OXFORD

Briefings in Bioinformatics, 20(2), 2019, 384–389

doi: 10.1093/bib/bbx071
Advance Access Publication Date: 30 June 2017
Paper

Establishing a distributed national research infrastructure providing bioinformatics support to life science researchers in Australia

Maria Victoria Schneider, Philippa C. Griffin, Sonika Tyagi, Madison Flannery, Saravanan Dayalan, Simon Gladman,

Maria V. Schneider is the Deputy Director of EMBL Australia Bioinformatics Resource (EMBL-ABR) and Associate Professor at the University of Melbourne, Australia. She has a keen interest in data-driven science and open science and best practice in data life cycle, FAIR data, tools and bioinformatics training.
Philippa C. Griffin is a Bioinformatician/Research Fellow at EMBL-ABR: Melbourne Bioinformatics Node, University of Melbourne, Australia. She uses bioinformatics approaches to tackle questions around species ecology, evolution and climate response.
Sonika Tyagi is the Bioinformatics Supervisor at the Australian Genome Research Facility Ltd, in Melbourne. She is also the Head of the EMBL-ABR: AGRF Node. She develops bioinformatics tools and applications using machine learning, data mining and statistical approaches. She has a keen interest in open source and bioinformatics training. She is the champion for EMBL-ABR Key Area Training Coordination.
Madison Flannery was a developer at the EMBL-ABR: MelBioinf Node supporting EMBL-ABR Hub with ToolsAU and STM.
Saravanan Dayalan is the Lead Scientist (Bioinformatics and Biostatistics) at Metabolomics Australia, the University of Melbourne. He is part of EMBL-ABR: MA Node and currently the champion for EMBL-ABR Key Area Standards Coordination. He is interested in large-scale bioinformatics solutions, such geographically distributed LIMS and database systems, biostatistical methods and big data.
Simon Gladman is a research bioinformatician working on the Genomics Virtual Laboratory (GVL) and member of EMBL-ABR: MelBioinf Node, University of Melbourne, Australia.
Nathan Watson-Haigh is a Senior Research Fellow in Bioinformatics at the University of Adelaide. He is EMBL-ABR Key Area Tools Coordinator.
Philipp E. Bayer is part of the Edwards group in the School of Biological Sciences, University of Western Australia. Philipp's research focuses on plant genomics for the study of plant breeding and evolution. He is part of EMBL-ABR: UWA Node.
Michael Charleston is Associate Professor in Bioinformatics at the University of Tasmania. He is the Head of UTAS Node for EMBL-ABR and a member of the GOBLET Standards Committee. His research interests include phylogenetics, co-evolution, molecular evolution, the analysis of biological networks and modelling the ecology and behaviour of social insects.
Ira Cooke is Senior Lecturer in Bioinformatics at the Comparative Genomics Centre and Department of Molecular and Cell Biology, James Cook University. He is also the Head of Node for EMBL-ABR: JCU Node. He is interested in invertebrate genomes, particularly corals and cephalopods. He also develops software tools for proteomics that make better use of draft genome and transcriptome data for non-model organisms.
Rob Cook is the Chief Executive Officer of the Queensland Cyber Infrastructure Foundation Ltd (QCIF). QCIF facilitates and coordinates advanced eResearch services and information technology infrastructure for research in its member universities. He is the Head of Node for EMBL-ABR: QCIF Node.
Richard J. Edwards is Senior Lecturer in Bioinformatics at the University of New South Wales, Sydney, Australia. He works on the development and application of tools for protein interaction motif discovery (SLiMSuite), and genomics with PacBio long-read DNA sequencing. He is fascinated with the molecular basis of evolutionary change, and how we can harness the genetic sequence patterns left behind to make useful predictions about contemporary biological systems. He is part of EMBL-ABR: SBI Node.
David Edwards is a Professor in the School of Biological Sciences and the Institute of Agriculture at the University of Western Australia. His research activities include the characterization of complex plant genomes, translational genomics and genome informatics, with a focus on wheat, *Brassica* and chickpea crops. He is the Head of the UWA Node for EMBL-ABR.
Dominique Gorse is Director of QFAB Bioinformatics (Brisbane, Australia) where he leads the development of QFAB's platform for integrated and accessible bioinformatics, which is designed to support large multi-institution research projects and to provide advanced bioinformatics services to the biotechnology, pharmaceutical, clinical and research communities. He is part of EMBL-ABR: QCIF Node.
Malcolm McConville is Director of the Bio21 Institute, University of Melbourne, the Head of the Metabolomics Australia at the Bio21 Institute and the Head of Bioinformatics of Metabolomics Australia. He is the Head of Node for EMBL-ABR: MA Node.
David Powell leads the Monash Bioinformatics Platform and is the Head of Node for EMBL-ABR: Monash Node.
Marc R. Wilkins directs two centres at UNSW: the NSW Systems Biology Initiative and the Ramaciotti Centre for Genomics. His current research interests include (i) the role of protein methylation in the proteome, (ii) the dynamics of protein interaction networks and (iii) the analysis of next-generation sequencing data. He is the Head of Node for EMBL-ABR: SBI Node.
Andrew Lonie is the Director of EMBL-ABR, Director of the Melbourne Bioinformatics at the University of Melbourne, Australia and the Node Head of the Melbourne Bioinformatics Node of EMBL-ABR.

