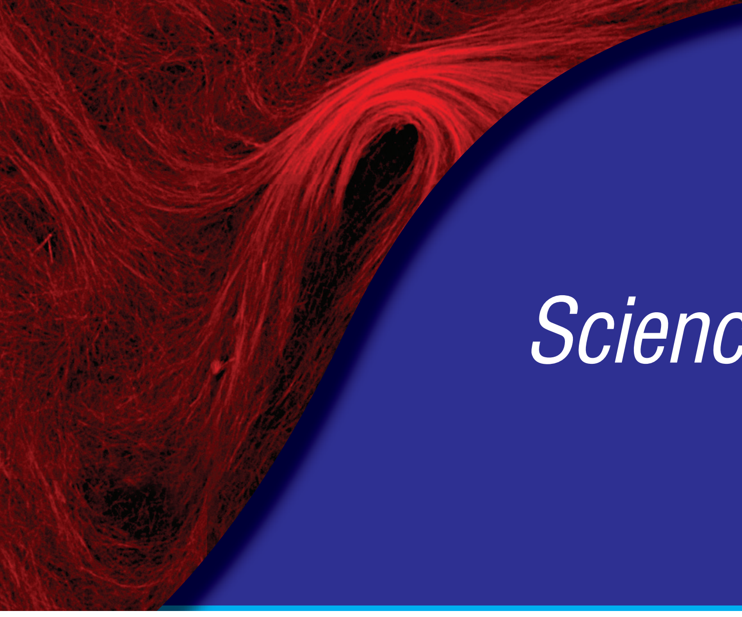


Science Community Focus

# Tackling Big Data Challenges in Bioscience

To help organisations and businesses benefit from the ever-increasing availability of large and complex datasets the UK's Biotechnology and Biological Sciences Research Council (BBSRC) has invested £7.5 million in new infrastructure. This has included investment in UK-based organisations to tackle bioscience big data challenges.  During the AAAS 2015 Annual Meeting (Feb 13), BBSRC announced that The Genome Analysis Centre (TGAC) with partner Institutes had been awarded £6 million for three joint projects;

1) Big Data Infrastructure for Crop Genomics

2) iPlant UK – creating a cyber-infrastructure for plant sciences

3) Establishing the infrastructure for functional annotation of farmed animal genomes

The University of Dundee will also lead on a 4th project to develop a repository for scientific image data that can be accessed via open source platforms

The new funding will improve the storage and curation of enormous datasets that will unlock untold discoveries in important areas like health, agriculture and sustainable fuels. Biological discovery is increasingly being driven by ground-breaking technologies, such as high-throughput genomic analysis and next generation biological imaging, which generate massive amounts of data. The three projects being led by TGAC are:

*Dr Sarah Ayling, Crop Genomics and Diversity Group Leader, TGAC*

### Big Data Infrastructure for Crop Genomics

Led by Crop Genomics and Diversity Group Leader at TGAC Dr Sarah Ayling, with EMBL-EBI, this project has been awarded £2 million to develop an open-source platform to enable users to access genetic and characteristic variation data and to perform analyses.

Recent advances in sequencing technologies and computational tools have made it possible to sequence the genetic information of some of the world's most important crop species, such as rice, barley, rapeseed, maize, soya and wheat. These crops constitute a substantial part of the daily food intake for most of the population of the world and any improvements in the breeding for more efficient and nutritious varieties will have a direct impact on ensuring global food security.

The platform will be developed using open source principles and publicly available data. This novel platform for crop bioinformatics will promote opportunities for collaborative work with R&D groups in industry, research and academia. The availability of data generated by publicly funded resources, and the concomitant development of new, production-quality tools will lower the barriers to information-enabled crop science, stimulating new opportunities for research and application. The platform will also open up new opportunities for the UK bioinformatics community, traditionally focused on biomedical applications, by developing alternative career paths around biotechnology and agri-food.

Dr Ayling said: "Producing enough food to feed the world's growing population under changing climatic conditions is an enormous challenge. The development of this crop bioinformatics platform will support the use of genomics technologies to explore genetic diversity for crop species and help to speed up the breeding process, producing more sustainable crops sooner."

*Dr Federica Di Palmer, Director of Science/ Head of Vertebrates and Health Genomics, TGAC*

### Establishing the Infrastructure for Functional Annotation of Farmed Animal Genomes

The project, co-led by Director of Science/Head of Vertebrates and Health Genomics at TGAC Dr Federica Di Palma, with The Roslin Institute at the University of Edinburgh and EMBL-EBI, has been awarded £1.9 million to develop an infrastructure to deliver reference genomes to enable research into economically important animals.

Our knowledge of the functional elements and in particular of the regulatory sequences within these animal genomes is limited. Identifying the functional elements within the genome and the consequence of variation in these functional sequences is essential. This funding will establish hardware and compute capacity at TGAC, The Roslin Institute, and EMBL-EBI, together with software, to enable the functional annotation of animal genomes.

