***DRAFT: Overview: Recommendations from the Minimal Standards Working Group for High Throughput Adaptive Immune Receptor Repertoire (AIRR) Sequencing Studies

Florian Rubelt, Christian Busse, Danny Douek, Lindsay Cowell, Uri Hershberg, Marie-Paule LeFranc, Brian Corrie, Ahmad Chan, Chaim Schramm, Bojan Zimonja, Jerome Jaglale, Chris Murawsky, Nina Luning Prak^, Steven H. Kleinstein^
^Minimal Standards Working Group co-chairs

Overview. As high throughput experiments become more prevalent in the field of Immunology and elsewhere, there is an increased need for collective organization of data and standardized methods of data reporting. No current standards exist for adaptive immune-receptor repertoire (**AIRR**) sequencing data. Data and metadata formats need to be harmonized so that data from different experiments can be mined. Once recovered, the mined data need to have sufficient descriptive metadata in order to be useful. To fulfill these unmet needs, the Minimal Standards Working Group (**MSWG**) proposes a set of minimal standards that we recommend journals adopt, and that could form the requirements for submission to a public data repository. These standards consist of the following six critical elements:

1. The experimental study design including sample data relationships (*e.g.*, which raw data file(s) relate to which sample, which samples are technical, which are biological replicates).
2. The essential sample annotation including experimental factors and their values (*e.g.*, the set of markers used to sort the cell population being studied).
3. Sufficient annotation of the amplicon being sequenced that would allow the raw data to be transformed into the processed sequences (*e.g.*, barcodes, primers, unique molecular identifiers).
4. The raw data for each sequencing run (*e.g.*, FASTQ)
5. The essential laboratory and data processing protocols (*e.g.,* software tools with version numbers, quality thresholds, primer match cutoffs, *etc.*) that were used to obtain the final processed data.
6. The final processed adaptive immune-receptor sequences for the set of samples in the experiment (study) (*e.g.*, the set of sequences used for V(D)J assignment).

Methods and Results. A detailed list of data elements that comprise the minimal standards was identified and revised over the course of several meetings of the MSWG since the last AIRR Community meeting in June 2015. Members of the working group performed literature searches and reached out to existing data repositories (including dbGAP/SRA, IEDB, ImmPort, VDJServer, iReceptor and sciReptor) to obtain their input. Currently the six main elements contain in total about 90 data fields and are defined in **Table 1**. The relationships of these standards with existing database elements are described in **Table 2**.

Requested Feedback from AIRR attendees. The overall goal for this AIRR Community meeting is to develop a vetted checklist of data elements that will comprise the recommended data standard that journals, and ultimately data repositories and the NIH, would use. During the 2016 AIRR Community meeting, the MSWG will seek detailed feedback on:

a) Which standards are required for publication (**Table 3**)
b) Which standards are required for data upload into a repository (**Table 4**)
c) What format the data should take to allow for integrated searching across repositories (data element naming conventions, controlled vocabularies: **Round table with Common Repository Working Group at the end of the session**)