AIRR Common Repository Working Group (CRWG)

•••

AIRR Community Meeting | Sunday, December 3, 2017

Members:

Meredith Ashby
Felix Breden
Richard Bruskiewich
Tania Bubela
Syed Ahmad Chan Bukhari
Scott Christley
Brian Corrie
Lindsay Cowell
John Harting

Rob Holt
David Klatzmann
Uri Laserson
Nishanth Marthandan
Bjoern Peters
Adrien Six
Adrian Thorogood
Corey Watson
Yariv Wine

CRWG’s Mission

To promote and facilitate the creation of common repositories that facilitate open ***deposition***, ***access***, and ***sharing/reuse*** of IG and TCR AIRR-seq datasets.

CRWG’s Goals (2016-2017)

→ Refine and ratify the AIRR CRWG 2016 Recommendations Document

(available here: https://github.com/airr-community/common-repo-wg/blob/master/recommendations.md)

→ Develop consistent consent documents that are compliant with best practices for open sharing of AIRR-seq data.

→ Develop a framework for an AIRR Community “Data Commons” for AIRR-seq datasets.

Progress

→ Refine and ratify the AIRR CRWG 2016 Recommendations Document

Progress

→ Refine and ratify the AIRR CRWG 2016 Recommendations Document

Recommendations 1 - 3: The default data sharing policy should be to deposit data in a public domain database with no restrictions over deposit, access, storage, curation, and use.

Progress

→ Refine and ratify the AIRR CRWG 2016 Recommendations Document

Recommendations 1 - 3: The default data sharing policy should be to deposit data in a public domain database with no restrictions over deposit, access, storage, curation, and use.

Recommendations 5, 10 - 11: Dedicated AIRR repositories should be established for hosting processed repertoire-sequencing data and annotations to facilitate data queries and cross-study meta-analyses. An intermediate distributed architecture is recommended.

Progress

→ Refine and ratify the AIRR CRWG 2016 Recommendations Document

Recommendations 1 - 3: The default data sharing policy should be to deposit data in a public domain database with no restrictions over deposit, access, storage, curation, and use.

Recommendations 5, 10 - 11: Dedicated AIRR repositories should be established for hosting processed repertoire-sequencing data and annotations to facilitate data queries and cross-study meta-analyses. An intermediate distributed architecture is recommended.

Recommendations 4, 6 - 9: Compliant repositories will adhere to operational criteria put forth by AIRR working groups.

Progress

→ Refine and ratify the AIRR CRWG 2016 Recommendations Document

Recommendations 1 - 3: The default data sharing policy should be to deposit data in a public domain database with no restrictions over deposit, access, storage, curation, and use.

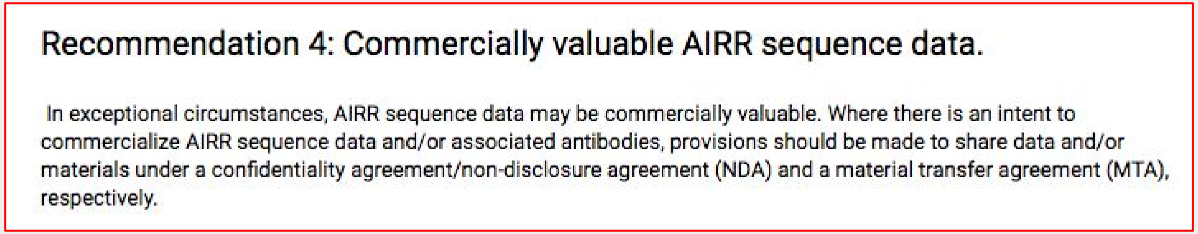
Recommendations 5, 10 - 11: Dedicated AIRR repositories should be established for hosting processed repertoire-sequencing data and annotations to facilitate data queries and cross-study meta-analyses. An intermediate distributed architecture is recommended.

Recommendations 4, 6 - 9: Compliant repositories will adhere to operational criteria put forth by AIRR working groups.

Recommendations 12 - 13: AIRR sequencing studies should also deposit data in IEDB and ImmPort, where appropriate.

Progress

→ Refine and ratify the AIRR CRWG 2016 Recommendations Document

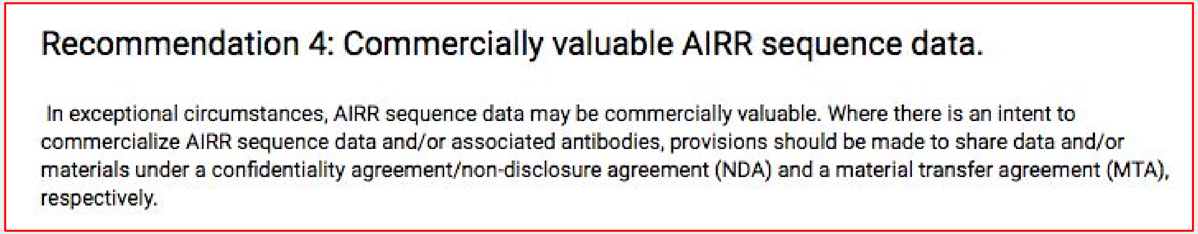


Recommendation 4: Commercially valuable AIRR sequence data.

