a word index. Each word, or "preferred term" is then given a large number of semantic relationships (including synonyms) connecting the classifications, code sets, thesauri, and lists of controlled terms compiled in the UMLS Metathesaurus (Aronson, 2001; Schuyler et al., 1993). The Metathesaurus aims to create meaningful relationships between already existing vocabularies, meaning that it does not have an ambition of structuring data or reducing redundancy among overlapping vocabularies.

The OBO Foundry is a collaborative development creating a more unified *is-a* hierarchy of terms covering all domains of scientific research (Smith et al., 2007). Its overall aim is to pursue a strategy where ontologies are accepted into the foundry provided that they conform to the shared principles of the OBO Foundry.

The eB3Kit relies on ontologies based on OBO principles (Brochhausen et al., 2013) to federate and exchange data using the MIABIS (Merino-Martinez et al., 2016) and clinical data will to a great extent by local stakeholders using components of The Unified Medical Language System (UMLS).

2 Aim of thesis

The thesis aims to enable and to accelerate research cycles (see figure 3). This overall goal is achieved by helping researchers to more efficiently transfer and utilise information. The specific contributions presented in this thesis can be divided into three core components:

- The development of the Molecular Methods database.
- The creation of the eB3Kit bioinformatics component.
- Investigations into the interaction between researchers and technology platforms to evaluate how to address the needs of the end users.

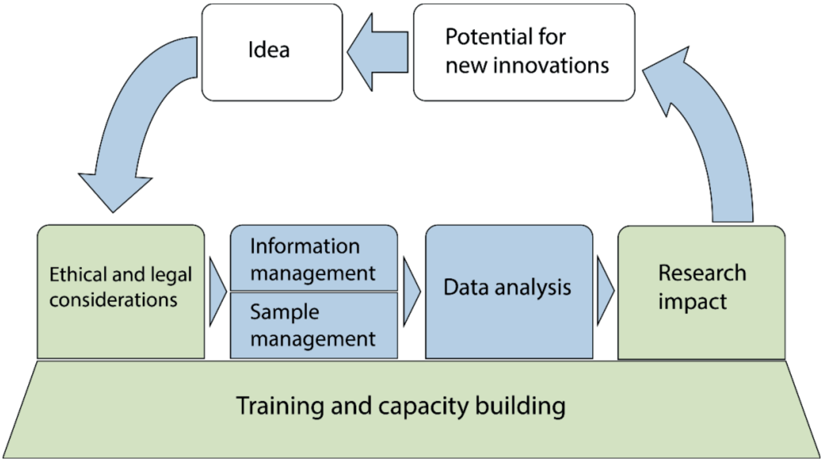

Idea
Potential for new innovations
Ethical and legal considerations
Information management
Sample management
Data analysis
Research impact
Training and capacity building

Figure 3. The research cycle model adopted by the B3Africa project. The eB3Kit provides an informatics platform covering the technical solutions necessary to manage information, samples and data analysis as well as share this information with external stakeholders on a secure platform (blue components). The social components of the B3Africa project (green) serves to integrate this technical capacity into the local research environment by providing training and expertise that allow researchers to use the technology, overcome ethical and legal barriers and make the most of their research. These two components strengthen local research communities which can use the platform to innovate and produce new innovations or scientific discoveries (white).

3 Summary of research & results from papers I - V

3.1 Paper I: Workshop on laboratory protocol standards for the molecular methods database

The Molecular Methods Database was created in 2009 and initially funded by the EMERALD project (Beisvåg et al., 2011) as a central resource for methods and protocols focusing on microarray-based technologies. Throughout its lifetime the database has undergone several revisions changing the database from a software application to a web server and changing how the user interacts with the database (table 2). Since its third version which was created by the author based on the conclusions of paper I the protocol database has supported several European and Swedish research projects, hosting protocols used in numerous scientific publications (Daebeler et al., 2017; Medhat et al., 2017; Allen et al., 2016; de Koning, 2016; Marikanty et al., 2016; Saxena et al., 2013; Librizzi et al., 2012) and more than 40 000 user sessions by external users have been logged since 2013 when Google Analytics was implemented on the website.

Table 2. *Iterations of MolMeth. Work on version 3 and 4 are covered in this thesis with data collected based on version 4.*

Version	Features	Reason for replacement	Code platform
1	Desktop application developed to exchange protocols and experience between laboratories.	Rapid development of Ajax and Web applications made software solutions outdated as purely web based applications offer superior cross platform compatibility and freedom of movement.	Unknown
2	Web based platform with focus on semantic integration with other websites.	Modern content management systems and other factors out-competed Ruby on Merb which made the situation unsustainable unless a full development team was to be recruited.	Ruby on Merb
3	Web based platform with focus on semantic integration with other websites.	The migration to Drupal 7 reduced maintenance costs but the requirements devised by informaticians and ontologists created a learning curve that laboratory researchers were unwilling to commit to. A new interface built on the existing database was therefore necessary.	Drupal 7
4	Web based platform with semantic integration and large focus on user friendliness and search engine optimisation for maximum visibility.	Current version, all technical objectives are achieved, and researchers find the website by searching for protocols of interest. A lack of sufficient incentives to publish protocols is however evident and restricts growth.	Drupal 7

Version 4 (2013-current) and version 3 (2012-2013) were supported by the Swedish biobanking infrastructure (www.BBMRI.se) and the EU FP7 project Affinomics (www.affinomics.org) producing antigen targets and novel binders for a variety of purposes. Antibodies are a fast moving area where major batch effects are known to occur, causing issues with reproducibility and the calibration of protocols (Baker, 2015). Researchers reliant on antibodies can therefore greatly benefit from the rapid dissemination of technical information about viable assays and the suitability of new batches of antibodies being produced. At the same time, the publication of research projects is taking longer and longer as more and more data is required for publication (Vale, 2015), meaning that the feedback cycle for new technology or quality issues is slowing down. Encouraging the public dissemination of technical data and the formation of loose collaborative networks with researchers alerting each other of issues as well as troubleshooting methods by using a shared web portal was therefore deemed a potential solution to the issues.

In biobanks, the research environment is more stable, but there are collaborative benefits in having similar workflows adapted at a large number of biobanks. National networks such as the national Biobank and Biomolecular Research Infrastructures (BBMRI) in Europe (Yuille et al., 2007) and the Canadian Tumour Repository Network (CTRNet) (Matzke et al., 2012) therefore often engage in training and support efforts to promote the development of biobanks abroad as well as on the national level (Matimba et al., 2016; Cohen et al., 2013). It was therefore deemed important by BBMRI.se to provide a technical platform where researchers could publish and compare protocols without constraints regarding pre-publication review and estimates of scientific impact required by peer-reviewed publications. Despite the significantly different challenges faced by the two projects, both projects made provisions to use MolMeth to create active user communities sharing technical information between experts to improve productivity and reduce error rates within the two fields.

3.1.1 Description of the Molecular Methods database

MolMeth relies on an open submission system where users may register on the site and immediately start publishing protocols. A commercially available anti-spam system is used to categorise submissions as spam, or non-spam and administrators evaluate new protocols post-publication to conduct basic quality checking as well as re-categorising spam/non-spam if necessary.

Registered users can comment and rate protocols which help users to find the highest rated protocols in each sector. Protocols are never removed from MolMeth, but a revisioning system is used to provide researchers with a simple but useful tool to update and manage protocols.

Users can update any protocol they have submitted and also grant other members the right to edit protocols. When a protocol is edited, a revision is created. Each revision is saved with a timestamp and accessible by clicking the “revision” tab at the top of a protocol. By referring to specific time stamps, users can ensure that they always cite the correct iteration of a protocol while still providing other researchers with access to both older and more updated versions of the protocol. Outdated or erroneous protocols can thereby be labelled as such, but older versions are still available to researchers who wish to access such content later.

3.1.2 Current status and sustainability

MolMeth was first established in 2009 but has undergone several major overhauls where protocols have been re-curated, and user accessibility improved. Reading and posting protocols is free and all protocols are published using a Creative Commons 3 license, which means that all material can be freely distributed as long as it is properly cited.

The two versions developed throughout the work of this doctoral project has been built using Drupal 7, an open source Content Management System (CMS) to ensure the longevity of the site as extensive documentation and commercial support is available for future development and maintenance. The website is currently being run on a hosted by the Swedish University of Agricultural Sciences, and backups of the site are being made daily. As the website relies on standard off the shelf software, there is a minimal need for the in-house development of new tools and software the as these are provided by the Drupal user community.

3.1.3 Traffic and activity

Rebasing MolMeth to Drupal 7 allowed provided support for integration with Google Analytics and systematic tracking of visitors started in 2013. During the four-year period between 13, May 2013 and 13 May 2017 almost 12 000 sessions per year have been registered using Google Analytics. The by far most common way to find MolMeth is by search queries on search engines (65.4 %) with direct links (16.5 %) and referrals (14.0 %) providing most of the remaining visits (figure 4)

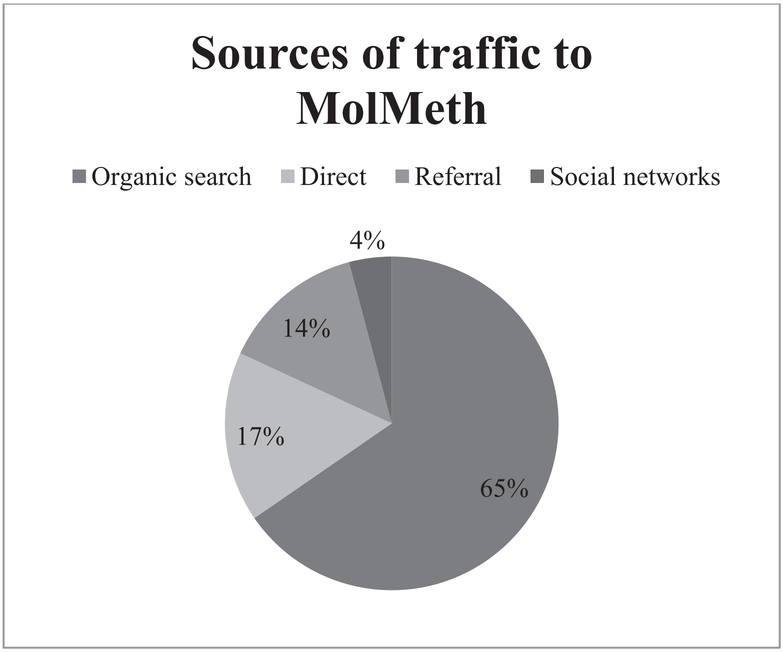


Figure 4. Sources of traffic to www.molmeth.org.

MolMeth is designed to make information easily accessible using search engines like Google and users are expected to arrive immediately at the right protocol rather than visiting the front page. This is reflected in the user behaviour where the front page is only the third most visited page (table 3). Protocol popularity is very unevenly distributed and among the 584 unique URLs accessed by searchers the ten most influential pages have received 26.5 % of the search engine hits (table 3).

Table 3. *During the period 13th May 2013 to 13 May 2017 a total number of 46 966 sessions by 38 760 different users were measured by Google Analytics. Of the 30 671 sessions with a measured target page, 26.5 % targeted one of the ten most popular web pages on MolMeth, and the remainder were spread out among 584 valid* URL:s *at MolMeth.*

Target page	Sessions	Pages/session
/protocol/processing-blood-specimens	2 697	1.13
/protocol/collecting-plasma-whole-blood	2 092	1.15
/ (front page)	1 258	4.05
/protocol/situ-proximity-ligation-assay-pla-protocol-using-duolink	1 196	1.16
/protocol/situ-proximity-ligations-assay-pla-protocol	890	1.22
/protocol/collecting-and-processing-saliva	868	1.26
/protocol/vivo-biotinylation-protocol-avitagged-proteins	844	1.25

Target page	Sessions	Pages/session
/protocol/protocol-decontamination-instruments-reagents-and-laboratory-areas-challenging-dna-analysis	707	1.22
/protocol/click-chemistry-antibody-dna-conjugation-protocol	700	1.23
/protocol/collection-white-blood-cells	699	1.17
Total number of sessions in top 10	8 138	
Total number of sessions in measurement	30 671	

Age distribution shows a strong bias towards younger researchers with 86 % of site visitors being 44 years old or younger (figure 5). The gender distribution is relatively equal with a slight overrepresentation of female visitors (53 %).

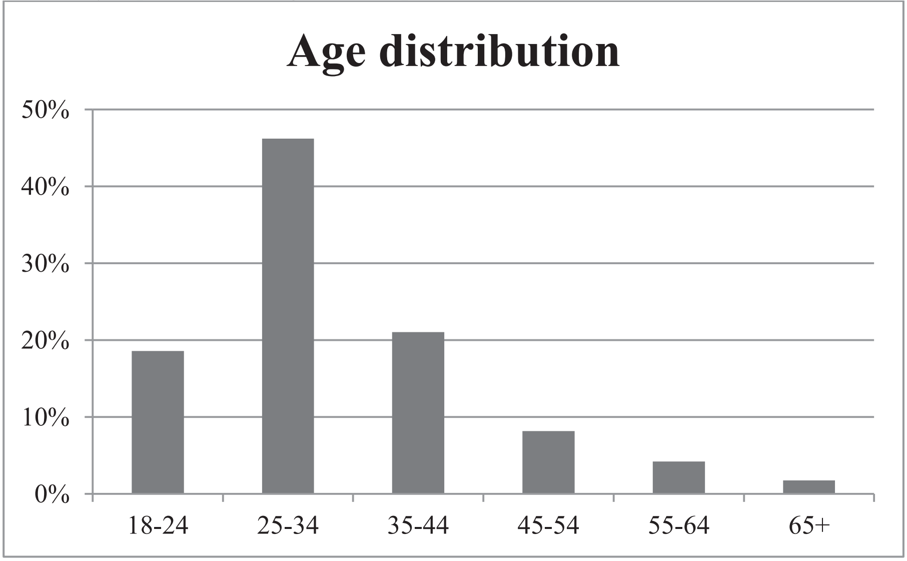


Figure 5. Google Analytics was able to provide age data about 28.8 % of site visitors.

The USA is the most common location for visitors to MolMeth (27.7 %) followed by Russia (9.6 %) and Sweden (4.3 %). Upon deeper analysis using Google Analytics, it is, however, clear that usage patterns in Sweden and Russia differ from normal usage patterns (table 4). In Sweden, this is related to a high number of partners providing direct links to the front page (www.molmeth.org) and high activity by partners at SLU and Karolinska Institutet to publish protocols. In Russia, a careful analysis (data not shown) show that the number of genuine visits most likely are only a few hundred while the rest are caused by various spamming programs manipulating Google analytics. This includes an attack directed via the social networking site Reddit which contributed almost half the visits generated by social media during the survey period.

Table 4. *Usage patterns across the globe. Web crawlers (mainly hosted in Russia or from an unknown location) disproportionally visits the front page compared to normal user behaviour.*

Country (100 % of total sessions)	Sessions	Front page visits	Percentage
United States	13022	1712	13.1%
Russia	4489	4 491	100.0%
Sweden	3469	2024	58.3%
United Kingdom	2648	388	14.7%
India	1851	60	3.2%
Germany	1764	280	15.9%
(not set)	1380	547	39.6%
Canada	1228	70	5.7%
France	1073	159	14.8%
Japan	1025	107	10.4%
Total top 10	31949	9838	30.8%
Total	46966		

MolMeth hosts over 10 000 user sessions per year and its reliance on the Drupal content management system with limited customisation means that it can be operated at a very low cost to handle periods with limited or no funding. Based on user feedback the concept of an open protocol platform is also appreciated by researchers looking for support to find protocols and standard operating procedures. Generating the user engagement necessary to create a sustainable web 2.0 platform with an active user community publishing and discussing protocols have however been harder than anticipated. Attempts have been made to lower the threshold for publishing protocols on MolMeth by creating a simple uploading process sacrificing much of the structured data fields in favour of free text fields which can be filled in by copy and pasting from Word documents.

3.1.4 Representation of protocols using the EXACT ontology

The EXACT ontology (Soldatova et al., 2008) is closely related to the Robot Scientist project (King et al., 2009) and can in a controlled environment provide a perfect representation of laboratory protocols. In real life experiments using protocols and machine learning techniques, the ontology is estimated to cover 85 % of the typical experimental actions described in previously unseen protocols (Soldatova et al., 2014).

Both MolMeth and EXACT (Soldatova et al., 2014) have however been struggling to find a way to represent protocols in a manner accepted by biologists. Despite significant effort to simplify user interactions with the systems uptake of ontology-based protocols have failed to achieve significant traction among the intended user community. In MolMeth version 3 a simplified annotator tool was created to help researchers separate metadata, information about preconditions/experimental conditions and the experimental actions described by the protocol (Trollvad et al., 2012). This annotation allowed MolMeth to represent all relevant information contained in the experiment actions of an EXACT protocol using approximately one-tenth of the number of lines used in an EXACT protocol (see table 5 for a comparison) provided that writers follow two simple rules:

- Each sentence should only describe a single experimental action.
- Each sentence should begin with the verb describing the experimental action.

Table 5. *Comparison of EXACT, MolMeth V3 text and an example of a free text description.*

	MolMeth	Free text
EXACT		
operating procedure: grow yeast culture	Protocol: Grow yeast cell culture	Materials and methods: Growth of yeast cell culture
pre-condition: sealed yeast colonies plate in cold room	reagent: sealed yeast colonies plate stored in cold room	Yeast cultures were taken from cold storage and a single yeast colony was added to a 500 ml conical flask and incubated for 12-24 h in a shaking incubator (30 C, 200 rpm)
pre-condition: YPD media bottle in cold room	reagent: YPD media	
	equipment: 500 ml conical flask	
experiment action: move	equipment: pipette	
object: YPD media bottle	equipment shaking incubator	
start location: in store	equipment: inoculating loop	
end location: in laminar flow hood		
	Add 50 ml YPD medium to a 500 ml conical flask using a pipette	
experiment action: move		

	MolMeth	Free text
EXACT		
object: conical flask	Transfer a single yeast colony from the plate to the 500 ml conical flask using the inoculating loop	
start location: in store		
end location: in laminar flow hood	Incubate at 30 C for 12-24 h in shaking incubator at 200 rpm	
experiment action: move		
object: sealed yeast colonies plate		
start location: in cold room		
end location : in laminar flow hood		
experiment action: add		
component 1: YPD medium		
volume: 50ml		
start container: YPD media bottle		
end container: 500ml conical flask		
equipment: pipette		
experiment action: rename		
old name: 500ml conical flask		
new name: YPD conical flask		
experiment action: add		
component 1: single yeast colony		
volume: small volume		
precision: N/A		
start container: yeast single colonies plate		
end container: YPD conical flask		

EXACT	MolMeth	Free text
equipment: inoculating loop		
experiment action: rename		
old name:YPD conical flask		
new name:yeast culture flask		
experiment action: incubate		
object: yeast culture flask		
equipment: shaking incubator		
rpm: 200		
temp: 30C		
time interval: 12-24h		

The proposed annotation system did however not increase user engagement in any measurable manner, leading to the development of MolMeth version 4 which allows researchers to copy and paste word documents into the website without any rewriting at all.

3.1.5 Summary

By comparing the initial plans for MolMeth outlined in paper I and research proposals the project has as of yet reached a point with a combination of achieved and unfulfilled objectives:

- Create a platform to make protocols readily available – achieved.
- Create a platform that is sustainable in a technical and financial perspective – achieved.
- Create a platform where structured data regarding laboratory information is made available – objective removed to increase user-friendliness.
- Create a platform with an active and sustainable user community – unfulfilled.

Even with a functional, technical solution and an acceptable number of traffic of visitors looking for information MolMeth has failed to establish an active user community generating content on its own. To better understand the mechanisms behind such a community work on paper II was initiated.

3.2 Paper II: Scholarly publication in the digital age, an investigation into why novel publishing concepts have failed to disrupt the market of scientific journals.

Paper II covers 42 different platforms for social networking aimed at researchers. Out of the 42 platforms, only six websites maintained active user communities that could be defined as "active" in the sense that members got a response from other members active on the site when publishing content. Of the six successful websites, three are focused on helping researchers to access articles of interest and advertise their articles (ResearchGate, Mendeley and Academia.edu) and one obtains members by requiring students participating in the International Genetically Engineered Machine (iGEM) competition to contribute protocols to the Wikipedia like web page OpenWetWare. The two remaining websites, Biostars and myExperiment are targeting bioinformaticians in line with traditions from programming and computer science rather than "wet work" in Life Science.

Based on the evaluation in paper II we conclude that despite significant efforts the scholarly communication in Life Science remain highly similar to how it was conducted in the pre-digital era as devised by Garvey and Griffith (Garvey and Griffith, 1972) (figure 6) but with small modification to how researchers gain awareness of new publications (figure 7).

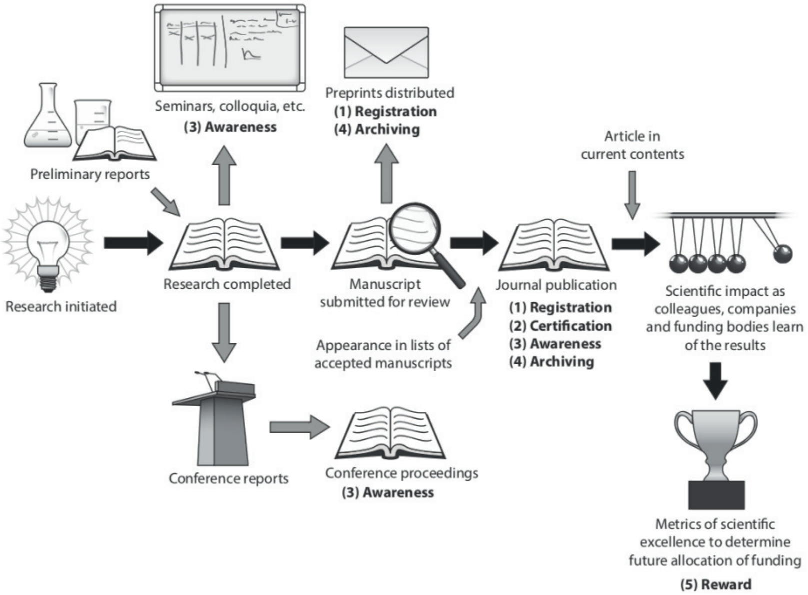

Seminars, colloquia, etc.
(3) Awareness
Preprints distributed
(1) Registration
(4) Archiving
Preliminary reports
Article in current contents
Research initiated
Research completed
Manuscript submitted for review
Journal publication
(1) Registration
(2) Certification
(3) Awareness
(4) Archiving
Scientific impact as colleagues, companies and funding bodies learn of the results
Appearance in lists of accepted manuscripts
Conference reports
Conference proceedings
(3) Awareness
Metrics of scientific excellence to determine future allocation of funding
(5) Reward

Figure 6. Scholarly communication as explained by Garvey & Griffith in 1972 with an overlay of the academic process (marked 1-5) based on Roosendal and Guerts. To justify funding, promotion and recognition researchers are expected to contribute to scientific progress. As a researcher, it is, therefore, necessary to register discoveries (1), have them accepted by peers (2) and present the results to the wider research community (3) in a format accessible to future researchers (4). Reward metrics may be straightforward (*e.g.*, reward for publishing in the 'right' journal, with high acceptance in the community) or implicit (*e.g.*, being awarded a promotion based on a successful track record of conference presentations and publication of significant articles) but all four criteria are necessary to estimate scientific impact and justify the future allocation of resources (5).

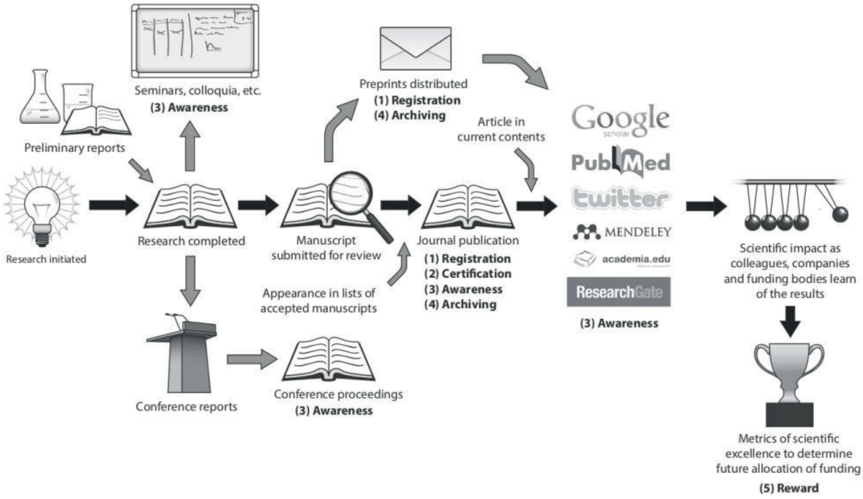


Figure 7. The current status of scientific publishing. Publication in peer-reviewed journals remains the only method to achieve certification, awareness and archiving, but search engines and scientific, social networks provide new channels for researchers to raise awareness of new articles. This makes scientists less reliant on the popularity of the journal to reach potential readers making it more attractive to publish raw data and methods, as long as the content is delivered in a citeable format.

Correlation between the journal impact factor and citation rate of an article is low (Bornmann and Leydesdorff, 2017) and seem to fall further as digitalisation of communication allows researchers to increasingly rely on searchers and social media rather than journal subscriptions to drive article readership, (Lozano et al., 2012) creating opportunities for alternative means of publication and lowering barriers of entry for new journals. Peer-reviewed articles are likely to remain a cornerstone of research, but even the decision of a single researcher to not pursue a traditional journal publication for his pre-print (Coop, 2016) has been commented upon in Nature (Singh Chawla, 2017). Furthermore, citations are increasingly based on DOI numbers, and researchers in high-energy physics are already experimenting with citeable datasets (Herterich and Dallmeier-Tiessen, 2016). Historically high-energy physics has been at the forefront for previous developments in digital publication (Lozano et al., 2012) with the popularisation of preprints (Gentil-Beccot et al., 2010; Till, 2001) which are now accepted and even encouraged in Life Science (Callaway and Powell, 2016; NIH Grants, n.d.) as well. Suggesting that services such as Zenodo (www.zenodo.org) and Figshare (www.figshare.com) which are currently seeing early adoption activity very well may have a bright future ahead of them.

As the survey suggests life science researchers are unlikely to adopt new technology unless there are clear incentives to do so. Given the current focus on articles and citation rates it is therefore not surprising that the most successful websites have been websites which allow life scientists to boost their visibility and their efficiency by marketing their articles and find suggestions on articles to read, thereby reducing the time they have to spend finding information of relevance to them, and making communication faster.

For a website like MolMeth this means that the barrier of entry to the scientific publication market has been reduced and that citeability and the ability of researchers to build awareness of their research are key factors to success. Obtaining DOI numbers and integrating a protocol website with established journals would, therefore, be key factors to success in the further development of MolMeth which is a development pattern currently followed by protocols.io (Teytelman et al., 2016).

3.3 Paper III: Legal & ethical compliance when sharing biospecimen

DNA is one of the most widely studied analytes and can be extracted from practically any sample commonly held in a biobank. Stored samples and derived data must, therefore, be treated with care in regards to privacy, or in the case of many biological samples, ownership rights. This means that ethical, legal and social implications (ELSI) can be a significant barrier when engaging in collaborative projects (The Expert Advisory Group on Data Access, 2014; van Panhuis et al., 2014).

Bioethics is an interdisciplinary subject (Silber, 1982) just like bioinformatics but its historical relationship to life science is complex as the roots of bioethics is not among natural scientists but among people trained in a philosophy who began to teach, write, and profoundly influence life science (Pellegrino, 1999). Bioethicists have always sought a normative role in medical and biological research (Pellegrino, 1999; Silber, 1982) and have been successful in doing so as the field has achieved an almost extra-legal status with bioethicists functioning as experts in research ethics committees and institutional review boards (Abbott and Grady, 2011; Angell et al., 2006). Delegating research ethics to experts has however, despite significant resources being devoted to the task, not provided much of a tangible benefit to the research process (Dixon-Woods et al., 2016; Abbott and Grady, 2011) and also risk alienating the researchers who conduct sensitive research but do not consider bioethics as being a part of their competence (Johnsson et al., 2014).

Paper III was conceived as a collaboration between SLU Global Bioinformatics Centre and the Centre for Research Ethics & Bioethics at Uppsala University to provide better support for researchers who need to deal with ethical and legal compliance as their biobank grows. The resulting paper is a primer for active researchers who need help to deal with early stages of biobank planning and to understand common scientific discourse in bioethics papers.

Key conclusions from the report and implications for the eB3Kit are:

- To accommodate international collaboration, it is necessary to bridge the gap between national legal frameworks. This is usually done by designating experts and organisations who determine if material transfer agreements are able to protect the rights of the donors in accordance with what they could expect when giving their consent for samples to be stored for future usage. – Standardised templates to encourage a more expedient review process has previously been suggested in other projects (Thompson et al., 2014) and the eB3Kit provides a standardised technology platform making not only formatting but also the underlying security measures more standardised.
- Collaboration is substantially more likely to be accepted between nations where the respective authorities have had the possibility to become familiar with each other's customs and traditions. Identifying successful precedents by other researchers participating in collaborative projects can therefore significantly reduce the time necessary to access samples. – More mature biobanks using the eB3Kit can serve as role models for other biobanks and provide a template for biobank governance as well as applications to review boards. By including ethics parameters in the data model, this data can be spread more efficiently within networks of eB3Kit users (Merino et al., 2016).
- Different institutions define terms such as consent, informed consent and broad consent differently. This means that an 'informed consent' at one institution may not be accepted as truly informed by another. Under such circumstances, researchers are likely to face a situation where the strictest interpretation in terms of data protection or privacy becomes the governing one. – Issues of terminology will be a significant challenge and the data model developed by B3Africa (Merino et al., 2016) will need to provide a mapped vocabulary to bridge the gap between terminologies and support comparisons between applications.
- There is a conflict between reciprocity, anonymity and the right to not know. Research must, therefore, be planned and conducted in accordance

with what the donors could reasonably expect when donating their samples and giving their consent. – eB3Kits will be deployed in biobanks at various levels of maturity and networks such as ESBB and BCNet (Mendy et al., 2015) may be more suitable for this kind of early stage support even if the eB3Kit provides a comprehensive hands on training environment similar to how we use the eBiokit for bioinformatics training.

Since the study was completed the European Union General Data Protection Regulation (GDPR) (European Union, 2016) has been published providing, for the first time, a legally binding set of regulations on data protection covering a large number of countries. From a technical perspective many of the terms used in the regulation are however still open for interpretation by expert bodies and there is also room for national legislation making the GDPR somewhat different in each country (Litton, 2017). The training requirements, even for researchers within the EU will therefore remain but the further development of the data model will benefit as it can be partially based on the GDPR.

3.4 Paper IV: Supporting the development of biobanks in low and medium income countries

Paper IV describes the work of the B3Africa consortium, and how social as well as technical aspects are taken into account when building an informatics platform for research integrated biobanks. The project is based on experience from a several different projects contributing to the development of research infrastructure using technology and knowledge developed in several different projects concerning the development of biobanking and bioinformatics in Africa (Bendou et al., 2017; Hernández-de-Diego et al., 2017; Müller et al., 2017; IARC, 2016; Merino-Martinez et al., 2016; Mulder et al., 2016; Mendy et al., 2015; Abayomi et al., 2013; Klingström, 2013; Norlin et al., 2012; Fuxelius et al., 2010).

The overall aim is to provide a robust and sustainable technology platform for biobanks. Currently, active biobanks in low- and medium-income countries have mainly been created to support specific programmes targeting major health issues in the countries hosting the biobank. Consequently, biobanks are often ill-prepared to collaborate independently and maintain operations after the end of a project (Mendy et al., 2015), thus leading to the loss of valuable infrastructure and disassembly of the expert team necessary to run it (figure 8).

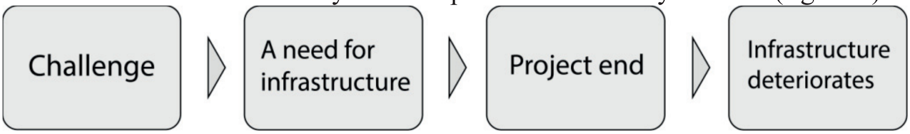


Figure 8. *The life cycle of a biobank if sustainability is not achieved.*

Research groups in resource-constrained settings were particularly vulnerable to loss of key personnel as they are reliant on a small number of individuals and lack the means to replace lost competencies quickly. An appropriately configured informatics system can in most cases negate these issues as data is stored in a structured manner and can be retrieved in a format accessible to non-specialists. This means that even if the data protection officer is lost her work can be retrieved and evaluated during future reviews and application, if the biobank core staff is lost samples will remain accessible and appropriately annotated in the freezers and if the bioinformatician leaves all workflows can not only be replicated but also re-used and applied to new datasets. Under ideal circumstances, a biobank operating eB3Kit should be able to hibernate to such an extent that even if all personnel leave the project, a researcher familiar with the eB3Kit should be able to activate the platform and retrieve data or ensure that precious or dangerous samples are still kept in the biobank. This means that the eB3Kit needs to contain structured data and the necessary tools to:

- Support the collection of samples in compliance with the ethical and regulatory framework.
- Ensure that samples are kept in compliance with the ethical and regulatory framework.
- Allow staff to organise collections, studies and standard operating procedures (SOP).
- Conduct sample acquisition and metadata management in each study.
- Track the processing and storage of samples.
- Handle the retrieval, retention, processing and destruction of samples and/or data.
- Integrate and catalogue external data using MIABIS or other data transfer formats.
- Support the retrieval of relevant data from all the above tasks to conduct data analysis.

The bioinformatics module will have access to all data from the clinician’s request to the start of the analytical examination, thus providing an unprecedented coverage of pre-analytical variables that may influence the analysis. This data access is provided in collaboration with the other work packages as well as built on the experience gained in paper I and III to ensure that relevant data is retrieved from the other components of the eB3Kit and made available for data analysis with appropriate safeguards.

3.5 Paper V: Galaksio, a user-friendly workflow-centric front end for Galaxy

Galaksio provides a workflow-centric graphical user interface (GUI) for data analysis in the eB3Kit and is described in paper V of the thesis. The aim of the tool within the context of the eB3Kit is to reduce the workload of bioinformaticians working at research integrated biobanks but can also be deployed independently as a user-friendly portal to any Galaxy server. Galaksio runs on a Python web server installed on the eB3Kit and connects to the Galaxy workflow management system using the Galaxy API.

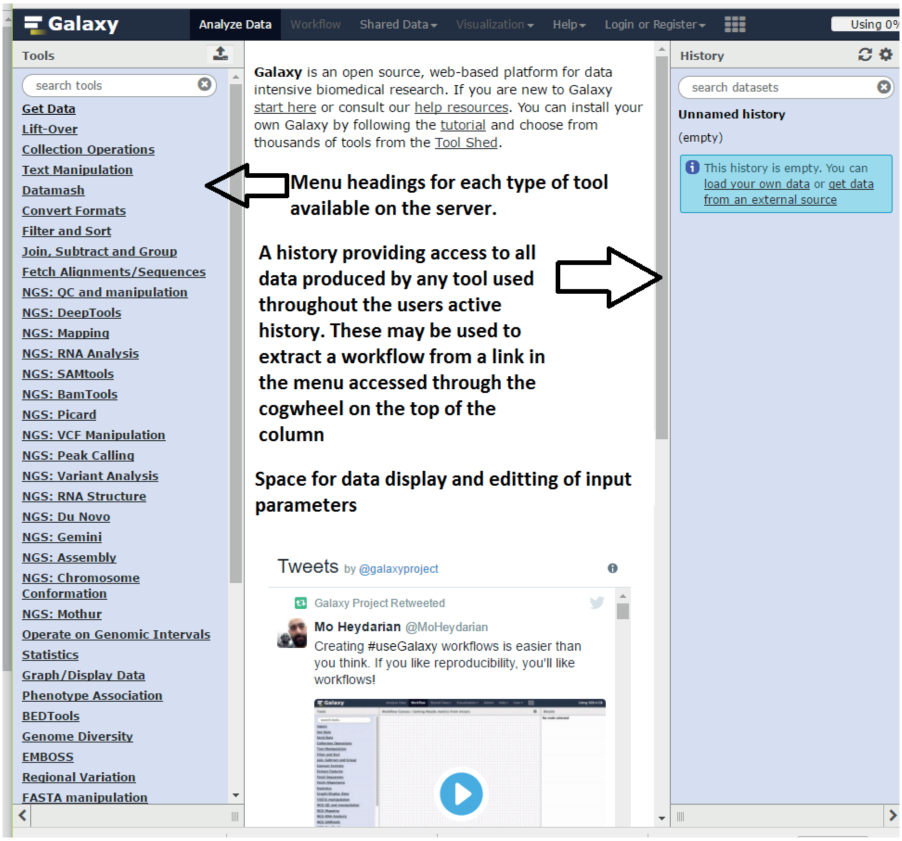
The Galaksio interface is created based on previous experience with Galaxy which has been installed in the eBiokit and used in several training & capacity building projects (Mulder et al., 2016; Atwood et al., 2015; Fuxelius et al., 2010). During this work, it has become evident that the Galaxy workflow management can provide significant benefits when implemented as it allows workflows to be shared and ensures that data is handled reproducibly. A lack of stakeholder analysis has however created similar issues as with the MolMeth database in the sense that researchers acknowledge the overall benefits of the system but do not use it due to a lack of relevant incentives on a more personal level.

Our in-house analysis concluded that a more layered approach would make the system more attractive for the users. Work related to bioinformatics can largely be separated into three distinct tasks:

- Using bioinformatics tools or workflows.
- Developing bioinformatics tools and workflows.
- Creating and managing the infrastructure for data analysis.

The graphical user interface provided by Galaxy provides enables the first two tasks but does not take into consideration that users completing the two tasks are likely to have significantly different backgrounds. Users developing workflows for others are likely to have bioinformatics training and even if there is no exact definition of what a bioinformatician needs to know (Vincent and Charette, 2015) bioinformaticians tend to prefer working in a command-line environment due to its flexibility. Implementing tools in a graphical user interface with controlled input/out of data between tools limits this flexibility which means that bioinformaticians are unlikely to adopt Galaxy for their personal needs(Oliphant, 2016; Budd et al., 2015). Researchers looking to use pre-made workflows on the other hand come from a wide variety of backgrounds and in most cases this mean that they will have little or no experience of the tools they intend to use (Smith, 2013). Rich and flexible user options are therefore of limited value while a user-friendly graphical user interface is of high importance (Kumar and Dudley, 2007).

Despite these different user needs Galaxy provides the same interface to both user groups (figure 9) generating a far from optimal user experience for novice users using the graphical user interface.



Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources. You can install your own Galaxy by following the tutorial and choose from thousands of tools from the Tool Shed.

Menu headings for each type of tool available on the server.

A history providing access to all data produced by any tool used throughout the users active history. These may be used to extract a workflow from a link in the menu accessed through the cogwheel on the top of the column

Space for data display and editting of input parameters

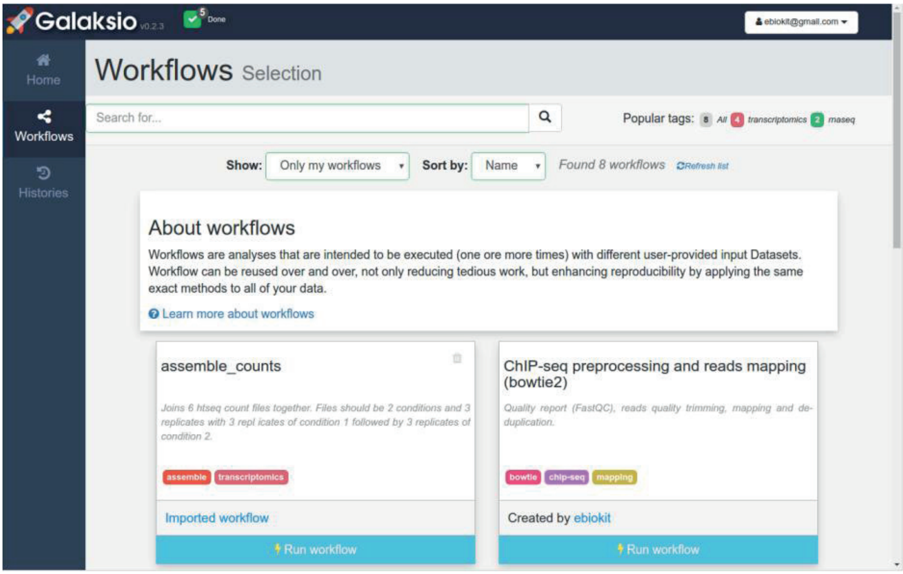
This history is empty. You can load your own data or get data from an external source

Creating #useGalaxy workflows is easier than you think. If you like reproducibility, you'll like workflows!

