



Review of Data-Intensive Bioscience



Table of Contents

Executive Summary	3
Impact of Data-Intensive Bioscience	4
Background	6
Aims and scope of the review	6
Review process	7
Wider context	7
Bioscience is now a data-rich discipline with growing sophistication	7
Barriers and needs highlighted by the community	9
Key Findings and Recommendations	10
Skills	10
Professionalisation	11
Digital infrastructure	12
Software tools and resources	13
Data sharing	14
Community coordination and capacity building	15
Multidisciplinarity	16
Next Steps	16
Acknowledgements	17
Members of the Expert Group	17

Published November 2020

Cover image Getty Images

Executive Summary

The volume and complexity of biological data is growing with the ongoing development of transformative technologies such as next-generation sequencing and high-resolution imaging. Bioscience researchers are now regularly employing computationally dependent analysis and modelling approaches to process data ‘at scale’ while also increasingly benefitting from access to and re-use of data to accelerate discovery. Such work allows researchers to explore unprecedented research questions leading to major advances in frontier knowledge discovery. Additionally, these data-driven technologies are essential for addressing key challenges underpinning a healthy, prosperous and sustainable future.

Recognising the opportunities presented by the ever-increasing amount of data available to bioscience researchers, in 2019 UKRI-BBSRC initiated a review of the area to underpin our forwards strategy. Data-intensive bioscience is a recognised strength for the UK, both from a scientific and technical perspective. The review found strong evidence for the pervasiveness of data-rich approaches within contemporary bioscience research. Associated with this prevalence were a range of challenges impacting both data specialists and the wider research and innovation community. In considering the evidence gathered from our community, the Expert Group that provided advice to UKRI-BBSRC through the review process has developed seven key recommendations to support the continued expansion of bioscience as a data-intensive discipline.

- Recommendation 1: UKRI-BBSRC should take specific actions to increase the UK capacity in mathematical and computational skills within the biosciences.
- Recommendation 2: UKRI-BBSRC should catalyse the establishment of professional roles to support data-intensive research within independent research organisations.
- Recommendation 3: UKRI-BBSRC should take a leadership role in building coherent digital infrastructure provision for the biosciences.
- Recommendation 4: UKRI-BBSRC should significantly increase its investment in provision of high-quality software and data resources for the research community.

- Recommendation 5: UKRI-BBSRC should update its data sharing policy to broaden its coverage and improve its implementation.
- Recommendation 6: UKRI-BBSRC should establish a programme to build capacity in data-intensive bioscience through networking and strategic investment in key areas.
- Recommendation 7: UKRI-BBSRC should ensure its peer review processes fully embed data-intensive research as a way of working.

UKRI-BBSRC extends our sincere thanks to the research community for the inputs provided and to the Expert Group who have guided and advised throughout the review process. The recommendations provide a clear framework to guide our strategy, supported by UKRI-BBSRC Council. Our next steps will be to take forward each of these recommendations by developing an implementation plan, also exploring opportunities to work with other UKRI partners and organisations. While some of the recommendations can be addressed through policy or operational changes it is acknowledged that for others significant increases in investment will be required. The evidence from the review will therefore also be factored into UKRI-BBSRC’s longer term strategy, forming part of a case for increased investment in the biosciences more broadly.

Impact of Data-Intensive Bioscience

At the outset of this review we recognised that 'data-intensive' approaches to bioscience are increasingly pervasive, a trend that is expected to continue over the coming years. This section provides a flavour of some of the varied and exciting data-intensive research supported by UKRI-BBSRC. The highlighted projects are all the result of collaborative multidisciplinary partnerships between researchers and organisations.

Agri systems research to enhance livelihoods in developing countries

Dr Marion Pfeifer, Newcastle University

Work supported through the Global Challenges Research Fund is combining agricultural, social and ecological data to quantify how integrated landscape management can enhance benefits to and from agriculture in tropical landscapes of Sub-Saharan Africa, taking into account trade-offs with loss of land and potential increased exposure of crops to pests. The project will consolidate scattered evidence of benefits to biodiversity, soil quality and crop yield of retaining natural habitat, building on long-term partnerships with rural farmers, agribusiness, researchers and government in Tanzania. ☈

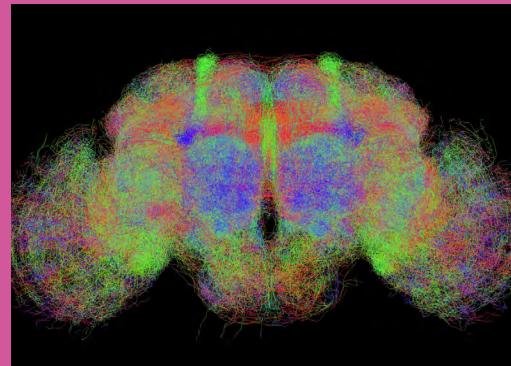


Credit: Marion Pfeifer

Understanding the complexity of the brain

Professor Daniel Coca, The University of Sheffield

The fruit fly brain observatory, a project supported by UKRI-BBSRC and the US National Science Foundation (NSF), is allowing researchers at the University of Sheffield and Columbia University to investigate the complex relationship between genes, brain structure, function and behaviour. Integration and modelling of diverse data types is emerging as a key requirement to make sense of the vast range of biological data, stimulating novel avenues of research and enabling a richer and more integrated understanding of how living systems function. ☈



Credit: Daniel Coca

Deep learning fights tuberculosis in cattle

Professor Mike Coffey, SRUC

Deep learning is a powerful technique that can provide novel insights from large scale data. A project supported via UKRI-BBSRC's Responsive Mode is analysing spectral data from millions of national milk records from thousands of herds and combining this with other bovine TB (bTB) data to accurately predict the bTB status of dairy cows, contributing to early management and diagnosis of a disease that is estimated to cost the UK dairy industry around £175 million a year. ☈



Credit: Mike Coffey

Data-driven modelling of microbiomes

Professor Orkun Soyer, University of Warwick

Anaerobic digestion is a green-energy technology that uses communities of microbes to convert organic waste into methane. UKRI-BBSRC supported research is using data-driven approaches to analyse and model the functions and dynamics of these microbiomes. By developing a better understanding of these industrial processes, researchers aim to better control them and increase their efficiency, enhancing use of resources in the bioeconomy. ☈