© The Author 2017. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Nathan Watson-Haigh, Philipp E. Bayer, Michael Charleston, Ira Cooke, Rob Cook, Richard J. Edwards, David Edwards, Dominique Gorse, Malcolm McConville, David Powell, Marc R. Wilkins and Andrew Lonie

Corresponding author: Maria Victoria Schneider, University of Melbourne, Australia. Tel.:+61 3 8344 1395; E-mail: mvschneiderg@gmail.com; Andrew Lonie, University of Melbourne, Australia. Tel.: +61 3 8344 1395; E-mail: alonie@unimelb.edu.au

Abstract

EMBL Australia Bioinformatics Resource (EMBL-ABR) is a developing national research infrastructure, providing bioinformatics resources and support to life science and biomedical researchers in Australia. EMBL-ABR comprises 10 geographically distributed national nodes with one coordinating hub, with current funding provided through Bioplatforms Australia and the University of Melbourne for its initial 2-year development phase. The EMBL-ABR mission is to: (1) increase Australia's capacity in bioinformatics and data sciences; (2) contribute to the development of training in bioinformatics skills; (3) showcase Australian data sets at an international level and (4) enable engagement in international programs. The activities of EMBL-ABR are focussed in six key areas, aligning with comparable international initiatives such as ELIXIR, CyVerse and NIH Commons. These key areas—Tools, Data, Standards, Platforms, Compute and Training—are described in this article.

Key words: EMBL-ABR; distributed network; bioinformatics infrastructure; data-driven analysis; bioinformatics service; bioinformatics training; data-driven science; national; capability; services

Introduction

The surge of bioinformatics infrastructures at national and international levels has been nothing short of transformational in the past decade. In a wide range of countries and jurisdictions, efforts to establish and support bioinformatics infrastructure for the life sciences have started, and continue to expand. Examples can be found in Europe, such as the Dutch Techcentre for Life Sciences (DTL) in the Netherlands, the German Network for Bioinformatics Infrastructure (deNBI) and ELIXIR at the European level; in the United States, including the National Institutes for Health's 'Big Data to Knowledge' (BD2K) and the National Science Foundation's CyVerse; and in Canada with Bioinformatics/Computational Biology Framework (B/CB).

The establishment of ELIXIR has acted as a catalyst for many European countries that either had some existing level of national bioinformatics infrastructure or had to develop this *de novo*. ELIXIR is a distributed infrastructure for life science information which aims to coordinate, integrate and sustain bioinformatics resources across its member states—including those at the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI)—to enable users in academia and industry to access vital data, tools, standards, compute and training services for their research [1]. A different approach is taken in the United States, where life science, agriculture and biomedical researchers can access multiple major bioinformatics projects and resources that act at the national level to provide bioinformatics infrastructure: (a) the National Center for Biotechnology Information (NCBI) advances science and health by providing access to biomedical, agricultural and genomic information [2]; (b) CyVerse is a dynamic virtual organization that provides scientists across all life sciences disciplines with computational infrastructure to handle huge data sets and complex analyses, thus enabling data-driven discovery [3]; (c) the NIH Commons is a virtual platform for biomedical data, software, metadata and workflow discovery, management, sharing and reuse [4]; (d) BD2K is a trans-NIH initiative established to enable biomedical research as a digital research enterprise, to facilitate discovery and support new knowledge and to maximize community engagement [5]; (e) the AgBioData working group aims to coordinate bioinformatics capability and expertise in the United States Department of Agriculture.