Figure 9. Hundreds of tools are available but hidden in the menu to the left and the history to the right expose all data generated by workflows. In the middle, a cramped space remains open to provide access to input parameters for each tool used.

As the GUI is the dominant form of user access (Galaxy Questionnaire Results by Manuel Corpas & Rafael Jimenez, data not published) and crucial to automating bioinformatics tasks work package four have developed the Galaksio interface described in paper V (figure 10). Each workflow covers a dedicated task and Galaksio retrieves all the workflow data from the Galaxy server. All actions accessible to the user in Galaksio (see table 6) are directly sent to the Galaxy server using the Galaxy API and processed on the Galaxy server. Meaning that there should be no conflicts between Galaksio and other

Galaxy extensions such as Bioblend (Sloggett et al., 2013), Cloudman (Afgan et al., 2010) and Pulsar (Afgan et al., 2015).



Workflows Selection

Search for...

Popular tags: 8 All 4 transcriptomics 2 rnaseq

Show: Only my workflows Sort by: Name Found 8 workflows Refresh list

About workflows

Workflows are analyses that are intended to be executed (one ore more times) with different user-provided input Datasets. Workflow can be reused over and over, not only reducing tedious work, but enhancing reproducibility by applying the same exact methods to all of your data.

Learn more about workflows

assemble_counts

Joins 6 htseq count files together. Files should be 2 conditions and 3 replicates with 3 repl icates of condition 1 followed by 3 replicates of condition 2.

assemble transcriptomics

Imported workflow

Run workflow

ChIP-seq preprocessing and reads mapping (bowtie2)

Quality report (FastQC), reads quality trimming, mapping and de-duplication.

bowtie chip-seq mapping

Created by ebiokit

Run workflow

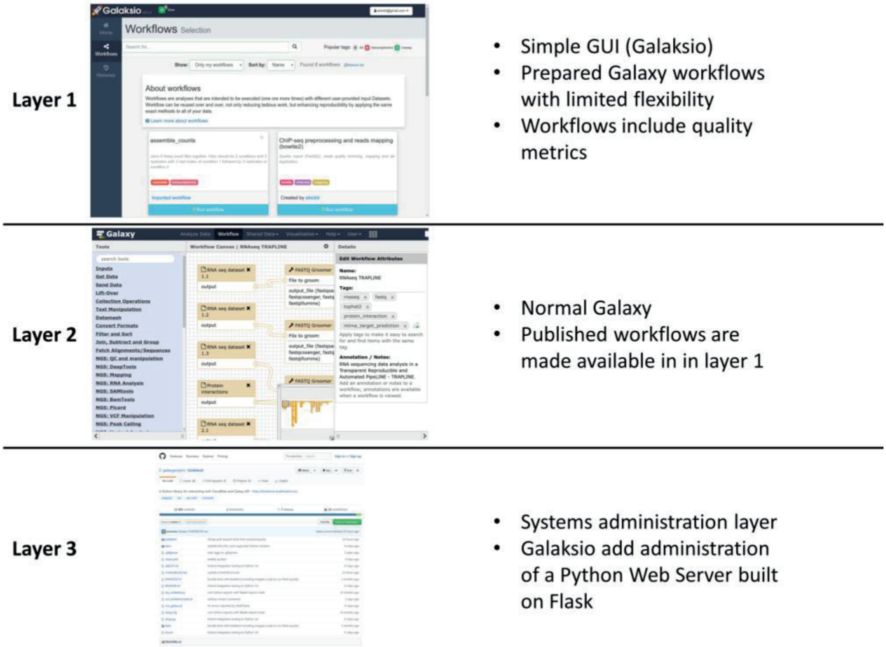
Figure 10. Galaksio provides access to a clean workflow-centric view providing access annotated workflows describing necessary input data. Galaxy histories and a homepage are available in the menu to the left.

Table 6. *Table of Galaksio features.*

Feature	Category	Implemented	Planned
User sign-in/out	Users	X	
User sign-up	Users	X	
Workflow listing	Workflows	X	
Workflow importing	Workflows	X	
Workflow execution	Workflows	X	
Workflow creation	Workflows		X
Simultaneous execution of workflows	Workflows	X	
Recovering previous executions	Workflows	X	
Help and description for tools in workflow	Workflows	X	
Input selection and parameter configuration	Workflows	X	
History selection	History	X	
History creation	History		X
History deletion	History		X
Dataset uploading	Dataset manipulation	X	
Dataset downloading	Dataset manipulation	X	

Feature	Category	Implemented	Planned
Dataset deletion	Dataset manipulation	X	
Dataset collection creation	Dataset manipulation	X	
Dataset collection deletion	Dataset manipulation		X
Tool execution	Tools		X

The result is what we call a layered approach allowing users to work with bioinformatics at the level they are comfortable with. Researchers at the biobank using defined workflows use Galaksio, bioinformaticians can develop workflows and integrate tools using the default Galaxy interface or high-level programming languages such as Python through Bioblend and system administration is only burdened by the addition of a separate Python Web Server built using Flask (figure 11).



Layer 1

- Simple GUI (Galaksio)
- Prepared Galaxy workflows with limited flexibility
- Workflows include quality metrics

Layer 2

- Normal Galaxy
- Published workflows are made available in in layer 1

Layer 3

- Systems administration layer
- Galaksio add administration of a Python Web Server built on Flask

Figure 11. The layered approach.

This means that a biobank lacking dedicated bioinformaticians can still operate effectively using automated workflows and retrieve these from public Galaxy servers or by collaborations with the Galaxy community and other eB3Kit users. Furthermore, researchers using these automated workflows will have their work tracked by the eB3Kit and provenance data is exported from

Galaxy to the file management system of the eB3Kit. Meaning that even if a researcher leaves the biobank his or her work will remain reproducible.

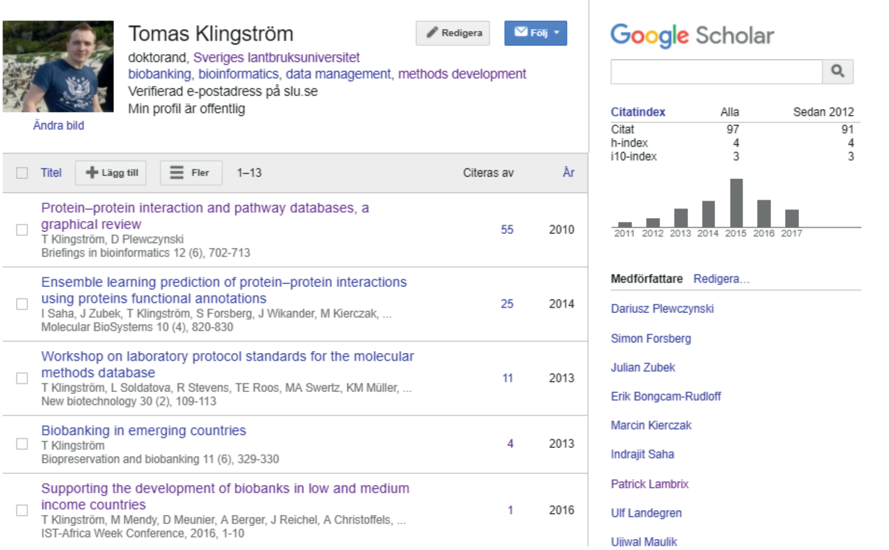
4 Discussion

The adoption of technology in research is hard to predict as the motivations underlying their implementation almost invariably constitute a wicked problem. Open sharing of data is a widely shared value in research and public databases like MolMeth can reach thousands of researchers each year even without formal certification of the protocols through peer review. Getting researchers to engage and contribute is, however, a more difficult task, and most entrepreneurs in social sharing of data have failed to propose a successful user proposition to motivate researchers to contribute. Likewise, proponents of best practice standards in a variety of subjects have found themselves struggling to reach widespread traction. With the eB3Kit it is possible to provide an out-of-the-box environment with integration between the components necessary to track, manage and share data. Which may be useful for quality management, large-scale research and beneficial to the careers of researchers using the sharing capacity to maximise the utilisation of their samples or results.

4.1 Evaluating the results of MolMeth and social networking in research

A lack of reproducibility is one of the major issues of research today, and numerous initiatives have been made to introduce new means of improving record keeping as well as improving reproducibility. It has however proven extremely difficult to make researchers embrace data sharing and open record keeping for laboratory methods. Based on discussions most people seem very interested in finding documented methods, but there has been a lack of incentives to motivate researchers into compiling their protocols into a shareable format. This problem is not unique to MolMeth and most of the social networks covered in paper II as well as published technology-oriented solutions based on the EXACT2 (Soldatova et al., 2014, p. 2), the BioAssay Template (BAT) (Clark et al., 2016) and several commercial workflow systems have failed as experimental biologists persist with notebooks and spreadsheets for personal usage (Kazic, 2015).

This behaviour stands in stark contrast to openness as being one of the central values of science (Royal Society (Great Britain), 2012; Merton, 1942) which makes it somewhat surprising that technologies that enable the sharing of data and methods find it so hard to achieve widespread adoption. Empirical research (Anderson et al., 2007), as well as the results of the aforementioned projects, does, however, support our rather cynical view of researchers as rational rather than idealistic agents of science as outlined in paper II. There is a lack of solid empirical data on successful means to promote open sharing (Rowhani-Farid et al., 2017) but paper II provides a theoretical model that help us evaluate the rational incentives for researchers to share information. An important conclusion in paper II is that the journal, despite the common perception of a publish or perish culture (Harley and Acord, 2011; Fanelli, 2010; Neill, 2008; Fuyuno and Cyranoski, 2006), probably is not an end in itself. Rather the journal has received its premier position in life science due to its unique ability to combine, registration, certification, awareness-building and archiving in a single unit of information. If this theory proves true, it is likely that ongoing initiatives attaching digital object identifiers (DOI) to laboratory protocols (Teytelman et al., 2016), datasets (Herterich and Dallmeier-Tiessen, 2016) and genome reports (Smith, 2016) will become increasingly important in the coming decade. With DOI numbers, search engines and access to social networks researchers can register their findings and build awareness by other means than journal circulation and collect the relevant information using tools such as Google Scholar profiles which automatically aggregate publications as well as citation counts (figure 12).



Tomas Klingström

doktorand, Sveriges lantbruksuniversitet

biobanking, bioinformatics, data management, methods development

Verifierad e-postadress på slu.se

Min profil är offentlig

Ändra bild

□ Titel	1–13 Citeras av	År
□ Protein–protein interaction and pathway databases, a graphical review T Klingström, D Plewczynski Briefings in bioinformatics 12 (6), 702-713	55	2010
□ Ensemble learning prediction of protein–protein interactions using proteins functional annotations I Saha, J Zubek, T Klingström, S Forsberg, J Wikander, M Kierczak, ... Molecular BioSystems 10 (4), 820-830	25	2014
□ Workshop on laboratory protocol standards for the molecular methods database T Klingström, L Soldatova, R Stevens, TE Roos, MA Swertz, KM Müller, ... New biotechnology 30 (2), 109-113	11	2013
□ Biobanking in emerging countries T Klingström Biopreservation and biobanking 11 (6), 329-330	4	2013
□ Supporting the development of biobanks in low and medium income countries T Klingström, M Mendy, D Meunier, A Berger, J Reichel, A Christoffels, ... IST-Africa Week Conference, 2016, 1-10	1	2016

Citatindex	Alla	Sedan 2012
Citat	97	91
h-index	4	4
i10-index	3	3

Medförfattare Redigera...

Dariusz Plewczynski

Simon Forsberg

Julian Zubek

Erik Bongcam-Rudloff

Marcin Kierczak

Indrajit Saha

Patrick Lambrix

Ulf Landegren

Ujjwal Maulik

Figure 12. View of the public Google Scholar profile of the author. Articles can be manually added, but the Google web crawler technology provides a high-quality aggregator collecting article, conference proceedings and abstracts into a single comprehensive view along with citation metrics (available at https://scholar.google.com/citations?user=KdTvA-wAAAAJ).

4.2 Wicked problems

It is evident, but hard to explain, the discrepancy between the ideals of researchers and how they act in regards to open research. The existence of a normative dissonance between the ideals of researchers and their actions has previously been described in the empirical literature (Anderson et al., 2007) even if paper II provides a, to our knowledge, unique example of its practical implications for the development of technology platforms in research.

Recognising this normative dissonance and its implications is a necessary step to evaluate the development of websites and standards for sharing data. When advocating open data sharing researchers tend to focus on the reduction of technical barriers and overall societal benefits (McKiernan et al., 2016; Voytek, 2016) which based on our conclusions is an idealistic but inefficient way of promoting open data sharing. Unfortunately, empirical data on studies on alternative means to promote data sharing are lacking and the only empirically tested method to successfully encourage open data sharing in health and medical research is by "rewarding" researchers with open data badges (Rowhani-Farid et al., 2017; Kidwell et al., 2016).

One way to approach this issue is by accepting the issue as an example of a "wicked" problem (Rittel and Webber, 1973), meaning that the problem is subjective and its solutions cannot be quantitatively evaluated as the issue, and its solutions must be assessed from the perspective of each stakeholder (see table 7 for a full list of the characteristics of a wicked problem as summarised and explained by Ritchey (Ritchey, 2013) based on Ritten and Webbers original work).

Table 7.*The criteria used to define a "wicked" problem.*

Criteria	Explanation
1. There is no definite formulation of a wicked problem.	"The information needed to *understand* the problem depends upon one's idea for *solving* it. This is to say: in order to *describe* a wicked problem in sufficient detail, one has to develop an exhaustive inventory for all the conceivable solutions ahead of time."
2. Wicked problems have no stopping rules.	In solving a tame problem, "… the problem-solver knows when he has done his job. There are criteria that tell when *the* solution or *a* solution has been found". With wicked problems you never come to a "final", "complete" or "fully correct" solution - since you have no objective criteria for such. The problem is continually evolving and mutating. You stop when you run out of resources, when a result is subjectively deemed "good enough" or when we feel "we've done what we can…"
3. Solutions to wicked problems are not true-or-false, but better or worse.	The criteria for judging the validity of a "solution" to a wicked problem are strongly stakeholder dependent. However, the judgments of different stakeholders …"are likely to differ widely to accord with their group or personal interests, their special value-sets, and their ideological predilections." Different stakeholders see different "solutions" as simply better or worse.
4. There is no immediate and no ultimate test of a solution to a wicked problem.	"… any solution, after being implemented, will generate waves of consequences over an extended - virtually an unbounded - period of time. Moreover, the next day's consequences of the solution may yield utterly undesirable repercussions which outweigh the intended advantages or the advantages accomplished hitherto."
5. Every solution to a wicked problem is a "one-shot operation"; because there is no opportunity to learn by trial-and-error, every attempt counts significantly.	"… *every* implemented solution is consequential. It leaves "traces" that cannot be undone … And every attempt to reverse a decision or correct for the undesired consequences poses yet another set of wicked problems … ."

Criteria	Explanation
6. Wicked problems do not have an enumerable (or an exhaustively describable) set of potential solutions, nor is there a well-described set of permissible operations that may be incorporated into the plan.	"There are no criteria which enable one to prove that all the solutions to a wicked problem have been identified and considered. It may happen that no solution is found, owing to logical inconsistencies in the 'picture' of the problem."
7. Every wicked problem is essentially unique.	"There are no *classes* of wicked problems in the sense that the principles of solution can be developed to fit *all* members of that class." …Also, …"Part of the art of dealing with wicked problems is the art of not knowing too early which type of solution to apply."
8. Every wicked problem can be considered to be a symptom of another [wicked] problem.	Also, many internal aspects of a wicked problem can be considered to be symptoms of other internal aspects of the same problem. A good deal of mutual and circular causality is involved, and the problem has many causal levels to consider. Complex judgements are required in order to determine an appropriate *level of abstraction* needed to define the problem.
9. The causes of a wicked problem can be explained in numerous ways. The choice of explanation determines the nature of the problem's resolution.	"There is no rule or procedure to determine the 'correct' explanation or combination of [explanations for a wicked problem]. The reason is that in dealing with wicked problems there are several more ways of refuting a hypothesis than there are permissible in the [e.g. physical] sciences."
10. [With wicked problems,] the planner has no right to be wrong.	In "hard" science, the researcher is allowed to make hypotheses that are later refuted. Indeed, it is just such hypothesis generation that is a primary motive force behind scientific development (Ritchey, 1991). Thus one is not penalised for making hypothesis that turn out to be wrong. "In the world of … wicked problems no such immunity is tolerated. Here the aim is not to find the truth, but to improve some characteristic of the world where people live. Planners are liable for the consequences of the actions they generate …"

The basic question of "how should researchers communicate their findings and ideas" is in itself a wicked problem and provides an example of how wicked problems lack a clear stop where an issue can be seen as solved (point 2 in table 7). Publication in peer-reviewed journals became the favoured option at a time when the number of publications and active researchers increased rapidly while two-way communication was limited by physical constraints (Harley and Acord, 2011). As society progresses, other opportunities and challenges present themselves. From a bioinformatics perspective, the most significant trend is perhaps that high-throughput research and improving means of digital communication (Akerman, 2006) generate large datasets with complex data that cannot be adequately described within the confines of a journal article. Such datasets present not only a technical challenge but a social one as well as the publication of such datasets are significant achievements in their own right that must be recognised in the academic system. Furthermore they often represents major collaborative efforts and are likely to generate significant scientific impact by secondary research without the active involvement of the original data creators (Altman et al., 2015).

Providing access to data sharing models and integration with internationally accepted ontologies through a federated framework is therefore not enough to encourage data sharing in the B3Africa project. We must also recognise that data is a valuable commodity and accept that even if openness is an important value in research (Anderson et al., 2007) we must also provide rational incentives for researchers to share their data as they are unlikely to do so otherwise (Stephan, 2012). This complexity was underestimated at the start of the MolMeth project and even if more systematic reviews provide a more clear formulation of potential barriers that prevent sharing (van Panhuis et al., 2014) it is largely dependent on the individual stakeholder which obstacles are relevant in each particular case. Accepting the complexity of wicked problems also makes it less surprising that around 90 % of all attempts to develop social networks for researchers have failed. Most attempts fail, but a few provide solutions that are sufficiently adapted to specific demographics that they become sustainable as they do not only contribute to the abstract concept of scientific benefit but also provide benefits to the individuals using the network.

As Digitial Object Identifier numbers (DOI-number) provided by CrossRef are becoming a de facto standard for citeability in research (Lammey, 2016) alternative resources for publishing scientific material become more attractive. Based on the model proposed in paper II, citeability is a key factor, meaning that journal publications are likely to face increased competition from citeable resources such as citeable datasets (Herterich and Dallmeier-Tiessen, 2016), preprints and published protocols provided that they have a DOI attached to them. Protocols.io (Teytelman et al., 2016) is a recently announced website with the support of private equity that can serve as a valuable test for the citeability hypothesis as each protocol is given a DOI. Evaluating its development may, therefore, provide valuable insights on future developments in data sharing platforms with implications for websites such as MolMeth and the development of communities communicating using data models supported by the eB3Kit.

4.3 Improving quality in study design and data analysis

Errors preventing the reproducibility of research can roughly be divided into 25 % study design, 50 % failure during the sample processing and 25 % poor statistics and data reporting (Freedman et al., 2015). Study design and data analysis are therefore concepts closely related to each other with ad hoc post-processing an important sources of errors.

To consult a statistician after a study is initiated is all too often like asking a veterinarian to examine your dead dog, the post-mortem may be interesting, but it will not help the dog. If used appropriately the eB3Kit will contribute to improving the quality of research. Studies on human-machine interactions show that there is a significant bias against automated methods in favour of human "hands on" management unless the error rate of the machine is very low (Lee and See, 2004). With automated workflows for data analysis and a provenance system tracking samples throughout the system, bioinformaticians can provide workflows with support for automatic testing of confounding factors influencing the study and workflows to evaluate the statistical power and suitability of a workflow using simulated data before initiating an experiment.

Furthermore, the reproducibility of an automated workflow management system like the eB3Kit bioinformatics module can be extended to include the entire process from sample to final results as long as the biobank provide persistent storage of samples. Most research institutions are essentially collections of independent laboratories, each run by principal investigators who head a team of trainees. This scheme has ancient roots and a track record of success (Hyman, 2017) but also makes it hard to create sustainable environments where data and samples are made available for follow-up research (Dangl et al., 2010). This is especially true in low and medium income countries where successful researchers often leave the country as grants dry up, or new positions at more attractive locations become available (Adewole et al., 2014; Kasper and Bajunirwe, 2012; Dodani, 2005).

A quality infrastructure is, however, no guarantee for quality research unless appropriate management practices are in place (Simeon-Dubach et al., 2012; Grizzle et al., 2011; H. M. Moore et al., 2011; Compton, 2007). The bioinformatics module provides an automated infrastructure for *reproducible* data analysis which in combination with training provided by BCNet, H3Africa, H3Bionet and other organisations will provide the means for continuous quality management and maximise the likelihood of generating *replicable* results.

Currently 19-50 % of the current medical literature contains statistical flaws (Thiese et al., 2015; Ercan et al., 2007) and experiences in biobanking suggest that review by external experts evaluating the standardisation and documentation of procedures can have a significant positive influence on reproducibility (di Donato, 2014; Freedman and Inglese, 2014; Matzke et al., 2012). Aims to standardise data analysis are however in direct contradiction to current training practices in bioinformatics which are heavily dependent on “on demand training”. Meaning that each bioinformatician assembles his or her unique training program in the form of short workshops and web open courses (Atwood et al., 2015; Leek and Peng, 2015). Despite the fact that the need for comprehensive training programs at the university level for quantitative biology and bioinformatics have been known for a long time (Atwood et al., 2015; National Research Council (US) Committee on Undergraduate Biology Education to Prepare Research Scientists for the 21st Century, 2003).

The bioinformatics module of the eB3Kit can contribute to this solution by allowing experts to create automated workflows accessed through the Galaksio interface. These automated workflows provide several benefits compared to the *ad hoc* analysis used by researchers today.

- Clearly defined start and end points forcing the researchers to consider their study design to be compatible with the statistical methods available.
- Pre-defined workflows provide a comprehensive framework for updating existing knowledge. As new tools are developed, they replace legacy tools in workflows. Researchers can compare new workflows versus old workflows simply by retrieving original raw data from the eB3Kit and compare the output of new workflows with the legacy output.
- Provenance is maintained as the entire process from study design to data analysis is maintained within the same system supporting retrospective evaluation of analysis decisions.

An ideal study is from a statistical perspective a study where all potential outcomes are explored prior to data collection. Such an ideal study would provide a complete data set with information recording observed variables, heterogeneity within the population, biases in sample collection and a statistical data plan regarding where power levels have guided the size of the studied population and *a priori* decision has been made regarding statistical significance and tests of association.

Such planning does, however, require significant work, and as with other previously described “wicked” problems, it is not surprising that the research community consistently fails to achieve such a high level of quality. The eB3Kit bioinformatics component can here make a significant contribution by reducing the threshold to adopt such practices by providing “appropriate data management by design” and access to workflows that can be used to simulate experimental results. Thereby strengthening research applications and contributing to more efficient usage of resources by supporting funding applications that include feasibility tests using simulated data in workflows.

4.4 The automation of research and reliance on specialists

As research moves further and further away from the one gene-one enzyme era (Bussard, 2005; Beadle and Tatum, 1941) into the Omics era (Gomez-Cabrero et al., 2014; Tan et al., 2009a). Researchers are pushed into an environment where we are forced to deal with challenges requiring knowledge in a wide variety of fields ranging from law and ethics to functional biology, molecular biology and bioinformatics. Ideally, this would be met by a corresponding increase in the knowledge of the researchers. Unfortunately, this is unlikely to happen as the margin effect of knowledge acquisition declines over time and research shows that domain knowledge in one field does not greatly accelerate how we deal with challenges in other areas unless they are closely related (Hambrick and Oswald, 2005; Hambrick and Engle, 2002).

Collaboration in interdisciplinary teams allows researchers to focus on their own core skills but places high demands on trust and collaboration. Communication between fields, or even subfields, is not always straightforward with tacit as well as tangible norms favouring specialisation and distribution of credit to principal investigators (Fiore, 2008). Leading to situations like the one described for the Linnaeus Centre for Bioinformatics in the introduction as stakeholder could not reconcile their respective aims. It is therefore not surprising that work in interdisciplinary fields tends to deviate into the establishment of new research fields with dedicated university programs and career tracks (Canuel et al., 2015; Wightman and Hark, 2012; Tan et al., 2009b). As the field matures, the interdisciplinary field is provided with more and more specialists educated in the field, providing the subject with its distinct specialisation, potentially alienating it from the research communities it originated. One way to enable specialisation of researchers while maintaining support to researchers in related fields is by the continuous automation of tasks. Reviewers of the Galaksio article (paper V), as well as informal discussions with peers, have however revealed that many researchers raise concerns regarding the costs and benefits of automation related to the quality of work and the importance of researchers learning new skills.

4.4.1 Automation versus training more people in bioinformatics

Automation is a global trend, and in research, it both accelerates research and allows researchers to focus on their own fields of research. With modern NGS technology most sequencing can be performed in a highly automated process only requiring extracted DNA and (comparatively) little technical knowledge of the user. The trend of automation is in other words already well established in life science and can easily be inferred from concepts related to the division of labour which is a core concept of modern economy. The original theory of division of labour state that *"The division of labour, however, so far as it can be introduced, occasions, in every art, a proportionable increase of the productive powers of labour"* (Smith, 1786) and it applies to research as well. When applied to research and bioinformatics can be exemplified in the following way:

- ➢ Alternative A is to hire a person who has been trained in accordance with a recognised curriculum where secondary and tertiary education has prepared the person to work in bioinformatics.
- ➢ Alternative B is to hire a person who possesses a marketable skill in another research subject but is willing to undergo training to become a bioinformatician. As the person possess a marketable skill, the alternative cost of having an employment using that skill must be covered while learning to become a bioinformatician

To obtain a marketable skill for a research subject in Sweden (i.e. be eligible for a position as a PhD student) requires ten years of primary school, three years of secondary school and at least five years of tertiary education. The economic value of such a degree in the private sector is approximately 325 000 to 360 000 SEK per year (37 500 - 41 600 USD at current conversion rate) which is roughly equivalent to a first-year American postdoc and significantly less than a graduate working in the private sector (Stephan, 2012). To become proficient in a subject like bioinformatics specialisation often occurs at the secondary school level (Wightman and Hark, 2012) suggesting that an additional 3-8 years of education is a reasonable estimate of the time necessary to retrain a biologist into a competent bioinformatician with sufficient programming skills.

This means that to convert a "research biologist" into a "bioinformatician" we extend the education time of a researcher by 17-44 % while reducing their active years as a researcher by 3-8 years. Even at the lower end of the training estimate, such training implies an alternative cost of ~1 000 000 MSEK that must be covered by the employer or the researcher.

There are however three significant counterpoints towards this line of reasoning:

- A biologist retrained into a bioinformatician is a more versatile researcher than a pure expert on bioinformatics.
- A significant part of this time may be spent by “learning on the job” or be equivalent to the time the bioinformatician needs to learn another subject.
- Expertise in one field provides general skills that make it easier to transition into another field.

Even if a researcher, in theory, is more versatile, it is hard, if not impossible, to maintain a competitive level of expertise in a field without working actively in it as skills quickly become obsolete or atrophy (Bapna et al., 2013; Allen and Velden, 2002). Attending a standardised curriculum equivalent to an undergraduate program also means that the researcher will return to the school bench and thereby reduce the time spent applying previously obtained theoretical skills in an applied environment which is an important part of skills development (Boshuizen, 2003). In comparison “learning on the job” provides better opportunities to maintain practical skills in their main subject but does not fit well with the ideals of standardisation described in the previous section and still requires a significant investment of time and effort.

The final argument is based on optimism towards how academic education provides generalisable skills to the student. Such optimism is widespread, and SLU is, for example, deducting points from tertiary education level courses when included in a PhD degree. Unfortunately, evidence overwhelmingly supports a less optimistic truth where obtaining domain knowledge in one area does not significantly strengthen the ability to acquire domain knowledge in an unrelated field (Hambrick and Oswald, 2005; Hambrick and Engle, 2002).

Furthermore, it can be argued that given the lack of trained bioinformaticians and ad-hoc nature of training, many potential bioinformaticians are already being recruited from the biomedical community. So even if it is possible to find skilled researchers who can be trained on the job with limited negative side effects, this is a resource that is already being heavily utilised. Meaning that even with significant investments in bioinformatics education the shortage of bioinformaticians will remain which makes automation a necessity.

4.4.2 Does automated bioinformatics reduce quality?

Automation always implies a loss of controlled compared to production in the hands of a skilled artisan. Despite this downside automation and simplification have a history of improving quality over the long term as machines, when properly configured and provided with the appropriate raw materials, provide superior quality products as they are able to repeat step after step without rest or boredom. When evaluating statistics literature it is clear that human intervention is a common source of error with solutions suggested being primarily about increased peer-review, more education and semi-automated checklists checking for common errors (Lang and Altman, 2015; Peng, 2015; McNutt, 2014; Nolan, 2000). Researchers may therefore prefer “artisan statistics” as they are mainly worried about missing on important results due to false negatives. From a community perspective, however, automation of data analysis provides significant benefits in regards to data provenance, comparability and reproducibility as well as reducing the risk of inappropriate ad-hoc processing of data. Systems such as Galaxy which provides modular workflows and saving data in every step may therefore be an acceptable compromise between the personal incentives of a researcher and the needs of the community as the system enables researchers to extract data and analyse it using customised workflows when deemed necessary (Goecks et al., 2010).

The strongest argument in favour of automation may however be the simple realisation that bioinformatics has been heavily automated ever since its inception and even a single human genome would be impossible to analyse without the automation of routine tasks (Ewing et al., 1998) many cutting edge projects have published guidelines advising researchers to rely on FASTQ files with automatically assigned Phred-quality scores rather than the more expansive sequence read format (SRF) when working with large volumes of data (Van der Auwera et al., 2013; Clarke et al., 2012). When advancing automation from a per-tool basis to a per-workflow basis we must therefore realise that bioinformatics is already heavily automated and should realise that it is more relevant to discuss *how we automate bioinformatics* rather than *if we should do it* and draw inspiration from how increased automation has affected other information-heavy fields such as healthcare while doing so (Nolan, 2000).

As with all wicked problems no specific solution can be considered optimal for all stakeholders. Rather each proposed solution must be treated as a simulation where an unknown number of subjects evaluate the proposition and by their actions communicate the approval or avoidance of the proposition. In this context, we believe that automation and division of labour is a highly attractive solution for research integrated biobanks operating in a resource-constrained setting. These biobanks often have valuable sample collections, a need for an improved informatics infrastructure and can benefit from standardised solutions in molecular biology and bioinformatics as their novelty lies in their samples and study design rather than the creation of novel algorithms or hardware solutions.

5 Conclusions

It has recently been proposed that to sustain the wealth of a nation the key is to maintain a high level of productivity among its skilled workers (Malmberg et al., 2017). These results are consistent with my conclusion that it would be hugely beneficial for the research community if researchers are enabled to focus on their core areas of competence (where they are skilled) rather than constantly expand outside their field of expertise as it allows them to maximise their productivity. Unfortunately, modern research is highly competitive (Edwards and Roy, 2017; Lawrence, 2009) with incentives pushing researchers towards a "one project one researcher" approach preferring temporary staff performing ad-hoc solutions. Rather than making long-term investments in staff scientists who can specialise within key fields and support a large number of projects within a research institution (Hyman, 2017; Stephan, 2012).

As the academic reward system is unlikely to change quickly it is necessary to adapt any technical solutions to this reality. One way to do this is to adopt a matrix model for researchers similar to the one employed by many companies to create cross-functional teams (Barlett and Ghoshal, 1990). Research integrated biobanks using the eB3Kit provide a perfect fit for such an academic model as the infrastructure facilitates the transfer of information when necessary and automation of tasks when possible.

The common perception of research is skewed in favour of major landmark projects such as the Human Genome Organization (HUGO) project where scientific knowledge in a large number of fields is pushed forward at breakneck speed. Most research projects, however, are not the HUGO project, and overly ambitious research projects are undone due to delays, the loss of core members and unexpected results leading to *ad hoc* solutions drastically increasing the risk of errors. By recognising our limitations and focus on providing novelty within our own area of expertise we can greatly increase research quality, reduce time to publication and resolve many of the conflicts we see today between research and supporting tasks.

In a traditional matrix organisation employees are divided into permanent groups by their competence/function but assigned to work in cross-functional teams led by project managers, creating a matrix structure as the organisational structure is displayed in a pattern where employees report both vertically to their function manager and horizontally to their project managers. Research organisations could adopt similar structures for interdisciplinary research projects (see figure 13) to improve how we distribute workloads. In such an organisation, novel research should mainly be carried out by principal investigators who are the ones generating novel results by innovating within their field, while using automated systems or validated methods to handle tasks outside their area of competence. This means that an expert in a subject like bioinformatics may spend significant time helping researchers running standardised and validated workflows but also engage in the development of workflows on his own and make them available to colleagues as automated workflows in Galaksio as soon as they can be considered validated (see figure 13).

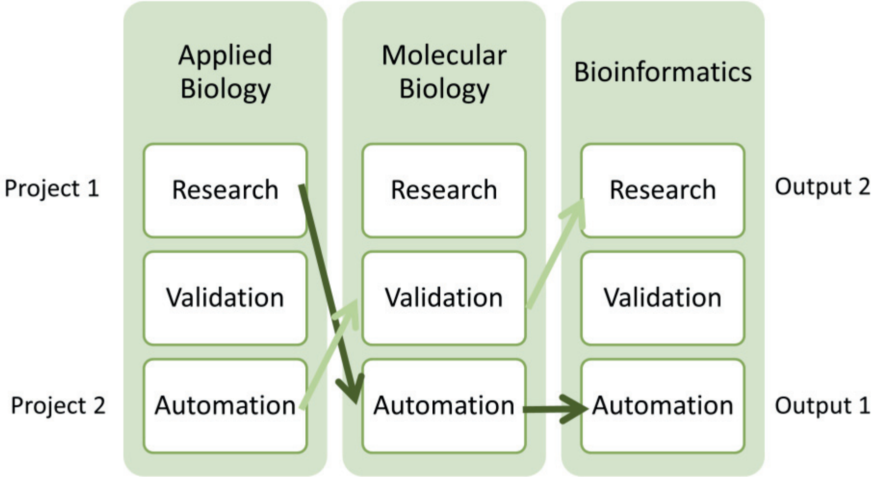


Figure 13. A matrix model dividing researchers into three functions (applied biology, molecular biology and bioinformatics) with two projects being performed within the organisation. In project one the chart indicates that a researcher specialised in applied science is the principal investigator and relying on automated support services using prepared workflows to generate results using prepared assays and data analysis workflows. In the second project, the bioinformatician or the molecular biologist would be the principal investigator as a new experimental method is validated and a data analysis pipeline being created using samples provided by the biobank rather than in a new research project.

Open sharing of data is a widely shared value in research and open databases like MolMeth can reach thousands of researchers each year even without formal peer review. Getting researchers to engage and contribute more actively have however proven to be a difficult task and most entrepreneurs aiming to encourage the social sharing of data have failed to achieve traction among the research community. Likewise proponents of best practice standards in a variety of subjects have found themselves struggling to reach widespread traction.

The aims of the B3Africa project are in light of these findings a highly ambitious project as we aim to produce an informatics platform which significantly alters how our users manage their day-to-day work. Our combination of technical solutions and development of soft skills does, however, provide strong incentives that may motivate researchers to use the platform as intended. That this approach has a good chance of becoming successful is not only supported by internal beliefs but also by the external advisory committee and feedback at conferences. At the 2017 Annual General meeting professor Ames Dhai, director of the Steve Biko Centre for Bioethics at the University of Witwatersrand in South Africa made the following assessment of the Be3Africa project:

- B3 Africa provides for meaningful research infrastructure partnerships because not only does it make provision for infrastructure, but also for the development of human capital and training and capacity building programs towards African intellectual leadership
- Currently, biobanks in Africa serve rather as collection centres for biological materials with progress in the bioinformatics aspect being non-existent or at a very nascent stage. With its focus on making robust the bioinformatics infrastructure and building capacity of African scientists in this discipline, it responds to an African need, fosters sustainability and reduces a dependency on well-resourced regions.

Integration of bioinformatics with a platform for biobanking is not an obvious decision for most biobanks in Europe as biobanks tend to operate as independent organisations. The evaluations of the eB3Kit do however show that fears of exploitation or the discovery of compromising errors are significant obstacles to sharing data or samples. Unpopular results are often challenged and relevant data questioning the scientific basis for widely supported causes such as the distribution of vitamin A and deworming pills have been withheld for years due to fears that any minor errors in data-handling may lead to the entire project being discredited and ignored (Hawkes, 2012). Such fears combined with the obvious risks of results being “scooped” by researchers with superior analytics capacity make it hard to convince researchers to share data openly

It should also be noted that many of the characteristics of research integrated biobanks can be extended to research institutions active in agricultural science and applied biology (such as the Swedish University of Agricultural science) as well. We may have cutting-edge facilities covering some aspects of molecular biology and bioinformatics, but overall our strength lies in the extensive knowledge possessed by researchers engaged in the study of the land, the animals and the plants surrounding us. With unique access to research facilities and research stations such as Lövsta and Grimsö we can obtain samples along with phenotype data in a manner matched by few other universities. Despite these advantages advances in Next Generation Sequencing and other –omics technologies have created a significant strain on our ability to manage the information collected and generated by modern research. A more stringent division of labour would therefore not only benefit research integrated biobanks in low- and medium income countries but can also contribute to significantly increase the research productivity of SLU.

6 Future prospects

A natural progression of this thesis will be to incorporate the eB3Kit and the Galaksio interface in the daily research at SLU. The eB3Kit is developed to work on a community level with researchers sharing workflows in Galaxy and creating new tools using the BiobankApps catalogue (Müller et al., 2017). Such tool-oriented articles are often rewarded with high citations rates (Park et al., 2015; Van Noorden et al., 2014) making it an attractive way of advancing your career as a young researcher.

Research automation can in itself also become a key strategic area for SLU and its aspirations to engage in an increased number of international collaborations. By adopting the division of labour within the university, we get the opportunity to build robust and reliable structures to exchange not only results but methodological knowledge in a scalable manner. This can, in turn, be used to strengthen collaboration with external actors through SLU Global and other research initiatives. The eB3Kit can thereby serve as a platform for innovation in several fields of biological and medical research as outlined below.

6.1 Deployment of automated bioinformatics support

A survey among PhD students at SLU made in anticipation of funding for the bioinformatics support infrastructure indicates that aid in data handling is a key priority to improve productivity. According to the survey, PhD students normally allocate 51-60 % of their time on data transformation and interpretation. In comparison they believe that roughly 31-40 % of the time is necessary while the remaining time is spent compensating for a lack of knowledge, tools or other issues.

With a labour force of PhD students equivalent to 277 full-time positions this implies that PhD students at SLU spend 50 000-150 000 hours per year on bioinformatics task that they could have allocated to other tasks if they had access to better support and training. Given the wide variety of research tasks performed by PhD students, it is unlikely that such gains can ever be achieved by automation. The numbers do however provide some insight on how bioinformatics training and platforms can increase productivity of SLU employees. Deployment of automated bioinformatics support is therefore likely to be a cost efficient investment to increase productivity and should be combined with studies to evaluate the amount of time saved by PhD students and other categories of researchers at SLU.

6.2 Deployment of Galaksio service interfaces for Galaxy servers

The development of the Galaksio user interface is an important part of the B3Africa project but can be installed independently of the eB3Kit. To demonstrate this capability Galaksio instance connected to a PhenoMeNal Galaxy server (van Rijswijk et al., 2017), will be installed on Google cloud to demonstrate how researchers can create an easily accessible and scalable cloud computing environment for metabolomics.

The Galaksio interface will also be highlighted in the Galaxy newsletter and is currently being considered for usage in connection with the main public Galaxy server (www.usegalaxy.org). The system is also under consideration for more specialised deployments creating easily accessible portals for bioinformatics projects. Promoting the usage of reproducible Galaxy workflows is at the heart of the Galaxy project (Goecks et al., 2010), but the user interface is not optimised for swift utilisation of workflows and reporting. Galaksio helps to resolve these issues and encouraging researchers to put together and publish comprehensive workflow packages may help to generate a positive feedback loop increasing overall usage of Galaxy (figure 14).