Credit: Getty Images

New software to exploit DNA-sequencing technology

Professor Matt Loose, The University of Nottingham

The rapid development of DNA sequencing has been underpinned by new software to allow the vast amounts of data to be analysed effectively. UKRI-BBSRC supported the development of software to analyse data from the latest 'long read' nanopore sequencing technologies. In addition, UKRI-BBSRC has supported the development of adaptive sampling, whereby individual molecules can be sequenced from a larger pool enabling scientists to quickly and easily target individual regions of larger genomes. This is helping scientists produce longer and more complete genomes, with applications across biology, environmental science, biomedicine, and beyond. ↗

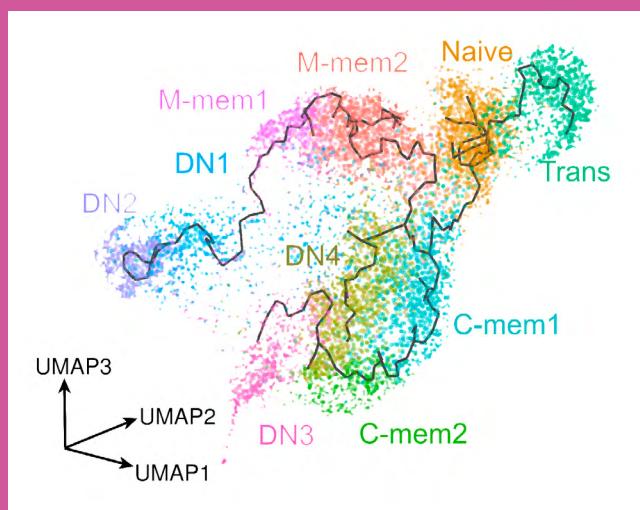
Credit: Getty Images



Mapping antibody class switch mechanisms and function

Professor Franca Fraternali, King's College London

A UKRI-BBSRC supported 'longer larger' grant is investigating the mechanisms and consequences of changing constant regions in the antibody structure, with implications for the design of new drugs or vaccines. Interdisciplinary collaboration between immunologists and computational biologists is key to tackling the large gap in our understanding of the molecular mechanisms playing a role in different B cell states and the associated antibody changes. ↗



Credit: Franca Fraternali

Sequencing the wheat genome

Professor Anthony Hall, Earlham Institute

UKRI-BBSRC supported researchers have played an important role, in collaboration with international scientists, in assembling and annotating of agronomically important wheat genomes from across the globe. Knowing the sequence of these genomes, enables us to use wheat as a model crop species and changes the way research and breeding can be done. The identification of genes and networks controlling important traits links crop researchers back to the huge knowledge base of basic plant science research. Together these substantial datasets are accelerating science discovery and giving us the tools to meet growing global demands for new higher yielding, more sustainable, disease resistant and healthier wheat varieties. ↗



Credit: Pixabay/Creative Commons CC0

Understanding the collective behaviour of genes

Professor Constanze Bonifer, University of Birmingham

In the 'post-genomic' era a major challenge is to understand how regulation of all the genes in mammals is orchestrated to generate specific cell types, which is only possible by adopting sophisticated data-driven approaches. UKRI-BBSRC supported research into the differentiation of blood cells has demonstrated how networks of transcription factors and chromatin components regulate cell fate decisions, with broad implications for both biology and biomedicine. ↗

Credit: Getty Images



Background

UKRI-BBSRC initiated the review of data-intensive bioscience in recognition of the huge growth and importance of 'data' within the biosciences, along with the opportunities this presents to undertake important and exciting research across the spectrum of science relevant to UKRI-BBSRC's mission.

Undertaking the review was timely for two reasons. Firstly, while UKRI-BBSRC has been active in supporting data-intensive research it had been some years since a more holistic, cross-cutting look at the area had been attempted. During this period the nature of the science and the wider context, such as the emergence of data science in other disciplinary areas, had developed substantially. Bioscience researchers are now regularly employing computationally dependent analysis and modelling approaches to process data 'at scale' while also increasingly benefitting from access to and re-use of data to accelerate discovery. There have also been significant community initiatives, such as the establishment of the FAIR¹ Data guiding principles to make data 'Findable Accessible Interoperable and Re-usable', which we expect the recommendations from this review to make an important contribution towards.

Secondly, UKRI-BBSRC was beginning to consider themes that would contribute to our long-term strategy and preparations for the Government Spending Review, building on the first UKRI-BBSRC Delivery Plan after the formation of UKRI. Data-intensive bioscience is an area of recognised UK strength, presenting an excellent possibility to build on both technical and scientific leadership. Greater insight into opportunities and challenges relating to data-intensive bioscience, which is consistently highlighted by our research and innovation communities as being of future importance, would help UKRI-BBSRC identify ways of helping realise the UK's full potential in this area. This would ultimately form part of a case for increased investment in the biosciences more broadly and enable researchers to make exciting new discoveries and address key challenges.

This final report sets out the key findings of the review. The major challenges emerging from the various inputs to the review have been synthesised into seven high-level recommendations, which have been endorsed by UKRI-BBSRC Council. While these

recommendations are directed to UKRI-BBSRC, we recognise they will be of broader relevance and interest. For example, universities and research institutes may utilise these insights in considering their role and strategy towards developing robust and sustainable approaches to support their researchers undertaking data-intensive science. We hope the review will be well received by the research and innovation community and will stimulate further discussions throughout the UK. We look forward to continuing our engagement with you in this area as we begin to progress plans towards implementation.

Aims and scope of the review

At the outset of the review it was recognised that data-intensive bioscience is a broad area and that it would be difficult to tightly define the research and innovation activity encompassed therein. Our high-level aims in embarking on the review were to:

- strengthen UKRI-BBSRC's understanding of the research landscape in this area
- identify current issues and future needs within the UK research community
- ensure UKRI-BBSRC is well positioned to support future requirements.

Through the review process we recognised that there were two kinds of data driven science supported by UKRI-BBSRC. The first was widespread data-rich bioscience, involving generation and analysis of large-scale data of various types. In addition, there was also a more 'data-intensive' form of research, with a stronger emphasis on more sophisticated analysis approaches owing to the scale and complexity of the data used and questions being addressed. Hard delineation between the two is difficult given the broad diversity of bioscience in this area but nevertheless, we established the following working definition for data-intensive bioscience to help guide the review:

Research: Bioscience involving computationally intensive analysis of large scale and/or complex data. In particular:

- Research that is primarily focused on generating new biological insight through large-scale data analysis and integration
- Innovation in data-intensive bioscience through the development of new approaches, methods and software

¹ <https://www.force11.org/group/fairgroup/fairprinciples>

Skills: The specific skill sets (computational/mathematical/statistical/curatorial) required to undertake research involving complex datasets.