Australia, although large geographically, has a relatively small number of bioscientists: 30–35 000 full-time equivalent based on our estimate of 30–35% of the total ∼100 000 Australian researcher workforces [6], compared with ∼500 000 bioscientists in the European Union, which is about half the area of Australia. The Australian biomedical/bioinformatics community is concentrated in several large, but geographically separated research precincts, usually in the major state/national capital cities, and face-to-face meetings between collaborators across states realistically require air travel—typically 1 h (e.g. Melbourne, Sydney or Adelaide) to 4.5 h (Perth–Brisbane or Sydney). Although relatively well temporally aligned with Asian countries [e.g. China (0–2 h) and India (2.5–5.5 h)], Australia has substantial time differences with Europe (9–11 h) and North America (14–16 h). This often means that considerable logistic effort is required for international collaborations, with meetings that are either early in the morning or late at night.

These factors create a set of challenges when it comes to the maintenance of important collaborations at a national or international level, and Australia's geographic characteristics have of course influenced the development of national infrastructure in support of bioinformatics and biosciences research.

Like many countries, Australia has a long but somewhat inconsistent history of providing national bioinformatics infrastructure, often in the context of numerous institutional and state-based efforts. An early example is the Australian National Genome Information Service [6], started in 1991 but which ceased operations in 2009, which itself evolved out of the Sydney University Sequence Analysis Interface. Shortly after, a

Bio-Mirror [7] was established at Australian National University, providing high-speed access to common hosted data sets, which of course was (and remains) an important factor in Australian digital research infrastructure access because of intercontinental network bandwidth constraints. Significant investment in national bioinformatics was made through the Australian Research Council Centre of Excellence in Bioinformatics [8] from 2003 to 2010, which in turn provided a natural context for the establishment of the EBI Mirror project at the University of Queensland in 2010. Concurrently, in 2006, a major infrastructure investment through the National Collaborative Research Infrastructure Scheme (NCRIS) [9] was made in omics data platform infrastructure through Bioplatforms Australia [10], and the Australian Bioinformatics Facility [11], hosted by the Centre for Comparative Genomics at Murdoch University, was established in support of that data infrastructure. As the discipline of bioinformatics grew, several initiatives to form an Australian Bioinformatics Network (ABN) were made in 2006 and 2012 [12] to foster the developing bioinformatics community and its connection to the broader research community; in 2014, ABN evolved into the Australian Bioinformatics And Computational Biology Society (ABACBS) [13] in response to the maturing of bioinformatics in Australia as a discipline. As ABN evolved in support of bioinformatics, so did the EBI Mirror project in support of bioinformatics infrastructure, in 2012, becoming the Bioinformatics Resource Australia of EMBL, Australia, also known as BRAEMBL [14]. The project evolved from this original concept to also encompass services across the areas of tools and training across Australia in 2015. In February 2016 the project was relaunched as a national network of bioinformatics infrastructure services: EMBL Australia Bioinformatics Resource (EMBL-ABR).

Access to hardware, bioinformatics services and data remain a priority in biosciences, and the challenges involved across the biological domains are shared, so implementing solutions at national level means reducing redundancy of effort and ensuring longer-term sustainability for the maintenance and further developments [15–18]. EMBL-ABR is an evolving response towards a federated network of bioinformatics infrastructure in Australia to address these challenges and provide a suitable 'digital ecosystem' that leverages on existing NCRIS capabilities such as Research Data Services, the National eResearch Collaboration Tools and Resources (Nectar), Australian National Data Service (ANDS) and BPA and is also linked and immersed in the cutting-edge developments that are taking place overseas.