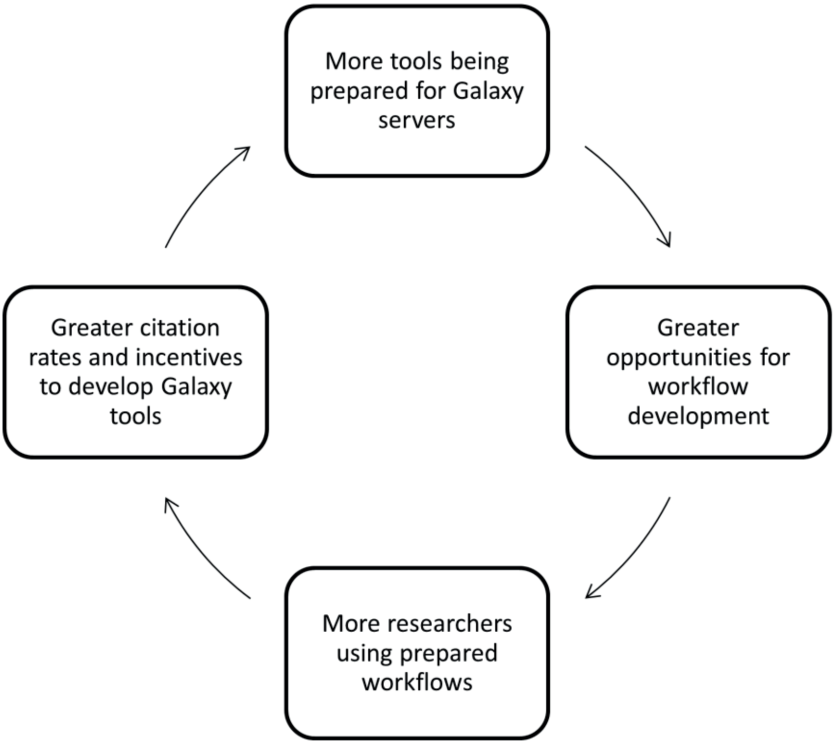

More tools being prepared for Galaxy servers
Greater opportunities for workflow development
More researchers using prepared workflows
Greater citation rates and incentives to develop Galaxy tools

Figure 14. Envisaged positive feedback loop for Galaxy usage by introducing a workflow-oriented user interface.

6.3 Using structured databases and high-throughput technologies for replication studies

In an ideal research environment positive as well as negative results accumulate in scientific articles until researchers feel confident enough to canonise theories as facts (Nissen et al., 2016). Unfortunately, given the skewed incentives of researchers, there is a severe bias in favour of positive results (Moonesinghe et al., 2007; Ioannidis, 2005) which distorts scientific reporting. A way to bypass these issues would be to promote harmonised methods and ensure that relevant metadata describing the samples is available to the research community.

With access to high-throughput analysis methods and structured databases, replication studies can be conducted in tandem with one's own original research provided that three key criteria are achieved.

- The biological systems under investigation must be sufficiently similar.

- The sample handling and analytical process must yield comparable results.
- There must be a system or an a priori knowledge of the fact that an experiment may serve as a replication of previous studies.

In theory, deep learning techniques (Angermueller et al., 2016) should enable researchers to sift through large datasets and identify experiments similar enough to be evaluated as replicates automatically. For such an approach to be realistic, it is, however, necessary that there is a degree of harmonisation of sampling methods and analytical techniques and that sufficient metadata is made available for the algorithms to identify functionally similar projects.

Federated networks of biobanks and analytics oriented service platforms would provide such an environment supporting replication studies at an unprecedented scale. Hypothesis generation and experimentation by robot scientists have already been explored (King et al., 2004) but a similar approach could be feasible on already existing data in a federated network of eB3Kits. In such a network, data would systematically be returned to biobanks or dedicated databanks. Making data available for hypothesis testing after research projects have been published.

6.4 Bridging the gap between applied research and ethics

There is a historical divide between applied research, bioethics and law as described in section 1.5 and 3.3. For researchers this results in frustration, and potentially alienation as projects are delayed for reasons beyond their control or understanding (Johnsson et al., 2014).

Embedding researchers with training in life science into collaborative projects with ethicists or lawyers can help reduce this frustration and contribute to developments in the respective fields. Research on threat modelling presented in a context relevant to lawyers and bioethicists could therefore be of tremendous value to the respective research communities. High impact articles on privacy related issues and emerging threats to the anonymity of donors are sometimes published with significant influence on policymaking (Erlich and Narayanan, 2014; Gymrek et al., 2013; Homer et al., 2008; Zerhouni and Nabel, 2008) but there is also evidence of scaremongering taking advantage of information-privacy experts or bioethicist lacking the knowledge necessary to evaluate claims in the field (Reardon, 2017). Given the phrasing of consent forms and priorities in data protection is unlikely that threat modelling can be conducted on samples donated for health research but public datasets and aggregated data can provide a sufficient substitute for topics such as the risk of genetic discrimination by insurance companies or employers.

Since the publication of paper III, Legal & ethical compliance when sharing biospecimen, the authors have been invited to write a review in the *British Medical Bulletin,* participated in writing a proposal for the EU H2020 call "The ethical dimensions of IT technologies: a European perspective on security and human rights aspects" and declined an invitation to participate in a competing proposal for the same call. Continuing this research therefore looks like an attractive proposition in the intersection between ethics, law and bioinformatics.

6.5 Automated biomolecular research

If widely adopted MolMeth would have provided a unique opportunity to deploy machine learning methods on semi-structured text to develop automated parsers able to read and annotate laboratory protocols. A lack of source data means that such a project is not yet possible, but the promotion of increasingly demanding laboratory standards (Neururer et al., 2016) create increasingly powerful incentives to describe the sample management process better.

Being able to integrate MolMeth features into the compliance process of the eB3Kit could thereby resolve many of the limitations of the current website as standard operating procedures or protocols could be retrieved directly from the document repository of a laboratory following CEN/TS standards for biobanks. By accessing published data generated from biobanks and comparing variations between protocols it is then possible to identify key parameters influencing the quality of samples and confounding factors contributing to irreplicable findings in research. A pilot study regarding the factors influencing fragmentation of DNA is currently ongoing and will serve as a template for the design of quality management workflows in the eB3Kit.

6.6 Development of a Phenomics platform

Connecting phenotype data with genotype data is of immense value to biobanks and life science in general. It is therefore unfortunate that SLU still lacks a system to connect recordings from Lövsta research station or the university animal hospital with genomic data generated in the laboratory.

We have previously collaborated with partners from International Bull Evaluation (located at SLU), the European Molecular Biology Laboratory, Wageningen University, the University of Bari, the Italian National Research Council and University of Manchester to develop such a platform but failed to gain funding for the proposed Federated Phenomics Data project for livestock research (FePheD). Renewed attempts based on experience from the eB3Kit and the Biobank Cloud project which has developed Hops (Ismail et al., 2017) may, however, be approved in the future.

A phenotype platform built on Hops would be highly extendable and able to support pre-processing of resource intensive data from sources such as 3D camera imaging for long term storage. Integrating the phenomics platform with the eB3Kit would combine the system with a graphical user interface for researchers, enabling researchers to identify relevant data from the phenomics platform and integrating it with other resources such as laboratory information systems using the STATegra experiment management system.

A pilot project to integrate the eB3Kit with Hops is scheduled for the autumn at Karolinska Institutet and will provide valuable experience for a potential future implementation of a similar system at SLU.

6.7 Curing Ebola and other tropical diseases

Ebola is perhaps the most famous and fear-inducing of tropical diseases. The fearsome spread of Ebola during short periods of time create a “boom and bust” environment where international researchers and medical staff arrive at the scene, collect samples, do their work and then disappear almost as quickly as they came. Attempting to organise research ad hoc during a major health crisis is impossible and the lack of coordinated sample management constitutes a biohazard as well as a wasted opportunity to produce long term solutions (Abayomi, Katz, et al., 2016).

With the portability of the eB3Kit and its modular design, the kit is perfect for challenges where time is of importance. With the eB3Kit being deployed at several locations in Africa many researchers will already have prior experience of working with the kit and a new kit can quickly be cloned from an existing implementation. At a time of crisis an eB3Kit can therefore rapidly be made available and provide an informatics platform for sample management at health care institutions during a disease outbreak. Early responders thereby get access to a system for structured sample storage which reduces a currently serious biohazard in the form of poorly maintained sample collections with potentially life threatening content while enabling long term stakeholders to make better use of the samples for future research.

7 Concluding remarks

As humans, we are biased in favour of innovations that allow us to carry out our work in the same way as before albeit in a more efficient manner. This bias in combination with the inherent difficulties of gaining acceptance for solutions to wicked problems contributes to the remarkable resilience of the academic system of peer-review and career advancement. At a time when global patterns of business and consumption are being disrupted by industrialisation, globalisation and digitalisation, the same trends have merely accelerated academic communication while leaving the overall structure intact.

The eB3Kit fits into this conservative environment as it helps researchers to collect more data, analyse it faster and register their results in scientific journals. At the same time the eB3Kit may enable a more radical transformation in how we handle and perceive data. Access to data is not only limited by passwords and access rights. With an ever increasing amount of publications it is impossible for researchers to keep track of potentially useful data unless it is made available in structured and searchable data formats. Making structured data from studies more readily accessible for analysis and reuse may therefore help to break the current paradigm of one researcher, one project, one main author, and significantly increase the productivity as well as the quality of research.

For example, an extensively documented set of samples collected for applied biology may generate data that is reused for biomolecular research on quality parameters, replication studies for unrelated phenotypes, as a validation dataset for the development of new bioinformatics tools and as a resource for information about genetic variance in threat modelling. To make such a future possible it is however necessary to ensure that it is rational for researchers to share high quality datasets. Researchers are however reluctant to share data if it requires a significant amount of time or financing to share it, if it carries a risk that they will be scooped to future publications or if it may inadvisable due to ethical or legal concerns. One way to encourage sharing and efficient sample utilisation is by separating the task of sample collection from the analysis of samples. Organising samples into biobanks was rated as one of the top 10 ideas changing the world right now by Time Magazine in 2009. Such investments are however generally restricted to fields and regions where strategic funding over long periods is available as the time from start up to measurable results spans many years from the establishment of a biobank. Distributing eB3Kits provide researchers with access to a cost efficient informatics platform for the establishment of biobanks and thereby make such investment feasible in regions and fields without access to the strategic funding currently necessary to establish a biobank.

Concerns regarding the replicability and quality of research also make it important for research groups to adopt robust informatics platforms. Important scientific discoveries may be delayed for long periods of time as researchers delay publication over concerns that minor errors may cast doubt on important but controversial findings. Many researchers also delay publications as they are unwilling to make unique datasets available until they are confident that all relevant research opportunities have been exhausted, fearing that they otherwise may be scooped by other researcher groups better equipped to analyse the data. The eB3Kit enables researcher groups to better manage this risk and quickly sift through their data to identify features of scientific interest using widely accepted workflows.

Using the eB3Kit may therefore generate a cascade effect of benefits to research groups who adopt it. Improved quality management and data analysis enables them to publish their findings more quickly. The published datasets may then after publication serve as beacons for the researchers who published them as stringent quality measures and relevant metadata improves the credibility of results, making the data and results more likely to generate secondary publications by other researchers, boosting the scientific impact of the original findings. Deployment of the eB3Kit into research institutions can thereby significantly contribute to research productivity and enable research groups with limited resources to produce competitive results influencing local communities as well as global scientific knowledge.

References

Abayomi, A., Christoffels, A., Grewal, R., Karam, L. A., Rossouw, C., Staunton, C., Swanepoel, C. and van Rooyen, B. (2013) 'Challenges of Biobanking in South Africa to Facilitate Indigenous Research in an Environment Burdened with Human Immunodeficiency Virus, Tuberculosis, and Emerging Noncommunicable Diseases', *Biopreservation and Biobanking*, vol. 11, no. 6, pp. 347–354 [Online]. DOI: 10.1089/bio.2013.0049.

Abayomi, A., Gevao, S., Conton, B., Deblasio, P. and Katz, R. (2016) 'African civil society initiatives to drive a biobanking, biosecurity and infrastructure development agenda in the wake of the West African Ebola outbreak', *Pan African Medical Journal*, vol. 24 [Online]. DOI: 10.11604/pamj.2016.24.270.8429 (Accessed 17 June 2017).

Abayomi, A., Katz, R., Spence, S., Conton, B. and Gevao, S. M. (2016) 'Managing dangerous pathogens: challenges in the wake of the recent West African Ebola outbreak', *Global Security: Health, Science and Policy*, vol. 1, no. 1, pp. 51–57 [Online]. DOI: 10.1080/23779497.2016.1228431.

Abbott, L. and Grady, C. (2011) 'A Systematic Review of the Empirical Literature Evaluating IRBs: What We Know and What We Still Need to Learn', *Journal of Empirical Research on Human Research Ethics: An International Journal*, vol. 6, no. 1, pp. 3–20 [Online]. DOI: 10.1525/jer.2011.6.1.3.

Adewole, I., Martin, D. N., Williams, M. J., Adebamowo, C., Bhatia, K., Berling, C., Casper, C., Elshamy, K., Elzawawy, A., Lawlor, R. T., Legood, R., Mbulaiteye, S. M., Odedina, F. T., Olopade, O. I., Olopade, C. O., Parkin, D. M., Rebbeck, T. R., Ross, H., Santini, L. A., Torode, J., Trimble, E. L., Wild, C. P., Young, A. M. and Kerr, D. J. (2014) 'Building capacity for sustainable research programmes for cancer in Africa', *Nature Reviews Clinical Oncology*, vol. 11, no. 5, pp. 251–259 [Online]. DOI: 10.1038/nrclinonc.2014.37.

Adoga, M. P., Fatumo, S. A. and Agwale, S. M. (2014) 'H3Africa: a tipping point for a revolution in bioinformatics, genomics and health research in Africa', *Source Code for Biology and Medicine*, vol. 9, no. 1, p. 10 [Online]. DOI: 10.1186/1751-0473-9-10.

Afgan, E., Baker, D., Coraor, N., Chapman, B., Nekrutenko, A. and Taylor, J. (2010) 'Galaxy CloudMan: delivering cloud compute clusters', *BMC Bioinformatics*, vol. 11, no. Suppl 12, p. S4 [Online]. DOI: 10.1186/1471-2105-11-S12-S4.

Afgan, E., Coraor, N., Chilton, J., Baker, D., Taylor, J. and The Galaxy Team (2015) 'Enabling cloud bursting for life sciences within Galaxy: Enabling Cloud Bursting for Life Sciences within Galaxy', *Concurrency and Computation: Practice and Experience*, vol. 27, no. 16, pp. 4330–4343 [Online]. DOI: 10.1002/cpe.3536.

Akerman, R. (2006) 'Technical solutions: Evolving peer review for the internet', *Nature* [Online]. DOI: 10.1038/nature04997 (Accessed 16 October 2014).

Alberts, B., Kirschner, M. W., Tilghman, S. and Varmus, H. (2014) 'Rescuing US biomedical research from its systemic flaws', *Proceedings of the National Academy of Sciences*, vol. 111, no. 16, pp. 5773–5777 [Online]. DOI: 10.1073/pnas.1404402111.

Allen, J. and Velden, R. van der (2002) 'When do skills become obsolete, and when does it matter?', in *Research in Labor Economics*, Bingley, Emerald (MCB UP), vol. 21, pp. 27–50 [Online]. DOI: 10.1016/S0147-9121(02)21004-3 (Accessed 2 June 2017).

Allen, M., Bjerke, M., Edlund, H., Nelander, S. and Westermark, B. (2016) 'Origin of the U87MG glioma cell line: Good news and bad news', *Science Translational Medicine*, vol. 8, no. 354, p. 354re3-354re3 [Online]. DOI: 10.1126/scitranslmed.aaf6853.

Altman, M., Borgman, C., Crosas, M. and Matone, M. (2015) 'An introduction to the joint principles for data citation: An Introduction to the Joint Principles for Data Citation', *Bulletin of the American Society for Information Science and Technology*, vol. 41, no. 3, pp. 43–45 [Online]. DOI: 10.1002/bult.2015.1720410313.

Anderson, M. S., Martinson, B. C. and De Vries, R. (2007) 'Normative Dissonance in Science: Results from a National Survey of U.S. Scientists', *Journal of Empirical Research on Human Research Ethics: An International Journal*, vol. 2, no. 4, pp. 3–14 [Online]. DOI: 10.1525/jer.2007.2.4.3.

Andersson, L. (2016) 'Domestic animals as models for biomedical research', *Upsala Journal of Medical Sciences*, vol. 121, no. 1, pp. 1–11 [Online]. DOI: 10.3109/03009734.2015.1091522.

Angell, E. and Dixon-Woods, M. (2009) 'Do research ethics committees identify process errors in applications for ethical approval?', *Journal of Medical Ethics*, vol. 35, no. 2, pp. 130–132 [Online]. DOI: 10.1136/jme.2008.025940.

Angell, E., Sutton, A. J., Windridge, K. and Dixon-Woods, M. (2006) 'Consistency in decision making by research ethics committees: a controlled comparison', *Journal of Medical Ethics*, vol. 32, no. 11, pp. 662–664 [Online]. DOI: 10.1136/jme.2005.014159.

Angermueller, C., Pärnamaa, T., Parts, L. and Stegle, O. (2016) 'Deep learning for computational biology', *Molecular Systems Biology*, vol. 12, no. 7, p. 878 [Online]. DOI: 10.15252/msb.20156651.

Anon (2015) 'Data overprotection', *Nature*, vol. 522, no. 7557, pp. 391–392 [Online]. DOI: 10.1038/522391b.

Aronson, A. R. (2001) 'Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.', *Proceedings of the AMIA Symposium*, pp. 17–21.

Arrowsmith, J. (2011) 'Trial watch: Phase II failures: 2008–2010', *Nature Reviews Drug Discovery*, vol. 10, no. 5, pp. 328–329 [Online]. DOI: 10.1038/nrd3439.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. and Sherlock, G. (2000) 'Gene Ontology: tool for the unification of biology', *Nature Genetics*, vol. 25, no. 1, pp. 25–29 [Online]. DOI: 10.1038/75556.

Astrin, J. J. and Betsou, F. (2016) ‘Trends in Biobanking: A Bibliometric Overview’, *Biopreservation and Biobanking*, vol. 14, no. 1, pp. 65–74 [Online]. DOI: 10.1089/bio.2015.0019.

Atwood, T. K., Bongcam-Rudloff, E., Brazas, M. E., Corpas, M., Gaudet, P., Lewitter, F., Mulder, N., Palagi, P. M., Schneider, M. V., van Gelder, C. W. G. and GOBLET Consortium (2015) ‘GOBLET: The Global Organisation for Bioinformatics Learning, Education and Training’, Welch, L. (ed), *PLOS Computational Biology*, vol. 11, no. 4, p. e1004143 [Online]. DOI: 10.1371/journal.pcbi.1004143.

Baker, M. (2015) ‘Reproducibility crisis: Blame it on the antibodies’, *Nature*, vol. 521, no. 7552, pp. 274–276 [Online]. DOI: 10.1038/521274a.

Baker, M. (2016) ‘1,500 scientists lift the lid on reproducibility’, *Nature*, vol. 533, no. 7604, pp. 452–454.

Bapna, R., Langer, N., Mehra, A., Gopal, R. and Gupta, A. (2013) ‘Human Capital Investments and Employee Performance: An Analysis of IT Services Industry’, *Management Science*, vol. 59, no. 3, pp. 641–658 [Online]. DOI: 10.1287/mnsc.1120.1586.

Barlett, C. A. and Ghoshal, S. (1990) ‘Matrix management: not a structure, a frame of mind’, *Harvard Business Review*, vol. 68, no. 4, pp. 138–145.

Beadle, G. W. and Tatum, E. L. (1941) ‘Genetic Control of Biochemical Reactions in Neurospora’, *Proceedings of the National Academy of Sciences*, vol. 27, no. 11, pp. 499–506 [Online]. DOI: 10.1073/pnas.27.11.499.

Begley, C. G. and Ellis, L. M. (2012) ‘Drug development: Raise standards for preclinical cancer research’, *Nature*, vol. 483, no. 7391, pp. 531–533 [Online]. DOI: 10.1038/483531a.

Beisvåg, V., Kauffmann, A., Malone, J., Foy, C., Salit, M., Schimmel, H., Bongcam-Rudloff, E., Landegren, U., Parkinson, H., Huber, W., Brazma, A., Sandvik, A. and Kuiper, M. (2011) ‘Contributions of the EMERALD project to assessing and improving microarray data quality’, *BioTechniques*, vol. 50, no. 1, pp. 27–31 [Online]. DOI: 10.2144/000113591.

Bendou, H., Sizani, L., Reid, T., Swanepoel, C., Ademuyiwa, T., Merino-Martinez, R., Meuller, H., Abayomi, A. and Christoffels, A. (2017) ‘Baobab Laboratory Information Management System: Development of an Open-Source Laboratory Information Management System for Biobanking’, *Biopreservation and Biobanking*, vol. 15, no. 2, pp. 116–120 [Online]. DOI: 10.1089/bio.2017.0014.

Betsou, F., Lehmann, S., Ashton, G., Barnes, M., Benson, E. E., Coppola, D., DeSouza, Y., Eliason, J., Glazer, B., Guadagni, F., Harding, K., Horsfall, D. J., Kleeberger, C., Nanni, U., Prasad, A., Shea, K., Skubitz, A., Somiari, S., Gunter, E. and International Society for Biological and Environmental Repositories (ISBER) Working Group on Biospecimen Science (2010) ‘Standard Preanalytical Coding for Biospecimens: Defining the Sample PREanalytical Code’, *Cancer Epidemiology Biomarkers & Prevention*, vol. 19, no. 4, pp. 1004–1011 [Online]. DOI: 10.1158/1055-9965.EPI-09-1268.

Björk, B.-C. and Solomon, D. (2013) ‘The publishing delay in scholarly peer-reviewed journals’, *Journal of Informetrics*, vol. 7, no. 4, pp. 914–923 [Online]. DOI: 10.1016/j.joi.2013.09.001.

Blackburn, H. (2012) ‘Genetic Selection and Conservation of Genetic Diversity*: Genetic Selection and Conservation of Genetic Diversity’, *Reproduction in Domestic Animals*, vol. 47, pp. 249–254 [Online]. DOI: 10.1111/j.1439-0531.2012.02083.x.

Bodenreider, O. (2004) ‘The Unified Medical Language System (UMLS): integrating biomedical terminology’, *Nucleic Acids Research*, vol. 32, no. 90001, p. 267D–270 [Online]. DOI: 10.1093/nar/gkh061.

Bornmann, L. and Leydesdorff, L. (2017) ‘Skewness of citation impact data and covariates of citation distributions: A large-scale empirical analysis based on Web of Science data’, *Journal of Informetrics*, vol. 11, no. 1, pp. 164–175 [Online]. DOI: 10.1016/j.joi.2016.12.001.

Boshuizen, H. P. A. (ed.) (2003) *Expertise development: the transition between school and work: conference*, Heerlen, Open Universiteit Nederland.

Brochhausen, M., Fransson, M. N., Kanaskar, N. V., Eriksson, M., Merino-Martinez, R., Hall, R. A., Norlin, L., Kjellqvist, S., Hortlund, M., Topaloglu, U., Hogan, W. R. and Litton, J.-E. (2013) ‘Developing a semantically rich ontology for the biobank-administration domain’, *Journal of Biomedical Semantics*, vol. 4, no. 1, p. 23 [Online]. DOI: 10.1186/2041-1480-4-23.

Budd, A., Corpas, M., Brazas, M. D., Fuller, J. C., Goecks, J., Mulder, N. J., Michaut, M., Ouellette, B. F. F., Pawlik, A. and Blomberg, N. (2015) ‘A Quick Guide for Building a Successful Bioinformatics Community’, Shehu, A. (ed), *PLOS Computational Biology*, vol. 11, no. 2, p. e1003972 [Online]. DOI: 10.1371/journal.pcbi.1003972.

Bussard, A. E. (2005) ‘A scientific revolution?’, *EMBO reports*, vol. 6, no. 8, pp. 691–694 [Online]. DOI: 10.1038/sj.embor.7400497.

Callaway, E. and Powell, K. (2016) ‘Biologists urged to hug a preprint’, *Nature*, vol. 530, no. 7590, pp. 265–265 [Online]. DOI: 10.1038/530265a.

Canuel, V., Rance, B., Avillach, P., Degoulet, P. and Burgun, A. (2015) ‘Translational research platforms integrating clinical and omics data: a review of publicly available solutions’, *Briefings in Bioinformatics*, vol. 16, no. 2, pp. 280–290 [Online]. DOI: 10.1093/bib/bbu006.

Ceusters, W., Smith, B. and Goldberg, L. (2005) ‘A terminological and ontological analysis of the NCI Thesaurus’, *Methods of Information in Medicine*, vol. 44, no. 4, pp. 498–507.

Chandrasekaran, B., Josephson, J. R. and Benjamins, V. R. (1999) ‘What are ontologies, and why do we need them?’, *IEEE Intelligent Systems*, vol. 14, no. 1, pp. 20–26 [Online]. DOI: 10.1109/5254.747902.

Chang, J. (2015) ‘Core services: Reward bioinformaticians’, *Nature*, vol. 520, no. 7546, pp. 151–152 [Online]. DOI: 10.1038/520151a.

Clark, A. M., Litterman, N. K., Kranz, J. E., Gund, P., Gregory, K. and Bunin, B. A. (2016) ‘BioAssay Templates for the semantic web’, *PeerJ Computer Science*, vol. 2, p. e61 [Online]. DOI: 10.7717/peerj-cs.61.

Clarke, L., Zheng-Bradley, X., Smith, R., Kulesha, E., Xiao, C., Toneva, I., Vaughan, B., Preuss, D., Leinonen, R., Shumway, M., Sherry, S. and Flicek, P. (2012) ‘The 1000 Genomes Project: data management and community access’, *Nature Methods*, vol. 9, no. 5, pp. 459–462 [Online]. DOI: 10.1038/nmeth.1974.

Cohen, Y., Almog, R., Onn, A., Itzhaki-Alfia, A. and Meir, K. (2013) ‘Establishing and Sustaining a Biorepository Network in Israel: Challenges and Progress’, *Biopreservation and Biobanking*, vol. 11, no. 6, pp. 331–338 [Online]. DOI: 10.1089/bio.2013.0046.

Compton, C. (2007) ‘Getting to personalized cancer medicine: Taking out the garbage’, *Cancer*, vol. 110, no. 8, pp. 1641–1643 [Online]. DOI: 10.1002/cncr.22966.

Coop, G. (2016) ‘Does linked selection explain the narrow range of genetic diversity across species?’, [Online]. DOI: 10.1101/042598 (Accessed 19 May 2017).

Crosswell, L. C. and Thornton, J. M. (2012) ‘ELIXIR: a distributed infrastructure for European biological data’, *Trends in biotechnology*, vol. 30, no. 5, pp. 241–242.

Crosthwaite, J. (1995) ‘Moral expertise: A problem in the professional ethics of professional ethicists.’, *Bioethics*, vol. 9, no. 4, pp. 361–379 [Online]. DOI: 10.1111/j.1467-8519.1995.tb00312.x.

Daebeler, A., Bodelier, P. L. E., Hefting, M. M., Rütting, T., Jia, Z. and Laanbroek, H. J. (2017) ‘Soil warming and fertilization altered rates of nitrogen transformation processes and selected for adapted ammonia-oxidizing archaea in sub-arctic grassland soil’, *Soil Biology and Biochemistry*, vol. 107, pp. 114–124 [Online]. DOI: 10.1016/j.soilbio.2016.12.013.

Dangl, A., Demiroglu, S. Y., Gaedcke, J., Helbing, K., Jo, P., Rakebrandt, F., Rienhoff, O. and Sax, U. (2010) ‘The IT-infrastructure of a biobank for an academic medical center’, *Studies in Health Technology and Informatics*, vol. 160, no. Pt 2, pp. 1334–1338.

De Souza, Y. G. and Greenspan, J. S. (2013) ‘Biobanking past, present and future: responsibilities and benefits’, *AIDS*, vol. 27, no. 3, pp. 303–312 [Online]. DOI: 10.1097/QAD.0b013e32835c1244.

DiEuliis, D., Johnson, K. R., Morse, S. S. and Schindel, D. E. (2016) ‘Opinion: Specimen collections should have a much bigger role in infectious disease research and response’, *Proceedings of the National Academy of Sciences*, vol. 113, no. 1, pp. 4–7 [Online]. DOI: 10.1073/pnas.1522680112.

Directorate-General for Research and Innovation (2016) *Strategy Report on Research Infrastructures: Roadmap 2016*, Brussels, Belgium, European Commission [Online]. Available at https://ec.europa.eu/research/infrastructures/pdf/esfri/esfri_roadmap/esfri_roadmap_2016_full.pdf.

Dixon-Woods, M., Foy, C., Hayden, C., Al-Shahi Salman, R., Tebbutt, S. and Schroter, S. (2016) ‘Can an ethics officer role reduce delays in research ethics approval? A mixed-method evaluation of an improvement project’, *BMJ Open*, vol. 6, no. 8, p. e011973 [Online]. DOI: 10.1136/bmjopen-2016-011973.

Dodani, S. (2005) ‘Brain drain from developing countries: how can brain drain be converted into wisdom gain?’, *Journal of the Royal Society of Medicine*, vol. 98, no. 11, pp. 487–491 [Online]. DOI: 10.1258/jrsm.98.11.487.

di Donato, J.-H. (2014) ‘Quality Management Systems: how do biobanks benefit?’, Leipzig, Germany [Online]. Available at http://www.esbb.org/leipzig/download/speakers/diDonato.pdf (Accessed 1 June 2017).

Doucet, M., Becker, K. F., Björkman, J., Bonnet, J., Clément, B., Daidone, M.-G., Duyckaerts, C., Erb, G., Haslacher, H., Hofman, P., Huppertz, B., Junot, C., Lundeberg, J., Metspalu, A., Lavitrano, M., Litton, J.-E., Moore, H., Morente, M., Naimi, B.-Y., Oelmueller, U., Ollier, B., Parodi, B., Ruan, L., Stanta, G., Turano, P., Vaught, J., Watson, P., Wichmann, H.-E., Yuille, M., Zaomi, M., Zatloukal, K. and Dagher, G. (2016) ‘Quality Matters: 2016 Annual Conference of the National Infrastructures for Biobanking’, *Biopreservation and Biobanking* [Online]. DOI: 10.1089/bio.2016.0053 (Accessed 8 January 2017).

Dove, E. S., Townend, D., Meslin, E. M., Bobrow, M., Littler, K., Nicol, D., de Vries, J., Junker, A., Garattini, C., Bovenberg, J., Shabani, M., Levesque, E. and Knoppers, B. M. (2016) 'Ethics review for international data-intensive research', *Science*, vol. 351, no. 6280, pp. 1399–1400 [Online]. DOI: 10.1126/science.aad5269.

Edwards, M. A. and Roy, S. (2017) 'Academic Research in the 21st Century: Maintaining Scientific Integrity in a Climate of Perverse Incentives and Hypercompetition', *Environmental Engineering Science*, vol. 34, no. 1, pp. 51–61 [Online]. DOI: 10.1089/ees.2016.0223.

Ercan, I., Yazıcı, B., Yang, Y., Özkaya, G., Cangur, S., Ediz, B. and Kan, I. (2007) 'Misusage of statistics in medical research', *Eur J Gen Med*, vol. 4, no. 3, pp. 128–134.

Erlich, Y. and Narayanan, A. (2014) 'Routes for breaching and protecting genetic privacy', *Nature Reviews Genetics*, vol. 15, no. 6, pp. 409–421 [Online]. DOI: 10.1038/nrg3723.

European Union (2016) 'Directive 95/46/EC (General Data Protection Regulation)', *Official Journal of the European Union*, vol. 59.

Ewing, B., Hillier, L., Wendl, M. C. and Green, P. (1998) 'Base-Calling of Automated Sequencer Traces Using *Phred.* ?I. Accuracy?Assessment', *Genome Research*, vol. 8, no. 3, pp. 175–185 [Online]. DOI: 10.1101/gr.8.3.175.

Fanelli, D. (2010) 'Do Pressures to Publish Increase Scientists' Bias? An Empirical Support from US States Data', Scalas, E. (ed), *PLoS ONE*, vol. 5, no. 4, p. e10271 [Online]. DOI: 10.1371/journal.pone.0010271.

Fiore, S. M. (2008) 'Interdisciplinarity as Teamwork: How the Science of Teams Can Inform Team Science', *Small Group Research*, vol. 39, no. 3, pp. 251–277 [Online]. DOI: 10.1177/1046496408317797.

Fjällbrant, N. (2012) 'Scholarly communication-historical development and new possibilities',.

Freedman, L. P., Cockburn, I. M. and Simcoe, T. S. (2015) 'The Economics of Reproducibility in Preclinical Research', *PLOS Biology*, vol. 13, no. 6, p. e1002165 [Online]. DOI: 10.1371/journal.pbio.1002165.

Freedman, L. P. and Inglese, J. (2014) 'The Increasing Urgency for Standards in Basic Biologic Research', *Cancer Research*, vol. 74, no. 15, pp. 4024–4029 [Online]. DOI: 10.1158/0008-5472.CAN-14-0925.

Fuxelius, H., Bongcam, E. and Jaufeerally, Y. (2010) 'The contribution of the eBioKit to Bioinformatics Education in Southern Africa', *EMBnet.journal*, vol. 16, no. 1, p. 29 [Online]. DOI: 10.14806/ej.16.1.173.

Fuyuno, I. and Cyranoski, D. (2006) 'Cash for papers: putting a premium on publication', *Nature*, vol. 441, no. 7095, pp. 792–792 [Online]. DOI: 10.1038/441792b.

Garvey, W. D. and Griffith, B. C. (1972) 'Communication and information processing within scientific disciplines: Empirical findings for Psychology', *Information Storage and Retrieval*, vol. 8, no. 3, pp. 123–136 [Online]. DOI: 10.1016/0020-0271(72)90041-1.

Geller, J., Morrey, C. P., Xu, J., Halper, M., Elhanan, G., Perl, Y. and Hripcsak, G. (2009) 'Comparing inconsistent relationship configurations indicating UMLS errors', *AMIA ... Annual Symposium proceedings. AMIA Symposium*, vol. 2009, pp. 193–197.

Gentil-Beccot, A., Mele, S. and Brooks, T. C. (2010) 'Citing and reading behaviours in high-energy physics', *Scientometrics*, vol. 84, no. 2, pp. 345–355 [Online]. DOI: 10.1007/s11192-009-0111-1.

Glasziou, P., Altman, D. G., Bossuyt, P., Boutron, I., Clarke, M., Julious, S., Michie, S., Moher, D. and Wager, E. (2014) 'Reducing waste from incomplete or unusable reports of biomedical research', *The Lancet*, vol. 383, no. 9913, pp. 267–276 [Online]. DOI: 10.1016/S0140-6736(13)62228-X.

Goecks, J., Nekrutenko, A., Taylor, J. and Galaxy Team, T. (2010) 'Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences', *Genome Biology*, vol. 11, no. 8, p. R86 [Online]. DOI: 10.1186/gb-2010-11-8-r86.

Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., Ballestar, E., Bongcam-Rudloff, E., Conesa, A. and Tegnér, J. (2014) 'Data integration in the era of omics: current and future challenges', *BMC Systems Biology*, vol. 8, no. Suppl 2, p. I1 [Online]. DOI: 10.1186/1752-0509-8-S2-I1.

Grizzle, W. E., Bell, W. C. and Sexton, K. C. (2011) 'Issues in collecting, processing and storing human tissues and associated information to support biomedical research', Srivastava, S. and Grizzle, W. E. (eds), *Cancer Biomarkers*, vol. 9, no. 1–6, pp. 531–549 [Online]. DOI: 10.3233/CBM-2011-0183.

Groeneveld, L. F., Gregusson, S., Guldbrandtsen, B., Hiemstra, S. J., Hveem, K., Kantanen, J., Lohi, H., Stroemstedt, L. and Berg, P. (2016) 'Domesticated Animal Biobanking: Land of Opportunity', *PLOS Biology*, vol. 14, no. 7, p. e1002523 [Online]. DOI: 10.1371/journal.pbio.1002523.

Gruber, T. R. (1995) 'Toward principles for the design of ontologies used for knowledge sharing?', *International Journal of Human-Computer Studies*, vol. 43, no. 5–6, pp. 907–928 [Online]. DOI: 10.1006/ijhc.1995.1081.

Gymrek, M., McGuire, A. L., Golan, D., Halperin, E. and Erlich, Y. (2013) 'Identifying Personal Genomes by Surname Inference', *Science*, vol. 339, no. 6117, pp. 321–324 [Online]. DOI: 10.1126/science.1229566.

Hallmans, G. and Vaught, J. B. (2011) 'Best Practices for Establishing a Biobank', in Dillner, J. (ed), *Methods in Biobanking*, Totowa, NJ, Humana Press, vol. 675, pp. 241–260 [Online]. DOI: 10.1007/978-1-59745-423-0_13 (Accessed 31 May 2017).

Hambrick, D. Z. and Engle, R. W. (2002) 'Effects of Domain Knowledge, Working Memory Capacity, and Age on Cognitive Performance: An Investigation of the Knowledge-Is-Power Hypothesis', *Cognitive Psychology*, vol. 44, no. 4, pp. 339–387 [Online]. DOI: 10.1006/cogp.2001.0769.

Hambrick, D. Z. and Oswald, F. L. (2005) 'Does domain knowledge moderate involvement of working memory capacity in higher-level cognition? A test of three models', *Journal of Memory and Language*, vol. 52, no. 3, pp. 377–397 [Online]. DOI: 10.1016/j.jml.2005.01.004.

Harley, D. and Acord, S. K. (2011) 'Peer Review in Academic Promotion and Publishing: Its Meaning, Locus, and Future', [Online]. Available at http://www.escholarship.org/uc/item/1xv148c8.

Hawkes, N. (2012) ‘Deworming debunked’, *BMJ*, vol. 346, no. jan02 1, pp. e8558–e8558 [Online]. DOI: 10.1136/bmj.e8558.

Hayes, B. J., Lewin, H. A. and Goddard, M. E. (2013) ‘The future of livestock breeding: genomic selection for efficiency, reduced emissions intensity, and adaptation’, *Trends in genetics: TIG*, vol. 29, no. 4, pp. 206–214 [Online]. DOI: 10.1016/j.tig.2012.11.009.

Hernández-de-Diego, R., Boix-Chova, N., Gómez-Cabrero, D., Tegner, J., Abugessaisa, I. and Conesa, A. (2014) ‘STATegra EMS: an Experiment Management System for complex next-generation omics experiments’, *BMC Systems Biology*, vol. 8, no. Suppl 2, p. S9 [Online]. DOI: 10.1186/1752-0509-8-S2-S9.

Hernández-de-Diego, R., de Villiers, E. P., Klingström, T., Gourlé, H., Conesa, A. and Bongcam-Rudloff, E. (2017) ‘The eBioKit, a stand-alone educational platform for bioinformatics’, Ouellette, F. (ed), *PLOS Computational Biology*, vol. 13, no. 9, p. e1005616 [Online]. DOI: 10.1371/journal.pcbi.1005616.

Herterich, P. and Dallmeier-Tiessen, S. (2016) ‘Data Citation Services in the High-Energy Physics Community’, *D-Lib Magazine*, vol. 22, no. 1/2 [Online]. DOI: 10.1045/january2016-herterich (Accessed 19 May 2017).

Heymann, D. L. (2017) ‘One health: science, politics and zoonotic disease in Africa’, *The Lancet Infectious Diseases*, vol. 17, no. 1, p. 37 [Online]. DOI: 10.1016/S1473-3099(16)30567-9.

Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J. V., Stephan, D. A., Nelson, S. F. and Craig, D. W. (2008) ‘Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays’, Visscher, P. M. (ed), *PLoS Genetics*, vol. 4, no. 8, p. e1000167 [Online]. DOI: 10.1371/journal.pgen.1000167.

Hyman, S. (2017) ‘Biology needs more staff scientists’, *Nature*, vol. 545, no. 7654, pp. 283–284 [Online]. DOI: 10.1038/545283a.

IARC (2016) *BCNet Launch* [Online]. Available at http://www.iarc.fr/en/media-centre/iarcnews/pdf/BCNet%20Launch.pdf.

Ioannidis, J. P. A. (2005) ‘Why Most Published Research Findings Are False’, *PLoS Medicine*, vol. 2, no. 8, p. e124 [Online]. DOI: 10.1371/journal.pmed.0020124.

Ismail, M., Gebremeskel, E., Kakantousis, T., Berthou, G. and Dowling, J. (2017) ‘Hopsworks: Improving User Experience and Development on Hadoop with Scalable, Strongly Consistent Metadata’, IEEE, pp. 2525–2528 [Online]. DOI: 10.1109/ICDCS.2017.41 (Accessed 1 September 2017).

Jain, M., Olsen, H. E., Paten, B. and Akeson, M. (2016) ‘The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community’, *Genome Biology*, vol. 17, no. 1 [Online]. DOI: 10.1186/s13059-016-1103-0 (Accessed 31 May 2017).

Johnsson, L., Eriksson, S., Helgesson, G. and Hansson, M. G. (2014) ‘Making researchers moral: Why trustworthiness requires more than ethics guidelines and review’, *Research Ethics*, vol. 10, no. 1, pp. 29–46 [Online]. DOI: 10.1177/1747016113504778.