In exceptional circumstances, AIRR sequence data may be commercially valuable. Where there is an intent to commercialize AIRR sequence data and/or associated antibodies, provisions should be made to share data and/or materials under a confidentiality agreement/non-disclosure agreement (NDA) and a material transfer agreement (MTA), respectively.

Progress

→ Refine and ratify the AIRR CRWG 2016 Recommendations Document



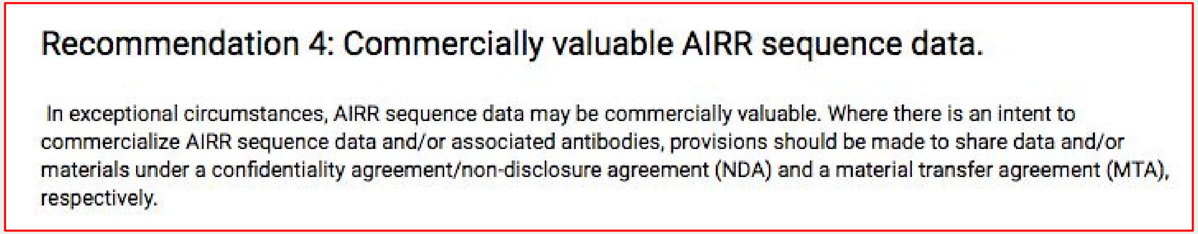
Recommendation 4: Commercially valuable AIRR sequence data.

In exceptional circumstances, AIRR sequence data may be commercially valuable. Where there is an intent to commercialize AIRR sequence data and/or associated antibodies, provisions should be made to share data and/or materials under a confidentiality agreement/non-disclosure agreement (NDA) and a material transfer agreement (MTA), respectively.

In January of 2017, we held a vote on a modified CRWG Recommendations document.

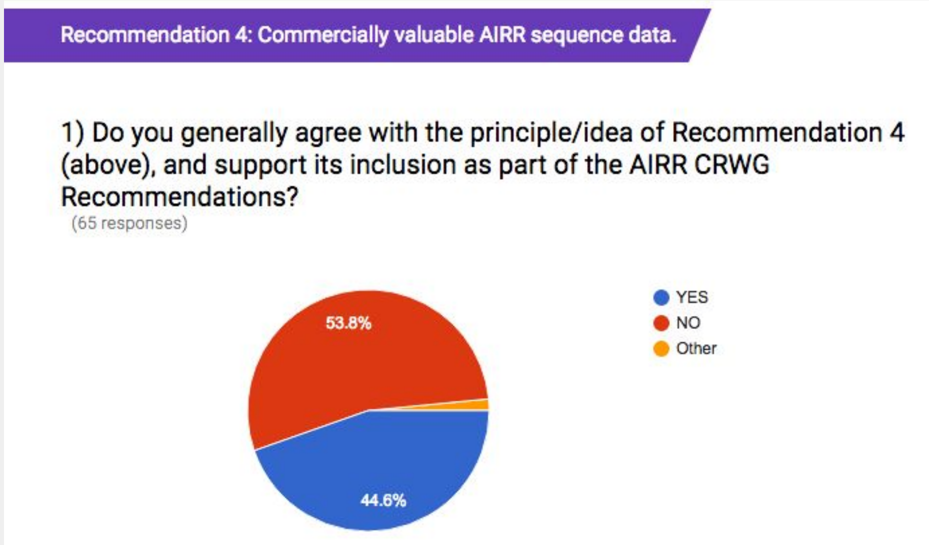
3 questions; 65 respondents

Progress



Recommendation 4: Commercially valuable AIRR sequence data.

In exceptional circumstances, AIRR sequence data may be commercially valuable. Where there is an intent to commercialize AIRR sequence data and/or associated antibodies, provisions should be made to share data and/or materials under a confidentiality agreement/non-disclosure agreement (NDA) and a material transfer agreement (MTA), respectively.

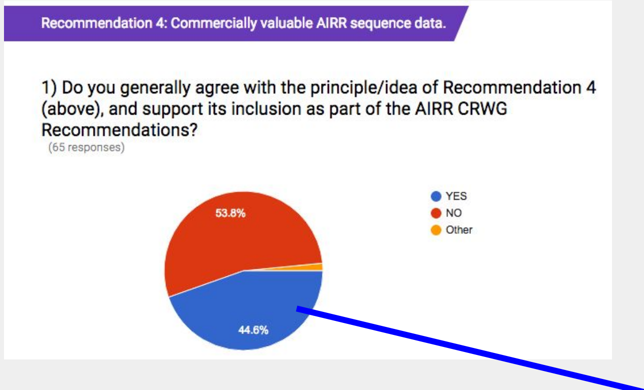


Recommendation 4: Commercially valuable AIRR sequence data.

1) Do you generally agree with the principle/idea of Recommendation 4 (above), and support its inclusion as part of the AIRR CRWG Recommendations?

