iReceptor – A case study in the challenges/opportunities in Canadian DRI

Felix Breden, Scientific Director, iReceptor, SFU and Brian Corrie, Technical Director, iReceptor, SFU

The new Canadian Digital Research Infrastructure (DRI) organization consists of three pillars, Research Data Management (**RDM**), Advanced Research Computing (**ARC**), and Research Software (**RS**). We would agree that all of these pillars are critically important to modern research. To date, in Canada, funding for the development of research projects as well as obtaining support for projects across the pillars has been disjointed. Each of the organizations that provide funding and support in one or more of **RDM**, **ARC**, and **RS** pillars have been successful, and yet the disjointed nature of this funding and support makes it challenging for researchers to navigate the DRI ecosystem. The pieces of the puzzle are there but combining them into a coherent picture is challenging, even for researchers who are technically sophisticated. For those that are new to DRI this is a daunting task! NDRIO is in a unique position to coalesce and unify the DRI landscape in Canada and to broaden the DRI community.

In this white paper we use the development of the iReceptor Platform from 2014 – 2020 as a use case to consider the challenges faced by researchers undertaking the development of a complex RS project in the current Canadian DRI landscape as well as the challenges and opportunities faced by NDRIO in supporting DRI in Canada.

iReceptor as a case study

Next-generation sequencing (NGS) allows the characterization of the adaptive immune receptor repertoire (B-cell and T-cell receptor sequences, or AIRR) in exquisite detail. These large-scale AIRR-seq data sets have rapidly become critical to vaccine development, understanding the immune response in autoimmune and infectious disease, and monitoring novel therapeutics against cancer. Since 2014, a grass roots, international community (the AIRR Community - www.airr-community.org)[1] has been working towards establishing standards and recommendations for obtaining, analyzing, curating, comparing, and sharing NGS AIRR-seq data sets [2][3][4]. Using these standards, the AIRR Community Working Groups have established an international network of AIRR-seq repositories (the AIRR Data Commons) that contain AIRR-seq data that are findable, accessible, interoperable, and reusable (FAIR)[5]. The AIRR Community was co-founded by SFU's iReceptor co-PIs Dr. Felix Breden and Dr. Jamie Scott (the first AIRR Meeting was held at SFU[1]), and the iReceptor team is extremely active in leading and participating in the standards developments of the AIRR Community.

The iReceptor Data Integration Platform (iReceptor)[6] provides an implementation of the AIRR Data Commons envisioned by the AIRR Community. iReceptor is a Distributed Data Management system and Scientific Gateway[2] for mining and analyzing "Next Generation" sequence data from immune responses. The main goal of iReceptor is to lower the barrier to immune genetics researchers who need to federate large, distributed, AIRR-seq data sets to answer complex questions about the immune response.

iReceptor and RDM

As a distributed data management system with the goal of making AIRR-seq data FAIR[7], iReceptor is an **RDM** platform. iReceptor provides three levels of **RDM** infrastructure. Firstly, iReceptor has developed

[1] https://www.antibodysociety.org/the-airr-community/meetings/airr-community-meeting-i/
[2] https://gateway.ireceptor.org

the iReceptor Turnkey repository[3], a docker based AIRR standards compliant repository software stack that allows researchers to download, install, curate, and share their own AIRR-seq data. Secondly, iReceptor maintains and operates two AIRR-seq data repositories as part of the AIRR Data Commons, the iReceptor Public Archive (IPA) and AIRR COVID-19 repositories. These repositories utilize a resource allocation on Compute Canada's *Arbutus* open stack cloud platform and the iReceptor Turnkey software to deliver these services. Thirdly, the iReceptor Gateway is a web interface that allows researchers to find, federate, analyze, and reuse AIRR-seq data from across the entire AIRR Data Commons. Again, the iReceptor Gateway runs on the Compute Canada *Arbutus* cloud platform. The iReceptor platform has benefited from CANARIE's **RDM** funding program[4], with a currently active **RDM** project underway.

iReceptor and ARC

As a platform for finding, federating, and analyzing AIRR-seq data, iReceptor relies heavily on **ARC** infrastructure for its success. As described above, iReceptor makes extensive utilization of its cloud-based resource allocation on Compute Canada to operate its repository and web services. In addition, iReceptor has both a storage and compute allocation on Compute Canada's *Cedar* platform. The storage allocation is used for archival and active storage of the data that is curated in the iReceptor repositories on *Arbutus.* The compute allocation is currently used primarily for running the data through the iReceptor bioinformatics curation pipeline.