Kasper, J. and Bajunirwe, F. (2012) ‘Brain drain in sub-Saharan Africa: contributing factors, potential remedies and the role of academic medical centres’, *Archives of Disease in Childhood*, vol. 97, no. 11, pp. 973–979 [Online]. DOI: 10.1136/archdischild-2012-301900.

Kazic, T. (2015) ‘Ten Simple Rules for Experiments’ Provenance’, *PLOS Computational Biology*, vol. 11, no. 10, p. e1004384 [Online]. DOI: 10.1371/journal.pcbi.1004384.

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., Kennett, C., Slowik, A., Sonnleitner, C., Hess-Holden, C., Errington, T. M., Fiedler, S. and Nosek, B. A. (2016) ‘Badges to Acknowledge Open Practices: A Simple, Low-Cost, Effective Method for Increasing Transparency’, Macleod, M. R. (ed), *PLOS Biology*, vol. 14, no. 5, p. e1002456 [Online]. DOI: 10.1371/journal.pbio.1002456.

King, G. (1986) ‘How not to lie with statistics: Avoiding common mistakes in quantitative political science’, *American Journal of Political Science*, pp. 666–687.

King, R. D., Rowland, J., Aubrey, W., Liakata, M., Markham, M., Soldatova, L. N., Whelan, K. E., Clare, A., Young, M., Sparkes, A., Oliver, S. G. and Pir, P. (2009) ‘The Robot Scientist Adam’, *Computer*, vol. 42, no. 7, pp. 46–54 [Online]. DOI: 10.1109/MC.2009.270.

King, R. D., Whelan, K. E., Jones, F. M., Reiser, P. G. K., Bryant, C. H., Muggleton, S. H., Kell, D. B. and Oliver, S. G. (2004) ‘Functional genomic hypothesis generation and experimentation by a robot scientist’, *Nature*, vol. 427, no. 6971, pp. 247–252 [Online]. DOI: 10.1038/nature02236.

Klingström, T. (2013) ‘Biobanking in Emerging Countries’, *Biopreservation and Biobanking*, vol. 11, no. 6, pp. 329–330 [Online]. DOI: 10.1089/bio.2013.1161.

Klingstrom, T., Bongcam-Rudloff, E. and Reichel, J. (2017) ‘Legal &amp; ethical compliance when sharing biospecimen’, *Briefings in Functional Genomics* [Online]. DOI: 10.1093/bfgp/elx008 (Accessed 8 May 2017).

Klingstrom, T., Mendy, M., Meunier, D., Berger, A., Reichel, J., Christoffels, A., Bendou, H., Swanepoel, C., Smit, L., Mckellar-Basset, C., Bongcam-Rudloff, E., Soderberg, J., Merino-Martinez, R., Amatya, S., Kihara, A., Kemp, S., Reihs, R. and Muller, H. (2016) ‘Supporting the development of biobanks in low and medium income countries’, IEEE, pp. 1–10 [Online]. DOI: 10.1109/ISTAFRICA.2016.7530672 (Accessed 23 August 2016).

Klingström, T., Soldatova, L., Stevens, R., Roos, T. E., Swertz, M. A., Müller, K. M., Kalaš, M., Lambrix, P., Taussig, M. J., Litton, J.-E., Landegren, U. and Bongcam-Rudloff, E. (2013) ‘Workshop on laboratory protocol standards for the molecular methods database’, *New Biotechnology*, vol. 30, no. 2, pp. 109–113 [Online]. DOI: 10.1016/j.nbt.2012.05.019.

de Koning, D.-J. (2016) ‘RAD Sequencing of Diverse Accessions of Lepidium campestre, a Target Species for Domestication as a Novel Oil Crop’, *Plant and Animal Genome XXIV Conference*, Plant and Animal Genome.

Konopka, B. M. (2015) ‘Biomedical ontologies—A review’, *Biocybernetics and Biomedical Engineering*, vol. 35, no. 2, pp. 75–86 [Online]. DOI: 10.1016/j.bbe.2014.06.002.

Kousta, S., Ferguson, C. and Ganley, E. (2016) ‘Meta-Research: Broadening the Scope of PLOS Biology’, *PLOS Biology*, vol. 14, no. 1, p. e1002334 [Online]. DOI: 10.1371/journal.pbio.1002334.

Kumar, S. and Dudley, J. (2007) ‘Bioinformatics software for biologists in the genomics era’, *Bioinformatics*, vol. 23, no. 14, pp. 1713–1717 [Online]. DOI: 10.1093/bioinformatics/btm239.

Lagos-Quintana, M., Rauhut, R., Lendeckel, W. and Tuschl, T. (2001) ‘Identification of novel genes coding for small expressed RNAs’, *Science (New York, N.Y.)*, vol. 294, no. 5543, pp. 853–858 [Online]. DOI: 10.1126/science.1064921.

Lammey, R. (2016) ‘Preprints are go at Crossref!’, *Crossef Blog* [Online]. Available at https://www.crossref.org/blog/preprints-are-go-at-crossref/ (Accessed 19 May 2017).

Lang, T. A. and Altman, D. G. (2015) ‘Basic statistical reporting for articles published in Biomedical Journals: The “Statistical Analyses and Methods in the Published Literature” or the SAMPL Guidelines’, *International Journal of Nursing Studies*, vol. 52, no. 1, pp. 5–9 [Online]. DOI: 10.1016/j.ijnurstu.2014.09.006.

Langseth, H., Gislefoss, R. E., Martinsen, J. I., Dillner, J. and Ursin, G. (2016) ‘Cohort Profile: The Janus Serum Bank Cohort in Norway’, *International Journal of Epidemiology*, p. dyw027 [Online]. DOI: 10.1093/ije/dyw027.

Lawrence, P. A. (2009) ‘Real Lives and White Lies in the Funding of Scientific Research’, *PLoS Biology*, vol. 7, no. 9, p. e1000197 [Online]. DOI: 10.1371/journal.pbio.1000197.

Lawson, T. (2004) ‘A conception of ontology’, *Mimeograph, University of Cambridge*.

Lee, J. D. and See, K. A. (2004) ‘Trust in Automation: Designing for Appropriate Reliance’, *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 46, no. 1, pp. 50–80 [Online]. DOI: 10.1518/hfes.46.1.50_30392.

Lee, R. C., Feinbaum, R. L. and Ambros, V. (1993) ‘The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14’, *Cell*, vol. 75, no. 5, pp. 843–854.

Leek, J. T. and Peng, R. D. (2015) ‘Opinion: Reproducible research can still be wrong: Adopting a prevention approach: Fig. 1.’, *Proceedings of the National Academy of Sciences*, vol. 112, no. 6, pp. 1645–1646 [Online]. DOI: 10.1073/pnas.1421412111.

Librizzi, M., Longo, A., Chiarelli, R., Amin, J., Spencer, J. and Luparello, C. (2012) ‘Cytotoxic Effects of Jay Amin Hydroxamic Acid (JAHA), a Ferrocene-Based Class I Histone Deacetylase Inhibitor, on Triple-Negative MDA-MB231 Breast Cancer Cells’, *Chemical Research in Toxicology*, vol. 25, no. 11, pp. 2608–2616 [Online]. DOI: 10.1021/tx300376h.

Litton, J.-E. (2017) ‘We must urgently clarify data-sharing rules’, *Nature*, vol. 541, no. 7638, pp. 437–437 [Online]. DOI: 10.1038/541437a.

Lozano, G. A., Larivière, V. and Gingras, Y. (2012) ‘The weakening relationship between the impact factor and papers’ citations in the digital age’, *Journal of the American Society for Information Science and Technology*, vol. 63, no. 11, pp. 2140–2145 [Online]. DOI: 10.1002/asi.22731.

Malentacchi, F., Ciniselli, C. M., Pazzagli, M., Verderio, P., Barraud, L., Hartmann, C. C., Pizzamiglio, S., Weisbuch, S., Wyrich, R. and Gelmini, S. (2015) ‘Influence of pre-analytical procedures on genomic DNA integrity in blood samples: The SPIDIA experience’, *Clinica Chimica Acta*, vol. 440, pp. 205–210 [Online]. DOI: 10.1016/j.cca.2014.12.004.

Malmberg, H., Stockholm University, Institute for International Economic Studies and Department of Economics (2017) *Humancapital in development accounting and other essays in economics*, Stockholm, Stockholms Universitet.

Marikanty, R., Gupta, M., Cherukuvada, S., Kompella, S., Prayaga, A., Konda, S., Polisetty, R., Idris, M., Rao, P., Chandak, G. and Dakshinamurty, K. (2016) ‘Identification of urinary proteins potentially associated with diabetic kidney disease', *Indian Journal of Nephrology*, vol. 26, no. 6, p. 434 [Online]. DOI: 10.4103/0971-4065.176144.

Matimba, A., Tybring, G., Chitereka, J., Zinyama-Gutsire, R., Dandara, C., Bürén, E., Dhoro, M. and Masimirembwa, C. (2016) 'Practical Approach to Biobanking in Zimbabwe: Establishment of an Inclusive Stakeholder Framework', *Biopreservation and Biobanking*, vol. 14, no. 5, pp. 440–446 [Online]. DOI: 10.1089/bio.2015.0043.

Matzke, E. A. M., O'Donoghue, S., Barnes, R. O., Daudt, H., Cheah, S., Suggitt, A., Bartlett, J., Damaraju, S., Johnston, R., Murphy, L., Shepherd, L., Mes-Masson, A.-M., Schacter, B. and Watson, P. H. (2012) 'Certification for Biobanks: The Program Developed by the Canadian Tumour Repository Network (CTRNet)', *Biopreservation and Biobanking*, vol. 10, no. 5, pp. 426–432 [Online]. DOI: 10.1089/bio.2012.0026.

Maurice-van Eijndhoven, M. H. T. (2014) *Genetic variation of milk fatty acid composition between and within dairy cattle breeds*, Wageningen, Wageningen University.

McKiernan, E. C., Bourne, P. E., Brown, C. T., Buck, S., Kenall, A., Lin, J., McDougall, D., Nosek, B. A., Ram, K., Soderberg, C. K., Spies, J. R., Thaney, K., Updegrove, A., Woo, K. H. and Yarkoni, T. (2016) 'How open science helps researchers succeed', *eLife*, vol. 5 [Online]. DOI: 10.7554/eLife.16800 (Accessed 1 June 2017).

McNutt, M. (2014) 'Raising the bar', *Science*, vol. 345, no. 6192, pp. 9–9 [Online]. DOI: 10.1126/science.1257891.

Medhat, E., Marzaban, R. N., Dwedar, R. A., Reda, A. M., Rashid, L. and Al-Enezi, T. (2017) 'Validity of Salivary Polymerase Chain Reaction in Diagnosis of Helicobacter pylori Among Egyptian Patients':, *Infectious Diseases in Clinical Practice*, vol. 25, no. 2, pp. 76–81 [Online]. DOI: 10.1097/IPC.0000000000000460.

Mendy, M., Caboux, E., Sylla, B. S., Dillner, J., Chinquee, J., Wild, C. and BCNet survey participants (2015) 'Infrastructure and Facilities for Human Biobanking in Low- and Middle-Income Countries: A Situation Analysis', *Pathobiology*, vol. 81, no. 5–6, pp. 252–260 [Online]. DOI: 10.1159/000362093.

Merino, R., Reichel, J. and Amatya, S. (2016) *D1.2 Data model for personal data protection and sharing and integration*, Annual report of the B3Africa project, Swedish University of Agricultural Sciences.

Merino-Martinez, R., Norlin, L., van Enckevort, D., Anton, G., Schuffenhauer, S., Silander, K., Mook, L., Holub, P., Bild, R., Swertz, M. and Litton, J.-E. (2016) 'Toward Global Biobank Integration by Implementation of the Minimum Information About BIobank Data Sharing (MIABIS 2.0 Core)', *Biopreservation and Biobanking*, vol. 14, no. 4, pp. 298–306 [Online]. DOI: 10.1089/bio.2015.0070.

Merton, R. K. (1942) 'A note on science and democracy', *J. Legal & Pol. Soc.*, vol. 1, p. 115.

Moen, T., Torgersen, J., Santi, N., Davidson, W. S., Baranski, M., Ødegård, J., Kjøglum, S., Velle, B., Kent, M., Lubieniecki, K. P., Isdal, E. and Lien, S. (2015) 'Epithelial Cadherin Determines Resistance to Infectious Pancreatic Necrosis Virus in Atlantic Salmon', *Genetics*, vol. 200, no. 4, pp. 1313–1326 [Online]. DOI: 10.1534/genetics.115.175406.

Moonesinghe, R., Khoury, M. J. and Janssens, A. C. J. W. (2007) 'Most Published Research Findings Are False—But a Little Replication Goes a Long Way', *PLoS Medicine*, vol. 4, no. 2, p. e28 [Online]. DOI: 10.1371/journal.pmed.0040028.

Moore, H. M., Compton, C. C., Alper, J. and Vaught, J. B. (2011) ‘International Approaches to Advancing Biospecimen Science’, *Cancer Epidemiology Biomarkers & Prevention*, vol. 20, no. 5, pp. 729–732 [Online]. DOI: 10.1158/1055-9965.EPI-11-0021.

Moore, Helen M., Kelly, A. B., Jewell, S. D., McShane, L. M., Clark, D. P., Greenspan, R., Hayes, D. F., Hainaut, P., Kim, P., Mansfield, E., Potapova, O., Riegman, P., Rubinstein, Y., Seijo, E., Somiari, S., Watson, P., Weier, H.-U., Zhu, C. and Vaught, J. (2011) ‘Biospecimen Reporting for Improved Study Quality (BRISQ)’, *Journal of Proteome Research*, vol. 10, no. 8, pp. 3429–3438 [Online]. DOI: 10.1021/pr200021n.

Mulder, N. J., Adebiyi, E., Alami, R., Benkahla, A., Brandful, J., Doumbia, S., Everett, D., Fadlelmola, F. M., Gaboun, F., Gaseitsiwe, S., Ghazal, H., Hazelhurst, S., Hide, W., Ibrahimi, A., Jaufeerally Fakim, Y., Jongeneel, C. V., Joubert, F., Kassim, S., Kayondo, J., Kumuthini, J., Lyantagaye, S., Makani, J., Mansour Alzohairy, A., Masiga, D., Moussa, A., Nash, O., Ouwe Missi Oukem-Boyer, O., Owusu-Dabo, E., Panji, S., Patterton, H., Radouani, F., Sadki, K., Seghrouchni, F., Tastan Bishop, Ö., Tiffin, N., Ulenga, N. and The H3ABioNet Consortium (2016) ‘H3ABioNet, a sustainable pan-African bioinformatics network for human heredity and health in Africa’, *Genome Research*, vol. 26, no. 2, pp. 271–277 [Online]. DOI: 10.1101/gr.196295.115.

Müller, H., Malservet, N., Quinlan, P., Reihs, R., Penicaud, M., Chami, A., Zatloukal, K. and Dagher, G. (2017) ‘From the evaluation of existing solutions to an all-inclusive package for biobanks’, *Health and Technology*, vol. 7, no. 1, pp. 89–95 [Online]. DOI: 10.1007/s12553-016-0175-x.

National Research Council (US) Committee on Undergraduate Biology Education to Prepare Research Scientists for the 21st Century (2003) *Bio2010: Transforming Undergraduate Education for Future Research Biologists*, The National Academies Collection: Reports funded by National Institutes of Health, Washington (DC), National Academies Press (US) [Online]. Available at http://www.ncbi.nlm.nih.gov/books/NBK43511/ (Accessed 3 May 2017).

Neill, U. S. (2008) ‘Publish or perish, but at what cost?’, *Journal of Clinical Investigation*, vol. 118, no. 7, pp. 2368–2368 [Online]. DOI: 10.1172/JCI36371.

Nel, P. (2004) *Dr. Seuss: American icon*, New York, Continuum.

Neururer, S. B., Hofer, P. and Göbel, G. (2016) ‘An IT-Supported Evaluation Tool for Biobanks Based on International Guidelines to Improve the Biosample Quality’, *Stud Health Technol Inform*, vol. 223, pp. 46–53.

Nicholls, S. G., Hayes, T. P., Brehaut, J. C., McDonald, M., Weijer, C., Saginur, R. and Fergusson, D. (2015) ‘A Scoping Review of Empirical Research Relating to Quality and Effectiveness of Research Ethics Review’, Bayer, A. (ed), *PLOS ONE*, vol. 10, no. 7, p. e0133639 [Online]. DOI: 10.1371/journal.pone.0133639.

NIH Grants (n.d.) ‘Reporting Preprints and Other Interim Research Products’, Announcement [Online]. Available at https://grants.nih.gov/grants/guide/notice-files/NOT-OD-17-050.html (Accessed 19 May 2017).

Nissen, S. B., Magidson, T., Gross, K. and Bergstrom, C. T. (2016) ‘Publication bias and the canonization of false facts’, *eLife*, vol. 5 [Online]. DOI: 10.7554/eLife.21451 (Accessed 29 September 2017).

Nolan, T. W. (2000) ‘System changes to improve patient safety’, *BMJ : British Medical Journal*, vol. 320, no. 7237, pp. 771–773.

Nordgren, J. and Uppsala universitet (2007) *Quality and renewal 2007: an overall evaluation of research at Uppsala University 2006/2007*, Uppsala, Uppsala universitet.

Norlin, L., Fransson, M. N., Eriksson, M., Merino-Martinez, R., Anderberg, M., Kurtovic, S. and Litton, J.-E. (2012) ‘A Minimum Data Set for Sharing Biobank Samples, Information, and Data: MIABIS’, *Biopreservation and Biobanking*, vol. 10, no. 4, pp. 343–348 [Online]. DOI: 10.1089/bio.2012.0003.

Notter, D. R. (1999) ‘The importance of genetic diversity in livestock populations of the future.’, *Journal of Animal Science*, vol. 77, no. 1, p. 61 [Online]. DOI: 10.2527/1999.77161x.

Oliphant, T. (2016) ‘Anaconda and Hadoop --- a story of the journey and where we are now.’, *Technical Discovery*, Blog [Online]. Available at http://technicaldiscovery.blogspot.se/2016/03/anaconda-and-hadoop-story-of-journey.html (Accessed 7 April 2017).

Ouzounis, C. A. and Valencia, A. (2003) ‘Early bioinformatics: the birth of a discipline--a personal view’, *Bioinformatics*, vol. 19, no. 17, pp. 2176–2190 [Online]. DOI: 10.1093/bioinformatics/btg309.

van Panhuis, W. G., Paul, P., Emerson, C., Grefenstette, J., Wilder, R., Herbst, A. J., Heymann, D. and Burke, D. S. (2014) ‘A systematic review of barriers to data sharing in public health’, *BMC Public Health*, vol. 14, no. 1 [Online]. DOI: 10.1186/1471-2458-14-1144 (Accessed 9 May 2017).

Papatheodorou, S. I., Trikalinos, T. A. and Ioannidis, J. P. A. (2008) ‘Inflated numbers of authors over time have not been just due to increasing research complexity’, *Journal of Clinical Epidemiology*, vol. 61, no. 6, pp. 546–551 [Online]. DOI: 10.1016/j.jclinepi.2007.07.017.

Park, S., Venkat, A., Gopinath, A. and Kang, J. (2015) ‘Quantitative Analysis of the Trends Exhibited by the Three Interdisciplinary Biological Sciences: Biophysics, Bioinformatics, and Systems Biology’, *Journal of Microbiology & Biology Education*, vol. 16, no. 2, pp. 198–202 [Online]. DOI: 10.1128/jmbe.v16i2.949.

Pellegrino, E. D. (1999) ‘The origins and evolution of bioethics: some personal reflections’, *Kennedy Institute of Ethics Journal*, vol. 9, no. 1, pp. 73–88.

Peng, R. (2015) ‘The reproducibility crisis in science: A statistical counterattack’, *Significance*, vol. 12, no. 3, pp. 30–32 [Online]. DOI: 10.1111/j.1740-9713.2015.00827.x.

Pettersson, I. and Söder, I. (2016) *Standard för svensk indelning av forskningsämnen 2011, uppdaterad augusti 2016*, Stockholm, Universitetskanslersämbetet [Online]. Available at http://www.scb.se/contentassets/10054f2ef27c437884e8cde0d38b9cc4/standard-for-svensk-indelning--av-forskningsamnen-2011-uppdaterad-aug-2016.pdf.

Powers, M. (2005) ‘Bioethics as Politics: The Limits of Moral Expertise’, *Kennedy Institute of Ethics Journal*, vol. 15, no. 3, pp. 305–322 [Online]. DOI: 10.1353/ken.2005.0023.

Prinz, F., Schlange, T. and Asadullah, K. (2011) ‘Believe it or not: how much can we rely on published data on potential drug targets?’, *Nature Reviews Drug Discovery*, vol. 10, no. 9, pp. 712–712 [Online]. DOI: 10.1038/nrd3439-c1.

Reardon, S. (2017) ‘Geneticists pan paper that claims to predict a person’s face from their DNA’, *Nature*, vol. 549, no. 7671, pp. 139–140 [Online]. DOI: 10.1038/nature.2017.22580.

Rhoads, A. and Au, K. F. (2015) ‘PacBio Sequencing and Its Applications’, *Genomics, Proteomics & Bioinformatics*, vol. 13, no. 5, pp. 278–289 [Online]. DOI: 10.1016/j.gpb.2015.08.002.

van Rijswijk, M., Beirnaert, C., Caron, C., Cascante, M., Dominguez, V., Dunn, W. B., Ebbels, T. M. D., Giacomoni, F., Gonzalez-Beltran, A., Hankemeier, T., Haug, K., Izquierdo-Garcia, J. L., Jimenez, R. C., Jourdan, F., Kale, N., Klapa, M. I., Kohlbacher, O., Koort, K., Kultima, K., Le Corguillé, G., Moschonas, N. K., Neumann, S., O'Donovan, C., Reczko, M., Rocca-Serra, P., Rosato, A., Salek, R. M., Sansone, S.-A., Satagopam, V., Schober, D., Shimmo, R., Spicer, R. A., Spjuth, O., Thévenot, E. A., Viant, M. R., Weber, R. J. M., Willighagen, E. L., Zanetti, G. and Steinbeck, C. (2017) ‘The future of metabolomics in ELIXIR’, *F1000Research*, vol. 6, p. 1649 [Online]. DOI: 10.12688/f1000research.12342.1.

Ritchey, T. (2013) ‘Wicked Problems, Modelling Social Messes with Morphological Analysis.’, *Acta Morphologica Generalis*, vol. 2, no. 1.

Rittel, H. W. J. and Webber, M. M. (1973) ‘Dilemmas in a general theory of planning’, *Policy Sciences*, vol. 4, no. 2, pp. 155–169 [Online]. DOI: 10.1007/BF01405730.

Robb, J. A., Gulley, M. L., Fitzgibbons, P. L., Kennedy, M. F., Cosentino, L. M., Washington, K., Dash, R. C., Branton, P. A., Jewell, S. D. and Lapham, R. L. (2014) ‘A Call to Standardize Preanalytic Data Elements for Biospecimens’, *Archives of Pathology & Laboratory Medicine*, vol. 138, no. 4, pp. 526–537 [Online]. DOI: 10.5858/arpa.2013-0250-CP.

Rogers, J., Carolin, T., Vaught, J. and Compton, C. (2011) ‘Biobankonomics: A Taxonomy for Evaluating the Economic Benefits of Standardized Centralized Human Biobanking for Translational Research’, *JNCI Monographs*, vol. 2011, no. 42, pp. 32–38 [Online]. DOI: 10.1093/jncimonographs/lgr010.

Rounge, T. B., Lauritzen, M., Langseth, H., Enerly, E., Lyle, R. and Gislefoss, R. E. (2015) ‘microRNA Biomarker Discovery and High-Throughput DNA Sequencing Are Possible Using Long-term Archived Serum Samples’, *Cancer Epidemiology Biomarkers & Prevention*, vol. 24, no. 9, pp. 1381–1387 [Online]. DOI: 10.1158/1055-9965.EPI-15-0289.

Rowhani-Farid, A., Allen, M. and Barnett, A. G. (2017) ‘What incentives increase data sharing in health and medical research? A systematic review’, *Research Integrity and Peer Review*, vol. 2, no. 1 [Online]. DOI: 10.1186/s41073-017-0028-9 (Accessed 28 May 2017).

Royal Society (Great Britain) (2012) *Science as an open enterprise: open data for open science.*, London, Royal Society.

Salman, R. A.-S., Beller, E., Kagan, J., Hemminki, E., Phillips, R. S., Savulescu, J., Macleod, M., Wisely, J. and Chalmers, I. (2014) ‘Increasing value and reducing waste in biomedical research regulation and management’, *The Lancet*, vol. 383, no. 9912, pp. 176–185 [Online]. DOI: 10.1016/S0140-6736(13)62297-7.

Satel, S. (2015) ‘The Bioethics dilemma’, *Pacific Standard*, no. Web Issue [Online]. Available at https://psmag.com/social-justice/steven-pinker-and-the-real-value-of-bioethicists (Accessed 18 May 2017).

Saxena, R. K., Aziz, A. S., Kalekar, M. G., Mol, J., Suryakar, A. N., Tabita, B., Shirahatti, R. V. and Medikeri, R. S. (2013) ‘Presence of Helicobacter pylori Detected by PCR in Saliva of Male Smokers and Non Smokers with Chronic Periodontitis.’,.

Scherzinger, G. and Bobbert, M. (2017) ‘Evaluation of Research Ethics Committees: Criteria for the Ethical Quality of the Review Process’, *Accountability in Research*, vol. 24, no. 3, pp. 152–176 [Online]. DOI: 10.1080/08989621.2016.1273778.

Schuyler, P. L., Hole, W. T., Tuttle, M. S. and Sherertz, D. D. (1993) ‘The UMLS Metathesaurus: representing different views of biomedical concepts.’, *Bulletin of the Medical Library Association*, vol. 81, no. 2, pp. 217–222.

Silber, T. J. (1982) ‘Bioethics: An Interdisciplinary Enterprise’, *Journal of Religion and Health*, vol. 21, no. 1, pp. 21–28.

Simeon-Dubach, D., Burt, A. D. and Hall, P. A. (2012) ‘Quality really matters: the need to improve specimen quality in biomedical research: Editorial’, *The Journal of Pathology*, vol. 228, no. 4, pp. 431–433 [Online]. DOI: 10.1002/path.4117.

Simera, I., Altman, D. G., Moher, D., Schulz, K. F. and Hoey, J. (2008) ‘Guidelines for Reporting Health Research: The EQUATOR Network’s Survey of Guideline Authors’, *PLoS Medicine*, vol. 5, no. 6, p. e139 [Online]. DOI: 10.1371/journal.pmed.0050139.

Simera, I., Moher, D., Hirst, A., Hoey, J., Schulz, K. F. and Altman, D. G. (2010) ‘Transparent and accurate reporting increases reliability, utility, and impact of your research: reporting guidelines and the EQUATOR Network’, *BMC Medicine*, vol. 8, no. 1 [Online]. DOI: 10.1186/1741-7015-8-24 (Accessed 1 June 2017).

Singh Chawla, D. (2017) ‘When a preprint becomes the final paper’, *Nature* [Online]. DOI: 10.1038/nature.2017.21333 (Accessed 19 May 2017).

Sloggett, C., Goonasekera, N. and Afgan, E. (2013) ‘BioBlend: automating pipeline analyses within Galaxy and CloudMan’, *Bioinformatics*, vol. 29, no. 13, pp. 1685–1686 [Online]. DOI: 10.1093/bioinformatics/btt199.

Smith, A. (1786) ‘An Inquiry into the Nature and Causes of the Wealth of Nations.’,.

Smith, B. (2004) ‘Beyond concepts: ontology as reality representation’, *Proceedings of the third international conference on formal ontology in information systems (FOIS 2004)*, IOS Press, Amsterdam, pp. 73–84.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Scheuermann, R. H., Shah, N., Whetzel, P. L. and Lewis, S. (2007) ‘The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration’, *Nature Biotechnology*, vol. 25, no. 11, pp. 1251–1255 [Online]. DOI: 10.1038/nbt1346.

Smith, B., Köhler, J. and Kumar, A. (2004) ‘On the Application of Formal Principles to Life Science Data: a Case Study in the Gene Ontology’, in Rahm, E. (ed), *Data Integration in the Life Sciences*, Berlin, Heidelberg, Springer Berlin Heidelberg, vol. 2994, pp. 79–94 [Online]. DOI: 10.1007/978-3-540-24745-6_6 (Accessed 26 May 2017).

Smith, B., Kusnierczyk, W., Schober, D. and Ceusters, W. (2006) ‘Towards a reference terminology for ontology research and development in the biomedical domain’, *Proc. of KR-MED 2006*, pp. 57–66.

Smith, D. R. (2013) ‘The battle for user-friendly bioinformatics’, *Frontiers in Genetics*, vol. 4 [Online]. DOI: 10.3389/fgene.2013.00187 (Accessed 23 May 2017).

Smith, D. R. (2016) ‘Goodbye genome paper, hello genome report: the increasing popularity of “genome announcements” and their impact on science: Table 1.’, *Briefings in Functional Genomics*, p. elw026 [Online]. DOI: 10.1093/bfgp/elw026.

Soldatova, L. N., Aubrey, W., King, R. D. and Clare, A. (2008) ‘The EXACT description of biomedical protocols’, *Bioinformatics*, vol. 24, no. 13, pp. i295–i303 [Online]. DOI: 10.1093/bioinformatics/btn156.

Soldatova, L. N., Nadis, D., King, R. D., Basu, P. S., Haddi, E., Baumlé, V., Saunders, N. J., Marwan, W. and Rudkin, B. B. (2014) ‘EXACT2: the semantics of biomedical protocols’, *BMC Bioinformatics*, vol. 15, no. Suppl 14, p. S5 [Online]. DOI: 10.1186/1471-2105-15-S14-S5.

Soo, C. C., Mukomana, F., Hazelhurst, S. and Ramsay, M. (2017) ‘Establishing an academic biobank in a resource-challenged environment’, *South African Medical Journal*, vol. 107, no. 6, p. 486 [Online]. DOI: 10.7196/SAMJ.2017.v107i6.12099.

Stephan, P. (2012) ‘Research efficiency: Perverse incentives’, *Nature*, vol. 484, no. 7392, pp. 29–31 [Online]. DOI: 10.1038/484029a.

Tan, T., Lim, S., Khan, A. M. and Ranganathan, S. (2009a) ‘A proposed minimum skill set for university graduates to meet the informatics needs and challenges of the “-omics” era’, *BMC Genomics*, vol. 10, no. Suppl 3, p. S36 [Online]. DOI: 10.1186/1471-2164-10-S3-S36.

Tan, T., Lim, S., Khan, A. M. and Ranganathan, S. (2009b) ‘A proposed minimum skill set for university graduates to meet the informatics needs and challenges of the “-omics” era’, *BMC Genomics*, vol. 10, no. Suppl 3, p. S36 [Online]. DOI: 10.1186/1471-2164-10-S3-S36.

Teitelbaum, M. S. (2008) ‘RESEARCH FUNDING: Structural Disequilibria in Biomedical Research’, *Science*, vol. 321, no. 5889, pp. 644–645 [Online]. DOI: 10.1126/science.1160272.

Teytelman, L., Stoliartchouk, A., Kindler, L. and Hurwitz, B. L. (2016) ‘Protocols.io: Virtual Communities for Protocol Development and Discussion’, *PLOS Biology*, vol. 14, no. 8, p. e1002538 [Online]. DOI: 10.1371/journal.pbio.1002538.

The Expert Advisory Group on Data Access (2014) *Establishing incentives and changing cultures to support data access*, London, United Kingdom, the Wellcome Trust [Online]. Available at https://wellcome.ac.uk/sites/default/files/establishing-incentives-and-changing-cultures-to-support-data-access-eagda-may14.pdf (Accessed 15 October 2017).

The H3Africa Consortium, Matovu, E., Bucheton, B., Chisi, J., Enyaru, J., Hertz-Fowler, C., Koffi, M., Macleod, A., Mumba, D., Sidibe, I., Simo, G., Simuunza, M., Mayosi, B., Ramesar, R., Mulder, N., Ogendo, S., Mocumbi, A. O., Hugo-Hamman, C., Ogah, O., El Sayed, A., Mondo, C., Musuku, J., Engel, M., De Vries, J., Lesosky, M., Shaboodien, G., Cordell, H., Pare, G., Keavney, B., Motala, A., Sobngwi, E., Mbanya, J. C., Hennig, B., Balde, N., Nyirenda, M., Oli, J., Adebamowo, C., Levitt, N., Mayige, M., Kapiga, S., Kaleebu, P., Sandhu, M., Smeeth, L., McCarthy, M. and Rotimi, C. (2014) ‘Enabling the genomic revolution in Africa’, *Science*, vol. 344, no. 6190, pp. 1346–1348 [Online]. DOI: 10.1126/science.1251546.

Thiese, M. S., Arnold, Z. C. and Walker, S. D. (2015) ‘The misuse and abuse of statistics in biomedical research’, *Biochemia Medica*, pp. 5–11 [Online]. DOI: 10.11613/BM.2015.001.

Thompson, R., Johnston, L., Taruscio, D., Monaco, L., Béroud, C., Gut, I. G., Hansson, M. G., 't Hoen, P.-B. A., Patrinos, G. P., Dawkins, H., Ensini, M., Zatloukal, K., Koubi, D., Heslop, E., Paschall, J. E., Posada, M., Robinson, P. N., Bushby, K. and Lochmüller, H. (2014) 'RD-Connect: An Integrated Platform Connecting Databases, Registries, Biobanks and Clinical Bioinformatics for Rare Disease Research', *Journal of General Internal Medicine*, vol. 29, no. S3, pp. 780–787 [Online]. DOI: 10.1007/s11606-014-2908-8.

Till, J. E. (2001) 'Predecessors of preprint servers', *Learned Publishing*, vol. 14, no. 1, pp. 7–13 [Online]. DOI: 10.1087/09531510125100214.

Trollvad, S., Ali Khan, I. and Malm, P. (2012) *Project MolMeth A management system for molecular analysis protocols*, Final Report, Uppsala University, Department of Information Technology, Uppsala University [Online]. Available at https://it.uu.se/edu/course/homepage/lims/vt13/ex-project-reports/finalReportLIMS(v7).pdf (Accessed 27 May 2017).

Uppsala universitet (2011) *Quality and Renewal 2011: Kvalitet och Förnyelse 2011 (KoF11) : an overall evaluation of research at Uppsala University 2010/2011*, Uppsala, Uppsala universitet.

Vale, R. D. (2015) 'Accelerating scientific publication in biology', *Proceedings of the National Academy of Sciences*, vol. 112, no. 44, pp. 13439–13446 [Online]. DOI: 10.1073/pnas.1511912112.

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S. and DePristo, M. A. (2013) 'From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline: The Genome Analysis Toolkit Best Practices Pipeline', in Bateman, A., Pearson, W. R., Stein, L. D., Stormo, G. D., and Yates, J. R. (eds), *Current Protocols in Bioinformatics*, Hoboken, NJ, USA, John Wiley & Sons, Inc., p. 11.10.1-11.10.33 [Online]. DOI: 10.1002/0471250953.bi1110s43 (Accessed 25 June 2017).

Van Noorden, R., Maher, B. and Nuzzo, R. (2014) 'The top 100 papers', *Nature*, vol. 514, no. 7524, pp. 550–553 [Online]. DOI: 10.1038/514550a.

Vincent, A. T. and Charette, S. J. (2015) 'Who qualifies to be a bioinformatician?', *Frontiers in Genetics*, vol. 6, p. 164 [Online]. DOI: 10.3389/fgene.2015.00164.

Voytek, B. (2016) 'The Virtuous Cycle of a Data Ecosystem', Bourne, P. E. (ed), *PLOS Computational Biology*, vol. 12, no. 8, p. e1005037 [Online]. DOI: 10.1371/journal.pcbi.1005037.

de Vries, J., Tindana, P., Littler, K., Ramsay, M., Rotimi, C., Abayomi, A., Mulder, N. and Mayosi, B. M. (2015) 'The H3Africa policy framework: negotiating fairness in genomics', *Trends in Genetics*, vol. 31, no. 3, pp. 117–119 [Online]. DOI: 10.1016/j.tig.2014.11.004.

Warnich, L., I. Drogemoller, B., S. Pepper, M., Dandara, C. and E.B. Wright, G. (2011) 'Pharmacogenomic Research in South Africa: Lessons Learned and Future Opportunities in the Rainbow Nation', *Current Pharmacogenomics and Personalized Medicine*, vol. 9, no. 3, pp. 191–207 [Online]. DOI: doi:10.2174/187569211796957575.

Weeks, W. B., Wallace, A. E. and Kimberly, B. C. S. (2004) ‘Changes in authorship patterns in prestigious US medical journals’, *Social Science & Medicine*, vol. 59, no. 9, pp. 1949–1954 [Online]. DOI: 10.1016/j.socscimed.2004.02.029.

Wightman, B., Ha, I. and Ruvkun, G. (1993) ‘Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans’, *Cell*, vol. 75, no. 5, pp. 855–862.

Wightman, B. and Hark, A. T. (2012) ‘Integration of bioinformatics into an undergraduate biology curriculum and the impact on development of mathematical skills’, *Biochemistry and Molecular Biology Education*, vol. 40, no. 5, pp. 310–319 [Online]. DOI: 10.1002/bmb.20637.

Yuille, M., van Ommen, G.-J., Brechot, C., Cambon-Thomsen, A., Dagher, G., Landegren, U., Litton, J.-E., Pasterk, M., Peltonen, L., Taussig, M., Wichmann, H.-E. and Zatloukal, K. (2007) ‘Biobanking for Europe’, *Briefings in Bioinformatics*, vol. 9, no. 1, pp. 14–24 [Online]. DOI: 10.1093/bib/bbm050.

Zerhouni, E. A. and Nabel, E. G. (2008) ‘Protecting Aggregate Genomic Data’, *Science*, vol. 322, no. 5898, p. 44a–44a [Online]. DOI: 10.1126/science.1165490.

Popular science summary

New technologies in life science constantly create new opportunities for researchers to generate data for analysis. In life science, this increased output has outpaced developments in sample collection, processing power, and storage space. As a result researchers struggle to keep up to date with new technologies while also struggling to handle increasingly complex challenges related to logistics and information technology.

Research infrastructures such as the Swedish Science for Life Laboratory, numerous biobanks and the National Bioinformatics Infrastructure Sweden have been established to support researchers in handling these issues. Dedicated infrastructures enable researchers to work more effectively but can only offer limited customisation for the specific needs of individual researchers. This is especially noticeable in fields such as bioinformatics which apart from biological knowledge require an understanding of computer science and statistics. In many countries there is also a lack of long term funding to establish infrastructures. SLU Bioinformatics Centre therefore works to develop networks and tools to empower researchers across the world and leverage the capacity of bioinformatics into their research. This thesis describes studies made to support the development of the eB3Kit in the B3Africa project. With the eB3Kit researcher gains access to a comprehensive informatics platform supporting researchers with sample collection, data analysis and data storage in compliance with good research practice and laws regarding personal privacy.

To make technical platforms relevant for researchers it is important that it addresses the needs of researchers on an individual level. New technology and the development of science makes researchers increasingly reliant on team efforts and collaboration but academic achievements are still judged on an individual level. This makes a solution like the eB3Kit attractive as it enable researchers to automate work outside their own field of expertise. Researchers can thereby quickly establish structured sample collections and use bioinformatics in their work while maximising the time available for specialisation within their chosen fields of expertise.

Populärvetenskaplig sammanfattning

Ny teknik skapar hela tiden nya möjligheter för forskare att planera och genomföra experiment. Dessa nya metoder gör det möjligt för forskare att producera mer och helt nya typer av data än tidigare vilket gör analysarbetet allt mer komplicerat. Inom livsvetenskaperna har dessutom förmågan att generera data ökat snabbare än kapaciteten för provinsamling, dataanalys och lagring. Eftersom forskare i hög grad är specialiserade inom sina respektive ämnesområden skapar detta problem då datahantering och logistik blir en allt mer krävande del av arbetet.

För att stödja forskning har viktiga infrastrukturer såsom Science for Life Laboratory, biobanker och den nationella bioinformatikinfrastrukturen NBIS etablerats i Sverige för att stödja forskare med krävande delmoment. Sådana infrastrukturer kräver långsiktig finansiering och kan bara i mycket begränsad omfattning anpassas efter enskilda forskares behov. Det här är särskilt tydligt inom bioinformatik som bygger på en blandning av biologi, datavetenskap och statistik. Många andra länder saknar även de resurser som krävs för att etablera långsiktigt hållbara infrastrukturer. SLU Global Bioinformatics Centre samarbetar därför med forskargrupper i många länder i syfte att stödja deras arbete och etablera nätverk med bioinformatiker som kan analysera biologiska data. Denna avhandling beskriver arbetet med att studera vilket stöd forskare behöver och använda dessa slutsatser för att bidra till utvecklingen av plattformen ”the eB3Kit” inom ramen för B3Africa-projektet. Målet är att skapa en teknisk plattform som ger forskare stöd med provhantering, dataanalys och datalagring i enlighet med lagar och riktlinjer för integritetsskydd och kvalitetsarbete.

