Bioinformatics Training: towards a Programme of National impact for Australia

Vicky Schneider, University of Melbourne, AU
Gabriella Rustici, University of Cambridge, UK
Jason Williams, Cold Spring Harbor, USA

Table of Contents

1. Background ........................................1
2. Why do we need more training? ........2
3. Types of training to be organised ......3
   3.1 Efficient dissemination of Knowledge through Webinars: tailored to AU needs ..3
   3.2 Case studies focusing on Australia datasets..................................................3
      3.2.1 Joint online e-learning courses with EMBL-EBI ......................................3
      3.2.2 Creation of Join training materials with ELIXIR UK and Uni Cambs........3
      3.2.2 Australian Led Training Modules .. 3
   3.3 Training Capacity in Australia..........4
      3.3.1 Collaborating with existing experienced TtT global efforts ..............4
      3.3.2 Open Science & Bioinformatics.....4
4. Desirable training infrastructure .......4
5. Impact and Quality Assessment ........5
6. What is the minimum resourcing needed for making this happen? ........5
7. Short term activities and events .......5
   2017 ..................................................5
      Galaxy hackathon materials at Uni Camb .........................................................5
      ELIXIR UK Impact workshop at Uni Camb, ......................................................5
   Train the Trainer on Clinical Bioinformatics ......................................................6

1. Background
The advent of high-throughput technologies has greatly increased the demand from life scientists to learn how to use bioinformatics resources and tools. This worldwide phenomenon is also widespread in Australia. It is important to provide new opportunities and harmonise existing training efforts across Australia to maximise their reach and scope, as well as to combine experience and best practice from Australia with international initiatives. Vicky Schneider (UoM) has collected the interest and input from global partners and initiatives interested in collaborating and working towards a scalable national-level program for Bioinformatics training across Australia. This currently includes EMBL-EBI Training Team (Key person: Paul Flicek); University of Cambridge & Institute of Continuing Education (key person: Gabriella Rustici); CyVerse & Carpentries (Key person: Jason Williams); ELIXIR Training (Key persons: Gabriella Rustici, Celia van der Gelder, Patricia Palagi), and key individuals from the ELIXIR training scene such as Allegra Via (head of ELIXIR TtT taskforce), Frederik University of Cambridge & University of Melbourne workshop on Skills for Open and Scalable Data Analysis .......................6
Biostatistics Course from Uni Camb ..........6
Non model organisms genomics analysis course .....................................................6
Standards & BioSchemas........................6
BioSchemas Planning Meeting, Hinxton 6
2018 ..........................................................6
Implementation of Webinar programme .................................................................6
Launch of ftf training for the Non model organisms genomics analysis ....................6
Skills for Open and scalable data analysis (computing-Nectar and GVL) .................7
Biostatistics Course from Uni Camb .......7

About the Authors ................................. 8
Coppens (chair of the Elixir Galaxy Working Group).

Our proposal to EMBL-ABR and BPA consists of tri-partite training regime which bridges and brings to Australia existing efforts successfully implemented overseas as well as increasing the skills and competencies of bioinformatics trainers from Australia. The proposed set of actions also maximizes the impact and scalability of training, ensuring Australian bioscience research community needs (end users of bioinformatics resources and tools) guide the topical range of the training implemented.

BPA has already established a successful Train the Trainer (TtT) programme with CSIRO model which includes the EBI (http://www.bioplatforms.com/bioinformatics-training/) and we propose here an expansion from this model, to ensure it takes advantage of the major progress and efforts happening within ELIXIR as well as other relevant partners in USA.

It is important to also mention the efforts from the current bioinformatics community body in Australia, the Australian Bioinformatics and Computational Biology Society (ABACBS), http://www.abacbs.org/ which was incorporated in 2014, subsuming the student and early career research society COMBINE as a sub-committee and taking over responsibility for the Australian Bioinformatics Network, an online information hub on jobs, news and events in Australian bioinformatics that was initially instigated by EMBL Australia/CSIRO/BPA. ABACBS has a dedicated subcommittee to Education and Training in which EMBL-ABR is represented by its key area training coordinator Sonika Tyagi.