(65 responses)

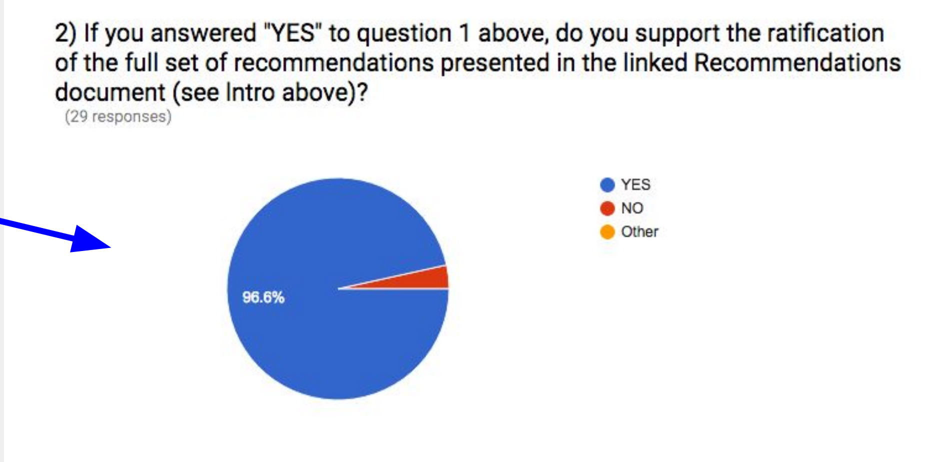
Progress



Recommendation 4: Commercially valuable AIRR sequence data.

1) Do you generally agree with the principle/idea of Recommendation 4 (above), and support its inclusion as part of the AIRR CRWG Recommendations?

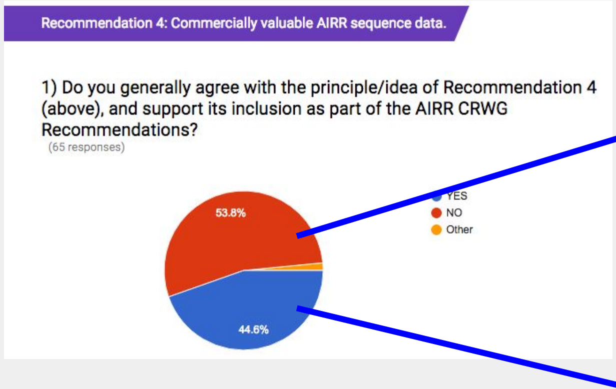
(65 responses)



2) If you answered "YES" to question 1 above, do you support the ratification of the full set of recommendations presented in the linked Recommendations document (see Intro above)?

(29 responses)

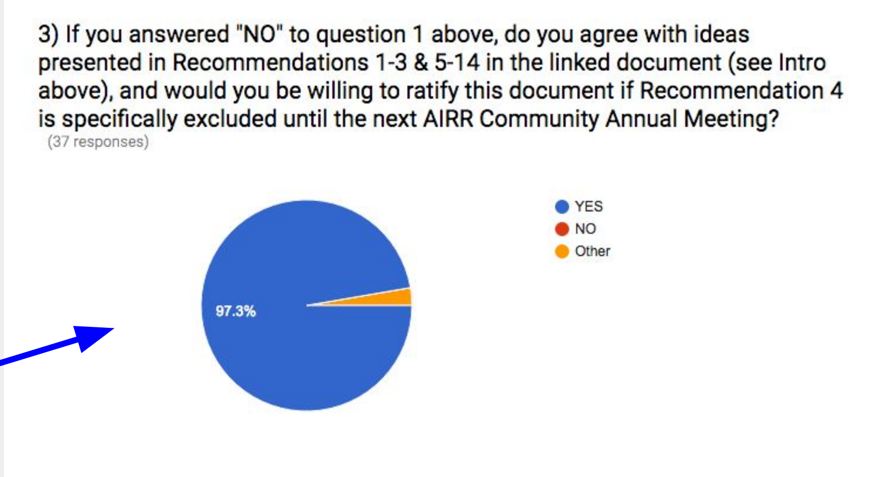
Progress



Recommendation 4: Commercially valuable AIRR sequence data.

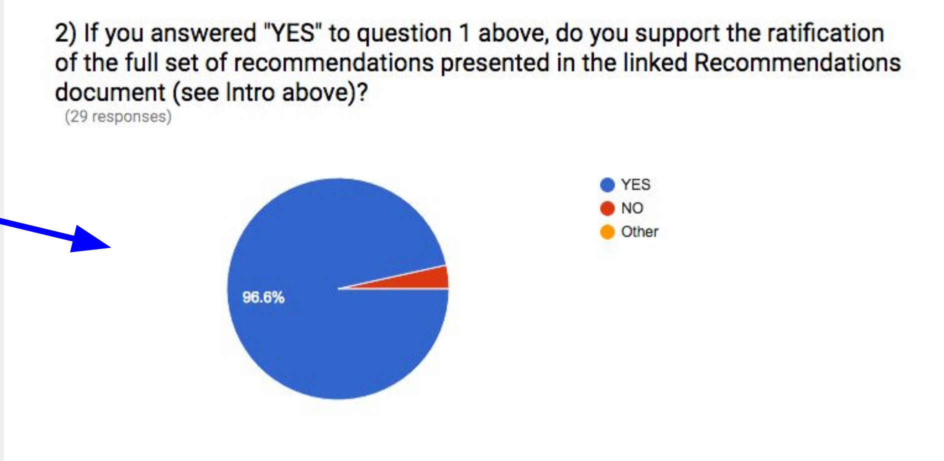
1) Do you generally agree with the principle/idea of Recommendation 4 (above), and support its inclusion as part of the AIRR CRWG Recommendations?