Resources: The databases, software and computational infrastructure needed to support data processing, analysis, modelling and sharing.

Review process

Given the breadth and complexity of data-intensive bioscience it was decided the review should take a consultative approach with the bioscience research and innovation community, rather than attempt to fully 'map' the UK landscape. The area of infrastructure had also recently been examined extensively as part of a broader UKRI programme² and we did not seek to duplicate that work. The review thus drew on three key lines of evidence:

- analysis of UKRI-BBSRC's funding portfolio and existing strategic approaches to the area
- a community questionnaire run between April and July 2019 as well as some targeted follow-on engagement
- a workshop held on 18 September 2019 in Manchester to discuss key issues in greater depth with members of the bioscience research community.

In total, 164 inputs from individuals or groups were received as part of the review process, covering 64 research organisations.

A small Expert Group was established to steer the review and provide advice to UKRI-BBSRC (see end of document for membership). The Group held four meetings to develop the consultation approaches, synthesise the inputs received, and agree the key recommendations and final report.

Wider context

The review sits in the wider context of the expansion of data science and the open science agenda. The Expert Group was conscious of several related initiatives and activities, such as the work of the BEIS Open Research Data Taskforce³. The review findings are aligned to the goals of these initiatives and the recommendations stemming from the review should be considered as part of UKRI-BBSRC's response.

Bioscience is now a data-rich discipline with growing sophistication

Bioscience has clearly undergone a seismic shift in the scale and complexity of data being exploited by the research and innovation community in the past decade. Analysis of UKRI-BBSRC's funding portfolio has revealed that at least 50% of research grants now involve large-scale biological data and can be considered 'data-rich'. This encompasses an increasingly diverse range of sequencing, omics, imaging, and other data producing platforms that require robust mathematical and computational analysis to extract biological knowledge. Within Responsive Mode, data-rich research has more than doubled in the last ten years, to over 50% of applications (Figure 1).

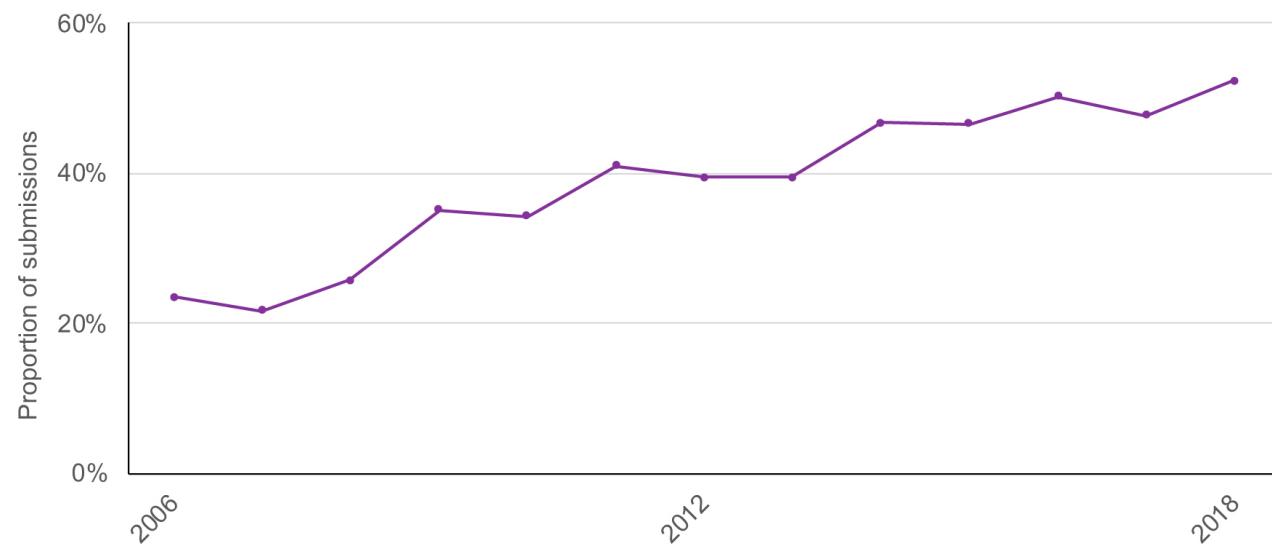


Figure 1: Trend in 'data-rich' applications submitted to Responsive Mode based on analysis of project summaries

2 <https://www.ukri.org/research/infrastructure/>

3 <https://www.gov.uk/government/publications/open-research-data-task-force-final-report>

Its reach has also extended, with data-rich research now being well represented across all four Responsive Mode Research Committees.

While it is now relatively straightforward to generate large-scale data, the level of expertise and sophistication in the analysis of biological data varies significantly within the research community. Analysis of Responsive Mode submissions found ~10% evidenced use of advanced approaches such as development of new methods and algorithms, use of artificial intelligence and deep learning, data integration and mathematical modelling approaches, though this is difficult to measure precisely.

As part of the review the members of the research and innovation community were asked to comment on how they saw data-intensive research in biology

changing over the next five years, and what new and exciting science might be enabled; some of the key themes are illustrated in Figure 2. In the consultation there were consistent messages about:

- the expanding overall volume and diversity of data, with data analysis considered to be a key bottleneck for bioscience in the coming years
 - current and emerging research technologies such as artificial intelligence, phenotyping, spatial genomics, laboratory automation, cryo-EM and *in vivo* imaging presenting new data challenges
 - multi-dimensional/ multi-scale approaches, with respondents anticipating use of a broad range of data types that would need to be integrated in order to derive quantitative insights into dynamic biological phenomena.



Figure 2: Word cloud illustrating key research themes from the consultation

Many respondents considered that data-intensive analyses would increasingly be used to guide the development of research questions and experiments. It would help researchers elaborate hypotheses and avenues of investigation that would not otherwise be apparent through qualitative insights and traditional approaches to hypothesis generation.

The research and innovation community envisage greater use of sophisticated approaches in the coming years as they seek to deal with the data deluge and maximise the insight that can be obtained from their data to answer increasingly complex questions. These new types of data analysis methods will bring insights that are conceptually very different to what most biologists are used to. They will require people to learn, understand, and appreciate systems-level concepts such as robustness, noise/heterogeneity, and signal processing functions.

