EMBL-ABR Node Application Form: AGRF

Table of Contents

Part A: Summary of EMBL-ABR Node ................................................................. 2
   A1: Synopsis ........................................................................................................ 2
   A2: Snapshot of services/activities ............................................................... 2

Part B: Administrative details .......................................................................... 2
   B1: Organisation(s) involved in the Node .................................................... 2
   B2: Contact details of Head of Node ............................................................. 3

Part C: Operational description of Node ......................................................... 4
   C1: Description of services/activities to be provided .................................. 4
   C2: Description of those involved, what their proposed role is for this activity and their relevant experience ................................................................. 5
   C3: Service provision tables ......................................................................... 6
   C4: Node Impact .............................................................................................. 6
   C4: Node Impact .............................................................................................. 8

11th July 2016
Part A: Summary of EMBL-ABR Node

A1: Synopsis

*Please describe the aim, structure and domain focus of this node; the services/activities it is already providing, and the services/activities it will provide with EMBL-ABR Hub support*

This node focuses on three key areas: tools, training and compute. It will center on Training activities targeted to life scientists working on non model organisms. Within Tools it will be an active participant in activities dealing with sharing databases, tools and pipelines developed at AGRF and via collaborations. In Compute it will concentrate its efforts towards cutting-down data transaction time and human errors by pushing the data directly to a high performance computing environment.

A2: Snapshot of services/activities

*Please indicate the services/activities to be provided by the Node (indicate all appropriate)*

<table>
<thead>
<tr>
<th>Data:</th>
<th>Tools:X</th>
<th>Training:X</th>
<th>Platforms:</th>
<th>Compute:X</th>
</tr>
</thead>
</table>

Part B: Administrative details

B1: Organisation(s) involved in the Node

- *Please list the people involved in the Node.*

<table>
<thead>
<tr>
<th>Who is involved from this node (name, Affiliation, e-mail)</th>
<th>Who is involved from other nodes in the activities involved (name, affiliation, e-mail)</th>
</tr>
</thead>
</table>
| Dr Sonika Tyagi  
Bioinformatics Supervisor  
sonika.tyagi@agrf.org.au  
Australian Genome Research Facility Ltd.  
Melbourne | Simon Gladman, EMBL-ABR@VLSCI node |
B2: Contact details of Head of Node

Name: Dr. Sonika Tyagi
Job Title: Bioinformatics Supervisor
Organisation: Australian Genome Research Facility Ltd.

E-mail: sonika.tyagi@agrf.org.au
Telephone: Phone: +61 3 9345 2664
Part C: Operational description of Node

C1: Description of services/activities to be provided

Please provide details for the Activity proposed for each of the EMBL-ABR Focus areas (Data, Tools, Platforms, Compute, Training) it covers, including a title, description, target users and standards. Also mentioned desired EMBL-ABR support input and/or development where relevant. Please describe each area separate if you are cover more than one using the following categories.

Focus Area: Training
Title: non-model organisms bioinformatics focused training
Description:
How it will be provided: In collaboration with the EMBL-ABR hub and using the templates and best practice guidelines tutorials, we will develop interactive materials that will be available online as well as for face-to-face training. It will also form part of the core “Training & Learning” program from EMBL-ABR.
Who the users will be: The focus user-group are ECRs among AGRF clients who are working with non-model organisms. Its scalability to national level will be derived from the production of of materials to be placed online for other nodes to adapt and use. Our participation in Face to face course will be harmonised with EMBL-ABR Hub efforts which will follow a cost-recovery participation fee.
Standards: We will follow goblet and other state of the art initiatives. For example, a training material for genome level assemblies has already been developed in collaboration with the TGAC UK.

Focus Area: Compute
Title: access to high performance compute to whole genome level studies
Description:
How it will be provided: Ideally we would like efficient access to data generated by high throughput machines outside of AGRF, eg. collaborative data generated using Garvan’s HiSeqX system, or PacBio Sequel at WEHI, for downstream analysis. Currently, the data is moved from the data generation centre to AGRF for further analysis on portable drives. The data is then moved to AGRF servers for analysis, and sent back to the client either via the FTP server or again on portable drives. As a node of EMBL-ABR, we could cut-down data transaction time and human errors by pushing the data directly to a high performance computing environment at the EMBL-ABR hub or a node specialized in this from the data generation center. At the same time we can increase both the speed and efficiency of the “data-to-knowledge generation process” by running analysis on the EMBL-ABR servers. Eventually, life scientists will be able to access their analyzed data and perform further data mining; most likely using graphical user interface enabled platform tools hosted on the hub, and then submit data directly to public databases (eg EMBL-EBI).
Who the users will be: AGRF clients working with whole genome sequencing/re-sequencing. Analysis can be done using tools and pipeline set up in EMBL-ABR tools resources. The national benefit of this activity is the actual development of the workflow to solve high volumes of data transfer which can per se be adopted by other EMBL-ABR nodes in Australia.
Standards: We will follow peer reviewed best practice standards to analyze and manage the data. Per user quote and access time limits might apply.