(65 responses)



3) If you answered "NO" to question 1 above, do you agree with ideas presented in Recommendations 1-3 & 5-14 in the linked document (see Intro above), and would you be willing to ratify this document if Recommendation 4 is specifically excluded until the next AIRR Community Annual Meeting?

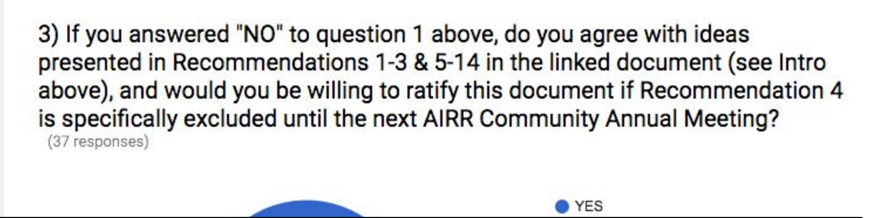
(37 responses)



2) If you answered "YES" to question 1 above, do you support the ratification of the full set of recommendations presented in the linked Recommendations document (see Intro above)?

(29 responses)

Progress



3) If you answered "NO" to question 1 above, do you agree with ideas presented in Recommendations 1-3 & 5-14 in the linked document (see Intro above), and would you be willing to ratify this document if Recommendation 4 is specifically excluded until the next AIRR Community Annual Meeting?

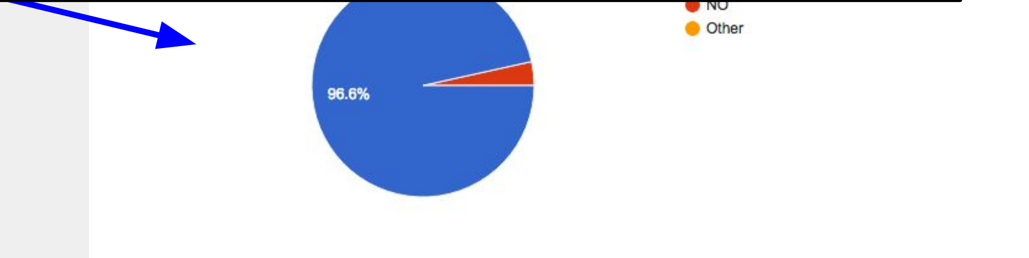
(37 responses)

SUMMARY:

Majority against the inclusion of “Recommendation 4”

(debate tabled until Monday, December 4)

Majority in favor of Recommendations document (sans “Rec. 4”)



Progress

→ Develop consistent consent documents that are compliant with best practices for open sharing of AIRR-seq data.

Progress

→ Develop consistent consent documents that are compliant with best practices for open sharing of AIRR-seq data.

1. Where data sharing presents a risk to individual privacy, data producers should seek **“broad consent”**: consent to access by qualified researchers for future, unspecified purposes with appropriate governance. The ethical basis of broad consent depends on the following elements:
   - Appropriate privacy and security safeguards are in place across the data trajectory
   - Participants are informed of limits to anonymity and confidentiality.
   - Participants are informed of governance arrangements

Progress

→ Develop consistent consent documents that are compliant with best practices for open sharing of AIRR-seq data.

1. Where data sharing presents a risk to individual privacy, data producers should seek **“broad consent”**: consent to access by qualified researchers for future, unspecified purposes with appropriate governance. The ethical basis of broad consent depends on the following elements:
    - Appropriate privacy and security safeguards are in place across the data trajectory
    - Participants are informed of limits to anonymity and confidentiality.
    - Participants are informed of governance arrangements

2. When broad consent is not practical because of regulatory requirements, ethical concerns, or the preferences of the research population, an alternative approach is to **maintain a link with the participant to enable specific consent to future access** and re-use of data.

Progress

→ Develop consistent consent documents that are compliant with best practices for open sharing of AIRR-seq data.

1. Where data sharing presents a risk to individual privacy, data producers should seek **“broad consent”**: consent to access by qualified researchers for future, unspecified purposes with appropriate governance. The ethical basis of broad consent depends on the following elements:
   - Appropriate privacy and security safeguards are in place across the data trajectory
   - Participants are informed of governance arrangements
   - Participants are informed of limits to anonymity and confidentiality.

2. When broad consent is not practical because of regulatory requirements, ethical concerns, or the preferences of the research population, an alternative approach is to **maintain a link with the participant to enable specific consent to future access** and re-use of data.

3. Consent and governance processes should be transparent.

Progress

→ Develop consistent consent documents that are compliant with best practices for open sharing of AIRR-seq data.

1. Consent Elements - Prospective Consent to Data Sharing (GA4GH Consent Policy 2015)
2. Sharing Legacy Data - When is Re-consent, Notification, or Ethics Waiver Required? (GA4GH Consent Policy 2015)
3. Consent to International Data Sharing: Template Consent Forms and Clauses

Progress

→ Develop a framework for an AIRR Community “Data Commons” for AIRR-seq datasets.