Den övergripande slutsatsen från studierna är att ny teknik skapar svårigheter inom universitetsforskning då akademisk meritering bedöms på individnivå samtidigt som ny teknik hela tiden skapar behov av större team och samarbeten. För att komma förbi detta är det viktigt att tekniska lösningar gör det möjligt att automatisera processer utanför det egna kompetensområdet och att forskare därigenom ges möjlighet att specialisera sig inom det egna forskningsområdet. Med eB3Kit får forskare möjlighet att etablera strukturerade provsamlingar även med begränsade investeringar. Vilket gör det möjligt att snabbt etablera biobanker vid sjukdomsutbrott och öka produktiviteten likväl som kvaliteten hos etablerade forskargrupper då tidsödande arbete utanför det egna kompetensområdet minimieras.

Acknowledgements

As a doctoral student in Sweden there are many things to be thankful for. It is a privilege to live in a geopolitically stable and wealthy region and my entire education has been provided for by the Swedish tax payers. More specifically this doctoral project has been funded the Swedish Research Council grant agreement no. 829-2009-6285 for work on the Molecular Methods database within BBMRI.se.

By the European Community's 7th Framework Programme (FP7/2007–2013) grant agreement no. 313010 for work with ethics and quality management as a part of the BBMRI-LPC project and the B3Africa project is supported by the European Union s Horizon 2020 research and innovation programme under grant agreement No 654404.

On a more personal level I have enjoyed the support of a loving family, good friends, great teachers, inspiring colleagues and admirable supervisors throughout slightly more than three decades of continued learning.

Meeting Report

Workshop on laboratory protocol standards for the molecular methods database

Tomas Klingström, Larissa Soldatova, Robert Stevens, T. Erik Roos and Morris A. Swertz, Kristian M. Müller, Matúš Kalaš[1,2], Patrick Lambrix, Michael J. Taussig, Jan-Eric Litton, Ulf Landegren and Erik Bongcam-Rudloff[1,2,*], Erik.Bongcam@slu.se

Management of data to produce scientific knowledge is a key challenge for biological research in the 21st century. Emerging high-throughput technologies allow life science researchers to produce big data at speeds and in amounts that were unthinkable just a few years ago. This places high demands on all aspects of the workflow: from data capture (including the experimental constraints of the experiment), analysis and preservation, to peer-reviewed publication of results. Failure to recognise the issues at each level can lead to serious conflicts and mistakes; research may then be compromised as a result of the publication of non-coherent protocols, or the misinterpretation of published data. In this report, we present the results from a workshop that was organised to create an ontological data-modelling framework for Laboratory Protocol Standards for the Molecular Methods Database (MolMeth). The workshop provided a set of short- and long-term goals for the MolMeth database, the most important being the decision to use the established EXACT description of biomedical ontologies as a starting point.

Introduction

The Molecular Methods database (MolMeth) is a structured database intended to provide researchers with an efficient resource to create, develop and publish life science laboratory protocols. It is available as an early beta version at http://www.molmeth.org. Using MolMeth, a researcher should be able to find relevant lab protocols quickly, and to use one or more protocols to create an individualised workflow. The user should also be able to publish the protocols through MolMeth to make them available as peer-reviewed articles via stable accession numbers. To achieve such a goal, it is important to have a common schema and vocabulary to describe laboratory protocols, along with a common understanding of the entities in the domain to make descriptions of those protocols: ontologies are now commonly used to provide such common understandings of the entities within a field of interest. With the use of an effective ontology, built using accepted standards, it becomes easier to create a coherent environment, where laboratory protocols are integrated with resources made available by biobanks, scientific literature and best-practice protocols supported by commercial providers.

This workshop, held in Uppsala on November 15th, 2011, served to initiate the effort to create an ontological data-modelling framework for MolMeth. It was divided into a series of lectures, and a session to receive input from experts regarding the development of MolMeth and its supporting ontologies. The six lectures were focused on the development of ontologies, relevant examples of ontologies, on web services that could provide examples and standards to which the MolMeth developers could adhere, and flexible database implementations for local laboratories to adopt easily.

Key points from lectures

Agile development of an ontology

The Robot Scientist project presented by Dr Larisa Soldatova (Aberystwyth University, UK) aims to develop a computer system capable of planning and conducting its own experiments as well as interpreting the results [1,2]. To support this system the scientists at Aberystwyth University have created the LABORS and EXACT ontologies [3] to provide open access to Robot Scientist experimental data (LABORS) and laboratory protocols that can be interpreted by a

1871-6784/$ - see front matter http://dx.doi.org/10.1016/j.nbt.2012.05.019

fully automated system (EXACT). This fully computerised environment does not possess the ability to process and interpret natural language that we take for granted when writing laboratory protocols intended for other human beings. EXACT can therefore be considered an effective 'upper limit' on the amount of information necessary to replicate laboratory experiments as all information is explicitly recorded (Fig. 1 for an example comparing natural language and EXACT instructions). For humans it is, however, possible to remove excess information as we can understand many steps implicitly and are likely to miss important information hidden behind unnecessary text blocks. It would for example be highly unfortunate if a researcher read the 'remove lid' instruction whilst missing 'in a fume hood'. Therefore it is necessary to carefully evaluate the EXACT ontology and to improve it according to human needs.

Dr Robert Stevens (University of Manchester, UK) presented a set of guidelines for the quick and efficient development of the basic input to form an ontology. He also reported practical experience from his work with the Software Ontology Project [4] and via collaborators from the Ontology for Biomedical Investigations (OBI) [5]. This framework consists of several key points for the agile development of ontologies:

- Iterative and incremental;
- Evolving requirements and solutions;
- Self-organising and cross-functional teams;
- Short time boxes; rapid and responsive development;
- Doing what is important first;
- Users are embedded in the process as first class citizens;
- Test driven; regular and frequent builds.

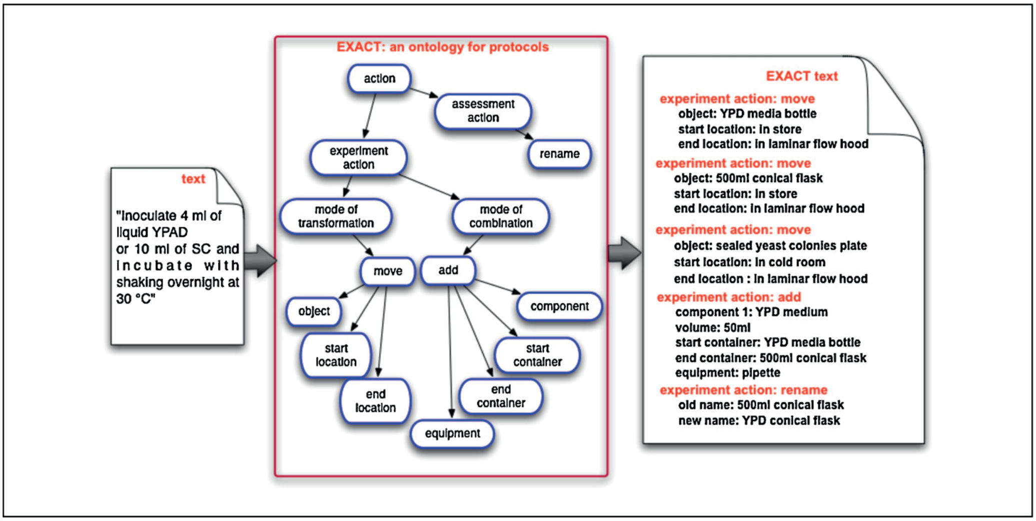
To achieve this, MolMeth will develop its ontology out of EXACT by removing ontological terms necessary to a computer but implicitly understood by a human researcher. The ontology will then be iteratively improved according to input from the users and external developers.

It is not uncommon that such development leads to unintended defects in the ontology network as the ontology is remodelled and extended. Dr Patrick Lambrix (Linköping University, Sweden) addressed many of these concerns as he presented the RepOSE environment for repairing ontologies [6]. Syntactic defects such as misspellings can be easy to find, but issues such as inconsistencies and missing connections are harder to find because they require extensive domain expertise and the careful study of the ontology. RepOSE automates much of this process that removes much time consuming manual labor from the process.

Existing web systems with integrated ontologies

The Embrace Data and Methods (EDAM) ontology [7,8] presented by Matúš Kalaš (University of Bergen, Norway) has been developed to support the categorisation of bioinformatics resources, such as eventually the web services collected in Biocatalogue [9]. Integration of MolMeth with ontologies such as EDAM and with information standards under the Minimum Information for Biological and Biomedical Investigations (MIBBI) [10] is also desirable to enhance the end user experience. Such integration in combination with the possibility of collaborating with organisations such as the Registry of Standard Biological parts were discussed by Dr Kristian Müller (University of Potsdam, Germany), who is part of a team creating services for laboratory protocols on smart phones. Their mobile app enables downloading of preformatted protocols from the Internet and provides interactive features such as note taking, barcode reading, countdown timing, time stamping, and log file generation. The protocols are stored in an XML property list (plist) format and are used by the iGEM team of the University of Potsdam.

Erik Roos and Dr Morris Swertz (University Medical Center Groningen, The Netherlands) presented the MOLGENIS application suite [11]. At its core there is a generic data structure named 'Observation Object Model (Observ-OM)' [12] to capture any phenotypic observation and the provenance of the protocols used to produce them, a structure particularly well suited to use the ontologies above in daily research practice.



"Inoculate 4 ml of liquid YPAD or 10 ml of SC and incubate with shaking overnight at 30 °C"

experiment action: move
object: YPD media bottle
start location: in store
end location: in laminar flow hood
experiment action: move
object: 500ml conical flask
start location: in store
end location: in laminar flow hood
experiment action: move
object: sealed yeast colonies plate
start location: in cold room
end location : in laminar flow hood
experiment action: add
component 1: YPD medium
volume: 50ml
start container: YPD media bottle
end container: 500ml conical flask
equipment: pipette
experiment action: rename
old name: 500ml conical flask
new name: YPD conical flask

FIGURE 1

Description of how natural language is converted to an EXACT protocol.

Observ-OM therefore includes extensive use of ontological references for unambiguous protocol, protocol-parameter and observed value definitions using the OntoCAT framework [13] to automatically retrieve ontology terms from resources like BioPortal and OLS, all developed in collaboration with EU-GEN2PHEN (http://www.observ-om.org). Many applications have built on this core with more underway such as:

- AnimalDB for the management and observation of laboratory animals (http://www.animaldb.org);
- A Next Generation Sequencing LIMS for resequencing laboratories and the analysis protocols surrounding this data with an application to the Genome of the Netherlands (770 Dutch whole genomes) project (http://www.nlgenome.nl);
- An International Dystrophic Epidermolysis Bullosa patient registry, where the model is used to report protocols and observations for phenotypic, clinical and genetic features (http://www.deb-central.org) [14];
- Interestingly, this also includes a computational framework to capture and run computational protocols using exactly the same model as wet lab protocols (should this run on with the previous bullet point or be a separate one?);
- The 'xQTL workbench' for the observation of genetic quantitative trait loci in genome wide linkage and association studies (GWL, GWAS) in human and model organism populations (http://www.xqtl.org).

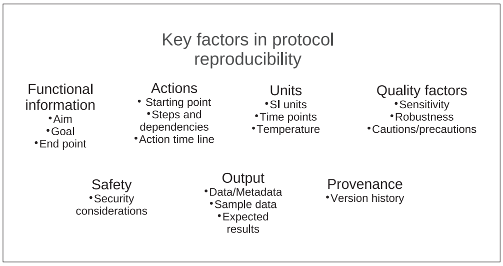


FIGURE 2

Information deemed necessary to create a fully reproducible protocol during the first workshop session. During the workshop the information was divided into distinct classes as shown in the figure.

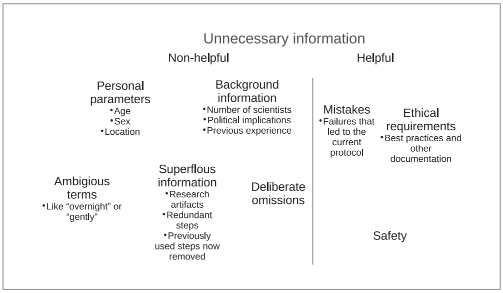


FIGURE 3

Information deemed unnecessary for creation of a fully reproducible protocol. During the workshop the information was divided into distinct classes as shown in the figure.

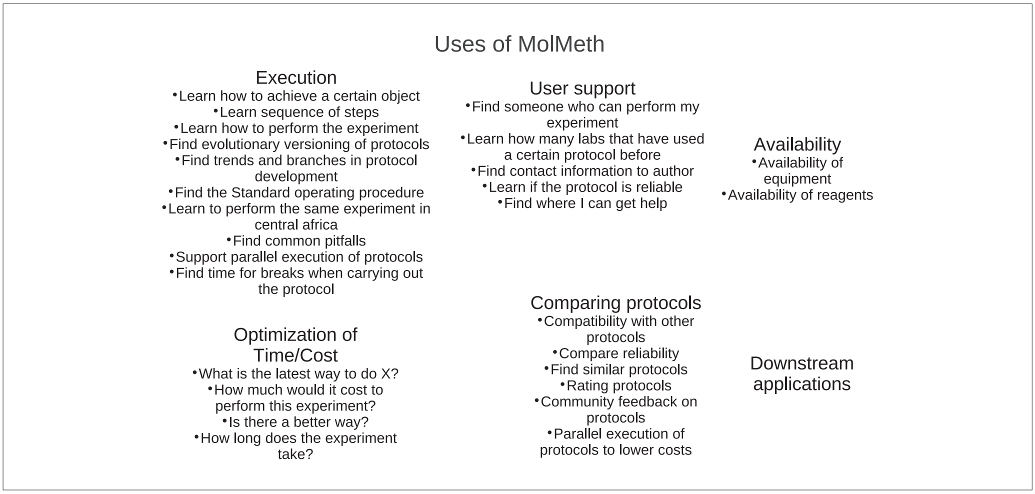

Uses of MolMeth
Execution
•Learn how to achieve a certain object
•Learn sequence of steps
•Learn how to perform the experiment
•Find evolutionary versioning of protocols
•Find trends and branches in protocol development
•Find the Standard operating procedure
•Learn to perform the same experiment in central africa
•Find common pitfalls
•Support parallel execution of protocols
•Find time for breaks when carrying out the protocol
User support
•Find someone who can perform my experiment
•Learn how many labs that have used a certain protocol before
•Find contact information to author
•Learn if the protocol is reliable
•Find where I can get help
Availability
•Availability of equipment
•Availability of reagents
Optimization of Time/Cost
•What is the latest way to do X?
•How much would it cost to perform this experiment?
•Is there a better way?
•How long does the experiment take?
Comparing protocols
•Compatibility with other protocols
•Compare reliability
•Find similar protocols
•Rating protocols
•Community feedback on protocols
•Parallel execution of protocols to lower costs
Downstream applications

FIGURE 4

User expectations of the MolMeth database based on input from the Workshop.

Underlying all these applications is the MOLGENIS automatic software generator that makes it possible to rapidly generate new databases with fully functional web user interfaces from an XML model [15]. In practice this means that local groups can easily customise the MOLGENIS + Observ-OM application suite taking the best components from the existing applications such as those listed above.

A great future enhancement would be to enable exchange of protocol metadata between each application and MolMeth whilst converging Observ-OM with MolMeth common data schema. Moreover, this model driven approach can be perfectly adapted to the needs of the MolMeth database to create extensions where natural language is combined with generic methods to convert protocols between natural language and formal languages such as EXACT.

Discussion

The workshop delivered input for the further development of MolMeth. The discussion session was divided into three sections to determine the scope of the supporting ontologies and expectations of the MolMeth database itself. The first section consisted of a brainstorming session where researchers paired up to identify information necessary to make a laboratory protocol fully reproducible. The suggested key points were then clustered and eight key areas were identified as necessary to create a fully reproducible protocol (Fig. 2).

The second session served to identify information commonly available in laboratory protocols but not essential to render the protocol reproducible. Two major classes were quickly identified to divide information into 'beneficial information unrelated to reproducibility' and 'non-beneficial or harmful information'. These classes were then further divided to create a clustering similar to that arising from the first session (Fig. 3).

The final session was conducted to secure input from ontologists and wet lab experts regarding their expectations for the MolMeth database and how they would like to use it (Fig. 4). To keep in line with the principles of Agile development these ideas will be implemented iteratively when the first fully functional version of the database is published.

Conclusions

The workshop provided a set of short term and long-term goals for the MolMeth database as well as some highly valuable advice, the most immediate being the decision to use EXACT as a starting point for the ontology.

In the Agile development of software and ontologies, emphasis is placed on quickly creating a fully functional core platform to allow early user input into the development. The platform is then gradually improved as new functions and modules are added to the core platform. MolMeth is currently making its first steps through this development process that can be visualised in Fig. 5. In this first pre-launch iteration of the MolMeth database the following functionalities will be implemented:

- A user-friendly environment for the publication of protocols.
- A user-friendly environment for finding, reading and downloading protocols.
- Features to find and contact protocol authors or other users with experience of the protocol.
- A back end ontology enabling users to compare and switch sections of protocols *in silico*, to develop protocols suitable for their own needs.
- A versioning system allowing users to access referenced protocols, earlier protocols and later developments of the same protocol.

Further development will then be carried out based on future workshops, contact with end users and unsatisfied requests summarised in Fig. 4.

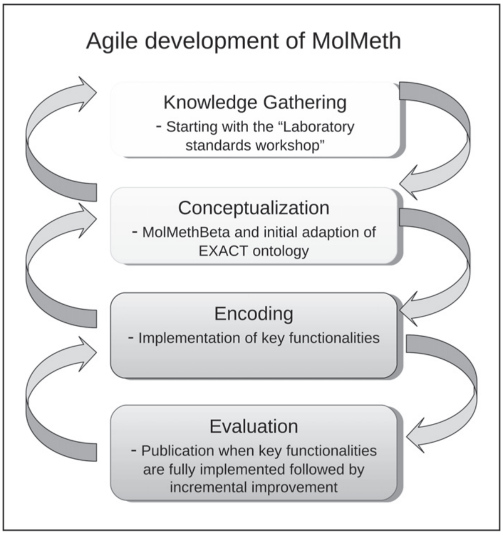

Agile development of MolMeth
Knowledge Gathering
- Starting with the "Laboratory standards workshop"
Conceptualization
- MolMethBeta and initial adaption of EXACT ontology
Encoding
- Implementation of key functionalities
Evaluation
- Publication when key functionalities are fully implemented followed by incremental improvement

FIGURE 5

Planned development process of MolMeth (image based on presentation made by Dr Robert Stevens).

References

1 Soldatova, L.N. *et al.* (2006) An ontology for a robot scientist. *Bioinformatics* 22, e464–e471
2 King, R.D. *et al.* (2004) Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature* 427, 247–252
3 Soldatova, L.N. *et al.* (2008) The EXACT description of biomedical protocols. *Bioinformatics* 24, i295–i303
4 Malone J., *et al.*, *Software Ontology Project* (http://softwareontology.wordpress.com/)
5 *OBI Ontology: The Ontology for Biomedical Investigations* (http://obi-ontology.org/page/Main_Page)
6 Lambrix, P. *et al.* (2009) RepOSE: an environment for repairing missing ontological structure. In *The Semantic Web*, (Vol. 5926) (Gómez-Pérez, A., Yu, Y., Ding, Y., eds) pp. 365–366, Springer Berlin Heidelberg
7 Pettifer, S. *et al.* (2010) The EMBRACE web service collection. *Nucleic Acids Res.* 38, W683–W688
8 EDAM Ontology (http://edamontology.org)
9 Bhagat, J. *et al.* (2010) BioCatalogue: a universal catalog of web services for the life sciences. *Nucleic Acids Res.* 38, W689–W694
10 Taylor, C.F. *et al.* (2008) Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat. Biotechnol.* 26, 889–896
11 Swertz, M.A. *et al.* (2010) The MOLGENIS toolkit: rapid prototyping of biosoftware at the push of a button. *BMC Bioinform.* 11, S12
12 Adamusiak, T. *et al.* (2012) Universal syntax solutions for the integration, search, and exchange of phenotype and genotype information. *Human Mutat.* 33 (5), 867–873 http://dx.doi.org/10.1002/humu.22070
13 Adamusiak, T. *et al.* (2011) OntoCAT – simple ontology search and integration in Java, R and REST/JavaScript. *BMC Bioinform.* 12, 218

14 van den Akker, P.C. *et al.* (2011) The international dystrophic epidermolysis bullosa patient registry: an online database of dystrophic epidermolysis bullosa patients and their COL7A1 mutations. *Hum. Mutat.* 32, 1100–1107

15 Swertz, M.A. and Jansen, R.C. (2007) Beyond standardization: dynamic software infrastructures for systems biology. *Nat. Rev. Genet.* 8, 235–243

Tomas Klingström
Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden

Larissa Soldatova
Department of Computer Science, Aberystwyth University, Wales, UK

Robert Stevens
School of Computer Science, University of Manchester, UKT. Erik Roos

Morris A. Swertz
Genomics Coordination Center, Dept. of Genetics & Groningen Bioinformatics Center, University Medical Center Groningen & University of Groningen, Groningen, The Netherlands

Kristian M. Müller
Institute for Biochemistry and Biology, University of Potsdam, Potsdam, Germany

Matúš Kalaš[1,2]
[1]*Computational Biology Unit, Uni Computing, 5008 Bergen, Norway*
[2]*Department of Informatics, University of Bergen, 5008 Bergen, Norway*

Patrick Lambrix
Department of Computer and Information Science/Swedish e-Science Research Centre, Linköping University, Sweden

Michael J. Taussig
Babraham Bioscience Technologies, Cambridge, UK

Jan-Eric Litton
Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

Ulf Landegren
Department of Immunology, Genetics and Pathology, SciLifeLab Uppsala, Uppsala University, Sweden

Erik Bongcam-Rudloff[1,2]
[1]*Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden*
[2]*Department of Immunology, Genetics and Pathology, SciLifeLab Uppsala, Uppsala University, Sweden*

II

Scholarly publication in the digital age, an investigation into why novel publishing concepts have failed to disrupt the market of scientific journals.

Tomas Klingström, Babette Regierer, Teresa K. Attwood, Erik Bongcam-Rudloff.

Abstract

Digital technologies have changed the ways in which society in general, and science professionals in particular, communicate. Senior researchers send and receive hundreds of e-mails per day; many write or read blogs; many visit social networking sites; many 'tweet'. Similarly, all major journals use the Internet to communicate their latest issues/volumes, landmark articles, *etc*. Despite all this the time-consuming nature of journal publication including tasks like peer-review, separation of articles into scheduled issues remains a cornerstone of scholarly communication in the life sciences. The persistence of this labour-intensive system of communication suggests that factors other than rapid dissemination are crucial to how researchers engage in scholarly communication.

To gain insights into these factors, and the opportunities offered by digital communication, we have developed a framework for evaluating how social networking and other 'Web 2.0 technologies' have been used to increase the efficacy of scholarly communication and applied it on life science research. Of the Web-based services that we surveyed based on the framework, around 15 % have been able to maintain active user-bases; by comparing the unique value propositions of those with active and inactive user bases, it was possible to identify incentives that motivate researchers to use them. In particular, the study showed that web 2.0 technology:

- Have primarily been used to boost the visibility and efficiency of researchers by increasing awareness of their own published work and reducing the time they have to spend finding information of relevance to them;

- have lowered the entry barriers to publishing scientific articles, as search engines and social networks have made article discovery much easier, and journal circulation and impact factors less important for article *visibility*;

- that there is a current trend where innovators in the field develop tools to make scientific outputs "citeable" by attaching digital object identifiers to scientific outputs other than articles published in peer reviewed journals.

Introduction

Scientific publications are an integral part of academic life: they are used to convey research findings to peers, to assert priority, and as quality metrics for academic promotion committees and grant-awarding bodies (1,2).

As the number of journals publishing scientific articles increases(3), the more important it becomes for individual researchers to stand out and produce 'good' articles (*i.e*., articles that are highly cited and/or published in prestigious journals with high impact factors (4–7). In competitive fields, researchers may therefore be tempted to 'game the system (8), by publishing as much as possible, while withholding information that provides them with a competitive edge to support further publications (9).

This article focuses on how social-networking and other platforms can support productive and constructive publishing strategies for individual researchers, and also support the research community as a whole. In this context, many attempts have been made to stimulate social change using information technology (10) , to mimic the success with which social networks, and other such platforms reliant on user-generated content (hereafter referred to as 'Web 2.0'), revolutionised online behaviour in the early 2000s and subsequently altered the dynamics of the news media market (11). The expectation was that a similar revolution might transform the process of scientific publishing (5,10,12). However, so far, the norms of scholarly communication have remained remarkably resilient to change, and peer-reviewed journals remain the gold standard for measuring the 'quality' of academic outputs in the life sciences (1).

Aiming to gain insights into why some platforms have been more successful than others, we studied the user incentives offered by 45 Web- or social-networking tools tailored for life scientists identified in a peer analysis of the collaborative protocol management website Molecular Methods database(13). The initial study was conduct by searching for "social network for scientists" and similar terms to identify websites marketing themselves as such or being mentioned as social networks for scientists in articles. This initial search was then expanded by searching for articles writing about the previously identified websites and mentioning other, not yet identified, websites meeting the criteria described below. By differentiating those that have been successful (*i.e*., with active user communities) from those that have not gained traction, we have been able to identify key features and incentives that motivate researchers to adopt new scientific communication technologies.

Defining a framework of evaluation

The theoretical framework underlying the survey and its conclusions is divided into three separate segments: 1) a definition of virtual communities (or social networks from communication theory (14); 2) a view of information sharing as a gift from the field of anthropology (15), and 3) the Garvey-Griffith model of scholarly communication, which was initially developed for the field of psychology, but is also applicable in other fields of research (16). The Garvey-Griffith model was selected as it provides a comprehensive overview over the high level tasks of an active researcher but predates the development of virtual communities and widespread distribution of data (17), enabling us to use it as a baseline to compare what virtual communities have offered to scholarly communication.

Defining a social network

Candidate websites for inclusion were based on an initial Google search for potential competitors and/or role models for the development of Molecular Methods database, a website for the publication and crowd-sourcing of laboratory protocols (13). The initial aim of the peer-analysis was not to write a full scientific paper on the topic. Given the lack of similar studies and our experience of developing the Molecular Methods database we do however believe that many entrepreneurs and research innovators may benefit from a study on successful motivators for researchers based on comparisons between successful and not so successful attempts. Websites and tools described as social networks for scientists or as comparable to such a website/tool were considered for inclusion in the study. Final inclusion was based on the following criteria used to define virtual communities (14):

(1) the website must offer users an interface for interacting with each other;
(2) interactions between several communicators must be supported;
(3) there must be a sustained user-base not affiliated with the founders; and
(4) there must exist a virtual common-public-space where a significant portion of interactive computer-mediated group communication occurs. The existence of such a virtual settlement is evidence of the existence of a related virtual community.

Information sharing as a gift

Virtual communities are characterised by the exchange of information. The procedure of giving away material, texts or advice can be described, in this context, as a gift-giving process based on a mixture of altruism and self interest (17). This process of sharing information can be divided into two distinct practices (15); the first is usual among explorers or enthusiasts who share information on almost anything; the second is dominant among professionals who rely on having information constantly at their disposal journalists, freelance artists, and programmers are mentioned in anthropological works (15), but researchers in rapidly developing fields, such as the life sciences, also fit this profile.

Aside from needing social interaction, these professionals require a constant influx of new information, which they achieve by creating networks of contacts. Providing information as a gift within these networks forms an important step in the establishment of a virtual community. Sharing information online comes at very low distribution cost, and usually without control mechanisms to limit access (*e.g.*, via a payment system); instead, information is gifted, with an anticipated and undetermined future reciprocity (18).

In research, information in the form of publications and conference presentations is a major carrier of value (6,8,19); it is therefore of particular interest to analyse which components scientists are willing to “gift away” and what incentives the networks themselves provide to content-producing members.

The Garvey-Griffith model for scholarly communication

In 1972, Garvey and Griffith outlined a model for scientific communication for the field of psychology (16). Figure 1 illustrates the comprehensive picture they devised for how advances in scientific research are communicated in scientific papers, in preprints, seminars and conference presentations. Peer-reviewed articles are indexed, and comprehensive reviews provide researchers with overviews of the state-of-art in various fields of science.

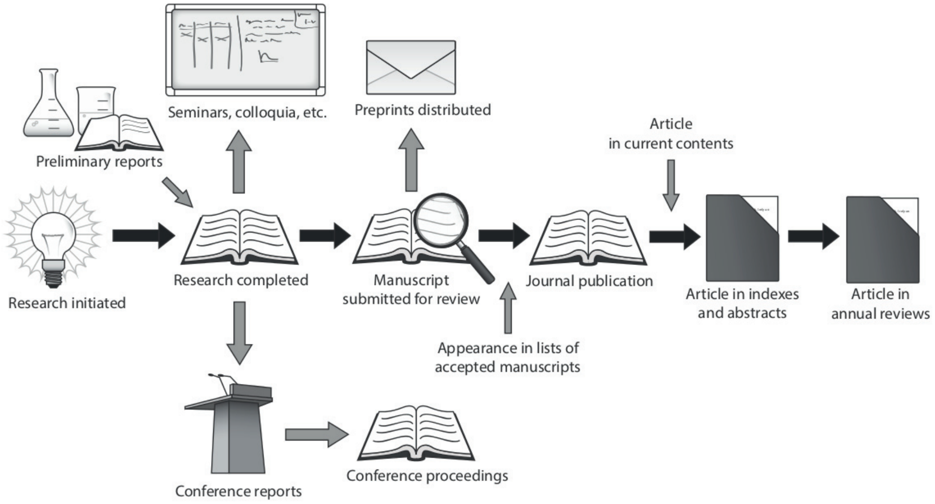


Figure 1. The Garvey-Griffith model of scholarly communication. Research is conducted in discrete steps, where the results of each (notable) step are communicated in conferences, seminars and, most importantly, journal publications.

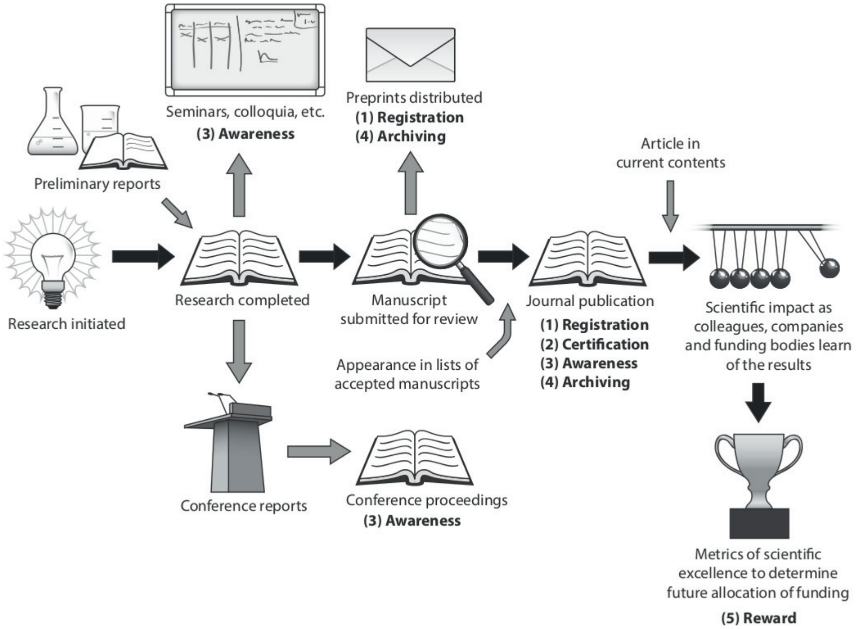
A joint model for scholarly communication in a digital environment

A weakness of the Garvey-Griffith model is that it accurately portrays *how* people share information, but fails to account for *why* they do so. For many researchers, altruism and idealism serve as major motivations to contribute information. But, regardless of such ideals, it is also necessary for researchers to work with a rational dissemination strategy to secure their recognition as originators of knowledge; in return, this supports the procurement of future funding and their promotion in the academic environment.

Funding, promotion and recognition are rewards necessary for researchers to be able to maintain their research activities. Roosendal and Guerts (20) have identified five steps in the academic process that must be fulfilled before a researcher is eligible to receive the rewards needed to continue their work. These have been summarised by Van de Sompel *et al.* (21) as follows:

- registration, which allows claims of precedence for a scholarly finding;
- certification, which establishes the validity of a registered scholarly claim;
- awareness, which allows actors in the scholarly system to remain aware of new claims and findings;
- archiving, which preserves the scholarly record over time; and
- rewarding, the explicit or implicit assignment of funding, promotion or recognition as a leading researcher.

By assigning these five steps to different phases in the Garvey-Griffith model, we produce a framework for assessment of rational incentives for researchers to participate in scholarly communication and/or scientific social networks (Figure 2). When comparing the modes of communication, it becomes evident how peer-reviewed publication is the mode of communication that alone provides registration, certification, awareness building and archiving, which preserves the scholarly record over time. Conferences, seminars and colloquia are all important in helping researchers raise awareness of their research, but they do not provide a persistent and easily accessible form of storage suitable for the registration and archiving of results unless compiled in conference proceedings which is common practice in computer science and handled in a manner similar to journal articles in other research fields (22). Preprints, on the other hand, make it easy to quickly register new discoveries, and make them accessible to a wider community; however, the lack of certification makes it hard to distinguish credible results from unsupported claims.



Seminars, colloquia, etc.
(3) Awareness
Preprints distributed
(1) Registration
(4) Archiving
Preliminary reports
Article in current contents
Research initiated
Research completed
Manuscript submitted for review
Journal publication
(1) Registration
(2) Certification
(3) Awareness
(4) Archiving
Scientific impact as colleagues, companies and funding bodies learn of the results
Appearance in lists of accepted manuscripts
Conference reports
Conference proceedings
(3) Awareness
Metrics of scientific excellence to determine future allocation of funding
(5) Reward

Figure 2. To justify funding, promotion and recognition researchers are expected to contribute to scientific progress. As a researcher it is therefore necessary to register new discoveries (1), have them accepted by peers (2) and present the results to the wider research community (3) in a format accessible to future researchers (4). Reward metrics may be straightforward (*e.g.*, reward for publishing in the 'right' journal, with high acceptance in the community) or implicit (*e.g.*, being awarded a promotion based on a successful track record of conference presentations and publication of significant articles) but all four criteria are necessary to estimate scientific impact and justify the future allocation of resources (5).

Material & Methods

In this work, we compiled a list of social networks and Web tools dedicated to using Web 2.0 technologies (23). By identifying features offered by each, and comparing those with active user-bases with those that have failed to attract significant audiences, we gain insights into the motivations and incentives that encourage researchers to adopt new technologies and share information.

The relative success of each website has been estimated by commonly used Web metrics used in marketing research and current activity. As an external and easily verified metric, we also measured the Google keyword search volume for the name of each service, as many users perform keyword searches rather than writing out full URLs into the address window (24).

Key metrics were selected based on tools commonly used in web analytics and used to evaluate overall website popularity by Alexa toolbar traffic estimates, Google Pagerank, Compete and MozRank. Data were initially collected using MozBar 2.63 and SearchStatus 1.46 on 29 October 2013. A repeat measurement was carried out on 25 February 2016 using WebRank SEO 3.3.7. The data behind the measurement services are not available to external researchers; therefore, we also cross-validated each service against the others to minimise the risk of skewed data (see Table 1). In the 2016 update, metrics from popular social media were also included in the comparison as studies on Altmetrics show a strong correlation between citation rate and activity on Twitter/Facebook (25), thus indicating that they are used by researchers for research related tasks.

Alexa toolbar measures Web traffic based on an international panel of Internet users who have installed the toolbar. Data are normalised to correct for demographic inaccuracies, and the ranking is based on Web traffic in the previous three months, the most popular website is ranked number 1 (26).

Compete measures Web traffic based on a panel consisting of more than two million Internet users in the United States of America (USA). Data are then normalised to correctly reflect the demographics of the USA, and sites are ranked based on their relative popularity, again, the most popular website is ranked number 1 (27).

Page rank estimates site popularity based on the number of inbound links, and popularity of the websites hosting the inbound links. Sites are linked on a logarithmic scale from 1-10, 10 being the highest possible value. Public Page rank data are no longer provided by Google, and hence ranks were not updated between the two surveys (28).

MozRank is similar to page rank, but provides more granular data (1-100), and is more frequently updated. Data are also processed so that links made on pages with a large number of outbound links are less valuable (29). Owing to changed terms of usage MozRank data were unavailable in the 2016 survey.

Google search is a popular search tool via which many users reach websites, rather than directly accessing them by laboriously typing out their URLs (24). Google keywords provide access to the average monthly search volume on keywords, and may, in cases where the site names differ from popular search terms unrelated to the site, provide an estimate of website popularity.

Facebook is the largest social networking website in the world with over 1.7 billion monthly active users (30). Content is commonly shared on Facebook in the form of links and aggregated statistics on how often users share content from a domain therefore provide an estimate of user engagement.

Linkedin is the largest professional networking website in the world with 106 million monthly users (31) sharing or posting content, in a manner similar to Facebook.

StumbleUpon is a private company providing a “discovery engine” that finds and recommends web content to users based on the preferences of similar users; it does not provide exact usage statistics, but was estimated in January 2015 to drive 0.5 % of all website traffic (32).

Google+ is operated by search-engine giant, Google; user activity is hard to measure, as Google+ membership has been bundled with other services provided by Google. In January 2016, it was estimated that links provided by members using Google+ generated 0.04% of the total traffic on the Web (32).

Results

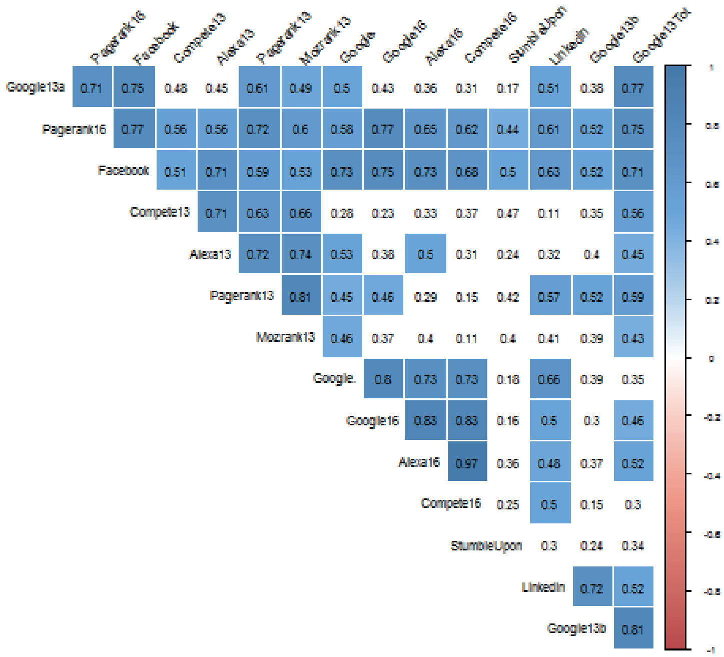


Figure 3, Spearman rank correlation on data from 29 October 2013, showing whether rankings are consistent throughout the sample: ‘1’ indicates a perfect positive correlation (blue); ‘0’, no correlation; and ‘-1’, a negative correlation (red). Squares with a white background indicate that the compared variables did not pass the significance threshold (p < .05). The strongest correlation is dark blue

Comparison by Spearman rank correlation was performed to evaluate whether websites performed similarly, based on the selected Web metrics. 17 of the investigated websites were excluded, as the social networks had either been removed from the Web or used a business model where the network was inaccessible to outsiders. Spearman rank correlation (Figure 3) and scatter plots (see Supplementary Material 1) indicate that Google search frequency, the Alexa toolbar and Facebook engagement all generate a high correlation, despite relying on different measurement techniques to estimate web site popularity. Correlations also remain statistically significant between years ($p < 0.5$), supporting the initial conclusions based on the manual evaluation, indicating that websites with activity in 2013 remained the most active in 2016 (Supplementary Material 2, Table 3).

Manual evaluation of websites was conducted in 2013 and 2016 with user accounts registered on the sites for the purpose of evaluation, covering a four-week periods and manually counting the number of submissions made to the website by external users. At the end of the period site popularity was estimated using the above-mentioned tools. A community was deemed active if users not affiliated by the websites submitted content during the four-week periods in October 2013 and February 2016, ending at the evaluation date listed above. Manual evaluation was consistent with the quantitative data generated using the website analytics tools for large websites.