Barriers and needs highlighted by the community

The review highlighted a range of issues that represent current and future barriers to progress. In the questionnaire, responses to the question 'what are the biggest issues that need to be addressed over the next five years?' (Figure 3) showed that infrastructure, software, training and staff were the most consistently cited challenges. Responses to the follow up question 'what should the priority areas of action for UKRI-BBSRC be?' (Figure 4) particularly highlighted training and infrastructure as well as software and the potential to stimulate the area with funding initiatives. They also registered peer review and policy as areas for improvement.

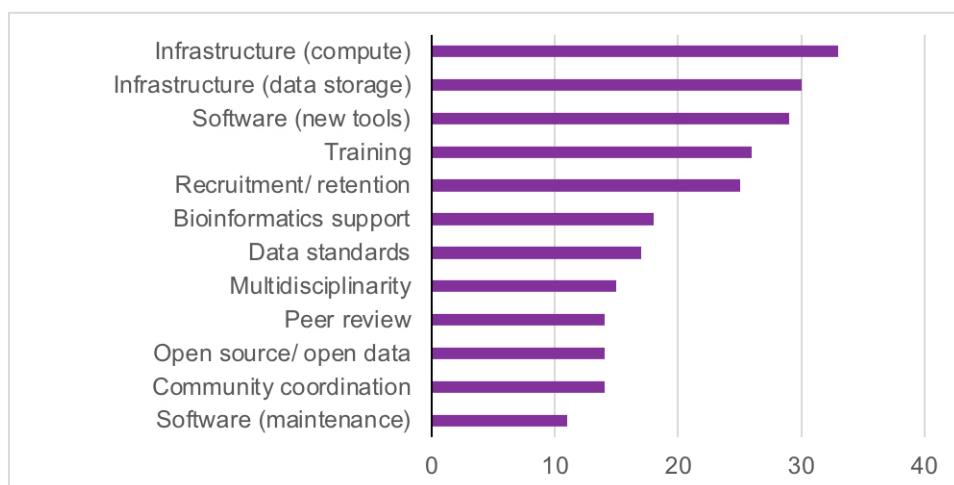


Figure 3: Top 12 issues highlighted in questionnaire responses

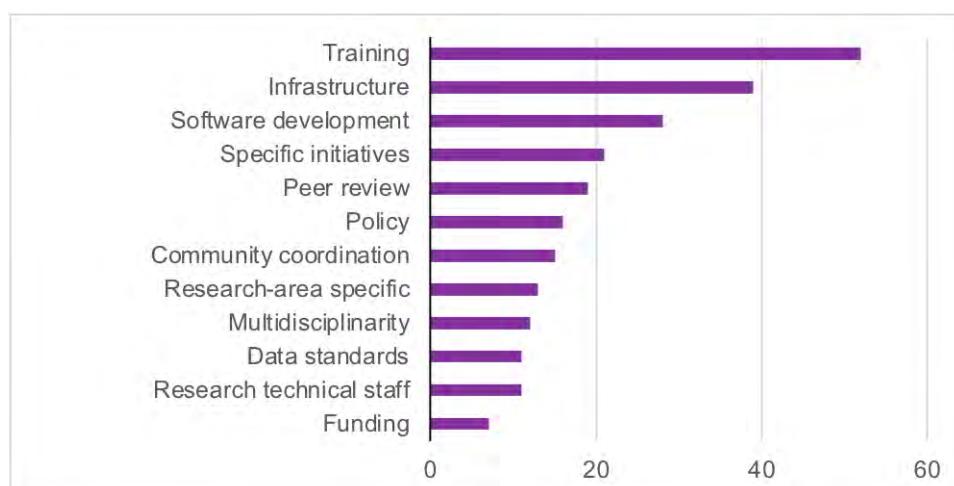


Figure 4: Top 12 areas suggested for UKRI-BBSRC action

Key Findings and Recommendations

The following sections summarise the key findings of the review, presented under seven thematic areas along with key recommendations to UKRI-BBSRC developed by the Expert Group that provided expert advice and steered the review process.

Skills

The findings of the review add to the clear evidence of high demand for researchers with quantitative skills in bioscience. 'Bio-informatician' continues to be listed on the Home Office shortage occupation list⁴. The need for computational and mathematical skills is a focus across sectors, as evidenced by the work of groups and organisations such as the Digital Skills Taskforce⁵, the Office for Artificial Intelligence⁶, and the OECD⁷.

The review highlighted significant ongoing demand for training, through a variety of mechanisms that reflect the range of skills required under the broad umbrella of data-intensive research. Skills in data science are no longer limited to specialists but a core requirement for the whole bioscience community. Scale of training is a key factor; there is evidence of good initiatives in the community but there is a need to train many more people, across all career stages. The time required to acquire proficiency is also a significant factor that needs to be considered and factored into future activities in this area.

Within academia the recruitment and retention of staff with advanced quantitative skills is a significant issue. It is not possible to match salaries offered by big industry players. Greater flexibility and interchange between academia and industry, particularly with SMEs, could facilitate skills development. To aid both recruitment and retention there is a need to strengthen career pathways and professionalise careers for individuals with important technical skills such as software development and data management. Such career models are currently not commonplace within research organisations such as universities and research institutes, though there are successful examples within individual organisations and from initiatives such as UKRI-EPSRC's Research Software Engineer pools.

The Expert Group highlighted the accessibility of existing skills as another key issue, with the need to have a UKRI-BBSRC funding system that is supportive of flexible approaches to share or bring in relevant expertise (e.g. industry consultancy, researcher placements and interchange).

In particular, the important role of technical professionals, who work within a research group or in various types of shared facilities, needs to be recognised as they work across a lot of research groups and are often involved in training researchers.

There is a strong case for expanded UKRI-BBSRC support for data-intensive research skills. The Expert Group recommended using several mechanisms together:

- Doctoral Training Partnership and Centre for Doctoral Training models, providing both broad development of data-intensive bioscience related skills and in-depth specialisation
- programmes like UKRI-BBSRC's Strategic Training Awards for Research Skills⁸ scheme (STARS) that enable community focused training in specific areas
- short term researcher placements and interchange, which would enable specialist skills transfer at different career stages and provide adequate time for individuals to acquire relevant skills and build professional networks
- support for training of and by 'core' technical professionals (flexibly defined).

UKRI-BBSRC should also do more to recognise the skills shortage through its grant application guidance and assessment processes, including:

- stronger signalling to include and support upskilling costs in proposals
- supporting some appointments above the standard salary points, to recognise the competitive market for staff with strong quantitative expertise.

Recommendation 1: UKRI-BBSRC should take specific actions to increase the UK capacity in mathematical and computational skills within the biosciences.