Focus Area: Tools
Title: To share databases, tools and pipelines developed at AGRF and via collaborations
Description: AGRF has a dynamic environments in terms of needs for data analysis and interpretation. This requires building tailored pipelines, and sometimes specific tools/extensions to
existing tools. We would welcome to exchange knowledge and experience in this across EMBL-ABR network. We also would like to be part of EMBL-ABR efforts towards a sustainable environment for Australia’s produced tools/pipelines.

**How it will be provided:** Stand-alone tools as binary version or using a version controlled system (such as github/bitbucket) following best practice protocols. At the moment there is more than one type of tool and database available for any given bioinformatics task. The most relevant tools for the Australian life sciences community connected with the EMBL-ABR, and importantly, any tools developed and maintained by Australian researchers (e.g., GBS data analysis pipeline, Metagenomics pipelines) will be made available via EMBL-ABR network. Best practice workflow and technical support will also be provided. Hosting of integrative, multi-omics, platforms tailored to Australia-specific research programs will follow.

**Who the users will be:** Obviously AGRF clients, and of course relevant national and international audiences.

**Standards:** software development practices, detailed instructions for installation, testing and troubleshoot will be provided as well as sustainable maintenance plan.

---

**C2 Description of those involved, what their proposed role is for this activity and their relevant experience**

**NAME:** Sonika Tyagi

**Role:** Bioinformatics Trainer, Data Analyst, Researcher

**Responsibilities:** Training activities, liase on compute AGRF needs, Tools pipelines developers and adopters.

**Experience:** ST has PhD in Bioinformatics and more than 10 years of work experience in genomics, especially in statistical data analysis, developing and applying bioinformatics tools and pipelines. The tools and pipelines developed by ST are open source and are heavily used by the researchers. ST is currently a Bioinformatics Supervisor at the Australian Genome Research Facility Ltd. (AGRF) and specializes in RNAseq and DNAseq analysis services. In the current role ST has worked on >100 RNAseq and DNAseq analysis service projects and have provided consultancy starting from experiment design, sequencing and analysis requirements to generating publication ready data presentations. ST has also co-author 16 journal articles resulting from her work in the current role and previous projects involving high throughput sequencing data analysis of Cancer genomes, exomes and sequences of coding and non-coding RNAs. As bioinformatics expert ST is also involved in bioinformatics teaching and training programs working in collaboration with CSIRO, Bioplatforms Australia, EMBL-EBI and TGAC UK. ST has developed and delivered modules at over 25 hands-on NGS bioinformatics trainings on topics varying from NGS Quality Control, RNAseq, ChIPSeq, Cancer genomics and denovo assembly at various Australian universities and Research Institutes.
C3: Service provision tables

Domain area of services/activities to be provided by the Node

- assign a number from 1 to 5 in order of priority for each service/activity in each knowledge domain, with 1 indicating the most relevant and 5 being the least.
- Use N/A if field is not applicable

<table>
<thead>
<tr>
<th></th>
<th>Biology</th>
<th>Health</th>
<th>Food</th>
<th>Environment</th>
<th>Soil &amp; Water</th>
<th>Other (please specify)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DATA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOOLS</td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>TRAINING</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PLATFORMS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COMPUTE</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

C4: Node Impact

Briefly describe the expected impact of this node at the local, national, and international level. Include 1-2 sentences describing the added value for Australia in having this bioinformatics node.

Scientific and technical impact and how will this be assessed (State, National and international):
AGRF has an existing national reach in Australia, which will allow rapid dissemination of information and uptake by suitable researchers. Regular surveys, feedback programs and peer review are conducted under corporate ISO 17025:2005 accreditation, which will be extended to encompass review of these resources. In parallel, metrics such as usage statistics for tools and adoption of shared workflows and pipelines will also be collected with support of EMBL-ABR Hub.

A positive impact on the national bioinformatics community:
Bioinformatics is an interdisciplinary field and has become essential to unravel complex systems in plant, animal, environment and human health research. The high throughput technology to generate the biological data is evolving at a fast pace and bioinformatics skills to cope with the new surge of data are in great demand. There is national dearth of skilled bioinformatics to keep pace with the growing amount of biological data in the Australian research community. One way to cope with this shortage is by empowering life scientists to take ownership of the analysis of their own data through improved knowledge and understanding of bioinformatic approaches. This will not only enable wet lab scientists to re-connect with their data, but will also help them better communicate their requirements with bioinformaticians. At the same time, bioinformaticians will also benefit from training in areas of their inexperience and greater understanding of biology.
Whole genome sequencing of higher animals and plants requires ultra-deep sequencing for the downstream analysis. The movement of big data from sequencing centres to analysis servers, and eventually to users, is a big challenge at the moment. A central and secure analysis space, equipped with best-practice tools and pipelines, where data can be directly posted via secure ftp protocols, will enable fast data transfer, knowledge extraction from the data and easy access to the end user. The analysis and resulting biological knowledge in the form of new assemblies, annotations etc. can benefit from the easy central access to public databases and tools via the EMBL-ABR resources. The users can also directly submit data to these public repositories such as ENA at EMBL-EBI.