Through more recent developments through the iReceptor Plus project,[5] a joint CIHR/EU Horizon 2020 initiative, we anticipate being able to run large, complex analysis jobs on behalf of our international user community at one of several large computation centres with Compute Canada's *Cedar* being one of them (the other anticipated platforms being the Texas Advanced Computation Centre (TACC) and the European Open Science Cloud (EOSC)). iReceptor also relies on the **RS** bioinformatics software stack maintained by the Compute Canada **ARC** team. iReceptor has benefited from Compute Canada resource allocations from its inception, both for cloud compute, traditional compute, and storage. iReceptor has also benefited from the **ARC** focussed CFI Cyberinfrastructure Platform funding competition in 2015[6].

iReceptor and RS

As a platform that has been envisioned, developed, and maintained in Canada, iReceptor is a complex Canadian **RS** project. iReceptor has developed a full docker based AIRR compliant software stack, implementing a range of web services to query its repositories, and has developed a widely used web based Scientific Gateway. Each of these components is continuously updated as the AIRR standards change and as we learn more about the needs of our user community.

The iReceptor project has been fortunate in that it was able to establish a Research Software Engineering (**RSE**) team in 2014 and has been able to maintain that core team throughout the project. iReceptor was established in 2014 at the Center for Interdisciplinary Research in the Mathematical and Computational Sciences (IRMACS)[7], which offered **RS** technical support and access to **RSE** personnel. The ability to "share" **RSEs** with other projects at the IRMACS Centre was an indispensable catalyst for the project. iReceptor benefited from two rounds of funding from CANARIE **RS** funding programs, with the

[3] https://github.com/sfu-ireceptor/turnkey-service-php
[4] https://www.canarie.ca/canarie-awards-2m-to-research-teams-to-extend-the-interoperability-of-research-data/
[5] https://cordis.europa.eu/project/id/825821
[6] https://www.innovation.ca/about/press-release/government-canada-supports-data-rich-research
[7] http://www.sfu.ca/irmacs-archive/irmacs/infrastructure/gateways.html

first round of funding in 2014 critical to the establishment of the project. iReceptor has also utilized the CFI Cyberinfrastructure Platforms funding, and more recently both CANARIE **RDM** funding and CIHR/EU Horizon 2020 funding to maintain and continue to develop the iReceptor **RS** platform.

iReceptor Status

iReceptor as a platform has been in production since 2014. The AIRR Data Commons has grown from just over 500 million sequence annotations at the end of 2018 to over 4 billion sequence annotations from 60 different research groups/studies. Enabled by the existence of this platform, the iReceptor team (in collaboration with the AIRR Community) has focussed on curating and storing COVID-19 AIRR-seq data. As of December 2020, over 1 billion annotated sequences from 17 studies on COVID-19 patients[8] are available for search, exploration, and download through the iReceptor Gateway. To our knowledge the AIRR Data Commons contains all of the publicly available AIRR-seq COVID-19 data.

iReceptor is an actively used platform with over 400 user accounts with 90 users added in 2019. We have seen a dramatic increase in use as a result of the availability of COVID-19 data through the iReceptor platform. We have added over 250 new users in 2020 with over 100 users added in July and August alone. This was primarily a result of prominent COVID-19 studies/publications citing iReceptor as the place where either their data was available [9][10], or perhaps more importantly, where they found, accessed, and reused (FAIR) COVID-19 AIRR-seq in their research [11]. The iReceptor user community spans most continents as well as academia and industry. Of the over 200 new users from April – October 2020, 36 are from industry, 59 from Europe, 56 from North America, 27 from Asia, 8 from government agencies, and the remainder from Australia, South America, Africa, and non-profit organizations. Over 350 GB of AIRR-seq data has been downloaded through the platform for reuse by researchers since June 2020, again the bulk of these downloads driven by the availability of COVID-19 data.