EMBL-ABR structure

EMBL-ABR has a coordinating hub hosted by Melbourne Bioinformatics at the University of Melbourne and 10 nodes across Australia, which represent existing institutions and facilities that are already working in one or more of the six key areas. The overall administration, communication and outreach capacity are provided by the hub. Figure 1 shows the current structure of EMBL-ABR. The scientific remit and actual activities are spread across the nodes, with the aim to form a roadmap to develop bioinformatics resources as well as adopt existing solutions that can be federated in Australia once consistent funding is secured. Each node therefore has compiled a form describing their existing capabilities and how these are scalable at state and national level. Initial node descriptions are all published on EMBL-ABR website [19]. Hence, the overall set of EMBL-ABR activities depends on the actual areas of expertise and bioinformatics resources already present in the nodes.

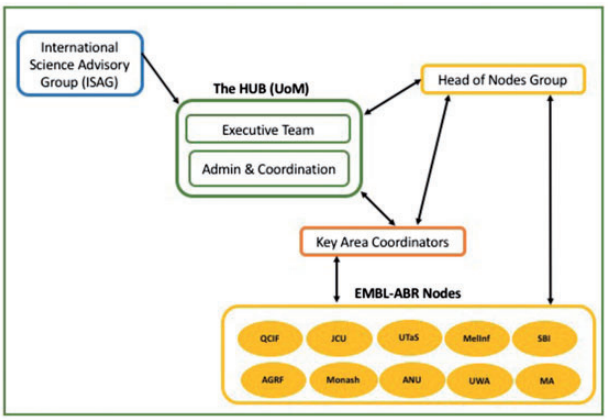

Figure 1. Structure of EMBL-ABR.

EMBL-ABR Key Areas (summarized in Table 1) map clearly to the ELIXIR Platform topics: Data, Tools, Compute, Training and Interoperability, the last of which is closely related to the EMBL-ABR Key Area of Standards. These key areas encompass all aspects of bioinformatics across specific research domains, encouraging expertise sharing and reuse for similar problems identified in distinct research domains. EMBL-ABR is also interacting with specific biological domains to address their respective needs, such as wheat annotation and bioinformatics, invertebrate omics and microbial bioinformatics.

EMBL-ABR Nodes are represented by the Head of the Nodes Group that together with the Hub executive team have a regular monthly meeting based on a shared and collaborative agenda to discuss the strategy as well as actual activities. The Hub executive team is in regular contact with the Head of the Nodes Group and sends updates through its mailing list. The agenda and minutes, plus nodes updates, are also posted online, in the key documents area, so anyone can stay informed and follow-up EMBL-ABR discussions [20].

EMBL-ABR relies on the goodwill and enthusiasm from its nodes and champions and has a group of Key Area Coordinators to collect what the Australian Biosciences needs are for these key areas and aid the development of EMBL-ABR activities and priorities to address such needs. The group has also established monthly meetings to discuss updates as well as solutions and advances in these areas that exist and could be of benefit to federate within Australia.

EMBL-ABR has also established an International Science Advisory Group (ISAG) that brings together 10 members, highly regarded scientists from around the world, to discuss and advise on EMBL-ABR strategy, activities and priorities. EMBL-ABR ISAG is composed of a mix between international bioinformatics infrastructure experts and Australian biological domain experts. The group met face to face in December 2016, and took an active participation in the first EMBL-ABR All-Hands meeting [21]. The review and feedback from the ISAG helps EMBL-ABR to align its strategy with international standards and needs of Australian life science researchers.

EMBL-ABR has direct links with ABACBS through its chair being a member of the EMBL-ABR ISAG, and the EMBL-ABR Training Coordinator being part of ABACBS's dedicated Education and Training subcommittee.

EMBL-ABR encourages Australian Bioinformaticians to become members of ABACBS, as it offers the best forum to

Table 1. EMBL-ABR Key areas mapped to the ELIXIR platform topics: Data, Tools, Compute, Training and Interoperability