4 <https://www.gov.uk/guidance/immigration-rules/immigration-rules-appendix-k-shortage-occupation-list>

5 <http://www.ukdigitalskills.com/>

6 <https://www.gov.uk/government/organisations/office-for-artificial-intelligence>

7 [http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=DSTI/STP/GSF\(2020\)6/FINAL&docLanguage=En](http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=DSTI/STP/GSF(2020)6/FINAL&docLanguage=En)

8 <https://webarchive.nationalarchives.gov.uk/20200302112440/https://bbsrc.ukri.org/funding/filter/stars/>

Professionalisation

Responses to the review consultation and discussions at the workshop point towards growing challenges relating to the ‘four Vs - volume, velocity, variety and veracity - of data’ in bioscience research. Consequently, there is an increasing need for professional research software engineers, bioinformaticians, data scientists, data curators and data stewards to support delivery of high-quality bioscience alongside skilled researchers. Data stewards, for example, can provide expert advice on good practice and can play a key role at the start and end of projects in ensuring high quality data management and sharing, in line with agreed community standards and data sharing policies. Some institutions have begun to address this need by establishing dedicated posts, but these institutional approaches and career pathways are not yet established consistently across the research landscape. Apprenticeship models might be suited to help develop the skills base in some of these areas but appear relatively untapped as a mechanism.

The Expert Group considered that UKRI-BBSRC should encourage research organisations to develop relevant professional support for data analysis, management and sharing. A variety of funding mechanisms should be explored to achieve this; for example, learning from initiatives run by UKRI-EPSRC for research software engineers to pilot and raise the professional profile of such positions within independent research organisations.

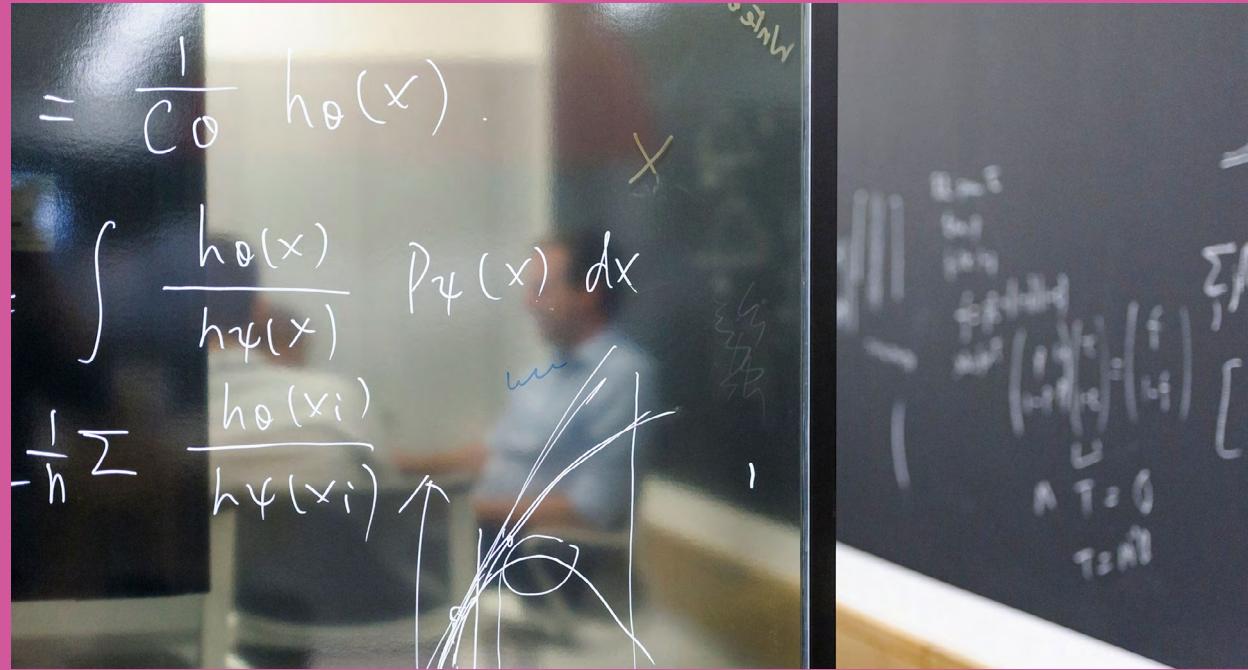
The Expert Group advised that UKRI-BBSRC should ensure it provides clear guidance to applicants and committee members regarding inclusion of professional support for data-intensive bioscience on grant applications. The inclusion of full or partial FTE roles associated with data analysis and management needs to be accepted as part of modern bioscience, in contrast to more common single-PDRA projects.

Recommendation 2: UKRI-BBSRC should catalyse the establishment of professional roles to support data-intensive research within independent research organisations.

Role of professional support

Organisations are recognising the need for professional support and training to help scientists deliver open and reproducible tools and research practices; for example, the research engineering team at the **Alan Turing Institute**  collaborates and contributes their expertise across the Institute’s programmes. The UKRI-supported **Software Sustainability Institute**  is raising the profile of the technical professionals who play a crucial role in delivering better and more sustainable software, while also acting as a focal point for sharing best practices between disciplines and supporting these professionals deliver digital skills training to researchers.

Credit: Alan Turing Institute



Digital infrastructure

Digital infrastructure was highlighted as an important area to underpin support for data-intensive disciplines. Scientific computing provision in bioscience is less developed in comparison to the physical sciences both locally (mid-range and central facilities within independent research organisations) and at a regional/national level. In both instances, traditional high-performance computing suited to simulation rather than data-intensive applications tends to dominate discussions of computing provision. Alongside computing capacity, data storage and associated software infrastructure to enable researchers to make data Findable, Accessible, Interoperable and Re-usable (FAIR⁹) was identified as an important and growing need. The advent of cloud computing presents new opportunities but also complex issues (e.g. procurement, sustainability) that the bioscience community has yet to fully work through but would benefit from further investigation and knowledge sharing.

The bioscience community needs greater engagement and coordination in planning for scientific computing and data infrastructures. A significant challenge is being able to clearly articulate the requirements of bioscience users both locally as well as on a regional and national level. Community capability and coordination both vary widely across the biosciences, with many users lacking the in-depth technical expertise and the professional networks to influence the provision of digital infrastructure resources.