The Good News

The iReceptor project has been fortunate in that the requirements to move from a concept to an active and successful platform have been available in the currently existing DRI ecosystem in Canada. The **ARC** resources required to operate and deliver its services have been available from Compute Canada, the IRMACS Centre at SFU, and SFU's research support services. **RDM** has always been a mature field in Canada, driven initially by Canadian Research Libraries (CARL) and more recently by Research Data Canada (RDC) at the national level. We were fortunate to have started our project at the IRMACS Centre where **RDM** was a central component of what SFU (as an institution - through the SFU Library) and the IRMACS Centre (as a research institute) supported[8]. Last, but certainly not least, we have managed to find, build, and maintain an **RS** team throughout the duration of the iReceptor project. Building and sustaining such a team requires funding and support. Starting with an initial CANARIE **RS** project and support from the IRMACS Centre, we have continued our efforts through a broad range of funding programs, including several CANARIE **RS** and **RDM** projects, a CFI Cyberinfrastructure **RS** Platform project, and a CIHR/EU Horizon 2020 initiative.

The Challenges and Opportunities

As suggested above, we believe that Canada is in a relatively good position from the perspective of having many of the required pieces of the DRI puzzle – the main challenge for any given researcher is

[8] http://www.sfu.ca/irmacs-archive/irmacs/infrastructure/datamanagement.html

finding the pieces of the puzzle that they need and putting them together to form the picture that they require to accomplish their research. In the current Canadian context, it is quite challenging to navigate the DRI ecosystem. Given that today almost all research is data driven, the breadth of the research community that needs to be supported is vast. This is both the opportunity and the challenge for NDRIO.

Defining NDRIO's scope

Although there is no right or wrong answer to what the NDRIO scope is, a critical challenge for NDRIO will be a concise definition of that scope. What role does NDRIO play and who and how does it partner with the other players (institutions, government agencies, funding bodies) to deliver on its DRI mandate. How far does NDRIO go towards being all things to all researchers related to DRI – or another way of looking at it, what is the line between what NDRIO does versus what the institutions do? Having a clear scope will benefit the entire DRI community.

- **Challenge**: Define very clearly the scope that NDRIO covers in the DRI landscape.
- **Opportunity**: Work with other institutions to provide top-to-bottom support for DRI in Canada.

NDRIO as a funder

Many of the roles of the organizations that NDRIO is replacing/subsuming/incorporating provide both infrastructure and support that researchers rely on as well as direct funding towards research programs. The funding landscape for researchers to use DRI in their research is complicated. NDRIO has an opportunity to revisit the DRI funding landscape to provide clarity around future funding, ensure that critical existing programs continue, as well as address any gaps that exist in the current landscape. This landscape is complicated, including infrastructure and operating funding (e.g. CFI), support for operations personnel (e.g. NSERC MSI), funding to the national research data infrastructures (e.g. CARL, CANARIE RDM), and funding for developing research software platforms (e.g. CANARIE RS and CFI Cyberinfrastructure). All of these are funding activities that are critical to the DRI ecosystem but fall outside traditional Tri-Council funding. With the consolidation of **RDM**, **ARC**, and **RS** within NDRIO, there is an opportunity for NDRIO to explore ways to better support national platforms that need to be sustained over long periods of time such as **RDM** platforms (e.g. FRDR) and domain specific **RS** platforms (e.g. iReceptor, Ocean Networks Canada, CANFAR, CBrain, GenAP, etc. [9]). Such platforms are critical to the DRI ecosystem but are difficult to sustain in our current funding landscape.