Website usage statistics

In total, 42 platforms were evaluated in the study (Table 1). Out of the 42 platforms only six websites maintained active user communities that could be defined as "active" in the sense that members got a response from other members not affiliated with the site responding. Out of the six successful websites three are focused on helping researchers to access articles of interest and advertise their own articles (ResearchGate, Mendeley and Academia.edu) and one obtain members by requiring students participating in the International Genetically Engineered Machine (iGEM) competition to contribute protocols to the Wikipedia like web page OpenWetWare as a part of the competition. The two remaining websites, Biostars and myExperiment are targeting bioinformaticians in line with traditions from programming and computer science rather than "wet work" in Life Science – see Table 1 for ranking, and Supplementary Material 2, Table 3 for raw data including all the metrics used to measure web site popularity.

Based on both the tool-based and manual evaluation, three websites clearly stand out from the rest, as they dominate the rankings (see Table 1). By evaluating the value proposition(33)offered to prospective members, it is possible to assess what each of the host companies considers to be the relevant key customer benefits.

	2016 Google search rank	2016 Alexa rank	2016 Facebook rank	2016 Google+ rank	2016 Linkedin rank	2013 Google search rank	2013 Alexa rank	2013 MozRank	1 year trend
Academia.edu	3	2	1	3	1	1	1	4	150%
Mendeley	2	4	2	1	3	2	4	2	0%
Scitable	NA	3	3	10	5	6	NA	NA	125%
Researchgate	1	1	4	2	2	21	18	27	-70%
MyScienceWork	10	7	5	5	25	9,5	5	18,5	100%
Labguru	*5*	*11*	*6*	*7*	*4*	*4*	*NA*	*18,5*	*200%*
Labroots.com	NA	9	7	8	6	11,5	7	15	-60%
OpenWetWare	8	8	8	12	19,5	8	12	6	30%
Researchblogging	22	18	9	16	8	3	3	3	400%
BiomedExperts	12	25	10	9	10	7	9	8	-30%
Sciweavers	9	10	11	4	13	15	6	12	0%
Protocol-online	7	6	12	17	14	22	10	11	-45%
BenchFly	19	15	13	15	12	18	NA	21	-25%
Science stage	21	12	14	13	25	23	8	5	-20%
Malariaworld	*20*	21	16	22,5	17	9,5	21	20	180%
Scientist Solutions	15	13	17	11	11	29	14	23	-50%
Biostars	6	5	18	14	25	13	11	16,5	200%
MyNetResearch	24	27	19	26	21	NA	16	23	NA
UniPhyHealth	*NA*	*26*	*20*	*22,5*	*18*	*NA*	*NA*	*NA*	*NA*
Epernicus	*11*	*17*	*22*	*21*	*15,5*	*11,5*	*NA*	*16,5*	*-10%*
myExperiment	*14*	22	22	29,5	15,5	16	13	10	-40%
Labspaces	27	16	23	19,5	19,5	24	15	13	-40%
Labslink	NA	24	24,5	26	25	30	NA	28	0%
Protocolsonline	*26*	20	24,5	26	25	27	22	30	50%
Sci mate	25	29	26	26	25	NA	NA	29	NA
Laboratree	*23*	*NA*	*27*	*26*	*25*	*25*	*NA*	*NA*	*-75%*
The Science advisory	NA	14	28	6	9	NA	20	14	0%

board										
2collab	NA	NA	NA	NA	NA	NA	NA	NA	NA	
BioCrowd	NA	NA	NA	NA	NA	NA	NA	25	NA	
GoingON	13	NA	NA	NA	NA	14	17	7		-60%
Jeffs bench	NA	NA	NA	NA	NA	NA	NA	NA	NA	
Kappaprime	NA	NA	NA	NA	NA	NA	NA	NA	NA	
Lablife	4	NA	NA	NA	NA	5	NA	NA		-20%
Labmeeting	18	NA	NA	NA	NA	17	NA	NA		-38%
My Sdscience	NA	NA	NA	NA	NA	28	23	23		0%
Nature network	NA	NA	NA	NA	NA	26	2	1		-50%
PaperCritic	NA	*28*	NA	NA	NA	NA	24	26		-50%
Science2point0	NA	NA	NA	NA	NA	NA	NA	NA	NA	
Scilink	*17*	NA	NA	NA	NA	19	NA	NA		-15%
Vivoweb	16	*19*	*15*	*18*	*7*	*20*	*19*	*9*		*-20%*

Table 1. Ranking of scientific social networks sorted by average rank. The five websites with the lowest (best) rank in each category have been marked green to show the most highly correlated category scores (a table using all available metrics is available in Supplementary Materials 1, Table 3). Rows with bold text indicate websites with active communities, grey backgrounds indicate social networks that were discontinued by February 2016. Rows with italic text indicate websites dedicated to private social networking within organisations. NA denote sites that lack a score from the measurement method used and are removed pairwise when calculating Spearmans's rho rank.

The value propositions offered by the top-ranked websites (as described on each website landing page) are:

Research Gate (34)

- Research visibility
- Connect and collaborate
- Stats and metrics

Academia.edu (35)

- Share your papers
- See analytics on your profile and papers
- Follow other people in your field

Mendeley (36)

- Organise your research
- Collaborate with others online
- Discover the latest research

Academia.edu and ResearchGate are almost identical in their customer offer, and similar to Facebook or Linkedin in their design. Mendeley is, in its basic form, a bibliography management tool but with social-networking features that allow users to discover and track new research, similar to Academia.edu and ResearchGate. Based on the Van de Sompel categorisation (21), the interpretation might be that two of the most successful platforms (Research Gate and Academia.edu) attract users by promising to increase awareness of their work, while Mendeley provides a variant of peer-to-peer certification, promoting awareness as friends and associates spread articles of interest to their collaborators.

Apart from these three highly successful platforms, three others have managed to maintain active communities in niche areas. These are:

Open Wetware, a Wiki-style community dedicated to managing laboratory information and sharing protocols supported by, among others, the BioBricks Foundation (37) – BioBricks offers a registry of standardised (safe, ethical, cost effective, free) biological components, which teams of students use in synthetic biology's iGem competition to build viable biological systems – teams that actively share information via Open Wetware are rewarded;

myExperiment, a website to find, use and share scientific workflows (38);

Biostars, a question-and-answer website where users can post questions, and casual visitors and subscribed members can answer. Questions and answers are voted up or down, which helps users to find the best solutions. This sort of website is popular with programmers; Biostars brings this concept to bioinformatics audiences.

Apart from these websites functioning as autonomous virtual communities it should also be noted that two websites have active communities but are largely reliant on users affiliated with the founders to provide content, which violates criterion (3) in the definition of a social network ("there must be a sustained user-base not affiliated with the founders").

Protocol-online, a website that publishes experimental protocols, and hosts a forum where members can discuss laboratory techniques and research-related questions. The protocol section sees low activity, but the question-and-answer forum is active, a high proportion of answers being provided by members affiliated with the website;

The Science Advisory Board, a community sponsored by Bioinformatics LLC, where users are rewarded for answering questions and surveys. The community consists of groups of users who can post articles and questions, a significant part of the communication coming from members affiliated with the website.

Discussion

The major limitation of the study is the difficulty of creating an exhaustive list of social networking websites. Few of the surveyed websites have an accompanying marker paper in indexed journals and Google searches are not repeatable over time making it hard, if not impossible to evaluate the proportion of suitable websites meeting the criteria being included. The nature of Google searches with its advanced metrics for prioritising relevant web sites does however mean that websites left out of the comparison are unlikely to have been widely popular at the time of the surveys. The second survey to evaluate the development between 2013 and 2016 further strengthens this conclusion as the relative popularity of websites remained relatively stable throughout the time covered by the two surveys. Indicating that widely visited and well known websites remained popular while less popular websites were either shut down or remained relatively unknown.

The majority of scientists are, in one way or another, operating online, surfing, searching for information, commenting, rating, and, last but not least, sharing their scientific results with others as reflected by the correlation between citation rates and high exposure in social media(25,39). Nevertheless, only three out of 40 social networks developed for open scientific communication online (see Table 1 and Supplementary Materials 2, Table 1 for full list and features) have achieved widespread influence in several research fields of life science, with active user-bases from a variety of scientific fields. This is, however, consistent with the typical pattern of what economists label as 'winner takes it all markets' (40). In such markets, it is necessary for companies to make significant investments in marketing, as well as infrastructure, in order to gain profits of scale that make it unattractive for competitors to fight for a market share (41). From a user perspective, this might be reasonable, as the work-cost associated with sharing information online is the same, regardless of the size of the audience, but the likelihood of reciprocation is far greater in a community with a million users than in a community with small numbers of users. It is therefore of little surprise that all three major networks have been highly successful in obtaining financing (Mendeley 2.13 M USD (42), ResearchGate undeclared (43) and Academia.edu 2.2 M USD in early funding (44)) for expanding the networks at an early stage in their development.

All three platforms provide a combination of broadcasting capabilities and filtering, as scientists can both recommend their own articles and promote awareness of articles written by others. This provides an efficient solution to one of the major paradoxes of modern science; for researchers, it is important to publish as much as possible in high-impact journals (5,6,8), in order to advance their career and to acquire research funding; at the same time, scientists struggle to cope with the information deluge published in their fields, and strive to minimise the time reading less important articles (20).

None of the smaller social networks provide these broadcasting and filtering services as efficiently as the larger ones. But, by finding attractive niches, three of the small networks have been able to create viable networks, despite the dominance of the larger ones. Open Wetware focus on publishing protocols and workflows that would be hard to publish elsewhere – several of the networks that failed to gain traction have tried to offer similar services, but Open Wetware is unique as it is an integrated part of the BioBricks infrastructure, supported by eager participants in the iGem competition, who are rewarded for publishing material in Open Wetware (45) but also remains relevant for other researchers accessing the website.

The two remaining networks (Biostars and myExperiment) are dedicated to bioinformatics, an informatics-related sub-field of the life sciences in which computers are used to manage, analyse and help interpret biological data (46). The bioinformatics community is strongly influenced by traditions in programming where virtual communities are as old as the Internet itself (47), meaning that social networking sites for bioinformaticians may be considered an extension of traditions in programming rather than a case of technology adoption among life science researchers.

In general the websites included in the survey had a focus on facilitating interactions between researchers, either by directly promoting the creation of virtual communities or by encouraging researchers to share material and then use the networking features to discuss the materials presented. Indicating that there was a widespread belief in the importance of speeding up communications among entrepreneurs in the field starting companies at a time when social networks such as Facebook and Twitter rose to prominence among a mainstream audience.

In comparison more recent ventures seem to follow a different path with entrepreneurs investing efforts into systems that allow researchers to register their findings and make them more easily citeable to a wider audience. This development is closely related to the development of digital object identifiers (48) or “DOI-numbers” which are nowadays ubiquitous for scientific journals and reference management systems (including Mendeley) use them to automatically retrieve metadata for citations and generate bibliographies. There is however no technical limitations restricting DOI-numbers to traditional journals and entrepreneurs in scholarly communication now provide websites making it possible to register scientific outputs other than journals using these identifiers and associating them with relevant metadata. In 2016 Crossref announced that they would provide DOI numbers to preprints (49) and other forms of research outputs can now also easily be published in an easily referenced manner using DOI numbers using websites such as Protocols.io (50) for laboratory protocols, datasets at Zenodo (www.zenodo.org) and any kind of “research output” at Figshare (www.figshare.com).

It is yet too early to evaluate which ones of these may become significant contributors to the system of scholarly communication, but it is clear that while the websites covered by our survey aimed to change communication, this newer generation of websites is more oriented towards helping researchers integrate new kinds of research outputs into a very traditional system of academic recognition.

Implications for future development and research

Creators of social networks for scientists have often found that participation rates are lower than anticipated. Leading to most web sites covered by the study either being shut down or languishing with empty or inactive communities (51,52). This survey is to our knowledge the first of its kind and provides an insight into how technology has been adopted in life science which as a field appears more conservative in its means of communication than for example physics and data science.

From the survey it is evident that the most successful social networks have been the ones helping researchers to achieve the four criteria necessary for qualifying scientists for academic recognition (registration, certification, awareness and archiving). In environments where traditional communication is restricted to print media or verbal presentations, scientists need to publish in widely distributed and well-known journals in order to be recognised and rewarded. But access to dedicated social networks and communication platforms has increased the accessibility of scientific results and knowledge (53). However, in order to become eligible for promotions or to receive funding, this notoriety must be coupled with publication in scientific journals(1,2).

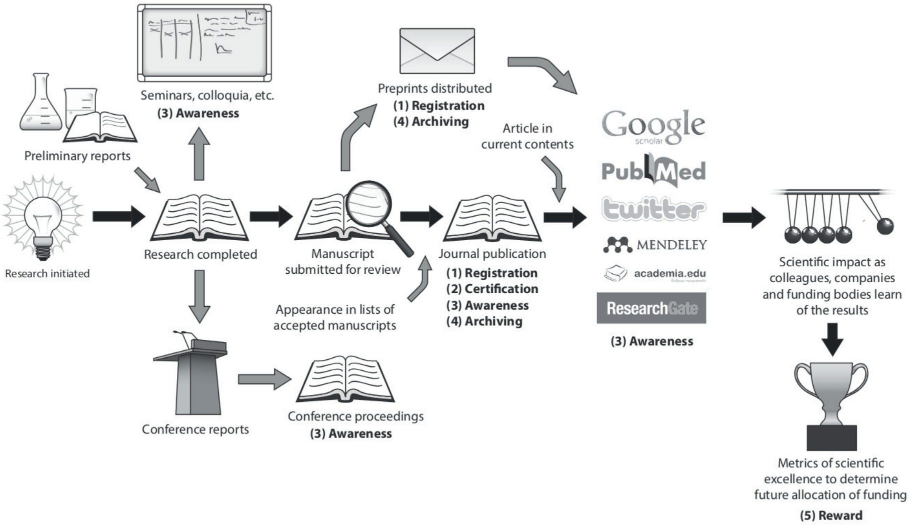


Figure 4. Current status of scientific publishing in life science. Publication in peer reviewed journals remains the only method to achieve certification, awareness and archiving, but search engines and scientific social networks provide new channels for researchers to raise awareness of new articles. This makes scientists less reliant on the popularity of the journal to reach potential readers.

Journals have so far maintained a position as the mainstay of scientific communication, as they provide a 'one stop service' in scientific communication. The value of this contribution is evident as, relative to other sectors of the publishing business (*e.g.*, novels and weekly news magazines), profit margins and returns on equity are 5-15 % higher, (54). The profitability of this market has so far not been significantly threatened, as open-access journals (*e.g.*, PLoS One), and currently successful scientific social networks, have built their success by adhering to the traditional academic recognition process; at the same time, they provide other advantages, such as faster review and support for researchers to raise awareness of their work, thereby sustaining a system in which the value they contribute can be converted into considerable profit margins.

This conservatism is also evident when looking at engagement levels on social networking websites. Nature has launched ambitious projects such as Nature Networks (discontinued) and Scitable (http://www.nature.com/scitable), providing high-quality material to the public, but without achieving major user engagement. But even early adopters to new platforms such as PeerJ (a publisher offering extremely low-cost open-access publishing in return for providing at least one peer-review per year) (55) display a similar unwillingness to step outside the traditional structure of scholarly communications. Apart from its revolutionary approach to review, with life-time users and low costs, PeerJ still functions mainly as a traditional scientific journal, despite offering Web 2.0 features, such as a Q&A section, modelled on the popular Stack Overflow system. In its first 14 months of existence, the Q&A section only received questions from 15 members, and answers from 49 members, despite being highly visible on a website where 1,512 members had registered and 410 had participated as non-anonymous reviewers in the traditional peer-review process (a more time-consuming task than responding to questions in the Q&A section). By 18 May 2016, the number of members had grown to 33,052 (an increase by over 2,000 %), but the number of members who had responded in the open Q&A forum had only increased to 52 (+6 %), suggesting that adapting to new ways of communication is not high on the list of priorities of new members.

Based on the evaluation of platform success and other factors described in this article, we can see that there has been no Web 2.0 revolution, and new technology has only been successful when used to support the current system of scholarly communication. There may yet be a significant upheaval in the publishing business, as new methods for building awareness and certification allow researchers to bypass publication in scientific journals and still achieve all four requirements to be eligible for academic rewards, such as funding or career promotion (Figure 5). However, for such an upheaval to take place, it is necessary that systems are put in place to enable external funding bodies and academic review boards to evaluate these metrics just as they review publication metrics today

For researchers, this means that new technologies, at the moment, are most readily useful in the 'awareness building' stage of scholarly communication. By sharing information about interesting articles with colleagues online, we gain acceptance within communities, we gain access to relevant articles efficiently, and have opportunities for promoting our own articles.

The long-term effects of new technologies are harder to predict and the development of websites making it possible to register scientific output other than articles for citations may yet generate significant shifts in scholarly communication. Journal publication is optimised for a different era of communication, and the rapid pace of research, combined with a limited pool of reviewers, has put the system under serious strain (1,56).

Openness is a core value of science (57) but our results indicate that researchers are unwilling to commit the time and effort to openly sharing data online unless it contributes to the already established norms for scholarly communication. With publication metrics already being considered an important factor by many researchers (58) and many researchers being open to alternative metrics as well to expand their resumes (59) it may be increasingly attractive for researchers to register and archive their researcher in unconventional ways without going through the established process of pre-publication review.

Evidence of this is still limited to individual cases and even the decision by a single researcher to not commit to journal publication for a pre-print accepted as valid by peers has been deemed newsworthy enough to cover in a Nature editorial (60). Such a development would however be a logical conclusion based on our model integrating the Garvey-Griffith model with the 5 steps to scientific reward by Roosendal and Guerts.

The lack of a controlled environment makes it hard to draw definite conclusions. But our model provides a logical explanation for the current lack of radical transformation of scholarly communication despite how electronic communication has disrupted and radically changed other traditionally printed media (11). Furthermore it provides the foundation for a theory that will be tested over the coming years as we see the effect of preprints and other citeable media play out in the research community. Based on our model we suggest that:

- Preprints and other data will be increasingly cited as infrastructure enabling researchers to both find and cite alternative means of publication of data is readily available.
- As a result external or alternative certification bodies conducting peer review (10,61) will become increasingly important.

Without external certification bodies, it will be virtually impossible for academic institutions to properly reward scientific material published outside traditional journals. Citeability in itself does however create renewed opportunities for entrepreneurs to challenge traditional journal publications and if external certification outside of the established journal system become widely recognised by academic institutions, then a revolutionary change in scholarly communication could disrupt scientific journals and their profitability just as news media is currently changing.

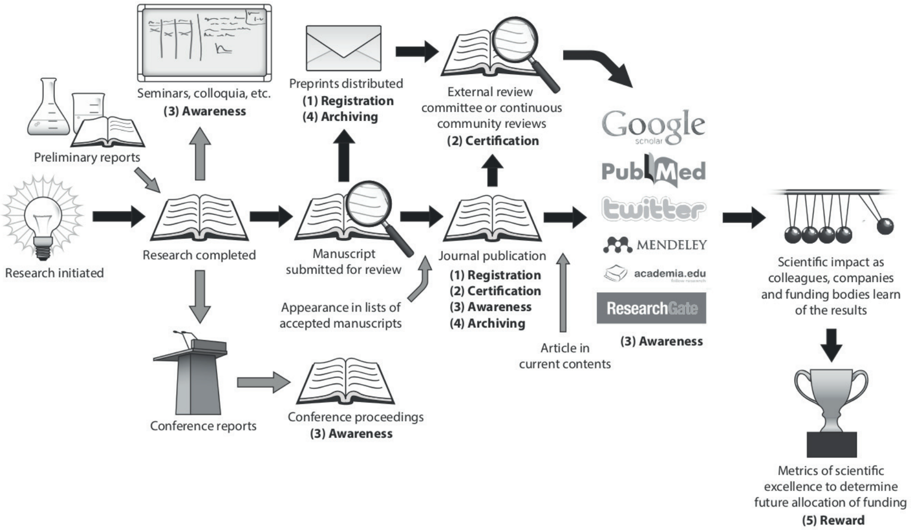


Figure 5. A future scenario where publication in scientific journals is no longer the single gold standard as articles can obtain the steps of registration (1), certification (2), awareness (3) and archiving (4) without

publication in traditional scientific journals. In such a scenario peer review moves into the hands of formal accreditation bodies, external committees like Peerage of science, or scientific social network platforms, such as PubMed Commons, Paper Critic (Mendeley) or ResearchGate, providing relatively reliable certification regardless of the status or reputation of the archive hosting the article.

References

1. Harley D, Acord SK. Peer Review in Academic Promotion and Publishing: Its Meaning, Locus, and Future. 2011; Available from: http://www.escholarship.org/uc/item/1xv148c8

2. Leeuwen TNV, Moed HF. Funding decisions, peer review, and scientific excellence in physical sciences, chemistry, and geosciences. Res Eval. 2012;21(3):189–198.

3. Larsen PO, von Ins M. The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. Scientometrics. 2010 Sep;84(3):575–603.

4. Harley D. The Influence of Academic Values on Scholarly Publication and Communication Practices. J Electron Publ [Internet]. 2007 May 30 [cited 2014 Dec 12];10(2). Available from: http://hdl.handle.net/2027/spo.3336451.0010.204

5. Brembs B, Button K, Munafò M. Deep impact: unintended consequences of journal rank. Front Hum Neurosci [Internet]. 2013 [cited 2014 Dec 12];7. Available from: http://journal.frontiersin.org/Journal/10.3389/fnhum.2013.00291/full

6. González-Alcaide G, Valderrama-Zurián JC, Aleixandre-Benavent R. The Impact Factor in non-English-speaking countries. Scientometrics. 2012 Mar 23;92(2):297–311.

7. Fuyuno I, Cyranoski D. Cash for papers: putting a premium on publication. Nature. 2006 Jun 15;441(7095):792–792.

8. Brischoux F, Cook TR. Juniors Seek an End to the Impact Factor Race. BioScience. 2009 Sep;59(8):638–9.

9. Archambault É, Larivière V. History of the journal impact factor: Contingencies and consequences. Scientometrics. 2009 Jun;79(3):635–49.

10. Hames I. The changing face of peer review. Sci Ed. 2014 Feb 13;1(1):9–12.

11. Mitchell A, Rosenstiel T. State of the news media 2015. Pew Res Cent Journal Media. 2015;

12. Simmons LW, Tomkins JL, Kotiaho JS, Hunt J. Fluctuating paradigm. Proc R Soc B Biol Sci. 1999 Mar 22;266(1419):593–5.

13. Klingström T, Soldatova L, Stevens R, Roos TE, Swertz MA, Müller KM, et al. Workshop on laboratory protocol standards for the molecular methods database. New Biotechnol. 2013 Jan;30(2):109–13.

14. Jones Q. Virtual-Communities, Virtual Settlements &amp; Cyber-Archaeology: A Theoretical Outline. J Comput-Mediat Commun. 2006 Jun 23;3(3):0–0.

15. Bergquist M, Ljungberg J. The power of gifts: organizing social relationships in open source communities. Inf Syst J. 2001;11(4):305–320.

16. Garvey WD, Griffith BC. Communication and information processing within scientific disciplines: Empirical findings for Psychology. Inf Storage Retr. 1972;8(3):123–36.

17. Rheingold H. The virtual community: Finding commection in a computerized world. Addison-Wesley Longman Publishing Co., Inc.; 1993.

18. Kollock P. The Economies of Online Cooperation: Gifts and Public Goods in Cyberspace. In: Kollock P, Smith M, editors. Communities in Cyberspace [Internet]. 11 New Fetter Lane, London EC4P 4EE: Routledge; 1999. p. 220–239. Available from: http://www.sscnet.ucla.edu/soc/faculty/kollock/papers/economies.htm

19. Linton JD, Tierney R, Walsh ST. Publish or Perish: How Are Research and Reputation Related? Ser Rev. 2011 Dec;37(4):244–57.

20. Roosendaal HE, Geurts PATM. Forces and functions in scientific communication: an analysis of their interplay. Vol Nuit Antoine St-Exupéry Ch 19. 1997.

21. Van de Sompel H, Payette S, Erickson J, Lagoze C, Warner S. Rethinking Scholarly Communication: Building the System that Scholars Deserve. -Lib Mag [Internet]. 2004 Sep [cited 2013 Oct 22];10(9). Available from: http://www.dlib.org/dlib/september04/vandesompel/09vandesompel.html

22. Freyne J, Coyle L, Smyth B, Cunningham P. Relative status of journal and conference publications in computer science. Commun ACM. 2010 Nov 1;53(11):124.

23. Thorin SE. Global Changes in Scholarly Communication. In: Ching HS, Poon PWT, McNaught C, editors. eLearning and Digital Publishing [Internet]. Berlin/Heidelberg: Springer-Verlag; [cited 2014 Mar 7]. p. 221–40. Available from: http://link.springer.com/10.1007/1-4020-3651-5_12

24. McKenna G. Experiment Shows Up To 60% Of “Direct” Traffic Is Actually Organic Search [Internet]. Search Engine Land. 2014 [cited 2017 Sep 7]. Available from: http://searchengineland.com/60-direct-traffic-actually-seo-195415

25. Thelwall M, Haustein S, Larivière V, Sugimoto CR. Do Altmetrics Work? Twitter and Ten Other Social Web Services. Bornmann L, editor. PLoS ONE. 2013 May 28;8(5):e64841.

26. Alexa Internet - Our Data [Internet]. [cited 2013 Nov 13]. Available from: http://www.alexa.com/help/traffic-learn-more

27. Our Data | Compete [Internet]. [cited 2013 Nov 13]. Available from: https://www.compete.com/about-compete/our-data/

28. Facts about Google and Competition [Internet]. [cited 2013 Nov 13]. Available from: http://www.google.com/competition/howgooglesearchworks.html

29. MozRank - Learn SEO - Moz [Internet]. [cited 2013 Nov 13]. Available from: http://moz.com/learn/seo/mozrank

30. Facebook Inc. Facebook Reports Second Quarter 2016 Results [Internet]. [cited 2016 Sep 13]. Available from: https://s21.q4cdn.com/399680738/files/doc_financials/2016/Facebook-Reports-Second-Quarter-2016-Results.pdf

31. LinkedIn Corporation. Analyst-Sheet_PostTax_IE_0802.pdf [Internet]. 2016 [cited 2016 Sep 13]. Available from: https://s21.q4cdn.com/738564050/files/doc_financials/quarterly/2016/Q2/Analyst-Sheet_PostTax_IE_0802.pdf

32. Wong D. In Q4, Social Media Drove 31.24% of Overall Traffic to Sites [REPORT]. Shareaholic Reports, Social Media. 2016.

33. Porter ME, Advantage C. Creating and sustaining superior performance. Compet Advant. 1985;167.

34. ResearchGate [Internet]. [cited 2013 Nov 15]. Available from: http://www.researchgate.net/

35. Academia.edu - Share research [Internet]. [cited 2013 Nov 14]. Available from: http://www.academia.edu/

36. Free reference manager and PDF organizer | Mendeley [Internet]. [cited 2013 Nov 14]. Available from: http://www.mendeley.com/

37. Smolke CD. Building outside of the box: iGEM and the BioBricks Foundation. Nat Biotechnol. 2009 Dec;27(12):1099–102.

38. De Roure D, Goble C, Stevens R. The design and realisation of the Virtual Research Environment for social sharing of workflows. Future Gener Comput Syst. 2009 May;25(5):561–7.

39. Hawkins CM, Hunter M, Kolenic GE, Carlos RC. Social Media and Peer-Reviewed Medical Journal Readership: A Randomized Prospective Controlled Trial. J Am Coll Radiol. 2017 May;14(5):596–602.

40. Noe T, Parker G. Winner Take All: Competition, Strategy, and the Structure of Returns in the Internet Economy. J Econ Htmlent Glyphamp Asciiamp Manag Strategy. 2005 Mar;14(1):141–64.

41. Hand JRM. Evidence on the Winner-takes-all Business Model: The Profitability Returns-to-scale of Expenditures on Intangibles Made by U.S. Internet Firms, 1995-2001. SSRN Electron J [Internet]. 2001 [cited 2013 Nov 15]; Available from: http://www.ssrn.com/abstract=292099

42. Mendeley | CrunchBase Profile [Internet]. [cited 2013 Nov 15]. Available from: http://www.crunchbase.com/company/mendeley

43. ResearchGate [Internet]. [cited 2013 Nov 15]. Available from: http://www.researchgate.net/aboutus.AboutUsPressPage.html?page=aboutus.AboutUsPressPage2010September08

44. Academia.edu | CrunchBase Profile [Internet]. [cited 2013 Nov 15]. Available from: http://www.crunchbase.com/company/academia-edu

45. Labtimes: Bench philosophy: OpenWetWare [Internet]. [cited 2013 Dec 6]. Available from: http://www.labtimes.org/labtimes/method/methods/2009_02.lasso

46. Dudley JT, Butte AJ. A Quick Guide for Developing Effective Bioinformatics Programming Skills. Lewitter F, editor. PLoS Comput Biol. 2009 Dec 24;5(12):e1000589.

47. Ouzounis CA, Valencia A. Early bioinformatics: the birth of a discipline--a personal view. Bioinformatics. 2003 Nov 22;19(17):2176–90.

48. Paskin N. Digital Object Identifiers for scientific data. Data Sci J. 2005;4:12–20.

49. Lammey R. Preprints are go at Crossref! [Internet]. Crossef Blog. 2016 [cited 2017 May 19]. Available from: https://www.crossref.org/blog/preprints-are-go-at-crossref/

50. Teytelman L, Stoliartchouk A, Kindler L, Hurwitz BL. Protocols.io: Virtual Communities for Protocol Development and Discussion. PLOS Biol. 2016 Aug 22;14(8):e1002538.

51. Kling R, McKim G, King A. A bit more to it: Scholarly Communication Forums as Socio-Technical Interaction Networks. J Am Soc Inf Sci Technol. 2003 Jan 1;54(1):47–67.

52. Harley D. Scholarly Communication: Cultural Contexts, Evolving Models. Science. 2013 Oct 4;342(6154):80–2.

53. Ware M. Scientific publishing in transition: an overview of current developments [Internet]. Mark Ware Consulting Ltd; 2006 Sep. Available from: http://www.stm-assoc.org/helpful-articles-reports-messa/

54. The market for scientific technical and medical journals - oft396.pdf [Internet]. [cited 2013 Nov 18]. Available from: http://www.oft.gov.uk/shared_oft/reports/media/oft396.pdf

55. PeerJ - How it works [Internet]. [cited 2014 May 19]. Available from: https://peerj.com/about/how-it-works/

56. Akerman R. Technical solutions: Evolving peer review for the internet. Nature [Internet]. 2006 [cited 2013 Oct 17]; Available from: http://www.nature.com/nature/peerreview/debate/nature04997.html

57. Anderson MS, Martinson BC, De Vries R. Normative Dissonance in Science: Results from a National Survey of U.S. Scientists. J Empir Res Hum Res Ethics Int J. 2007 Dec;2(4):3–14.

58. Abbott A, Cyranoski D, Jones N, Maher B, Schiermeier Q, Van Noorden R. Metrics: Do metrics matter? Nature. 2010 Jun 17;465(7300):860–2.

59. Piwowar H, Priem J. The power of altmetrics on a CV: The Power of Altmetrics on a CV. Bull Am Soc Inf Sci Technol. 2013 Apr;39(4):10–3.

60. Singh Chawla D. When a preprint becomes the final paper. Nature [Internet]. 2017 Jan 20 [cited 2017 May 19]; Available from: http://www.nature.com/doifinder/10.1038/nature.2017.21333

61. Hettyey A, Griggio M, Mann M, Raveh S, Schaedelin FC, Thonhauser KE, et al. Peerage of Science: will it work? Trends Ecol Evol. 2012 Apr;27(4):189–90.

III

OXFORD

Briefings in Functional Genomics, 2017, 1–7

doi: 10.1093/bfgp/elx008
Review paper

Legal & ethical compliance when sharing biospecimen

Tomas Klingstrom, Erik Bongcam-Rudloff, and Jane Reichel

Corresponding author: Tomas Klingström, SLU-Global Bioinformatics Centre, Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Sweden. Tel.: +4618-672126; E-mail: tomas.klingstrom@slu.se
Postal address: Inst för Husdjursgenetik, Box 7023, 75007 Uppsala, Sweden
Visiting address: Ulls väg 26, Uppsala

Abstract

When obtaining samples from biobanks, resolving ethical and legal concerns is a time-consuming task where researchers need to balance the needs of privacy, trust and scientific progress. The Biobanking and Biomolecular Resources Research Infrastructure-Large Prospective Cohorts project has resolved numerous such issues through intense communication between involved researchers and experts in its mission to unite large prospective study sets in Europe. To facilitate efficient communication, it is useful for nonexperts to have an at least basic understanding of the regulatory system for managing biological samples. Laws regulating research oversight are based on national law and normally share core principles founded on international charters. In interview studies among donors, chief concerns are privacy, efficient sample utilization and access to information generated from their samples. Despite a lack of clear evidence regarding which concern takes precedence, scientific as well as public discourse has largely focused on privacy concerns and the right of donors to control the usage of their samples.
It is therefore important to proactively deal with ethical and legal issues to avoid complications that delay or prevent samples from being accessed. To help biobank professionals avoid making unnecessary mistakes, we have developed this basic primer covering the relationship between ethics and law, the concept of informed consent and considerations for returning findings to donors.

Key words: ethics; biobank; sample access; genomics; DNA

Introduction

The Biobanking and Biomolecular Resources Research Infrastructure-Large Prospective Cohorts project has provided valuable experience on the issues of sample access through its open calls to provide funding for accessing biobanked samples. For the cohorts who did participate in various projects, attaining ethics approval and sorting the legal issues were time-consuming but not insurmountable tasks because of intense communication between involved researchers and experts.

The risk of biobank samples being used in an inappropriate manner has received increasing attention in scientific discourse. In comparison, the threat of under-utilization of samples or an inability to return the benefits of research to donors has received relatively little attention, despite also being among the chief concerns of interviewed donors [1]. Furthermore, the genomic revolution means that pretty much any sample can be considered to contain potentially identifiable personal data in the form of DNA. Researchers therefore face an intricate extra-legal regulatory system complete with steering documents

Tomas Klingström is a PhD student at the Swedish University of Agricultural Sciences (SLU) and employed at SLU-Global Bioinformatics Centre. His main research deals with data integration and quality verification of the scientific process from sample procurement to data analysis working in the BBMRI-lpc, BBMRI.se and B3Africa projects.
Erik Bongcam-Rudloff is a professor in Bioinformatics at the Swedish University of Agricultural Sciences (SLU) and the head of the SLU-Global Bioinformatics Centre. His main research deals with the development of bioinformatics solutions for the Life Sciences community including nematode, plant, animal and human genomics research; numerous projects on metagenomics; and the building of LIMS and Biobank systems based on open-source elements (www.b3africa.org).
Jane Reichel is a professor in Administrative Law at the Faculty of Law and is tied part-time to the Centre for Research Ethics and Bioethics, both at Uppsala University. She is currently vice dean and chairman of the research committee at the Faculty of Law. Her research focuses on the globalization and Europeanization of administrative law, mainly in the areas of transparency and data protection in cross-border biomedical research.

(ethics guidelines), overseeing bodies (research ethics committees) and formal procedures (informed consent) [2] when attempting to access samples.

Although laws regulating research oversight have been implemented differently in every country, there is a similarity of core principles founded on international charters such as the Helsinki Declaration. International consortia have translated these core principles into policies, procedures, tools and governance that facilitate interoperability between biobanks across national borders in a manner acceptable to national law makers [3–5], thereby enabling the scientific community to operate despite a lack of clarity and international agreements that may provide a stable and enabling environment for international collaboration [6, 7].

As biobanks mature, priorities tend to shift [8], and it is not uncommon that biobanks find themselves prevented from providing samples due inappropriate decisions taken several years earlier. These complications are often the result of requests with unforeseen requirements causing uncertainties if given consents are sufficient and how or if information from new research projects should be returned to the donors. The primer therefore covers how these obligations are governed under international agreements and national law, the practice of establishing this relationship by the concept of informed consent and the difficulties on deciding when and what information should be provided to sample donors.

Hard and soft law, the key to international collaboration

The national legal framework of biobanking is often substantially different even between countries of comparable jurisdictional systems [9]. To accommodate international collaboration, it is therefore necessary to rely on 'soft law' or extra-legal means to bridge the gap between the national legal systems, which operate on a 'one nation, one law, one project' approach [10].

When dealing with such matters, it is therefore important to understand and recognize how research is regulated by a combination of 'hard law' and 'soft law' where the terms can be defined as follows:

Hard law: Binding legal instruments, either in the form of international law (conventions, treaties or agreements) or national law (statutory law). International law is often drafted in a more general form and subsequently implemented in national law. For the individual researcher, it is most often the national statutory law that regulates the legality of actions.

Soft law: Nonbinding instruments such as guidelines and codes of conducts that may lay down suitable and commonly accepted ways to deal with a matter. Soft law in different forms varies in form from openly phrased to rather strictly defined rules, bearing close resemblance to hard law.

Hard law is codified in legal text, which makes it relatively straightforward for a trained expert to access and identify the relevant laws. Soft law is on the other hand more flexible but makes it harder to find and understand the regulatory mechanisms, as it allows governmental and nongovernmental experts to update regulations and standards without requiring active engagement of law-making bodies, and often these experts may be specified in hard law as bodies tasked with providing legally binding regulations and decisions. Funding bodies are becoming an increasingly important source of soft law by enforcing contracts requiring certain guidelines or procedures to be followed by researchers to be eligible for funding.

For European researchers, an important source of this kind of regulation is the European Union (EU) funding programs managed by the European Commission. It requires applicants to state in their proposal that they will conform to specific standards [11] where failure to comply mean that the researcher will not be eligible to receive the funds provided by the grant.

Similar approaches are not only used for international projects but are also a way for national agencies to harmonize activities in nations where legislation is done at a regional or state level. For example, in the United States, the National Research Council stipulates the following for the international transfer of embryonic stem cells:

> If a U.S.-based investigator collaborates with an investigator in another country, the ESCRO committee may determine that the procedures prescribed by the foreign institution afford protections consistent with these guidelines, and the ESCRO committee may approve the substitution of some of or all of the foreign procedures for its own. [12]

These guidelines are defined by one selected group of experts (the National Research Council) who delegate decisions to another group of experts [the Embryonic Stem Cell Research Oversight (ESCRO) Committee], which is charged with deciding if there is a comparable set of checks and balances in the partner country in the form of a, yet to be identified, third group of experts. These guidelines are a good example of how a soft law approach with several layers reduces transparency in return for increased flexibility, as guidelines, review committees and research practitioners make up an ever-changing system of stakeholders. Under such circumstances, collaboration is substantially more likely to be accepted between nations where the respective authorities have had the possibility to become familiar with each other's customs and traditions, and above all, where the legal requirements applicable to the matter have been enacted as a result of international agreements. A lack of trust, harmonization or the local preferences of the committee may therefore significantly affect the outcome of an application for the transfer of data or samples. Decisions by judicial authorities covering one of the partners in a collaboration may also have an immediate impact on international collaboration, as certain procedures are deemed to be in conflict with national law. The EU has, for example, chosen a high standard for data protection, as seen in the recent Safe Harbor-ruling from the Court of Justice of the European Union (C-362/14), where the US level of protection was found not to uphold an adequate protection.

However, most modern national laws are based on an ambition to adhere to a common set of core principles derived from the declaration of human rights and international declarations such as the Declaration of Helsinki [13]. This means that even if there is yet little legal harmonization between countries. There is a strong case for researchers to argue that institutional review boards should take into account decisions from review boards in other countries, in a soft version of a principle of mutual recognition.

Consent as the basis of international collaboration

The signed consent form provides a receipt that verify that the donor has been provided with sufficient information to make an informed consent when donating his or her samples.

Modern regulations regarding informed consent were codified in an international setting by the Helsinki declaration and Nuremberg code [14] as a result of the horrors in World War II and subsequent development. Respect for the autonomy of research subjects and their right to refuse participation in research does however have a much longer history in research [15] even if modern researchers may find certain practices troubling or even barbaric. For example, in the mid-19th century in America, it was considered acceptable for a slave owner to obtain consent for invasive experimental surgery from slaves [16]. While it for a modern person is hard, if not impossible to accept slavery or the concept of 'a consenting slave'. From an academic context, this intuitive protest can be interpreted as an example of how we instinctively respect that a person in a position of dependence cannot make a truly autonomous decision [17]. The concept of donors as autonomous agents is one of the key concepts of modern research, and the question of identifying what information and freedom is necessary before a person can make an autonomous decision is therefore central to all forms of biobanking and genomic research with human participants.