Why bother? We can all just query SRA/GenBank …

1. The query interfaces to NCBI repositories are not optimized for AIRR-relevant queries.

Why bother? We can all just query SRA/GenBank …

1. The query interfaces to NCBI repositories are not optimized for AIRR-relevant queries.
2. Only raw data in SRA.
   a. You will need to download that data, curate it to properly post-process (e.g., barcodes and primers), and run through an analysis pipeline.
   b. Many users do not want to duplicate this effort.
   c. Many users do not have the resources or expertise to do this.
   d. Many queries will be more efficient on processed data.

Why bother? We can all just query SRA/GenBank …

1. The query interfaces to NCBI repositories are not optimized for AIRR-relevant queries.

2. Only raw data in SRA.
   a. You will need to download that data, curate it to properly post-process (e.g., barcodes and primers), and run through an analysis pipeline.
   b. Many users do not want to duplicate this effort.
   c. Many users do not have the resources or expertise to do this.
   d. Many queries will be more efficient on processed data.

3. It is unclear if GenBank will host all processed data (e.g., unproductive). Genbank will only store a small subset of annotations.

Why bother? We can all just query SRA/GenBank …

1. The query interfaces to NCBI repositories are not optimized for AIRR-relevant queries.

2. Only raw data in SRA.
   a. You will need to download that data, curate it to properly post-process (e.g., barcodes and primers), and run through an analysis pipeline.
   b. Many users do not want to duplicate this effort.
   c. Many users do not have the resources or expertise to do this.
   d. Many queries will be more efficient on processed data.

3. It is unclear if GenBank will host all processed data (e.g., unproductive). Genbank will only store a small subset of annotations.

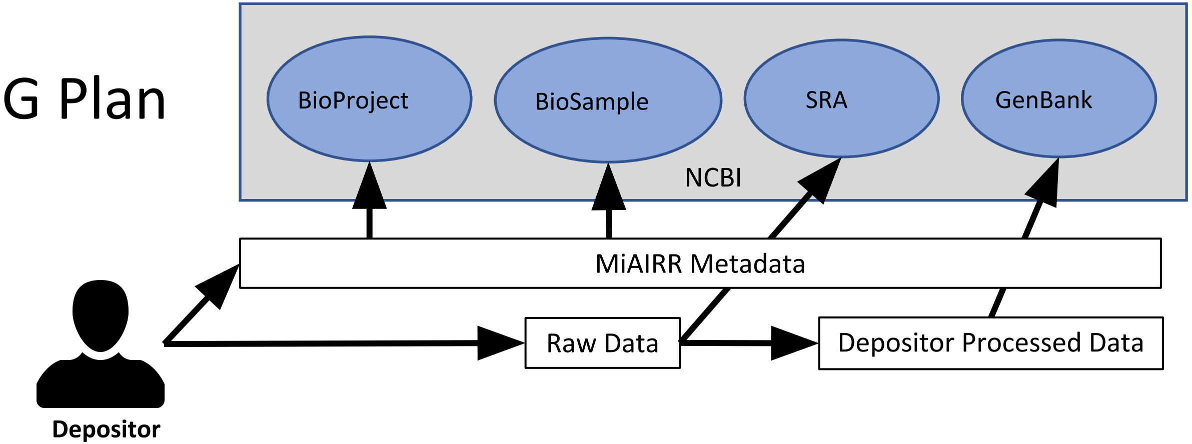
4. There is a desire to run different processing pipelines (e.g., with different germline databases, VDJ calling algorithms, etc.) - Genbank will only provide the depositor version. AIRR repositories can provide standardized processing pipelines for specific applications.