- **Challenge**: Consider the full spectrum of DRI funding and ensure that important funding streams that currently exist are not lost in the creation of NDRIO. Carefully consider continuity for the funding programs that currently exist and the platforms and projects they support.
- **Opportunity**: Consider the existing organizations that are being merged and identify any funding gaps that exist that the creation of NDRIO as an organization can fill.

NDRIO as a Service Provider

The range of services that NDRIO as an organization could fill are extensive given the breadth of DRI. These encompass all the services that the existing **ARC**, **RDM**, and **RS** organizations provide as well as any services that are missing in the gaps between these organizations. In addition, NDRIO as an organization can choose to focus on providing substantial support for a narrow set of services or

[9] https://www.oceannetworks.ca/, https://www.canfar.net/en/, http://www.cbrain.ca/index.html, https://genap.ca/

broader support for a wider range of services. This will be a challenging balance to strike. Researchers' know their research domain but will have varying experience with DRI. NDRIO has an opportunity to build a DRI community, helping researchers at all levels of the DRI ecosystem to determine an appropriate DRI solution and guiding and supporting them in utilizing or developing that solution.

- Challenge: Be practical and pragmatic in determining what services can and should be supported by NDRIO within the DRI ecosystem.
- Opportunity: Build a DRI community within Canada. Focus service development on services that are likely to maximize the efficient use of the entire DRI ecosystem – help the researcher find the DRI pieces of the puzzle and assemble them into a solution that makes a difference.

The iReceptor Team looks forward to working with NDRIO in building such a DRI community in Canada!

References

[1] F. Breden *et al.*, "Reproducibility and Reuse of Adaptive Immune Receptor Repertoire Data," *Front. Immunol.*, vol. 8, p. 1418, Nov. 2017.

[2] F. Rubelt *et al.*, "Adaptive Immune Receptor Repertoire Community recommendations for sharing immune-repertoire sequencing data," *Nat. Immunol.*, vol. 18, no. 12, 2017.

[3] J. A. Vander Heiden *et al.*, "AIRR Community Standardized Representations for Annotated Immune Repertoires.," *Front. Immunol.*, vol. 9, p. 2206, 2018.

[4] S. Christley *et al.*, "The ADC API: a web API for the programmatic query of the AIRR Data Commons," *Front. Immunol.*

[5] M. D. Wilkinson *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship," *Sci. Data*, vol. 3, p. 160018, Mar. 2016.

[6] B. D. Corrie *et al.*, "iReceptor: A platform for querying and analyzing antibody/B-cell and T-cell receptor repertoire data across federated repositories," *Immunol. Rev.*, vol. 284, no. 1, pp. 24–41, Jul. 2018.

[7] B. Corrie *et al.*, "iReceptor: A case study in the importance of standards for data sharing," *Scientific Gateways Conference 2019*, 2019. [Online]. Available: https://osf.io/2f98g/. [Accessed: 17-Oct-2019].

[8] J. K. Scott and F. Breden, "The adaptive immune receptor repertoire community as a model for FAIR stewardship of big immunology data," *Current Opinion in Systems Biology*, vol. 24. Elsevier Ltd, pp. 71–77, 01-Dec-2020.

[9] C. Schultheiß *et al.*, "Next-Generation Sequencing of T and B Cell Receptor Repertoires from COVID-19 Patients Showed Signatures Associated with Severity of Disease," *Immunity*, vol. 53, no. 2, pp. 442-455.e4, Aug. 2020.

[10] L. Paschold *et al.*, "SARS-CoV-2 specific antibody rearrangements in pre-pandemic immune repertoires of risk cohorts and COVID-19 patients," *J. Clin. Invest.*, Oct. 2020.

[11] P. Meysman, A. Postovskaya, N. De Neuter, B. Ogunjimi, and K. Laukens, "Tracking SARS-CoV-2 T cells with epitope-T-cell receptor recognition models," *bioRxiv*, p. 2020.09.09.289355, Sep. 2020.