When establishing a new biobank, it is important to rely on forward-looking consent procedures to ensure the future viability of the sample collection. A large number of different forms of consent have been proposed in scientific literature. But in practice, consent forms likely available to a biobank would need to result in a presumed, broad or specific kind of consent (Table 1). In bioethicist literature, concepts such as 'tiered' or 'dynamic' consent are suggested as compromises between specific or broad forms of consent. In practice, these forms of consent can either be broad or specific depending on whether the components of the consent are widely or narrowly specified. It is however not always possible or feasible to obtain information from a known, informed and willing donor. In some cases, a presumed consent is necessary, and several ethicists also argue that a consent can never be truly informed unless strict requirements are met [18–20].

When looking at large biobank infrastructures, a broad consent is favored among the major infrastructures [21–23] even if there still is debate among ethicists on how broad a consent can be while still maintaining the autonomy of the donor [24]. The dominance of broad consent in infrastructures based on soft law is in this context a good example of how soft law solutions allow society to adapt more quickly to new possibilities and risks compared with hard law where important laws may be debated for years before implementation [7].

Specific consent is by its nature reactive, as it is impossible to request specific consent for purposes not yet foreseen. As a response to this issue, proponents of specific consent have made numerous proposals where modern communication technology makes it possible to repeatedly (or dynamically) ask donors for consent [25]. Thus, initial consent only needs to cover foreseeable research, while new projects are made possible by a renewed consent, thereby, in the opinion of its proponents, creating a balance between maximizing the value of samples and the necessary safeguards to ensure that consent is truly informed.

However, research rarely takes place in clearly defined modules, and there is often a continuum where it is hard to define the acceptable threshold for clarity, which requires new consent [26]. In practice, this means that a biobank will require a similar independent ethics review board, regardless of if the biobank operates under a legislation requiring specific, broad or any other form of consent.

Recent research further underlines the support for a broad consent among biobank experts [20], but even a broad consent is limited in how much freedom may be given to researchers to initiate new projects. That an administrative framework remains in place for the sample collection and that the new research does not change the overall aims or governance structure are core conditions and may be regarded as a minimal set of regulations for a broad consent to remain valid [27]. For European needs, Carlo Petrini at the Bioethics unit of the Presidentã s office in Italy has conducted a bibliographical study of the requirements necessary to operate a biobank under a broad consent in Europe [28], suggesting that the following requirements must be met:

- Adequate sample coding procedures are used.
- Adequate procedures for personal data protection are used.
- The importance of the research aim is sufficient to justify conducting the study and is evaluated on a case-by-case basis by an ethics committee.
- The sensitivity of the data is evaluated on a case-by-case basis. Genetic information varies in sensitivity based on its significance, ranging from stringent protection to a lesser degree of protection.
- Generic research results are always released without specifically identification of individual subjects.
- 'Opt-out' consent is allowed for subsequent or secondary studies. Every subject must be guaranteed the possibility of withdrawing consent at any time.
- Participants must have adequate means of involvement, such as encouraging participant consultation or communicating information through the mass media before project initiation. The multiple modes of involvement should be complementary as opposed to mutually exclusive. It is especially important that forms of direct participation also be available, for example, by having population representatives serve on the ethics committees that will decide on the approval of the research before it begins.
- Measures to ensure transparency and supervision must be in place. Adequate supervisory, procedural and technical systems are necessary to guarantee information protection. Further, it is highly advisable to have external and independent supervisory bodies monitoring procedural correctness.

The reporting of planned or incidental findings

Another controversial subject with far-reaching consequences for sample availability is whether researchers should be obliged to return information on findings to the donor [29]. There is currently no overall consensus on when to tell and when not to tell participants of incidental findings [30]. Careful planning of procedures to satisfy local or national expectations is therefore necessary to ensure that donor interests are managed properly. In cases where a study is based on samples, not yet collected, researchers can, and should, plan ahead to ensure that donors are properly informed at the time of consent on reporting procedures. For studies on samples already collected or where clinically relevant findings are incidental in nature, it instead becomes necessary for study manager to base their reporting procedures on their own judgment or guidelines provided by local experts or governing boards suited to the task.

Based on the conflicting opinions described by researchers conducting systematic reviews of the field, it would be foolhardy to claim that practitioners and ethicists are anywhere near a consensus in the field [21–23, 29]. It may however be

Table 1. Forms of consent described in literature

Generalized category	Type of consent	Definition	Authors	Disagreement
No consent given	Presumed	Consent is presumed to have been given by donors to use their samples and information for all research unless they actively choose to opt out	Master *et. al*. and Hofman	
	Passive/tacit/silent consent	Presuming that the persons object if they do not consent	Hofman	
	Hypothetical consent	Consent under the presumption that a person would have consented to the treatment or research were she or he able to consent	Hofman	
A broad or specific consent	Future/deferred consent	Postponing the consent procedure	Hofman	
An extremely broad consent	General/blanket/open consent	Donors can actively consent once for the current study and all future research involving the general use of their samples and information	Master *et. al*., Hofman and Salvaterra *et al*.	Salvaterra refer to this as broad consent
May be either broad or specific depending on how the consent is formulated and the definition used by the reviewers	Broad	Donors can actively consent once for the current study and all future research within a broad field, e.g. cancer, diabetes or heart disease	Master *et. al*., Hofman and Salvaterra *et al*.	Salvaterra refer to this as partially restricted consent
	Delegated trustee	Donors can transfer consent to a trustee who is at arm's distance from the biobank and consents on behalf of donors	Master *et al*.	
	Third-party oversight	Donors can actively consent to a general, broad or other model, but an ethics board must approve the study before the commencement of research using stored samples and information. This approach is emerging as a common component of biobanking governance schemes	Master *et al*.	
	Tiered	Donors can actively consent once for the current study and choose one or more broad fields of research or other options, i.e. whether they would be willing to have their samples used in research that result in commercialization. Other terms: line item or multilayered consent	Master *et al*.	
	Re-consent	Donors are informed and are required to consent to the current study and to each future research study involving the use of their samples and information	Master *et al*.	
	Specific informed consent	Allows the use of biological specimens and related data only in immediate research; forbids any future study that is not foreseen at the time of the original consent	Salvaterra *et al*.	

Note: Terms used in literature are not always univocal and may also be used with different levels of specificity. In the table, the specific definitions described by the authors have been clustered into more general of consent described in accordance with this article. The more specific definitions are listed in the column 'Definition', and the terms used to name them are outlined in 'Type of consent' and 'Disagreement'.

possible to break down disclosure into two dimensions to separate situations where researchers are closer to consensus from areas where there still is severe disagreement (Figure 1).

Given this four-field breakdown and preceding information, ethicists are at least approaching a consensus on the lower left and upper right corners. Which mean that incidental findings with a high level of actionability and clinical validity should, if possible, be reported back to the donor [31] and findings of low validity and actionability should not be reported to the donors (2, upper right corner) ahead of [31] and (3, lower left corner). There is however no consensus on whether it is a moral necessity to actively look for such genes in genetic data, and many researchers also feel uncertain when judging if specific markers are actionable and clinically valid [31]. To support clinicians, the American College of Medical Genetics has taken initiatives to support researchers to reduce these difficulties with lists of

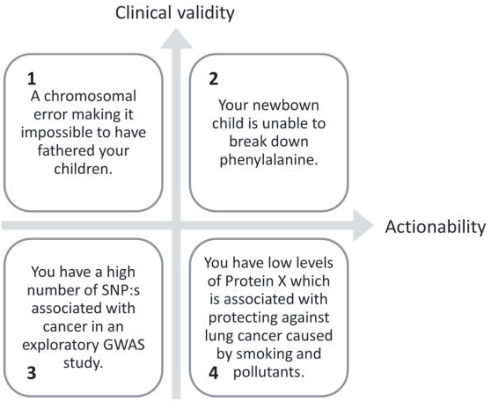

Clinical validity
Actionability
1
A chromosomal error making it impossible to have fathered your children.
2
Your newbown child is unable to break down phenylalanine.
You have a high number of SNP:s associated with cancer in an exploratory GWAS study.
3
You have low levels of Protein X which is associated with protecting against lung cancer caused by smoking and pollutants.
4

Figure 1. A breakdown of potential situations encountered when conducting genetic analysis on collected samples and practical examples of cases clearly belonging to each quarter. Support for returning information to the donor is strong when a finding is both reliable (possessing a high level of clinical validity) and actionable (the donor can act on the given information) as in the example given in Square 2 and, there is little support for providing information that is neither reliable nor actionable (Square 3). Decisions are harder and in greater need of consideration when the reliability of findings is low (Square 4) or when there is little the donor can do about the situation (Square 1).

valid and actionable genetic biomarkers [32], which can be consulted by clinicians to determine if incidental findings should be reported. The procedures for how and if findings are to be reported to the donor should be outlined to the donor at least by the time of consent, thereby helping to set donor expectations and define their future relationship with their donated samples.

This means that the researchers, when developing the consent form, must take care to ensure the long-term viability of the biobank and balance their obligations to donors with the scientific needs of the project. A high level of reciprocity cannot, for example, be offered in a biobank where a large portion of the research is expected to be conducted by external researchers limited to anonymized data to maintain privacy. It is therefore necessary that researchers make important decisions such as coding [33] versus anonymization before contacting potential donors for consent. Failure to do so may otherwise result in major issues in the future, as national laws on privacy or obligations outlined in the consent form may prevent the efficient usage of biospecimen.

Concluding remarks

International collaboration relies on soft law connecting national legal systems, which creates an environment that is inconsistent, unfair and often lacking in transparency. But replacing the soft law with hard law may be even worse, as a codification of overly restrictive standards into law may stifle or outright halt scientific progress in regions within the jurisdiction of such laws [7]. Furthermore, it is unlikely that hard law solutions would be able to possess the necessary flexibility to keep up the pace with the rapid advancement of research and genomics.

As a researcher, it is easy to become frustrated and avoid engaging in such a complex, and ever-changing field of work. But despite calls for harmonization, it is unlikely that issues will be solved in the immediate future. There are significantly different legal traditions [34–37] as well as variation in public perception [38, 39] of research. Taken together, this makes it a perhaps insurmountable task to reach harmonization of national laws regarding biological samples and data protection. The legal obligations of biobank professionals concerning consent and reciprocity are therefore likely to change over time and remain areas associated with a high risk of interfering with the individual goals and aims of researchers.

In this context, adhering to best practices contributes to the long-term value of samples, as new implementations of soft law instruments and codified law are likely to take established best practices in consideration. Guidance and templates provided by international organizations such as International Society for Biological and Environmental Repositories (ISBER, www.isber.org), Global Alliance for Genomics and Health (http://genomicsandhealth.org), the Asian Network of Research Resource Centers (http://anrrc.org), the Biobanking and BioMolecular resources Research Infrastructure-European Research Infrastructure Consortium (www.bbmri-eric.eu) and the Human Heredity and Health in Africa (http://h3africa.org), here, form a platform for harmonization as well as generating the opportunities to build the mutual trust necessary to enable the transfer of samples or data. The role and function of these soft law tools must however take into account the constitutional aspect of the bioethical framework involving several human rights. Traditionally, these rights, and especially the limiting of the rights, are usually thought to be best regulated by democratically elected parliaments [40]. These international soft law tools do thus not supersede national authorities and courts, but their status as internationally recognized authorities may provide considerable support in achieving approval from institutional review boards acting under mandate from national laws.

It is therefore in the best interest of researchers to respect and promote core principles codified by international conventions and organizations. Connecting local interpretations on law to an international context also makes it easier to compare decisions and encourage the development of trust that is necessary for collaboration using sensitive genomic data. It is therefore advisable for biobank builders to adopt a system of governance where:

- The ethical standards set forth by the Global Alliance for Genomics and Health are upheld [5].
- Samples are stored and managed in accordance with the internationally recognized ISBER standards for best practice [41].
- Sharing is handled in a manner compliant with the International Charter of principles for sharing bio-specimens [42].

This does not preclude researchers from having to abide by the national law of each state involved in international research collaborations and is far from an exhaustive list of tools to support international sharing of samples. But it may provide an international research project with a common foundation and framework, which make the project more easily acceptable to the national authorities charged with reviewing projects.

The inherent adaptability of soft law also mean that international collaboration through soft law mechanisms may steadily improve, as experience is gained among stakeholders and thus alleviate the need for global governance via codified hard law solutions within the field. If given time to adapt, researchers and associated organizations might instead be able to contribute to a bottom-up harmonization of a soft global bioethical framework.

Key Points

- To accommodate international collaboration, it is necessary to bridge the gap between national legal frameworks. This is usually done by designated experts and organizations who determine if material transfer agreements are able to protect the rights of the donors in accordance with what they could expect when giving their consent for samples to be stored for future usage.
- Collaboration is substantially more likely to be accepted between nations where the respective authorities have had the possibility to become familiar with each other's customs and traditions. Identifying successful precedents by other researchers participating in collaborative projects can therefore greatly reduce the time necessary to access samples.
- Different institutions define terms such as consent, informed consent and broad consent differently. This mean that an 'informed consent' at one institution may not be accepted as truly informed by another. Under such circumstances, researchers are likely to face a situation where the strictest interpretation in terms of data protection or privacy becomes the governing one.
- There is a conflict between reciprocity, anonymity and the right to not know. Research must therefore be planned and conducted in accordance with what the donors could reasonably expect when donating their samples and giving their consent.

Funding

This work was financed by the BBMRI-LPC and the B3Africa projects. BBMRI-LPC is supported by the European Community's 7th Framework Programme (FP7/20072013) grant agreement no. 313010, B3Africa is supported by the European Union s Horizon 2020 research and innovation programme under grant agreement No 654404.

References

1. Hoeyer K. The ethics of research biobanking: a critical review of the literature. *Biotechnol Genet Eng Rev* 2008;**25**:429–52.
2. Johnsson L, Eriksson S, Helgesson G, *et al*. Making researchers moral: why trustworthiness requires more than ethics guidelines and review. *Res Ethics* 2014;**10**:29–46.
3. Global Alliance for Genomics and Health. Global_Alliance_White_Paper_3_June_2013.pdf. genomicsandhealth.org
4. Budimir D, Polašek O, Marušić A, *et al*. Ethical aspects of human biobanks: a systematic review. *Croat Med J* 2011;**52**:262–79.
5. Global Alliance for Genomics and Health. Framework for Responsible Sharing of Genomic and Health-Related Data, Read Online. Ontario, Canada: Global Alliance for Genomics and Health, 2014. www.genomicsandhealth.org.
6. Knoppers BM. Biobanking: international norms. *J Law Med Ethics* 2005;**33**:7–14.
7. Editorial. Data overprotection. *Nature* 2015;**522**:391–2.
8. Simeon-Dubach D, Watson P. Biobanking 3.0: evidence based and customer focused biobanking. *Clin Biochem* 2014;**47**:300–8.
9. Kiehntopf M, Krawczak M. Biobanking and international interoperability: samples. *Hum Genet* 2011;**130**:369–76.
10. Kaye J. From single biobanks to international networks: developing e-governance. *Hum Genet* 2011;**130**:377–82.
11. European Commission. Ethics http://ec.europa.eu/programmes/horizon2020/en/h2020-section/ethics (16 December 2016, date last accessed).
12. Final report of the National Academies' Human Embryonic Stem Cell Research Advisory Committee and 2010 Amendments to the National Academies' Guidelines for Human Embryonic Stem Cell Research, 2010. Washington, DC: The National Academies Press.
13. World Medical Association. World Medical Association declaration of Helsinki: ethical principles for medical research involving human subjects. *JAMA* 2013;**310**:2191.
14. Weindling P. The origins of informed consent: The International Scientific Commission on medical war crimes, and the nuremburg code. *Bull Hist Med* 2001;**75**:37–71.
15. Vollmann J, Winau R. Informed consent in human experimentation before the Nuremberg code. *BMJ* 1996;**313**:1445–7.
16. Wall LL. The medical ethics of Dr J Marion Sims: a fresh look at the historical record. *J Med Ethics* 2006;**32**:346–50.
17. Sjostrand M, Eriksson S, Juth N, *et al*. Paternalism in the name of autonomy. *J Med Philos* 2013;**38**:710–24.
18. Hofmann B. Broadening consent–and diluting ethics? *J Med Ethics* 2009;**35**:125–9.
19. Salvaterra E, Lecchi L, Giovanelli S, *et al*. Banking together. A unified model of informed consent for biobanking. *EMBO Rep* 2008;**9**:307–13.
20. Master Z, Campo-Engelstein L, Caulfield T. Scientists' perspectives on consent in the context of biobanking research. *Eur J Hum Genet* 2015;**23**:569–74.
21. Hansson MG. Ethics and biobanks. *Br J Cancer* 2009;**100**:8–12.
22. Petrini C. 'Broad' consent, exceptions to consent and the question of using biological samples for research purposes different from the initial collection purpose. *Soc Sci Med* 2010;**70**:217–20.
23. Simon CM, L'heureux J, Murray JC, *et al*. Active choice but not too active: public perspectives on biobank consent models. *Genet Med* 2011;**13**:821–31.
24. Master Z, Nelson E, Murdoch B, *et al*. Biobanks, consent and claims of consensus. *Nat Methods* 2012;**9**:885–8.
25. Karlsen JR, Solbakk JH, Holm S. Ethical endgames: broad consent for narrow interests; open consent for closed minds. *Camb Q Healthc Ethics* 2011;**20**:572–83.
26. Shickle D. The consent problem within DNA biobanks. *Stud Hist Philos Biol Biomed Sci* 2006;**37**:503–19.
27. Steinsbekk KS, Kåre Myskja B, Solberg B. Broad consent versus dynamic consent in biobank research: is passive participation an ethical problem? *Eur J Hum Genet* 2013;**21**:897–902.
28. Presidenza del Consiglio dei Ministri. Collection of Biological Samples for Research Purposes: Informed Consent, 2009. Rome: Istituto Superiore di Sanità (ISS).
29. Christenhusz GM, Devriendt K, Dierickx K. To tell or not to tell? A systematic review of ethical reflections on incidental findings arising in genetics contexts. *Eur J Hum Genet* 2013;**21**:248–55.
30. Viberg J, Hansson MG, Langenskiöld S, *et al*. Incidental findings: the time is not yet ripe for a policy for biobanks. *Eur J Hum Genet* 2014;**22**:437–41.
31. Bradbury A, McCormick J, Robson MA. Changing practice: the controversy over obligations to return incidental findings in genomic sequencing. ASCO Annual Meeting. Alexandria (USA): ASCO Daily News, 2014.

32. Green RC, Berg JS, Grody WW, *et al*. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet Med* 2013;**15**:565–74.
33. Hunter LE, Hopfer C, Terry SF, *et al*. Reporting actionable research results: shared secrets can save lives. *Sci Transl Med* 2012;**4**:143cm8.
34. Zika E, Paci D, Schulte in den Bäumen T, *et al*. Biobanks in Europe Prospects for Harmonisation and Networking, 2010. Seville: The Institute for Prospective Technological Studies (IPTS).
35. Watson PH, Ravid R, Eng CB, *et al*. What are the main roadblocks to transnational biobank collaboration, and how can we overcome them? *Biopreserv Biobank* 2011;**9**:213–16.
36. Chen H, Pang T. A call for global governance of biobanks. *Bull World Health Organ* 2015;**93**:113–17.
37. Lind A-S, Reichel J, Österdahl I. Transparency in EU research governance? A case study on cross-border biobanking. Information and Law in Transition – Freedom of Speech, the Internet, Privacy and Democracy in the 21st Century, 2015. Stockholm: Liber.
38. Gaskell G, Gottweis H, Starkbaum J, *et al*. Publics and biobanks: pan-European diversity and the challenge of responsible innovation. *Eur J Hum Genet* 2013;**21**:14–20.
39. Ewing AT, Erby LAH, Bollinger J, *et al*. Demographic differences in willingness to provide broad and narrow consent for biobank research. *Biopreserv Biobank* 2015;**13**:98–106.
40. Reichel J. The need for a legitimate regulatory regime in bioethics: a global and European perspective. *Scand Stud Law* 2013;**58**:197–216.
41. Campbell LD, Betsou F, Garcia DL, *et al*. Development of the ISBER best practices for repositories: collection, storage, retrieval and distribution of biological materials for research. *Biopreserv Biobank* 2012;**10**:232–233.
42. Mascalzoni D, Dove ES, Rubinstein Y, *et al*. International charter of principles for sharing bio-specimens and data. *Eur J Hum Genet* 2015;**23**:721–728.

IV

Supporting the development of biobanks in low and medium income countries

Tomas KLINGSTRÖM[1], Maimuna MENDY[2], Dominique MEUNIER[2], Anouk BERGER[2], Jane REICHEL[3], Alan CHRISTOFFELS[4], Hocine BENDOU[4], Carmen SWANEPOEL[5], Lemoene SMIT[6], Campbell MCKELLAR-BASSET[6], Erik BONGCAM-RUDLOFF[7], Jonas SÖDERBERG[7], Roxana MERINO-MARTINEZ[8], Suyesh AMATYA[8], Absolomon, KIHARA[9], Steve KEMP[9], Robert REIHS[10] and Heimo MÜLLER[10].

[1]SLU-Global Bioinformatics Centre, Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden Ulls väg 26, Uppsala, Sweden
Tel +4618-672126, Email: tomas.klingstrom@slu.se
[2]International Agency for Research on Cancer, Lyon, France
[3]Faculty of Law & Centre of Research Ethics & Bioethics, Uppsala University, Uppsala, Sweden
[4]South African National Bioinformatics Institute, University of the Western Cape, Bellville, South Africa
6BikaLabs, Western Cape, South Africa
[7]SLU-Global Bioinformatics Centre, Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden
[8]Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
[9]International Livestock Research Institute, Nairobi, Kenya
[10]Institute for Pathology / Center for Medical Research, Graz, Austria

Abstract: Biobanks are an organized collection of biological material and associated data. They are a fundamental resource for life science research and contribute to the development of pharmaceutical drugs, diagnostic markers and to a deeper understanding of the genetics that regulate the development of all life on earth.

Biobanks are well established in High Income Countries (HIC) and are rapidly emerging in Low and Middle Income Countries (LMIC). Surveys among biobanks operating in a LMIC setting indicate that limited resources and short term funding tied to specific projects threaten the sustainability of the biobanks. Fit-for-purpose biobanks targeting major societal challenges such as HIV and Malaria provide an excellent basis for integrating biobanks with the available research communities in LMIC regions. But to become sustainable for the future it is important that biobanks become an integrated part of local research communities. To achieve this, the cost of operating biobanks must be lowered, templates must be developed to support local ethics committees and researchers must be given the opportunity to build experience in successfully operating biobank based research projects.

The B3Africa consortium is based on these conclusions and set up to support biobank based research by creating a cost efficient Laboratory Information Management System (LIMS) for developing biobanks and also contribute to the training and capacity building in the local research community. The technical platform called the eB3Kit is open source and consists of a LIMS and a bioinformatics module based on the eBiokit that allow researchers to take control over the analysis of their own data. Along with the technical platform the consortium will also contribute training and support for the associated infrastructures necessary to regulate the ethical and legal implications of biobank based research.

Keywords: biobank, low- and medium-income countries, biobank and cohort building, eInfrastructure, open source open source software, ethical, legal and social issues, bioinformatics.

Introduction

A biobank is a collection of biologic material and associated data obtained from a population or cohort of individuals/subjects and stored in an organized system. The associated data (which can be linked back to the subject who provided the sample) include epidemiologic, clinical, and lifestyle data. Biobanks are increasingly considered an important platform for medical and scientific research. They provide key resources for studying the etiology and molecular mechanisms of diseases and for developing potential diagnostic biomarkers. Biobanking also contributes significantly to the development of personalized drug treatment through translational research.

Biobanking facilities are well established in High Income Countries (HIC) with efforts from international organisations such as the International Society for Biological and Environmental Repositories (ISBER), the European, Middle Eastern & African Society for Biopreservation & Biobanking (ESBB), the Biobanking and Biomolecular resources Research Infrastructure (BBMRI-ERIC), the US National Institute of Health-National Cancer Institute (NIH-NCI) and the International Agency for Research on Cancer (IARC). These organisations have developed international guidelines and protocols which have contributed to this development[1], [2]. However, in many Low and Middle Income Countries (LMICs) the uptake of the available resources has moved at a slower pace. In many of these settings, standard guidelines and protocols to regulate the collection, management, sharing and use of biological samples for research purposes are not being utilized.

IARC launched the LMIC Biobank and Cohort Network (BCNet) in 2013[3] with the overall objectives to promote capacity-building in LMIC biobanks and to increase opportunities for training and funding. BCNet aims to raise awareness among stakeholders, communities, and decision-makers about the benefits of biobanking as an important infrastructure for research.

A situational analysis of infrastructures and facilities was conducted, in order to gather information on biobanking activities, research infrastructures and resources. Twenty seven institutions from sixteen LMICs participated, including twenty-two institutions from eleven African countries.

Results from the survey showed that, although information on biobanking activities in Africa is limited, biological resource management infrastructure is being developed[4]. Biobanks were introduced through specific programmes targeted at major health issues affecting the populations of these countries (e.g. HIV treatment programmes) and their sustainability is seriously threatened after the project ends.

Lack of access to electronic record management systems and organised databases has made it difficult for the linkage of biobank records to other health related services. In cases where electronic databases are available, they are not harmonised with locally used software to allow direct linkage to associated databases such as cancer registries, clinical and treatment records databases, etc.

The lack of harmonization amongst databases and procedures limits opportunities for research collaborations and less than 30% of the centres are involved in scientific research collaboration.

Technical consideration in selecting and maintaining databases and cost implications with respect to acquisition, installation and management of Laboratory Information Management Systems (LIMS) were limiting factors affecting progress. Many do not have access to dedicated facilities to support the biobank LIMS or personnel with the relevant expertise.

On the other hand, biomedical research is producing a huge amount of big data generated from high-throughput experiments associated to biobanked samples. The lack of infrastructures and software platform supporting storage and analysis of this data is creating a bottleneck in biomedical research[5].

The challenge of ethical, legal and social issues (ELSI) related to biobanking was also highlighted in this survey, as well as in previous reports[6], [7]. Although the different ethics and scientific committees have established various mechanisms to deal with ELSIs in scientific research, in most cases, the committees lack the experience and the regulatory framework to adequately review applications specific to biobanking or biobank projects. This situation creates a barrier for the effective use of biobank samples in these countries as well as international collaboration.

There is also the need to increase the level of participation of African researchers within Africa in a fair and transparent manner.

To address these anomalies, the Human Heredity and Health in Africa (H3Africa) Initiative project has developed a policy framework to promote fair collaboration between scientists in Africa and with the international community[8]. This policy particularly highlight key areas to focus on such as sample and data sharing as well as the importance of a data and biospecimen access committee.

In this article we present the solutions that are being proposed by the EU-H2020 B3Africa project. The project builds on the conclusions from the BCNet and other projects such as the H3Africa Initiative[8]. It proposes an innovative solution, integrating available open-source software, services and tools for biobanking, bioinformatics, ethics, regulation and training.

The main goal of the project is to create a collaboration framework that bridges European and African biomedical research and provides a technical platform (eB3Kit) that implements and integrates the necessary components of this framework including: biobanking, bioinformatics, education, training and dissemination.

Defining an ethical and regulatory framework

There are two basic notions of bioethics that can be considered universally accepted, the need for informed consent and ethical approval to allow a safe and legitimate handling of biological samples and data within research[9]. The B3Africa project builds upon these two notions as well as through a Model Data Management Policy (MDMP) based on the legal framework. The MDMP will provide a set of formal rules, criteria and priorities that should guarantee a consistent ascertainment of all requirements that should be fulfilled by the platform and by the users of the B3Africa platform.

Four main risks have been identified within the project; a) The B3Africa project encounters a complex legal landscape which is difficult to navigate. This creates a risk that the legal tools used are incorrect. b) Some of the prospective users of the B3Africa platform may be situated in states with underdeveloped bioethical legislation, where procedures that would be considered unethical elsewhere, remain legal. c) In states with underdeveloped bioethical legislation, it is likely that the governing infrastructure is also underdeveloped, for example institutional control and audit of biobanks, research ethics boards, etc. d) If the bioethical regulations are too strict, research in itself may be hampered, especially if all national laws of the participating states must be upheld in practice. The potential limitations for the future use of research data according to ethical regulation and guidelines, for example the (African Union) AU convention on cyber security and personal data protection, could lead to global national control, unless there is specific provision for data access for future scientific research.

Legal framework

All processing of biomedical data within the B3Africa project should adhere to two basic principles, *informed consent* and *ethical approval*, both in regards to processing within a state and for cross-border sharing. The legal implementations of the principles are carried out within each national legal order, by national authorities enacting administrative decisions applicable within the state. For sharing data between two states, there is a need to find a model for connecting the ethical approval from the sending state to the receiving state[10]. International administrative law provides two approaches that can be applied[11]. First, a common rule can be established for all entities to apply, together with an obligation for all to accept each other decisions (home state control). Secondly, the collaboration can be built on a conflict of law-approach, leaving each state to decide for themselves how to govern the issue at hand and to develop tools to connect to other administrative orders, for example via agreements in each individual case.

The B3Africa project will employ both strategies to safeguard compliance with relevant law set out above, the home state control approach and the conflict-of-law approach, in a model adapted to the project. A common understanding of the principles of informed consent and ethical approval will function as threshold-principles, basic requirements that all parties must follow in their internal work. These requirements should thus be upheld by all and will be implemented in the project via the Model Data Management Policy (MDMP) and the Data Model, explained below. When collaboration cross-borders, each transfer of data must be governed by an individual legal tool, for example a data transfer agreement. Standard versions of data transfer agreements, especially targeting transfer of medical data, may be drafted in the project, for the parties to use or to adapt according to their own needs. In developing of this part of the framework, inspiration will be taken from other relevant international research projects and infrastructure, such as the work within the common service ethical, legal and societal issues of BBMRI-ERIC (http://bbmri-eric.eu), IARC (http://www.iarc.fr) the H3Africa project (http://h3africa.org) and the RD-Connect project (http://rd-connect.eu).

Model Data Management Policy (MDMP)

The MDMP is based on the B3Africa legal framework and will provide a set of formal rules, criteria and priorities that should guarantee a consistent ascertainment of all requirements that should be fulfilled by the platform and by the users of the B3Africa platform. This policy will be implemented as part of the B3Africa final product, the eB3Kit and will pave the way for managing the use of the eB3Kit in control and regulated environments beyond the project's life time.

The model will be based on common standards regarding informed consent and ethical approval. The definition of the concepts will be determined by the applicable law on site. If the applicable law does not regulate these issues, the minimum requirements set out within the legal framework of the project will be applied. For all the European users it will be based on the EU Data Protection Directive (and in due time, the General Data Protection Regulation). Central features of the MDMP are an *Adoption Committee* that is tasked with evaluating adoption requests from potential users. In order to enter the platform, each user needs an *Organization membership* as well as *Individual membership* for the person conducting the analysis. An application for an organizational membership is done via a standardized application form in which the requester provides information regarding the organization, (contact information, aim and purpose, if the organisation is private or public, means of funding, forms of supervision over the organisation, etc.). Once the B3Africa platformeB3Kit has been adopted by an organization or group, the individual requesters is also asked to provide information. Only accepted members can request store, management and analysis services. Before data is uploaded, standardized documents with the approvals by ethics committee or institute research boards for the research project, consent information from sample donors whose samples have generated the data, and, if applicable, for authorization of use of non-consented data must be uploaded.

Technology Description

B3Africa will provide a curated platform for open-source software in the biobank domain built upon the *BiBBoX* (biobank software in a box). Within *BiBBoX* software components are pre-installed and configured ready to go with minimal IT effort. The *BiBBoX* core module will cover the core functionality necessary to operate a biobank and the repository will also contain docked software tools such as a bioinformatics module based on the eBiokit[12]. In addition APIs and interfaces will be specified to integrate open source software solutions in the areas electronic health records (EHR), patient and study management, imaging and data integration and analysis. The first version of *BiBBoX* is already accessible at http://bibbox.org/.

BiBBoX architecture

The *BiBBoX* system architecture is built on top of a virtual machine and docker containers, see figure 1. A lightweight central component (green part in figure 1) will provide functionality for the deployment of software tools, a central ID und user management and a user interface based on the Liferay portal (www.liferay.com).

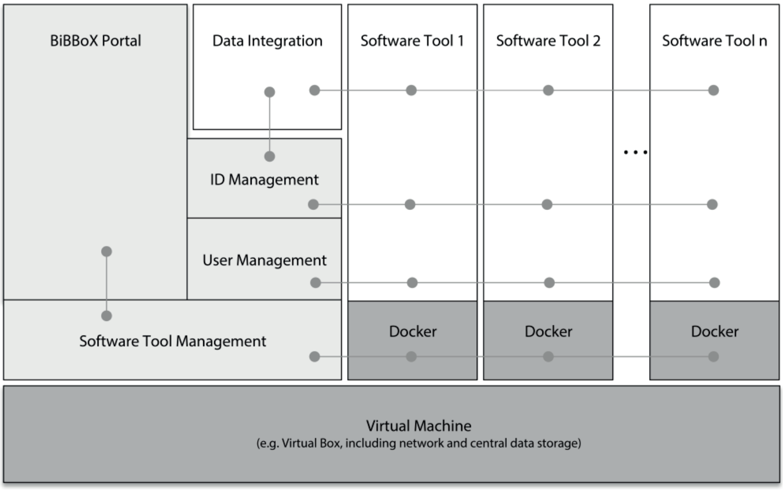


Figure 1: The BiBBoX System Architecture, central component in green, integration software in blue and docked software in white.

Data exchange between software tools and ID management will follow the MIABIS recommendations.

The Minimum Information About BIobank data Sharing (MIABIS) was developed in 2012 by the Biobanking and BioMolecular Resources Research Infrastructure of Sweden (BBMRI.se). In 2013 a working group was formed under BBMRI-ERIC to continue the development of MIABIS through a multi-country governance process. MIABIS is the “de facto” biobank information standard for BBMRI-ERIC community and has been widely accepted within Europe and beyond.

The minimum information guidelines consist of a collection of components with associated attributes representing relevant concepts from biobanking and biomedical research and can be used to integrate the most relevant building blocks of the biomedical research ecosystem (https://github.com/MIABIS/miabis/wiki).

MIABIS has been implemented in several projects and e-infrastructures as BiobankCloud (http://www.biobankcloud.com/), RD-Connect (http://rd-connect.eu/), BioMedBridges (http://www.biomedbridges.eu/), BCNet Catalogue (http://bcnetcat.iarc.fr/), BBMRI-ERIC Directory (http://bbmri-eric.eu/bbmri-eric-directory-2.0), among others.

B3Africa platform will implement MIABIS as part of the data model for representing and sharing biobank and research data which will lead to a wider and more efficient use of valuable bio-resources.

Based on previous requirement analysis work [13], [14], we specified core functionality for the *BiBBoX* and identified an initial set of core modules for the first *BiBBoX* release.

BiBBoX core module

The *BiBBoX* core module consists of software tools for the organization of samples and related data in the context of a collection / study protocol. It includes functionalities for sample acquisition and sample metadata management, sample processing, sample storage, sample and data retrieval/distribution (provided by BIKA http://www.bikalabs.com/) as well as data integration and cataloguing. In particular the following functionality will be supported:

Organize Collections, Studies and SOPs: Each collection has to be based on a specific study design, follow well-documented rules, and describe responsible stakeholders (including their roles), sample types, data formats, ontologies and business processes. The *BiBBoX* core module will provide functionality to handle all this information in a version-controlled manner and will provide a user management system that defines users and what actions they are entitled to perform in the system.

Sample Acquisition and Sample Metadata Management: Each sample and aliquot in a collection is related to a specific patient / donor and to a medical / study event. Basic information about patient / donors and medical / study data objects will be available in the *BiBBoX* core module, at least global unique identifiers for patients and medical/study events will be provided for each sample and aliquot. For sample acquisition in the field we plan to integrate the ILRI monitoring system as software component[15].

Sample Processing: In a typical biobank environment various kinds of procedures are carried out on samples, e.g. preservation procedures, generating aliquots and more complex processing such as deviation of the original material. To ensure that the fate of all of the sample material is known, each aliquot and sub-sample will be tracked and their relationship to the parent sample will be recorded. Even if all of the sample material has been consumed, this needs to be recorded, so that the processing history is known.

Sample Storage: After processing and validation samples are moved to a (long term) storage. The *BiBBoX* core module will assist users to organize shelf spaces and to find stored samples again when they need to retrieve them. The system will track sample locations and monitor freezing and thawing cycles, as well as disposal of them. For bigger biobanks automated systems and storage robots will be supported.

Data Integration and Cataloguing: The *BiBBoX* core module will provide functionality to import and integrate external data objects and generate a catalogue of physical and data objects available in the biobank. A catalogue will include metadata elements collected in the sample acquisition process including sample availability and access conditions. In addition biobanks metadata describing storage and quality parameters will be included. We plan to implement the catalogue functionality based on the Molgenis platform[16] providing both sample level and aggregated information as well as search functionality.

Sample / data retrieval and shipping, retention and destruction: The *BiBBoX* core module will provide functionality to manage samples in transit between source and destination, for both receiving samples and redistributing them to authorized recipients. Shipping is related to the whole chain of custody and storage inventory and the sample access management. In addition also a retention schedule for samples and data defining how long samples and associated sample records will be supported.

Administrative, business and management support: The *BiBBoX* core module will support administrative tasks of the biobank operation, including workflow and customer management. For this we will provide functionalities, which allow managers and auditors to monitor the biobank operation.

Docked bioinformatics module

The BiBBoX integrate a bioinformatics modules based on the eBiokit[12] which is a self-contained computing platform providing users with access to popular bioinformatics software and services. The local storage allows researchers to conduct advanced data analysis in a user friendly environment and access local copies of commonly used databases. Combined with extensive in-built tutorials this system allows non-specialists to process data generated from a biobank collection.

Data analysis is often a significant bottle neck in modern life science and the bioinformatics component in the eB3kit provides research institutions with the means to more efficiently distribute the work load. Junior researchers can perform routine operations and screening while simultaneously learning the basics of bioinformatics. Research institutions can then either use the eB3kit bioinformatics toolkit as a foundation for further specialization in bioinformatics or use the data produced as a basis for collaboration with research groups specialized in bioinformatics.

Education and capacity building

The eB3kit will contribute to establishing an open-innovation ecosystem in Africa and Europe. In doing so it is important that researchers in Low and Middle Income Countries have the infrastructure and know-how to participate in the co-creation and exploration of ideas. To bring a real added-value to the institutes and countries in which it will be implemented, the eB3Kit will therefore be supported by an extensive education and capacity building effort where different enabling factors have been considered.

First, the project recognizes the paramount importance of having the B3Africa concept validated by biobanks from both continents. This implies an active involvement of a different range of actors from involved biobanks, providing complementary technical, scientific, ethical, legal and political perspectives. Different actions are planned to engage those actors: ethics and legal issues meeting, stakeholder forums, technical jamborees.

In addition, the eB3kit will be installed and tested in real-life settings (use case), using a step-wise approach. Three centers involved in H3Africa were first selected to take part of the use case work package.

- National Health Laboratory Services - Stellenbosh University Biobank, Cape Town, South Africa
- Institute of Human Virology, Abuja, Nigeria
- Makerere University College of Health Sciences, Kampala, Uganda

The involvement of those three centers represents a great added-value for both the B3Africa and the H3Africa projects, as it reinforces synergies and complementarities. Since the beginning of the B3Africa project, the International Livestock Research Institute (Nairobi, Kenya) already partner of the project, has also volunteered to act as use case for non-human biobank setting.

BCNet, from which the B3Africa project could largely draw on, has also been identified as an important source for use case centers. With twenty-one members from seventeen countries (seventeen centers from thirteen African & Middle East countries, along with centers from two European countries and two Asian countries), this dynamic network represents great opportunities for testing and dissemination of the eB3kit. Equally, the eB3kit will provide a concrete solution to some of the challenges reported by BCNet members. Initially, four BCNet members will be take part in the use case work package:

- Breast Care International, Peace and Love Hospital, Kumasi, Ghana
- Medical Research Council, International Nutrition Group, Banjul, The Gambia
- National Cancer Institute, Vilnius, Lithuania
- Wroclaw Research Centre, EIT+ Biobank, Wroclaw, Poland

Furthermore, key professionals involved in the implementation of the eB3Kit will be trained. Courses and tailored learning materials will be developed based on a detailed learning needs assessment. Inputs from B3Africa partners in charge of developing the eB3Kit will be requested in order to take into account the specificities of the various tools included in the kit (technology and various components, required knowledge and skills, available learning resources, etc.). Professionals at use case institutions will also be consulted in order to take into account the various profiles of those responsible for the implementation of the eB3Kit ("focal points"), as well as end users of the kit. Besides knowledge and skills specific to eB3kit components (tool-specific), other competencies (core competencies) will be considered for an effective and sustainable use of the eB3Kit. Available learning resources/opportunities within B3Africa and related partners (i.e. H3Africa, ESBB, ISBER, BBMRI, etc.) will be identified, in order to maximise synergies between existing laboratory capacity building initiatives. Over the first years of the project, training initiatives will target the use case institutions listed above.