The BBSRC grant will provide key infrastructure for the three partner Institutes in the recently launched international collaborative Functional Annotation of Animal Genomes (FAANG) initiative. The FAANG initiative is concerned with addressing the need for high quality annotated genomes as key sources of information and is critical for contemporary research in the biological sciences. It is valuable not only to academic researchers, but also to scientists working in animal breeding, animal health and pharmaceutical industries. This project is concerned with the infrastructure for delivering high quality annotated reference genomes to enable research on economically important animals.

Dr Di Palma said: "High-quality, annotated genomes are essential for the research communities to develop the sophisticated molecular biology tools necessary to facilitate research studies in these economically important models. Farm animal genomic resources will not only facilitate research in basic animal biology, but will also aid developments in the animal health industries including animal breeding, food, and sustainable agriculture."

*Dr Robert Davey, Data Infrastructure and Algorithms Group Leader, TGAC*

### Collaborative Bioinformatics UK Infrastructure for Data-Intensive Plant Science

Co-led by Data Infrastructure and Algorithms Group Leader Dr Robert Davey and Head of Scientific Computing at TGAC Dr Tim Stitt, with University of Warwick, University of Liverpool, University of Nottingham, University of Arizona and the Texas Advanced Computing Centre, this project has been awarded £1.78 million to establish a UK iPlant node that will connect the UK with the US's cyberinfrastructure for the plant sciences. TGAC's National Capability in Genomics will support the computational infrastructure.

Plant science research generates huge volumes of data containing untold discoveries, which could help tackle

Dr Tim Stitt, Head of Scientific Computing, TGAC

global challenges in medicine, biofuels, biodiversity and agriculture. A current bottleneck to these discoveries is a lack of capacity to share enormous data files and analyse them in an efficient, user-friendly way.

The iPlant Collaborative is a virtual organisation funded by the US National Science Foundation (NSF) to create cyberinfrastructure for the plant sciences. Harnessing the power of some of the world's fastest supercomputers, iPlant provides huge cloud-based storage space and a virtual lab bench, which put global plant science data and online tools in one place. Users can share datasets and tools to analyse data with as many or as few people as they wish. Tools to analyse data developed by iPlant staff, or built by others, can be shared with the wider community in a similar manner to smartphone 'apps'.

The iPlant Collaborative is currently distributed across three US locations and in less than 10 years has amassed over 18,500 users. The BBSRC funding will extend this into an international collaboration by building a UK iPlant node at TGAC in Norwich, which provides National Capability for computational infrastructure. Software tools developed for specific plant science sequencing, systems biology and image analysis projects at the Universities of Warwick, Liverpool and Nottingham will be adapted by a dedicated team of programmers so that they can be integrated into iPlant UK. These will then be made freely and openly available for the wider plant science community to use.

Dr Davey said: "The deployment of the iPlant platform at TGAC, in conjunction with the expansion of National Capability hardware and collaboration with the iPlant US team, will provide a long-term data management, analysis and sharing hub for the UK plant science community. Infrastructure and training that empower researchers through robust, efficient and intuitive tools are vital for the UK to continue its advances into understanding and addressing key scientific challenges."

"The recent advances in data-generation technologies have opened up new opportunities for innovation in life sciences. Receiving this funding for three big data projects is a great opportunity for TGAC and places us at the forefront of new science. These projects will support our world-class Institute to continue its response to the increasing demand for more sophisticated computing platforms and algorithms for data analysis in the plant and animal genomics communities," said Mario Caccamo, Director of TGAC.

## Dundee leads `big data' challenge

*Professor Jason Swedlow, Open Microscopy Environment, College of Life Sciences, University of Dundee*

Of the £7.5 million BBSRC funding, £1.79 million has been allocated to a project led by Professor Jason Swedlow, University of Dundee, to build a next generation image repository with open access for scientists and the public around the world.

Imaging in the life sciences is used to understand the behaviour of organisms, the formation of embryos, the structure and dynamics of cells, and the function and interactions of molecules that are the building blocks of life. However, imaging datasets are complex, diverse in character or content, and often extremely large. This means that they are rarely shared or published.

Professor Swedlow and his Open Microscopy Environment (OME) team in the College of Life Sciences at Dundee will work with the European Bioinformatics Institute (EMBL-EBI) and the University of Cambridge to create a next generation image data resource to host, serve, and make available original scientific image data that underpins life sciences research.

"We are seeing things that we have never been able to see before, thanks to a new generation of imaging technology and techniques," said Professor Swedlow. "But with these advances also come challenges. The new imaging techniques generate enormous amounts of complex data and so special tools are required to manage, share and analyse that data.

"At Dundee we have built several open software platforms that are used by scientists worldwide for accessing and managing these enormous datasets. This new project will take that work up another level."