2. Why do we need more training?

There is undoubtedly an ever increase need in bioinformatics skills and competencies from actual technologists to end users. This has been also identified by EMBL-ABR survey in 2016 across Australia (https://doi.org/10.6084/m9.figshare.4307768.v1), and resounds with the recent outcomes from Cold Spring Harbor Laboratory and CyVerse study across 704 Principal Investigators funded by the U.S. National Science Foundation (http://biorxiv.org/content/early/2017/02/14/108555) which concluded that training, not computational infrastructure, was the most unmet need for biological researchers. Critical factors contributing to training gaps among life science researchers in Australia (and worldwide) include:

1 - insufficient exposure to the principles and application of bioinformatics;
2 - the speed at which technologies, data analysis practices and tools develop, and the lag in how these are incorporated into the undergraduate and postgraduate curricula;
3 - a lack of support within some research groups or institutions for students and postdoctoral researchers to learn bioinformatics skills and tools.

The combined effect of these barriers is that researchers can find themselves involved with, or even designing, projects that require skills and experience they are currently lacking. Similarly, when it comes to the exploitation, integration and accessibility of data, researchers time is often inefficiently used as they lack the skills, and awareness of the bioinformatics needed to solve their challenges correctly in a reasonable amount of time. In
contrast to other sciences where the data types are few and uniform, the heterogeneity of biological data requires a diverse of skills (http://dx.doi.org/10.1371/journal.pbio.1002195). Hence it is important to for a successful nation-wide training regime to cover a breadth of subject matter. These problems are not unique to Australia; if anything, the challenges are amplified here due to its geographic location as well as temporal difference with events taking place in the northern hemisphere. We stress the need to participate in solutions already in place as a sensible and cost-effective way to enable internationalisation of the training activities, access, and exportability of efforts in Australia.

3. Types of training to be organised

3.1 Efficient dissemination of Knowledge through Webinars: tailored to AU needs

Based on information and communication technologies, and the existing series of the EMBL-EBI Training Webinar Series, work closely together and create a forum supported by EMBL-ABR with EMBL-EBI to gather questions and problems that the Australia Bioscience community has, with the aim of getting the most out of publicly available biological data. We will explore how to work across time barriers and ensure accessibility for Australia of the online training EMBL-EBI already has established. Similarly, University of Cambridge Postgraduate Bioinformatics is also keen to explore opening relevant courses and events via webinars to Australia and develop joint materials and efforts as needs of the audiences overlap.

3.2 Case studies focusing on Australia datasets

3.2.1 Joint online e-learning courses with EMBL-EBI

Explore a collaboration on training materials where EMBL-ABR contributes to actual online courses in “Train online”, bringing specific case studies of relevance to Australia Biosciences.

3.2.2 Creation of Join training materials with ELIXIR UK and Uni Cambs

Based on feedback from EMBL-ABR Training coordinator, BPA Training demand surveying, bioscientists working in diverse organisms and areas, and EMBL-ABR nodes training offering, the following 3 courses were identified among AU priority areas:

-Non model organisms genomics analysis
-Analysis of genomics variation and clinical implications
-Skills for Open and scalable data analysis (computing-Nectar and GVL)

3.2.2 Australian Led Training Modules

Based on the gaps and priority needs of the biosciences domains within Australia, actual creation of specific tutorials that can also be showcased and adopted by the overseas partners.

Rad-seq Data analysis

Rad-seq data analysis is of interest among population geneticists, comparative genomics, evolutionary biology and phylogenetics, hence of extreme importance among the Australian bioscientists studying diversity and evolution of species of relevance, Australian species. Hence, we believe is of much need and relevance for Australia, and will have traction internationally is the development of modular training for Rad-seq data analysis. Vicky Schneider (UoM)
and Sonika Tyagi (AGRF) has already discussed developing this within the context of Global organisation for Learning, Education and Training.

**Biostatistics for Bioinformatics users**
Contribution and development of materials in biostatistics to complement and expand on the exiting effort starting now at the University of Cambridge: Basic Statistics and Data handling course. This course would be delivered in person both in UK and Australia. The aim would be to create a put together a pool Australian based trainers that can deliver the training in collaboration with UK based trainers.

**Genome annotation**
Create the set of training materials for genome annotation in parallel to the developments of the projects UoM and EMBL-ABR are leading on:
- wheat genome manual annotation
- Australian mammals genome annotation
The actual process adopts Webapolo as the collaborative tool, which is something CyVerse is also developing, so we envisage working together in joint training materials and resources. Towards this goal, EMBL-ABR Hub will support the visit of Monica Munos-Torres http://www.berkeleybop.org/people/monica-munoz-torres/ to deliver actual training in genome annotation and webapolo, as well as participate on the Open Science SIG.