The Expert Group identified perceived barriers to applying for computing equipment and costing for computing and data storage resources in UKRI-BBSRC grant applications. Data storage is of particular importance given the role of data in underpinning the scientific record and UKRI-BBSRC's requirements to retain data for a ten-year period. Preservation of data for re-use by the research and innovation community ultimately enables cutting-edge bioscience but there is increasing awareness that these benefits must be balanced with practical considerations and the environmental impacts of data storage. The Expert Group considered that UKRI-BBSRC should develop clearer guidance on how applicants can cost computing and data storage within grant applications, particularly considering the rapid changes in cloud computing.

Noting the recent work of the UKRI infrastructure project¹⁰, it was considered that digital infrastructure required further action to ensure it does not

Investing in infrastructure

Digital infrastructure plays a key role in ensuring scientists have the data they need to make discoveries. A £45 million boost to data and building infrastructure from UKRI's Strategic Priorities Fund is supporting academic and industrial demand for open access to biological data at one of the world's largest centres, **EMBL-EBI**. The investment will support intensifying growth in data resources driven by new technologies such as single cell sequencing and cryo-electron microscopy, and help ensure data are FAIR (Findable, Accessible, Interoperable and Re-usable) for users globally. ☘



constrain data-intensive bioscience. In addition to investment, UKRI-BBSRC should bring together the research community into a more coherent but flexible 'ecosystem', including facility managers, HEI computing and other service providers, to better understand needs, issues and potential solutions that may have already been developed in some institutions and networks. An initial step forward would be to hold a meeting on scientific computing for the bio/life sciences, linking to a longer-term programme of coordination and networking (see Community, below), to help ensure the requirements of the bioscience community are well understood and clearly articulated in discussions on digital infrastructure provision, both within independent research organisations and regionally/nationally.

Recommendation 3: UKRI-BBSRC should take a leadership role in building coherent digital infrastructure provision for the biosciences.

9 FAIR Data principles: Findable, Accessible, Interoperable, Re-usable
<https://www.force11.org/group/fairgroup/fairprinciples>

10 <https://www.ukri.org/research/infrastructure/>

Software tools and resources

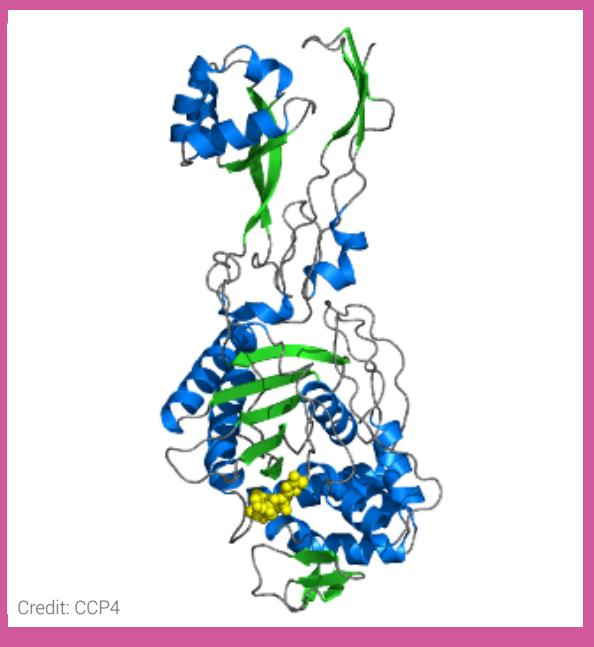
Software is integral to data-intensive bioscience and UKRI-BBSRC has had dedicated funding routes to support software and data resources for over ten years (Tools and Resources Development Fund (TRDF)¹¹ and Bioinformatics and Biological Resources Fund (BBR)¹²). Responses to the questionnaire and discussions at the workshop found these to be valued and important mechanisms. However, there was significant concern that these mechanisms have not kept pace with the substantial expansion of the field, so their overall scale is too small.

A further issue was that the structure of the TRDF and BBR Fund does not support a sufficiently broad range of activities required in this area. Examples of key activities include development of new methods

and software tools, maturation of experimental research software to a robust and usable tool for use by the wider research community, development and maintenance of databases from both a technical and data content/data curation perspective, and support for other forms of cyberinfrastructure (e.g. data standards, tools to facilitate workflows, gateway resources). Different specialist skillsets are required for these activities (e.g. in programming vs data curation) and different drivers underpin them. A limitation of the current model is that software and resources can often be under-developed from the perspective of the end user. The Expert Group concluded that UKRI-BBSRC's funding approach could be better designed to fully cover and incentivise the range of activities needed to support high quality software tools, data resources and cyberinfrastructure.

Long-term support for software

Collaborative Computational Project 4 (CCP4) is one of the leading sources of software and facilitator of research in computational methods for macromolecular crystallography and other biophysical techniques. Sustained investment over several decades has enabled CCP4 to support academic and commercial users worldwide, underpinning leading edge discovery research on protein structure and important sectors such as agri-tech and pharmaceutical R&D. As a community-based resource, CCP4 also plays a key role in education and training. ↗



Credit: CCP4

A significant increase in the current level of resources dedicated to this area is needed to meet both current and expanding future needs. The Expert Group recommended UKRI-BBSRC also re-examine its modes of support for the development and maintenance of software and resources.

UKRI-BBSRC should carefully design a funding process that accommodates the whole range and pathway of activities, recognising the differing nature of work needed at each stage. Supporting this pathway would help ensure that software and data resources are high quality, open, robust and reusable, underpinning research reproducibility.

The Expert Group noted that a careful balance needed to be struck between supporting innovative new approaches and investing in maintaining existing tools and resources. Use of follow-on funding models could help promote a smoother development pipeline, recognising that resources for the further development and maintenance of tools and resources can equal or exceed the original cost of early development work. Such models could also help retain skilled staff in a highly competitive environment and engage specialist technical expertise in certain areas (e.g. software engineers, sub-contractors) at the right stage of development.

Recommendation 4: UKRI-BBSRC should significantly increase its investment in provision of high-quality software and data resources for the research community.

