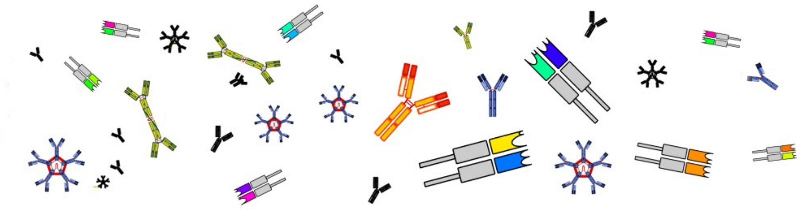


Deliverable 7.1

An updated MiAIRR standard

Work Package Number: 7

Work Package Title: Single cell data integration

Type: RTD

This project is funded by the European Union’s H2020 Research and Innovation Programme under Grant Agreement No. 825821 and Canadian Institutes of Health Research (CIHR).

Document Information

iReceptor Plus Project Information	
Project full title	Architecture and Tools for the Query of Antibody and T-cell Receptor Sequencing Data Repositories for Enabling Improved Personalized Medicine and Immunotherapy
Project acronym	iReceptor Plus
Grant agreement number	825821
Project coordinator	Prof. Gur Yaari
Project start date and duration	1st January, 2019, 48 months
Project website	https://www.ireceptor-plus.com

Deliverable Information	
Work package number	7
Work package title	Single cell data integration
Deliverable number	7.1
Deliverable title	An updated MiAIRR standard
Description	An updated MiAIRR standard containing a proper representation of single-cell AIRR-seq data and a corresponding API supporting single-cell queries. (M06, draft; M18, finalized; M42, revised)
Lead beneficiary	DKFZ
Lead Author(s)	Christian Busse
Contributor(s)	n/a
Revision number	1
Revision Date	2019-05-31
Status (Final (F), Draft (D), Revised Draft (RV))	D

Dissemination level (Public (PU), Restricted to other program participants (PP), Restricted to a group specified by the consortium (RE), Confidential for consortium members only (CO))	PU

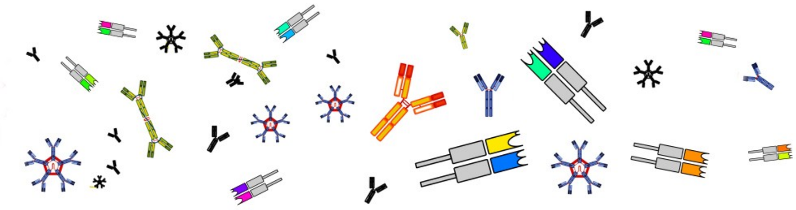
Document History			
Revision	**Date**	**Modification**	**Author**
v1	2019-05-31		CEB
v2	2019-06-16	Minor changes	BDC, FB
v3	2019-06-24	Review	EMF
final	2019-06-28	Merged changes and feedback	CEB

Approvals				
	Name	**Organisation**	**Date**	**Signature (initials)**
Coordinator	Dr. Gur Yaari	Bar Ilan University		
WP Leaders	Dr. Christian Busse	DKFZ	2019-06-28	CEB

This project is funded by the European Union's H2020 Research and Innovation Programme under Grant Agreement No. 825821 and Canadian Institutes of Health Research (CIHR).

Table of Contents

This project is funded by the European Union’s H2020 Research and Innovation Programme under Grant Agreement No. 825821 and Canadian Institutes of Health Research (CIHR).



Executive Summary

Heterogeneity of B and T cell populations is a well-known challenge in immunology and the main reason for the high relevance that single-cell resolution techniques have in the field. Most AIRR-seq data however, are derived from bulk populations of 10^5 to 10^7 cells and therefore do not allow us to decompose this underlying complexity. While various single-cell AIRR-seq workflows have been used in academic settings since 2013, they were not widely adopted due to their increased complexity. With the recent commercial availability of microfluidic single-cell encapsulation devices by various manufacturers, this situation is expected to change substantially within the next years.

The strength of single-cell workflows is based on their ability to use the individual cell as an "atomic" unit of the immune system and associate multiple data types with it:

1) AIRR-seq data for both Ig/TCR chains that make up the receptor heterodimer, thus providing chain-linkage information and complete receptor information.
2) Gene expression data describing the cell phenotype.
3) Experimental measurements for individual Ig/TCR reactivity.

The linkage of these three data types provides comprehensive information on the receptor and cell identity, thereby facilitating powerful and detailed analyses of phenotype-genotype-functionality correlations.

This deliverable will provide a community-approved extension to the existing MiAIRR standard that facilitates the annotation and storage of single-cell related information. In addition, it will provide an API to be implemented by iReceptor Plus repositories that allows the access to and use of this information.

This project is funded by the European Union’s H2020 Research and Innovation Programme under Grant Agreement No. 825821 and Canadian Institutes of Health Research (CIHR).

Deliverable description

Adaptive Immune Receptor Repertoire (AIRR) sequencing experiments are typically conducted in a "bulk" setting, i.e. the genes or transcripts encoding the immunoglobulin (Ig) or T cell receptor (TCR) chains are sequenced from samples containing 10^4 to 10^7 B or T cells, respectively. It is important to note that the actual antigen receptors are always heterodimers (e.g. the TCR containing one TCRα and one TCRβ chain) and both chains contribute to the specificity and sensitivity of antigen recognition. Therefore the loss of this pairing information (i.e. which chains were associated within a single cell) in bulk experiments is also the loss of the capability to infer the reactivity or even to reconstruct the receptor using recombinant expression systems.

Nevertheless, due to their simplicity, bulk AIRR-seq is still the most commonly used experimental setup, therefore the standardization efforts of the AIRR Community have until now focused on developing recommendations for metadata annotation (MiAIRR[1]) and more recently also for a uniform representation of the AIRR-seq data itself (DataRep[2]) with primarily such bulk setups in mind.