The new resource will be built using OME's open source technologies and will be housed at EMBL-EBI, which is the established home of molecular and structural life sciences data. The resource will interface with ELIXIR, Europe's research infrastructure for life science informatics. It will build links with established molecular and structural resources and work towards a seamless integration of these data, so that any scientist can easily browse, query and compute on genomic, structural and phenotypic data across several scales.

With an international reputation for excellence in imaging and microscopy the University recently announced the launch of the Dundee Imaging Facility, an £8 million resource helping to drive the University's objectives to develop transformational research in physical, life and medical sciences.

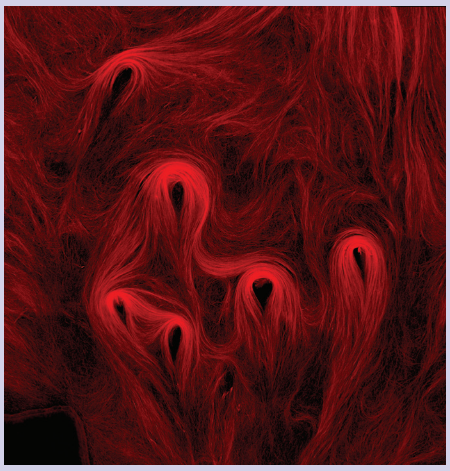
OME's work is part of this internationally recognised activity. Professor Swedlow founded OME with colleagues at MIT in 2000. Since then, OME has grown into an international consortium, based in Dundee, that has revolutionised the ability of researchers and industrial partners to handle, analyse, share and interpret vast amounts of image data.

Biological Sciences research at Dundee was rated top among universities in the UK in the 2014 Research Excellence Framework - the main test of research quality carried out by higher education funding bodies.

Professor Jackie Hunter, BBSRC Chief Executive, said, "This funding is one example of BBSRC strengthening investment in big data infrastructure so that scientists can access vast quantities of data to create the knowledge that will be needed to tackle the challenges of tomorrow.

"We experience problems coping with our own local data storage – videos, picture and other media take up huge amounts of space on our home computers. In life sciences, the data required for research is unimaginably larger and growing at unprecedented rates. The reference wheat genome takes up about 6 Gigabytes, for example, and a high resolution video of the human heart of just one patient can be around 50 Gigabytes, or the equivalent of 50 feature length films at standard definition.

"This data provides a mine of information that will help us now and in the future but it needs to be properly stored, curated and made easily accessible. These investments will help us achieve this in important areas, from discovering new drugs to breeding crops that are more resistant to climate change."



*Microtubules in vitro' by Ian Newton, College of Life Sciences, University of Dundee. This image was also voted winner of BBSRC's 'Images with Impact' competition, 2014.*

For more information, please contact:
Hayley London
Marketing & Communications Officer, The Genome Analysis Centre (TGAC).
Tel: +44 (0) 1603 450107, Email: Hayley.London@tgac.ac.uk

### About BBSRC

The Biotechnology and Biological Sciences Research Council (BBSRC) invests in world-class bioscience research and training on behalf of the UK public. Our aim is to further scientific knowledge, to promote economic growth, wealth and job creation and to improve quality of life in the UK and beyond.

Funded by UK Government, BBSRC invested over £484 million in world-class bioscience in 2013-14. We support research and training in universities and strategically funded institutes. BBSRC research and the people we fund are helping society to meet major challenges, including food security, green energy and healthier, longer lives. Our investments underpin important UK economic sectors, such as farming, food, industrial biotechnology and pharmaceuticals.

For more information about BBSRC, our science and our impact:
http://www.bbsrc.ac.uk

For more information about BBSRC strategically funded institutes:
http://www.bbsrc.ac.uk/institutes

### About TGAC

The Genome Analysis Centre (TGAC) is a world-class research institute focusing on the development of genomics and computational biology. TGAC is based on the Norwich Research Park and receives strategic funding from the Biotechnology and Biological Science Research Council (BBSRC) - £7.4 million in 2013/14 - as well as support from other research funders. TGAC is one of eight institutes that receive strategic funding from BBSRC. TGAC operates a National Capability to promote the application of genomics and bioinformatics to advance bioscience research and innovation.

TGAC offers state of the art DNA sequencing facility, unique by its operation of multiple complementary technologies for data generation. The Institute is a UK hub for innovative Bioinformatics through research, analysis and interpretation of multiple, complex data sets. It hosts one of the largest computing hardware facilities dedicated to life science research in Europe. It is also actively involved in developing novel platforms to provide access to computational tools and processing capacity for multiple academic and industrial users and promoting applications of computational Bioscience.

Additionally, the Institute offers a Training programme through courses and workshops, and an Outreach programme targeting schools, teachers and the general public through dialogue and science communication activities.

www.tgac.ac.uk