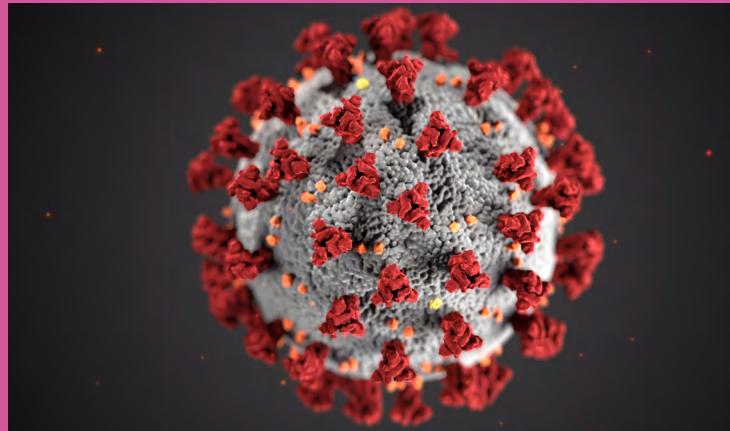
11 Tools and Resources Development Fund
<https://webarchive.nationalarchives.gov.uk/20200302110441/><https://bbsrc.ukri.org/funding/filter/2019-tools-resources-development-fund/>

12 Bioinformatics and Biological Resources Fund
https://webarchive.nationalarchives.gov.uk/20200302113839tf_<https://bbsrc.ukri.org/funding/filter/2019-bioinformatics-biological-resources-fund/>

Data sharing accelerates fight against COVID-19

Photo by CDC on Unsplash

Rapid and open sharing of research data has been essential to the global response to COVID-19. UKRI committed researchers it funds to sharing of findings, data and software arising from research into the coronavirus pandemic, in order to accelerate the effort to find effective solutions and treatments. Resources such as the **COVID-19 data portal** are playing a key role in facilitating data sharing globally.



Data sharing

UKRI-BBSRC has had a data sharing policy¹³ since 2006, recognising the broad scientific, economic and societal benefits that the availability of research data can bring, as well as the importance of data management in underpinning the reproducibility of UKRI-BBSRC funded research. The community consultation highlighted that the proportion of bioscience projects involving data integration and data re-use will increase in the coming years and that more needs to be done to enable this. Achieving this goal will require engagement with international initiatives to ensure effective integration and standardisation across the global bioscience community; for example, in provision of data sharing resources and the development of data standards and ontologies.

Data sharing is a key area in which UKRI-BBSRC needs to continue to provide discipline-focused leadership as a part of UKRI, and work to encourage and incentivise good practice within its community. Discussions at the community workshop highlighted a need to re-examine UKRI-BBSRC's data sharing policy and its implementation. The policy should be updated in several areas to better embed and reward good data sharing practices following the FAIR Data principles. The Expert Group agreed that the policy should be expanded to include other types of digital objects relating to reproducibility of data-intensive bioscience research; for example, protocols, data processing scripts and workflows. This revised policy will help underpin the current and future data-intensive research landscape.

Data Management Plans (DMPs) offer a mechanism to enhance data sharing practices, to emphasise the need to develop and support relevant digital skills (Recommendations 1 and 2) and to encourage planning for data storage infrastructure (Recommendations 3 and 4). A template for DMPs could help improve the quality of information provided, alongside further guidance and examples on what constitutes a 'good' DMP. Guidance on which data to share should be revised to support greater utility for data-intensive bioscience along with broader benefits, balanced with the financial and environmental impact of curation and storage. Enabling methods such as online and machine-readable DMPs should also be explored. Finally, there are apparent gaps in local knowledge within research organisations about appropriate sharing mechanisms and relevant community-focused databases in some areas, which present barriers to effective data sharing that need to be addressed.

Within the data sharing policy and community practices, greater emphasis should be placed on how data will be made available and re-usable by others. The Expert Group proposed that applicants' track record in data sharing should be explicitly requested by UKRI-BBSRC and considered as part of the grant assessment process, to underline its importance to the research process and incentivise good practice. Dipstick monitoring might also be explored as a mechanism both to monitor compliance and to understand the issues experienced by grant holders in making their research data available to others.

Recommendation 5: UKRI-BBSRC should update its data sharing policy to broaden its coverage and improve its implementation.

13 <https://bbsrc.ukri.org/about/policies-standards/data-sharing-policy/>

Community coordination and capacity building

It is clear from the inputs to the review that there are common challenges within the research base, where potential opportunities for enhanced community networking can be identified. In particular emerging research areas where there are methodological challenges, areas where there is a need to develop common community practices, and where skills issues cut across particular research domains, community coordination could provide much needed support and cohesion. UKRI-BBSRC's approaches to developing such networks have to date been relatively ad hoc and responsive to community need (recent examples include BioimagingUK¹⁴ and the UK plant phenotyping network¹⁵). However, the Expert Group considered that a more active approach to identifying potential areas could be useful.

An example of an area with a disconnect between acknowledged need for community action and willingness or mechanisms to take this forward is data standards, resulting in standards often lagging behind community need. Reasons for this disconnect include the absence of leadership in standards development, particularly in less mature areas, and lack of subsequent standard ownership and uptake by the relevant community. However, it is accepted that appropriate standards often require

a period of time to be agreed and established. The role of data infrastructures in supporting implementation and adoption is also important, and efforts in this area need to join up with relevant international activity.

UKRI-BBSRC should consider opportunities to build community-based capacity in further areas relevant to data-intensive bioscience, e.g. to build and support relevant communities in emerging areas, to address cross-cutting issues, and increase coordination and knowledge exchange. This should involve active engagement with the research community to identify areas that would benefit from this type of approach through regular horizon scanning exercises, the Bioscience Big Ideas Pipeline, and dialogue with other organisations. UKRI-BBSRC should also continue to engage with international funding agencies and support other organisations, such as ELIXIR¹⁶, to facilitate international community activities where these will add value.

Recommendation 6: UKRI-BBSRC should establish a programme to build capacity in data-intensive bioscience through networking and strategic investment in key areas.

Transnational community coordination

The **UK ELIXIR Node** brings together 18 UK organisations to coordinate and provide training and services so that life sciences researchers can more easily discover, distribute, analyse, and store data, as well as exchange expertise and agree on standard approaches. It is a part of the broader ELIXIR distributed infrastructure for life sciences information, which aims to coordinate and develop vital bioinformatics resources such as databases and portals, toolkits, software, training materials and computing across Europe. ☺



14 <https://www.rms.org.uk/network-collaborate/bioimaginguk-network.html>

15 <https://www.phenomuk.net/>

16 <https://elixir-europe.org/>

Multidisciplinarity

A consequence of the increasing prevalence of data-rich bioscience research is that the proportion of research effort on grants dedicated to more sophisticated analysis of large-scale data will continue to increase. Peer review in this area is a clear concern for the research community, owing to issues surrounding multidisciplinary proposals and assessment panel composition. A further issue reported was that data-driven bioscience research can be perceived differently compared to more mechanistic studies owing to the nature of the research approach, with the former often not framed by a single hypothesis or question from the outset.