However, over the last years the use of experimental techniques providing single-cell resolution has seen a constant increase, mainly due to the off-the-shelf availability of microfluidic platforms. Therefore, the goal of WP7 is to allow iReceptor Plus to store and query single-cell data and harness the unique properties associated with it.

Deliverable 7.1 builds the foundations for this goal and standardizes the type of single-cell information stored by a repository and the ways this information can be accessed (i.e. queried upon or retrieved) by external applications. The work on this will proceed via a number of steps:

- *Storage*: The minimal information standard on single-cell experiments needs to facilitate the storage of cell phenotype data and Ig/TCR reactivity. While there is currently no common standard that would address these points, the MiAIRR data standard does provide a framework for the common levels of metadata (study, subject, sample, processing). Therefore we are developing a single-cell extension (hereafter: scXT) of the MiAIRR data standard that will

[1] Nat Immunol 18:1274 (2017) doi: 10.1038/ni.3873

[2] Front Immunol 9:2206 (2018) doi: 10.3389/fimmu.2018.02206

allow the storage of these data types. This development will proceed in three steps: a draft that is feature-complete **(M06, D7.1)**, a tested final version (M18, D7.1) and – if required – a revised version after two years (M42, D7.1).

- *Access*: MiAIRR scXT will only specify which information an application can expect to find in a repository, not how to access it. However, standards for access are required to ensure interoperability between a repository and an application as well as between connected repositories. Therefore we will also develop a REST API that will allow external access to single-cell information. This development will proceed in three steps: A draft of the API based on MiAIRR scXT (M12, no deliverable), the finalized specification (M18, D7.1) and the implementation into the repository framework (M24, D2.2).

The coordinated development of these standards (both scXT and the API) with other stakeholders of the AIRR Community is paramount for their adoption by the wider community. This avoids parallel – and potentially incompatible – developments that could result in iReceptor Plus becoming an insular solution (“silo”).

The initial approach to the scXT was based on the following three elements, of which (2) and (3) are the essential novel components of the scXT, while (1) maps to the existing MiAIRR key `cell_id`:

1) A field that holds a unique identifier that encodes chain pairing information and represents an individual cell (hereafter: UID_{cell})
2) A record holding flow cytometric information referring to the UID_{cell}.
3) A record holding reactivity information referring to the UID_{cell}.

However, based on consultations with the AIRR Community it became clear that a specification focusing only on these items would be too narrow and only sufficient for a limited set of technology platforms. The three central aspects lacking would be:

a) While flow cytometry remains an important platform for single-cell isolation and combined phenotyping, there are other techniques like cite-Seq and single-cell transcriptomes that are increasingly used. Therefore the scXT should be able to represent this type of data.
b) A number of experimental platforms allow to the inference of pairing of the receptor chains that are associated on the clonal level from bulk populations. While this approach does have restrictions based on its stochastic nature and

the requirement for limiting-dilution setups, it is occasionally used in diagnostic approaches. Therefore the annotation of paired chains outside of a single-cell context – as it would be suggested by the use of a UID_{cell} – should be supported.

c) While the classical assumption in immunology is that a single B or T cell always and only expresses a single type of receptor, there are numerous examples of cells with at least two different receptor species. The most common example for this are T cells that due to incomplete allelic exclusion can contain two functionally rearranged *TRA* loci, which upon expression give rise to two TCRα chains and hence two different TCRs. Similarly, the dual expression of Igκ and Igλ light chains has been observed in subpopulations of the B cell lineage. Also in this case the encoding of chain pairing via the association of chains with a single cell is not a clean solution.

To address these concerns for the planed scXT, the originally stated elements were extended as follows:

1) An identifier to indicate association of two chains in a single receptor (UID_{rec}) independent of the method used to determine association.
2) An identifier to indicate association of individual chains with a cell of origin (UID_{cell}).
3) A record holding gene expression information referring to the UID_{cell}. The record is structured in a technology-agnostic manner, so that it can hold e.g. either flow cytometric data or a single-cell transcriptome profile.
4) A record holding reactivity information referring to the UID_{rec}.

It should be noted that scXT assumes that (1) and (2) are connected to an individual sequence (i.e. chain) and that the extension does neither specify nor require an intermediate structure mapping UID_{rec} to UID_{cell}. The main motivation behind this is lower complexity, which is important for general-purpose repositories (e.g. INSDC) to hold this information. However, dedicated AIRR-seq repositories can choose to implement such structure within their schema to speed-up access.

With these additions, the scXT draft has now been ratified as a provisional standard by the AIRR Community on 2019-05-15 during its IV. Community Meeting in Genoa, Italy. The full specification is available here.

Summary and future work

The AIRR Standards single-cell extension (AIRR-scXT) provides a comprehensive framework to annotate the most commonly used aspects of single-cell AIRR-seq experiments. With the approval of AIRR-scXT as a provisional standard by the AIRR Community, we will now start working on designing the API to query the information provided by these fields. The API will receive its first implementation in sciReptor and will provide access to its SQL database backend. This will provide testing for the suitability of AIRR-scXT as well as the first step towards Deliverable 7.4. The finalized AIRR-scXT will be then submitted to the responsible AIRR Community committees in March next year, for the adoption as final standard. In parallel, also the API will be submitted to the Common Repository Working Group for review and approval.

In addition to this, we are working with the AIRR Community Minimal Standards Working Group to include the scXT in an upcoming manuscript, to achieve even broader dissemination in the scientific community.

This project is funded by the European Union’s H2020 Research and Innovation Programme under Grant Agreement No. 825821 and Canadian Institutes of Health Research (CIHR).