3.3 Training Capacity in Australia

3.3.1 Collaborating with existing experienced TtT global efforts

**Carpentries TtT**
Experience from TtT can be drawn from the Carpentries who have a certification program for their global cohort of volunteer instructors. Belinda Weaver, EMBL-ABR: QCIF node is a member of the Software Carpentry Steering Committee and is a certified instructor trainer. She would be a Carpentries advocate for scaling up a TtT program for EMBL-ABR throughout Australia.

**Elixir Train the Trainer Taskforce**
Explore potential methods for increasing training capacity by expanding the existing TtT programme to join in with trainers for activities 3.1, 3.2.1 and 3.2.2, as well as ensure participation on the TtT programme from ELIXIR led by Allegra Via (ELIXIR-IT). Vicky Schneider (UoM) is already involved and bringing the Australian focus to how the developed activities can be federated and implemented for Australia.

3.3.2 Open Science & Bioinformatics
Collaborate in shared activities and materials re-usability, for the Open Science training activities (including academic software management and shareability) led by Gabriella Rustici (Institute Continued Education) with Vicky Schneider (UoM). Current course in this area covers clinical bioinformatics. Open Science is also the theme for EMBL-ABR SIG in 2017, and the visit of Monica Munos-Torres from Berkeley Bioinformatics Open-Source Projects group (BBOP) will be instrumental to launch EMBL-ABR SIG and forge our efforts in line to initiatives and activities already happening in this area (e.g. BOSC).

4. Desirable training infrastructure
It is important to create an adequate impact framework and resource these efforts with the infrastructure that accounts for 1) mobility and time of trainers; 2) event management; 3) facilities
Bioinformatics Training: towards a Programme of National impact for Australia

March 2017

for face-to-face training; 4) VM environments for training modules. Australia is in a leading position when it comes to VM and the cloud for bioinformatics training. Discussions are already happening between UoM, EMBL-ABR and ELIXIR-Be (incl. the ELIXIR Galaxy Working Group) to learn from Australia about the mechanisms and resources used and how transferable are these to the European Training activities. The adoptability of the VM through the GVL as a scalable training infrastructure provides a mechanism to track materials produced and standardisation of the components to maximise share ability, traceability and actual adoption and use.

5. Impact and Quality Assessment

Vicky Schneider has started the process of having Australia’s activities included within the impact and quality assessment study conducted by ELIXIR-UK node (Louise Bellis and Gabriella Rustic). Vicky Schneider (UoM) will be working on sharing the information from EMBL-ABR training providers and other relevant and interested sources (e.g. ABACBS, BPA/CSIRO TtT) to contribute towards the development of best practice in impact and quality assessment of training, which already has involved USA (the Carpentries), South Africa (Nicky Mulder), ELIXIR, and through the University of Melbourne (Vicky Schneider) Australia. Our primary action for this will be to build on existing work (e.g. feedback from postgraduate courses already collected across EMBL-ABR nodes, BPA workshops) in collaboration with ELIXIR efforts led by ELIXIR-UK/University of Cambridge. The latter are currently implementing, monitoring, analysing and reporting on a comprehensive system of training metrics/key performance indicators across the ELIXIR training, which we will in parallel adopt to bring the impact metrics system to Australia Bioinformatics training.

6. What is the minimum resourcing needed for making this happen?

This proposal takes advantage of Vicky Schneider’s (proposal lead) role in creating and enhancing relationships and exchanges between AU bioscientists and the international bioinformatics community. To this, Australia would also need to support the mobility of Australian trainers overseas and some mobility also within Australia. The programme is design with the idea of having presence in the Australian pool across its states (EMBL-ABR nodes can provide such spread) so that the scalability derives from enabling the local trainers that are formed to support the local communities by accessing and working at the global level. Ideally, it would be important to consider the need to resource one to two FTEs (also in kind from existing nodes) that can work with Vicky on aspects such as technical requirements and implementation for the VM environments (from implementation to scalability and interoperability aspects); actual materials creation and liaising with AU content experts, training management (including events and facilities admin), impact and long term feedback capturing.

7. Short term activities and events

2017

Galaxy hackathon materials at Uni Camb in May (22-24) with Simon Gladman (Melbourne Bioinformatics and EMBL-ABR Node), and
ELIXIR UK Impact workshop at Uni Camb, September with Sonika Tyagi (EMBL-ABR Node) and Vicky Schneider (UoM and EMBL-ABR Hub)

Train the Trainer on Clinical Bioinformatics in July (3-5) (Vicky Schneider (UoM and EMBL-ABR Hub), and ideally another trainer from AU?)