Progress

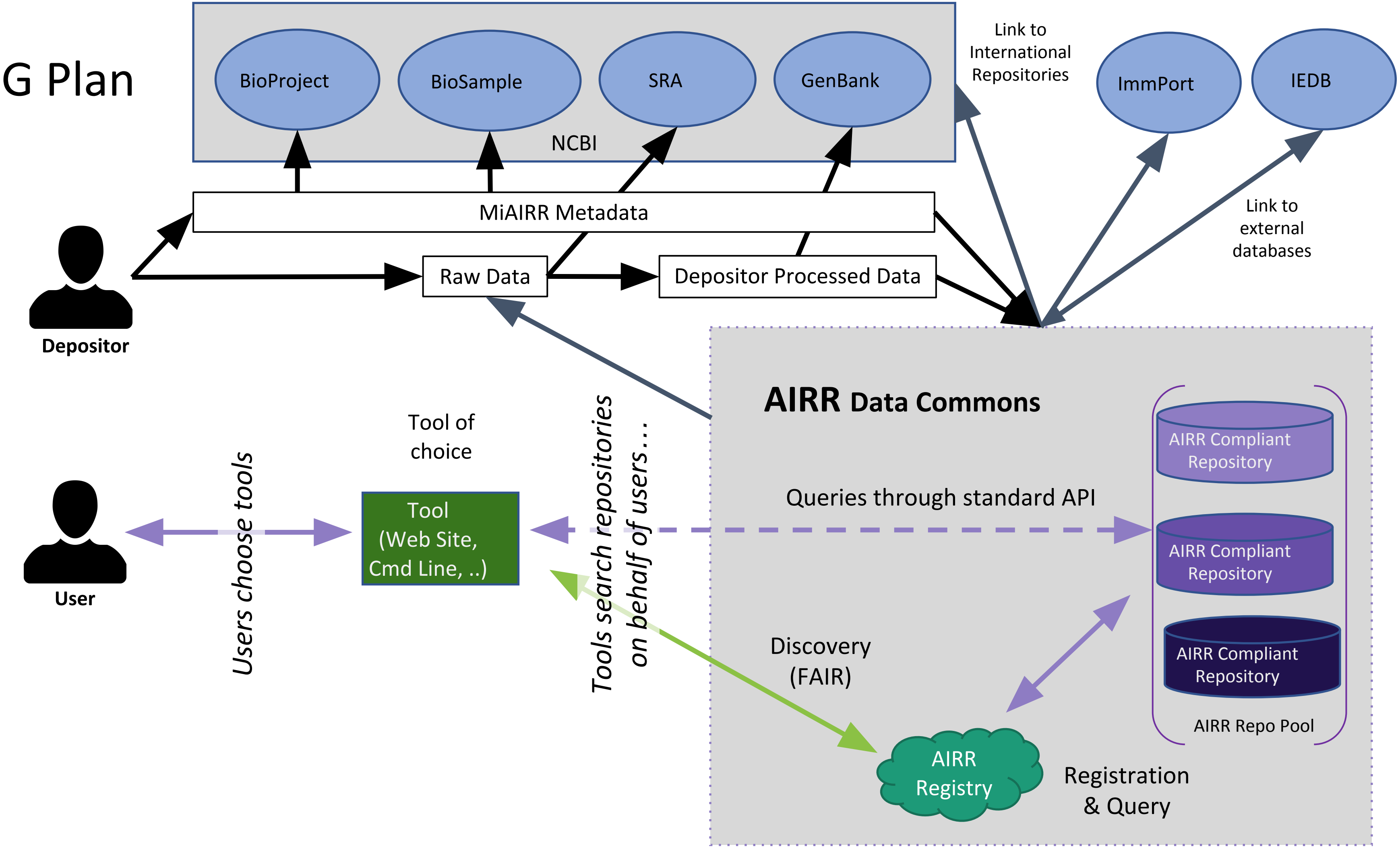
→ Develop a framework for an AIRR Community “Data Commons” for AIRR-seq datasets.

Recommendation 10: The dedicated AIRR repositories (Recommendation 5) should comprise a system of ***multiple, distributed repositories*** supported by a centralized registry consistent with an intermediate distributed model as described in http://science.sciencemag.org/content/350/6266/1312.full.

CRWG Plan


BioProject
BioSample
SRA
GenBank
NCBI
MiAIRR Metadata
Raw Data
Depositor Processed Data
Depositor

CRWG Plan



BioProject

BioSample

SRA

GenBank

NCBI

Link to International Repositories

ImmPort

IEDB

Link to external databases

MiAIRR Metadata

Raw Data

Depositor Processed Data

Depositor

AIRR Data Commons

AIRR Compliant Repository

AIRR Compliant Repository

AIRR Compliant Repository

AIRR Repo Pool

Queries through standard API

Tool of choice

Tool (Web Site, Cmd Line, ..)

User

Users choose tools

Tools search repositories on behalf of users…

Discovery (FAIR)

AIRR Registry

Registration & Query

Progress

→ Develop a framework for an AIRR Community "Data Commons" for AIRR-seq datasets.

Recommendation 7: The AIRR Working Groups should collaboratively develop **operational criteria for compliant repositories**. … The operational criteria should include implementation of:

1. Standardized data elements with exact (computable) specifications;
2. A standardized data submission process (including standardized data and metadata formats);
3. A standardized set of queries;
4. A system for assigning unique identifiers that ensures coordination among repositories/registritries ...

Example queries - what do users want to query for?