Key area	Aims	EMBL-ABR Nodes working in this Key Area
Training	• Provide Hub training on critical topics not covered elsewhere • Coordination and dissemination of node end-user training activities (STM, iAnn) • Enable improved access to EBI training resources and international experts	All EMBL-ABR Nodes
Data	• Provide advice on best practice data management and data life cycle in the Biosciences • Showcase Australian data at an international level • Catalogue and deliver the bioinformatics data Australia needs	ANU, MA, Melbourne Bioinformatics, QCIF, SBI, UTAS, UWA, AGRF
Standards	• Foster adoption of standardized file formats, metadata, vocabularies and identifiers by the Australian Bioscience community • Bring Australian needs in standards development to the attention of international efforts • Ensure Australian input into development and where appropriate coordination of international standards activities	MA, AGRF, QCIF
Tools	• Assess Australian needs for bioinformatics software hosting, maintenance and support • Implement a registry of Australian bioinformatics tools, ToolsAU • Contribute to the development and dissemination of bioinformatics tools	AGRF, ANU, MA, Melbourne Bioinformatics, SBI, UTAS, UWA
Platforms	• Leverage existing expertise in platform development • Enable research access to bioinformatics platforms that link multiple tools, facilitate data sharing and access and record analysis pipelines	ANU, JCU, MA, Melbourne Bioinformatics, Monash, QCIF, UTAS, UWA
Compute	• Assist in the design, architecture and delivery of compute infrastructure • Support network activities to ensure interoperability across Australia and with international efforts	AGRF, Melbourne Bioinformatics, Monash, QCIF

participate and discuss bioinformatics career developments and opportunities and to advance awareness and recognition of this ever-expanding field. ABACBS is also the place for listings for all job opportunities for bioinformaticians across the Australian research community.

EMBL-ABR mission and remit

EMBL-ABR aspires to:

1. increase Australia's capacity to collect, integrate, analyse, exploit, share and archive the large heterogeneous data sets, now part of modern life sciences research;
2. contribute to the development and delivery of training in data, tools, platforms and international standards, to enable Australia's life science researchers to undertake research in the age of big data;
3. showcase Australian research and data sets at an international level; and
4. enable engagement in international programs that create, deploy and develop best practice approaches to data management, software tools and methods, computational platforms and bioinformatics services.

These four objectives contribute overall to create a sustainable bioinformatics infrastructure network that reflects the specific geographical and temporal patterns of those working in Australia.

The life sciences and biomedical bioinformatics communities are critically dependent on the tools needed to store, find, access, annotate, enrich, visualize, integrate and interpret data. Tool development is extremely dynamic, partly because of the fast pace of advancement in data production technologies and the consequent spawning of new file formats. Recent efforts [22–27] aim to describe best practice for academic software. Tools should have exposure, so that redundancy can be avoided, and sustainable plans can be in place for maintenance and/or expansion of functions, as the field and users' needs evolve. During 2016, EMBL-ABR set up a Tool registry, ToolsAU, powered by the ELIXIR Tools and Data registry [28, 29], which allows users to filter and display the tools created in Australia or with Australian involvement. The current registry (3 October 2016) shows >50 entries for Australian Tools: https://www.embl-abr.org.au/tools/toolsau/. EMBL-ABR is also contributing to the community effort developing open-source principles to promote software quality and sustainability [30]. A Search for Training Material (STM) was implemented to collect Australian training material resources and enhance discoverability of such materials [31].

The quality and utility of data relies on the existence and adoption of standards, shared formats and mechanisms for biological researchers to share and annotate the data, so it can be easily searchable, conveniently linked and consequently used for further biological analysis and discovery [32]. One of the biggest challenges is the inherent heterogeneity of biological data types [32–34]. Effective solutions that reduce the challenges associated with data volume and complexity are being developed worldwide [35, 36]. However, awareness from data producers and consumers of best practice in data management, as well as intuitive tools that allow researchers to easily implement such practices, remains lacking [37–39]. The development of the FAIR principles (Findable, Accessible, Interoperable and

Reusable) [40] and recent efforts internationally on applying such principles are critical for ensuring use and reuse [41, 42].

To gather an overview of the spectrum of data Australia contributes in terms of domains in biology, systems and model organisms as well as the type of data, EMBL-ABR launched a BioSharing Collection, so that our quantification of available data would also capture the parameters established by BioSharing and the larger international community in terms of best practice. This is an ongoing activity and as of 3 October 2016 listed 15 databases [43].

EMBL-ABR has also been active in initiating key collaborations for training, standards and federation of services with international partner, including EMBL-EBI, GOBLET, University of Cambridge Bioinformatics Training team, several ELIXIR Nodes (Belgium, Italy and UK) and CyVerse in United States.