Finally, learning materials developed throughout the implementation of the project will be produced as standalone generic resources organised as a standardized learning environment such as the eBioKit used in the Pan African Bioinformatics Network for H3Africa (H3ABionet). Resources will be more widely disseminated through awareness raising actions on project activities and results, as well as dissemination of project deliverables and outcomes to relevant stakeholders, including at the policy level. Besides communication activities of the B3Africa project, a course will be organised towards the end of the project and will target professionals from other biobanks interested in the use the eB3kit.

All the above will enhance the ability to conduct training for new and/or developing biobanks and therefore strongly contribute to create a sustainable network of biospecimen repositories infrastructures interacting and sharing knowledge between Europe and Africa.

Maximizing the social and economic impact of biobanks in Africa

The surveys among biobanks operating in a LMIC environment show that biobanks have so far mainly been created to support specific programmes targeting major health issues in the countries hosting the biobank[4]. Such programmes may be successful in achieving their aims and also set a precedent for effective biobank management in the region. But experience in the field show that these benefits may quickly be lost unless the biobanks evolve into a sustainable infrastructure embedded in the local research community (figure 2).

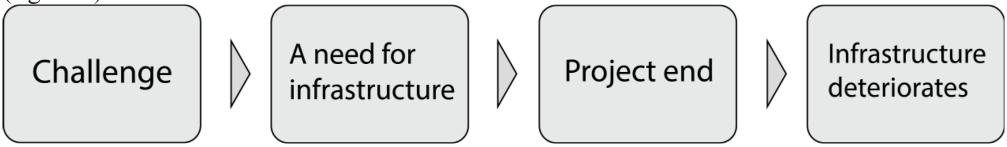


Figure 2, the life cycle of a biobank if sustainability is not achieved.

To ensure sustainability it is important for projects like the B3Africa project to take a holistic perspective towards biobanking and its role in research. Research is not a continual flow of progress from point A to point B but rather an endless number of research cycles where ideas are conceived, tested and hopefully contribute to the shared knowledgebase of society. Biobanks contribute to this process by facilitating the efficient management of biological samples and associated data. In a fit-for-purpose biobank this task is well defined and the project justifying the creation of a biobank can be expected to handle the research tasks outside simple storage and management.

For a biobank to become sustainable it must strengthen the local research community enough to justify its costs. The B3Africa project achieves this by not only focusing on a technical solution for tracking samples and associated information but also to support researchers all the way from the ideation stage to knowledge generation. Such an integration into a local research community cannot be limited to technical solutions but must also incorporate social and legal factors in order to ensure its integration into the local community into the program. The B3Africa project therefore consist of a combination of the open source software package called the eB3Kit and a social integration program aimed towards enabling communities adopting the software(figure 3).

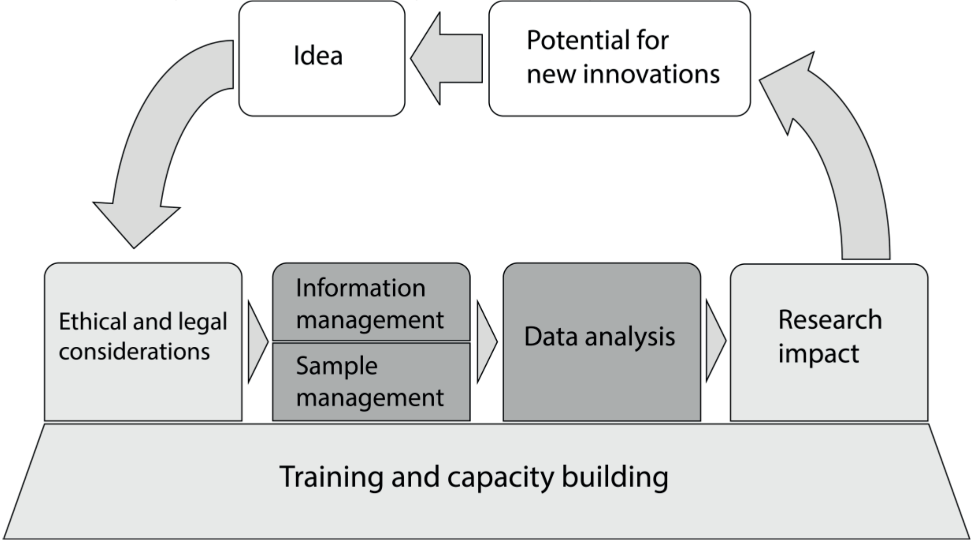

Idea
Potential for new innovations
Ethical and legal considerations
Information management
Sample management
Data analysis
Research impact
Training and capacity building

Figure 3,the B3Africa project is built upon a technology platform in the form of the eB3Kit (blue) and social components (green).Together the components contribute to strengthen African and European research communities in LMIC by reducing the thresholds in the research cycle..

The successful completion of the B3Africa project will therefore not only be reliant on production and availability of a high quality open source LIMS in the form of the eB3Kit, but also on its successful integration into local research communities across the Europe and Africa. Major initiatives such as H3Africa and BCNet provide a valuable opportunity to embed biobanking as a natural component of research in a LMIC setting. But smaller local initiatives serve as a complement to major programs as they can draw upon earlier experience and implement them in local research communities[17]. The eB3Kit provide the information technology necessary to provide the services of a biobank in a robust and cost efficient management. Combined with the training and ELSI components this allows the small scale creation of biobanks that can be integrated into the global biobanking networks through the regional networks initiated by H3Africa and BCNet.

[1] E. Caboux, A. Plymoth, P. Hainaut, International Agency for Research on Cancer, and IARC World Forum, *Common minimum technical standards and protocols for biological resource centres dedicated to cancer research*. Lyon, France; Geneva: International Agency for Research on Cancer ; Distributed by WHO Press, 2007.

[2] L. D. Campbell, F. Betsou, D. L. Garcia, J. G. Giri, K. E. Pitt, R. S. Pugh, K. C. Sexton, A. P. N. Skubitz, and S. B. Somiari, "Development of the *ISBER Best Practices for Repositories: Collection, Storage, Retrieval and Distribution of Biological Materials for Research*," *Biopreservation and Biobanking*, vol. 10, no. 2, pp. 232–233, Apr. 2012.

[3] IARC, "BCNet Launch," *BCNet Launch*, 16-Jan-2016. [Online]. Available: http://www.iarc.fr/en/media-centre/iarcnews/pdf/BCNet%20Launch.pdf.

[4] M. Mendy, E. Caboux, B. S. Sylla, J. Dillner, J. Chinquee, C. Wild, and BCNet survey participants, "Infrastructure and Facilities for Human Biobanking in Low- and Middle-Income Countries: A Situation Analysis," *Pathobiology*, vol. 81, no. 5–6, pp. 252–260, Mar. 2015.

[5] BiobankCloud, "Attacking the Biobank Bottleneck." 18-Jan-2016.

[6] A. Abayomi, A. Christoffels, R. Grewal, L. A. Karam, C. Rossouw, C. Staunton, C. Swanepoel, and B. van Rooyen, "Challenges of Biobanking in South Africa to Facilitate Indigenous Research in an Environment Burdened with Human Immunodeficiency Virus, Tuberculosis, and Emerging Noncommunicable Diseases," *Biopreservation and Biobanking*, vol. 11, no. 6, pp. 347–354, Dec. 2013.

[7] European Commission and Directorate-General for Research and Innovation, *Biobanks for Europe: a challenge for governance*. Luxembourg: EUR-OP, 2012.

[8] J. de Vries, P. Tindana, K. Littler, M. Ramsay, C. Rotimi, A. Abayomi, N. Mulder, and B. M. Mayosi, "The H3Africa policy framework: negotiating fairness in genomics," *Trends in Genetics*, vol. 31, no. 3, pp. 117–119, Mar. 2015.

[9] M. Ruffert and S. Steinecke, *The Global Administrative Law of Science*, vol. 228. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.

[10] A.-S. Lind, J. Reichel, and I. Österdahl, "Transparency in EU research governance? A case study on cross-border biobanking.," in *Information and Law in Transition – Freedom of Speech, the Internet, Privacy and Democracy in the 21st Century*, Liber, 2015.

[11] H. Wenander, "A toolbox for administrative law cooperation beyond the state," in *Administrative law beyond the state: Nordic perspectives*, Leiden : Stockholm: Martinus Nijhoff Publishers ; Liber, 2013.

[12] H. Fuxelius, E. Bongcam, and Y. Jaufeerally, "The contribution of the eBioKit to Bioinformatics Education in Southern Africa," *EMBnet.journal*, vol. 16, no. 1, p. 29, Sep. 2010.

[13] Müller H., Reihs R., Zatloukal K., Jeanquartier F., Marino-Martinez R., van Enckevort D., and Swertz M., "State-of-the-Art and Future Challenges in the Integration of Biobank Catalogues.," in *Smart Health Lecture Notes in Computer Science*, LNCS 8700, Berlin, Heidelberg: Springer, 2015, pp. 261–273.

[14] Müller H, Reihs R, and Zatloukal K, "User experience and Usability requirements for NGS workflows in clinical applications.," in *Proceedings of the Medicon 2016 conference*, Cyprus, 2016.

[15] M. Norling, A. Kihara, and S. Kemp, "Web-Based Biobank System Infrastructure Monitoring Using Python, Perl, and PHP," *Biopreservation and Biobanking*, vol. 11, no. 6, pp. 355–358, Dec. 2013.

[16] M. A. Swertz, M. Dijkstra, T. Adamusiak, J. K. van der Velde, A. Kanterakis, E. T. Roos, J. Lops, G. A. Thorisson, D. Arends, G. Byelas, J. Muilu, A. J. Brookes, E. O. de Brock, R. C. Jansen, and H. Parkinson, "The MOLGENIS toolkit: rapid prototyping of biosoftware at the push of a button," *BMC Bioinformatics*, vol. 11 Suppl 12, p. S12, 2010.

[17] T. Klingström, "Biobanking in Emerging Countries," *Biopreservation and Biobanking*, vol. 11, no. 6, pp. 329–330, Dec. 2013.

Galaksio, a user friendly workflow-centric front end for Galaxy

Tomas Klingström[1*], Rafael Hernández-de-Diego[1*] and Erik Bongcam-Rudloff[1]

1. SLU-Global Bioinformatics Centre, Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Sweden.

* These authors contributed equally to the work.

Abstract

There is a severe shortage of statisticians and bioinformaticians available in research. As universities fail to cover the increasing need of graduates with the necessary skills, *ad hoc* training and workshops have become commonplace but are insufficient to cover the needs. Technical solutions that distribute the workload more efficiently between researchers with a different education background (e.g., computer scientists and biologists) are therefore necessary to cover some of this shortage.

Galaksio provides a workflow-centric graphical user interface for the Galaxy Workflow Management system easy to use for biologists and medical researchers who need to run routine tasks in bioinformatics. Combined with back end tools such as BioBlend, CloudMan and Pulsar, Galaksio provides a novel, layered approach to Galaxy making it easier to divide research tasks to researchers depending on their skills in interdisciplinary subjects such as bioinformatics and computational science.

Galaksio is developed by the B3Africa project for the eB3Kit but functions as a stand-alone server that can be configured for to be connected to any Galaxy server using the Galaxy API. Galaksio can be downloaded at: https://github.com/fikipollo/galaksio.

Key points

- Galaksio is built to provide a more layered approach to Galaxy, providing a simplified user interface based on workflows.
- Galaksio reduces the workload of bioinformaticians as routine tasks can be performed with minimal training. The presentation of workflows also provides a comprehensive overview of necessary input data as well as methodological changes to the end user.
- Galaksio can be used to rapidly deploy new services. Public Galaxy servers are a powerful tool to support collaborative research and Galaksio provides a more lightweight user interface for researchers who wish to make a specific project or workflow available.

Introduction

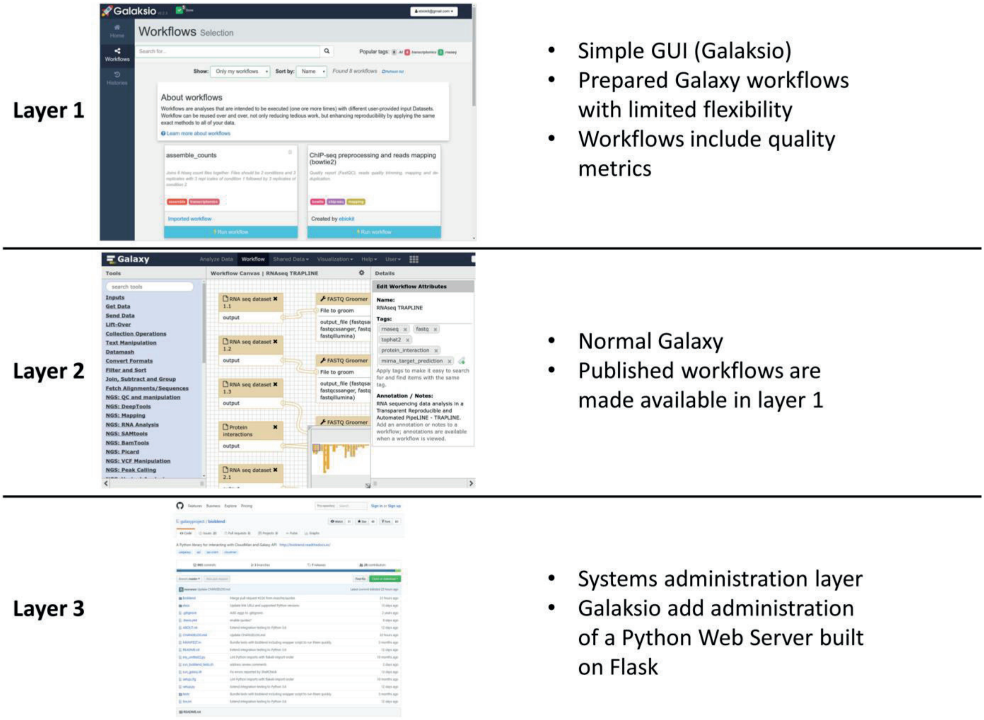
Galaxy is a widely supported workflow management system used in bioinformatics (Goecks *et al.*, 2010; Leipzig, 2016; Tastan Bishop *et al.*, 2015; Atwood *et al.*, 2015) to facilitate accessible and reproducible research. One of the main aims of Galaxy is to provide access to bioinformatic analysis tools for experimentalists with limited expertise in programming (Atwood *et al.*, 2015; Blankenberg *et al.*, 2010). Nevertheless, our experience with Galaxy, gained by implementing it in the eBiokit (Hernández-de-Diego *et al.*, 2017) and by using Galaxy in several training and capacity building projects (Fuxelius *et al.*, 2010; Atwood *et al.*, 2015; Mulder *et al.*, 2016) shown us that many potential Galaxy users find themselves in a bit of a conundrum when trying to use Galaxy. Researchers skilled enough in bioinformatics to install and configure tools prefer command line tools, whereas less advanced users are left on their own struggling to find and combine tools using the user interface provided by Galaxy. Therefore, many research groups remain reliant on in-house scripts maintained by a small number of bioinformaticians spending significant time on providing *ad hoc* support to other researchers in the group. To provide an attractive technology platform for researchers it was therefore deemed necessary to provide a more simplified, workflow-centric model of operations. In the workflow-centric model researchers with limited bioinformatics training are provided with prepared workflows and default input parameters, while more advanced users can create and modify workflows using the normal Galaxy GUI. This allows research teams to work in a more efficient way. Trained bioinformaticians can adapt and develop tools and then provide the finished workflows for routine analysis to lab researchers.

In standard Galaxy all users rely on the same GUI, despite significantly different education background and expertise. Trained bioinformaticians often rely on a set of skills dependent on education decisions taken by students several years ahead of enrolling at a university (Wightman and Hark, 2012) while other researchers may have little or no formal training. Given the complexities of training needs, influential stakeholders such as the US National Research Council has therefore concluded that bioinformatics research is likely to be carried out by two disparate groups of researchers: quantitative biologists, who work at the interface of mathematical/computer science and biology, and research biologists, who need familiarity with a range of mathematical and computational concepts without necessarily being an expert (National Research Council (US) Committee on Undergraduate Biology Education to Prepare Research Scientists for the 21st Century, 2003)

We therefore present Galaksio, a solution based on the Galaxy API and a Python web server, that we have developed to provide a layered access to Galaxy functions that facilitate the work of research biologists through an easy-to-use web interface, while the default Galaxy interface is used by bioinformaticians to create new workflows and systems administration tasks that are facilitated by packages created by other researchers such as BioBlend (Sloggett *et al.*, 2013), CloudMan (Afgan *et al.*, 2010) and Pulsar (Afgan *et al.*, 2015). With Galaksio, all data is managed within the normal Galaxy workflow management system and user credentials are passed on to the Galaxy server to manage user privileges, meaning that Galaksio can be used to access all workflows created on a normal Galaxy server using the command line tools implemented on the server.

Thanks to Galaksio, the Galaxy user's experience can be managed at three different levels: 1) a layer suited to research biologists (i.e., users using tools); 2) a layer suited to bioinformaticians (i.e., users developing tools); 3) a layer suited to computer scientists (i.e., users developing the environment tools work in) (Figure 1).

This approach is currently being implemented in the B3Africa project using the eB3Kit which includes Galaksio and relies on these resources to connect the relatively light weight Mac Pro Server, commonly hosting the eB3Kit, to external computing resources (Klingstrom *et al.*, 2016).



Layer 1

- Simple GUI (Galaksio)
- Prepared Galaxy workflows with limited flexibility
- Workflows include quality metrics

Layer 2

- Normal Galaxy
- Published workflows are made available in layer 1

Layer 3

- Systems administration layer
- Galaksio add administration of a Python Web Server built on Flask

Figure 1. This figure shows the layered approach used by Galaksio and implemented in the eB3Kit to divide labour more efficiently between researchers with different background.

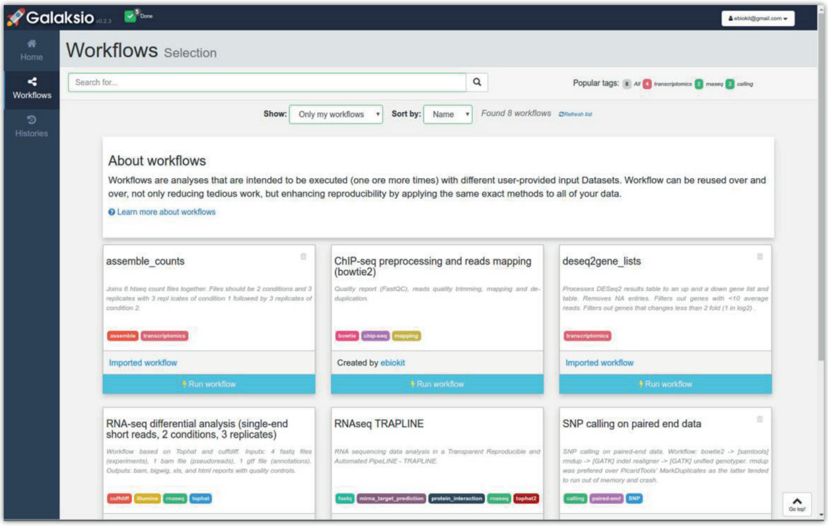
Materials, methodologies and techniques

Galaksio has been designed as a multiuser web application and is divided in two components: the server side application and the web interface for users.

The server side, which is built on Python Flask server(http://flask.pocoo.org/), is responsible for accessing the Galaxy data using the tools provided by the Galaxy application programming interface (API) (Blankenberg *et al.*, 2010; Goecks *et al.*, 2010). The Galaksio web interface has been developed using AngularJS(https://angular.io) and Bootstrap(http://getbootstrap.com), both popular HTML, CSS, and JavaScript cross-browser frameworks for developing responsive and user-friendly web applications. The exchange of the data between clients and the server is handled using asynchronous JavaScript and XML (AJAX) communication.

Results

Galaksio is free to use and is distributed under the GNU General Public License, Version 3. A public copy of the application is hosted at the SLU facilities as part of the eBioKit platform (http://ebiokit.eu/) and source code is available at GitHub(https://github.com/fikipollo/galaksio), allowing other laboratories to browse, propose code reviews, and download the code in order to set up their own instance of the application. Additionally, Galaksio can easily be installed using Docker(https://www.docker.com), an open-source virtualisation software that provides a lightweight, stand-alone, portable, and ready-to-execute package that includes the software and all the dependencies necessary to run the application independently of the operating system installed on the server. Documentation for the project can be found at the ReadTheDocs platform(https://galaksio.readthedocs.io).



About workflows

Workflows are analyses that are intended to be executed (one ore more times) with different user-provided input Datasets. Workflow can be reused over and over, not only reducing tedious work, but enhancing reproducibility by applying the same exact methods to all of your data.

Learn more about workflows

assemble_counts

Joins 6 htseq count files together. Files should be 2 conditions and 3 replicates with 3 repl icates of condition 1 followed by 3 replicates of condition 2.

assemble transcriptomics

Imported workflow

Run workflow

ChIP-seq preprocessing and reads mapping (bowtie2)

Quality report (FastQC), reads quality trimming, mapping and de-duplication.

bowtie chip-seq mapping

Created by ebiokit

Run workflow

deseq2gene_lists

Processes DESeq2 results table to an up and a down gene list and table. Removes NA entries. Filters out genes with <10 average reads. Filters out genes that changes less than 2 fold (1 in log2) .

transcriptomics

Imported workflow

Run workflow

RNA-seq differential analysis (single-end short reads, 2 conditions, 3 replicates)

Workflow based on Tophat and cuffdiff. Inputs: 4 fastq files (experiments), 1 bam file (pseudoreads), 1 gtf file (annotations). Outputs: bam, bigwig, xls, and html reports with quality controls.

cuffdiff illumina rnaseq tophat

RNAseq TRAPLINE

RNA sequencing data analysis in a Transparent Reproducible and Automated PipeLINE - TRAPLINE.

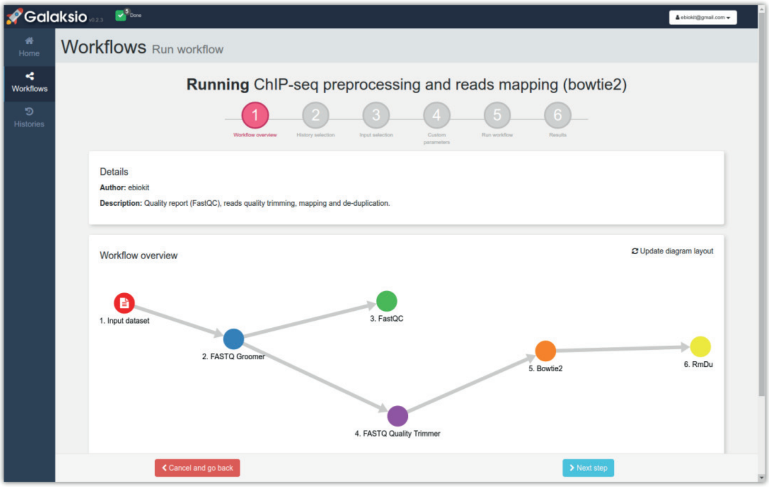
fastq mirna_target_prediction protein_interaction rnaseq tophat2

SNP calling on paired end data

SNP calling on paired-end data. Workflow: bowtie2 -> [samtools] rmdup -> [GATK] indel realigner -> [GATK] unified genotyper. rmdup was prefered over PicardTools' MarkDuplicates as the latter tended to run out of memory and crash.

calling paired-end SNP

Figure 2. The figure shows the graphical interface for the workflow selection in Galaksio.



Running ChIP-seq preprocessing and reads mapping (bowtie2)

Details

Author: ebiokit

Description: Quality report (FastQC), reads quality trimming, mapping and de-duplication.

Figure 3. The figure shows the Galaksio web interface that is presented to the user after the selection of a workflow.

Figure 2 shows the Galaksio's GUI for biologists. Using this interface users can run any workflow implemented in the associated Galaxy instance in just few clicks and get a clear image of the analysis steps included in the selected workflow (Figure 3). The user interface allows the user to customise the execution of pre-selected tools, the uploading of the necessary files, the downloading of the results, and the execution of several workflows simultaneously in the background.

Table 1 provides an overview of all the developed features in the current Galaksio version. As all interactions with Galaxy are managed through the Galaxy API, the Galaksio implementation can be hosted independently as a separate server sending commands to any available Galaxy server. This includes public servers such as the popular usegalaxy.org website(https://usegalaxy.org/). Information on the connected server is provided when logging in via the Galaksio interface. It should however be noted that Galaksio, while light-weight in itself, is completely dependent on the speed of the Galaxy server when returning workflows and any user restrictions defined by the Galaxy server.

Feature	Category	Implemented	Planned
User sign-in/out	Users	X	
User sign-up	Users	X	
Workflow listing	Workflows	X	
Workflow importing	Workflows	X	
Workflow execution	Workflows	X	
Workflow creation	Workflows		X
Simultaneous execution of workflows	Workflows	X	
Recovering previous executions	Workflows	X	
Help and description for tools in workflow	Workflows	X	
Input selection and parameter configuration	Workflows	X	
History selection	History	X	
History creation	History		X
History deletion	History		X
Dataset uploading	Dataset manipulation	X	
Dataset downloading	Dataset manipulation	X	
Dataset deletion	Dataset manipulation	X	
Dataset collection creation	Dataset manipulation	X	
Dataset collection deletion	Dataset manipulation		X
Tool execution	Tools		X

Table 1. Implemented and planned features for Galaksio.

Use case

Due to delays in achieving approval for tool wrappers created by the Galaksio team, an alternative use case has been created with much appreciated support from Marius van den Beek at the Institut Curie, Paris, France. The test dataset is available from the Zenodo data repository (Freeberg and Heydarian, 2016) but all data can also be imported from usegalaxy.org.

History containing dataset collections: https://usegalaxy.org/u/tomkl/h/galaksio-use-case-mouse-chip-seq-data.

Main workflow: https://usegalaxy.org/u/tomkl/w/copy-of-imported-parent-workflow-chipseq

Subworkflow: https://usegalaxy.org/u/tomkl/w/copy-of-imported-chipseqtutorialchild1

The workflows can be imported inside Galaksio by any users logged into a Galaksio server connected to https://usegalaxy.org. Other use cases will be added with the addition of “Galaksio use case” in the name of the workflow to make them easy to be identified in the Galaksio’s repository. Issues are tracked using the Galaksio repository on GitHub (https://github.com/fikipollo/galaksio/issues) and external contributions are welcome.

Discussion

Compared to the clearly defined classes of "research biologist" and "quantitative biologist", proposed by the US National Research Council, bioinformatics has developed into a field where its practitioners share a number of characteristics, but none of which are essential enough to characterise what a bioinformatician truly is (Vincent and Charette, 2015). Many people may therefore be highly skilled and productive researchers in bioinformatics, despite very limited skills in one or more of the core competencies associated with being a bioinformatician (Smith, 2015). Due to the shortage of comprehensive university programmes in the field (Williams and Teal, 2017; Atwood *et al.*, 2015), most researchers currently active in bioinformatics have participated in a number of courses, workshops and self-learning sessions that, step by step, has taken them to a skill level where they may be considered qualified bioinformaticians or quantitative biologists. Such a self-organised curriculum encourages bioinformaticians to obtain exactly the skills necessary to complete their own projects but with limited consideration for auxiliary skills such as code documentation and a deeper understanding of computer science.

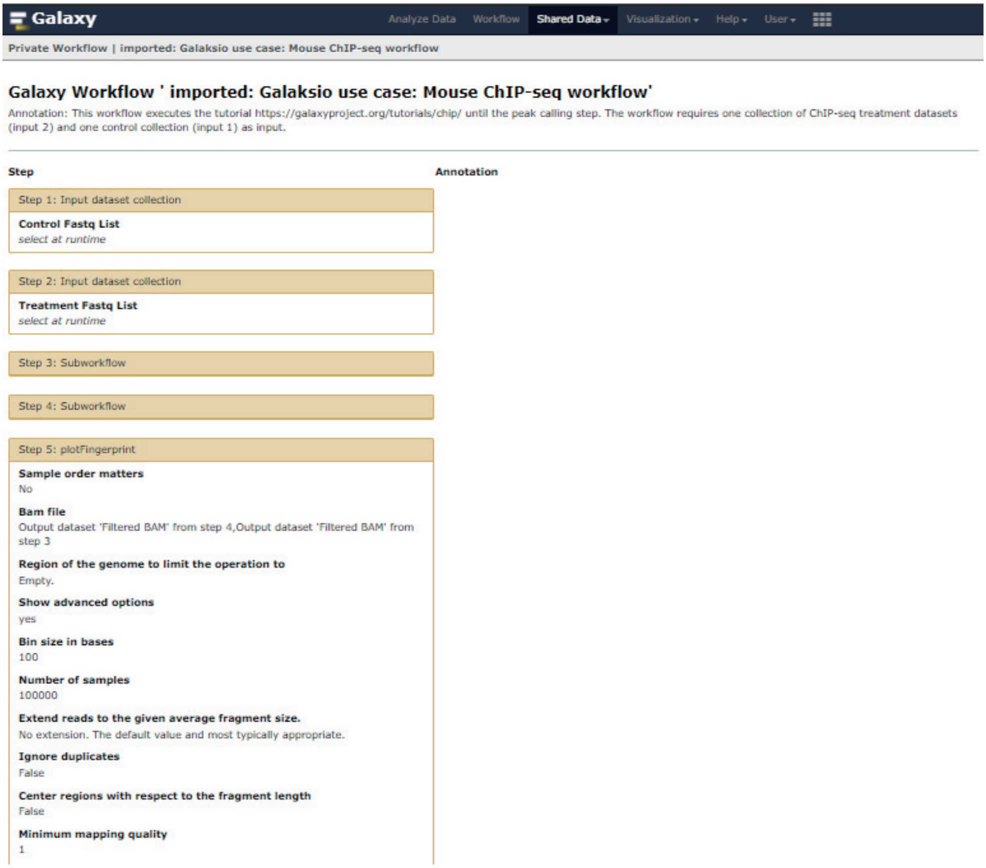
As a result of this self-motivated style of learning, significant delays occur when new technologies emerge if they require significant retraining of practitioners before becoming fully competitive with the new solution. This is perhaps most evident in the slow adoption of distributed computing systems such as Hadoop(http://hadoop.apache.org/). While significant investments in large Hadoop infrastructures has been made, the production of bioinformatics tools to use them has been delayed as bioinformatics tools are developed by bioinformaticians focused on high-level languages which, until recently, had limited support for Hadoop. Thereby delaying the adoption of distributed computing in bioinformatics (Oliphant, 2016).

The Galaksio interface itself is tailored towards enhancing user friendliness for biologists and medical researchers with limited IT-skills. The implementation of such a tool is a necessary step towards a multi-layered approach to Galaxy which allows distribution of labour not only between biologists and bioinformaticians, but also between "scripting" bioinformaticians and bioinformaticians with a strong background in computer science. Enabling researchers with the latter form of education background to provide access to more advanced computation tools by creating tools such as BioBlend (Sloggett *et al.*, 2013), CloudMan (Afgan *et al.*, 2010) and Pulsar (Afgan *et al.*, 2015) connect the Galaxy workflow management system to more powerful computation resources.

A common objection to user-friendly and automated systems such as Galaksio is the fear that automation can increase the error rate or can reduce the willingness of researchers to learn bioinformatics properly. Automation is however one of the core concepts of advanced research ever since the introduction of the automated sequencing (Smith et al., 1986). Indeed, without the automation of routine tasks even the sequencing and analysis of a single genome would be an impossible task (Ewing et al., 1998). The relevance of automation within specific research tasks is perhaps best demonstrated by the common reliance on FASTQ files, with automatically assigned phred-quality scores, rather than the more expansive sequence read format (SRF) when working with large volumes of data (Clarke et al., 2012; Van der Auwera et al., 2013). With Galaksio automation is moved from a per-tool basis to a per-workflow basis and it is therefore appropriate to not only look at the risks that a further automation of tasks can bring, but also to evaluate how the current state of automation is facilitated in bioinformatics and other IT heavy fields. As an example, in healthcare the data management is seen as a way to reduce error rates and three key factors to success have been proposed for automation to be beneficial (Nolan, 2000):

- the system should prevent errors;
- procedures must be transparent so that they may be intercepted;
- procedures should be designed to mitigate the adverse effects of errors when they are not detected and intercepted.

Current practices in research are far from optimal when considering these three criteria for automation of bioinformatics. When dealing with bioinformatics tasks beyond their expertise, biologists may prefer commercial software that provides a more comprehensive, but also expensive platform with a dependency on proprietary software (Pabinger et al., 2014; Smith, 2015b). As an alternative they may rely on outsourcing computing tasks to collaborators. Other biologists take the course of establishing their own curriculum of training as previously discussed. Some of these researchers may, over time, become proficient bioinformaticians but even in the best case scenario researchers are likely to produce a number of papers based on ad-hoc scripting with low transparency and potentially serious errors, unlikely to be caught by reviewers. In comparison, prepared workflows accessed in Galaxy or Galaksio limits the time spent on ad-hoc scripting and provide a comprehensive file history with source data and the individual steps used to generate the final results that greatly improve the reproducibility of the results (see Figure 4).



Private Workflow | imported: Galaksio use case: Mouse ChIP-seq workflow

Galaxy Workflow ' imported: Galaksio use case: Mouse ChIP-seq workflow'

Annotation: This workflow executes the tutorial https://galaxyproject.org/tutorials/chip/ until the peak calling step. The workflow requires one collection of ChIP-seq treatment datasets (input 2) and one control collection (input 1) as input.

Step

Annotation

Step 1: Input dataset collection

Control Fastq List
select at runtime

Step 2: Input dataset collection

Treatment Fastq List
select at runtime

Step 3: Subworkflow

Step 4: Subworkflow

Step 5: plotFingerprint

Sample order matters
No

Bam file
Output dataset 'Filtered BAM' from step 4,Output dataset 'Filtered BAM' from step 3

Region of the genome to limit the operation to
Empty.

Show advanced options
yes

Bin size in bases
100

Number of samples
100000

Extend reads to the given average fragment size.
No extension. The default value and most typically appropriate.

Ignore duplicates
False

Center regions with respect to the fragment length
False

Minimum mapping quality
1

Figure 4. The figure displays a report generated by Galaxy by exporting a workflow after running a ChIP-seq use case.

The downside of Galaksio is that it does not provide a natural exposure to the command line environment. However, Galaksio provides a comprehensive overview of any workflow available in the Galaxy system. If used properly Galaksio can therefore also serve as a training tool to explain theoretical concepts prior to coding exercises and function as a road map for researchers aiming to improve their skills in bioinformatics and build their own workflows step-by-step using the command line.

Conclusions

Galaksio does not replace the role of trained bioinformaticians in a research environment. It does however allow bioinformaticians to automate routine tasks and promote transparency in research as researchers with limited, or no, bioinformatics training can run best practice procedures and automatically generate the data necessary for others to evaluate their work. Such automation of routine tasks have contributed positively to the productivity and to the reduction of error rates in other information heavy fields (Horsfall, 1992; Leek and Peng, 2015; Nolan, 2000). Automation can thereby reduce the work load of expert bioinformaticians and provide them with the freedom to target more challenging tasks as well as develop a curriculum for the evaluation and training of colleagues with basic or intermediate training (Peng, 2015).

References

Afgan E, Baker D, Coraor N, Chapman B, Nekrutenko A, *et al.* (2010) Galaxy CloudMan: delivering cloud compute clusters. *BMC Bioinformatics* **11** (Suppl 12), S4. http://dx.doi.org/10.1186/1471-2105-11-S12-S4

Afgan E, Coraor N, Chilton J, Baker D, Taylor J, *et al.* (2015) Enabling cloud bursting for life sciences within Galaxy: Enabling Cloud Bursting for Life Sciences within Galaxy. *Concurr. Comput. Pract. Exp.* **27** (16), 4330–4343. http://dx.doi.org/10.1002/cpe.3536

Atwood TK, Bongcam-Rudloff E, Brazas ME, Corpas M, Gaudet P, *et al.* (2015) GOBLET: The Global Organisation for Bioinformatics Learning, Education and Training. *PLOS Comput. Biol.* **11** (4), e1004143. http://dx.doi.org/10.1371/journal.pcbi.1004143

Blankenberg D, Kuster GV, Coraor N, Ananda G, Lazarus R, *et al.* (2010) Galaxy: A Web-Based Genome Analysis Tool for Experimentalists. *Current Protocols in Molecular Biology*. John Wiley & Sons, Inc., Hoboken, NJ, USA, Hoboken, NJ, USA,

Freeberg M and Heydarian M (2016) Training Material For Chip-Seq Analysis. http://dx.doi.org/10.5281/zenodo.197100

Fuxelius H, Bongcam E, and Jaufeerally Y (2010) The contribution of the eBioKit to Bioinformatics Education in Southern Africa. *EMBnet.journal* **16** (1), 29. http://dx.doi.org/10.14806/ej.16.1.173

Goecks J, Nekrutenko A, Taylor J, and Galaxy Team T (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* **11** (8), R86. http://dx.doi.org/10.1186/gb-2010-11-8-r86

Hernández-de-Diego R, de Villiers EP, Klingström T, Gourlé H, Conesa A, *et al.* (2017) The eBioKit, a stand-alone educational platform for bioinformatics. *PLOS Comput. Biol.* **13** (9), e1005616. http://dx.doi.org/10.1371/journal.pcbi.1005616

Horsfall K (1992) The human impact of library automation University of South Australia Library,.

Klingstrom T, Mendy M, Meunier D, Berger A, Reichel J, *et al.* (2016) Supporting the development of biobanks in low and medium income countries. IEEE, pp. 1–10

Leek JT and Peng RD (2015) Opinion: Reproducible research can still be wrong: Adopting a prevention approach: Fig. 1. *Proc. Natl. Acad. Sci.* **112** (6), 1645–1646. http://dx.doi.org/10.1073/pnas.1421412111

Leipzig J (2016) A review of bioinformatic pipeline frameworks. *Brief. Bioinform.* **18** (3), 530–536. http://dx.doi.org/10.1093/bib/bbw020

Mulder NJ, Adebiyi E, Alami R, Benkahla A, Brandful J, *et al.* (2016) H3ABioNet, a sustainable pan-African bioinformatics network for human heredity and health in Africa. *Genome Res.* **26** (2), 271–277. http://dx.doi.org/10.1101/gr.196295.115

National Research Council (US) Committee on Undergraduate Biology Education to Prepare Research Scientists for the 21st Century (2003) Bio2010: Transforming Undergraduate Education for Future Research Biologists National Academies Press (US), Washington (DC),.

Nolan TW (2000) System changes to improve patient safety. *BMJ* **320** (7237), 771–773.

Oliphant T (2016) Anaconda and Hadoop --- a story of the journey and where we are now. http://technicaldiscovery.blogspot.se/2016/03/anaconda-and-hadoop-story-of-journey.html (accessed 7 April 2017).

Peng R (2015) The reproducibility crisis in science: A statistical counterattack. *Significance* **12** (3), 30–32. http://dx.doi.org/10.1111/j.1740-9713.2015.00827.x

Sloggett C, Goonasekera N, and Afgan E (2013) BioBlend: automating pipeline analyses within Galaxy and CloudMan. *Bioinformatics* **29** (13), 1685–1686. http://dx.doi.org/10.1093/bioinformatics/btt199

Smith DR (2015) Broadening the definition of a bioinformatician. *Front. Genet.* **6**, 258. http://dx.doi.org/10.3389/fgene.2015.00258

Tastan Bishop O, Adebiyi EF, Alzohairy AM, Everett D, Ghedira K, *et al.* (2015) Bioinformatics Education--Perspectives and Challenges out of Africa. *Brief. Bioinform.* **16** (2), 355–364. http://dx.doi.org/10.1093/bib/bbu022

Vincent AT and Charette SJ (2015) Who qualifies to be a bioinformatician? *Front. Genet.* **6**, 164. http://dx.doi.org/10.3389/fgene.2015.00164

Wightman B and Hark AT (2012) Integration of bioinformatics into an undergraduate biology curriculum and the impact on development of mathematical skills. *Biochem. Mol. Biol. Educ.* **40** (5), 310–319. http://dx.doi.org/10.1002/bmb.20637

Williams JJ and Teal TK (2017) A vision for collaborative training infrastructure for bioinformatics: Training infrastructure for bioinformatics. *Ann. N. Y. Acad. Sci.* **1387** (1), 54–60. http://dx.doi.org/10.1111/nyas.13207

Acknowledgements

This work was financed by the B3Africa project. B3Africa is supported by the European Union’s Horizon 2020 research and innovation programme under grant agreement No 654404.

We would also like to acknowledge the much appreciated help provided by Marius van den Beek at the Institut Curie, Paris, France and other active users in the Galaxy chat.