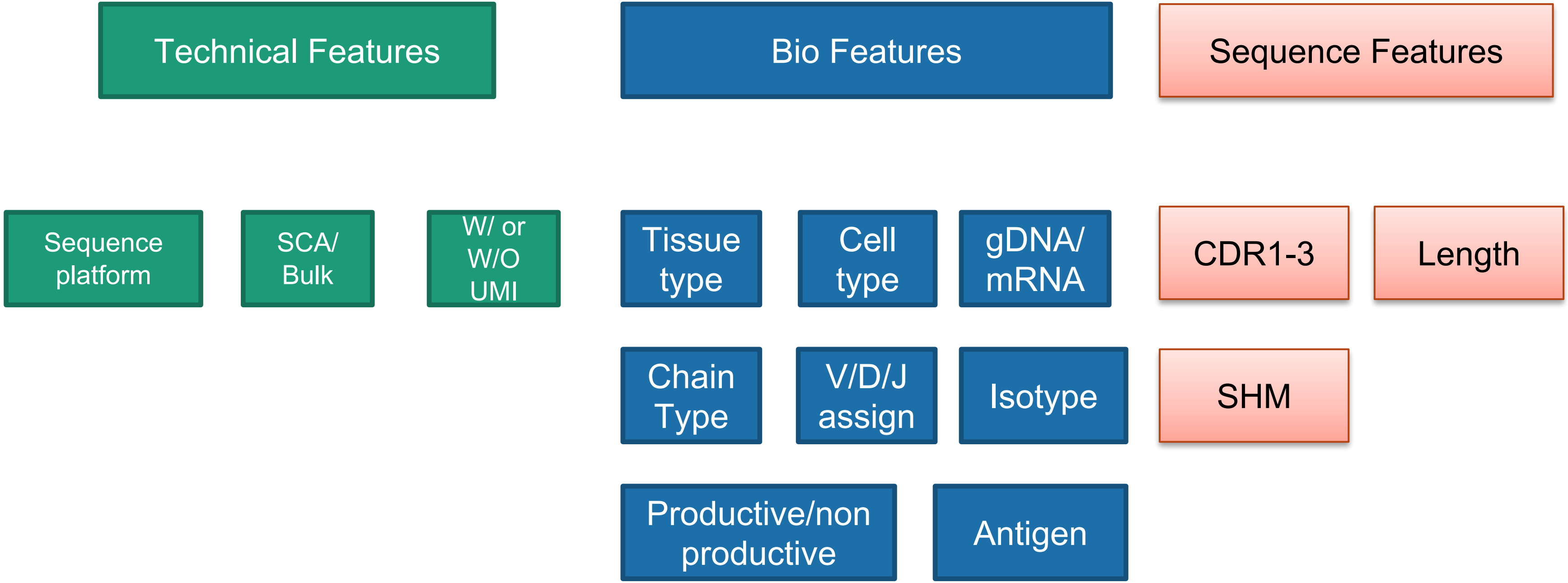
- What human full length TCR-beta sequences have CDR3 region: “GTGGTNEKL”?
- What human full length IgH sequences have been found in patients with an autoimmune diagnosis?
- What is the antibody IG heavy chain V usage in people who have diabetes?
- Give me all the anti-HIV antibody sequences that use IGHV1-69 in HIV infected individuals?
- Return repertoires from cancer patients where we have pre- and post-immunotherapy peripheral blood (or tumor biopsy) repertoires.

What human full length TCR-beta sequences have CDR3 region: “GTGGTNEKL”?

Query parameters:
Donor species = human
Sequence type = TCR beta chain
Sequence feature: CDR3 = “GTGGTNEKL”

Translation to AIRR terms is non trivial:

- organism = “homo sapiens” → NCBI taxonomy ID or string match?
- AND
- v_call *contains* “TRB” → string matching or hierarchical organization (ontology) of genes
- AND
- junction_aa contains “GTGGTNEKL”



Technical Features

Sequence platform

SCA/ Bulk

W/ or W/O UMI

Bio Features

Tissue type

Cell type

gDNA/ mRNA

Chain Type

V/D/J assign

Isotype

Productive/non productive

Antigen

Sequence Features

CDR1-3

Length

SHM

Key repeating elements to prioritize for computationally precise standardization

Donor species (e.g., homo sapiens)

Donor health status (e.g., diabetes)

Sequence type (e.g., TRB)

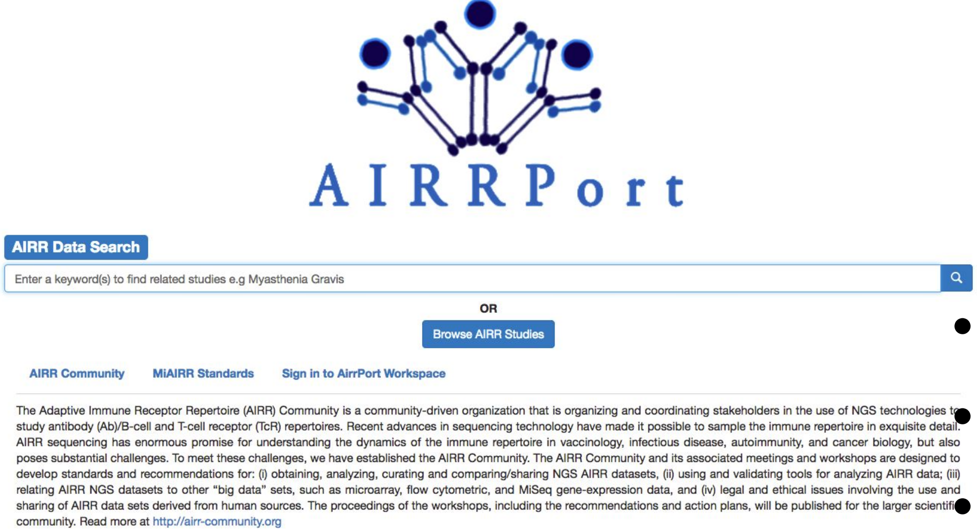
Gene usage (e.g., IGHV1-69)

CDR3 sequence (e.g., “CASSYIKLN”)

Receptor specificity (e.g., HIV virus)