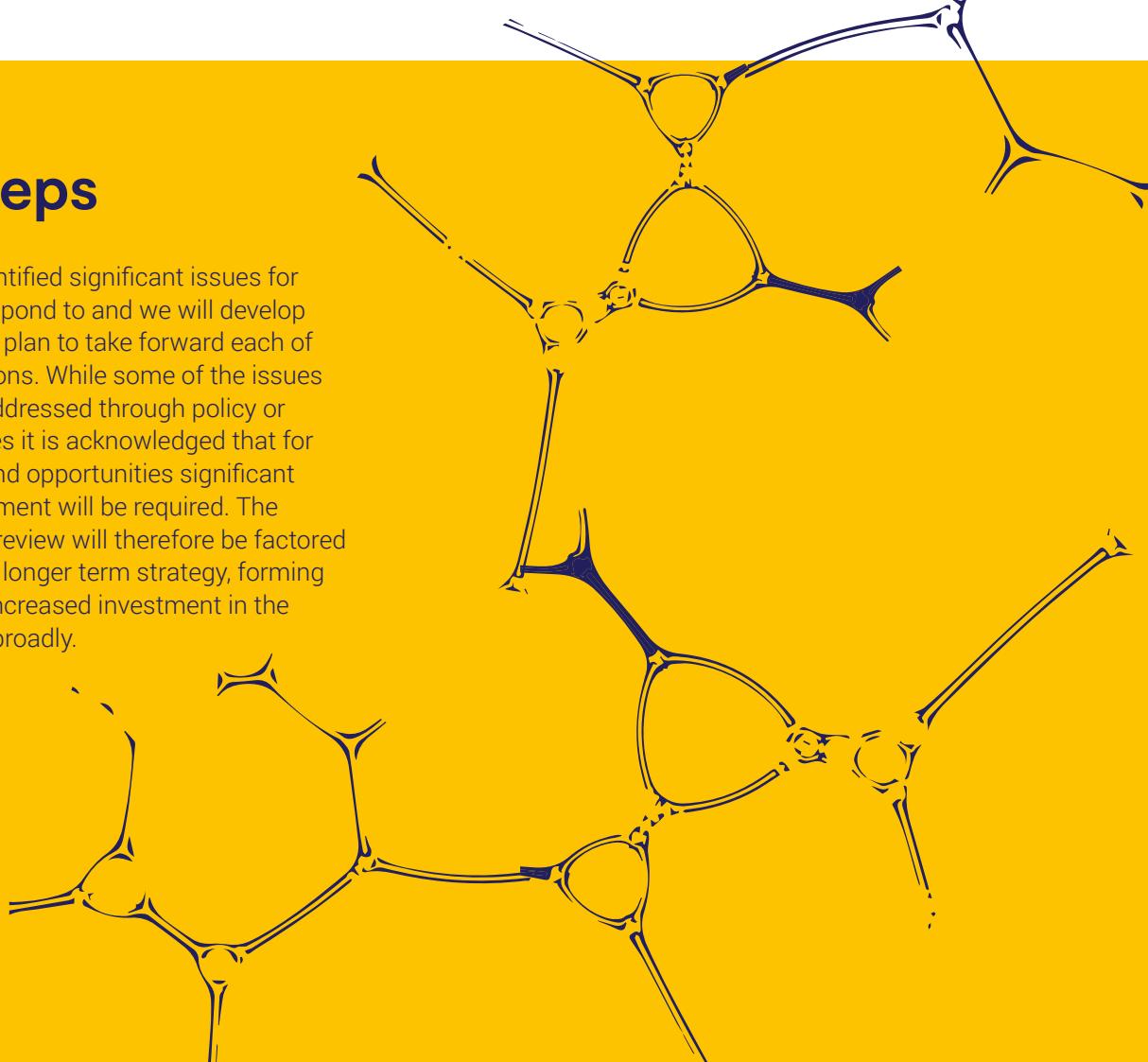
UKRI-BBSRC should examine its current peer review approaches and make appropriate adjustments to ensure scientific excellence in data-intensive bioscience is appreciated and valued throughout its peer review system. It is important that both assessment criteria and peer review guidance recognise that data-driven research and the

development of associated methods, tools and resources are an integral part of contemporary bioscience. UKRI-BBSRC should also ensure that reviewers and panel members have appropriate expertise in quantitative approaches alongside relevant knowledge of the bioscience area under investigation. This should be regularly reviewed as the field further develops. UKRI-BBSRC should also work with UKRI-EPSRC to ensure that research at the interface between the two Councils is well supported.

Recommendation 7: UKRI-BBSRC should ensure its peer review processes fully embed data-intensive research as a way of working.

Next Steps

The review has identified significant issues for UKRI-BBSRC to respond to and we will develop an implementation plan to take forward each of the recommendations. While some of the issues identified can be addressed through policy or operational changes it is acknowledged that for other challenges and opportunities significant increases in investment will be required. The evidence from the review will therefore be factored into UKRI-BBSRC's longer term strategy, forming part of a case for increased investment in the biosciences more broadly.



Acknowledgements

UKRI-BBSRC gratefully acknowledges the substantial and constructive input from the members of the Expert Group. Additional advice was provided by UKRI-BBSRC advisory groups, particularly the UKRI-BBSRC Council and the Exploiting New Ways of Working Strategy Advisory Panel. A particular thank you goes to Mr Neil Chue Hong (University of Edinburgh), for his engagement.

We are indebted to all members of the research and innovation community who responded to the community questionnaire, participated in the community workshop or provided other inputs to the review. Information collected via these routes represents a substantial part of the data upon which the recommendations in this report are built.

Members of the Expert Group

Professor Andrew Millar

University of Edinburgh
UKRI-BBSRC Council member and
Chair of the Expert Group

Professor Charlotte Deane

University of Oxford

Professor Katherine Denby

University of York

Professor Andrew French

The University of Nottingham

Professor Christine Orengo

University College London

Dr Helen Parkinson

European Bioinformatics Institute – EMBL-EBI

Professor Magnus Rattray

The University of Manchester

Dr Malcolm Skingle

GlaxoSmithKline
UKRI-BBSRC Council member

Professor Eileen Wall

SRUC, Scotland's Rural College



Biotechnology and
Biological Sciences
Research Council

Polaris House
North Star Avenue
Swindon
Wiltshire
SN2 1UH

www.bbsrc.ukri.org