The Australian added value of the proposed Activity:

AGRF has been involved in biological data generation and research programs involving Australia animals, crop-plants, bacterial, fungi and corals, soil microbiomes and melanoma sequencing in collaboration with the Bioplatforms Australia. New data integration platform development initiatives are already in progress.

Why should this be done in Australia with EMBL-ABR
[Please describe any other similar efforts occurring nationally or internationally, how this activity differs and why it is important to do it now and in Australia]

In general research data management is a big challenge for many Australian life scientists. A central, secure, authorized data storage, management and access with archiving and sharing facility is the demand of the hour for Australian life science community. Organizations such as Bioplatforms Australia, have initiated a number of national level collaborative projects and the Australia specific data repositories are being developed and some are already accessible to the end user. AGRF is one of the partners on these projects. Such data complemented with required compute, tools and platforms require a national level collaboration to extract the valuable biology from these datasets. There has been a great effort by various groups to run nation-wide bioinformatics trainings focused mainly at wet-lab scientist. Some of the new data generation platforms such as PacBio and Oxford Nanopore are being used more often (Illumina was the basis of most of the training material and tools previously) and the newer sequencers are being deployed by AGRF independently or in collaboration with research institutes (for example latest 10X addition with University of Melbourne). We will have to adapt to the changing data types and develop and deliver training modules to meet the new specific requirements. A set up similar to ELIXIR in Europe, would be a great example to follow in order to address specific challenges, user requirements and technical support needed by the Australian life scientist.
C4: Node Impact

Briefly describe the expected impact of this node at the local, national, and international level. Include 1-2 sentences describing the added value for Australia in having this bioinformatics node.

Scientific and technical impact and how will this be assessed (State, National and international):
Regular survey, feedback programs and peer review will assess node impact.

A positive impact on the national bioinformatics community:

Bioinformatics is an interdisciplinary field and has become essential to unravel complex systems in plant, animal, environment and human health research. The high throughput technology to generate the biological data is evolving at a fast pace and Bioinformatics skills to cope with the new surge of data is in great demand. There is dearth of skilled bioinformaticians to keep pace with the growing amount of biological data in Australian research community. One way to cope with this shortage is by empowering life scientists to take ownership of the analysis of their own data through improved knowledge and understanding of bioinformatic approaches. This will not only enable wet lab scientists to re-connect with their data but will also help them better communicate their requirements with their fellow bioinformaticians. At the same time bioinformaticians will also benefit from training in out of their expertise as well as by understanding the biology side in greater depth.

Whole genome sequencing of higher organisms and plant requires greater depth for the downstream analysis. The movement of the large size of the data from sequencing centre to analysis server and eventually to user end is a big challenge at the moment. A central, more secure, analysis server with already installed best practice tools and pipelines where data can be directly posted via secure ftp protocols will be enable fast data transfer, knowledge extraction from the data and easy access to
The analysis and resulting biological knowledge in the form of new assemblies, annotations etc. can benefit by the easy central access to all public databases and tools via the EMBL-ABR resources.

The Australian added value of the proposed Activity:

AGRF has been involved in biological data generation and research programs involving Australian animals, bacterial strains, soil microbiome profiling, melanoma sequencing in collaboration with Bioplatforms Australia. A few collaborative data integration platform development initiatives are already in progress.

Why should this be done in Australia with EMBL-ABR
[Please describe any other similar efforts occurring nationally or internationally, how this activity differs and why it is important to do it now and in Australia]

In general research data management is a big challenge for many Australian life scientists. A central, secure, authorized data storage, management and access, archiving and sharing facility is the demand of the hour for Australian life science community. Organizations such as Bioplatforms Australia, have initiated a number of national level collaborative projects and the Australia specific data repositories are being developed and some are already accessible to the end user. Such data complements with required compute, tools and platforms require a national level collaboration to extract the valuable biology from these datasets. There has been a great effort by various groups to run national wide bioinformatics trainings focused mainly at wet-lab scientist. Some of the new data generation platforms such as Pacbio, Oxford Nanopore are being used more often (Illumina was the basis of most of the training material and tools previously) and newer sequencer such as 10X has already arrived in Australia. We will have to adapt to the changing data types and develop and deliver training modules meet the new specific requirements. A set up similar to ELIXIR in Europe would be a great example to follow to address specific challenges, user requirements and technical support needed by the Australian life scientist.