In Progress Implementations

- iReceptor
- VDJServer
- AIRRPort
- IEDB
- Klatzman group



AIRRPort

AIRR Data Search

Enter a keyword(s) to find related studies e.g Myasthenia Gravis

OR

Browse AIRR Studies

AIRR Community   MiAIRR Standards   Sign in to AirrPort Workspace

The Adaptive Immune Receptor Repertoire (AIRR) Community is a community-driven organization that is organizing and coordinating stakeholders in the use of NGS technologies to study antibody (Ab)/B-cell and T-cell receptor (TcR) repertoires. Recent advances in sequencing technology have made it possible to sample the immune repertoire in exquisite detail. AIRR sequencing has enormous promise for understanding the dynamics of the immune repertoire in vaccinology, infectious disease, autoimmunity, and cancer biology, but also poses substantial challenges. To meet these challenges, we have established the AIRR Community. The AIRR Community and its associated meetings and workshops are designed to develop standards and recommendations for: (i) obtaining, analyzing, curating and comparing/sharing NGS AIRR datasets, (ii) using and validating tools for analyzing AIRR data; (iii) relating AIRR NGS datasets to other "big data" sets, such as microarray, flow cytometric, and MiSeq gene-expression data, and (iv) legal and ethical issues involving the use and sharing of AIRR data sets derived from human sources. The proceedings of the workshops, including the recommendations and action plans, will be published for the larger scientific community. Read more at http://airr-community.org

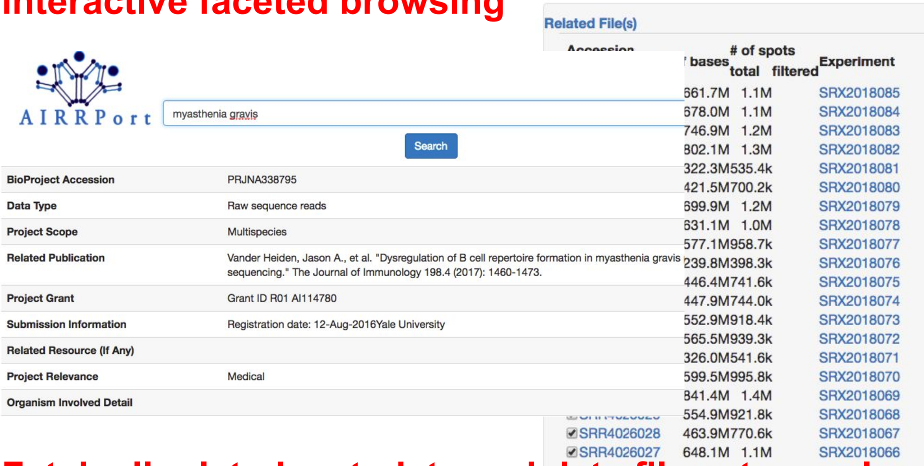
https://airrport.org

Salient Features

- A centralized portal to discover AIRR Studies
- Currently, it fetches AIRR data from the NCBI
- You can register your data through BioProjectID
- You can make your already submitted data to

  MiAIRR compliant through user-friendly interface
- AIRRPort is a prototype yet.
- Future work will include more AIRR related repos

interactive faceted browsing



AIRRPort

myasthenia gravis

Search

BioProject Accession	PRJNA338795
Data Type	Raw sequence reads
Project Scope	Multispecies
Related Publication	Vander Heiden, Jason A., et al. "Dysregulation of B cell repertoire formation in myasthenia gravis sequencing." The Journal of Immunology 198.4 (2017): 1460-1473.
Project Grant	Grant ID R01 AI114780
Submission Information	Registration date: 12-Aug-2016Yale University
Related Resource (If Any)	
Project Relevance	Medical
Organism Involved Detail	

Related File(s)

Accession	bases	# of spots total	filtered	Experiment
	661.7M	1.1M		SRX2018085
	678.0M	1.1M		SRX2018084
	746.9M	1.2M		SRX2018083
	802.1M	1.3M		SRX2018082
	322.3M	535.4k		SRX2018081
	421.5M	700.2k		SRX2018080
	699.9M	1.2M		SRX2018079
	631.1M	1.0M		SRX2018078
	577.1M	958.7k		SRX2018077
	239.8M	398.3k		SRX2018076
	446.4M	741.6k		SRX2018075
	447.9M	744.0k		SRX2018074
	552.9M	918.4k		SRX2018073
	565.5M	939.3k		SRX2018072
	326.0M	541.6k		SRX2018071
	599.5M	995.8k		SRX2018070
	841.4M	1.4M		SRX2018069
	554.9M	921.8k		SRX2018068
☑SRR4026028	463.9M	770.6k		SRX2018067
☑SRR4026027	648.1M	1.1M		SRX2018066

Fetch all related metadata and data files at one place

so what you think about AIRRPort tell us at twitter
@airrportdb.
To discuss further contact: {Ahmad.chan, steven.kleinstein}@yale.edu

Progress

→ Develop a framework for an AIRR Community "Data Commons" for AIRR-seq datasets.

- I want to put my data somewhere … I want to host an AIRR-compliant repository.
- "Hosting" a repository means you
  - can run a database that scales with your data size,
  - can run a public HTTP/REST API server that responds to queries, and
  - (optionally) have a batch submission/queue system for performing large queries.

Progress

→ Develop a framework for an AIRR Community “Data Commons” for AIRR-seq datasets.

- Tier 1: Have the infrastructure to host a large repository with data from multiple labs and institutions.
- Tier 2: Have the infrastructure to host a repository with my institution’s data.
- Tier 3: Have the infrastructure to host a repository with my lab’s data.
- Tier 4: Unable to host a repository.

Goals for the coming year

- Formalize and document the computable specifications for query elements
- Identify a query language
- Decide what should be returned for each query
  - Download locally
  - Transfer to an analysis server (skip intervening download/upload steps)