Key Points

- EMBL-ABR was launched on February 2016 and includes 10 nodes distributed across Australia.
- EMBL-ABR has a direct connection with the user communities by including biological domain experts across its structure, including its International Advisory group.
- EMBL-ABR activities and remit are spread across six key areas: Data, Tools, Platforms, Standards, Compute and Training.
- EMBL-ABR has collected during 2016 information about bioinformatics training events and materials produced in Australia and Tools created in Australia, by federation of existing solutions (Biotools.org and STM) rather than recreating such efforts from scratch.
- EMBL-ABR is fostering and promoting awareness of Australian bioinformatics and its bioscience overall across a variety of international institutions and collaborations.

Acknowledgements

The authors are grateful to the ISAG for their time and availability; Jason Williams and Paul Flicek for their regular time and availability for discussions; Rafael Jimenez for his effort and proactive engagement with EMBL-ABR; Allegra Via and Gabriella Rustici for their inclusivity and availability to explore and discuss training collaborations; Frederik Coppens for the discussions and exploration on possible collaborations around virtualization solutions from Australia; Susanna Sansone and Peter McQuilton for their support and welcoming approach when it comes to discussions on standards and best practices across databases and policies; Fiona Kerr, Karin Diamond and Claudia Curcio for the constant support of the EMBL-ABR Hub; Helen Gardiner and Christina Hall for their hard work and continuous support on EMBL-ABR communication (for the hub and network) and also for providing useful comments and edits to this manuscript; and Ben Moran for his fast action when it comes to the EMBL-ABR website.

Funding

The University of Melbourne and Bioplatforms Australia (BPA) via an Australian Government NCRIS investment (to EMBL-ABR).

References

[1]. Crosswell LC, Thornton JM. ELIXIR: a distributed infrastructure for European biological data. *Trends Biotechnol* 2012;**30**(5):241–2.
[2]. Lindberg DA. Internet access to the National Library of Medicine. *Eff Clin Pract* 2000;**3**(5):256–60.
[3]. Merchant N, Lyons E, Goff S, *et al*. The iPlant collaborative: cyberinfrastructure for enabling data to discovery for the life sciences. *PLoS Biol* 2016;**14**(1):e1002342.
[4]. Bonazzi V. NIH Commons Overview, Framework & Pilots - Version 1, 2015. https://datascience.nih.gov/sites/default/files/CommonsOverviewFrameWorkandCurrentPilots281015_508.pdf.
[5]Data Science at NIH. https://datascience.nih.gov/bd2k/about
[6]. ANGIS. https://en.wikipedia.org/wiki/ANGIS.
[7]. Biomirror. http://biomirror.aarnet.edu.au/biomirror/about-biomirrors.txt (3 May 2017, date last accessed).
[8]. Australian Research Council (ARC) Centre of Excellence in Bioinformatics. http://bioinformatics.org.au/documents/ACB-annual-report-2010.pdf (3 May 2017, date last accessed).
[9]. National Collaborative Research Infrastructure Scheme. https://www.education.gov.au/national-collaborative-research-infrastructure-strategy-ncris (3 May 2017, date last accessed).
[10]. Bioplatforms Australia (BPA). http://www.bioplatforms.com/(3 May 2017, date last accessed).
[11]. Australian Bioinformatics Facility. https://researchdata.ands.org.au/australian-bioinformatics-facility/11524 (3 May 2017, date last accessed).
[12]. Australia Bioinformatics Network (ABN). http://australianbioinformatics.net/our-history/
[13]. Australian Bioinformatics and Computational Biology Society. http://www.abacbs.org/about/ (3 May 2017, date last accessed).
[14]. BRAEMBL. http://www.imb.uq.edu.au/braembl (3 May 2017, date last accessed).
[15]. Cook CE, Bergman MT, Finn RD, *et al*. The European bioinformatics institute in 2016: data growth and integration. *Nucleic Acids Res* 2016;**44**(D1):D20–6.
[16]. Stephens ZD, Lee SY, Faghri F, *et al*. Big data: astronomical or genomical? *PLoS Biol* 2015;**13**(7):e1002195.
[17]. Laurence M, Dalgleish R, Thorisson GA, *et al*. The use of bioresources for promoting their sharing in scientific research. *Gigascience* 2013;**2**:7.
[18]. Lampa S, Dahlö M, Olason PI, *et al*. Lessons learned from implementing a national infrastructure in Sweden for storage and analysis of next-generation sequencing data. *Gigascience* 2013;**2**:9.
[19]. EMBL-ABR Nodes Descriptions. https://www.embl-abr.org.au/node-descriptions-forms_2016/ (3 May 2017, date last accessed).
[20]. EMBL-ABR Head of Nodes Group. https://www.embl-abr.org.au/head-nodes-group/ (3 May 2017, date last accessed).
[21]. EMBL-ABR All hands 2016. https://www.embl-abr.org.au/all-hands-mtg-2016/ (3 May 2017, date last accessed).
[22]. Stodden V, Guo P, Ma Z. Toward reproducible computational research: an empirical analysis of data and code policy adoption by journals. *PLoS One* 2013;**8**:e16800. doi: 10.1371/journal.pone.0067111
[23]. Shamir L, Wallin JF, Allen A, *et al*. Practices in source code sharing in astrophysics. *Astron Comput* 2013;**1**:54–8.
[24]. Poline JB, Breeze JL, Ghosh S, *et al*. Data sharing in neuroimaging research. *Front Neuroinform* 2012;**6**:9.