University of Cambridge & University of Melbourne workshop on Skills for Open and Scalable Data Analysis, May, with Simon Gladman (Melbourne Bioinformatics and EMBL-ABR Node), Frederik Coppens (VIB, ELIXIR-BE) and others. Gabriella Rustici (University of Cambridge) and Vicky Schneider (Melbourne University) as main organisers. 25th of May 2017

Annotation and Open Science workshops, Australia
Vicky organising with the hub admin support, Monica Munos-Torres visit and per the proposed model of support from the nodes (one way trip internal + local accommodation and subsistence) her roadshow of annotation and OP workshops across AU. Vicky will be working with Pip Griffin on Monica’s activities for her week in Melbourne 26-30 June 2017

Biostatistics Course from Uni Camb collection of case studies from AU, Vicky Schneider (Melbourne University)

Non model organisms genomics analysis course
Start defining course structure and content, Vicky Schneider (University of Melbourne together with key content experts from Australia) and liaison with EMBL-EBI and ELIXIR-UK

Standards & BioSchemas
Vicky Schneider is also the GOBLET Standards Committee chair and as such bringing Australia to an active voice in the initiatives around standards from biocuration to training materials and dissemination, Vicky is also involved within Bioschemas and will continue working in this area and closely interacting with the Committee members and other relevant stakeholders. We will be working on discoverability and accessibility of exiting training materials from Australia. Vicky has added EMBL-ABR as a provider in TeSS and started to add materials as well as discuss with TeSS what are the requirements to get this process automated. In the long term this may represent a more viable and sustainable resource compared to the existing EMBL-ABR Search for Training materials tool http://stm.embl-abr.org.au/.


BioSchemas Planning Meeting, Hinxton
Mon 6 March 2017 to Wed 8 March 2017 EMBL - EBI, Hinxton. Vicky taking part as EMBL-ABR and UoM. DONE

2018
Implementation of Webinar programme
Launch of ftf training for the Non model organisms genomics analysis
Analysis of genomics variation and clinical implications
Start defining course structure and content, Vicky Schneider (University of Melbourne together with key content
experts from Australia) and liaison with EMBL-EBI and ELIXIR-UK

Skills for Open and scalable data analysis (computing-Nectar and GVL)
Implement actual course both in UK and AU. Vicky Schneider (University of Melbourne together with key content experts from Australia) and liaison with EMBL-EBI and ELIXIR-UK

Biostatistics Course from Uni Camb
Run and create the Australia based pool of trainers in liaison with their UK equivalents, led by Gabriella Rustici (University of Cambridge) and Vicky Schneider (University of Melbourne).
About the Authors

Gabriella Rustici

Gabriella Rustici is the Head of the Bioinformatics Training Program of the School of Biological Sciences of the University of Cambridge. In this role, she conceives, designs and implements a broad range of high-quality undergraduate and postgraduate hands-on training courses focused on bioinformatics and computational biology. Previously, she worked for 7 years at the European Bioinformatics Institute as Research and Training Coordinator, where she developed the scientific content of over 100 training events, covering a wide range of topics from introduction to bioinformatics to advanced analysis of high-throughput data, including 6 EMBO-funded practical courses.

Jason Williams

Jason Williams is Assistant Director, External Collaborations, at Cold Spring Harbor Laboratory’s DNA Learning Center where he works to spread hands-on biology education internationally. As Education, Outreach, and Training lead for CyVerse – the U.S. National Cyber infrastructure for Life Science – Jason provides training and support to scientists and educators, helping them leverage the most advanced tools and best methods for research and education in data-intensive biology. Jason organises, instructs, and speaks at more than dozen bioinformatics workshops and conferences annually. Additionally, he serves on several committees and boards for projects that advance science and science education including his service on the Steering Committee of the Software Carpentry Foundation (Chair, 2016), and as an instructor for Software and Data Carpentry – organisations that centre around scientists teaching scientists computational best practice.

Vicky Schneider

Vicky Schneider is an Associate Professor at the University of Melbourne. Vicky is also the Deputy Director, EMBL Australia Bioinformatics Resource (https://www.embl-abr.org.au/) commenced in February 2016. Previously, in March 2013 Vicky joined The Genome Analysis Centre (TGAC)’s Senior Management Team and head the 361° Division. In previous years Vicky was responsible for the strategic coordination and implementation of the EMBL-EBI’s User-Training program, providing training for the scientific users of EMBL-EBI’s data services. Prior to joining EMBL-EBI in 2007 Vicky held an Assistant Professor position at the University of Bern and at the Institute for Aquatic Sciences (EAWAG), with postdocs at the University of Zurich and University of Rome (Torvergata).