[25]. Stodden VC. Trust your science? Open your data and code. *Amstat News* 2011;**409**:21–2.
[26]. Ince DC, Hatton L, Graham-Cumming J. The case for open computer programs. *Nature* 2012;**482**:485–8.
[27]. McKiernan EC, Bourne PE, Brown CT, *et al*. How open science helps researchers succeed. *eLife* 2016;**5**:e16800.
[28]. Ison J, Rapacki K, Ménager H, *et al*. Tools and data services registry: a community effort to document bioinformatics resources. *Nucleic Acids Res* 2016;**44**(D1):D38–47.
[29]. Pettifer S, Thorne D, McDermott P, *et al*. An active registry for bioinformatics web services. *Bioinformatics* 2009;**25**:2090–1.
[30]. Open Source Software recommendations for research. https://github.com/SoftDev4LS/open-source-software.
[31]. Search for Training Materials (STM). http://stm.embl-abr.org.au/(3 May 2017, date last accessed).
[32]. Lapatas V, Stefanidakis M, Jimenez RC, *et al*. Data integration in biological research: an overview. *J Biol Res* 2015;**22**:9.
[33]. Schneider MV, Jimenez RC. Teaching the fundamentals of biological data integration using classroom games. *PLoS Comput Biol* 2012;**8**(12):e1002789.
[34]. Gomez-Cabrero D, Abugessaisa I, Maier D, *et al*. Data integration in the era of omics: current and future challenges. *BMC Syst Biol* 2014;**8(Suppl 2)**:I1.
[35]. Ma'ayan A, Rouillard AD, Clark NR, *et al*. Lean big data integration in systems biology and systems pharmacology. *Trends Pharmacol Sci* 2014;**35**(9):450–60.
[36]. Goble C, Stevens R. State of the nation in data integration for bioinformatics. *J Biomed Inform* 2008;**41**(5):687–93.
[37]. Nekrutenko A, Taylor J. Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nat Rev Genet* 2012;**13**:667–72.
[38]. Barone L, Williams J, Micklos D. Unmet needs for analyzing biological big data: a survey of 704 NSF principal investigators. *BioRxiv*. 2017; 1–12. https://doi.org/10.1101/108555.
[39]. Schneider MV, Flannery M, Griffin P. Survey of Bioinformatics and Computational Needs in Australia 2016.pdf. Figshare, 2016. https://dx.doi.org/10.6084/m9.figshare.4307768.v1
[40]. Wilkinson MD, Dumontier M, Aalbersberg IJ, *et al*. The FAIR guiding principles for scientific data management and stewardship. *Sci Data* 2016;**3**:160018.
[41]. Rodríguez-Iglesias A, Rodríguez-González A, Irvine AG, *et al*. Publishing FAIR data: an exemplar methodology utilizing PHI-base. *Front Plant Sci* 2016;**7**:641.
[42]. Nature Genetics Editorial. FAIR principles for data stewardship. *Nat Genet* 2016;**48**:343.
[43]. EMBL-ABR BioSharing Collection. https://biosharing.org/collection/EMBLABR (3 May 2017, date last accessed).